I.J. Information Engineering and Electronic Business, **2025, 3, 99-153**
Published Online on June 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2025.03.07

Agile Technology of Information Data Engineering for Intelligent Analysis of the Happiness Index and Life Satisfaction in Known World Cities

Yuriy Ushenko*
Department of Computer Science, Educational and Research Institute of Physical, Technical and Computer Sciences, Yuriy Fedkovych Chernivtsi National University, 58012, Ukraine
E-mail: y.ushenko@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0003-1767-1882
*Corresponding Author

Victoria Vysotska
Department of Information Systems and Networks, Institute of Computer Sciences and Information Technologies, Lviv Polytechnic National University, Lviv, 79013, Ukraine
E-mail: Victoria.A.Vysotska@lpnu.ua
ORCID iD: https://orcid.org/0000-0001-6417-3689

Daryna Zadorozhna
Department of Information Systems and Networks, Institute of Computer Sciences and Information Technologies, Lviv Polytechnic National University, Lviv, 79013, Ukraine
E-mail: daryna.zadorozhna.sa.2022@lpnu.ua
ORCID iD: https://orcid.org/0009-0006-2711-6865

Mariia Spodaryk
Department of Information Systems and Networks, Institute of Computer Sciences and Information Technologies, Lviv Polytechnic National University, Lviv, 79013, Ukraine
E-mail: mariia.spodaryk.sa.2022@lpnu.ua
ORCID iD: https://orcid.org/0000-0002-2829-0164

Zhengbing Hu
School of Computer Science, Hubei University of Technology, Wuhan, China
E-mail: drzbhu@gmail.com
ORCID iD: https://orcid.org/0000-0002-6140-3351

Dmytro Uhryn
Department of Computer Science, Educational and Research Institute of Physical, Technical and Computer Sciences, Yuriy Fedkovych Chernivtsi National University, 58012, Ukraine
E-mail: d.ugryn@chnu.edu.ua
ORCID iD: https://orcid.org/ 0000-0003-4858-4511

Received: 24 January, 2025; Revised: 01 March, 2025; Accepted: 05 April, 2025; Published: 08 June, 2025

Abstract: This paper presents the development of an intelligent information system for analysing the happiness index and life satisfaction based on sociological survey data from various countries. The research addresses the need to improve the accuracy and efficiency of social research by integrating data mining and machine learning methods – specifically K-means clustering and multiple regression analysis – into the system design. The proposed module enables automated classification of countries and cities by life satisfaction levels, allowing stakeholders to make informed decisions on urban planning and social policy. The system also facilitates the identification of favourable living environments, providing valuable insights into the social, economic, and environmental factors affecting well-being. The experimental results on real-world datasets confirm the module's effectiveness and predictive capabilities.

Index Terms: Happiness Index; Life Satisfaction; Intelligent Analysis; Sociological Data; K-Means Clustering; Multiple Regression; Favourable Living Environment; Data Mining; Urban Development; Decision Support System

1. Introduction

In the conditions of the rapid pace of informatization of society, successes in the sociological science development are closely related to the development of innovative directions of intelligent data analysis [1-3]. A promising direction for increasing the efficiency of sociological research is the use of smart systems for analysing sociological data obtained during surveys, which allow for automated processing and analysis of primary sociological information [4-6]. Among the existing approaches to collecting and analysing sociological information, two main approaches can be conditionally distinguished [7-10]. The first traditional approach involves developing a questionnaire and manually conducting a survey with subsequent input and processing of information using general-purpose software tools, such as Microsoft Excel and Google Sheets, or specialized software (Statistica, SPSS, MatLab). This approach is very laborious and does not provide prompt adjustment of data parameters and their required quality, preliminary analysis of survey data, and reliability of operations. The second approach involves conducting surveys taking into account the organization's web systems (SurveyMonkey, Google Forms). At the same time, the researcher's target audience is Internet users. However, it is pretty challenging to ensure the quality, efficiency of the survey, and representativeness of the sample. The data analysis stage must be performed in a third-party program. Therefore, there is a need to develop ways to increase the correctness of the results of sociological research based on improving the quality of measuring empirical information through the creation of a special automated information system with built-in methods of intelligent data analysis. It determined the goal of the work, which is to increase the efficiency of conducting sociological research by developing a smart system for analysing sociological data using classification and clustering algorithms. In accordance with the goal, the following tasks were formulated:

- to investigate the theoretical foundations of sociological research and methods of processing and analysing sociological information;
- to justify the choice of tools for developing a system for analysing sociological data;
- to develop and implement a software implementation of an intelligent system for analysing the level of happiness based on the processing of data from sociological surveys in different countries of the world.

The study hypothesises that the integration of machine learning methods (in particular, K-means clustering algorithms and multiple linear regression) into a specialised information system will allow automated, objective and practical analysis of the level of happiness and satisfaction with life in cities and countries, which will improve decision-making in the field of social policy and urban planning.

The main research question of the study is how to increase the efficiency of analysing sociological data on the level of happiness and life satisfaction in different countries and cities using machine learning and clustering. This question focuses on how combining machine learning techniques (clustering, regression) with automated processing of sociological data can help improve the accuracy of social welfare analysis.

These formulations are based on a well-defined research objective: to develop an intelligent system for the analysis of sociological data, as well as on the tasks, methods and justifications of the research.

The development of a module for determining a favourable environment for human life and development is becoming increasingly relevant due to growing urbanization and the complexity of problems related to the urban environment. In the modern world, the happiness index is considered an important indicator that helps determine the level of comfort and well-being of the population in different cities and countries. This index helps to understand what factors affect people's quality of life. It can also provide information for making decisions regarding policy, infrastructure, education, medicine and other aspects of urban development. It can contribute to improving living conditions and creating a favourable environment for human development in cities. When developing the module, machine learning methods such as linear regression, clustering, correlation analysis and others can be applied, which allow you to obtain predictions and categories based on existing data. It will enable you to identify patterns and trends that affect the happiness index and help you understand the characteristics of different places, which can be helpful for urban development planning. In addition, this module can contribute to the involvement of the public and stakeholders in the decision-making process, as well as raise public awareness of the importance of an enabling environment for living and development in cities. Overall, the development of a module for determining the enabling environment for living and development of people in a town is of great importance to society. It can help local and national governments, researchers, organizations and other stakeholders identify problems and develop strategies for urban planning, infrastructure, social services and other essential aspects of urban development. In particular, using such a module, it is possible to conduct analysis not only by country but also by city if the relevant data are available. It will allow us to obtain a more detailed picture of the standard of living and well-being of the population in specific regions and cities, which will help to better understand the specifics of local conditions and respond to the identified problems accordingly for countries such as Ukraine, the development and application of such a module may be of particular importance, as it

will contribute to improving the quality of life and ensuring the sustainable development of urban environments. The application of machine learning methods and data analysis in this area may open up new opportunities for information support for decision-making at all levels of city management and state structures. The goal of the project is to develop a module that uses machine learning methods to analyse various factors that affect the quality of life and level of happiness of people and to determine favourable conditions for creating an environment that promotes human development in cities. Therefore, to achieve the goal, the following tasks need to be solved:

- review the principles and stages of determining a favourable environment for human life and development;
- analyse existing approaches to determining a favourable environment for human life and development;
- formulate tasks for developing a module for determining a favourable environment for human life and development based on the happiness index;
- develop algorithms based on machine learning to determine a favourable environment for human life and development based on the happiness index;
- implement a software module for determining a favourable environment for human life and development based on the happiness index;
- verify the effectiveness and accuracy of the developed module on real data.

The focus of this study is on exploring the interconnections among multiple factors – economic, social, environmental, and infrastructural – that influence both the quality of life and the overall happiness of the population. Additionally, the study examines machine learning techniques that enable the analysis and prediction of these complex relationships. The subject of the study is the study of machine learning methods, such as K-means clustering and multiple regression models, for analysing and predicting various factors that affect the quality of life and the level of happiness of people and determining a favourable environment for the development of people in cities.

The methodological basis of the study is general scientific methods of data mining, which allows for a comprehensive study of the subject and object of the study to explore the main approaches to processing primary sociological information and its mining analysis using data classification and clustering algorithms.

The research methods in this project are machine learning methods, in particular, clustering and regression analysis methods. For data clustering, the K-means algorithm is used, which allows data to be divided into groups with similar characteristics. Multiple linear regression is used to predict the happiness index using a regression model, which will enable you to find dependencies between several variables and the target value. The study also uses correlation and covariance analysis methods to study the dependencies between variables. The practical value of the study is that it can help identify factors that contribute to improving the quality of life and development of people in cities. The application of machine learning methods, such as K-means and the multiple regression model, allows you to automatically analyse large amounts of data and determine the dependencies between various indicators of quality of life and the level of happiness of the population. The results obtained can be helpful in making decisions on policies and infrastructure that will contribute to improving the quality of life and increasing the level of happiness of the population. In addition, the study can open new opportunities for further research in the field of social policy and psychology.

2. Related works

2.1. Organization and conduct of sociological research in the information society

Sociological research is a system of sequential, logical methodological, methodological and organizational procedures, the purpose of which is to obtain scientific knowledge about a specific social object. Conducting sociological research allows you to confirm guesses and conjectures collect and evaluate information about the phenomena being studied. It is a connecting link between objective reality and theoretical knowledge, which allows you to establish new patterns of development of structural elements or society as a whole [11-12]. According to the goals, sociological research is divided into applied and fundamental. The first study aims to solve specific problems in society. The organization of sociological research of the essential type is carried out in order to determine and analyse social trends and patterns of development, and it is designed to solve complex problems in society. The following types of sociological research are distinguished by duration:

- express – from 1 week to 1 month;
- short-term – 2-6 months;
- medium-term – 0.5-3 years;
- long-term – 3 years or more.

The depth of analysis distinguishes the following types of sociological research:

- Descriptive – allows you to form a holistic picture of the processes and phenomena under study. They are carried out according to a clear program. The object of analysis is mainly a large community of people with specific professional, social and demographic characteristics;

- Exploratory – solve the simplest problems. They are most often used as a preliminary stage of large-scale analysis in cases where the object, subject or program of analysis is poorly studied or not studied at all;
- Analytical – Describe social phenomena and their components, determine the mechanisms of their functioning, and determine the causes of their occurrence. The preparation, organization and conduct of a sociological study of this type require serious efforts and professionalism on the part of the researcher – he must be able to correctly interpret the complex information received and draw informed conclusions.

Conducting and organizing research in sociology is carried out consistently. The stages of analysis, which replace each other, are organizationally autonomous and, at the same time, substantively interconnected:

- preparation for research;
- collection of primary sociological information;
- preparation and processing of collected data;
- analysis of information, summarization, formulation of recommendations and conclusions.

Sociological research is conducted in a particular environment of society and is aimed at finding truly existing facts on the principle of "here and now". Each research activity is unique, but some theoretical generalizations are not excluded. To obtain the necessary social facts, the researcher must penetrate an inevitable social reality - distribute questionnaires to all respondents. If this is a questionnaire survey, explain the content and purpose of the study, the rules for filling out questionnaire data, etc. Expert telephone surveys or interviews require direct communication between researchers and respondents. "Penetration" of the objects under study also occurs when other methods of collecting social facts are used.

In the context of the formation of an information society, methods and means of obtaining and analysing sociological data are changing. It makes it necessary to investigate new, recently emerged problems and find ways to solve them. The formation of a single digital space can be the key to the deterioration of integration processes in society. On the one hand, the Internet has become a very important channel for disseminating information that is trusted by the population. On the other hand, the need for reliable verification of the information received, for the possibility of its critical understanding, has become acute.

To obtain objective, operational, and reliable information about the state of society, it is proposed that timely and objective monitoring of the development of its structures be organized. For monitoring, it is suggested to use a research complex (an automated information system and the associated methodology of sociological research) designed not only for the study of social capital and other structures of society but also for the formation of the necessary culture of thinking, scientifically substantiated adoption of specific management decisions. The key task of improving the measurement system in sociology in this context is the development of a special automated information system to increase the correctness of the results of sociological research based on improving the quality of measurement of empirical information.

On the one hand, there is a need to strengthen the use and development of modern methods of collecting and analysing sociological information modelling social phenomena, including the use of various scaling methods, proximity measurement, pairwise comparisons, comparisons in triads, as well as different software products, models and algorithms for text analysis, data mining, and other modern methods developed in global and domestic practice. On the other hand, the rapid pace of development of information and communication technologies and computing capabilities has led to the development and widespread use of modern methods of Data Mining, the theory of artificial intelligence, and such a branch of science as computational sociology [13]. On the other hand, the rapid pace of development of information and communication technologies and computing capabilities has led to the development and widespread use of modern methods of Data Mining, the theory of artificial intelligence, and such a branch of science as computational sociology [13]. Currently, various large organizations are engaged in research on the level of happiness, using multiple methods of tracking and measuring the level of happiness. One of these methods is the determination of the Legatum Prosperity Index - an index calculated by the Legatum Institute, based on data from surveys by the Gallup Institute, the World Development Indicator, the International Telecommunication Union, the ranking of state failure, world governance indicators, data from WHO, Freedom House, WVS, and the Centre for Systemic Peace. The index is calculated according to the following indicators: economic development, business development, quality of governance, education, health, security and personal freedom, social capital, and the environment. Another method for determining the level of happiness in the world and countries is Happiness and Life Satisfaction (authors Ortiz-Ospina E., Rosera M.). They calculate several indicators: subjective life satisfaction and the number of people who call themselves happy. These researchers pay great attention to the dynamics of these indicators in different countries. Another major study of the level of happiness in the world is called the Happy Planet Index (HPI). This indicator is calculated using the formula:

$$I_{HPI} = \frac{i_{life\ expectancy} * i_{well-being\ subjectivity} * i_{social\ inequality}}{i_{ecology\ state}} \quad (1)$$

Most countries that are in high positions in other happiness rankings are not in the highest positions in this ranking because, despite the high indicator of subjective well-being, low indicators of social inequality and high life expectancy in this country can be extremely polluted environmental environment. The dynamics of changes in the level of happiness in different countries and regions are different due to the various levels of economic, ecological and social development of countries. Fig. 1 shows the dynamics of changes in the level of happiness in the USA and India using the integral index of the World Happiness Report for the period from 2012 to 2025.

World Happiness Report – The World Happiness Report is the most famous and popular report of the UN Sustainable Development Solution Network (UN Sustainable Development Solutions Network). This report, based on Gallup poll data and statistical materials, provides an integral indicator of the level of happiness in a particular country or region. It is calculated using GDP per capita, healthy life expectancy, freedom to choose a life path, generosity/trust, level of perception of corruption, and level of expression of positive and negative emotions. Among the existing approaches to collecting and analysing sociological information, two main approaches can be distinguished. Let us describe them in more detail. The survey will be conducted using the traditional method, which is based on developing a questionnaire, manually performing the study, and entering and processing information using software. As a rule, at the processing stage, either widely available general-purpose software tools (such as spreadsheet packages with a number of mathematical functions, Microsoft Excel, Google Sheets) or specialized software (Statistica, SPSS, MatLab) are used. With this approach, it is impossible to ensure prompt adjustment of sampling parameters, ensure the necessary quality of the sample, conduct preliminary analysis of survey data, and ensure the reliability of operations performed by the interviewer. It should also be noted that it is difficult to ensure the absence of errors during data entry, and the entry operation itself is very laborious.

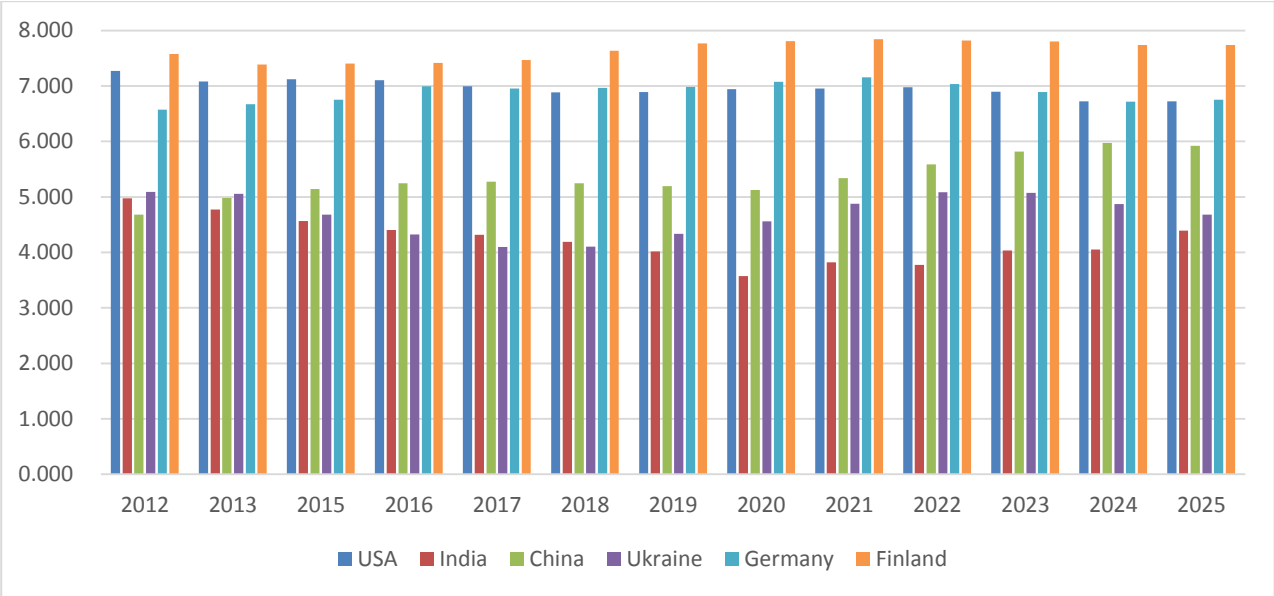


Fig. 1. Dynamics of changes in the happiness index for some countries of the world for 2012-2025 (https://worldhappiness.report/analysis/)

Conducting taking into account web-based systems for organizing surveys (SurveyMonkey, Google Forms). At the same time, the target audience of the researcher is Internet users (which significantly worsens the quality of the sample). Still, it isn't easy to ensure the quality, efficiency of conducting the survey, and representativeness of the sample. In addition, the data analysis stage must be performed in a third-party program.

Thus, there is a need to develop ways to increase the correctness of the results of sociological research based on improving the quality of empirical information processing through the creation of a special automated information system with built-in methods of intelligent data analysis.

We will conduct a detailed analysis of well-known indices such as the World Happiness Report, OECD Better Life Index, and Happy Planet Index and compare them with the proposed approach. Below is a detailed critical analysis of the mentioned happiness indices in the context of their comparison with the model developed in the study.

A critical review of happiness indices and their comparison with the research model

1. The World Happiness Report (WHR) is used in the study as the basis of the data. The metric is based on Gallup World Polls with a large sample in more than 150 countries. It takes into account both objective and subjective factors: GDP, life expectancy, corruption, social support, etc. The index has a high regularity of updating and wide international recognition. The weak side is that happiness ratings are subjective due to the "Cantril Ladder" scale (from 0 to 10), which depends on culture, mood, and mentality. Globalised parameters do not take into account local cultural priorities (for example, spirituality or family values). Limited internal structure – all factors add up without complex interactions. The proposed model in the study tries to cluster countries according to the similarity of features, that is, it identifies

structural types of countries, and does not just rank them. The possibility of further classification and forecasting, which is absent in the WHR, is taken into account.

2. The OECD Better Life Index (BLI) covers a broader set of factors: housing, employment, work/life balance, education, safety, etc. The metric allows individual weighting of factors through the OECD interface. However, it is focused mainly on OECD member countries, most of which are economically developed, so the method is unsuitable for global analysis. It also does not contain a single integral index, which complicates comparisons between countries. The proposed system automates the detection of patterns and allows you to train classification algorithms, while the OECD BLI is a comparison tool, not an analytics or forecasting tool.

3. The Happy Planet Index (HPI) takes into account the environmentally sustainable component – the ecological footprint combined with well-being. The metric focuses less on the economy and more on quality of life and longevity. However, it does not include key social aspects: freedom of choice, corruption, and social support. It also has a narrow set of variables, which does not give a complete picture of happiness as a complex phenomenon. The proposed model is more flexible and extensible, allowing the integration of additional variables, including environmental ones, if necessary.

Table 1. Happiness indices and their comparison with the research model

Criterion	WHR	BLI	HPI	Proposed model
Happiness Data	Yes	No (indirectly)	Yes	Yes
Subjective factors	Yes	Partially	Yes	Yes
Objective indicators	Yes	Yes	Partially	Yes
Global Reach	Yes	Limited (OECD)	Yes	Yes
Integral index	Yes	No	Yes	Yes (with regression)
Forecasting/Classification	No	No	No	Yes
Adapting to machine learning	Limited	Limited	No	Full

The table clearly highlights the strengths/weaknesses of the existing indices. The proposed model is an analytics tool capable of adapting to new variables, performing clustering, forecasting and integration with information systems. Existing indices provide a base, and the proposed system offers flexible algorithmic tools for processing and analysing this data.

2.2. Principles and stages of determining a favourable environment for human life and development

International organizations and research institutions calculate various happiness indices using a range of methodologies. Among these, the most recognized and widely adopted alternative to GDP is the Human Development Index (HDI). This index incorporates not only GDP per capita but also key indicators such as access to education, life expectancy, and several other socio-economic factors. The HDI is compiled annually by analysts from the United Nations Development Programme (UNDP) in collaboration with a team of independent international experts. The index is based on statistical data from national agencies and international bodies, supplemented by analytical research. UNDP has been publishing Human Development Reports since 1990. In creating the HDI rankings, a wide array of elements is considered – including the state of human rights and civil liberties, opportunities for civic participation, access to social protection, levels of territorial and social mobility, cultural development, access to healthcare and information, crime rates, and more. Despite its strengths and comprehensive scope, the HDI also has notable limitations. It relies on national averages, which fail to capture inequality in the distribution of resources. Additionally, it overlooks environmental factors and does not account for the spiritual and moral dimensions of human development. As a result, although the HDI uses a broad set of indicators, there are ongoing efforts to refine the methodology and develop a more accurate and universally applicable measure of happiness.

One notable attempt to develop an alternative happiness metric was made in 2006. The British research organization New Economics Foundation, in collaboration with several international bodies and independent experts, introduced the Happy Planet Index (HPI). This composite indicator evaluates how effectively countries and regions around the world enable their citizens to live happy, fulfilling lives. The index is grounded in the belief that people inherently strive to live long, healthy lives free from hardship and that governments should prioritize the well-being of their citizens while managing environmental resources responsibly. Based on this philosophy, the HPI incorporates three core elements: subjective well-being (life satisfaction), life expectancy, and environmental impact – measured through the ecological footprint per capita, which reflects the biologically productive land required to sustain consumption. In essence, the purpose of the index is to assess how well public policy balances environmental stewardship with human well-being.

The primary objective behind the creation of the World Happiness Index was to measure key aspects of well-being, including income, housing, employment, environment, education, work-life balance, safety, life satisfaction, civic involvement, health, and sense of community. In the standard model, each of these sub-indicators is given equal weight. Additionally, numerous global rankings of happiness levels are based on population surveys conducted using a wide range of methodologies. One of the most well-known is the “Quality of Life Index” or “Well-being Index,” whose methodology was developed by psychologist and Nobel laureate Daniel Kahneman. The data for this index is drawn from the global Gallup survey and incorporates a variety of indicators such as personal health, optimism, fulfilment of

basic societal needs, civic participation, trust in institutions, youth development, corruption perception, and more. The survey is conducted at the national level with a consistent core questionnaire across all countries. Respondents answer a series of questions regarding their housing, access to food, rule of law, personal financial situation, health status, and level of trust in governmental institutions. Based on the responses, the population is categorized into three groups: "suffering" (low life satisfaction), "struggling" (moderate life satisfaction), and "thriving" (high life satisfaction). The proportion of people in each group is used to establish a country's ranking.

Another problem is that all these alternative indices cannot be calculated as quickly and regularly as GDP. There is no single and generally accepted method for studying happiness, and the question of what happiness is and how to measure it remains open. In the materials of the section "Economics of Happiness" of the journal "Science and Life", when analysing reports from such publications as "New Scientist" (England), "Geo", "Natur + Cosmos" and "VGI-Nahrichten" (Germany), "American Scientist", "Discover" and "Popular Science" (USA), "Recherche", "Science et Vie" and "Sciences et Avenir" (France), it was concluded that in the world today there are only 15 scientific definitions of the concept of "happiness". Survey methods are recognized as the most effective methods for assessing happiness. The earliest happiness questionnaire was conducted by the American psychologist J.B. Watson at the beginning of the 20th century. Later, this topic was studied by such scientists as E. L. Thorndike (1940s), Andrews and White (1976), and Campbell et al. (1976). Fordis developed a measure of happiness in 1988, Brandstetter developed the Affective Balance Questionnaire in 1991, and Diener developed the Satisfaction with Life Scale (SWLS) in 1996, which, according to Argyle, are the most indicative and valid methods for assessing happiness, along with the Oxford Happiness Inventory (OR; OR; Argyle et al., 1989). Thus, having analysed the history of the formation of a view on subjective satisfaction with the quality of life and having studied the level of happiness as one of the criteria for assessing the effectiveness of government decisions, the following conclusions can be drawn.

The GDP indicator is still the primary and most developed criterion of the country's economic development. However, its growth, according to well-known politicians, economists, sociologists and psychologists, can also be carried out at the expense of "growing unhappiness". The question of the linear dependence of happiness on well-being remains open. It has been proven that the subjective perception of life satisfaction has a more objective effect on the social situation than on the real state of affairs. In this regard, the scientific community is increasingly raising the question of the inadequacy of using purely economic indicators as an assessment of the efficiency and effectiveness of various socio-economic measures. The importance of the level of happiness of the population as an alternative to the gross domestic product indicator and a criterion for assessing the effectiveness of state policy is recognized by politicians, sociologists, economists and psychologists around the world. There is gradually an understanding that "happiness surveys can serve as an important auxiliary tool for the formation of state policy", which conducts interstate comparisons, as well as numerous questionnaire surveys of the population, which allow for a deeper study of this problem. The areas studied within the framework of various index methodologies usually include economic development, environmental protection, promotion of national culture, efficiency of public administration, sustainable development, security, political rights of people, state of ecology, accessibility of services of social institutions, expenditures on research, education, culture and sports, ability of people to participate in public life, degree of territorial and social life – population mobility, etc. Analysis of the experience of the practical application of index methods for assessing the level of happiness allows us to conclude about the significant shortcomings of this approach. First of all, these include the following:

- The integral indicator does not reflect the problems in the direction of private indices. Therefore, none of the indexes can be used once to assess the country's position in the world.
- At the same time, the country's characteristics, cultural and ethnic differences, historical experience, and current situation are not taken into account.
- Happiness indices cannot be calculated as quickly and regularly as GDP.
- There are no sufficiently reliable reasons to believe that this or that issue of choosing a method for assessing the level of happiness remains relevant. At the moment, there are many different indices and assessments of happiness. With the help of another component of the index, the data assigned to it by researchers have exactly the weight.
- Index methods for assessing the level of happiness are most often based on average state indicators that do not reflect asymmetry in the distribution of goods and do not take into account some factors and issues of human spiritual and moral development.

Alternative indices and indicators of the level of happiness perform an auxiliary function in relation to GDP but are not yet able to completely replace it. However, the very fact of the formation in recent years of a serious intellectual movement associated with attempts to study happiness using scientific methods indicates the special importance of this problem and the growing interest in it on the part of society, science and government.

2.3. Processing and analysis of primary sociological information

Processing and analysis of the collected primary sociological information consists of quantitatively assessing the influence of various factors on the development of social processes in different spheres of society. Primary sociological details can be processed using multiple methods of statistics and data mining: clustering, classification, regression,

correlation, and factor analysis [14-15]. The processed information can be presented in tables, graphs, diagrams, pictures, and schemes that allow interpreting the collected data, analysing and identifying specific dependencies, drawing conclusions, and developing recommendations. However, processing sociological information is possible only if the characteristics of the phenomenon under study are quantitatively measured. Therefore, processing primary statistical information involves converting categorical data into numerical data. Sociological measurement is a procedure by which the qualitative characteristics of a social phenomenon or object under study are compared with a certain standard, and a numerical value is obtained. The standard of measurement is a scale that the sociologist himself creates in the research process. Scaling gives quantitative certainty to the qualitative characteristics being studied. With the help of scaling, qualitatively heterogeneous social characteristics lead to comparable quantitative indicators. In the scaling process, the external signs of the phenomenon (object) under study are first identified, that is, those properties and characteristics that are subject to observation and measurement. A specific set of variables characterizes each sign. The indicators of variation of each of these variables are indicators - characteristics available for observation and measurement.

In order to select an indicator, it is necessary to first interpret and operationalize the phenomenon under study. Suppose interpretation allows us to determine the direction of analysis in which quantitative information should be collected, and operationalization. In that case, the types of this information, then defining indicators, help us to choose the methods and forms of its collection. In addition, indicators make it possible to correctly formulate the research toolkit and determine the structures of answers to the questions posed. Each indicator has specific characteristics, which are included in the toolkit as options for answering the question. Arranged in a particular sequence, they form a measurement scale. A combination of several indicators forms an index. The index can be expressed using a simple or weighted arithmetic mean of each answer option in the scale or using the difference between high and low, positive and negative manifestations of the feature. Scaling together with indexing form a procedure called quantification in sociology. Quantification is a procedure for measuring and quantitatively expressing qualitative features and relations of social objects – for example, a measurement in points of positive emotions, satisfaction, etc.

To build a measurement scale, first find the range of judgment, that is, determine the extreme values of the manifestation of a particular feature (maximum and minimum), which is called establishing its continuum (duration). Then, the continuum is divided into parts. It is the procedure for establishing granularity or gradation of the scale. Verbal formulation of guessed (possible) answers to questions are only points on a continuous scale of judgments. If a feature of a phenomenon is measured, reflected by the operational concept of "satisfaction", then the positions of the measurement scale can be the characteristics of such a subjective indicator as the "degree of satisfaction": "partially satisfied", "more satisfied than dissatisfied", "completely dissatisfied", "rather dissatisfied than satisfied", "dissatisfied".

The number of gradations determines the so-called sensitivity of the scale - its ability to reveal the respondent's attitude to various aspects of the studied social phenomenon with an appropriate degree of differentiation. So, thanks to scaling, it becomes possible not only to record the presence or absence of a qualitative feature but also to measure it, that is, to assess the degree of its manifestation. Therefore, scaling performs three functions: classification, ranking, and the introduction of metrics, which measure the intensity of the manifestation of the social features being studied and determine the difference in such intensity. Let us dwell in more detail on the characteristics of measurement scales.

- A nominal scale contains a list of characteristics of an object or phenomenon, which are sometimes mutually exclusive. These scales measure such objective characteristics as gender, nationality, marital status, age, work experience, and qualifications, as well as the subjective attitude of respondents to certain aspects of a social phenomenon or process.
- A rank scale, or order scale, is formed by cumulation (addition, accumulation) based on the ordering of a scale of names. Rank scales allow you to order the properties being studied from the most to the least significant or vice versa. A rank scale in general can be presented as follows:

a) the most negative (negative) answer;
b) the negative answer;
c) rather negative than affirmative (positive) answer;
e) neutral answer;
f) rather affirmative than negative answer;
g) affirmative answer;
g) the most affirmative answer.

Most closed questions from various questionnaires are, in fact, rank scales.

- An interval (metric) scale is formed on the basis of a rank scale by assigning a certain number of points to each position. An example of such a scale:

a) does not satisfy at all (– 1);
b) instead does not satisfy than satisfies (– 0.5);
c) it is difficult to answer (0);

d) instead satisfies than does not satisfy (+ 0.5);
f) completely satisfies (+1).

Unlike a rank scale, an interval scale allows not only the order of the manifestations of the property being studied but also the calculation of the difference between individual positions of the scale, that is, the determination of intervals.

Thus, the processing and analysis of primary sociological information are concentrated on assigning numerical values to the qualitative characteristics of the phenomena, processes and objects under study in accordance with specific rules. This operation is carried out on the basis of an established (as a result of interpretation and operationalization) system of indicators in the form of empirical indicators and mathematical indices. It was called sociological measurement.

The qualitative and quantitative characteristics used in this case reproduce the structure and dynamics of the studied social phenomena, processes and objects, forming a complex system of social indicators. They consist of indicators that record the qualitative aspects (presence or absence of a feature) and indices or coefficients that record the quantitative aspects (intensity of manifestation of a feature).

The effectiveness of the study does not depend on the volume of information collected but, on the depth, and comprehensiveness of the analysis. The analysis begins with checking the instrument for accuracy, completeness, and filling quality. Checking for accuracy of filling involves identifying errors in the answers and correcting them. After the instrument has been inspected and the part of it that does not meet the specified criteria has been removed, the information is coded, that is, its formalization. The coding procedure consists of assigning specific conditional numbers - codes to each answer option. As a result of coding, all information is converted into a numerical system. Coding begins at the beginning of the study when specific codes are assigned to the answer options that are embedded in the instrument itself in closed and semi-closed questions. Coding of answers to open questions occurs after the survey. To do this, all the answers are written out, the frequency of repetition of answers of a specific content is determined, and their classification is carried out. That is, the answers are grouped into particular semantic group variants and a formalized list is developed - a codifier, with the help of which different variants of answers are encoded. In this case, two coding systems are used:

- ordinal with end-to-end numbering of all positions;
- positional with the numbering of variants only within one question.

Generalization of information begins with grouping respondents by one selected indicator. The groups thus obtained, homogeneous in composition, become the object of analysis. The use of combinational grouping (distribution of respondents by two or more indicators) allows for a deeper analysis. Depending on the research objectives, such grouping can be structural, typological or analytical. Structural grouping is a classification by a certain indicator that is objectively inherent in the entire set of data; typological - by an indicator created by the researcher himself or subjective, analytical - occurs when grouping is carried out by two or more indicators in order to determine their interaction. In this way, attributive and variational distribution series are obtained, to which statistical analysis methods are applied:

- analysis of average values, dispersion, correlation coefficient and conjugation;
- analysis of percentage distribution, graphical analysis of distribution;
- analysis of the correlation dependence of quantitative and qualitative characteristics;
- meaningful interpretation of data, and determination of causal relationships.

In the process of processing and analysing primary sociological information of the integral scale, indices are indicators of the intensity of manifestation of a sign or the level of development of a process, which is determined by a particular scale. Studying statistical relationships, a sociologist comes across various correlations between a factor and a resulting sign. Intensive development of sociological science requires expansion of methodological and information support of sociological research.

2.4. Existing approaches review and analysis to determining a favourable environment for people's life and development

In [16], it was mentioned that happiness can be a good indicator of how well a society is functioning. It becomes essential because Bentham [17] said that the best society is one where citizens are the happiest. Several studies have been conducted on the positive aspects and issues of happiness in policymaking [18-19]. As mentioned earlier, happiness can be defined based on various factors. Unfortunately, these factors have been analysed manually [20-21]. The complexity of the factor leads to expensive analysis costs. Therefore, automated analysis is needed. The complexity of the factor leads to expensive analysis costs. Thus, automated analysis is required, which can provide a more objective and rapid identification of happiness factors. By applying machine learning and data analysis methods, it is possible to significantly simplify and automate the process of assessing and predicting happiness based on available data.

In addition, automatic analysis can help identify new relationships and patterns between different factors that affect people's happiness. It can contribute to the development of new strategies and approaches in policymaking to improve the quality of life of citizens. As mentioned earlier, there are a large number of features that are used to predict national

happiness. However, some features may not have a significant contribution to the forecast. Therefore, it is unreasonable to use all features to analyse happiness. In this work, factor analysis is used to explore related features. Factor analysis aims to determine the contribution of a particular feature. This technique does not focus on dimensionality reduction. Therefore, no removed features will be present. The works in [22-23] presented the advantages of factor analysis. The first mentioned advantage is the ability to detect latent dimensions or constructs that cannot be done using direct analysis. In addition, such an approach is easy to use and inexpensive in terms of resources.

Due to its ability to learn from experience, machine learning is used in various fields. Support vector machine is one of the powerful machine learning algorithms. Support vector machine (SVM) is a learning method used to correctly classify unseen data. It is a learning method commonly used to correctly classify unseen data. This technique has been used in various research fields due to its excellent performance. To perform the classification task, a support vector machine constructs a hyperplane that divides the data into different categories [21]. One of the crucial advantages of a support vector machine is its ability to handle data scarcity. In addition, a support vector machine is able to effectively learn the boundaries of complex solutions in a high-dimensional feature space. Due to the complex features used to predict national happiness, it is essential to apply a technique with the ability to deal with complex features.

Thus, the use of support vector machines (SVM) in the study of the enabling environment for people to live and develop in a city can help identify relevant factors and the relationships between them. It will provide a better understanding of which aspects of society and infrastructure affect citizens' happiness and can be used to improve living standards [24-26]. The SVM method can be beneficial when analysing data with a limited number of samples and high dimensionality of features, as is often the case in happiness studies. Due to its ability to make flexible decisions in a high-dimensional feature space, SVM can help identify complex patterns and relationships between factors that affect people's happiness in cities.

Clustering methods are also essential machine learning tools for detecting structure in data sets and can be applied to determine the enabling environment for people to live and develop. These methods help group observations based on their similarities, which can reveal patterns and relationships between different aspects of life and happiness.

- K-means is one of the most well-known clustering methods and is widely used in various scientific studies [26]. In [27], the authors applied the k-means method to cluster countries based on the happiness index and various socio-economic indicators, which allowed them to identify common characteristics and differences between them.
- Hierarchical clustering methods build a tree-like structure of clusters and can help identify deep relationships between observations. In [28], the authors used hierarchical clustering to analyse the quality of life indicators at different levels of society, which helped to understand various aspects that affect people's happiness.
- The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method is a robust clustering algorithm that can detect clusters of various shapes and sizes, as well as detect noise in the data. In [29], the authors applied DBSCAN to analyse socio-economic data that affect people's quality of life and happiness in different regions. This study helped to identify groups of regions with similar characteristics and differences that can be used to develop effective development strategies and welfare policies.
- Agglomerative clustering combines the closest clusters in a hierarchical order. In the article [30], the authors used agglomerative clustering to analyse data on housing conditions and environmental factors that affect the quality of life in different cities. The results helped to develop strategies to improve living conditions for residents.
- Spectral clustering uses graphs and spectral properties to divide data into clusters. In the study [31], the authors used spectral clustering to identify similar places of residence based on happiness indicators and socio-economic characteristics, which helped to identify the needs of the local population and contributed to the development of a favourable living environment.
- The application of various clustering methods in scientific research confirms the importance of these methods for analysing a favourable environment for human life and development. They help to identify relationships and patterns between different aspects of happiness and can be used to improve human living conditions and development in cities.

In the article [32], researchers conduct a spatial econometric analysis of the happiness of Canadian cities using data from quality-of-life surveys. The authors apply spatial autoregressive models (SAR) to determine the impact of various factors on the happiness of city dwellers. The results show that income, unemployment, nationality, and other demographic characteristics are important factors in the happiness level of city dwellers. The article [33] examines the key aspects of success in online shopping, which may be related to the quality of life in cities as people increasingly turn to online shopping to meet their needs. The researchers use a hybrid approach that combines machine learning and statistical methods, including regression analysis, to identify key factors that affect the success of online shopping. They find that convenience, security, and service quality are essential factors for a successful online shopper.

Both papers use regression models to investigate different aspects of a conducive environment for people to live and develop. On the one hand, Páez and Scott [32] focus on a spatial analysis of happiness in cities, examining the impact of economic and demographic factors on life satisfaction. The use of spatial autoregressive models allows us to take into account the interrelationships between towns. It contributes to a better understanding of the factors that

influence happiness at the city level. On the other hand, Huang et al. [33] study issues related to the quality of life in cities from the perspective of online shopping. They use a hybrid approach that combines regression analysis and machine learning to identify key factors of success in online shopping that can reflect convenience, security, and service quality, which contribute to improving the quality of life in cities. These articles show how regression models can be used to investigate various aspects of the enabling environment for people to live and develop, including happiness and convenience, which are essential components of quality of life. The use of regression models and machine learning can help researchers better understand and predict the factors that influence the enabling environment for people to live and develop and provide recommendations for policymakers and urban project developers.

In summary, research in the field of determining the enabling environment for people to live and develop has actively used machine learning methods, in particular clustering algorithms such as k-means and multiple regression models. The application of K-means allows you to detect hidden structures in the data and group objects with similar characteristics, which can contribute to improving urban planning. Multiple regression models help to study the relationships between different factors and predict their impact on quality of life and happiness. The application of these methods allows us to obtain valuable conclusions about the factors that contribute to the creation of a favourable environment for human life and development and to develop effective strategies in urban planning and social policy.

2.5. Setting the task for developing a module for determining a favourable environment for human life and development

The development of a module for determining a favourable environment for the life and development of people based on the happiness index is very relevant due to the complexity of the problems of the urban environment. The happiness index is considered an important indicator that helps determine the level of comfort and well-being of the population in different cities and countries. The module can use machine learning methods, such as linear regression, clustering, and others, which allow for predictions and categories based on existing data. It helps to identify patterns and trends that affect the happiness index and can be useful for urban development planning. The module can involve the public and stakeholders in the decision-making process and raise public awareness of the importance of a favourable environment for life and development in cities. For Ukraine, the development of such a module is significant, as it will contribute to improving the quality of life and ensuring sustainable development of urban environments. The development of a module for determining a favourable climate for the life and development of people is a very relevant task in the modern world. The application of machine learning and data analysis methods in this area can open up new opportunities for information support for decision-making at all levels of city management and government structures. The Happiness Index helps to determine the level of comfort and well-being of the population in different cities and countries, which can contribute to improving living conditions and creating a favourable environment for the development of people in cities. The development of the module can also contribute to the involvement of the public and stakeholders in the decision-making process, as well as raising public awareness of the importance of a favourable environment for life and development in cities. Therefore, this project can be of great importance for the development of towns and ensuring sustainable development in general. The research aims to develop a module that uses machine learning methods to analyse various factors that affect the quality of life and level of happiness of people and to determine favourable conditions for creating an environment that promotes human development in cities. Therefore, to achieve the goal, the following tasks need to be solved:

- review the principles and stages of determining a favourable environment for human life and development;
- analyse existing approaches to determining a favourable environment for human life and development;
- formulate tasks for developing a module for determining a favourable environment for human life and development based on the happiness index;
- develop algorithms based on machine learning to determine a favourable environment for human life and development based on the happiness index;
- implement a software module for determining a favourable environment for human life and development based on the happiness index;
- verify the effectiveness and accuracy of the developed module on real data.

3. Material and Methods

3.1. Methods of intellectual analysis of sociological data

In the modern technological world, the development of sociological science is closely related to the use of new information technologies [34]. Successes in the field of artificial intelligence have determined the vector of further technological development and initiated the process of intellectualization of technical means, which, in particular, led to the emergence of a new paradigm of cognition - data mining. The main idea of Data Mining is that data stores information that is invisible from a familiar perspective.

Data Mining is not one but a whole set of different methods for identifying patterns, hypotheses, patterns and knowledge: searching for associative rules, classification, forecasting, clustering, etc. [35]. The choice of a specific method depends on the type of data and what information needs to be obtained. Before starting work with algorithms and methods of Data Mining, it is necessary to understand what data is required, whether there is a need for their

pre-processing and what kind. Now, you can find ready-made sets of sociological data on the Internet. They are usually unprepared and have outliers, omissions, and problematic data. The found data set should be used in the future, together with new data, with a sufficiently high degree of reliability. For further sociological research of the data set obtained as a result of a survey on the level of happiness in different parts of the world, it is advisable to identify the internal structure of the data by applying clustering methods. Having identified clusters of countries related to happiness and analysing the distribution of countries in the world by level of happiness, it is advisable to build a classifier that allows the classification of new countries. The application of classification and clustering methods requires determining the measures of proximity between objects, which can be determined for both numerical and categorical scales [36]. Below is a clear justification of why it is advisable to use the methods of clustering K-means and multiple linear regression in the context of analysing sociological data on the level of happiness and life satisfaction, including within the framework of this study.

1. Let's describe the rationale for choosing K-means. In sociological research, it is often necessary to identify hidden groups or typologies of countries/respondents by similarity in different social, economic and psychological characteristics. Manual classification with a large number of features is subjective and ineffective. K-means is simple and effective for large amounts of numerical data, which is typical for arrays of sociological information (happiness index, GDP per capita, social support, etc.). The algorithm allows you to automatically divide countries/cities into clusters that have similar indicators of happiness and factors that affect it. It also provides interpretation of the results: the average values of the clusters make it possible to understand the standard features of the groups. At the same time, it is computationally inexpensive, which allows for fast processing of extensive data. Alternatives can be DBSCAN or hierarchical clustering. They are better at complex cluster shapes, but they are less interpreted, less scalable, and more challenging to set up. K-means is the optimal compromise between simplicity, speed and information content in conditions of limited computing resources and a large amount of data.

2. Let's describe the rationale for choosing multiple linear regression. The basis is that you need to understand which factors (independent variables) have the most significant impact on the level of happiness (dependent variable), as well as predict this level for new countries/regions. Multiple regression makes it possible to establish linear relationships between happiness and influential factors (income, health, freedom of choice, level of corruption, etc.). The algorithm provides coefficients of influence of each factor, which allows you to draw reasonable conclusions and formulate policy recommendations. It is also suitable for predicting happiness index values under given conditions. In parallel, it has simple interpretations of the results, which are vital for sociological analysis. Alternatives can be nonlinear models (e.g., decision trees, neural networks), which can give better accuracy but are poorly interpreted, which is critical in the social sciences. Regression is a classic tool of sociology that allows you to draw informed, statistically sound conclusions.

The methods of K-means and multiple regression were chosen due to the balance between accuracy, interpretation, ease of implementation and practical application in the sociological analysis of large data sets. They allow solving two key tasks: identifying typical groups (clustering) and determining influencing/forecasting factors (regression). For data presented by numerical metric features, the following measures are most often used as a measure of dissimilarity:

1) The formula for the Euclidean distance between two points $A = (x_1, y_1, \dots, z_1)$ and $= (x_2, y_2, \dots, z_2)$ in n-dimensional space is:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + \cdots + (z_2 - z_1)^2}. \quad (2)$$

Or, more generally, for n-dimensional space:

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}. \quad (3)$$

where $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$ are the data set objects and a_i and b_i are the i-th attribute of the data set objects, and n is the number of variables. It is a classical metric that defines the "straight line" distance between two points (data set objects) in space.

2) The formula for the square of the Euclidean distance between two points $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$ in n-dimensional space is:

$$d^2(A,B) = \sum_{i=1}^{n}(a_i - b_i)^2. \quad (4)$$

This expression is often used in cluster analysis, machine learning, and statistics when the root in the formula is not needed, for example, to compare distances. An explanation of why it is sometimes more convenient to work with the square of the distance:

1. Simpler calculations, i.e. calculating the square of the distance, does not require taking the square root, which:

- reduces computational complexity;
- speeds up the execution of algorithms, especially on large data sets.

In clustering or nearest neighbour search, we often compare which of the distances is smaller, and the root here does not change the result.

2. Mathematical convenience, in particular, in many statistical methods (e.g., the least squares method) or machine learning algorithms (e.g., in k-means) the square of the distance:

- allows you to avoid derivatives of the root during optimization;
- behaves better when differentiating – the derivative of the square is simpler.

3. Penalty for large deviations, i.e., the square of the distance "penalizes" larger distances more strongly:

- if the distance is doubled, the square will increase fourfold;
- it can be helpful if you need to take larger deviations into account.

4. Linear algebra and spaces, in particular, in many problems (for example, when working with vectors in spaces), the square of the Euclidean distance can often be expressed in terms of the scalar product:

$$d^2(A,B) = \langle A-B, A-B \rangle. \quad (5)$$

It is convenient for algebraic transformations, such as in PCA (Principal Component Analysis). In short, the square of the Euclidean distance is faster, easier to differentiate, and often more mathematically compact. All of these factors make it popular in real-world problems. For data represented by categorical attributes, as a measure of dissimilarity of objects $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$ we use:

- Hamming distance, which is equal to the number of attributes whose values for objects differ and is a metric;
- The percentage of disagreement, which is calculated as the ratio of the number of non-coincident features to the total number of categorical features:

$$d(A,B) = \frac{l}{n} \quad (6)$$

where n is the total number of categorical features; l is the number of non-matching features for which $a_i \neq b_i$, $i \in \{1,2,\dots,n\}$.

Before applying clustering and classification methods, the features of the sociological survey dataset must be normalized – brought to a single numerical range.

Among the clustering methods for identifying groups of related countries based on sociological survey data, the k-means algorithm was chosen [37]. The action of this algorithm is aimed at minimizing the objective function:

$$J = \sum_{j=1}^{k} \sum_{i:c_i=j} |\boldsymbol{x_i} - \mu_j|^2 \quad (7)$$

The centre of gravity of the cluster is a point in the feature space with coordinates that are the arithmetic mean values of the corresponding features of the countries included in the cluster.

The k-means algorithm divides the set of N points $\{x_1, x_2, \dots, x_N\} \subset R^n$ into k clusters so as to minimize the sum of the squares of the distances of the points to the centres of their clusters.

1. Initialization. Randomly select k cluster centres (centroids):

$$\mu_1, \mu_2, \dots, \mu_k \in R^n \quad (8)$$

2. Assign points to clusters. Each point $\boldsymbol{x_i}$ is assigned to the nearest centre:

$$c_i = \arg\min_{j \in \{1,\dots,k\}} |x_i - \mu_j|^2 \quad (9)$$

where c_i is the cluster index to which the point x_i, $|x_i - \mu_j|^2$ is the square of the Euclidean distance.

3. Update cluster centres. For each cluster j, owe update the centre as the average of all points belonging to it:

$$\mu_j = \frac{1}{N_j} \sum_{i:c_i=j} x_i \quad (10)$$

where N_j is the number of points in cluster j, $\sum_{i:c_i=j} x_i$ is the sum of all points assigned to cluster$j$.

4. Iteration. We repeat steps 2–3 until the centres stop changing (or change very little) – i.e., convergence.

5. The objective function (which is being minimized). The algorithm minimizes the sum of the squares of the distances of all points to the centres of their clusters:

$$J = \sum_{j=1}^{k} \sum_{i:c_i=j} |x_i - \mu_j|^2$$

It is the loss function or objective function that k-means optimizes. Key points explained:

- We use the square of the distance to avoid the root and speed up calculations.
- The centre of a cluster is the centre of mass (average) of its points.
- The algorithm does not guarantee a global minimum but usually gives good results.

The k-means algorithm is iterative with the following stages:

1) the value of *k* is set - the number of clusters into which the data set must be divided;
2) *k* initial centres of gravity of the clusters are randomly selected;
3) the distances from all objects to the centres of gravity of the clusters are calculated;
4) each object is assigned to the cluster with the nearest centre of gravity;
5) the centres of gravity of the clusters are recalculated according to their current composition;
6) return to point 3 and repeat points 4 and 5 until there is no movement of objects from cluster to cluster – at this point, the algorithm stops working.

The primary classification methods that can be used to analyse sociological data include the following: support vector method, k-nearest neighbours' method, Naïve Bayes, decision trees, etc. The k-nearest neighbours' algorithm was used to build a classifier of countries by level of happiness.

The k-nearest Neighbours (kNN) method is a popular classification algorithm and one of the most understandable approaches to classification. The algorithm is able to select k nearest neighbours among all objects that are similar to a new object. Based on the classes of nearest neighbours, a decision is made regarding the class to which the new object will belong. To classify a new object using the k-nearest neighbours' algorithm, it is necessary [38]:

- Set the number k - the number of nearest neighbours;
- Choose a method for calculating proximity (distances between objects) and calculate the distance to each of the objects in the training set;
- Select k objects with the minimum distance to them;

The class of the new object is the class that occurs most often among the nearest neighbours (in simple unweighted voting) or the class that received the most significant number of votes (in weighted voting).

The weak point of this algorithm is the selection of the coefficient k - the number of objects that will be considered similar (neighbours). If we take k = 1, the algorithm will lose its generalization ability (i.e., the ability to give the correct result for data that has not been encountered before), and a new record will be assigned to the class closest to it. If we set the value too high, many local features will not be detected. The advantages of the kNN algorithm are as follows:

- The algorithm is resistant to anomalous outliers because the probability of such a record being included in the number of k-nearest neighbours is small. If this happens, the impact on voting (especially weighted) (at k>2) will also be insignificant;
- The software implementation of the algorithm is relatively simple;
- The result of the algorithm is easy to interpret. Experts in various fields are pretty clear about the logic of the algorithm based on finding similar objects;
- The ability to modify the algorithm by using the most suitable combination of functions and metrics allows you to tailor the algorithm to a specific task.

Disadvantages of the kNN algorithm:

- The data set used for the algorithm must be representative;
- The model cannot be "separated" from the data, to classify a new object. All examples must be used - this feature greatly limits the use of the algorithm.

The following is a justification for why clustering (K-means is the best approach, as well as an appropriate and effective method for identifying patterns related to the level of happiness and life satisfaction, taking into account the nature of the data and the goals of the study.

1. Nature of data: multidimensional, quantitative, cross-country indicators. The dataset contains numerous quantitative indicators (GDP per capita, social support, life expectancy, freedom of choice, generosity, level of corruption) that together form a multidimensional socioeconomic space in which each country is a point. In this case, clustering makes it possible to identify groups of countries with similar profiles of well-being and happiness. Also, the

method allows you to find hidden structures in the data that a simple comparison of individual variables cannot see.

2. Absence of a priori classes. Unlike classification problems, in which it is necessary to assign objects to pre-known classes (for example, "happy"/"unhappy"), in this study, there are no fixed categories of happiness, but only numerical indicators. Clustering allows you to automatically form groups without prior labelling and identify country typologies according to aggregate indicators of well-being.

3. Easy and understandable interpretation of clustering results, since each cluster has average values of indicators, which allows describing typical features of "groups of happy countries", "moderately happy", etc. It is also possible to visualise data in 2D/3D space, which simplifies the perception of the results by politicians, government officials, and the public.

4. Clustering as a basis for further analytics. The study provides not only for clustering, but also for the construction of a classifier based on the results of clustering (the method of k-closest neighbours). It makes it possible to classify new countries according to already identified groups and assess changes in the position of countries in the welfare area when updating the data.

5. Clustering helps you make informed decisions. Comparison of clusters with factors (economy, social support, corruption) allows you to determine which combinations of factors most often accompany a high level of happiness. At the same time, this will enable you to create recommendations for regional policy – for example, how a country can "move" from a cluster of less happy to happier by changing specific indicators.

Clustering is the best approach in this study, as it allows you to identify natural groups of countries without predefined labels and works adaptively with multidimensional data. It generates interpreted results suitable for analysing, predicting and supporting social decision-making. The method also serves as a basis for building further classification and regression models.

3.2. Problem statement

Having analysed the subject area of research in the field of sociology, the main approaches to processing sociological information, and the methods of its analysis, it was concluded that it is necessary to develop an intelligent system for analysing sociological information using clustering and data classification algorithms. The object of the work is the process of processing and analysing sociological information. The subject of the work is software tools for processing sociological survey data and methods of classifying and clustering data. The purpose of the work is to increase the efficiency of conducting sociological research by developing an intelligent system for analysing sociological data using classification and clustering algorithms. Achieving the set goal requires solving the following tasks:

- to investigate the theoretical foundations of conducting sociological research and methods for processing and analyzing sociological information;
- to justify the choice of tools for developing a system for analysing sociological data;
- to develop and implement software for an intelligent system for analysing the level of happiness based on processing data from sociological surveys in different countries of the world.

The tasks set define the main functionality of the developed system, which should provide the opportunity to:

- Download survey data from a CSV format file;
- View data by the user;
- Perform normalization of downloaded data;
- Perform clustering of normalized data using the k-means algorithm;
- View by the user the characteristics of countries that fall into each cluster and the average values of the features of each cluster;
- Based on the clustering performed, taking into account the distribution of countries into groups of related objects, a classifier using the k-nearest neighbours' algorithm is built, and new objects are classified.

3.3. Algorithm for categorization of countries by happiness index

Data visualization does not have unnecessary decorative elements. It is aimed at presenting large amounts of information, taking into account possible relationships. In Data Science, data visualization is used not only for the visual presentation of results in the form of understandable graphs but also as a method of rapid prototyping. Within the framework of data mining, analysts and Data Scientists use a variety of visual representations to reveal hidden relationships and dependencies, maximize immersion in the data, select the most critical variables, identify deviations and anomalies, test the main hypotheses and develop initial models.

EDA [39-40] is a component of data preparation for ML modelling, which includes generating features after sampling and cleaning the dataset. At the same time, EDA allows the data scientist to make sure that the interpretation of the results is correct and that they are applicable to the business context. For business users, EDA can serve as a means of checking the correctness of assumptions and questions. EDA is also an essential tool for validating data and assessing its compliance with business goals. When working with large amounts of data, EDA is beneficial because a dataset can contain information from different sources with different levels of accuracy and detail. In practice, EDA can

lead to interesting business insights, such as discovering the relationship between the amount of the check and the time of day or the correlation between the number of visitors and weather conditions. EDA in Data Science is based on statistics and probability theory, which includes probability distributions of variables, correlation matrices, factor and discriminant analysis, and multivariate scaling. EDA is performed using specialized mathematical programs such as SAS, MATLAB, KNIME, Weka, and Orange, systems such as RStudio, original Python scripts, or built-in formulas in spreadsheet editors such as Excel and Google Sheets. For data visualization, you should choose the appropriate type of graph, taking into account the purpose of the analysis and/or presentation of information, such as comparing different indicators, demonstrating the distribution of data, identifying the composition and structure of data, or relationships between variables. For data visualization, sometimes more than 20 different types of charts are used, which differ in their purpose and capabilities. The choice of a specific chart for data visualization depends on the number of variables and periods being analysed. The following types of charts are often used in Data Science:

- a histogram, which displays the distribution of data within a continuous interval or period;
- a scatter plot, which allows you to detect a correlation between two variables;
- a boxplot, which displays groups of numerical data through quartiles;
- a heat matrix, which helps to detect correlations between several variables;
- a bubble plot, which displays the relationships between variables through their location and proportions.

A correctly selected type of chart for data visualization meets the following criteria [41]:

- Conciseness - the ability to simultaneously display many different types of data;
- Relativity and proximity – the ability to demonstrate clusters, relative sizes of groups, their similarities and differences, and outliers;
- Concentration and context – the ability to easily and quickly interact with the selected object by interactively viewing it (displaying structure and relationships);
- Scalability – the ability to easily and quickly change the size of the representation;
- User convenience due to maximum visibility and support for intuitive actions to identify patterns.

Therefore, visualization is needed not only for the visual presentation of results but also for the development of preliminary hypotheses, as well as validation of initial data. EDA, or exploratory data analysis, is an essential stage in preparing a dataset for ML modelling and other data mining techniques. The choice of a graph for visualization depends on the goal (comparison of variables, identification of relationships, representation of composition and structure or demonstration of statistical distribution) and the analysed categories (multivariate analysis, time series or correlation of several indicators). In the process of EDA (English Exploratory Data Analysis), it is essential to understand what tools and methods can be used for data analysis. One of the most common methods is visualization, which allows you to present data in a convenient and easy-to-understand form. One of the types of visualization is categorization, which will enable you to divide data into specific groups according to a given criterion. In this case, let's consider the algorithm for categorizing countries by the happiness index. This algorithm uses quartiles to determine the happiness categories of countries. It allows you to analyse the data in more detail and draw conclusions about the level of happiness in different countries. Algorithm for categorizing countries by happiness index:

- The first quartile (Q1) of the Happiness Index is determined;
- The median (Q2) of the Happiness Index is calculated;
- The third quartile (Q3) of the Happiness Index is determined;
- Using a for loop, the program iterates through all rows in the DataFrame data. For each country, a happiness category is determined based on the Happiness Index value:
  - If the Happiness Index is less than or equal to Q1, the country is assigned the 'Sad' category;
  - If the Happiness Index is greater than Q1 and less than Q2, the country is assigned the 'Moderately Sad' category;
  - If the Happiness Index is greater than Q2 and less than Q3, the country is assigned the 'Happy' category;
  - If the Happiness Index is greater than or equal to Q3, the country is assigned the 'Very Happy' category.

The algorithm for categorizing countries by the Happiness Index allows for a more detailed analysis of the data and concludes the level of happiness of countries depending on the value of their Happiness Index. As a result of the algorithm, each country is assigned a happiness category that can be used for further research and analysis. For example, comparing happiness categories with other indicators such as GDP, education, healthcare, and others can help identify factors that influence the level of happiness in countries.

3.4. Data Clustering Algorithm Using KMeans

Clustering is the process of dividing elements of a data set into groups based on their similarities. These groups are called clusters. One of the elements to be clustered is called an object. For each object, a feature vector is created that describes it:

$$X = \{x_1, x_2, \ldots, x_m\} \tag{11}$$

Objects in clustering algorithms are described by feature vectors, which consist of individual components x_i. These components can be quantitative (e.g., point coordinates) or qualitative (e.g., colours). The number of components determines the dimensionality of the feature space, and the set of all possible feature vectors is denoted as the feature space.

$$A = \{X_1, X_2, \ldots, X_n\} \tag{12}$$

For the correct operation of clustering algorithms, it is necessary to normalize the characteristics, that is, to bring them to the same scale. A cluster is a group of objects close to each other from the set A. The distance $\rho(x_i, y_i)$ between objects x_i and y_i is determined using the selected metric in the space of characteristics. To speed up the clustering process, you can reduce the dimensionality of the space of characteristic vectors by highlighting the most essential properties of the objects (for example, using the principal component method). The choice of metric depends on the space in which the objects are located and the implicit characteristics of the clusters. Euclidean distance

$$\rho(x_i, y_i) = \sqrt{\sum_{k=1}^{n}(x_{ik} - y_{ik})^2} \tag{13}$$

The squared Euclidean distance is a metric used to measure the distance between two points in n-dimensional space. It is one of the most well-known and frequently used metrics. The squared Euclidean distance between two points x_i and y_i is calculated as the sum of the squares of the differences between the corresponding coordinates of these points:

$$\rho(x_i, y_i) = \sum_{k=1}^{n}(x_{ik} - y_{ik})^2 \tag{14}$$

where x_{ik} and y_{ik} are the k-th coordinates of the vectors x_i and y_i.

When using the square of the Euclidean distance as a clustering metric, greater weight is given to nearby objects, that is, those that are located closer to a given point. It can be helpful in cases where nearby objects are more similar to each other than distant ones. However, if the distances between objects are large, the square of the Euclidean distance can lead to an underestimation of distant objects and an excess of weight for nearby objects. The Manhattan distance or city block distance is one of the metrics in n-dimensional space, which is defined as the sum of the absolute differences between the corresponding coordinates of two points. The name of the metric arose because, in city blocks, the distances between points are measured along straight lines that are parallel to and perpendicular to the streets. The distance between two points (x_1, y_1) and (x_2, y_2) in n-dimensional space will be determined by the formula:

$$\rho(x_i, y_i) = \sum_{k=1}^{n}|x_{ik} - y_{ik}|^2 \tag{15}$$

where $|x_1 - x_2|$ та $|y_1 - y_2|$ are the absolute differences between the corresponding coordinates of the points. This metric is widely used in computer vision, urban transport routing, urban planning projects and other areas.

The k-means method (k-means) [42-43] is an iterative data clustering algorithm that determines k clusters by grouping objects into k groups or clusters based on the similarity between them. The approach works as follows:

- The number of clusters k is specified by the user;
- k objects are randomly selected to become the cluster centres;
- Each object is assigned to the nearest cluster based on the distance between the object and the cluster centre;
- The arithmetic mean of all objects in each cluster is calculated, and a new cluster centre is selected;
- Each object is reassigned to the nearest cluster based on the new cluster centre;
- The process continues until the variability between the previous and current location of the cluster centres is less than a particular threshold value or until the maximum number of iterations is reached.

The result of the k-means method is k clusters that contain objects with similar characteristics. This algorithm is fast and efficient to use, but it also has some drawbacks, including dependence on the initial choice of cluster centres and vulnerability to outliers. The following is a pseudocode for clustering the happiness index using the k-means method. It explains in detail each step of the algorithm, from displaying the data and encoding it to deriving information about the data frame and visualizing the correlation and covariance matrices. It also discusses in detail the use of the elbow method to determine the optimal number of clusters and train a k-means model with a specified number of clusters. It will be helpful for those who want to cluster data and explore the dependencies between their attributes using correlation and covariance matrices. Pseudocode for clustering the happiness index:

- Display the first rows of the data frame "data" and its index;
- Encode the column "Regional_Indicator" using LabelEncoder;
- Output the values of different categories in "Regional_Indicator";
- Create a K-Means model to cluster the data;
- Visualize the results using the elbow method (KElbowVisualizer) to determine the optimal number of clusters;
- Train a K-Means model with a specified number of clusters;
- Get the cluster labels for each observation in the data frame "data";
- Add the cluster labels to the data frame "data" as a new column "Labels";
- Output information about the data frame "data";
- Calculate the correlation using the Pearson and Spearman methods;
- Visualize the Pearson correlation matrix using a heatmap;
- Visualize the Spearman correlation matrix using a heatmap;
- Calculate the covariance matrix for the data frame "data";
- Visualize the covariance matrix using a heatmap.

So, the pseudocode for clustering the happiness index data was described. It was shown how to code categorical variables and how to use the k-means model for clustering data. It also demonstrated the use of the elbow method to determine the optimal number of clusters and visualize the results using a heatmap. Next, the correlation analysis was performed using the Pearson and Spearman methods, and the visualization of the corresponding matrices was done using a heat map. Finally, the covariance matrix was calculated and visualized using a heatmap. All these steps will help researchers better understand the data and find interesting relationships between their parameters.

3.5. Algorithm for predicting the happiness index in Ukraine using linear multiple regression

Simple linear regression is a method that allows an expert or statistician to predict the value of one variable based on data on another variable. This method is applicable only to two continuous variables - an independent and a dependent. The independent variable is used to calculate the value of the dependent variable. The multiple regression model extends this method to the case when more than one explanatory variable is used. Multiple linear regression (MLR) [44-45] is a statistical method that uses several explanatory variables to predict the values of the dependent variable. The primary purpose of this method is to model the linear relationship between the descriptive and dependent variables. MLR is an extension of ordinary least squares (OLS) regression because it includes more than one explanatory variable. The MLR formula consists of the y-intercept, the slope coefficients for each explanatory variable, and a model error term.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon \tag{16}$$

where, for $i = n$ observations: p_i is the dependent variable; x_i is explanatory variables; β_0 is y-intercept (constant term); $\boldsymbol{\beta_p}$ is slope coefficients for each explanatory variable; ε is the model error term (also known as residuals).

The multiple regression model assumes that the dependent variables are linearly related to the independent variables. In addition, the independent variables are only slightly correlated with each other. To obtain accurate results, observations are randomly and independently selected from the entire population. The residuals are also expected to be normally distributed with mean zero and variance σ. The coefficient of determination (R^2) is used to measure how much variation in the independent variables explains the variation in the outcomes. However, increasing the number of predictors in an MLR model always increases the value of R^2, regardless of whether these predictors are related to the outcome variable. Therefore, R^2 alone cannot indicate which predictors should be included in the model and which should be excluded. The value of R^2 can range from 0 to 1, where zero means that any of the independent variables cannot predict the outcome, and one indicates that the outcome can be expected without error using the independent variables.

RMSE (Root Mean Squared Error) is a mathematical metric for measuring the root mean square error of a forecasting model, used to compare the quality of different forecasting models. It is calculated as the square root of the mean square of the difference between the predicted values and the actual values. The formula for RMSE is:

$$S_{RMSE} = sqrt\left(\left(\frac{1}{n}\right) \bullet sum\left(\left(y_{pred} - y_{true}\right)^2\right)\right) \tag{17}$$

where y_{pred} is a vector of predicted values; y_{true} is a vector of actual values; n is the number of observations.

RMSE allows us to determine how far the predicted values deviate from the actual values. The smaller the RMSE value, the better the model predicts. However, RMSE has a unit of measurement that depends on the unit of measurement of the input data, so comparing RMSE between different models is possible only within the same input data.

The following is a pseudocode describing the algorithm for predicting the happiness index in Ukraine using linear regression. Here is a verbal description of the algorithm, according to the pseudocode:

- Import the necessary libraries (pandas, LinearRegression, mean_squared_error);
- Assume that the data frame df has already been loaded and processed;
- Create a linear regression model;
- Extract the necessary variables and target data from df;
- Divide the data into variables and target values;
- Train a linear regression model using training data;
- Evaluate the model by comparing the predicted values with the actual values;
- Calculate and display the RMSE error and the coefficient of determination R2 to assess the quality of the model;
- Find the expected happiness index for Ukraine using data about Ukraine and the trained linear regression model;
- Display the predicted happiness index for Ukraine.

This algorithm helps to create a linear regression model to predict the happiness index based on selected variables and evaluate the quality of the model using metrics such as RMSE and R^2. So, the pseudocode of the algorithm for predicting the happiness index in Ukraine using linear regression is given. The algorithm includes the steps of creating a linear regression model, selecting the necessary variables and target data, training the model, assessing the quality of the model, and predicting the happiness index for Ukraine. The RMSE and R^2 metrics are used to determine the quality of the model. This algorithm is helpful in predicting the happiness index in Ukraine and can be used in other areas to predict various indicators using linear regression.

3.6. Description and analysis of the data set

The World Happiness Report [46] is a landmark analytical document that reflects the current state of global well-being. Its authority continues to grow internationally as governments, civil society organizations and institutions increasingly use happiness indicators as a basis for policymaking. Leading experts in fields such as economics, psychology, survey analysis, government statistics, health, and public administration demonstrate the possibility of effectively using well-being indicators to assess the progress of countries. The report examines the state of happiness in the world and reveals how the modern science of happiness explains both individual and national differences in life satisfaction. The rankings and indicators of happiness are based on data from the Gallup Global Survey. The table compares the overall happiness score with the contribution of six key factors – GDP per capita, social support, life expectancy, freedom of choice, corruption and generosity – to improving life in each country compared to a hypothetical environment called Dystopia (a country with the lowest possible scores on each of the six factors). These components do not affect the overall score, but they do provide insight into why some countries rank higher in the rankings than others. The database contains 149 records (rows) and 20 columns with different characteristics. Below is a brief description of the dataset:

- Country_Name: country name (object);
- Regional_Indicator: a regional indicator (object);
- Ladder_Score: ladder score (float64);
- Standard_Error_Ladder_Score: ladder score standard error (float64);
- Upper_Whisker: upper whisker (float64);
- Lower_Whisker: lower whisker (float64);
- Logged_GDPper_Capita: log GDP per capita (float64);
- Social_Support: social support (float64);
- Healthy_Life_Expectancy: healthy life expectancy (float64);
- Freedom_Make_Life_Choices: freedom to make life choices (float64);
- Generosity: generosity (float64);
- Perceptions_Corruption: perception of corruption (float64);
- Ladder_Score_Dystopia: Dystopia ladder score (float64);
- E_Log_GDPper_Capita: Expected log GDP per capita (float64);
- E_Social_Support: Expected social support (float64);
- E_Healthy_Life_Expectancy: Expected healthy life expectancy (float64);
- E_Freedom_Make_Life_Choices: Expected freedom to make life choices (float64);
- E_Generosity: Expected generosity (float64);
- E_Perceptions_Corruption: Expected perception of corruption (float64);
- Dystopia_Residual: Dystopian residue (float64).

Overall, we have no missing values (null), and the data types include 18 numeric columns (float64) and two text columns (object). For further analysis, you can explore statistical measures, relationships between features, and distributions. The mean and standard deviations (std) of the main numeric features are:

- Perceptions_Corruption: mean – 0.727450, standard deviation – 0.179226;
- Generosity: mean – (-0.015134), standard deviation – 0.150657;
- Freedom_Make_Life_Choices: mean – 0.791597, standard deviation – 0.113332;
- Healthy_Life_Expectancy: mean – 64.992799, standard deviation – 6.762043;
- Social_Support: mean – 0.814745, standard deviation – 0.114889;
- Logged_GDPper_Capita: mean – 9.432208, standard deviation – 1.158601;
- Standard_Error_Ladder_Score: mean – 0.058752, standard deviation – 0.022001;
- Ladder_Score: mean – 5.532839, standard deviation – 1.073924.

The minimum (min) and maximum (max) values are also given for each column. You can use this information for further analysis, such as the distribution of different characteristics and finding possible correlations between them. Quartiles also help to assess the distribution of the data. Next, the data was filtered for countries with low levels of social support, which have a value less than or equal to the mean value of social support (Fig. 2a). As a result, a list of 50 countries was obtained where social support is lower than the mean value.

These countries include different regions such as Africa, Asia, Latin America and Eastern Europe. It is important to note that countries with low levels of social support may have different political, economic and cultural contexts that affect the social support score. Next, we will analyse countries with high levels of social support, which have a value greater than or equal to the mean value of social support (Fig. 2b). As a result, we obtained a list of 50 countries where social support is higher than the mean value. These countries include different regions such as Europe, North and South America, Asia and Oceania. It is important to note that countries with high levels of social support may have different political, economic and cultural contexts that affect the social support score. Additional analysis can be conducted to find out which factors, such as GDP per capita, corruption, freedom to choose life decisions, etc., correlate with high levels of social support. It can help to better understand what factors influence social support in these countries and how they can support or improve it. Ukraine is also included in the list of countries with high levels of social support. It indicates that in Ukraine, a significant part of the population has access to social support, which can include various social programs provided by the state or non-governmental organizations, as well as family support, friends and acquaintances.

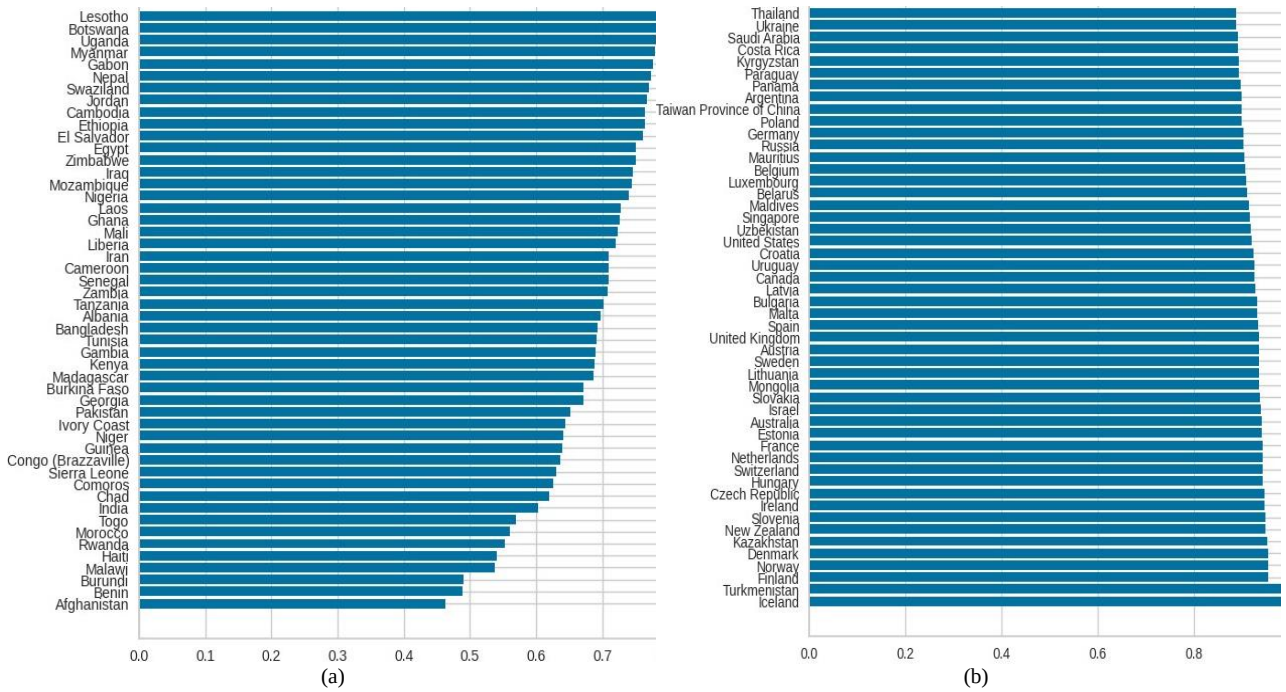


Fig. 2. (a) Countries with low social support and (b) countries with high social support

The study used data from the World Happiness Report (WHR). We will clearly describe the sources of the sociological survey data and the methods of collecting or assessing the representativeness of the sample in the context of scientific reliability. Below is a detailed description of how the

The data used in the World Happiness Report is based on the annual global Gallup World Poll, a large-scale research project covering more than 160 countries. Collection methodology:

– The number of respondents is about 1000 people per country every year.

– The sample is usually representative of the adult population (aged 15+), taking into account gender, age, and region of residence.

– Data collection methods are based on personal interviews (PAPIs or CAPIs) in countries with low Internet access and telephone interviews in developed countries or countries with a high level of urbanisation.

– The language of the interview is the local language of the country.

– The main question about happiness is based on the Cantril Ladder, where the respondent rates their life on a scale from 0 (the worst possible life) to 10 (the best possible life).

The representativeness of data for the world population is built on an extensive international coverage. The survey is conducted in 95-99% of the world's population. The methodology is adapted to regional conditions (different collection channels). The sample is representative at the country level, not just at the global level. Standardised questionnaires ensure comparability between countries. Although the data from the World Happiness Report is the largest international source on happiness, its representativeness at the global level is generally high, but at the national/cultural level it can be limited due to:

– uneven depth of coverage of the population,
– cultural and methodological distortions,
– the standard fixed interpretation of "happiness".

Table 2. Limitations of representativeness

Restriction	Explanation
Small sample per country	~1000 respondents are not always enough for countries with large and diverse populations.
Uneven coverage of regions	In some countries in Africa, the Middle East or Oceania, data is not collected every year.
Accessibility to technology	Surveys via phone or the Internet do not take into account people without access to communication.
Cultural biases	Respondents in different cultures interpret the happiness scale differently (the effect of modesty or exaggeration).
Linguistic/semantic features	The meanings of the words "happiness" and "contentment" may differ in languages, despite the translation.
Socially Desirable Answers	Respondents may underestimate or overestimate due to social pressure or expectations.

The data was obtained from the World Happiness Report, based on the results of the annual Gallup World Poll. The sample includes approximately 1000 people per country and is representative of the adult population. However, limitations of representativeness may be related to uneven coverage of regions, cultural interpretations of the concept of "happiness", and potential social biases in responses.

3.7. Clustering and classification of sociological data

The features of the dataset are presented in different numerical ranges. Therefore, to perform clustering and classification of data, it is necessary to normalize the feature values. To do this, click the Normalization button. After that, the normalized attribute values will be displayed in the table (Fig. 3).

After normalization, the system provides for clustering using the k-means algorithm to identify the internal structure of the data. To do this, click the k-means clustering button. The result of the clustering will be displayed visually in the form of a graph and in a table, in which the number of the clusters to which it falls will be indicated opposite each country.

We can see the graph of the clustering results, in which you can change the angle to choose the best viewing point for the clustering method data in the three-dimensional feature space (Fig. 4a), which does not fully reflect the result obtained since the analysed data set contains six features. The graphs below show the clustering result in the coordinate system of the features: Social support, Freedom to make life choices, and Perceptions of corruption.

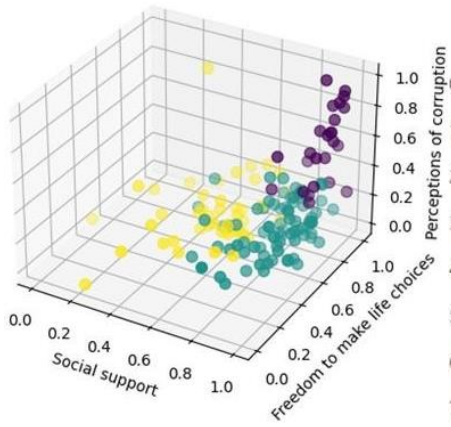
	Country name	og GDP per capit	Social support	althy life expectar	om to make life cl	Generosity	eptions of corru
0	Afghanistan	0.195685909	0.230318769	0.233826851	0.0	0.237331442	0.002299084
1	Albania	0.590009226	0.536638253	0.743820535	0.666328665	0.300146796	0.047567669
2	Algeria	0.614219244	0.738580768	0.655132002	0.121083791	0.20869103	0.242310211
3	Argentina	0.669279403	0.886903934	0.746847458	0.751280312	0.123023486	0.113314588
4	Armenia	0.525981002	0.668518111	0.681883926	0.545351205	0.188176825	0.196222045
5	Australia	0.852747353	0.954495542	0.898747449	0.897019777	0.570315521	0.630195457
6	Austria	0.857230474	0.928841512	0.879698509	0.870322665	0.448409281	0.527524494

Fig. 3. Normalized feature values of the dataset

For a generalized display of the clustering results, a table was formed and displayed that contains the average values of each feature of the clusters - centroids (Fig. 4b). Here 0 – Log GDP per capita, 1 – Healthy life expectancy at birth, 2 – Social support, 3 – Freedom to make life choices, 4 – Generosity, 5 – Perceptions of corruption, 6 – Positive affect, 7 – Negative affect. Analysing the data obtained, we see that the countries classified in cluster 1 are characterized by the level of GDP per capita, social support, freedom to make life choices and generosity. In cluster 3, these characteristics are slightly lower but satisfactory, and in cluster 2, they are the lowest. For cluster 1, the positive effect is

the highest, and the adverse impact is the lowest. For cluster 2, the opposite is true: the positive impact is the lowest, and the adverse effect is the highest. For cluster 3, the value of these features is in the middle. Since the level of happiness focuses on the social, urban and natural environment and the emotional perception of events, it can be argued that cluster 1 includes countries with a high level of happiness, cluster 3 – with an average level, and cluster 2 – with a low level of happiness. Based on this, it can be seen that the highest level of perception of corruption characterizes countries with a high level of happiness. The expected healthy life expectancy at birth for countries with a high level of happiness turned out to be the highest and the lowest – for countries with a low level of happiness. For a more detailed analysis of the clustering results, the system displays a table indicating which cluster each country falls into (Fig. 4c). The table sorts the countries alphabetically, which is convenient for searching and further analysing the data. Countries with a high level of happiness include Canada, the Netherlands, Sweden, Switzerland, the United Kingdom, and the United States. Countries with an average level of happiness include Turkey and Latvia. Countries with a low level of happiness include Pakistan, Azerbaijan and Laos. Having analysed the composition of the clusters, we can conclude that the level of happiness in a country correlates with the level of its economic development. Having the results of clustering, we can proceed to classify countries that were not involved in the analysis. To do this, we will establish a correspondence:

- class 1 will correspond to cluster 1 (high level of happiness);
- class 2 will correspond to cluster 2 (low level of happiness);
- class 3 will correspond to cluster 3 (medium level of happiness).



(a)

	Cluster 1	Cluster 2	Cluster 3
0	0.859330692	0.315940103	0.652545413
1	0.479598949	0.358506213	0.270465124
2	0.926174958	0.550911609	0.830756852
3	0.859525404	0.363138611	0.709231427
4	0.876653112	0.559376879	0.684852865
5	0.628432575	0.206157406	0.15785944
6	0.861894825	0.354570083	0.620884001
7	0.32960967	0.856171118	0.606653908

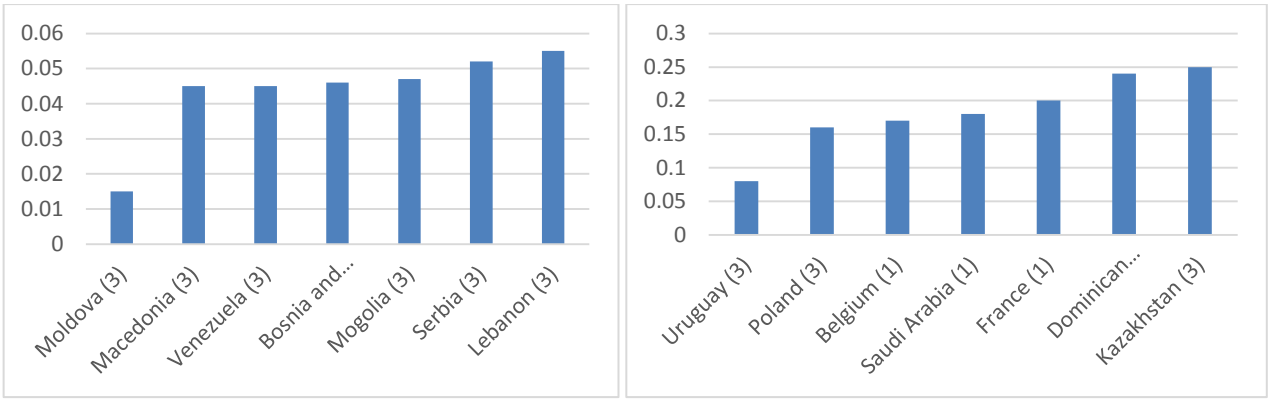
(b)

	Country name	Cluster
0	Afghanistan	3
1	Albania	2
2	Algeria	2
3	Argentina	2
4	Armenia	2
5	Australia	1
6	Austria	1
7	Azerbaijan	2
8	Bahrain	1
9	Bangladesh	3

(c)

Fig. 4. (a) Initial display of the cluster graph, (b) defined cluster centroids and (c) table showing the countries belonging to the clusters

To classify countries, click the k-NN classification button. After that, the algorithm for the k-nearest neighbours will be launched, and the countries that turn out to be closest to the one being classified will be displayed in a graphic form. Fig. 5a shows the results of the countries that turned out to be closest to Ukraine. As we can see, all countries belong to the third class. Therefore, using unweighted voting, Ukraine can be attributed to countries with an average level of happiness. Applying the classifier to classify the country Estonia, we found that it is close to one country of the first class and six countries of the third class. Therefore, it can also be attributed to countries with an average level of happiness (see Fig. 5b).



(a) (b)

Fig. 5. The result of the classification of the country Ukraine/Estonia by level of happiness

Below is a detailed analysis of potential problems with the quality of sociological data in the context of the study, based on the general principles of sociological data processing and the structure of the data set from the World Happiness Report.

1. Missing values. In real sociological datasets, respondents may miss answers. Some indicators may be missing for specific countries or regions. In the data from the World Happiness Report, for example, missing values are not found in this particular set (Overall, we have no missing values (null)). But it's important to note that this only applies to the processed version, and raw survey data (especially online) often contains significant gaps. If the data has been cleaned beforehand, this may lead to a loss of objectivity – the removed countries/respondents with problematic data may have specific characteristics. It can potentially skew happiness index scores, especially if the omissions are not random. If there are missing values, it is necessary to analyse the type of omissions: MCAR, MAR or MNAR. It is also essential to consider padding methods: averages, regression imputation, or the use of skip-resistant models (e.g., XGBoost).

2. Noisy data is random or inconsistent values in surveys, for example, when a respondent answers frivolously, clicks on random answers, or technical errors occur when collecting. In a sociological context, noise can be abnormal values, such as a very high happiness index in countries with a critical social situation (due to emotional perception). Also, too much variance in the data is considered noise, which distorts classification and regression models. As a result of noise intrusion, the accuracy of clustering decreases (K-means is sensitive to noise and emissions). Noise can be misleading when predicted by regression models. The main recommendation is to apply the analysis of the distributions of variables (boxplot, histogram) to detect outliers. Noise reduction methods such as emission removal, logarithmic transformation, or robust models (for example, RANSAC) are necessary.

3. Bias is a systematic distortion in the sample that affects the representativeness of the results. Possible sources of prejudice are geographical, socioeconomic and cultural. Geographical bias is that some regions of the world (especially developing countries) may be underrepresented. Socioeconomic bias stems from the fact that people with access to the Internet (online surveys) are predominantly those with higher income/education. Cultural bias is based on the fact that the concept of "happiness" can have different meanings in different cultures (subjectivity of the answer). The consequence of the presence of such biases is that the comparison between countries becomes incorrect or distorted. A parallel regression model may overestimate or underestimate the impact of certain factors (e.g., GDP vs. social support). It is then necessary to carry out a stratified analysis by region or income level and use bias correction methods, such as weight modelling or resampling (SMOTE, bootstrap). It is also necessary to take into account culturally specific interpretations of the happiness index.

Although there are no apparent gaps in the processed World Happiness Report dataset, real sociological data can still potentially contain structural distortions in the form of noise, biases, and anomalies. Without an in-depth analysis of these factors, classification and prediction results may be unreliable or biased.

The study uses data from the World Happiness Report, which is one of the most well-known global sources for measuring happiness. However, despite its authority, the use of this dataset has a number of significant limitations. Below is a detailed analysis of the main limitations of the global happiness dataset.

1. Regional data gaps. Some regions (especially countries in Africa, the Middle East, and parts of Asia) may be poorly represented or completely absent from the dataset, due to political instability, lack of resources for data collection, and lack of access to respondents. Therefore, the data are not entirely representative of the global comparison. Also, clustering and regression models can shift towards countries with full, stable and regular dimensions, such as European or North American states.

2. Cultural bias in the perception of happiness. The concept of "happiness" is deeply culturally conditioned and can have different meanings in different societies. In Western cultures, happiness is often associated with personal achievement and self-realisation, and in East Asian cultures, it is associated with harmony and collective harmony. In some traditional cultures, there is stability and a lack of stress. Therefore, survey results based on standardised questionnaires may not reflect the real level of life satisfaction in the context of local values. Accordingly, it is possible to underestimate/overestimate the happiness index due to different interpretations of scales (for example, extreme ratings are avoided).

3. Bias in data collection methods. Most of the data in the World Happiness Report comes from Gallup polls, which are often conducted over the phone or the Internet. They also have limitations in the sample: people without access to communication/network are not taken into account. Therefore, social minorities, people with low incomes or rural residents may be underrepresented. Urbanised, educated, wealthy respondents may prevail in the sample, distorting average ratings.

4. Data aggregation at the national level. All indicators in the set are averaged across countries, without taking into account internal regional diversity. The level of happiness in countries with high social or regional inequality (e.g. Brazil, India, USA) can be misleading because it does not show the difference between population groups.

5. Time variability and instability of indicators. Data collection takes place once a year, which does not allow tracking rapid changes in sentiment caused, for example, by wars, pandemics, or economic crises. There is an impression of stability that may not exist – in particular, for countries with active changes (for example, Ukraine after 2022).

Although the global data from the World Happiness Report is user-friendly and standardised, its use is accompanied by certain structural limitations, including:

– regional and demographic gaps in the sample;
– cultural heterogeneity in the perception of happiness;
– averaging and loss of parts;
– possible distortions caused by data collection methods.

These limitations should be taken into account when interpreting the results and formulating policy recommendations. Otherwise, the conclusions may be unreliable or even false.

Below is a detailed explanation of how exactly the K-means clustering algorithm for grouping countries by happiness indicators works with multidimensional data. We will also discuss the key assumptions and limitations of its application. K-means processes multidimensional happiness data. The study analyses a dataset containing several numerical variables such as: Logged_GDP_per_Capita, Social_Support, Healthy_Life_Expectancy, Freedom_to_Make_Life_Choices, Generosity, Perceptions_of_Corruption. These features form a multidimensional space, where each country is a point with coordinates according to its values for each variable. The K-means algorithm does the following:

Step 1. Data normalisation — all variables are brought to the same scale so that none dominates.
Step 2. Cluster initialisation, in particular, randomly specifies the number of cluster centres (k).
Step 3. Calculating distances, i.e. for each point (country), the Euclidean distance to each centre of the cluster is calculated.
Step 4. Assignment to a cluster, for example, each country is assigned to the nearest centre.
Step 5. Renewal of cluster centres, in particular, centres move to the middle position of the new groups.
Step 6. Repetitions, i.e. stages 3–5, are repeated until the centres stop changing (convergence).

Thus, the algorithm works efficiently in spaces with many features, bringing together countries with similar socioeconomic profiles.

Assumptions and limitations of K-means that affect the applicability of the results:

1. Assumptions about the shape of clusters: spherical only. K-means works best when the clusters are roughly the same size, spherical (radial) shape, and have the same dispersion across all traits. In real happiness data (with many factors of different nature), these assumptions are often violated. For example, countries with high GDP but low freedom can create elongated or uneven clusters that K-means will incorrectly combine or split.

2. The use of Euclidean distance. The algorithm applies the Euclidean metric as a measure of similarity between objects. Such a metric does not take into account correlations between features. For example, if GDP and social support are firmly linked, the Euclidean distance can overestimate the difference between countries that differ only slightly in one of the factors.

3. Sensitivity to the choice of the number of clusters (k). The algorithm requires you to set k in advance, which is not always intuitive. The manuscript mentions the use of the "elbow method", but this method does not guarantee optimal k. Choosing a bad k can lead to mixed or artificial clusters that do not reflect real social structures.

4. Sensitivity to emissions and noise. K-medium is not emission-resistant. A country with atypical values (e.g., a very high level of freedom with a low GDP) may distort the cluster centre calculation and be misclassified. In sociological data, emissions are commonplace, especially in countries with non-standard development models (for example, Bhutan or Singapore).

5. Linearity and lack of consideration of complex relationships. The algorithm does not detect nonlinear dependencies between variables. Many factors of happiness have a nonlinear influence (for example, the effect of GDP growth on happiness has saturation).

The K-means algorithm provides a quick and interpretable grouping of countries in a multidimensional social space. However, its performance is limited by assumptions about cluster shape, uniformity of variance, metric data properties, and the absence of outliers. These limitations reduce the reliability of conclusions if they are not taken into account in the research methodology and are the next future of direct research.

To ensure the reliability, reproducibility and generalizability of the research results, it is advisable to use the following verification methods:

– K-fold cross-validation for the regression model.
– Train/Test Split or Stratified Cross-Validation for the classifier.
– Internal clustering quality metrics for K-means.
– Testing on new or synthetic data as proof of the model's ability to work in real-world conditions.

Below is a proposal for a methodology for verifying the reliability and generalizability of the obtained research results, which is critical for evaluating the reliability and generalizability of models.

Step 1. Cross-Validation for the Multiple Regression Model. Since the paper uses multiple linear regression to predict the happiness index, it is advisable to use k-fold cross-validation:

Step 1. Data from the World Happiness Report (149 countries) are broken down into k subsets (e.g., k = 10).

Step 2. In each iteration, k – 1 part is used to train the regression model, and 1 part is used for validation (testing).

Step 3. The procedure is repeated k times, which allows you to estimate the average error of the model on new (previously unseen) data.

The evaluation metrics are: RMSE (root of standard error), MAE (mean absolute error) and R^2 (coefficient of determination). The advantage of Phase 1 is that it provides a stable assessment of model quality on different pieces of data. It makes it possible to detect overfitting.

Step 2. Classifier Test (k-NN) on Deferred Data (Train/Test Split). After clustering K-means, the article creates a classifier of countries by the level of happiness using the method of k-nearest neighbours (k-NN). To assess its generalizability:

Step 1. Divide the data set into training and test samples, for example: 80% for training the classifier, 20% for testing.

Step 2. Train k-NN on the training data.

Step 3. Test on deferred data and evaluate the quality of classification.

The evaluation metrics are Accuracy, Precision / Recall / F1-score (in case of imbalance) and Confusion matrix. An alternative is to apply stratified k-fold cross-validation so that each subset contains all cluster classes proportionally.

Step 3. Clustering validation (K-means). Although K-means is an unsupervised method, there are approaches to assessing the quality of clustering, including internal validation methods. The Silhouette Score measures how similar objects are to their cluster and different from others. The Davies–Bouldin Index assesses the compactness and separability of clusters. Calinski-Harabasz Index is the ratio of intervariance and intracluster variance. External validation methods (if an accurate/benchmark is available) consist of comparing clusters with existing country categories (e.g., by income level or region).

Step 4. Testing on new or synthetic data to test how the model performs on data it has not seen before.

Step 1. Collect or generate new country data (e.g. from a later World Happiness Report).

Step 2. Predict happiness using a constructed regression model or classify countries through a classifier.

Step 3. Compare the predictions with the actual data of the new year.

4. Experiments, Results and Discussion

4.1. Forms and methods of presentation and preliminary statistical processing of numerical data of time series

The study developed an intelligent system for analysing happiness in Python using the libraries pandas, scikit-learn, and matplotlib. Deployment – local or in an environment such as Jupyter Notebook. Architecture – single-level, without complex APIs or databases, with the ability to classify and analyse.

The type of environment is a cross-platform web application and a desktop application. Possible platforms:

– Jupyter Notebook or Google Colab – for interactive analysis and prototyping (very likely, since there are links to Pandas, Matplotlib, and KMeans).

– Web-based GUI — for visualising results and user interaction (possibly with Flask or Streamlit).

– Backend system or console application – to run clustering, classification, normalisation, and graph output.

The following Python tools were used in the study:

– pandas – for processing tables (data from a CSV file);
– matplotlib.pyplot – for plotting;
– sklearn.cluster.KMeans – implementation of K-means clustering;
– sklearn.linear_model. LinearRegression – for regression;
– sklearn.neighbors.KNeighborsClassifier – a classifier based on kNN.

So, the implementation is based on a standard stack for machine learning in Python.

Table 3. Libraries and tools used

Category	Library/Tool	Appointment
Data Processing	pandas	Reading, filtering, and normalising data
Visualisation	matplotlib.pyplot	Cluster graphs, bar charts, charts
Clustering	sklearn.cluster.KMeans	Basic clustering algorithm
Regression	sklearn.linear_model	Building forecast models
Classification	sklearn.neighbors	k-closer neighbors to predict the cluster
Evaluation of models	sklearn.metrics	R ? classification accuracy (presumably)
Normalisation	sklearn.preprocessing	Feature scaling

Logical architecture of the system:

Process 1. Downloading data from a CSV/Excel file → pandas.
Process 2. Preprocessing: Filtering → normalising → eliminating gaps (if applicable).
Process 3. K-means clustering → automatic country grouping.
Process 4. Visualisation of clusters → colour graphs, diagrams.
Process 5. Building a regression model → determining the factors that affect happiness.
Process 6. Training of the kNN classifier → the ability to predict the belonging of new countries to the cluster.
Process 7. Output results via console or simple interface (Streamlit / Jupyter).

5. Possible extensions/directions of automation:
– Integration with Streamlit or Flask to create a simple web interface.
– Application of Docker for portability.
– Storing results in SQLite / JSON.
– Automatic updating of data from worldhappiness.report.

The paper considers methods of analysing statistical data and time series that are used to identify patterns and trends in the processes under study. The main attention is paid to methods of pre-processing, visualization, smoothing and correlation analysis to obtain quantitative and graphical characteristics of the data. Pre-processing of data begins with their organization in the form of files and tables. The data is organized in formats convenient for further processing, for example, in the form of a table of the "object - property" or "serial number - value" type. The file format can be presented in the form of text files with the extension .dat or .txt. Such data is imported into the Excel spreadsheet, where they are processed. In particular, tables are compressed, which provides a compact presentation of information and facilitates their perception. Graphical representation of data is an essential stage of visualization. The main types of graphs include histograms and cumulates. Histograms are used to analyse the distribution of sample values and allow you to see the structure of the data, including their mode and variation. When constructing histograms, the boundaries of the intervals and the frequency of values are determined. Formulas, such as the Sturges or Scott formula, can be used to select the number of intervals. Cumulative represents the integral distribution of values and allows you to estimate the accumulated frequency. They are constructed as graphs that show the percentage accumulation of data in the sample.

One of the key stages of analysis is smoothing time series, which allows you to detect trends by eliminating random fluctuations. The moving average method involves replacing each value of the series with the average value of neighbouring points in a specific window. It allows you to reduce fluctuations and make the trend more expressive. The weighted moving average, unlike the simple one, takes into account the weighting coefficients for each value, giving more weight to points located closer to the centre of the window. Median filtering is an alternative approach and is used to eliminate anomalous values, leaving the primary trend unchanged. It is especially effective for data with outliers, as it replaces each value with the median of neighbouring points in a given interval.

Correlation analysis is used to study the relationships between features in data. The main tools of this analysis include a correlation field, which graphically displays the relationship between two variables, and a correlation coefficient, which measures the strength and direction of this relationship. Autocorrelation is used to analyse time series, which allows you to assess the interdependence of values at different points in time. A correlation matrix is also calculated, which contains the correlation coefficients between all possible pairs of variables in the sample. An important task is to plot autocorrelation functions, which help to identify cyclicality or seasonality in the data.

Hierarchical agglomerative cluster analysis is used to group objects by similar features. The algorithm creates a hierarchical cluster structure, which is presented in the form of dendrograms. Grouping is carried out based on calculating the distances between objects, which allows you to detect structure in multidimensional data. Clustering helps to identify separate groups with similar characteristics and analyse their differences.

Methods for detecting trends in time series are based on smoothing and mathematical modelling. They include analytical approaches based on approximating data with functions such as linear, polynomial, or exponential models. In addition, algorithmic smoothing methods are used to estimate trends by calculating average values or weighting coefficients for individual points in the series.

In general, the work introduces the basics of statistical analysis and data visualization, which are necessary for processing experimental results and building forecasts. It allows you to master the methods of smoothing, correlation analysis, and clustering, which are essential tools for studying patterns and data behaviour.

4.2. Data pre-processing and results presentation

The data was converted to CSV format for ease of processing. The reporting table does not contain empty cells, which allows for easy processing (Fig. 6). Link to the dataset: https://www.kaggle.com/datasets/emirhanai/city-happiness-index-2024. The first diagram (Fig. 7a) shows the dependence of the noise level (Decibel_Level) on the air quality index (Air_Quality_Index). The X-axis shows the air quality index, and the Y-axis shows the noise level in decibels. There is a tendency for the noise level to increase with the increase in the air quality index. The distribution of points shows a noticeable relationship between these indicators, where higher values of the air quality index can be associated with increased noise levels.

The second diagram (Fig. 7b) illustrates the relationship between the healthcare index (Healthcare_Index) and the cost of living index (Cost_of_Living_Index). The X-axis shows the cost of living index, and the Y-axis shows the healthcare index. The graph shows an increasing trend, indicating a positive correlation. Higher indicators in the field of healthcare usually accompany higher cost of living. However, the scatter of the points suggests the presence of certain deviations, which may indicate the influence of additional factors.

	City	Month	Year	Decibel_Level	Traffic_Density	Green_Space_Area	Air_Quality_Index	Happiness_Score	Cost_of_Living_Index	Healthcare_Index
0	New York	January	2024	70	High	35	40	6.5	100	80
1	Los Angeles	January	2024	65	Medium	40	50	6.8	90	75
2	Chicago	January	2024	60	Medium	30	55	7.0	85	70
3	London	January	2024	55	High	50	60	7.2	110	85
4	Paris	January	2024	60	High	45	65	6.9	95	80

Fig. 6. Fragment of the reporting table

The polar diagram (Fig. 8) displays the data collection activity by month. It is organized in the form of a radial grid, where each sector corresponds to a particular month of the year, and the distance from the centre indicates the level of activity. The graph shows that the highest level of activity is observed in March, with some decrease in February and April. The summer months (June, July, August) demonstrate a significantly lower level of activity, and the lowest value falls in October. The filled area emphasizes the general distribution of activity, indicating its concentration in the first quarter of the year. Such a graph makes it easy to identify seasonal trends and peak periods in data collection.

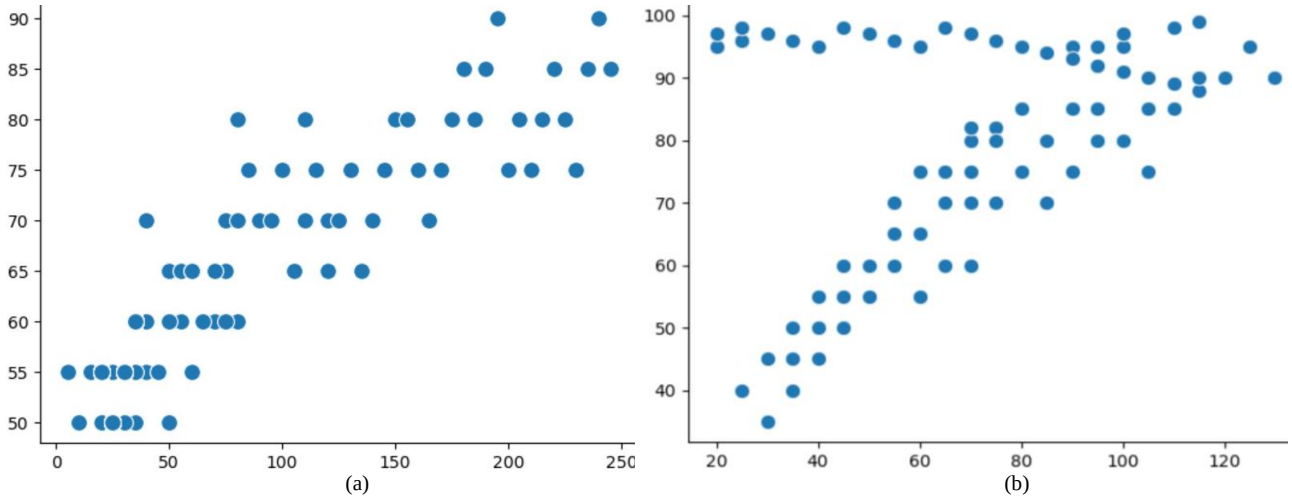


Fig. 7. Diagram in Cartesian coordinate system: a- dependence of air quality indicator X on noise indicator Y and b - dependence of cost of living indicator X on health indicator Y

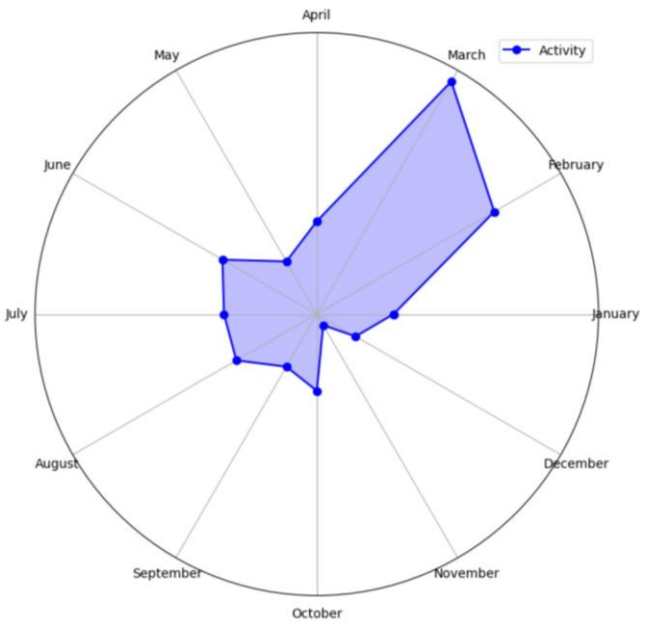


Fig. 8. Diagram of collected data by months in polar coordinate system

The table (Fig. 9) contains descriptive statistics for the leading indicators of the dataset. It includes the average value (mean), standard deviation (std), minimum (min) and maximum (max) values, as well as the coefficient of variation (cv %) in per cent. The average value (mean) shows the typical level for each indicator. For example, the average noise level (Decibel_Level) is 56.78, and the average air quality index (Air_Quality_Index) is 38.04.

	Year	Decibel_Level	Green_Space_Area	Air_Quality_Index	Happiness_Score	Cost_of_Living_Index	Healthcare_Index
mean	2026.082569	56.779817	1085.366972	38.036697	-44.865505	30.458716	93.086239
std	1.652363	6.856402	756.993165	36.300656	42.407240	21.082180	10.550034
min	2024.000000	50.000000	5.000000	5.000000	-122.900000	20.000000	35.000000
max	2029.000000	90.000000	2425.000000	245.000000	8.600000	130.000000	99.000000
cv (%)	0.081555	12.075422	69.745366	95.435878	-94.520813	69.215591	11.333613

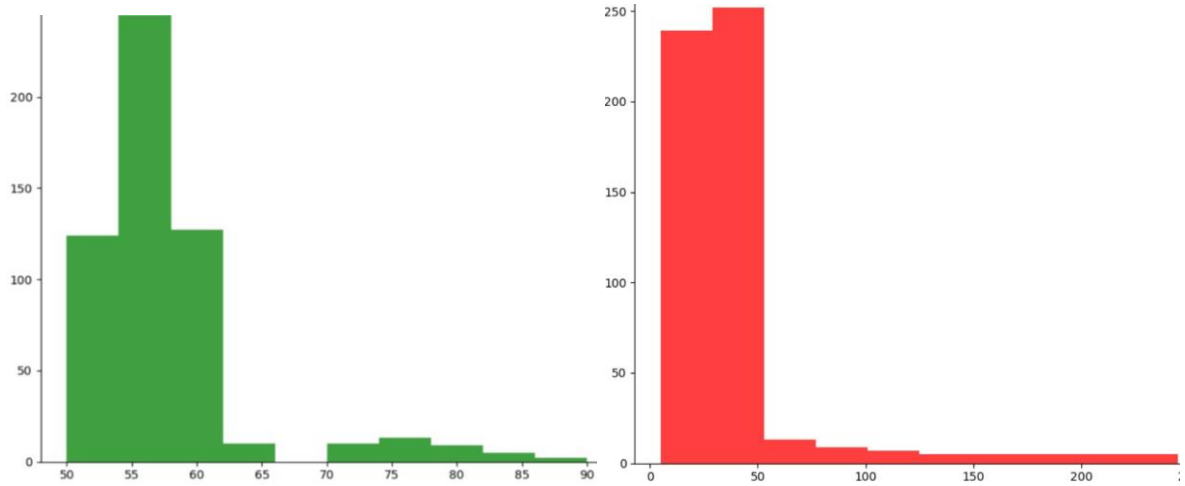
Fig. 9. Table of descriptive statistics

The standard deviation (std) indicates the degree of dispersion of the data. The most significant deviation is observed in the happiness index (Happiness_Score) – 42.41, which means considerable variability of the data.

The minimum (min) and maximum (max) values demonstrate the range of indicators. For example, the cost of living index (Cost_of_Living_Index) varies from 20 to 130, while the level of healthcare (Healthcare_Index) varies from 35 to 99. The coefficient of variation (cv %) characterizes the relative variability of the data in percentage. The most significant variability is observed for the air quality index (95.44%) and the cost of living (69.22%). It indicates substantial deviations from the average values in these indicators. In contrast, the healthcare index has the lowest variability (11.33%), which means the relative stability of this indicator.

The histograms (Fig. 10) display the distributions of the leading indicators from the dataset, allowing a visual assessment of the frequency of values in each column. The first histogram (Fig. 10a) shows the distribution of the noise level (Decibel_Level). Most of the values are concentrated in the range of 50–60 dB, which indicates a high concentration of data in this interval. A small number of values exceed 70 dB, which may indicate the presence of emissions or anomalous points. The second histogram (Fig. 10b) displays the distribution of the air quality index (Air_Quality_Index). Most of the values are below 50, but there are a small number of cases with very high values, up to 245, which may indicate emissions or specific pollution cases. The third histogram (Fig. 10c) shows the distribution of the cost of living index (Cost_of_Living_Index). The bulk of the data is concentrated in the range of 20–40, indicating a low cost of living for most observations. However, there are a few higher values, indicating the presence of diverse economic conditions.

The fourth histogram (Fig. 10d) shows the distribution of the healthcare index (Healthcare_Index). The data is mainly concentrated around 100, indicating the high quality of healthcare services in the sample. Low values are rare, which may indicate specific regions with poorer healthcare infrastructure. In general, all graphs (Fig. 10) show strong asymmetry of the distributions, especially for the air quality and cost of living indices, which may require logarithmic transformation or other normalization for further analysis.



(a) (b)

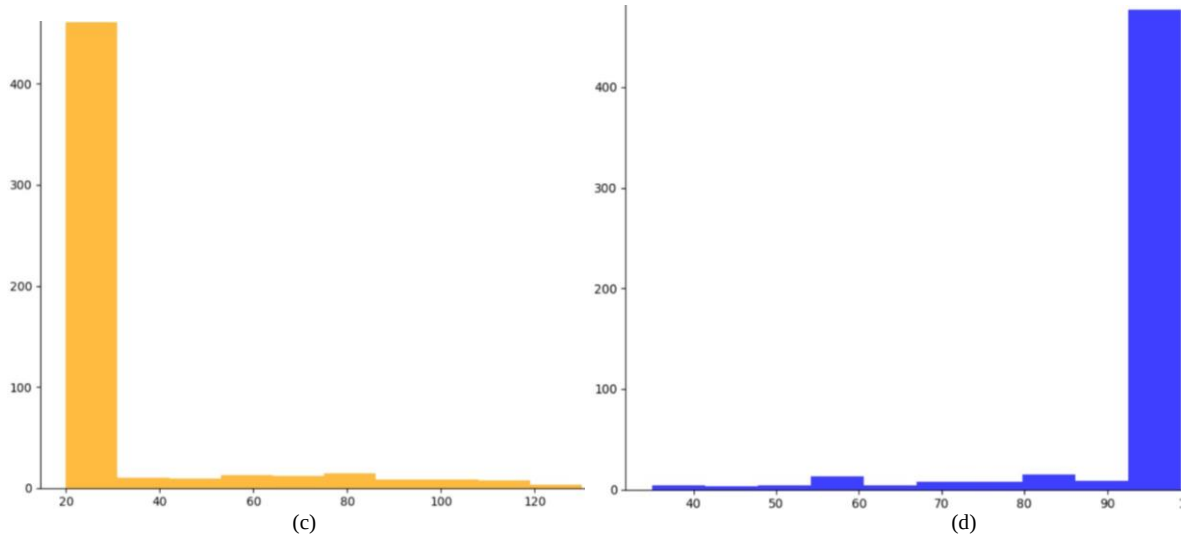


Fig. 10. Histograms of the distribution of indicators: a - noise, b - air quality, c - cost of living and d - health

The graph (Fig. 11) presents boxplots for the four indicators, illustrating the presence of outliers in the data. Decibel_Level has a small number of outliers that exceed the upper limit of the range, but the bulk of the values are concentrated in the range of 50–60 dB.

The Air_Quality_Index shows a significant number of outliers that are well above the upper limit, reaching values above 200. It indicates strong asymmetry and potential outliers. The Cost_of_Living_Index also has numerous outliers located above the main range, although most values are in the range of 20–40. It suggests possible regions with abnormally high costs of living. The Healthcare_Index shows a relatively compact distribution, but some outliers are present below the main range, which may indicate areas with poor healthcare.

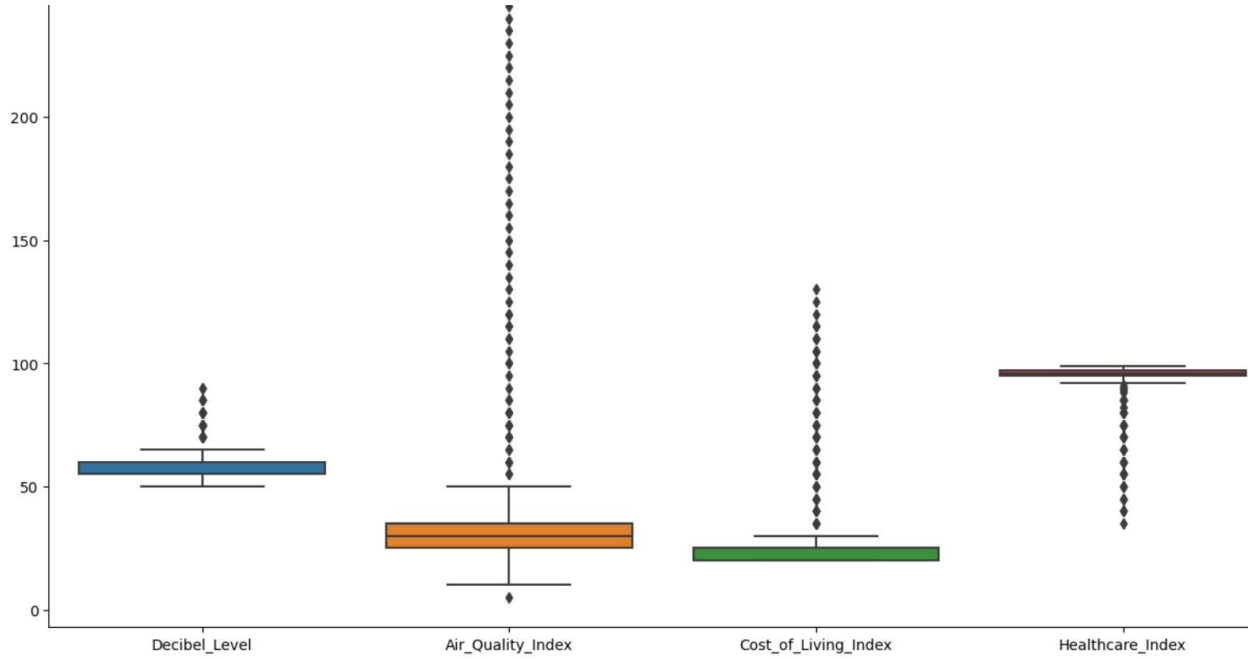
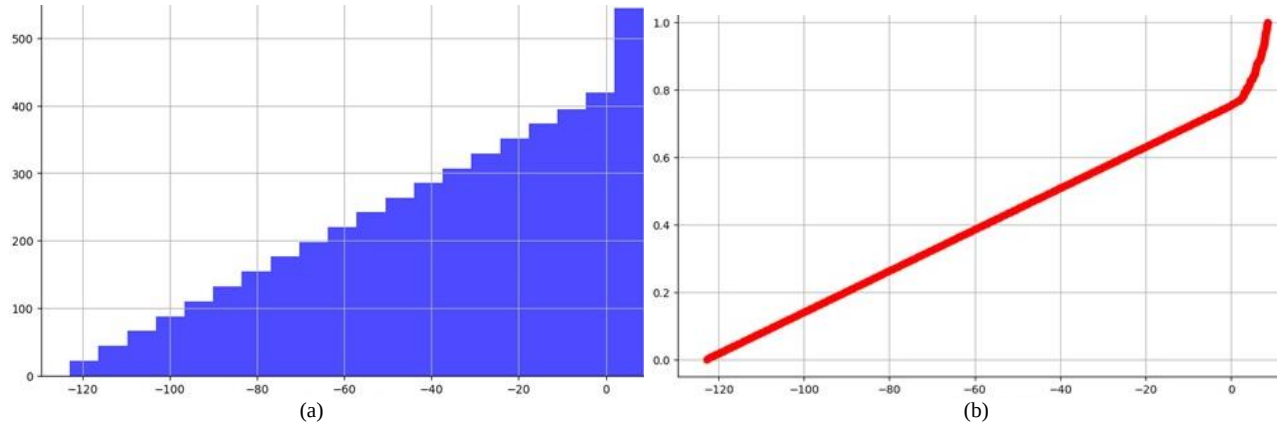


Fig. 11. Box plot of emissions

The first chart (Fig. 12a) presents the cumulated happiness level (Happiness_Score) based on the histogram data. It shows the accumulated frequency of values, which allows us to assess how the data is distributed in the population. The graph shows a gradual increase in frequency with small jumps, indicating a uniform distribution of most of the data with a sharp rise at the right end of the range. It shows a larger cluster of values closer to the upper limits of the happiness indicator.



(a) (b)

Fig. 12. Cumulative happiness level (according to histogram data and integral percentage)

The second chart (Fig. 12b) shows the cumulative happiness level plotted as an integral percentage. It shows the normalized distribution function in the form of a probability curve. The curve gradually increases, approaching 1 (100%), indicating a stable accumulation of values. However, in the last segments, there is a sharp increase, indicating the presence of relatively small but significant outliers in the positive region. These charts help to assess the nature of the data distribution and identify trends, including asymmetry or clustering of values. They are helpful in understanding the structure of the data before further analysis or modelling.

The chart (Fig. 13) illustrates the cumulative sum for the indicator of green space area (Green_Space_Area). It shows the gradual accumulation of green space area values as they increase. The chart has the form of an increasing curve, indicating a nonlinear distribution of the data. The initially slow growth suggests a small number of records with low area values. In contrast, the more rapid growth towards the end shows a significant accumulation of large values of green area. This plot is helpful in estimating the total amount of green area and identifying thresholds beyond which a rapid increase in contribution to the total is observed. It can also indicate the presence of clusters or regions with concentrated large areas of green area.

The graph (Fig. 14) shows the smoothing of Decibel_Level indicators according to the Kendall formula for different parameters (Kendall 3, 5, 7, 9, 11, 13, 15). The original series shows high volatility, while the smoothed series becomes more stable with increasing Kendall parameters. It indicates the effectiveness of smoothing in reducing data noise. For higher values of the parameter, an almost linear behaviour is observed, which ignores short-term fluctuations.

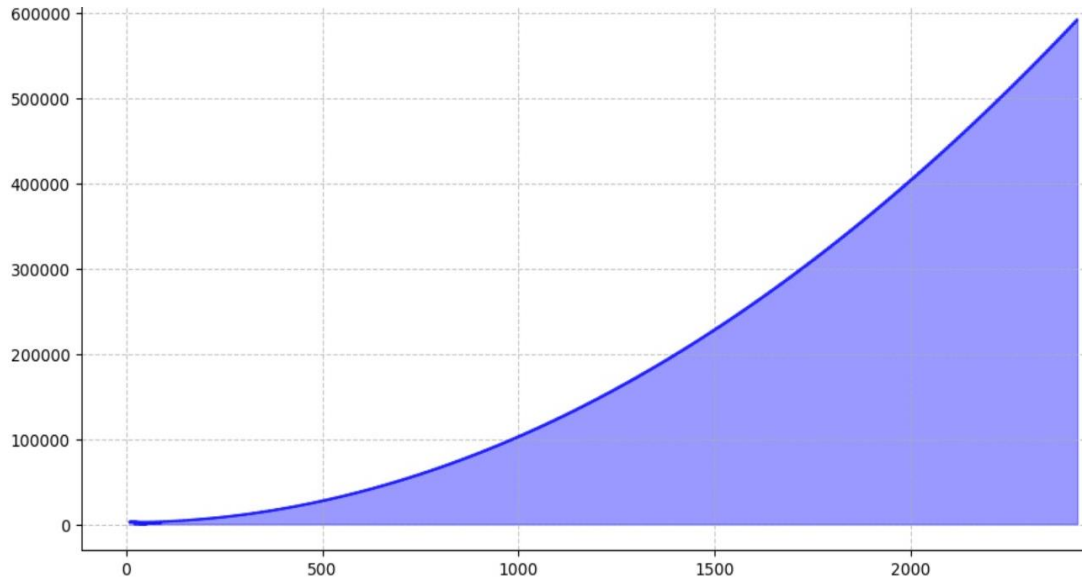


Fig. 13. Cumulative sum for the indicator of green space area (Green_Space_Area)

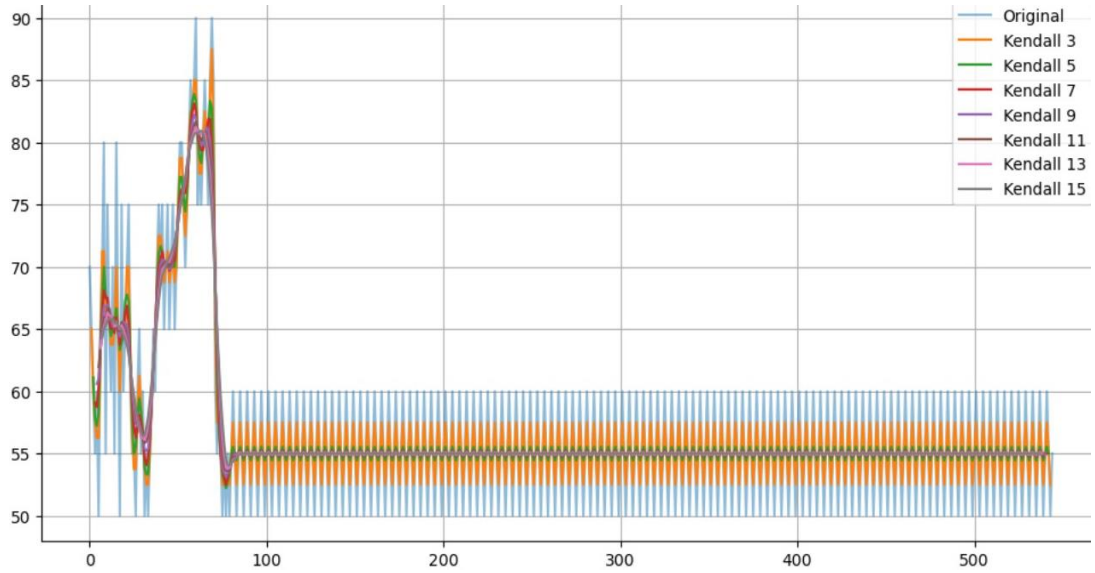


Fig. 14. Smoothing Decibel Level indicators using Kendall formulas

The diagram (Fig. 15) shows the correlation matrix for the original and smoothed data. The original series has a strong correlation with the Kendall 3 series (0.95), but with increasing parameters, the correlation gradually decreases, reaching the lowest values with Kendall 15 (0.81). It indicates that with an increasing degree of smoothing, the data lose their exact correspondence to the original, which is the expected result of smoothing.

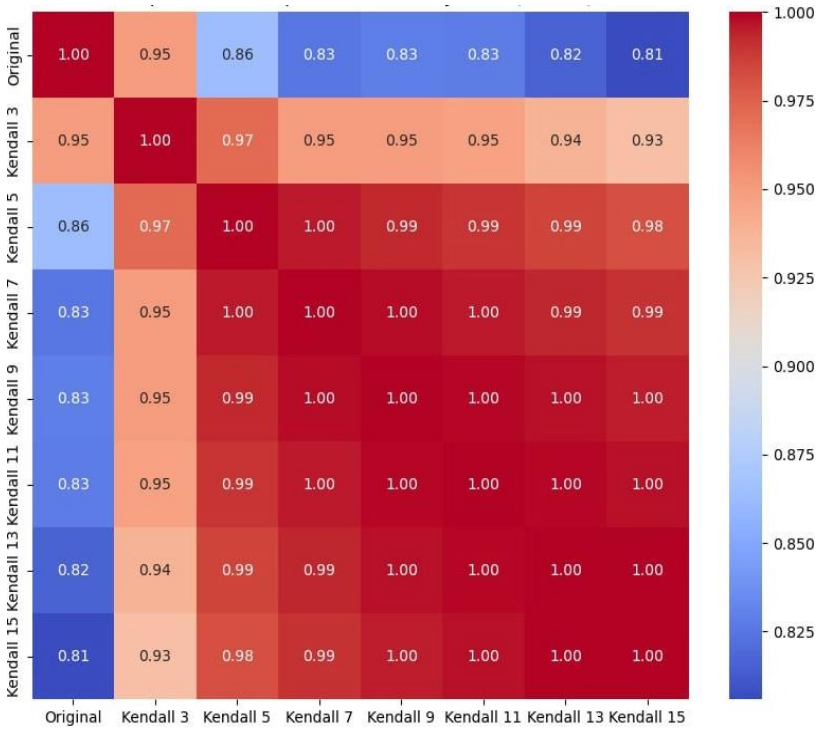


Fig. 15. Correlation matrix for smoothing by Kendall formulas

The graph (Fig. 16) shows the autocorrelation for the original series and the smoothed data. For the original data, the autocorrelation decreases sharply at the initial lags, which indicates high variability. In the smoothed series, the autocorrelation increases with increasing Kendall parameters, demonstrating a more pronounced dependence between neighbouring values. It means that smoothing strengthens the temporal relationships in the series, making it more predictable. In conclusion, Kendall smoothing is effective in reducing noise and increasing autocorrelation, but it is essential to choose the parameter carefully to avoid losing important information. The correlation matrix allows us to assess the degree of similarity between series, and the autocorrelation demonstrates the effect of smoothing on the temporal structure of the data.

Smoothing with Pollard formulas

a) smooth the data using the smoothing interval sizes w = 3, 5, 7, 9, 11, 13, 15. We should get seven columns in a row;

b) smooth the data using the smoothing interval size w = 3, then smooth the obtained smoothed data again, but use the smoothing interval size w = 5. We continue smoothing the received data with the smoothing interval w = 7 and so on until w = 15. We should get seven columns in a row.

c) in both cases, find the number of turning points and the correlation coefficients between the original values and the smoothed ones for each smoothing.

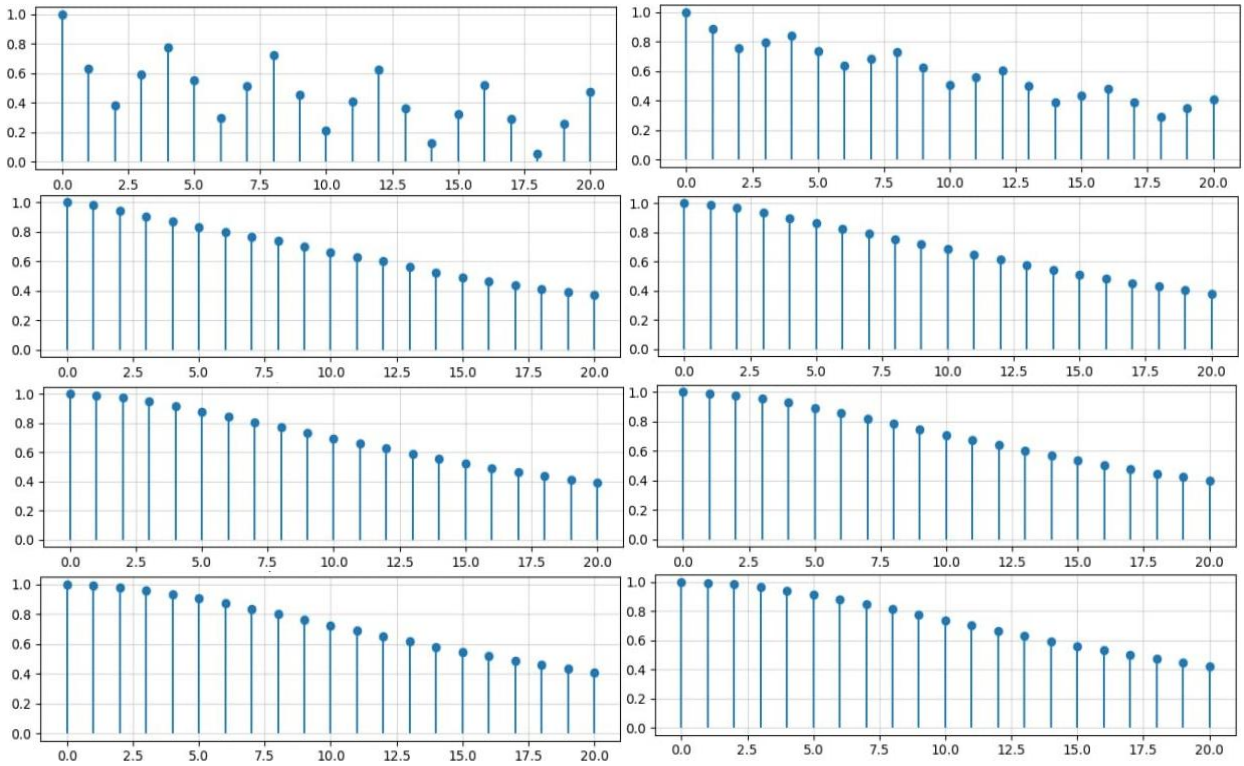


Fig. 16. Autocorrelation (right-swept and top-down for indicators original, w=3, w=5, w=7, w=9, w=11, w=13, w=15)

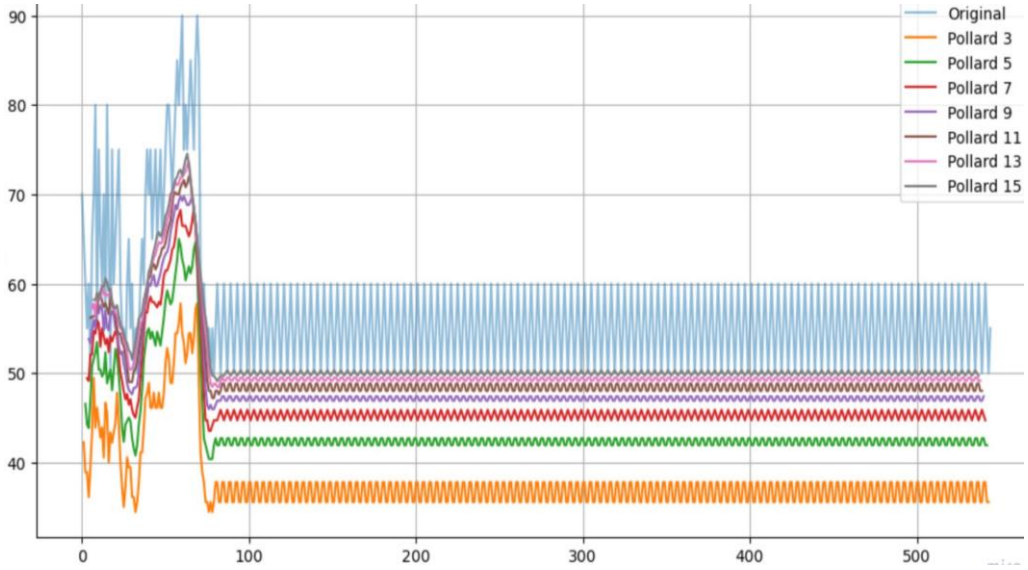


Fig. 17. Smoothing Decibel_Level indicators using Pollard formula

The graph (Fig. 17) demonstrates the smoothing of the noise level (Decibel_Level) using the Pollard formula for different parameters. The blue line represents the original data, which has high variability and strong fluctuations, especially at the initial indices. The other lines display the smoothed data for different values of the Pollard parameter (from 3 to 15). As the smoothing parameter value increases, the curves become less fluctuating, reflecting a gradual decrease in the influence of local variations and noise. For example, the Pollard 3 line (orange) shows significant smoothing but loses some detail. While Pollard 15 (purple) preserves more detail, it still reduces the noise level compared to the original data. This approach allows you to choose the optimal level of smoothing depending on the needs of the analysis. The graph shows how you can effectively reduce fluctuations and emphasize the primary trend in the noise level data.

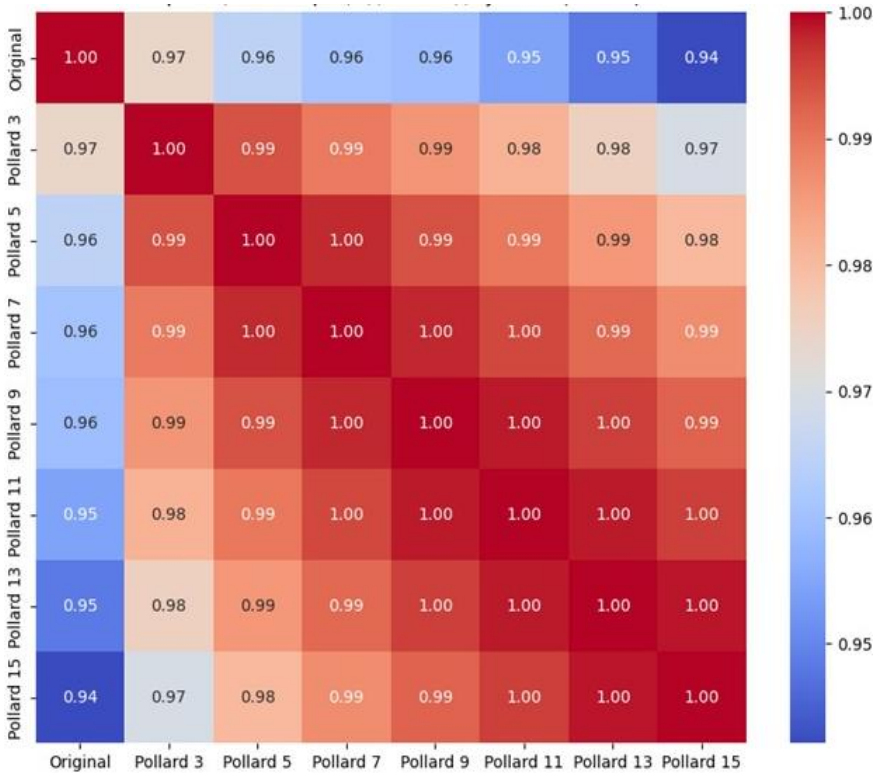


Fig. 18. Correlation matrix for smoothing using Pollard formula

The diagram (Fig. 18) presents the correlation matrix for different levels of smoothing performed using the Pollard formula. The matrix demonstrates the relationship between the original data (Original) and the smoothed series for Pollard parameters (from 3 to 15). The correlation values vary from 0.94 to 1.00, which indicates a strong relationship between the original and smoothed data. Darker red cells indicate high correlation coefficients (close to 1), and blue ones indicate slightly lower values. The correlation between the original data and the smoothed series decreases with increasing Pollard parameters. For example, for Pollard 3, the correlation with the original data is 0.97, and for Pollard 15, it is 0.94. It confirms that stronger smoothing reduces the similarity with the original data but still maintains a high level of dependence. The high correlation between the smoothed series (values 0.98–1.00) shows that the Pollard method provides stable smoothing even when the parameter is changed. It allows you to choose the smoothing parameter depending on the desired level of detail without significantly distorting the data structure.

The diagram (Fig. 19) contains autocorrelation plots for the original data (Original) and the smoothed series for different Pollard parameters (from 3 to 15). The autocorrelation shows the relationship between the current and previous values of the time series (lags). All plots show a gradual decrease in autocorrelation with increasing lag, indicating the presence of a trend or serial dependence in the data. The original data (Original) has the highest autocorrelation values at slight lags, indicating a strong relationship between neighbouring values. As the Pollard parameter increases, the autocorrelation becomes less pronounced, especially for higher lags. It suggests that the smoothing reduces short-term fluctuations, making the series more stable. The plots for Pollard 3 and Pollard 5 show relatively high values of autocorrelation even at considerable lags, while Pollard 15 shows significant smoothing, reducing autocorrelation for long lags. It confirms that Pollard smoothing gradually eliminates short-term fluctuations while preserving the underlying structure of the data and long-term trends. The higher the smoothing parameter, the more the series is smoothed, and the autocorrelation is reduced.

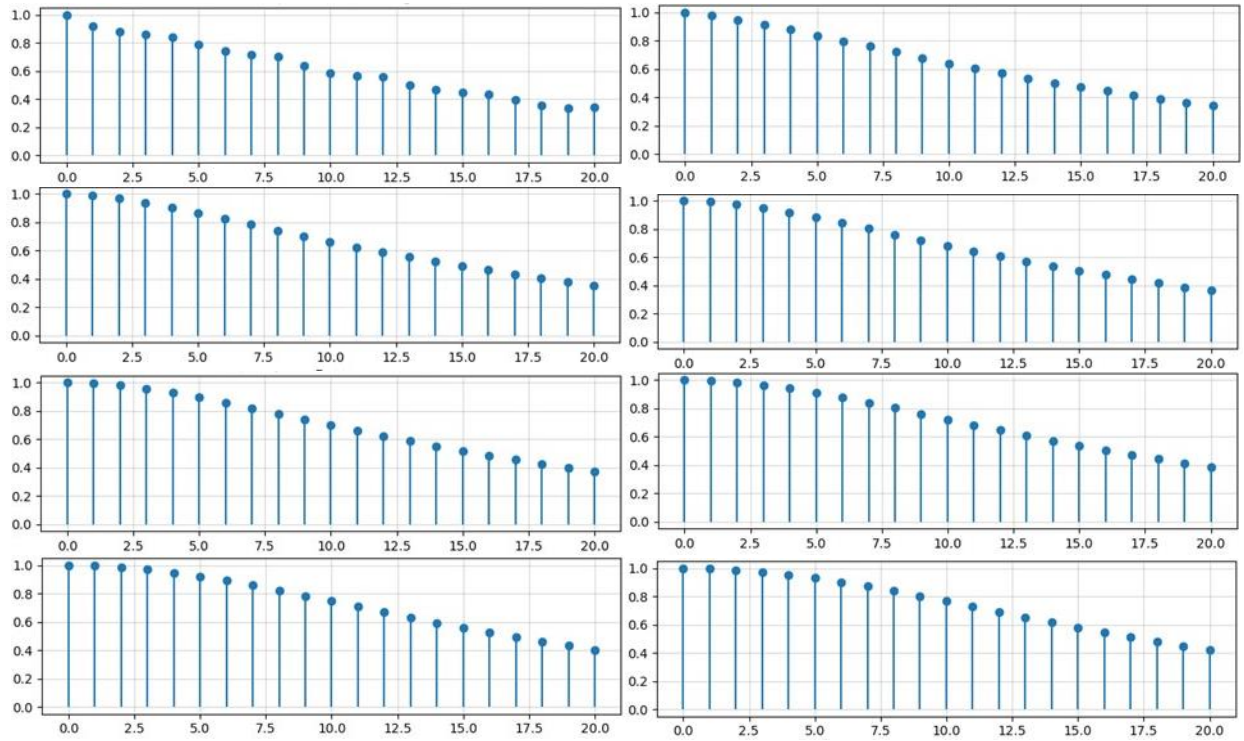


Fig. 19. Autocorrelation plots for smoothing using Pollard's formula (scroll to the right and top to bottom for the indices original, w=3, w=5, w=7, w=9, w=11, w=13, w=15)

Fig. 20 presents the results of hierarchical cluster analysis. The dendrogram (Fig. 20) illustrates the process of combining objects into clusters depending on their similarity. The distance between objects, displayed on the vertical axis, shows the level of differences between them. Different colours of the lines reflect the formed clusters, indicating groups of objects with similar characteristics. A clear cluster structure suggests the presence of several well-defined groups in the sample. The proximity matrix (Fig. 21a) contains numerical values of the distances between objects, which are used to assess similarity. It shows how close or far apart individual objects are from each other. Low values indicate similarity, while high values indicate differences.

The clustering results table (Fig. 21b) shows which cluster each object belongs to. In this example, the objects are distributed among five clusters. The cluster sizes (Fig. 21c) show the number of objects in each of them. The largest cluster contains 179 objects, while the smallest – 35, which indicates different densities of grouping. The cluster centres (Fig. 22) represent the average values for each indicator within the corresponding cluster. They help to understand the characteristic features of each group. For example, specific clusters have higher levels of air pollution or other parameters, which allows the identification of differences between groups. The results of cluster analysis will enable the identification of leading groups of objects with similar characteristics, which can be helpful for further study of trends and detection of anomalies in the data.

The bar graph (Fig. 23) shows the average value of the area of green zones (Green_Space_Area) for each year in the period from 2024 to 2029. The graph demonstrates a clear trend of growth in the location of green zones with each subsequent year. Starting from a small average value in 2024, the indicator increases significantly in 2025 and 2026, reaching a peak in 2029. Such a graph indicates a tendency to expand the area of green zones during the analysed period, which may indicate active measures for greening or infrastructure development in the studied regions. The growth is gradual and stable, which emphasizes the positive dynamics in the use of space for green spaces.

The table (Fig. 24) presents the maximum and minimum values for each indicator, along with the respective cities in which these values were recorded.

Decibel_Level has a maximum value of 90.0 in Dhaka city and a minimum value of 50.0 in Berlin, indicating a significant difference in noise level between these cities.

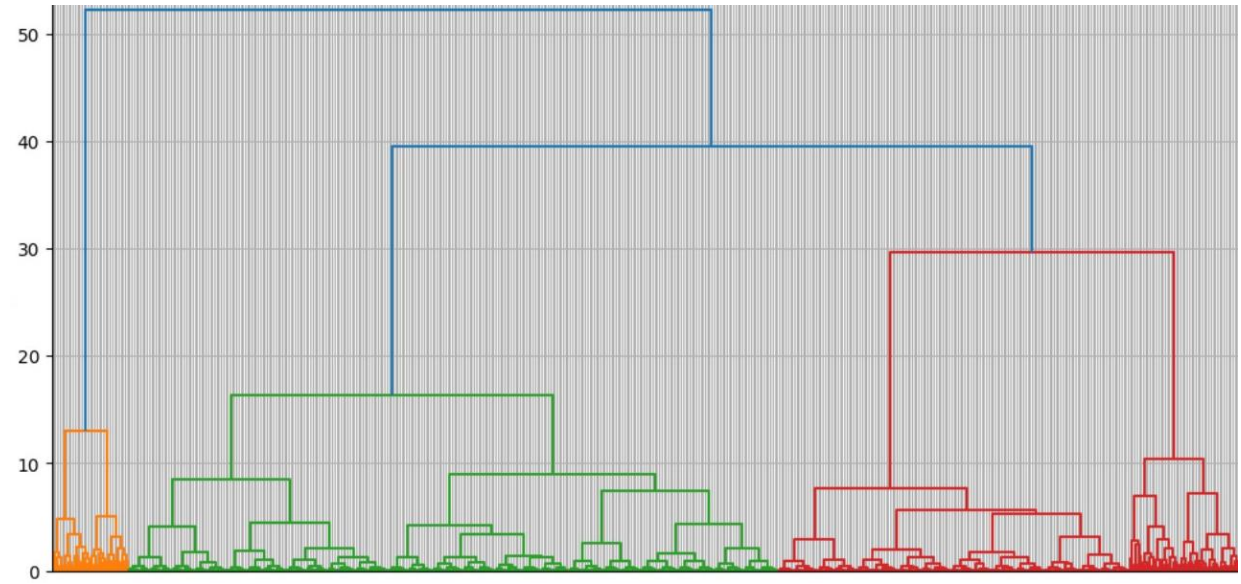


Fig. 20. Dendrogram of hierarchical cluster analysis

	0	1	2	3	4	5	6	\
0	0.000000	1.029237	1.926036	2.355885	1.631837	3.109877	1.737506	
1	1.029237	0.000000	0.912884	2.002279	0.992604	2.468242	2.087707	
2	1.926036	0.912884	0.000000	1.996626	1.096326	2.125563	2.658198	
3	2.355885	2.002279	1.996626	0.000000	1.133167	1.742613	1.630664	
4	1.631837	0.992604	1.096326	1.133167	0.000000	1.883591	1.691602	

	7	8	9	...	535	536	537	538	\
0	1.101434	2.996585	2.440548	...	6.654279	6.271731	6.109424	6.217870	
1	1.231284	2.900021	2.316555	...	6.263357	5.978448	5.837322	5.893817	
2	1.887515	3.226566	2.440257	...	6.129347	5.948749	5.855692	5.841857	
3	2.434067	4.781406	0.906631	...	6.288983	6.135126	6.234270	6.069837	
4	1.630625	3.702628	1.715413	...	6.095844	5.886284	5.838403	5.796155	

	539	540	541	542	543	544
0	6.679395	6.298433	6.136893	6.244922	6.704641	6.325259
1	6.290038	6.006457	5.866070	5.922354	6.316844	6.034585
2	6.156688	5.976978	5.884432	5.870729	6.184150	6.005325
3	6.315543	6.162410	6.261181	6.097536	6.342229	6.189816
4	6.123238	5.914712	5.867126	5.825152	6.150753	5.943256

[5 rows x 545 columns]

	Object	Cluster
0	1	5
1	2	5
2	3	5
3	4	5
4	5	5

Cluster	
1	35
2	119
3	179
4	161
5	51

Fig. 21. Proximity matrix table, clustering results table and cluster size table

Decibel_Level	Green_Space_Area	Air_Quality_Index	Happiness_Score	Cost_of_Living_Index	Healthcare_Index
2.9935139893082594	-1.4073374120734983	3.390469550144158	1.1613075945613394	1.0091507416417256	-3.4236280034244264
-0.14941571881577112	1.3440777365636516	-0.25403485681458965	-1.384240535492941	-0.4027913551535466	0.29876194708936504
-0.32506691751096584	0.4175313565108644	-0.3140195394928395	-0.3918779871688821	-0.36790965575791196	0.3400411200264741
-0.2734239931155432	-0.7239747124430025	-0.2926693215410207	0.8160830224932487	-0.36384876504205127	0.3253487135698894
0.298347490967707	-1.350325987071046	0.29201905967084807	1.23207158815353	2.6872034285931	-0.5681214696719676

Fig. 22. Cluster centre table

Green_Space_Area reaches a maximum of 2425.0 in Hokitika, while a minimum of 5.0 is recorded in Dhaka, indicating significant differences in the availability of green spaces.

Air_Quality_Index shows a maximum value of 245.0 in Jakarta and a minimum of 5.0 in Napier, indicating significant differences in air quality. Happiness_Score has a maximum of 8.6 in Napier and a minimum of -122.9 in Hokitika, which may indicate anomalies or outliers in the data.

Cost_of_Living_Index shows a maximum value of 130.0 in Hong Kong and a minimum of 20.0 in Kawerau, indicating significant economic differences. Healthcare_Index has a maximum of 99.0 in Oslo and a minimum of 35.0 in Phnom Penh, indicating substantial differences in the quality of healthcare. Month_Num shows a maximum of 12.0 in Matakana and a minimum of 1.0 in New York, corresponding to the number of months in a year. This table provides key information about the extreme values. It allows you to identify the regions with the highest and lowest indicators for each parameter, which can be helpful for further analysis.

The heatmap (Fig. 25) shows the correlation matrix for the leading indicators in the dataset. Correlation values range from -1 to 1, where 1 means a perfect positive correlation, -1 means a perfect negative correlation, and 0 means no correlation. Year and Green_Space_Area have a strong positive correlation (0.98), indicating that green space has increased over time. Decibel_Level and Air_Quality_Index have a strong positive correlation (0.86), which may indicate a relationship between noise levels and air quality. Happiness_Score and Green_Space_Area show a perfect negative correlation (-1), which may suggest that these factors have opposing effects. Healthcare_Index and Air_Quality_Index have a strong negative correlation (-0.94), which may indicate better healthcare in areas with better air quality. Cost_of_Living_Index and Happiness_Score have a moderate positive correlation (0.46), which may suggest that a higher cost of living is associated with greater satisfaction. Decibel_Level and Healthcare_Index have a negative correlation (-0.85), which may indicate that areas with higher noise levels have worse health outcomes. This heatmap allows us to quickly assess the relationships between variables and identify potential dependencies for further analysis.

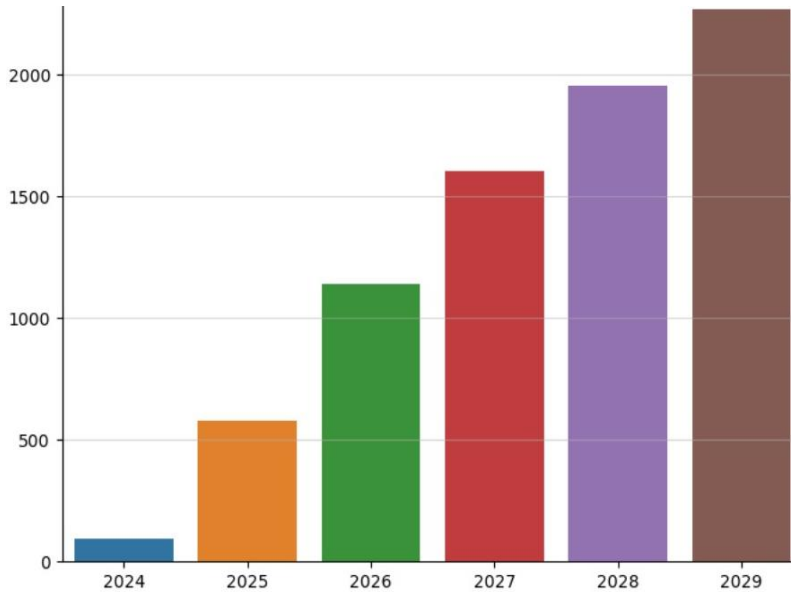


Fig. 23. The average value of green areas for each year

	Column	City with Max	Max Value	City with Min	Min Value
0	Decibel_Level	Dhaka	90.0	Berlin	50.0
1	Green_Space_Area	Hokitika	2425.0	Dhaka	5.0
2	Air_Quality_Index	Jakarta	245.0	Napier	5.0
3	Happiness_Score	Napier	8.6	Hokitika	-122.9
4	Cost_of_Living_Index	Hong Kong	130.0	Kawerau	20.0
5	Healthcare_Index	Oslo	99.0	Phnom Penh	35.0
6	Month_Num	Matakana	12.0	New York	1.0

Fig. 24. Maximum and minimum values for each indicator

The graph (Fig.26) presents the distribution of happiness levels for different levels of traffic density in the form of a boxplot. For low (Low) and medium (Medium) traffic densities, there is significant variability in happiness levels, with a median well below zero and an extensive range of values, indicating a high level of dispersion. High and Very High traffic densities exhibit a narrow distribution of values with minimal variability, indicating consistently low levels of happiness in heavy traffic conditions. The overall trend shows a decrease in variability as traffic density increases, which may indicate a consistently negative impact of high road congestion on happiness.

This heat map (Fig. 27) illustrates the geographical distribution of the Air Quality Index (Air_Quality_Index) in the 100 cities around the world that show the most significant variations in the indicators. The intensity of the colour on the map reflects the level of air pollution: blue shades indicate low or medium levels of pollution. In contrast, yellow and red shades signal high levels of pollution, which may indicate environmental problems in these regions. The map clearly shows a region in Oceania, in particular New Zealand, which shows high levels of pollution, as seen in the bright yellow and green areas. It indicates a localized problem that needs further investigation. Southeast Asia, particularly in cities such as Jakarta and Bangkok, has moderate levels of pollution, which is likely due to heavy traffic and industry. In Europe, including megacities such as London, Paris and Berlin, pollution levels are moderate, consistent with the typical levels for densely populated urban areas. North America, represented by the cities of Los Angeles, Toronto and New York, shows lower levels of pollution, which may be the result of effective environmental policies and emission controls. In Africa and South America, particularly Cape Town and Rio de Janeiro, pollution levels remain low, which is likely due to less industrialization in these regions.

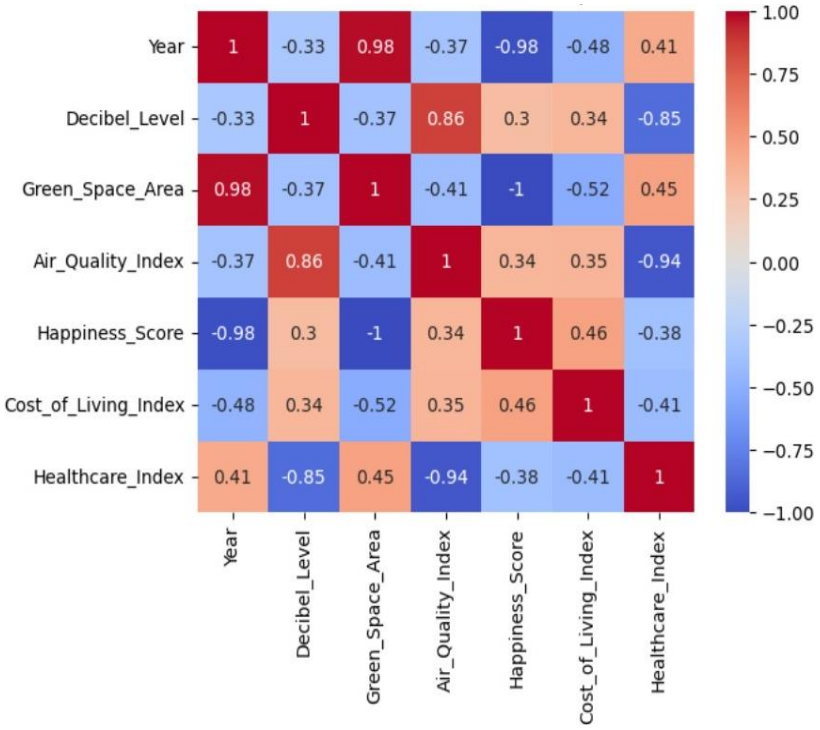


Fig. 25. Correlation matrix

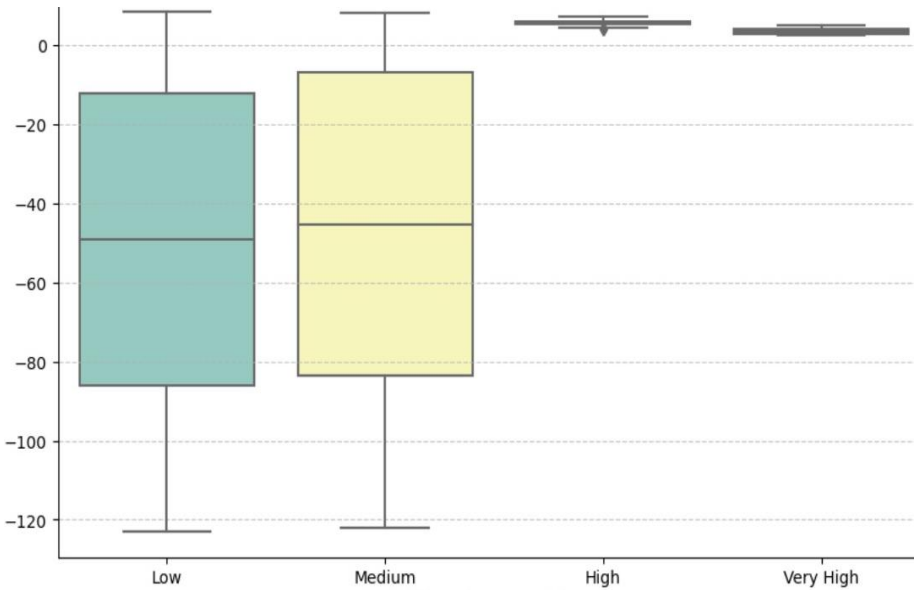


Fig. 26. Distribution of happiness levels for different levels of traffic density (X = happiness level, Y = traffic density)

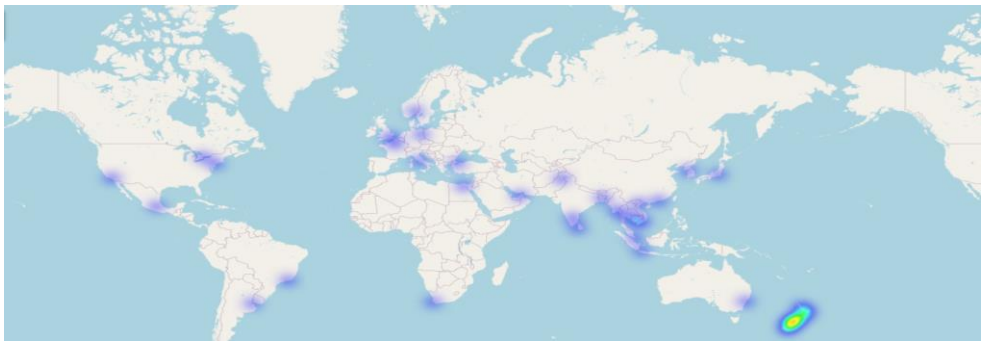


Fig. 27. Heat map

The map shows significant variability in pollution levels between continents. The highest levels of pollution are observed in Asia and Oceania, indicating the need for further analysis of the causes and factors influencing this level. Selected areas with increased pollution levels could become the subject of environmental studies and the development of air quality improvement programs. Additional analysis will help to study the relationship between pollution and indicators such as noise levels or cost of living. The results of this visualization can be used to inform the implementation of emission reduction measures in the worst-performing cities, as well as to promote environmentally friendly practices in areas with low pollution levels. For further research, it is advisable to build a time series to track the dynamics of pollution changes, analyze its impact on public health, and compare the correlations between air quality and socio-economic indicators of cities. These steps will allow for a deeper understanding of the environmental situation and contribute to the development of effective measures to improve it. The graph (Fig. 28) shows the results of data clustering using the DBSCAN algorithm based on two indicators: Decibel_Level (noise level) and Air_Quality_Index (air quality index). The colours in the graph indicate different clusters and dark purple dots (cluster -1) indicate anomalies or noisy data.

The algorithm identified several clear clusters. For example, yellow dots formed a group with high noise levels (75–80 dB) and moderate air quality (50–150), which may indicate regions with increased noise pollution. Green and blue clusters correspond to areas with lower noise levels (50–60 dB) and better air quality indicators, which may indicate cleaner areas. Visualization allows you to identify anomalies and patterns in the data, helping to assess the relationship between noise levels and air quality. It can be helpful for further analysis of environmental conditions and the development of measures to improve the quality of life in critical regions. The graph (Fig. 29) illustrates the detection of anomalies in the relationship between noise levels (Decibel_Level) and air quality (Air_Quality_Index). The points on the graph are divided into two categories: standard (marked in blue) and abnormal (marked in red). The analysis shows that most anomalies are observed at high noise levels (above 70 dB) and air quality values exceeding 150. It may indicate specific areas with intense noise and air pollution, such as industrial areas or transport hubs. At the same time, normal points prevail in the lower noise range (50–60 dB), indicating more stable environmental conditions with better air quality. The detected anomalies may signal problem regions that require additional research or the implementation of measures to improve ecological conditions. The graph demonstrates the relationship between noise and air pollution, which allows the identification of potentially risky areas for further monitoring and analysis.

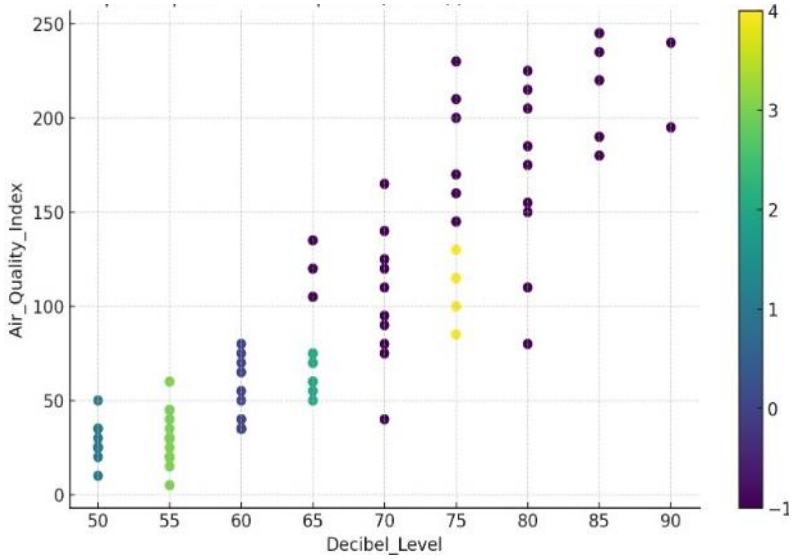


Fig. 28. Spatial clustering graph using DBSCAN

The research presented in the paper covers comprehensive data analysis, focusing on visualization, clustering and anomaly detection. It started with data pre-processing, including normalization and preparation for further statistical and spatial analysis methods. Particular attention was paid to the construction of graphs, heat maps and diagrams, which allowed us to identify the main trends and relationships in the data sets. The results of clustering, in particular hierarchical analysis, revealed groups of cities with similar characteristics, such as noise levels, air quality, cost of living and health indicators. The dendrogram showed clear boundaries of the clusters, which helped to identify regions with a high density of similar features. The proximity matrix table demonstrated the degree of similarity between objects, which became the basis for building clusters. The results of the analysis also indicated significant differences between groups, which is helpful for further management strategies and resource planning.

Spatial analysis included heat maps that visualized the geographic location of cities with high and low air pollution. It allowed us to identify critical areas with increased pollution levels in Asia, especially in the cities of Jakarta and Bangkok, as well as to highlight more environmentally friendly regions, for example, in South America and Africa. This approach helped form the basis for assessing environmental risks and planning measures to improve ecological conditions.

An essential part of the work was to identify anomalies in the data that illustrate the relationship between noise levels and air quality. Anomalies were observed in areas with high noise levels and significant air pollution, indicating possible problems in industrial and transport centres. Identification of such zones has practical significance for further monitoring and implementation of environmental measures.

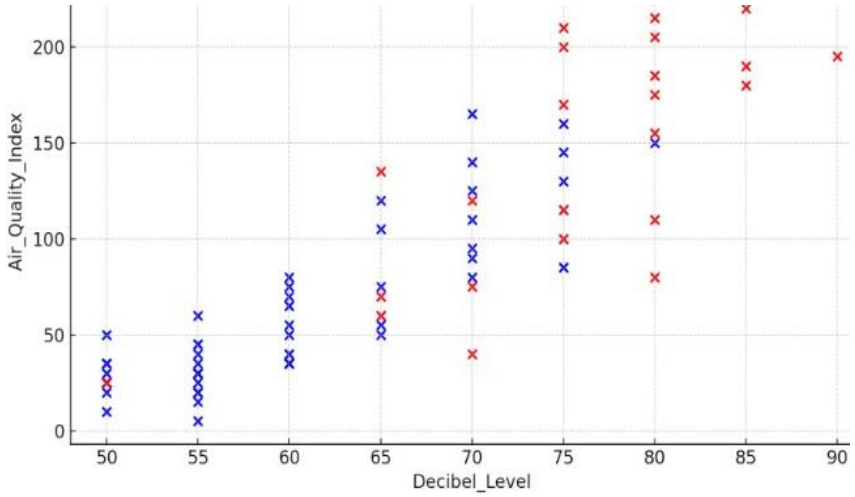


Fig. 29. Anomaly detection graph: noise level and air quality (blue – normal, red – anomalies)

Analysis of correlations between indicators allowed us to identify dependencies that can be useful for forecasting and management. For example, a positive relationship between noise levels and air pollution indicates a mutual influence of these factors. At the same time, the negative correlation between the health index and air quality emphasizes the importance of environmental conditions for the well-being of the population.

The results obtained demonstrate the practical application of statistical analysis, clustering, and anomaly detection methods to understand patterns in large data sets. They can be used as a basis for further research aimed at studying factors affecting the quality of life and the environment, as well as for developing measures to improve the environmental situation in individual regions.

Consider the scatter plot (Fig. 30a) using the `px.scatter` function from the Plotly Express library. The x-axis in the plot represents the attribute "Healthy life expectancy" from the data set, and the y-axis represents the "Ladder score". The points on the graph are shaded according to the column "Regional indicator" of the data set. The size of each dot is proportional to the Social Support characteristic, and the size_max argument sets the maximum size of the dots to 20. The graph also contains tooltips that display the name of each country when you hover over the corresponding dot.

Finally, the template argument sets the background of the graph to a simple white colour. The graph shows the relationship between the two variables and how they vary across regions, as well as the importance of social support to the overall well-being of a country. Using a logarithmic scale for the X-axis can help emphasize the relationship between the two variables for countries with low healthy life expectancy.

The correlation matrix analysis (Fig. 30b) showed a strong direct relationship between the happiness index and GDP per capita (0.79) and social support (0.76). A moderate relationship exists between the happiness index and freedom to choose life (0.61). Weaker relationships are observed between the other parameters.

(a) (b)

Fig. 30. (a) Relationship between life expectancy and happiness rating in different regions of the world (X – life expectancy and Y – happiness rating) and (b) Correlation matrix

There is a positive (Fig. 31) relationship between the Log GDP per capita and the Happiness Index. This means that countries with higher GDP per capita tend to have higher happiness indexes. Different colours of the dots in the diagram represent different regional indicators. Using the colours, you can track trends within various regions of the world. Some areas have higher levels of happiness and GDP, while others have lower ones. Overall, the diagram confirms the results of the correlation analysis, which indicates a strong positive relationship between the Happiness Index and Log GDP per capita. Next, we calculate the first (Q1), second (Q2) and third (Q3) quartiles of the Happiness Index for the data set data_(Fig. 32). Based on these values, each country is assigned a specific happiness category - 'Sad', 'Moderately Sad', 'Happy' and 'Very Happy' - in a new column 'label' in the DataFrame data.

The result of executing the code is a DataFrame data, which now contains a new column 'label' with the happiness category for each country. It helps to better understand and analyse the distribution of happiness among different countries based on the Happiness Index. Next, we will display a column chart (Fig. 32) that shows the distribution of countries by happiness categories ('Sad', 'Moderately Sad', 'Happy' and 'Very Happy'), sorted by the Happiness Index values. Each happiness category has a different colour according to the regional indicator. Using the hover_data parameter, the chart shows the name of the country and its Happiness Index when you hover the cursor over the column. The results of this chart help to analyse the distribution of countries by happiness levels according to regional indicators. It allows you to compare different regions and countries according to their happiness categories and to determine which areas have higher or lower levels of happiness overall. The chart (Fig. 33a) shows the relationship between the Happiness Index and the freedom to choose life for each country. The chart is divided into four columns according to the happiness categories: 'Sad', 'Moderately Sad', 'Happy' and 'Very Happy'.

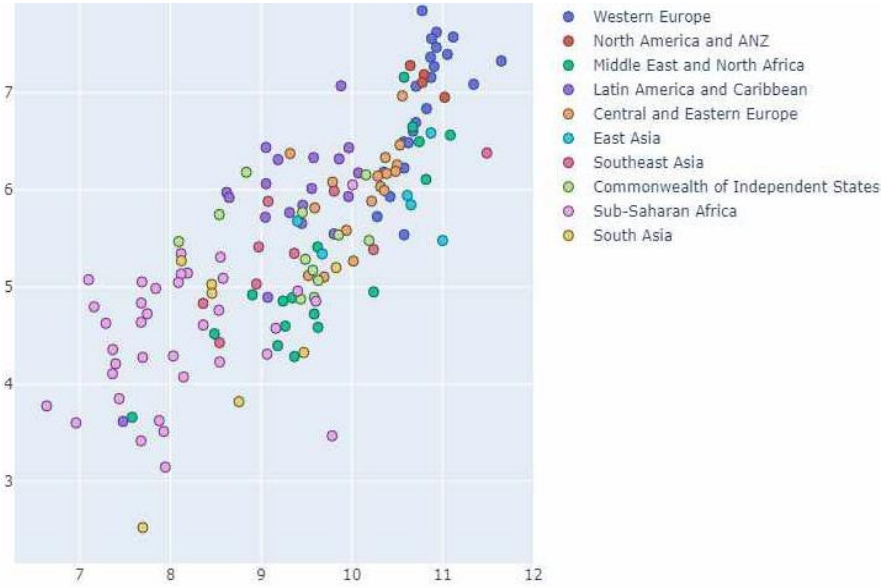


Fig. 31. Dependence of the happiness index on the world's GDP per capita (X is the logarithmic GDP per capita, and Y is the happiness index)

```
Q1 = np.percentile(data_, 25, interpolation = 'midpoint')
Q2 = np.mean(data_)
Q3 = np.percentile(data_, 75, interpolation = 'midpoint')
for i in range(len(data_)):
    if data['Ladder score'][i]<=Q1:
        data['label'][i] = 'Sad'
    if data['Ladder score'][i]>Q1 and data['Ladder score'][i]<Q2:
        data['label'][i] = 'Moderately Sad'
    if data['Ladder score'][i]>Q2 and data['Ladder score'][i]<Q3:
        data['label'][i] = 'Happy'
    if data['Ladder score'][i]>=Q3:
        data['label'][i] = 'Very Happy'
```

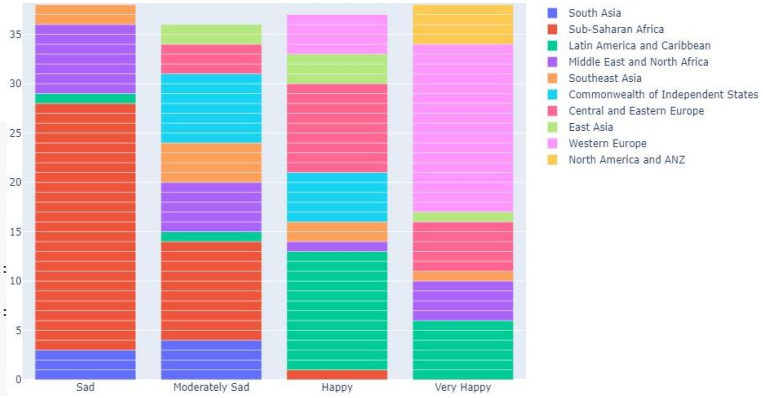
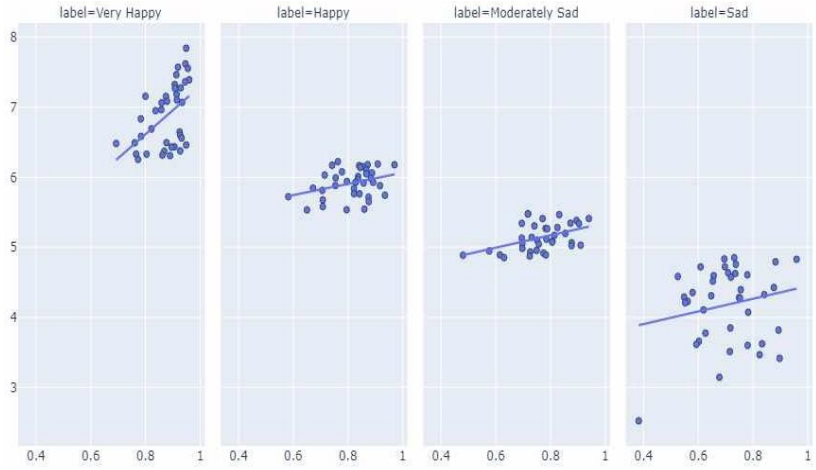


Fig. 32. Distribution of countries by happiness categories, taking into account regional indicators

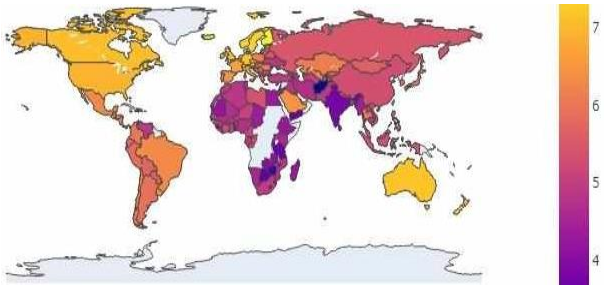
Each point on the chart represents a country, and its location reflects the value of the Happiness Index and Life Freedom. A trend line has also been added for each happiness category using the least squares (OLS) method. Results:

1. In the 'Sad' category, a specific positive correlation can be observed between Life Freedom and the Happiness Index, which is reflected in the trend line.
2. For the 'Moderately Sad' category, the correlation between Life Freedom and the Happiness Index is also positive but weaker compared to the 'Sad' category.
3. In the 'Happy' category, a positive correlation is observed between Life Freedom and the Happiness Index, although the scatter of the points is wider than in the previous categories.
4. Finally, for the 'Very Happy' category, the correlation between Life Freedom and the Happiness Index is weak and positive, and the points are located relatively close to each other.

Overall, this diagram confirms the existence of a positive relationship between freedom of life choices and the Happiness Index for different categories of happiness. However, the strength of the relationship may vary.



(a)



(b)

Fig. 33. Relationship between the Happiness Index and Life Freedom by Happiness Categories and the World Happiness Index

Finally, a choropleth diagram (Fig. 33b) is displayed, which shows the distribution of the Happiness Index around the world. Each country is coloured according to its Happiness Index value. The higher the index value, the lighter the colour of the country on the map. It allows us to visualize the level of happiness in different countries and facilitates comparisons between them. General observations from the map include:

- Northern European countries (Scandinavian countries) have high Happiness Index values and light colours on the map;
- Western Europe, North America and Australia also have relatively high levels of happiness;
- African countries, especially those located south of the Sahara, have low Happiness Index values and dark colours on the map;
- Central Asia and some countries in the Middle East also have low levels of happiness.

This diagram helps to get a general idea of the geographical distribution of happiness in the world and can encourage further research and analysis of the reasons for such distribution. Analysis of the results of the World Happiness Index shows a direct relationship between GDP per capita, social support and freedom to choose life. High correlation values indicate a close relationship between economic well-being and happiness. Visualizations help to observe the geographical distribution of happiness and compare the level of happiness in different countries.

4.3. Clustering countries by the World Happiness Index

The k-means model will try to find optimal cluster centres for this data, which can help reveal structure in the data and identify groups of countries with similar characteristics based on happiness index, economic indicators, social support, and other factors. With KElbowVisualizer, you estimate the optimal number of clusters for the k-means model using the Elbow Method. This method is based on visual analysis of a graph, where the X-axis shows the number of clusters (k), and the Y-axis shows the sum of the squares of the distances between each object and the centre of its cluster (inertia). The idea is to find a value of k at which the inertia stops decreasing significantly, and the graph takes the shape of an elbow. After running KElbowVisualizer (Fig. 34), the graph shows the dependence of inertia on the number of clusters, allowing you to determine the optimal value of k. Based on this analysis, you can choose the optimal number of clusters for your k-means model based on the determined "elbow" in the graph. In this case, the elbow plot shows the optimal value of k = 4, with the sum of the squares of the distances from the clusters to the cluster centres equal to 1420.893. It means that the data for the countries according to the World Happiness Index can best be divided into 4 clusters that minimize the intragroup variation and maximize the intergroup variation. With this result, you can better analyse the dependencies and differences between countries with different levels of happiness.

Next, the cluster labels will be correlated with other values in order to identify relationships between clusters and other data parameters using the Pearson and Spearman methods. Correlation analysis will help determine which characteristics have the most significant impact on the clustering of countries according to the World Happiness Index. The results obtained can be used to understand common trends and differences between groups of countries, as well as to identify potential policy directions aimed at increasing the level of happiness of the population. This diagram shows the correlation matrix between each pair of columns in the data. The correlation values are placed in cells using a colour scale: blue corresponds to a negative correlation, black to a zero correlation, and red to a positive correlation. The brighter the colours, the stronger the correlation between the corresponding columns.

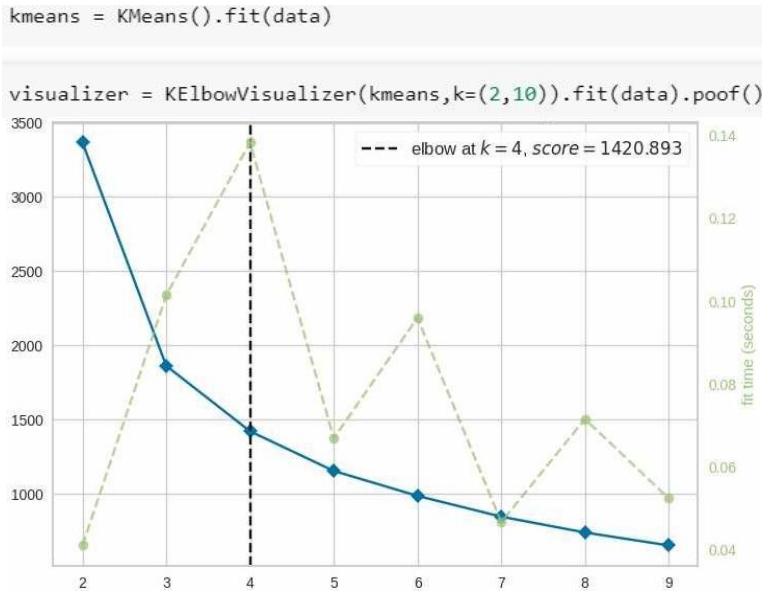
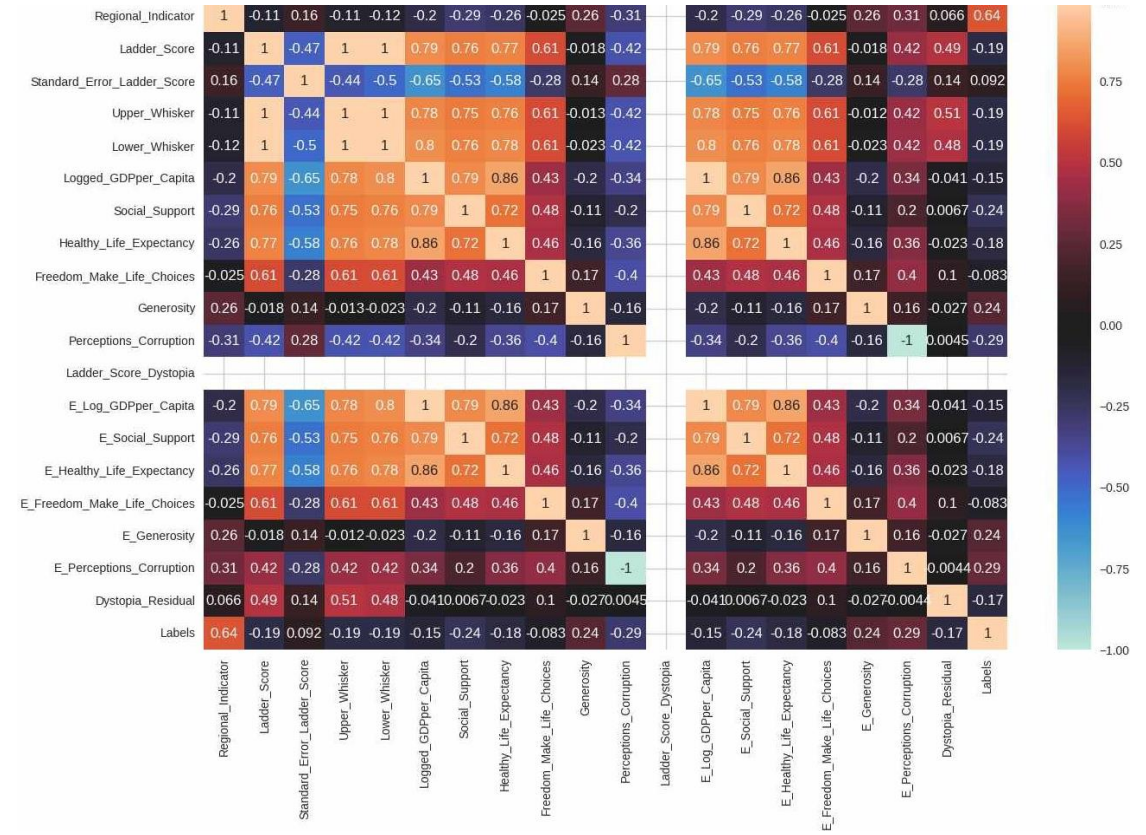
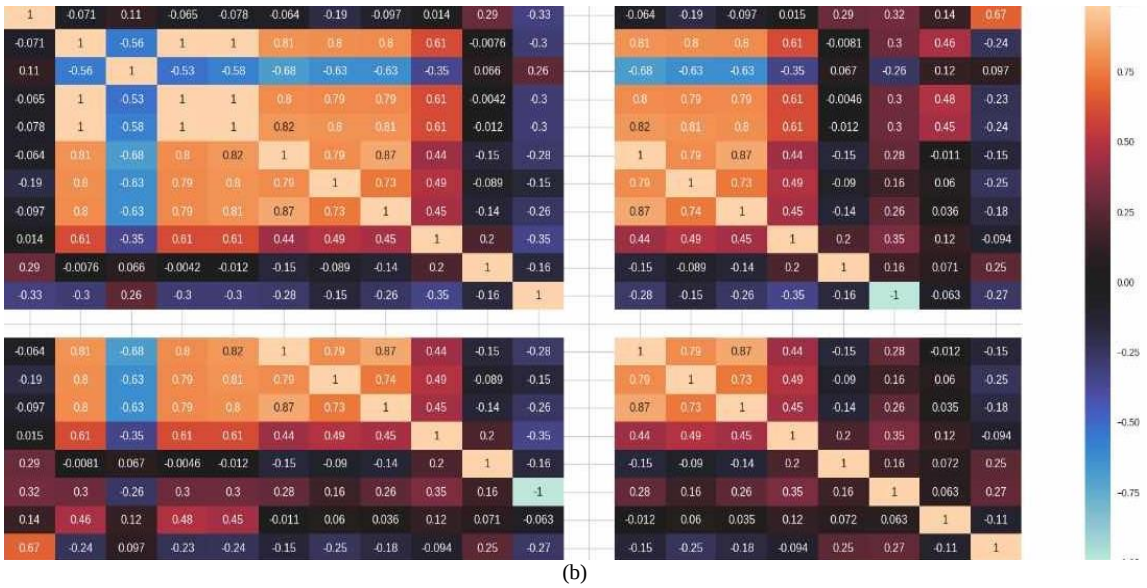


Fig. 34. Inertia vs. Cluster Count Diagram

The correlation matrix shows (Fig. 35a) that the Labels parameter is most correlated with the Generosity parameters (correlation coefficient 0.24), i.e. there is a positive relationship between these parameters, and Perceptions_Corruption (correlation coefficient -0.29), i.e. there is a negative relationship between these parameters. It is also noticeable that the correlation between Labels and the other parameters is slight to moderate, with correlation coefficient values from -0.29 to 0.24. It may indicate that the label parameter does not depend very much on the other parameters or that there are complex and multidimensional relationships between them. The parameter "Labels" represents the cluster category of countries, which was formed based on the distribution of their indicators regarding the level of happiness and well-being.



(a)



(b)

Fig. 35. (a) Pearson correlation matrix of Labels to other parameters and (b) Spearman correlation matrix of Labels to other parameters, where from top to bottom and from left to right Regional Indicator, Ladder Score, Standard Error Ladder Score, Upper Whisker, Lower Whisker, Logged GDPper Capita, Social Support, Healthy Life Expectancy, Freedom Make Life Choices, Generosity, Perceptions Corruption, Ladder Score Dystopia, E Log, GDPper Capita, E Social Support, E Healthy Life Expectancy, E Freedom Make Life Choices, E Generosity, E Perceptions Corruption, Dystopia Residual, Labels

According to the correlation matrix (Fig. 35b), it has the highest correlation with the parameter "Regional_Indicator" (0.67), which may indicate a significant influence of the region on the formation of these clusters. It also has a negative correlation with the indicators Ladder_Score (-0.24), Social_Support (-0.25) and Perceptions_Corruption (-0.27), which may indicate that these indicators are essential for determining clusters. However, it is worth noting that the correlation analysis does not allow us to conclude the cause-and-effect relationships between the parameters but only shows the degree of their relationship.

Next, a covariance was performed with respect to the Labels parameter to investigate the dependence between the parameters and to find out how they affect the clustering of countries by their level of happiness. Covariance allows us to measure the degree of dependence between two variables. If the value of the covariance between the parameters is high, then a change in one parameter is associated with a change in the other parameter. That is, they are interdependent. For this reason, examining the covariance between parameters can help to find the most critical parameters that affect the clustering of countries by their level of happiness. The covariance matrix (Fig. 36) lists the covariances of each variable with respect to the Labels variable. The following conclusions can be drawn from the analysis:

- The Ladder_Score variable has the most significant impact on the Labels variable since the covariance between them is the largest compared to the other variables;
- The Healthy_Life_Expectancy variable also has a significant impact on the Labels variable since the covariance between them is the largest among the other variables except for the Ladder_Score;
- The Logged_GDPper_Capita, Social_Support, Freedom_Make_Life_Choices, and Dystopia_Residual variables have a weak impact on the Labels variable since their covariances are almost zero;
- The variables Generosity and Perceptions_Corruption have negative covariance with the variable Labels, meaning that their values decrease as the value of Labels increases. However, the covariances of these variables are very weak.

Next, we will create a scatter plot graph (Fig. 37a) of the dependence between the indicators "Social_Support" and "Freedom_Make_Life_Choices" using the Seaborn library. Each point on the graph corresponds to one country, and the colour of the point reflects its level of happiness, which is indicated in the "Labels" column. The graph helps to visualize the dependence between the two indicators and their relationship to the level of happiness of countries.

The scatter plot (Fig. 37b) shows the relationship between Social_Support (social support) and Healthy_Life_Expectancy (life expectancy and quality of life) for each country depending on its assigned label. Each cluster has its color. Cluster 0 is characterized by high levels of Social_Support (from 0.7 to 0.95) and high levels of Healthy_Life_Expectancy (from 63 to 72). Cluster 1 has lower levels of Social_Support (from 0.4 to 0.8) and lower levels of Healthy_Life_Expectancy (from 53 to 56). Cluster 2 has a high level of Social_Support (from 0.8 to 0.98) and a high level of Healthy_Life_Expectancy (from 70 to 76). Cluster 3 has a lower level of Social_Support (from 0.4 to 0.86) and a medium level of Healthy_Life_Expectancy (from 55 to 63). With the help of this graph, you can see which countries are included in each cluster and compare their levels of Social_Support and Healthy_Life_Expectancy.

Fig. 36. Covariance matrix Labels to other parameters, where from top to bottom and from left to right Regional Indicator, Ladder Score, Standard Error Ladder Score, Upper Whisker, Lower Whisker, Logged GDPper Capita, Social Support, Healthy Life Expectancy, Freedom Make Life Choices, Generosity, Perceptions Corruption, Ladder Score Dystopia, E Log, GDPper Capita, E Social Support, E Healthy Life Expectancy, E Freedom Make Life Choices, E Generosity, E Perceptions Corruption, Dystopia Residual, Labels

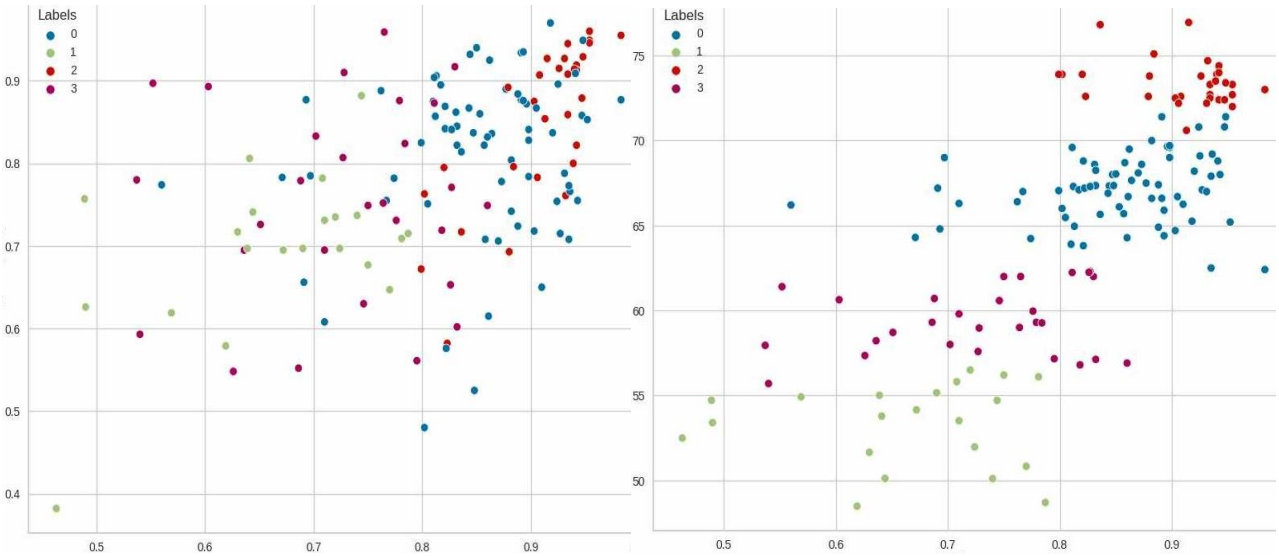


Fig. 37. (a) Relationship between social support and freedom of choice in different regions of the world (X – Social Support and Y- Freedom Make Life Choices) and (b) - plot of the relationship between Social_Support (X) and Healthy_Life_Expectancy (Y)

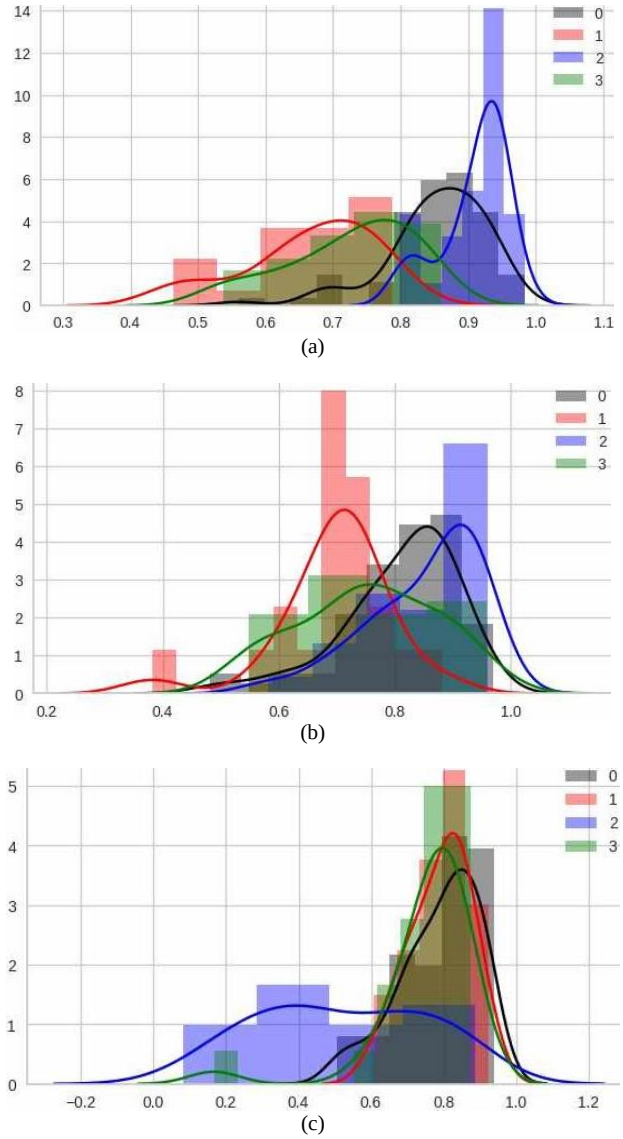


Fig. 38. Density plots of different data groups for the corresponding indicators as (a) Social Support, (b) Freedom to Make Life Choices and (c) Perceptions of Corruption

Let us present a histogram (Fig. 38) with density graphs of different data groups for the indicator "Social_Support". Each group has its separate density graph, marked with its colour and label. The name of the graph title is "Label". The graph allows you to compare the distribution of this indicator for different clusters and understand whether there are any differences in the distribution of values between the groups.

During clustering, the KMeans method was used with the number of clusters equal to 4. Various graphs were created using the Seaborn library to visualize the results. It was found that the clusters have different characteristics depending on the region in which the country is located. In addition, it was found that the clustering was carried out on the basis of such features as Social_Support, Healthy_Life_Expectancy, Freedom_Make_Life_Choices and Perceptions_Corruption. In particular, Social_Support and Healthy_Life_Expectancy were decisive in separating the clusters. The clustering results can be used for further analysis and comparison of countries from different regions depending on their attractiveness for living and standard of living.

4.4. Predicting the value of the happiness index

Now, we will use linear regression to predict the value of the happiness index based on the selected features. We will use this method to build a model for predicting the happiness index value based on selected features and evaluate the quality of the model using the R-squared and mean-square error metrics. In this code, the variables var_array and target_array contain the features and the target indicator, respectively, for further work with them. The to_numpy() method converts the data to NumPy arrays to make it more convenient to process them in further calculations. The result shows that the features have a size (1479, 5), which means that we have 1479 rows with five features each, and the target indicator has a size (1479, 1). That is, it has 1479 values in the vector.

```
#convert to be array
var_array = final_df[df_var_final].to_numpy()
target_array = final_df[df_target_final].to_numpy()
print('shape of final feature:',var_array.shape)
print('shape of target:',target_array.shape)
```

```
shape of final feature: (1479, 5)
shape of target: (1479, 1)
```

```
# split the data into test and train
X_train, X_test, y_train, y_test = model_selection.train_test_split(
    var_array,
    target_array,
    train_size=0.80,
    random_state=0
    )
```

Fig. 39. Program code

This code splits the data into training and test sets using the `train_test_split()` function from the `model_selection` module of the Scikit-learn library. The `var_array` and `target_array` variables contain the features and the target variable, respectively. The function parameters specify that the training set will consist of 80% of the randomly selected data, and the remaining 20% will be used for testing. The `random_state` parameter sets a random starting state for the pseudo-random number generator, which ensures repeatability of the results. The resulting training and test data sets will be stored in the variables `X_train`, `X_test`, `y_train`, and `y_test`.

Next, we derive the data dimensions for the training and test datasets and their corresponding labels. The data dimension is essential to ensure that the correct data has been loaded into the model. In this case, the code shows the dimensionality of the variables for the training and test datasets, which were split using the train_test_split() function of the scikit-learn library. The output of this code indicates the number of data samples (rows) and the number of variables (columns) for each dataset. These results show that the data was split into training and test sets in a ratio of 80:20, respectively. The size of the training set is 1183 observations, and the size of the test set is 296 observations. Separate arrays for the features and their corresponding labels were created for each of the sets. Next, we create a linear regression object and train it on the training data to find the coefficients of a linear function that finds the relationship between the independent variables and the dependent variable. In this case, the independent variables are the features, passed as `X_train`, and the dependent variable is the happiness index, passed as `y_train`. The `lm` object contains trained coefficients that can be used to predict the happiness index for new data.

```
# check the shape data
print('Shape Data X Train:')
print(X_train.shape)
print('\nShape Data X Test:')
print(X_test.shape)
print('\nShape Data y Train:')
print(y_train.shape)
print('\nShape Data y Test:')
print(y_test.shape)
```

```
Shape Data X Train:
(1183, 5)

Shape Data X Test:
(296, 5)

Shape Data y Train:
(1183, 1)

Shape Data y Test:
(296, 1)
```

```
#making model
lm = linear_model.LinearRegression()
```

```
#fit model
lm.fit(X_train, y_train)
```

```
LinearRegression
LinearRegression()
```

Fig. 40. Program code

This code contains the results of a linear regression simulation.

```
# model result
print('Coefficients:\n Social support, Healthy life, Freedom, Generosity, Perceptions of corruption \n',lm.coef_)
print('Intercept:',lm.intercept_)
```

```
Coefficients:
 Social support, Healthy life, Freedom, Generosity, Perceptions of corruption
 [[ 3.32440659  0.0655999   0.92166755  0.55388933 -0.93104098]]
Intercept: [-1.37392068]
```

Fig. 41. Program code

The first five values are output for the regression coefficients, which indicate how much the variable affects the dependent variable. So, in this case, we see that the coefficient for Social support is larger than all the others, so this variable has the most significant impact on the happiness index. The other coefficients are also positive, which means that higher values of these variables are associated with higher values of the happiness index. The last line outputs the value of a constant that indicates the baseline level of the happiness index when all variables are equal to zero. The linear regression model looks like this:

$$I_{Ladder\ score} = -1.3739 + 3.3244 \times I_{Social\ support} + 0.0656 \times I_{Healthy\ life} + 0.9217 \times I_{Freedom} + + 0.5539 \times I_{Generosity} - 0.9310 \times I_{Perceptions\ of\ corruption} \quad (18)$$

It means that with an increase of one unit in the values of the variables "Social support", "Healthy life", "Freedom", and "Generosity", the expected value of the happiness index increases by 3.3244, 0.0656, 0.9217 and 0.5539 respectively. With an increase of one unit in the value of "Perceptions of corruption", the expected value of the happiness index decreases by 0.9310. The value of the constant -1.3739 indicates the expected value of the happiness index when the values of all variables are zero.

```
#predic data
y_train_pred = lm.predict(X_train)
y_test_pred = lm.predict(X_test)
target_array_pred = lm.predict(var_array)
```

```
# check the prediction data & real data
print('Real Data')
print(y_test[:10])
print('\n Predicted Data')
print(y_test_pred[:10])
print('\n Diff')
print(y_test[:10]-y_test_pred[:10])
```

Fig. 42. Program code

This code performs prediction for a linear regression model built on the training data and applies it to predict the happiness index value for the test data and all data. Then, the first 10 values of the actual data, the predicted data, and the difference between them are output. The results demonstrate that the model predicts the happiness index value quite well, as the difference between the actual and predicted data is relatively small for the first 10 values. However, larger deviations may be noticeable for other data. The results in the diagram (Fig.43) indicate that linear regression can be helpful in predicting the happiness index values based on five input features. However, the deviations between the actual and predicted happiness index values turned out to be quite significant, with some deviation values greater than 1, which may indicate the limited effectiveness of such a model in predicting accurate happiness index values. It is worth considering using other machine learning methods to obtain more accurate predictions.

This command will build a histogram of the residuals. The residuals represent the difference between the actual value of the output variable (in the test set) and the predicted value. The histogram (Fig. 44) can show how well the model is able to predict the values of the output variable. If the histogram of the residuals has a normal distribution, this means that the model is adequate. If the histogram has some other shape, it may be a reason to improve the model.

The regression model estimation using the coefficient of determination (R^2) showed a value of 0.67, which means that a linear dependence on the input variables can explain 67% of the variance of the target variable. The RMSE value for the training sample is 0.58, and for the test sample - 0.60. This means that the average prediction error for the test sample is about 0.60. Based on the trained linear regression model, predictions of the happiness index for 2021 were made (before the start of the war in Europe on the territory of Ukraine, which significantly negatively affects the indicator in many European Union countries) based on the latest available data on social support, healthy life expectancy, freedom, generosity and perception of corruption in countries around the world. For this, the coefficients obtained as a result of training the model were used. In our case, the RMSE is 0.5478, which means that our model has an average error of 0.5478 points of the happiness index (Fig. 44). This is a reasonably small error. Still, you can try to improve the model by adding more data or using other algorithms to improve the accuracy. The value of R^2 = 0.7380 means that our model explains about 73.8% of the variation in the happiness index. It indicates that the model has a moderate ability to predict the happiness index, but there is potential for improvement. The predicted happiness index

for Ukraine is 5.5044. It is an absolute value that reflects the level of happiness in Ukraine according to the model's predictions. To better understand the result, you can compare it with the happiness indices of other countries or with previous values of the happiness index for Ukraine. So, a linear regression model was used to predict the happiness index in Ukraine based on several parameters, such as social support, healthy life expectancy, freedom to choose a life path, generosity, and perception of corruption. The model was trained on data from previous years and modelled for 2021 and showed a moderate ability to predict the happiness index, with an R^2 of about 0.7380 and an RMSE of 0.5478. Also, the expected happiness index for Ukraine is determined to be 5.5044 in 2021 (actual 5.084). This means that, based on our model and the parameters used, Ukraine's socio-economic development can be assessed as moderate compared to other countries. However, it is worth remembering that the accuracy of the prediction can be improved with more data, the use of different algorithms, or optimization of the model's hyperparameters. Also, due to the limitations of the model and historical data, the actual happiness index may differ from the predicted one due to unforeseen events or changes in the country's conditions.

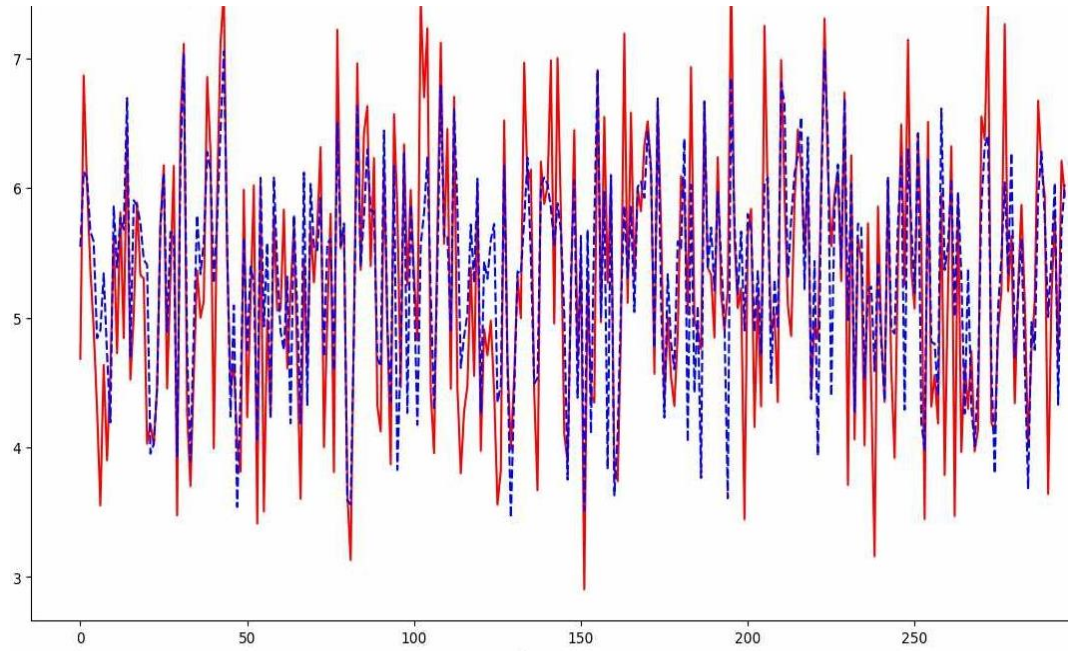


Fig. 43. Comparison of actual and predicted values (X – Happiness Index and Y - Observation), where red is the exact value, and purple is the expected value.

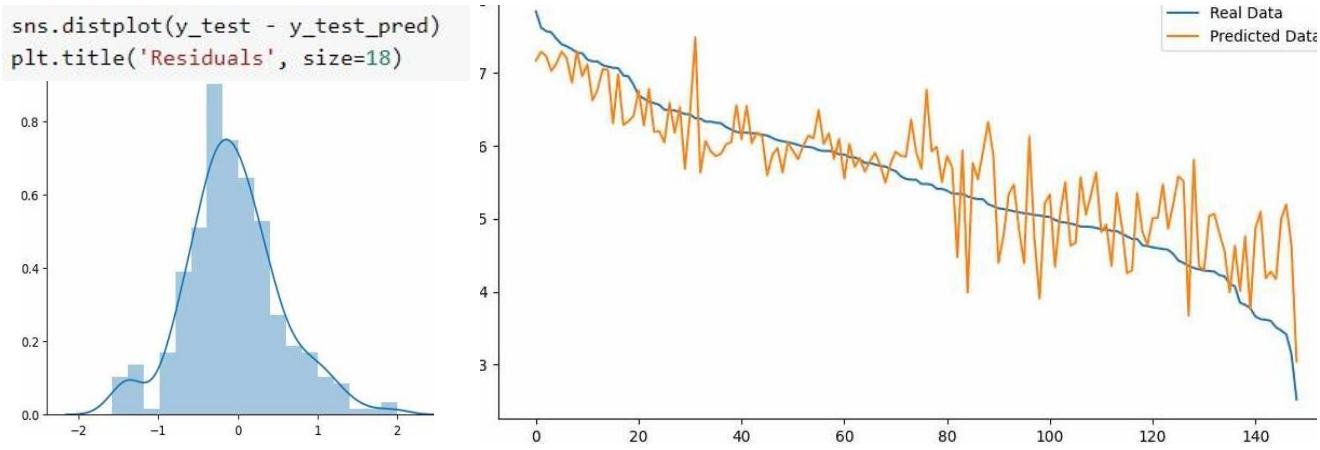


Fig. 44. Histogram of residuals and comparison of actual and predicted values

We will describe a universal approach to assessing happiness based on a global index (Gallup World Poll / World Happiness Report), taking into account cultural, regional or individual characteristics. A one-size-fits-all approach ignores cultural specificities and social contexts, which can lead to distorted or non-representative outcomes. It is necessary to apply a multilevel model that takes into account: personal characteristics (age, gender, education), cultural context (values, traditions) and macroeconomic indicators. The next step is to group countries by region/civilisation zones and apply appropriate weights or adjustments. Next, it is necessary to analyse the sensitivity of the results to weight changes to see how the happiness index changes when the cultural context changes. It can significantly affect the

accuracy of the interpretation of the results. Here's how this can be remedied by developing a more sensitive, culturally adaptive approach and conducting an analysis.

An approach to measuring happiness, taking into account cultural, regional and personal differences:

Step 1. An adaptive multilevel analysis model (Multilevel Happiness Index) that normalises or calibrates the happiness score according to the context of each level before integrating it into the overall indicator:

Level 1. Individual (personal differences), such as gender, age, education, employment, and psycho-emotional state. For example, in young people, happiness is more often associated with freedom, while in older people, it is associated with health and stability.

Level 2. Cultural / Regional context, such as Collectivist vs individualist cultures, as well as social religiosity, political system, and values. For example, in East Asian countries, people may underestimate happiness for reasons of modesty (the so-called "social desire").

Level 3. Global (unified indicator), such as GDP, health, freedom of choice, social support, trust, etc.

Step 2. Calculation of adjusted indicators from the World Happiness Report set. Let's analyse the variables from the existing set of Logged GDP per Capita, Social Support, Healthy Life Expectancy, Freedom to make life choices, Generosity, and Perceptions of Corruption. Let's group the data by regional blocks (for example, according to the UN classification) — Europe, Asia, Africa, and the America.

Table 4. An example of calculating adjusted indicators from the World Happiness Report set

Region	GDP	Support	LifeExp	Freedom	Generosity	Corruption	HappyIndex (averaged)
Western Europe	10.8	0.95	75.1	0.87	0.22	0.13	7.45
Eastern Asia	9.2	0.84	72.5	0.56	0.15	0.12	5.80
Africa	7.4	0.56	58.2	0.47	0.28	0.28	4.13
South. Americas	9.0	0.91	69.0	0.72	0.35	0.22	6.25

Step 3. Analysis of results taking into account regional contexts. European countries have high levels of happiness, where almost all factors have high values. East Asia shows lower levels of happiness despite relatively high economic performance, which may be due to cultural restraints on emotions and understatement of responses. African countries show a low happiness score due to the low value of almost all factors. Still, at the same time, generosity is higher than in Europe, which indicates value differences. South America is an example of a region with a strong emotional response, a high level of social support, and generosity that compensates for economic constraints.

Step 4. Cultural adaptation through weight factors. For each region/culture, different weights can be applied to the factors to better reflect the local understanding of happiness. For example, in Asia, there is a higher weight for social harmony and stability, in Africa, for religion and community, in the United States, for freedom of choice, and in Europe, for social support and health.

The developed intelligent system for analysing the level of happiness and life satisfaction based on socioeconomic data has a vast potential for practical application. Below is a detailed description of where, how and for whom such a system can be implemented. The developed system can be implemented in state institutions and public authorities, research institutions and universities, international organisations and charitable foundations, think tanks and consulting companies, etc.

Table 5. Implementation of developed systems

Environment	Appointment	Purpose of application
State institutions and public authorities	– Ministries of Social Policy, Economy, Healthcare, Education. – Local authorities (municipalities) will analyse regional welfare.	– Monitoring the level of happiness/life satisfaction of the population. – Identification of risk groups. – Formation of social programs adapted to specific regions.
Research institutions and universities	– Sociological, economic, and demographic research. – Educational programs (master's/PhD courses).	– Conducting research on machine learning in the humanities. – Development of your models based on local or regional data. – Teaching students interdisciplinary data analysis.
International organisations and charitable foundations	– UN, WHO, UNICEF, World Bank, OECD. – Non-Governmental Organisations engaged in the assessment of the quality of life, human rights, and well-being. For example, the protection of human rights (for example, Amnesty International), ecology (Greenpeace), education, culture, health, assistance to refugees, socially vulnerable groups, anti-corruption initiatives and public control over the authorities.	– Comparative analysis between countries/regions. – Identification of priority areas for assistance. – Creating reports using analytical models.
Think tanks, consulting companies	Any think tanks, consulting companies	– Analytics for governments, cities, corporations. – Forecasting the effectiveness of social reforms.

The technical implementation is based on the implementation of an interactive web platform, an exclusive analytical system, and/or a built-in module for the Smart City digital platform. An interactive web platform is implemented as a web application (for example, based on Python + Streamlit or Flask). The user uploads CSV files with data. The system automatically clusters and visualises the results, and then makes cluster forecasting for new objects. The desktop analytical system is designed for sociologists/analysts working with local data. It can integrate with Excel, as well as the ability to calculate clusters "on the spot". The built-in module in the Smart City digital platform enables automated collection of social data from regions, the definition of a "happiness index" at the city/district level, and supports visual dashboards for local authorities.

Organisational implementation:

– Conducting workshops and training for government officials and analysts.
– Software licensing or open access to the prototype.
– Pilot implementation in one region, with further scaling.
– Collect feedback and adapt the model to local needs (for example, taking into account the cultural context).

Table 6. Target users

User Category	Need	How the system helps
Analysts/Sociologists	Identification of social patterns	Clustering, Visualisation, Comparison
Government officials/managers	Data-driven policymaking	Identification of risk groups, recommendations
Scientists/Teachers	Well-being modelling, research preparation	Data flexibility, forecasting
International organisations	Country comparison, progress monitoring	Objective metrics and groupings
Business / Consulting	Analysis of the social environment	Forecasting trends in consumers/regions

The system has a high potential for real implementation by authorities, scientific institutions, and international organisations. Its advantage is adaptability, scalability, ease of integration, and support for machine learning. For a successful application, it is essential to develop a user-friendly interface, provide high-quality documentation, and evaluate the results in the context of local conditions. Below is a detailed description of the potential implementation of the system and the criteria for measuring its effectiveness.

The system is implemented as a prototype in Python (pandas, sklearn, matplotlib) and deployed in a local environment (for example, Jupyter Notebook or Google Colab). It is designed for semi-automatic analysis: the user downloads data, selects clustering parameters, runs classification, and receives graphs.

Table 7. Technical implementation options

Platform Type	Example of implementation	Appointment
Jupyter Notebook	System prototype for researchers	Manual code run, experiments
Web service	With Flask / Streamlit	Interactive analysis without coding
Desktop application	Tkinter, PyQt, or Electron (via the web)	Standalone version of the system

Because the system performs clustering, classification, and regression analysis, performance can be measured through a combination of metrics of model quality, performance, and user satisfaction:

A. Mathematical/algorithmic metrics like Clustering (K-means), Regression (Linear Regression), and classification (kNN). For clustering (K-means), Silhouette Score (degree of cluster separation, optimally > 0.5), Davies–Bouldin Index (lower value = better segmentation) and stability of clusters when restarted with different initialisations were used. For regression (Linear Regression), R^2 (coefficient of determination as a fraction of explained variance) and MAE/RMSE (mean/standard error of predictions) were used. For classification (kNN), accuracy (proportion of correct classifications) and Precision / Recall / F1-score (in case of cluster imbalance) were used.

B. Technical metrics such as the time to perform clustering/forecasting on a certain amount of data, the ability to process new records without re-learning, and the flexibility of data loading (CSV/Excel formats).

C. User metrics (when the system is public) as an assessment of the usability of the interface (according to surveys), training time for staff/researchers to use the system, and feedback on usefulness in analytics/decision-making.

The system is implemented as a prototype in Python with support for clustering, regression, and visualisation. Performance is measured using classic machine learning metrics (R^2, Silhouette Score, classification accuracy), as well as technical and custom metrics.

We will describe the prospects for further research in the direction of intellectual analysis of happiness and life satisfaction based on socioeconomic data, which is an essential final element of scientific work. Below is a detailed answer, which covers theoretical, methodological, technical and applied areas of development and prospects for further research.

1. Deepening of the cultural and psychological context through the integration of psychological and cultural variables (e.g. collectivism index, level of religiosity, gender equality index) into the analysis. It is necessary to develop regionally oriented models of happiness adapted to the specifics of continents or countries. It is also necessary to investigate the influence of personal characteristics (age, gender, life events) on individual levels of happiness — the transition from the macro level (country) to the micro level (person).

2. Expansion of data and sources of information based on the addition of time series to analyse changes in the happiness index over the years. It is necessary to integrate with alternative data sources, in particular, social networks (sentiment analysis by texts), open government data, and satellite or geolocation sources (for example, the level of lighting as an indicator of development). At the same time, it is necessary to support the use of raw microdata in surveys, not only aggregated national values.

3. Development of artificial intelligence models through the replacement of basic algorithms (K-means, linear regression) with more powerful models, in particular, neural networks (MLP, RNN) for predicting happiness, gradient boosting (XGBoost, LightGBM), and deep clustering or Bayesian models for more sensitive grouping. It is necessary to investigate nonlinear relationships and interactions between variables (for example, the combined effect of income and freedom). The use of interpretable AI solutions, such as SHAP analysis, is supported to explain the influence of factors on happiness.

4. Interactive and scalable software solutions based on the development of a full-fledged information and analytical platform that will allow dynamically updating data, comparing regions and forming policy recommendations. It is necessary to implement interfaces for ordinary users, sociologists, and managers with priority settings (for example, "importance of freedom" and "priority of family").

5. Political and practical application through the development of the Social Policy Efficiency Index – assessment of the impact of state decisions on the level of happiness. It is necessary to create predictive models to assess the effect of a change in a social indicator (for example, what will happen to happiness with a decrease in corruption by 20%). Next, it was necessary to apply the result of the development in Smart City – an assessment of well-being in cities based on local data.

6. Interdisciplinary research is based on a combination of approaches from sociology, psychology, economics, computer science, and philosophy. It is necessary to analyse the topic of "digital happiness" – the impact of digitalisation, technology, and social networks on the level of life satisfaction.

The model of happiness index analysis proposed in the study is a promising basis for further research. The following steps should be aimed at:

– deepening of the context (cultural, personal);
– technological improvement of models and implementations;
– expansion of application in governmental, educational, and social policy;
– transition to individualised and dynamic approaches to measuring happiness.

5 Conclusions

In the modern world, creating a favourable environment for human life and development is one of the most critical problems, especially in the context of growing urbanization. The happiness index, which is considered an important indicator, helps to determine the level of comfort and well-being of the population in different cities and countries. The development of a module for determining a favourable environment for human life and development based on the happiness index and machine learning methods can help identify patterns and trends that affect the happiness index, as well as raise awareness among the public and stakeholders about the importance of a favourable environment for human life and development in cities. The application of machine learning and data analysis methods in this area can open up new opportunities for information support for decision-making at all levels of city management and government agencies.

Research in the field of determining a favourable environment for human life and development is fundamental and relevant, as it can contribute to improving the quality of life and creating a favourable environment for human development in cities. To achieve these goals, machine learning methods are used, in particular clustering algorithms such as k-means and multiple regression models. The application of these methods allows us to obtain valuable conclusions about the factors that contribute to the creation of a favourable environment for human life and development and to develop effective strategies in urban planning and social policy. In general, research in the field of determining a favourable environment for human life and development is an essential tool for improving the lives of the population and the development of cities.

This paper examines methods for studying and analysing happiness index data. Clustering algorithms, such as k-means that allow grouping countries by happiness level and regression models that can be used to predict the happiness index, were investigated. Pseudocode of the algorithm for clustering happiness index data and the algorithm for predicting the happiness index in Ukraine using linear regression were described. Correlation and covariance analysis between different indicators were also conducted, which allowed for finding dependencies between them and understanding their impact on the happiness level. As a result of these studies, it can be concluded that machine learning and data analysis methods are helpful for studying and analysing the happiness index, which helps to better understand the factors that affect the quality of life and happiness level of the population in different countries.

Based on the developed algorithms, a software implementation of the module was carried out to determine a favourable environment for life and human development based on the happiness index. Analysis of the world happiness index showed a close relationship between GDP per capita, social support, freedom of life choice and the level of

happiness in countries. The correlation values confirm this relationship. In particular, the correlation between GDP and happiness was high (0.79). The visualizations helped to compare happiness levels in different countries and see the geographical distribution of happiness.

During clustering, 4 clusters were identified depending on the features Social_Support, Healthy_Life_Expectancy, Freedom_Make_Life_Choices and Perceptions_Corruption. Social_Support and Healthy_Life_Expectancy were decisive for separating the clusters. The clustering results can be used to compare and analyze countries from different regions.

A linear regression model was used to predict the happiness index in Ukraine using data from previous years and the parameters of social support, healthy life expectancy, freedom to choose a life path, generosity and perception of corruption. The model showed moderate ability to predict the happiness index, with R^2=0.702 and RMSE=0.394. These results can be used to further study and analyse the factors affecting the happiness level in Ukraine.

Therefore, defining a favourable environment for human life and development continues to gain increasing importance in our society. The application of machine learning methods allows us to obtain valuable conclusions about the factors that contribute to the creation of a favourable environment for human life and development and to develop effective strategies in urban planning and social policy.

The use of intelligent systems for analysing sociological data is a promising direction for increasing the efficiency of conducting sociological research. The work conducted allows us to draw the following conclusions.

It was established that the analysis of sociological data is carried out in a particular environment of society by collecting questionnaire data related to various aspects of social life. In the conditions of the formation of an information society, methods and means of obtaining and analysing sociological data are changing. The rapid pace of development of information and communication technologies and computing capabilities led to the emergence of computer sociology and the development and widespread use of modern methods of intelligent analysis of sociological data.

One of the areas of sociological research is the analysis of sociological data in order to determine the level of happiness in different countries of the world. The main approaches to analysing the level of happiness in different countries of the world are considered, and the features of analysing sociological data are determined. It was established that there are approaches that provide for the calculation of happiness indices based on the data obtained, which allow determining the happiness rating in a particular country: The Legatum Prosperity Index, Happiness and Life Satisfaction, and Happy Planet Index. The application of Data Mining methods to the processing of sociological survey data makes it possible to use clustering methods to divide countries into related groups according to indicators that determine the level of happiness. The construction of a classifier makes it possible to classify new countries. The choice of sociological data analysis methods for clustering and classification – k-means and k-nearest neighbours' algorithms - is justified, and their main stages are revealed.

To develop a system for intelligent analysis of sociological data, the cross-platform integrated development environment PyCharm was used, which is a powerful tool for developing applications, supports the programming languages selected for development, provides a convenient interface, has extended functionality, a code editor and other helpful tools for developers. The project was developed in Python using the Pandas, Matplotlib, SciPy, NumPy, PyQt5, Math, and Sklearn libraries. The Pandas library was used to import data from various file formats and perform data merging, reformatting, and cleaning. The Matplotlib library and SciPy built on it were used for data visualization, editing and storage. The Sklearn library was used to implement the main algorithms for cluster data analysis and classification. An intelligent sociological data analysis system was developed, and software was implemented to allow the downloading of sociological survey data on the standard of living in different countries of the world from a CSV file. To perform intelligent data analysis, the system provides for normalization of the downloaded data and their clustering using the k-means algorithm. Clustering allows you to establish groups of related countries by a group of characteristics that determine the level of happiness of people living in the country: GDP per capita in purchasing power parity, social support, freedom of choice, generosity, perception of corruption, positive influence, and negative affect. The user has the opportunity to view the characteristics of the countries that fall into each cluster and the average values of the characteristics of each cluster. Based on the results of the clustering, the system provides for the construction of a classifier using the k-nearest neighbours algorithm and the classification of new objects. The tasks have been completed. However, in the future, the functionality of the system can be expanded by more detailed configuration of the parameters of the selected clustering and classification algorithms.

Acknowledgement

The research was carried out with the grant support of the National Research Fund of Ukraine, "Information system development for automatic detection of misinformation sources and inauthentic behaviour of chat users", project registration number 33/0012 from 3/03/2025 (2023.04/0012). Also, we would like to thank the reviewers for their precise and concise recommendations that improved the presentation of the results obtained.

References

[1] V. Vysotska, "Analytical Method for Social Network User Profile Textual Content Monitoring Based on the Key Performance Indicators of the Web Page and Posts Analysis," CEUR Workshop Proc., vol. 3171, pp. 1380–1402, 2022.

[2] M. Bublyk, V. Feshchyn, L. Bekirova, and O. Khomuliak, "Sustainable Development by a Statistical Analysis of Country Rankings by the Population Happiness Level," CEUR Workshop Proc., vol. 3171, pp. 817–837, 2022.

[3] L. Chyrun, I. Kis, and V. Vysotska, "Content monitoring method for cut formation of person psychological state in social scoring," in Proc. Int. Conf. Computer Sciences and Information Technologies (CSIT), 2018, pp. 106–112. doi: 10.1109/STC-CSIT.2018.8526624.

[4] R. Yurynets, Z. Yurynets, O. Budiakova, L. Gnylianska, and M. Kokhan, "Innovation and Investment Factors in the State Strategic Management of Social and Economic Development of the Country: Modeling and Forecasting," CEUR Workshop Proc., vol. 2917, pp. 357–372, 2021.

[5] T. Batiuk, V. Vysotska, R. Holoshchuk, and S. Holoshchuk, "Intelligent System for Socialization of Individual's with Shared Interests based on NLP, Machine Learning and SEO Technologies," CEUR Workshop Proc., vol. 3171, pp. 572–631, 2022.

[6] V. Vysotska et al., "Disinformation, Fakes and Propaganda Identifying Methods in Online Messages Based on NLP and Machine Learning Methods," Int. J. Comput. Netw. Inf. Secur., vol. 16, no. 5, pp. 57–85, 2024. doi:10.5815/ijcnis.2024.05.06.

[7] A. Tryhuba et al., "A deep neural network model for predicting the competitive score of social projects for community development," CEUR Workshop Proc., vol. 3711, pp. 55–74, 2024.

[8] S. Popereshnyak, V. Zhebka, and A. Vecherkovskaya, "Methodology for Countering Malicious Information on Social Networks," CEUR Workshop Proc., vol. 3688, pp. 108–121, 2024.

[9] M. Biryukova et al., "Social Computing of the Social Well-being of Refugees and Internally Displaced Persons in Ukraine Using Data Mining Methods," CEUR Workshop Proc., vol. 3403, pp. 410–422, 2023.

[10] O. Morushko, N. Khymytsia, and V. Teslyuk, "Remote Selection of Staff Based on Socionic Analysis of Social Network Content," CEUR Workshop Proc., vol. 3171, pp. 138–149, 2022.

[11] M. Amuchastegui, K. Birch, and W. Kaltenbrunner, "The Intersections between Sociology and STS: A Big Data Approach," Sociol. Perspect., vol. 66, no. 5, pp. 868–887, 2023.

[12] A. Rahman and M. G. Muktadir, "SPSS: An imperative quantitative data analysis tool for social science research," Int. J. Res. Innov. Soc. Sci., vol. 5, no. 10, pp. 300–302, 2021.

[13] K. Plummer, Sociology: The Basics, Routledge, 2021.

[14] A. Purwanto and Y. Sudargini, "Partial least squares structural equation modeling (PLS-SEM) analysis for social and management research: a literature review," J. Ind. Eng. Manag. Res., vol. 2, no. 4, pp. 114–123, 2021.

[15] C. Maione, D. R. Nelson, and R. M. Barbosa, "Research on social data by means of cluster analysis," Appl. Comput. Inform., vol. 15, no. 2, pp. 153–162, 2019.

[16] J. Ko, C. K. Leung, and X. Chen, "Economic crises and happiness: Empirical insights from 134 countries (2008–2019)," Dev. Sustain. Econ. Finance, vol. 6, p. 100040, 2025.

[17] J. Bentham, An Introduction to the Principles of Morals and Legislation, 3rd ed., Oxford: Clarendon, 1996 (orig. 1789), pp. 68–73.

[18] L. B. Aknin and A. V. Whillans, "Helping and happiness: A review and guide for public policy," Soc. Issues Policy Rev., vol. 15, no. 1, pp. 3–34, 2021.

[19] V. Voukelatou et al., "Measuring objective and subjective well-being: dimensions and data sources," Int. J. Data Sci. Anal., vol. 11, no. 4, pp. 279–309, 2021.

[20] J. O. Andersson, "Explaining Finnish Economic and Social Success–And Happiness," Studia Europejskie, vol. 26, no. 4, pp. 177–198, 2022.

[21] A. Jannani, N. Sael, and F. Benabbou, "Predicting quality of life using machine learning: Case of world happiness index," in Proc. 4th Int. Symp. Adv. Elect. Commun. Technol. (ISAECT), IEEE, 2021.

[22] G. Dutschke et al., "Factors contributing for organisational happiness: content, exploratory and confirmatory factorial analysis," Eur. J. Int. Manag., vol. 22, no. 3, pp. 376–406, 2024.

[23] N. Fitriana, F. D. Hutagalung, Z. Awang, and S. M. Zaid, "Happiness at work: A cross-cultural validation of happiness at work scale," PLoS ONE, vol. 17, no. 1, p. e0261617, 2022.

[24] L. Li, X. Wu, M. Kong, D. Zhou, and X. Tao, "Towards the Quantitative Interpretability Analysis of Citizens Happiness Prediction," in IJCAI, pp. 5094–5100, 2022.

[25] S. Khemraj et al., "Prediction of world happiness scenario effective in the period of COVID-19 pandemic, by artificial neuron network (ANN), support vector machine (SVM), and regression tree (RT)," NVEO J., vol. 8, no. 4, pp. 13944–13959, 2021.

[26] E. Detrinidad and V.-R. López-Ruiz, "The Interplay of Happiness and Sustainability: A Multidimensional Scaling and K-Means Cluster Approach," Sustainability, vol. 16, no. 22, p. 10068, 2024.

[27] X. Zhang, J. A. Noah, R. Singh, J. C. McPartland, and J. Hirsch, "Support vector machine prediction of individual Autism Diagnostic Observation Schedule (ADOS) scores based on neural responses during live eye-to-eye contact," Sci. Rep., vol. 14, no. 1, p. 3232, 2024.

[28] A. Chakraborty and C. P. Tsokos, "A real data-driven clustering approach for countries based on happiness score," Amfiteatru Econ., vol. 23, no. 15, pp. 1031–1045, 2021.

[29] M. M. Ulkhaq and A. Adyatama, "Clustering countries according to the world happiness report 2019," Eng. Appl. Sci. Res., vol. 48, no. 2, pp. 137–150, 2021.

[30] P. Kumar et al., "Urban housing: a study on housing environment, residents' satisfaction and happiness," Open House Int., vol. 46, no. 4, pp. 528–547, 2021.

[31] A. I. Jigani, C. Delcea, M. S. Florescu, and L. A. Cotfas, "Tracking Happiness in Times of COVID-19: A Bibliometric Exploration," Sustainability, vol. 16, no. 12, p. 4918, 2024.

[32] T. Ziogas et al., “Happiness, space and place: Community area clustering and spillovers of life satisfaction in Canada,” Appl. Res. Qual. Life, vol. 18, no. 5, pp. 2661–2704, 2023.
[33] C. J. Verma, Z. Illés, and D. Kumar, “An investigation of novel features for predicting student happiness in hybrid learning platforms–An exploration using experiments on trace data,” Int. J. Inf. Manag. Data Insights, vol. 4, no. 1, p. 100219, 2024.
[34] H. Leitgöb, D. Prandner, and T. Wolbring, “Big data and machine learning in sociology,” Front. Sociol., vol. 8, p. 1173155, 2023.
[35] W. Zhang and D. Zinoviev, Introduction to Quantitative Social Science with Python, Chapman and Hall/CRC, 2024.
[36] X. Wu et al., “Aligning Bytes with Bliss: Integrating Happiness Computing with Sociological Insight,” in Int. Conf. Adv. Data Min. Appl., Singapore: Springer Nature Singapore, 2024.
[37] G. D. Garson, Data Analytics for the Social Sciences: Applications in R, Routledge, 2021.
[38] Y. Zhang et al., “Big data and human resource management research: An integrative review and new directions for future research,” J. Bus. Res., vol. 133, pp. 34–50, 2021.
[39] M. G. Majumder, S. D. Gupta, and J. Paul, “Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis,” J. Bus. Res., vol. 150, pp. 147–164, 2022.
[40] A. Dhariwal et al., “ResistoXplorer: a web-based tool for visual, statistical and exploratory data analysis of resistome data,” NAR Genomics Bioinform., vol. 3, no. 1, p. lqab018, 2021.
[41] C. Gaston-Breton et al., “Pleasure, meaning or spirituality: Cross-cultural differences in orientations to happiness across 12 countries,” J. Bus. Res., vol. 134, pp. 1–12, 2021.
[42] R. Ghezelbash et al., “Genetic algorithm to optimize the SVM and K-means algorithms for mapping of mineral prospectivity,” Neural Comput. Appl., vol. 35, no. 1, pp. 719–733, 2023.
[43] Y. Wang et al., “A systematic review on affective computing: Emotion models, databases, and recent advances,” Inf. Fusion, vol. 83, pp. 19–52, 2022.
[44] J. D. Barrera, “Mathematics Academic Performance: Multiple Regression Analysis Model,” Int. J. Multidiscip. Appl. Bus. Educ. Res., vol. 5, no. 10, pp. 3905–3910, 2024.
[45] S. Oh et al., “Deep learning model based on expectation-confirmation theory to predict customer satisfaction in hospitality service,” Inf. Technol. Tour., vol. 24, no. 1, pp. 109–126, 2022.
[46] World Happiness Report 2012–2025. [Online]. Available: https://worldhappiness.report/analysis/

Authors’ Profiles

Yuriy Ushenko is a Professor in the Computer Science Department at Chernivtsi National University, Chernivtsi, Ukraine. Research Interests: Data Mining and Analysis, Computer Vision and Pattern Recognition, Optics & Photonics, and Biophysics.

Victoria Vysotska is a Professor at the Information Systems and Networks Department of Lviv Polytechnic National University. She defended her Doctoral degree in Technical Science, speciality in «structural, applied and mathematical linguistics», in 2023 on the topic "Analysis and synthesis of computational linguistic systems for processing Ukrainian textual content" Also, she received a PhD degree in Information Technologies from Lviv Polytechnic National University in 2014. She has currently published more than 500 publications. Her main research interests are focused on the identification of disinformation/fakes/propaganda, detection of sources of disinformation and inauthentic behaviour (bots) of coordinated groups based on artificial intelligence, NLP, computer linguistics, data science, system analysis, information technologies, machine learning, cyber security.

Daryna Zadorozhna is an ambitious senior student at Lviv Polytechnic National University, studying in the Department of Information Systems and Networks with a specialization in Business Analysis. She is an enthusiastic future professional with a deep interest in Marketing, Communication Strategy, and the role of Psychology in modern communication. Throughout her academic journey, Daryna has been developing a strong foundation in combining technical expertise with business-driven thinking. Her interest in data visualization and analytical methods further enriches her ability to generate meaningful insights in both business and communication contexts.

Mariia Spodaryk is a dedicated senior student pursuing an undergraduate degree in the Department of Information Systems and Network at Lviv Polytechnic National University, majoring in Business Analysis. She is a passionate emerging professional with a strong interest in Project Management, Data Analytics and Artificial Intelligence. Alongside her academic pursuits, she is an active Project Manager. Her goal is to bridge the gap between technical insight and business strategy, making her a valuable asset in any tech-oriented environment.

Zhengbing Hu: Prof., Deputy Director, International Center of Informatics and Computer Science, Faculty of Applied Mathematics, National Technical University of Ukraine "Kyiv Polytechnic Institute", Ukraine. Adjunct Professor, School of Computer Science, Hubei University of Technology, China. Visiting Prof., DSc Candidate in National Aviation University (Ukraine) from 2019. Primary research interests: Computer Science and Technology Applications, Artificial Intelligence, Network Security, Communications, Data Processing, Cloud Computing, Education Technology.

Dmytro Uhryn graduated from Yuriy Fedkovych Chernivtsi National University, Chernivtsi. Currently, he is a Doctor of Technical Sciences and associate professor at Yuriy Fedkovych Chernivtsi National University. His research interests are data mining, information technologies for decision support, swarm intelligence systems, and industry-specific geographic information systems.

How to cite this paper: Yuriy Ushenko, Victoria Vysotska, Daryna Zadorozhna, Mariia Spodaryk, Zhengbing Hu, Dmytro Uhryn, "Agile Technology of Information Data Engineering for Intelligent Analysis of the Happiness Index and Life Satisfaction in Known World Cities", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.17, No.3, pp. 99-153, 2025. DOI:10.5815/ijieeb.2025.03.07