Ministry of Education and Science of Ukraine
Sumy State University
Educational and Scientific Institute of Business, Economics and Management
Department of Economic Cybernetics

BACHELOR'S QUALIFICATION WORK

on the topic “FORECASTING THE IMPACT OF THE COUNTRY'S DIGITALIZATION LEVEL ON ITS ECONOMIC DEVELOPMENT”

Completed student of 4th (course number) course, group AB-81a.en (group code)
Specialties 051 “Economics” (Business analytics)
M.A. Litsman
(student's last name, initials)
Supervisor Dr. Sc. in Economics, professor,
H.M. Yarovenko
(position, degree, last name, initials)

Sumy-2022

Ministry of Education and Science of Ukraine
Sumy State University
Educational and Scientific Institute of Business, Economics and Management
Department of Economic Cybernetics

APPROVE
Head of the Department
Dr. Econ. Sciences, Professor
_____ Kuzmenko O.V.
“__”_____2022

TASK
FOR THE BACHELOR'S QUALIFICATION WORK
in the direction of training 051 Economic (Business analytics)
student 4th year of the group АБ-81а.en

Litsman Maryna Andriivna

1. Topic of the work: analysis and forecasting the impact of the country's digitalization level on its economic development approved by order of the university 0382-III from 15.03.2022.
2. The deadline for the student to submit the completed work "_"_____ 2022.
3. The purpose of the work is to analyze the impact of the country's digitalization level on its economic development to build the forecast models.
4.The object of the study is the economic indicators of the world.
5. The subject of research is statistical, analytical and software tools used to identify economic factors that are affected by the level of digitalization of the country.
6. Article in a professional journal of category B and implementation in the discipline of the educational process "Introduction to Business Analytics" and "Forecasting of socio-economic processes"..
7. Indicative plan of qualification work, terms of submission of sections to the head and the maintenance of tasks for performance of the set purpose

8. Consultations on work:

Chapter	Consultant	Signature, data	
		Task issued by	Task accepted by
1		H.M. Yarovenko.	M.A. Litsman
2		H.M. Yarovenko.	M.A. Litsman.
3		H.M. Yarovenko.	M.A. Litsman.

9. Date of issue of the task“___”_____20__ p.

Supervisor _____ _H.M. Yarovenko __
Signuture Initials, surname

Received the task to perform _____ _M.A. Litsman ____
Signuture Initials, surname

ABSTRACT

of the qualifying work

for obtaining the educational and qualification level “bachelor”

Litsman Maryna Andriivna

ANALYSIS AND FORECASTING THE IMPACT OF THE COUNTRY'S DIGITALIZATION LEVEL ON ITS ECONOMIC DEVELOPMENT

The fourth industrial revolution led to the invention and implementation into various spheres of life of artificial intelligence, the Internet of Things and Services. We can say that the economy is being transformed into a digital one.

The main advantages of digitalization are: increasing productivity by reducing manual labor and thus improving product quality.

On the other hand, digitalization can increase unemployment due to robotization of processes, people can be not fine with.

But today the main problem is that digitalization highly affects on the level of cybercrime. That is why the topic of the diploma is relevant, which is devoted to determining the impact of the level of digitalization of the country on the factors of its economic development in order to identify those who have the closest relationships, which will build models for forecasting.

The object of this study is the economic indicators of the world, such as GDP, life expectancy, income, etc.

The subject of the study is statistical, analytical and software tools used to identify economic factors that are affected by the level of digitalization of the country.

Methods of research – analysis of empirical databases for 138 countries on their economic and digital development in 2019.

Information base – World Bank Internet resource and the e-Governance Academy Foundation; Python programming language documentation.

The main contribution of the work is the results of the study were tested by publishing 1 article in a professional journal of category B and implementation in the discipline of the educational process "Introduction to Business Analytics" and "Forecasting of socio-economic processes".

Keywords: digitalization, economic development, cluster analysis, principal components method, polynomial regression.

The content of the qualification work is presented on 31 pages. The references consist from 30 names, placed on 3 pages. The work contains 20 figures, 2 appendices.

Year of performance of qualification work – 2022. Year of protection of work – 2022.

CONTENT

INTRODUCTION

The digital world has evolved at a huge rate in the last decade. The development of the Internet, mobile communications, and online services is a basic tool for shaping the digital economy.

These processes affect all sectors of the economy and social activities, manufacturing, health care, education, finance, transport, and so on. It is known that not all countries are developing equally and an important feature is the rapid increase in the digital divide, which threatens to lag behind. For any country, the manufacturing sector and maintaining its own technological level is a strategically important national task for the development of the economy, services and ensuring the growth of income and national welfare. [1].

That is why the topic is devoted to statistical analysis and modeling of the process of identifying economic factors that affect the level of digitalization of the country. Accordingly, the relevance of the topic allowed to determine the object and subject of research.

The object of this study is the economic indicators of the world, such as GDP, life expectancy, income, etc.

The subject of the study is statistical, analytical and software tools used to identify economic factors that are affected by the level of digitalization of the country.

– The object and subject of the study was determined by its purpose. The purpose of the research is to study the impact of digitalization of countries on their economic development, which is to identify clusters of countries with similar trends, as well as to build models for forecasting the development of individual economic factors depending on the level of digitalization. To achieve this goal it is necessary to implement the following tasks:

– describe the essence of digitalization of the country;

– identify economic factors that are affected by the level of digitalization of the country;

– develop a conceptual research model;

– calculate and analyze basic statistics;

– visualize the main factors;

– perform correlation analysis and apply the principal components method;

– to carry out a cluster analysis of countries by the level of digitalization, taking into account the level of economic development;

– build models for predicting the impact of digitalization on the factors of economic development.

The information and factual base of the study consisted of: empirical databases for 138 countries on their economic and digital development in 2019. The source of indicators presented in the paper is the World Bank Internet resource and the e-Governance Academy Foundation; Python programming language documentation used for calculations. An array of data was generated and cleaned up for missing values, duplicates, anomalous emissions, etc.

The results of the study were tested by publishing 1 article in a professional journal of category B and implementation in the discipline of the educational process "Introduction to Business Analytics" and "Forecasting of socio-economic processes".

1 ANALYSIS OF THE PROCESS OF IDENTIFYING ECONOMIC FACTORS AFFECTED BY THE LEVEL OF DIGITALIZATION OF THE COUNTRY

1.1 The essence of digitalization of the country

Digitalization is the introduction of modern digital technologies in various spheres of life and production. Globalization is a concept of economic activity based on digital technologies implemented in various spheres of life and production. And this concept is widely implemented in all countries.

Why haven't all countries reached digitalization yet and why isn't it working as globally as we would like? There is one small nuance: to digitize all countries, you need to start electrifying them. People in Africa or the northernmost parts of our planet, for example, find it difficult to explain the advantage of a smart refrigerator that will check the freshness of products and order new ones if necessary. Especially if these people store all the products in the basement and grow everything in their own gardens. They simply do not understand modern technologies.

Apparently, the main area where digitalization is sought in all countries is the economy, which is gradually becoming digital today. That is, all data is processed digitally.

The digital economy is an economy based on technology using leading computing devices and technology. This type of economy is sometimes confused with the Internet economy or the web economy. Digital economy is the sale, production and delivery of goods (services and documentation) via the Internet.

The term "digital economy" first appeared in 1995 in the scientific works of Professor of Management D. Topscott from Canada and quickly spread all over the word, replacing such economic sciences and concepts as "New Economy", "Web Economy", "Internet Economy" , "Network Economy", and giving this term a more specific meaning. According to the scientist, in comparison with the traditional market, the advantages of digitalization should include: the virtual nature of economic relations; lack of physical weight of products, the equivalent of which will be the

amount of information; low level of costs for the production of electronic goods and less space that will be occupied by electronic means and media; instant global data exchange via the Internet; emergence of new digital currencies.

The transformation into a digital economy allows citizens to access services and goods faster and easier. Consequently, a huge contribution to the development of the digital economy of a state is planned.

The digital economy of the state, which focuses on a new type of information and telecommunication technologies, is a key and modernized sign of sustainable economic development of the country. There are objective reasons for this: the use of computers in all spheres of life, the absolute and full use of mobile devices, the growing network dependence in society, the constant need for the most "digital" workers in the labor market.

If we take into account the updates in the economic dictionary, then such terms as cyberphysical systems, sensor technologies, big data analysis technologies, etc. have been established. For developing countries, the reorientation of the economy to a digital way of functioning is accelerating.

The digital economy can be defined as an economy based on the transfer of data, which due to modern opportunities for their generation, analysis and further decision-making become a valuable resource. Provision of digitalization is carried out in the information and computer environment by means, tools and objects of information infrastructure.

If we classify the countries of the world according to the level of digitalization, we can distinguish 4 levels of population development: 1) constrained - the level of digitalization of the economy 1-29% (65 countries - Ghana and the lion's share of Africa, Hungary, Vietnam, etc.); 2) emerging - the level of digitalization of the economy 29-39% (19 countries - Tatarstan, Kazakhstan, China, etc.); 3) transitional - the level of digitalization of the economy 39-49% (28 countries - Peru, Brazil, Mexico, Ukraine, etc.); 4) advanced development - the degree of digitalization of the economy more than 49% (37 countries - the United States, Finland, Sweden, Norway, France, Germany, etc.) [2]

According to a study by analysts of the UN Development Program according to PWC criteria, the economy of Ukraine was classified as the third level, which may confirm the reliability of increasing the possibility of ICT in the country. In terms of gross domestic product per capita, Ukraine ranks 133rd and 84th in the HDI rankings.

Also, it can be noted that the fourth level is developed countries, where the level of digitalization is higher. These countries are characterized by a high level of GDP per capita and an indicator of human development. Thus, the level of GDP per capita and the human development indicator are directly proportional to the level of the country's digital economy.

In Ukraine, digitalization in the potential of opportunities can become a basis for stimulating economic growth, the basis of a new path of development in terms of depletion of traditional raw materials for the domestic economy. It is clear that the key to the success of digital modernization of the economy is a comprehensive study of the digitalization process [3].

1.2 Characteristics of economic factors affected by the level of digitalization of the country

At the beginning of the study it is necessary to determine the indicators that will be used in the calculation process and which will characterize, on the one hand, the degree of digitalization of countries, and on the other hand, the level of their economic development. The Digital Development Level (DDL) index, determined by the e-Governance Academy Foundation, was chosen for the first feature, based on the degree of development of information and communication technologies and network readiness of countries to implement and use the latest technologies. This indicator shows how fast the digital society in the country is developing. The higher its value, the more powerful are the processes of digitalization and informatization of the country.

For the second feature, 11 indicators were selected, which are used in various scientific studies to analyze the economic development of countries:

– Gross Domestic Product Per Capita (GDP) – gross domestic product per capita. This is a key indicator that characterizes the level of economic development and represents the sum of gross value added of all resident producers, including all taxes, but excluding subsidies on products, depreciation of assets, depletion of natural resources spent on production;

– Inflation in Consumer Prices (ICP) – inflation rate, expressed in consumer prices. This is the most commonly used indicator for characterizing the economic development of countries, showing the percentage of changes in the prices of the consumer basket of goods and services consumed by households. The highest inflation rates indicate low rates of economic development and inefficiency of public administration in economic development and policy;

– Unemployment Total (UT) – unemployment rate. It is a part of the labor force that does not have a job, but is at the stage of its active search. Its high level is typical of countries with low economic and social development;

– Vulnerable Employment Total (VET) – level of vulnerable employment. Shows the level of self-employed and family workers without receiving compensation for their work. High values of this indicator are typical for countries with low rates of economic development and typical for economies with a large agricultural sector and low level of industry;

– Government Expenditure On Education (GEE) – government spending on education. Characterize the level of economic development and provides for public expenditures, financed by transfers from international sources of government, for the formation of a quality education system in the country;

– Revenue Excluding Grants (REG) – income excluding grants. This is an indicator that includes all the country's cash receipts from taxes, social security contributions and other revenues (fines, fees, rents and income from property or sales). It can probably indirectly depend on the level of digitalization of the country as a source of additional income;

– High-technology Exports (THE) – export of high technologies. It is an indicator that characterizes the level of development of the industry sector related to

research and development in the field of high technology, which indicates the high rate of development of the computer, digital and information technology industry;

– General Government Final Consumption Expenditure (GGFCE) – final consumption expenditure of the general government sector. Related to the government's current expenditures on goods and services, national defense and security expenditures, which also include cybersecurity expenditures in the country;

– Life Expectancy at Birth (LE) – life expectancy. It is an indicator of the quality of life, level of economic and social development of countries. Its highest values correspond to economically developed countries, and the lowest are characteristic of the least developed countries;

– Ease of Doing Business Score (EDB) – assessment of ease of doing business. It is an important indicator of how many countries have created the conditions for organizing and conducting business by different economic entities. Its value, which is close to 100, indicates the most favorable conditions for economic activity, which provides increased economic development through increased tax payments, increased gross domestic product, increased employment, etc;

– Wage and Salaried Workers (WSW) – hired and paid workers. Shows the total number of specialists who are employed in the public and private sectors and receive remuneration in the form of salaries, bonuses, commissions, etc. The highest value corresponds to countries with a high level of economic development.

Empirical data of selected indicators were taken for 138 countries in 2019 from the official Internet resource World Bank and e-Governance Academy Foundation. We import input using the Python programming language [4], namely the Pandas library, which provides ample opportunities for data analysis (Fig. 1.1):

	Country	DDL	GDP	ICP	UT	VET	WSW	GEE	REG	THE	EDB
0	Afghanistan	19.50	516.747871	2.300000	11.73	79.360001	17.809999	10.253860	1.303926e+01	0.0	44.06497
1	Albania	48.74	5246.292306	1.620887	11.70	51.200001	45.730000	13.435470	4.253680e+11	5055767.0	67.74847
2	Algeria	42.81	3306.858208	2.415131	12.83	27.640001	67.709999	16.549191	0.000000e+00	9027398.0	48.59758
3	Angola	22.69	1776.166868	17.100000	7.70	73.710000	21.490000	6.467230	2.026102e+01	77770742.0	41.28838
4	Armenia	55.06	4266.018074	1.211436	20.21	33.070001	66.029999	8.682490	1.566552e+12	28749973.0	74.49401

THE	EDB	GGFCE	LE
0.0	44.06497	0.000000e+00	64.833
5055767.0	67.74847	1.797136e+09	78.573
9027398.0	48.59758	2.916543e+10	76.880
77770742.0	41.28838	5.927841e+09	61.147
28749973.0	74.49401	2.027079e+09	75.087

Figure 1.1 - Import input data

1.3 Conceptual research model

In order to systematize the future analysis, a conceptual model of the study was created (Fig. 1.2). In the first step, indicators are selected, statistical analysis is performed using the Python programming language. The data are analyzed by calculating the main statistical characteristics, then they are grouped by specific characteristics and the distribution of each of the characteristics is built. A correlation matrix is calculated, which identifies variables with a strong relationship that is taken into account in the regression process. To determine the map of clusters of countries, the method of main components will be applied to the factors that characterize economic development, inorder to eliminate multicollinearity between them and reduce the dimensionality ofdata.

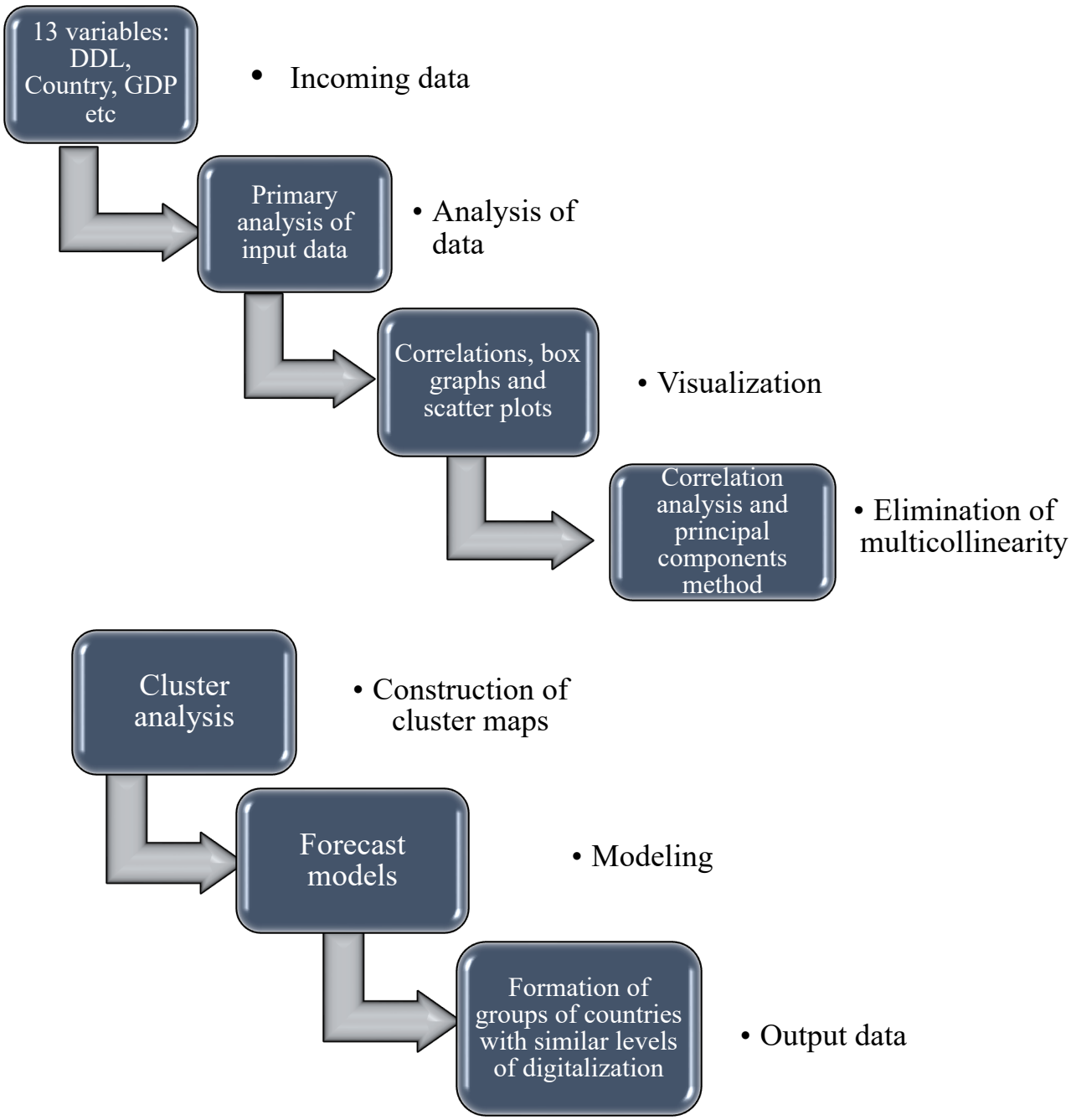


Figure 1.2 - Conceptual research model

The main components are selected. The number of the most optimal clusters is determined on the basis of the Elbow Method. In the last step, forecast models are built that demonstrate the impact of digitalization on the most correlated factors that characterize economic development.

2 STATISTICAL ANALYSIS OF THE INFLUENCE OF THE LEVEL OF DIGITALIZATION OF THE COUNTRY ON ITS ECONOMIC DEVELOPMENT

2.1 Calculation and analysis of basic statistics

We will perform a statistical analysis, which will determine the main statistical characteristics of each numerical feature, using the method describe (): count - the total number of observations, mean - the average value of the sample, max - maximum value, min - minimum value, std - standard deviation, 25% - the first quartile, 50% - the second quartile, 75% - the third quartile (Fig. 2.1).

	DDL	GDP	ICP	UT	VET	WSW	GEE	REG	THE	EDB
count	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000	138.000000	1.380000e+02	1.380000e+02	138.000000
mean	51.857391	15834.450702	4.765994	7.626159	33.797102	62.953116	13.404825	1.838869e+13	1.777898e+10	66.617098
std	18.866114	20791.411642	13.996997	4.894541	25.559928	25.028640	5.311128	1.676105e+14	7.076923e+10	12.205490
min	11.300000	238.990726	-2.540315	0.310000	0.140000	7.400000	0.000000	0.000000e+00	0.000000e+00	32.685520
25%	35.175000	2397.379487	0.499698	4.462500	10.712500	45.670000	10.605022	2.714367e+01	1.123318e+07	59.131052
50%	52.085000	6103.306413	2.428515	6.255000	27.985001	68.134998	13.273515	8.524265e+10	2.234174e+08	68.835555
75%	66.590000	20365.874941	4.023606	9.370000	51.487501	85.032498	16.326357	1.046536e+12	4.357825e+09	76.478655
max	84.170000	116014.602497	150.322724	28.740000	92.250000	99.589996	28.839149	1.955102e+15	7.576827e+11	86.764650

Figure 2.1 - Statistical characteristics of numerical features

The following conclusions were obtained from Figure 2.1:

1) average total income, excluding grants - 1.838869, standard deviation - 1.676105, minimum and maximum values of 0 and 1.955102, respectively;

2) the average value of inflation and consumer prices - 4.76599, the standard deviation - 13.99699, the minimum and maximum values - -2 and 150, respectively.

Find the missing values in this object of a number of each numeric feature using the isnull () method. Analyzing Fig. 2.2, there are no missed objects in this row.

```
Country    0
DDL        0
GDP        0
ICP        0
UT         0
VET        0
WSW        0
GEE        0
REG        0
THE        0
EDB        0
GGFCE      0
LE         0
dtype: int64
```

Figure 2.2 - Missing values in rows

2.2 Visualization of the main factors

We are initializing the environment for visualization, for which we will import the most important libraries and tools, which will allow us to build better graphics (Fig. 2.3):

```
import numpy as np
import pandas as pd
pd.options.display.max_columns = 12
#Disable warnings in Anaconda
import warnings
warnings.simplefilter('ignore')
#We will display plots right inside Jupiter Notebook
%matplotlib inline
import matplotlib.pyplot as plt
#We will use the Seaborn Library
import seaborn as sns
sns.set()
#Graphics in SVG format are more sharp and legible
%config InlineBackend.figure_format = "svg"
from pylab import rcParams
rcParams['figure.figsize'] = 5,4
```

Figure 2.3 - Initialization of the software environment

Universal analysis considers only one variable over time. Analyzing a variable independently, as a rule, researchers are mostly interested in the distribution of its values, so we ignore other variables in the data set. Consider different statistical types of variables and perform a visual analysis [5].

Quantitative features acquire ordered numerical values. These values can be discrete, like integers, or continuous, like real numbers, and usually express counting or measuring.

The seaborn library was used to visualize the data. This is a top-level API library based on the matplotlib library. Seaborn contains more adequate default graphics settings. A "complex" type of scatterplotmatrix (Fig. 2.4) was used for the graphs. This visualization will help to see different variables in one picture, such as GDP per capita, overall unemployment, and an indicator of inflation and consumer prices. Figure 2.4 shows a fragment of the distribution graphs. Appendix B provides graphs for all variables.

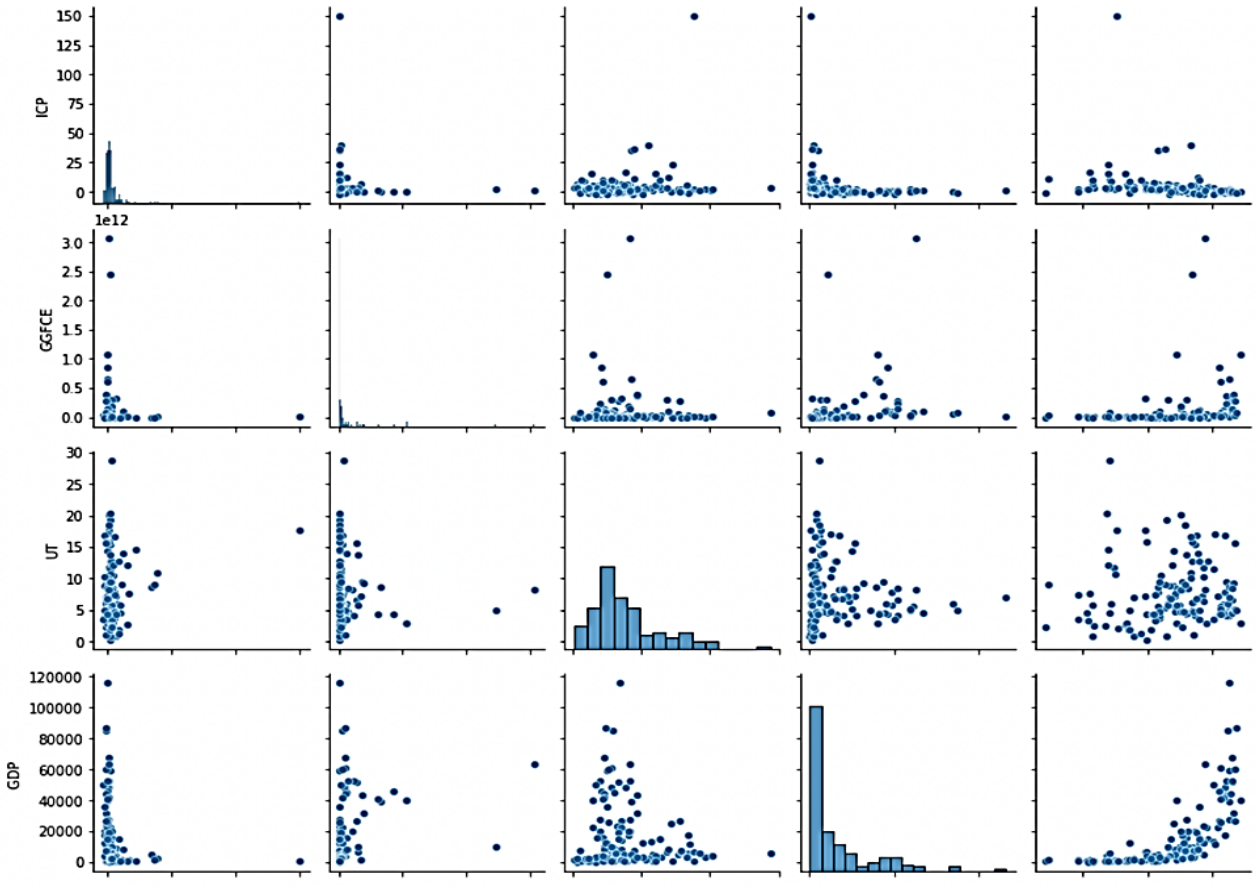


Figure 2.4 - Fragment of the scattering matrix

As we can see, the histograms of the distribution of functions are located on the diagonal of the graph matrix. The rest of the charts are regular scatter charts for the corresponding pairs of objects, which are the most obvious example of the relationship

between two quantitative variables. The dot on the diagram means the values of two variables for one observation at a time [6].

Thus, histograms show that the variables do not meet the law of normal distribution, which must be taken into account in the process of constructing regressions, and scatter plots show the presence of emissions.

A box graph, also known as a Whiskers graph, was used to represent variables to display the sum of a set of data values that have properties such as minimum quartile, median, third quartile, and maximum. The box diagram creates a box from the first quartile to the third quartile, and there is also a vertical line that passes through the box along the median. Here, the x-axis indicates the data to be constructed, and the y-axis indicates the frequency distribution [7].

Figure 2.5 shows a graph of the level of digital development. Visualization of box graphs for other variables is presented in Appendix B.

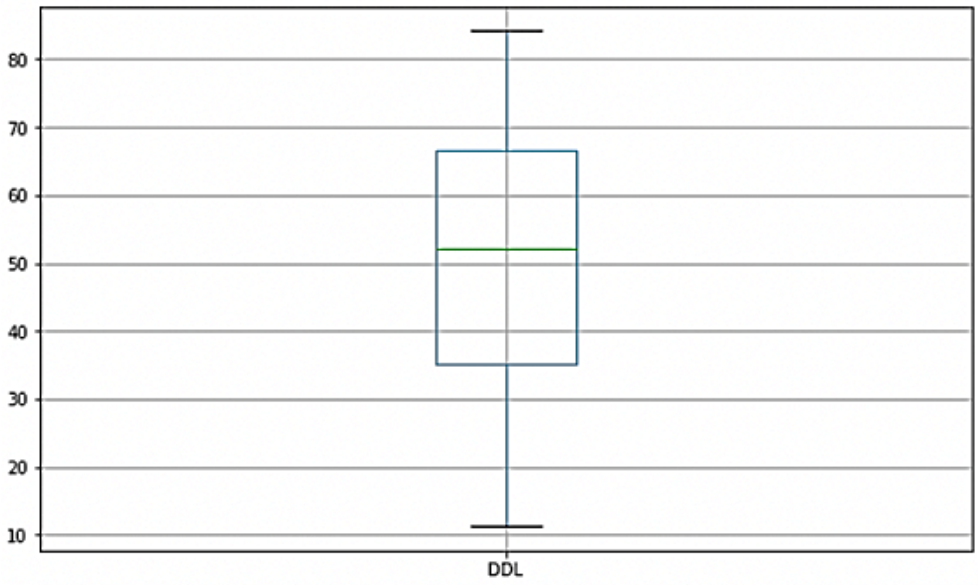


Figure 2.5 - Box graph of the digital development level index

From Visualization 2.5 you can see that there are no emissions. As for other indicators, there are anomalous values for: the level of inflation, expressed in consumer prices (Fig. A.2); final consumption expenditure of the general government sector (Fig. A.3); export of high technologies (Fig. A.5); income excluding grants (Fig. A.6);

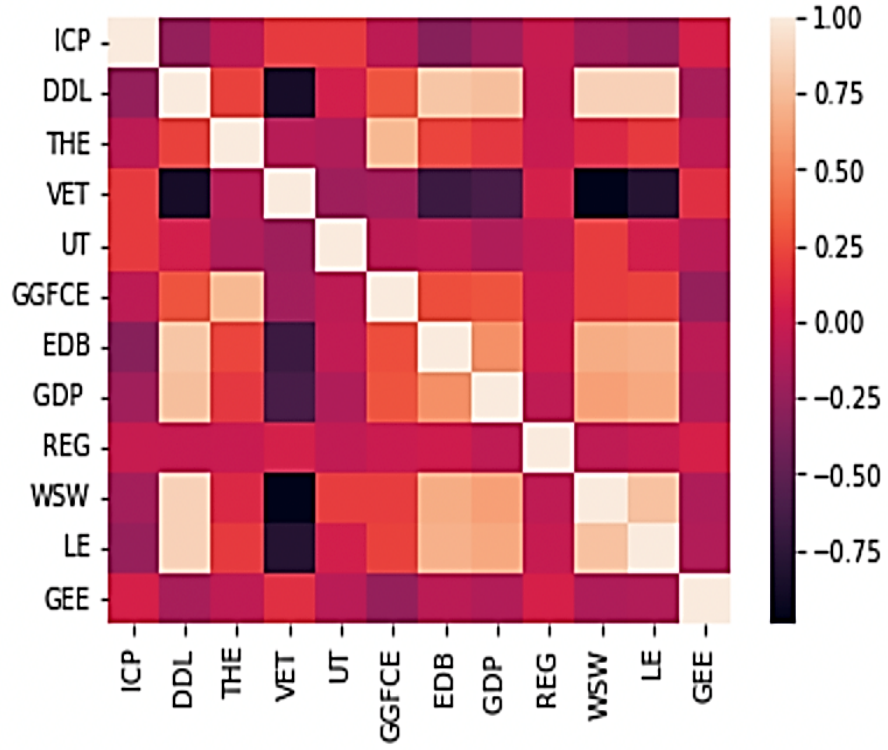
government spending on education (Fig. A.8); unemployment (Fig. A.9); estimates of ease of doing business (Fig. A.10) and life expectancy (Fig. A.11).

Emissions indicate the heterogeneity of the series, but this is due to the fact that data are collected for different countries with different levels of economic development, ie there is a significant difference in the development of the most developed countries and those classified as least developed, which determines emissions.

2.3 Correlation analysis and principal components method

Correlation analysis is a study of the stochastic relationship between quantities that are random. Correlation (from the Latin Correlatio - relationship) is a statistical relationship between random variables, which is probabilistic [8]. The main task of the analysis is to find out the existence of a significant dependence of one variable on others. It is also an estimate of correlation coefficients; checking the significance of sample correlation coefficients or correlation ratios [9].

For example, if we compare the indicators of variables in the presented data set, we have that in most cases it is a positive correlation or direct relationship. If the growth of one variable is carried out by decreasing the values of another, then these are signs of negative correlation, compared to our variables, we have no negative correlation. Zero is a correlation in the absence of a variable. However, a zero overall correlation can only indicate the absence of a linear relationship, and not the absence of any statistical relationship at all. Some variables have zero correlation. Consider the ratio of quantitative features, for which we construct a correlation matrix in the form of a thermal map (Fig. 2.6). It is important to know this information because there are algorithms such as linear and logistic regression that do not order highly correlated input variables. We use the corr () method, which calculates the correlation between each pair of functions.



	DDL	GDP	ICP	UT	VET	WSW	GEE
DDL	1.000000	0.769181	−0.253898	0.042690	−0.864934	0.862447	−0.166523
GDP	0.769181	1.000000	−0.197353	−0.123026	−0.615715	0.616550	−0.117658
ICP	−0.253898	−0.197353	1.000000	0.182204	0.187311	−0.183726	0.060792
UT	0.042690	−0.123026	0.182204	1.000000	−0.211589	0.189453	−0.090955
VET	−0.864934	−0.615715	0.187311	−0.211589	1.000000	−0.996381	0.139400
WSW	0.862447	0.616550	−0.183726	0.189453	−0.996381	1.000000	−0.137218
GEE	−0.166523	−0.117658	0.060792	−0.090955	0.139400	−0.137218	1.000000
REG	−0.023722	−0.059167	−0.020325	−0.046836	0.055211	−0.057335	0.063797
THE	0.215006	0.166847	−0.060918	−0.127999	−0.095615	0.099339	−0.058921
EDB	0.799712	0.545734	−0.314673	−0.046454	−0.674374	0.678091	−0.083462
GGFCE	0.290228	0.287648	−0.067638	−0.064451	−0.187410	0.193775	−0.251719
LE	0.863871	0.657560	−0.239601	0.048150	−0.796437	0.787626	−0.119078

	REG	THE	EDB	GGFCE	LE
DDL	−0.023722	0.215006	0.799712	0.290228	0.863871
GDP	−0.059167	0.166847	0.545734	0.287648	0.657560
ICP	−0.020325	−0.060918	−0.314673	−0.067638	−0.239601
UT	−0.046836	−0.127999	−0.046454	−0.064451	0.048150
VET	0.055211	−0.095615	−0.674374	−0.187410	−0.796437
WSW	−0.057335	0.099339	0.678091	0.193775	0.787626
GEE	0.063797	−0.058921	−0.083462	−0.251719	−0.119078
REG	1.000000	−0.011071	0.024691	0.000251	−0.024255
THE	−0.011071	1.000000	0.231369	0.729089	0.181407
EDB	0.024691	0.231369	1.000000	0.264590	0.697520
GGFCE	0.000251	0.729089	0.264590	1.000000	0.212717
LE	−0.024255	0.181407	0.697520	0.212717	1.000000

Figure 2.6 - Correlation matrix

The results presented in Figure 2.6 show that there is a strong correlation between the level of digital development and factors such as gross domestic product,

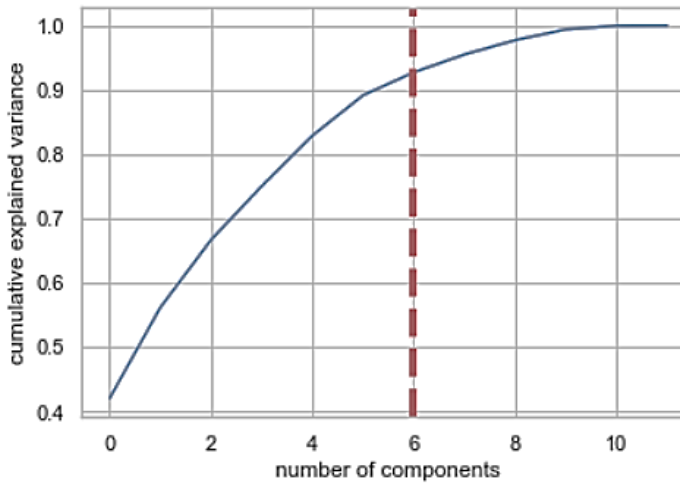
vulnerable employment, hired and paid workers, assessment of ease of doing business and life expectancy. correlations more than 0.7), which indicates the formation of close interactions between individual factors of economic development and digitalization.

At the same time, it can be seen that there is also a strong correlation between these economic indicators, which is due to similar general trends in these factors. Other indicators correlate at the weak (0 - 0.3) and medium (0.3 - 0.7) levels. A high level of correlation will give biased estimates, so this should be taken into account when constructing an analysis.

For the clustering procedure we leave the composition of indicators at the primary level, but since there is a strong correlation between some of them, ie there is a phenomenon of multicollinearity, we use the principal components method, which will reduce the dimensionality of initial data, eliminate multicollinearity and take into account which are weakly correlated. The latter is possible due to the fact that the input data are spatial rather than temporal, so even minor effects of indicators can contribute to more accurate cluster identification for a given country.

PCA – principal component analysis – a method in statistics for factor analysis that uses orthogonal transformation of a set of observations with possibly related variables and is used to eliminate multicollinearity in an array of variables [10].

To apply the method of main components, we use only those indicators that characterize the level of economic development of the country. Its results are presented in Figure 2.7, where we can see that the most optimal for the study are six vectors for which the accumulated variance is 0.8914, and the level of significance of each of the components exceeds 0.05. Appendix C provides program code for executing the Principal Component Method.



	Cumulative Variance Ratio	Explained Variance Rati
0	0.419872	0.41987
1	0.561056	0.14118
2	0.666309	0.10525
3	0.749948	0.08363
4	0.828524	0.07857
5	0.891385	0.06286

Figure 2.7 - The result of the principal components method

Using the pca.transform method, the values of the 6 main components were obtained (Fig. 2.8).

```
array([[-3.66358191e+00,  3.61365566e-01,  7.29161138e-01,
        -5.63075183e-01,  8.47847611e-01, -4.72369068e-01],
       [-5.26411427e-01, -3.84241503e-01,  2.30579007e-01,
        -4.30227141e-02,  2.35768318e-01, -6.43389239e-01],
       [-6.82175004e-01, -8.30205054e-01,  6.65223220e-01,
         2.84961799e-01, -2.25346146e-01, -9.25803748e-01],
       [-3.82229292e+00,  6.19272907e-01,  1.08710423e+00,
        -6.69823043e-01,  8.69599084e-01,  8.96216036e-01],
       [ 3.44321336e-01, -1.04746178e+00,  1.76279032e+00,
         1.32676609e-01,  1.05036399e+00, -1.37530421e+00],
       [ 2.97778280e+00, -2.21945119e-01, -4.77245106e-01,
        -1.26718994e-01, -2.55173470e-01,  5.26973088e-01],
       [ 2.84855814e+00, -3.61674137e-01, -3.70163174e-01,
        -3.61053341e-01,  7.09786578e-02,  7.57570631e-01],
       [-7.20898523e-01,  2.28370393e-01, -3.70813027e-01,
```

Figure 2.8 - The value of the obtained components

3 CLUSTER ANALYSIS AND MODELS OF FORECASTING THE INFLUENCE OF THE LEVEL OF DIGITALIZATION OF A COUNTRY ON ITS ECONOMIC DEVELOPMENT

3.1 Cluster analysis of countries by level of digitalization taking into account the level of economic development

The data obtained from the application of the principal components method and the digital development level indicator formed a set of data that will be used for cluster analysis. In this case, the level of digital development will be input, and the resulting components of economic development indicators - output, ie take into account the possibility of the impact of digitalization processes on a set of factors of economic development. To optimize the clustering procedure, we use the "Elbow Method", which will determine the optimal number of clusters. Its results are presented in Figure 3.1. Appendix D provides program code for executing the Elbow Method.

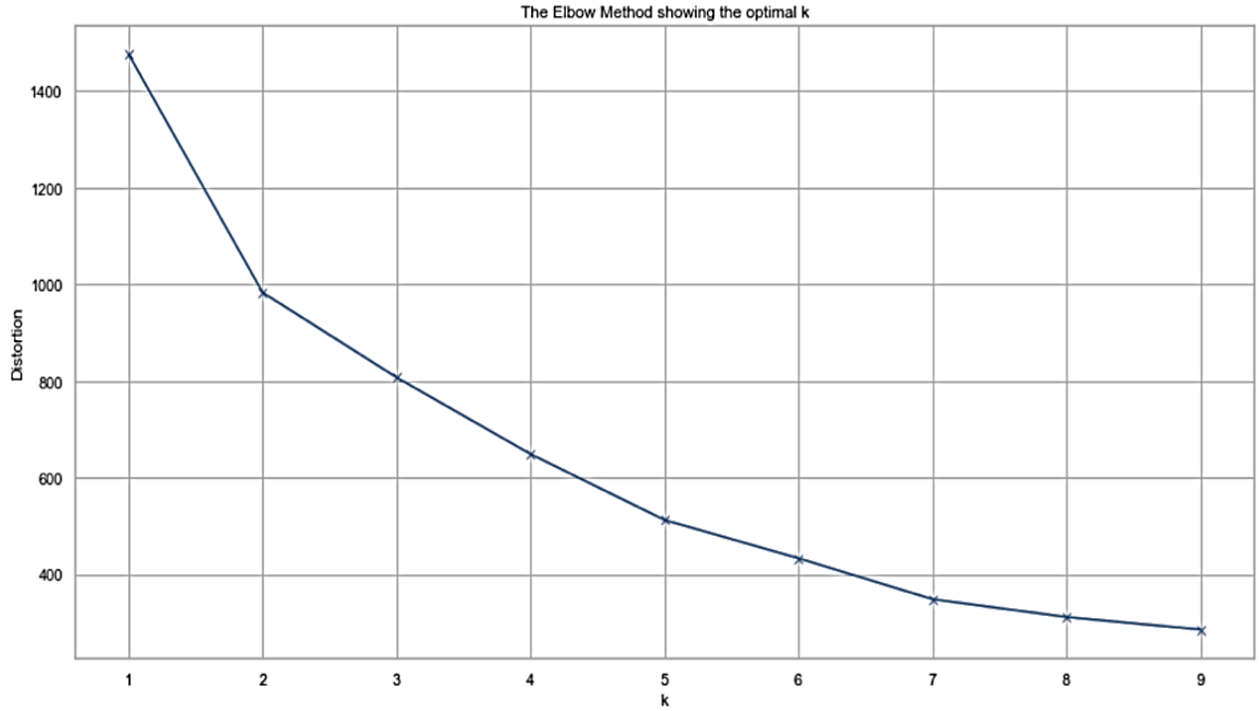


Figure 3.1 - Results of the application of "Elbow Method"

In Figure 3.1 we can see that the points of change of the curve direction are 2, 5 and 7 clusters. The use of two and five clusters for these 138 countries is impractical,

as the resulting clusters will not demonstrate an adequate distribution of countries in terms of digital and economic development. Therefore, for the further clustering procedure we use 7 clusters, the feasibility of which will also be confirmed by the lowest level of maximum and average quantization error, the value of which goes to zero (Figure 3.2).

The clustering process was performed using the k – means method using the Deductor Studio Academic analytical package. Deductor is a technology platform from Loginom to create applied analytical solutions that use the latest methods of data extraction, manipulation, visualization, clustering, forecasting and other technologies of data mining [11].

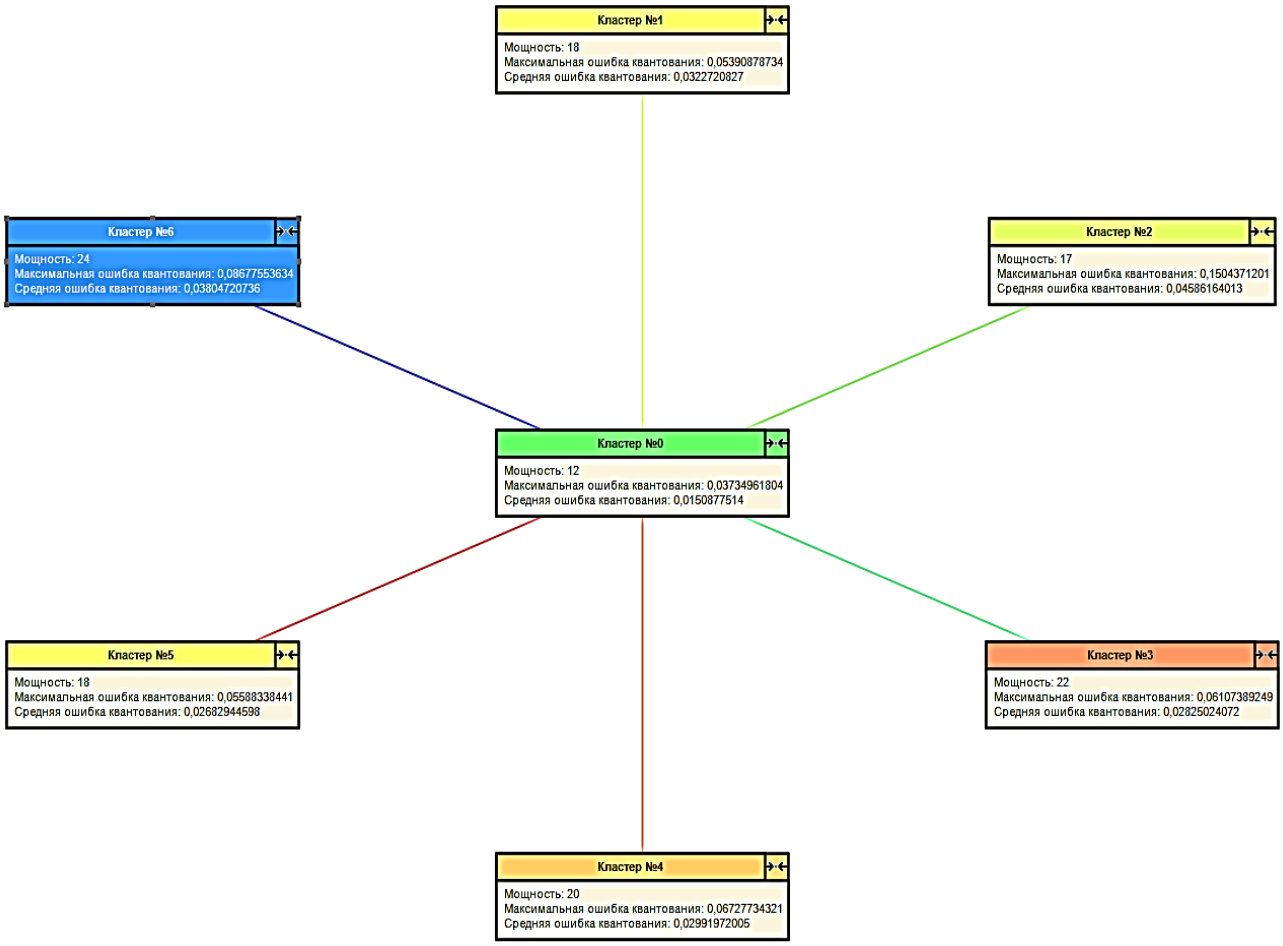


Figure 3.2 - Cluster diagram of countries taking into account quantization errors

We visualize the obtained clusters of countries with the built-in capabilities of the MS Excel software product and build a map (Figure 3.3).

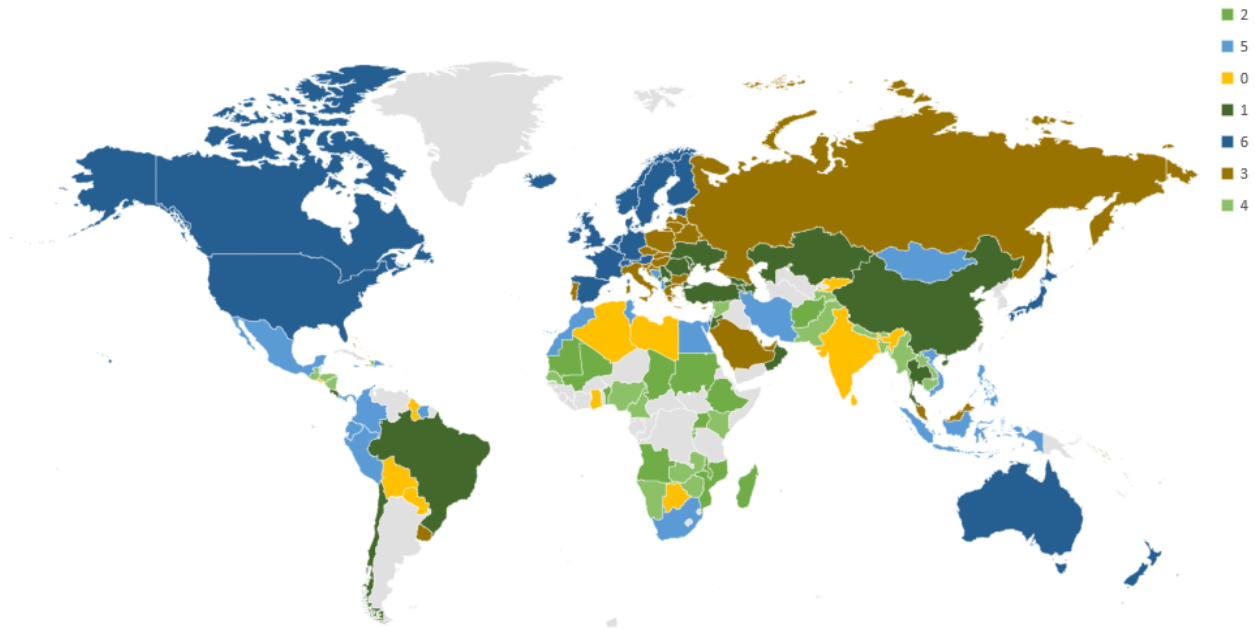


Figure 3.3 - Map of clusters of countries taking into account the impact of the level of their digitalization on economic development

The sixth cluster was the most powerful, which included 25 countries: Australia, the United States, Great Britain, Austria, Canada, Germany, France, Norway, the Netherlands, Finland and others. It is characterized by the highest average level of digitization - 78.7720, a standard error of 0.6642, a standard deviation of 3.3209. This segment is represented by countries with the highest level of economic development, life expectancy, employment, ease of doing business.The least powerful is the zero cluster (Figure 3.3), which includes only 12 countries: Algeria, Bolivia, Botswana, El Salvador, Ghana, Guyana, India, Kyrgyzstan, Libya, Paraguay, Sri Lanka, Tonga. Its average value of the digitization level is 41.8917, the standard error is 0.3902, the standard deviation is 1.3515, and the significance is 93%. It should be noted that the countries of this segment are among those that are economically developing and have a level of digitalization below average.

3.2 Construction of models for predicting the impact of digitalization on the factors of economic development

Linear regression is a regression model that is very often used in statistics to explain the dependence of one variable, which is Y, on another or many other variables, X, which are factors.

The linear regression model is the most widely used and studied in econometrics. Namely, the properties of the estimates of the parameters obtained by different methods under certain assumptions about the probabilistic characteristics of factors and random errors of the model are studied. Boundary (asymptotic) properties of estimates of nonlinear models are also derived based on the approximation of the latter by linear models. It should be noted that from an econometric point of view, linearity by parameters is more important than linearity by model factors [12].

Linear regression equations in general:

$$f(x, b) = b_0 + b_1 * x_1 + b_2 * x_2 +_{...} b_k * x_k, \quad (3.1)$$

where: b_k – regression parameters (coefficients),

x_k – regressors (model factors),

k – number of model factors.

The linear regression coefficients show the rate of change of the dependent variable for this factor. Appendix F provides program code for executing the Liner Regressions.

```
X = df.drop(['DDL','Country','Cluster'], axis=1)
y = df['DDL']
```

```
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
predictions = model.predict(X)

print_model = model.summary()
print(print_model)
```

Figure 3.4 - Construction of linear regression

As a result, the following characteristics of linear regression were obtained (Fig. 3.5):

OLS Regression Results

Dep. Variable:	DDL	R-squared:	0.912
Model:	OLS	Adj. R-squared:	0.905
Method:	Least Squares	F-statistic:	131.6
Date:	Mon, 24 Jan 2022	Prob (F-statistic):	5.33e-62
Time:	17:34:14	Log-Likelihood:	-433.00
No. Observations:	138	AIC:	888.0
Df Residuals:	127	BIC:	920.2
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0031	0.001	-2.666	0.009	-0.005	-0.001
GDP	0.0002	3.46e-05	6.805	0.000	0.000	0.000
ICP	0.0029	0.038	0.075	0.940	-0.073	0.079
UT	-0.0034	0.113	-0.031	0.976	-0.227	0.220
VET	-0.4050	0.077	-5.272	0.000	-0.557	-0.253
WSW	-0.1528	0.098	-1.555	0.122	-0.347	0.042
GEE	-0.1381	0.099	-1.389	0.167	-0.335	0.059
REG	1.378e-15	3e-15	0.460	0.646	-4.55e-15	7.31e-15
THE	1.007e-11	1.07e-11	0.941	0.349	-1.11e-11	3.12e-11
EDB	0.3970	0.063	6.303	0.000	0.272	0.522
GGFCE	-3.816e-14	2.14e-12	-0.018	0.986	-4.27e-12	4.19e-12
LE	0.6310	0.122	5.175	0.000	0.390	0.872

Omnibus:	13.170	Durbin-Watson:	2.110
Prob(Omnibus):	0.001	Jarque-Bera (JB):	17.758
Skew:	-0.545	Prob(JB):	0.000139
Kurtosis:	4.379	Cond. No.	9.49e+15

Figure 3.5 - Linear regression

The value of p for most variables is less than 0.05, except 7, so we will remove them from the model (Fig. 3.6):

```
X2 = df.drop(['DDL','Country', 'GGFCE','UT','ICP','REG','THE','GEE','WSW','Cluster'], axis=1)
y = df['DDL']
```

Figure 3.6 - Elimination of variables that are not statistically significant

The result was the following linear regression (Fig. 3.7):

OLS Regression Results

Dep. Variable:	DDL	R-squared:	0.909
Model:	OLS	Adj. R-squared:	0.907
Method:	Least Squares	F-statistic:	333.3
Date:	Mon, 24 Jan 2022	Prob (F-statistic):	2.96e-68
Time:	17:34:14	Log-Likelihood:	-435.06
No. Observations:	138	AIC:	880.1
Df Residuals:	133	BIC:	894.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.6426	10.306	-2.100	0.038	-42.028	-1.258
GDP	0.0002	3.23e-05	7.301	0.000	0.000	0.000
VET	-0.2458	0.034	-7.311	0.000	-0.312	-0.179
EDB	0.4062	0.059	6.890	0.000	0.290	0.523
LE	0.6897	0.137	5.043	0.000	0.419	0.960

Omnibus:	11.889	Durbin-Watson:	2.072
Prob(Omnibus):	0.003	Jarque-Bera (JB):	14.046
Skew:	-0.558	Prob(JB):	0.000891
Kurtosis:	4.095	Cond. No.	5.47e+05

Figure 3.7 - Linear regression

Linear regression equation:

$$f(x, b) = 0{,}002 + (-0{,}2458) * x_1 + 0{,}4062 * x_2 + 0{,}6897 * x_3 \qquad (3.2)$$

We can draw conclusions from linear regression. The level of digitalization of the country is influenced by such economic factors as: the level of GDP per capita; life expectancy; the level of vulnerable employment without receiving compensation for work; assessment of ease of doing business.

In the last step, we will build forecast models for those indicators that have the highest value of the correlation between the level of digital development and individual factors that characterize the level of economic development. The following indicators were: GDP per capita (0.7692); total life expectancy at birth (0.8639); ease of doing business (0.7997); vulnerable employment (-0.8649). For prediction we use linear, quadratic and cubic regressions.

Figure 3.8 presents three models - linear, quadratic, cubic regression, which show the dependence of gross domestic product on the level of its digitization, as well as coefficients of determination to assess their quality.

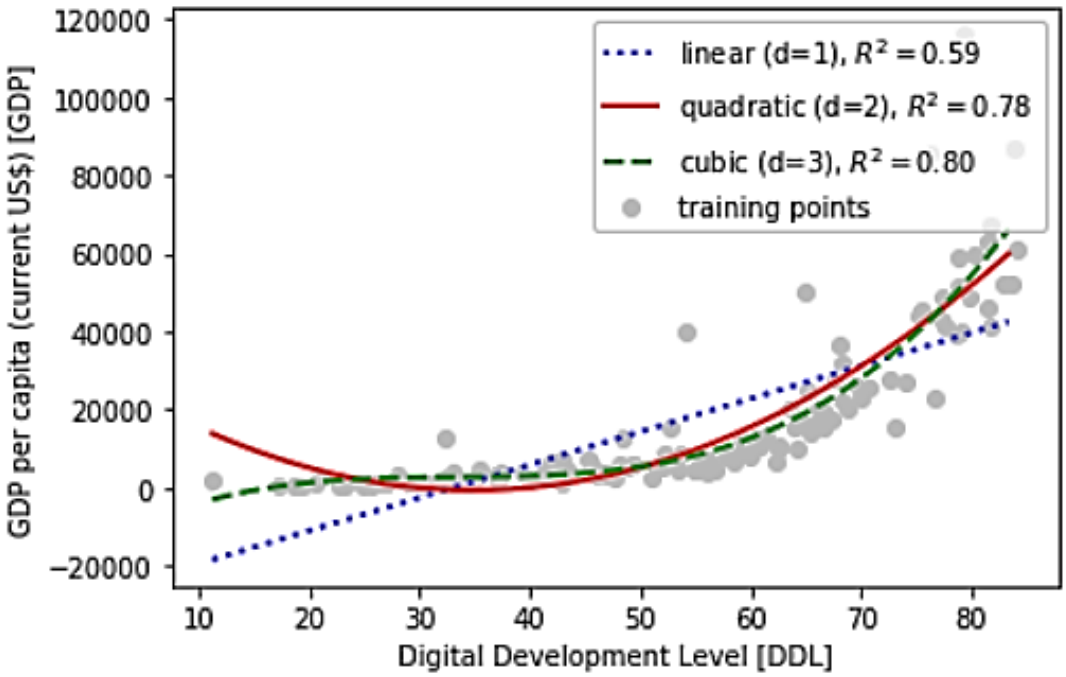


Figure 3.8 - Models of forecasting gross domestic product per capita depending on the level of its digitization

Its highest value is shown by cubic regression, and the lowest - linear. This criterion has the property - to increase with the number of factors in the model. But in this figure we can see that it is the cubic model more accurately reproduces the nature of the dependence of gross domestic product of the country on the level of its digitization. Also, the rms error is smaller for the cubic model. Therefore, it is expedient to predict, and its equation will look like (3.3):

$$y_i = 0{,}4996x^3 - 49{,}9211x^2 + 1682{,}9673x - 16485{,}0826 \qquad (3.3)$$

Figure 3.9 shows linear, quadratic, cubic regressions, which show the dependence of the level of vulnerable employment in the country on the level of its digitalization.

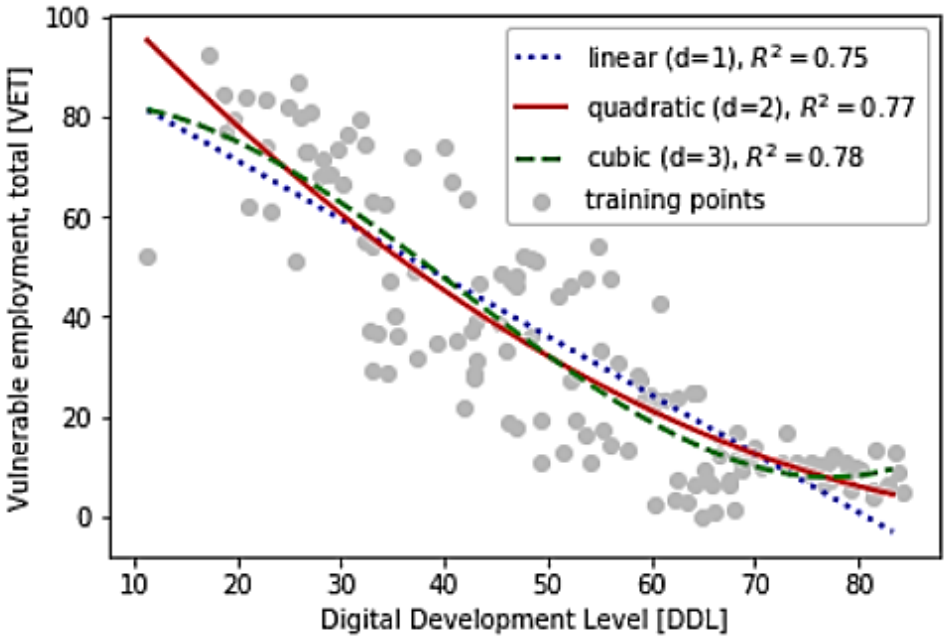


Figure 3.9 - Models for forecasting the level of vulnerable employment in the country depending on the level of its digitalization

The calculated values of the coefficients of determination are quite close for the three models and correspond to the good quality of the forecast models. When comparing the values of the root mean square error, it turned out that the values obtained by cubic and quadratic models are quite close. But if we predict these models a few steps ahead, it turns out that the cubic model is quite sensitive and shows a sharp increase compared to the quadratic model. Therefore, the latter is more suitable for forecasting vulnerable employment (equation (3.4)):

$$y_i = 0{,}0111x^2 - 2{,}3139x + 119{,}9222, \qquad (3.4)$$

Figure 3.10 shows models for forecasting the volume of employees and paid employees of the country depending on the level of its digitalization.

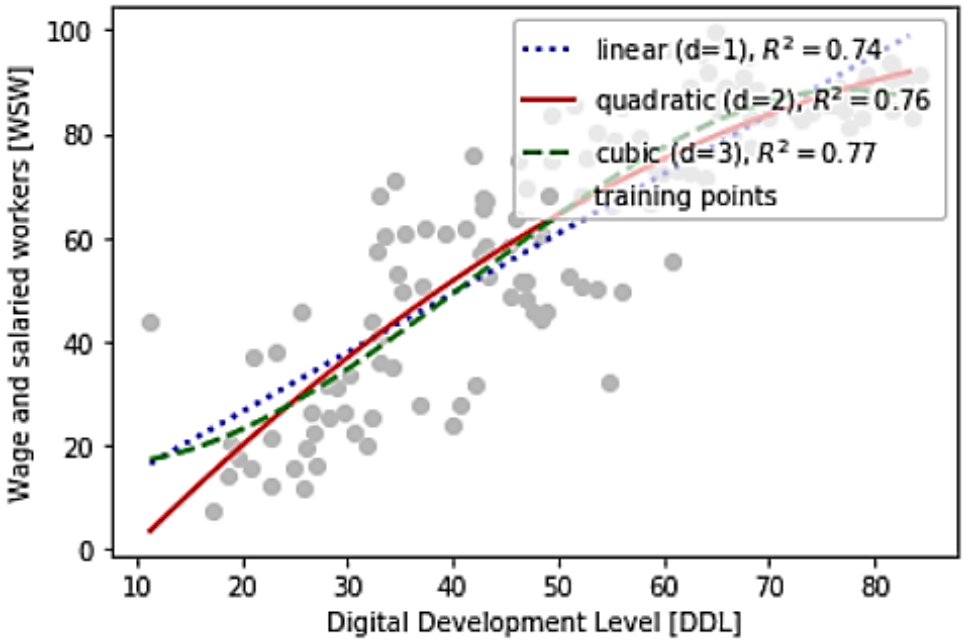


Figure 3.10 - Models for forecasting the number of employees and paid employees of the country depending on the level of its digitalization.

The obtained values of the coefficients of determination for the constructed models are close and testify to the good quality of the forecast models. The cubic model has the smallest value of standard error, but when using it for forecasting, it shows a rather pessimistic version of the dependence, so in this case it is better to use a quadratic model that is not so sensitive. Therefore, its equation will take the form (3.5):

$$y_i = -0{,}0104x^2 + 2{,}2091x - 20{,}0242. \quad (3.5)$$

Linear, quadratic and cubic models were also constructed to predict the impact of the level of digitalization on the assessment of ease of doing business (Figure 3.11). The values of the coefficients of determination are close to the three models and demonstrate the quality of predictive models above average, which is a sufficient indicator for real statistics. Similar values of standard error were also obtained. When using the built models for forecasting a few steps ahead, it was found that all three show similar trends in the forecast, so in practice we can use three models, although the most accurate is the cubic model (equation (3.6)):

$$y_i = -0{,}000018x^3 - 0{,}000898x^2 + 0{,}766258x + 33{,}157690. \quad (3.6)$$

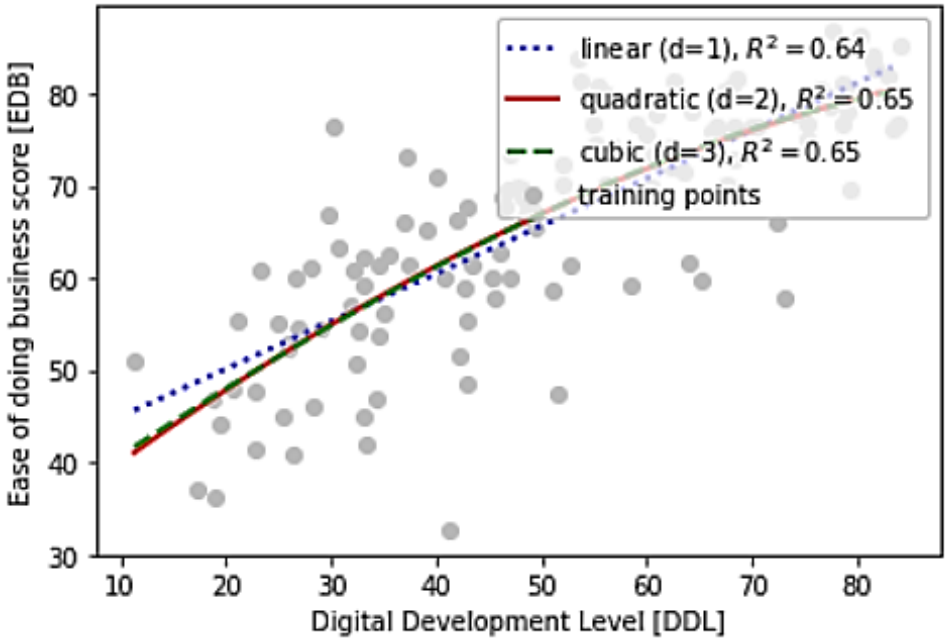


Figure 3.11 - Models of forecasting the assessment of the ease of doing business in the country depending on the level of its digitization

Figure 3.12 presents models for forecasting life expectancy in the country depending on the level of its digitalization, which demonstrate almost the same quality. When forecasting a few steps ahead, they also show the same trends and roughly the same forecast values.

We can conclude that all three equations are quite suitable for prediction (equation (3.7) - (3.9)):

$$y_i = 0{,}3061x + 58{,}0875, \quad (3.7)$$

$$y_i = -0{,}0016x^2 + 0{,}4709x + 54{,}4289, \quad (3.8)$$

$$y_i = 0{,}000021x^3 - 0{,}004729x^2 + 0{,}614802x + 52{,}476635. \quad (3.9)$$

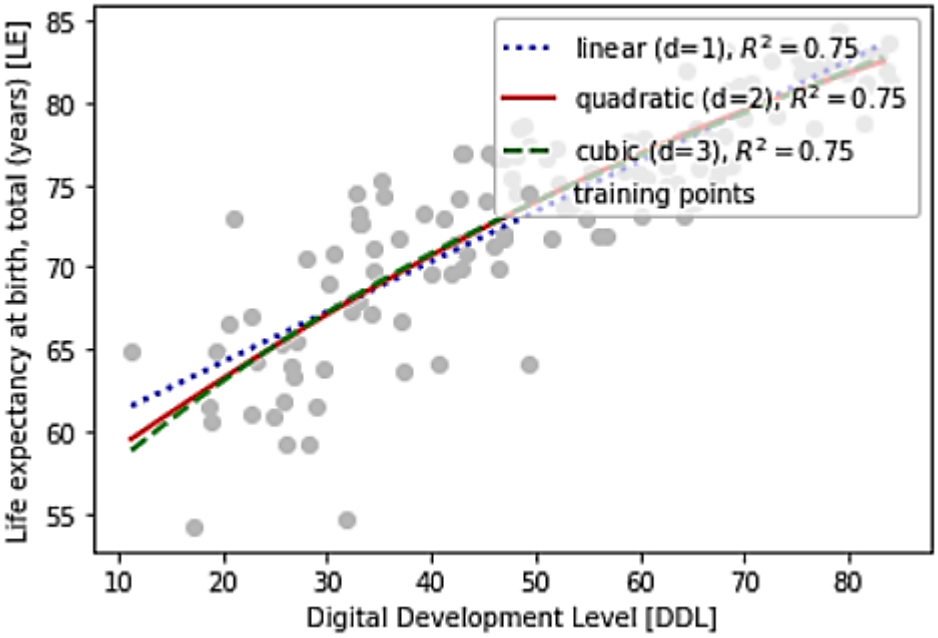


Figure 3.12 - Models of forecasting life expectancy in the country depending on the level of its digitization

CONCLUSIONS

This study reveals the problem of determining the impact of the level of digitalization of the country on the factors that characterize its economic development. The indicator of the level of digital development, which was selected for 138 countries in 2019, was identified as an indicator of impact. The factors that characterize economic development were: gross domestic product per capita, inflation, expressed in consumer prices, unemployment, the level of vulnerable employment, government spending on education, income excluding grants, high technology exports, final costs consumption of the general government sector, life expectancy, assessment of ease of doing business and employees.

For this purpose, statistical analysis was performed and cluster analysis was built using the Python programming language. The data were analyzed by calculating the main statistical characteristics, then they were grouped by specific characteristics and the distribution of each of the characteristics was built. As a result, it was found that the distribution of variables is different, which will be taken into account in the simulation process. A correlation matrix was calculated in which variables with a strong correlation were identified, which was taken into account in the regression construction process.

As a result of constructing a correlation matrix for selected factors, it was found that the most correlated are gross domestic product per capita, life expectancy, assessment of ease of doing business, vulnerable employment and employees and paid workers. This is due to the fact that these indicators have a general trend. It was also found that these factors also have a close linear relationship with the level of digitization, which indicates the possibility of their mutual influence.

To determine the clusters of countries, the principal components method was applied to the factors that characterize economic development, in order to eliminate multicollinearity between them and reduce the dimensionality of data. As a result, six main components were selected. Based on the Elbow Method, it was determined that

the most optimal will be the division of data into 7 clusters. The implementation of k-means clustering allowed to identify clusters that were formed at the expense of countries close in both the level of digitalization and the level of economic development.

In the last step, forecast models were built that demonstrate the impact of digitalization on the most correlated factors that characterize economic development. It is established that for forecasting the gross domestic product of the country and ease of doing business it is advisable to use a cubic log, vulnerable employment, the number of employees - square, the total life expectancy can be used linear, quadratic and cubic models.

REFERENCES

1. The essence of digitalization of the country: website. URL: https://razumkov.org.ua/uploads/article/2020_digitalization.pdf (date of application: 25.03.2022)

2. Classification of countries by type of digitalization: website. URL: https://www.strategyand.pwc.com/media/file/Strategyand_Maximizing–the–Impact–of–Digitization.pdf; Human Development Report 2016. Human Development for Everyone. N. Y. : UNDP, 2016. 286 p. (date of application: 25.03.2022)

3. Ukraine's place in world economic development. Retrieved from: – Диба М. І. Диджиталізація економіки: світовий досвід та можливості розвитку в Україні / М. І. Диба, Ю. О. Гернего // Фінанси України. – 2018. – № 7. – С. 50–63. (date of application: 25.03.2022)

4. Kashnitskiy Y. Open Machine Learning Course. Topic 1. Exploratory Data Analysis with Pandas. *Medium.com* : website. URL: https://medium.com/open–machine–learning–course/open–machine–learning–course–topic–1–exploratory–data–analysis–with–pandas–de57880f1a68. (date of application: 25.03.2022)

5. Polusmak E. Open Machine Learning Course. Topic 2. Visual Data Analysis with Python. *Medium.com* : website. URL: https://medium.com/open–machine–learning–course/open–machine–learning–course–topic–2–visual–data–analysis–in–python–846b989675cd. (date of application: 25.03.2022)

6. Scatter diagram: website URL: https://seaborn.pydata.org/generated/seaborn.scatterplot.html (date of application: 25.03.2022)

7. Box graphics: website: URL: https://www.geeksforgeeks.org/box–plot–in–python–using–matplotlib/ (date of application: 25.03.2022)

8. Correlation analysis: website. URL: Correlation analysis Archived Wayback Machine. – Geostatistics (course of lectures from Khomyak M.M.) (date of application: 27.03.2022)

9. Correlation heat map. *Stackoverflow.com* : website. URL: https://stackoverflow.com/questions/39409866/correlation–heatmap. (date of application: 27.03.2022)

10. Main component method: website. URL: Онлайн руководство «Метод Главных Компонент (PCA)» (date of application: 27.03.2022)

11. Deductor: website URL: https://korusconsulting.ru/platforms/business–analytics/deductor/(date of application: 27.03.2022)

12. Linear regression website URL: https://znaimo.com.ua/Лінійна_регресія (date of application: 27.03.2022)

13. Havryliuk V., Hromyk A., Semenets I., Pylypiuk T., Motsyk R., Kostyakova A. Digitalization of territorial and economic systems at the regional level. *Regional Science Inquiry*. 2021. Vol. 13, no. 2. P. 209–226.

14. Zhang S., Ma X., Cui Q. Assessing the Impact of the Digital Economy on Green Total Factor Energy Efficiency in the Post-COVID-19 Era. *Frontiers in Energy Research*. 2021. Vol. 926, no. 798922. DOI:10.3389/fenrg.2021.798922.

15. Sternik S.G., Safronova N.B. Financialization of Real Estate Markets as a Macroeconomic Trend of the Digital Economy. *Studies on Russian Economic Development*. 2021. Vol. 32, no. 6. P. 676–682. DOI: 10.1134/S1075700721060149.

16. Rodionov D., Zaytsev A., Konnikov E., Dmitriev N., Dubolazova Y. Modeling changes in the enterprise information capital in the digital economy. *Journal of Open Innovation: Technology, Market, and Complexity*. 2021. Vol. 7, no. 3. Article number 166. DOI: 10.3390/joitmc7030166.

17. Didenko N., Skripnuk D., Kikkas K., Kalinina O., Kosinski E. The impact of digital transformation on the micrologistic system, and the open innovation in logistics. *Journal of Open Innovation: Technology, Market, and Complexity*. 2021. Vol. 7, no. 2. Article number 115. DOI: 10.3390/joitmc7020115.

18. Khachaturyan A.A. Unemployment and Other Social Threats of the Digital Economy. *Studies on Russian Economic Development*. 2021. Vol. 32, no. 3. P. 297–304. DOI: 10.1134/S1075700721030151.

19. Ahmed E.M. Modelling Information and Communications Technology Cyber Security Externalities Spillover Effects on Sustainable Economic Growth. *Journal of the Knowledge Economy*. 2021. Vol. 12, no. 1. P. 412–430. DOI: 10.1007/s13132-020-00627-3.

20. Aleksandrova A., Khabib M.D. The role of information and communication technologies in a country's GDP: a comparative analysis between developed and developing economies. *Economic and Political Studies*. 2021. DOI: 10.1080/20954816.2021.2000559.

21. Hak M., Devčić A., Budić H. Determinants of digital taxation in european union. *WSEAS Transactions on Business and Economics*. 2021. Vol. 18. P. 1319–1329. DOI: 10.37394/23207.2021.18.122.

22. Andriychuk O. Shaping the new modality of the digital markets: The impact of the DSA/DMA proposals on inter-platform competition. *World Competition*. 2021. Vol. 44, no. 3. P. 261–286.

23. Zhashkenova R., Pritvorova T., Talimova L., Mazhitova S., Dauletova A., Kernebaev A. Analysis of the transformation of higher educational institutions through entrepreneurship in the conditions of digitalization. *Academy of Accounting and Financial Studies*. 2021. Vol. 25, no. 4. P. 1–10.

24. Python Notes for Professionals: website. URL: https://books.goalkicker.com/PythonBook/ (date accessed 10.06.2021)

25. Shaw Z. Learn Python the hard way: a very simple introduction to the terrifyingly beautiful world of computers and code. Crawfordsville: Addison- Wesley, 2014. 306 p. URL: https://learntocodetogether.com/learn-python-the-hard- way-free-ebook-download/ (date accessed 27.03.2022)

26. McKinney W. Pandas: powerful Python data analysis toolkit, 2021. 3325 p. URL: https://pandas.pydata.org/docs/pandas.pdf (date accessed 27.03.2022)

27. Sedhain S. Web framework for Python: Django, 2006. 190 p. URL: https://www.programmer-books.com/wp-content/uploads/2018/08/Django-Book-Web-framework-for-Python.pdf (date accessed 27.03.2022)

28. Forcier J., Bissex P., Chun W. Python Web Development with Django. Pearson Education inc, 2009. 405 p. URL: https://freepdf-books.com/download/?file=4536 (date accessed 27.03.2022)

29. Dauzon S., Bendoraitis A., Ravindran A. Django: Web Development with Python. Birmingham: Packt Publishing Ltd, 2016. 717 p. URL: http://englishonlineclub.com/pdf/Django%20-%20Web%20Development%20with%20Python%20(Learning%20Path)%20[EnglishOnlineClub.com].pdf (date accessed 27.03.2022)

30. Python Notes for Professionals: website. URL: https://books.goalkicker.com/PythonBook/ (date accessed 27.03.2022)

Appendix A

SUMMARY

Litsman M. A. Analysis and forecasting the impact of the country's digitalization level on its economic development. Qualifying work of the bachelor. Sumy State University, Sumy, 2022.

Indicators were selected. The data was analyzed by calculating the main statistical characteristics. A correlation matrix was calculated. It was determined the map of clusters of countries.

The main components were selected. The number of the most optimal clusters was determined on the basis of the Elbow Method. In the last step, forecast models were built that demonstrate the impact of digitalization on the most correlated factors that characterize economic development.

Keywords: digitalization, economic development, cluster analysis, principal components method, polynomial regression.

АНОТАЦІЯ

Ліцман М. А. Аналіз та прогнозування впливу рівня цифровізації країни на її економічний розвиток. Кваліфікаційна робота бакалавра. Сумський державний університет, Суми, 2022.

Були відібрані показники. Дані проаналізовано шляхом розрахунку основних статистичних характеристик. Була розрахована кореляційна матриця. Була визначена карта кластерів країн.

Підібрано основні компоненти. Кількість найбільш оптимальних кластерів визначали на основі “Ліктьового” методу. На останньому кроці були побудовані прогнозні моделі, які демонструють вплив цифровізації на найбільш корельовані фактори, що характеризують економічний розвиток.

Ключові слова: цифровізація, економічний розвиток, кластерний аналіз, метод головних компонентів, поліноміальна регресія.

Appendix B

Visualization of variables

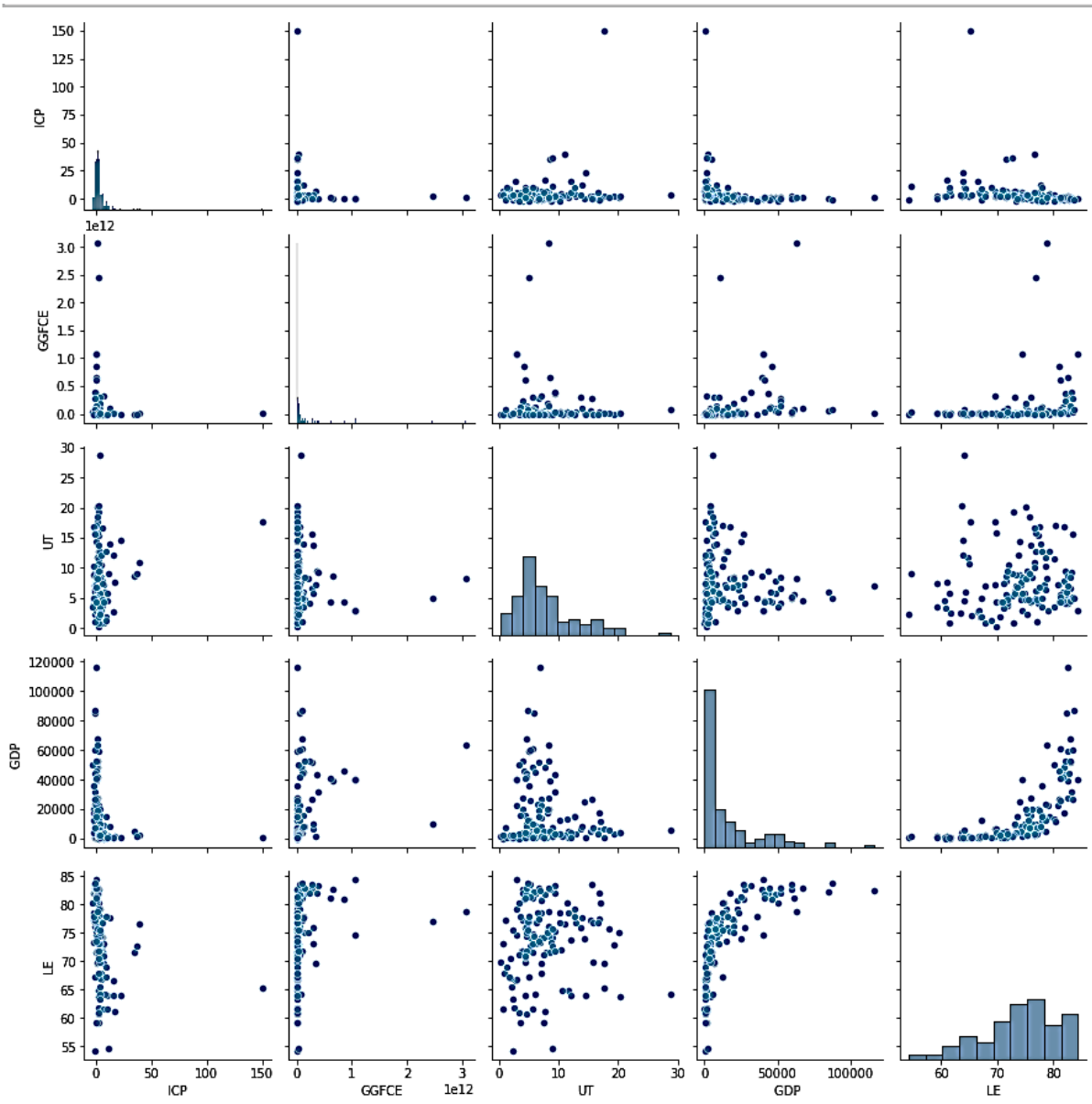


Figure B.1 - Matrix scattering matrix

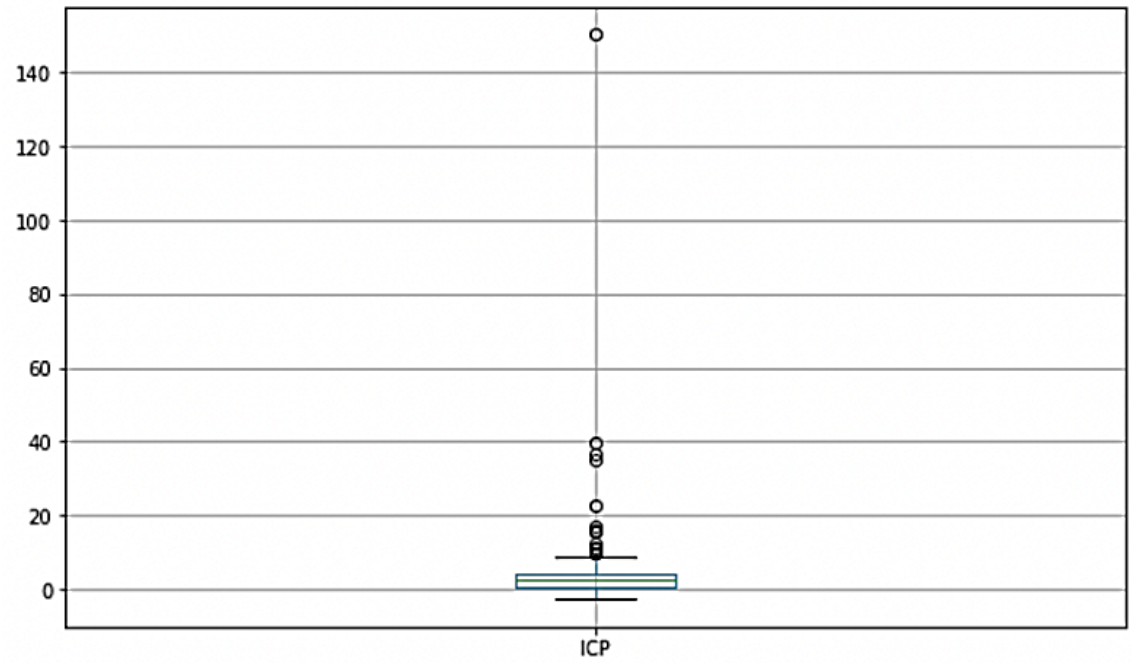


Figure B.2 - Box graph of inflation, expressed in consumer prices

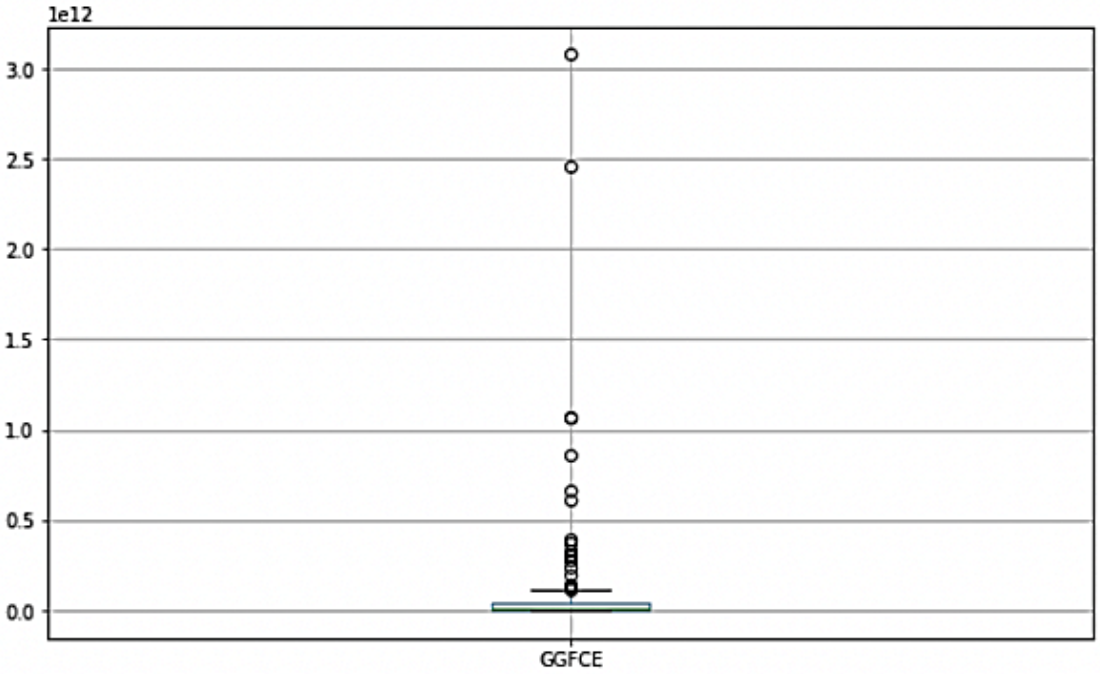


Figure B.3 - Box schedule of final consumption expenditures of the general government sector

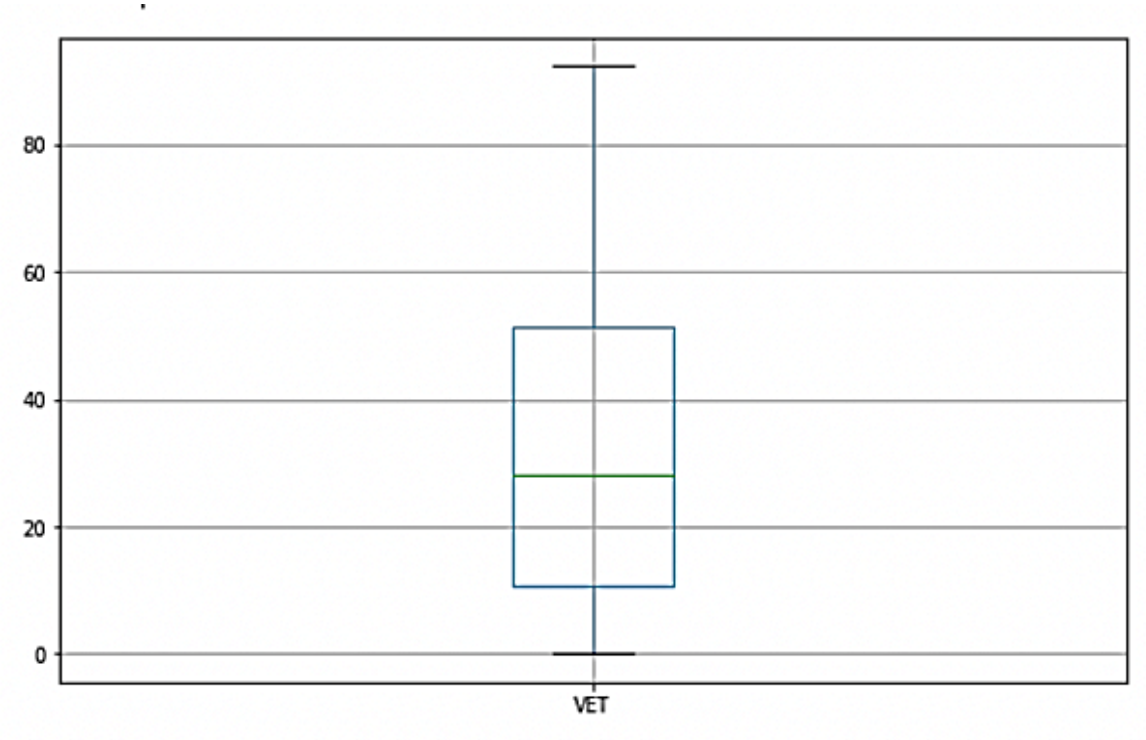


Figure B.4 - Box schedule of variable vulnerable employment

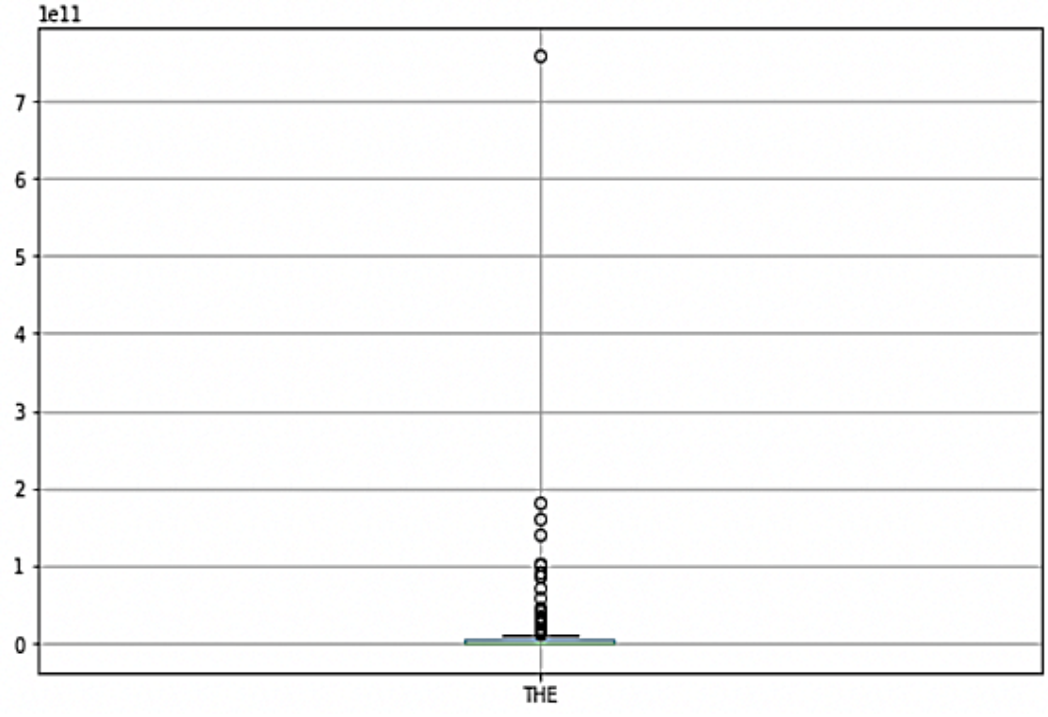


Figure B.5 - Box chart of high technology exports

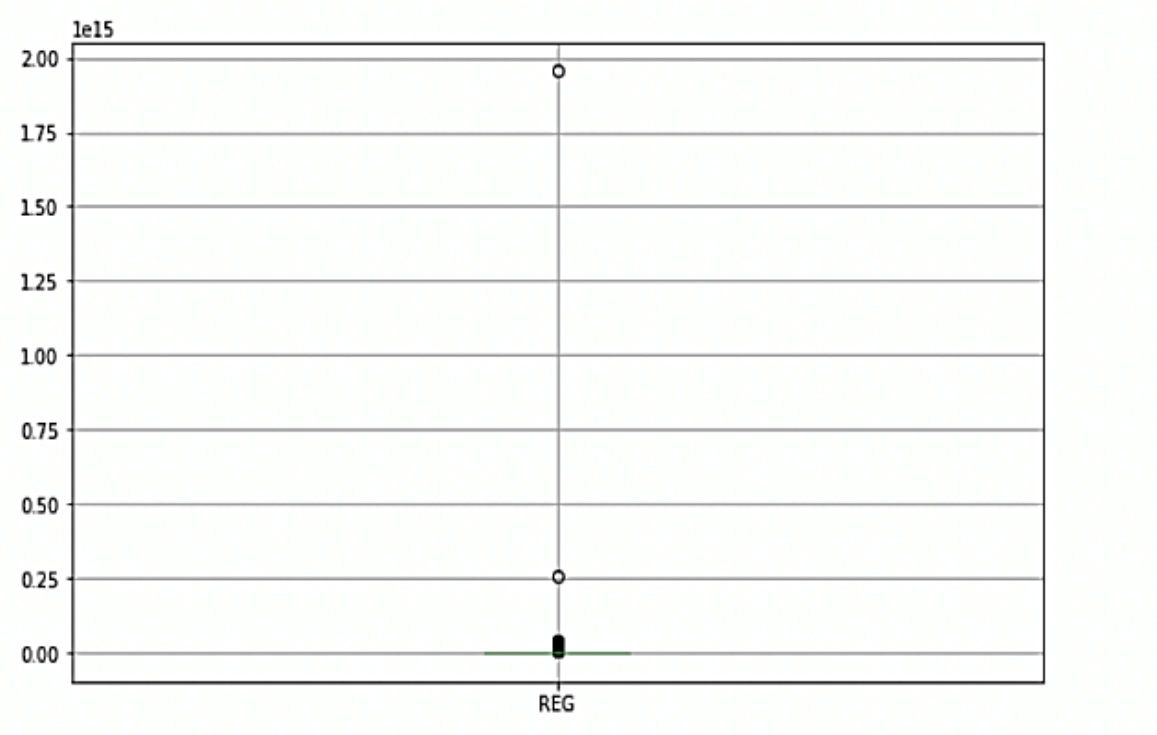


Figure B.6 - Box revenue schedule excluding grants

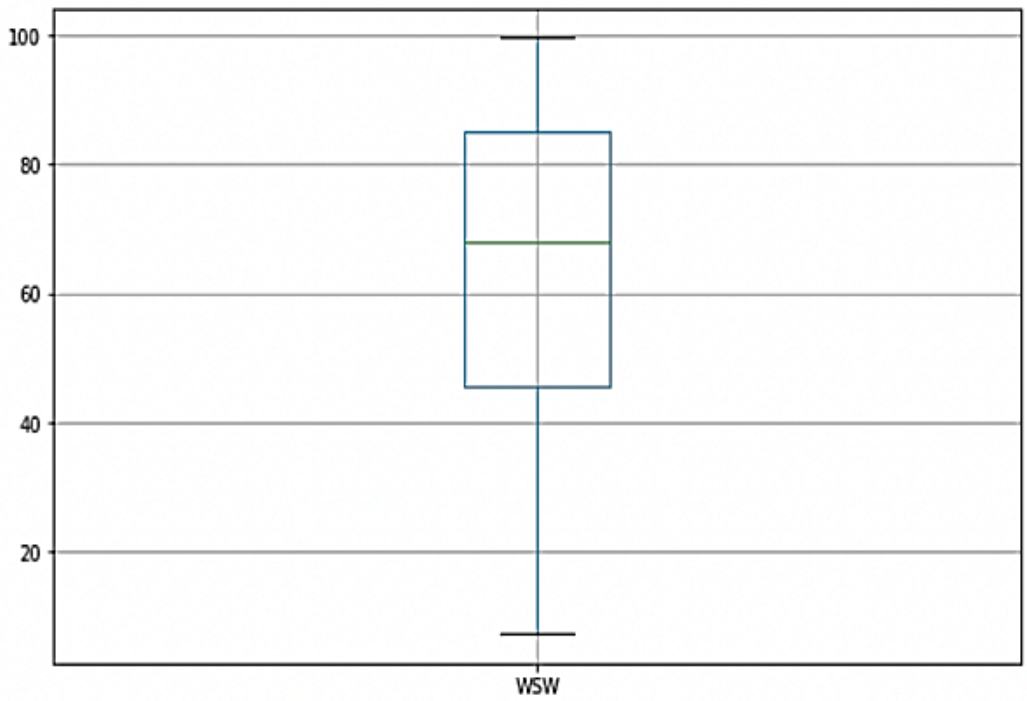


Figure B.7 - Box schedule of employees' variable

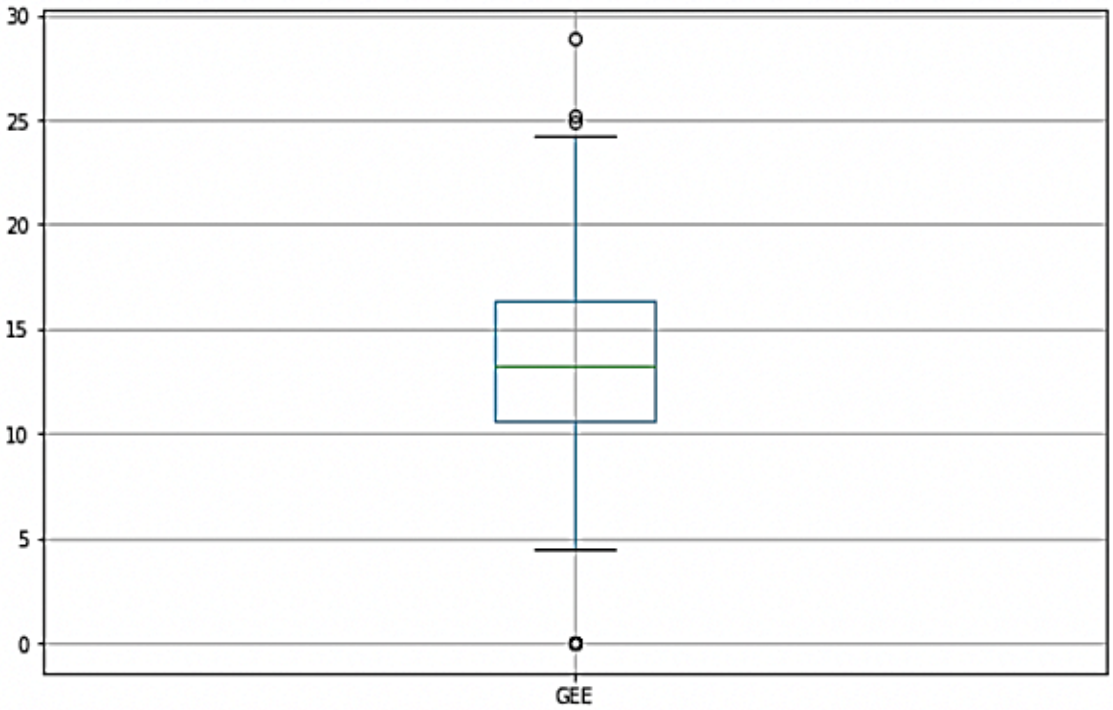


Figure B.8 - Box schedule of government spending on education

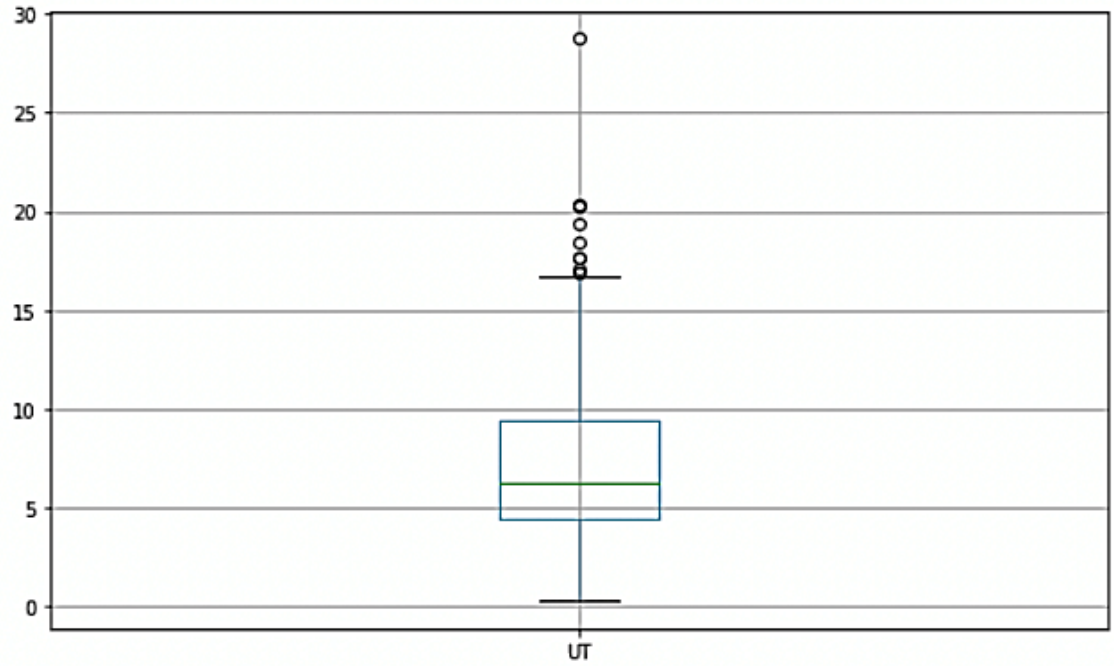


Figure B.9 - Box graph of the unemployment rate

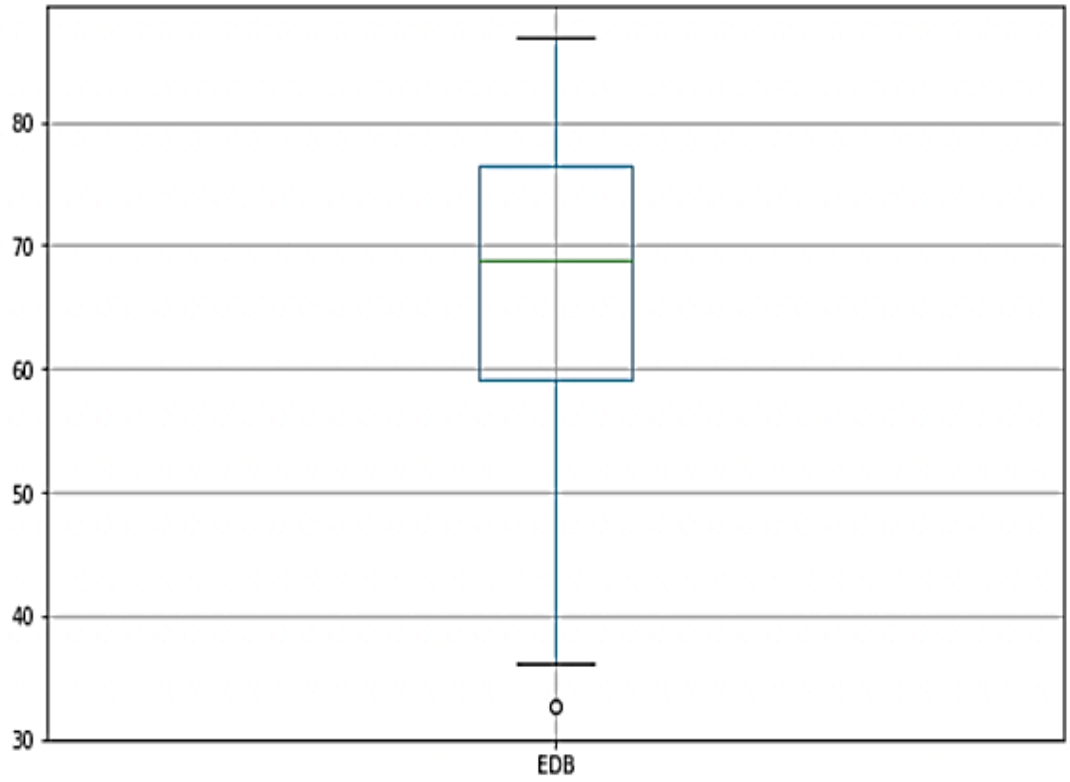


Figure B.10 - Box schedule for assessing the ease of doing business

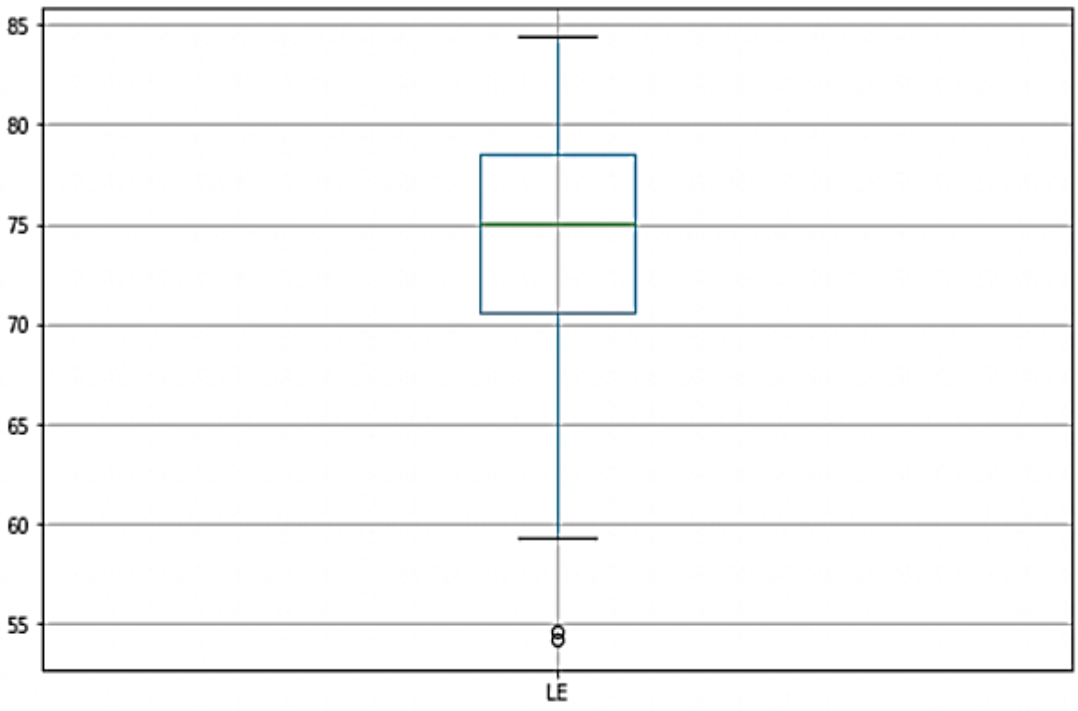


Figure B.11 - Box graph of life expectancy

Appendix C

Program code for executing the Principal Component Method

In [114]:
```
df_new = df.drop(['Country'], axis=1)
```

In [115]:
```
sc=StandardScaler()

scaler = sc.fit(df_new)
df_new_scaled = scaler.transform(df_new)
```

In [116]:
```
from sklearn.decomposition import PCA
pca_test = PCA(n_components=12)
pca_test.fit(df_new_scaled) #pca_test.fit(trainX_scaled)
sns.set(style='whitegrid')
plt.plot(np.cumsum(pca_test.explained_variance_ratio_))
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance')
plt.axvline(linewidth=4, color='r', linestyle = '--', x=6, ymin=0, ymax=1)
display(plt.show())
evr = pca_test.explained_variance_ratio_
cvr = np.cumsum(pca_test.explained_variance_ratio_)
pca_df = pd.DataFrame()
pca_df['Cumulative Variance Ratio'] = cvr
pca_df['Explained Variance Ratio'] = evr
display(pca_df.head(6))
```

In [117]:
```
pca = PCA(n_components=6)
pca.fit(df_new_scaled)
df_pca = pca.transform(df_new_scaled)
```

Appendix D

Program code for executing the Elbow Method

In [135]:
```
distortions = []
K = range(1,10)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(df_pca)
    distortions.append(kmeanModel.inertia_)
```

In [122]:
```
plt.figure(figsize=(16,8))
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```

Appendix E

Program code for executing the Nonliner Regressions

In [10]:
```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
```

In [27]:
```
X = df[['DDL']].values
y = df['GDP'].values

regr = LinearRegression()

# create quadratic features
quadratic = PolynomialFeatures(degree=2)
cubic = PolynomialFeatures(degree=3)
X_quad = quadratic.fit_transform(X)
X_cubic = cubic.fit_transform(X)

# fit features
X_fit = np.arange(X.min(), X.max(), 1)[:, np.newaxis]

regr = regr.fit(X, y)
y_lin_fit = regr.predict(X_fit)
linear_r2 = r2_score(y, regr.predict(X))
linear_MSE = mean_squared_error(y, regr.predict(X))

regr = regr.fit(X_quad, y)
y_quad_fit = regr.predict(quadratic.fit_transform(X_fit))
quadratic_r2 = r2_score(y, regr.predict(X_quad))
quadratic_MSE = mean_squared_error(y, regr.predict(X_quad))

regr = regr.fit(X_cubic, y)
y_cubic_fit = regr.predict(cubic.fit_transform(X_fit))
cubic_r2 = r2_score(y, regr.predict(X_cubic))
cubic_MSE = mean_squared_error(y, regr.predict(X_cubic))

# plot results
plt.scatter(X, y, label='training points', color='lightgray')

plt.plot(X_fit, y_lin_fit,
         label='linear (d=1), $R^2={:.2f}$'.format(linear_r2),
         color='blue',
         lw=2,
         linestyle=':')

plt.plot(X_fit, y_quad_fit,
         label='quadratic (d=2), $R^2={:.2f}$'.format(quadratic_r2),
         color='red',
         lw=2,
         linestyle='-')

plt.plot(X_fit, y_cubic_fit,
         label='cubic (d=3), $R^2={:.2f}$'.format(cubic_r2),
         color='green',
         lw=2,
         linestyle='--')

plt.xlabel('Digital Development Level [DDL]')
plt.ylabel('GDP per capita (current US$) [GDP]')
plt.legend(loc='upper right')

print("MSE_linear:",linear_MSE)
print("MSE_quadratic:",quadratic_MSE)
print("MSE_linear:",cubic_MSE)
```

```
MSE_linear: 175247854.87860048
MSE_quadratic: 96394870.12435687
MSE_linear: 86297802.7220103
```

In [29]:

```
X = df[['DDL']].values
y = df['WSW'].values

regr = LinearRegression()

# create quadratic features
quadratic = PolynomialFeatures(degree=2)
cubic = PolynomialFeatures(degree=3)
X_quad = quadratic.fit_transform(X)
X_cubic = cubic.fit_transform(X)

# fit features
X_fit = np.arange(X.min(), X.max(), 1)[:, np.newaxis]

regr = regr.fit(X, y)
y_lin_fit = regr.predict(X_fit)
linear_r2 = r2_score(y, regr.predict(X))
linear_MSE = mean_squared_error(y, regr.predict(X))

regr = regr.fit(X_quad, y)
y_quad_fit = regr.predict(quadratic.fit_transform(X_fit))
quadratic_r2 = r2_score(y, regr.predict(X_quad))
quadratic_MSE = mean_squared_error(y, regr.predict(X_quad))

regr = regr.fit(X_cubic, y)
y_cubic_fit = regr.predict(cubic.fit_transform(X_fit))
cubic_r2 = r2_score(y, regr.predict(X_cubic))
cubic_MSE = mean_squared_error(y, regr.predict(X_cubic))

# plot results
plt.scatter(X, y, label='training points', color='lightgray')

plt.plot(X_fit, y_lin_fit,
         label='linear (d=1), $R^2={:.2f}$'.format(linear_r2),
         color='blue',
         lw=2,
         linestyle=':')

plt.plot(X_fit, y_quad_fit,
         label='quadratic (d=2), $R^2={:.2f}$'.format(quadratic_r2),
         color='red',
         lw=2,
         linestyle='-')

plt.plot(X_fit, y_cubic_fit,
         label='cubic (d=3), $R^2={:.2f}$'.format(cubic_r2),
         color='green',
         lw=2,
         linestyle='--')

plt.xlabel('Digital Development Level [DDL]')
plt.ylabel('Wage and salaried workers, total (% of total employment) [WSW]')
plt.legend(loc='upper right')
print("MSE_linear:",linear_MSE)
print("MSE_quadratic:",quadratic_MSE)
print("MSE_linear:",cubic_MSE)
```

```
MSE_linear: 159.32041869547942
MSE_quadratic: 146.68129107840934
MSE_linear: 140.01227528070402
```

In [28]:

```
X = df[['DDL']].values
y = df['VET'].values

regr = LinearRegression()

# create quadratic features
quadratic = PolynomialFeatures(degree=2)
cubic = PolynomialFeatures(degree=3)
X_quad = quadratic.fit_transform(X)
X_cubic = cubic.fit_transform(X)

# fit features
X_fit = np.arange(X.min(), X.max(), 1)[:, np.newaxis]

regr = regr.fit(X, y)
y_lin_fit = regr.predict(X_fit)
linear_r2 = r2_score(y, regr.predict(X))
linear_MSE = mean_squared_error(y, regr.predict(X))

regr = regr.fit(X_quad, y)
y_quad_fit = regr.predict(quadratic.fit_transform(X_fit))
quadratic_r2 = r2_score(y, regr.predict(X_quad))
quadratic_MSE = mean_squared_error(y, regr.predict(X_quad))

regr = regr.fit(X_cubic, y)
y_cubic_fit = regr.predict(cubic.fit_transform(X_fit))
cubic_r2 = r2_score(y, regr.predict(X_cubic))
cubic_MSE = mean_squared_error(y, regr.predict(X_cubic))

# plot results
plt.scatter(X, y, label='training points', color='lightgray')

plt.plot(X_fit, y_lin_fit,
         label='linear (d=1), $R^2={:.2f}$'.format(linear_r2),
         color='blue',
         lw=2,
         linestyle=':')

plt.plot(X_fit, y_quad_fit,
         label='quadratic (d=2), $R^2={:.2f}$'.format(quadratic_r2),
         color='red',
         lw=2,
         linestyle='-')

plt.plot(X_fit, y_cubic_fit,
         label='cubic (d=3), $R^2={:.2f}$'.format(cubic_r2),
         color='green',
         lw=2,
         linestyle='--')

plt.xlabel('Digital Development Level [DDL]')
plt.ylabel('Vulnerable employment, total (% of total employment) (modeled ILO estimate) [VET]')
plt.legend(loc='upper right')
print("MSE_linear:",linear_MSE)
print("MSE_quadratic:",quadratic_MSE)
print("MSE_linear:",cubic_MSE)
```

```
MSE_linear: 163.3689660536247
MSE_quadratic: 148.83145090812062
MSE_linear: 142.03192039001053
```

In [30]:

```
X = df[['DDL']].values
y = df['EDB'].values

regr = LinearRegression()

# create quadratic features
quadratic = PolynomialFeatures(degree=2)
cubic = PolynomialFeatures(degree=3)
X_quad = quadratic.fit_transform(X)
X_cubic = cubic.fit_transform(X)

# fit features
X_fit = np.arange(X.min(), X.max(), 1)[:, np.newaxis]

regr = regr.fit(X, y)
y_lin_fit = regr.predict(X_fit)
linear_r2 = r2_score(y, regr.predict(X))
linear_MSE = mean_squared_error(y, regr.predict(X))

regr = regr.fit(X_quad, y)
y_quad_fit = regr.predict(quadratic.fit_transform(X_fit))
quadratic_r2 = r2_score(y, regr.predict(X_quad))
quadratic_MSE = mean_squared_error(y, regr.predict(X_quad))

regr = regr.fit(X_cubic, y)
y_cubic_fit = regr.predict(cubic.fit_transform(X_fit))
cubic_r2 = r2_score(y, regr.predict(X_cubic))
cubic_MSE = mean_squared_error(y, regr.predict(X_cubic))

# plot results
plt.scatter(X, y, label='training points', color='lightgray')

plt.plot(X_fit, y_lin_fit,
         label='linear (d=1), $R^2={:.2f}$'.format(linear_r2),
         color='blue',
         lw=2,
         linestyle=':')

plt.plot(X_fit, y_quad_fit,
         label='quadratic (d=2), $R^2={:.2f}$'.format(quadratic_r2),
         color='red',
         lw=2,
         linestyle='-')

plt.plot(X_fit, y_cubic_fit,
         label='cubic (d=3), $R^2={:.2f}$'.format(cubic_r2),
         color='green',
         lw=2,
         linestyle='--')

plt.xlabel('Digital Development Level [DDL]')
plt.ylabel('Ease of doing business score [EDB]')
plt.legend(loc='upper right')
print("MSE_linear:",linear_MSE)
print("MSE_quadratic:",quadratic_MSE)
print("MSE_linear:",cubic_MSE)
```

```
MSE_linear: 53.310075531304015
MSE_quadratic: 51.72855079910013
MSE_linear: 51.71504631488099
```

In [31]:

```
X = df[['DDL']].values
y = df['LE'].values

regr = LinearRegression()

# create quadratic features
quadratic = PolynomialFeatures(degree=2)
cubic = PolynomialFeatures(degree=3)
X_quad = quadratic.fit_transform(X)
X_cubic = cubic.fit_transform(X)

# fit features
X_fit = np.arange(X.min(), X.max(), 1)[:, np.newaxis]

regr = regr.fit(X, y)
y_lin_fit = regr.predict(X_fit)
linear_r2 = r2_score(y, regr.predict(X))
linear_MSE = mean_squared_error(y, regr.predict(X))

regr = regr.fit(X_quad, y)
y_quad_fit = regr.predict(quadratic.fit_transform(X_fit))
quadratic_r2 = r2_score(y, regr.predict(X_quad))
quadratic_MSE = mean_squared_error(y, regr.predict(X_quad))

regr = regr.fit(X_cubic, y)
y_cubic_fit = regr.predict(cubic.fit_transform(X_fit))
cubic_r2 = r2_score(y, regr.predict(X_cubic))
cubic_MSE = mean_squared_error(y, regr.predict(X_cubic))

# plot results
plt.scatter(X, y, label='training points', color='lightgray')

plt.plot(X_fit, y_lin_fit,
         label='linear (d=1), $R^2={:.2f}$'.format(linear_r2),
         color='blue',
         lw=2,
         linestyle=':')

plt.plot(X_fit, y_quad_fit,
         label='quadratic (d=2), $R^2={:.2f}$'.format(quadratic_r2),
         color='red',
         lw=2,
         linestyle='-')

plt.plot(X_fit, y_cubic_fit,
         label='cubic (d=3), $R^2={:.2f}$'.format(cubic_r2),
         color='green',
         lw=2,
         linestyle='--')

plt.xlabel('Digital Development Level [DDL]')
plt.ylabel('Life expectancy at birth, total (years) [LE]')
plt.legend(loc='upper right')
print("MSE_linear:",linear_MSE)
print("MSE_quadratic:",quadratic_MSE)
print("MSE_linear:",cubic_MSE)
```

```
MSE_linear: 11.25739080102837
MSE_quadratic: 10.95476398974836
MSE_linear: 10.93764885326559
```

Appendix F

Program code for executing the Liner Regressions

In [33]:
```
X = df.drop(['DDL','Country','Cluster'], axis=1)
y = df['DDL']
```

In [34]:
```
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
predictions = model.predict(X)

print_model = model.summary()
print(print_model)
```

In [35]:
```
X2 = df.drop(['DDL','Country', 'GGFCE','UT','ICP','REG','THE','GEE','WSW','Cluster'], axis=1)
y = df['DDL']
```

In [36]:
```
X2 = sm.add_constant(X2)
model2 = sm.OLS(y, X2).fit()
predictions = model2.predict(X2)

print_model2 = model2.summary()
print(print_model2)
```