Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics

Statistical analysis of COVID-19 development in the Czech Republic and selected countries

Linda Krčmářová

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

Linda Krčmářová

Economics and Management
Economics and Management

Thesis title

Statistical analysis of COVID-19 development in the Czech Republic and selected countries

---

Objectives of thesis

The main objective of this thesis is to assess the development of COVID-19 in selected countries.

Methodology

The methodology will be based on data acquisition, study and examination of the data. There will be used cluster analysis to obtain countries with similar characteristics and then for the cluster with the Czech Republic will be analysed the development of COVID-19 measures.

The proposed extent of the thesis

30 – 40 pages

Keywords

Big Data, Data mining, Predictive analysis

Recommended information sources

ABBOTT, D. Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst. Indiana: John Wiley & Sons, Incorporated, 2014. ISBN 9781118727935

DHAENENS, C. – JOURDAN, L. Metaheuristics for Big Data. John Wiley & Sons, Incorporated, 2016. ISBN 9781119347583

HINDLS, R. *Statistika pro ekonomy.* Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

KOTU, V. – DESHPANDE, B. Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner. Elsevier Science & Technology, 2014. ISBN 9780128016503

LAROSE, D. T. – LAROSE, C. D. Data Mining and Predictive Analytics. John Wiley & Sons, Incorporated, 2015. ISBN 9781118868676

LARSON, R. – FARBER, E. *Elementary statistics : picturing the world.* Boston: Pearson Prentice Hall, 2015. ISBN 9780321693624.

SIEGEL, E. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. John Wiley & Sons, Incorporated, 2016. ISBN 9781119153658

Expected date of thesis defence

2020/21 SS – FEM

The Bachelor Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 20. 1. 2021

prof. Ing. Libuše Svatošová, CSc.

Head of department

Electronic approval: 22. 1. 2021

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 25. 01. 2021

Declaration

I declare that I have worked on my bachelor thesis titled " Statistical analysis of COVID-19 development in the Czech Republic and selected countries" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break copyrights of any their person.

In Prague on 11. 3. 2021 _____

Acknowledgement

I would like to thank the supervisor of this thesis Ing. Tomáš Hlavsa, Ph. D. for his patient assistance and constructive advice in the making of this thesis.

Statistical analysis of COVID-19 development in the Czech Republic and selected countries

Abstract

This Bachelor thesis is focused on the analysis of the development of COVID-19 in the Czech Republic and other selected countries. The Literature Review concerns mainly the explanation of Big data and with it connected techniques, such as Processing and Analysis of Big data as well as Predictive analysis. Lastly, COVID-19 is introduced and its development in the world is explained. Measures taken by the governments and other factors influencing the spread and course of the disease are also defined in the theoretical part and later are analyzed in the practical part of the thesis. The second part of this paper is concerned with own research. Firstly, the hierarchical clustering analysis is performed in order to create a group of countries with similar characteristics as Czech Republic. Then, data from those countries about the spread and development of COVID-19 are analyzed and the fluctuations are compared with the established measures and other factors influencing the course of the pandemic in the specific territory. As the most influential elements were identified the population density and the number of already occurring cases in the specified country.

Keywords: Big Data, Hierarchical clustering, data analysis, COVID-19, pandemics

Statistická analýza vývoje COVID-19 v České republice a dalších vybraných zemích

Abstrakt

Bakalářská práce je zaměřena na analýzu vývoje nemoci COVID-19 v České republice a dalších vybraných zemích. Teoretická část se zaměřuje především na vysvětlení pojmu Big data a s tím spojených technik, jako je zpracování a analýza Big Data, ale také prediktivní modelování. V závěru první části práce je představen COVID-19 a je znázorněn jeho vývoj ve světě. Tato kapitola se také zabývá vysvětlením zavedených nařízení a dalších faktorů, které mají vliv na šíření a průběh této nemoci. Tyto faktory jsou později analyzovány v praktické části. Druhá část této práce je zaměřena na vlastní výzkum. Nejprve bylo provedeno hierarchické shlukování, s cílem získat skupinu zemí z podobnými charakteristikami, jako má Česká republika. Data o šíření a vývoji COVID-19 ve vybraných zemích byla poté analyzována. Byly zde srovnávány především výkyvy v počtu případů a úmrtí v jednotlivých zemích, s nařízeními a dalšími okolnostmi majícími vliv na vývoj pandemie ve vybrané oblasti. Jako nejvlivnější faktory byly identifikovány hustota obyvatelstva a počet již vyskytujících se případů v konkrétní zemi.

Klíčová slova: Big Data, Hierarchické shlukování, analýza dat, COVID-19, pandemie

Table of contents

List of pictures

1 Introduction

This thesis aims to describe and analyze the development of the novel coronavirus, COVID-19, in the Czech Republic and other countries with similar characteristics. For the exploration of the development of the disease, Big data is used and therefore, the theoretical part deals with the main concepts of Big data and connected subjects such as Data mining, Predictive analytics, Exploratory data analysis, and others.

Data is viewed as one of the most precious commodities of the modern world. Owning data means having power. Companies and businesses are using data about their customers to predict future sales, banks and insurance companies are using this information to determine whether or not it will be profitable to invest in their client, or as in our case, it can be used to assess the usefulness of the measures implemented by the government and even further, to predict future development of the observed variable.

The outbreak of the new pandemic, SARS-CoV-2, took the world by surprise, and it created a wave of shock and panic in the early months of 2020. As the disease spread the countries implemented certain measures in order to stop or at least reduce the transmission in their region. Besides restrictions established by the governments, there are other factors that influence the development of the situation in a specific country. Population density, number of performed tests, or, as was recently discovered, the "tightness" of the culture and social norms all have impact on the number of cases which emerge in each country.

The group of countries which were to be analyzed was determined by a Hierarchical cluster analysis based on 5 factors – Government Effectiveness, Human Development Index, Demographic Structure, Consumer Price Index and the proportion of Import and Export to the total production of the specified country. Those attributes aimed to classify the countries from multiple angles. Some of the selected indexes evaluate the countries from the economical point of view whilst others look more closely on the social perspective and the quality of life of the citizens. This broad variety of indexes was purposefully chosen in order to group the countries which are as similar as possible as a whole, not only in one aspect. Because of the similarity in multiple attributes, it can be expected that the development of the pandemic was comparable.

The practical part compares the situation in the selected countries and aims to explain the diversity both with differences in measures and other factors which are diverse among

the countries. Both the restrictions implemented, and the additional aspects will be observed and analyzed and additionally, the correlations between the individual factors and the development of the spread in the specific country will be evaluated in order to assess the importance of individual elements.

2 Objectives and Methodology

2.1 Objectives

The main objective of this thesis is to assess the development of COVID-19 in multiple countries. The selection of the countries is based on the Hierarchical Cluster Analysis in order to obtain a sample of states with similar characteristics. As the sample for the final analysis was chosen the cluster containing Czech Republic. The main goal is to examine the differences in the development in the chosen countries by analyzing the relationship between the determinants of the rate of spread (such as number of new cases, mortality rate) and factors and measures affecting it.

2.2 Methodology

2.2.1 Exploratory data analysis

When computing any statistical analysis, it is essential to start with exploratory data analysis. It is the "approach to data analysis where the features and characteristics of the data are reviewed… without attempting to apply any particular model to the data" (SpringerLink, 2020). We can divide the measures into two main categories, measures of central tendency and measures of dispersion and skewness. Among the main representatives of measures of central tendency (location) belongs the arithmetic and geometric mean, median, mode and five number summary. To the dispersion measures belongs the range, variance, standard deviation, and others. (SEEMON, 2014)

2.2.2 (Single) Linear regression

As defined in a book *Introduction to Linear Regression Analysis*, Simple Linear Regression is a model with a single regressor *x* that has a relationship with a response *y* that is a straight line. (MONTGOMERY et. al., 2012) In other words, it is a model which aims to describe the changes in the dependent variable (*y*) by the changes in the independent or so-called explanatory variable (*x*). The simplest formula is as follows:

$$y = f(x)$$

or

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where there are two unknown constants, β_0 and β_1 (the slope) and ε is a random error component. (MONTGOMERY et. al., 2012)

However, this thesis is mainly concerned with Coefficient of correlation (r) and Coefficient of determination (r^2). r describes the linear relationship between the two variables in the model. Its values are $-1 \leq r \leq 1$ where if the number is negative there is a inverse relationship between the variables, meaning that if *y* increases, *x* decreases. If the coefficient of correlation has a positive value, the relationship between the variables is positive, therefore, when the value of one increases the value of the other rises too. Additionally, the absolute value of the coefficient determines the strength of the relationship between the variables. The higher the absolute value of *r* the stronger is the relationship. The coefficient of determination states the proportion of variation in *y* (the dependent variable) which is explained by the variation in the regressor *x*. It is important to note that $0 \leq r^2 \leq 1$ and the higher it is, the more variability is explained. (DUPUIS et. al., 2002) (MONTGOMERY et. al., 2012)

2.2.3 **Cluster analysis**

Cluster Analysis is a group of methods which aim to find structures in data and arrange them into homogenous groups (clusters). Data with similar characteristics is put in the same cluster and therefore this analysis divides a population (sample) into multiple samples with common aspects. There are multiple approaches to cluster analysis which use various methods and processes to divide the population. This thesis focuses on the Hierarchical analysis, more specifically the use of Ward clustering method with Euclidean distance as dissimilarity measure.

2.2.3.1 Hierarchical clustering

Hierarchical clustering is a type of clustering where in the beginning each object forms its own cluster (containing only itself). In a successive order, the closest clusters are merged until they all form one group which encloses all of the objects. (GOVAERT et. al., 2013)

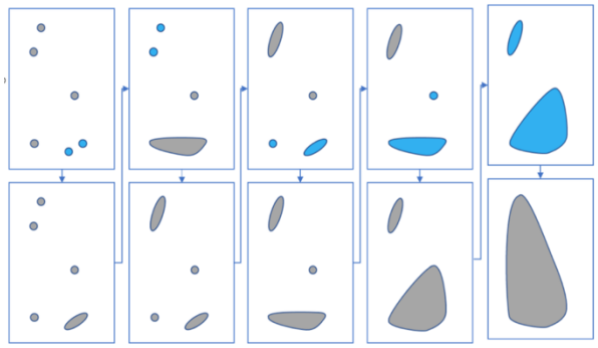

Figure 1 - Hierarchical clustering (BOCK, 2021)

A common output of a hierarchical cluster analysis is a *dendrogram* (see Figure 2) where the closest objects (the ones which are connected at first) are linked with the shortest node (in this case *E* and *F),* the further the distance between the variables is, the longer is the node in the dendrogram. (BOCK, 2021)

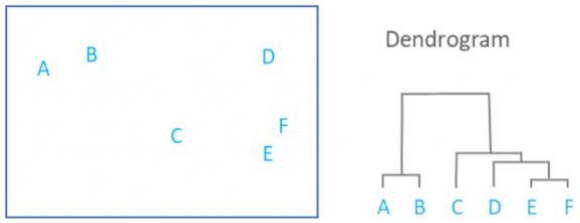

Figure 2 - Hierarchical clustering - Dendrogram (BOCK, 2021)

2.2.3.2 Ward's minimum variance method

Ward's minimum variance method is a hierarchical clustering model which aims to create clusters while minimizing the variance within the individual clusters. As all hierarchical clustering models, it starts with *n* clusters where each object creates its own cluster. When making a new cluster it always has to consider all the possible clusters and choose the one that creates minimal intra-cluster variance. (PLOS ONE, 2017) (HU et. al., 2018)

3 Literature Review

3.1 Big Data

3.1.1 Definition

As explained by J. S. Ward and A. Baker in their paper called *Undefined by Data: A Survey of Big Data Definitions:"* "Big" implies significance, complexity and challenge. Unfortunately, the term "big" also invites quantification and therein lies the difficulty in furnishing a definition*."* (WARD et. al., 2013) We can see that it is not easy to come up with a simple explanation which depicts all characteristics of Big data. One of the possible interpretations can be that the term *Big data* refers to an extensive amount of data which is diversified and come from multiple different sources. The origin of data collected mainly depends on the industry. Data can be gathered using sensors, social media sites, cellphone GPS signals, transactions records, etc. (DHAENENS et. al., 2016)

Big data on its own is not of a great value. Their true worth comes up when they are analyzed and patterns between variables are discovered. Process which searches for these patterns and makes the data valuable is called *Data mining* and will be explained later.

3.1.1.1 The "Three Vs"

As the main characteristics of Big data, three main dimensions can be established:

Data Volume: Can be explained as the size or the amount of data. (DHAENENS et. al., 2016) (LANEY, 2001)

Data Velocity: Is the rapid rate at which data is generated and the related pace with which the analysis should take place. (GANDOMI et. al., 2015) (LANEY, 2001)

Data Variety: Showcases that data is not collected from a single source, there are millions of sources which use different techniques and technologies to gather data.

(DHAENENS et. al., 2016) (LANEY, 2001)

According to many sources except the classic “Three Vs” explained by D. Laney in his 2001 work, we can take into consideration three more dimensions:

Veracity: This feature shows unreliability of some sources of data, such as decisions made by customer because of sentimental reasons - it involves human judgement which doesn’t have to correspond with predicted behavior.

Variability (and complexity): The first of these components, variability, takes into account the inconsistency of data, especially seasonality (periodic peaks and throughs). Complexity involves the challenge of connecting data from different sources to eliminate redundancy and cleaning them.

Value: As mentioned before, raw unprocessed data is relatively, to its size, worthless but when processed and analyzed its value grows exponentially. (GANDOMI et. al., 2015)

3.1.2 Data mining

Refers to the utilization of data in order to find patterns, which can be further used, in it. Sometimes it is also interpreted as knowledge discovery, machine learning, and predictive analytics. The second of those interpretations is closely related to database systems, visualization, exploratory data analysis, etc. (KOTU et. al., 2014)

Enormous amount of data is daily collected but without Data mining they would be just stored without any use. Data mining is a set of techniques which enable the observer to take gathered data and by using algorithms find patterns and describe how one or more variables are related to (or affect) other variables. It takes incomprehensible data and transforms them into knowledge (information), which can be used by businesses, governments, and others. (LAROSE et. al., 2015), (KOTU et. al., 2014)

3.1.3 Data storage and management

Companies usually use one or more relational management systems for storing data which allows them to keep track of what data is stored and where. These systems are usually not adapted for storage of Big data, so they require technologies which are able to deal with scalability, variety of data, velocity of data, etc. This trend shifts companies toward using

open-source alternatives which support all of the characteristics of data. The transportation of big data has to be also included in its management so data can travel from its source to data centers (databases). (DHAENENS et. al., 2016)

One of the other types of data storing is ***Data warehousing*** which is able to collect, store and manage data while using analysis to obtain knowledge. Data in warehouses is structured explicitly for query and analysis. This concept was introduced in 1980s and due to its ability to make predictions, create trends and analyze historical data it became highly popular. (The Pathologies of Big Data, 2009) (KRISHNAN, 2013)

3.2 Processing data

3.2.1 Sources and types of data

There are innumerable sources from which data can be acquired such as Monitoring sensors, Social media sites, Online digital pictures, Transaction records, Application logs, Call data records, etc. (ACHARI, 2015) Each source type requires specific processing of data before the data can be used for analysis because of particular data type. We can divide data types into three main categories which ought to be handled differently - *Text analytics, Audio analytics, Video analytics.*

3.2.1.1 Text analytics

Text analytics (text mining) is a technique that is used for data extraction from written data. It can take its inputs from many sources such as social networks, news, documents, etc. Among the main methods for text analytics belong: *Information extraction, Text summarization, Question answering and Sentiment analysis (opinion mining).* (GANDOMI et. al., 2015)

3.2.1.2 Audio analytics

Audio analytics refers to a method which extracts data from audio or human speech *(speech analysis).* This technique is mainly implemented in call centers where thousands of hours of calls from customers are analyzed with the aim of improving the customer service, performance of workers and, therefore, to increase profits. (GANDOMI et. al., 2015)

3.2.1.3 Video analytics

In comparison to other data collection techniques, video analysis is the one that is the least developed. (PANIGRAHI et. al., 2010) There are multiple techniques for both real-time and pre-recorded videos. Problem which arises with analysis of videos is the size of the data files which is considerably larger in comparison to other types of data. (GANDOMI et. al., 2015)

3.2.2 **Erroneous and missing values**

Before the data can be used for analytics it has to be tested for any errors or outliers which can negatively influence the results of the analysis. It is important to detect those values and exclude them from the system to acquire accurate test results.

An ***Outlier*** can be defined as "an observation that lies outside the overall pattern of a distribution" (MOORE et. al., 1998) This definition can be furtherly developed as "a point which falls more than 1,5 times the interquartile range above the third quartile or below the first quartile" (MathWorld. A Wolfram Web Resource, 2020). An outlier can be both a result of an error and a genuine value that is against the trend of the remaining data. In both cases it needs to be eliminated.

Erroneous values can be a result of data entry error, human caused error, etc. Invalid values have an impact on the quality and accuracy of the model. To purify the model from these values, organizations have to use data cleansing practices and transformation techniques. These processes enhance the quality of the dataset. (KOTU et. al., 2014)

Missing values create analogical problem as erroneous values and outliers, it decreases the quality and reliability of a dataset and the results of test. One of the most common solutions of how to deal with missing values is the basic omission of the records with missing values. However, this technique can lead to omitting values which are crucial, or which create important patterns or connections. As explained in *Data mining and Predictive analysis* by D. Larose and C. Larose, there are four approaches suggesting how the solution:

1. Replace the missing value with a constant, specified by the analyst.
2. Replace the missing value with the field mean (for numeric variables) or the mode (for categorical variables).

3. Replace the missing values with a value generated at random from observed distribution of the variable.
4. Replace the missing values with imputed values based on other characteristics of the record.

(LAROSE et. al., 2015)

3.2.3 Process of Big Data

On its own, data used for analysis is without any value. But their importance emerges when they are put into context and are analyzed. Businesses need processes with the highest efficiency possible to transform immense amounts of data into organized observations with meaning from which they can derive conclusion beneficial for their organization.

Processing of Big data can be divided in two main categories – Data Management and Analytics of Data. The process of data management has three steps. First is the Data Acquisition and Recording which simply means collecting data, and its compression and filtration. Second step is mainly concerned with extraction of the required information from the data and its cleaning. In this phase we also have to deal with erroneous and missing values which are likely to affect our analysis and, later on, predictions. After this step, data should be in a suited structure for analysis. (LABRINIDIS et. al., 2012) Last phase of the Data management is Integration, Aggregation and Representation of data.

Second part of the processing of Big data is Analytics, where we can find two stages, Modelling and Analysis, and Interpretation, which is the final step necessary to obtain useful information from data. Without correct interpretation data can be translated into erroneous conclusion. (GANDOMI et. al., 2015)

3.3 Predictive analytics

3.3.1 Definition and characteristics

Predictive analytics can be defined as “Technology that learns from experience (data) to predict the future behavior of individuals in order to derive better decisions.” (SIEGEL, 2016) It is a field of study of its own and it’s used in everyday life to make billions of decisions. In comparison with forecasting which makes aggregate predictions of the future,

Predictive analysis tells us the individual values (values for individual inputs/outputs). (SIEGEL, 2016)

Predictive analytics uses automatized algorithms to search for patterns or relationships in data. We can distinguish between two main types of algorithms used for prediction - *Supervised learning* and *Unsupervised learning.* (ABOTT, 2014)

3.3.1.1 Supervised learning

Sometimes also called *Predictive modeling*, uses the target variable as the *supervisor*. The target variable is the parameter which is unknown and is needed to be observed and whose values are unknown, and we want to predict them.

There are two fundamental algorithms which are used for analysis.

1) *Classification* - used for categorical target variables
2) *Regression* - used for continuous target variables (ABOTT, 2014)

3.3.1.2 Unsupervised learning

Unsupervised learning, or so-called *Descriptive modelling* or *Segmentation*, has no target variable. The data inputs are clustered or grouped and are given a label to specify to which group the record belongs. (ABOTT, 2014)

Furtherly we can divide algorithms based on the knowledge of distribution in the data. *Parametric models* expect known distributions in the data (not necessarily normal distribution) and it attempts to find linear relationships among the variables. *Non-parametric* models are usually machine-learning algorithms which are not expected to assume distributions. These models are more flexible and less time consuming in comparison with Parametric models. Usually, the *Parametric models* are assumed to be more reliable because extensive properties of data are known which increases the probability of finding the optimal solution. (ABOTT, 2014)

3.4 Utilization in a specific field

The utilization and application of Big data creates new possibilities for analyzing data in countless scientific fields. The practical part of this paper focuses on its usage in assessing the effectiveness of measures implemented by the governments in order to diminish the spread of a global pandemic – COVID-19 and also on the evaluation of other factors influencing the development of the disease in the specific countries.

When a government issues a new law, or as in our case a preventive measure, it is crucial to analyze its impact and whether the implementation produced the desired effect. Big data analysis serves as a mean to check and compare current values of the observed phenomenon with figures from the period before the legislation changed. If the development of the data signifies the desired outcome it indicates the efficiency of the measure. Furtherly, predictive analysis can be used to forecast the progress of the situation and the possible effects of the potential regulations. Nevertheless, it is crucial to take into consideration any other aspects which may have affected the situation both before and after the implementation of the law.

3.5 COVID-19

SARS-CoV-2, COVID-19 or so called "coronavirus" is a novel type of coronavirus which emerged in Wuhan, China at the end of the year 2019. Among its main symptoms belongs cough, fever, fatigue, headache, loss of taste and/or smell, and other. The virus spreads between people in close contact via small respiratory droplets in aerosol which can be sprayed when a person talks, coughs, sneezes, etc. Due to the ease of its spread, it has transmitted to the whole world and on March 11, 2020 the World Health Organization characterized COVID-19 as a pandemic. Many countries went into lockdown in order to control the spread of the virus in their territory. Governments also implemented other measures such as an obligation to wear a facemask in public, prohibition of gatherings and social events, closing of schools, shops, and also, in most cases, closing their borders. (World Health Organization, 2020)

3.5.1 **Data**

Data used for the analysis come from a publicly available dataset by *Our World in Data* (Our World in Data, 2021) which is daily updated. It provides data on new cases and deaths (both absolute values, as well as per million), reproduction rate, number of administered tests, etc. Additionally, it provides information on number of Intensive Care Unit (ICU) patients and hospital admissions, number of administered vaccines, current positive rate, and others. Values used for our analysis are from January 2021.

3.5.2 **Development of COVID-19**

Before analyzing the sample data, it is crucial to get a global perspective by looking on the development of the pandemic in the world. As stated before, the novel type of coronavirus emerged in Wuhan, China at the end of the year 2019. *Figure 3* depicts the situation in the whole world in the span of one year, starting in January 2020 and ending in January 2021. We can see that the first major increase occurred in March 2020, but the governments fairly quickly contained the spread so that the number of new cases remained constant at a certain level and it stayed this way for 2 months. During the summertime there was a continuous slow rise and at the end of September 2020 came the second wave of new cases. The wave pattern was seen in other pandemics and specialists were therefore suspecting that COVID-19 would show the same behavior. The major factor influencing the significant increases in the observed cases is the human factor – mainly the behavior of the citizens. In the period when the second wave occurred ended the summer holidays and people again started to go to work and school. Additionally, the weather got colder on the northern hemisphere and therefore people started gathering more inside. Despite the immediate response of the governments, it took more than a month to slow down the increase because of the delay between a policy change and the response of the behavior of the spread. (John Hopkins Medicine, 2020) (TIME, 2020) Regardless of the effort and implemented policies, most of the countries are in the current time, at the beginning of February 2021, still in a very desperate situation with stable high daily increase.

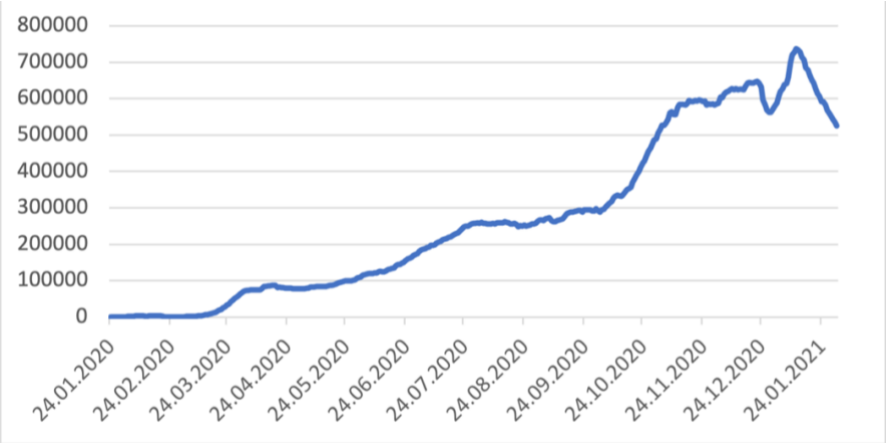

Figure 3 - Daily new confirmed cases (rolling 7-day average)

((Our World in Data, 2021), own elaboration)

3.5.3 **Preventive measures**

This chapter describes the preventive measures which were implemented by most of the governments in the world at some point in the battle against COVID-19. The data as well as explanations of those measures are taken and interpreted from a publicly available dataset by The Oxford Covid-19 Government Response Tracker, OxCGRT, a project created by the *Blavatnik School of Government.* (HALE et. al., 2020). These restrictions aim to slow down the spread of the disease and consequently, the number of newly infected citizens. For each factor there are several levels of strictness which were implemented. These levels as well as their meaning are described in the following paragraphs.

1. *Closing of schools* = specifies the status of schools in the country
   - 0–3; 0 = no measures, 3 = all levels of schools closed
2. *Closing of workplaces* = specifies the status of workplaces of people (including all sectors – office jobs, stores, institutions, …)
   - 0–3; 0 = no measures, 3= closing or work from home required for all but essential workplaces (such as grocery stores, medical staff, etc.)
3. *Cancelling public events* = restrictions on organizing and holding public events
   - 0–2; 0 = no measures, 2 = cancelling required

4. *Restrictions on gatherings* = prohibits gatherings of people exceeding certain count
   - 0–4; = 0 = no restrictions, 4 = restrictions on gatherings of 10 or less people
5. *Limitations on public transport* = measure limiting the public transport, both reduction in number of available lines and volume in transportation
   - 0–2; 0 = no limitations, 2 = closing or prohibiting most citizens from using it required
6. *Stay-at-home requirements* = lockdown, demand on citizens to stay home, strictness determines the conditions for leaving one's house
   - 0–3; 0 = no measures, 3 = not possible to leave the house with minimal exceptions required
7. *Intra-national travel restrictions* = limits the travelling between certain cities or regions
   - 0–2; 0 = no measures, 2= internal movement restrictions in place
8. *International travel restrictions* = limits international travel
   - 0–4; 0 = no restrictions, 4 = ban on arrivals from all regions or total border closure
9. *Facial coverings* = determines the recommendation or obligation to cover one's face in public spaces
   - 0–4; 0 = no policies, 4 = face covering required outside one's home at all times regardless of the location or presence of other people

3.5.4 **Other factors**

Besides the preventive measures implemented, there are multiple factors affecting the spread and course of COVID-19 in the individual countries. Scientific research has shown that among those factors belongs the number of performed tests, population density and also the rate of airport traffic. The number of tests performed is very important since when more tests are done, more positive cases are uncovered. To avoid the misreading of data because

of this phenomenon, it is useful to observe not only the number of confirmed cases but also the positivity rate (how many percent of the performed tests had a positive result). The effect of population density is described more deeply in the following chapter (*4.6.3 Relationship between population density and number of new cases)*. Airport traffic had the most impact in the beginning of the spread because the main epicenters occurred in places with high traffic, movement and interaction of people. Other discriminatory factors are for example the healthcare index and the proportion of older people in the population. (SATAYAKI et. al., 2020)

As mentioned previously, tourism and especially aerial transport has an impact on the spread of the disease. The first problem which, especially airplane travel, entails is the transport of a possibly infected person to another country, where by then could've been no cases so far. The advance and extent of international travel creates an ideal medium for the spread of the virus to the whole world therefore countries with extensive tourism are the ones which are the most susceptible to an extensive outbreak. A study conducted by Farzanegan, et. al. discovered that with a 1% increase in tourism observed in the last decade in a country, the number of confirmed COVID-19 cases and deaths rose by 1,2% and 1,4% respectively in that country. (FARZANEGAN et. al., 2020)

Furtherly, a new study shows that cultural differences, more specifically the willingness to follow rules, have a significant impact on the development of COVID-19 in the specific country. Some countries, usually those which historically had to cope with chronic threats (such as famines, invasions, diseases, natural disasters, etc.), can be identified as "tighter" since they had to develop stricter social norms in order to survive. As the study shows, these countries reported significantly less cases and deaths per million than those countries which can be identified as "loose". (GELFAND et. al., 2021)

4 Practical Part

4.1 Main Objective

The main objective of the practical part is to assess the development of COVID-19 in selected countries. The allocation of countries into groups is based on cluster analysis, where five factors were used as a criterion for the division. This paper aims to observe and explain factors which had an impact on the spread and also the progress of the disease in the specified countries.

As mentioned in the early paragraphs of this thesis, there is no universal definition of what Big Data is and how voluminous the dataset has to be in order to be considered a Big Data dataset. For the analysis in the practical part multiple datasets of various sizes were used. Some of them were relatively small containing only tenths of rows, however most of the datasets comprised thousands of inputs and those can be classified as Big Data. The data used originate from various sources and mainly text analytics was used. Additionally, predictive analysis was used in the form of the Linear Regression analysis, representing supervised learning, and the Hierarchical cluster analysis which represents unsupervised learning

4.2 Criteria used for Cluster analysis

The cluster analysis is based on five factors which are as follows:

1. Government Effectiveness

Government Effectiveness is an indicator issued by the World Bank and is a part of the *Worldwide Governance Indicators.* This specific index should describe the impression of: "quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies" (WORLDBANK, 2020). The values describing Government Effectiveness range from –2,5 which signifies very weak performance of the government to +2,5 which means very strong.

In our case were taken data from year 2010 to 2019 and they were averaged to get the most reliable number as possible.

2. Human Development Index

Human Development Index (HDI) is a measure developed by the United Nations and it measures key dimensions of human development – health, knowledge (education) and standard of living. More specifically, it takes data such as life expectancy at birth, mean of years of schooling, national income per capita, etc. to assess the level of growth in specific countries. It is calculated by normalizing each of the component to a set scale between 0 and 1 and creating a geometric mean[1] of those values. (Investopedia, 2020) (United Nations Development Programme Human Development Reports, 2020)

Among its main critiques belong that it does not take into consideration inequalities, poverty or human security. Additionally, the correlation between the Human Development Index and GNI (or GDP) per capita is very high and therefore it can be perceived as redundant for a certain types for analysis, however, for our purposes is HDI more meaningful since it takes into consideration more dimensions of human life.

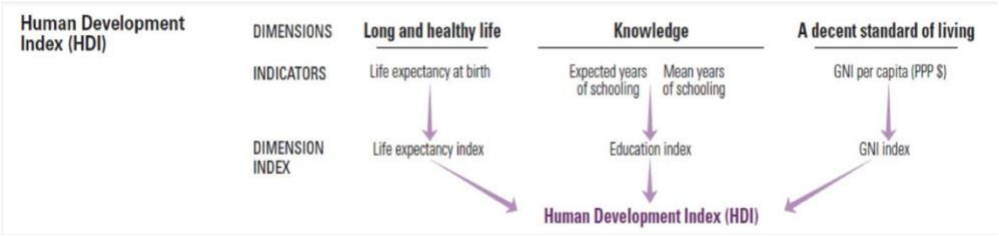

Figure 4 - Human Development Index (United Nations Development Programme Human Development Reports, 2020)

Similarly as with the first criteria, the Government Effectiveness, the final number used for the actual analysis is an average of multiple historical values of HDI (2010–2019).

3. Demographic structure

This factor describes the composition of the population of the specific state, specifically the age structure of it. Demographic structure can be defined as the proportion of each age group in the total population of the country (AKSOY et. al., 2015)

[1] „The *Geometric mean* $\sqrt{ab}$ of positive real numbers *a* and *b* (…) gives the length of the side of the square with the same area as the rectangle with sides of length *a* and *b.*" (LAWSON et. al., 2001)

For the purpose of this paper, we are working with three age group categories, Ages 0–14; Ages 15–64; and Ages 65+. Data which is used are from the year 2019 in order to keep the analysis as up to date as possible. The numbers used for the analysis signify the percentage representation of each group in the total population of the territory.

4. Consumer Price Index

Consumer Price Index (CPI) "is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services" (U. S. Bureau of Labor Statistics, 2021). In other words, the index is based on the comparison of prices of the base period (year) and another selected period (year) either in the future or in the past. Our base year is 2010 and as the final analyzed variable is used the average of indexes of years 2011 to 2019 which are expressed as a percentage. This shows us an average level of inflation in the span of the last 9 years.

5. Imports and Exports

Imports can be defined as "the value of foreign goods and services bought by a country's households, firms, government agencies, and other organizations in a given period of time". It can be influenced by multiple factors such as National Income, Trade barriers, Real Exchange rate, etc. (Economics Online, 2021). To put the number into a perspective, for the analysis were used Imports expressed as a percentage of the total production of the specific country.

On the other hand, we have exports. Those are defined as "the value of goods and services produced by a country's firms in a given period of time and which are sold abroad". It can refer to both (final) physical goods and services. (Economics Online, 2021). Likewise with imports, for the cluster analysis were used percentual value of exports to total production of the particular territory.

4.3 Preparation of data for the analysis

Because of the diversity of sources of the primary data for the analysis, it was necessary to sort the database. Firstly, were deleted fields with descriptive information about the datasets. Then, the tables had to be sorted alphabetically and manually checked so that

they all label the countries with the same name. Additionally, the datasets were modified, so that they all contain the same countries. Some states were omitted in certain datasets and therefore, they were deleted from all of the other files. The database also had to be checked for missing values. If for a certain country were missing values which are crucial for the success of the analysis, again, the country had to be excluded from all of the datasets. Furthermore, in tables describing *Government Effectiveness*, *Human Development Index*, and *Consumer Price Index* were computed averages of values from multiple years. This should contribute to the reliability of the data.

4.4 Cluster analysis - Application

As a sample for the cluster analysis were used the countries of the European Union and the European Economic Area. Few of them had to be excluded from the dataset because of missing values. The clustering analysis was performed on 36 countries

The analysis was performed by *SAS® Studio* where Euclidean distance was specified as the dissimilarity measure and as a method for the clustering analysis was selected the Ward minimum-variance. The fundamental features of this method is described in the Methodology part of this thesis. *Figure 5* displays the results in the form of a dendrograph. We can see that the first country with which was Czech Republic matched is Slovenia, then it was connected with Lithuania-Croatia cluster, etc. As the final sample for our further analysis was chosen a cluster of 14 countries including: Portugal, Italy, Greece, Malta, Latvia, Estonia, Slovenia, Czech Republic, Lithuania, Croatia, Poland, Hungary, Romania and Bulgaria. This size of the cluster was selected so that it provides sufficient number of countries for the performance of various analyses but small enough so that each country can be observed individually, and the data can be sorted manually. Additionally, one level of increase in the sample size would lead to a cluster of 25 countries which was rejected because of the size of the possible sample which would also decrease the level of similarity among the selected countries.

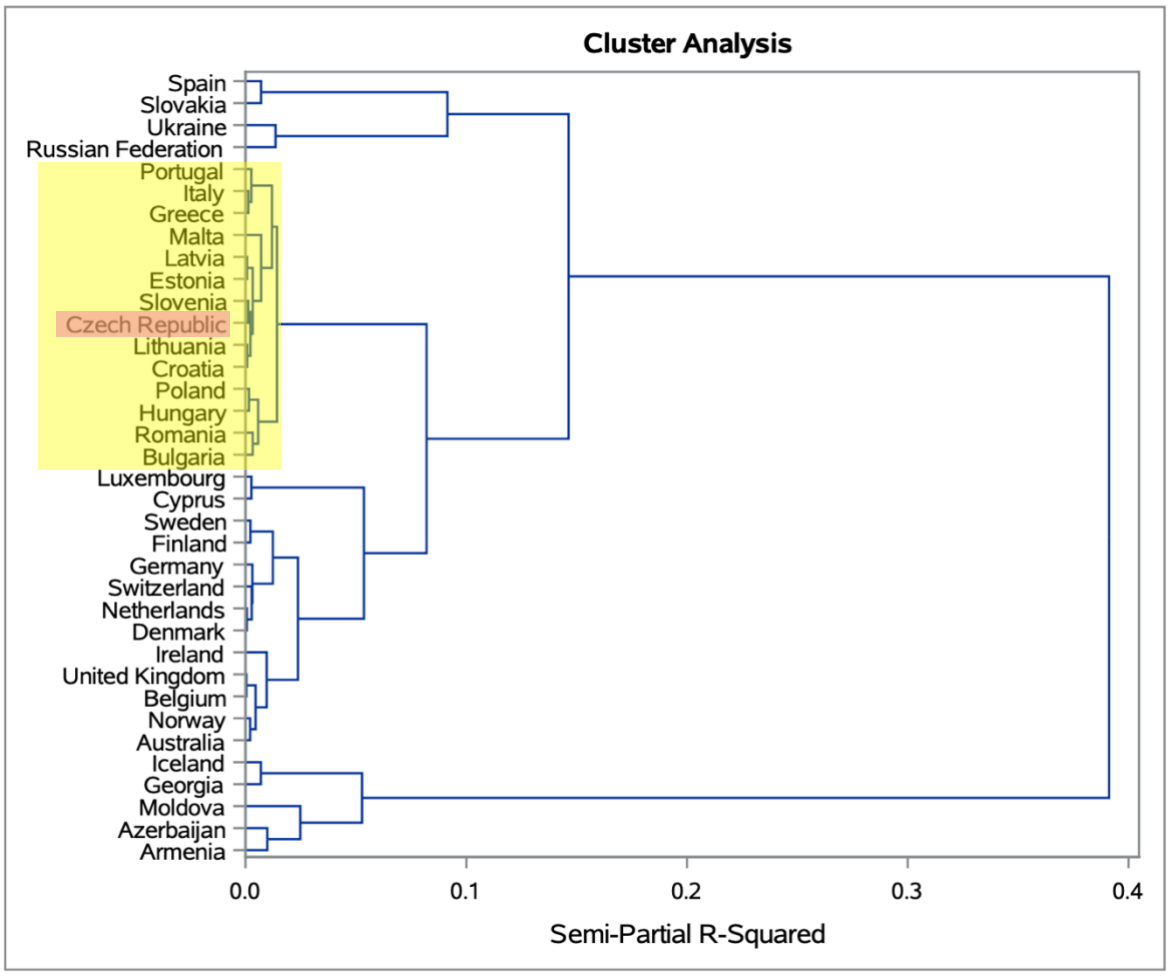

Figure 5 - Results of cluster analysis, own elaboration

4.5 Exploratory data analysis

Following part describes the findings of Exploratory data analysis performed on the indexes according to which was the sample clustered. Both values from the whole sample as well as only from the cluster are examined and crucial observations are presented for each of the criteria.

4.5.1 Government Effectiveness

As mentioned in a previous chapter, the values for the Government Effectiveness index range from –2,5 which signifies the effectiveness being very low, to +2,5 which suggest very high performance of the government. The mean value in the chosen cluster is 0,7 which

suggests neither very high nor low effectivity. This can have an impact on the carrying out of the measures and when communicating the new restrictions to the public, and later, endorsing them.

4.5.2 Human Development Index

The Human Development Index ranges from 0 to 1 where the more developed the country is, the higher is its HDI. The analyzed cluster has a mean value of HDI equal to 0,859 which is not very different from the average HDI value for the whole sample (0,849). This shows that the development is comparable in all EU and EEA countries. It can be due to the fact, that EU sets standards for the countries which are part of it and it also helps the countries to increase and maintain certain factors, such as the standard of living, at a certain level.

4.5.3 Demographic structure

For our analysis is the population divided among three main age groups: below 15 years; aged 15–64; and older than 65. The average for the first, second and third group of the selected cluster (in *Figure 6* marked "*Cluster"*) is 14,79%, 64,71% and 20,50% respectively. *Figure 6* shows the comparison of distribution of the population of the selected cluster and other countries. When compared to the whole sample the selected cluster has slightly higher portion of elderly population than is the average of the whole European area. This factor can contribute to the final number of deaths caused by the coronavirus since elderly people are affected more by it.

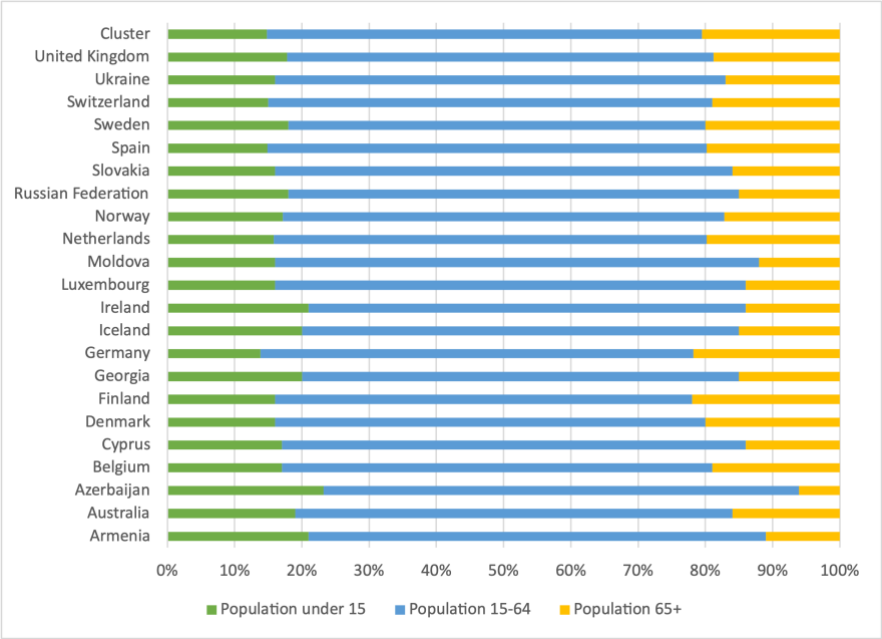

Figure 6 - Demographics ((Our World in Data, 2021), own elaboration)

4.5.4 Consumer Price Index

CPI is the most widely used inflation measure, therefore, in other words, it indicates the purchasing power of one's money. In our case it shows the development of prices in from the base year (2010). The average CPI from our cluster is 108,85 which indicates an average of 8% growth in prices over the span of 9 years (2010–2019). All of the countries in our group have Consumer Price Index higher than 100 and therefore, since 2010 (the base year) an increase in price occurred in all of them.

4.5.5 Import and Export

The values of Export and Import for the analysis were represented as a percentage of total production of that specific country. The average in the cluster for Imports and Exports is 90,09% and 95,34% respectively. This indicates that the countries are in trade surplus since the exports are higher than imports.

4.6 Analysis and comparison of development in the selected countries

The aim of this part is to explain the difference in the progress of COVID-19 by comparing the factors, which have impact on the spread or the progress of the disease, as

well as the preventive measures introduced by the governments. There are 9 specific preventive procedures which are taken into consideration and which were put into action at different time in each territory. As described in a previous chapter *(3.5.3 Preventive measures*) there are multiple levels of strictness of the restrictions which were enforced. In the following paragraphs are visualized and compared variables characterizing the development of the pandemic in the selected countries and they are correlated with the selected measures and other factors which were proven to have influence on the situation of the specific country.

4.6.1 Comparison of total increase per million in individual countries

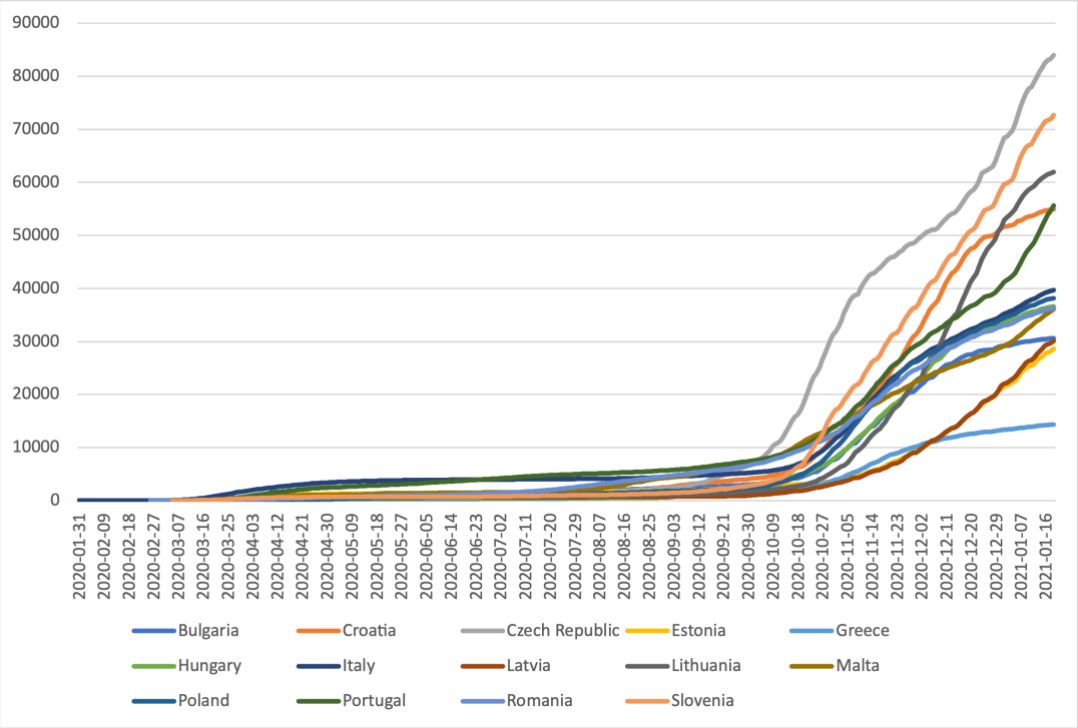

Figure 7 - Development of COVID-19 in selected countries – timeline ((Our World in Data, 2021), own elaboration)

As a first step, it is crucial to compare the overall development in the selected countries. For this purposes data from *Our World in Data* (Our World in Data, 2021) was used. It is a publicly available data source with daily updates to the values and therefore, it provides up do date information about the spread of COVID-19 in the world as well as other important information which are connected to it. *Figure 7* shows the progress on a timeline

starting from January 2020 to January 2021. On the vertical axis is the cumulative number of cases per million citizens since the beginning of the pandemic. We can see that the situation evolved very similarly in the vast majority of the chosen countries.

The lowest overall increase in our cluster has Greece where the total increase per million citizens is approximately 15 000. It is even more remarkable when the fact that it has the second oldest population in Europe is taken into account. Additionally, it is a very popular tourist destination with millions of travelers coming each year. According to *TIME magazine* and *AP News*, the probable key for their success is the early implementation of restrictions regarding gatherings of big groups of people and closing of schools. The reason for their fast response is the bad state of the health care system in the country because it wouldn't be able to cope with an extensive outbreak. (TIME, 2020) (AP News, 2020)

On the exact opposite side of the spectrum is Czech Republic, where the numbers skyrocketed during the second wave of the pandemic. The government created anti-epidemiological evaluation system *PES*[2] which should determine the severity of the situation in the country and based on that will the measures be regulated. (Ministry of Health of the Czech Republic, 2020) Even though this system was put into action in the middle of November, by the time of the making of this thesis (February 2021) the situation has not yet improved. One of the explanations can be the reluctance of the citizens to obey the rules set by the government – people still gather, don't obey rules about face coverings and drinking in public and even in some cases the owners secretly opened their restaurants and pubs to the public. (iDNES, 2020) This behavior has impact on the situation regardless which additional measures and restrictions are put in place by the government.

[2] PES refers to *Proti-epidemiologický systém* (*Anti-epidemiological system)*

4.6.2 **Development in the number of cases per million – April 2020**

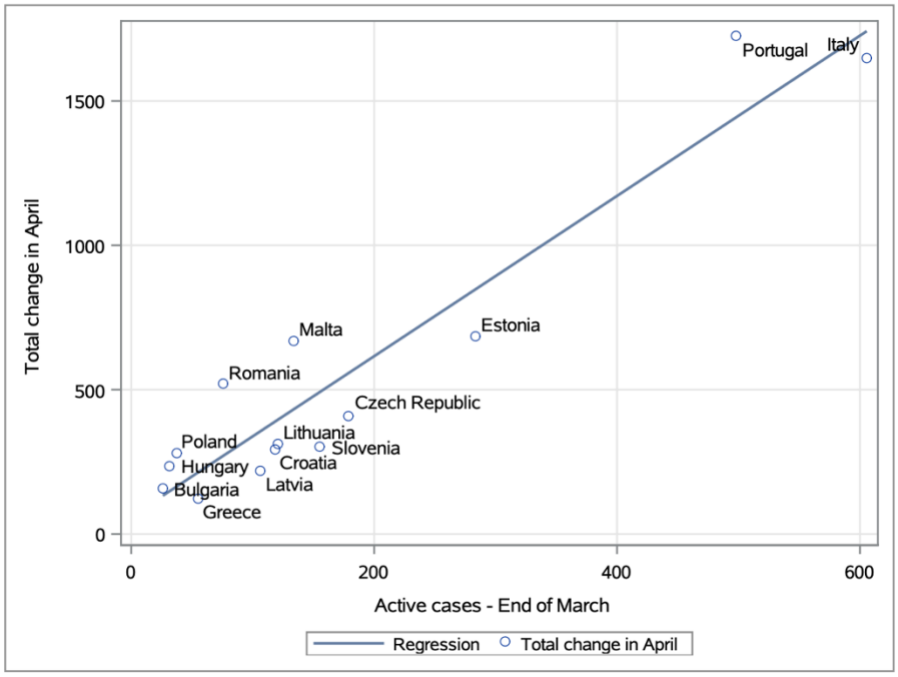

Figure 8 - Development in the number of cases per million in April 2020 ((Our World in Data, 2021), own elaboration)

Figure 8 shows the development of the situation regarding COVID-19 in the selected countries by comparing two variables. On the horizontal axis is the number which signifies the sum of the newly discovered cases in the last week of March 2020 recalculated per one million citizens. This number describes the situation at the beginning of the observed period and is an approximation of the number of active cases in the specific territory. It was proven that a person is the most infectious in the first 5-7 days after the symptoms show and therefore in this period the positive patients spread the disease the most. (CEVIK et. al., 2020) (Becker's Hospital Review, 2020) Additionally, the data from a whole week was used in order to eliminate the possible variations in the number of tests performed on the specific day of the week and with it connected misinterpretation of data. On the vertical axis is the change in the total number of confirmed cases per million which occurred during April 2020. It also displays the Linear Regression line which depicts the relationship among the variables.

We can see that not only the relationship among the variables exists but that it is very strong since all of the countries lay closely along the line. If we perform a Linear Regression

analysis, we get an r^2=0,8966. As described in the methodology, r^2 is the coefficient of determination and it tells us how the differences in one variable (dependent variable) can be explained by difference in another variable (independent variable). In this analysis it depicts that 89,66% of the changes in the dependent variable (the total change in April 2020) is explained by the change in the independent/explanatory variable (number of Active cases in March). This means, that the change in this time period is majorly influenced by the number of cases at the beginning of the period and only approximately 10% is influenced by other factors. We can assume that among aspect influencing the increase belong the measures and restrictions taken as well as other predispositions and factors influencing the countries.

To assess which measures are effective we can compare countries which started at the same level of active cases at the end of the March 2020, but which increase during April 2020 differs. As an example of those countries were chosen Croatia, Lithuania and Malta, all starting at around 120 active cases per million of citizens. Croatia and Lithuania had a total increase in April of around 300 new cases per million whereas Malta had more than double. If we compare the restrictions in place, all three countries are very similar. The analysis of the measures implemented had to be done manually. It was observed and compared when did the individual countries implemented each measure and which level of strictness they enforced. Most of the measures were put into action in the middle of March 2020 and also the severity of the restrictions implemented is comparable. The reason for Malta's poorer situations seems to be the high population density which is 25times higher than it is in the other two countries. The topic of population density is analyzed in the following chapter.

4.6.3 **Relationship between population density and number of new cases**

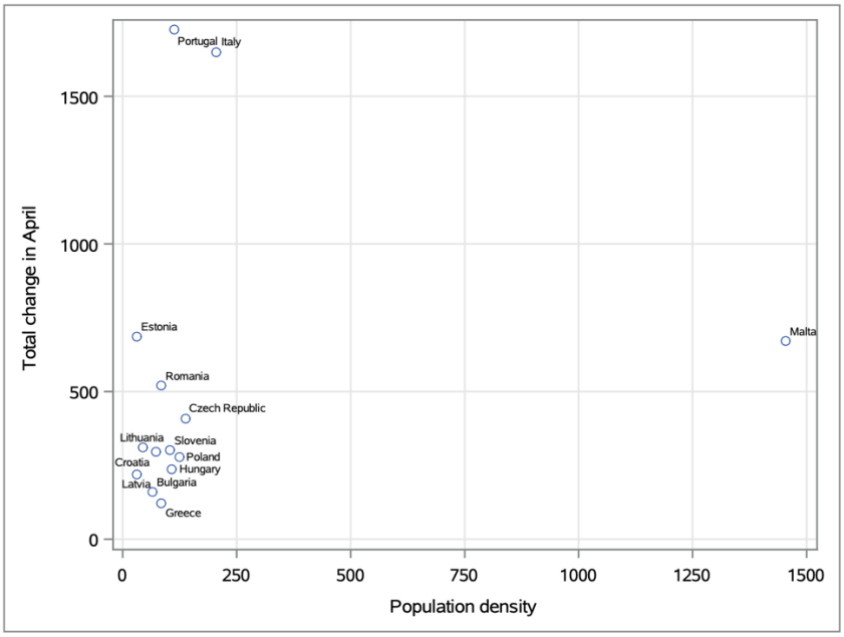

Figure 9 - Population density and increase in cases

((Our World in Data, 2021), own elaboration)

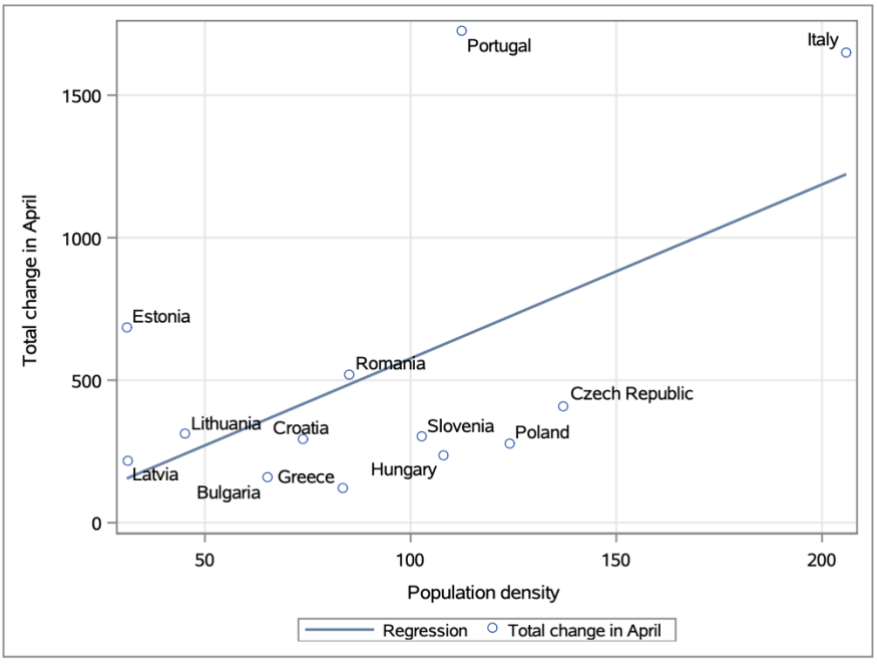

Figure 10 - Population density and increase in cases (without Malta)

((Our World in Data, 2021), own elaboration)

As described in the theoretical part, studies show that the relationship between the population density and number of cases of COVID-19 exists. (SATAYAKI et. al., 2020) The comparison of countries with similar population density but varying increase of confirmed cases can spotlight which measurements were crucial to limit the spread.

Figure 9 shows the relationship between the population density in individual countries and the change in the number of cases in April 2020. We can see that the population density for most of the countries ranges from 30 to 200 citizens per kilometer square and only Malta has much higher population density of approximately 1300 people per km-sq. In order to display data more clearly *Figure 10* was created. This scatter plot displays the same data, but Malta was excluded from the sample.

In this second graph, the relationship among the two variables is more evident. If we perform Linear Regression analysis, we get an $r^2 = 0{,}3003$, therefore, approximately 30% of the variation in the increase in April 2020 is explained by the change in population density. Also it has to be considered, that population density does not depict the proportion of people living in the cities versus those living in rural areas and other important factors which have significant impact on the outcome. The disease spreads much more in cities because of crowded communal areas such as public transport whereas in the villages where people come to less contact the spread is not so fast.

As discussed before, Baltic states like Latvia and Estonia have very low population density which highly contributed to the low spread of COVID-19 in those countries.

Even though Portugal has comparable population density as countries like Czech Republic, Poland, and Hungary the outbreak of the disease was much more severe there. If we compare the measures issued in the individual countries, we find out that restrictions regarding the closing of public spaces such as schools and workplaces were implemented at approximately the same time in all of those countries. The prohibition of public events and gatherings was announced approximately a week later in Portugal in comparison with Czech Republic and Hungary, but the biggest difference is in the obligation to cover one's face in public. From these four countries, Czechia was the first to adopt this restriction, followed almost a full month later by Poland and another two weeks later by Hungary. Portugal implemented this measure at the beginning of May when there have already been more than 13 000 cases of the disease and the daily increase was around 300 new cases. (In comparison, Czech Republic, Hungary and Poland issued this restriction when there was a

total of 464, 2 583, and 7 918 cases respectively.) Additionally, Portugal suffered from the lack of medical equipment (similarly as other countries) since the government believed that the disease won't reach their country. (El País, 2020)

Second largest increase during April 2020 was in Italy. Italy was one of the first European countries where was the novel coronavirus detected. On top of that, it has relatively high population density of 205 citizens per square kilometer (EU's average is at 112 people/km-sq) and has one of the oldest populations in the world. Czech Republic has the most similar population density (137ppl/km-sq) so we will use it for the comparison of the applied measures. Even though it is the closest match, it has to be taken into consideration that it is a big difference and therefore it is not fully comparable. In Czech Republic was the first case discovered on March 1, which is approximately one month later than the first case in Italy. Because of that will not be compared the exact dates of the implementation but the number of days that passed between the first discovered case and the implementation of the measures.

Overall, Czech Republic was much faster in the implementation of all of the measures by approximately 10 days. Additionally, the schools were closed after 10 days after the first discovered case in Czechia whereas it took a whole month in Italy. The biggest difference is, once again, in the restrictions obliging people to cover their face in public areas. Czech Republic issued this measure approximately two weeks after the detection of the first case whereas in Italy it took them more than two months. Since the coronavirus is transmitted through respiratory droplets emitted by an infected person, wearing a facemask significantly reduces the spread of the disease in communal areas and therefore this measure is seen as crucial.

4.6.4 **Deaths per million by country**

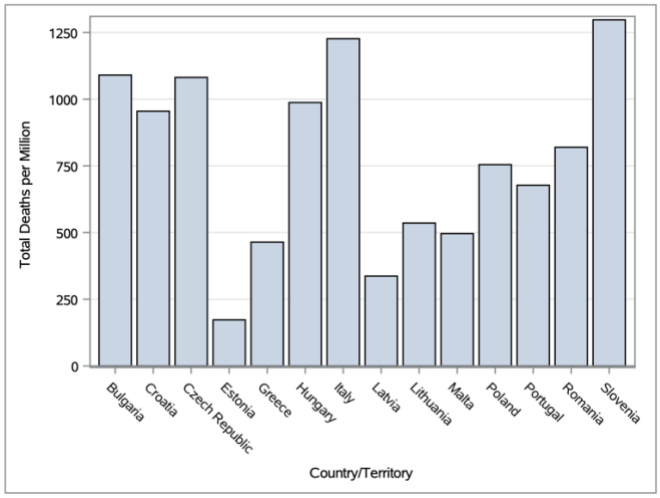

Figure 11 - Deaths per million by country ((Our World in Data, 2021), own elaboration)

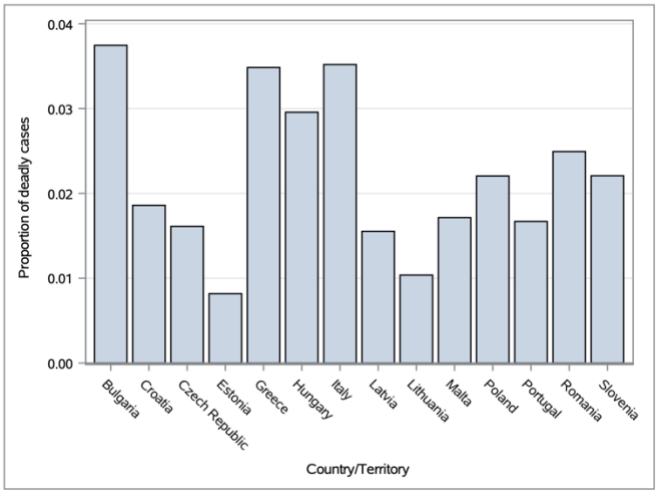

Figure 12 – Mortality rate ((Our World in Data, 2021), own elaboration)

Figure 11 is a bar chart showing the total number of deaths per million associated with COVID-19 to the date December 31, 2020. We can see that the number ranges from the minimum of approximately 200 deaths per million in Estonia to almost 1 300 in Slovenia.

To put the numbers into perspective, *Figure 12* was created. It shows the mortality rate of the disease, meaning the proportion of cases that were fatal from the total number of discovered cases. Even in this case, Estonia still remains the country with the lowest death count. On the other hand, Slovenia which showed a high number of deaths per million in

Figure 11 shows relatively lower number in *Figure 12.* The highest proportions of deaths according to this analysis have Bulgaria, Greece and Italy.

Estonia's success in containing the disease projected also on the final number of deaths reported in the country. Its fast response to the crisis and implementation of restrictions such as lockdown had the desired effect, and the numbers remain very low in comparison to other countries. Additionally, it can be due to high digitalization of the country even prior to the crisis. This helped when forming an information portal for the citizens and foreigners as well as when developing a tracking software. (Vontobel, 2020) On the other hand, when compared with countries which surround Estonia, its performance is comparable to their results. One of the possible causes of that can be the low population density of those countries. (LRT English, 2020) The influence of population density on the spread and course of COVID-19 was discussed in the previous chapter.

Italy's high portion of deadly cases is most probably caused by the high percentage of older citizens. It has approximately 23% of citizens who are older than 65 which is more than double of the worlds average (slightly over 9% (The World Bank, 2019)). A similar situation occurs in other countries from our cluster. Since the demographics was one of the factors upon which was the clustering analysis based, all of the countries in the cluster have significantly larger portion of the elderly citizens than the rest of the world (the cluster averaging on 20,5%).

If the Linear Regression analysis is performed on our sample, it reveals that not even 2% of the change in the number of deaths is explained by the percentage of the elderly. There are several factors which can contribute to this result, which contradicts the beliefs of the experts. Firstly, it is the small size of our sample. We are observing only 14 countries which are very similar and therefore it lacks the wide range of values which would be otherwise observed when studying a broad spectrum of countries. Additionally, it is the lack of standardization in determining which deaths are caused by COVID-19 and which were a result of a different disease or cause, but the person tested positive for the Coronavirus.

The example of overreporting is Belgium where the government wanted to be as transparent as possible and reported not only the deaths which were proven to be caused by COVID-19 but also those where the suspicion was. (NPR, 2020) In other countries we can encounter the opposite phenomenon – underreporting. This can occur for various reasons – political, lack of testing or misdiagnosis. An example of underreporting can be North Korea

(DPRK) where even though, that in the neighboring China were reported hundreds of cases, were reported none. (TIME, 2020)

The varying death rate is influenced by multiple factors and therefore it cannot be accounted to only one aspect. Circumstances such as environment, culture and hygiene contribute to the final number as well as other factors including the demographics and government. (SATAYAKI et. al., 2020) The only measure which will with great probability affect the final number of deadly cases is the vaccine. After a certain period of time, it will be possible to test if there is a negative relationship between the number of vaccines administered and count of deaths but currently this data is not available since the vaccination of the population has begun recently. (SKYNEWS, 2020)

5 Results and Discussion

The analysis performed signals that there is no one measure nor other singular factor which would determine the spread and course of the disease in a specific country. It is the combination of both the measures implemented as well as the mean of recording cases and deaths, attitude of the citizens towards the restrictions and finally the culture and social norms. The factors which stand out the most are the number of active cases and the population density of the specific country.

The reason for the importance of the situation at the beginning of the observed period is that the number of infected people grows exponentially. Experts say that without any intervention the number of cases of people infected with COVID-19 doubles every three to four days. (BBC, 2020) This highly increases the importance of an in-time rapid action because the lower the number of currently infected people is, the less people they can infect. *Figure 8* displays this relationship and the future analysis of the available data even shows that almost 90 % of the change in the number of cases that occurred in April 2020 is explained by the number of "active" cases at the end of March. Therefore, it can be expected that when comparing multiple countries, the variance in the number of previously infected patients (especially the number of "active" cases) has very significant impact on the development of the situation in following periods.

Population density is another factor which was proven to have impact on the spread of COVID-19 not only by our analysis but also by experts. Our analysis shows, that 30% of the change in the dependent variable (Change in the number of cases in April 2020) is explained by the population density of the territory. This finding is also supported by the research conducted by Satayaki and Ghoshs (SATAYAKI et. al., 2020) where population density was determined as one of the key aspects when determining the spread of the novel coronavirus.

Another indicator which was observed is the mortality rate. In order to get proportional data was instead of the absolute value of deaths per million citizens observed the percentage of fatal cases out of the total detected cases. The country with the lowest mortality rate was Estonia where, as the main factors influencing their success, were identified the quickness of the response to the crisis and the high technical and digitalization standards of the country as well as the low population density of the territory. On the other hand, as the country with the highest mortality rate was identified Italy. Even though it wasn't proven by the analysis

completed on our sample, there the main contributing factor is most probably the high proportion of elderly citizens among the country's population. The reasons for the more severe course of the disease are mainly the pre-existing conditions of those people, such as diabetes, heart disease, etc. which are more common among the elderly citizens. Additionally, the decrease in immunity and biological aging also contribute to the severeness of the illness. (MUELLER et. al., 2020)

The measures introduced in the theoretical part of the thesis were observed as a part of each analysis but were not proven as critical. In majority, there were several restrictions introduced at the same time and therefore, the effect couldn't be attributed to a single measure. Additionally, the effect of a measure is not immediate which complicates the process of assessing its effect.

In addition to factors which were more closely examined in the practical part of the thesis, there are other attributes which significantly influence the spread and/or the course of COVID-19. Among those factors belongs the previously mentioned tourism and air-travel, hygiene, "tightness" of the culture, and others. Additionally, it has to be taken into account that the means of recording the number of new cases and which deaths are to be attributed to Coronavirus are not similar in all of the countries. Therefore, either over- or under-reporting can occur which may have an impact on the reliability of the data and the analyses performed on those data.

6 Conclusion

The main objective of this thesis was to assess and compare the development of COVID-19 in selected countries. The grouping of countries was based on Hierarchical cluster analysis and the cluster with Czech Republic was picked as a sample for further analysis. It was presumed, that the similarity of countries is great since the cluster analysis was based on multiple criteria assessing the countries from various angles. The first criteria was the effectivity of government activities and communication of new legislation, which enables us to analyze the effect of the individual measures implemented without questioning the differences in the means of translating the restrictions to the public. Furtherly, the countries were clustered on the basis of both economical and socio-economical indexes as well as well-being indicators and the demographics of the selected territory.

The data signifying the development of the virus were firstly compared among the selected countries in order to study the overall pattern of the spread. This analysis also displayed which countries perform better, have lower number of cases per million, and which worse. Even in this part were introduced some factors which influence the spread such as the response time of the governments leading to the lowest numbers observed (Greece) or on the other hand the effect of unwillingness of Czech citizens to obey the rules set by the government.

Secondly, was analyzed the increase in the number of cases observed in April 2020 and the effect of two factors on it – namely the number of "active" cases and the population density. It was discovered that the number of cases in the preceding period before the observed period is crucial and significantly influences the final number. This discovery is important for future analysis. When analyzing multiple territories where the situation differs the observer should take the number of active cases into considerations, in order to avoid misinterpretation of data and be able to fully understand it. The comparison and correlation of population density was also proven. Since COVID-19 is transmitted through secretions and respiratory droplets through the air, in places where the concentration of people is higher the chance of transmitting the disease is higher. This benefited the countries with low population density, such as Baltic states in our cluster, and on the other hand was disastrous for densely populated areas such as Hong Kong or the New York City.

The mortality rate is another indicator which was observed in the practical part of this thesis. Even though the relationship between the percentage of elderly citizens in the population and the mortality rate was not proven by the linear regression analysis of the specified sample, it is confirmed by previously cited scientific research. It corresponds with the finding of Italy's high mortality rate since Italy has more than double the world's average of citizens over the age of 65.

To improve our understanding of the disease and its spread, further research can focus on the comparison of a bigger sample of countries or contrasting the clusters with each other. When comparing the data from the whole clusters, the reliability of the results would improve, because of the larger sample, and therefore, the applicability of them would also increase. The variance in data provided by this set of countries would also provide a mean to observe and define factors influencing the spread and perform further analysis of their correlation to number of new cases or mortality rate.

7 Bibliography

MONTGOMERY, D. C., E. A. PECK a G. G. VINING, 2012. *Introduction to Linear Regression Analysis*. 5. John Wiley & Sons, 2012. 0470542810, 9780470542811.

GOVAERT, G., M. NADIF a G. GOVAERT, 2013. *Co-Clustering: Models, Algorithms and Applications*. John Wiley & Sons, Incorporated. ISBN 9781118649497.

WARD, J. S. a A. BAKER, 2013. In: *Cornell University* [online].2013 [cit. 2019-May-23]. Available at: https://arxiv.org/pdf/1309.5821.pdf

DHAENENS, C. a L. JOURDAN, 2016. *Metaheuristics for Big Data*. John Wiley & Sons, Incorporated. ISBN 9781119347583.

GANDOMI, A. a M. HAIDER, 2015. Science Direct. In: *Beyond the hype: Big data concepts, methods and analytics* [online]. 2. April. 2015 [cit. 2019-May-23]. Available at: https://www.sciencedirect.com/science/article/pii/S0268401214001066

KOTU, V. a B. DESHPANDE, 2014. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Elsevier Science & Technology. ISBN 9780128016503.

LAROSE, D. T. a C. D. LAROSE, 2015. *Data Mining and Predictive Analytics*. John Wiley & Sons, Incorporated. ISBN 9781118868676.

ACHARI, S., 2015. *Hadoop Essentials*. Packt Publishing, Limited . ISBN 9781784390464.

MOORE, D. S. a G. MCCABE, 1998. *Introduction to the Practice of Statistics, 3rd Edition*. W H Freeman & Co. ISBN 9780716734581.

LABRINIDIS, A. a H. V. JAGADISH, 2012. ACM Digital Library. In: *Challenges and Opportunities with Big Data* [online].2012 [cit. 2019-May-24]. Available at: https://hpc-forge.cineca.it/files/CoursesDev/public/2014/Tools_Techniques_Data_Analysis/papers/p2032-labrinidis.pdf

SIEGEL, E., 2016. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. John Wiley & Sons, Incorporated. ISBN 9781119153658.

ABOTT, D., 2014. *Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst*. John Wiley & Sons, Incorporated. ISBN 9781118727935.

SATAYAKI, R. a G. PREETAM, 2020. PlosOne. In: *Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking* [online]. 23. October. 2020

[cit. 2021-February-02]. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241165#abstract0

AKSOY, Y. et al., 2015. *Demographic Structure and Macroeconomic Trends* [Electronic Document]. ISSN 1579-8666.

SpringerLink, 2020. *Exploratory Data Analysis* [online]. accessed 2020 [cit. 2020-December-27]. Available at: https://doi.org/10.1007/978-0-387-32833-1_136

PANIGRAHI, B. K., A. ABRAHAM a S. DAS, 2010. *Computational intelligence in power engineering*. Springer. ISBN 3642140130.

U. S. Bureau of Labor Statistics, 2021. *Consumer Price Index* [online]. accessed 2021 [cit. 2021-January-07]. Available at: https://www.bls.gov/cpi/

Economics Online, 2021. *Imports* [online]. accessed 2021 [cit. 2021-January-07]. Available at: https://www.economicsonline.co.uk/Definitions/Imports.html

SEEMON, T., 2014. *Basic Statistics*. Alpha Science Internation. ISBN 9781783320301.

DUPUIS, F. a L. LAURENCELLE, 2002. *Statistical Tables, Explained and Applied*. World Scientific Publishing Company. ISBN 9789812777669.

BOCK, T., 2021. DISPLAYRBlog. In: *What is Hierarchical clustering?* [online].acessed 2021 [cit. 2021-February-02]. Available at: https://www.displayr.com/what-is-hierarchical-clustering/

PLOS ONE, 2017. *Generalising Ward's Method for Use with Manhattan Distances* [online] [cit. 2021-February-19]. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0168288

HU, Y., K. LI a A. MENG, 2018. github. In: *Agglomertive Hierarchical Clustering using Ward Linkage* [online]. 7. December. 2018 [cit. 2021-February-19]. Available at: https://github.com/Anranmg/project506/tree/master/project506/final

LANEY, D., 2001. Meta Group: Application Delivery Strategies. In: *3D Data Management: Controlling Data, Volume, Velocity, and Variety* [online]. 6. February. 2001 [cit. 2019-May-25]. Available at: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

The Pathologies of Big Data, 2009 [online] [cit. 2019-May-23]. Available at: http://www.informatica.uniroma2.it/upload/2018/IA2/The%20Pathologies%20of%20Big%20Data.pdf

KRISHNAN, K., 2013. *Data Warehousing in the Age of Big Data*. Elsevier Science & Technology. ISBN 9780124059207.

MathWorld. A Wolfram Web Resource, 2020. *Outlier* [online]. accessed 2020 [cit. 2019-May-29]. Available at: http://mathworld.wolfram.com/Outlier.html

World Health Organization, 2020. *Archived? WHO Timeline - COVID-19* [online] [cit. 2021-February-02]. Available at: https://www.who.int/news/item/27-04-2020-who-timeline---covid-19

Our World in Data, 2021. *Coronavirus Pandemic (COVID-19) – the data* [online]. accessed 2021 [cit. 2021-January-20]. Available at: https://ourworldindata.org/coronavirus-data

John Hopkins Medicine, 2020. *Coronavirus Second Wave? Why Cases Increase* [online] [cit. 2021-February-02]. Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus

TIME, 2020. *Europe's Second Wave of COVID-19 is Being Driven by Two Countries. Here's Why* [online] [cit. 2021-February-02]. Available at: https://time.com/5902172/europe-coronavirus-second-wave-belgium-czech-republic/

HALE, T. et al., 2020. GitHub. In: *Oxford Covid-19 Government Response Tracker (OxCGRT)* [online].2020 [cit. 2021-February-02]. Available at: https://github.com/OxCGRT/covid-policy-tracker

FARZANEGAN, M. R. et al., 2020. *International Tourism and Outbreak of Coronavirus (COVID-19): A Cross-Country Analysis* [Article]. 3. July. 2020.

GELFAND, M. J. et al., 2021. The Lancer Planetary Health. In: *The relationship between cultural tightness–looseness and COVID-19 cases and deaths: a global …* [online]. 29. January. 2021 [cit. 2021-February-02]. Available at: https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(20)30301-6/fulltext

WORLDBANK, 2020. Worldbank. In: *Worldwide Governance Indicators* [online].accessed 2020 [cit. 2020-December-27]. Available at: https://info.worldbank.org/governance/wgi/Home/Documents

Investopedia, 2020. *Human Development Index (HDI)* [online]. 29. December. 2020 [cit. 2021-January-04]. Available at: https://www.investopedia.com/terms/h/human-development-index-hdi.asp

United Nations Development Programme Human Development Reports, 2020. *Human Development Index (HDI)* [online]. accessed 2020 [cit. 2021-January-04]. Available at: http://hdr.undp.org/en/content/human-development-index-hdi

Economics Online, 2021. *Exports* [online]. accessed 2021 [cit. 2021-January-07]. Available at: https://www.economicsonline.co.uk/Definitions/Exports.html

TIME, 2020. *Greece Has an Elderly Population and a Fragile Economy. How Has It Escaped the Worst of the …* [online] [cit. 2021-February-17]. Available at: https://time.com/5824836/greece-coronavirus/

AP News, 2020. *Greece imposes lockdown to avoid worst at hospitals* [online] [cit. 2021-February-17]. Available at: https://apnews.com/article/greece-imposes-virus-lockdown-14c94c700513fc16dfdb26137542e6df

Ministry of Health of the Czech Republic, 2020. *The epidemiological situation will newly be depicted by the PES evaluation system* [online] [cit. 2021-February-04]. Available at: https://koronavirus.mzcr.cz/en/the-epidemiological-situation-will-newly-be-depicted-by-the-pes-evaluation-system/

iDNES, 2020. *Češi na podzim porušovali opatření více. Města řeší až stovky oznámení Zdroj: …* [online] [cit. 2021-February-04]. Available at: https://www.idnes.cz/zpravy/domaci/mesta-policie-spravni-rizeni-nouzovy-stav-poruseni-vladniho-narizeni-praha-brno.A201203_093934_domaci_lre

CEVIK, M. et al., 2020. *SARS-CoV-2, SARS-CoV-1 and MERS-CoV viral load dynamics, duration of viral shedding and …*. 29. July. 2020.

Becker's Hospital Review, 2020. *COVID-19 patients most infectious 2 days before, 5 days after symptoms emerge, analysis finds* [online] [cit. 2021-February-24]. Available at: https://www.beckershospitalreview.com/infection-control/covid-19-patients-most-infectious-2-days-before-5-days-after-symptoms-emerge-analysis-finds.html

El País, 2020. *Portugal and Spain: same peninsula, very different coronavirus impact* [online] [cit. 2021-February-18]. Available at: https://english.elpais.com/spanish_news/2020-05-11/portugal-and-spain-same-peninsula-very-different-coronavirus-impact.html

Vontobel, 2020. *Estonia's digital strategies in the fight against the Coronavirus* [online] [cit. 2021-February-18]. Available at: https://www.vontobel.com/en-int/impact/estonias-digital-strategies-in-the-fight-against-the-coronavirus-21306/

LRT English, 2020. *Low population density is a blessing for Baltic states during Covid-19 pandemic – analysis* [online] [cit. 2021-February-18]. Available at: https://www.lrt.lt/en/news-in-english/19/1170208/low-population-density-is-a-blessing-for-baltic-states-during-covid-19-pandemic-analysis

The World Bank, 2019. *Population ages 65 and above (% of total population)* [online] [cit. 2021-February-17]. Available at: https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS

NPR, 2020. *Why Belgium's Death Rate Is So High: It Counts Lots Of Suspected COVID-19 Cases* [online] [cit. 2021-February-02]. Available at: https://www.npr.org/sections/coronavirus-live-updates/2020/04/22/841005901/why-belgiums-death-rate-is-so-high-it-counts-lots-of-suspected-covid-19-cases?t=1611252421408&t=1612256722387

TIME, 2020. *North Korea Says It Has No Coronavirus – Despite Mounting Clues to the Contrary* [online] [cit. 2021-February-04]. Available at: https://time.com/5794280/north-korea-coronavirus/

SKYNEWS, 2020. *COVID-19 vaccine will 'substantially reduce deaths' - and UK will have up to four jabs to use by …* [online]. skynews, 10. December. 2020 [cit. 2021-February]. Available at: https://news.sky.com/story/covid-19-vaccine-will-substantially-reduce-deaths-and-uk-will-have-up-to-four-jabs-to-use-by-mid-2021-12155970

BBC, 2020. *Exponential growth bias: The numerical error behind Covid-19* [online] [cit. 2021-February-19]. Available at: https://www.bbc.com/future/article/20200812-exponential-growth-bias-the-numerical-error-behind-covid-19

MUELLER, A. L., M. S. MCNAMARA a D. A. SINCLAIR, 2020. *Why does COVID-19 disproportionately affect older people*. 29. May. 2020.

LAWSON, J. D. a Y. LIM, 2001. Taylor and Francis Online. In: *The Geometric Mean, Matrices, Metrics and More* [online].2001 [cit. 2021-January-04]. Available at: https://www.tandfonline.com/doi/abs/10.1080/00029890.2001.11919815

8 Appendix

Appendix 1 – Exploratory data analysis - Government Effectiveness

Analysis Variable : GE GE					
Mean	Std Dev	Minimum	Maximum	N	Range
0.6995742	0.3933443	-0.1938858	1.1336883	14	1.3275741

Appendix 2 – Exploratory data analysis – Human Development Index

Analysis Variable : HDI HDI					
Mean	Std Dev	Minimum	Maximum	N	Range
0.8588714	0.0314017	0.8041000	0.9260000	14	0.1219000

Appendix 3 – Demographics (data for comparison of cluster with other countries)

Country	Pop. under 15	Pop. 15-64	Pop. 65+
Armenia	21	68	11
Australia	19	65	16
Azerbaijan	23	70	6
Belgium	17	64	19
Cyprus	17	69	14
Denmark	16	64	20
Finland	16	62	22
Georgia	20	65	15
Germany	14	65	22
Iceland	20	65	15
Ireland	21	65	14
Luxembourg	16	70	14
Moldova	16	72	12
Netherlands	16	65	20
Norway	17	65	17
Russian Federation	18	67	15
Slovakia	16	68	16
Spain	15	66	20
Sweden	18	62	20
Switzerland	15	66	19
Ukraine	16	67	17
United Kingdom	18	64	19
Cluster	14,79	64,71	20,50

Appendix 4 – Exploratory data analysis – Demographics (only cluster)

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	Range
Population under 15	Population under 15	14.7857143	1.0509023	13.0000000	16.0000000	14	3.0000000
Population 15-64	Population 15-64	64.7142857	1.0690450	63.0000000	67.0000000	14	4.0000000
Population 65+	Population 65+	20.5000000	1.2860195	18.0000000	23.0000000	14	5.0000000

Appendix 5 – Exploratory data analysis – Consumer Price Index

Analysis Variable : CPI CPI					
Mean	Std Dev	Minimum	Maximum	N	Range
108.8592602	3.0344261	102.2321320	114.1396380	14	11.9075060

Appendix 6 – Exploratory data analysis – Import and Export

Variable	Label	Mean	Std Dev	Minimum	Maximum	N	Range
Export Share in Total Products (	Export Share in Total Products (%)	95.3378571	1.8560967	91.0500000	97.9900000	14	6.9400000
Import Share in Total Products (	Import Share in Total Products (%)	90.0900000	6.4366714	69.5700000	96.3800000	14	26.8100000

Appendix 7 – Data for figure 8

Country/Territory	Active cases - End of March	Total change in April
Bulgaria	26,05	159,316
Croatia	118,14	294,5
Czech Republic	178,73	408,442
Estonia	283,444	685,996
Greece	54,782	122,517
Hungary	31,572	236,327
Italy	605,605	1648,495
Latvia	106,564	218,428
Lithuania	120,486	311,503
Malta	133,623	670,383
Poland	37,256	279,18
Portugal	498,298	1726,243
Romania	75,425	519,553
Slovenia	154,888	301,597

Appendix 8 – Results of Linear regression – Figure 8

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	3087868	3087868	104.01	<.0001
Error	12	356252	29688		
Corrected Total	13	3444120			

Root MSE	172.30113	**R-Square**	0.8966
Dependent Mean	541.60571	**Adj R-Sq**	0.8879
Coeff Var	31.81302		

Appendix 9 – Data for figure 9

Country/Territory	**Total change in April**	**Population density**
Bulgaria	159,316	65,18
Croatia	294,5	73,726
Czech Republic	408,442	137,176
Estonia	685,996	31,033
Greece	122,517	83,479
Hungary	236,327	108,043
Italy	1648,495	205,859
Latvia	218,428	31,212
Lithuania	311,503	45,135
Malta	670,383	1454,037
Poland	279,18	124,027
Portugal	1726,243	112,371
Romania	519,553	85,129
Slovenia	301,597	102,619

Appendix 10 – Results of Linear regression – Figure 10

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	1029056	1029056	4.72	0.0525
Error	11	2397205	217928		
Corrected Total	12	3426261			

Root MSE	466.82725	**R-Square**	0.3003
Dependent Mean	531.69977	**Adj R-Sq**	0.2367
Coeff Var	87.79903		

Appendix 11 – Data for figure 11 and 12

Country/Territory	**Total Deaths per Million**	**Proportion of deadly cases**
Bulgaria	1090,316	0,037
Croatia	954,871	0,019
Czech Republic	1081,335	0,016
Estonia	172,63	0,008
Greece	464,163	0,035
Hungary	987,231	0,030
Italy	1226,542	0,035
Latvia	336,655	0,016
Lithuania	535,578	0,010
Malta	495,992	0,017
Poland	754,467	0,022
Portugal	677,277	0,017
Romania	819,589	0,025
Slovenia	1297,301	0,022

Appendix 12 – Results of Linear regression – Figure 11

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	24663	24663	0.19	0.6670
Error	12	1521292	126774		
Corrected Total	13	1545955			

Root MSE	356.05381	**R-Square**	0.0160
Dependent Mean	778.13907	**Adj R-Sq**	-0.0661
Coeff Var	45.75709		