Czech University of Life Sciences

Faculty of Economics and Management

Department of Statistics

Diploma Thesis

Understanding the Covid-19 pandemic: A machine learning approach

Prepared by: Rediet Shimeles Mengistu

Thesis Supervisor: Ing. Tomáš Hlavsa, Ph.D.

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Rediet Mengistu

Systems Engineering and Informatics
Informatics

Thesis title

Understanding the Covid-19 pandemic: A machine learning approach

---

Objectives of thesis

The main objective of this thesis is to analyze the given data for the purpose of investigating the Covid-19 pandemic.
Partial goals of the thesis also include;

- Cluster countries based on the spread pattern
- Cluster countries based on measures taken by governments
- Studying how the pandemic affected Socio-Economic status of the clusters.

Methodology

A dataset from https://ourworldindata.org/ and other credible sources will be used for analysis purpose of this thesis. In order to achieve the above mentioned goals of the analysis of the chosen dataset, Python is to be utilized since the language consists of a large number of analytic libraries which include packages like Numerical computing, Data analysis, Statistical analysis, Visualization and Machine learning.

The proposed extent of the thesis

60 – 80 pages

Keywords

Big Data, Data Visualisation, Data Analysis, Python, Coronavirus, Pandemic

---

Recommended information sources

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

Aggarwal, C. C. & Reddy, C. K., 2014. Data Clustering and Applications. s.l.:CRC Press.

Carlsson-Szlezak, P., Reeves, M. & Swartz, P., 2020. What Coronavirus Could Mean for the Global Economy. Harvard Business Review , p. 10.

TUFFÉRY, Stéphane. Data Mining and Statistics for Decision Making. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8.

WHO, 2020. CoronaVIrus. [Online] Available at: https://www.who.int/health-topics/coronavirus#tab=tab_1

Expected date of thesis defence

2020/21 SS – FEM

The Diploma Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 23. 11. 2020

prof. Ing. Libuše Svatošová, CSc.

Head of department

Electronic approval: 24. 11. 2020

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 26. 03. 2021

Declaration

I declare that I have worked on my bachelor thesis titled " Understanding the Covid-19 pandemic: A machine learning approach" 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.

Prague, 30th March 2021 Rediet Shimeles Mengistu

Acknowledgment

First and foremost, I want to express my gratitude to God Almighty for providing me with opportunities and guidance in my life. Next, I'd like to express my heartfelt appreciation to the Czech Republic's government for providing me with a scholarship to pursue my master's degree.

I'd like to thank my supervisor, Ing. Tomáš Hlavsa, Ph.D., for his insightful comments and positive suggestions, which profoundly aided the improvement of my work. Finally, I'd like to thank my family, especially my mom, without whom I would not have been able to accomplish this.

Understanding the Covid-19 pandemic: A machine learning approach

Abstract

Since the onset of the Covid-19 pandemic, scientists have been working frantically to learn more about the coronavirus's behaviour in order to keep it from spreading further. That's where big data comes in, delivering much-needed data to aid in the battle against the pandemic. This diploma thesis seeks to assess the covid-19 pandemic by using machine learning techniques with the goal of adding to the much-needed information about the virus.

This thesis tackles the first objective of clustering countries based on spread pattern by using hierarchical clustering algorithm. Five distinct clusters of countries of different sizes are uncovered when assessing the spread pattern. These clusters showed different patterns of the spread of the virus in different time segments.

This study also investigates how the pandemic impacted the clusters from an economic education and health standpoint after grouping countries together based on the virus's distribution pattern. Growth rate and unemployment rate are used as metrics to assess the economic effects of the pandemic. Both metrics suggest that the pandemic has had a negative effect on the economy of the clusters. To investigate the effect of the pandemic on education, school closures in each cluster for the year 2020 were analyzed and linked to the availability of internet in that cluster to show how students without internet were affected. Finally, health impact of Covid-19 are investigated, demonstrating how hospitals found in each cluster were overburdened because of the pandemic.

As the coronavirus advanced, governments took a variety of approaches to dealing with the outbreak with the intention of influencing the virus's trajectory. Hierarchical clustering is also implemented here based on the stringency of the government response and 3 distinct clusters were found. The clusters are further investigated with respected to individual policies member countries have followed.

Keywords: Big Data, Python, Coronavirus, Pandemic, New Cases, Stringency Index

Porozumění pandemii Covid-19: využití strojového učení

Abstrakt

Od vzniku pandemie Covid-19 vědci intenzivně pracují na tom, aby se dozvěděli více o chování koronaviru, aby se zabránilo jeho dalšímu šíření. S tím přicházejí velká data, která moho poskytnou mnoho potřebných informací, aby pomohla v boji proti pandemii. Tato diplomová práce se snaží posoudit pandemii covid-19 pomocí technik strojového učení s cílem hlouběji vytěžit dostupná o chování viru a reakce jednotlivých států na ně.

Tato práce se zabývá shlukováním zemí na základě vybraných indikátorů pomocí hierarchického klastrového algoritmu. Při hodnocení modelu šíření je odkryto pět odlišných klastrů zemí různých velikostí. Tyto shluky vykazovaly různé vzorce šíření viru v různých časových segmentech.

Tato práce dále zkoumá, jak pandemie ovlivnila klastry z hlediska ekonomického, vzdělávácího a zdraví. Míra růstu a míra nezaměstnanosti se používají jako metriky k hodnocení ekonomických dopadů pandemie. Oba tyto metriky naznačují, že pandemie měla negativní dopad na ekonomiku. Za účelem zkoumání vlivu pandemie na vzdělávání byly analyzovány uzávěry škol v každém klastru pro rok 2020 včetně dostupností a pokrytí internetem ve sledovaných klastrech, aby se ukázalo, jak byli ovlivněni studenti bez internetu. Nakonec jsou zkoumány dopady Covid-19 na zdraví, což ukazuje, jak byly nemocnice v každém klastru přetíženy kvůli pandemii.

Jak koronavirus postupoval, vlády zvolily různé přístupy k řešení ohniska s úmyslem ovlivnit trajektorii viru. Hierarchické shlukování je zde také implementováno na základě přísnosti vládní reakce a byly nalezeny 3 odlišné shluky. Klastry jsou dále zkoumány s ohledem na jednotlivá opatření, které státy zavedly s cílem dalšího šíření viru.

Klíčová slova: Big Data, Python, Coronavirus, Pandemic, New Cases, Stringency Index,

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SARS-CoV-2 – Severe acute respiratory syndrome coronavirus 2

EDA – Exploratory Data Analysis

WHO – World Health Organization

ICU – Intensive Care Unit

1. Introduction

(Merriam-Webster, 2021) defines big data as an "accumulation of data that is too large and complex for processing by traditional database management tools". Despite the fact that big data has permeated our daily lives and will continue to do so with the availability of low data storage and smart devices across the globe, its effect on decision-making in many disciplines will continue to grow. (O. Austin & Kusumoto, 2016)

A new coronavirus (SARS-CoV-2) appeared in December 2019 with an acute respiratory syndrome (COVID-19) outbreak in humans, based in Wuhan, China. (Zhou, et al., 2020) The planet was taken captive by the pandemic of COVID-19, and people around the world were shocked by the rapid invasion and dissemination of this virus worldwide. Governments around the world have scrambled to get the spread of the virus under control, of which some have unfortunately not been successful. "Optimal decision making in the context of Covid-19 pandemic is a complex process that requires to deal with a significant amount of uncertainty and the severe consequences of not reacting timely and with the adequate intensity." (Alamo, et al., 2020) This is where big data comes in handy by offering useful knowledge based on the information we have about the virus, which can then be used to make important decisions to mitigate the consequences of the pandemic. Investigating the pandemic retrospectively from the its beginning is also very crucial in the much-needed preparation for the future pandemics that may emerge in our planet.

The purpose of this thesis is to examine the pandemic in order to obtain some insight on which countries had a relatively the same pattern of spread of the virus. This can be accomplished by a clustering study of new cases identified globally on a daily basis. The socio-economic effects of the pandemic on the clusters shall also be studied

It is also the purpose of the thesis do clustering analysis based on steps taken by governments around the world to address the transmission of the virus.

2. Objectives and Methodology

2.1. Objectives

The main objective of this thesis is to analyze the given data for the purpose of investigating the Covid-19 pandemic.

Partial goals of the thesis also include:

- Cluster countries based on the spread pattern
- Cluster countries based on measures taken by governments
- Studying how the pandemic affected Socio-Economic status of the clusters.

2.2. Methodology

Certain steps must be taken in order for the objectives to be accomplished. The study therefore begins with the collection of information and data on the Covid-19 subject. The following are the general steps followed which will later be used as a guidance in the practical part of this thesis.

Step-1: Selecting Technology

Python is chosen for analysis purposes for this thesis for the reason that it is an open source and has many open-source libraries that can be leveraged for the purpose of the analysis. (Python, n.d.)

For the exploratory data analysis, clustering analysis, and data visualization used in this thesis, different packages and libraries were imported. Some of them are presented as follows.

- **Pandas-** Pandas among other aspects, it provides efficient, expressive, and scalable data frameworks that enable data manipulation and analysis. (Pandas, n.d.)One of these constructs is the DataFrame or what we usually call dataset. It facilitates the process of the importing of the dataset helps on the manipulation.
- **Numpy -** NumPy is a Python library that helps us to deal with arrays. (Numpy, n.d.)NumPy arrays can be viewed and modified easily by processes.
- **Matplotlib -** Matplotlib is a Plotting library that helps us to construct fixed, dynamic, and interactive visualizations. (matplotlib, n.d.)
- **Scikit-learn -** Regression, clustering, and statistical modelling are only a handful of the valuable methods in the sklearn library for machine learning. (scikit-learn, n.d.)This particular package comes in handy for the Agglomerative Clustering activities that will be needed later in this paper.
- **Scipy** – Scipy is a python library that is based on the NumPy extension which lets us access and visualize data. (SciPy, n.d.)

Step-2: Data Gathering and Importing

The first and important step taken before gathering data is identifying the relevant variable for the analysis. After the proper variable is identified, careful data gathering is undertaken. The data is then imported uploaded to the analysis environment which in this case is Jupyter.

Step-3: Processing of Data

The raw data that is imported from sources might contain missing values, irregularities and outliers which requires us to somehow process the data to make it ready for analysis. Records containing missing values are removed from the dataset as they may case in accurate result of the analysis. Outliers and irregularities on the other hand are reduced by scaling the data which was achieved by calculating moving averages.

Step-4 Exploratory Data analysis

Data clean-up is accompanied by exploratory data Analysis, which extracts required information from the data to be used. Graphical representation of the information extracted of exploratory data analysis is used for better understanding of each dataset.

The type of data analysis to be done on the dataset, which in our case is clustering analysis, is decided in this phase after getting enough insights about the data set.

Step-5 Clustering Algorithm Selection

When choosing a clustering algorithm, we must have task's priorities, targets and the behaviour of the dataset in mind.

For the purpose of this thesis Hierarchical algorithm will be chosen for the following reasons. (Tuffery, 2011)

1. We do not know the number of clusters prior to the analysis and hierarchical clustering is suitable in such cases in that we can cut off the hierarchy at any given level depending on our choice of clusters.
2. It can uncover clusters of various shapes using different distance calculation methods

Step-6 Clustering Analysis

After choosing the right algorithm, the proper clustering algorithms, the clustering algorithm is run on the dataset with the intention of extracting groups of countries with similar behaviour you we can study the most common exhibited behaviours instead of individually examining every country around the world.

The achieve the objective of generating valid and meaningful clusters we need to be identify the optimal number of clusters. Silhouette Score is used as guidance to choose the number of clusters in this thesis.

The silhouette plot shows how similar each point in one cluster is to points in adjecent clusters, and thus allows guides to determine parameters like cluster count. This analysis is used to measure how close each object in one cluster is close to another objects in another cluster. (Ogbuabor & Ugwoke, F. N, 2018) Silhouette score on the other hand is calculated using the mean distance (a) inside the cluster and the mean distance (b) to adjacent cluster for each sample. (Scikit-learn, n.d.).

$$Silhouette\ Score = \frac{(b-a)}{\max(a,b)}$$

The value for silhouette score ranges from -1 to 1. The more Silhouette Score near 1 the clearer the cluster boundaries. (Ogbuabor & Ugwoke, F. N, 2018)

Step-7 Further examination of clusters

After creating validated cluster out of the dataset, the next is task is assessing the behaviour of the cluster. Descriptive statistics will mainly be used to describe each cluster well. Other supporting datasets will also be imported to examine the characteristics of the cluster.

Step- 8 Conclusion

Finally, conclusion will be drawn based on the finding of the analysis.

3. Literature Review

3.1. Big Data

Big data is rather a new term that originated from the need of industry leads, such as Yahoo, Google, and Facebook, to analyze large amounts of data. (Garlasu, et al., 2020) Many scholars who have researched on big data, associate big data with the storage capacity technology has reached for the given period of time. For instance Manyika et al. (J. Manyika, et al., 2011) defined big data as datasets whose size is outside the capacity of typical database software tools to capture, store, manage and analyze. This definition does not quantify the amount of data that qualifies a data to be considered as big data since the invention of new advanced technologies will mean that the big data is not big anymore.

Mashinaidze et al. (Mashingaidze, K & Backhouse, J., 2017) seem to support this definition by referring to big data as a dataset with sizes beyond the ability of regularly used software tools to capture, curate, manage and process data within a reasonable amount of time. From these definitions we can learn that with amount of data increasing in an exponential rate as it is doing now, the development of more effective softwares is important for the growth of the utilization of big data because it makes big data easier to handle and cheaper to store.

However, Gantz et al. (Gantz, J. & E. Reinsel, 2011) views big data from not just perspective of volume but also its nature of organization, diversity and application by defining big data as "a new generation of technologies and architectures, designed to economically extract the value from very large volumes of a wide variety of data by enabling the high-velocity capture, discovery, and/or analysis."

The utilization of Big Data is turning out to be normal nowadays by the organizations to beat their competitors. In many industries, existing contenders utilize the procedures resulting from the broke down data to contend, create and develop businesses. That is when big data comes in handy by providing large amount of data with variety.

Facebook	• 22% of the world's total population Every day, 100 million hours of video content are viewed, and 300 million images are updated on Facebook, which is used by 22% of the world's population. • 510,000 remarks are posted every 60 seconds, and 293,000 statuses are changed every 60 seconds. • Five new profiles are generated every second.
YouTube	There are 1,300,000,000 YouTube consumers. 300 hours of video are posted every minute. Almost 5 billion videos are viewed every day. Per day, YouTube attracts 30 million visitors.
LinkedIn	LinkedIn has a customer base of 467 million people. 3 million people post content on a weekly basis. About 19.7 million presentations have been shared on SlideShare.
Twitter	Twitter has 317 million users 500 million tweets are tweeted per day
Instagram	95 million photos are uploaded per day

Snapchat	400 million snaps are shared on Snapchat per day, 9,000 photos are shared every second

Table 1: Statistical summary of data *(source :* (Sreenivasan, 2017)

Data can be available for processing in forms of

Structured data- is an organized data in column and rows in a way that the data is easily accessible in a relational or object orientation. Structured data has various datatypes, the data are arranged in a well-defined way and the entries can interact with computer very easily. (Praveen & Chandra, 2017) This type of data can include any data put in a table with rows and columns.

Unstructured data: This type of data has no predetermined data model structure or format. Unstructured data are data that have no fixed data model, and are not put together in a fixed pre-defined manner and these kinds of data cannot be stored in a table. (Wieringa, 2016) Unstructured type of data includes simple texts and emails.

Semi structured data: Semi structured data lays somewhat between the above two. This type of data cannot be stored in a table with columns and rows but it has an orderly structure but the structure and content are put together that it has no clear shape which gives us the reason to classify the structure of semi-structured data as irregular. (Lin, et al., 2018)An XML file can be an instance in this case.

	Structured data	Semi-structured data	Unstructured data
Technology	RDBMS	XML, RDF	Text, image, video
Schema	Flexible dependent	Flexible	No predefined schema
Scalability	Difficult	Simply scalable schema	Highly scalable
Robustness	Highly robust	Not reliable for robustness	Not robust

Query handling	Structured query handling	Support textual inquiries only	Query is possibly on any mode
Version control	Easy	Not commonly used	Versioned as whole
Transaction control	Highly controlled	Transaction management adapted from RDBMS	No transaction model and n concurrency

Table 2: Comparison of different types of data adopted form (Kwok Tai Chui, et al., 2010)
Characteristics of big data

Laney (Laney, 2001) characterizes big data by 3Vs theory i.e. volume, variety and velocity. According to Laney, these three terms are defines as, **Volume**: with generation and collection of lots of data, data becomes progressively big; **velocity**: specifies suitability of big data, specifically data collection and analysis must be held quickly and timely. **variety**: the various types of data which include semi-structed and unstructured data as well as traditional structural data. Zikopoulos et al (Paul Zikopoulos, et al., 2011), supports Laney by adding the term veracity. He defines Veracity as the dependability and uncertainty characteristic in some resources of data.

1. Volume: This is the first characteristics that comes to most people's mind when they think of big data. It refers to "ever increasing amounts of data". (Ylijoki & Porras, 2016)These large amounts of data can be of sizes of terabytes to zettabyte.
2. Velocity : refers to the rate by which data is being generated. It shows the data generation rate and generation speed. (Rui Han, et al., 2014)
3. Variety: as the name suggests variety of data is of wide range. The data collected by a company can be of any of the types listed above however 90% of data generated is in unstructured form. (Panel Ishwarappaa & J.Anuradhab, 2015)
4. Veracity: refers to the uncertainty of data, also to inconsistencies found in the data. It shows the degree to which the data is trusted to make a decision based on it (HADI, et al., 2015) from which we can conclude that veracity covers origin, authenticity and Integrity of the data.
5. Value: the collected data can have all of the above characteristics but if it can't be converted to a valuable information, it is useless. (Panel Ishwarappaa & J.Anuradhab, 2015) describe big data value as "the extent to which the data collected, after analysis, can contribute to the intended purpose." So big data should be valuable enough to return the investment of storage.

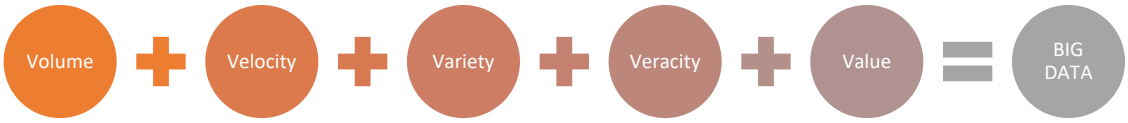

Figure 1: 5Vs of Big data. (Own visualization)

3.1.1. Big data analytics

Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information . (Verma, et al., 2016) Big data analytics is a process of inspecting, differentiating, and transforming big data with the goal of identifying useful information, suggesting conclusion and helping to take accurate decisions. . (Chunarkar-Patil & Bhosale, 2018)

Types of analytics

Every analysis starts asking a question we want our data to answer.

a. Perspective – prescriptive analytics look at a set of possible actions and recommends actions based on descriptive and predictive analyses of complex data (IBM, n.d.) It is the type of analytics that prescribes what action to take to get rid of a future problem or take advantage of .
b. Predictive - Predictive analytics uses the understanding of the past to make predictions about the future (Josh & Mujawar, 2015) From the definition we can learn that the predictive analytics of big data helps the assessment of future outcomes by identifying past patterns.

c. Diagnostic- Diagnostic analytics is an application of data analytics to investigate the causes and effects of situations. (Baum, et al., n.d.) Diagnostic analytics helps an analyst to isolate the root cause of the subject to be analyzed.
d. Descriptive - It is the simplest class of analytics that allows one to break down big data into smaller, more useful portions of information. (Josh & Mujawar, 2015)

3.1.2. Big data analytic techniques

It has already put that big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes.

Type	Methods	Description
Machine Learning	Supervised Learning	Learning models that are conditioned using labelled data points are used in supervised learning approaches to forecast upcoming scenarios. The supervised learning models are trained with labeled data points and evaluated using different approaches. For data classification and clustering, supervised learning methods are often used. However, supervised learning algorithms are limited in their ability to manage information changes in large datasets.
	Unsupervised Learning	Unsupervised Learning is a machine learning methodology in which the model does not require the user's supervision. Instead, it encourages the model to work independently to uncover previously undetected trends and knowledge. It is primarily concerned with unlabeled information.

	Semi-Supervised Learning	The semi-supervised learning models are built from labelled data points and modified in real time based on input from positively predicted events. Semi-supervised learning models' adaptive behavior allows them to deal with information change.
	Deep Learning	supervised and unsupervised learning methods are represented hierarchically in deep learning models. Deep learning models work well for vast amounts of high-dimensional data. When it comes to processing large amounts of data, deep learning models are a good option.
Data Mining	Classification	The classifiers are used to predict the object class of nominal data points and can be built with or without learning models.
	Regression Analysis	The regression analysis methods are based on statistical theories and are used to establish a relationship between given data points.
	Descriptive Statistics	The descriptive statistical methods are used to produce summary statistics using basic statistical operations over whole input data

Table 3: Types of Data Analysis Table (Rehman, et al., n.d.)

3.2. Cluster Analysis

Because of its various applications to summarization, learning, segmentation, and target marketing, the topic of data clustering has received a lot of attention in the data mining and machine learning literature. (Aggarwal & Reddy, 2014) As the name suggests clustering is grouping objects based on common attributes they have or exhibit. According to (K.Sasirekha & P.Baby, 2013) the aim of cluster analysis is to divide a set of N objects into C clusters in such a way that cluster objects are similar to each other and objects in different clusters are distinct. (Omran, et al., 2007) define

clustering analysis as Clustering is the method of detecting interesting patterns or groups within data sets based on some measure of resemblance. All of the above definitions give us the general purpose of clustering zooming out from individual objects and looking at scenarios from a bigger perspectives.

3.2.1. Types of Clustering techniques

There are various ways and algorithms in which we can group objects or data together. The following are the most commonly used clustering techniques.

3.2.1.1. Connectivity-Based Clustering (Hierarchical Clustering)

Hierarchical clustering as a cluster analysis approach that attempts to construct a cluster hierarchy. (Rani & Rohil, 2013) This constructed hierarchy of clusters has different level from which we can cut to get certain number of clusters which can explain why it is not important to pre-specify the number of clusters. Dendrogram show the hierarchical relationship between clusters , and can be modified by the user to achieve the desired clustering number. (Xiaofei Ma & Satya Dhavala, 2018). (Xiaofei Ma & Satya Dhavala, 2018) goes on to state the dendrogram layout offers a simple way to investigate the interaction between individuals at all granularity levels. "A dendrogram over a finite set X is defined to be nested family of partitions, usually represented graphically as a rooted tree." (Carlsson & Memoli, 2010) The rooted tree is graphically presented as follows.

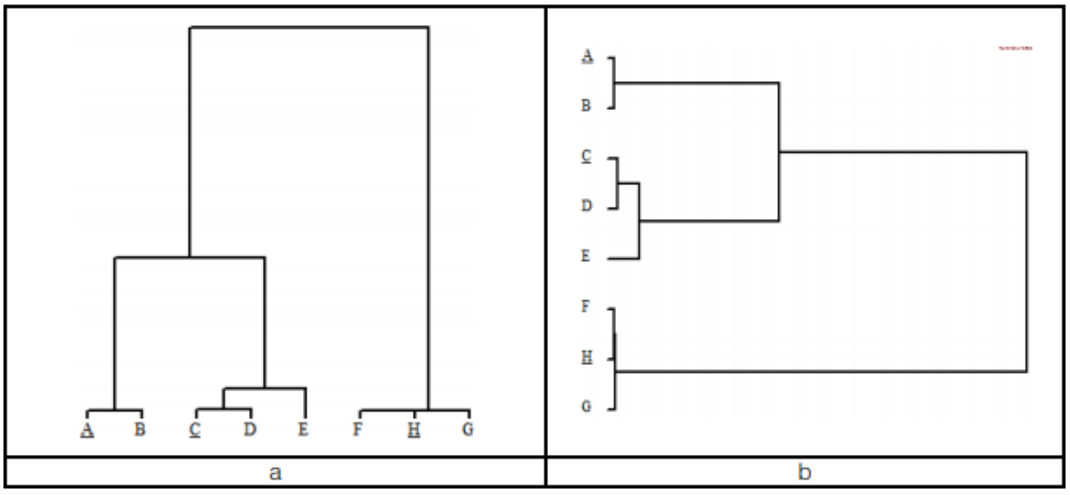

Figure 2: Vertical and Horizonal view of dendrograms (source: (Aljumily, 2016))

As can be seen from figure 2, hierarchical clustering have a vertical and horizonal orientation. In a series of steps, these labels are grouped into clusters. The letters (A,B,C,D,E,F,G and H) are representative or labels of the vectors. In a series of steps, these labels are grouped into clusters. In divisive hierarchical approach, starts with one cluster and the cluster is broken up into smaller clusters in continuous iteration. (Reddy, et al., 2017) On the contrary, agglomerative approach starts with individual objects and merges them with steps of iteration until one super cluster remains. (Reddy, et al., 2017) The most common agglomerative hierarchical clustering methods include single linkage clustering method, complete clustering method and average clustering method. In single linkage clustering method, the distance between two clusters is measured as the minimum distance between this group, meaning the distance between the closes members of this cluster is taken into consideration.

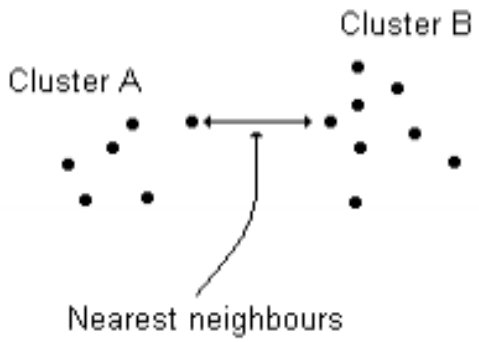

Figure 3: Single-Linkage Clustering (source: (Aljumily, 2016))

Unlike single linkage clustering, in complete clustering approach, the distance between two clusters is considered as the distance between the members of the clusters which are the furthest apart.

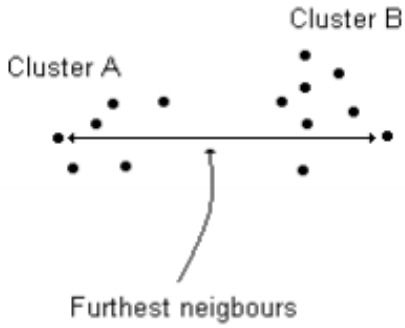

Figure 4: Complete Clustering (Source: (Aljumily, 2016))

In average clustering method the distance is calculated as the average of the distances between all the datapoints in the two clusters.

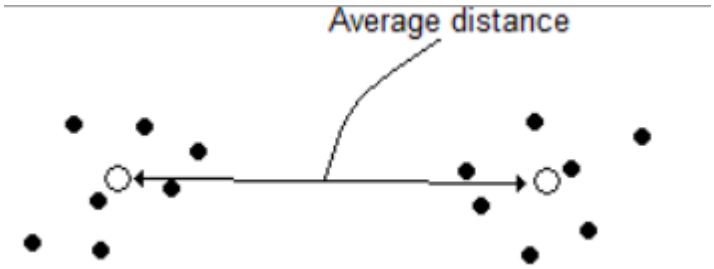

Figure 5: Average Clustering (Aljumily, 2016)

The general form of agglomerative hierachical clustering algorithm is as follows: *(Tuffery, 2011)*

Step 1. The initial clusters are the data points.

Step 2. The distances between clusters are calculated.

Step 3. The two clusters which are closest together are merged and replaced with a single cluster.

Step 4. We repeat steps 2 and 3 until there is only one cluster containing all of the data points.

3.2.1.2. Centroid Based Clustering:

The overall theme of each of the centroid-based algorithms is the feature of measuring the distance measure between the objects in the data set. (Uppada, 2014)According to (Auxiliadora Sarmiento, 2019) The distance between cluster centers is used to merge samples in center-based clustering

approaches. A medoid or a centroid may be found in the center of a cluster in this instance. K-means clustering is an example of centroid based clustering. Since any data point falls into just one partition (cluster), K-Means belongs to the class of clustering algorithms known as hard partitioning algorithms. (Abbott, 2014)

The algorithm of K-means clustering iterate as follows (Omran, et al., 2007)

1. Randomly locate the K cluster centroid
2. **Repeat**
    a. **For** each pattern Z_p in the data set **do**
        i. calculate its membership $u(m_k \backslash z_p)$ to each centroid mk and its weight $w(z_p)$
    b. endloop

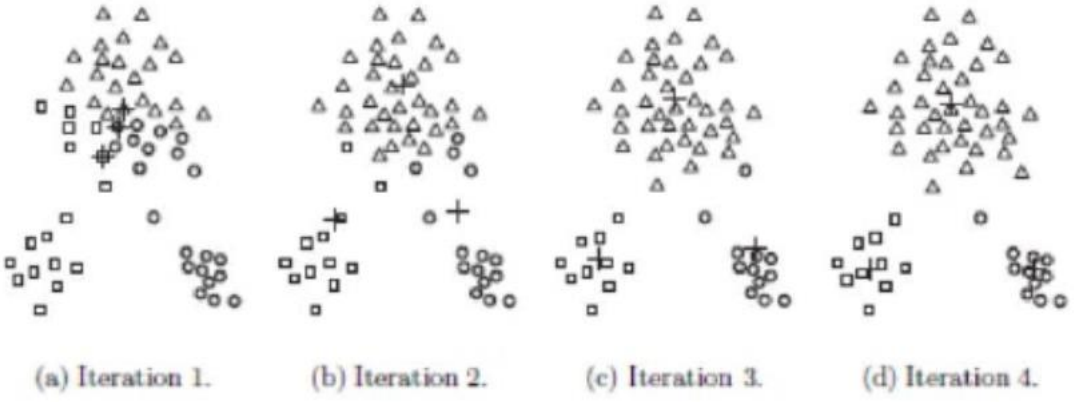

Figure 6: Architecture of K-means clustering Algorithm(Source: (Reddy, et al., 2017))

3.2.1.3. Density-based Clustering (Model-based Methods)

The discovery of clusters of entities that form neighboring, dense regions in the data field is referred to as density-based clustering. (Ester, et al., 1996) To classify clusters, they depend on spatially varying densities. Clusters of high density areas are surrounded by low density areas. Usually, the data space is explored at a comparatively high degree of granularity, and the dense regions of the data space are “put together” into an arbitrary structure using a postprocessing phase. (Aggarwal & Reddy, 2014)

3.3. Covid-19 Pandemic

A new coronavirus (SARS-CoV-2) appeared in December 2019 with an acute respiratory syndrome (COVID-19) outbreak in humans, based in Wuhan, China. (Zhou, et al., 2020) The planet was taken captive by the pandemic of COVID-19, and people around the world were shocked by the rapid invasion and dissemination of this virus worldwide. This outbreak of coronavirus disease pneumonia 2019 (COVID-19) was declared a pandemic by the World Health Organization (WHO) on 11 March 2020. (MBBS, et al., 2020)

Pathogen transmission from a vertebrate animal to a human, also known as zoonotic spillover, represents a global public health burden, which while associated with multiple outbreaks, still remains a poorly understood phenomenon. (Plowright, et al., 2017)) The full genetic sequence of SARS-CoV-2 from the early human cases and the sequences of many other virus isolated from human cases from China and all over the world since then show that SARS-CoV-2 has an ecological origin in bat populations. (WHO, 2020) Coronaviruses are positive-sense RNA viruses having an extensive and promiscuous wide range of natural hosts and affect multiple systems. Coronaviruses can cause clinical diseases in humans that may extend from the common cold to more severe respiratory diseases like SARS and MERS. (Dhama, et al., 2020) The COVID 19 pandemic has caused severe disruptions to the existing way of life. As of February 12th 2021, more than 111 million people have been infected with the virus and more than 2.46 million lives have been lost. (WorldoMeter, 2021)

3.3.1. Clinical Manifestation and Diagnosis

The most commonly reported symptoms are fever, cough, myalgia or fatigue, pneumonia, and complicated dyspnea, whereas less common reported symptoms include headache, diarrhea, hemoptysis, runny nose, and phlegm producing cough. (Adhikari, et al., 2020)

3.3.2. Coronavirus and the economy

Coronavirus has been extremely detrimental to the world economy. Numerous lockdowns, curfews, cancellation of flights and the shutting down of businesses has brought world economy a very difficult period. The impact is being felt across multiple industries including the aviation industry, hospitality industry, global financial markets and the pharmaceutical industry just to name a few. The S&P reduced its projection of the global economic growth to just 0.4%, from expected 3.3%,

which would be the slowest growth since the 1982 economic crisis. (Boshkoska & Jankulovski, 2020)

Social distance, self-isolation and travel constraints have contributed to a decline in the work force in all economic sectors and have led to a loss of many jobs. (Nicola, et al., 2020)

Responses from governments to the pandemic has introduced many lockdowns that had negative impacts to the manufacturing industry. China, which was a very important center of manufacturing of everything ranging from spare parts to heavy machinery, was the source of many raw materials for a lot of countries. For instance, 95% on the supply of electric batteries and 80% of the raw materials for the active components of a drug in the health industry come from China or Asia. Another obstacle is that dependency on raw materials or products. (Boshkoska & Jankulovski, 2020)

The service industry has also been severely impacted by the pandemic., with hourly workers facing potentially overwhelming difficulties. (Nicola, et al., 2020)

The aviation, tourism and hospitality industries have been extremely disadvantaged due to the several rounds of lockdowns that have been imposed by world governments. After the introduction of the first bans in March 2020 the air traffic reduced between 48% and 61%. (Boshkoska & Jankulovski, 2020) This happened as a result of the travel restrictions imposed by various countries around the world.

Leadership from governments is critical in the effort to mitigate the Covid-19 pandemic. However the responses variations have created debate as policymakers and publics deliberate over the level of response that should be pursued and how quickly to implement them or roll them back, and as public health officials lean in real time the measures that are more or less effective. (Hale, et al., 2020) . Economic responses include income support, debt relief for households, fiscal measures and giving international support. Health system responses were supported by public information campaigns, testing policies, contact tracing, emergency investment in healthcare and investment in Covid-19 vaccines. (Hale, et al., 2020)

4. Practical Part

The practical section of this thesis evaluates the most prevalent distribution trends shown by countries around the world, and then investigates how the pandemic impacted the socio-economic status of groups of countries that displayed the same trend. Finally, we'll look at how countries around the world responded to the virus's sudden invasion.

4.1. Examination of the Spread Pattern of Coronavirus

Lessons can be learned from the coronavirus pandemic in terms of early adoption of ways to combat pandemics and preparedness for diseases before they get out of hand. Understanding the covid-19 pandemic's spread trend in countries will help decide which countries managed the virus's spread successfully and which did not which will be critical in upcoming researches aimed at identifying the best practices for coping with a rapidly spreading pandemic.

Cluster Analysis can be used to identify clusters of countries that have encountered a similar virus outbreak trajectory in their populations. The clusters created by the clustering study are then further analysed to discover more knowledge about each cluster and its behaviour, rather than trying to analyse each nation one by one.

4.1.1. Data preparation

To achieve the objective of extracting meaningful clusters, dataset from https://ourworldindata.org/ is used. This dataset has a variable time span that ranges from January 22, 2020 to Decemper 31, 2020 and is published by Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The most relevant columns for the analysis are presented in the table below.

Column Name	Data type	Description
Date	Date	Shows the date of record
Location	String	Presents the country of the record

Continent	String	Continent of country of the record
New_cases	Integer	Absolute new cases recorded each day from January 22 2020 to December 31st 2020.
New_cases_per_million	Integer	New cases divided by a million of a given population from December 31st , 2020

Table 4 Data Set Description (Daily confirmed cases. Data source: https://ourworldindata.org/)

4.1.1.1. Cleaning up the dataset

Dataset clean up is a crucial step of clustering analysis as any missing values and/or outliers can easily alter the structure of the final cluster. Outliers can disturb a structure by forming cluster of their own because they do not belong to other clusters. The following two steps are taken to clean up the dataset in order to get more accurate clusters at the end.

1. Any records with missing value of Country are removed as the main objective of this analysis is to group countries together.
2. Continental values are removed as these entries are observed to contain regional (continental) data as opposed to country level data that is needed for our analysis.
3. The dataset is scaled just before the clustering algorithm is run so that we have more standardized data. The scaling is done by calculating moving 7day moving average. of each country which helps with the smooth out of outliers and inconsistencies which may alter the outcome of our analysis.

4.1.2. Exploratory Data Analysis

Exploratory data analysis is performed on the aforementioned dataset in order to get a better understanding of data dynamics and to identify predicted outliers or anomalous events. The data

collection was analyzed using multiple EDA methods, and the results were visualized as follows to provide a more detailed image of the COVID19 pandemic.

4.1.2.1. Moving Average of New Cases

The worldwide twenty-day new cases moving average is calculated for the purpose emphasizing long-term trends of the spread of the virus.

Although the rate varies from time to time, it can be seen from figure 7 that new cases show the trend of increasing from the start of the pandemic to the end of the year 2020. The steep slop between mid-March and the end of May and between October and January 2021, show that the number of new cases worldwide rose very sharply.

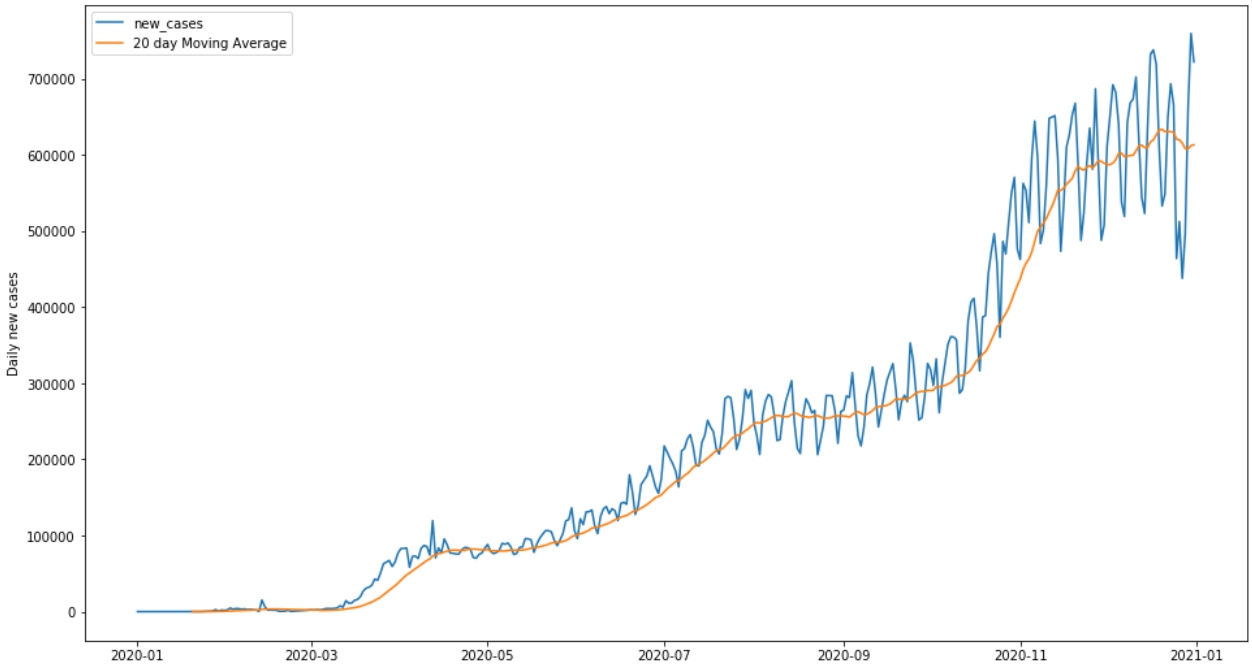

Figure 7: 20 day moving average of world wide cases (Data source: https://ourworldindata.org/)

4.1.2.2. Top 10 countries with the highest total number of coronavirus cases

The countries with the most reported Covid-19 cases are depicted in figure 8. With a total reported Covid-19 cases of almost 20 million people until the end of 2020, the United States is at the top of the list. India's Coronavirus infection rate also surpassed 10 million, making it the world's second most infected area. Brazil closely follows by having total cases of well above 6 million cases. The

rest of the top ten countries with the most coronavirus cases are the Russia, France, United Kingdom, Italy, Spain, Germany, and Colombia.

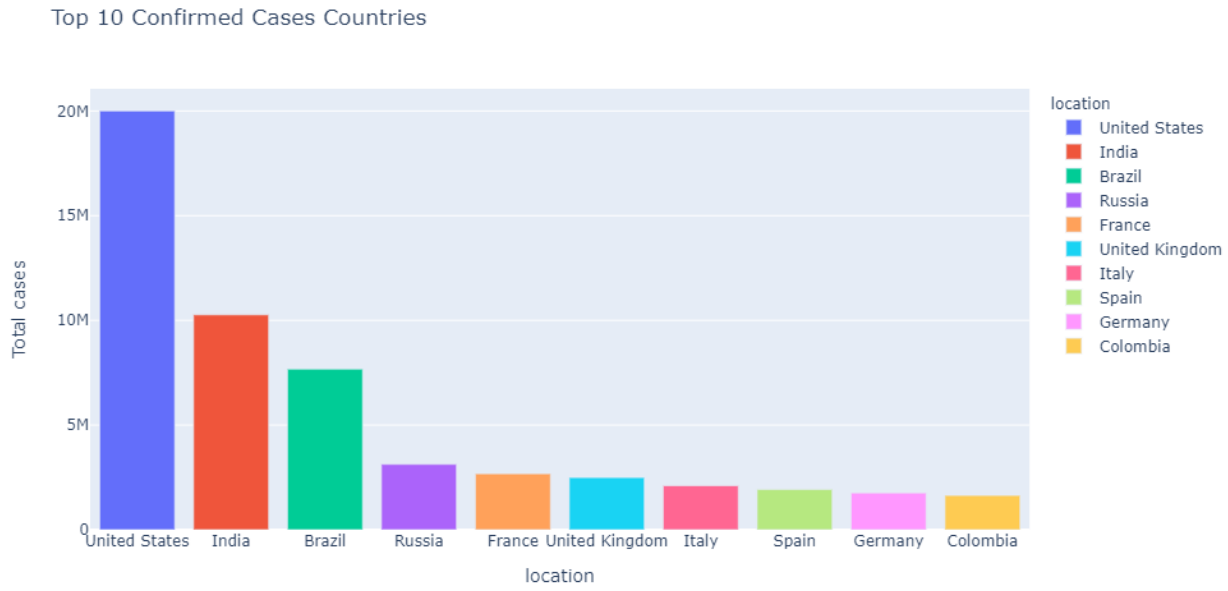

Figure 8: Top 10 countries with the highest total number of coronavirus cases (Data source: https://ourworldindata.org/)

4.1.2.3. Top 10 countries - per million residents

The total number of cases per million people in a country determines how widespread the disease is at a given moment in time. The top three nations are Andorra, Montenegro, and Luxembourg. The rest of the top ten hardest affected locations on per million people bases are San Marino, Czechia, the United States, Slovenia, Panama, Georgia and Liechtenstein.

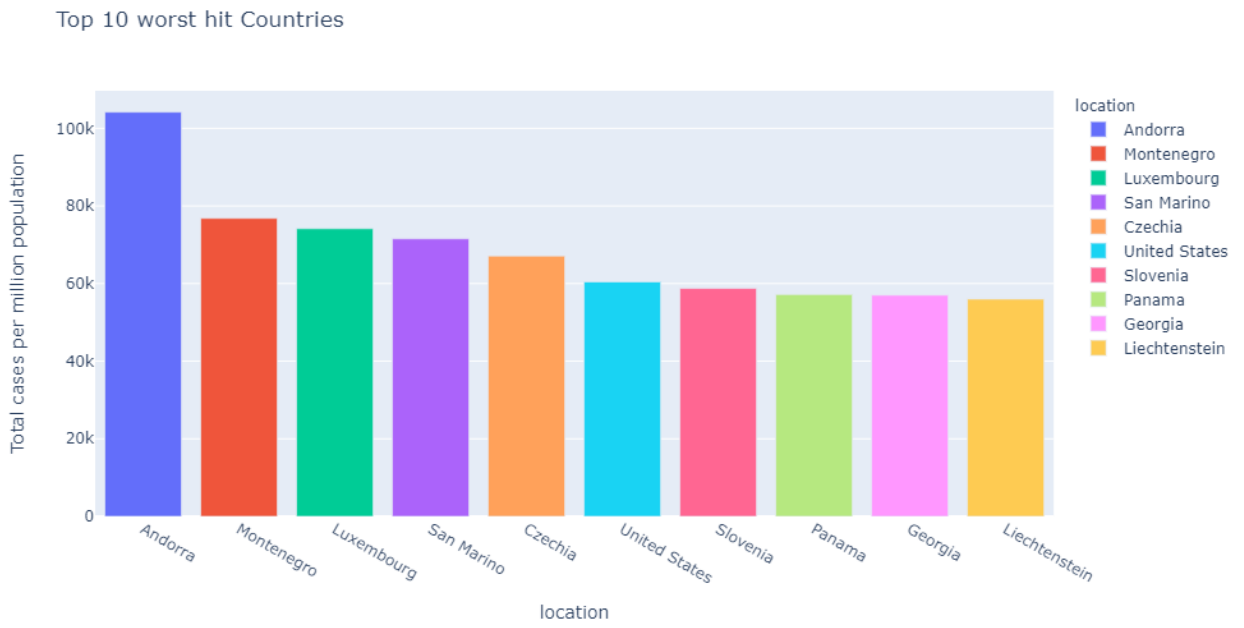

Figure 9: Top 10 countries with the highest total number cases of per million residents (Data source: https://ourworldindata.org/)

4.1.2.4. Total Cases worldwide

The graph below illustrates the increase in the number of people infected with coronavirus from the start of the pandemic to December 31st, 2020. At the end of 2020, the cumulative number of cases have surpassed 80 million.

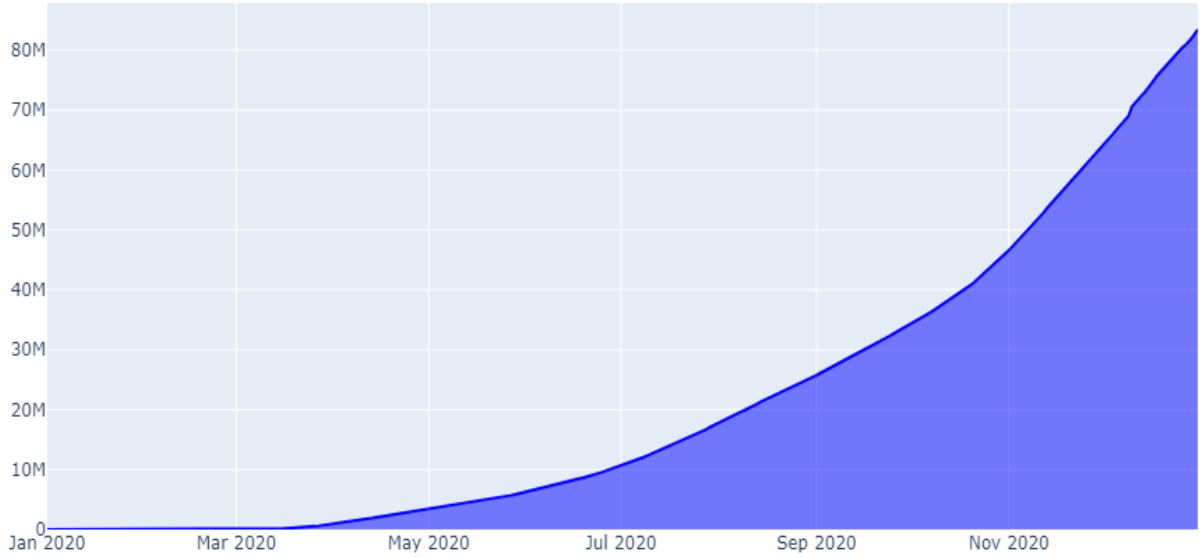

Figure 10: Total cases worldwide (Data source: https://ourworldindata.org/)

4.1.2.5. Case distribution by continent

The pie chart below displays the contribution of individual continents to the overall combined Covid-19 cases. Europe, which accounts for 28.8% of all positive cases globally, tops the other continents in terms of overall cases. North America comes in close second with 27.8 percent, while Asia comes in third with 24 percent. South America, Africa, and Oceania combine to account for the remaining 19.38% of overall total cases of Covid-19.

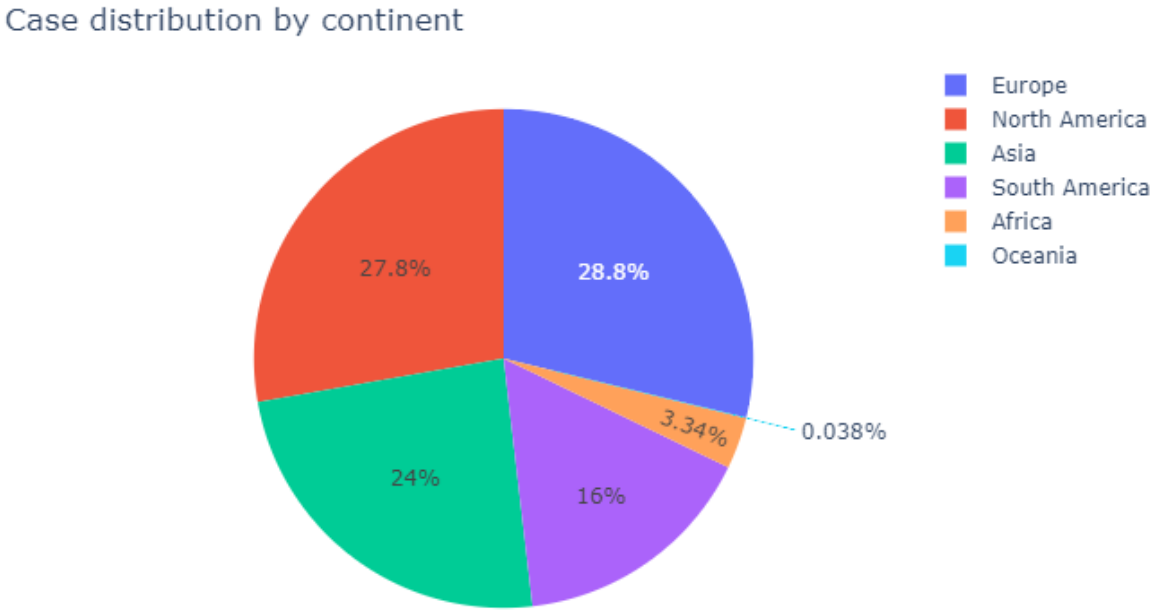

Figure 11: Case distribution by continent (Data source: https://ourworldindata.org/)

4.1.3. Variable Selection

The number of daily confirmed cases reported by each country can be used to evaluate the virus's spread pattern, as it indicates the number of infected people in a specific location. These daily incremental values of new cases can be studied to evaluate the pattern of the pandemic's progression in each country. We utilize new confirmed cases of each country for the clustering

analysis in two different ways. The first is the absolute number of newly reported coronavirus cases and the second is the number of new cases divided by 1 million people in the country.

Since we are comparing completely different population sizes, merely attempting to analyze the absolute number of new cases cannot provide us with much information. This is because there might be data imbalance that may occur from big populations reporting many more cases than smaller countries where the virus has affected large percentage of the population.

For subsequent cluster analysis, the variable "New cases per million" is chosen for the reason that it is a relative number that also takes the size of the residents of the country into consideration. In conclusion, we will use the following variables for the purpose of the clustering analysis

$\mathbf{X_1}$ => The name of country (Location)

$\mathbf{X_2}$ => Date

$\mathbf{X_3}$ => New cases per million

4.1.4. Extracting clusters

As shown in the methodology part of this diploma thesis, hierarchical clustering algorithm is chosen to be run on the chosen variables. The graphical representation illustrates the truncated form of the algorithm's output.

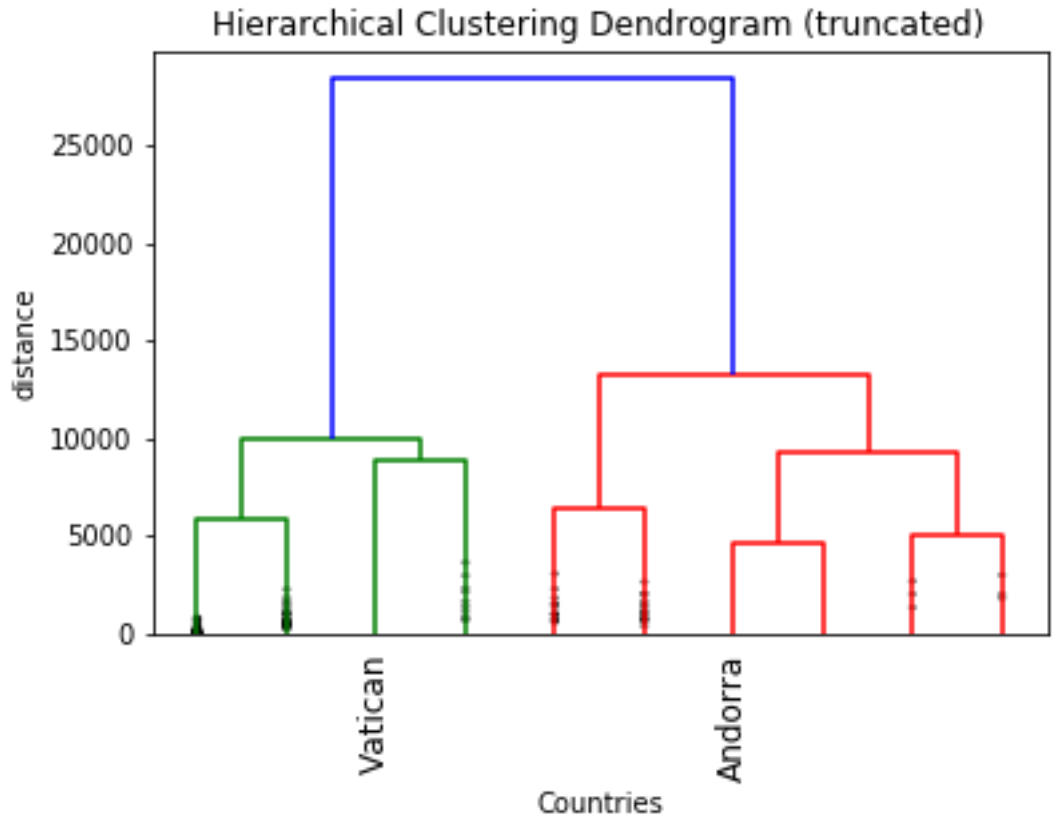

Figure 12: Truncated Dendrogram (Data source: https://ourworldindata.org/)

The above diagram shows the truncated structure of the clusters where the initial set containing all the countries is split in each step until we reach individual countries at the bottom. The vertical axis shows the distance between each cluster. It can be seen from the figure that Vatican and Andorra have a distinct distance that they would form their own clusters if the dendrogram was to be cut at a distance of just above 5000. To decide the cutting pint, i.e to decide the number of clusters, we will use a technique called silhouette scoring which is shown in detail again in the methodology section of this thesis. The figure below shows the Silhouette Score for each potential number of clusters.

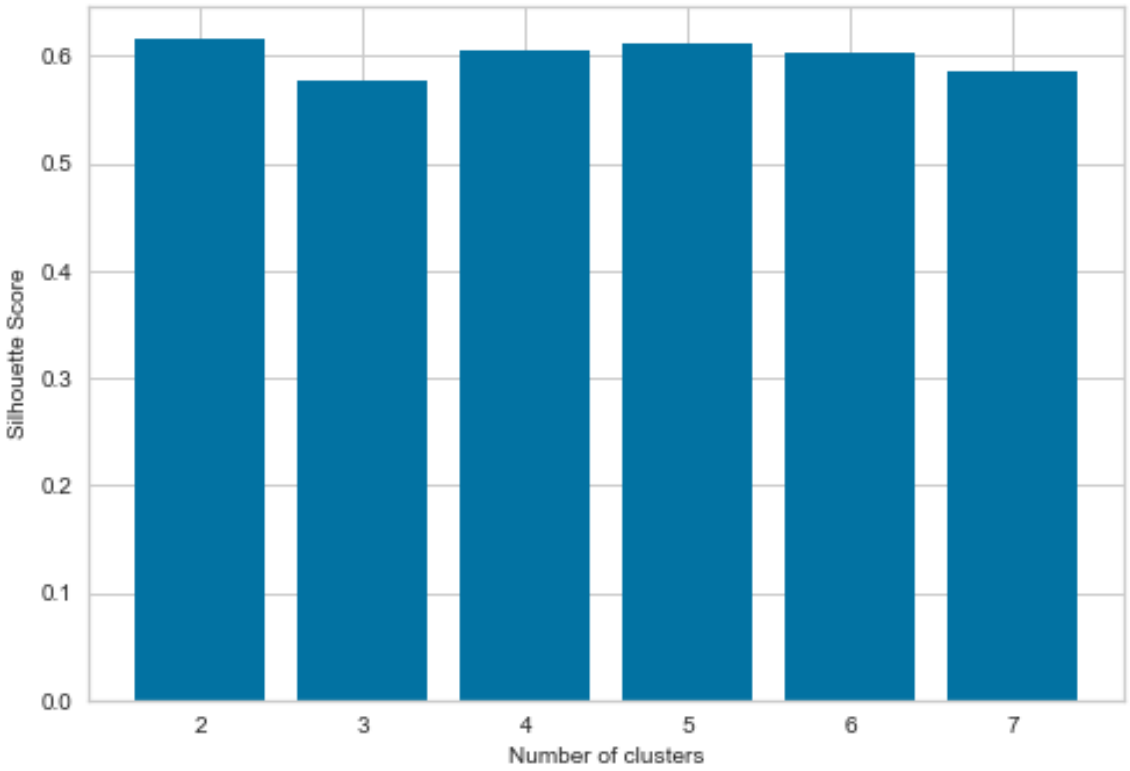

Figure 13: Silhouette Score (Data source: https://ourworldindata.org/)

Based on Figure 8 we can learn that we can get clusters of similar quality if we have 2 or 5 clusters. A cluster number of 5 will be chosen so we will be able to dissect each chunk for further investigation instead of just having two large distinct clusters.

After deciding the number of clusters, we go back to the truncated dendrogram diagram and decide a cutting point that would result in the generation of 5 clusters. In our case, cut off distance of 9000 is used to generate our intended 5 clusters.

Next, the graphical representation of each cluster will be shown and judgments will be made based on observations of the graph. The vertical axis of all these graphs presents the daily 7 day moving average the countries clustered together. The complete list of the cluster with their respective member location can be found in Appendix 1.

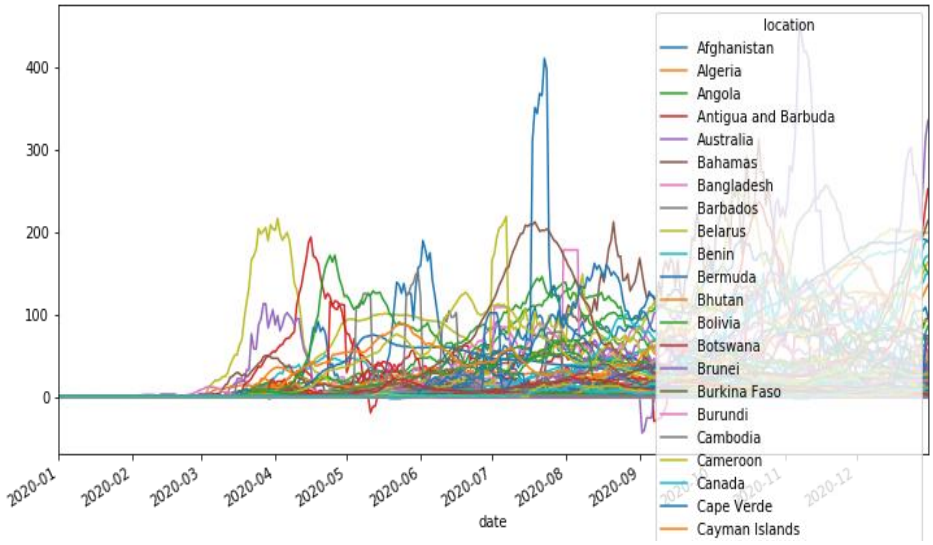

*Figure 14: Spread pattern **cluster -1** (Data source: https://ourworldindata.org/)*

Cluster one is composed of 140 countries making it the most populous cluster. This cluster showed a maximum 7 day moving average of just over 300 throughout the year of 2020 except for some outlying instances in August and November.

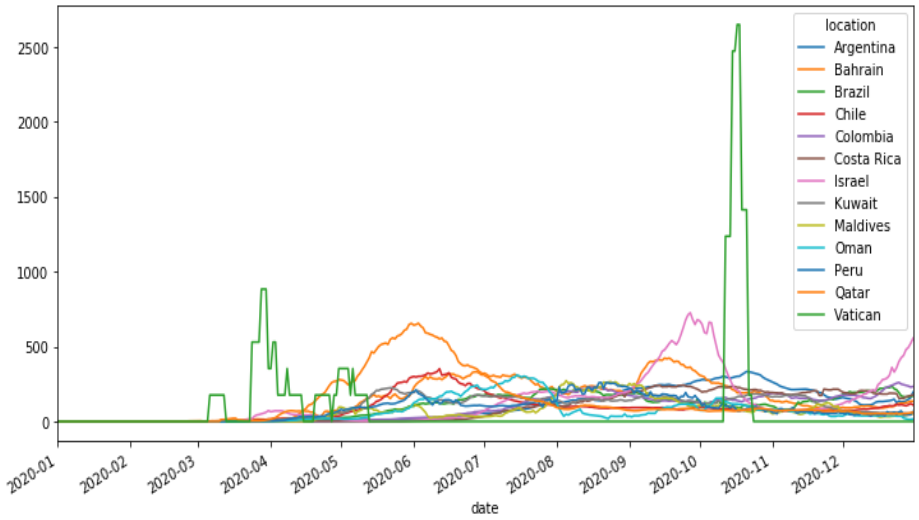

*Figure 15: Spread pattern **cluster -2** (Data source: https://ourworldindata.org/)*

Cluster 2 is composed of 13 countries. Except for some infection spurts observed in end of March and in October, this cluster has shown a relatively constant number of per million population infections. Vatican appears to be a heavy outlier in this group in that it showed reported zero cases for the majority of the year of 2020 and showing a 7 day moving average of per million cases as high as 2500.

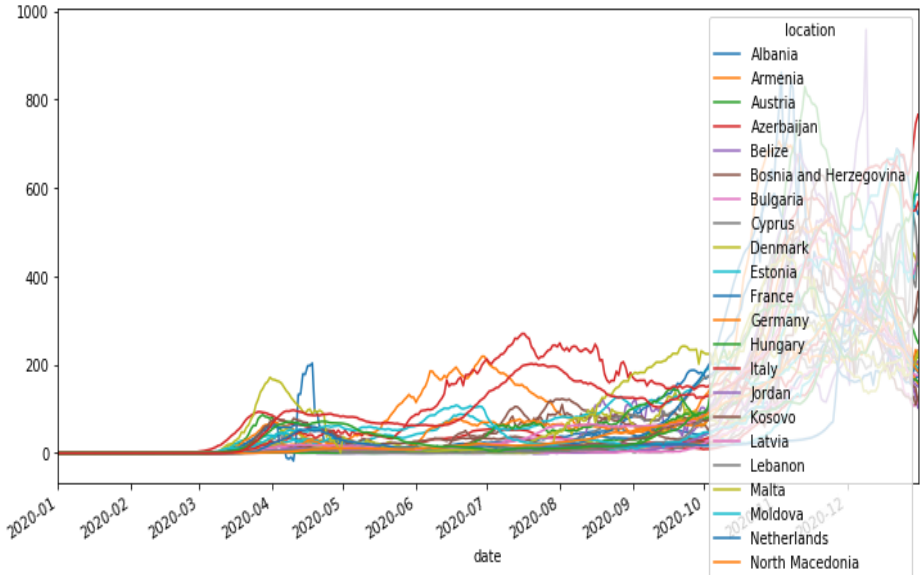

*Figure 16: Spread pattern **cluster -3** (Data source: https://ourworldindata.org/)*

Cluster 3 cluster has 35 members. Even if the cluster has kept moving average of daily per million infection below 400 until the fall of 2020, cases started to take off after October where some of the members reached daily per million cases of more than 900. This increase in infection happened in the fall seems to slow down later the same year for this cluster.

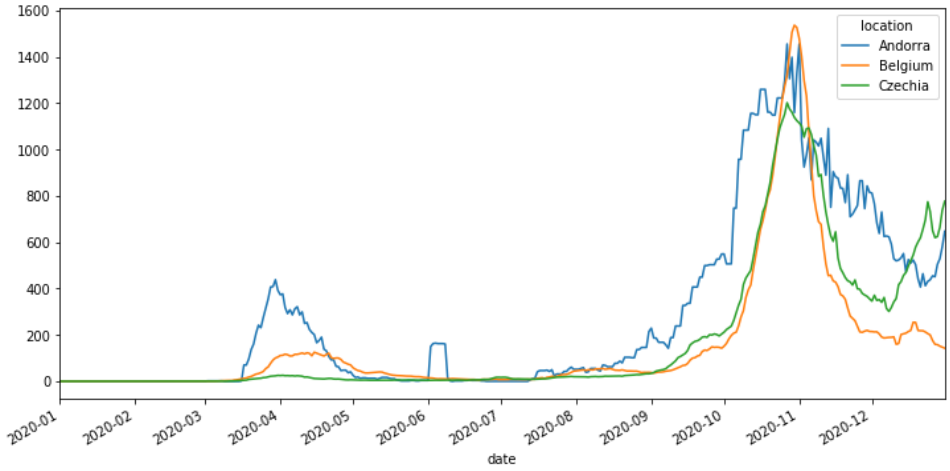

Figure 17: Spread pattern cluster -4 (Data source: https://ourworldindata.org/)

Cluster 4 is made up of just 3 countries. For this cluster, daily per million cases were extremely low from the beginning of the pandemic to the end the summer of 2020. However cases sharply rose around September where member of cluster reached a 7 day moving average of per million cases of 1400 in October.

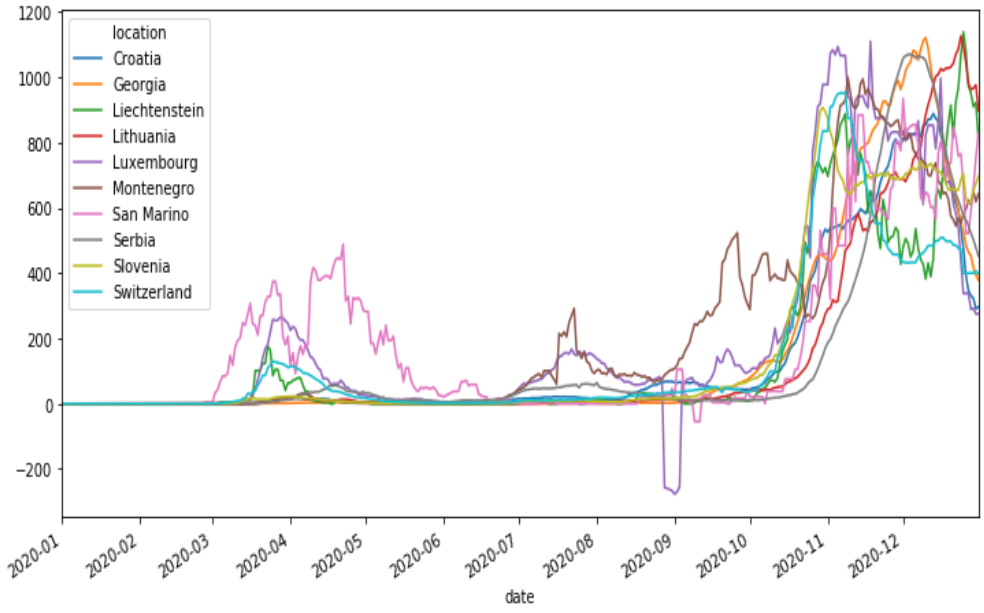

Figure 18: Spread pattern cluster -5 (Data source: https://ourworldindata.org/)

Cluster 5 has 10 members. The trend of the cluster looks a bit noisy where there is no certain pattern on the increase and decrease of cases. However, it can be seen clearly that cases started increasing

by the mid of October. And after fall, increasing infection rated did not decrease in the rate we have seen in cluster 3 and 4. The most notable scenario in this section is the fact that it has a negative moving average of new per million case the around September 2020. This is because of some countries performing data reconciliation or auditing consolidation activities that result in massive numbers of cases or deaths being removed from overall records. (WHO, 2020)

4.2. Impact Covid-19 on the Socio-Economic Status of clusters

The COVID-19 pandemic, according to the United Nations Development Programme (UNDP, 2020), is something more than a health emergency; it is wreaking havoc on communities and economies, but the scale of the consequences can vary from country to country.

In this portion of the thesis, we will look at the Socio-Economic impact of COVID-19 in terms of clusters we have extracted based on spread pattern. The effect of the COVID-19 crisis on populations, markets, and disadvantaged communities must be assessed in order to advise and adjust government and partner responses. (UNDP, 2020) . However, (Howe, et al., 2012) states that there will be no standard metric of Socio-economic status that is suitable for all experiments and contexts; the strengths and weaknesses of a given indicator will most likely differ depending on the study issue. According to American Psychological association (APA, 2015), Socio-economic status can be examined based on individuals' income, occupation and educational achievement. By taking the APA assessment as an inspiration, we will try to look at the impact of Covid-19 from three Socio-Economic indicators namely;

1. The economy
2. Education
3. Health

4.2.1. Impact of Covid-19 on the Economy

In this section we will try to check if that the spread of the pandemic in terms of case numbers and economic factors. We can divide the economic effects in terms of the following indicators;

1. Growth rate
2. Unemployment

4.2.1.1. Impact on Growth Rate

To study how the growth rate of the clusters was impacted by the pandemic, we can plot the growth rate of each cluster in 2020 and 2019 and observe if there is any noticeable difference.

Data description

Column	Description
Country	The name of the country
Code	Code that represent the country
2020	The growth rate of the country in the year 2020
2019	The growth rate of the country in the year 2019

Table 5: Dataset description - Growth rate by country (Data source: https://www.imf.org/en/Publications/WEO/Issues/2021/01/26/2021-world-economic-outlook-update)

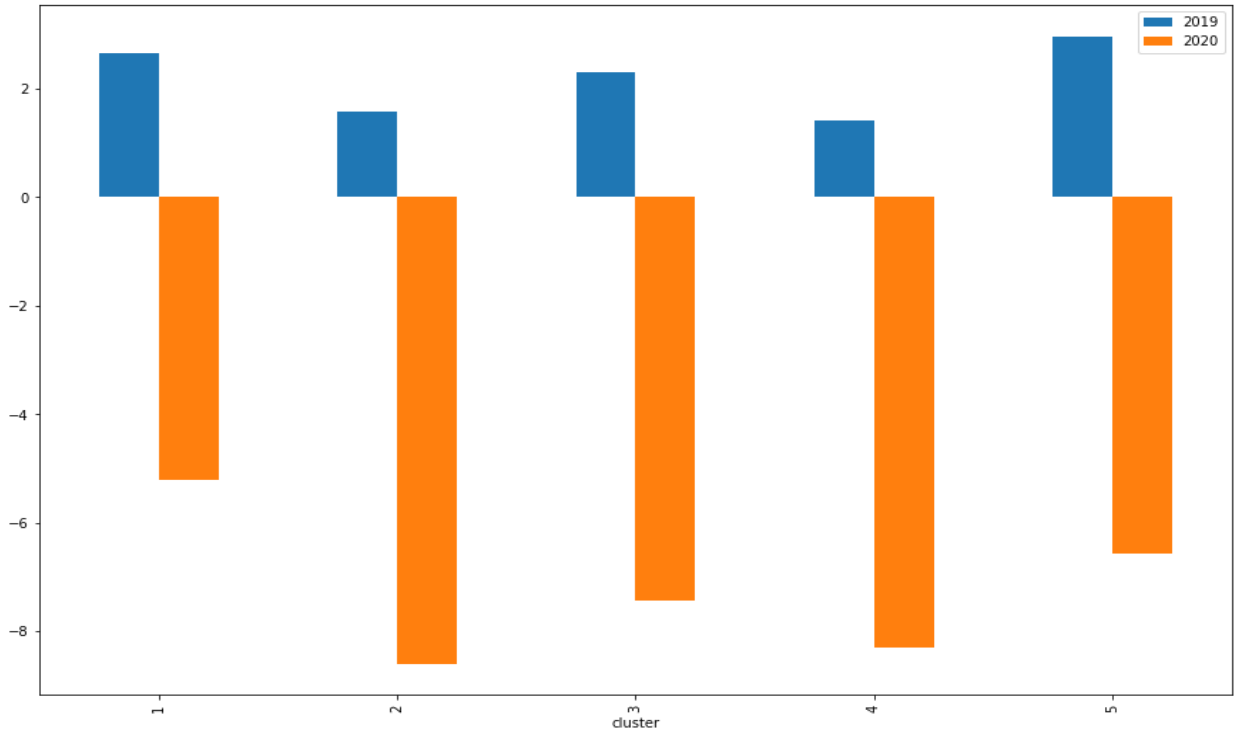

Figure 19:Growth rate comparison between 2019 and 2020 (Data Source: https://www.imf.org/en/Publications/WEO/Issues/2021/01/26/2021-world-economic-outlook-update)

As can be seen form Fig 19, Cluster one and three had an average of growth rate more than 2 % in 2019. However the average reduced in these clusters to reach an economic contraction of 5.2% and 7.4% respectively in 2020. Cluster two and four both had an average growth rate of more than 1% in 2019 which sunk to an average decrease of 8% in the year of 2020. Cluster 5 had also an average increase of growth rate of 2.5% in 2019 only to have its average reach contraction of 6.56% by 2020.

From the above analysis, we can see that the pandemic has brought an unfavorable effect in all of the clusters given that each cluster recorded a negative average growth rate. However, cluster 2 appears to the hardest hit cluster in terms of growth rate given that it recorded the highest decrease of average growth rate.

4.2.1.2. Impact on Unemployment

The impact of the economy can also be seen from the perspective of unemployment as it represents an economy's failure to provide jobs for those who wish to work but are unable to do so, despite the fact that they are available and actively finding work. (International Labour Organization, n.d.).

The aim of this section of the thesis is to look at the effect of the coronavirus on the shift in unemployment rates in individual countries between 2019 and 2020.

The reason that only 2020 data is not used in this study is that countries that already had a high unemployment rate before the outbreak of coronavirus may skew the data from countries that experienced a large increase in that year.

Data description

Column	**Description**
Country	The name of the country
Code	Code that represent the country

2019	The Unemployment rate of the country in the year 2020
2020	The unemployment rate of the country in the year 2019

Table 6: Dataset Description: Unemployment (Data Source: https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS)

The examination starts with deducting each country's unemployment rate in 2020 from its 2019 corresponding value to calculate the change in unemployment rate. After obtaining the change in the unemployment rate from 2019 to 2019, descriptive statistics is used to analyse how each cluster performed in 2020 compared to 2019.

Figure 7 shows the change in unemployment rate recorded in the year 2020.

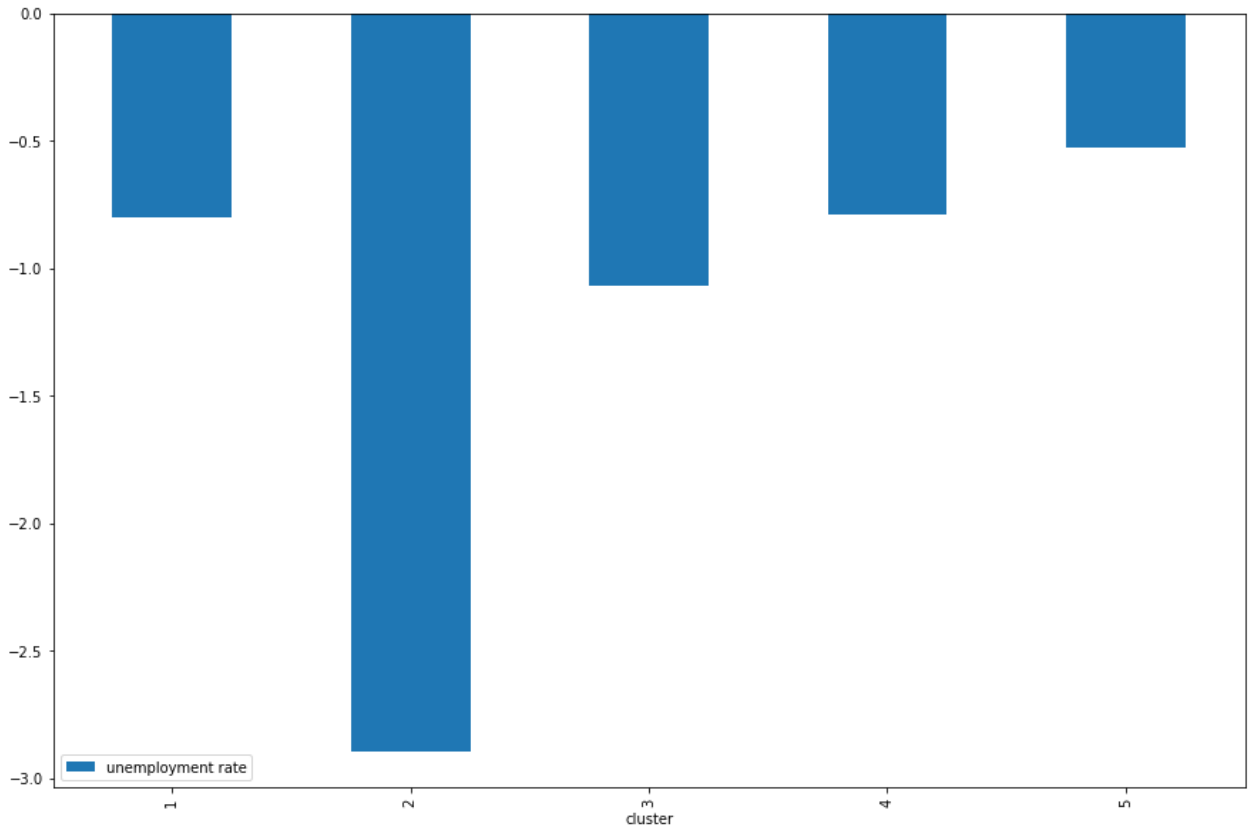

Table 7: Dataset Description: Unemployment (Data Source: https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS)

The above figure clearly shows that all the clusters showed an average increase in unemployment. Cluster 2 appears to have been hit by the pandemic hard with average unemployment rate within the cluster increasing by 2.89%.

4.2.2. Impact on Education

Many countries have shut schools, colleges, and universities as a result of the COVID-19 pandemic health crisis. (Burgess & Sievertsen, 2020) By the 7th of May 2020, 177 countries' schools and universities were already suspended, affecting 1 268 164 088 students, or 72.4 percent of overall enrolled students. (Techson, 2020). For this reason, some educational institutions switched to online learning method which comes with its requirements. Since online learning is entirely reliant on electronic devices and the internet, teachers and students with poor internet connectivity can be left out. (Adedoyin & Soykan, 2020)

In this portion of this thesis, we will assess how clusters did in regards with the closing down of schools in the year 2020. It is also bringing the individual internet availability in the cluster to the picture to assess if online teaching can be feasible.

The data Oxford COVID-19 Government Response Tracker (**OxCGRT**) compiles widely accessible data on government reaction metrics such as school closures are used in these indicators will be used for the purpose of this analysis. (OWID, 2021) This data ordinally ranks the strictness of policies taken by governments around the world in the following manner (OWID, 2021)

School closures:

0 - No measures

1 - recommend closing

2 - Require closing (only some levels or categories, eg just high school, or just public schools)

3 - Require closing all levels

No data – blank

Dataset Description

Column	Description
Entity	The name of the country
Code	Code that represent the country
Day	The date of the data record
School closures	The rate of the school closure policy (0, 1, 2,3 and no data)

Table 8: Dataset description: School closures(Data source: https://ourworldindata.org/)

Figure 20 presents the daily median of the school closure level of each cluster. It summarizes how the policy of the cluster countries on school closures changed throughout the year of 2020 to get some insight out on how each cluster behaved in terms of school closure.

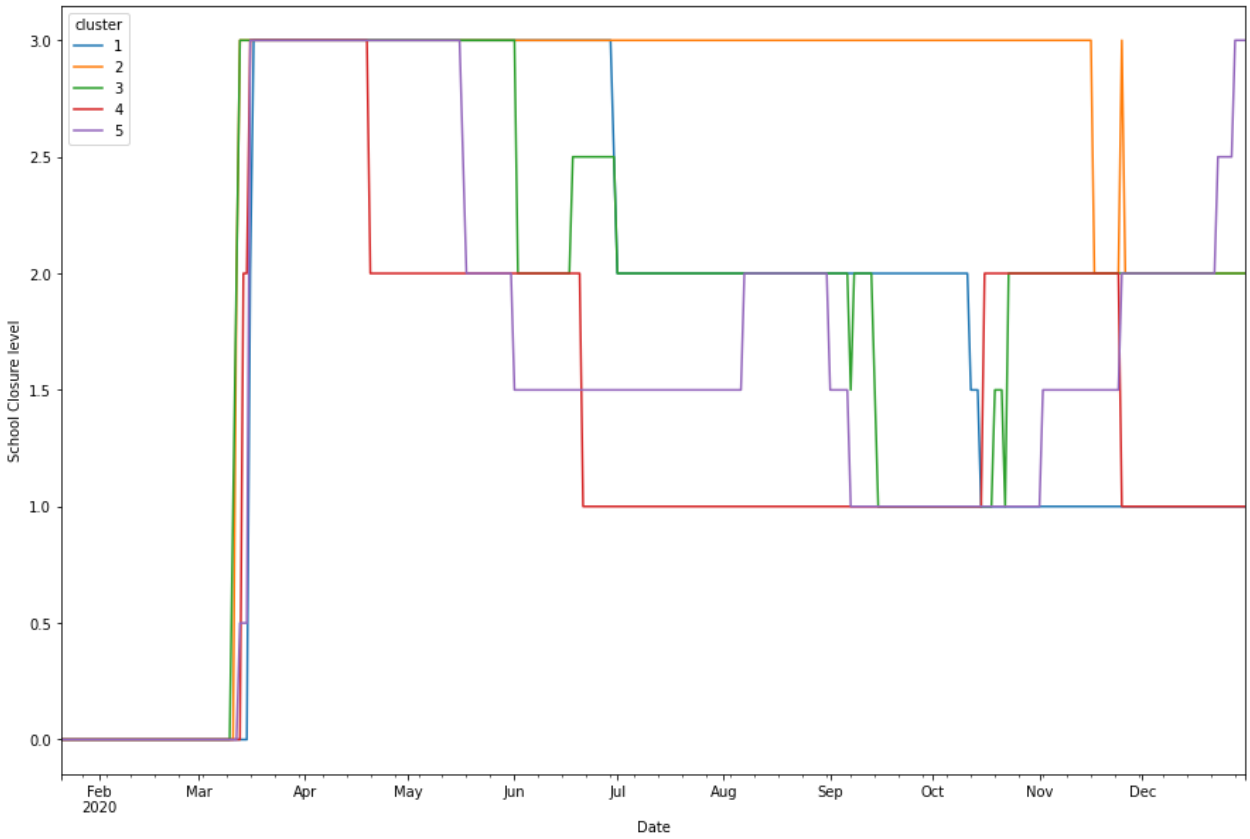

Figure 20: School closures of each cluster (Data source: https://ourworldindata.org/)

Figure 20 shows that all levels of schools were closed in at least half of the members of all of the clusters in mid-march. However how long the shutdown stayed varies from cluster to cluster. Cluster 2 showed the longest closed down of all levels of schools where at least half of its member countries closed all levels of schools up until November of 2020. Cluster 1 and 4 loosened the restriction and relaxed it to the closed down of some levels of schools (level 2) by late May and July respectively. Cluster 3 and 5 also loosened their restriction to level 2 in June

The fact that schools were recommended to be closed in at least half of the members of all clusters through the year 2020 begs the question how students were able to continue their studies from home as internet infrastructure is not available for everyone for everyone. From data that is collected from world bank, let us see how the clusters are doing regarding internet coverage throughout the countries.

Dataset Description

Column	**Description**
Country	The name of the country
Internet coverage	Country internet coverage by %

Table 9: Dataset Description: Internet coverage by country Data Source: https://datacatalog.worldbank.org/individuals-using-internet-population

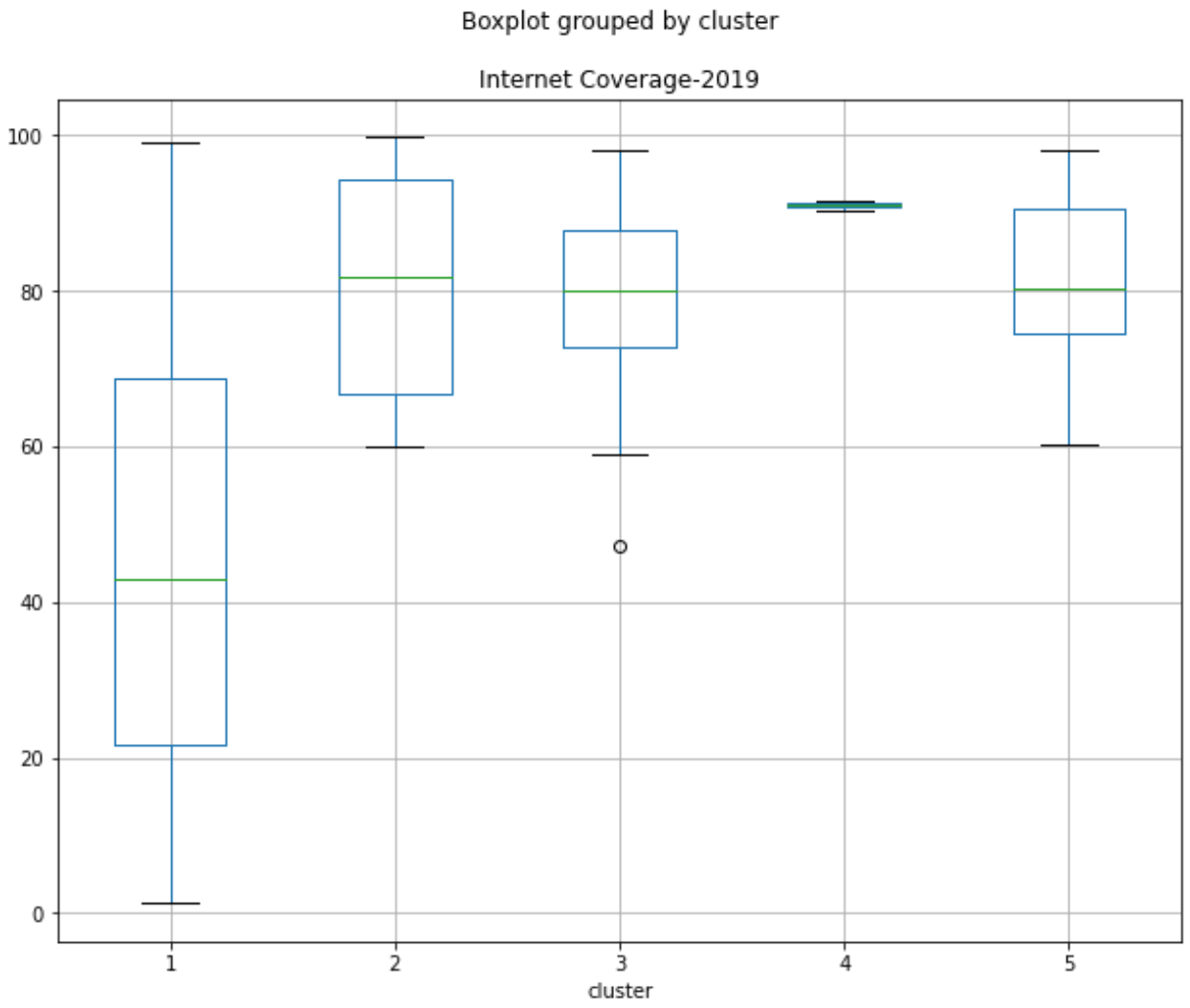

Figure 21: Box plot - Internet coverage by cluster (Data Source: https://datacatalog.worldbank.org/individuals-using-internet-population)

The above data shows the range, median and quartile information of a given the clusters. We can take cluster 1 for example, where internet coverage stands at only 45% average for the year 2019. It is also important to note that 50% of the member countries of this cluster has an internet access of below 45%. For the information that we extracted in figure 20 on average, this cluster experienced closing down of some levels of schools for half of the year of 2020. These students are highly likely to suspend their studies if it is not given in person. The rest of the clusters have relatively higher coverage of internet with lower 50% of their members population having 60% to 80% internet availability.

From this analysis we can safely conclude covid-19 jeopardized the quality and availability of education by forcing students to adapt to unprecedented learning situation, even more so by forcing some of them to suspend their studies.

Impact on Health Services

As we have seen form the EDA done above, more than 80 million people have been affected by the virus for the year 2020 and the world was seeing daily new cases as high as seven hundred thousand by the end of the year. The surge of too many critically ill patients in such a brief amount of time has almost saturated hospital capacity. (Soria, et al., 2020) This overwhelming of the hospitals causes a significant percentage of healthcare workers to have elevated levels of depression and anxiety, as well as low levels of well-being. (Mahyijari, et al., 2020).

Moreover, based on the experiences of countries who experienced the first coronavirus wave, many countries have implemented health strategies and tailored their services to the fact of this global health emergency. (Buheji, et al., 2020) For example, “To free up critical hospital personnel and beds, all non-emergent and elective surgeries and procedures were cancelled” (Kaye, et al., 2020).

The section aims to explore the COVID-19 hospitalization and ICU admission estimate the burden taken by the healthcare facilities of each cluster.

Dataset Description

Column	**Description**
Entity	The name of the country
Code	Code that represent the country
Day	The date of the data record
Hospitalization	Weekly number of hospitalized infected people
ICU admission	Weekly number of ICU admitted people

Table 10: Dataset Description: Hosptalization and ICU admission rate (Data source: https://ourworldindata.org/)

The following 5 charts show the plot of average weekly new hospitalizations and weekly new ICU admissions in each cluster.

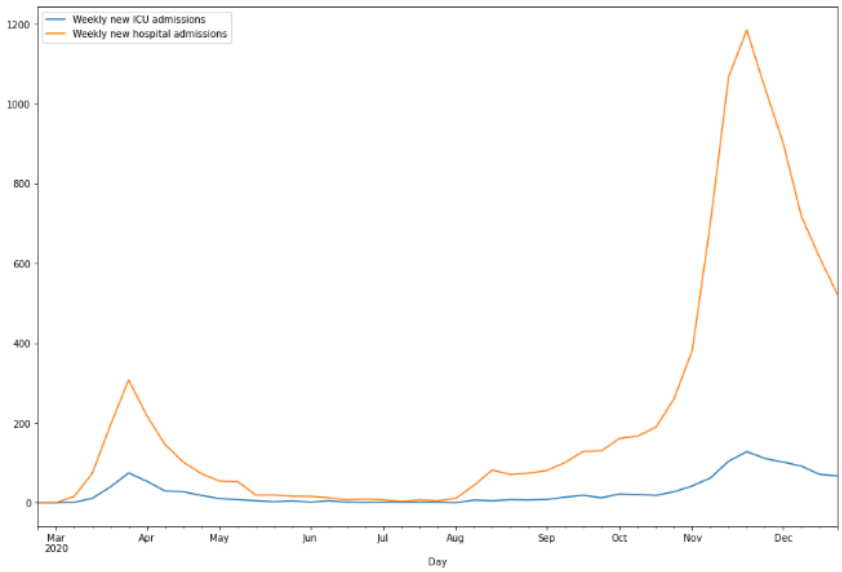

Figure 22: Cluster 1 average daily Hospitalization and ICU admission per million population basis (Data source: https://ourworldindata.org/)

Cluster 1 has kept it under average new weekly hospitalization of 400 people however the numbers increased to almost 1200 average new weekly hospitalizations by the end of November. Weekly ICU admission in this cluster was kept under 100 until late November where admission reached its peak in 2020.

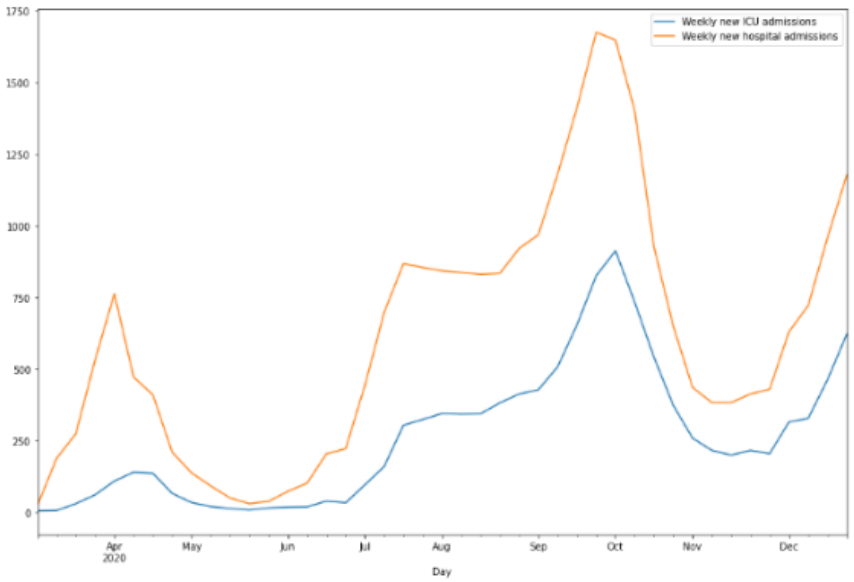

Figure 23: Cluster 2 average daily Hosptalization and ICU admission rate per million population basis (Data source: https://ourworldindata.org/)

Cluster two showed much higher hospital admission than cluster one throughout the year 2020. This cluster exhibited an increased number of hospitalizations in April which quickly decreased in May. Hospitalizations and ICU admissions where the highest in this cluster in late September where they surpassed average weekly numbers of 1500 and 750 admissions respectively. It is also important to show that this cluster shows a very high hospital admission to ICU admission ratio showing most patients admitted in hosptial are also going to the ICU.

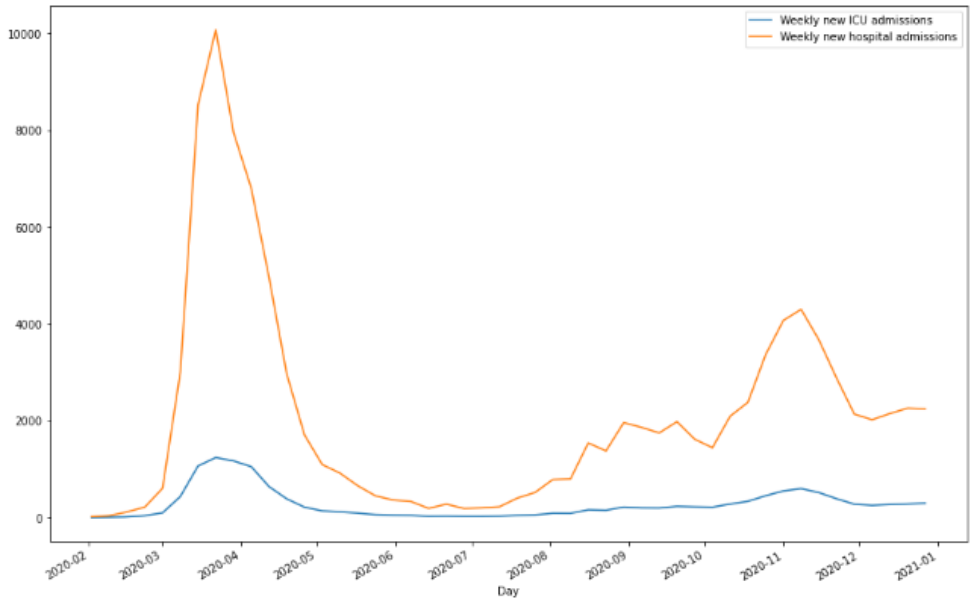

Figure 24: Cluster 3 average daily Hosptalization and ICU admission rate per million population basis (Data source: https://ourworldindata.org/)

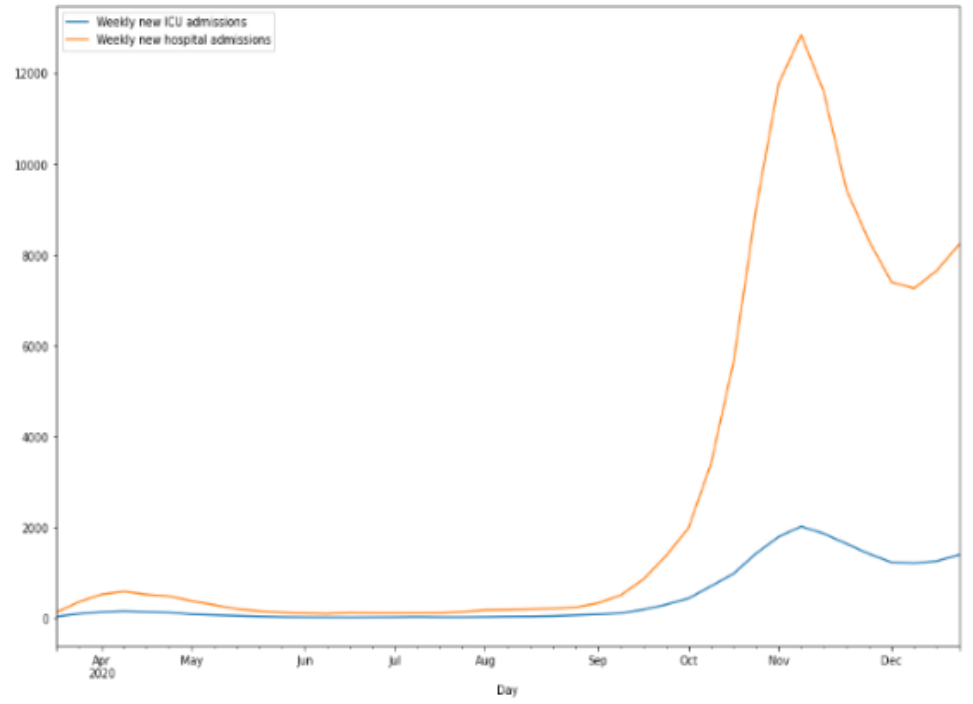

Figure 25: Cluster 4 average weekly Hosptalization and ICU admission rate (Data source: https://ourworldindata.org/)

Cluster 3 and 4 have exhibited high number of new hospitalizations at times almost reaching 10 thousand and 12 thousand new average hospitalizations per week. The infection for cluster 4, average weekly new ICU admission even reached average of 2000 in October of 2020. Hospital admission to ICU admission ratio is lower in this two clusters as compated to cluster 4.

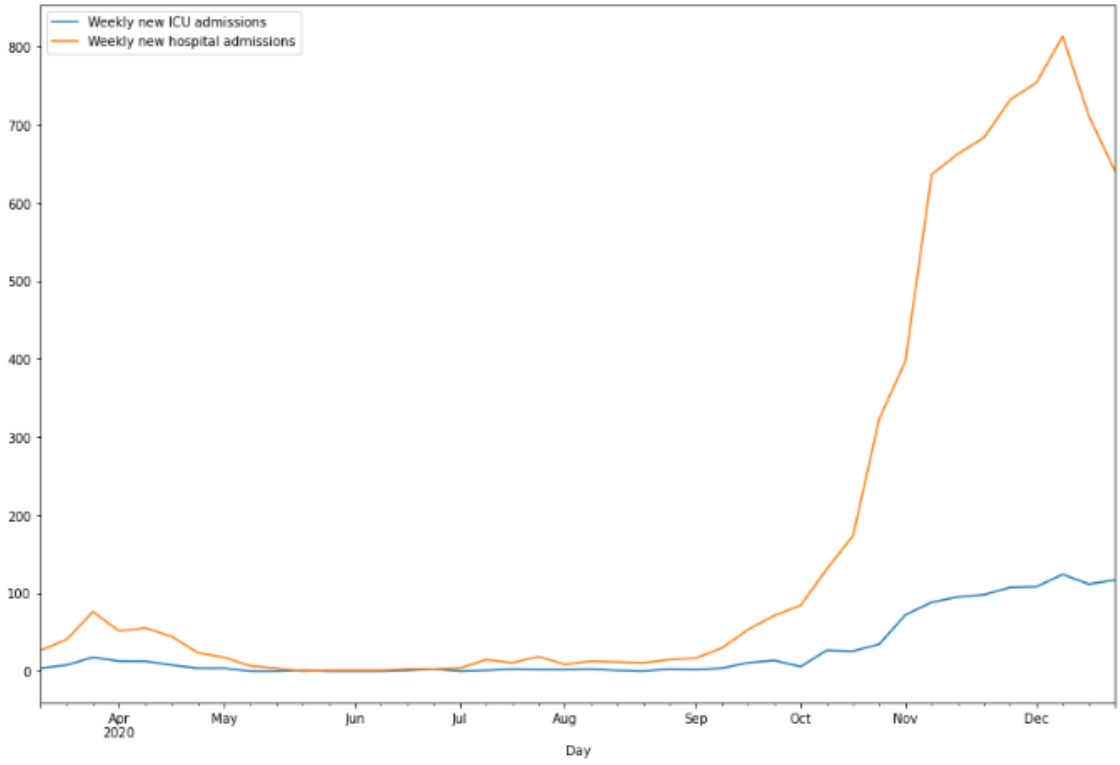

Figure 26: Cluster 5 average weekly Hosptalization and ICU admission rate (Data source: https://ourworldindata.org/)

This cluster kept average weekly new hospitalizations under 100 people until almost October. The maximum new average new hospitalization was recorded to be around 800 in December.

Key takeaway

We have seen that cluster 5 received the least number of hospitalization and ICU admission throughout the pandemic. On the contrary, cluster 3 and 4 have had the worst numbers in terms of hospitalization and ICU admission rate. It is important to note that as these are absolute numbers and the level of occupancy of hospitals depends on the capacity of individual countries grouped in

the clusters. If these hospitalizations happen to be above the threshold of the capacity of one country could handle in these clusters, it will have a high impact in the health facilities work flow , the mental health of healthcare professionals who are the frontline workers and other non-covid patients who have to deal with delayed, or canceled treatments.

4.3. Clustering Analysis based on the Government Measures

Governments around the world have scrambled to get the spread of the virus under control, of which some have unfortunately not been successful. However, “Optimal decision making in the context of Covid-19 pandemic is a complex process that requires to deal with a significant amount of uncertainty and the severe consequences of not reacting timely and with the adequate intensity.” (Alamo, et al., 2020)

According to (Nadeem Ashraf, 2020) government measures taken to mitigate coronavirus can be of two types. The first category is Social distancing interventions where Closure in schools, offices, parks, and mass transportation, among other factors, are examples of. The second category of measures is containment and health response which includes public awareness programs, testing, and quarantining policies.

Comparing a country's government steps against the reported cases is vital evidence for understanding the best practices for managing the virus and preventing a catastrophic pandemic that will inevitably strike our world.

In this step, we'll group the countries based on the policies implemented by governments around the world.

4.3.1. Government Measure Metrics

Based on (Nadeem Ashraf, 2020) classification of social distancing measures, Face coverings, Stay at Home Restrictions, School and workplace closures are the metrics that will be used to measure the degree of actions taken by different nations in different point of time. Testing Policy will also be added to take containment and health responses into consideration in this thesis.

Dataset description

Column	Description
Entity	The name of the country
Code	Code that represent the country

Day	The date of the data record
face_coverings	Country policy on masks Values can be 0,1,2,3 or 4
School_closures	Country policy on school closures values can be 0,1,2 or 3
Work_place closures	Country policy on work place closures values can be 0,1,2 or 3
Testing_policy	Testing Policy of the country values can be 0,1,2 or 3

Table 11: Dataset Description: Government measures (Data source: https://ourworldindata.org/)

Table 12, presents the Oxford COVID-19 Government Response Tracker (**OxCGRT**) government response metrics. (OWID, 2021) where it puts countries into a predefined categories that the countries fit into depending on their situation in regards with implemented mitigation policy. The indicators are further explained as follows.

1. **Face Coverings(Mask mandate):** Countries are placed in the following orders depending on their respective governments stand on masks. (OWID, 2021)

   0- No policy

   1- "Recommended"

   2- "Required in some specified shared/public spaces outside the home with other people present, or some situations when social distancing not possible"

   3- "Required in all shared/public spaces outside the home with other people present or all situations when social distancing not possible"

   4- "Required outside the home at all times regardless of location or presence of other people. "

2. **School Closures** – The following are the categories used to measure the level of school closure in countries. (OWID, 2021)

   0 – "No measures"

   1 – "recommend closing"

   2 – "Require closing (only some levels or categories)"

   3 – "Require closing all levels"

3. **Workplace Closures**- The level of this recommendation in terms of workplace closures is presented as follows. (OWID, 2021)

0 – "No measures"

1 – "recommend closing (or work from home)"

2 – "require closing (or work from home) for some sectors or categories of workers"

3 – "require closing (or work from home) all but essential workplaces (eg grocery stores, doctors)"

4. **Testing Policy**: (OWID, 2021)

0 – "No testing policy"

1 – "Only those who both (a) have symptoms AND (b) meet specific criteria (eg key workers, admitted to hospital, came into contact with a known case, returned from overseas)"

2 – "testing of anyone showing COVID-19 symptoms"

3 – "open public testing (eg "drive through" testing available to asymptomatic people)"

4.3.2. The Czech Republic Perspective

To understand the above mentioned government responses indexes better, let us take a look how the Czech Republic applied the above-mentioned steps from the start of the pandemic to the end of 2020. Line graph can be utilized to see how Czech Republic's policy on testing, facial coverings, workplace closures and school closures varied over time in 2020.

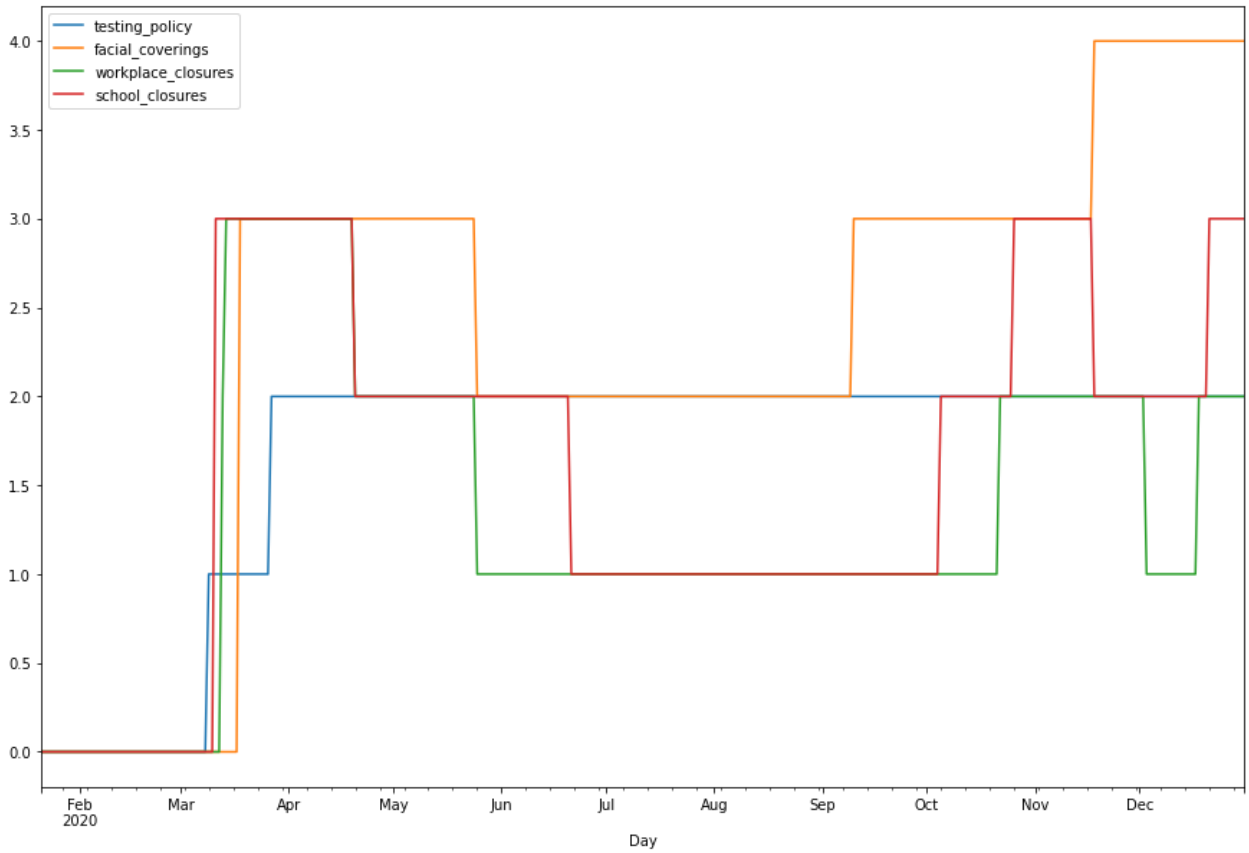

Figure 27: Government measures - The Czech Republic Case (Data source: https://ourworldindata.org/)

As can be seen in the illustration above, responses started to be taken by the beginning of March. Face covering is the latest intervention to be adopted, requiring it in all shared/public spaces outside of the house. All levels of schools and workplaces were required to be closed right from the begging. All schools remained closed until May, when only a few levels of education reopened. After may the required closing down of workplaces were also eased to just recommendation of working from home for non-essential workers.

The summer had the loosest restriction in terms of closing down schools (which should be closed down any way in the summer), workplace closures and face coverings. However, these measures were tightened back again in September and October. This can be correlated to the initial cluster analysis we had we have done on based on the spread pattern. The Czech Republic is a member of cluster 4 where we saw the cases rising begging of September which justifies why the measures where tightened back again in September.

It also can be seen that the testing policy of Czechia stayed relatively the same with the level of 2 where testing is done to anyone showing COVID-19 symptoms.

We can examine if there is a strong relationship between the measures taken by the Czech Republic by using correlation matrix. Since our data consists of ordinal ranks of government responses, Spearman Rank correlation matrix will be utilized. (Mukaka, 2012) Figure 21 depicts the strength of the relationship between the given variables in the form of a heat map grid.

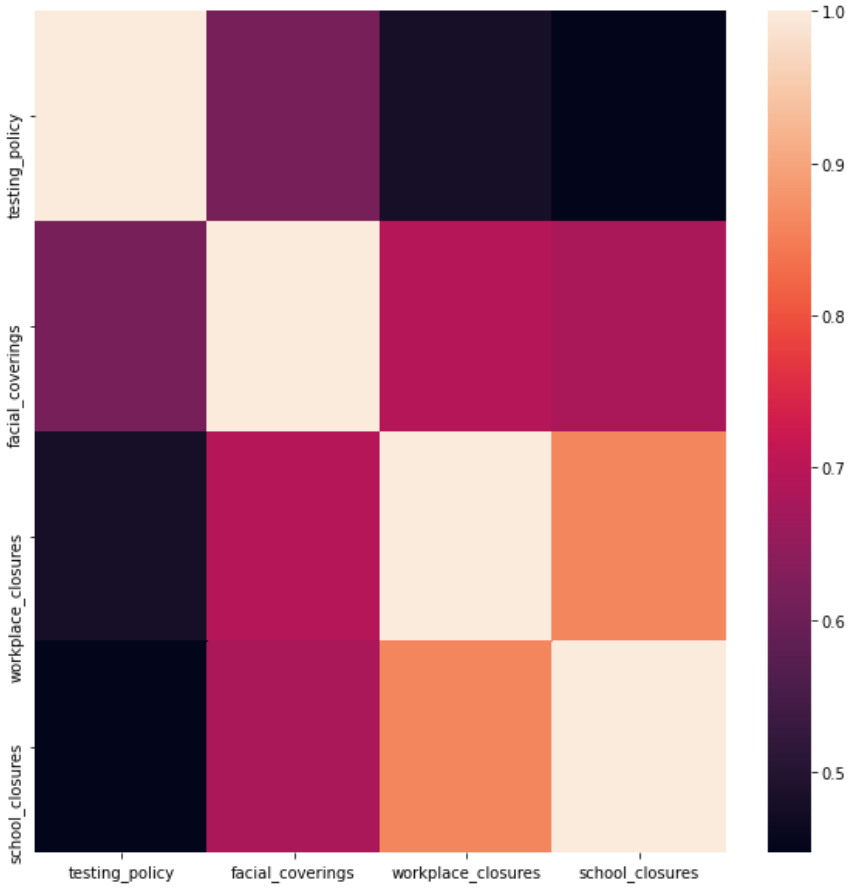

Figure 28:Correlation of Government measures - The Czech Republic Case (Data source: https://ourworldindata.org/)

As can be seen from the figure, we do not have a negative correlation suggesting that all the variables are moving in the same direction. We can rate the strength of the relationships as very

high positive, high positive, moderate positive, low positive and negligible correlation depending on the correlation value they have. (DE, et al., 2003)

From the correlation heatmap shown above, the correlation 0.44 to 0.86.(See Appendix 3) School and Workplace closure have the highest correlation value of 0.86 indicating a high positive relationship between them. This means tougher workplace closures are highly related to tougher school closures. Facial covering and testing policy, Facial coverings and workplace closures, facial coverings and school closures have correlation value of approximately 0.6 suggesting a moderate positive correlation. Testing policy has correlation of 0.48 and 0.45 with Workplace closures and school closures respectively indicating a low positive relationship.

Now that we have clarity how the individual government measures work, we can introduce the more generic Stringency Index concept. Government measure Stringency index is an aggregate index that covers the measures stated above and other measures taken by the governments around the world to measure the overall strictness of the government responses to covid-19. (University of Oxford and BLAVATIK SCHOOL OF GOVERNMENT, 2021).

The index ranges from 0 to 100, with 100 being the strictest. Eight of the regulation metrics outline containment and termination strategies, such as school closures and transport restrictions. Four of the metrics monitor economic policies, such as stipends. The COVID-19 test regime and emergency health spending are two of the metrics (H1-H7) that track health system policies. (University of Oxford and BLAVATIK SCHOOL OF GOVERNMENT, 2021). The following graph shows the Czech Republic Government response over all stringency throughout the year 2020.

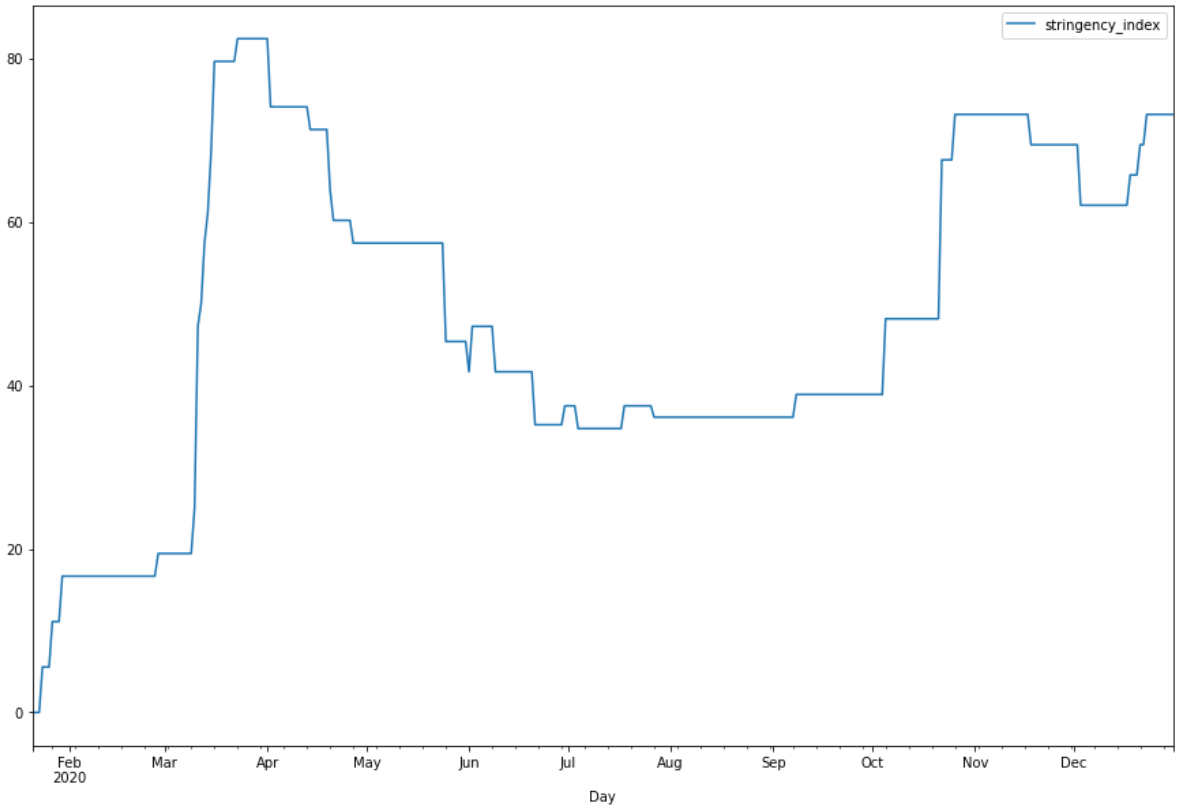

Figure 29: Czech Republic government response stringency (Data source: https://ourworldindata.org/)

Figure 8 presents how the government measures started tightening up by mid March 2020 and how it started to loosen up shortly after. The Czech Republic had the loosest restriction over the summer until October where the strictness increased again. This information is validated by the assessment done on Figure 26 where we saw school closure, workplace closure and mask mandate policies were loosened over the summer of 2020.

4.3.3. Clustering Countries based on Government measures

We have seen how different government measures were imposed and lifted from the Czech Republic perspective in the previous section. However, in this section of the thesis we will zoom out a little bit and investigate how the strictness of the measures changed around the world. Clustering counties which have shown the same trajectory of strictness will help us understand were the common pattern of the measures to be able to achieve that the variable stringency index will be utilized since it encompasses various responses taken by governments. The individual responses such as school closure, workplace closure, facial coverings and testing policy will later come in the picture to further analyze each cluster.

Dataset Description

Column	Description
Entity	The name of the country of the record
Code	The code that represents the country
Day	The date of the record
Stringency	The strictness of the government measure of the given country in the given date

Clustering Analysis

The very first step of the analysis is to clean up the dataset, it is very important to identify if there are countries with zero reported government response stringency. The countries with zero reported response stringency will be hard to interpret since one cannot conclusively tell if these countries have 0 stringency because they actually have no government restriction imposed or if there was imply no data. For that reason, we will get rid of countries with zero reported stringency throughout the year of 2020 from the dataset before we continue with our analysis.

After the dataset is properly cleaned up, clustering analysis can follow and run a hierarchical clustering algorithm on the dataset. The output of the algorithm is presented as follows.

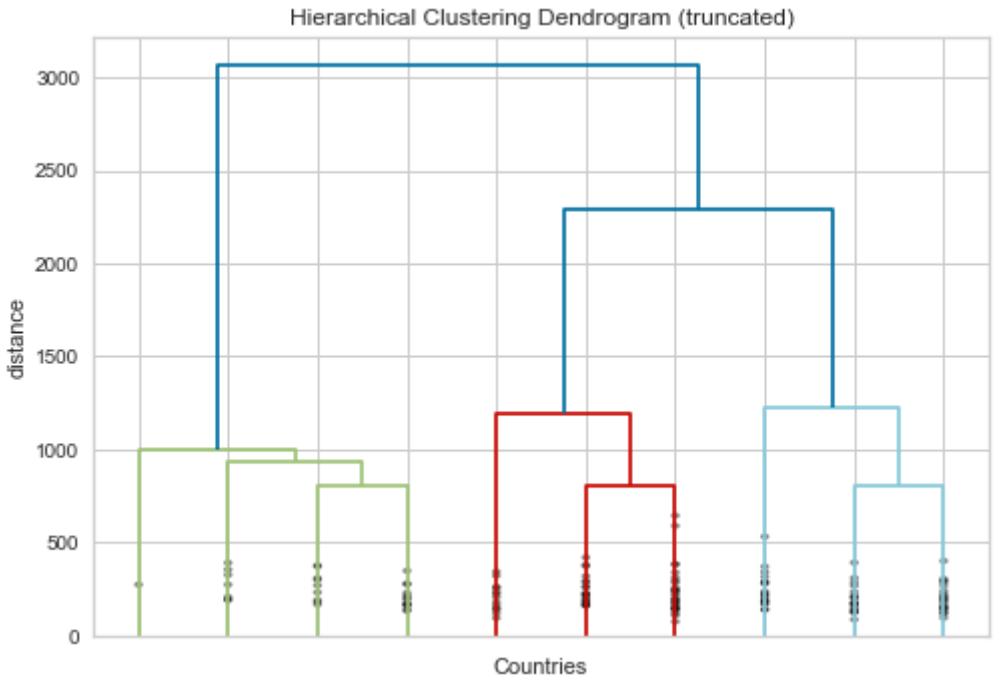

Figure 30: Truncated dendrograms – Government measure stringency (Data source: https://ourworldindata.org/)

Figure 29 shows the hierarchy of the clusters uncovered by the algorithm. To be able to decide the optimal number of cluster and subsequently decide the cut off point, silhouette scores can be used again.

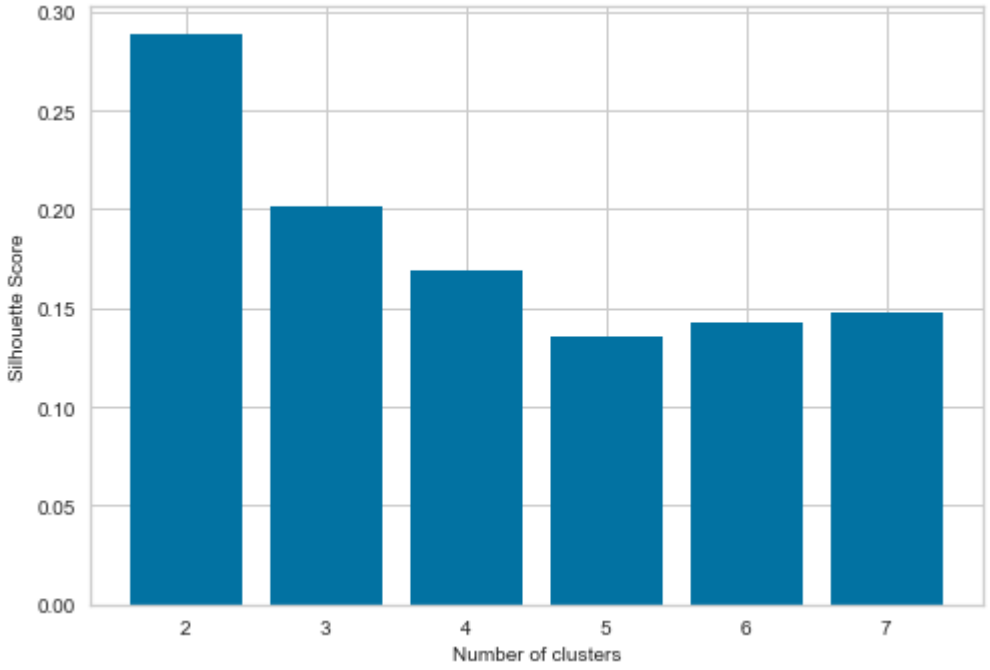

Figure 31: Silhouette Score: Government response stringency (Data source: https://ourworldindata.org/)

As we can see from the silhouette score, we will have the most separated clusters if generate two clusters. When we try to put the silhouette score into context, the first cluster will be consisting of the green labeled in the countries dendrogram (See fig 29) and the second cluster will contain the red and blue labeled countries combined together. However is it apparent from fig 29 that cluster 2 is very large and can be split int two clusters (red and blue) to get more detailed clusters which leaves us with total of three clusters. However, we have to check if having 3 clusters is actually valid and for that the Figure 30 should be check again with 3 clusters. From the silhouette score, having 3 clusters is still valid with a score of just above 0.2. Therefore, we will cut off the truncated diagram on fig 19 at the distance of 1500 to extract 3 clusters. The following figures are the lines graph of stringency index of countries in their respective cluster. The complete list of countries under each cluster can found in Appendix 1.

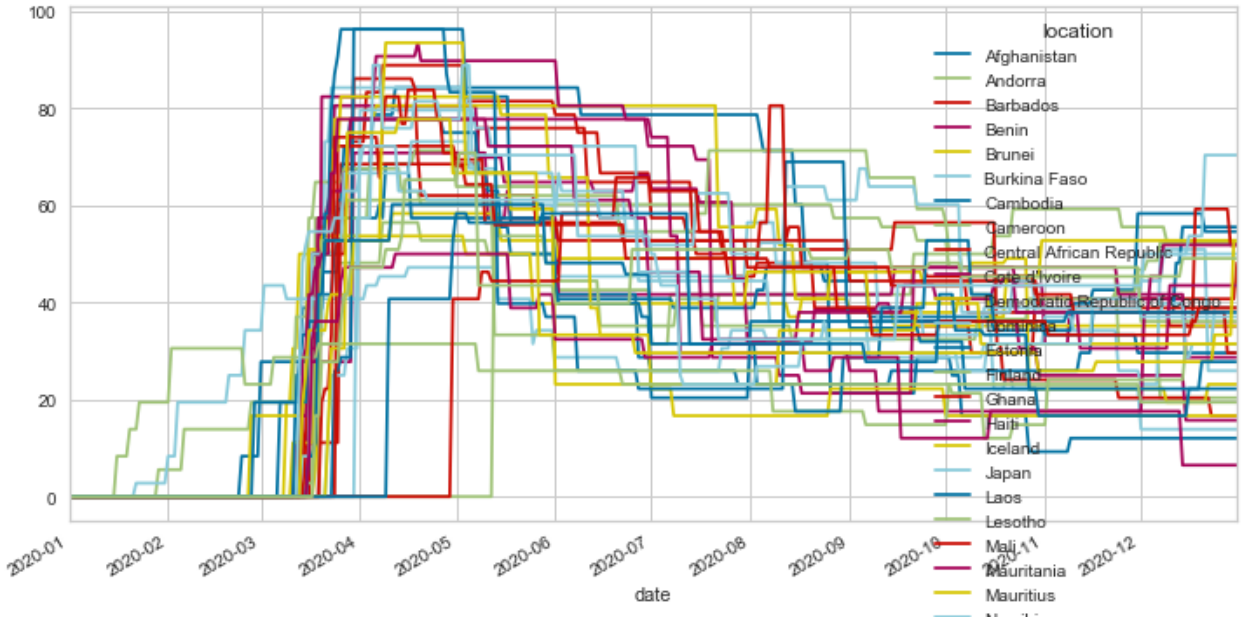

Figure 32: Cluster 1- Government response stringency

Cluster 1 is the least populated cluster with 38 countries grouped under it. From the pattern of line graph shown on Figure 31, it can be seen that government measures were very strict from the begging of the pandemic until around mid May where the ristrictions started to loosen up from that point on. The strictness does not increase once it was loosened up in may in this cluster .

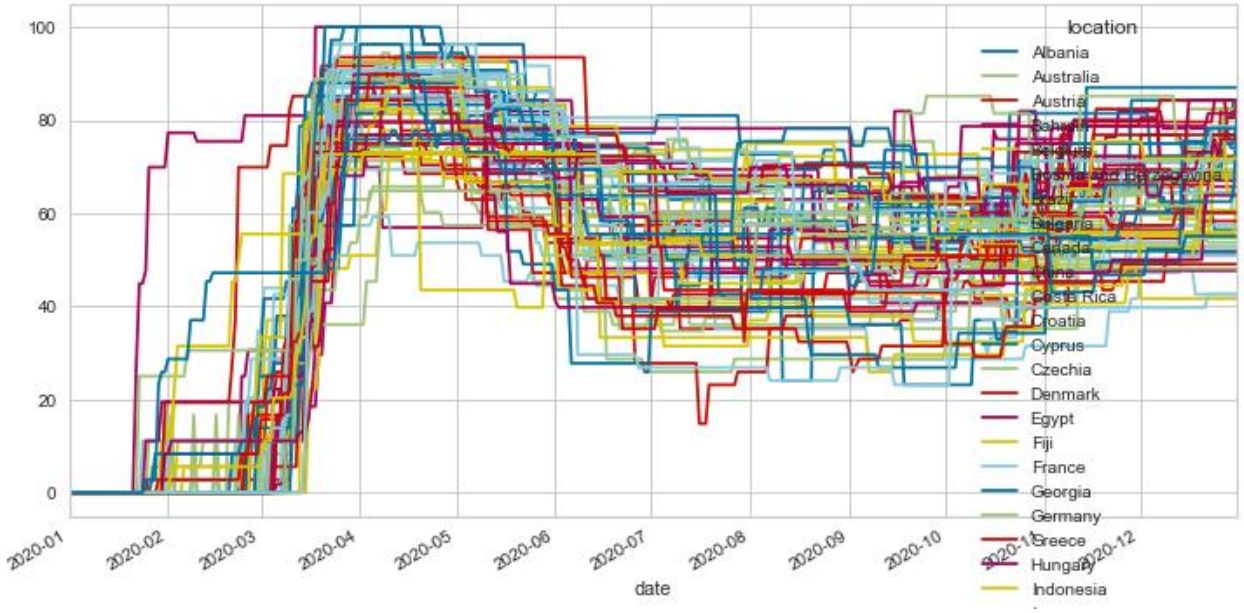

Figure 33: Cluster 2- Government response stringency

The above illustration, figure 32, shows the change in the strictness of government measures for countries in cluster two. Compared with cluster one, cluster two has relatively more number countries that made it above the stringency or strictness level of 60. Similar to cluster one measures were tight from the beginning of the pandemic to around may and these measure were loosened over the summer. However the loosened up measures started to tighten up by the end of the summer. The Czech Republic whose government response was shown earlier is member of this cluster.

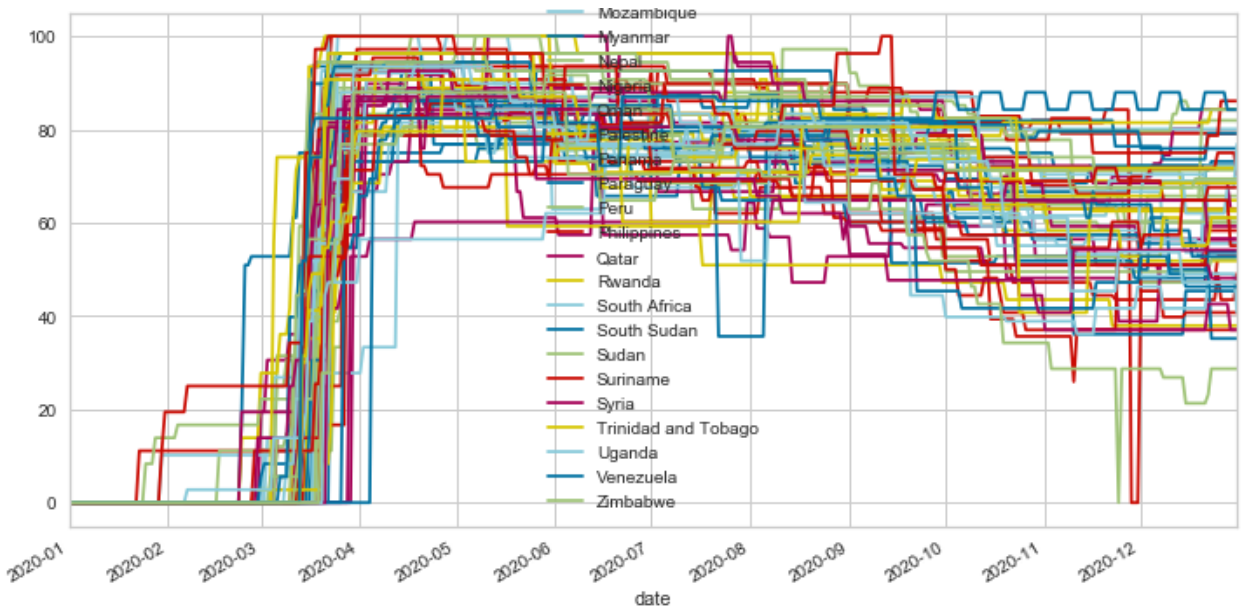

Figure 34: Cluster 3- Government response stringency

Cluster 3 is distinct from the above two clusters in that the measure continues to be strict until the fall of 2020 where some of the cluster countries started to loosen restrictions up going into the winter of 2020. This cluster has 67 member which makes it the most populous cluster.

5. Results and Discussion

Cluster Comparison – Spread Pattern

The average daily per million population cases in each cluster are seen below in figure 34, cascaded up on each other. The key goal of this graphical representation is to compare the clusters side by side and study the virus's spread pattern in the clustered locations.

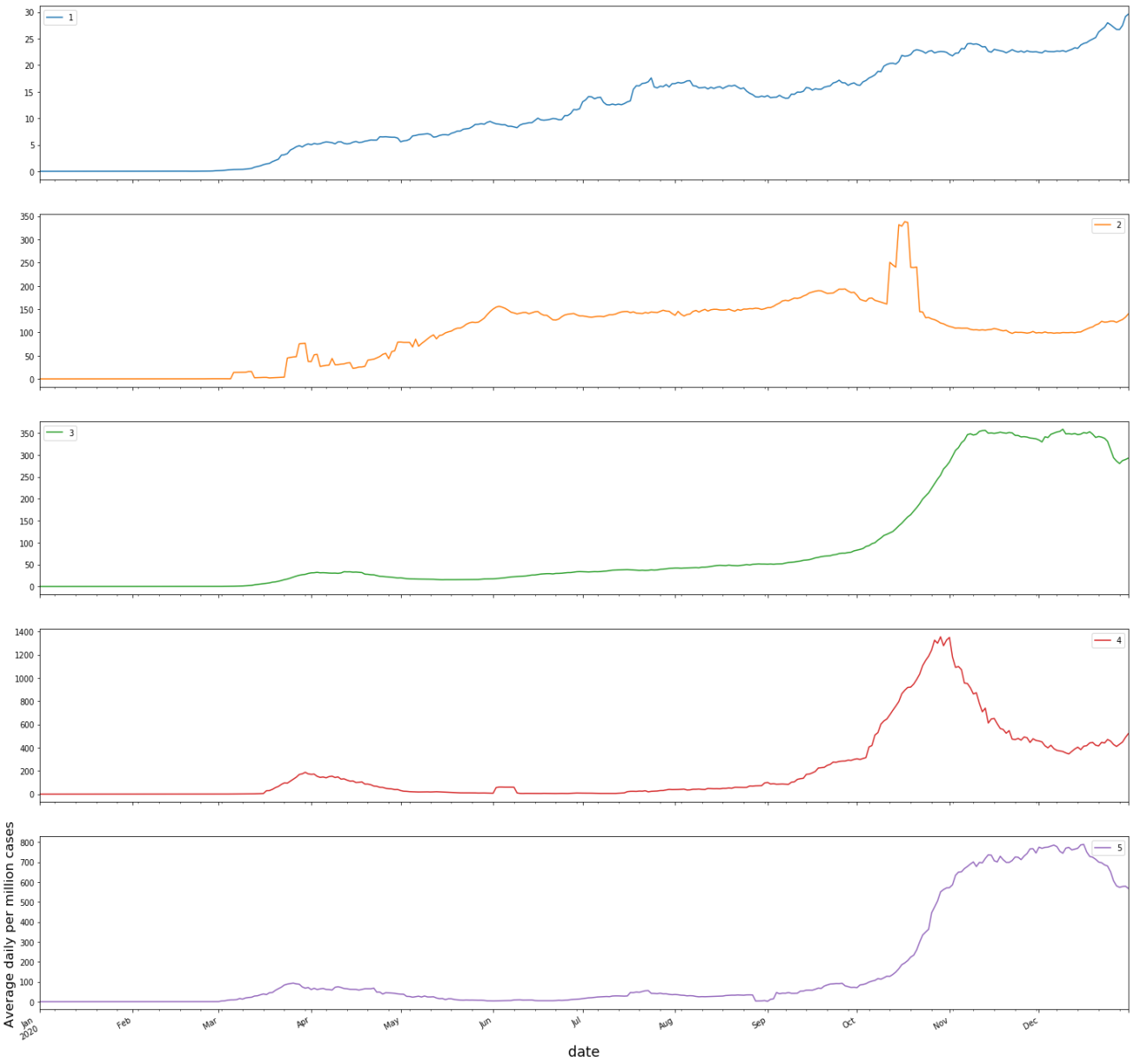

Figure 35: Spread Pattern Cluster Comparison (Data source: https://ourworldindata.org/)

As of December 31st, 2020, Cluster 1 had an average new daily per million case of below 30. Until October, the daily average number of new infected people per million population in this group was only less than 25.

Cluster 2 has relatively high number of infection rate compared to cluster 1. Except for a sudden increase in cases in October, this cluster relatively had a constant daily new per million cases of 150 throughout the year of 2020. It showed a maximum of average daily per million cases of 350 in October.

Cluster 3 also stayed in the same range with cluster 2 in 2020. However they had different patterns. Compared to cluster 2, the third cluster had a relatively lower cases until October and kept it below daily average of 100 new per million.

Cluster 4 and 5 both showed high increase in infection rate where they reached daily average of per million infection of 800 and 1200 respectively in fall of 2020 despite having really low count in earlier seasons.

From the above comparison, it can be concluded that cluster 1 has managed the spread of the virus well. Cluster 4 and 5, although they did manage to control the virus until the end of the summer, infections skyrocketed form that point to the end of the year of 2020 making them on the most affected clusters.

The cluster mapping below is presented to help visualize how the clusters have spread around the world. All African countries are grouped in cluster 1. The same goes for almost all Asian countries whereas Europe is the most diverse in terms of cluster with the majority being grouped in cluster 3. Most south American countries are clustered in cluster 2 where North America has been put in cluster one except for USA which is grouped in cluster 3.

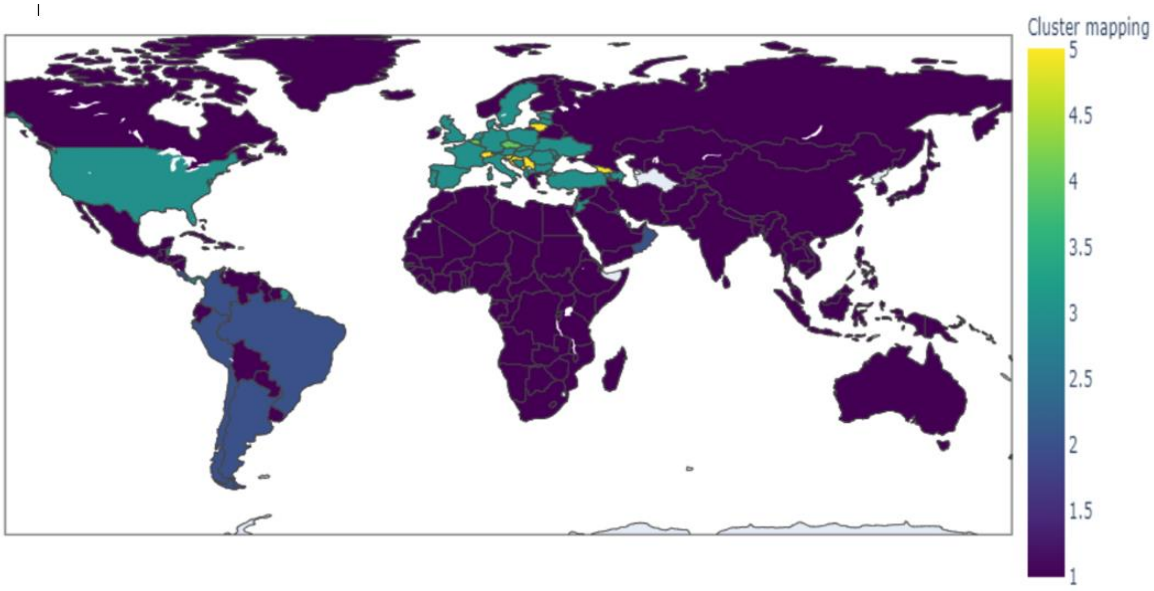

Figure 36: Spread Pattern Cluster location (Data source: https://ourworldindata.org/)

Analyzing the clusters based on Stringency index

In the practical part we have seen how the stringency, or the strictness of government measures changed in each cluster. However, it requires to dive a little deeper than just the stringency index to understand what measures exactly changed. Now, this section focuses on breaking the stringency index down to the individual measures we have seen for the case of Czech Republic.

The three clusters extracted based on government response stringency index will be assessed based on school closures, workplace closures, facial covering policy and test policy to learn more about the clusters. We cannot calculate the daily average of each cluster as we did while assessing the spread pattern given that we are dealing with ordinary variables in this case. Instead we will calculate the daily median of each index for each cluster to describe the behavior of the cluster with respect to the above-mentioned responses.

Cluster 1

In Figure 36 shows the daily median of this cluster on testing policy ranking which is shown increasing after May of 2020 to reach the point of half of the memebers were tesing anyone showing COVID-19 symptoms.. Face mask policy is also shown to increase throughout the year of 2020. However it is apparent from the graph figure 36,that school and government closures dramatically

being lifter after the April of 2020 and half of the member of this cluster continued to follow just a closure reccomendation until the end of the year.

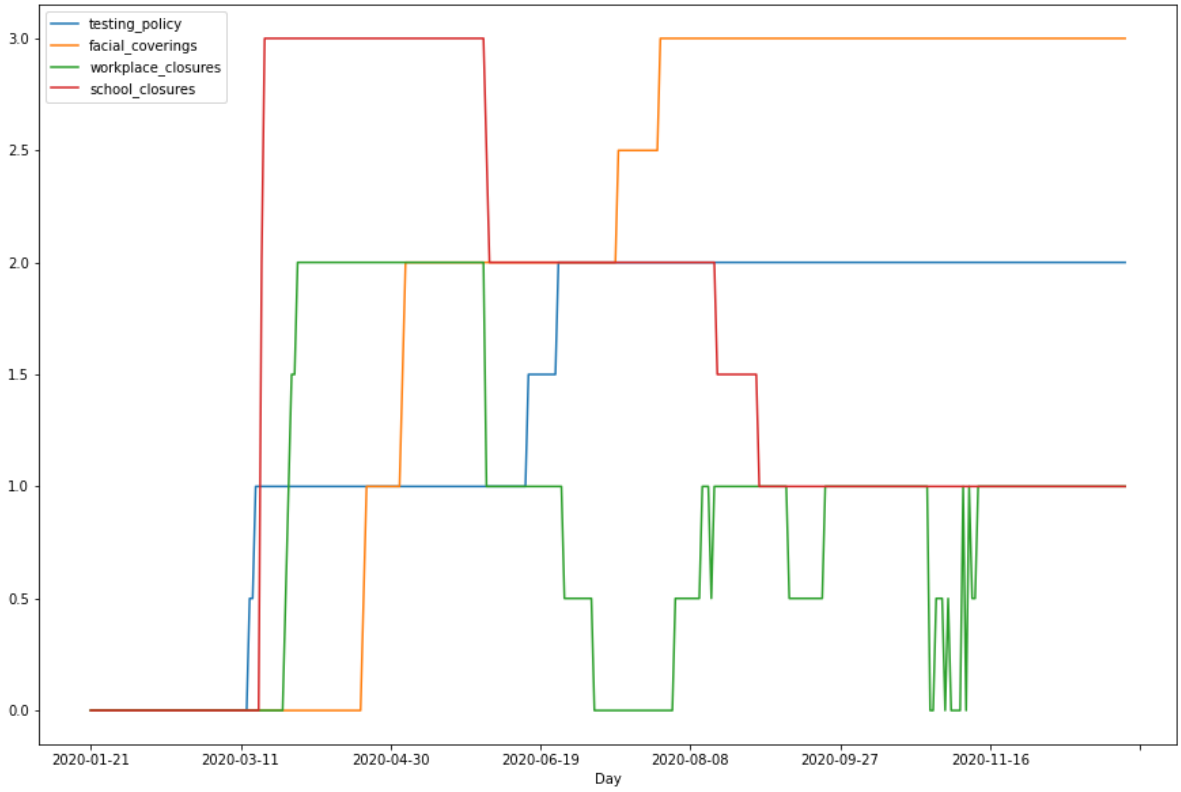

Figure 37: Daily median of various government measures – Cluster 1(Data source: https://ourworldindata.org/)

Cluster 2

From figure 7, we can onserve that measures regarding school and workplace were tightened to the higest measures in atleast half of the countries in this cluster. It is also seen that loosened up measures after May did not go as loose as Cluster 1 countires and ever were tightened back up again. The same as cluster 1 this cluster shows a trajectory of increased mask recommendation policy over time.

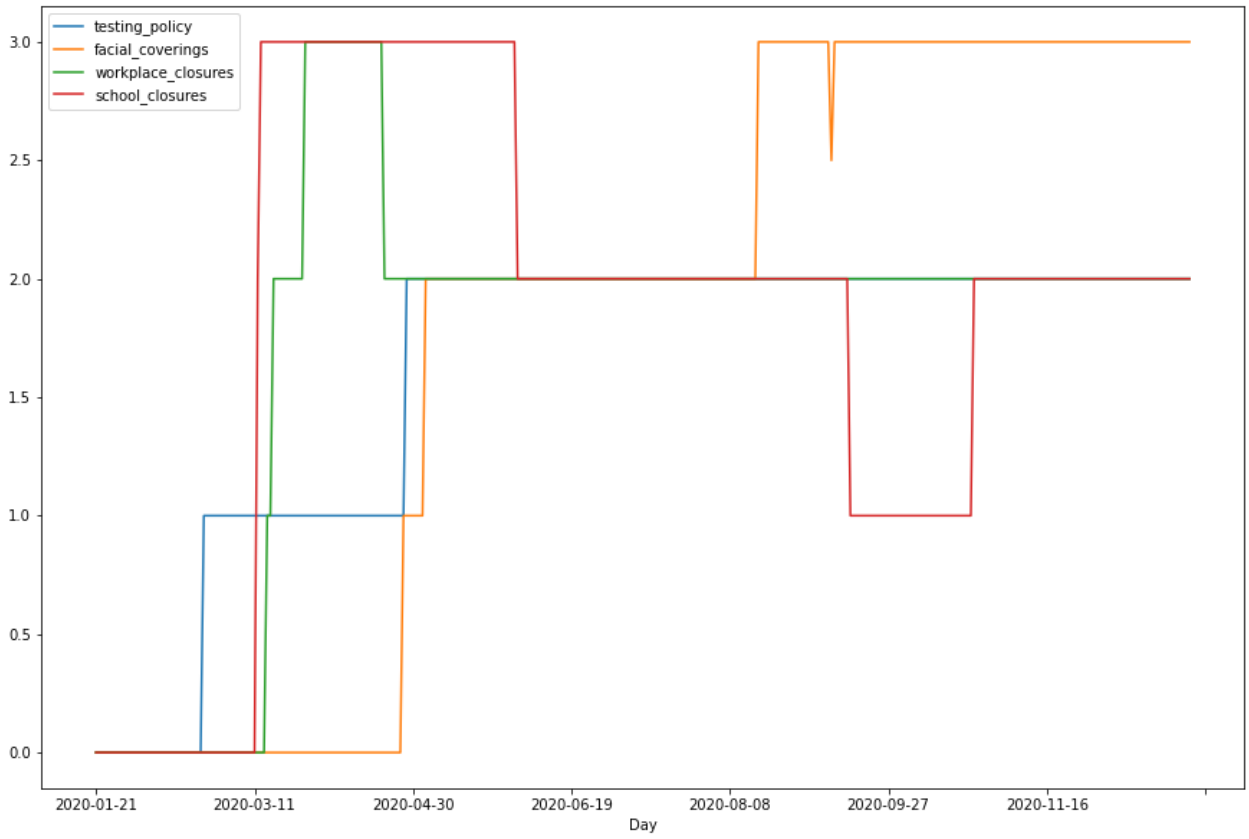

Figure 38: Daily median of various government measures – Cluster 2(Data source: https://ourworldindata.org/)

Cluster 3

In cluster 3 in it shown in fig 38 that similar to cluster 2, the measures were tightened up to highest point but unlike the other clusters, the measures continued to be the same until the fall of 2020. For half of the countries in this cluster face covering started to be required in all share spaces outside home in April and continued to be the same until the end of the year.

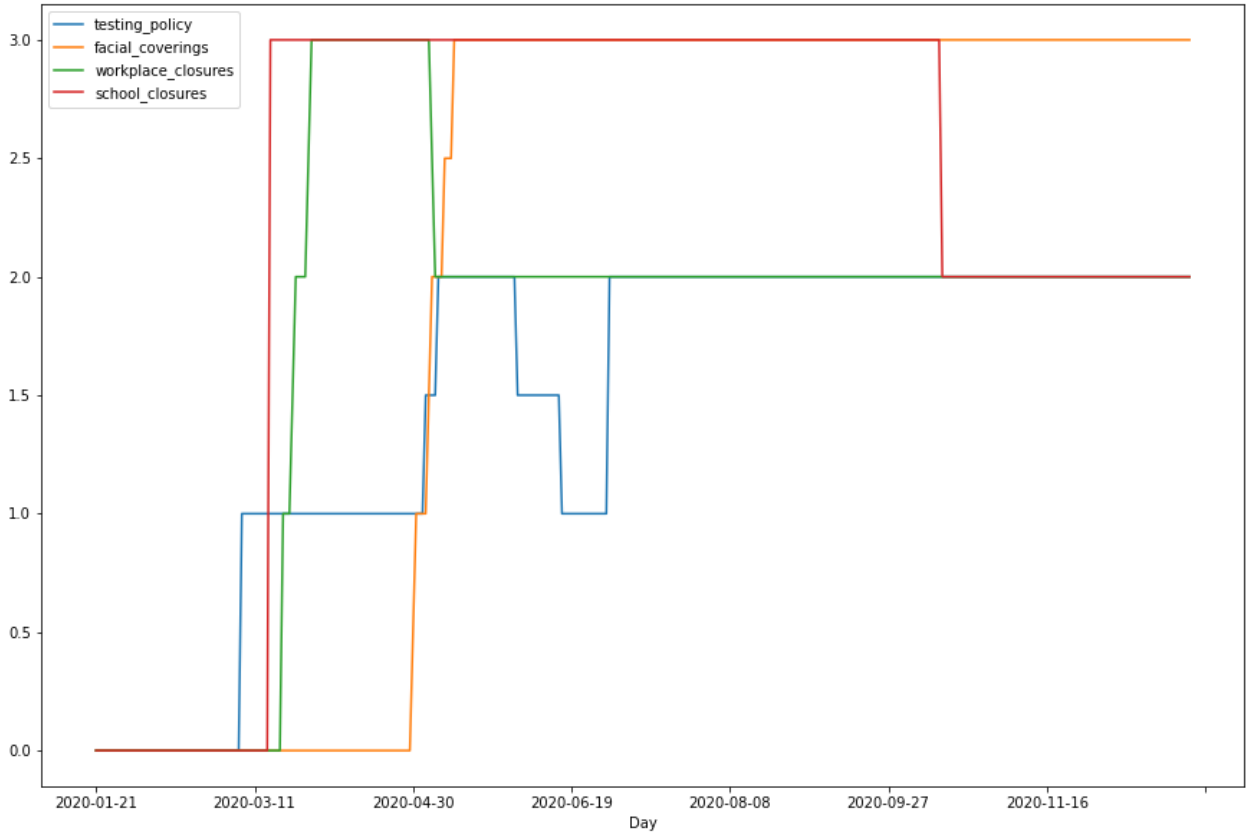

Figure 39: Daily median of various government measures – Cluster 3(Data source: https://ourworldindata.org/)

6. Conclusion

As our world has been struck by an unprecedented virus, Coronavirus, big data driven research should be able to provide insights into the pandemic's transmission and control measures, which will better guide current and future epidemiological decision-making.

This diploma thesis used machine learning techniques to understand the COVID-19 pandemic. In order achieve this objective various data sets were utilized. Exploratory data analysis done in one of the datasets revealed that the United States was the worst hit country in 2020 in terms of absolute COVID-19 cases while Andorra was number one in terms of COVID-19 cases per 1 million population. Moreover, the EDA revealed that in 2020, Europe contributed 28.8% of the total covid-19 cases reported worldwide which is higher than aggregated case reported by any other ocontinent.

Cluster analysis was done on countries around the world based on new cases per million population of the countries to identify the most common trajectory of the spread of the virus. After running hierarchical clustering algorithm, five most common trajectories exhibited around the world were discovered. The clusters are of different sizes and showed different spread pattern. The first cluster , Cluster 1 , was found to be the least affected by showing a maximum moving average of only 300 new cases per million throughout the year 2020. On the contrary cluster 4 and cluster 3 were found to be the most hit countrier where moving average of new cases per million reached 1600 and 1200 by the end of the year 2020.

After grouping countries based on the virus's transmission, the socio-economic impact of the pandemic was analysed in each cluster. The study focused on some aspects of the economy, education and health services of the clusters. Based on the descriptive statistics done on growth and unemployment rate of clusters, it was shown that cluster 2 was the worst performing cluster with respect to the economy. In terms of education, school closures were compared against the internet availability of each cluster. In that regard, cluster 1 was the worst performing cluster in that it had a median of 45% of coverage of internet and the schools some levels of school were closed for at least half of 2020 and all levels were closed down from the beginning of the pandemic to May 2020. The fact was found to be problematic as online learning cannot be an option during close down of schools. The level of hospitalization and ICU admission rate was also studied in

each cluster to show how the occupancy of the health care services of each cluster will increase because of the pandemic. Though in a different time, Cluster 3 and 4 experienced the highest hospitalizations in 2020 where the numbers reached 10000 and 12000 respectively.

Finally, the government measures taken to contain the transmission of the virus were assessed. The three most common response stringency variations were uncovered. Cluster 1 in this case shown an increased restriction at the beginning of the pandemic and the measures were quickly loosened and the restriction was not increased after the May of 2020. Cluster 2 also showed the increased restriction from the beginning of the pandemic until around May and the restriction were loosened until the end of the summer. Unlike cluster 1, the response measures of cluster 2 countries started to be tightened back again by of the same year. Cluster 3 contrary to the other 2 clusters, the restriction continued until the end of September and measures started to be loosened back again after that until the end of the year 2020.

Future research may analyse the possible relationship between the clusters produced based on the virus's spread pattern and the clusters created based on the stringency index to assess which steps were actually effective in combating the virus's spread.

Furthermore, as more data becomes available, it will be important to analyse the pandemic's socioeconomic impact in the clusters using more measures such as the human development index, the number of recorded domestic violence cases during lockdowns, and other critical data that will help us understand the pandemic's effects from a variety of perspectives.

7. References

Baum, J. et al., n.d. Applications of Big Data analytics and Related Technologies in Maintenance—Literature-Based Research. *Department of Economics, University of Applied Sciences Zwickau, 08012 Zwickau, Germany;.*

Buheji, M. et al., 2020. The Extent of COVID-19 Pandemic Socio-Economic Impact on Global Poverty. A Global Integrative Multidisciplinary Review. *American Journal of Economics.*

Chunarkar-Patil, P. & Bhosale, . A., 2018. Big data analytics. *Open Access Journal of Science* , 2(5).

Kaye, A. D. et al., 2020. Economic impact of COVID-19 pandemic on healthcare facilities and systems: International perspectives.

Kwok Tai Chui, Miltiadis D. Lytras & Ryan Wen Liu, 2010. *nnovations, Algorithms, and Applications in Cognitive Informatics and Natural Intelligence (Advances in Computational Intelligence and Robotics).* s.l.:s.n.

Mahyijari, N. A., Badahdah, A. & Khamis, F., 2020. The psychological impacts of COVID-19: a study of frontline physicians and nurses in the Arab world. *Cambridge University Press.*

Omran, M. G., Engelbrecht, A. P. & Salman, A., 2007. *An Overview of Clustering Methods.* [Online]
Available at:
https://www.researchgate.net/publication/220571682_An_overview_of_clustering_methods
[Accessed 22 3 2021].

Praveen, S. & Chandra, U., 2017 . Influence of Structured, Semi-Structured, Unstructured data on various data models. *International Journal of Scientific & Engineering Research,* 8(12).

Ylijoki, O. & Porras, J., 2016. Perspectives to Definition of Big Data: A Mapping Study and Discussion. *Journal of Innovation Management,* Volume 4.

Abbott, D., 2014. *Applied Predictive Analytics.* Indianapolis, Indiana: John Wiley & Sons, Inc.

Adedoyin, O. B. & Soykan, E., 2020. Covid-19 pandemic and online learning: the challenges and opportunities. *Interactive Learning Environments.*

Adhikari, S. P. et al., 2020. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. Volume 9.

Aggarwal, C. C. & Reddy, C. K., 2014. *Data Clustering Algorithms and Applications.* Newyork: Tylor & Francis Group.

Alamo, T., Reina, D. G. & Gata, P. M., 2020. Data-Driven Methods to Monitor, Model, Forecast and Control Covid-19 Pandemic: Leveraging Data Science, Epidemiology and Control Theory. *arXiv.*

AL, J. J. V. B. E., 2020. Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour.*

Aljumily, R., 2016. Agglomerative Hierarchical Clustering: An Introduction to Essentials.. *Global Journal of HUMAN-SOCIAL SCIENCE: G Linguistics & Education* , 16(3).

APA, 2015. *Measuring Socioeconomic Status and Subjective Social Status.* [Online]
Available at: https://www.apa.org/pi/ses/resources/class/measuring-status

Auxiliadora Sarmiento, I. F. ,. I. D.-D. S. C., 2019. Centroid-Based Clustering with αβ-Divergences. *MDPI.*

Boshkoska, M. & Jankulovski, N., 2020. *Coronavirus Impact on Global Economy,* Bitola: University of Targu Jiu.

Burgess, S. & Sievertsen, H. H., 2020. *Schools, skills, and learning: The impact of COVID-19 on education.* [Online]
Available at: https://voxeu.org/article/impact-covid-19-education

Carlsson , G. & Memoli, F., 2010. Characterization, Stability and Convergence of Hierarchical Clustering Methods. *Journal of Machine Learning Research* .

Chen, B., Ting, K. M., Washio , T. & Zhu , Y., 2018. Local contrast as an effective means to robust clustering against varying densities. *Mach learn.*

DE, H., Wiersma , W. & SG, J., 2003. *Applied statistics for the behavioral sciences.* Boston: Houghton.

Ester, M., Kriegel, H.-P., Sander, J. & Xu, . X., 1996. *A Density-Based Algorithm for Discovering Clusters.* [Online]
Available at: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Gantz, J. & E. Reinsel, 2011. Extracting Value from Chaos. *IDC's Digital Universe Study.*

Garlasu, D. et al., 2020. A big data implementation based on Grid computing. pp. 1-4.

HADI, H. J., SHNAIN, A. H., HADISHAHEED, S. & AHMAD, A. H., 2015. BIG DATA AND FIVE V'S CHARACTERISTICS. *International Journal of Advances in Electronics and Computer Science,* Volume 2.

Hale, T. et al., 2020. *Variation in government responses to COVID-19* , Oxford: Blavatnik School of Government, University of Oxford.

Howe, L. D. et al., 2012. Measuring socio-economic position for epidemiological studies in low- and middle-income countries: a methods of measurement in epidemiology paper. *International Journal of Epidemiology* .

IBM, n.d. Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics. *IBM Software Thought Leadership WHite Paper.*

International Labour Organization, n.d. *Indicator description: Unemployment rate.* [Online]
Available at: https://ilostat.ilo.org/resources/concepts-and-definitions/description-unemployment-rate/

J. Manyika, et al., 2011. Big data: The next frontier for innovation, competition, and productivity.

Josh, A. & Mujawar, S., 2015. Data Analytics Types, Tools and their Comparison. *International Journal of Advanced Research in Computer and Communication Engineering,* 4(2).

K.Sasirekha & P.Baby, 2013. Agglomerative Hierarchical Clustering Algorithm- A Review. *International Journal of Scientific and Research Publications.*

Laney, D., 2001. Data Management:Controlling Data Volume, Velocity and Variety. *Application Delivery Strategies.*

Lin, Y. et al., 2018. A Method of Extracting The Semi-structured Data Implication Rules. *ScienceDirect.*

Mashingaidze, K & Backhouse, J., 2017. The relationships between definitions of big data, business intelligence and business analytics: A literature review.. *Inderschience Publishers* .

matplotlib, n.d. *Documentation.* [Online]
Available at: https://matplotlib.org/

MBBS, P. D., MRCOG., 2020. Coronavirus disease 2019 (COVID-19) pandemic and pregnancy. *American Journal of Obstetrics and Gynecology.*

Merriam-Webster, 2021. *Dictionary.* [Online]
[Accessed 27 03 2021].

Mukaka, M., 2012. A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal.*

Nadeem Ashraf, B., 2020. Economic impact of government interventions during the COVID-19 pandemic: International evidence from financial markets. *Journal of Behavioral and Experimental Finance.*

Nicola, M. et al., 2020. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International Journal of Surgery.*

Numpy, n.d. *Documentation.* [Online]
Available at: https://numpy.org/doc/stable/

O. Austin, C. & Kusumoto, F. M., 2016. The application of Big Data in medicine: current implications and future directions. *Journal of Interventional Cardiac Electrophysiology.*

Ogbuabor, G. & Ugwoke, F. N, 2018. CLUSTERING ALGORITHM FOR A HEALTHCARE DATASET USING SILHOUETTE SCORE VALUE. *International Journal of Computer Science & Information Technology (IJCSIT)*, 04.Volume 10.

OWID, 2021. *School-Closures.* [Online]
Available at: https://ourworldindata.org/grapher/school-closures-covid

Pandas, n.d. *Documentation.* [Online]
Available at: https://pandas.pydata.org/docs/

Panel Ishwarappaa & J.Anuradhab, 2015. A BRIEF INTRODUCTION ON BIG DATA 5VS CHARACTERISTICS AND HADOOP TECHNOLOGY. *International conference On Intellegent Computing, Communication and Convergence* .

Paul Zikopoulos, et al., 2011. *Harness the Power of Big Data.* s.l.:s.n.

Piiroinen, J., 2018. Density-based clustering, Master's Thesis. *University of Eastern Finland* .

Plowright, R. et al., 2017. Pathways to zoonotic spillover. 15(502-510).

Python, n.d. *About.* [Online]
Available at: https://www.python.org/about/

Rani, Y. & Rohil, D. H., 2013. A Study of Hierarchical Clustering Algorithm. *International Journal of Information and Computation Technology.*.

Reddy, M. V., Vivekananda, M. & Satish, R. U. V. N., 2017. Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering. *International Journal of Computer Science Trends and Technology (IJCST)* , 5(5).

Rehman, M. H. u., Chang, V., Batool, A. & Wah, . T. Y., n.d. Big Data Reduction Framework for Value Creation in Sustainable. *Faculty of Computer Science and Information Technology, University of Malaya, Suzhou Business School, Xi'an Jiaotong Liverpool University, Suzhou, China,Department of Computer Science, Iqra University, Islamabad, Pakistan.*

Rui Han, Xiaoyi Lu & iangtao Xu, 2014. *Big Data Benchmarks, Performance Optimization, and Emerging Hardware.* s.l.:s.n.

scikit-learn, n.d. *Home.* [Online]
Available at: https://scikit-learn.org/stable/
[Accessed 01 03 2021].

Scikit-learn, n.d. *sklearn.metrics.silhouette_score¶*. [Online]
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
[Accessed 01 02 2021].

SciPy, n.d. *Documentation.* [Online]
Available at: https://www.scipy.org/

Soria, A. et al., 2020. The high volume of patients admitted during the SARS-CoV-2 pandemic has an independent harmful impact on in-hospital mortality from COVID-19. *PLOS ONE.*

Sreenivasan, R. R., 2017. Characteristics of Big Data – A Delphi study. *Faculty of Business Administration-Memorial University of Newfoundland.*

Techson, M., 2020. *Blog.* [Online]
Available at: https://blog.angular.io/version-11-of-angular-now-available-74721b7952f7
[Accessed 3 3 2021].

Thorén, E. & Filip Brännlund, . S., n.d. *Usage of ANgular from developer's perspective,* Karlskrona: Blekinge Institute of Technology .

Tuffery, S., 2011. *Dara Mining and Statistics for Decision Making.* s.l.:John Wiley & SOns .

UNDP, 2020. *Socio-economic Impact of Covid-10.* [Online]
Available at: https://www.undp.org/content/undp/en/home/coronavirus/socio-economic-impact-of-covid-19.html

University of Oxford and BLAVATIK SCHOOL OF GOVERNMENT, 2021. *Coronavirus Government Response Tracker.* [Online]
Available at: https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker

Uppada, S. K., 2014. Centroid Based Clustering Algorithms- A Clarion Study. *International Journal of Computer Science and Information Technologies,.*

Verma, J. P., Agrawal, . S., Patel, . B. & Patel, A., 2016. BIG DATA ANALYTICS: CHALLENGES AND APPLICATIONS FOR TEXT, AUDIO, VIDEO, AND SOCIAL MEDIA

DATA. *International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI),*.

WHO, 2020. *Coronavirus disease (COVID-19) - Diruation Report - 153.* [Online]
Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200621-covid-19-sitrep-153.pdf?sfvrsn=c896464d_2

Wieringa, P. d. J. E., 2016. Unstructured data: Can its power be unleashed?. *Univeristy of groningen: Faculty of economics and business.*

WorldoMeter, 2021. *Coronavirus.* [Online]
Available at: https://www.worldometers.info/coronavirus/

Xiaofei Ma & Satya Dhavala, 2018. Hierarchical Clustering with Prior Knowledge. *Amazon.com Inc.*.

ZEHRA, A., 2020. SPREAD PATTERN ANALYSİS OF COVID-19 WİTH THE HELP OF ORDİNARY DİFFERENTİAL EQUATİON. *INTERNATIONAL CONFERENCE ON COVID-19 STUDIES* .

Zhou, F. et al., 2020. *Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study*. [Online]
Available at: https://pubmed.ncbi.nlm.nih.gov/32171076/
[Accessed 25 03 2021].

8. Appendix

8.1. Appendix 1 : Cluster members

Cluster	Members	
1	Afghanistan, Algeria, Angola, Antigua and Barbuda, Australia, Bahamas, Bangladesh, Barbados, Belarus, Benin, Bermuda, Bhutan, Bolivia, Botswana, Brunei, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Cayman Islands, Central African Republic, Chad, China, Comoros, Congo, Cote d'Ivoire, Cuba, Democratic Republic of Congo, Djibouti, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Faeroe Islands, Fiji, Finland, Gabon, Gambia, Ghana, Gibraltar, Greece, Greenland, Grenada, Guatemala, Guernsey, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hong Kong, Iceland, India, Indonesia, Iran, Iraq, Ireland, Isle of Man, Jamaica, Japan, Jersey, Kazakhstan, Kenya, Kyrgyzstan, Laos, Lesotho, Liberia, Libya, Madagascar, Malawi, Malaysia, Mali, Marshall Islands, Mauritania, Mauritius, Mexico, Micronesia (country), Monaco, Mongolia, Morocco, Mozambique, Myanmar, Namibia, Nepal, New Zealand, Nicaragua, Niger, Nigeria, Norway, Pakistan, Papua New Guinea, Paraguay, Philippines, Russia, Rwanda, Saint Helena, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa, Sao Tome and Principe, Saudi Arabia, Senegal, Seychelles, Sierra Leone, Singapore, Solomon Islands, Somalia, South Africa, South Korea, South Sudan, Sri Lanka, Sudan, Suriname, Syria, Taiwan, Tajikistan, Tanzania, Thailand, Timor, Togo, Trinidad and Tobago, Tunisia, Uganda, United Arab Emirates, Uruguay, Uzbekistan, Vanuatu, Venezuela, Vietnam, Yemen, Zambia, Zimbabwe	
2	Argentina, Bahrain, Brazil, Chile, Colombia, Costa Rica, Israel, Kuwait, Maldives, Oman, Peru, Qatar, Vatican	
3	Albania, Armenia, Austria, Azerbaijan, Belize, Bosnia and Herzegovina, Bulgaria, Cyprus, Denmark, Estonia, France, Germany, Hungary, Italy, Jordan, Kosovo, Latvia, Lebanon, Malta, Moldova, Netherlands, North Macedonia, Palestine, Panama, Poland, Portugal, Romania, Slovakia, Spain, Sweden, Turkey, Ukraine, United Kingdom, United States	
4	Andorra, Belgium, Czechia	
5	Croatia, Georgia, Liechtenstein, Lithuania, Luxembourg, Montenegro, San Marino, Serbia, Slovenia, Switzerland	

Table 12: Spread Pattern Cluster members(Data source: https://ourworldindata.org/)

Cluster	Members
1	Afghanistan, Andorra, Barbados, Benin, Brunei, Burkina Faso, Cambodia, Cameroon, Central African Republic, Cote d'Ivoire, Democratic Republic of Congo, Dominica, Estonia, Finland, Ghana, Haiti, Iceland, Japan, Laos, Lesotho, Mali, Mauritania, Mauritius, Namibia, New Zealand, Niger, Papua New Guinea, Senegal, Seychelles, Sierra Leone, Somalia, Taiwan, Tajikistan, Tanzania, Timor, Uruguay, Yemen, Zambia
2	Albania, Australia, Austria, Bahrain, Belgium, Bosnia and Herzegovina, Brazil, Bulgaria, Canada, China, Costa Rica, Croatia, Cyprus, Czechia, Denmark, Egypt, Fiji, France, Georgia, Germany, Greece, Hungary, Indonesia, Iran, Ireland, Israel, Italy, Jordan, Kosovo, Latvia, Lebanon, Lithuania, Luxembourg, Malaysia, Malta, Moldova, Monaco, Mongolia, Morocco, Netherlands, Norway, Pakistan, Poland, Portugal, Romania, Russia, San Marino, Saudi Arabia, Serbia, Singapore, Slovakia, Slovenia, South Korea, Spain, Sri Lanka, Sweden, Switzerland, Thailand, Togo, Tunisia, Turkey, Ukraine, United Arab Emirates, United Kingdom, United States, Uzbekistan, Vietnam
3	Algeria, Angola, Argentina, Azerbaijan, Bahamas, Bangladesh, Belize, Bhutan, Bolivia, Botswana, Cape Verde, Chad, Chile, Colombia, Congo, Cuba, Djibouti, Dominican Republic, Ecuador, El Salvador, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Guatemala, Guinea, Guyana, Honduras, India, Iraq, Jamaica, Kazakhstan, Kenya, Kuwait, Kyrgyzstan, Liberia, Libya, Madagascar, Malawi, Mexico, Mozambique, Myanmar, Nepal, Nigeria, Oman, Palestine, Panama, Paraguay, Peru, Philippines, Qatar, Rwanda, South Africa, South Sudan, Sudan, Suriname, Syria, Trinidad and Tobago, Uganda, Venezuela, Zimbabwe

Table 13: Stringency Index cluster members (Data source: https://ourworldindata.org/)

8.2. Appendix 2 : Economic Indicators Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
1	110.0	-5.204545	8.355665	-66.7	-7.10	-4.40	-1.500	26.2
2	12.0	-8.600000	4.292700	-18.6	-10.45	-7.05	-5.725	-4.5
3	31.0	-7.429032	4.316495	-25.0	-8.45	-6.10	-4.750	-3.6
4	1.0	-8.300000	NaN	-8.3	-8.30	-8.30	-8.300	-8.3
5	9.0	-6.566667	3.518167	-12.0	-9.00	-5.80	-5.000	-1.8

Table 14: 2020 Growth rate summary statistics(Data Source: https://www.imf.org/en/Publications/WEO/Issues/2021/01/26/2021-world-economic-outlook-update)

2019 Growth rate summary statistics

	count	mean	std	min	25%	50%	75%	max
1	110.0	2.658182	4.784347	-35.0	1.10	3.00	5.300	9.9
2	12.0	1.583333	2.036411	-2.1	0.70	1.45	2.475	5.7
3	31.0	2.306452	2.441166	-6.9	1.55	2.20	3.500	7.6
4	1.0	1.400000	NaN	1.4	1.40	1.40	1.400	1.4
5	9.0	2.966667	1.356466	1.1	2.30	2.90	3.900	5.1

Table 15: 2019 Growth rate summary statistics (Data Source: https://www.imf.org/en/Publications/WEO/Issues/2021/01/26/2021-world-economic-outlook-update)

8.3. Appendix 3 : Spearman Rank Correlation- Government Measures

	testing_policy	facial_coverings	workplace_closures	school_closures
testing_policy	1.000000	0.613949	0.481775	0.447345
facial_coverings	0.613949	1.000000	0.692863	0.680235
workplace_closures	0.481775	0.692863	1.000000	0.860795

school_closures	0.447345	0.680235	0.860795	1.000000

Table 16: Spearman Rank correlation - Government Measures(Data source: https://ourworldindata.org/)