Registration number: 210132797
Word Count: 3269
**excluding front page content, abstract, content table, tables, figure captions, and appendix*

Measuring COVID-19 Impact on Economy with Countries Clustering, PCA, and Multiple Linear Regression

Abstract

COVID-19 has spread worldwide and affects every part of our lives, including the economy. Owing to the importance of the economy sector, this essay performs comprehensive research to assess the variables that affect economic growth during the COVID-19 pandemic and predict the economic growth based on the assessed variables. The research uses a comprehensive method including clustering, dimensionality reduction, visual analysis, and regression. The result shows that the service sector, the total number of deaths, life expectancy, human development index, and stringency index are integral variables in directing economic growth during the COVID-19 pandemic for countries clustering and economic growth prediction.

Contents

1. Introduction

The world is facing a severe COVID-19 Pandemic brought by the SARS-CoV-2 Virus and its variants. The spread of this disease has affected not only the healthcare system of countries, but the impact can be seen across critical sectors, including food supply (Aday & Aday, 2020), education (Daniel, 2020), global diplomacy (Davies & Wenham, 2020), and especially economy (Verma et al., 2021)

The spread A pandemic can lead to the shortfall of the economic development in countries. As Prager et. al (2017) argue that a pandemic might negatively affect GDP of countries, Price-Smith, A.T (2009) also found that a 3-4% decrement of GDP might happen.

With the impact the COVID-19 brought to the economic sector, a comprehensive assessment needs to be conducted to measure how big the impact of COVID-19 pandemic brought to our economy and variables that contribute to it.

This research aims to answer these questions:

1. How are the countries clustered based on disease prevalence, economic, and healthcare and wellbeing indicators related to covid19?
2. Relationship between economic performance (measured in GDP Growth Rate) and other covid-19 related variables.
3. How is the performance of linear regression algorithm measured in different set of variables and clusters?

This essay is organised as follows. The first section discusses the introduction, the background, and the experiment's goal. The second section will discuss the methodology that consists of dataset explanation, pre-processing method, and analytic method. Section 3 analyse the results and findings from used algorithms. Then, concluding remarks are offered in Section 5.

2. Methodology

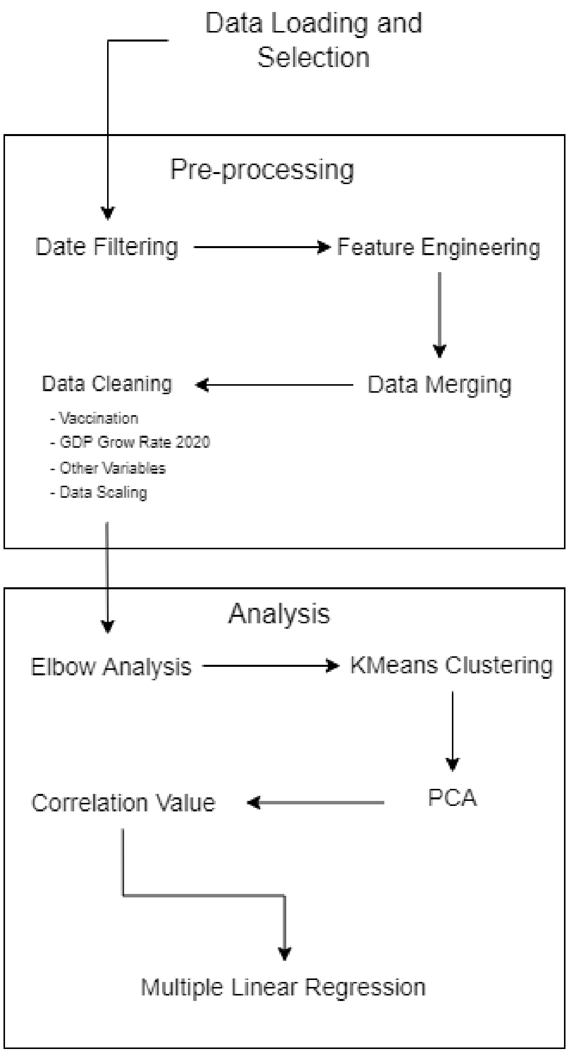


Figure 1: Code Flow

2.1 Data Selection and Data Loading

Firstly, The datasets are loaded to our environment. There are two methods used to load the data. The first one uses the *read.csv()* function, which loads the Covid19 database from the local drive as the data has been downloaded from its original source; https://github.com/owid/covid-19-data/tree/master/public/data. The second method is utilising World Development Indicator (WDI) from R package WDI.

There are two main data sources used in this research. The first one is the COVID19 Dataset provided in the course. The second one is World Bank Data. There are 13 variables used in this research which some of them used for analysis and other as the indicator. Disease prevalence indicators comprise of total cases per million, total deaths per million, total vaccinated per hundred, and stringency index. Economic prevalence entails five indicators, GDP per capita, GDP growth rate, service sector percentage of GDP, industrial sector percentage of GDP, and agriculture percentage of GDP. The healthcare and wellbeing indicators consist of life expectancy, human development index, population density, and hospital beds per thousand.

Table 1 : List of Variables

Prevalence	Source	Variable Name	Description
Disease Prevalence	Covid 19 Dataset	total_cases_per_million	Cumulative confirmed case of COVID-19
		total_deaths_per_million	Cumulative deaths attributed to COVID-19
		total_vaccinated_per_million	Total number of COVID-19 vaccination doses administered per 100 people in the total population
		hospital_beds_per_thousand	Hospital beds per 1,000 people, most recent year available since 2010
		stringency_index*	Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response) (Oxford Dataset) *will be feature engineered to describe mean and max stringency index
Wellbeing		life_expectancy	Life expectancy at birth in 2019
		human_development_index	A composite index measuring a long and healthy life, knowledge and a decent standard of living.

		population_density	Number of people divided by land area, measured in square kilo meters
Economy		gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars)
	World Bank	GDP_growth_rate_2020	Gross Domestic Product growth (annual %)
		service_sect	Services, value added (% of GDP)
		industrial_sect	Industry (including construction), value added (% of GDP)
		agriculture_sect	Agriculture, forestry, and fishing, value added (% of GDP)

2.2 Pre-processing

The second step is pre-processing and divided into four parts: Feature Engineering, Data Merging, Data Filtration, and Data Cleaning. Due to its constantly changing values, we need to summarise the stringency index variable. Feature Engineering process is conducted on stringency index variable to produce mean and maximum stringency index. This process is utilising the piping method, *group_by()*, *summarise()*, *mean()*, and *max()* functions. After that, *merge()* function from generic base **R** was used to combine WDI data, COVID19 data, and stringency index data with country name variable (location) as a pairing key between datasets. The *merge()* function is chosen to avoid column redundancy. All the variables were then filtered for a specific date, 31 December 2020, due to the relevancy with our regression goal, which is to predict the 2020 GDP Growth rate. However, there are some quality issues in the datasets, such as missing value and the presence of outliers. The Interquartile Range method was used to remove outliers from the data. Data points lies outside the Q1 and Q3 are considered as outliers and deleted. Figure 2 depicts comparison between dataset before and after outliers deletion.

The last part of data pre-processing is to remove missing values from the dataset. Missing values can be quite troublesome especially when we will conduct mathematical or statistical analysis to our data, the missing values may lead to computational error and unexpected values and incorrect conclusion. (Zainal Abidin et al., 2018)

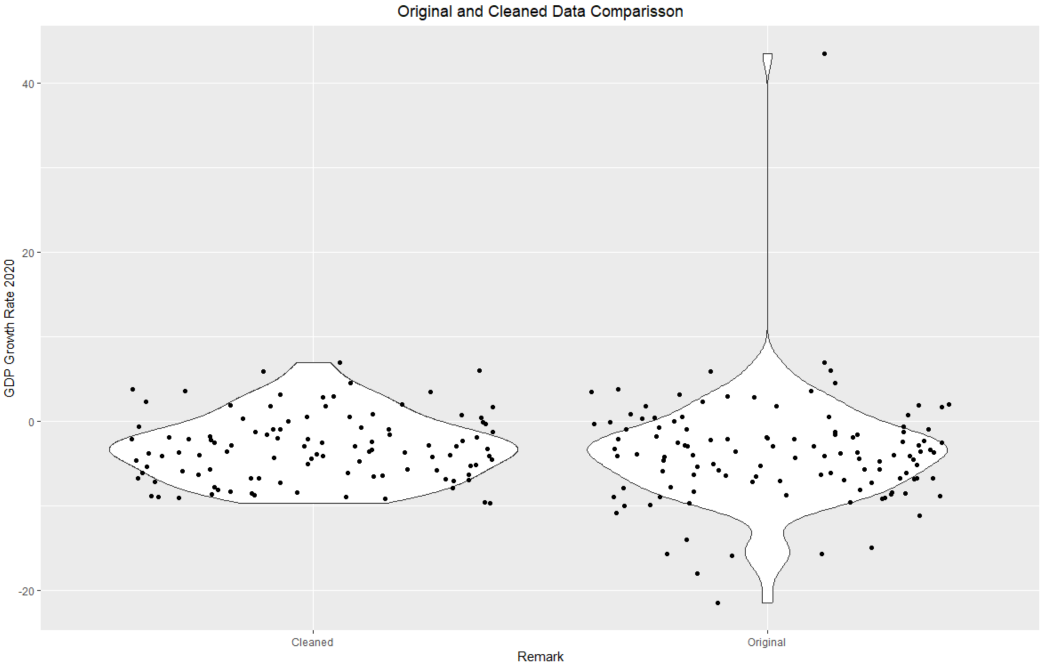


Figure 2: Before and After Pre-Processing Data Points Comparison

After the data are cleaned, scaling process is conducted. The scaling is essential especially when we are dealing with distance-based clustering method such as KMEANS and DBSCAN. The performance of clustering method which are implemented on scaled data grants more accuracy and efficiency. (Mohamad, I and Usman, D. 2013).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Equation 1: Scaling Formula

2.3 Analysis

The aim of this part is to bring out insights from our dataset to answer the goals mentioned in the introduction section

2.3.1 Elbow Analysis

Since the KMEANS algorithm perquisite the initial cluster number, an elbow analysis is conducted. The elbow analysis is a heuristic method to determine the optimum cluster number based on the average distance between data points within the clusters. Intuitionally, as the cluster increases, the fit will also increase. To determine the most optimum cluster number, we need to find the 'sharpest elbow' from the graph. However, in the practice of identifying a sharp elbow, it is not always possible to pinpoint it with precision.

The elbow analysis conducted by utilising *fviz_nbclust()* function from *factorextra* library. This function can compute three analysis methods, elbow, silhouette, and gap statistic, that can be utilised by different types of clustering analysis such as K-means, K-medoids, and HCUT. In this experiment, we use *kmeans* as the value for *the FUNcluster* parameter, *wss* as the value for *method* parameter to estimate the optimal number of cluster (*wss* stands for a total within the sum of the square), and 10 as value for *k.max*, which indicates the maximum cluster number that wants to be analysed.

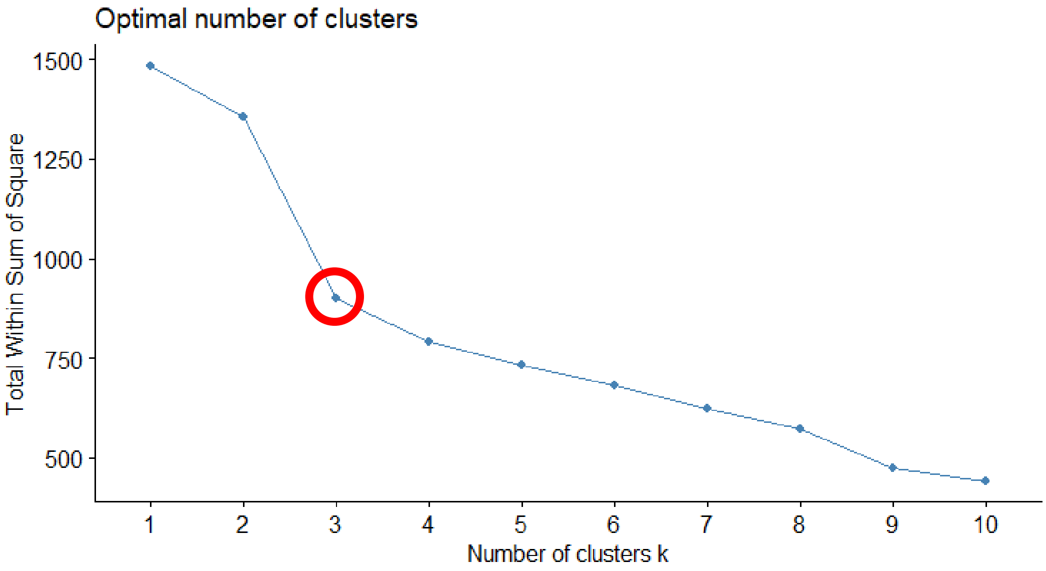


Figure 3: Elbow Analysis

Based on the elbow analysis graph, 3 is chosen as the optimum number of clusters used for the KMEANS clustering. Number 3 was chosen due to a significant decrement of the total within the sum of square value from 2 to 3, but there is no more significant decrement after 3. Hence, over-fit will occur if the cluster is larger than 3.

2.3.2 KMEANS Clustering

KMEANS clustering is chosen due to its effectiveness and ubiquitous use case for clustering numerical dataset. (Mathur B, Ant P, Kaushik M, 2014). The algorithm will produce a high intra-cluster similarity with low inter-cluster similarity; hence, the data points from each cluster is easier to distinguish from each cluster. As the result from prior elbow analysis, 3 is set as the value for krange parameter so the KMEANS algorithm will produce three different clusters. For *criterion* and *usepam* parameter, *asw* and *TRUE* are chosen due to the small size of dataset. The *FALSE* value used for *scaling* parameter as the scaling process has been done in the pre-processing part.

2.3.3 Principal Component Analysis

Due to the number of variables used in the clustering analysis, it is impossible to visualise multidimensional data in a 2d plane; hence, we conduct the Principal Component Analysis to reduce the dimensionality of our data. The output of this process is two additional dimensions or variables which can be used to assist visualisation and in predictive models. In the last part of the analysis method (regression), the performance of the regression method that uses the original variable with the one that uses the Principal Component dimension will be compared.

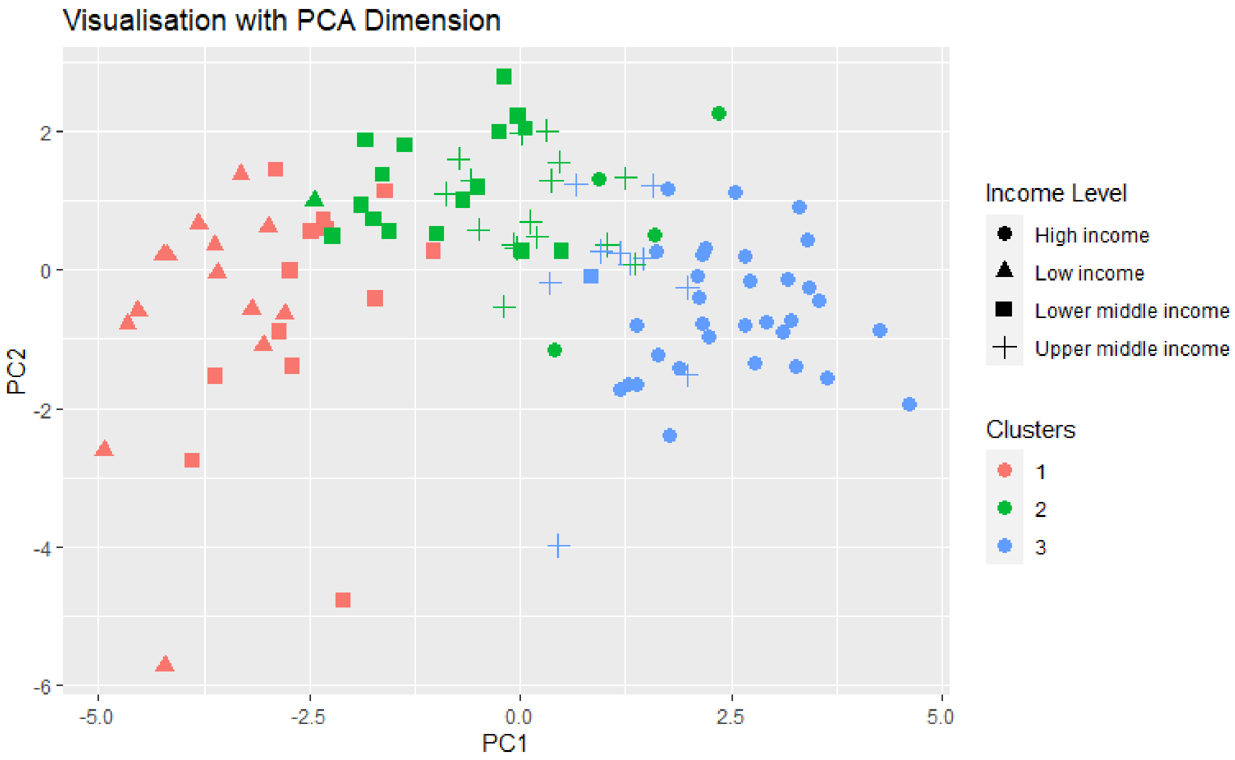


Figure 4: Cluster Spread on PCA Visualisation

The Principal Component Analysis part was conducted using *prcomp()* function from the default "stats" library. There are only two parameters used for the PCA which consist of the data and the column.

2.3.4 Correlation Value

Correlation value has a pivotal role in determining which variable is best for predicting GDP Growth Rate 2020. However, variables with a high correlation value do not always guarantee to be used in the regression; this is due to the multicollinearity rule, which dictates the correlation value between independent variables that are used for regression cannot exceed a certain value to achieve a more reliable result with narrower confidence interval and increase the statistical significance of regression coefficients(Vatcheva et al., 2016)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Equation 2: Pearson Correlation Formula

Pearson correlation coefficient is utilised to measure the relationship strength between each variable. The strength of the relationship between X and Y variable is measured through equation X. The closer the correlation value to zero, then the weaker the correlation between the variables. As the correlation value get closer to 1, it indicates a strong positive correlation between the two variables, which if one value is increases the other one also increases. In contrast, strong negative correlation relationship occurs as the number getting closer to -1, which if one value is increases the other one decreases.

To produce correlation matrix, *corplot()* and *cor()* function from *corrplot* package is utilised. First we run *cor()* function to generate correlation matrix, then the matrix is used an input parameter to the *corplot()* function.

2.3.5 Multiple Linear Regression

After acquiring the PCA dimension, the clusters, and the variables with high correlation values, several multiple linear regression experiments are done to check which variable and method produce the highest accuracy. The quality of the attempts is measured with Adjusted R-Squared Value and P-Value. This step aims to compare the performance of multiple linear regression on several regression attempts. The variation of the attempts is listed in Table 2. Generally, there are three kinds of attempt, the first one is to assess the performance of the regression on the un-clustered data, the second one is done on each cluster individually, and the last attempt is to measure the regression performance on the un-clustered data with PCA variables as predictor.

P-value and Adjusted R Squared Value is used to score the performance of each experiment/attempt. P-value describes the statistical significance of the model; if the value is less than 0.05, the model is considered statistically significant. Adjust R Squared Value preferred over the R Squared value due to the difference in the number of variables used in this experiment.

The Multiple Linear Regression part is conducted using *lm()* function from R *stats* package with the dependent variable as input parameter, followed by independent variables. On this part, variable GDP Growth Rate 2020 (%) is used as the representation of countries' economic performance.

Table 2: List of Experiment Attempted

Attempt	Independent Variable
Cluster 1 Only	• Total deaths per million • Human development index • Service sector • Max stringency
Cluster 2 Only	• Total deaths per million • Human development index • Service sector • Max stringency
Cluster 3 Only	• Total deaths per million • Human development index • Service sector • Max stringency
All Data	• Total deaths per million • Human development index • Service sector

	• Max stringency
All Data	• PC1 • PC2

3. Result and Discussion

3.1 Cluster Analysis

Table 3: Produced Clusters

Cluster	Countries
Cluster 1	Afghanistan, Benin, Burkina Faso, Burundi , Cameroon ,Central African Republic , Djibouti ,Eswatini, Ethiopia, Ghana, Guinea, Haiti, Kenya, Liberia, Malawi, Mali, Mozambique, Myanmar, Nicaragua, Niger, Pakistan, Sudan, Tajikistan, Tanzania, Togo, Uganda, Zambia
Cluster 2	Albania, Algeria, Azerbaijan, Bangladesh, Bolivia, Botswana, China, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Gabon, Guatemala, Honduras, India, Indonesia, Jordan, Kuwait, Madagascar, Malaysia, Mexico, Mongolia, Morocco, Nepal, Paraguay, Philippines, Saudi Arabia, South Africa, Sri Lanka, Thailand, Tunisia, United Arab Emirates, Uruguay, Uzbekistan, Vietnam, Zimbabwe
Cluster 3	Australia, Austria, Bahrain, Belarus, Belgium, Bosnia and Herzegovina, Brazil, Bulgaria, Chile, Croatia, Cyprus, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Kazakhstan, Latvia, Lithuania, Luxembourg, Malta, Moldova, Netherlands, Norway, Poland, Portugal, Qatar, Romania, Serbia, Singapore, Slovenia, Sweden, Switzerland, Turkey, Ukraine, United Kingdom

3.1.1 Cluster Analysis

As shown in Table 3, Cluster 1 comprises 27 countries. This cluster contains low income and lower-middle-income countries. Cluster 2 comprises 37 countries. This cluster contains mixed income level countries, from low-income to high-income. Cluster 3 comprises 43 countries, as shown in Table 2. This cluster mostly contains high-income countries. The result suggest separation between the clusters on the used variables as shown in Table 4. The clusters which mostly consist of certain income level countries positioned accordingly in the tensor plane. However, due to sheer number of variables, the PCA visualisation presented to shown separation between clusters in Figure 3.

Table 4: Clusters mean of Variables

Cluster mean of variables for COVID-19			
Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**
total deaths per million	24.2212963	223.9962973	593.4922326
total cases per million	1312.801444	10541.24243	33655.31784
total vaccinations per hundred	0	0.008378378	0.393953488
population density	104.9234444	152.0863243	374.2573721
GDP per capita (USD)	2669.713481	14876.19376	36993.22272
hospital beds per thousand	0.988888889	2.076756757	4.568139535
life expectancy (years)	63.79962963	73.62459459	79.50465116
human development index	0.522851852	0.728648649	0.880372093
GDP growth rate 2020 (%)	0.647054116	-4.805326861	-4.057180228
service sector (% of GDP)	45.609193	54.64003087	62.75930756
industry sector (% of GDP)	23.80314228	29.12222073	23.86338353
agriculture sector (% of GDP)	23.00100524	10.02896921	3.010384087
mean stringency index	53.255581	67.41520526	57.34238428
max stringency index	75.68555556	91.34108108	82.39627907

3.1.2 Cluster Analysis in PCA Visualisation

Inferring to Figure 3, which visualise the data points on PCA Axes, the separation between clusters can be seen. However, there are no clear boundaries between clusters. Cluster 1 positioned in the lower side of PC 1 and PC2, cluster 2 position in the middle between cluster 1 and 3 which located on the positive direction of PC1. In addition to that, Figure 3 also depicts the spread of the countries based on the income and cluster. Taking into account the mean values on Table 4, we can infer that the result of the PCA and clustering is related especially on the PC1 Axis.

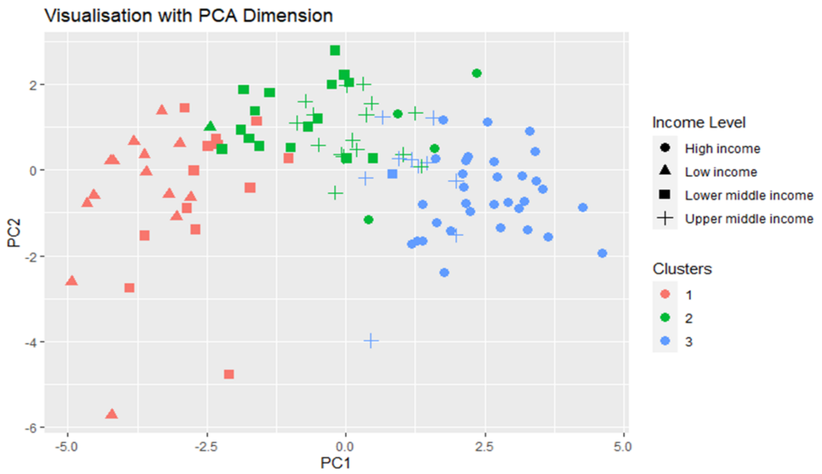
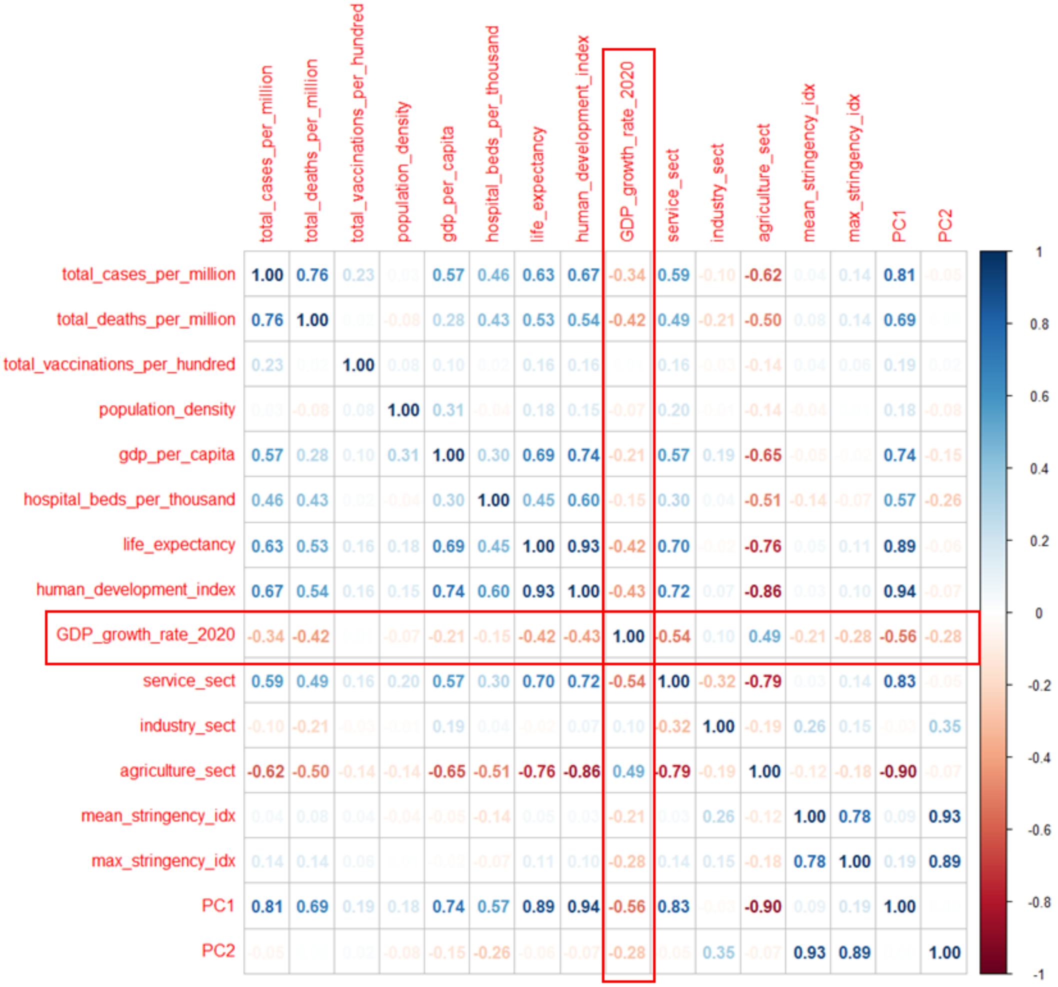


Figure 5 Visualisation with PCA

3.1 Correlation Analysis

Figure X describe the correlation value between each variable. Based on the correlation matrix depicted in Figure X, it can be interpreted that a high negative correlation exists between GDP Growth rate 2020 and service sector (-0.54), Principal Component 1 or PC1 (-0.56), human development index (-0.43), life expectancy (-42), and total deaths per million (-0.42). A moderate negative correlation value exists between the GDP Growth rate 2020 and total cases per million (-0.34). The correlation matrix shows weak negative correlation values exist for GDP per capita (-0.21), hospital beds number per thousand (-0.15), Principal Component 2 or PC2 (-0.28), and both stringency index values. On the other side, there is a strong positive correlation between GDP Growth and the agriculture sector (0.49).



	total_cases_per_million	total_deaths_per_million	total_vaccinations_per_hundred	population_density	gdp_per_capita	hospital_beds_per_thousand	life_expectancy	human_development_index	GDP_growth_rate_2020	service_sect	industry_sect	agriculture_sect	mean_stringency_idx	max_stringency_idx	PC1	PC2
total_cases_per_million	1.00	0.76	0.23	0.03	0.57	0.46	0.63	0.67	-0.34	0.59	-0.10	-0.62	0.04	0.14	0.81	-0.05
total_deaths_per_million	0.76	1.00	0.02	-0.08	0.28	0.43	0.53	0.54	-0.42	0.49	-0.21	-0.50	0.08	0.14	0.69	0.00
total_vaccinations_per_hundred	0.23	0.02	1.00	0.08	0.10	0.02	0.16	0.16	[illegible]	0.16	-0.03	-0.14	0.04	0.06	0.19	0.02
population_density	0.03	-0.08	0.08	1.00	0.31	-0.04	0.18	0.15	-0.07	0.20	-0.01	-0.14	-0.04	0.01	0.18	-0.08
gdp_per_capita	0.57	0.28	0.10	0.31	1.00	0.30	0.69	0.74	-0.21	0.57	0.19	-0.65	-0.05	-0.02	0.74	-0.15
hospital_beds_per_thousand	0.46	0.43	0.02	-0.04	0.30	1.00	0.45	0.60	-0.15	0.30	0.04	-0.51	-0.14	-0.07	0.57	-0.26
life_expectancy	0.63	0.53	0.16	0.18	0.69	0.45	1.00	0.93	-0.42	0.70	-0.02	-0.76	0.05	0.11	0.89	-0.06
human_development_index	0.67	0.54	0.16	0.15	0.74	0.60	0.93	1.00	-0.43	0.72	0.07	-0.86	0.03	0.10	0.94	-0.07
GDP_growth_rate_2020	-0.34	-0.42	[illegible]	-0.07	-0.21	-0.15	-0.42	-0.43	1.00	-0.54	0.10	0.49	-0.21	-0.28	-0.56	-0.28
service_sect	0.59	0.49	0.16	0.20	0.57	0.30	0.70	0.72	-0.54	1.00	-0.32	-0.79	0.03	0.14	0.83	-0.05
industry_sect	-0.10	-0.21	-0.03	-0.01	0.19	0.04	-0.02	0.07	0.10	-0.32	1.00	-0.19	0.26	0.15	-0.03	0.35
agriculture_sect	-0.62	-0.50	-0.14	-0.14	-0.65	-0.51	-0.76	-0.86	0.49	-0.79	-0.19	1.00	-0.12	-0.18	-0.90	-0.07
mean_stringency_idx	0.04	0.08	0.04	-0.04	-0.05	-0.14	0.05	0.03	-0.21	0.03	0.26	-0.12	1.00	0.78	0.09	0.93
max_stringency_idx	0.14	0.14	0.06	0.01	-0.02	-0.07	0.11	0.10	-0.28	0.14	0.15	-0.18	0.78	1.00	0.19	0.89
PC1	0.81	0.69	0.19	0.18	0.74	0.57	0.89	0.94	-0.56	0.83	-0.03	-0.90	0.09	0.19	1.00	0.00
PC2	-0.05	0.00	0.02	-0.08	-0.15	-0.26	-0.06	-0.07	-0.28	-0.05	0.35	-0.07	0.93	0.89	0.00	1.00

Figure 6: Correlation Matrix

Deduced from the correlation matrix, countries which their economy rely on service sectors such as hospitality and tourism, transportation services, Food and beverage services, financial services are countries hit hardest by the COVID-19 Pandemic in terms of GDP growth in 2020. These findings are in line with Fernandes (2020), arguing that the service sector is the most negatively affected by this Pandemic. As several countries perform lockdown, the global travel industry, such as airlines and cruise companies, face a massive reduction in their activities. Hospitality sectors like tourism, restaurants, hotels, casinos, and hotels are deserted as people avoid public spaces. In addition to this, stock markets collapsed in early 2020, indicating the pandemic effect on the financial service sector like banks and stock markets. The relationship between the service sector and the GDP growth rate can be seen in Figure 7. Interestingly, the agriculture sector positively correlates with the GDP growth in 2020, suggesting that countries that rely more on agriculture sectors like farming, fishing, meat and dairy products, and plantations perform better in their GDP growth. This finding contradicts Workie et al (2020), arguing that the change in the food consumption pattern due to lockdowns decrease the output of the industry sector. However, the strong positive correlation is also possibly strengthened by the fact that those farms and plantations are usually located in less dense areas, which minimises the virus' transmission. The relationship between the GDP Growth rate 2020 and the agriculture sector can be seen in Figure 8.

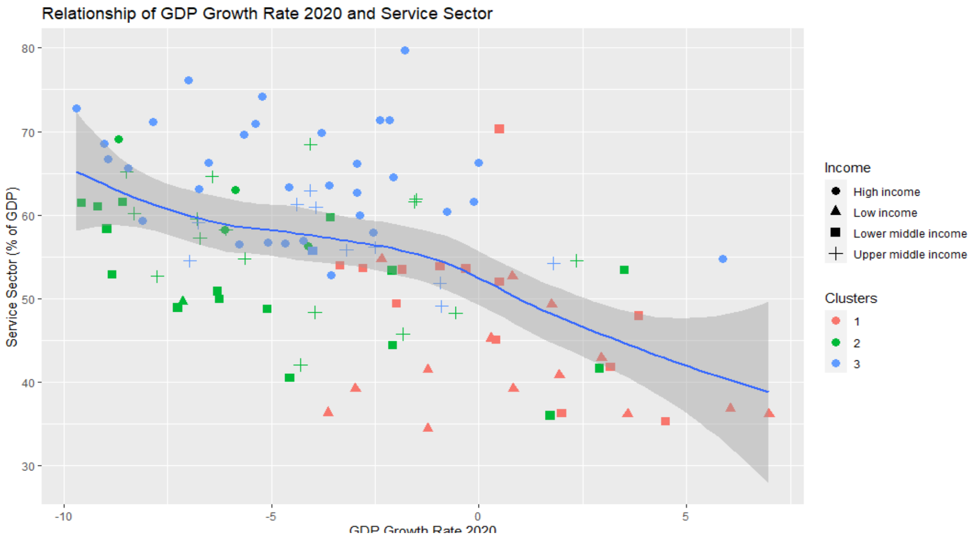


Figure 7 : Relationship of GDP Growth Rate 2020 and Service Sector

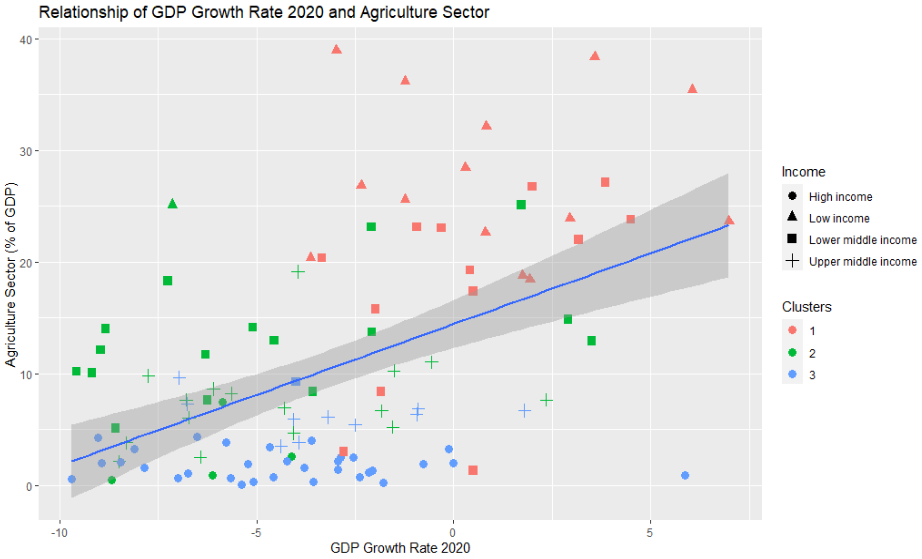


Figure 8: Relationship of GDP Growth Rate 2020 and Agriculture Sector

Regarding disease prevalence variables, we can see that both total cases and total deaths are inversely correlated with GDP Growth Rate in 2020. Figure 9 suggests that the higher covid case and mortality rate in a country, the more severe the effect on their economic sector. However, there is a possible bias in the number of deaths and cases. As shown in figure 9 and 10, countries from cluster 1, which primarily consist of low and lower-middle-income countries, recorded a tiny number of deaths and cases. The small number on the mentioned covid prevalence indicator is possible due to the underreporting case that often occurs in low and lower-middle-income countries subjected to the government's lack of healthcare system funding. (Rizvi et al., 2021).

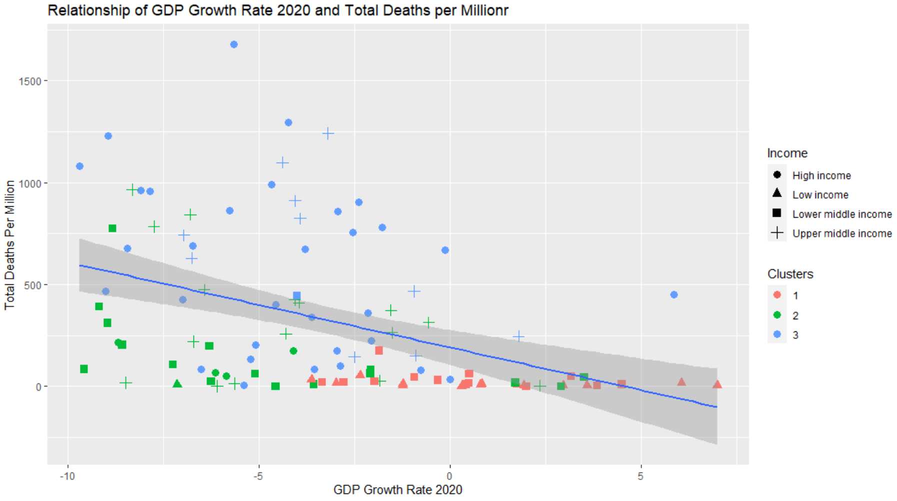


Figure 9: Relationship of GDP Growth Rate 2020 and Total Deaths

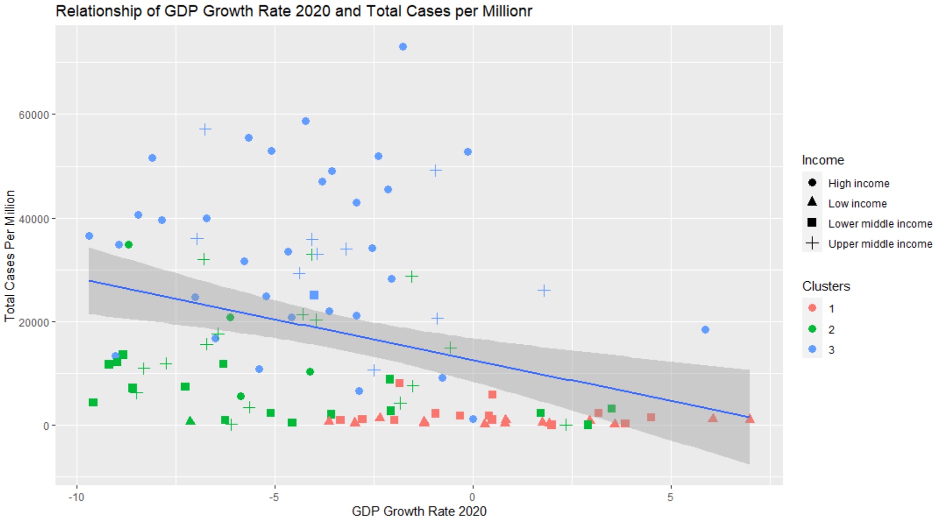


Figure 10: Relationship of GDP Growth Rate and Total Cases

The correlation matrix analysis on healthcare and wellbeing indicators shows that life expectancy and human development index have a strong negative correlation with GDP Growth Rate in 2020. The relationship can be seen in Figures 8 and 9.

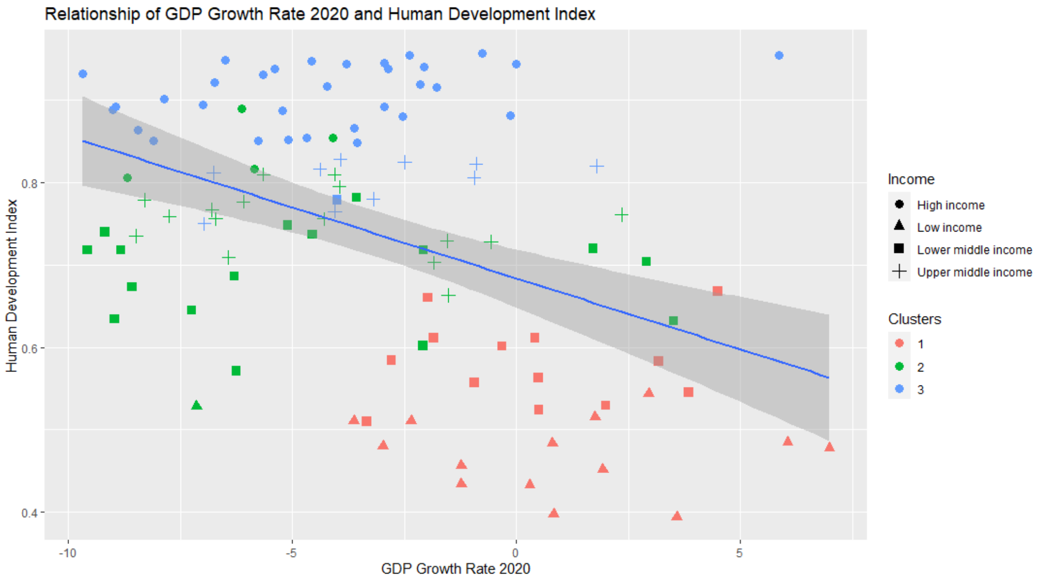


Figure 11: Relationship of GDP Growth Rate and Human Development Index

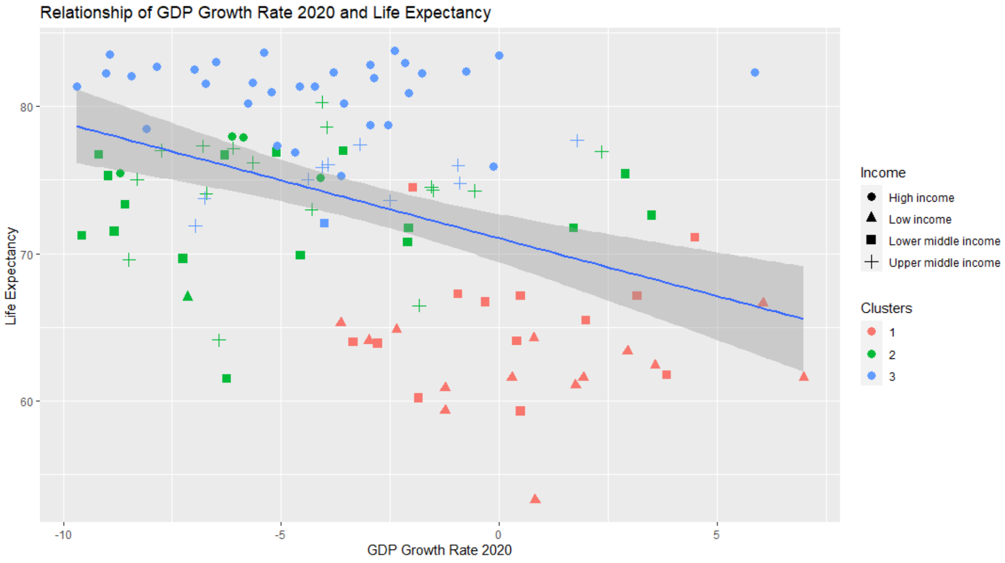


Figure 12: Relationship of GDP Growth Rate 2020 and Life Expectancy

Referring to the correlation results, figures 7 to 12, and table 4 (Cluster mean of variables), we can infer that cluster 1 outperformed other clusters for GDP Growth Rate in 2020. This is because cluster 1 has the highest percentage of the agricultural sector, a variable that resistant to pandemic effect, and the lowest score for service sector, total deaths and cases, life expectancy, and human development index. However, we cannot directly conclude any causal relationship between those variables to GDP Growth Rate 2020. Figure 13 depicts the boxplot of each cluster against the GDP Growth Rate Value.

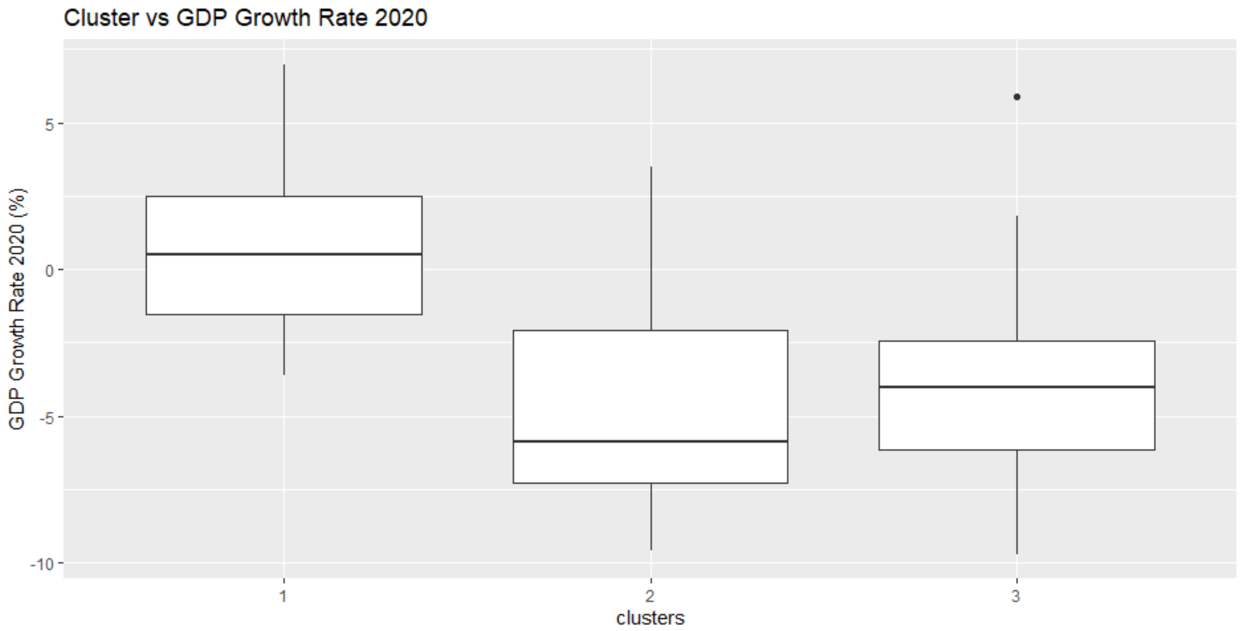


Figure 7: Cluster vs GDP Growth Rate Boxplot

3.3 Multiple Linear Regression Analysis

Variables used for this step are acquired based on the result of the correlation matrix. However, not every strongly correlated value can be used to predict GDP Growth Rate 2020. This is due to the multicollinearity issue, which may affect the higher variance from coefficient, thus giving us an inaccurate picture of the relationship between the two factors. (Farrar, D & Glauber, R. 1967). Hence only total numbers of death, human development index, service sector percentage to GDP and maximum stringency index are used to predict GDP Growth Rate 2020. Other variables with solid, moderate, or low correlation to GDP Growth Rate 2020 with correlation score larger than 0.7 or smaller than -0.7 with either total number of deaths, human development index, service sector percentage to GDP and maximum stringency will not be used to prevent multicollinearity.

Table 5: Attempt Result

Attempt	Variable Used	P-Value	Adjusted R^2 Value
Attempt 1: Cluster 1 Only	• Total deaths per million • Human development index • Service sector • Max stringency	0.287	0.06302
Attempt 2: Cluster 2 Only	• Total deaths per million • Human development index • Service sector	0.0068	0.3717

	• Max stringency		
Attempt 3: Cluster 3 Only	• Total deaths per million • Human development index • Service sector • Max stringency	0.08716	0.1614
Attempt 4: All Data	• Total deaths per million • Human development index • Service sector • Max stringency	2.26E-09	0.3364
Attempt 5: All Data	• PC1 • PC2	1.501E-07	0.2466

The result of the attempts is shown in Table 5. Based on the table, the most statistically significant model is from attempt 4 (2.26E-09), followed by Attempt 5 (1.501E-07) and Attempt 2 (0.0068). However, attempt 1 (0.287) and 3 (0.08716) is not statistically significant. Model with the highest Adjusted R Squared Value is attempt 2 (0.3717), followed by attempt 4 (0.3364), attempt 5 (0.2466), and attempt 3 (0.1614) and attempt 1 (0.063).

Based on the result above, using Principal Component Variable as a representation of other variables decreases the statistically significant and Adjusted R Squared Value compared to the model that uses original variables. This is due to the PCA's properties that trades-off dimensionality reduction and information loss. Based on this result, although PCA can be used for data visualisation, using it as variable in regression cause information loss and decrease the adjust r squared value. However, PCA might be useful in dealing with larger variable number to minimise the computational cost performed by regression.

The result of comparing clustered regression with standard regression show that standard regression performed better. Despite attempt 2, which is done on cluster 2 is having larger Adjusted R Squared Value, the adjusted R Squared Value of attempt 1 and 3 is low and not statistically significant. The reason is indeed performing regression on clustered data may perform better since the similarity of data points within the same cluster is larger. However, due to the bias mentioned in section 3.2, the performance of other regression attempted on cluster 1 and 3 is lower compared to cluster 2 and the combination of all clusters.

Figures 14 and 15 visualise the actual vs predicted value of GDP Growth Rate 2020. In-line with results in Table 5, Cluster 2 shows more datapoints located closer to the regression line which indicates greater regression performance compared to other clusters.

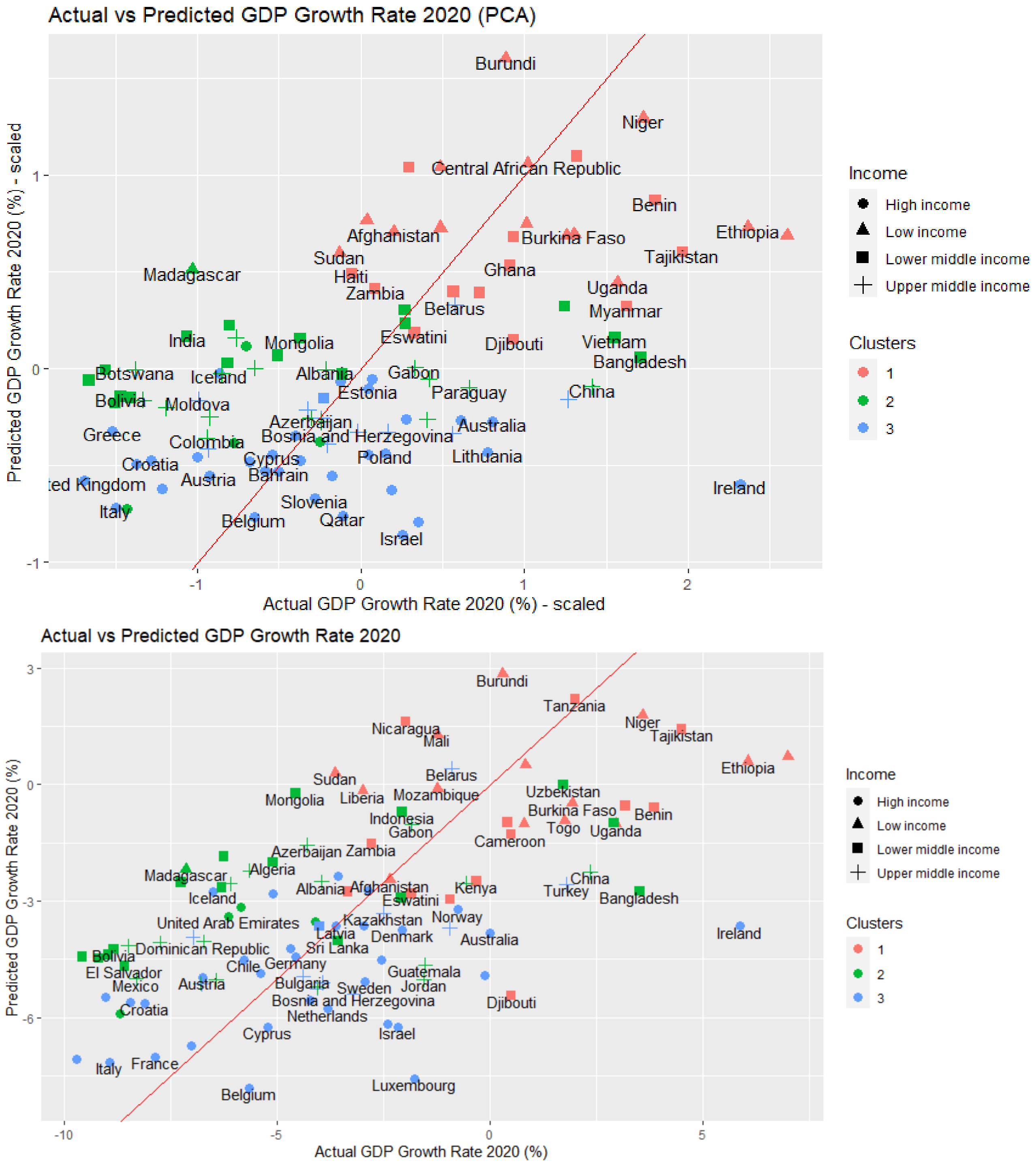


Figure 8 : PCA and Original Variable Comparison

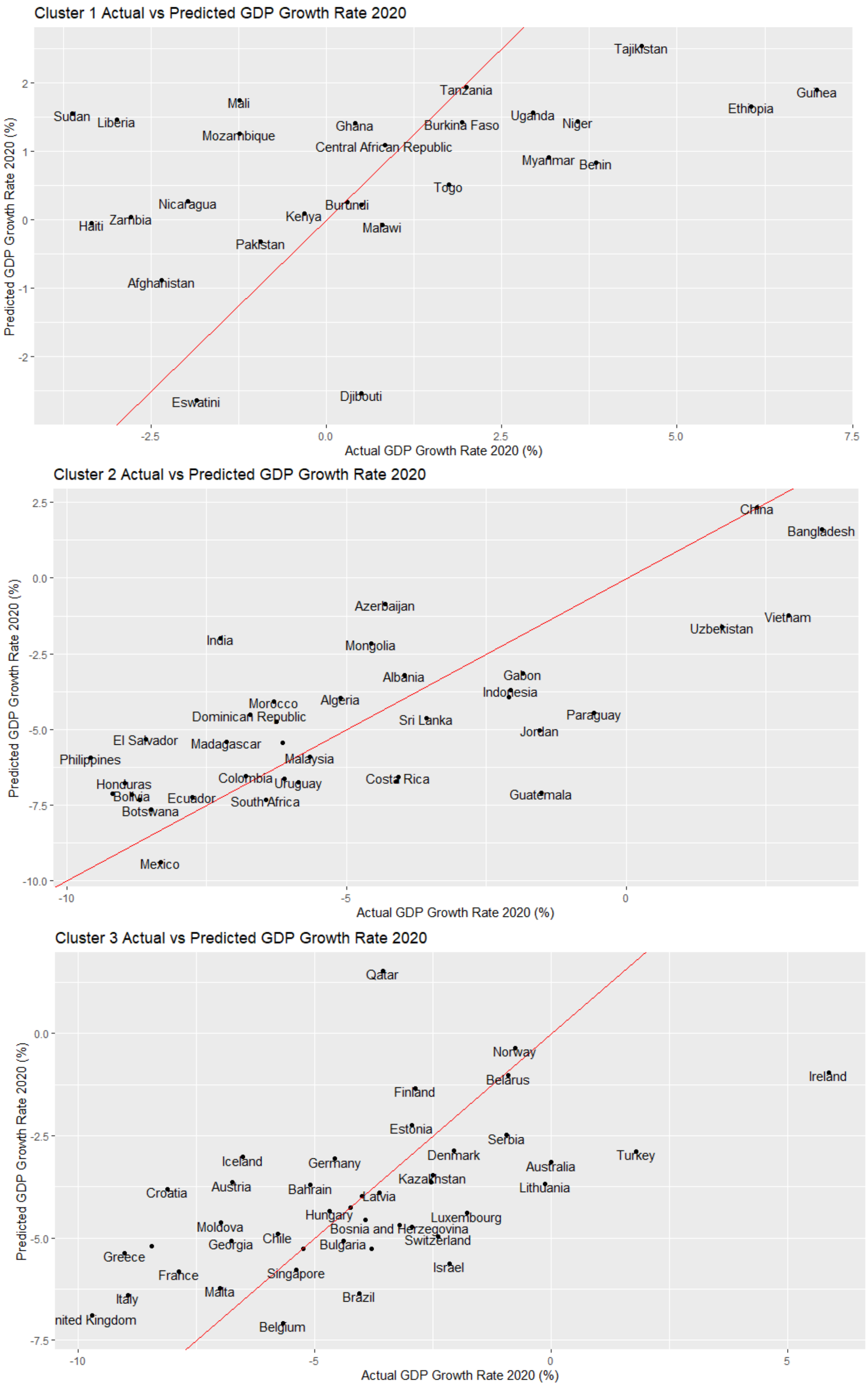


Figure 9: Clusters Comparison

4. Conclusion

4.1 Summary

3 clusters produced by the KMEANS clustering method. The clusters produced by KMEANS algorithm separate the data mainly on the service sector, agriculture sector, total deaths, human development index variables, and life expectancy. Separation of the clusters can be seen using PCA visualisation. In addition to that, the features produced by PCA method also can be used as independent variables in regressing GDP Growth Rate with relatively good results. The service sector is the most correlated variable with GDP Growth Rate, followed by total deaths and cases, life expectancy, and human development index. The high correlation value with the service sector is mainly due to the pandemic's impact on service sectors like hospitality and leisure, finance, education, and transport. PCA decrease the regression's performance due to information loss. However, the comparison between clustered regression and the original Dataset is inconclusive due to bias embedded in the cluster. Underreported cases in low-income countries are one possible suspect of this bias. This bias leads to the bad performance of the regression analysis conducted on clusters 1 and 3.

4.2 Weakness and Further Research

Deriving from the bias issue, additional variables need to be included to ensure that the bias occurred. The additional variables can be the percentage of expenditure for the health care system and other healthcare prevalence variables. In addition to that, adding the number of variables can be used to assess the performance of PCA for regressing because the effectiveness of PCA and other dimensional reduction methods can be seen more clearly with a more significant number of clusters. Other clustering methods can also be conducted to better separate the clusters due to the KMEANS weakness in separating non-spherical clusters. For example, DBSCAN has a better performance in separating clusters within a cluster.

5. Bibliography

Aday, S., & Aday, M. S. (2020). Impact of COVID-19 on the food supply chain. In *Food Quality and Safety* (Vol. 4, Issue 4, pp. 167–180). Oxford University Press. https://doi.org/10.1093/fqsafe/fyaa024

Daniel, S. J. (2020). Education and the COVID-19 pandemic. *Prospects*, *49*(1–2), 91–96. https://doi.org/10.1007/s11125-020-09464-3

Davies, S. E., & Wenham, C. (2020). Why the COVID-19 response needs international relations. *International Affairs*, *96*(5), 1227–1251. https://doi.org/10.1093/ia/iiaa135

Farrar, D. E., & Glauber, R. R. (1967). *Multicollinearity in Regression Analysis: The Problem Revisited* (Vol. 49, Issue 1). https://about.jstor.org/terms

Fernandes, N. (2020). *Economic effects of coronavirus outbreak (COVID-19) on the world economy*. https://ssrn.com/abstract=3557504

The World Bank, World Development Indicators (2012). *GDP growth (annual %)* [Data file]. Retrieved from http://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG

The World Bank, World Development Indicators (2012). *Manufacturing, value added (% of GDP)* [Data file]. Retrieved from http://data.worldbank.org/indicator/NV.IND.MANF.ZS

The World Bank, World Development Indicators (2012). *Services, value added (% of GDP)* [Data file]. Retrieved from http://data.worldbank.org/indicator/NV.SRV.TOTL.ZS

The World Bank, World Development Indicators (2012). *Industry (including construction), value added (% of GDP)* [Data file]. Retrieved from http://data.worldbank.org/indicator/NV.IND.TOTL.ZS

The World Bank, World Development Indicators (2012). *Agriculture, forestry, and fishing, value added (% of GDP)* [Data file]. Retrieved from http://data.worldbank.org/indicator/ NV.AGR.TOTL.ZS

Johor Bahru, U., Darul Ta, J., bin Mohamad, I., Usman, D., & Bahru, J. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303.

Owidbot. (2021). *Data on COVID-19 (coronavirus) by Our World in Data* [Data set]. https://github.com/owid/covid-19-data/tree/master/public/data

Prager, F., Wei, D., & Rose, A. (2017). Total Economic Consequences of an Influenza Outbreak in the United States. *Risk Analysis*, *37*(1), 4–19. https://doi.org/10.1111/risa.12625

Rizvi, S. A., Umair, M., & Cheema, M. A. (2021). Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons and Fractals*, *151*. https://doi.org/10.1016/j.chaos.2021.111240

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1). https://doi.org/10.1088/1757-899X/336/1/012017

Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale, Calif.)*, *6*(2). https://doi.org/10.4172/2161-1165.1000227

Verma, P., Dumka, A., Bhardwaj, A., Ashok, A., Kestwal, M. C., & Kumar, P. (2021). A Statistical Analysis of Impact of COVID19 on the Global Economy and Stock Index Returns. *SN Computer Science*, *2*(1). https://doi.org/10.1007/s42979-020-00410-w

Workie, E., Mackolil, J., Nyika, J., & Ramadas, S. (2020). Deciphering the impact of COVID-19 pandemic on food security, agriculture, and livelihoods: A review of the evidence from developing countries. *Current Research in Environmental Sustainability*, *2*, 100014. https://doi.org/10.1016/j.crsust.2020.100014

Zainal Abidin, N., Ritahani Ismail, A., & Emran, N. A. (2018). Performance Analysis of Machine Learning Algorithms for Missing Value Imputation. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 9, Issue 6). www.ijacsa.thesai.org

6. Code

```
#Libraries
library(tidyverse)
library(dplyr)
library(ggplot2)
library(WDI)
library(corrplot)
library(GGally)
library(factoextra)
library(cluster)
library(fpc)
library(ggpubr)
library(MASS)
library(lubridate)
library(ggstatsplot)

#WDI Loading
new_wdi_cache<-WDIcache()

WDI<-WDI(country = "all",
         indicator = c("NY.GDP.MKTP.KD.ZG",
                       "NV.IND.MANF.ZS",
                       "NV.SRV.TOTL.ZS",
                       "NV.IND.TOTL.ZS",
                       "NV.AGR.TOTL.ZS"),
         start = 2020,
         end = 2020,
         extra = TRUE,
         cache = new_wdi_cache)

#COVID19 Data Loading
covidData<-read.csv("covidData.csv")

covidDF<-data.frame(covidData[,(colnames(covidData) %in% c('iso_code',
                                                            'continent',
                                                            'location',
                                                            'date',

'total_cases_per_million',

'total_deaths_per_million',

'total_vaccinated_per_hundred',
                                                            'gdp_per_capita',

'life_expectancy',

'human_development_index',

'population_density',

'hospital_beds_per_thousand',

'total_vaccinations_per_hundred'))])

#MFeature Engineering on Stringency Index Value
covidDF_stringency<-data.frame(covidData[, (colnames(covidData) %in%
c('iso_code',

'location',

'date',

"stringency_index"))])

stringency_2020<-covidDF_stringency %>% filter(date <= as.Date("2020-12-31"))

#Value for maximum and mean stringency index
stringency_mean<-stringency_2020 %>%
  group_by(location) %>%
  summarise(stringency_index = mean(stringency_index, na.rm = TRUE))

colnames(stringency_mean)<-c("location", "mean_stringency_idx")

stringency_max<-stringency_2020 %>%
  group_by(location) %>%
  summarise(stringency_index = max(stringency_index, na.rm = TRUE))

colnames(stringency_max)<-c("location", "max_stringency_idx")
```

```
#Creating new column names for WDI Data
WDI_DF<-data.frame(WDI$country,
                   WDI$NY.GDP.MKTP.KD.ZG,
                   WDI$NV.SRV.TOTL.ZS,
                   WDI$NV.IND.TOTL.ZS,
                   WDI$NV.AGR.TOTL.ZS)

colnames(WDI_DF)<-c("location",
                    "GDP_growth_rate_2020",
                    "service_sect",
                    "industry_sect",
                    "agriculture_sect")



#Date Filtering and Datasets Merging
covidDF<-filter(covidDF, date=="2020-12-31")
covidDF<-merge(covidDF, WDI_DF,by = "location")
covidDF<-merge(covidDF, stringency_mean, by = "location")
covidDF<-merge(covidDF, stringency_max, by = "location")


#Dealing with data quality
#NA Value Replacement
covidDF<-mutate_at(covidDF, c("total_vaccinations_per_hundred"), ~replace(.,
is.na(.),0.0))

#NA Value Deletion
covidDF<-na.omit(covidDF)


#Outliers Deletion with IQR Method
Q<-quantile(covidDF$GDP_growth_rate_2020, probs = c(.25, .75), na.rm = FALSE)
iqr<-IQR(covidDF$GDP_growth_rate_2020)
up<-Q[2]+1.5*iqr #upper interquartile
low<-Q[2]-1.5*iqr #lower interquartile

covidDF_no_outlier<-subset(covidDF,
                           covidDF$GDP_growth_rate_2020>low &
                             covidDF$GDP_growth_rate_2020<up) #Deleting
datapoints outside interquartile range


#Comparing dataset before and after cleaning
covidDF_no_outlier_check<-covidDF_no_outlier
covidDF_check<-covidDF

covidDF_no_outlier_check$remark<-"Cleaned"
covidDF_check$remark<-"Original"

covidDF_check_ori_and_no_outliers<-rbind(covidDF_no_outlier_check,
covidDF_check)

ggplot(covidDF_check_ori_and_no_outliers, aes(remark, GDP_growth_rate_2020))+
  geom_boxplot()+
  geom_jitter()+
  ggtitle("Original and Cleaned Data Comparisson")+
  xlab("Remark")+
  ylab("GDP Growth Rate 2020")+
  theme(plot.title = element_text(hjust = 0.5))

#Correlation Matric for All Variables
cor_data1<-select_if(covidDF_no_outlier, is.numeric)
cor = cor(cor_data1)
corrplot(cor, method = 'number')

#Data Scaling
covidDF_clean<-na.omit(covidDF_no_outlier)
covidDF_sclaed<-scale(covidDF_clean[,c(4:17)])

#Elbow Analysis
fviz_nbclust(
  covidDF_sclaed,
  FUNcluster = kmeans,
  method = "wss",
  diss = NULL,
  k.max = 10,
  nboot = 100,
  verbose = interactive(),
  barfill = "steelblue",
  barcolor = "steelblue",
  linecolor = "steelblue",
  print.summary = TRUE)

#KMEANS Algorithm with 3 clusters
pamk_result<-pamk(covidDF_sclaed,krange=3,criterion="asw", usepam=TRUE,
                  scaling=FALSE, alpha=0.001, diss=inherits(data, "dist"),
                  critout=FALSE, ns=10, seed=NULL)

clusters<-data.frame(pamk_result$pamobject$clustering)

covidDF_clean<-cbind(covidDF_clean, clusters)
colnames(covidDF_clean)[18]<-"clusters"

covidDF_clean_scaled<-data.frame(cbind(covidDF_clean[,1:3], covidDF_sclaed,
clusters))

#correlation matrix check before and after scaling (result : same correlation
value between variables)
cor_data_clean<-select_if(covidDF_clean, is.numeric)
cor = cor(cor_data_clean)
corrplot(cor, method = 'number')

cor_data_clean_scaled<-select_if(covidDF_clean_scaled, is.numeric)
cor = cor(cor_data_clean_scaled)
corrplot(cor, method = 'number')




#Principal COmponent Analysis
covidDF_clean_scaled_for_PCA<-subset(covidDF_clean_scaled, select=-
c(GDP_growth_rate_2020))
pca<-prcomp(covidDF_clean_scaled_for_PCA[,4:16])
```

```
covidPCA<-data.frame(covidDF_clean_scaled,
                     PC1=pca$x[,1],
                     PC2=pca$x[,2])

colnames(covidPCA)[18]<-"clusters"

#Visualise Datapoints in the PCA Axes
ggplot(complete_data,aes(PC1, PC2))+
  geom_point(aes(col=as.character(clusters), shape = income), size = 3)+
  ggtitle("Visualisation with PCA Dimension")+
  xlab("PC1")+
  ylab("PC2")+
  labs(col = "Clusters", shape = "Income Level")

WDI_select<-data.frame(WDI$country, WDI$income)
colnames(WDI_select)<-c("location", "income")

complete_data<-merge(covidPCA, WDI_select, by = )
complete_data<-subset(complete_data, location != "World")

complete_data_clean<-merge(covidDF_clean, WDI_select, by = )
complete_data_clean<-subset(complete_data_clean, location != "World")

cor_data_clean_scaled_PCA<-select_if(complete_data, is.numeric)
cor_data_clean_scaled_PCA<-subset(cor_data_clean_scaled_PCA, select=-
c(clusters))
cor = cor(cor_data_clean_scaled_PCA)
corrplot(cor, method = 'number')




#Scatter Plot Service Sector vs GDP Growth Rate
ggplot(complete_data_clean,aes(GDP_growth_rate_2020, service_sect))+
  geom_point(aes(col=as.character(clusters), shape = income), size = 3)+
  geom_smooth()+
  ggtitle("Relationship of GDP Growth Rate 2020 and Service Sector")+
  xlab("GDP Growth Rate 2020")+
  ylab("Service Sector (% of GDP)")+
  labs(col = "Clusters", shape = "Income")

#Scatter Plot Total Deaths vs GDP Growth Rate
ggplot(complete_data_clean,aes(GDP_growth_rate_2020,
total_deaths_per_million))+
  geom_point(aes(col=as.character(clusters), shape = income), size = 3)+
  geom_smooth(method = lm)+
  ggtitle("Relationship of GDP Growth Rate 2020 and Total Deaths per
Millionr")+
  xlab("GDP Growth Rate 2020")+
  ylab("Total Deaths Per Million")+
  labs(col = "Clusters", shape = "Income")

#Scatter Plot Total Cases vs GDP Growth Rate
ggplot(complete_data_clean,aes(GDP_growth_rate_2020,
total_cases_per_million))+
  geom_point(aes(col=as.character(clusters), shape = income), size = 3)+
  geom_smooth(method = lm)+
  ggtitle("Relationship of GDP Growth Rate 2020 and Total Cases per
Millionr")+
  xlab("GDP Growth Rate 2020")+
  ylab("Total Cases Per Million")+
  labs(col = "Clusters", shape = "Income")

#Scatter Plot Agriculture Sector vs GDP Growth Rate
ggplot(complete_data_clean,aes(GDP_growth_rate_2020, agriculture_sect))+
  geom_point(aes(col=as.character(clusters), shape = income), size = 3)+
  geom_smooth(method = lm)+
  ggtitle("Relationship of GDP Growth Rate 2020 and Agriculture Sector")+
  xlab("GDP Growth Rate 2020")+
  ylab("Agriculture Sector (% of GDP)")+
  labs(col = "Clusters", shape = "Income")

#Scatter Plot Human Development Index vs GDP Growth Rate
ggplot(complete_data_clean,aes(GDP_growth_rate_2020,
human_development_index))+
  geom_point(aes(col=as.character(clusters), shape = income), size = 3)+
  geom_smooth(method = lm)+
  ggtitle("Relationship of GDP Growth Rate 2020 and Human Development
Index")+
  xlab("GDP Growth Rate 2020")+
  ylab("Human Development Index")+
  labs(col = "Clusters", shape = "Income")

#Scatter Plot Life Expectancy vs GDP Growth Rate
ggplot(complete_data_clean,aes(GDP_growth_rate_2020, life_expectancy))+
  geom_point(aes(col=as.character(clusters), shape = income), size = 3)+
  geom_smooth(method = lm)+
  ggtitle("Relationship of GDP Growth Rate 2020 and Life Expectancy")+
  xlab("GDP Growth Rate 2020")+
  ylab("Life Expectancy")+
  labs(col = "Clusters", shape = "Income")


#Boxplot Clusters and GDP Growth Rate
ggplot(complete_data_clean, aes(as.character(clusters),
GDP_growth_rate_2020))+
  geom_boxplot()+
  ggtitle("Cluster vs GDP Growth Rate 2020")+
  xlab("clusters")+ylab("GDP Growth Rate 2020 (%)")



#Create data frames for each clusters
cluster1<-filter(complete_data_clean, clusters == 1)
cluster2<-filter(complete_data_clean, clusters == 2)
cluster3<-filter(complete_data_clean, clusters == 3)

#correlation matrix for cluster 1
cor_data_m1<-select_if(cluster1, is.numeric)
cor = cor(cor_data_m1)
corrplot(cor, method = 'number')

#correlation matrix for cluster 2
cor_data_m2<-select_if(cluster2, is.numeric)
```

```
cor = cor(cor_data_m2)
corrplot(cor, method = 'number')

#correlation matrix for cluster 3
cor_data_m3<-select_if(cluster3, is.numeric)
cor = cor(cor_data_m3)
corrplot(cor, method = 'number')

#Linear Regression Model for Each Cluster
#Multiple Linear Regression for Attempt 3 (Cluster 1)
model1<-lm(formula = GDP_growth_rate_2020~
             total_deaths_per_million+
             human_development_index+
             service_sect+
             max_stringency_idx,
           data = cluster1)

summary(model1)

#Multiple Linear Regression for Attempt 3 (Cluster 2)
model2<-lm(formula = GDP_growth_rate_2020~
             total_deaths_per_million+
             human_development_index+
             service_sect+
             max_stringency_idx,
           data = cluster2)

summary(model2)

#Multiple Linear Regression for Attempt 3 (Cluster 3)
model3<-lm(formula = GDP_growth_rate_2020~
             total_deaths_per_million+
             human_development_index+
             service_sect+
             max_stringency_idx,
           data = cluster3)

summary(model3)



##Multiple Linear Regression for Attempt 5 (PC1 and PC2 Variables)
model_all_PCA<-lm(formula = GDP_growth_rate_2020~
                    PC1+
                    PC2,
                  data = complete_data)

summary(model_all_PCA)

#Multiple Linear Regression for Attempt 4 (All clusters with original
variables)
model_all_clean<-lm(formula = GDP_growth_rate_2020~total_deaths_per_million+
                      human_development_index+
                      service_sect+
                      max_stringency_idx, data = covidDF_clean)
```

```
summary(model_all_clean)


#Plotting Performance from each attempts


gdp_prediction_cluster1<-data.frame(location = cluster1$location,
                                actual = cluster1$GDP_growth_rate_2020,
                                pred = model1$fitted.values)

gdp_prediction_cluster2<-data.frame(location = cluster2$location,
                                    actual = cluster2$GDP_growth_rate_2020,
                                    pred = model2$fitted.values)

gdp_prediction_cluster3<-data.frame(location = cluster3$location,
                                    actual = cluster3$GDP_growth_rate_2020,
                                    pred = model3$fitted.values)

#Plotting for Attempt 1 (Cluster 1)
cluster1_plot<-ggplot(data = gdp_prediction_cluster1, aes(actual, pred))+
  geom_point()+
  geom_abline(mapping=aes(slope=1, intercept = 0), color='red')+
  geom_text(aes(label = location), vjust = 0, nudge_y=-0.1, check_overlap =
TRUE)+
  ggtitle("Cluster 1 Actual vs Predicted GDP Growth Rate 2020")+
  xlab("Actual GDP Growth Rate 2020 (%)")+
  ylab("Predicted GDP Growth Rate 2020 (%)")
cluster1_plot

#Plotting for Attempt 2 (Cluster 2)
cluster2_plot<-ggplot(data = gdp_prediction_cluster2, aes(actual, pred))+
  geom_point()+geom_abline(mapping=aes(slope=1, intercept = 0), color='red')+
  geom_text(aes(label = location), vjust = 0, nudge_y=-0.2, check_overlap =
TRUE)+
  ggtitle("Cluster 2 Actual vs Predicted GDP Growth Rate 2020")+
  xlab("Actual GDP Growth Rate 2020 (%)")+
  ylab("Predicted GDP Growth Rate 2020 (%)")
cluster2_plot

#Plotting for Attempt 3 (Cluster 3)
cluster3_plot<-ggplot(data = gdp_prediction_cluster3, aes(actual, pred))+
  geom_point()+geom_abline(mapping=aes(slope=1, intercept = 0), color='red')+
  geom_text(aes(label = location), vjust = 0, nudge_y=-0.2, check_overlap =
TRUE)+
  ggtitle("Cluster 3 Actual vs Predicted GDP Growth Rate 2020")+
  xlab("Actual GDP Growth Rate 2020 (%)")+
  ylab("Predicted GDP Growth Rate 2020 (%)")
cluster3_plot

#Plotting for Attempt 5 (PCA Regression)
gdp_prediction_all_pca<-data.frame(location = complete_data$location,
                                   actual =
complete_data$GDP_growth_rate_2020,
                                   pred = modelPCA$fitted.values)
```

```
gdp_prediction_all_pca<-merge(complete_data, gdp_prediction_all_pca, by =
"location")

all_pca<-ggplot(data = gdp_prediction_all_pca, aes(actual, pred))+
  geom_point(aes(col=as.character(clusters), shape = income), size = 3)+
  geom_abline(mapping=aes(slope=1, intercept = 0), color='red')+
  ggtitle("Actual vs Predicted GDP Growth Rate 2020 (PCA)")+
  xlab("Actual GDP Growth Rate 2020 (%) - scaled")+
  ylab("Predicted GDP Growth Rate 2020 (%) - scaled")+
  geom_text(aes(label = location), vjust = 0, nudge_y=-0.05, check_overlap =
TRUE)+
  labs(col = "Clusters", shape = "Income")

all_pca

#Plotting for Attempt 4 (Unclustered with Original Variable)
gdp_prediction_all_clean<-data.frame(location = complete_data_clean$location,
                                     actual =
complete_data_clean$GDP_growth_rate_2020,
                                     pred = model_all_clean$fitted.values)

gdp_prediction_all_clean<-merge(gdp_prediction_all_clean, WDI_select,by =
"location")
gdp_prediction_all_clean<-merge(gdp_prediction_all_clean,
complete_data_clean, by = "location")

all_clean<-ggplot(data = gdp_prediction_all_clean, aes(actual, pred))+
  geom_point(aes(col=as.character(clusters), shape = income.y), size = 3)+
  geom_abline(mapping=aes(slope=1, intercept = 0), color='red')+
  ggtitle("Actual vs Predicted GDP Growth Rate 2020")+
  xlab("Actual GDP Growth Rate 2020 (%)")+
  ylab("Predicted GDP Growth Rate 2020 (%)")+
  geom_text(aes(label = location), vjust = 0, nudge_y=-0.3, check_overlap =
TRUE)+
  labs(col = "Clusters", shape = "Income")
geom_text(aes(label = location))
all_clean
```