MA335-Coursework

The study of the relationship between WDI indicators and the severity of the COVID-19 pandemic

Created by: Khwanchanok Chumkhun

Registration Number: 2110866

Date: Mar 31st, 2022.

Table of Contents

Introduction

This study aims to investigate the relationship between the World Development Indicators (WDI), and the severity of the COVID-19 pandemic. The dataset includes a selection of the WDI, derived from a primary World Bank database, and information about the casualties of the COVID-19 pandemic in terms of the Number of COVID deaths per 1M people as of 1st March 2022 (*variable name: Covid_deaths*).
In the first task, univariate analysis and bivariate analysis will be used to examine the descriptive statistics of the dataset. And the result will be shown
After that, the clustering technique will be applied to study the relation between WDI data and the Continent.
Then, in the third task, after classifying the *Covid_deaths* into a binary variable, a logistic regression model will be implemented to predict COVID causalities.
Next, same as the previous task, the *Covid_deaths* will be transformed to be a classification variable but this time with 4 levels, and then implementing Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA) and logistic regression to study this multiclass classification problem.
Finally, as a conclusion, the ability to use WDI data to predict the causalities of similar pandemics will be discussed. Moreover, the response to COVID-19 pandemic of countries with similar economic profiles will be summarized in this section.
All statistical studies in this report will be conducted using the R language with MS excel and MS word support to improve the visualization.

< Word count : 1895 words >

Task1

Task: Analyse using descriptive statistics (both graphical and numerical representations) the dataset project data.csv. Generate an appropriate table as a summary and appropriate graphs-for example boxplots, histograms, and scatterplots.

The main purpose of this task is to analyze the dataset using descriptive statistics.
Before starting this task, data preparation processes are required to clean this 185 rows x 20 columns dataset. ‘Covid.deaths’ column which represents the number of COVID deaths per 1M people should be converted the variable type from character to be numeric. Moreover, the NA values (missing value) in

each column are replaced by mean or median depending on the existence of the outliers. (Replacing by the median in case data contains outliers).

After the data preparation processes, the univariate analysis is implemented to clarify the min, max, mean, median, also standard deviation of the WDI numeric variables. And the result is shown in the Table1 below.

Table:1 Summary of each numerical variable in the dataset (see full table in apprendices)

	min	max	median	mean	sd
Covid.deaths	3	6252	711	1143.38	1219.88
Life.expec	54.24	85.08	74.23	73.01	7.32
Elect.access	6.72	100	100	85.73	25
Net.nat.income	-14.38	50.17	3.65	4.1	5.18
Net.nat.income.capita	-17.35	47.25	2.64	2.85	4.98
Mortality.rate	1.6	82.4	13.05	19.94	18.79
Primary	54.73	120.45	97.41	94.61	10.44

The result in the table1 descript raw data in form of descriptive statistics. 2 main summary statistics (the centre of the data and the data dispersion) are shown. The mean and median show the centre of the data, min, max together with sd show how data is dispersed.

For the Covid.deaths in this table can show that the range of min and max is large. While there is a country where people only are dead 3 per 1M, in another country people are dead from covid 6252 per 1M.

Figure1: Comparison of Number of COVID deaths per 1M people in each Continent.

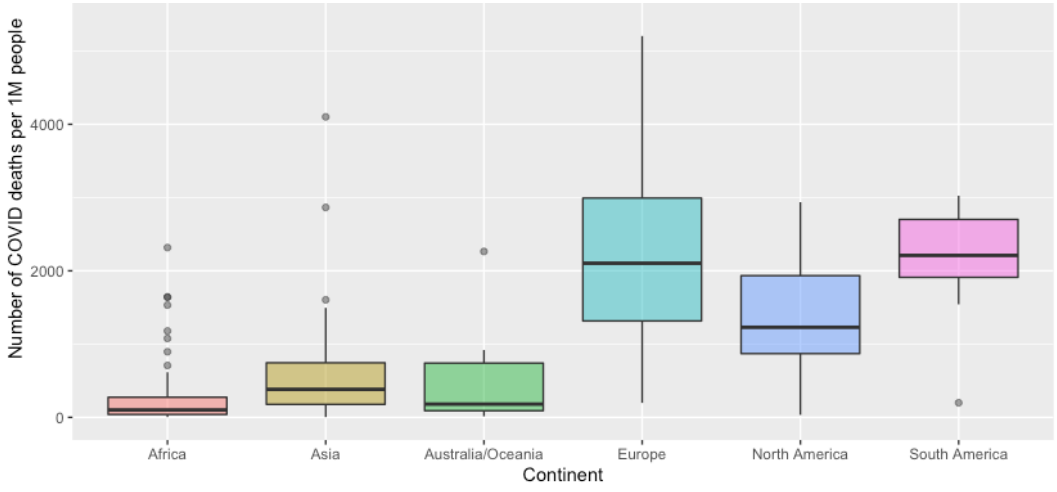


Figure1 shows the ratio of people who died from COVID in each Continent. On average, Europe and South America are the 2 continents that have the highest rate of COVID deaths. While the median of deaths rate from COVID is the lowest one. But the boxplot shows that the distribution of data in Europe is normal, the distribution is symmetric below and upper median line, moreover there is no outlier in the data. Whereas the distribution of death ratio in Africa is abnormal, almost all data are distributed above the median line and data contains many outliers.

Task2

Task: Implement clustering algorithms (leaving out Continent, and Covid deaths). Comment on the results of clustering in relation to the Continent variable.

This task is to implement clustering algorithms by leaving out Continent, and Covid deaths. To proceed this task, first, fviz_nbclust is used to find the optimal number of clusters by using the elbow method which calculates the total sum of square within the cluster. This method selects the optimal number of clusters k at the point of a change of slope from steep to shallow. And the result presents in the figure below shows that the optimal number of clusters of this dataset is k=3.

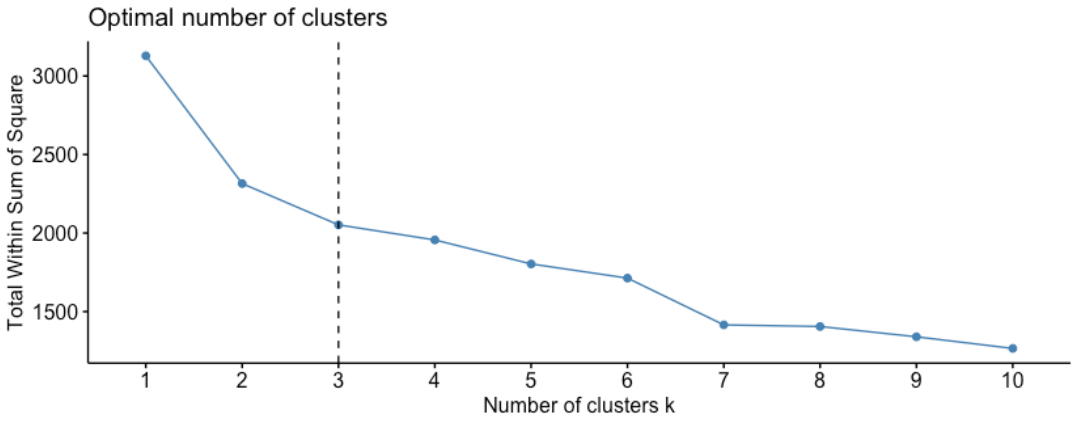


Figure2: The optimal number of clusters by K-means algorithm

After getting the optimal number of clusters, the k-means algorithm is implemented to group the country that has a similar development indicator level into the same cluster. Then we do visualize and summarize the members of each cluster separate by continent and the result is shown below.

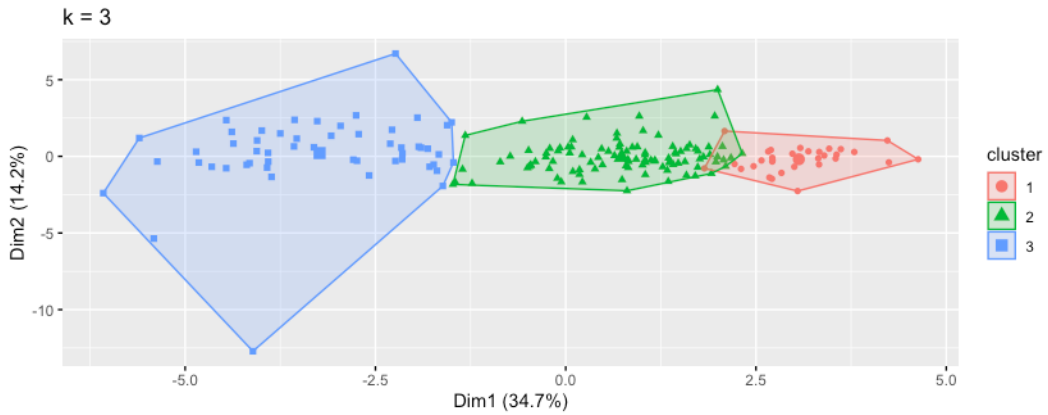


Figure:3 WDI Clustering result (see the plot with more no. of K in apprendice)

Table:2 Summary of each numerical variable in the dataset

	Cluster1	Cluster2	Cluster3
Africa	0	10	41
Asia	6	31	9
Australia/Oceania	2	2	2
Europe	21	25	0
North America	3	20	1
South America	0	12	0
Total	**32**	**100**	**53**

32 countries are belonged to cluster1, while the other 100 and 53 countries are in cluster2 and cluster 3 respectively. Regarding the continent, we can see that all countries in South America have the same development indicator, which can say that they all have a similar economic profile level. Whereas the most diversity of development indicator is Australia/Oceania, in which all 6 countries are distributed in all 3 clusters at the same ratio.

Task3

Task: Transform Covid deaths into a binary variable. Fit a logistic regression model using the remaining variables to predict high COVID causalities. Describe the produced model and comment on what it demonstrates.

In the beginning, the Covid_deaths variable which is now numerical data should be converted to a binary classification variable using the mean (=1143.378) as a cutoff point. The data in which the value is higher than its mean is classified as the ‘Up’ class, whereas the remaining are in the ‘Down’ group. There are 72 data and 113 data in Up and Down classification respectively.

After that, the 2 features selection algorithms, Wrap: Forward and Backward, are carried out. This is an essential process to identify the important variables and eliminate the irrelevant features from the model to avoid overfitting, make the model interpretable and it can also reduce the computation cost in case of huge data. And by comparing the AIC, the Forward algorithm, which provides less AIC score of 158.62, is selected to be the way for the feature selections of this data set (AIC of Backward is 160.15). And by implementing the logistic regression model between the response variable and 7 predictors, the model and the summary result are as below.

Logistic regression moedel : Covid_deaths ~ Pop.growth + Mortality.rate + Unemployment + Health.exp + Comp.education + Life.expec + Pop.total

Table3: Result of summary of the logistic regression model.

Coefficients:					
	Estimate	Std. Error	Z value	Pr(>|z|)	
(Intercept)	8.26E+00	5.95E+00	1.387	0.16531	
Pop.growth	-7.20E-01	2.64E-01	-2.729	0.00636	**
Mortality.rate	-1.30E-01	3.98E-02	-3.275	0.00106	**
Unemployment	1.60E-01	6.02E-02	2.662	0.00777	**
Health.exp	2.95E-01	1.09E-01	2.705	0.00684	**
Comp.education	2.01E-01	9.66E-02	2.085	0.03705	*
Life.expec	-1.47E-01	7.77E-02	-1.895	0.05805	.
Pop.total	-4.28E-09	5.10E-09	-0.839	0.40127	

The estimate column shows the change of response value in association with the change of each predictor. For instance, in this case, the increase of the predictor variable Pop.grownth is affected with an average change of -0.72 in the log-odds of the response variable (Covid_Deaths).

P-Value which is related to the z-value, explains how much each predictor is able to predict the response variable. And here, we can say that we have strong evidence that Pop.growth, Mortality.rate, Unemployment and Heath.exp are associated with the response and must be included in the model.

After the modelling process is done, the model evaluation is conducted, In this case, we will use the same dataset we use to train the model to evaluate the accuracy of the prediction. (In the actual case we should separate training data and evaluate data but, in this report, it will be done in Task 4.) Using the confusion matrix then calculate the mean of the prediction accuracy which the prediction accuracy of this model is about 77%. (This process will be conducted and presented in task4).

Task4

Task: Transform Covid deaths into a categorical variable with 4 possible labels. Consider whether some manipulation of the dataset should be implemented before applying your learning algorithms. Implement QDA, LDA and logistic regression for this multiclass classification problem.
Compare the results using appropriate validation techniques and performance metrics.

For this task, some manipulation of the dataset is required. The dataset is randomly separate into 2 group, training data and validation data. Training data is 80% of all data in dataset, this data is used to train or build the model. Whereas the validation data is, the remaining 20%, is used to evaluate the accuracy of the model.
Same as in task3, the Covid_deaths variable which is numerical data should be converted to a classification variable. But this time it is divided into 4 classes using the 1st quartile, 2nd quartile, 3rd quartile and the 4th quartile (maximum) as indicators. The data in which the value is lower than its 1st quartile, 2nd quartile, 3rd quartile and the 4th quartile value is classified as the 'Q1', 'Q2', 'Q3' and 'Q4' class respectively. And again, the features selection algorithm is performed, in this task, the Boruta algorithm which is a wrapper built around the random forest classification algorithm is implemented. And 10 important predictor variables are selected as shown in the figure below.

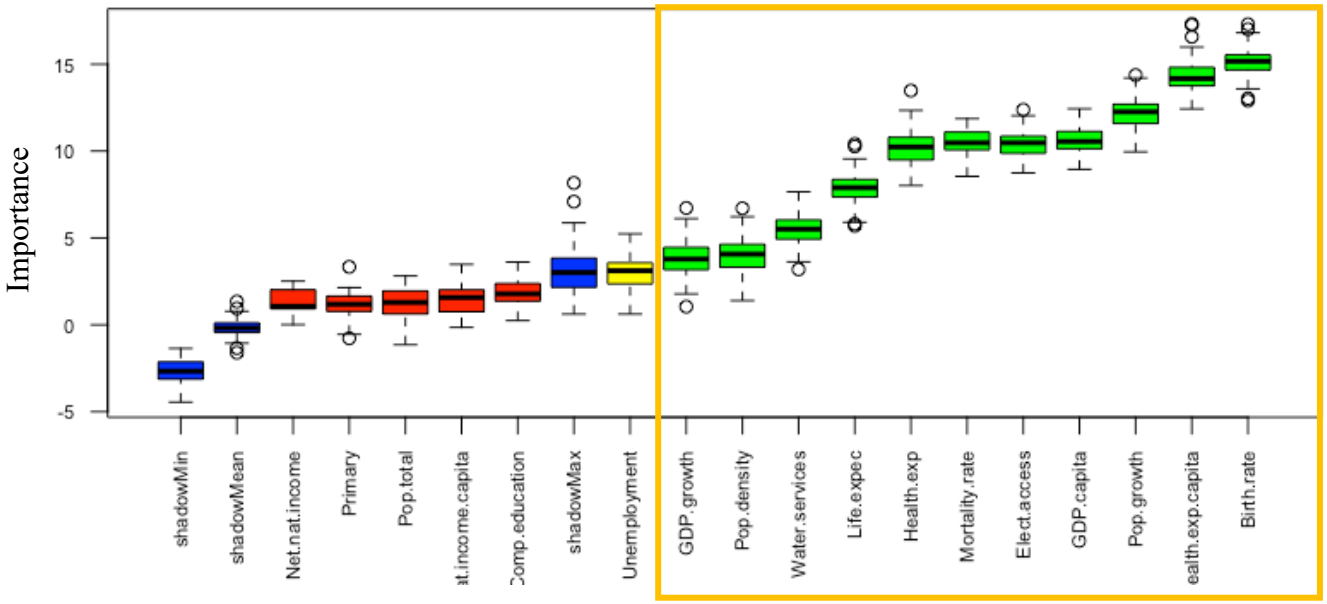


Figure:4 Features selection by Boruta algorithm

Subsequently, these 11 predictors to build are utilized to implement the LDA, QDA, and logistic regression models from training data.

After the models are built, the prediction process is conducted by using evaluation data, the confusion matrix is used to summarize the prediction result of the classification models. The matrix can explain not just only the error that is made by the classifier but also can tell what type of error is made (True but predict as False or False predict as True). And the comparison of the prediction result of LDA, QDA and logistic regression models is presented below.

Table:4 comparison of the prediction result

		Actual Value											
		LDA				QDA				GLM			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Predict Value	Q1	7	1	0	0	8	3	5	0	7	1	0	0
	Q2	1	4	3	3	0	3	2	3	1	3	1	3
	Q3	0	1	5	1	0	0	0	0	0	3	6	0
	Q4	1	3	1	5	1	3	2	6	1	2	2	6
	Accuracy	0.58				0.47				0.61			

From the summary table above, it can conclude that logistic regression for this multiclass shows the best result as a model to create a multiclassification prediction model for this dataset, having an accuracy of prediction of 61%.

The accuracy might be increased by comparing the features selection process with multiple methods then selecting the best one. On the other hand, since data represented the WDI of each country in the world the number of rows of data could not be added for more accuracy of the model.

Task5

Task: Discuss to what extent WDI data can be used to predict the causalities of similar pandemics. Summarise what can be learnt from this data about the response to COVID-19 pandemic of countries with similar economic profiles.

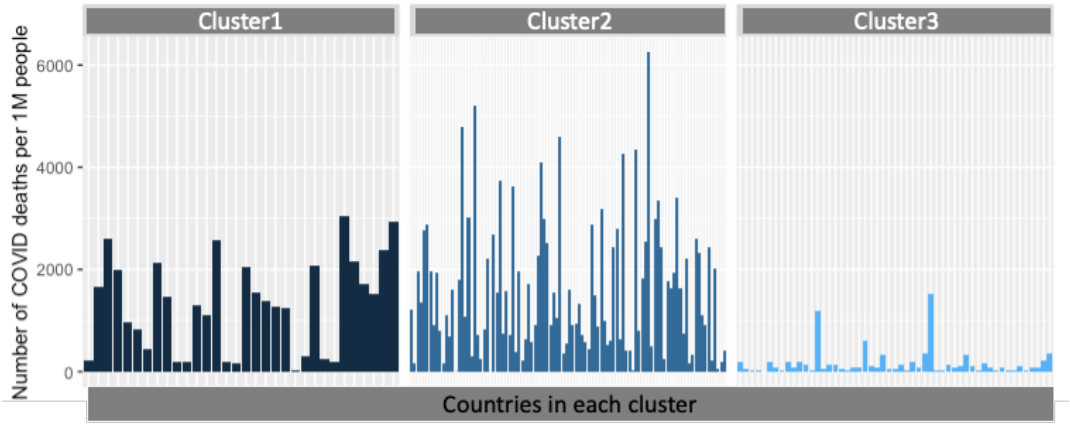


Figure:5 Comparison of Covid_Death in each country grouping by Cluster

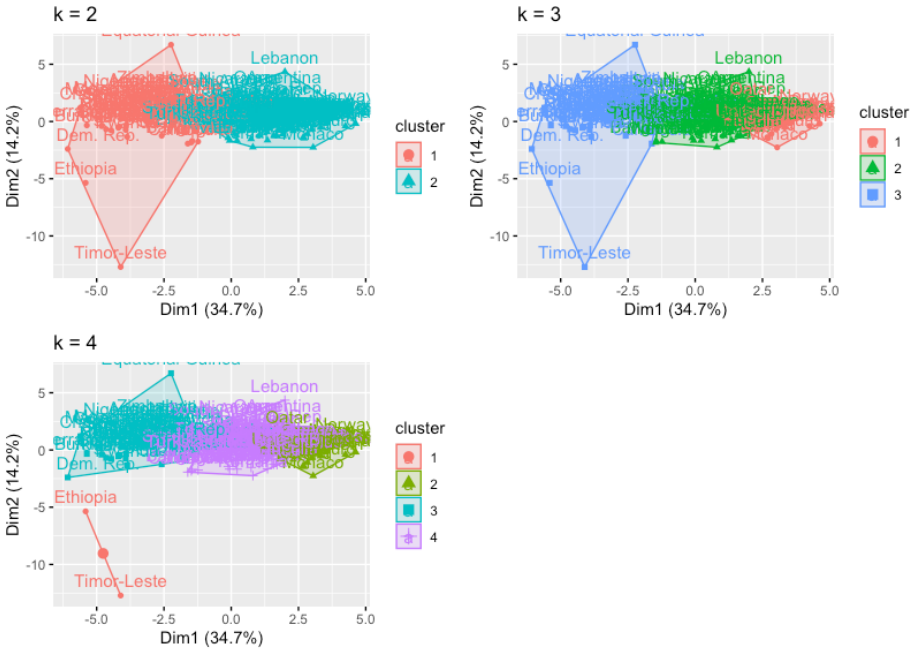
The studies in this report can illustrate that the WDI can be utilized as predictor data to predict the COVID pandemic causalities. The above figure shows that the countries that have a similar economic profile have a slightly similar ratio of COVID deaths. It might be because of the countries with similar economic profiles have similar public health and educational system. By the way, if similar pandemics occur in the future the WDI data could be preliminarily utilized to predict the causalities. There are other factors that can impact the pandemic's causalities, such as the government rule, the culture or behaviour of people in the country. As conclusion, we cannot say that WDI is the best data set to predict the situation since the prediction result showed only 56% in the previous task. But to some extent, we can utilize WDI to predict the pandemic's causalities can be one of the preliminary options.

Appendices

[Full table of Table1]

	min	max	median	mean	sd
Covid.deaths	3	6252	711	1143.38	1219.88
Life.expec	54.24	85.08	74.23	73.01	7.32
Elect.access	6.72	100	100	85.73	25
Net.nat.income	-14.38	50.17	3.65	4.1	5.18
Net.nat.income.capita	-17.35	47.25	2.64	2.85	4.98
Mortality.rate	1.6	82.4	13.05	19.94	18.79
Primary	54.73	120.45	97.41	94.61	10.44
Pop.growth	-1.61	4.47	1.16	1.24	1.12
Pop.density	2.07	19223.98	84.2	377.32	1634.91
Pop.total	33706	1407745000	9370390.5	41003338.6	148876303
Health.exp.capita	19.85	10921.01	370.11	1116.28	1836.1
Health.exp	1.53	16.77	6.24	6.41	2.46
Unemployment	0.1	28.47	5.35	6.56	4.18
GDP.growth	-7.16	19.54	2.65	2.9	3.02
GDP.capita	228.21	189487.1	6731.24	18261.55	28298.06
Birth.rate	5.9	45.64	17.57	19.39	9.75
Water.services	5.58	100	73.6	73.6	22.49
Comp.education	5	17	10	9.89	2.35

[Visualization of K-means = 2,3 and 4]



[R code]

```
library(knitr)
library(dplyr)
library(dsEssex)
library(tidytext)
library(tidyverse)
library(class)
library(factoextra)
library(gridExtra)
library(AICcmodavg)
library(caret)
library(Boruta)
library(ggplot2)
library(psych)
library(GGally)
library(MASS)
library(caret)
library(tidyr)
library(nnet)

# set working directory
setwd("/Volumes/Macintosh HD - Data/MA335_Final")
#load data
Project_data <- read.csv("project_data.csv")
# confirm data
str(Project_data)
attach(Project_data)
############## Data Preparation ##############################
#Clean Covid.deaths and change variable type from character to be numeric
Project_data$Covid.deaths <- Covid.deaths %>% str_replace_all(",","") %>% as.numeric()
Project_data$Comp.education <- as.numeric(Comp.education)
```

```
str(Project_data)
#Check outliers and NA by boxplot
sum(is.na(Project_data))
num_only <- Project_data[,3:20]
boxplot(num_only)
#Replace NA by mean or median
Project_data$Life.expec[is.na(Life.expec)] <- mean(Project_data$Life.expec,na.rm = T)
Project_data$Net.nat.income[is.na(Net.nat.income)] <- median(Net.nat.income,na.rm = T)
Project_data$Net.nat.income.capita[is.na(Net.nat.income.capita)] <-
median(Net.nat.income.capita,na.rm = T)
Project_data$Mortality.rate[is.na(Mortality.rate)] <- median(Mortality.rate,na.rm = T)
Project_data$Primary[is.na(Primary)] <- median(Primary,na.rm = T)
Project_data$Pop.growth[is.na(Pop.growth)] <- median(Pop.growth,na.rm = T)
Project_data$Pop.density[is.na(Pop.density)] <- median(Pop.density,na.rm = T)
Project_data$Pop.total[is.na(Pop.total)] <- median(Pop.total,na.rm = T)
Project_data$Health.exp.capita[is.na(Health.exp.capita)] <- median(Health.exp.capita,na.rm = T)
Project_data$Health.exp[is.na(Health.exp)] <- median(Health.exp,na.rm = T)
Project_data$Unemployment[is.na(Unemployment)] <- median(Unemployment,na.rm = T)
Project_data$GDP.growth[is.na(GDP.growth)] <- median(GDP.growth,na.rm = T)
Project_data$GDP.capita[is.na(GDP.capita)] <- median(GDP.capita,na.rm = T)
Project_data$Birth.rate[is.na(Birth.rate)] <- mean(Birth.rate,na.rm = T)
Project_data$Water.services[is.na(Water.services)] <- mean(Water.services,na.rm = T)
Project_data$Comp.education[is.na(Comp.education)] <- median(Comp.education,na.rm = T)
#recheck that all NA are removed
sum(is.na(Project_data))
############## Task1 ############################
#Create describe table and select only required variable
des.table <- describe(round(num_only,2))[,c(8,9,4,3,5)]
des.table
#plot covid_deaths by continent using boxplot
ggplot(num_only, aes(x=Covid.deaths, y=Continent, fill=Continent)) +
```

```
  geom_boxplot(alpha = 0.5, show.legend = FALSE)  +
  coord_flip() + xlab("Number of COVID deaths per 1M people")
############## Task2 ############################
#leaving out Continent, and Covid deaths and convert covid death to factor
set.seed(333)
Project_data$Country.Name <- as.factor(Project_data$Country.Name)
Data.2 <- Project_data[,-c(2,3)]
#Change countries to be rowname and scale data
rownames(Data.2) <- Data.2[,1]
Data.2 <- Data.2[,-1]
Scale_data <- scale(Data.2)
#plot the optimal number of clusters ( k-means )
fviz_nbclust(scale(Data.2), kmeans,method = "wss",k.max = 10) + geom_vline(xintercept = 3, linetype
= 2)
#Define K-means as several value
kmeans2 <- kmeans(Scale_data, centers = 2, nstart = 20)
kmeans3 <- kmeans(Scale_data, centers = 3, nstart = 20)
kmeans4 <- kmeans(Scale_data, centers = 4, nstart = 20)
kmeans3$cluster
#plot the clustering grouping by each no. of k
f1 <- fviz_cluster(kmeans2, geom = "point", data = Scale_data) + ggtitle("k = 2")
f2 <- fviz_cluster(kmeans3,  geom = "point",data = Scale_data) + ggtitle("k = 3")
f3 <- fviz_cluster(kmeans4,  geom = "point",data = Scale_data) + ggtitle("k = 4")
f2
grid.arrange(f1, f2, f3, nrow = 2)
#To get conclusion table of each cluster grouping by continent
con.clus <-tapply(kmeans3$cluster,Continent, table)
con.clus
############## Task3 ############################
#Define cutoff value (mean value)
Cutoff <- mean(Project_data$Covid.deaths,)
```

```
Cutoff
#Assign Down or Up for each covid death
Project_data$lh = "Down"
Project_data$lh[Project_data$Covid.deaths>Cutoff] = "Up"
#check no of member of the up and down group
table(Project_data$lh)
#convert to factor data
Project_data$lh <- as.factor(Project_data$lh)
str(Project_data)
# select only 17 predictor variable
X3 <- Project_data[,-c(1,2,3,21)]
str(X3)
Y3 <- Project_data[,21]
Y3
#============ Feature Selection  =========
# Apply feature selection Forward method
model1 <- glm(Y3~1,family = binomial(link="logit"),data = X3)
summary(model1)
Forward_step <- step(model1,scope =
~Birth.rate+Elect.access+GDP.capita+GDP.growth+Health.exp+Health.exp.capita+Life.expec+Mortalit
y.rate+Net.nat.income+Net.nat.income.capita+Primary+Pop.density+Pop.total+Pop.growth+Unemploy
ment+Water.services+Comp.education, method="forward")
summary(Forward_step)
# Apply feature selection Backward method
model2 <- glm(formula = Project_data$lh~.,family = binomial(link="logit"),data =X3)
Backward_step <- step(model2,method="backward")
summary(Backward_step)
#Prediction the result and create Confusion Table
contrasts(Project_data$lh)
f.glm.probs <- predict(Forward_step ,type="response")
f.glm.predicted <- rep("Down",185)
```

```
f.glm.predicted[f.glm.probs>0.5]="Up"
table(f.glm.predicted, Project_data$lh)
mean(f.glm.predicted==Project_data$lh)
############## Task4 ############################
#prepare indicator of each classification
x <- summary(Project_data$Covid.deaths)
q1 <- x[2] #1st quartile
q2 <- x[3] #2nd quartile
q3 <- x[5] #3rd quartile
q4 <- x[6] # maximun
#Assign each data to class base on indicator
Project_data$QQ4<-
factor(ifelse(Project_data$Covid.deaths<q1,"Q1",ifelse(Project_data$Covid.deaths<q2,"Q2",ifelse(Proje
ct_data$Covid.deaths<q3,"Q3","Q4"))))
Project_data$QQ4
table(Project_data$QQ4)
str(Project_data)
# select only column that will be used
Data_No4 <- Project_data[,-c(1,2,3,21)]
str(Data_No4)
# Separate data for training and evaluation 80% for Train and 20% for Validation
set.seed(3456)
trainIndex <- createDataPartition(Data_No4$QQ4, p = .8, list = FALSE,times = 1)
Train <- Data_No4[trainIndex,]
Valid <- Data_No4[-trainIndex,]
attach(Train)
attach(Valid)
str(Train)
str(Valid)
# Feature selection by Boruta
set.seed(123)
```

```
booruta.fit <-Boruta(Train$QQ4~.,data=Train, doTrace=2)
decision<-booruta.fit$finalDecision
signif <- decision[booruta.fit$finalDecision %in% c("Confirmed")]
print(signif)
plot(booruta.fit, cex.axis=.7, las=2, xlab="", main="Variable Importance")
attStats(booruta.fit)
summary(signif)
#build LDA model with selected features
model_boruta.lda <-  lda(QQ4~ Life.expec  +  Elect.access  +  Mortality.rate  +  Pop.growth  +
Pop.density +Health.exp.capita  +  Health.exp  +  GDP.growth +  GDP.capita  +  Birth.rate  +
Water.services  , data = Train)
model_boruta.lda
# predict lda and create confusion matrix
lda.predicted <- predict(model_boruta.lda,Valid)
lda.table <- table(lda.predicted$class,Valid$QQ4)
m.lda <- mean(lda.predicted$class==Valid$QQ4)
m.lda
#build QDA model with selected features
##### boruta
model_boruta.qda <-  qda(QQ4~ Life.expec  +  Elect.access  +  Mortality.rate  +  Pop.growth  +
Pop.density +Health.exp.capita  +  Health.exp  +  GDP.growth+  GDP.capita  +  Birth.rate  +
Water.services  , data = Train)
model_boruta.qda
# predict qda and create confusion matrix
qda.predicted <- predict(model_boruta.qda,Valid)
qda.table <- table(qda.predicted$class,Valid$QQ4)
m.qda <- mean(qda.predicted$class==Valid$QQ4)
m.qda
#build GLM model with selected features
```

```
glm.test <- multinom(QQ4~ Life.expec  +  Elect.access  +  Mortality.rate  +  Pop.growth  +
Pop.density +Health.exp.capita  +  Health.exp +  GDP.growth  +  GDP.capita  +  Birth.rate  +
Water.services  , family = binomial, data = Train)
summary(glm.test)
# predict glm and create confusion matrix
predict.glm.test <- predict(glm.test,Valid)
glm.table <- table(predict.glm.test,Valid$QQ4)
glm.table
m.glm <- mean(predict.glm.test==Valid$QQ4)
m.glm
#Create table to compare 3 models
conf.tbl <- cbind(lda.table,qda.table,glm.table)
conf.tbl
acc.tbl <- cbind(m.lda,m.qda,m.glm)
acc.tbl <- round(acc.tbl,digits = 2)
acc.tbl
############## Task5 ##############################
# prepare data to visualize the covid death in each country grouping by cluster in task 3
Data.5 =
data.frame(Project_data$Country,kmeans3$cluster,Project_data$Covid.deaths,kmeans3$cluster)
rownames(Data.5) <- c(1:185)
Data.5$kmeans3.cluster = as.factor(Data.5$kmeans3.cluster)

#Plot the covid death in each country grouping by cluster in task 3
Data.5 %>% ggplot(aes(Project_data$Country, Project_data.Covid.deaths, fill = kmeans3$cluster)) +
  geom_col() + facet_wrap(~ kmeans3$cluster, scales = "free_x") +
  theme(axis.text.x = element_text(angle = 90)) +
  ylab("Number of COVID deaths per 1M people") + xlab("") +
guides(fill=guide_legend(title="Cluster"))
```