MA335 Final Project

Relationship between World Development Indicators and causalities of the Covid-19 pandemics.

Nuttarun Kunratvej

Registration number: 2110374

Date: 31 March 2021

Contents

Introduction

The objective report is to investigate the relevance of WDI indicators, collected from the World Bank database and the epidemic of COVID-19 in each country. The data is collected using economic, population, birth rate, health care, service and education of the area as an indicator. Logistic Regression, LDA and QDA are statistical models used in the prediction and analysis part as it is a classification problem that will divide into two ways; binary variable and multiple variables. Finally, we will summarize what we have gained from this data regarding the response for covid-19 in countries with having similar backgrounds.

Problem 1:

	Country	Continent	Covid.deaths	Life.access	Elect.access	Net.nat.income
Min	(Length: 185)	(Length: 185)	3	54.24	6.721	-14.379
Q1	(Class: Character)	(Class: Character)	167	67.92	83.500	2.451
Median	-	-	711	74.23	100.000	3.653
Mean	-	-	1143	73.01	85.731	4.104
Q3	-	-	1830	77.97	100.000	5.090
Max	-	-	6252	85.08	100.000	50.172
	Net.nat.income.capita	Mortality.rate	Primary	Pop.growth	Pop.density	Pop.total
Min	-17.347	1.60	54.73	-1.6095	2.071	3.371e+04
Q1	1.136	5.60	94.46	0.4374	35.893	2.125e+06
Median	2.637	13.05	97.41	1.1647	84.195	9.370e+06
Mean	2.849	19.94	94.61	1.2406	377.319	4.100e+07
Q3	4.081	30.50	98.71	1.9763	217.008	3.037e+07
Max	47.252	82.40	120.45	4.4687	19223.976	1.408e+09
	Health.exp.capita	Health.exp	Umployment	GDP.growth	GDP.capita	Birth.rate
Min	19.85	1.525	0.10	-7.157	228.2	5.90
Q1	103.03	4.495	4.99	1.402	2246.6	10.50
Median	370.11	6.243	5.35	2.653	6731.2	17.57
Mean	1116.28	6.413	6.56	2.905	18261.5	19.39
Q3	1062.39	7.834	6.33	4.745	19701.3	26.79
Max	10921.01	16.767	28.47	19.536	189487.1	45.64
	Water.services	Comp.education				
Min	5.581	5.000				
Q1	73.600	9.000				
Median	73.600	10.000				
Mean	73.600	9.892				
Q3	93.776	11.000				
Max	100.000	17.000				

Table 1: summary of the data set

The WDI indicators dataset provides various variables. This dataset contains 185 rows of refers countries and 20 columns of the country name, continent and indicators. The data was rearranged to be in perfect condition; not available data (NA) was changed into a mean if data does not have outliers and median if there are outliers, these are to avoid the effect in other values in the same column. As the summary table above, the total amount of covid death has large differentiation between 3 and 6,252 which occurs from management and problem-solving

in each country which has different potential to handle the situation. Obviously, Health expenditure per capita, population density and GDP per capita have differed immensely. Although, spending on treatment should be at the same level due to the various diseases having the same in spreading the infection while other values in the dataset have the least difference in the median and mean, this show that the dataset is relatively balanced information.

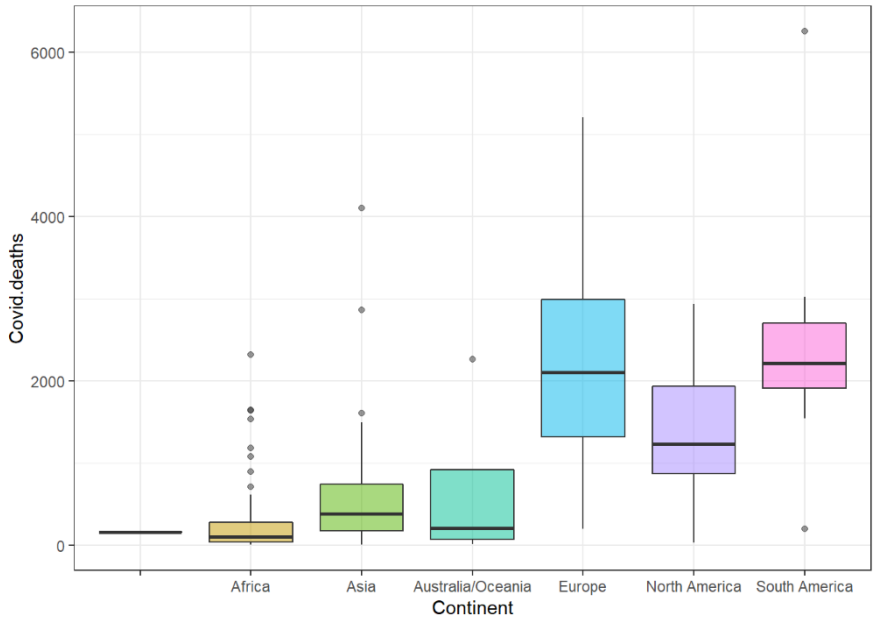

Figure 1: boxplot between continent and amount of death by covid

Boxplot has shown the rate of covid death in 6 continents and the island. Europe has the most balanced distribution of information; there are no outliers or no statistically significant in the data. While in Africa, there is a huge difference in deaths from COVID; some have the least mortality, but some have a high rate. It may be concluded that the sample group is abnormal or different from the authentic group. This may cause the analysis to be inaccurate. Australia/Oceania is expected to have skewed to the left graph, this is because the datasets are very different and mean and medium are not related. Although from the figure, Europe and America seem to have a high rate of covid death; In addition to the data in the tables, it may occur from the management of each country for example lockdown or manage an endemic disease.

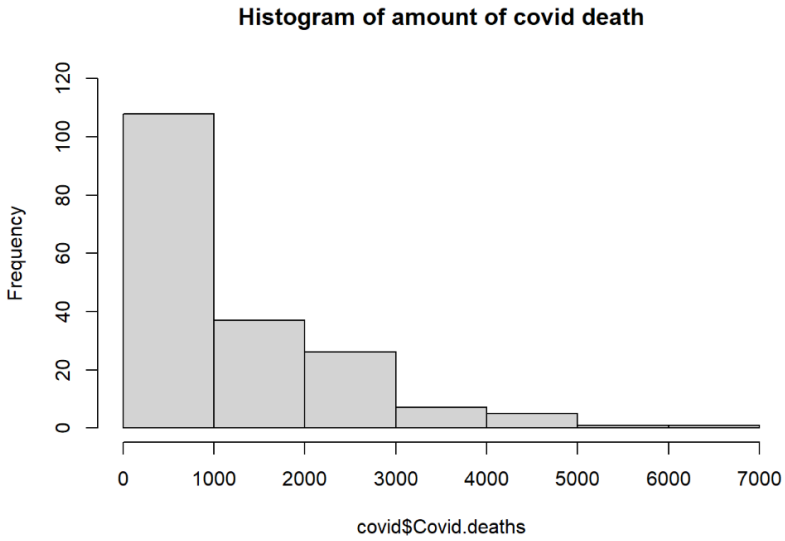

Figure 2: amount of covid death in frequency of countries

The figure showed the amount of death rate from covid-19 and the frequency in each country, apparently, the amount of the first thousand has risen high compared to other levels which this statement will be used in another question. This can be analyzed that death from covid above 1000 is a huge amount and in the range, 5000-7000 barely happens; only 2 cases occur. The death rate of covid over 1000 is almost half of the total numbers. This could be said that the frequency of deaths from COVID is in the range of 0-1000.

Problem 2:

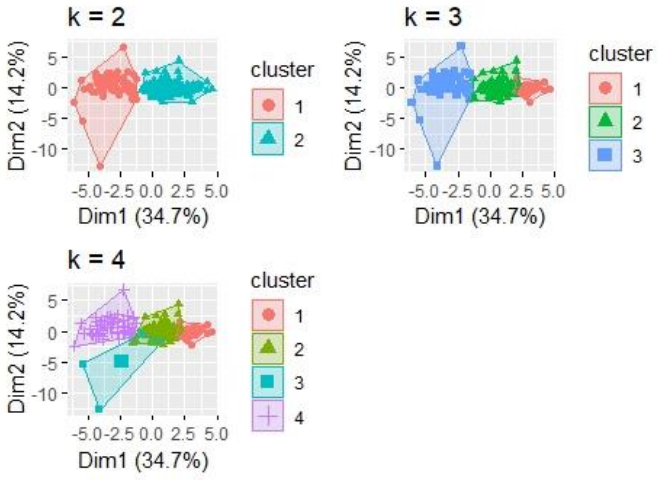

Figure 3: Cluster plot between each country and WDI data

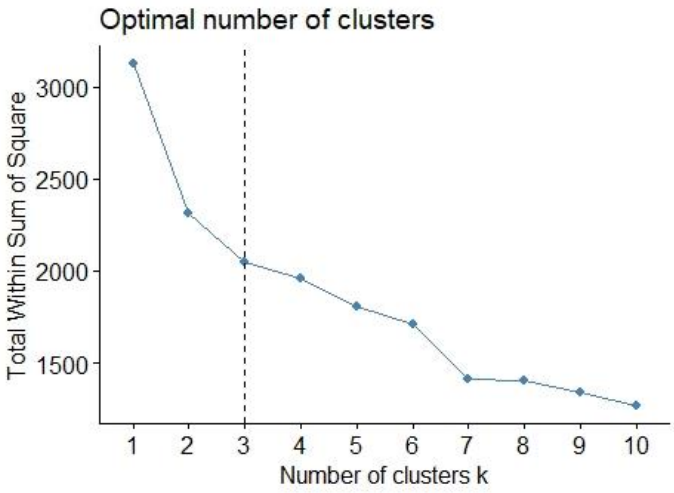

Figure 4: Optimal number of clusters

Clustering data grouping the countries with similar economic profiles by using the k-mean method where each country will belong in each group. K-mean will assign the centroid to be the centre point of each group and the data near each centroid will be attracted to the centroid. Sometimes, the data would be overlap due to the varieties set of the data. The value of k was selected based on the Elbow method which is the optimal number of cluster K is the change of slope which is 3.

	Island	Africa	Asia	Australia	Europe	North America	South America
1	-	-	6	2	21	3	-
2	-	10	31	2	25	20	12
3	1	41	9	1	-	1	-

Table 2: Clustering in each continents

In this case, can be concluded by having a relationship with the continent variable that most of the countries are arranged to be in cluster 2 where there are least countries in cluster 1, these means 3 groups of cluster refer to the similar of the countries profit; this classification may show that each group have similar effects in dealing with the epidemic as well as the overall well-being of the people of the country. Although, there were not completely separated from each others due to the k-means, it is the most suitable value.

Problem 3:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.08593	1.35770	-3.009	0.002617	**
Mortality.rate	-0.08441	0.02403	-3.513	0.000442	***
Unemployment	0.26196	0.07019	3.732	0.000190	***
Comp.education	0.27749	0.09749	2.846	0.004423	**
Pop.growth	-0.63841	0.25067	-2.547	0.010871	*
Health.exp	0.17736	0.09939	1.784	0.074356	.

Table 3: logistic model for covid death group by 1000

This report set the number of covid death over 1000 people to be a large number since in question 1 the histogram shows that 0 to 1000 seem to be the most frequent. Then, the forward method was used in this question to find the fit variables that can predict the response variable; selected prediction variables one by one to test whether it works with the model or not. From the method, mortality rate, unemployment, compulsory education, population growth and current health expenditure were chosen. A coefficient estimate is used to predict the change in log-odds between response variables and predictor variables. In this case, the mortality rate is decreased by 0.08441 will affect the result of the response variables as well as an unemployment rate of 0.26196 will be calculated in the result of the response variables. Z-value can be calculated by dividing the coefficient estimate by standard error.

The important part is the analysis p-value; it tells the ability of prediction variables toward the response variable. As a result, mortality and unemployment rate is closest to zero which means we can reject the null hypothesis with very strong evidence, compulsory has a 0.1% of p-value and can be interpreted that it has strong evidence against the null hypothesis, population growth can be rejected the null hypothesis at 5% significant level and current health expenditure have some evidence, rather weak evidence to rejected the null hypothesis.

Null deviance and residual deviance are used to calculate chi-square statistics to measure how well the model can use; this model can be calculated the p-value correlated with chi-square to be as near as zero so is highly recommended and useful to use this model. The last value to test the efficiency of this model is the Akaike information criterion or AIC, the lowest value can be assumed as the better module that can fit the data and the reason to choose the forward model is it has a lower value compared to the backward model.

glm.predicted.covid	Down	Up
Down	90	18
Up	18	59

Table 4: Prediction table

When inserting the code to predict the high rates of covid death, it can be interpreted that the true negative rate was 0.833 and the sensitivity of 0.766; the prediction was correct 80.54% of the time.

Problem 4:

In this part, the covid death data will be grouped into 4 groups; divided into quartile 1, mean and quartile 3 or first 25 percent, 50 percent, 75 percent and 100 percent. These would make the close amount of data in each group. First, this report will be discussed in the linear discriminant analysis (LDA), from the code can be interpreted that 45.4% is in quantile 1 or 45.4% of covid death data is in this group; the most country has under the approximate amount of 570 people who died by covid. While 15.68%, 21.62% and 17.30% are in group 2, 3 and 4 respectively. Coefficient of linear discriminant use to create equations of LDA model, for example, LD1 would be:

-5.446043e-02*Birth.rate+1.464761e-01*Comp.education+4.251783e-03*Elect.acces-1.680799e-06*GDP.captical-5.974042e-02*GDP.growth+1.156956e-01*Health.exp-7.274599e-05*Health.exp.captical-1.314911e-01*Life.expec-5.865040e-02*Mortality.rate+3.902174e-02*Net.nat.income-1.693599e-02*Net.nat.income-6.049809e-03*Primary+6.091062e-06*Pop.density-1.475380e-09*Pop.total-2.134673e-01*Pop.growth+8.885823e-02*Unemplyment+1.596102e-02*Water.services

The additional, proportion of trace will indicate the percentage of achievement by each linear discriminant function which LD1 that have been written in the formula above is the most successful quotation with 83.62 percent.

In Quadratic Discriminant Analysis (QDA), although the prior probability of the group has almost the same value as LDA, QDA is suitable with a large amount of data and the variance of the classifier is not a major concern which the result of the prediction will be discussed in detail later in this question.

Whilst, logistic regression for 4 groups are different from binary variables since this report wants to compare three models so the full model will be applied in this logistic regression. The result came out to be life expectancy by birth and the mortality rate is closest to zero so the null hypothesis is rejected with very strong evidence. Unemployment and water service turn out to be near zero with a 0.1 significant level so the null hypothesis is rejected. In compulsory education, Electricity access, Health expenditure and Health expenditure per capita have a 1% significant level so they have evidence against the null hypothesis while others do not have evidence to reject the null hypothesis and are not suitable for this response variable. The chi-square score to the p-value of this model is almost zero same as the one above in question 3.

LDA					QDA				
	Truth					Truth			
lda.class (predicted)	0	1	2	3	qda.class (predicted)	0	1	2	3
0	66	3	4	1	0	67	0	1	0
1	7	14	7	1	1	7	26	7	0
2	10	9	22	9	2	8	1	19	0
3	1	3	7	21	3	2	2	13	32

GLM				
	Truth			
glm (predicted)	0	1	2	3
0	59	2	0	0
1	12	1	4	2
2	7	5	6	1
3	6	21	30	29

Table 5: Cross Validation

The prediction in these 3 models can be interpreted as LDA has 66.49% accuracy of all time, QDA has 77.83% while logistic regression has only 51.35% and these can be shown in the table above; LDA has the 66 correct predicted at 0 class and 14, 22, 21 in class 1, 2 and 3 respectively compared to QDA with the 67, 26, 19, 32 in class 0, 1, 2 and 3 respectively and the worst prediction is logistic regression at 4 variables of 59, 1, 6 and 29 correct prediction in class 0, 1, 2 and 3 respectively; refer to the previous section that mentions about the large data is suitable for using QDA to predicted, the result was in accordance with the theory.

Problem 5:

It is expected that this data can be used in predicted the causalities of similar pandemics since this data compile useful information related to the economy, population, health and services which can forecast the quality of life of people in each country. Although, the management approach in the country contributes causes of infections in different countries.

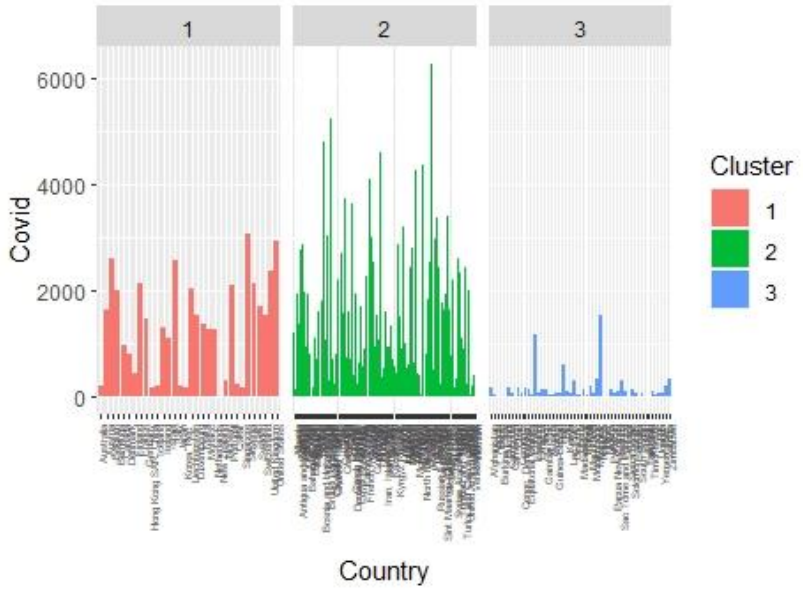

Figure 5: graph between WDI cluster and covid

The figure shows the cluster of countries that have the same economy and clustering 2 seems to have a covid death rate more than others group. If you look deeply into the economic situation, cluster 2 seem to be the group of developing countries which mean the amount of

tourist and international travel will be at a high level, but there is no systematic arrangement and profile as group 1. In addition, group 2 having the largest number of countries; that means the countries with have infected with the covid-19 is quite high. Countries in cluster 3 had the lowest number of covid death and it seem to have the lowest amount of tourist which may be concluded that covid came with the international travel and contact with others countries which is the issue of the economic profile.

Conclusion:

This report can conclude can be a preliminary summary as the cause of infection is likely to come from many factors, such as information provides by WDI, opening the country, the leadership of the country at that time, public health management or cooperation within the country is also important to reduce infection and transmission. It turns out like this because same cluster of WDI indicator still have different rate of covid death and may be less or more than others countries in different cluster and also the cluster have some overlap in the data.

Appendix

```
library(tidyr)

library(tidyverse)

library(tidytext)

library(ggplot2)


covid <- read.csv("project_data.csv", header = TRUE)

#rearrange the data

covid$Comp.education <- as.numeric(gsub(",", "..", covid$Comp.education))

covid$Covid.deaths <- as.numeric(gsub(",", "", covid$Covid.deaths))

#use boxplot to check outliers if yes then use median to replace NA otherwise use means

covid$Covid.deaths[is.na(covid$Covid.deaths)] = mean(covid$Covid.deaths, na.rm=TRUE)

covid$Life.expec[is.na(covid$Life.expec)] = mean(covid$Life.expec, na.rm=TRUE)

covid$Elect.access[is.na(covid$Elect.access)] = median(covid$Elect.access, na.rm=TRUE)

covid$Net.nat.income[is.na(covid$Net.nat.income)] = median(covid$Net.nat.income,
na.rm=TRUE)

covid$Net.nat.income.capita[is.na(covid$Net.nat.income.capita)] =
median(covid$Net.nat.income.capita, na.rm=TRUE)

covid$Mortality.rate[is.na(covid$Mortality.rate)] = median(covid$Mortality.rate, na.rm=TRUE)

covid$Primary[is.na(covid$Primary)] = median(covid$Primary, na.rm=TRUE)

covid$Pop.growth[is.na(covid$Pop.growth)] = median(covid$Pop.growth, na.rm=TRUE)

covid$Pop.density[is.na(covid$Pop.density)] = median(covid$Pop.density, na.rm=TRUE)

covid$Pop.total[is.na(covid$Pop.total)] = median(covid$Pop.total, na.rm=TRUE)
```

```
covid$Health.exp[is.na(covid$Health.exp)] = median(covid$Health.exp, na.rm=TRUE)

covid$Health.exp.capita[is.na(covid$Health.exp.capita)] = median(covid$Health.exp.capita, na.rm=TRUE)

covid$Unemployment[is.na(covid$Unemployment)] = median(covid$Unemployment, na.rm=TRUE)

covid$GDP.growth[is.na(covid$GDP.growth)] = median(covid$GDP.growth, na.rm=TRUE)

covid$GDP.capita[is.na(covid$GDP.capita)] = median(covid$GDP.capita, na.rm=TRUE)

covid$Birth.rate[is.na(covid$Birth.rate)] = mean(covid$Birth.rate, na.rm=TRUE)

covid$Water.services[is.na(covid$Water.services)] = mean(covid$Water.services, na.rm=TRUE)

covid$Comp.education[is.na(covid$Comp.education)] = median(covid$Comp.education, na.rm=TRUE)


covid = covid[-186,]


#1


#make boxplot graph with the data of continents and covid

ggplot(covid, aes(x=Covid.deaths, y=Continent, fill=Continent)) +

  geom_boxplot(alpha = 0.5, show.legend = FALSE) +

  theme_bw() + coord_flip()


#histogram of covid death

hist(covid$Covid.deaths, breaks = 5, ylim = c(0,120), main = "Histogram of amount of covid death")
```

```
#2

library(factoextra)

set.seed(278613)

covid = covid %>% rename(Country = X.ปฟCountry.Name)

#set country column into factor for doing the index
covid$Country <- as.factor(covid$Country)

#pull out the covid and continents column
covid02 <- covid[,c(1,4:20)]

#assign index
rownames(covid02) <- covid02[,"Country"]

#delete country column
covid02 <- covid02 %>%
  select(-Country)
```

```
#standardisation the data

data1 = scale(covid02)


#doing k-means

kmeans2 <- kmeans(data1, centers = 2, nstart = 20)

kmeans3 <- kmeans(data1, centers = 3, nstart = 20)

kmeans4 <- kmeans(data1, centers = 4, nstart = 20)


#plot k-means data

fviz_cluster(kmeans2, data = data1)

fviz_cluster(kmeans3, data = data1)

fviz_cluster(kmeans4, data = data1)

f1 <- fviz_cluster(kmeans2, geom = "point", data = data1) + ggtitle("k = 2")

f2 <- fviz_cluster(kmeans3, geom = "point", data = data1) + ggtitle("k = 3")

f3 <- fviz_cluster(kmeans4, geom = "point", data = data1) + ggtitle("k = 4")


library(gridExtra)

#arrange the graph

grid.arrange(f1, f2, f3, nrow = 2)


#optimal number of clusters

fviz_nbclust(data1, kmeans, method = "wss")+

  geom_vline(xintercept = 3, linetype = 2)
```

```
#extract data between cluster and continents
tapply(kmeans3$cluster, covid$Continent, table)

#3
#set the amount less than 1000 to be down and greater than to be up
covid$Covid.deaths.1000 = "Down"
covid$Covid.deaths.1000[covid$Covid.deaths>1000] = "Up"
#check the up and down group
table(covid$Covid.deaths.1000)
#set class for covid 1000
covid$Covid.deaths.1000 <- as.factor(covid$Covid.deaths.1000)

#pull out the column
X = covid[c(-1:-3,-21,-22)]

#apply forward method
model0 = glm(covid$Covid.deaths.1000~1,data=X,family = binomial(link = "logit"))
step1 = step(model0, scope=~ Birth.rate+Comp.education+Elect.access+GDP.capita+
             GDP.growth+Health.exp+Health.exp.capita+Life.expec+Mortality.rate+
             Net.nat.income+Net.nat.income.capita+Primary+Pop.density+Pop.total+
             Pop.growth+Unemployment+Water.services, method="forward")
summary(step1)
```

```
#test the predicted accuracy

glm.probs.covid <- predict(step1,type="response")

glm.predicted.covid <- rep("Down",185)

glm.predicted.covid[glm.probs.covid>0.5]="Up"


table(glm.predicted.covid, covid$Covid.deaths.1000)


mean(glm.predicted.covid==covid$Covid.deaths.1000)


#4

#set group for covid by quantile and mean (4 variables)

covid$Covid.deaths.grp = ifelse(covid$Covid.deaths <= 570, "0", "3")

covid$Covid.deaths.grp = ifelse(covid$Covid.deaths <= 1140 & covid$Covid.deaths > 570, "1",
covid$Covid.deaths.grp)

covid$Covid.deaths.grp = ifelse(covid$Covid.deaths <= 2280 & covid$Covid.deaths > 1140,
"2", covid$Covid.deaths.grp)


#pull out the column

A = covid[c(-1:-3,-21,-22)]


#set the class to covid group variable

covid$Covid.deaths.grp =  as.factor(covid$Covid.deaths.grp)
```

```
set.seed(1)

#LDA

library(MASS)

lda.fit<-lda(Covid.deaths.grp~Birth.rate+Comp.education+Elect.access+GDP.capita+
        GDP.growth+Health.exp+Health.exp.capita+Life.expec+Mortality.rate+
        Net.nat.income+Net.nat.income.capita+Primary+Pop.density+Pop.total+
        Pop.growth+Unemployment+Water.services,data=covid)
lda.fit

#QDA

qda.fit<-qda(Covid.deaths.grp~Birth.rate+Comp.education+Elect.access+GDP.capita+
        GDP.growth+Health.exp+Health.exp.capita+Life.expec+Mortality.rate+
        Net.nat.income+Net.nat.income.capita+Primary+Pop.density+Pop.total+
        Pop.growth+Unemployment+Water.services,data=covid)
qda.fit

#logistic regression
```

```
model001 =
glm(covid$Covid.deaths.grp~Birth.rate+Comp.education+Elect.access+GDP.capita+
        GDP.growth+Health.exp+Health.exp.capita+Life.expec+Mortality.rate+
        Net.nat.income+Net.nat.income.capita+Primary+Pop.density+Pop.total+
        Pop.growth+Unemployment+Water.services,data=A,family = binomial(link = "logit"))
summary(model001)


#CV test


install.packages("caret")
library(caret)



trControl <- trainControl(method = "cv", number = 5)
lda.fit01 <- train(Covid.deaths.grp~Birth.rate+Comp.education+Elect.access+GDP.capita+
          GDP.growth+Health.exp+Health.exp.capita+Life.expec+Mortality.rate+
          Net.nat.income+Net.nat.income.capita+Primary+Pop.density+Pop.total+
          Pop.growth+Unemployment+Water.services,
        method = "lda",
        trControl = trControl,
        metric = "Accuracy",
        data = covid)
lda.pred01 <- predict(lda.fit01,covid)
```

```
table(lda.pred01, covid$Covid.deaths.grp)

mean(lda.pred01==covid$Covid.deaths.grp)



trControl <- trainControl(method = "cv", number = 5)

qda.fit01 <- train(Covid.deaths.grp~Birth.rate+Comp.education+Elect.access+GDP.capita+

                     GDP.growth+Health.exp+Health.exp.capita+Life.expec+Mortality.rate+

                     Net.nat.income+Net.nat.income.capita+Primary+Pop.density+Pop.total+

                     Pop.growth+Unemployment+Water.services,

                   method = "qda",

                   trControl = trControl,

                   metric = "Accuracy",

                   data = covid)

qda.pred01 <- predict(qda.fit01,covid)

table(qda.pred01, covid$Covid.deaths.grp)

mean(qda.pred01==covid$Covid.deaths.grp)


glm.probs <- predict(model001,type="response")

glm.predicted <- rep("0",185)

glm.predicted[glm.probs>0.25]="1"

glm.predicted[glm.probs>0.50]="2"

glm.predicted[glm.probs>0.75]="3"
```

```
table(glm.predicted, covid$Covid.deaths.grp)

mean(glm.predicted==covid$Covid.deaths.grp)


#5

#make a data frame to use in the graph

AA = kmeans3$cluster

q5 = data.frame(covid$Country,kmeans3$cluster,covid$Covid.deaths,AA)


#index the data

rownames(q5) <- c(1:185)


q5$kmeans3.cluster = as.factor(q5$kmeans3.cluster)


library(ggplot2)

library(tidyr)


#plot the graph

q5 %>% ggplot(aes(covid.Country, covid.Covid.deaths, fill = kmeans3.cluster)) +

  geom_col() + facet_wrap(~ AA, scales = "free_x") +

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1,size = rel(0.5))) +

  ylab("Covid") + xlab("Country") + guides(fill=guide_legend(title="Cluster"))
```