

Modelling experimental and observational data

FINAL PROJECT REPORT



March 31, 2022

Bushra Hashmi (2110387)

UNIVERSITY OF ESSEX

qw2ws2w

# Abstract:

*This project is part of the course Modelling and Experimental data-MA335. We are trying to find out the relationship between the covid deaths and WDI(World Development Indicators). Covid deaths are considered as the response variable in this report. Conclusion this report complies with the descriptive statistic, clustering algorithm, logistic regression model, QDA, LDA, and logistic regression conclusively; all the analyses show mixed results.*

Table of Contents

[Abstract: 1](#_Toc99595371)

[1. Introduction: 2](#_Toc99595372)

[1a. Preliminary Analysis: 2](#_Toc99595373)

[2. Question 1: 2](#_Toc99595374)

[3. Question 2: 3](#_Toc99595375)

[3a. Euclidean distance 3](#_Toc99595376)

[3b. K-Mean Algorithm: 4](#_Toc99595377)

[3c. Elbow Method: 4](#_Toc99595378)

[3d. Clustering in relation to the Continent variable: 5](#_Toc99595379)

[4. Question:3: 5](#_Toc99595380)

[4a. Logistic Regression Model: 5](#_Toc99595381)

[5. Question 4: 6](#_Toc99595382)

[5a. Logistic Regression: 7](#_Toc99595383)

[5b. LDA and QDA: 7](#_Toc99595384)

[6. Question 5: 8](#_Toc99595385)

[7. Conclusion: 8](#_Toc99595386)

[8. References: 8](#_Toc99595387)

[9. Appendices: 9](#_Toc99595388)

Count of words:2000(excluding Appendix)

# 1. Introduction:

The covid-19 pandemic has changed the world dynamics over the last two years. Worlds indicators static’s have changed due to the number of covid deaths every day. This report uses the provided data of project\_data.csv, which includes the data of countries of the world with selective indicators of World Development Indicators (WDI); the data of covid deaths are considered the response variable in this report. This report contains the descriptive statistic, clustering algorithm, logistic regression model, QDA, LDA and logistic regression for multiclass classification problems.

## 1a. Preliminary Analysis:

This project uses different libraries to exploit data like pastecs, ggstatsplot, dplyr, ggplot2, MASS, ISLR, dplyr, factoextra, summarytools, gridExtra, caTools, caret and nnet. Firstly, we cleaned our data set by handling the missing values. We replace the missing values of x with the mean of the observed values of x. Moreover, we change the category of comp. education and Covid. deaths into the numeric category to further proceed with descriptive analysis.

# 2. Question 1:

For the numerical representation of the data, this report performs the descriptive analysis of the data. There is more than one way to get a descriptive data analysis; this report uses the stat. desc(), using this code, we get the standard deviation, mean, median, range, quartile value, and the complete evaluation *(Table 01)* below.

*Table1: Summary of Descriptive Data Analysis*

A close-up of a document

Description automatically generated with medium confidence

From the summary, we can evaluate that Nbr. Null value is 0.000, which explains that there is no missing value in the data set. Continents and country names are in a character data type. The mean is the average of a set of given numbers, and the standard deviation explains how to scatter the data corresponding with the mean.

For the graphical representation of the data set, a box plot is used for explanation. We drive the relationship between one of the WDI indicators concerning covid deaths in the continents; the below graph *(Figure 1)* represents the graphical representation of seven continents of the world, covid deaths in on the x-axis, and population growth is on the y-axis. The data represent the mixed result of covid deaths in different continents with respect to population growth for a specific time.

From the graph *(Figure 1),* we can conclude that in Asia, the ratio of the number of covid deaths in contrast to population growth is the same for the specific period, but in Australia, the number of covid deaths was more as compared to population growth in a specified year.

Chart, box and whisker chart

Description automatically generated  
*Figure 1: Boxplot between the covid deaths and population growth in terms of Continent*

# 3. Question 2:

## 3a. Euclidean distance

Clustering of observations is used to compute the similarity or dissimilarity between each pair of observations. The clustering method can be affected by the presence of outliers’ responses to the question of the clustering algorithm. First, we need to find out the distance between the observations. For that, this report performs the Euclidean distance method by using the get\_dist and fviz\_dist in R with the help of package “factoextra”.

From *Figure 2*, we can observe that Light blue or turquoise colour represents a small distance; on the other hand, orange colour indicates the big distance between observations. Smaller distance shows greater similarity, and larger distance indicates more dissimilarity.

A picture containing text

Description automatically generated  
*Figure 2: Clustering Distance Measure*

## 3b. K-Mean Algorithm:

Before applying the clustering algorithms to the data set, we need to standardise our data set so that all the variables will be on the same scale; the k-mean algorithm is used to minimise the cluster variation. It splits the data set into K no of clusters, it rehashes the cycle until it doesn’t track down the best cluster, and we try different values of k and then decide the best at the end. The fviz\_cluster function visualises results to see a clear picture of clusters.

We try the value at 2,3 and 4, and at K-means clustering with 4 clusters result, we find out that the value of between\_SS / total\_SS is low at 41.0%. This ratio accounts for the total sum of squares of the data points between the clusters (*Reference, Appendices*)

## 3c. Elbow Method:

The elbow method looks at the percentage of variance explained as a function of the number of clusters and chose a number of clusters. Adding another cluster doesn’t give much better modelling of the data.

Chart, line chart

Description automatically generated  
*Figure 3. Optimal Number of Cluster*

From the above Graph *(Figure 3),* we can observe that k=4, the between\_ss/total\_ss ratio tends to change slowly and remain less changing compared to other k’s. So for this data, k=4 should be a good choice for the number of clusters

## 3d. Clustering in relation to the Continent variable:

This report uses the bind function to connect the continent variable again with the clustered data set. From *(Table 2)* we can observe that each continent falls into different 4 clusters from the results. But only Asian and African countries are part of the 1cluster, Europe only falls in cluster 2 or 4, and South America doesn’t fall in any cluster except 4th. On the other hand, in North America, only one country falls in the 3rd cluster.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster | Continent | | | | | |
| Africa | Asia | Australia/Ocenia | Europe | North America | South America |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 6 | 2 | 21 | 3 | 0 |
| 3 | 41 | 7 | 2 | 0 | 1 | 0 |
| 4 | 9 | 32 | 2 | 25 | 20 | 12 |

*Table 2. Clustering in Relation to Continental Variable*

# 4. Question:3:

To answer this question, we change the variable covid. deaths into binary form, The first thing to do is install and load the ISLR package,

## 4a. Logistic Regression Model:

Logistic regression is a method for fitting a regression curve, y = f(x) when y is a categorical variable. The typical use of this model is predicting y given a set of predictors x. The predictors can be continuous, categorical or a mix of both. In this question, we first converted our y, i.e. covid. deaths into a binary variable from numeric that can assume either the value 1 or 0.

We split the data into two chunks: training and testing set. The training set will fit our model, which we will be testing over the testing set. We specify the parameter *family=binomial* in the *glm()* function.

*Table 3. Summary of Logistic Regression Model*

Table

Description automatically generated

From the above summary (*Table 3*), we get the values of the estimate, standard errors, z-score, and p-values on each coefficient. None of the coefficients is significant here. All the p-value show insignificant probabilities. It also gives you the null and residual deviance (the deviance for the model with all the predictors). There’s a minimal difference between the deviance and degrees of freedom. The AIC value is 102.39

For logistic regression prediction analysis, we take glm. Probability > 0.5, and glm predicted calls “1”; otherwise, it calls “0” the predicted mean value, identifying 64.6% chances of high or increasing COVID causalities.

# 5. Question 4:

To solve this question, we first divide the data into four categories. We categorise them as Coviddeaths < 250 tends to 0, then Covid.deaths < 1000 = 1 than < 2000=2 and above 2000 is 3.

## 5a. Logistic Regression:

To carry out this model, we utilise the MASS bundle. Before carrying out this calculation, we need to eliminate exceptions from the dataset and afterwards perform normalisation; Multinomial logistic regression is used to model nominal outcome variables. The log odds of the outcomes are modelled as a linear combination of the predictor variables. nnet package required to estimate a multinomial logistic regression model. This model-running output includes some iteration history, and the final negative log-likelihood is 21.366755. The multinorm mean value represents 42.73351, and the multinorm predicted mean value represents 0.8478261.

5b. LDA and QDA:

Linear Discriminant Analysis used when a linear boundary is required between classifiers, and Quadratic Discriminant Analysis is used to find a non-linear boundary between classifiers. LDA and QDA work better when the response classes are separable, and the distribution of X=x for all classes is normal. **QDA performed worse than LDA since it ﬁt a more ﬂexible classifier than the necessary,** same condition we experienced in our model.

From *(Figure 5),* we can analyse the value of group 0 and group 1. The probability is increasing in group 1. It represents covid deaths will increase over time.LDA predicted mean value is 85.8%. On the other hand, the QDA mean value represents 88.5%

Chart, histogram

Description automatically generated  
*Figure 5. Linear Discriminant Analysis*

Caret library is used for confusion matrices of LDA and QDA showing very high accuracies; the LDA shows an accuracy of 0.8486, i.e. 84.8%, QDA has demonstrated the accuracy of 0.8162, i.e. 81.6%, and logistic regression has a mean of 84.7%, representing that model is completely fitted with the data set and the model is accurate. The best-fitted model for our project is LDA because of the accuracy of the predicted value and actual value. The prediction table is included in the appendix part of this report.

# 6. Question 5:

According to WHO, results from the latest COVID-19 deaths reported now exceed 3.3 million; based on the excess mortality estimates produced for 2020, we are likely facing a significant undercount of total deaths directly and indirectly attributed to COVID-19.

Many countries still lack functioning civil registration and vital statistics systems with the capacity to provide accurate, complete and timely data on births, deaths and causes of death. A recent assessment of health information systems capacity in 133 countries found that the percentage of registered deaths ranged from 98% in the European region to only 10% in the African region, which is very difficult to believe that, but the point of consideration is that in underdeveloped counties covid 19 is not the only pandemic living there, malaria, dengue, ebola are significantly present.

Countries also use different processes to test and report COVID-19 deaths, making comparisons difficult. We learned from this data set that there is no comparison between the COVID-19 pandemic of countries with similar economic profiles. In Japan, the covid death was 194; in France, the rate was 2,121, although both are developed countries of the world and included in the G7 (group of seven countries), although both the countries have the same life expectancy rate. In terms of health expenditure per capita associated with the government is almost the same for both countries. Japan has 4360.4, And France has 4491.7. there is some similarity observed in covid deaths concerning continents, but this similarity is not enough to judge the results as this report already compiles the effect in question 2.

# 7. Conclusion:

In the end, the linear model results of this dataset represent the model is not best suited as all the variables are insignificant to each other and have a small R2 value of 0.2358. for controlling the covid deaths, more market research and in-depth study are required, as most of the WDI indicators are not dependent on or impact this research’s response variable.

# 8. References:

* James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert. (2013) An introduction to statistical learning: with applications in R, New York: Springer. vol. Springer texts in statistics, Chapter 4 (Sections 4.1-4.3)
* https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/
* <https://www.datacamp.com/community/tutorials/logistic-regression-R>
* <https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess->
* https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/mortality

# 9. Appendices:

**Text

Description automatically generated**

Chart, scatter chart

Description automatically generated

Text

Description automatically generated

Quadratic Discriminant Analysis

185 samples

17 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 185, 185, 185, 185, 185, 185, ...

Resampling results:

Accuracy Kappa

0.8104437 0.5846137

> lda\_matrix

Cross-Validated (5 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

Reference

Prediction 0 1

0 23.2 3.8

1 11.4 61.6

Accuracy (average) : 0.8486

> qda\_matrix

Cross-Validated (5 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

Reference

Prediction 0 1

0 27.0 10.8

1 7.6 54.6

Accuracy (average) : 0.8162

#=============R Coading#=============

#=======Libraries=========#

library(pastecs)

library(ggstatsplot)

library(dplyr)

library(ggplot2)

library(MASS)

library(ISLR)

library(dplyr)

library(factoextra)

library(summarytools)

library(gridExtra)

library(caTools)

library(caret)

library(nnet)

#=======Preliminary Analysis==========

FP<-read.csv("project\_data.csv",header=TRUE)

str(FP)

#changing data format

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

FP$Elect.access[is.na(FP$Elect.access)] = mean(FP$Elect.access, na.rm=TRUE)

FP$Net.nat.income[is.na(FP$Net.nat.income)] = mean(FP$Net.nat.income, na.rm=TRUE)

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

FP$Mortality.rate[is.na(FP$Mortality.rate)] = mean(FP$Mortality.rate, na.rm=TRUE)

FP$Primary[is.na(FP$Primary)] = mean(FP$Primary, na.rm=TRUE)

FP$Pop.growth[is.na(FP$Pop.growth)] = mean(FP$Pop.growth, na.rm=TRUE)

FP$Pop.density[is.na(FP$Pop.density)] = mean(FP$Pop.density, na.rm=TRUE)

FP$Pop.total[is.na(FP$Pop.total)] = mean(FP$Pop.total, na.rm=TRUE)

FP$Health.exp.capita[is.na(FP$Health.exp.capita)] = mean(FP$Health.exp.capita, na.rm=TRUE)

FP$Unemployment[is.na(FP$Unemployment)] = mean(FP$Unemployment, na.rm=TRUE)

FP$GDP.growth[is.na(FP$GDP.growth)] = mean(FP$GDP.growth, na.rm=TRUE)

FP$Health.exp[is.na(FP$Health.exp)] = mean(FP$Health.exp, na.rm=TRUE)

FP$GDP.capita[is.na(FP$GDP.capita)] = mean(FP$GDP.capita, na.rm=TRUE)

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

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

#change comp.education into numeric

FP$Comp.education<-as.numeric(FP$Comp.education)

FP$Comp.education[is.na(FP$Comp.education)] = mean(FP$Comp.education, na.rm=TRUE)

typeof(FP$Comp.education)

#Change Covid deaths into numeric

FP$Covid.deaths<-as.numeric(FP$Covid.deaths)

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

str(FP)

#descriptive summary

stat.desc(FP)

FP$Continent[FP$ï..Country.Name=="Solomon Islands"]<-"Australia/Oceania" #assigning continent name which is empty previosuly

summary(FP)

Summary\_FP<-descr(FP, headings = FALSE, stats = "common") # most common descriptive statistics

#========Graphical Analysis of Data==========

FP%>%ggplot(aes(`Covid.deaths`, `Pop.growth`, colour=`Continent`)) +geom\_boxplot()

#+================Clustering=============

#removing variable covid deaths and continent

FP2 <- FP %>%

select(-Continent, -Covid.deaths,-ï..Country.Name)

FP4 <- FP %>%

select(-Continent,-ï..Country.Name)

par(mfrow=c(1,8))

for(i in 1:8) {

hist(FP4[,i], main=names(FP4)[i])}

colnames(FP2)

distance.Euclidean<-get\_dist(FP2)

fviz\_dist(distance.Euclidean,gradient=list(low="#00AFBB",mid="white",high="#FC4E07"))

distance.corr<-get\_dist(FP2,stand=TRUE,method="pearson")

fviz\_dist(distance.corr,gradient=list(low="#00AFBB",mid="white",high="#FC4E07"))

FP3<-scale(FP2)

#=====K-means clustering, because we need a data into k cluster foam

set.seed(123)

kmeans2<-kmeans(FP3,centers=2,nstart=20)

kmeans3<-kmeans(FP3,centers=3,nstart=20)

kmeans4<-kmeans(FP3,centers=4,nstart=20)

kmeans4

f1 <- fviz\_cluster(kmeans2, geom = "point", data = FP3) + ggtitle("k = 2")

f2 <- fviz\_cluster(kmeans3, geom = "point", data = FP3) + ggtitle("k = 3")

f3 <- fviz\_cluster(kmeans4, geom = "point", data = FP3) + ggtitle("k = 4")

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

#Determining the optimal number of clusters

fviz\_nbclust(FP3, kmeans, method = "wss")+

geom\_vline(xintercept = 4, linetype = 2)

#binding continent again

B<-cbind(FP$Continent,FP2)

cluster=as.integer(kmeans4$cluster)

H<-cbind(B,cluster)

table(kmeans4$cluster,Continent)

#=========Logistic regression=========

#=====Transform Covid deaths into a binary variable.

set.seed(4500)

GLM\_copy = subset(FP, select = -c(Continent, ï..Country.Name) )

GLM\_copy$Covid.deaths<-ifelse(GLM\_copy$Covid.deaths>mean(GLM\_copy$Covid.deaths), 1, 0) #converting covid.deaths into binary variable

GLM\_copy$Covid.deaths<- as.factor(GLM\_copy$Covid.deaths)

GLM\_copy

#=========split into training and tested data=====

table(GLM\_copy$Covid.deaths)

train <-((FP$GDP.growth)<3)

train

GLM\_train<-GLM\_copy[!train, ]

typeof(GLM\_train)

GLM\_test<-GLM\_copy[!train,]

#========Applting GLM Function

glm.fits<-glm(Covid.deaths~., data = GLM\_copy, family = binomial, subset = train)

summary(glm.fits)

#GLM Prediction and mean

glm.probs<-predict(glm.fits, GLM\_test, type="response")

glm.pred<-rep(0,82)

glm.pred[glm.probs>0.5]=1

table(glm.pred, GLM\_test)

mean(glm.pred==GLM\_test$Covid.deaths)

#===============LDA and QDA===========#

# =====divide into 4 possible labels

set.seed(4500)

LR\_FP <- FP

LR\_FP <- subset(LR\_FP, select = -c(Continent,ï..Country.Name) )

LR\_FP$Covid.deaths[is.na(LR\_FP$Covid.deaths)] = mean(LR\_FP$Covid.deaths, na.rm=TRUE)

LR\_FP$Covid.deaths <- ifelse((LR\_FP$Covid.deaths < 250) , 0 ,

ifelse((LR\_FP$Covid.deaths < 1000), 1,

ifelse((LR\_FP$Covid.deaths < 2000), 2, 3)))

#==Logistic Regression

LR\_FP$Covid.deaths=as.factor(LR\_FP$Covid.deaths)

LR\_FP\_train <-((LR\_FP$GDP.growth)<3)

LR\_FP\_train2<-LR\_FP[LR\_FP\_train,]

LR\_FP\_test\_data<-LR\_FP[-LR\_FP\_train,]

Multinorm\_FP <- multinom(Covid.deaths~.,data = LR\_FP\_train2)

summary(Multinorm\_FP)

#====Logistic Regression Prediction

multinorm\_Predict <- predict(Multinorm\_FP, LR\_FP\_test\_data)

multinorm\_Predict

mean(multinorm\_Predict==LR\_FP\_test\_data$Covid.deaths)

#=======LDA

set.seed(5500)

lda\_fit <- lda(Covid.deaths~.,data=LR\_FP\_train2)

lda\_fit

plot(lda\_fit)

lda\_predicted=predict(lda\_fit, LR\_FP\_test\_data)

lda\_predicted

mean(lda\_predicted$class==LR\_FP\_test\_data$Covid.deaths)

#======QDA

set.seed(5000)

qda\_fit <- train(Covid.deaths~.,data=LR\_FP, method="qda")

qda\_fit

plot.train(qda\_fit)

qda\_predicted=predict(qda\_fit, LR\_FP\_test\_data)

qda\_predicted

mean(qda\_predicted==LR\_FP\_test\_data$Covid.deaths)

#=========cross validation

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

#confusion Matrix of LDA

set.seed(7000)

lda\_control<-train(Covid.deaths~., method="lda", trControl=trControl,metric="Accuracy", data=LR\_FP)

lda\_control\_predict<-predict(lda\_control,LR\_FP)

lda\_matrix<-confusionMatrix(lda\_control)

lda\_matrix

#=========confusion Matrix of QDA

set.seed(8000)

qda\_control<-train(Covid.deaths~., method="qda", trControl=trControl,metric="Accuracy", data=LR\_FP)

qda\_control\_predict<-predict(qda\_control,LR\_FP)

qda\_matrix<-confusionMatrix(qda\_control)

qda\_matrix

#=============End.============#