Covid19 Analysis

Numrah Alauddin

12/10/2021

#Library Defined  
  
knitr::opts\_chunk$set(error = TRUE)  
library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

library(factoextra)

## Warning: package 'factoextra' was built under R version 4.1.2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:Hmisc':  
##   
## src, summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

Comment: library(Hmisc): it contains many functions useful for data analysis, high-level graphics,utility operations, functions for computing sample size library(factoextra): is used to extract and visualize the output of exploratory multivariate data analysis. library(dplyr):

#File that have been called for the data  
  
data <- read.csv("C:/Users/ARSALAN IQBAL/Downloads/COVID19\_line\_list\_data.csv")  
data <- select(data,-c(X.1:X.6))  
data <- select(data,-c(X))  
data\_copied <- data  
describe(data)

## data   
##   
## 20 Variables 1085 Observations  
## --------------------------------------------------------------------------------  
## ï..id   
## n missing distinct Info Mean Gmd .05 .10   
## 1085 0 1085 1 543 362 55.2 109.4   
## .25 .50 .75 .90 .95   
## 272.0 543.0 814.0 976.6 1030.8   
##   
## lowest : 1 2 3 4 5, highest: 1081 1082 1083 1084 1085  
## --------------------------------------------------------------------------------  
## case\_in\_country   
## n missing distinct Info Mean Gmd .05 .10   
## 888 197 197 1 48.84 54.99 2.00 4.00   
## .25 .50 .75 .90 .95   
## 11.00 28.00 67.25 110.30 153.65   
##   
## lowest : 1 2 3 4 5, highest: 365 443 875 925 1443  
##   
## Value 0 20 40 60 80 100 120 140 160 180 200  
## Frequency 215 241 137 81 84 40 22 19 22 19 1  
## Proportion 0.242 0.271 0.154 0.091 0.095 0.045 0.025 0.021 0.025 0.021 0.001  
##   
## Value 280 300 360 440 880 920 1440  
## Frequency 1 1 1 1 1 1 1  
## Proportion 0.001 0.001 0.001 0.001 0.001 0.001 0.001  
##   
## For the frequency table, variable is rounded to the nearest 20  
## --------------------------------------------------------------------------------  
## reporting.date   
## n missing distinct   
## 1084 1 43   
##   
## lowest : 02/01/20 02/02/20 02/03/20 02/04/20 02/05/20   
## highest: 2/24/2020 2/25/2020 2/26/2020 2/27/2020 2/28/2020  
## --------------------------------------------------------------------------------  
## summary   
## n missing distinct   
## 1080 5 967   
##   
## lowest : confirmed COVID-19 pneumonia patient No.11 in Tianjin: female, 55, symptom onset on 01/23/2020, hospitalized on 01/23/2020, confirmed on 01/26/2020 confirmed COVID-19 pneumonia patient No.12 in Tianjin: female, 79, symptom onset on 01/24/2020, hospitalized on 01/24/2020, confirmed on 01/26/2020 confirmed COVID-19 pneumonia patient No.13 in Tianjin: female, 19, symptom onset on 01/19/2020, hospitalized on 01/20/2020, confirmed on 01/26/2020 confirmed COVID-19 pneumonia patient No.14 in Tianjin: male, 71, Wuhan resident, visited Malaysia from 01/19/2020 to 01/25/2020, arrived in Tianjin on 01/25/2020, symptom onset on 01/25/2020, hospitalized on 01/25/2020, confirmed on 01/26/2020 confirmed imported COVID-19 pneumonia patient in Gansu: female, 20, lives in Wuhan, arrived in Gansu on 01/18/2020, symptom onset on 01/19/2020, visit clinic on 01/24/2020, hospitalized on 01/24/2020.   
## highest: new recovered imported COVID-19 pneumonia patient in Beijing: female, returned to Beijing from Wuhan on 01/08/2020, symptom onset afterwards, recovered on 01/24/2020. new recovered imported COVID-19 pneumonia patient in Beijing: male, returned to Beijing from Wuhan on 01/08/2020, symptom onset afterwards, recovered on 01/25/2020. Second confirmed imported COVID-19 pneumonia patient in Guangxi: male, 46, in contact with individuals from Wuhan before symptom onset. symptom onset on 01/20/2020. Second confirmed imported COVID-19 pneumonia patient in Liaoning: male, 40, works in Wuhan, visit Fushun, Liaoning on 01/12/2020, symptom onset on 01/14/2020, visit clinic in Fushun Dalian on 01/19/2020. Second confirmed imported COVID-19 pneumonia patient in Sichuan: male, 57, Wuhan resident, visited Sichuan on 01/15/2020, symptom onset on 01/16/2020 and hospitalized.   
## --------------------------------------------------------------------------------  
## location   
## n missing distinct   
## 1085 0 156   
##   
## lowest : Afghanistan Aichi Prefecture Alappuzha Algeria Amiens   
## highest: Yunnan Zabaikalsky Zaragoza Zhejiang Zhuhai   
## --------------------------------------------------------------------------------  
## country   
## n missing distinct   
## 1085 0 38   
##   
## lowest : Afghanistan Algeria Australia Austria Bahrain   
## highest: Thailand UAE UK USA Vietnam   
## --------------------------------------------------------------------------------  
## gender   
## n missing distinct   
## 902 183 2   
##   
## Value female male  
## Frequency 382 520  
## Proportion 0.424 0.576  
## --------------------------------------------------------------------------------  
## age   
## n missing distinct Info Mean Gmd .05 .10   
## 843 242 85 0.999 49.48 20.79 22.0 25.0   
## .25 .50 .75 .90 .95   
## 35.0 51.0 64.0 75.0 78.9   
##   
## lowest : 0.25 0.50 1.00 2.00 4.00, highest: 86.00 87.00 89.00 91.00 96.00  
## --------------------------------------------------------------------------------  
## symptom\_onset   
## n missing distinct   
## 563 522 70   
##   
## lowest : 01/02/20 01/03/20 01/04/20 01/05/20 01/06/20   
## highest: 2/22/2020 2/23/2020 2/24/2020 2/25/2020 2/26/2020  
## --------------------------------------------------------------------------------  
## If\_onset\_approximated   
## n missing distinct Info Sum Mean Gmd   
## 560 525 2 0.123 24 0.04286 0.08219   
##   
## --------------------------------------------------------------------------------  
## hosp\_visit\_date   
## n missing distinct   
## 507 578 60   
##   
## lowest : 01/01/20 01/03/20 01/05/20 01/06/20 01/08/20   
## highest: 2/24/2020 2/25/2020 2/26/2020 2/27/2020 2/28/2020  
## --------------------------------------------------------------------------------  
## exposure\_start   
## n missing distinct   
## 128 957 39   
##   
## lowest : 01/03/20 01/06/20 01/08/20 01/09/20 01/10/20   
## highest: 2/15/2020 2/17/2020 2/19/2020 2/20/2020 2/21/2020  
## --------------------------------------------------------------------------------  
## exposure\_end   
## n missing distinct   
## 341 744 52   
##   
## lowest : 01/02/20 01/03/20 01/04/20 01/05/20 01/06/20   
## highest: 2/21/2020 2/22/2020 2/23/2020 2/24/2020 2/25/2020  
## --------------------------------------------------------------------------------  
## visiting.Wuhan   
## n missing distinct Info Sum Mean Gmd   
## 1085 0 2 0.437 192 0.177 0.2916   
##   
## --------------------------------------------------------------------------------  
## from.Wuhan   
## n missing distinct Info Sum Mean Gmd   
## 1081 4 2 0.37 156 0.1443 0.2472   
##   
## --------------------------------------------------------------------------------  
## death   
## n missing distinct   
## 1085 0 14   
##   
## lowest : 0 02/01/20 1 2/13/2020 2/14/2020  
## highest: 2/24/2020 2/25/2020 2/26/2020 2/27/2020 2/28/2020  
##   
## 0 (1022, 0.942), 02/01/20 (1, 0.001), 1 (42, 0.039), 2/13/2020 (1, 0.001),  
## 2/14/2020 (1, 0.001), 2/19/2020 (2, 0.002), 2/21/2020 (2, 0.002), 2/22/2020 (1,  
## 0.001), 2/23/2020 (4, 0.004), 2/24/2020 (1, 0.001), 2/25/2020 (2, 0.002),  
## 2/26/2020 (3, 0.003), 2/27/2020 (2, 0.002), 2/28/2020 (1, 0.001)  
## --------------------------------------------------------------------------------  
## recovered   
## n missing distinct   
## 1085 0 32   
##   
## lowest : 0 02/02/20 02/04/20 02/05/20 02/06/20   
## highest: 2/24/2020 2/25/2020 2/26/2020 2/27/2020 2/28/2020  
## --------------------------------------------------------------------------------  
## symptom   
## n missing distinct   
## 270 815 108   
##   
## lowest : chest discomfort chills cold, fever, pneumonia cough cough with sputum   
## highest: throat pain, chills throat pain, fever tired vomiting, cough, fever, sore throat vomiting, diarrhea, fever, cough   
## --------------------------------------------------------------------------------  
## source   
## n missing distinct   
## 1085 0 85   
##   
## lowest : å¤®è§†æ–°é—» ABC ABC News æ–°æµª Al Arabiya   
## highest: Wa.de Washington Examiner Xin Hua Net Yahoo News Yonnhap News Agency  
## --------------------------------------------------------------------------------  
## link   
## n missing distinct   
## 1085 0 490   
##   
## lowest : http://behdasht.gov.ir/news/%DA%A9%D8%B1%D9%88%D9%86%D8%A7+%D9%88%DB%8C%D8%B1%D9%88%D8%B3/199807/%D8%AF%D8%B1+%D8%B1%D9%88%D8%B2%D9%87%D8%A7%DB%8C+%DA%AF%D8%B0%D8%B4%D8%AA%D9%87+735+%D8%A8%DB%8C%D9%85%D8%A7%D8%B1+%D8%A8%D8%A7+%D8%B9%D9%84%D8%A7%D8%A6%D9%85+%D8%B4%D8%A8%D9%87+%D8%A2%D9%86%D9%81%D9%84%D9%88%D8%A2%D9%86%D8%B2%D8%A7+%D8%AF%D8%B1+%DA%A9%D8%B4%D9%88%D8%B1+%D8%A8%D8%B3%D8%AA%D8%B1%DB%8C+%D8%B4%D8%AF%D9%86%D8%AF+%D8%A8%D8%B1+%D8%A7%D8%B3%D8%A7%D8%B3+%D8%A2%D8%AE%D8%B1%DB%8C%D9%86+%D9%86%D8%AA%D8%A7%DB%8C%D8%AC+%D8%A2%D8%B2%D9%85%D8%A7%DB%8C%D8%B4+%D9%87%D8%A7+%D8%A7%D8%A8%D8%AA%D9%84%D8%A7%DB%8C+13+%D9%85%D9%88%D8%B1%D8%AF+%D8%AF%DB%8C%DA%AF%D8%B1+%D8%A8%D9%87+%DA%A9%D9%88%D9%88%DB%8C%D8%AF19+%D9%82%D8%B7%D8%B9%DB%8C+%D8%A8%D9%87+%D9%86%D8%B8%D8%B1+%D9%85%DB%8C+%D8%B1%D8%B3%D8%AF http://english.alarabiya.net/en/News/gulf/2020/02/25/Number-of-Kuwait-coronavirus-cases-rises-to-eight-KUNA.html http://sxwjw.shaanxi.gov.cn/art/2020/1/27/art\_9\_67483.html http://wjw.beijing.gov.cn/xwzx\_20031/wnxw/202001/t20200121\_1620353.html http://wjw.sz.gov.cn/wzx/202001/t20200120\_18987787.htm   
## highest: https://www3.nhk.or.jp/nhkworld/en/news/20200116\_23/ https://www3.nhk.or.jp/nhkworld/en/news/20200124\_14/ https://www3.nhk.or.jp/nhkworld/en/news/20200126\_31/ https://www3.nhk.or.jp/nhkworld/en/news/20200130\_02/ https://www3.nhk.or.jp/nhkworld/en/news/20200131\_01/   
## --------------------------------------------------------------------------------

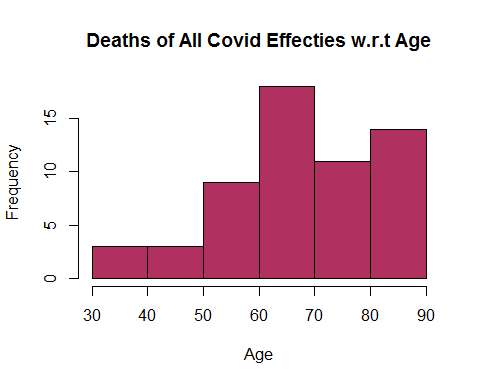
data\_copied$death\_dummy <- as.integer(data$death !=0)  
data\_copied <- select(data\_copied,-c(2,9,10,11,12,13,18,19))  
data\_copied <- na.omit(data\_copied)  
summary(data\_copied)

## ï..id reporting.date summary location   
## Min. : 1 Length:821 Length:821 Length:821   
## 1st Qu.: 240 Class :character Class :character Class :character   
## Median : 450 Mode :character Mode :character Mode :character   
## Mean : 469   
## 3rd Qu.: 682   
## Max. :1085   
## country gender age visiting.Wuhan   
## Length:821 Length:821 Min. : 0.50 Min. :0.0000   
## Class :character Class :character 1st Qu.:35.00 1st Qu.:0.0000   
## Mode :character Mode :character Median :51.00 Median :0.0000   
## Mean :49.82 Mean :0.1778   
## 3rd Qu.:64.00 3rd Qu.:0.0000   
## Max. :96.00 Max. :1.0000   
## from.Wuhan death recovered link   
## Min. :0.0000 Length:821 Length:821 Length:821   
## 1st Qu.:0.0000 Class :character Class :character Class :character   
## Median :0.0000 Mode :character Mode :character Mode :character   
## Mean :0.1827   
## 3rd Qu.:0.0000   
## Max. :1.0000   
## death\_dummy   
## Min. :0.00000   
## 1st Qu.:0.00000   
## Median :0.00000   
## Mean :0.07065   
## 3rd Qu.:0.00000   
## Max. :1.00000

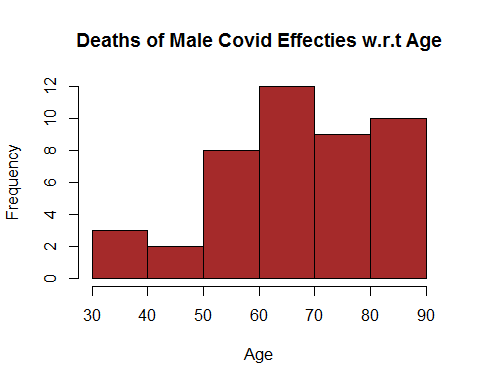
Comment:

Summary() function displays the data after cleaning it from the junk data i.e. N/A data or NaN data is omitted to increase the accuracy of data. Thus, increasing the efficiency of results. All the statistical data is displayed in the last of summary. All columns that has integer values resulted in these statistical data. This data will further be used to conclude promising results.

#Deaths of the patients  
  
death\_all = subset(data\_copied,death\_dummy == 1)  
hist(death\_all$age, col='maroon',main='Deaths of All Covid Effecties w.r.t Age',xlab="Age")



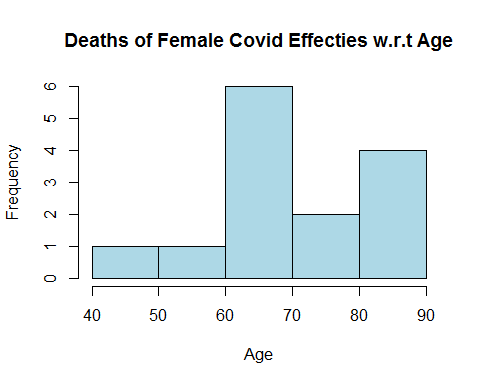
#Gender wise Deaths (Male)  
  
men <- subset(data\_copied, gender == "male")  
mendeath <- subset(men,death\_dummy == 1)  
hist(mendeath$age, col='brown',main='Deaths of Male Covid Effecties w.r.t Age',xlab="Age")



Comment:

Histogram above shows that male counts of death increases with the increase of age. Thus, age factor imposes a great impact on death due to COVID virus. Therefore, we can construct a result that greater the age, greater will the chance of death due to virus.

#Gender wise Deaths (Female)  
  
women <- subset(data\_copied, gender == "female")  
womendeath <- subset(women,death\_dummy == 1)  
hist(womendeath$age, col='light blue',main='Deaths of Female Covid Effecties w.r.t Age',xlab="Age")



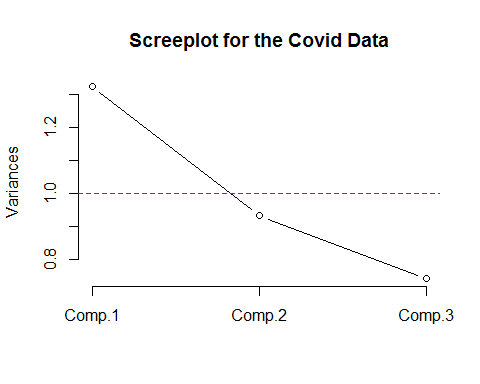
Comment:

Histogram above shows that counts of overall death increases with the increase of age. Thus, age factor imposes a great impact on death due to COVID virus. Therefore, we can construct a result that greater the age, greater will the chance of death due to virus.

#Called Principle Component Analysis on our data  
  
head(data\_copied)

## ï..id reporting.date  
## 1 1 1/20/2020  
## 2 2 1/20/2020  
## 3 3 1/21/2020  
## 4 4 1/21/2020  
## 5 5 1/21/2020  
## 6 6 1/21/2020  
## summary  
## 1 First confirmed imported COVID-19 pneumonia patient in Shenzhen (from Wuhan): male, 66, shenzheng residence, visited relatives in Wuhan on 12/29/2019, symptoms onset on 01/03/2020, returned to Shenzhen and seek medical care on 01/04/2020, hospitalized on 01/11/2020, sample sent to China CDC for testing on 01/18/2020, confirmed on 01/19/2020. 8 others under medical observation, contact tracing ongoing.  
## 2 First confirmed imported COVID-19 pneumonia patient in Shanghai (from Wuhan): female, 56, Wuhan residence, arrived in Shanghai from Wuhan on 01/12/2020, symptom onset and visited fever clinic on 01/15/2020, laboratory confirmed on 01/20/2020  
## 3 First confirmed imported cases in Zhejiang: patient is male, 46, lives in Wuhan, self-driving from Wuhan to Hangzhou on 01/03/2020, symptom onset 01/04/2020, hospitalized on 01/17/2020, sample deliver to China CDC for testing on 01/20/2020, test positive on 01/21/2020.  
## 4 new confirmed imported COVID-19 pneumonia in Tianjin: female, age 60, recently visited Wuhan, visited fever clinic on 01/19/2020 in Tianjin then quarantined immediately.  
## 5 new confirmed imported COVID-19 pneumonia in Tianjin: male, age 58, visited fever clinic on 01/14/2020.  
## 6 First confirmed imported COVID-19 pneumonia patient in Chongqing (from Wuhan): female, age 44, symptoms onset on 01/15/2020, laboratory confirmed on 01/21/2020.  
## location country gender age visiting.Wuhan from.Wuhan death  
## 1 Shenzhen, Guangdong China male 66 1 0 0  
## 2 Shanghai China female 56 0 1 0  
## 3 Zhejiang China male 46 0 1 0  
## 4 Tianjin China female 60 1 0 0  
## 5 Tianjin China male 58 0 0 0  
## 6 Chongqing China female 44 0 1 0  
## recovered  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## link  
## 1 http://wjw.sz.gov.cn/wzx/202001/t20200120\_18987787.htm  
## 2 https://www.weibo.com/2372649470/IqogQhgfa?from=page\_1001062372649470\_profile&wvr=6&mod=weibotime&type=comment  
## 3 http://www.zjwjw.gov.cn/art/2020/1/21/art\_1202101\_41786033.html  
## 4 https://m.weibo.cn/status/4463235401268457?  
## 5 https://m.weibo.cn/status/4463235401268457?  
## 6 http://wsjkw.cq.gov.cn/tzgg/20200121/249730.html  
## death\_dummy  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0

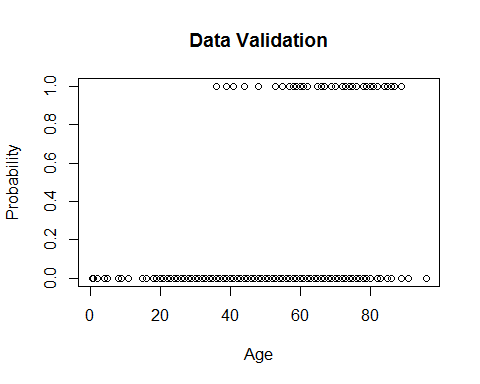
data\_copiedpc <- data\_copied[,c(7,8,9)]  
pc.data <- princomp(data\_copiedpc, cor = TRUE)  
screeplot(pc.data,type="l", main="Screeplot for the Covid Data")  
abline(1,0,col= 'red',lty=2)



Comment:

Screeplot here indicates the out of 3 principle components i.e. (i) age, (ii) visiting wuhan and (iii) from wuhan, data majorly shows deviation from the Age components. It shows that whole data majorly variates with the age variable.

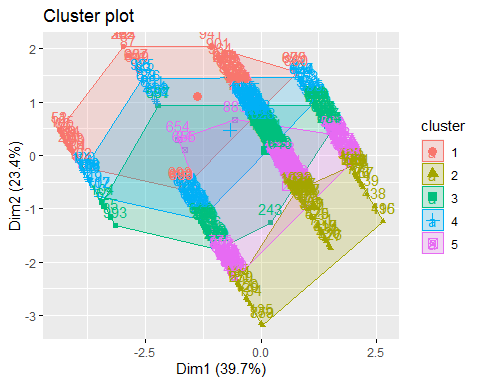
#Regression and Data Validation  
  
death\_dummy <- as.integer(data$death !=0)  
plot(data\_copied$age,data\_copied$death\_dummy,main="Data Validation",xlab = "Age",ylab = "Probability")



Comment:

Apart from Regression, above figure shows that the data does not have any error. Patient/effected person is either death due to virus contraction or survived after contracted the virus. Also, it shows that majority of the patients were recovered from virus while the death ones are concentrated in older ages above 35-years.

#Clustering of Age, From Wuhan, Visiting Wuhan and Deaths  
  
data\_ultra <- select(data\_copied,-c(1,2,3,4,5,6,10,11,12))  
data\_ultra <- na.omit(data\_ultra)  
km <- kmeans(data\_ultra, centers = 5, nstart = 100)  
fviz\_cluster(km, data = data\_ultra)



Comment:

K-mean clustering shows the deviation of the prescribed 04 principle components from their centroid values i.e. Age, Visiting Wuhan, From Wuhan and Deaths. 5 clusters are made and displayed on right side of the cluster plot. To clean the data that included unoccupied column, we used omit command which ultimately deletes the rows with one or more vacant columns.