



UNIVERSITY OF KARACHI

DEPARTMENT OF COMPUTER SCIENCE

PROJECT REPORT: DATA WAREHOUSING AND DATA MINING

ACADEMIC YEAR: 2020-2021

MCS FINAL(MORNING)

PROJECT TITLE: COVID19 DATA ANALYSIS AND VISUALIZATION

GROUP MEMBERS:

1. Numrah Alauddin	P19101052
2. Saba Islam	P19101059
3. Sheikh Muhammad Shayan Iqbal	P19101066
4. Amir Raza	P19101006

SUBMITTED TO

SIR TEHSEEN AHMED JILANI

Table of Contents

1. ABSTRACT	3
2. INTRODUCTION	4
COVID-19 DATA ANALYSIS AND VISUALIZATION:	4
3. LITERATURE REVIEW	5
i. The Analysis of Spread of Corona Virus	6
ii. Comparison Table	6
iii. Comparison Of Death Ratio	8
4. VISUALIZATION OF COVID-19 DATA	9
i. Covid-19 data	10
ii. Multilevel Table	10
iii. Describe data	13
iv. Summary of our data	13
5. CONFUSION MATRIX	19
Why you need Confusion matrix?	19
a) Outcomes of the confusion Matrix	20
b) Confusion Matrix Of Our Covid-19 Data	20
c) Explanation	22
d) Accuracy Test	22
6. CONCLUSION	22
i. Impact of COVID-19 by Age	22
Age is <i>not</i> just a number	23
7. Appendix	28

1. ABSTRACT

Data mining is one of the most promising and ever-changing fields in the field of data analysis. The current paper reviews various papers published on COVID-19 using data mining techniques to address this epidemic in terms of definition, testing and solution. The current paper reviews the work done by various authors using data mining techniques. The paper contributes specifically to the literature by filling the gaps in the review of work related to COVID-19. The following page presents a brief overview of all the events that took place after the plague began. A brief review of SARS-CoV-2 attempted to be a summary of the paper provided. The idea of working for SARS-CoV-2 to represent the effects of the epidemic around the world , is how it changes people's behavior, lifestyle, and their perceptions of nature.

Keywords: COVID-19, data mining, reviews, epidemics, diseases.

2. INTRODUCTION

COVID-19 DATA ANALYSIS AND VISUALIZATION:

Since December 2019 the world has been plagued by a deadly virus caused by the novel coronavirus called acute acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease associated with this virus is known as COVID-19.

The COVID-19 outbreak was first reported in Wuhan, China and has spread to more than 50 countries. The WHO has declared COVID-19 as the International Emergency Medical Concern (PHEIC) on January 30, 2020. Naturally, a growing infectious disease involves the rapid spread, endangerment of many lives, and thus urgent action to prevent this disease. social level. The lethal effect of COVID-19 is conducting a large number of studies aimed at understanding the various features of the epidemic. Although there is no vaccine, much effort has been put into understanding the spread of the disease in various parts of the world. The rate at which the disease has spread worldwide requires immediate solutions to understand and measure the progression of the disease.

According to the World Health Organization (2020), the 2019 coronavirus novel SARS-CoV-2 caused an outbreak of pneumonia in Wuhan, China, leading to the 2019-2020 coronavirus epidemic announced by the World Health Organization (WHO). It belongs to the small family Orthocoronavirinae. It is different from Middle East Respiratory Syndrome (MERS) and the more severe Corona Virus (SARS-CoV). The study of (Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Yu T) suggested that infected patients show clinical manifestations of dry cough, fever, confusion, sore throat, rhinorrhea, chest pain, dyspnea, Pneumonia is common in the brain, nausea, vomiting and diarrhea. The SARS-CoV-2 virus known as COVID-19 can be fatal. This occurs when the initial severity of the disease causes severe alveolar injury and progressive respiratory failure and a mortality rate of 2%.

On January 30, 2020, the World Health Organization (WHO) declared the outbreak of COVID-19 as the sixth international public health emergency, following H1N1 (2009), polio (2014), Ebola in West Africa (2014), Zika (2016), and Ebola in the Democratic Republic of Congo (2019) (Yoo JH, 2020). Since the advent of COVID-19 and its impact on the continents of developing and developing countries, there have been numerous research papers published on various aspects of COVID-19. In addition to the research that is being done on immunization, drug treatment and other clinical aspects, much of the research work is being done with patients as well as fully recovered patients; patients associated with illness and viral event etc. A comprehensive analysis is being done on people recovering to determine how they can cope with these events. Data scientists around the world are busy making sense of available data and predicting the near future. Finding the pattern of trends, the selection of features, the prediction strategies used internally and externally to come to a conclusion (Rajan Gupta, Saibal Kumar, 2020).

3. LITERATURE REVIEW

Coronaviruses belong to the family Coronaviridae according to the Nidovirales system. Corona represents crown-shaped spikes on the outside of the virus; therefore, it was named as coronavirus. Coronaviruses are small in size (65-125 nm in diameter) and contain single-stranded RNA as a nucleic acid, size ranging from 26 to 32kbs in length, as shown by the small coronavirus family groups are alpha (α), beta (β), gamma (γ) and delta (δ) coronavirus. Acute Respiratory Syndrome coronavirus, H5N1 influenza A, H1N1 2009 and Middle East Respiratory Syndrome Coronavirus cause acute lung injury (ALI) and acute Respiratory Distress Syndrome (ARDS) leading to lung failure and the result in death. These viruses were thought to infect only the animals until the world saw the outbreak of a serious respiratory disease (SARS) caused by SARS-CoV, in 2002 in Guangdong, China (N. Zhong, B. Zheng, Y. Li, L. Poon, Z. Xie, K. Chan, et al, 2003).

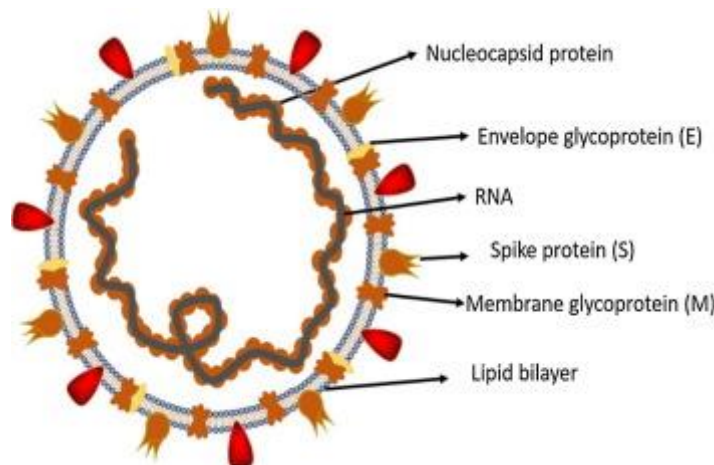


Fig. 1

Historically, SARS-CoV (2003) infected 8098 people with a mortality rate of 9%, in 26 countries around the world, on the other hand, the novel coronavirus (2019) infected 120,000 people with a mortality rate of 2.9%, in all 109 countries, to date this writing. It shows that the SARS-CoV-2 transmission rate is higher than SARS-CoV and the reason may be the event of genetic reunification of S protein in the RBD area of SARS-CoV-2 is likely to improve its transmission capacity. In this review article, we briefly discuss the transmission of human coronaviruses. We also discussed diseases associated with the biological features of SARS and MERS with a particular focus on COVID-19.

i. The Analysis of Spread of Corona Virus

The research analysis suggested that the source of the source and transmission was important to determine in order to develop infection prevention strategies. In the case of SARS-CoV, researchers initially focused on raccoon dogs and palm civets as a last resort. (B. Kan, M. Wang, H. Jing, H. Xu, X. Jiang, M. Yan, et al, 2005).

However, a study by B.J. Zheng, Y. Guan, K.H. Wong, J. Zhou, K.L. Wong, B.W Young, et al in 2004 reported that only samples isolated from the civets of the food market showed positive effects of viral RNA detection, suggesting that palm palms may be secondary hosts. In 2001 the samples were separated from healthy people in Hongkong and molecular testing showed a 2.5% antiretroviral rate against SARS-coronavirus. These indicators suggest that SARS-coronavirus may be circulating in humans before the outbreak in 2003.

In addition, in 2018, Eltahir, et al, Later in Rhinolophus bats were also found to have antibodies against SARS-CoV that elevated bats as a source of viral replication. Middle East Respiratory Syndrome (MERS) coronavirus first appeared in 2012 in Saudi Arabia, MERS-coronavirus is also associated with beta-coronavirus and having camels as a zoonotic source or primary host.

In a recent study by R. In 2020 it was discovered that MERS-coronavirus was also detected in bats in Pipistrellus and Perimyotis. bat strains support the statement that they are not snakes but only bats can be important ponds as shown in the comparison data table. 1.

ii. Comparison Table

	SARS-CoV	SARS-CoV-2	Year
Features			
Emergence date	November 2002	December 2019	2003, 2004,
Area of emergence	Guangdong, China	Wuhan, China	2020
Date of fully controlled	July 2003	Not controlled yet	
Key hosts	Bat, palm civets and Raccon dogs	Bat	2011, 2020
Number of	26	109	2020

	SARS-CoV	SARS-CoV-2	Year
Features			
countries infected			
Entry receptor in humans	ACE2 receptor	ACE2 receptor	2020
Sign and symptoms	and fever, malaise, myalgia, headache, diarrhoea, shivering, cough and shortness of breath	Cough, fever and shortness of breath	2003,2020
Disease caused	SARS, ARDS	SARS, COVID-19	2003, 2020
Total infected patients	8098	123882	2020
Total recovered patients	7322	67051	
Total died patients	776 (9.6% mortality rate)	4473 (3.61% mortality rate)	

iii. Comparison Of Death Ratio

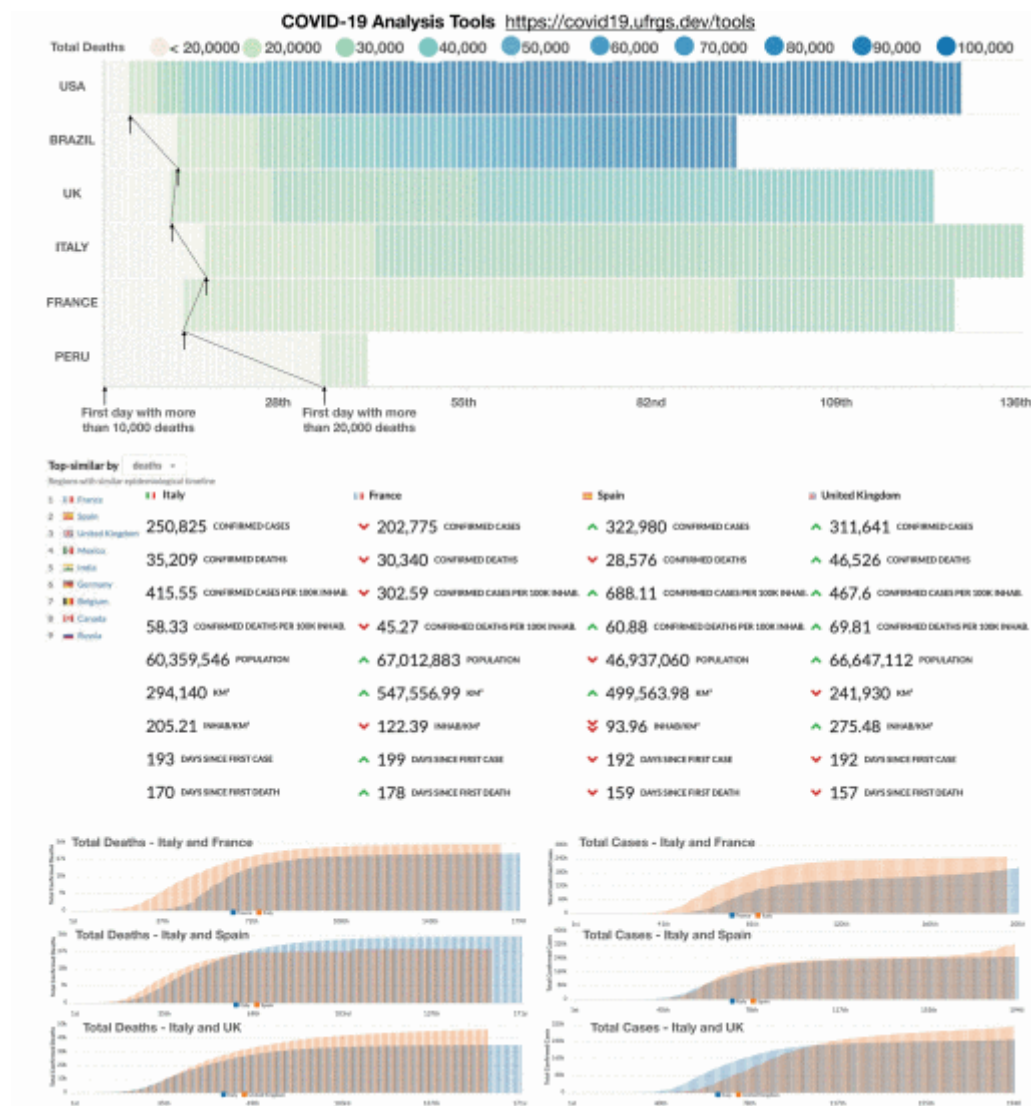


Figure 2.

Description: (Top) Heatmap matriculants are useful in comparing the time series as the mortality rate of different countries. Columns may be aligned on the first date after reaching a certain limit, which allows us to compare when countries pass certain test sites. (Below) Searches for locations with similar mortality rates in Italy.

A large collection of community-developed dashboards and interactive tools for COVID-19 are available. The best places to start looking at a data hub hosted by the Tableau and the top 100 R resources organized by Soetewey. In addition, in-depth analysis is available on sites, such as Our World in Data, Bing, and the COVID Tracking Project, among others. After

developing the Brazilian dashboard, we have devoted our efforts to building a set of tools to compare the spread of COVID-19 data to different regions of the world. We have collected data from the website, and that observation has various charts that support multiple visualizations in a single chart. Since the epidemic is located in different parts of the world, the article allows the user to align a time series of data with a specific data chain beyond a certain limit (e.g., after 100 cases). This representation is useful for observation when different sites pass certain test sites taken from (covid19.ufrgs.dev).

4. Visualization of Covid 19 Data



```

1 |
2 rm(list=ls())
3 library(Hmisc)
4 library(factoextra)
5 library(dplyr)
6

```

line 2 is removing the Global environments values that previously given clear all previous values.
line 3,4 and 5 is calling library to give output which is needed to run our code



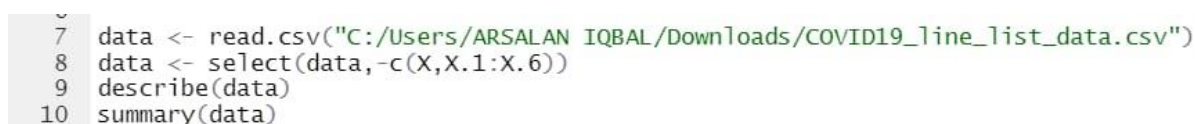
```

1 |
2 rm(list=ls())
3 library(Hmisc)
4 library(factoextra)
5 library(dplyr)
6 |
7 data <- read.csv("C:/Users/ARSALAN IQBAL/Downloads/COVID19_line_list_data.csv")
8 data <- select(data, -c(X,X.1:X.6))

```

Line 7 is calling our data or reading our file from the source address given.

Line 8 is removing variables from our data because variables have no values and still in our data so we just remove that variables.



```

7 data <- read.csv("C:/Users/ARSALAN IQBAL/Downloads/COVID19_line_list_data.csv")
8 data <- select(data, -c(X,X.1:X.6))
9 describe(data)
10 summary(data)

```

Line 9 is describing our data frame.

Line 10 is giving summary. It is a generic function used to produce result summaries of the results of various model fitting functions.

i. COVID-19 DATA:

i.Id	reporting.date	summary	location	country	gender	age	visiting.Wuhan	from.Wuhan	death	recovered	death_dummy
49	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 61, sym...	Wuhan, Hubei	China	male	61	0	1	1	0	1
50	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 69, sym...	Wuhan, Hubei	China	male	69	0	1	1	0	1
51	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 69, pneu...	Wuhan, Hubei	China	male	69	0	1	1	0	1
52	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 69, sym...	Wuhan, Hubei	China	male	69	0	1	1	0	1
53	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 66, sym...	Wuhan, Hubei	China	male	66	0	1	1	0	1
54	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 75, sym...	Wuhan, Hubei	China	male	75	0	1	1	0	1
55	1/22/2020	Death from COVID-19 pneumonia in Wuhan: female, 48, sy...	Wuhan, Hubei	China	female	48	0	1	1	0	1
56	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 62, sym...	Wuhan, Hubei	China	male	62	0	1	1	0	1
57	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 66, coug...	Wuhan, Hubei	China	male	66	0	1	1	0	1
58	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 61, sym...	Wuhan, Hubei	China	male	61	0	1	1	0	1
59	1/22/2020	Death from COVID-19 pneumonia in Wuhan: female, 62, ho...	Wuhan, Hubei	China	female	62	0	1	1	0	1
60	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 65, sym...	Wuhan, Hubei	China	male	65	0	1	1	0	1
61	1/22/2020	Death from COVID-19 pneumonia in Wuhan: female, 60, sy...	Wuhan, Hubei	China	female	60	0	1	1	0	1
62	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 53, sym...	Wuhan, Hubei	China	male	53	0	1	1	0	1
63	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 66, sym...	Wuhan, Hubei	China	male	66	0	1	1	0	1
64	1/22/2020	Death from COVID-19 pneumonia in Wuhan: female, 70, ho...	Wuhan, Hubei	China	female	70	0	1	1	0	1
65	1/22/2020	Death from COVID-19 pneumonia in Wuhan: male, 64, sym...	Wuhan, Hubei	China	male	64	0	1	1	0	1
82	1/23/2020	new death from COVID-19 pneumonia, female, 65, death on...	Hubei	China	female	65	0	0	1	0	1
83	1/23/2020	new death from COVID-19 pneumonia, female, 69, sympto...	Hubei	China	female	69	0	0	1	0	1
84	1/23/2020	new death from COVID-19 pneumonia, male, 36, symptom ...	Hubei	China	male	36	0	0	1	0	1
85	1/23/2020	new death from COVID-19 pneumonia, male, 73, symptom ...	Hubei	China	male	73	0	0	1	0	1
86	1/23/2020	new death from COVID-19 pneumonia, female, 70, sympto...	Hubei	China	female	70	0	0	1	0	1
87	1/23/2020	new death from COVID-19 pneumonia, male, 61, symptom ...	Hubei	China	male	61	0	0	1	0	1
88	1/23/2020	new death from COVID-19 pneumonia, female, 65, hospital...	Hubei	China	female	65	0	0	1	0	1

Showing 1 to 25 of 58 entries. 12 total columns.

ii. MULTILEVEL TABLE

(Cases record from Starting first month Of Covid-19)

COUNTRY	CITY	Total Cases	DEATH	RECOVERY
CHINA	Shenzhen, Guangdong	1	0	1
	Shanghai	2	0	2
	Zhejiang	1	0	1
	Tianjin	22	0	22
	Chongqing	1	0	1
	Sichuan	7	0	7
	Beijing	18	0	18
	Shandong	3	0	3
	Yunnan	19	0	19
	Sichuan	1	0	1
	Jiangxi	3	0	3
	Macau	4	0	4
	Liaoning	4	0	4
	Fujian	1	0	1
	Guizhou	1	0	1
	Shanxi	35	0	35
	Ningxia	1	0	1
	Guangxi	5	0	5
	Henan	4	0	4
	Hebei	1	0	1
	Jiangsu	1	0	1
	Heilongjiang	2	0	2
	Jilin	2	0	2
	Wuhan	40	39	1

	Hunan	5	0	5
	Guizhou	1	0	1
	Gansu	1	0	1
	Xinjiang	3	0	3
	Inner Mongolia	1	0	1
	Hechi, Guangxi	2	0	2
	Gansu	2	0	2
	Ningxia	1	0	1
	Guizhou	1	0	1
	Jiangxi	1	0	1
AUSTRALIA	NSW	4	0	4
	VICTORIA	4	0	4
	QUEENSLAND	5	0	5
	SOUTH AUSTRALIA	2	0	2
CAMBODIA	Preah Sihanouk	1	0	1
CANADA	TORONTO	5	0	5
	VANCOUVER	7	0	7
FINLAND	LAPLAND	1	0	1
FRANCE	PARIS	5	2	3
	BORDEAUX	1	0	1
	ANNECY	3	0	3
	AMIENS	1	0	1
	STRASBOURG	1	0	1
	NANTES	1	0	1
	MONTPELLIER	1	0	1
	BREST	1	0	1
	LYON	1	0	1
	NICE	1	0	1
GERMANY	BAVARIA	5	0	5
	Baden-Wuerttemberg	3	0	3
	Tubingen	2	0	2
	North Rhine-Westphalia	3	0	3
	HESSE	1	0	1
HONG KONG	KOWLOON	4	1	3
	HONG KONG	79	1	78
	FO TAN	1	0	1
	Kwun Tong	2	0	2
	Yau Ma Tei	1	0	1
	Tsing Yi	1	0	1
	Kwai Chung	2	0	2
	Zhuhai	1	0	1
	Wan Chai	1	0	1
	Ngau Chi Wan	1	0	1
ITALY	ROME	1	0	1
	JAPAN	15	1	14
	TOKYO	25	1	24
	Aichi Prefecture	7	0	7
	Chiba Prefecture	11	0	11
	Fukuoka Prefecture	2	0	2
	Gifu Prefecture	2	0	2

JAPAN	Haneda	1	0	1
	Hokkaido	47	1	46
	Ishikawa	5	0	5
	Kanagawa	7	1	6
	Kumamoto City	4	0	4
	Kumamoto Prefecture	1	0	1
	Wakayama Prefecture	12	0	12
	Sapporo	5	0	5
	Sagamihara	9	0	9
	Nara Prefecture	2	0	2
	Nagoya City	20	0	20
	Kyoto	2	0	2
	Osaka Prefecture	3	0	3
	Mie	1	0	1
	Okinawa Prefecture	3	0	3
	Nagano Prefecture	1	0	1
LEBANON	LEBANON	1	0	1
MALAYSIA	Johor	8	0	8
	MALAYSIA	14	0	14
	LAGKAWI	1	0	1
NEPAL	Kathmandu	1	0	1
PHILLIPINES	Manila	2	1	1
	PHILLIPINES	1	0	1
SINGAPORE	SINGAPORE	90	0	90
SOUTH KOREA	SOUTH KOREA	90	9	81
	SEOUL	2	0	2
SPAIN	Andalusia	7	0	7
	Barcelona	3	0	3
	Castellon	1	0	1
	Castile and Leon	1	0	1
	Tenerife	1	0	1
	Valencia	1	0	1
	Zaragoza	1	0	1
Sri Lanka	Sri Lanka	1	0	1
SWEDEN	Jonkoping	1	0	1
Switzerland	BERN	1	0	1
TAIWAN	TAIWAN	31	1	30
Thailand	Thailand	16	0	16
DUBAI	UAE	7	0	7
UK	LONDON	1	0	1
USA	Washington	1	0	1
	Illinois	2	0	2
	Massachusetts	1	0	1
	California	2	0	2
Vietnam	Ho Chi Minh City	2	0	2
	Vinh Phuc	6	0	6

iii. DESCRIBE DATA:

```

Console terminal
R 4.1.1 - ~/Covid Project/
> describe(data)
data
  20 variables      1085 observations
-----
i..id
  n missing distinct    Info    Mean    Gmd
1085      0      1085      1      543      362
.05      .10      .25      .50      .75      .90
55.2     109.4     272.0    543.0    814.0    976.6
.95
1030.8

lowest :      1      2      3      4      5, highest: 1081 1082 1083 1084 1085
-----
case_in_country
  n missing distinct    Info    Mean    Gmd
888      197      197      1     48.84    54.99
.05      .10      .25      .50      .75      .90
2.00      4.00     11.00    28.00    67.25   110.30
.95
153.65

lowest :      1      2      3      4      5, highest: 365 443 875 925 1443

0 (215, 0.242), 20 (241, 0.271), 40 (137, 0.154), 60 (81,
0.091), 80 (84, 0.095), 100 (40, 0.045), 120 (22, 0.025), 140
(19, 0.021), 160 (22, 0.025), 180 (19, 0.021), 200 (1,
0.001), 280 (1, 0.001), 300 (1, 0.001), 360 (1, 0.001), 440
(1, 0.001), 880 (1, 0.001), 920 (1, 0.001), 1440 (1, 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
-----

```

iv. SUMMARY OF DATA:

```

> summary(data)
i..id      case_in_country  reporting.date  summary      location
Min.   : 1  Min.   : 1.00  Length:1085  Length:1085  Length:1085
1st Qu.:272 1st Qu.: 11.00  Class :character  Class :character  Class :character
Median :543 Median : 28.00  Mode :character  Mode :character  Mode :character
Mean   :543 Mean   : 48.84
3rd Qu.:814 3rd Qu.: 67.25
Max.   :1085 Max.   :1443.00
NA's   :197

country    gender    age    symptom_onset
Length:1085 Length:1085  Min.   : 0.25  Length:1085
Class :character Class :character 1st Qu.:35.00  Class :character
Mode :character  Mode :character Median :51.00  Mode :character
Mean   :49.48
3rd Qu.:64.00
Max.   :96.00
NA's   :242

If_onset_approximated hosp_visit_date exposure_start exposure_end
Min.   :0.0000  Length:1085  Length:1085  Length:1085
1st Qu.:0.0000  Class :character  Class :character  Class :character
Median :0.0000  Mode :character  Mode :character  Mode :character
Mean   :0.0429
3rd Qu.:0.0000
Max.   :1.0000
NA's   :525

visiting_wuhan  from_wuhan  death  recovered  symptom
Min.   :0.000  Min.   :0.0000  Length:1085  Length:1085  Length:1085
1st Qu.:0.000  1st Qu.:0.0000  Class :character  Class :character  Class :character
Median :0.000  Median :0.0000  Mode :character  Mode :character  Mode :character
Mean   :0.177  Mean   :0.1443
3rd Qu.:0.000  3rd Qu.:0.0000
Max.   :1.000  Max.   :1.0000
NA's   :4

source  link
Length:1085 Length:1085
Class :character Class :character
Mode :character  Mode :character

```

```

7 data <- read.csv("C:/Users/ARSALAN IQBAL/Downloads/COVID19_line_list_data.csv")
8 data <- select(data,-c(X,X.1:X.6))
9 describe(data)
10 summary(data)
11 data_copied <- data
12 data_copied$death_dummy <- as.integer(data$death !=0)

```

Line 11 is saving our data in "data_copied" data frame as a duplicate.

Line 12 we adding variable "death_dummy" in in data_copied and adding values from its death column, we did this because in death variable we have dates and we want numeric data 0 or 1.

```

11 data_copied <- data
12 data_copied$death_dummy <- as.integer(data$death !=0)
13
14
15 data_copied <- select(data_copied,-c(2,9:13,18:20))
16 data_copied <- na.omit(data_copied)

```

Line 15 now we will remove the variable that are not usable for us right now.

for eg:- 1.case.in.country, 2.symptom.onset, 3.if.onset.approximated, 4, hosp.visitdate, 5.exposure.start, 6.exposure.end, 7.symptom, 8.source, 9.link

Line 16 we just clean our data by using command "na.omit". it will remove the empty row from our data and give us the usefull data that we can work on it without missing anything.

```

17
18 death_all = subset(data_copied,death_dummy == 1)
19 alive_all = subset(data_copied,death_dummy == 0)
20

```

Line 18 we will calculate the death to know how many death occurs because of Covid.

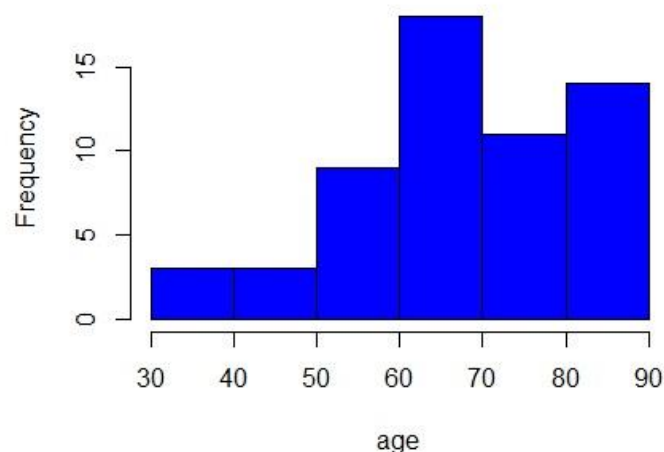
Line 19 we will calculate the data to know how many people is alive in this period of Covid.

```

21 hist(death_all$age, col="blue",xlab="age",main='Histogram of deaths and age')
22

```

Histogram of deaths and age




```

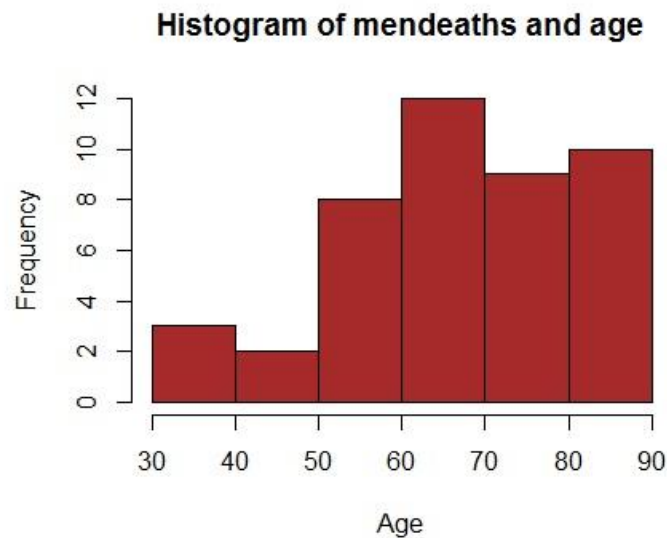
23 men <- subset(data_copied, gender == "male")
24 women <- subset(data_copied, gender == "female")
25 mendeath <- subset(men, death_dummy == 1)
26 hist(mendeath$age, col='brown',main='Histogram of mendeaths and age',xlab="Age")
27

```

Line 23 and 24 we separate the data by gender wise.

Line 25 we calculate the death of male by Covid.

Line 26 we plot a Histogram by Command "Hist()"



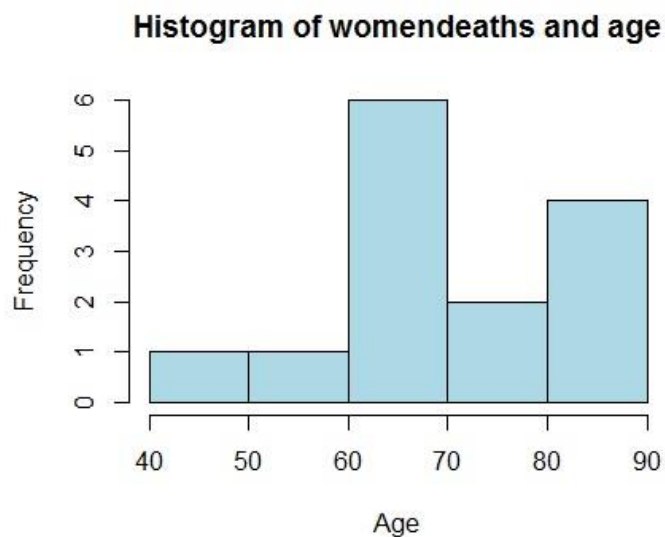
```

27
28 womendeath <- subset(female, death_dummy == 1)
29 hist(womendeath$age, col='light blue',main='Histogram of womendeaths and age',xlab="Age")
30

```

Line 28 we will calculate the female deaths by covid.

Line 29 we plot histogram of women deaths along with the Age.



```

31 malealive <- subset(male,death_dummy == 0)
32 femalealive <- subset(female,death_dummy == 0)
33 |

```

Line 31 we will calculate the male person's who are alived in this Covid period.

Line 32 we will calculate the female person's who are alived in this Covid period.

OBSERVATION

data_copied	821 obs. of 12 variables
death_all	58 obs. of 12 variables
female	347 obs. of 12 variables
femalealive	333 obs. of 12 variables
femaledeath	14 obs. of 12 variables
male	474 obs. of 12 variables
malealive	430 obs. of 12 variables
maleddeath	44 obs. of 12 variables

```

35 mean_maleddeath <- mean(male$death_dummy)
36 mean_femaledeath <- mean(female$death_dummy)
37

```

Line 35 it will calculate the mean of male deaths.

Line 36 it will calculate the mean of female deaths.

OUTPUT

values	
mean_femaled...	0.0403458213256484
mean_maleddea...	0.0928270042194093

```

38 mean_age_death <- mean(death_all$age)
39 mean_age_alive <- mean(alive_all$age)

```

Line 38 it will calculate the mean of Age of Dead Peoples.

Line 39 it will calculate the mean of Age of Alive Peoples.

OUTPUT

values	
mean_age_alive	48.3938401048493
mean_age_death	68.5862068965517


```

44 #Principle Component Analysis
45
46 head(data_copied[,c(4,7,8,9,12)])
47 tail(data_copied[,c(4,7,8,9,12)])

```

Line 46 it will give you the starting values of the given column that we have selected for output and that are "Location Age From.Wuhan Visiting.Wuhan and Death Dummy".

Line 47 it will give you the ending values of the given column that we have selected for output and that are "Location Age From.Wuhan Visiting.Wuhan and Death Dummy".

OUTPUT

```

> head(data_copied[,c(4,7,8,9,12)])
  location age visiting.wuhan from.wuhan death_dummy
1 shenzhen, Guangdong 66      1         0          0
2      Shanghai 56      0         1          0
3      Zhejiang 46      0         1          0
4      Tianjin 60      1         0          0
5      Tianjin 58      0         0          0
6      Chongqing 44      0         1          0
> tail(data_copied[,c(4,7,8,9,12)])
  location age visiting.wuhan from.wuhan death_dummy
1027 Andalusia 25      0         0          0
1028 Andalusia 58      0         0          0
1030 Zaragoza 27      0         0          0
1031 Jonkoping 25      1         0          0
1053 Lebanon 45      0         0          0
1085 Bern 70      0         0          0
> |

```

```

48 write.table(data_copied[,c(4,7,8,9,12)],file="pca data table1.csv",sep = ",",row.names = F)
49 sdatapc <- data_copied[,c(7,8,9,12)]
50

```

Line 49 it will write our table or data in to a file and save it in my computer.

Line 50 now we jjust remove the column location because we need numeric data and its alphatical data and save to sdatapc data frame.

```

51 pc.data <- princomp(sdatapc, cor=TRUE)
52 names(pc.data)
53 summary(pc.data)

```

Line 51 princomp performs a principal components analysis on the given numeric data matrix and returns the results as an object of class princomp.

Line 52 Functions to get or set the names of an object.

Line 53 It will generate summary and summary is a generic function used to produce result summaries of the results of various model fitting functions

OUTPUT

```

> pc.data <- princomp(sdatapc, cor=TRUE)
> names(pc.data)
[1] "sdev"      "loadings"  "center"    "scale"     "n.obs"     "scores"    "call"
> summary(pc.data)
Importance of components:
              Comp.1      Comp.2      Comp.3      Comp.4
Standard deviation  1.2607955  0.9665549  0.9434363  0.7655679
Proportion of Variance 0.3974013  0.2335571  0.2225180  0.1465236
Cumulative Proportion 0.3974013  0.6309584  0.8534764  1.0000000
> |

```

Name	Type	Value
pc.data	list [7] (S3: princomp)	List of length 7
sdev	double [4]	1.261 0.967 0.943 0.766
loadings	double [4 x 4] (S3: loadings)	0.4838 -0.4497 0.4933 0.5660 0.7126 0.1725 -0.6706 0.1125 0.0862 -0.8123 ...
center	double [4]	49.8203 0.1778 0.1827 0.0706
scale	double [4]	17.929 0.382 0.386 0.256
n.obs	integer [1]	821
scores	double [821 x 4]	-0.91974 1.26325 0.99343 -1.08163 0.04057 0.93947 1.30006 -1.28403 -1.68150 ...
call	language	princomp(x = sdatapc, cor = TRUE)

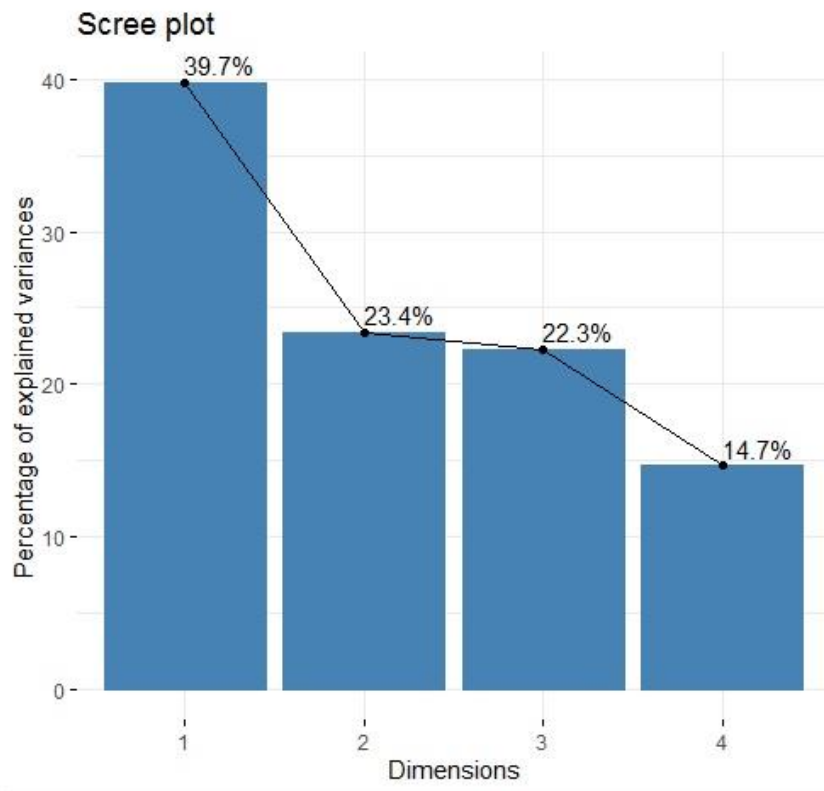
```

54
55 fviz_eig(pc.data,addlabel=TRUE)
56

```

Line 55 it will plot the screeplot of the pc data.

OUTPUT



5. CONFUSION MATRIX

The confusion matrix is a way of measuring the performance of a learning machine. It is a type of table that helps you to know the functionality of the partition in the test data set so that the true values are known. In simple terms, “Confusion matrix is a performance measure of machine learning algorithm”.

By visualizing the matrix of confusion, one can determine the accuracy of a model by looking at diagonal values to estimate the value of precise separation.

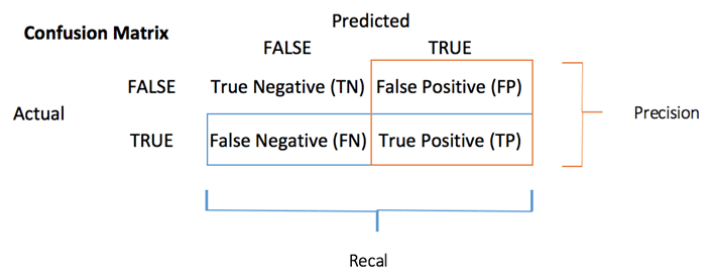
Why you need Confusion matrix?

- It shows how any differentiation model is confused when creating predictions.
- The confusion matrix not only gives you an idea of the mistakes that are made for your student but also the types of mistakes that are set.
- This classification helps you to overcome the limit of using the accuracy of the sections themselves.
- Each column of the confusion matrix represents the conditions of the predicted phase.

- Each line of the matrix of confusion represents the conditions of the real phase.
- It not only provides insight into classified errors but also errors.

a) OUTCOMES OF THE CONFUSION MATRIX

The confusion matrix visualizes the accuracy of a class divider by associating real and predicted categories. The binary confusion matrix is square:



- TP: True Positive: Predicted values appropriately predicted as actual positive
- FP: Predicted values inaccurately predicted an actual positive. i.e., Negative values predicted as positive
- FN: False Negative: Positive values anticipated as negative
- TN: True Negative: Predicted values appropriately predicted as an actual negative

b) CONFUSION MATRIX OF OUR COVID-19 DATA

```

77 table(data_copied$death_dummy)
78 mymodel=glm(death_dummy~visiting.wuhan+from.wuhan+age,family=binomial,data_copied)
79 print(mymodel)
80 pmymodel=predict(mymodel,data_copied)
81 tab=table(pmymodel>0.5,data_copied$death_dummy)
82 print(tab)
83 accuracy <- sum(diag(tab))/sum(tab)*100
84 print(accuracy)
85

```

Line 77 it will show the tables of death and that are how many people alive and how many peoples are dead.

Line 78 we have applied generalized linear model for the outcome of confusion matrix.

Line 79 print my model data.

Line 80 predict the data model.

Line 81 make the confusion matrix table.

Line 82 print the table.

Line 83 find the accuracy of the model.

Line 84 Print the accurany of the model.

OUTPUT

```

> table(data_copied$death_dummy)

 0  1
763 58
> mymodel=glm(death_dummy~visiting.wuhan+from.wuhan+age,family=binomial,data_copied)
> print(mymodel)

Call: glm(formula = death_dummy ~ visiting.wuhan + from.wuhan + age,
          family = binomial, data = data_copied)

Coefficients:
(Intercept)  visiting.wuhan    from.wuhan         age
   -8.1557      -0.9827       2.1295       0.0823

Degrees of Freedom: 820 Total (i.e. Null);  817 Residual
Null Deviance: 419.2
Residual Deviance: 291.3    AIC: 299.3
> pmymodel=predict(mymodel,data_copied)
> tab=table(pmymodel>0.5,data_copied$death_dummy)
> print(tab)

      0  1
FALSE 763 47
TRUE   0 11
> accuracy <- sum(diag(tab))/sum(tab)*100
> print(accuracy)
[1] 94.27527
> |

```

c) Explanation

The output shows the values of confusion table as:

- TN: True Negative = 763, i.e. these patients predicted as they will not die and they actually not died.
- FP: False Positive =47, i.e. these patients predicted as they will die but they did not.
- FN: False Negative =0, i.e. these patients predicted as they will not die but they died.
- TP: True Positive =11, i.e. these patients predicted as they will die and they actually died.

d) ACCURACY TEST

One of the most important parameters in determining the accuracy of division problems, is how the model usually predicts the correct output and can be measured as the number of true predictions made by the classifier over the total number of calculations made by classifier.

Calculated accuracy from the confusion matrix with the following formula:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The output shows the accuracy test result as **94.27**, which shows that our model predicts output 94% correctly which is the high ratio and good for our further data mining analysis.

6. Conclusion

COVID-19 is spreading rapidly around the world. Within a few months, the mortality and morbidity levels reached an unexpected level. Doctors are working to develop treatments and vaccines to prevent the spread of the disease. A terrible situation is yet to come. However, if we take just one step of isolation, it can save the whole community and the risk will decrease dramatically. This is a situation where each person should take steps to reduce the risk by staying indoors and taking personal action. Transmission, the contact can be disinfected only if proper hand washing procedures are followed and individuals take precautionary measures to protect other people from the invading virus. The world has great potential for public health, and the various sectors can work together to address the challenges of public-private partnerships and policy initiatives.

i. Impact of COVID-19 by Age

This study aims to integrate existing research into the effects of the COVID-19 epidemic, as well as related interventions for isolation and prevention strategies, in adults. The second objective is to investigate the effects of the COVID-19 epidemic, as well as the associated measures of isolation and prevention measures, in older adults with Alzheimer's disease and related dementia.

While COVID-19, a disease caused by the new coronavirus, can lead to hospitalization and even death in young and middle-aged adults, creating more serious health problems for adults over the age of 60 - with deadly consequences especially for those 80 years and older. This is due to the low level of health conditions that exist in older people. Conditions such as diabetes, heart disease, and other chronic illnesses can lead to severe symptoms and complications. Additionally, as people get older, their immune system gradually loses its strength, which means they are at risk of infection of any kind, especially new ones like COVID-19.

Adults are apparently among the groups most at risk of illness and death for COVID-19. Data from China after the first few months of the Wuhan outbreak highlighted the risk of serious illness and death of COVID-19 in the elderly, with a mortality rate of nearly 15 percent in people aged 80 and older, compared with a total of 2.3 percent in the general population.

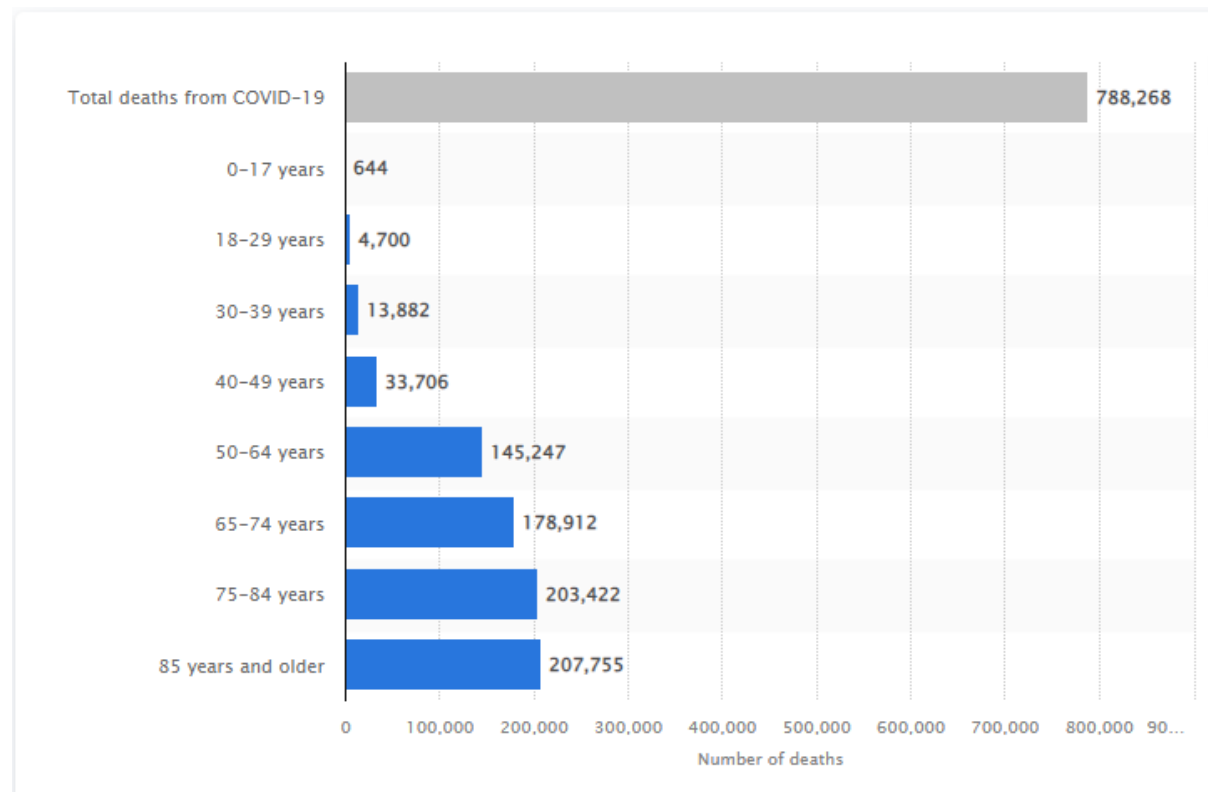
Age is not just a number

Older adults are the main victims of the COVID-19 novel outbreak and older people in Long Term Care Facilities (LTCFs) are particularly vulnerable to death. This Report presents a qualitative study of the impact of the COVID-19 outbreak in Wuhan during the early stages of the epidemic, which focuses on the effects of increased mortality among older people over the age of 80 and its relationship to LTCFs. The study of growth patterns shows an increase in legal status in the early stages of the epidemic through unequal behavior between different regions and an increase in deaths due to the different effects of COVID-19. However, the incidence rate of COVID-19 does not fully explain the effect of the impact on adult mortality among different regions. We explain the volume correlation between adult mortality and the number of people on LTCFs confirming the significant impact of COVID-19 on LTCFs. Adding a link between LTCFs and undiagnosed conditions as well as the effects of health system malfunctions is also being observed. Our results confirm that LTCFs did not play a protective role for older adults during the epidemic, but when the number of older people living on LTCFs increased the number of common and COVID-19-related deaths. We also observed that problem management in LTCFs disrupted the tracking of COVID-19 prevalence and promoted an increase in deaths not directly caused by SARS-CoV-2.

While we should be very careful in dividing the age when it comes to determining who needs intensive care by providing services to young people and not the elderly it is obviously the most important factor in predicting the chances of survival of COVID-19. In fact, this is one of the strongest consensus among scientists. In a good country, that alone should justify the expectation that older people will be more effective in following community health recommendations. However, the issue is not so straightforward.

COVID-19 is not the only disease known to have serious side effects in the elderly. This is a complication of pneumococcal infection or fever, among others. In both cases, there are preventive measures available.

That being said, the mortality rate of COVID-19 among the elderly is, in fact, much higher than other diseases such as pneumococcal disease or fever. In addition, the response of governments and citizens and the measures taken to reduce the effects of the disease are to a completely different degree. Therefore, it is only reasonable to expect that older people will be more effective when it comes to following the recommendations of public health institutions and the government. The following graph shows the death rate by age group.



References

- A. Soetewey, "Top 100 R resources on novel COVID-19 coronavirus", [online] Available: <https://www.statsandr.com/blog/top-r-resources-on-covid-19-coronavirus>.
- B. Kan, M. Wang, H. Jing, H. Xu, X. Jiang, M. Yan, *et al.* **Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms.** J Virol, 79 (18) (2005), pp. 11892-11900
- B.J. Zheng, Y. Guan, K.H. Wong, J. Zhou, K.L. Wong, B.W.Y. Young, *et al.* **SARS-related virus predating SARS outbreak, Hong Kong** Emerg Infect Dis, 10 (2) (2004), p. 176
- C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, *et al.* **Clinical features of patients infected with 2019 novel coronavirus in Wuhan China.** The Lancet (2020)
- C. Paden, M. Yusof, Z. Al Hammadi, K. Queen, Y. Tao, Y. Eltahir, *et al.* **Zoonotic origin and transmission of Middle East respiratory syndrome coronavirus in the UAE.** Zoonoses Public Health, 65 (3) (2018), pp. 322-333
- Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Yu T. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. The Lancet. 2020; 395(10223):507-513.
- Dey SK, Rahman MM, Siddiqi UR, Howlader A. Analysing the Epidemiological Outbreak of COVID-19: A Visual Exploratory Data Analysis (EDA) Approach. Journal of Medical Virology; 2020.
- F. Pontes, F. Pinto, W. Leuschner and J. L. D. Comba, "COVID-19 analysis tools at Instituto de Informática-ufrgs", [online] Available: <https://covid19.ufrgs.dev>

Hui DS, I Azhar E, Madani TA, Ntoumi F, Kock R, Dar O, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases*. 2020; 91:264–6.

J.F.-W. Chan, S. Yuan, K.-H. Kok, K.K.-W. To, H. Chu, J. Yang, *et al.* **A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster.** *Lancet* (2020).

N. Zhong, B. Zheng, Y. Li, L. Poon, Z. Xie, K. Chan, *et al.* **Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003.** *The Lancet*, 362 (9393) (2003), pp. 1353-1358

R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, *et al.* **Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.** *The Lancet* (2020)

Rajan Gupta, Saibal Kumar. Trend Analysis and Forecasting of COVID-19 outbreak in India. Rajan Gupta, Saibal Kumar PalmedRxiv, 2020.

Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Agha R. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *International Journal of Surgery*; 2020.

Tian X, Li C, Huang A, Xia S, Lu S, Shi Z, et al. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *bioRxiv*; 2020.

Vara V. Coronavirus outbreak: The countries affected. 11 MARCH 2020; Available from: <https://www.pharmaceutical-technology.com/features/coronavirus-outbreak-the-countries-affected/>.

Yoo JH. The fight against the 2019-nCoV outbreak: an arduous march has just begun. *J Korean Med Sci.* 2020;35: e56 10.

7. Appendix

Code:

```
rm(list=ls())
library(Hmisc)
library(factoextra)
library(dplyr)

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))
describe(data)
summary(data)
data_copied <- data
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)

death_all = subset(data_copied, death_dummy == 1)
alive_all = subset(data_copied, death_dummy == 0)

hist(death_all$age)

men <- subset(data_copied, gender == "male")
women <- subset(data_copied, gender == "female")
menddeath <- subset(men, death_dummy == 1)
womenddeath <- subset(women, death_dummy == 1)

menalive <- subset(men, death_dummy == 0)
womenalive <- subset(women, death_dummy == 0)

mean_menddeath <- mean(men$death_dummy)
mean_womenddeath <- mean(women$death_dummy)

mean_age_death <- mean(death_all$age)
mean_age_alive <- mean(alive_all$age)

hist(menddeath$age, col='brown')
hist(womenddeath$age, col='light blue')

t.test(alive_all$age, death_all$age, alternative="two.sided",
conf.level = 0.95)

t.test(men$death_dummy, women$death_dummy, alternative="two.sided",
conf.level = 0.99)

#Principle Component Analysis

head(data_copied)
data_copiedpc <- data_copied[,c(7,8,9)]
pc.data <- princomp(data_copiedpc, cor = TRUE)
names(pc.data)
```

```
summary(pc.data)
eigenvectors <- pc.data$loadings
eigenvalues <- pc.data$sdev *pc.data$sdev
screeplot(pc.data,type="l", main="screeplot for the covid data")
abline(1,0,col= 'red',lty=2)

#REGRESSION

plot(data_copied$age,data_copied$death_dummy)

#CLUSTERING

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)
```