Covid- 19 Analysis using R Programming

Big Data

Department of Computer Science and Engineering

SRM University, AP

Nov – 2021

Pawan Aditya M - AP18110010085 - CSE B
Bhavana Keerthi Sri Nettam - AP18110010098 - CSE B
Mokshagna Kalki Suhas Amaraneni - AP18110010114 - CSE B
Mekapothula Sai Sathwik - AP18110010091 - CSE B

Abstract

Corona Virus may be diagnosed quickly and efficiently with effective COVID screening, by reducing the load on healthcare systems. These are intended to aid medical personnel across the globe in triaging patients, particularly in areas where healthcare resources are few. So from the data of the covid positive we have analysed and gained some effective knowledge and verified some prediction about it. COVID-19 was found in all of the people who were tested. The data in the test set came from the next week. sex, age years, and the location where it was most impacted, main features of this analysis is to get the needed information and prove some assumption on the covid 19 so from the data after we pre-process the data we have first taken the death rate and then we have to prove some assumption about the covid 19 that is the older people are most likely to die of covid than the younger once

Introduction:

2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. Early on, many of the patients in the outbreak in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring. At this time, it’s unclear how easily or sustainably this virus is spreading between people

With proper screening, COVID-19 may be detected swiftly and efficiently, decreasing the burden on healthcare systems. Prediction models that combine a variety of characteristics to predict the likelihood of infection have been developed in the hopes of assisting medical practitioners throughout the world in triaging patients, particularly in countries with limited healthcare resources.

Controlling an infectious illness like COVID-19 is a crucial, time-sensitive, yet challenging task. As research is geared toward vaccines and countries scurry to establish public health measures to minimize the spread of the illness, the health of the world population is possibly the most crucial aspect. These measures have taken the form of local or national lockdowns in most nations throughout the world, when people are urged or obliged to stay at home unless they have a strong reason not to, such as educational or medical reasons, or if they are unable to work from home. The consequences of attempting to manage COVID-19, on the other hand, are not limited to the health industry.

Problem survey:

The main features of this analysis is to get the needed information and prove some assumption on the covid 19 so from the data after we pre-process the data we have first taken the death rate and then we have to prove some assumption about the covid 19 that is the older people are most likely to die of covid than the younger once and then male has for fatality rate then the female.

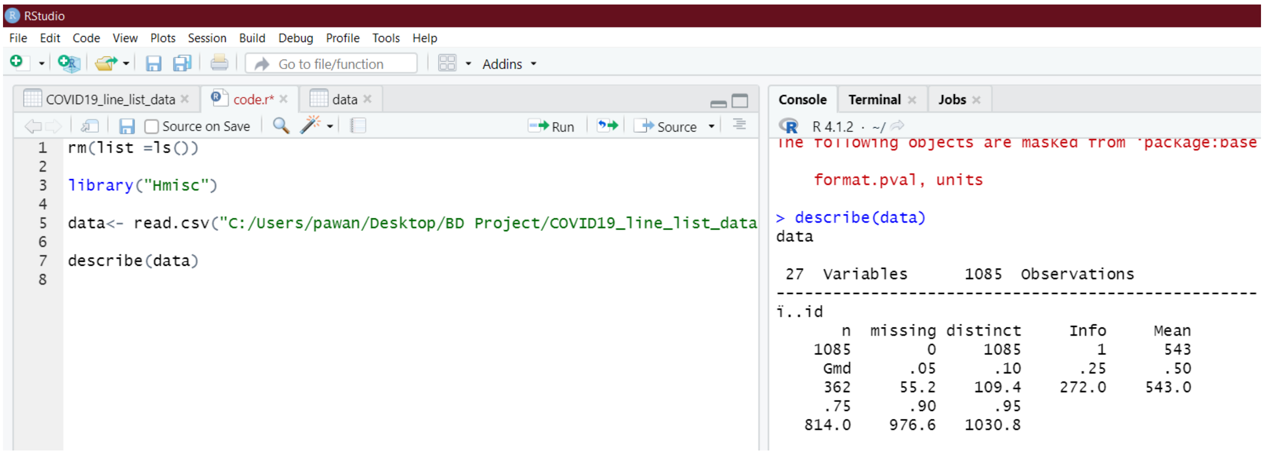
So these we have considered to be determined from the data we have.

Dataset Description:

This dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. The dataset is taken from johns Hopkins university dashboard on Kaggle. The dataset has attributes like age, gender, date got effected, date tested positive, date tested negative is alive or dead.

Implementation:

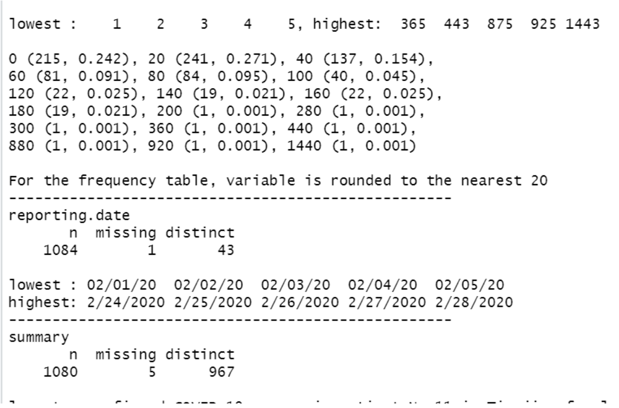
We have first imported the dataset to the R studio, the dataset file was a CSV file and then we just observed for dataset what are the attributes and so then I have used DESCRIBE function to dataset to get description of the dataset by the r studio.



```
rm(list =ls())

library("Hmisc")

data<- read.csv("C:/Users/pawan/Desktop/BD Project/COVID19_line_list_data

describe(data)
```

```
The following objects are masked from 'package:base

    format.pval, units

> describe(data)
data

 27  Variables      1085  Observations
---------------------------------------------------
ï..id
       n  missing distinct     Info     Mean
    1085        0     1085        1      543
     Gmd      .05      .10      .25      .50
     362     55.2    109.4    272.0    543.0
     .75      .90      .95
   814.0    976.6   1030.8
```

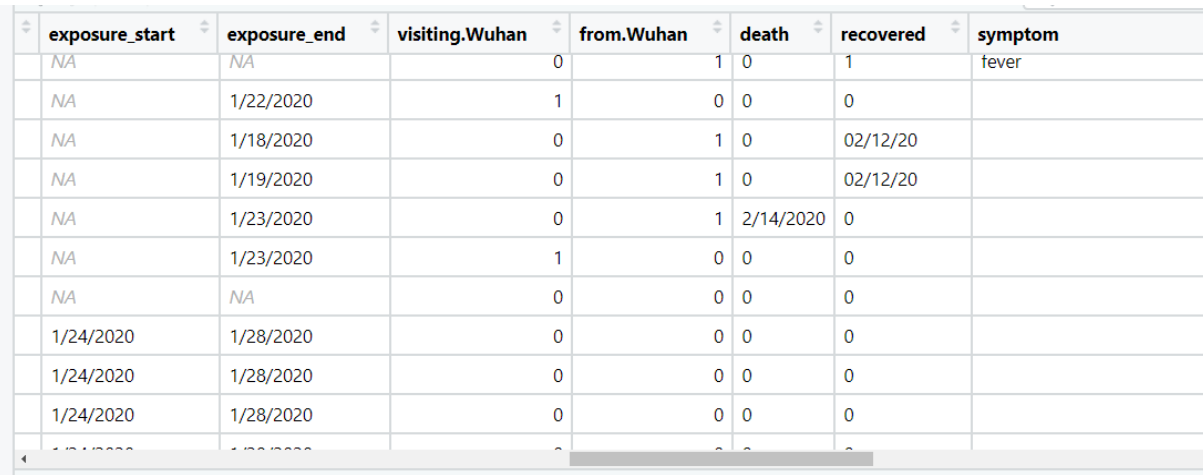
So we can see there were total 27 attributes and 1085 entries of covid cases.



```
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
---------------------------------------------------
summary
       n  missing distinct
    1080        5      967
```

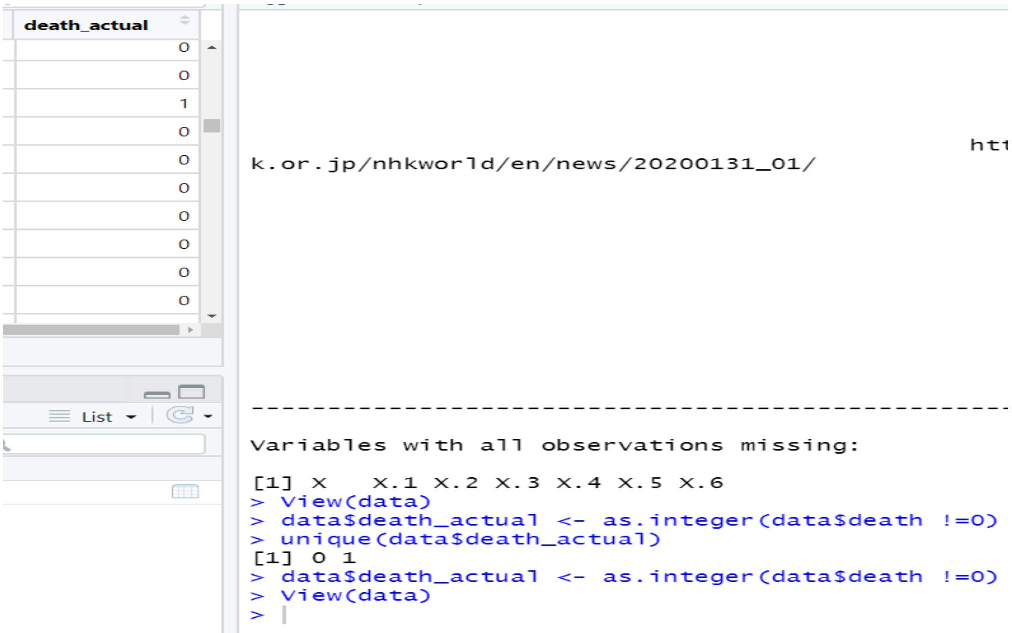
The describe function also return the total null or missing values present in the data set.

So now to find out the death rate of the covid 19 we took the total number of deaths by total number of effected people so while observing the data set in the death attribute we found that if the person is dead then it is 1 if the person is alive its 0 and some entries we found that some values has date of death than 1 or 0.



exposure_start	exposure_end	visiting.Wuhan	from.Wuhan	death	recovered	symptom
NA	NA	0	1	0	1	fever
NA	1/22/2020	1	0	0	0	
NA	1/18/2020	0	1	0	02/12/20	
NA	1/19/2020	0	1	0	02/12/20	
NA	1/23/2020	0	1	2/14/2020	0	
NA	1/23/2020	1	0	0	0	
NA	NA	0	0	0	0	
1/24/2020	1/28/2020	0	0	0	0	
1/24/2020	1/28/2020	0	0	0	0	
1/24/2020	1/28/2020	0	0	0	0	

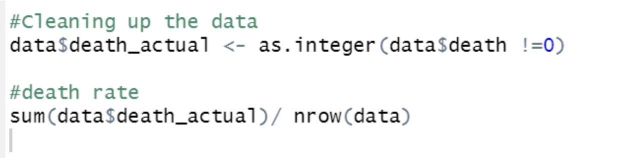
So to remove those we created a new attribute actual deaths and then if the 0 is not present we considered the person is dead and if 0 is present the person is alive by 0 and 1.



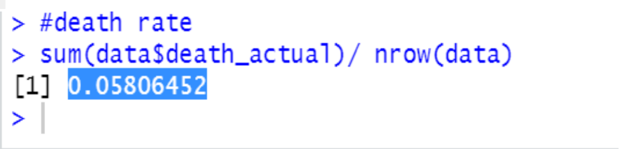
death_actual
0
0
1
0
0
0
0
0
0
0

```
                                                          ht1
k.or.jp/nhkworld/en/news/20200131_01/

----------------------------------------------------------

Variables with all observations missing:

[1] X    X.1 X.2 X.3 X.4 X.5 X.6
> View(data)
> data$death_actual <- as.integer(data$death !=0)
> unique(data$death_actual)
[1] 0 1
> data$death_actual <- as.integer(data$death !=0)
> View(data)
> 
```

The we calculated the death rate

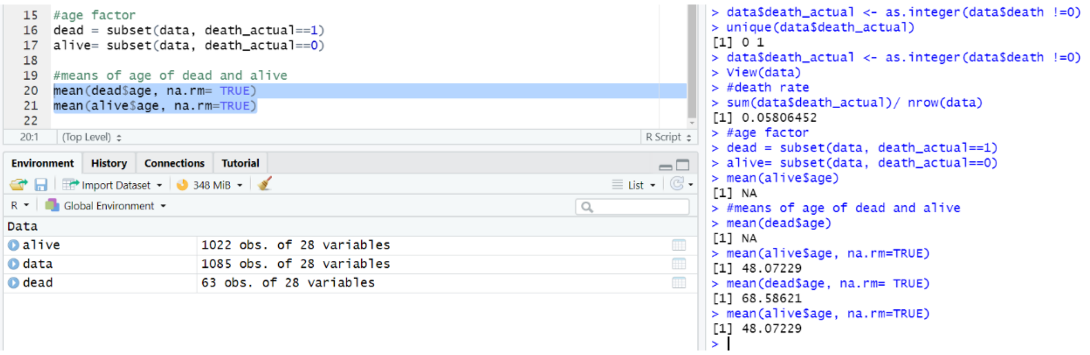


```
#Cleaning up the data
data$death_actual <- as.integer(data$death !=0)

#death rate
sum(data$death_actual)/ nrow(data)
```



```
> #death rate
> sum(data$death_actual)/ nrow(data)
[1] 0.05806452
> 
```

The we verified some assumption that were coming that early time that is older people are more likely to die than the younger once so then we calculated the mean of the age of dead people and mean of the age of the alive people we got the mean of dead people was 68.5 years old but where as the alive mean age was 48.8 years old we can conclude that the older one are more likely to die.

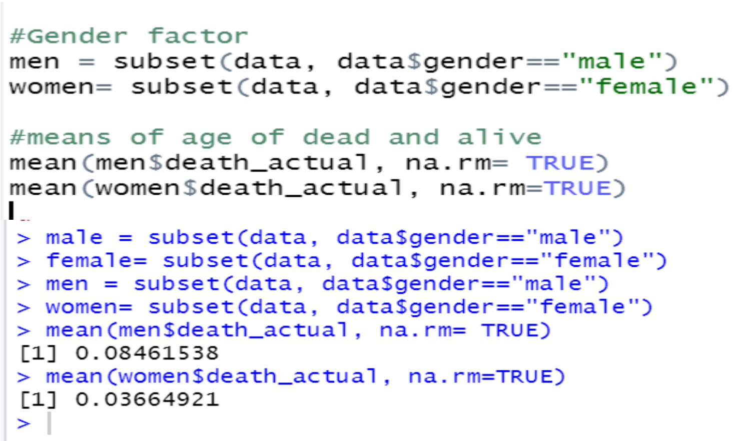


```
#age factor
dead = subset(data, death_actual==1)
alive= subset(data, death_actual==0)

#means of age of dead and alive
mean(dead$age, na.rm= TRUE)
mean(alive$age, na.rm=TRUE)
```

```
Data
alive        1022 obs. of 28 variables
data         1085 obs. of 28 variables
dead         63 obs. of 28 variables
```

```
> data$death_actual <- as.integer(data$death !=0)
> unique(data$death_actual)
[1] 0 1
> data$death_actual <- as.integer(data$death !=0)
> View(data)
> #death rate
> sum(data$death_actual)/ nrow(data)
[1] 0.05806452
> #age factor
> dead = subset(data, death_actual==1)
> alive= subset(data, death_actual==0)
> mean(alive$age)
[1] NA
> #means of age of dead and alive
> mean(dead$age)
[1] NA
> mean(alive$age, na.rm=TRUE)
[1] 48.07229
> mean(dead$age, na.rm= TRUE)
[1] 68.58621
> mean(alive$age, na.rm=TRUE)
[1] 48.07229
> |
```

Then we verified the gender factor that is the male has for fatality rate then the female.



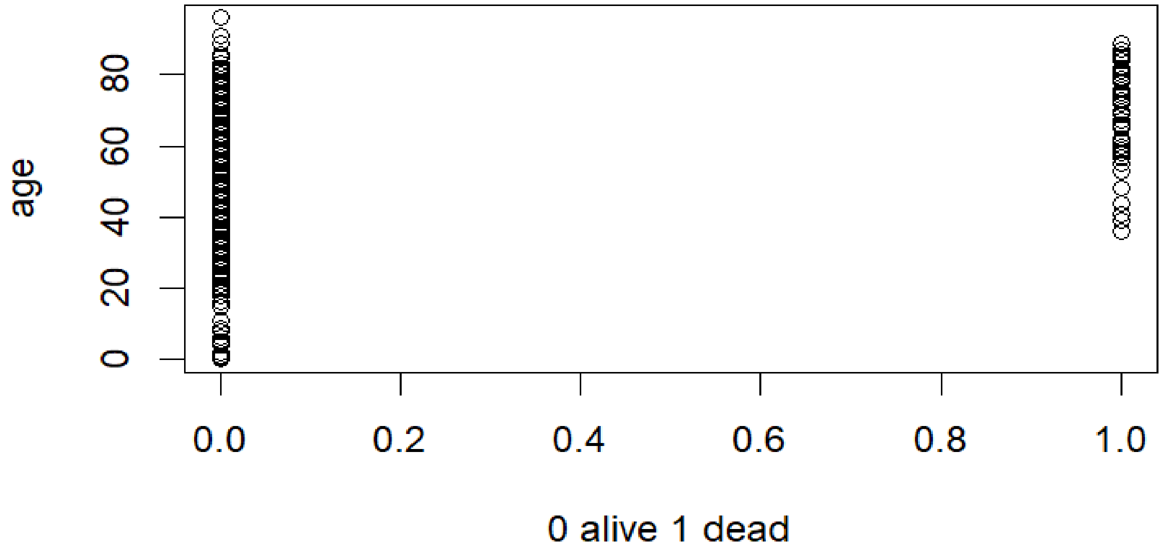
```
#Gender factor
men = subset(data, data$gender=="male")
women= subset(data, data$gender=="female")

#means of age of dead and alive
mean(men$death_actual, na.rm= TRUE)
mean(women$death_actual, na.rm=TRUE)
```

```
> male = subset(data, data$gender=="male")
> female= subset(data, data$gender=="female")
> men = subset(data, data$gender=="male")
> women= subset(data, data$gender=="female")
> mean(men$death_actual, na.rm= TRUE)
[1] 0.08461538
> mean(women$death_actual, na.rm=TRUE)
[1] 0.03664921
> |
```

By same method we calculated the mean of men who were died and mean of women who were died. And we got means as 8.4% for men and 3.6% for women.

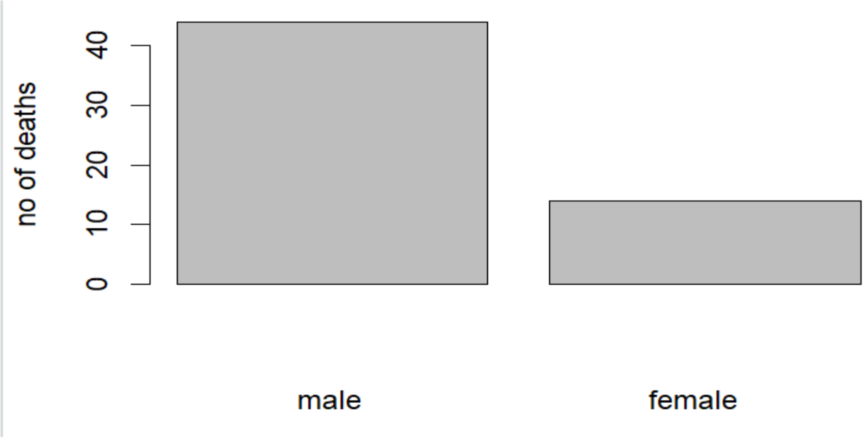
Graphs:

- Graphs for the age factor of died of covid19



We can observe that the more no. of death are on the upper side of the age y axis that means the older are more likely to die.

- Graph for the gender factor who died of covid19



Here we can observe the male death were more than the female death which means men are more likely to die than women.

Conclusion:

In conclusion we can say the death rate of covid 19 was around 5.8% from the data we have analysed.

Then we concluded that the older age people are more likely to die of covid19.

The we also concluded that the men are 4.5% more likely to die than the women these were the conclusion from the analysis of the covid dataset.