STATISTICAL ANALYSIS OF COVID-19 DATA USING R

A Project Report

SUBMITTED IN COMPLETE FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF TECHNOLOGY

IN

[MATHEMATICS AND COMPUTING]

SUBMITTED BY:

PRANAV PUNDEER

2K20/MC/099

UNDER THE SUPERVISION OF

DR. R SRIVASTAVA

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

DELHI TECHNOLOGICAL UNIVERSITY

(FORMERLY Delhi College of Engineering)

Bawana Road, Delhi – 110042

ABSTRACT

This project report will mainly be taking up the following studies:

Using the ***covid19india*** library in rstudio to test that if the onset of the vaccination drive in India has been successful in depreciating the daily cases of covid-19 and the daily deaths caused by the same. We will be observing the data accumulated from the june of 2021 till the current time period.

Also, using the data obtained from a comma separated values (csv) file containing covid cases from 2020 from across the globe , we will test the notion that older people in the population who contracted the virus at a given time, are more likely to die because of it, as compared to the younger members of the same population sample who also acquired the virus during the same period.

We will also address and test the idea that mortality due to the covid-19 infection is higher in males than that in females belonging to the same population sample.

DELHI TECHNOLOGICAL UNIVERSITY

(FORMERLY Delhi College of Engineering)

Bawana Road, Delhi – 110042

ACKNOWLEDGEMENT

We would like to convey our sincerest gratitude and regards to Professor **R Srivastava** for his continuous support and guidance, without which the current project would not have been possible.

INTRODUCTION

Ever since the advent of the covid-19 vaccines, there was hope that the doses would help in controlling the spread of the virus, bringing down the daily case count and would lower the mortality numbers significantly. The general public, ourselves included, always only had a *notion* about this change that the vaccination drive would bring about. There was no statistical evidence, atleast at our levels, that would back our intuition.
But, now that we have enough data that was gathered over the past 10 months, we can actually ***statistically*** prove our intuition; that the vaccines, in real time, had an impact on the case count and the death count in the country. We will be using a **Linear Regression model** to understand the relation between the response variable(daily deaths/daily cases) and the explanatory variable(no. of vaccine doses completed).

Also, during the pandemic, medical professionals consistently indicated that the elderly population of the nation was extremely susceptible to the virus and that this segment would be prone to a higher mortality rate. These claims were biologically supported but lacked a statistical clearance to them. Therefore, using the data we have gathered, we will try to build a relationship between the age of a population sample and the deaths within the population sample to support this claim.

Lastly, medical reports that surfaced during the pandemic suggested that a male who has contracted the novel coronavirus, is more likely to suffer a lethal impact as compared to a female patient who has contracted the virus within the same timeline.

To test the last two hypotheses, we will be using **Student t-test** to disprove the null hypotheses of the two cases respectively.

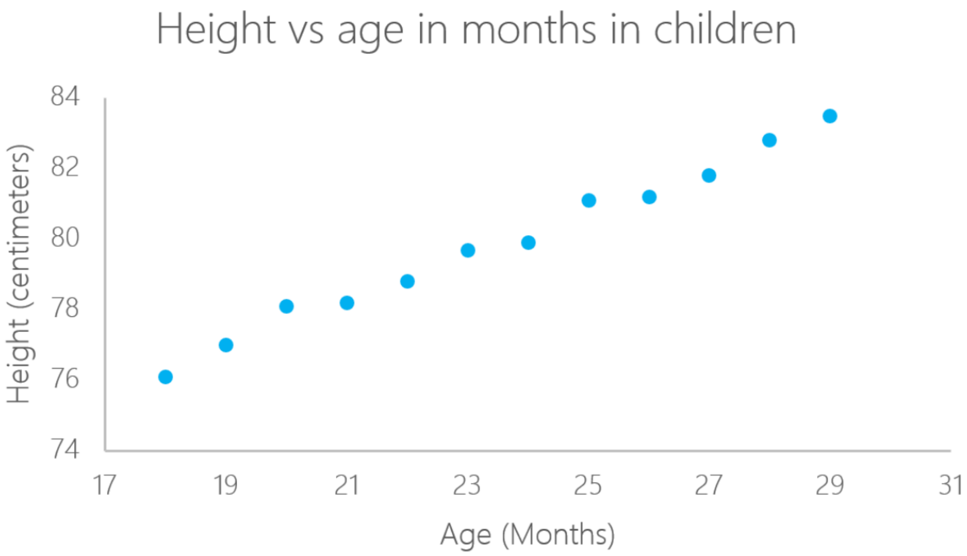
METHODOLOGY AND TECHNIQUES

1) Linear Regression

A Linear Regression is a statistical model that analyzes the relationship between a response variable, often termed as *y* and one or more explanatory variables, often taken as x. In simple terms, this model frames how a certain variable will respond and grow with respect to another independent variable. It is a very general perception that as a person ages, his/her height should increase too, i.e., the older the person is, the taller he/she will be. This is a perfect example of a linear regression which relates two independent variables, height and age, in a linear relationship. In this particular example, you can calculate the height of a child if you know his/her age:

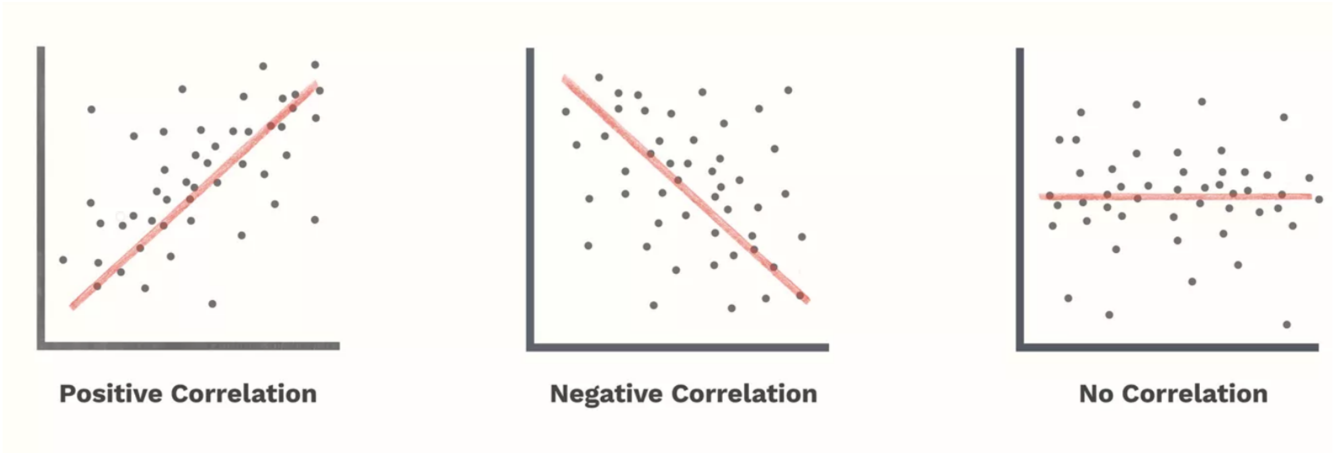
Height=a+Age*b

In this case, "a" and "b" are called the **intercept** and the **slope** respectively. With the same example, "a" or the intercept, is the value from which you start measuring. Newborn babies with zero months are not zero centimeters necessarily; this is the function of the intercept. The slope measures the change of height with respect to the age in months. In general, for every month older the child is, his or her height will increase with "b"**.**



2) *Correlation Coefficient*

Correlation coefficient is a measure that determines the degree to which the movement of 2 independent variables x and y is associated. It determines how strongly x and y are related(linearly), either directly or inversely. The possible range of values for the coefficient is between -1.0 and 1.0 . A correlation coefficient with value -1.0 indicates a **perfect negative correlation** and a value of 1.0 indicates a **perfect positive correlation**. Likewise, a correlation coefficient with value less than 0 indicates a negative relationship and with values greater than 0 indicates a positive relationship. A coefficient of 0 indicates that there exists **no relationship** at all between our two independent variables.



3) *Student t-test for Hypothesis Testing*

A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another. A t-test can only be used when comparing the means of two groups. The t-test assumes that our data:

1. are independent.
2. are (approximately) normally distributed.
3. have a similar amount of variance within each group being compared.

How to decide which t-test should we be using on our data? This decision rests on two factors:

1. Whether the groups being compared come from a single population or two different populations.
2. Whether you want to test the difference in a specific direction.

- If the groups come from the same population, for example, measuring the before and after parameters after a treatment, we use a **paired t-test**.
- If the groups come from 2 different population samples, for example elements from two different species, genders etc., we perform a **two-sample t-test**. Furthermore, if we only care whether the two populations are different from one another irrespective of the direction, we use a **two tailed test**. If, however, we are interested in knowing whether one population mean is greater or lesser than the other, we will use a **one tailed test**.

Interpretation of the above test's result is dependent on 3 main parameters:

1. ***t-value:*** A larger t-value shows that the difference between group means is greater than the pooled standard error, indicating a more significant difference between the groups.

2. ***p-value:*** this helps us in rejecting or accepting our initial null hypothesis. If the resultant p-value is less than the critical p-value of 0.05, we can safely reject the

null hypothesis in favor of the alternative hypothesis. If the p-value is greater than the critical value, we accept the null hypothesis. A p-value of 0.05 is ambiguous and cannot be decided upon.

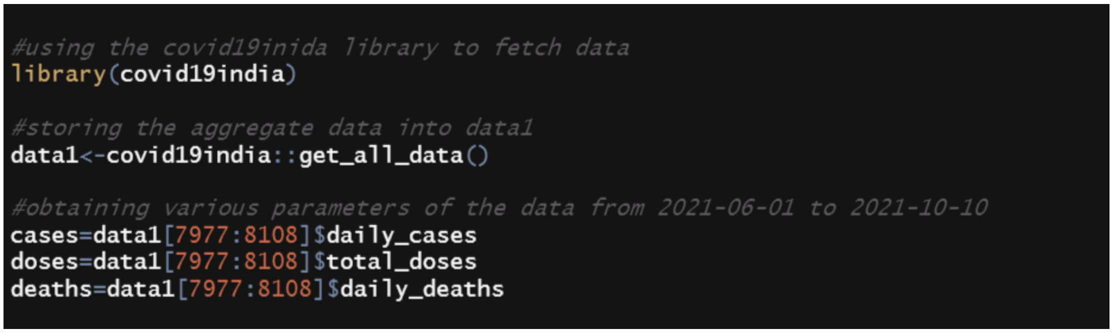
3. ***Degree of freedom:*** Degrees of freedom is related to our sample size, and shows how many 'free' data points are available in our test for making comparisons. The greater the degrees of freedom, the better our statistical test will work.

MODELLING THE PROBLEMS IN RSTUDIO

- As discussed, we will first model the data obtained from the ***covid19inida*** **library** in RStudio to check the effectiveness of the vaccination drive in bringing down the daily covid-19 cases and death count in the country from after June 2021. We will do so by constructing linear regression models for two sets of data: ***daily covid cases vs. total vaccine doses given on that day,*** and ***daily deaths due to covid-19 vs. total vaccine doses given on that day.*** Furthermore, we will find the correlation coefficients for the above data sets. Finally, we interpret the results based on the obtained value of the correlation coefficients, p-values, slopes and intercepts of the linear regression models.

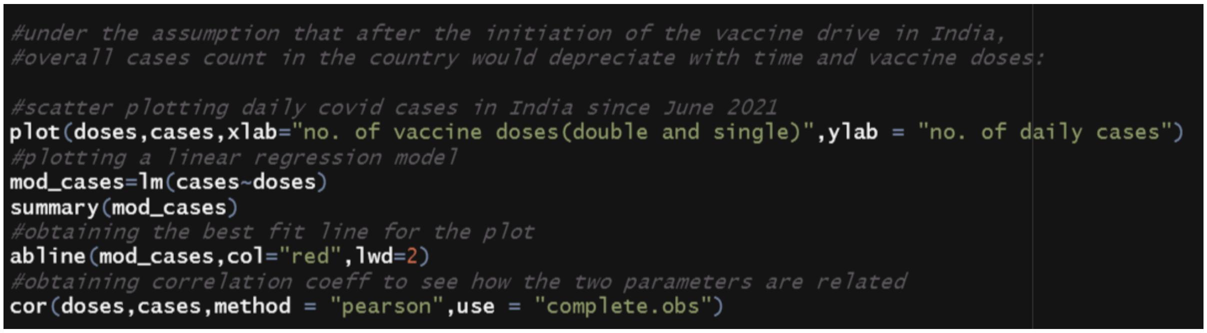
I. DAILY COVID-19 CASES VS DAILY COUNT OF VACCINES ADMINISTERED

1) *obtaining and storing relevant data in variables*

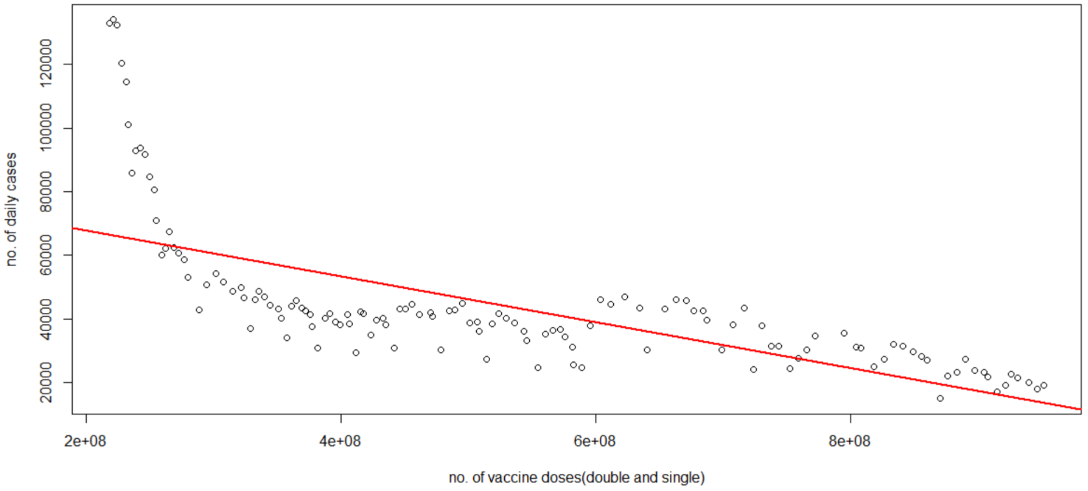


```
#using the covid19inida library to fetch data
library(covid19india)

#storing the aggregate data into data1
data1<-covid19india::get_all_data()

#obtaining various parameters of the data from 2021-06-01 to 2021-10-10
cases=data1[7977:8108]$daily_cases
doses=data1[7977:8108]$total_doses
deaths=data1[7977:8108]$daily_deaths
```

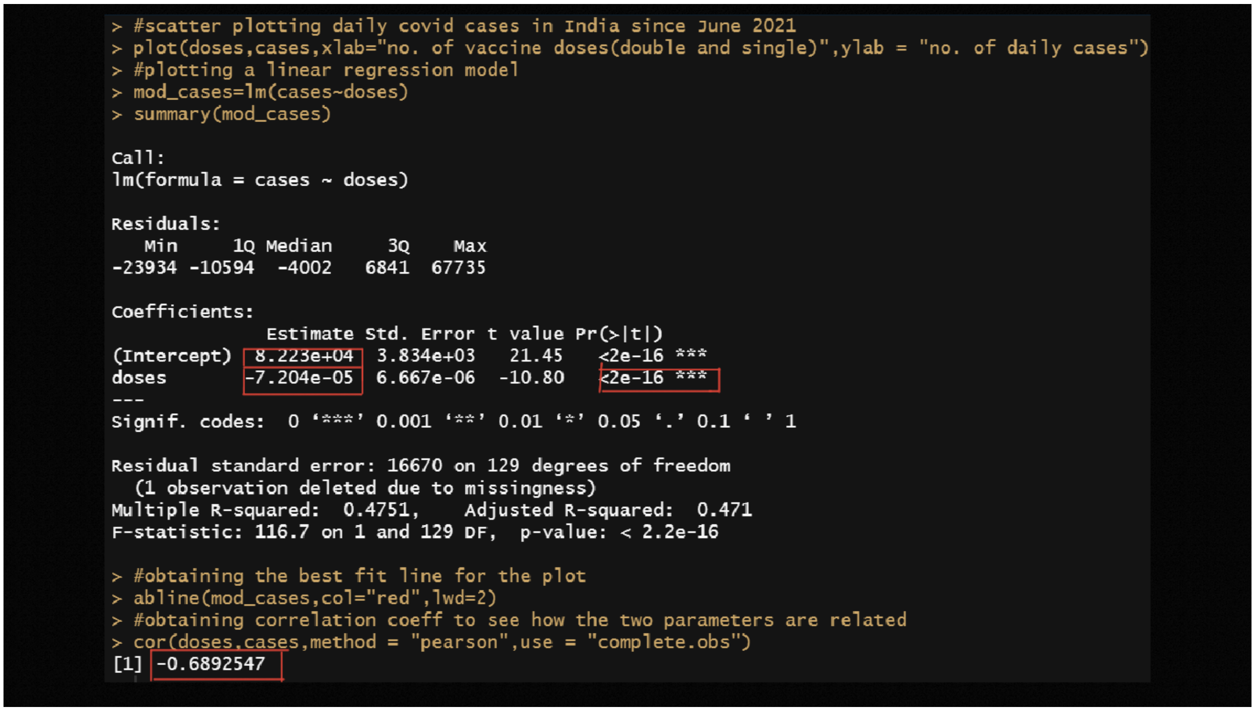
2) *Plotting a linear regression model for no. of daily covid case vs no. of daily vaccine doses administered and finding the correlation coefficient for the same.*



```
#under the assumption that after the initiation of the vaccine drive in India,
#overall cases count in the country would depreciate with time and vaccine doses:

#scatter plotting daily covid cases in India since June 2021
plot(doses,cases,xlab="no. of vaccine doses(double and single)",ylab = "no. of daily cases")
#plotting a linear regression model
mod_cases=lm(cases~doses)
summary(mod_cases)
#obtaining the best fit line for the plot
abline(mod_cases,col="red",lwd=2)
#obtaining correlation coeff to see how the two parameters are related
cor(doses,cases,method = "pearson",use = "complete.obs")
```



3) Interpreting the obtained plot and linear regression summary, and the value for correlation coefficient



```
> #scatter plotting daily covid cases in India since June 2021
> plot(doses,cases,xlab="no. of vaccine doses(double and single)",ylab = "no. of daily cases")
> #plotting a linear regression model
> mod_cases=lm(cases~doses)
> summary(mod_cases)

Call:
lm(formula = cases ~ doses)

Residuals:
   Min     1Q Median     3Q    Max 
-23934 -10594  -4002   6841  67735 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.223e+04  3.834e+03   21.45   <2e-16 ***
doses       -7.204e-05  6.667e-06  -10.80   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16670 on 129 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4751,	Adjusted R-squared:  0.471 
F-statistic: 116.7 on 1 and 129 DF,  p-value: < 2.2e-16

> #obtaining the best fit line for the plot
> abline(mod_cases,col="red",lwd=2)
> #obtaining correlation coeff to see how the two parameters are related
> cor(doses,cases,method = "pearson",use = "complete.obs")
[1] -0.6892547
```

***Intercept value for lm*: 82230**

***Slope value for lm*: - 7.2 *10^-5**

***p-value for lm*: 2*10^-16 (<<0.05)**

Correlation coefficient : -0.6892547

R-squared value: 0.4751

Interpretations:

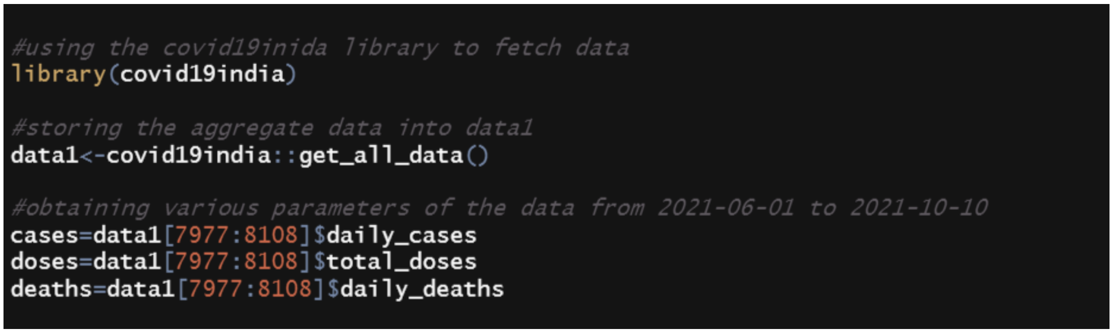
1. The **p-value** suggests that we can reject the null hypothesis which would state that the daily number of vaccines administered is irrelevant and does not affect the number of daily reported cases in the country. Hence, the explanatory variable ***does in fact,*** influences the response variable.

2. The best fitted line on the regression has a **negative slope** which indicates that the number of daily cases depreciates linearly with increasing number of vaccine doses administered. The **R-squared** value is very near to 0.5 which indicates that our model is a good fit for the data.

3. A **negative correlation coefficient** also solidifies the idea that the number of daily cases and the number of vaccine doses administered are inversely related to one another and that as more vaccinations are carried out, less cases of the virus are reported on a daily basis

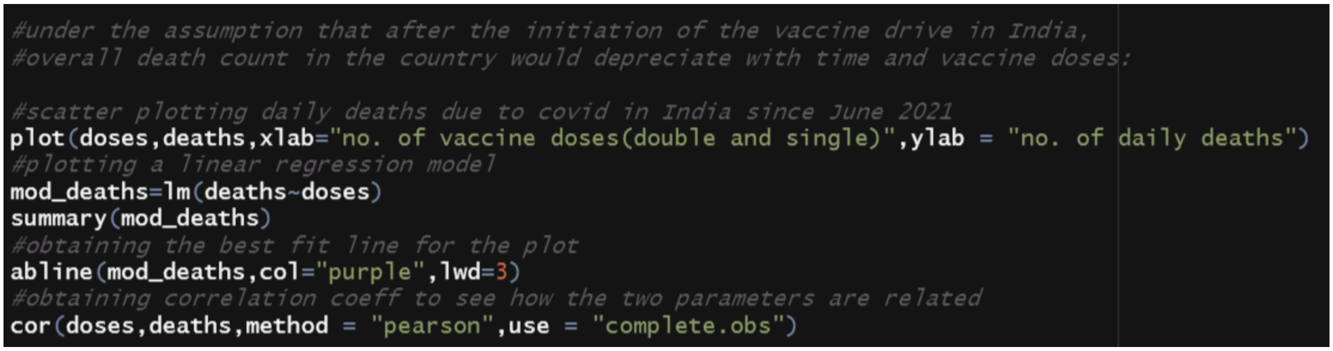
II. DAILY DEATH COUNT VS DAILY COUNT OF VACCINES ADMINISTERED

1) obtaining and storing relevant data in variables

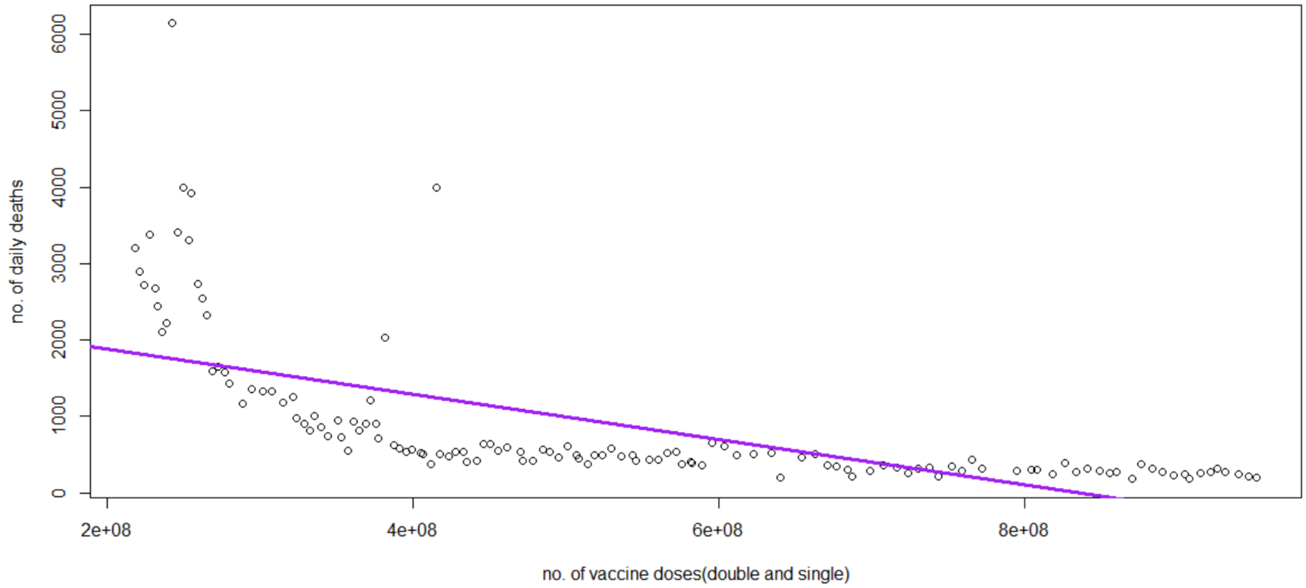


```
#using the covid19inida library to fetch data
library(covid19india)

#storing the aggregate data into data1
data1<-covid19india::get_all_data()

#obtaining various parameters of the data from 2021-06-01 to 2021-10-10
cases=data1[7977:8108]$daily_cases
doses=data1[7977:8108]$total_doses
deaths=data1[7977:8108]$daily_deaths
```

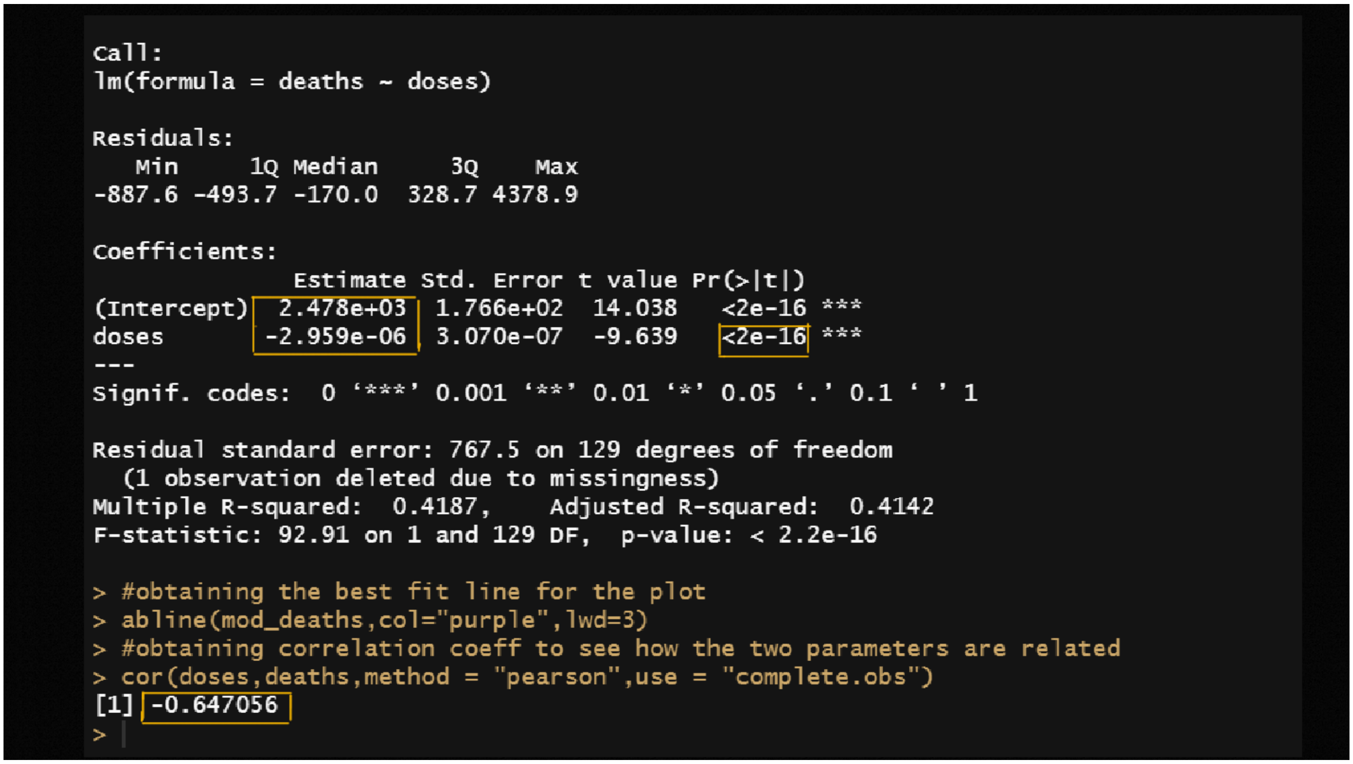
2) plotting a linear regression model for no. of daily deaths due to covid-19 vs the no. of daily vaccine doses administered and finding the correlation coefficient for the same.



```
#under the assumption that after the initiation of the vaccine drive in India,
#overall death count in the country would depreciate with time and vaccine doses:

#scatter plotting daily deaths due to covid in India since June 2021
plot(doses,deaths,xlab="no. of vaccine doses(double and single)",ylab = "no. of daily deaths")
#plotting a linear regression model
mod_deaths=lm(deaths~doses)
summary(mod_deaths)
#obtaining the best fit line for the plot
abline(mod_deaths,col="purple",lwd=3)
#obtaining correlation coeff to see how the two parameters are related
cor(doses,deaths,method = "pearson",use = "complete.obs")
```



3) Interpreting the obtained plot and linear regression summary, and the value for correlation coefficient



```
Call:
lm(formula = deaths ~ doses)

Residuals:
   Min     1Q Median     3Q    Max 
-887.6 -493.7 -170.0  328.7 4378.9 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.478e+03  1.766e+02  14.038   <2e-16 ***
doses       -2.959e-06  3.070e-07  -9.639   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 767.5 on 129 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4187,	Adjusted R-squared:  0.4142 
F-statistic: 92.91 on 1 and 129 DF,  p-value: < 2.2e-16

> #obtaining the best fit line for the plot
> abline(mod_deaths,col="purple",lwd=3)
> #obtaining correlation coeff to see how the two parameters are related
> cor(doses,deaths,method = "pearson",use = "complete.obs")
[1] -0.647056
> 
```

***Intercept value for lm*: 2478**

***Slope value for lm*: - 2.959 *10^-6**

***p-value for lm*: 2*10^-16 (<<0.05)**

Correlation coefficient : -0.647056

R-squared value: 0.4187

Interpretations:

1. The **p-value** suggests that we can reject the null hypothesis which would state that
2. the daily number of vaccines administered is irrelevant and does not affect the number of reported daily deaths due to the virus in the country. Hence, the explanatory variable ***does in fact,*** influences the response variable.

3. The best fitted line on the regression has a **negative slope** which indicates that the number of deaths per day depreciates linearly with increasing number of vaccine

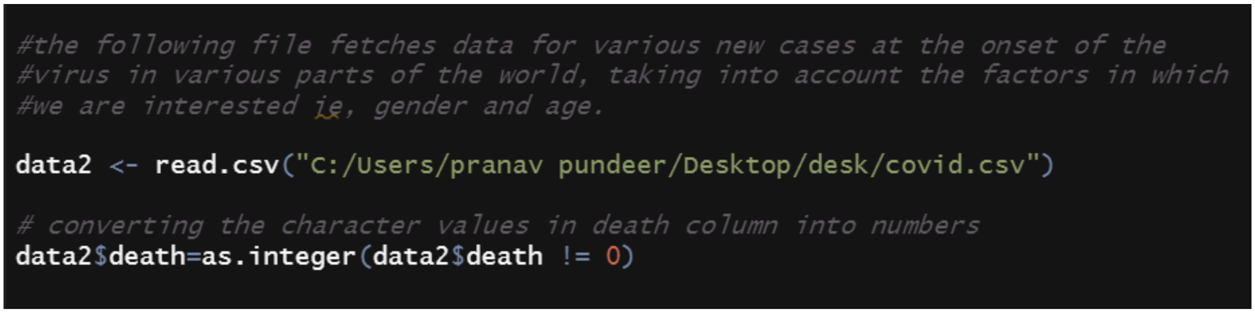
doses administered. The **R-squared** value is 0.4187 which indicates that our model is a near to good fit for the data.

4. A **negative correlation coefficient** also solidifies the idea that the number of daily reported deaths and the number of vaccine doses administered are inversely related to one another and that as more vaccinations are carried out, less people die because of the infection.

- Now, we shall address the notion that the mortality in a population of covid infected people also depends upon factors like ***age*** and the ***gender*** of the person. To perform this, data obtained from a csv file containing the details of covid cases in the year 2020 from around the world, will be modelled using a Student's t-test to disprove the null hypotheses in the respective cases. We will be using a **Two sample , one sided t-test** because our aim is to obtain which among the 2 populations (younger-elder or male-female) has more mean deaths, which gives us a directional aspect.

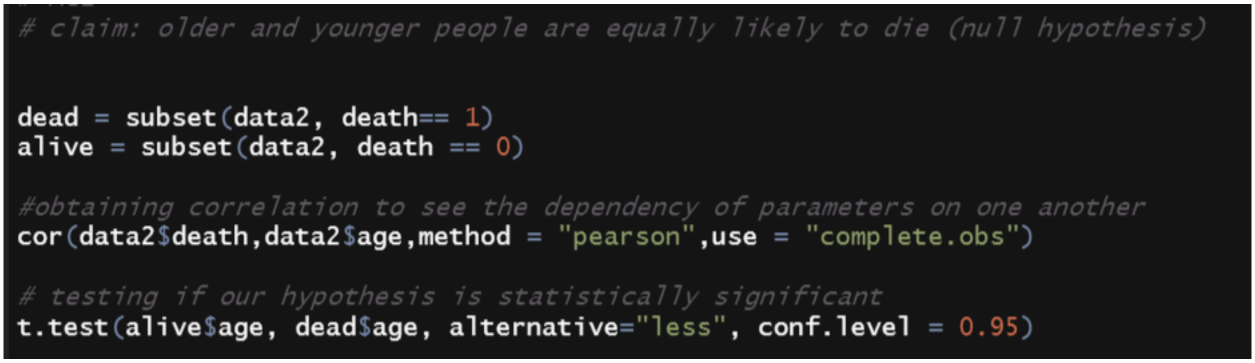
I. COMPARING MORTALITY IN DIFFERENT AGE GROUPS

1) obtaining data from csv file



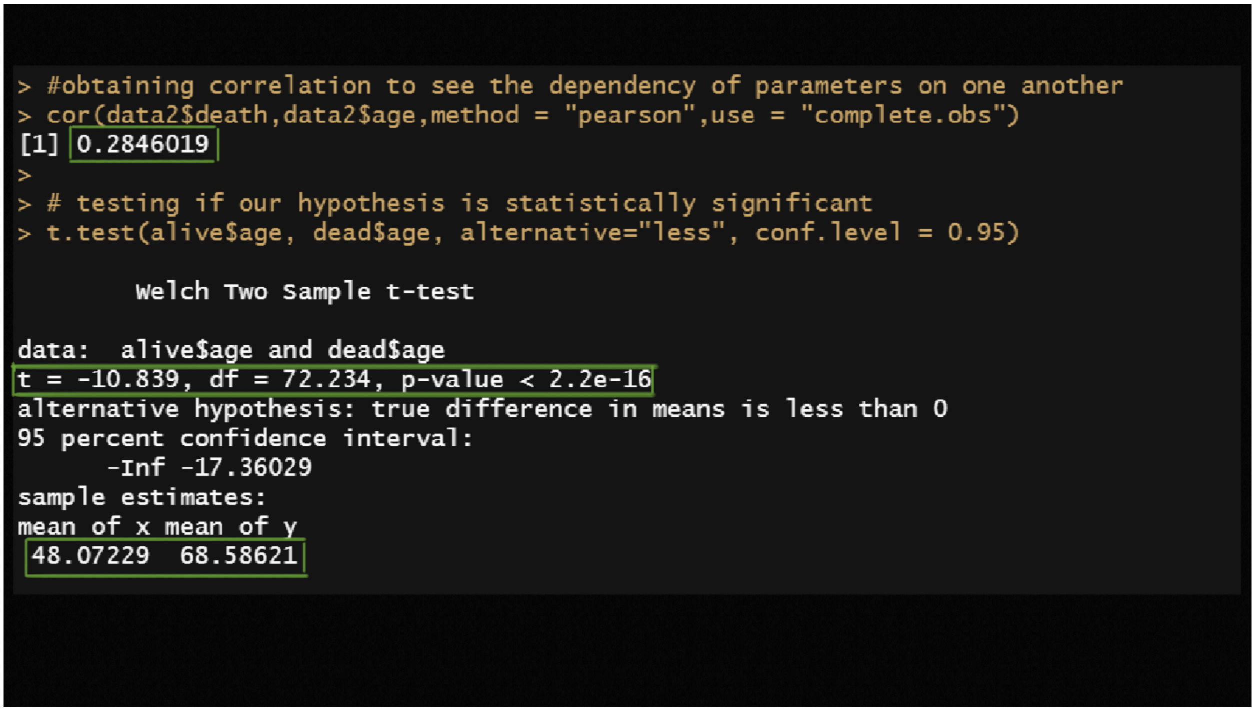
```
#the following file fetches data for various new cases at the onset of the
#virus in various parts of the world, taking into account the factors in which
#we are interested ie, gender and age.

data2 <- read.csv("C:/Users/pranav pundeer/Desktop/desk/covid.csv")

# converting the character values in death column into numbers
data2$death=as.integer(data2$death != 0)
```

2) conducting student's t-test on the data for age and death in the given sample



```
# claim: older and younger people are equally likely to die (null hypothesis)


dead = subset(data2, death== 1)
alive = subset(data2, death == 0)

#obtaining correlation to see the dependency of parameters on one another
cor(data2$death,data2$age,method = "pearson",use = "complete.obs")

# testing if our hypothesis is statistically significant
t.test(alive$age, dead$age, alternative="less", conf.level = 0.95)
```

Here, we used the **two sample t-test** to prove our hypothesis that the number of deaths in the young population is less than the number of deaths in the elderly population, over the null hypothesis which claims that both the age groups are equally likely to die due to the infection. The result obtained is as:



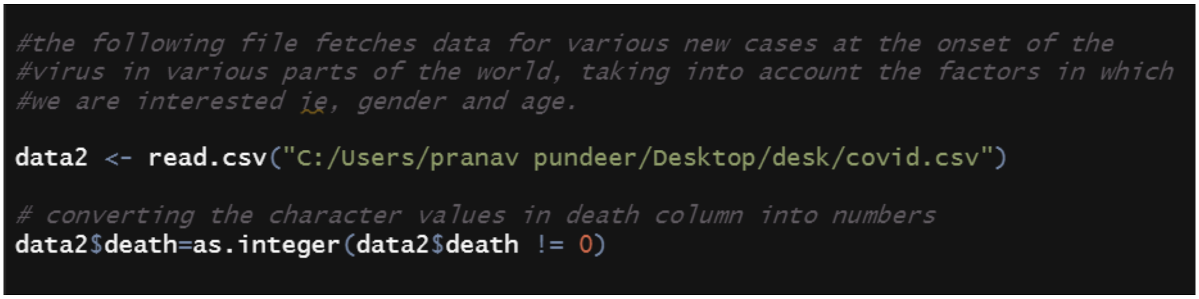
```
> #obtaining correlation to see the dependency of parameters on one another
> cor(data2$death,data2$age,method = "pearson",use = "complete.obs")
[1] 0.2846019
>
> # testing if our hypothesis is statistically significant
> t.test(alive$age, dead$age, alternative="less", conf.level = 0.95)

        Welch Two Sample t-test

data:  alive$age and dead$age
t = -10.839, df = 72.234, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -17.36029
sample estimates:
mean of x mean of y
 48.07229  68.58621
```

3) *interpreting the test values and result*

1. The first thing to observe is that the correlation coefficient is **0.2846,** which implies that there is a **direct relation** between the age and the respective number of deaths associated with that age. As the age increases, the number of deaths for that particular age in the sample also increases.
2. A larger t-value shows that the difference between group means is greater than the pooled standard error, indicating a more **significant difference between the groups**.
3. Also, the true difference between the means will be at least **17.3** years 95% of the time. This means that the mean age of the people who survive will be **17.3** years less than the mean age of the people who die due the infection, 95% of the time.
4. A degrees of freedom value of **72.234** indicates that our data has enough free data points for making comparisons so that our test runs just fine.
5. The p-value of **2.2* 10^-16** which is very less than the critical value of 0.05 indicates that we can safely reject the null hypothesis. Therefore, our alternative hypothesis, that younger people are less likely to die due to the infection as compared to the elders, stands true.

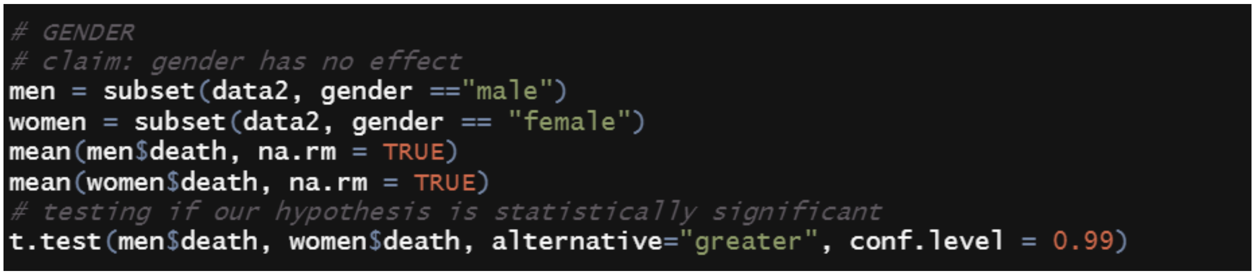
II. COMPARING MORTALITY IN MALES AND FEMALES

1) *obtaining data from csv file*



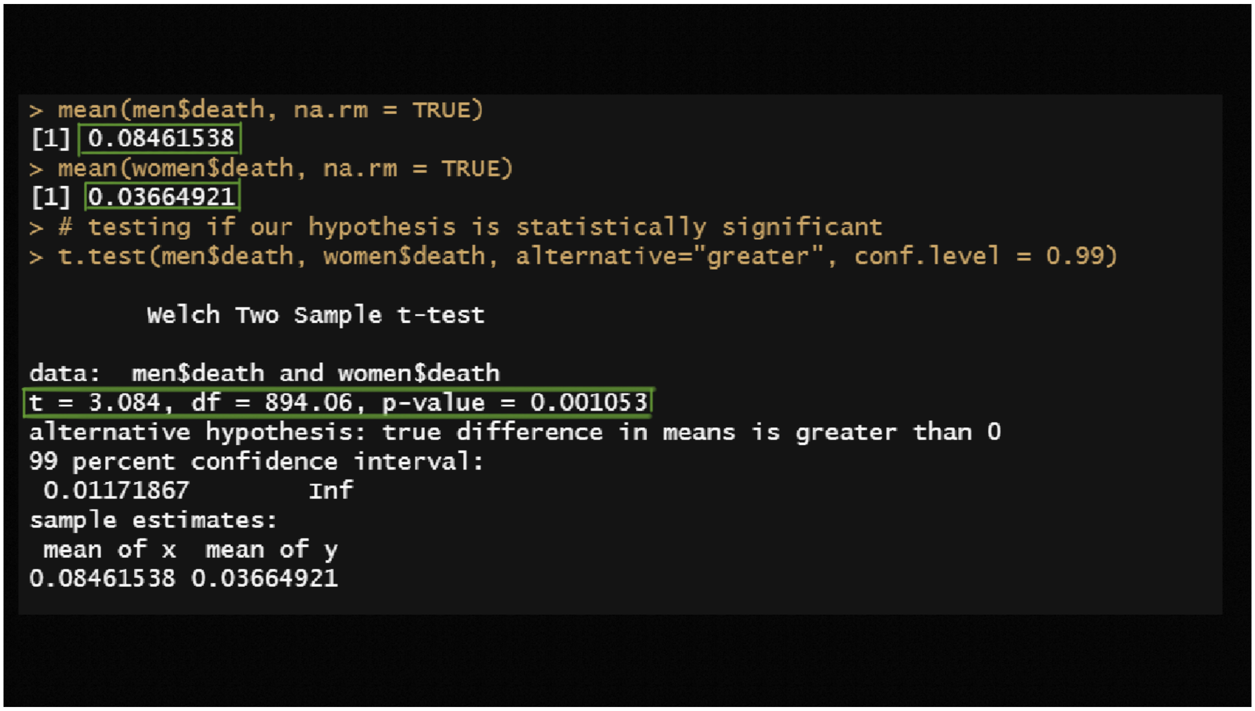
```
#the following file fetches data for various new cases at the onset of the
#virus in various parts of the world, taking into account the factors in which
#we are interested ie, gender and age.

data2 <- read.csv("C:/Users/pranav pundeer/Desktop/desk/covid.csv")

# converting the character values in death column into numbers
data2$death=as.integer(data2$death != 0)
```

2) *conducting student's t-test on the data for age and death in the given sample*



```
# GENDER
# claim: gender has no effect
men = subset(data2, gender =="male")
women = subset(data2, gender == "female")
mean(men$death, na.rm = TRUE)
mean(women$death, na.rm = TRUE)
# testing if our hypothesis is statistically significant
t.test(men$death, women$death, alternative="greater", conf.level = 0.99)
```

Here, we used the **two sample t-test** to prove our hypothesis that the number of deaths in the male population is greater than the number of deaths in the female population, over the null hypothesis which claims that both the gender groups are equally likely to die due to the infection. The result obtained is as:



```
> mean(men$death, na.rm = TRUE)
[1] 0.08461538
> mean(women$death, na.rm = TRUE)
[1] 0.03664921
> # testing if our hypothesis is statistically significant
> t.test(men$death, women$death, alternative="greater", conf.level = 0.99)

        Welch Two Sample t-test

data:  men$death and women$death
t = 3.084, df = 894.06, p-value = 0.001053
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 0.01171867        Inf
sample estimates:
 mean of x  mean of y
0.08461538 0.03664921
```

3) interpreting the test values and results

1. A t-value of **3.084** shows that the difference in the group means is not *as* significant as we would have liked.
2. However, the p-value of **0.001053** indicates that we can still reject the null hypothesis safely. Therefore, our alternative hypothesis, that males are more likely to die due to the infection as compared to females of the same sample population, stands true.
3. A degrees of freedom value of **894.06** indicates that our data has more than enough free data points for making comparisons so that our test runs just fine.
4. The confidence interval shows that 99% of the time, the difference between the male population and female populations mean deaths will at least be 1.171%. That is, a male is **1.171%** more likely to die due to covid-19 than a female.

CONCLUSION

We would like to conclude our study by finalizing and drawing upon the following findings from the data analysis:

- ***Vaccination drive has significantly helped in the reduction of daily count of covid-19 cases in India.***
- ***Vaccinating masses has also helped in reducing the daily death toll in the nation manifolds.***
- ***An older person is more susceptible to die due to contracting the novel coronavirus when compared to a younger individual.***
- ***Even though not significantly enough, yet, a male who has contracted the virus is more likely to die because of it when compared to a female patient.***

REFERENCES

- http://r-statistics.co/Statistical-Tests-in-R
- data-flair.training/blogs/hypothesis-testing-in-r/
- https://www.scribbr.com/
- https://stackoverflow.com/questions/23050928/error-in-plot-new-figure-margins-too-large-scatter-plot
- https://www.investopedia.com/
- https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/
- https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/