Take Home Exam

De, Abhijit

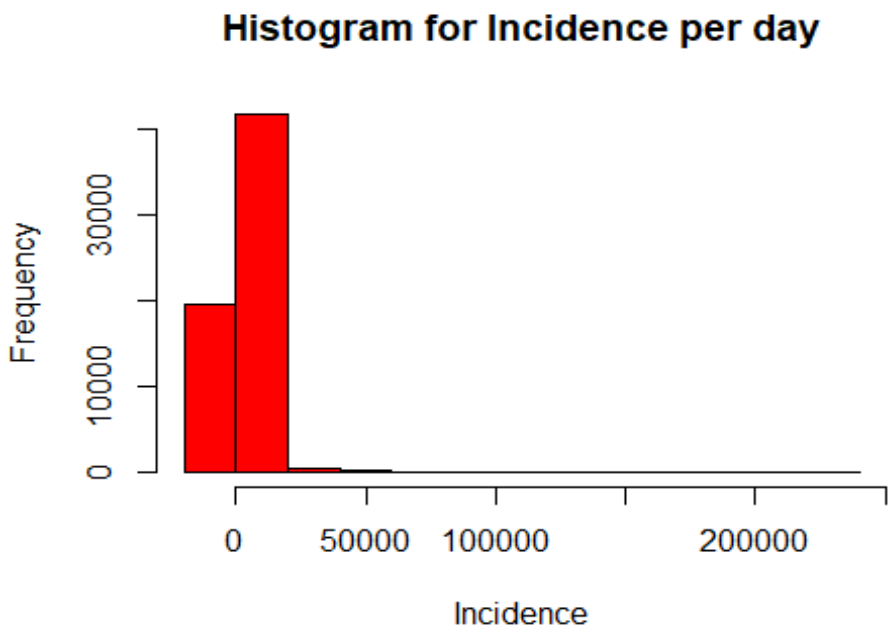
Data is collected for COVID 19 from the given link - https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide for analysis.

```
#Loading the packages.

library(gridExtra)
library(knitr)
library(tinytex)
library(forecast)

#reading the downloaded covid data from the link given, assigning the data fr
ame to "covid_data".

getwd()

## [1] "C:/Users/JM933JS/Downloads/MSDS/Stat/r_code/MSDS401"

setwd("C:/Users/JM933JS/Downloads/MSDS/Stat/r_code/MSDS401")

covid_data <- read.csv("data.csv")


# c) Using str() function to verify the structure of "covid_data", it has 619
00 observations and 12 variables.

str(covid_data)

## 'data.frame':    61900 obs. of  12 variables:
##  $ dateRep                                                    : chr  "14/12
/2020" "13/12/2020" "12/12/2020" "11/12/2020" ...
##  $ day                                                        : int  14 13
12 11 10 9 8 7 6 5 ...
##  $ month                                                      : int  12 12
12 12 12 12 12 12 12 12 ...
##  $ year                                                       : int  2020 2
020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ cases                                                      : int  746 29
8 113 63 202 135 200 210 234 235 ...
##  $ deaths                                                     : int  6 9 11
10 16 13 6 26 10 18 ...
##  $ countriesAndTerritories                                    : chr  "Afgha
```

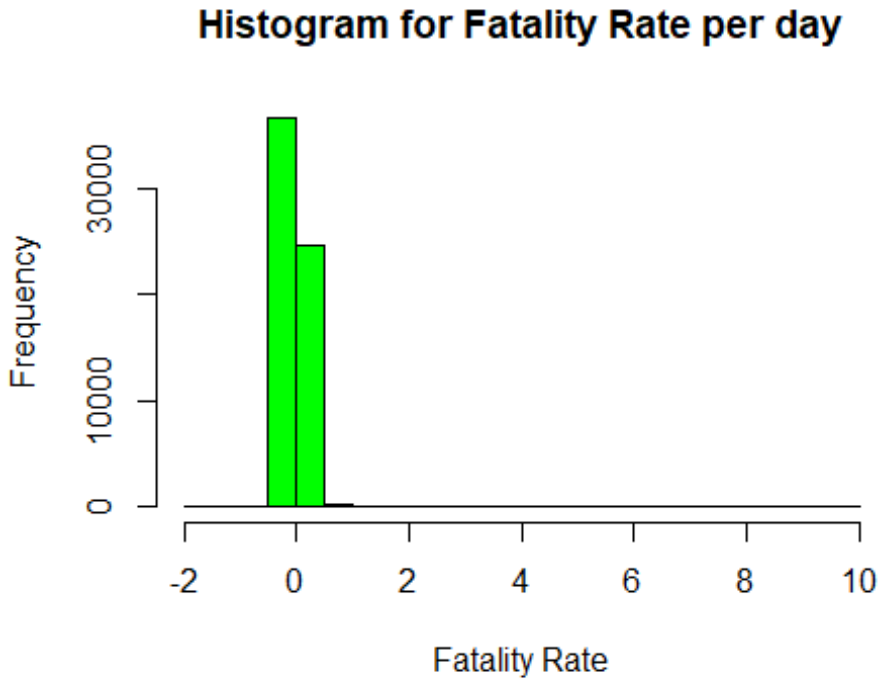
```
nistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ geoId                                                    : chr  "AF" "
AF" "AF" "AF" ...
##  $ countryterritoryCode                                     : chr  "AFG"
"AFG" "AFG" "AFG" ...
##  $ popData2019                                              : int  380417
57 38041757 38041757 38041757 38041757 38041757 38041757 38041757 38041757 38
041757 ...
##  $ continentExp                                             : chr  "Asia"
"Asia" "Asia" "Asia" ...
##  $ Cumulative_number_for_14_days_of_COVID.19_cases_per_100000: num  9.01 7
.05 6.87 7.13 6.97 ...
```

Question 1: Descriptive Statistics: Do an Exploratory Data Analysis (EDA) and provide appropriate summary statistics / visualizations to help understand the spread of the disease (incidence) as well as its fatality rate.

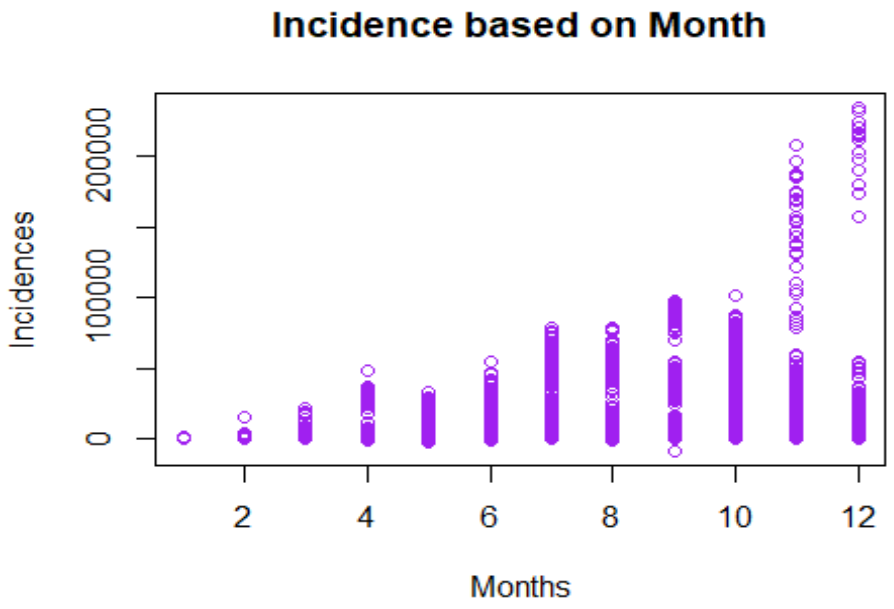
```
#calculating incidence per day which I am assuming is equal to number of case
s per day
incidence <- covid_data$cases
summary(incidence)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -8261       0      15    1155     273  234633

#calculating fatality rate by dividing number of deaths with number of cases
per day
fatality_rate <- covid_data$deaths/covid_data$cases
summary(fatality_rate)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -1.869   0.000   0.008     Inf   0.028     Inf   19167

#there are few records with "NA" values I am converting "NA" values to "0" fo
r calculation
fatality_rate[is.na(fatality_rate)] = 0

#plotting histogram for number of incidence per day
par(mfrow = c(1,1))
hist(incidence,main = "Histogram for Incidence per day",xlab = "Incidence",co
l = "red")
```



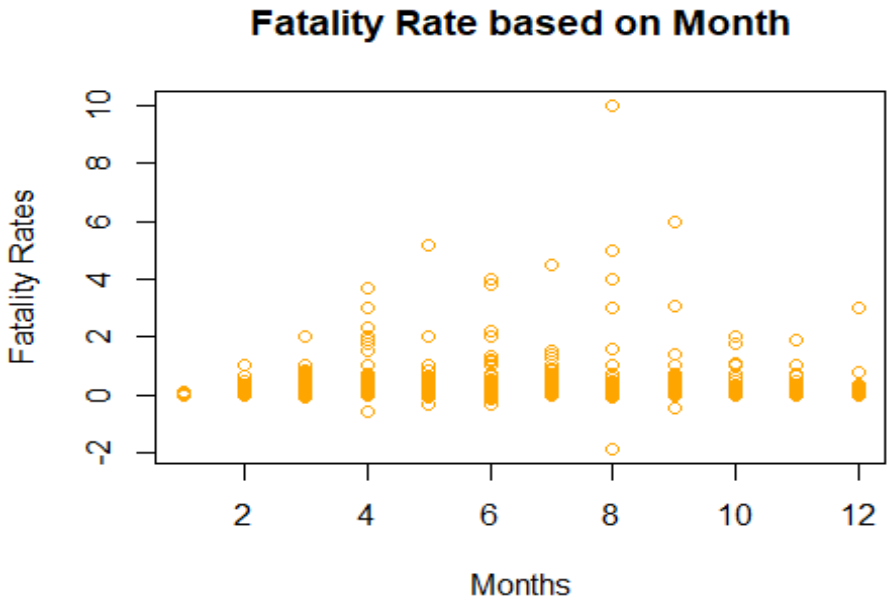

```r
#plotting histogram for fatality rate per day
par(mfrow = c(1,1))
hist(fatality_rate,main = "Histogram for Fatality Rate per day",xlab = "Fatal
ity Rate",col = "green")
```



```r
#plotting scatter plot for number of incidence based on month
par(mfrow = c(1,1))
plot(covid_data$month,incidence,main = "Incidence based on Month",xlab = "Mon
ths",ylab = "Incidences",col = "purple")
```




```r
#plotting scatter plot for fatality rate based on month
par(mfrow = c(1,1))
plot(covid_data$month,fatality_rate,main = "Fatality Rate based on Month",xla
b = "Months",ylab = "Fatality Rates",col = "orange")
```



Discussion for Question 1: Spread of disease, or incidences, can be calculated using cases, assuming 1 case = 1 incidence per day. Fatality rate is the number of deaths per number of cases in a day. Summary statistics is generated, along with histograms of Incidences and Fatality Rates, and scatter plots of Incidences and Fatality Rates by month. Summary statistics provide the minimum, maximum and average number of cases of covid-19 and fatality rate. Histogram shows the concentration of cases and fatality rate. Scatter plot shows the month wise cases and month wise fatality rate of covid-19 disease.

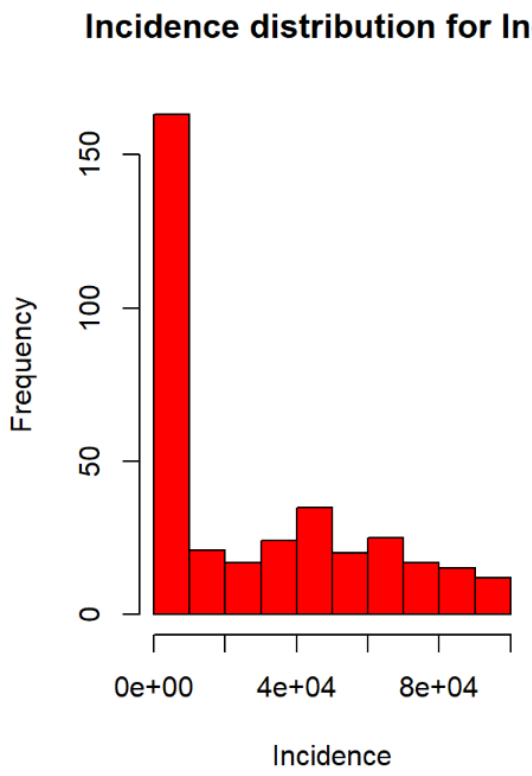
Question 2: Inferential Statistics: Pick 2 countries and compare their incidence and fatality rates using hypothesis testing and confidence interval methods.

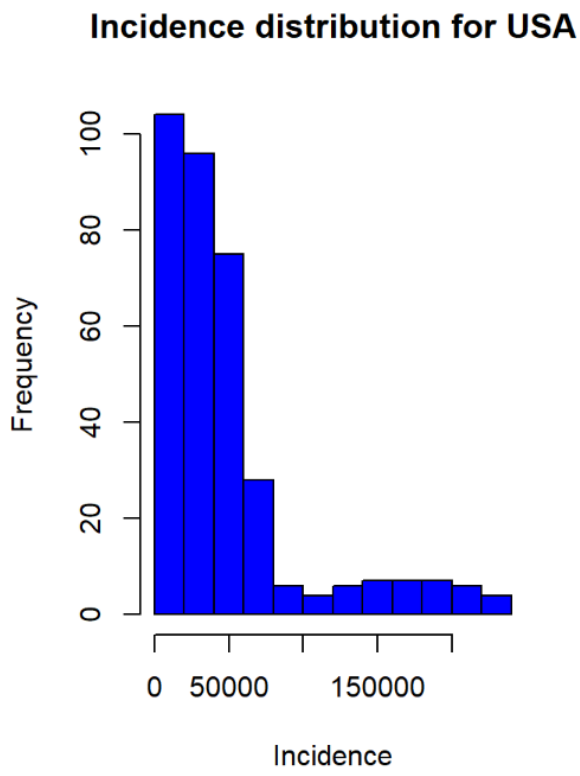
```
#I have taken India and United States of America countries to perform inferen
tial statistics

#Extracting records for India
covid_data_india <- subset(covid_data,covid_data$countriesAndTerritories=="In
dia")
str(covid_data_india)
```

```
## 'data.frame':    349 obs. of  12 variables:
##  $ dateRep                                                    : chr  "14/12
/2020" "13/12/2020" "12/12/2020" "11/12/2020" ...
##  $ day                                                        : int  14 13
12 11 10 9 8 7 6 5 ...
##  $ month                                                      : int  12 12
12 12 12 12 12 12 12 12 ...
##  $ year                                                       : int  2020 2
020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ cases                                                      : int  27071
30254 30006 29398 31521 32080 26567 32981 36011 36652 ...
##  $ deaths                                                     : int  336 39
1 442 414 412 402 385 391 482 512 ...
##  $ countriesAndTerritories                                    : chr  "India
" "India" "India" "India" ...
##  $ geoId                                                      : chr  "IN" "
IN" "IN" "IN" ...
##  $ countryterritoryCode                                       : chr  "IND"
"IND" "IND" "IND" ...
##  $ popData2019                                                : int  136641
7756 1366417756 1366417756 1366417756 1366417756 1366417756 1366417756 136641
7756 1366417756 1366417756 ...
##  $ continentExp                                               : chr  "Asia"
"Asia" "Asia" "Asia" ...
##  $ Cumulative_number_for_14_days_of_COVID.19_cases_per_100000: num  33.1 3
4 34.8 35.6 36.6 ...
```

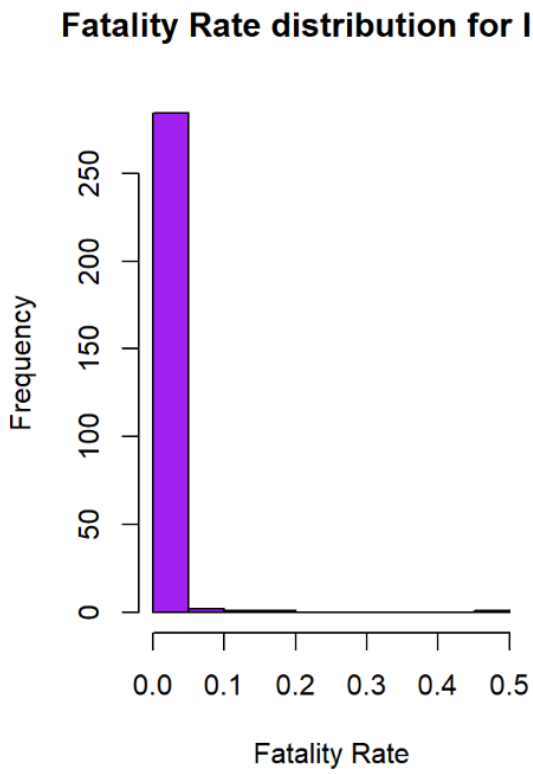
```
#Extracting records for United States of America
covid_data_usa <- subset(covid_data,covid_data$countriesAndTerritories=="Unit
```

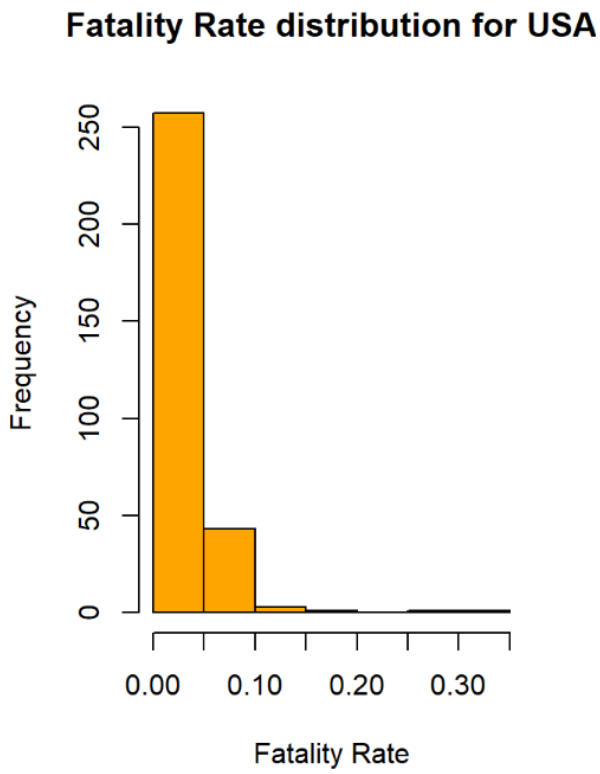
```
ed_States_of_America")
str(covid_data_usa)

## 'data.frame':    350 obs. of  12 variables:
##  $ dateRep                                                  : chr  "14/12
/2020" "13/12/2020" "12/12/2020" "11/12/2020" ...
##  $ day                                                      : int  14 13
12 11 10 9 8 7 6 5 ...
##  $ month                                                    : int  12 12
12 12 12 12 12 12 12 12 ...
##  $ year                                                     : int  2020 2
020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ cases                                                    : int  189723
216017 234633 224680 220025 217344 197334 173432 211933 231930 ...
##  $ deaths                                                   : int  1340 2
315 3343 2748 3124 2564 1433 1111 2203 2680 ...
##  $ countriesAndTerritories                                  : chr  "Unite
d_States_of_America" "United_States_of_America" "United_States_of_America" "U
nited_States_of_America" ...
##  $ geoId                                                    : chr  "US" "
US" "US" "US" ...
##  $ countryterritoryCode                                     : chr  "USA"
"USA" "USA" "USA" ...
##  $ popData2019                                              : int  329064
917 329064917 329064917 329064917 329064917 329064917 329064917 329064917 329
064917 329064917 ...
##  $ continentExp                                             : chr  "Ameri
ca" "America" "America" "America" ...
##  $ Cumulative_number_for_14_days_of_COVID.19_cases_per_100000: num  873 85
7 839 830 794 ...

#calculating incidence per day for India
covid_data_india_incidence <- covid_data_india$cases
summary(covid_data_india_incidence)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     106   15413   28321   50210   97894

#calculating incidence per day for USA
covid_data_usa_incidence <- covid_data_usa$cases
summary(covid_data_usa_incidence)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   16995   33587   46448   56788  234633

#histogram plotting of incidence for both countries
par(mfrow = c(1,2))
hist(covid_data_india_incidence,main = "Incidence distribution for India",xla
b = "Incidence",col = "red")
hist(covid_data_usa_incidence,main = "Incidence distribution for USA",xlab =
"Incidence",col = "blue")
```






```
par(mfrow = c(1,1))

#calculating fatality rate per day for India
covid_data_india_fatality <- covid_data_india$deaths/covid_data_india$cases
summary(covid_data_india_fatality)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00000 0.01243 0.01487 0.02123 0.02469 0.50000      60

#calculating fatality rate per day for USA
covid_data_usa_fatality <- covid_data_usa$deaths/covid_data_usa$cases
summary(covid_data_usa_fatality)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00000 0.01078 0.01722 0.02763 0.03277 0.33333      44

#histogram plotting of fatality rate for both countries
par(mfrow = c(1,2))
hist(covid_data_india_fatality,main = "Fatality Rate distribution for India",
xlab = "Fatality Rate",col = "purple")
hist(covid_data_usa_fatality,main = "Fatality Rate distribution for USA",xlab
= "Fatality Rate",col = "orange")
```






```
par(mfrow = c(1,1))

# using t.test for hypothesis testing and calculation of confidence interval
for incidence of both countries
t.test(covid_data_usa_incidence,covid_data_india_incidence,alternative = "two
.sided",conf.level = 0.95)

##
##  Welch Two Sample t-test
##
## data:  covid_data_usa_incidence and covid_data_india_incidence
## t = 5.6485, df = 561.79, p-value = 2.573e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11823.37 24429.96
## sample estimates:
## mean of x mean of y
##  46447.87  28321.20

# using t.test for hypothesis testing and calculation of confidence interval
for fatality rate of both countries
t.test(covid_data_usa_fatality,covid_data_india_fatality,alternative = "two.s
ided",conf.level = 0.95)

##
##  Welch Two Sample t-test
```

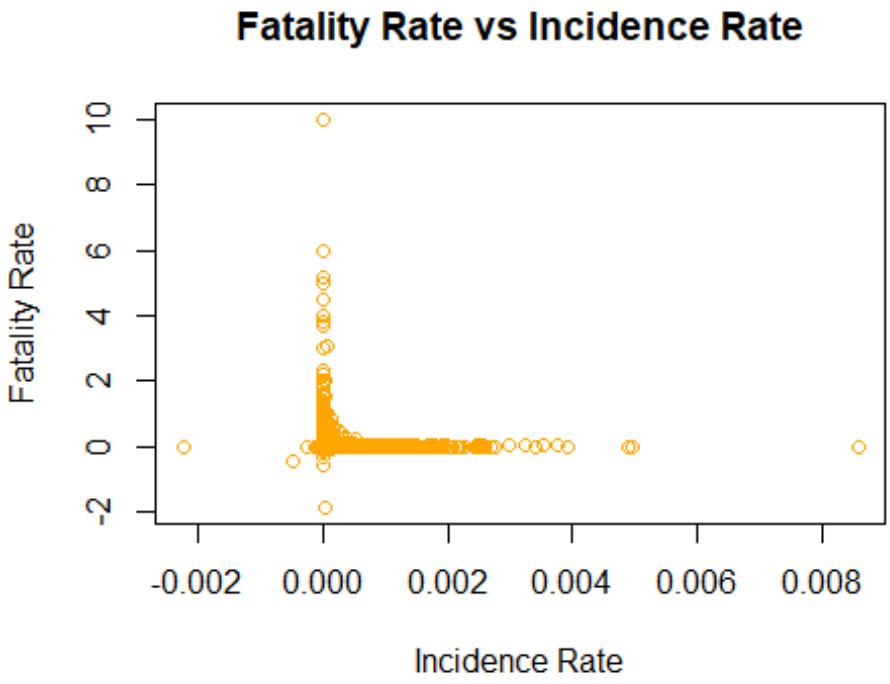
```
##
## data:  covid_data_usa_fatality and covid_data_india_fatality
## t = 2.3714, df = 592.81, p-value = 0.01804
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.001098985 0.011695363
## sample estimates:
##  mean of x  mean of y
## 0.02762786 0.02123069
```

Discussion for Question 2: Incidence t test: p value (2.573e-08) < 0.05 alpha level of significance. So, we can reject the null hypothesis, there is significant difference between the incidence of two countries. The 95% confidence interval is (11823.37, 24429.96). The 95% confidence interval shows that out of 100, 95 times the difference between the mean incidence number of countries lies between the given confidence band.

Fatality Rate t test: p value (0.01804) < 0.05 alpha level of significance. So, we can reject the null hypothesis, there is a significant difference between the fatality rate of two countries. The 95% confidence interval is (0.001098985 0.011695363). The 95% confidence interval shows that out of 100, 95 times the difference between the mean fatality rate of countries lies between the given confidence band.

Question 3: Correlation: Pick all the countries and evaluate the relationship between incidence rates and fatality rates. Compute the correlation coefficient, if relevant.

```
#calculating incidence rate by dividing number of cases with popData2019
incidence_rate <- covid_data$cases/covid_data$popData2019
summary(incidence_rate)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -0.00226  0.00000  0.00000  0.00005  0.00003  0.00859      123

#there are few records with "NA" values I am converting "NA" values to "0" fo
r calculation
incidence_rate[is.na(incidence_rate)] = 0

#plotting scatter plot for incidence rate and fatality rate
par(mfrow = c(1,1))
plot(incidence_rate,fatality_rate,main = "Fatality Rate vs Incidence Rate",xl
ab = "Incidence Rate",ylab = "Fatality Rate",col = "orange")
```




```
#using cor.test for correlation check
cor.test(incidence_rate,fatality_rate,alternative = "two.sided",method = "pea
rson",conf.level = 0.95)

##
##  Pearson's product-moment correlation
##
## data:  incidence_rate and fatality_rate
## t = NaN, df = 61898, p-value = NA
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  NaN NaN
## sample estimates:
## cor
## NaN

#using cor for correlation coefficient calculation
cor(incidence_rate,fatality_rate,method = "pearson")

## [1] NaN
```

Discussion for Question 3: The plot shows that there is no relation between fatality rate and incidence rate. So, correlation coefficient cannot be computed which we checked using correlation test and correlation coefficient is "NaN".

Question 4: Regression: Pick United States. Look at the time series of cases and time series of deaths. Use a regression model to predict the number of cases and the number of deaths for the next 5 days in the future.

```
#Extracting records for United States of America
covid_data_usa <- subset(covid_data,covid_data$countriesAndTerritories=="Unit
ed_States_of_America")
str(covid_data_usa)
```

```
## 'data.frame':    350 obs. of  12 variables:
##  $ dateRep                                                      : chr  "14/12
/2020" "13/12/2020" "12/12/2020" "11/12/2020" ...
##  $ day                                                          : int  14 13
12 11 10 9 8 7 6 5 ...
##  $ month                                                        : int  12 12
12 12 12 12 12 12 12 12 ...
##  $ year                                                         : int  2020 2
020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ cases                                                        : int  189723
216017 234633 224680 220025 217344 197334 173432 211933 231930 ...
##  $ deaths                                                       : int  1340 2
315 3343 2748 3124 2564 1433 1111 2203 2680 ...
##  $ countriesAndTerritories                                      : chr  "Unite
d_States_of_America" "United_States_of_America" "United_States_of_America" "U
nited_States_of_America" ...
##  $ geoId                                                        : chr  "US" "
US" "US" "US" ...
##  $ countryterritoryCode                                         : chr  "USA"
"USA" "USA" "USA" ...
##  $ popData2019                                                  : int  329064
917 329064917 329064917 329064917 329064917 329064917 329064917 329064917 329
064917 329064917 ...
##  $ continentExp                                                 : chr  "Ameri
ca" "America" "America" "America" ...
##  $ Cumulative_number_for_14_days_of_COVID.19_cases_per_100000: num  873 85
7 839 830 794 ...
```

```
#selecting ARIMA model for time series analysis on number of cases for USA
arima_cases_usa <- auto.arima(covid_data_usa$cases)

#forecasting number of cases for next 5 days with 80% and 95% confidence inte
rval including lower and upper interval
forecast_cases_usa <- forecast(arima_cases_usa,h = 5)
print(forecast_cases_usa)
```

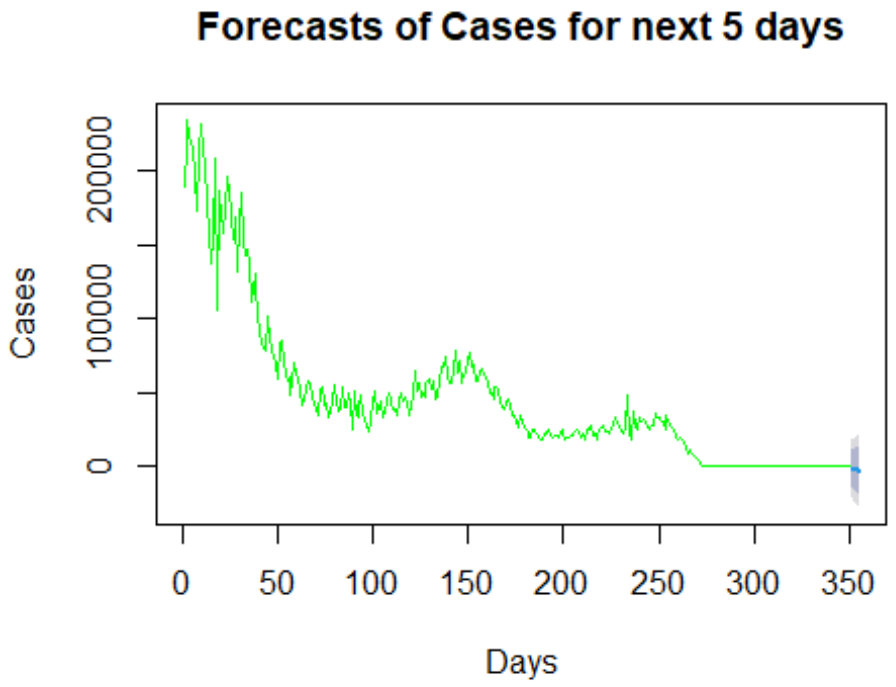
```
##     Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## 351      -1172.371 -13418.30 11073.56 -19900.91 17556.17
## 352      -1719.404 -15098.98 11660.17 -22181.70 18742.89
## 353      -2304.964 -17001.04 12391.11 -24780.68 20170.75
```

```
## 354      -2696.014 -18039.62 12647.59 -26162.03 20770.01
## 355      -3177.786 -19531.23 13175.66 -28188.21 21832.64
```

```
#forecasting number of cases for next 5 days with 95% confidence and upper va
lue
print(forecast_cases_usa$upper[,2])
```

```
## Time Series:
## Start = 351
## End = 355
## Frequency = 1
## [1] 17556.17 18742.89 20170.75 20770.01 21832.64
```

```
#plotting forecasting of cases for next 5 days
par(mfrow = c(1,1))
plot(forecast_cases_usa,main = "Forecasts of Cases for next 5 days",col = "gr
een",xlab = "Days",ylab = "Cases")
```

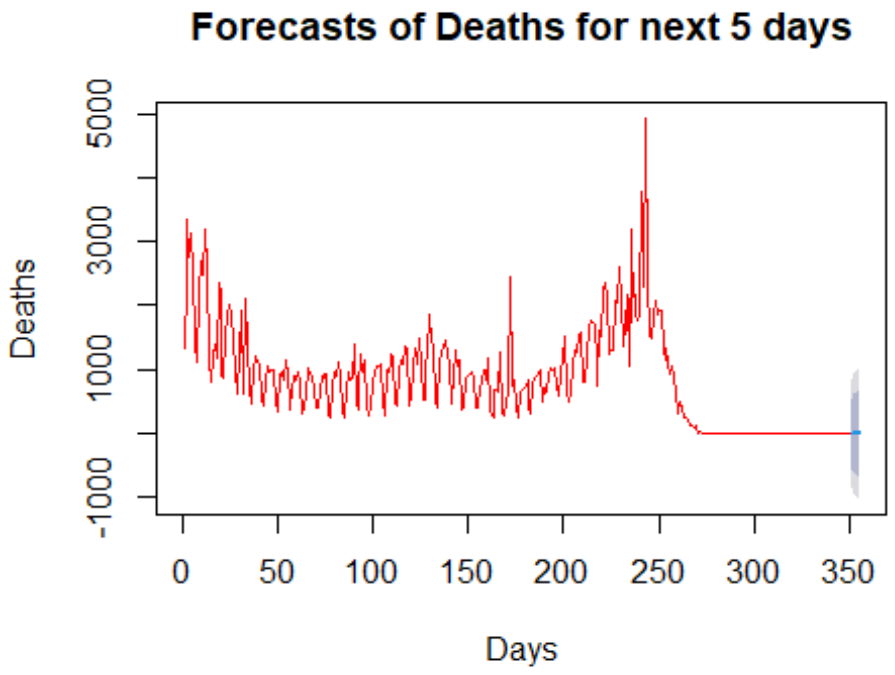



```
#selecting ARIMA model for time series analysis on number of deaths for USA
arima_deaths_usa <- auto.arima(covid_data_usa$deaths)
```

```
#forecasting number of deaths for next 5 days with 80% and 95% confidence int
erval including lower and upper interval
forecast_deaths_usa <- forecast(arima_deaths_usa,h = 5)
print(forecast_deaths_usa)
```

```
##     Point Forecast     Lo 80    Hi 80      Lo 95     Hi 95
## 351   5.343748e-05 -542.7712 542.7713  -830.0970  830.0971
```

```
## 352   7.052427e-05 -605.3609 605.3610  -925.8196  925.8197
## 353   7.598782e-05 -631.6214 631.6216  -965.9816  965.9818
## 354   7.773480e-05 -649.6878 649.6880  -993.6118  993.6120
## 355   7.829341e-05 -665.2750 665.2751 -1017.4503 1017.4505

#forecasting number of deaths for next 5 days with 95% confidence and upper v
alue
print(forecast_deaths_usa$upper[,2])

## Time Series:
## Start = 351
## End = 355
## Frequency = 1
## [1]  830.0971  925.8197  965.9818  993.6120 1017.4505

##plotting forecasting of deaths for next 5 days
par(mfrow = c(1,1))
plot(forecast_deaths_usa,main = "Forecasts of Deaths for next 5 days",col = "
red",xlab = "Days",ylab = "Deaths")
```



Discussion for Question 4: The ARIMA model has been selected to do the Time Series Analysis. This particular model forecasts based on its previous values. We can see from both the graphs, the forecasting for next 5 days and purple line at the end of the graph represent the projected values. Forecasting values for cases and deaths are provided with 80% and 95% confidence bands. As well as generated predicted values for cases and deaths for next 5 days with 95% confidence interval with upper values.