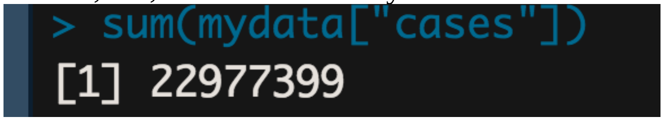
Gurjus Singh

MSDS 401: Applied Stats With R

The Rise of COVID-19

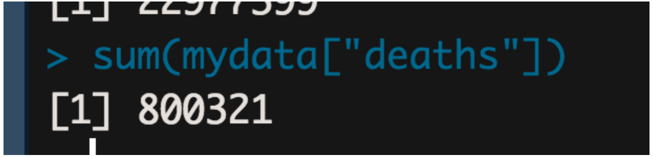
1.)

Some summary statistics I found for the data set are that the sum for all the cases in the dataset was 22,977,399 for the entire year since COVID started.



```
> sum(mydata["cases"])
[1] 22977399
```

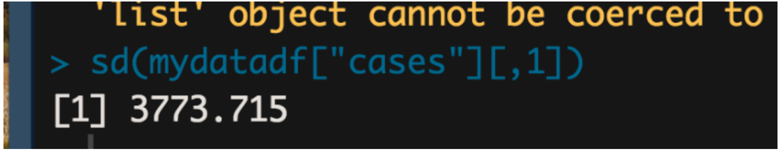
The total deaths were 800,321.



```
> sum(mydata["deaths"])
[1] 800321
```

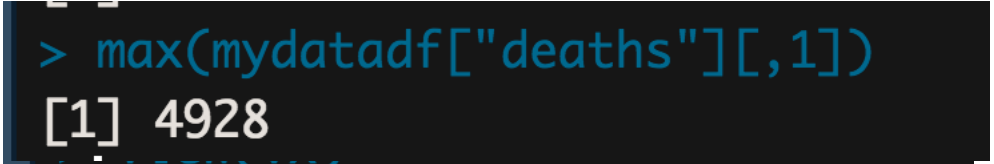
This makes a fatality rate of 800,321/22,977,399 which equals 3.48%.

The standard deviation of the number of cases 3,773.715

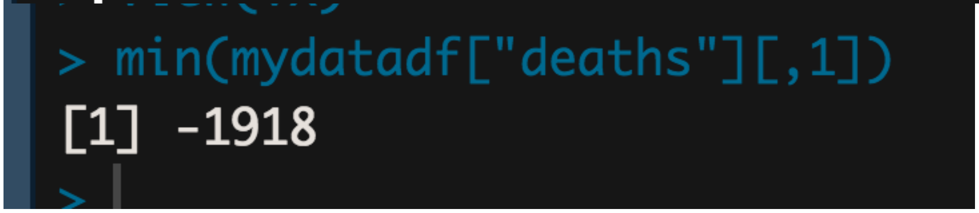


```
  'list' object cannot be coerced to
> sd(mydatadf["cases"][,1])
[1] 3773.715
```

The max number of deaths is 4,928 while the minimum is -1918

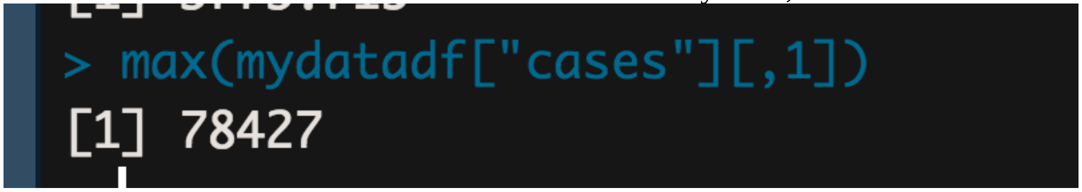


```
> max(mydatadf["deaths"][,1])
[1] 4928
```



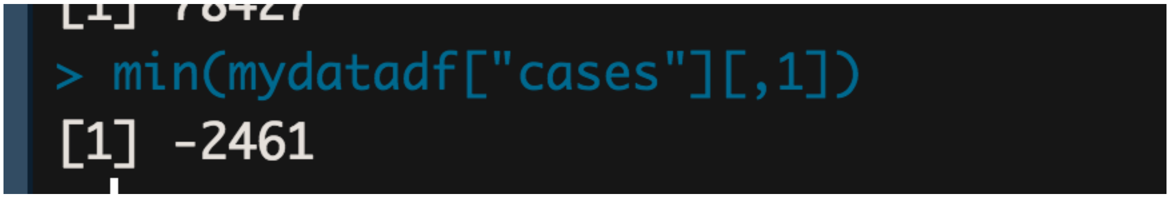
```
> min(mydatadf["deaths"][,1])
[1] -1918
>
```

The max of the dataset of the number of cases for 1 day is 78,427



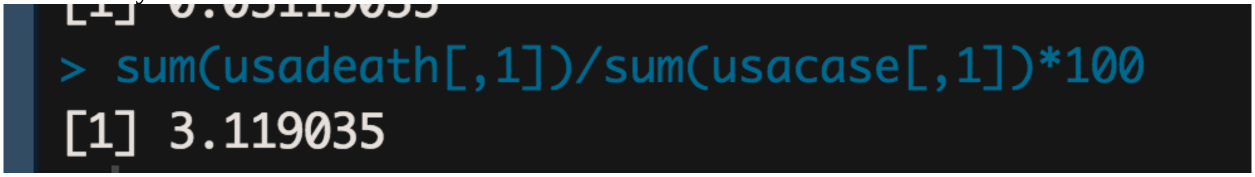
```
> max(mydatadf["cases"][,1])
[1] 78427
```

The minimum of the cases per day is -2461



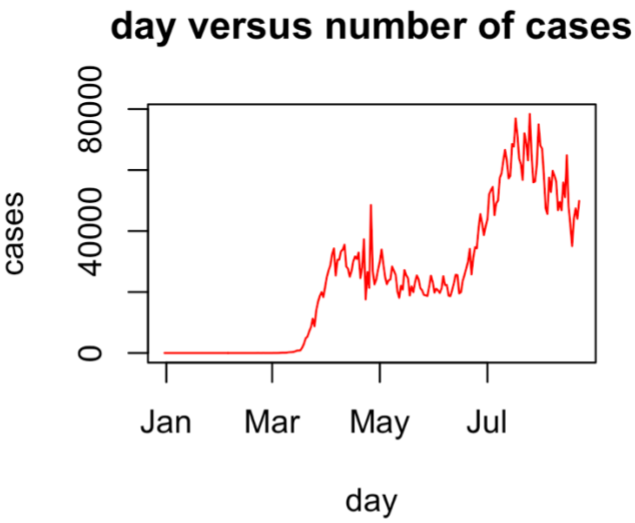
```
[1] 78427
> min(mydatadf["cases"][,1])
[1] -2461
```

The fatality rate for the USA is at 3.11% that is deaths over total cases makes a 3% death rate.

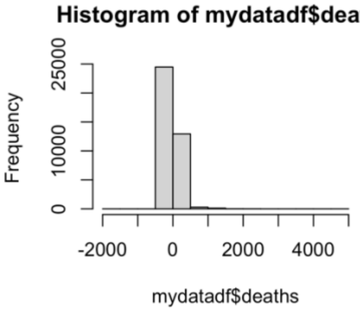


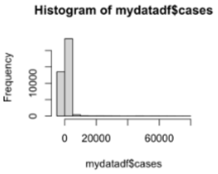
```
[1] 0.03119035
> sum(usadeath[,1])/sum(usacase[,1])*100
[1] 3.119035
```

In my first data visualization I show a line graph of the cases per USA starting in January and going till August. As you can see the number of cases rise and now starting to ease out as we enter August.



Here are histograms for deaths and cases for each country over all the days from January – August

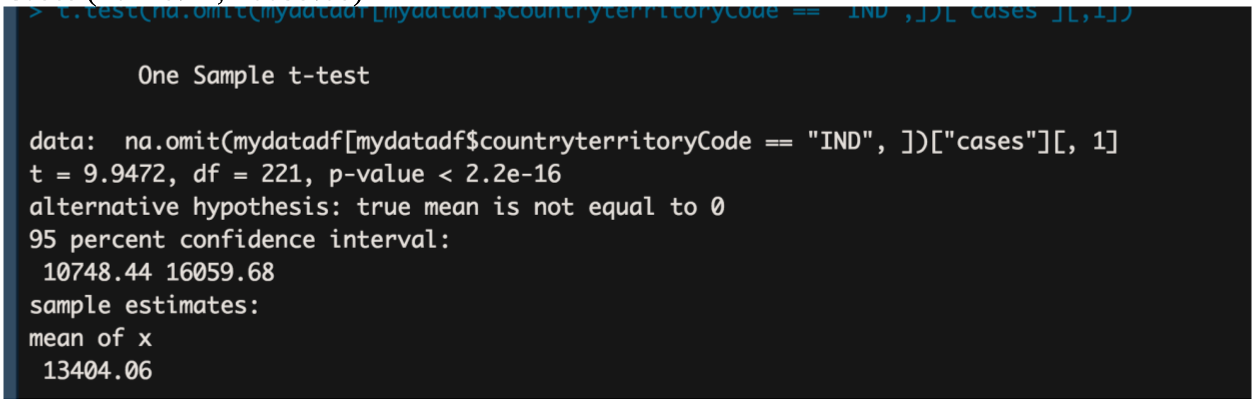




2.)
Two countries I got the 95% confidence interval for using t.test are India and USA cases and deaths.

For India the confidence interval deaths and cases are below:

Cases (10748.44, 16059.68)

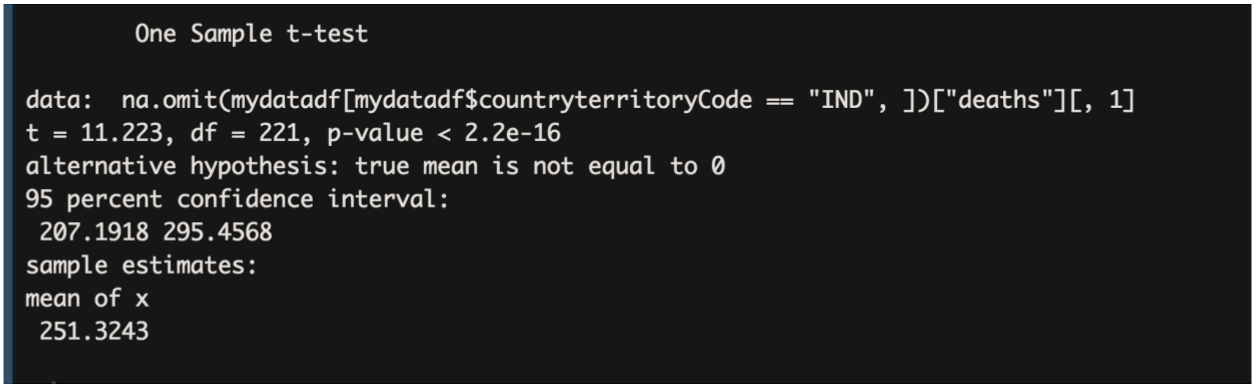


```
> t.test(na.omit(mydatadf[mydatadf$countryterritoryCode == "IND",])["cases"][,1])

        One Sample t-test

data:  na.omit(mydatadf[mydatadf$countryterritoryCode == "IND", ])["cases"][, 1]
t = 9.9472, df = 221, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 10748.44 16059.68
sample estimates:
mean of x
 13404.06
```

Deaths(207, 295)

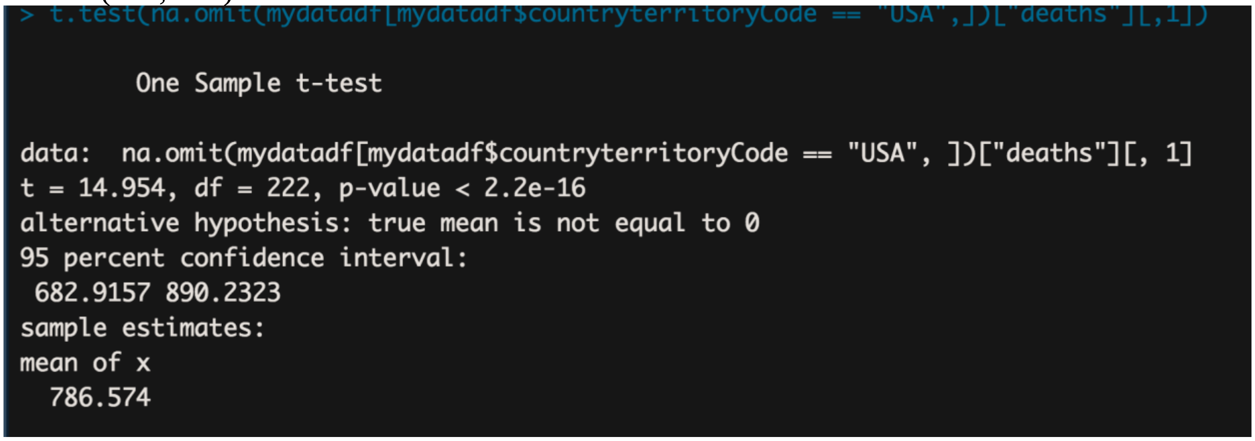


```
        One Sample t-test

data:  na.omit(mydatadf[mydatadf$countryterritoryCode == "IND", ])["deaths"][, 1]
t = 11.223, df = 221, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 207.1918 295.4568
sample estimates:
mean of x
 251.3243
```

USA:

Deaths (682, 890)

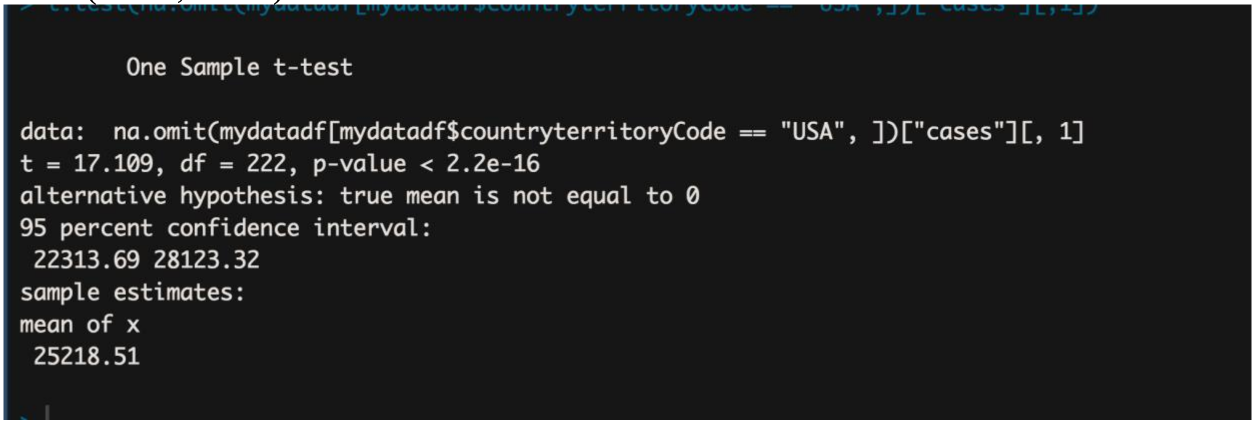


```
> t.test(na.omit(mydatadf[mydatadf$countryterritoryCode == "USA",])["deaths"][,1])

        One Sample t-test

data:  na.omit(mydatadf[mydatadf$countryterritoryCode == "USA", ])["deaths"][, 1]
t = 14.954, df = 222, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 682.9157 890.2323
sample estimates:
mean of x
  786.574
```

Cases (22313, 28123)

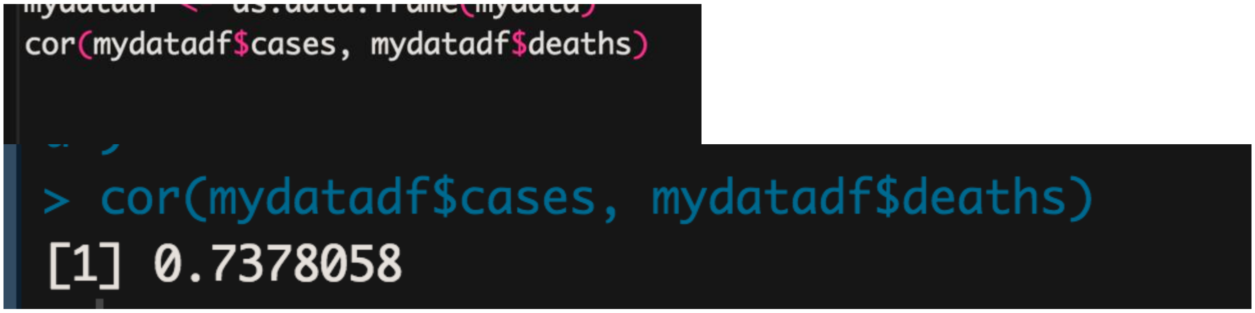


```
> t.test(na.omit(mydatadf[mydatadf$countryterritoryCode == "USA",])["cases"][,1])

        One Sample t-test

data:  na.omit(mydatadf[mydatadf$countryterritoryCode == "USA", ])["cases"][, 1]
t = 17.109, df = 222, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 22313.69 28123.32
sample estimates:
mean of x
 25218.51
```

3.)
The relationship is between cases and death rate is mildly correlated it is not highly correlated in that the correlation is not close to 1 but in the middle at 0.738.

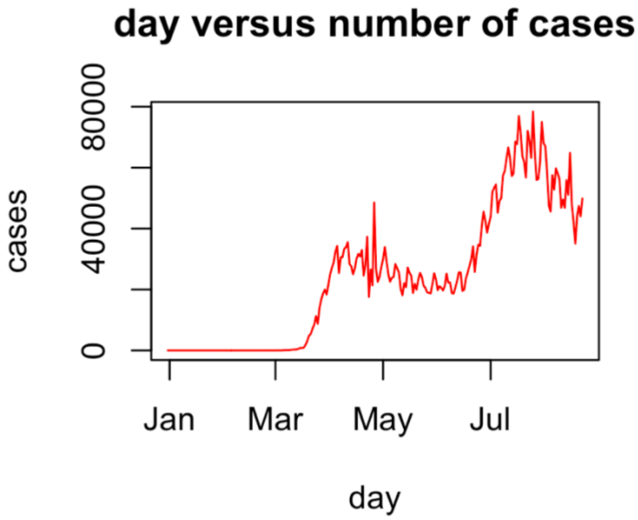


```
mydatadf <- as.data.frame(mydata)
cor(mydatadf$cases, mydatadf$deaths)
```

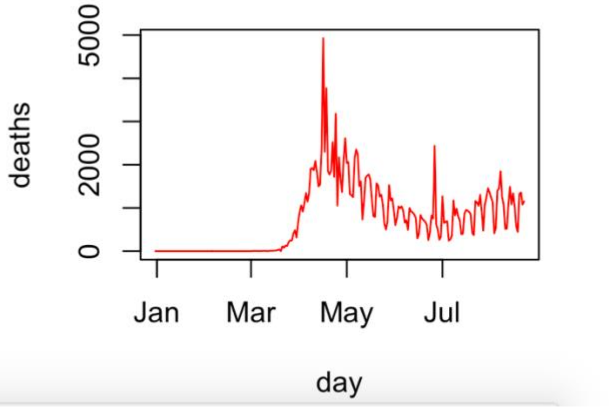
```
> cor(mydatadf$cases, mydatadf$deaths)
[1] 0.7378058
```

4.)

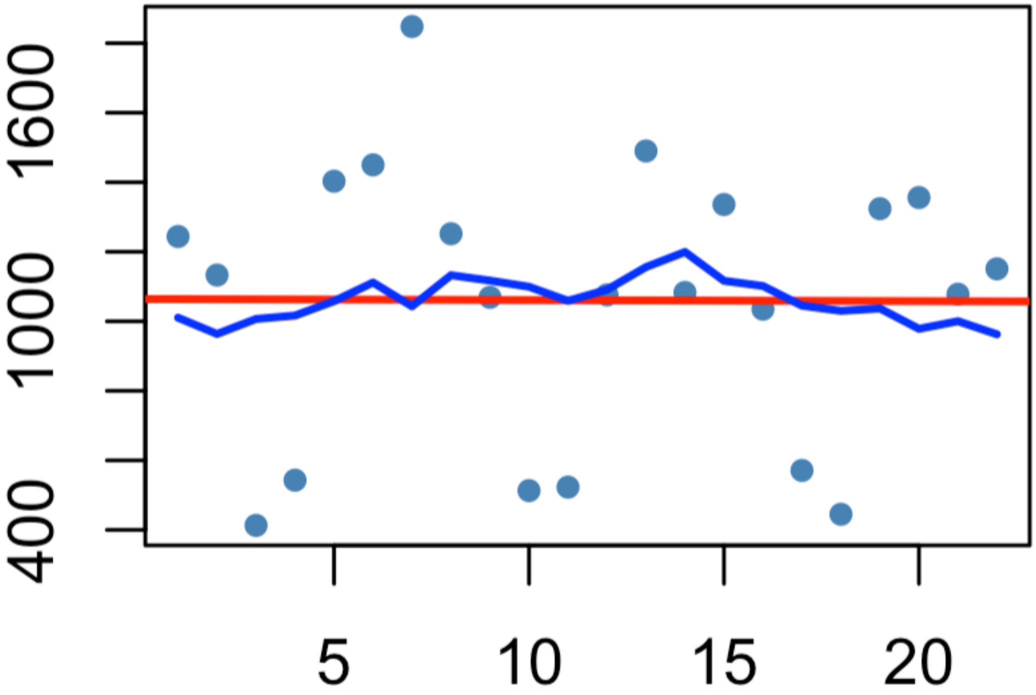
Time series of number of cases



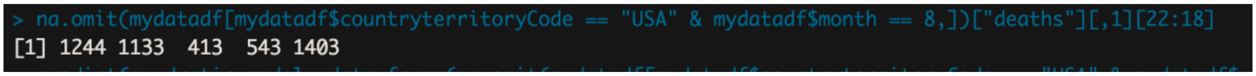
Time series of number of deaths



Here is the how the model fits the data using the lm function and also making it non-linear.

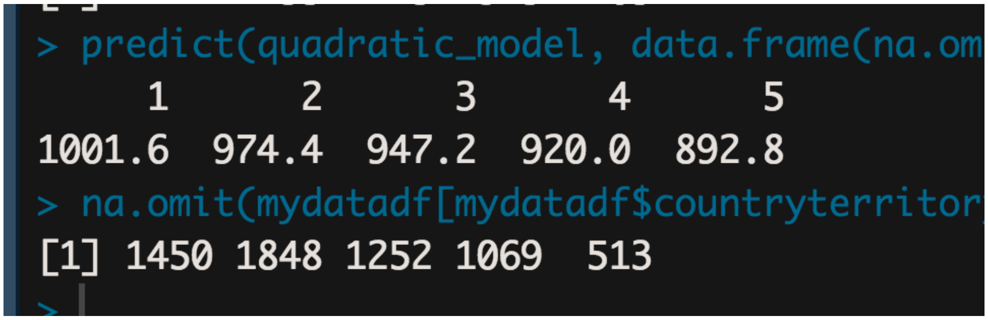


Training data based on deaths 5 days



```
> na.omit(mydatadf[mydatadf$countryterritoryCode == "USA" & mydatadf$month == 8,])["deaths"][,1][22:18]
[1] 1244 1133  413  543 1403
```

Prediction of next five days versus actual data



```
> predict(quadratic_model, data.frame(na.om
      1       2       3       4       5
1001.6   974.4   947.2   920.0   892.8
> na.omit(mydatadf[mydatadf$countryterritor
[1] 1450 1848 1252 1069  513
>
```