Brian Moran
Data 101 Final Project
May 5th, 2020

Covid-19 Data Review

Url- https://ourworldindata.org/coronavirus
Data- https://covid.ourworldindata.org/data/owid-covid-data.csv
Purpose:
The Data:

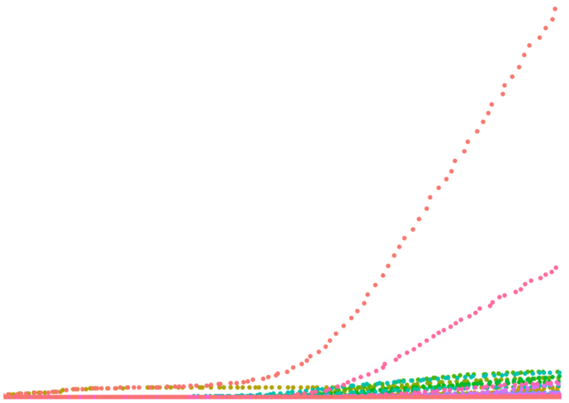
I believe this data to be very trustworthy due to the sources history is being a reliable source for accurate data sets. Their data sets are also kept open source so viewers can test and look at the data themselves if they feel graphs to be misleading.

Inside the CSV there are 14,502 elements over 16 different variables, including an ISO code, Country location, date, total_cases, new_cases, total_deaths, new_deaths, total_cases_per_million, new_cases_per_million, total_deaths_per_million, new_deaths_per_million and others.

The data is updated daily as new results are released. For this example, the data was last updated May 1st.

Graphing

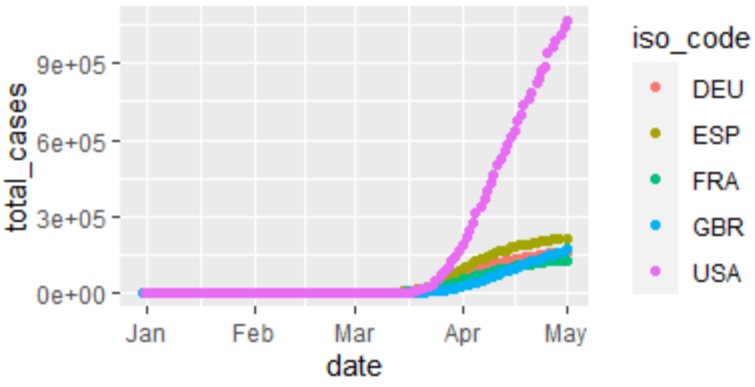
I first graphed all of the countries in the data set to get an idea of how it would look when graphing. Even with color separating by country, the data was extremely difficult to visualize with over 14,000 elements in the graph. I decided to simply use the top 5 countries with the highest amount of cases for the data study.



```
covidUS = subset(covid, location == "United States")
covidUK = subset(covid, location == "United Kingdom")
covidES = subset(covid, location == "Spain")
covidFR = subset(covid, location == "France")
covidGR = subset(covid, location == "Germany")

covidMain <- merge(covidUS, covidUK)
covidMain <- merge(covidMain, covidES)
covidMain <- merge(covidMain, covidFR)
covidMain <- merge(covidMain, covidGR)
```

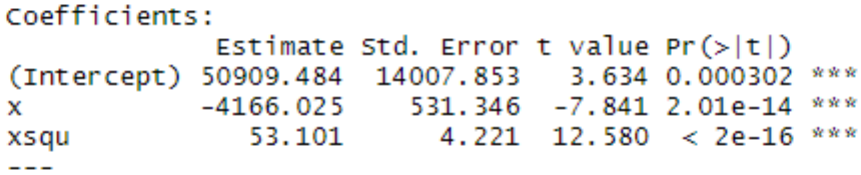
In this graph, I organized the data by highest number of cases and selected the top five countries: {United States, United Kingdom, France, Spain, Germany}. Placing these five subgroups into one graph made the data much better for visualization. Seeing the data visualization and how the countries all seem to follow a similar path as time moves on, I completed a regression analysis test to predict where the total case numbers could potentially spread.



```
ggplot(covidMain, aes(x=date
, y=total_cases ,
color=iso_code)) +
geom_jitter()
```

Regression Analysis Five Countries

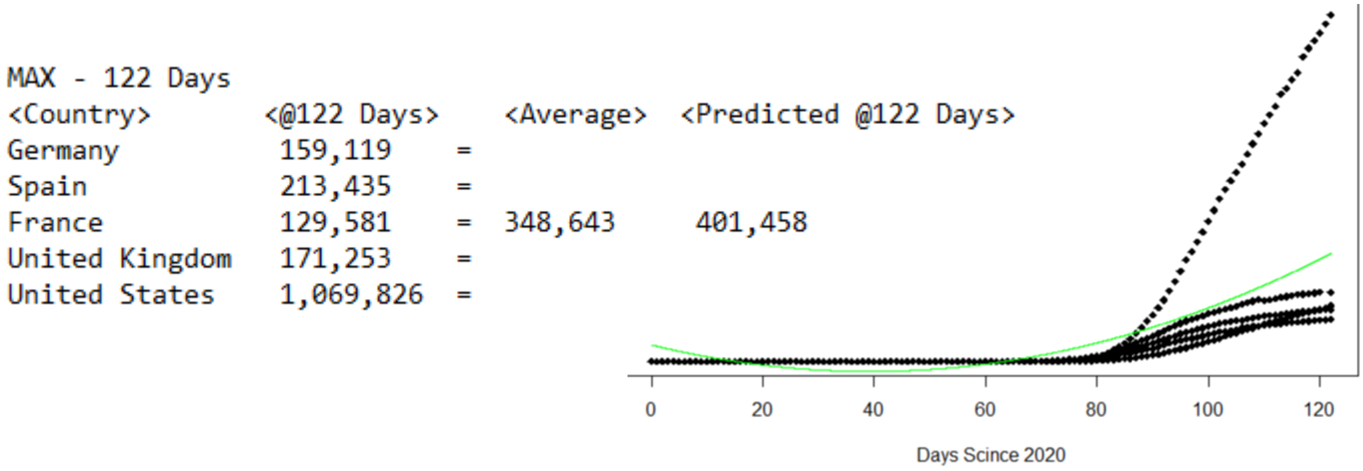
The date format of the elements were in a “2019-12-31” fashion so I first created a new column that was the days since the first in the data set. I first set up separate x and y variables for the number of days past and total cases respectively. I also made a x^2 variable as well as a x^3 variable for the regression test. Using lm() for simple regression models I started with a quadratic function using the formula lm(y~x+x^2).



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50909.484  14007.853   3.634 0.000302 ***
x           -4166.025    531.346  -7.841 2.01e-14 ***
xsqu           53.101      4.221  12.580  < 2e-16 ***
---
```

```
startdate <- as.Date("2019-12-31","%Y-%m-%d")
covidMain$days <- difftime(covidMain$date , startdate,
units="days")
covidMain$days <- as.numeric(covidMain$days)
x <- covidMain$days
y <- covidMain$total_cases
xsqu <- covidMain$days^2
xcub <- covidMain$days^3
```

Science all P-values are below .05, they can all be considered significant and be used for testing the regression.



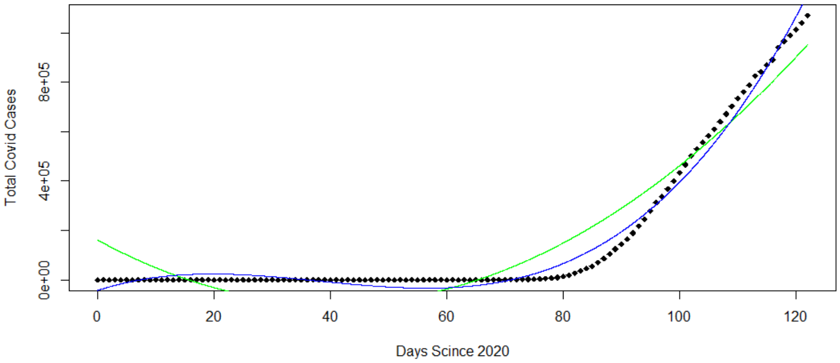
```
MAX - 122 Days
<Country>        <@122 Days>    <Average>  <Predicted @122 Days>
Germany           159,119     =
Spain             213,435     =
France            129,581     =  348,643     401,458
United Kingdom    171,253     =
United States     1,069,826   =
```

```
xval <- seq(min(x),max(x),0.01)
yval <- predict(output, list(x
= xval, xsqu = xval^2))

lines(xval,yval, col = "green")

output <- lm(y~x+xsqu)
summary(output)
```

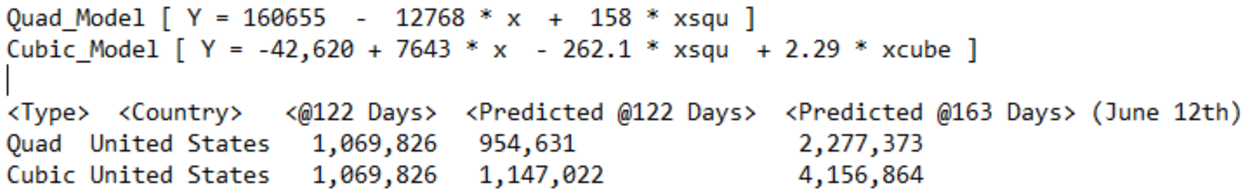
After seeing the result of what the regression was testing, I decided that it would be far more accurate to focus in and study one trend (country) in particular. Doing a regression test on the five countries together gives an idea of what the five countries will tend to average all of the results together instead of predicting each properly. To get a more in depth study I decided to focus just on the United States.



Regression Analysis United States

I began by doing the same as before and started with a quadratic model to get an idea of the data.
After graphing, I decided to try a cubic model to get more accurate results. I attempted an x^4 but the data showed it to be insignificant. (Cubic)(Quadratic)

Seeing that the cubic model seems to hold true to the data a lot more than the quadratic, it is alarming when looking at the predicted results for 41 days later.



```
Quad_Model [ Y = 160655  -  12768 * x  +  158 * xsqu ]
Cubic_Model [ Y = -42,620 + 7643 * x  - 262.1 * xsqu  + 2.29 * xcube ]
|
<Type>  <Country>   <@122 Days>  <Predicted @122 Days>  <Predicted @163 Days> (June 12th)
Quad  United States   1,069,826   954,631                2,277,373
Cubic United States   1,069,826   1,147,022              4,156,864
```

Analysis

It is interesting to see how different predictions can lead to such different results depending on how a test is done. Both of the predictions end in convex upwards slopes which increase exponentially. The quadratic function ends more generously with a softer upwards curve putting the US at about 2 million cases by June 12th. Although the cubic function is more accurate for the data being shown, it has a steep upwards curve at the end that seems to not match with the ending data and results in creating an extremely high prediction for 163 days (4 million).

This prediction demonstrates that something based off of strictly previous data is a difficult task. For factors like coronavirus there are far too many factors for a simple cubic regression analysis to predict the numbers. A much more complex predictor such as another similar virus with similar statistics can yield more accurate results.