STA 3000 Shakir Mahabirsingh

An Analysis of Covid-19 on Global & Country Data

Summary: The Coronavirus epidemic has caused a state of disarray around the globe. After finding a dataset that tallied global cases and deaths daily from December 31st, 2019 - April 5th, 2020, I knew I'd be able to pull some useful info from it. I looked at this data and sought to answer a few questions pertaining to the global impact of the virus. These questions were what country had the highest infection, which had the highest mortality, does higher infection lead to higher mortality, when (day/month) did these countries really start to see a sharp incline in infection and deaths? Given global data, can I quantify the relationship between infection and death? I also wanted to see if I could use a linear regression to find a rough estimate for the amount of deaths that the world could possibly reach. I wanted to focus on four main countries in further analysis and used The United States, India, Italy, and Spain. Each of these countries implemented very different social practices to fight the infection and I wanted to see what the data would reflect.

Data Description: The dataset was found on Kaggle.com and is under the UNCOVER COVID-19 data challenge. It is sourced and updated from multiple data scientists across the globe. It tallies the daily updates on cases (infection) and deaths for every country from December 31st, 2019 to April 5th , 2020. It can be found with the link below:

https://www.kaggle.com/roche-data-science-coalition/uncover/version/3#current-data-on-the-geographic-distribution-of-covid-19-cases-worldwide.csv

Data Manipulation: After reading the initial data into RStudio, I knew there were some columns I wasn't going to need. These were the *geoid* and *countrycode.* These were just country identifiers and wouldn't serve a purpose in my analysis. I formed a new dataset without these columns by using *select* to remove them. Then I needed to clean my data in case there were any NAs that would disrupt my analysis. I also needed the data to read correctly from the earliest date to the latest date in proper order. This I did by using *arrange* to first organize by year, month, then date. This resulting dataset displays all the global info for the first recorded day then moves onto the following day.

Later in my analysis I start to pull the case and death data specific to the four countries I'm focusing on. I get the info and data I want by using *filter* and *select.* I assign these to a variable to make it easy to call and obtain the info.

After subsetting the data for each specific country, I created a table to show their relation to the world and their death amounts. Finally, I wanted to show a time frame scatterplot for each country and used *filter* and *select* again to get the case and death data over the total allotted time.

Data Visualization: I wanted to create a linear regression to see the prediction effect **cases** and **population** had on **deaths**. I initially built a simple linear regression model with both predictors. I knew I wanted to test this against another model and created a model only using population data as a predictor. I thought that maybe countries with larger populations would obviously see higher deaths tolls given the mortality rates globally. After running these two models, the model keeping both predictors was a better fit. It had a higher R-Squared, which means the dependent variable's variance is better explained given the included independent variables. A higher R-Squared means a better explained prediction variance. While my model *mdl* is a good enough fit, it could be better with the addition of variables more closely related to deaths. No model is perfect but the more explainable variance, the better.

Based off the model:

- For every one-unit change in Cases, Deaths will increase by 0.04507761
- For every one-unit change in PopData2018, Deaths will increase by -0.000000005880054 (inverse relationship)

This follows general intuition which is a good sign. It's notable that based off this data and model, cases have a **+4.5%** affect on deaths globally. We've now quantified the relationship between cases and deaths. Plotting Deaths to Cases provides a visual representation of this affect. I then used R's *predict* function to develop a graph to show the further deaths and cases. I attached confidence interval bands to show that the data will fit into those lines.

After subsetting the data for each country, I found their cases, deaths and their proportion to the greater world data.

- The US has 312,237 total cases & 8,501 total deaths
- India has 3,374 total cases & 77 total deaths
- Italy has 124,632 total cases & 15,362 total deaths
- Spain has 124,736 total cases & 11,744 total deaths

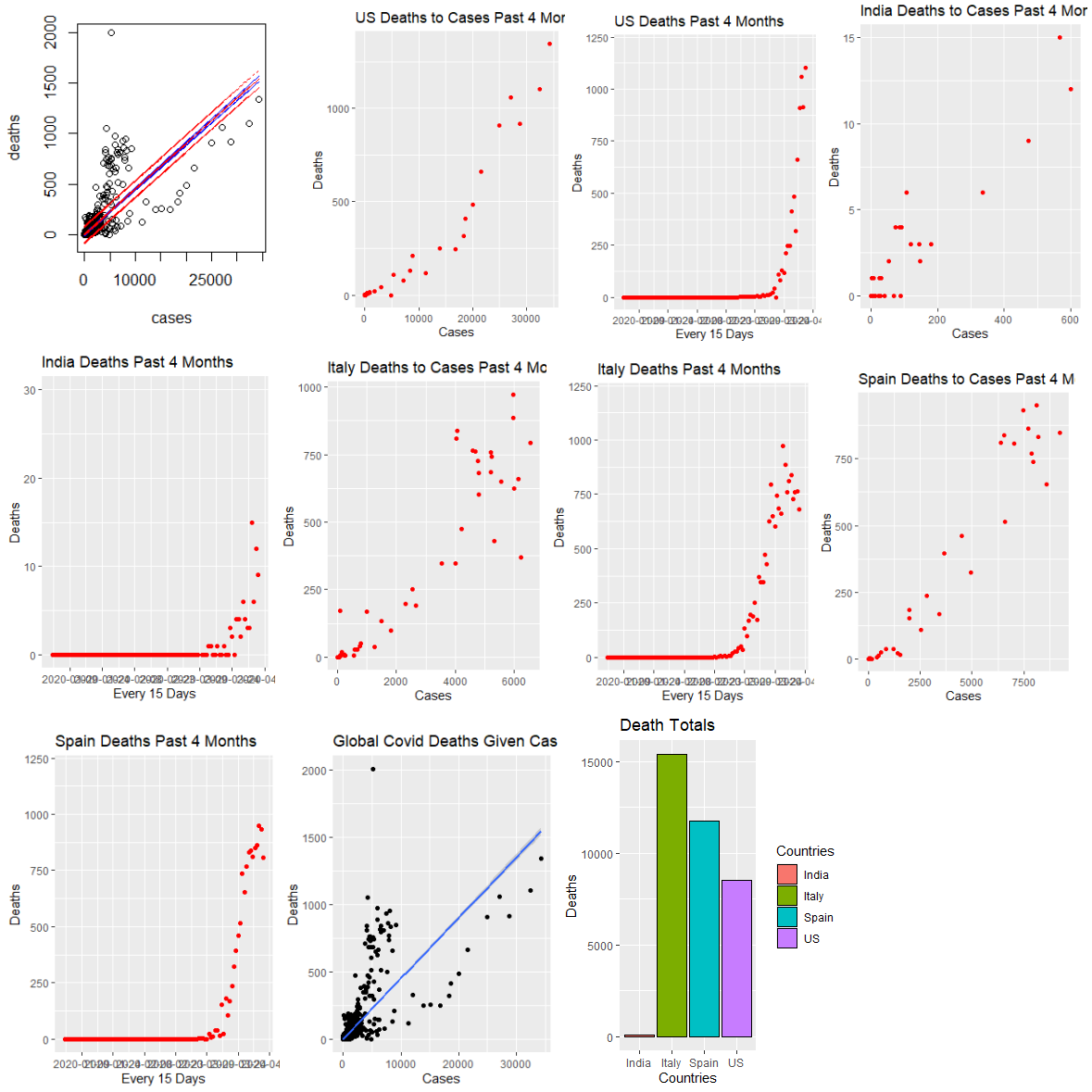
Cases to the World:

- The US has 26%, India has 0.28%, Italy has 10.61%, and Spain has 10.61%

Deaths to the World:

- The US has 13.2%, India has 0.11%, Italy has 23.85%, Spain has 18.23%

I've attached all my graphs below to show the outcomes of the analysis discussed.



Conclusion: Analyzing the world data answered and quantified the questions I originally started with. We can see that although the US seems to have the highest infection rate (given most cases), it is actually Italy who has the highest mortality rate (given most deaths). I think we can relate this to Italy's loose social distancing and enforcement of Covid-19 safety measures. While the US has the most cases, our quickly enacted measures saved a lot of deaths from happening. We can also see Spain is having a hard time slowing the rate of deaths given they have roughly the same amount of cases as Italy. I am led to believe that India is either misreporting or flat out not reporting virus cases because it seems very unlikely that the most populated country per

capita has the lowest measurements in every category. Given the news that China and some other countries are trying to hide the total number of deaths and cases, India doing the same seems very likely. It seems that the end of March is where these countries started to see the sharp rise of cases. This data is devastating to see but the more it can be analyzed and visualized, hopefully the more knowledgeable the public can become to further prevent higher mortality and infection.