**Name: Mohammad Siddiqui Date: 05/04/2021**

**Exploratory Data Analysis of COVID19 in New Jersey**

This project was critical, in my opinion, in light of the current situation. Following a thorough review of my project plan. I made the decision to work towards the same goal, which was to obtain well-analyzed data for Covid-19 cases in New Jersey. Hence, I decided to give it my best shot working on this project.

The first goal was to compile a comprehensive dataset that would enable us to focus on Coronavirus cases in our home state of New Jersey. I been able to find a few datasets on Kaggle, the home of datasets, after doing some analysis. The most difficult part was deciding which datasets would better help me reflect the data I needed to display I was still on the lookout for datasets that were dynamic and regularly updated on a daily basis. I was fortunate enough to obtain the NYTimes dataset for the states and counties, which was a raw CSV file, allowing us to create a few of our charts and represent a significant amount of data, which can be very useful given the current circumstances. However, I am unable to locate such a dataset for displaying the number of cases at a county level and were forced to rely on a CSV file downloaded locally from NJ government sources, ensuring that the information reflected in the charts was as accurate as possible.

Any dataset I want to use may or may not contain the information I need. There is always the possibility of mistakes and typos. But the first step was to clean up the data and make it as accurate as possible. I had to practice data cleaning and rearing techniques. This is where the Pandas kit came in useful. I used pandas read\_csv function to import the data as a data frame. I could not delete rows, edit entries, add columns, delete columns, restructure the dataset, or something else because the data was imported into a data frame. First thing I did after reducing the dataset to include entries only within the state of NJ was to add a column to dataset for everyday new cases and everyday new deaths. It was a very simple mathematical approach which was basically just subtracting the current count with the previous count in the both the number of cases and deaths. After that, I focused on estimating an average of the number of cases and counts. Since the counts were rapidly shifting, the daily average turned out to be very unpredictable and was not reflecting the true values. To reduce this uncertainty, I added the average of the previous four days to the overall average. This also helped us balance out the mean line a little.

The next step was to research New Jersey's population to determine how many cases per million people there are in our culture. I got this information from the Census Bureau's website. However, this time the dataset was not so clean, the columns were not named right, the rows that NaN or NULL values, the sequencing was off. So, the data cleaning was not simply filtering out the erroneous data but also to bring the dataset to usable format and structure. Again, I made us of Pandas data frames to help us bring this dataset to life and were able to extract the population of NJ from the census bureau’s data till the year of 2019. The number was 8,882,190.

To display the number of cases per million population, I added two more columns to our previous dataset. Which means how many people in our state have coronavirus for every million people. This number, like all the others I have measured, is dynamic, meaning it will change with each new entry to the Covid-19 dataset on the NYTimes website. And the same procedure was repeated for the number of deaths per million.

The next step was to transform this data into interactive graphs. I've worked with matplotlib to create maps, but the issue is that they're not real-time charts, as we've seen on numerous websites with charts and graphs that return information when you select or click on them. As a result, I used the Plotly software. The express library with the plotly package was used to plot these graphs.

Finding the number of cases and breaking it down at a county level was the last challenge I wanted to focus on. As previously mentioned, this dataset was not dynamic, so I used the most recent version as of May 9, 2020. The goal was to find out how to reflect this onto a map of New Jersey divided into counties, which was the most challenging aspect of the task. For this pulled in a json file I obtained on github which had the map of USA broken into counties of all the states with the coordinates. Then compared a cleaned dataset against the county values and showcased the breakdown of cases on a map. Honestly, I was not targeting towards making a map which showed cases of all dates but studying the documentation of plotly package I was able to achieve that.

In the future, I would like to learn more about machine learning and try to make predictions using the datasets I have. Generalizing these graphs and creating different situations in which the number of cases increases, or decreases will be amazing. Also, I am trying to compile a dataset on how many people have recovered from this virus and how many cases are still awaiting results, but I cannot find it on a reputable data base like NJ.gov. But I would like to keep working on this project because the possibilities for researching this virus seem limitless, and conditions change on a daily basis.