Final Project Documentation
CS-204

Alejandro Perez
Stefano Biasini
Hisham Aladil

Part A

We had to collect raw data about COVID-19 since March 1st, 2020 from the United States. With some research we found and collected from the CDC website an excel file containing all the COVID-19 data for the United States and its Territories. The data collected was from January 1st, 2020 until November 15th, 2020(day we found the data). We had to remove data from January 1st, 2020 to Feb 29 , 2020, after removing it we started manipulating and playing with this data.

To read the data from the excel file we used Pandas, a data analysis toolkit, this toolkit is very popular, and it has a wide variety of functions to use. We used the wrapper function read_csv which reads data from a csv file. After reading the data we need it to store it somewhere, so we stored in a dataframe which is another function from Pandas. We had to remove NaN values which we got from previously removing the data, we used fillna() fuction and set the parameter to 0 now all our NaN values are 0. We then converted the dataframe to a list for better manipulation of the data. After this we were ready for the next part of the project. The data collected consisted of each state COVID-19 cases, cases hospitalized, and positiveIncrease of each day.

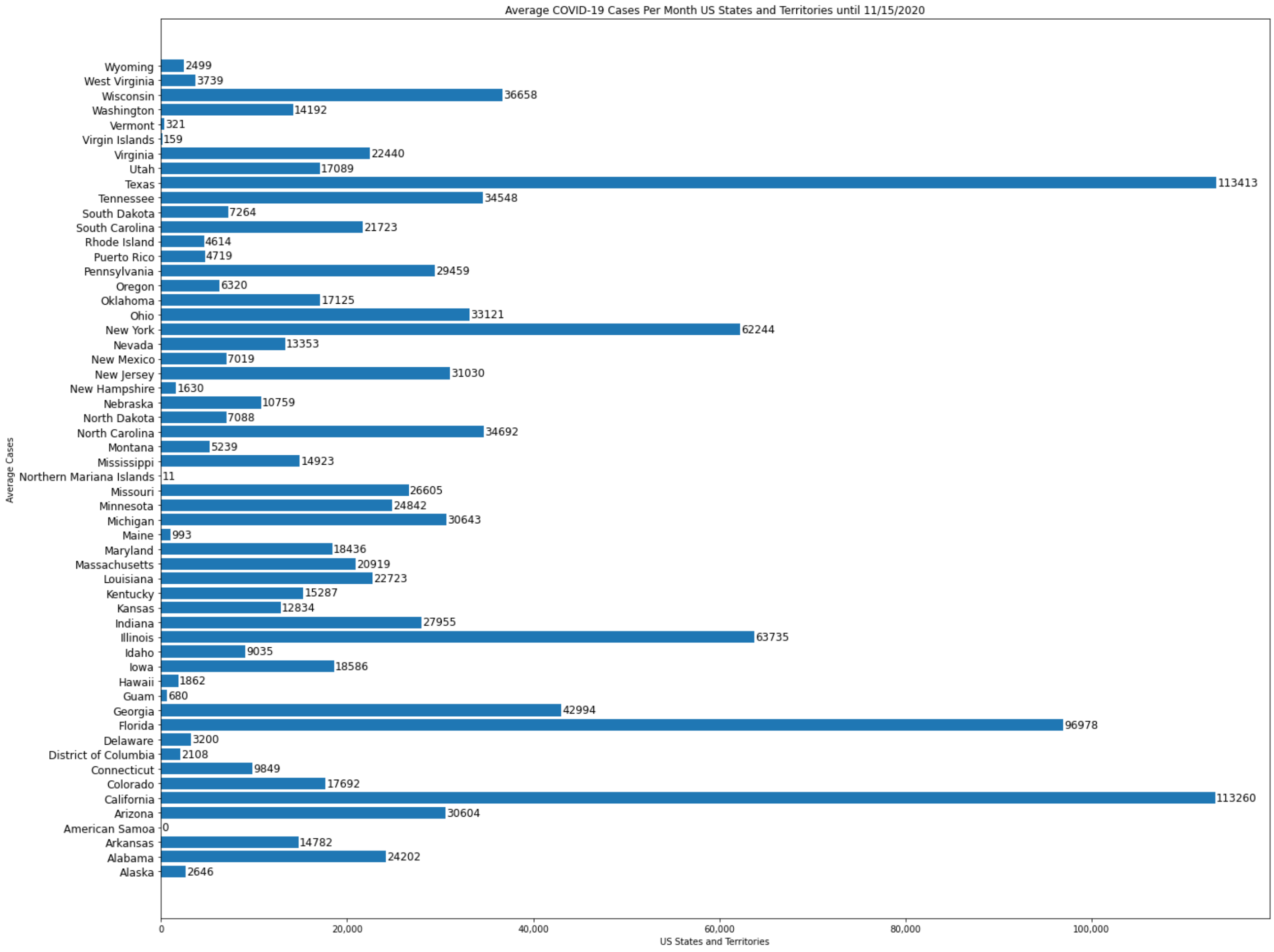
Part B

In this part, we had to display the average number of positive cases per month per state. So, we created an average list in which we stored the total cases divided by the number of months from March until November (which are 9 months).

We used list comprehension to store the average cases in a dictionary with key as state and values as the averages. After that, we calculated the "Average Cases Per Month" by using the average list and the state list. In the code this part is commented out because It was no longer needed as to it was only used to make sure outputs were correct.

We then created a methodology which we use all over the project, its best to just explain it here, we needed to make the data understandable for every person including non-programmers to be able to read and understand, so displaying it in a dictionary was not the right way. We decided on using a graph in which our data will be plotted. We got every state and territory full label and store it in a list, better for display rather than just the abbreviation of the state. We then got the average list and made a horizontal bar graph with the actual data label value at the end of the bar, we then made the x labels more readable and finally we created the title and x, y labels of the graph. (See Figure 1)

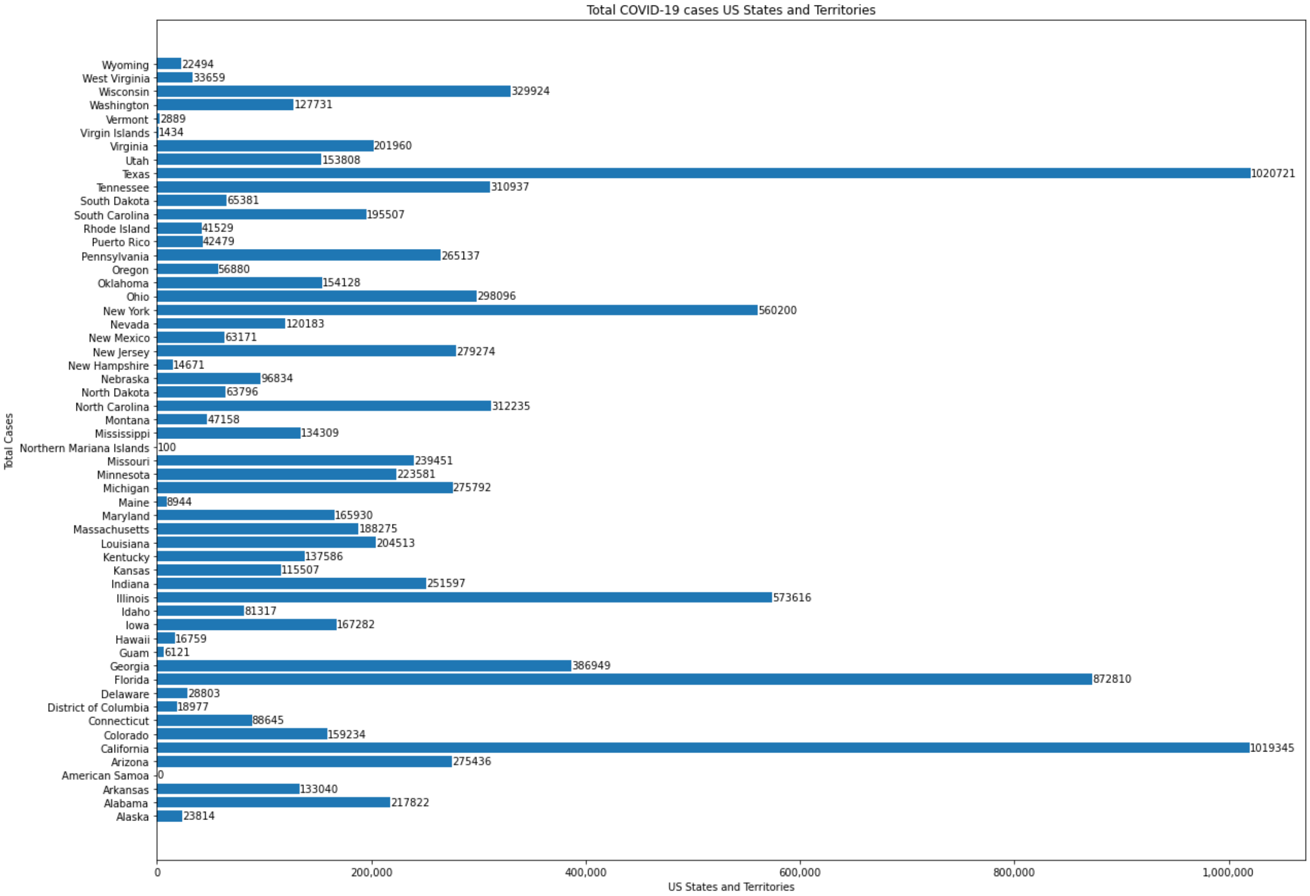
Figure 1



Part C

In this part, we had to display the total number of positive cases per state. In the code, we had to consider the total cases. So, we took each state with the cases related to it and the output was the representation of the total number of positive cases per state from March 1st, 2020 until November 15th, 2020. We then followed the previously mentioned methodology to plot the data. (See Figure 2)

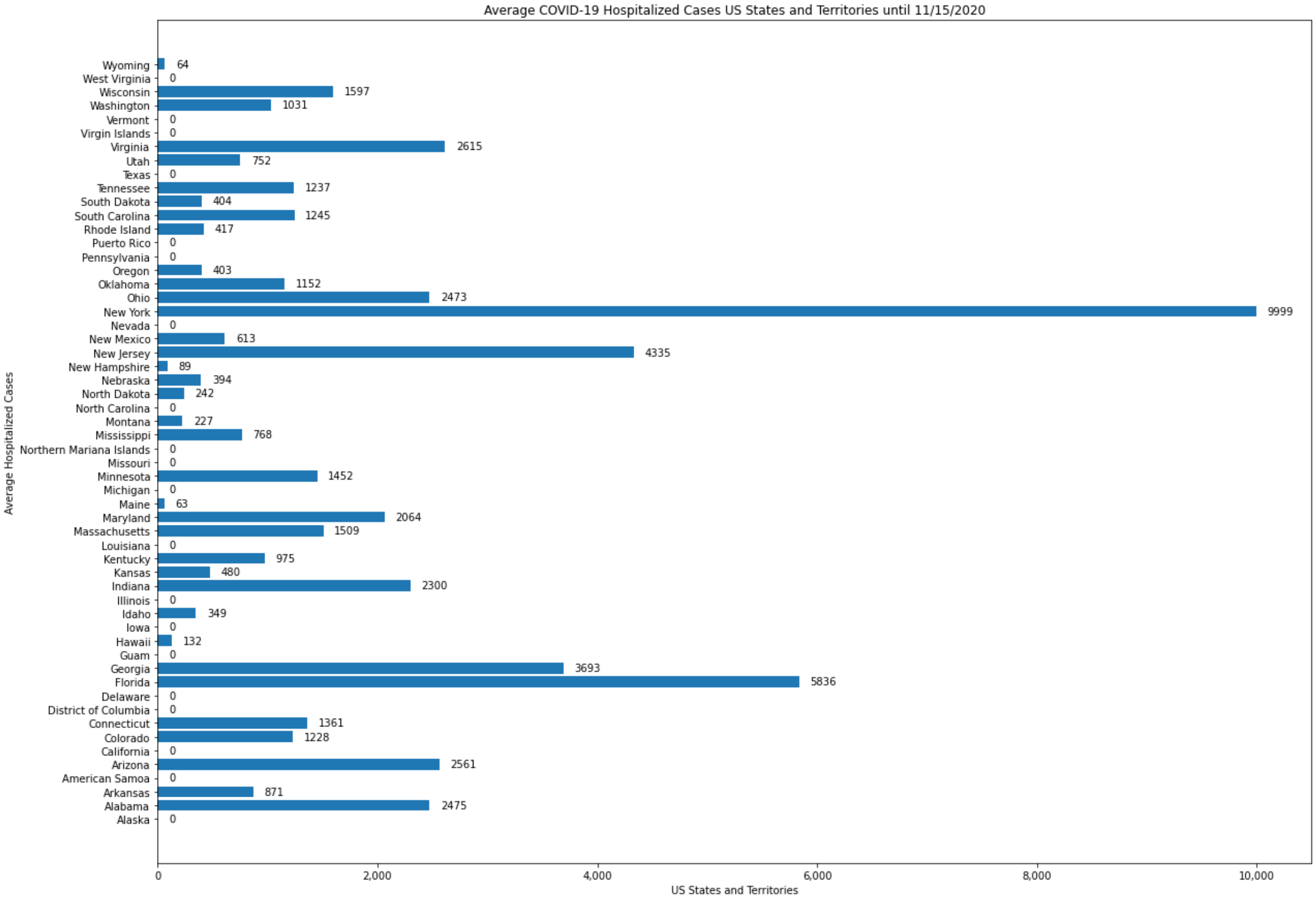
Figure 2



Part D

In this part, we had to display the average number of hospitalized people per month per state. First, we created an average list in which we put all the hospitalized data for all the period analyzed divided them by 9 (number of months from March until November. After that, we calculated the "Average Hospitalized Cases Per Month" by using the average list and the state list. The output was the representation of the average number of hospitalized people per month per state from March 1st, 2020 until November 15th, 2020. We then followed the previously mentioned methodology to plot the data. (See Figure 3)

Figure 3

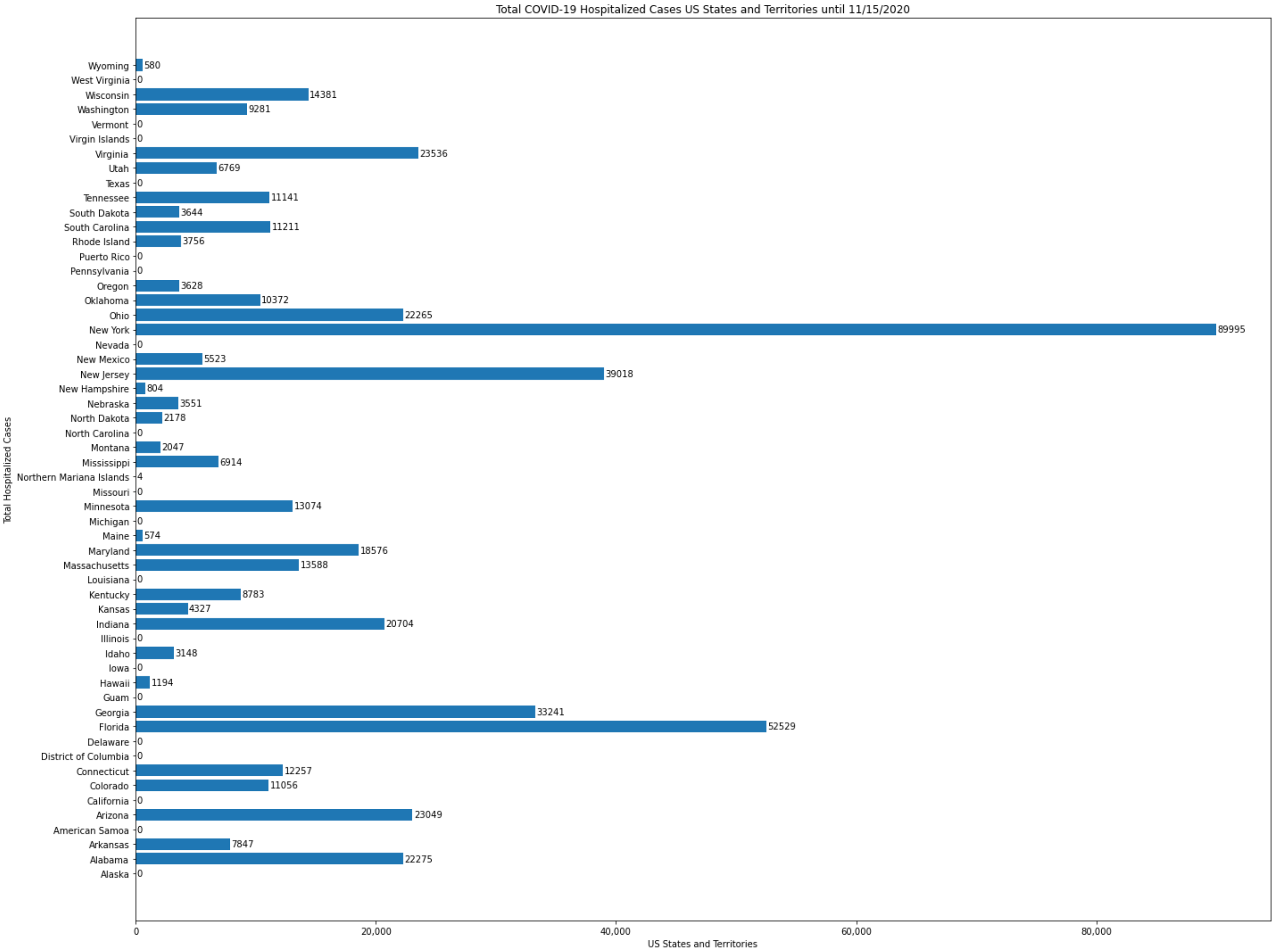


NOTE: 0 values in this plot are NaN values

Part E

In this part, we had to display the total number of hospitalized people per state. In the code, we had to consider the total hospitalized cases. So, we took each state with the hospitalized cases related to it and the output was the representation of the total number of hospitalized people per state from March 1st, 2020 until November 15th, 2020. We then followed the previously mentioned methodology to plot the data. (See Figure 4)

Figure 4

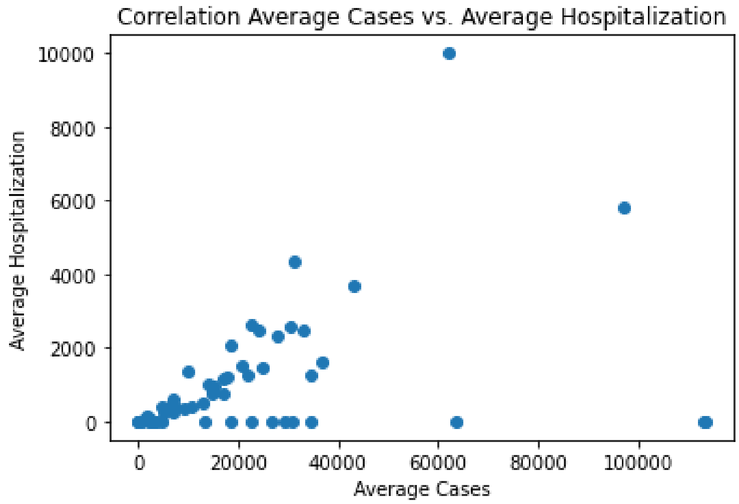


NOTE: 0 values in this plot are NaN values.

Part F

We had to check for correlation between part B and D and explain results found. To make this correlation we used a simple scatter plot with our two-average list, average cases, and average hospitalization. Our output we find that there was significantly positive correlation between the average total cases and the average hospitalized cases.(See Figure 5)

Figure 5



Part G

Our task for this part was to display the mean and standard deviation of daily positive cases for the state of Florida. Since we had a state column, we needed to find every iteration were FL string appeared, then with this same iteration get the positive increase. To do this we used list comprehension to get ith elements from state list and get only the daily positive increase values in the state ith element. Doing this helped us a lot in getting the daily positive increase of cases only in the state of Florida. With this we could then find the mean and the standard deviation. For this we used the mean and standard deviation functions already given by statistic module. We just plugged in our list as parameters.

Output:

```
Mean Daily Positive Cases =  3369.92277992278
Standard Deviation Daily Positive Cases =  3296.976351717038
```

Part H

For this part of the project which consisted of finding out, if the data for daily cases in the state of Florida followed normal distribution, we made a histogram graph with normal distribution curve to get a better look of the data if it follows a normal distribution using the daily positive cases of Florida between each day between March 1st to November 15th. Based on the plot we saw a clear bell-shaped normal curve centered around the mean with little deviation demonstrating a normal distribution. (See Figure 6)

After going over our output with the professor, there was a flaw in our graph our standard deviation values in the graph were off. Our problem was that our standard deviation in the graph were off and we figured that the problem was the np.random.normal function, which returns a list of random samples drawn from a normal distribution, however we were not able to figure out why it was causing it. Therefore, as suggested, we then proceeded in using a normal bar chart and not using a datagram chart. The normal bar graph showed no sign of a bell curve so our daily positive cases for Florida does not follow a normal distribution. (See Figure 7)

Figure 6

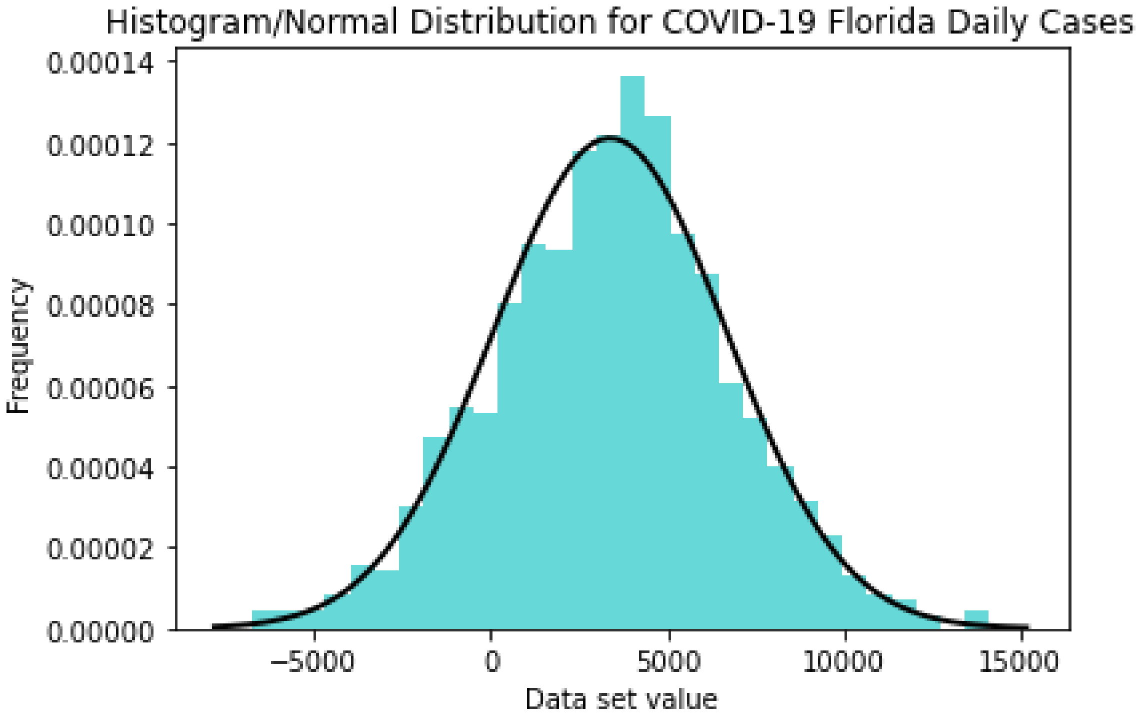
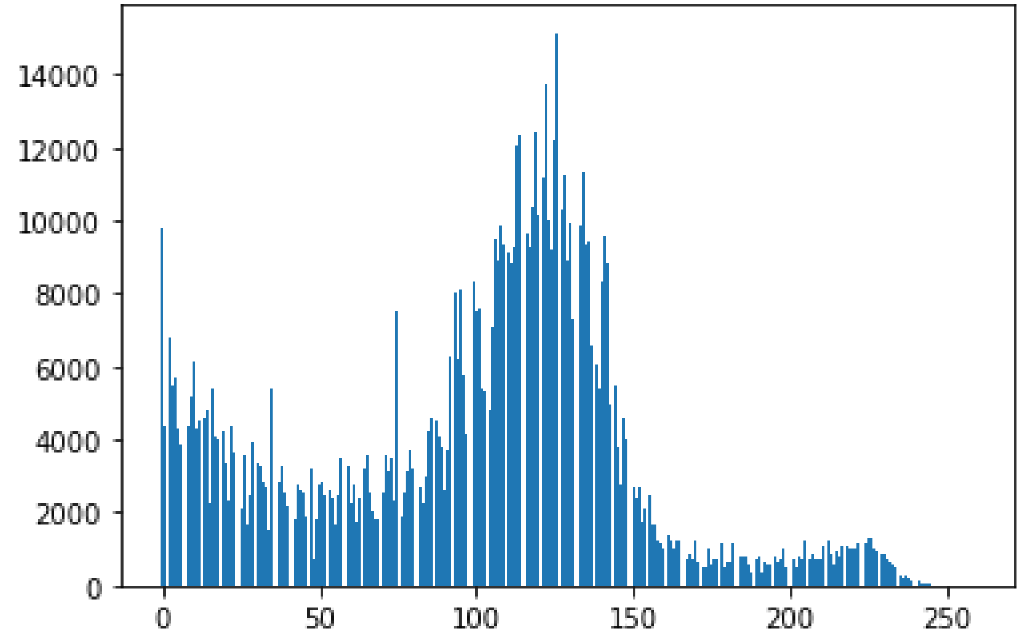

Histogram/Normal Distribution for COVID-19 Florida Daily Cases
Frequency
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
−5000
0
5000
10000
15000
Data set value

Figure 7



Part I

For this part of the project, we had to use a regression method to project number of cases for Florida for the month of December and January (assuming there is no vaccine). To find this we used functions presented in class and incorporated them into our program. First, we used the least square fit to find the best line fit for our alpha and beta. We made a scatter plot of our data points which are the positive daily cases for the state of Florida in one month. Our Y hat is the new cases, and our X is the month of December or/and January. (See Figure 8 for plot)

Output:

alpha 2539.4193548387093

beta 117.74798387096774

r-squared 0.44181370107677054

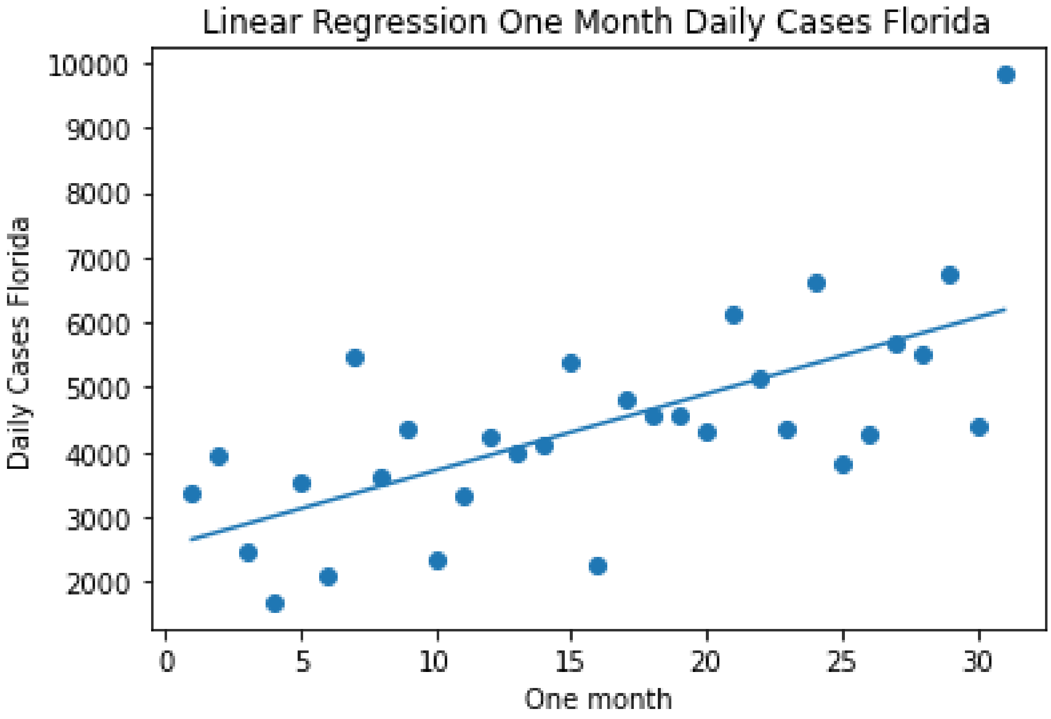
Equation: Y = 117.7*X + 2539

New Cases = 117.7*31 + 2539 = 6,187.7 //Month of December

New Cases = 117.7*62 + 2539 = 9,836.4 //Month of December and January

So, we conclude that the new cases will be increased to 6,188 after 31 days. Also, we conclude that the new cases will be increased to 9,836 after 62 days on the state of Florida.

Figure 8


Linear Regression One Month Daily Cases Florida
Daily Cases Florida
One month