Ritika Rajiv & Nicholas Scuccimarra

Professor Carter

MA 346: Data Science

December 3, 2020

Github repository : https://github.com/ritikarajiv/Project-2

Project Two

Curve Fitting COVID Data Versus Google Mobility Data

I. *Overview*

The introduction of Coronavirus-19 to the world has resulted in countless changes to the way in which many people live their lives, and has also caused questions to arise about how exactly to track the behavior of the virus as it still plagues countries throughout the world. It is no question that the United States in particular has been hit the hardest by the virus, with case counts and deaths still climbing as 2021 approaches. In particular, COVID-19 has piqued the interest of Data Scientists and Analysts, all trying to find a way to predict the spread of the virus by modeling it against other variables that may be correlated.

One promising set of data that could be used to model the rate of spread, and possibly predict its path into the future, is Google's newly released *Mobility Dataset*. This data takes into account the average movement of the population on any given day. This mobility value can be broken up into a few distinct categories, each of which pertain to the average movement of the population within these categories. Google's mobility data has the following mobilities listed: transit stations, retail & recreation, workplaces, and parks. Each of the mobilities is a percent change from the baseline, the baseline being the mobility value from 2019.

Our variable of interest in this analysis will be the reproduction rate of the Coronavirus on any given day in the United States. The reproduction rate is a useful metric in trying to predict the overall spread of any given virus, and according to Marwan Al-Raeel it is, "the expected cases which (are) generated by one infectious case in a population"(Al-Raeel, 2020). Reproduction rate is often used in SIR models (Susceptible-Infected-Recovered), and can give us a general idea if the population is overcoming a pandemic or not. Generally speaking, if the reproduction rate is greater than one, the virus is still rapidly spreading. If it is less than one, the spread of the virus is slowing, and moving towards dying out. The question that is being answered in this analysis is whether or not mobility data can be used to help predict the reproduction rate. In other words, will an increase in mobility result in an increase in reproduction rate as well?

II. Data Cleaning & Merging

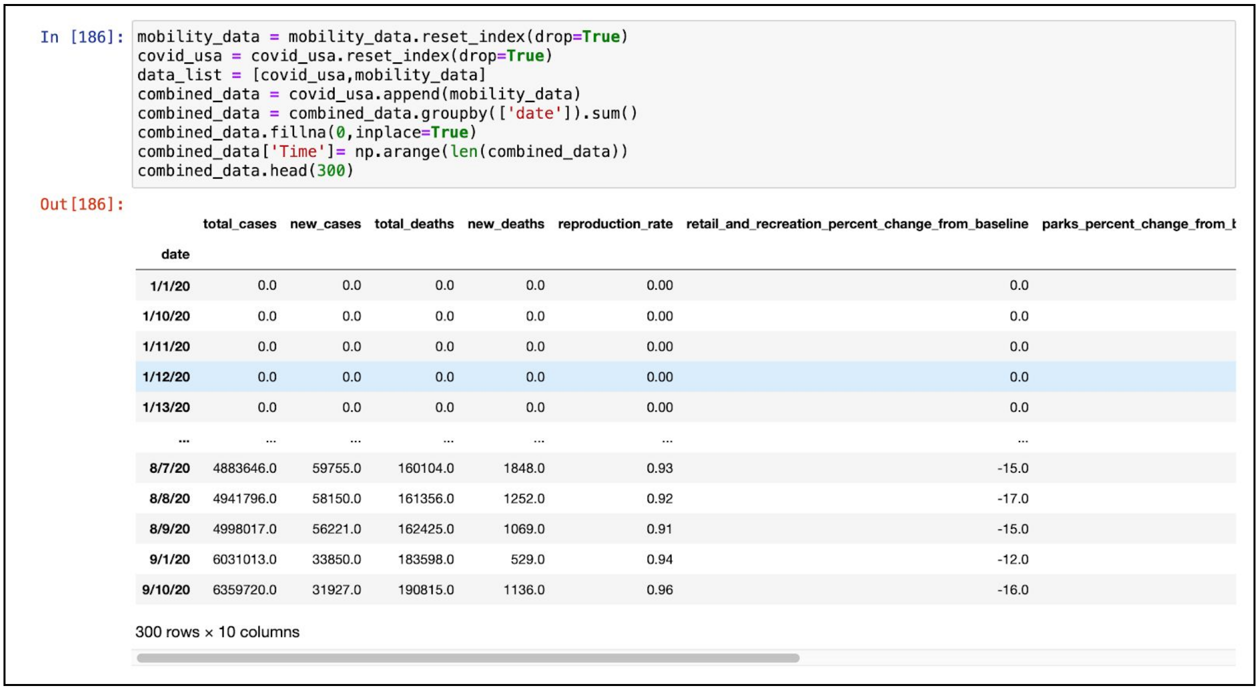
An important first step in data analysis is to clean the data to make sure it only contains values we are interested in, and is consistent in its data-types with no missing values. As stated above, the two datasets were obtained from different sources, so they have to be merged together to form one large dataset to work with. From the COVID Dataset only six of the columns were extracted:

- Date: given date of record in XX/XX/XXX format.
- New_cases: number of new cases counted on any given date
- Total_cases: cumulative cases
- New_deaths:number of new deaths counted on any given date
- Total_deaths: cumulative deaths
- Reproduction_rate: variable of interest.

From the mobility dataset, there were eight columns extracted:

- Country_region_code:
- Country_region:
- Sub_region_1:
- Date: given date of record in XX/XX/XXX format.
- Workplaces_percent_change_from_baseline: mobility in places of work.
- Transit_stations_percent_change_from_baseline: mobility on public transportation.
- Retail_and_recreation_percent_change_from_baseline: mobility in entertainment.
- Parks_percent_change_from_baseline: mobility in public parks.

After checking the dataset for no 'NaN' values, and reducing the datasets to only necessary columns, the two files can now be merged into a singular table. The following code was used to combined the mobility and COVID data:



```
In [186]: mobility_data = mobility_data.reset_index(drop=True)
          covid_usa = covid_usa.reset_index(drop=True)
          data_list = [covid_usa,mobility_data]
          combined_data = covid_usa.append(mobility_data)
          combined_data = combined_data.groupby(['date']).sum()
          combined_data.fillna(0,inplace=True)
          combined_data['Time']= np.arange(len(combined_data))
          combined_data.head(300)
```

Out[186]:

	total_cases	new_cases	total_deaths	new_deaths	reproduction_rate	retail_and_recreation_percent_change_from_baseline	parks_percent_change_from_
date							
1/1/20	0.0	0.0	0.0	0.0	0.00	0.0	
1/10/20	0.0	0.0	0.0	0.0	0.00	0.0	
1/11/20	0.0	0.0	0.0	0.0	0.00	0.0	
1/12/20	0.0	0.0	0.0	0.0	0.00	0.0	
1/13/20	0.0	0.0	0.0	0.0	0.00	0.0	
...	...	...	...	...	...	...	
8/7/20	4883646.0	59755.0	160104.0	1848.0	0.93	-15.0	
8/8/20	4941796.0	58150.0	161356.0	1252.0	0.92	-17.0	
8/9/20	4998017.0	56221.0	162425.0	1069.0	0.91	-15.0	
9/1/20	6031013.0	33850.0	183598.0	529.0	0.94	-12.0	
9/10/20	6359720.0	31927.0	190815.0	1136.0	0.96	-16.0	

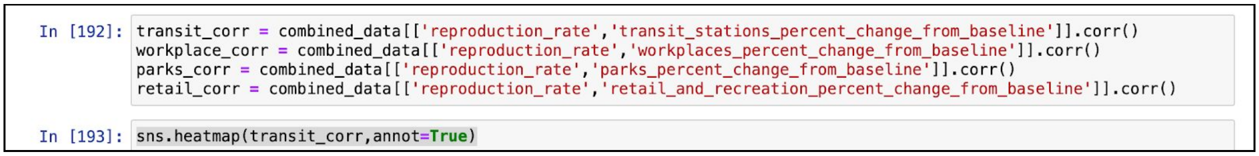
300 rows × 10 columns

Now that our data is all in a singular pandas dataframe, it is almost ready to be plotted and observed. Notice the range of the various columns of interest, there are different scales depending on the type of variable. Due to this, it is evident that a normalization of our columns will have to occur to bring them all on a scale from zero to one.

III. *Observing Relationships Between Variables*

Once the two datasets were combined together, it was now important to visualize the data to see if there were any relationships between our given reproduction rate, and any type of mobility data given. In this analysis, there is a step-by-step process by which we observe these variables. The first step is making a simple line plot of our variable which can be found in **Appendix I**. It can be observed from the line graphs created that each of the graphs for mobility are similar in shape, which would make sense due to social distancing guidelines affecting all of the mobilities. Furthermore, we can observe from the reproduction rate line plot that around day 120 (starting on January 1, 2020) there is a large increase in reproduction rate, around a similar time all of the mobilities experienced their maximum values as well.

A common way to observe if two variables are related with one another is to calculate the correlation between the two. The next step in the process is to test the correlation of each of the mobilities with the reproduction rate variable. To achieve this goal the following code was used:



```
In [192]: transit_corr = combined_data[['reproduction_rate','transit_stations_percent_change_from_baseline']].corr()
          workplace_corr = combined_data[['reproduction_rate','workplaces_percent_change_from_baseline']].corr()
          parks_corr = combined_data[['reproduction_rate','parks_percent_change_from_baseline']].corr()
          retail_corr = combined_data[['reproduction_rate','retail_and_recreation_percent_change_from_baseline']].corr()

In [193]: sns.heatmap(transit_corr,annot=True)
```

Each of the four correlations were put into specific variables associated with type of mobility's correlation with the reproduction rate. Heatmaps were created for each of the four correlations by using the heatmap function in seaborn; an imported package (**Appendix II**). An important observation to make is that all of the mobilities have a negative correlation with reproduction rate other than park mobility. The strongest correlation seems to be between reproduction rate and transit mobility, with a value of -0.44. Seeing that this is the highest correlation, a great next step is to plot both the reproductive rate and transit mobility on the same graph. It is important to first turn these two variables; along with the other three mobilities, into *rolling seven day averages.* A rolling seven day average will not only let us observe how the data behaves on a week-by-week basis, but it will also 'smooth' out the disparities in the data if there were any large fluctuations between days. Once the rolling seven day averages are calculated, they are plotted to a line graph along with a plot of the average of all four mobilities (**Appendix III**). From the plotted mobilities we can observe that, for starters, workplace mobility sees the largest peaks and troughs. It has the highest starting mobility, and at the most recent date it is the lowest, this can most likely be attributed to the movement to work-from-home. Furthermore, we can see that park mobility has had the biggest spike recently, which could be attributed to the population's general desire to be back outdoors and move around. Although there are slight differences in each of the mobilities, they do follow a similar pattern, and of course, the average of the four falls in the middle of all other plots. This being said, using mean mobility will be a great predictor variable for us to observe how all the mobilities affect reproduction rate without having to flood our model with variables.

IV. Creating the Best-Fit Model

In order to answer the original question; can the reproduction rate be predicted well through the use of mobility data, the best possible method would first be to create a series of multiple linear regression models to try and determine how to get the best fit model. Linear regression was chosen as the statistical method of comparison for this project because they illustrate the relationship between two variables. Therefore, four models were created to better predict the best-fit model. When generating a linear model, obviously reproduction rate was used as the response variable; in particular Reproduction 7-Day Avg. In python, in order to generate the multiple linear regression models necessary the package *linear_model* is imported from the *sklearn* library. Additionally, the p-value calculation had to be generated from a different package; *ttest_ind* had to be imported from *scipy.stats*. The first model created has four predictor variables: total_cases, new_cases, new_deaths, and Transit Mobility 7-Day Avg. Below is a table that displays some of the results from model 1:

Model 1: Summary		
Variable	**Coefficient**	**P-Value**
total_cases	-9.75e(-09)	6.88e(-58)
new_cases	6.60e(-07)	1.11e(-60)
new_deaths	-4.43e(-05)	1.39e(-70)
Transit Mobility 7-Day Avg	-3.64e(-01)	1.44e(-187)

From this, it is observed that all of our p-values are significant, or below our alpha threshold of 0.05. This being said, all of the variables used to predict the Reproduction 7-Day Avg. are significant predictors. The calculated y-intercept is 1.49, and R-Squared is equal to roughly 19.64%. This R-Squared value is particularly low for what we would like to see from a linear regression model.

To dive a bit deeper into the first model, we will now see how the model behaves with the sole predictor of Transit Mobility. By eliminating three other predictor variables we expect to see a drop in R-Squared regardless, but by how much is the question. The second model describes the relationship between Transit Mobility and the Reproduction 7-Day Avg, as shown below:

Model 2: Summary		
Variable	**Coefficient**	**P-Value**

Transit Mobility 7-Day Avg	-0.305	1.44e(-187)

Again, it is evident that our p-value is below the significance level. The R-square was also small at 17.53% and a y-intercept of 0.42. The p-value is naturally the same as it was in model 1, yet even though there were three variables removed from our model, the R-Squared value only suffered a slight dip of a little over 2%. This being said, the Transit Mobility does a better job at explaining the variance than the other predictors.

Like Transit Mobility, the other mobilties were also plotted and this was used to create another predictive model using all of them. In the third model, variables of Park Mobility 7-Day Avg, Retail Mobility 7-Day Avg, Transit Mobility 7-Day Avg, Workplace Mobility 7-Day Avg are compared with the Reproduction 7-Day Average.

Model 3: Summary		
Variable	**Coefficient**	**P-Value**
Park Mobility 7-Day Avg	0.06	1.01e(-178)
Retail Mobility 7-Day Avg	-0.34	8.54e(-179)
Transit Mobility 7-Day Avg	0.04	1.44e(-187)
Workplace Mobility 7-Day Avg	-0.09	4.34e(-178)

As seen above, the p-values are lower values than our significance model thus portraying that the linear regression model is the best statistical method to compare these variables. This model actually provides the best R-squared value of all the models at roughly 23.18%, and a y-intercept of 0.46.

Now that we have seen how a multiple linear regression model will behave with all the mobilities as separate predictor variables, now the average of all the mobility columns will be taken and put into its own column. The last model demonstrates the comparison between Mean Mobility and reproduction rate. This model was generated to show how column averaging of all mobilities affect the model.

Model 4: Summary		
Variable	**Coefficient**	**P-Value**
Mean Mobility	-0.437	7.93e-45

The p-value remains below the alpha, with the model having a slightly lower R-squared value of 18.62% and a y-intercept of 0.50. Out of all the models, the one containing all of the mobilities' seven day average has the best R-Squared value, although low, still having all significant predictors.

V. Conclusions

In all of the models created throughout the course of this analysis, the p-values for our mobility parameter; whether it be a sole mobility value or the mean mobility, were all significant. Even though the p-values indicate that mobility is a significant predictor for reproduction rate, the R-Squared values calculated are strikingly low. As stated above, R-Squared will represent the percent of the variance in the response variable explained through the predictor variables in the model. How does this occur, in other words, how do predictor variables that do not explain the variance also be significant predictors of a response variable? A simple answer to this is external factors, when considering the reason why reproduction rate spikes up to a maximum value then shoots back down can be attributed to government regulation directly impacting mobility and reproduction rate.

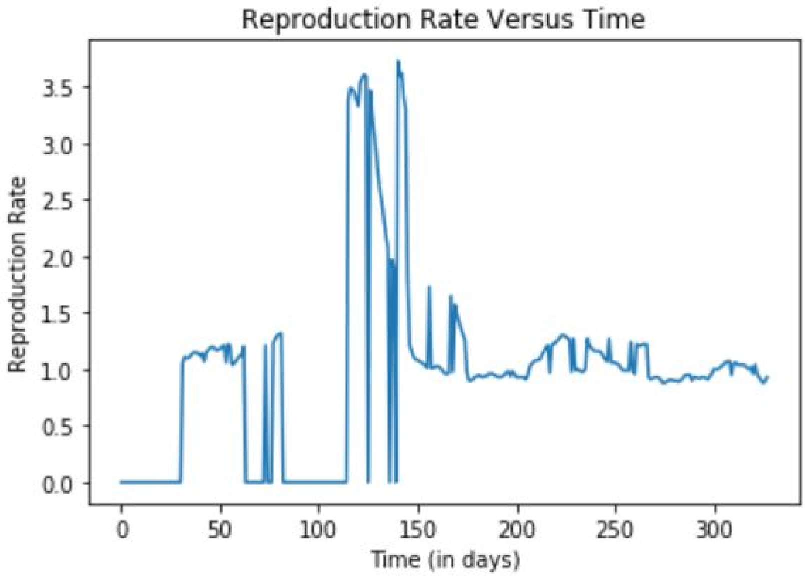
If the analysis were to be continued, a great way to expand upon this would be to create an SIR Model (Susceptible-Infected-Recovered Model). Looking at model 1, the coefficient for our new_cases variable is a positive number, showing that for every one person increase in new_cases, there will be an estimated increase in reproduction rate of 6.60e(-07). This makes intuitive sense seeing that if case counts rise, there is also most likely an increase in any given person's transmission onto more people. The opposite is true for the new_deaths variable. Where every new death recorded will result in -4.43e(-05) change in reproduction rate. This also makes sense, seeing that the population is decreasing, thus leaving less people in the pool to be infected. An SIR model would help an individual conduct an analysis of how reproduction rate will progress as our susceptible population decreases, and the infected and/or recovered increases.

From there, it would be useful to see how the mobility parameters interact with the elements of the SIR model to observe more relationships.

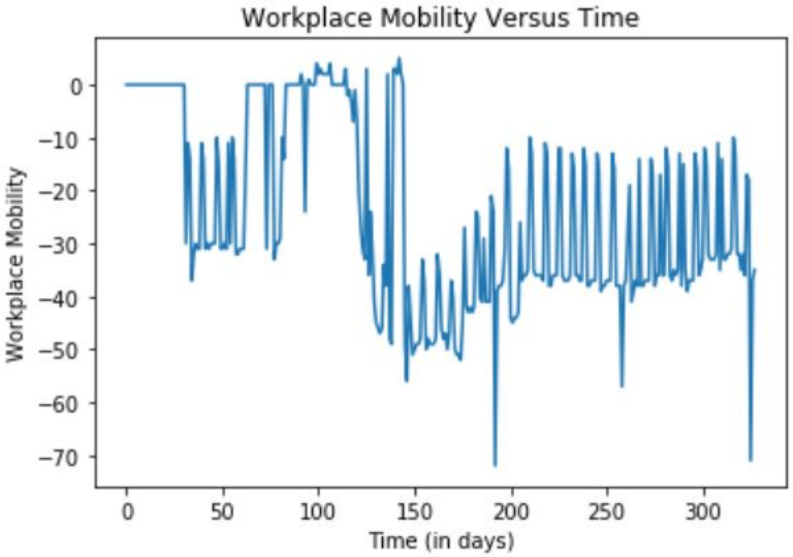
VI. Appendix

APPENDIX I: Line Plots of Reproduction Rate and Mobilities

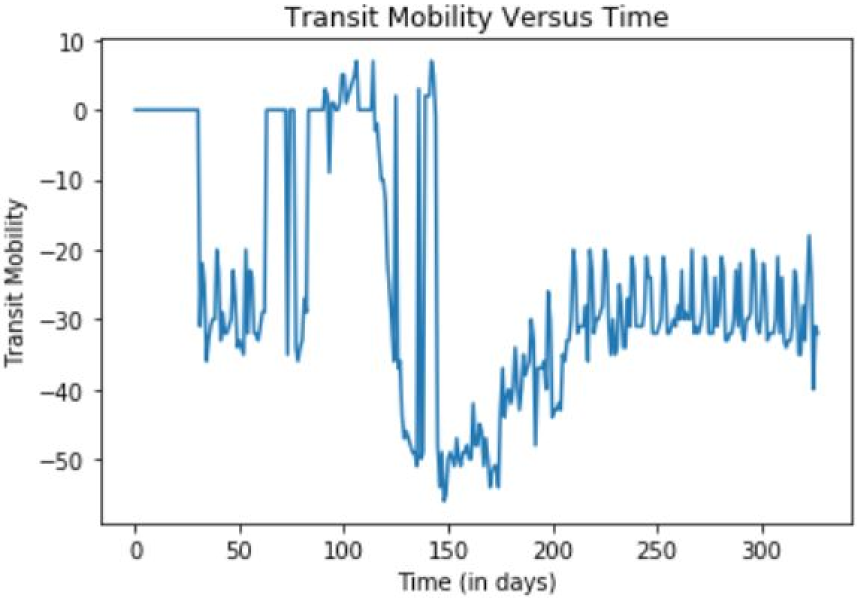
Reproduction Rate VS. Time



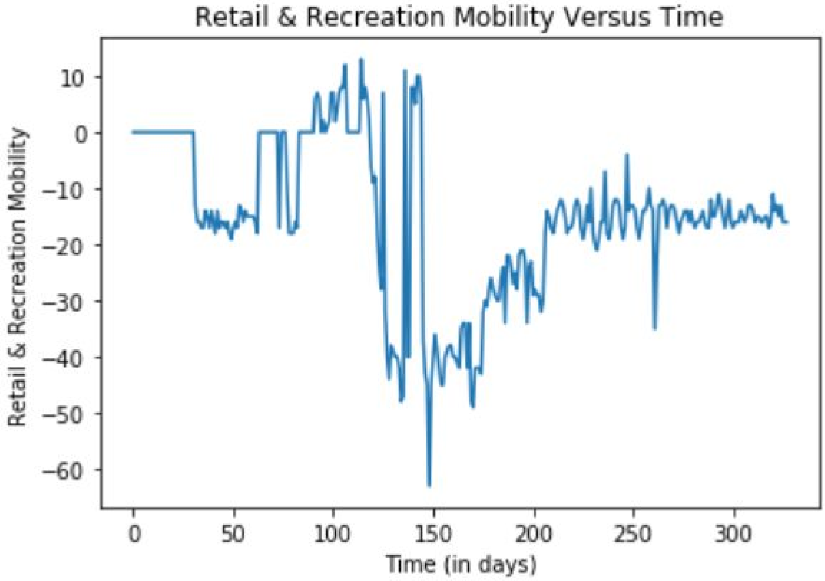
Workplace Mobility VS. Time



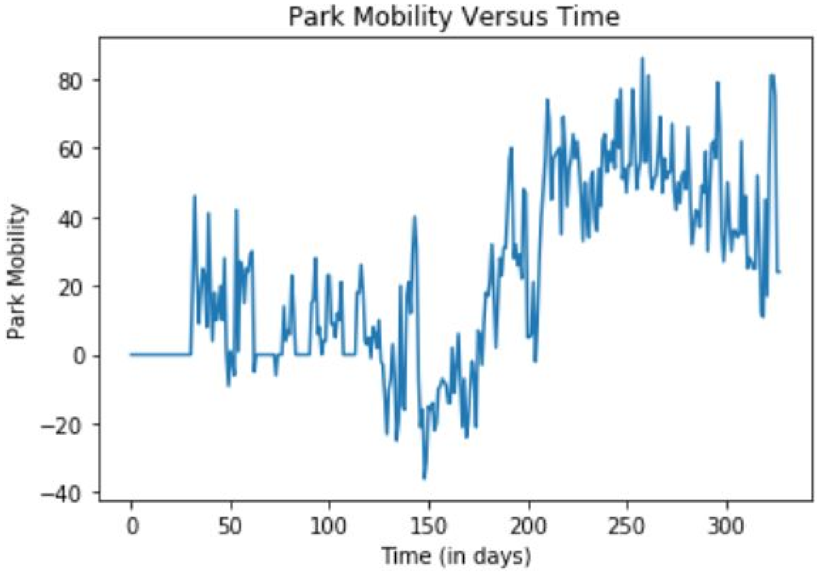
Transit Mobility VS Time



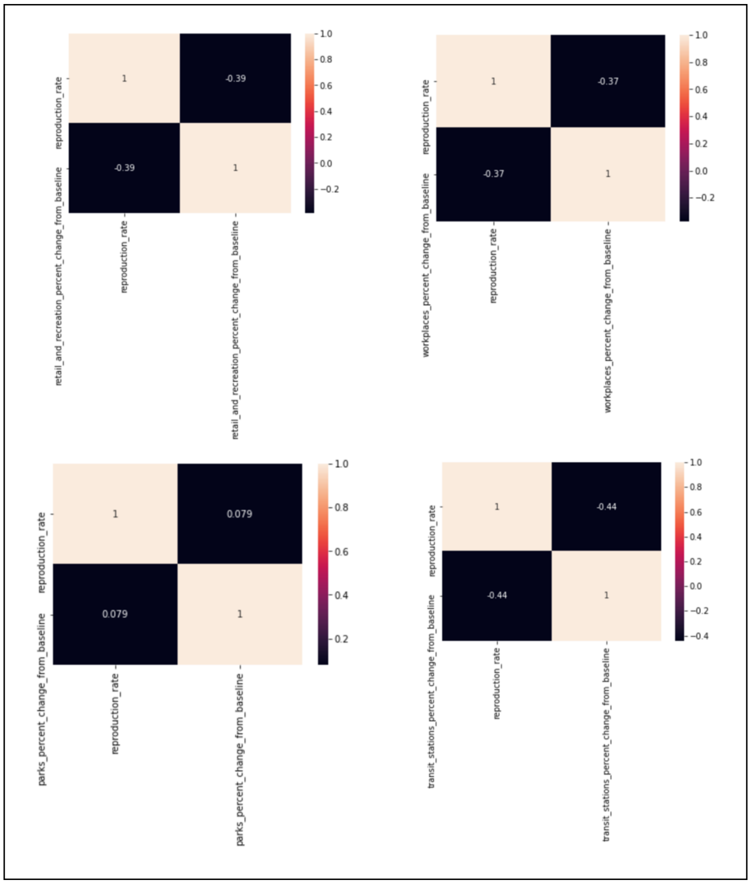
Retail & Recreation Mobility VS Time



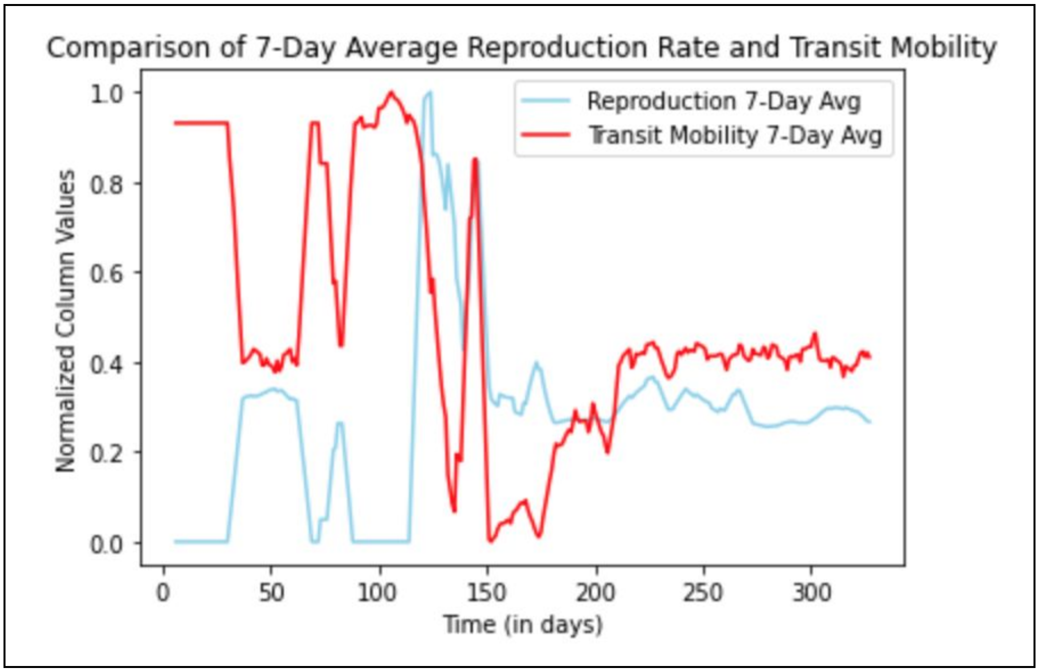
Park Mobility VS. Time



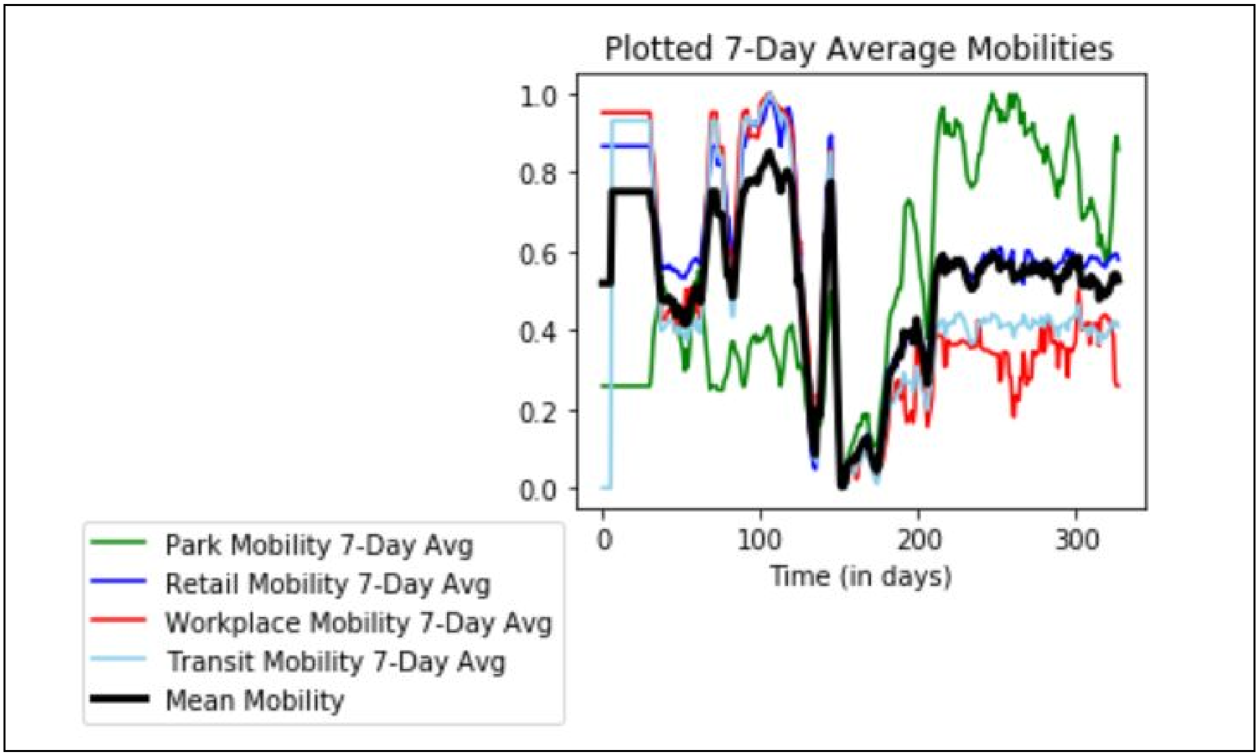
APPENDIX II: Correlation Heatmaps of Each Mobility with Reproduction Rate



APPENDIX III: Reproduction Rate Plotted with Transit Mobility (7-Day Average)



APPENDIX IV: Individual Mobilities Plotted with Average Mobility (7-Day Average)



Works Cited

Al-Raeei, Marwan. “The basic reproduction number of the new coronavirus pandemic with mortality for India, the Syrian Arab Republic, the United States, Yemen, China, France, Nigeria and Russia with different rate of cases.” *Clinical epidemiology and global health*, 10.1016/j.cegh.2020.08.005. 20 Aug. 2020, doi:10.1016/j.cegh.2020.08.005

Linka, Kevin et al. “The reproduction number of COVID-19 and its correlation with public health interventions.” *medRxiv : the preprint server for health sciences* 2020.05.01.20088047. 7 Jul. 2020, doi:10.1101/2020.05.01.20088047. Preprint.

Rath, Smita et al. “Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model.” *Diabetes & metabolic syndrome* vol. 14,5 (2020): 1467-1474. doi:10.1016/j.dsx.2020.07.045