1. Group Name

Team Mango

1. Project Participants

Franklin Geeng (fg1077)

Nikolai Sannikov (ns4463)

Yuxiang Shi (ys3700)

1. Project Description
2. Project Overview:

Based on the most recent discussions in the field of epidemiology, the most advertised approach to contain the spread of COVID-19 community transmission is isolation of its members in their residences by reducing non-essential social interactions. For several weeks this approach has been adopted by many of the countries including the United States, and some conclusions can be drawn. In this work our team will assess the current outcomes of social isolation in the United States and develop a predictive model.

The most obvious indicator that people are staying at their homes is a change in their movement patterns. Therefore, the overall goal of this project is to explore the impact of people’s movements in the United States on infections and deaths from the coronavirus. Our analysis involves the following four subtasks:

* Identify patterns in change in mobility (Franklin Geeng)
* Determine whether there is a correlation between change in mobility and change in infection rate (Franklin Geeng)
* Measure how dangerous each county is by using mobility data

(Yuxiang Shi)

* Determine Spread Patterns Determined By Socioeconomic Standing (Nikolai Sannikov)

The impacts will be assessed primarily using time-series data of infections and deaths compiled by the New York Times and time-series data on mobility as measured by cell phone locations from the Descartes Lab. Both datasets have county-level spatial resolution and day-level temporal resolution. Moreover, both datasets contain a unique state-county fips number for geospatial indexing and joining of the datasets.

Other supporting datasets include U.S. Census population data to compute proportions of county populations that were infected and USDA data on unemployment and median household income.

The subtasks and the proposed methodologies for addressing them are discussed in more detail below.

1. Identify patterns in change in mobility (Franklin Geeng)

Anecdotally, different regions and different subpopulations of the United States began social isolation at different times. College students from Florida and Texas were infamously recorded partying in late March as east and west coast communities were locking down. Even as of mid-April, some Michigan residents are protesting the lockdown.

The goal of this task is to identify patterns in change in mobility. This is accomplished using the Descartes Labs mobility index m50\_index time series of each county using the k-means clustering algorithm implemented in the Python sklearn library.

To compute the m50\_index, Descartes Labs sampled a small proportion of people present in a county each day and first calculated the raw maximum distance in a convex hull of cell phone locations for each sampled person. These raw distances were scaled by the median of such a value for weekdays in the period of 2/17 through 3/7 and the median percentage value of the samples in a county was reported as m50\_index.[[1]](#footnote-1) The index includes travel outside of the county and sampled persons who leave a county are still included in the computation of m50\_index.

For each county, the m50\_index on each day from 3/1 through 4/23 is treated as a feature on which counties were clustered. The number of clusters is varied to observe any clusters that emerge. The cluster centroids and members are plotted to see the emerging mobility patterns as well as how well the cluster centroids characterize the behavior of the members. The observed patterns feed into the analysis to observe any correlation between a change in mobility and a change in infection rates.

1. Determine whether there is a correlation between change in mobility and change in infection rate (Franklin Geeng)

The Descartes Labs mobility index m50\_index is taken as a proxy for social contact. It is assumed that the amount of social contact increases with an increasing m50\_index. Thus, it is hypothesized that after a period of decrease in a county’s m50\_index, there will be a corresponding decrease in the infection rate of that county. This hypothesis is tested by comparing the average change in infection rate of each county over two weeks following two weeks of approximately normal average normal mobility and following two weeks of reduced mobility. The change in infection rate is calculated as the instantaneous second derivative of the NYTimes cumulative infection count time series dataset, otherwise referred to as the ‘infection acceleration.’

1. Measure how dangerous each county is by using mobility data (Yuxiang Shi)

Instinctively, a county with more confirmed cases may be more dangerous than a county with less confirmed cases. However, if we all stayed at home, the possibility that we get infected are both low. So we can measure this in another perspective, since people are infected by getting connection with others, by using mobility data, we have the distance people travel in each county(m50) and the percent of normal m50 in the region(m50\_index), for coronavirus cases data, we have the cumulative confirmed cases (cases) and deaths(deaths) each day. Divide m50 by cumulative confirmed cases, we will get a constant P which represents every P miles people in that county travel, there will be one more confirmed case due to the travel.

The reason why we are using cumulative data is that the distance people travel in one day will not have an immediate impact on the confirmed cases that day. But it will contribute to the future growing cases.

1. Analyze how socio-economic parameters determine the patterns of COVID-19 spread (Nikolai Sannikov)

As it is often discussed in the media, COVID-19 spread patterns can be drastically different: explosive growth is observed in some regions, while other ones maintain a slow and steady growth. Some regions even get barely affected by COVID-19. Besides mobility, socio-economic factors may be determining which pattern of spread a county will follow.

It is a trivial task to analyze, how well a county is doing in terms of economics, the most commonly used metrics are unemployment and median income. Counties can be easily clustered based on their economic performance.

A more challenging task is to analyze whether there is a distinct pattern within these clusters. Unfortunately, raw COVID-19 spread has a very high parameter space: the most trivial approach is to treat a number of new cases each day as an independent variable. In this case, the observations will be too noisy to draw any conclusions and will cause difficulties in deriving an averaged pattern.

In this work, this problem is proposed to be solved by switching from a space of raw observations to a space of polynomial approximation parameters. In this case a fixed-degree polynomial is used to approximate the number of daily new COVID-19 cases, what reduces the size of a space for each county from each day to 4 parameter values for a degree-3 polynomial function. Besides smoothing for a better perception by a human observer, values of this parameter space can be easily averaged for each cluster, resulting in an averaged approximation for each cluster.

In this case, the hypothesis stated above can be assessed in 2 ways: if it is false, socio-economic clusters averaged functions will not differ. If it is true, these functions will be clearly different. As a supporting experiment, a few counties can be excluded from deriving a cluster-based average. If the cluster-based average for the corresponding cluster will reflect the major trends in the recorded COVID-19 spread data, the hypothesis can be accepted.

1. Datasets

The datasets that will be used are:

|  |
| --- |
| New York Times Covid-19 Cases  <https://github.com/nytimes/covid-19-data>  Description:  The New York Times is releasing a series of data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time. We are compiling this time series data from state and local governments and health departments in an attempt to provide a complete record of the ongoing outbreak. Since late January, The Times has tracked cases of coronavirus in real time as they were identified after testing. Because of the widespread shortage of testing, however, the data is necessarily limited in the picture it presents of the outbreak. We have used this data to power our maps and reporting tracking the outbreak, and it is now being made available to the public in response to requests from researchers, scientists and government officials who would like access to the data to better understand the outbreak. |
| US Mobility Data  <https://github.com/descarteslabs/DL-COVID-19>  Description:  Mobility statistics (representing the distance a typical member of a given population moves in a day) at the US admin1 (state) and admin2 (county) level. A technical report describing the motivation behind this work with methodology and definitions is available at descarteslabs.com/mobility-v097. We intend to update the data in this repository regularly. |
| US Country-Level Population Data, 2010-2019  <https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html#par_textimage_70769902>  Description:  US county level populations from 2010 through 2019. Includes other information such as birth, death, and immigration, and migration rates |
| USDA [Unemployment and median household income for the U.S., States, and counties, 2007-18](https://www.ers.usda.gov/webdocs/DataFiles/48747/Unemployment.xls?v=2564.4)  <https://www.ers.usda.gov/webdocs/DataFiles/48747/Unemployment.xls?v=2564.4>  Description :  This dataset is a snapshot of economic performance of counties in a period from 2007 to 2018. The most interesting metrics are unemployment, median household income, and percentage of median household income of the corresponding state. |

1. Data Analysis
2. Github Repository

A Github repository containing the raw and cleaned datasets and the Jupyter notebooks for data cleaning and analysis is located here: <https://github.com/DDDMingo/Big-Data-Final-Project-Team-Mango>

1. Identify patterns in change in mobility (Franklin Geeng)[[2]](#footnote-2)

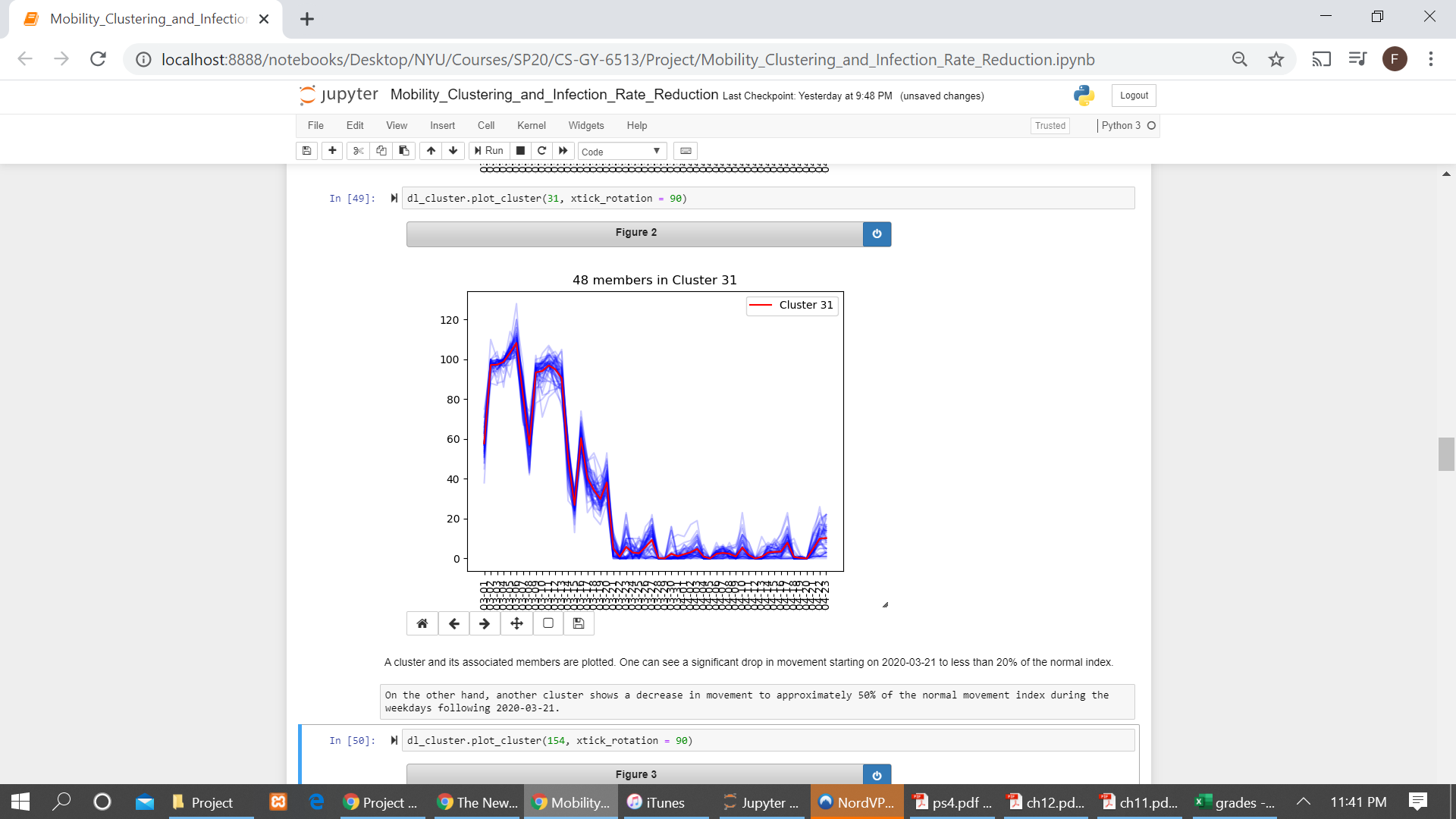
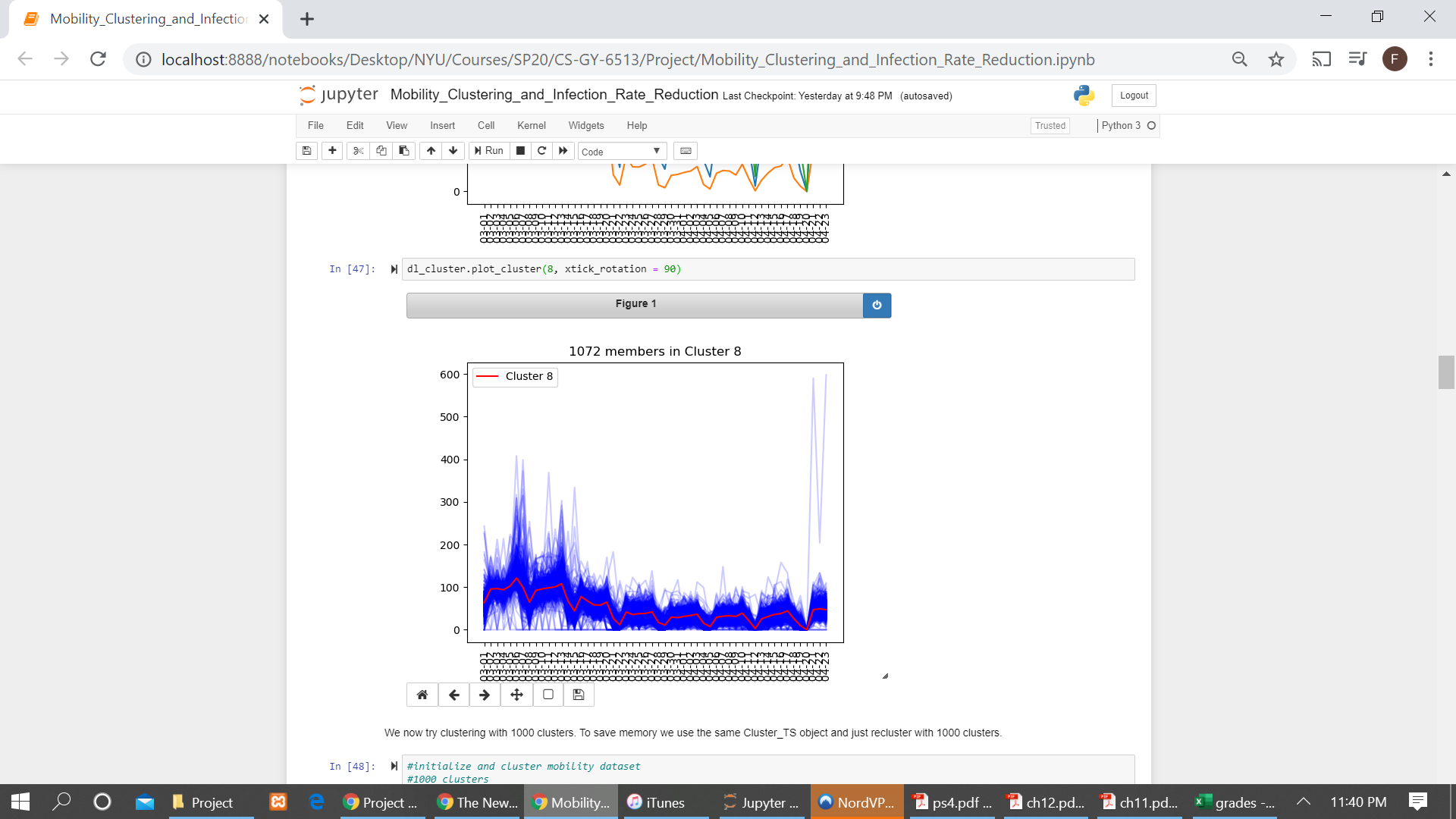
Before proceeding to look for any correlation between mobility and infection, an exploratory analysis of the mobility data was performed to understand any patterns that are present. This analysis helps to select periods in the months of March and April to examine change in infection rate, to be discussed in the next section.

The data examined was the Descartes Labs mobility index m50\_index data for the period of 3/1 through 4/23. The m50\_index was computed from distances measured by the anonymized cell phone locations of samples of persons. The dataset provides the m50\_index for 2722 counties in the United States on each day from 3/1 through 4/23, excluding 4/20, which was missing from the dataset. Zeros were added for all counties for 4/20 to account for the missing values. As discussed above, the m50\_index is the median of the sampled raw maximum convex hull distance normalized by the median such value for weekdays from 2/17 through 3/7, the ‘normal’ mobility period. This is actually a potential issue with the data in the usage for this analysis because one can expect that mobility patterns on weekends will differ from weekday patterns and yet the weekend raw distances are still normalized by the weekday value, so we cannot observe if there was a change in weekend mobility patterns. Moreover, due to expected seasonal or monthly changes in mobility patterns, a more appropriate normalizing factor might be the median raw distance from the same day or month of the previous several years. Unfortunately, this data is not available. Ultimately, the Descartes Labs is taken at face value as reliable for the proceeding analyses.

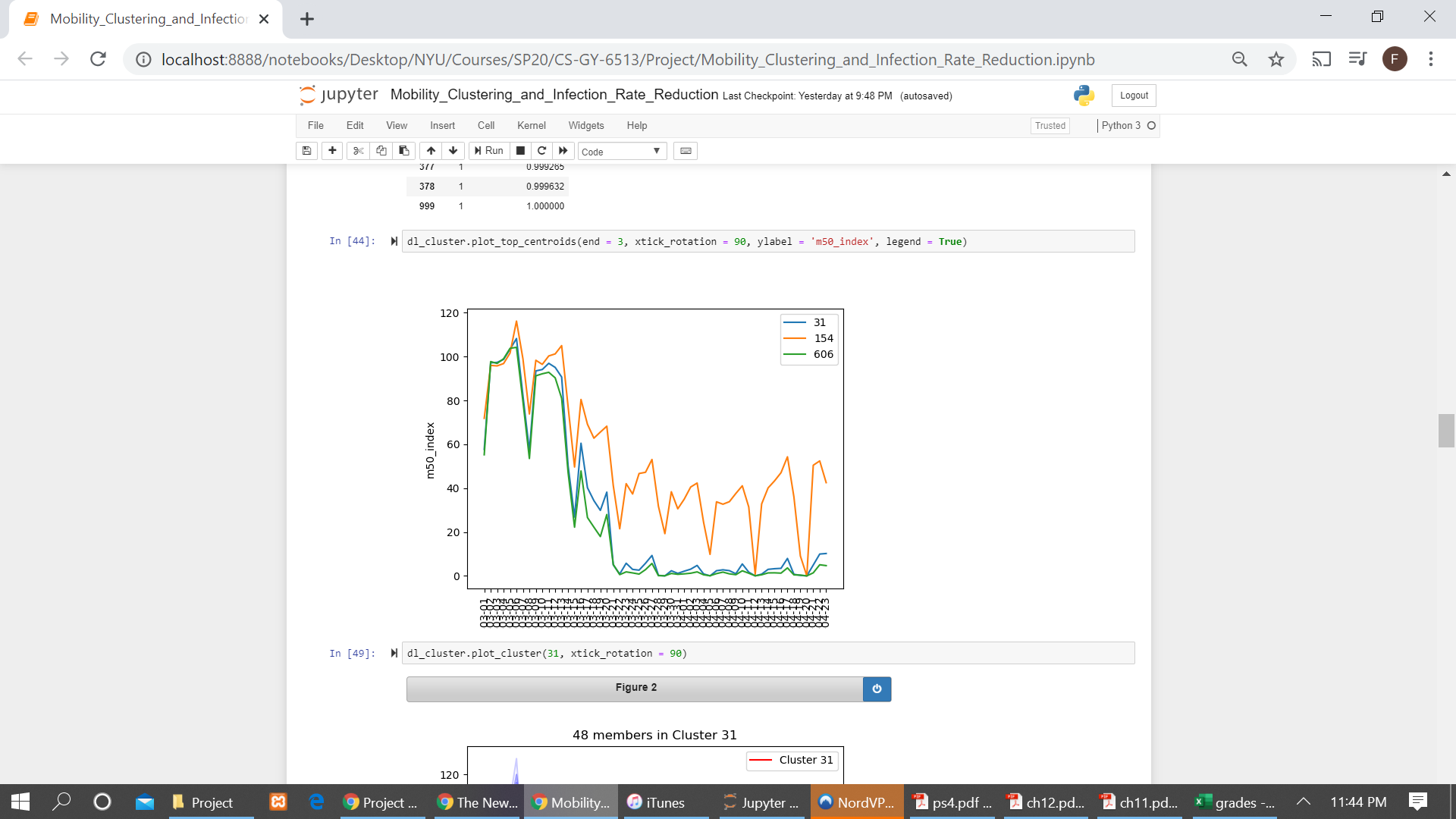
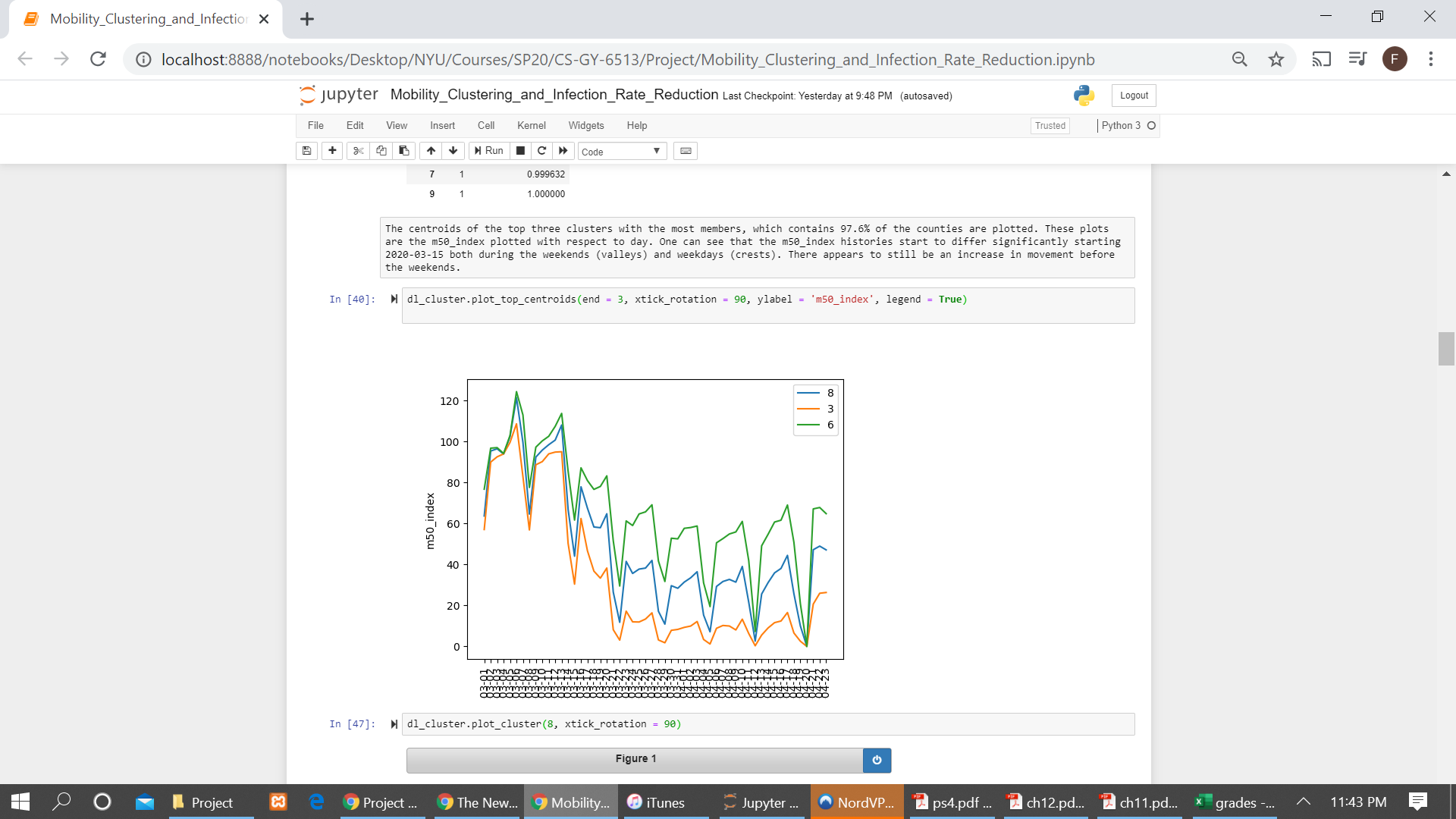
The method for identifying mobility patterns is to cluster counties using the k-means clustering algorithm implemented in the sklearn Python library with the features being the daily m50\_index values. The k-means clusters points by iteratively assigning cluster membership of points to their closest cluster centroid and then recomputing the cluster centroids based on the new cluster membership.

Many combinations of k-means clustering parameters including the number of clusters, maximum number of iterations, and convergence tolerance were tried to determine the number-of-clusters parameter where true clusters emerge. The Cluster\_TS was implemented to maximize the amount of code reuse and assist with visualizing and exploring the clusters. A Cluster\_TS object contains the raw data, cluster centroids, cluster assignments, and other data relevant to the analysis. The Cluster\_TS class also contains useful methods for plotting the clusters with the greatest number of members and all the members of a cluster.

The number-of-clusters parameter was varied from 10 to 2000, with the intraclass variance decreasing with increasing number-of-clusters parameter. In other words, members of classes were more similar to the cluster centroid as the number of clusters was increased. This is illustrated below, with plots of cluster members for a single cluster for number-of-cluster equal to 10 on the left and 1000 on the right.



The numbers of counties in each county can also be seen to decrease with an increasing number of clusters. Finally, the centroids of the most populated clusters also change, as shown below. Again, the number of clusters is 10 and 1000 on the left and right, respectively. The clusters on the left contain approximately 98% of the counties whereas the clusters on the right only contain approximately 4% of counties. The choice of the number of clusters is somewhat subjective. 1000 clusters was selected as the ‘correct’ number of clusters because of the closer similarity of cluster members while maintaining a relatively small but not miniscule number of clusters with more than one member (200/1000).



Several insights can be gleaned from this analysis. First of all, the county-level mobility time series are all very different, so much so that only a small number of coherent clusters form after running the k-means algorithm. In fact, when clustering with 1000 clusters, there are 800 clusters with only a single member.

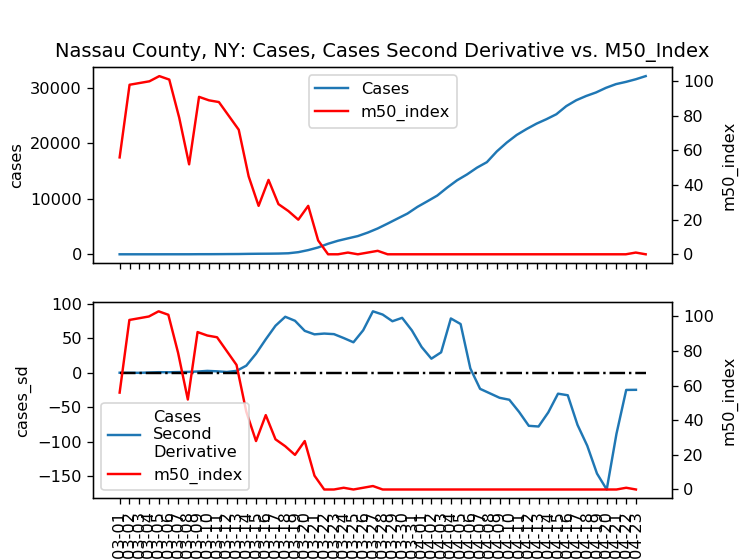
Second of all, of the clusters that emerge, the differentiating factor of the clusters is the mobility pattern following 3/14/2020, with some clusters of counties significantly reducing mobility starting around 3/23/2020. This large drop in mobility corresponds to the spate of statewide stay-at-home orders issued starting around 3/23/2020[[3]](#footnote-3). The correlation of this drop in mobility with change in infection rate will be investigated in the next section.

1. Determine whether there is a correlation between change in mobility and change in infection rate (Franklin Geeng)[[4]](#footnote-4)

An important question with regards to the coronavirus pandemic response is regarding how effective social distancing and isolation has been at reducing the spread of infections. Social interactions are difficult to accurately measure without an Orwellian surveillance system that tracks everyone at every moment of the day. In lieu of access to such a system, the Descartes Labs mobility dataset is used as a proxy measurement of the amount of social interaction occurring in counties with the assumption that a decrease in mobility implies a decrease in social interactions. The mobility dataset is admittedly imperfect for quantifying social interactions because it is possible for individuals to travel large distances without having any social interactions as is the case with, for example, driving. It is also possible for individuals to stop traveling large distances only to have many social interactions near their home, for example with neighbors.

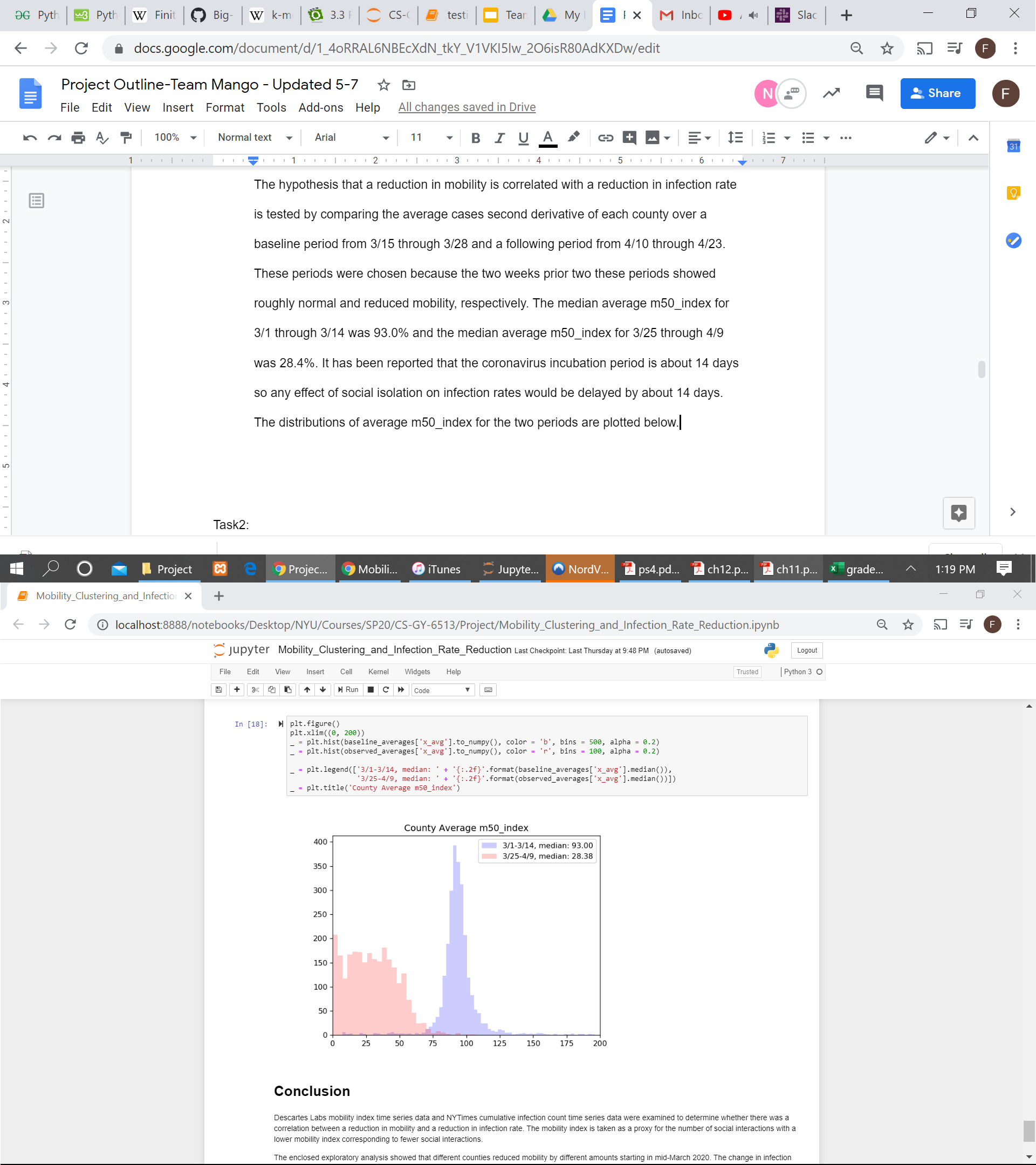
With these issues in mind, infection rates were observed to be decreasing following a period of reduced mobility. The quantity for observing a change in infection rate is the instantaneous second derivative of the NYTimes cumulative infection counts dataset. The instantaneous second derivative in this context represents the acceleration of the infection on a given day i.e. the rate at which the number of new cases per day changes on a given day. A negative second derivative means that the number of new cases per day is decreasing, the desired outcome of the lockdown across the United States.

To illustrate the second derivative of cases, we first focus on Nassau county, NY, which had a significant number of infections numbering at approximately 30,000 as of 4/23/2020. Plotting the number of cases against the m50\_index, we see the number of new cases per day increased starting on 3/18 but then decreasing starting around 4/7, that is, the slope of the cases series becomes progressively shallower with each day starting around 4/7. This is confirmed by the periods with positive and negative second derivatives of cases starting on those dates. One can also see that the period with a negative second derivative of cases starts approximately two weeks after the m50\_index drops to nearly zero starting around 3/22. This is in line with the hypothesis that infection rates should decrease following a period of social isolation.

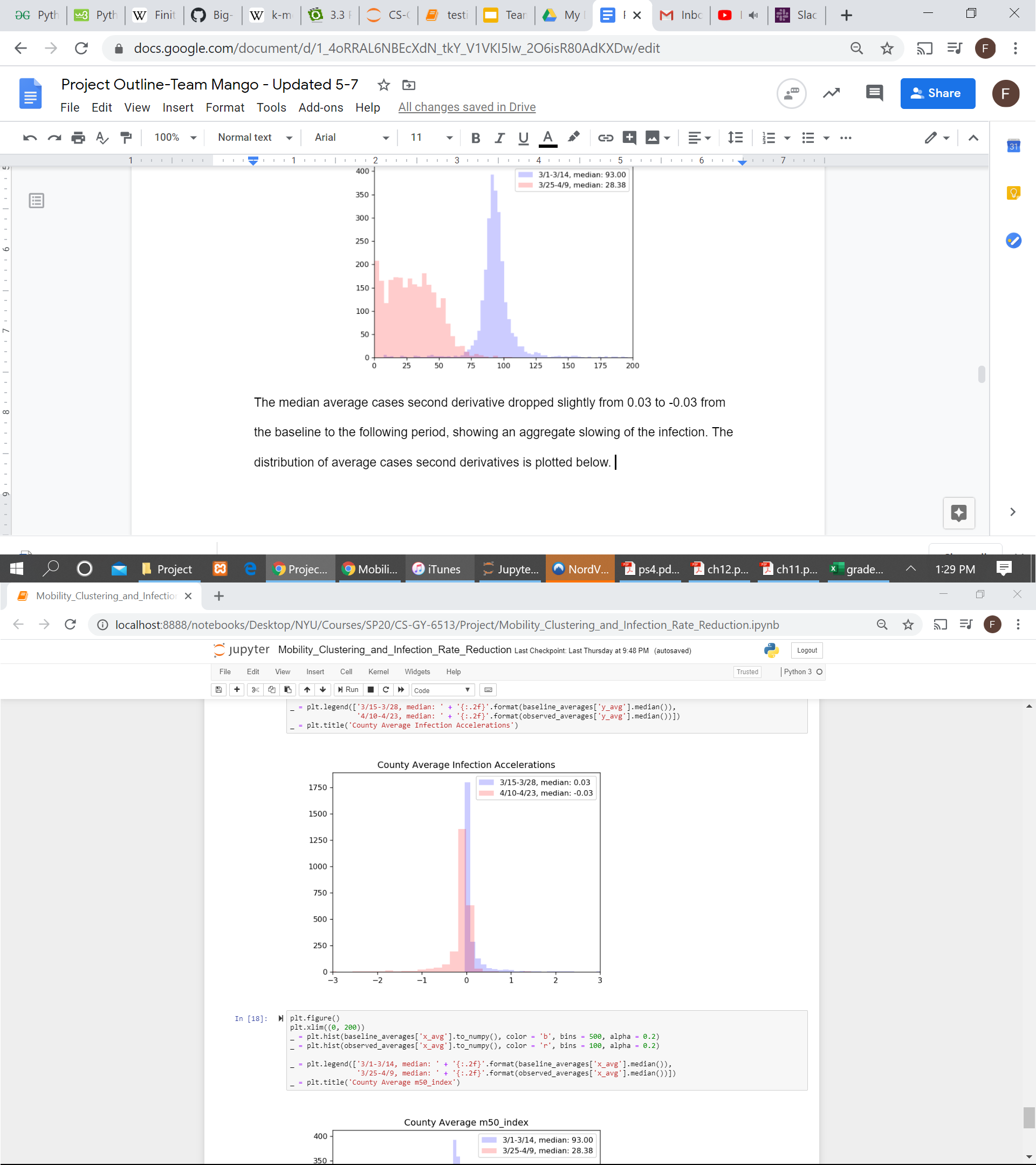


To calculate the instantaneous second derivative of the infection counts, the gradient function in the Python numpy library is applied twice to the smoothed cumulative infection count time series for each county in the NYTimes dataset. The infection count time series must be smoothed before the second derivative calculation because of the NYTimes infection counts are not always updated daily and can contain revisions that add or remove cases, which results in artificially high instantaneous case second derivatives that in turn will artificially affect the average over a period. What is returned is another time series where for each county and on each day there is a value for the instantaneous second derivative of the infection counts. The numpy gradient function employs the central difference formulation, which leads to reduced error compared to forward and backward difference formulations. Continuing with the theme of code reuse, several functions to calculate the moving average, set the range of dates of interest, and calculate the second derivative are defined.

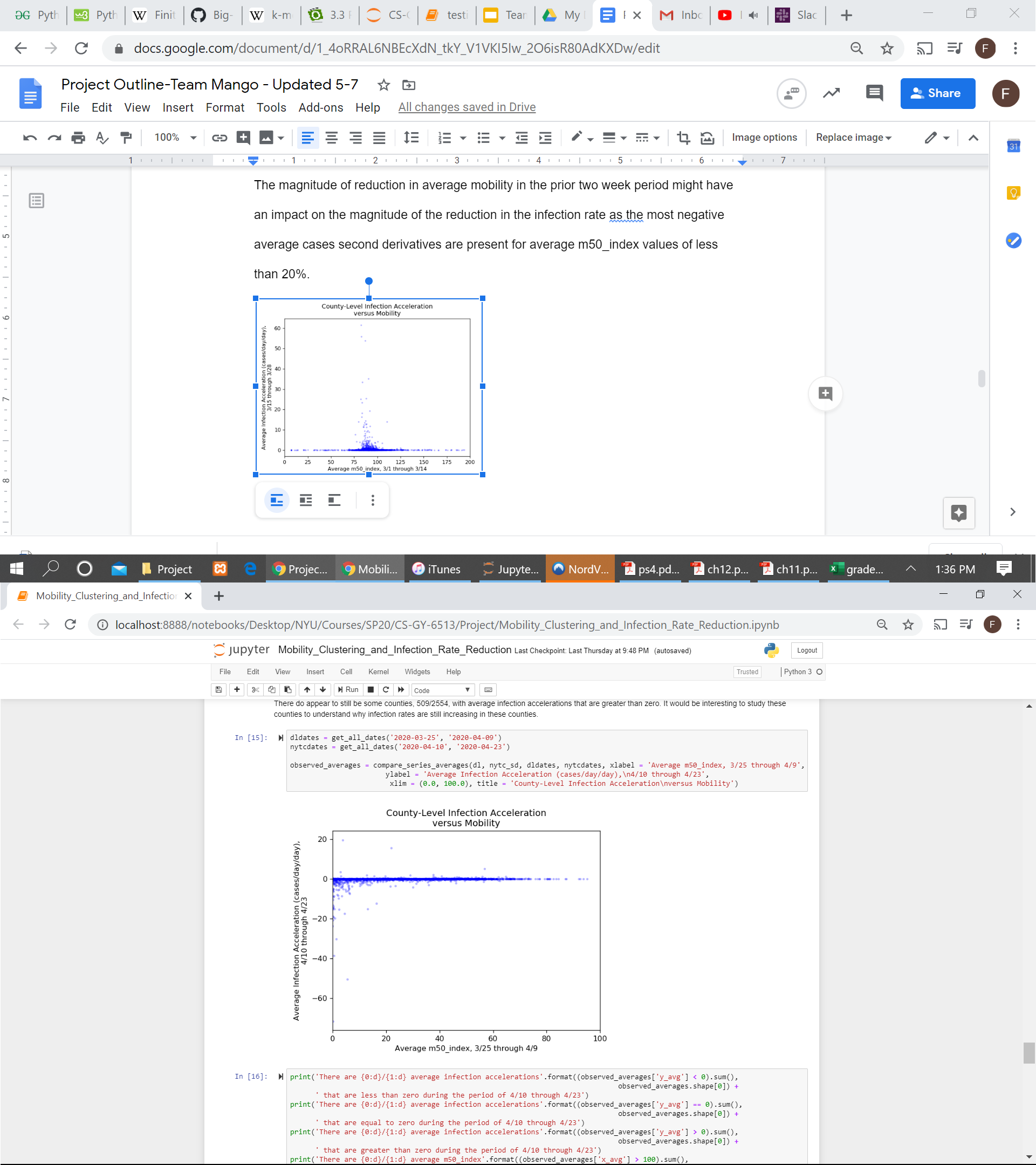
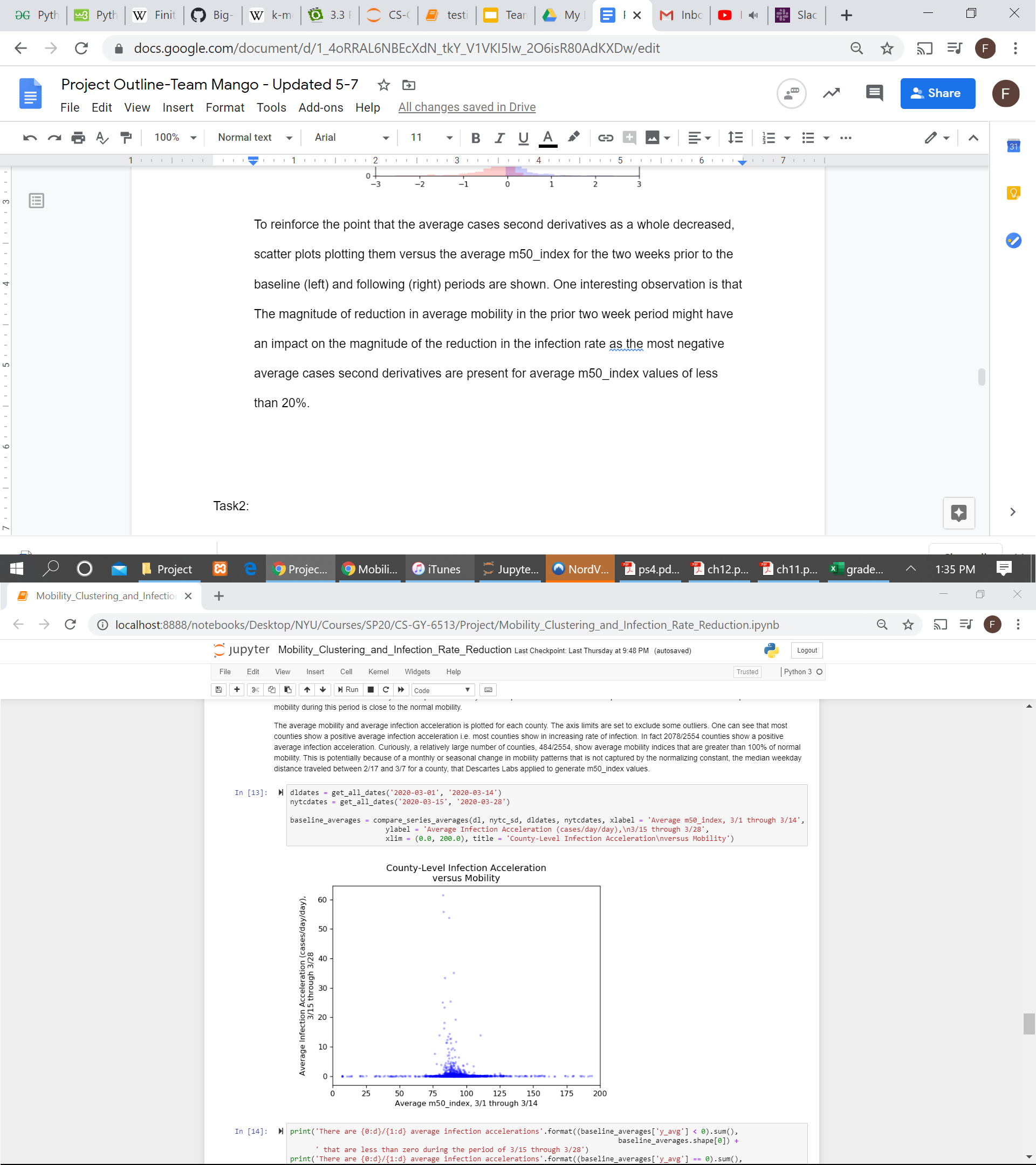
The hypothesis that a reduction in mobility is correlated with a reduction in infection rate is tested by comparing the average cases second derivative of each county over a baseline period from 3/15 through 3/28 and a following period from 4/10 through 4/23. These periods were chosen because the two weeks prior to these two periods showed roughly normal and reduced mobility, respectively. The median average m50\_index for 3/1 through 3/14 was 93.0% and the median average m50\_index for 3/25 through 4/9 was 28.4%. It has been reported that the coronavirus incubation period is about 14 days so any effect of social isolation on infection rates would be delayed by about 14 days. The distributions of average m50\_index for the two periods are plotted below.



The median average cases second derivative dropped slightly from 0.03 to -0.03 from the baseline to the following period, showing an aggregate slowing of the infection. The distribution of average cases second derivatives is plotted below, focusing on the region with greatest mass of points. Only 46/2554 counties show negative average cases second derivatives for the baseline period versus 1885/2554 during the following period.



To reinforce the point that the average cases second derivatives as a whole decreased, scatter plots plotting them versus the average m50\_index for the two weeks prior to the baseline (left) and following (right) periods are shown. One interesting observation is that the magnitude of reduction in average mobility in the prior two week period might have an impact on the magnitude of the reduction in the infection rate as the most negative average cases second derivatives are present for average m50\_index values of less than 20%. Another interesting observation is that despite the overall nationwide lockdown, there appear to be some counties with accelerating infections (average second derivative is greater than zero) during the period of 4/10 through 4/23.



In summary, the infection cases second derivative was computed as a metric for measuring the rate of change of infection rates to determine whether the outbreak was accelerating or decreasing. The average cases second derivative was shown to decrease from mostly positive values during a baseline period with normal prior mobility to mostly negative values for a following period with reduced prior mobility, indicating a correlation between the nationwide reduction in mobility and the reduction in the number of new cases per day. Moreover, a greater average reduction in infection rate was observed for counties with a greater average reduction in mobility. Future studies may seek to understand whether there is a causal relationship between reduced mobility and reduced infection rate (to rule out other explanations such as the establishment of a herd immunity) and seek to understand whether it is the case that infections in some counties are still accelerating despite the nationwide lockdown.

1. Measure how dangerous each county is by using mobility data(Yuxiang Shi) [[5]](#footnote-5)

Procedure:

1. We store confirmed cases and mobility data in separate dictionaries. And sum up the m50 parameter if they share the same fips number.

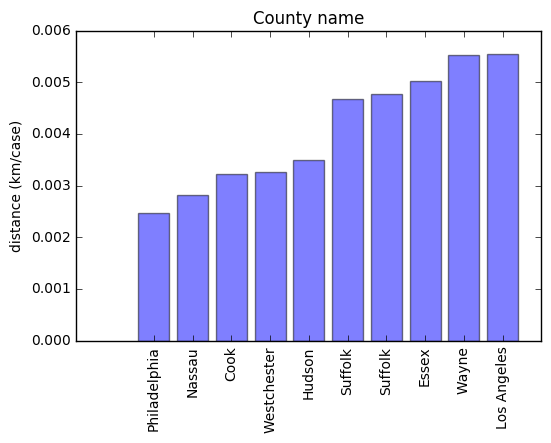
2. We choose counties that exist in both case dictionaries and m50 dictionaries in case one of them doesn't have corresponding data. And then, we divide the sum of m50 by the number of confirmed cases, therefore, for every county we get a constant P, which represents P km/case.

3. We use this computed constant P to represent how dangerous a county might be, since if P gets smaller, it means that every P kilometers people in that county move(it will be every person in that county), there will be one more confirmed case caused by the movement.

Result:

As we can expect, the result will also vary from county to county since we have large differences in confirmed numbers for different counties. Here is the distribution of computed the number of counties is 2554, the mean of P is 50.708342 (km/case), and std will be 248.071232(km/case), the minimum value of P is 0.002478 (km/case), and the maximum value will be 11057.744000 (km/case). P at 25% position is 2.110413 (km/case), P at 50% position is 10.777292 (km/case), P at 75% position is 41.605350 (km/case).

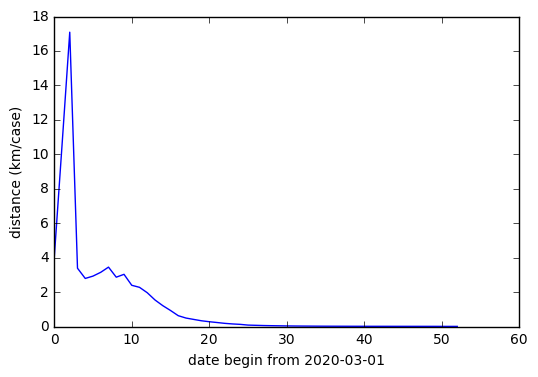
We can also take a look at the top 10 counties that have the smallest value of P.



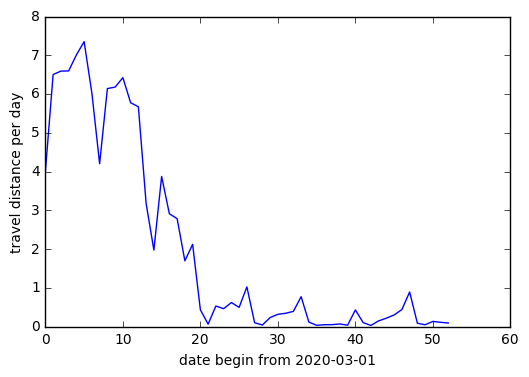
The P seems to be a little too small here, if we think about it. It’s because of the lack of mobility data.

The cases were recorded when every county had its first case, like LA, its first case appears on 1/26. However the mobility data began from 3/26. And movements from 1/26 to 3/26 also contributed to the growth of later growing confirmed cases.

We also look at Los Angeles to see how P changes from time to time.

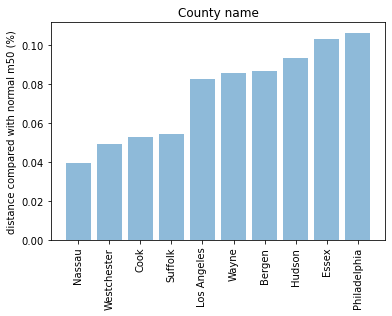


As we can see here, with time going by, P becomes smaller in general.



Here is the graph of travel distance made by people in Los Angeles, the growing distance at first may explain why P becomes larger at first.

If we think further, the transportation methods also vary from county to county, people who live in cities tend to travel less if they are using public transportation and people who live in countries who travel via their own cars probably move more than people who live in cities. So, we also explored the m50\_index which represents the distance compared with normal m50 in the region. We use the similar procedure to sum the m50\_index up and then divide it by the number of confirmed cases.



And the top 10 counties that have the smallest m50\_index/cases are different from what we got previously. There may be some reasons like even if some people traveled more than others, as long as they stayed in their own cars, it will not contribute to the spread of the virus.

# COVID-19 Spread Patterns Determined By Socioeconomic Standing (Nikolai Sannikov)[[6]](#footnote-6)

The goal of this part of the project is to analyze the COVID-19 spread patterns in various counties and to assess how the socio-economic standing may determine these patterns. The main hypothesis is that counties similar in their economic and social background prior to the pandemic, are likely to have similar COVID-19 spread patterns.

Based on this hypothesis, the following approach is suggested:

1. Retrieve data regarding the growth of the number of infected people for some set of counties

2. Approximate the COVID-19 incidence growth rate with a fixed-degree polynomial function (i.e.: approximate every county growth rate with a degree-3 polynomial)

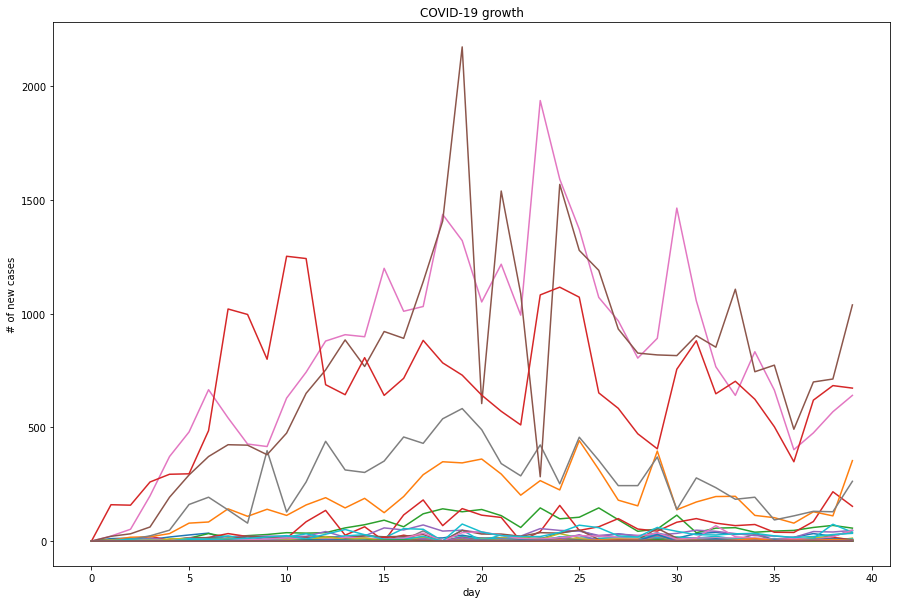
3. Split counties in several clusters basing on their properties

4. Find an average approximation for samples in each cluster

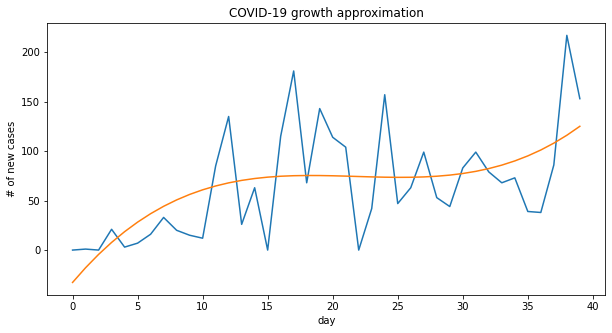
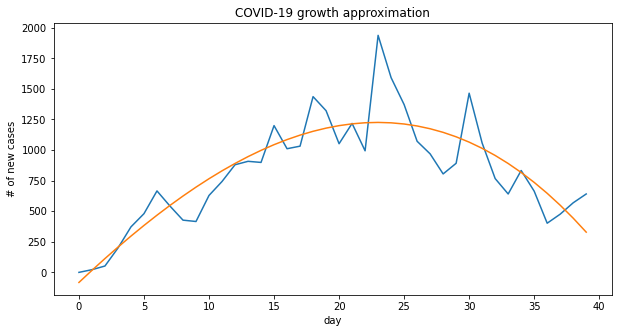
As a result, there will be an averaged function that shows the general behavior of COVID-19 spread in this cluster of counties with less stress given to an absolute number of sick people. If socio-economic data is not sufficient to determine the COVID-19 spread patterns, cluster-averaged approximation functions will not differ significantly and the hypothesis can be rejected.

This hypothesis can be declared to be reasonable if the clustered counties will have recognizable similar patterns in the growth rate approximation, while counties belonging to the different clusters will show different spread patterns. Each previously unknown county can be assigned to a cluster of best fit. After that the way pandemic unfolds is likely to be somewhat similar to the cluster-specific averaged approximation.

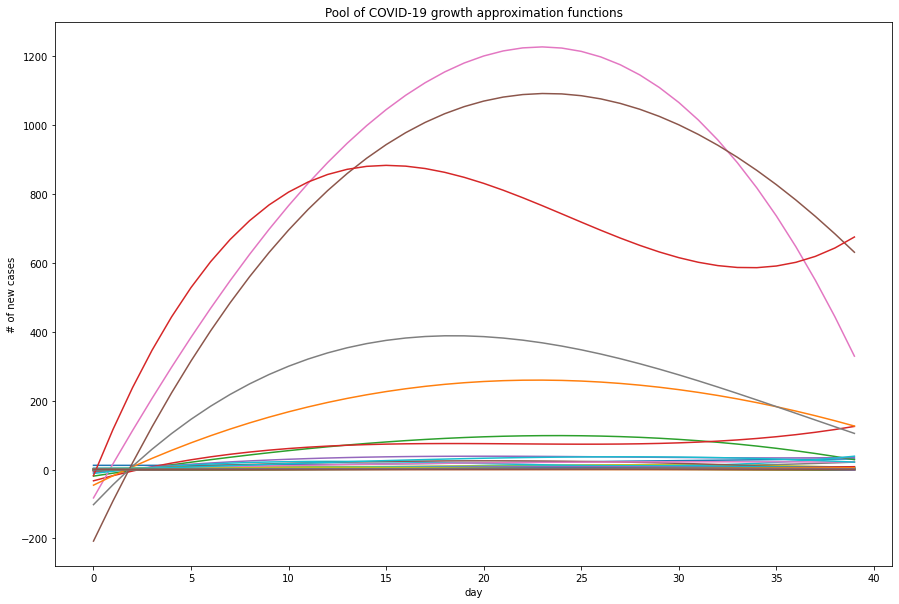
This approach can be well illustrated using counties in the state of New York. The problems inherent to the raw data can be clearly seen in the chart below:



While it is clear that the counties have different spread patterns, a more detailed analysis is impeded by a high noisiness of the raw data. Finding an averaged values for similar counties will keep the noise and general patterns may be lost. As proposed in methodology, these observations were approximated using a degree-3 polynomial. As it can be seen for the 3 specific counties, this model successfully smoothens the data and preserves the general trends:



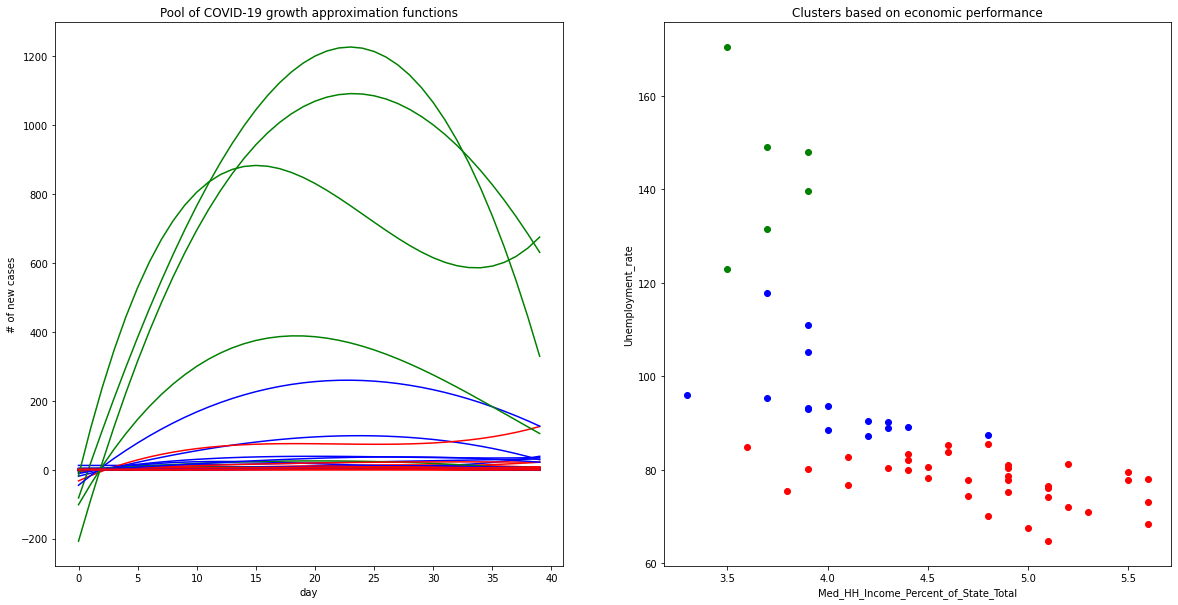
Therefore, a pool of functional approximations was created, one for each county:



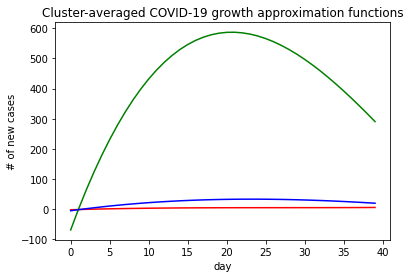
Now it is time to cluster the counties to observe how socio-economic standing dictates the pattern of COVID-19 spread. The parameters used in the clustering were:

* Median Household Income - evaluates general level of life
* Unemployment rate - estimates the percentage of poor people with a limited access to resources
* Med HH Income Percent of State Total - shows how important this county is on a state level

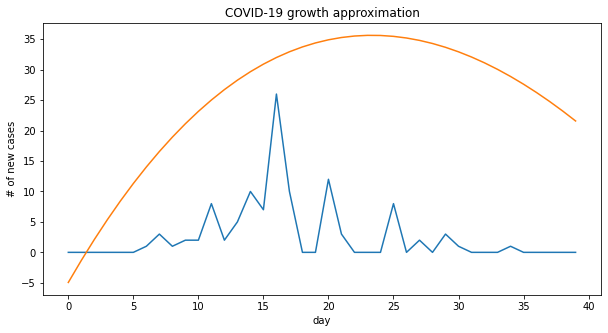
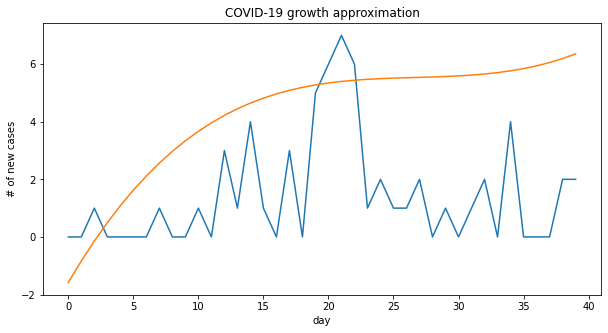
These parameters reflect general socio-economic standing of a county and demonstrate the degree of involvement of a county in a state economy. Clustering of only NY counties show the following result:



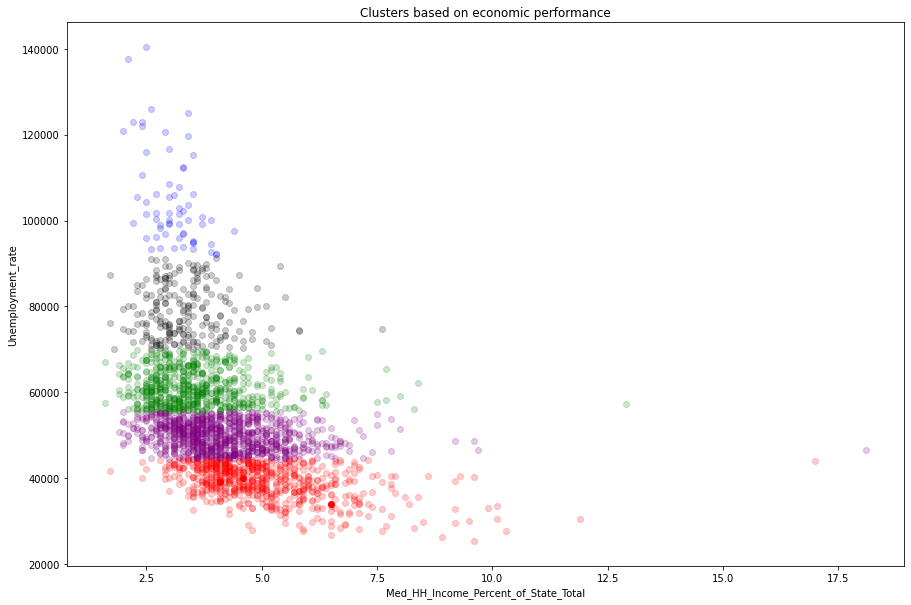
Clearly, the 3 clusters used in this analysis have a drastically different spread patterns on a fine-grain level. Cluster-based average models were constructed by taking a mean value of each polynomial parameter. As a result, the NY clusters have the following averaged models:



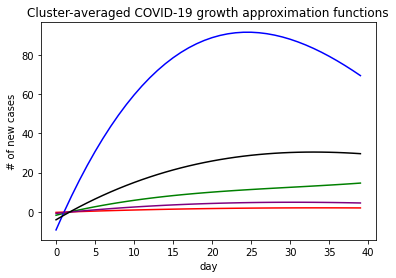
As it can be seen, socio-economic clusters demonstrate a different behavior in the averaged models. If the clusters are constructed without several counties, the averaged model sufficiently reflect the general trends of the recorded data:



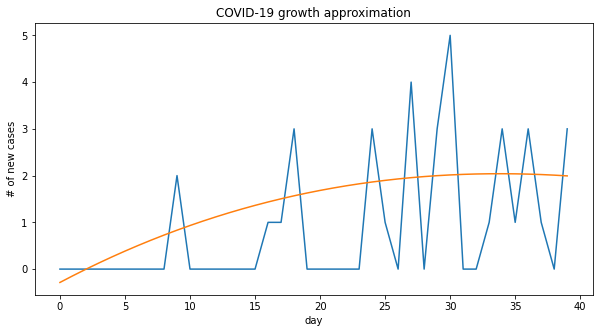
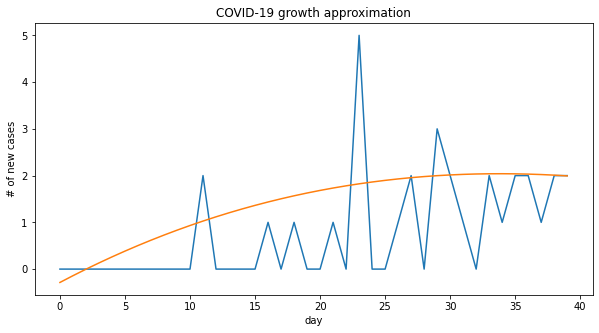
While NY data is exemplary to illustrate the used methodology, the hypothesis can be proved or disproved using the data for all counties in the US. Since the number of counties significantly increases, 5 clusters were used. The pool of model was constructed, and counties were clustered based on the came parameters:



With these clusters, cluster-based average models look as follows:



Clearly, COVID-19 spread patterns differ significantly for different clusters. Therefore, socio-economic standing of a county is an important factor in assessing the COVID-19 spread. While it is proved that clusters have different COVID-19 patterns, the question is whether these clusters are sufficient to determine the pattern. Cluster-based average look as follows compared to the excluded counties’ recorded spread:



As it can be seen, cluster-based average sufficiently reflects the county’s spread pattern.

Based on the reported results, it can be concluded that the counties within each cluster demonstrate drastically different patterns of COVID-19 spread. Socio-economic standing plays a major role in modelling and predicting COVID-19 spread. Splitting counties in clusters based on economic standing may be useful in a regression model that will predict the future spread in areas that COVID-19 has not yet fully reached, based on economically similar counties in highly affected regions.

1. Mobility Changes in Response to COVID-19, Warren and Skillman, March 2020 <https://www.descarteslabs.com/wp-content/uploads/2020/03/mobility-v097.pdf> [↑](#footnote-ref-1)
2. The analysis is contained in Part 1 of the Mobility\_Clustering\_and\_Infection\_Rate\_Reduction.ipynb Jupyter notebook [↑](#footnote-ref-2)
3. ‘See Which States and Cities Have Told Residents to Stay at Home’, New York Times, April 20, 2020, <https://www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html> [↑](#footnote-ref-3)
4. The analysis is contained in Part 2 of the Mobility\_Clustering\_and\_Infection\_Rate\_Reduction.ipynb Jupyter notebook [↑](#footnote-ref-4)
5. The analysis is contained in the Data-Analysis.ipynb Jupyter notebook [↑](#footnote-ref-5)
6. The analysis is contained in the spread\_functional\_approximation.ipynb Jupyter notebook [↑](#footnote-ref-6)