Investigation on Covid-19 spread in US states: Relationship between spread of the disease and Google mobility data & classification of patterns of spread between states

Course Project:

623 – Big Data

Authors:

Javad Honarvar

Hanif Kolahdoozan

University of Waterloo

Department of Engineering

Management Science

Spring 2020

Abstract

In response to the coronavirus disease 2019 (COVID-19) pandemic, almost all national governments have applied lockdown restrictions to reduce the infection rate. There is a huge body of literature on different aspects of Covid-19, ranging from estimations of spread and predictions to analysis of whether there is any relationship between this virus and any other variable. One of the most important analyses is finding out the relationship between mobility restrictions and spread.

In this project, we studied the Google mobility data for the past 7 months and its relationship and effect on the evolvement of Covid-19 disease in all states of the US. We chose the US for several reasons. First, because the US has been one of the hardest-hit countries in the world so far and had at least two major peaks in their new case numbers. Moreover, Google mobility data are more detailed and accurate in the US than in any other place in the world. Covid-19 disease statistics have been gathered from several other sources like John Hopkins University Covid-19 page and New York times Github repository which are a reliable source of statistics of Coronavirus.

Several statistical analyses had been made on data with different variations, time lags, and in different states, and finally a significant relationship between mobility data (which is a clear reflector of lockdown policies) and disease numbers observed through a regression model. Finally, Covid-19 cases in 13 states were explained by google mobility report and the corresponding R^2 was more than 0.65. After that, a comprehensive study on k-means clustering has done which made interesting intuitions.

Keywords: Google mobility, k-means clustering, Regression, Covid-19

Table of Contents

Contents

Literature Review

Introduction

After 8 months of the official declaration of the Covid-19 pandemic in December 2019, every country in the world tried to manage this catastrophe in their way, so the result of the performance of each country in controlling the disease is pretty different. Surprisingly some most developed countries in the world with sophisticated health infrastructures like the US and UK have the worst results in the fight with the Corona virus. In the absence of any effective cure and vaccination, such difference in countries numbers reveals that social behaviors and managerial decisions have a deep effect on the control of pandemic. Lockdown and opening rules are not simple decisions. For national or local government one day of lockdown means disappearing millions or billions of dollars of tax incomes and so many social consequences like bankrupt companies and lost jobs. On the other hand, the surge of the disease means the incalculable cost of health care and deaths, more harmful lockdown decisions, and frustrated society.

In such circumstances, tough managerial decisions like lockdown and opening decisions need solid facts that support great consequences, tough because of the complicated nature of bringing society behavior in numbers and prepare fact out of them is so complicated task.

Fortunately, with the help of GPS systems and sharing data with providers like Google Maps, social commuting data are more available than anytime before. Google company has recently released the mobility data of most of the countries which are valuable sources of data for study the societies behavior over time.

General Statement

As of March 23, 2020, 85% of new COVID-19 cases are reported in the United States and Europe. Countries that were initially heavily impacted by this pandemic (e.g. China and South Korea) have been successful at limiting the number of newly transmitted cases through massive testing as well as strict mobility and travel restrictions (lai et al, 2020; Chinazzi et al, 2020). Italy, which experienced the earliest and the most large scale outbreak of COVID-19 in Europe, enacted similar restrictions on citizens' mobility on March 8 and, as of March 21, 2020, began to show a decline in the reported number of new infections (Burn-Murdoch et al, 2020). Through the Italian government's nationwide lockdown, a 50% reduction in mobility within and between provinces was measured using large scale anonymized location data (Bajardi et al, 2020). This reduction in mobility and the quarantine measures imposed on the epicenters of the epidemic is expected to considerably change the trajectory of the pandemic in Italy (Cereda et al, 2020). South Korea's announced strategy of large-scale testing, uni_ed public messaging encouraging social distancing, and the use of face masks, along with contact tracing and early isolation of infectious individuals appears to have achieved a top-down mobility restriction enforced in Italy and China. While most

other European countries have now enacted population mobility restrictions similar to those in Italy, the delayed timing of these policies has left several European countries (in particular, Spain, France, and the United Kingdom) vulnerable to the rapidly advancing pandemic. Now, the pandemic has been more serious in the U.S. which requires a better understanding, inspection, and focus.
Before going further, specifically on U.S. Covid-19 curve and data, we need to take a look at google community mobility reports and have an insight over this helpful report.

What is Google Mobility Report?
According to the official google website in response to what google mobility report is, the following explanation has been written:

> "As global communities respond to COVID-19, we've heard from public health officials that the same type of aggregated, anonymized insights we use in products such as Google Maps could be helpful as they make critical decisions to combat COVID-19. These Community Mobility Reports aim to provide insights into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential."

How COVID-19 Community Mobility Data has been constructed?

COVID-19 Community Mobility Reports provide insights into changes in mobility patterns. These reports use anonymized, aggregated data to chart movement trends over time by geography, as well as by place categories, showing trends over several weeks. This works in a similar way to existing Google products and features. For example, Google Maps uses aggregated, anonymized data to show how busy certain types of places are, including when a local business tends to be the most crowded. Public health officials have suggested this same type of aggregated, anonymized data could also be helpful as they make critical decisions to combat COVID-19. The COVID-19 Community Mobility Reports provide insights into what has changed in response to work-from-home, stay-at-home, and other recommended policies aimed at flattening the curve of the COVID-19 pandemic. They analyze trends in visits made to high-level categories of places, including workplaces, retail and recreational venues, groceries and pharmacies, parks, transit centers, and places of residence. Each version of the report will show trends over several weeks, with the most recent data representing 48 hours prior.

Related Works

Although from the very beginning of spreading the COVID-19, scholars and scientists have tried to reach a model best describes the behavior of this virus, not much accurate answers, results, or modeling have been made and it is yet to be done. In the following section, we will discuss related works with a concentration on two major approaches: Forecasting Modeling and Google Mobility Report Implications.

Forecasting Modellings

a group of Chinese scholars developed a modified stacked auto-encoder for modeling the transmission dynamics of the epidemics. They applied this model to real-time forecasting the confirmed cases of Covid-19 across China. The data were collected from January 11 to February 27, 2020, by WHO. Furthermore, they employed the latent variables in the auto-encoder and clustering algorithms to group the provinces/cities for investigating the transmission structure (Zixin Hu et al, 2020).

In their publication, using the multiple-step forecasting, the estimated average errors of 6-step, 7-step, 8-step, 9-step and 10-step forecasting were 1.64%, 2.27%, 2.14%, 2.08%, 0.73%, respectively. They also predicted that the time points of the provinces/cities entering the plateau of the forecasted transmission dynamic curves varied, ranging from Jan 21 to April 19, 2020. The 34 provinces/cities were grouped into 9 clusters (Zixin Hu et al, 2020).

Another approach that has been done in Argentina, were using models from an exponential smoothing family with multiplicative error and multiplicative trend components turned into a forecast of novel Coronavirus. In this research, the author has provided five different forecasting periods and suggests that while in China Mainland the coronavirus will decrease, outside of China, the number of cases will soon double (Petropoulos and Makridakis, 2020)!

Mobility Report Implications

In the very first months of happening this outbreak, many tried to focus on mobility restriction effects on stopping the spread. One of the most famous researches belongs to the Laboratory for the Modeling of Biological and Sociotechnical Systems at Northeastern University. This study states that by 23 January 2020, the epidemic had already spread to other cities within mainland China and travel quarantine around Wuhan was not successful and has only modestly delayed the spread of disease to other areas of mainland China. Furthermore, the modeling shows that outside of China, additional travel limitations (up to 90% of traffic) have only a modest effect unless paired with public health interventions and behavioral changes that can facilitate a considerable decline in disease transmissibility (Chinazzi et al, 2020).

Italian scholars started working on different aspects of Covid-19 spread. In one of these works, an exploration of how variations in mobility relate to some basic and essential economic variables has been given. According to their approach, they showed that a decline in connectivity is stronger in municipalities with low average individual income and high-income inequality. Simultaneously, they demonstrated that mobility restrictions had a higher impact on municipalities with higher fiscal capacity. Finally, they concluded with three main points: 1. the lockdown seems to unequally affect the poorer part of the population. 2. Reduction in mobility

and connectivity induced by the lockdown is more pronounced for municipalities with stronger fiscal capacity. And Finally, 3. The distribution of income plays an indispensable role; Municipalities where inequality is higher experience more pronounced mobility contractions (Bonaccorsi et al, 2020).
Working from home is another issue, investigated by some researchers. In one of them, a survey distributed to over 5,000 working-age adults in the U.S., resulted that from 8.2% of the workforce worked totally from home in February 2020, 35.2% practiced working from home in May 2020. Another important outcome is that highly educated, high-income, and white individuals were much more likely to shift to remote work and to maintain employment following the virus outbreak. They also claimed that with available estimates of the potential number of home-based workers, a large majority (almost 71%) of US workers could work from home (Bick et al, 2020).

Exploratory Data Analysis

In this part of the work, an overview of the following items is provided:

- Data Structure
- Variables & Features
- A brief explanation of Data Cleaning
- Charts, Graphs, and Figures of Data

The reason to do this in a separate part is to have a deep concentration on what data is and how we can work on it. Additionally, we need to clarify some points about the methods which have been employed for analysis.

Data Structure

The data has consisted of 10 columns for 54 states

Variable	Definition	min	mean	Max	sd
lag case	Daily cases of COVID-19 in U.S. states	0	180	15300	1246.96
Uninsured population	The percentage of uninsured people	2.8	8.194	17.7	3.043
Total Tests	Total tests take in states	2	434111.23	6414321	737891.7
date	Date	2020-02-15	2020-05-17	2020-07-20	-
retail	Mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters	-77	-21.35	33	18.381

groceries	Mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies	-62	-1.59	51	13.16
parks	Mobility trends for places like local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens	-77	60.62	502	82.53
transit	Mobility trends for places like public transport hubs such as subway, bus, and train stations	-81	-25.33	62	23.40
work	Mobility trends for places of work	-78	-32.8184	10	14.93239
residential	Mobility trends for places of residence	-5	11.24238391	33	6.682564482

Variables & Features

As Google company declared in the mobility data page, most governments use similar aggregated data to Google map data, to develop an estimate of society mobility and make critical decisions about the Covid-19 issues. Google mobility data is novel source of data which recently become available by the Google and contains average mobility data of most of the regions of the world, in six categories with daily updates. Until the date of writing this report (31 July 2020) the mobility data until 29th July is available which is a csv file, with 1365701 rows and 14 columns wich are:

```
['country_region_code', 'country_region', 'sub_region_1', 'sub_region_2',
       'metro_area', 'iso_3166_2_code', 'census_fips_code', 'date',
       'retail_and_recreation_percent_change_from_baseline',
       'grocery_and_pharmacy_percent_change_from_baseline',
       'parks_percent_change_from_baseline',
       'transit_stations_percent_change_from_baseline',
       'workplaces_percent_change_from_baseline',
       'residential_percent_change_from_baseline']
```

First 7 columns (‘country_region_code’ to ‘census_fip_code’) are related to geograohical point of data, date shows data of data and last six columns are mobility columns as it could be infered from their name which related to difference in percentage change of society mobility in those kind of places. For example, for a sample record like this:

```
mobility_raw.loc[1000]
Out[173]:
country_region_code                                         AE
country_region                                 United Arab Emirates
sub_region_1                                            Sharjah
```

```
sub_region_2                                                    NaN
metro_area                                                      NaN
iso_3166_2_code                                              AE-SH
census_fips_code                                                NaN
date                                                     2020-03-02
retail_and_recreation_percent_change_from_baseline                3
grocery_and_pharmacy_percent_change_from_baseline                 4
parks_percent_change_from_baseline                                5
transit_stations_percent_change_from_baseline                    -3
workplaces_percent_change_from_baseline                           5
residential_percent_change_from_baseline                          1
Name: 1000, dtype: object
```

The record above means that in the day March 2nd of 2020 in the Sharjah region in UAE, in comparison to a baseline, there were 3 percent more in retail and recreation places, 4 percent more in grocery and pharmacy, 5 percent more in parks, 3 percent less in transit stations, 5 percent more in the workplace in one percent more mobility in residential places mobility. These data are out of processed, anonymized and accumulated data users share with Google Maps. The baseline day is the median value from the 5 weeks Jan 3 – Feb 6, 2020.

Because of date columns, the nature of the dataset is time series and should be considered in further analysis. Also, according to instructions provided by Google about this dataset, it is highly recommended to not to compare and analyze data of different countries with each other since there are significant differences between aria divisions in different regions and how people use and share location history with Google applications.

Each high-level category in this data set consists of many other sub-categories that are aggregated with the distancing logic and rationality is behind the categorization of places for Covid-19. For example, grocery and pharmacy are coming together since these two are among the essential places to go. Also, two categories of parks and transition are aggregated from these subcategories:

Parks	Transit stations
Public garden	Subway station
Castle	Sea port
National forest	Taxi stand
Camp ground	Highway rest stop
Observation deck	Car rental agency

Data Cleaning

An important note in preparing the dataset is to figure out how many lag days do people face for incubation. According to the incubation time of the disease is between 2 and 12 days (95% interval; see Lauer et al. (2020)). Given this, it is to be expected that any changes in mobility will

have a lagged effect on the discovery of new cases. For this reason, lagged moving averages of the mobility indicators are calculated. Furthermore, it is possible that mobility and reports of new cases of COVID-19 are endogenous, if the public adjust their mobility according to reports of the incidence. Therefore, in addition to being consistent with an incubation period, the use of lagged indicators also helps to break this potential indigeneity.

For this purpose, we also tried to reach the best time lag that can explain better the COVID-19 incidence. Therefore, we employed 3 different lagging days, including **9 days**, **15 days,** and once with **20 days**. Finally, we chose 15 days lag, because this amount demonstrated better performance on our regression model.

Regarding data wrangling, in the first step the null data queried which shows significant amount of null data in sum columns:

```
mobility_raw.isna().sum()
```

```
country_region_code                                   1273
country_region                                           0
sub_region_1                                         31218
sub_region_2                                        323288
metro_area                                         1356488
iso_3166_2_code                                    1043367
census_fips_code                                    938526
date                                                     0
retail_and_recreation_percent_change_from_baseline  468243
grocery_and_pharmacy_percent_change_from_baseline   496436
parks_percent_change_from_baseline                  709558
transit_stations_percent_change_from_baseline       688560
workplaces_percent_change_from_baseline              60189
residential_percent_change_from_baseline            670628
```

with majority null cells in columns metro_area, iso_3166_2_code, and census_fips_code it is obvious that these columns do not have reliable data and should be filtered. Also, there is a significant amount of null data in mobility columns which according to Google instruction those nulls are because of a lack of sufficient amount of data and avoiding violating anonymity in certain dates and areas.

By analyzing data and implementing several filters, finally we reached a part of the dataset which is related to states of the US. By applying this filter and change columns names for simplicity, a data set by the shape of 9 columns and 8364 rows which is related to the mobility records of states in the United States for the last 7 months (beginning of pandemic till the time of this report).

In this subset of data, the number of null data is significantly reduced and only parks data have 52 null data which:

```
mobility.isna().sum()
```

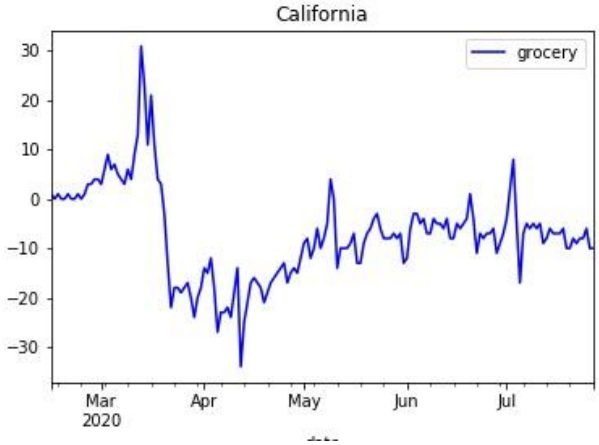
```
country_region_code      0
state                    0
date                     0
retail                   0
grocery                  0
parks                   52
transit                  0
workplace                0
residential              0
```

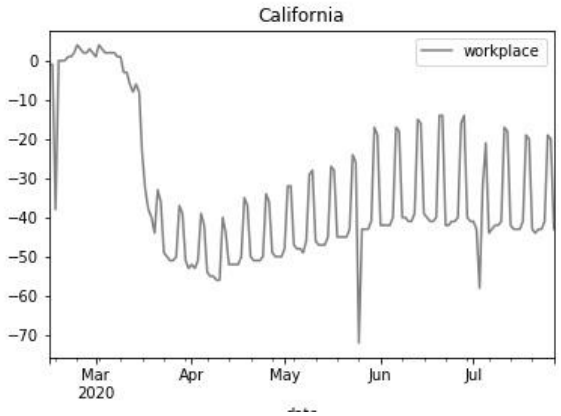
mobility.shape

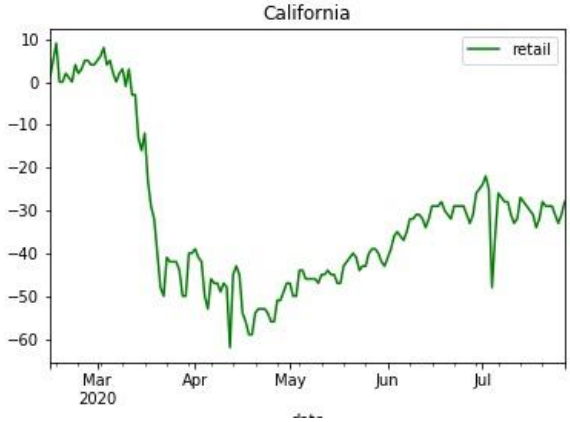
```
(8364, 9)
```

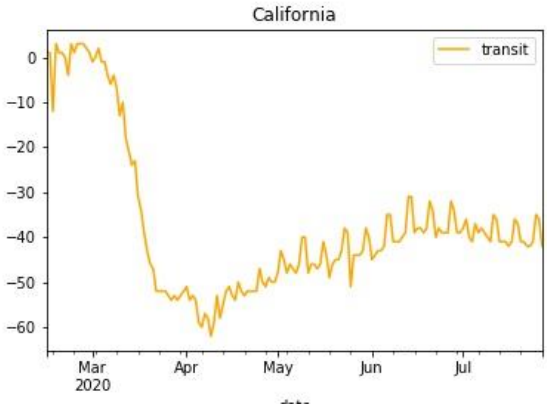
Charts, Graphs, and Figures

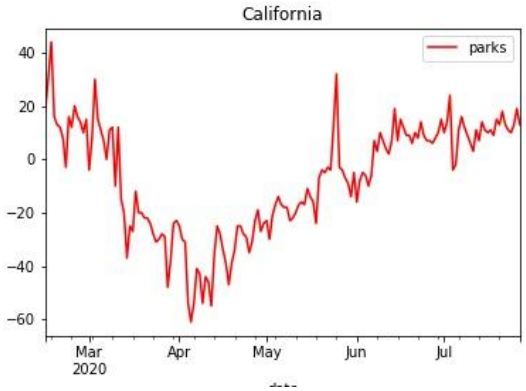
the following graphs illustrate mobility change against a baseline for a sample of the states, California. It represents that in different locations like parks, residential, grocery, workplace, retail, and transit.

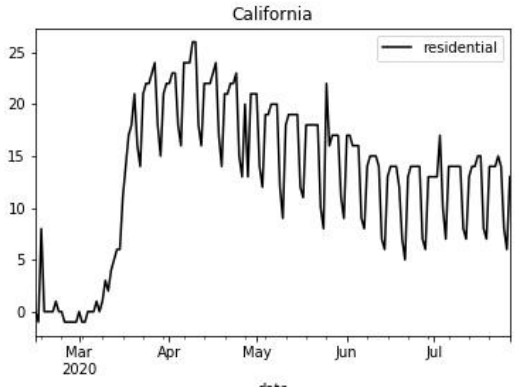












Supervised Learning Model

We employed a regression model as a supervised learning model to investigate how much of daily Covid-19 cases are explained by the mobility report features.

The output of the regression model for states those which have more than 0.65 Adjusted R squared are as follows:

$$Cases = \beta_0 + \beta_1(retail) + \beta_2(lagcases) + \beta_3(parks) + \beta_4(grocery) + \beta_5(transit) + \beta_6(workplace) + \beta_7(residnetial)$$

state	Required	retail	laccases	parks	grocery	transit	workplace	residential
North Carolina	0.874	9.34	1.11	-0.87	2.05	-5.25	-0.31	9.34
South Carolina	0.866	2.72	1.19	-2.09	-5.31	12.56	-11.12	-15.09
Texas	0.866	58.60	1.58	-10.70	-18.79	-31.73	-26.94	-95.05
California	0.859	35.48	1.39	-11.93	1.93	-50.66	-22.69	-138.44
Utah	0.773	1.98	1.03	-0.12	2.92	-0.99	-0.83	1.14
Nevada	0.752	12.41	1.17	-1.31	-1.80	-3.01	-8.27	-15.79
Alabama	0.737	-2.77	0.85	-3.03	-8.06	24.72	-16.43	-23.88
Florida	0.734	132.07	1.53	-72.28	41.91	-75.49	-4.59	-144.61
Arizona	0.703	84.57	0.84	-5.00	-12.47	-73.90	-13.51	-44.19
Arkansas	0.703	-0.73	0.02	0.99	-9.15	18.30	-12.53	-25.75
Pennsylvania	0.664	-23.84	-0.08	-1.23	2.16	26.53	-11.27	26.52
Oklahoma	0.662	-1.80	0.90	-1.69	-6.44	20.89	-11.59	-8.47
Rhode Island	0.657	-2.97	-0.04	-0.04	-1.60	0.62	0.75	7.42

Unsupervised Learning Model:

In the implementation of unsupervised learning models, according to literature and time-series feature we employed k-means clustering, which is an appropriate way to cluster different states in the case of Covid-19. It is important to note that we used

In the implementation of unsupervised learning models, according to literature and time-series feature we employed k-means clustering, which is an appropriate way to cluster different states in the case of Covid-19. It is important to note that we used two clustering approaches which are described as follows:

1- Clustering states based on the cases and deaths per population in the specific date which is a normal 2 dimension clustering task
2- Clustering states based on the daily case curves which are a time series clustering task.

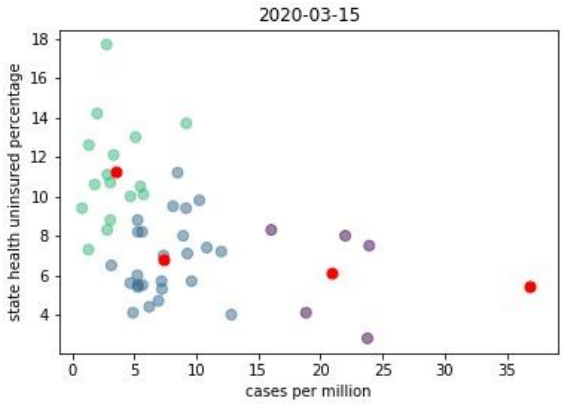
K-means Clustering based on Test cases per million and Uninsured percentage of the population in each state

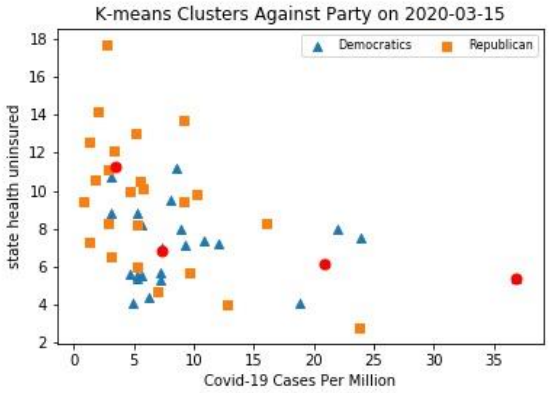
To get a deep understanding of how Covid-19 disease growth in different societies with different levels of health services we looked for the statistics of health insurances of each state in the US which one of the main indicators were the percentage of the uninsured population, which shows the vulnerable percentage of the population with less access to health services. Then we run K-means clustering on the data to classify states based on the number of cases per million and uninsured percentage of the population to see if the states with less health service coverage (more uninsured percentage) would become along those who will eventually with more cases per million or not.

Since different states have different populations, for comparison between their Corona-virus indicators should be considered their population, so by dividing indicators (new cases and deaths) by population and multiply by million (to have more touchable numbers), new data columns developed. By feeding new columns to the K-means clustering function of the Scikit library in Python 3, for each date, there would be a different cluster.

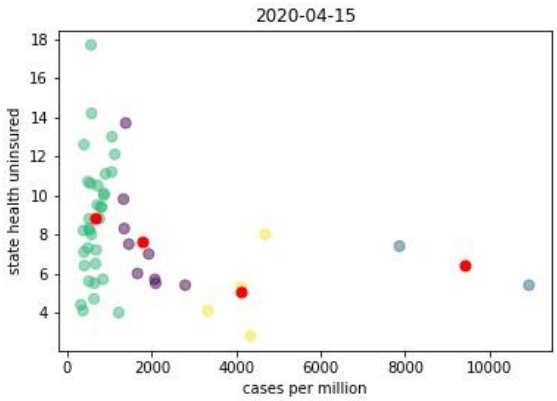
For this purpose, 4 clusters considered and code runs for 5 different dates in monthly intervals. The results are shown in the tables below:

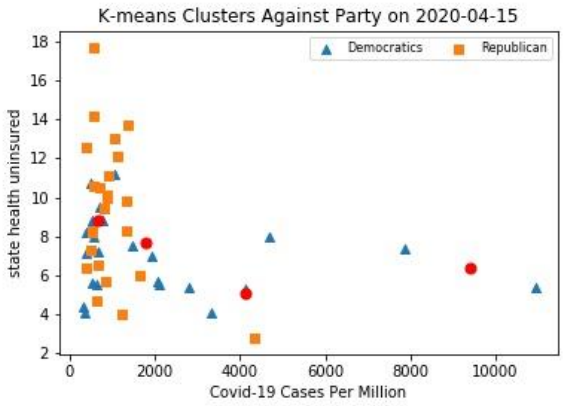
Another important point of our study is its political perspective. According to arguments between President Donald Trump and governor Andrew Cuomo at the beginning of Covid-19 spread in New York and the difference between their approaches in the management of the situation, the idea of analysis cases in terms of which party policies have been governed on corresponding state developed and following figures are represented.



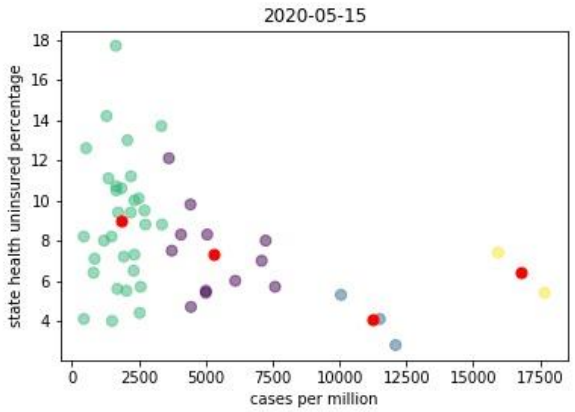


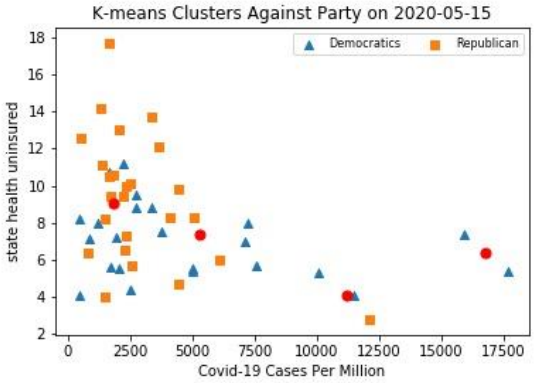
group number	number of states	Democratic	Republican	average Covid-19 cases per miilion	Average uninsured percentage of population
1	5	60%	40%	11706.09	9.6
2	25	68%	32%	18275.69	6.4
3	16	13%	88%	2804.97	6.9
4	1	100%	0%	7406.30	8.2



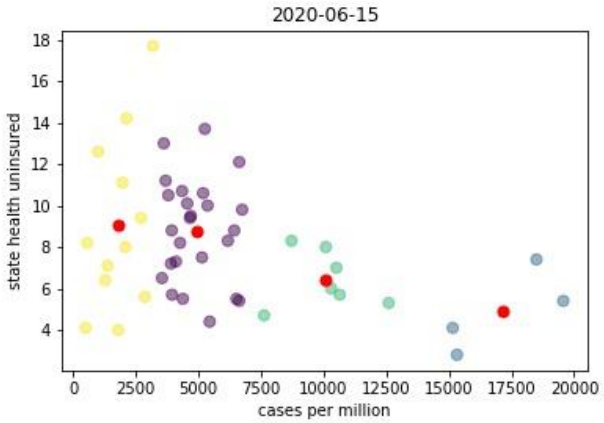


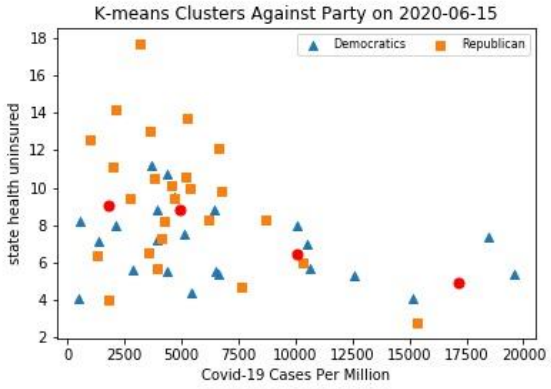
group number	number of states	Democratic	Republican	average Covid-19 cases per million	Average uninsured percentage of the population
1	9	56%	44%	20.92	6.1
2	2	100%	0%	7.41	6.8
3	33	39%	61%	3.48	11.3
4	4	75%	25%	36.85	5.4



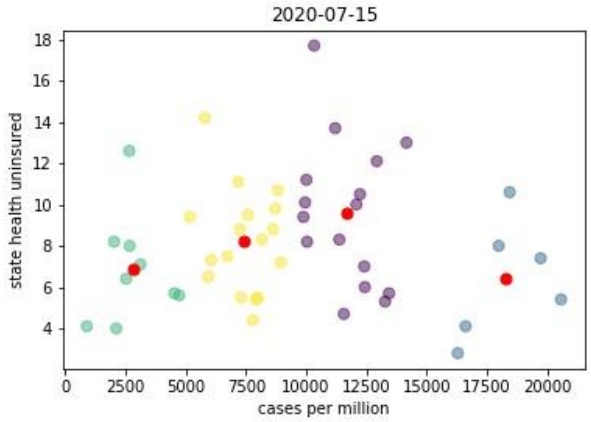


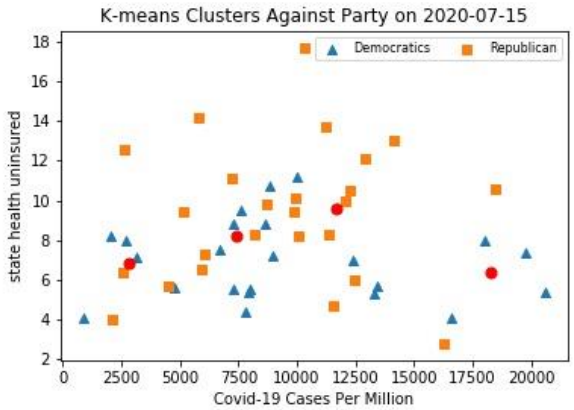
group number	number of states	Democratic	Republican	average Covid-19 cases per million	Average uninsured percentage of the population
1	12	50%	50%	5281.94	7.4
2	3	67%	33%	11220.38	4.1
3	31	42%	58%	1846.10	9.0
4	2	100%	0%	16800.14	6.4





group number	number of states	Democratic	Republican	average Covid-19 cases per million	Average uninsured percentage of the population
1	25	44%	56%	4927.96	8.8
2	4	75%	25%	17138.49	4.9
3	7	57%	43%	10074.56	6.4
4	12	42%	58%	1797.74	9.0





group number	number of states	Democratic	Republican	average Covid-19 cases per million	Average uninsured percentage of the population
1	16	25%	75%	11706.09	9.6
2	6	67%	33%	18275.69	6.4
3	9	56%	44%	2804.97	6.9
4	17	59%	41%	7406.30	8.2

Red dots: centers of each cluster.
Colors: indicate each cluster members

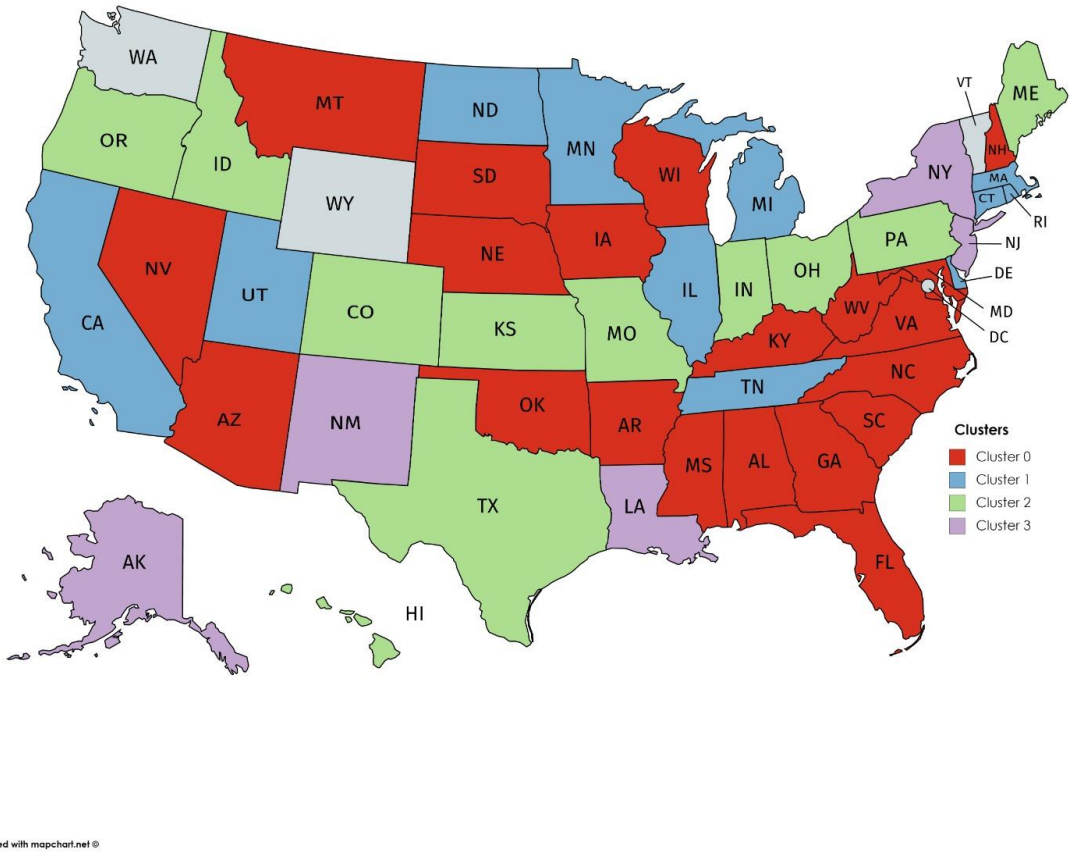
Highlights:

- As of the beginning of the outbreak to May 15, more democratic states have more cases even with low uninsured percentages (8 democratic states against 2 republican states.)
- As of July 15, both parties seem to have the same amount of states in right-hand clusters, meaning that both of them in the second wave of Covid-19 have a high rate of cases regardless of their corresponding uninsured percentage

According to mentioned highlights and as we can see from March 15th to June 15th, however, the states with a more uninsured fraction of population are among those which have republican governors and at the same time suffer less from Covid-19, in the last period, as of July 15th, they are no longer in left-hand side clusters, and we can find them all spread out evenly in large amounts of Covid-19 cases.

From March 15th to June 15th we can see the same pattern in which states with the less uninsured percentage of the population have higher cases but in the second wave of Covid-19 as of July 15th left-hand clusters are more skewed to the right which fails the initial assumption. Besides, the first and fourth clusters' centroids on July 15th have the same percentage of the uninsured population, therefore checking clustered scatter plots visually it looks to have enough evidence

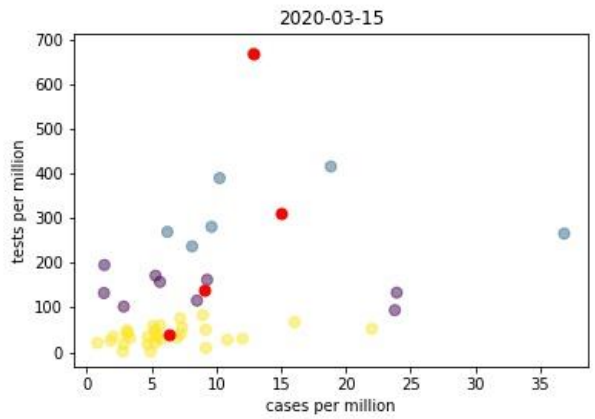
to claim that there is no particular pattern between the uninsured percentage of the population and people's social distancing behavior.

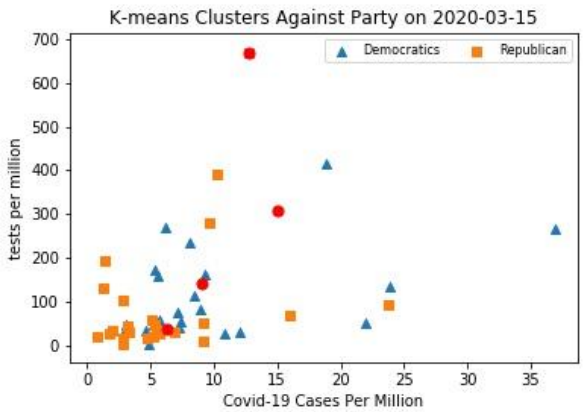


Before starting the other K-means clustering, according to the above figure which has been drawn based on July 15th data, you can easily find different states based on corresponding clusters. The first cluster has 19 states, second cluster 10 states, third one 10 states, and the last one includes only 5 states.

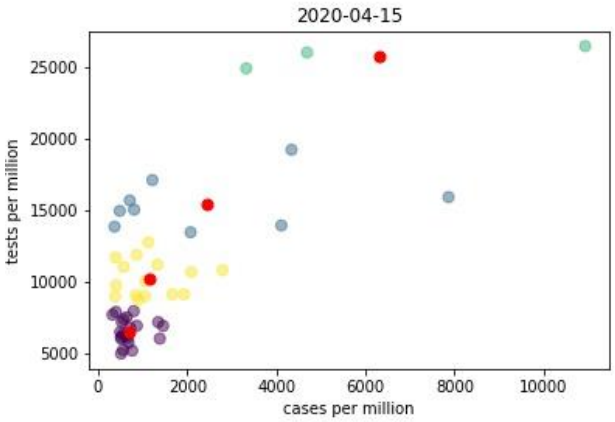
K-means Clustering based on Test cases per million and Cases per million

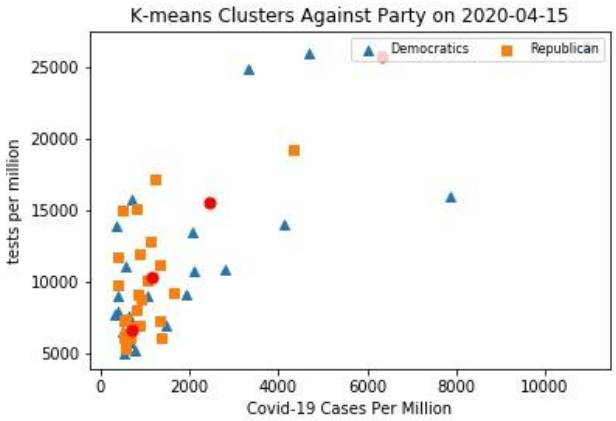
Mass testing considered one of the effective actions governments can take to control the disease. To get a deeper insight into how different levels of testing can classify states based on their testing numbers and Covid-19 spread, test per million numbers versus Covid-19 case per million data gathered and fed into K-means clustering algorithm. Plots of 5 dates are provided below which provides better insight into the evolution of the disease and the number of tests per million for each state.



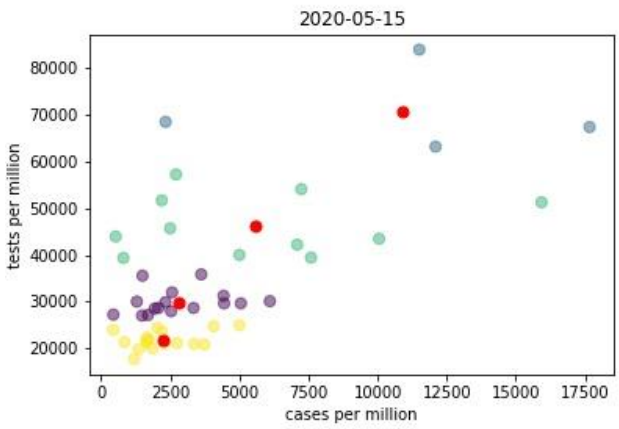


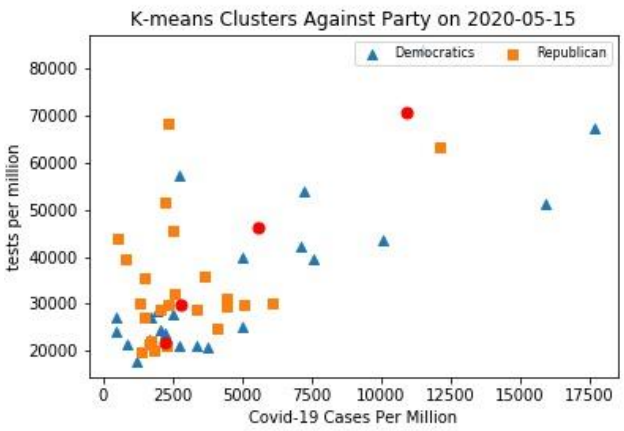
group number	number of states	Democratic	Republican	average Covid-19 cases per million	Average tests per million
1	9	56%	44%	9.1	140.3
2	6	67%	33%	15.0	309.5
3	1	0%	100%	12.8	668.3
4	31	45%	55%	6.4	38.0



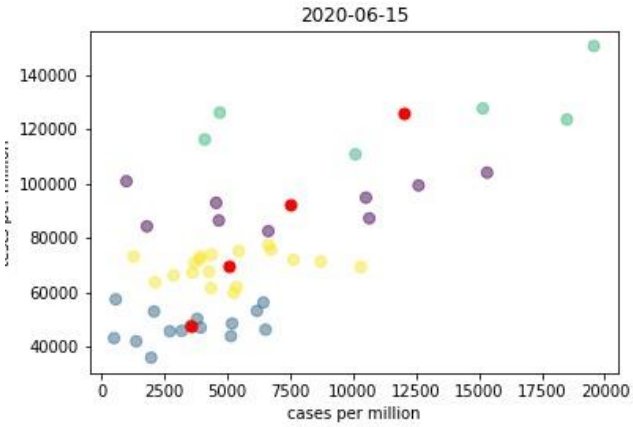


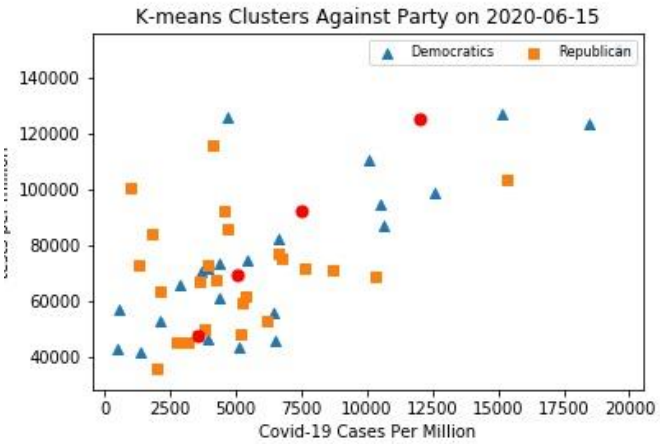
group number	number of states	Democratic	Republican	average Covid-19 cases per million	Average tests per million
1	21	43%	57%	2801.21	29915.82
2	9	56%	44%	10901.66	70742.30
3	3	100%	0%	5603.26	46195.89
4	15	40%	60%	2224.21	21832.46



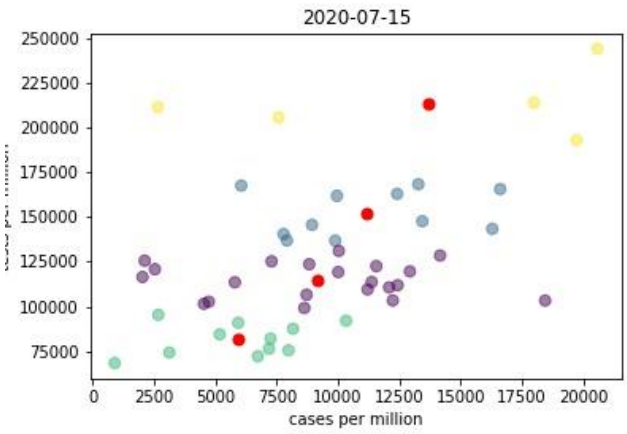


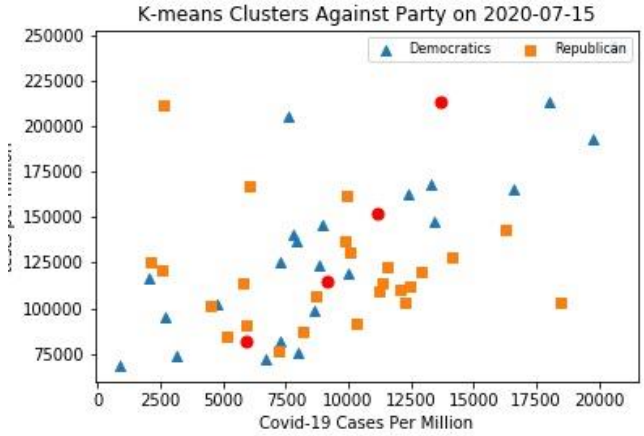
group number	number of states	Democratic	Republican	average Covid-19 cases per miilion	Average tests per million
1	16	25%	75%	2801.2	29915.8
2	4	50%	50%	10901.7	70742.3
3	11	64%	36%	5603.3	46195.9
4	17	59%	41%	2224.2	21832.5





group number	number of states	Democratic	Republican	average Covid-19 cases per million	Average tests per million
1	9	44%	56%	7530.3	92461.1
2	15	53%	47%	3550.9	47685.5
3	6	83%	17%	12024.1	125813.0
4	18	33%	67%	5037.0	69520.6





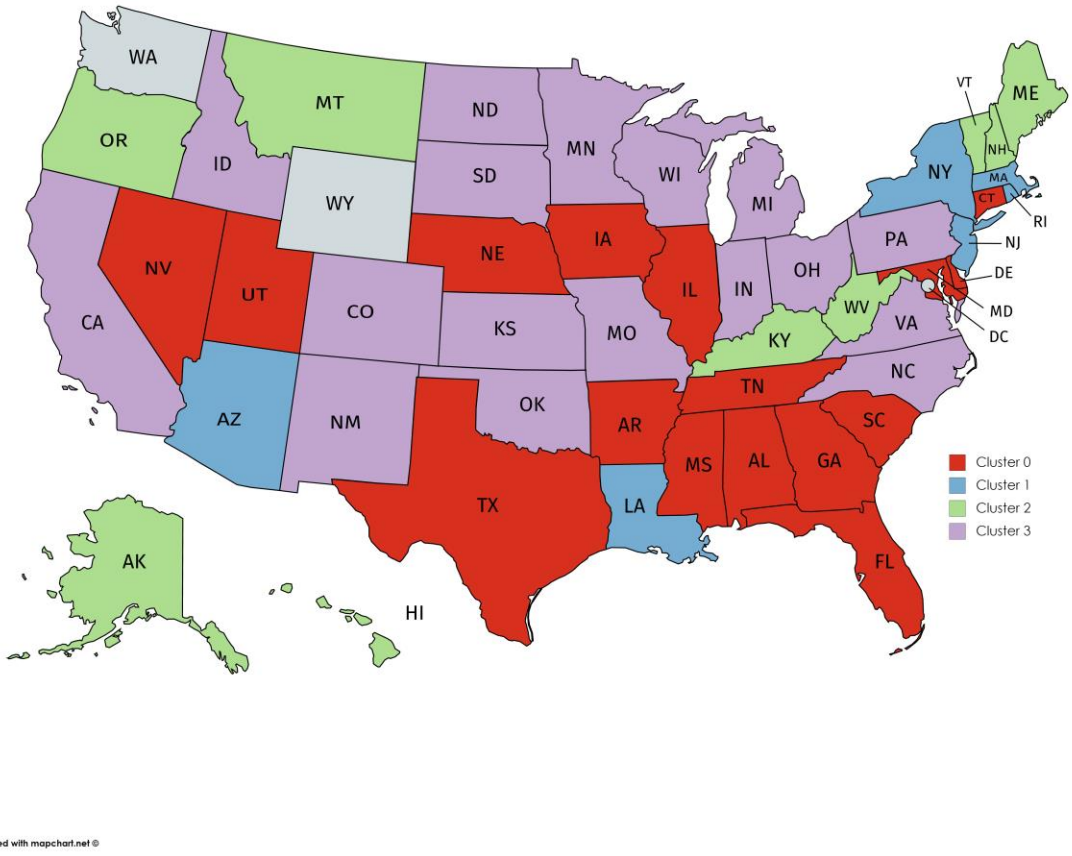
group number	number of states	Democratic	Republican	average Covid-19 cases per miilion	Average tests per million
1	9	44%	56%	9124.6	114695.1
2	15	53%	47%	11140.6	152469.1
3	6	83%	17%	5946.2	81719.2
4	18	33%	67%	13706.2	213594.9

Highlights:

- At the very beginning of the spread, most states, regardless of the governors' party, are in lower clusters.
- In last month, data is more spread out and you can find clusters are more based on cases.

According to above and concerning the fact the more tests, the better distinction between high case states and low case states (data is more spread out in upper clusters rather lower ones), we may be able to see that more democratic parties could provide tests in helped the distinction for policymakers to have better insight toward a level of emergency in each state.

From scatters above it could be seen that at the first months of spreading of the disease each cluster are tightly around the center point (red dot) and a great amount of correlation between test numbers and case number could be seen which means states with more tests have more cases which imply more testing doesn't have an impact on the control of the disease, but as the time passes in the 15th July it could be seen that there is a great amount of spread of data on X-axis especially on the latest cluster with highest tests per millions which shows there are states with a high percentage of testing that have fewer cases per million than others and could control the disease better than other states.



Before starting a new section, in the above figure, which has been drawn based on July 15th data, you can easily find different states based on corresponding clusters. The last cluster has 17 states, 16 states first cluster, 5 states second cluster, and 9 states in the third cluster.

2- Clustering of new daily cases curves.

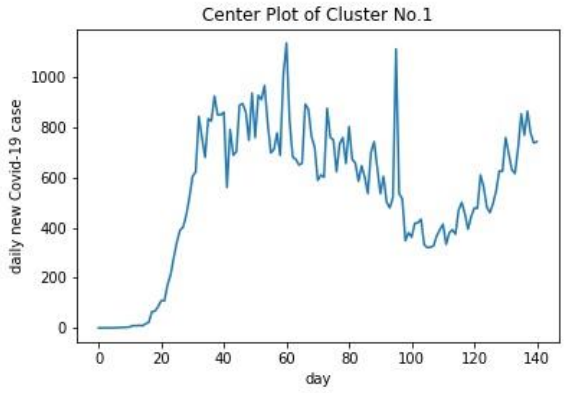
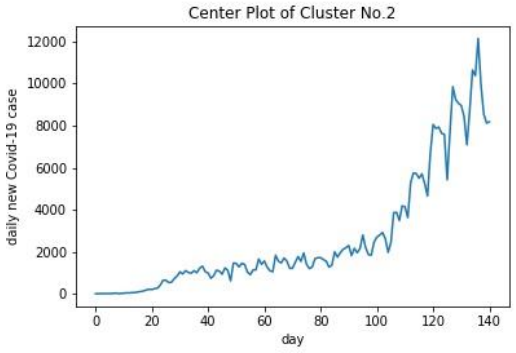
Time-series clustering using the k-means algorithm is an interesting concept. As mentioned in the related work section, there are significant applications like anomaly detection and signal processing when clustering technics applied on time series data. In signal detestation for instance, by moving a window function on the date which takes fixed lengths of data and then feed theses chunks of data to clustering algorithms, since clustering algorithms consider them as fixed vectors would eventually classify them to similar shapes that would be typical signals of data.

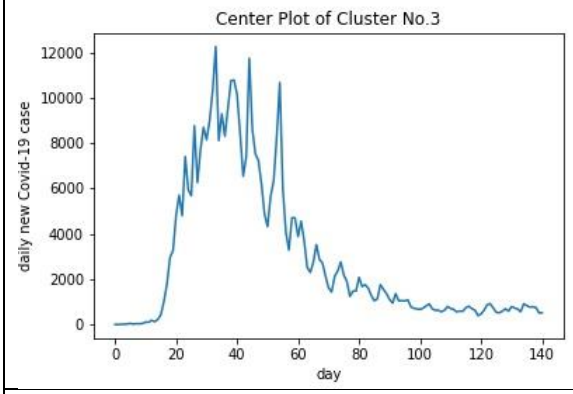
In this experiment with Covid-19 data, we considered the four months daily case data of each state as 1 × 140 vectors, and feed them to a K means clustering algorithm, since this algorithm considers these time-series data as vectors and would classify states with similar patterns of daily cases. these clustering tasks would give a clear insight through how we can study different states in their fight with the pandemic.

The plot of Center values of each cluster represents the trend of daily cases for that class. Several clustering runs with a different number of clusters run on the data and finally, we visually observed that 8 clusters would bring a distinct classification out of data. Plots of this classification and final table provided below:

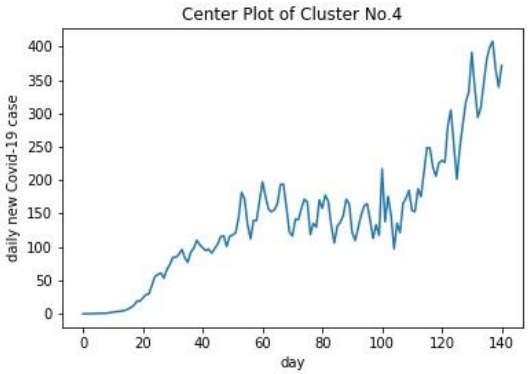
It should be mentioned that for visibility reasons, the scale of y-axes in plots are different which should be considered in comparing clusters together. Some clusters have similar patterns like clusters 2 & 4, but their y-axes scale is far different, so states that belong to each of these clusters have a completely different number of daily cases.

The X-axis in plots below is day number from March 1st to July 20th and below each center plot of the cluster is the name of members classified in that cluster. This analysis can help policymakers to classify states for applying new policies and doing analysis.

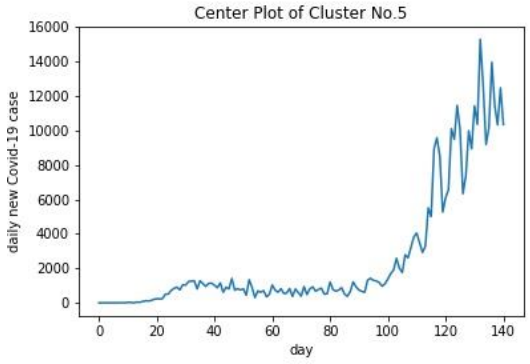
<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>Connecticut Ohio Maryland
Indiana Pennsylvania
Virginia Michigan</td>
<td>California Texas</td>
</tr>
</table>



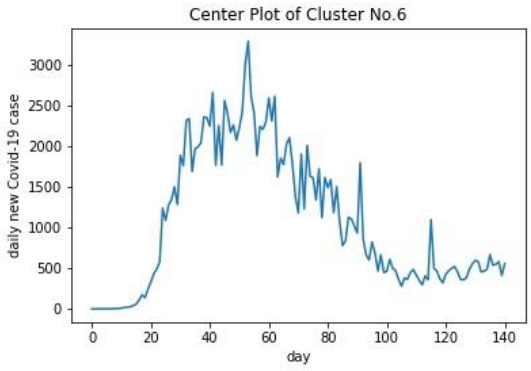
New York



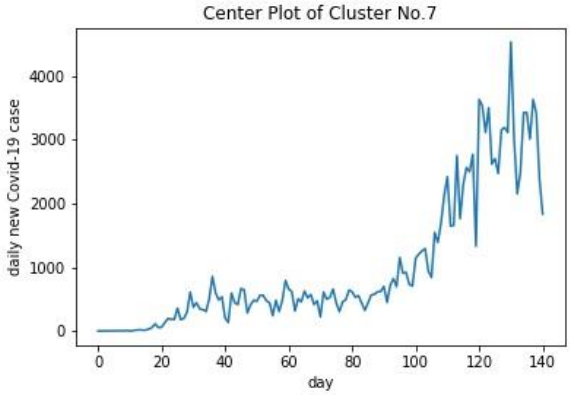
Alaska	Nebraska	Kansas
Arkansas	Nevada	Kentucky
Colorado	New Hampshire	Maine
Delaware	New Mexico	Minnesota
Hawaii	North Dakota	Mississippi
Idaho	Oklahoma	Missouri
Iowa	Oregon	Montana
Rhode Island	Vermont	Utah
South Dakota	West Virginia	Wisconsin



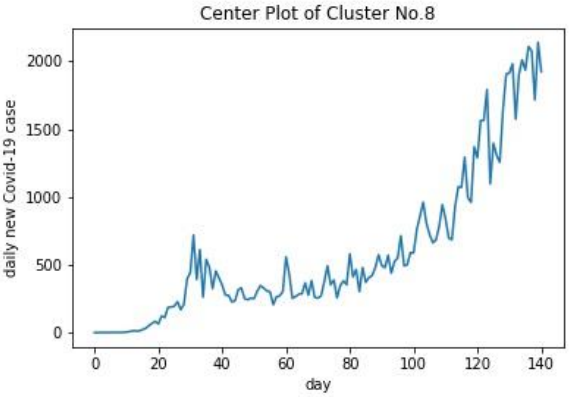
Florida



Illinois	Massachusetts	New Jersey



Arizona	Georgia



Alabama	Louisiana	North Carolina
South Carolina	Tennessee	

Conclusion

Although there is a huge demand to work on Covid-19 data across the world, unfortunately, there are barriers in this way that can negatively affect the quality of the studies. For instance, in our study, we barely could find an official set of data, containing many useful and meaningful features by the U.S. Government. If we had the chance to work on a more comprehensive dataset with many different variables, better models could have been built, tested, and offered. We believe that in ease of access to data, the world could stay more effective against this virus.

As a conclusion to draw, Covid-19 cases could have been explained by google community mobility report with meaningful percentages between 0.86 and 0.874 in North Carolina, South Carolina, Texas, and California. 9 other states have recorded R Squared above 0.65. Regarding this fact, there might be some intuition behind each variable of google mobility report. For instance, going to parks have increased during this period which makes sense that people tend to replace indoor party or gatherings with outdoor activities with social distancing practices. Or you can see that residential have a negative coefficient in most of the states, illustrating the effect of physical distancing.

In terms of unsupervised learning, employing clustering, we built three different k-means clustering models for the following datasets:

- Covid-19 cases against the percentage of uninsured population among states
- Covid-19 cases against total taken tests among states
- Political perspective in the context of clustering Covid-19 cases against uninsured people
- Political perspective in the context of clustering Covid-19 cases against total taken tests

We also provided another major investigation by clustering time series of each state Covid-19 cases concerning the shape of curves in 8 different clusters. This also can help decision-makers to have a deep understanding of circumstances in each state and the country as a whole.

We believe that this step can be adopted in future studies, and we are enthusiastically following your feedback to enhance future studies.

References

Hu, Z., Ge, Q., Jin, L., & Xiong, M. (2020). Artificial intelligence forecasting of covid-19 in china. *arXiv preprint arXiv:2002.07112*.

Lai, S., Ruktanonchai, N. W., Zhou, L., Prosper, O., Luo, W., Floyd, J. R., ... & Yu, H. (2020). Effect of non-pharmaceutical interventions for containing the COVID-19 outbreak in China. *medRxiv*.

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... & Viboud, C. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, *368*(6489), 395-400.

J. Burn-Murdoch, C. Tilford, K. Fray, S. Bernard,
The latest _gures as the pandemic spreads, https : / / www . ft . com / content /a26fbf7e-48f8-11ea-aeb3-955839e06441 (2020).

Pepe, E., Bajardi, P., Gauvin, L., Privitera, F., Lake, B., Cattuto, C., & Tizzoni, M. (2020). COVID-19 outbreak response: a first assessment of mobility changes in Italy following national lockdown. *medRxiv*.

Cereda, D., Tirani, M., Rovida, F., Demicheli, V., Ajelli, M., Poletti, P., ... & Magoni, M. (2020). The early phase of the COVID-19 outbreak in Lombardy, Italy. *arXiv preprint arXiv:2003.09320*.

Bonaccorsi, G., Pierri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F., ... & Pammolli, F. (2020). Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences*, *117*(27), 15530-15535.

Bick, A., Blandin, A., & Mertens, K. (2020). Work from home after the COVID-19 Outbreak.

Data Set Recourses:

- Daily new cases and daily deaths per state in U.S.
  https://github.com/nytimes/covid-19-data
- Daily tests taken per state.
  https://covidtracking.com/data/download
- Google Mobility Report Data
  https://www.google.com/covid19/mobility/