APPLIED DATA SCIENCE CAPSTONE – WEEK 4

ANALYZING THE OUTBREAK OF COVID-19 IN THE UNITED STATES AND USING MACHINE LEARNING ALGORITHMS TO PREDICT CONFIRMED CASES IN THE COMING DAYS

BACKGROUND

The coronavirus COVID-19 pandemic is the defining global health crisis of our time and the greatest challenge we have faced since the second World War. Since its emergence in Asia late last year, the virus has spread to every continent except Antarctica. COVID-19, short for "coronavirus disease 2019," is the official name given by the World Health Organization. As many as 213 countries and territories have registered COVID-19 cases, and the entire world is buzzing with uncertainty and questions. At the time of writing this report, there are over 4,713,026 confirmed cases of COVID-19 across the globe. The COVID-19 pandemic has been greatly affecting people's lives and the world's economy. Among many infection related questions, governments and people are most concerned with (i) when the COVID-19 outbreak will peak; (ii) how long the outbreak will last and (iii) how many people will eventually be infected.

Through this project, we will try to understand and analyze the trend of COVID 19 in different states in the US, predict its damage so that effective measures can be taken against it. The forecasts obtained from this project can help inform public health decision-making by projecting the likely impact in coming days. This understanding can help the government gauge the current level f preparedness and devote appropriate resources and medical services in regions requiring critical attention.

The stakeholders for the project are US Government and Health Department as this analysis would help them determine the resources and medical services required in the near future. This report is also beneficial to all the people who have been affected by the lockdown.

DATA

To perform the required analysis, data has been extracted from two sources, namely, Kaggle and Foursquare.
COVID-19 dataset has been taken from Kaggle. The features of the dataset are as follows :

- **Sno** - Serial number
- **ObservationDate** - Date of the observation in MM/DD/YYYY
- **Province/State** - Province or state of the observation (Could be empty when missing) **Country/Region** - Country of observation
- **Last Update** - Time in UTC at which the row is updated for the given province or country. Confirmed - Cumulative number of confirmed cases till that date
- **Deaths** - Cumulative number of deaths till that date

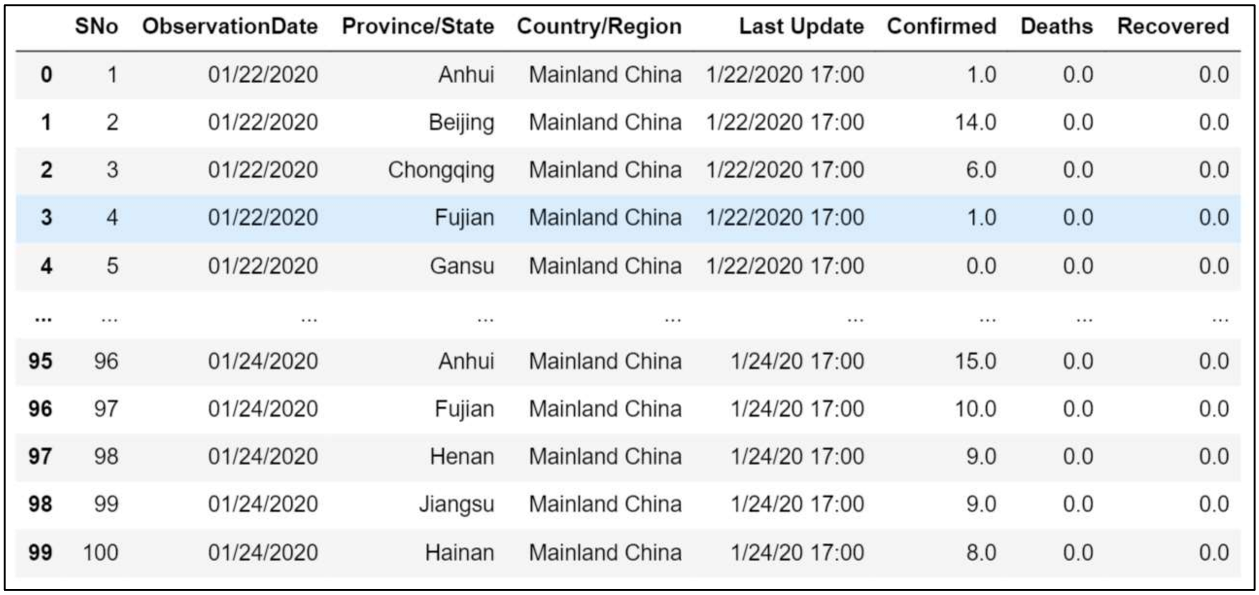
- **Recovered** - Cumulative number of recovered cases till that date

Datasource : https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Foursquare Location Dataset

It provides data about different places all around the world.

Features and Examples: Analysis of different geographic locations including exploring places, finding trending locations near a venue, exploring users and their reviews about places can be done through Foursquare location dataset.



	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
95	96	01/24/2020	Anhui	Mainland China	1/24/20 17:00	15.0	0.0	0.0
96	97	01/24/2020	Fujian	Mainland China	1/24/20 17:00	10.0	0.0	0.0
97	98	01/24/2020	Henan	Mainland China	1/24/20 17:00	9.0	0.0	0.0
98	99	01/24/2020	Jiangsu	Mainland China	1/24/20 17:00	9.0	0.0	0.0
99	100	01/24/2020	Hainan	Mainland China	1/24/20 17:00	8.0	0.0	0.0

1 METHODOLOGY

The methodology adopted for this project includes the following :

- Data Collection
- Exploratory Data Analysis
- Data Classification
- Graphical Visualization
- Insights and Conclusions

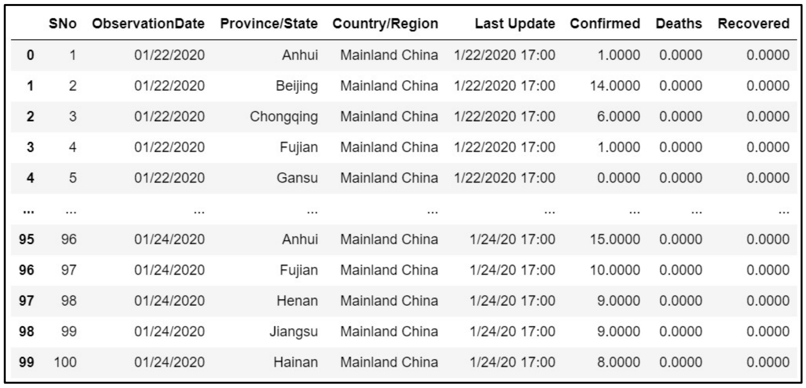
The following pages contain description of all the steps involved in the methodology.

1.1 Data Collection

The data is captured through various sources like Kaggle and FourSquare API in the form of excel spreadsheets (csv format). Let us elaborate on the captured Datasets in detail.

Covid-19 World Data:

Here is the Covid-19 data that has been taken from Kaggle. Dataset contains 25959 rows and 8 columns. The dataset contains data for 223 countries and 355 states all over the globe.

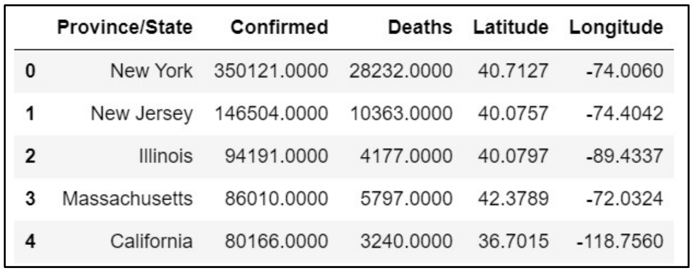


	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0000	0.0000	0.0000
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0000	0.0000	0.0000
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0000	0.0000	0.0000
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0000	0.0000	0.0000
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0000	0.0000	0.0000
...	...	...	...	...	...	...	...	...
95	96	01/24/2020	Anhui	Mainland China	1/24/20 17:00	15.0000	0.0000	0.0000
96	97	01/24/2020	Fujian	Mainland China	1/24/20 17:00	10.0000	0.0000	0.0000
97	98	01/24/2020	Henan	Mainland China	1/24/20 17:00	9.0000	0.0000	0.0000
98	99	01/24/2020	Jiangsu	Mainland China	1/24/20 17:00	9.0000	0.0000	0.0000
99	100	01/24/2020	Hainan	Mainland China	1/24/20 17:00	8.0000	0.0000	0.0000

Figure 1 Covid-19 World Data

Location Data:

We have deployed OpenCageGeocode API to determine the location coordinates for each province/state in the United States.

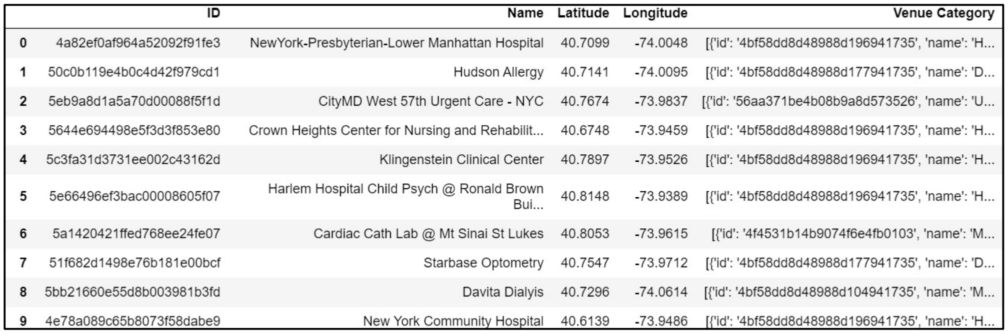


	Province/State	Confirmed	Deaths	Latitude	Longitude
0	New York	350121.0000	28232.0000	40.7127	-74.0060
1	New Jersey	146504.0000	10363.0000	40.0757	-74.4042
2	Illinois	94191.0000	4177.0000	40.0797	-89.4337
3	Massachusetts	86010.0000	5797.0000	42.3789	-72.0324
4	California	80166.0000	3240.0000	36.7015	-118.7560

Figure 2 State wise Location Data

Foursquare Hospital Data:

Using the Foursquare API, we will collect healthcare data for the state with highest number of confirmed cases in the United States.



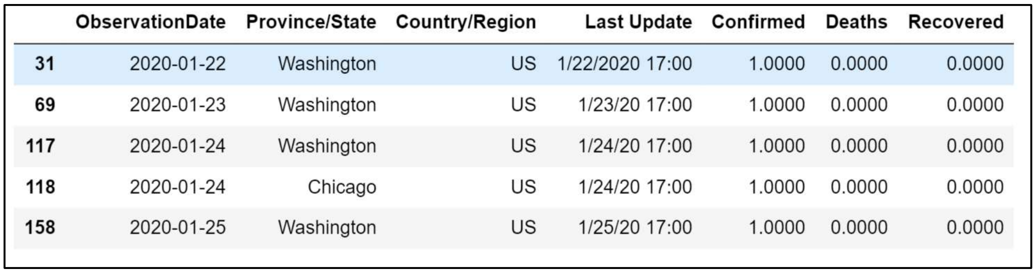
	ID	Name	Latitude	Longitude	Venue Category
0	4a82ef0af964a52092f91fe3	NewYork-Presbyterian-Lower Manhattan Hospital	40.7099	-74.0048	[{'id': '4bf58dd8d48988d196941735', 'name': 'H...
1	50c0b119e4b0c4d42f979cd1	Hudson Allergy	40.7141	-74.0095	[{'id': '4bf58dd8d48988d177941735', 'name': 'D...
2	5eb9a8d1a5a70d00088f5f1d	CityMD West 57th Urgent Care - NYC	40.7674	-73.9837	[{'id': '56aa371be4b08b9a8d573526', 'name': 'U...
3	5644e694498e5f3d3f853e80	Crown Heights Center for Nursing and Rehabilit...	40.6748	-73.9459	[{'id': '4bf58dd8d48988d196941735', 'name': 'H...
4	5c3fa31d3731ee002c43162d	Klingenstein Clinical Center	40.7897	-73.9526	[{'id': '4bf58dd8d48988d196941735', 'name': 'H...
5	5e66496ef3bac00008605f07	Harlem Hospital Child Psych @ Ronald Brown Bui...	40.8148	-73.9389	[{'id': '4bf58dd8d48988d196941735', 'name': 'H...
6	5a1420421ffed768ee24fe07	Cardiac Cath Lab @ Mt Sinai St Lukes	40.8053	-73.9615	[{'id': '4f4531b14b9074f6e4fb0103', 'name': 'M...
7	51f682d1498e76b181e00bcf	Starbase Optometry	40.7547	-73.9712	[{'id': '4bf58dd8d48988d177941735', 'name': 'D...
8	5bb21660e55d8b003981b3fd	Davita Dialyis	40.7296	-74.0614	[{'id': '4bf58dd8d48988d104941735', 'name': 'M...
9	4e78a089c65b8073f58dabe9	New York Community Hospital	40.6139	-73.9486	[{'id': '4bf58dd8d48988d196941735', 'name': 'H...

Figure 3 Foursquare Medical Facilities Data

1.2 Exploratory Data Analysis

1.2.1 Data Preprocessing

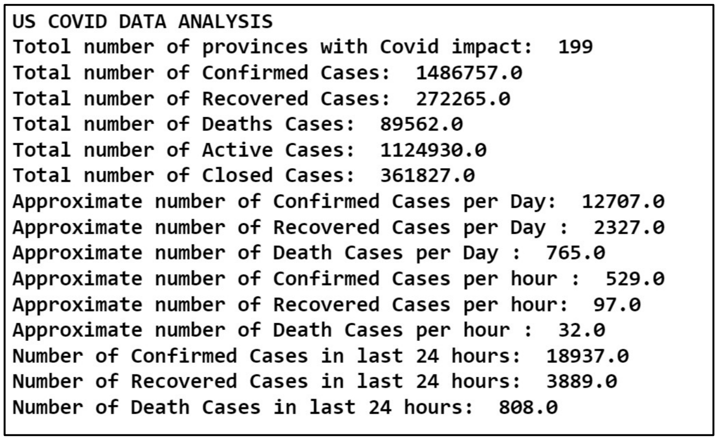
Let us filter the World data to obtain information relevant to the United States. After cleaning the dataset we get the following dataframe:



	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
31	2020-01-22	Washington	US	1/22/2020 17:00	1.0000	0.0000	0.0000
69	2020-01-23	Washington	US	1/23/20 17:00	1.0000	0.0000	0.0000
117	2020-01-24	Washington	US	1/24/20 17:00	1.0000	0.0000	0.0000
118	2020-01-24	Chicago	US	1/24/20 17:00	1.0000	0.0000	0.0000
158	2020-01-25	Washington	US	1/25/20 17:00	1.0000	0.0000	0.0000

Figure 4 US Covid-19 Data

Extracting basic insights from this dataset, we get the following results:



```
US COVID DATA ANALYSIS
Totol number of provinces with Covid impact:  199
Total number of Confirmed Cases:  1486757.0
Total number of Recovered Cases:  272265.0
Total number of Deaths Cases:  89562.0
Total number of Active Cases:  1124930.0
Total number of Closed Cases:  361827.0
Approximate number of Confirmed Cases per Day:  12707.0
Approximate number of Recovered Cases per Day :  2327.0
Approximate number of Death Cases per Day :  765.0
Approximate number of Confirmed Cases per hour :  529.0
Approximate number of Recovered Cases per hour:  97.0
Approximate number of Death Cases per hour :  32.0
Number of Confirmed Cases in last 24 hours:  18937.0
Number of Recovered Cases in last 24 hours:  3889.0
Number of Death Cases in last 24 hours:  808.0
```

Figure 5 US Covid-19 Data – Analysis

1.2.2 Statistical Analysis and Data Visualization

Now we will plot the daily increase in the number of deaths, confirmed cases and recovered cases in the United States.

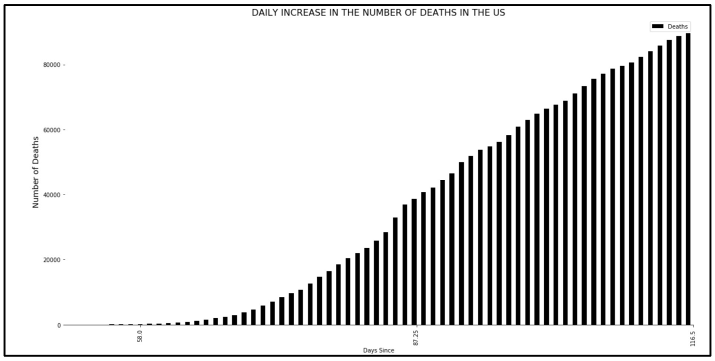


Figure 6 Daily increase in the number of deaths in US

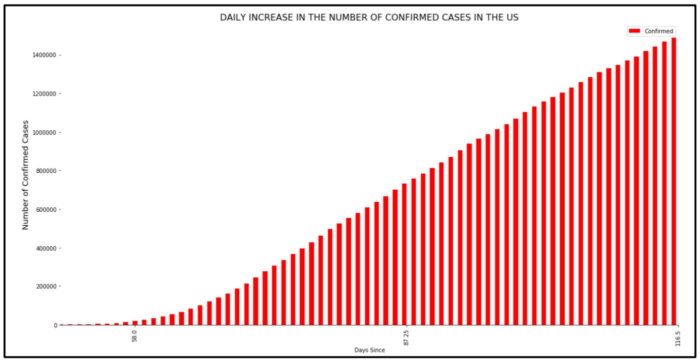


Figure 7 Daily increase in the number confirmed cases in US

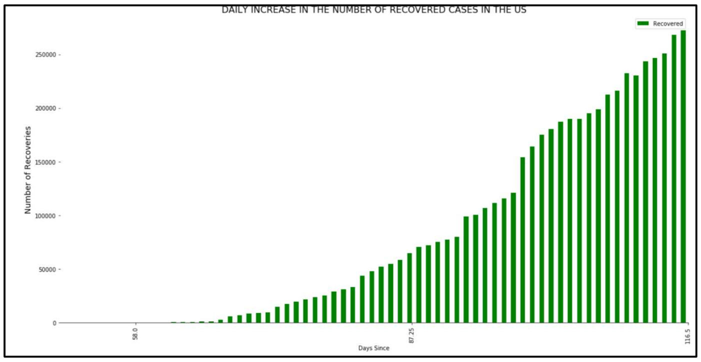


Figure 8 Daily increase in the number of recovered cases in US

Now plotting the weekly increase in the number of confirmed cases, deaths and the number of recovered cases in the US.

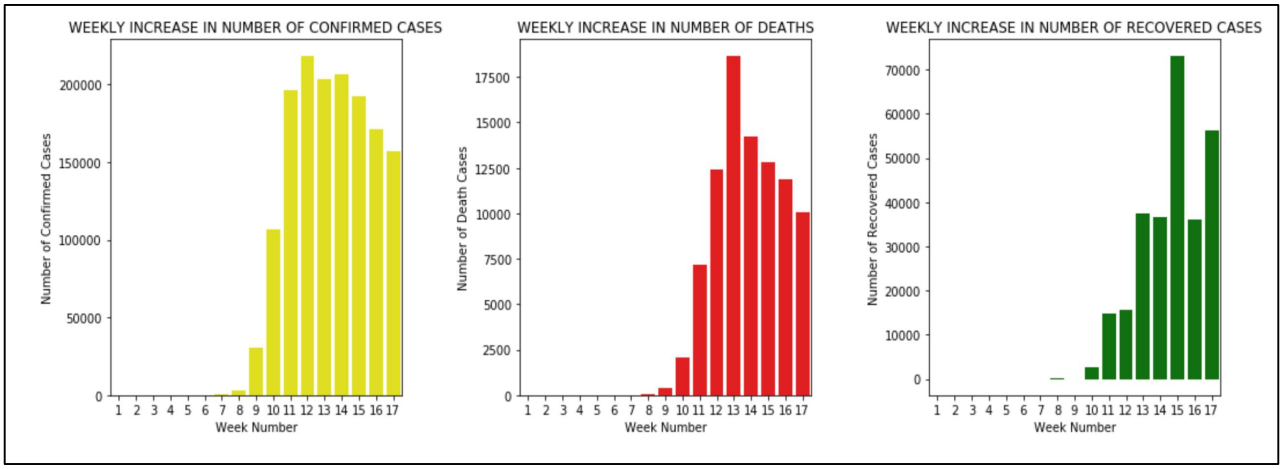
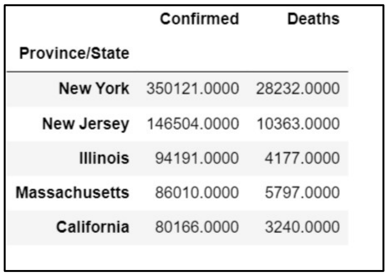


Figure 9 Weekly increase in the number of cases in US

Now, we will perform state wise analysis on the US dataset. Grouping the data according to state gives the following dataframe:



	Confirmed	Deaths
Province/State		
New York	350121.0000	28232.0000
New Jersey	146504.0000	10363.0000
Illinois	94191.0000	4177.0000
Massachusetts	86010.0000	5797.0000
California	80166.0000	3240.0000

Figure 9 State wise Data

Now we will focus on developing insights from the state wise data.

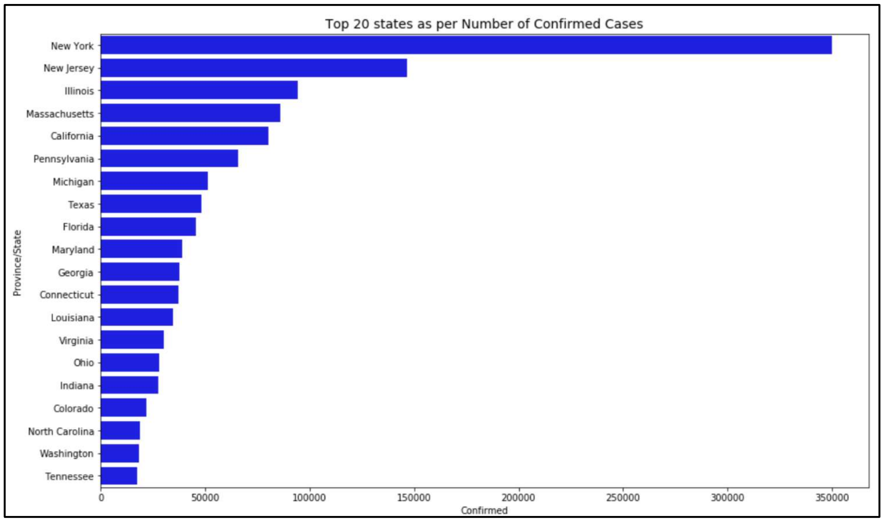


Figure 10 Top 20 states as per number of confirmed cases

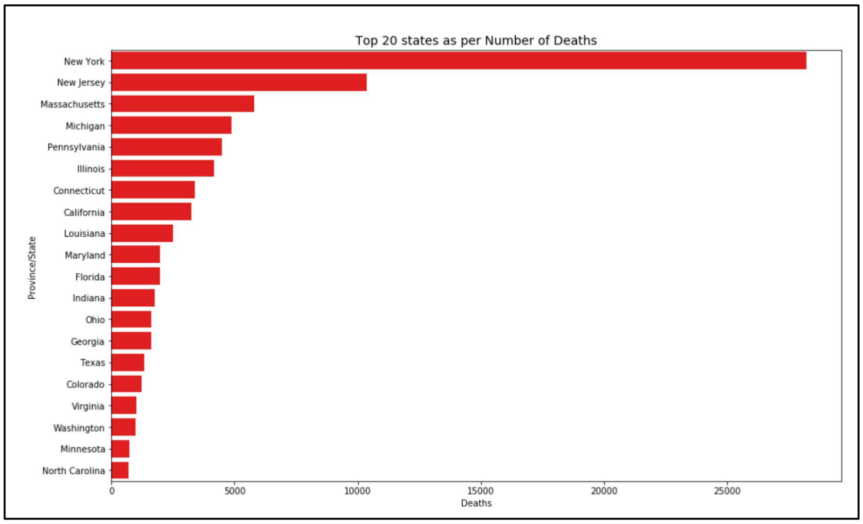


Figure 11 Top 20 states as per number of deaths

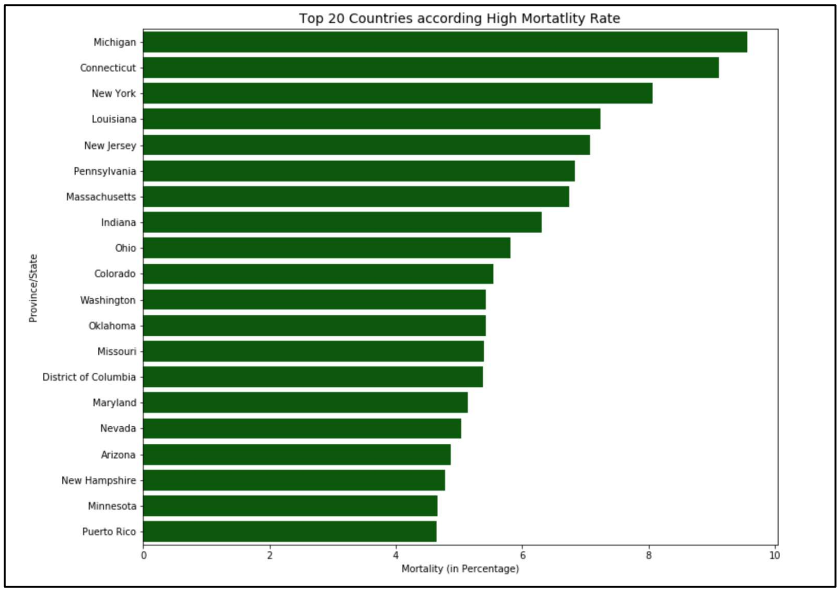


Figure 12 Top 20 states as per mortality rates

Next, we will be plotting the number of cases in different states of United States on a US map using plotly.

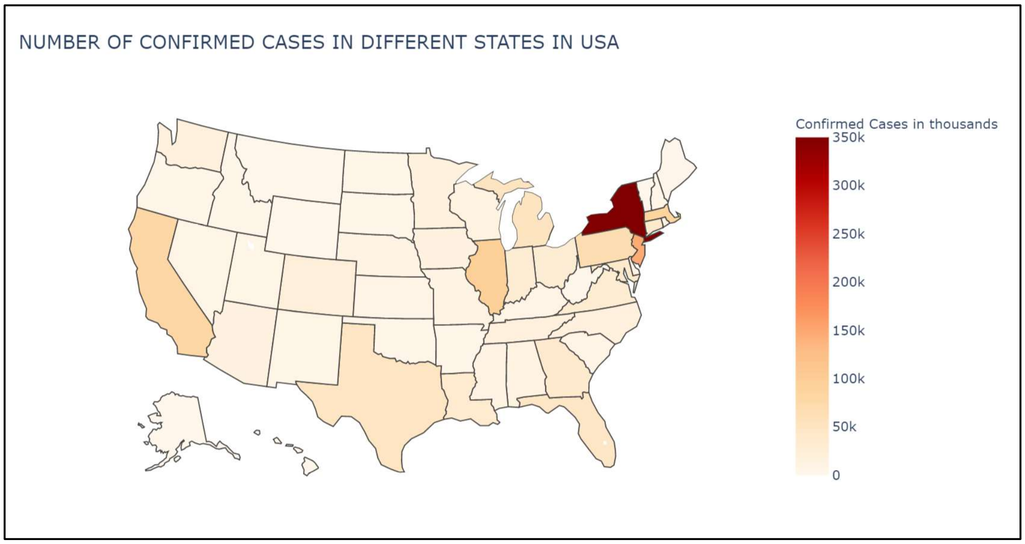

NUMBER OF CONFIRMED CASES IN DIFFERENT STATES IN USA
Confirmed Cases in thousands
350k
300k
250k
200k
150k
100k
50k
0

Figure 13 Number of confirmed cases in different states in USA

Figure 13 shows that New York, Illinois and California have relatively higher number of confirmed cases as compared to other states in United States. The same trend was observed in the number of deaths in different states in USA. This can be observed in Figure 14.

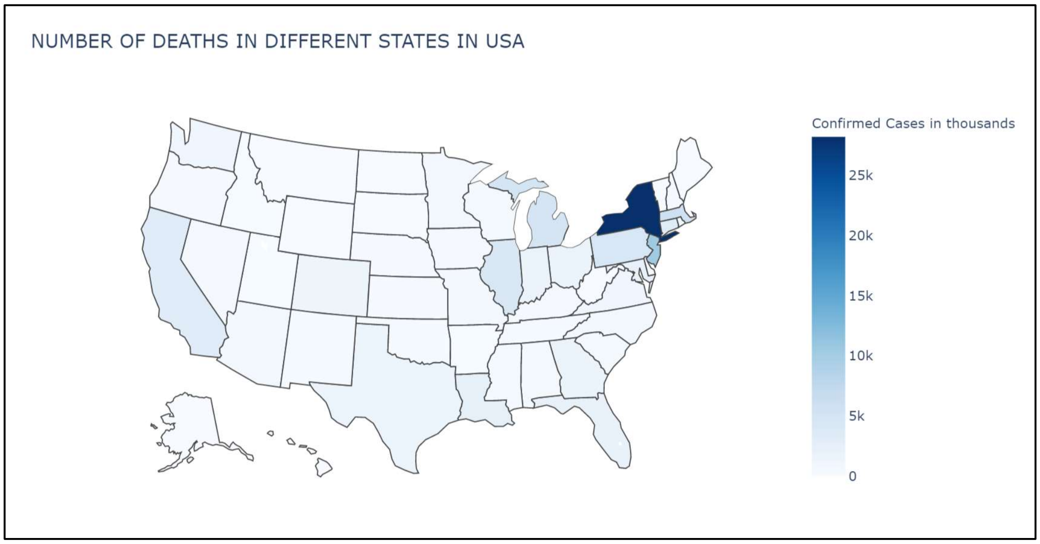

NUMBER OF DEATHS IN DIFFERENT STATES IN USA
Confirmed Cases in thousands
25k
20k
15k
10k
5k
0

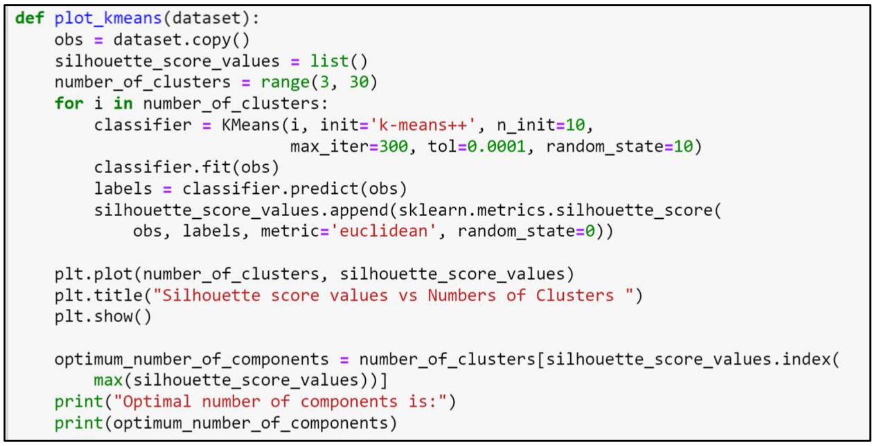
Figure 14 Number of deaths in different states in USA

1.3 Data Classification:

Let us now use K- means clustering to identify clusters within the dataset and partition the dataset. We will be using Elbow and silhouette method for finding the optimal value of k.

1.3.1 K-Means Clustering

For this we will use the Confirmed cases and Deaths as attributes. Python library sklearn will be used for fitting the model and finding clusters. The data was first normalized to obtain accurate results.

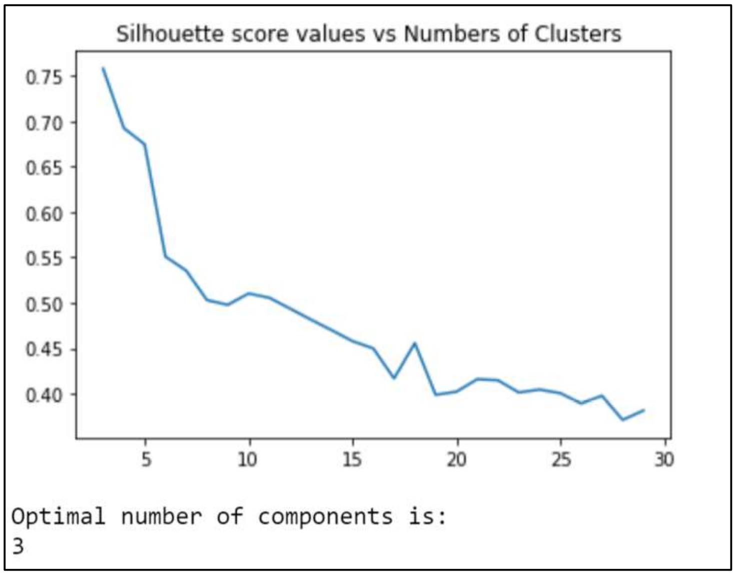


```python
def plot_kmeans(dataset):
    obs = dataset.copy()
    silhouette_score_values = list()
    number_of_clusters = range(3, 30)
    for i in number_of_clusters:
        classifier = KMeans(i, init='k-means++', n_init=10,
                            max_iter=300, tol=0.0001, random_state=10)
        classifier.fit(obs)
        labels = classifier.predict(obs)
        silhouette_score_values.append(sklearn.metrics.silhouette_score(
            obs, labels, metric='euclidean', random_state=0))

    plt.plot(number_of_clusters, silhouette_score_values)
    plt.title("Silhouette score values vs Numbers of Clusters ")
    plt.show()

    optimum_number_of_components = number_of_clusters[silhouette_score_values.index(
        max(silhouette_score_values))]
    print("Optimal number of components is:")
    print(optimum_number_of_components)
```

Figure 15 K Means Clustering

1.3.2 Parameter Optimization:

Using the elbow method, we determine the optimal number of k to be 3.



```
Optimal number of components is:
3
```

Figure 16 Elbow Method: Parameter Optimization

1.4 Cluster Analysis:

We will be analyzing the clusters obtained using scatter plots. Every color in these scatter plots represent a different cluster. Cluster 0 is a set of states which are very less affected, with comparatively low number of deaths and confirmed cases. Cluster 1 belongs to states which are worst affected with high number of deaths and confirmed cases. Cluster 2 is set of states which are severely affected, with high number of deaths and confirmed cases.

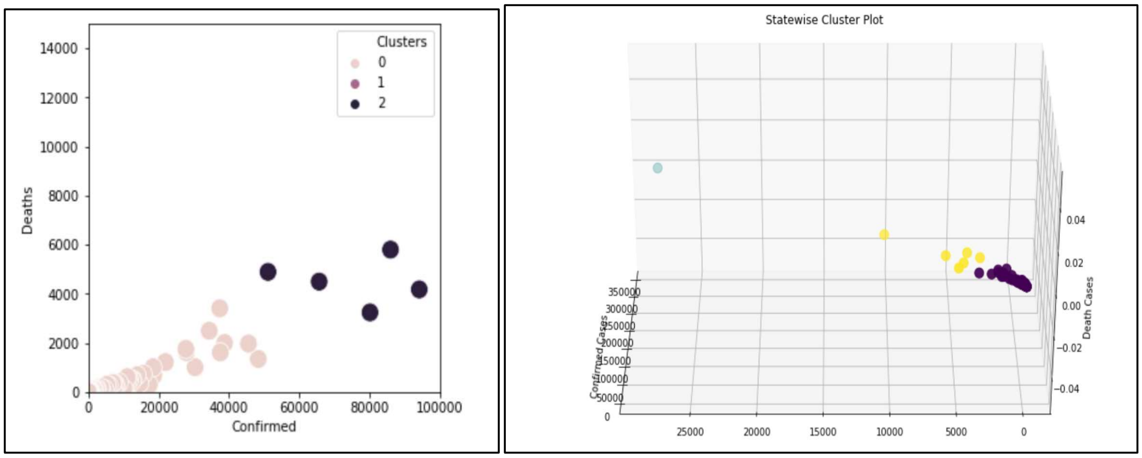
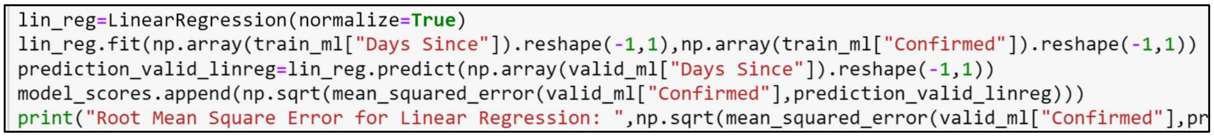


Figure 17 Elbow Method: Parameter Optimization

1.5 Using ML models for making predictions for the coming days:

1.5.1 Linear Regression

In this step we utilized Linear Regression and Polynomial Linear Regression to predict the number of confirmed cases in the coming days. Sklearn library in python was used to fit the two models and predict the values. Figure 18 shows the code used for fitting the model and making predictions.



```
lin_reg=LinearRegression(normalize=True)
lin_reg.fit(np.array(train_ml["Days Since"]).reshape(-1,1),np.array(train_ml["Confirmed"]).reshape(-1,1))
prediction_valid_linreg=lin_reg.predict(np.array(valid_ml["Days Since"]).reshape(-1,1))
model_scores.append(np.sqrt(mean_squared_error(valid_ml["Confirmed"],prediction_valid_linreg)))
print("Root Mean Square Error for Linear Regression: ",np.sqrt(mean_squared_error(valid_ml["Confirmed"],pr
```

Figure 18 Linear Regression for predicting confirmed cases

Linear regression does not fit the model well, which can be observed from the graph below.

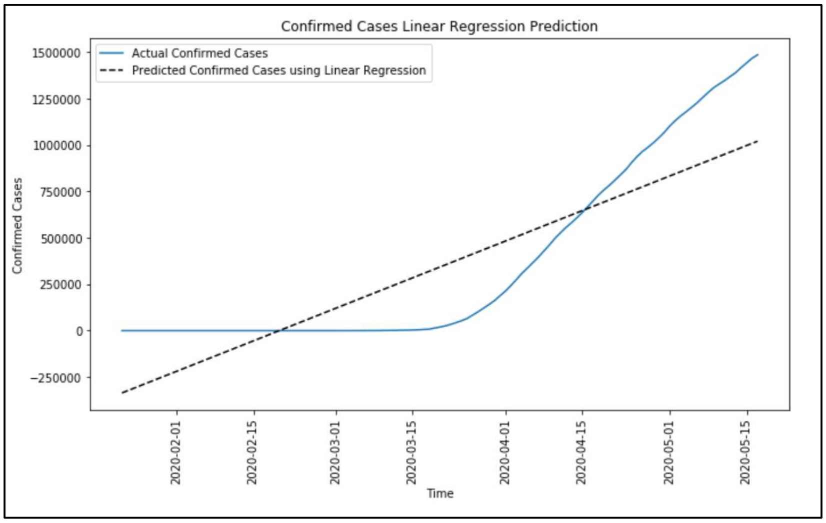
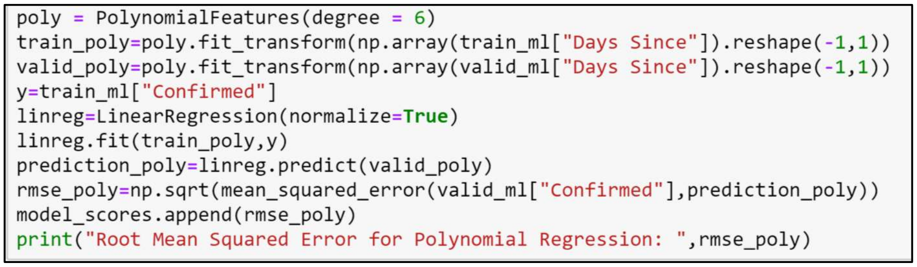


Figure 19: Confirmed Cases vs Predicted cases using Linear Regression

1.5.2 Polynomial Regression

Next polynomial regression was used to fit the model and make predictions.



```
poly = PolynomialFeatures(degree = 6)
train_poly=poly.fit_transform(np.array(train_ml["Days Since"]).reshape(-1,1))
valid_poly=poly.fit_transform(np.array(valid_ml["Days Since"]).reshape(-1,1))
y=train_ml["Confirmed"]
linreg=LinearRegression(normalize=True)
linreg.fit(train_poly,y)
prediction_poly=linreg.predict(valid_poly)
rmse_poly=np.sqrt(mean_squared_error(valid_ml["Confirmed"],prediction_poly))
model_scores.append(rmse_poly)
print("Root Mean Squared Error for Polynomial Regression: ",rmse_poly)
```

Figure 20 Polynomial Regression

Polynomial regression has a much lower value of Root Mean Square error as compared to linear regression. Polynomial Regression fits the model well as can be seen from Figure 21. The predictions obtained by both the models have been summarized in the results section.

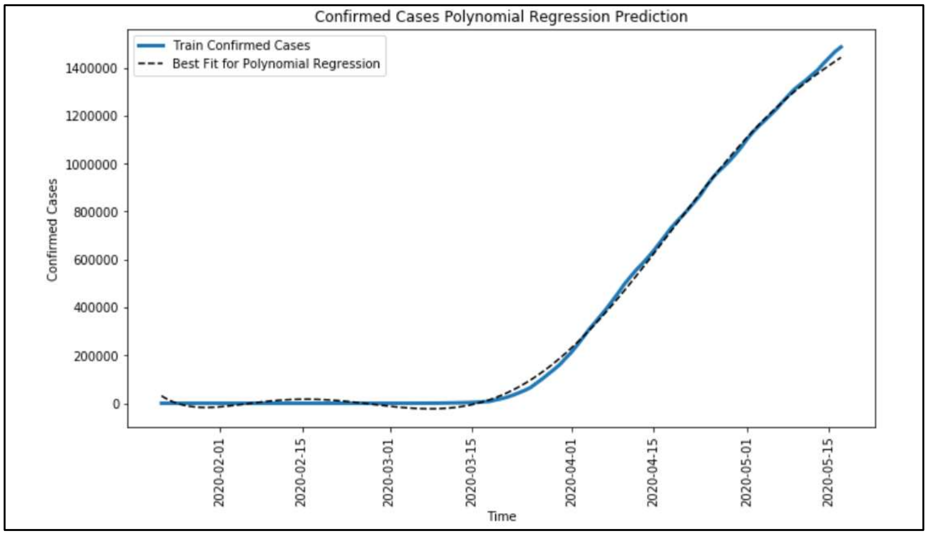
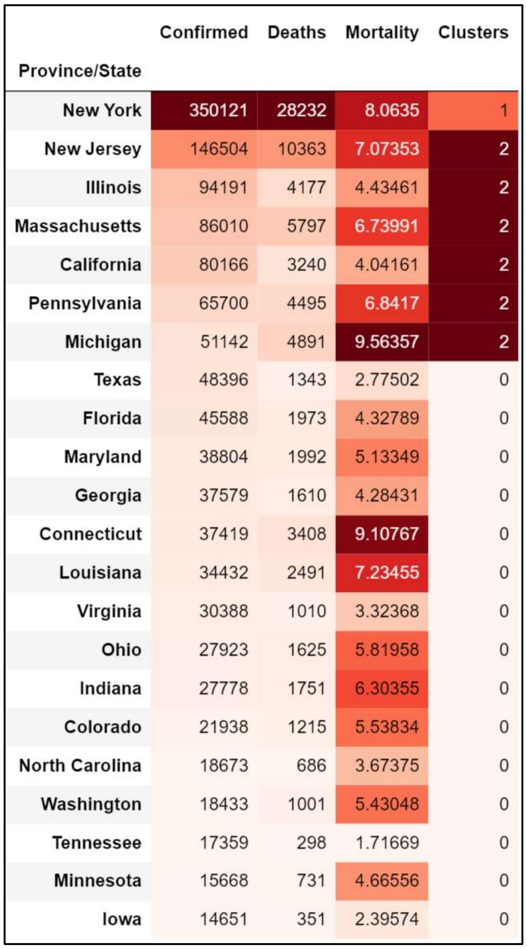


Figure 21: Confirmed Cases vs Predicted cases using Polynomial Regression

2 RESULTS

Below are the datasets attributed to each cluster identified:

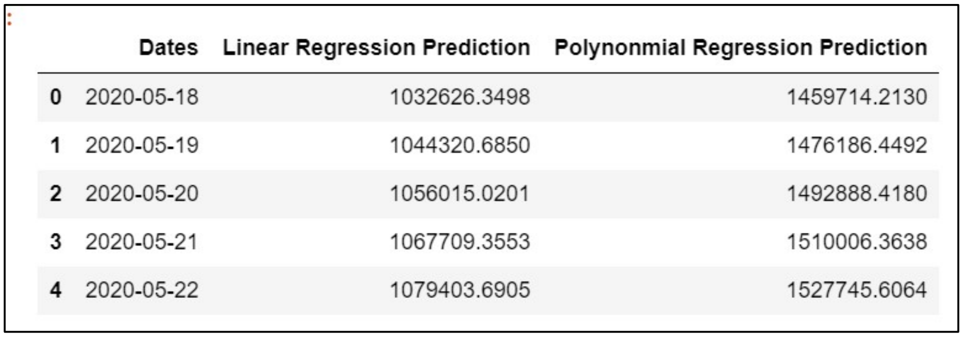
Figure 22 gives labels for Top 20 states according to confirmed cases in our dataframe. Cluster 0 is a set of states which are very less affected, with comparatively low number of deaths and confirmed cases. Cluster 1 belongs to states which are worst affected with high number of deaths and confirmed cases, this includes states like Illinois, California, Pennsylvania etc. Cluster 2 is set of states which are severely affected, with high number of deaths and confirmed cases, this cluster includes only one state New York which has been worst affected by Covid 19, with highest number of confirmed cases and deaths. Next, we will be fetching the latitude and longitude of New York using geocode and analyzing medical facilities in New York.



Province/State	Confirmed	Deaths	Mortality	Clusters
New York	350121	28232	8.0635	1
New Jersey	146504	10363	7.07353	2
Illinois	94191	4177	4.43461	2
Massachusetts	86010	5797	6.73991	2
California	80166	3240	4.04161	2
Pennsylvania	65700	4495	6.8417	2
Michigan	51142	4891	9.56357	2
Texas	48396	1343	2.77502	0
Florida	45588	1973	4.32789	0
Maryland	38804	1992	5.13349	0
Georgia	37579	1610	4.28431	0
Connecticut	37419	3408	9.10767	0
Louisiana	34432	2491	7.23455	0
Virginia	30388	1010	3.32368	0
Ohio	27923	1625	5.81958	0
Indiana	27778	1751	6.30355	0
Colorado	21938	1215	5.53834	0
North Carolina	18673	686	3.67375	0
Washington	18433	1001	5.43048	0
Tennessee	17359	298	1.71669	0
Minnesota	15668	731	4.66556	0
Iowa	14651	351	2.39574	0

Figure 2 Cluster Summary

Figure 23 shows the predictions for the total number of confirmed cases in the US for the coming days.



	Dates	Linear Regression Prediction	Polynonmial Regression Prediction
0	2020-05-18	1032626.3498	1459714.2130
1	2020-05-19	1044320.6850	1476186.4492
2	2020-05-20	1056015.0201	1492888.4180
3	2020-05-21	1067709.3553	1510006.3638
4	2020-05-22	1079403.6905	1527745.6064

Figure 23 Predictions of number of confirmed cases

Using FourSquare API for New York to analyze medical facilities in the state. The data was extracted for all the medical facilities in New York.

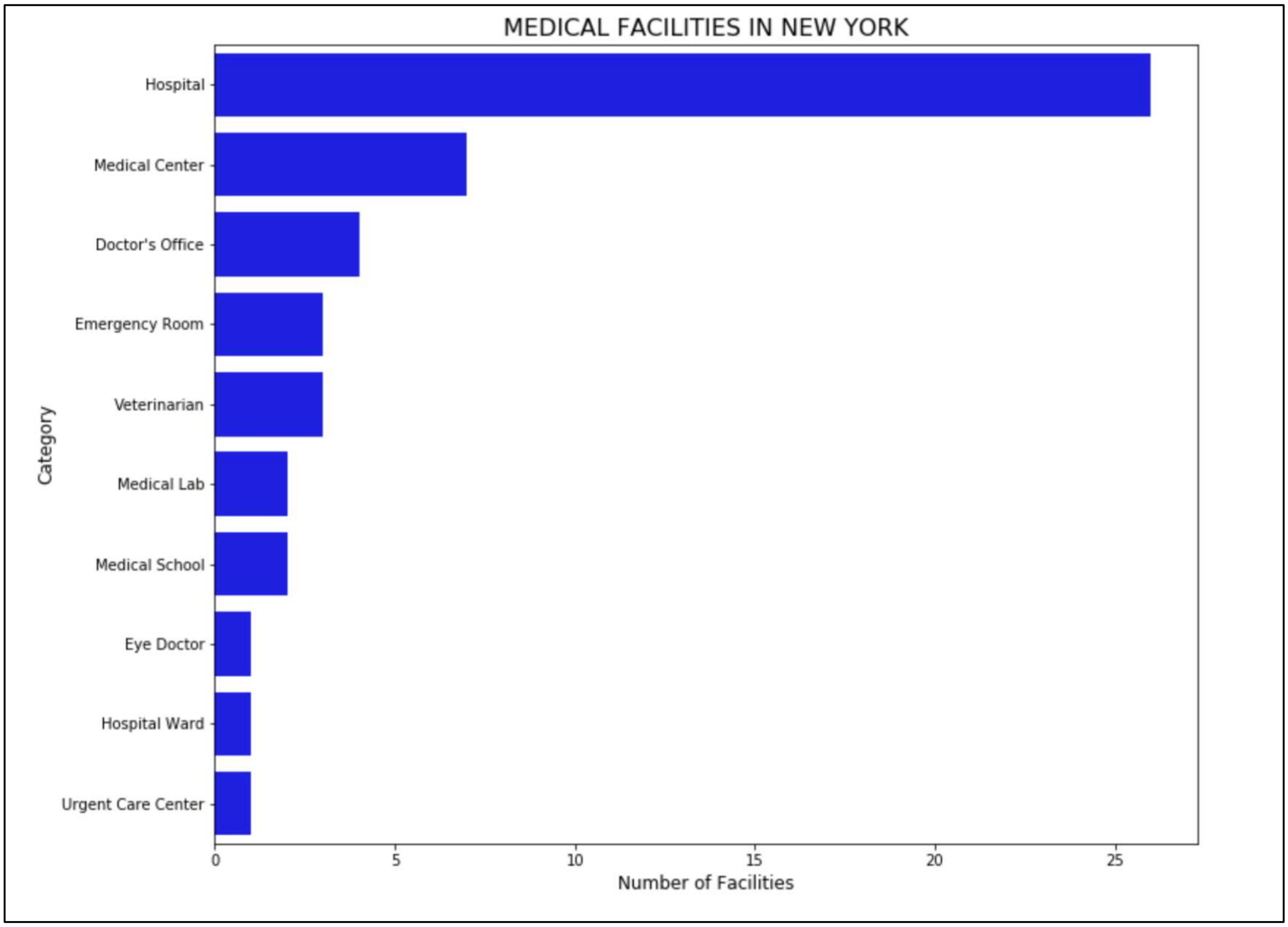


Figure 24: Analyzing Medical Facilities in New York

3 DISCUSSION

In this analysis, I analyzed the dataset to identify states with high deaths, confirmed cases and mortality rates. We also used K-Means to separate US states into three clusters based on number of cases in each state. Cluster 0 is a set of states which are very less affected, with comparatively low number of deaths and confirmed cases. Cluster 1 belongs to states which are worst affected with high number of deaths and confirmed cases. Cluster 2 is set of states which are severely affected, with high number of deaths and confirmed cases. We also analyzed the medical facilities available in New York, with highest number of confirmed cases using FourSqaure API. We also predicted the number of confirmed cases in the United States for the coming days using Linear Regression and Polynomial Regression.

Improvement Scope

- Accuracy of predictions can be improved by using other predictive models like SVM, Gradient Boost etc.

- **Foursquare API** has not been updated and does not contain the entire list of hospitals available in New York.

4 RECOMMENDATIONS

The stakeholders of the project which include US government and the US health department are already doing their best to flatten the curve of this pandemic and various relief packages are released for all affected people. Through this project they can identify which state belongs to which cluster and allocate medical resources and services accordingly. The predictions made for the coming days can be used by the health representatives to understand the trends and make efforts to mitigate the impact of COVID 19.

5 CONCLUSION

The purpose of this project is to gather actionable insights from COVID 19 data and predict the number of cases in the coming days. The trends show the number of cases is likely to increase in the near future. However, several measures can be taken by an individual to protect oneself and mitigate the impact of the pandemic. They are as follows:

1. Wash your hands often with soap and water for at least 20 seconds especially after you have been in a public place, or after blowing your nose, coughing, or sneezing.

2. If soap and water are not readily available, use a hand sanitizer that contains at least 60% alcohol. Cover all surfaces of your hands and rub them together until they feel dry.

3. Avoid touching your eyes, nose, and mouth with unwashed hands. 4. If surfaces are dirty, clean them: Use detergent or soap and water prior to disinfection.

We, on an individual level, can help fight this pandemic by keeping ourselves healthy and following the directives laid by respective governments.