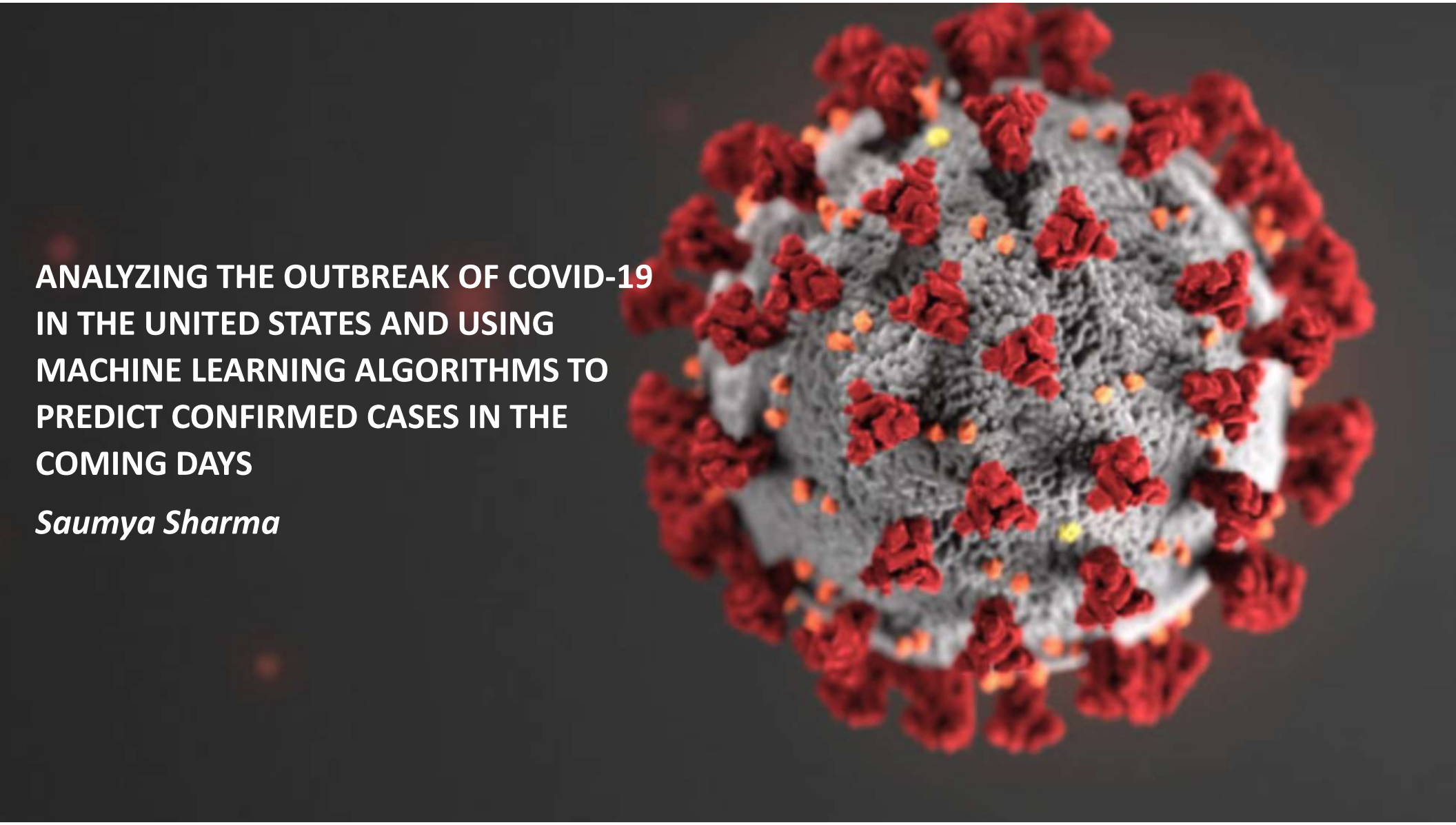
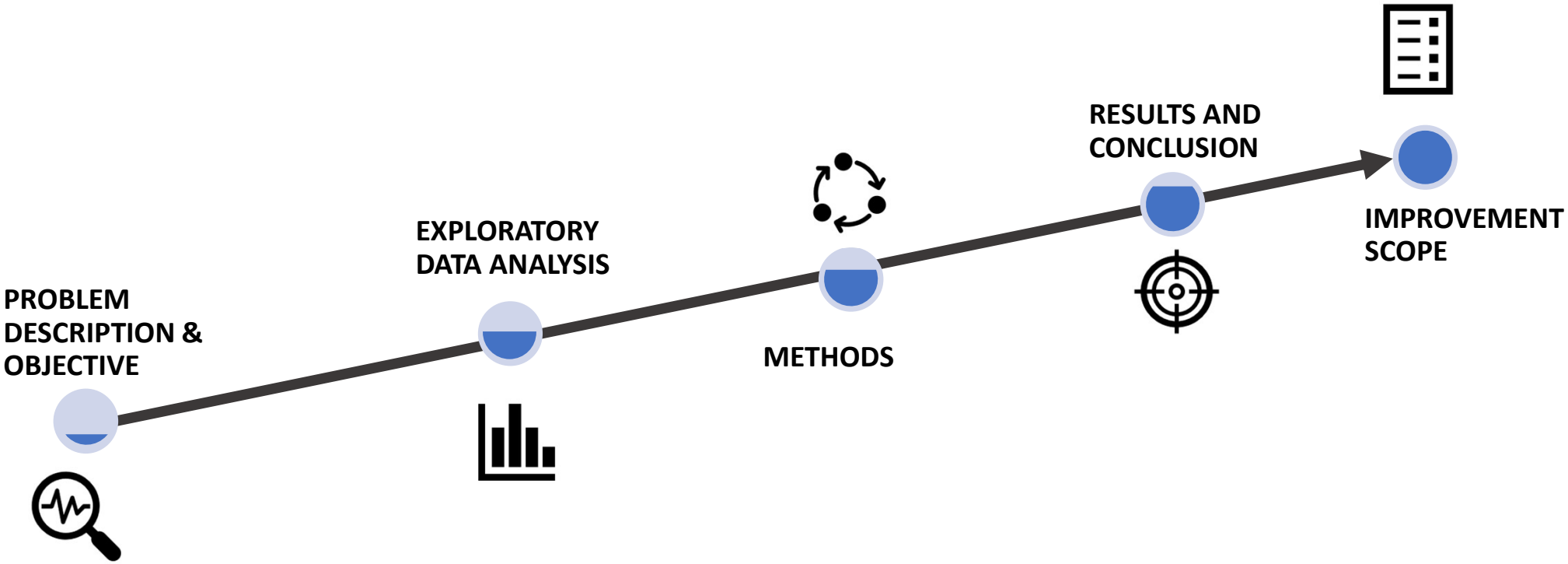
ANALYZING THE OUTBREAK OF COVID-19 IN THE UNITED STATES AND USING MACHINE LEARNING ALGORITHMS TO PREDICT CONFIRMED CASES IN THE COMING DAYS

Saumya Sharma



ROADMAP



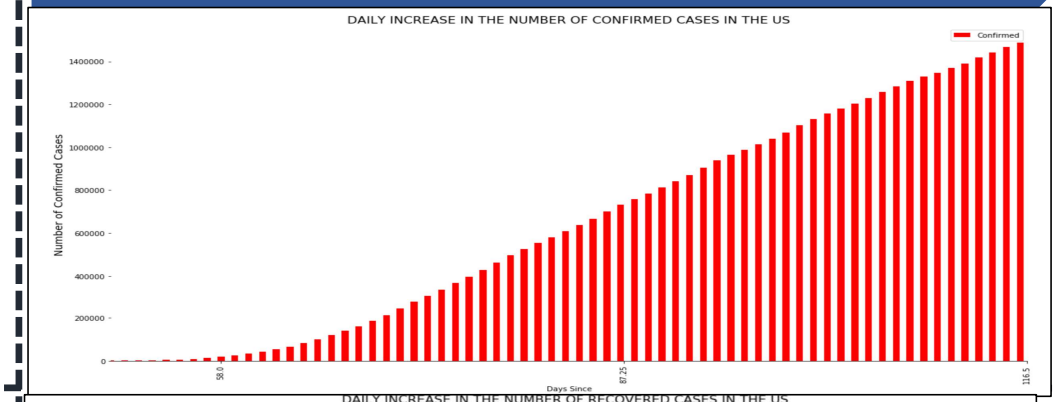
OVERVIEW

METHODOLOGY – Exploratory Data Analysis

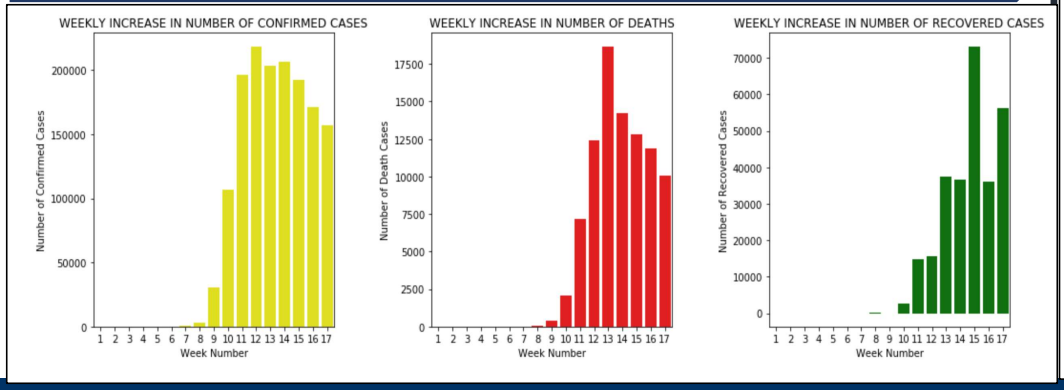
Dataset Preparation

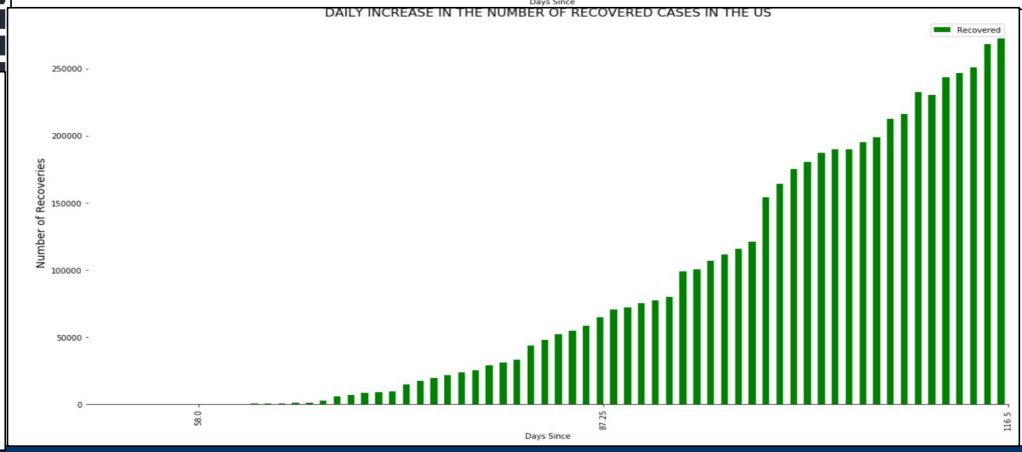
	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0000	0.0000	0.0000
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0000	0.0000	0.0000
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0000	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

Daily increase in the number of cases



Weekly Increase in number of cases

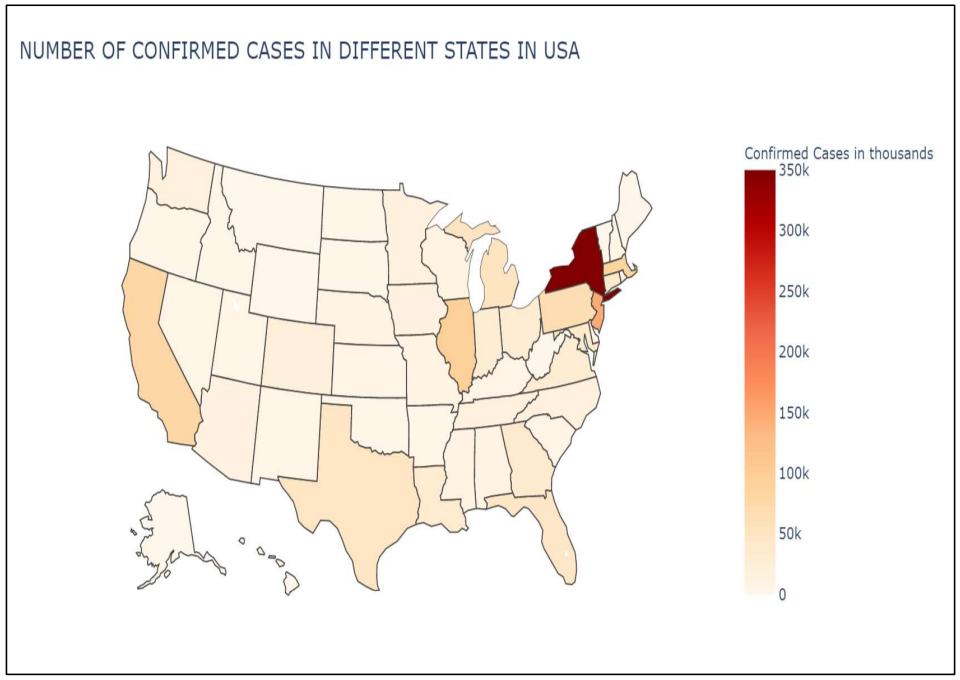




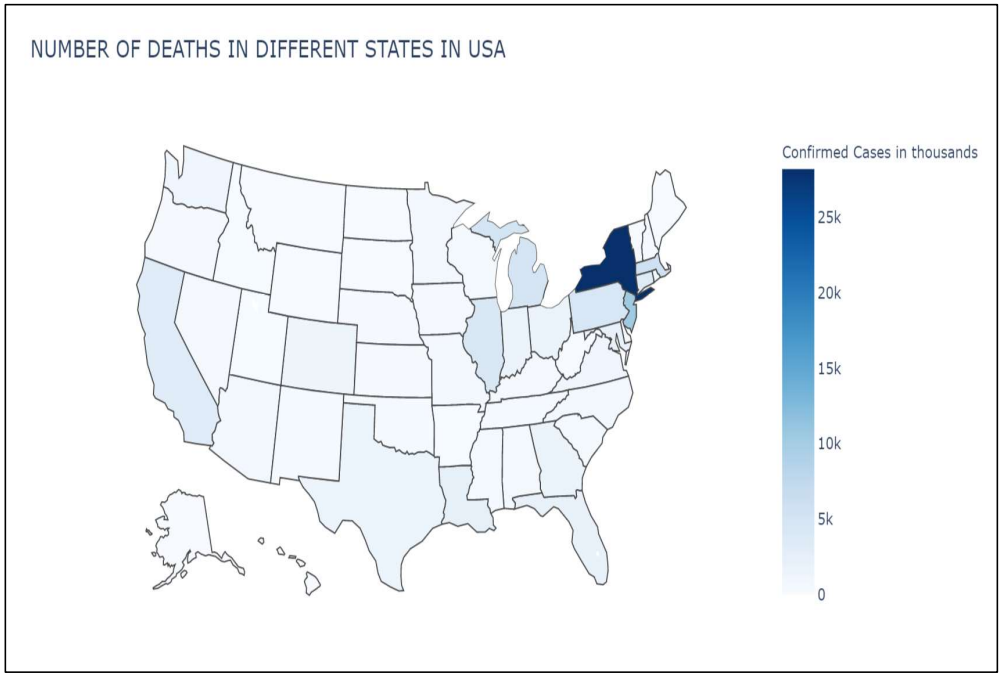
METHODOLOGY

METHODOLOGY – Exploratory Data Analysis

Number of Confirmed Cases in different states in USA

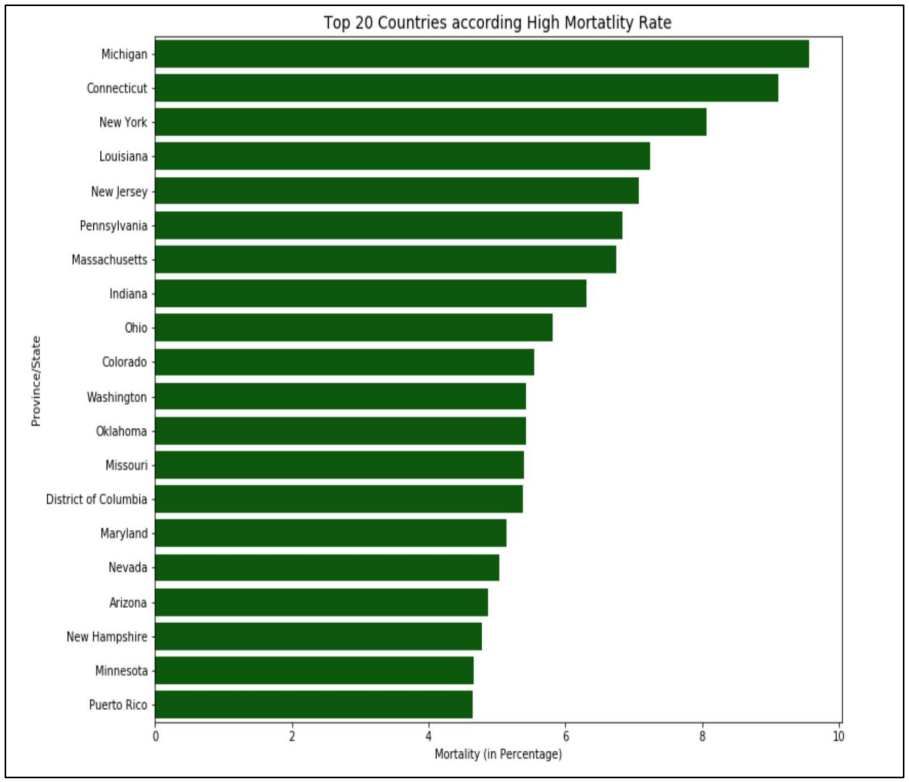


Number of Deaths in different states in USA

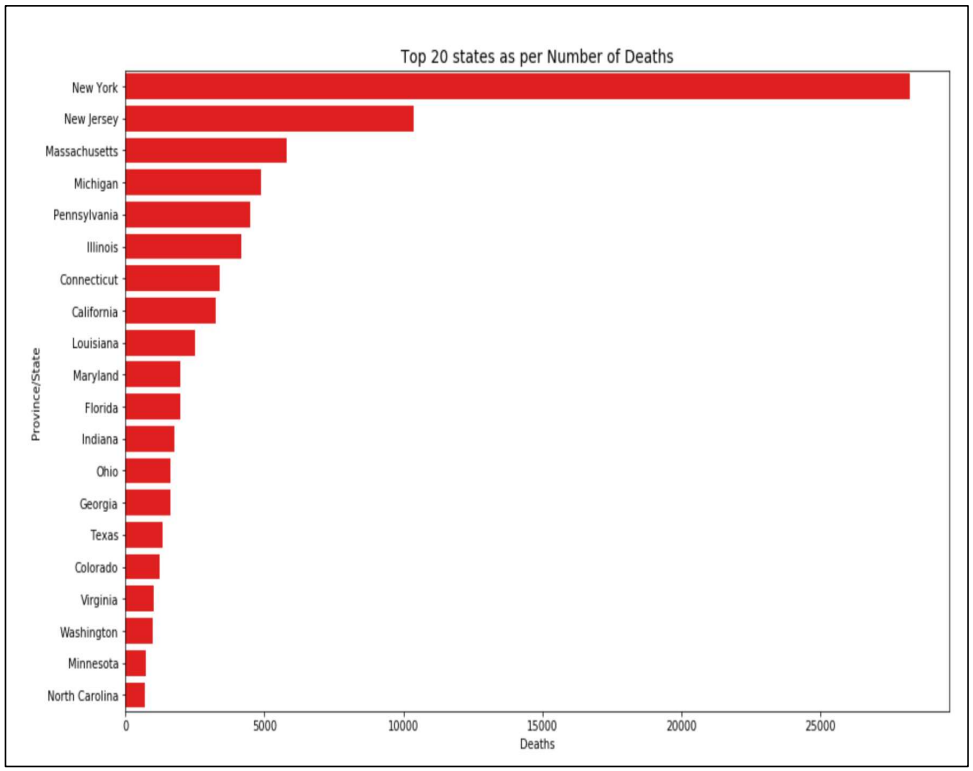


METHODOLOGY – Exploratory Data Analysis

TOP 20 states with high mortality rate



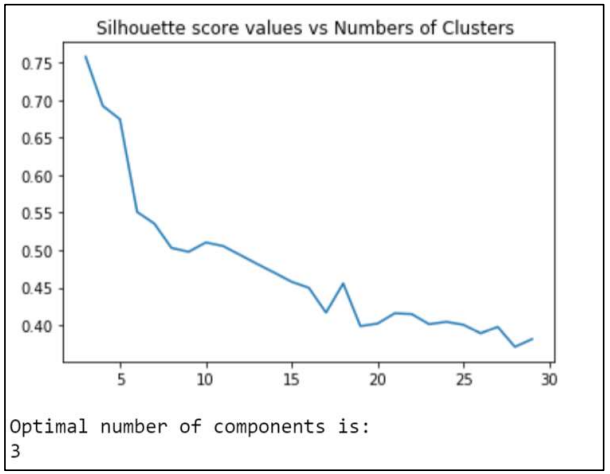
TOP 20 states with high deaths



METHODOLOGY – K Means Clustering

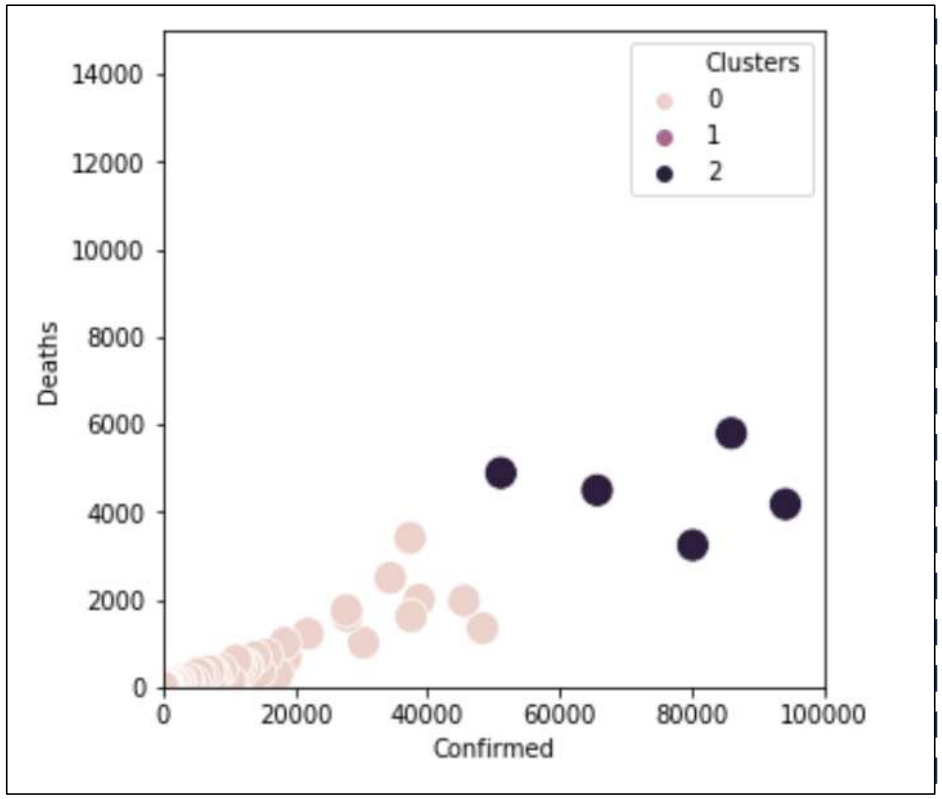
```python
def plot_kmeans(dataset):
    obs = dataset.copy()
    silhouette_score_values = list()
    number_of_clusters = range(3, 30)
    for i in number_of_clusters:
        classifier = KMeans(i, init='k-means++', n_init=10,
                            max_iter=300, tol=0.0001, random_state=10)
        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)
```

K Means Clustering



Parameter Optimization

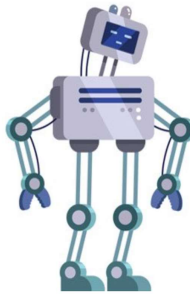
METHODOLOGY – Cluster Analysis



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

Cluster Analysis

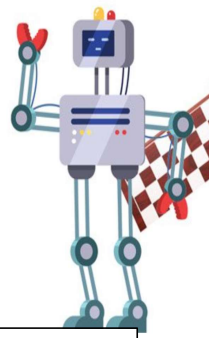
METHODOLOGY – Prediction Modeling

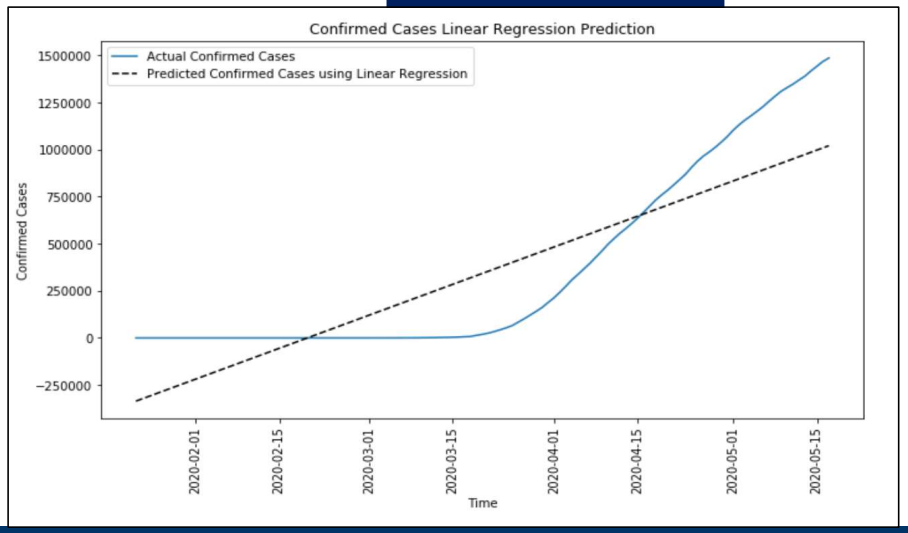


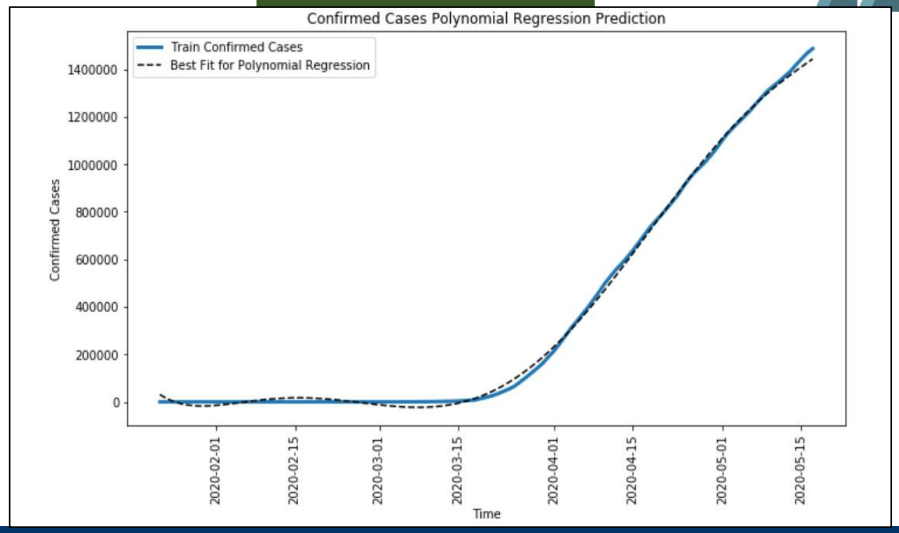
Predictive Modeling

Linear Regression

Polynomial Regression



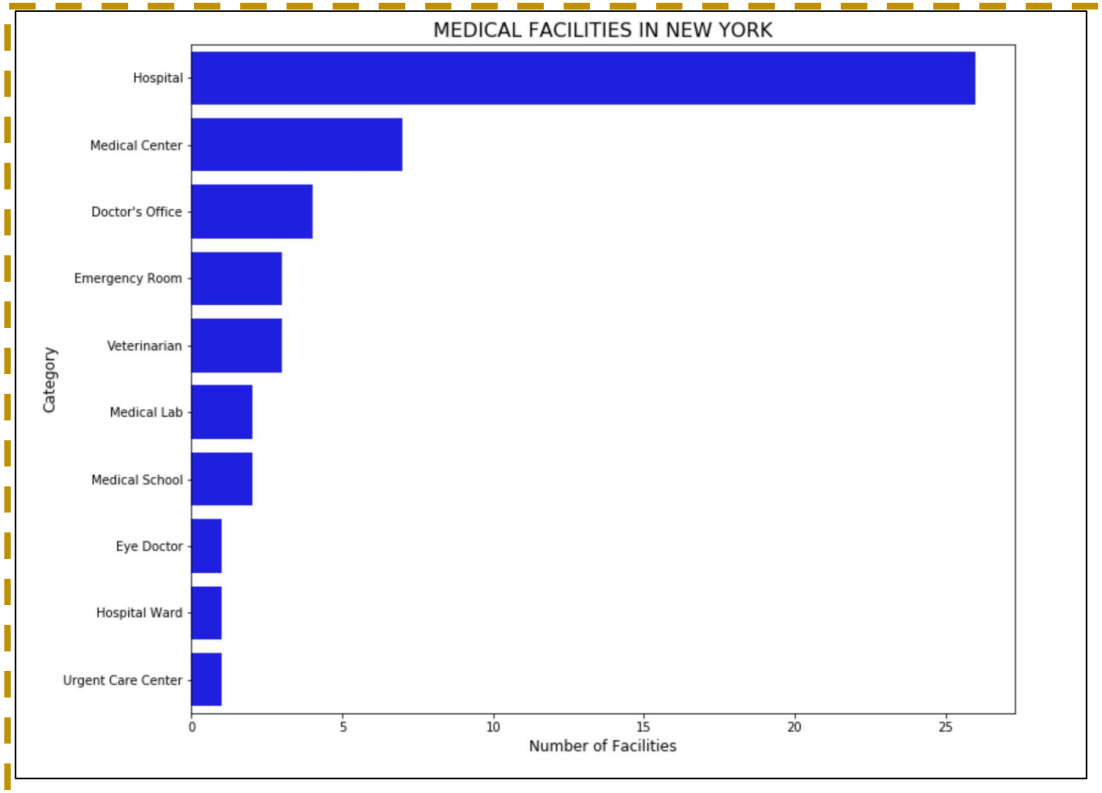




RESULTS

	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

Predictions Using Linear Regression and Polynomial Regression



Medical facilities in New York

CONCLUSION

The purpose of this project was 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.

FUTURE 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.

THANK YOU