INTRODUCTION

COVID-2019 is an infectious disease that comes from large family of viruses that cause illness ranging from the common cold to more severe diseases called Coronaviruses (CoV). Other examples of coronaviruses are: Middle East Respiratory Syndrome and Severe Acute Respiratory Syndrome.COVID-2019 was discovered in 2019 in Wuhan, China. The most common symptoms of COVID-19 are fever, tiredness, and dry cough. Some patients may have aches and pains, nasal congestion, runny nose, sore throat or diarrhea. These symptoms are usually mild and begin gradually (WHO, 2019). This project will help people to understand cov-2019 and the statistics behind it.

In this project, I will use Johns Hopkins dataset to achieve the following:

1. Data Preprocessing
2. Preforming exploratory data analysis on the world data, especially China data where the outbreak started, using Plotly, Matplotlib and seaborn
3. Visualizing the geospatial data concerning china using Folium
4. Building Linear Regression model to forecast the recovered cases

The aforementioned parts will be explained in details in the methodology section.

DATA

As mentioned above, the data used is Johns Hopkins dataset which contain the following fields:

1. County/Region
2. Province/State
3. Latitude
4. Longitude
5. Confirmed: Number of Confirmed Cases
6. Recovered: Number of Recovered Cases
7. Deaths
8. Date: Date of the report

This is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL). This set contains

sufficient amount of data that will help to carry on the aforementioned tasks (Gabriel Preda, 2019/2020).

METHODOLOGY

This section is divided into four parts which are: Data Preprocessing, exploratory data analysis (EDA), geospatial data visualization and Linear Regression model to forecast the recovered cases.

I. Data Preprocessing

At this stage, I checked the missing data:

```
covid_2019.isna().sum()
```

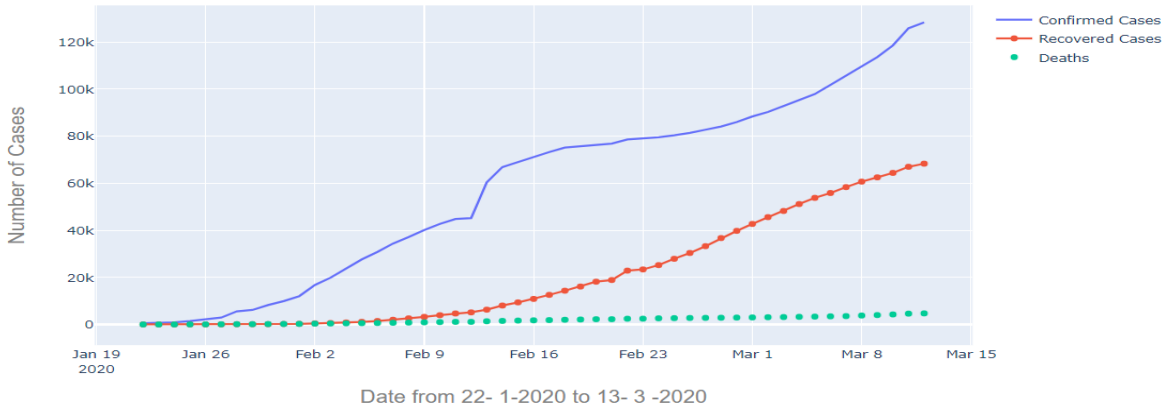
```
Out[186]: Country/Region        0
          Province/State     1924
          Latitude              1
          Longitude             1
          Confirmed            19
          Recovered           388
          Deaths              441
          Date                  0
          dtype: int64
```

If the provinces is null, changed to the name of the corresponding county/region. Dropped the row with missing latitude and longitude.

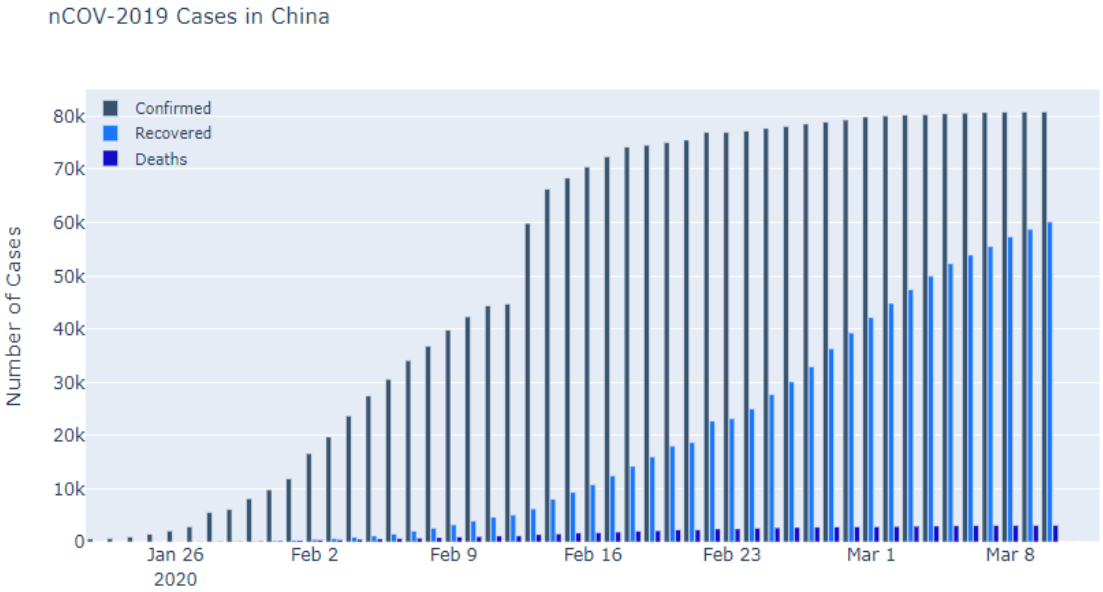
The missing in columns (Confirmed, recovered, Deaths) was replaced by 0.

II. Exploratory Data Analysis (EDA)

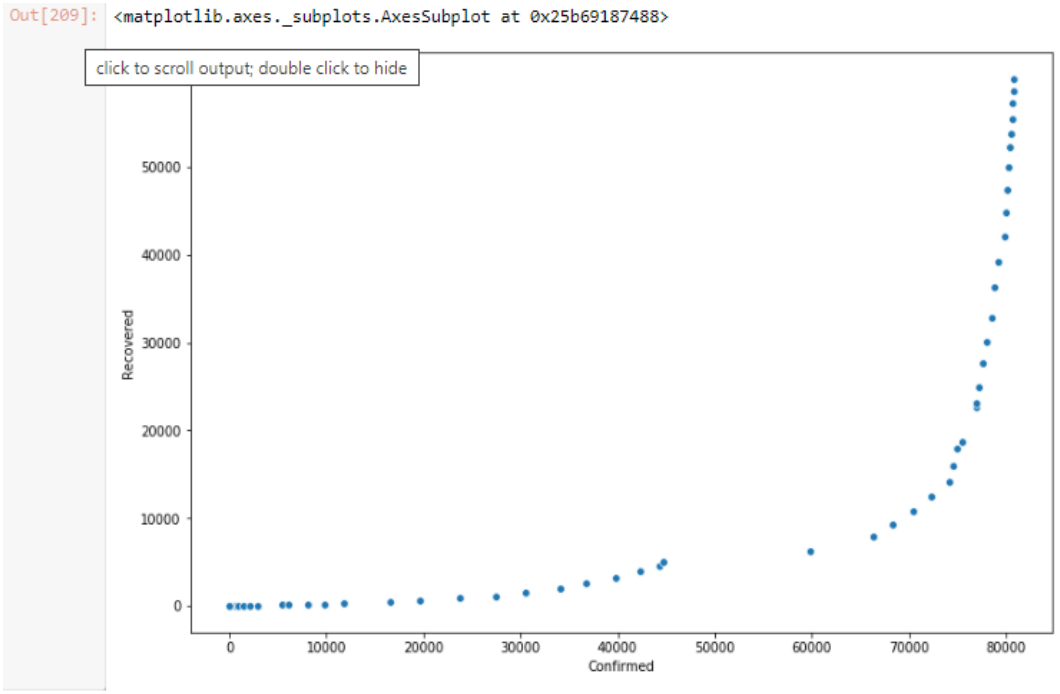
- Visualized the confirmed, recovered, and death cases worldwide.



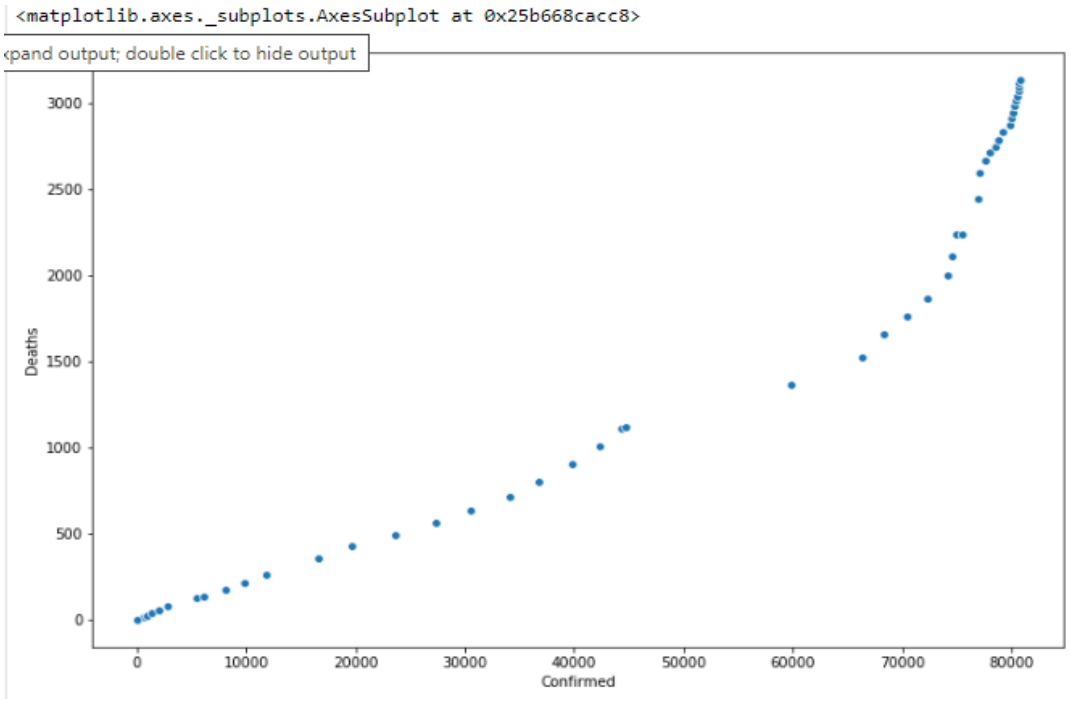
- Visualized the confirmed, recovered, and death cases only in China



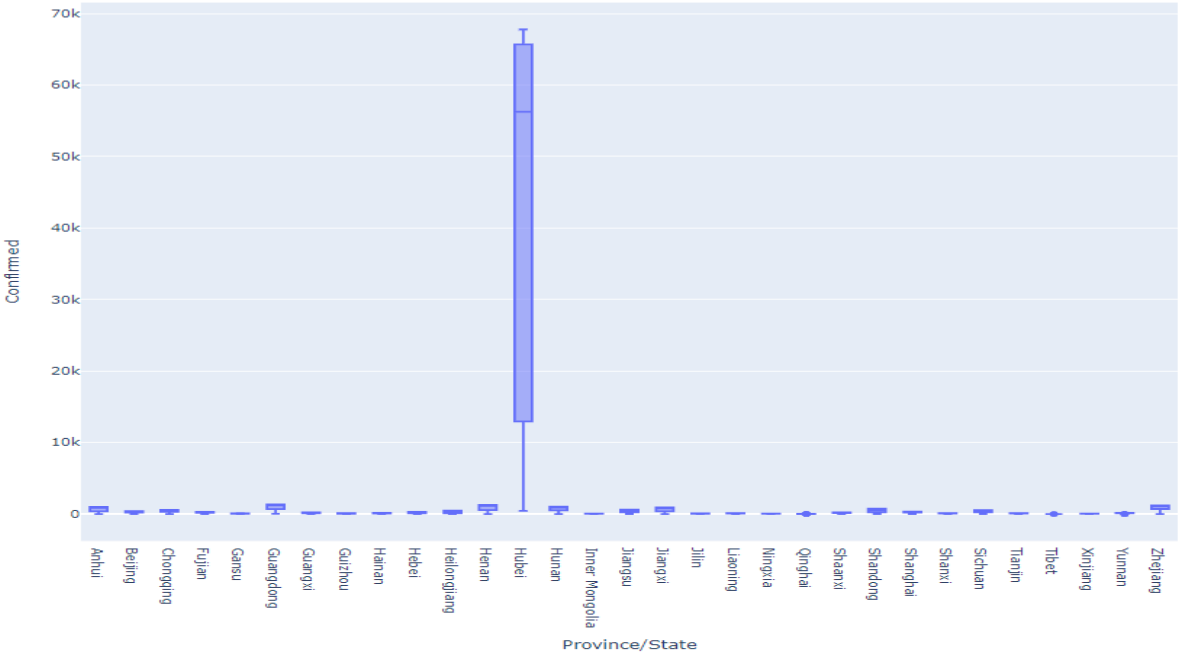
- Confirmed VS recovered cases in China



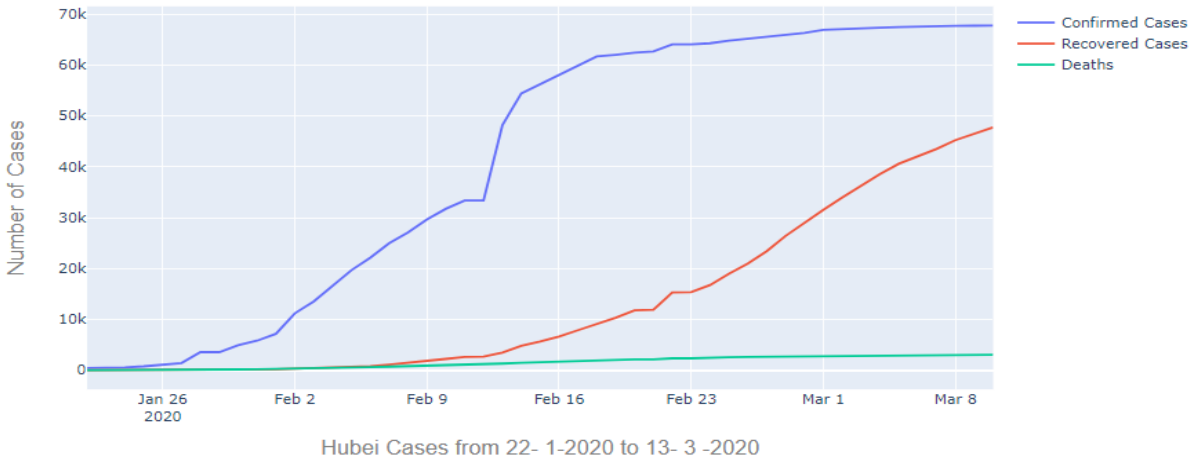
- Confirmed VS deaths cases in China



- Boxplot to show the confirmed cases grouped by provinces. Hubei has the highest number of cases.

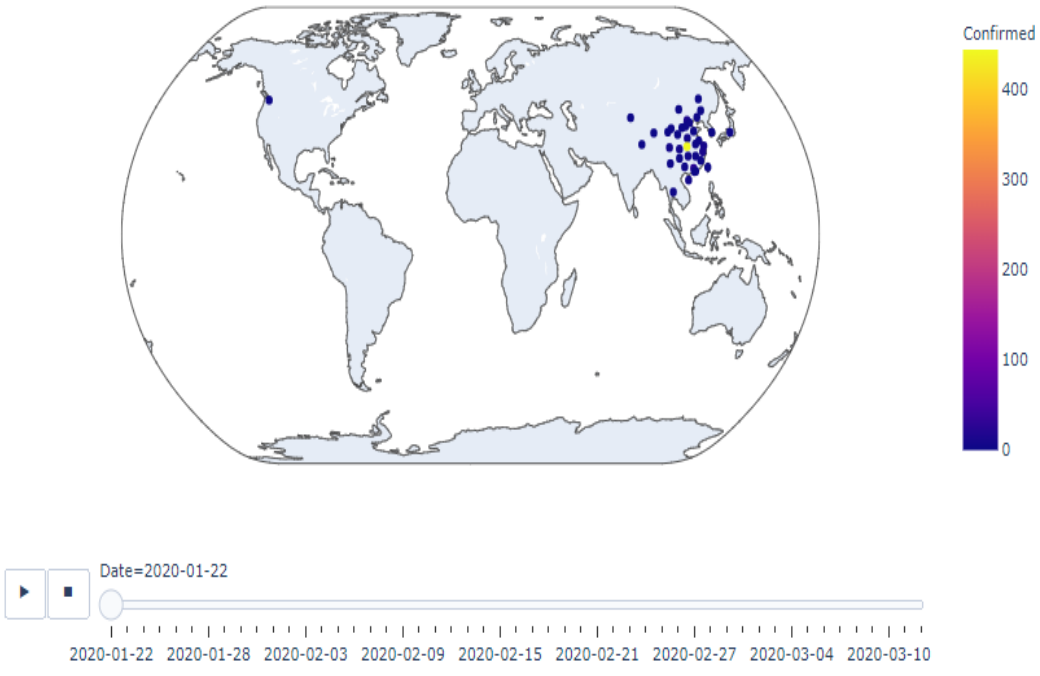


- Visualized the confirmed, recovered, and death cases only in Hubei

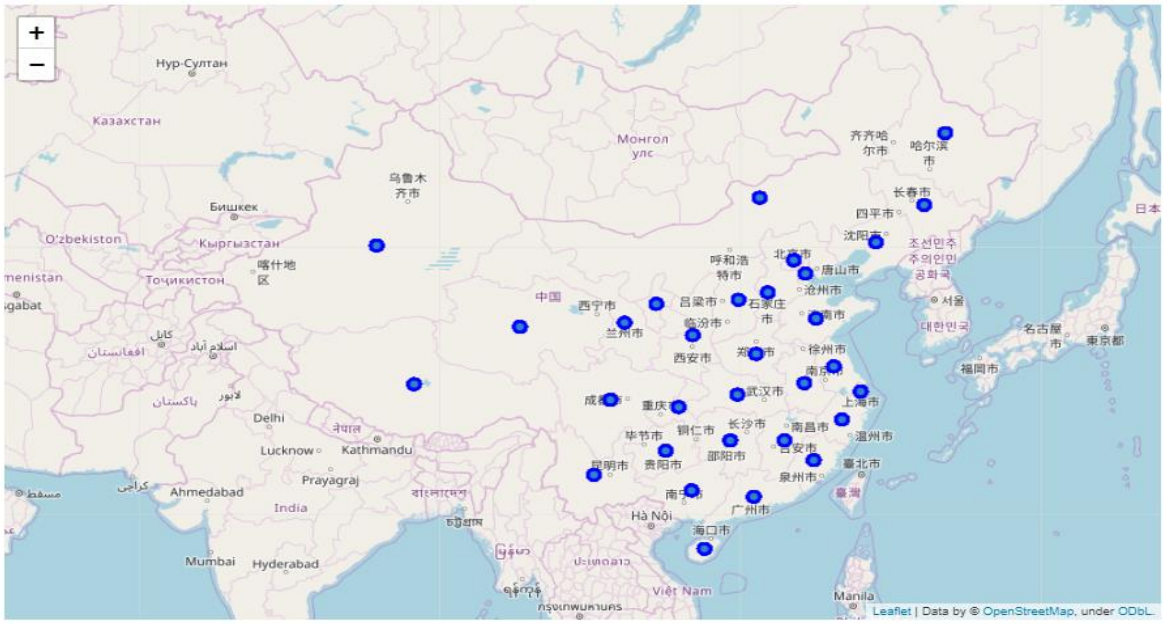


III. Geospatial Data Visualization

Visualization of the outbreak of the virus with time slider using Folium



ut[36]:



IV. Linear Regression Model

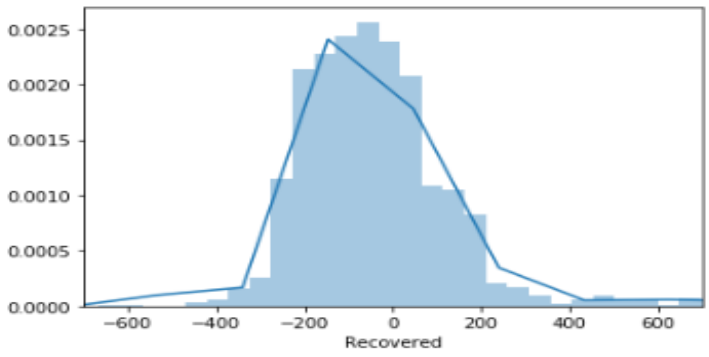
I have build a regression model to predict the number of recovered cases. I divided the data into two groups: training set which consists of 3451 instances and testing set which consists of 1701 instances.

The following figure shows MSE and MAE.

In [49]:
```
print('MAE:', mean_absolute_error(test_pred, y_test))
print('MSE:',mean_squared_error(test_pred, y_test))
```

```
MAE: 247.2532891323024
MSE: 693826.2898917153
```

In [50]:
```
fig, ax = plt.subplots()
sns.distplot((y_test- test_pred), ax=ax, bins =500)
ax.set_xlim(-700,700)
plt.show()
```



Results and Discussion

After analyzing cov-2019, the data shows the increase of the confirmed cases as well as the deaths. The number of the recovered cases is greater than the deaths. Regarding China data, the outbreak started from Hubei province which has the highest number of cases. The spread of the virus is fast and it keeps on increasing in the next period.

Conclusion

In this project, analyze the novel cov-2019 . preformed exploratory data analysis on the world data, especially China data where the outbreak started. Used visualization libraries such as: Plotly, Matplotlib and seaborn. Visualized the geospatial data worldwide as well as China. Build Linear Regression model to forecast the recovered cases.

References

Gabriel Preda. 2019/2020. coronavirus-2019ncov. *Kaggle.* [Online] 2019/2020. https://www.kaggle.com/gpreda.

WHO. 2019. Coronavirus. *World Health Organization.* [Online] 2019. https://www.who.int/health-topics/coronavirus.