Q2 Report

This question requires us to predict the daily covid cases and deaths in the US over a given time period. To do this I have calculated the difference in total covid cases and deaths between each day.

I have used referenced from Aggarwal et al. [1] “Forecast and prediction of COVID-19 using machine learning” and Al-Anzi et al. [2] “Forecast and prediction of COVID-19 using machine learning”. In both approaches an AutoRegressive Integrated Moving Average model (ARIMA) is used.

ARIMA is a time-series forecasting approach where past values are used to predict the future. However typically time series data sets tend to be non stationary or have seasonality. Non stationary means that the mean and standard deviation varies through the dataset.

To account for this, the ARIMA model has 3 parameters , p ,d and q. P is the amount of lags that are used to predict the current value. D is the order of differencing required to bring the time series to stationary state by removing non-stationarity and seasonality. Q represents the order of the moving average of the model. This is used to smooth out any random effects.

To identify the range of values to be used for p,d and q I have taken reference from [1] and [2]. In [1] the values tested are (p,d,q) = (5,1,0),(3,1,0) and (1,1,0) respectively. In [2] the auto_arima model is used to automatically obtain a model and the corresponding p,d,q values. The values are found to be 1,2,3.

Nevertheless, while keeping these references in mind, I used the partial autocorrelation and autocorrelation graphs to obtain ranges of values for my model. This work can be seen in the Jupyter notebook.

For the prediction of daily cases, the best RMSE value that I could reach was 70190 using p = 1, d = 1 and q = 1. For the prediction of daily deaths, the best RMSE value that I could reach was 879. These results are not very promising however it is in the same magnitude of the real values.

In order to try as an extension for the current algorithm, I try to combine the ARIMA algorithm with a Gradient Boosting Regressor. A gradient boosting regressor works by sequentially adding predictors to an ensemble, each one correcting the residual error of the previous estimator. Essentially it tries to fit the residual at each step

To incorporate this, I used the prediction from the ARIMA model for the validation set, then divided this predicted by 2. Fitted the first half of the prediction with the first half of the ground truth validation set to train by GBRT. Then I used the GBRT along with the latter half of the prediction to predict for the latter half of the ground truth validation set.

This allows me to try and reduce the residual error between the prediction and the ground truth values. This trained GBRT would then be used to adjust the ARIMA prediction from 20th Nov to 6th Dec. Nevertheless, it didn’t show much improvement as the rmse for the prediction of deaths only decreased to 837.

My Values for daily cases for 30th Nov to 6th December:

```
106465., 106797., 107127., 107454., 107777., 108099., 108423
```

Values from John Hopkins database

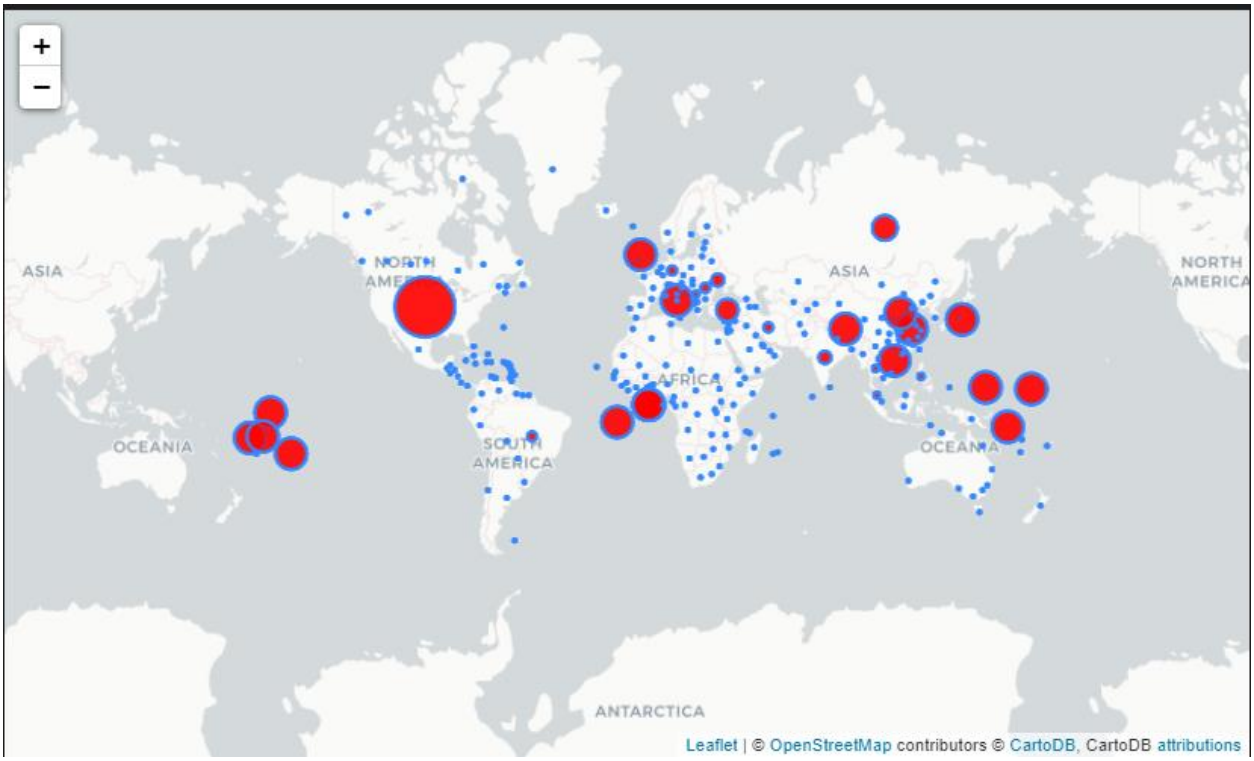
109294,139357,138718,153245,34572,197449

My values for death cases for 30th Nov to 6th December:

```
1489., 1490., 1491., 1492., 1493., 1494., 1495
```

1555,1966,1350,2149,490,161,1351

Task 2:



[1] "Forecast and prediction of COVID-19 using machine learning", Aggarwal et al. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8138040/pdf/main.pdf

[2] "On the accuracy of ARIMA based prediction of COVID-19 spread", Al-Anzi et al.

[3] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

[4] pmdarima documentation

[5] folium documentation