Data Analysis based on Situational Awareness

Ashwin Pathak (apathak60)
Megha Sankhlecha(msankhlecha3)

Definition

- Situational awareness is being aware of what is happening around you in terms of where you are, where you are supposed to be, and whether anyone or anything around you is a threat to your health and safety.

Motivation

- We want to extend the personal idea of situational awareness with real-time corroboration of data of events.
- We want to put forth a systematic analysis of real-time events to correlate and stitch the previously formed notion of awareness and how it affects the current situations.

Goals

- Disaster management and pandemic data analysis for situational awareness with focus on preparedness
- We want to majorly focus on the community wide awareness of the disasters.

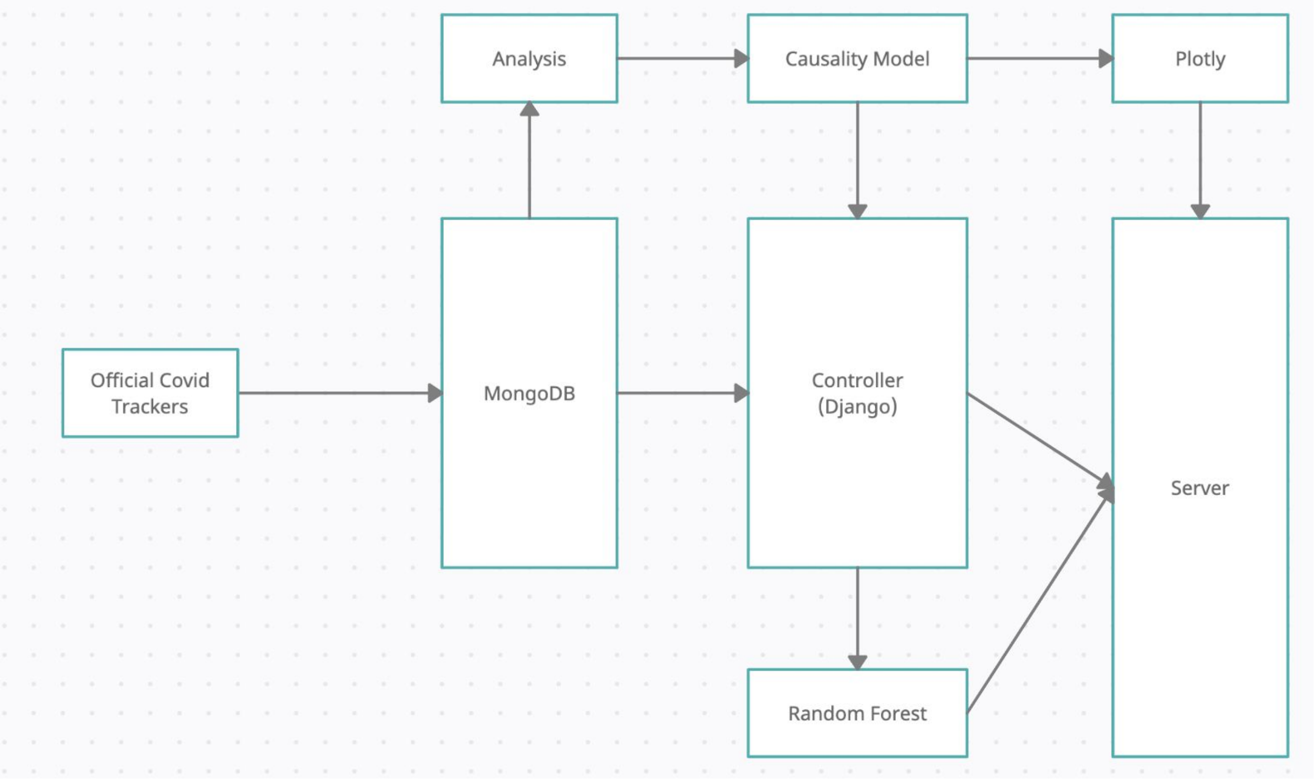
Work Plan

- We divide our work into following modules :
    - Data Collection :
        - Collect data from publicly available COVID dashboards provided by John Hopkins University, World Health Organization etc.
    - Data Analysis :
        - Analyse the data and draw meaningful inferences on our objectives.

Proposed Work And Deliverables

- We divide our work into following modules :
    - Visualization :
        - Present the data in graphs for understanding underlying patterns in the data.
    - Web Application :
        - Based on the inputs, a MVC based application to render the collected information intuitively.

Design Architecture



Work Plan

Phase Distribution

- Phase 1 : 9/25 - 10/21
  - Requirement analysis
  - Data collection
- Phase 2 : 10/22 - 11/20
  - Data Analysis
  - Drawing inference and conclusions
  - Visualization
- Phase 3 : 11/21 - 12/2
  - Web Application

Methodology

- **Automated Data Population**
  - Built an automated bot that will fetch data incrementally from covid dashboards
- **Data Analysis**
  - We have divided our analysis into time and space
  - We have further categorized our analysis into the following:
    - Countries
    - States
    - Cities
  - This kind of categorization helps us in visualizing the covid-19 cases in the space-time perspective.

- **Stationarity in time-series analysis**
  - Statistical properties of a process generating a time series do not change over time.
  - It does not mean that the series does not change over time, just that the *way* it changes does not itself change over time.
  - Stationary processes are easier to analyze.
  - Helps in ensuring ubiquity in time series analysis, making the ability to understand, detect and model it.
- **Null hypothesis**
  - There is no relationship between two population parameters, i.e., an independent variable and a dependent variable.
  - Sample observations result purely from chance.

- **Alternative Hypothesis**
    - Denoted by H1 or Ha
    - Sample observations are influenced by some non-random cause.
- **Augmented Dickey - Fuller Test**
    - Statistical test used to test whether a given Time series is stationary or not.
    - Most commonly used statistical tests when it comes to analyzing the stationary of a series.
    - The p-value obtained should be less than the significance level (say 0.05) in order to reject the null hypothesis
- **KPSS Test**
    - Checks for stationarity of a series around a deterministic trend.
    - Used to analyse the stationarity of a series.

- **Difference Method**
  - We observed by using the ADF Test and KPSS Test that our series is not stationary.
  - The p-values are all well above the 0.05 alpha level, so we cannot reject the null hypothesis.
  - Therefore, we transform the time series to be stationary by difference method.
  - Running the ADF Test and KPSS Test again gives us the stationary series.
- **Vector Autoregression**
  - Generalisation of the univariate autoregressive model for forecasting a vector of time series
  - It comprises one equation per variable in the system. The right hand side of each equation includes a constant and lags of all of the variables in the system.
  - The VAR class assumes that the passed time series are stationary.
  - We choose the lag order as 15 as that is most relevant to the covid-19 spread cycle.

- **Residuals**
  - A residual is the vertical distance between a data point and the regression line. Each data point has one residual. They are:
    - Positive if they are above the regression line,
    - Negative if they are below the regression line,
    - Zero if the regression line actually passes through the point

- **Durbin Watson Test**
  - Measure of autocorrelation (also called serial correlation) in residuals from regression analysis.
  - Autocorrelation is the similarity of a time series over successive time intervals.
  - A value of 2.0 means that there is no autocorrelation detected in the residuals.

Granger Causality Test

- Granger causality is a statistical concept of causality that is based on prediction.
- According to Granger causality, if a signal X_1 "Granger-causes" (or "G-causes") a signal X_2, then past values of X_1 should contain information that helps predict X_2 above and beyond the information contained in past values of X_2 alone.

Prediction using regression models

- We used Machine Learning regression models to predict the cases and deaths based on the observed data among countries.
- We tried the following regression models
  - SVM
  - Linear Regression
  - Logistic Regression
  - Random Forest
  - Decision Trees

Web Application

- We created a MVC based framework to dynamically fetch the updated data and ingest it in the dashboard we have created.
- We have used Django for our MVC framework.
- The modularity is defined as follows :
    - We use MongoDB as our database.
    - We use python for our controller.
    - We use HTML, CSS, JS for our view.

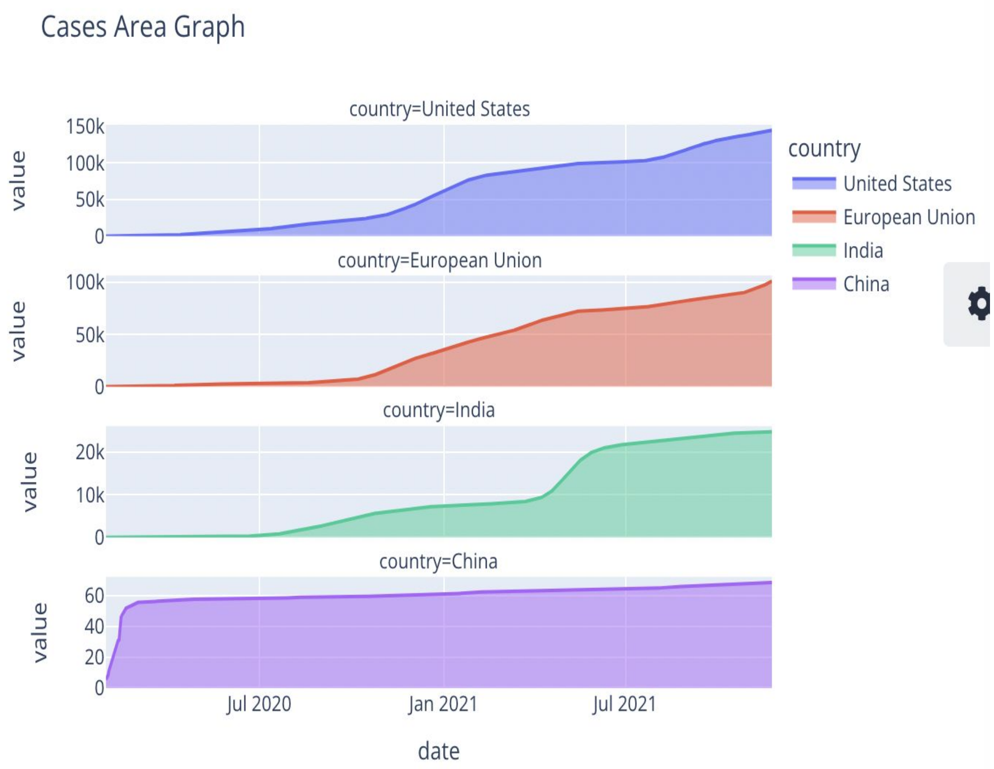
Deliverables

- Web Applications for projection of insights
- Automated Data Gathering Bot with incremental fetching optimization
- Mathematical Model for situational awareness for pandemic data

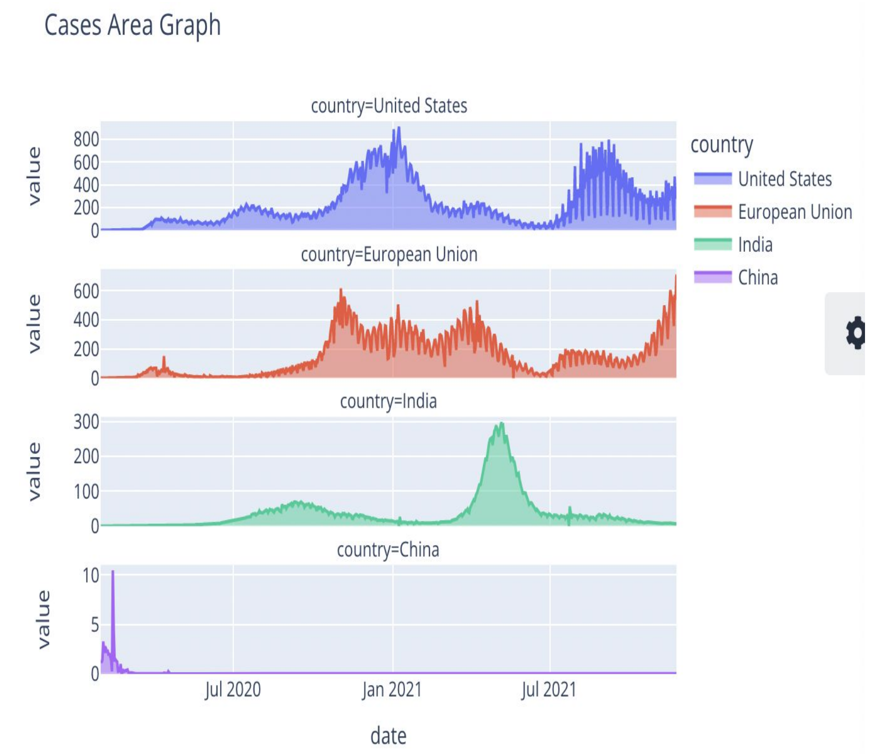
Results

- We take into account the following countries
    - India
    - China
    - European Union
    - United States

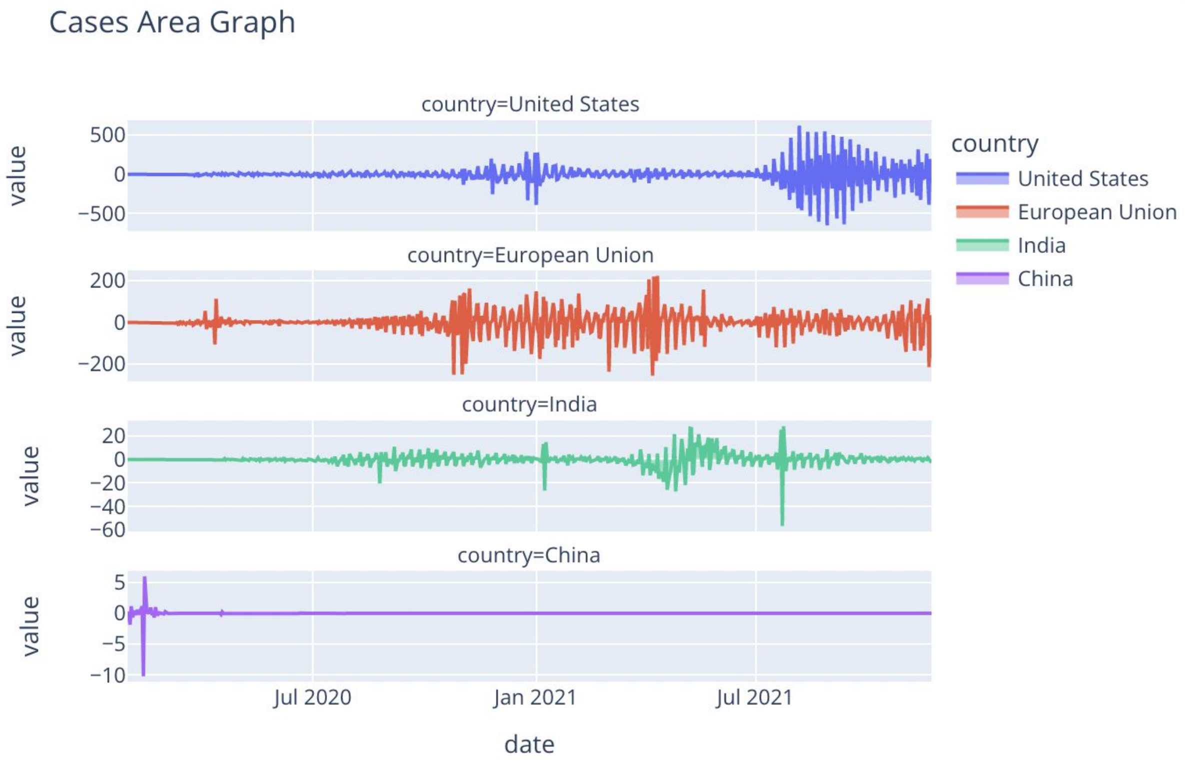
Total COVID-19 cases



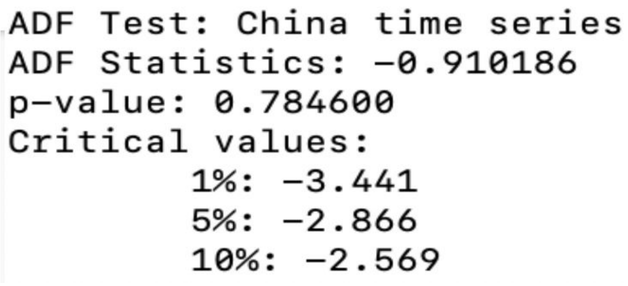
COVID-19 cases everyday



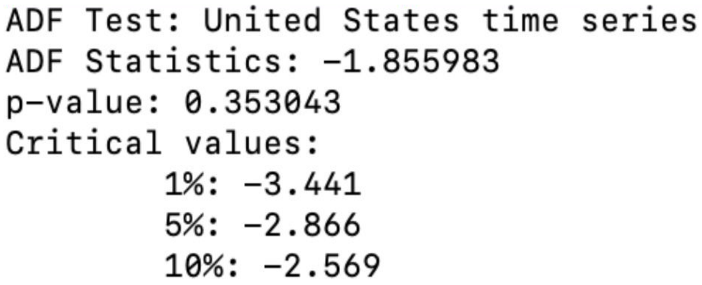
Cases after Difference Method Application



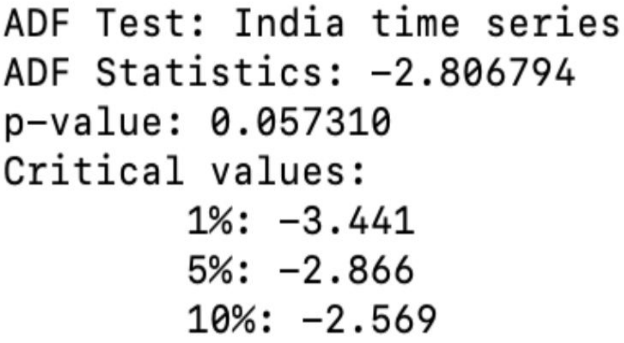
ADF Test Output



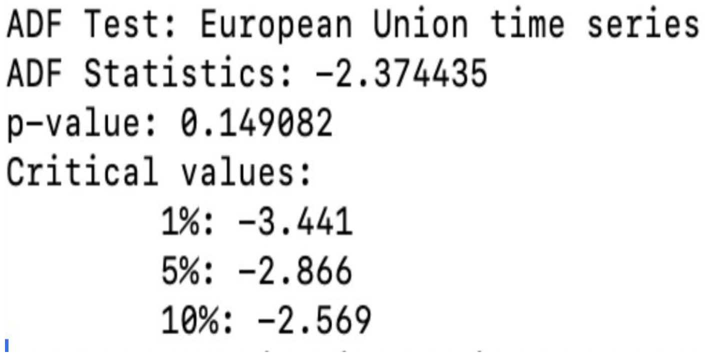
```
ADF Test: China time series
ADF Statistics: -0.910186
p-value: 0.784600
Critical values:
        1%: -3.441
        5%: -2.866
        10%: -2.569
```



```
ADF Test: United States time series
ADF Statistics: -1.855983
p-value: 0.353043
Critical values:
        1%: -3.441
        5%: -2.866
        10%: -2.569
```

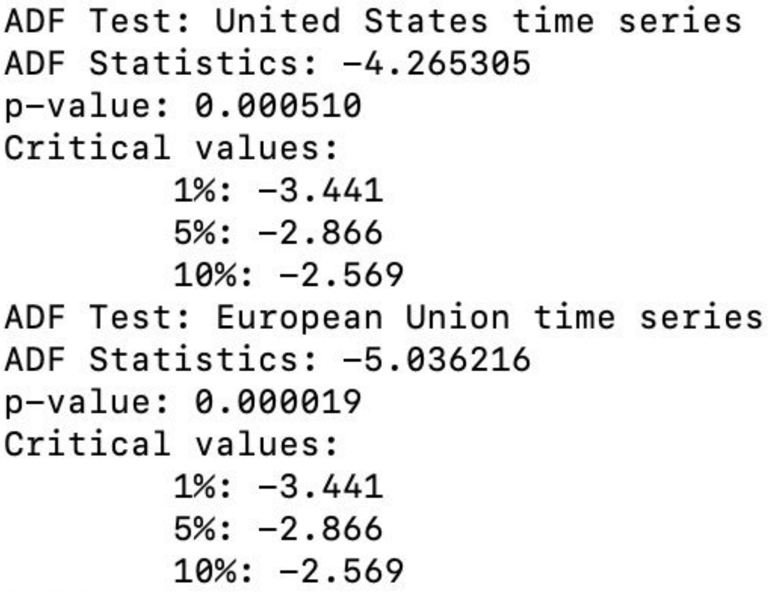


```
ADF Test: India time series
ADF Statistics: -2.806794
p-value: 0.057310
Critical values:
        1%: -3.441
        5%: -2.866
        10%: -2.569
```

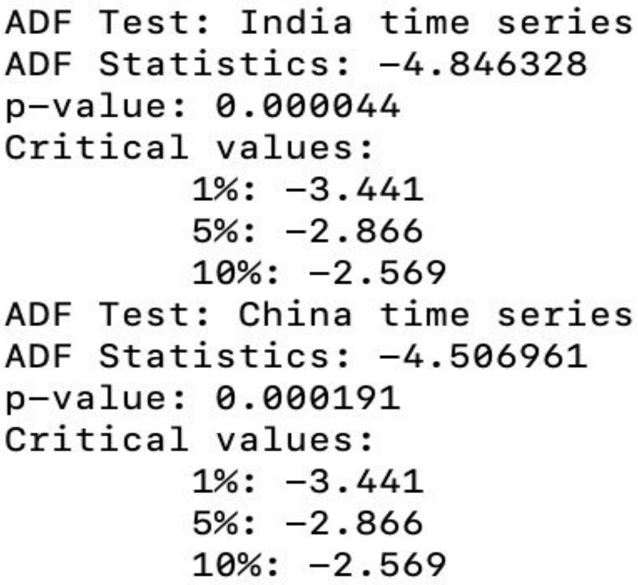


```
ADF Test: European Union time series
ADF Statistics: -2.374435
p-value: 0.149082
Critical values:
        1%: -3.441
        5%: -2.866
        10%: -2.569
```

ADF Test on transformed data

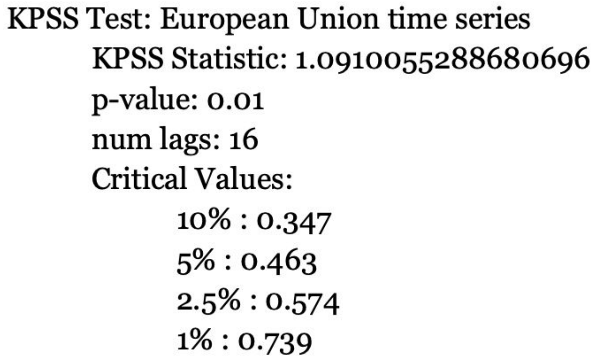


```
ADF Test: United States time series
ADF Statistics: -4.265305
p-value: 0.000510
Critical values:
        1%: -3.441
        5%: -2.866
        10%: -2.569
ADF Test: European Union time series
ADF Statistics: -5.036216
p-value: 0.000019
Critical values:
        1%: -3.441
        5%: -2.866
        10%: -2.569
```

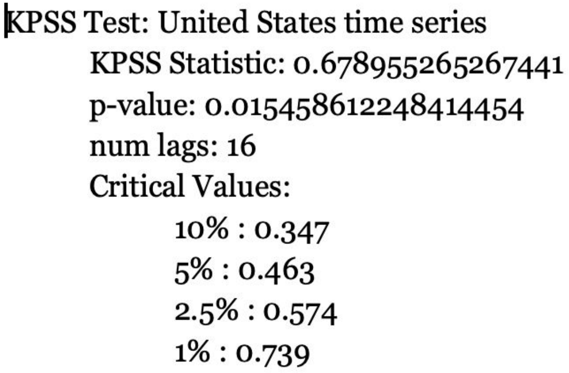


```
ADF Test: India time series
ADF Statistics: -4.846328
p-value: 0.000044
Critical values:
        1%: -3.441
        5%: -2.866
        10%: -2.569
ADF Test: China time series
ADF Statistics: -4.506961
p-value: 0.000191
Critical values:
        1%: -3.441
        5%: -2.866
        10%: -2.569
```

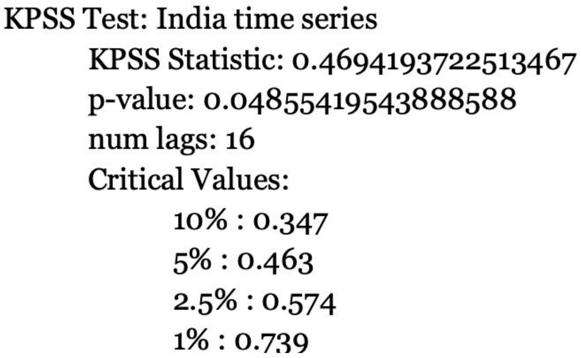
KPSS Test Result



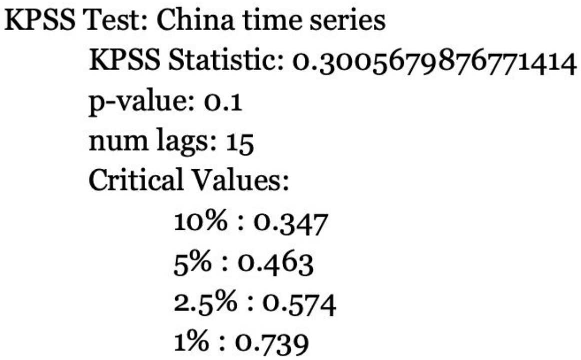
```
KPSS Test: European Union time series
        KPSS Statistic: 1.0910055288680696
        p-value: 0.01
        num lags: 16
        Critical Values:
                10% : 0.347
                5% : 0.463
                2.5% : 0.574
                1% : 0.739
```



```
KPSS Test: United States time series
        KPSS Statistic: 0.678955265267441
        p-value: 0.015458612248414454
        num lags: 16
        Critical Values:
                10% : 0.347
                5% : 0.463
                2.5% : 0.574
                1% : 0.739
```

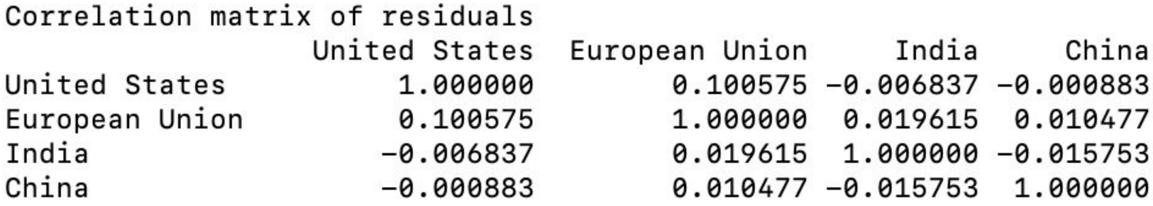


```
KPSS Test: India time series
        KPSS Statistic: 0.4694193722513467
        p-value: 0.04855419543888588
        num lags: 16
        Critical Values:
                10% : 0.347
                5% : 0.463
                2.5% : 0.574
                1% : 0.739
```



```
KPSS Test: China time series
        KPSS Statistic: 0.3005679876771414
        p-value: 0.1
        num lags: 15
        Critical Values:
                10% : 0.347
                5% : 0.463
                2.5% : 0.574
                1% : 0.739
```

VAR Model Prediction

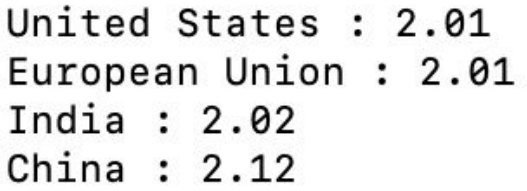


```
Correlation matrix of residuals
                United States  European Union     India     China
United States        1.000000        0.100575 -0.006837 -0.000883
European Union       0.100575        1.000000  0.019615  0.010477
India               -0.006837        0.019615  1.000000 -0.015753
China               -0.000883        0.010477 -0.015753  1.000000
```

biggest correlation is between the European Union and United States

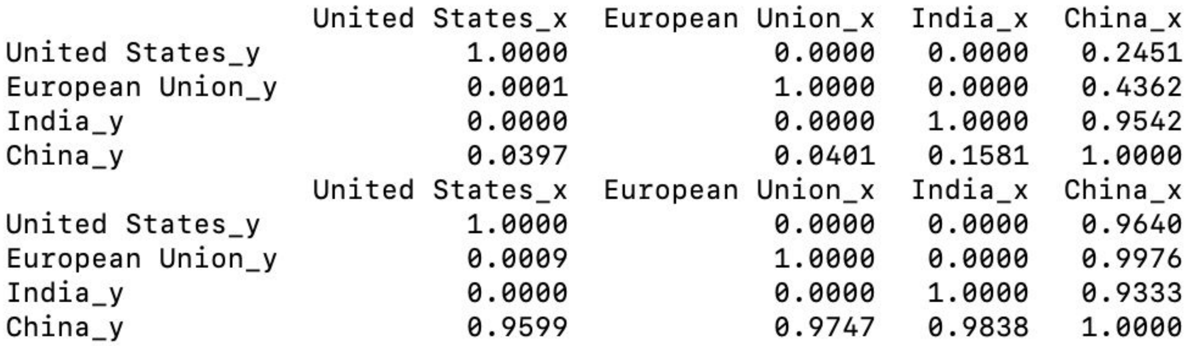
Durbin Watson statistic

This means that there is no autocorrelation in the residuals.



```
United States : 2.01
European Union : 2.01
India : 2.02
China : 2.12
```

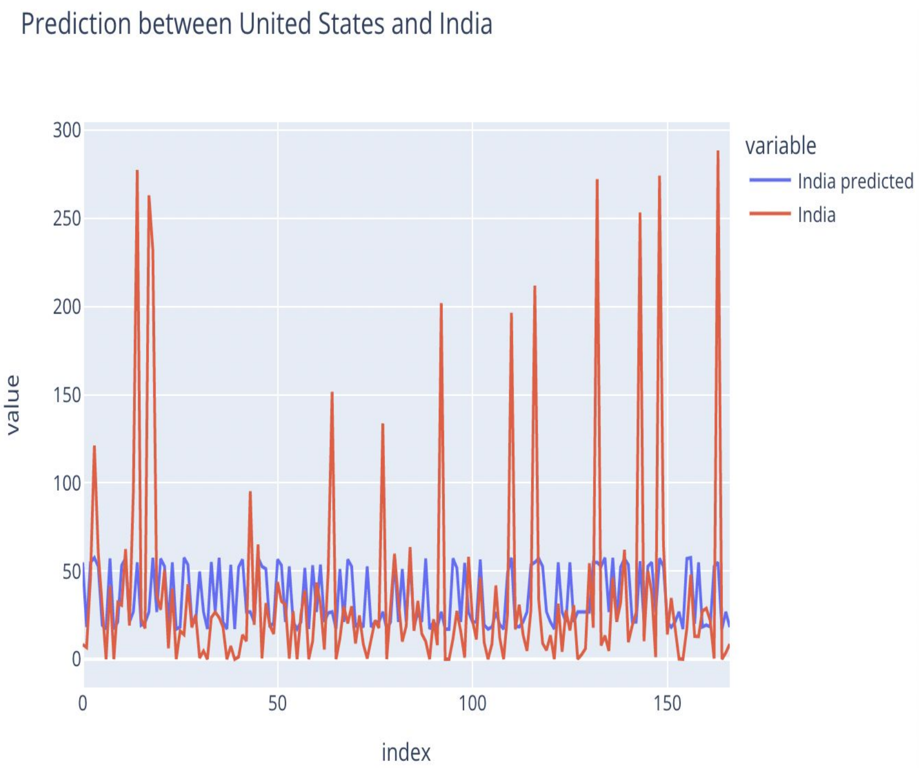
Granger Causality Results



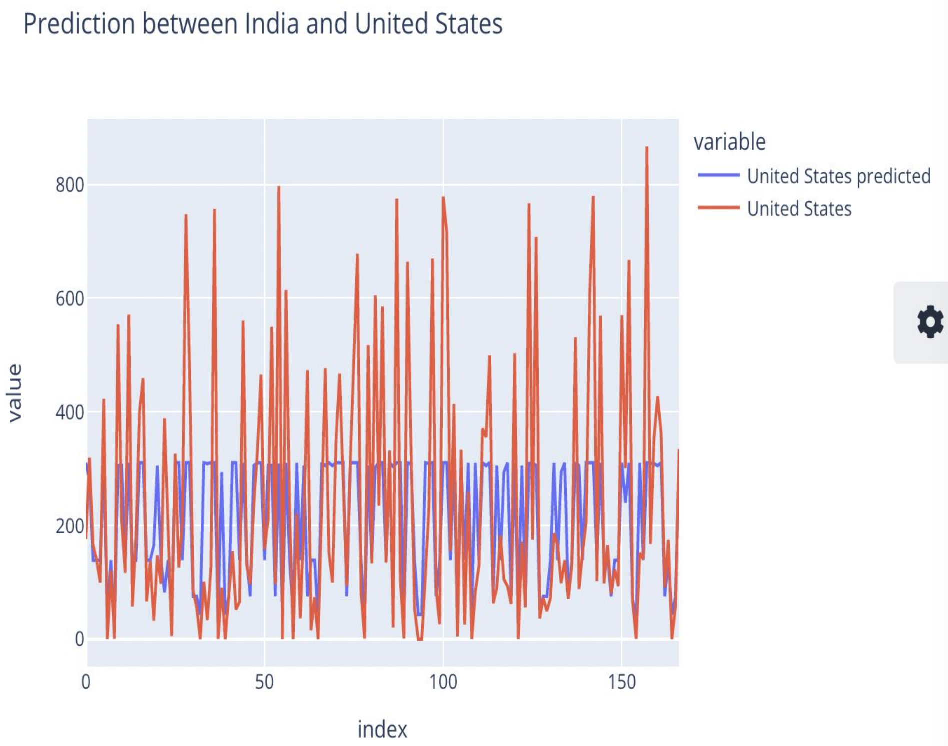
```
                  United States_x  European Union_x  India_x  China_x
United States_y            1.0000            0.0000   0.0000   0.2451
European Union_y           0.0001            1.0000   0.0000   0.4362
India_y                    0.0000            0.0000   1.0000   0.9542
China_y                    0.0397            0.0401   0.1581   1.0000
                  United States_x  European Union_x  India_x  China_x
United States_y            1.0000            0.0000   0.0000   0.9640
European Union_y           0.0009            1.0000   0.0000   0.9976
India_y                    0.0000            0.0000   1.0000   0.9333
China_y                    0.9599            0.9747   0.9838   1.0000
```

- From the second table above, it can be observed that China does not cause any Covid-19 cases. Additionally, the European Union causes India and the United States and vice-versa.
- We have generated similar results for deaths in countries and states too.

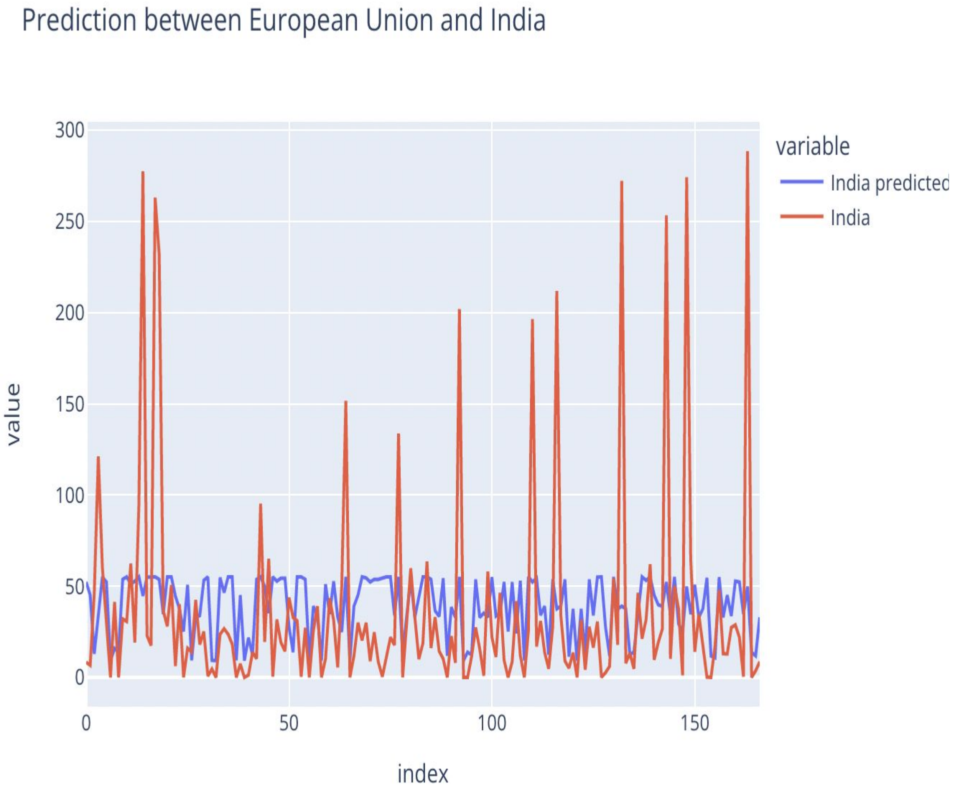
ML Regression Models - Random Forest among countries



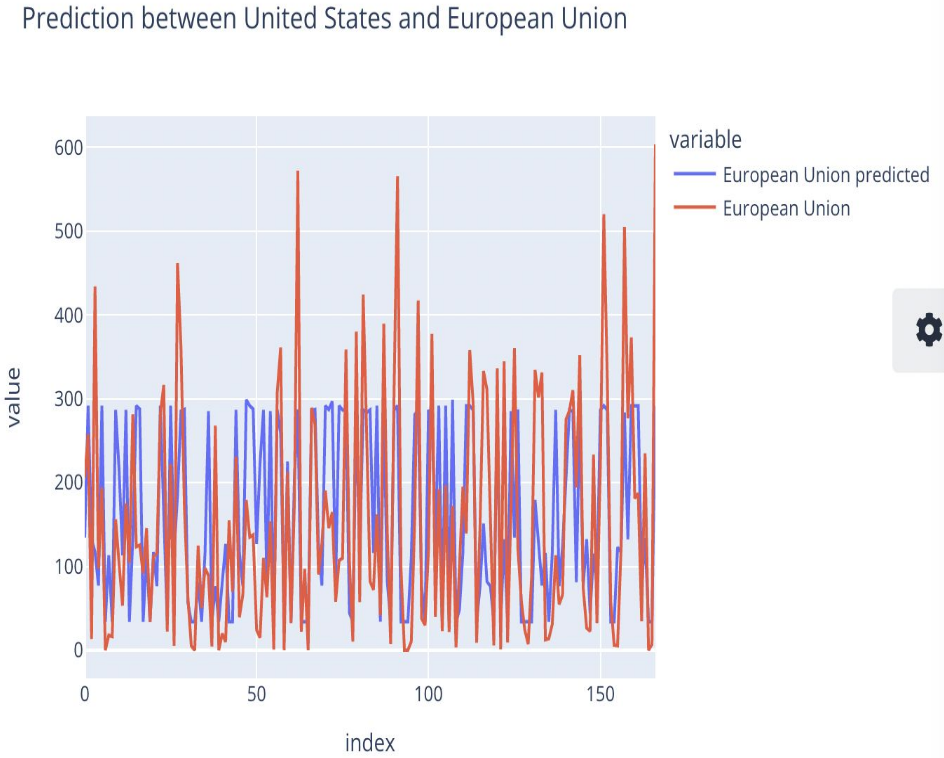
Prediction between United States and India



Prediction between India and United States



Prediction between European Union and India



Prediction between United States and European Union

Future Work

1. We have started the analysis on a subset of countries and states. This can be extended to a larger extent.
2. The analysis can be extended to a finer grain to a city level and intra-city level which can give a lot of insights on the situational spread of covid-19.
3. Due to the limited scope of the project, the analysis is done on deaths and cases. We can extend it to hospitalizations, gender, age-group, etc:

References

- https://github.com/nytimes/covid-19-data
- https://hgis.uw.edu/virus/
- https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
- https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4
- https://data.ct.gov/Health-and-Human-Services/COVID-19-Cases-and-Deaths-by-Age-Group/ypz6-8qyf
- https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3
- https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku
- https://www.google.com/covid19/mobility/
- https://github.com/CSSEGISandData/COVID-19
- https://covid19.who.int
- https://covid19.who.int/info/
- https://coronavirus.jhu.edu/map.html
- https://www.ecdc.europa.eu/en
- https://www.researchgate.net/publication/340516704_Analyzing_Situational_Awareness_through_Public_Opinion_to_Predict_Adoption_of_Social_Distancing_Amid_Pandemic_COVID-19
- https://www.liebertpub.com/doi/pdf/10.1089/hs.2020.0194
- https://www.sciencedirect.com/science/article/pii/S1871402120302332?casa_token=8hqPLHhUnlYAAAAA:HE-ncsn-XfLttw6Iyut29AxLeVE2m4MOOu-P6b01Y147n9gxn1aCbrKqGIyU1RUsg_34tJMyyAQ

Questions?