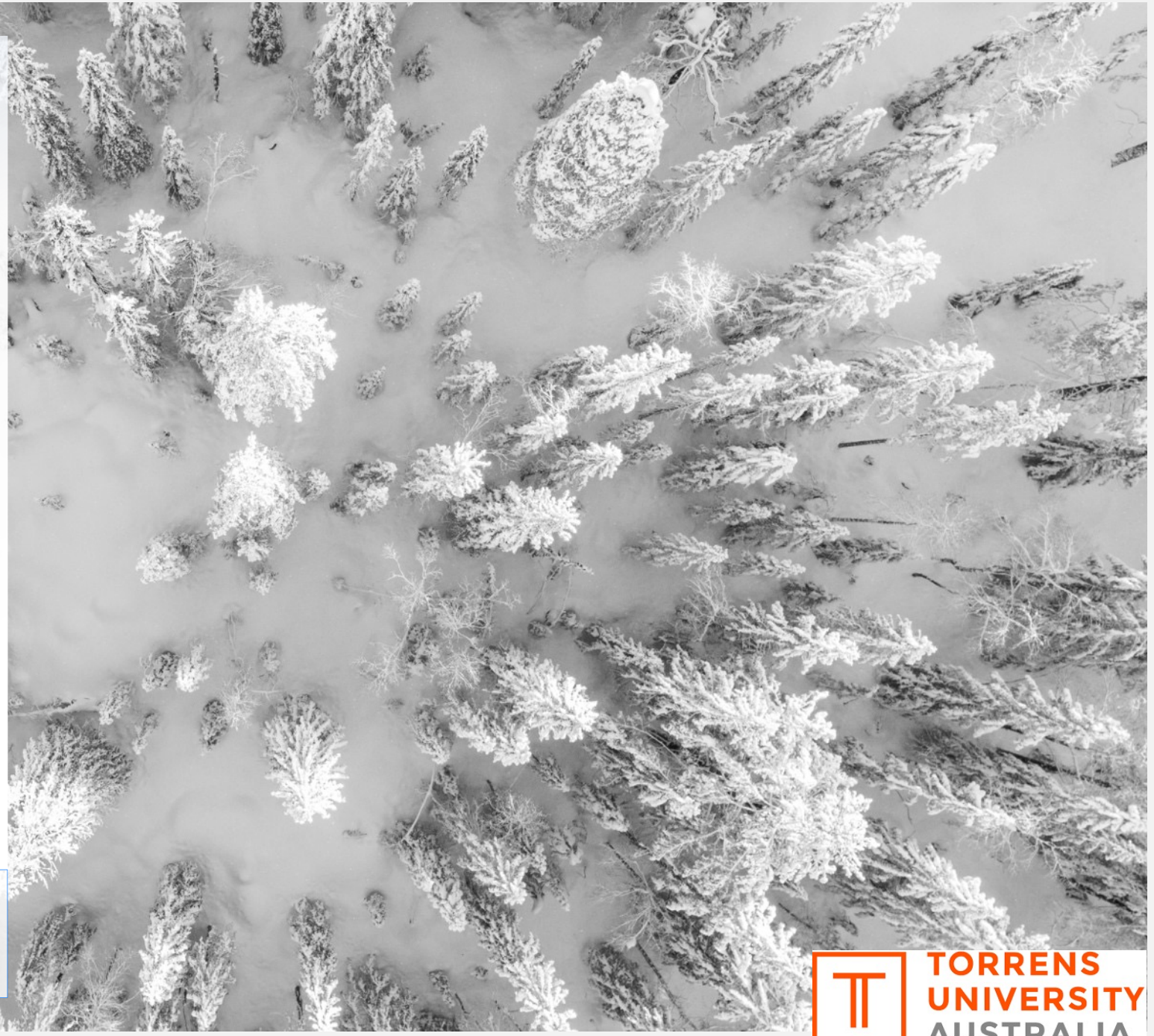
COVID 19 - BIG DATA ANALYSIS

Kunal Kanade - A00061993



TORRENS UNIVERSITY AUSTRALIA

Index

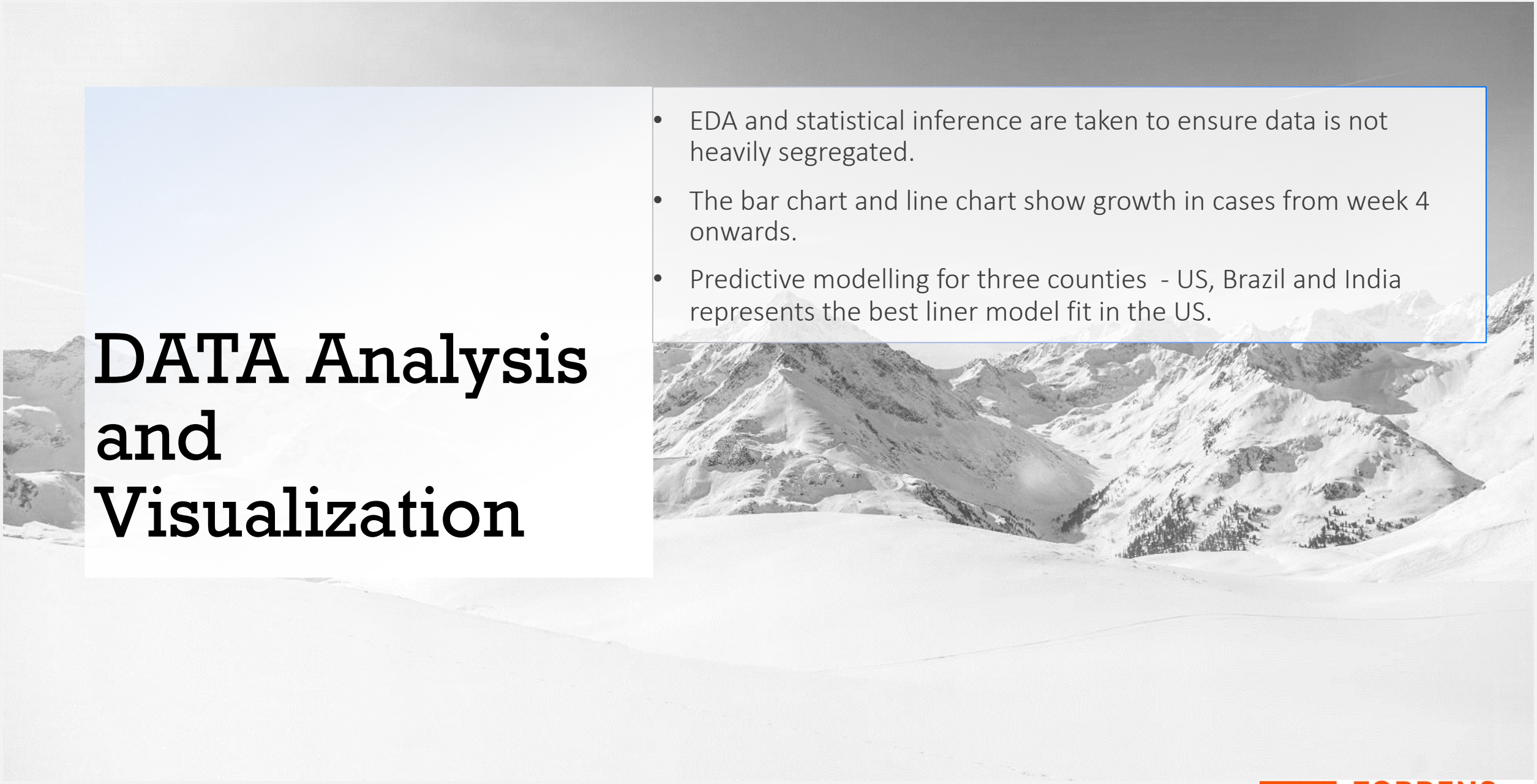
1. Data Preparation
2. Data Analysis Visualization
3. Insight and Recommendation

DATA Preparation

- The data set consists of 267 rows and 5 columns (Province/Country/Latitude/Longitude/Confirmed).
- Dates columns which are in columns need to be transposed.
- NULL values present in Province columns need to be removed or dropped.
- Use of panda and NumPy for data cleaning and transformation.
- Datetime function and week conversion needs to be done.

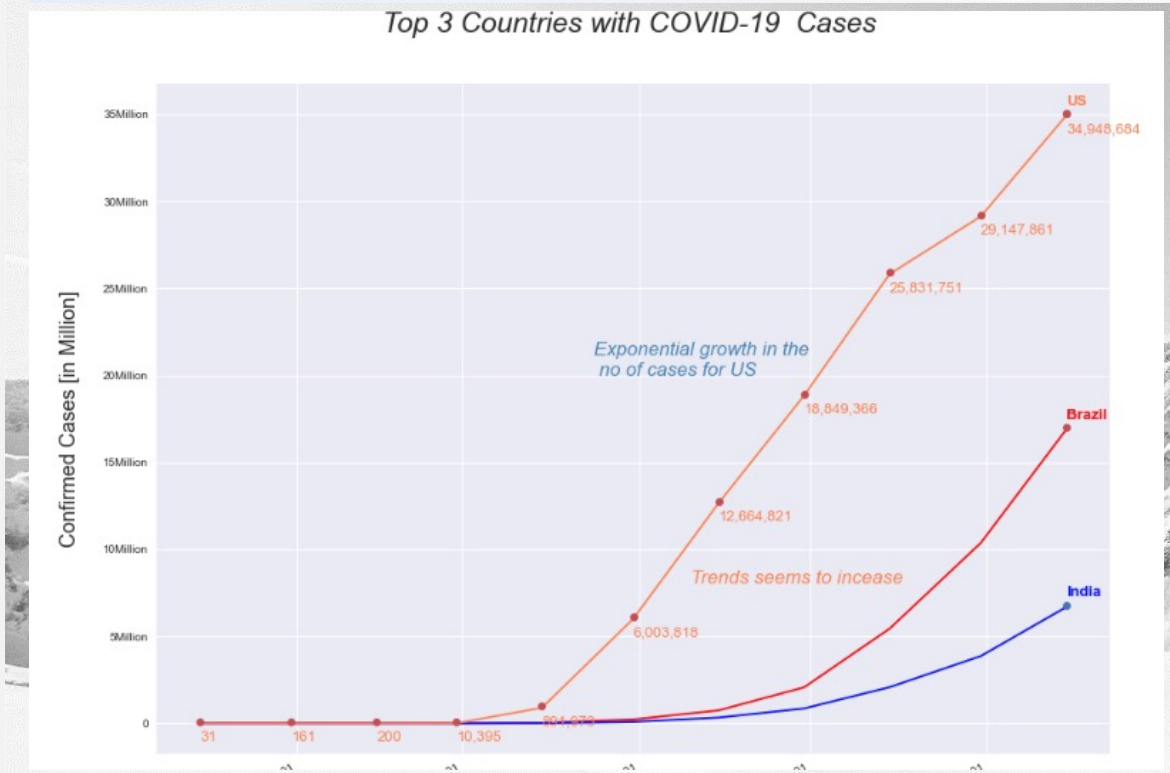
DATA Analysis and Visualization

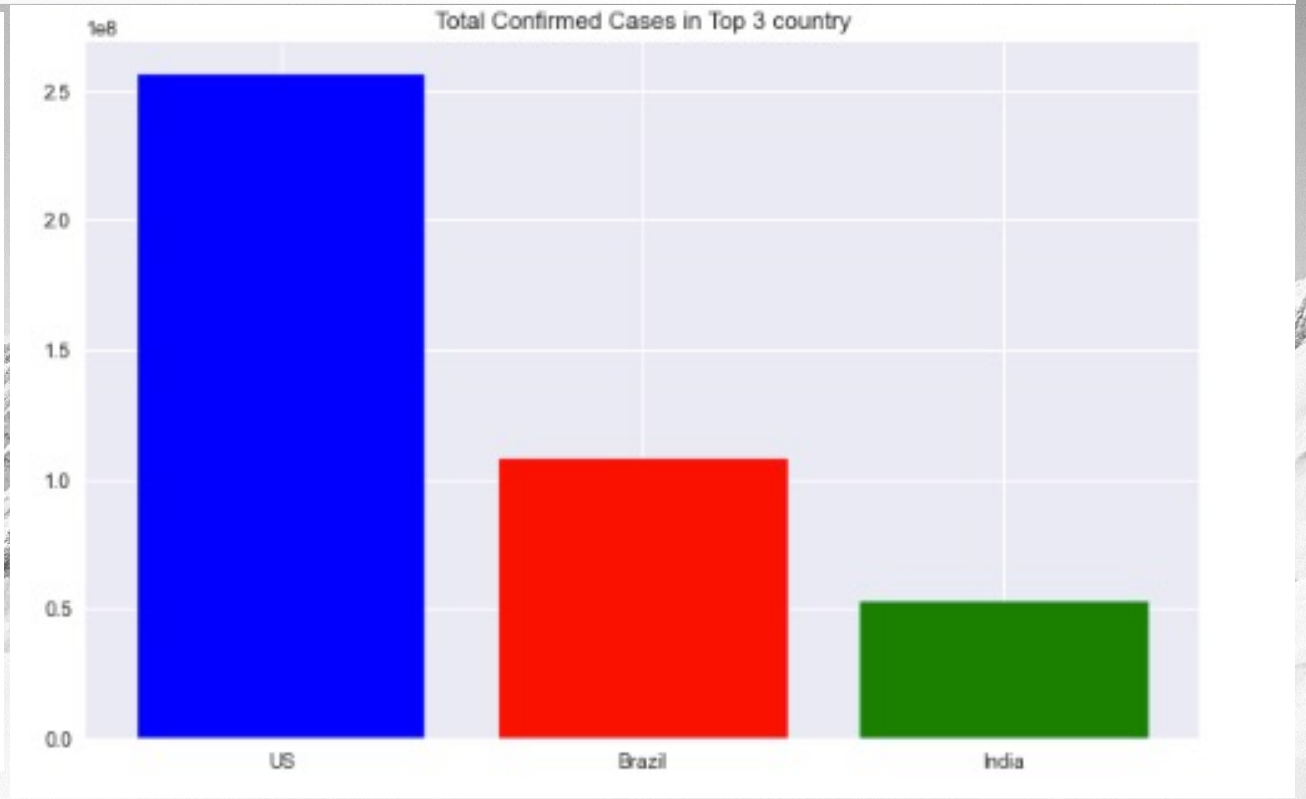
- EDA and statistical inference are taken to ensure data is not heavily segregated.
- The bar chart and line chart show growth in cases from week 4 onwards.
- Predictive modelling for three counties - US, Brazil and India represents the best liner model fit in the US.



TORRENS UNIVERSITY AUSTRALIA

Exploratory Data Analysis

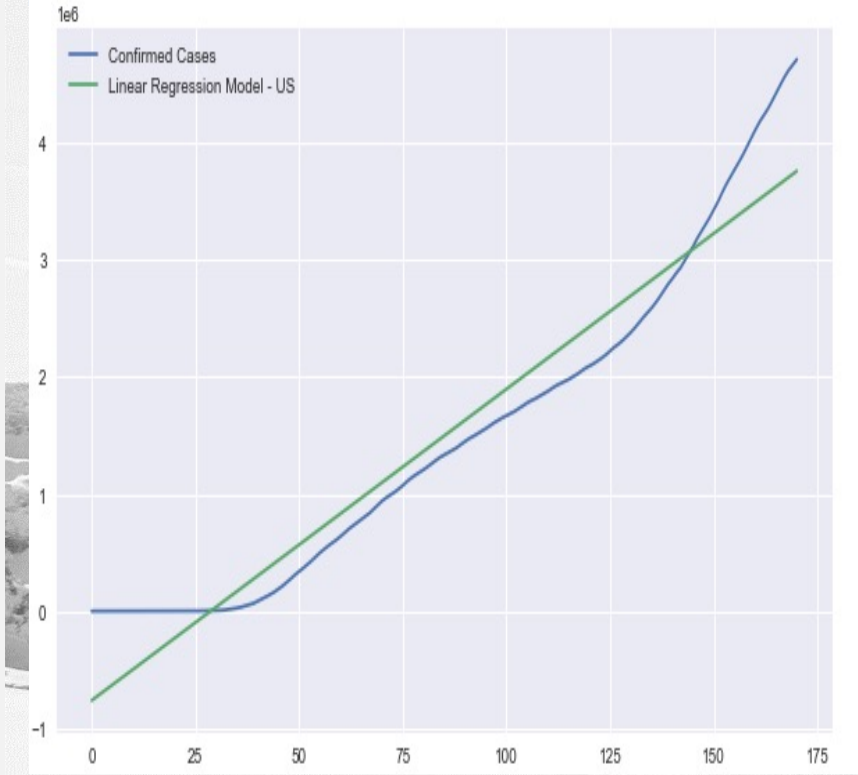


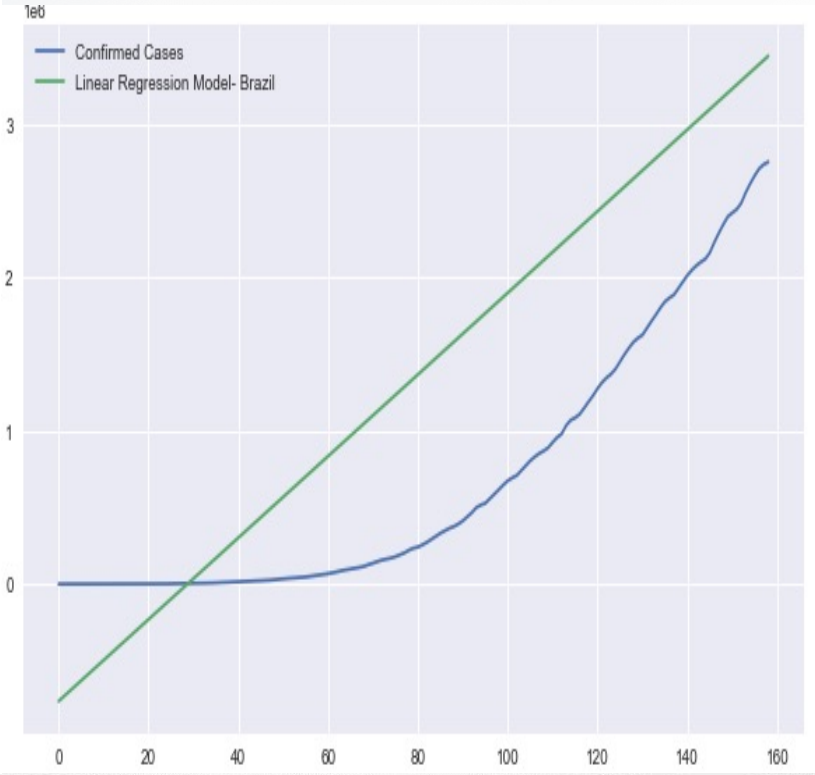


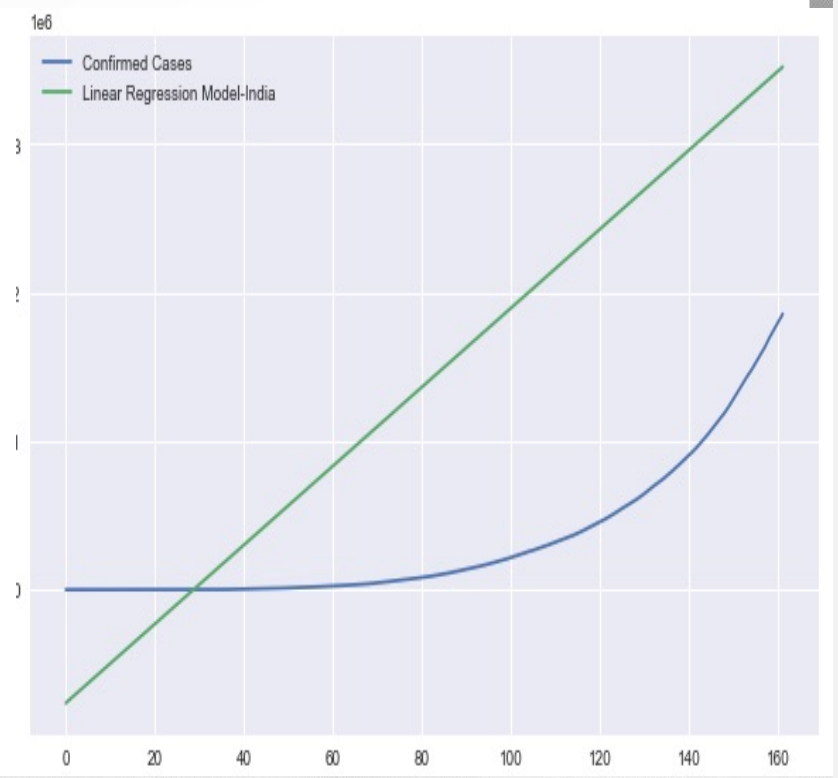
- The bar chart and line chart show growth in cases from week 10 onwards.

TORRENS UNIVERSITY AUSTRALIA

Liner Regression Model – Top 3 Countries



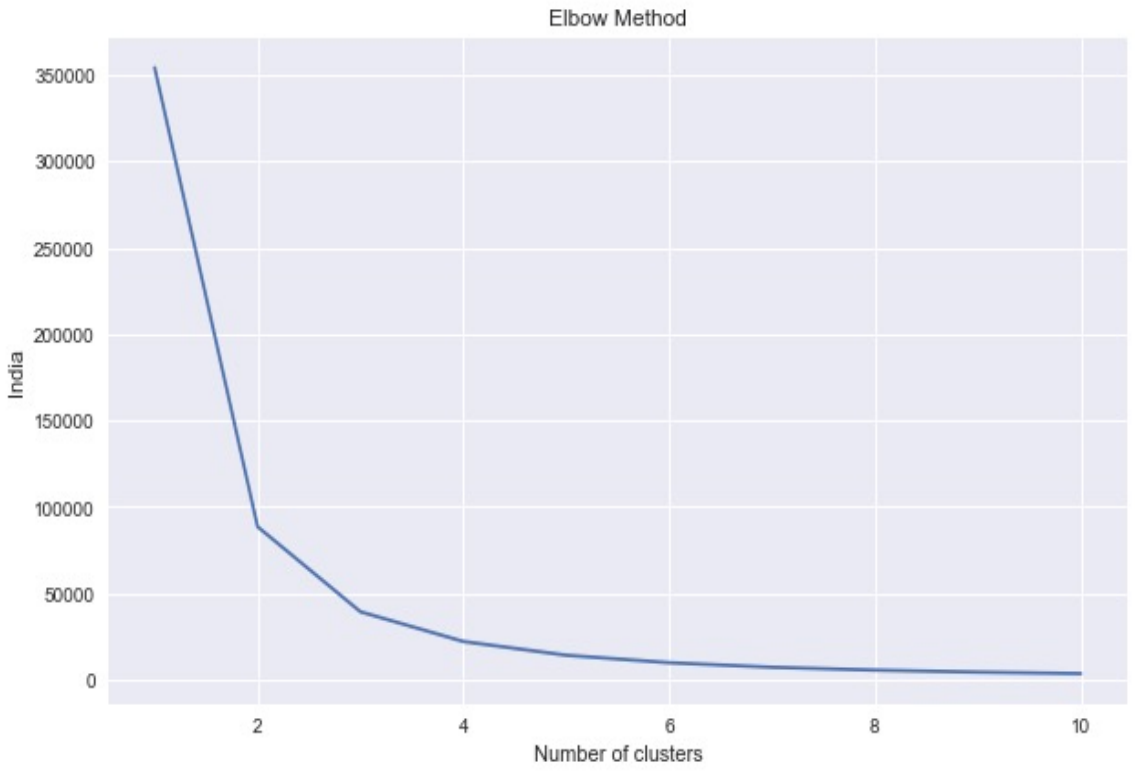


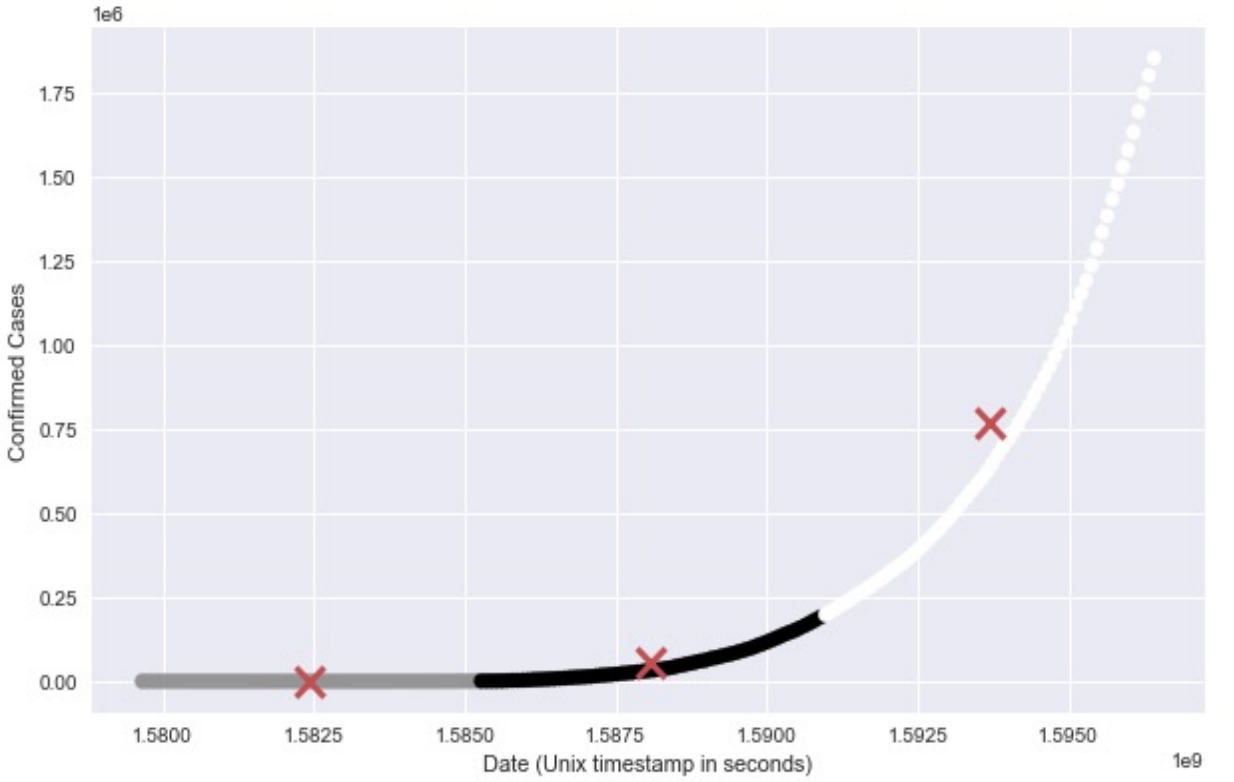


- The regression model suggests the best model for US and Brazil with above 82% accuracy, whereas the India prediction model suggests 69%.

TORRENS UNIVERSITY AUSTRALIA

K Mean Clusters - India





- As per Elbow Method, we have selected k=3 for clustering and we put this clustering on our India dataset.
- Another diagram represents the three clusters based on date and weekly timestamp, as week 1 to week 10 is ideal for 1st cluster with a number of cases, week 10 to week 19 is the second cluster with cases on the rise and lastly week 19 onwards peak started and has been put under the third cluster.

TORRENS UNIVERSITY AUSTRALIA

Insight and Recommendation

- Cases are on the rise for most of all countries and preventive measures should be taken to avoid them. As per our model cases rises from week 4 in many countries and the top three countries represent the same scenario.
- Although China can be seen in decline because of the preventive measure taken by them.
- As per clustering, we need more data to conclude which area has much influence on spreads.
- The border does have the influence of spread if the US has been quarantining measures, then the spread would have been avoided.
- Our data needs cases of recovered and vaccination taken to gauge more accurate results for our model.

Thank You

TORRENS UNIVERSITY AUSTRALIA