UNIVERSITY OF BONN

LAB DATA SCIENCE IN PRACTICE

Prediction of Covid-19 Spread

Author:
Prajakta Bhujbal

Supervisor:
Dr. Elena Demidova

MA-INF 4325 - Lab Data Science in Practice: Spatio-Temporal Data Analytics

Department of Computer Science

Chapter 1

Introduction

1.1 Introduction

The COVID-19 widespread caused by coronavirus, is an on-going pandemic. The current episode was caused by a strain of coronavirus that had not already been found anyplace within the world until it was discovered in December 2019. A total of 122,435,351 confirmed cases of the coronavirus (COVID-19) have been recorded from 219 countries and territories around the world, with 2,704,440 deaths.[10] This virus continues to influence people in every part of the world, but some countries are experiencing high levels of infection, although others have managed to control it to some degree. There are a lot of questions and unusual scenarios that come at the side this undiscovered infection.

However, we can use Data Science practices to address certain questions and assess the pandemic situation. Machine learning algorithms can provide effective models to predict the trend of infections. We'll look at Time Series models like ARIMA and VAR to see if these models can anticipate future patterns for some of the worst-affected countries. To avoid the spread of the pandemic, governments had to enforce a number of measures, including lockdowns, international travel bans, and online teaching at schools and universities. We will build a model to see how these preventive measures affect the virus transmission.

1.2 Research Questions

The aim of this project is to examine the following aspects of the pandemic and answer the following research questions using Time Series modelling approaches:

- **Predicting the growth and trend of COVID-19** pandemic in highly affected countries
- **Identifying how government policies encoded using stringency index have influenced the spread of coronavirus**
- Correctly predict and estimate the percentage of the total population that will be impacted in the future

This will encourage policymakers to change their policies accordingly and prepare ahead of time for preventive measures such as public health messaging, raising community awareness, and expanding the healthcare system's capacity.

1.3 Related Work / Literature

The theoretical work related to this study is discussed in this section. To study the patterns and predict the spread of infectious diseases, machine learning and statistical methods, of which time series forecasting is a subset, have been successfully applied in the past. My first step was to collect as much information as possible in order to get a good start on this project. I looked at all the aspects of Spatio-Temporal Analysis and came across some relevant research work. I found a variety of papers in which researchers had used Time Series modeling to forecast the spread of the current pandemic.

One study's results on a comparative study of time series methods to forecast percentage of active cases per population for Covid-19 were published in a research paper.[9] In [6], researchers have developed and compared the performance of time series models and applied it to 187 countries for an empirical comparison. They believed that a one-to-one relationship exists between a data pattern and the most accurate time series model. Study included tracked population outflow from Wuhan using nationwide cell phone data and connected it to Covid-19 infection counts by location.[3]

Using the time series models, a forecasting study is done for reported Covid-19 cases, deaths, and recoveries in the United States and Italy. They performed point forecasts and PIs for both countries seven days or one week ahead.[2] A comparison of three different forecasting models was conducted to estimate the Covid-19 case count in the next 10 days, in order to assist Asia Pacific countries in preventing this breakthrough by taking appropriate precautions ahead of time. The model includes RNN, LSTM, and GRU compromises deep learning layers to dynamically extract the features from previous knowledge and predict new patterns symmetrically.[7]

1.4 Datasets

The primary dataset used in this case study is the **Covid-19 Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2** by O. Wahltinez.

This repository attempts to assemble the largest Covid-19 epidemiological database in addition to a powerful set of expansive covariates. It includes open, publicly sourced, licensed data relating to demographics, economy, epidemiology, geography, health, hospitalizations, mobility, government response, weather, and more. Moreover, the data merges daily time-series, +20,000 global sources, at a fine spatial resolution, using a consistent set of region keys.

More details about the dataset are available at: https://goo.gle/covid-19-open-data

Some of the features ("Confirmed New", "Recovered New", "Deaths New") are integrated into above dataset from following source: **Covid-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.**

The dataset can be found at https://www.kaggle.com/gpreda/coronavirus 2019ncov

Chapter 2

Data Acquisition and Pre-processing

2.1 Data Acquisition

The datasets required for this study were acquired from two different sources, the first one is GCP dataset, collected at Google Cloud Platform and the other is JHU dataset provided by John Hopkins University. These two datasets were then integrated to make the final dataset. The original dataset is about 2GB in size and contains data for over 248 countries with 108 features. The data is updated every day, and the earliest records for each country is available from beginning, that is January 1, 2020.

2.2 Data Pre-Processing

After acquiring the dataset, I began with data pre-processing at the start of the first iteration. I followed a series of steps to obtain the necessary results, as this is a big dataset to work with. And I repeated those steps until the dataset was clean enough to proceed with EDA and Model Building. Since the real-world dataset includes values that are inaccurate, contradictory, or even incomplete, to remove any anomalies in the data, I implemented the following steps:

2.2.1 Read and Understand the Dataset

After reading through the dataset, I looked for basic details such as the dataset's shape and how the first few records of the dataset look like to get an overall idea of the dataframe.

2.2.2 Identifying and handling incorrect, inconsistent or unnecessary values

The names of the columns such as **"3166-1-alpha-2"** and **"3166-1-alpha-3"** does not make any sense. As a result, I looked at the values of those columns and decided to drop them. I also decided to drop unnecessary features such as **"noaa station", "noaa distance", "wikidata"** and **"aggregation level"**.

Other column values such as **"key", "datacommons" and "CD KEY"** were also examined. And I discovered that some of the columns, such as - **"country name" and "CD KEY"**, **"key", "country code" and "3166-1-alpha-2"**, **"datacommons" and "3166-1-alpha-3"** have similar values. Also I would not need this information or these

features. As a result, the columns **"date"** and **"country name"** were the only ones I retained.

2.2.3 Identifying and handling the missing values

- **If multiple features/columns are missing for any particular record/row. We will drop such a row.**
- **If multiple records/rows are missing for any particular column/feature. We will drop that column/feature.**

When this isn't possible in practice, and the column/record in question is critical for making predictions, I used other imputation methods.

I estimated the percentage of missing values for each feature. **There are 22 features that are missing more than 60 percent of the records.** And I chose to drop those columns. Then I proceeded with removing the records for countries which had more than three months of data missing. Just one country's HDI value was absent, and that was Ireland. **The HDI value for Ireland was determined to be 0.955, so I replaced NaN with 0.955.**

I double-checked that the **"Confirmed", "Recovered" and "Deaths"** columns has no relationships with any other relevant columns **("new confirmed", "new deceased", "total confirmed", "total deceased").** There are irregularities between the numbers. As a result, I made the decision to drop those incorrect valued columns.

Statistical Imputations: I used statistical imputations to fill the NaN values for remaining features.

- **Weather related features :** I used the country-wise mean for "average temperature","minimum temperature","maximum temperature","rainfall","dew point","relative humidity" features.
- **Govt. Policy related features :** I used the country-wise mean for "mobility retail and recreation","mobility grocery and pharmacy","mobility parks","mobility transit stations", "mobility workplaces","mobility residential" features.
- **Mobility related features:** I used the country-wise mean for features like "school closing","workplace closing","cancel public events", etc.

2.2.4 Data Integration with JHC(Johns Hopkins CSSE) dataset:

At this point, there are no more wrong or missing values in the dataset. However, we still lack information on the **"Recovered" and "Newly Recovered"** cases. These features were taken from the JHC (Johns Hopkins CSSE) dataset.

2.2.5 Data Summary and Datatype Checks

Now that the final, clean dataset with around 34111 rows and 69 columns was ready to use, I checked some basic details for each feature, such as index and column dtypes, non-null values, and memory usage, and also converted the date column's datatype to datetime64 datatype. I also looked up statistical information on each variable to get a basic understanding of how each numeric variable was distributed.

Chapter 3

Exploratory Data Analysis

3.1 Exploratory Data Analysis

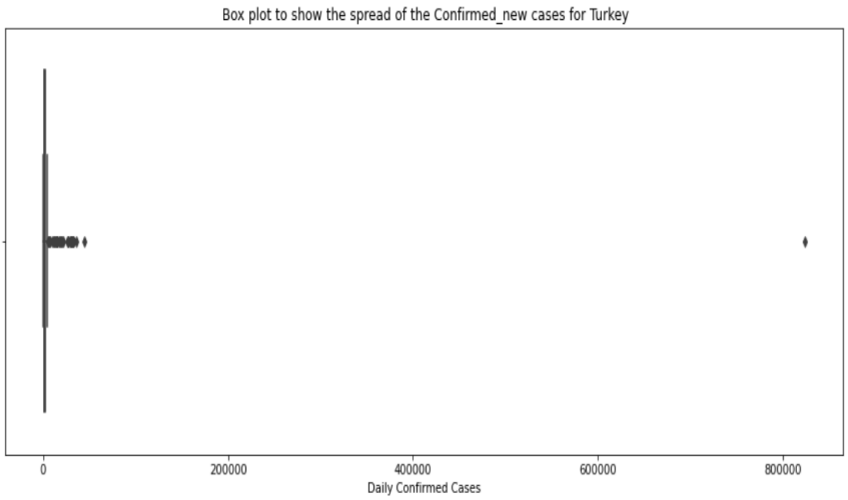
Exploratory Data Analysis involves analyzing data, finding trends in the data, and identifying distributions using visualization techniques to assist in the formulation of hypothesis or inference. Based on this hypothesis, we train our model and use it to make predictions.

While doing exploratory data analysis on this covid dataset, my goal was to gain a comprehensive understanding of each feature, which could then be used for more advanced data analysis or modeling. The methods implemented in the EDA helped in interpretation of things like standard deviations and distribution of feature values.

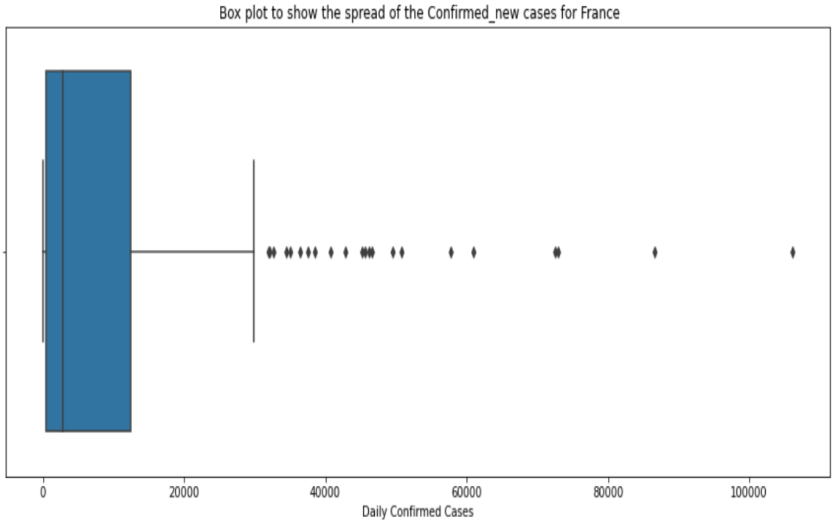
3.2 Outlier Treatment

After seeing the plotted distribution of daily Confirmed Cases for different countries, I discovered that some of the countries have some outliers. After that, I went online and double-checked the highest confirmed cases for those countries.[12] After confirmation, I removed those outliers. **Boxplot visualizations helped me in identifying outliers for each country.**

There was one record in particular for Turkey's daily confirmed cases that was over 800,000. However, Turkey's highest daily reported case count, according to https://www.worldometers.info/ is around 40000.



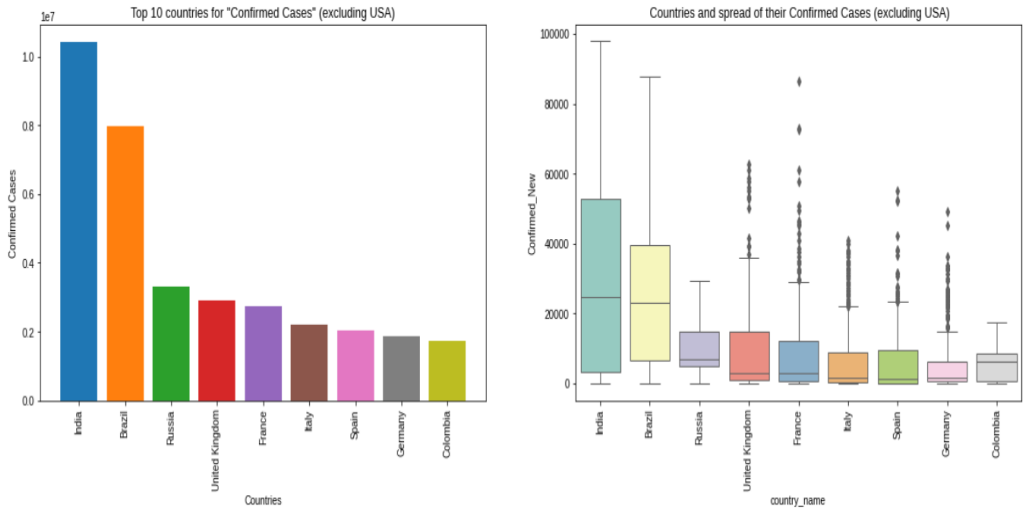
Similarly for France, **there was one record stating that over 100000 confirmed Covid-19 cases were registered on that particular day.** But, the highest number of confirmed cases registered for France is around 80000 according to worldometer.[12]



3.3 Identifying severely affected countries due to Covid-19

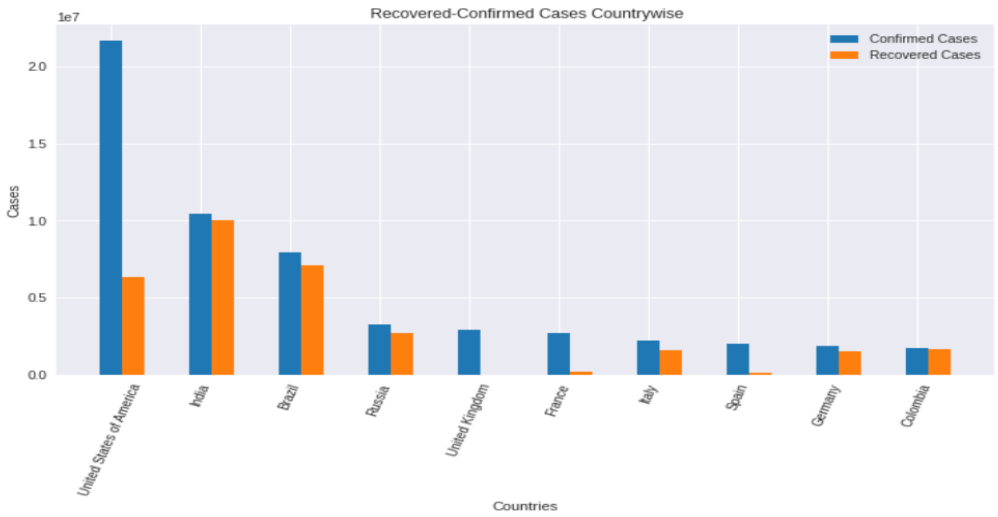
I decided to identify countries that have been seriously impacted by the pandemic and the spread of virus in those respective countries, cases that have been Confirmed. **For each country, I produced a bar plot for the total number of confirmed cases over a year, as well as a box plot for the distribution of those confirmed cases for that respective country.**

The top ten countries with the most infected Covid-19 cases are shown below, along with the spread of number of reported cases in each country. **I've purposely taken the United States off this list because I'm going to discuss it separately.**



3.4 Identifying countries with maximum deaths due to Covid-19

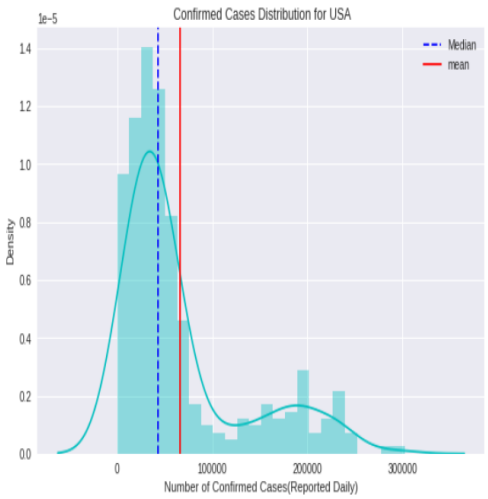
For the severely affected countries, **I used the bar plot to compare Confirmed Cases and Recovered Cases.** Largest the gap between Confirmed and Recovered Cases, highest the mortality rate due to covid-19 for that particular country.

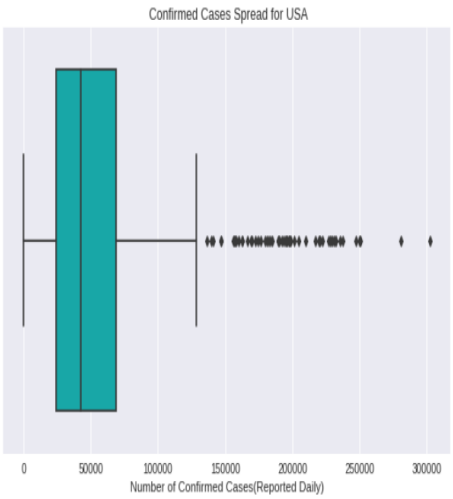


The United States of America has the highest mortality rate due to covid-19, as seen in the comparative plot above. Also, the countries like France, Italy, and Spain are vulnerable to covid-19.

3.5 Analysis for USA

Now, I moved on to the research for the United States of America. The target variable, "Confirmed New" is the number of newly identified Covid cases on daily basis. For the United States, I plotted the distribution of Confirmed Cases as well as the spread of Confirmed Cases.

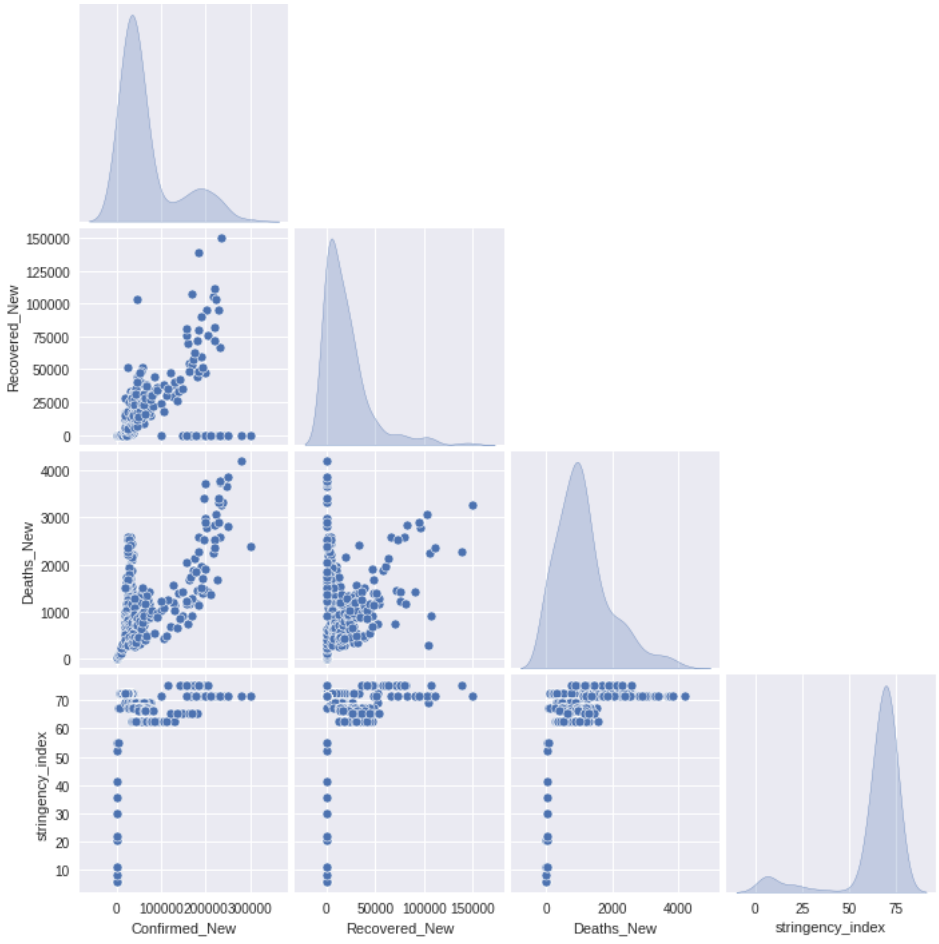




Most of the daily reported "Confirmed New" cases are below 90000. The distribution is right skewed for the country USA. Also, the data points are far spread out from the mean, which indicates high variance in the "Confirmed Cases" (85% of the cases are below 157162 and remaining 15% are between 157162 and 302506).

3.5.1 Correlation between "Confirmed New", "Recovered New", "Deaths New", and "Stringency index"

The correlation between the Stringency index and Confirmed New, Recovered New, and Deaths New for USA is visualized below using pairplot.



Strigency index rose in accordance with the regular number of Reported Cases. That is to say, as the number of infected people increased, policies became more stringent.

3.5.2 Checking Multicollinearity

Since the Stringency Index is made up of all other social indicators. I wondered if there is some multicollinearity between all of the predictor variables.

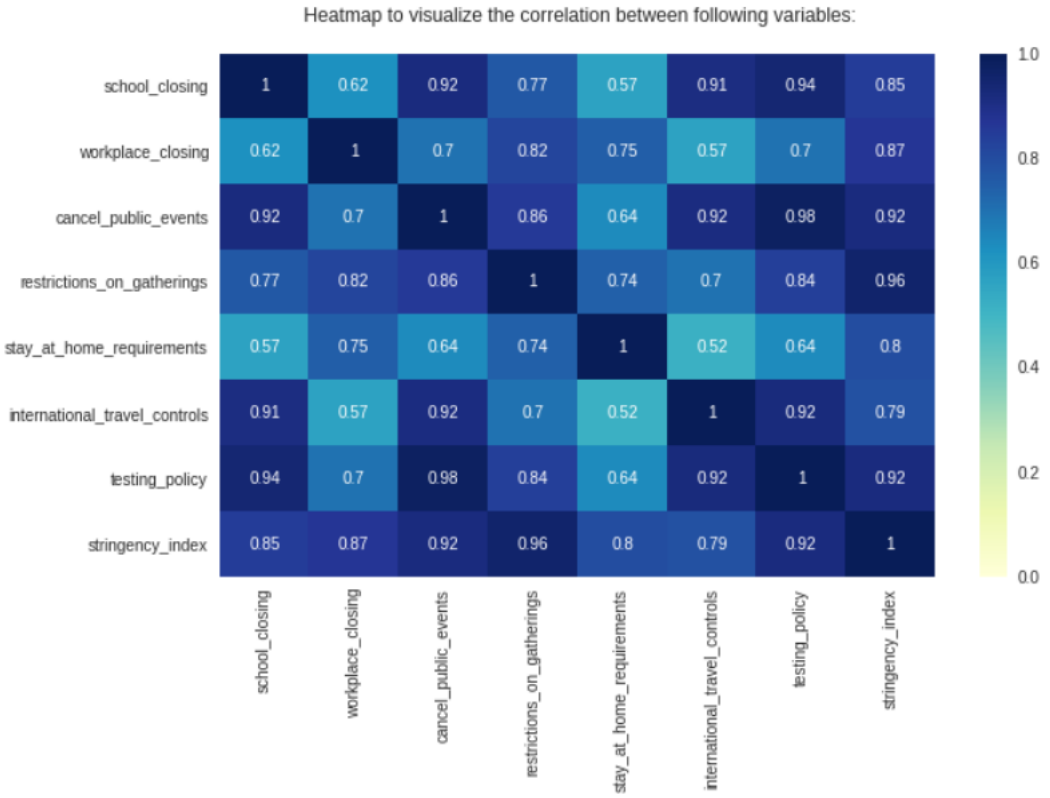
- **Variance Inflation Factor method from statsmodels:** The variance inflation factor (VIF) measures the extent of correlation between one predictor and the

other predictors in a model. A value of 1 means that the predictor is not correlated with other variables. Higher the value, larger the correlation of a variable with other variables.

	variables	VIF
0	school_closing	160.414649
1	workplace_closing	47.142639
2	cancel_public_events	471.422825
3	restrictions_on_gatherings	175.767594
4	stay_at_home_requirements	22.968691
5	international_travel_controls	47.890725
6	testing_policy	724.896210
7	stringency_index	1112.441089

The value of VIF is very high for each of the above variables. **But we can specifically notice that the value for stringency index is highest amongst them.**

- **Plotting Heatmap to visualize the correlation**



"Strigency Index" is highly correlated with all of the other variables. In addition, all of the other variables are fairly correlated with one another. **This indicates that there exist a multicollinearity between all of these variables.**

3.6 Other Features

I concentrated on variables that encodes the values of the government's response policies in this pandemic situation because the study's focus required it. I did, however, look into the relationship between other variables related to mobility and weather, to see whether they had any association with target variables.

To figure out how they are correlated, I generated a heatmap for the input and target variables for mobility and weather data.

We see that the values for correlation of dependent variables with mobility variables and with weather variables are very low. So we can say there is negligible correlation between dependent and independent variables for mobility data and weather data.

Chapter 4

Model Building

4.1 Brief Overview:

We've gone through all of the necessary steps for pre-processing and exploratory data analysis that I've implemented so far. Those measures are critical for moving forward with model construction. ARIMA and VAR are the two Time Series models that I researched and developed. I also tested the models performance on test datasets and used them to do the forecast.

4.2 ARIMA

ARIMA stands for 'AutoRegressive Integrated Moving Average'. It is a time series forecasting algorithm which uses the past values of the time series to predict the future values. **An ARIMA model is characterized by 3 terms: p, d, q where,**

- **p is the order of the AR term**
- **q is the order of the MA term**
- **d is the number of differencing required to make the time series stationary**

Since we only consider one time dependent variable, that is "Confirmed New", this falls under **Univariate Time Series Analysis/Forecasting.**

4.2.1 Stationary or Non-Stationary :

To build an ARIMA model, the first step is to ensure that the time series is stationary. I started by deciding whether the series is stationary or non-stationary.

There are three methods we can implement to check the stationarity of a time series.

Method 1 : Visual Test

Method 2 : Split the dataset into groups and calculate the mean and variance for each group

Method 3 : ADF (Augmented Dickey Fuller) Test

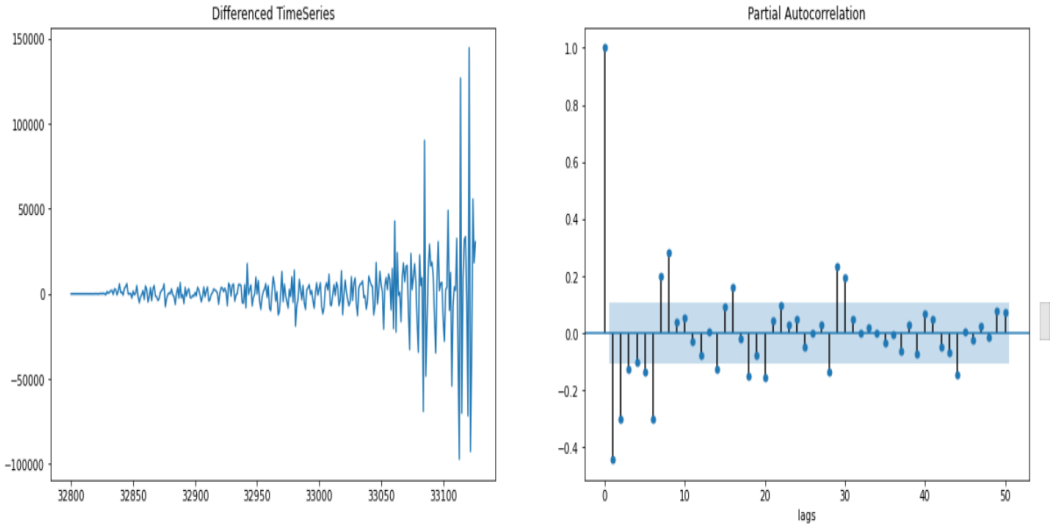
```
Results of Dickey-Fuller Test:
Test Statistic                  0.675235
p-value                         0.989330
#Lags Used                     16.000000
Number of Observations Used    311.000000
Critical Value (1%)             -3.451553
Critical Value (5%)             -2.870879
Critical Value (10%)            -2.571746
dtype: float64

 Weak evidence against null hypothesis,indicating it is non-stationary
```

In above example, the p-value > critical value (0.98 > 0.05), which implies that the series is non-stationary.

I used the **Differencing** method to make the time series stationary and then ran the ADF test on the differenced dataset. **The p-value was below significant threshold value, meaning that it was now stationary.**

4.2.2 Order of the AR term (p)

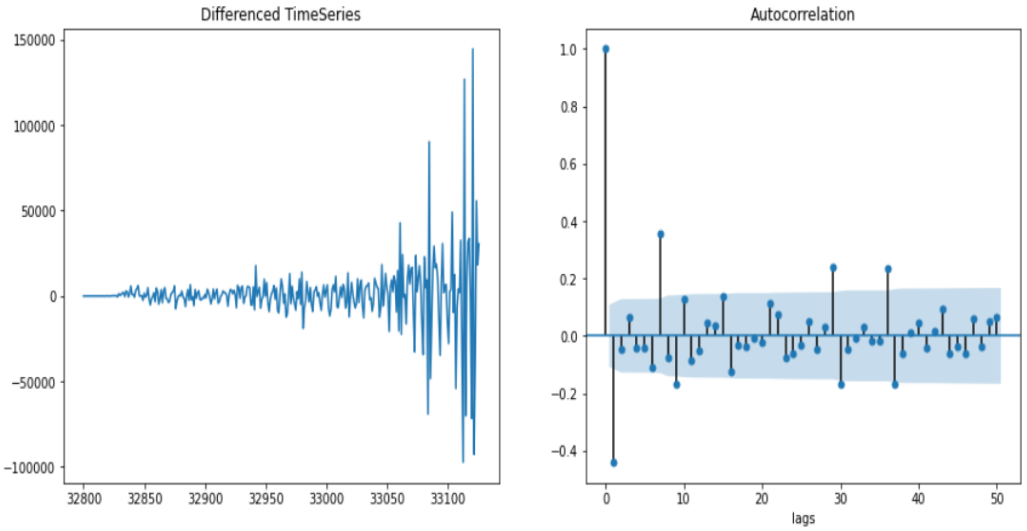
To find the order of AR term, I plotted the Partial Autocorrelation (PACF) plot.



Based on above PACF plot of differenced time series, I decided to have Auto Regressive model with lags 1,2,6,7,8,9

4.2.3 Order of the MA term (q)

To find the order of MA term, I plotted the Autocorrelation (ACF) plot.

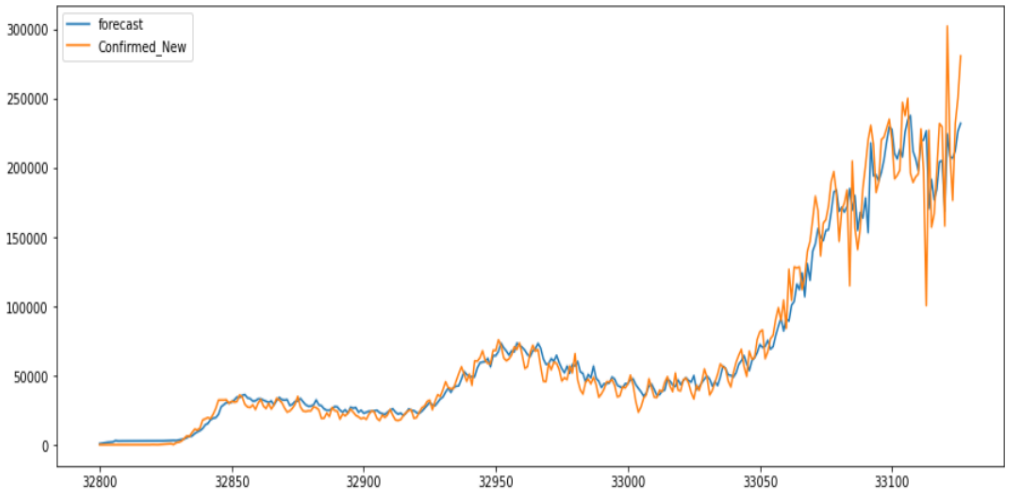


Based on above ACF plot of differenced time series, I decided to have MA model with lags 1,7.

So finally, I developed an ARIMA model(p,d,q) having parameter values as below: p = 1,2,6,7,8,9 and q = 1,7 and d = 1

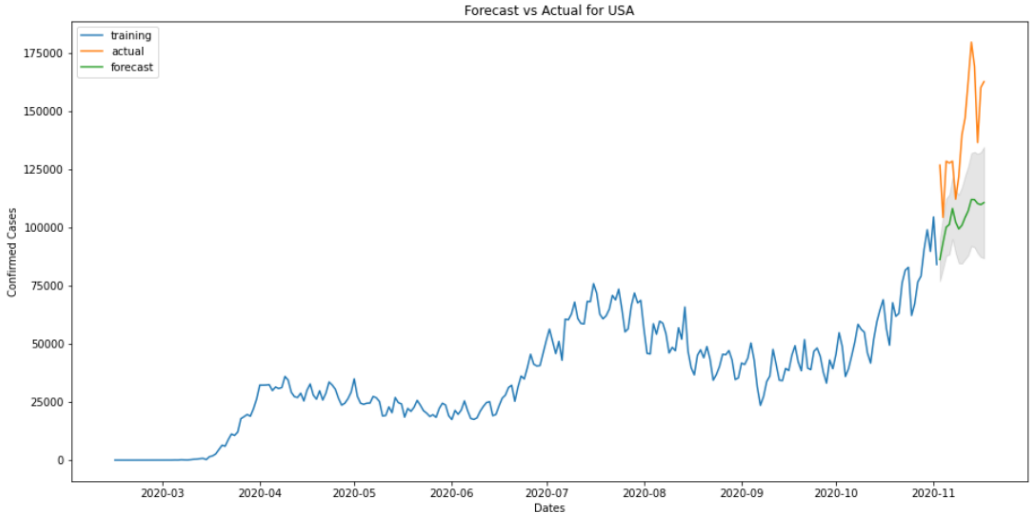
4.2.4 Model Fitting to the data

I used the parameters p=6, q=1, and d=1 to fit the model to the data. We can see in below diagram the model is well fitted to the data.



4.2.5 Evaluation(Prediction) on Test Set

I used this trained model to predict Confirmed Cases in the United States for the next 15 days.



4.2.6 Performance Measures

The commonly used accuracy metrics to judge forecasts are:

```
Mean Absolute Percentage Error is : 24.89429899223602
Mean Absolute Error is : 36631.310674239765
Root Mean Squared Error is :  40402.74319900152
Correlation Matrix between actual and forecast
 [[1.         0.72082206]
 [0.72082206 1.        ]] 0.720822060046947
Correlation between actual and forecast
 0.720822060046947
```

Around 24.89 % MAPE implies that the model is about 75.1 % accurate in predicting the next 15 observations.

4.3 VAR

VAR stands for **'Vector AutoRegression'.** In VAR model, each variable is a linear function of the past values of itself and the past values of the other variables.

This falls under **Multivariate Time Series Analysis/Forecasting** because it involves more than one time dependant variable.

4.3.1 Stationary or Non Stationary

Before we train a VAR model on training dataset, we need to make sure that the data is stationary. I have used ADF Test to check the stationarity of dataset, similar to ARIMA.

4.3.2 Order (p) of VAR model

For this model I considered following features to make predictions of Confirmed Cases for USA : "school closing", "workplace closing","international travel controls","testing policy","stringency index".

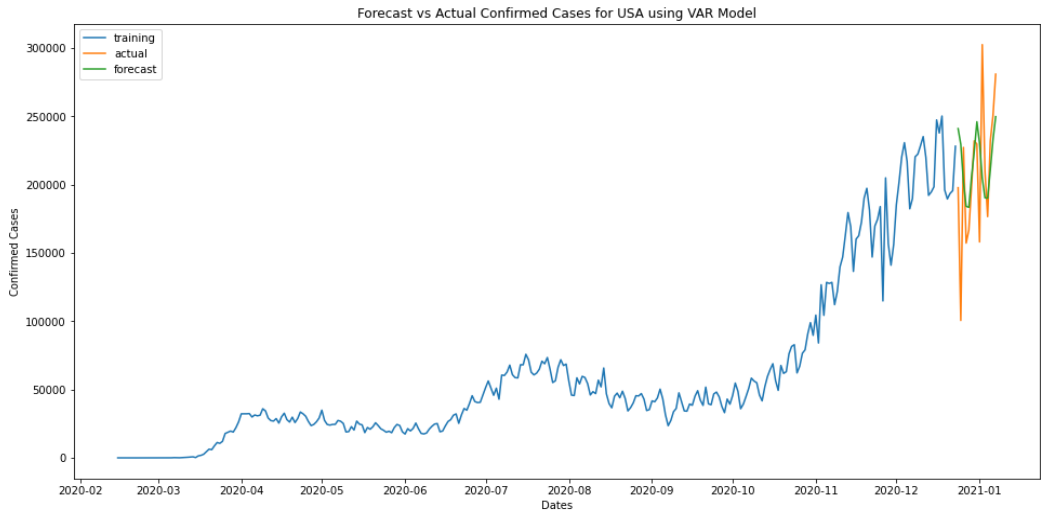
I fit increasing orders of VAR models iteratively until I found the one with the lowest AIC to find the right order of VAR model. And I chose the order with the lowest AIC value.

VAR Order Selection (* highlights the minimums)

	AIC	BIC	FPE	HQIC
0	2.035	2.109	7.649	2.064
1	1.605	2.123	4.977	1.812
2	0.9159	1.879	2.500	1.301
3	-58.75	-57.34	3.059e-26	-58.19
4	-54.13	-52.28	3.102e-24	-53.39
5	-55.76	-53.47	6.096e-25	-54.84
6	-57.08	-54.34	1.634e-25	-55.98
7	-53.47	-50.28	6.080e-24	-52.19
8	-48.99	-45.36	5.376e-22	-47.54
9	-48.38	-44.31	9.975e-22	-46.75
10	-49.55	-45.03	3.127e-22	-47.74
11	-49.75	-44.78	2.598e-22	-47.76
12	-94.85*	-89.44*	6.797e-42*	-92.69*

I choose to begin with a lag order of 7 because the value at lag 8 is rising in comparison to the value at lag 7.

accuracy was not optimal when the order of VAR was equal to 7. So I played with various order values. And it was at order 12 that the best accuracy was found.



4.3.3 Performance Measures

The commonly used accuracy metrics to judge forecasts are:

```
Mean Absolute Percentage Error is : 21.664461237751826
Mean Absolute Error is : 36394.16577987939
Root Mean Squared Error is :  50084.9817370483
```

Around 21.6% MAPE implies that the model is about 78.4% accurate in predicting the next 15 observations.

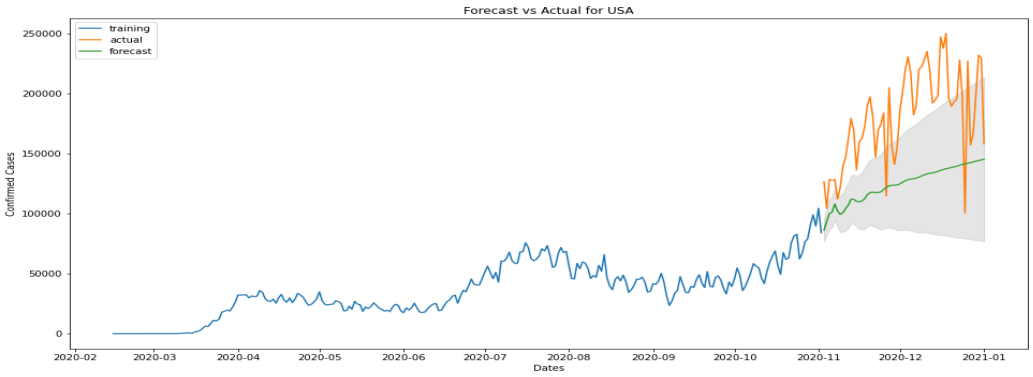
4.4 Evaluation of Metrics

Performance Metrics for ARIMA and VAR Models:

Performance Metrics of different Models				
Model Name	MAPE	RMSE	MAE	Predictor
ARIMA	24.89%	40402.74	36631.31	Confirmed cases
VAR	21.6%	50084.98	36394.16	Confirmed cases and other

4.5 Observations on Comparative Study of VAR and ARIMA

ARIMA has correctly captured the trend as we can see in above prediction diagram, however the prediction was done for 15 days only. When I forecast for two months with the same model, the prediction curve flattens out after a while. As a result, I believe ARIMA is great for making short-term predictions.



With the dataset used in this case study, VAR provided better accuracy for Covid-19 prediction. The variables used in VAR were strongly correlated with one another.

Chapter 5

REFERENCES

5.1 Bibliography

[1] *Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm* Cafer Mert Ye ¸s ilkanat Science Teaching Department, Artvin Çoruh University, Artvin, Turkey. 20 August 2020
[2] *Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy*, Emrah Gecili, Assem Ziady, Rhonda D. Szczesniak
[3] *Population flow drives spatio-temporal distribution of COVID-19 in China*, Jayson S.Jia, Xin Lu, Yun Yuan, Ge Xu, Jianmin Jia,, 29 April, 2020
[4] Shreshth Tuli, Shikhar Tuli , Rakesh Tuli , SukhpalSingh Gill. *Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing* Queen Mary University of London, UK, 12 May 2020
[5]Kumar, Naresh, and Seba Susan. *"COVID-19 Pandemic Prediction using Time Series Forecasting Models."* arXiv preprint arXiv:2009.12176 (2020).
[6] *Tailoring time series models for forecasting coronavirus spread:Case studies of 187 countries*, Leila Ismail,Huned Materwala,Taieb Znati,Sherzod Turaev,Moien A.B. Khan, 24 September 2020
[7] Rauf, H.T., Lali, M.I.U., Khan, M.A. et al.*Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks.* Pers Ubiquit Comput (2021). https://doi.org/10.1007/s00779-020-01494-0
[8] Propagation analysis and prediction of the COVID-19 Lixiang Li, Zihang Yang, Zhongkai Dang, Cui Meng, Jingze Huang, Hao Tian Meng,Deyu Wang, Guanhua Chen, Jiaxuan Zhang, Haipeng Peng, University of Posts and Telecommunications, Beijing 100876, China
[9] *COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population*, Vasilis Papastefanopoulos , Pantelis Linardatos and Sotiris Kotsiantis, Department of Mathematics, University of Patras, 26504 Patras, Greece, 3 June 2020.
[10] World Health Organization, WHO Coronavirus (COVID-19) Dashboard, 2020, https://covid19.who.int/.
[11]. Robert Koch Institute COVID Situation report
[12] Worldometer,https://www.worldometers.info/coronavirus/