Coronavirus

Project

4/17/2020

Introduction

The Coronavirus is a serious concern around the globe. In this report we analyzed the dataset Coronavirus which has been collected from WHO (World Health Organization) by doing exploratory data analysis followed by performing time series analysis.

Data and Planning

The Coronavirus dataset contains 70,020 observations and 7 variables which gives the information about the number of confirmed positive cases, number of deaths and number of recovered cases due to Coronavirus for different countries from January 22,2020 to April 21, 2020.

Analysis

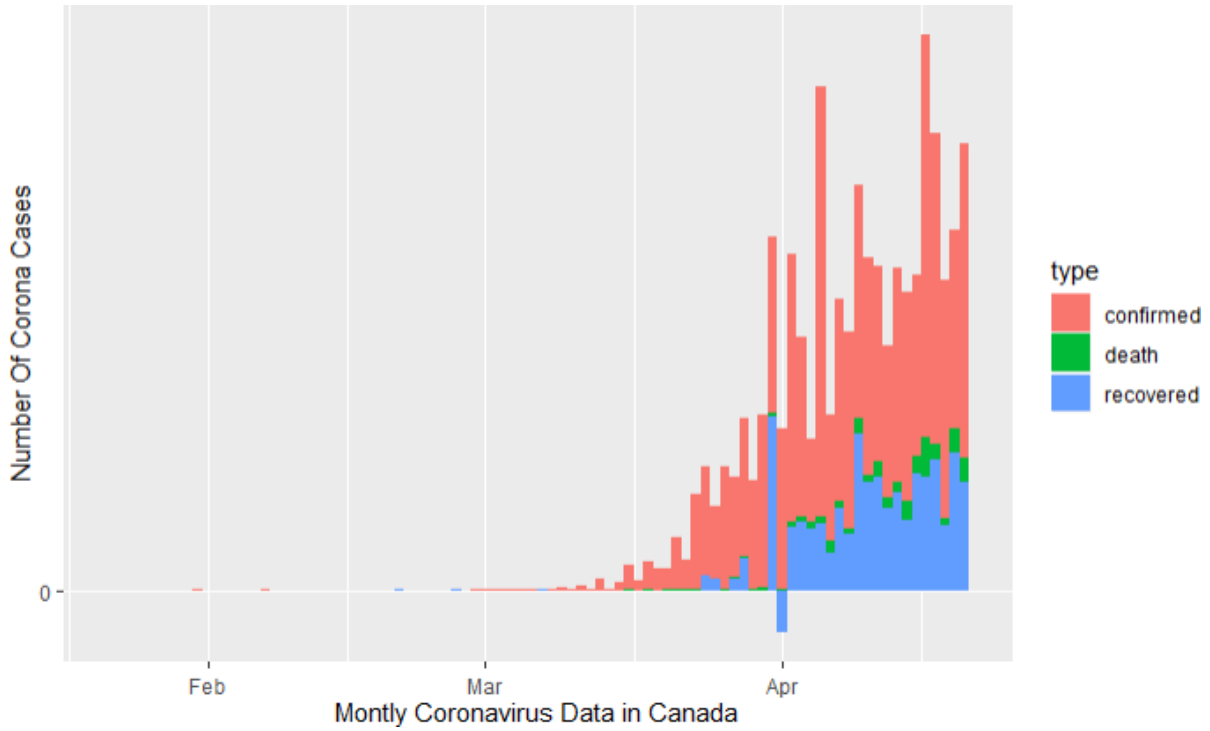
After exploring the dataset,in order to continue our analysis further we have filtered the data by selecting the country as “CANADA” and created a new dataset “Canada_D” from the original Coronavirus dataset.The Coronavirus data for Canada has 2,790 observations.

Data Cleaning / Tidying Data: Getting the Data ready

By looking at our new dataset “Canada_D”, we can say that there are no missing values in our data. Our filtered data is not so clean for our analysis. So we have created a new dataset “Tidy_data” by making our dataset “Canada_D” wider by increasing the number of columns and decreasing the number of rows.The three new columns confirmed, deaths, recovered have been added and this makes our data tidy.The “Tidy_data” contains 1440 observations and 8 variables as seen below.

Data Analysis

For better understanding, let’s visualize our data by plotting it.From the below graph we can say that number of positive corona cases are steadily increasing from 1st week of march. The number of recovered cases are very less when compared to number of confirmed cases. The number of deaths are very low as of now but we cannot predict what might happen. Let’s check the correlation between number of deaths and number of confirmed cases due to Coronavirus.



Correlation

As the data is not normally distributed (fromm Q-Q plot). We are Running kendall's correlation test on the data which gives us the below results. The number of deaths is significantly positively correlated with the number of confirmed Coronavirus cases, r = 0.511. A correlation of 0.511 represents a medium effect explaining 26.11% of the variance.

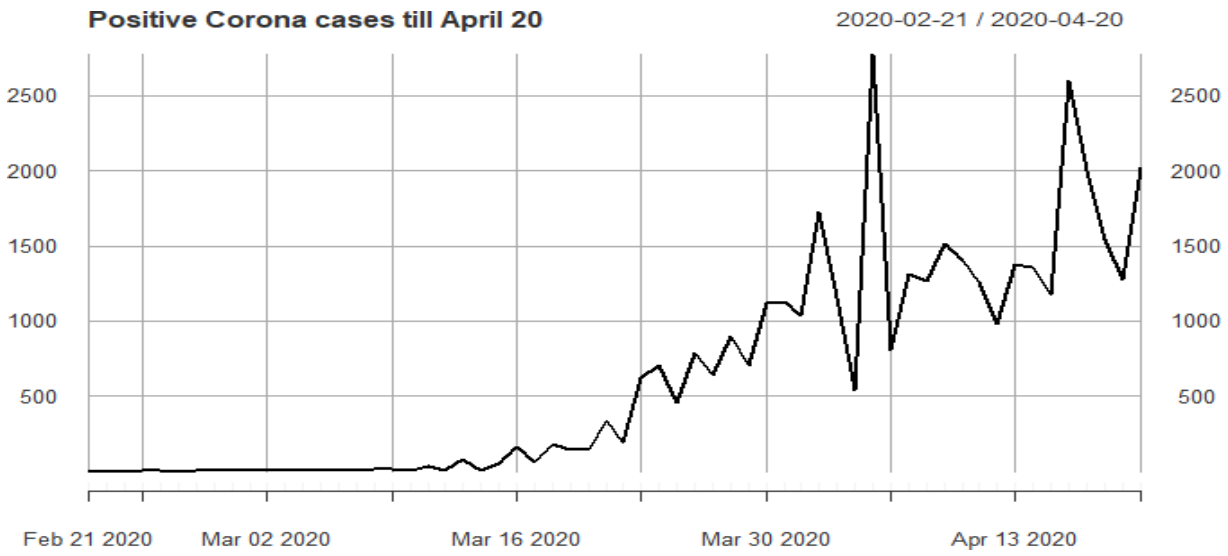
From the correlation graph (rmd code), we can conclude that these two variables are slightly linearly related but this relationship is not casual.Lets continue our analysis to predict the number of confirmed cases in future based on our data by time series forecasting.

Time Series Analysis

In order to perform time series analysis on our data, our data should be in time series format. We filtered the data i.e. confirmed number of cases w.r.t date and transformed this into time series data. We have removed the first 30 rows in our data because the number of confirmed cases are mostly zero during that period of time, as it is in the month of January. This will not affect our analysis so we are considering data from february 21,2020 to April 21,2020. So, our time series data is assigned to a variables "ts_data".

Exploratory Data Analysis on Time Series data

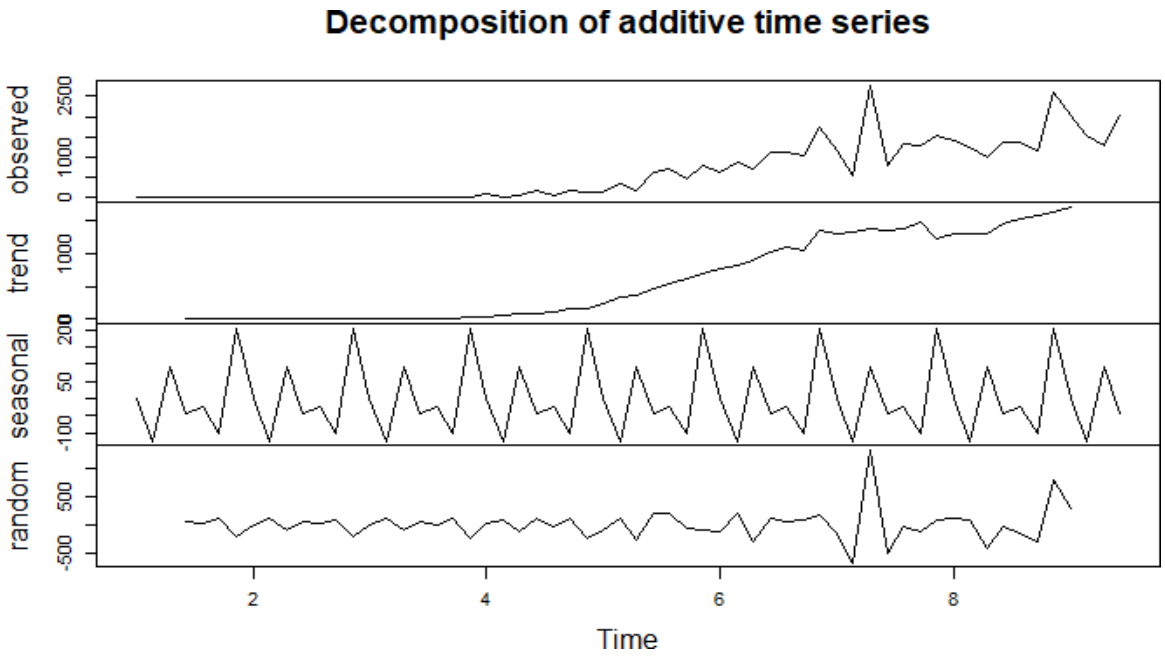
Let's visualize our time series data for better understanding. From the below graph we can say for the first 3-4 weeks starting from february 21,2020 the number of positive corona cases are less than 5 and has not increased. But the number of cases are increasing slowly from March 16,2020 and we got the highest peak 2778 positive corona cases have been registered on April 5, 2020 and then there is a slight decrease in the number of cases but then its not constant this is changing with time.



From the above figure, We limit our attention to trended and non-seasonal models as the Coronavirus is a univariate time series which is not having any seasonal affect because this is not a stockmarket data or any other data which has seasonal affect.We assume that the trend will continue indefinitely in the future because of the ongoing situation around the globe. We opt for an ARIMA model with additive error and additive trend components. Even if in some cases an multiplicative trend model gives lower information criteria values, we opted for the additive trend model given the asymmetric risks involved as we believe that it is better to err to the positive direction.

Time Series Decomposition;

We have set the frequency of the time series object "ts_data" to weekly for decomposing the time series. After decomposition we can visualize the underlying categories of patterns.



The plot above shows the original time series (top), the estimated trend component (second from top), the estimated seasonal component (third from top), and the estimated irregular component random (bottom). We see that the estimated trend component shows a steady increase after 4 weeks.

TEST STATIONARITY OF THE TIME SERIES

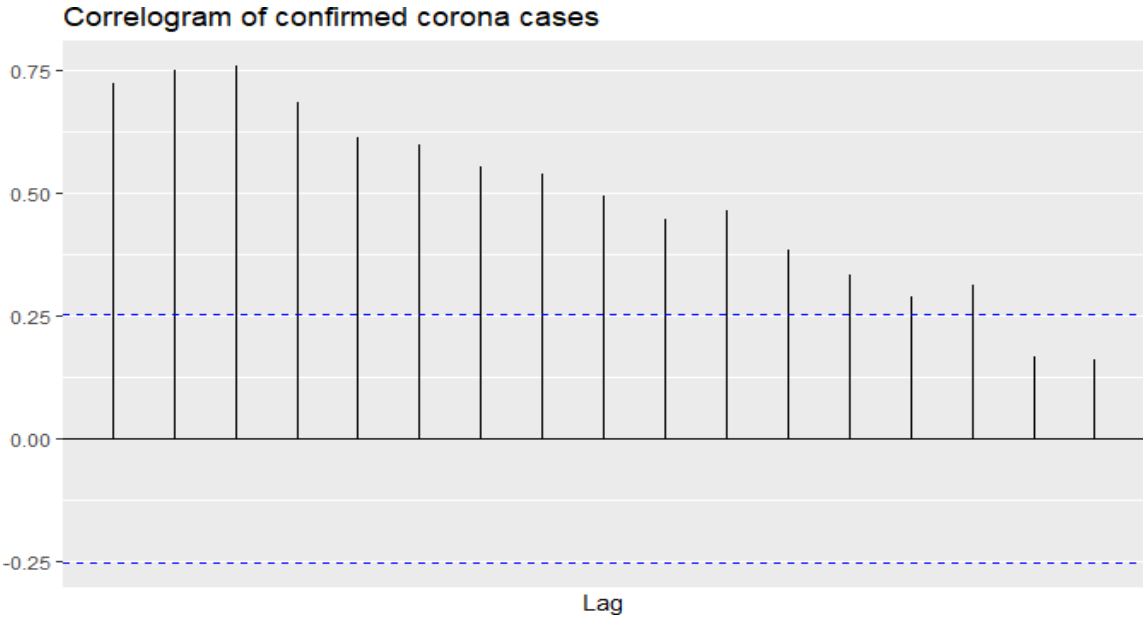
In order to fit arima models, the time series is required to be stationary. We will test the stationarity of the time series by two methods

Test stationarity of the time series (Augmented Dickey-Fuller Test-ADF):

The null hypothesis H0: The time series is non stationary, The alternative hypothesis HA: that the time series is stationary. As per the test results above, the p-value is 0.4371 which is >0.05. So, we will accept the null hypothesis that the time series is not stationary.

Test stationarity of the time series (Autocorrelation):

We will test the stationarity of the time series by autocorrelation function (acf). This plots the correlation between a series and its lags. As the ACF decreases slowly , we can say that our time series is non stationary.



TRANSFORMATION OF TIME SERIES

Ideally, we want to have a stationary time series for modelling. Of course, not all of them are stationary, but we can make different transformations to make them stationary.In our analysis, we checked by using difference transformation and BoxCox transformation to convert our data into stationary time series data. By using BoxCox transformation we can say that our data has converted into slightly stationary but not completely.

FIT A TIME SERIES MODEL

From the linear model graph (rmd) we can say that this may not be the best model to fit as it doesn't capture the seasonality and multiplicative or additive effects over time.

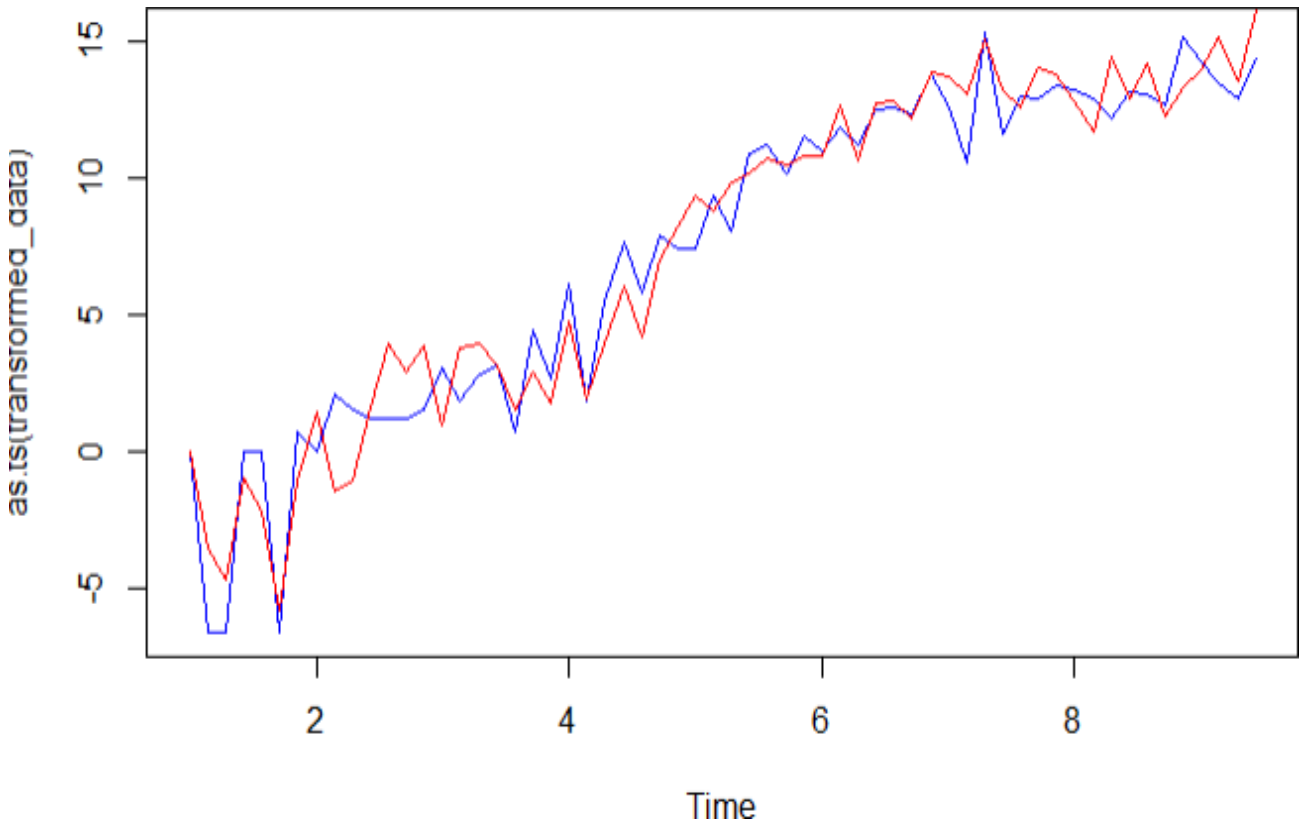
ARIMA MODEL

We are using auto.arima function to fit our model.Note we have used the ARIMA modeling procedure as referenced (rstudio-pubs-static.s3.amazonaws.com/311446_08b00d63cc794e158b1f4763eb70d43a.html).The results of the model are present as below.

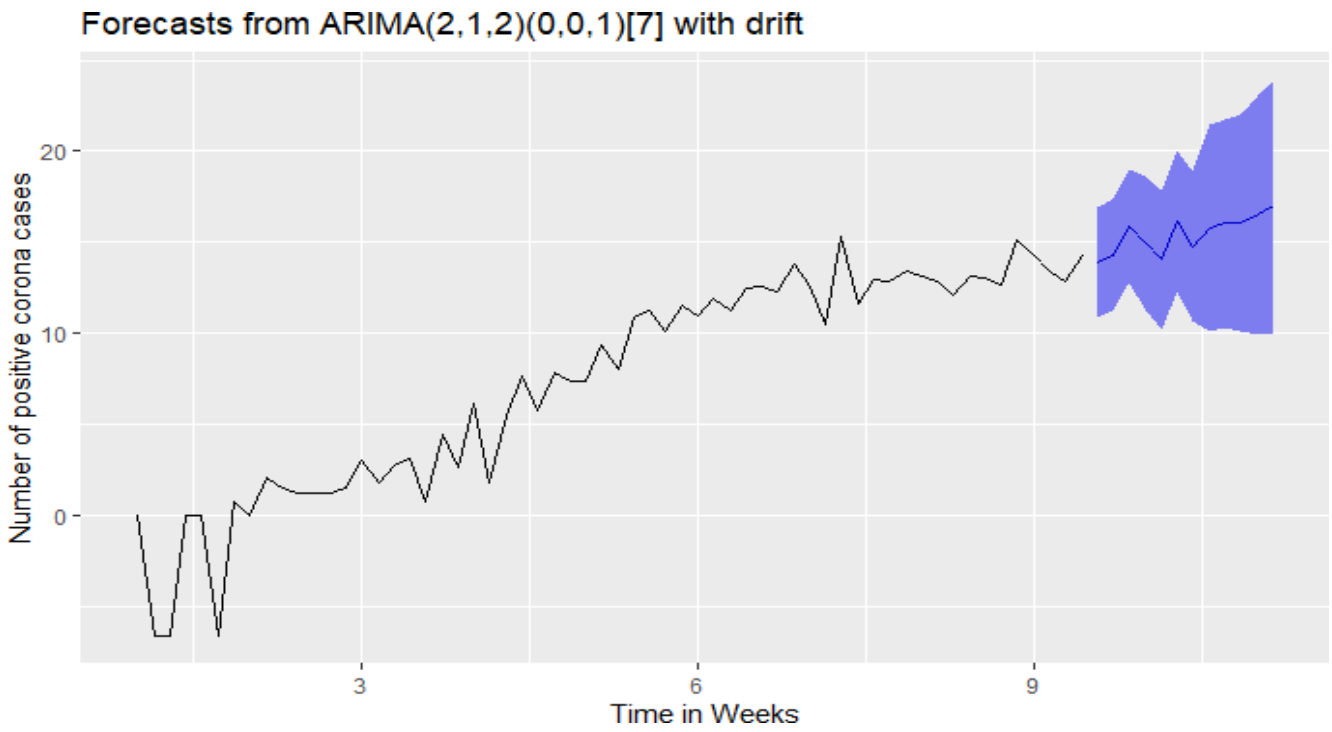
```
Series: transformed_data
ARIMA(2,1,2)(0,0,1)[7] with drift

Coefficients:
          ar1      ar2      ma1     ma2    sma1   drift
      -0.2696  -0.4929  -0.8355  0.4730  0.9130  0.3341
s.e.   0.3218   0.1960   0.3437  0.2472  0.5511  0.1257

sigma^2 estimated as 2.322:log likelihood=-111.97
AIC=237.93   AICc=240.13   BIC=252.47
```

MODEL VALIDITY

Let's check how much our fitted data fits the original data by plotting it.From the below graph, we can say that the blue color line indicates the original data and red color line indicates the fitted data. Our model approximately fits our original data. From the below plot, we can say that the residual plots appear to be centered slightly around 0 as noise, with no pattern the arima model is a fairly good fit.



CALCULATE FORECASTS



FORECASTING VALUES (for the next 2-3 weeks)

	Point Forecast <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
9.571429	1791.670	619.1053	4475.038
9.714286	2032.032	713.1118	5017.356
9.857143	3285.046	1229.7002	7727.400
10.000000	2455.843	713.5826	6955.211
10.142857	1879.084	481.6602	5808.151

CONCLUSION

From our analysis, we can conclude that Coronavirus is indeed a disease to be eradicated from the world as it can be seen from the predicted positive Coronacases numbers. We can also say that the number of confirmed corona cases in the future are going to increase steadily by visualizing from the forecasting model graph.As per our model, we can say that in the coming 1-3 weeks i.e. from April 22nd to May 9th the number of cases are going to increase between 12,626 and 1,21,589. In a best situation considering all the changes happening and the precautions taken by the people by staying at home, the predicted number can be low.