Final Project MSCI 718 - Statistics for Data Analytics

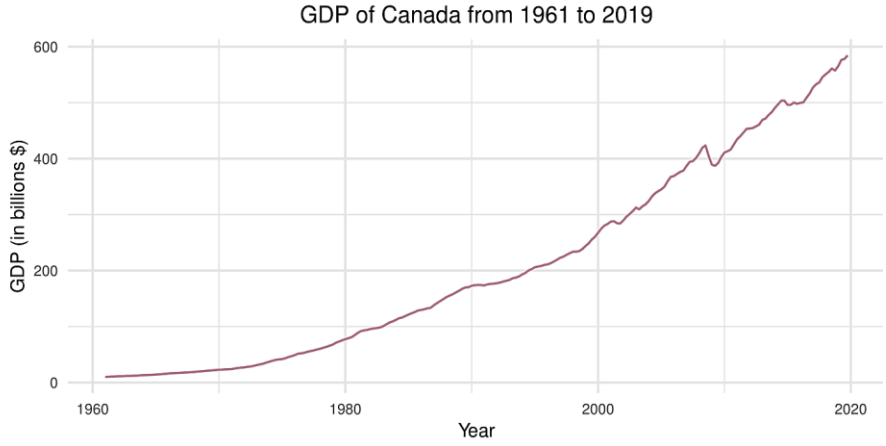
Aim: Forecasting the quarterly GDP of Canada for the year 2020 using ARIMA (Time Series Analysis)

Introduction to ARIMA

ARIMA stands for Auto Regressive Integrated Moving Average. Auto Regressive (AR) terms refer to the lags of the differenced series, Moving Average (MA) terms refer to the lags of errors and (I) is the number of differences used to make the time series stationary. Time Series data is an aggregate/combination of four components (Data, Trend, Seasonality and Irregularity). These components are useful abstraction that are aggregated/combined either additively or multiplicatively and facilitate the selection of forecasting methods.

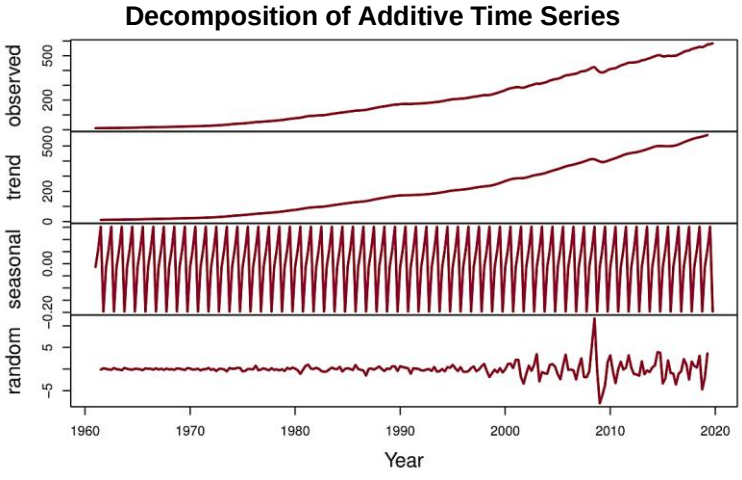
Exploring the data

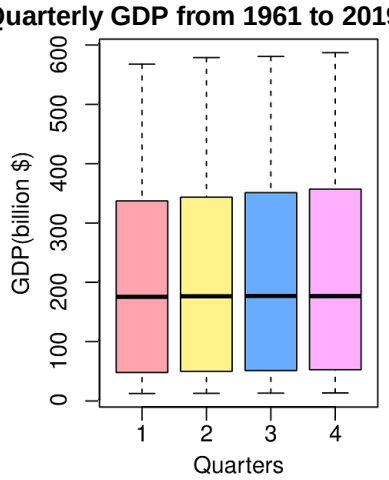
The dataset consists GDP (in billions $) of Canada **from 1961 to 2019**, segregated **quarterly** (Q1 to Q4). The data was obtained from https://fred.stlouisfed.org/, developed by Federal Reserve Bank of St. Louis. There are **no missing values** (checked using is.na(gdp)) and **no duplicates** (checked using gdp[duplicated(gdp),]). Our GDP data is converted into univariate time series data format and plotted as shown. It is observed from the graph that the data points follow an overall upward trend with no outliers and constant variance, but mean is not constant rather mean increases from 1960 to 2019.



Identifying trend and seasonality

Time Series Decomposition: Decomposing the time series involves tying to separate the time series into the below individual components shown in the plot. From the below graphs, we draw the following inferences:





Trend: The decomposed time series plot suggests that GDP increase over time with each year which is an indication of an increasing linear trend.
Seasonality: The boxplot suggests the absence of seasonality, with GDP showing no seasonal increase or decrease in any of the 4 quarters.

Testing the Assumptions

The following **assumptions** of ARIMA model should be tested before fitting the model:

1) **Data should be univariate** –ARIMA works on a single variable. Auto-regression is all about regression with the past values. For this project, data is univariate.
2) **Data should be stationary** – The properties of the series do not depend on the time when it is captured (constant variance and mean). A white noise series or cyclic behavior series can also be considered as stationary. This is tested using **autocorrelation plots (using ACF function)** and the two tests: ***Augmented Dickey-Fuller Test (ADF)*** and ***Kwiatkowski–Phillips–Schmidt–Shin unit root test (KPSS)***.

Testing Stationarity of the time series:

1) **Augmented Dickey-Fuller Test (ADF)**
   Null hypothesis, Ho: The time series is non-stationary
   Alternative hypothesis, Ha: The time series is stationary

```
## Augmented Dickey-Fuller Test
##
## data:  gdp3
## Dickey-Fuller = -0.88604, Lag order = 6, p-value = 0.9529
## alternative hypothesis: stationary
```

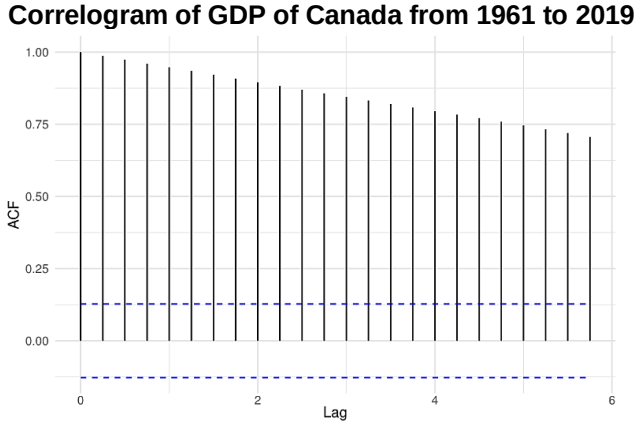
   As per the test results above, the p-value is 0.9529 which is >0.05 so, we do not reject the null in favor of the alternative hypothesis and conclude that time series is **not** stationary.

2) **Kwiatkowski–Phillips–Schmidt–Shin test (KPSS)**
   Null hypothesis, Ho: The time series is stationary
   Alternative hypothesis, Ha: The time series is non-stationary due to the presence of a unit root

```
## KPSS Test for Level Stationarity
##
## data:  gdp3
## KPSS Level = 4.6226, Truncation lag parameter = 4, p-value = 0.01
```

   As per the test results above, the p-value is 0.01 which is less than 0.05 so, we reject the NULL hypothesis, concluding that the time series is **not** stationary.

3) **Autocorrelation plot (ACF)**
   ACF function plots the correlation between a series and its lags i.e. previous observations with a 95% confidence interval shown by blue dashed lines. If the autocorrelation crosses the dashed blue line, it means that specific lag is significantly correlated with current series. The following plot further proves **non-stationarity** as it is in the form of a slowly decaying function.

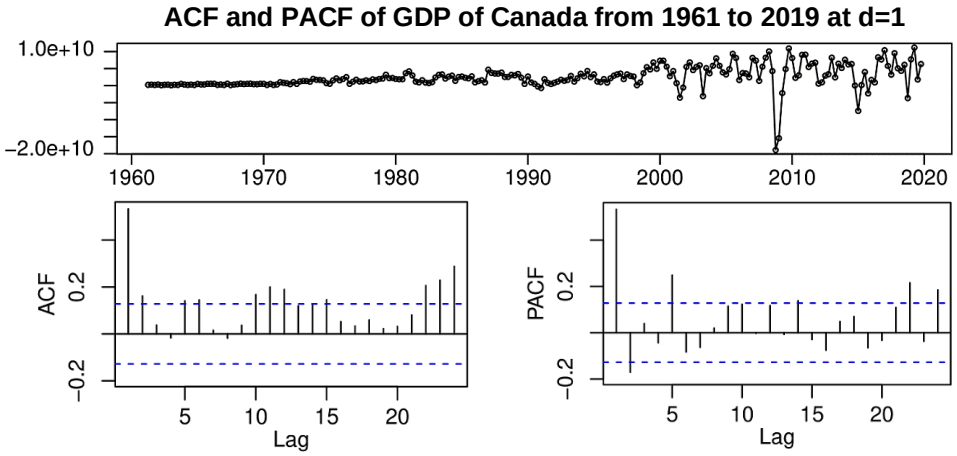


Transforming the Data

Making the data stationary

To achieve stationarity, the data is differentiated i.e. difference between consecutive observations. Using (diff) function, the data is differentiated twice and Autocorrelation and ADF and KPSS tests are checked again.

1) **d=1**, i.e. differentiating data once



After differentiating once (i.e. d=1), we can still see autocorrelation in the ACF and PACF plots. Hence, further differentiation is required as the data is still not stationary.

2) **d=2**, i.e. differencing data twice

Running the ADF and KPSS tests on the double differentiated data:

a) Augmented Dickey-Fuller Test (ADF)

Null hypothesis, Ho: The time series is non-stationary

Alternative hypothesis, Ha: The time series is stationary

```
## Augmented Dickey-Fuller Test
##
## data: d2
## Dickey-Fuller = -8.6699, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

The p-value is 0.01 which is <0.05 therefore we reject the null hypothesis in favor of the alternative hypothesis and conclude that time series is stationary.

b) Kwiatkowski–Phillips–Schmidt–Shin test (KPSS)

Null hypothesis, Ho: The time series is stationary

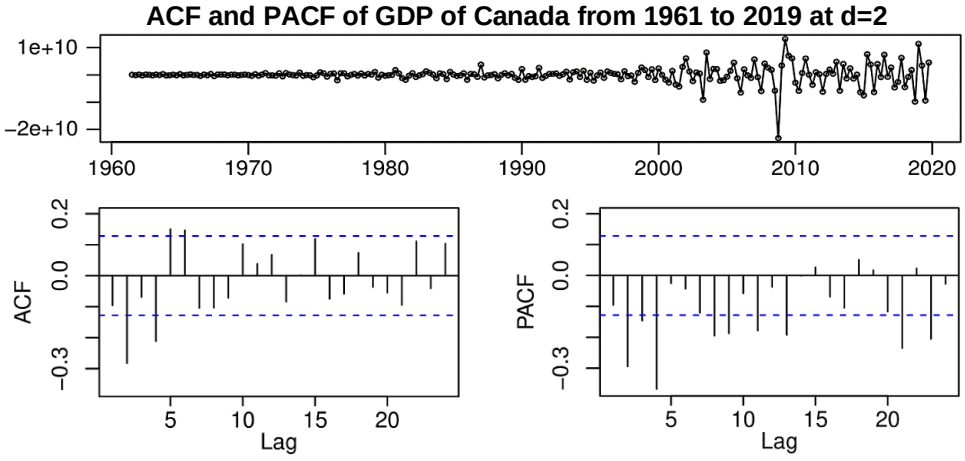
Alternative hypothesis, Ha: The time series is non-stationary due to the presence of a unit root

```
## KPSS Test for Level Stationarity
##
## data:  d2
## KPSS Level = 0.011557, Truncation lag parameter = 4, p-value = 0.1
```

The p-value is 0.1 which is greater than 0.05 so, we do not reject the NULL hypothesis, concluding that the time series is stationary.

Hence, the data is now stationary.

c) Examining the ACF and PACF plots for stationary data



Interpreting the above graph

1) From the above ACF and PACF charts we see that the first lags are not significant, and rest of the lags do not show a decaying pattern, thus, we do not require Auto Regressive Model (AR).
2) As there are few spikes in the series and rest all are zero, we can use a Moving Average model (MA).
3) Here we see that ACF gives max of 2 postive significant lags. Hence q=2 is a possible model.
   Secondly, PACF gives us no positive significant lags. Hence p=0 is a possible model.
4) Since the autocorrelation is negative, we should add an SMA term in the model (i.e Q term). A pure SMA(1) process has spikes in the PACF at lags s, 2s, 3s, etc., while the ACF cuts off after lag s, which is the case here. Thus, we will be fitting SMA(1), i.e. Q=1 model.

Model Selection and Fitting

Understanding p, d, q and P, D, Q values of the model

1) p, d & q
Used to specify non-seasonal components of the model. (If multiple values of p and q are provided, the one which minimizes AIC will be chosen.)
p= The order of the non-seasonal auto-regressive (AR) terms.
d= The order of integration for non-seasonal differencing (values selected via repeated KPSS tests).
q= The order of the non-seasonal moving average (MA) terms.

2) P, D & Q
Used to specify seasonal components of the model. (If multiple values of P and Q are provided, the one which minimizes AIC will be chosen.)
P= The order of the seasonal auto-regressive (SAR) terms.
D= The order of integration for seasonal differencing (values selected via repeated heuristic tests)
Q= The order of the seasonal moving average (SMA) terms

Selecting the best model

1) Now that all the assumptions are met and stationarity has been achieved, we will fit the ARIMA model on the differentiated time series data to forecast the 2020 GDP.
2) As observed above using the ACF & PACF plots for stationary data, we need to differentiate the data twice (d=2); and ACF & PACF functions helped us determine that a MA model will be a better fit than AR model or a combination of AR-MA model. Also, we saw the need to have and SMA(1) term. Thus, keeping Q=1.
3) Therefore, we will check which model is the best fit for p=0 (as AR is not required), d=2 and different values of q. Usually we prefer p+q should be less than and equal to 3 else there is a risk of over-fitting. Hence, we will check different models for q={1,2,3} keeping p=0 and d=2.

```
## [1] 10871.35
## [2] 10820.89
## [3] 10821.23
## [4] 10821.54
```

From the above results, we observe that model #2 gives us the lowest AIC value (aic = 10820.89). Hence we proceed with this model. The selection was verified by using auto.arima function as well (auto.arima(gdp3)) and the same model was suggested by R.

Model summary

The p, d, q and P, D, Q components of chosen model #2

```
##      p      d      q      P      D      Q  Frequency
##      0      2      2      0      0      1          4
```

Summary of the model:

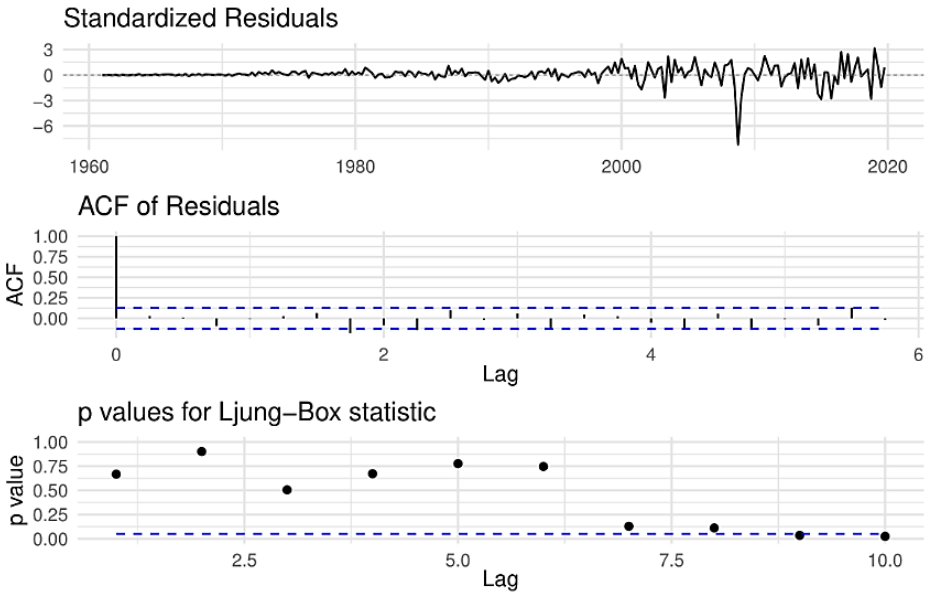
```
##
##  Call:
```

```
##  arima(x = gdp3, order = c(0, 2, 2), seasonal = c(0, 0, 1))
##
##  Coefficients:
##                ma1       ma2      sma1
##            -0.4227   -0.4987   -0.2479
##    s.e.     0.0532    0.0552    0.0757
##
##  sigma^2 estimated as 6.777e+18:  log likelihood = -5406.44,  aic = 10820.89
##
##  Training set error measures:
##                          ME          RMSE          MAE         MPE       MAPE      MASE         ACF1
##  Training set      293934019    2592290650   1475496203   0.3118554  0.7957327 0.5024098  0.02777016
```

Goodness of Fit

Testing the goodness of fit for the selected ARIMA model

1) Using Residuals plot



The residual plots appear to be centered around 0 as noise, with no pattern. So, the arima model is a good fit.

2) Using Ljung-Box test to test Autocorrelation

Null hypothesis, Ho: Autocorrelations are not significantly different from zero
Alternative hypothesis, Ha: Autocorrelations are significantly different from zero

```
##
## Box-Ljung test
##
## data:  arimaModel2$residuals
## X-squared = 0.18432, df = 1, p-value = 0.6677
```

Ljung-Box test p-value is > 0.05, hence we do not reject the NULL hypothesis. Therefore, there is little evidence of non-zero autocorrelations and we can assume the residuals are white noise.

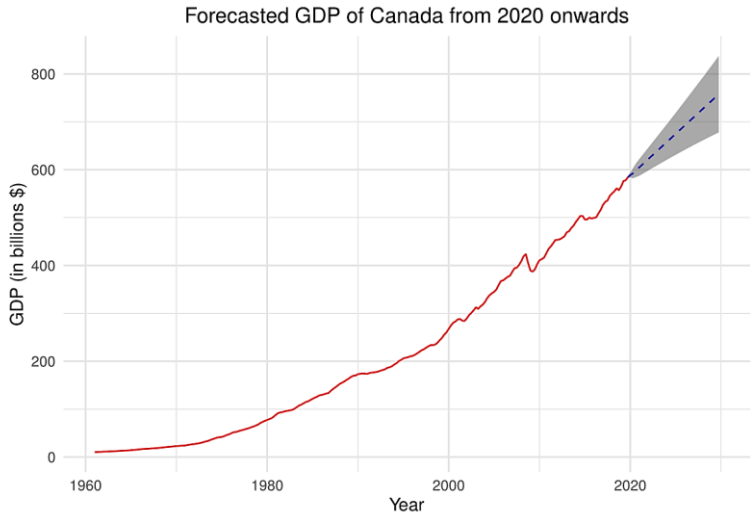
After proving that the model is a good fit, we proceed to forecasting.

Forecasting

1) Predicted values of GDP of Canada for Quarters 1,2,3 and 4 for the year 2020

```
##          Point      Forecast        Lo 80        Hi 80        Lo 95        Hi 95
##  2020      Q1  589009460750  585673139779  592345781720  583906998142  594111923358
##  2020      Q2  591686062730  585455204049  597916921411  582156786706  601215338754
##  2020      Q3  596586785486  588259511956  604914059015  583851319628  609322251343
##  2020      Q4  600874291910  590733873151  611014710669  585365859717  616382724103
```

2) Plot for predicting the GDP for next 10 years (from 2020 onwards) with 95% confidence interval



Model Validation

Testing the model accuracy by assessing the fit of the forecasted object over training and testing period.

Training period: 1961-2004 which is about 75% of the data
Testing period: 2005-2019 which is about 25% of the data

##	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
## Training set	144733869	1328845725	813261525	0.207	0.724	0.108	0.025	NA
## Test set	-11993956193	18321397379	16106777051	-2.421	3.462	2.142	0.917	2.561

The MASE (Mean Absolute Scaled Error) of test set is 2.142 and MAPE (Mean Absolute Percentage Error) is 3.462, which indicates that percentage of error in the model is about 3.5%. So, the model predicts the GDP with ~96.5% accuracy. Thus, the model selection is justified with 96.5 % of accuracy in predicting the GDP.

Conclusion, Practical Significance and Future Work

1) GDP is considered one of the best measures to assess how well the economy is performing. Also, it is a vital basis for government to set up economic developmental strategies and policies. Therefore, an accurate prediction of GDP is necessary to get an insightful idea of the future trend of an economy. So, ARIMA was chosen as it gives the short-run forecast for a large amount of data with high precision.

2) Using ARIMA model, Canada's GDP is forecasted to reach $600 billion until the end of 2020. The GDP was forecasted with 96.5% accuracy (however extreme circumstances like recession or pandemic can give significantly different results, but these events occur very rarely). We expect that the Canadian GDP will continue to raise according to the forecasted values form our model. Moreover, the data was segregated into training set and testing set to further validate the model's accuracy.

3) For this project, the data was first converted to a time series data with one variable (GDP in billions $) for year 1961 to 2019 with frequency = 4 (considering four quarters in year). The various assumptions of trend/seasonality were tested, and the data was double differenced to make it stationary. By ACF and PACF plots, multiple models were created and the one with lowest AIC value was fitted for forecasting.

4) ARIMA was chosen as it can handle time series data with trends and seasonality. Other conventional methods such as Holt Exponential Smoothing or Holt-Winter Exponential Smoothing can be used when all the characteristics (such as seasonality) of data are known. In future, forecasting time series could be done by considering uncertainty bounds, change point and anomaly detection.

5) The findings of this project have important implications for policy makers to formulate better policy to deal with foreign direct investment (FDI) and foreign institutional investment (FII) using future economic condition of Canada more precisely in advance.