# Crude Oil Price Prediction

MSA 8200 - Predictive Analytics - Final Project Report


**TEAM - 4**

| **Team Member** | **Email ID** |
|---|---|
| Manoj Velu | mvelu1@student.gsu.edu |
| Nivethitha Avarampalayam Manoharan | navarampalayammanoh1@student.gsu.edu |
| Varshini Vaisnavi Srinivasan | vsrinivasan3@student.gsu.edu |


## Project Type:

This project is an "application - flavor" where we try to predict the crude oil price in different countries.

## Problem Statement:

Develop a robust predictive model to forecast crude oil prices, leveraging historical data and relevant market indicators. The objective is to provide accurate predictions that enable stakeholders to make informed decisions regarding investments, trading strategies, and risk management in the volatile global oil market. The ultimate goal is to enhance market participants' ability to anticipate price movements.

## Goal and Motivation:

The motivation behind predicting crude oil prices in different countries lies in the economic significance of oil, which determines the health of the world economy. This volatility in the market necessitates effective risk management strategies and underscores the importance of strategic decision-making. Moreover, advancements in data analytics and predictive modeling offer potential for enhanced forecasting accuracy, further emphasizing the need for reliable predictions in this critical sector.

## Dataset

(Dataset link) – Data from Organization for Economic Co-operation and Development.

The dataset contains 8237 rows and 8 columns. We selected Canada and Germany locations for forecasting crude oil production due to Canada's upward trend and Germany's downward trend, facilitating a comparative analysis of diverse market dynamics across North America and Europe. We used location, Time, and Value columns for this project. The Columns we are interested in our project are:

| **Column Name** | **Description** |
|---|---|
| Location | 3 letter country code |
| Subject | Subject which the data is related(Oil Prod-Oil and Petroleum) |
| Measure | Unit of energy |
| Time | Year of data collected |
| Value | Value of oil production |

## Data Pre-processing:

The analysis reveals that both the Canadian and Germany datasets exhibit no null values. Moreover, no instances of dirty data or negative values were detected.

## Time Series Analysis:

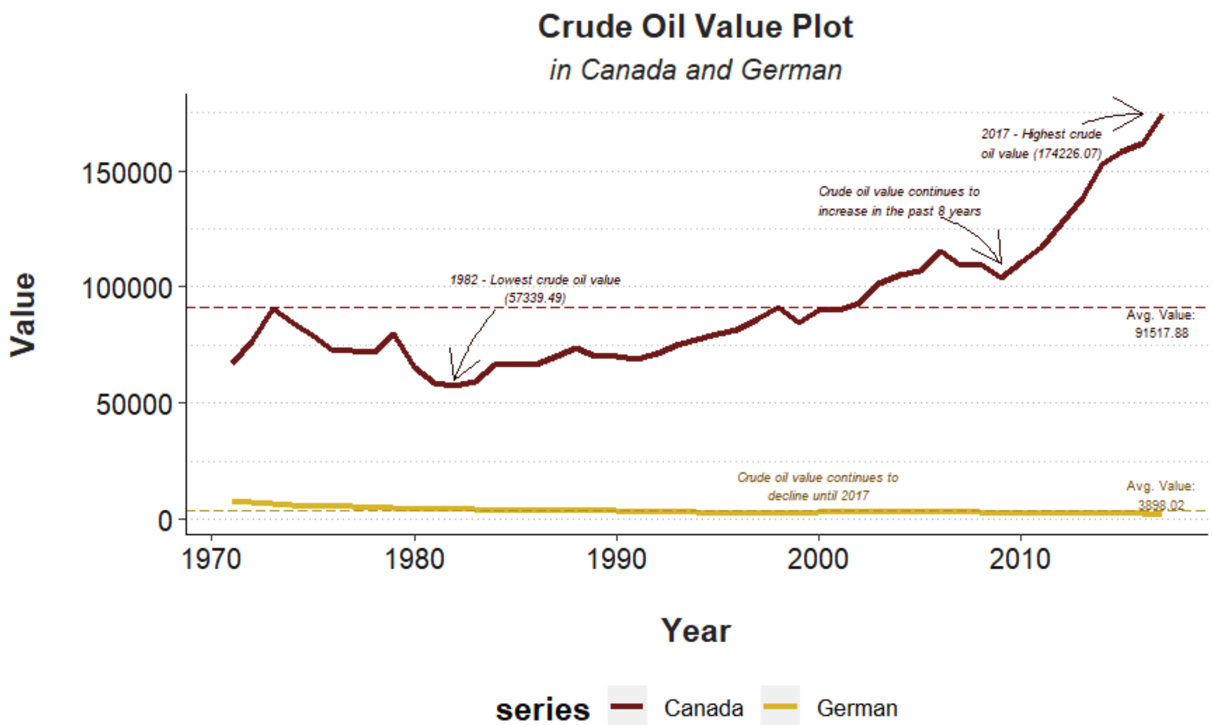
### Time Series Analysis Components:

- **Trend:** Indicates the overall tendency of data to increase or decrease over time.
- **Seasonal Variation:** Represents periodic fluctuations occurring within a year's duration.
- **Cyclical Variation:** Non-seasonal fluctuations that follow a predictable cycle.
- **Irregular/Random Variation:** Occurs due to random or irregular factors influencing the analyzed variable.

### Types of Time Series:

- **Stationary:** Characterized by constant mean, variance, and autocorrelation throughout the observation period.

- **Non-Stationary:** Exhibits changing mean, variance, and autocorrelation over the observation period.




We can observe that Canada exhibits an upward trend with no seasonality, and its time series plot is non-stationary. Conversely, Germany shows a downward trend with no seasonality, and its time series plot is stationary.

## Statistical Tests:

To prove that this time series plot is non-stationary or stationary, we used Mann-Kendall trend test, unit root test, autocorrelation function (ACF) plot.

### Mann-Kendall trend test:

**Test Hypothesis:** H0: no trend available; H1: trend available

if p-value is less than 0.05, reject H0

For Canada:

```
##
##  Spearman's rank correlation rho
##
## data:  myts_can and time(myts_can)
## S = 3334, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8072387
```

For Germany:

```
##
##  Spearman's rank correlation rho
##
## data:  myts_deu and time(myts_deu)
## S = 33178, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.918247
```

For both Canada and Germany, the p-values are below 0.05, indicating a trend is available in either series, thus leading to the rejection of the null hypothesis (H0).

## Augmented Dickey-Fuller (ADF) test:

**Test Hypothesis:** H0: not stationary; H1: stationary

if p-value is less than 0.05, reject H0

For Canada:

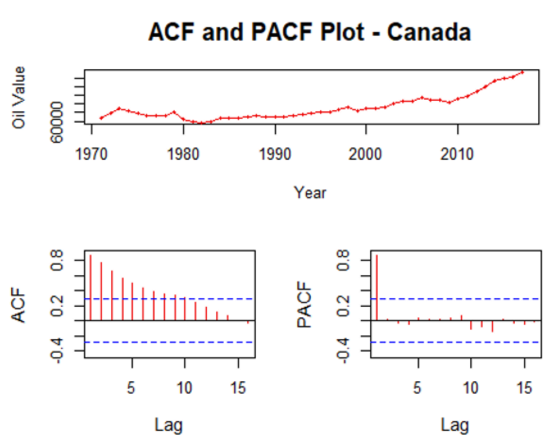
```
##
##  Augmented Dickey-Fuller Test
##
## data:  myts_can
## Dickey-Fuller = -0.031821, Lag order = 3, p-value = 0.99
## alternative hypothesis: stationary
```

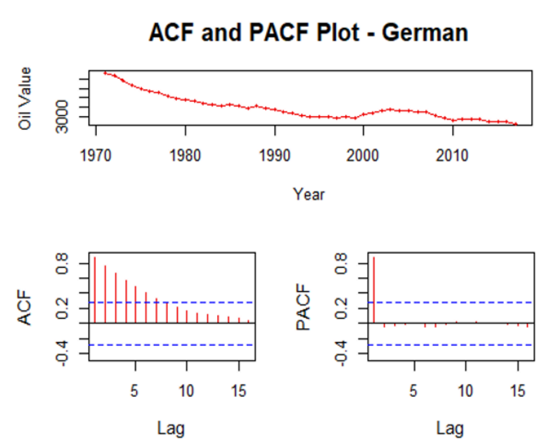
For Germany:

```
##
##  Augmented Dickey-Fuller Test
##
## data:  myts_deu
## Dickey-Fuller = -2.4953, Lag order = 3, p-value = 0.3767
## alternative hypothesis: stationary
```

For both Canada and Germany, the p-values are greater than 0.05, indicating the series is not stationary, thus leading to the acceptance of the null hypothesis (H0).

## ACF and PACF Plots:



Neither the series for Canada nor for Germany is stationary, and Autocorrelation for both countries remains significant for the first several lags and dies exceptionally slow.

# Forecasting Techniques:

## Simple Exponential Smoothing:

The Simple Exponential Smoothing (SES) model is a time series forecasting method that assigns exponentially decreasing weights to past observations. It is characterized by a single smoothing parameter, alpha ($\alpha$), which controls the rate at which the importance of past observations decreases.

For $\alpha = 0$, the forecasted values are based solely on the average of past data.

For $\alpha = 1$, the forecasted values are based solely on the most recent observation.

Implemented SES model using the ses function in R with alpha set to 0.8.Initial level estimated using simple averaging approach.

### For Canada:

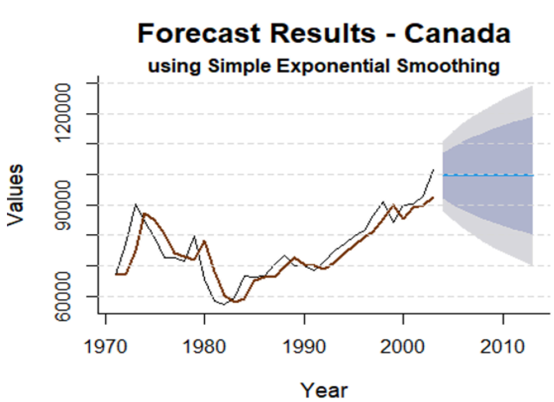
```
##
## Forecast method: Simple exponential smoothing
##
## Model Information:
## Simple exponential smoothing
##
## Call:
##  ses(y = train_can, initial = "simple", alpha = 0.8)
##
##   Smoothing parameters:
##     alpha = 0.8
##
##   Initial states:
##     l = 67007.824
```

```
##
##   sigma:  5828.136
## Error measures:
##                      ME     RMSE      MAE      MPE     MAPE     MASE
ACF1
## Training set 1236.493 5828.136 4559.323 1.147407 5.979463 1.035264
0.2532003
```

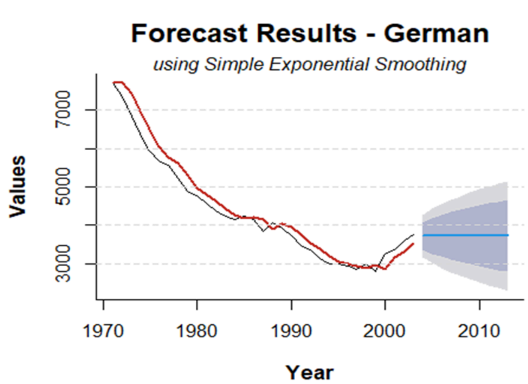
### For Germany:

```
##
## Forecast method: Simple exponential smoothing
##
## Model Information:
## Simple exponential smoothing
##
## Call:
##  ses(y = train_deu, initial = "simple", alpha = 0.8)
##
##   Smoothing parameters:
##     alpha = 0.8
##
##   Initial states:
##     l = 7723.824
##
##   sigma:  281.4892
## Error measures:
##                     ME     RMSE     MAE       MPE     MAPE     MASE
ACF1
## Training set -151.6778 281.4892 237.831 -2.988587 5.463883 1.130951
```

### For Canada:




### For Germany:



## Error, Trend, and Seasonal (ETS):

The ETS (Error, Trend, and Seasonal) model is a forecasting method that takes into account past values to predict future ones. It extends the trend observed in the historical data to make forecasts. This model incorporates three main components: Error, Trend, and Seasonal patterns.

- Error: Represents the random fluctuations or noise in the data that cannot be attributed to the trend or seasonality.
- Trend: Captures the long-term movement or directionality in the data.
- Seasonal: Accounts for recurring patterns or fluctuations that occur at regular intervals, such as daily, weekly, or yearly patterns.

The forecasts generated by the ETS model provide the median of the forecast distributions, indicating the central tendency of the predicted values. This model helps in understanding and predicting future trends based on the historical behavior of the time series data.The ETS model was implemented using the ets function in R. No specific parameters were set, allowing the model to determine optimal settings.

### For Canada:

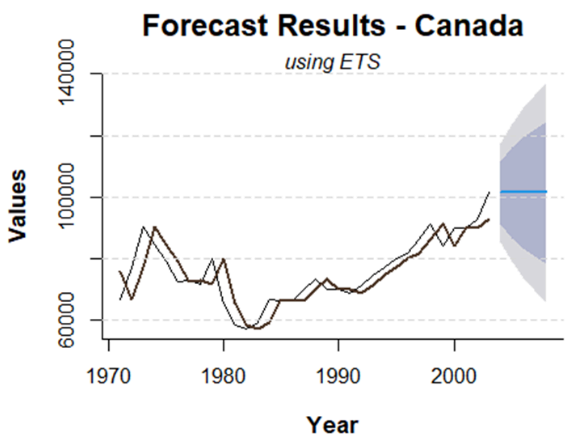
```
## .: ETS Summary :.
summary(ets_can)
## ETS(M,N,N)
##
## Call:
##  ets(y = train_can)
##
##   Smoothing parameters:
##     alpha = 0.9999
##
##   Initial states:
##     l = 76120.3458
##
##   sigma:  0.0793
##
##      AIC     AICc      BIC
## 692.6571 693.4847 697.1466
##
## Training set error measures:
##                     ME     RMSE      MAE       MPE    MAPE     MASE
ACF1
```

```
## Training set 767.3676 5827.863 4546.747 0.5738026 5.97587 1.032409
0.01688343
```

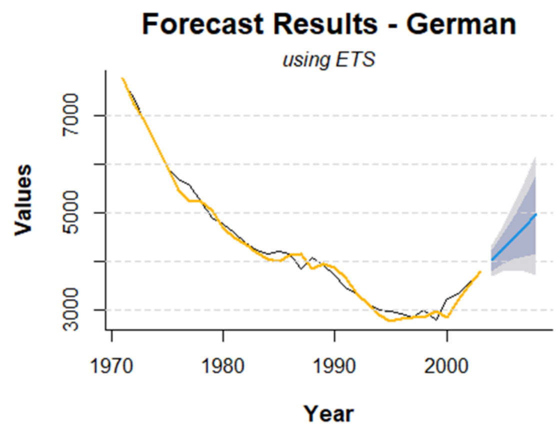
### For Germany:

```
## .: ETS Summary :.
summary(ets_deu)
## ETS(A,A,N)
##
## Call:
##  ets(y = train_deu)
##
##   Smoothing parameters:
##     alpha = 0.5541
##     beta  = 0.4904
##
##   Initial states:
##     l = 8243.654
##     b = -483.6936
##
##   sigma:  163.9537
##
##      AIC     AICc      BIC
## 457.6933 459.9156 465.1759
##
## Training set error measures:
##                     ME     RMSE      MAE      MPE     MAPE      MASE
ACF1
## Training set 44.55097 153.6963 114.3732 1.099157 2.951933 0.5438756 -
0.0209253
```

**For Canada:**




**For Germany:**




## Moving Average:

The Moving Average model is a basic forecasting technique that predicts the next period's value to be equal to the last period's actual value. It operates on the assumption that the most recent data point is the best indicator of future outcomes. However, it does not take into account trends or seasonal patterns present in the data. Thus, while simple and easy to implement, the Moving Average model may not capture more complex patterns that could affect future values.The Moving Average model was implemented using the naive function in R.The model uses a simple approach, assuming that future values will be similar to recent observed values.

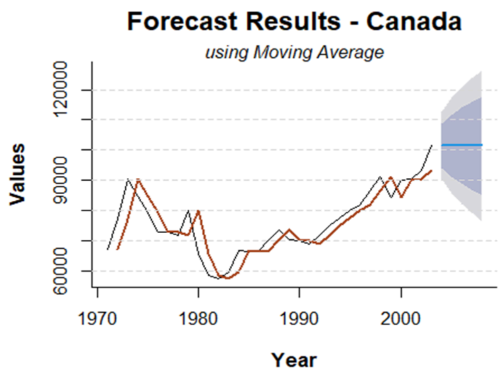
**For Canada:**

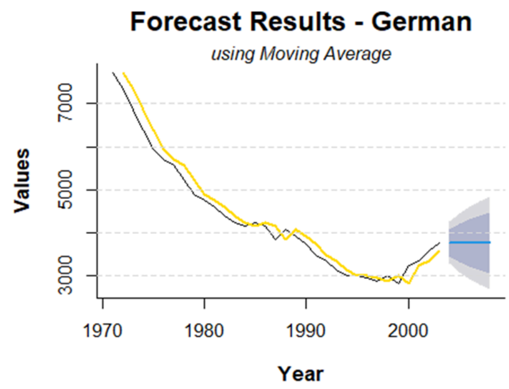
```
## .: MA Summary :.

summary(naive_can)

##
## Forecast method: Naive method
##
## Model Information:
## Call: naive(y = train_can)
##
## Residual sd: 5694.7495
##
## Error measures:
##                   ME    RMSE      MAE      MPE     MAPE MASE     ACF1
## Training set 1076.061 5694.75 4404.019 1.016679 5.737568    1 0.111976
##
```

**For Germany:**

```
## .: MA Summary :.

summary(naive_deu)

##
## Forecast method: Naive method
##
## Model Information:
## Call: naive(y = train_deu)
##
## Residual sd: 246.3645
##
## Error measures:
##                     ME     RMSE      MAE       MPE     MAPE MASE      ACF1
## Training set -123.6361 246.3645 210.2929 -2.394366 4.873583    1 0.4624619
##
```

**For Canada:** **For Germany:**




## Accuracy Comparison of SES, ETS, and MA Models:

```
Canada :

.: SES Accuracy :.
                    ME      RMSE       MAE       MPE      MAPE     MASE      ACF
Training set  1236.493  5828.136  4559.323  1.147407  5.979463 1.035264 0.253200
Test set     14716.848 18012.611 14716.848 12.210294 12.210294 3.341686 0.472409
.: ETS Accuracy :.
                    ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training Set  767.3676 5827.863 4546.747 0.5738026 5.975870 1.0324087 0.01688343
Test Set     4985.9911 7934.319 6757.898 3.4763913 5.141536 0.9325808 0.24255024

.: Moving Average Accuracy :.
                    ME     RMSE       MAE       MPE      MAPE    MASE      ACF1
Training set  1076.061  5694.75  4404.019  1.016679  5.737568 1.00000 0.1119760
Test set     12926.325 16581.87 12926.325 10.632897 10.632897 2.93512 0.4724096


German :

.: SES Accuracy :.
                    ME     RMSE     MAE        MPE      MAPE     MASE      ACF1
Training set -151.6778 281.4892 237.831  -2.988587  5.463883 1.130951 0.6027797
Test set     -672.9970 784.2853 672.997 -24.234747 24.234747 3.200283 0.7821702

.: ETS Accuracy :.

                    ME     RMSE      MAE        MPE     MAPE      MASE       ACF
Training Set  44.55097 153.6963 114.3732  1.0991570 2.951933 0.5438756 -0.020925
Test Set     -18.80508 195.9271 152.4963 -0.6469876 5.608312 1.1592404  0.377814

.: Moving Average Accuracy :.

                    ME     RMSE      MAE        MPE      MAPE     MASE      ACF1
Training set -123.6361 246.3645 210.2929  -2.394366  4.873583 1.000000 0.4624619
Test set     -720.9379 825.7905 720.9379 -25.836006 25.836006 3.428255 0.7821702
```

We could observe that for both Canada and Germany ETS model has lesser RMSE, MAE and other error measures. Hence, of these SES, ETS and MA models, ETS model is doing better and has higher accuracy.

# ARIMA Models:

The ARIMA (AutoRegressive Integrated Moving Average) model is a popular time series forecasting method that combines autoregressive (AR), differencing (I), and moving average (MA) components.

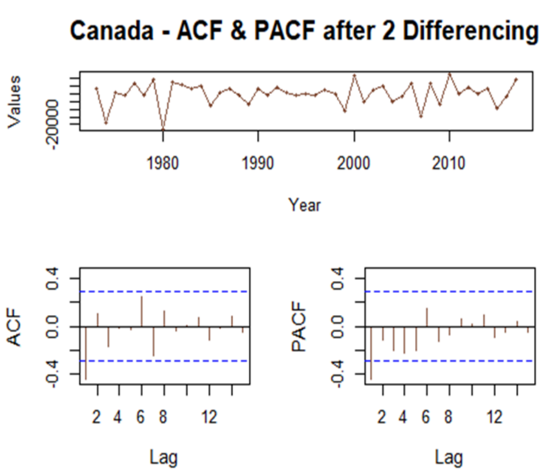
- AutoRegressive (AR): This component predicts future values based on linear regression of past observations.
- Integrated (I): This component represents the differencing of the time series data to make it stationary, removing trends and seasonality.
- Moving Average (MA): This component predicts future values based on the weighted sum of past prediction errors.

ARIMA models are versatile and can capture a wide range of time series patterns, including trends, seasonality, and autocorrelation. They are widely used in various fields for time series forecasting due to their flexibility and effectiveness in capturing complex data patterns.

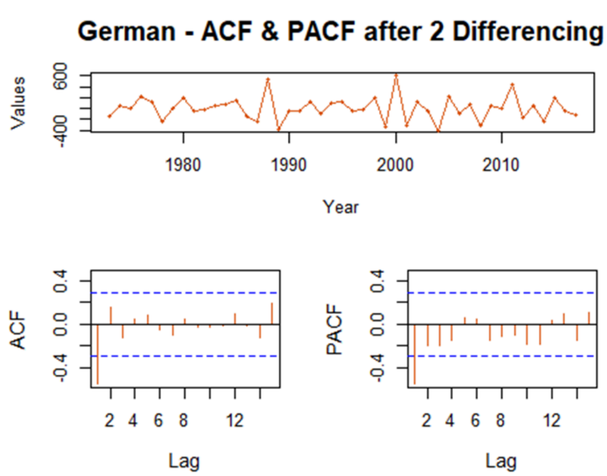
In order to get stationary time series, we are doing Non-seasonal differencing.

## ACF and PACF Plot after 2 Non-Seasonal Differencing:

### For Canada:




### For Germany:




For both Canada and Germany, the ACF and PACF show cut off after lag 1, indicating a need for differencing of 2. Therefore, ARIMA (1,2,1) models are suggested. However, for comparison, ARIMA(1,2,2) will be used for both countries.

## For Canada:

```
Canada :

.: ARIMA 1 Accuracy :.
Series: new_can
ARIMA(1,2,1)

Coefficients:
         ar1      ma1
      0.1696  -0.8882
s.e.  0.1766   0.0911

sigma^2 = 37092228:  log likelihood = -455.61
AIC=917.22   AICc=917.8   BIC=922.64

Training set error measures:
                   ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set 252.8001 5825.414 4299.228 -0.1479867 4.886964 0.8291267 0.00569091
```

```
.: ARIMA 2 Accuracy :.
Series: new_can
ARIMA(1,2,2)

Coefficients:
          ar1     ma1      ma2
      -0.8972  0.0526  -0.7435
s.e.   0.3939  0.4261   0.3774

sigma^2 = 38788111:  log likelihood = -456.06
AIC=920.12   AICc=921.12   BIC=927.35

Training set error measures:
                   ME    RMSE      MAE        MPE     MAPE     MASE      ACF1
Training set 186.7825 5887.42 4415.271 -0.2168723 5.035719 0.851506 0.1556451
```

## For Germany:

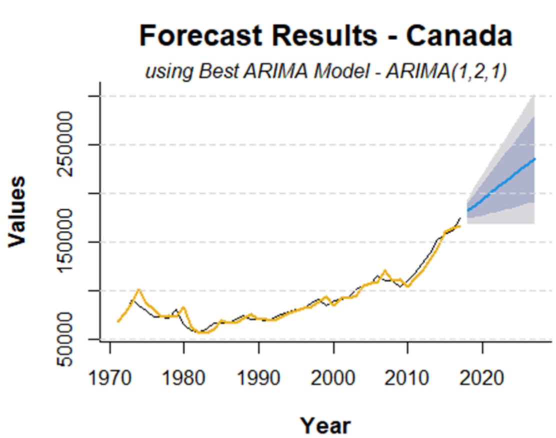
```
>> German :
.: ARIMA 1 Accuracy :.Series: new_ger
ARIMA(1,2,1)

Coefficients:
          ar1      ma1
      -0.1665  -0.5389
s.e.   0.2585   0.2405

sigma^2 = 31510:  log likelihood = -296.16
AIC=598.31   AICc=598.9   BIC=603.73

Training set error measures:
                   ME   RMSE      MAE       MPE     MAPE     MASE        ACF1
Training set 15.60203 169.79 125.0024 0.3461047 3.630782 0.664504 -0.01252289
```

```
.: ARIMA 2 Accuracy :.Series: new_ger
ARIMA(1,2,2)

Coefficients:
          ar1      ma1      ma2
      -0.5923  -0.1146  -0.2854
s.e.   0.6412   0.6698   0.4881

sigma^2 = 32077:  log likelihood = -296.04
AIC=600.07   AICc=601.07   BIC=607.3

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE     MASE        ACF1
Training set 16.40772 169.3058 125.0218 0.3611468 3.621868 0.664607 -0.01074029
```

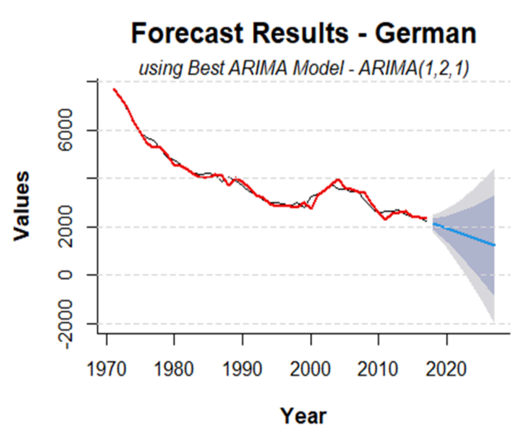
We observe that for both countries ME, RMSE, MAE and other error measures are lesser for ARIMA(1,2,1) model compared to ARIMA(1,2,2) model. Also, the AIC, AICc, and BIC values are lower for ARIMA(1,2,1). Hence, we conclude that ARIMA(1,2,1) is the best model.

## ARIMA(1,2,1) Model:

### For Canada:




### For Germany:



## SARIMA:

The SARIMA (Seasonal AutoRegressive Integrated Moving Average) model is an extension of the ARIMA model that specifically accounts for seasonality in time series data.

- Seasonal (S): This component captures periodic patterns that repeat over fixed intervals, such as daily, weekly, or yearly seasonality.
- AutoRegressive (AR): Similar to ARIMA, this component models the relationship between an observation and a number of lagged observations in the time series.
- Integrated (I): This component deals with differencing the time series to make it stationary, removing trends and seasonality.
- Moving Average (MA): Like ARIMA, this component models the relationship between an observation and a residual error term from a moving average model applied to lagged observations.

SARIMA models are particularly useful for time series data with predictable seasonal patterns. They offer improved forecasting accuracy by explicitly modeling and incorporating these seasonal fluctuations into the forecasting process.

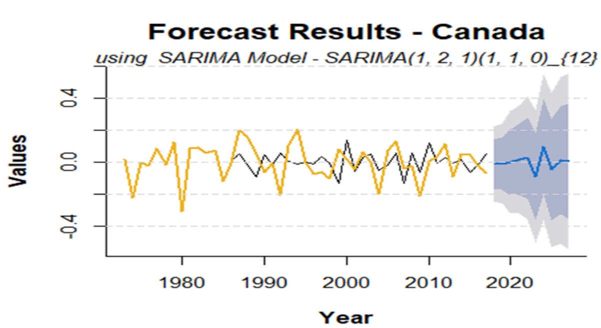
### For Canada:

```
Canada :
.: SARIMA Canada Accuracy :.
Call:
arima(x = difflog_can, order = c(1, 2, 1), seasonal = list(order = c(1, 1, 0),
    period = 12))

Coefficients:
          ar1      ma1     sar1
      -0.7267  -0.9937  -0.7039
s.e.   0.1297   0.1128   0.1274

sigma^2 estimated as 0.01386:  log likelihood = 12.77,  aic = -17.54

Training set error measures:
                      ME       RMSE        MAE       MPE     MAPE      MASE       ACF1
Training set -0.00137748 0.09770008 0.06926844 -39.91396 368.4274 0.6422468 0.02740034
```

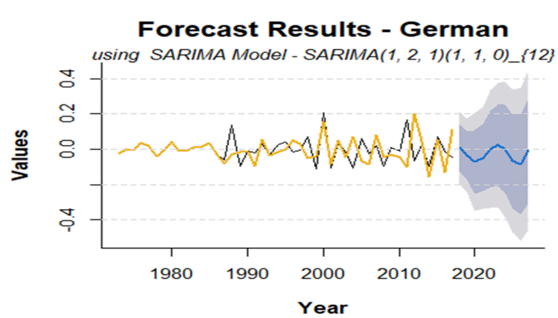
### For Germany:

```
German :

.: SARIMA German Accuracy :.
Call:
arima(x = difflog_deu, order = c(1, 2, 1), seasonal = list(order = c(1, 1, 0),
    period = 12))

Coefficients:
          ar1      ma1     sar1
      -0.7905  -0.9956  -0.3429
s.e.   0.1031   0.0950   0.2234

sigma^2 estimated as 0.01036:  log likelihood = 20.51,  aic = -33.02

Training set error measures:
                     ME       RMSE        MAE      MPE     MAPE      MASE       ACF1
Training set 0.00450028 0.08448714 0.05180688 24.36487 137.0602 0.6017318 -0.3908892
```

### For Canada:

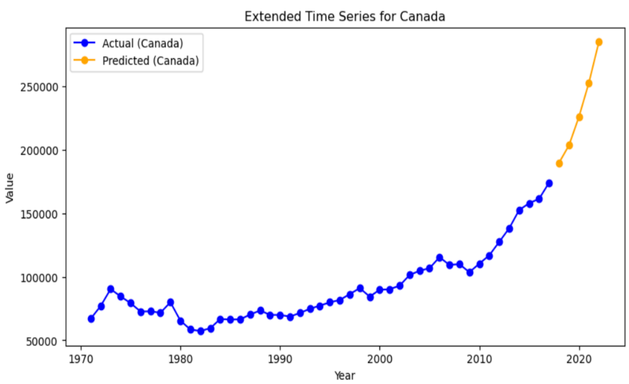


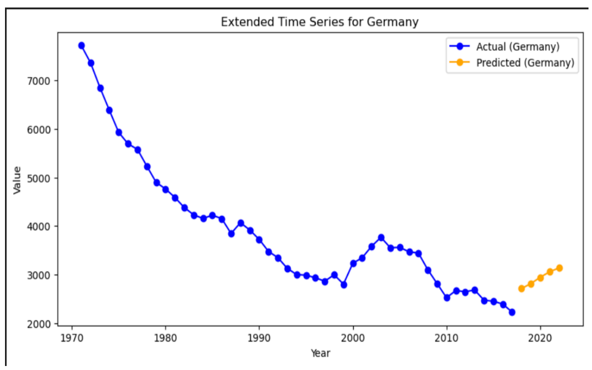

### For Germany:



## LSTM:

LSTM (Long Short-Term Memory) is a type of deep learning model used for time series forecasting. It's effective because it can capture long-term dependencies, nonlinear patterns, and handle variable-length sequences. LSTM models are trained on historical data to make predictions for future time steps. They have shown success in various forecasting tasks, including stock prices and weather predictions.We used the last 3 time steps to make predictions, and our final training loss was 0.0177.




## Conclusion:

Among the models we've implemented, ETS generally outperforms SES and Moving Average. However, SARIMA and LSTM models show even better performance with lower losses. It's worth noting that LSTMs usually need a larger dataset for training compared to simpler models like SARIMA. Given our limited data, an LSTM might not be the most suitable option.

## Group Members' Contributions:

### Manoj Velu:

Led the data preparation efforts, ensuring our datasets were clean and ready for analysis. He conducted statistical tests to validate our findings and provided valuable insights during the time-series analysis. His expertise was instrumental in implementing the LSTM model and analyzing its performance. Additionally, Manoj played a key role in evaluating various models, offering insightful analysis on their suitability for our forecasting tasks.

### Nivethitha Avarampalayam Manoharan:

Focused on time-series analysis, extracting meaningful insights from our data. She specialized in implementing the ARIMA model,a powerful tool for time-series forecasting and also worked on SARIMA, which extends ARIMA to handle seasonal data.

Her time-series analysis greatly contributed to our understanding of the underlying patterns and trends in the data.

**Varshini Vaisnavi Srinivasan:**

Analyzed time-series data and uncovering important trends and patterns. She implemented the Simple Exponential Smoothing (SES) model, providing a foundational method for our forecasting efforts. She also contributed to implementing the ETS model, which considers error, trend, and seasonal components in forecasting. Her work on the Moving Average (MA) model added another dimension to our forecasting toolkit, offering a simple yet effective approach to prediction.

## Software & Source Code:

**Software Used:** R Studio, jupyter notebook

**Source Code:** **https://github.com/Rambo1806/Predictive-Analytics-Final-Project**

**Dataset:** Dataset link