Report

In this report we aim to perform a thorough time series analysis on the datasets:

- Electric Production
- Johnson & Johnson.

The analysis involves the implementation of various models, including exponential smoothing, AR, MA, ARMA, ARIMA, SARIMA, and LSTM on the datasets.

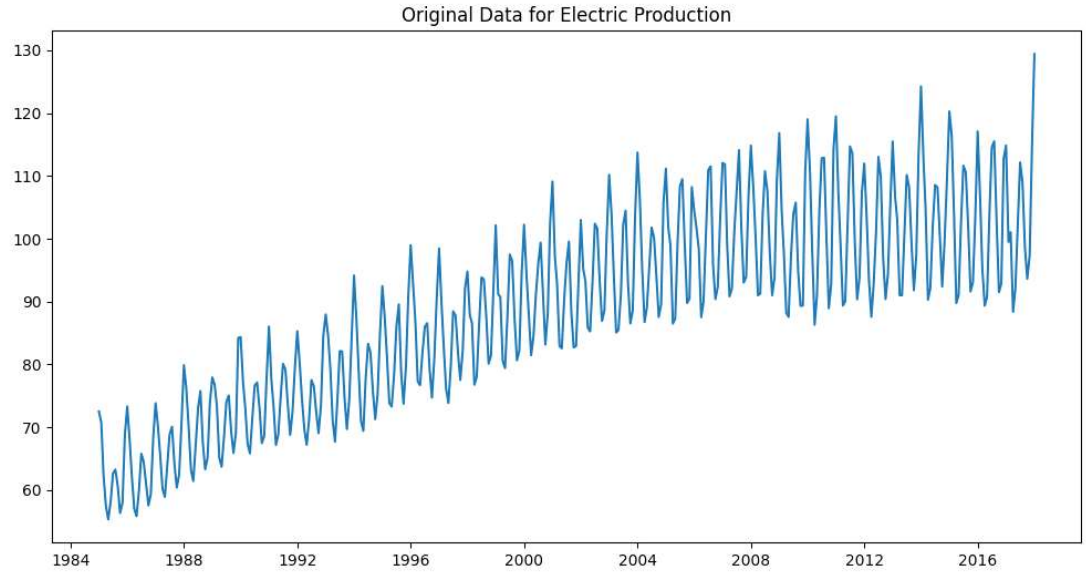
This analysis covers stationarity testing, assessing seasonality and Non stationarity, making predictions for subsequent years, evaluating model performance, plotting forecasted versus original values, comparing forecasts, and providing an out-of-sample forecast for the next four years.

Introduction to the Electric Production dataset:

The electric production dataset represents monthly electric production data spanning from January 1985 to January 2018. It comprises of monthly observations, capturing fluctuations in electric production over a time. Below is a snapshot of the initial data:

DATE	IPG2211A2N
1985-01-01	72.5052
1985-02-01	70.6720
1985-03-01	62.4502
1985-04-01	57.4714
1985-05-01	55.3151
...	...
2017-09-01	98.6154
2017-10-01	93.6137
2017-11-01	97.3359
2017-12-01	114.7212
2018-01-01	129.4048

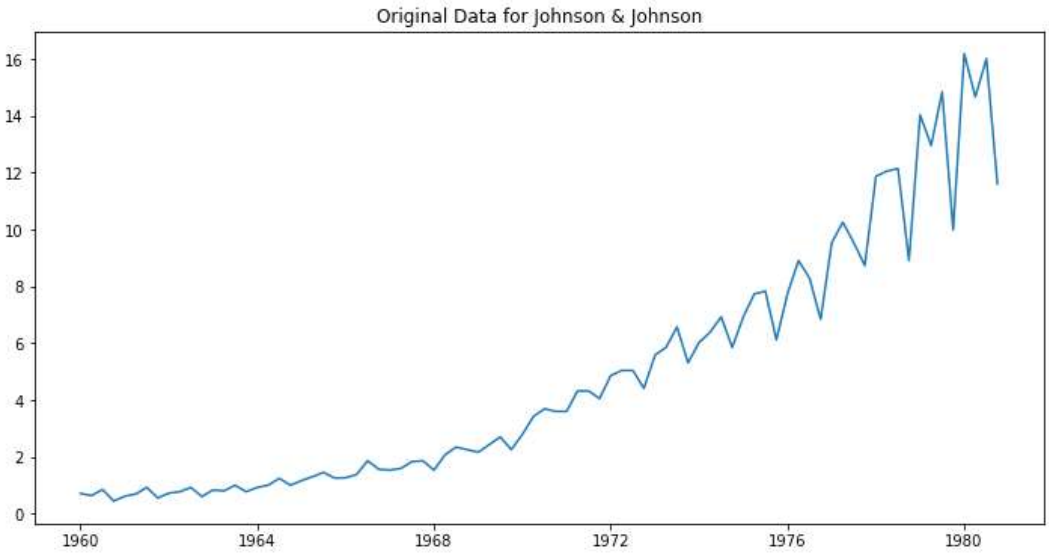
[397 rows x 1 columns]



Intro to the Johnson & Johnson (JJ) dataset:

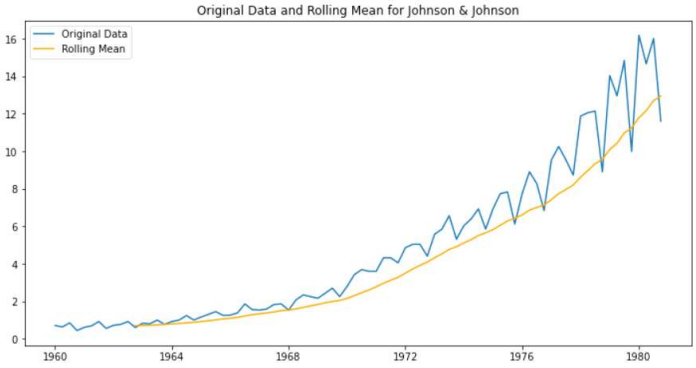
The Johnson & Johnson Quarterly Earnings dataset spans from 1960 to 1980, providing a chronological record of quarterly earnings. With a quarterly frequency, the dataset consists of 84 observations, capturing vital financial information during each quarter.

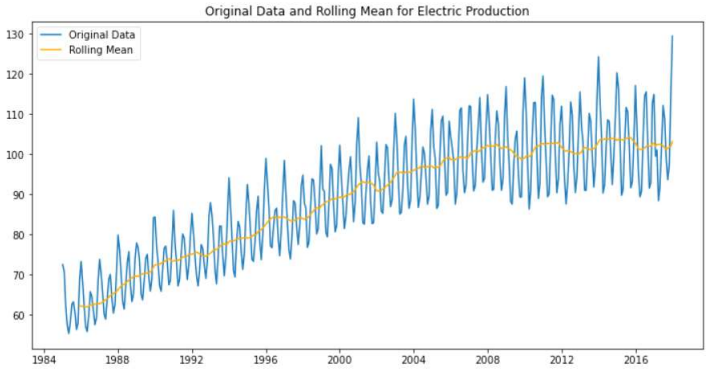
```
         date   data
0  1960-01-01   0.71
1  1960-04-01   0.63
2  1960-07-02   0.85
3  1960-10-01   0.44
4  1961-01-01   0.61
..        ...    ...
79 1979-10-01   9.99
80 1980-01-01  16.20
81 1980-04-01  14.67
82 1980-07-02  16.02
83 1980-10-01  11.61

[84 rows x 2 columns]
```



Stationarity Testing:

Method 1) Assessing stationarity through mean:

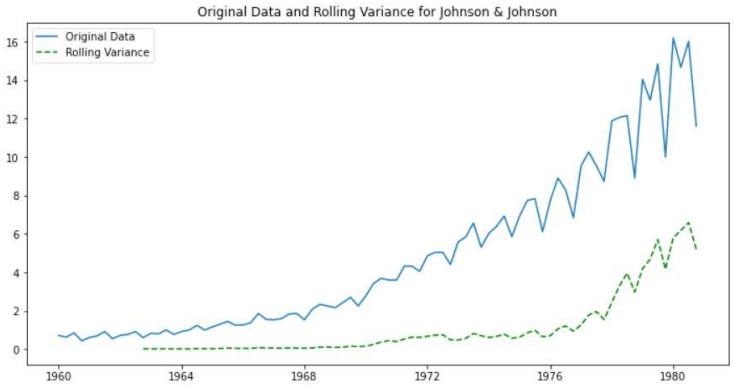


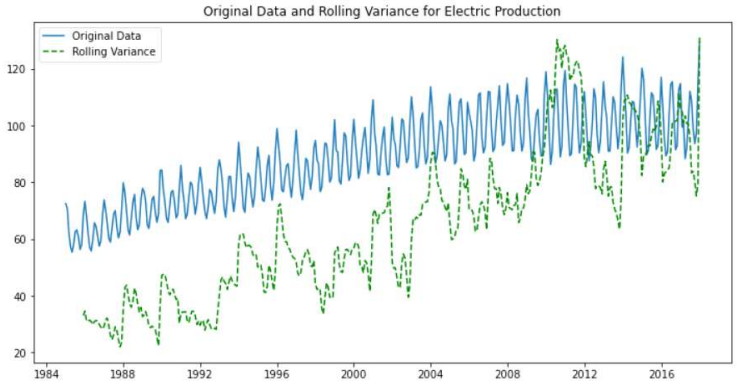


By looking at the plot of original data and mean and we can clearly tell that mean/trend of both data sets are increasing, this tells us that the data is not stationarity as constant mean is

required for stationarity.

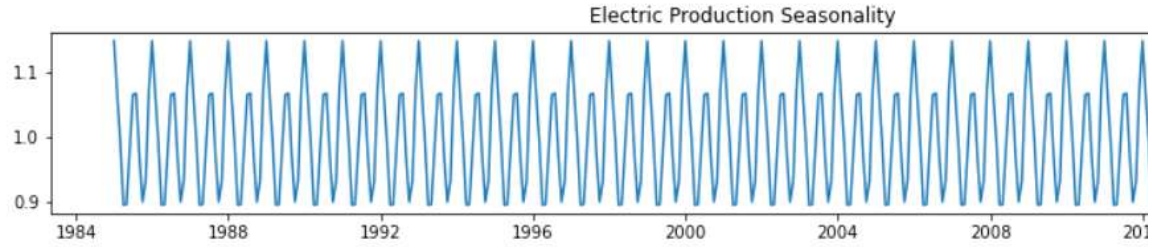
Method 2) We check for constant variance as it indicates stationarity:

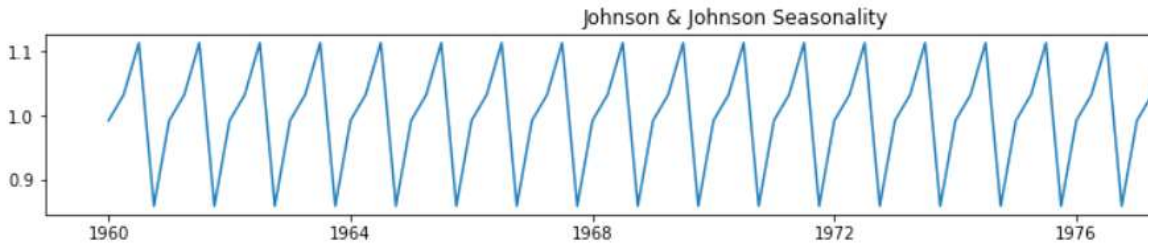




As we can see for both the time series, variance is not constant, so the data is not stationary.

Method 3) By performing seasonal decomposition we get the seasonal component of the time series, and if we see clear signs of seasonality, the time series is not stationary:





Method 4) ADF Test:

The Augmented Dickey-Fuller (ADF) test is a statistical test used in econometrics and time series analysis to assess whether a given time series is stationary.

The ADF test specifically focuses on detecting the presence of a unit root in a time series. A unit root is a feature of non-stationary time series data where the process has a root equal to 1. The presence of a unit root suggests that the time series is non-stationary.

The ADF test involves estimating the parameters of a regression equation with a lagged dependent variable and then performing statistical tests to determine whether the coefficient on the lagged variable is significantly different from zero. The null hypothesis of the test is that a unit root is present, indicating non-stationarity. Rejection of the null hypothesis suggests stationarity.

Johnson & Johnson:

```
ADF Statistic: 2.7420165734574735
p-value: 1.0
Critical Values: {'1%': -3.524624466842421, '5%': -2.9026070739026064, '10%': -2.5886785262345677}
Johnson & Johnson is Non-stationary (fail to reject the null hypothesis)
```

Electrical production:

```
ADF Statistic: -2.25699035004725
p-value: 0.18621469116586592
Critical Values: {'1%': -3.4476305904172904, '5%': -2.869155980820355, '10%': -2.570827146203181}
Electric Production is Non-stationary (fail to reject the null hypothesis)
```

Critical Values: These are thresholds. If the calculated test statistic is more extreme than the critical value, or if the p-value is below a chosen significance level (e.g., 0.05), you reject the null hypothesis in favor of stationarity.

P-Value: If p-value is low (typically < 0.05), you reject the null hypothesis; if high, you fail to reject.

Thus, from these different tests we conclude that the time series are not stationary.

Making our time-series Stationary:

Before we can use models like Exponential Smoothing, AR, MA or ARMA, we must first make the time-series stationary.

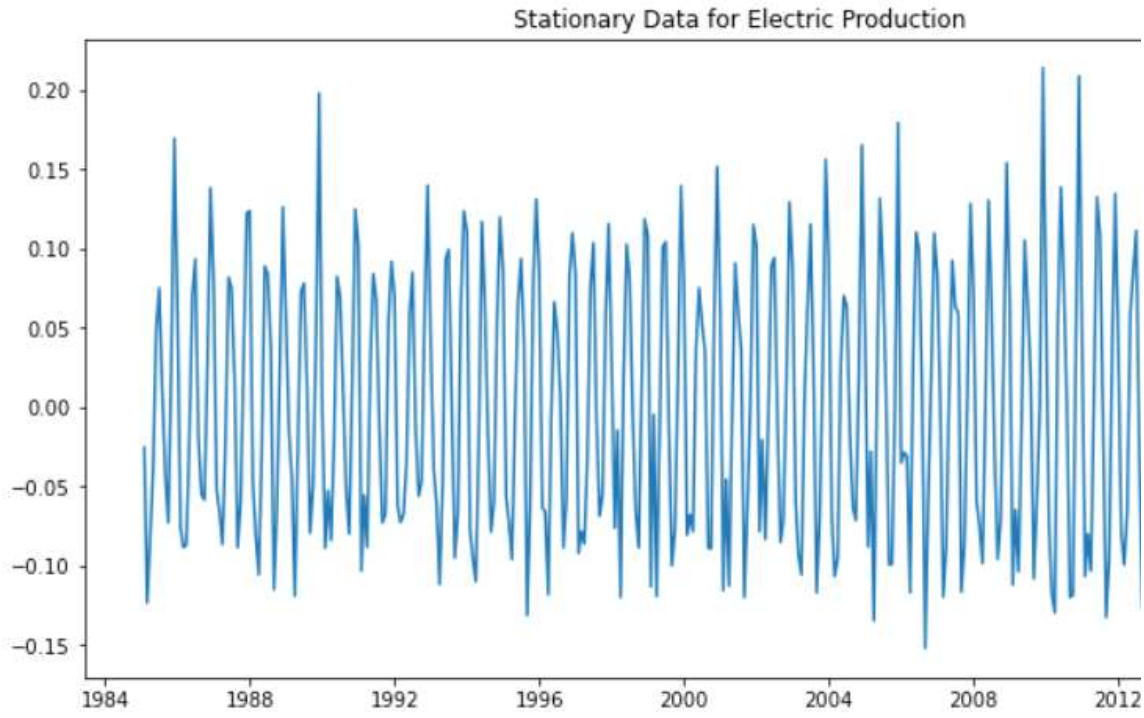
Why do we do this?

Stationary time series simplify analysis, stabilize statistical properties, and improve forecasting accuracy while non-stationary data often has trends or seasonality. Converting it removes these, enhances model performance, enables accurate testing, and avoids spurious correlations.

We use Log differencing to make the time series stationary.

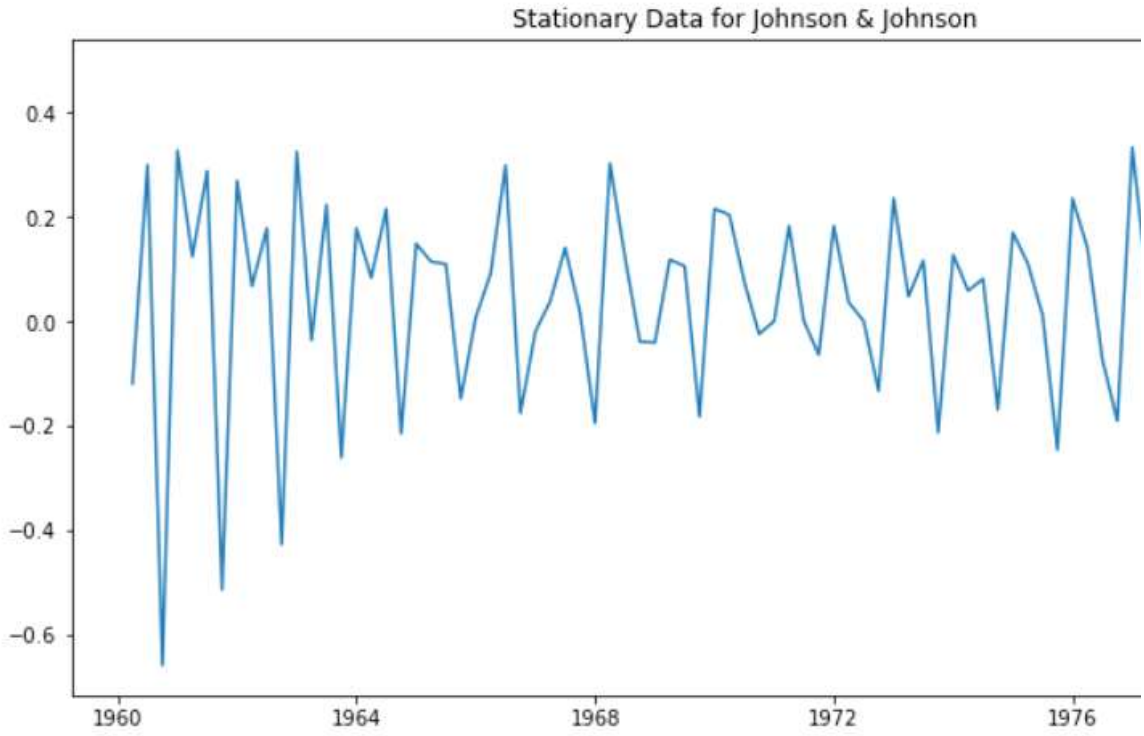
In Log differencing we take the natural logarithm of a time series and then differencing the resulting series. This method is often used to stabilize the variance and make the data more stationary.

Electric Production after log-diff:




```
ADF Statistic: -6.748333370019134
p-value: 2.9951614981160046e-09
Critical Values: {'1%': -3.4476305904172904, '5%': -2.869155980820355, '10%':
Electric Production is Stationary (reject the null hypothesis)
```

Johnson & Johnson after log-diff:




```
ADF Statistic: -4.317043945811843
p-value: 0.00041497314044405543
Critical Values: {'1%': -3.518281134660583, '5%': -2.899878185191432, '10%': -
Johnson & Johnson is Stationary (reject the null hypothesis)
```

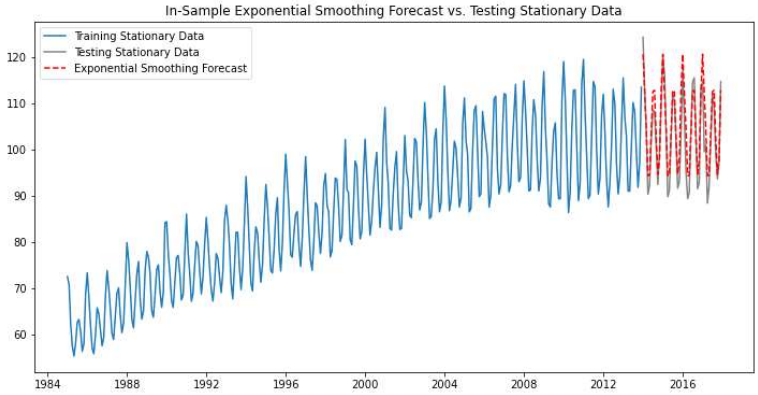
Both are stationary after log diff.

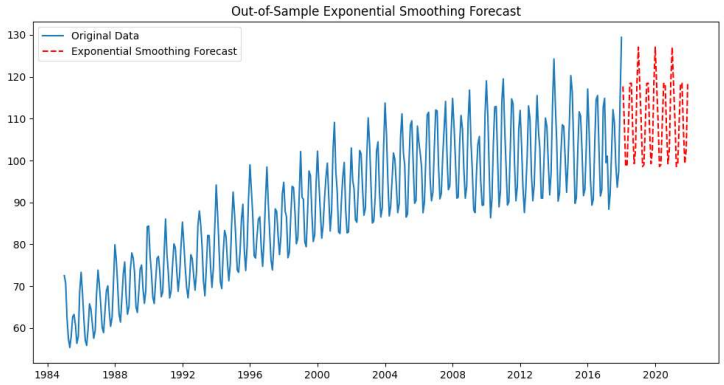
Forecasting (For the sake of brevity we will be showing only a few plots here):

Exponential Smoothing Model:

Exponential smoothing is a forecasting method that gives more weight to recent data, with exponentially decreasing weights as observations age. It's effective for time series data with trends or seasonality.

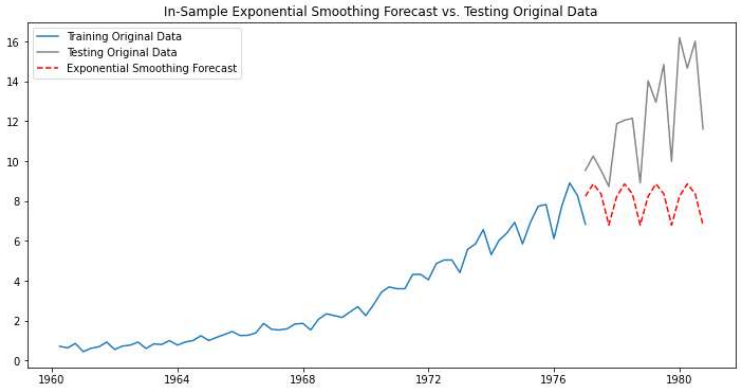
Electric Production:

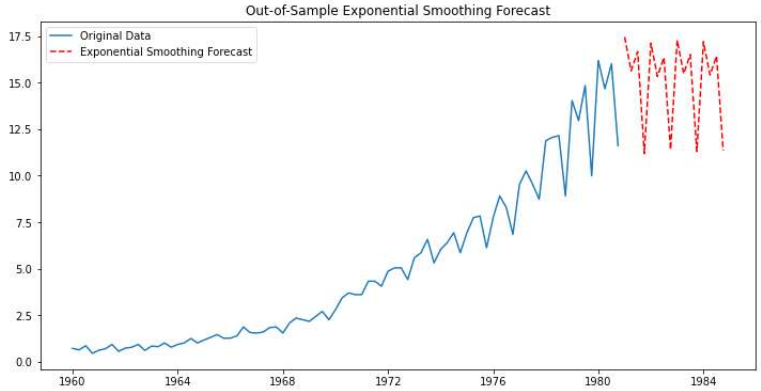





```
RMSE (Exponentional Smoothing): 4.012466567379388
R-squared (Exponentional Smoothing): 0.8217665517603081
MSE (Exponentional Smoothing): 16.09988795433733
MAPE (Exponentional Smoothing): 0.030820873086746373%
```

Johnson & Johnson:





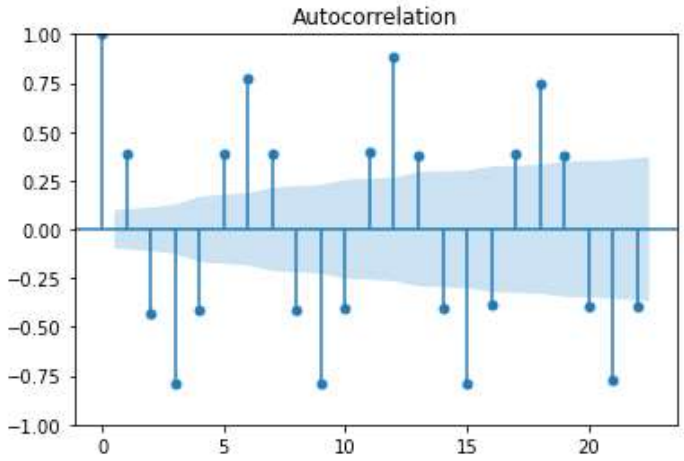

```
RMSE (Exponentional Smoothing): 4.560663603670137
R-squared (Exponentional Smoothing): -2.5549169439378763
MSE (Exponentional Smoothing): 20.79965250584148
MAPE (Exponentional Smoothing): 0.3128292376702754%
```

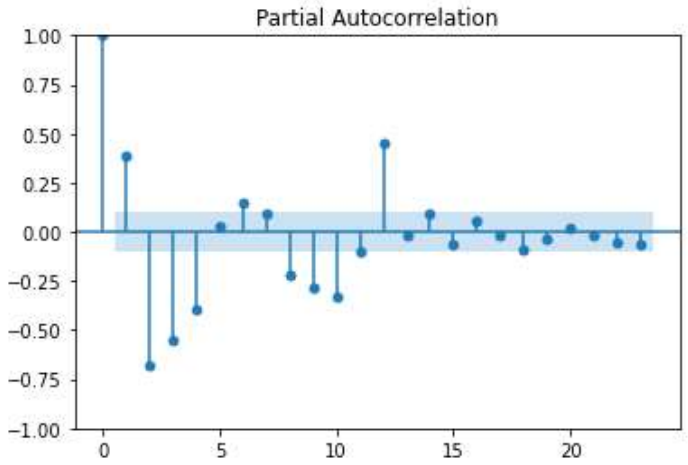
Choosing p and q with ACF and PACF:

An ACF measures and plots the average correlation between data points in time series and previous values of the series measured for different lag lengths.

A PACF is similar to ACF except that each partial correlation controls for any correlation between observations of a shorter lag length.

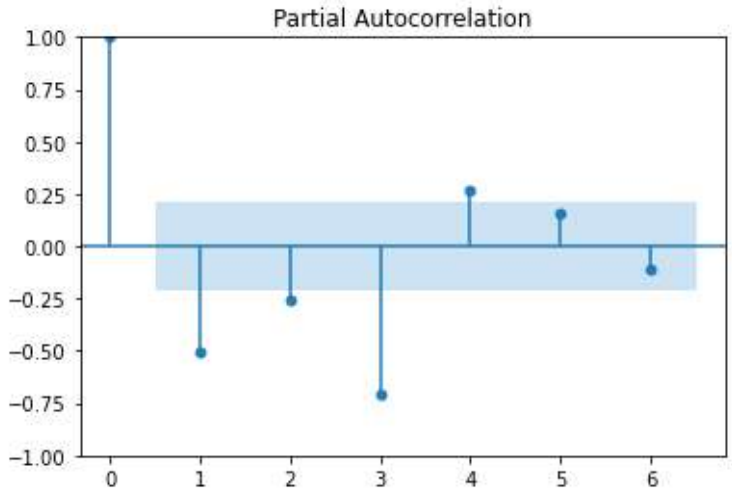
Electric Production:

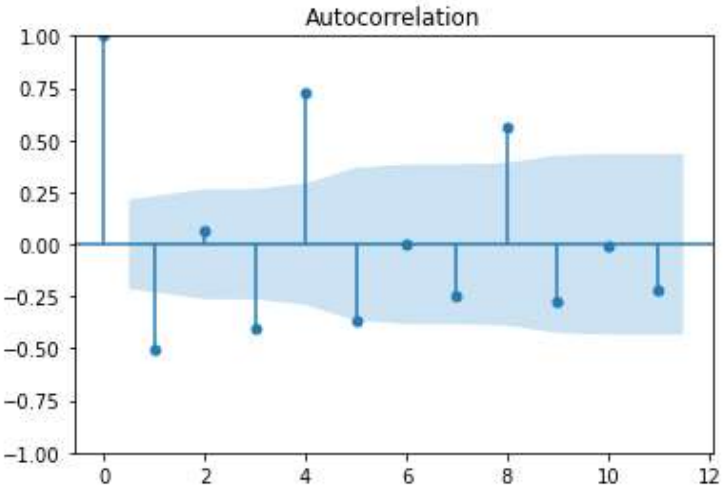




When figuring out how many past values should influence our current one (p) for an AR model and for the MA model how many past errors should affect the current value (q), we checked the PACF and ACF plots. The PACF plot showed the last significant value at lag 23, so we chose p=23. For the ACF plot, the last significant value was at lag 22, so we chose q=22. These numbers, 23 for AR and 22 for MA, are what we're picking based on what stood out in the plots.

Johnson & Johnson:



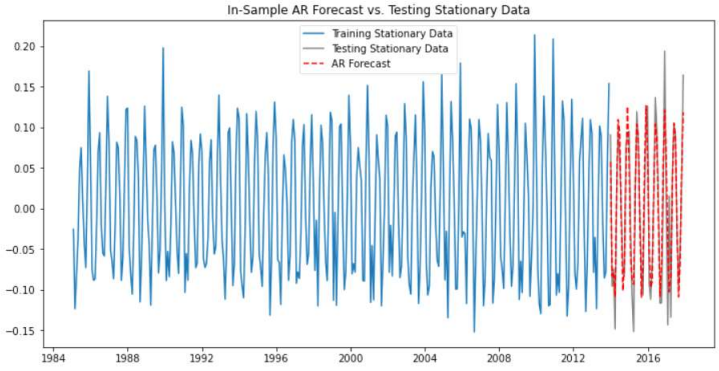


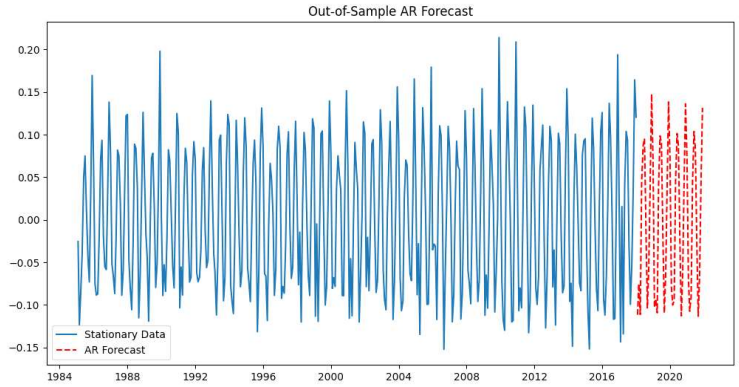
The PACF plot showed the last significant value at lag 6, so we chose p=6. For the ACF plot, the last significant value was at lag 11, so we chose q=11.

AR Model:

An AR (AutoRegressive) model is a type of time series model that describes the relationship between a time series and its past values.

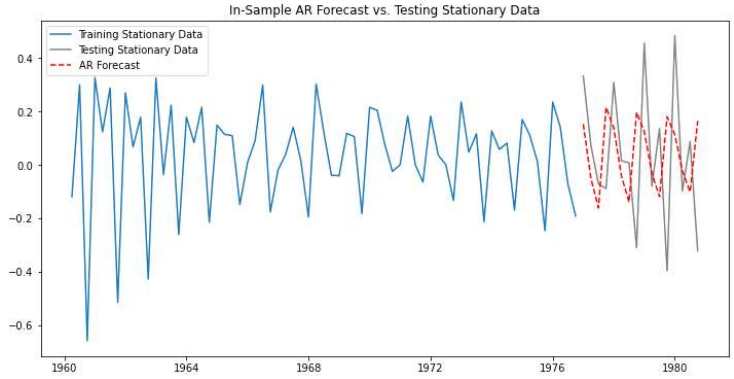
Electric Production:

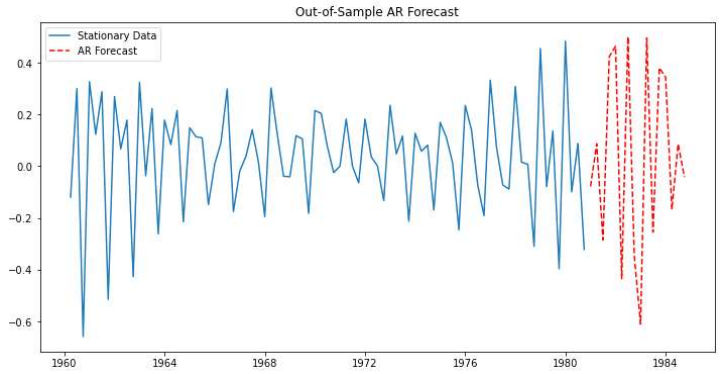





```
Optimal lag order (p) based on PACF: 23
RMSE (AR): 0.036761266979860645
R-squared (AR): 0.8452573595148535
MSE (AR): 0.0013513907499645924
MAPE (AR): 0.7015495563203937%
```

Johnson & Johnson:





```
Optimal lag order (p) based on PACF: 6
RMSE (AR): 0.2938320891004463
R-squared (AR): -0.3195325851019717
MSE (AR): 0.0863372965851326
MAPE (AR): 2.649977480667639%
```

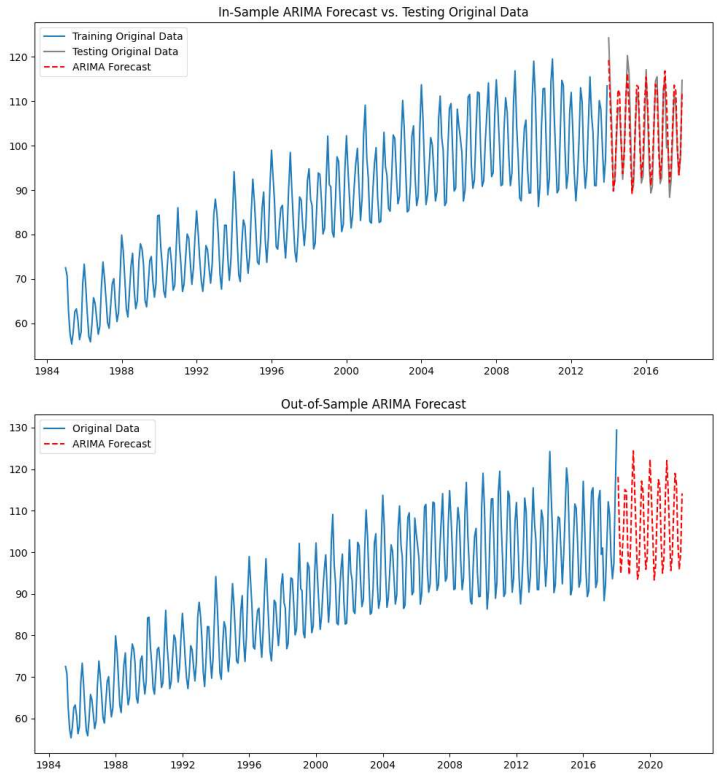
ARIMA Model:

ARIMA is a powerful time series forecasting model that integrates components from AutoRegressive (AR), Integrated (I), and Moving Average (MA) models. It is designed to capture a wide range of time series patterns, including trends and seasonality. ARIMA is particularly effective when dealing with non-stationary time series data.

The Integrated component represents the differencing of the time series to achieve stationarity.

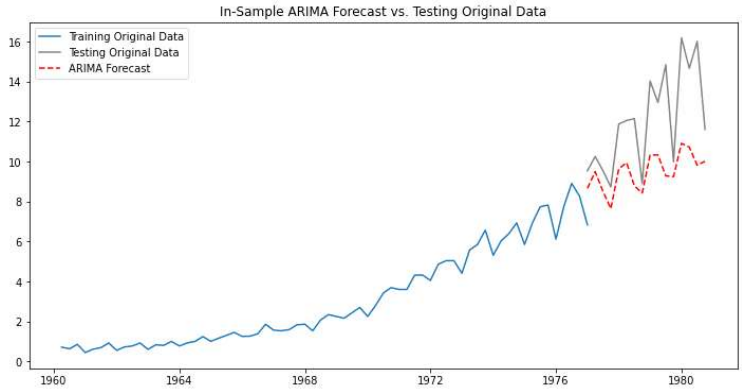
The 'd' parameter denotes the number of differencing operations applied to make the time series stationary.

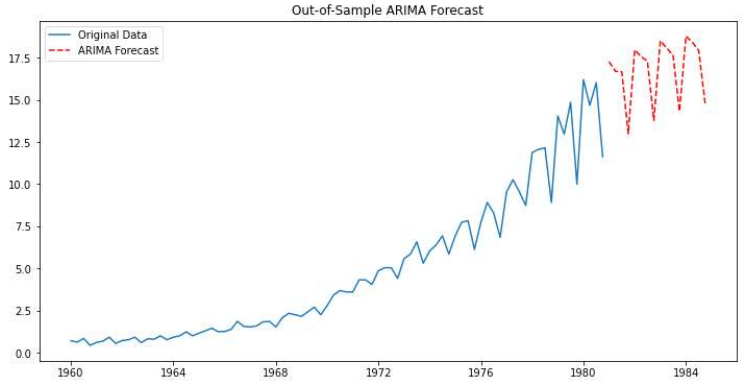
Electric Production:



```
Optimal lag order (p, d, q) based on PACF, ADF Test and ACF: (23, 1, 22)
RMSE (ARIMA): 3.3732849255064594
R-squared (ARIMA): 0.8740284692166124
MSE (ARIMA): 11.37905118864912
MAPE (ARIMA): 0.025418587210538898%
```

Johnson & Johnson:



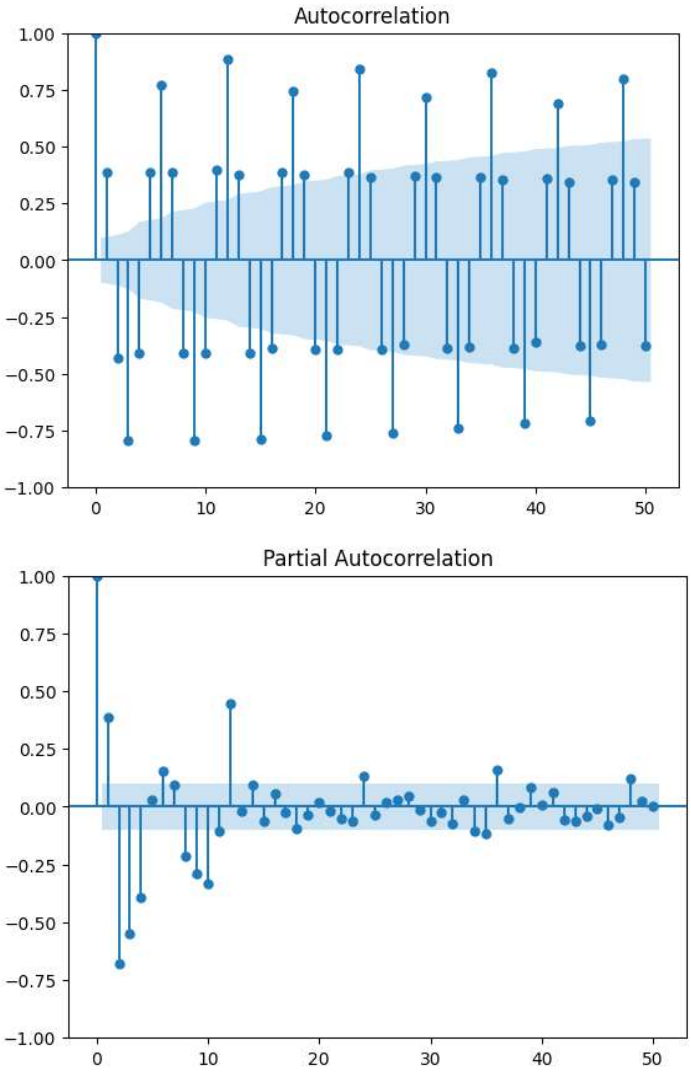


SARIMA Model:

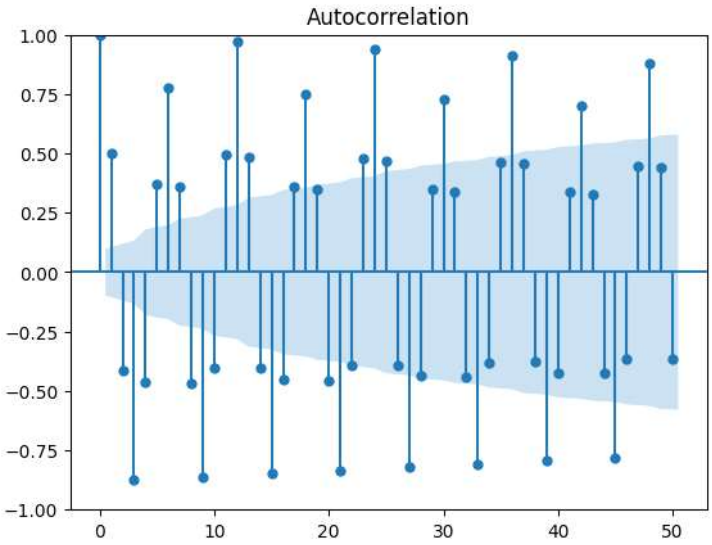
SARIMA (Seasonal AutoRegressive Integrated Moving Average) is an extension of the ARIMA model that incorporates seasonality. It's particularly useful for time series data that exhibit periodic patterns.

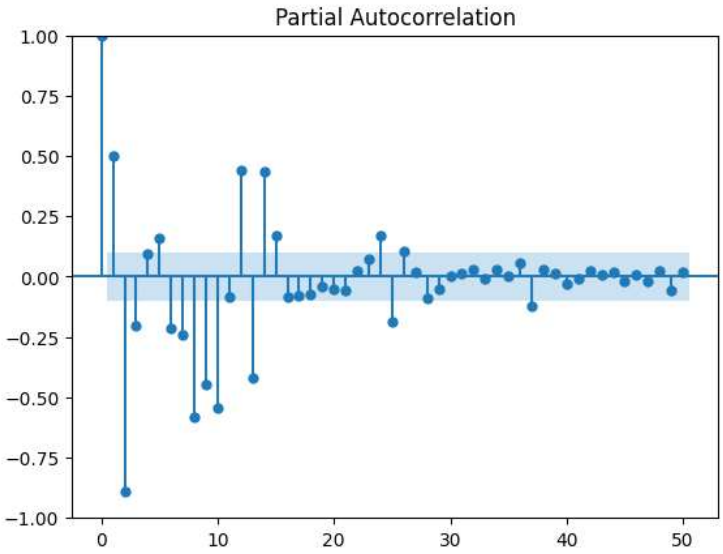
Choosing p and q with ACF and PACF and P and Q with Seasonal ACF and PACF:

For Electric Production stationary data:

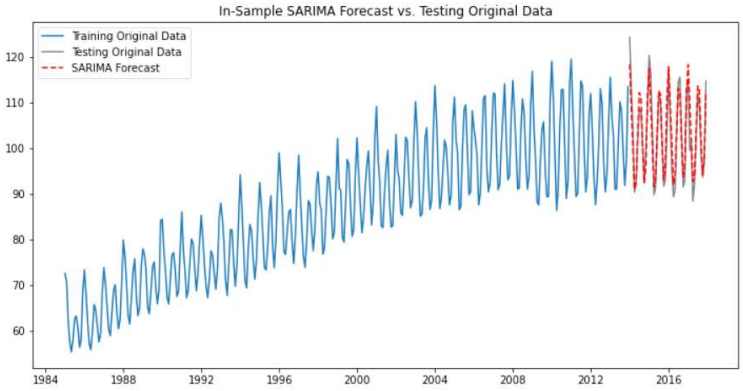


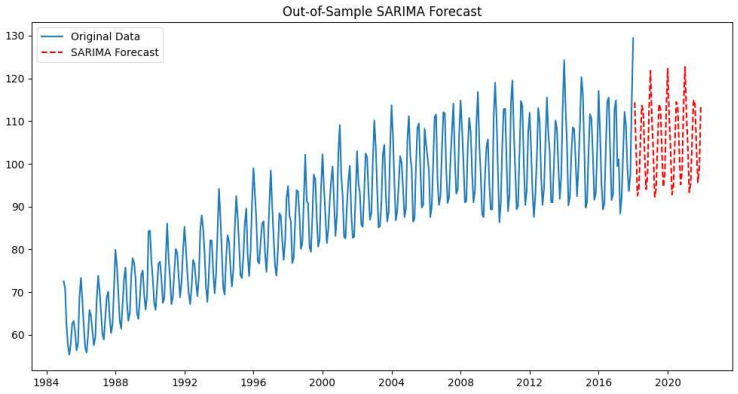
For Electric Production seasonal component:



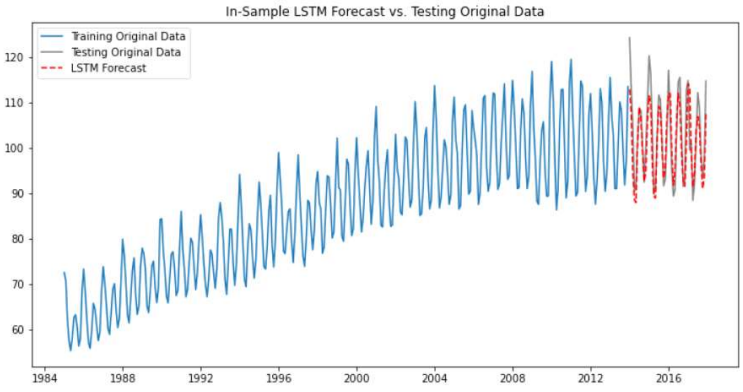


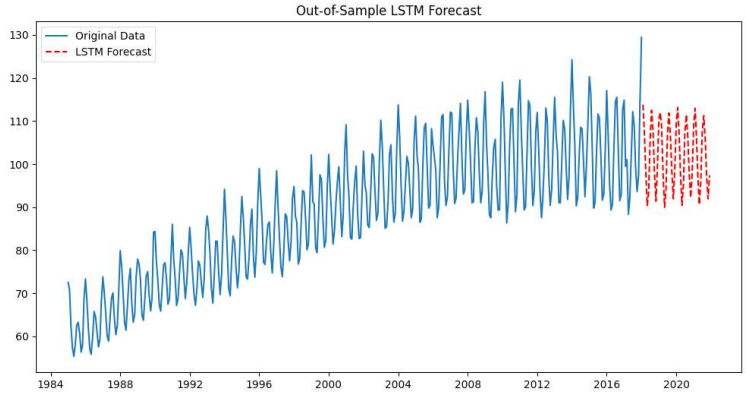
P, Q, p and q were picked based on trial and error.


In-Sample SARIMA Forecast vs. Testing Original Data
Training Original Data
Testing Original Data
SARIMA Forecast
120
110
100
90
80
70
60
1984
1988
1992
1996
2000
2004
2008
2012
2016


Out-of-Sample SARIMA Forecast
Original Data
SARIMA Forecast
130
120
110
100
90
80
70
60
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020

LSTM Model:


In-Sample LSTM Forecast vs. Testing Original Data
Training Original Data
Testing Original Data
LSTM Forecast
120
110
100
90
80
70
60
1984
1988
1992
1996
2000
2004
2008
2012
2016



Conclusion:

To conclude, we will compare the forecasting performance of various models based on both visualization and accuracy metrics.

1)AR model

- **RMSE:** 0.0367
- **R-squared:** 0.8454
- **MSE:** 0.0014
- **MAPE:** 0.7023%

2)MA model:

- **RMSE:** 0.0805
- **R-squared:** 0.2572
- **MSE:** 0.0065
- **MAPE:** 0.9493%

3)ARMA model:

- **RMSE:** 0.0377
- **R-squared:** 0.8372
- **MSE:** 0.0014
- **MAPE:** 0.7133%

4)Exponential Smoothing Model:

- **RMSE:** 4.0125
- **R-squared:** 0.8218
- **MSE:** 16.0999
- **MAPE:** 0.0308%

5)ARIMA model:

- **Optimal lag order (p, d, q):** (23, 1, 22)

- **RMSE:** 3.3733
- **R-squared:** 0.8740
- **MSE:** 11.3791
- **MAPE:** 0.0254%

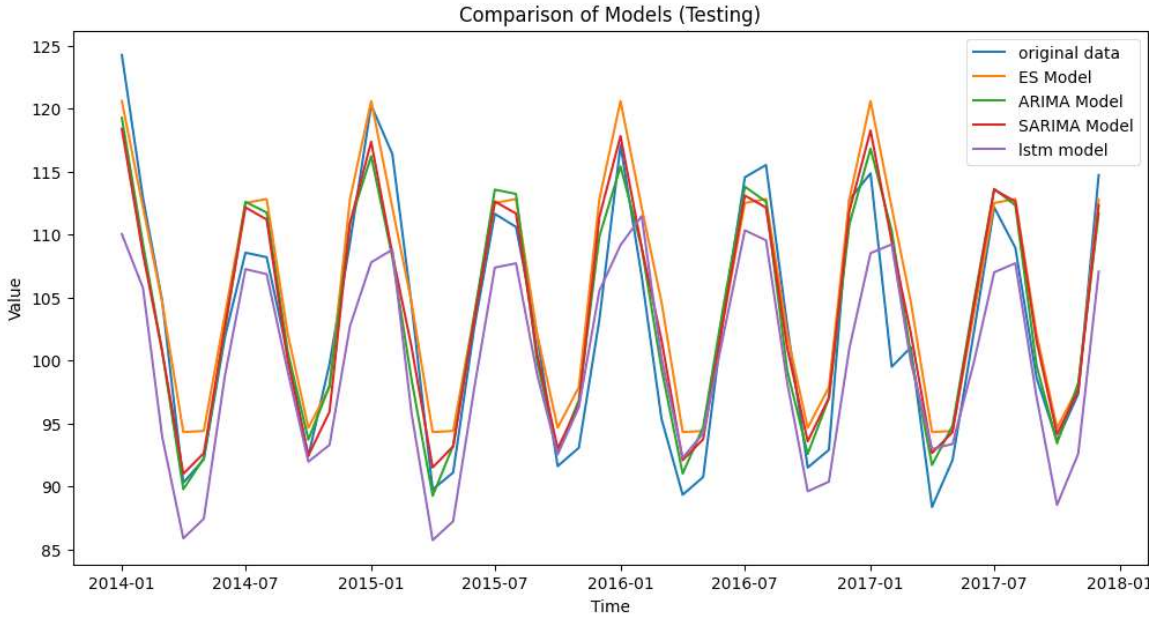
6)SARIMA model:

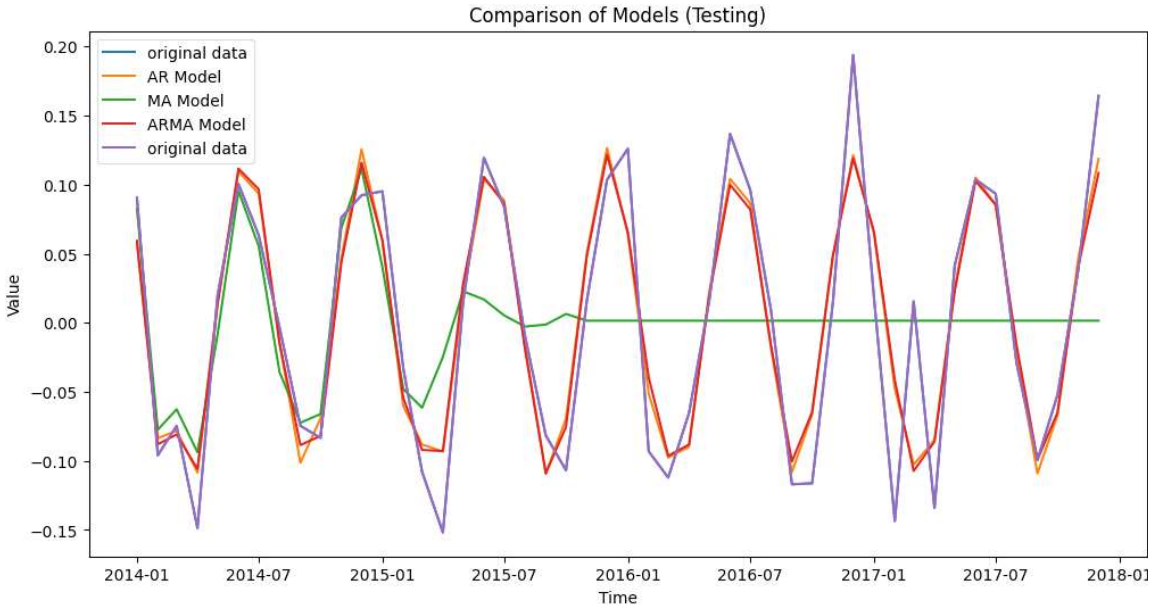
- **Optimal lag order (p, d, q):** (3, 1, 1)
- **Optimal seasonal lag order (P, D, Q):** (1, 0, 1)
- **RMSE:** 3.4018
- **R-squared:** 0.8719
- **MSE:** 11.5722
- **MAPE:** 0.0257%

7)LSTM model:

- **RMSE:** 5.2177
- **R-squared:** 0.6986
- **MSE:** 27.2243
- **MAPE:** 0.0388

Visual representation:





Thoughts:

Considering both accuracy metrics and visual inspection, the ARIMA model appears to outperform other models, exhibiting the lowest RMSE and MAPE values, as well as a high R-squared value. The AR model also performs well, demonstrating a good balance between accuracy and simplicity. The LSTM model, while providing reasonable results, appears to have higher RMSE and MAPE compared to the ARIMA model.

It is recommended to consider the trade-off between model complexity and forecasting accuracy when selecting the final model for practical applications. Additionally, sensitivity analysis and further tuning of hyperparameters may enhance the performance of the models.