SUMMER PROJECT

Name: Sudipta Das
Roll No.: 22N0075
Course: MSc. ASI
Topic: Time series
Guide: Prof. Sanjeev V. Sabnis

Objective:

The objective of time series modelling for monthly average price of potato chips is to develop a reliable and accurate forecasting model that can capture the underlying patterns and trends in the price fluctuations over time. By analyzing historical price data, the goal is to identify seasonality, trends, and other temporal dependencies that influence the potato chips' pricing behaviour.

Data description:

The dataset was taken from the website of FRED(Federal Reserve Economic Data). It consists monthly average price of potato chips in US cities. The price was measured in US dollar per 16 ounces. It has data from 1980 to 2020.

Data View:

First 10 observations:

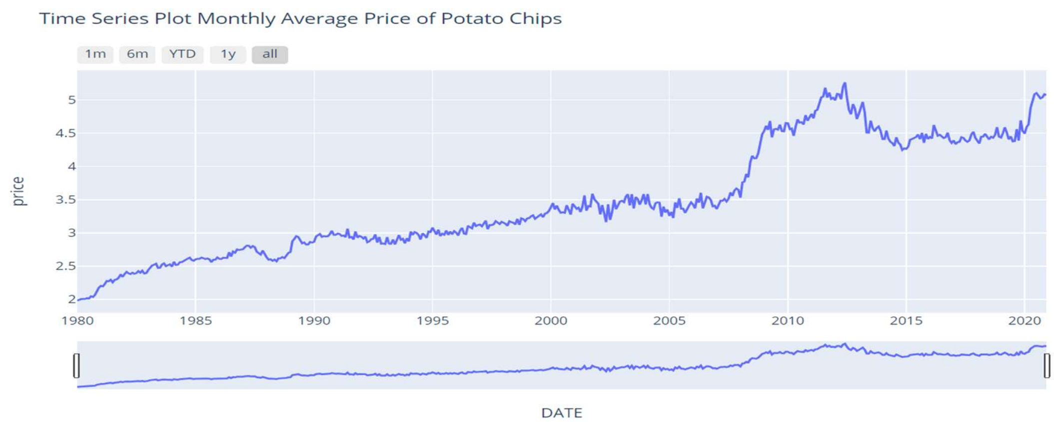
	DATE	price
0	1980-01-01	1.981
1	1980-02-01	1.994
2	1980-03-01	2.003
3	1980-04-01	2.006
4	1980-05-01	2.006
5	1980-06-01	2.018
6	1980-07-01	2.012
7	1980-08-01	2.046
8	1980-09-01	2.035
9	1980-10-01	2.066

Data Information:

There is no null or missing present in this data. It has total 492 observations.

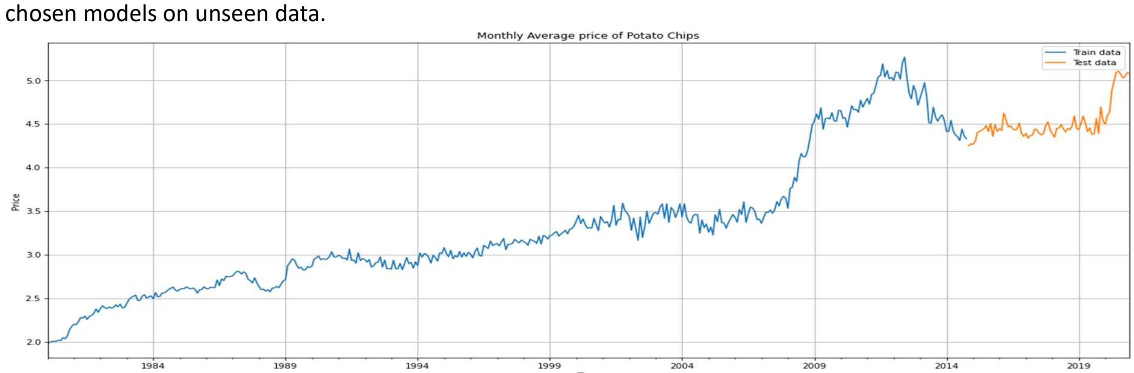
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 492 entries, 0 to 491
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   DATE    492 non-null    datetime64[ns]
 1   price   492 non-null    float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 7.8 KB
```

Plot of the Data:

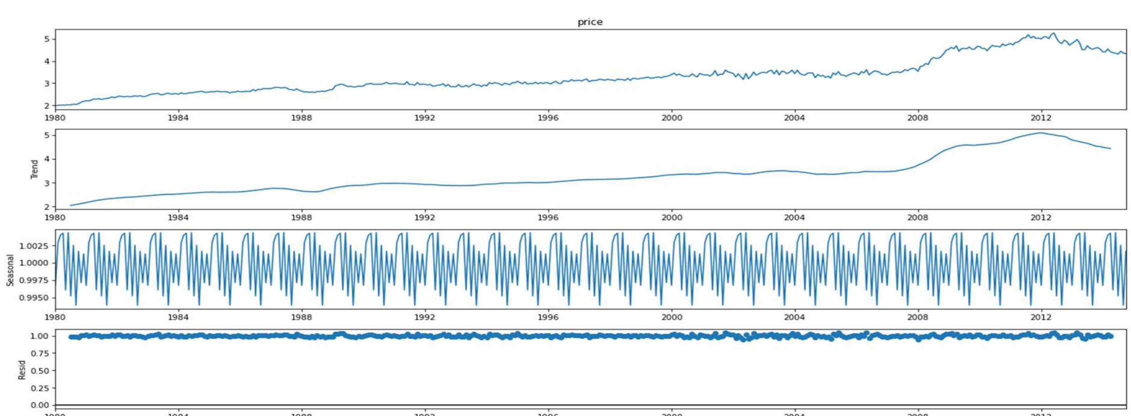


Based on above plot, it is evident that the plot exhibits an upwards trend over time. This characteristic indicates that the data is non-stationary.

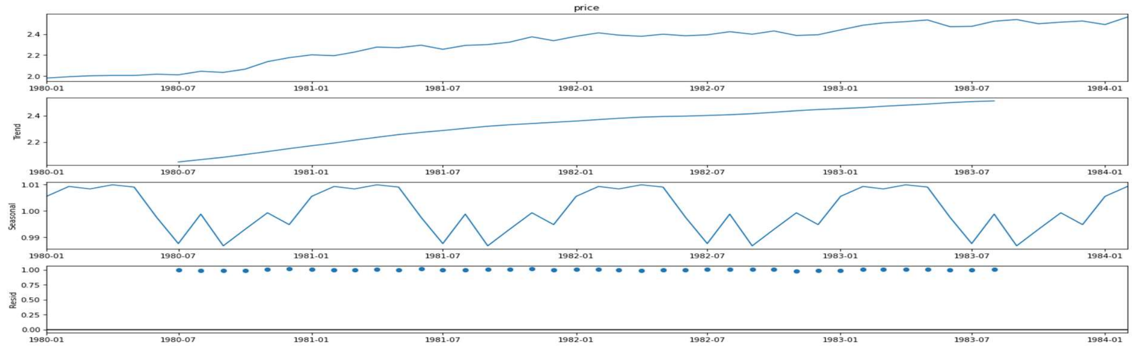
The dataset will be divided into two distinct subsets “Train” dataset and “Test” dataset. We will allocate the first 85% of the entire data to the Train dataset and remaining 15% to the Test dataset. The Train dataset will be utilized to train and develop time series models. The Test dataset will be utilized to check performance of the chosen models on unseen data.



Decomposition understanding:



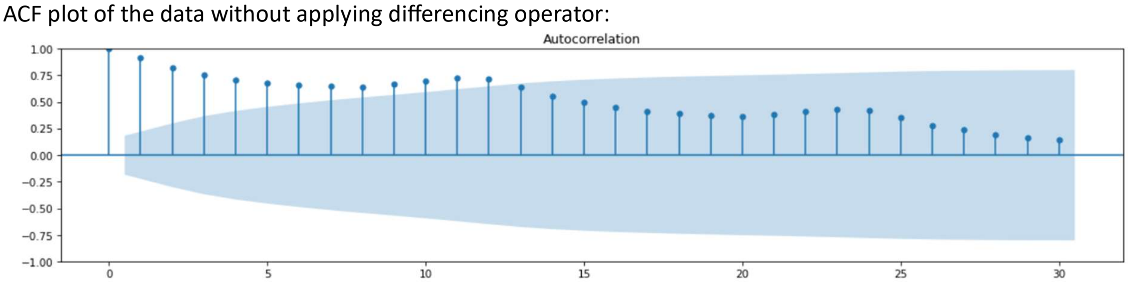
A closer view



Based on above plots we can infer that the data exhibits an upward trend also it contains seasonal pattern with periods 12(months).

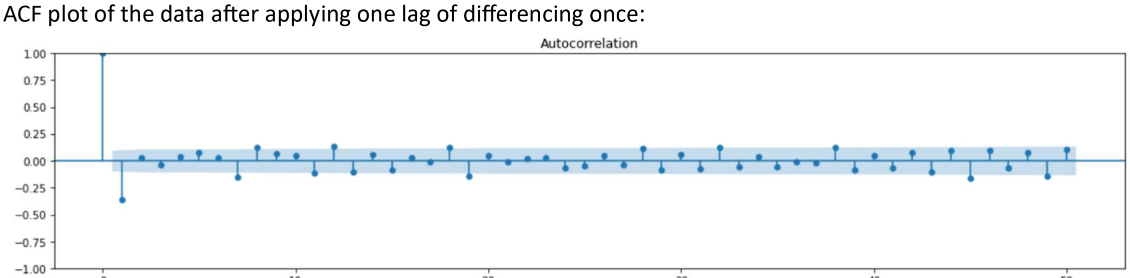
Stationarity:

ACF plot of the data without applying differencing operator:



The above plot shows that the auto-correlation function decaying slowly. It indicates that the data is non-stationary.

ACF plot of the data after applying one lag of differencing once:



As we can see that after applying one lag of differencing operator once the auto-correlation function die out after lag 1. So, data may become stationary.

Let’s confirm it using ADF test:

ADF test of the data without applying differencing operator:

```
ADF Statistic: -1.182860
p-value: 0.680925
Critical Values:
        1%: -3.447
        5%: -2.869
        10%: -2.571
```

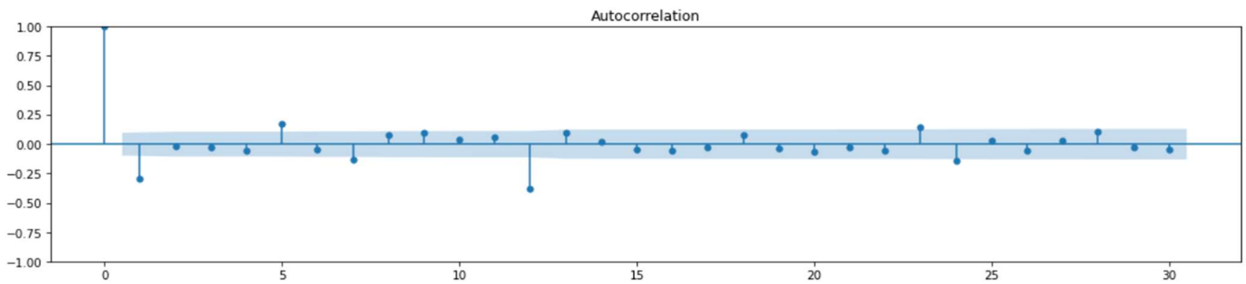
ADF test of the data after applying one lag of differencing operator once:

```
ADF Statistic: -3.985024
p-value: 0.001491
Critical Values:
        1%: -3.447
        5%: -2.869
        10%: -2.571
```

As p-value of the ADF test of the data after applying one lag of differencing once is less than 0.05, we can conclude that the stationarity achieved by using one lag of differencing once.

As, we have seen before that the data has seasonal pattern of period 12. To eliminate the seasonality from the data I had applied 12 lags of differencing to the trend eliminated data:

Appling differencing of lag 12 once



The above ACF plot die out after lag 1 and it has only one significant seasonal lag (at lag 12). So, we have eliminated seasonality from data.

ADF test after applying 12 lags of differencing on trend eliminated data:

```
ADF Statistic: -12.176268
p-value: 0.000000
Critical Values:
        1%: -3.497
        5%: -2.891
        10%: -2.582
```

<u>Model Parameters:</u>

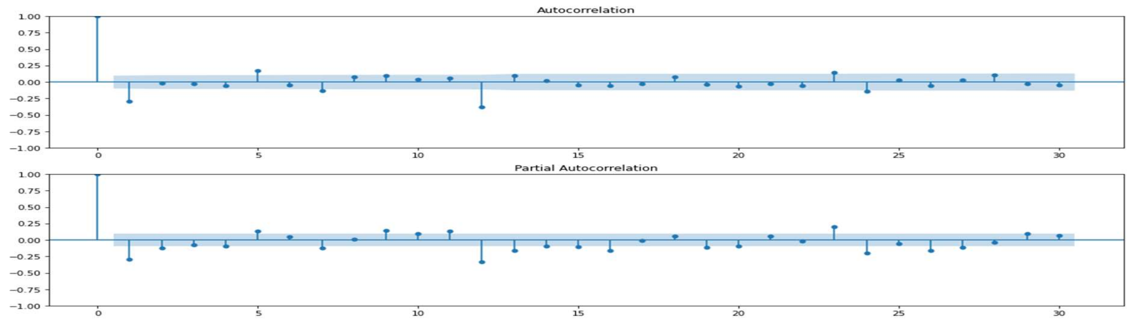
- **<u>SARIMA (p,d,q)x(P,D,Q,S):</u>**
  SARIMA stands for Seasonal Autoregression Integrated Moving Average. When a time series data contains seasonal pattern, we should model that time series data using SARIMA. A SARIMA model has total 7 parameters which are following

  p: non-seasonal autoregression parameter
  q: non-seasonal moving average parameter
  d: number of times we should apply differencing operator of lag one to eliminate the trend of data

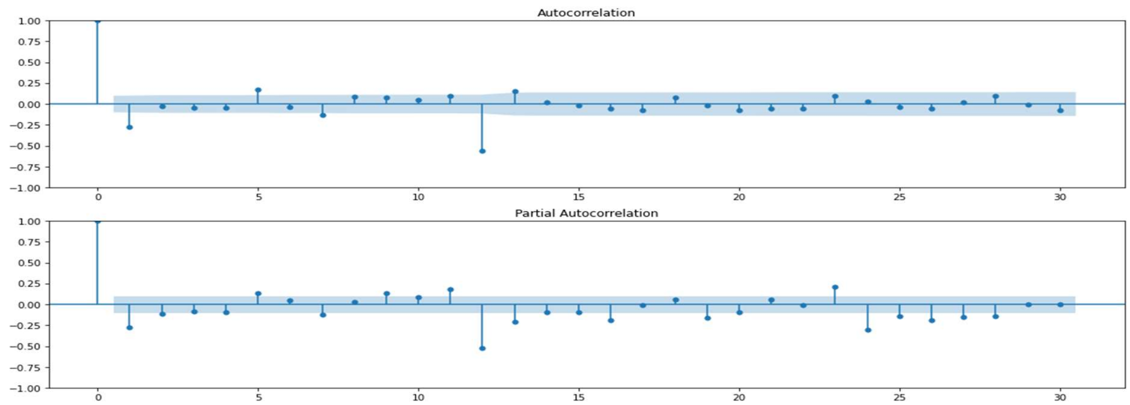
  P: seasonal autoregression parameter
  Q: seasonal moving average parameter
  S: period of the seasonal pattern
  D: number of times we should apply differencing operator of lag S to eliminate the seasonal pattern of data

So, we have to find all those 7 parameters to train a SARIMA model.

Here I had applied 12 lags of differencing once on top of one lag of differencing once:



Here I had applied 12 lags of differencing twice on top of one lag of differencing once:



So, potential parameters are as follows:

$$p = 1\,; q = 1\,; d = 1$$
$$P = 1{,}2\,; Q = 1\,; D = 1{,}2\,; S = 12$$

So, potential SARIMA models are as follows:

1. $SARIMA(1,1,1) \times (1,1,1,12)$
2. $SARIMA(1,1,1) \times (1,2,1,12)$
3. $SARIMA(1,1,1) \times (2,1,1,12)$
4. $SARIMA(1,1,1) \times (2,2,1,12)$

Modelling:

$Model1: SARIMA(1,1,1) \times (1,1,1,12)$

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                price   No. Observations:                418
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood                 468.230
Date:                              Mon, 31 Jul 2023   AIC                           -926.461
Time:                                      09:54:06   BIC                           -906.442
Sample:                                  01-01-1980   HQIC                          -918.537
                                       - 10-01-2014
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0095      0.124     -0.077      0.939      -0.252       0.233
ma.L1         -0.3485      0.114     -3.064      0.002      -0.571      -0.126
ar.S.L12       0.0789      0.052      1.516      0.129      -0.023       0.181
ma.S.L12      -0.9930      0.216     -4.607      0.000      -1.415      -0.571
sigma2         0.0053      0.001      4.667      0.000       0.003       0.007
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                77.16
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               6.34   Skew:                            -0.21
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.10
===================================================================================
```

$Model2: SARIMA(1,,1,1) \times (2,1,1,12)$

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                price   No. Observations:                418
Model:             SARIMAX(1, 1, 1)x(2, 1, 1, 12)   Log Likelihood                 471.052
Date:                              Mon, 31 Jul 2023   AIC                           -930.105
Time:                                      10:02:37   BIC                           -906.081
Sample:                                  01-01-1980   HQIC                          -920.596
                                       - 10-01-2014
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0066      0.126     -0.052      0.959      -0.254       0.241
ma.L1         -0.3454      0.117     -2.957      0.003      -0.574      -0.116
ar.S.L12       0.0775      0.051      1.527      0.127      -0.022       0.177
ar.S.L24      -0.1267      0.045     -2.795      0.005      -0.215      -0.038
ma.S.L12      -0.9879      0.115     -8.562      0.000      -1.214      -0.762
sigma2         0.0052      0.001      8.338      0.000       0.004       0.006
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                68.66
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               6.20   Skew:                            -0.16
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.99
===================================================================================
```

$Model3: SARIMA(1,1,1) \times (1,2,1,12)$

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                price   No. Observations:                418
Model:             SARIMAX(1, 1, 1)x(1, 2, 1, 12)   Log Likelihood                 356.661
Date:                              Mon, 31 Jul 2023   AIC                           -703.323
Time:                                      10:06:32   BIC                           -683.454
Sample:                                  01-01-1980   HQIC                          -695.449
                                       - 10-01-2014
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0076      0.120      0.064      0.949      -0.227       0.242
ma.L1         -0.3812      0.109     -3.487      0.000      -0.595      -0.167
ar.S.L12      -0.3977      0.044     -8.966      0.000      -0.485      -0.311
ma.S.L12      -0.9913      0.470     -2.111      0.035      -1.912      -0.071
sigma2         0.0084      0.004      2.139      0.032       0.001       0.016
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               154.04
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               6.81   Skew:                            -0.41
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.95
===================================================================================
```

$Model4: SARIMA(1,1,1) \times (2,2,1,12)$

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                price   No. Observations:                  418
Model:             SARIMAX(1, 1, 1)x(2, 2, 1, 12)   Log Likelihood                 377.005
Date:                              Mon, 31 Jul 2023   AIC                           -742.011
Time:                                      10:08:00   BIC                           -718.168
Sample:                                  01-01-1980   HQIC                          -732.562
                                       - 10-01-2014
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0487      0.119      0.411      0.681      -0.184       0.281
ma.L1         -0.4222      0.111     -3.817      0.000      -0.639      -0.205
ar.S.L12      -0.5331      0.055     -9.775      0.000      -0.640      -0.426
ar.S.L24      -0.3375      0.053     -6.394      0.000      -0.441      -0.234
ma.S.L12      -0.9944      0.643     -1.547      0.122      -2.254       0.265
sigma2         0.0074      0.005      1.551      0.121      -0.002       0.017
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               190.63
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               5.94   Skew:                            -0.56
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.22
===================================================================================
```

We will choose model based on least AIC score:

Model	AIC score
$SARIMA(1,1,1) \times (1,1,1,12)$	-926.461
$SARIMA(1,1,1) \times (2,1,1,12)$	-930.105
$SARIMA(1,1,1) \times (1,2,1,12)$	-703.323
$SARIMA(1,1,1) \times (2,2,1,12)$	-742.011

Model diagnostics:

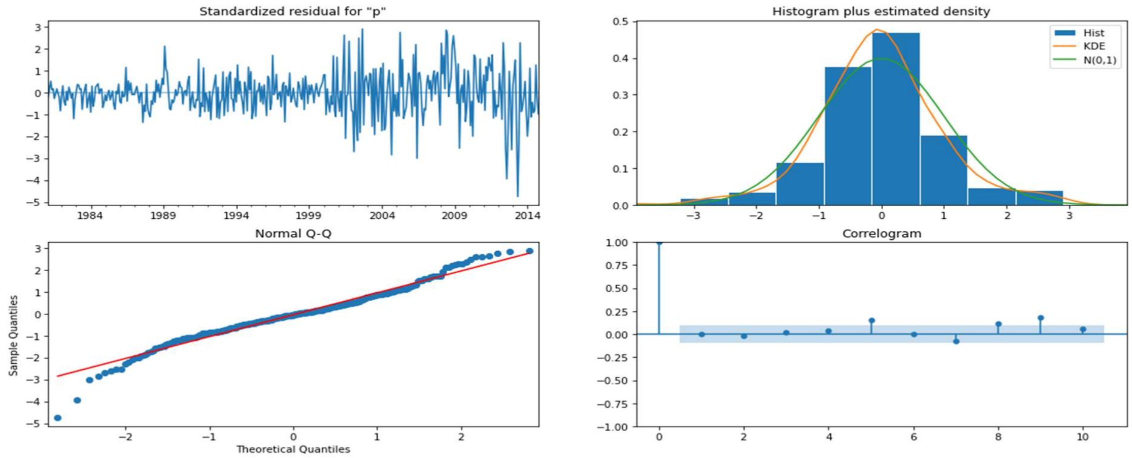
- **Ljung-Box test:**
  We will use Ljung-Box test in model diagnostics part to identify whether the residuals of a model are auto-correlated or not. The null hypothesis of this test is “there is no autocorrelation in the residuals” And the alternative hypothesis is “there is significant autocorrelation in the residuals”

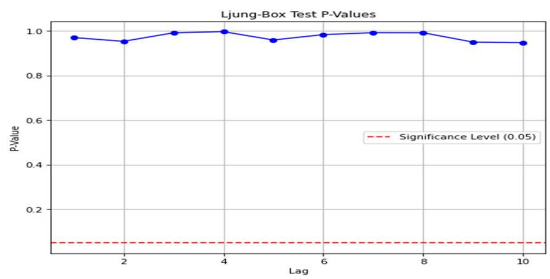
  $$H_0 : There\ is\ no\ auto-correlation$$
  $$H_a : There\ is\ significant\ auto-correlation$$

- **Breusch pagan test:**
  We will use Breusch pagan test in the model diagnostics part to identify whether the residuals had constant variance or not.

  $$H_0 : The\ variance\ of\ the\ residuals\ is\ constant\ (Homoscedasticity)$$
  $$H_a : The\ variance\ of\ the\ residuals\ is\ not\ constant\ (Heteroscedasticity)$$

$SARIMA(1,1,1) \times (1,1,1,12)$

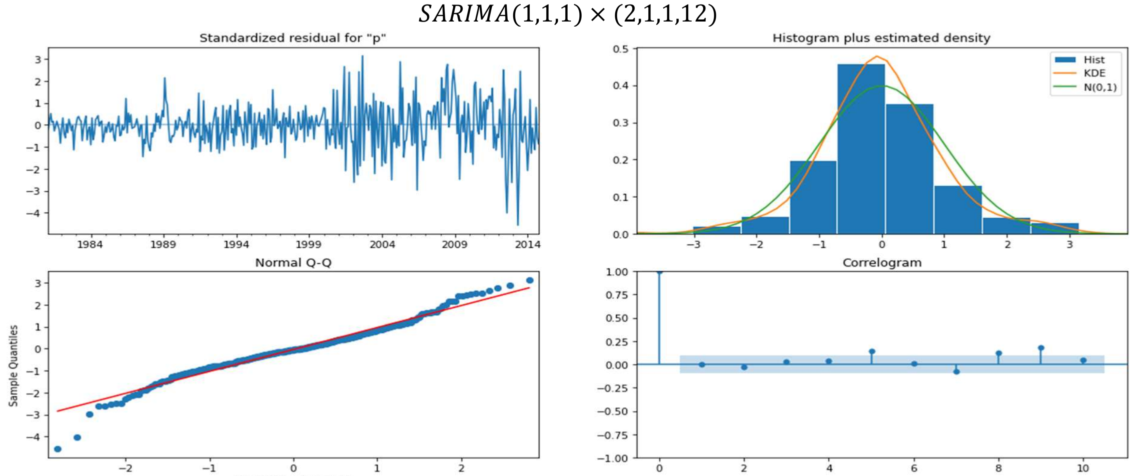


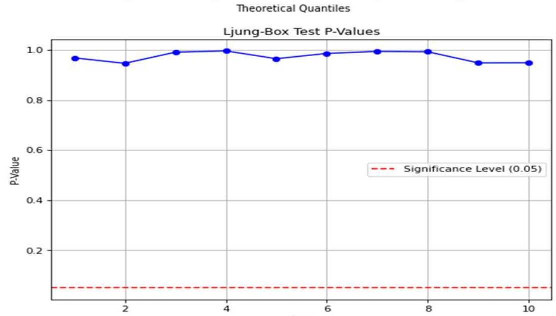



```
# Breusch Pagan test
_, p_value, _, _ = het_breuschpagan(resid1, train)
# Test the null hypothesis at a certain significance level (e.g., 0.05)
alpha = 0.05
if p_value < alpha:
    print("The residuals are likely heteroscedastic (reject null hypothesis).")
else:
    print("The residuals are likely homoscedastic (fail to reject null hypothesis).")
```

The residuals are likely homoscedastic (fail to reject null hypothesis).

$SARIMA(1,1,1) \times (2,1,1,12)$

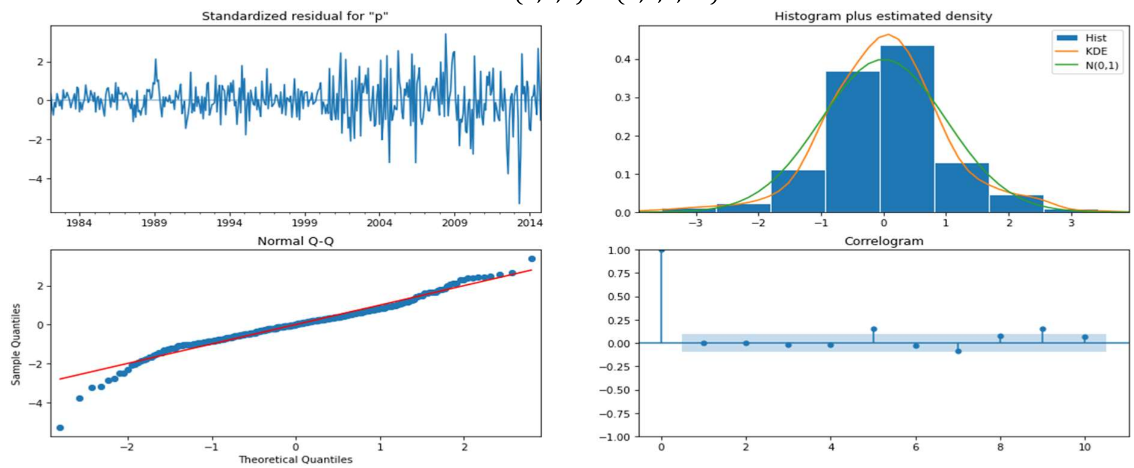


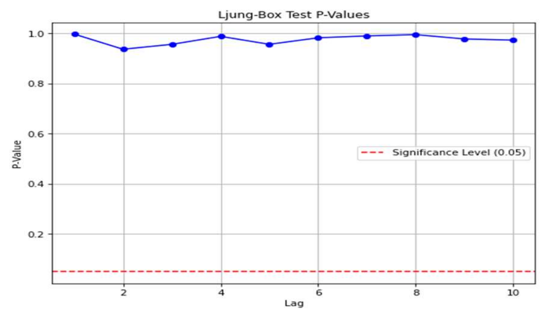



```
# Breusch Pagan test
_, p_value, _, _ = het_breuschpagan(resid2, train)
# Test the null hypothesis at a certain significance level (e.g., 0.05)
alpha = 0.05
if p_value < alpha:
    print("The residuals are likely heteroscedastic (reject null hypothesis).")
else:
    print("The residuals are likely homoscedastic (fail to reject null hypothesis).")
```

The residuals are likely homoscedastic (fail to reject null hypothesis).

$SARIMA(1,1,1) \times (1,2,1,12)$

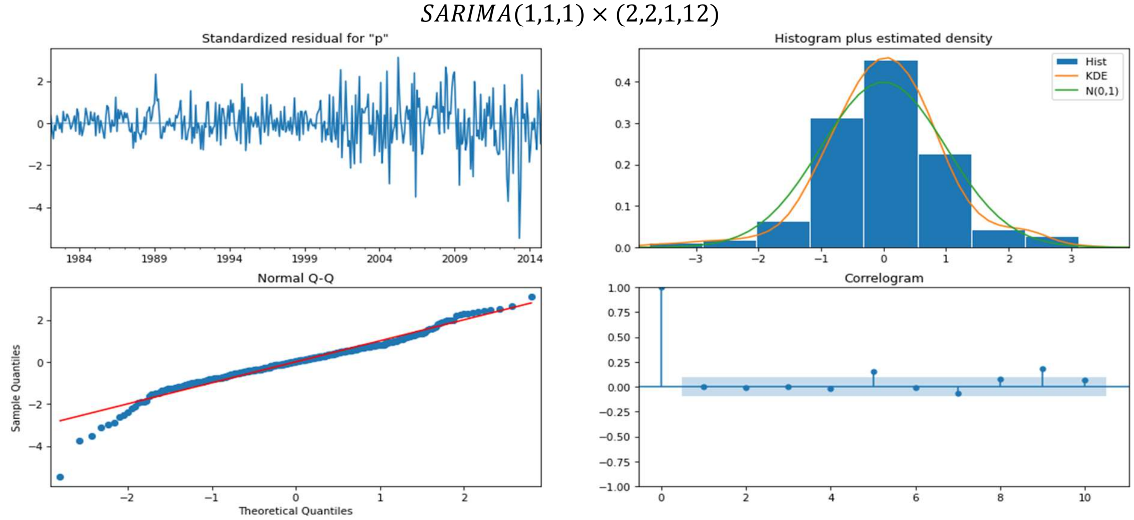


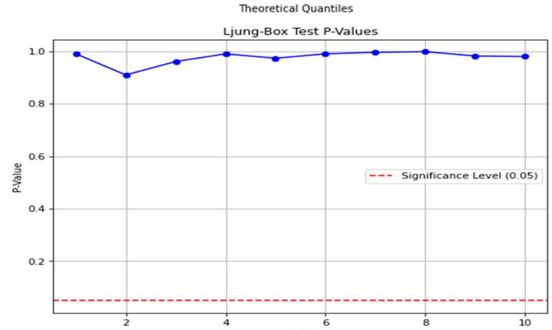



```python
# Breusch Pagan test
_, p_value, _, _ = het_breuschpagan(resid3, train)
# Test the null hypothesis at a certain significance level (e.g., 0.05)
alpha = 0.05
if p_value < alpha:
    print("The residuals are likely heteroscedastic (reject null hypothesis).")
else:
    print("The residuals are likely homoscedastic (fail to reject null hypothesis).")
```

```
The residuals are likely homoscedastic (fail to reject null hypothesis).
```

$SARIMA(1,1,1) \times (2,2,1,12)$






```python
# Breusch Pagan test
_, p_value, _, _ = het_breuschpagan(resid4, train)
# Test the null hypothesis at a certain significance level (e.g., 0.05)
alpha = 0.05
if p_value < alpha:
    print("The residuals are likely heteroscedastic (reject null hypothesis).")
else:
    print("The residuals are likely homoscedastic (fail to reject null hypothesis).")
```

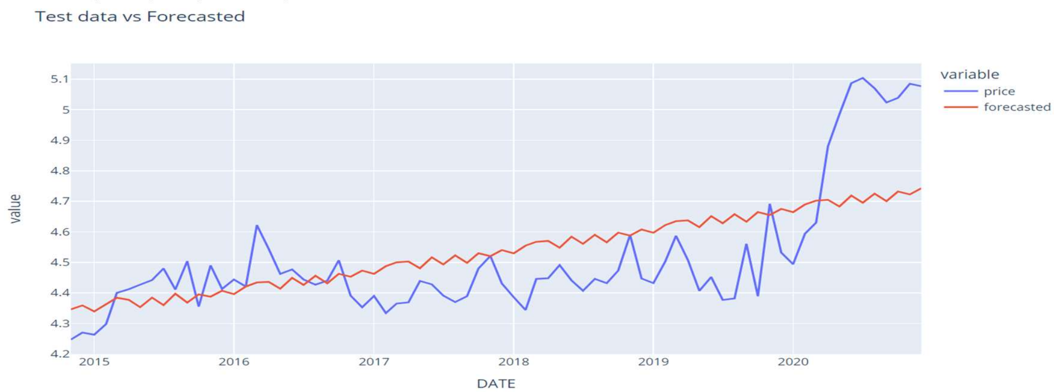
```
The residuals are likely homoscedastic (fail to reject null hypothesis).
```

From above analysis we can conclude that the residuals of each model follow white noise process. So, all the four models satisfy the model assumptions.

Forecast:

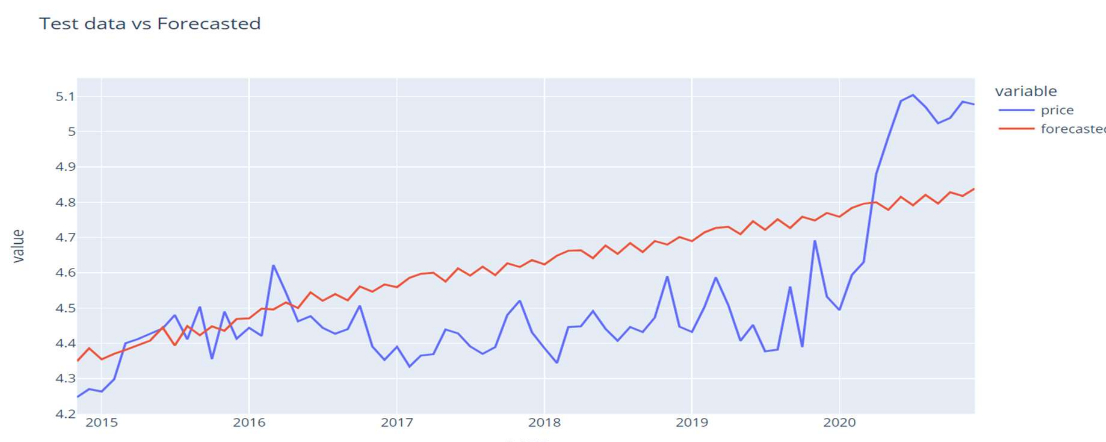
We are going to forecast using SARIMA(1,1,1)x(1,1,1,12) and SARIMA(1,1,1)x(2,1,1,12) models as this two models satisfies model assumptions and have least AIC scores.

- $SARIMA(1,1,1) \times (1,1,1,12)$



DATE	price	forecasted
2014-11-01	4.247	4.345922
2014-12-01	4.270	4.359009
2015-01-01	4.263	4.339379
2015-02-01	4.298	4.362527
2015-03-01	4.400	4.384150
2015-04-01	4.412	4.377324
2015-05-01	4.427	4.353221
2015-06-01	4.442	4.384583
2015-07-01	4.480	4.359820
2015-08-01	4.411	4.397216

- $SARIMA(1,1,1) \times (2,1,1,12)$



DATE	price	forecasted
2014-11-01	4.247	4.349094
2014-12-01	4.270	4.385790
2015-01-01	4.263	4.354380
2015-02-01	4.298	4.370068
2015-03-01	4.400	4.380979
2015-04-01	4.412	4.395448
2015-05-01	4.427	4.407591
2015-06-01	4.442	4.444891
2015-07-01	4.480	4.393507
2015-08-01	4.411	4.448533

<u>MAPE score:</u>
MAPE stands for Mean Absolute Percentage Error, and it is a statistical measure used check performance of a model.

The formula for MAPE is as follows:

$$MAPE = \frac{1}{n}\sum\left(\left|\frac{Actual - Forecas}{Actual}\right|\right) \times 100$$

Model	MAPE
$SARIMA(1,1,1) \times (1,1,1,12)$	2.78
$SARIMA(1,1,1) \times (2,1,1,12)$	3.73

<u>Conclusion:</u>

Finally, we can conclude that the model $SARIMA(1,1,1) \times (1,1,1,12)$ performed well on our data with 2.78% mean absolute error on unseen data.