Time Series Analysis and Forecasting of Beer Prices in Australia: 1960-2010

Sonu Gupta(22N0062) & Anik Paul(22N0070)

Summary

1 Abstract

This project aims to analyze quarterly data on beer prices in Australia spanning from 1960 to 2010, employing time series analysis techniques to understand the underlying patterns and dynamics. The primary objective is to identify and fit appropriate time series models to the dataset, thereby enabling the forecasting of future beer prices. The analysis involves exploring the temporal dependencies, trends, and seasonality present in the data to select the most suitable model. Various time series models, including autoregressive integrated moving average (ARIMA), seasonal ARIMA (SARIMA), and possibly more sophisticated models such as exponential smoothing methods, will be considered and evaluated based on their ability to capture the historical price variations accurately. Additionally, diagnostics tests will be employed to assess the adequacy of the selected model and ensure its robustness. Ultimately, leveraging the chosen model, this study aims to provide reliable forecasts of beer prices in Australia, offering valuable insights for stakeholders in the brewing industry, policymakers, and consumers.

2 Theory

2.1 Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are essential tools in time series analysis for understanding the correlation structure within a time series dataset.

ACF: The ACF measures the correlation between a time series and its lagged values. Mathematically, it is defined as:

$$\rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\sqrt{\mathrm{Var}(y_t) \cdot \mathrm{Var}(y_{t-k})}}$$

where ρ_k represents the autocorrelation at lag k, y_t is the time series at time t, and Cov and Var represent the covariance and variance, respectively.

PACF: The PACF measures the correlation between a time series and its lagged values while controlling for the intermediate lags. It helps identify the direct effect of a lag on the current value of the series. Mathematically, it can be calculated using the Durbin-Levinson recursion formula.

2.2 Model Selection based on ACF and PACF

The ACF and PACF plots provide insights into the potential structure of the time series data, aiding in model selection. For instance:

- A rapidly decaying ACF and a significant cutoff in the PACF suggest an AR(p) model.
- A rapidly decaying PACF and a significant cutoff in the ACF suggest an MA(q) model.
- Both slowly decaying ACF and PACF suggest a possible integrated component, leading to an ARIMA model.

2.3 ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal ARIMA)

ARIMA: ARIMA models are a class of models that capture autocorrelation in a time series. They consist of three main components: Autoregression (AR), Differencing (I), and Moving Average (MA). The general form of an ARIMA model of order (p, d, q) is:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t$$

where Y_t is the observed value at time t, c is a constant, $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive parameters, $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average parameters, ϵ_t is white noise, and d is the degree of differencing.

SARIMA: SARIMA extends the ARIMA model to incorporate seasonal components. It includes additional seasonal autoregressive (SAR) and seasonal moving average (SMA) terms. A SARIMA model is denoted as $(p, d, q) \times (P, D, Q)_s$, where (p, d, q) are the non-seasonal components and $(P, D, Q)_s$ are the seasonal components with period s.

2.4 Ljung-Box Test and Portmanteau Test

The Ljung-Box test and Portmanteau test are statistical tests used to evaluate whether the residuals of a time series model exhibit autocorrelation at various lags.

Ljung-Box Test: The Ljung-Box test is based on the sum of squares of autocorrelations of the residuals up to a certain lag. The test statistic is:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

where $\hat{\rho}_k$ are the sample autocorrelations of the residuals at lag k, n is the sample size, and h is the number of lags being tested.

Portmanteau Test: The Portmanteau test is similar to the Ljung-Box test but uses a different formulation to test for autocorrelation in residuals. It provides a single test statistic, which is compared to the Chi-squared distribution with degrees of freedom equal to the number of lags being tested.

Both tests are used to assess whether the residuals are independently and identically distributed (iid), which is a key assumption of many time series models. A significant result suggests the presence of autocorrelation in the residuals, indicating that the model might be misspecified.

2.5 Metrics for accuracy checking

Mean Squared Error (MSE):

The Mean Squared Error (MSE) is a measure of the average squared difference between the actual and predicted values in a dataset. It is calculated as the average of the squared residuals.

Mathematically, the MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- n is the number of observations,
- y_i represents the actual value of the target variable for observation i,
- $\hat{y}_i$ represents the predicted value of the target variable for observation i.

Mean Absolute Percentage Error (MAPE):

The Mean Absolute Percentage Error (MAPE) is a measure of the average percentage difference between the actual and predicted values in a dataset. It is calculated as the average of the absolute percentage errors.

Mathematically, the MAPE is defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

where:

- n is the number of observations,
- y_i represents the actual value of the target variable for observation i,
- $\hat{y}_i$ represents the predicted value of the target variable for observation i.

3 Data Description

The data is collected from CRAN Library. The data contains quarterly data on beer price in Australia from 1960 to 2010.

	Quarter	Beer
0	1956 Q1	284
1	1956 Q2	213
2	1956 Q3	227
3	1956 Q4	308
4	1957 Q1	262

Figure 1: First 5 rows of the dataset

The plot of the data is like this:

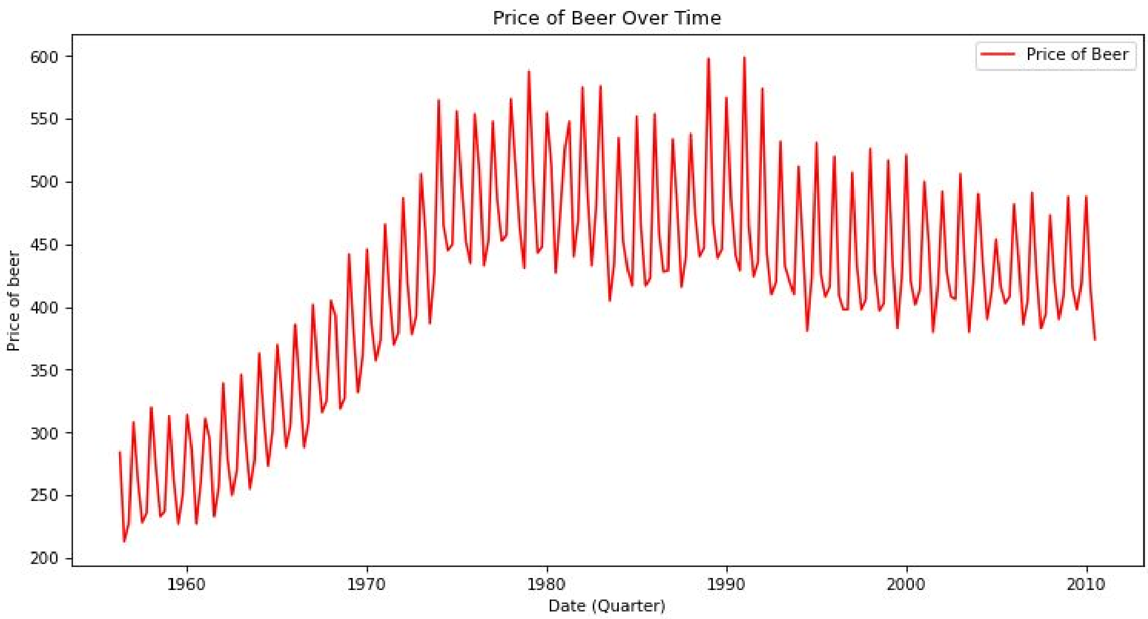

Figure 2: Plot of the dataset

4 Analysis Methodologies

4.1 Data Splitting

The dataset was divided into training and testing sets, with the first 80% of the data allocated for training and the remaining 20% for testing.

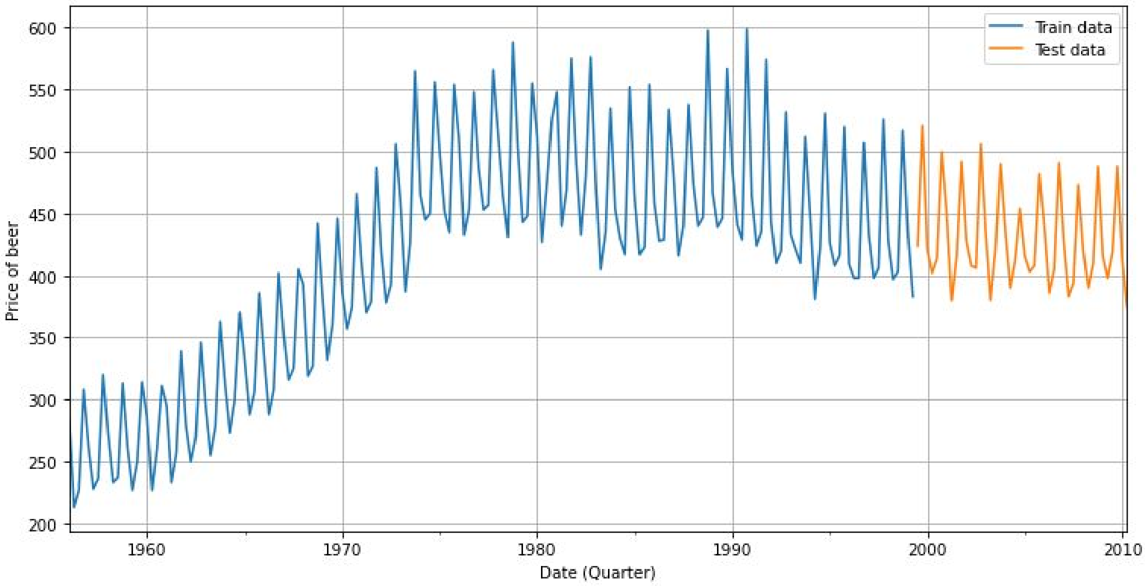

Figure 3: Train-test split of the dataset

4.2 Trend Analysis

To assess the presence of trend in the dataset, a decomposition technique was employed to separate the time series into trend, seasonality, and error components.

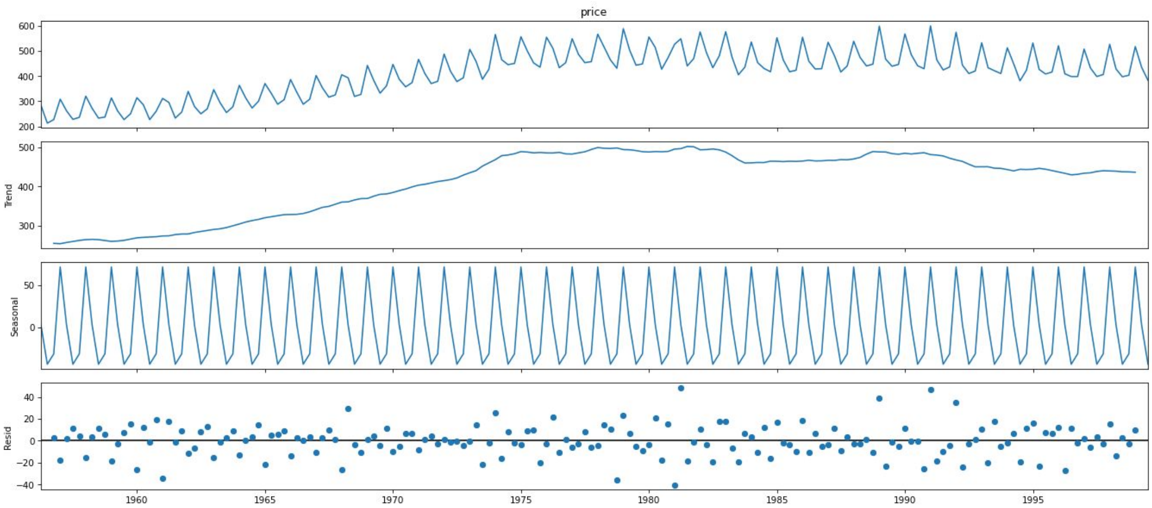

Figure 4: Decomposition of Time series

4.3 Autocorrelation and Partial Autocorrelation Analysis

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were generated to examine the autocorrelation structure of the time series data. Significant periodic spike patterns were observed in the ACF plot.

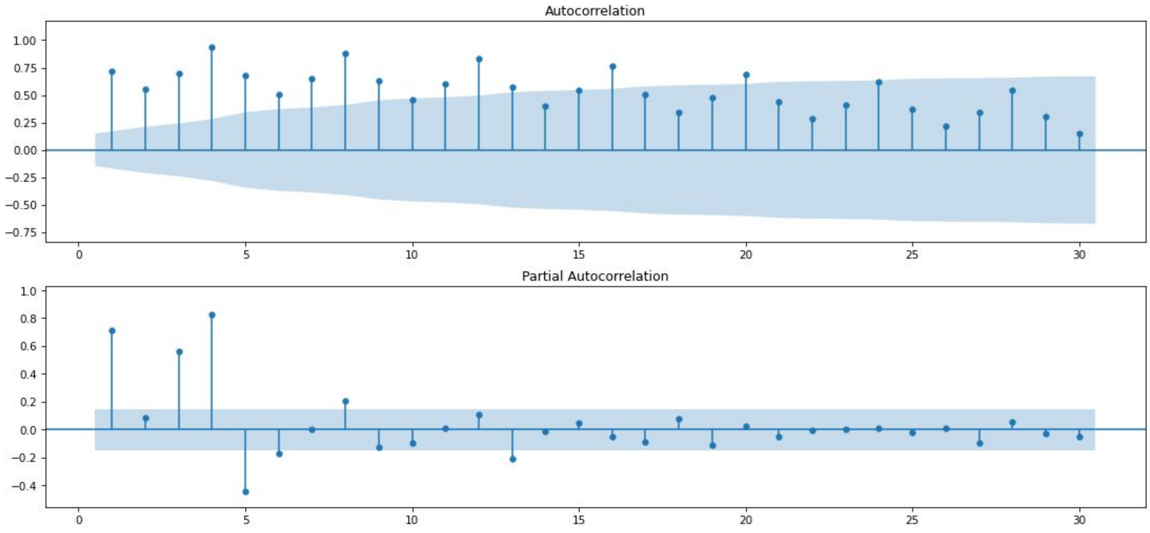

Figure 5: ACF & PACF of the whole dataset

4.4 Differencing

First-order differencing was applied to the dataset to remove the trend component and reduce the non-seasonal lags in autocorrelation. Subsequently, ACF and PACF plots were re-examined to assess the effectiveness of differencing.

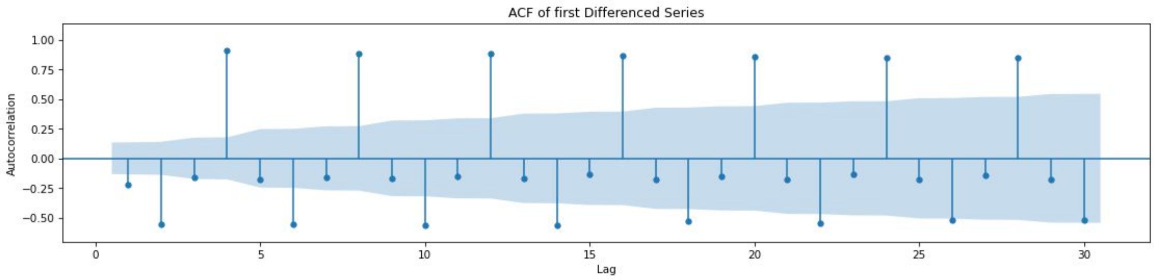

Figure 6: ACF of First order differenced data

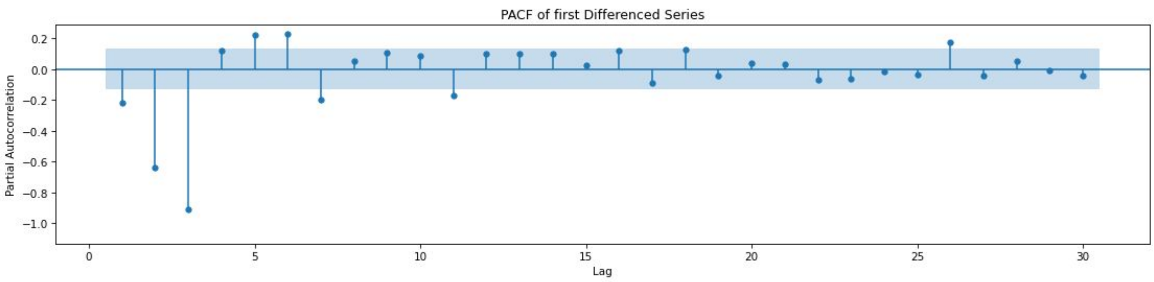

Figure 7: PACF of First order differenced data

4.5 Seasonal Differencing

To address the seasonality observed in the dataset, seasonal differencing of period 4 was performed. This process aimed to eliminate the seasonal component from the time series.

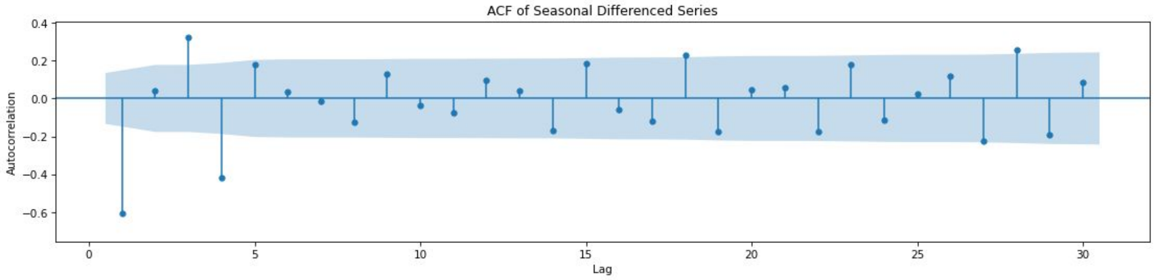

Figure 8: ACF of seasonality removed data

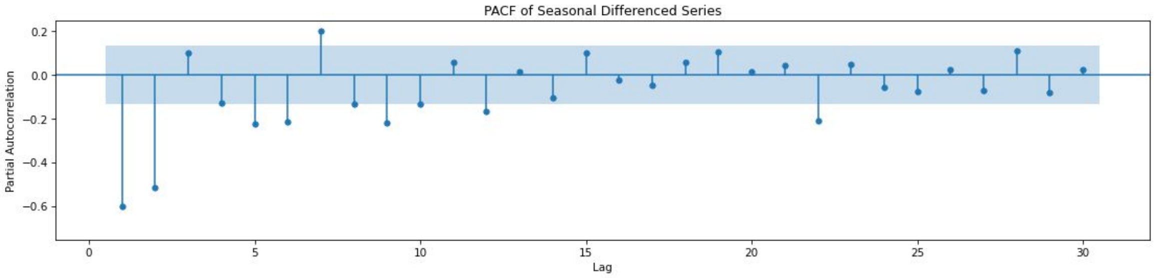

Figure 9: PACF of seasonality differenced data

4.6 Model Selection

Based on the ACF and PACF plots post-differencing, two candidate ARIMA models were identified: ARIMA(3,1,3) and ARIMA(3,1,2). These models were selected considering their ability to capture the autocorrelation structure and seasonal patterns observed in the data.

5 Result

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  price   No. Observations:                  174
Model:               SARIMAX(3, 1, 2)   Log Likelihood                -726.260
Date:                Thu, 02 May 2024   AIC                           1466.521
Time:                        00:57:46   BIC                           1488.471
Sample:                    03-31-1956   HQIC                          1475.428
                         - 06-30-1999
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      3.1816      2.169      1.467      0.142      -1.069       7.433
ar.L1         -0.9009      0.038    -23.461      0.000      -0.976      -0.826
ar.L2         -1.0036      0.007   -144.315      0.000      -1.017      -0.990
ar.L3         -0.8985      0.039    -23.270      0.000      -0.974      -0.823
ma.L1         -0.1166      0.052     -2.247      0.025      -0.218      -0.015
ma.L2          0.6947      0.053     13.174      0.000       0.591       0.798
sigma2       297.7777     30.176      9.868      0.000     238.633     356.922
===================================================================================
Ljung-Box (L1) (Q):                   0.17   Jarque-Bera (JB):                 3.03
Prob(Q):                              0.68   Prob(JB):                         0.22
Heteroskedasticity (H):               2.03   Skew:                            -0.14
Prob(H) (two-sided):                  0.01   Kurtosis:                         3.59
===================================================================================
```

Figure 10: Summary of ARIMA(3,1,2)

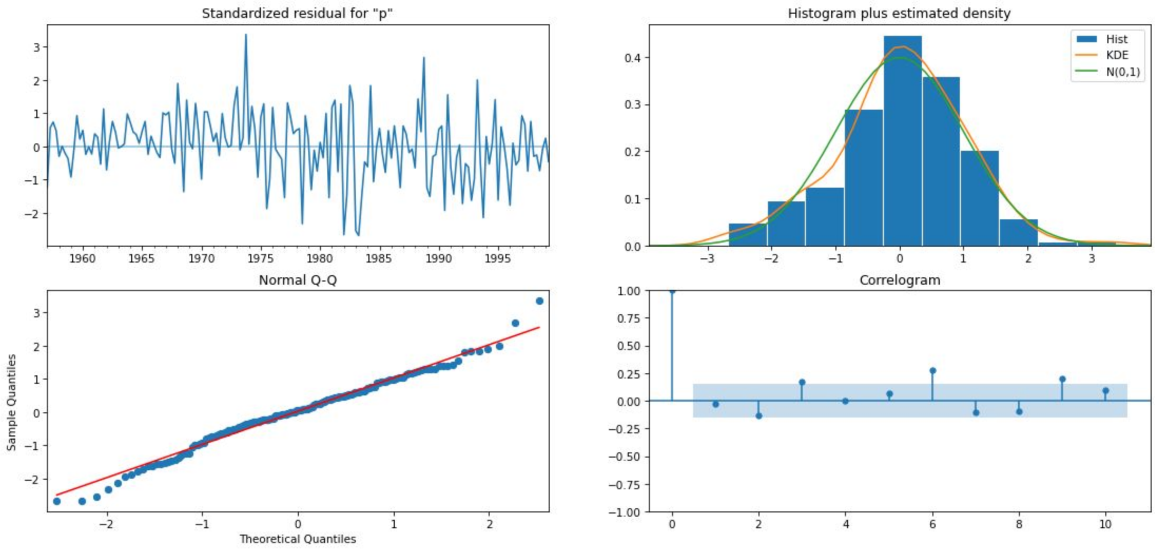

Figure 11: Model Adequacy Checking of ARIMA(3,1,2)

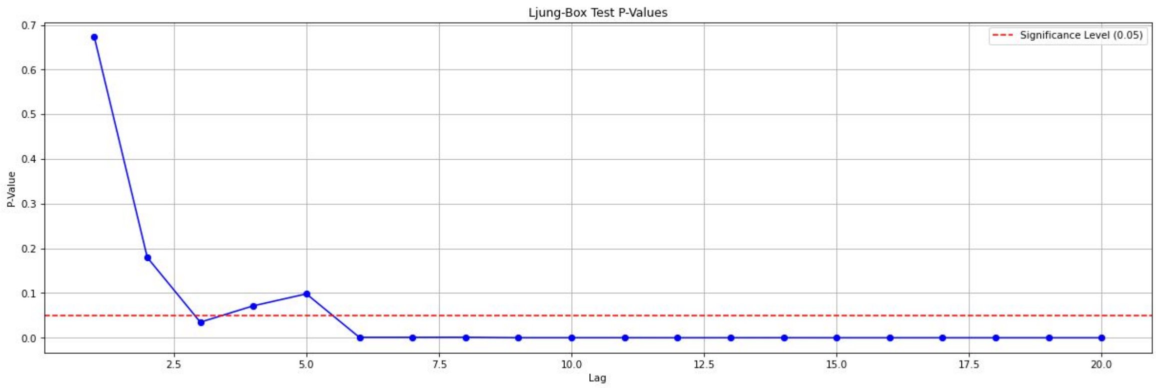

Figure 12: Ljung box test for ARIMA(3,1,2)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  price   No. Observations:                  174
Model:               SARIMAX(3, 1, 3)   Log Likelihood                -720.591
Date:                Thu, 02 May 2024   AIC                           1457.182
Time:                        01:15:34   BIC                           1482.221
Sample:                    03-31-1956   HQIC                          1467.343
                         - 06-30-1999
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      3.4867      1.932      1.805      0.071      -0.300       7.273
ar.L1         -0.8563      0.048    -17.996      0.000      -0.950      -0.763
ar.L2         -1.0047      0.005   -222.365      0.000      -1.014      -0.996
ar.L3         -0.8555      0.048    -17.959      0.000      -0.949      -0.762
ma.L1         -0.2350      0.083     -2.840      0.005      -0.397      -0.073
ma.L2          0.8158      0.051     15.845      0.000       0.715       0.917
ma.L3         -0.1616      0.083     -1.952      0.051      -0.324       0.001
sigma2       292.0594     28.430     10.273      0.000     236.338     347.780
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                 8.09
Prob(Q):                              0.77   Prob(JB):                         0.02
Heteroskedasticity (H):               2.16   Skew:                            -0.19
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.00
===================================================================================
```

Figure 13: ARIMA(3,1,3)

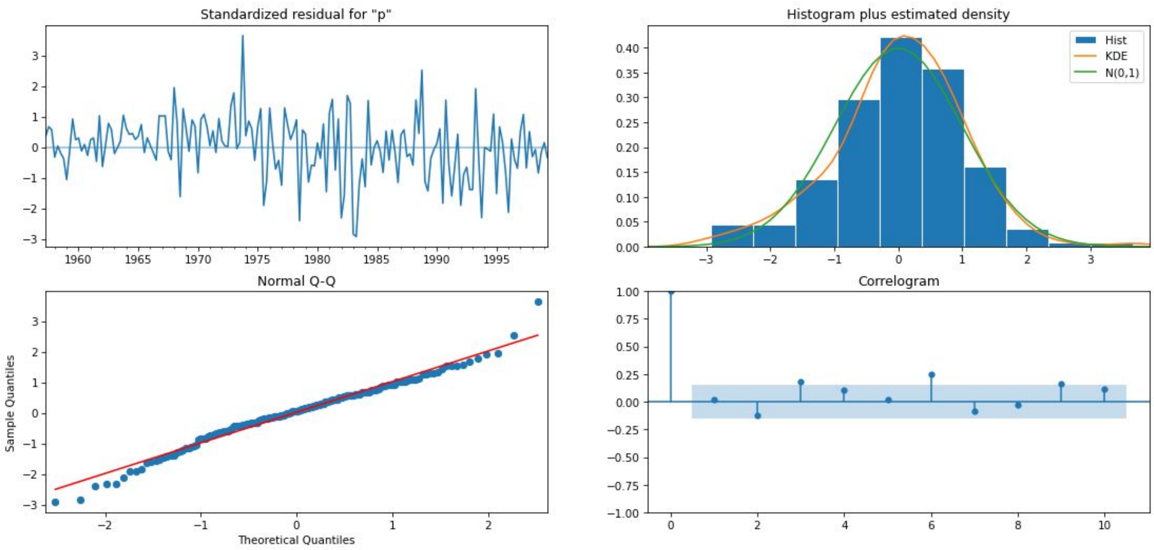

Figure 14: Model Adequacy checking for ARIMA (3,1,3)

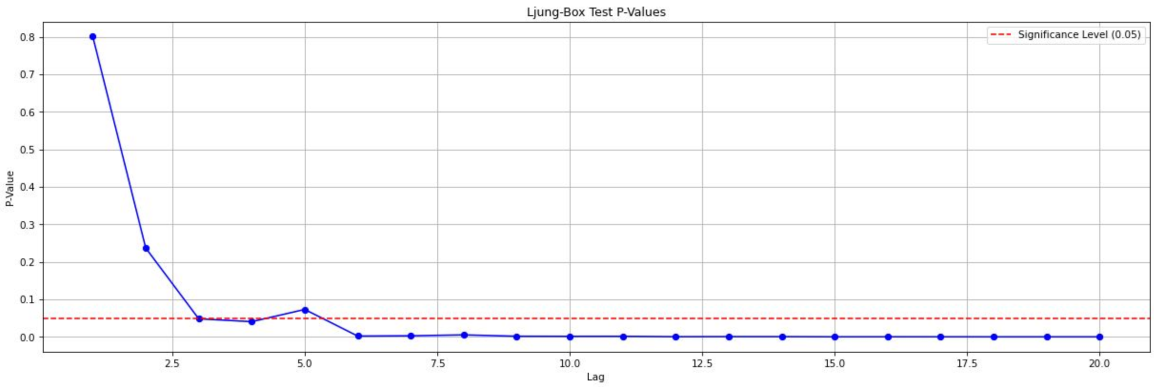

Figure 15: Ljung Box test for ARIMA(3,1,3

5.1 Forecasting using ARIMA(3,1,3)

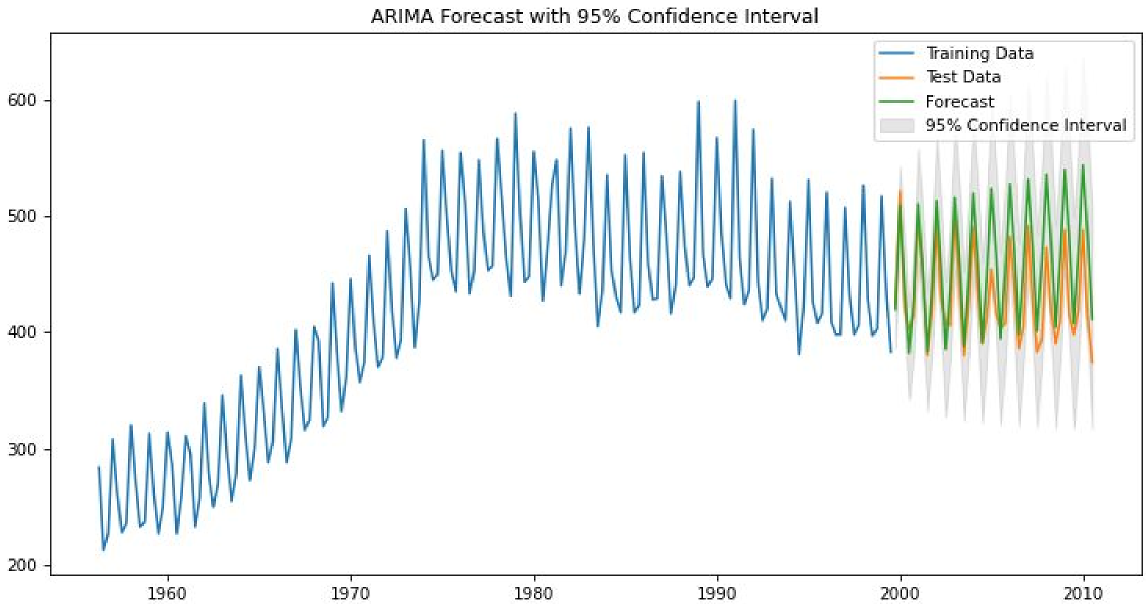

Figure 16: Forecasting using ARIMA(3,1,3)

```
Mean Squared Error (MSE): 1405.3869321572079
Mean Absolute Percentage Error (MAPE): 7.31 %
```

Figure 17: MSE of ARIMA model

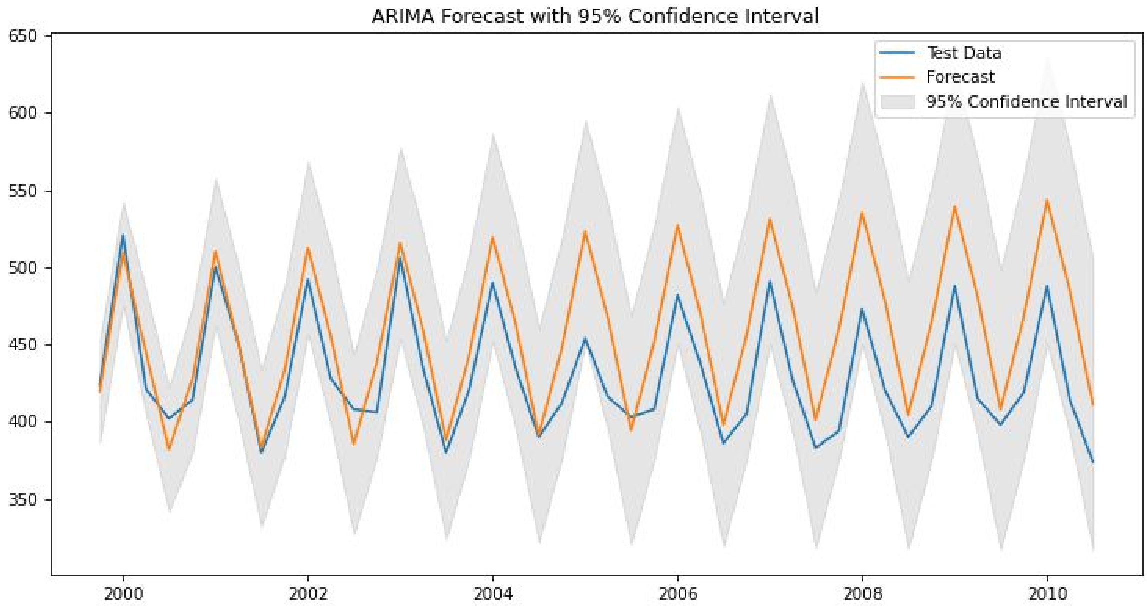

Figure 18: Magnified view

```
                                    SARIMAX Results
==========================================================================================
Dep. Variable:                               price   No. Observations:                174
Model:             SARIMAX(3, 1, 3)x(1, 1, [1], 4)   Log Likelihood               -669.488
Date:                             Thu, 02 May 2024   AIC                          1358.976
Time:                                     02:50:21   BIC                          1389.790
Sample:                                 03-31-1956   HQIC                         1371.487
                                      - 06-30-1999
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -0.2568      0.398     -0.645      0.519      -1.038       0.524
ar.L1         -1.3423      0.191     -7.017      0.000      -1.717      -0.967
ar.L2         -1.3185      0.165     -8.003      0.000      -1.641      -0.996
ar.L3         -0.3719      0.144     -2.588      0.010      -0.653      -0.090
ma.L1          0.3906      0.208      1.875      0.061      -0.018       0.799
ma.L2          0.4109      0.094      4.393      0.000       0.228       0.594
ma.L3         -0.4260      0.127     -3.364      0.001      -0.674      -0.178
ar.S.L4        0.1955      0.130      1.502      0.133      -0.060       0.451
ma.S.L4       -0.7975      0.070    -11.416      0.000      -0.934      -0.661
sigma2       234.7199     22.374     10.491      0.000     190.867     278.572
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                10.15
Prob(Q):                              0.92   Prob(JB):                         0.01
Heteroskedasticity (H):               2.40   Skew:                             0.00
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.23
===================================================================================
```

Figure 19: Summary of Sarima(3,1,3)(1,1,1)4

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                price   No. Observations:                174
Model:             SARIMAX(3, 1, 3)x(0, 1, [1, 2], 4)   Log Likelihood              -652.363
Date:                              Thu, 02 May 2024   AIC                         1324.726
Time:                                      02:50:21   BIC                         1355.288
Sample:                                  03-31-1956   HQIC                        1337.138
                                       - 06-30-1999
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -0.3049      0.467     -0.654      0.513      -1.219       0.610
ar.L1         -1.2835      0.206     -6.232      0.000      -1.687      -0.880
ar.L2         -1.2649      0.194     -6.520      0.000      -1.645      -0.885
ar.L3         -0.3250      0.159     -2.042      0.041      -0.637      -0.013
ma.L1          0.3263      0.220      1.480      0.139      -0.106       0.758
ma.L2          0.4075      0.093      4.393      0.000       0.226       0.589
ma.L3         -0.4505      0.122     -3.697      0.000      -0.689      -0.212
ma.S.L4       -0.5413      0.098     -5.549      0.000      -0.732      -0.350
ma.S.L8       -0.2086      0.100     -2.086      0.037      -0.405      -0.013
sigma2       235.6173     22.795     10.336      0.000     190.940     280.295
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 9.16
Prob(Q):                              0.97   Prob(JB):                         0.01
Heteroskedasticity (H):               2.19   Skew:                             0.08
Prob(H) (two-sided):                  0.01   Kurtosis:                         4.17
===================================================================================
```

Figure 20: Summary of Sarima(3,1,3)(0,1,2)4

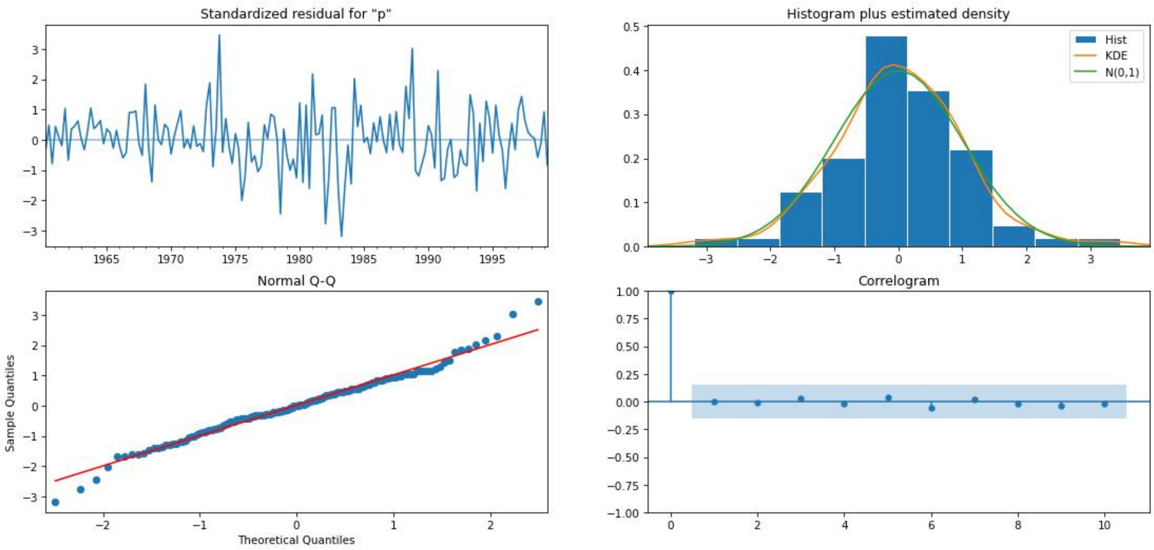

Figure 21: Model Adequacy for Sarima

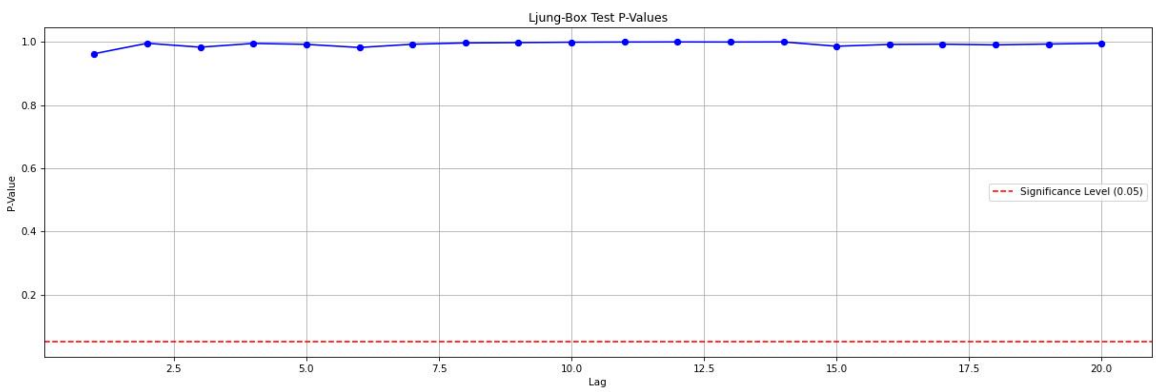

Figure 22: Ljung Box test for Sarima model

5.2 Forecasting using Sarima model

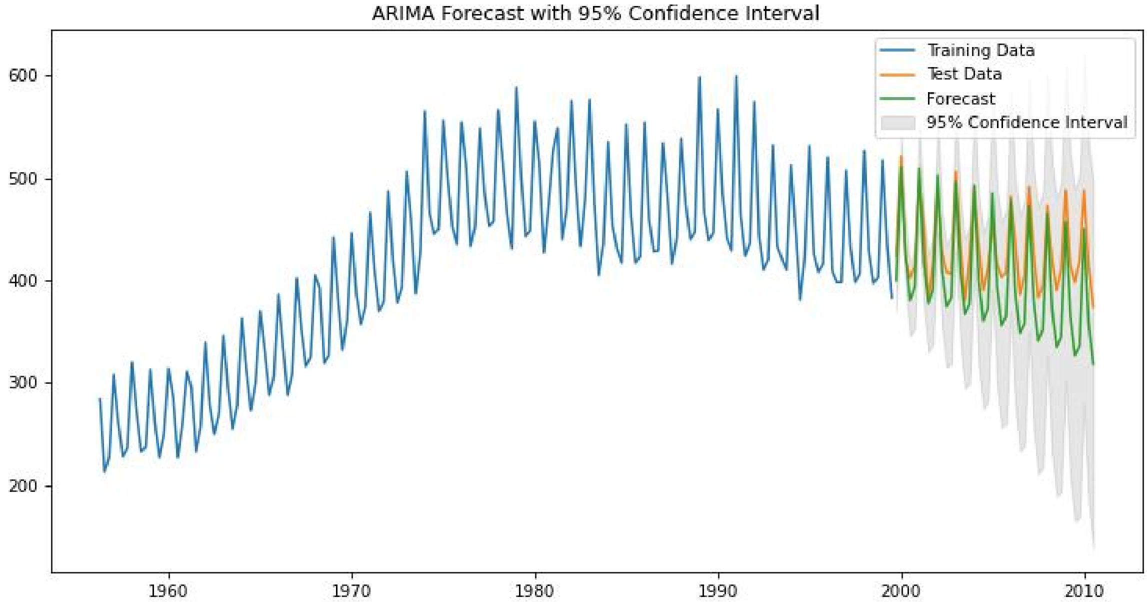

Figure 23: Forecasting using Sarima Model

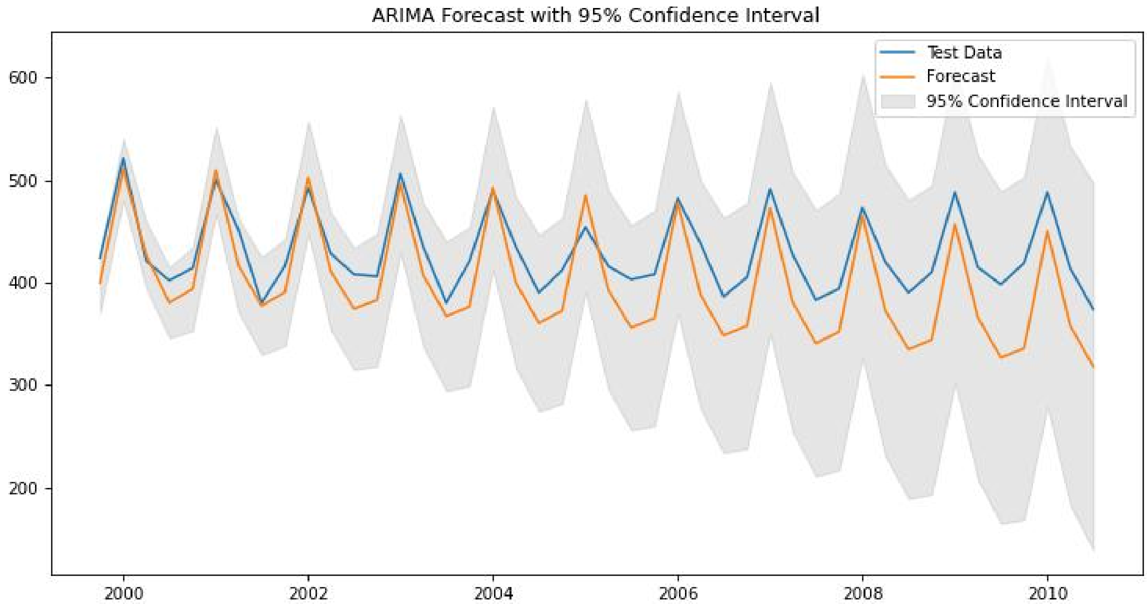

Figure 24: Magnified View

```
Mean Squared Error (MSE): 1427.8832585608397
Mean Absolute Percentage Error (MAPE): 7.82 %
```

Figure 25: MSE of forecasted values using Sarima Model

6 Conclusion

Utilized visualizations of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots on the first difference of the time series data to identify potential ARIMA models. Two promising ARIMA models, (3, 1, 2) and (3, 1, 3), were obtained through this process. Further analysis involved visualizing the ACF and PACF plots on the seasonal difference of the differenced time series data. This led to the identification of two seasonal ARIMA models, namely (3, 1, 3) with seasonal order (1, 1, 1, 4) and (3, 1, 3) with seasonal order (10, 1, 2, 4). The data was decomposed into trend-cycle, seasonal, and remainder error terms to better understand its underlying components. Assessment of model accuracy was performed using Ljung-Box p-value, Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE) metrics. These evaluations provided insights into the predictive performance of both ARIMA and seasonal ARIMA models.