Forecasting Beer Production in Australia for 1996

Applied Time Series Analysis-5306
Group Number 24
Team Members:

Kiran Ala – 1002170943
Vengaiah Chowdary Madineni-1002172698

Sarath Chintakunta- 1002186989

OUTLINES

- Introduction
- Data Exploration and Preparation
- Data Transformation
- EDA
- Data Splitting and Model Fitting
- Model Evaluation
- Model Evaluation

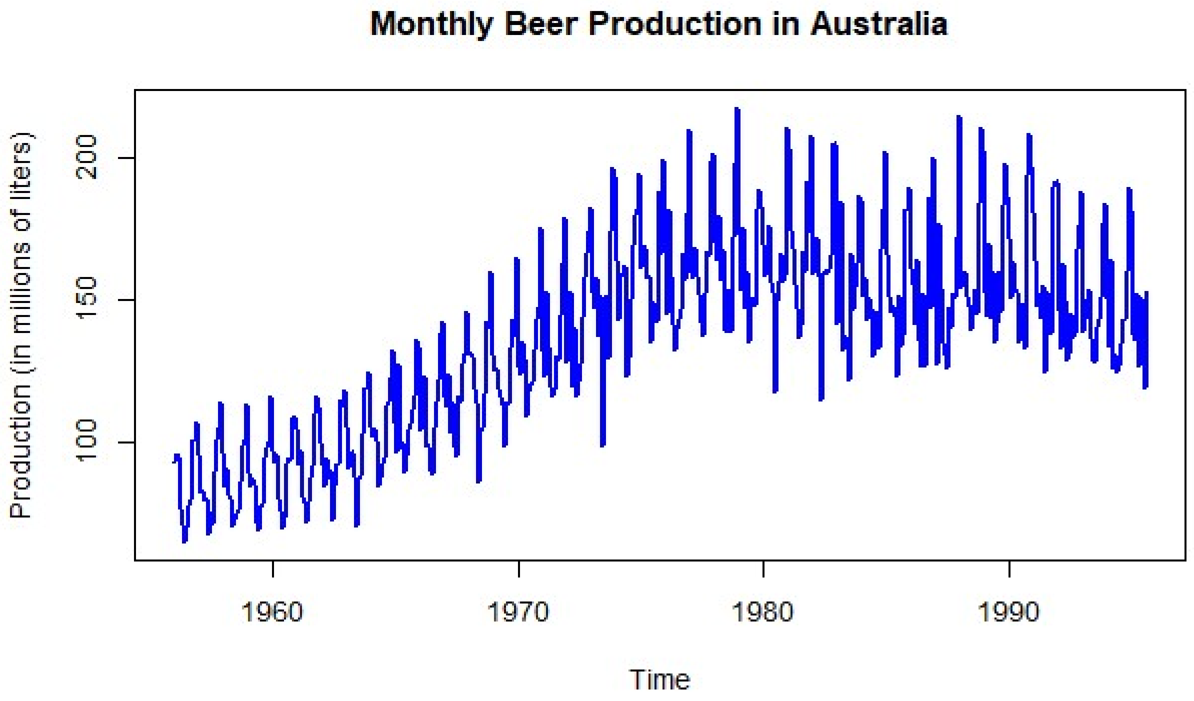
Introduction

This project aims to apply time series forecasting methods to predict monthly Australian beer production for the year 1996 accurately. Using historical data from 1956 to 1995, we strive to model trends, seasonality, and patterns in beer production to produce reliable forecasts.

Data Exploration and Preparation

1. **Data Loading and Initial Inspection**:
   - We loaded the dataset, which contains two columns: Month (indicating the month and year) and Monthly.beer.production (beer production in millions of liters).
   - Initial exploration using functions like head() and str() revealed 476 observations spanning from January 1956. The data showed an upward trend with clear seasonal patterns.
2. **Time Series Conversion**:
   - The data was converted into a time series object (ts) with a frequency of 12, representing monthly data.
   - This allowed us to visualize beer production trends and seasonal patterns over the decades.
   - **Title**: "Monthly Beer Production in Australia" for clear communication of the graph's purpose.
   - **X-Axis**: Represents the timeline from 1956 to the early 1990s.
   - **Y-Axis**: Represents beer production in millions of liters.

Observations from the Plot:



- **Upward Trend**: Beer production steadily increased from 1956 until the late 1970s, indicating growth in production over time.
- **Seasonality**: Clear cyclical patterns are visible, suggesting seasonal variations in beer production.

Decomposition of Time Series for Monthly Beer Production

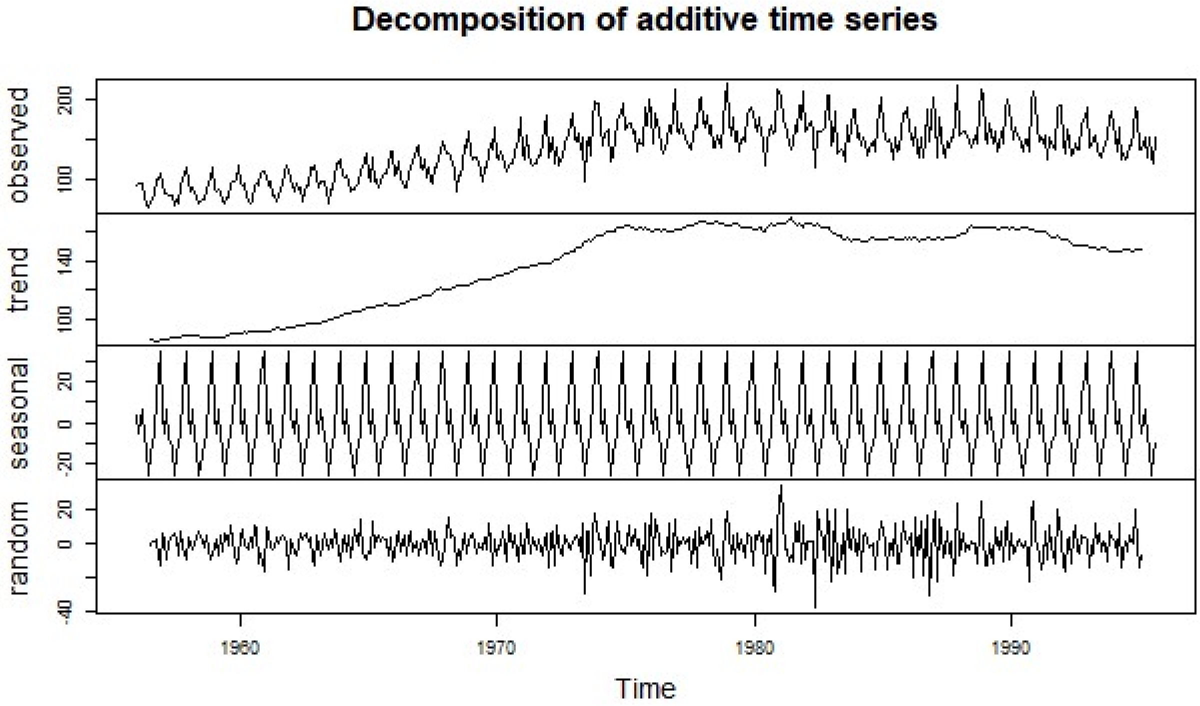
To gain deeper insights into the components of Australian monthly beer production over time, we performed a time series decomposition using an additive model.

The **Observed** plot shows.

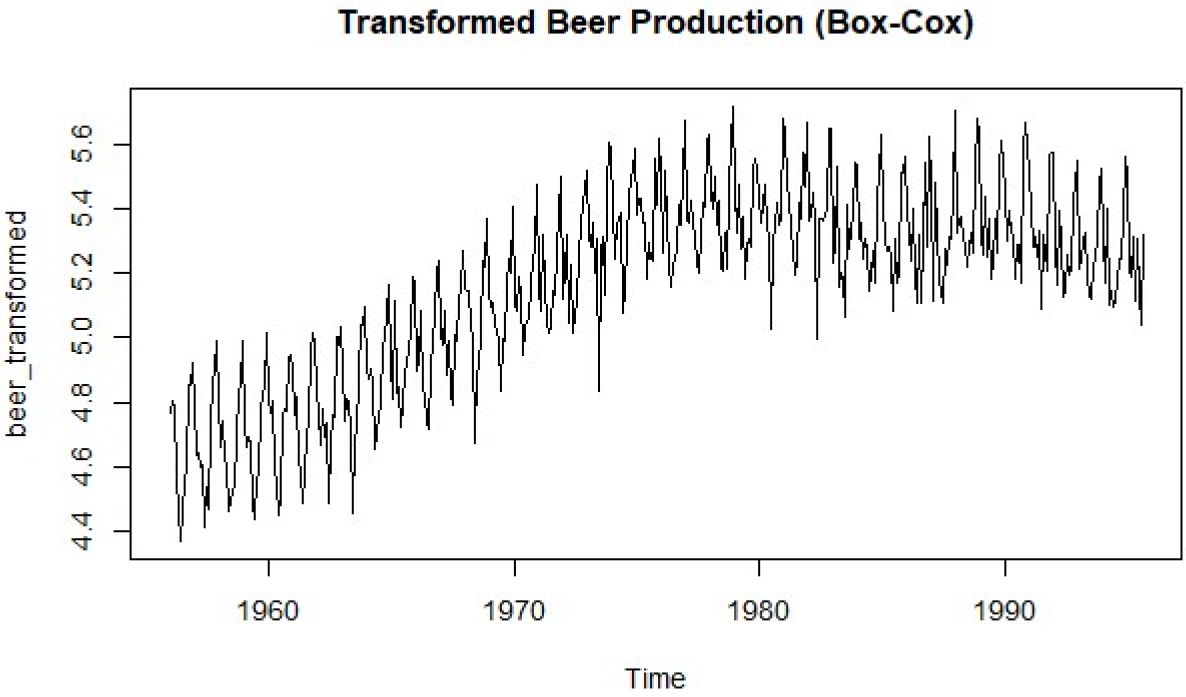
- The **Trend** component shows steady growth in beer production from the 1950s to the late 1970s, followed by stabilization and a slight decline in the 1990s.
- The **Seasonal** component remains consistent throughout, indicating strong, recurring seasonal

patterns.

- The **Random** component displays irregular variations, likely due to external factors such as economic shifts or unforeseen events.



Data Transformation and Stationarity



1. **Box-Cox Transformation**:
   To improve the quality of our analysis and prepare the dataset for further statistical modeling, we applied the **Box-Cox transformation**

- To stabilize the variance in the data, we applied the Box-Cox transformation. The optimal lambda was calculated as approximately 0.022, which ensured that the transformed data satisfied the stationarity requirements.

2. **Differencing**:
   - First-order differencing was performed to remove the trend, and seasonal differencing was applied to eliminate recurring seasonal patterns.
   - The transformed series showed constant mean and variance, making it stationary and ready for modeling.
3. **Stationarity Testing**:
   - We conducted the Augmented Dickey-Fuller (ADF) test, which confirmed the stationarity of the seasonally differenced series. The test statistic (-15.987) and p-value (0.01) supported the rejection of the null hypothesis of non-stationarity.

Exploratory Analysis of Lag Relationships

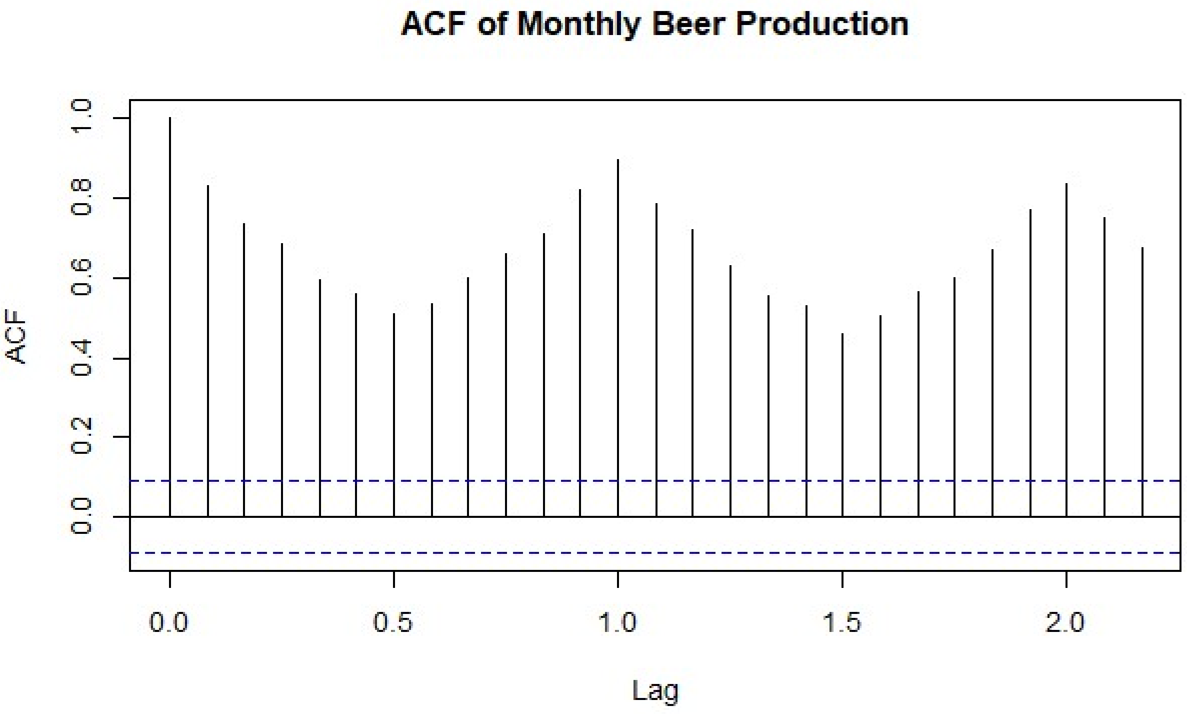
We took a period of 24 Months where the interval from 0.0 to 1.0 represents 12 months where each lag representing a month.

Since the data shows monthly beer production, we used a seasonal lag of 12 months to capture yearly patterns in the data. Beer production often increases during certain times of the year, like summer or holidays. The ACF plot shows spikes at lags of 12, 24, and so on, confirming this seasonality. Including this lag was important for building an accurate model and improving forecast reliability.

1)Autocorrelation Function (ACF) Plot

To better understand the structure and dependencies in the time series data, we conducted an Autocorrelation Function (ACF)

- The **Autocorrelation Function (ACF)** plot showed significant correlations at seasonal lags (multiples of 12 months), indicating strong seasonal dependencies.



- We used the acf() function to compute and plot the autocorrelation coefficients for the time series data (AUS_beer_ts) at various lags.

- The horizontal blue lines in the plot represent the confidence intervals. Any bars extending beyond these lines indicate statistically significant autocorrelation at those lags.

Lag Structure:

- The ACF plot shows a strong correlation at lags that correspond to 12 months (one year), indicating clear seasonality in beer production.
- Significant correlations at smaller lags suggest that recent months' beer production is strongly influenced by the immediately preceding months.

Seasonality:

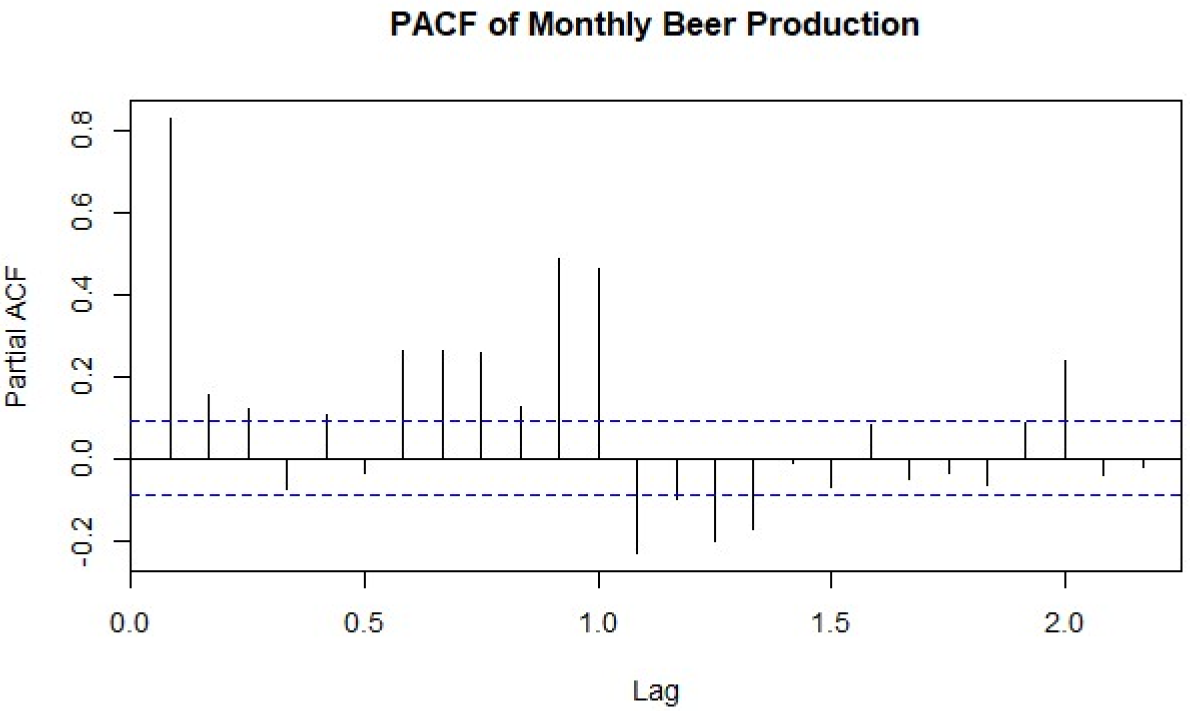
- The regular, periodic spikes in the ACF plot confirm a consistent seasonal pattern in beer production, which aligns with expected trends in demand based on seasons.

Persistence:

- High autocorrelation at lower lags implies that the time series has a memory effect, meaning past production values influence future values.

2) Partial Autocorrelation Function (PACF) Plot

- The **Partial Autocorrelation Function (PACF)** plot revealed significant lag relationships, particularly at lower-order lags, supporting the presence of both seasonal and non-seasonal components in the data.



As with the ACF plot, **blue horizontal lines** represent confidence intervals. Bars that extend beyond these lines indicate significant partial autocorrelations at those lags.

To complement the analysis of the autocorrelation function (ACF), we computed and plotted the **Partial Autocorrelation Function (PACF).** This step provided additional insights into the relationships between lags in the data.

Significant Lags:

- The PACF plot shows significant spikes at the first few lags, suggesting that these lags have a direct and strong influence on the current month's beer production.
- The sharp drop in partial autocorrelations after the first few lags indicates that higher-order lags

contribute less directly to the current value.

Seasonal Patterns:

- Similar to the ACF, periodic spikes in the PACF suggest seasonal relationships in the data, though the influence decreases with higher lags.

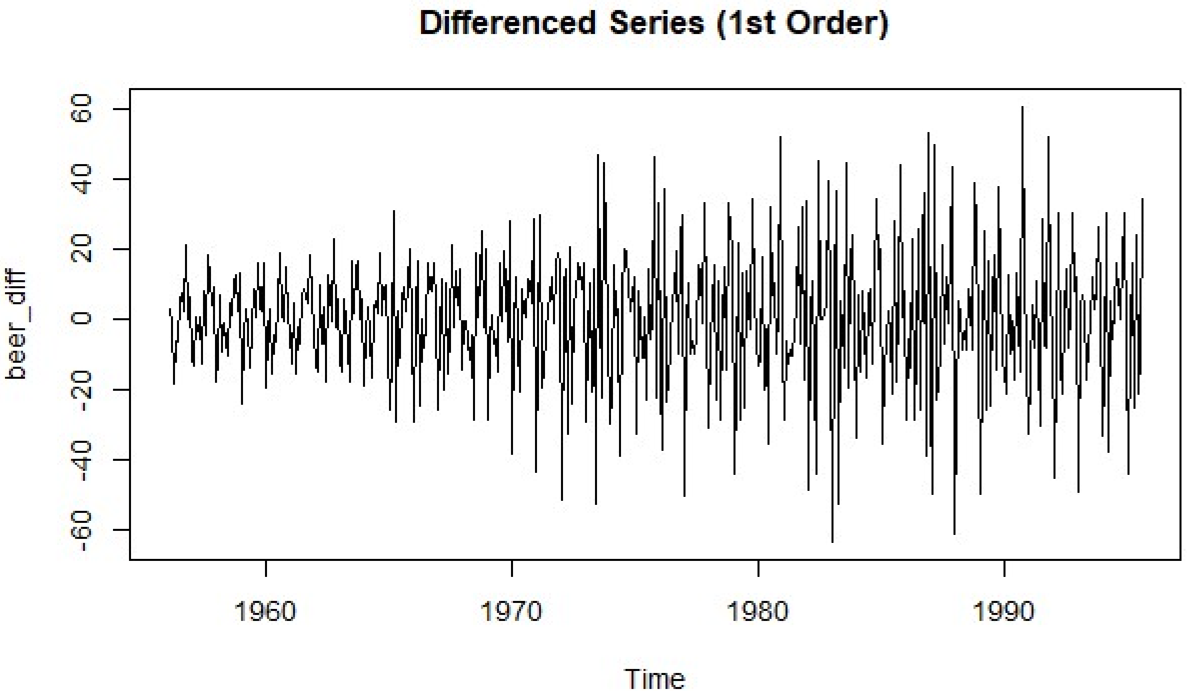
Lag 1 Importance:

- The first lag has the strongest partial autocorrelation, indicating a dominant influence of the previous month on the current month's production.

Comparison to ACF:

- While the ACF showed the cumulative influence of past lags, the PACF isolated the direct relationships. Together, these plots provide a complete picture of the temporal dependencies in the data.

Differencing the Time Series to Achieve Stationarity



To prepare the data for modeling, we applied first-order differencing. This transformation addresses non-stationarity in the data, which is a key requirement for many time series forecasting models.
The above plot of the differenced series was created, titled **"Differenced Series (1st Order)"**, to visually assess the stationarity of the transformed data.

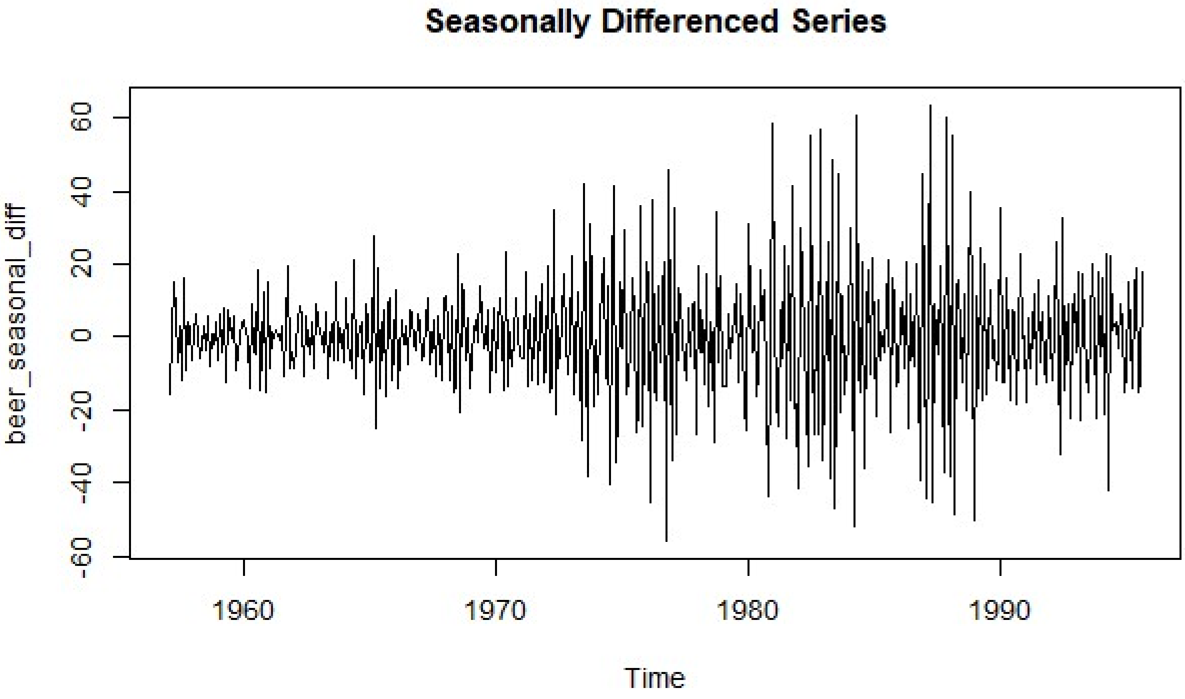
Differenced Series:

- The plot of the differenced series shows that the long-term trend has been removed, leaving behind fluctuations around a relatively constant mean.
- The variance appears more stable, indicating progress toward achieving stationarity.

Fluctuations:

- The differenced series captures short-term variations in beer production, making it suitable for further analysis.

Seasonal Differencing of the Time Series



We applied seasonal differencing to the already differenced series.

Seasonal Differencing:

- We applied the diff() function with a **lag of 12**, as the data is monthly, and the season repeats every 12 months.

The plot of the seasonally differenced series shows:

- Fluctuations around a consistent mean, with reduced seasonal patterns.
- A relatively stable variance compared to the original or first-order differenced series.

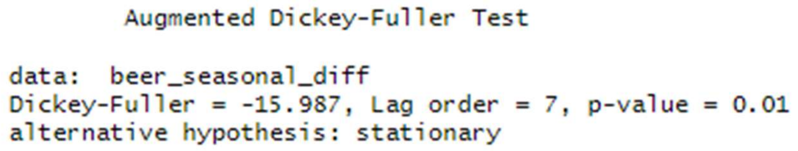
The data now appears more stationary, making it better suited for modeling.

Rechecking Stationarity Using the Augmented Dickey-Fuller (ADF) Test

After applying seasonal differencing to the time series, we used the Augmented Dickey-Fuller (ADF) Test to confirm that the transformed series is now stationary

- We loaded the tseries package to use the adf.test() function, which performs the Augmented Dickey-Fuller Test.
- The test was applied to the **seasonally differenced series (beer_seasonal_diff)**.

Output Interpretation:

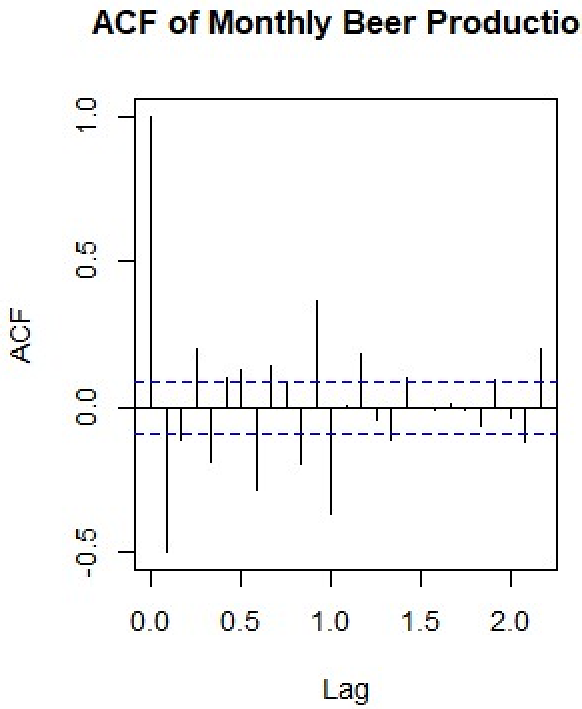


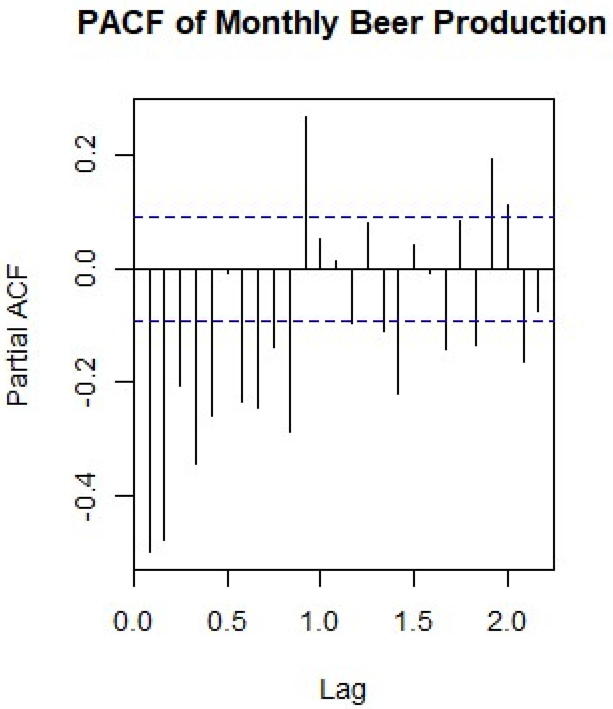
```
	Augmented Dickey-Fuller Test

data:  beer_seasonal_diff
Dickey-Fuller = -15.987, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

P-Value:

- The p-value is **0.01**, which is less than the commonly used threshold of 0.05. This indicates strong evidence against the null hypothesis (non-stationarity).
- The test confirms that the seasonally differenced series is stationary, with no significant trends or seasonal patterns remaining.

Reanalysis of ACF and PACF for Seasonally Differenced Series





After confirming that the seasonally differenced time series is stationary using the Augmented Dickey-Fuller (ADF) test, we revisited the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. These plots are essential for understanding the lag relationships in the stationary data and guiding model selection.

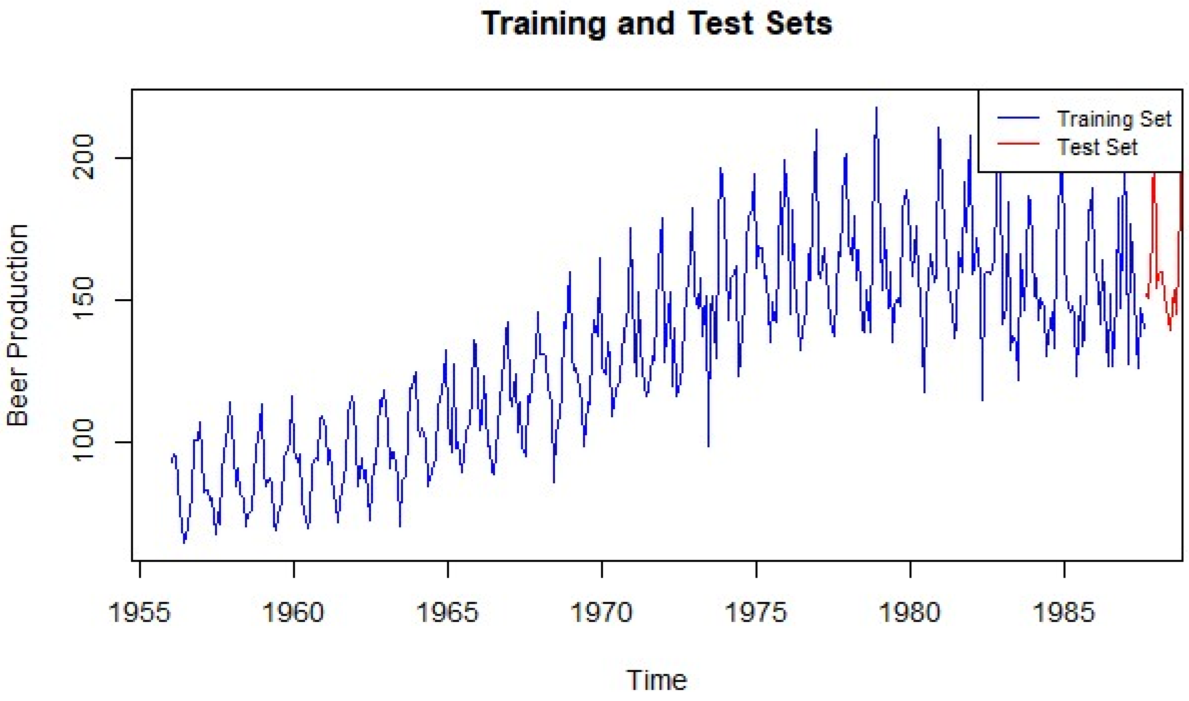
ACF Plot:

- The ACF shows significant spikes at specific lags, particularly at seasonal lags (multiples of 12 for monthly data). These spikes suggest the presence of seasonal moving average (SMA) components.
- The remaining lags die off quickly, which is a sign of stationarity.

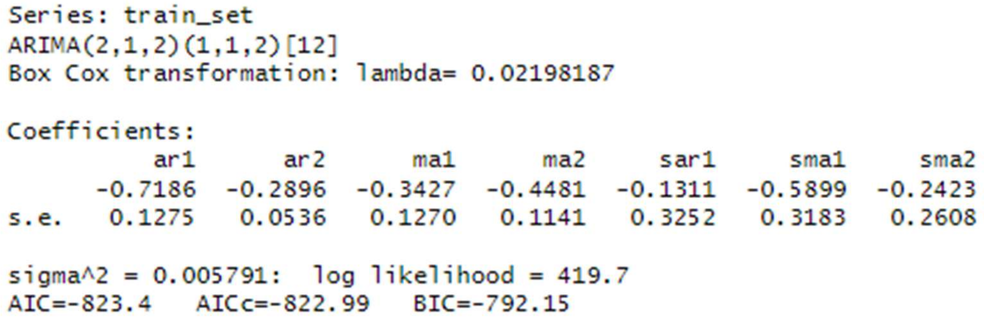
PACF Plot:

- The PACF shows significant spikes at lower lags, indicating the influence of non-seasonal and seasonal autoregressive (AR) components.
- The gradual tapering off at higher lags further supports the presence of seasonality.

Data Splitting and Model Fitting



1. **Data Splitting**:
   - The dataset we split into a **training set** (80% of the data) and a **test set** (20%). The training set was used to fit the models, and the test set was reserved for model evaluation.
   - A visual plot distinguished the training data (blue) and test data (red) for clarity.
2. **Model Fitting**:



```
Series: train_set
ARIMA(2,1,2)(1,1,2)[12]
Box Cox transformation: lambda= 0.02198187

Coefficients:
          ar1      ar2      ma1      ma2     sar1     sma1     sma2
      -0.7186  -0.2896  -0.3427  -0.4481  -0.1311  -0.5899  -0.2423
s.e.   0.1275   0.0536   0.1270   0.1141   0.3252   0.3183   0.2608

sigma^2 = 0.005791:  log likelihood = 419.7
AIC=-823.4   AICc=-822.99   BIC=-792.15
```

   - **SARIMA Model**:
     - Using auto.arima(), a SARIMA model was selected as ARIMA(2,1,2)(1,1,2)[12].
     - The model captured both non-seasonal (AR and MA terms) and seasonal components effectively.
     - A manual SARIMA model (1,1,1)(1,1,1)[12] was also tested to compare performance.

The SARIMA model **(2,1,2)(1,1,2)[12]** combines non-seasonal and seasonal components to handle both trends and yearly patterns in the data:

Non-Seasonal Part ((2,1,2)):

2 AR terms: Uses the last two observations for predictions.

1 differencing: Removes trends in the data.

2 MA terms: Accounts for past forecast errors.

Seasonal Part ((1,1,2)[12]):

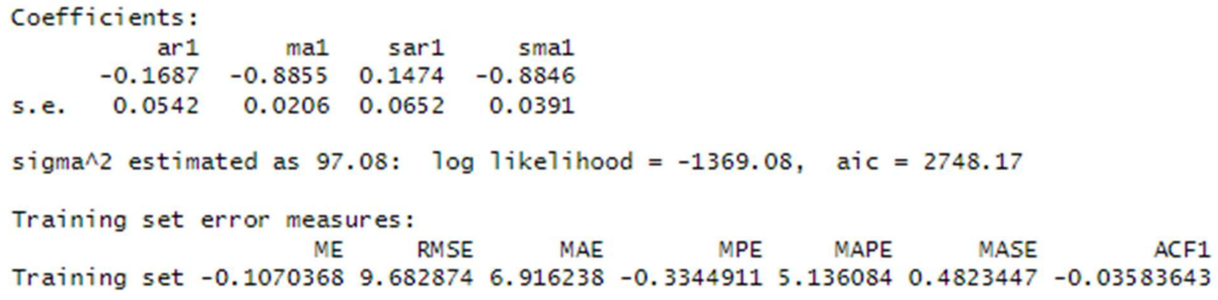
1 SAR term: Considers one observation from the same month in the previous year.

1 seasonal differencing: Removes yearly patterns.

2 SMA terms: Incorporates past seasonal forecast errors.

[12]: Specifies the seasonal frequency (12 months in a year).

Fitting a SARIMA Model with Updated Parameters



```
Coefficients:
          ar1      ma1    sar1     sma1
      -0.1687  -0.8855  0.1474  -0.8846
s.e.   0.0542   0.0206  0.0652   0.0391

sigma^2 estimated as 97.08:  log likelihood = -1369.08,  aic = 2748.17

Training set error measures:
                     ME     RMSE      MAE        MPE     MAPE      MASE        ACF1
Training set -0.1070368 9.682874 6.916238 -0.3344911 5.136084 0.4823447 -0.03583643
```

We manually specified and fitted a SARIMA model with updated parameters to compare its performance and validate its suitability for the data.

Coefficients:

- The coefficients for the AR (ar1), MA (ma1), SAR (sar1), and SMA (sma1) terms were estimated. Each coefficient has an associated standard error (s.e.), which indicates the variability of the estimate.

Model Fit:

- The log likelihood and AIC values represent the model's goodness of fit. A lower AIC value indicates a better model.
- This model had an AIC of **2748.17**.

Error Metrics:

The model's error metrics were:

- RMSE (Root Mean Square Error): 9.68
- MAE (Mean Absolute Error): 6.91
- MAPE (Mean Absolute Percentage Error): 5.14%

These metrics show how well the model performed on the training data. A lower value indicates better performance.

- **ETS Model**:
  - The Exponential Smoothing (ETS) model ETS(A,A,A) was fitted, which uses additive components for error, trend, and seasonality.
  - The model's smoothing parameters and initial states were optimized to match the data's characteristics.

Smoothing Parameters:

- **Alpha (α)**: 0.093 (controls the smoothing of the level component).
- **Beta (β)**: 0.0058 (controls the smoothing of the trend component).
- **Gamma (γ)**: 0.0001 (controls the smoothing of the seasonal component).

Error Metrics:

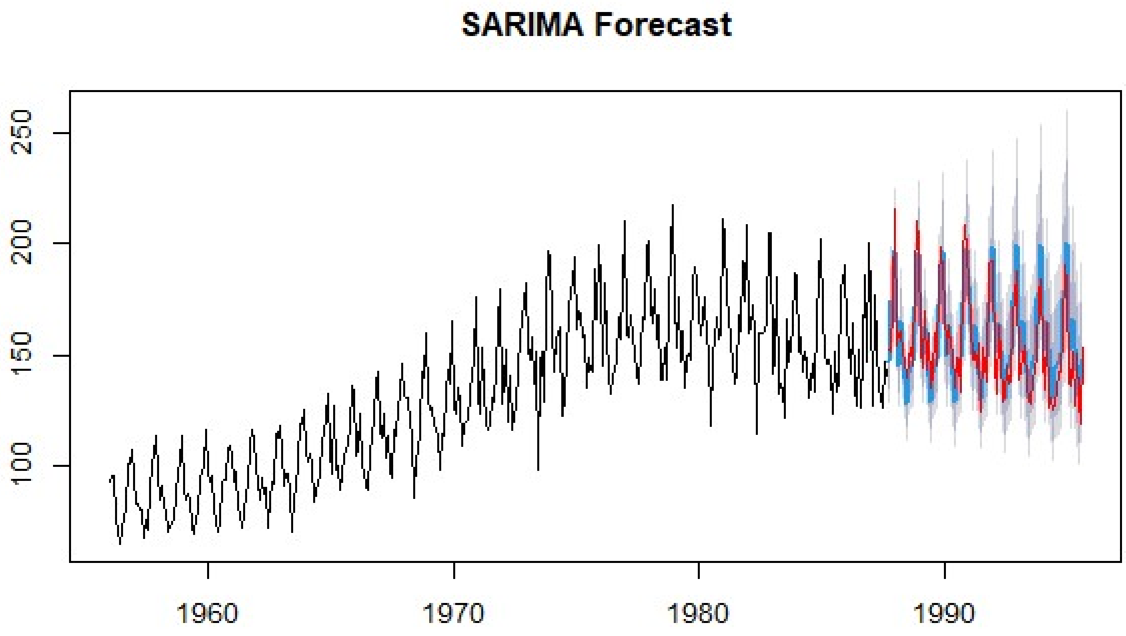
- RMSE: 9.45
- MAE: 6.96
- MAPE: 5.24%
- These metrics indicate how well the model captures the patterns in the training data.

Model Evaluation

Evaluating SARIMA and ETS Models by Forecasting and Comparing Predictions.

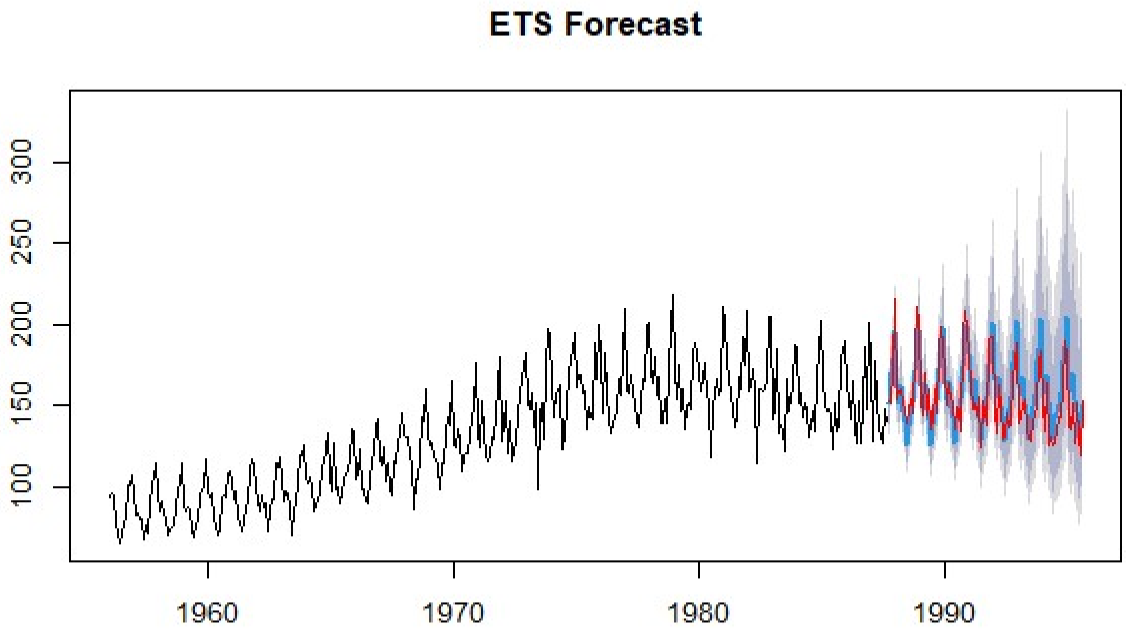
After fitting both SARIMA and ETS models to the training data, we evaluated their performance by forecasting the values for the test set.

SARIMA Forecast Plot:



- The SARIMA model's predictions align closely with the actual test set values, capturing both the trend and seasonal patterns.
- The forecast intervals widen as we move further into the future, reflecting increased uncertainty.

ETS Forecast Plot:



- The ETS model also captures the trend and seasonality but appears to produce wider prediction intervals compared to SARIMA, particularly for long-term forecasts.

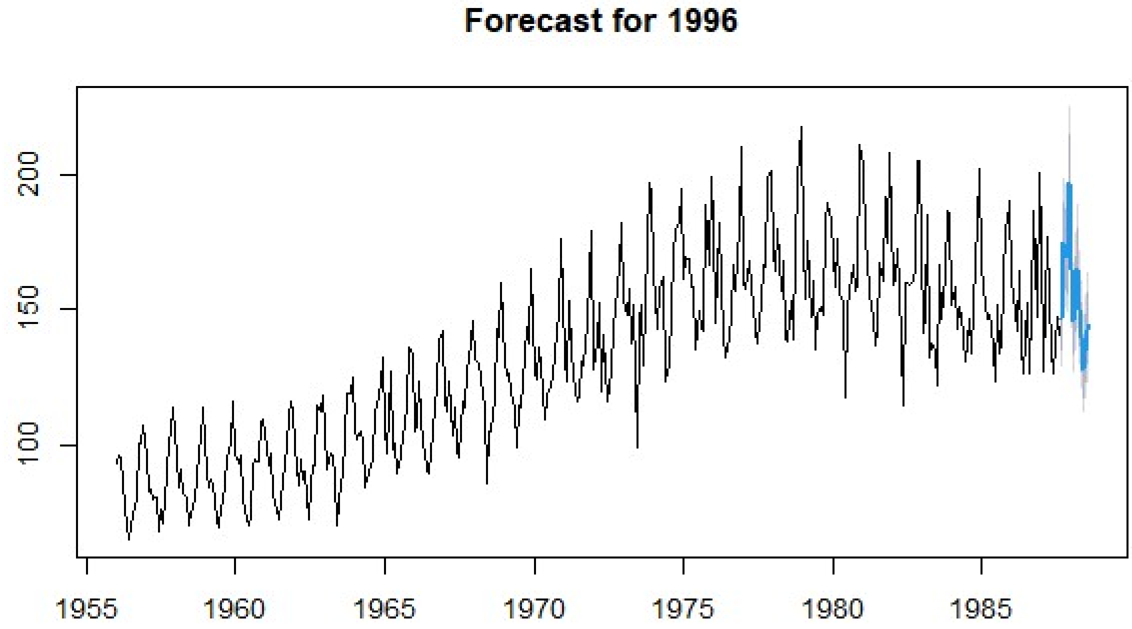
- While it predicts seasonal patterns, the alignment with the test set values is slightly less accurate than SARIMA.

The SARIMA and ETS models both captured the trend and seasonality in the beer production data, as shown in the forecasts. However, the SARIMA model appeared to have a closer alignment with the test set values and narrower prediction intervals.

COMPARISION

1. **Forecasting and Visualization**:
   - Both SARIMA and ETS models were used to forecast the test set period. The forecasts were plotted, and actual test set values were overlaid for comparison:
     - The **SARIMA model** captured the trend and seasonal patterns accurately with narrower prediction intervals.
     - The **ETS model** also performed well but showed slightly wider prediction intervals, indicating less confidence in long-term predictions.
2. **RMSE Comparison**:
   - The RMSE (Root Mean Square Error) was calculated for both models:
     - SARIMA RMSE: **13.11**
     - ETS RMSE: **13.79**
   - The lower RMSE of SARIMA confirmed it as the more accurate model for beer production forecasting.

Forecasting for 1996



1. **Using the Best Model**:
   - Based on the evaluation, the SARIMA model was selected as the best-performing model for forecasting future values.
   - A 12-month forecast for 1996 was generated using the SARIMA model. The forecast plot showed:
     - Predicted values (blue line) continuing the trend and seasonal patterns.
     - Prediction intervals, reflecting uncertainty, widened further into the future.
2. **Insights from the Forecast**:

- The SARIMA forecast provided realistic and reliable projections for beer production in 1996, showing periodic peaks and troughs consistent with historical patterns.

Conclusion

1. **Key Findings**:
    - The beer production dataset exhibited clear trends and strong seasonality, effectively modeled using SARIMA and ETS approaches.
    - The SARIMA model outperformed the ETS model based on RMSE and prediction accuracy.
2. **Recommendations**:
    - The SARIMA model should be used for future forecasting tasks as it provided better accuracy and narrower prediction intervals.
    - The model can aid in production planning, inventory management, and understanding seasonal demand variations.