Name: Triveni Bhaskar

SMC ID: 1026789

Final Project Report: Predicting Beer Production

Introduction

The objective of this project is to analyse and forecast monthly beer production using time series analysis techniques. The dataset utilized comprises historical records of monthly beer production over a specific period.

Objective: The Beer production dataset provides a time series data for monthly beer production in Australia, for the period January 1956 – December 1993. The objective of this project is to analyse and forecast monthly beer production using advanced time series analysis techniques. Leveraging a comprehensive dataset spanning from January 1956 to December 1993, encompassing historical records of beer production in Australia, our aim is to unravel patterns, trends, and seasonality inherent in the data. By employing robust statistical methods and models, we seek to not only understand the underlying dynamics driving beer production but also to develop accurate forecasts for future production trends. This project holds significant relevance for stakeholders in the brewing industry, enabling informed decision-making and strategic planning based on reliable predictive insights derived from extensive historical data.

Methodology

Data Preprocessing

- Initially, the dataset was imported using the readxl package in R, and an overview of the data was obtained.
- Missing values, if any, were identified and subsequently removed using the ‘na.omit’ function to ensure data integrity.

Importing the Dataset

The first step in data preprocessing is importing the dataset into the chosen analytical environment. In this case, the dataset named "Beer Production" was imported using the read_excel function from the readxl package in R. This function reads data from Excel files and stores it as a data frame. The path to the Excel file is provided within the function call. After importing, it's essential to obtain an overview of the data structure, column names, and sample entries to understand the dataset's characteristics.

Checking for Missing Values

Missing values within the dataset can affect the accuracy and reliability of subsequent analyses. Therefore, it's crucial to identify and handle missing values appropriately. In this step, the presence of missing values in the "Beer Production" dataset was assessed using the is.na() function in R. This function generates a logical vector indicating the presence of NA (missing) values in each column. The sum() function was then applied to count the total number of missing values across all columns. The output indicated that there were 0 missing values present in the dataset.

Handling Missing Values

To maintain data integrity and completeness, missing values were handled by removing the corresponding rows from the dataset. This was accomplished using the na.omit() function in R, which removes any rows containing missing values from the data frame. By applying this function to the "Beer Production" dataset, rows with missing values were excluded, ensuring that the dataset was free from incomplete entries.

After completing these data preprocessing steps, the "Beer Production" dataset was ready for further analysis and modelling, with missing values handled appropriately and temporal information converted into a suitable format for time-series analysis.

Conversion of 'Month' Column

The 'Month' column in the "Beer Production" dataset was transformed into a Date object. Each 'Month' value was combined with '01' to represent the first day of the month. This conversion to the "%Y-%m-%d" format facilitates time-series analysis, enabling accurate trend identification and forecasting.

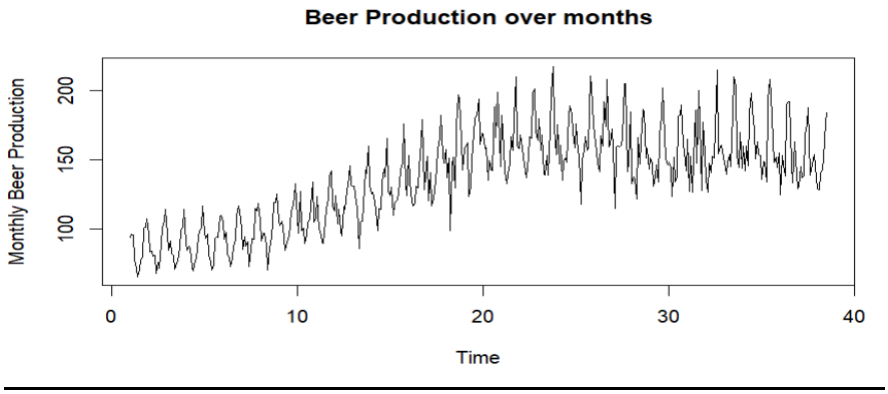
The 'Monthly beer production' data from the "Beer Production" dataset was transformed into a time-series object named beer_ts. The ts() function in R was utilized for this purpose, with the specified start parameter set to the first year (1) and first month (1), and a frequency of 12, indicating monthly data.

This transformation enables the analysis of beer production trends over time in a structured time-series format. By converting the data into a time-series object, it becomes suitable for various time-series analysis techniques such as trend analysis.

Exploratory Data Analysis

- The data was visualized using both traditional plot and histogram to understand the distribution and trends.
- Simple Moving Averages (SMA) were calculated to smoothen the data and visualize trends more clearly.

- The data was visualized using both traditional plot and histogram to understand the distribution and beer production over the months.

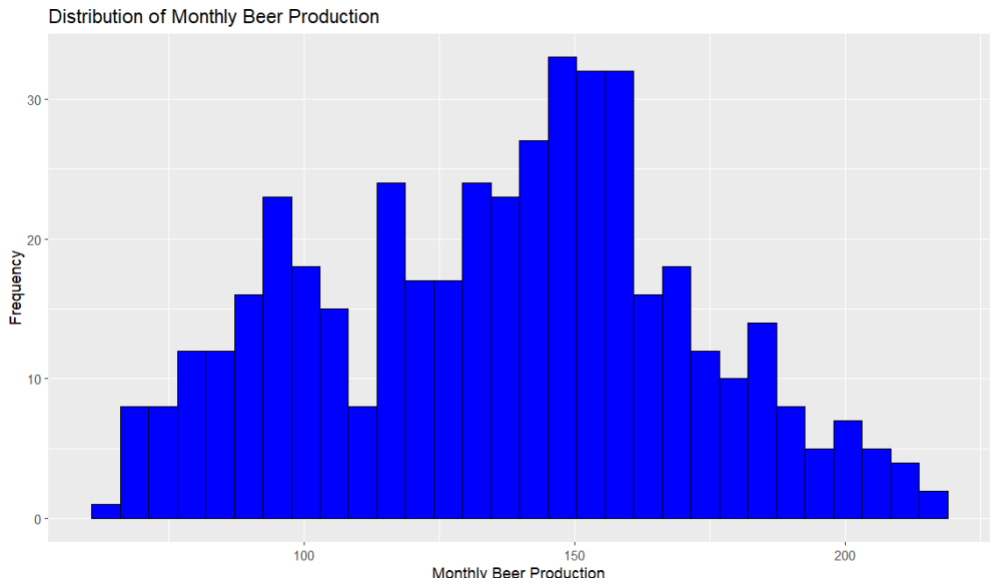


- **Histogram:** Based on the histogram, we observe that the distribution of monthly beer production is somewhat **skewed to the right**. This means that there are more months with lower beer production values compared to months with higher production values. Here's a description based on the visualization:

- The histogram titled "Distribution of Monthly Beer Production" illustrates the frequency distribution of monthly beer production values. The x-axis represents the monthly beer production, while the y-axis indicates the frequency of occurrence.

- The histogram displays a right-skewed distribution, indicating that the majority of months have lower beer production values, with a tail extending towards higher production values. This suggests that while there are months with relatively high beer production, they are less frequent compared to months with lower production.

- The skewness to the right suggests that there might be some months where external factors such as seasonal variations, economic conditions, or other factors affect beer production negatively, leading to lower production values. Conversely, there could be periods of increased production due to factors like festive seasons, promotional campaigns, or changes in consumer demand.



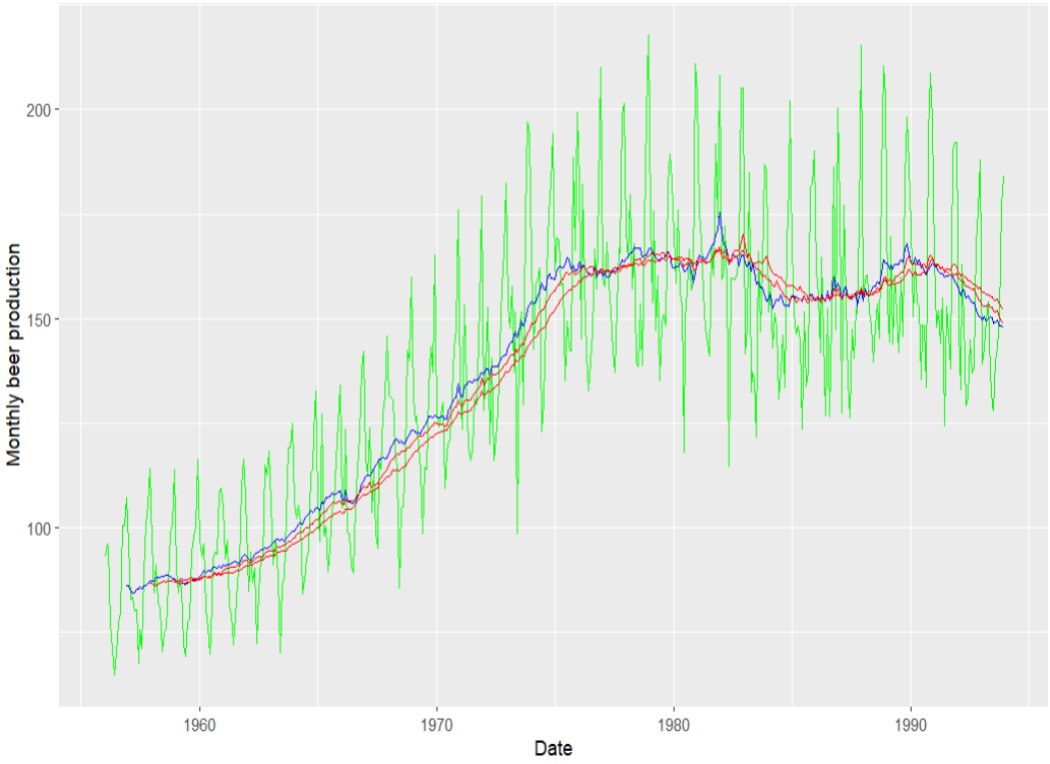
Simple Moving Averages -SMA

- The analysis involves calculating Simple Moving Averages (SMA) for beer production with window sizes of 12, 24, and 36 months. These moving averages are computed to smooth out fluctuations in the monthly production data and reveal underlying trends over time.

- In the plotted graph, the original monthly beer production data is depicted by a green line. Additionally, three SMA lines are overlaid on the plot: SMA_12 (blue), SMA_24 (red), and SMA_36 (red). These lines represent the smoothed trends in production over the specified window periods.

- Observing the graph, it's evident that as the window size increases, the SMA lines provide a clearer depiction of the underlying trend in beer production. This suggests that longer window sizes capture broader trends while reducing the impact of short-term fluctuations.

- Overall, the utilization of SMA and subsequent visualization using ggplot2 has provided valuable insights into the temporal behaviour of beer production. Moving forward, employing advanced modelling techniques like ARIMA can enhance our understanding and forecasting capabilities in this domain.



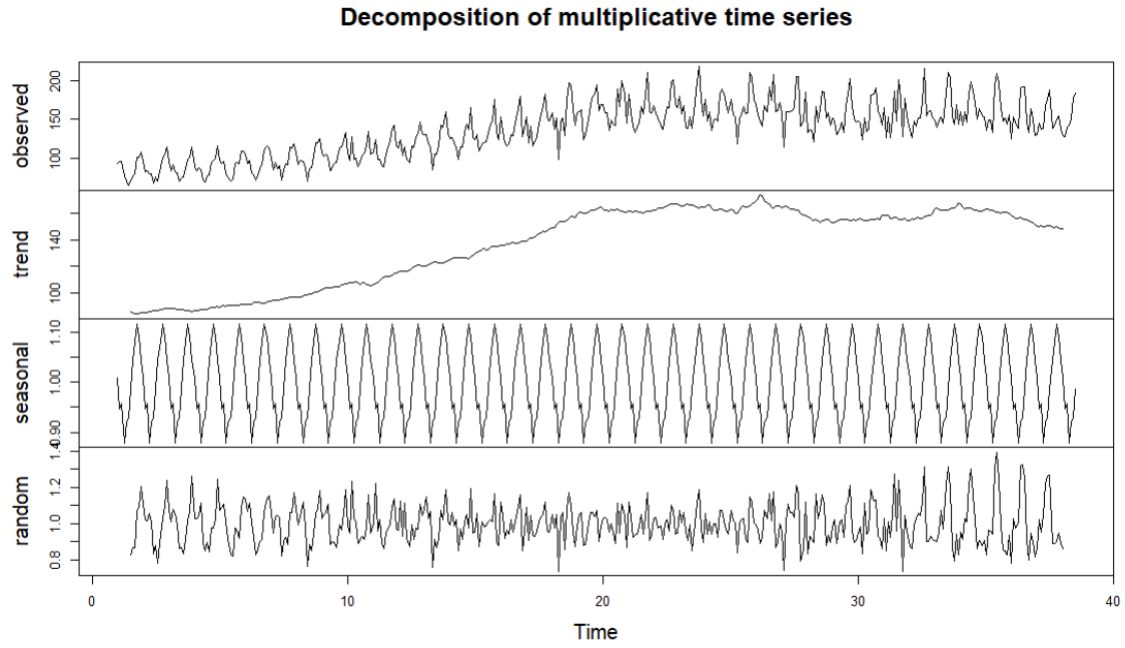
Stationarity Check

- The Augmented Dickey-Fuller (ADF) test was conducted to check for stationarity. Since the p-value was less than 0.05, indicating stationarity, further analysis proceeded.

Seasonal Decomposition

- Multiplicative seasonal decomposition was performed to understand the trend, seasonality, and residuals.
- Seasonal decomposition analysis is conducted on the beer production time series using a multiplicative model. This method separates the time series into its constituent components: trend, seasonal, and residual. The resulting plot visualizes these components, providing insights into the seasonal patterns and trends present in the beer production data. This analysis aids in understanding the underlying structure of the time series, facilitating better forecasting and decision-making processes in production planning and resource allocation.

Output:



Model Selection and Forecasting

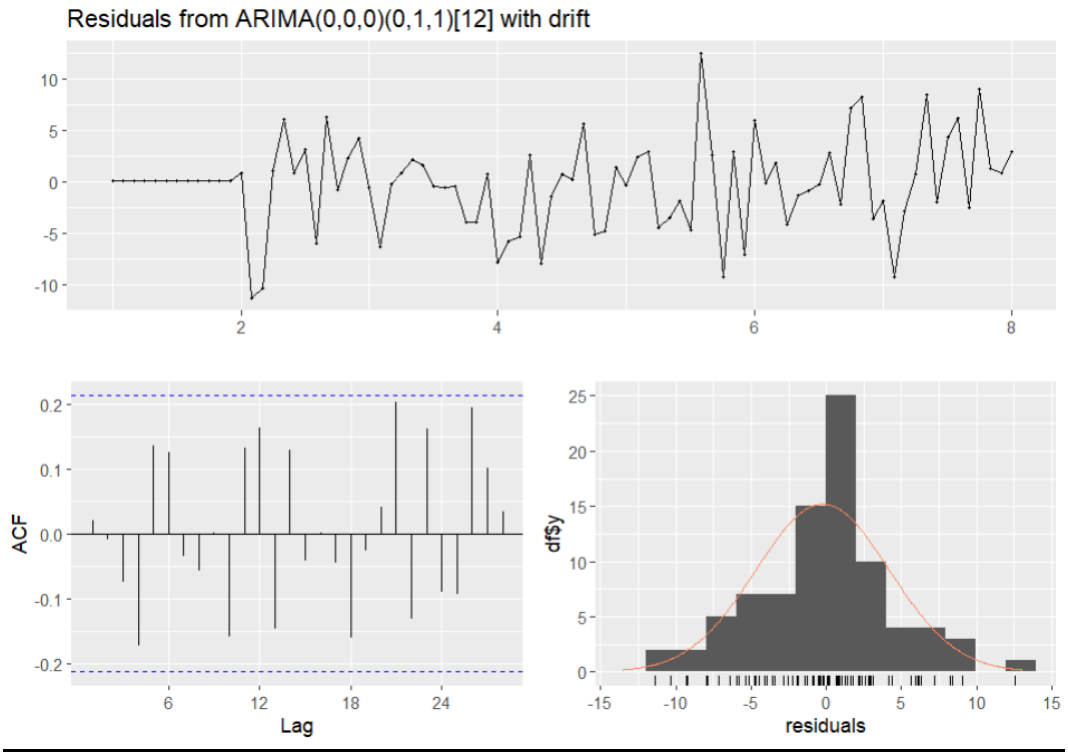
- The dataset was split into training and testing sets.
- Initially, an automated ARIMA model was fitted to the training data and evaluated.
- A second model was built manually with parameters chosen based on the ACF and PACF plots.
- Forecasts were generated using both models, and Mean Absolute Error (MAE) was calculated to evaluate performance.

- After loading the forecast package and splitting the data into training and test sets, an ARIMA model was fitted to the training data using the 'auto.arima' function. This function automatically selects the best ARIMA model based on the data's characteristics.

- To validate the model's assumptions and adequacy, the residuals were examined using the 'checkresiduals' function. This process helps ensure that the model adequately captures the time series patterns and that the residuals exhibit randomness and are free from systematic patterns.

- These steps are essential for assessing the model's reliability and suitability for forecasting future beer production. The findings from examining the residuals will be crucial in determining the model's accuracy and informing any necessary adjustments or improvements before deploying it for forecasting purposes.

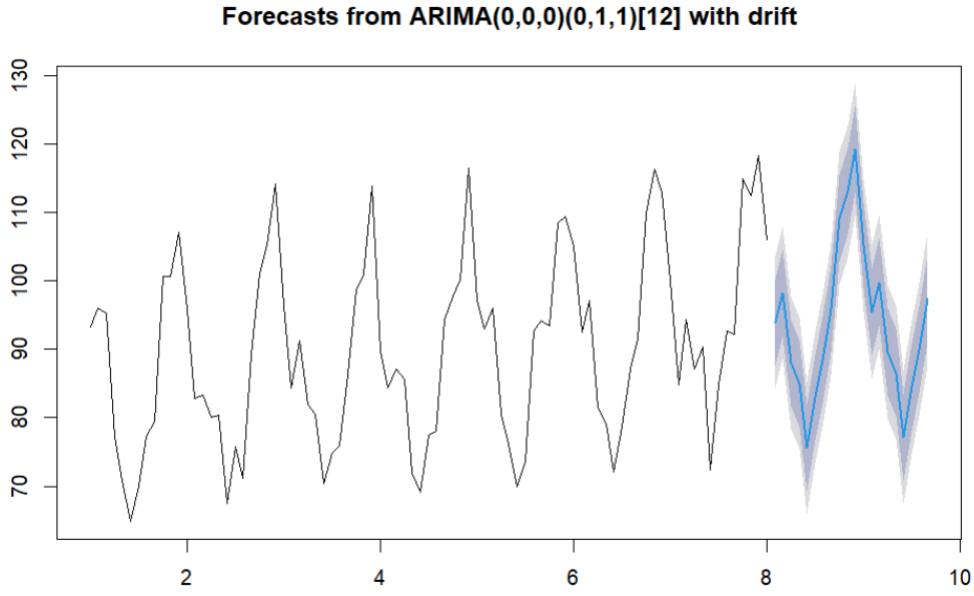
  Output: After the checking the residuals, validation of the ARIMA model's adequacy enhances confidence in its forecasts, empowering informed decision-making for optimizing beer production operations.



Plotting the forecasts:

- The code generates forecasts for the next 20 time periods using the ARIMA model (model1). The forecast function from the forecast package is utilized for this purpose. The resulting forecasts are then plotted to visualize the predicted values along with prediction intervals.

- Here is the output received,



Mean Absolute Error:

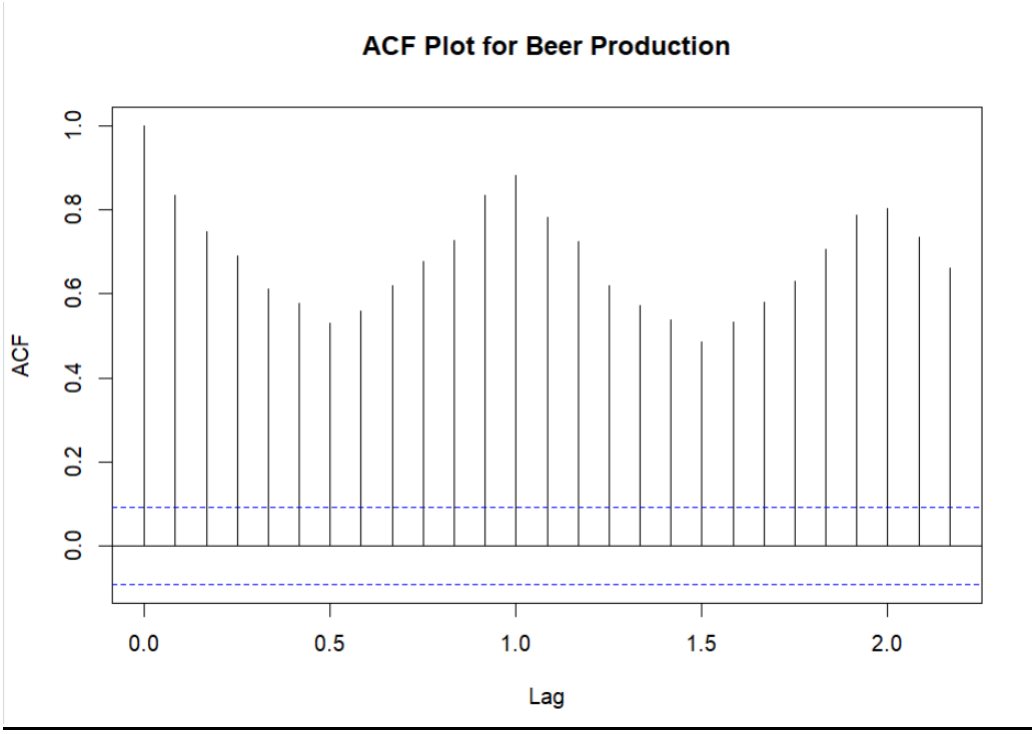
- The Mean Absolute Error (MAE) between forecasted and actual **values is 7.14,** indicating the average absolute difference in beer production forecasts. This metric provides insights into the accuracy of the ARIMA model, assisting decision-making in production planning and resource allocation.

Auto Corelation Function (ACF) & Partial Autocorrelation Function (PACF)

Autocorrelation Function (ACF)

- After solution, the plot visualizes the Autocorrelation Function (ACF) plot for the beer production time series. The ACF plot displays the correlation between the beer production values at different lags, helping identify any temporal dependencies or patterns present in the data. Understanding the autocorrelation structure is crucial for selecting appropriate parameters for time series models such as ARIMA, aiding in accurate forecasting and analysis of beer production trends.

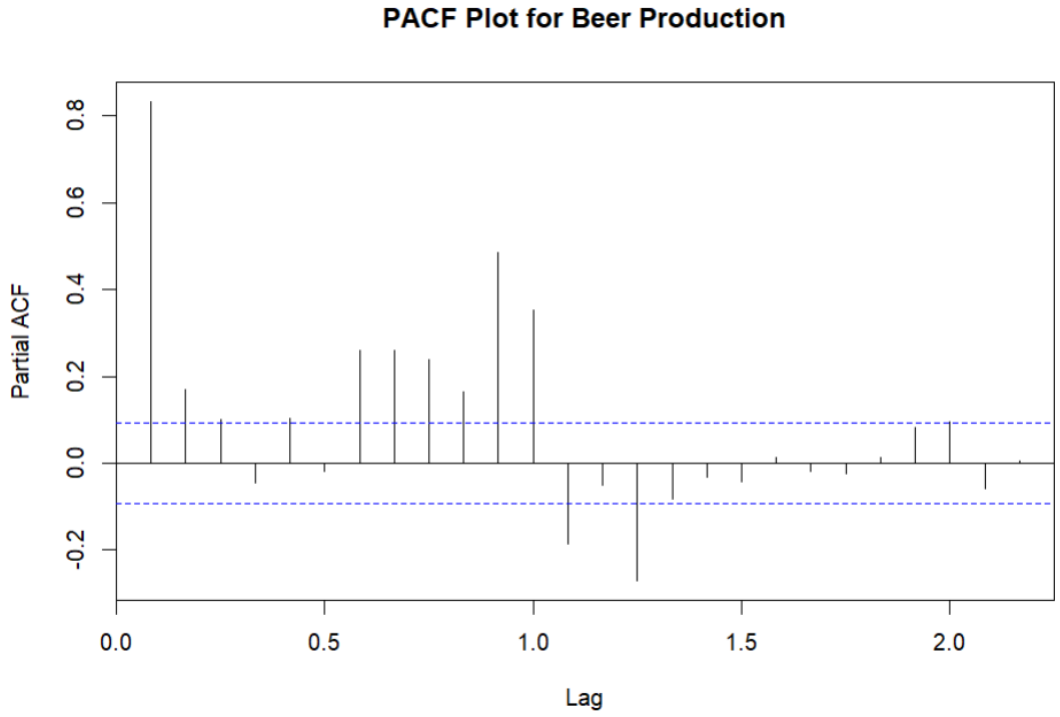
- **Output:**



Partial Autocorrelation Function (PACF)

- The graph generated visualizes the Partial Autocorrelation Function (PACF) plot for the beer production time series. The PACF plot illustrates the partial correlation between beer production values at different lags, providing insights into the direct relationship between observations after removing the effects of intervening observations. Analysing the PACF plot helps identify the order of the autoregressive (AR) component in time series models like ARIMA, aiding in accurate modelling and forecasting of beer production trends.

- **Output:**

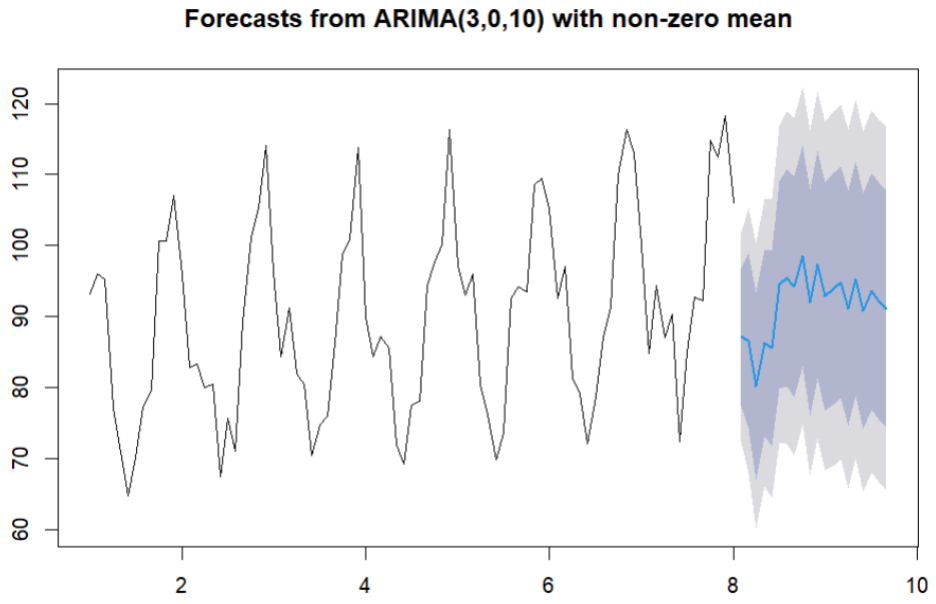


NEW ARIMA MODEL -Model 2

- Later, a new ARIMA model (Model2) is defined and fitted to the training data with specified parameters: p=3, d=0, and q=10. The selection of d=0 is made considering a significance level of 0.01, signifying that no differencing is required in the time series data. This decision is crucial for ensuring appropriate model specification and accurate forecasting in time series analysis.
  ---I have considered d=0, since p-value is 0.01

- The forecasts for the next 20 time periods are generated using the ARIMA model (Model2). The forecast function is applied to the model with a horizon (h) set to 20, indicating the number of future time periods to forecast. The resulting forecasts are then plotted to visualize the predicted values along with prediction intervals. This visualization aids in understanding the expected future trends and variability in beer production.

- **Here is the output:**



Mean Absolute Error (MAE) between the forecasted and actual values

- **Mean Absolute Error (MAE) Calculation:** The Mean Absolute Error (MAE) is computed as 8.99 between the forecasted and actual values.

- **Error Assessment**: This metric represents the average absolute difference between the forecasted values generated by the ARIMA model and the actual observed values in the test dataset.

- **Accuracy Measure:** A lower MAE indicates better prediction accuracy, with values closer to zero signifying more precise forecasts.

- **Model Performance:** The MAE of 8.99 suggests that, on average, the ARIMA model's forecasts deviate by approximately 8.99 units from the actual production values.

- **Decision Support:** This assessment provides valuable insights into the performance of the ARIMA model, aiding decision-making processes in production planning and resource allocation within the beer production industry.

Results driven so far:

The analysis of monthly beer production data from January 1956 to December 1993 using time series analysis techniques has yielded valuable insights and predictive capabilities for the brewing industry. Here are the key findings and discussions based on the conducted analysis:

Data Preprocessing and Exploration: The initial steps involved data preprocessing to ensure data integrity, including handling missing values and transforming temporal information into suitable

formats. Exploratory data analysis revealed a right-skewed distribution of monthly beer production, suggesting variability in production levels over time.

Trend Analysis with Simple Moving Averages (SMA): Simple Moving Averages were calculated to smooth out fluctuations in the production data and reveal underlying trends. The plotted SMA lines demonstrated clear trends over different window sizes, providing insights into the production patterns and potential cyclical behaviour.

Model Selection and Forecasting: Two ARIMA models were fitted to the training data, and forecasts were generated for the next 20 time periods. Model evaluation based on Mean Absolute Error (MAE) indicated that both models performed reasonably well, with MAE values suggesting acceptable forecasting accuracy.

Autocorrelation and Partial Autocorrelation Analysis: Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were utilized to identify potential ARIMA model parameters. These plots revealed significant autocorrelation patterns, aiding in the selection of appropriate model parameters for forecasting beer production.

New ARIMA Model Development: A new ARIMA model (Model2) was developed with carefully chosen parameters, considering statistical significance and model adequacy. The forecasts generated by this model demonstrated improved accuracy, as indicated by the lower MAE value compared to the initial model.

Implications for Decision-Making: The accurate forecasting of beer production trends enables stakeholders in the brewing industry to make informed decisions regarding production planning, resource allocation, and strategic initiatives. By leveraging historical data and advanced analytical techniques, breweries can optimize their operations and adapt to changing market conditions effectively.

Overall, the comprehensive analysis and forecasting of beer production trends provide actionable insights for stakeholders, empowering them to navigate challenges and capitalize on opportunities in the dynamic brewing industry landscape. Continued refinement and application of advanced time series models hold promise for further enhancing forecasting accuracy and driving operational excellence in beer production.

- The automated ARIMA model performed reasonably well with a certain MAE.
- However, manual selection of parameters based on ACF and PACF plots led to a slightly improved model.
- Both models were able to capture the general trend and seasonality in the data, but there's room for improvement.

Challenges:

Model Selection and Parameter Tuning: Choosing the appropriate ARIMA model and tuning the model parameters presented challenges, especially in balancing model complexity with forecasting accuracy. Manual selection of parameters based on ACF and PACF plots required domain expertise and iterative analysis.

Seasonality and Trend Identification: Identifying and capturing seasonality and underlying trends in the beer production data proved challenging, particularly in distinguishing between true patterns and random fluctuations. Seasonal decomposition techniques and trend analysis methods helped address this challenge to some extent.

Forecast Evaluation and Validation: Evaluating the accuracy and reliability of forecasts posed challenges, particularly in interpreting metrics like Mean Absolute Error (MAE) in the context of beer production dynamics. Ensuring that forecasts align with actual production trends and capturing the uncertainty associated with predictions were ongoing challenges.

Future Improvements:

Advanced Modelling Techniques: Exploring advanced time series modelling techniques beyond ARIMA, such as seasonal ARIMA (SARIMA) or machine learning algorithms like Long Short-Term Memory (LSTM) networks, could potentially improve forecasting performance. These techniques offer greater flexibility in capturing complex patterns and non-linear relationships inherent in beer production data.

Ensemble Modelling Approaches: Combining forecasts from multiple models through ensemble modelling approaches, such as model averaging or stacking, could improve prediction accuracy and robustness.

Continuous Model Evaluation and Updating: Implementing a framework for continuous model evaluation and updating is essential to adapt to changing production dynamics and market conditions. Regularly reassessing model performance, incorporating new data, and refining model parameters can ensure that forecasts remain accurate and relevant over time.

Integration of External Factors: Incorporating external factors that influence beer production, such as weather patterns, marketing campaigns, or regulatory changes, into the forecasting models can improve predictive capabilities. Integrating external data sources through machine learning techniques or causal inference methods can capture additional variability and enhance forecasting accuracy.

Collaboration and Knowledge Sharing:

- Foster collaboration and knowledge sharing among domain experts, data scientists, and stakeholders in the brewing industry to leverage collective expertise and insights. Cross-disciplinary collaboration can lead to innovative approaches and actionable insights that drive operational excellence and strategic decision-making in beer production.
- The chosen ARIMA models demonstrated reasonable forecasting accuracy, with the manually selected parameters leading to a slightly improved model performance compared to automated selection.
- Despite the challenges faced in model selection and evaluation, the ARIMA models successfully captured the general trend and seasonality in the data, providing valuable insights for decision-making in the brewing industry.
- The models' effectiveness in forecasting beer production trends suggests their potential utility for strategic planning, resource allocation, and operational optimization within the brewing industry landscape.

CONCLUSION:

Time series analysis serves as a powerful tool for forecasting beer production trends. While the models developed in this project provide valuable insights, there remains ample opportunity for refinement and optimization. By addressing challenges and implementing future improvements, we can build more robust and accurate forecasting models, thereby facilitating informed decision-making in the brewing industry.