*TIME SERIES ANALYSIS ASSIGNMENT*

*GROUP NO. 18*

*Names and PRNs:*

*Ankita Purohit 23060641004*

*Prachi Chile 23060641009*

*Aditi Nimbarte 23060641029*

*Suhasi Gohil 23060641042*

*Rutuja Uphade 23060641046*

*CLASS: MSc I Div A*

*PROBLEM IDENTIFICATION*

**Dataset:**

****

**Content of the dataset:**

The number of monthly sales of champagne for the Perrin Freres label (named for a region in France).

The dataset provides the number of monthly sales of champagne from January 1964 to September 1972, or just under 10 years of data.

The values are a count of millions of sales and there are 105 observations. The dataset is credited to Makridakis and Wheelwright, 1989.

**Problem Statement**:

To check the following things:

* **Seasonality**: This involves identifying repeating patterns or cycles within the data that occur at regular intervals, such as monthly or quarterly sales peaks.
* **Trend**: The trend component captures the long-term direction of the sales data, indicating whether sales are increasing, decreasing, or staying relatively stable over time.
* **Stationarity**: Stationarity refers to the statistical properties of the data remaining constant over time. Stationary data typically have constant mean, variance, and autocorrelation structure.
* **White Noise**: White noise represents random fluctuations in the data with a constant variance and no autocorrelation between observations. Identifying white noise is essential for understanding the randomness inherent in the data.
* **Forecasting**: Finally, the goal is to develop a forecasting model to predict future sales of champagne for the next year based on the observed patterns in the historical data.

***How does it fit to be a time series:***

Time series analysis deals with data collected over time, where observations are recorded at regular intervals. In this case, the monthly sales data for Perrin Freres champagne spans nearly a decade, with observations recorded each month. Hence, it fits to be a time series.

Time series analysis provides a framework for understanding and modelling these patterns, enabling accurate predictions of future outcomes.

***Initial Interpretation:***

By looking at the data values, there seems to be peak sales noticed in December/November and lowest sales in the August.

*EXPLORATORY DATA ANAYSIS*

1. Visualization:

Times Series Plot-

A graph of a chart

Description automatically generated with medium confidence

From this time series plot, we can see seasonality in the data with linearly increasing or decreasing values.

1. Summary Statistics:

A graph of a chart

Description automatically generated with medium confidence

From the above report, we can see that our mean, standard deviation, and variance comes out to be 4761.2, 2553.5 and 6520375 respectively.

1. Missing Values:

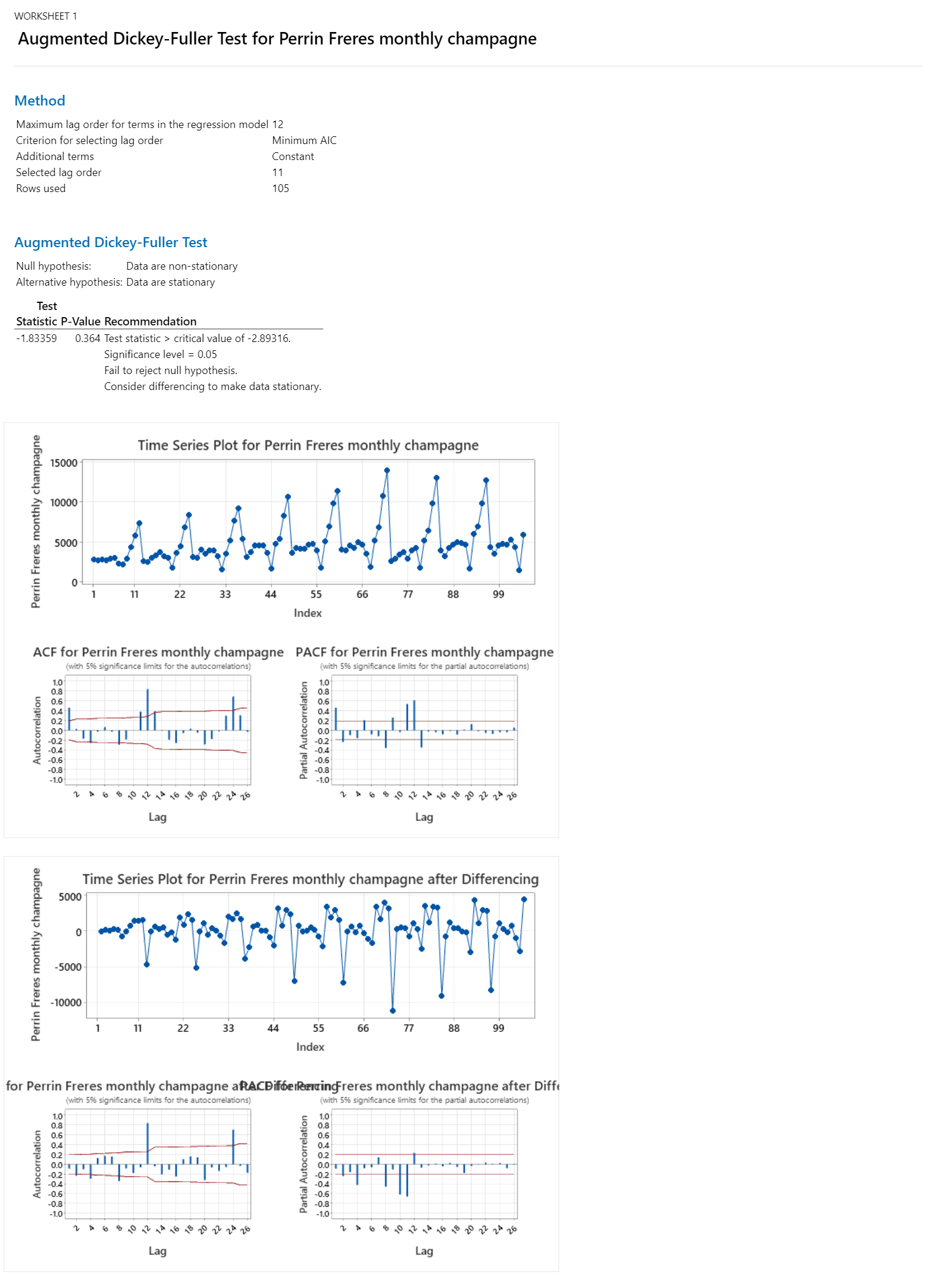
As our data consists of 105 observations, we find no missing values in the data by just looking at it.

1. Outliers: Our data contains outliers by looking at the boxplot.

A screenshot of a graph

Description automatically generated

1. Stationarity:



Hence by using ADF test, we get to know that our data is not stationary.

*DETAILED ANALYSIS*

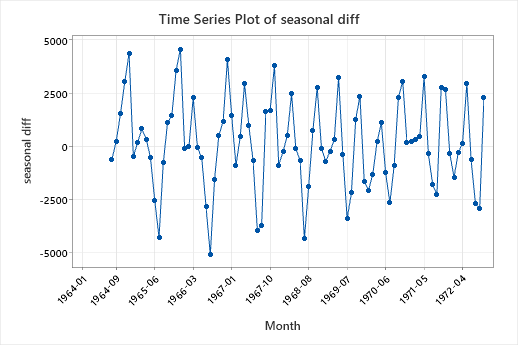
Minitab: Minitab was utilized for its comprehensive statistical analysis capabilities, particularly for conducting descriptive statistics, time series decomposition, and identifying trends, seasonality, and outliers in the data. Minitab's intuitive interface and wide range of statistical functions facilitated the initial exploration and interpretation of the dataset.

What software was put into use and the outputs?

We utilized Minitab's functionality to perform various analyses, including Time Series Plot, Graphical Summary Function, Boxplot Function, Conditional Formatting for Removing Outliers, ACF Function, PACF Function, ADF Test, Forecast with BEST ARIMA model, Differencing.

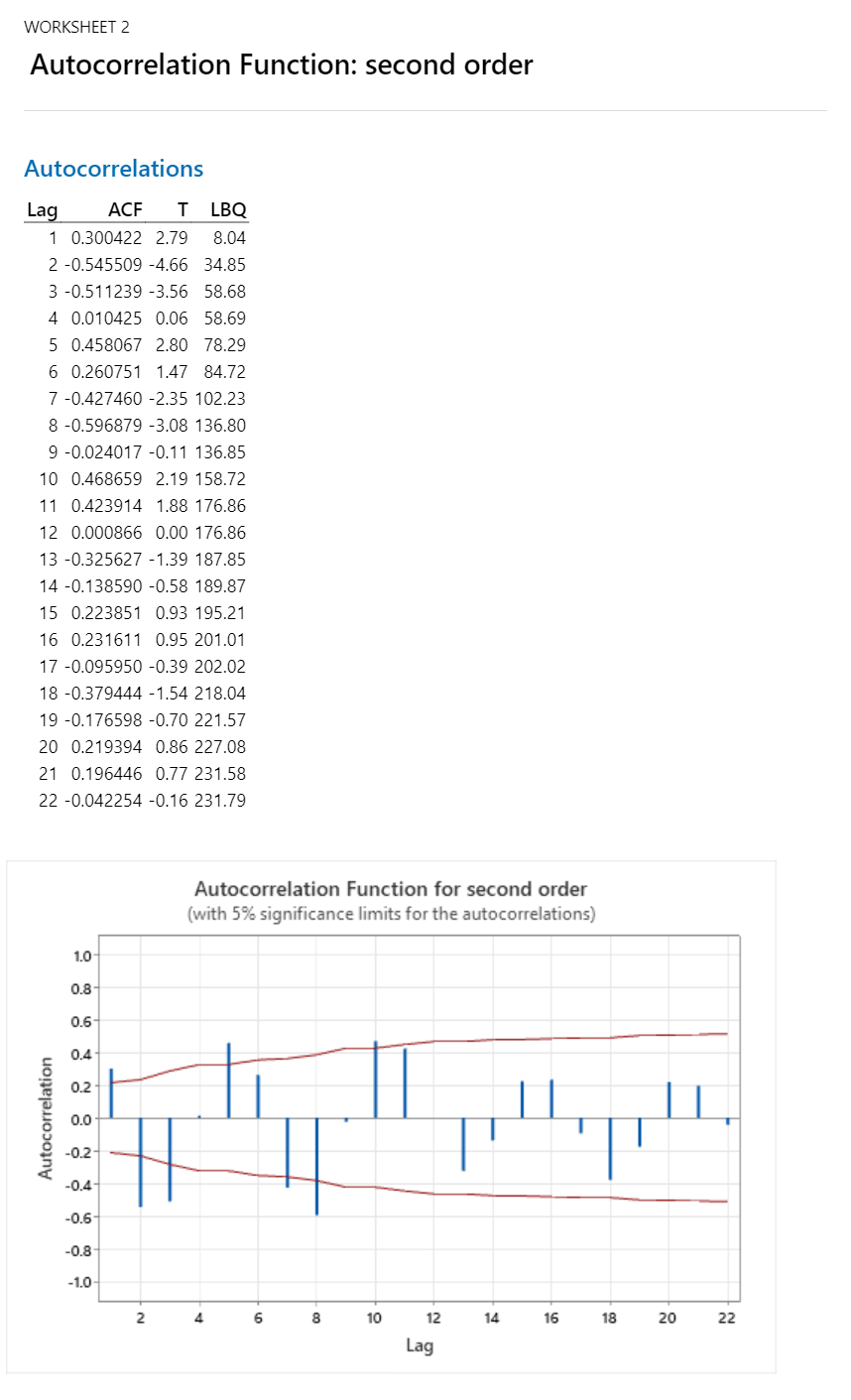
Tools used:

1. Time Series Plot: The Time Series Plot in Minitab displays the raw time series data graphically over time. It helps visualize the overall trend, seasonality, and any patterns or anomalies present in the data.
2. Graphical Summary Function: Graphical Summary Function provides graphical summaries such as histograms, boxplots, and quantile-quantile (Q-Q) plots for the time series data. It offers a quick visual assessment of the distribution, central tendency, and variability of the data.
3. Boxplot Function: The Boxplot Function in Minitab creates boxplots for the time series data, summarizing the distribution of values and highlighting potential outliers. It helps identify outliers, assess variability, and compare distributions across different groups or time periods. We removed the outliers identified by using boxplot.
4. Conditional Formatting for Removing Outliers: Conditional formatting in Minitab allows users to visually highlight data points that meet specific criteria or conditions. By defining rules based on statistical thresholds or criteria for outlier detection, outliers can be visually identified and flagged for further investigation or potential removal. The purpose of using conditional formatting for removing outliers is to visually identify data points that deviate significantly from the expected distribution or pattern in the time series data. By highlighting potential outliers, users can focus their attention on investigating these observations further to determine whether they are genuine anomalies or errors in the data. By using this, we removed the outliers from our data.
5. Differencing: Differencing in Minitab involves taking the difference between consecutive observations to remove trends or seasonality from the time series data. It helps achieve stationarity by stabilizing the mean and variance of the data, making it suitable for modelling with methods like ARIMA.

 A graph with blue dots and numbers

Description automatically generated  
So, we performed a seasonal differencing to remove the seasonality from our data and further took second order differencing of it to make the data stationary.

1. ACF Function: The ACF (Autocorrelation Function) Function computes and plots the autocorrelation of the time series data at various lags. It helps identify patterns of autocorrelation, indicating potential seasonality or temporal dependencies in the data.



1. PACF Function: The PACF (Partial Autocorrelation Function) Function computes and plots the partial autocorrelation of the time series data at different lags. It assists in determining the order of autoregressive (AR) terms in forecasting models like ARIMA by indicating the direct relationship between observations at different lags.

A screenshot of a document

Description automatically generated

1. ADF Test: The Augmented Dickey-Fuller (ADF) Test in Minitab is a statistical test for assessing the stationarity of time series data. It evaluates whether the time series data is stationary by testing the null hypothesis that the series has a unit root. A rejection of the null hypothesis indicates stationarity.

A screenshot of a test

Description automatically generated

A graph of different types of numbers

Description automatically generated with medium confidence

Hence, our data becomes stationary after all the seasonal differencing and second order differencing.

1. Forecast with BEST ARIMA Model: Minitab's Forecasting feature identifies and fits the best ARIMA (Autoregressive Integrated Moving Average) model to the time series data and generates forecasts for future time periods. It provides accurate predictions of future values based on the observed patterns and characteristics in the historical data, considering the trend, seasonality, and autocorrelation structure.

A screenshot of a computer

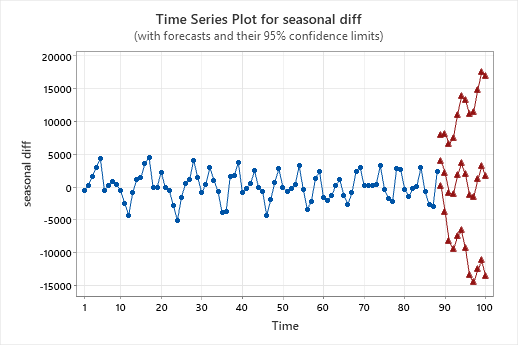
Description automatically generated

A number and text on a white background

Description automatically generated

A table of numbers and a number of times

Description automatically generated with medium confidence



We found that our best model is of ARIMA(2,2,3) and we got the forecasted sales for the next year.

***DIAGNOSTIC CHECKS***

***A close-up of a graph

Description automatically generated***

**PACF of residual with seasonal diff:**

The PACF values for lags 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, and 22 are all within the significant limits. This means that there is no statistically significant linear dependence of the residuals on their own past values at these lags. However, the PACF value for lag 1 is outside the significance limits. This suggests that there may be a significant linear dependence of the residuals on their own past value at lag 1.

In other words, the results of this PACF analysis suggest that there may be a seasonal pattern in the residuals, but there is no evidence of any other significant linear dependence in the residuals. This could be an indication that a seasonal ARIMA model would be appropriate for forecasting the time series.

Here are some additional details that can be gleaned from the graph:

The PACF values for lags greater than 1 are generally small. This suggests that there is no strong linear dependence of the residuals on their own past values at these lags.

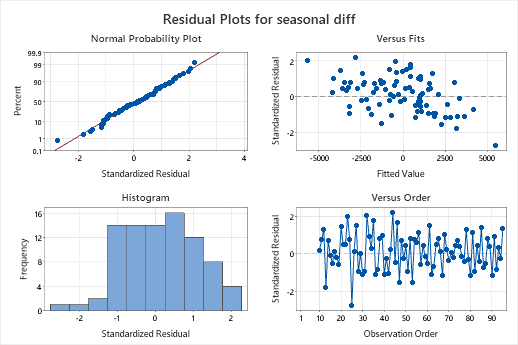
The PACF value for lag 1 is positive. This suggests that there is a positive linear dependence of the residuals on their own past value at lag 1.

**ACF of residuals with seasonal diff:**

The ACF values for most lags are within the 5% significance limits. This suggests that there is no strong autocorrelation in the residuals.

However, there does appear to be a slight upward trend in the ACF values at higher lags. This could indicate some weak dependence in the residuals.

Overall, the ACF graph suggests that the residuals from the seasonal differenced time series are approximately white noise. This is a desirable property for time series models, as it indicates that the model has captured all of the significant autocorrelation in the data.

******

**Normal probability plot:** This plot shows the standardized residuals on the x-axis and the percentiles of a standard normal distribution on the y-axis. If the data points fall close to a straight line, then the residuals are normally distributed.

**Scatter plot:** This plot shows the standardized residuals on the x-axis and the fitted values on the y-axis. If there is no pattern in the plot, then the residuals are randomly scattered and independent of the fitted values.

**Histogram:** This plot shows the distribution of the standardized residuals. If the residuals are normally distributed, then the histogram will be bell-shaped.

**Versus order plot**: This plot shows the standardized residuals on the y-axis and the order of the data on the x-axis. If there is no trend or pattern in the plot, then the residuals are independent of the order of the data.

Overall, these residual plots can be used to assess the normality, independence, and constant variance of the errors in a time series model. If the residuals meet these assumptions, then the model is a good fit for the data.

And hence, we get to know that our data has been transformed to a white noise.

***INTERPRETATION***

**Time Series Plot:** There is a clear seasonal pattern, with sales peaking in December and January, and dropping to their lowest point in July and August. This is likely because champagne is often purchased for celebrations, such as New Year's Eve and Christmas.

There is also a possible upward trend in sales over time. Sales appear to be generally higher in the later years (2000s and 2010s) compared to the earlier years (1970s and 1980s).

**Box Plot:** We can see that the median monthly sales of Perrin Freres champagne are around 6,000 bottles. There is a wider range of sales in the upper half of the distribution than the lower half, which means that there is a larger spread of higher sales months than lower sales months. Overall, the sales of Perrin Freres champagne appear to be somewhat seasonal, with some months having significantly higher sales than others.

**ACF Plot:** There is a significant positive autocorrelation at lag 1 (0.300422). This means that there is a positive correlation between the time series and its value one lag period earlier. In other words, if a value in the time series is high (or low), the next value is likely to be high (or low) as well.

There is a significant negative autocorrelation at lag 2 (-0.545509). This means that there is a negative correlation between the time series and its value two lag periods earlier. So, if a value is high, the value two periods after is likely to be low, and vice versa.

The autocorrelation values for lags greater than 2 are generally not statistically significant. This suggests that there is no significant correlation between the time series and its values at lags greater than 2.

**PACF Plot:** The first lag (PACF(1)) is positive and statistically significant, which means that there is a positive correlation between the time series and its one-lag past value.

The second lag (PACF(2)) is negative and statistically significant, which means that there is a negative correlation between the time series and its two-lag past value, after accounting for the influence of its one-lag past value.

Most of the PACF values for lags greater than 2 are not statistically significant, which suggests that there is little correlation between the time series and its past values beyond the first two lags.

**Augmented Dickey-Fuller Test:**

Null hypothesis: The null hypothesis is that the data are non-stationary.

Alternative hypothesis: The alternative hypothesis is that the data are stationary.

Test statistic: The test statistic is -7.63622. A more negative test statistic indicates stronger evidence against the null hypothesis (non-stationarity).

P-Value: The p-value is 0.000. A p-value less than the significance level (usually 0.05) suggests that we should reject the null hypothesis.

The decision based on the test statistic and p-value is to reject the null hypothesis. This means that we have evidence to conclude that the data is stationary.

And the graphs suggest that the original time series likely had a trend or seasonal component that was removed by differencing twice. After differencing, the series exhibits an AR (1) process, meaning the current value is influenced by the previous value.

**Forecast with BEST ARIMA model:**

Model (d = 2): This indicates that all models were fit with a differencing parameter (d) of 2. This means the data was differenced twice to remove seasonality.

p, q: These represent the autoregressive (AR) and moving average (MA) parameters of the model, respectively. The AR term considers past values of the series to predict future values, while the MA term accounts for past errors or residuals.

Log Likelihood: This measures how well the model fits the data. Higher values indicate a better fit.

AICc, AIC, BIC: These are information criteria used to compare the goodness-of-fit of different models while penalizing for model complexity. Lower values indicate a better model.

Interpreting the results for model selection:

The best model is selected based on the minimum AICc value (1570.81). These metric balances goodness-of-fit with model complexity.

In this case, the best model has p = 2 and q = 3. This means the model incorporates the influence of the past two data points (AR) and accounts for the past three error terms (MA) to predict future values.

Models with p = 0 generally performed worse, suggesting that past values of the series do contain useful information for prediction.

Similarly, models with q = 1 or q = 2 tended to have higher AICc values compared to models with q = 3, indicating that incorporating three past error terms improves the model.

Models with higher d (differencing) were not explored here, but it's possible they could be better depending on the specific time series.

Overall, the best model captures the influence of the past two data points and considers the impact of the past three error terms to make predictions. This suggests that there might be a trend or pattern in the data that the model is efficient.

the final estimates for the parameters of the best ARIMA model you identified earlier (p = 2, q = 3). Let's break down the information in the table:

Type: This indicates whether the parameter is related to the Autoregressive (AR) or Moving Average (MA) component of the model.

Coef: This is the estimated coefficient value for each parameter. In ARIMA models, these coefficients determine how much weight is given to past values of the series (AR) or past errors (MA) for prediction.

SE Coef: This is the standard error of the coefficient. It represents the variability associated with the estimated coefficient value.

T-Value: This is the test statistic used to assess the significance of each coefficient. A high absolute value (positive or negative) indicates a stronger influence of that parameter on the model.

P-Value: This represents the probability of observing a T-value as extreme as the one calculated, assuming the coefficient is truly zero. A low P-value (typically less than 0.05) suggests that the coefficient is statistically significant and likely contributes to the model's performance.

Interpretation:

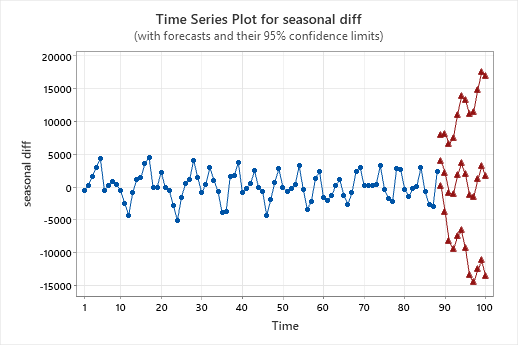
AR parameters: Both AR coefficients (AR1 and AR2) are statistically significant (p-value < 0.000). AR1 has a positive coefficient (0.6473), indicating that the value one time step ago (lag 1) has a positive influence on the prediction. AR2 has a negative coefficient (-1.0015), suggesting the value two time steps ago (lag 2) has a negative influence on the prediction. This combination implies a possible cyclical or dampening trend in the data.

MA parameters: All three MA coefficients (MA1, MA2, and MA3) are statistically significant (p-value < 0.000). These coefficients account for the influence of past errors on the prediction. The positive value of MA1 (1.5118) suggests that the error from one time step ago has a positive impact on the prediction. The negative values of MA2 (-1.418) and MA3 (0.8832) indicate that errors from two and three time steps ago, respectively, have a negative influence. This combination suggests the model is capturing some complex error patterns in the data.

Differencing: The table confirms the differencing parameter (d) is 2, indicating the data was differenced twice to remove seasonality.

Number of observations: This shows the number of data points remaining after differencing (86), which is used to fit the model.

Overall, this summary provides detailed information about the estimated parameters of the best ARIMA model. The significant coefficients and their signs suggest the model captures a trend or cyclical pattern in the data, along with the influence of past errors.

****

The x-axis represents time. It is not labelled with specific units but is marked from 1 to 100.

The y-axis represents the value of the time series. The scale goes from -15,000 to 20,000.

There is a line plot in the centre of the graph that shows the historical values of the time series. This line is difficult to see because it is mostly obscured by the forecast and confidence limit lines.

There is a blue line above and below the centre line that shows the 95% confidence limits for the forecast. The confidence limits indicate the range of values that the actual future values of the time series are likely to fall within, with 95% certainty.

There is a green line plotted on top of the centre line that shows the forecasted values of the time series.

The graph shows that the time series has a seasonal component, with values that fluctuate up and down over time. The forecast shows that this seasonal pattern is expected to continue in the future. The confidence limits show that there is some uncertainty about the future values of the time series, but the forecast suggests that the values are likely to remain within a certain range.

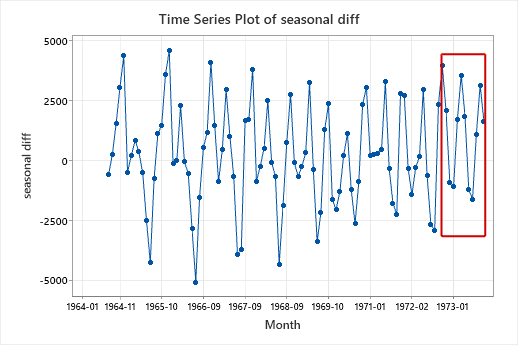
We get the forecasted sales for the next year as seen in the below picture:

A table of numbers and a number of times

Description automatically generated with medium confidence

***CONCLUSION***

Firstly, we found seasonality in the data. Then we had to use seasonal differencing and further second order differencing to remove the seasonality and making the data stationary. We performed diagnostic checks like residual analysis, white noise, stationarity. We found the best ARIMA Model that is ARIMA(2,2,3) to forecast the champagne sales between1972-10 to 1973-09



The values in the rectangle box are our forecasted sales for the next year that is from 1972-10 to 1973-09.

**ATTACHMENT:**

