**Problem**

For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

Read the data as an appropriate Time Series data and plot the data.

Graphical user interface, chart, histogram

Description automatically generated

1. Sparkling wine data is available from 1980 until 1995.
2. Data is presented to us as, YYYY-MM format.
3. Wine sales had seen an upward trend until 1988, at this point, Sales have reached an all-time high.
4. 1988 till 1995, the Sales trend maintained the increasing pattern.

Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

1. Data presented to us has two columns YearMonth (object type) and Sparkling (int64 type) a total of 187 rows.
2. No nulls were observed and YearMonth data will be changed to DateTime object in YYYY-MM-DD format with a frequency of M.
3. Mean of sales is at 2402, standard deviations of sales is at 1295, minimum sales observed is at 1070, 50% of sales is around 1874 and all time high sales is at 7242.

Yearly box plot

Chart, box and whisker chart

Description automatically generated

1. Normal distribution pattern is observed with an all-time high in 1987. Each year has one outlier, sales are higher at one point of time in that year.

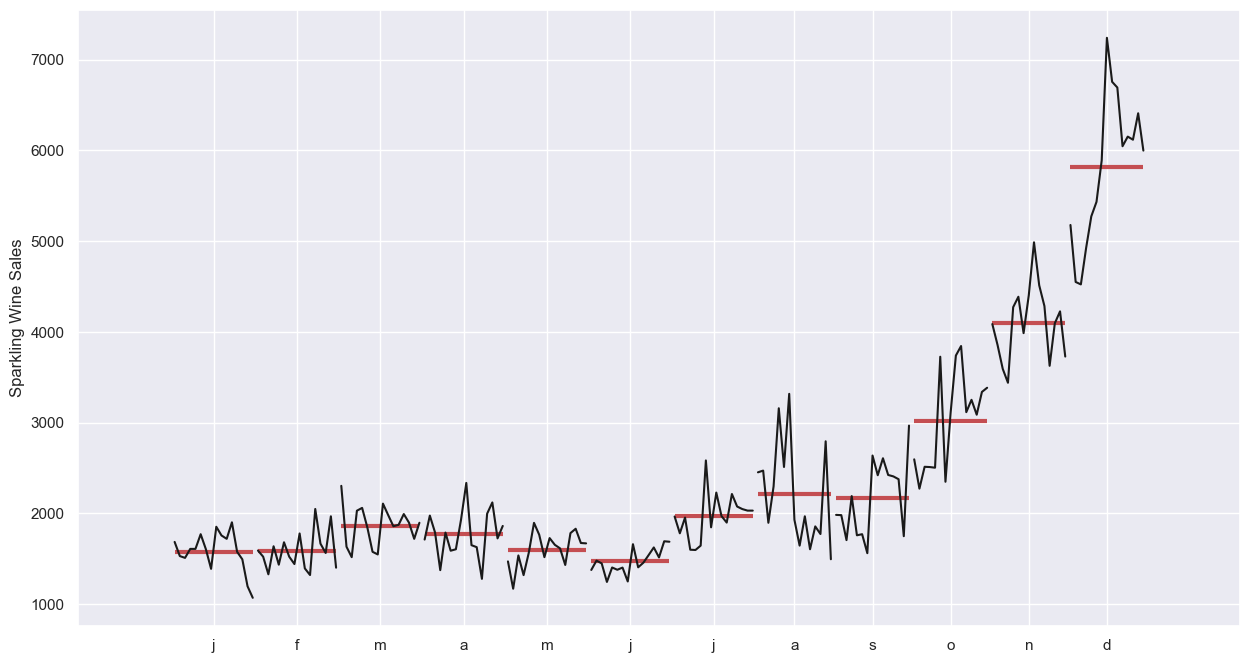
Monthly box plot

Chart

Description automatically generated

1. Monthly box plot is left skewed with an increase in sales during festive season of December.

Aggregated monthly plot



1. All time high was observed in the December months of all years with a rapid increase in sales towards the end of the month.
2. Pattern of sales increase is observed in the last 2 quarters of every year.

Time series decomposition

A picture containing graphical user interface

Description automatically generated

1. Multiplicative seasonality is observed.
2. Raw data shows a slight uptick at sales in 1987.
3. Trend presents us a plot with a half yearly averaged data, has a smooth Sales uptick in 1987.
4. Seasonality is observed to be constant across the time range.
5. Residuals are observed to follow the distribution as the trend with high residuals in 1987.

Split the data into training and test. The test data should start in 1991.

Chart, histogram

Description automatically generated

1. Training data includes observations until 1991 and Testing data includes observations from 1991 and above.

Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Linear Regression

Chart, histogram

Description automatically generated

1. Test predictions for the linear regression model have a uniform result that you can observe in the above plot with a green horizontal line across the Test data.
2. This model does not provide accurate test observation.
3. RMSE score is 1389.135.

Naïve Model

Chart, histogram

Description automatically generated

1. Naïve model set the last observed result as a predicted result.
2. Green line depicts the predicted result on test data, which seems ideal, but if the last observation has a low sales, then this result would be opposite.
3. This is not an accurate prediction with an increase in RMSE of 3864.

Simple Average Model

Chart, histogram

Description automatically generated

1. Simple average model will calculate the average of observations and presents it as a predicted output.
2. With an RMSE of 1275.08, the predicted result has a less error compared to regression and naïve models, however the prediction on the test data does not seem ideal.

Moving Average Model

Histogram

Description automatically generated with low confidence

Histogram

Description automatically generated with medium confidence

1. Moving average was calculated at 2, 4, 6, and 8 trailing months.
2. 2-point MovingAverage model’s test accuracy is better than other models with lowest RMSE of 813.4. 4-point MA Rmse is 1156.58, 6-point MA Rmse is 1283.92 and 9-point MA Rmse is 1346.27.

Observed model’s plot

Graphical user interface, chart, histogram

Description automatically generated

1. Naïve model predicts high values and Simple average predicts low values on Test data.
2. Best model until now is 2-point moving average model with overlapping predictions on Test data with RMSE of 813.4.

Simple Exponential Smoothing (SES)

Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.