**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.

***SPARKLING WINE***

**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**

Chart, histogram

Description automatically generated

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 Model**

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)**

Chart, histogram

Description automatically generated

1. After applying SES, Smoothing-Level/Alpha value is applied at 0.0496, Smoothing-Trend and Smoothing-Seasonal values are NaN.
2. RMSE score for **SES is 1316.135** and SES produces a horizontal line over the Test Predictions, this is not the best model to predict as it does not cover the variability of the test data.

**Simple Exponential Smoothing (SES) with various smoothing levels**

Table

Description automatically generated

1. I have tried alpha values from 0.3 to 0.9 and found 0.3 to be the optimal alpha value in the context of Test RMSE.
2. **Test RMSE score for 0.3 is the lowest** of all other alpha values.

Graphical user interface

Description automatically generated

1. Plot shows that the Alpha 0.3 did a better job in predicting the Test data than Alpha=1, even though it’s a horizontal line and predictions would be the same across the Test time range.

**Double Exponential Smoothing (DES)**

Table

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

1. Best alpha and beta values are 0.3 and 0.3 which produced a best Test RMSE of 18259.110, which however is worst performing of all the models produced till now.

**Triple Exponential Smoothing (TES)**

Graphical user interface

Description automatically generated

1. With default auto prediction, the above green colour line specifies the Test predictions with alpha=0.11, beta=0.0617 and gamma=0.395, this configuration has an RMSE score of 469.65.

**TES – Best Alpha, Beta and Gamma**

Table

Description automatically generated

1. Best **Alpha, Beta and Gamma values are 0.3, 0.3 and 0.3**, which produced the best **Test RMSE of 343.61**.
2. RMSE score for 0.3 is better than the default.

**Plot best of SES, DES and TES**

Graphical user interface, text

Description automatically generated

1. DES scores the least in comparison with other smoothing methods, TES with 0.3, 0.3 and 0.3; alpha, beta and gamma values produces an accurate prediction on the test data.

**TES predictions on Full Data**

Graphical user interface, application

Description automatically generated

1. Forecast from 1995 to 2000 produces an increasing Trend with an **RMSE score of 421.309** which predicts the highest point of Sales predicted at the 20th century.

**Margin of Error**

**Graphical user interface

Description automatically generated**

1. Plot shows the test predictions with varying confidence interval (margin of error).
2. Peaks are predicted with higher confidence, the dips are varying in their travel for different confidence intervals.

**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.**

1. p-value obtained from adFuller test with 175 number of observations is 0.601.
2. **As p-value is higher than the set-alpha 0.05, H0-Time series is not-stationary is Rejected and hence the alternate Hypothesis Ha-Time series is stationary is accepted.**

**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.**

**Automated ARIMA**

Text, letter

Description automatically generated

1. Automated ARIMA model was built with p and q ranging from 0-3 and d = 1.
2. **AR (p) => 2, I (d) => 1 and MA (q) => 2; provides a least and best AIC score of 2210.62**.

**ARIMA Automated Result (2,1,2) Test RMSE**

1. ARIMA (2,1,2) produced an RMSE score of 1374.311, which is not the best RMSEs obtained till now.

**Automated SARIMA for 6 Seasonality**

Text, letter

Description automatically generated

1. Automated SARIMA for 6 seasonality (1,1,2) (2,0,2,6) yielded lowest AIC score of 1727.66.

**SARIMA Automated Result (1,1,2) (2,0,2,6) Test RMSE**

1. SARIMA (1,1,2)(2,0,2,6) produced a Test RMSE score of 459.64. This is better than ARIMA Test RMSE.

**Automated SARIMA for 12 Seasonality**

Text, letter

Description automatically generated

1. Automated SARIMA for 12 seasonality (1,1,2) (1,0,2,12) yielded lowest AIC score of 1555.58.
2. This value is lower than SARIMA 6 AIC value.

**SARIMA Automated Result (1,1,2) (1,0,2,12) Test RMSE**

1. SARIMA (1,1,2)(1,0,2,12) produced a Test RMSE score of 546.44. SARIMA 6 RMSE score was least than the SARIMA 12 there by, having an accurate prediction on the Test data.

**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.**

**ACF and PACF**

**ACF (q)**

Chart, histogram

Description automatically generated

1. ACF value derived from the above plot is 2, hence the q value will be 2.

**PACF (p)**

Chart

Description automatically generated

1. PACF value derived from the above plot is 3, hence the p value will be 3.

**Manual ARIMA at (p,d,q) => (3,1,2)**

1. Test RMSE score obtained for (3,1,2) is 1378.09 which is not the best RMSE score obtained till now.

**Manual SARIMA (3,1,2) (3,1,2,6) for 6 seasonality**

1. Test RMSE for Manual SARIMA (3,1,2)(3,1,2,6) is 933.005, which is not the best RMSE score obtained till now.

**Manual SARIMA for (3,1,2) (3,1,2,12) for 12 seasonality**

1. Test RMSE for Manual SARIMA (3,1,2) (3,1,2,12) is 329.53, which is best RMSE score obtained till now.
2. This model has the capability to forecast the future timeline of 12 months.

**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.**

Table

Description automatically generated

Chart, bar chart

Description automatically generated

1. SARIMA of 12 seasonality produces the best RMSE score, next in line is the TripleExponentialSmoothing and after that it’s SARIMA with 6 seasonality.

**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.**

Chart

Description automatically generated

1. SARIMA (3,1,2) (3,1,2,12)RMSE obtained for the full data is 578.95.
2. Predicted future time range is 1995-08-31 to 1996-07-31.
3. In the above plot the optimal model => SARIMA (3,1,2)(3,1,2,12) predicts the peaks under good confidence, whereas, the dips have a range of confidence intervals with the SARIMA model being the average of all intervals.

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

1. Quarter end result will impact the next quarters performance as the PACF value is cut off at 3, If company can do well at the end of quarter the next quarter is expected to perform better as per the model.
2. The immediate 2 months can impact the 3rd month sales performance. It’s mandate to sustain good sale numbers for the first 2 months in a quarter to have a positive impact on the 3rd month. This is obtained from the q value of 2.
3. This forecast is expected to remain consistent for a season of 12 months, if company improves the sales more than 3 months then the dip sales also will improve as the confidence interval is high and same applies to deteriorate if the sales decrease more than 3 months.
4. Forecasted model predicts that, the sales shall maintain the trend as the current season for the next 12 months, with a chance of improve/deteriorate based on the month on month performance of the company.

***ROSE WINE***

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

Graphical user interface

Description automatically generated

1. Rose 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 1981, at this point, Sales have reached an all-time high.
4. From then till 1995 Rose wine sales were declining.
5. This tells us that the demand of wine was higher when it was introduced, and it declined during the next years.
6. 1981 till 1995, the Sales trend maintained the decreasing 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 Rose (int64 type) a total of 187 rows.
2. There were 2 nulls, I have used interpolate method to fill the nulls.
3. No nulls were observed and YearMonth data will be changed to DateTime object in YYYY-MM-DD format with a frequency of M.
4. 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. With an all-time high in 1981, followed with a decline each year, in years 1990-1995, the density of Rose wine customers declined rapidly by their number and sales.

**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**

A picture containing text, antenna

Description automatically generated

1. All time high was observed in the December months of all years with a rapid decrease in sales towards the end of the month.
2. Pattern of sales increase is observed in the 7th month of every year, and followed with an all-time high in December.

**Time series decomposition**

Graphical user interface

Description automatically generated with medium confidence

1. Multiplicative seasonality is observed.
2. Raw data shows a slight uptick at sales in 1981.
3. Trend presents us a plot with a half yearly averaged data, has a smooth Sales uptick in 1981, followed with a smooth decline.
4. Seasonality is observed to be constant across the time range.
5. Residuals are observed to follow the mean.

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

Graphical user interface, chart

Description automatically generated

1. Training data includes observations until 1991 and Testing data includes observations from 1991 and above, Train data has 132 rows and Test data has 55 rows.

**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 Model**

Graphical user interface, chart

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 15.268**.

**Naïve Model**

Graphical user interface, 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 is incorrect, but if the last observation has a low sales, then this result would be matching the declining trend.
3. This is not an accurate prediction with an increase in **RMSE of 79.71**.

**Simple Average Model**

Graphical user interface, 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 53.46**, the predicted result has a less error compared to naïve model, however the prediction on the test data is incorrect and it does not capture the decreasing trend of the Test data.

**Moving Average Model**

A picture containing histogram

Description automatically generated

A picture containing chart

Description automatically generated

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 11.529**. **4-point MA Rmse is 14.451, 6-point MA Rmse is 14.566 and 9-point MA Rmse is 14.727**.
3. There by 2 point MovingAverage model captures the variance in data both for Train and Test with a lowest RMSE than other models till this point.

**Observed model’s plot**

Graphical user interface, chart, histogram

Description automatically generated

1. Naïve, Simple Average model predictions are outright incorrect and does not capture variance.
2. Linear regression model captures the flow of trend in a flat line, which will produce incorrect predictions when used.
3. Best model until now is 2-point moving average model with overlapping predictions on Test data with **RMSE of 11.529**.

**Simple Exponential Smoothing (SES)**

Graphical user interface, chart

Description automatically generated

1. After applying SES, Smoothing-Level/Alpha value is applied at 0.0987, Smoothing-Trend and Smoothing-Seasonal values are NaN.
2. RMSE score for **SES is 36.796** and SES produces a horizontal line over the Test Predictions, this is not the best model to predict as it does not cover the variability of the test data.

**Simple Exponential Smoothing (SES) with various smoothing levels**

Table

Description automatically generated

1. I have tried alpha values from 0.3 to 0.9 and found 0.3 to be the optimal alpha value in the context of Test RMSE.
2. **Test RMSE score for 0.3 is the lowest** of all other alpha values.

Graphical user interface

Description automatically generated

1. Plot shows that the Alpha 0.3 did a better job in predicting the Test data than Alpha=1, both of them do not capture variance in data, however for alpha 0.3, it’s a horizontal line and predictions would be the same across the Test time range.

**Double Exponential Smoothing (DES)**

Table

Description automatically generated

Graphical user interface, chart

Description automatically generated

1. Best alpha and beta values are 0.3 and 0.3 which produced a best Test RMSE of 265.56, which however is worst performing of all the models produced till now.
2. As DES does not consider Seasonality it’s expected to proceed with the last observed level, trend and indicate an up ticking prediction.

**Triple Exponential Smoothing (TES)**

Graphical user interface, application

Description automatically generated

1. With default auto prediction, the above green colour line specifies the Test predictions with alpha=0.0646, beta=0.053 and gamma=0.0, this configuration has a Test RMSE score of 21.22.

**TES – Best Alpha, Beta and Gamma**

Table

Description automatically generated

1. Best **Alpha, Beta and Gamma values are 0.3, 0.3 and 0.4**, which produced the best **Test RMSE of 10.169**.
2. RMSE score for 0.3, 0.3 and 0.4 is better than the default.

**Plot best of SES, DES and TES**

Graphical user interface, application

Description automatically generated

1. DES scores the least in comparison with other smoothing methods, TES with 0.3, 0.3 and 0.4; alpha, beta and gamma values produce an accurate prediction on the test data with Test RMSE at 10.169.

**TES predictions on Full Data**

Graphical user interface, application

Description automatically generated

1. Forecast from 1995 to 2000 produces an decreasing Trend with an **RMSE score of 21.013** which predicts the lowest point of Sales at the end of the 20th century.
2. This RMSE is higher than the it’s TES’s Test RMSE, however, predicted pattern matches the declining trend and captures the declining variance.

**Margin of Error**

Graphical user interface, chart, line chart

Description automatically generated

1. Plot shows the test predictions with varying confidence interval (margin of error).
2. TES model is the average of all the high and low confidence intervals. This model will provide a ideal value without overshooting the previous trend.

**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.**

1. p-value obtained from adFuller test with 173 number of observations is 0.343.
2. **As p-value is higher than the set-alpha 0.05, H0-Time series is not-stationary is Rejected and hence the alternate Hypothesis Ha-Time series is stationary is accepted.**

**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.**

**Automated ARIMA**

Text

Description automatically generated

1. Automated ARIMA model was built with p and q ranging from 0-3 and d = 1.
2. **AR (p) => 3, I (d) => 1 and MA (q) => 3; provides a least and best AIC score of 1273.194**.

**ARIMA Automated Result (3,1,3) Test RMSE**

1. ARIMA (3,1,3) produced an RMSE score of 15.98, which is not the best RMSEs obtained till now.

**Automated SARIMA for 6 Seasonality**

Text

Description automatically generated

1. Automated SARIMA for 6 seasonality (1,1,2) (2,0,2,6) yielded lowest AIC score of 1041.65.

**SARIMA Automated Result (1,1,2) (2,0,2,6) Test RMSE**

1. SARIMA (1,1,2)(2,0,2,6) produced a Test RMSE score of 27.36. This is not better than ARIMA Test RMSE.

**Automated SARIMA for 12 Seasonality**

Text, letter

Description automatically generated

1. Automated SARIMA for 12 seasonality (0,1,2) (2,0,2,12) yielded lowest AIC score of 887.937.
2. This value is lower than SARIMA 6 AIC value.

**SARIMA Automated Result (0,1,2) (2,0,2,12) Test RMSE**

1. SARIMA (1,1,2)(2,0,2,12) produced a Test RMSE score of 26.928. SARIMA 6 RMSE score was least than the SARIMA 12 there by, having an accurate prediction on the Test data.

**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.**

**ACF and PACF**

**ACF (q)**

Chart, box and whisker chart

Description automatically generated

1. ACF value derived from the above plot is 2, hence the q value will be 2.

**PACF (p)**

Timeline

Description automatically generated

1. PACF value derived from the above plot is 4, hence the p value will be 4.

**Manual ARIMA at (p,d,q) => (4,1,2)**

1. Test RMSE score obtained for (4,1,2) is 33.949 which is higher than ARIMA (3,1,3).

**Manual SARIMA (4,1,2) (4,1,2,6) for 6 seasonality**

1. Test RMSE for Manual SARIMA (4,1,2)(4,1,2,6) is 19.69, which is not the best RMSE score obtained till now.

**Manual SARIMA for (4,1,2) (4,1,2,12) for 12 seasonality**

1. Test RMSE for Manual SARIMA (4,1,2) (4,1,2,12) is 17.528, which is not the best RMSE score obtained till now.

**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.**

Graphical user interface, text, application, email

Description automatically generated

Chart, bar chart

Description automatically generated

1. 0.3,0.3,0.8 TripleExponentialSmoothing produced the best RMSE and will have a good forecast ability.
2. DES model was the worst with highest RMSE score.
3. After TES, 2 to 9 point Moving Average models produced best RMSE scores in that order.

**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.**

Graphical user interface, chart

Description automatically generated

1. Forecast from 1995 to 2000 produces an decreasing Trend with an **RMSE score of 21.013** which predicts the lowest point of Sales at the end of the 20th century.
2. This RMSE is higher than its TES’s Test RMSE, however, predicted pattern matches the declining trend and captures the declining variance.
3. Plot shows the test predictions with varying confidence interval (margin of error).
4. TES model is the average of all the high and low confidence intervals. This model will provide an ideal value without overshooting the previous trend.

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

1. Forecasted model predicts that, the sales shall maintain the declining trend as the current season for the next 12 months, with a chance of improve/deteriorate based on the month on month performance of the company. Even with a higher confidence of TES model the sales will decline with increase in the level.
2. Rose wine has not seen an uptick since 1981 and from then it’s declining, even the forecasted model predicts that it will have declining sales.
3. Seasonal factor has no major affect the on Sales of Rose Wine.
4. To have an increase in Sales, company has to work on creating a Seasonal effect by changing the taste and branding of the Wine.
5. While Season, company can add promotional offers to increase the volume of the Wine sales.
6. Performing points 4 and 5 will ensure the Level and Trend are taken care of, which will indirectly impact the Seasonality in Years to come.
7. If not taken care of, Rose Wine will have near zero Sales volume in 21st century.