TIME SERIES FORECASTING

ON

Different Wines OF ABC Estate Wines



https://www.foodnetwork.com/healthyeats/holidays/2016/12/which-sparkling-wines-are-worth-your-holiday-jingle

Name- Debsmita Chakraborty

Batch-July ‘C”

TABLE OF CONTENTS

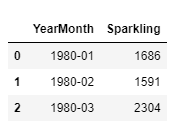
SUMMARY

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, we are tasked to analyse and forecast Wine Sales in the 20th century.

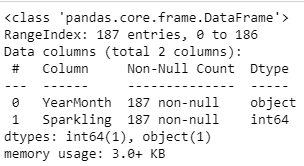
Data being of Sparkling wine and Rose wine.

INTRODUCTION

The file given to us is of two types of wine: Rose and Sparkling.



The data for Sparkling has:

* Data ranges from January 1980-July 1995. There are 187 rows in the data and 2 columns.
*  

The data has one object type column which is the Year Month and the other is the Sales column which is int64. The memory occupied by the data set is 3 KB. The data in 0-2 represents the head and the 184-186 is the tail of the data.

* No null values in the data:



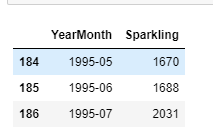
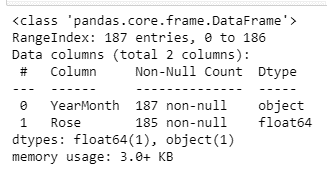


Fig a) Head and Tail-Sparkling

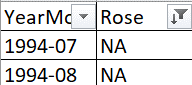
The data for Rose has:

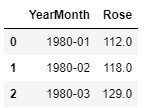
* Data ranges from January 1980-July 1995. There are 187 rows in the data and 2 columns.
*  

The data has one object type column which is the Year Month and the other is the Sales column which is int64. The memory occupied by the data set is 3 KB.

The data in 0-2 represents the head and the 184-186 is the tail of the data.

* There are two null values as below in the data for years mentioned:





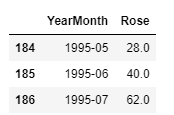


Fig a) Head and Tail-Rose

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

* The data has two wines namely Sparkling and Rose.
* Sparkling as shown above in fig a has no missing data but Rose has.
* This can also mean that Sparkling has been preferred throughout years but Rose was not a preference in July-August 1994.
* Let’s look at the data individually:

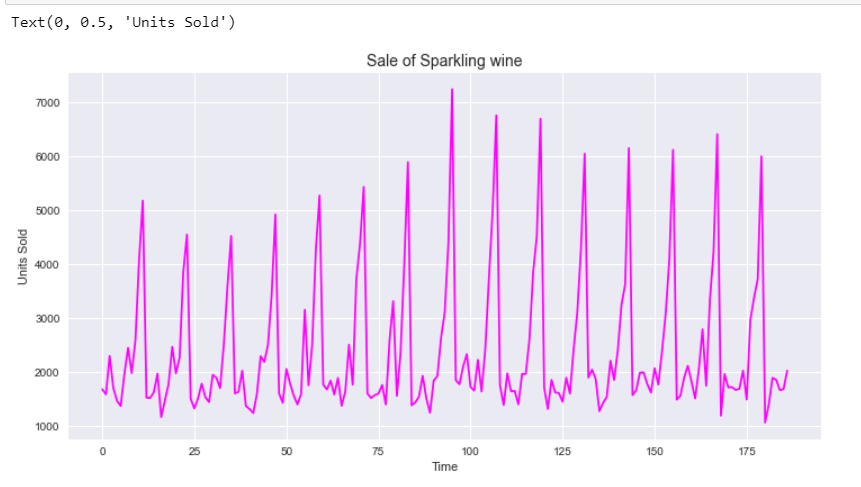


Fig 1.a Trend Seasonality - Sparkling

The data shows a trend though not consistent with significant seasonality as well. The trend is on the lower side initially and one’s a peak in of sales of 7000 post which its sales decline but not much.

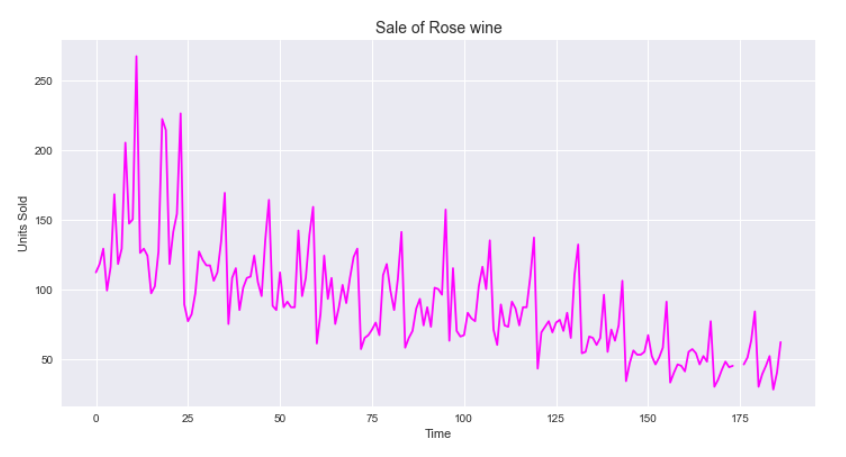


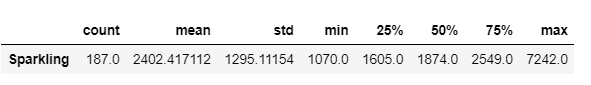
Fig 1.b Trend Seasonality - Rose

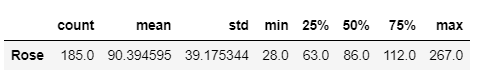
The data shows a declining trend with significant seasonality as well. The trend is on the lower side throughout. The sales are declining from 267 to 28 continuously.

* **The data for Rose has missing data so it has been imputed with Linear Method, the data being imputed in daily wise mean to get more accurate Prediction since we have monthly data of all years.**

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

* We will first see the exploratory Data Analysis when Linear Method and combining two models were not done.





For Sparkling we see the minimum and maximum has a lot of variances hence it can be interpreted to have outliers. The Description for Rose also has a lot of variances between minimum and maximum values which means a lot of outliers will exist.

* The outliers are as below:

Sparkling: **The yearly plot for outliers:**

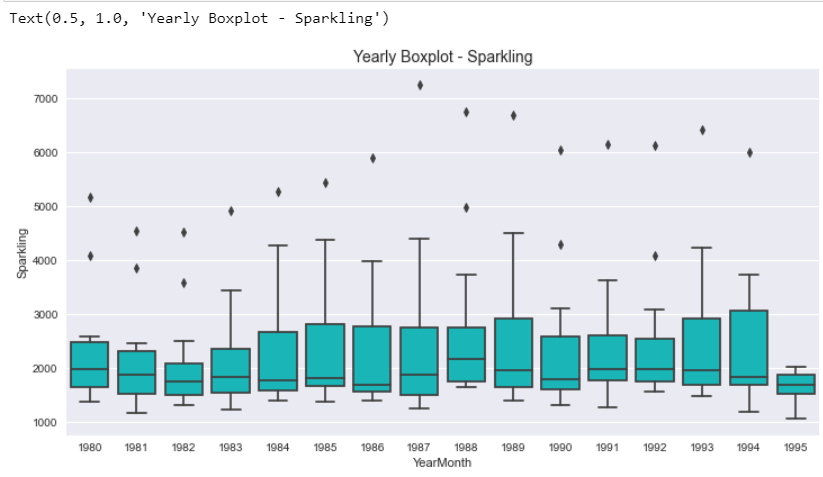


Fig 2.a Yearly box plot-Sparkling

INTERPRETATION-

1. The data shows an outlier in all the years is present.
2. The highest outlier is in the year 1987.
3. No outlier can be seen in 1995.

ROSE: **The yearly plot for outliers:**

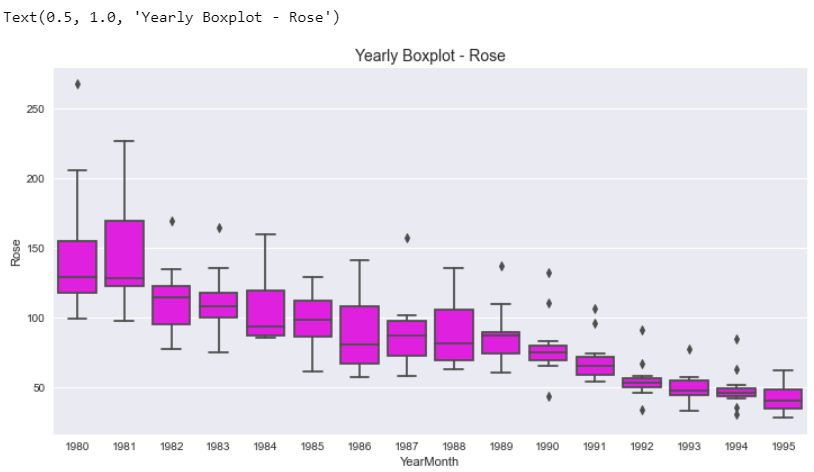


Fig 2.a Yearly box plot-Rose

INTERPRETATION-

1. The data shows an outlier in some of the years is present.
2. The highest outlier is in the year 1980.
3. No outlier can be seen in 1981, 1984,1985,1986,1988 and 1995.

Sparkling: **The Monthly plot for outliers:**

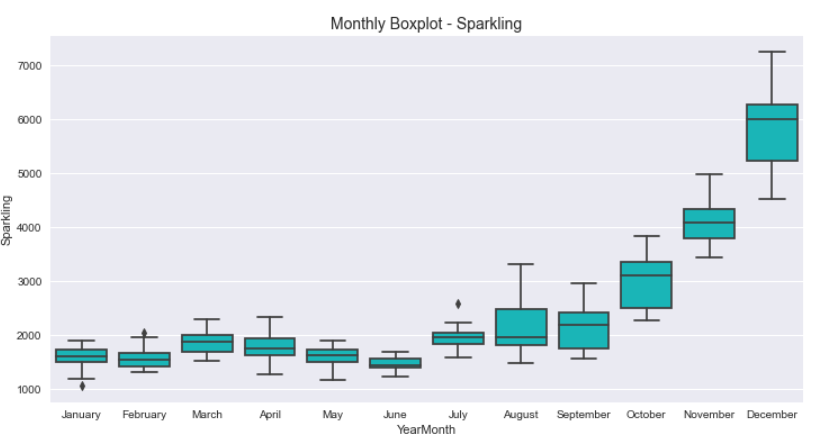


Fig 2.b Monthly box plot-Sparkling

INTERPRETATION-

1. We can see that we have a downward outlier in the month of January. February and July seem to have one outlier. Other months have no outliers.

Rose: **The Monthly plot for outliers:**

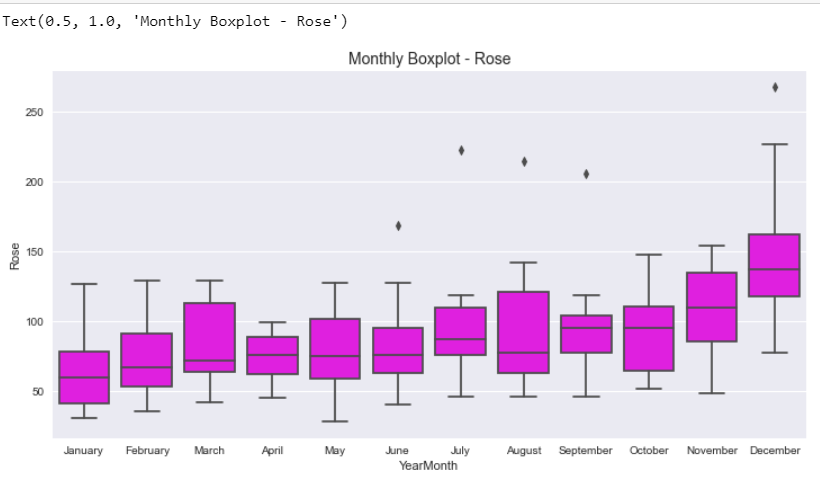
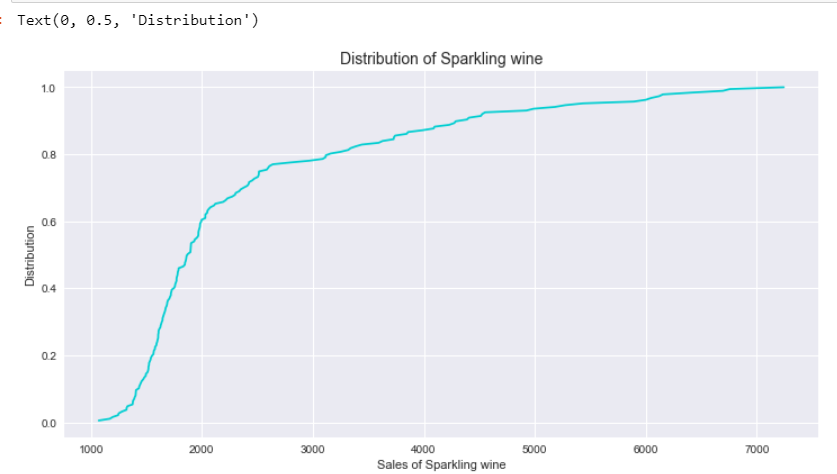
****

Fig 2.b Monthly box plot-Rose

INTERPRETATION-

1. We can see that we have an outlier in the months ranging from June to September and in December. Other months have no outliers.
2. The month of December has comparatively more sales.

* The distribution of Sparkling wine and Rose wine is as below:



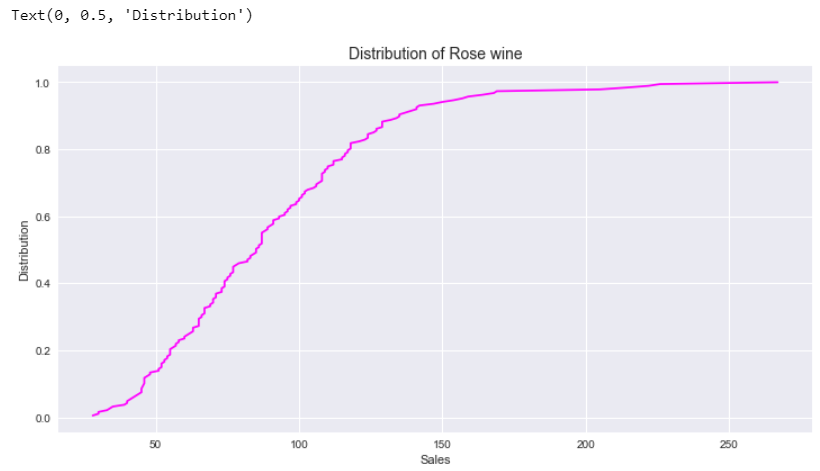
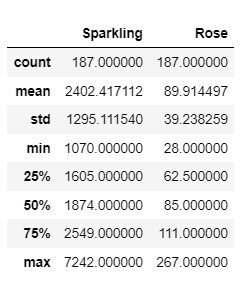


Fig 2.c Data Distribution-Sparkling & Rose

INTERPRETATION:

* The distribution shows in Rose wine starting from 28 and having a constant sale from 100-267 units. The sales seem to rise in small units so the graph looks flattened post 100 units of sales.
* For the distribution of Sparkling wine, we see the sales is rising upward throughout after 100 before which it did see a small fall in sales. The smallest sales are for units 1070 and the highest sales is for 7242 units.
* 

The data shows the minimum count of Sparkling being more as well s maximum sales is more for Sparkling. Rose doesn’t seem to outshine Sparkling in any of the mean, standard deviation or maximum value.

We can again conclude that Sparkling is most preferred at any given season or period.

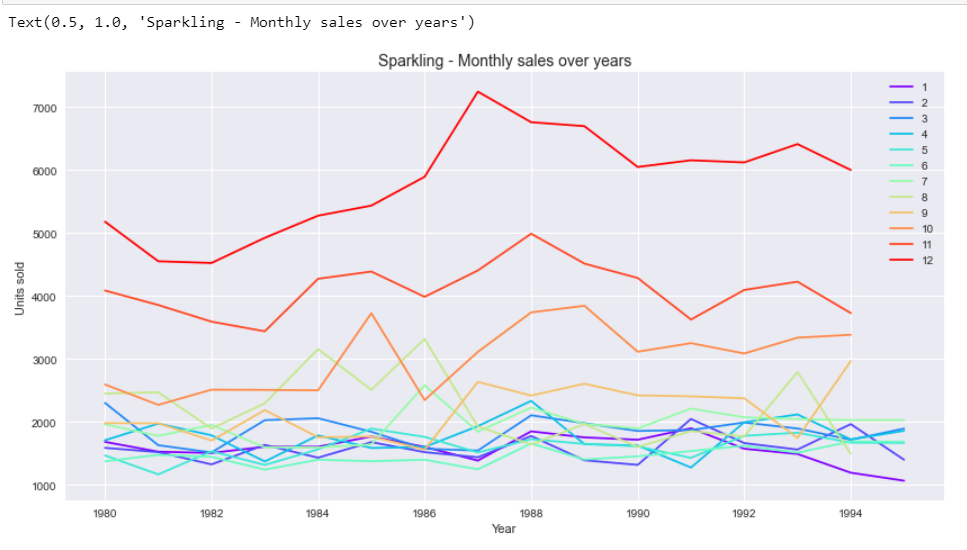
* 

Fig 2.d Data Distribution over years as per months-Sparkling & Rose

INTERPRETATION:

1. The monthly sales are maximum in December for Sparkling wines followed by November and October. We can say that sales more is October to December for Sparkling wines.
2. The sales being the lowest is for January. The sales are also low from all months ranging January to May. However, January is the least sales month.
3. January 1995 and May 1981 are the months-years of the least sales.
4. The sales for October to December have the highest sales of above 7000 as the graph represents for the years in between 1986 to 1988.

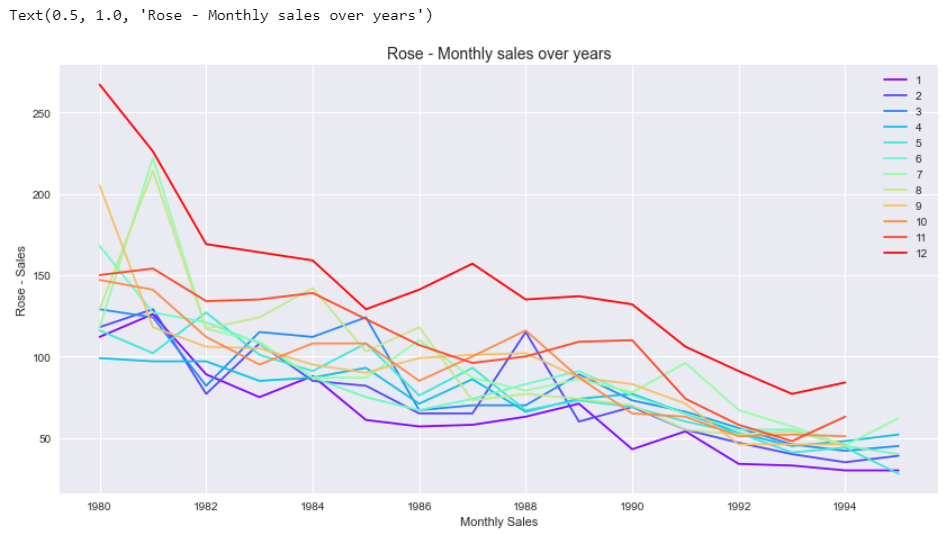
* 

Fig 2.d Data Distribution over years as per months-Rose

INTERPRETATION:

1. The monthly sales are maximum in December for Rose wines followed by July and August. We can say that seasonal season is October to December for Sparkling wines.
2. The sales being the lowest in May 1995. The sales are also low from all months ranging January to May. However, January is the least sales month.
3. May 1995 and January 1995 are the months-years of the least sales.
4. The sales for October to December have the highest sales of above 267 as the graph represents for the years in between 1980 to 1982.

Checking the data together for Sparkling and Rose:

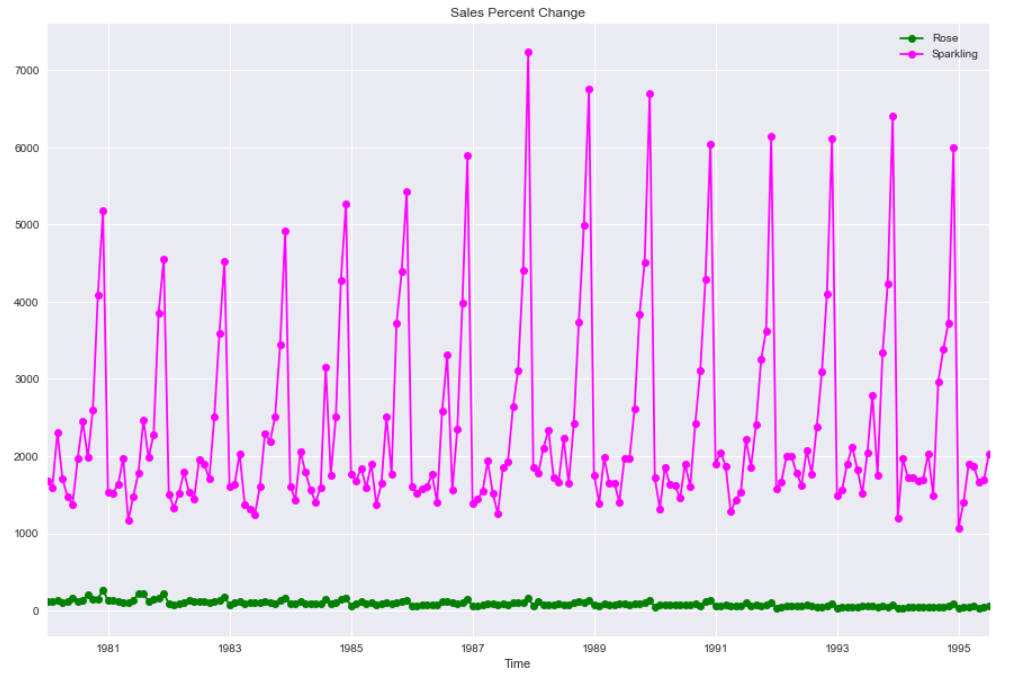


Fig 2.e Data Distribution over years as per months-Rose/Sparkling

INFERENCE:

* The data for Rose wine can be seen lower in sales throughout compared to Sparkling.
* The numbers also suggest the first five top sales month and unit wise that Rose is not much preferred like Sparkling is:

|  |  |
| --- | --- |
| Year Month | Rose |
| 1980-12 | 267 |
| 1981-12 | 226 |
| 1981-07 | 222 |
| 1981-08 | 214 |
| 1980-09 | 205 |

|  |  |
| --- | --- |
| Year Month | Sparkling |
| 1987-12 | 7242 |
| 1988-12 | 6757 |
| 1989-12 | 6694 |
| 1993-12 | 6410 |
| 1991-12 | 6153 |

Decomposing the Time Series is often done to help improve understanding of the time series, but it can also be used to improve forecast accuracy. We can usually identify an additive or multiplicative time series from its variation. If the magnitude of the seasonal component changes with time, then the series is multiplicative. Otherwise, the series is additive. Here we will perform both.

Additive Decomposition and Multiplicative Decomposition

**Sparkling:**

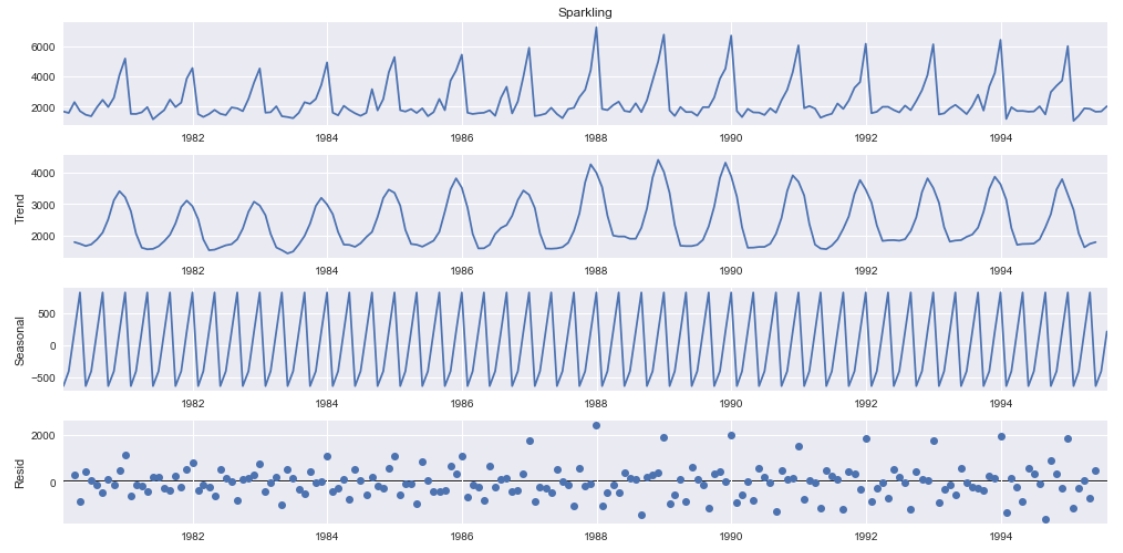


Fig 2.f Additive Decomposition-Sparkling

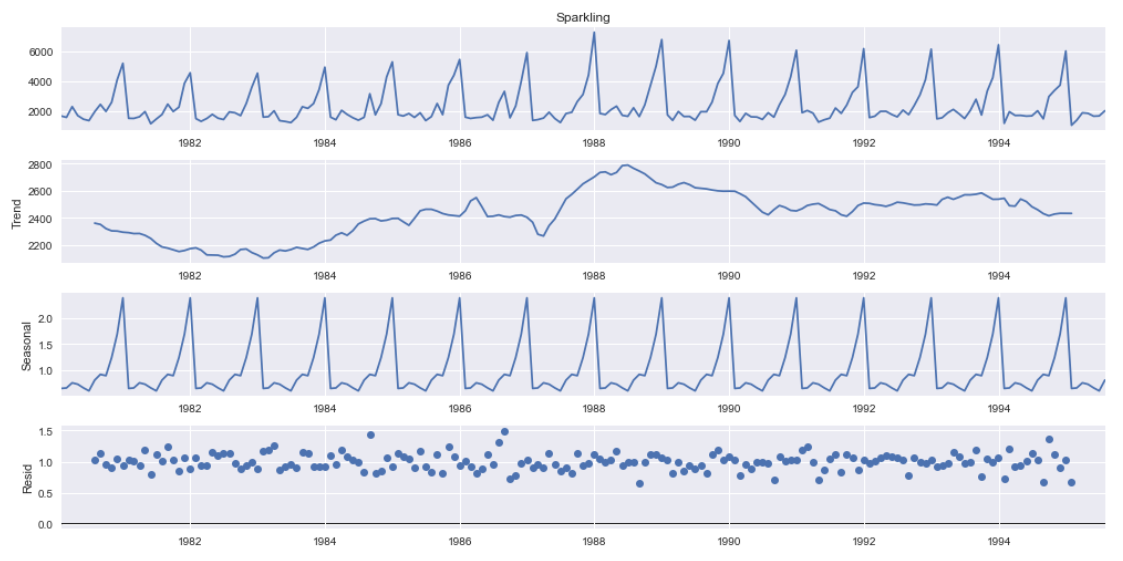


Fig 2.g Multiplicative Decomposition-Sparkling

INFERENCE:

* In additive model we can see the trend is consistent rising and falling quite consistently over all periods which is not likewise in Multiplicative model. The multiplicative model has a trend wherein we see a rise then a drop with a consistent rise until 1988 which has the highest peak of the model. Post 1990 we see a consistency in the data till 1994.
* The trend for both the models can be seen quite same which makes the data non-volatile since it doesn’t depend on seasonality.
* Additive residual is ranging from -1 to 2000 which makes us sure that multiplicative model is the best suit for the data as residual is less 0.5 to 1.5.
* More the residual more the errors in the model and lesser the residual better the model with non-volatile data. Volatility hampers seasonality which is why additive model shows frequent high and low in the decomposition under seasonality section.

**Rose:**

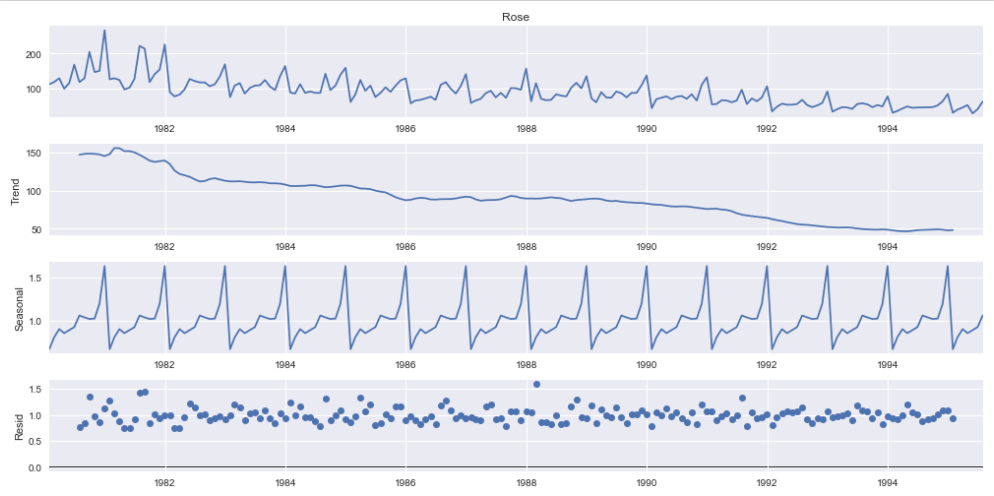
****

Fig 2.f Multiplicative Decomposition-ROSE

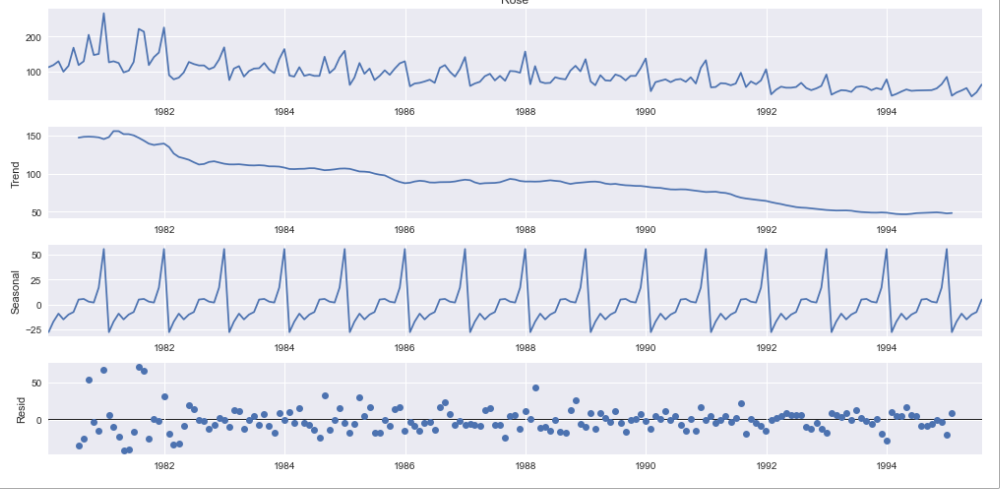


Fig 2.f Additive Decomposition-ROSE

INFERENCE:

* In additive model we can see the trend we see the trend is consistently falling which is also the same in multiplicative model. Exponential dips is seen between 1981-1983 and later 1991-1993.
* The trend for both the models can be seen quite same which makes the data non-volatile since it doesn’t depend on seasonality. The variance in seasonality for additive model is 25-50 and that for multiplicative is 16%
* Additive residual is ranging from 0 to 50 which makes us sure that multiplicative model is the best suit for the data as residual is less 0.5 to 1.5.
* More the residual more the errors in the model and lesser the residual better the model with non-volatile data. Volatility hampers seasonality which is why additive model shows frequent high and low in the decomposition under seasonality section.
* Also, to add, if seasonality peaks are consistently reducing altitude with trend, we will conclude saying for ROSE multiplicative model is the best.

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

The train and test have been split as asked in the question 1991 wise. This means 70% train and 20% test.

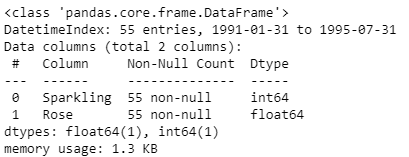
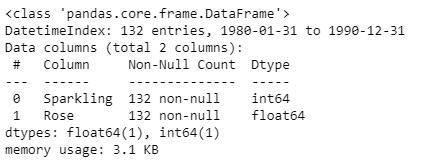


Fig 3. a Test DATA information

Fig 3. a Train DATA information

The data suggests no null value in train or test data. Also the Sales is an integer data type and the year is a float data type.

The train data has 55 columns and 2 columns. The test data has 55 rows and 2 columns:



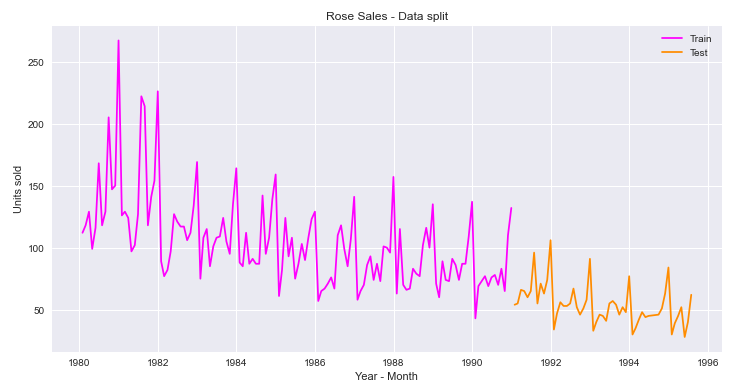


Fig 3. b Rose graphical representation of Train and Test

Inference: The Data for Train and Test for ROSE wine looks like above wherein we have data up to 1991 in Pink taken for TRAIN and Orange post 1991 as Test.

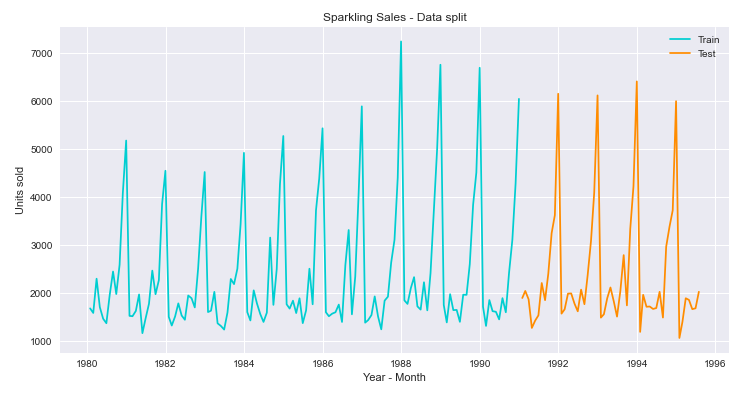


Fig 3. b Sparkling graphical representation of Train and Test

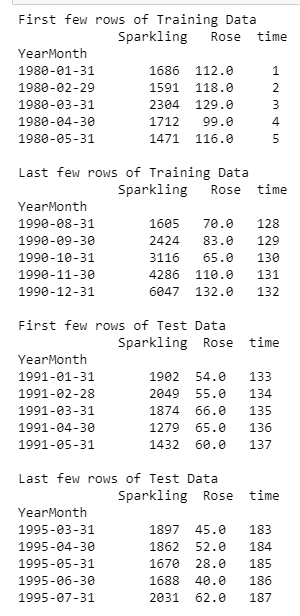
Inference: The Data for Train and Test for Sparkling wine looks like above wherein we have data up to 1991 in blue taken for TRAIN and Orange post 1991 as Test.

4. Build all the 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.

Exponential smoothing is a time series forecasting method for univariate data. Exponential smoothing is usually used to make short term forecasts, as longer-term forecasts using this technique can be quite unreliable. Simple (single) exponential smoothing uses a weighted moving average with exponentially decreasing weights.

LINEAR REGRESSION:

We use linear regression for time series analysis, it is used for predicting the result for time series as its trends. For example, here, we have a dataset of time series with the help of linear regression we can predict the sales with the time.



INFERENCE:

* We see that the training data set has dataset until 132 rows and the columns are for year-month and sales. The time column has been added for prediction.
* The Test Set has columns from 133-187 both for Rose and Sparkling wine.

Fig 4. a. Training and Test Data for Linear regression

Plot for LINEAR REGRESSION:

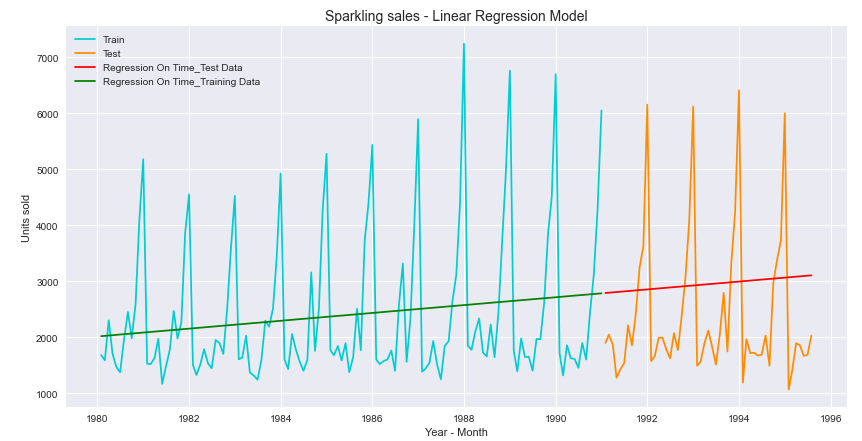


Fig 4. b. Training and Test Data for Linear regression- SPARKLING WINE

INFERENCE:

* We see that the forecast shows us an upward rising sale.
* The train and test data if seen in the image about the trend is consistent so we can expect the sales to be rising upward in future.
* 
* The RMSE value for Training data is 1279.322 which is too high and indicates the data to have errors and not have correct forecasting. The value for MAPE is 40.05 which means it is reasonable a forecasting but not great or accurate.
* The RMSE value of testing is higher with a MAPE of 50% error the forecasting is leaving us with.

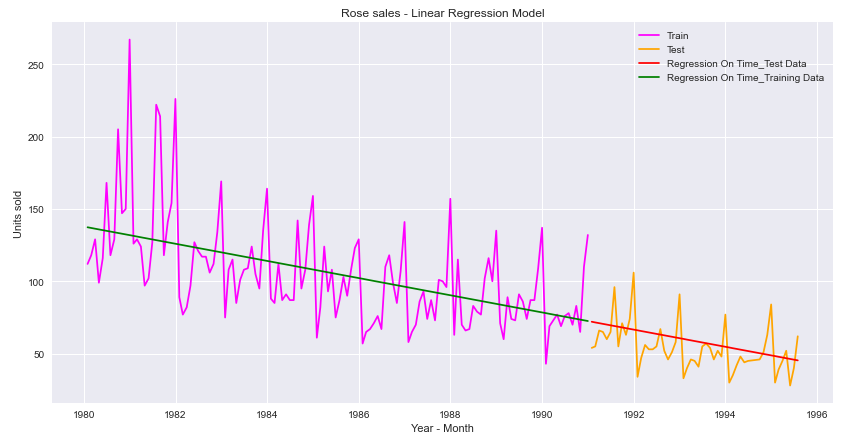


Fig 4. b. Training and Test Data for Linear regression- ROSE WINE

INFERENCE:

* We see that the forecast shows us a downward declining sale.
* The train and test data if seen in the image about the trend is inconsistent so we can expect the sales to go further down in future in the company under ROSE wine sales.
* 
* The RMSE value for Training data is 30.718 which is too high and indicates the data to have errors and not have correct forecasting. The value for MAPE is 21.22 which means it is reasonable a forecasting but not great or accurate.
* The RMSE value for Testing is lower than training confirming the forecasting leaving us an error of 22.82% or 23% error in prediction.

**Comparing linear regression results for both wines that we are selling ROSE will decline in future and the error percentage in forecast of Sparkling is more.**

Plot for NAÏVE FORECAST:

A naive forecast involves using the previous observation directly as the forecast without any change. It is often called the persistence forecast as the prior observation is persisted. This simple approach can be adjusted slightly for seasonal data.

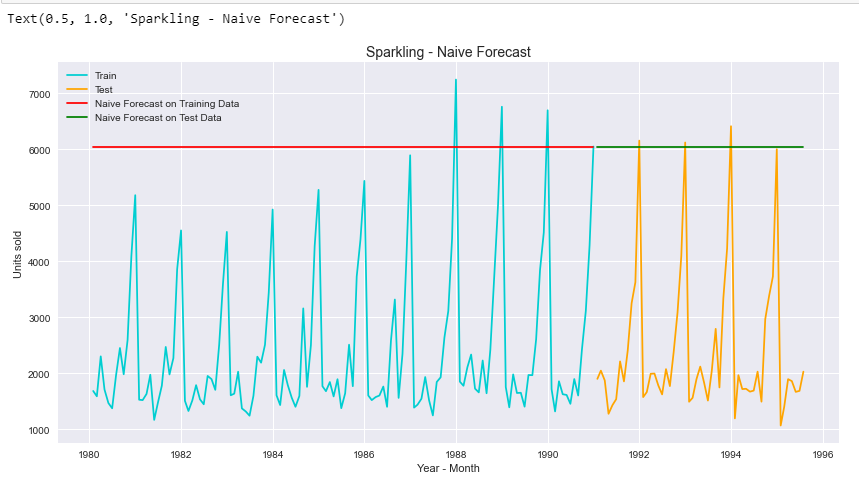


Fig 4. c. Training and Test Data for NAÏVE MODEL- SPARKLING WINE

INFERENCE:

We see that here the test and train data does have a trend. The trend in test is more than train since the volume taken there is less so the trend will impact the forecasting.

* The forecasting is a straight line because in NAÏVE we take today’s data for tomorrow and tomorrows data for the day after. Hence this will cause the data to look more or less the same.
* 
* The predictions under NAÏVE is completely vague and inaccurate with very high scores for RMSE model. The percentage of error or MAPE is also high which indicates the error in forecasting under training is 153% and under test it is 152.87 %.

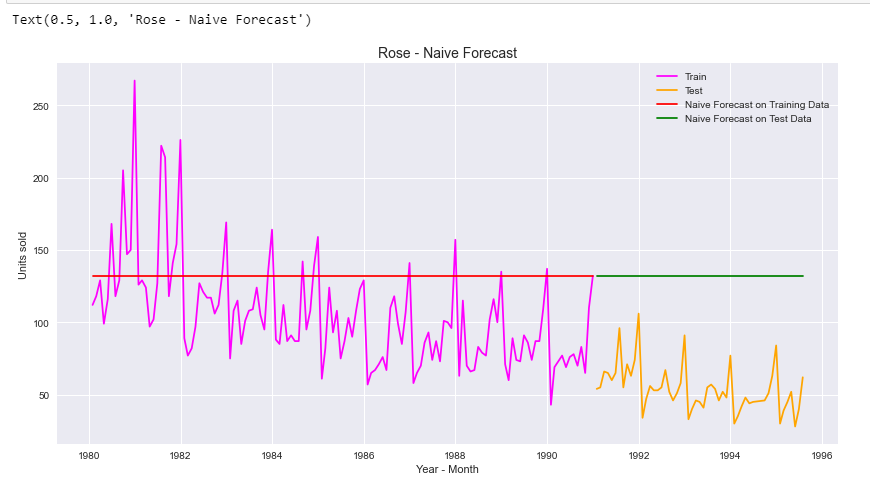


Fig 4. c. Training and Test Data for NAÏVE MODEL- ROSE WINE

INFERENCE:

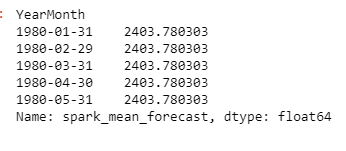
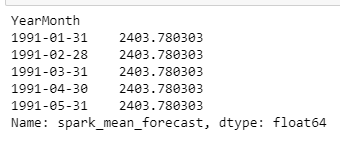
We see that here the test and train data does have a trend. The trend in test is more than train since the volume taken there is less so the trend will impact the forecasting.

* The forecasting is a straight line because in NAÏVE we take today’s data for tomorrow and tomorrows data for the day after. Hence this will cause the data to look more or less the same.
* 
* The predictions under NAÏVE is completely vague and inaccurate with very high scores for RMSE model. The percentage of error or MAPE % for Training data can be considered as a reasonable forecasting since scores is below 50%. However, for MAPE in testing we have high inaccuracy or error percentage of 145%.

**Comparing NAÏVE results for both wines that we are selling: The NAÏVE model is not suppose to be used or is an unfit for the data we have been provided with. It predicts the forecasting for both ROSE and SPARKLING with high error percentage making this test a failure to be used for our data.**

Plot for SIMPLE AVERAGE:

Such forecasting technique which forecasts the expected value equal to the average of all previously observed points is called Simple Average technique. We take all the values previously known, calculate the average and take it as the next value.

TEST

TRAIN

Fig 4. d. Training and Test Data for Simple average- Sparkling

The above train and test set is how our data looks post taking average of the data available. Since this simple average method is about mean of the values available with us, we see all the forecasting samples of the same value.

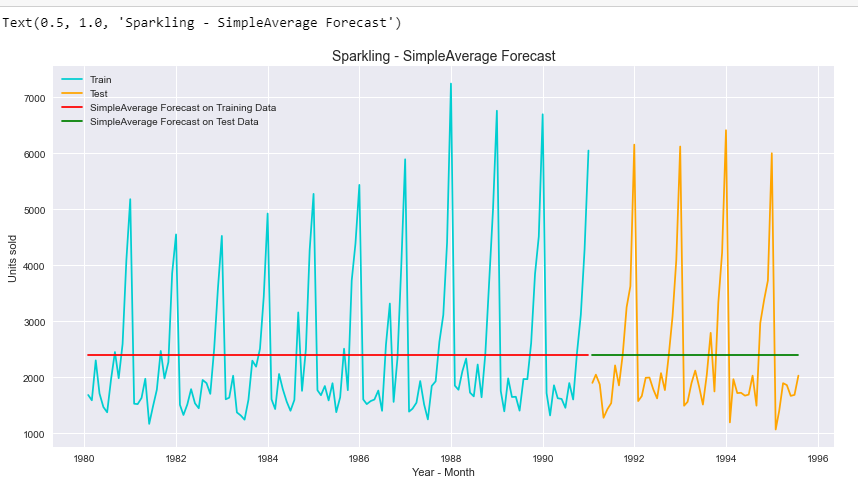
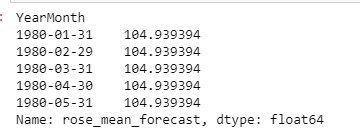
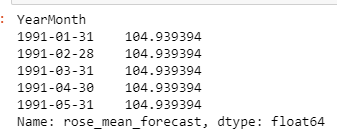


Fig 4. e. Simple average- Sparkling

INFERENCE:

* The data can be seen incapable of capturing any trend or seasonality. The same line throughout is because we have taken out the mean for the amounts.
* 

The RMSE scores for both training and testing is high which shows Simple Average cannot be used to forecast sales of our data. The error % for both training and testing data is below 50% which means its reasonable but cannot be relied on.

TRAIN

TEST

Fig 4. f. Training and Test Data for Simple average- ROSE

The above train and test set is how our data looks post taking average of the data available. Since this simple average method is about mean of the values available with us, we see all the forecasting samples of the same value.

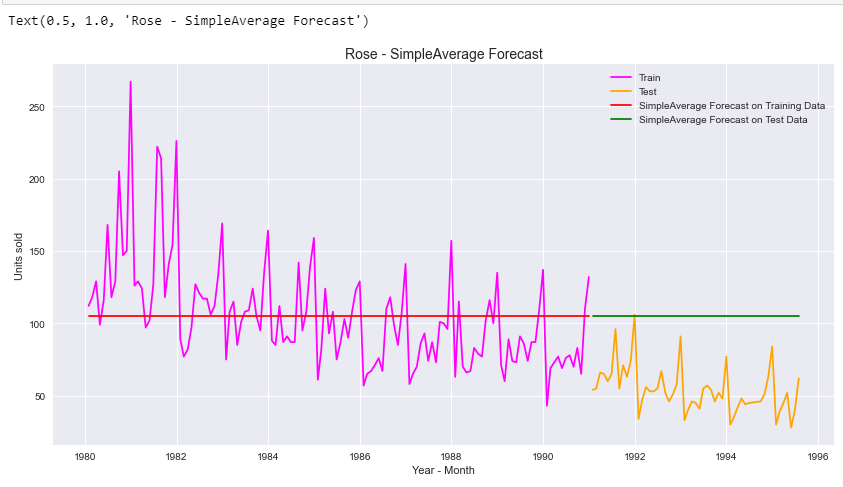


Fig 4. f. Simple average- ROSE

INFERENCE:

* The data can be seen incapable of capturing any trend or seasonality. The same line throughout is because we have taken out the mean for the amounts.

Training data and testing data is highly inconsistent with the prediction as we seen test is below the mean line.

* 

The RMSE scores for both training and testing is high which shows Simple Average cannot be used to forecast sales of our data. The error % for training is 25% which is considerate but testing data is higher than 50% which means the model cannot be trusted and the forecasting has high inaccuracy in its prediction.