PGP - Data Science and Business Analytics

TIME SERIES FORECASTING PROJECT REPORT

Akshaya Nallathambi

7th November, 2021

Table Of Contents

List of Figures

List of Tables

List of Graphs

Problem 1

Problem statement-

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.

Data Description-

SPARKLING WINE:

YearMonth: Date of sales

Sparkling: Wine sales

ROSE WINE:

YearMonth: Date of sales

Rose: Wine sales

Sample of the dataset-

SPARKLING WINE:

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471

FIGURE 1

There are 2 variables out of which 1 is int value and 1 is object value . The data given is for 187 individual sales. The dataset does not contain any null values.

ROSE WINE:

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

FIGURE 2

There are 2 variables out of which 1 is int value and 1 is object value. The data given is for 187 individual sales. The dataset contains null values.

Types of variables in the data frame-

SPARKLING WINE:

YearMonth	object	Categorical
Sparkling	int64	Continuous

TABLE 1

ROSE WINE:

YearMonth	object	Categorical
Rose	int64	Continuous

TABLE 2

1. Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.

The below figures give the first 5 rows of sample data from both Sparkling and Rose data set.

YearMonth	Sparkling
1980-01	1686
1980-02	1591
1980-03	2304
1980-04	1712
1980-05	1471

FIGURE 3

YearMonth	Rose
1980-01	112.0
1980-02	118.0
1980-03	129.0
1980-04	99.0
1980-05	116.0

FIGURE 4

The image below gives the basic information of the data set. It is clear that the variables have int and object data types with 2 columns and 187 rows. The dataset has no null values. The memory usage is 3.0+ KB.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   YearMonth   187 non-null    object
 1   Sparkling   187 non-null    int64
dtypes: int64(1), object(1)
memory usage: 3.0+ KB
```

FIGURE 5

The image below gives the basic information of the data set. It is clear that the variables have int and object data types with 2 columns and 187 rows. The dataset has 2 null values that will be removed. The memory usage is 3.0+ KB.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   YearMonth   187 non-null    object
 1   Rose        185 non-null    float64
dtypes: float64(1), object(1)
memory usage: 3.0+ KB
```

FIGURE 6

The below image gives the five point summary of the continuous variables in the data set. There is no need for scaling. There is skewness in the data when we take into account the standard deviation and the maximum value.

	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

FIGURE 7

The below image gives the five point summary of the continuous variables in the data set. There is no need for scaling. There is skewness in the data when we take into account the standard deviation and the maximum value.

	Rose
count	185.000000
mean	90.394595
std	39.175344
min	28.000000
25%	63.000000
50%	86.000000
75%	112.000000
max	267.000000

FIGURE 8

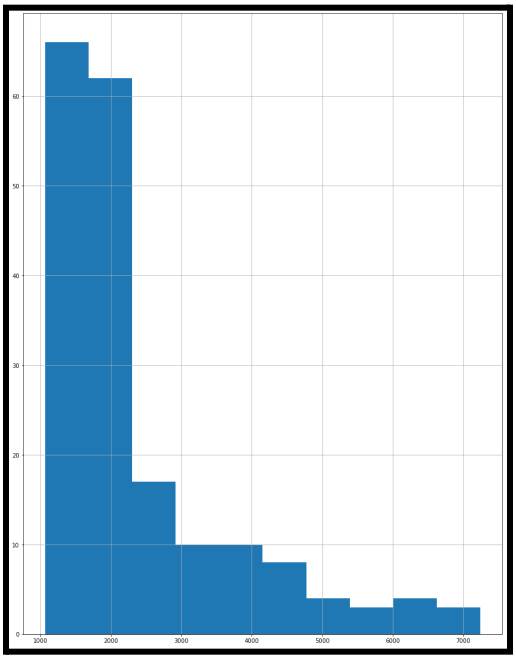
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Univariate analysis:

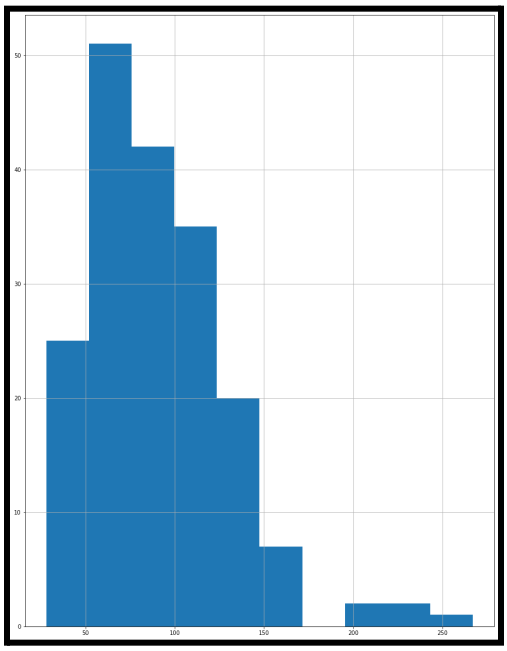
Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. [1]

The histograms are used for numerical variables and count plots are used for categorical variables to perform univariate analysis.

It is clear from the graph (*GRAPH 1 and 2*) that the numerical variable is skewed.



GRAPH 1 (Sparkling**)**



GRAPH 2 (Rose**)**

Skewness:

Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value. [2]

Skewness level	**Value**
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ±1

TABLE 3

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

The table below shows the skewness value corresponding to the variable in the given data set.

	Skewness
Sparkling	1.802999

FIGURE 9

	Skewness
Rose	1.256176

FIGURE 10

Both the data set has a positive value and it is above 1. Therefore, they are highly right skewed.

Bivariate analysis:

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. [3]

Box plots are used for categorical with numerical variables to perform bivariate analysis.

Box plot:

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell

you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed. [4]

Multiplicative decomposition:

The multiplicative decomposition argues that time series data is a function of the product of its components. Thus,

$$Y = T*S*R$$

where,

Y is the time series data

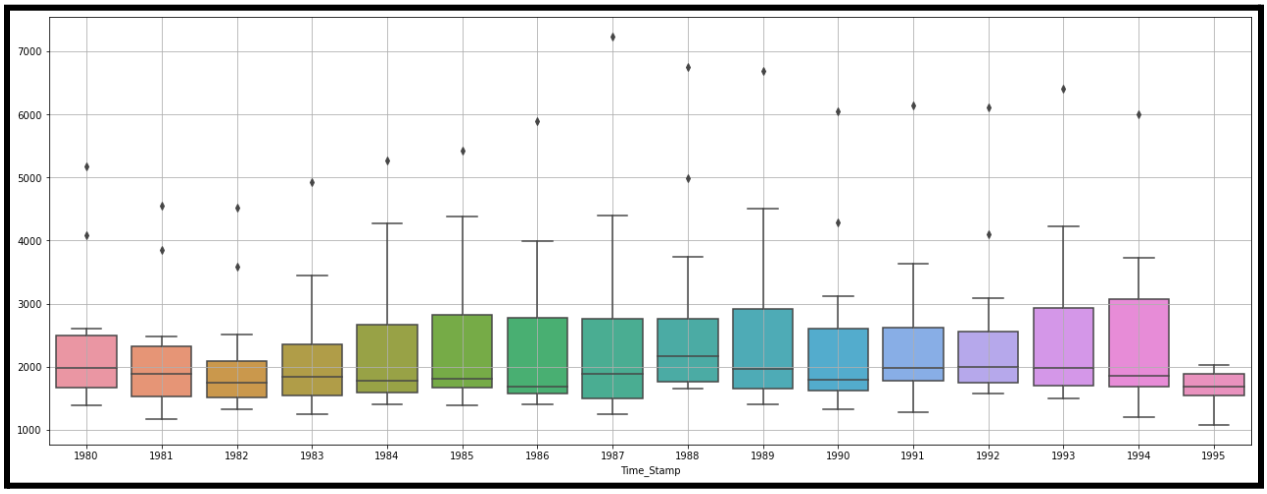
T is the trend-cycle component

S is the seasonal component

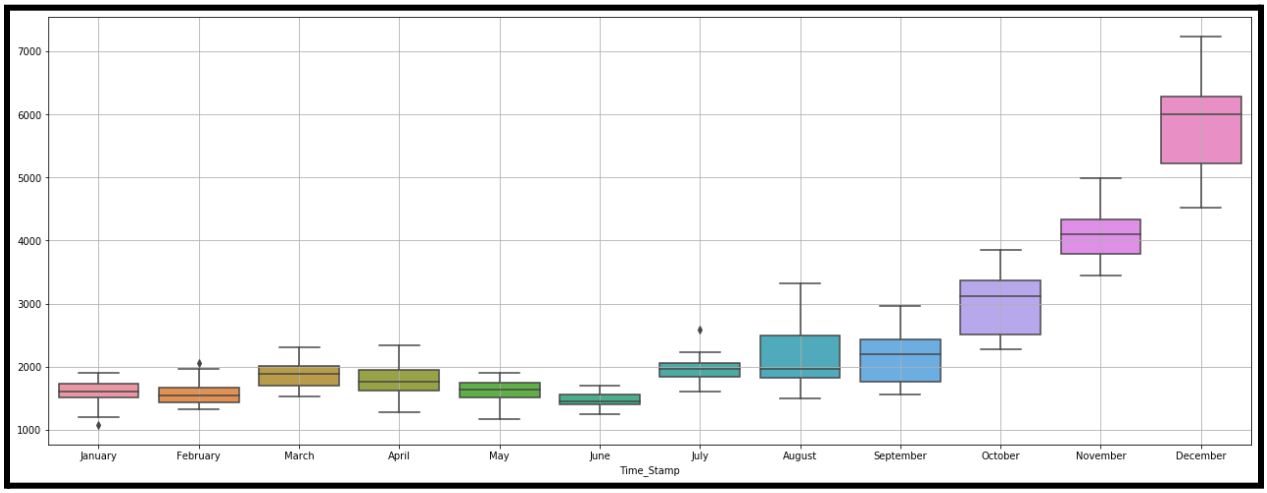
R is the remainder. [5]

SPARKLING WINE:

Box plots-



GRAPH 3



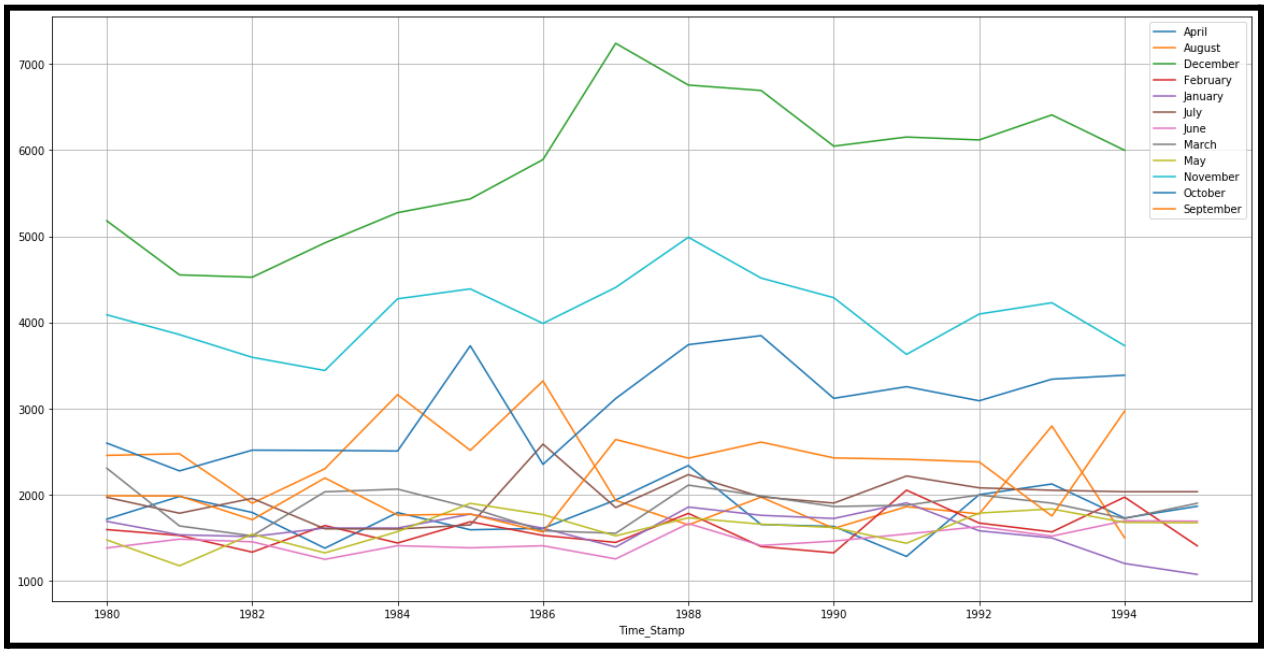
GRAPH 4

Pivot table-

Time_Stamp	April	August	December	February	January	July	June	March	May	November	October	September
Time_Stamp												
1980	1712.0	2453.0	5179.0	1591.0	1686.0	1966.0	1377.0	2304.0	1471.0	4087.0	2596.0	1984.0
1981	1976.0	2472.0	4551.0	1523.0	1530.0	1781.0	1480.0	1633.0	1170.0	3857.0	2273.0	1981.0
1982	1790.0	1897.0	4524.0	1329.0	1510.0	1954.0	1449.0	1518.0	1537.0	3593.0	2514.0	1706.0
1983	1375.0	2298.0	4923.0	1638.0	1609.0	1600.0	1245.0	2030.0	1320.0	3440.0	2511.0	2191.0
1984	1789.0	3159.0	5274.0	1435.0	1609.0	1597.0	1404.0	2061.0	1567.0	4273.0	2504.0	1759.0
1985	1589.0	2512.0	5434.0	1682.0	1771.0	1645.0	1379.0	1846.0	1896.0	4388.0	3727.0	1771.0
1986	1605.0	3318.0	5891.0	1523.0	1606.0	2584.0	1403.0	1577.0	1765.0	3987.0	2349.0	1562.0
1987	1935.0	1930.0	7242.0	1442.0	1389.0	1847.0	1250.0	1548.0	1518.0	4405.0	3114.0	2638.0
1988	2336.0	1645.0	6757.0	1779.0	1853.0	2230.0	1661.0	2108.0	1728.0	4988.0	3740.0	2421.0
1989	1650.0	1968.0	6694.0	1394.0	1757.0	1971.0	1406.0	1982.0	1654.0	4514.0	3845.0	2608.0
1990	1628.0	1605.0	6047.0	1321.0	1720.0	1899.0	1457.0	1859.0	1615.0	4286.0	3116.0	2424.0
1991	1279.0	1857.0	6153.0	2049.0	1902.0	2214.0	1540.0	1874.0	1432.0	3627.0	3252.0	2408.0
1992	1997.0	1773.0	6119.0	1667.0	1577.0	2076.0	1625.0	1993.0	1783.0	4096.0	3088.0	2377.0
1993	2121.0	2795.0	6410.0	1564.0	1494.0	2048.0	1515.0	1898.0	1831.0	4227.0	3339.0	1749.0
1994	1725.0	1495.0	5999.0	1968.0	1197.0	2031.0	1693.0	1720.0	1674.0	3729.0	3385.0	2968.0
1995	1862.0	NaN	NaN	1402.0	1070.0	2031.0	1688.0	1897.0	1670.0	NaN	NaN	NaN

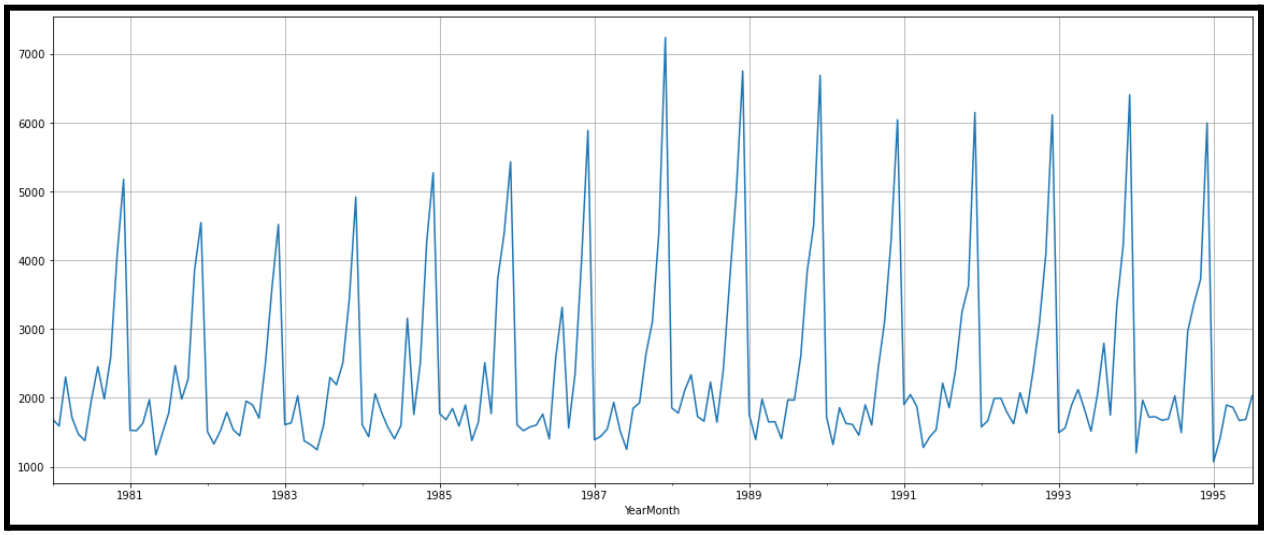
FIGURE 11

Monthly plots-



GRAPH 5

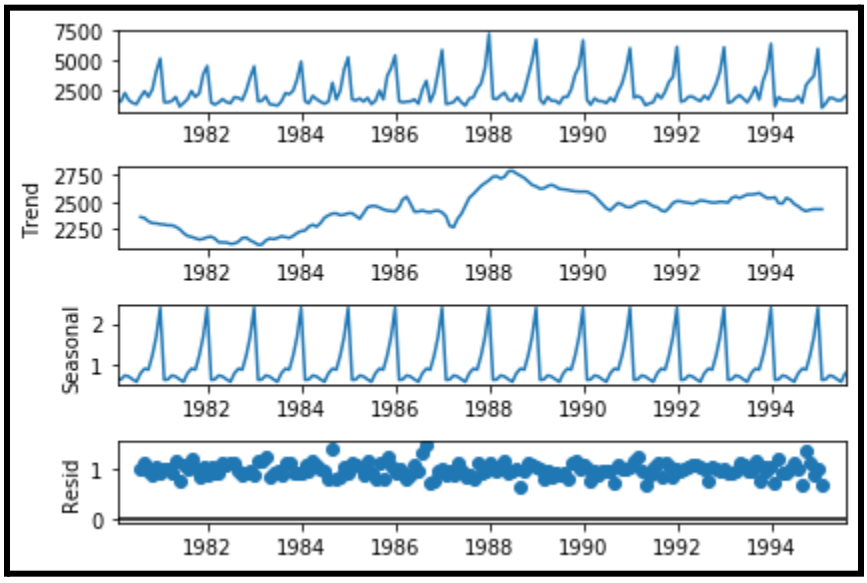
Yearwise plots-



GRAPH 6

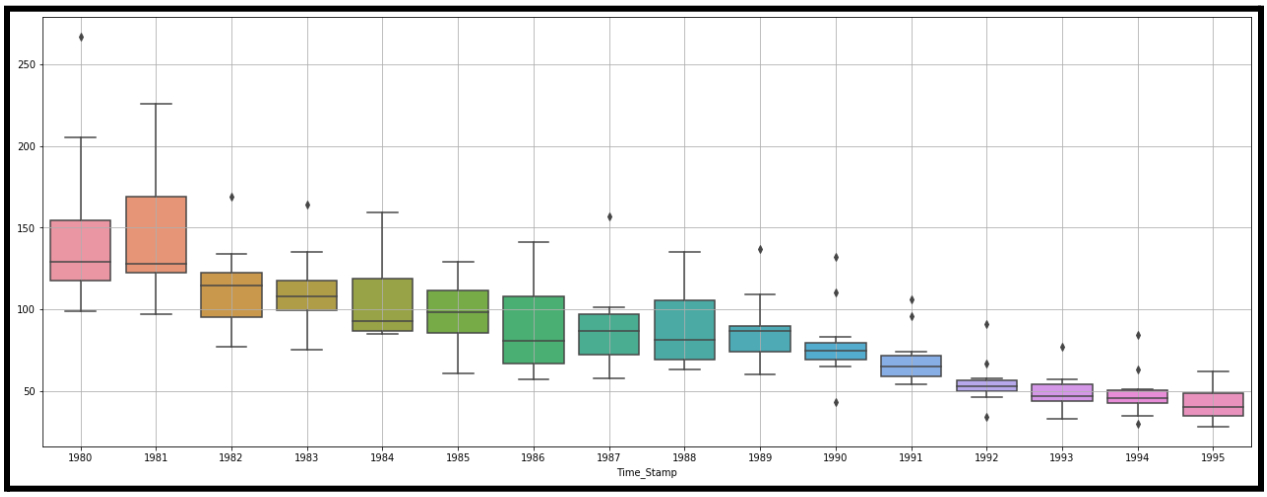
There doesn't seem any particular trend but it appears like a seasonal fluctuation and a pattern that follows in every cycle.

Multiplicative seasonal decomposition-

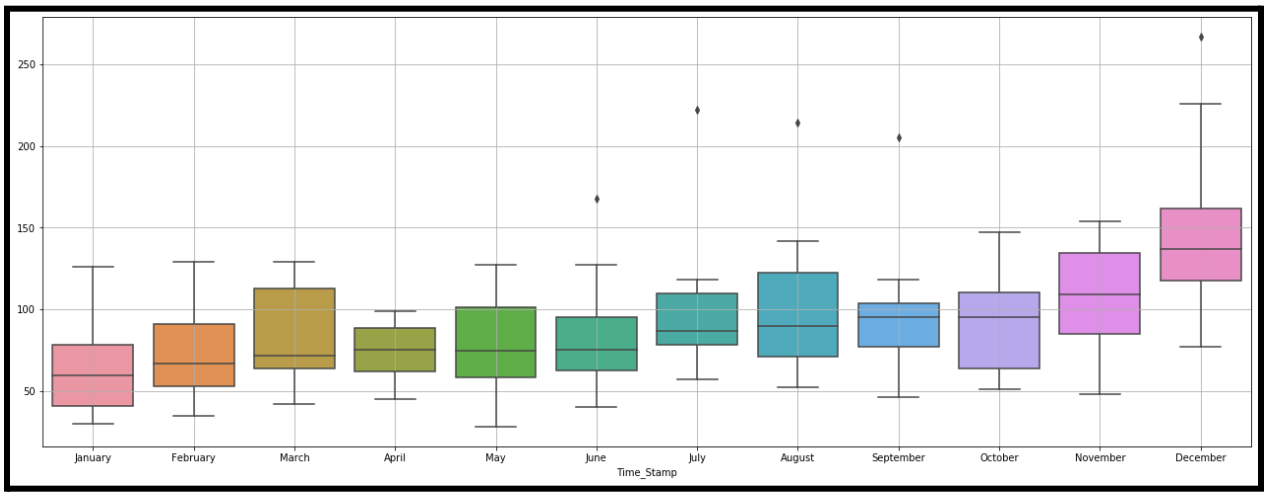


GRAPH 7

ROSE WINE:



GRAPH 8



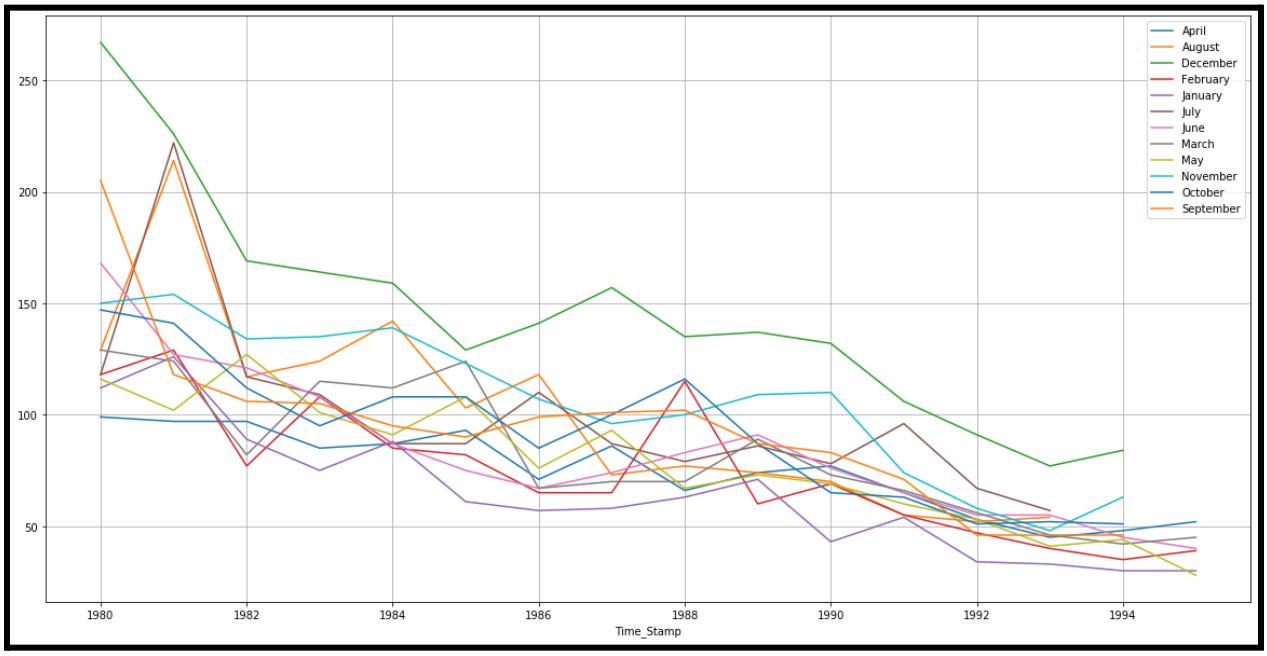
GRAPH 9

Pivot table-

Time_Stamp	April	August	December	February	January	July	June	March	May	November	October	September
Time_Stamp												
1980	99.0	129.0	267.0	118.0	112.0	118.0	168.0	129.0	116.0	150.0	147.0	205.0
1981	97.0	214.0	226.0	129.0	126.0	222.0	127.0	124.0	102.0	154.0	141.0	118.0
1982	97.0	117.0	169.0	77.0	89.0	117.0	121.0	82.0	127.0	134.0	112.0	106.0
1983	85.0	124.0	164.0	108.0	75.0	109.0	108.0	115.0	101.0	135.0	95.0	105.0
1984	87.0	142.0	159.0	85.0	88.0	87.0	87.0	112.0	91.0	139.0	108.0	95.0
1985	93.0	103.0	129.0	82.0	61.0	87.0	75.0	124.0	108.0	123.0	108.0	90.0
1986	71.0	118.0	141.0	65.0	57.0	110.0	67.0	67.0	76.0	107.0	85.0	99.0
1987	86.0	73.0	157.0	65.0	58.0	87.0	74.0	70.0	93.0	96.0	100.0	101.0
1988	66.0	77.0	135.0	115.0	63.0	79.0	83.0	70.0	67.0	100.0	116.0	102.0
1989	74.0	74.0	137.0	60.0	71.0	86.0	91.0	89.0	73.0	109.0	87.0	87.0
1990	77.0	70.0	132.0	69.0	43.0	78.0	76.0	73.0	69.0	110.0	65.0	83.0
1991	65.0	55.0	106.0	55.0	54.0	96.0	65.0	66.0	60.0	74.0	63.0	71.0
1992	53.0	52.0	91.0	47.0	34.0	67.0	55.0	56.0	53.0	58.0	51.0	46.0
1993	45.0	54.0	77.0	40.0	33.0	57.0	55.0	46.0	41.0	48.0	52.0	46.0
1994	48.0	NaN	84.0	35.0	30.0	NaN	45.0	42.0	44.0	63.0	51.0	46.0
1995	52.0	NaN	NaN	39.0	30.0	62.0	40.0	45.0	28.0	NaN	NaN	NaN

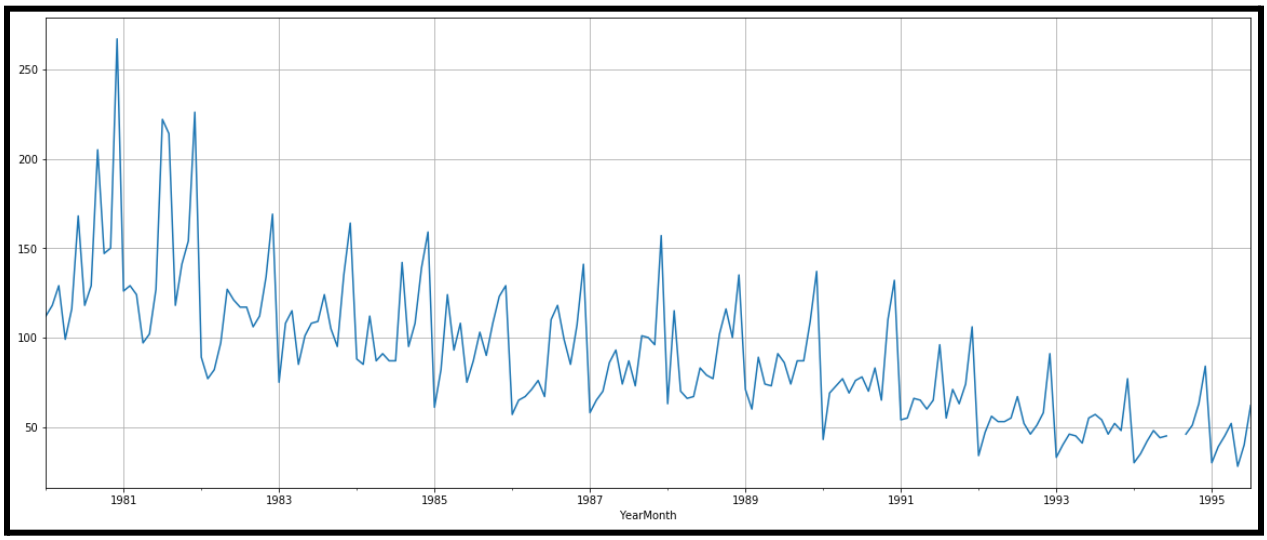
FIGURE 12

Monthly plots-



GRAPH 10

Yearwise plots



GRAPH 11

Decreasing pattern over the years along with a seasonal pattern.

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

Splitting the data using index year as 1991 and viewing few records –

SPARKLING WINE:

The dimensions of the train and test data is (132, 1) and (55, 1) respectively.

First few rows of Training Data

Time_Stamp	Sparkling
1980-01-31	1686
1980-02-29	1591
1980-03-31	2304
1980-04-30	1712
1980-05-31	1471

Last few rows of Training Data

Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

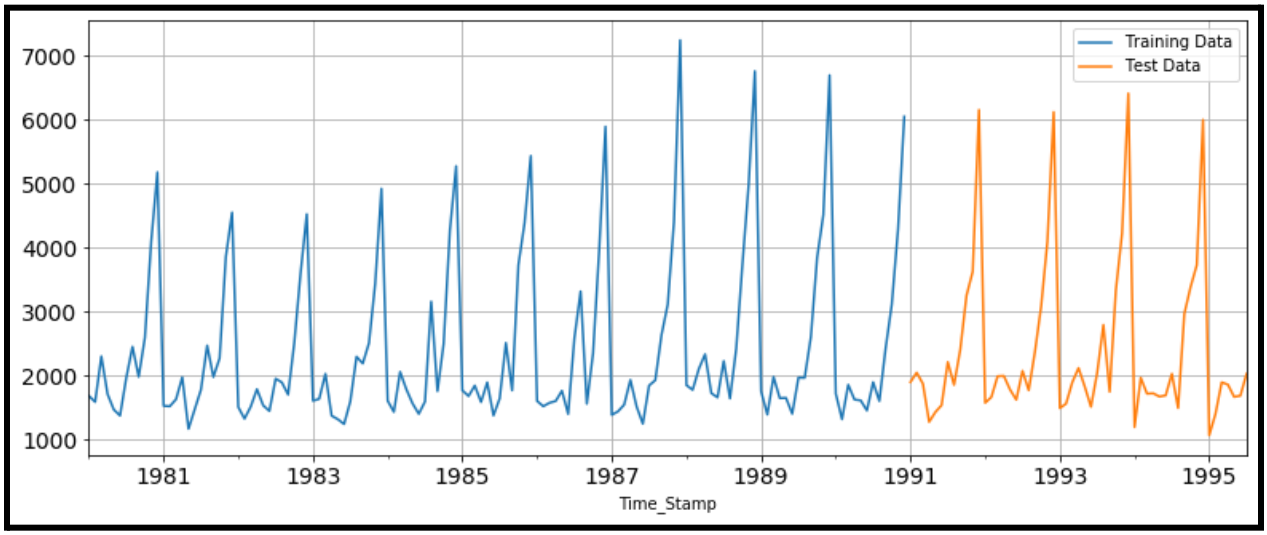
First few rows of Test Data

Time_Stamp	Sparkling
1991-01-31	1902
1991-02-28	2049
1991-03-31	1874
1991-04-30	1279
1991-05-31	1432

Last few rows of Test Data

Time_Stamp	Sparkling
1995-03-31	1897
1995-04-30	1862
1995-05-31	1670
1995-06-30	1688
1995-07-31	2031

FIGURE 13


Training Data
Test Data
7000
6000
5000
4000
3000
2000
1000
1981
1983
1985
1987
1989
1991
1993
1995
Time_Stamp

GRAPH 12

ROSE WINE:

The dimensions of the train and test data is (132, 1) and (53, 1) respectively.

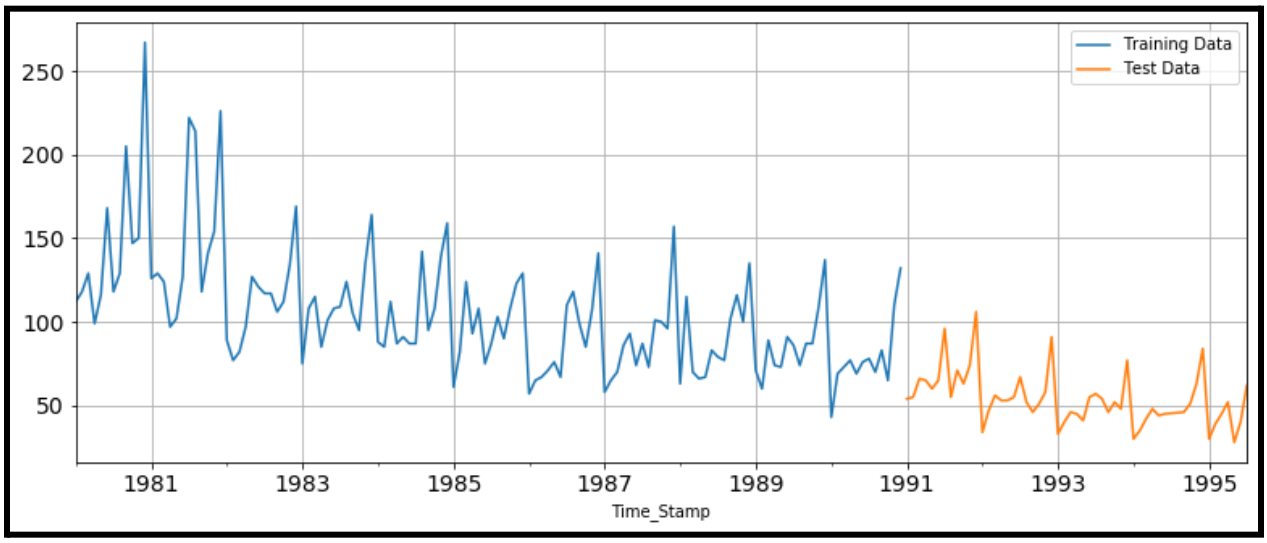
```
First few rows of Training Data
              Rose
Time_Stamp
1980-01-31  112.0
1980-02-29  118.0
1980-03-31  129.0
1980-04-30   99.0
1980-05-31  116.0

Last few rows of Training Data
              Rose
Time_Stamp
1990-08-31   70.0
1990-09-30   83.0
1990-10-31   65.0
1990-11-30  110.0
1990-12-31  132.0

First few rows of Test Data
             Rose
Time_Stamp
1991-01-31  54.0
1991-02-28  55.0
1991-03-31  66.0
1991-04-30  65.0
1991-05-31  60.0

Last few rows of Test Data
             Rose
Time_Stamp
1995-03-31  45.0
1995-04-30  52.0
1995-05-31  28.0
1995-06-30  40.0
1995-07-31  62.0
```

FIGURE 14



GRAPH 13

4. 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, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Linear Regression:

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. [6]

Naïve:

Naïve forecasting is the technique in which the last period's sales are used for the next period's forecast without predictions or adjusting the factors. Forecasts produced using a naïve approach are equal to the final observed value. [7]

Simple Average Forecast:

A simple moving average (SMA) calculates the average of a selected range of prices, usually closing prices, by the number of periods in that range.A simple moving average is a technical indicator that can aid in determining if an asset price will continue or if it will reverse a bull or bear trend. [8]

Moving Average Forecast:

A moving average is a technique to get an overall idea of the trends in a data set; it is an average of any subset of numbers. The moving average is extremely useful for forecasting long-term trends. You can calculate it for any period of time. For example, if you have sales data for a twenty-year period, you can calculate a five-year moving average, a four-year moving average, a three-year moving average and so on. Stock market analysts will often use a 50 or 200 day moving average to help them see trends in the stock market and (hopefully) forecast where the stocks are headed. [9]

RMSE:

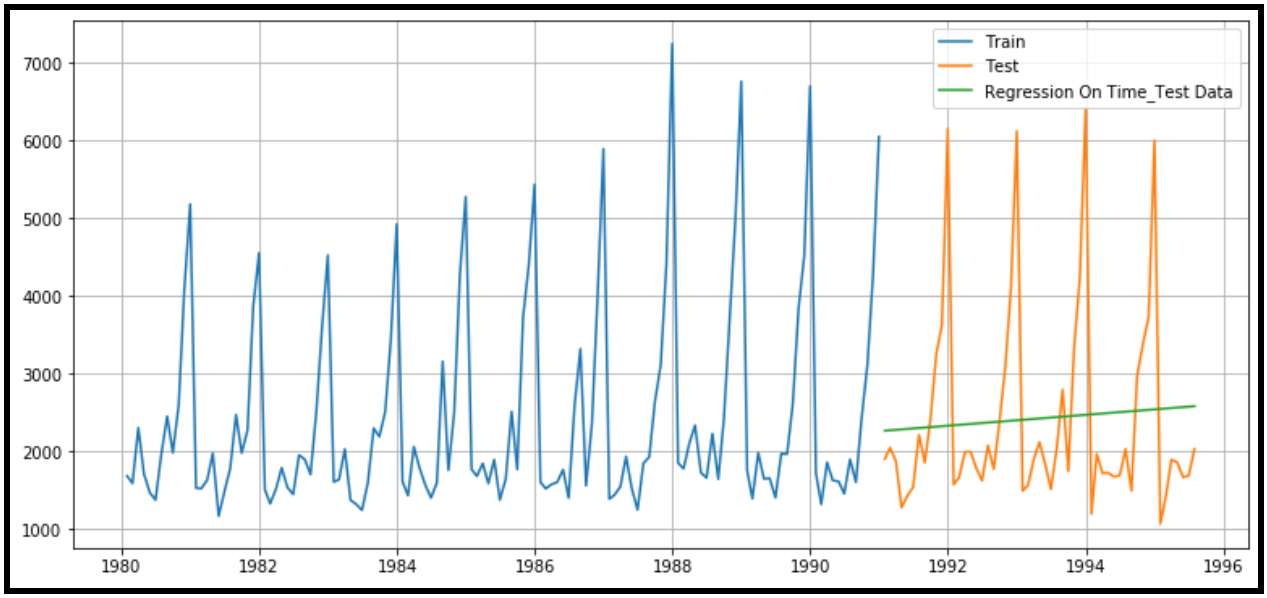
Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you

how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results. [10]

Building various smoothing models –

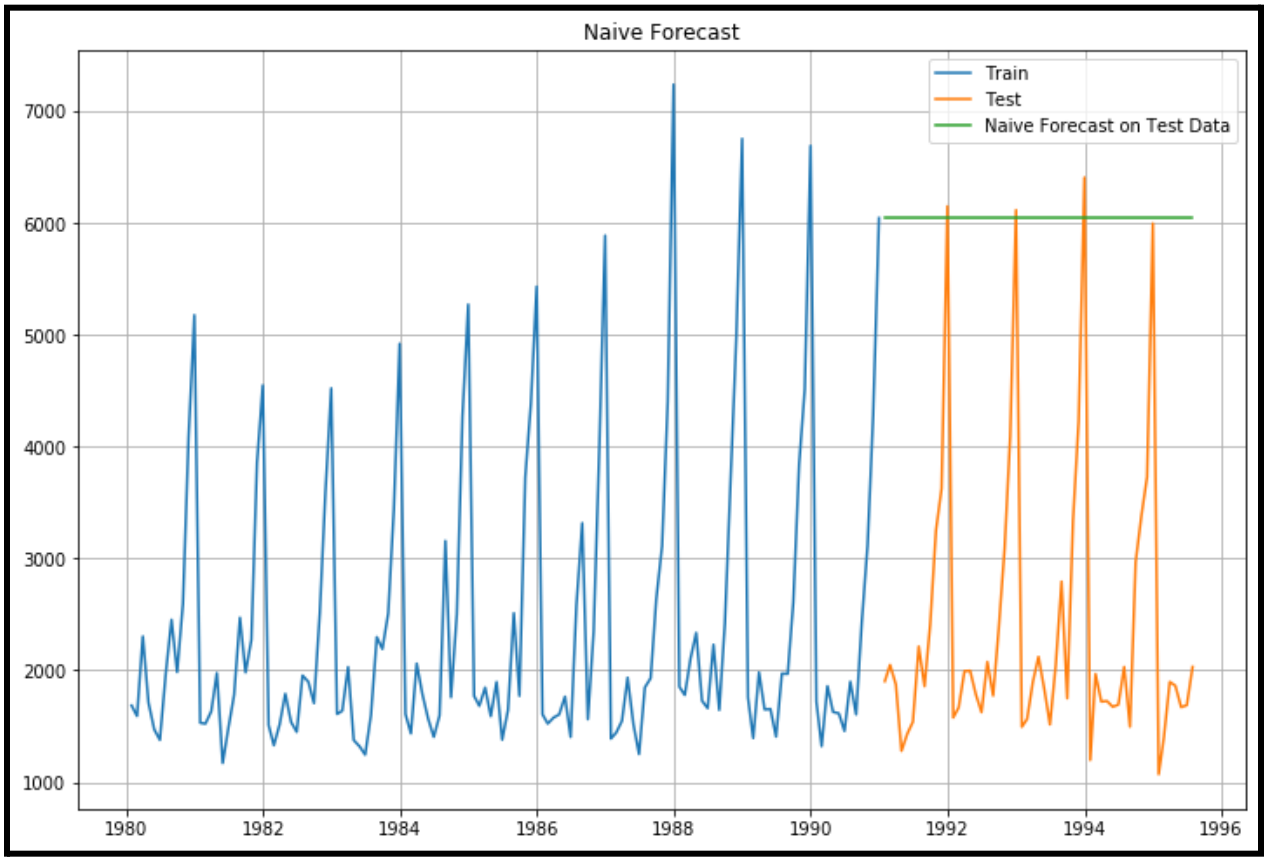
SPARKLING WINE:

1. Linear Regression-



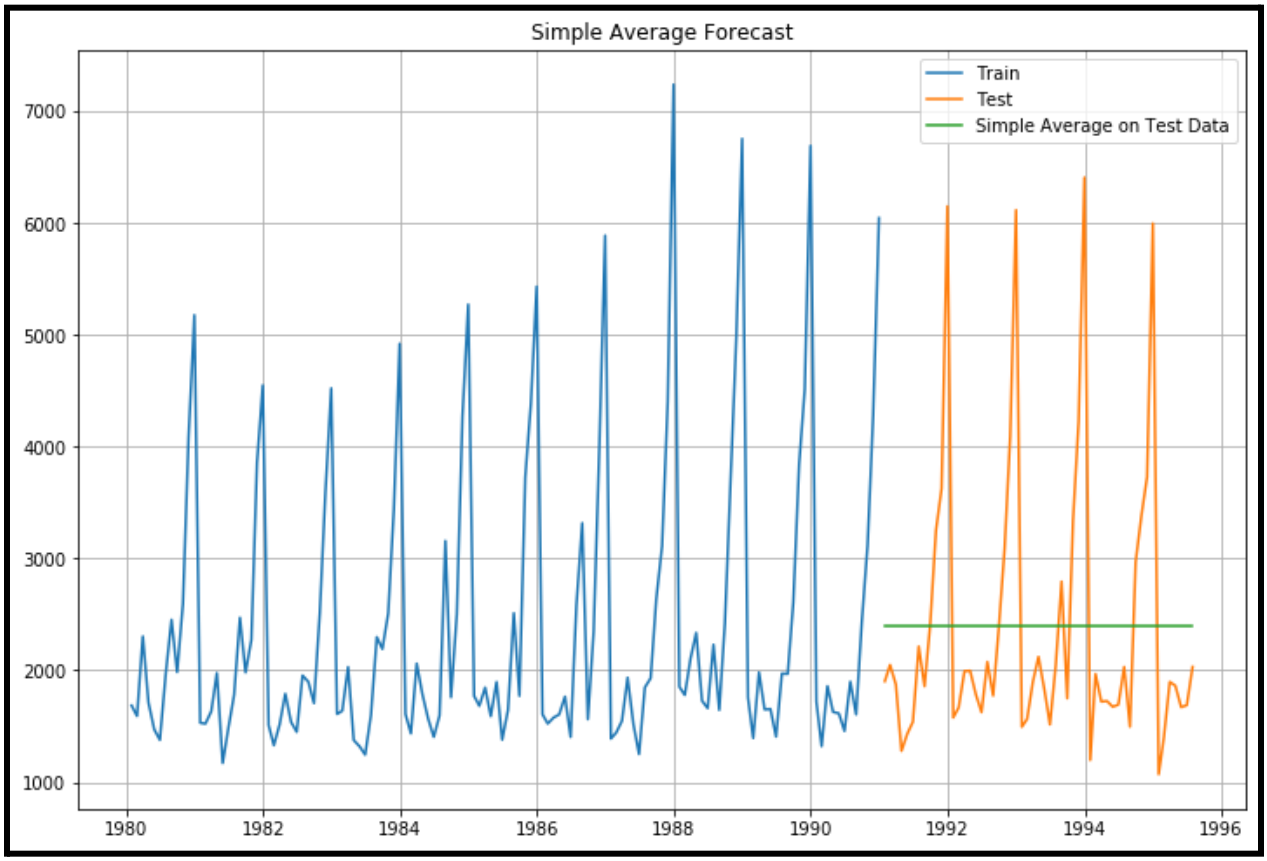
GRAPH 14

2. Naïve-



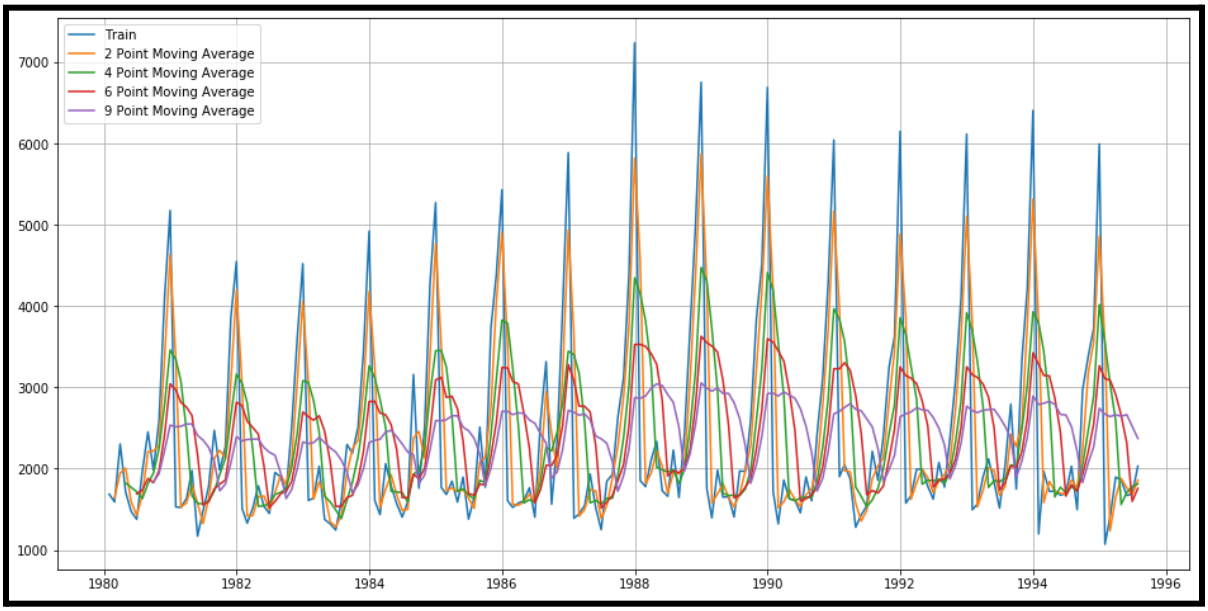
GRAPH 15

3. Simple Average Forecast-



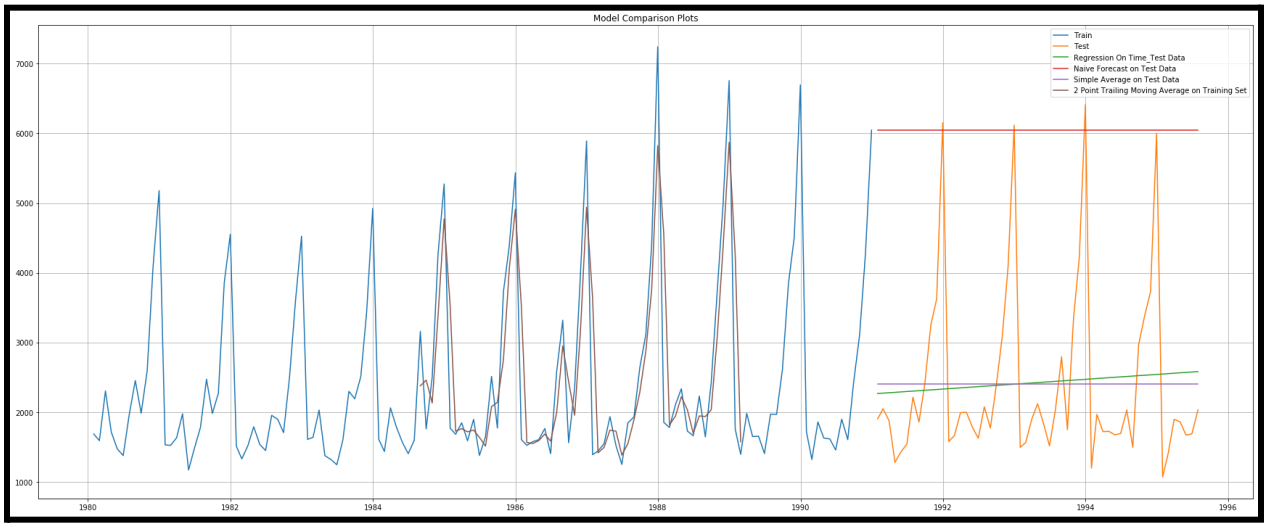
GRAPH 16

4. Moving Average Forecast-



GRAPH 17

All models comparison-



GRAPH 18

RMSE-

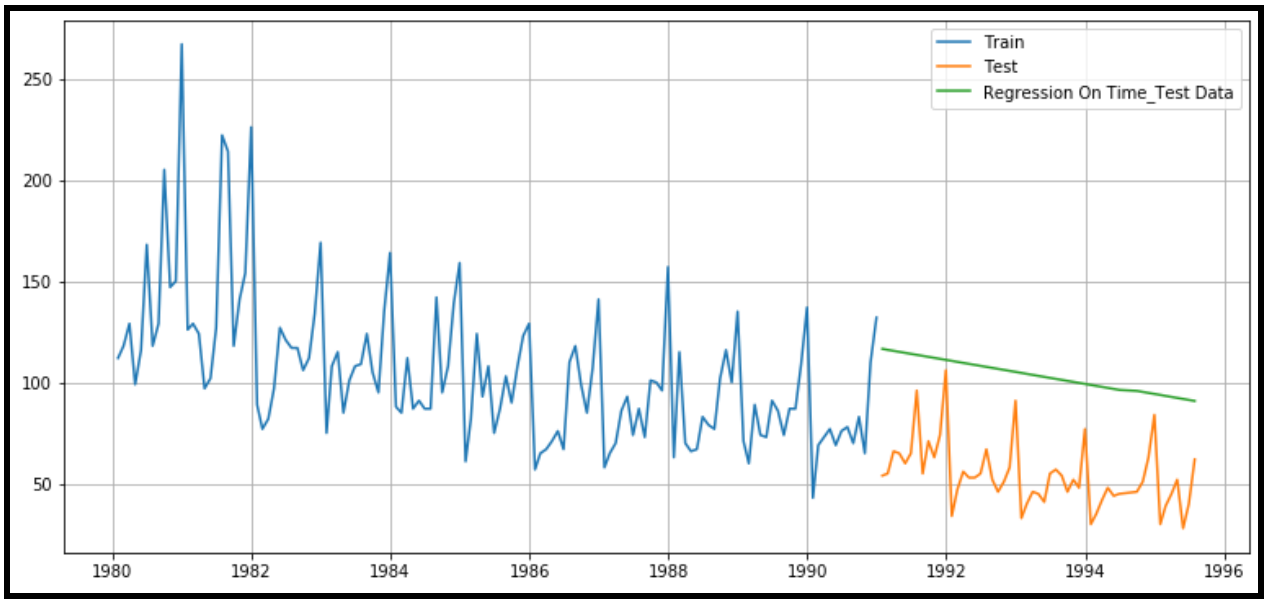
	Test RMSE
RegressionOnTime	1275.867052
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	2070.918532
4pointTrailingMovingAverage	1837.745348
6pointTrailingMovingAverage	1654.989198
9pointTrailingMovingAverage	1286.231497

FIGURE 15

Based on the RMSE value the Simple Average Model has the least value.

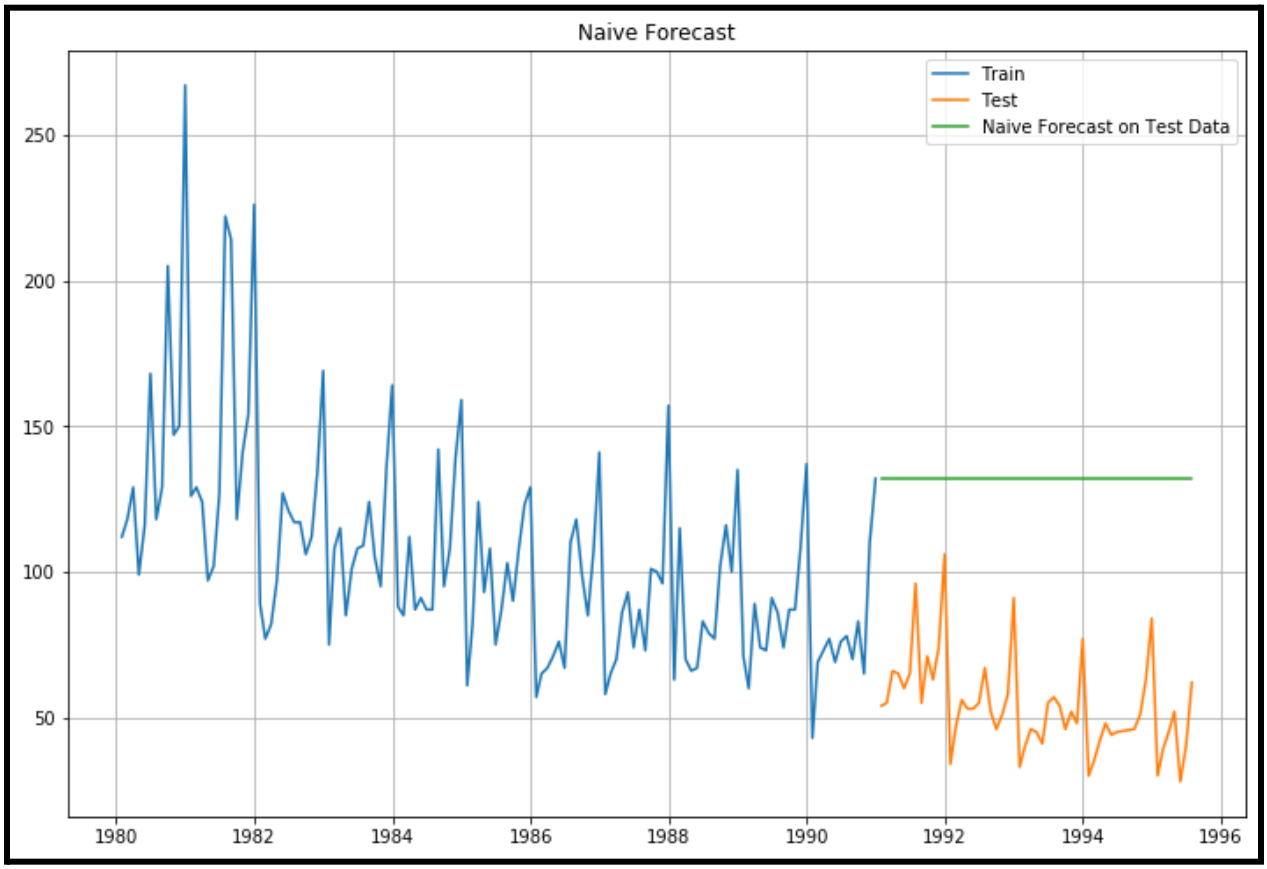
ROSE WINE:

1. Linear Regression-



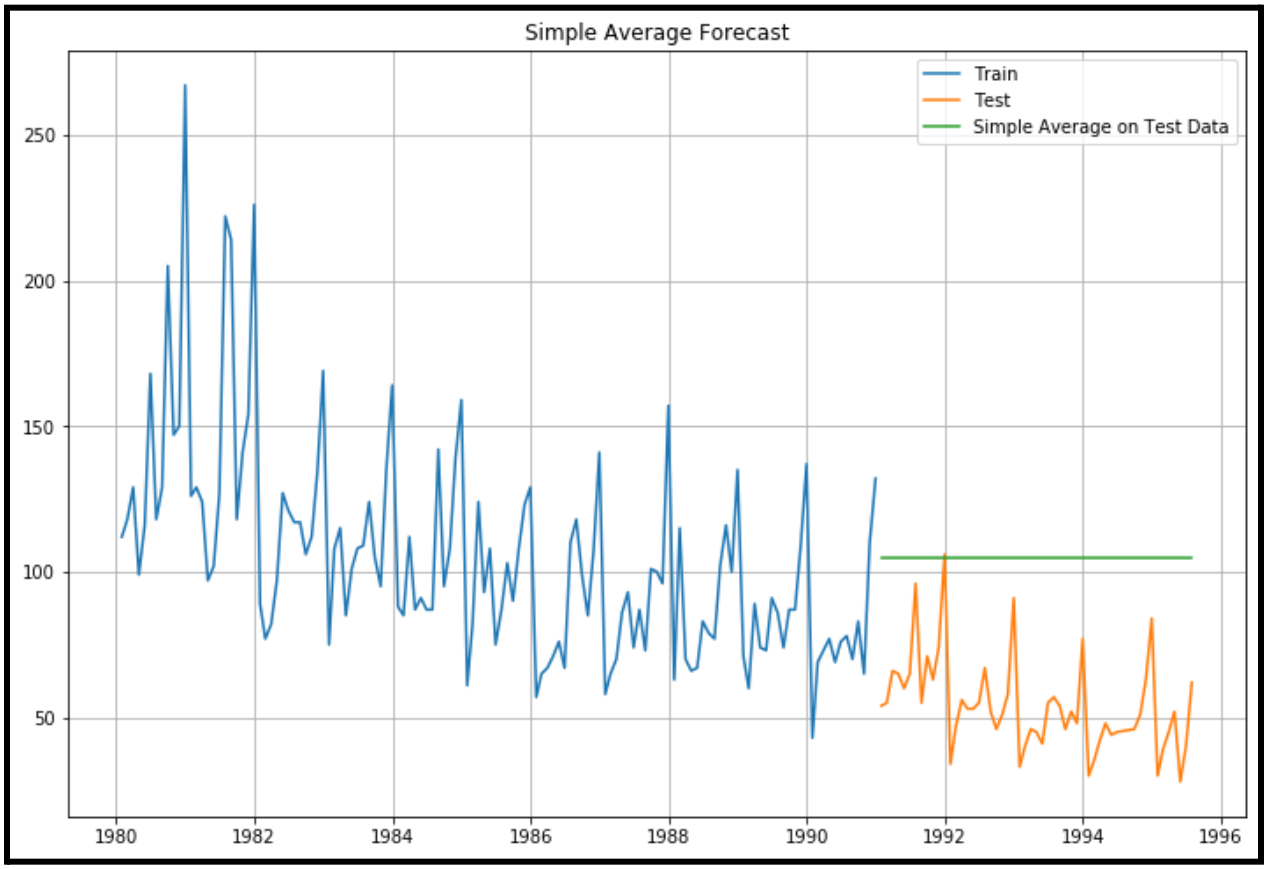
GRAPH 19

2. Naïve-



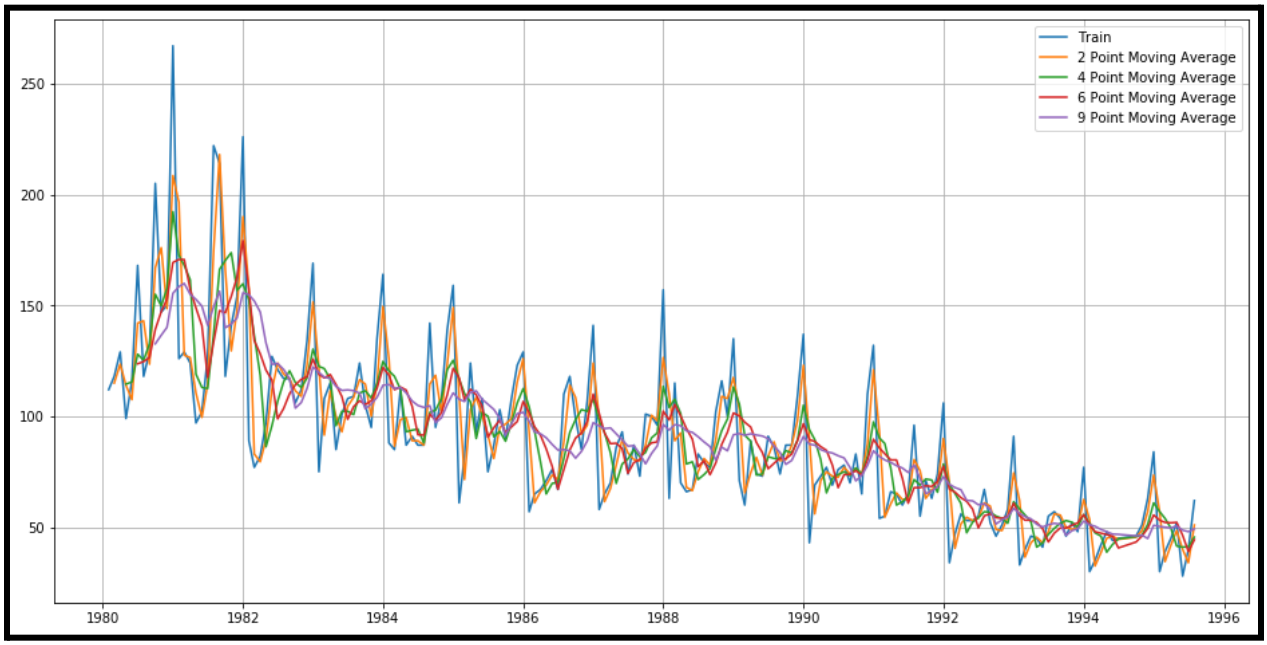
GRAPH 20

3. Simple Average forecast-



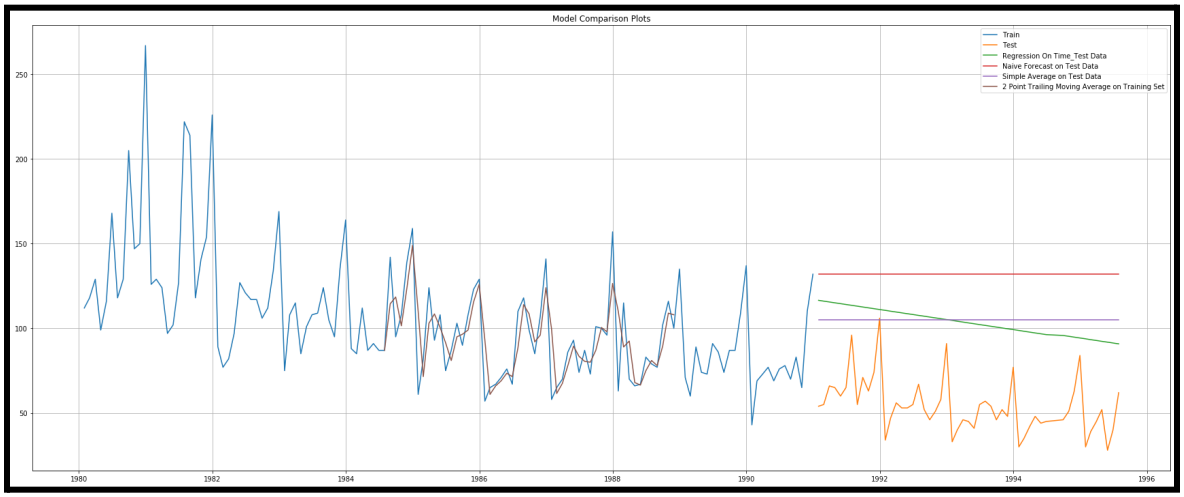
GRAPH 21

4. Moving Average forecast-



GRAPH 22

All models comparison-



GRAPH 23

RMSE-

	Test RMSE
RegressionOnTime	51.658895
NaiveModel	79.451540
SimpleAverageModel	53.221795
2pointTrailingMovingAverage	45.900271
4pointTrailingMovingAverage	43.335163
6pointTrailingMovingAverage	42.639228
9pointTrailingMovingAverage	42.343055

FIGURE 16

Based on the RMSE value the 9 point Trailing Moving Average Model has the least value.

5. 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.

The Augmented Dickey-Fuller test is an unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.

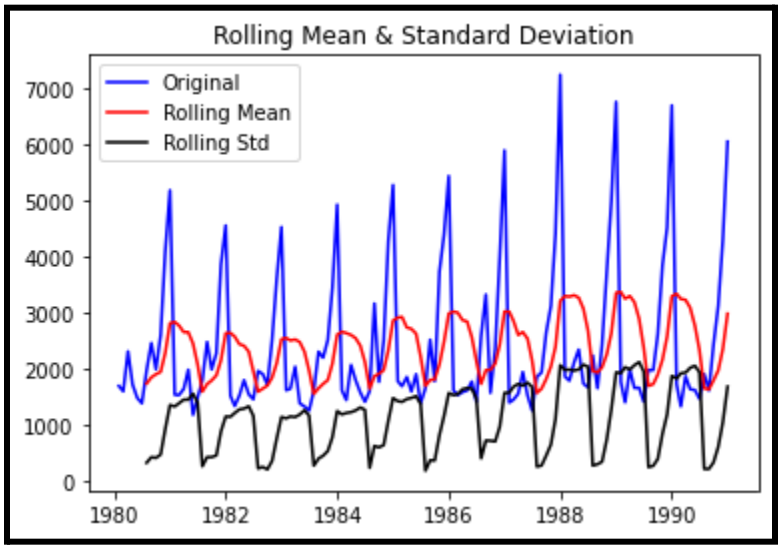
The hypothesis in a simple form for the ADF test is:

$H0$: The Time Series has a unit root and is thus non-stationary.

$H1$: The Time Series does not have a unit root and is thus stationary.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

SPARKLING WINE:

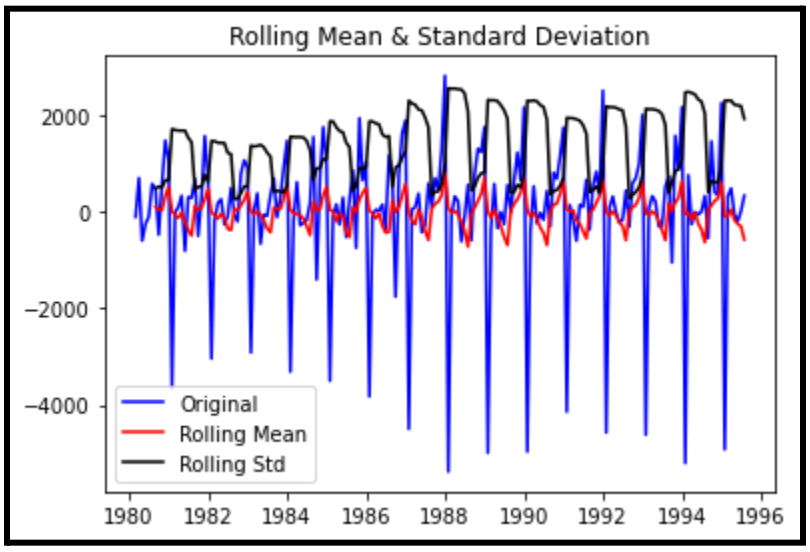


GRAPH 24

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.208926
p-value                          0.669744
#Lags Used                      12.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
dtype: float64
```

FIGURE 17

We see that at a 5% significant level the Time Series is non-stationary.



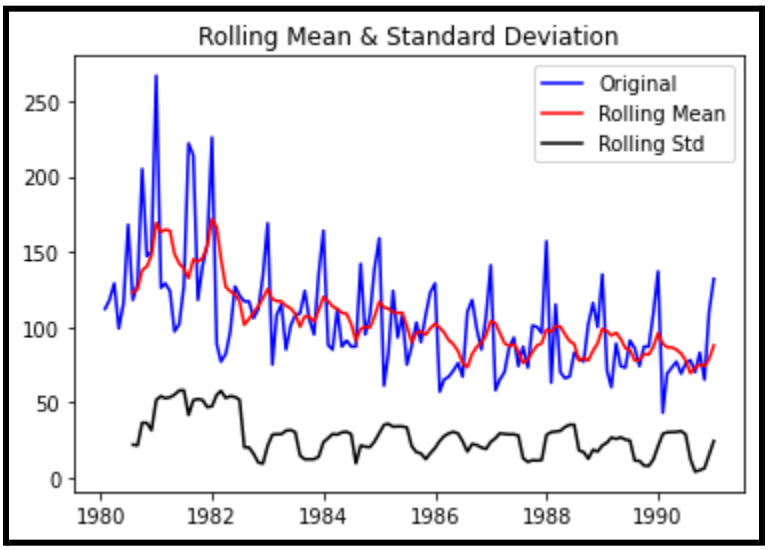
GRAPH 25

```
Results of Dickey-Fuller Test:
Test Statistic                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical Value (10%)             -2.575653
dtype: float64
```

FIGURE 18

After taking a difference of order 1 the Time Series is stationary.

ROSE WINE:

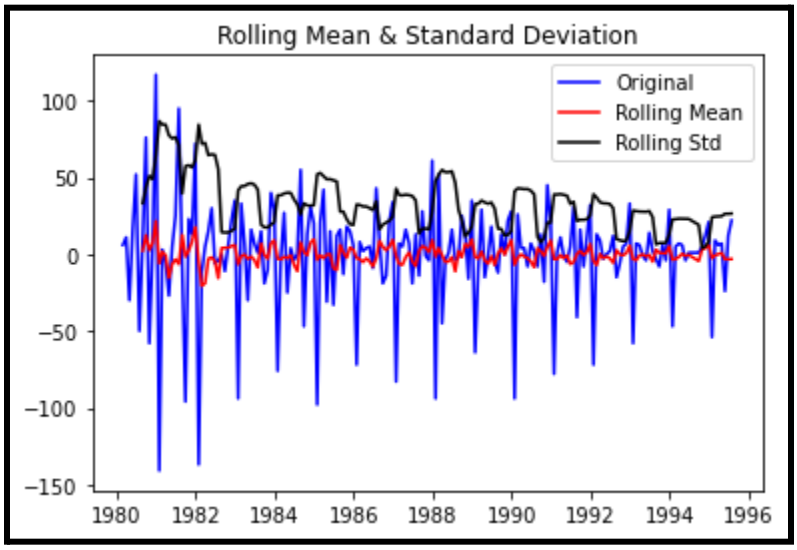


GRAPH 26

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.164250
p-value                          0.219476
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
dtype: float64
```

FIGURE 19

We see that at 5% significant level the Time Series is non-stationary.



GRAPH 27

```
Results of Dickey-Fuller Test:
Test Statistic                -8.167161e+00
p-value                        8.819858e-13
#Lags Used                     1.200000e+01
Number of Observations Used    1.710000e+02
Critical Value (1%)           -3.469181e+00
Critical Value (5%)           -2.878595e+00
Critical Value (10%)          -2.575863e+00
dtype: float64
```

FIGURE 20

After taking a difference of order 1 the Time Series is stationary.

6. 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.

ARIMA:

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts. ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration. [11]

SARIMA:

A seasonal autoregressive integrated moving average (SARIMA) model is one step different from an ARIMA model based on the concept of seasonal trends. In many time series data, frequent seasonal effects come into play. Take for example the average temperature measured in a location with four seasons. There will be a seasonal effect

on a yearly basis, and the temperature in this particular season will definitely have a strong correlation with the temperature measured last year in the same season. [12]

AIC:

Akaike's information criterion (AIC) compares the quality of a set of statistical models to each other. For example, you might be interested in what variables contribute to low socioeconomic status and how the variables contribute to that status. Let's say you create several regression models for various factors like education, family size, or disability status; The AIC will take each model and rank them from best to worst. The "best" model will be the one that neither under-fits nor over-fits. [13]

SPARKLING WINE:

ARIMA model-

```
ARIMA(0, 0, 0) - AIC:2271.203212328525
ARIMA(0, 0, 1) - AIC:2245.2688513920953
ARIMA(0, 0, 2) - AIC:2245.3432152954183
ARIMA(1, 0, 0) - AIC:2247.348274393555
ARIMA(1, 0, 1) - AIC:2245.9490918698066
ARIMA(1, 0, 2) - AIC:2246.012193247299
ARIMA(2, 0, 0) - AIC:2244.7999152004504
ARIMA(2, 0, 1) - AIC:2236.590818706324
ARIMA(2, 0, 2) - AIC:2200.9043938950617
```

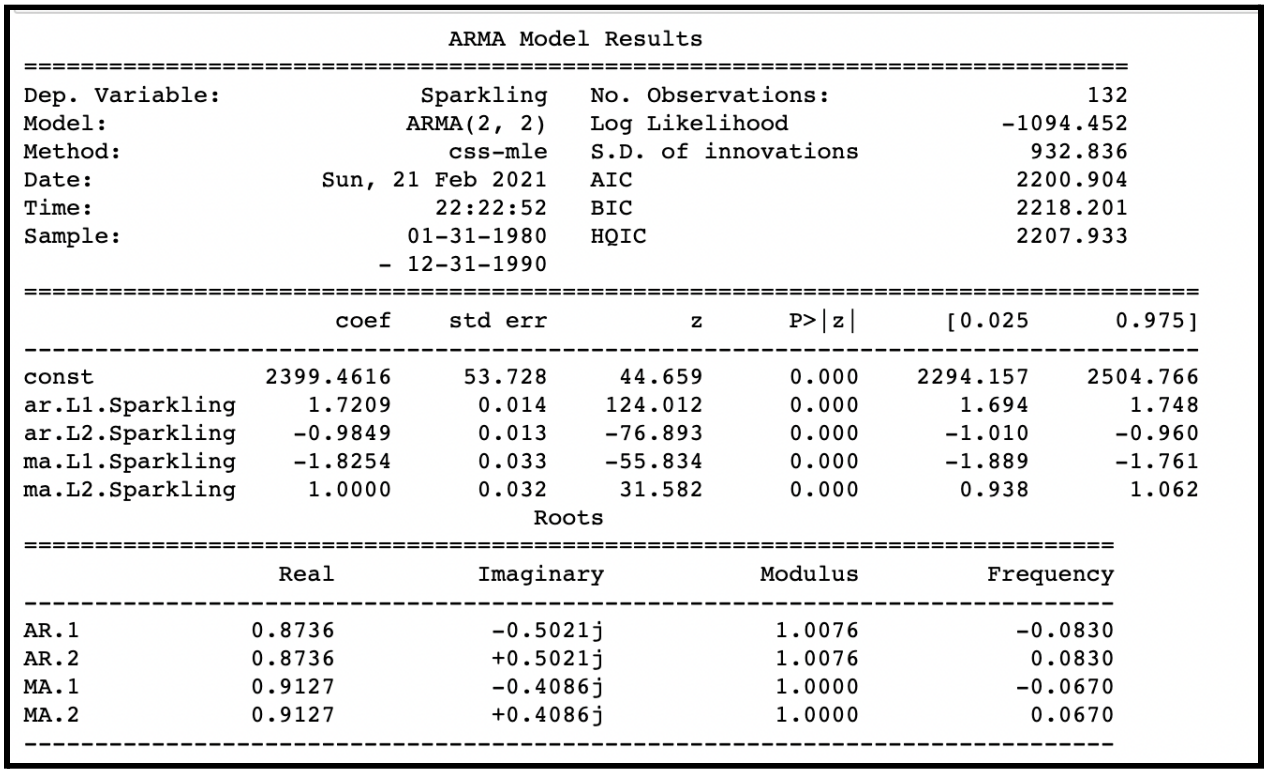
FIGURE 21

Sorted AIC values-

	param	AIC
8	(2, 0, 2)	2200.904394
7	(2, 0, 1)	2236.590819
6	(2, 0, 0)	2244.799915
1	(0, 0, 1)	2245.268851
2	(0, 0, 2)	2245.343215
4	(1, 0, 1)	2245.949092
5	(1, 0, 2)	2246.012193
3	(1, 0, 0)	2247.348274
0	(0, 0, 0)	2271.203212

FIGURE 22

Summary-



ARMA Model Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARMA(2, 2)	Log Likelihood	-1094.452
Method:	css-mle	S.D. of innovations	932.836
Date:	Sun, 21 Feb 2021	AIC	2200.904
Time:	22:22:52	BIC	2218.201
Sample:	01-31-1980 - 12-31-1990	HQIC	2207.933

	coef	std err	z	P>|z|	[0.025	0.975]
const	2399.4616	53.728	44.659	0.000	2294.157	2504.766
ar.L1.Sparkling	1.7209	0.014	124.012	0.000	1.694	1.748
ar.L2.Sparkling	-0.9849	0.013	-76.893	0.000	-1.010	-0.960
ma.L1.Sparkling	-1.8254	0.033	-55.834	0.000	-1.889	-1.761
ma.L2.Sparkling	1.0000	0.032	31.582	0.000	0.938	1.062

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.8736	-0.5021j	1.0076	-0.0830
AR.2	0.8736	+0.5021j	1.0076	0.0830
MA.1	0.9127	-0.4086j	1.0000	-0.0670
MA.2	0.9127	+0.4086j	1.0000	0.0670

FIGURE 23

SARIMA model-

	param	seasonal	AIC
80	(2, 0, 2)	(2, 0, 2, 7)	1950.790380
79	(2, 0, 2)	(2, 0, 1, 7)	1970.180782
47	(1, 0, 2)	(0, 0, 2, 7)	1970.814565
50	(1, 0, 2)	(1, 0, 2, 7)	1972.204363
74	(2, 0, 2)	(0, 0, 2, 7)	1972.694491

FIGURE 24

Summary-

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(2, 0, 2)x(2, 0, 2, 7)	Log Likelihood	-966.374
Date:	Sun, 07 Nov 2021	AIC	1950.747
Time:	18:28:13	BIC	1975.452
Sample:	01-31-1980 - 12-31-1990	HQIC	1960.775
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.7325	0.037	46.698	0.000	1.660	1.805
ar.L2	-1.0023	0.035	-28.751	0.000	-1.071	-0.934
ma.L1	-1.7926	0.158	-11.366	0.000	-2.102	-1.483
ma.L2	0.9545	0.150	6.360	0.000	0.660	1.249
ar.S.L7	0.7581	0.759	0.999	0.318	-0.730	2.246
ar.S.L14	0.2570	0.766	0.335	0.737	-1.245	1.759
ma.S.L7	-0.4284	0.731	-0.586	0.558	-1.861	1.004
ma.S.L14	-0.4884	0.635	-0.769	0.442	-1.734	0.757
sigma2	1.769e+06	1.25e-06	1.42e+12	0.000	1.77e+06	1.77e+06

Ljung-Box (L1) (Q):	0.33	Jarque-Bera (JB):	12.68
Prob(Q):	0.57	Prob(JB):	0.00
Heteroskedasticity (H):	2.44	Skew:	0.59
Prob(H) (two-sided):	0.01	Kurtosis:	4.12

FIGURE 25

RMSE-

	RMSE
ARIMA(2,0,2)	1004.957168
SARIMA(2,0,2)(2,0,2,7)	1004.957168

FIGURE 26

ROSE WINE:

ARIMA model-

```
ARIMA(0, 0, 0) - AIC:1324.8997029577333
ARIMA(0, 0, 1) - AIC:1305.468405768463
ARIMA(0, 0, 2) - AIC:1306.5866794770147
ARIMA(1, 0, 0) - AIC:1301.5463044353112
ARIMA(1, 0, 1) - AIC:1294.5105851813653
ARIMA(1, 0, 2) - AIC:1292.0532102477234
ARIMA(2, 0, 0) - AIC:1302.3460741768874
ARIMA(2, 0, 1) - AIC:1292.9371945613584
ARIMA(2, 0, 2) - AIC:1292.2480553299172
```

FIGURE 27

Sorted AIC values-

	param	AIC
5	(1, 0, 2)	1292.053210
8	(2, 0, 2)	1292.248055
7	(2, 0, 1)	1292.937195
4	(1, 0, 1)	1294.510585
3	(1, 0, 0)	1301.546304
6	(2, 0, 0)	1302.346074
1	(0, 0, 1)	1305.468406
2	(0, 0, 2)	1306.586679
0	(0, 0, 0)	1324.899703

FIGURE 28

Summary-

ARMA Model Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARMA(1, 2)	Log Likelihood	-641.027
Method:	css-mle	S.D. of innovations	30.999
Date:	Sun, 07 Nov 2021	AIC	1292.053
Time:	18:23:55	BIC	1306.467
Sample:	01-31-1980 - 12-31-1990	HQIC	1297.910

	coef	std err	z	P>|z|	[0.025	0.975]
const	107.8499	15.808	6.823	0.000	76.868	138.832
ar.L1.Rose	0.9861	0.018	53.661	0.000	0.950	1.022
ma.L1.Rose	-0.6873	0.098	-6.990	0.000	-0.880	-0.495
ma.L2.Rose	-0.2007	0.094	-2.129	0.033	-0.386	-0.016

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0141	+0.0000j	1.0141	0.0000
MA.1	1.1009	+0.0000j	1.1009	0.0000
MA.2	-4.5248	+0.0000j	4.5248	0.5000

FIGURE 29

SARIMA model-

	param	seasonal	AIC
47	(1, 0, 2)	(0, 0, 2, 7)	1105.275607
74	(2, 0, 2)	(0, 0, 2, 7)	1106.350698
50	(1, 0, 2)	(1, 0, 2, 7)	1111.303067
77	(2, 0, 2)	(1, 0, 2, 7)	1111.336875
53	(1, 0, 2)	(2, 0, 2, 7)	1113.130434

FIGURE 30

Summary-

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(1, 0, 2)x(0, 0, 2, 7)	Log Likelihood	-546.638
Date:	Sun, 07 Nov 2021	AIC	1105.276
Time:	18:38:00	BIC	1121.745
Sample:	01-31-1980	HQIC	1111.961
	- 12-31-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9953	0.001	998.588	0.000	0.993	0.997
ma.L1	-0.8199	0.127	-6.446	0.000	-1.069	-0.571
ma.L2	-0.1801	0.103	-1.752	0.080	-0.381	0.021
ma.S.L7	-0.1582	0.130	-1.219	0.223	-0.413	0.096
ma.S.L14	-0.0234	0.139	-0.168	0.867	-0.296	0.249
sigma2	753.7403	0.000	4.52e+06	0.000	753.740	753.741

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	4.95
Prob(Q):	0.98	Prob(JB):	0.08
Heteroskedasticity (H):	0.57	Skew:	0.35
Prob(H) (two-sided):	0.09	Kurtosis:	3.73

FIGURE 31

RMSE-

	RMSE
ARIMA(1,0,2)	45.088489
SARIMA(1, 0, 2)(0, 0, 2, 7)	45.088489

FIGURE 32

7. 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.

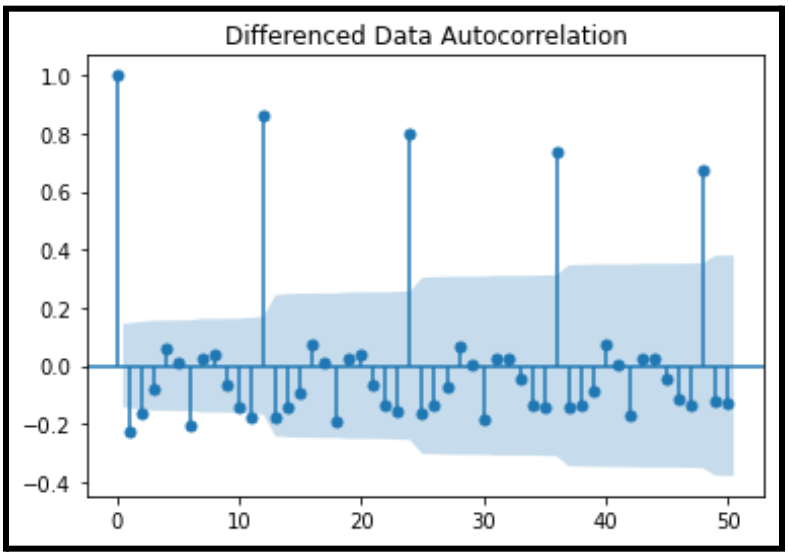
ACF and PACF-

ACF is an (complete) auto-correlation function which gives us values of auto-correlation of any series with its lagged values. We plot these values along with the confidence band and tada! We have an ACF plot. In simple terms, it describes how well the present value of the series is related with its past values. A time series can have components like trend, seasonality, cyclic and residual. ACF considers all these components while finding correlations hence it's a 'complete auto-correlation plot'.

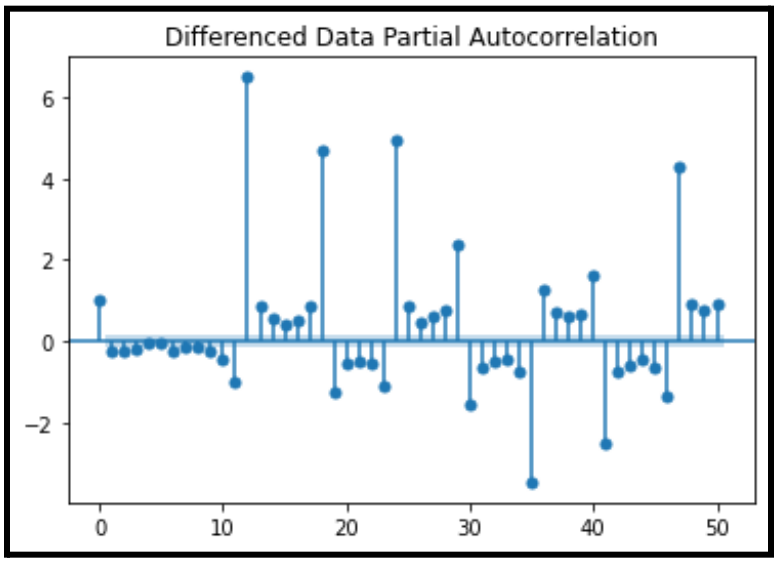
PACF is a partial auto-correlation function. Basically instead of finding correlations of present with lags like ACF, it finds correlation of the residuals (which remains after removing the effects which are already explained by the earlier lag(s)) with the next lag value hence 'partial' and not 'complete' as we remove already found variations before we find the next correlation. So if there is any hidden information in the residual which can be modeled by the next lag, we might get a good correlation and we will keep that next lag as a feature while modeling. Remember while modeling we don't want to keep too many features which are correlated as that can create multicollinearity issues. Hence we need to retain only the relevant features. [14]

SPARKLING WINE:

ACF and PACF-



GRAPH 28



GRAPH 29

Summary-

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(0, 1, 0)	Log Likelihood	-1132.791
Method:	css	S.D. of innovations	1377.911
Date:	Sun, 07 Nov 2021	AIC	2269.583
Time:	19:14:04	BIC	2275.333
Sample:	02-29-1980 - 12-31-1990	HQIC	2271.919

	coef	std err	z	P>|z|	[0.025	0.975]
const	33.2901	120.389	0.277	0.782	-202.667	269.248

FIGURE 33

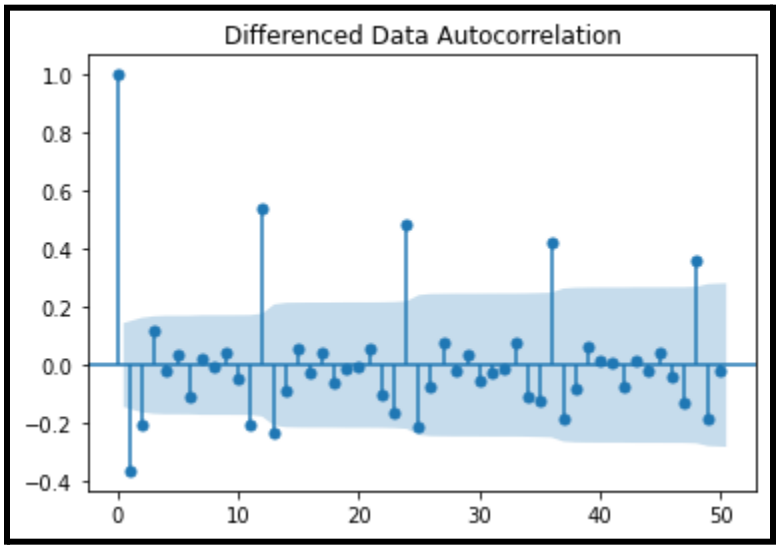
RMSE-

	RMSE
ARIMA(2,0,2)	1174.798984
ARIMA(0,1,0)	4779.154299

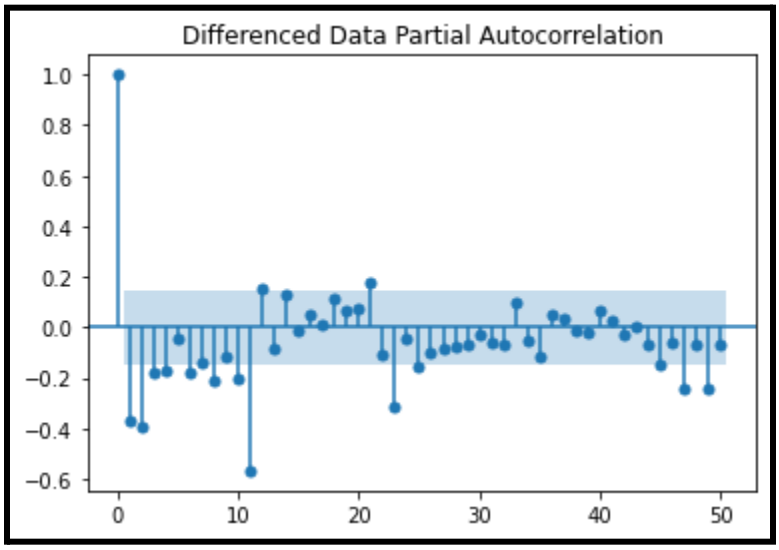
FIGURE 34

ROSE WINE:

ACF and PACF-



GRAPH 30



GRAPH 31

Summary-

ARMA Model Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARMA(0, 0)	Log Likelihood	-660.450
Method:	css	S.D. of innovations	36.034
Date:	Sun, 07 Nov 2021	AIC	1324.900
Time:	19:30:23	BIC	1330.665
Sample:	01-31-1980 - 12-31-1990	HQIC	1327.243

	coef	std err	z	P>|z|	[0.025	0.975]
const	104.9394	3.136	33.459	0.000	98.792	111.087

FIGURE 35

RMSE-

	RMSE
ARIMA(1,0,2)	45.088489
ARIMA(0,0,0)	53.221795

FIGURE 36

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

SPARKLING WINE:

	RMSE
ARIMA(2,0,2)	1.004857e+03
SARIMA(0,0,1)(2,0,2)7	1.004857e+03
SARIMAX(1,0,2)(2,0,2)7	3.186183e-13
SARIMAX_tvlm(1,0,2)(2,0,2)7	0.000000e+00

FIGURE 37

ROSE WINE:

	RMSE
ARIMA(1,0,2)	45.088489
ARIMA(0,0,0)	53.221795
SARIMA(1, 0, 2)(0, 0, 2, 7)	53.221795
SARIMAX_tvlm(1,0,2)(2,0,2)7	0.000000

FIGURE 38

9. 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.

SPARKLING WINE:

Summary-

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(1, 0, 2)x(2, 0, 2, 7)	Log Likelihood	-1449.392
Date:	Sun, 07 Nov 2021	AIC	2914.783
Time:	19:02:17	BIC	2939.870
Sample:	01-31-1980	HQIC	2924.963
	- 07-31-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.0006	0.001	869.609	0.000	0.998	1.003
ma.L1	-0.7296	0.158	-4.613	0.000	-1.040	-0.420
ma.L2	-0.2535	0.178	-1.422	0.155	-0.603	0.096
ar.S.L7	-0.1079	0.694	-0.155	0.876	-1.468	1.252
ar.S.L14	0.3977	0.428	0.928	0.353	-0.442	1.237
ma.S.L7	-0.1923	0.677	-0.284	0.776	-1.518	1.134
ma.S.L14	-0.6273	0.610	-1.029	0.304	-1.823	0.568
sigma2	1.869e+06	1.98e-07	9.46e+12	0.000	1.87e+06	1.87e+06

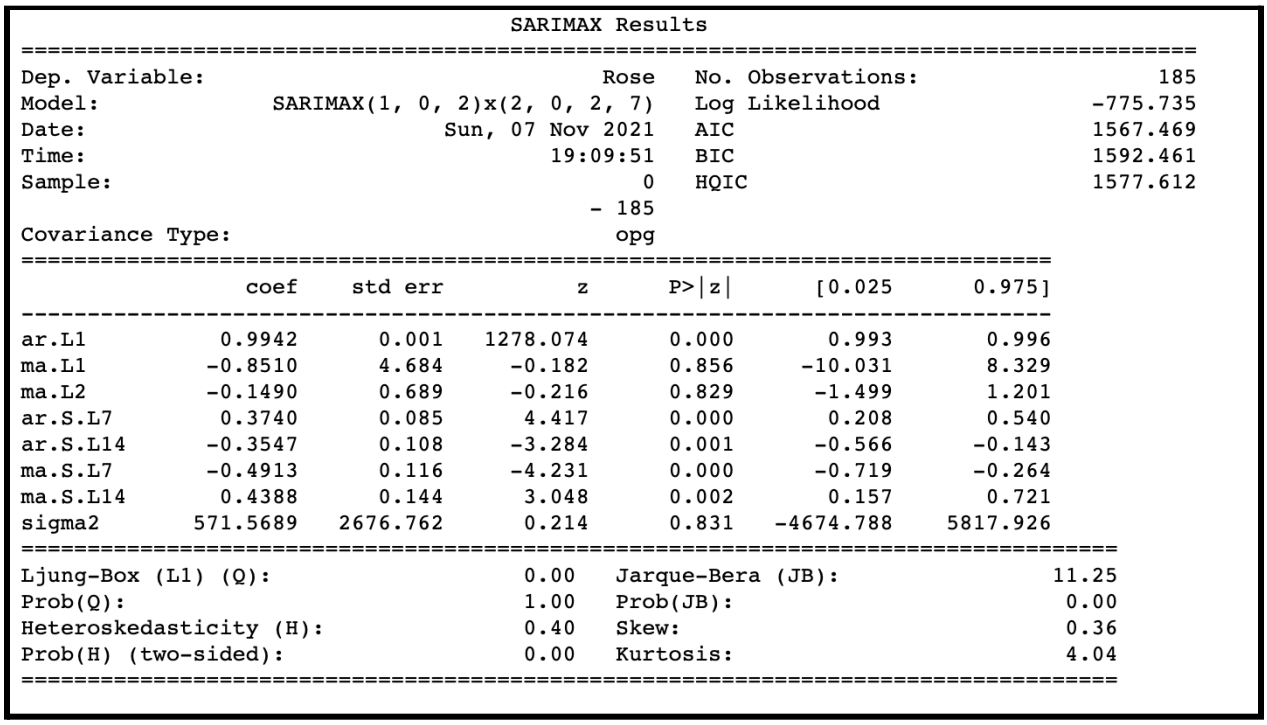
Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	30.40
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	1.77	Skew:	0.94
Prob(H) (two-sided):	0.03	Kurtosis:	3.88

FIGURE 39

RMSE of the Full Model 1174.7989843965743.

ROSE WINE:

Summary-



```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                Rose   No. Observations:                  185
Model:             SARIMAX(1, 0, 2)x(2, 0, 2, 7)   Log Likelihood                -775.735
Date:                           Sun, 07 Nov 2021   AIC                            1567.469
Time:                                   19:09:51   BIC                            1592.461
Sample:                                        0   HQIC                           1577.612
                                             - 185
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9942      0.001   1278.074      0.000       0.993       0.996
ma.L1         -0.8510      4.684     -0.182      0.856     -10.031       8.329
ma.L2         -0.1490      0.689     -0.216      0.829      -1.499       1.201
ar.S.L7        0.3740      0.085      4.417      0.000       0.208       0.540
ar.S.L14      -0.3547      0.108     -3.284      0.001      -0.566      -0.143
ma.S.L7       -0.4913      0.116     -4.231      0.000      -0.719      -0.264
ma.S.L14       0.4388      0.144      3.048      0.002       0.157       0.721
sigma2       571.5689   2676.762      0.214      0.831   -4674.788    5817.926
==========================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                11.25
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.36
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.04
==========================================================================================
```

FIGURE 40

RMSE of the Full Model 31.488627395270406.

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

- We can see that the Rose wine has a declining trend whereas Sparkling wine has an increasing/constant seasonality trend. It can be assumed that the Rose wine has a festival-related trend.
- The sales of both wines have been the highest during the months of October to December. It is important to match the demand and have much stock to cater to the expectations / demand.
- Rose wine could be reduced to effectively utilize the production and other materials cost.
- We are able to infer that triple exponential smoothing gives the best results for both wines. The sales of Sparkling wine has been highest during 3 months and it is important to effectively utilize the materials to cater to the demand and use the resources judiciously.
- Rose wine doesn't have many takers – we might infer that it caters to the need of only a special sect of people on certain occasions. Since the trend has been on the downward curve for the past several years, the sales of Rose wine need to be monitored constantly and match the supply only with the needs since overproduction can result in a hefty loss.

References

Websites-

[1] https://www.statisticshowto.com/univariate/

[2] https://www.spss-tutorials.com/skewness/

[3] https://en.wikipedia.org/wiki/Bivariate_analysis

[4] https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd5

[5] https://towardsdatascience.com/different-types-of-time-series-decomposition-396c09f92693

[6] https://en.wikipedia.org/wiki/Linear_regression

[7] https://www.avercast.in/blog/what-is-naive-forecasting-and-how-can-be-used-to-calculate-future-demand

[8] https://www.investopedia.com/terms/s/sma.asp

[9] https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/moving-average/

[10] https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

[11] https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[12] https://medium.com/@kfoofw/seasonal-lags-sarima-model-fa671a858729

[13] https://www.statisticshowto.com/akaikes-information-criterion/

[14] https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8

End of Project