Abubacker Siddhique I
Jun 2, 2024

Time Series Forecasting

Table of Contents

List Of Tables:

List Of Figures:

Define the problem and perform Exploratory Data Analysis:

Problem Statement:

We are provided with two datasets, with historical data encompassing the sales of different types of wines throughout the 20th century.

As an analyst at ABC Estate Wines, our objective is to delve into the data, analyze trends, patterns, and factors influencing wine sales over the course of the century which will help in enhancing sales performance, capitalize on emerging market opportunities, and maintaining a competitive edge in the wine industry.

Data Description:

The dataset consists of 187 entries and 2 columns. These columns represent the monthly sales for respective wines.

Read the data as an appropriate time series data:

Rose wine :

First few rows

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

Datatype

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    185 non-null    float64
dtypes: float64(1)
memory usage: 2.9 KB
```

Sparkling wine :

First few rows

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

Datatype

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

Figure 1

Plot the data :

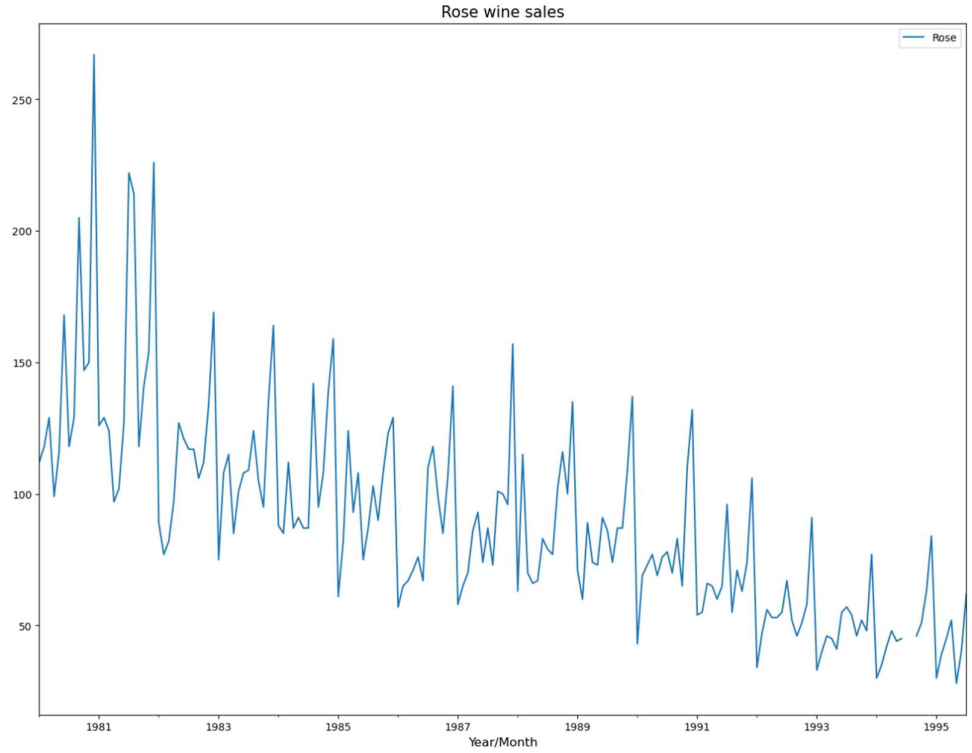

Figure 2

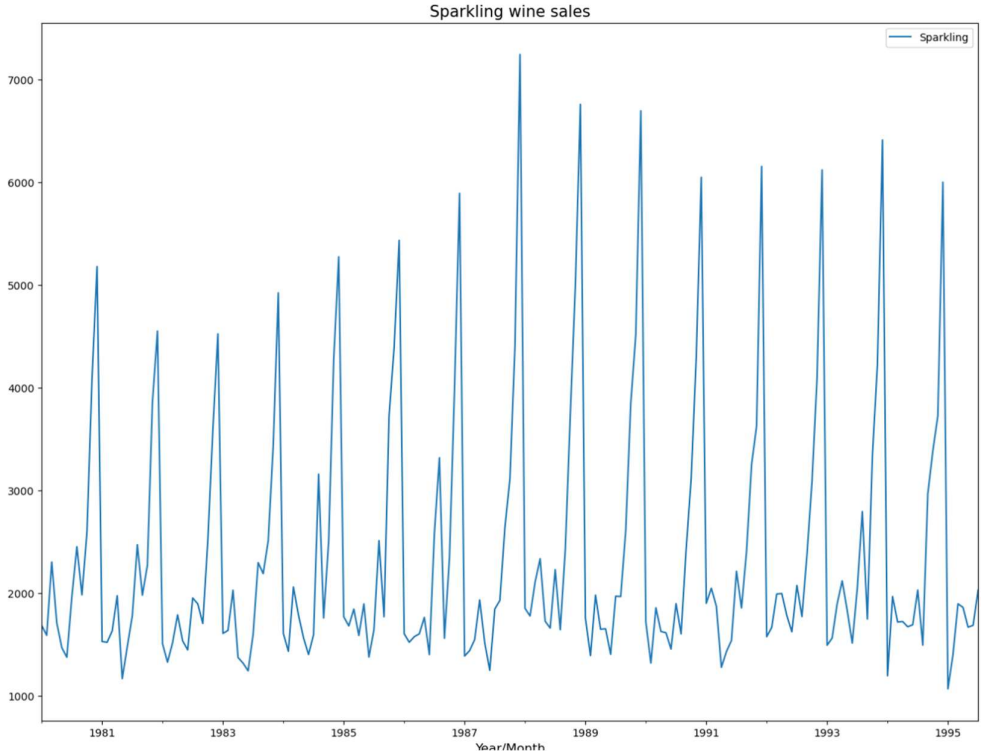

Figure 3

Exploratory Data Analysis:

Rose Wine:

	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 4

- We can see that there are 2 missing values for rose wine.The mean sales is 90.3 with a standard deviation of 39.17.
- The sales is spread over a wide range with minimum and maximum sales of 28 and 267 respectively.

Sales by Month:

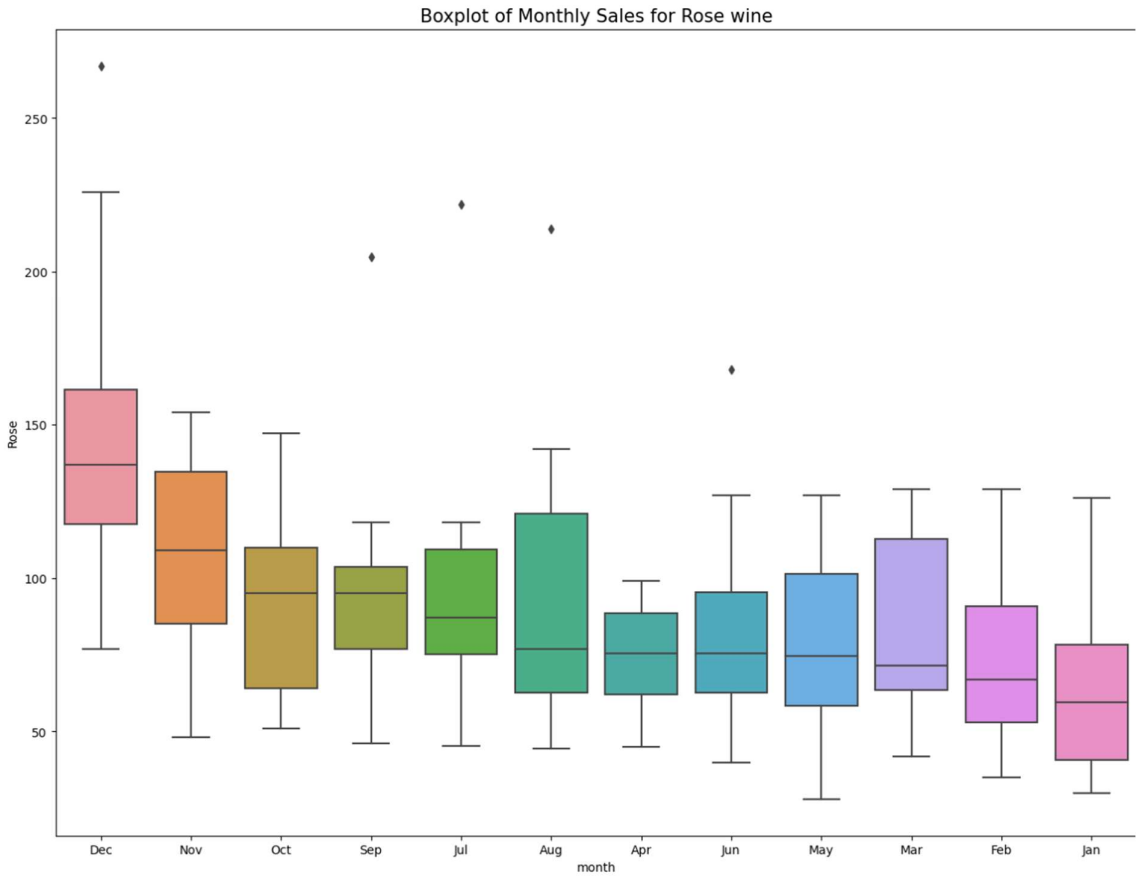

Figure 5

- From the above plot it is evident that the wine sales starts increasing towards the year end.
- It gradually starts increasing from July.
- December has the higest sales with a median of around 140.
- This may be due to the fact that December is a festive month and it is winter in most of the places.
- There are outliers present for Dec,Sep,Jun,July and Aug.

Sales by year:

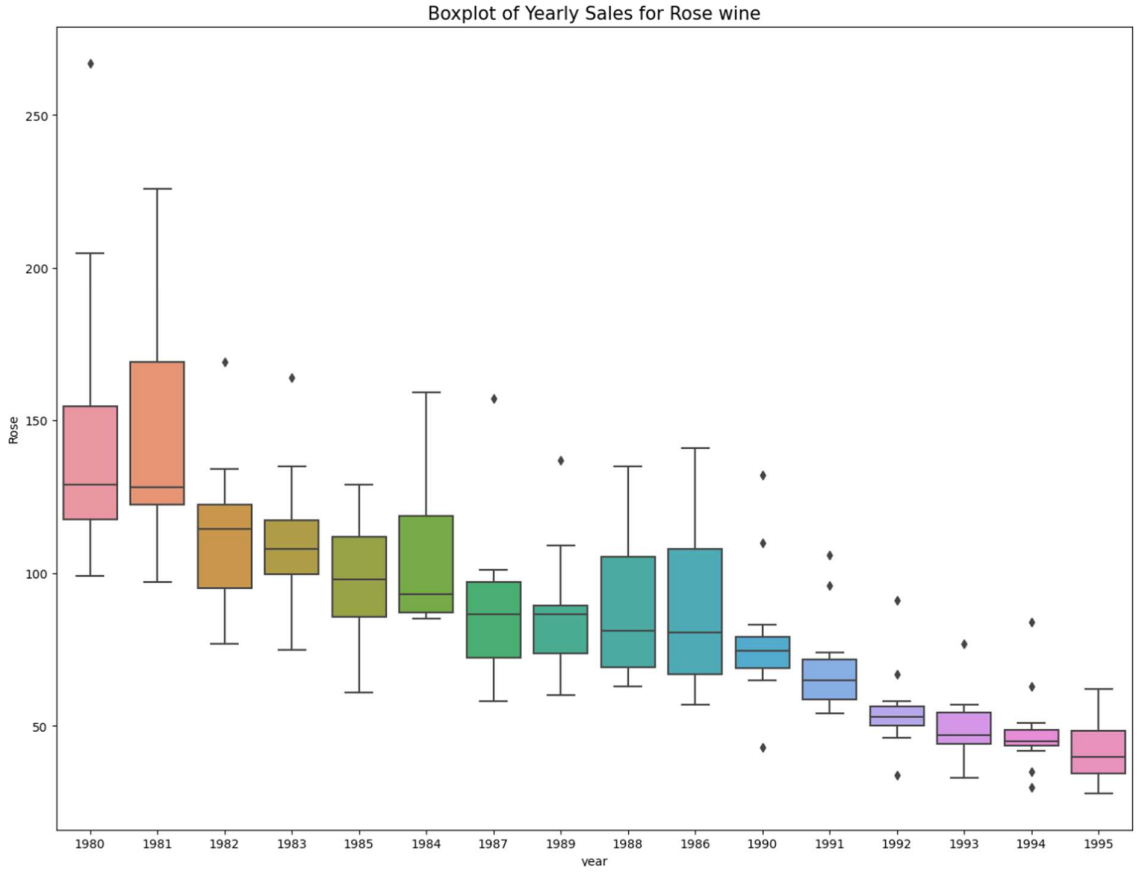

Figure 6

- From the above plot we can see a decreasing trend in wine sales.
- The year 1980 has the highest median in terms of sales while 1981 stands out for its wide spread, with a maximum sale reaching around 230.
- The year 1995 has the least sales with a median below 50.

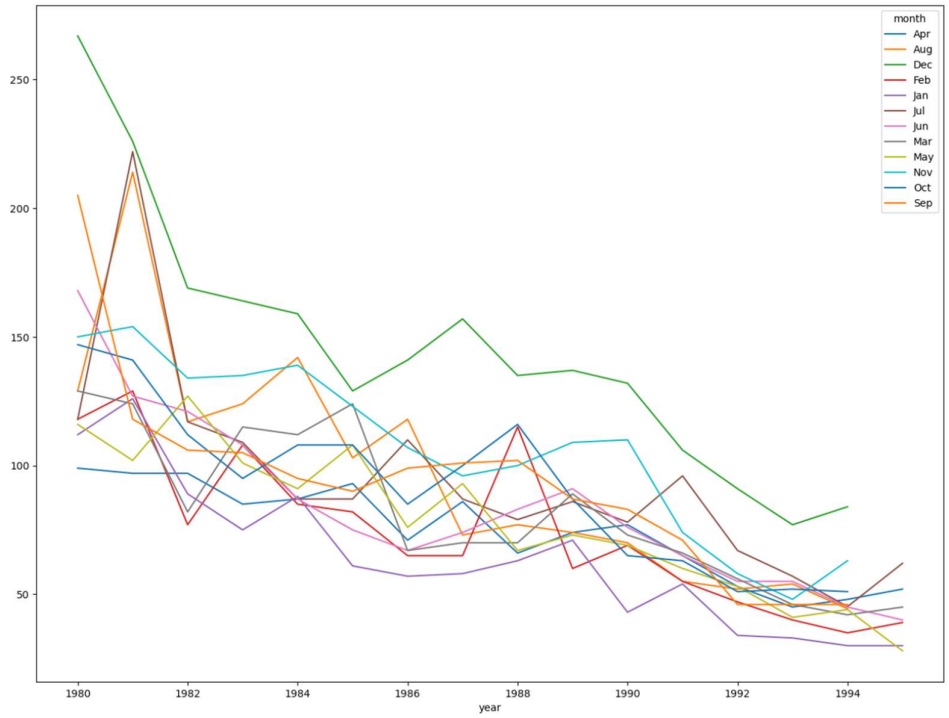

Figure 7

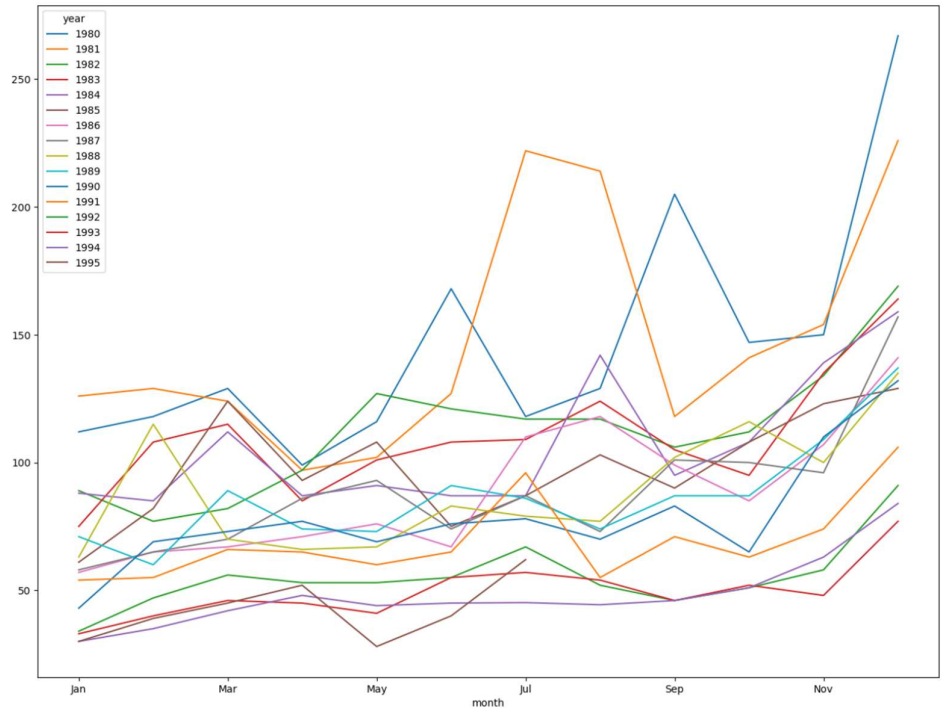

Figure 8

The above plots further explains the montly and yearly trend in sales.There is a significant difference in sales between December and the rest of the months.

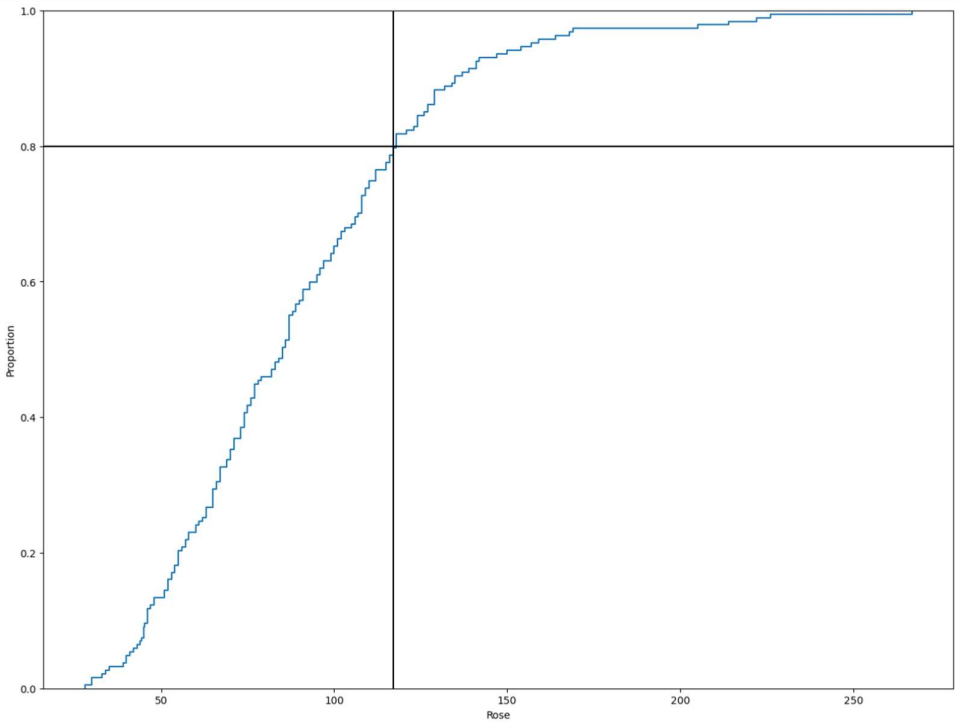

Figure 9

From the above plot we can see that around 80% of the sales is below 120.

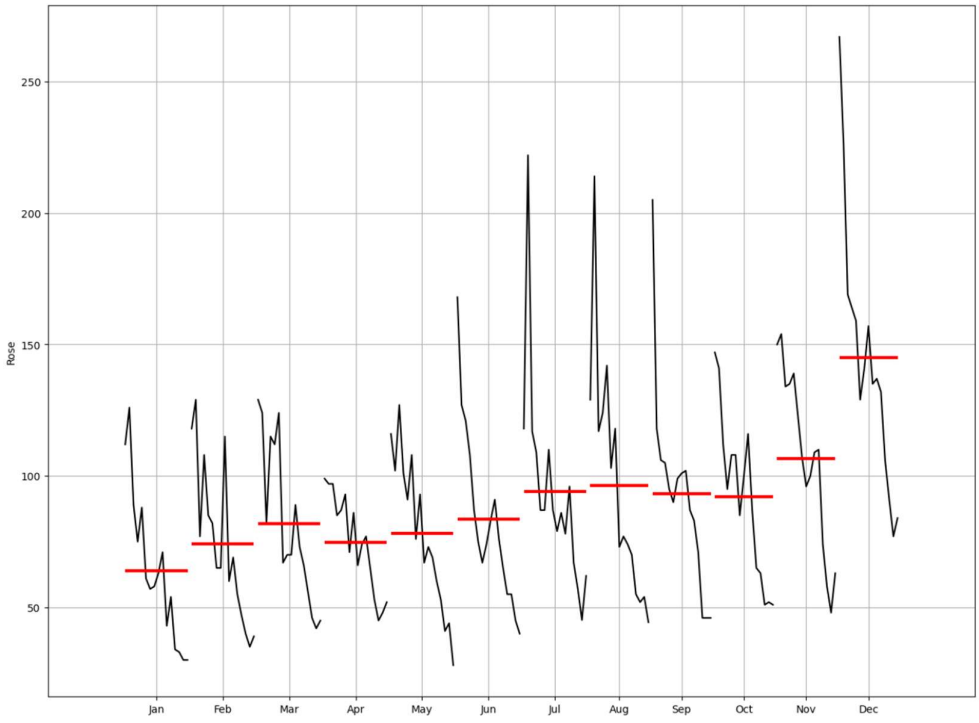

Figure 10

The month of December has the highest spread followed by Jun,July and Aug.

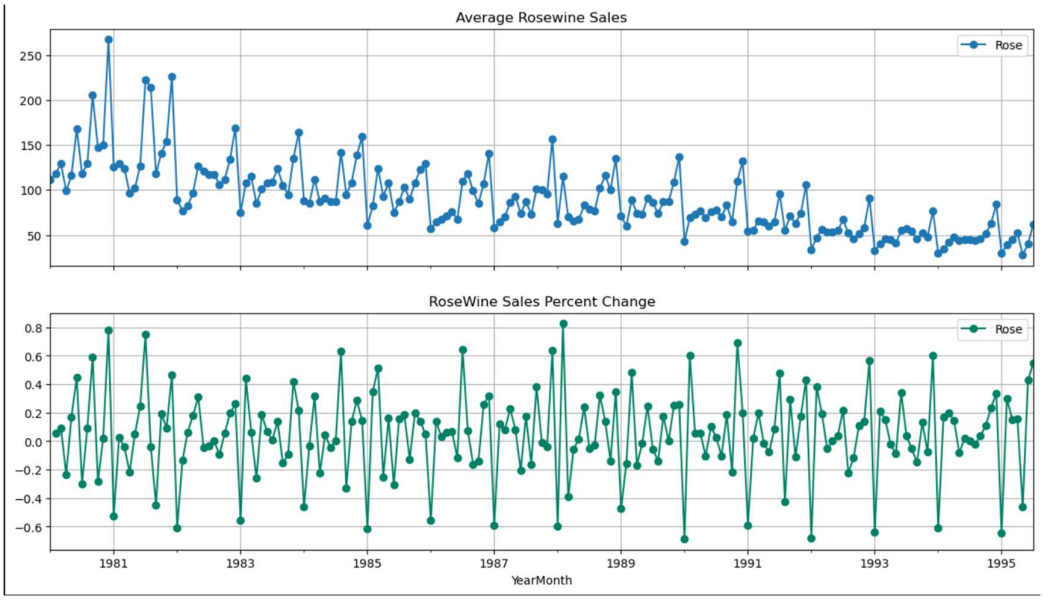

Figure 11

In terms of average sales, December consistently stands out with the highest values. Interestingly, despite the overall decrease in sales over the years, the percentage change remains relatively constant.

December sales:

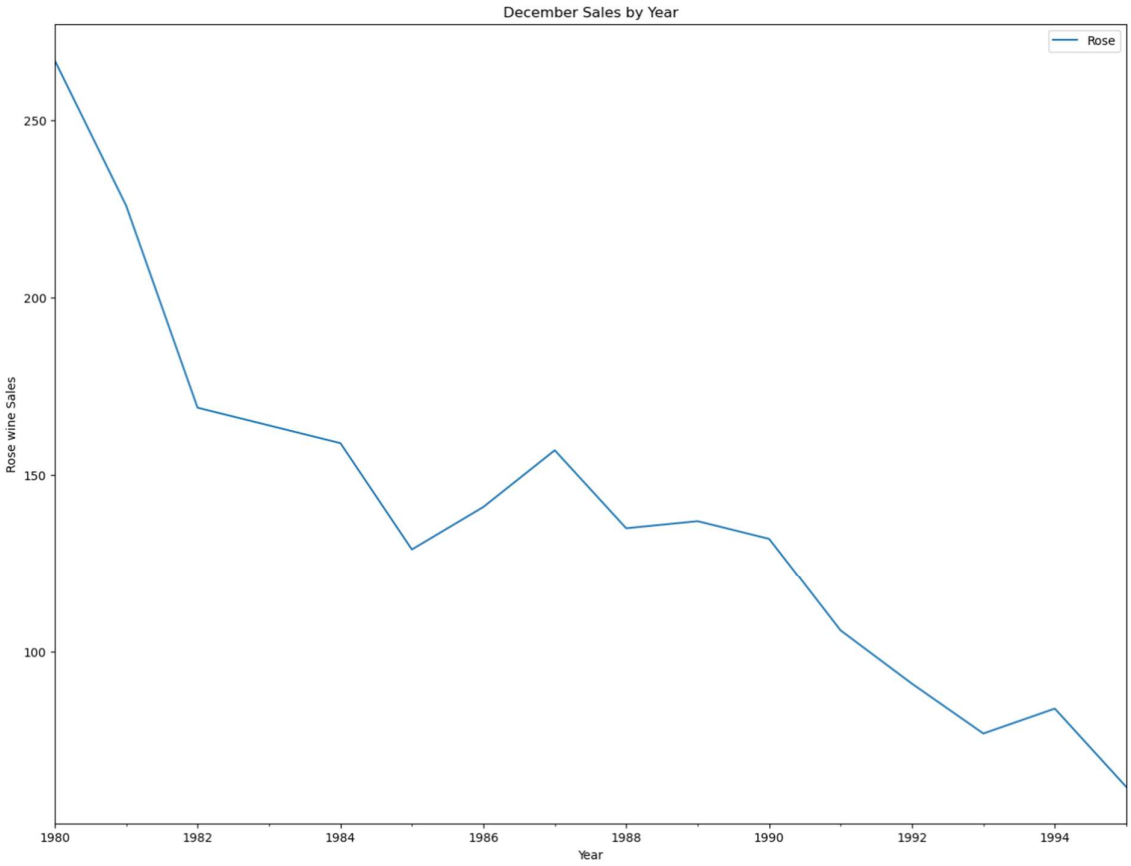

Figure 12

From the box plot, we observed that December had the highest sales every year. However, if we break it down further, we see that the sales for this most profitable month have been decreasing each year.

Sparkling Wine:

	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 13

- There are no missing values present.
- The mean sales is 2402 with a standard deviation of 1295.
- The 25^{th},50^{th} and 75^{th} peecentile values are 1605,1874,2549 respectively.

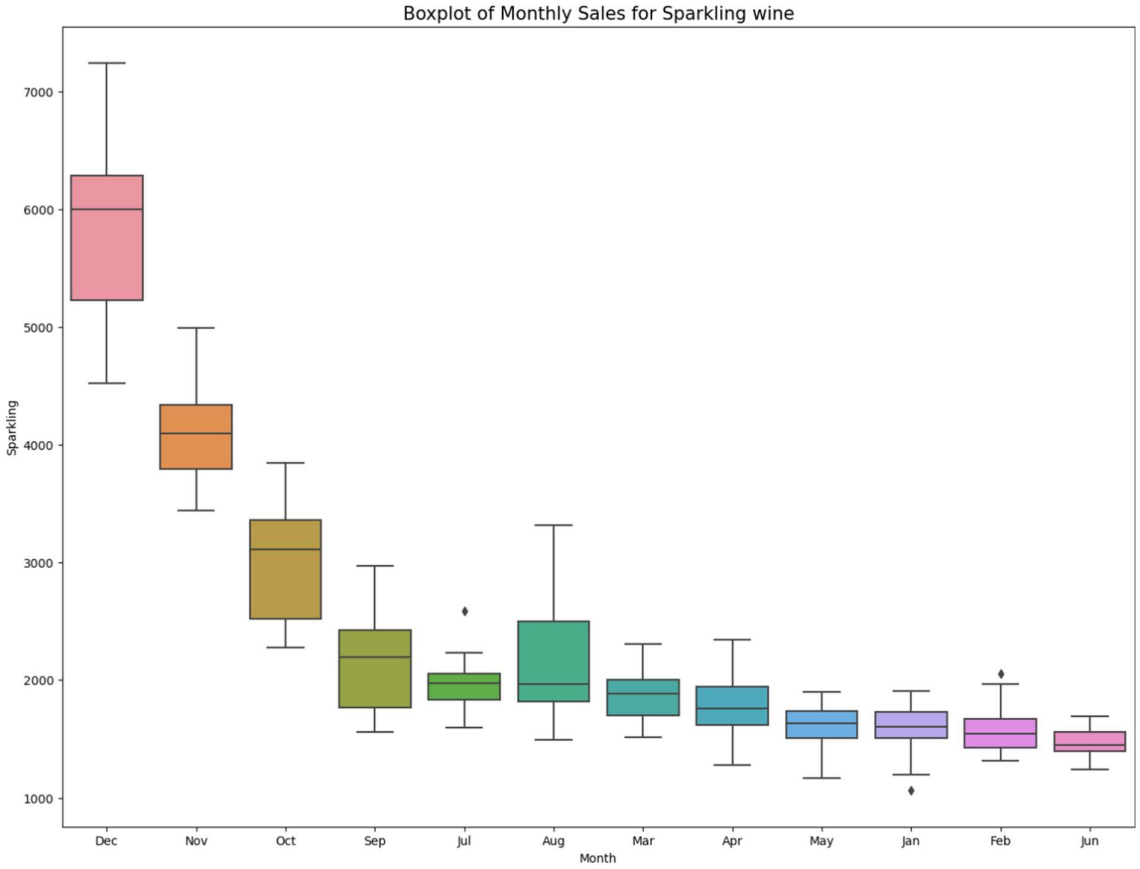

Figure 14

- The trend observed in rose wine sales also holds true for sparkling wine.
- December stands out in terms of sales as compared to the other months.
- There is a significant increase in sales towards the year-end.
- There is a 50% increase in sales between September and October. A similar trend is observed between November and December.
- Interestingly June has the least sales.

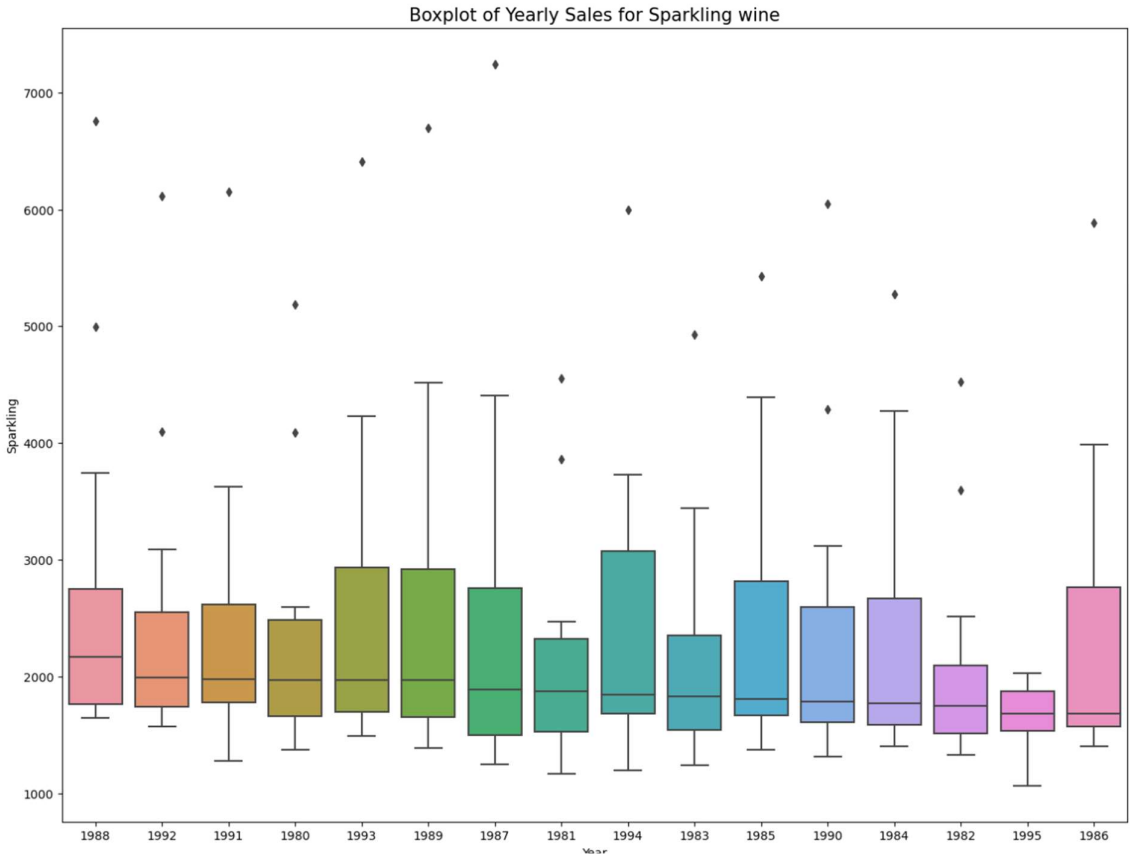

Figure 15

Unlike Rose wine,the sales of sparkling wine exhibit a steady trend, with occasional increases and decreases due to seasonality.

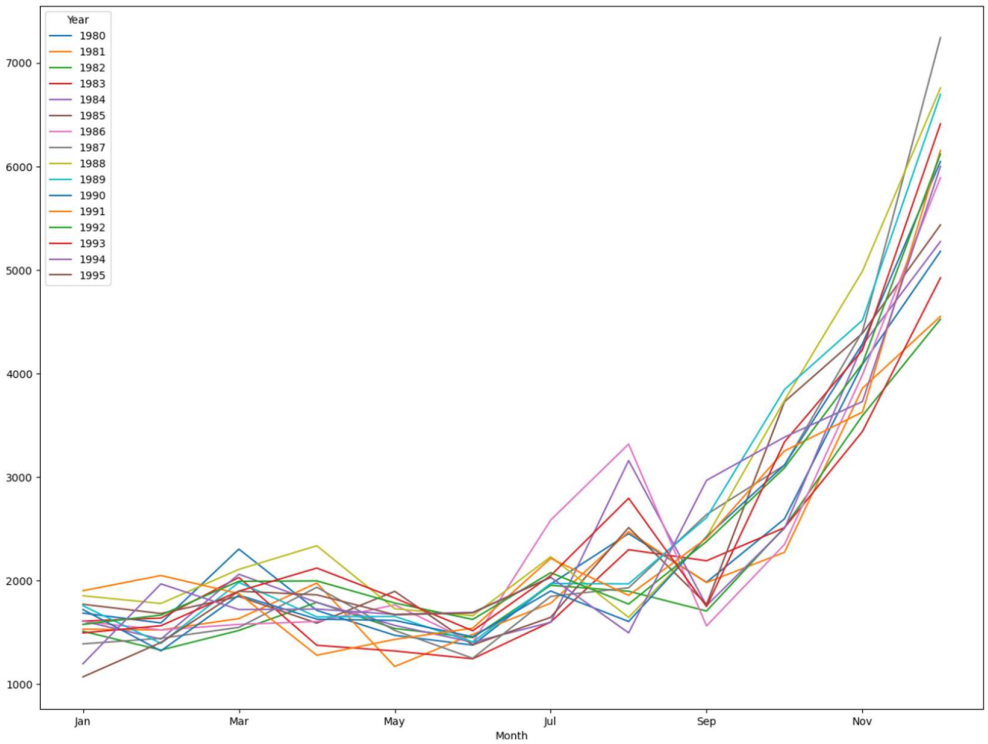

Figure 16

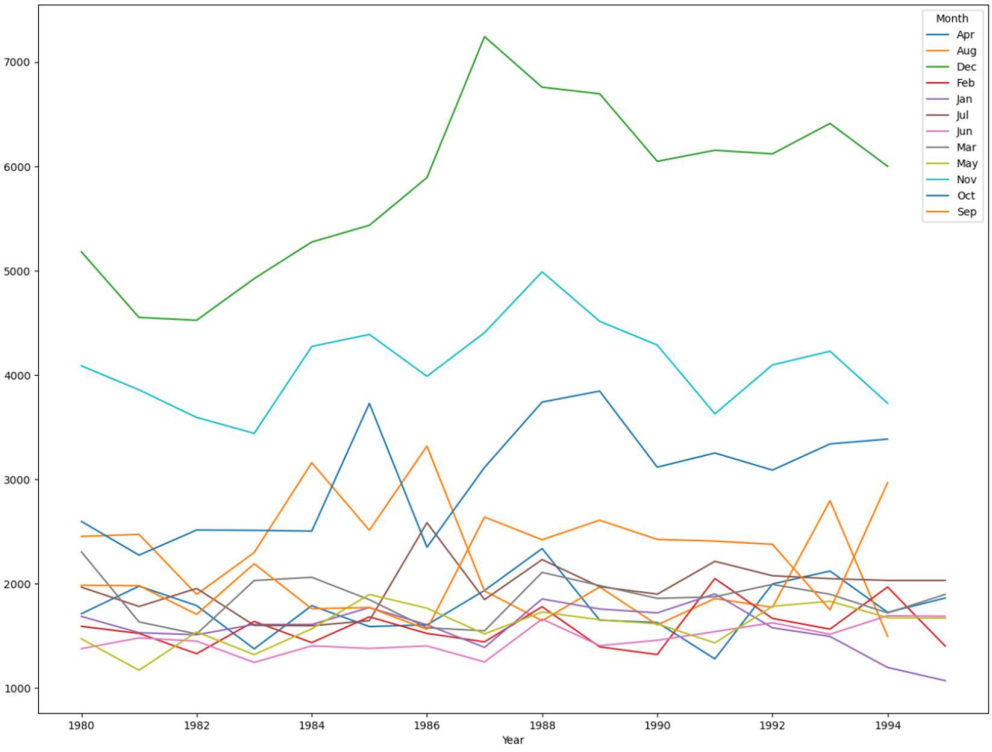

Figure 17

It is evident that,there is a significant increase in sales towards the year end especially in October,November & December.

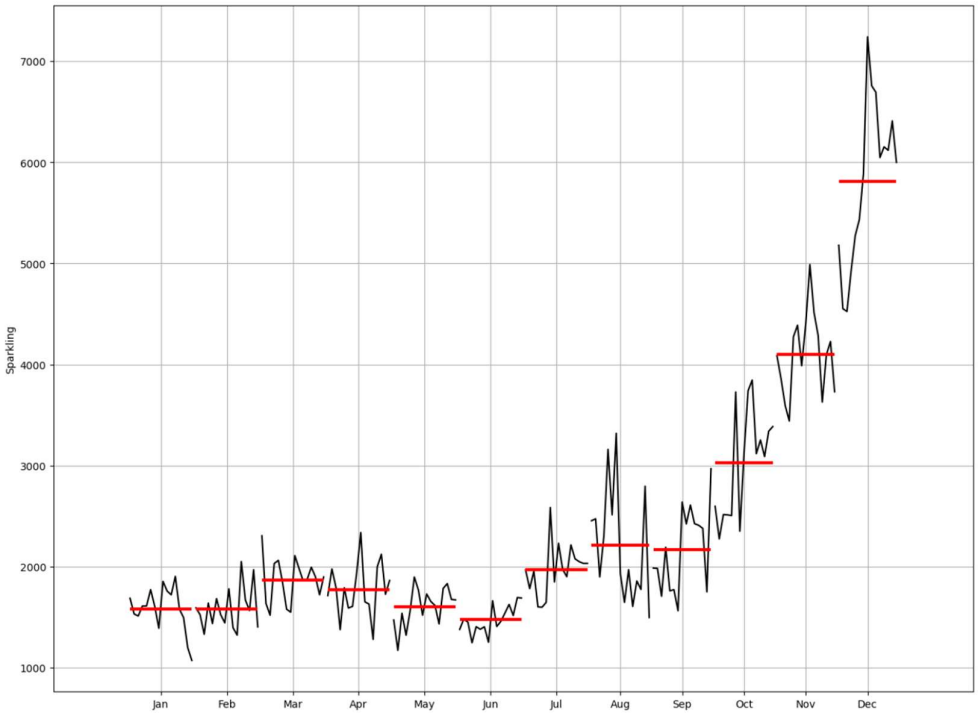

Figure 18

Indeed, the data consistently highlights December as the month with the widest spread, characterized by a median around 5900.

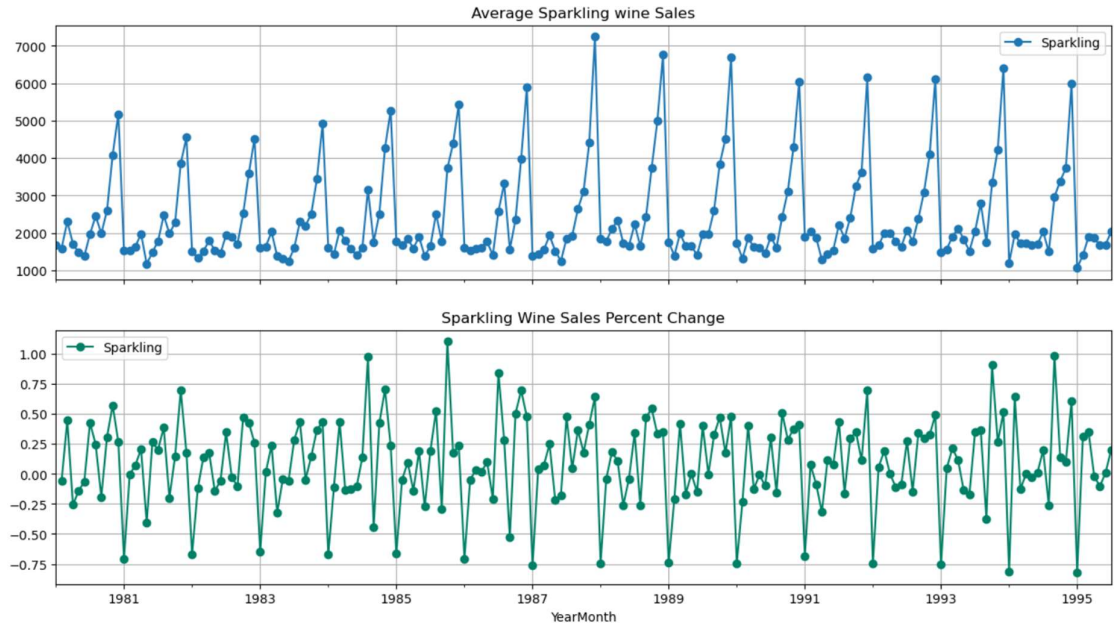

Figure 19

There is a steady increase in average sales of sparkling wine,with sales in 1988 being the highest.

Decomposition:

Sparkling Wine:

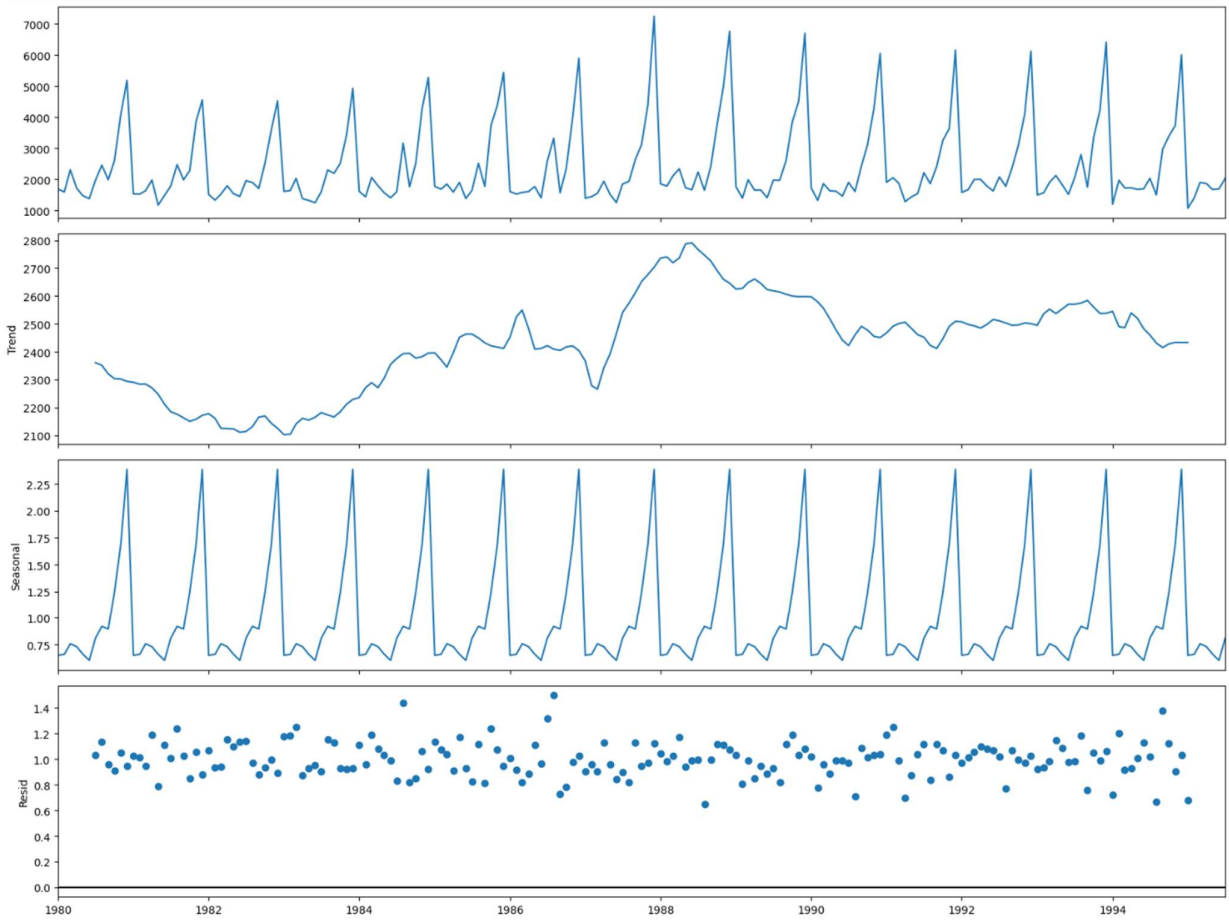

Figure 20

If we decompose the sales of sparkling wine using multiplicative seasonality,the errors or residuals are comparatively less.

Rose Wine:

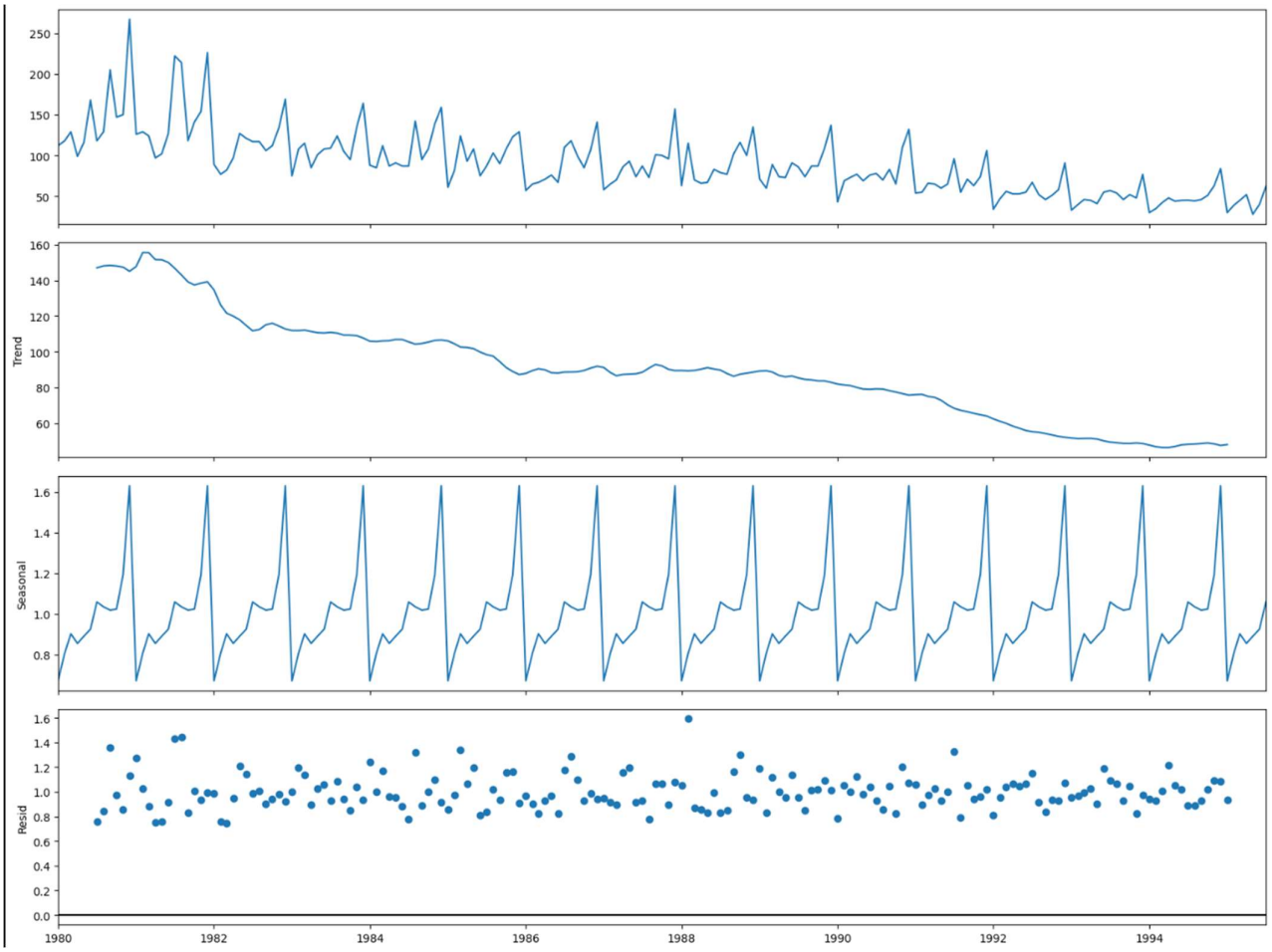

Figure 21

Similarly, incorporating multiplicative seasonality helps in reducing the error term for rose wine sales.

Data Pre-processing:

Missing value treatment:

- There are 2 missing values in the rose wine dataset.
- Using third-order polynomial interpolation to treat these missing values, as it can handle oscillations and inflection points better than quadratic interpolation.

After treating missing values:

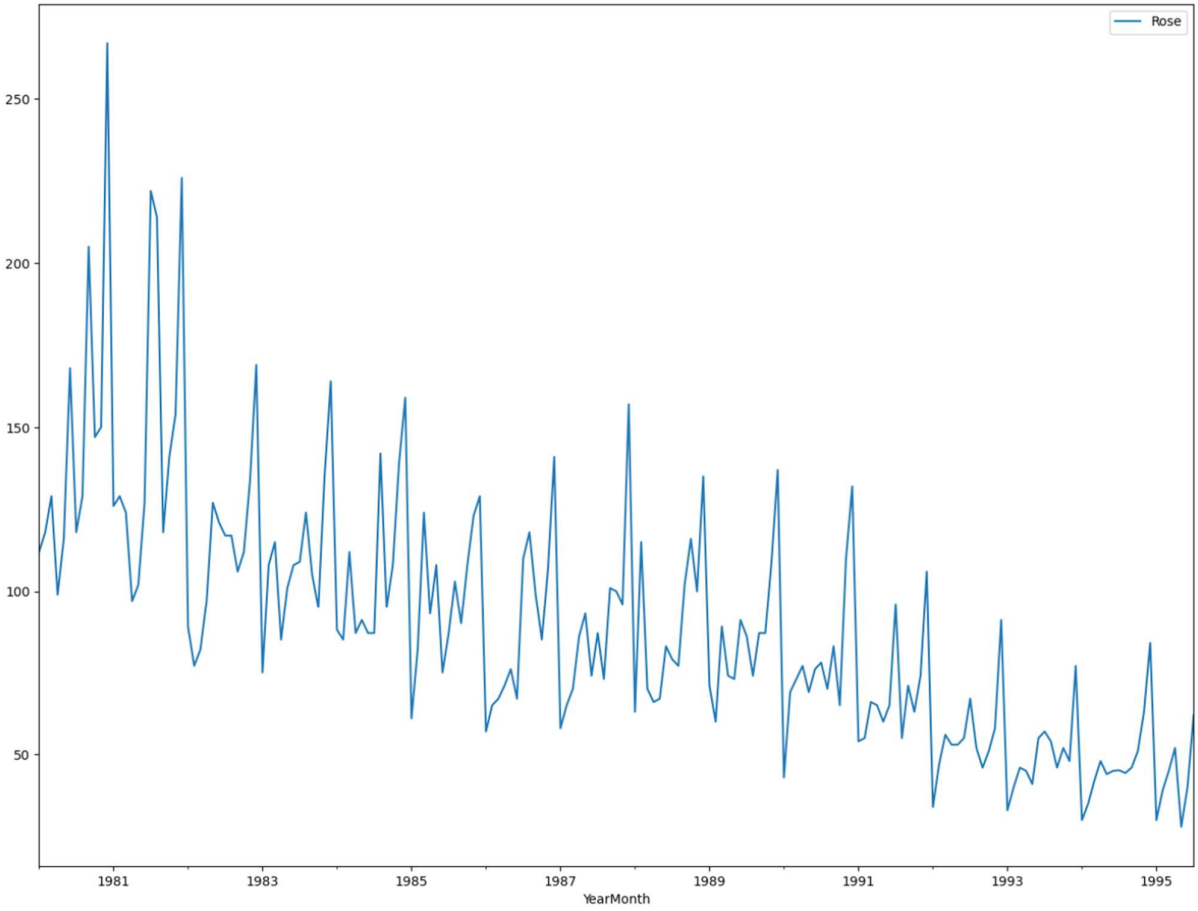

Figure 22

Train-test split:

We cannot split the dataset randomly as it will create a break in the time series and will cause an uneven time frame.

- Train data: Sales between 1980 to 1991.
- Test data: Sales between 1992 to 1995.

The last few rows of Train data:

YearMonth	Rose
1991-08-01	55.0
1991-09-01	71.0
1991-10-01	63.0
1991-11-01	74.0
1991-12-01	106.0

YearMonth	Sparkling
1991-08-01	1857
1991-09-01	2408
1991-10-01	3252
1991-11-01	3627
1991-12-01	6153

Figure 23

First few rows of Test data:

YearMonth	Rose
1992-01-01	34.0
1992-02-01	47.0
1992-03-01	56.0
1992-04-01	53.0
1992-05-01	53.0

YearMonth	Sparkling
1992-01-01	1577
1992-02-01	1667
1992-03-01	1993
1992-04-01	1997
1992-05-01	1783

Figure 24

Rose wine:

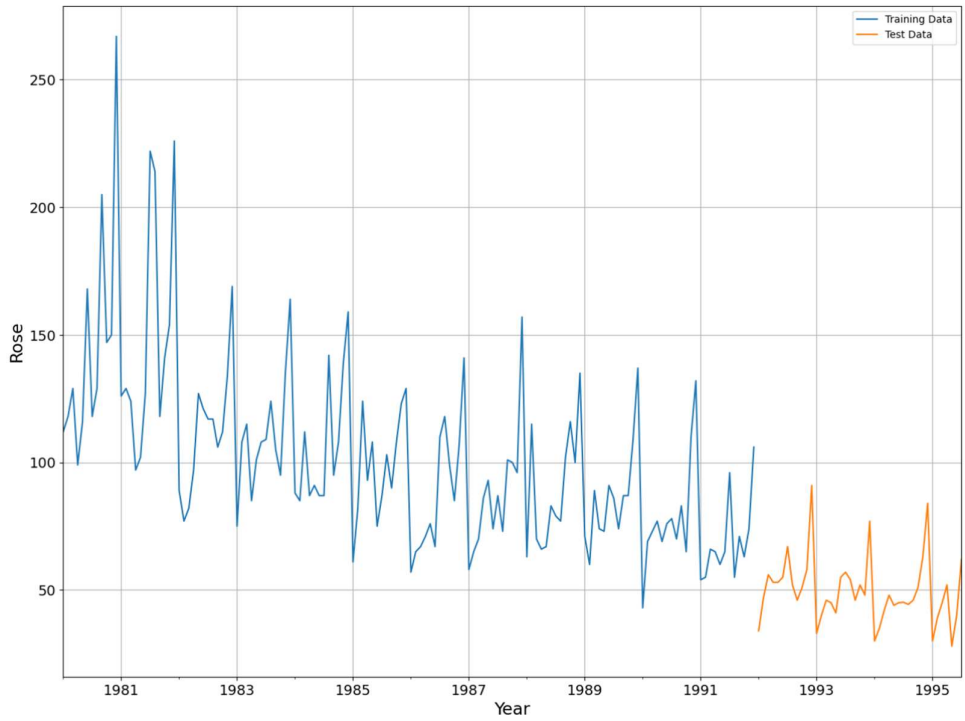

Figure 25

Sparkling wine:

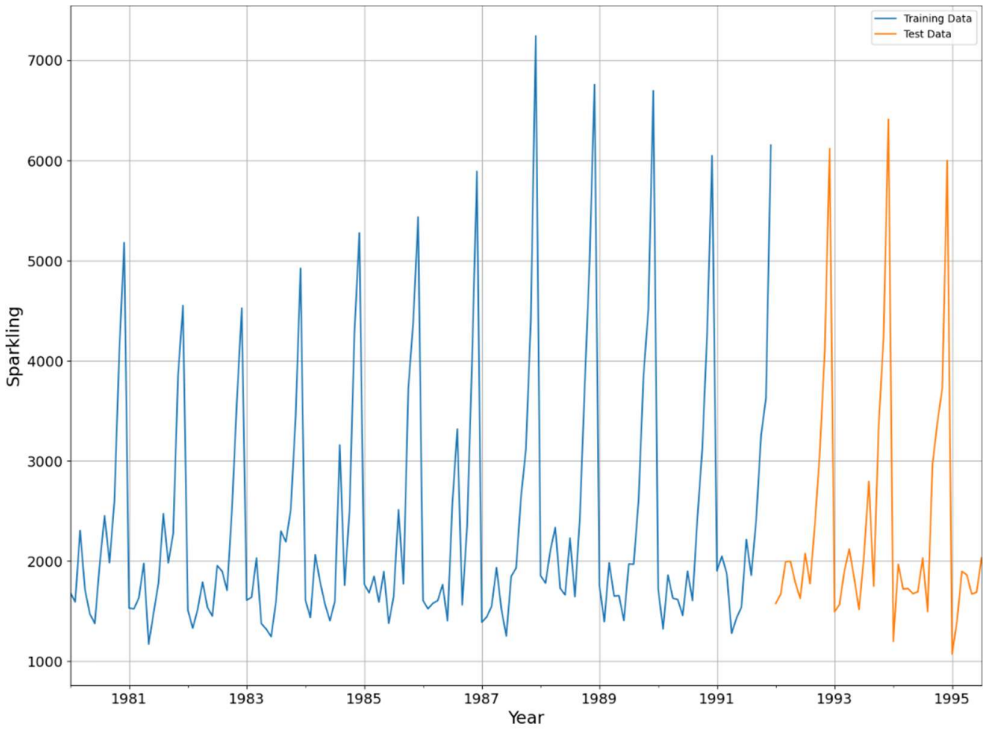

Figure 26

Model Building :

Linear regression:

Here we are going to regress the sales against every instance of time. Before doing that, we need to create instances for both train and test data.

A few rows of modified train data:

YearMonth	Rose	time
1980-01-01	112.0	1
1980-02-01	118.0	2
1980-03-01	129.0	3
1980-04-01	99.0	4
1980-05-01	116.0	5

Figure 27

YearMonth	Sparkling	time
1980-01-01	1686	1
1980-02-01	1591	2
1980-03-01	2304	3
1980-04-01	1712	4
1980-05-01	1471	5

Figure 28

Model evaluation:

For model evaluation, we are going to use root mean squared error. It measures the average deviation between the actual and the predicted values. The lower the RMSE, the better the model.

Rose wine:

For the Regression On Time forecast on the Test Data, RMSE is **14.921**.

Sparkling wine:

For the Regression On Time forecast on the Test Data, RMSE is **1356.62**.

Rose wine

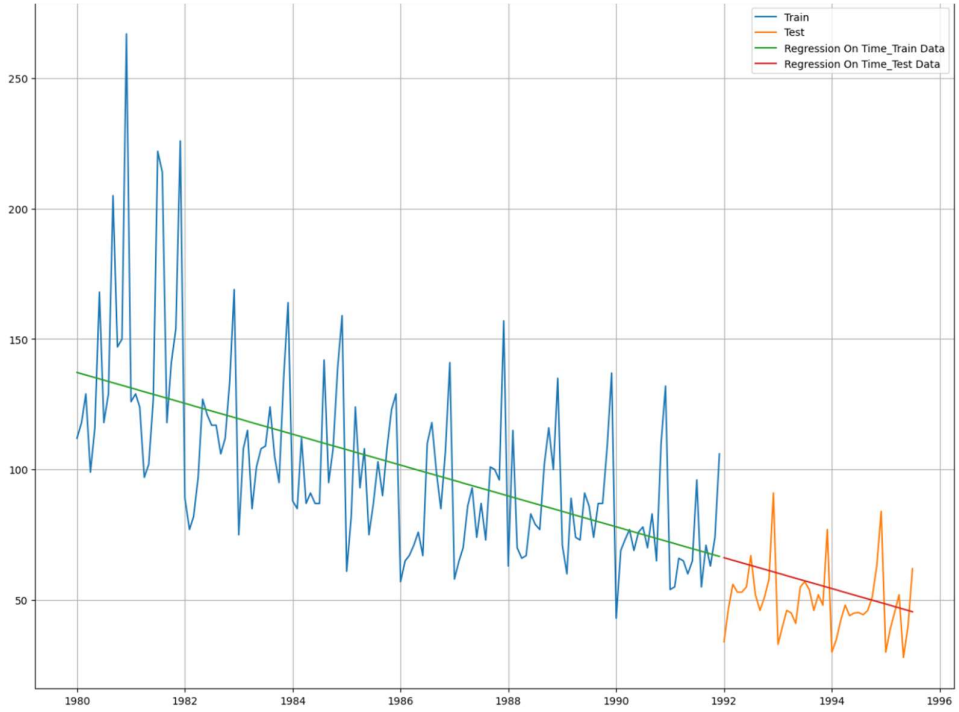

Figure 29

Sparkling wine:

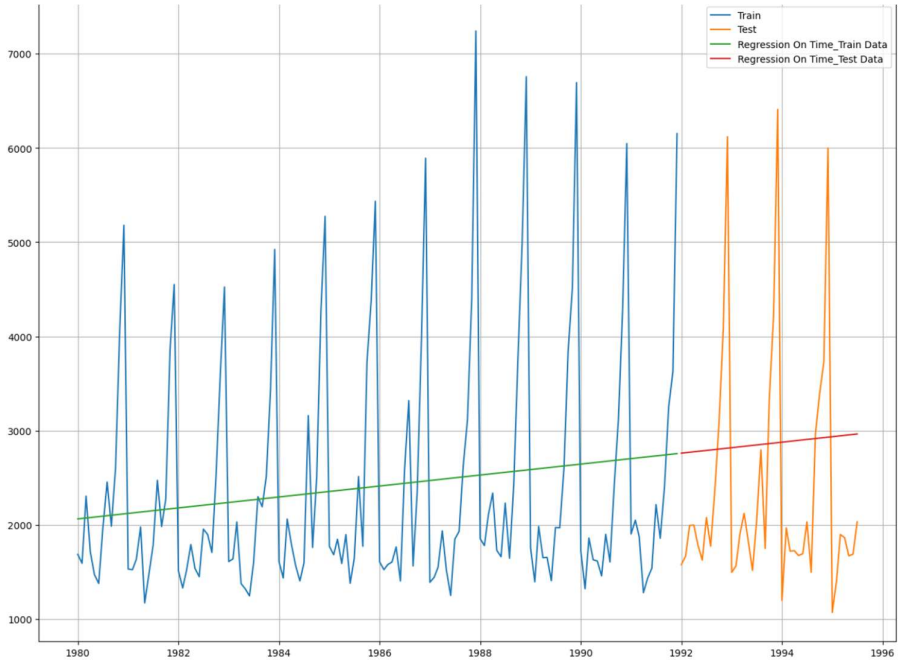

Figure 30

The model has captured only the trend but not the seasonality.

Simple Average:

Simple average model estimates the forecast by calculating the mean of all the past variables.

A few rows of test data:

YearMonth	Rose	Mean_Forecast
1992-01-01	34.0	101.958333
1992-02-01	47.0	101.958333
1992-03-01	56.0	101.958333
1992-04-01	53.0	101.958333
1992-05-01	53.0	101.958333

Figure 31

YearMonth	Sparkling	Mean_Forecast
1992-01-01	1577	2408.930556
1992-02-01	1667	2408.930556
1992-03-01	1993	2408.930556
1992-04-01	1997	2408.930556
1992-05-01	1783	2408.930556

Figure 32

Rose wine:

For the Simple Average forecast on the Test Data, RMSE is **53.97.**

Sparkling wine**:**

For the Simple Average forecast on the Test Data, RMSE is **1268.68.**

Predictions:

Rose wine:

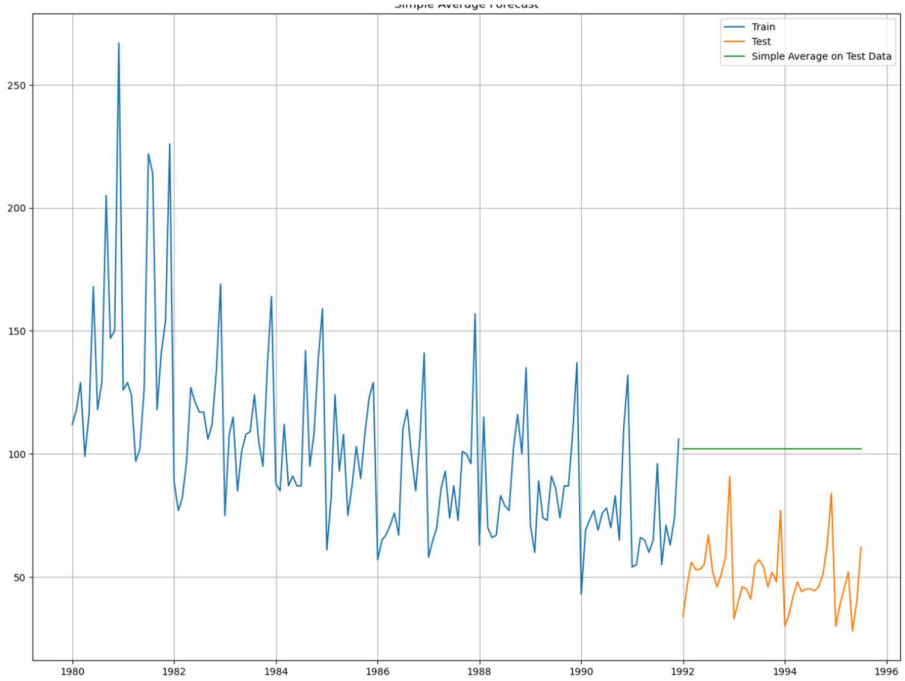

Figure 33

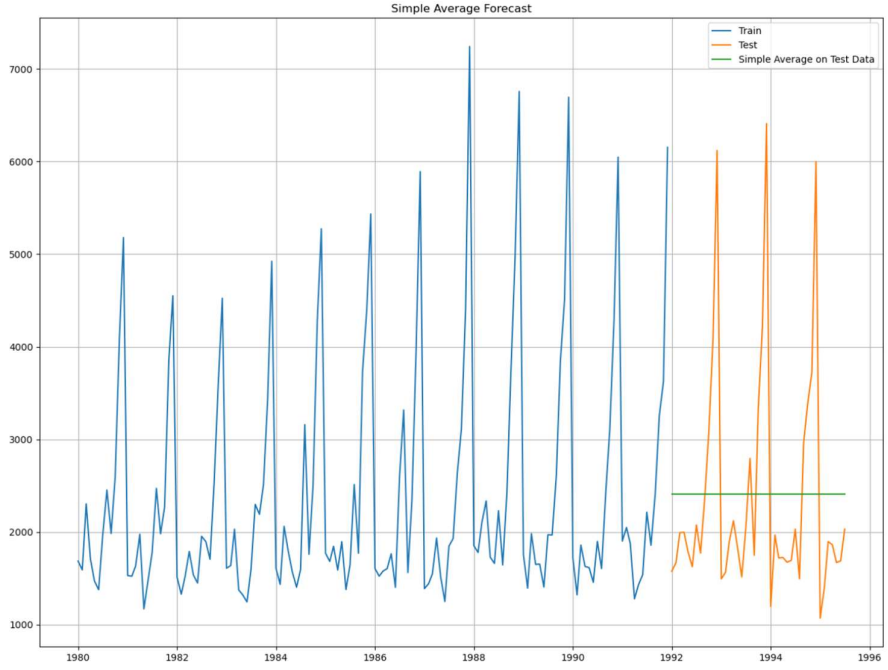

Figure 34

Moving Average:

This model calculates the rolling or moving mean for different time intervals. We are going to build models on intervals of 2,4,6,8 and 10.

Modified data:

YearMonth	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_8	Trailing_10
1980-01-01	112.0	NaN	NaN	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN	NaN	NaN
1980-04-01	99.0	114.0	114.50	NaN	NaN	NaN
1980-05-01	116.0	107.5	115.50	NaN	NaN	NaN
1980-06-01	168.0	142.0	128.00	123.666667	NaN	NaN
1980-07-01	118.0	143.0	125.25	124.666667	NaN	NaN
1980-08-01	129.0	123.5	132.75	126.500000	123.625	NaN
1980-09-01	205.0	167.0	155.00	139.166667	135.250	NaN
1980-10-01	147.0	176.0	149.75	147.166667	138.875	134.1

Figure 35

YearMonth	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_8	Trailing_10
1980-01-01	1686	NaN	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN	NaN
1980-06-01	1377	1424.0	1716.00	1690.166667	NaN	NaN
1980-07-01	1966	1671.5	1631.50	1736.833333	NaN	NaN
1980-08-01	2453	2209.5	1816.75	1880.500000	1820.000	NaN
1980-09-01	1984	2218.5	1945.00	1827.166667	1857.250	NaN
1980-10-01	2596	2290.0	2249.75	1974.500000	1982.875	1914.0

Figure 36

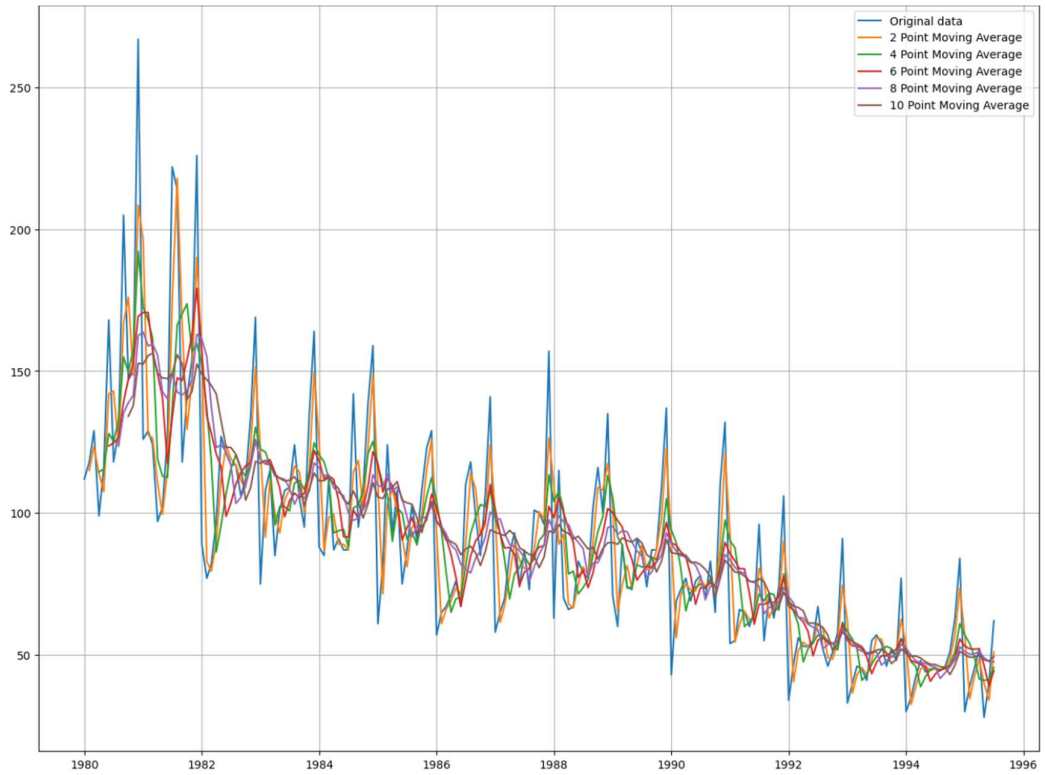

Figure 37

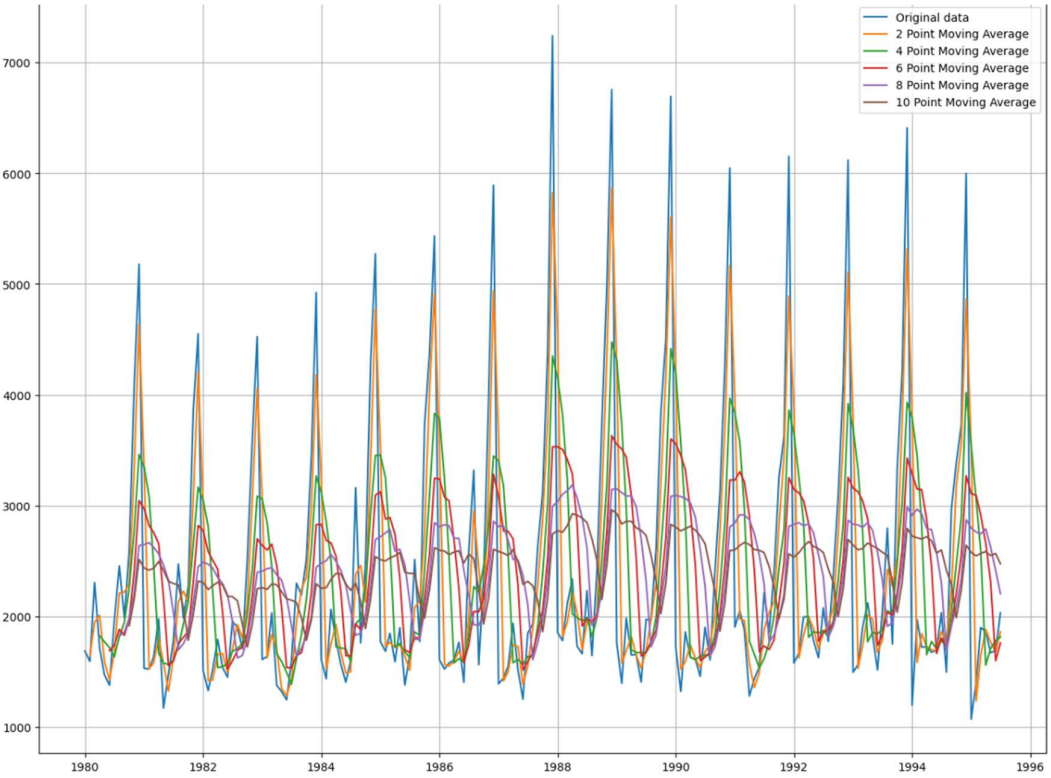

Figure 38

Test-Train data for moving average:

Rose wine:

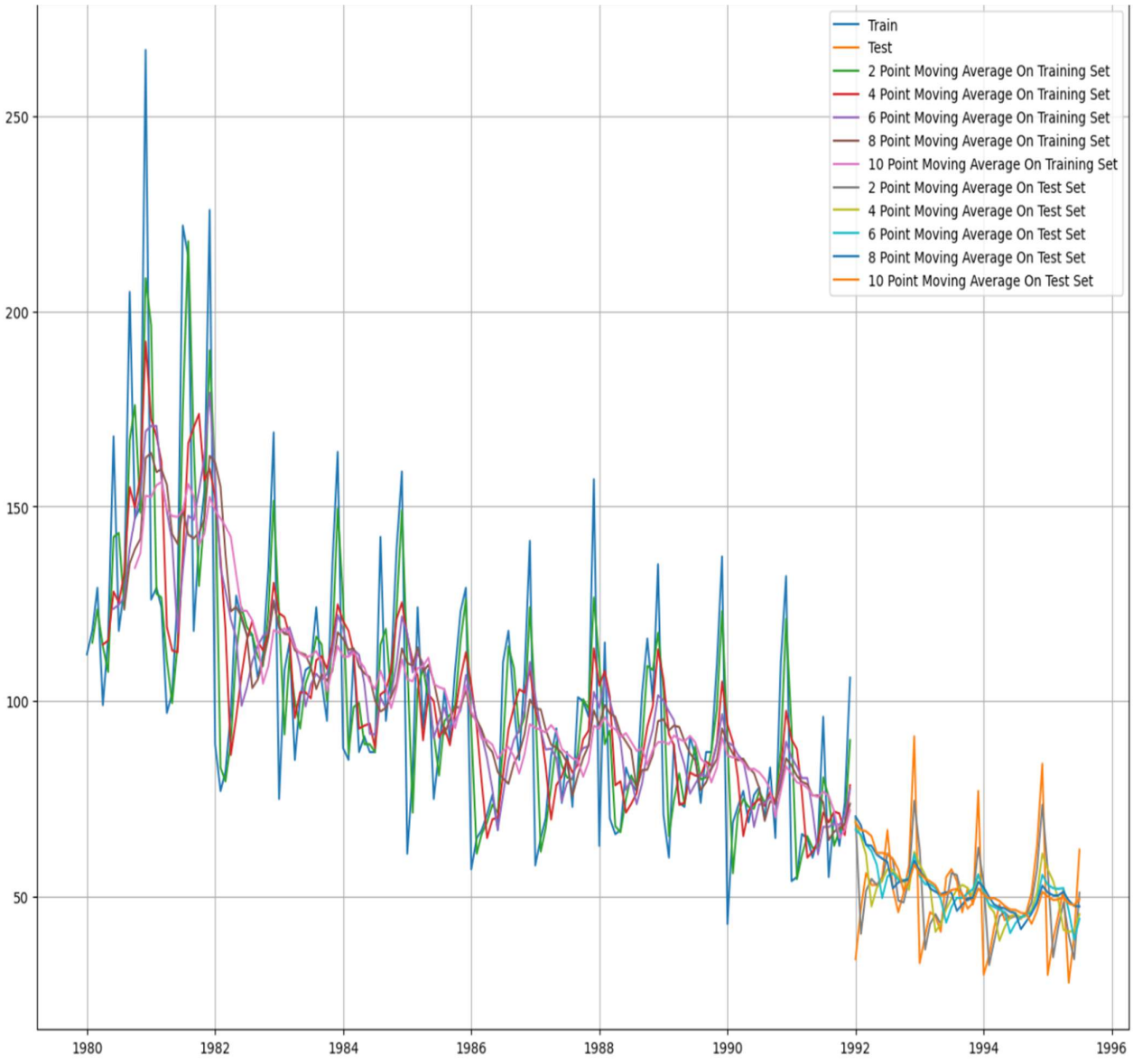

Figure 39

Sparkling wine:

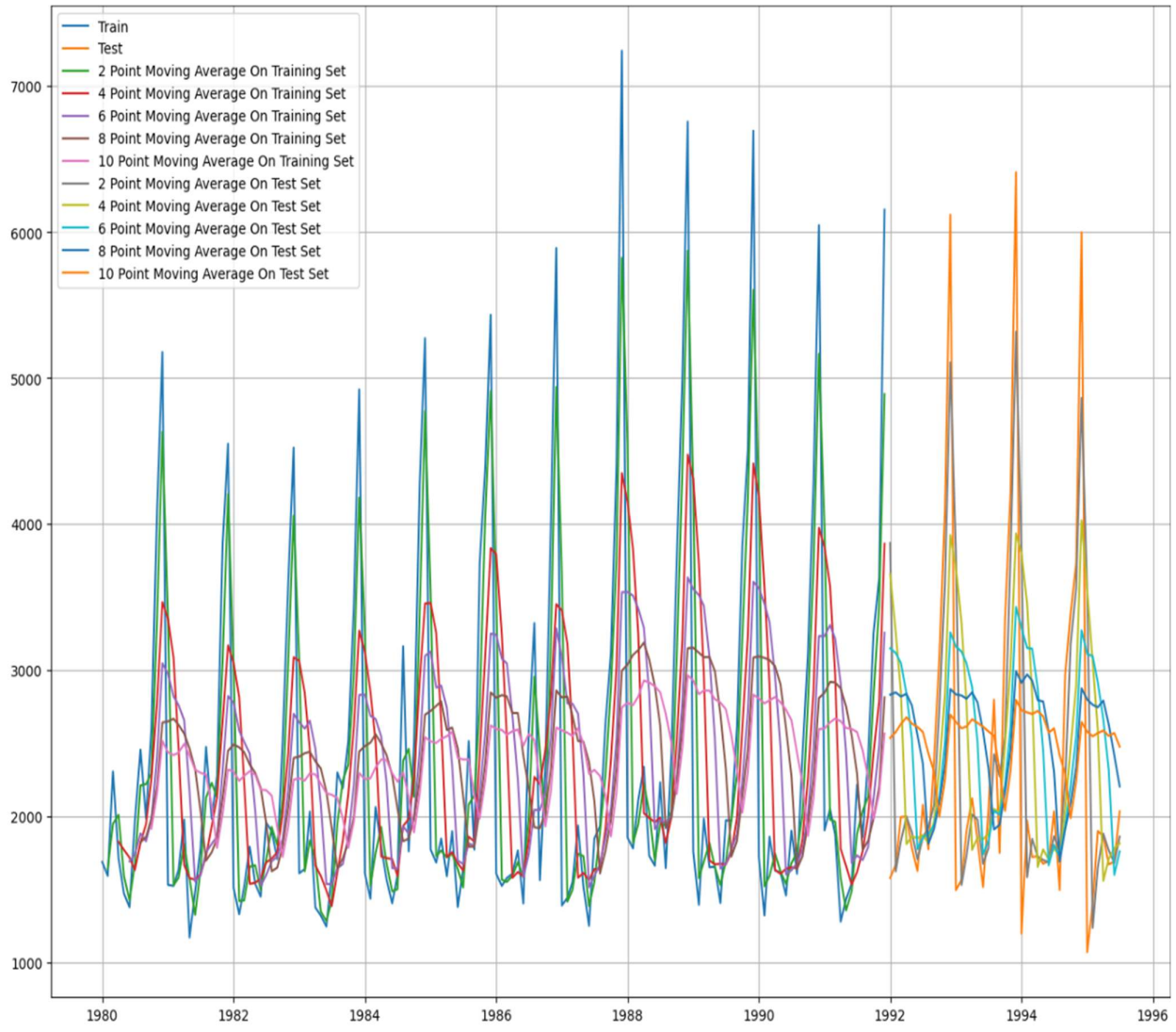

Figure 40

Rose wine:

- For the 2-point Moving Average Model forecast on the Test Data, **RMSE is 10.47**.
- For the 4-point Moving Average Model forecast on the Test Data, **RMSE is 13.01**.
- For the 6-point Moving Average Model forecast on the Test Data, **RMSE is 13.17**.
- For the 8-point Moving Average Model forecast on the Test Data, RMSE is **13.72.**
- For the 10-point Moving Average Model forecast on the Test Data, RMSE is **13.85.**

Sparkling wine:

- For the 2-point Moving Average Model forecast on the Test Data, **RMSE is 834.62.**
- For the 4-point Moving Average Model forecast on the Test Data, **RMSE is 1169.86.**
- For the 6-point Moving Average Model forecast on the Test Data, **RMSE is 1277.86.**
- For the 8-point Moving Average Model forecast on the Test Data, **RMSE is 1329.14.**
- For the 10-point Moving Average Model forecast on the Test Data, **RMSE is 1324.53.**

Simple exponential smoothening:

- This model is suitable for data with no trend and seasonality.
- Alpha or level is the weight given to each observation.
- Exponential Smoothing is the weighted average of past observations.

By using the autofit, the parameters obtained are:

Rose wine:

```
{'smoothing_level': 0.09972330553818615,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.35922330517218,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Figure 41

Sparkling wine:

```
{'smoothing_level': 0.056234830918812144,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1804.0444740358766,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Figure 42

Rose wine:

For Alpha =0.099 Simple Exponential Smoothing Model forecast on the Test Data, **RMSE is 30.09.**

Sparkling wine:

For Alpha =0.05 Simple Exponential Smoothing Model forecast on the Test Data, **RMSE is 1310.26.**

Predictions:

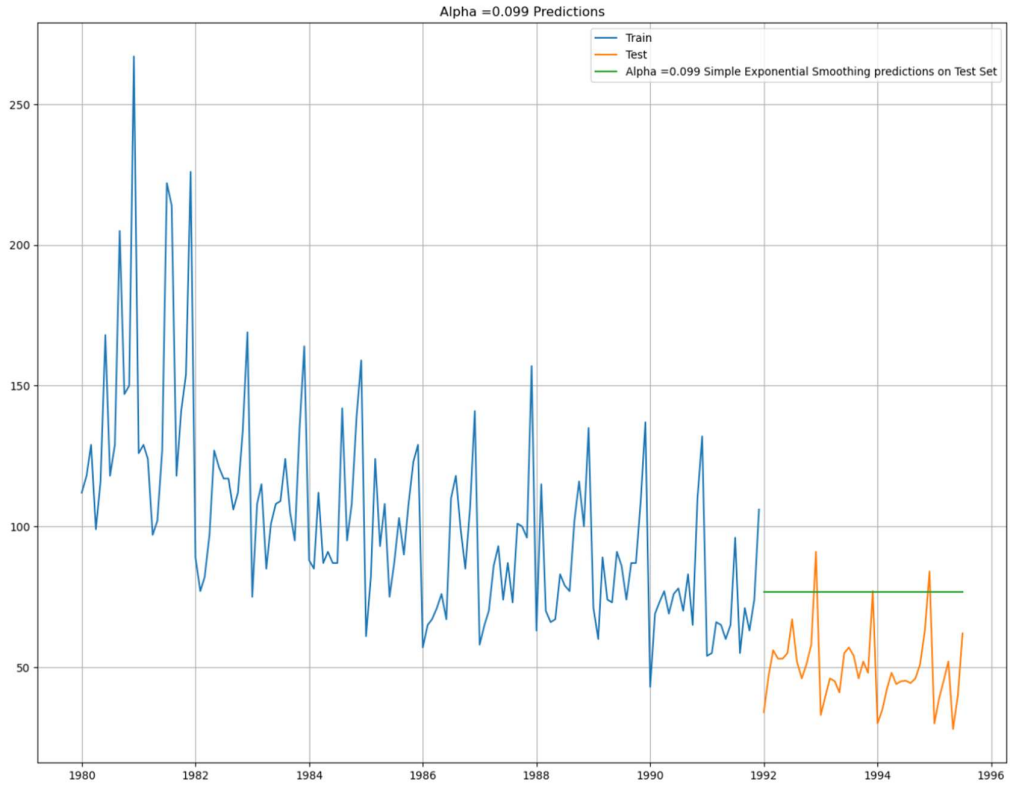

Figure 43

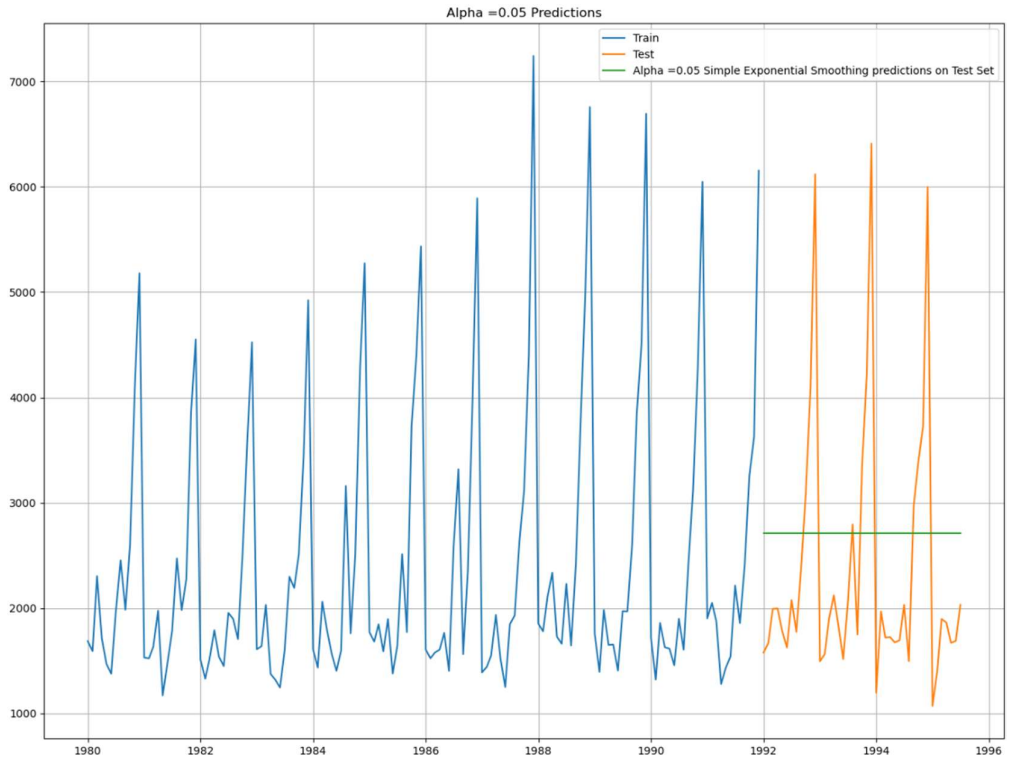

Figure 44

Double exponential smoothening:

This is suitable for data with trends but no seasonality.Using the autofit, the parameters obtained are

Rose wine:

- smoothing_level: 1.49e-08
- smoothing_trend: 3.59e-09.

Sparkling wine:

- smoothing_level: 0.66
- smoothing_trend: 0.0001.

Model evaluation:

Rose wine:

For Alpha =1.49e-08 and Beta = 3.59e-09 Double Exponential Smoothing Model forecast on the Test Data, **RMSE is 14.92.**

Sparkling wine:

For Alpha =0.66 and Beta = 0.0001 Double Exponential Smoothing Model forecast on the Test Data, **RMSE is 4773.35.**

Triple Exponential Smoothening Additive Seasonality:

This takes into account level, trend, and seasonality.

Rose wine:

- smoothing_level: 0.08
- smoothing_trend: 3.79e-07,
- smoothing_seasonal: 3.66e-05.

Sparkling wine:

- smoothing_level: 0.075
- smoothing_trend: 0.04
- smoothing_seasonal: 0.44.

YearMonth	Rose	predict
1992-01-01	34.0	33.631371
1992-02-01	47.0	44.562453
1992-03-01	56.0	52.337087
1992-04-01	53.0	42.027122
1992-05-01	53.0	49.215645

YearMonth	Sparkling	predict
1992-01-01	1577	1703.778080
1992-02-01	1667	1599.061319
1992-03-01	1993	1753.918784
1992-04-01	1997	1409.440014
1992-05-01	1783	1433.236770

Figure 45

Model evaluation:

Rose wine:

For Alpha =0.088, Beta = 3.79e-07, and Gamma = 3.66e-05 Triple Exponential Smoothing Model forecast on the Test Data, RMSE is **13.08.**

Sparkling wine:

For Alpha =0.075, Beta = 0.043, and Gamma = 0.445 Triple Exponential Smoothing Model forecast on the Test Data, RMSE is **368.57.**

Triple Exponential Smoothening Multiplicative Seasonality:

Rose wine:

- smoothing_level: 0.099
- smoothing_trend: 0.0003
- smoothing_seasonal: 1.09e-06.

Sparkling wine:

- smoothing_level: 0.076
- smoothing_trend: 0.076
- smoothing_seasonal: 0.342.

YearMonth	Rose	predict_multiplicative
1992-01-01	34.0	47.358533
1992-02-01	47.0	52.815662
1992-03-01	56.0	57.433606
1992-04-01	53.0	50.144011
1992-05-01	53.0	55.344216

YearMonth	Sparkling	predict_multiplicative
1992-01-01	1577	1720.344929
1992-02-01	1667	1612.471009
1992-03-01	1993	1795.028838
1992-04-01	1997	1531.065710
1992-05-01	1783	1508.856859

Figure 46

Model evaluation:

Rose wine:

For Alpha =0.09, Beta = 0.0003, and Gamma = 1.09e-06 Triple Exponential Smoothing Model forecast on the Test Data, RMSE is **8.45**.

Sparkling wine:

For Alpha =0.076, Beta = 0.076, and Gamma =0.342 Triple Exponential Smoothing Model forecast on the Test Data, RMSE is **347.34.**

Predictions:

Rose wine:

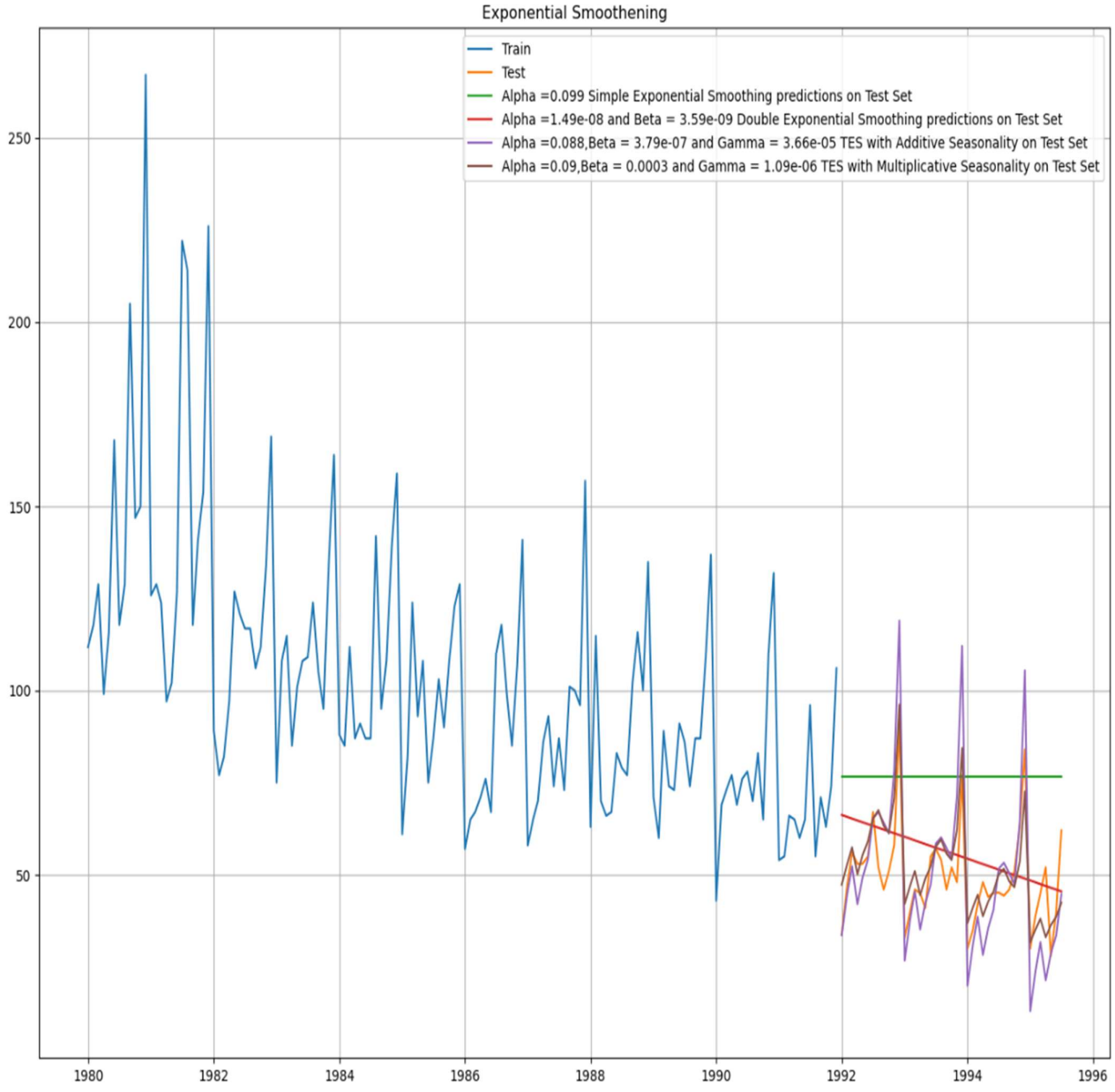

Figure 47

Sparkling wine:

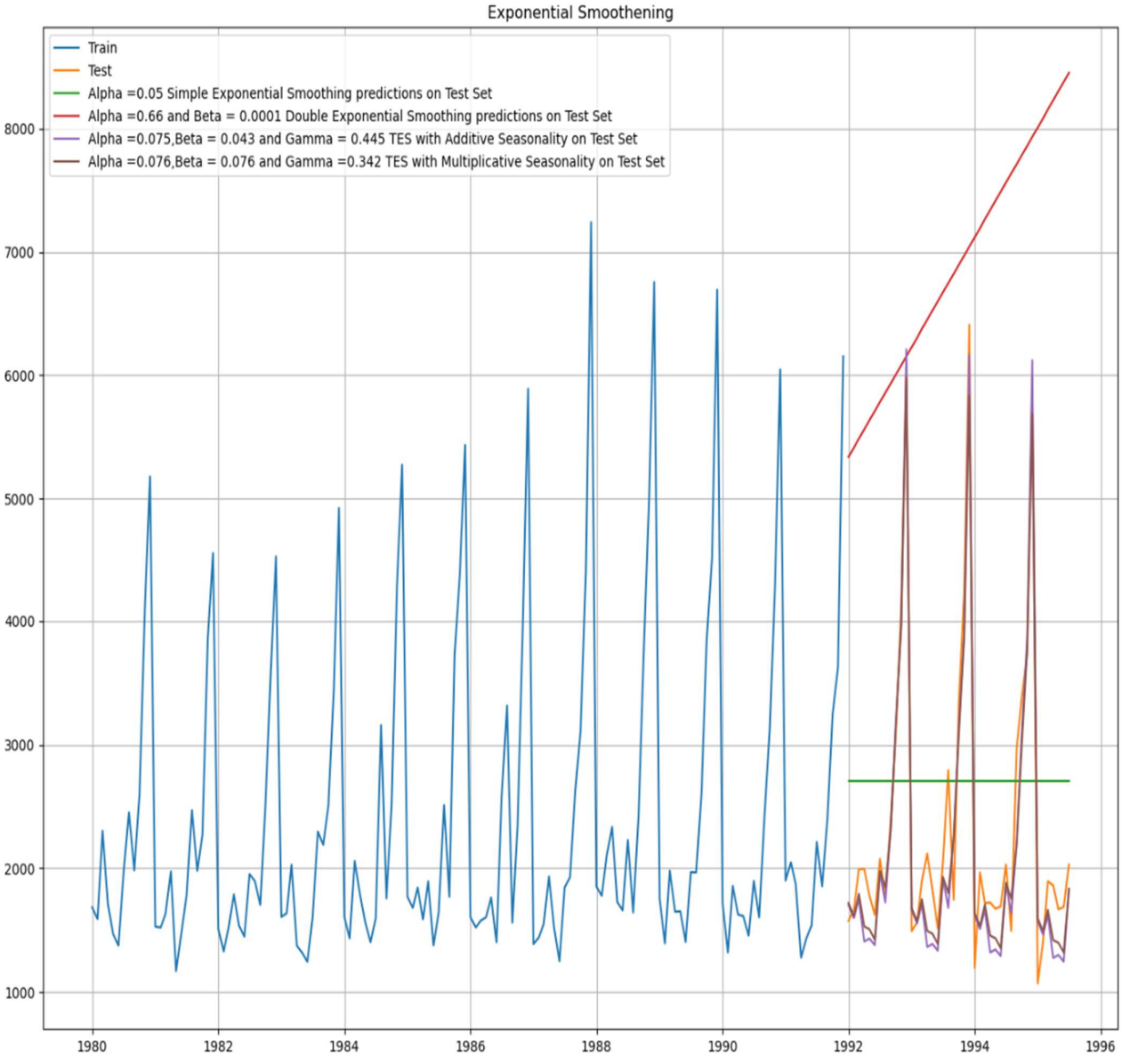

Figure 48

Check for Stationarity:

For building ARIMA and SARIMA models, the time series should be stationary.

A time series is said to be stationary when its statistical properties such as mean, variance, and standard deviation remains constant over time.

The Dickey-Fuller Test is a hypothesis test for checking stationarity.

- **Null Hypothesis** $H0$: Time Series is non-stationary.
- **Alternate Hypothesis** $\boldsymbol{Ha}$: Time Series is stationary.

So Ideally if p-value < 0.05 then the null hypothesis: TS is non-stationary is rejected else the TS is non-stationary is failed to be rejected.

Rose wine:

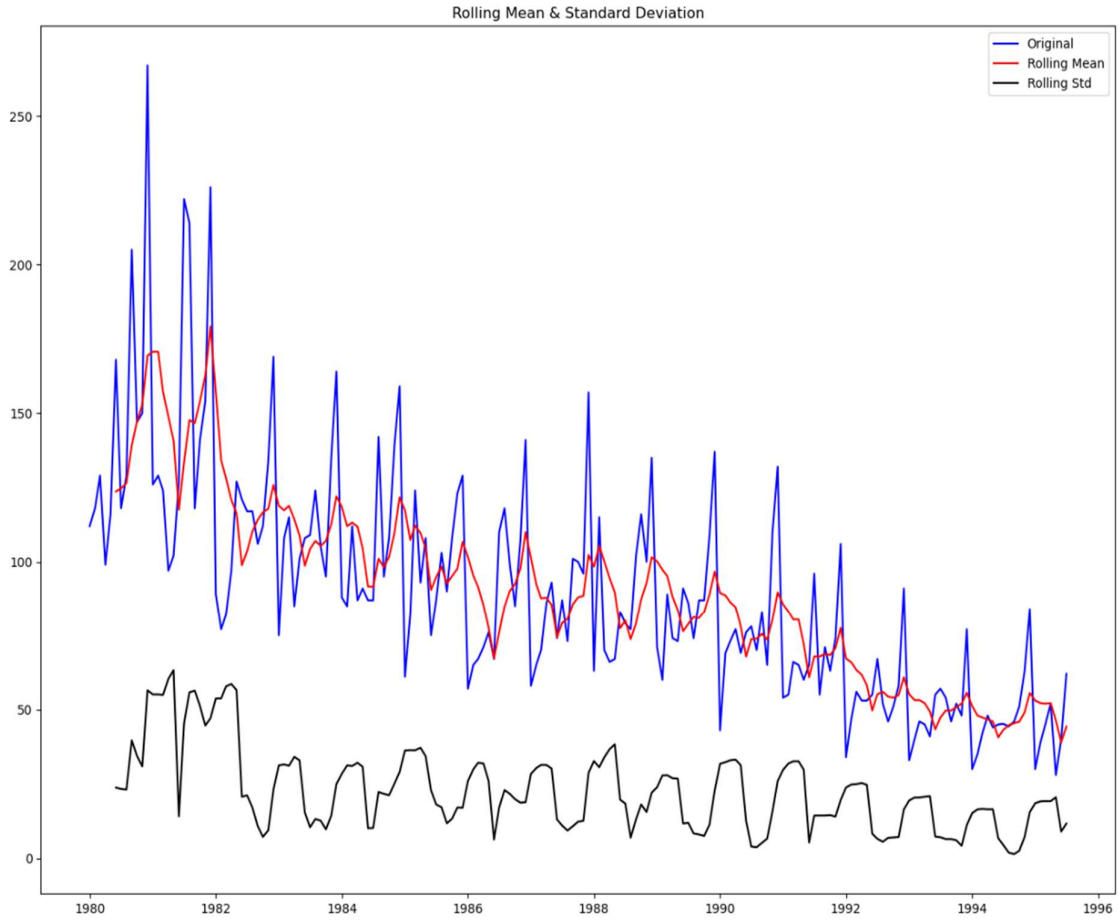

Figure 49

- Test Statistic = -1.87
- p-value= 0.34

We can see that the p-value is greater than 0.05, therefore the time series is not stationary

Sparkling wine:

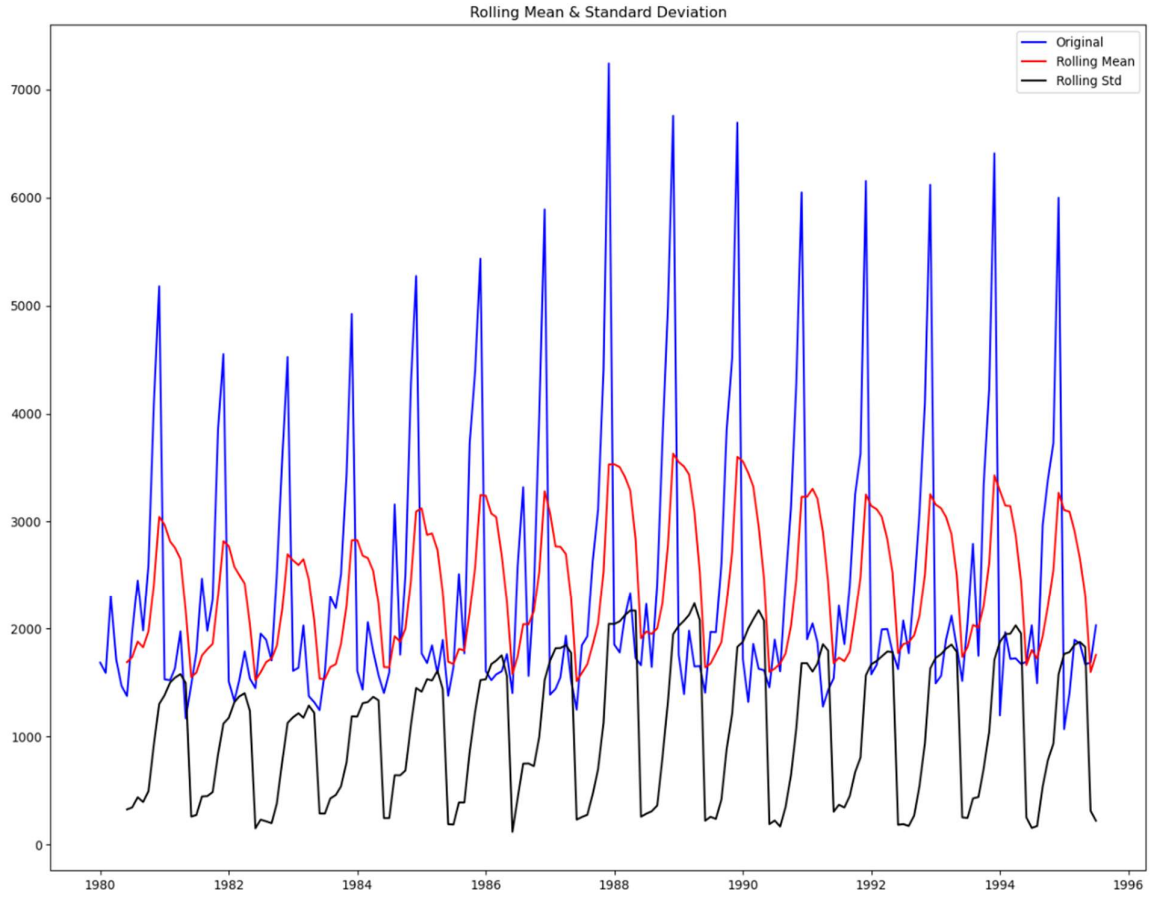

Figure 50

- DF test statistic is -1.36
- DF test p-value is 0.60

We can see that the p-value is greater than 0.05, therefore the time series is not stationary.

Making the time series stationary:

After differencing:

Rose wine:

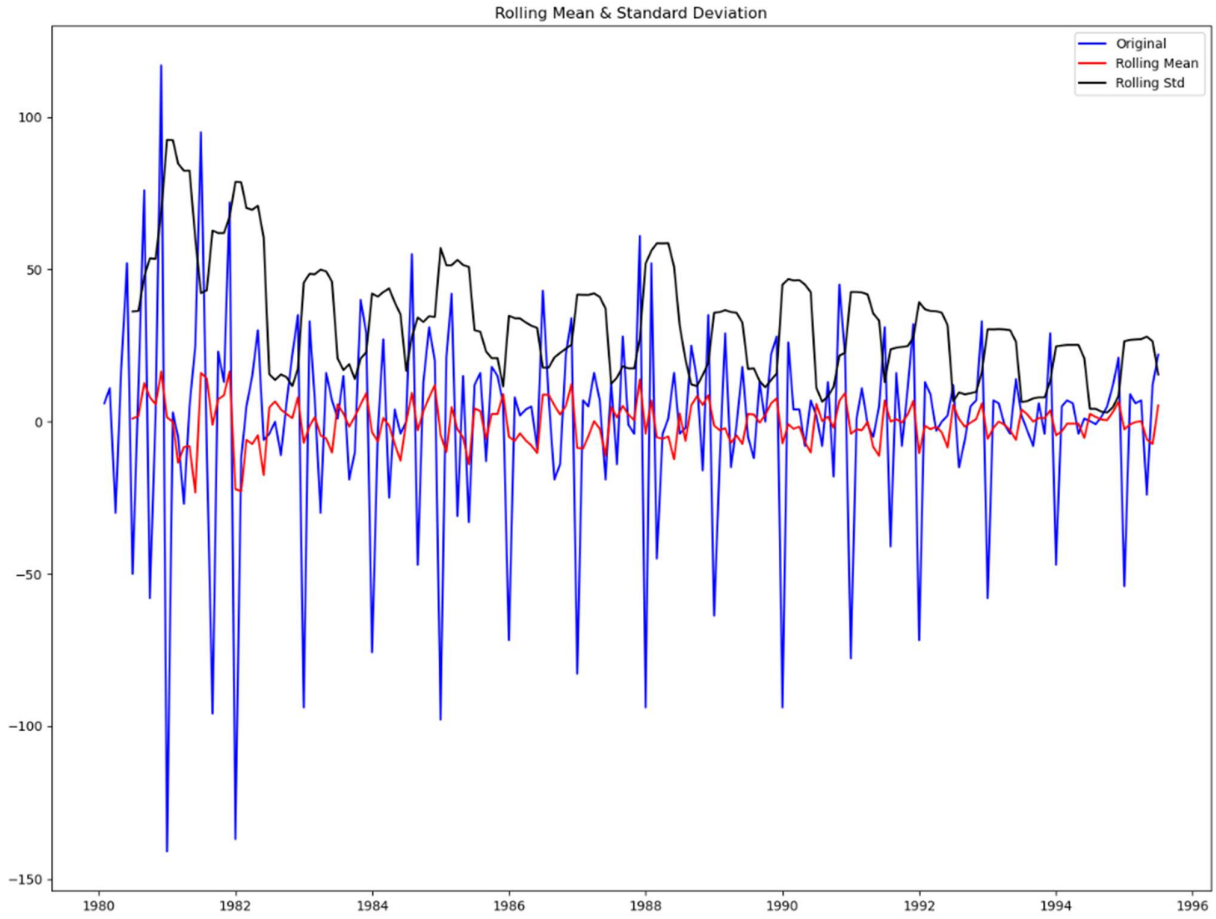

Figure 51

- Test Statistic = -8.04e+00
- p-value =1.81e-12

We can see that the first-order differencing of the time series makes it stationary, as the p-value is 1.81e-12 which is less than 0.05.

Sparkling wine:

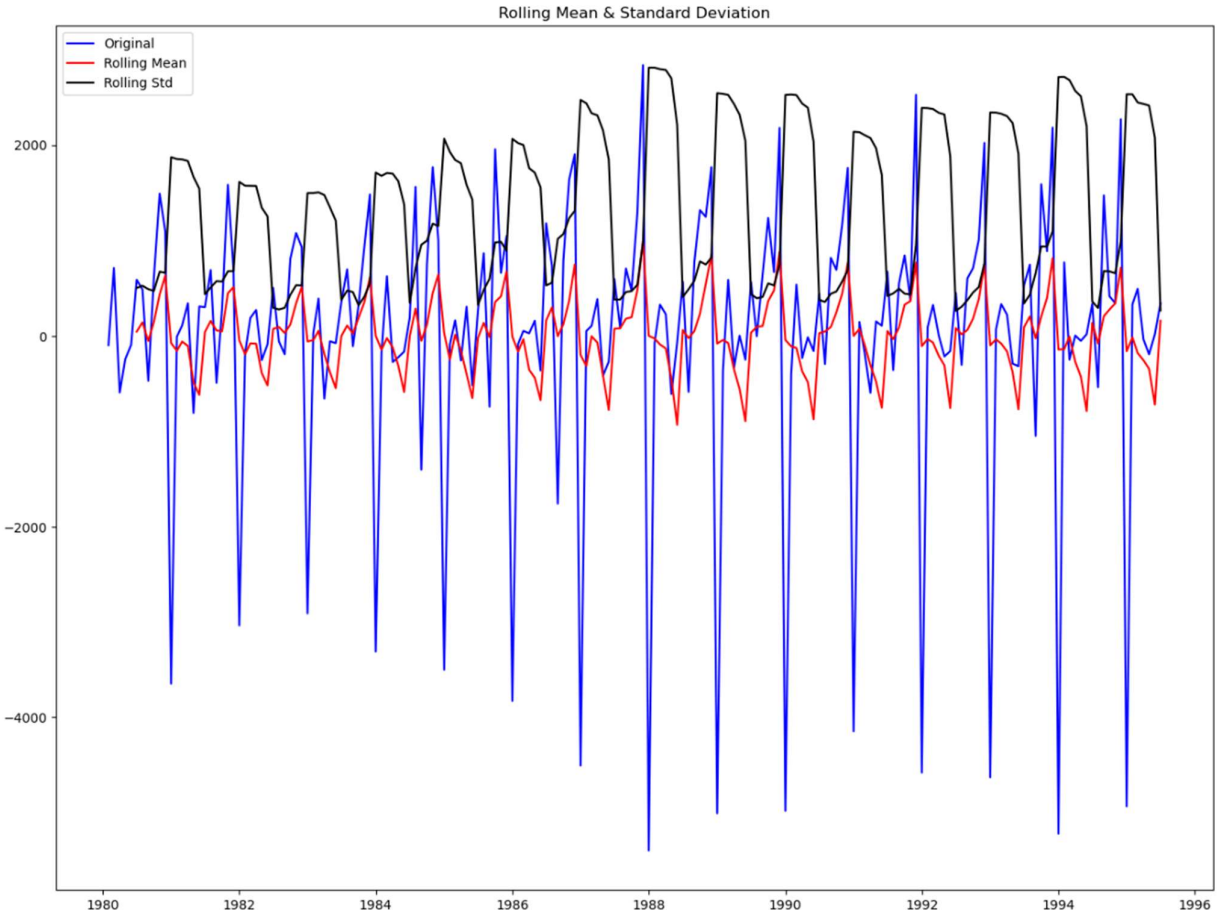

Figure 52

- Test Statistic = -45.05
- p-value =0.0

We can see that the first-order differencing of the time series makes it stationary, as the p-value is 0 which is less than 0.05.

Generate ACF & PACF Plot:

Rose wine:

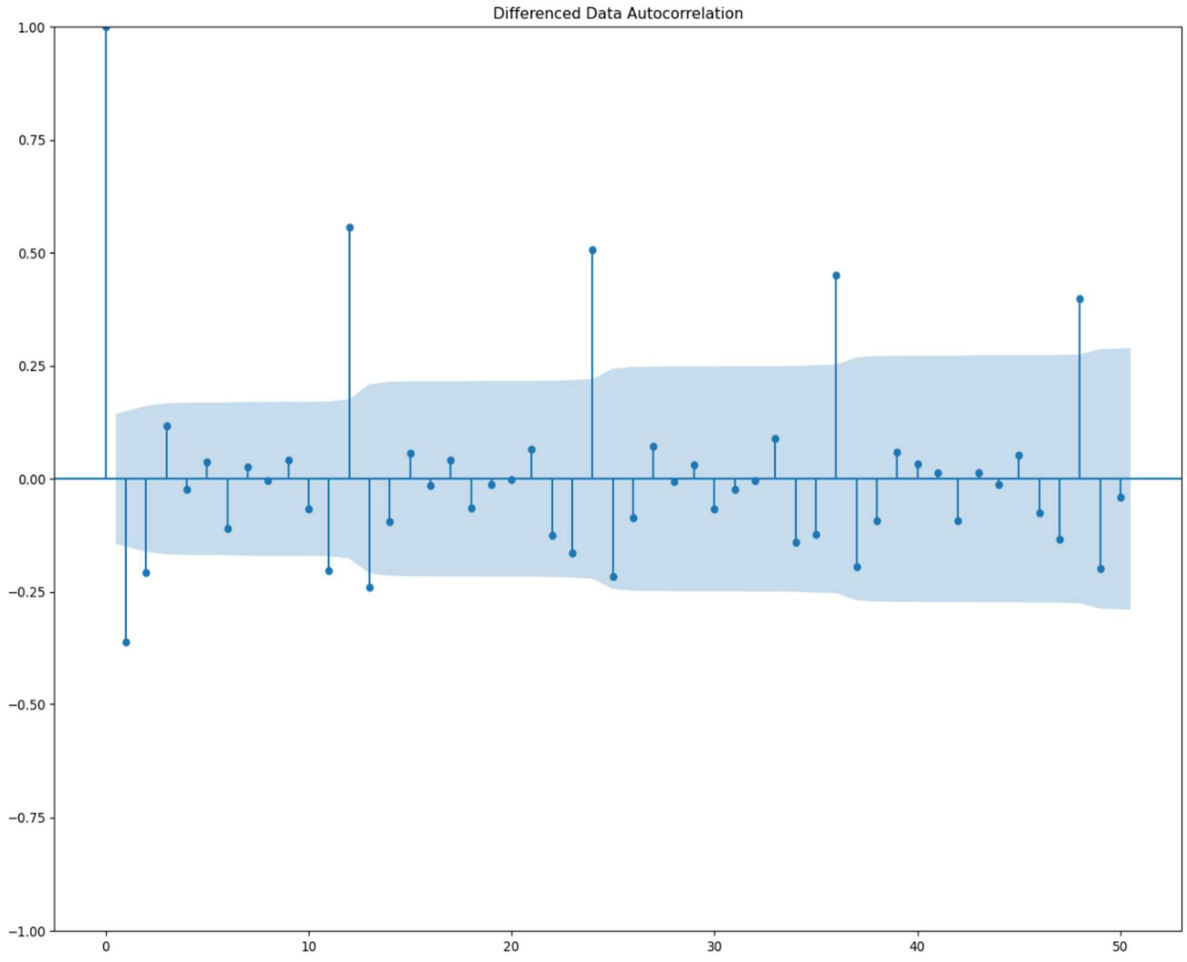

Figure 53

From the above ACF plot, we can see that there is no significant correlation in the differenced time series after lag 2. Therefore q=2.

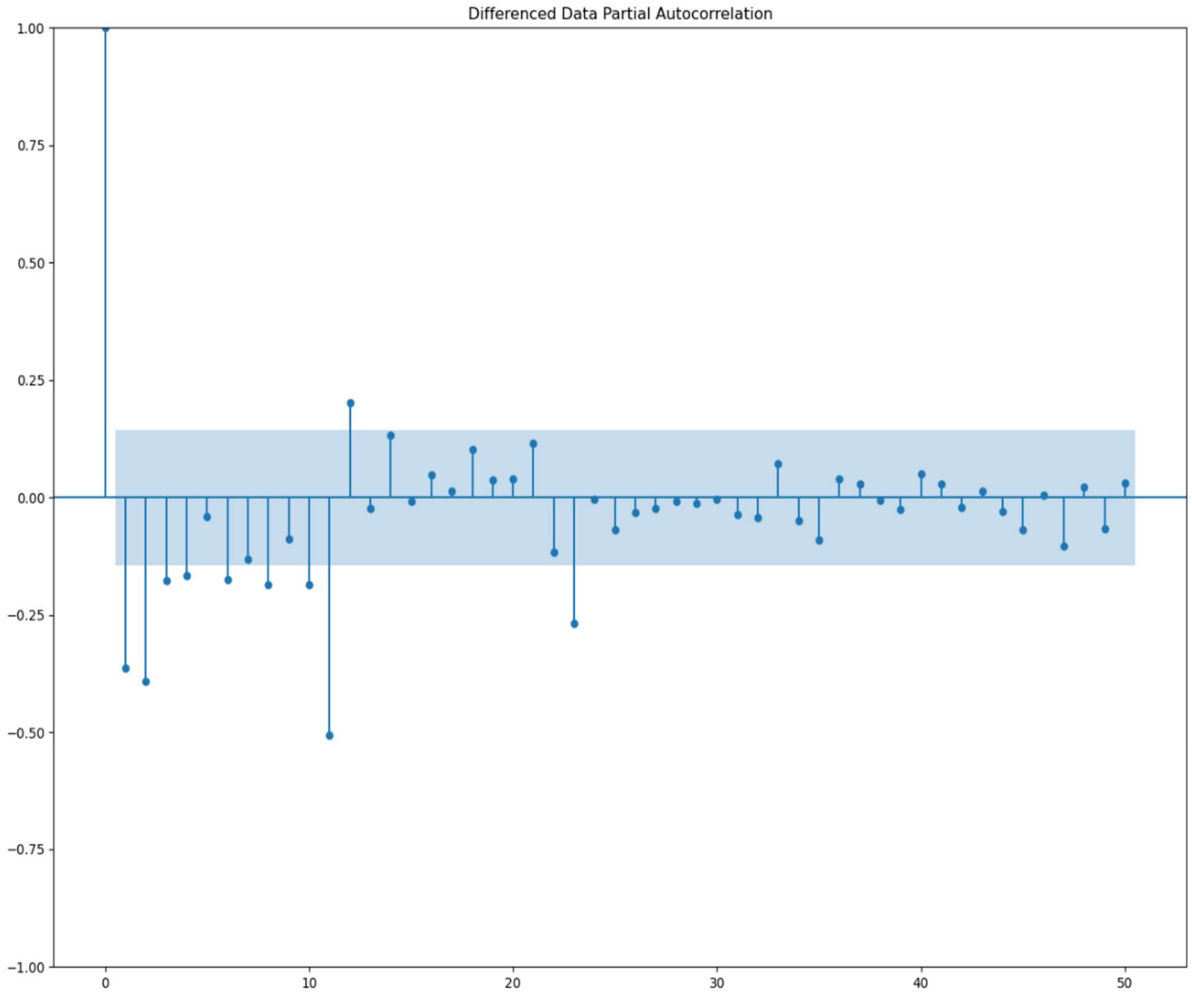

Figure 54

From the above PACF plot, we can see that there is no significant correlation in the differenced time series after lag 3. The lag 3 and lag 4 values are almost equal. Therefore p=3.

Sparkling wine:

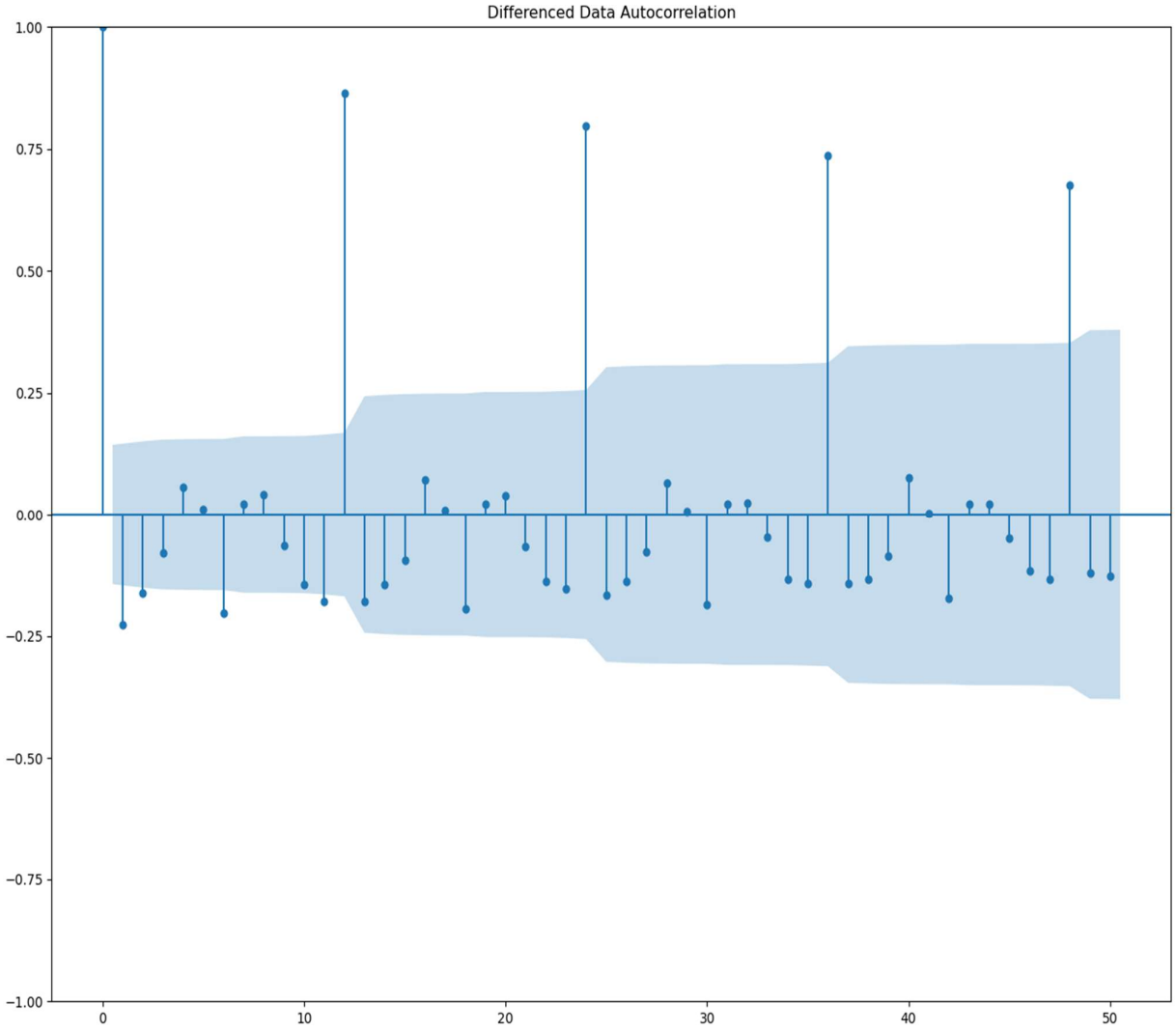

Figure 55

From the above ACF plot, we can see that there is no significant correlation in the differenced time series after lag 2. Therefore q=2.

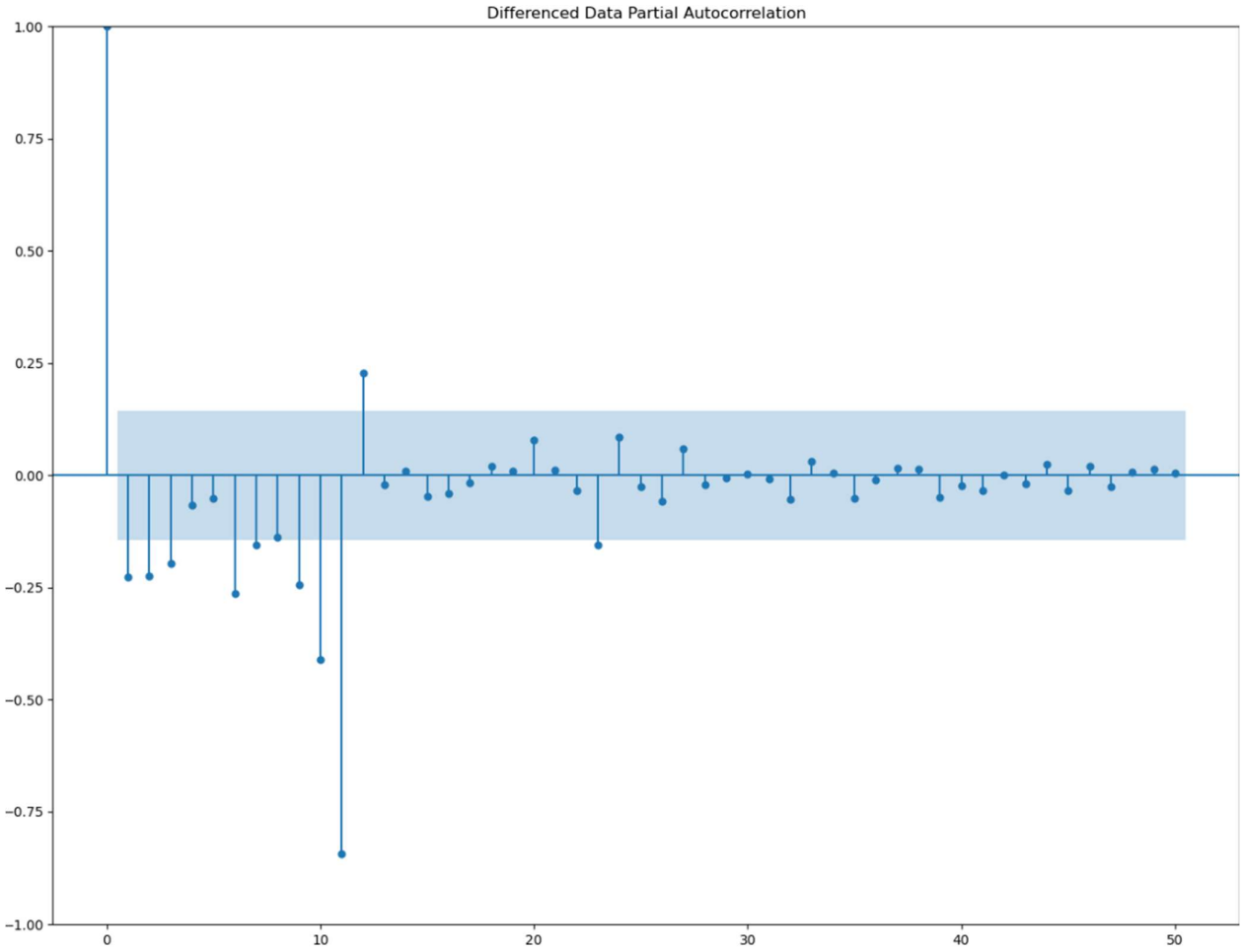

Figure 56

From the above PACF plot, we can see that there is no significant correlation in the differenced time series after lag 3. Therefore p=3.

Auto ARIMA:

Rose wine:

After running a loop for p values in the range from 0 to 3 and q in the range 0 to 2 with the value of differencing parameter d set to 1.

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
```

Figure 57

	param	AIC
5	(1, 1, 2)	1389.072404
2	(0, 1, 2)	1389.380794
4	(1, 1, 1)	1390.369694
7	(2, 1, 1)	1390.969079
8	(2, 1, 2)	1391.062173
10	(3, 1, 1)	1391.884221
1	(0, 1, 1)	1392.021422
11	(3, 1, 2)	1392.952680
9	(3, 1, 0)	1409.387611
6	(2, 1, 0)	1411.005725
3	(1, 1, 0)	1432.477510
0	(0, 1, 0)	1449.834956

Figure 58

The model with parameter 1,1,2 has the least Akaike Information Criteria (AIC).Building a model using those parameters.

Model summary:

```
==============================================================================
Dep. Variable:                      y   No. Observations:                  144
Model:                 ARIMA(1, 1, 2)   Log Likelihood                -690.536
Date:                Thu, 30 May 2024   AIC                           1389.072
Time:                        17:41:27   BIC                           1400.924
Sample:                             0   HQIC                          1393.888
                                  - 144
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4739      0.244     -1.942      0.052      -0.952       0.004
ma.L1         -0.2368      0.224     -1.057      0.290      -0.676       0.202
ma.L2         -0.6124      0.185     -3.318      0.001      -0.974      -0.251
sigma2       905.6674     80.144     11.300      0.000     748.588    1062.747
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                42.42
Prob(Q):                              0.82   Prob(JB):                         0.00
Heteroskedasticity (H):               0.35   Skew:                             0.82
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.10
===================================================================================
```

Figure 59

The coefficient of ma.L1 has a p-value of 0.290 which is greater than 0.05 which implies it is less significant.

Diagnostic plot:

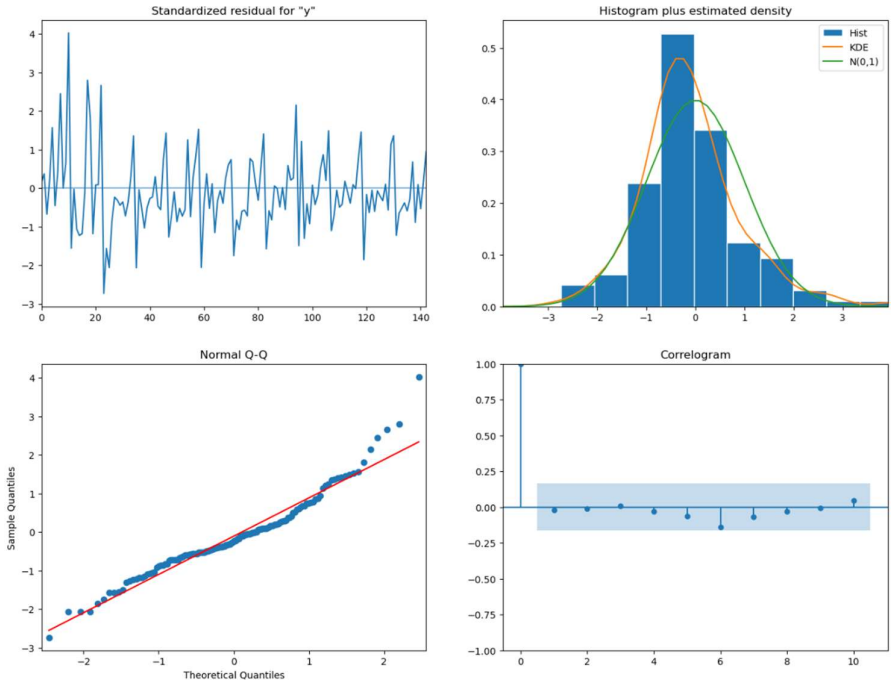

Figure 60

It is evident that the errors or residuals follow a normal distribution.

The RMSE on test data is **30.63.**

Sparkling wine:

After running a loop for p values in the range from 0 to 3 and q in the range 0 to 2 with the value of differencing parameter d set to 1.

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
```

Figure 61

	param	AIC
8	(2, 1, 2)	2419.166998
11	(3, 1, 2)	2435.454409
7	(2, 1, 1)	2438.871735
2	(0, 1, 2)	2439.471819
5	(1, 1, 2)	2439.712392
10	(3, 1, 1)	2440.427771
4	(1, 1, 1)	2440.486113
9	(3, 1, 0)	2466.346554
6	(2, 1, 0)	2468.713987
1	(0, 1, 1)	2470.903265
3	(1, 1, 0)	2474.923643
0	(0, 1, 0)	2476.745547

Figure 62

The model with parameter 2,1,2 has the least Akaike Information Criteria (AIC).Building a model using those parameters.

Model summary:

Dep. Variable:	y	No. Observations:	144
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1204.583
Date:	Thu, 30 May 2024	AIC	2419.167
Time:	17:41:01	BIC	2433.981
Sample:	0	HQIC	2425.187
	- 144		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3214	0.043	30.411	0.000	1.236	1.407
ar.L2	-0.5501	0.062	-8.906	0.000	-0.671	-0.429
ma.L1	-1.9912	0.105	-18.931	0.000	-2.197	-1.785
ma.L2	0.9995	0.106	9.473	0.000	0.793	1.206
sigma2	1.136e+06	1.88e-07	6.05e+12	0.000	1.14e+06	1.14e+06

Ljung-Box (L1) (Q):	0.07	Jarque-Bera (JB):	15.12
Prob(Q):	0.80	Prob(JB):	0.00
Heteroskedasticity (H):	2.03	Skew:	0.61
Prob(H) (two-sided):	0.02	Kurtosis:	4.03

Figure 63

Diagnostic plot:

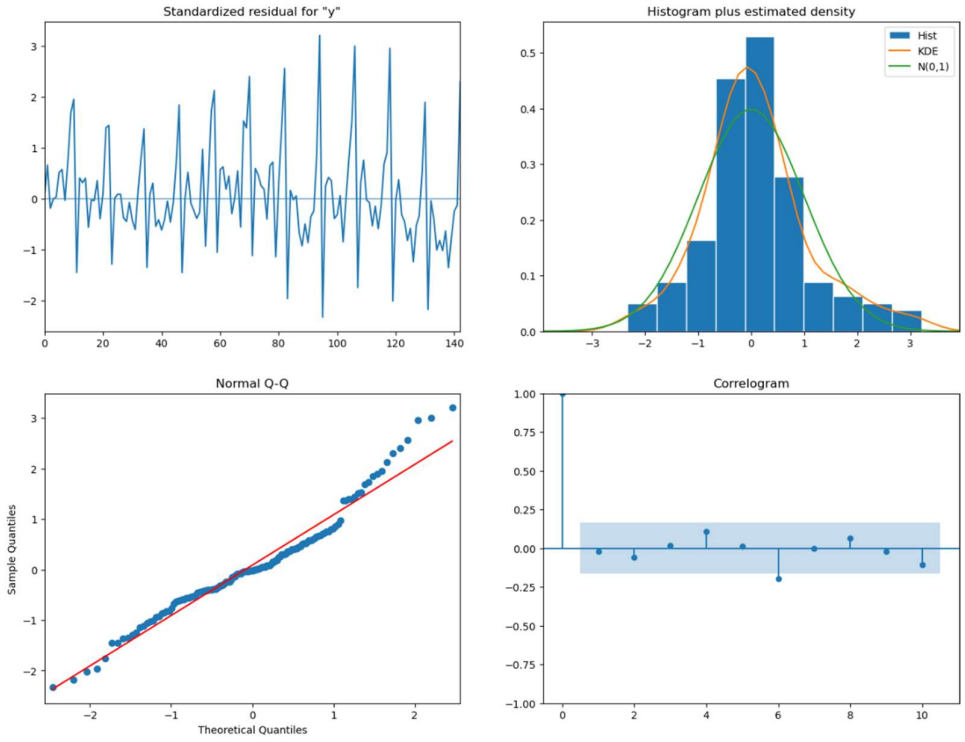

Figure 64

The residuals follow a normal distribution.

The RMSE on the test data is **1309.6.**

Manual ARIMA:

Rose wine:

Building a model using the cut-off values from the ACF and the PACF plot. The parameters are 3,1,2.

Model summary:

Dep. Variable:		Rose	No. Observations:			144
Model:		ARIMA(3, 1, 2)	Log Likelihood			-690.476
Date:		Thu, 30 May 2024	AIC			1392.953
Time:		17:41:32	BIC			1410.730
Sample:		01-01-1980	HQIC			1400.176
		- 12-01-1991				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4241	0.459	-0.924	0.356	-1.324	0.476
ar.L2	-0.0029	0.147	-0.020	0.984	-0.292	0.286
ar.L3	0.0340	0.105	0.325	0.745	-0.171	0.239
ma.L1	-0.2894	0.461	-0.628	0.530	-1.193	0.614
ma.L2	-0.5685	0.415	-1.369	0.171	-1.382	0.246
sigma2	904.9538	83.255	10.870	0.000	741.777	1068.131
Ljung-Box (L1) (Q):		0.03	Jarque-Bera (JB):			40.62
Prob(Q):		0.85	Prob(JB):			0.00
Heteroskedasticity (H):		0.35	Skew:			0.81
Prob(H) (two-sided):		0.00	Kurtosis:			5.05

Figure 65

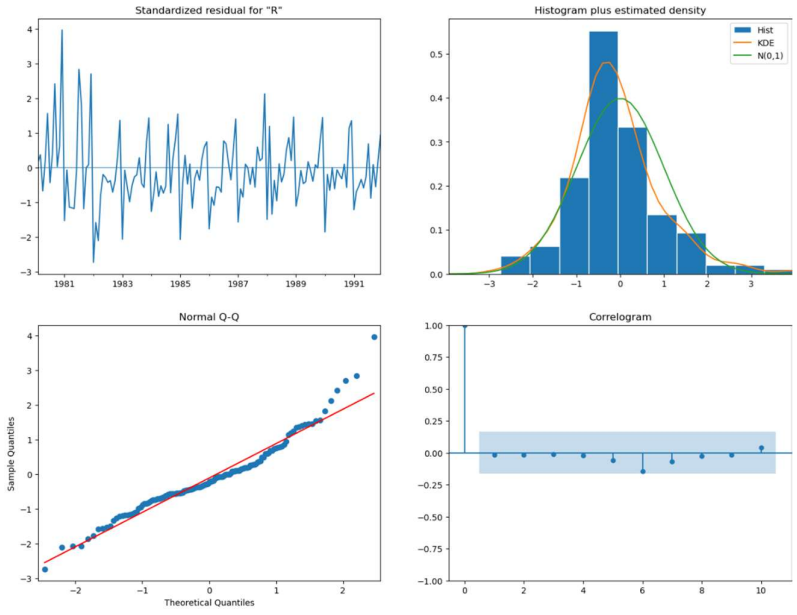

Figure 66

The RMSE on the test data is **30.75.**

Sparkling wine:

Building a model using the cut-off values from the ACF and the PACF plot. The parameters are 3,1,2.

Model summary:

```
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  144
Model:                 ARIMA(3, 1, 2)   Log Likelihood               -1211.727
Date:                Thu, 30 May 2024   AIC                           2435.454
Time:                        17:41:07   BIC                           2453.231
Sample:                    01-01-1980   HQIC                          2442.678
                         - 12-01-1991
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4392      0.028    -15.676      0.000      -0.494      -0.384
ar.L2          0.3417      0.059      5.790      0.000       0.226       0.457
ar.L3         -0.2191      0.031     -7.064      0.000      -0.280      -0.158
ma.L1      -7.914e-05      0.000     -0.257      0.797      -0.001       0.001
ma.L2         -0.9999      0.120     -8.303      0.000      -1.236      -0.764
sigma2     1.278e+06   9.44e-08   1.35e+13      0.000    1.28e+06    1.28e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                 6.78
Prob(Q):                              0.85   Prob(JB):                         0.03
Heteroskedasticity (H):               2.08   Skew:                             0.44
Prob(H) (two-sided):                  0.01   Kurtosis:                         3.60
===================================================================================
```

Figure 67

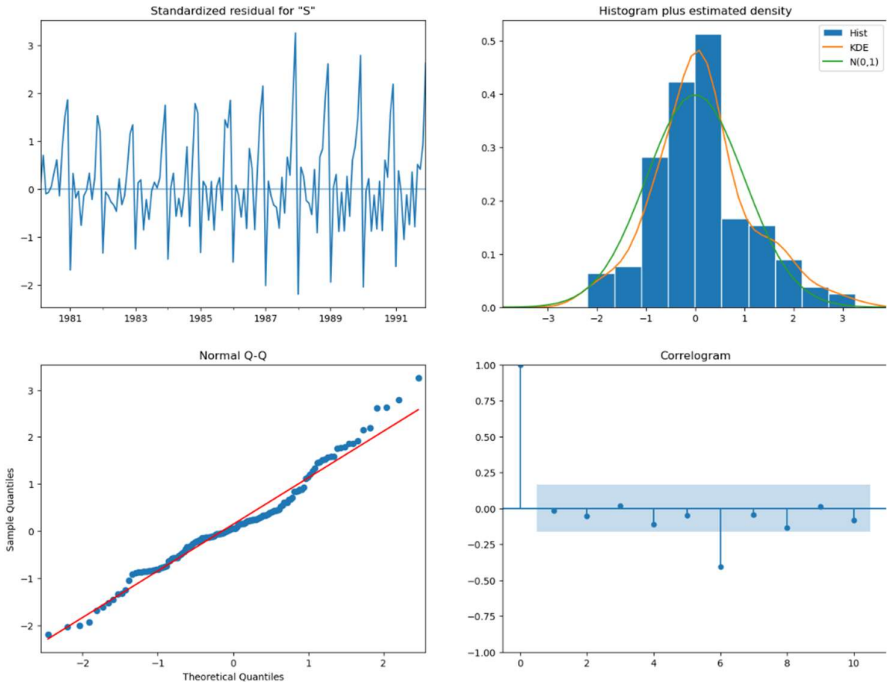

Figure 68

The RMSE on the test data is **1295.3.**

Auto SARIMA:

From the ACF plot, we can observe that there is a spike every 12 months. Therefore the seasonality factor is 12.

Rose wine:

Checking stationarity after the first order seasonal differencing.

- DF test statistic is -4.60
- DF test p-value is 0.00

Since p-value is less than 0.05 after first-order seasonal differencing, we are proceeding with D=1.

After running a loop for p values in the range from 0 to 3 and q in the range 0 to 2 with the value of differencing d and seasonal differencing parameter D set to 1.

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(0, 1, 0, 12)
Model: (1, 1, 1)(0, 1, 1, 12)
Model: (1, 1, 2)(0, 1, 2, 12)
Model: (2, 1, 0)(1, 0, 0, 12)
Model: (2, 1, 1)(1, 0, 1, 12)
Model: (2, 1, 2)(1, 0, 2, 12)
Model: (3, 1, 0)(1, 1, 0, 12)
Model: (3, 1, 1)(1, 1, 1, 12)
Model: (3, 1, 2)(1, 1, 2, 12)
```

Figure 69

	param	seasonal	AIC
262	(3, 1, 1)	(3, 1, 1, 12)	778.555044
286	(3, 1, 2)	(3, 1, 1, 12)	779.147418
263	(3, 1, 1)	(3, 1, 2, 12)	779.448586
287	(3, 1, 2)	(3, 1, 2, 12)	780.189164
190	(2, 1, 1)	(3, 1, 1, 12)	784.670223
...	...	...	...
24	(0, 1, 1)	(0, 0, 0, 12)	1373.315441
216	(3, 1, 0)	(0, 0, 0, 12)	1381.817920
144	(2, 1, 0)	(0, 0, 0, 12)	1392.720452
72	(1, 1, 0)	(0, 0, 0, 12)	1423.324811
0	(0, 1, 0)	(0, 0, 0, 12)	1440.682260

Figure 70

The model with parameter 3,1,1 and seasonal parameter 3,1,1,12 has the least Akaike Information Criteria (AIC). Building a model using those parameters.

Model summary:

```
Dep. Variable:                                 y   No. Observations:                144
Model:             SARIMAX(3, 1, 1)x(3, 1, 1, 12)   Log Likelihood                -380.278
Date:                            Thu, 30 May 2024   AIC                             778.555
Time:                                    17:54:41   BIC                             801.251
Sample:                                         0   HQIC                            787.715
                                              - 144
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0355      0.138      0.257      0.797      -0.235       0.306
ar.L2         -0.0303      0.131     -0.232      0.817      -0.287       0.226
ar.L3         -0.0603      0.115     -0.524      0.600      -0.286       0.165
ma.L1         -0.9294      0.080    -11.665      0.000      -1.086      -0.773
ar.S.L12       0.0404      0.123      0.328      0.743      -0.201       0.282
ar.S.L24      -0.0316      0.107     -0.296      0.767      -0.241       0.178
ar.S.L36      -0.0001      0.054     -0.002      0.998      -0.107       0.106
ma.S.L12      -0.7890      0.191     -4.121      0.000      -1.164      -0.414
sigma2       212.3547     39.017      5.443      0.000     135.883     288.826
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 4.20
Prob(Q):                              0.90   Prob(JB):                         0.12
Heteroskedasticity (H):               0.55   Skew:                             0.50
Prob(H) (two-sided):                  0.10   Kurtosis:                         3.28
===================================================================================
```

Figure 71

Coefficients with p values greater than 0.05 are less significant.

Diagnostic plot:

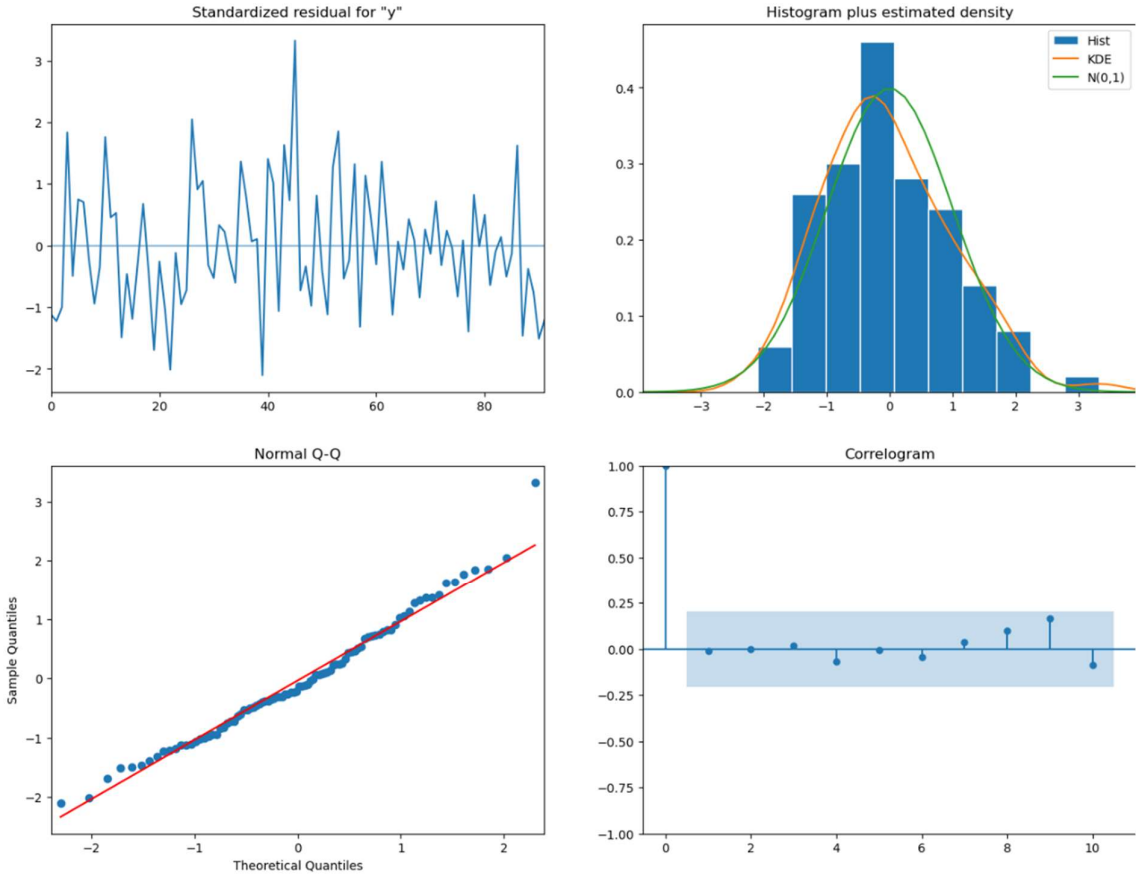

Figure 72

The residuals follow a normal distribution.

The RMSE on the test data is **11.89.**

Sparkling wine:

Checking stationarity after the first order seasonal differencing.

- DF test statistic is -5.11
- DF test p-value is 0.00

Since p-value is less than 0.05 after first-order seasonal differencing, we are proceeding with D=1.

After running a loop for p values in the range from 0 to 3 and q in the range 0 to 2 with the value of differencing d and seasonal differencing parameter D set to 1.

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 1, 1, 12)
Model: (0, 1, 2)(0, 1, 2, 12)
Model: (1, 1, 0)(1, 1, 0, 12)
Model: (1, 1, 1)(1, 1, 1, 12)
Model: (1, 1, 2)(1, 1, 2, 12)
Model: (2, 1, 0)(2, 1, 0, 12)
Model: (2, 1, 1)(2, 1, 1, 12)
Model: (2, 1, 2)(2, 1, 2, 12)
Model: (3, 1, 0)(3, 1, 0, 12)
Model: (3, 1, 1)(3, 1, 1, 12)
Model: (3, 1, 2)(3, 1, 2, 12)
```

Figure 73

	param	seasonal	AIC
129	(3, 1, 1)	(3, 1, 0, 12)	1390.803612
141	(3, 1, 2)	(3, 1, 0, 12)	1390.965334
130	(3, 1, 1)	(3, 1, 1, 12)	1392.747456
142	(3, 1, 2)	(3, 1, 1, 12)	1392.892464
143	(3, 1, 2)	(3, 1, 2, 12)	1393.614744
...	...	...	...
12	(0, 1, 1)	(0, 1, 0, 12)	1936.574300
108	(3, 1, 0)	(0, 1, 0, 12)	1942.735624
72	(2, 1, 0)	(0, 1, 0, 12)	1961.879432
36	(1, 1, 0)	(0, 1, 0, 12)	1993.765192
0	(0, 1, 0)	(0, 1, 0, 12)	2006.653982

144 rows × 3 columns

Figure 74

The model with parameters 3,1,1 and seasonal parameters 3,1,0,12 has the least Akaike Information Criteria (AIC). Building a model using those parameters.

Model summary:

```
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  144
Model:             SARIMAX(3, 1, 1)x(3, 1, [], 12)   Log Likelihood                -687.402
Date:                             Thu, 30 May 2024   AIC                           1390.804
Time:                                     17:49:43   BIC                           1410.978
Sample:                                          0   HQIC                          1398.946
                                               - 144
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2177      0.129      1.690      0.091      -0.035       0.470
ar.L2         -0.1339      0.131     -1.020      0.308      -0.391       0.123
ar.L3          0.0329      0.119      0.278      0.781      -0.199       0.265
ma.L1         -0.9575      0.054    -17.813      0.000      -1.063      -0.852
ar.S.L12      -0.3989      0.106     -3.754      0.000      -0.607      -0.191
ar.S.L24      -0.1981      0.134     -1.480      0.139      -0.460       0.064
ar.S.L36      -0.1129      0.104     -1.090      0.276      -0.316       0.090
sigma2       1.8e+05    2.5e+04      7.211      0.000    1.31e+05    2.29e+05
==========================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 20.98
Prob(Q):                              0.93   Prob(JB):                          0.00
Heteroskedasticity (H):               0.50   Skew:                              0.69
Prob(H) (two-sided):                  0.06   Kurtosis:                          4.90
==========================================================================================
```

Figure 75

Diagnostic plot:

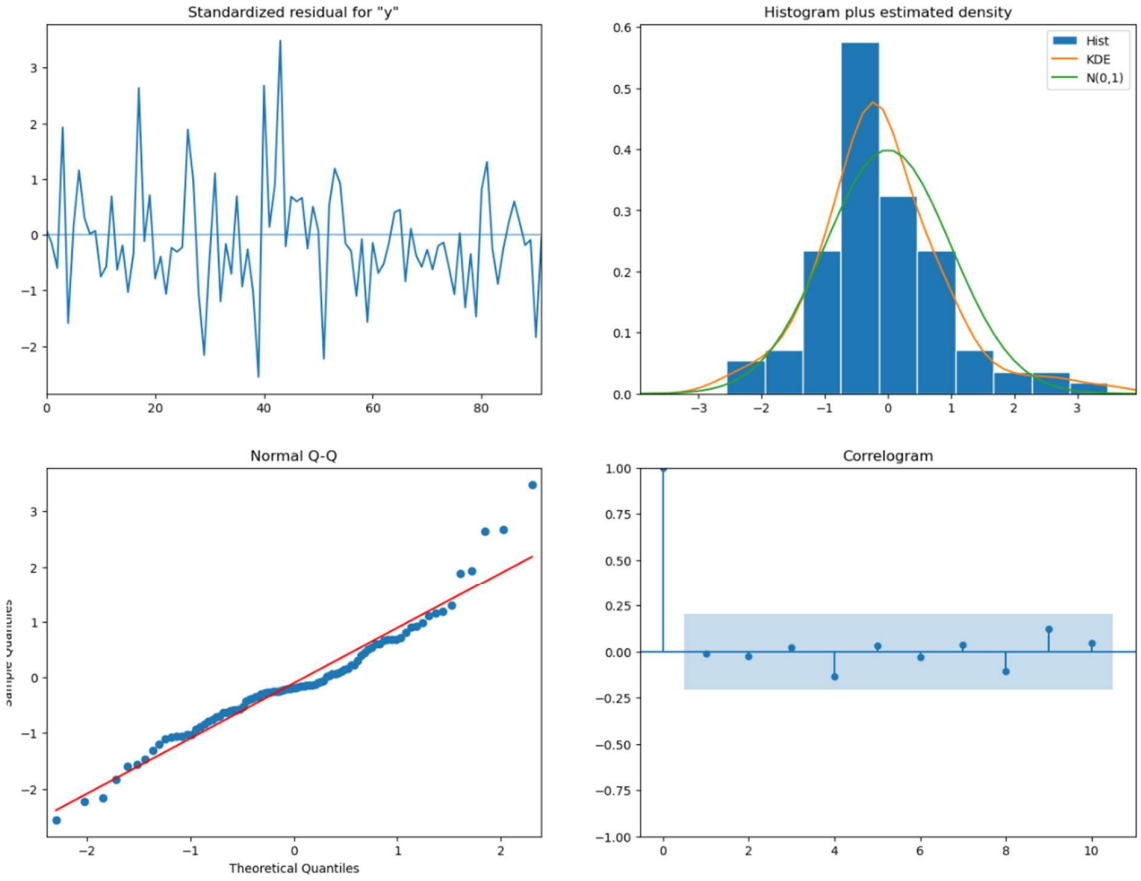

Figure 76

The RMSE on the test data is **344.66.**

Manual SARIMA:

Rose wine:

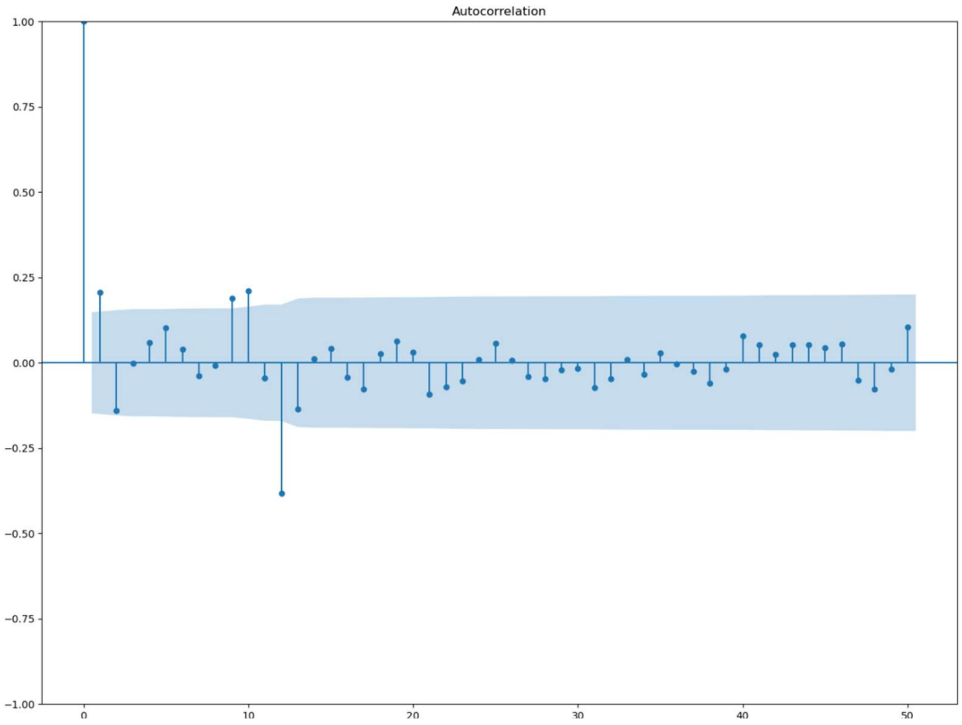

Figure 77

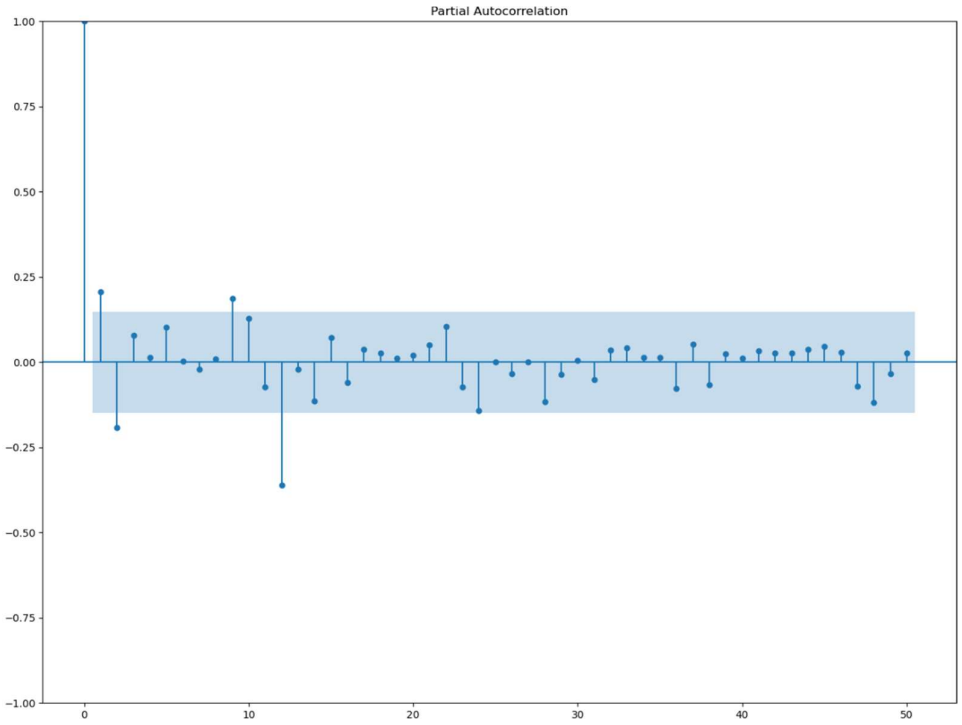

Figure 78

From the plot, there is no significant correlation after Q=1, P=2.

Building a model using the cut-off values from the ACF and the PACF plot.p=3,d=1,q=2,P=2, D=1,Q=1.

Model summary:

Dep. Variable:	y	No. Observations:	144
Model:	SARIMAX(3, 1, 2)x(2, 1, [1], 12)	Log Likelihood	-427.696
Date:	Fri, 31 May 2024	AIC	873.392
Time:	21:25:07	BIC	897.192
Sample:	0	HQIC	883.034
	- 144		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5912	0.446	1.325	0.185	-0.283	1.466
ar.L2	-0.0741	0.155	-0.477	0.633	-0.379	0.230
ar.L3	-0.1453	0.118	-1.229	0.219	-0.377	0.086
ma.L1	-1.4776	0.437	-3.383	0.001	-2.334	-0.622
ma.L2	0.5440	0.401	1.356	0.175	-0.242	1.330
ar.S.L12	0.0552	0.120	0.460	0.645	-0.180	0.290
ar.S.L24	-0.0410	0.030	-1.348	0.178	-0.101	0.019
ma.S.L12	-0.7338	0.157	-4.679	0.000	-1.041	-0.426
sigma2	203.4637	28.525	7.133	0.000	147.556	259.371

Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	5.24
Prob(Q):	0.76	Prob(JB):	0.07
Heteroskedasticity (H):	0.67	Skew:	0.43
Prob(H) (two-sided):	0.24	Kurtosis:	3.69

Figure 79

Diagnostic plot:

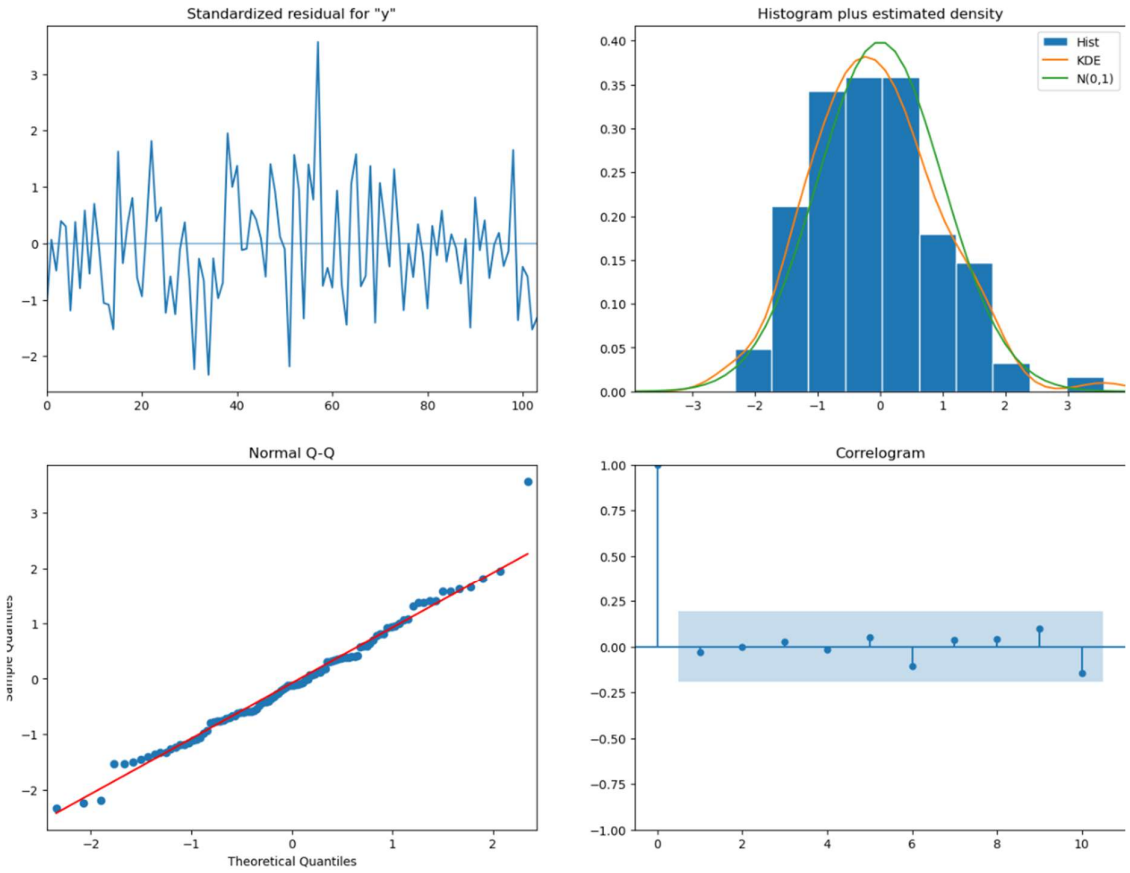

Figure 80

The RMSE on the test data is **10.08.**

Sparkling wine:

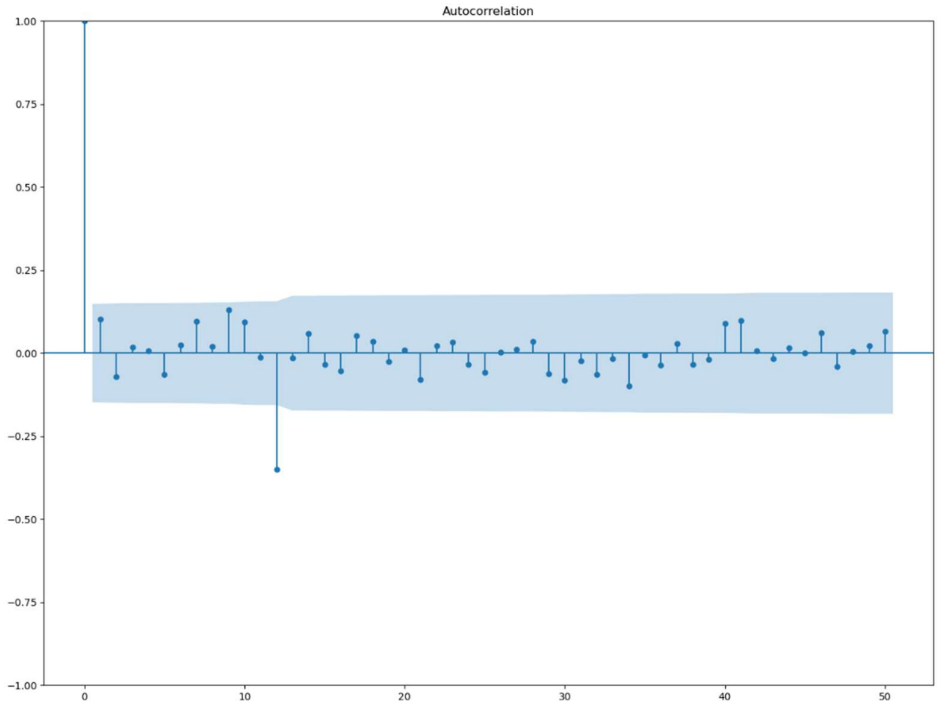

Figure 81

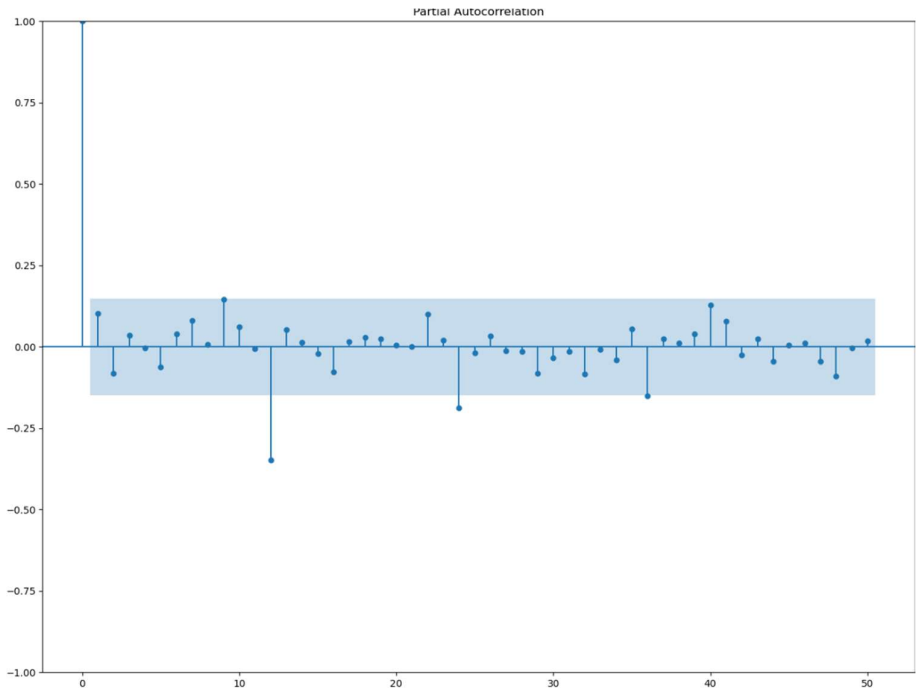

Figure 82

From the plot, there is no significant correlation after Q=0, P=0.

Building a model using the cut-off values from the ACF and the PACF plot.p=3,d=1,q=2,P=0, D=1,Q=0 and seasonality factor of 12.

Model summary:

Dep. Variable:	y	No. Observations:	144
Model:	SARIMAX(3, 1, 2)x(0, 1, [], 12)	Log Likelihood	-952.822
Date:	Fri, 31 May 2024	AIC	1917.644
Time:	21:14:13	BIC	1934.756
Sample:	0 - 144	HQIC	1924.597
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6967	0.083	-8.439	0.000	-0.859	-0.535
ar.L2	0.1080	0.081	1.334	0.182	-0.051	0.267
ar.L3	-0.0597	0.079	-0.755	0.450	-0.215	0.095
ma.L1	8.305e-06	79.866	1.04e-07	1.000	-156.535	156.535
ma.L2	-1.0000	0.112	-8.934	0.000	-1.219	-0.781
sigma2	1.624e+05	0.000	3.53e+08	0.000	1.62e+05	1.62e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	6.26
Prob(Q):	0.93	Prob(JB):	0.04
Heteroskedasticity (H):	0.98	Skew:	0.15
Prob(H) (two-sided):	0.96	Kurtosis:	4.04

Figure 83

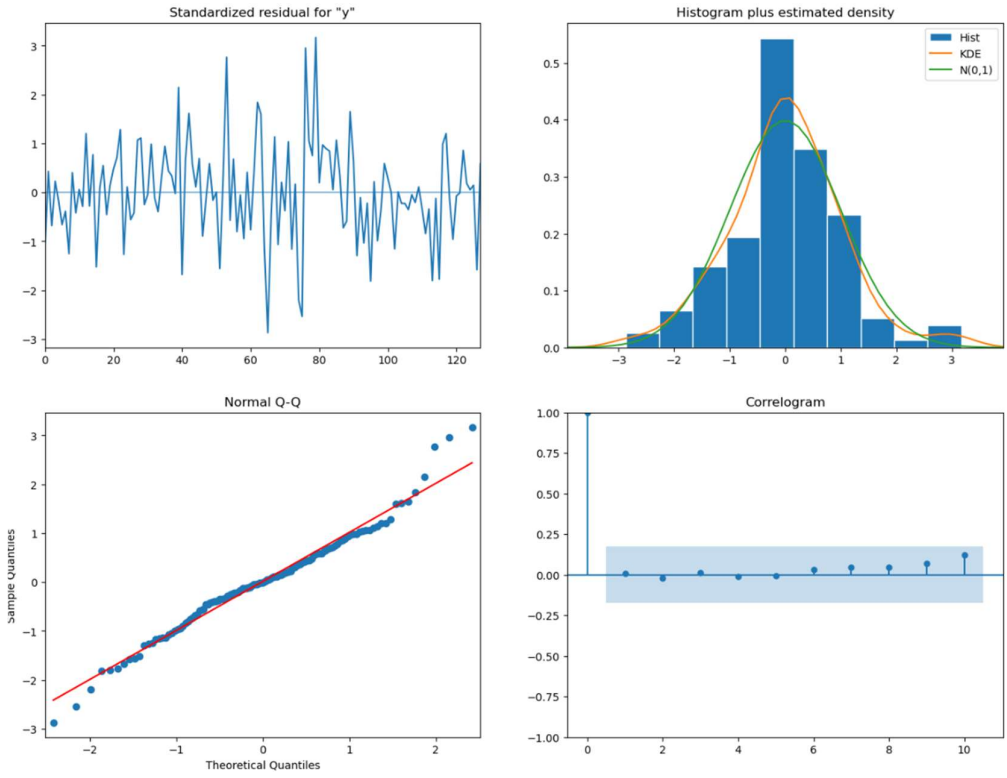

Figure 84

The RMSE on the test data is **424.8**.

Model Comparison:

Rose wine:

Model	Test RMSE
Alpha =0.09,Beta = 0.0003 and Gamma = 1.09e-06 TES Multiplicative	8.452768
Manual_SARIMA	10.08337
2pointTrailingMovingAverage	10.477689
Auto_SARIMA	11.891193
4pointTrailingMovingAverage	13.009639
Alpha =0.088,Beta = 3.79e-07 and Gamma = 3.66e-05 TES Additive	13.088792
6pointTrailingMovingAverage	13.179114
8pointTrailingMovingAverage	13.722744
10pointTrailingMovingAverage	13.856958
Alpha =1.49e-08 and Beta = 3.59e-09	14.920522
Linear Regression	14.920524
Alpha =0.099 Simple Exponential Smoothing	30.089953
Auto_ARIMA	30.634776
Manual_ARIMA	30.75278
Simple Average Model	53.979005

Table 1

Among the various models, the triple exponential smoothening with multiplicative seasonality has the least RMSE.

Sparkling wine:

Model	Test RMSE
Auto_SARIMA	344.668345
Alpha =0.076,Beta = 0.076 and Gamma =0.342 TES Multiplicative	347.346366
Alpha =0.075,Beta = 0.043 and Gamma = 0.445 TES Additive	368.571075
Manual_SARIMA	424.897543
2pointTrailingMovingAverage	834.625762
4pointTrailingMovingAverage	1169.865511
Simple Average Model	1268.683035
6pointTrailingMovingAverage	1277.869178
Manual_ARIMA	1295.371146
Auto_ARIMA	1309.634053
Alpha =0.05 Simple Exponential Smoothing	1310.262407
10pointTrailingMovingAverage	1324.529863
8pointTrailingMovingAverage	1329.140031
Linear Regression	1356.624538
Alpha =0.66 and Beta = 0.0001	4773.351788

Table 2

Among the various model,the Auto SARIMA has the least RMSE.

Rebuild the best model using the entire data:

Rose wine:

Building TES model with parameters Alpha =0.09,Beta = 0.0003 and Gamma = 1.09e-06.

Model summary:

ExponentialSmoothing Model Results

Dep. Variable:	Rose	No. Observations:	187
Model:	ExponentialSmoothing	SSE	48488.030
Optimized:	True	AIC	1071.339
Trend:	Additive	BIC	1123.037
Seasonal:	Multiplicative	AICC	1075.411
Seasonal Periods:	12	Date:	Fri, 31 May 2024
Box-Cox:	False	Time:	21:25:09
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.0995840	alpha	False
smoothing_trend	0.0003244	beta	False
smoothing_seasonal	1.0941e-06	gamma	False
initial_level	128.10997	l.0	True
initial_trend	-0.4882212	b.0	True
initial_seasons.0	0.8424334	s.0	True
initial_seasons.1	0.9564903	s.1	True
initial_seasons.2	1.0551926	s.2	True
initial_seasons.3	0.9427533	s.3	True
initial_seasons.4	1.0307352	s.4	True
initial_seasons.5	1.1166283	s.5	True
initial_seasons.6	1.2526803	s.6	True
initial_seasons.7	1.2899984	s.7	True
initial_seasons.8	1.2191305	s.8	True
initial_seasons.9	1.2042660	s.9	True
initial_seasons.10	1.3981984	s.10	True
initial_seasons.11	1.9274895	s.11	True

Figure 85

Calculating predictions for the next 12 months with 95% confidence interval:

	lower_CI	prediction	upper_ci
1995-08-01	15.467826	47.113637	78.759448
1995-09-01	12.284619	43.930430	75.576241
1995-10-01	11.161267	42.807078	74.452889
1995-11-01	17.372446	49.018257	80.664068
1995-12-01	34.987706	66.633517	98.279328

Figure 86

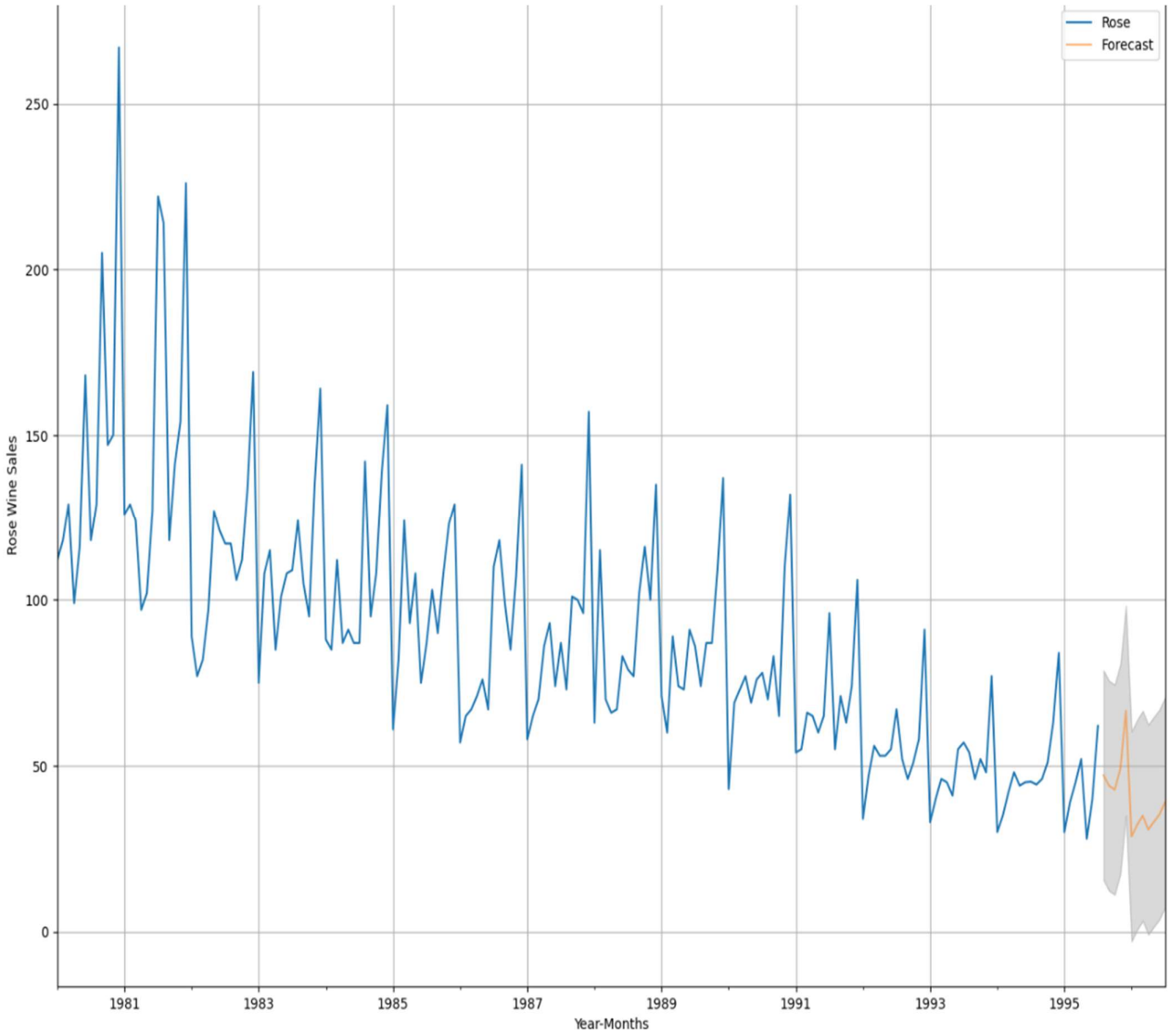

Figure 87

For sparkling wine:

Building an auto SARIMA model with parameters 3,1,1 and seasonal parameters 3,1,0,12 and forecasting for the next 12 months.

Model summary:

SARIMAX Results

Dep. Variable:	y	No. Observations:	187
Model:	SARIMAX(3, 1, 1)x(3, 1, [], 12)	Log Likelihood	-1001.463
Date:	Fri, 31 May 2024	AIC	2018.927
Time:	21:43:11	BIC	2042.169
Sample:	0 - 187	HQIC	2028.372
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1157	0.089	1.298	0.194	-0.059	0.290
ar.L2	-0.0820	0.110	-0.745	0.456	-0.298	0.134
ar.L3	0.0372	0.096	0.389	0.697	-0.150	0.225
ma.L1	-0.9636	0.035	-27.404	0.000	-1.033	-0.895
ar.S.L12	-0.5226	0.075	-6.958	0.000	-0.670	-0.375
ar.S.L24	-0.2550	0.114	-2.240	0.025	-0.478	-0.032
ar.S.L36	-0.1522	0.086	-1.779	0.075	-0.320	0.016
sigma2	1.617e+05	1.71e+04	9.458	0.000	1.28e+05	1.95e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	33.92
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	0.51	Skew:	0.67
Prob(H) (two-sided):	0.03	Kurtosis:	5.06

Figure 88

Prediction with 95% confidence interval.

Year/month	y	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-07-31		1891.069403	402.096316	1102.975105	2679.163700
1995-08-31		2473.030665	406.718116	1675.877806	3270.183525
1995-09-30		3296.362311	406.874395	2498.903150	4093.821471
1995-10-31		3857.620513	407.539448	3058.857872	4656.383154
1995-11-30		6118.451496	408.055535	5318.677344	6918.225648
1995-12-31		1198.509893	408.319079	398.219204	1998.800581
1996-01-31		1584.120086	408.613308	783.252718	2384.987454
1996-02-29		1838.038687	408.925559	1036.559320	2639.518055
1996-03-31		1846.216740	409.228600	1044.143422	2648.290058
1996-04-30		1678.517597	409.530092	875.853366	2481.181829
1996-05-31		1635.654199	409.832614	832.397035	2438.911363
1996-06-30		2013.290670	410.134836	1209.441163	2817.140178

Figure 89

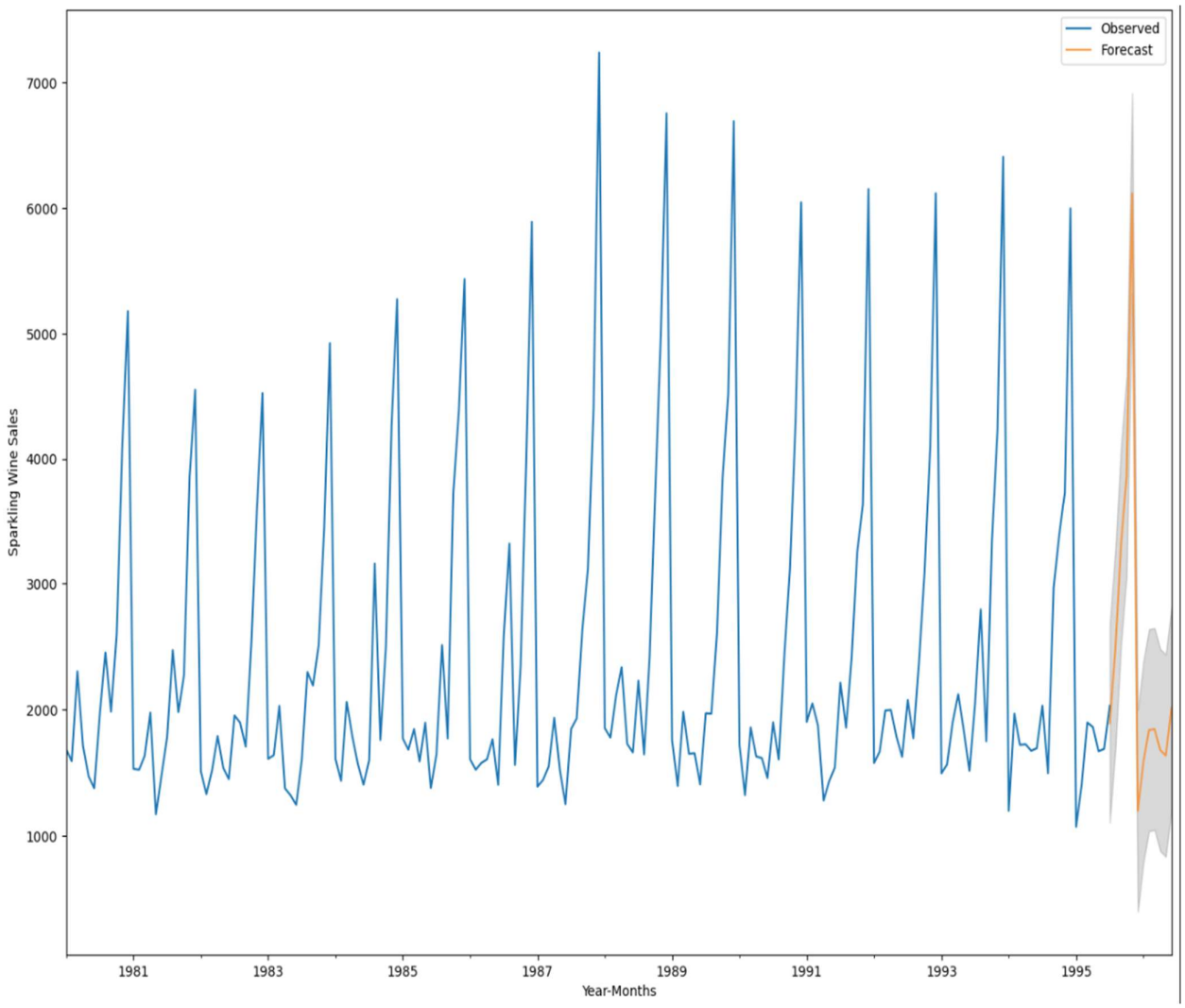
Observed
Forecast
Sparkling Wine Sales
7000
6000
5000
4000
3000
2000
1000
1981
1983
1985
1987
1989
1991
1993
1995
Year-Months

Figure 90

Actionable Insights & Recommendations:

Rose wine:

- The sales continue to follow a decreasing trend.
- Even though the sales in December are higher as compared to the other months, the sales for this most profitable month have been decreasing each year.
- The sales are highest during the year-end especially from September to December. They can introduce discounts during this period to attract new customers.
- Further they can conduct a public survey to understand the factors impacting their sales and ways to improve their product in terms of taste, packaging, selling outlets etc.
- They can introduce new marketing strategies and reach out to people during Q4(Jan, Feb, Mar), and Q1(Apr, May, Jun) as the sales are not significant during these months, and increasing the sales of these months would have a significant impact on the profit.
- To boost sales, they could introduce an alternative product at a lower cost alongside the rose wine during the start of the year.

Sparkling Wine:

- Unlike Rose wine, the sales of sparkling wine exhibit a steady trend, with occasional increases and decreases due to seasonality.
- There is a significant increase in sales towards the year-end.
- There is a 50% increase in sales between September and October. A similar trend is observed between November and December.
- As there is scope for improving profit substantially during the year-end, they can increase the productivity, number of outlets, and provide surprise shopping coupons for the festival during those seasons, which would attract more customers.
- During the off-season, the company can provide wines at a discounted price, free delivery or combo offers to improve sales.
- As December has the highest sales, they could introduce new packaging with attractive gift sets that include wines along with wine-related items like corkscrews, glasses, or other accessories wrap them up in holiday-themed packaging to make them even more appealing.
- Create promotional videos during these seasons on social media platforms like youtube, Instagram, etc.