Image courtesy: https://blog.reffascode.de/tag/machine-learning/

Time Series Forecasting Project

Business Report

Sept' 22, 2024

Authored by: Kartik Trivedi

List of Contents

List of Figures

List of Tables

Data Dictionary

Dataset 1

Column Name	Column Description	Data Type
YearMonth	Year and month of sale	object
Rose	Sales of Rose wine variant	Float64

Dataset 2

Name	Description	Data Type
YearMonth	Year and month of sale	object
Sparkling	Sales of Rose wine variant	Float64

Executive Summary

Background Information

ABC Estate Wines is a wine manufacturer that produces various distinct wine varieties. They possess sales data from the 20th century, which they aim to analyze for valuable insights and forecasts. By leveraging this data, the company seeks to improve sales performance and capitalize on emerging market opportunities, helping them maintain a competitive edge in the wine industry.

Business Objective

To stay competitive in the market, leveraging historical data to uncover meaningful insights, trends, and patterns is essential for identifying emerging opportunities and boosting sales performance. With this goal in mind, ABC Estate Wines seeks to use its historical sales data for time series forecasting, enabling them to make data-driven decisions and optimize their future sales strategies.

Problem Statement

The objective of this analysis is to examine ABC Estate Wines' sales data for various wine variants from the 20th century, utilizing data analytics and forecasting techniques to identify trends and patterns. These insights will help understand market dynamics and support informed strategic decisions, ultimately enhancing sales performance and maintaining a competitive edge in the wine industry.

Conclusion

ABC Estate Wines provided historical sales data for their Rosé and Sparkling wines. After a detailed analysis, we were able to uncover the following key insights.

Key Insights

1. In wine sales for both variants there was clear seasonality where January has the lowest sales and December has highest sales, additionally, throughout the year sales tend to consistently peaking in the month of December which coincides with the holiday season.

2. Variant specific insights:

- For Rose Wine:

1. Sales of Rosé wine have experienced a steady decline over the years. In December 1980, approximately 260 units were sold, but by 1994, this number had dropped to around 80 units, which is even lower than the lowest sales month in 1980.

- For Sparkling Wine:

- The demand for Sparkling wine has remained almost consistent for this entire time period.
- While year-on-year sales data for sparkling wine shows some fluctuations, these are likely influenced by factors beyond the general trend. A separate analysis could be conducted to

better understand the causes of these fluctuations. This would be crucial in improving the forecasting model's accuracy, especially since sparkling wine sales have consistently been higher than those of Rosé, and even more so now as the company faces a decline in Rosé sales.

3. Using the provided data, we built a forecasting model that predicts future wine sales by accounting for both trend and seasonality.

- For Rose Wine:
  After testing various models

	Test RMSE
2pointTrailingMovingAverage	11.529278
Alpha=0.0715,TripleExponentialSmoothing	14.249661
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
Alpha=1.49e-08,DoubleExponentialSmoothing	15.268944
SARIMA(0,1,2)(2,0,2,12)	26.928361
ARIMA(2,1,3)	36.817150
Alpha=0.1236,SimpleExponentialSmoothing	37.592212
RegressionOnTime	51.433312
SimpleAverage_Forecast	53.460570

Table1: Model Comparison

we selected the 2-point moving average model as the best-performing model for Rosé wine as it had the lowest RMSE value and forecasted for next 12 months.

	Rose	lower_CI	upper_CI
1995-08-01	51.000000	16.705127	85.294873
1995-09-01	56.500000	22.205127	90.794873
1995-10-01	53.750000	19.455127	88.044873
1995-11-01	55.125000	20.830127	89.419873
1995-12-01	54.437500	20.142627	88.732373
1996-01-01	54.781250	20.486377	89.076123
1996-02-01	54.609375	20.314502	88.904248
1996-03-01	54.695312	20.400439	88.990186
1996-04-01	54.652344	20.357471	88.947217
1996-05-01	54.673828	20.378955	88.968701
1996-06-01	54.663086	20.368213	88.957959
1996-07-01	54.668457	20.373584	88.963330

Table2: Confidence Interval

The plot below illustrates the forecasted sales of the above table represented by the orange line, for the next 12 months along with the historic sales data. The shaded region around it indicates the range within which sales are expected to fall with 95% confidence.

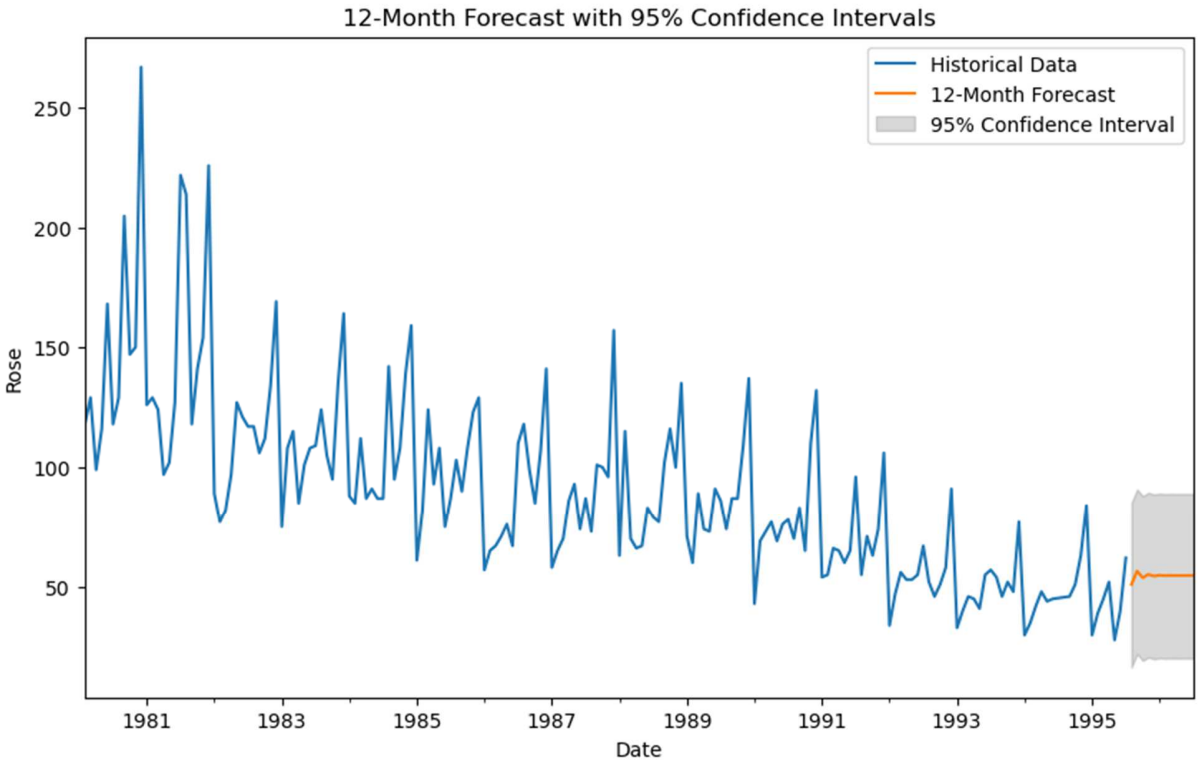


Figure1: Forecast Plot

- For Sparkling Wine:
  After testing various models

	Test RMSE
Alpha=0.111,TripleExponentialSmoothing	378.951023
SARIMA(3,1,3)(3,0,0,12)	611.473194
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
SimpleAverage_Forecast	1275.081804
RegressionOnTime	1275.867052
6pointTrailingMovingAverage	1283.927428
ARIMA(2,1,2)	1299.979665
Alpha=0.0395,SimpleExponentialSmoothing	1304.927405
9pointTrailingMovingAverage	1346.278315
Alpha=0.665,DoubleExponentialSmoothing	5291.879833

Table3: Model Comparison

we selected the Triple Exponential Smoothing Model (Holt-Winter's Linear Method) for Sparkling wine as it had the lowest RMSE value and forecasted for next 12 months.

	prediction	lower_CI	upper_CI
1995-08-31	1860.790317	1084.432878	2788.624368
1995-09-30	2470.931957	1694.574517	3398.766008
1995-10-31	3200.128073	2423.770633	4127.962124
1995-11-30	3806.537212	3030.179772	4734.371263
1995-12-31	5967.647889	5191.290450	6895.481940
1996-01-31	1224.569165	448.211726	2152.403216
1996-02-29	1600.114618	823.757178	2527.948669
1996-03-31	1861.874819	1085.517379	2789.708870
1996-04-30	1845.012697	1068.655258	2772.846748
1996-05-31	1681.565864	905.208425	2609.399915
1996-06-30	1635.125010	858.767571	2562.959061
1996-07-31	1993.194193	1216.836753	2921.028244

Table4: Confidence Interval

The plot below illustrates the forecasted sales of the above table represented by the orange line, for the next 12 months along with the historic sales data. The shaded region around it indicates the range within which sales are expected to fall with 95% confidence.

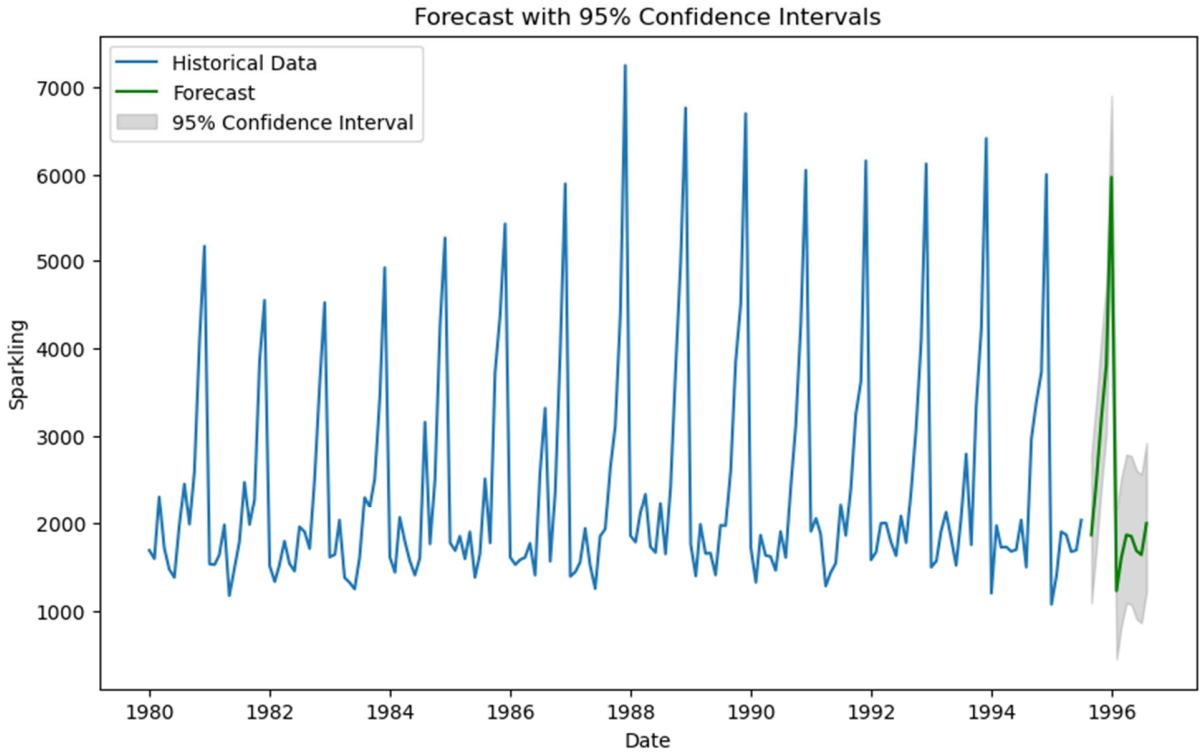


Figure2: Forecast Plot

Business Recommendations

1. The demand shows an increase from January to December, allowing ABC Wine Estate to manage inventory accordingly. In particular, demand rises significantly in December, making advanced planning and maintaining adequate stock level is crucial. This proactive approach will enable the company to fully capitalize on the peak sales season, maximizing revenue and minimizing the risk of stockouts.

2. The Rose Wine variant has experienced a steady decline in sales. To maintain a competitive edge in the wine industry, ABC Estate Wine is strongly advised to investigate the reasons behind this decline. It is essential to determine whether the decline is a broader industry trend or specific to the company. Based on these insights, necessary corrective actions should be taken to address the issue and enhance future performance.

3. Data for two wine variants were provided, showing that while sales for one variant have seen a steady decline, the other has remained relatively unchanged over a span of 15 years. This indicates that business growth is at best stagnant for these specific variants. If ABC Estate Wines produces other wine variants, it is recommended that similar analysis projects be conducted for those as well. This will enable the company to better understand its overall market position and plan its future course of action effectively.

Detailed Analysis

1.1 Background Information

ABC Estate Wines is a wine manufacturer that produces various distinct wine varieties. They possess sales data from the 20th century, which they aim to analyze for valuable insights and forecasts. By leveraging this data, the company seeks to improve sales performance and capitalize on emerging market opportunities, helping them maintain a competitive edge in the wine industry.

1.2 Business Objective

To stay competitive in the market, leveraging historical data to uncover meaningful insights, trends, and patterns is essential for identifying emerging opportunities and boosting sales performance. With this goal in mind, ABC Estate Wines seeks to use its historical sales data for time series forecasting, enabling them to make data-driven decisions and optimize their future sales strategies.

1.3 Problem Statement

The objective of this analysis is to examine ABC Estate Wines' sales data for various wine variants from the 20th century, utilizing data analytics and forecasting techniques to identify trends and patterns. These insights will help understand market dynamics and support informed strategic decisions, ultimately enhancing sales performance and maintaining a competitive edge in the wine industry.

1.4 METHODOLOGY

Import the libraries – Load the data – Check the structure of the data – Check the types of the data – Convert data to a time series data – Check for and treat (if needed) missing values – Check the statistical summary – Plot the data – Exploratory Data Analysis – Data Decomposition – Data Splitting – Model Building with Original Data – Predict values – Evaluate model – Check for Stationarity – Convert Data to Stationary – Model Building with Stationary Data – Predict values – Evaluate model – Compare model – Re-build Best Model for Entire Data – Forecast for next 12 months – Conclusion.

Key Points

1. **Data Collection**: ABC Estate Wines provided historical sales data for various types of wines, covering the period from 1980 to 1995.
2. **Data Cleaning and Pre-processing:** Dataset was checked for duplicates, missing values, bad data and outliers. There were missing values found for Rose wine data which were imputed using interpolate with method as 'linear'.
3. **Exploratory Data Analysis:** Time series data was analyzed using boxplot and different kinds of line plot to understand trend and seasonality in the data.

4. **Visualization Techniques:** In the report we have used boxplot, line plot, stem plot and histogram for analyzing time series data.
5. **Tools and Software:** We have carried out the analysis using programming language python on Jupyter notebook. For this analysis Python libraries Numpy, Pandas, Matplotlib, Seaborn, statsmodels and itertools were used.

1.5 Rose Wine Data

1.5.1 Data Overview

1. **Data Description:** Dataset has 187 rows and 2 columns.

```
shape of the dataset
------------------------------------------------------------------------------

(187, 1)
```

Table 5: Dataset Shape

2. **Dataset Information:** Of the 2 columns in the dataset, 1 is object type and 1 is float 64 types.

```
information of features
------------------------------------------------------------------------------
<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
```

Table 6: Dataset Information

3. **Missing Value Check:** There were no missing values in the dataset.

```
Number of rows with missing values:
------------------------------------------------------------------------------

Rose    2
dtype: int64
```

Table 7: Missing values information

4. **Duplicate Values:** Data was checked for duplicate values and no duplicates were found

```
checking for duplicates
--------------------------------------------------------------------------------
number of dupliacte rows: 0
```

Table 8: Data Duplicates

5. **Statistical Summary:**

```
statistical summary
--------------------------------------------------------------------------------
```

	count	mean	std	min	25%	50%	75%	max
Rose	185.0	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0

Table 9: Statistical Summary

Key observations

1. Dataset has 187 rows and 2 columns in which YearMonth column has object type data which should be datetime, we will convert it into datetime data time and set it as index to change data into a time series data.
2. There are missing values in the data which we will treat during pre-processing.

1.5.2 Exploratory Data Analysis

Plotting Data

With Missing Values

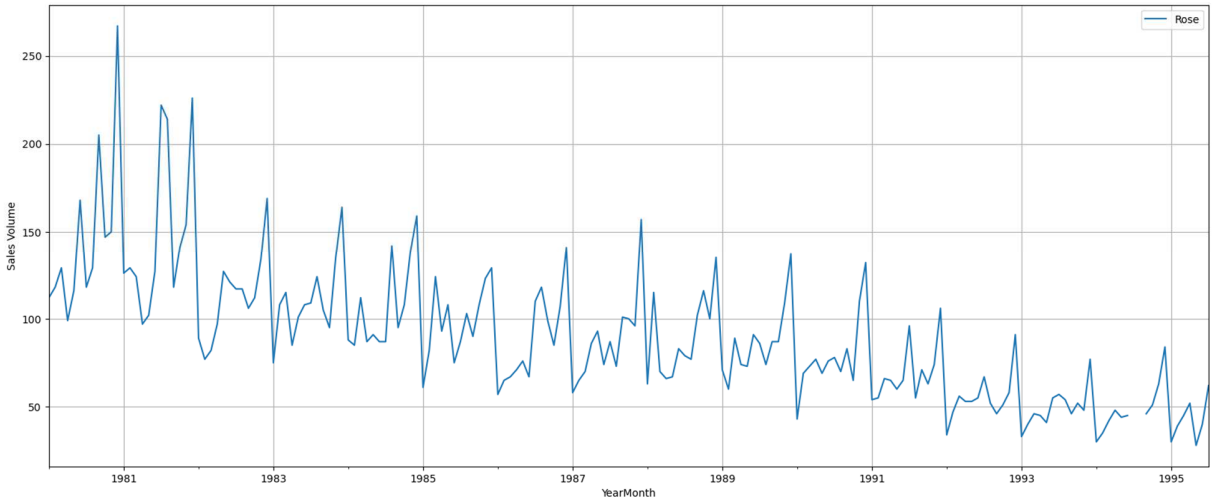


Figure3: Timeseries plot

After treating missing values

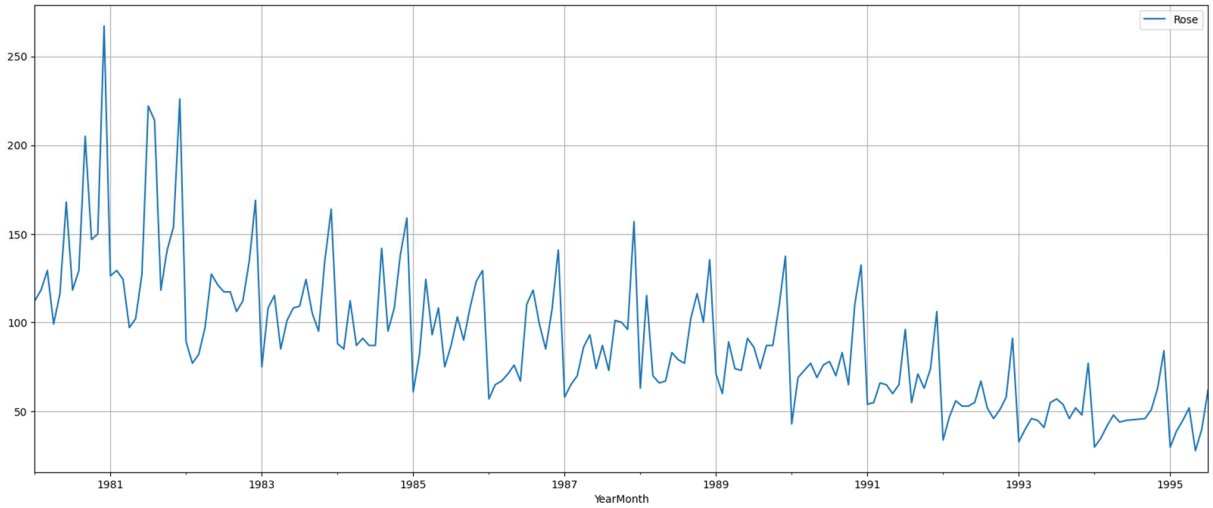


Figure4: Timeseries plot

Boxplot by Year

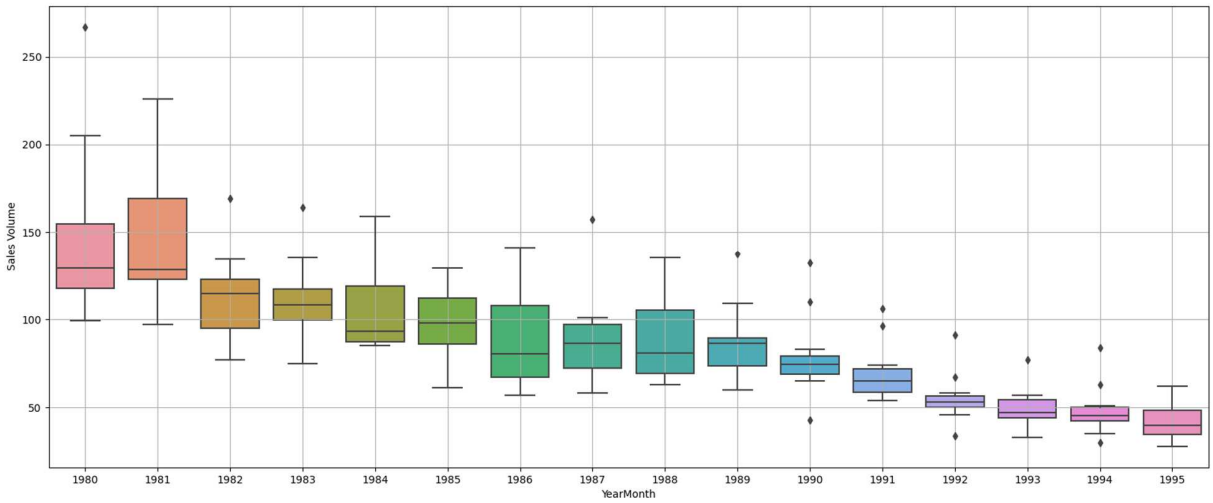


Figure5: Boxplot by year

Boxplot by Month

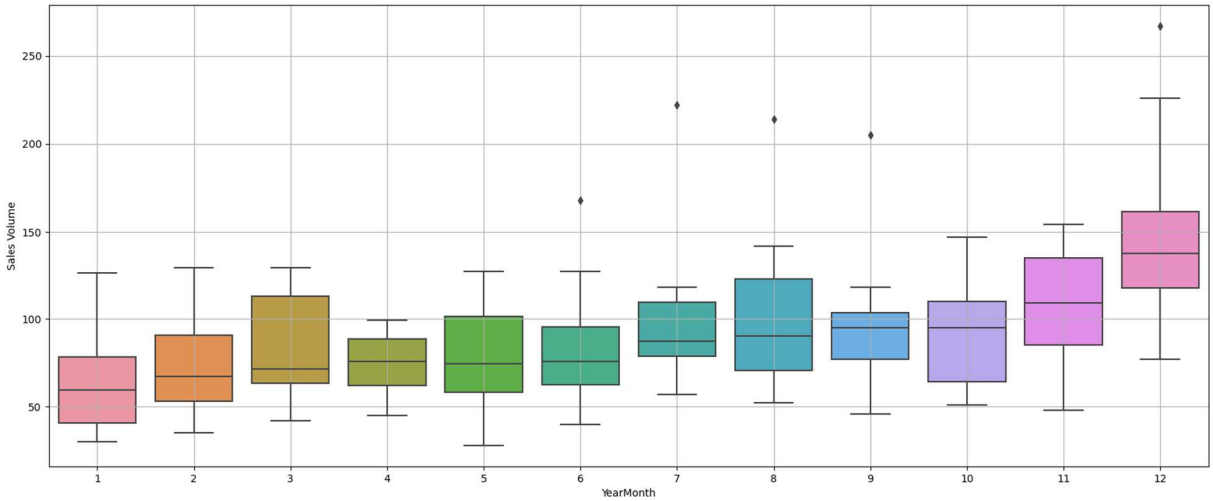


Figure6: Boxplot by month

Time Series Monthplot

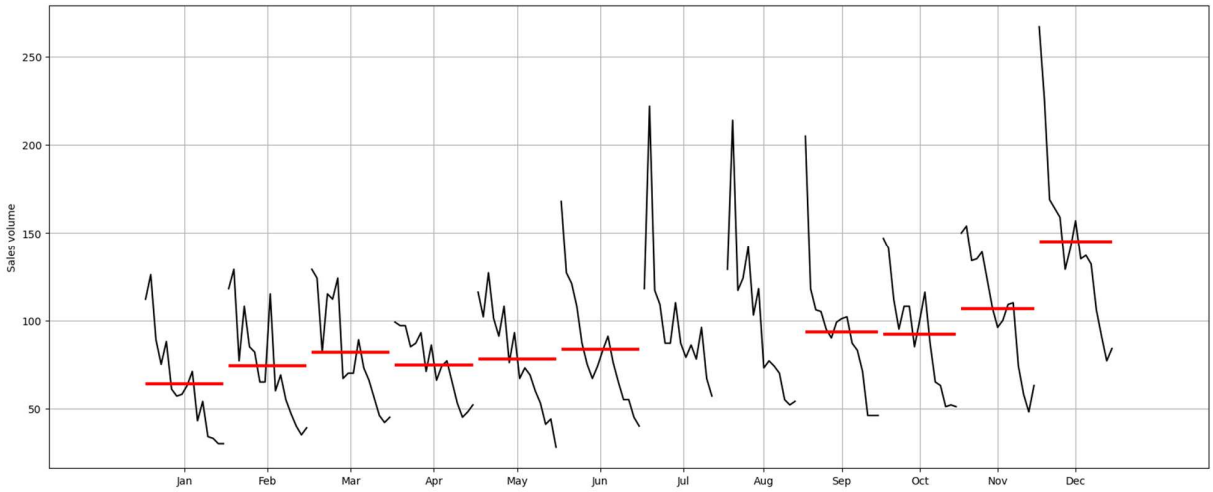


Figure7: Timeseries monthplot

Monthly Sales across Years

Pivot table across month and year

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.0	129.0	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.0	214.0	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.0	117.0	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.0	124.0	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.0	142.0	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.0	103.0	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.0	118.0	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.0	73.0	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.0	77.0	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.0	74.0	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.0	70.0	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.0	55.0	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.0	52.0	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.0	54.0	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	NaN	NaN	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.0	NaN	NaN	NaN	NaN	NaN

Table 10: Pivot table across month and year

Plot across month and year

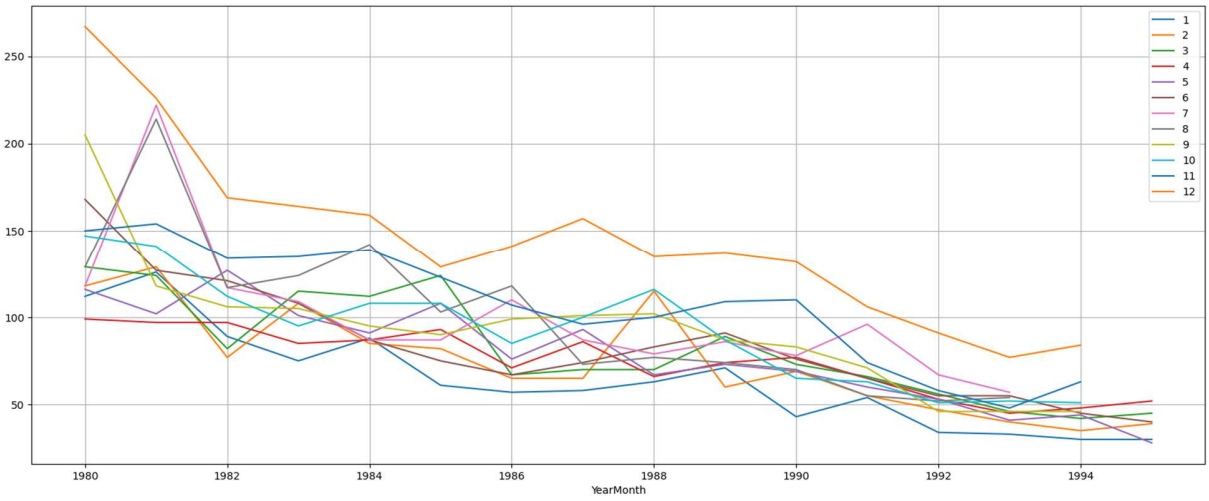


Figure8: Plot across month and year

Yearly Sales across Months

Pivot table across year and month

YearMonth	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
YearMonth																
1	112.0	126.0	89.0	75.0	88.0	61.0	57.0	58.0	63.0	71.0	43.0	54.0	34.0	33.0	30.0	30.0
2	118.0	129.0	77.0	108.0	85.0	82.0	65.0	65.0	115.0	60.0	69.0	55.0	47.0	40.0	35.0	39.0
3	129.0	124.0	82.0	115.0	112.0	124.0	67.0	70.0	70.0	89.0	73.0	66.0	56.0	46.0	42.0	45.0
4	99.0	97.0	97.0	85.0	87.0	93.0	71.0	86.0	66.0	74.0	77.0	65.0	53.0	45.0	48.0	52.0
5	116.0	102.0	127.0	101.0	91.0	108.0	76.0	93.0	67.0	73.0	69.0	60.0	53.0	41.0	44.0	28.0
6	168.0	127.0	121.0	108.0	87.0	75.0	67.0	74.0	83.0	91.0	76.0	65.0	55.0	55.0	45.0	40.0
7	118.0	222.0	117.0	109.0	87.0	87.0	110.0	87.0	79.0	86.0	78.0	96.0	67.0	57.0	NaN	62.0
8	129.0	214.0	117.0	124.0	142.0	103.0	118.0	73.0	77.0	74.0	70.0	55.0	52.0	54.0	NaN	NaN
9	205.0	118.0	106.0	105.0	95.0	90.0	99.0	101.0	102.0	87.0	83.0	71.0	46.0	46.0	46.0	NaN
10	147.0	141.0	112.0	95.0	108.0	108.0	85.0	100.0	116.0	87.0	65.0	63.0	51.0	52.0	51.0	NaN
11	150.0	154.0	134.0	135.0	139.0	123.0	107.0	96.0	100.0	109.0	110.0	74.0	58.0	48.0	63.0	NaN
12	267.0	226.0	169.0	164.0	159.0	129.0	141.0	157.0	135.0	137.0	132.0	106.0	91.0	77.0	84.0	NaN

Table 11: Pivot table across year and month

Plot across year and month

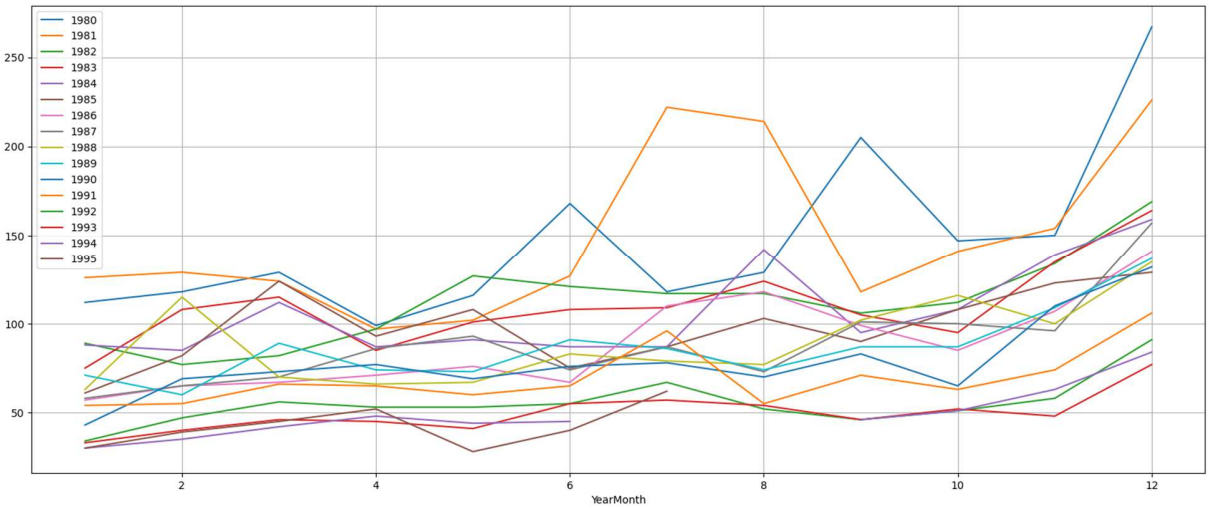


Figure9: Plot across year and month

Empirical Cumulative Sales Distribution

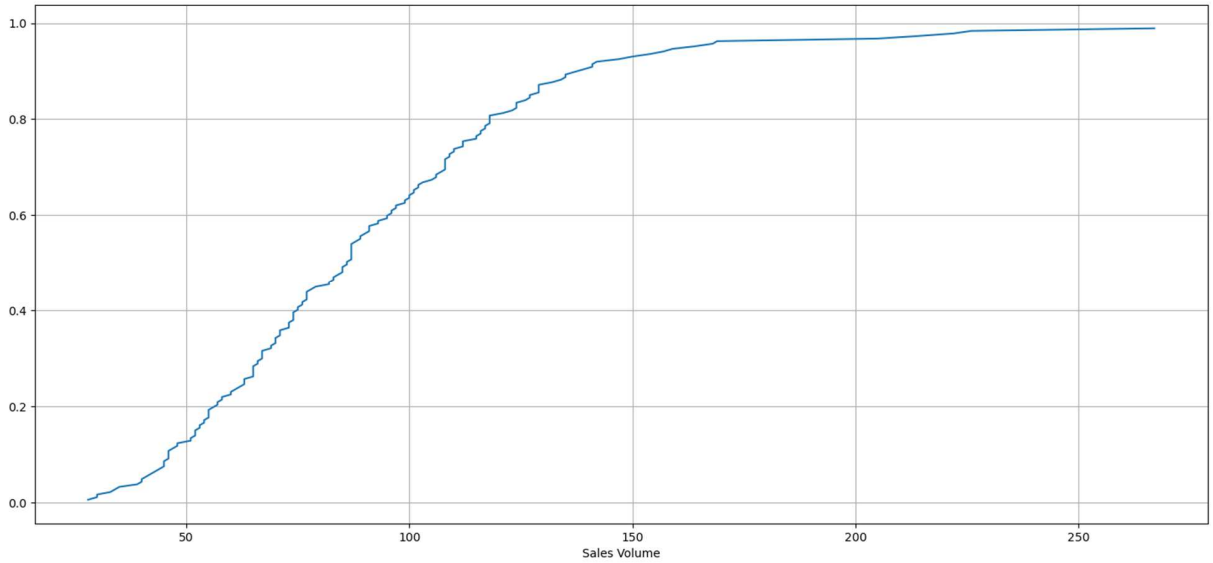


Figure10: Empirical cumulative sales distribution

Average Sales Volume per month and the month-on-month percentage change of Sales Volume

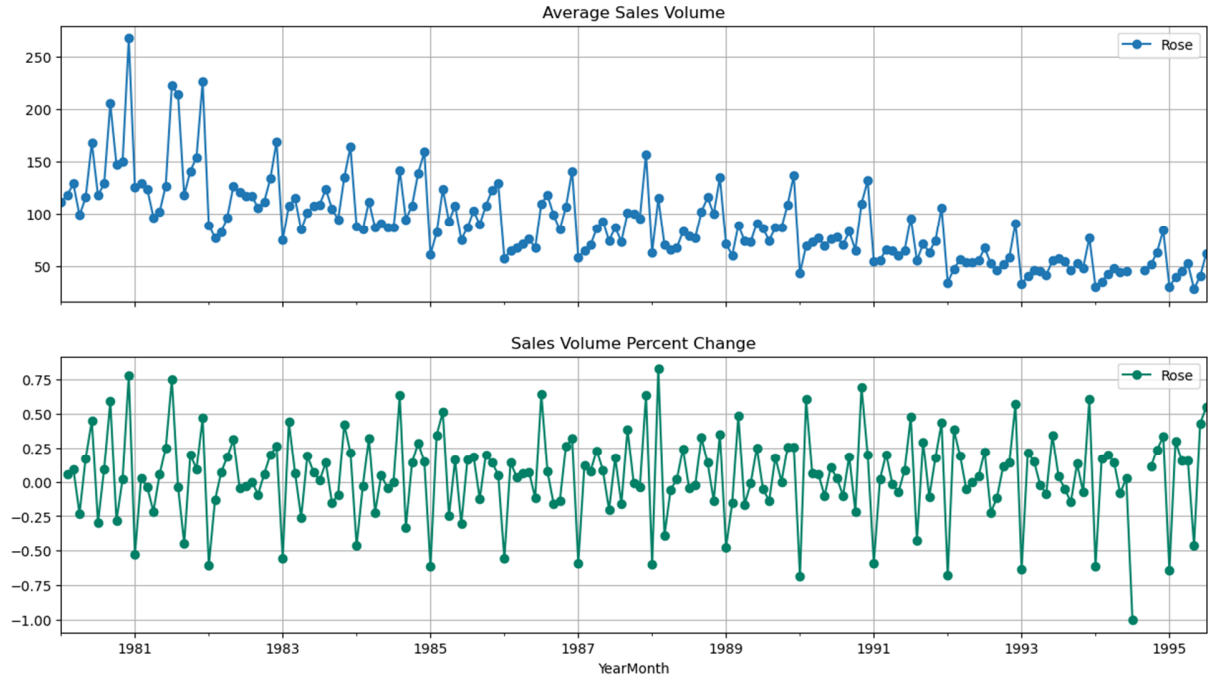


Figure11: Average sales volume and percentage change

Key Observations

1. Data is showing both trend and seasonality where trend is downward meaning from 1981 till 1995, we are seeing a steady decline in sales of Rose wine and this change is sales in multiplicative in nature.
2. Seasonally, the highest sales occur in December, while the lowest are observed in January. Sales typically show a steady increase from January through November, followed by a significant spike in December.

1.5.3 Data Decomposition

Using 'additive' model

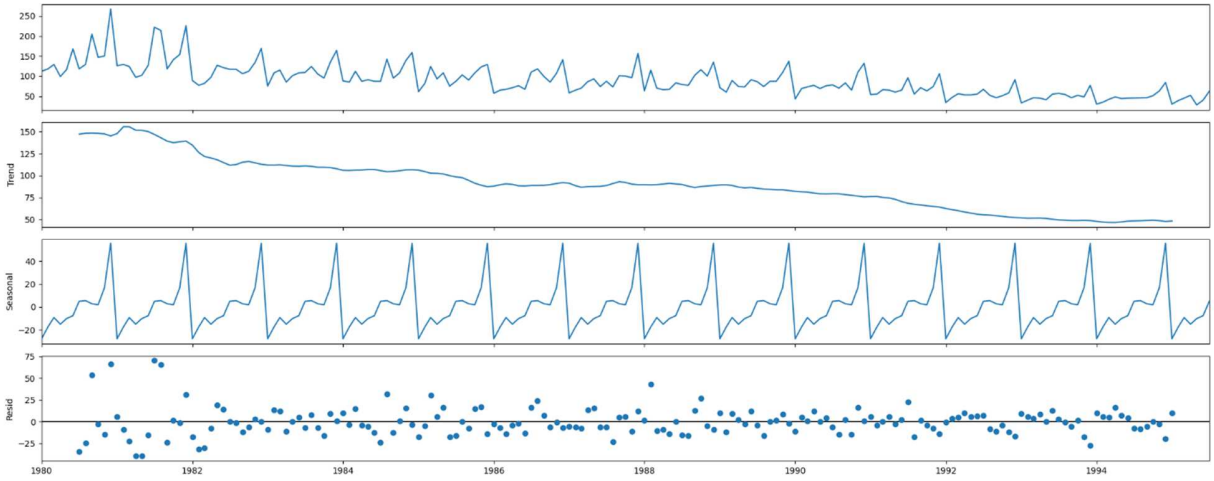


Figure12: Data Decomposition

Upon decomposing the data, we can observe both trend and seasonality in the data where like we have mentioned during the, EDA trend is downward in nature. The magnitude of fluctuations in the residuals decreases over time, indicating that a multiplicative model would be more suitable. Therefore, we will proceed with decomposition by specifying the model as multiplicative.

Using 'multiplicative' model

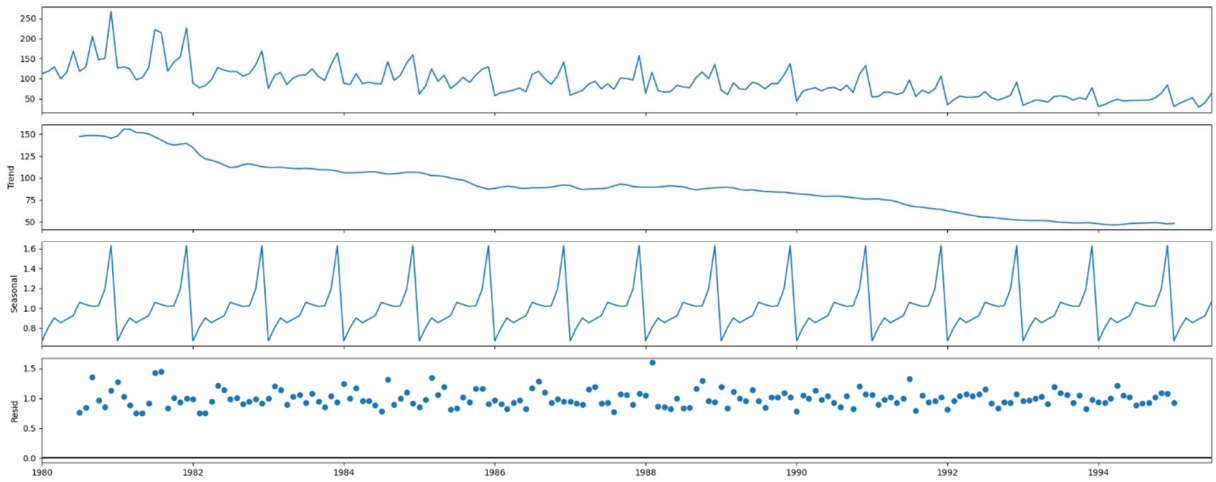


Figure13: Data Decomposition

Fluctuations in the residuals have become more consistent indicating that a multiplicative model is better.

Extracting trend, seasonality and residual from data

Trend		Seasonality		Residual	
YearMonth		YearMonth		YearMonth	
1980-01-01	NaN	1980-01-01	0.670111	1980-01-01	NaN
1980-02-01	NaN	1980-02-01	0.806163	1980-02-01	NaN
1980-03-01	NaN	1980-03-01	0.901164	1980-03-01	NaN
1980-04-01	NaN	1980-04-01	0.854024	1980-04-01	NaN
1980-05-01	NaN	1980-05-01	0.889415	1980-05-01	NaN
1980-06-01	NaN	1980-06-01	0.923985	1980-06-01	NaN
1980-07-01	147.083333	1980-07-01	1.058038	1980-07-01	0.758258
1980-08-01	148.125000	1980-08-01	1.035881	1980-08-01	0.840720
1980-09-01	148.375000	1980-09-01	1.017648	1980-09-01	1.357674
1980-10-01	148.083333	1980-10-01	1.022573	1980-10-01	0.970771
1980-11-01	147.416667	1980-11-01	1.192349	1980-11-01	0.853378
1980-12-01	145.125000	1980-12-01	1.628646	1980-12-01	1.129646
Name: trend, dtype: float64		Name: seasonal, dtype: float64		Name: resid, dtype: float64	

Table12: Extracted trend, seasonality and residual

1.5.4 Splitting Data

Here data is divided into train and test where train contains observations before 1991 and test has observations from 1991.

Train data

First few rows of Training Data

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

Last few rows of Training Data

YearMonth	Rose
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

Table13: Data Overview

Test data

Last few rows of Test Data

YearMonth	Rose
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

First few rows of Test Data

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Table14: Data Overview

1.5.5 Model Building - Original Data

We will build models using different modelling techniques for which the data could be directly used namely Linear Regression, Simple Average, Moving Averages and Exponential Modelling. For evaluation of each model, we will be using RMSE score.

Linear Regression Model

A regression model was created using LinearRegression() from linear_model in scikit-learn library. Model was used to make prediction on test data and predicted data was plotted with train and test data:

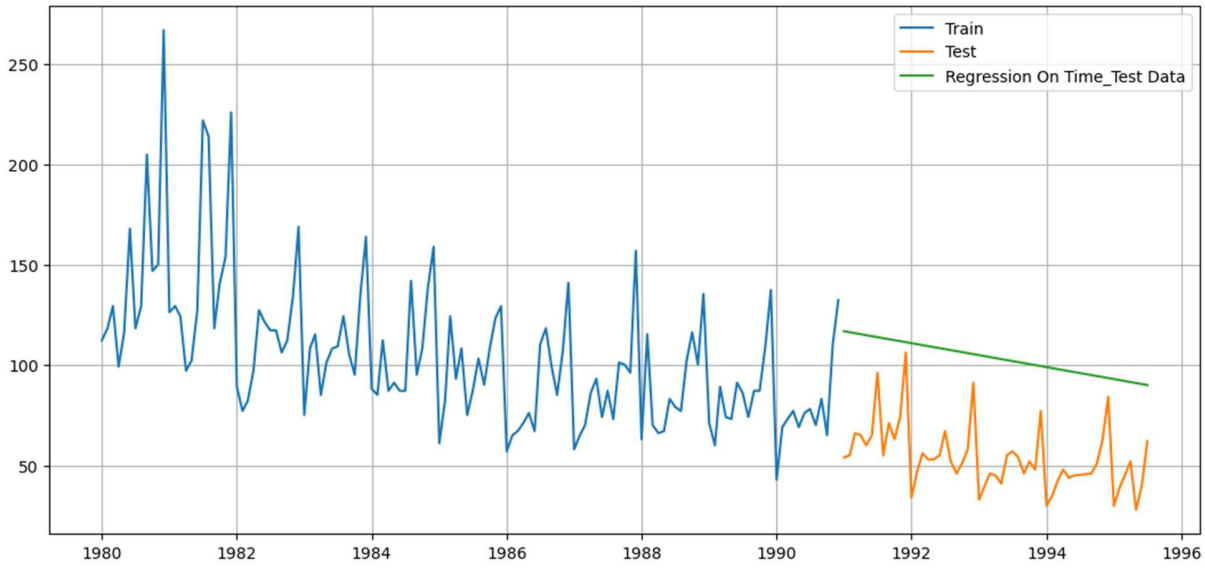


Figure14: Linear regression plot

This model is giving slanting line as prediction only considering trend and not accounting for seasonality.

Model Evaluation

```
For RegressionOnTime forecast on the Test Data,  RMSE is 51.43
```

Simple Average Model

In this model we consider the average of all the observations for the entire period as predicted observations. Based on this assumption we built a classification model where we calculated the mean of sales for train data and considered it as predicted value for test data. We plotted the predicted data along with train and test data.

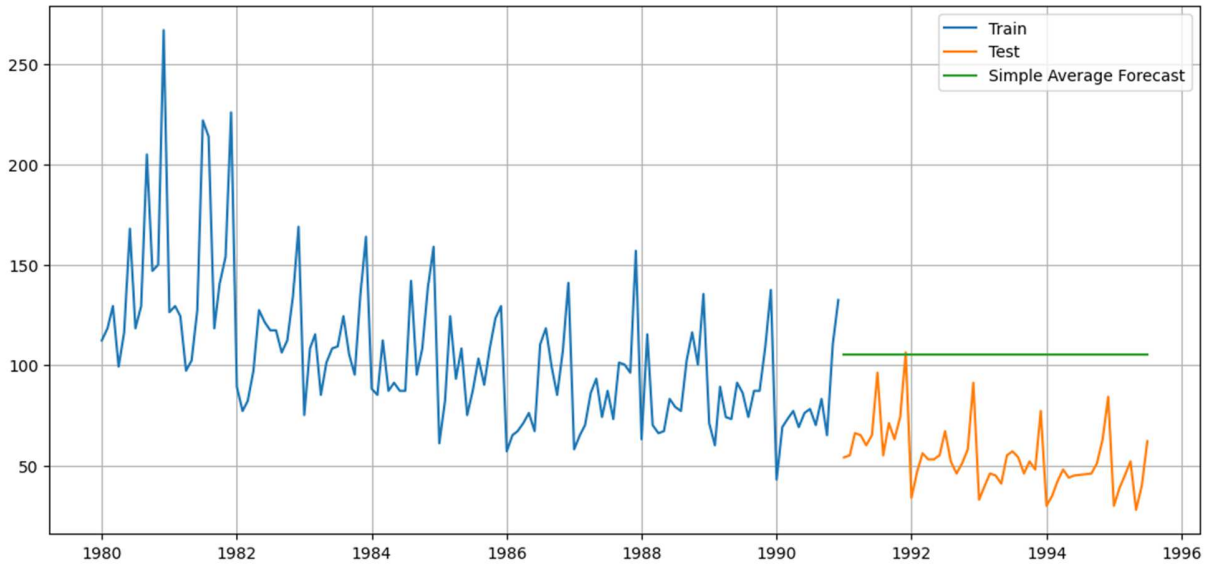


Figure15: Simple average forecast plot

This model is giving a straight line as cannot account for trend and seasonality.

Model Evaluation

```
For SimpleAverage forecast on the Test Data,  RMSE is 53.46
```

Moving Average Model

We developed a model using the trailing moving average method, where predictions are made by calculating the average of a specified number of previous data points. In this case, we considered four different trailing moving averages: 2, 4, 6, and 9. For each model, we used the corresponding number of previous data points (2, 4, 6, and 9) to build separate models over the entire dataset and plotted the moving average values alongside the train and test data.

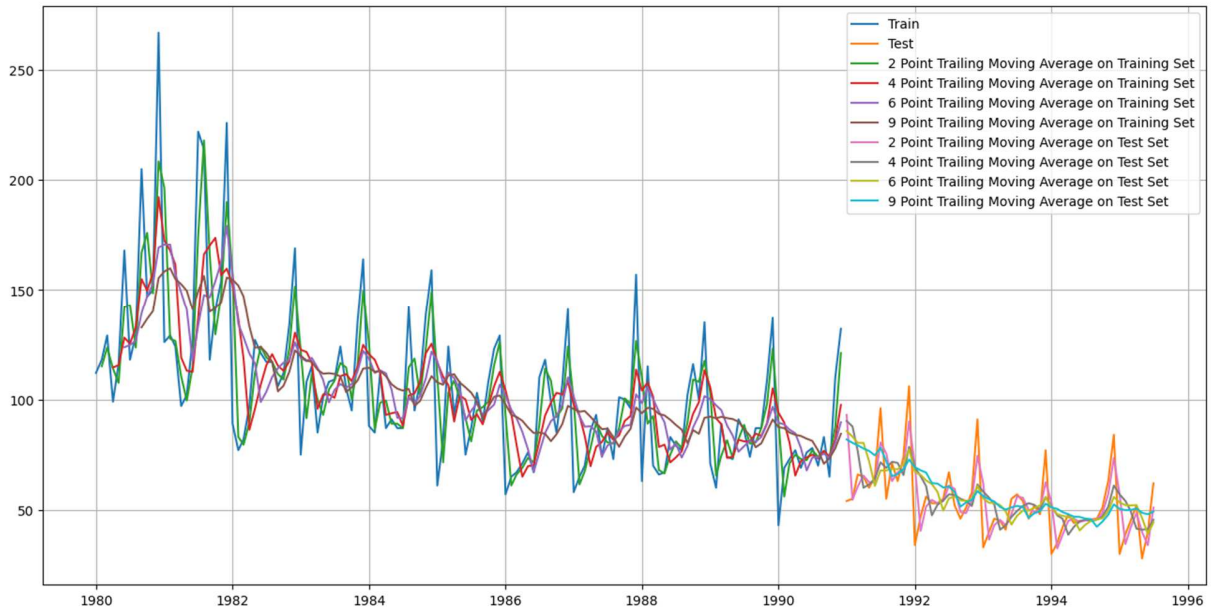


Figure16: Moving average forecast plot

Model Evaluation

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 11.529
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 14.451
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 14.566
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 14.728
```

Exponential Model

This model uses exponential smoothening to make predictions it has three parts
1. Single Exponential Smoothening: It just levels the data.
2. Double Exponential Smoothening (Holt's Linear Method): It accounts for trend in data.
3. Triple Exponential Smoothening (Holt-Winter's Linear Method): It accounts for both trend and seasonality in data.

Single Exponential Smoothening Model

We have built a single exponential smoothening model using SimpleExpSmoothing which is a part of tsa.api module in statsmodels library, this model calculates the parameters for best leveling which came as:

```
{'smoothing_level': 0.12362013444181875,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 112.0,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table15: Best Parameters

These parameters were used by model to make prediction on test data and predicted data was plotted with train and test data:

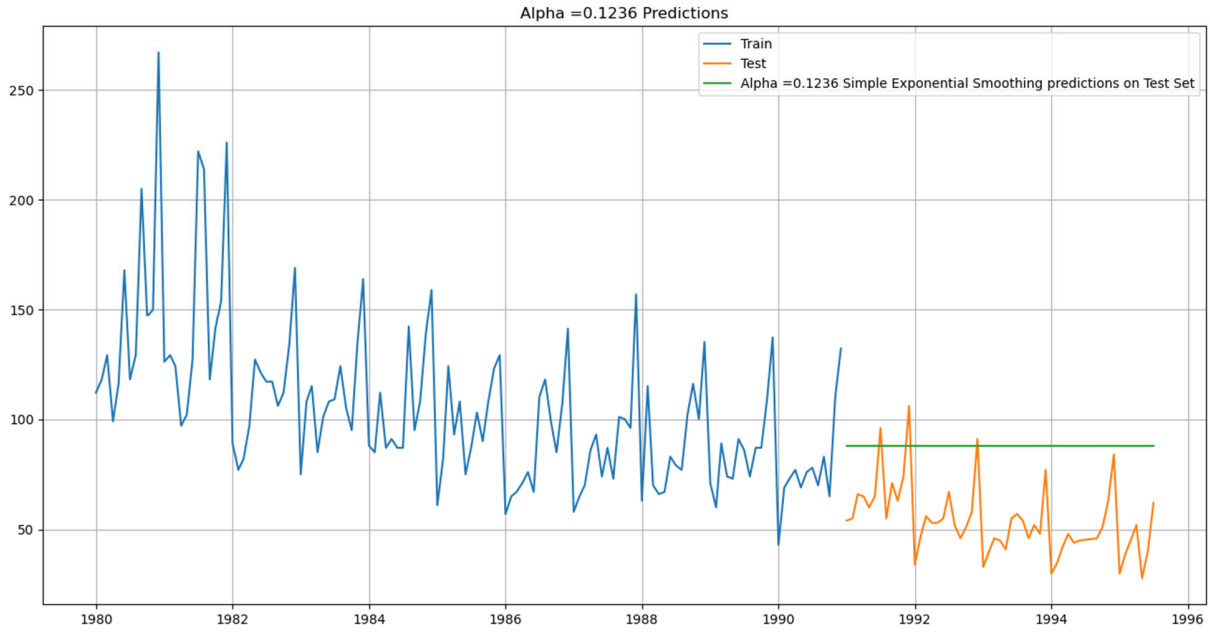


Figure17: Single exponential smoothening forecast plot

Since this model uses levelling to predict the data without considering trend and seasonality, we have a straight-line prediction.

Model Evaluation

```
For Alpha =0.1236 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 37.592
```

Double Exponential Smoothening

We have built a double exponential smoothening model using Holt which is a part of tsa.api module in statsmodels library, this model calculates the parameters for best leveling which came as:

```
{'smoothing_level': 1.4901161193847656e-08,
 'smoothing_trend': 1.6610391146660035e-10,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 137.81553690867275,
 'initial_trend': -0.4943781897068274,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table16: Best Parameters

These parameters were used by model to make prediction on test data and predicted data was plotted with train and test data:

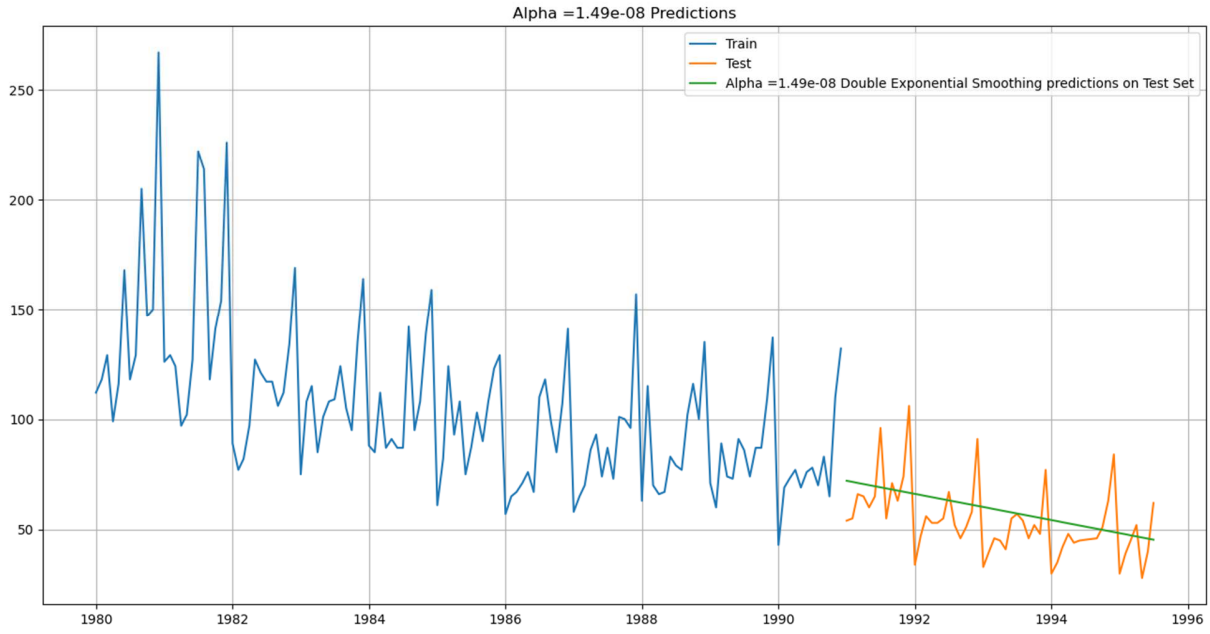


Figure18: Double exponential smoothening forecast plot

Since this model accounts for trend to predict the data without considering seasonality, we have a slanting line prediction.

Model Evaluation

```
For Alpha =1.49e-08 Double Exponential Smoothing Model forecast on the Test Data, RMSE is 15.269
```

Triple Exponential Smoothening

We have built a triple exponential smoothening model using ExponentialSmoothing which is a part of tsa.api module in statsmodels library, this model calculates the parameters for best leveling which came as:

```
{'smoothing_level': 0.08954054664605082,
 'smoothing_trend': 0.0002400108693915795,
 'smoothing_seasonal': 0.003466872515750747,
 'damping_trend': nan,
 'initial_level': 146.5570157826235,
 'initial_trend': -0.547196983509005,
 'initial_seasons': array([-31.17478463, -18.74839869, -10.76961776, -21.36741017,
       -12.63775539,  -7.27430333,   2.61279801,   8.69603625,
         4.79381122,   2.96110122,  21.05738849,  63.18279918]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table17: Best Parameters

These parameters were used by model to make prediction on test data and predicted data was plotted with train and test data:

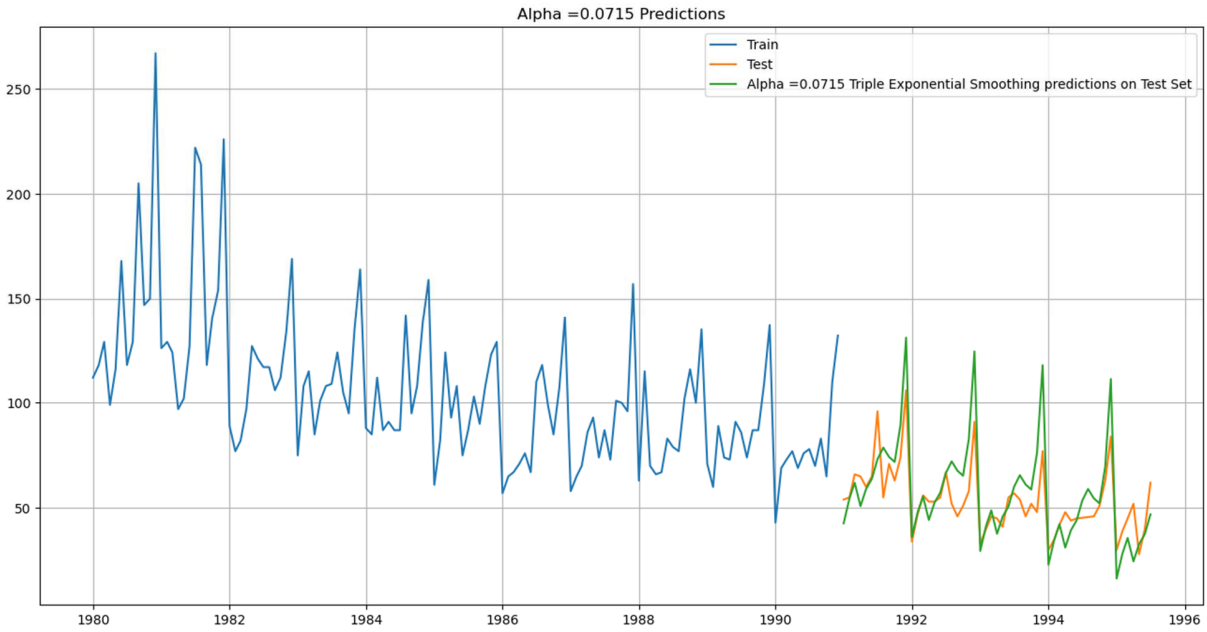


Figure19: Triple exponential smoothening forecast plot

Since this model accounts for both trend and seasonality to predict the data, it is to a great extent following same pattern as the test data.

Model Evaluation

```
For Alpha =0.0715 Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 14.250
```

1.5.6 Checking for Stationarity

Before applying ARIMA and SARIMA modelling we have to check if data is stationary as these models need stationary data for which we will do Dickey Fuller Test for which hypothesis are:

H0: Data is not stationary.

Ha: Data is stationary.

We performed the test and plotted it:

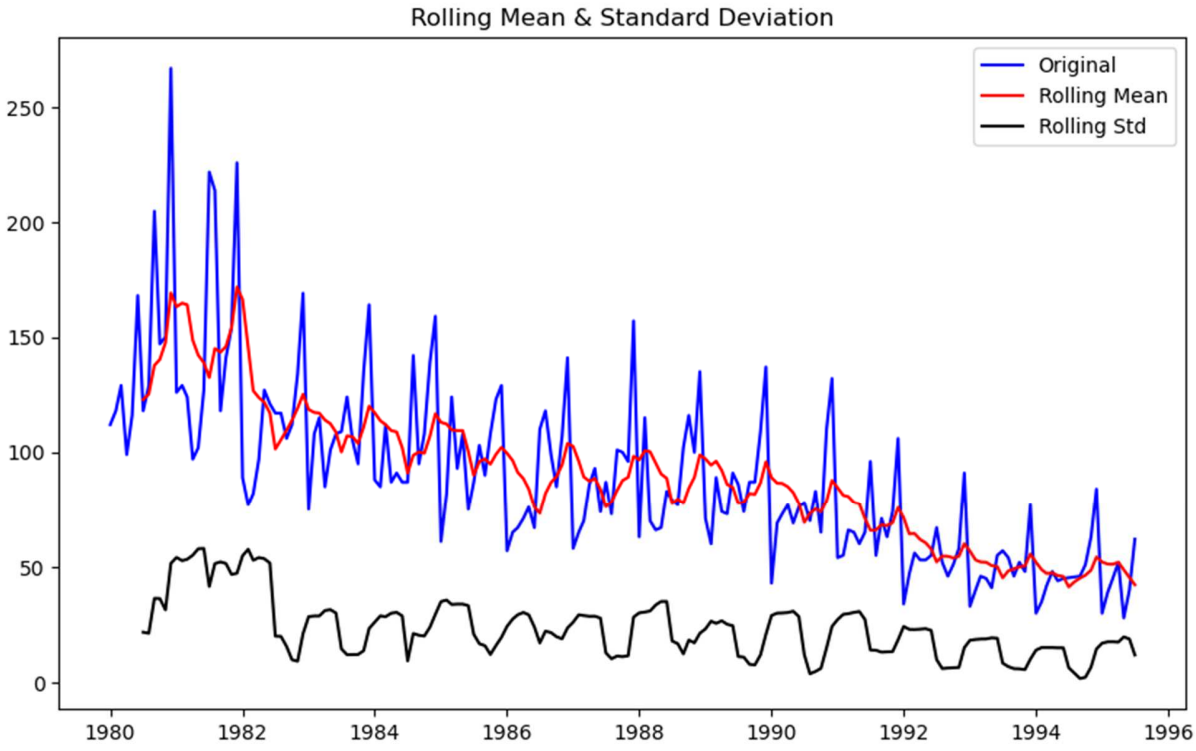


Figure20: Rolling mean and standard deviation plot

Result for this test were:

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
Weak evidence against the null hypothesis, indicating the series is non-stationary.
```

Table18: Dickey-Fuller test result

Since, p-value is greater than 0.05 means we cannot reject the null hypothesis that data is not stationary. We will have to make data stationary using differencing technique.

1.5.7 Making Data Stationary

Since the data exhibits seasonality, it's necessary to plot an autocorrelation plot to determine the appropriate seasonal differencing period. This will help in identifying the time lag where the seasonal patterns repeat, which is crucial for improving model accuracy.

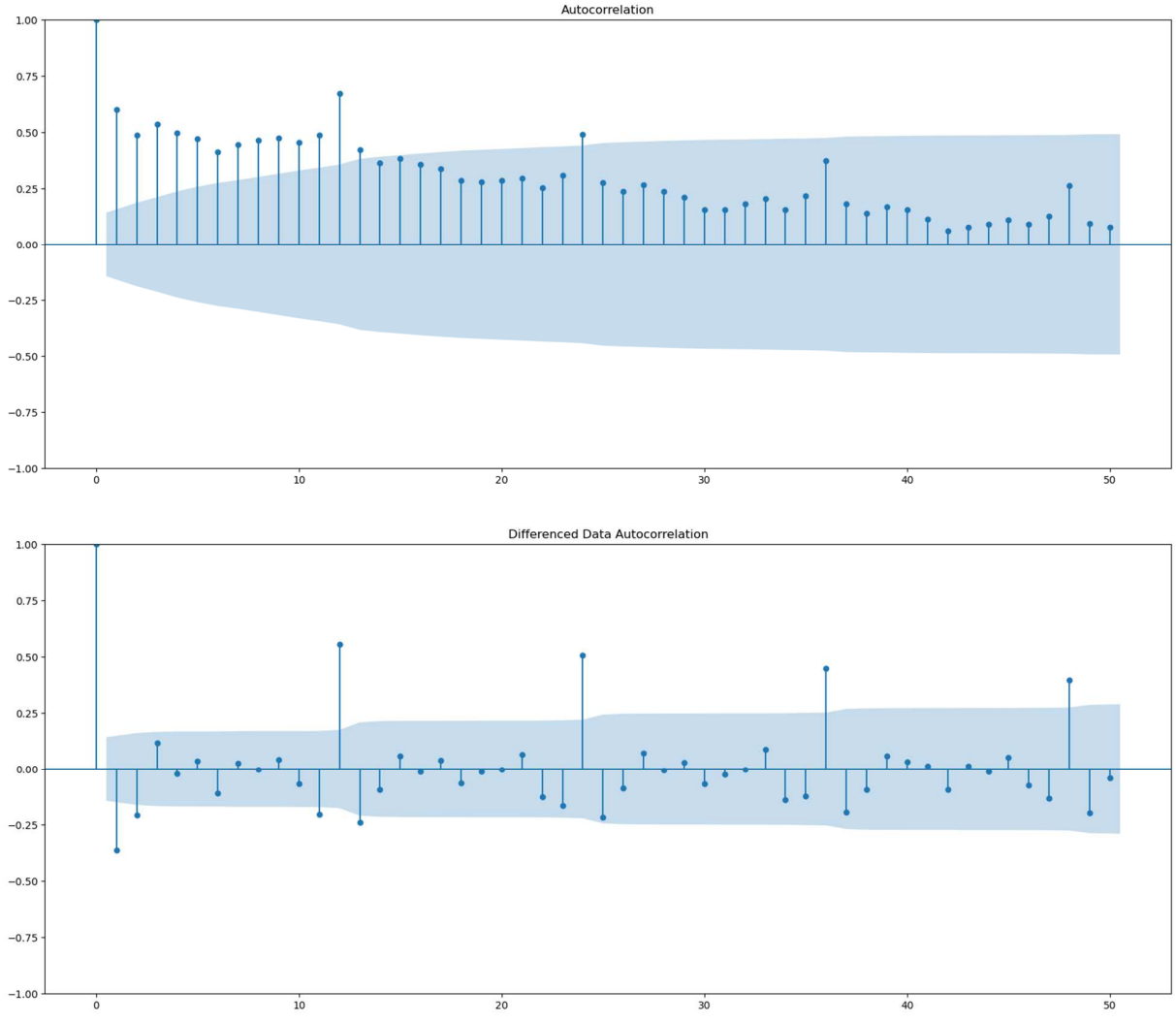


Figure21: Autocorrelation plot

From the above plot we can clearly identifying that seasonal patterns repeat after 12 months thus for differencing we will consider time lag of 12.

YearMonth	Rose	Seasonal_First_Difference
1980-01-01	112.0	NaN
1980-02-01	118.0	NaN
1980-03-01	129.0	NaN
1980-04-01	99.0	NaN
1980-05-01	116.0	NaN
1980-06-01	168.0	NaN
1980-07-01	118.0	NaN
1980-08-01	129.0	NaN
1980-09-01	205.0	NaN
1980-10-01	147.0	NaN
1980-11-01	150.0	NaN
1980-12-01	267.0	NaN
1981-01-01	126.0	14.0

Table19: First seasonal differencing

After differencing with time lag of 12 we will again conduct dickey fuller test to check for stationarity.

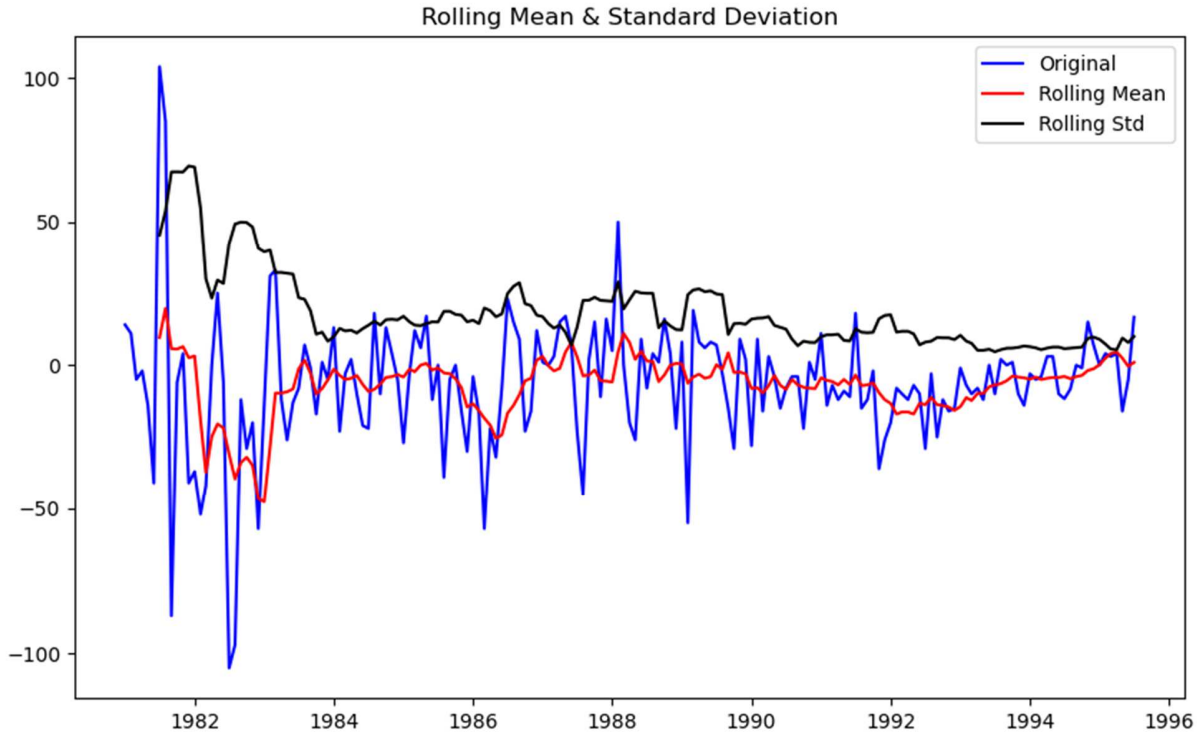


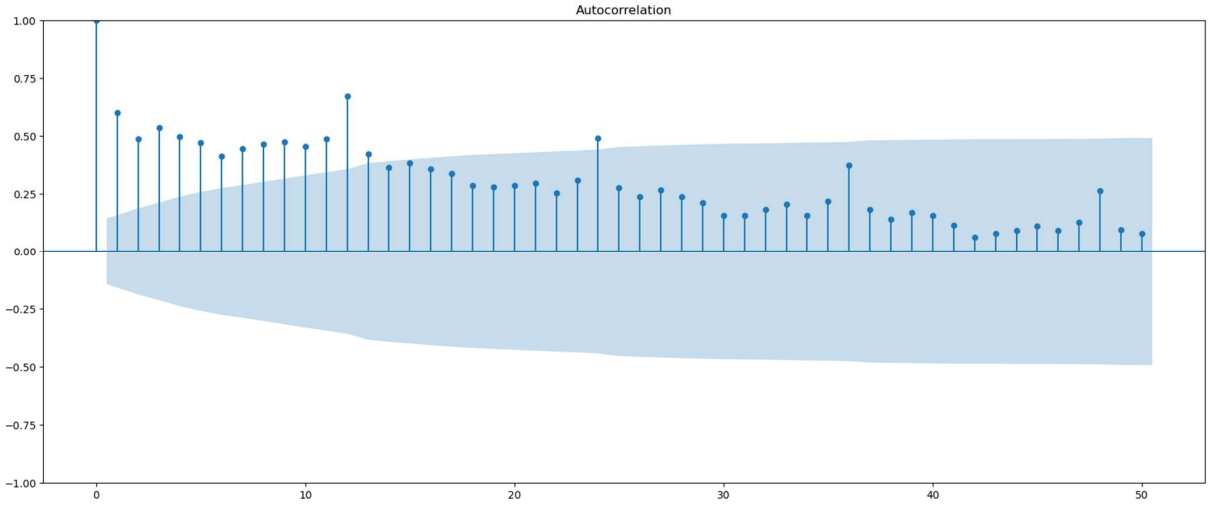
Figure22: Rolling mean and standard deviation plot

```
Results of Dickey-Fuller Test:
Test Statistic                  -4.255354
p-value                          0.000530
#Lags Used                      11.000000
Number of Observations Used    163.000000
Critical Value (1%)             -3.471119
Critical Value (5%)             -2.879441
Critical Value (10%)            -2.576314
dtype: float64
Strong evidence against the null hypothesis (H0), reject the null hypothesis. The series is stationary.
```

Table20: Dickey-Fuller test result

The data has become stationary now we can build ARIMA and SARIMA models on this stationary data. However, these models require moving average denoted by q, auto correlation denoted by p and differencing denoted by d as parameters here we have been able extract d value where d = 1 as d is the number of times differencing is done to make data stationary and in this case by first seasonal differencing we have a stationary data. Now to find p and q values we will have to plot Autocorrelation and Partial Autocorrelation Function.

1.5.8 Plot for Autocorrelation and Partial Autocorrelation Function



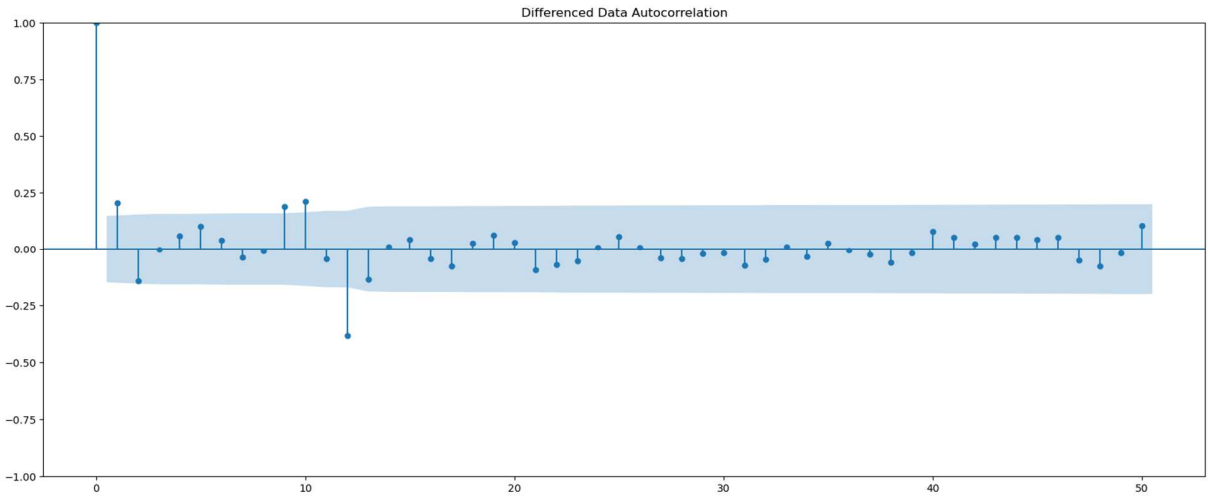
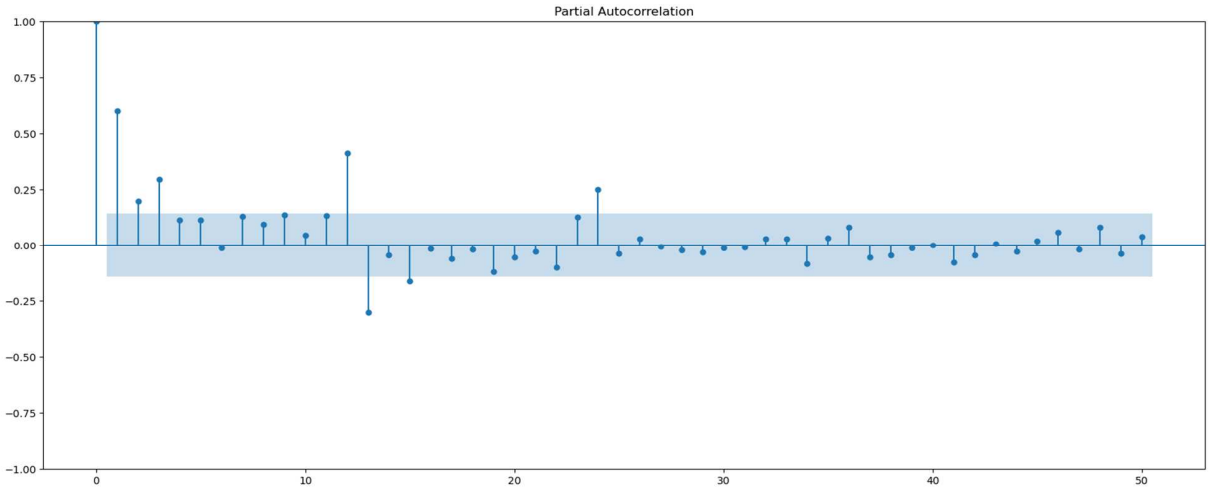


Figure23: Autocorrelation plot



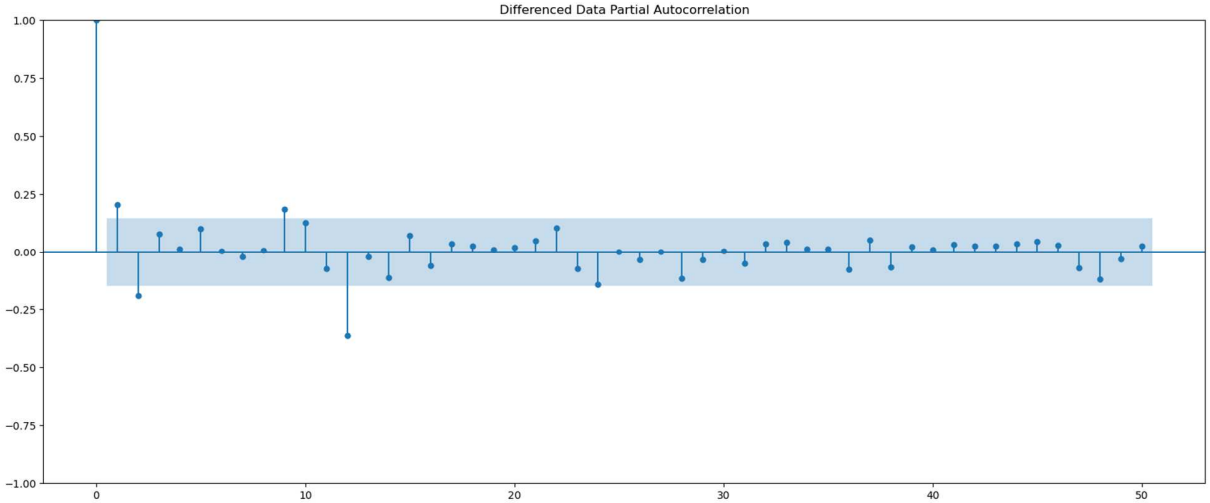


Figure24: Partial autocorrelation plot

Autocorrelation and Partial Autocorrelation Function are used to calculate optimum values of auto regression (AR) and moving average (MA) represented by p and q respectively. From the above plot we can conclude that optimum value of p should be 2 and q should be 1.

1.5.9 Model Building - Stationary Data

ARIMA Modelling

Though we have calculated optimum value of p and q using ACF and PACF plot, however, for building model we will be taking a range of p and q and calculate optimum values of p and q using auto ARIMA model.

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (0, 1, 4)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (1, 1, 4)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (2, 1, 4)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
Model: (3, 1, 4)
Model: (4, 1, 0)
Model: (4, 1, 1)
Model: (4, 1, 2)
Model: (4, 1, 3)
Model: (4, 1, 4)
```

Table21: Parameter combinations

Using these combinations, we ran the ARIMA model to calculate the AIC value of each combination. Best performing combinations were:

	param	AIC
10	(2, 1, 2)	2213.509212
15	(3, 1, 3)	2221.458953
14	(3, 1, 2)	2230.783429
11	(2, 1, 3)	2232.934772
9	(2, 1, 1)	2233.777626

Table22: Best performing parameters

In time series the best combination of p, d and q is the one with lowest AIC value for which we have sorted the above table in ascending order of AIC value based on which we can conclude that the best set of parameters are p =2, d =1 and q =3. Now we will manually build an ARIMA model using these best parameters.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sun, 22 Sep 2024   AIC                           1274.695
Time:                        14:36:25   BIC                           1291.946
Sample:                    01-01-1980   HQIC                          1281.705
                         - 12-01-1990
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6780      0.084    -20.047      0.000      -1.842      -1.514
ar.L2         -0.7289      0.084     -8.708      0.000      -0.893      -0.565
ma.L1          1.0448      0.653      1.601      0.109      -0.235       2.324
ma.L2         -0.7717      0.134     -5.742      0.000      -1.035      -0.508
ma.L3         -0.9045      0.592     -1.527      0.127      -2.065       0.256
sigma2       858.2216    549.150      1.563      0.118    -218.094    1934.537
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.44
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Table23: ARIMA result

Model Evaluation

```
For ARIMA(2,1,3) Model with  forecast on the Test Data, RMSE is %3.3f 36.81714953212603
```

SARIMA Modelling

In SARIMA models in addition to trend we also account for seasonality and to understand the seasonal parameter we have to use ACF plot.

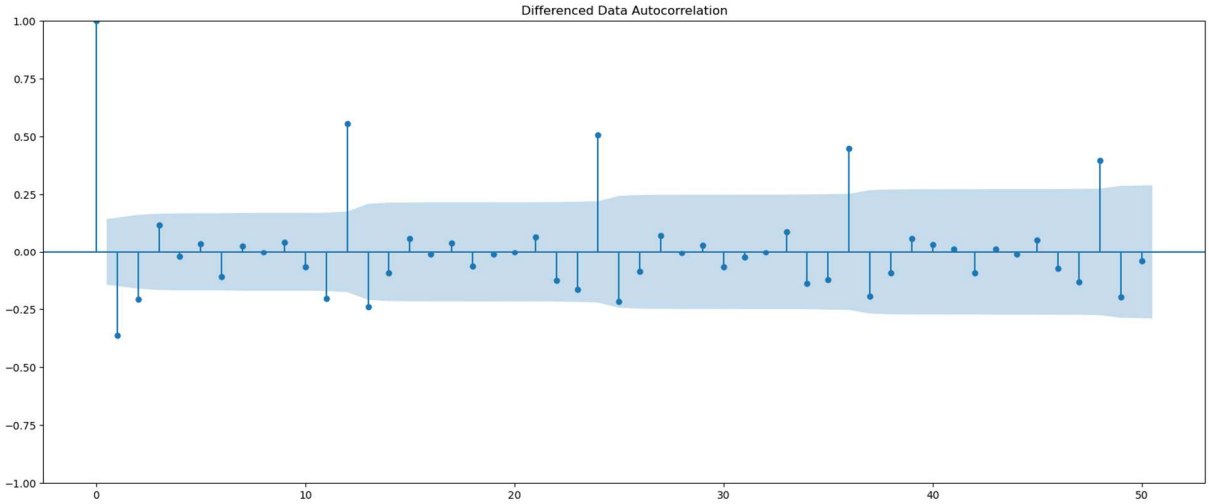


Figure25: Autocorrelation plot

We see that there can be a seasonality of 12. We will run our auto SARIMA models by setting as 12.

```
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)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

Table24: Parameter combinations

Using these combinations, we ran the SARIMA model to calculate the AIC value of each combination. Best performing combinations were:

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
53	(1, 1, 2)	(2, 0, 2, 12)	889.871767
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

Table25: Best performing parameters

From the above table which is sorted in ascending order of AIC score we can take the best values of parameters at (0, 1, 2) (2, 0, 2, 12). We will build the final SARIMA model using these parameters.

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-436.969
Date:	Sun, 22 Sep 2024	AIC	887.938
Time:	14:36:55	BIC	906.448
Sample:	01-01-1980 - 12-01-1990	HQIC	895.437
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8427	189.426	-0.004	0.996	-372.110	370.425
ma.L2	-0.1573	29.760	-0.005	0.996	-58.485	58.171
ar.S.L12	0.3467	0.079	4.375	0.000	0.191	0.502
ar.S.L24	0.3023	0.076	3.996	0.000	0.154	0.451
ma.S.L12	0.0767	0.133	0.577	0.564	-0.184	0.337
ma.S.L24	-0.0726	0.146	-0.498	0.618	-0.358	0.213
sigma2	251.3136	4.76e+04	0.005	0.996	-9.31e+04	9.36e+04

Ljung-Box (L1) (Q):	0.10	Jarque-Bera (JB):	2.33
Prob(Q):	0.75	Prob(JB):	0.31
Heteroskedasticity (H):	0.88	Skew:	0.37
Prob(H) (two-sided):	0.70	Kurtosis:	3.03

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Table26: SARIMA result

For these set of parameters, we plot a residual plot to check weather we have been able to extract all the trend and seasonality from data.

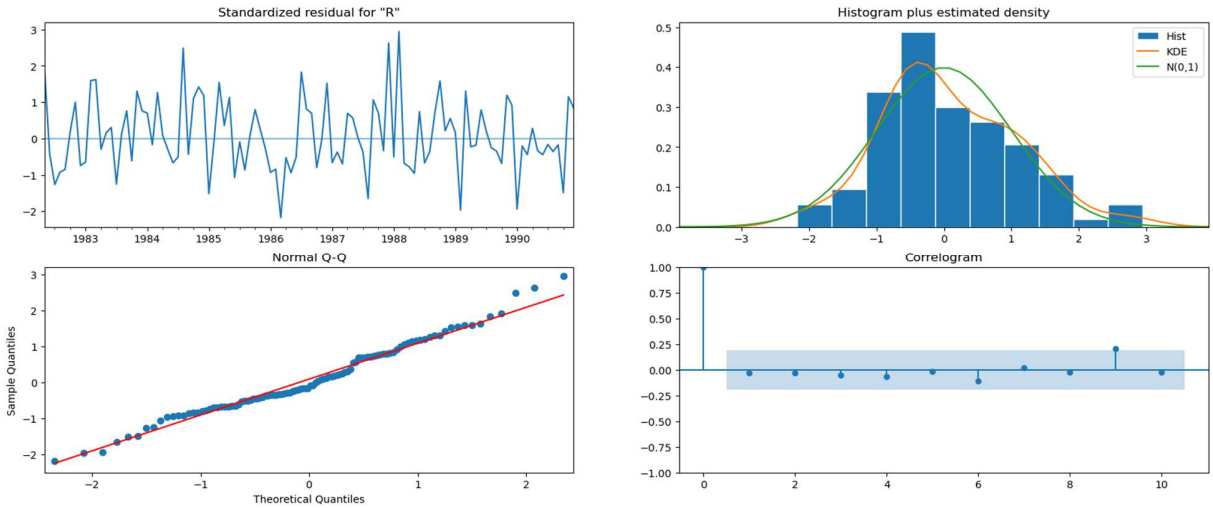


Figure26: Residual plot

In the above diagnostic plot residuals are randomly scattered meaning we have optimized the model now we can move ahead to evaluate it.

Model Evaluation

```
For SARIMA(0, 1, 2) (2, 0, 2, 12) Model with  forecast on the Test Data, RMSE is %3.3f 26.928360855038665
```

1.5.10 Model Comparison

We have created 11 models using different techniques compared each model's performance for test data using key metrics and have found that all the models are stable now we will compare these models with each other to find the best model based on their RMSE score for test data.

	Test RMSE
2pointTrailingMovingAverage	11.529278
Alpha=0.0715,TripleExponentialSmoothing	14.249661
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
Alpha=1.49e-08,DoubleExponentialSmoothing	15.268944
SARIMA(0,1,2)(2,0,2,12)	26.928361
ARIMA(2,1,3)	36.817150
Alpha=0.1236,SimpleExponentialSmoothing	37.592212
RegressionOnTime	51.433312
SimpleAverage_Forecast	53.460570

Table 27: Model comparison

From the above table we can conclude that 2pointTrailingMovingAverage model is the best performing model as it has the lowest RMSE score. For forecasting we will use this model.

1.5.11 Building the most optimum model on the Full Data

We rebuilt the 2pointTrailingMovingAverage model on the entire dataset as the final model and calculate the RMSE to evaluate the model performance.

Model Evaluation

```
For 2 point Moving Average Model forecast on the full Data,  RMSE is 17.498
```

Forecasting 12 months into the future

	Rose	lower_CI	upper_CI
1995-08-01	51.000000	16.705127	85.294873
1995-09-01	56.500000	22.205127	90.794873
1995-10-01	53.750000	19.455127	88.044873
1995-11-01	55.125000	20.830127	89.419873
1995-12-01	54.437500	20.142627	88.732373
1996-01-01	54.781250	20.486377	89.076123
1996-02-01	54.609375	20.314502	88.904248
1996-03-01	54.695312	20.400439	88.990186
1996-04-01	54.652344	20.357471	88.947217
1996-05-01	54.673828	20.378955	88.968701
1996-06-01	54.663086	20.368213	88.957959

Table28: Confidence interval

The table above shows the predicted sales for the upcoming months, along with a confidence interval. We estimate, with 95% confidence, that the actual sales will fall within this specified range.

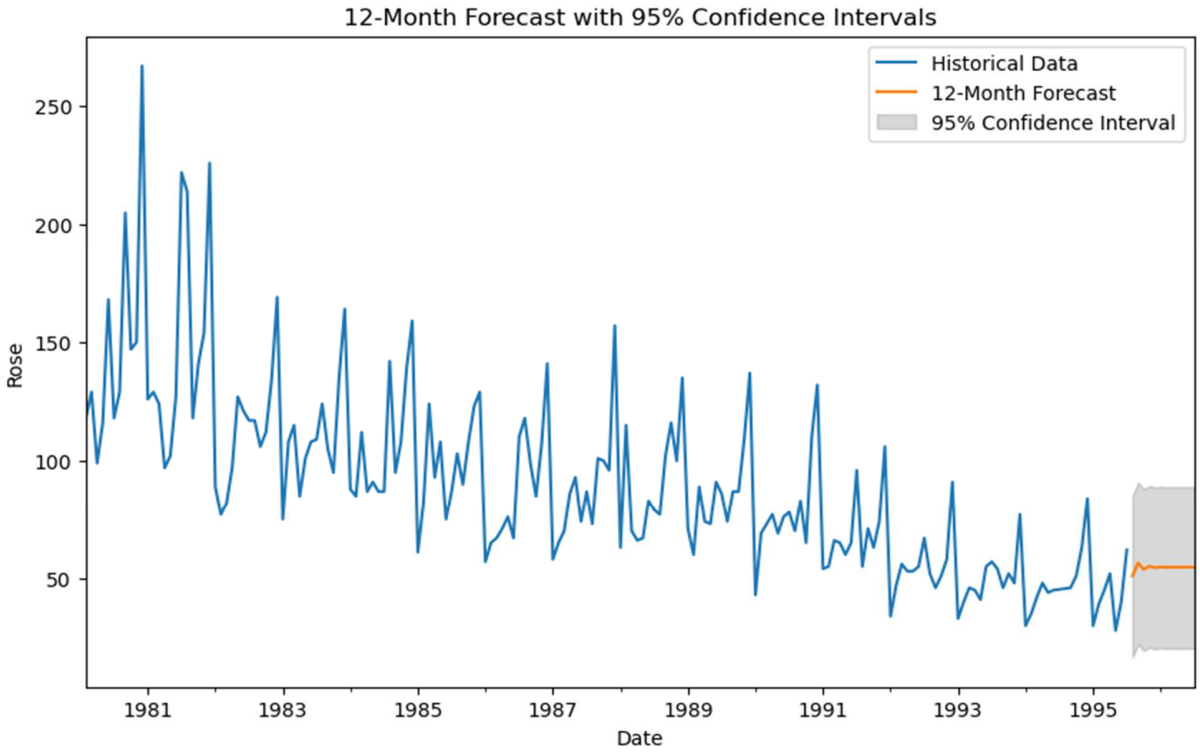


Figure27: forecast plot

The plot above illustrates the forecasted sales, represented by the orange line, for the next 12 months along with the historic data. The shaded region around indicates the range within which sales are expected to fall with 95% confidence.

1.6 Sparkling Wine Data

1.6.1 Data Overview

1. **Data Description**: Dataset has 187 rows and 2 columns.

```
shape of the dataset
--------------------------------------------------------------------------------

(187, 2)
```

Table29: Dataset Shape

2. **Dataset Information:** Of the 2 columns in the dataset, 1 is object type and 1 is float 64 types.

```
information of features
--------------------------------------------------------------------------------
<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: 3.1+ KB
```

Table30: Dataset Information

3. **Missing Value Check:** There were no missing values in the dataset.

```
Number of rows with missing values:
--------------------------------------------------------------------------------

YearMonth    0
Sparkling    0
dtype: int64
```

Table31: Missing values information

4. **Duplicate Values:** Data was checked for duplicate values and no duplicates were found

```
checking for duplicates
--------------------------------------------------------------------------------
number of dupliacte rows: 0
```

Table32: Data Duplicates

5. **Statistical Summary:**

```
statistical summary
--------------------------------------------------------------------------------
```

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

Table33: Statistical Summary

Key observations

1. Dataset has 2 rows and 187 observations in which YearMonth column should be of datetime datatype but is of object type. We will convert this feature to datetime and transform the data to a time series data.
2. There are no missing values in the dataset.

1.6.2 Exploratory Data Analysis

Plotting Data

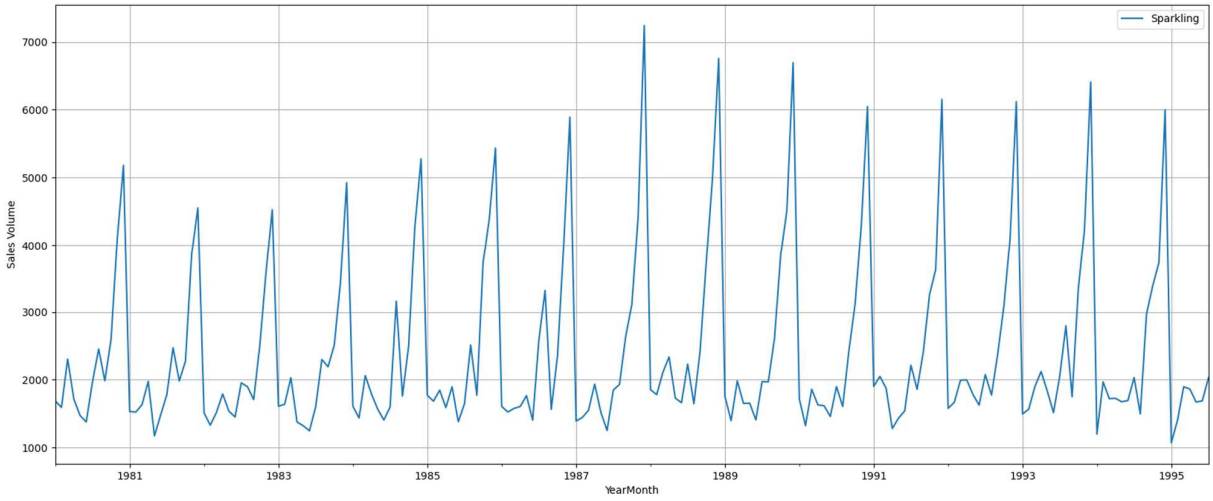


Figure28: Timeseries plot

Boxplot by Year

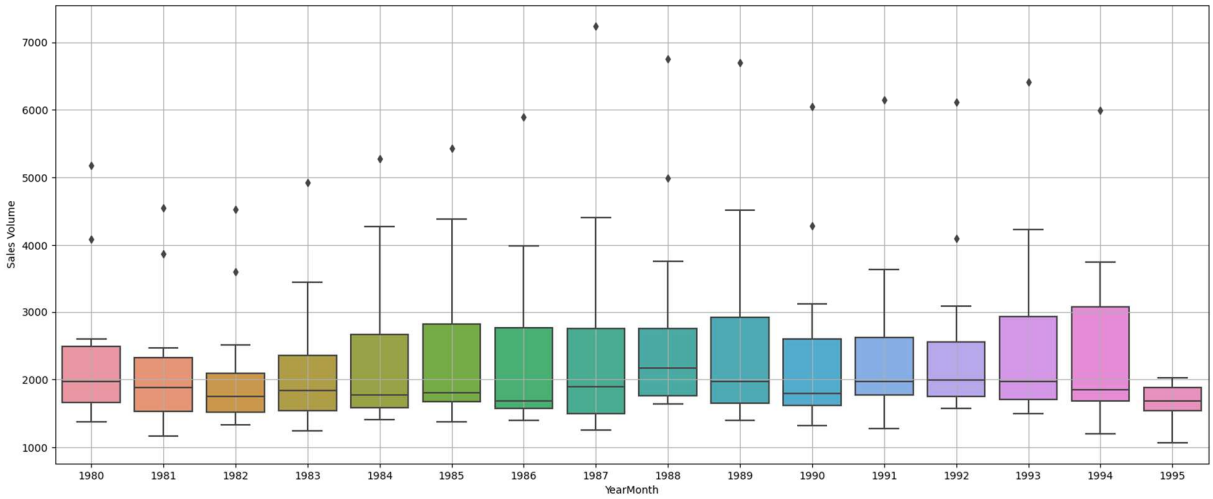


Figure29: boxplot by year

Boxplot by Month

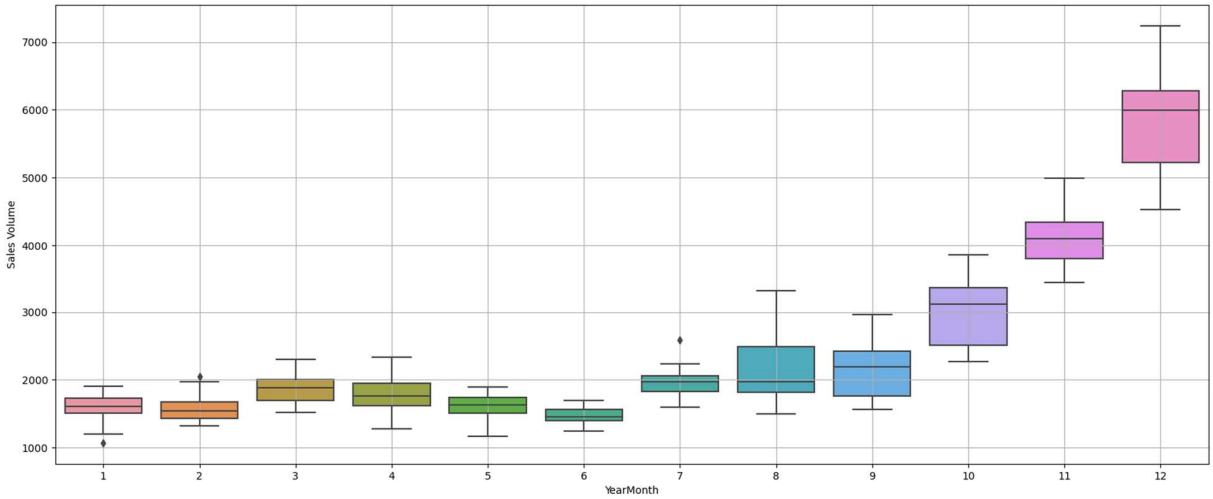


Figure30: boxplot by month

Time Series Monthplot

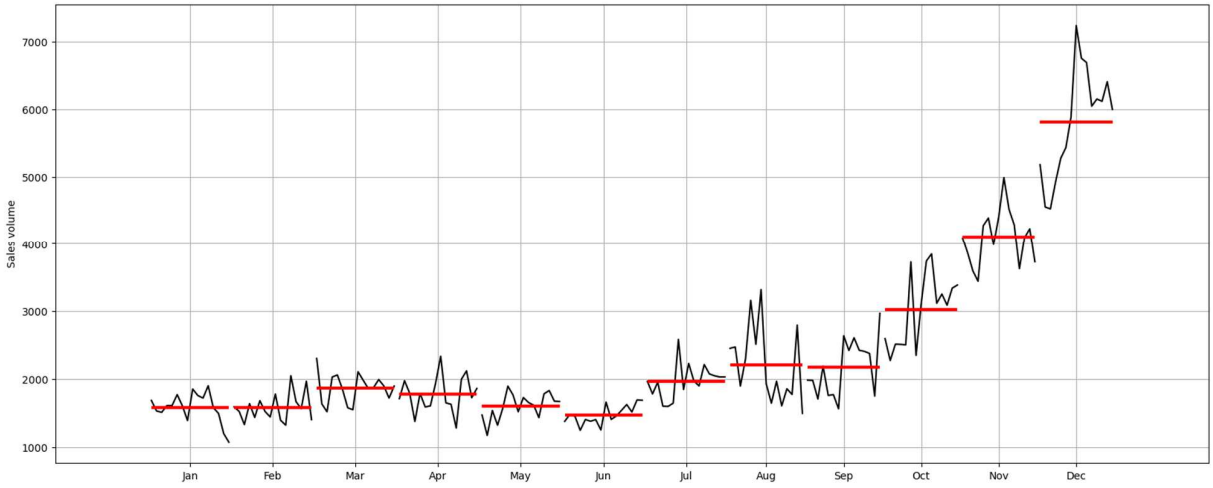


Figure31: Timeseries month plot

Monthly Sales across Years

Pivot table across month and year

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

Table34: Pivot table across month and year

Plot across month and year

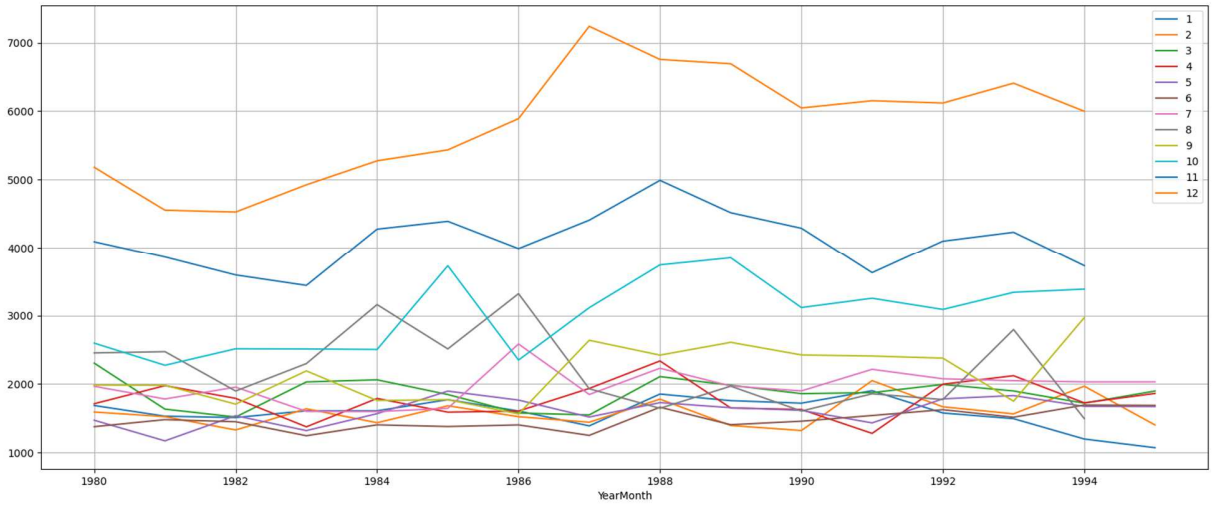


Figure32: Plot across month and year

Yearly Sales across Months

Pivot table across year and month

YearMonth	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
YearMonth																
1	1686.0	1530.0	1510.0	1609.0	1609.0	1771.0	1606.0	1389.0	1853.0	1757.0	1720.0	1902.0	1577.0	1494.0	1197.0	1070.0
2	1591.0	1523.0	1329.0	1638.0	1435.0	1682.0	1523.0	1442.0	1779.0	1394.0	1321.0	2049.0	1667.0	1564.0	1968.0	1402.0
3	2304.0	1633.0	1518.0	2030.0	2061.0	1846.0	1577.0	1548.0	2108.0	1982.0	1859.0	1874.0	1993.0	1898.0	1720.0	1897.0
4	1712.0	1976.0	1790.0	1375.0	1789.0	1589.0	1605.0	1935.0	2336.0	1650.0	1628.0	1279.0	1997.0	2121.0	1725.0	1862.0
5	1471.0	1170.0	1537.0	1320.0	1567.0	1896.0	1765.0	1518.0	1728.0	1654.0	1615.0	1432.0	1783.0	1831.0	1674.0	1670.0
6	1377.0	1480.0	1449.0	1245.0	1404.0	1379.0	1403.0	1250.0	1661.0	1406.0	1457.0	1540.0	1625.0	1515.0	1693.0	1688.0
7	1966.0	1781.0	1954.0	1600.0	1597.0	1645.0	2584.0	1847.0	2230.0	1971.0	1899.0	2214.0	2076.0	2048.0	2031.0	2031.0
8	2453.0	2472.0	1897.0	2298.0	3159.0	2512.0	3318.0	1930.0	1645.0	1968.0	1605.0	1857.0	1773.0	2795.0	1495.0	NaN
9	1984.0	1981.0	1706.0	2191.0	1759.0	1771.0	1562.0	2638.0	2421.0	2608.0	2424.0	2408.0	2377.0	1749.0	2968.0	NaN
10	2596.0	2273.0	2514.0	2511.0	2504.0	3727.0	2349.0	3114.0	3740.0	3845.0	3116.0	3252.0	3088.0	3339.0	3385.0	NaN
11	4087.0	3857.0	3593.0	3440.0	4273.0	4388.0	3987.0	4405.0	4988.0	4514.0	4286.0	3627.0	4096.0	4227.0	3729.0	NaN
12	5179.0	4551.0	4524.0	4923.0	5274.0	5434.0	5891.0	7242.0	6757.0	6694.0	6047.0	6153.0	6119.0	6410.0	5999.0	NaN

Table35: Pivot table across year and month

Plot across year and month

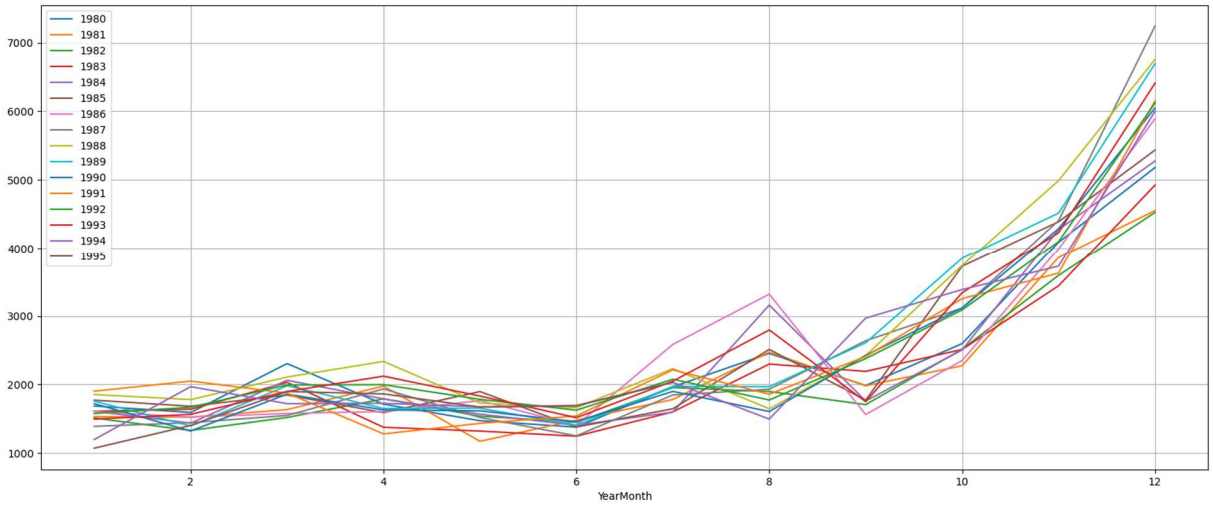


Figure33: Plot across year and month

Empirical Cumulative Sales Distribution

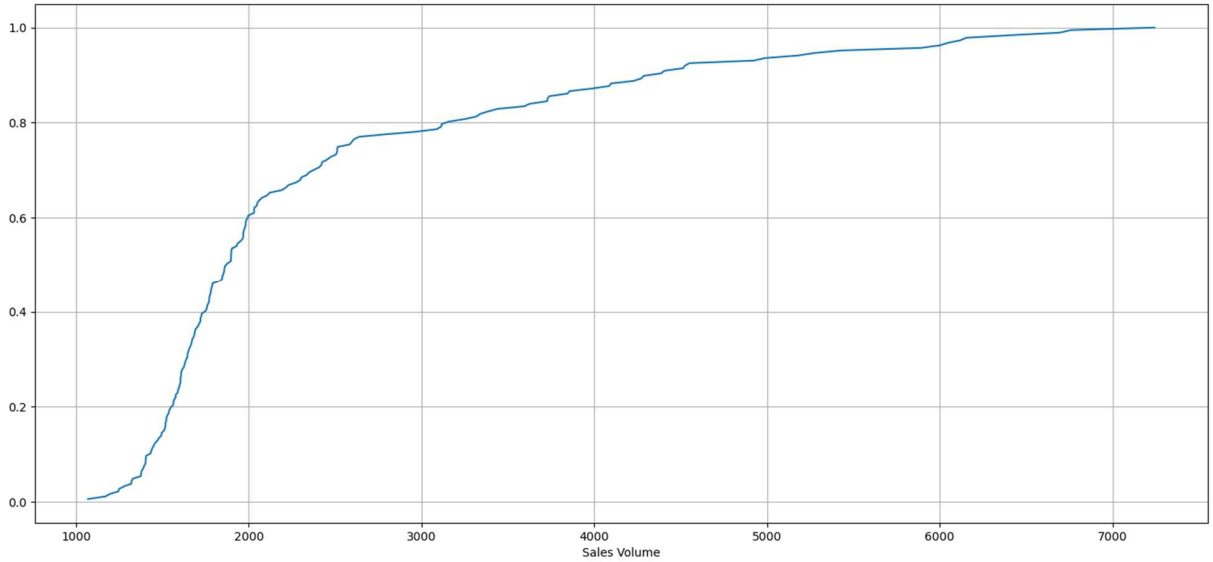


Figure34: Empirical cumulative month distribution

Average Sales Volume per month and the month-on-month percentage change of Sales Volume

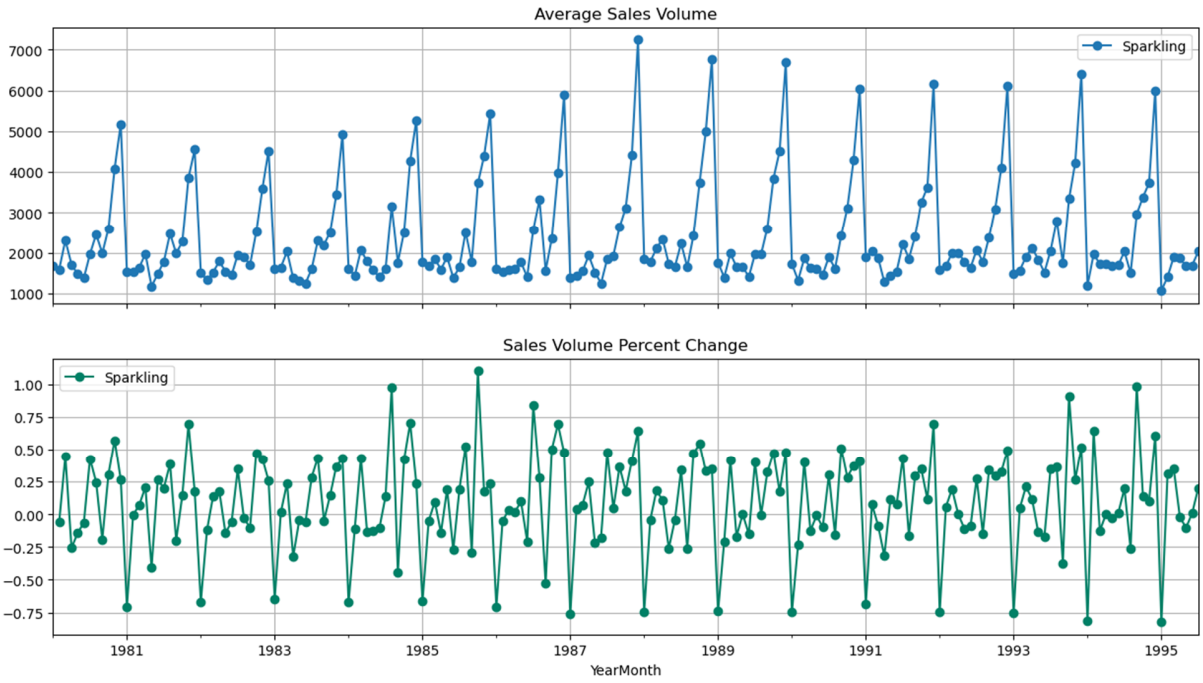


Figure35: Average sales volume and percentage change

Key Observations

1. Data is showing only seasonality as in year-on-year terms the sales tend to remain within a range and though there are some fluctuations these changes are not consistent; we will attribute these changes as error. In monthly terms though there is a set pattern being observed.
2. Seasonally, the highest sales occur in December, while the lowest are observed in January. Sales typically show a steady increase from January through November, followed by a significant jump in December.

1.6.3 Data Decomposition

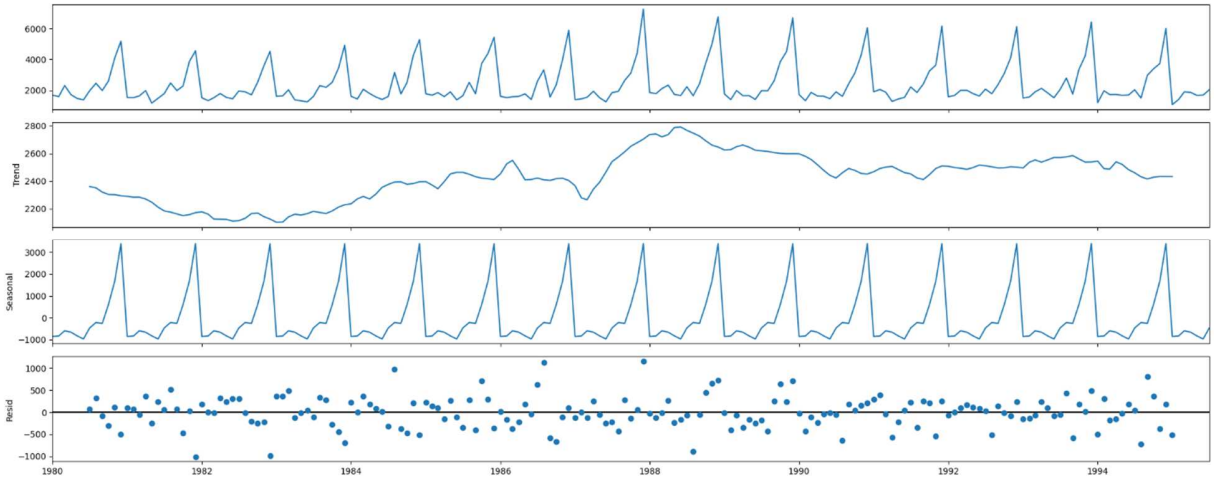


Figure36: Data Decomposition

the above plot reitrates what we had observed during the EDA that there is no set standard pattern in trend meaning there is no trend in the data but there is clear seasonality in the data.

Extracting trend, seasonality and residual from data

Seasonality		Trend		Residual	
YearMonth		YearMonth		YearMonth	
1980-01-01	-854.260599	1980-01-01	NaN	1980-01-01	NaN
1980-02-01	-830.350678	1980-02-01	NaN	1980-02-01	NaN
1980-03-01	-592.356630	1980-03-01	NaN	1980-03-01	NaN
1980-04-01	-658.490559	1980-04-01	NaN	1980-04-01	NaN
1980-05-01	-824.416154	1980-05-01	NaN	1980-05-01	NaN
1980-06-01	-967.434011	1980-06-01	NaN	1980-06-01	NaN
1980-07-01	-465.502265	1980-07-01	2360.666667	1980-07-01	70.835599
1980-08-01	-214.332821	1980-08-01	2351.333333	1980-08-01	315.999487
1980-09-01	-254.677265	1980-09-01	2320.541667	1980-09-01	-81.864401
1980-10-01	599.769957	1980-10-01	2303.583333	1980-10-01	-307.353290
1980-11-01	1675.067179	1980-11-01	2302.041667	1980-11-01	109.891154
1980-12-01	3386.983846	1980-12-01	2293.791667	1980-12-01	-501.775513
Name: seasonal, dtype: float64		Name: trend, dtype: float64		Name: resid, dtype: float64	

Table36: Extracted trend, seasonality and residual

1.6.4 Splitting Data

Here data is divided into train and test where train contains observations before 1991 and test has observations from 1991.

Train data

First few rows of Training Data

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

Last few rows of Training Data

YearMonth	Sparkling
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

Table37: Data Overview

Test data

First few rows of Test Data

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

Last few rows of Test Data

YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Table38: Data Overview

1.6.5 Model Building - Original Data

We will build models using different modelling techniques for which the data could be directly used namely Linear Regression, Simple Average, Moving Averages and Exponential Modelling. For evaluation of each model, we will be using RMSE score.

Linear Regression Model

A regression model was created using LinearRegression() from linear_model in scikit-learn library. Model was used to make prediction on test data and predicted data was plotted with train and test data:

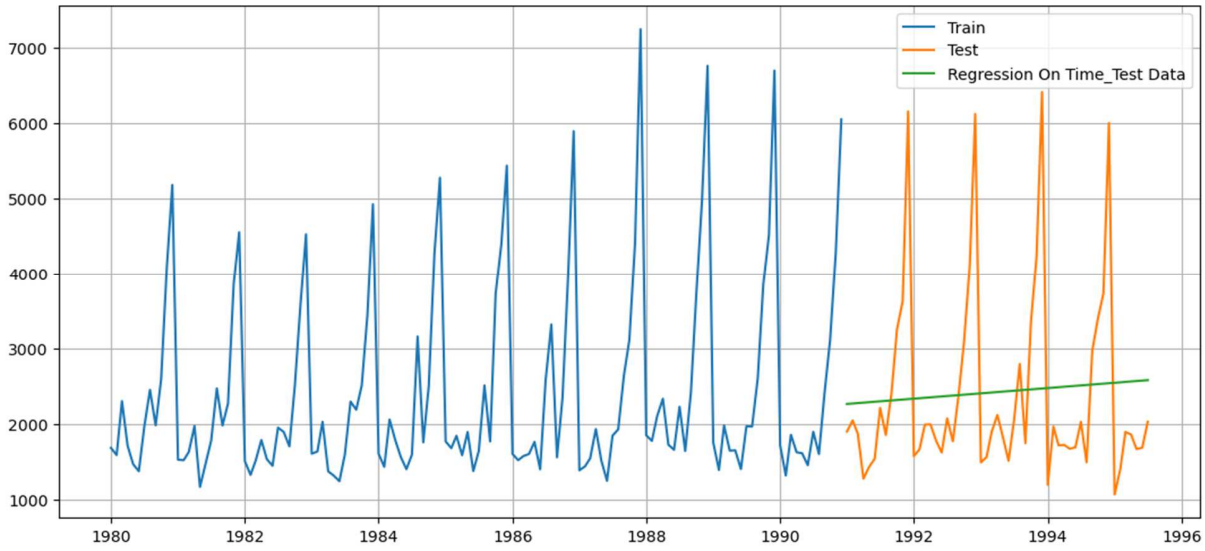


Figure37: Linear regression plot

This model is giving a slightly slanting line as prediction does not account for seasonality.

Model Evaluation

```
For RegressionOnTime forecast on the Test Data,  RMSE is 1275.87
```

Simple Average Model

In this model we consider the average of all the observations for the entire period as predicted observations. Based on this assumption we built a classification model where we calculated the mean of sales for train data and considered it as predicted value for test data. We plotted the predicted data along with train and test data.

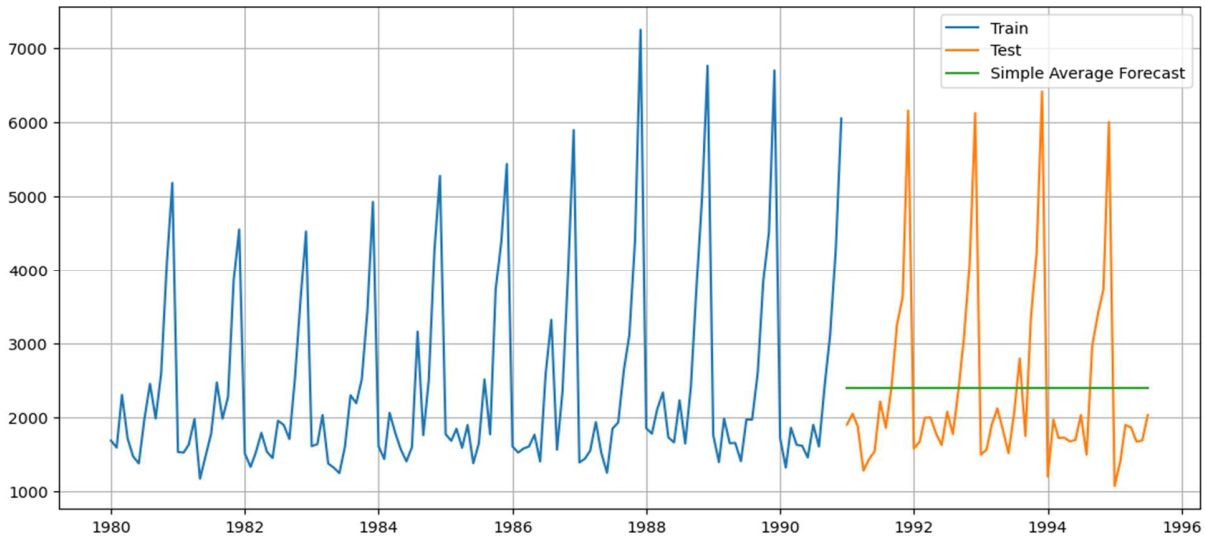


Figure38: Simple average plot

This model is giving a straight line as cannot account for seasonality.

Model Evaluation

```
For SimpleAverage forecast on the Test Data,  RMSE is 1275.08
```

Moving Average Model

We developed a model using the trailing moving average method, where predictions are made by calculating the average of a specified number of previous data points. In this case, we considered four different trailing moving averages: 2, 4, 6, and 9. For each model, we used the corresponding number of previous data points (2, 4, 6, and 9) to build separate models over the entire dataset and plotted the moving average values alongside the train and test data.

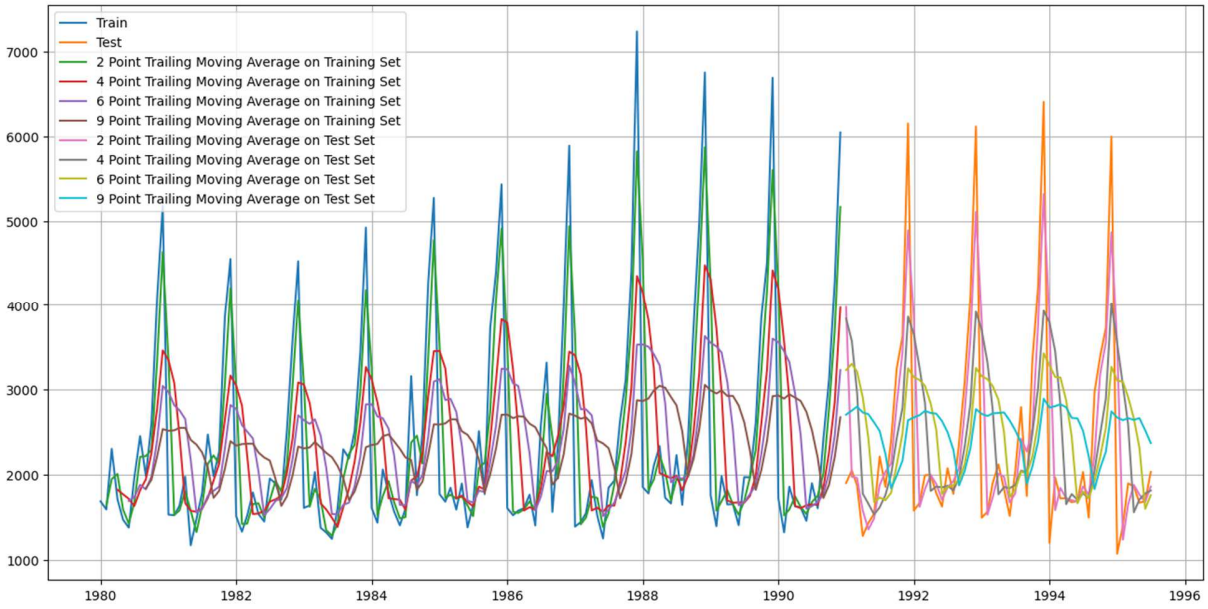


Figure39: Moving average plot

Model Evaluation

```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 813.401
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 1156.590
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 1283.927
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 1346.278
```

Exponential Model

This model uses exponential smoothening to make predictions it has three parts

1. Single Exponential Smoothening: It just levels the data.
2. Double Exponential Smoothening (Holt's Linear Method): It accounts for trend in data.

3. Triple Exponential Smoothening (Holt-Winter's Linear Method): It accounts for both trend and seasonality in data.

Single Exponential Smoothening Model

We have built a single exponential smoothening model using SimpleExpSmoothing which is a part of tsa.api module in statsmodels library, this model calculates the parameters for best leveling which came as:

```
{'smoothing_level': 0.03953488372093023,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1686.0,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table39: Best parameters

These parameters were used by model to make prediction on test data and predicted data was plotted with train and test data:

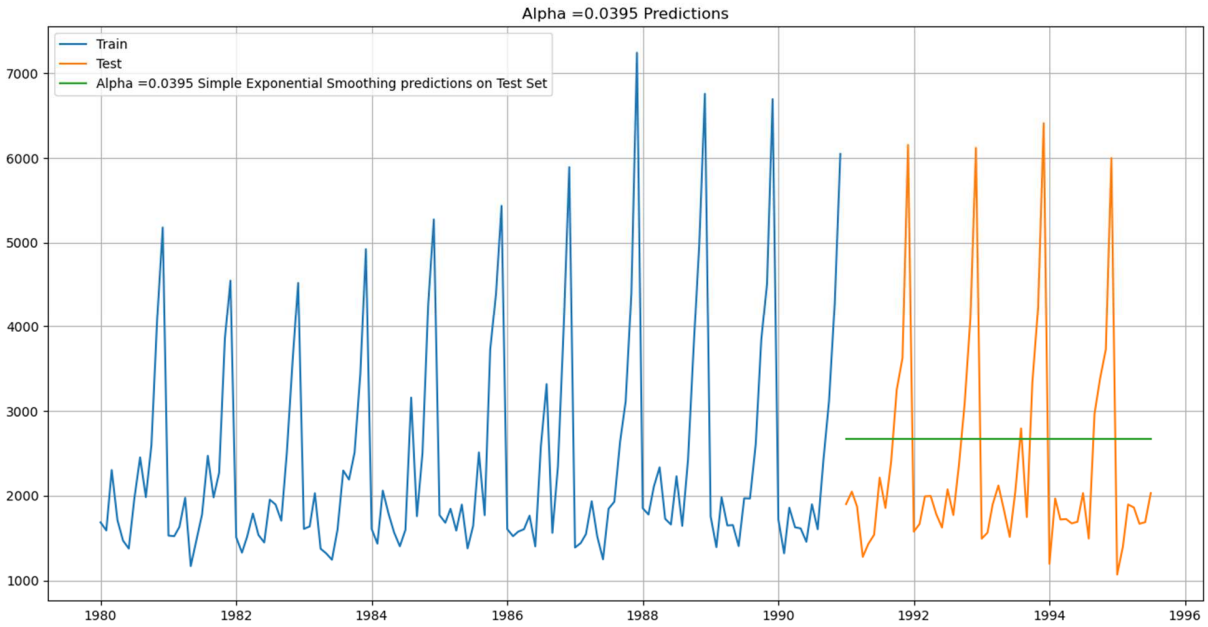


Figure40: Single exponential smoothening plot

Since this model uses levelling to predict the data without considering seasonality, we have a straight-line prediction.

Model Evaluation

```
For Alpha =0.0395 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1304.927
```

Double Exponential Smoothening

We have built a double exponential smoothening model using Holt which is a part of tsa.api module in statsmodels library, this model calculates the parameters for best leveling which came as:

```
{'smoothing_level': 0.6649999999999999,
 'smoothing_trend': 0.0001,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1502.1999999999991,
 'initial_trend': 74.87272727272739,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table40: Best parameters

These parameters were used by model to make prediction on test data and predicted data was plotted with train and test data:

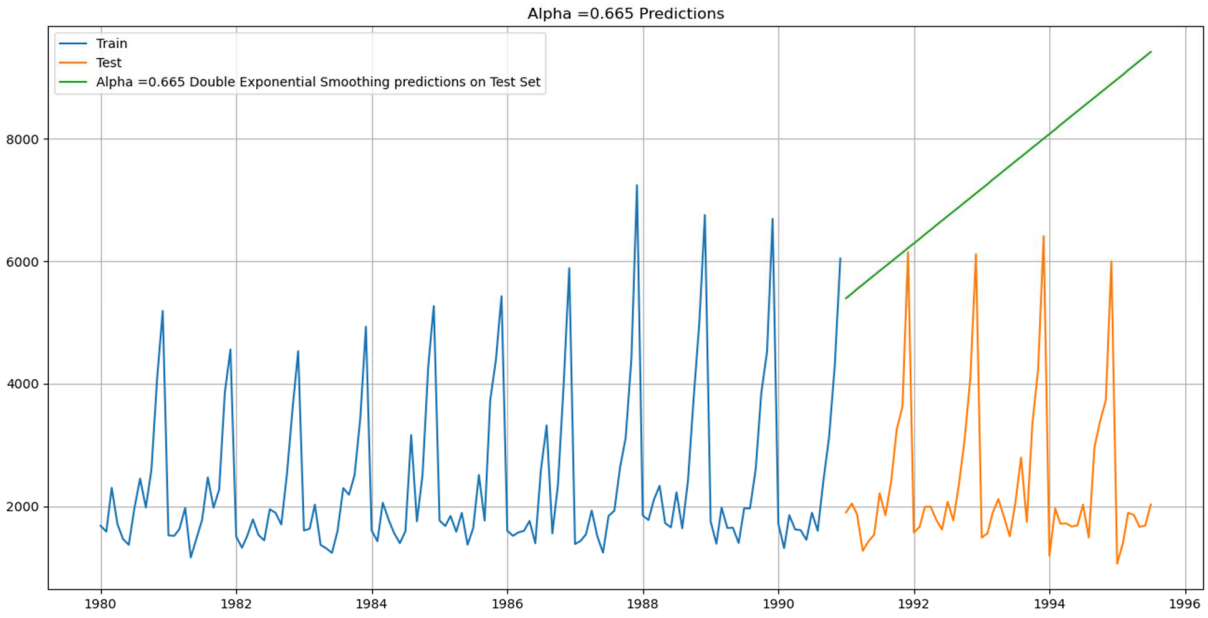


Figure41: Double exponential smoothening plot

Since this model accounts for trend to predict the data without considering seasonality, we have a slanting line prediction.

Model Evaluation

```
For Alpha =0.665 Double Exponential Smoothing Model forecast on the Test Data, RMSE is 5291.880
```

Triple Exponential Smoothening

We have built a triple exponential smoothening model using ExponentialSmoothing which is a part of tsa.api module in statsmodels library, this model calculates the parameters for best leveling which came as:

```
{'smoothing_level': 0.11127227248079453,
 'smoothing_trend': 0.012360804305088534,
 'smoothing_seasonal': 0.46071766688111543,
 'damping_trend': nan,
 'initial_level': 2356.577980956387,
 'initial_trend': -0.10243675533021725,
 'initial_seasons': array([-636.23319334, -722.9832009 , -398.64410813, -473.43045416,
       -808.42473284, -815.34991402, -384.23065038,   72.99484403,
       -237.44226045,  272.32608272, 1541.37737052, 2590.07692296]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Table41: Best parameters

These parameters were used by model to make prediction on test data and predicted data was plotted with train and test data:

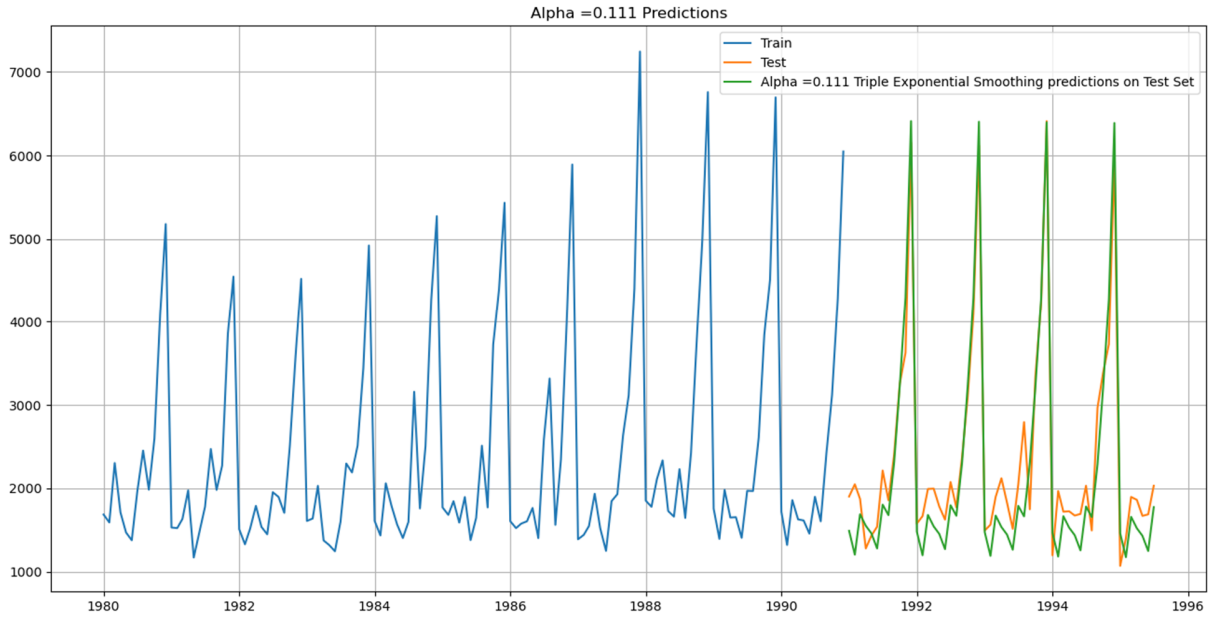


Figure42: Triple exponential smoothening plot

Since this model accounts for both trend and seasonality to predict the data, it is to a great extent following same pattern as the test data.

Model Evaluation

```
For Alpha =0.111 Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 378.951
```

1.6.6 Checking for Stationarity

Before applying ARIMA and SARIMA modelling we have to check if data is stationary as these models need stationary data for which we will do Dickey Fuller Test for which hypothesis are:

H0: Data is not stationary.

Ha: Data is stationary.

We performed the test and plotted it:

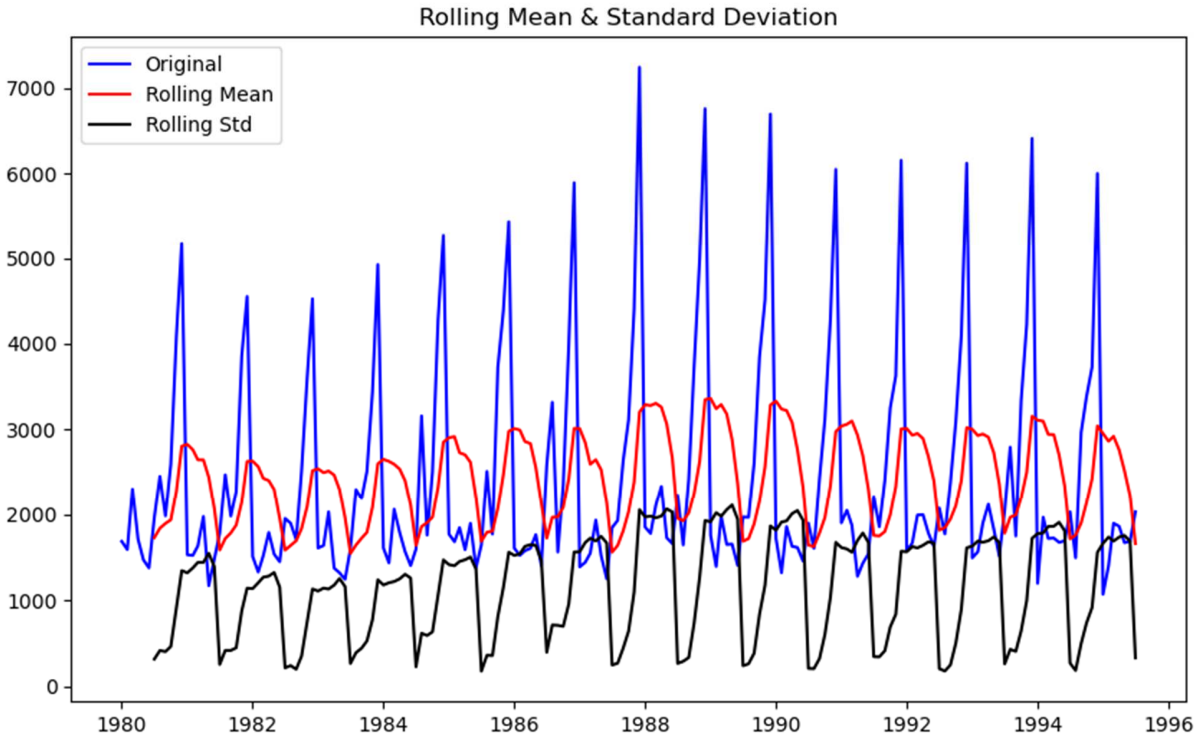


Figure43: Rolling mean and standard deviation plot

Result for this test were:

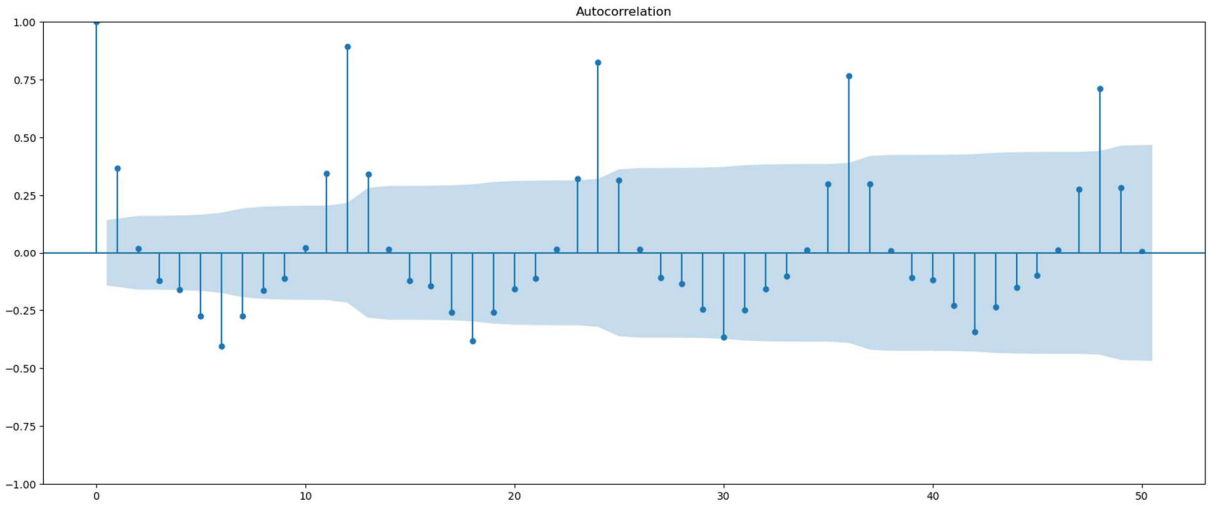
```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
Weak evidence against the null hypothesis, indicating the series is non-stationary.
```

Table42: Dickey-Fuller test

Since, p-value is greater than 0.05 means we cannot reject the null hypothesis that data is not stationary. We will have to make data stationary using differencing technique.

1.6.7 Making Data Stationary

Since the data exhibits seasonality, it's necessary to plot an autocorrelation plot to determine the appropriate seasonal differencing period. This will help in identifying the time lag where the seasonal patterns repeat, which is crucial for improving model accuracy.



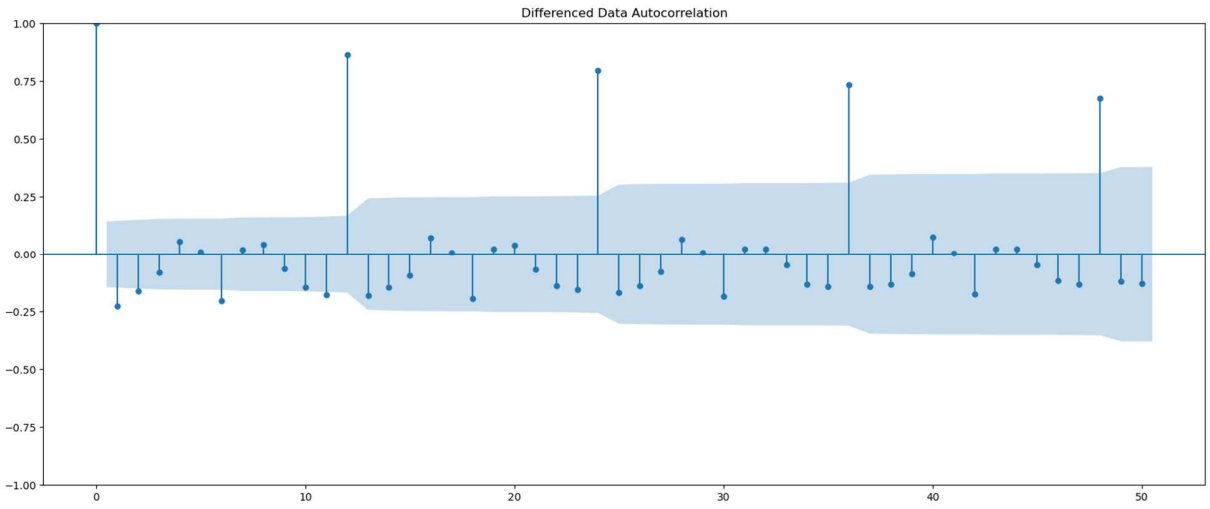


Figure44: Autocorrelation plot

From the above plot we can clearly identifying that seasonal patterns repeat after 12 months thus for differencing we will consider time lag of 12.

YearMonth	Sparkling	Seasonal_First_Difference
1980-01-01	1686	NaN
1980-02-01	1591	NaN
1980-03-01	2304	NaN
1980-04-01	1712	NaN
1980-05-01	1471	NaN
1980-06-01	1377	NaN
1980-07-01	1966	NaN
1980-08-01	2453	NaN
1980-09-01	1984	NaN
1980-10-01	2596	NaN
1980-11-01	4087	NaN
1980-12-01	5179	NaN
1981-01-01	1530	-156.0
1981-02-01	1523	-68.0

Table43: First seasonal differencing

After differencing with time lag of 12 we will again conduct dickey fuller test to check for stationarity.

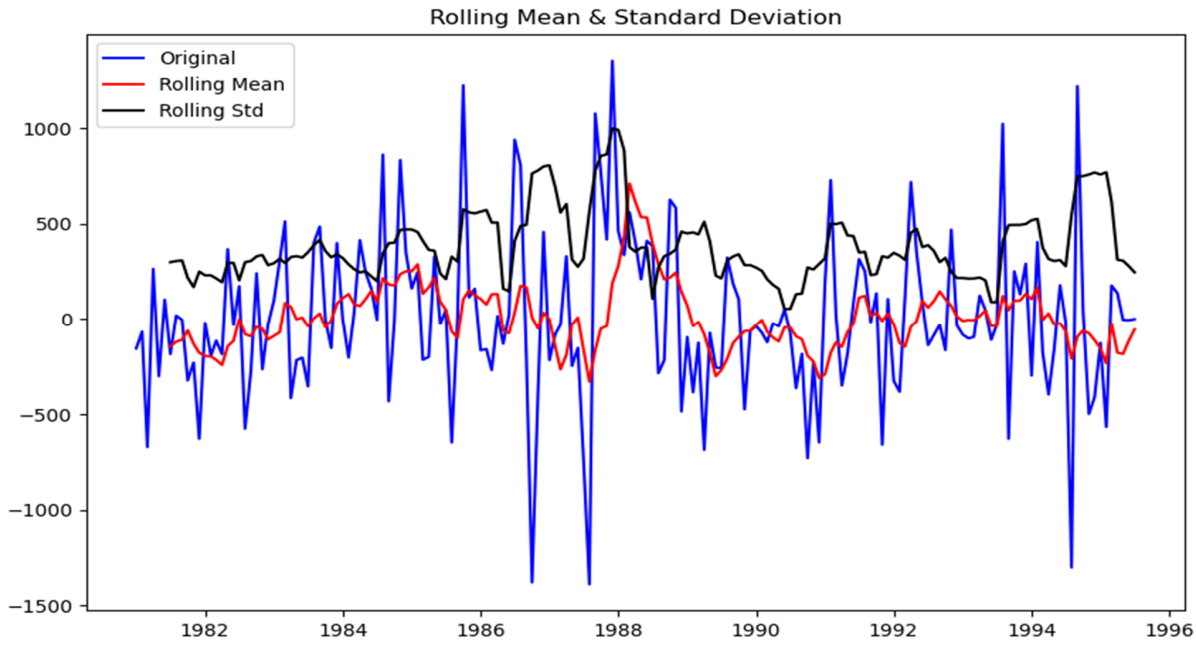


Figure45: Rolling mean and standard deviation plot

```
Results of Dickey-Fuller Test:
Test Statistic                  -4.460165
p-value                          0.000232
#Lags Used                      11.000000
Number of Observations Used    163.000000
Critical Value (1%)             -3.471119
Critical Value (5%)             -2.879441
Critical Value (10%)            -2.576314
dtype: float64
Strong evidence against the null hypothesis (H0), reject the null hypothesis. The series is stationary.
```

Table44: Dickey-Fuller test

The data has become stationary now we can build ARIMA and SARIMA models on this stationary data. However, these models require moving average denoted by q, auto correlation denoted by p and differencing denoted by d as parameters here we have been able extract d value where d = 1 as d is the number of times differencing is done to make data stationary and in this case by first seasonal differencing we have a stationary data. Now to find p and q values we will have to plot Autocorrelation and Partial Autocorrelation Function.

1.6.8 Plot for Autocorrelation and Partial Autocorrelation Function

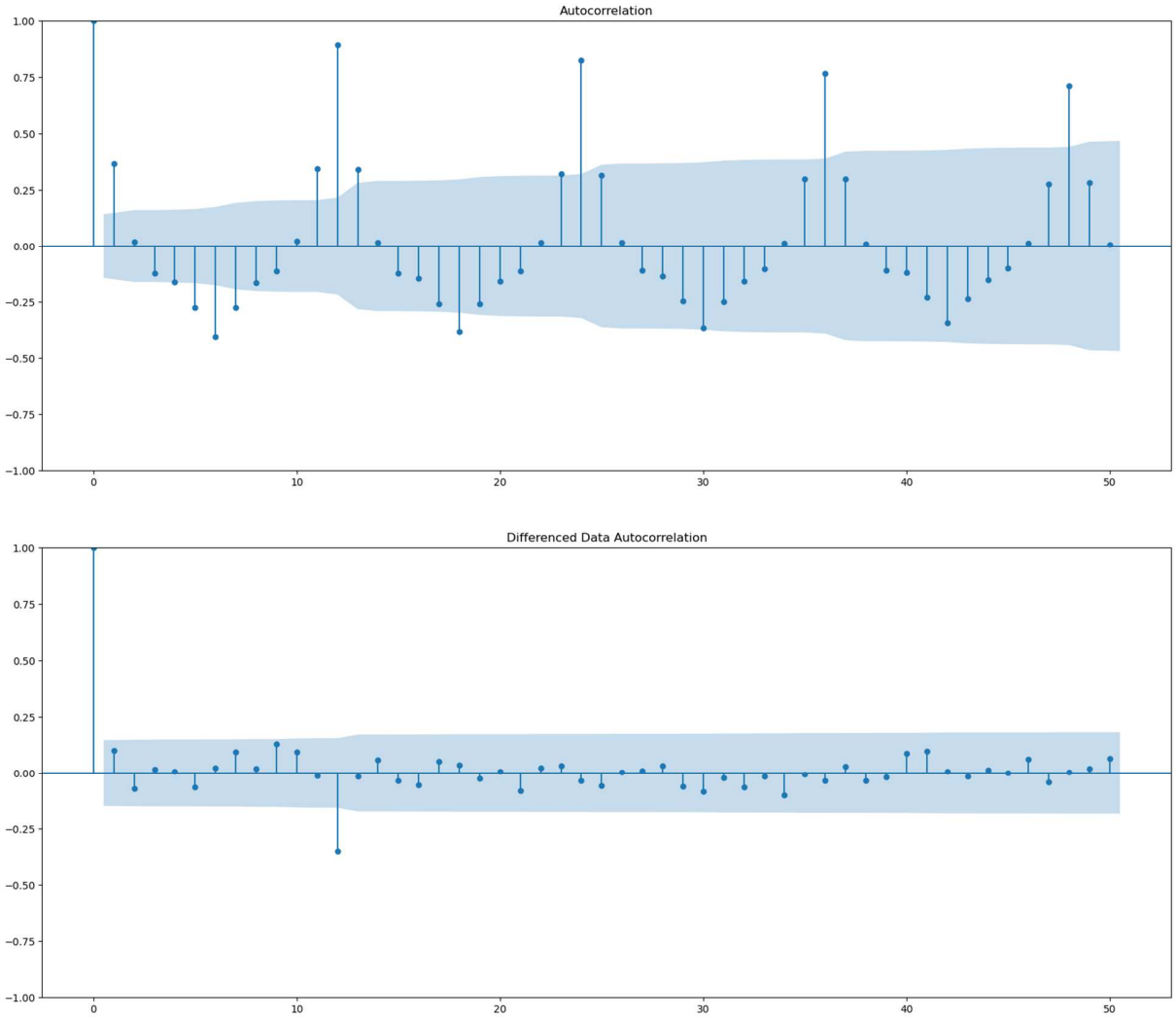


Figure46: Autocorrelation plot

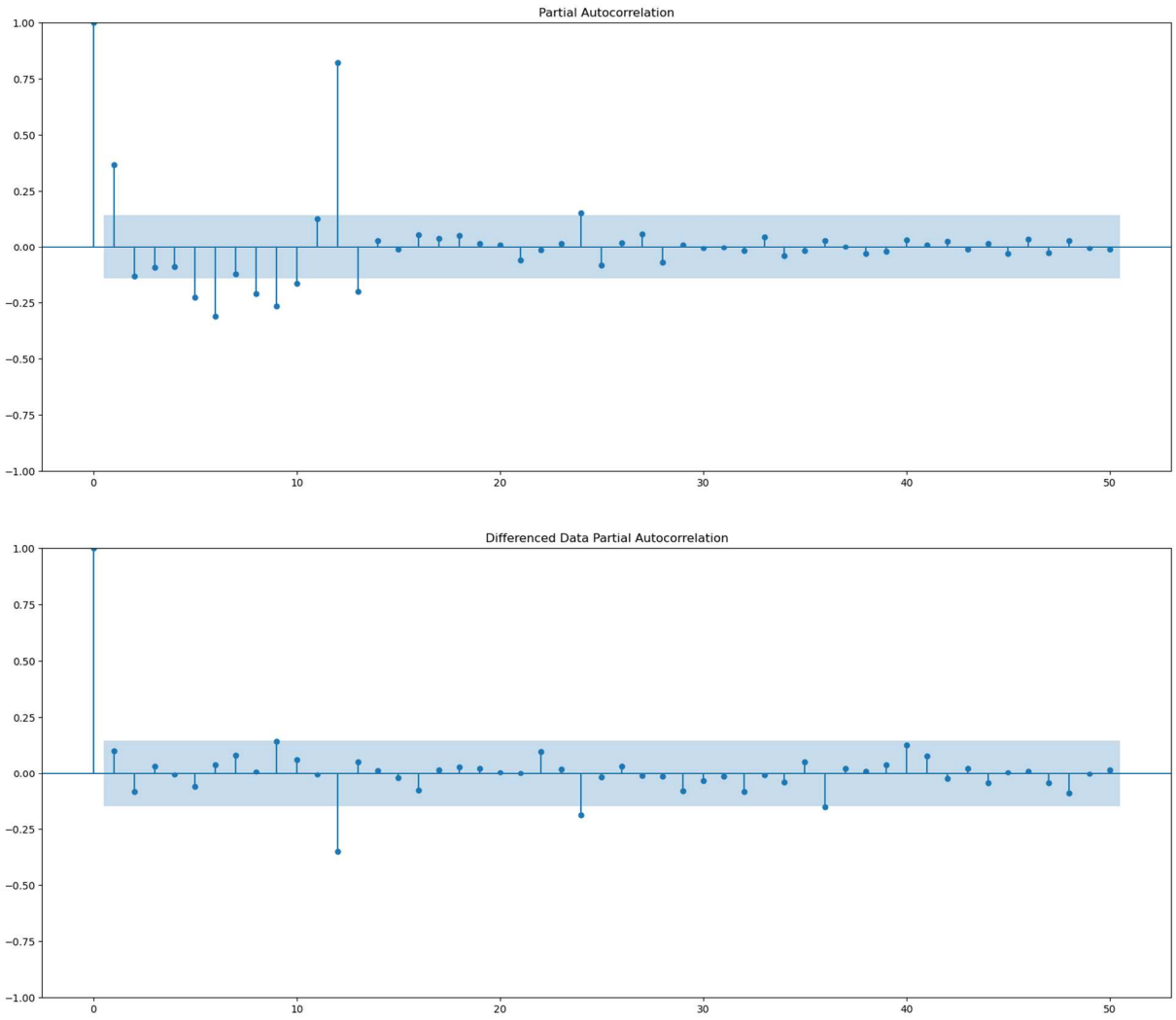


Figure47: Autocorrelation plot

Autocorrelation and Partial Autocorrelation Function are used to calculate optimum values of auto regression (AR) and moving average (MA) represented by p and q respectively. From the above plot we can conclude that optimum value of p should be 0 and q should be 1.

1.6.9 Model Building - Stationary Data

ARIMA Modelling

Though we have calculated optimum value of p and q using ACF and PACF plot, however, for building model we will be taking a range of p and q and calculate optimum values of p and q using auto ARIMA model.

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

Table45: Parameter combinations

Using these combinations, we ran the ARIMA model to calculate the AIC value of each combination. Best performing combinations were:

	param	AIC
10	(2, 1, 2)	2213.509212
15	(3, 1, 3)	2221.458953
14	(3, 1, 2)	2230.783429
11	(2, 1, 3)	2232.934772
9	(2, 1, 1)	2233.777626

Table46: Best performing parameters

In time series the best combination of p, d and q is the one with lowest AIC value for which we have sorted the above table in ascending order of AIC value based on which we can conclude that the best set of parameters are p =2, d =1 and q =2. Now we will manually build an ARIMA model using these best parameters.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Sun, 22 Sep 2024	AIC	2213.509
Time:	14:37:19	BIC	2227.885
Sample:	01-01-1980	HQIC	2219.351
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.781	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.741	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.218	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.109	0.000	0.785	1.215
sigma2	1.099e+06	1.99e-07	5.51e+12	0.000	1.1e+06	1.1e+06

Ljung-Box (L1) (Q):	0.19	Jarque-Bera (JB):	14.46
Prob(Q):	0.67	Prob(JB):	0.00
Heteroskedasticity (H):	2.43	Skew:	0.61
Prob(H) (two-sided):	0.00	Kurtosis:	4.08

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 7.7e+28. Standard errors may be unstable.

Table47: ARIMA result

Model Evaluation

For ARIMA(2, 1, 2) Model with forecast on the Test Data, RMSE is %3.3f 1299.9796651035053

SARIMA Modelling

In SARIMA models in addition to trend we also account for seasonality and to understand the seasonal parameter we have to use ACF plot.

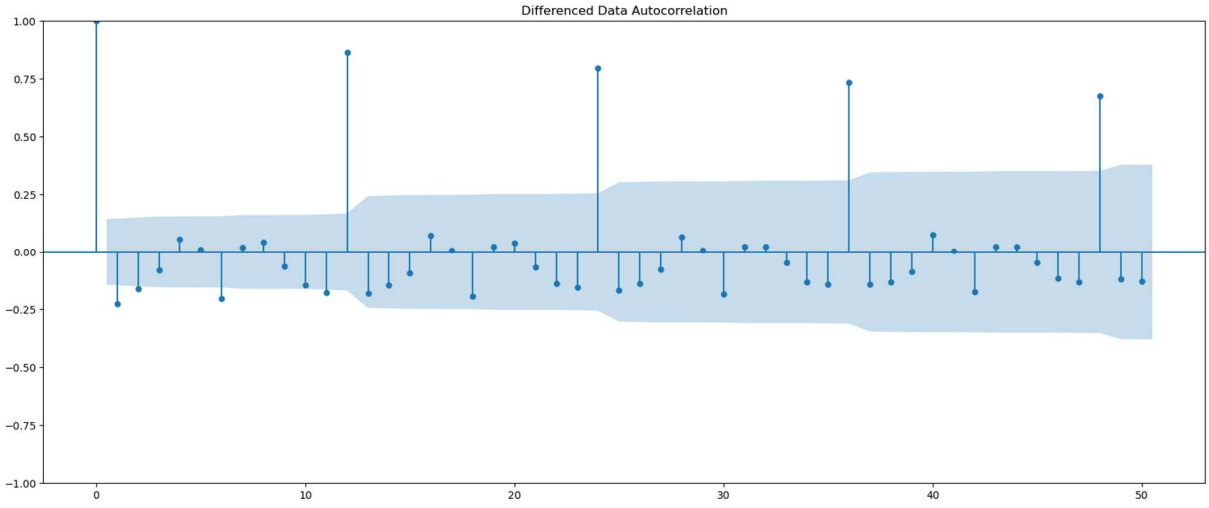


Figure48: Autocorrelation plot

We see that there can be a seasonality of 12. We will run our auto SARIMA models by setting as 12.

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

Table48: Parameter combinations

Using these combinations, we ran the SARIMA model to calculate the AIC value of each combination. Best performing combinations were:

	param	seasonal	AIC
115	(1, 1, 3)	(0, 0, 3, 12)	16.000000
252	(3, 1, 3)	(3, 0, 0, 12)	1387.497016
220	(3, 1, 1)	(3, 0, 0, 12)	1387.788331
237	(3, 1, 2)	(3, 0, 1, 12)	1388.602617
221	(3, 1, 1)	(3, 0, 1, 12)	1388.681484

Table49: Best performing parameters

From the above table which is sorted in ascending order of AIC score we can take the best values of parameters at (0, 1, 2) (2, 0, 2, 12). We will build the final SARIMA model using these parameters.

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(1, 1, 3)x(0, 0, 3, 12)	Log Likelihood	0.000
Date:	Sun, 22 Sep 2024	AIC	16.000
Time:	19:06:23	BIC	36.087
Sample:	01-01-1980	HQIC	24.104
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	6.3599	-0	-inf	0.000	6.360	6.360
ma.L1	2.4977	-0	-inf	0.000	2.498	2.498
ma.L2	-10.2544	1.287	-7.969	0.000	-12.777	-7.732
ma.L3	-7.9773	-0	inf	0.000	-7.977	-7.977
ma.S.L12	6.114e+13	-0	-inf	0.000	6.11e+13	6.11e+13
ma.S.L24	2.155e+14	-0	-inf	0.000	2.16e+14	2.16e+14
ma.S.L36	5.529e+13	-0	-inf	0.000	5.53e+13	5.53e+13
sigma2	1.614e+06	-0	-inf	0.000	1.61e+06	1.61e+06

Ljung-Box (L1) (Q):	nan	Jarque-Bera (JB):	nan
Prob(Q):	nan	Prob(JB):	nan
Heteroskedasticity (H):	nan	Skew:	nan
Prob(H) (two-sided):	nan	Kurtosis:	nan

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number inf. Standard errors may be unstable.

Table50: SARIMA result

On building the model using best parameters we are getting a warning stating that condition number is infinite meaning there is an issue with the accuracy and reliability of the model's parameter estimates due to numerical instability which would make this model unstable thus we will rebuild the model taking second best set of parameters which are (3, 1, 3) (3, 0, 0, 12).

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(3, 1, 3)x(3, 0, [], 12)	Log Likelihood	-683.749
Date:	Sun, 22 Sep 2024	AIC	1387.497
Time:	19:06:30	BIC	1412.715
Sample:	01-01-1980 - 12-01-1990	HQIC	1397.675
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6747	0.142	-11.819	0.000	-1.952	-1.397
ar.L2	-0.7438	0.258	-2.886	0.004	-1.249	-0.239
ar.L3	-0.0025	0.144	-0.017	0.986	-0.285	0.280
ma.L1	1.0551	0.191	5.514	0.000	0.680	1.430
ma.L2	-0.7782	0.172	-4.521	0.000	-1.116	-0.441
ma.L3	-0.9066	0.147	-6.147	0.000	-1.196	-0.618
ar.S.L12	0.5329	0.118	4.520	0.000	0.302	0.764
ar.S.L24	0.2786	0.116	2.399	0.016	0.051	0.506
ar.S.L36	0.2392	0.102	2.352	0.019	0.040	0.439
sigma2	1.527e+05	1.93e-06	7.9e+10	0.000	1.53e+05	1.53e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	4.32
Prob(Q):	0.93	Prob(JB):	0.12
Heteroskedasticity (H):	1.26	Skew:	0.30
Prob(H) (two-sided):	0.52	Kurtosis:	3.88

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.56e+26. Standard errors may be unstable.

Table51: SARIMA result

For these set of parameters, we plot a residual plot to check whether we have been able to extract all the trend and seasonality from data.

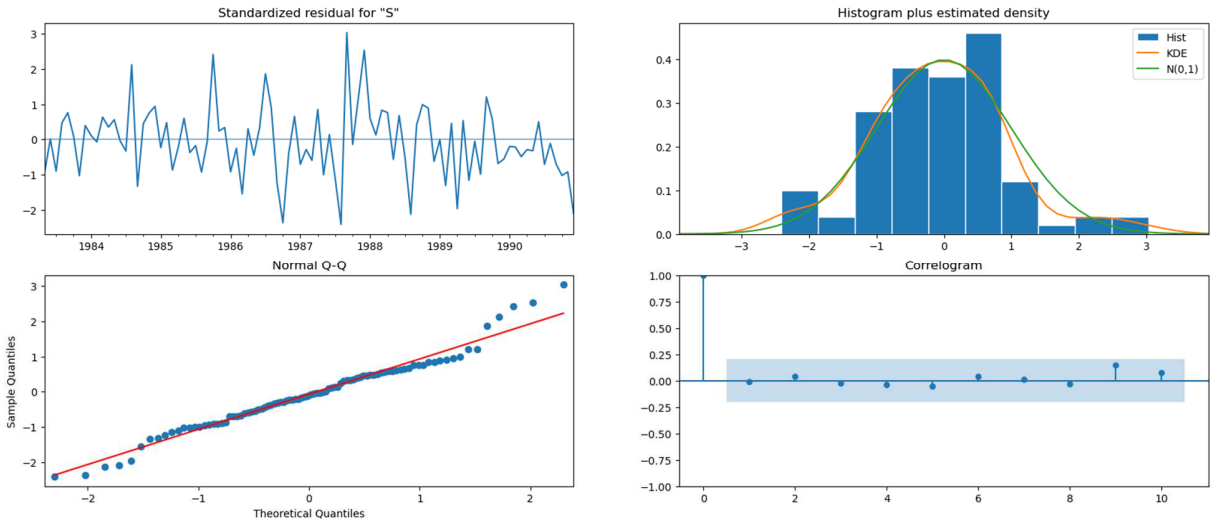


Figure49: Residual plot

In the above diagnostic plot residuals are randomly scattered meaning we have optimized the model now we can move ahead to evaluate it.

Model Evaluation

```
For SARIMA(3, 1, 3) (3, 0, 0, 12) Model with  forecast on the Test Data, RMSE is %3.3f 611.4731936284929
```

1.6.10 Model Comparison

We have created 11 models using different techniques compared each model's performance for test data using key metrics and have found that all the models are stable now we will compare these models with each other to find the best model based on their RMSE score for test data.

	Test RMSE
Alpha=0.111,TripleExponentialSmoothing	378.951023
SARIMA(3,1,3)(3,0,0,12)	611.473194
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
SimpleAverage_Forecast	1275.081804
RegressionOnTime	1275.867052
6pointTrailingMovingAverage	1283.927428
ARIMA(2,1,2)	1299.979665
Alpha=0.0395,SimpleExponentialSmoothing	1304.927405
9pointTrailingMovingAverage	1346.278315
Alpha=0.665,DoubleExponentialSmoothing	5291.879833

Table52: Model comparison

From the above table we can conclude that Triple Exponential Smoothing Model (Holt-Winter's Linear Method) is the best performing model as it has the lowest RMSE score. For forecasting we will use this model.

1.6.11 Building the most optimum model on the Full Data

We rebuilt the Triple Exponential Smoothing model on the entire dataset as the final model and calculate the RMSE to evaluate the model performance.

Model Evaluation

```
For Alpha=0.111,TripleExponentialSmoothing model forecast on the full Data,  RMSE is %3.3f 356.96820003228873
```

Forecasting 12 months into the future

	prediction	lower_CI	upper_CI
1995-08-31	1860.790317	1084.432878	2788.624368
1995-09-30	2470.931957	1694.574517	3398.766008
1995-10-31	3200.128073	2423.770633	4127.962124
1995-11-30	3806.537212	3030.179772	4734.371263
1995-12-31	5967.647889	5191.290450	6895.481940
1996-01-31	1224.569165	448.211726	2152.403216
1996-02-29	1600.114618	823.757178	2527.948669
1996-03-31	1861.874819	1085.517379	2789.708870
1996-04-30	1845.012697	1068.655258	2772.846748
1996-05-31	1681.565864	905.208425	2609.399915
1996-06-30	1635.125010	858.767571	2562.959061
1996-07-31	1993.194193	1216.836753	2921.028244

Table53: Confidence interval

The table above shows the predicted sales for the upcoming months, along with a confidence interval. We estimate, with 95% confidence, that the actual sales will fall within this specified range.

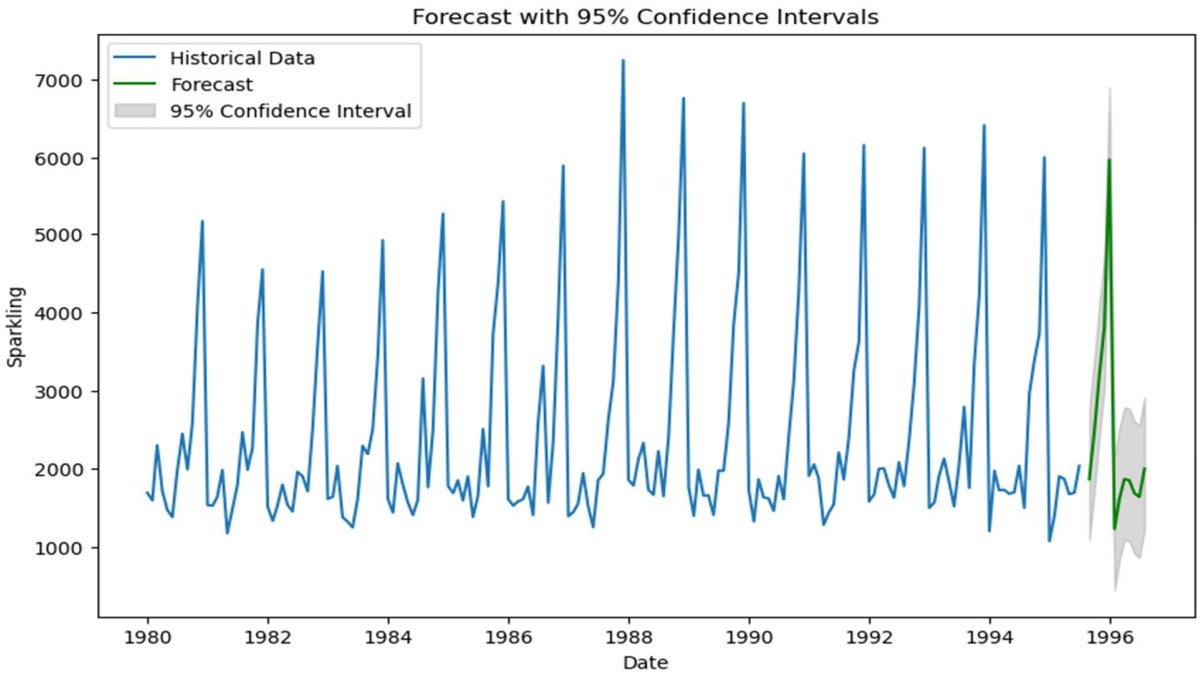


Figure50: Forecast plot

The plot above illustrates the forecasted sales, represented by the orange line, for the next 12 months along with the historic data. The shaded region around indicates the range within which sales are expected to fall with 95% confidence.

1.7 Conclusion

ABC Estate Wines provided historical sales data for their Rosé and Sparkling wines. After a detailed analysis, we were able to uncover the following key insights.

Key Insights

1. In wine sales for both variants there was clear seasonality where January has the lowest sales and December has highest sales, additionally, throughout the year sales tend to consistently peaking in the month of December which coincides with the holiday season.
2. Variant specific insights:

➢ For Rose Wine:

- Sales of Rosé wine have experienced a steady decline over the years. In December 1980, approximately 260 units were sold, but by 1994, this number had dropped to around 80 units, which is even lower than the lowest sales month in 1980.

➢ For Sparkling Wine:

- The demand for Sparkling wine has remained almost consistent for this entire time period.
- While year-on-year sales data for sparkling wine shows some fluctuations, these are likely influenced by factors beyond the general trend. A separate analysis could be conducted to better understand the causes of these fluctuations. This would be crucial in improving the forecasting model's accuracy, especially since sparkling wine sales have consistently been higher than those of Rosé, and even more so now as the company faces a decline in Rosé sales.

3. Using the provided data, we built a forecasting model that predicts future wine sales by accounting for both trend and seasonality. After testing various models, we selected the 2-point moving average model as the best-performing model for Rosé wine, and the Triple Exponential Smoothing Model (Holt-Winter's Linear Method) for Sparkling wine.

Business Recommendations

1. The demand shows a increase from January to December, allowing ABC Wine Estate to manage inventory accordingly. In particular, demand rises significantly in December, making advanced planning and maintaining adequate stock level is crucial. This proactive approach will enable the company to fully capitalize on the peak sales season, maximizing revenue and minimizing the risk of stockouts.
2. The Rose Wine variant has experienced a steady decline in sales. To maintain a competitive edge in the wine industry, ABC Estate Wine is strongly advised to investigate the reasons behind this decline. It is essential to determine whether the decline is a broader industry trend or specific to the company. Based on these insights, necessary corrective actions should be taken to address the issue and enhance future performance.

3. Data for two wine variants were provided, showing that while sales for one variant have seen a steady decline, the other has remained relatively unchanged over a span of 15 years. This indicates that business growth is at best stagnant for these specific variants. If ABC Estate Wines produces other wine variants, it is recommended that similar analysis projects be conducted for those as well. This will enable the company to better understand its overall market position and plan its future course of action effectively.