greatlearning
Power Ahead

BUSINESS REPORT ON

TIME SERIES FORECASTING

SPARKLING SALES DATASET

DATA SCIENCE AND BUSINESS ANALYTICS

By: Sanjam Preet Singh Bhullar

May 2023

Table of Contents

List of Figures

List of tables

PROBLEM STATEMENT

For this particular assignment, the data of Sparkling wine sales in the 20th century is to be analysed. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast wine sales in the 20th century.

1. Read the data as an appropriate time series data and plot the data.

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

Table 1: First 5 rows of dataset

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

Table 2: Last 5 rows of dataset

The Sparkling dataset shows wine sales at the end of each month.

We have the data starting from January 1980 to July 1995 – a duration of 14.5 years.

```
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
```

There is **one column** in the dataset, as has been seen in tables 1 and 2.

The dataset has **187 records** of Sparkling wine sales.

There are **no null values**.

Plotting the time series

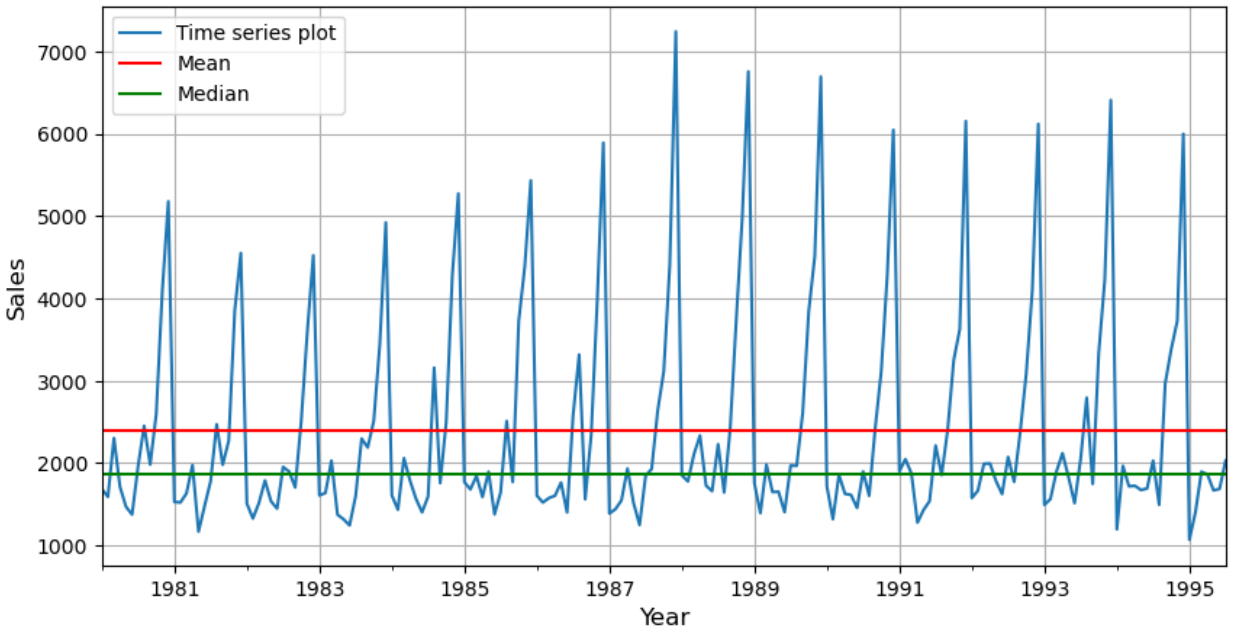

Figure 1: Time series plot

It can be seen from the plot that the **Sparkling wine sales have seasonality**. In other words, sales increase and decrease after a fixed time interval. The figure shows that sales are low at the start of a year and pick up pace at the end of the year. And this is the case every year.

The time series does not have a pronounced trend. From 1981 to 1988, the sales increase gradually. After 1988, the sales decrease slowly. In short, the sales increase and then decrease gradually.

The mean sales is more than the median sales.

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

Yearly sales

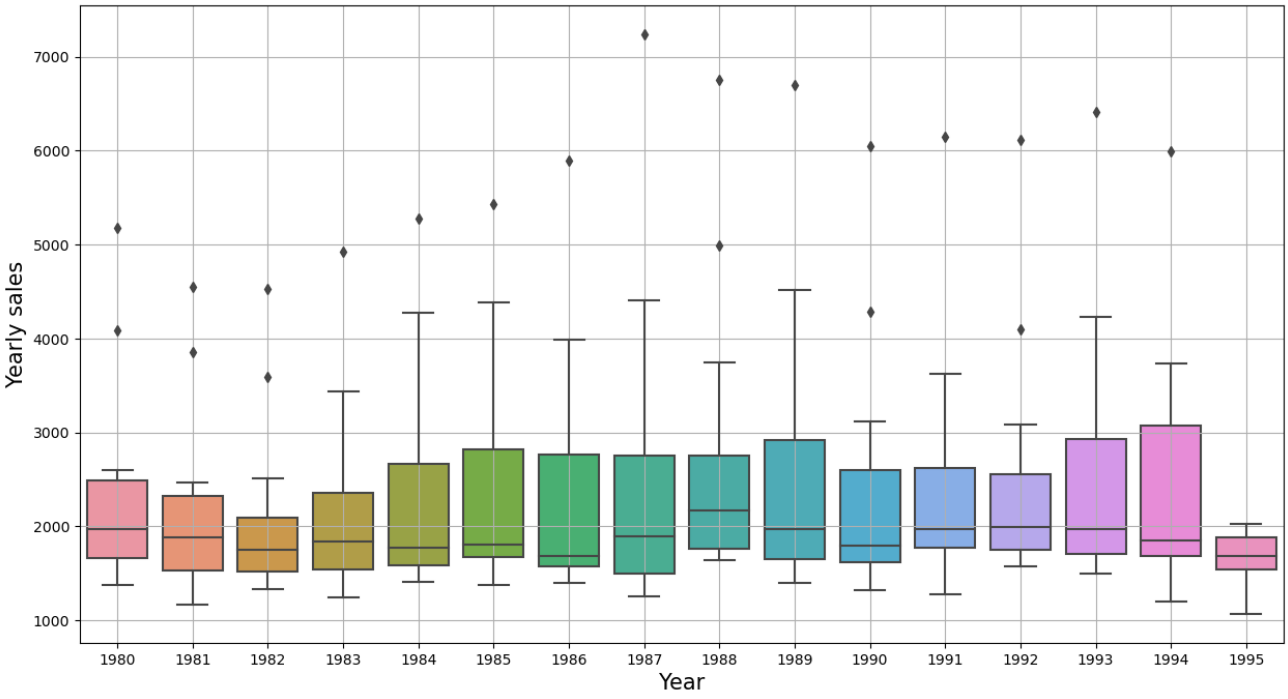

Figure 2: Yearly sale boxplot

The median yearly sales are low at the start of the 1980s. Sales increase in the late 1980s and fall again the 1990s.

This plot shows the mean sales of wine across years. The sales peak in 1988 and fall drastically in the subsequent years.

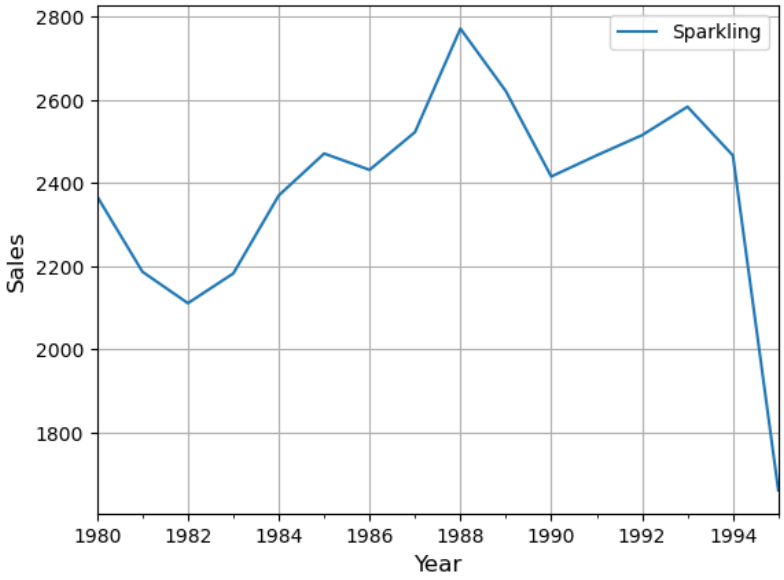

Figure 3: Yearly sale line plot

Quarterly sales

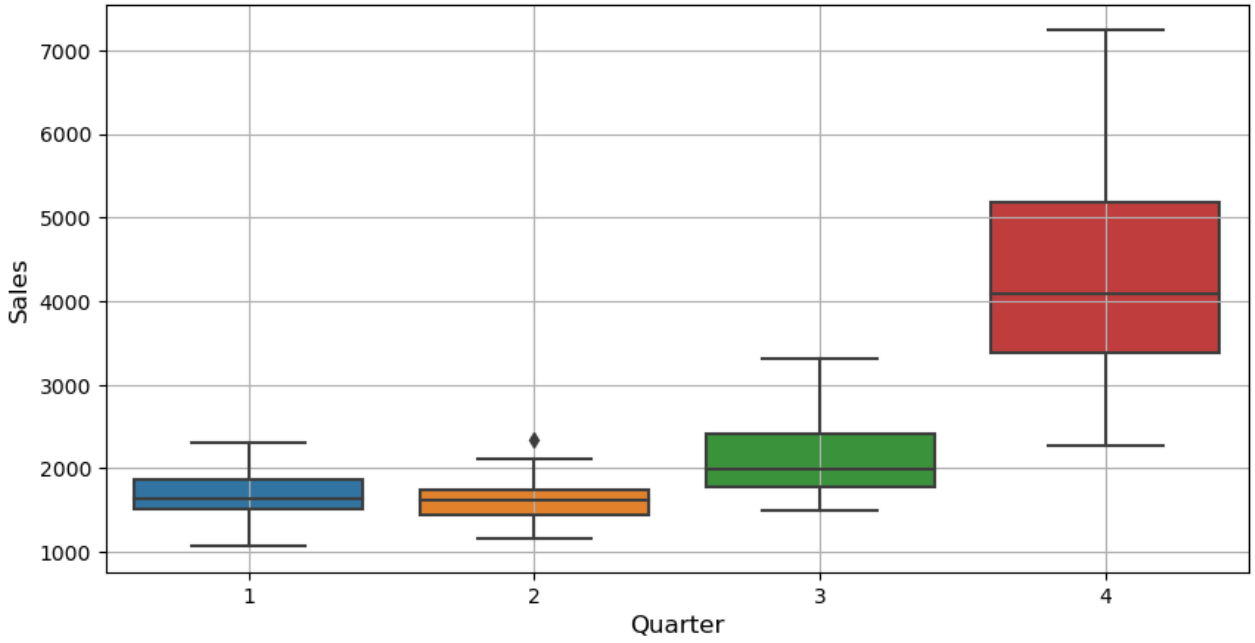

Figure 4: Quarterly sale boxplot

The median sales is the highest in quarter four and the lowest in quarter two.

Monthly sales

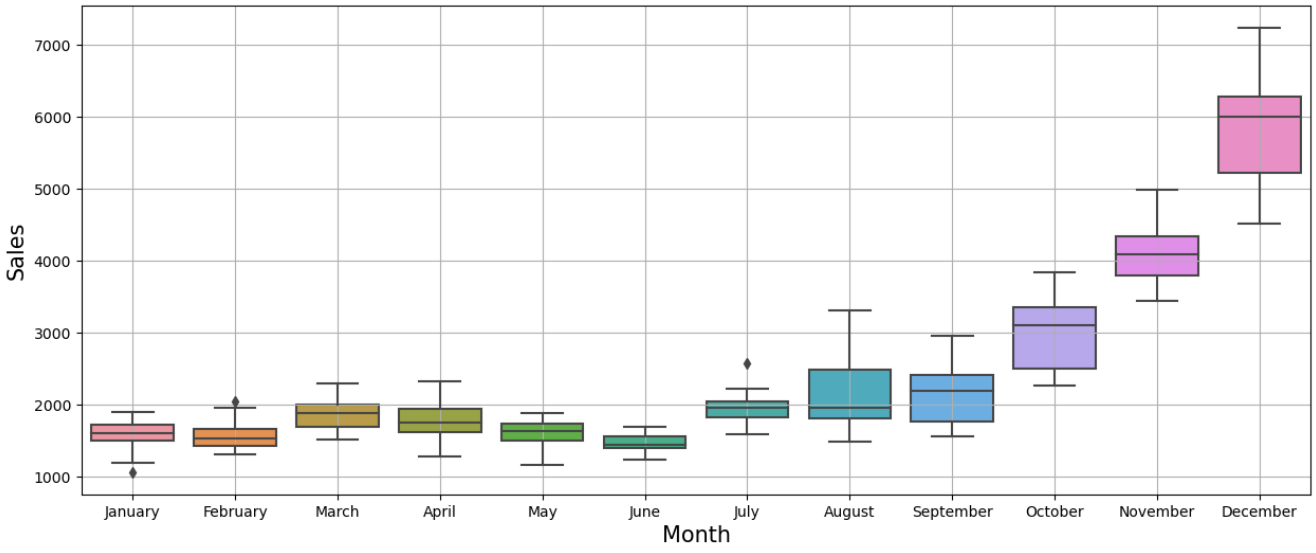

Figure 5: Monthly sale boxplot

The sales remain low at the start of the year, but pick up pace in the subsequent months. The last three months record high sales, with the highest being in December.

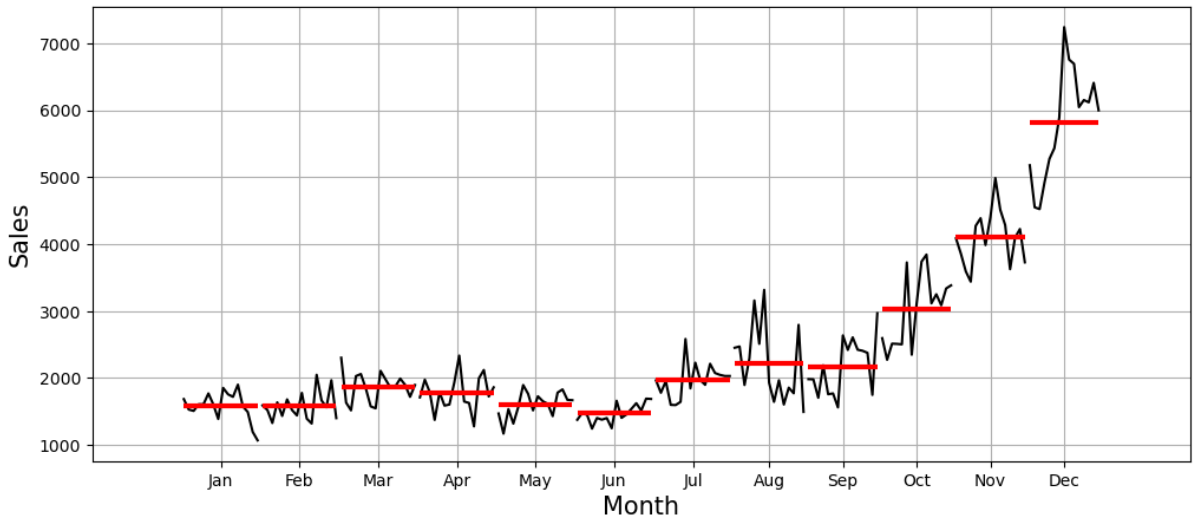

Figure 6: Month plot

This graph reiterates the observation drawn from Figure 5. **Sales are the highest in December**. The red lines denote median sales, which remain low in the initial months of a year and pick up pace in the last three months of the year.

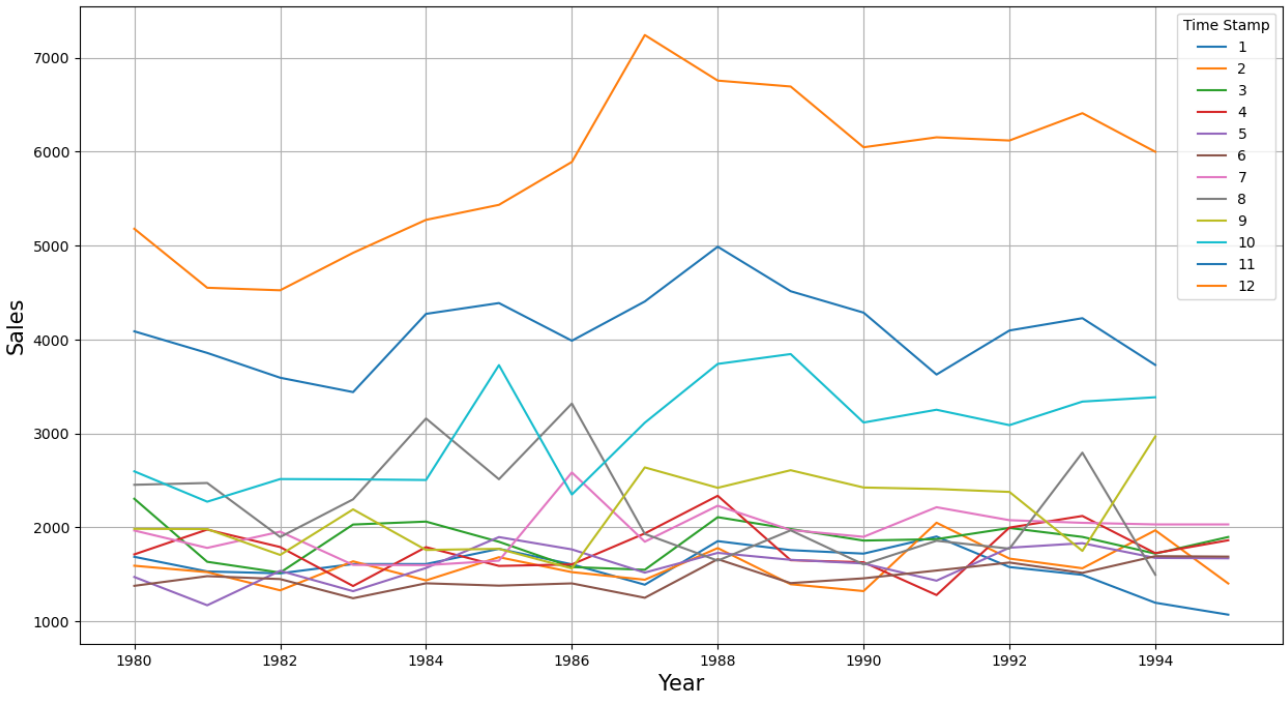

Figure 7: Monthly line plot

The line plot shows month-wise sales across years. Sales in December surpass the sales in other months.

Day-wise sales

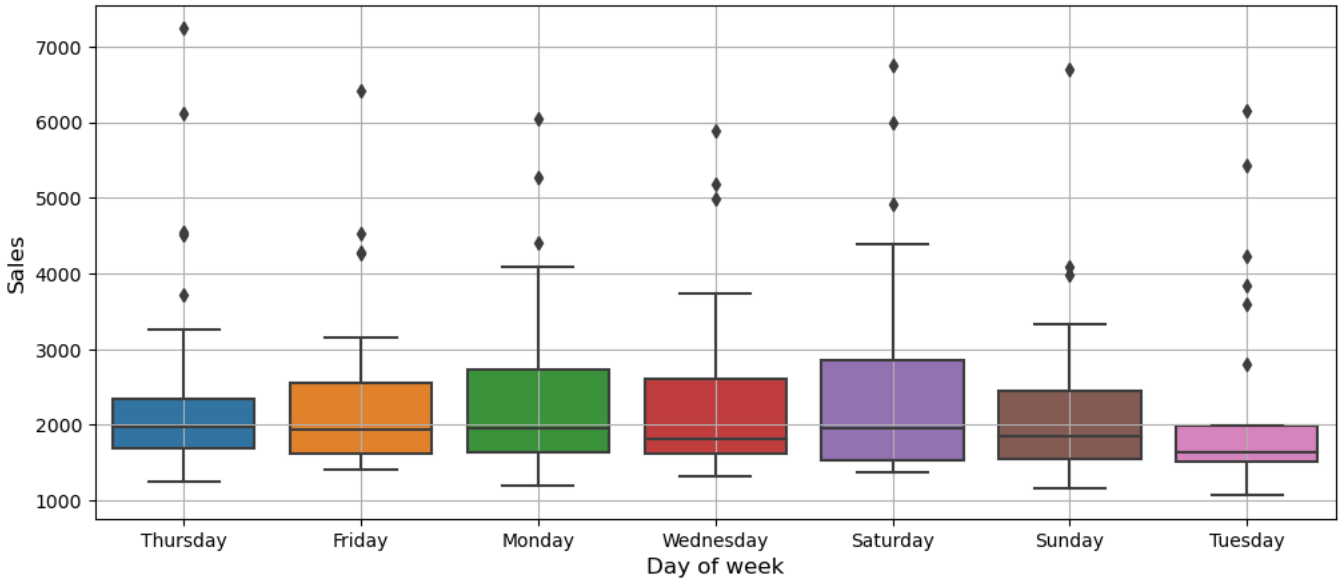

Figure 8: Day-wise sales boxplot

Median sales on four days of any week – Monday, Thursday, Friday and Saturday – remain the same. It is the **lowest on Tuesday**. Surprisingly, Sunday sales are lower than that of Saturday's. There are outliers on all days.

Empirical Cumulative Distribution Function (ECDF)

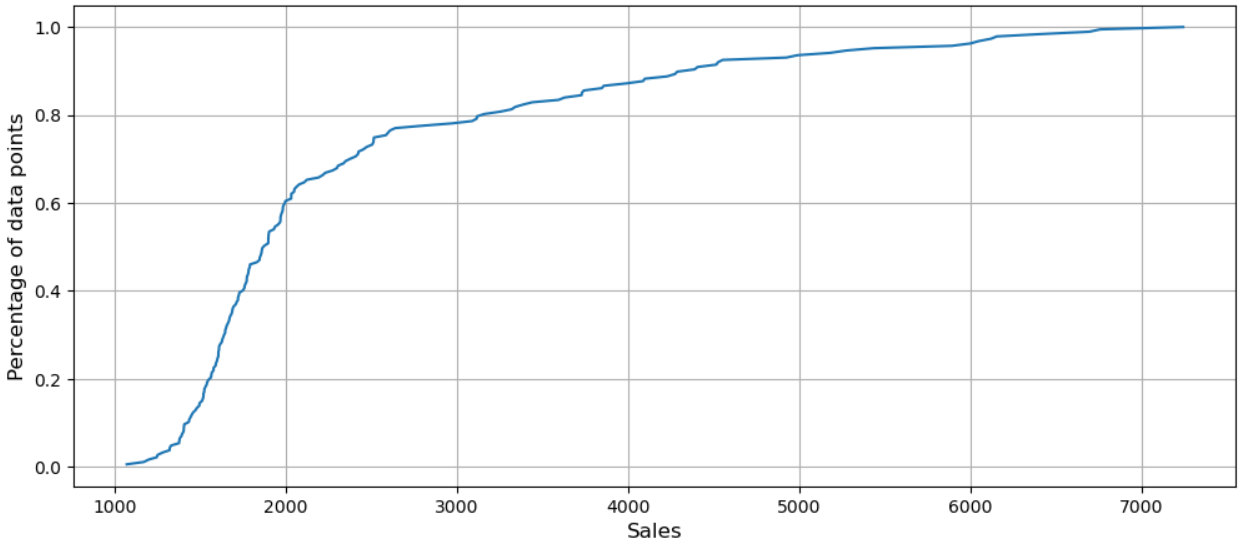

Figure 9: ECDF plot

The ECDF plot shows that 60 per cent of the data points have sale values up to 2,000 and 80% of the data points have sale values up to 3,000.

Average sales

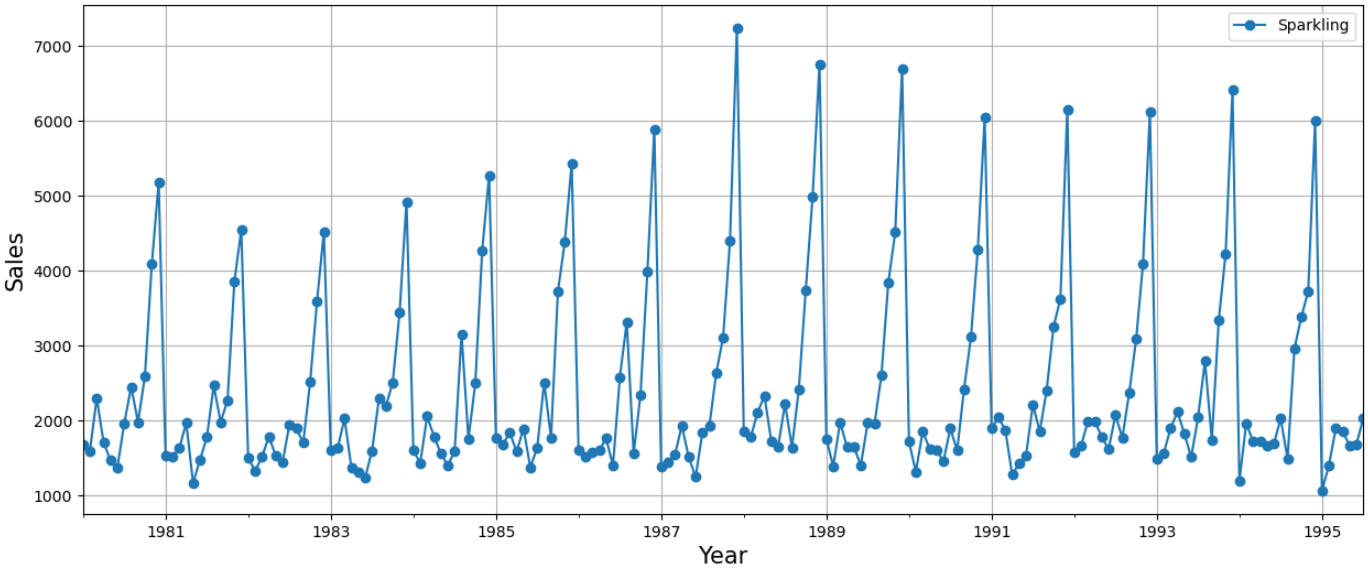

Figure 10: Average sales plot

Starting from 1982, average sales increased gradually till 1988, when Sparkling wine recorded the highest sales. In the subsequent years, it started falling only to increase later.

Sales percentage change

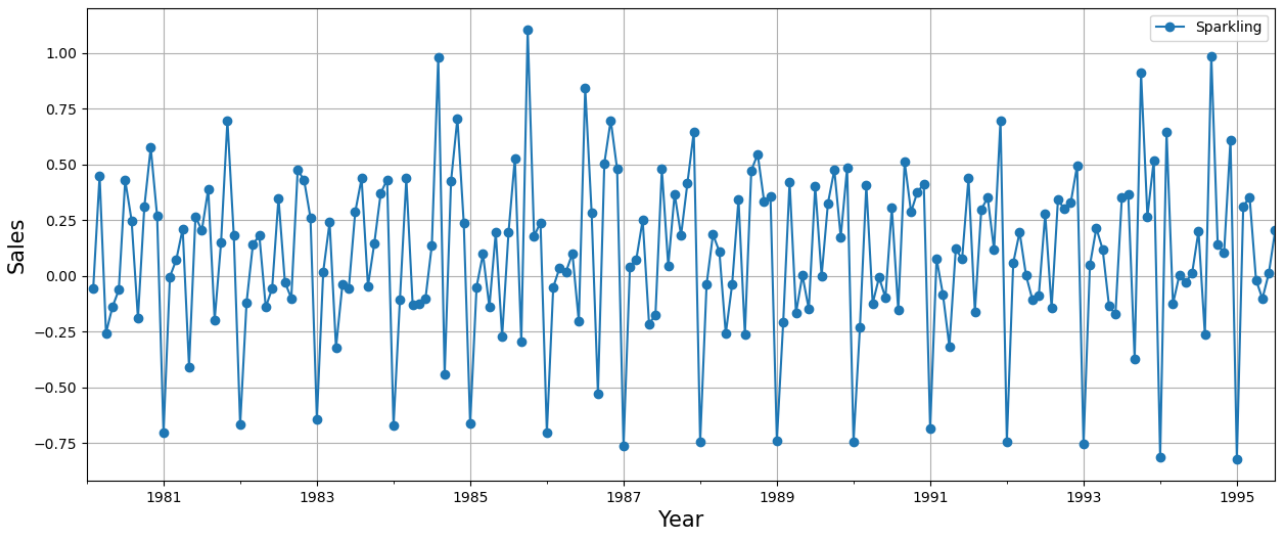

Figure 11: Sales percentage change plot

The graph shows month-on-month percentage change in sales. The slight trend that was visible in Figure 10 is not present in Figure 11.

Additive decomposition of time series

In the additive model, the time series is the sum of three components – trend, seasonality and residual. Trend and seasonality are systemic components, while residual is an irregular component. These two components are interpretable and can be estimated.

Trend shows the long-term movement of the time series, while **seasonality denotes the intra-year fluctuations** that are repeated over the entire length of the time series.

Residual is the "noise" element – something that happened in the past and is not expected to happen in the future.

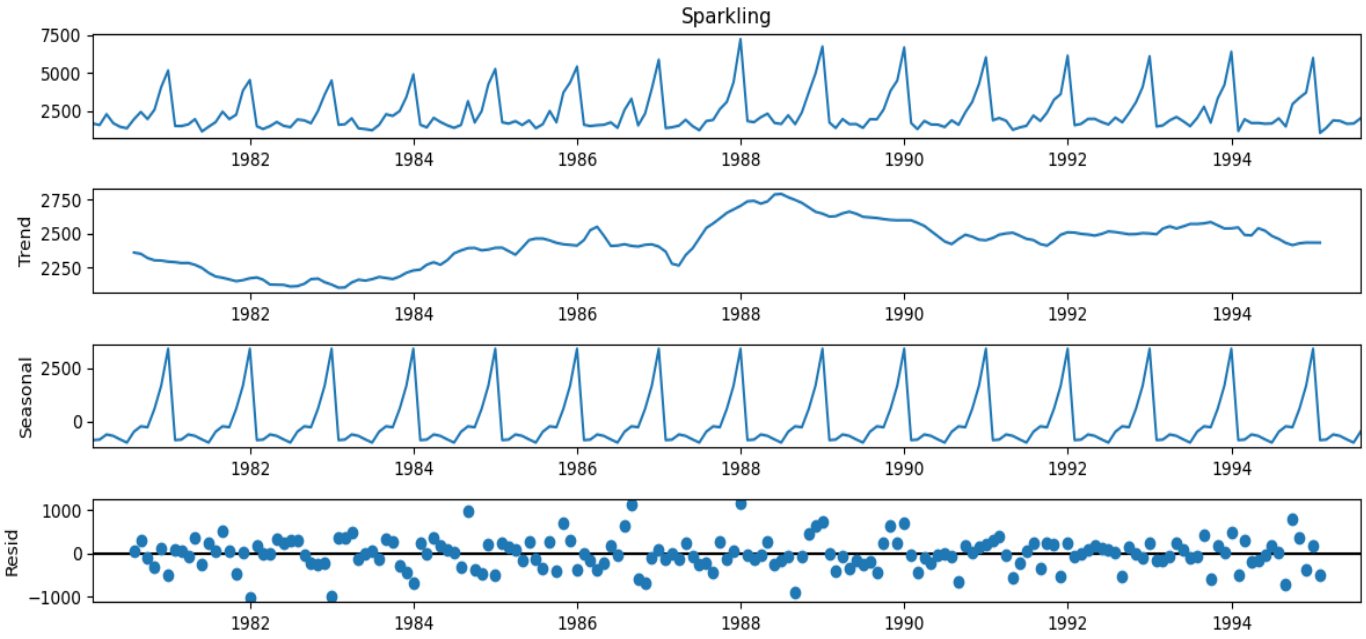

Figure 12: Additive decomposition

The first graph shows the entire time series, as was seen in Figure 1.

The second plot shows the trend over time. The sales decrease, then increase and decrease later.

The third plot shows the seasonality component of the time series. The same sales pattern is repeated each year.

The "noise" element, which can be seen in the fourth graph, is the random component which cannot be explained through systematic components. Residuals seem to follow a pattern. Residual values decrease and then increase.

Multiplicative decomposition of time series

In the multiplicative model, the time series is the multiplication of components – trend, seasonality and residual. It explains the non-linear change over time. Unlike the additive model, it tells the change in percentage terms.

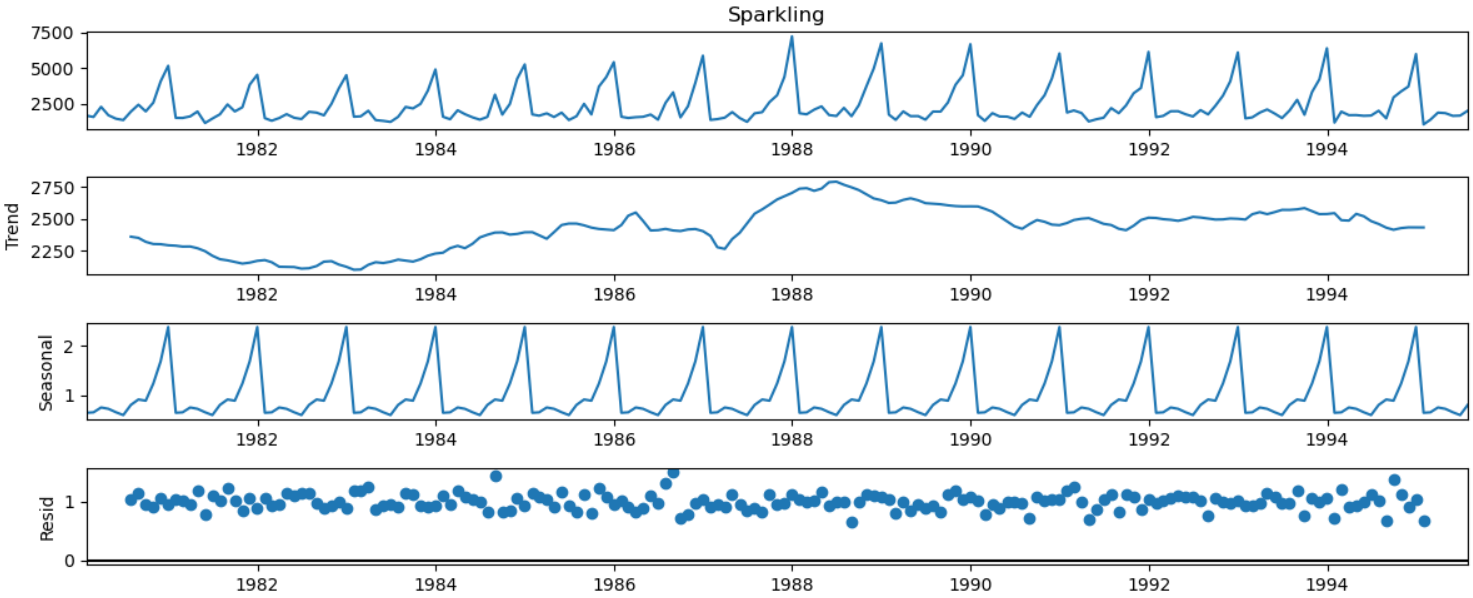

Figure 13: Multiplicative decomposition

The first three plots are the same, with the only difference being that seasonality is plotted against percentage change.

The residual points don't follow a pattern. Almost all values are centred on a single point

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

Unlike classification on regression technique, the data cannot be split randomly. The test set (unseen data) has to be the most recent one because of the ordered nature of data.

The data for Sparkling wine sales is split in such a manner that the train data should have records till 1990 and the test data starts from 1991.

Train set

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

Table 3: First 5 rows of train set

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

Table 4: Last 5 rows of train set

The train set contains **132 rows** and **1 column**.

Test set

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

Table 5: First 5 rows of test set

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

Table 6: Last 5 rows of test set

The test set has **55 rows** and **1 column**.

Plotting train-test graph

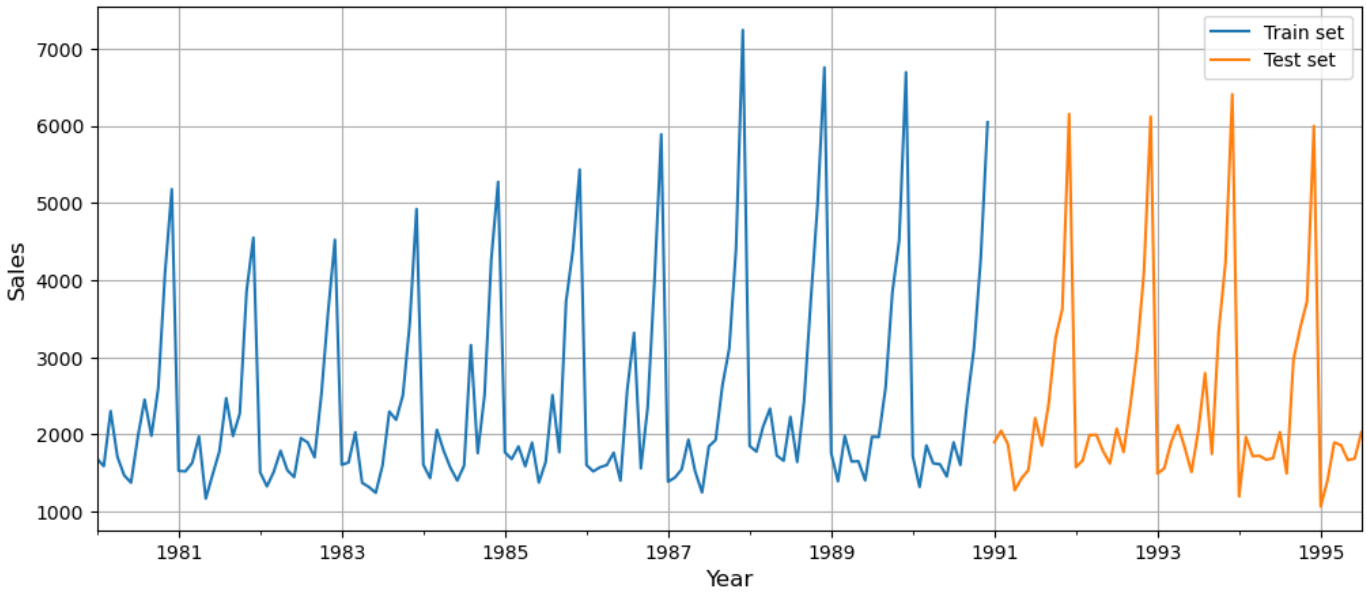

Figure 14: Train-test time series plot

The most recent sales records, starting from 1991, have been taken as the test set.

4. Build all exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, Naïve forecast models and simple average models should also be built on the training data and check the performance on the test data using RMSE.

Linear regression

Before building the regression model, we regress the wines sales variable against the order of the occurrence.

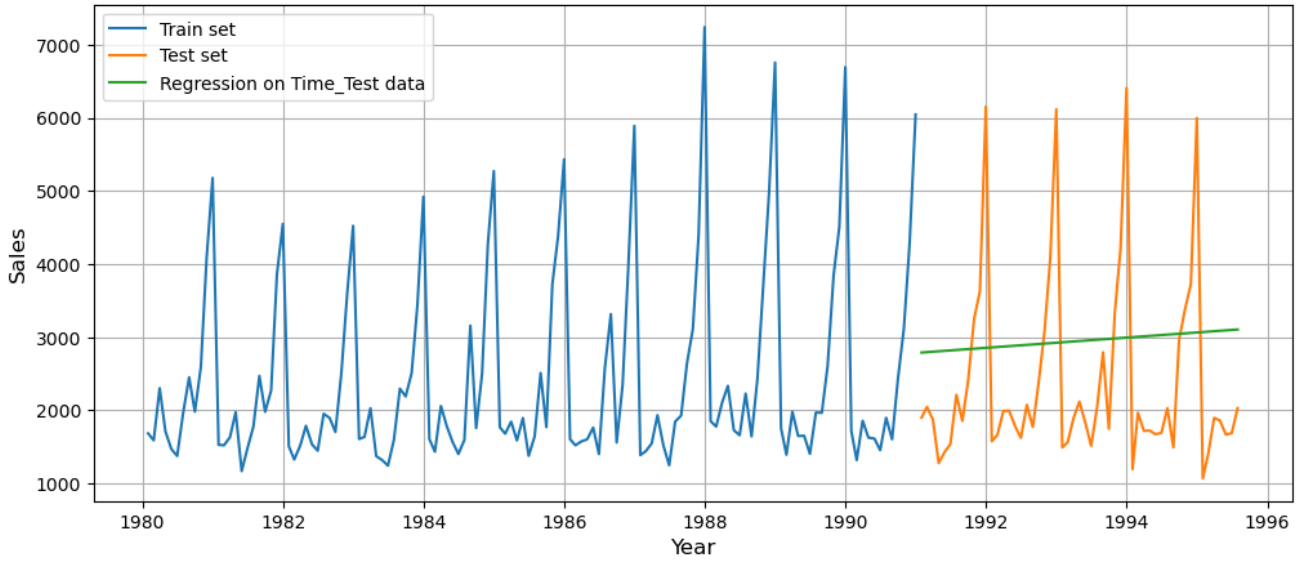

Figure 15: Regression on test set

The regression model captures the trend to some extent, but totally fails to factor in the seasonality component of the time series.

Root mean squared error (RMSE), which is a metric to measure the forecast accuracy, on the test set is 1389.135.

Naïve model

In the Naïve approach, **the last observed value is taken to forecast for future**. In other words, the last value from the train test is used to forecast for the future, i.e., the entire test period.

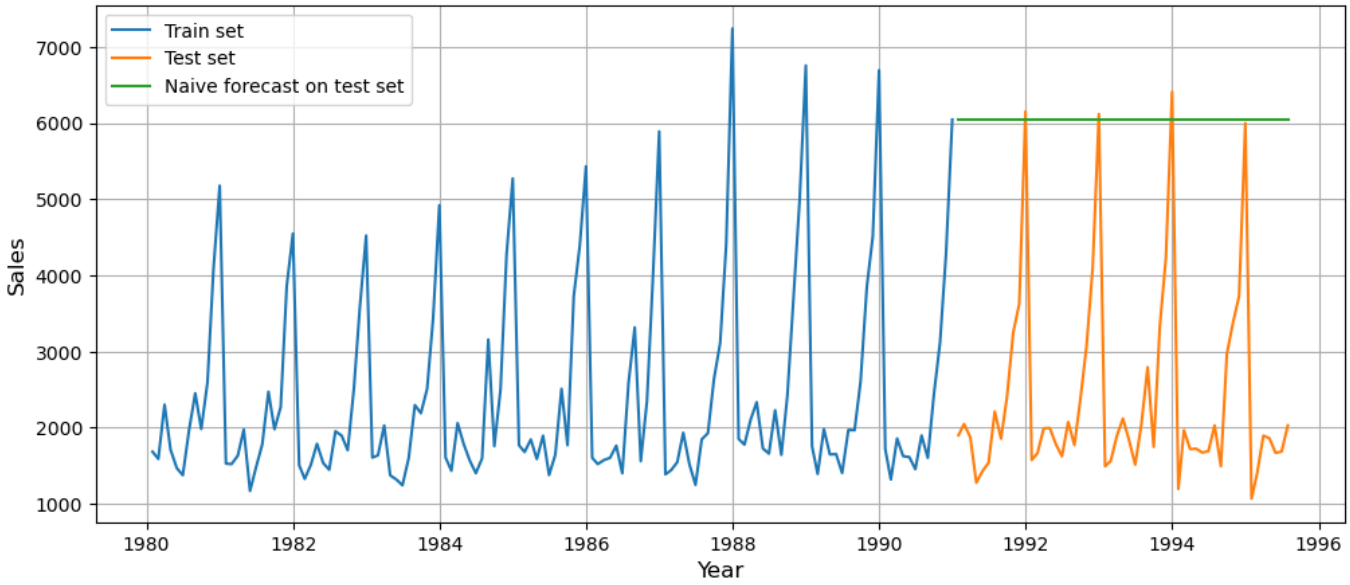

Figure 16: Naïve forecast on test set

The **forecast is constant**. It neither captures the trend nor the seasonality.

RMSE on test set is 3864.28

RMSE for the Naïve model is more than that of the regression model.

Simple average

For the simple average method, we will forecast by using the average of the training dataset values.

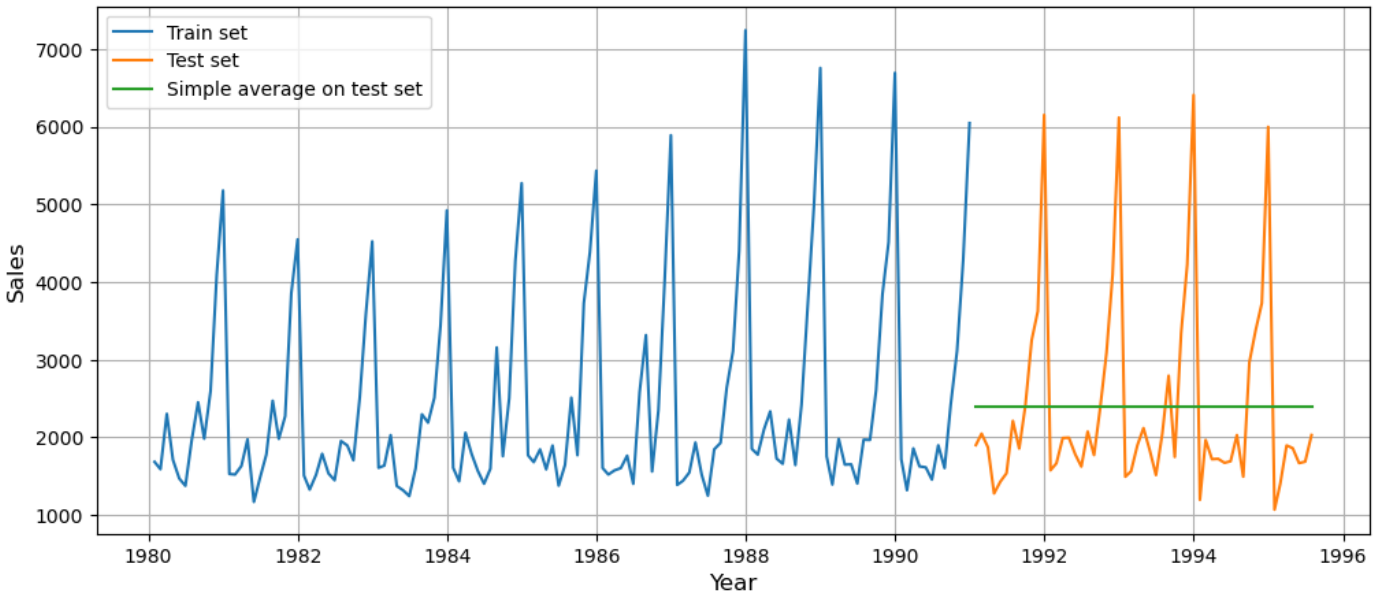

Figure 17: Simple average on test set

It can be observed that the forecast on the test set is constant. The simple average method neither factors in trend nor seasonality.

RMSE on test set is 1275.0818.

The RMSE value is the lowest out of the three models we have built so far. However, it cannot be used to make predictions for the future as the simple average model is too simple to capture the systematic components.

Moving average

In the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error).

For the Sparkling wine sale, we are going to take intervals of 2, 4, 6 and 9, and then average over the entire data.

Time Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1995-03-31	1897	1649.5	2592.00	2913.666667	2664.000000
1995-04-30	1862	1879.5	1557.75	2659.833333	2645.222222
1995-05-31	1670	1766.0	1707.75	2316.666667	2664.666667
1995-06-30	1688	1679.0	1779.25	1598.166667	2522.444444
1995-07-31	2031	1859.5	1812.75	1758.333333	2372.000000

Table 7: Moving average data sample

The table shows the last five rows of the dataset with rolling means taken at time intervals of 2, 4, 6 and 9.

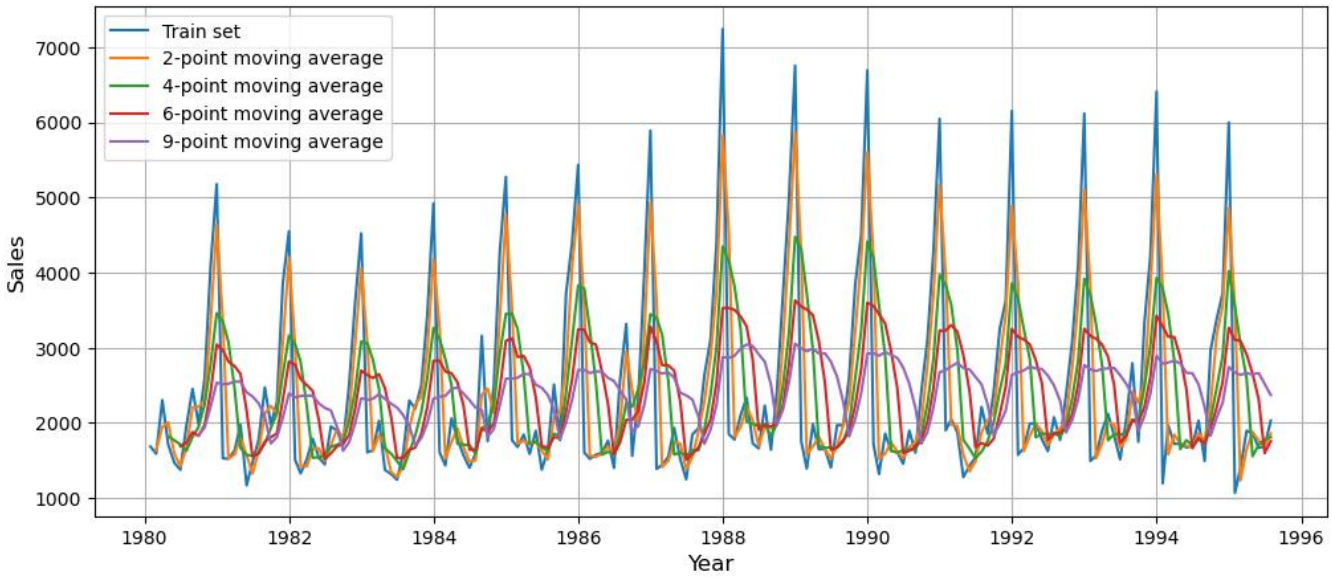

Figure 18: Moving average plot on entire dataset

The figure shows moving average on the entire time series.

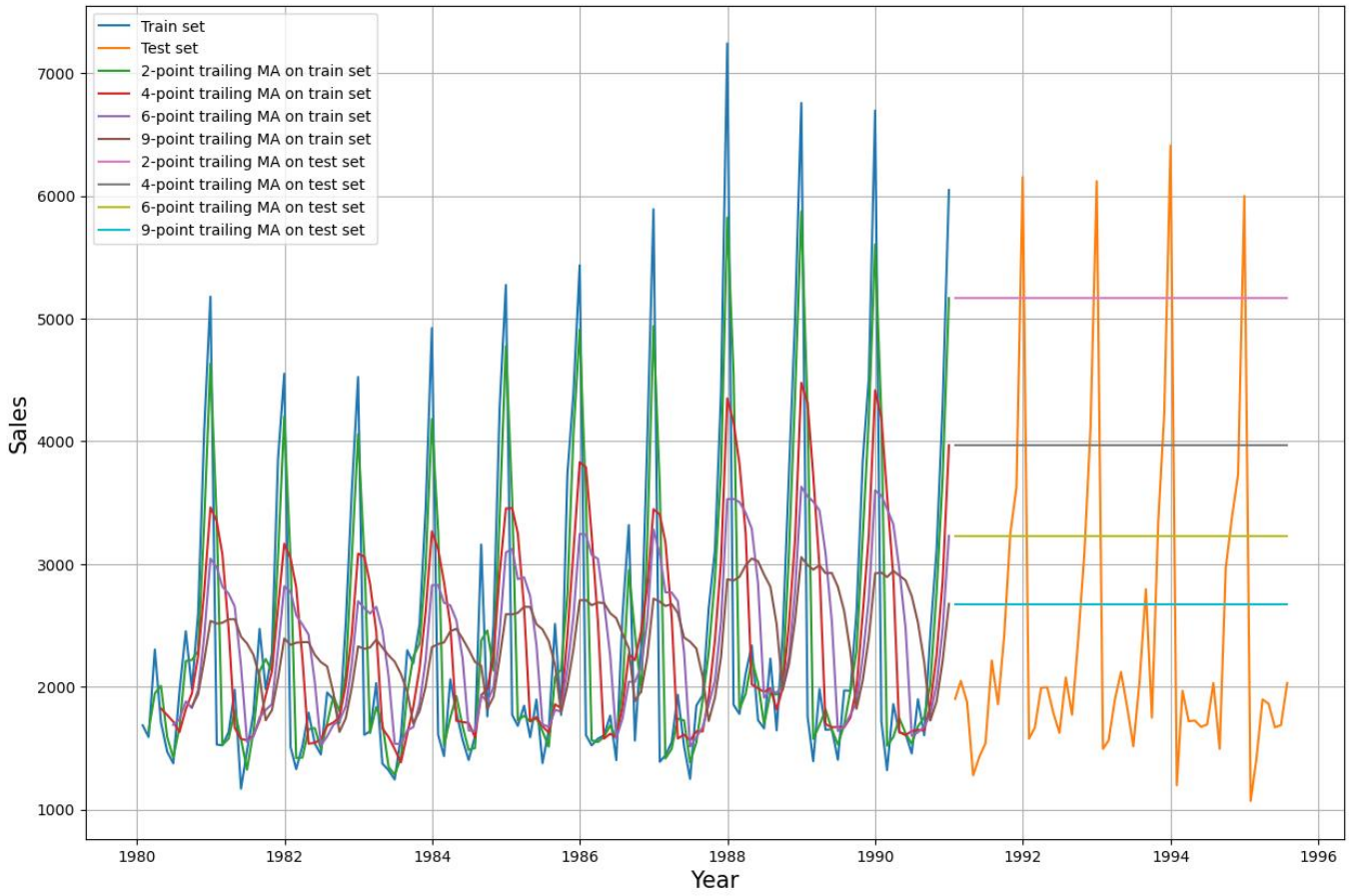

Figure 19: Moving average forecast on test set

The moving average forecast at different intervals gives a **constant prediction** for the future.

The 2-point rolling average forecasts higher sales.

The 9-point rolling average predicts the lowest sales.

RMSE for 2-point rolling moving average is 3046.976.

RMSE for 4-point rolling moving average is 2021.856.

RMSE for 6-point rolling moving average is 1521.611.

RMSE for 9-point rolling moving average is 1304.6189.

RMSE is the highest for interval 2 and the lowest for interval 9.

Simple Exponential Smoothing

The Simple Exponential Smoothing (SES) method is suitable for forecasting data with no clear trend or seasonal pattern.

Parameter alpha (α) is called the smoothing constant. It corresponds to level, which is the local mean. Its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Simple Exponential Smoothing.

If α is closer to 1, forecasts follow the actual observations more closely.

If α is closer to 0, forecasts are farther from the actual observations and the prediction line is smooth.

For the Sparkling wine sales, **α is taken to be 0.0496.** This value has been automatically generated by the model.

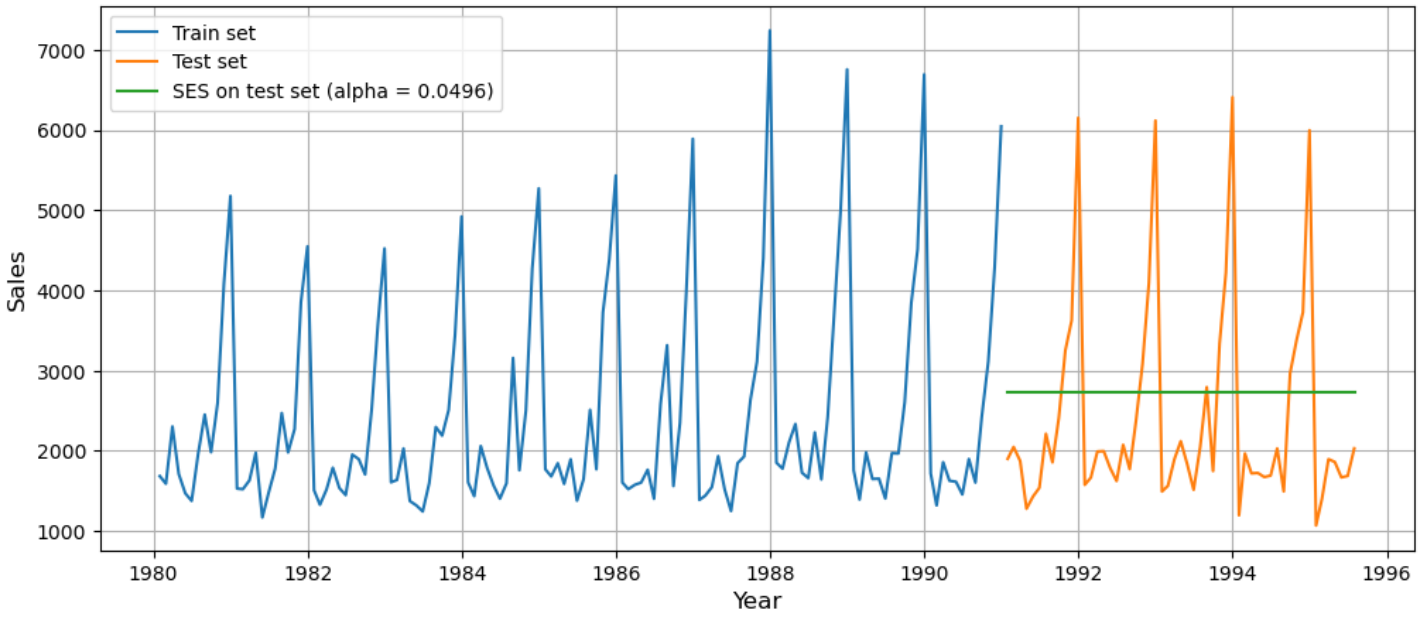

Figure 20: SES forecast (α = 0.0496)

The predictions for the future remain constant, neither capturing the trend nor seasonality.

RMSE on test set is 1316.0347.

We can set different values for α to check whether or not the performance of the SES model improves. The higher the alpha value, more weightage is given to the more recent observation. That means, what happened recently will happen again.

The details of the different values of α are given in the appendix.

Alpha Values	Train RMSE	Test RMSE
0.3	1359.511747	1935.507132
0.4	1352.588879	2311.919615
0.5	1344.004369	2666.351413
0.6	1338.805381	2979.204388
0.7	1338.844308	3249.944092
0.8	1344.462091	3483.801006
0.9	1355.723518	3686.794285

Table 8: RMSE for different α values

RMSE on the test data is the least for **α = 0.3**.

Therefore, another SES model will be built with α = 0.3.

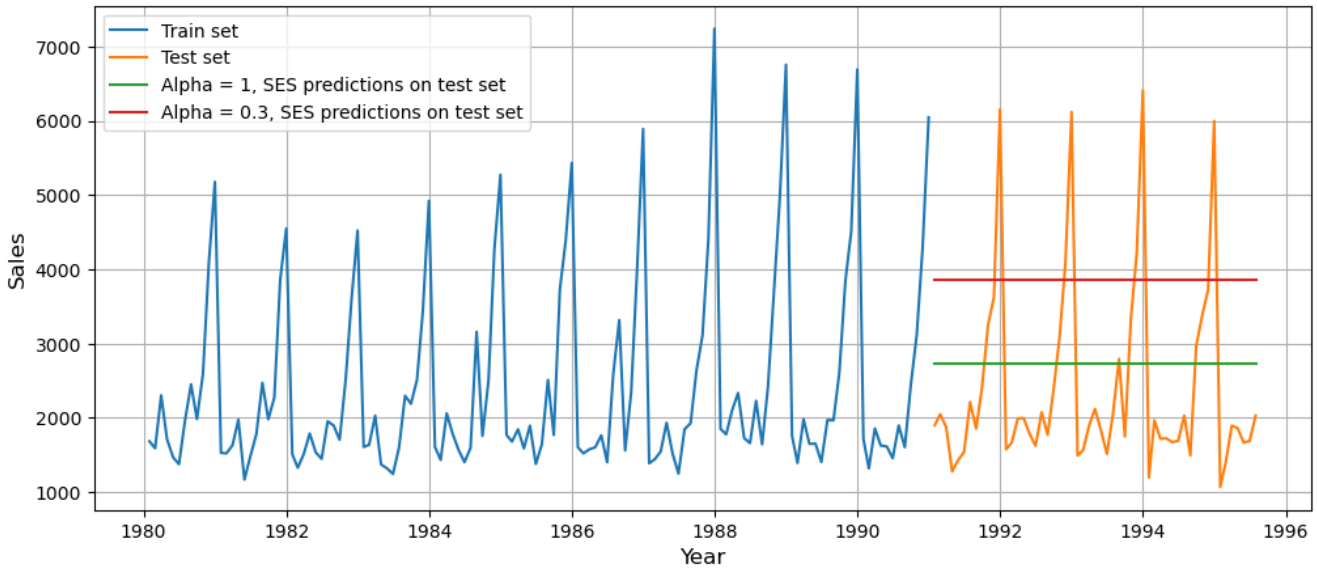

Figure 21: SES forecast (α = 0.3)

Even the optimised value of **α does not improve the performance of the model.**

RMSE on test set is 1935.507.

Double Exponential Smoothing

Double Exponential Smoothing (DES), also called the Holt's method, is an extension of the SES. It is applicable when data has trend, but no seasonality.

It has two smoothing parameters – level (α) and trend (β). Both α and β lie between 0 and 1.

For the DES model, the parameters are as follows:

α = 0.688, β = 0.0001

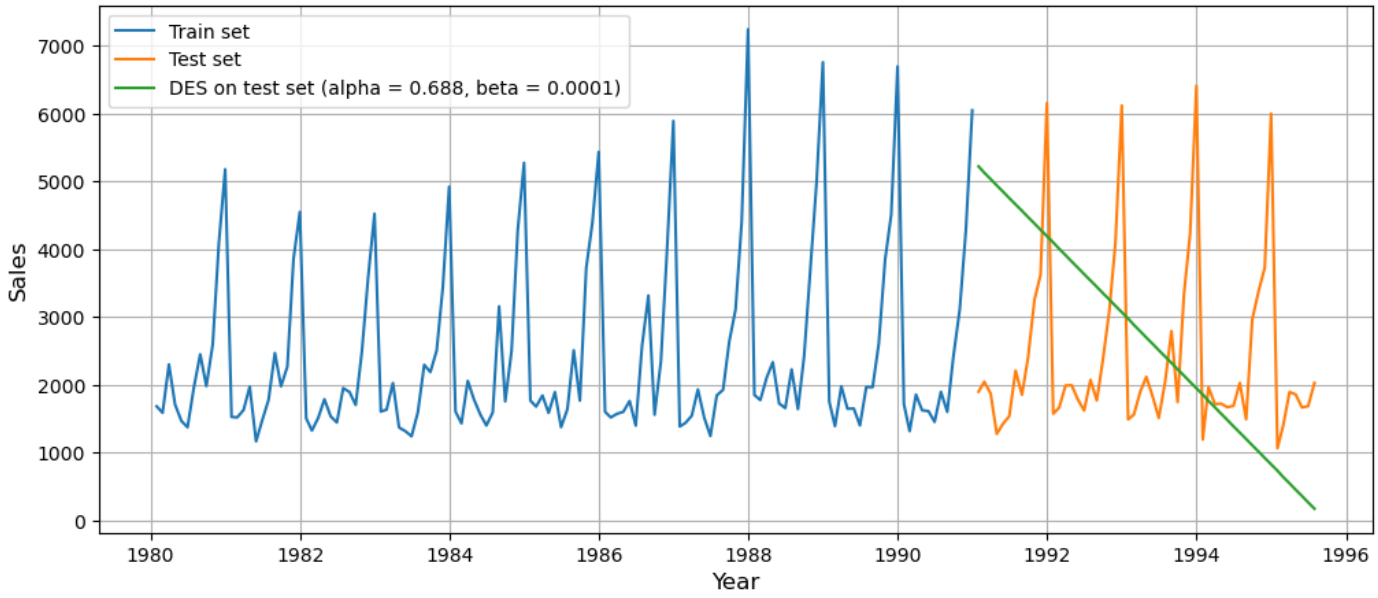

Figure 22: DES forecast (α = 0.688, β = 0.0001)

The DES model was expected to capture the trend, but it fails do so. The sales increase and then decrease gradually in the train data time series, but the forecast values have a decreasing trend.

RMSE on test set is 2007.239

We can set different values for α and β to check whether or not the performance of the DES model improves.

The details of ow the different values of α and β were chosen are given in the appendix.

Alpha values	Beta values	Train RMSE	Test RMSE
0.1	0.1	1382.520870	1778.564670
0.1	0.2	1413.598835	2599.439986
0.2	0.1	1418.041591	3611.763322
0.1	0.3	1445.762015	4293.084674
0.3	0.1	1431.169601	5908.185554

Table 9: RMSE for different α and β values

RMSE on the test set is the least for **α = 0.1** and **β = 0.1**.

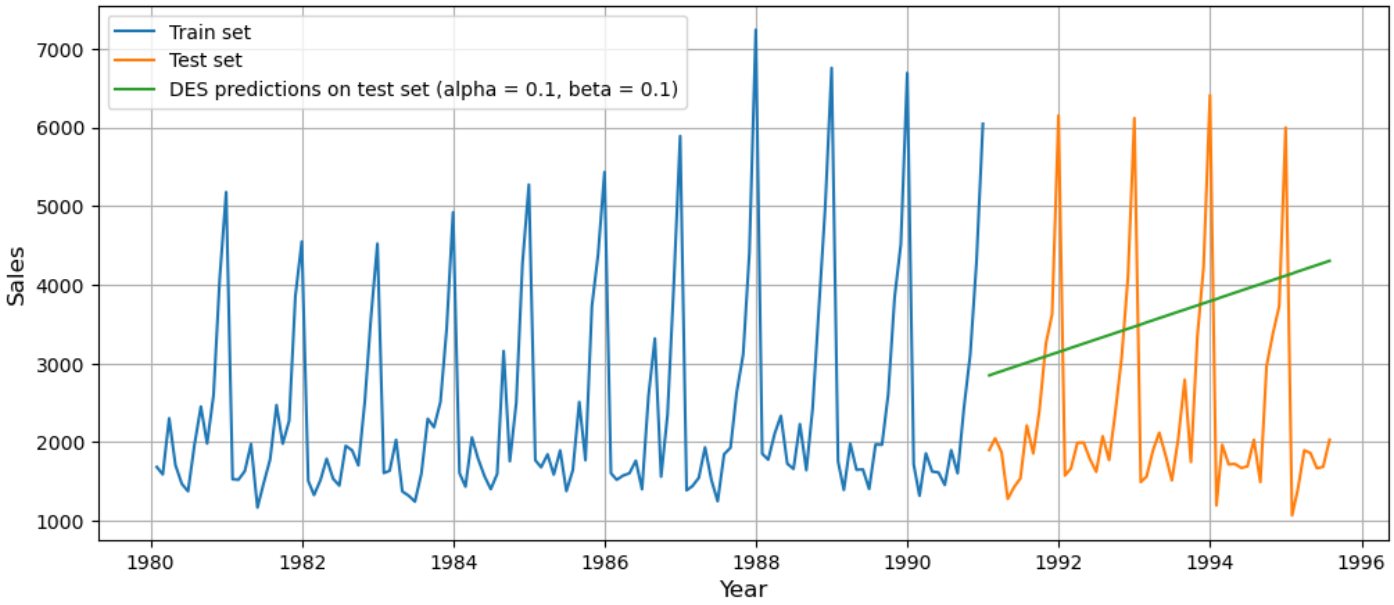

Figure 23: DES forecast (α = 0.1, β = 0.1)

The prediction is reversed as compared with the previous model. However, it still fails to capture the trend.

RMSE on tests set is 1778.565

The RMSE value is lower than that of the previous DES model.

Triple Exponential Smoothing

Triple Exponential Smoothing (TES), also called the Holt's Winter method, is applicable when data has trend and seasonality.

It has three smoothing parameters – level (α), trend (β) and seasonality (γ). The three parameters lie between 0 and 1.

Since seasonality is multiplicative and additive, we will build one model for each.

Multiplicative seasonality

For the TES model, the parameters are as follows:

α = 0.111, β = 0.493, γ = 0.362

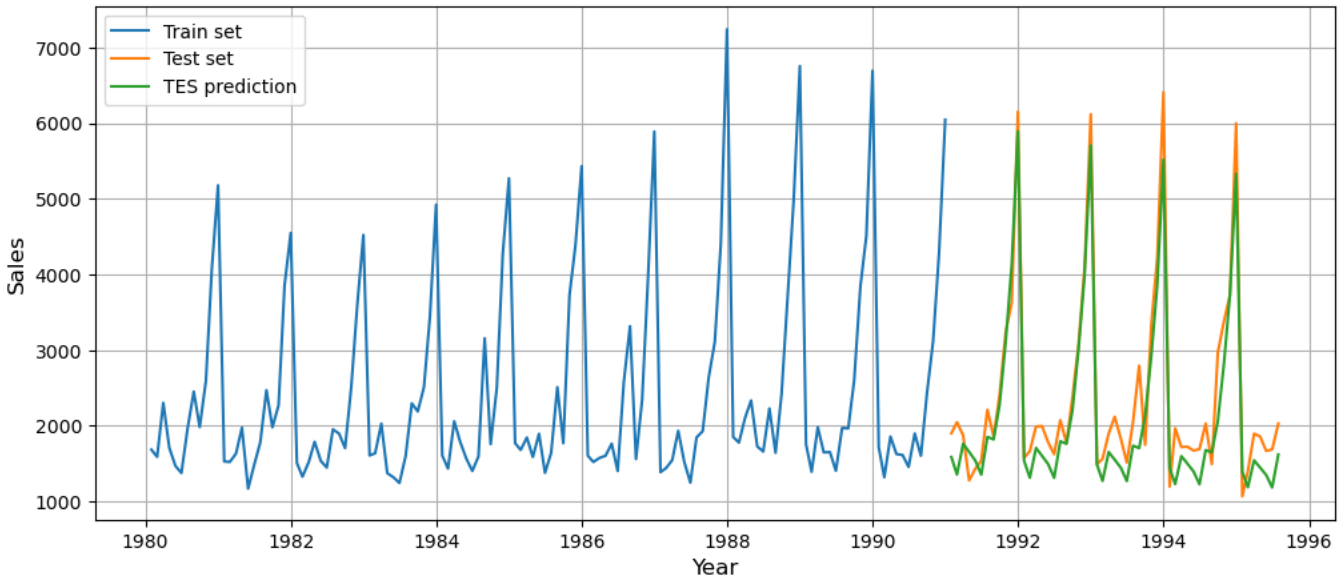

Figure 24: TES forecast (α = 0.111, β = 0.493, γ = 0.362)

The TES model captures the trend as well as the seasonality. The TES model seems to be the best model so far.

RMSE on test set is 403.126.

This is the least RMSE value so far.

Additive seasonality

For the TES model, the parameters are as follows:

α = 0.111, β = 0.124, γ = 0.461

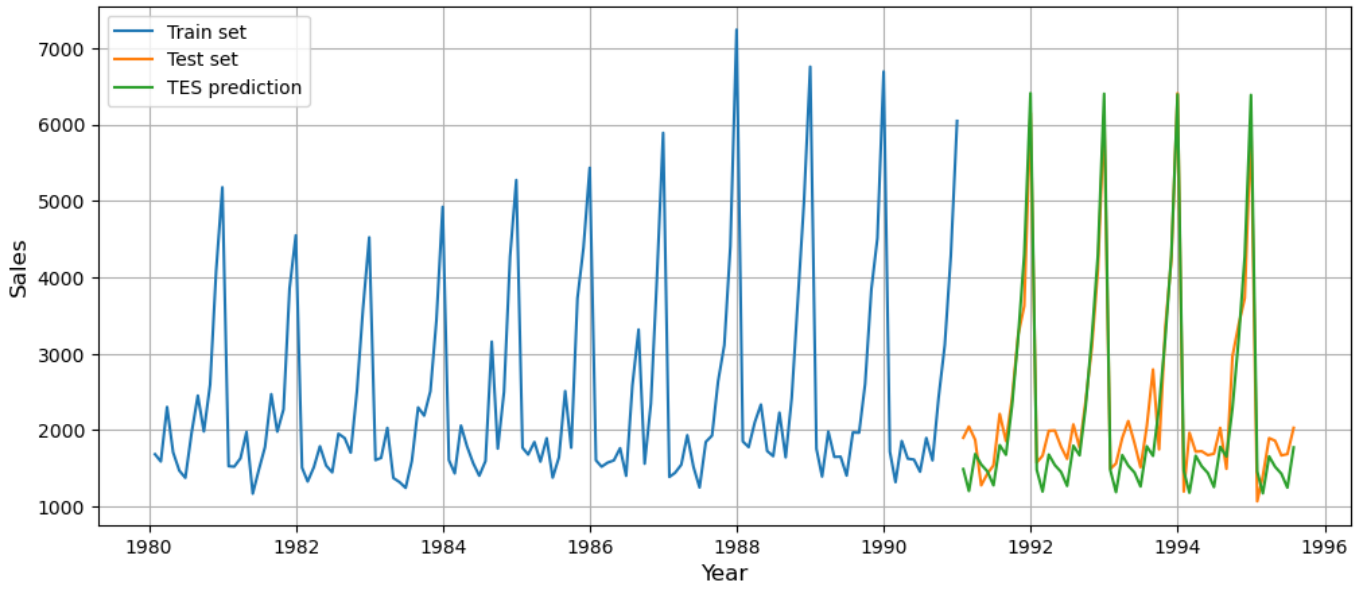

Figure 25: TES forecast (α = 0.111, β = 0.124, γ = 0.461)

The TES forecast on the test set captures both trend and seasonality.

RMSE on test set is 378.944

The RMSE score is less than the TES model with multiplicative seasonality. It means that TES model with additive seasonality performs better on the unseen data.

We can set different values for α, β and γ to check whether or not the performance of the TES model improves.

The details of how different values of α, β and γ were chosen are given in the appendix.

Alpha values	Beta values	Gamma values	Train RMSE	Test RMSE
0.8	1.0	0.3	790.740655	580.26611
0.8	1.0	0.3	790.740655	580.26611
0.8	1.0	0.3	790.740655	580.26611
0.8	1.0	0.3	790.740655	580.26611
0.8	1.0	0.3	790.740655	580.26611

Table 10: RMSE for different α, β and γ values

RMSE on the test set is the least for **α = 0.8**, **β = 1** and **γ = 0.3**.

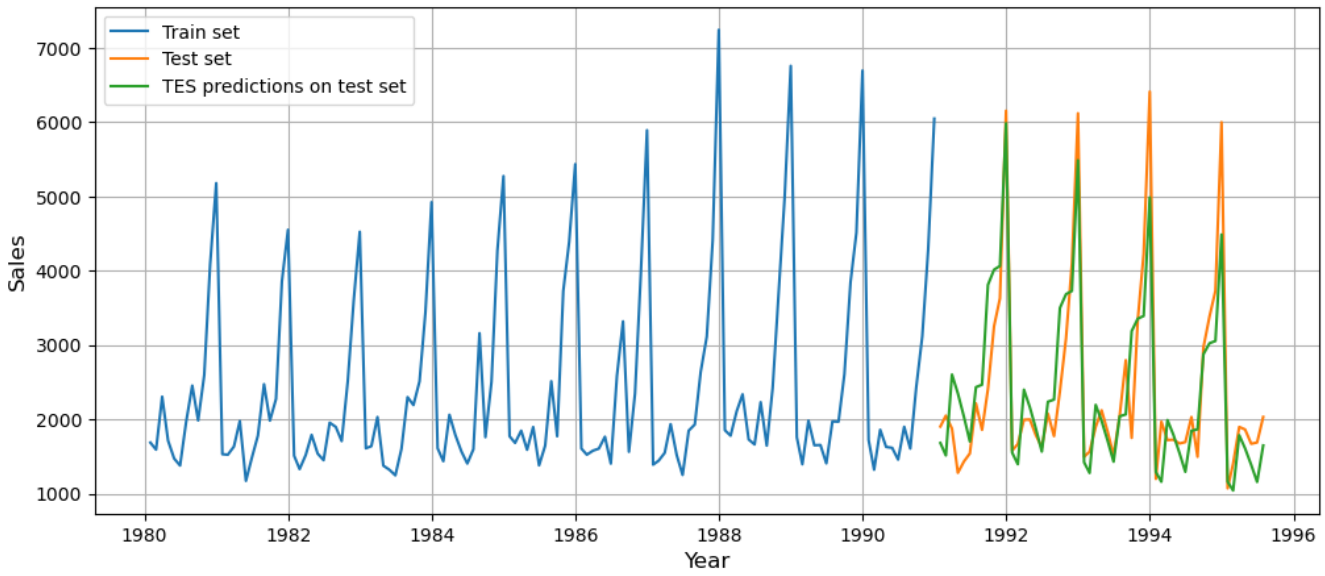

Figure 26: TES forecast (α = 0.8, β = 1, γ = 0.3)

The TES model built on the basis of the best parameters fails to capture the trend in entirety. The TES prediction on the unseen data gives a decreasing trend, which is not the case as denoted by the orange graph.

RMSE on test set is 580.266.

The value is more than that of the previous two TES models.

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

A time series is considered to be stationary when statistical properties such as the variance and (auto) correlation are constant over time.

Dickey-Fuller Test on the time series is run to check for stationarity of data.

Null hypothesis (H_0): Time series is non-stationary

Alternative hypothesis (H_1): Time series is stationary

If p-value < 0.05, the null hypothesis is rejected. This means that time series is stationary.

If p-value > 0.05, we fail to reject the null hypothesis. This means that time series is non-stationary.

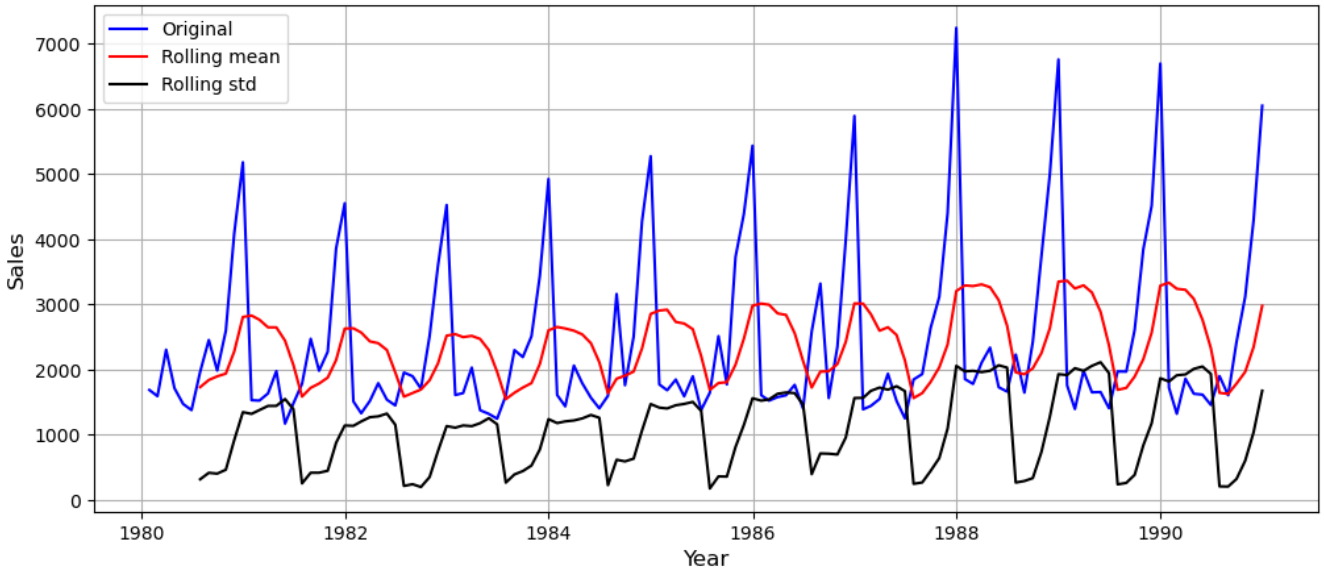

Figure 27: Non-stationary time series plot

The time series has trend and seasonality.

At α (level of significance) = 0.05, p-value is 0.669744.

Since the p-value > 0.05, we fail to reject the null hypothesis. This means that **time series is non-stationary**.

Differencing ('d') is done on a non-stationary time series data one or more times to convert it into a stationary time series data.

First-order differencing (d=1) is done where the difference between the current and previous (one lag before) series is taken and then checked for stationarity using the Dickey-Fuller test.

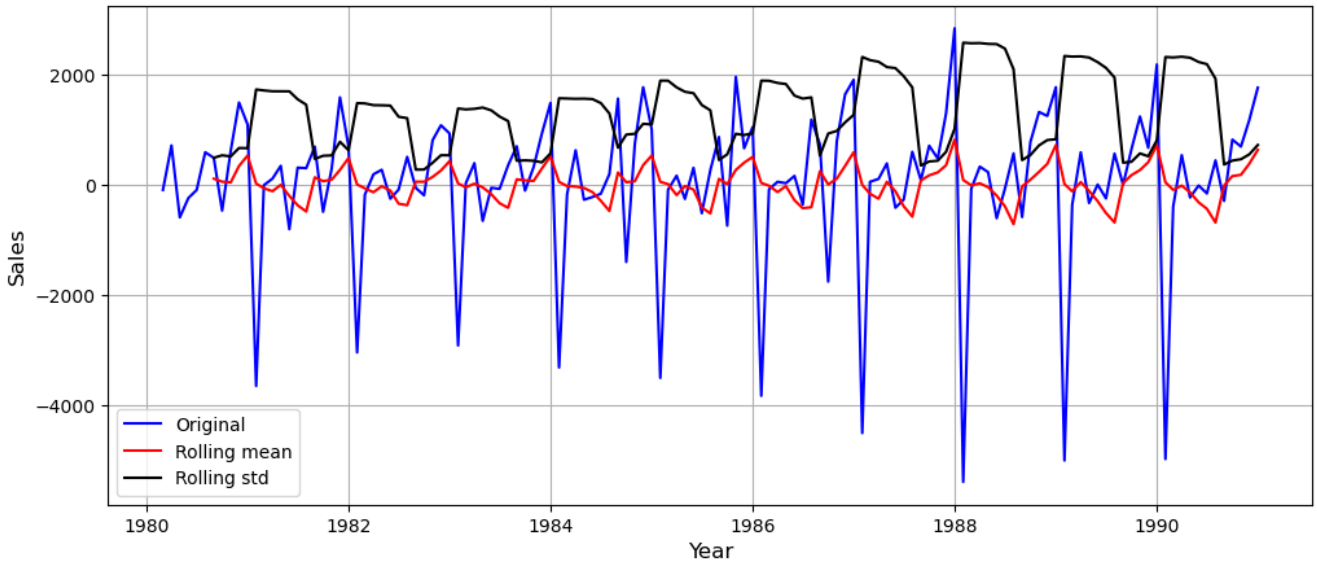

Figure 28: Stationary time series plot (d =1)

At α = 0.05, p-value is 2.280104e-12.

Since the p-value < 0.05, the null hypothesis is rejected. This means that **time series is stationary at d =1.**

6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

ARIMA model

An ARIMA model has three components:

1. **Autoregressive (AR) model**: AR models use previous time period values to predict the current time period values. The number of lag observations or autoregressive terms in an AR model is denoted by **'p'**. One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process.
2. **Integrated component**: It is the difference in the non-seasonal observations. It is denoted by **'d'**.
3. **Moving Average (MA) model**: MA models estimates the future values based on historical forecast errors. The size of the moving average window is denoted by **'q'**.

One of the approaches to find the order of 'p' and 'q' is the least value of **Akaike Information Criteria (AIC)**. The AIC values for different pairs of 'p' and 'q' are compared to find the optimum order for model-building.

For Sparling wine sales time series data, we can set different values for 'p' and 'q' to find the combination that gives the least AIC value. The details are given in the appendix.

Here, **d = 1** because at first-order differencing, the time series becomes stationary.

Params	AIC
(2, 1, 2)	2213.509213
(2, 1, 1)	2233.777626
(0, 1, 2)	2234.408323
(1, 1, 2)	2234.527200
(1, 1, 1)	2235.755095

The AIC value is the least for **p = 2**, **d = 1**, **q = 2**.

An ARIMA model will be built on the basis of these parameters.

Table 11: ARIMA parameters

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Thu, 11 May 2023	AIC	2213.509
Time:	09:39:27	BIC	2227.885
Sample:	01-31-1980 - 12-31-1990	HQIC	2219.351
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.782	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.740	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.215	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.108	0.000	0.785	1.215
sigma2	1.099e+06	2e-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

Table 12: ARIMA summary

The ARIMA model (2, 1, 2) has two AR terms and as many MA terms.

It can be observed that each variable has a p value. Each variable has a null (H_0) and an alternative (H_1) hypothesis.

H_0: Variable is not significant

H_1: Variable is significant

At 5% level of significance, p-values exceeding 0.05 will mean that an independent variable is not significant and, hence, can be dropped.

From the table, it can be seen that p-values for all variables are less than 0.05. Therefore, we reject the null hypothesis. Hence, **all four variables are significant**.

RMSE on test set is 1299.9798.

SARIMA model

A SARIMA model has the following components:

- Autoregressive (**p**) and moving average (**q**) components
- Seasonal autoregressive (**P**) and moving average (**Q**) components
- Ordinary and seasonal difference components of order **'d'** and **'D'**, respectively
- Seasonal frequency (**F**)

To find the seasonal parameter for the SARIMA model, we need to look at the **Autocorrelation Function (ACF) plot**, which summarises the correlation of an observation with lag values.

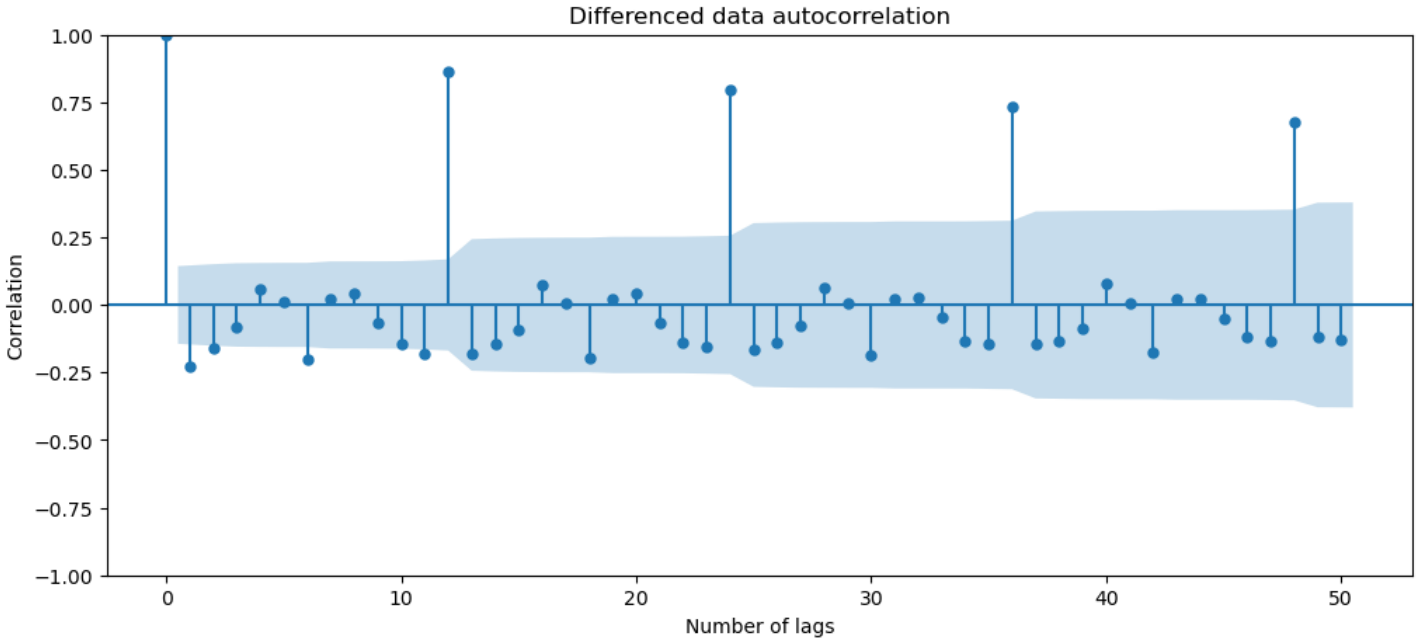

Figure 29: ACF plot

We see that there can be a seasonality of 12. Therefore, we will run an auto SARIMA model by setting the **seasonality (F) as 12** and **d =1 (for stationarity)**.

For other parameters, we can set different values for 'p', 'q', 'P' and 'Q' to find the combination that gives the least AIC value. The details of how the parameters were chosen are given in the appendix.

Param	Seasonal	AIC
(1, 1, 2)	(1, 0, 2, 12)	1555.584247
(1, 1, 2)	(2, 0, 2, 12)	1555.934564
(0, 1, 2)	(2, 0, 2, 12)	1557.121563
(0, 1, 2)	(1, 0, 2, 12)	1557.160507
(2, 1, 2)	(1, 0, 2, 12)	1557.340402

The AIC value is the least for the following combination of parameters.

p = 1, q = 2

P = 1, Q =2

d = 1, D = 0

Table 13: SARIMA parameters

On building a SARIMA model on the basis of the above parameters, we get the following results.

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(1, 0, 2, 12)	Log Likelihood	-770.792
Date:	Sat, 13 May 2023	AIC	1555.584
Time:	01:23:39	BIC	1574.095
Sample:	0	HQIC	1563.083
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6281	0.255	-2.463	0.014	-1.128	-0.128
ma.L1	-0.1041	0.225	-0.463	0.643	-0.545	0.337
ma.L2	-0.7276	0.154	-4.734	0.000	-1.029	-0.426
ar.S.L12	1.0439	0.014	72.841	0.000	1.016	1.072
ma.S.L12	-0.5550	0.098	-5.663	0.000	-0.747	-0.363
ma.S.L24	-0.1354	0.120	-1.133	0.257	-0.370	0.099
sigma2	1.506e+05	2.03e+04	7.400	0.000	1.11e+05	1.9e+05

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	11.72
Prob(Q):	0.84	Prob(JB):	0.00
Heteroskedasticity (H):	1.47	Skew:	0.36
Prob(H) (two-sided):	0.26	Kurtosis:	4.48

Table 14: SARIMA summary

The SARIMA model has the following components:

- One AR variable
- Two MA variables
- One seasonal AR variable with a lag of 12
- One seasonal MA variable with a lag of 12
- One seasonal MA variable with a lag of 24

The p-values of three variables are less than 0.05. This means that **three components out of six are significant**.

RMSE on test set is 528.607.

The RMSE value of the SARIMA model is less than that of the ARIMA model. In other words, **SARIMA model performs better on the unseen data.**

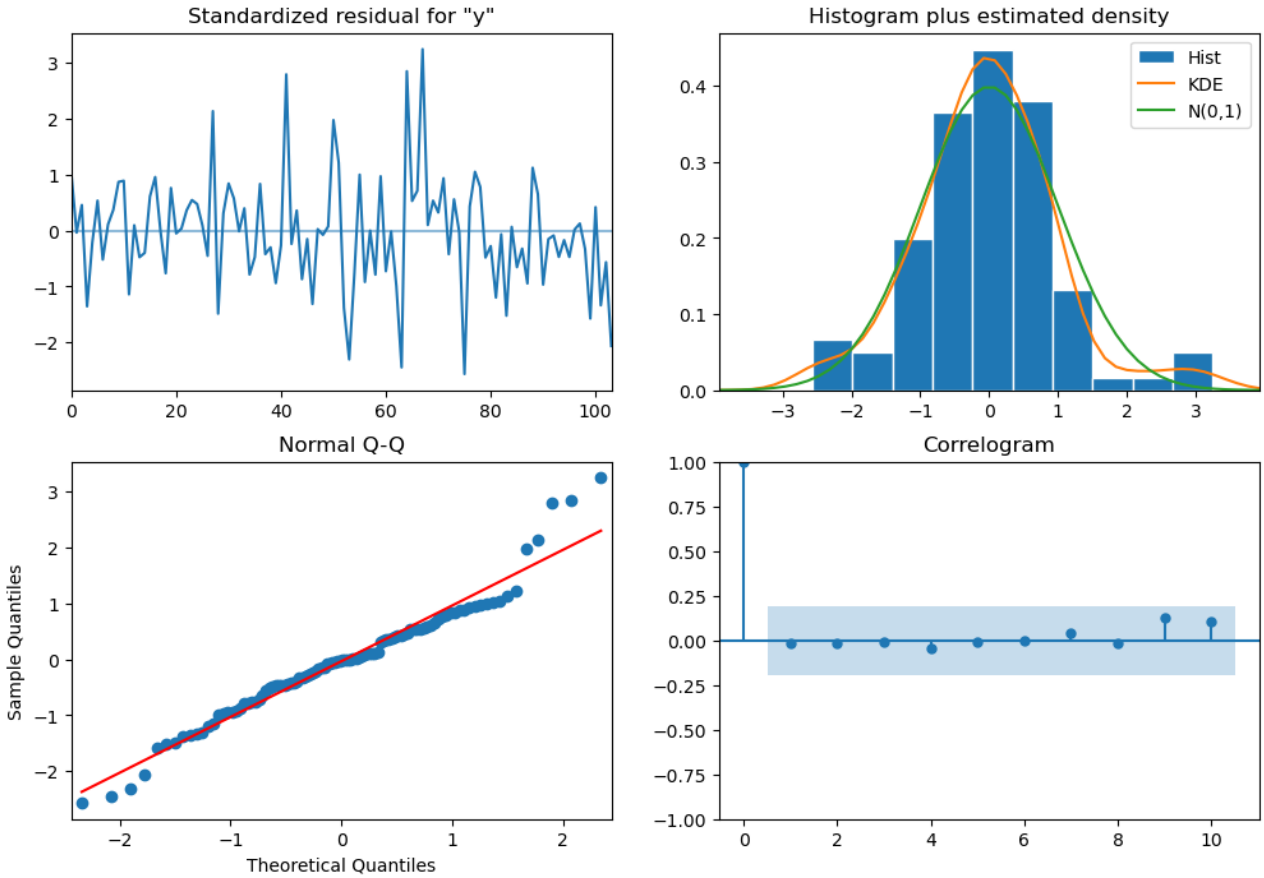

Figure 30: Diagnostic plots

The diagnostic graphs plot the residuals. All four diagnostics plots almost follow the theoretical numbers and thus we cannot observe any pattern from these plots.

The first plot shows the standardised error. Residuals are not constant.

The histogram and the normal Q-Q plot show that the residuals almost follow the normal distribution.

Correlogram tells us the correlation of residuals with time. It can be seen that all correlations are insignificant. This should be the case, as we do not want residuals to have correlation with time.

7. Build a table (create a data frame) with all models built along with their corresponding parameters and the respective RMSE values on the test data.

	Test RMSE
Triple Exp Smoothing (alpha = 0.111, beta = 0.124, gamma = 0.461)	378.944325
Triple Exp Smoothing (alpha = 0.111, beta = 0.493, gamma = 0.362)	403.125867
SARIMA (1, 1, 2) (1, 0, 2, 12)	528.607231
Triple Exp Smoothing (alpha = 0.8, beta = 1, gamma = 0.3)	580.266110
Simple Average Model	1275.081804
ARIMA (2, 1, 2)	1299.979855
9-point trailing Moving Average	1304.618912
Simple Exp Smoothing (alpha = 0.0496)	1316.034674
Regression on time	1389.135175
6-point trailing Moving Average	1521.611250
Double Exp Smoothing (alpha = 0.1, beta = 0.1)	1778.564670
Simple Exp Smoothing (alpha = 0.3)	1935.507132
Double Exp Smoothing (alpha = 0.688, beta = 0.0001)	2007.238526
4-point trailing Moving Average	2021.855880
2-point trailing Moving Average	3046.976092
Naive Model	3864.279352

Table 15: Comparison of all models

Of all the models built, the Triple Exponential Smoothing has the least RMSE. This model has the following parameters:

Level (α) = 0.111

Trend (β) = 0.124

Seasonality (γ) = 0.461

8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

The most optimum model based on the least RMSE is the Triple Exponential Smoothing. A full model will be built on the complete dataset and then we will predict for future 12 months.

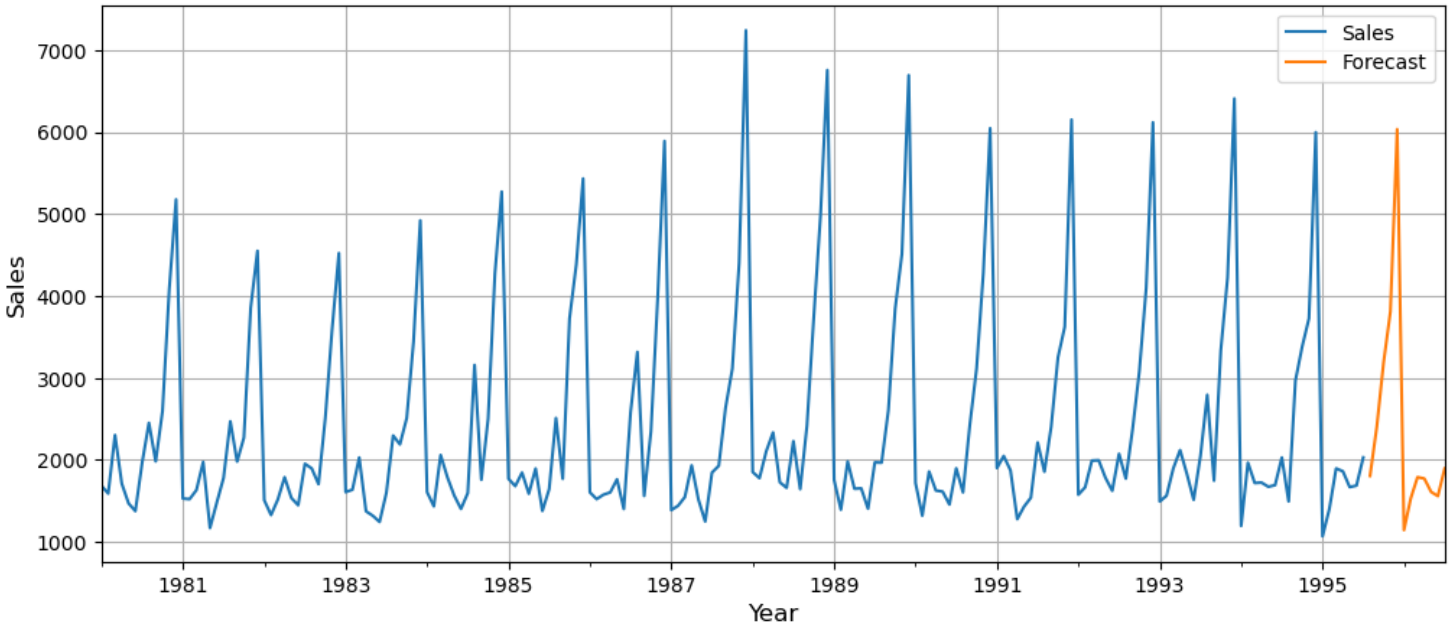

Figure 31: Sales forecast for 12 months

The graph shows the forecast for future 12 months. The prediction is in line with the time series data, as it captures both trend and seasonality.

However, this prediction (denoted by the orange graph) is very precise. In the real-world scenario, one cannot be sure about the forecast for future. Therefore, we will predict for future within a certain confidence interval.

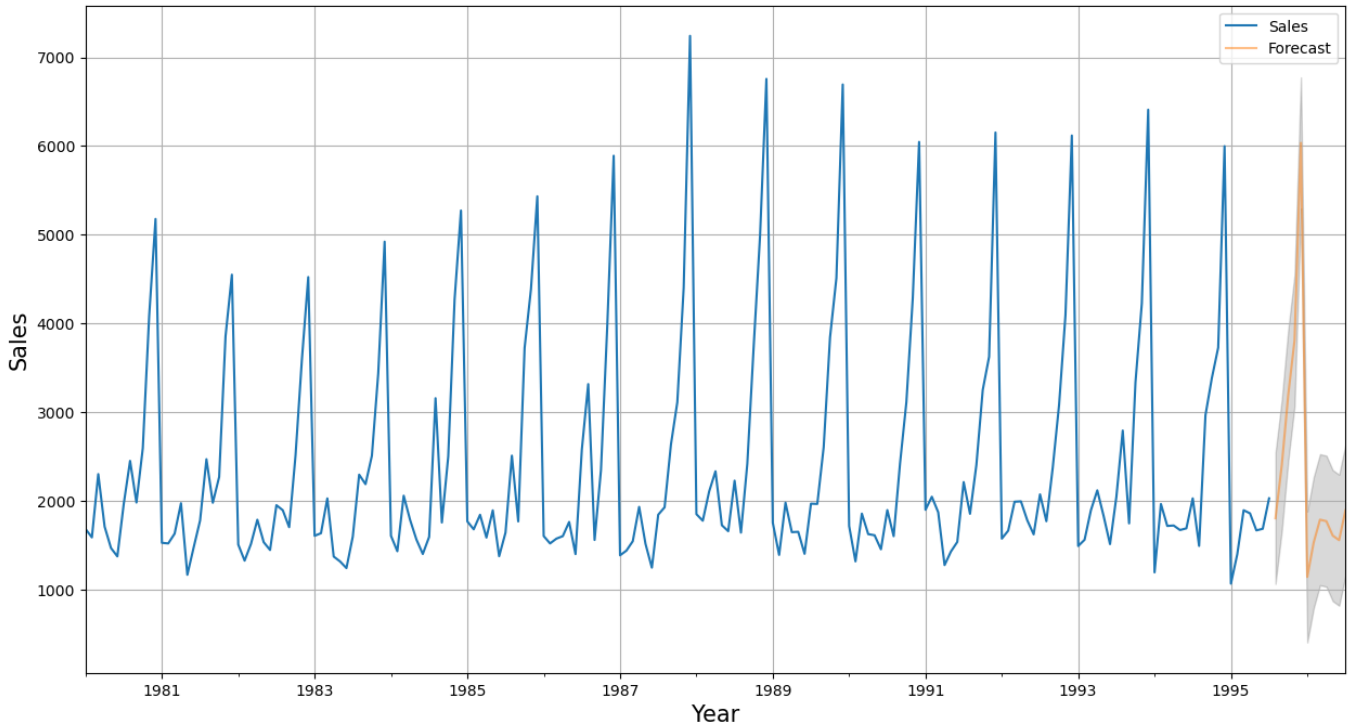

Figure 32: Sales forecast with confidence interval

The forecast values lie within the grey band, which represents 95 per cent confidence interval.

One assumption that we have made over here while calculating the confidence bands is that the standard deviation of the forecast distribution is almost equal to the residual standard deviation.

RMSE on the full model is 375.723.

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Exploratory Data Analysis summary

- The time series dataset does not have a pronounced trend. Sales increase and decrease gradually.
- As for the seasonality, sales are low at the start of a year and pick up pace at the end of a year. This pattern is repeated every year.
- The sales is the highest in quarter four and the lowest in quarter two.
- The demand for Sparkling wine shoots up in December. Sales at the end of a year are way more than the start of a year. Sales in January and February fail to go past 2,000 mark, but touches 6,000 in November and December.
- Median sales on four days of any week – Monday, Thursday, Friday and Saturday – remain the same. It is the **lowest on Tuesday**. Surprisingly, Sunday sales are lower than that of Saturday's.

Model-building summary

- Most of the models, be it Naïve, moving average or Double Exponential Smoothing, fail to capture the decreasing trend and seasonality.
- Triple Exponential Smoothing models are able to factor in both trend and seasonality.
- The SARIMA model, too, captures the systematic components. However, its RMSE value is more than that of the TES model with optimum parameters.

Recommendations

- Sales in quarter four should not be the company's concern, as demand is high at the end of the year, may be because of New Year's eve and Christmas.
- However, sales in January fall drastically when compared with December of the previous year. The company can continue to offer festive season discounts in January so that sales remain high at the start of the year as well.

- The company needs to focus on the first three quarters of the year. The advertising and marketing campaigns should be targeted at different segments of customers.
- Another strategy could be customer-specific social media campaigns.
- Wine tasting events are a good way to attract new customers. The company can invite celebrities and sommeliers for such events.

Appendix

Simple Exponential Smoothing

```
for i in np.arange(0.3, 1, 0.1):
    print(i)
```

Double Exponential Smoothing

```
for i in np.arange(0.1, 1.1, 0.1):
    for j in np.arange(0.1, 1.1, 0.1)
```

Triple Exponential Smoothing

```
for i in np.arange(0.3, 1.1, 0.1):
    for i in np.arange(0.3, 1.1, 0.1):
        for k in np.arange(0.3, 1.1, 0.1):
```

ARIMA model

```
p = q = range(0, 3)
d= range(1,2)
```

SARIMA model

```
p = q = range(0, 3)
d = range(1,2)
D = range(0,1)
P =Q = range(0,3)
```

Source: Great Learning logo that has been used on the cover page has been taken from one of the monographs provided by Great Learning