Project Report

Time Series Forecasting

Shubhadeep Bhattacharya
Program: PGP DSBA
Batch: October 2019

Table of Contents

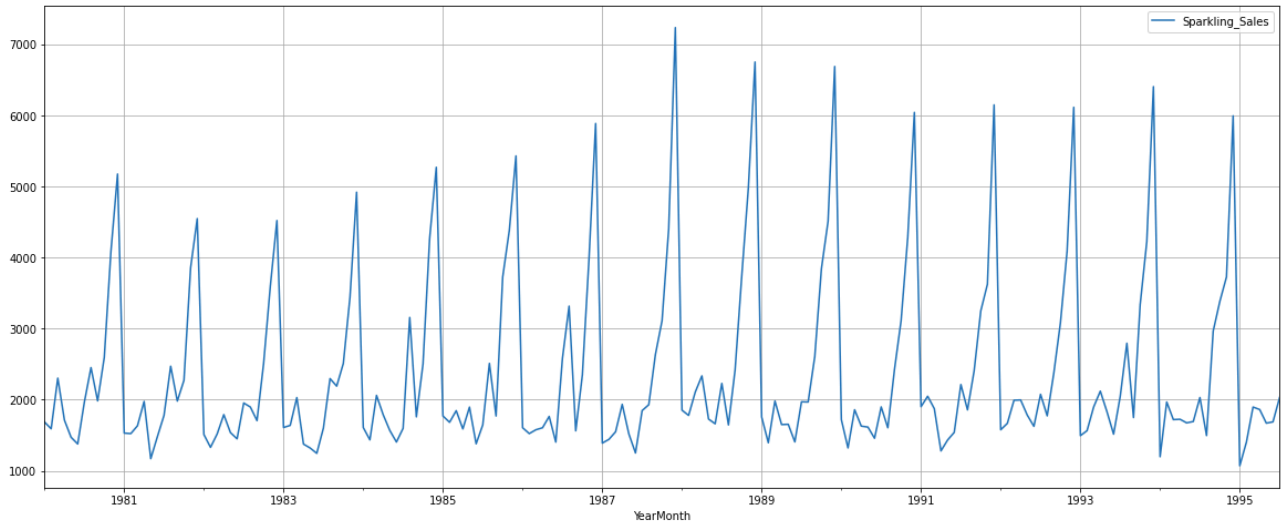
Part 1: Wine Sales analysis and forecast for dataset Sparkling.csv

1.1: Reading and Plotting Time Series Data

Question:
- Read the data as an appropriate Time Series data and
- Plot the data

- The given dataset contains details of Sales of Sparkling Wine over a period of 15 years and 7 months.
    - In all, there are 187 rows in the dataset that is a record of Monthly Sales of Sparkling Wine, from January 1980 to July 1995.
    - The .csv file given is read into a form of Time Series Pandas Dataframe, for the purpose of further analysis.
    - That effectively converts the Timeline of Months into the Index, and the Monthly Sales of Sparkling Wine into the Feature that we will analyse, in this exercise, as a Time Series.

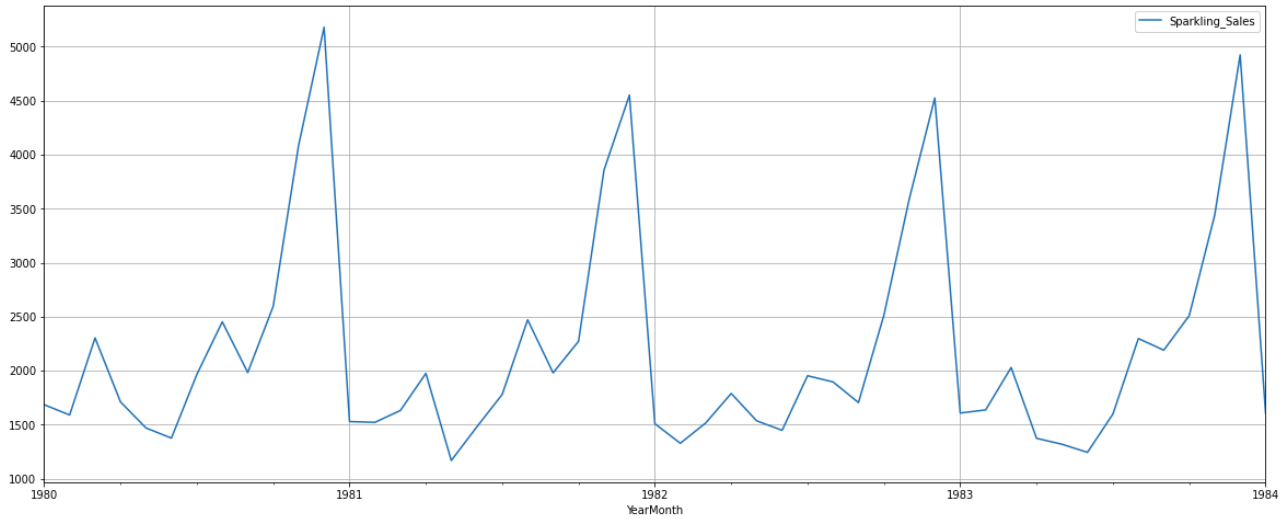
Plotting the Time Series data:



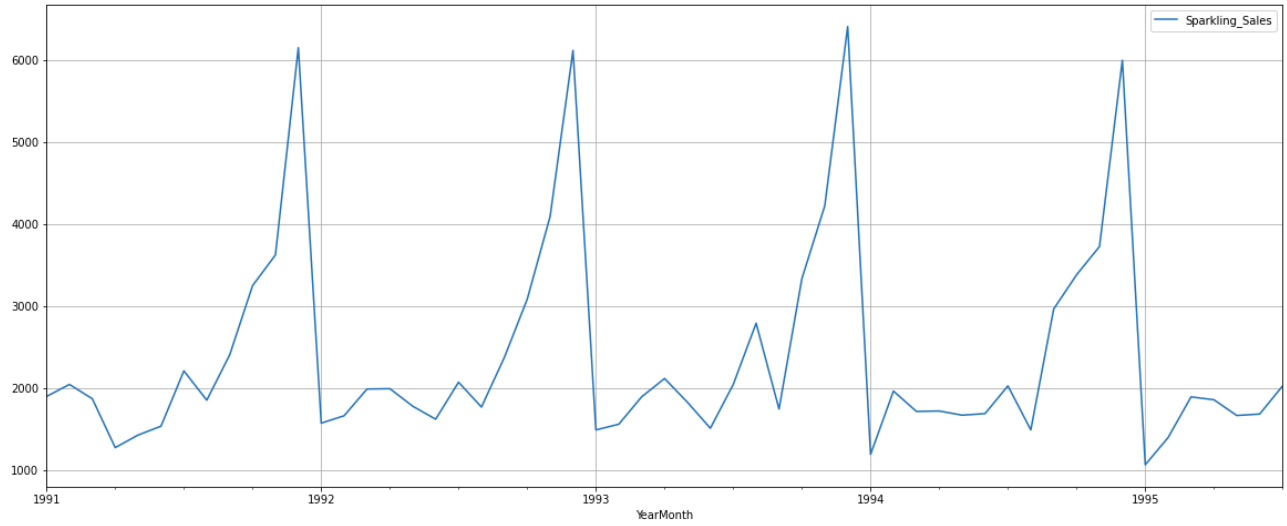
From the above plot, one gets the following impression:
1. There is a predominant seasonal pattern that repeats every year, with sales peaking in the last couple of months
2. There is no discernible overall trend, but there is slight growth in the first half of the period, and then a slight declining pattern in sales in the middle, and a steadying in latter half of this period.
3. There is no missing data in the series

The seasonal behaviour can be better discerned in the following plots, which are plots of the first 4 years and last ~ 4 and a half years.

The first 4 years, starting January 1980:



The last few years, starting January 1991:



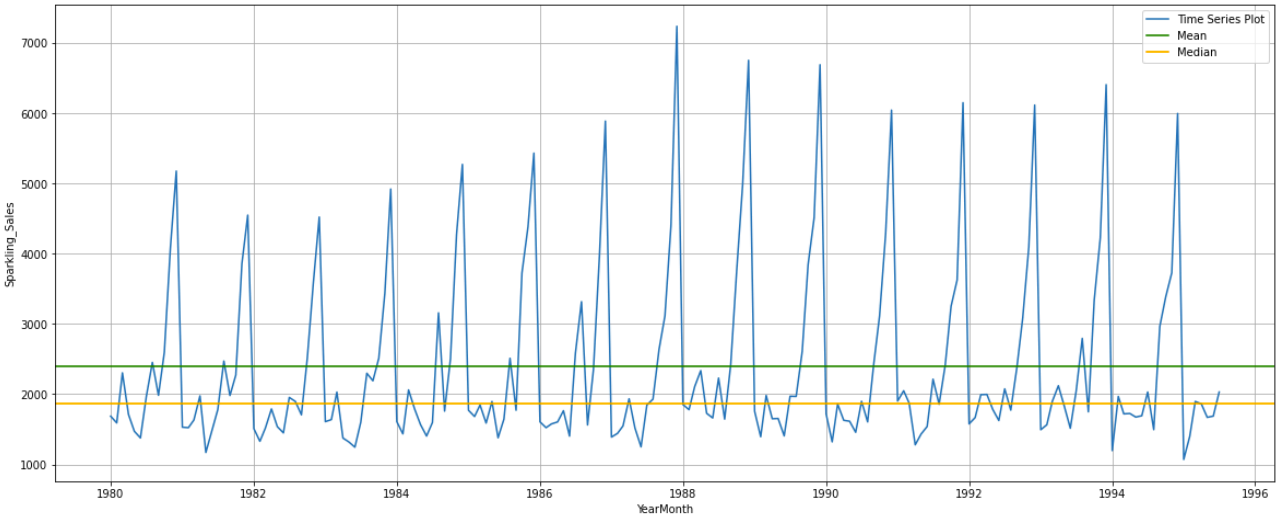
We will explore the data in detail in the next section.

1.2: Exploratory Data Analysis and Time Series Decomposition

Question:
- Perform appropriate Exploratory Data Analysis to understand the data, and also
- Perform Decomposition.

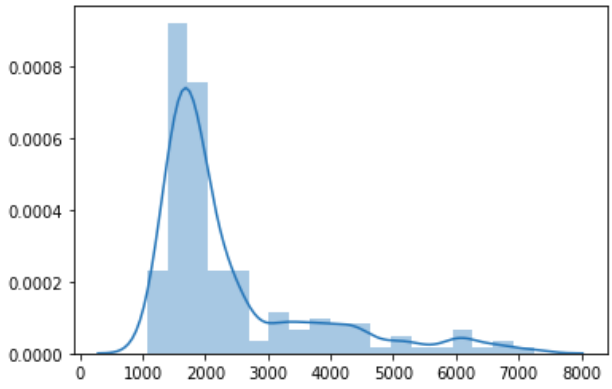
Overall Assessment:

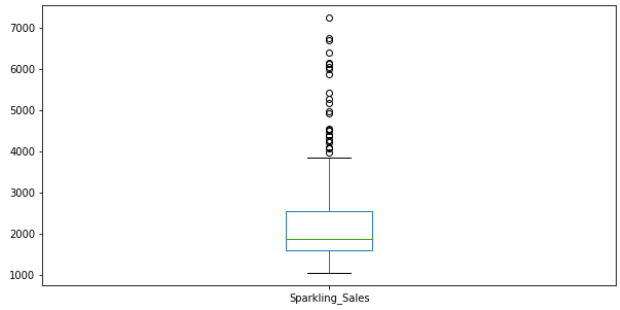
Monthly Sales of Sparkling Wine over a 15 year and 7 month period:



Summary of Sparkling_Sales:

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

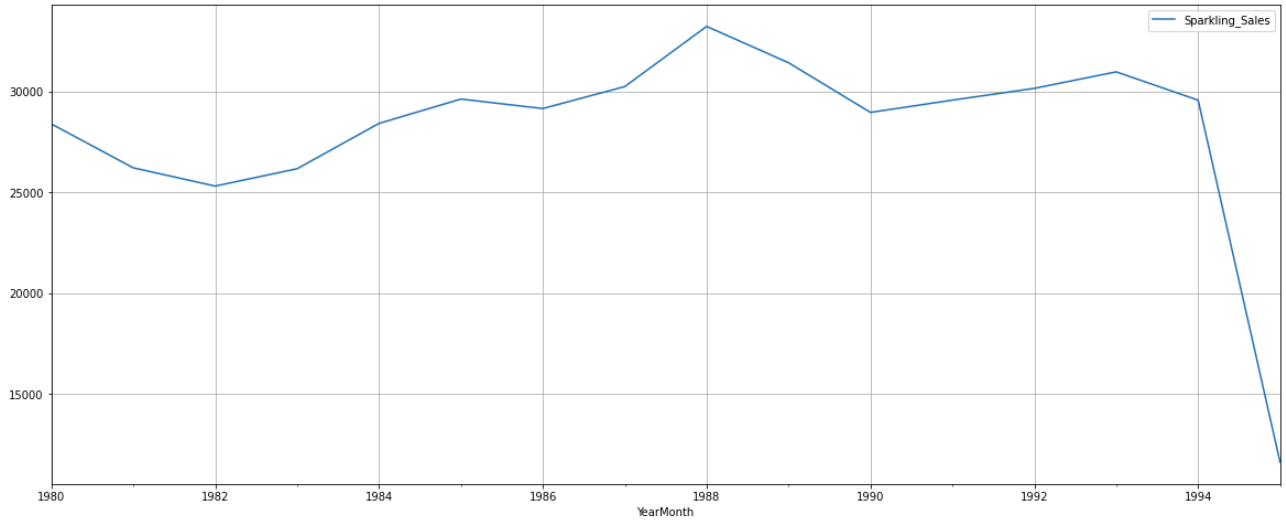




- The skewness in the distribution is evident in the shape of the histogram and boxplot.
- Most monthly sales values are in the 1000-2500 range, but a significant number of months also have very high sales.
- As evidenced from the seasonality in the Time Series, these outlier months are the last couple of months every year, indicating the holiday season sales. Sales in this period are significantly higher than the median monthly sales.

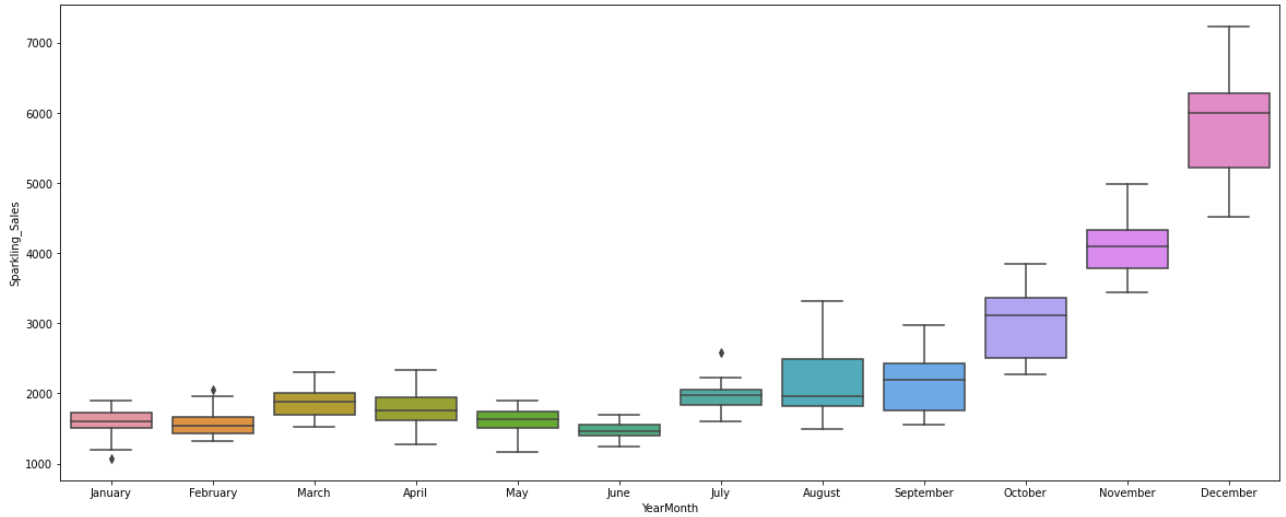
Studying the Overall Trend, by Resampling the data:

Annual Sales of Sparkling Wine over the 15 year and 7 month period:



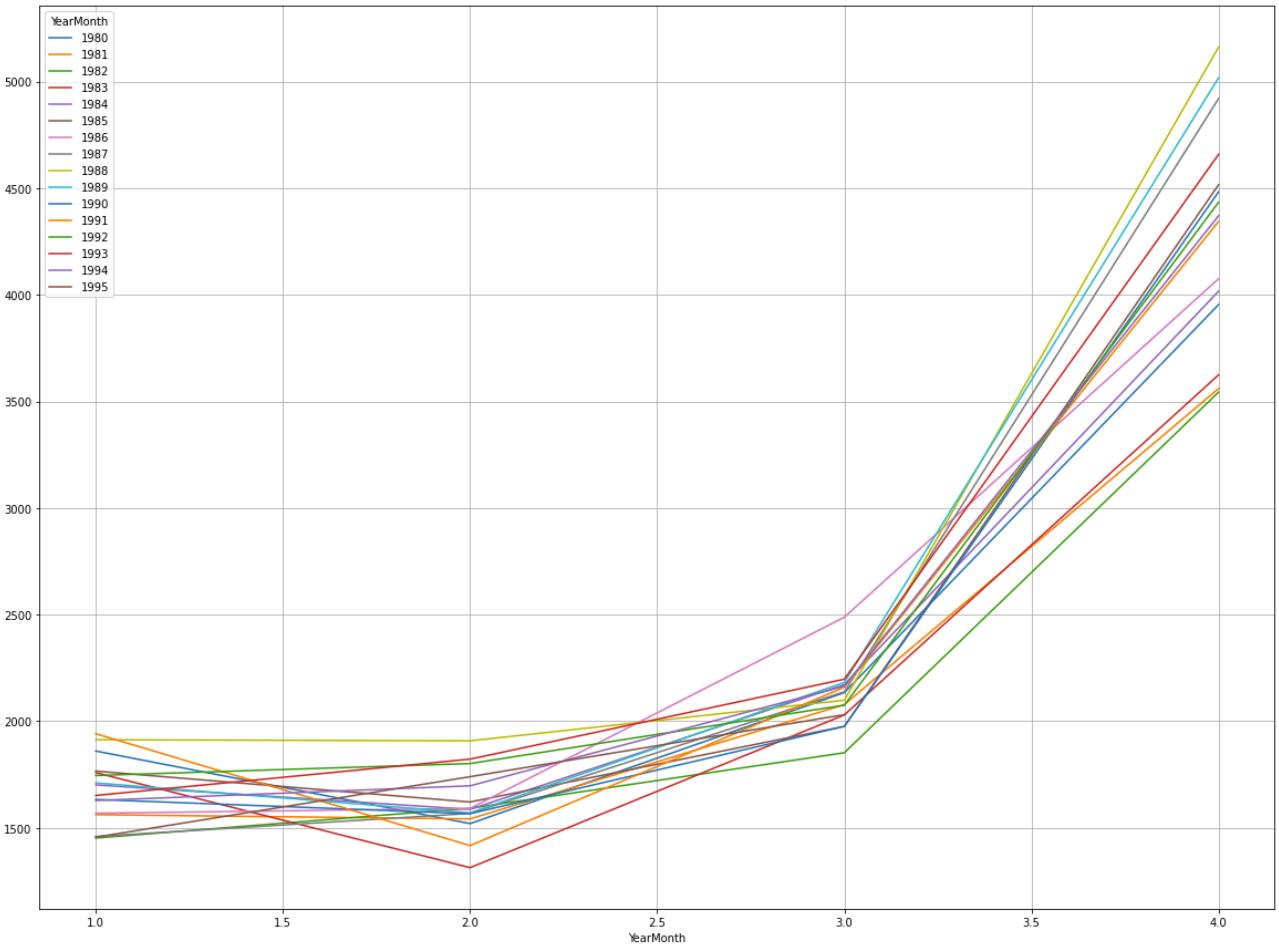
- After an initial decline in sales in the first two years of the given period, there is a steady growth in Sales of Sparkling Wine from year 1982 up to 1988, when it hit its peak of popularity.
- After another quick decline, the popularity of Sparkling Wine appears to have steadied in the latter years.

Seasonality of Wine Sales: using box plots to summarise Monthly Sales of Sparkling Wine across the 15 year and 7 months time period:

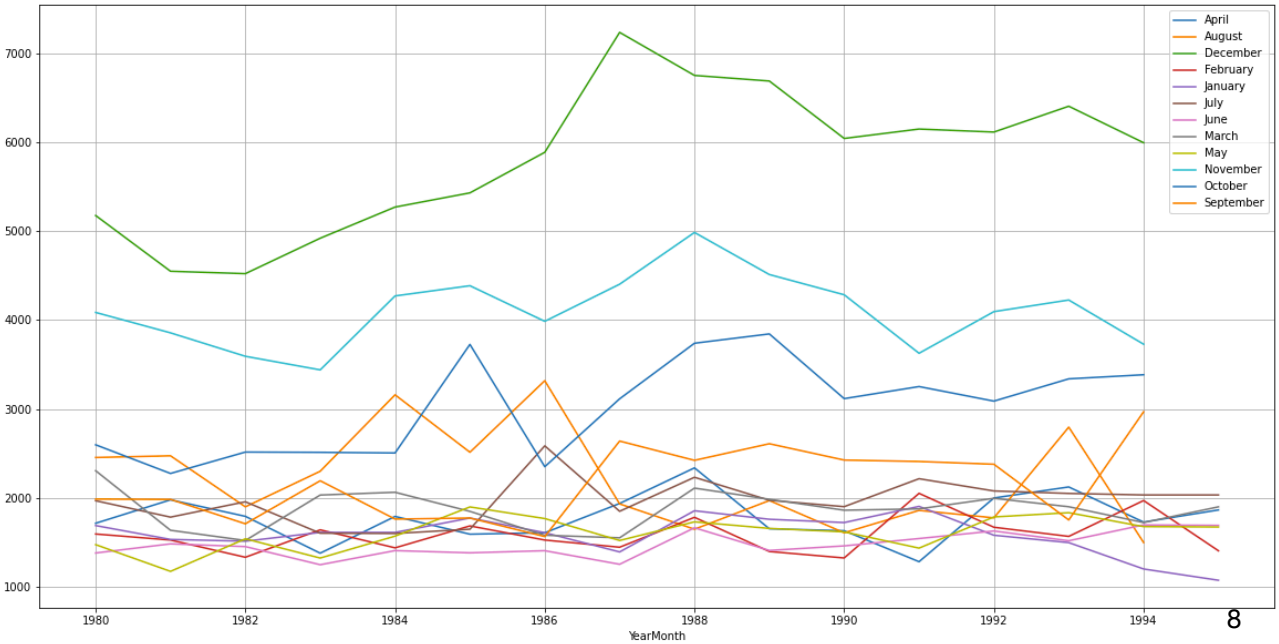


- As observed earlier, Sales of Sparkling Wine are highest in the last quarter, peaking in December.
- The first half of the year sees a significantly lower volume of sales, with March and April accounting for a slightly higher volume.
- Sales pick up post June, with August and September also being good months for Sparkling Wine Sales.

Seasonality of Wine Sales, viewed on a quarterly basis:
Sales of each Quarter, for every year in the dataset - shows the similar pattern of lower sales in the first 2 quarters, with a significant spike especially in the last quarter.

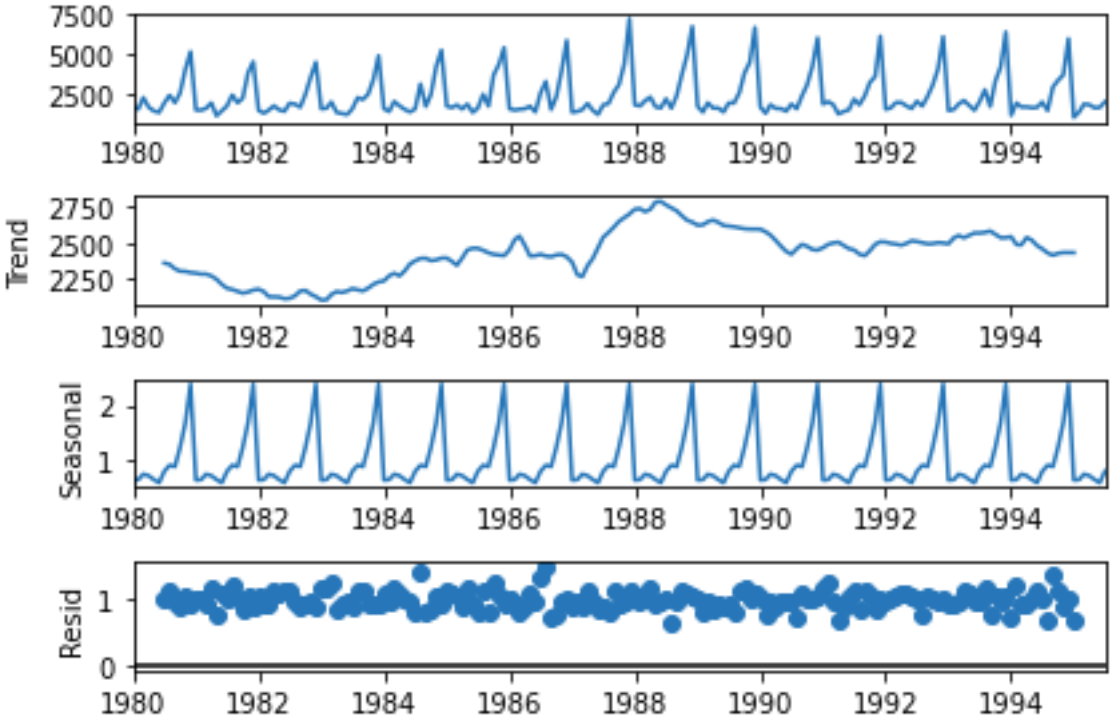


The following is a plot of Monthly Sales across years, which highlights the outsized contribution of December, November and October to the annual Sales of Sparkling Wine. December and November stand out as the 2 most lucrative months, consistently over all years.



Decomposition:

The pattern produced by the Time Series plot suggests a Multiplicative Seasonality. The seasonal pattern's width (frequency) and height (amplitude) vary across time periods. As we see a growth in sales, the seasonal effect also gets amplified. And when sales decline, the seasonal effects also appear diminished.



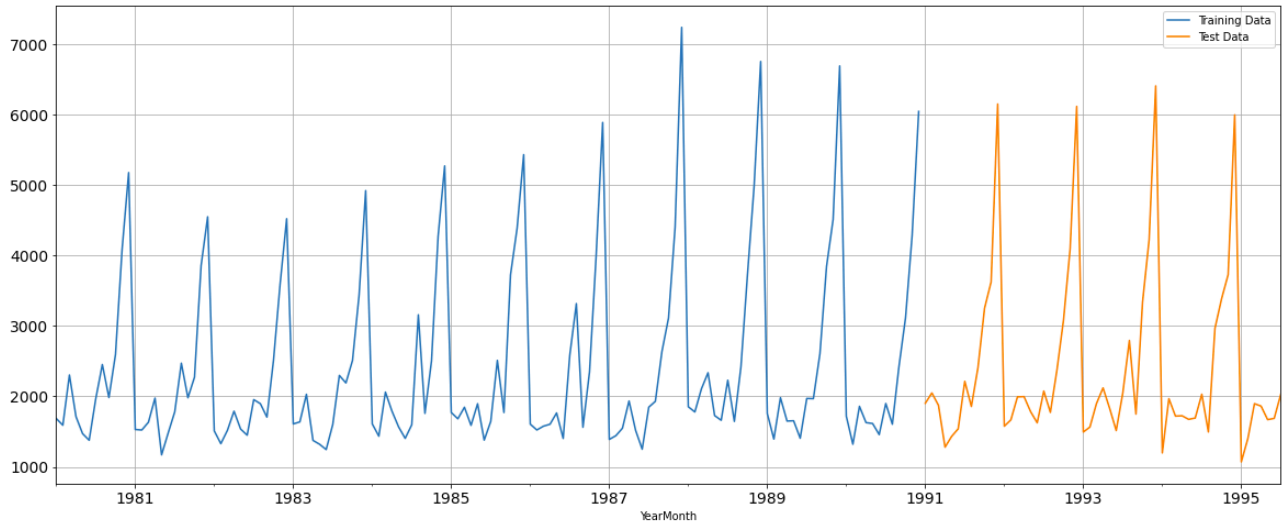
- Decomposition reiterates the predominance of the Seasonal factor.
- There is no clear trend in the data.
- Errors are largely concentrated, and appear to be distributed evenly along the 1 mark.

1.3: Splitting data into Training and Test datasets

Question:
- Split the data into training and test. The test data should start in 1991.

After splitting the data:
- 132 observations, starting January 1980 up to December 1990, comprises the Training Data.
- 55 observations, starting January 1991 up to July 1995, comprises the Test Data.

The split is depicted in the following plot:

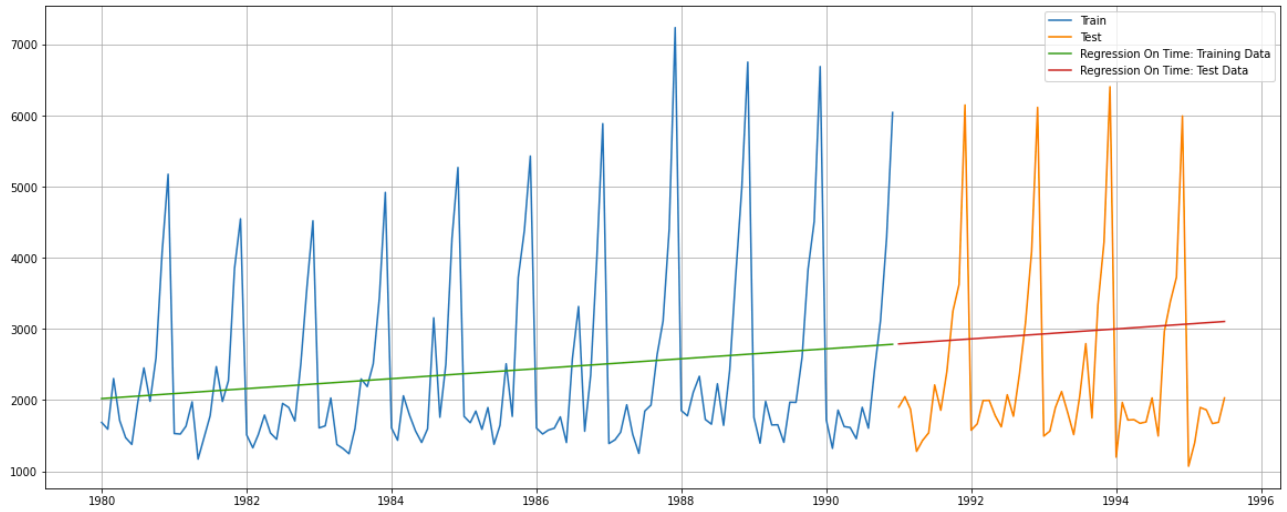


1.4: Model Building: Exponential Smoothing and other models

Question:
- Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.
- Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Model 1: Linear Regression

Model Output Visualised:



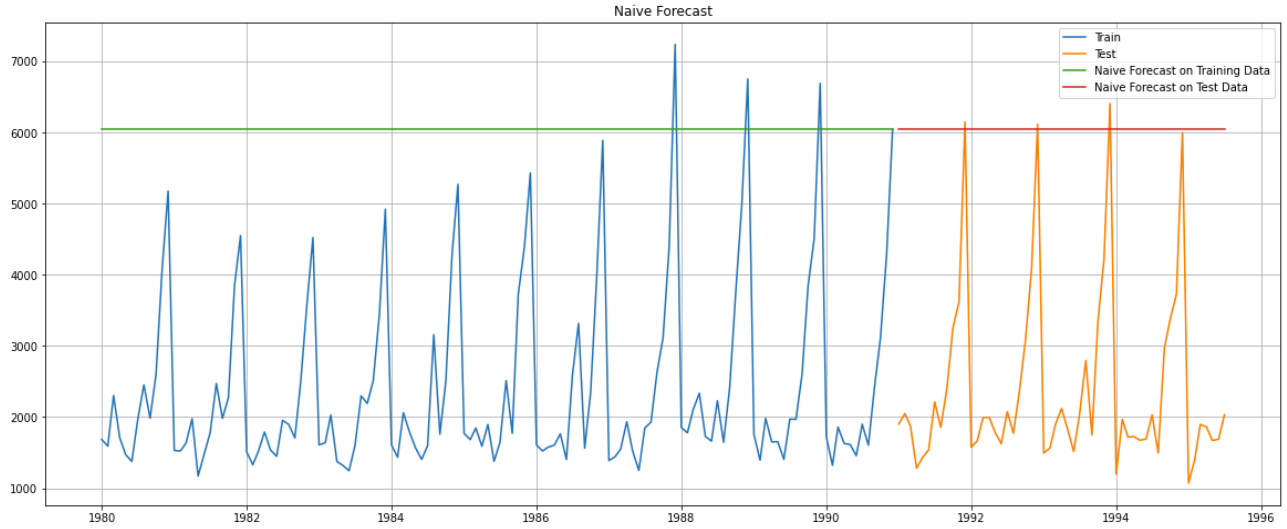
Performance Metrics:

	RMSE	MAPE
Training Data	1279.322	40.05
Test Data	1389.135	50.15

Observation:
A linear regression model is not a great fit for the data, since there is no clear overall trend, and seasonality is predominant. It therefore misses out on much of the variation, resulting in high RMSE scores.

Model 2: Naive

Model Output Visualised:



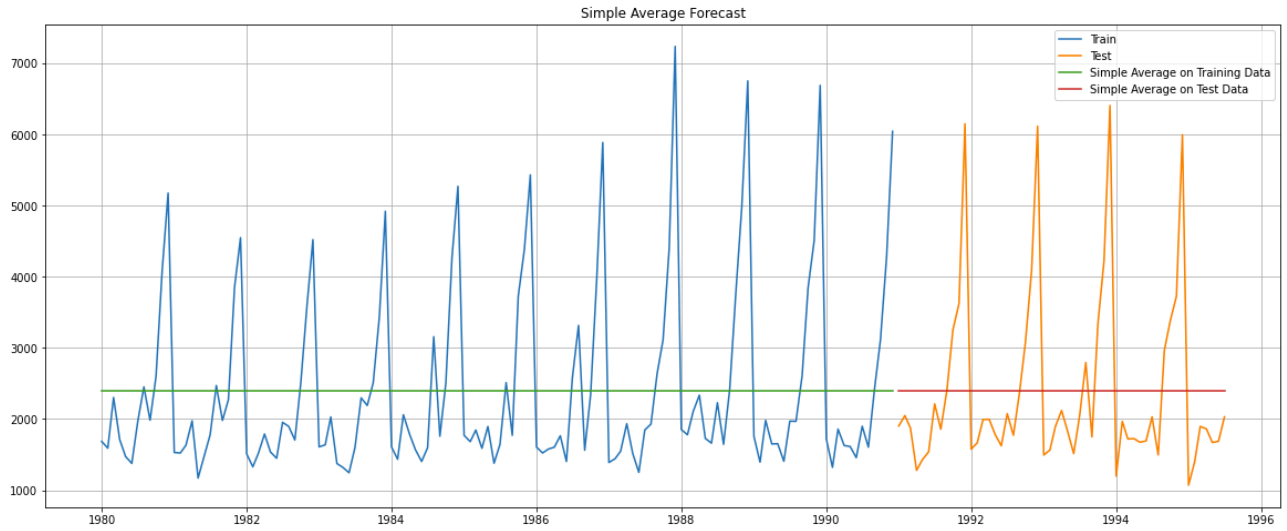
Performance Metrics:

	RMSE	MAPE
Training Data	3867.701	153.17
Test Data	3864.279	152.87

Observation:
The Naive model is dependent on the last observed value, which in our training data is the month of December. We know that the month of December records peak sales every year. So this value is clearly not representative of the dataset at large. Hence the expected very high RMSE scores.

Model 3: Simple Average

Model Output Visualised:



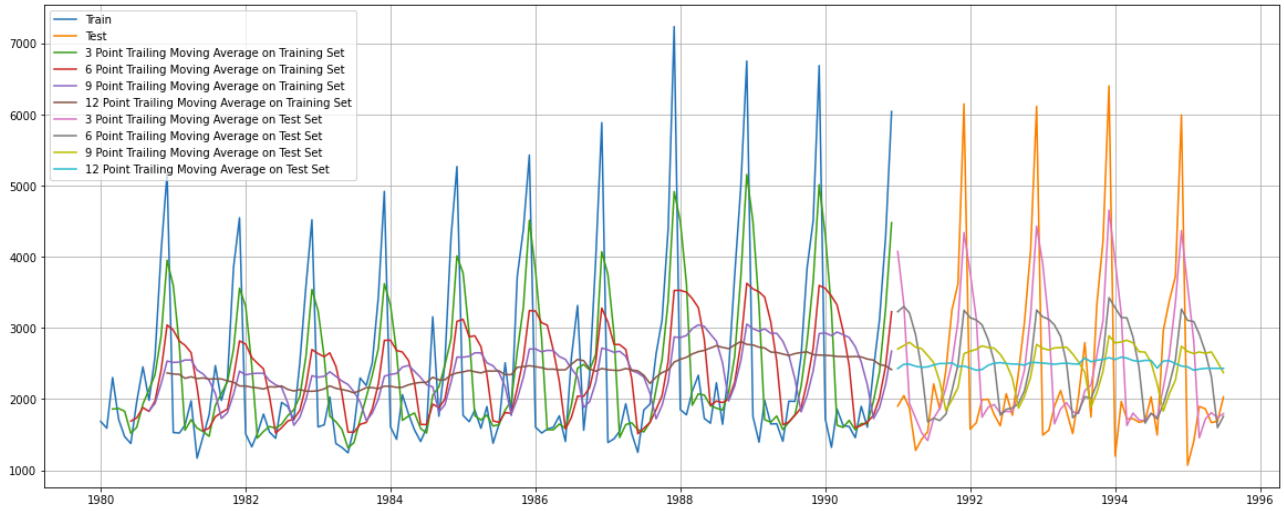
Performance Metrics:

	RMSE	MAPE
Training Data	1298.484	40.36
Test Data	1275.082	38.90

Observation:
A simple average model is not a great fit for the data, since seasonality is predominant in this time series. It therefore misses out on much of the variation, resulting in high RMSE scores.

Model 4: Moving Average/s

Model Output Visualised:



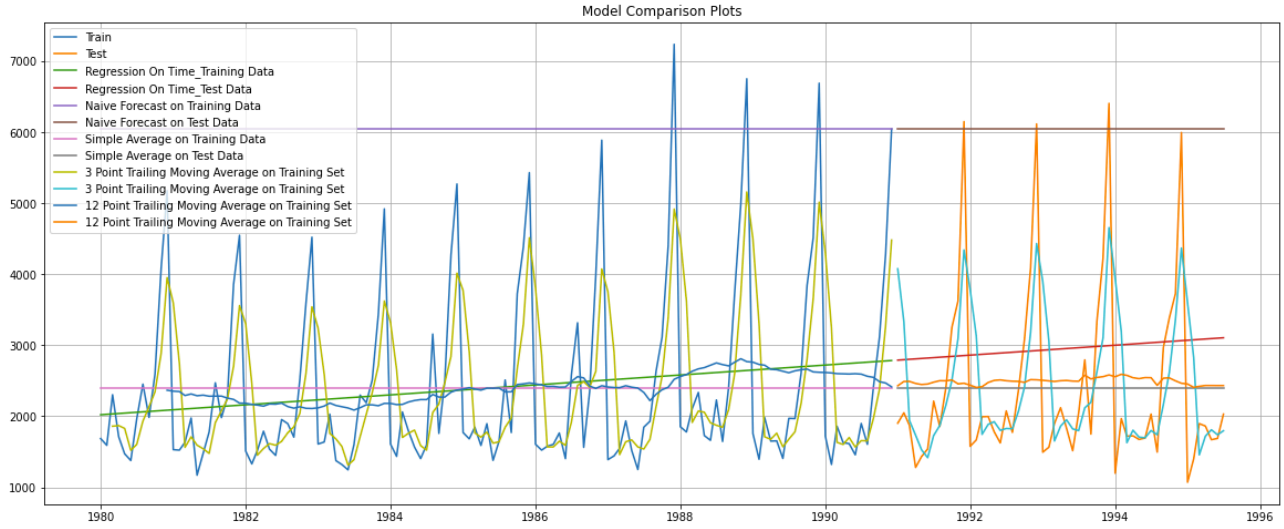
Performance Metrics on Test Data:

	RMSE	MAPE
3 point Moving Average	1028.606	29.73
6 point Moving Average	1283.927	43.86
9 point Moving Average	1346.278	46.86
12 point Moving Average	1267.925	40.19

Observation:
Moving Average Models are able to better track the variation of the time series, especially the 3 point MA in our case. To forecast for a year, we will need to use at least a 12-point MA, which however fails to capture the seasonal variation that is predominant in the data.

Model Comparison

The following is a visual depiction of how the models we've built thus far, perform.

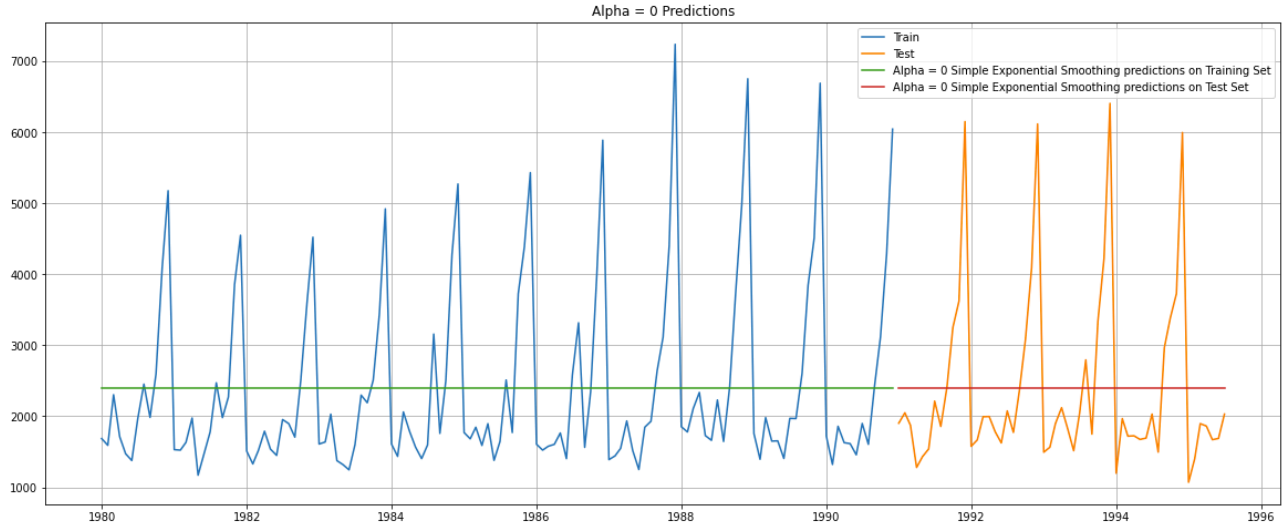


We will now proceed with the Exponential Smoothing Models.

Model 5: Single Exponential Smoothing (Auto-fit, alpha = 0)

SES, Alpha = 0

Model Output Visualised:



Performance Metrics on Test Data:

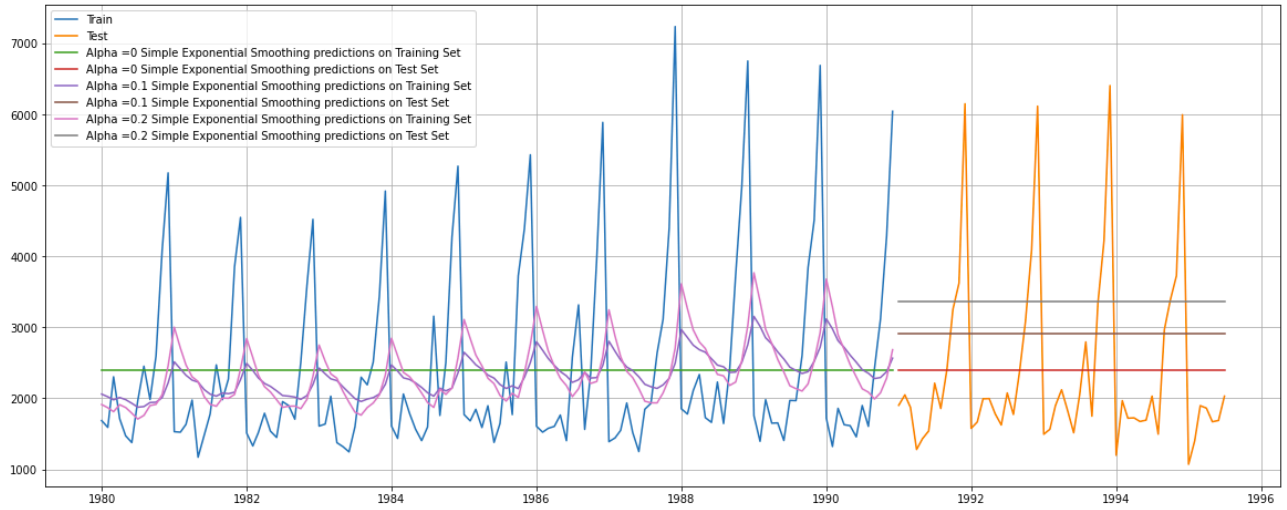
	RMSE	MAPE
Training Data	1028.606	29.73
Test Data	1283.927	43.86

Observation:
While this particular time series has no overall upward or downward trend, the level extrapolated by the SES model will be a useful component. However, since this time series has a predominant seasonal characteristic, it will not be able to capture the variation in the data.

Model 5a: Single Exponential Smoothing (using a Range of alpha values)

SES, Alpha ranging from 0.1 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0, 0.1 and 0.2)



Performance Metrics on Test Data:

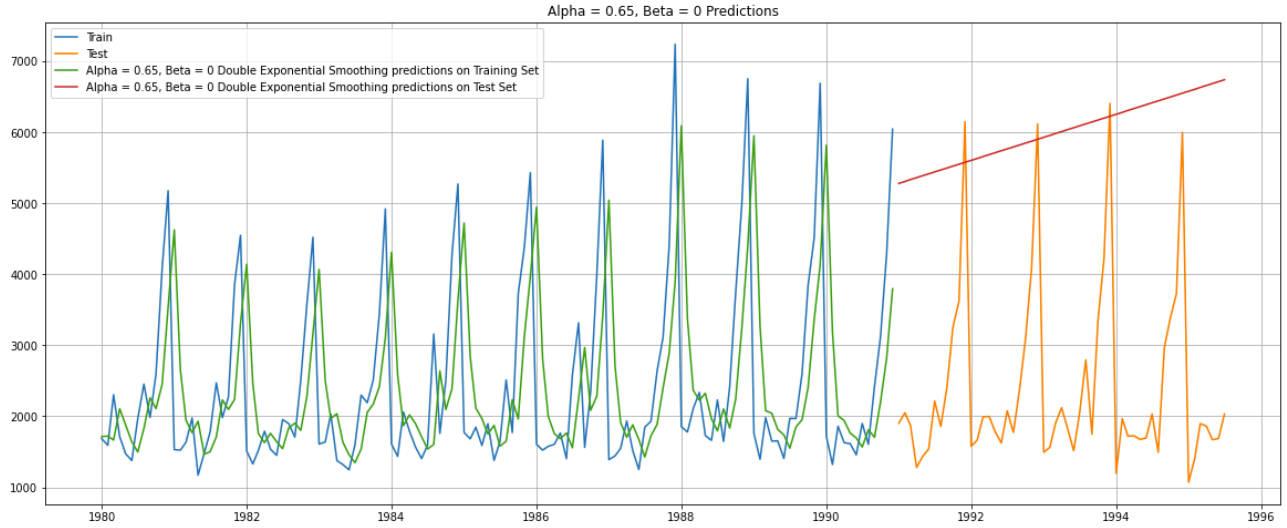
	RMSE	MAPE
Alpha = 0.1	1375.394	49.53
Alpha = 0.2	1595.207	60.46

Observation:
While this particular time series has no overall upward or downward trend, the level extrapolated by the SES model will be a useful component. However, since this time series has a predominant seasonal characteristic, it will not be able to capture the variation in the data.

Model 6: Double Exponential Smoothing (Auto-fit, alpha = 0.65, beta = 0)

DES (Holt's Model), Alpha = 0.65, Beta = 0

Model Output:



Performance Metrics on Test Data:

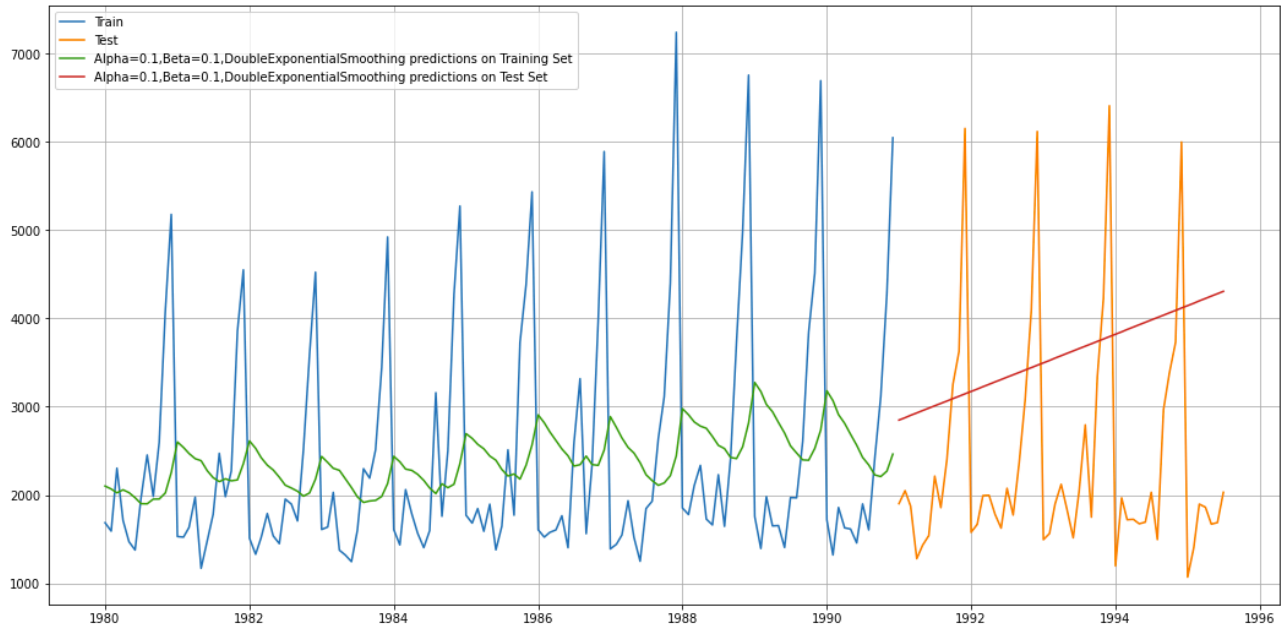
	RMSE	MAPE
Training Data	1337.484	39.11
Test Data	3850.969	152.06

Observation:
The Holt's model isn't a very good fit for this time series as the data has no trend, and strong seasonality, as evidenced by the high RMSE.

Model 6a: Double Exponential Smoothing (using a Range of alpha, beta values)

DES, Alpha ranging from 0.1 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and beta = 0.1)



Performance Metrics:

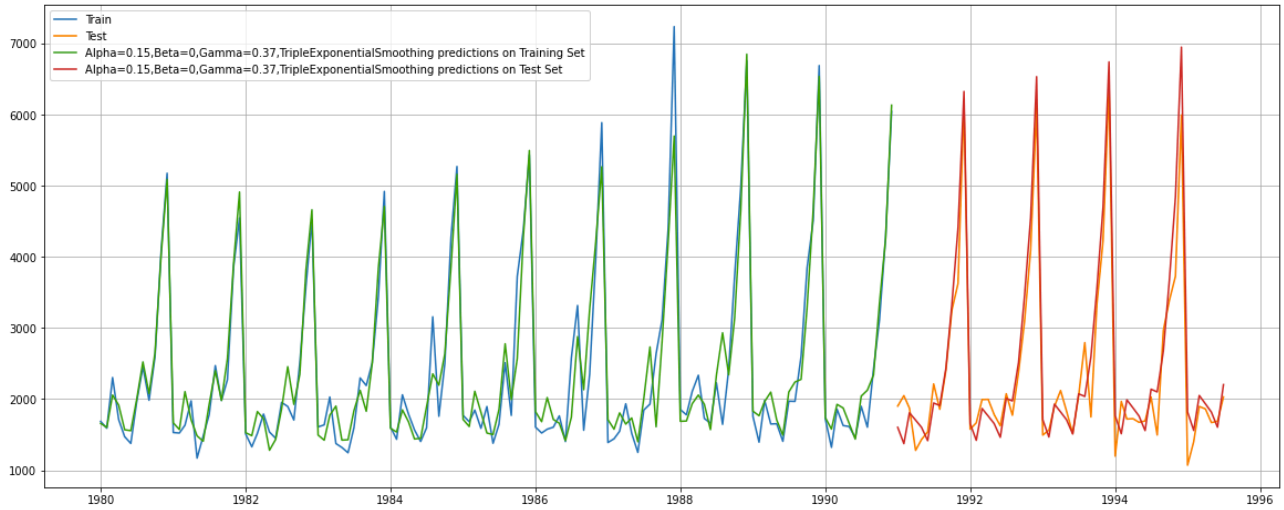
	RMSE	MAPE
Training Data	1363.474	44.26
Test Data	1779.425	67.23

Observation:
The Holt's model isn't a very good fit for this time series as the data has no trend, and strong seasonality, as evidenced by the high RMSE.

Model 7: Triple Exponential Smoothing (Auto-fit: Alpha=0.15, Beta=0, Gamma=0.37)

TES (Holt-Winters' Model), Alpha=0.15, Beta=0, Gamma=0.37

Model Output:



Performance Metrics on Test Data:

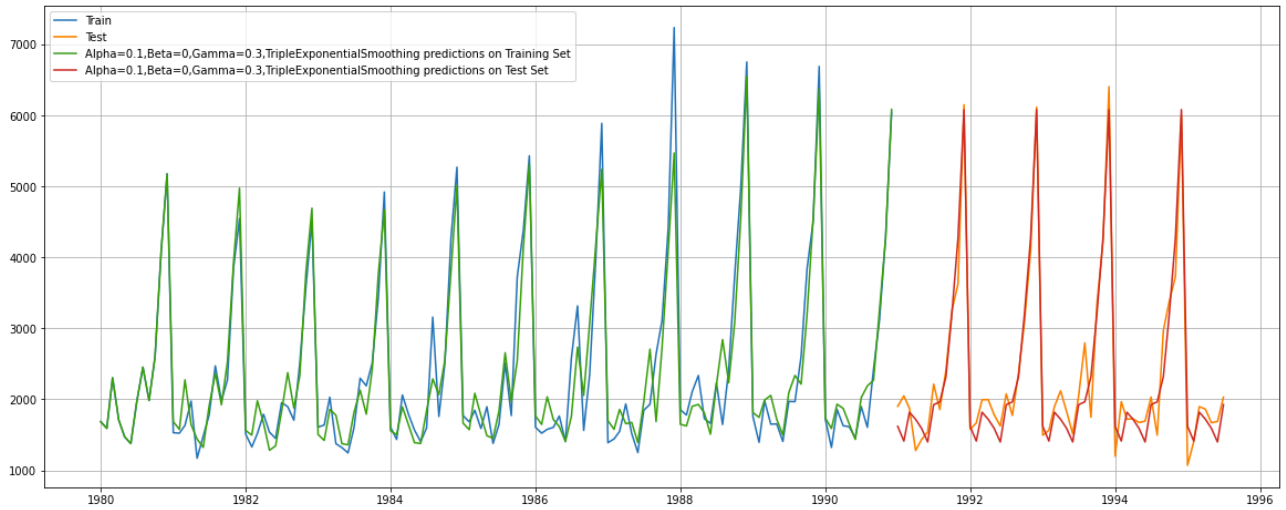
	RMSE	MAPE
Training Data	353.379	10.18
Test Data	384.159	11.94

Observation:
The Triple Exponential Smoothing model is a good fit for this time series, owing to the strong Seasonal component. This is corroborated by a low RMSE.

Model 7a: Triple Exponential Smoothing (using a Range of alpha, beta, gamma values)

TES, Alpha, Beta, Gamma ranging from 0 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and beta = 0.0, gamma = 0.3)



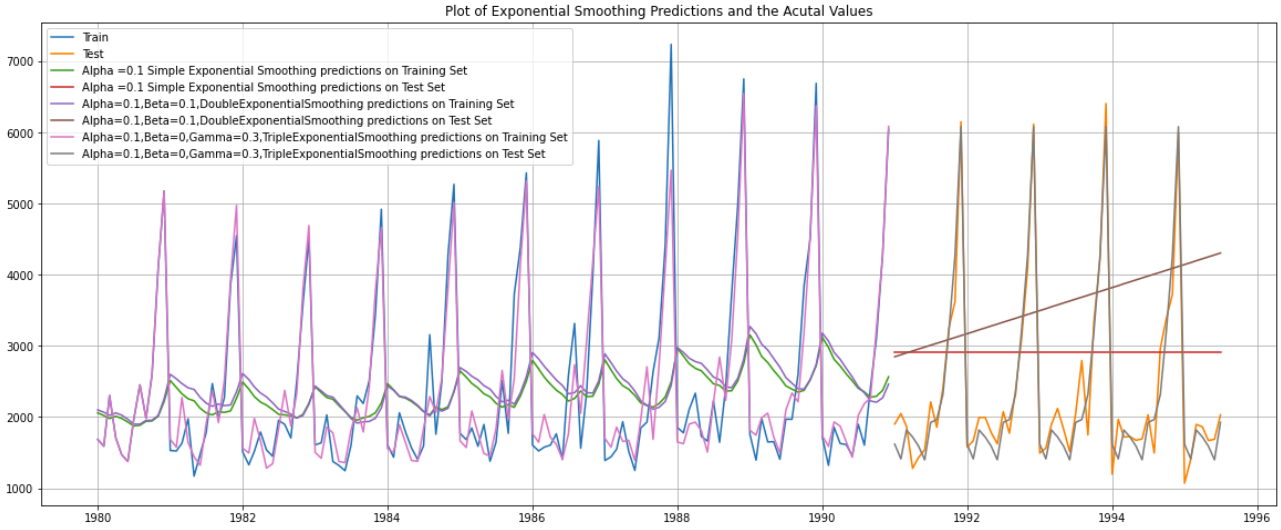
Performance Metrics:

	RMSE	MAPE
Training Data	362.136	10.18
Test Data	303.904	9.67

Observation:
The Triple Exponential Smoothing model is a good fit for this time series, owing to the strong Seasonal component. This is corroborated by a low RMSE.

A Consolidated Plot of all the Exponential Models built:

We see that the Triple Exponential Models are the best suited for this time series.



1.5: Stationarity Check

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

The **Augmented Dickey Fuller (ADF) Test** is a statistical test for affirming whether or not a time series is Stationary.

The Null Hypothesis H0 is: Time Series is Non-stationary
The Alternative Hypothesis H1 is: Time Series is Stationary

TEST 1: We administer the ADF test on the Original Time Series.

Results of Dickey-Fuller Test:
Test Statistic: -1.208926
p-value: 0.669744

With the resultant ADF test p-value at 0.67, we cannot reject the Null Hypothesis (at alpha 0.05).

We hence conclude that the Time Series is Non-stationary.

In order to make a Time Series Stationary, we need to transform the original series by taking a Difference of the original values. Usually, a 1 period difference suffices to transform a Non-Stationary series into a Stationary one.

TEST 2: We administer the ADF test on the new series - derived by taking a 1 period Difference of the original series.

Results of Dickey-Fuller Test:
Test Statistic: -8.005007e+00
p-value: 2.280104e-12

The resultant ADF p-value (0.000000000002) is significantly less than 0.05 (alpha).
We can hence reject the null hypothesis in the case of the new series, which is derived by differencing the original series over 1 period.

We conclude that at Difference 1, the time series is Stationary.

1.6: Model Building: Automated ARIMA / SARIMA

Question:
- 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.

Model 8: ARIMA (Lowest AIC parameters: p=2, d=1, q=2)

Of the ARIMA models generated using various combinations of parameters p and q, the model with the lowest AIC score was: ARIMA (2, 1, 2) with an AIC score of 2210.617708

Model Summary:

ARIMA Model Results

Dep. Variable:	D.Sparkling_Sales	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.309
Method:	css-mle	S.D. of innovations	1012.446
Date:	Wed, 12 Aug 2020	AIC	2210.618
Time:	14:32:11	BIC	2227.869
Sample:	02-01-1980 - 12-01-1990	HQIC	2217.628

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5856	0.517	10.814	0.000	4.573	6.598
ar.L1.D.Sparkling_Sales	1.2699	0.074	17.046	0.000	1.124	1.416
ar.L2.D.Sparkling_Sales	-0.5602	0.074	-7.618	0.000	-0.704	-0.416
ma.L1.D.Sparkling_Sales	-1.9983	0.042	-47.168	0.000	-2.081	-1.915
ma.L2.D.Sparkling_Sales	0.9983	0.042	23.560	0.000	0.915	1.081

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1335	-0.7073j	1.3361	-0.0888
AR.2	1.1335	+0.7073j	1.3361	0.0888
MA.1	1.0003	+0.0000j	1.0003	0.0000
MA.2	1.0013	+0.0000j	1.0013	0.0000

Performance Metrics on Test Data:

	RMSE	MAPE
ARIMA (2,1,2)	1374.761	48.37

Observation:
ARIMA does not factor the seasonal component which is very characteristic of this time series, and hence wouldn't be an ideal model.

Model 9: SARIMA (Lowest AIC parameters: p=2, d=1, q=2, P=0, D=1, Q=2)

Of the SARIMA models generated using various combinations of parameters p, q, P, Q and D, the model with the lowest AIC score was: SARIMA (1, 1, 2)x(0, 1, 2, 12) with an AIC score of 1382.347780

Model Summary:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-685.174
Date:	Wed, 12 Aug 2020	AIC	1382.348
Time:	15:02:14	BIC	1397.479
Sample:	0	HQIC	1388.455
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5507	0.287	-1.922	0.055	-1.112	0.011
ma.L1	-0.1612	0.235	-0.687	0.492	-0.621	0.299
ma.L2	-0.7218	0.175	-4.132	0.000	-1.064	-0.379
ma.S.L12	-0.4062	0.092	-4.401	0.000	-0.587	-0.225
ma.S.L24	-0.0274	0.138	-0.198	0.843	-0.298	0.243
sigma2	1.705e+05	2.45e+04	6.956	0.000	1.22e+05	2.19e+05

Ljung-Box (Q):	20.51	Jarque-Bera (JB):	13.48
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.89	Skew:	0.60
Prob(H) (two-sided):	0.75	Kurtosis:	4.44

Performance Metrics on Test Data:

	RMSE	MAPE
SARIMA (1,1,2) (0,1,2,12)	382.577	12.87

Observation:
A SARIMA model is better equipped for a time series with a very strong seasonal component. And this model matches up in performance to the TES model which is thus far the strongest model for this time series.

1.7: Model Building: ARIMA / SARIMA using ACF, PACF cut-offs

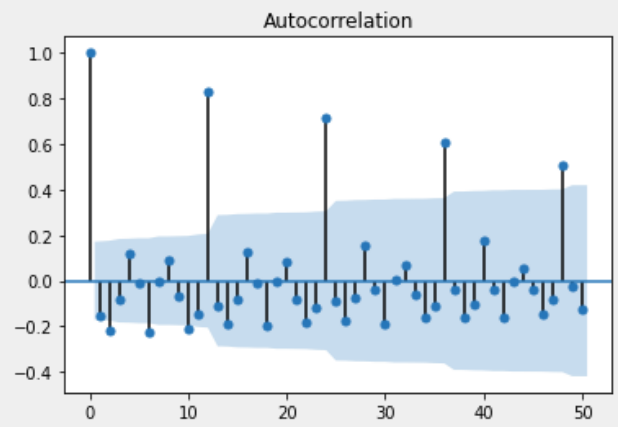
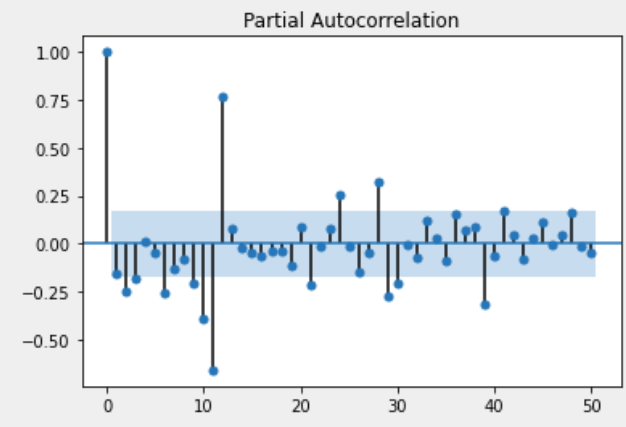
Question:
- Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data, and
- Evaluate this model on the test data using RMSE.

From the ADF test conducted in an earlier section, we know that the original time series is Non-stationary.

So the first step would be to transform the time series by differencing the original series over 1 period, and make it Stationary.

We then need to plot the ACF and PACF on the transformed stationary Time Series.

The following are the ACF and PACF plots:

ACF Plot	PACF Plot
	

The ACF plot:
- The cut-off appears right after lag 0: Lag 1 appears to be insignificant, though only marginally.
- The first significant point is at lag 2, after which the subsequent lags are again insignificant.
- Hence based on the ACF plot, we can assign 2 as the order of the MA component (q) of the ARIMA/SARIMA model.

The PACF plot:
- The cut-off here too appears right after lag 0: Lag 1 appears to be insignificant, though only marginally.
- The first significant point is at Lag 2.
- Lag 3 too appears to be significant, although it is right on the border.
- Hence based on the PACF plot, we can assign 3 as the order of the AR component (p) of the ARIMA/SARIMA model.

- We know that the Seasonal component is very apparent in the series. The ACF plot also clearly shows a pattern repeating every year - indicating strong seasonal behaviour.
- So a SARIMA model would be appropriate for modelling such a series. For employing the seasonal component we will accord value to a 12 period lag, which will map to a value of 1 to both P and Q.
- The value of d will be 1, as the series has been Differenced by 1 period, in order to make it stationary.

Hence based on the ACF and PACF plots, we can develop a SARIMA (3,1,2)x(1,1,1) model.

Model 9a: SARIMA (ACF, PACF plot parameters: p=3, d=1, q=2, P=1, D=1, Q=1)

Model Summary:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-685.174
Date:	Wed, 12 Aug 2020	AIC	1382.348
Time:	15:02:14	BIC	1397.479
Sample:	0	HQIC	1388.455
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5507	0.287	-1.922	0.055	-1.112	0.011
ma.L1	-0.1612	0.235	-0.687	0.492	-0.621	0.299
ma.L2	-0.7218	0.175	-4.132	0.000	-1.064	-0.379
ma.S.L12	-0.4062	0.092	-4.401	0.000	-0.587	-0.225
ma.S.L24	-0.0274	0.138	-0.198	0.843	-0.298	0.243
sigma2	1.705e+05	2.45e+04	6.956	0.000	1.22e+05	2.19e+05

Ljung-Box (Q):	20.51	Jarque-Bera (JB):	13.48
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.89	Skew:	0.60
Prob(H) (two-sided):	0.75	Kurtosis:	4.44

Performance Metrics on Test Data:

	RMSE	MAPE
SARIMA (1,1,2) (0,1,2,12)	393.01	13.27

Observations:
As seen in the previous instance, A SARIMA model is better equipped for a time series with very strong seasonal component. And this model, too, comes close to matching up in performance to the TES model.

1.8: Model performance comparison

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

All models, ranked according to their performance. The lowest Test RMSE score is the highest ranked or the best performing model.

		Test RMSE	Test MAPE
1	Triple Exponential Smoothing, Alpha=0.1,Beta=0,Gamma=0.3	303.904462	9.67
2	SARIMA (1, 1, 2) x (0, 1, 2, 12)	382.576723	12.87
3	Triple Exponential Smoothing, Alpha=0.15,Beta=0,Gamma=0.37	384.158614	11.94
4	SARIMA (3, 1, 2) x (1, 1, 1, 12)	393.099897	13.27
5	3 point Trailing Moving Average	1028.605756	29.73
6	12 point Trailing Moving Average	1267.92533	40.19
7	Simple Average Model	1275.081804	38.9
8	Simple Exponential Smoothing, Alpha=0	1275.081813	38.9
9	6 point Trailing Moving Average	1283.927428	43.86
10	9 point Trailing Moving Average	1346.278315	46.86
11	ARIMA (2, 1, 2)	1374.761115	48.37
12	Simple Exponential Smoothing, Alpha=0.1	1375.393526	49.53
13	Regression On Time	1389.135175	50.15
14	Single Exponential Smoothing, Alpha=0.2	1595.206839	60.46
15	Double Exponential Smoothing, Alpha=0.1,Beta=0.1	1779.42476	67.23
16	Double Exponential Smoothing, Alpha=0.65,Beta=0	3850.968926	152.06
17	Naive Model	3864.279352	152.87

1.9: Optimum Model and Forecasting

Question:
- 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.

From our comparison between models, **the top 2 models** are:
1. Triple Exponential Smoothing Model (alpha=0.1, beta = 0, gamma = 0.3)
2. SARIMAX (1,1,2)x(0,1,2,12).

We will re-build these 2 models on the complete data, and make 12 month forecasts.

Part A: TES model on complete data

Building a Triple Exponential Smoothing (alpha = 0.1, beta = 0, gamma = 0.3) model on the complete data, and forecasting for the next 12 months

Model Summary:

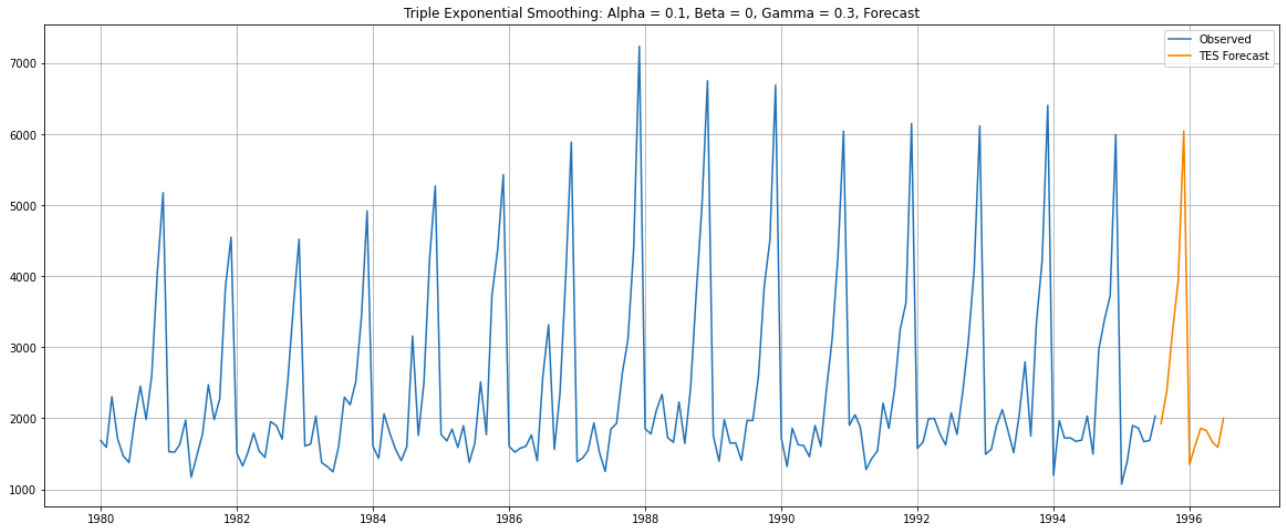
ExponentialSmoothing Model Results

Dep. Variable:	endog	No. Observations:	187
Model:	ExponentialSmoothing	SSE	23177578.963
Optimized:	True	AIC	2225.059
Trend:	Additive	BIC	2276.757
Seasonal:	Multiplicative	AICC	2229.130
Seasonal Periods:	12	Date:	Fri, 14 Aug 2020
Box-Cox:	False	Time:	20:47:45
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.1000000	alpha	False
smoothing_slope	0.000000	beta	False
smoothing_seasonal	0.3000000	gamma	False
initial_level	1580.0000	l.0	True
initial_slope	0.0100000	b.0	True
initial_seasons.0	1.0670886	s.0	True
initial_seasons.1	1.0069620	s.1	True
initial_seasons.2	1.4582278	s.2	True
initial_seasons.3	1.0835443	s.3	True
initial_seasons.4	0.9310127	s.4	True
initial_seasons.5	0.8715190	s.5	True
initial_seasons.6	1.2443038	s.6	True
initial_seasons.7	1.5525316	s.7	True
initial_seasons.8	1.2556962	s.8	True
initial_seasons.9	1.6430380	s.9	True
initial_seasons.10	2.5867089	s.10	True
initial_seasons.11	3.2778481	s.11	True

TES (alpha=0.1, beta = 0, gamma = 0.3) Model 12-month Forecast:

Timeline	Sparkling Wine Sales Forecast
1995-08-01	1923.569548
1995-09-01	2388.275735
1995-10-01	3209.004743
1995-11-01	3932.152285
1995-12-01	6049.534727
1996-01-01	1350.823521
1996-02-01	1624.617929
1996-03-01	1861.078429
1996-04-01	1824.834459
1996-05-01	1671.371518
1996-06-01	1590.584542
1996-07-01	1998.549369



Part B: SARIMA model on complete data

Building a SARIMAX (1,1,2)x(0,1,2,12) model on the complete data, and forecasting for the next 12 months

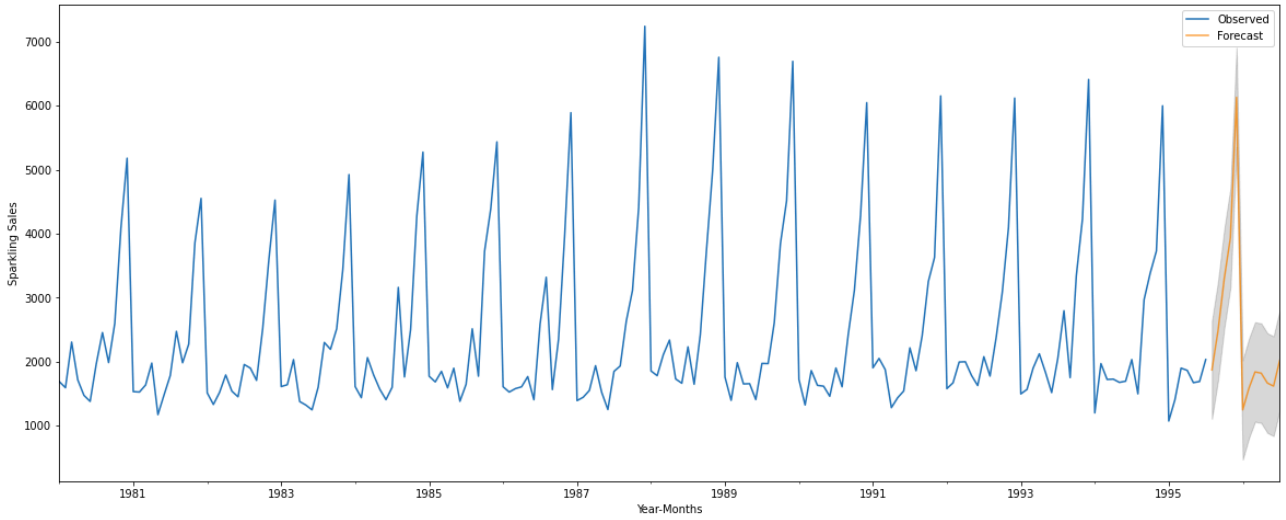
Model Summary:

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  187
Model:             SARIMAX(1, 1, 2)x(0, 1, 2, 12)   Log Likelihood               -1086.537
Date:                            Fri, 14 Aug 2020   AIC                           2185.074
Time:                                    21:04:10   BIC                           2203.017
Sample:                                         0   HQIC                          2192.364
                                              - 187
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5604      0.367     -1.528      0.127      -1.279       0.159
ma.L1         -0.2809      0.339     -0.830      0.407      -0.945       0.383
ma.L2         -0.6547      0.311     -2.102      0.036      -1.265      -0.044
ma.S.L12      -0.5443      0.068     -8.052      0.000      -0.677      -0.412
ma.S.L24      -0.0177      0.085     -0.207      0.836      -0.185       0.150
sigma2      1.515e+05   1.54e+04      9.820      0.000    1.21e+05    1.82e+05
===================================================================================
Ljung-Box (Q):                       18.57   Jarque-Bera (JB):                36.50
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.81   Skew:                             0.67
Prob(H) (two-sided):                  0.46   Kurtosis:                         5.04
===================================================================================
```

SARIMA model 12 month Forecast:

timeline	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	1869.770067	389.241605	1106.870539	2632.669594
1995-09-01	2484.479875	394.108109	1712.042176	3256.917575
1995-10-01	3294.052682	394.223309	2521.389194	4066.716169
1995-11-01	3932.866644	395.394534	3157.907597	4707.825691
1995-12-01	6131.680558	395.475843	5356.562149	6906.798966
1996-01-01	1245.188368	396.010406	469.022235	2021.354502
1996-02-01	1580.094096	396.241732	803.474572	2356.713619
1996-03-01	1837.50603	396.62706	1060.131278	2614.880783
1996-04-01	1818.652204	396.920827	1040.701677	2596.60273
1996-05-01	1664.131274	397.263937	885.508265	2442.754284
1996-06-01	1615.681123	397.578529	836.441526	2394.920721
1996-07-01	2016.79348	397.908527	1236.907098	2796.679862

SARIMA (1,1,2)x(0,1,2,12) model 12-month Forecast visualised:



1.10: Insights and Findings

Question:
- Comment on the model thus built and report your findings, and
- Suggest the measures that the company should be taking for future sales.

- The Sparkling Wine Sales have steadied over the last four years, and the levels don't show any significant growth or decline.
- The seasonal patterns recur consistently, especially the last quarter which sees the maximum sales every year.
- The models appear to forecast a similar range and pattern for the next 12 months. And given the low RMSE score, along with a certain consistency of past behaviour, the forecast looks dependable.
- The consumption pattern points to the fact that Sparkling Wines are most in demand in holiday and festive seasons, and is therefore positioned as a premium product, and meant for special occasions.
- To increase sales, we could look to capitalise on the last quarter demand, and create campaigns to push more consumption during the period. This would maintain the special-ness of the product.
- We could alternatively promote Sparkling Wine consumption during other festive periods and other seasons conducive for it. There are smaller spikes in consumption during March and June, which can be explored.
- Also, other geographies can be explored to tap into other festive seasons globally.

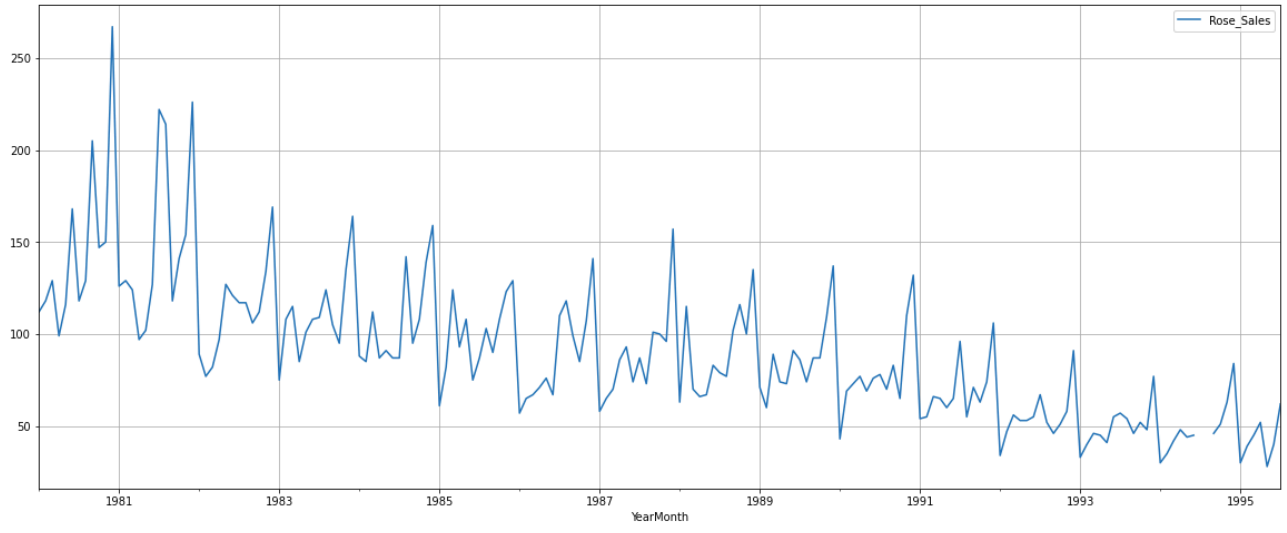
Part 2: Wine Sales analysis and forecast for dataset Rose.csv

2.1: Reading and Plotting Time Series Data

Question:
- Read the data as an appropriate Time Series data and
- Plot the data

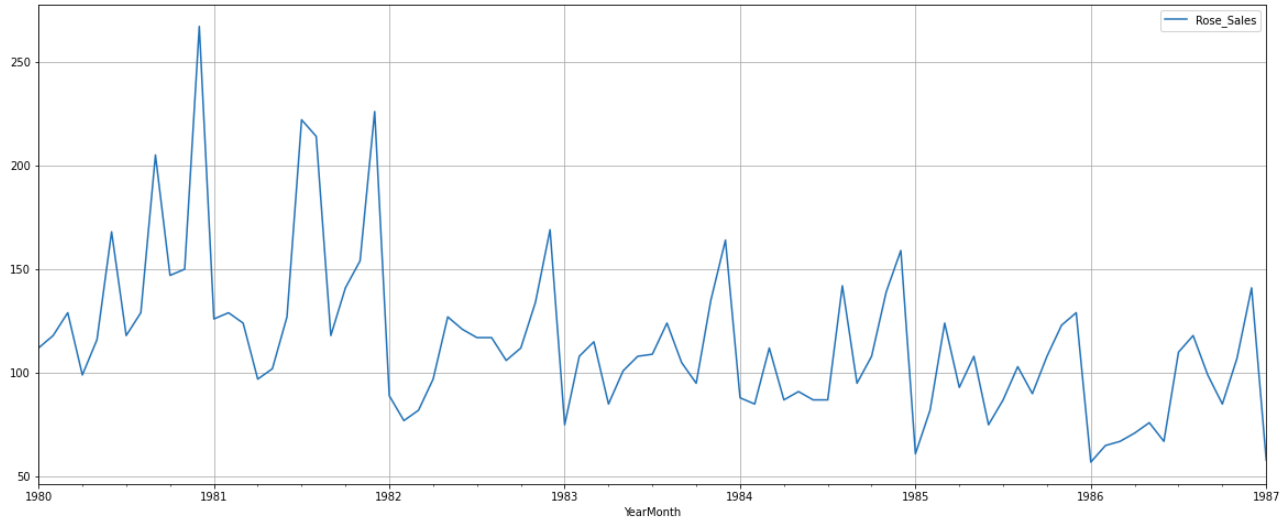
- The given dataset contains details of Sales of Rose over a period of 15 years and 7 months.
  - In all, there are 187 rows in the dataset that is a record of Monthly Sales of Rose, from January 1980 to July 1995.
  - The .csv file given is read into a form of Time Series Pandas Dataframe, for the purpose of further analysis.
  - That effectively converts the Timeline of Months into the Index, and the Monthly Sales of Rose into the Feature that we will analyse, in this exercise, as a Time Series.

Plotting the Time Series data:



From the above plot, one gets the following impression:
1. There is a discernible downward trend, indicating a steady decline in the sales and demand for Rose over the years.
2. There are missing data in the series - sales values for 2 months in 1994 are missing.
3. There is a seasonal behaviour at work, where sales peak in each year-end

The seasonal behaviour can be better discerned in the following plot, which depicts the first 7 years of the dataset, starting January 1980:



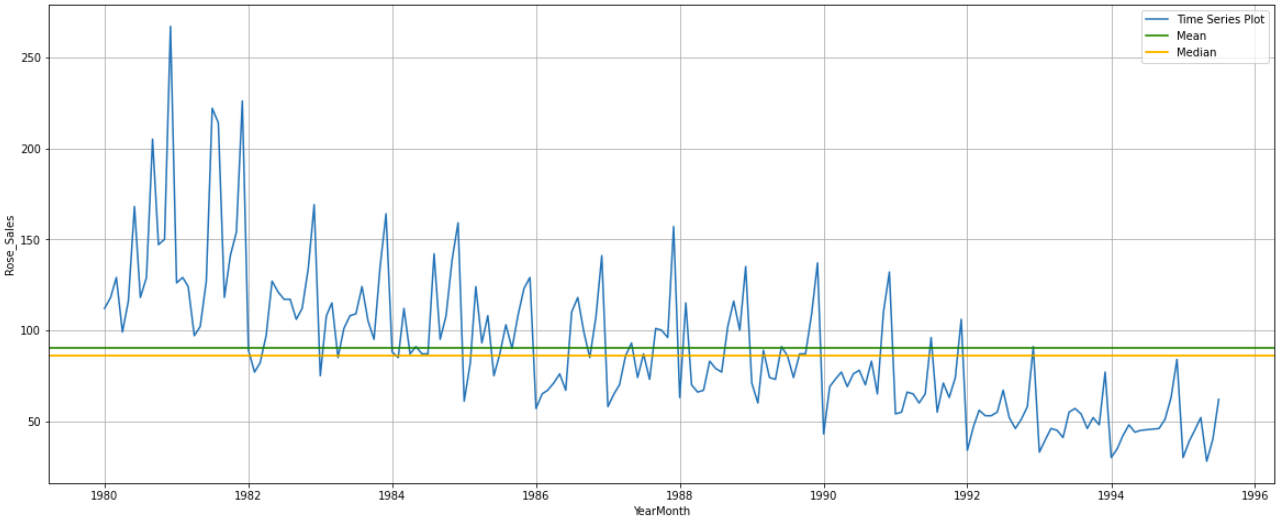
One notes that while a seasonal pattern is evident, the movement is uneven across years. This can be attributed to the Error or Unexplained component of the Time Series. The seasonal pattern does become steadier in the latter years.

2.2: Exploratory Data Analysis and Time Series Decomposition

Question:
- Perform appropriate Exploratory Data Analysis to understand the data, and also
- Perform Decomposition.

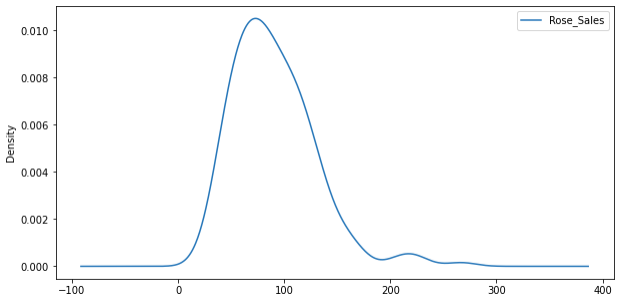
Overall Assessment:

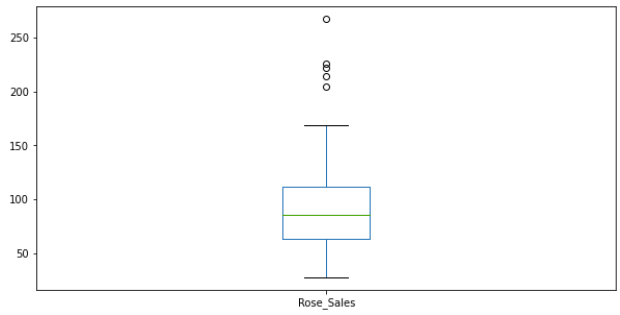
Monthly Sales of Rose over a 15 year and 7 month period:



Summary of Rose_Sales:

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

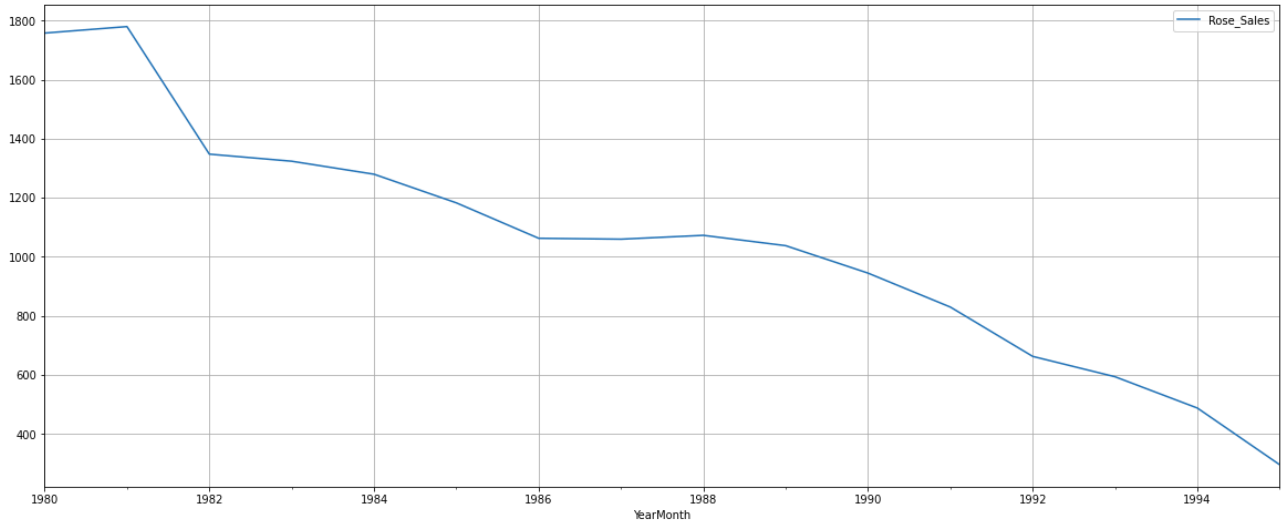




- The overall distribution is close to normal, as seen in the shape of the KDE, with a longer right tail, owing to some high sales figures in the initial years.
- The overall high range of monthly sales is largely due to the steady decline in Sales over the 15 + year period. Annual Range is a lot more limited.
- The most notable element of the data is the downward trend, indicating declining demand for Rose.

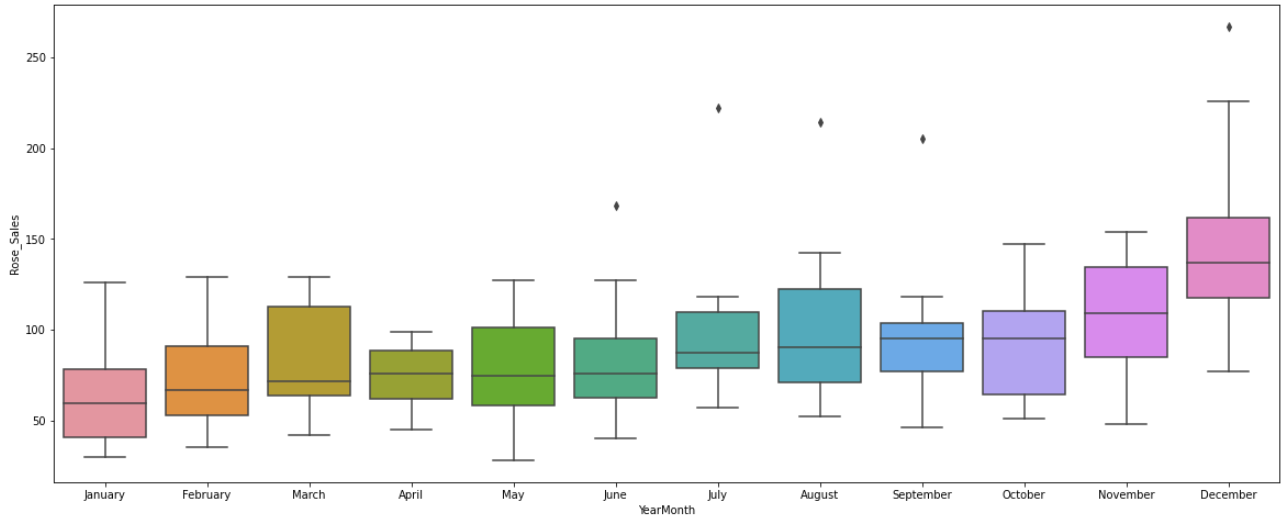
Studying the Overall Trend, by Resampling the data:

Annual Sales of Rose over the 15 year and 7 month period:



This plot makes the trend of declining Rose sales very evident. From 1991, the demand for Rose has consistently gone down, barring the period 1996 to 1998 where the Sales were steady briefly.

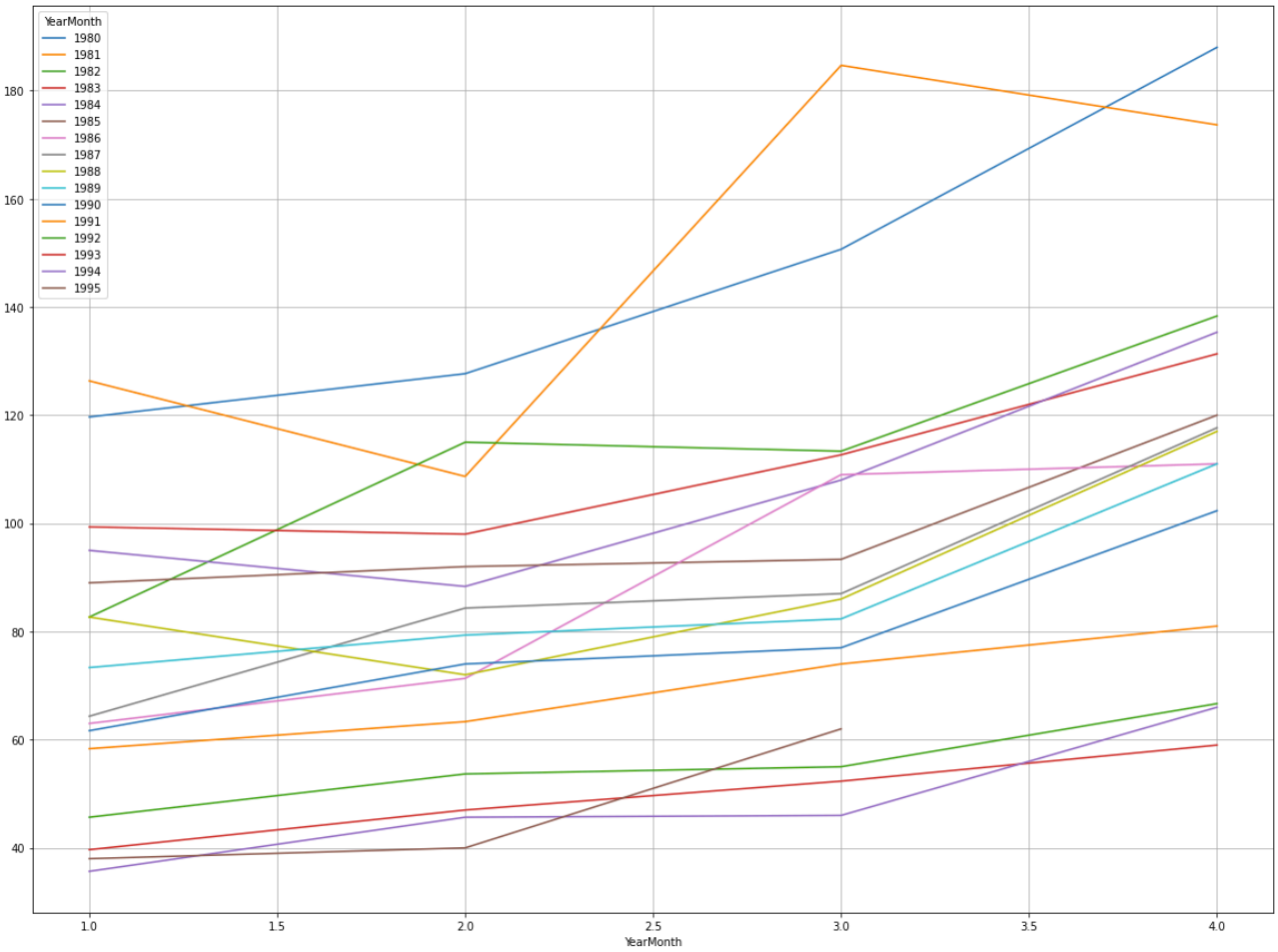
Seasonality of Wine Sales: using box plots to summarise Monthly Sales of Rose across the 15 year and 7 months time period:



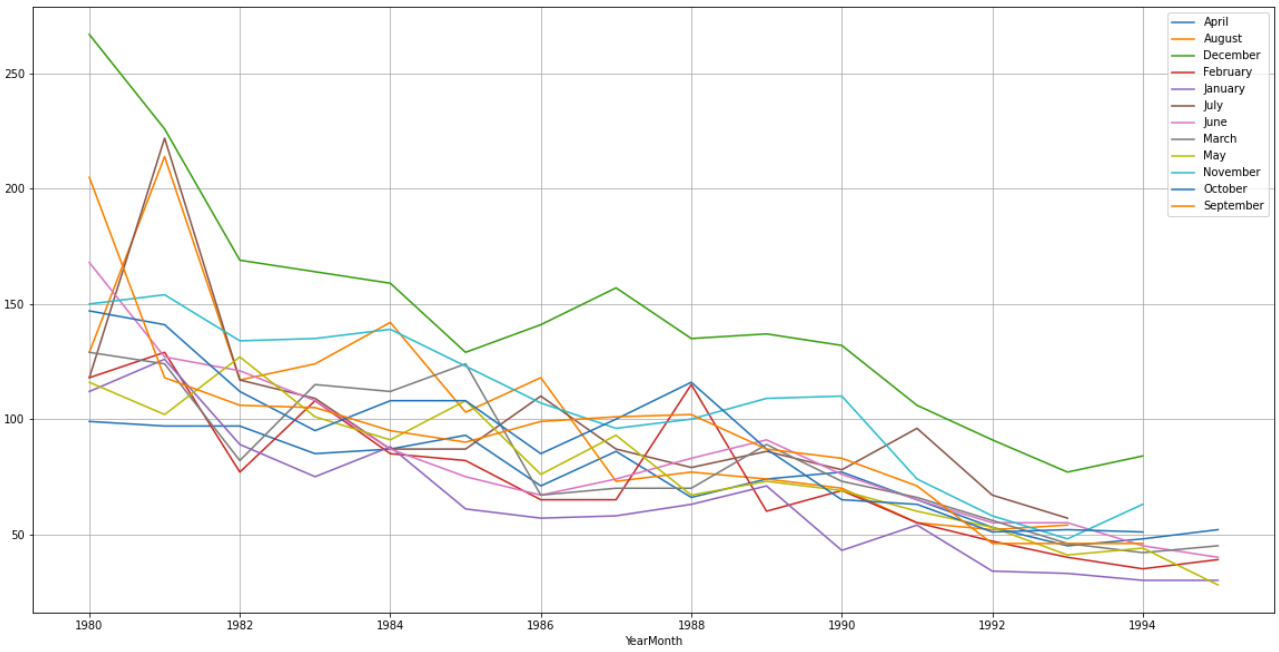
- As observed earlier, Sales of Rose are highest in the last quarter, peaking in December.
- The first half of the year sees a comparatively lower volume of sales, with March, May and June accounting for a slightly higher volumes.
- Sales pick up in the third quarter, August being a good month.

Seasonality of Wine Sales, viewed on a quarterly basis:

- Sales of each Quarter, for every year in the dataset - shows the similar pattern of lower sales but a steady rise through the first 3 quarters, with a spike in the last quarter.
- Only year 1981 exhibits a higher third quarter sale, which appears to be an anomaly.
- Interestingly, year 1995 shows a spike in demand, which is in contrast to the general pattern.



The following is a plot of Monthly Sales across years, which again brings out the declining annual Sales of Rose over the years. That said, December stands out as the most lucrative month, consistently over all years.



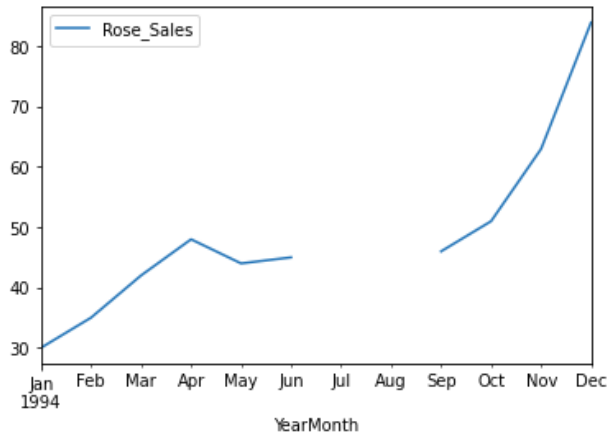
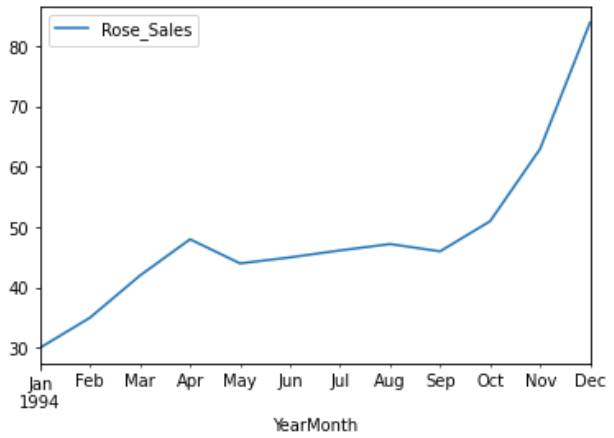
Missing Values: Imputation

The Sales Values for the months of July 1994 and August 1994 are missing in the dataset.

Inspecting the seasonal behaviour for the previous 2 years, there appears to be a marginal increase from June to July, and then a decrease in August and September.

It was also evident that the relative change in value across months was decreasing with each year - indicating that July 1994 and August 1994 would stay close to the June 1994 mark.

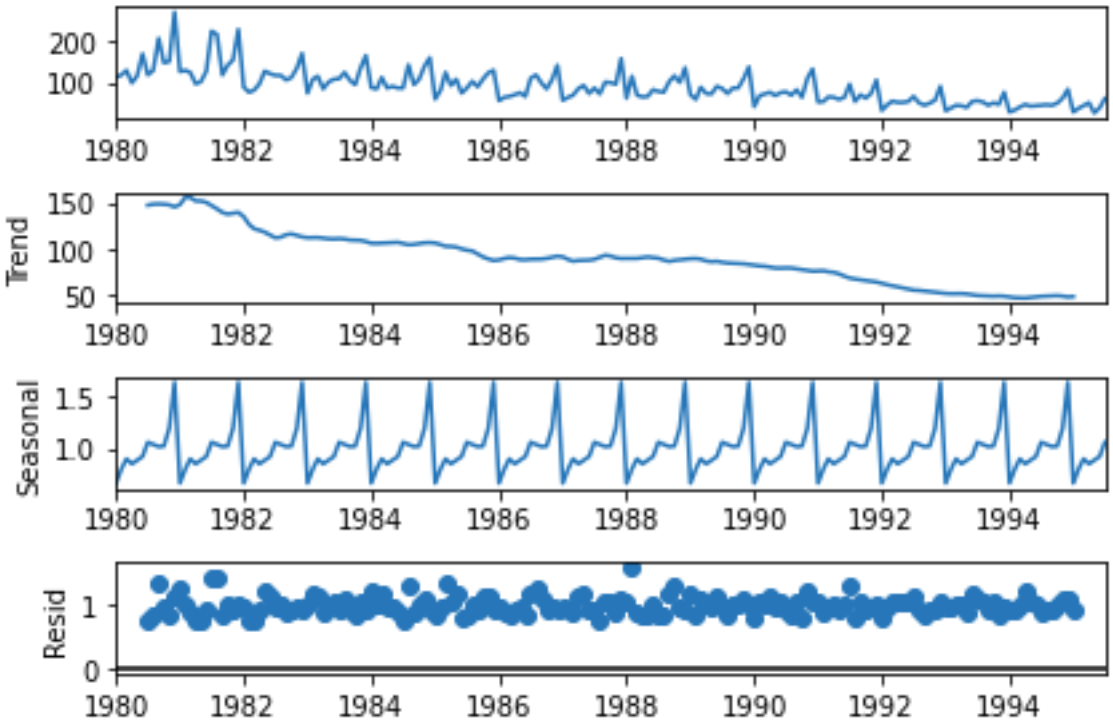
So we use the Spline, Order 1 method to interpolate the missing value of the 2 months.

Missing Values	Post Interpolation: using Spline, Order 1
	

Interpolated values:
1994-07-01 46.153199
1994-08-01 47.211982

Decomposition:

The pattern produced by the Time Series plot suggests a Multiplicative Seasonality.
The seasonal pattern's width (frequency) and height (amplitude) vary across time periods.
As we see a decline in sales, the seasonal effect also gets depressed.



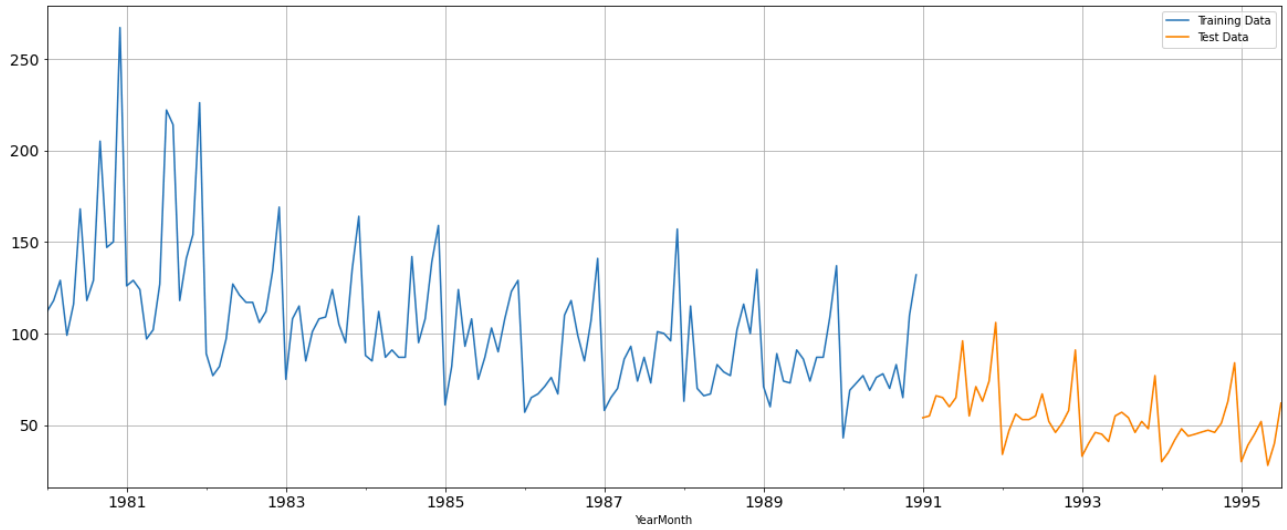
- Decomposition reiterates the predominance of the Trend factor.
- There is also a clear seasonality in the data.
- Errors are largely concentrated, and appear to be distributed evenly along the 1 mark.

2.3: Splitting data into Training and Test datasets

Question:
- Split the data into training and test. The test data should start in 1991.

After splitting the data:
- 132 observations, starting January 1980 up to December 1990, comprises the Training Data.
- 55 observations, starting January 1991 up to July 1995, comprises the Test Data.

The split is depicted in the following plot:

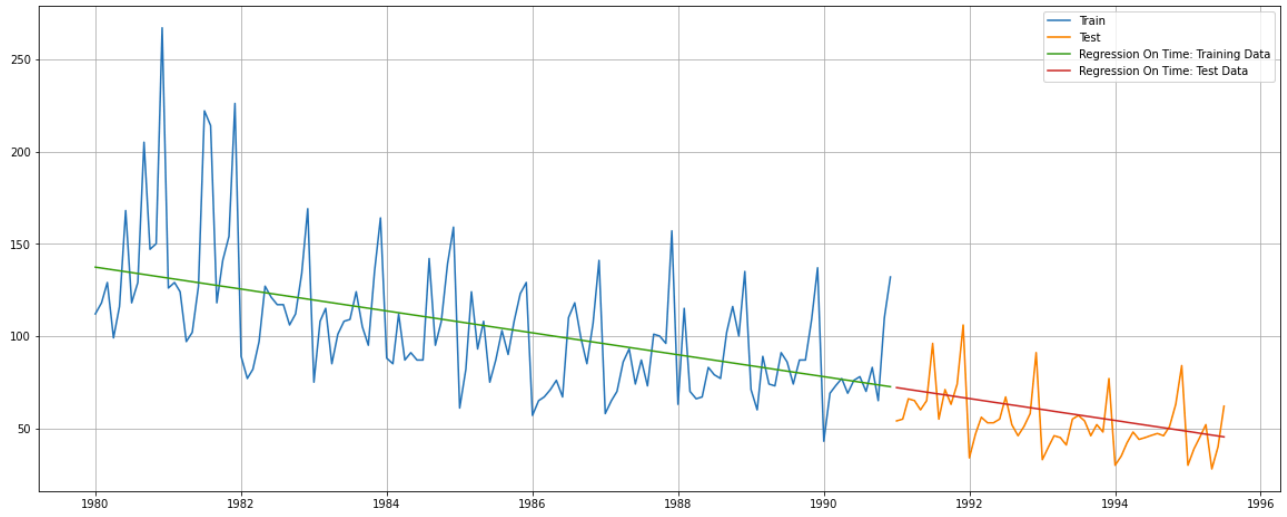


2.4: Model Building: Exponential Smoothing and other models

Question:
- Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.
- Other models such as regression, naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Model 1: Linear Regression

Model Output Visualised:



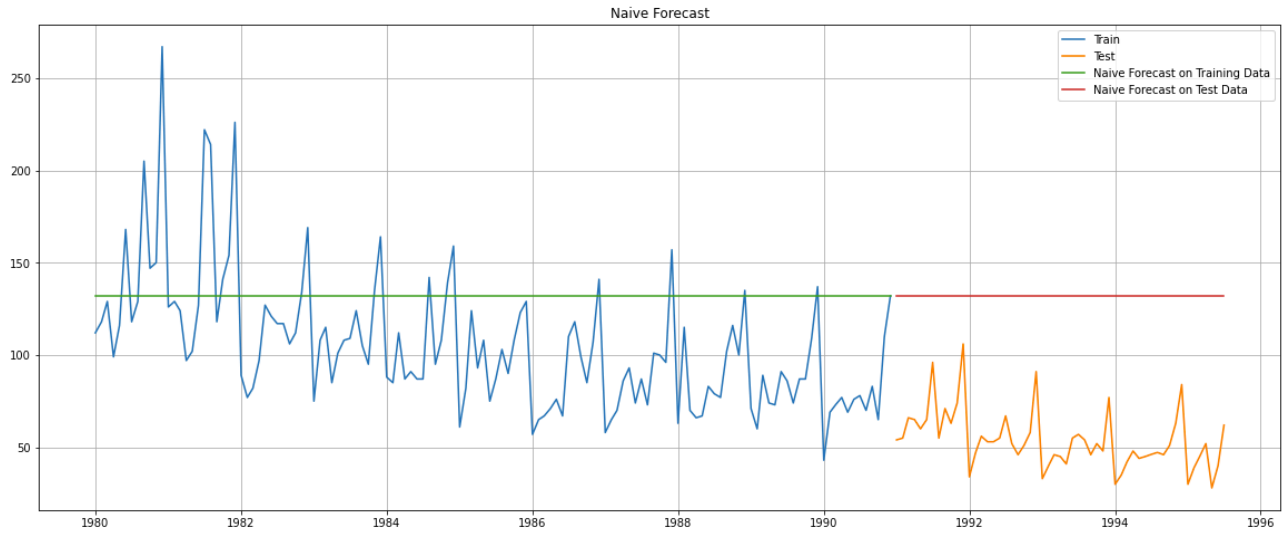
Performance Metrics:

	RMSE	MAPE
Training Data	30.718	21.22
Test Data	15.255	22.72

Observation:
A linear regression model captures the declining trend. It however is unable to capture the seasonal variation that is a characteristic feature of the time series.

Model 2: Naive

Model Output Visualised:



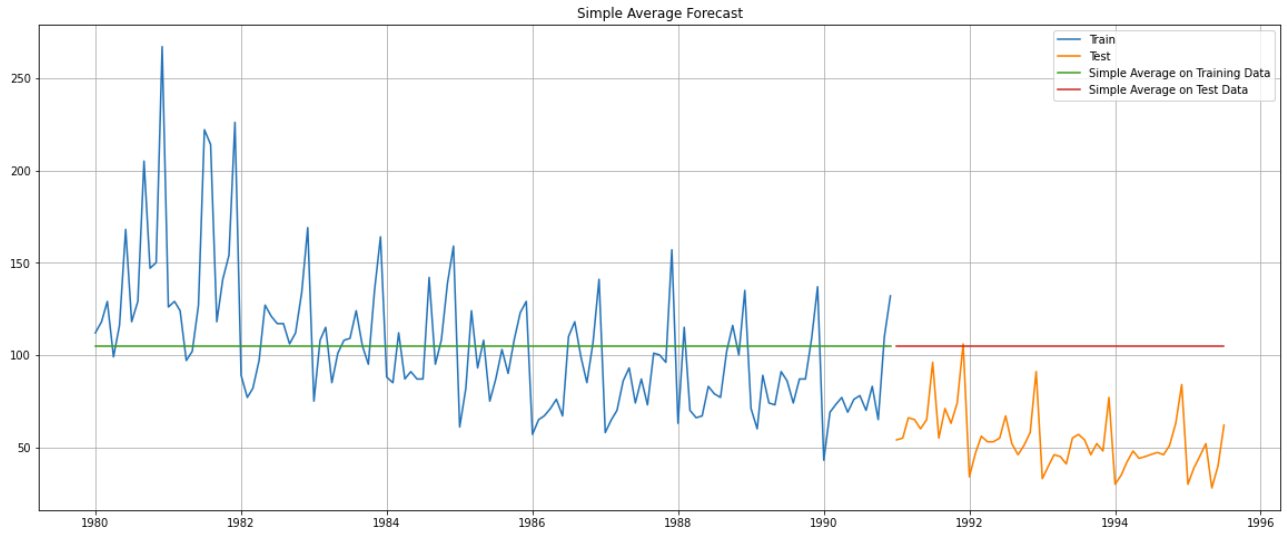
Performance Metrics:

	RMSE	MAPE
Training Data	45.064	36.38
Test Data	79.672	144.91

Observation:
The Naive model is dependent on the last observed value, which in our training data is the month of December. We know that the month of December records peak sales every year. So this value is clearly not representative of the dataset at large, also when the overall sales are declining. Hence the expected very high RMSE scores on the Test data.

Model 3: Simple Average

Model Output Visualised:



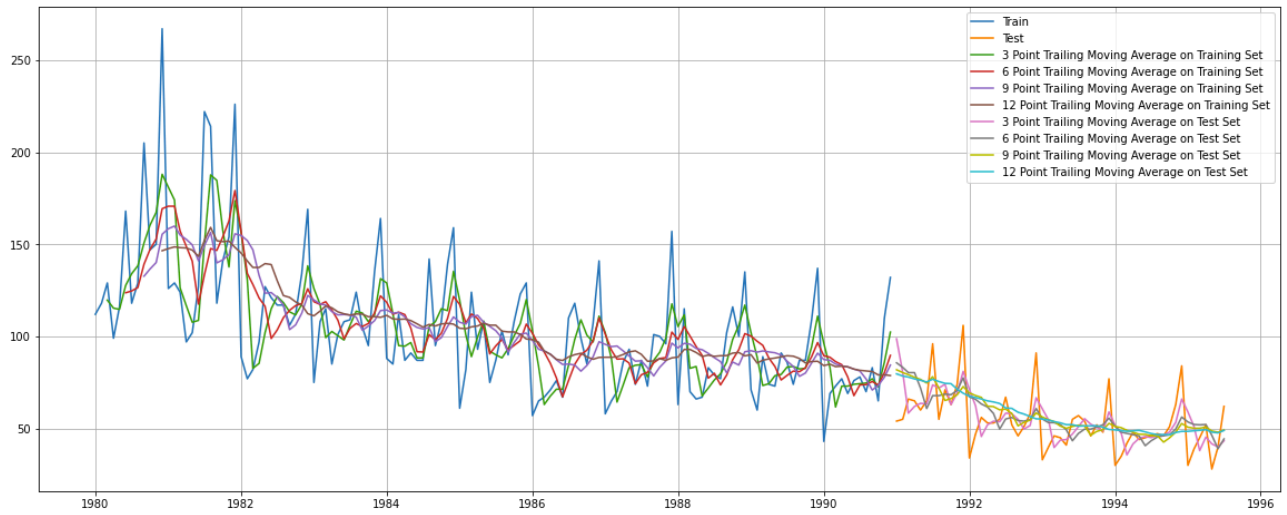
Performance Metrics:

	RMSE	MAPE
Training Data	36.034	25.39
Test Data	53.413	94.77

Observations:
A simple average model is not a great fit for the data, since it is unable to capture the trend, which in this case is one of decline. Also, seasonality is strong in this time series. It therefore misses out on much of the variation, resulting in high RMSE scores.

Model 4: Moving Average/s

Model Output Visualised:

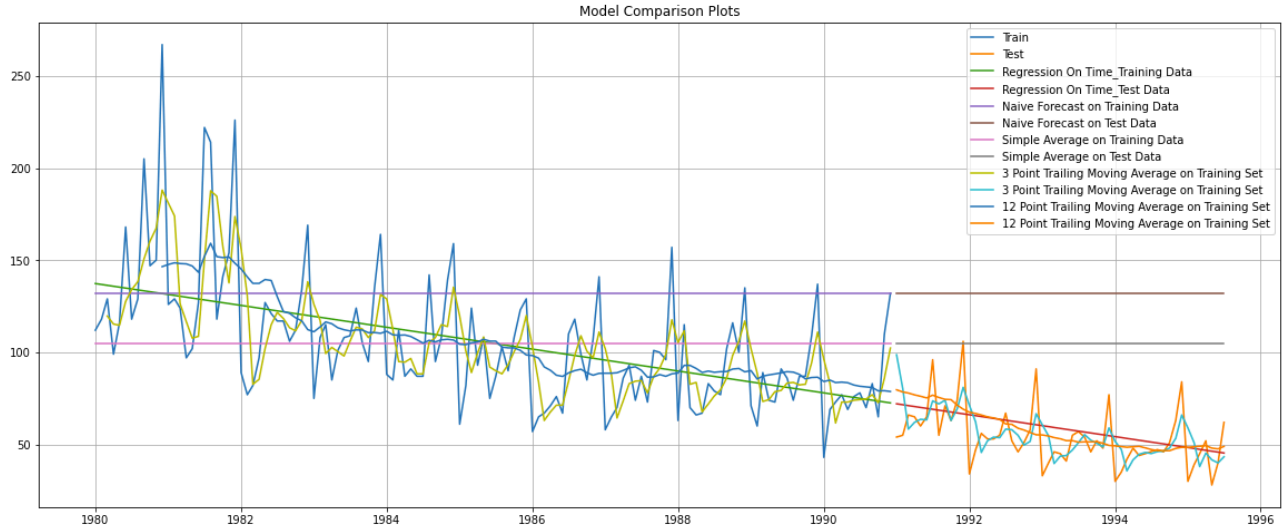


Performance Metrics on Test Data:

	RMSE	MAPE
3 point Moving Average	14.126	18.31
6 point Moving Average	14.555	20.82
9 point Moving Average	14.722	20.99
12 point Moving Average	15.233	22.03

Observation:
Moving Average Models are able to better track the variation of the time series. In this case it works especially as there is a clear trend. Plus the seasonal variation is also not very large. So one can see a relatively small difference in the performance of the 3-point and 12-point MA series.

Model Performance Comparison Visualised:

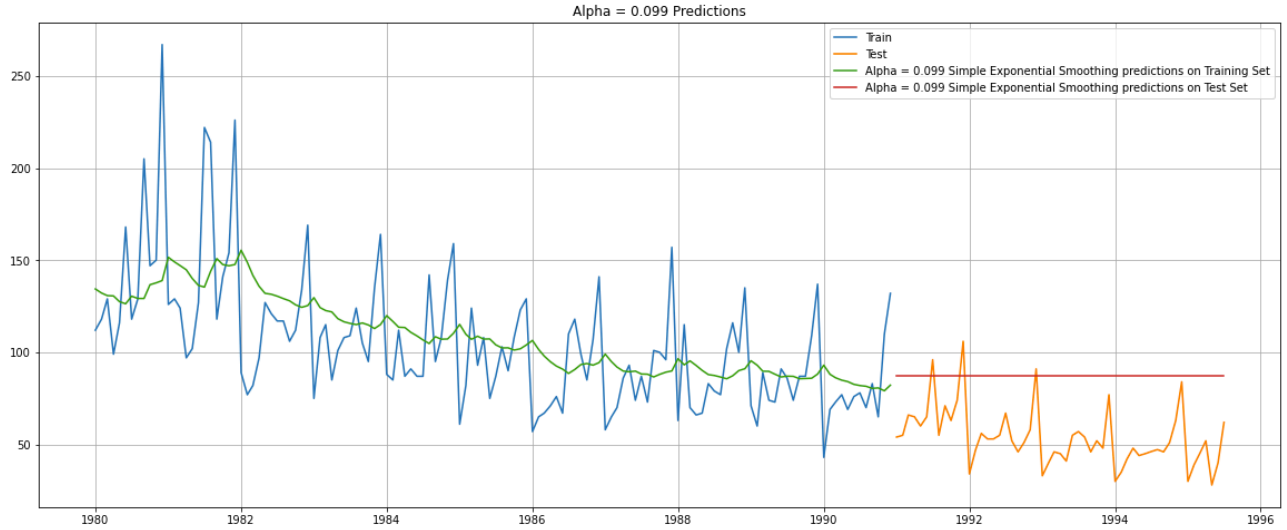


We will now proceed to the Exponential Models.

Model 5: Single Exponential Smoothing (Auto-fit, alpha = 0.099)

SES, Alpha = 0.099

Model Output Visualised:



Performance Metrics on Test Data:

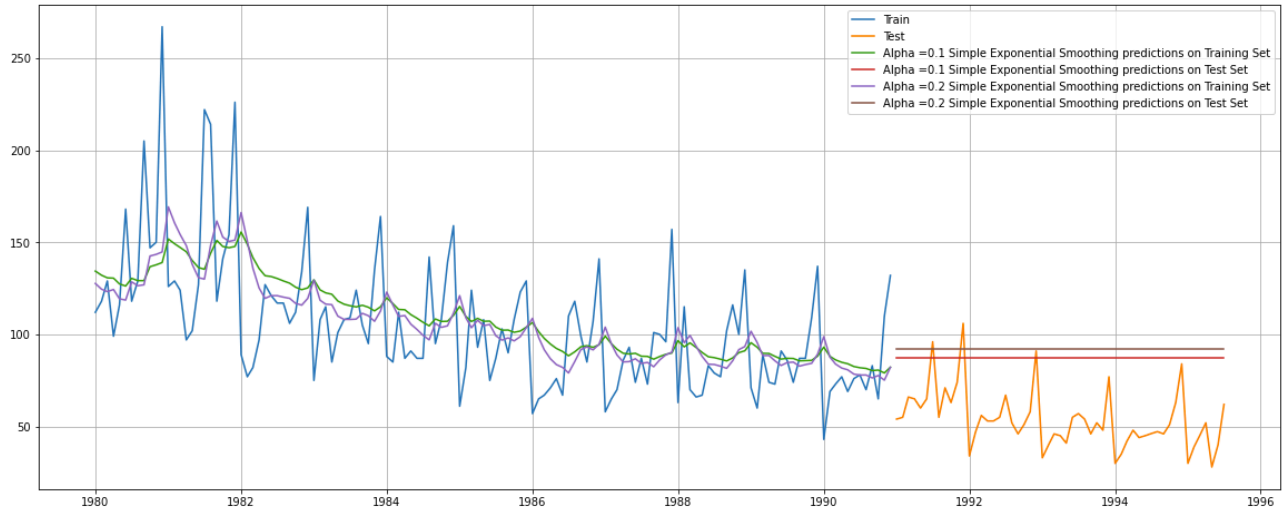
	RMSE	MAPE
Training Data	31.501	22.73
Test Data	36.748	63.75

Observation:
The level extrapolated by the SES model does not capture the trend (decline) or the seasonal variation of the time series.

Model 5a: Single Exponential Smoothing (using a Range of alpha values)

SES, Alpha ranging from 0.1 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and 0.2):



Performance Metrics on Test Data:

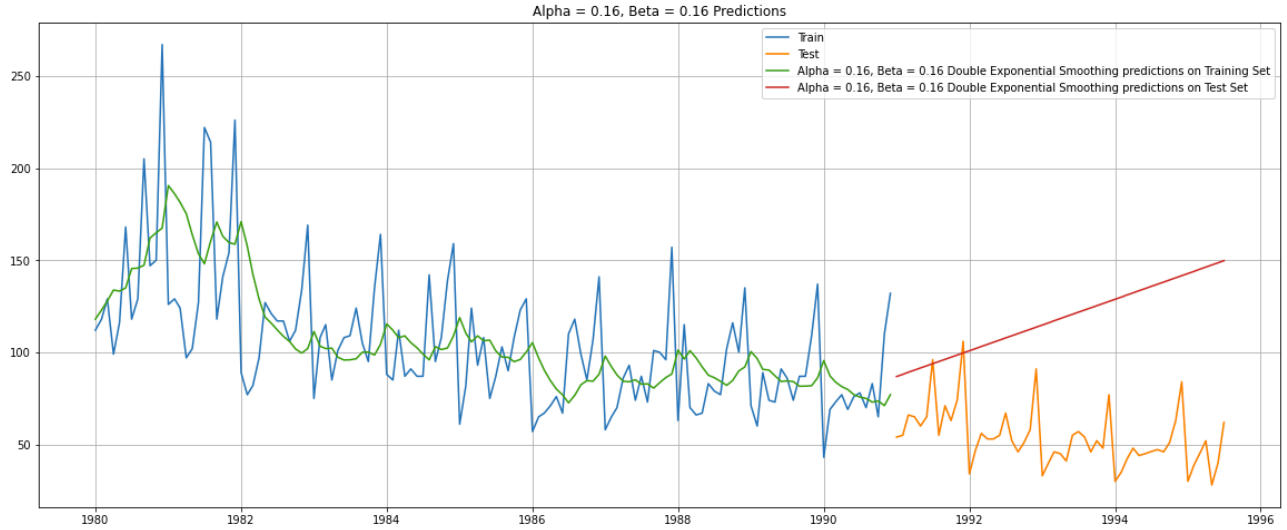
	RMSE	MAPE
Alpha = 0.1	36.780	63.81
Alpha = 0.2	41.314	72.07

Observations:
The level extrapolated by the SES model does not capture the trend (decline) or the seasonal variation of the time series. Hence SES will not be a good fit for the dataset.

Model 6: Double Exponential Smoothing (Auto-fit, alpha = 0.16, beta = 0.16)

DES, Alpha = 0.16, Beta = 0.16

Model Output Visualised:



Performance Metrics on Test Data:

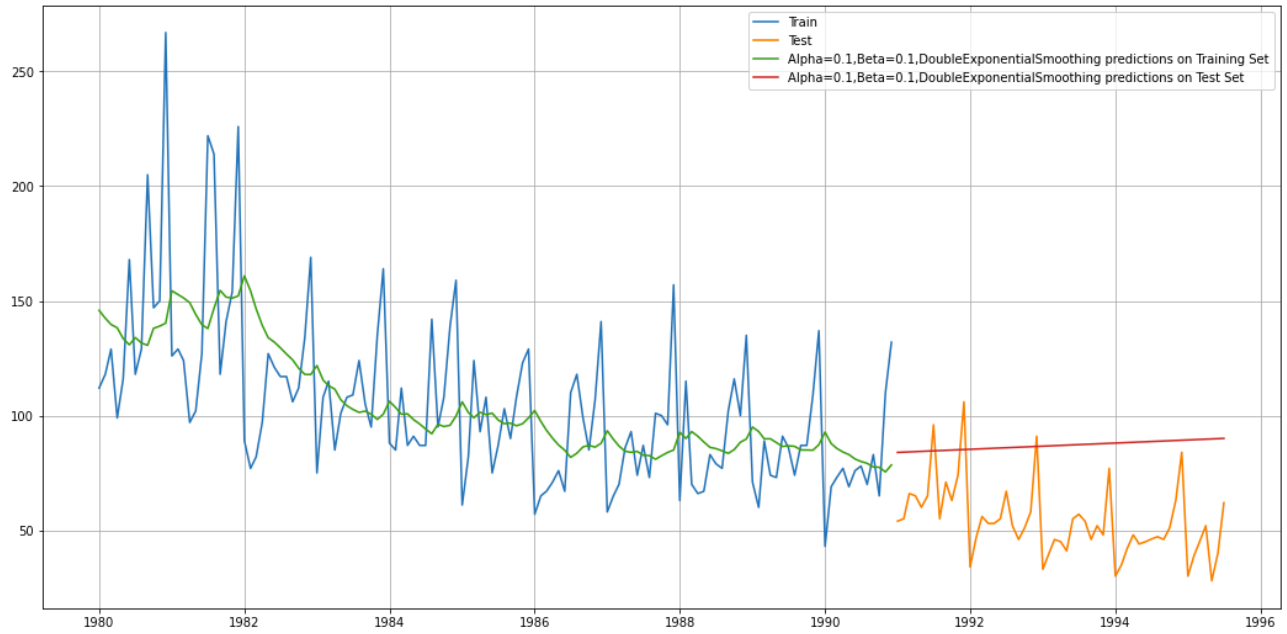
	RMSE	MAPE
Training Data	33.075	23.99
Test Data	70.517	120.07

Observation:
The Holt's model isn't a very good fit for this time series, as evidenced by the high RMSE. It is unable to capture the seasonality which is a strong component of the series. And in this case, it is unable to forecast the trend as well.

Model 6a: Double Exponential Smoothing (using a Range of alpha, beta values)

DES, Alpha and Beta ranging from 0.1 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and beta = 0.1)



Performance Metrics:

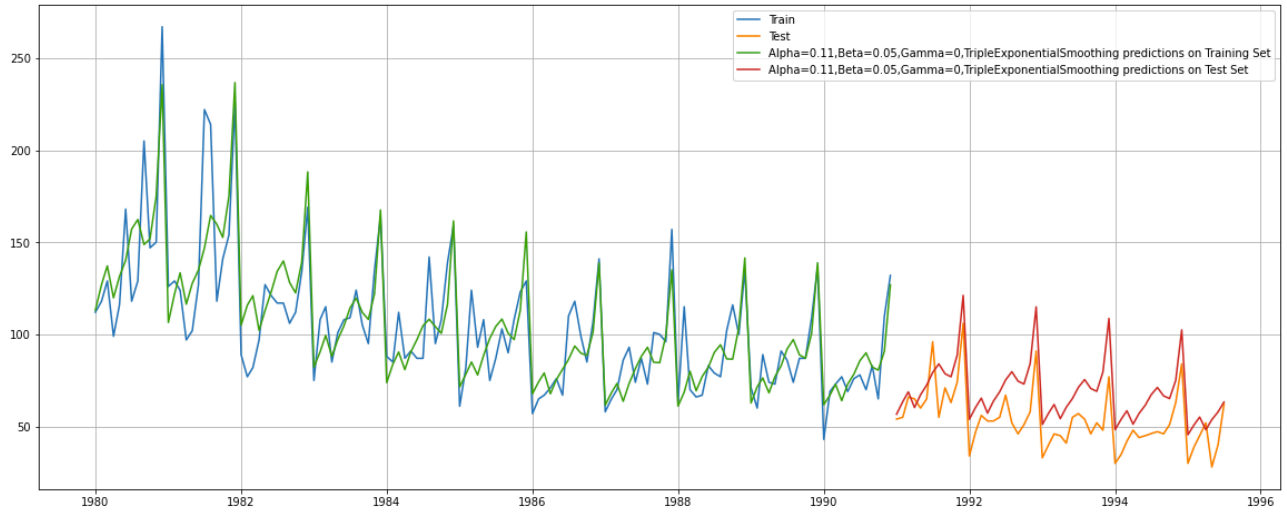
	RMSE	MAPE
Training Data	32.027	22.78
Test Data	37.008	63.89

Observation:
This model is an improvement over the earlier DES model. That said, it is unable to capture the seasonal variation which is a strong component of the series. And in this case, again, it is unable to capture the trend as well.

Model 7: Triple Exponential Smoothing (Auto-fit: Alpha=0.15, Beta=0, Gamma=0.37)

TES, Alpha=0.11, Beta=0.05, Gamma=0

Model Output:



Performance Metrics:

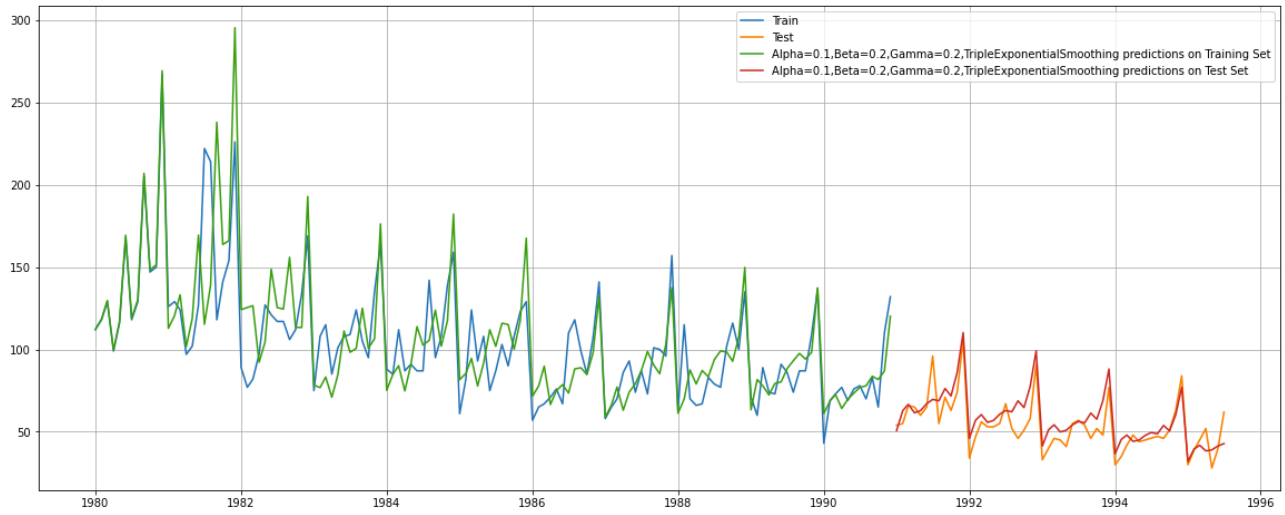
	RMSE	MAPE
Training Data	18.579	13.21
Test Data	17.311	28.78

Observation:
This Holt-Winters' model or the Triple Exponential Model is a better fit for this time series, as it has both a strong Trend and a strong Seasonal component. This is corroborated by a low RMSE.

Model 7a: Triple Exponential Smoothing (using a Range of alpha, beta, gamma values)

TES, Alpha, Beta, Gamma ranging from 0 to 1

Model Output with parameters with the lowest RMSE values: (alpha = 0.1 and beta = 0.2, gamma = 0.2)

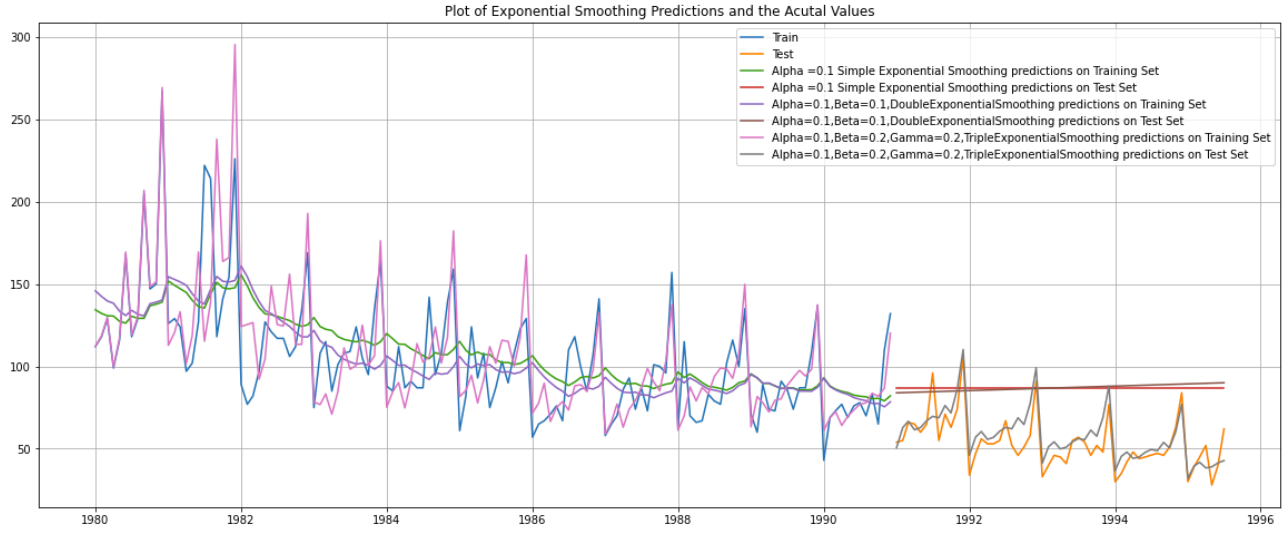


Performance Metrics:

	RMSE	MAPE
Training Data	24.366	15.36
Test Data	9.628	13.87

Observation:
This Triple Exponential Smoothing model is an improvement on the earlier auto-fit TES model. This, in fact, is the best performing model thus far, as evidenced by the low RMSE. This Holt-Winters' model or the TES Model is a good fit for this time series, as it has both a strong Trend and a strong Seasonal component.

A Consolidated Plot of all the Exponential Models built:



The Triple Exponential Models appear to be the strongest fit for this time series.

2.5: Stationarity Check

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

The **Augmented Dickey Fuller (ADF) Test** is a statistical test for affirming whether or not a time series is Stationary.

The Null Hypothesis H0 is: Time Series is Non-stationary
The Alternative Hypothesis H1 is: Time Series is Stationary

TEST 1: We administer the ADF test on the Original Time Series.

Results of Dickey-Fuller Test:
Test Statistic: -2.164250
p-value: 0.219476

With the resultant ADF test p-value at 0.21, we cannot reject the Null Hypothesis (at alpha 0.05).

We hence conclude that the Time Series is Non-stationary.

In order to make a Time Series Stationary, we need to transform the original series by taking a Difference of the original values. Usually, a 1 period difference suffices to transform a Non-Stationary series into a Stationary one.

TEST 2: We administer the ADF test on the new series - derived by taking a 1 period Difference of the original series.

Results of Dickey-Fuller Test:
Test Statistic: -6.592372e+00
p-value: 7.061944e-09

The resultant ADF p-value (0.000000007) is significantly less than 0.05 (alpha).
We can hence reject the null hypothesis in the case of the new series, which is derived by differencing the original series over 1 period.

We conclude that at Difference 1, the time series is Stationary.

2.6: Model Building: Automated ARIMA / SARIMA

Question:
- 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.

Model 8: ARIMA (Lowest AIC model parameters: p=3, d=1, q=3)

Of the ARIMA models generated using various combinations of parameters p and q, the model with the lowest AIC score was: ARIMA (3, 1, 3) with an AIC score of 1273.194115

Model Summary:

ARIMA Model Results

Dep. Variable:	D.Rose_Sales	No. Observations:	131
Model:	ARIMA(3, 1, 3)	Log Likelihood	-628.597
Method:	css-mle	S.D. of innovations	28.356
Date:	Sat, 15 Aug 2020	AIC	1273.194
Time:	14:26:20	BIC	1296.196
Sample:	02-01-1980	HQIC	1282.541
	- 12-01-1990		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4906	0.088	-5.547	0.000	-0.664	-0.317
ar.L1.D.Rose_Sales	-0.7242	0.086	-8.398	0.000	-0.893	-0.555
ar.L2.D.Rose_Sales	-0.7216	0.087	-8.328	0.000	-0.891	-0.552
ar.L3.D.Rose_Sales	0.2765	0.086	3.231	0.001	0.109	0.444
ma.L1.D.Rose_Sales	-0.0152	0.045	-0.342	0.733	-0.103	0.072
ma.L2.D.Rose_Sales	0.0152	0.044	0.343	0.732	-0.072	0.102
ma.L3.D.Rose_Sales	-1.0000	0.046	-21.863	0.000	-1.090	-0.910

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.5011	-0.8661j	1.0006	-0.3335
AR.2	-0.5011	+0.8661j	1.0006	0.3335
AR.3	3.6123	-0.0000j	3.6123	-0.0000
MA.1	1.0000	-0.0000j	1.0000	-0.0000
MA.2	-0.4924	-0.8704j	1.0000	-0.3319
MA.3	-0.4924	+0.8704j	1.0000	0.3319

Performance Metrics on Test Data:

	RMSE	MAPE
ARIMA (3,1,3)	15.984	26.04

Observation: ARIMA does not factor the seasonal component which is an important characteristic of this time series, and hence wouldn't be an ideal model for the series.

Model 9: SARIMA (Lowest AIC parameters: p=3, d=1, q=1, P=3, D=1, Q=1)

Of the SARIMA models generated using various combinations of parameters p, q, P, Q and D, the model with the lowest AIC score was: SARIMA (3, 1, 1) x (3, 1, 1, 12) with an AIC score of 681.362807

Model Summary:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(3, 1, 1)x(3, 1, 1, 12)	Log Likelihood	-331.681
Date:	Thu, 13 Aug 2020	AIC	681.363
Time:	00:05:15	BIC	702.801
Sample:	0	HQIC	689.958
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0173	0.151	0.114	0.909	-0.279	0.314
ar.L2	-0.0426	0.141	-0.302	0.763	-0.319	0.234
ar.L3	-0.0574	0.119	-0.484	0.628	-0.290	0.175
ma.L1	-0.9388	0.085	-11.104	0.000	-1.105	-0.773
ar.S.L12	0.0908	0.126	0.721	0.471	-0.156	0.337
ar.S.L24	-0.0437	0.108	-0.406	0.685	-0.254	0.167
ar.S.L36	-3.592e-05	0.053	-0.001	0.999	-0.104	0.103
ma.S.L12	-0.9998	237.117	-0.004	0.997	-465.741	463.741
sigma2	185.4128	4.4e+04	0.004	0.997	-8.6e+04	8.63e+04

Ljung-Box (Q):	42.97	Jarque-Bera (JB):	2.56
Prob(Q):	0.35	Prob(JB):	0.28
Heteroskedasticity (H):	0.56	Skew:	0.42
Prob(H) (two-sided):	0.13	Kurtosis:	3.22

Performance Metrics on Test Data:

	RMSE	MAPE
SARIMA (3,1,1) (3,1,1,12)	16.780	25.38

Observation:
A SARIMA model is generally better equipped for a time series with a strong seasonal component. However this model fails to improve on the ARIMA model score.

2.7: Model Building: ARIMA / SARIMA using ACF, PACF cut-offs

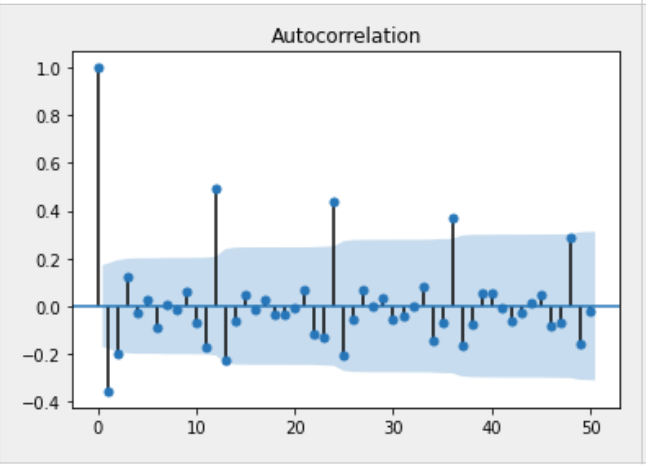
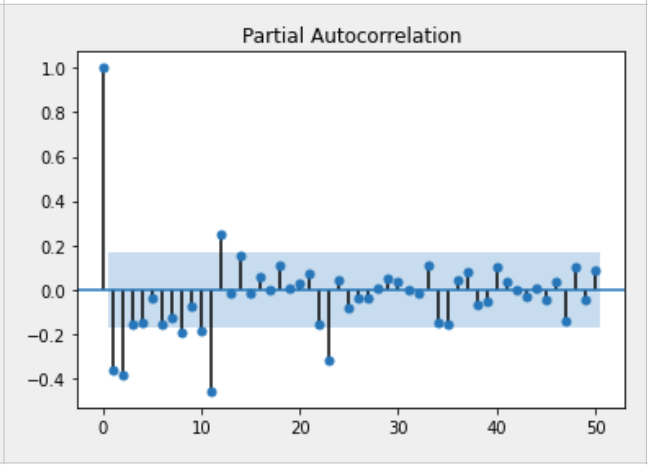
Question:
- Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data, and
- Evaluate this model on the test data using RMSE.

From the ADF test conducted in an earlier section, we know that the original time series is Non-stationary.

So the first step would be to transform the time series by differencing the original series over 1 period, and make it Stationary.

We then need to plot the ACF and PACF on the transformed stationary Time Series.

The following are the ACF and PACF plots:

ACF Plot	PACF Plot
	

The ACF plot:
- The cut-off is right after lag 2. It appears lag 2 is marginally above the border. Hence a possible value of q is 2.
- Hence from the ACF plot, we can assign 2 as the order of the MA component (q) of the ARIMA/SARIMA model.

The PACF plot:
- The cut-off here too appears right after lag 2
- Lag 3 appears to be insignificant, though only marginally.
- Hence from the PACF plot, we can assign 2 as the order of the AR component (p) of the ARIMA/SARIMA model.

We know that the Seasonal component is apparent in the series. The ACF plot also clearly shows a pattern repeating every year - indicating strong seasonal behaviour.

So a SARIMA model would be appropriate for modelling such a series. For employing the seasonal component we will accord value to a 12 period lag, which will map to a value of 1 to both P and Q.

The value of d will be 1, as the series has been Differenced by 1 period, in order to make it stationary.

Hence based on the ACF and PACF plots, we can develop a SARIMA (2,1,2)x(1,1,1, 12) model.

Model 9a: SARIMA (ACF, PACF plot parameters: p=2, d=1, q=2, P=1, D=1, Q=1)

Model Summary:

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(2, 1, 2)x(1, 1, [1], 12)	Log Likelihood	-450.847
Date:	Fri, 14 Aug 2020	AIC	915.693
Time:	23:26:33	BIC	934.204
Sample:	0	HQIC	923.192
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.1035	0.133	8.288	0.000	0.843	1.364
ar.L2	-0.3436	0.109	-3.150	0.002	-0.557	-0.130
ma.L1	-1.8152	0.105	-17.291	0.000	-2.021	-1.609
ma.L2	0.8668	0.093	9.276	0.000	0.684	1.050
ar.S.L12	-0.3877	0.069	-5.625	0.000	-0.523	-0.253
ma.S.L12	-0.0788	0.130	-0.606	0.545	-0.334	0.176
sigma2	338.2613	53.797	6.288	0.000	232.821	443.701

Ljung-Box (Q):	25.41	Jarque-Bera (JB):	0.03
Prob(Q):	0.96	Prob(JB):	0.98
Heteroskedasticity (H):	0.66	Skew:	0.04
Prob(H) (two-sided):	0.22	Kurtosis:	2.97

Performance Metrics on Test Data:

	RMSE	MAPE
SARIMA (2,1,2) (1,1,1,12)	13.275	17.53

Observation:
This SARIMA model appears to be an improvement over the last SARIMA model, going by the RMSE on the test data. The model is the second best model that we've developed.

2.8: Model performance comparison

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

All models, ranked according to their performance. The lowest Test RMSE score is the highest ranked or the best performing model.

		Test RMSE	Test MAPE
1	Triple Exponential Smoothing, Alpha=0.1,Beta=0.2,Gamma=0.2	9.628012	13.87
2	SARIMA (2,1,2) x (1,1,1,12)	13.275387	17.53
3	3 point Trailing Moving Average	14.12575	18.31
4	6 point Trailing Moving Average	14.554986	20.82
5	9 point Trailing Moving Average	14.72152	20.99
6	12 point Trailing Moving Average	15.232893	22.03
7	Regression On Time	15.255492	22.72
8	ARIMA (3,1,3)	15.983827	26.04
9	SARIMA (3,1,1) x (3,1,1,12)	16.780064	25.38
10	Triple Exponential Smoothing, Alpha=0.15,Beta=0,Gamma=0.37	17.310841	28.78
11	Single Exponential Smoothing, Alpha=0.099	36.748407	63.75
12	Single Exponential Smoothing, Alpha=0.1	36.780213	63.81
13	Double Exponential Smoothing, Alpha=0.1,Beta=0.1	37.007705	63.89
14	Simple Average Model	53.413298	94.77
15	Double Exponential Smoothing, Alpha=0.16,Beta=0.16	70.517385	120.07
16	Naive Model	79.672475	144.91

2.9: Optimum Model and Forecasting

Question:
- 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.

From our comparison between models, the **top 2 models** are:
1. Triple Exponential Smoothing, Alpha=0.1,Beta=0.2,Gamma=0.2
2. SARIMA (2,1,2) x (1,1,1,12)

We will re-build these 2 models on the complete data, and make 12 month forecasts.

Part A: TES model on complete data

Building a Triple Exponential Smoothing (alpha = 0.1, beta = 0.2, gamma = 0.2) model on the complete data, and forecasting for the next 12 months

Model Summary:

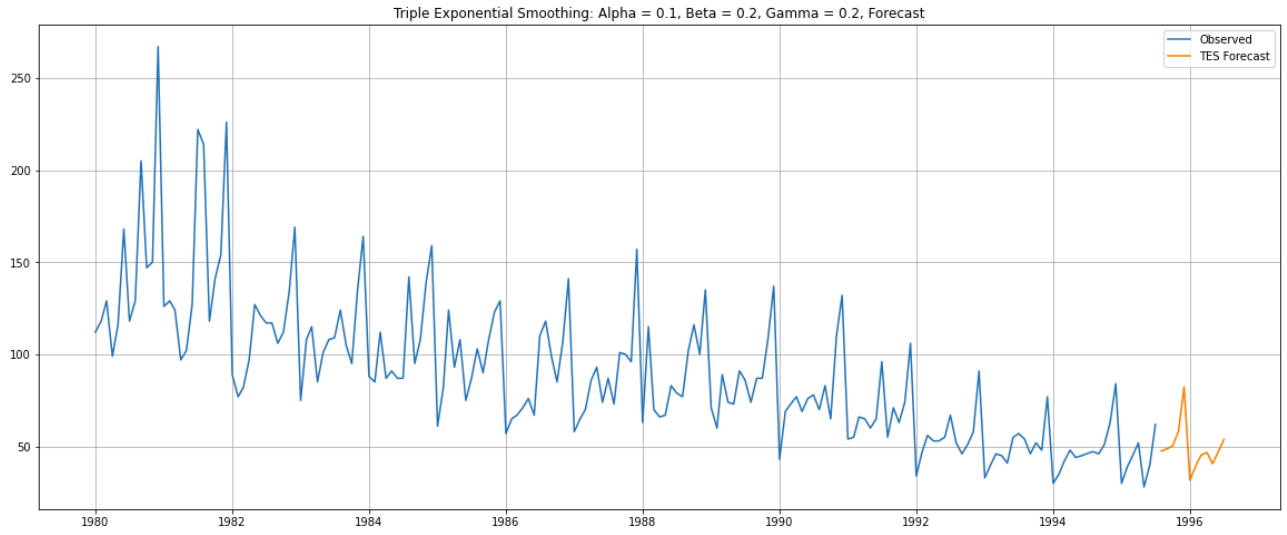
ExponentialSmoothing Model Results

Dep. Variable:	endog	No. Observations:	187
Model:	ExponentialSmoothing	SSE	81513.448
Optimized:	True	AIC	1168.477
Trend:	Additive	BIC	1220.174
Seasonal:	Multiplicative	AICC	1172.548
Seasonal Periods:	12	Date:	Fri, 14 Aug 2020
Box-Cox:	False	Time:	23:51:34
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.1000000	alpha	False
smoothing_slope	0.2000000	beta	False
smoothing_seasonal	0.2000000	gamma	False
initial_level	64.000000	l.0	True
initial_slope	0.1527778	b.0	True
initial_seasons.0	1.7500000	s.0	True
initial_seasons.1	1.8437500	s.1	True
initial_seasons.2	2.0156250	s.2	True
initial_seasons.3	1.5468750	s.3	True
initial_seasons.4	1.8125000	s.4	True
initial_seasons.5	2.6250000	s.5	True
initial_seasons.6	1.8437500	s.6	True
initial_seasons.7	2.0156250	s.7	True
initial_seasons.8	3.2031250	s.8	True
initial_seasons.9	2.2968750	s.9	True
initial_seasons.10	2.3437500	s.10	True
initial_seasons.11	4.1718750	s.11	True

TES (alpha = 0.1, beta = 0.2, gamma = 0.2) Model 12-month Forecast:

Timeline	Rose Sales Forecast
1995-08-01	47.552982
1995-09-01	48.746129
1995-10-01	50.277107
1995-11-01	58.269327
1995-12-01	82.302948
1996-01-01	31.700169
1996-02-01	39.432145
1996-03-01	45.380415
1996-04-01	46.71215
1996-05-01	40.645794
1996-06-01	47.20148
1996-07-01	53.867965



Part B: SARIMA model on complete data

Building a SARIMAX (2,1,2)x(1,1,1,12) model on the complete data, and forecasting for the next 12 months

Model Summary:

SARIMAX Results

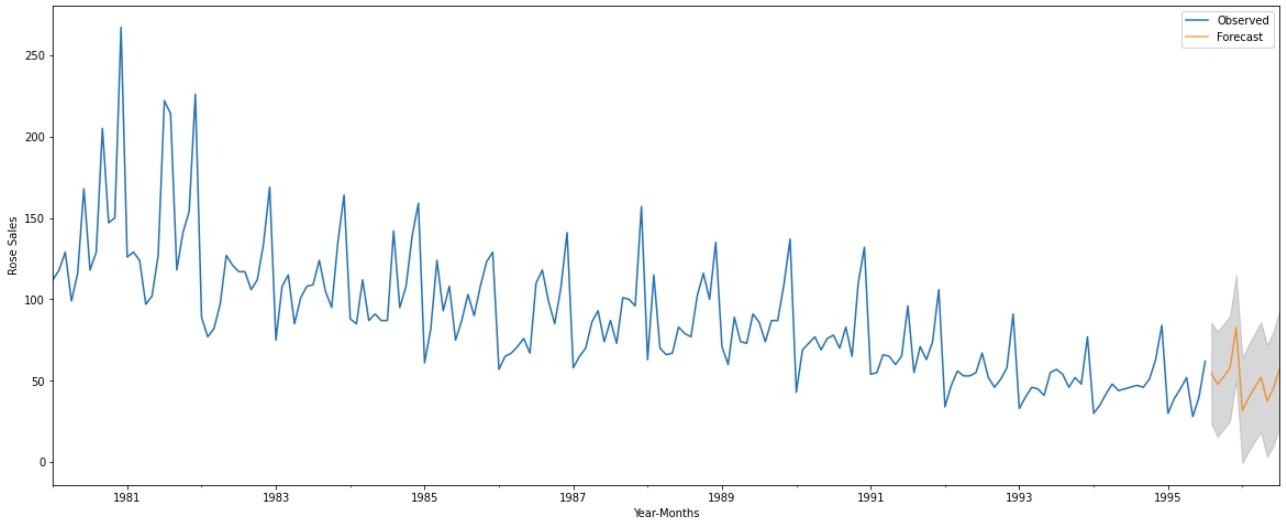
Dep. Variable:	y	No. Observations:	187
Model:	SARIMAX(2, 1, 2)x(1, 1, [1], 12)	Log Likelihood	-665.294
Date:	Fri, 14 Aug 2020	AIC	1344.588
Time:	23:45:34	BIC	1366.070
Sample:	0	HQIC	1353.312
	- 187		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.1081	0.093	11.951	0.000	0.926	1.290
ar.L2	-0.3257	0.080	-4.089	0.000	-0.482	-0.170
ma.L1	-1.8286	0.067	-27.351	0.000	-1.960	-1.698
ma.L2	0.8792	0.059	14.789	0.000	0.763	0.996
ar.S.L12	-0.3830	0.049	-7.757	0.000	-0.480	-0.286
ma.S.L12	-0.0828	0.093	-0.888	0.374	-0.265	0.100
sigma2	250.8891	26.925	9.318	0.000	198.118	303.660

Ljung-Box (Q):	35.53	Jarque-Bera (JB):	3.01
Prob(Q):	0.67	Prob(JB):	0.22
Heteroskedasticity (H):	0.22	Skew:	0.04
Prob(H) (two-sided):	0.00	Kurtosis:	3.67

SARIMA (2,1,2)x(1,1,1,12) model 12-month Forecast:

timeline	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	54.545303	15.839478	23.500497	85.59011
1995-09-01	47.917716	16.446711	15.682754	80.152677
1995-10-01	52.623395	16.455915	20.370395	84.876396
1995-11-01	57.646815	16.455942	25.393762	89.899869
1995-12-01	82.776318	16.466521	50.50253	115.050107
1996-01-01	31.838413	16.531662	-0.563049	64.239874
1996-02-01	39.28762	16.682039	6.591425	71.983816
1996-03-01	45.878564	16.915626	12.724547	79.032581
1996-04-01	52.236236	17.21173	18.501865	85.970606
1996-05-01	37.623006	17.546784	3.231942	72.01407
1996-06-01	44.997111	17.902037	9.909764	80.084459
1996-07-01	57.374141	18.265042	21.575316	93.172966



2.10: Insights and Findings

Question:
- Comment on the model thus built and report your findings, and
- Suggest the measures that the company should be taking for future sales.

- Looking at the pattern in the time series, one could infer that Rose wines have been going out of fashion for a while now. The time series shows a consistent decline in Sales of Rose.
- Perhaps there are other alternatives that are preferred in current times, in which case it may be worthwhile to focus more on the other festive wines in demand.
- But the last couple of years also shows a steadying of the decline. And in year 1995, the Sales in the first 2 quarters were surprisingly strong. This could mean that the interest in Rose might still be rekindled.
- The models appear to forecast a range and pattern for the next 12 months, which are similar to that of the past year. And given the low RMSE score, along with a certain consistency of past behaviour, the forecast looks dependable.
- The consumption pattern points to the fact that Rose Wines are most in demand in holiday and festive seasons. This brings them in competition with Sparkling Wines and other premium products.
- Perhaps Rose needs to be positioned differently, likely a level below premium. And then it can be promoted accordingly. The focus for Rose need not be the holiday season, because that is a very competitive space. A niche needs to be carved for Rose for sales and interest to pick up again.

Appendix

The related workings and code are appended in the following Jupyter notebooks:
TS_Project_Part_1_Sparkling_SB.ipynb
TS_Project_Part_2_Rose_SB.ipynb