Time Series Forecasting

SUBMISSION

nupur sarkar
GREAT LEARNING -PGPDSBA.O.DEC23.A | 18/08/2024

Contents

Figure 14(Sparkling Wine Sales (Processed) 13
Figure 16(Linear Regression for Rose wine) 15
Figure 17(Linear Regression Model for sparkling wine)....15
Figure 18(SA for Rose Wine) 16
Figure 19(SA for Sparkling wine)....16
Figure 20(Moving Average for Rose Wine) 17
Figure 21(Moving Average for Sparkling Wine) 17
Figure 22(Exponential Models for Rose wine) 18
Figure 23(Exponential Models for Sparkling Wine) 18
Figure 24(ACF and PACF Plots for Rose Wine) 20
Figure 25(ACF and PACF Plot for Sparkling Wine)....20

Figure 26(Auto ARIMA for Rose Wine)....21

Figure 27(Manual ARIMA for Rose Wine)....22

Figure 28(AUTO ARIMA for sparkling Wine)....22

Figure 29(Manual ARIMA for Sparkling wine)....23

Figure 30(Auto SARIMA for Rose Wine)....24

Figure 31(AUTO SARIMA for sparkling wine)....24

1. Define the problem and perform Exploratory Data Analysis

Context

As an analyst at ABC Estate Wines, we are presented with historical data encompassing the sales of different types of wines throughout the 20th century. These datasets originate from the same company but represent sales figures for distinct wine varieties. Our objective is to delve into the data, analyze trends, patterns, and factors influencing wine sales over the course of the century. By leveraging data analytics and forecasting techniques, we aim to gain actionable insights that can inform strategic decision-making and optimize sales strategies for the future.

Objective

The primary objective of this project is to analyze and forecast wine sales trends for the 20th century based on historical data provided by ABC Estate Wines. We aim to equip ABC Estate Wines with the necessary insights and foresight to enhance sales performance, capitalize on emerging market opportunities, and maintain a competitive edge in the wine industry.

1.1. Read the data as an appropriate time series data

	Month	Year	Rose_wine-Sales
1980-01-01	Jan	1980	112.0
1980-02-01	Feb	1980	118.0
1980-03-01	Mar	1980	129.0
1980-04-01	Apr	1980	99.0
1980-05-01	May	1980	116.0

Figure 1(Rose Wine Time series Data)

	Month	Year	Sparkling_wine-Sales
1980-01-01	Jan	1980	1686
1980-02-01	Feb	1980	1591
1980-03-01	Mar	1980	2304
1980-04-01	Apr	1980	1712
1980-05-01	May	1980	1471

Figure 3(Sparkling Wine Time series Data)

1.2. Plot the data

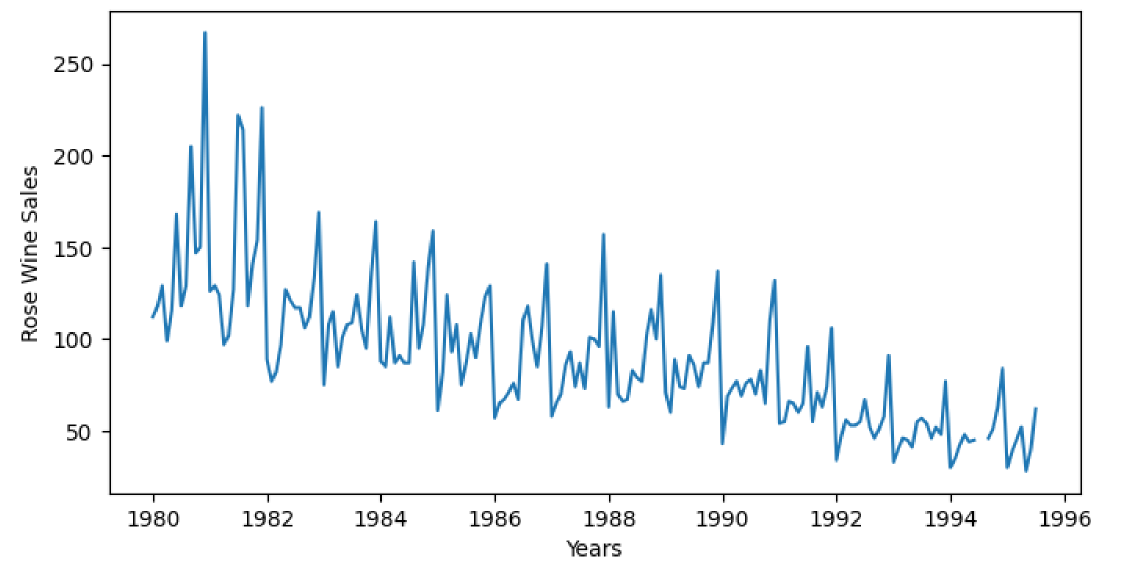


Figure 5(Rose-Wine Sales Plot)

Graph Description:

This plot shows downward Trend in sales in the years . Also it has some seasonality year wise.

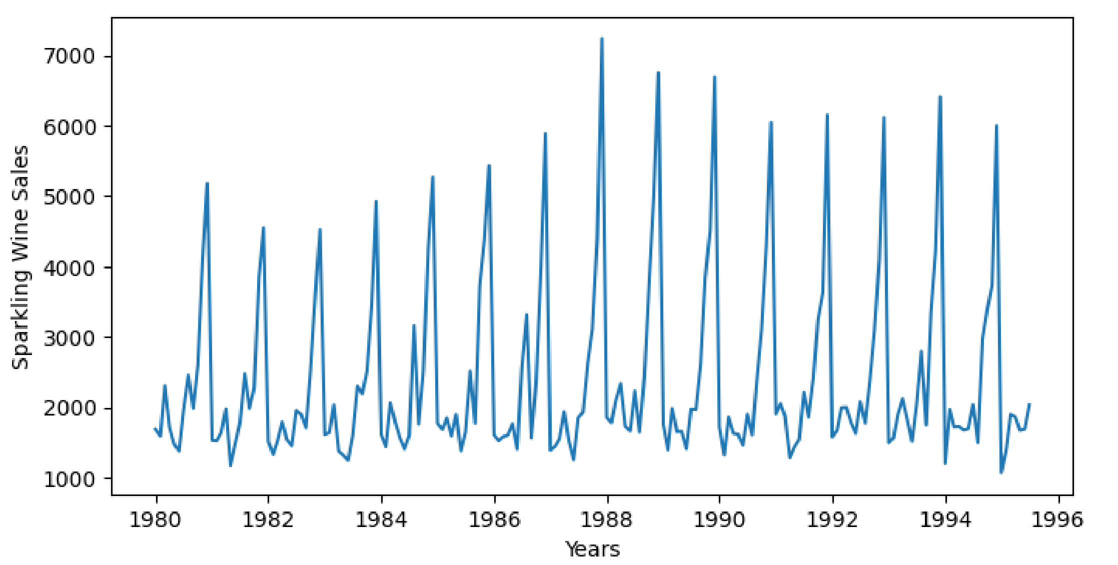


Figure 6(Sparkling-wine Sales Plot)

Graph Description:

This plot has a trend upward from 1980 to 1988 and then a downward trend from 1988 to 1966.

This also has some peaks which indicates the presence of seasonality in it.

1.3. Perform EDA

```
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Freq: MS
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Month            187 non-null    object
 1   Year             187 non-null    int32
 2   Rose_wine-Sales  185 non-null    float64
dtypes: float64(1), int32(1), object(1)
memory usage: 5.1+ KB
```

Figure 7(Data Structure for the given rose wine dataset)

We see that this dataset has 2 missing values in it

```
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Freq: MS
Data columns (total 3 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Month                 187 non-null    object
 1   Year                  187 non-null    int32
 2   Sparkling_wine-Sales  187 non-null    int64
dtypes: int32(1), int64(1), object(1)
memory usage: 5.1+ KB
```

Figure 8(Dat Structure for the given sparkling wine dataset)

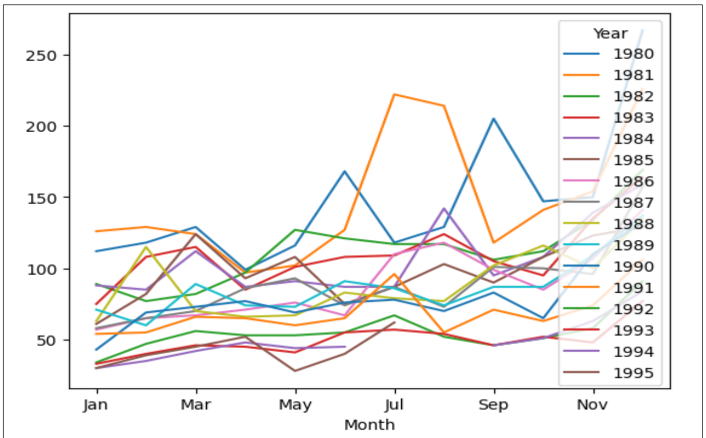


Figure 9(Month wise Rose wine Sales Plot for every year)

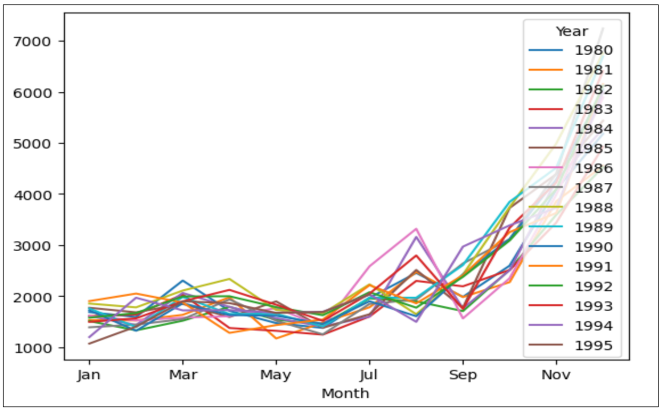


Figure 13(Month wise Sparkling wine Sales Plot for every year)

We see that there are no missing values in this dataset.

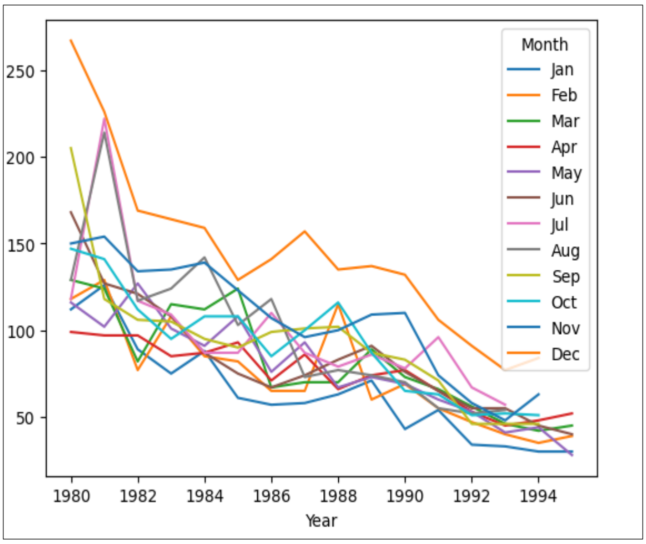


Figure 11(Rose wine Sales plot Year wise)

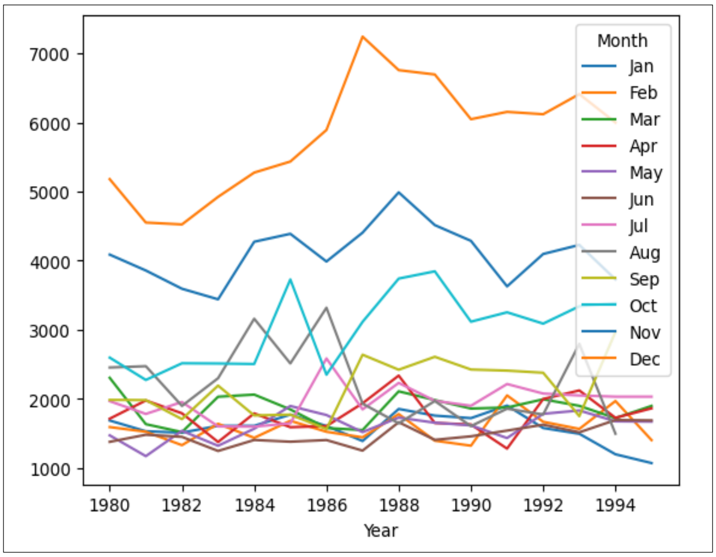


Figure 15(Month wise Sparkling wine Sales plot for every year)

Insights from the EDA:

This is a data for 16 years from 1980 to 1995 . So we observe that every year in the month of July there is a sudden rise of rose wine sales

1.4. Perform Decomposition

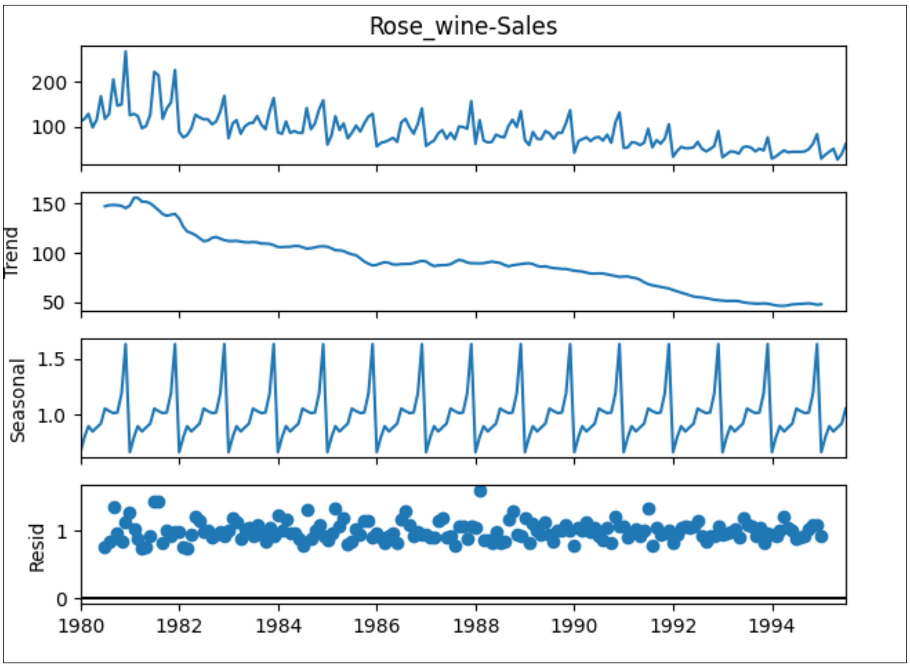


Figure 17(Decomposed Rise- wine Time series)

Figure 18(Decomposed Rise- wine Time series)

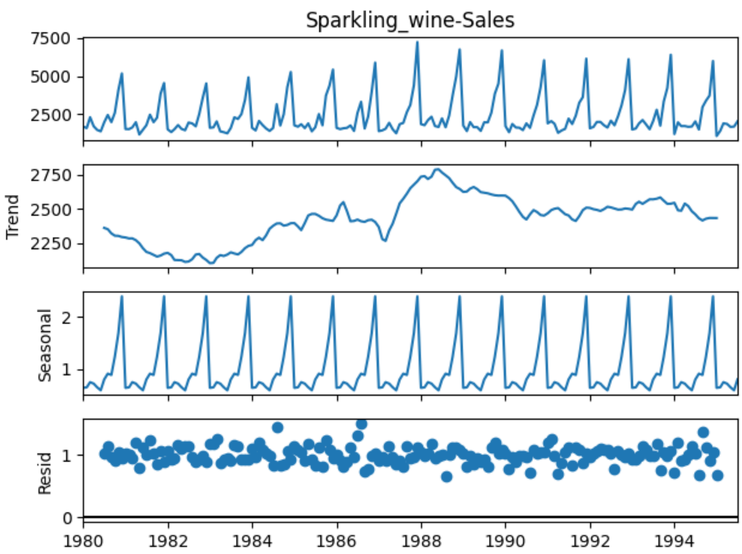


Figure 19(Decomposed Sparkling-wine Time series)

Figure 20(Decomposed Sparkling-wine Time series)

As per observation the rose wine Sales data is showing downward trend so the sales have fallen gradually with years. Seasonality is multiplicative and residual has no pattern showing which means they do not affect the time series. Now let's check Decomposed component of Sparkling-wine Sales Data

As per observation we can see that the sparkling wine data increasing trend till 1988 and then the rend decreases.After that. Seasonality is yearly and additive. The residual has no particular pattern

2. Data Pre-processing

2.1. Missing value treatment

Rose data has 2 missing values which are filled and treated with forward fill technique and Sparkling wine dataset has no missing values, so missing values not needed to be treated.

2.2. Visualize the processed data

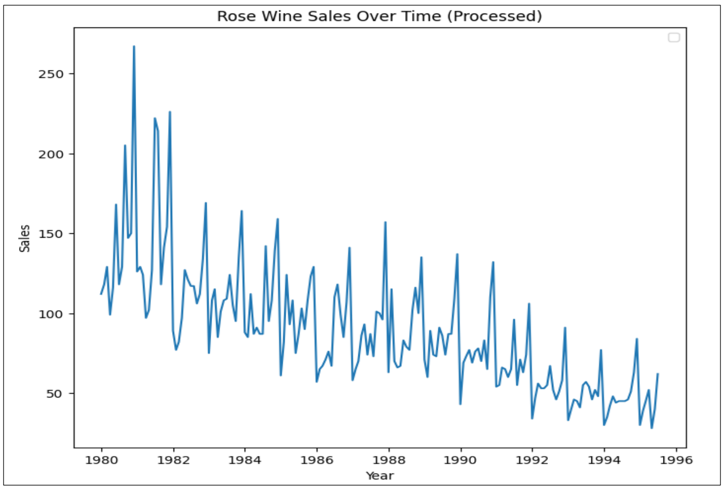


Figure 21(Rose Wine Sales Processed)

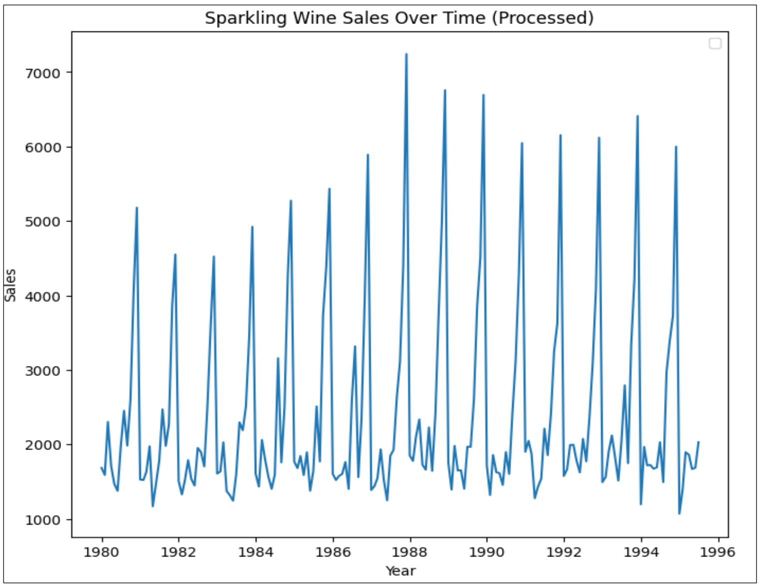


Figure 23(Sparkling Wine Sales (Processed)

2.3. Train-test split

Training Data

	Rose_wine-Sales
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0
...	...
1992-08-01	52.0
1992-09-01	46.0
1992-10-01	51.0
1992-11-01	58.0
1992-12-01	91.0

156 rows × 1 columns

Test Data

	Rose_wine-Sales
1993-01-01	33.0
1993-02-01	40.0
1993-03-01	46.0
1993-04-01	45.0
1993-05-01	41.0
1993-06-01	55.0
1993-07-01	57.0
1993-08-01	54.0
1993-09-01	46.0
1993-10-01	52.0
1993-11-01	48.0
1993-12-01	77.0
1994-01-01	30.0
1994-02-01	35.0
1994-03-01	42.0

Training Data

	Sparkling_wine-Sales
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471
...	...
1992-08-01	1773
1992-09-01	2377
1992-10-01	3088
1992-11-01	4096
1992-12-01	6119

156 rows × 1 columns

Test Data

	Sparkling_wine-Sales
1993-01-01	1494
1993-02-01	1564
1993-03-01	1898
1993-04-01	2121
1993-05-01	1831
1993-06-01	1515
1993-07-01	2048
1993-08-01	2795
1993-09-01	1749
1993-10-01	3339
1993-11-01	4227
1993-12-01	6410
1994-01-01	1197
1994-02-01	1968
1994-03-01	1720
1994-04-01	1725

Figure 25(Train_Test_Split for both wines)

3. Model Building- Original Data (Build forecasting models)

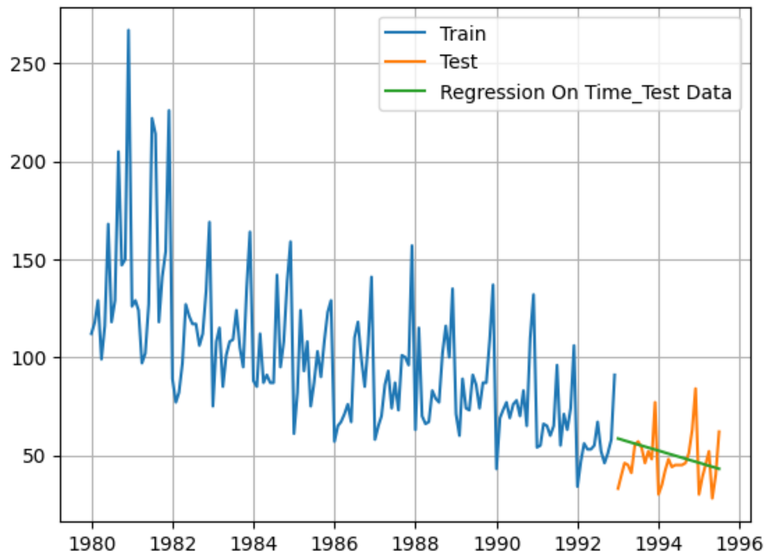
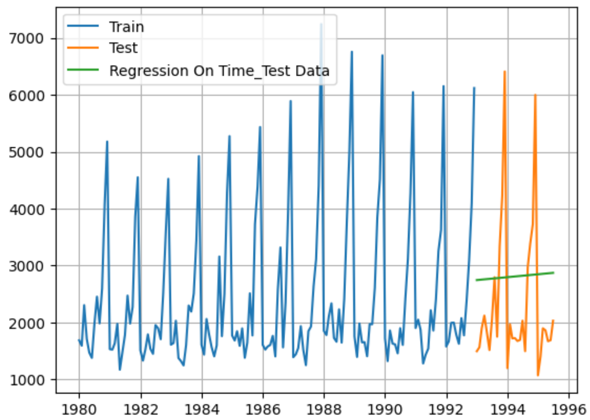


Figure 26(Linear Regression for Rose wine)

Figure 27(Linear Regression for Rose wine)




```
For RegressionOnTime forecast on the Test Data,  RMSE is 1342.993
```

	Test RMSE
RegressionOnTime	1342.992772

Figure 28(Linear Regression Model for sparkling wine)

3.1. Linear regression

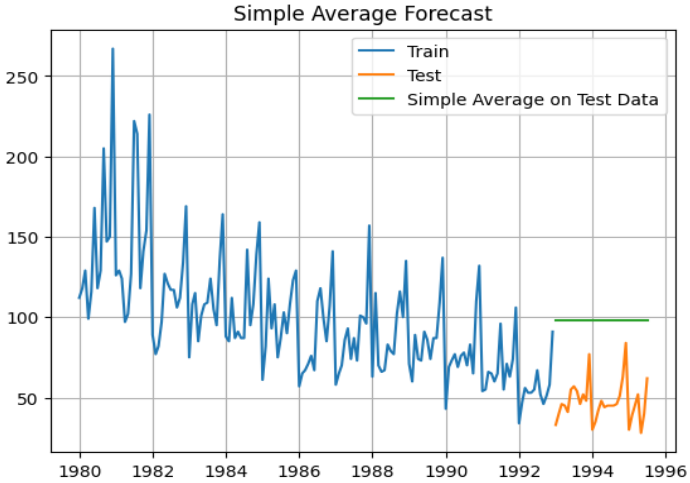


Figure 30(SA for Rose Wine)

3.2. Simple Average

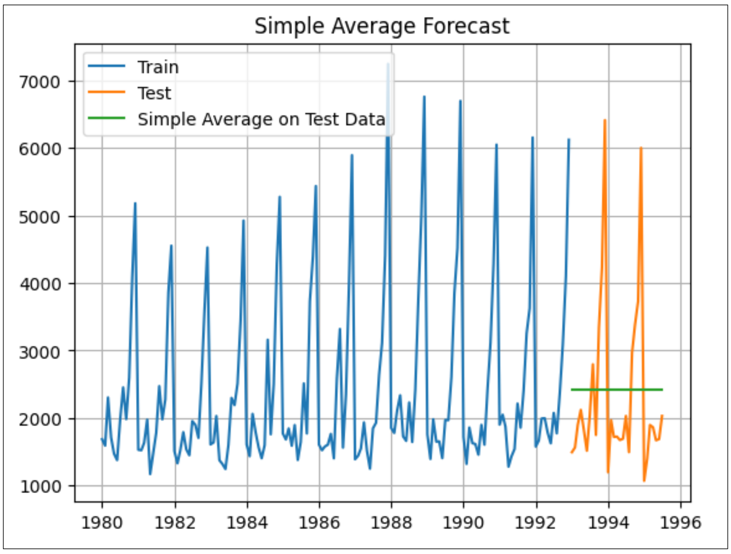


Figure 32(SA for Sparkling wine)

Figure 33(SA for Sparkling wine)

3.3. Moving Average

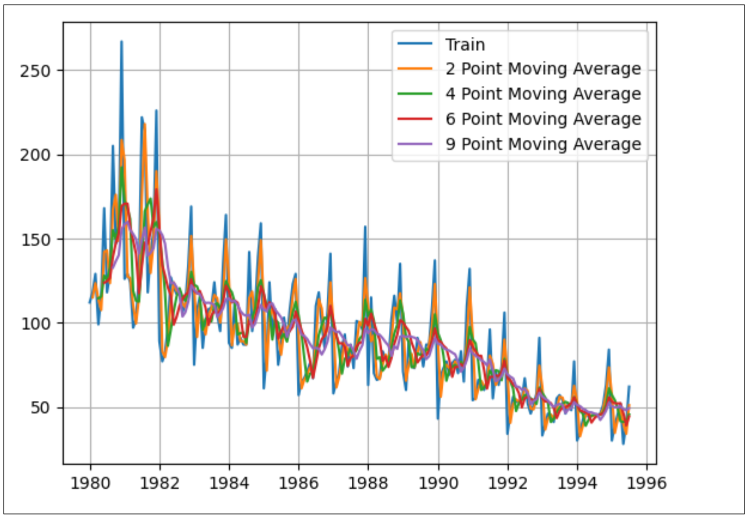


Figure 34(Moving Average for Rose Wine)

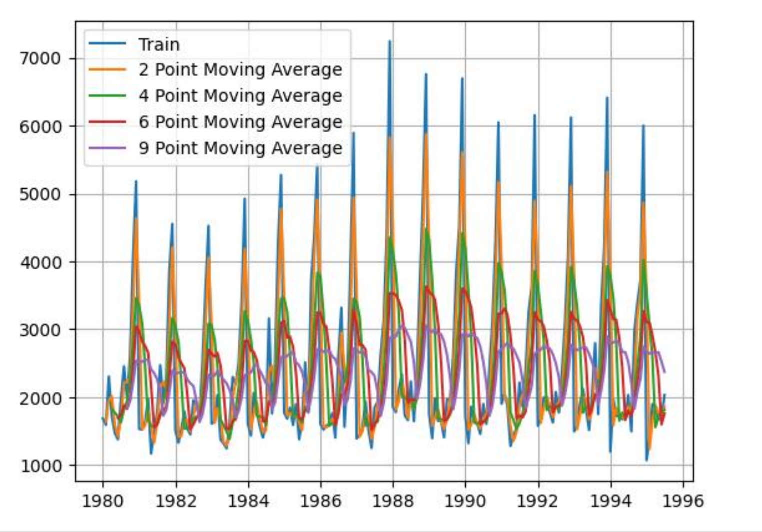


Figure 36(Moving Average for Sparkling Wine)

Figure 37(Moving Average for Sparkling Wine)

3.4. Exponential Models (Single, Double, Triple)

	Test RMSE
Alpha=0.99,SES	1294.706955
Alpha=1,Beta=0.0189:DES	4352.107605
Alpha=0.25,Beta=0.0,Gamma=0.74:TES	328.431198
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	322.969287

Figure 40(Exponential Models for Sparkling Wine)

Figure 41(Exponential Models for Sparkling Wine)

3.5. Check the performance of the models built

Performance of Exponential models for Rose Wine :

- **TES** with parameters **Alpha=0.74, Beta=2.73e-06, Gamma=5.2e-07, Gamma=0** has the lowest RMSE (9.95), suggesting it is the best-performing model.
- **DES** with **Alpha=1, Beta=0.0189** has an RMSE of 13.55, which is better than SES but not as good as the best TES model.
- **TES** with **Alpha=0.25, Beta=0.0, Gamma=0.74** has an RMSE of 15.49, which is better than SES but not as good as the best TES model.
- **SES** with **Alpha=0.99** has the highest RMSE (19.84), indicating it performs the worst among the listed models.

In conclusion, the TES model with parameters **Alpha=0.74, Beta=2.73e-06, Gamma=5.2e-07, Gamma=0** is the most accurate for the given dataset.

Performance of Exponential Models for Sparkling wine :

- **TES** with parameters **Alpha=0.74, Beta=2.73e-06, Gamma=5.2e-07, Gamma=0** has the lowest RMSE (322.97), suggesting it is the best-performing model among the ones listed.
- **SES** with **Alpha=0.99** also performs relatively well with an RMSE of 1294.71.

	Test RMSE
Alpha=0.99,SES	19.841992
Alpha=1,Beta=0.0189:DES	13.553460
Alpha=0.25,Beta=0.0,Gamma=0.74:TES	15.490694
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	9.945153

Figure 38(Exponential Models for Rose wine)

- **DES** appears to have the highest RMSE, indicating it might not be as suitable for this dataset.

The model with the lowest RMSE is generally considered the best in terms of accuracy for this specific test data.

4. Check for Stationarity

4.1. Check for stationarity

Rose Wine data :The data is not stationary as the p-value for the ADF test is greater than 0.05

Sparkling wine : The data is not stationary as the p-value for the ADF test is greater than 0.05

4.2. Make the data stationary (if needed)

Rose Wine data is made stationary with differencing of 7

Sparkling Wine data is made stationary using differencing 1

5. Model Building- Stationary Data

5.1. Generate ACF & PACF Plot and find the AR, MA values.

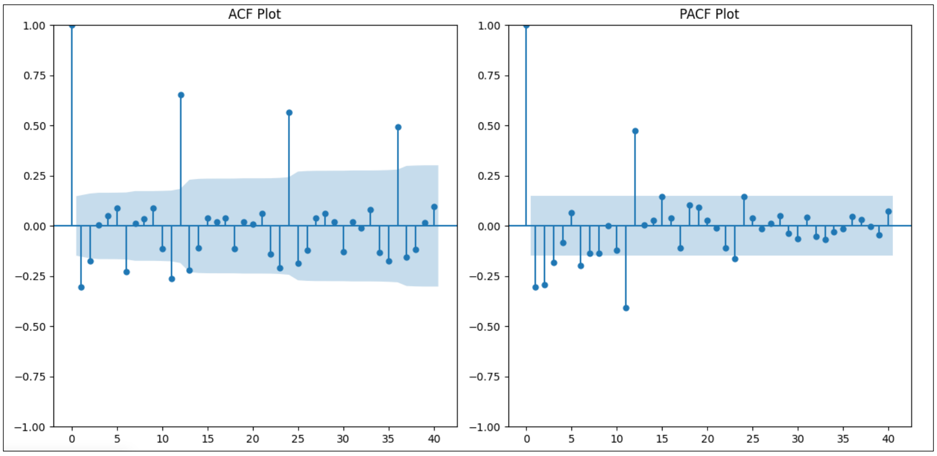


Figure 42(ACF and PACF Plots for Rose Wine)

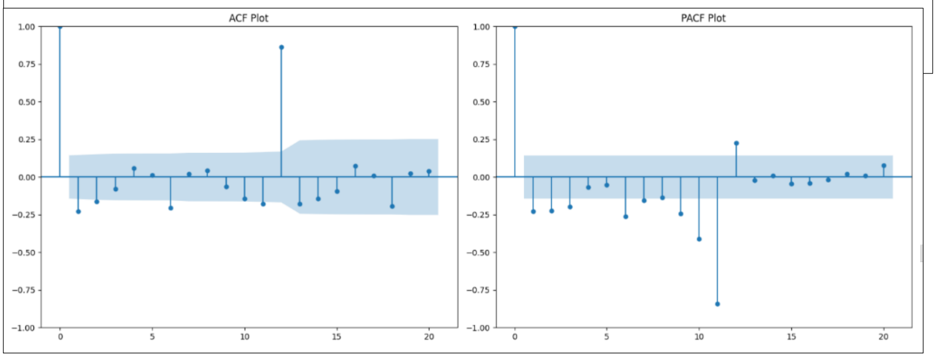


Figure 44(ACF and PACF Plot for Sparkling Wine)

Figure 45(ACF and PACF Plot for Sparkling Wine)

Insights :

For Rose wine :

AutoCorrelation Function (ACF): MA Process: For a Moving Average (MA) process of order q q, the ACF will show significant correlations up to lag q q and then cut off (drop to near zero) for higher lags. Identification: The ACF plot helps in identifying the order of the MA component by showing the number of significant lags.

Here the value of q is 2

Partial Auto-Correlation Function (PACF): AR Process: For an AutoRegressive (AR) process of order p p, the PACF will show significant correlations up to lag p p and then cut off for higher lags. Identification: The PACF plot helps in identifying the order of the AR component by showing the number of significant lags.

here the value of p is 3 and d =1

For Sparkling wine: From the above observation value of AR P= 3 and value of MA q = 2 and d=1

5.2 Build different ARIMA models- Auto ARIMA- Manual ARIMA

```
Best model:  ARIMA(3,0,1)(0,0,0)[0]
Total fit time: 15.841 seconds
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  175
Model:               SARIMAX(3, 0, 1)   Log Likelihood                -817.221
Date:                Sat, 17 Aug 2024   AIC                           1644.441
Time:                        10:51:20   BIC                           1660.265
Sample:                    01-01-1981   HQIC                          1650.860
                         - 07-01-1995
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.1875      0.060     19.780      0.000       1.070       1.305
ar.L2         -0.3222      0.105     -3.076      0.002      -0.527      -0.117
ar.L3          0.1343      0.082      1.646      0.100      -0.026       0.294
ma.L1         -0.8910      0.054    -16.483      0.000      -0.997      -0.785
sigma2       650.4739     59.891     10.861      0.000     533.091     767.857
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                37.11
Prob(Q):                              0.76   Prob(JB):                         0.00
Heteroskedasticity (H):               0.32   Skew:                             0.62
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.88
===================================================================================
```

Figure 46(Auto ARIMA for Rose Wine)

```
Best model:  ARIMA(0,0,1)(0,0,0)[0] intercept
Total fit time: 5.153 seconds
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  187
Model:               SARIMAX(0, 0, 1)   Log Likelihood               -1591.204
Date:                Sat, 17 Aug 2024   AIC                           3188.407
Time:                        13:09:25   BIC                           3198.100
Sample:                    01-01-1980   HQIC                          3192.335
                         - 07-01-1995
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept   2314.3727    249.342      9.282      0.000    1825.671    2803.075
ma.L1          0.3739      0.095      3.950      0.000       0.188       0.559
sigma2      1.479e+06    1.5e+05      9.837      0.000    1.18e+06    1.77e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.18   Jarque-Bera (JB):                57.83
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               1.84   Skew:                             1.10
Prob(H) (two-sided):                  0.02   Kurtosis:                         4.62
===================================================================================
```

Figure 50(AUTO ARIMA for sparkling Wine)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:       First_Order_Diff   No. Observations:                  175
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -816.804
Date:                Sat, 17 Aug 2024   AIC                           1645.608
Time:                        10:53:19   BIC                           1664.562
Sample:                    01-01-1981   HQIC                          1653.297
                         - 07-01-1995
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.1040      0.073     15.071      0.000       0.960       1.248
ar.L2         -0.3432      0.065     -5.246      0.000      -0.471      -0.215
ma.L1         -2.9529      0.981     -3.009      0.003      -4.876      -1.030
ma.L2          2.9505      1.961      1.505      0.132      -0.893       6.794
ma.L3         -0.9976      0.994     -1.003      0.316      -2.946       0.951
sigma2       637.7621    612.261      1.042      0.298    -562.248    1837.773
==============================================================================
```

Figure 48(Manual ARIMA for Rose Wine)

Figure 49(Manual ARIMA for Rose Wine)

```
                               SARIMAX Results
================================================================================
Dep. Variable:     Sparkling_wine-Sales   No. Observations:                  186
Model:                   ARIMA(3, 0, 3)   Log Likelihood               -1571.747
Date:                  Sat, 17 Aug 2024   AIC                           3159.494
Time:                          13:11:10   BIC                           3185.300
Sample:                      02-01-1980   HQIC                          3169.952
                           - 07-01-1995
Covariance Type:                    opg
================================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.9284      4.499      0.429      0.668      -6.890      10.747
ar.L1          0.4025      0.109      3.681      0.000       0.188       0.617
ar.L2         -0.9930      0.020    -49.506      0.000      -1.032      -0.954
ar.L3          0.4072      0.100      4.067      0.000       0.211       0.603
ma.L1         -0.9560      0.263     -3.632      0.000      -1.472      -0.440
ma.L2          0.9460      0.211      4.484      0.000       0.532       1.360
ma.L3         -0.9892      0.176     -5.617      0.000      -1.334      -0.644
sigma2      1.639e+06   4.81e-06   3.41e+11      0.000    1.64e+06    1.64e+06
==============================================================================
```

Figure 52(Manual ARIMA for Sparkling wine)

Figure 53(Manual ARIMA for Sparkling wine)

5.2. Build different SARIMA models- Auto SARIMA- Manual SARIMA

```
Best model:  ARIMA(0,0,1)(0,1,1)[12] intercept
Total fit time: 37.229 seconds
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  187
Model:             SARIMAX(0, 0, 1)x(0, 1, 1, 12)   Log Likelihood               -1284.819
Date:                            Sat, 17 Aug 2024   AIC                           2577.638
Time:                                    16:19:27   BIC                           2590.297
Sample:                                01-01-1980   HQIC                          2582.773
                                     - 07-01-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     23.0819     16.754      1.378      0.168      -9.755      55.919
ma.L1          0.1302      0.070      1.855      0.064      -0.007       0.268
ma.S.L12      -0.5468      0.052    -10.520      0.000      -0.649      -0.445
sigma2      1.374e+05   1.04e+04     13.210      0.000    1.17e+05    1.58e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                77.19
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.23   Skew:                             0.83
Prob(H) (two-sided):                  0.43   Kurtosis:                         5.80
===================================================================================
```

Figure 56(AUTO SARIMA for sparkling wine)

Figure 57(AUTO SARIMA for sparkling wine)

```
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  175
Model:             SARIMAX(3, 1, 1)x(1, 0, 1, 12)   Log Likelihood                -741.480
Date:                            Sat, 17 Aug 2024   AIC                           1496.960
Time:                                    10:55:09   BIC                           1519.073
Sample:                                01-01-1981   HQIC                          1505.931
                                     - 07-01-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2006      0.073      2.761      0.006       0.058       0.343
ar.L2         -0.1438      0.076     -1.900      0.057      -0.292       0.005
ar.L3         -0.1407      0.092     -1.532      0.125      -0.321       0.039
ma.L1         -0.8582      0.040    -21.255      0.000      -0.937      -0.779
ar.S.L12       0.9696      0.014     67.930      0.000       0.942       0.998
ma.S.L12      -0.6455      0.076     -8.475      0.000      -0.795      -0.496
sigma2       265.2772     24.653     10.761      0.000     216.959     313.595
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                26.44
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.19   Skew:                             0.03
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.91
===================================================================================
```

Figure 54(Auto SARIMA for Rose Wine)

Figure 55(Auto SARIMA for Rose Wine)

5.3. Check the performance of the models built

Performance of Sparkling Wine Models:

	Test RMSE
RegressionOnTime	1342.992772
SimpleAverageModel	1258.178011
2pointTrailingMovingAverage	769.337025
4pointTrailingMovingAverage	1122.415154
6pointTrailingMovingAverage	1298.119071
9pointTrailingMovingAverage	1451.127211
Alpha=1,Beta=0.0189:DES	4352.107605
Alpha=0.25,Beta=0.0,Gamma=0.74:TES	328.431198
Alpha=0.74,Beta=2.73e-06,Gamma=5.2e-07,Gamma=0:TES	322.969287
Best Auto ARIMA Model : ARIMA(3,0,0)	1832.865189
Best ARIMA Manual Model : ARIMA(3,0,2)	2265.774805
Best Auto SARIMA Model : SARIMAX(0, 0, 1)x(0, 1, 1, 12)	2628.991902

Figure 58(Performance of various models build for Sparkling wine dataset)

Performance of Rose Wine Models:

6. Compare the performance of the models

6.1. Compare the performance of all the models built

For Rose Wine

For sparkling Wine

The **Auto SARIMA model (SARIMAX(0, 0, 1)x(0, 1, 1, 12))** has a lower RMSE of **2628.99**, compared to the **manual SARIMA model (SARIMAX(2, 0, 1)x(2, 1, 1, 12))** with an RMSE of **2645.55**. This indicates that the Auto SARIMA model offers slightly better predictive accuracy. Therefore, **the Auto SARIMA model is the better choice based on RMSE**.

6.2. Choose the best model with proper rationale

For Rose Wine : Therefore, **the Manual SARIMA model is the better choice based on RMSE**

For sparkling wine:Therefore, **the Auto SARIMA model is the better choice based on RMSE**

6.3. Rebuild the best model using the entire data

Built in the IPYNB file.

6.4. Make a forecast for the next 12 months

7. Actionable Insights & Recommendations

7.1. Conclude with the key takeaways (actionable insights and recommendations) for the business