Retail Giant Sales Forecasting Assignment

Presented by

Yathestha Siddh

Business Problem Statement

Global Mart is an online supergiant store that has worldwide operations. This store takes orders and delivers across the globe and deals with all the major product categories — consumer, corporate, and home office.

As a sales manager for this store, you have to forecast the sales of the products for the next 6 months, so that you have a proper estimate and can plan your inventory and business processes accordingly.

Steps Covered in this Assignment to solve business problem

1. Import necessary libraries
2. Import shared sample time-series data
3. Data Preparation
4. The most profitable market segment using COV
5. Time Series Analysis
6. Building models to forecast sales and quantity for the most profitable market segment
7. Finding the optimum method in Smoothing Techniques and ARIMA Techniques
8. Conclusion:- The most suitable model, is based on the MAPE value comparison.

Data Preparation

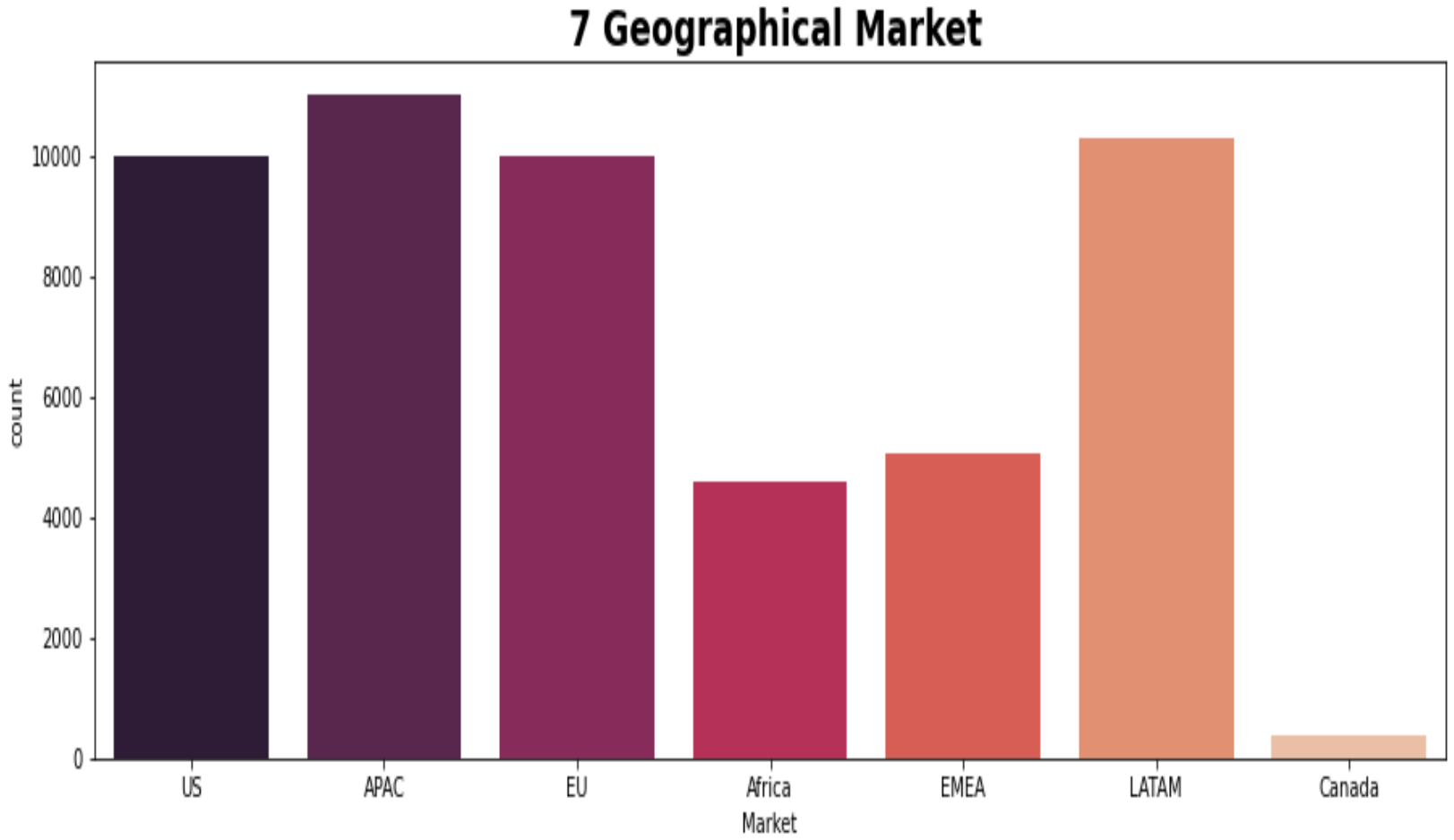
Data Attributes with Description

Attributes	Description
Order-Date	The date on which the order was placed
Segment	The segment to which the product belongs
Market	The market to which the customer belongs
Sales	Total sales value of the transaction
Profit	Profit made on the transaction

```
#Checking the datatypes
retail.info()
```

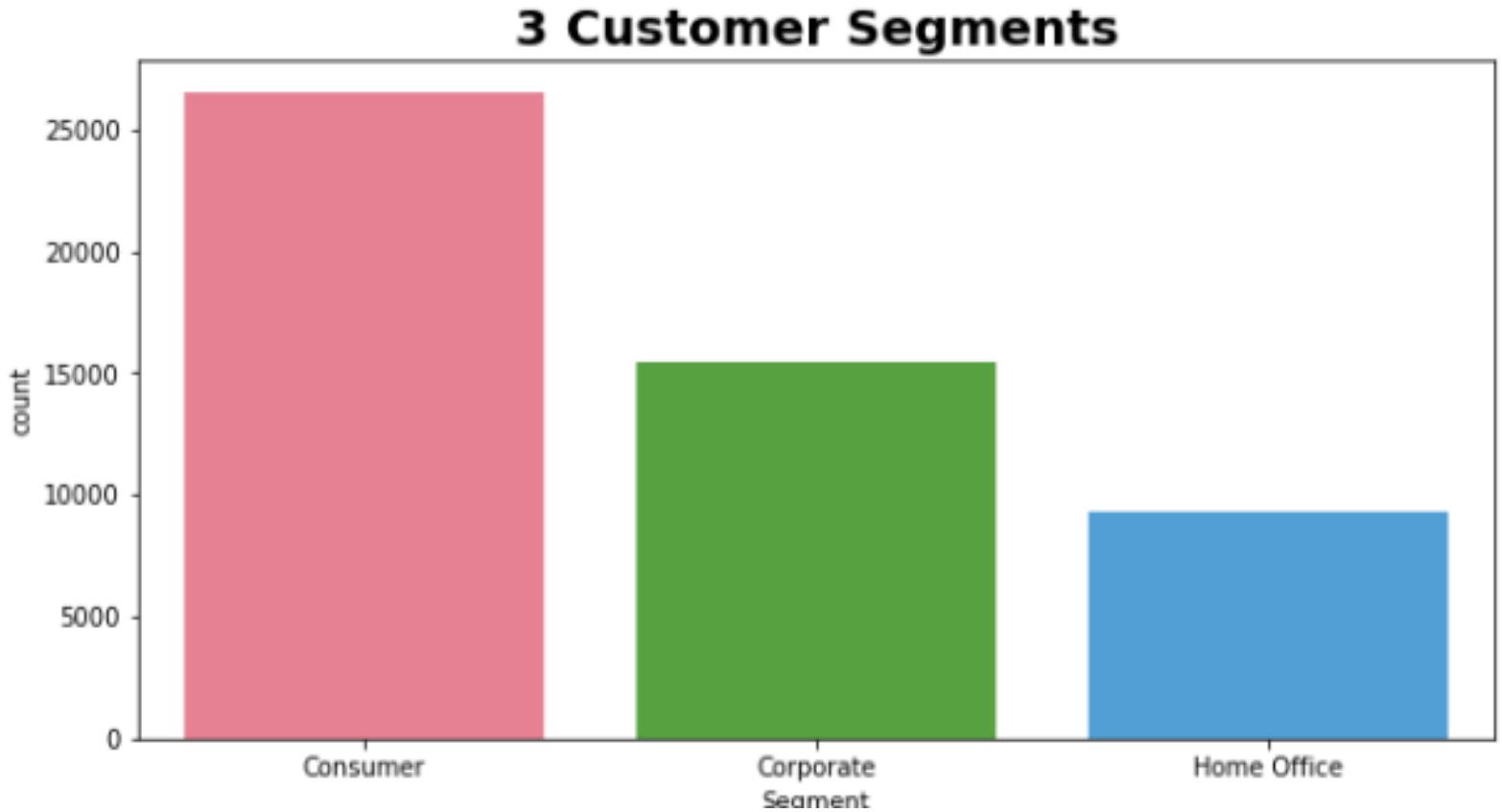
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51290 entries, 0 to 51289
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Order Date   51290 non-null  object 
 1   Segment      51290 non-null  object 
 2   Market       51290 non-null  object 
 3   Sales        51290 non-null  float64
 4   Profit       51290 non-null  float64
dtypes: float64(2), object(3)
memory usage: 2.0+ MB
```

Geographical Markets



Market
Africa
APAC (Asia Pacific)
Canada
EMEA(Middle East)
EU (European Union)
LATAM (Latin America)
US (United States)

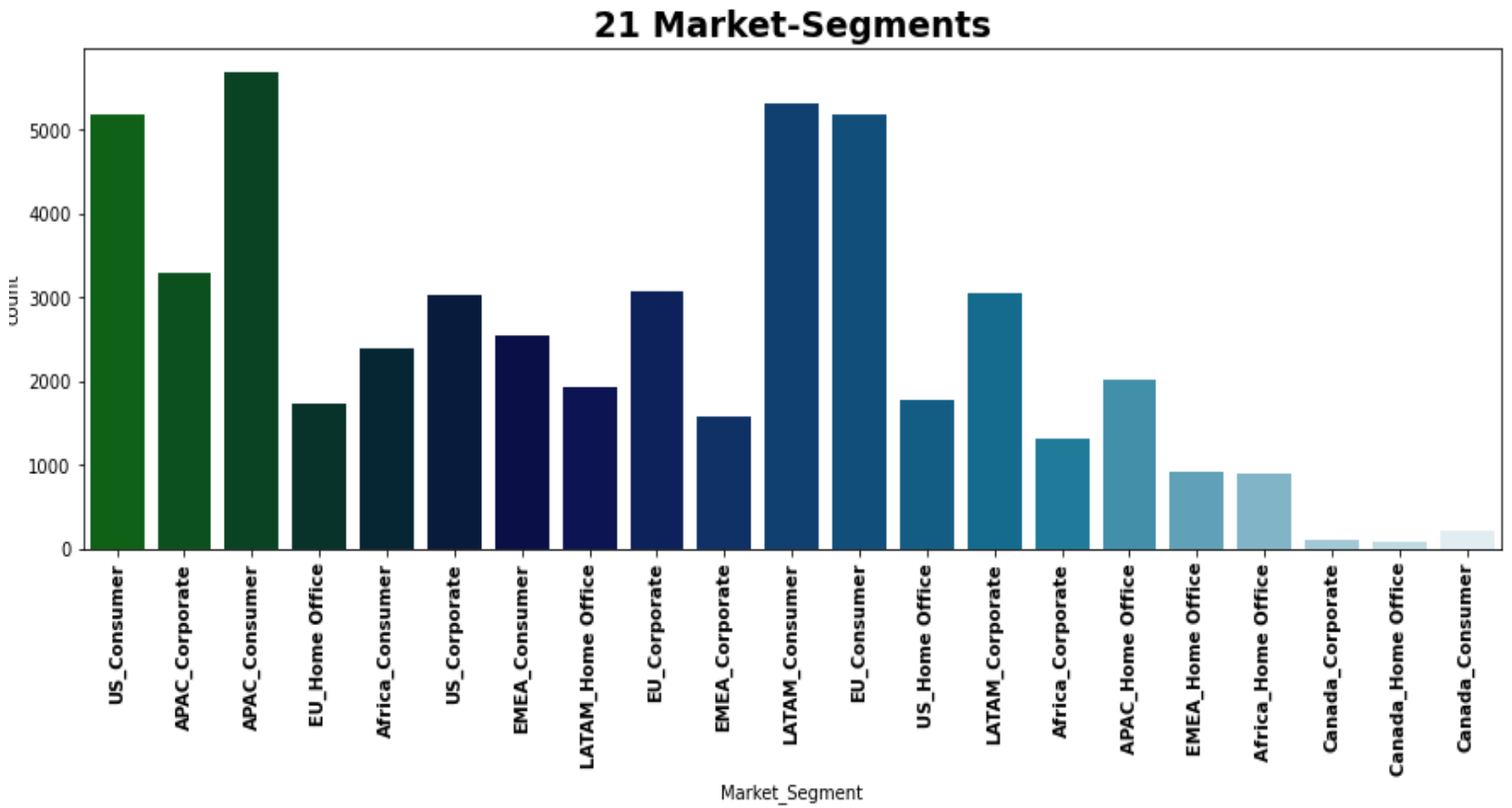
Customer Segments



Segment
Consumer
Corporate
Home Office

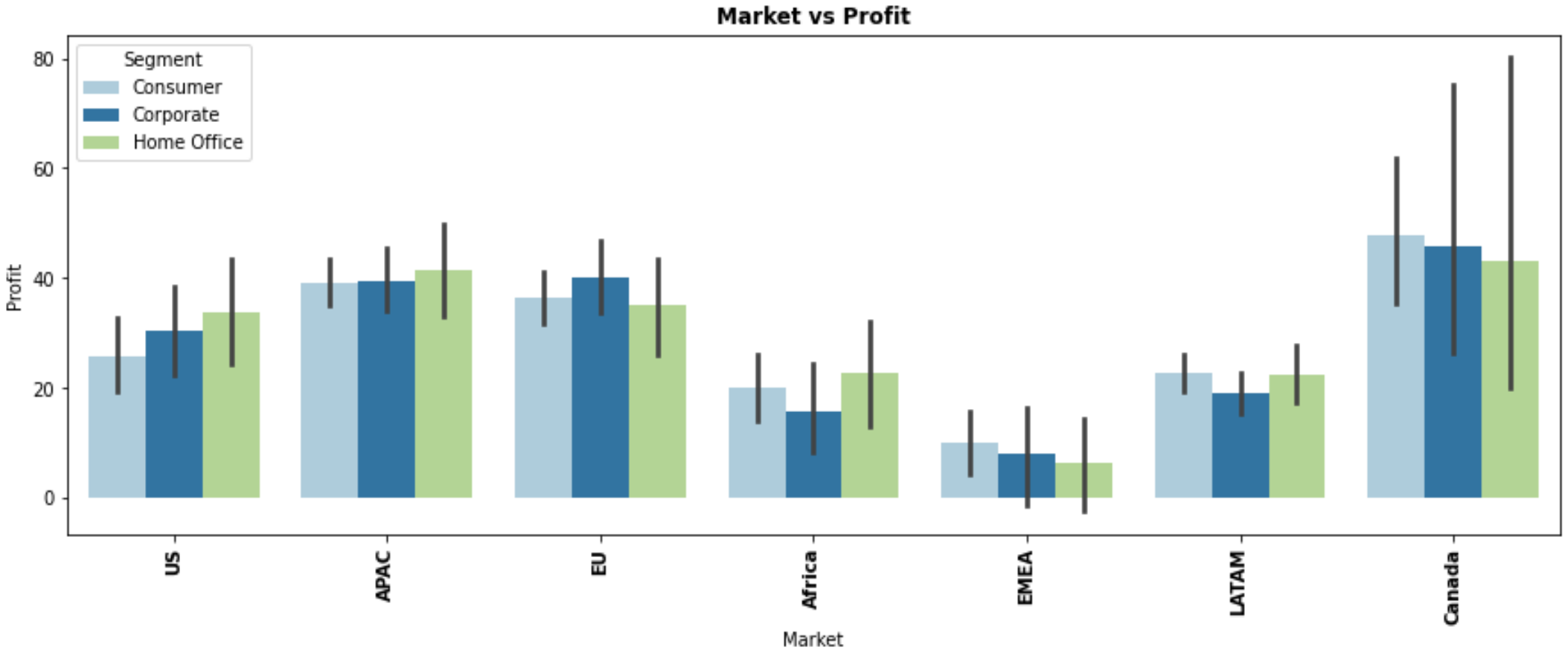
Market Segments

We can see that the there are 7 different geographical markets and 3 major customer segments using which we formed 21 Market-segments

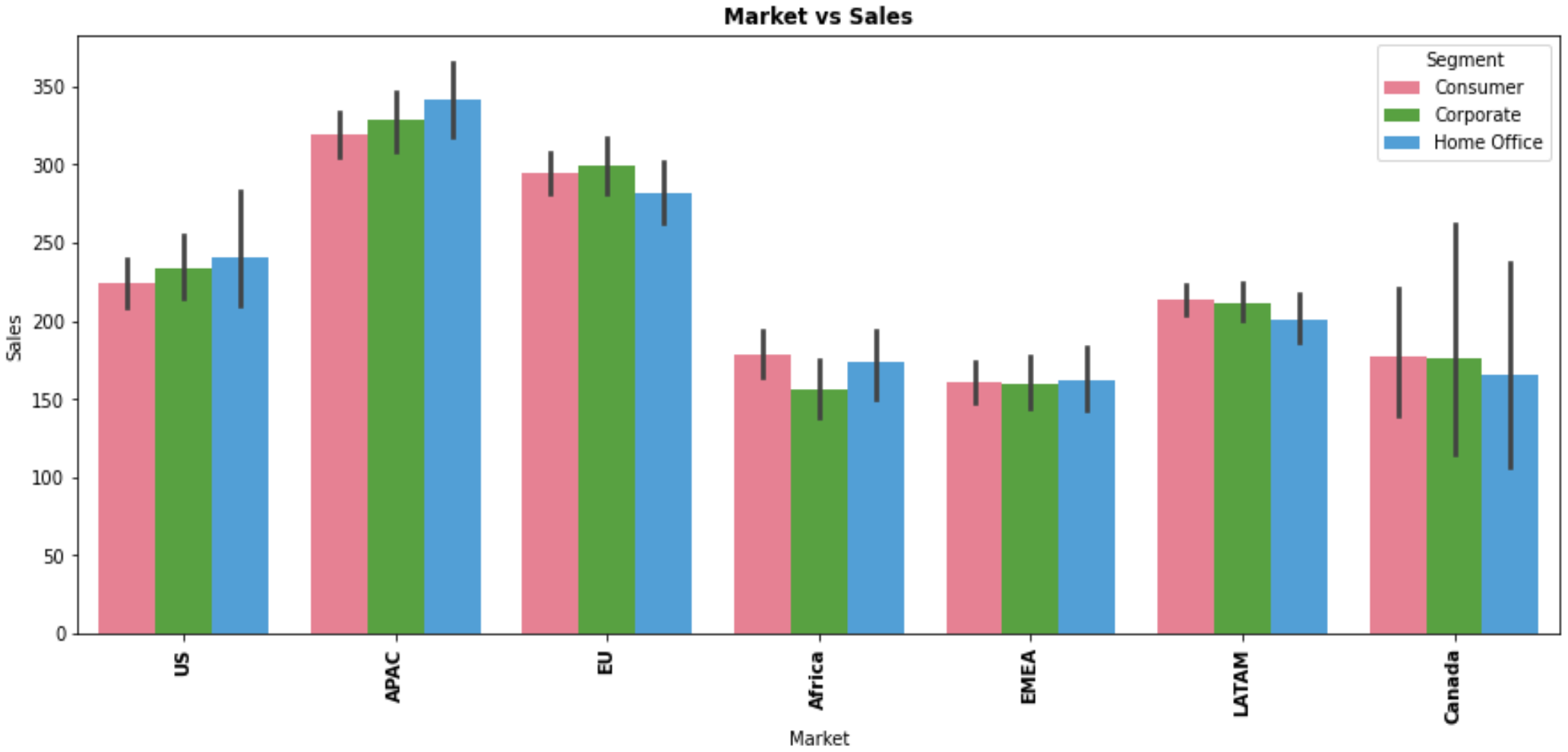


Market Segment
APAC_Consumer
LATAM_Consumer
US_Consumer
EU_Consumer
APAC_Corporate
EU_Corporate
LATAM_Corporate
US_Corporate
EMEA_Consumer
Africa_Consumer
APAC_Home Office
LATAM_Home Office
US_Home Office
EU_Home Office
EMEA_Corporate
Africa_Corporate
EMEA_Home Office
Africa_Home Office
Canada_Consumer
Canada_Corporate
Canada_Home Office

As We have observed from the plot, that Canadian markets in all the three segments have the most profit and EMEA markets have the lest profit



As We have observed from the above plot, the APAC market is having the highest Sales in all the three segments and EMEA and Africa have the lowest sales



Coefficient of variation calculated on the profit for the 21 market segments.

	Market_Segment	Mean	Std	CoV
0	APAC_Consumer	4400.894243	2300.457687	0.522725
1	APAC_Corporate	2574.919807	1364.837734	0.530051
12	EU_Consumer	3699.977143	2202.282289	0.595215
15	LATAM_Consumer	2295.555697	1569.632686	0.683770
13	EU_Corporate	2216.299429	1600.336696	0.722076
16	LATAM_Corporate	1122.633016	990.360880	0.882177
14	EU_Home Office	1224.456536	1148.627937	0.938072
2	APAC_Home Office	1511.088314	1523.508658	1.008219
18	US_Consumer	2686.740912	2715.031412	1.010530
19	US_Corporate	1754.199083	1880.200775	1.071829
20	US_Home Office	1132.065762	1272.476439	1.124030
17	LATAM_Home Office	818.398941	957.275713	1.169693
6	Canada_Consumer	225.987632	282.555788	1.250315
3	Africa_Consumer	957.707000	1254.932072	1.310351
7	Canada_Corporate	90.980294	162.493114	1.786025
4	Africa_Corporate	412.617571	780.566850	1.891744
5	Africa_Home Office	377.221071	759.322203	2.012937
8	Canada_Home Office	118.003750	279.632866	2.369695
9	EMEA_Consumer	423.960286	1124.552711	2.652495
10	EMEA_Corporate	182.642643	1160.698430	6.355024
11	EMEA_Home Office	84.231366	651.283095	7.732073

a) **Lowest CoV is 0.52272 and the highest is 7.732073**

b) **Lowest CoV “0.52272” is for Market Segment "APAC_Consumer“**

c) **We could infer from the CoV comparison of each profitable Market Segment, that the corresponding Market Segment is "APAC_Consumer" is the most profitable.**

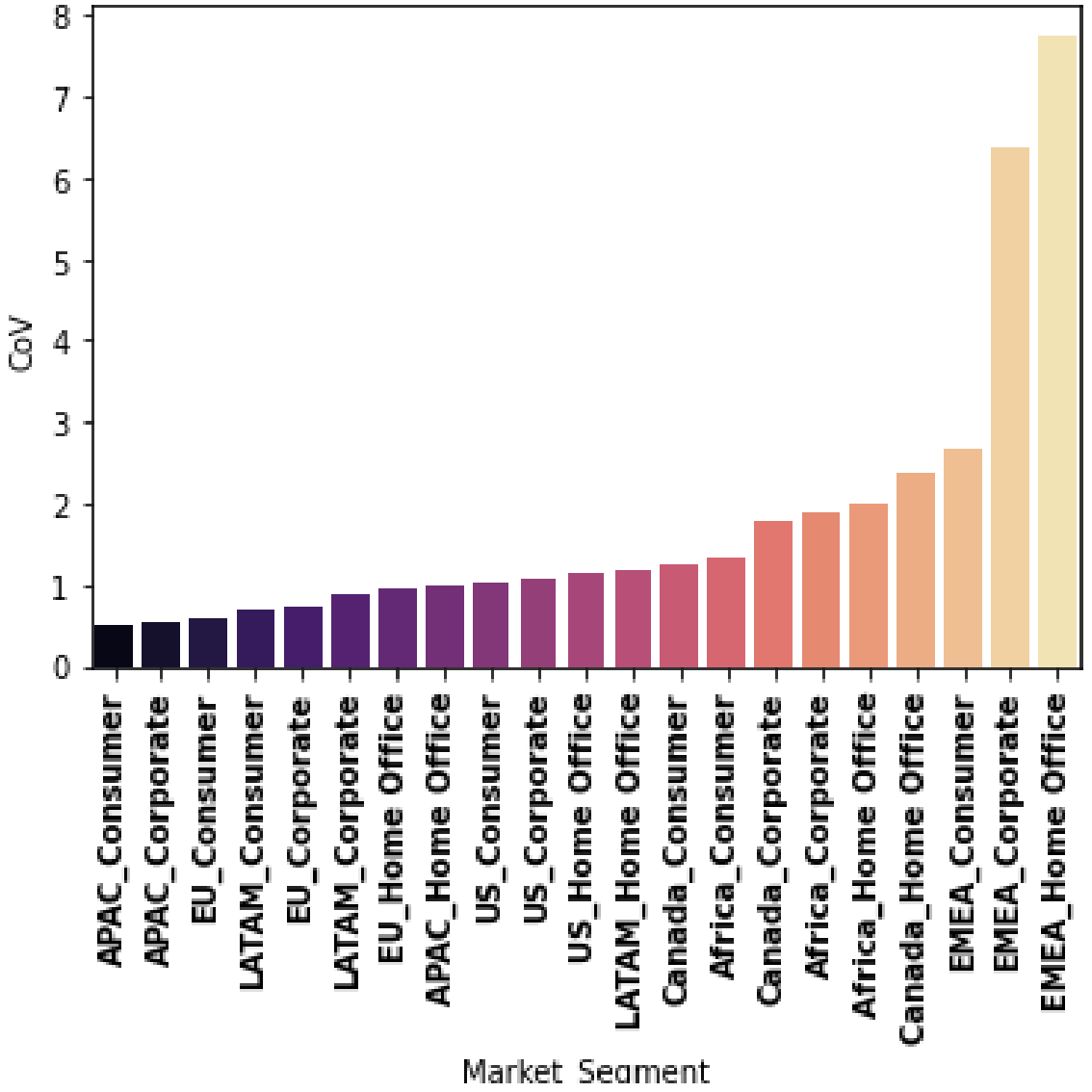
Why we are using the Coefficient of variation calculated to identify the most profitable market segment

It is meaningless to compare the 21 market segment's profits based on the standard deviation and their mean as these value vary a lot.

A better metric to compare the variance between the segment we will use the coefficient of variation which will normalize the standard deviation with the mean and give us a comparative figure based on which we could easily identify the most profitable market segment.

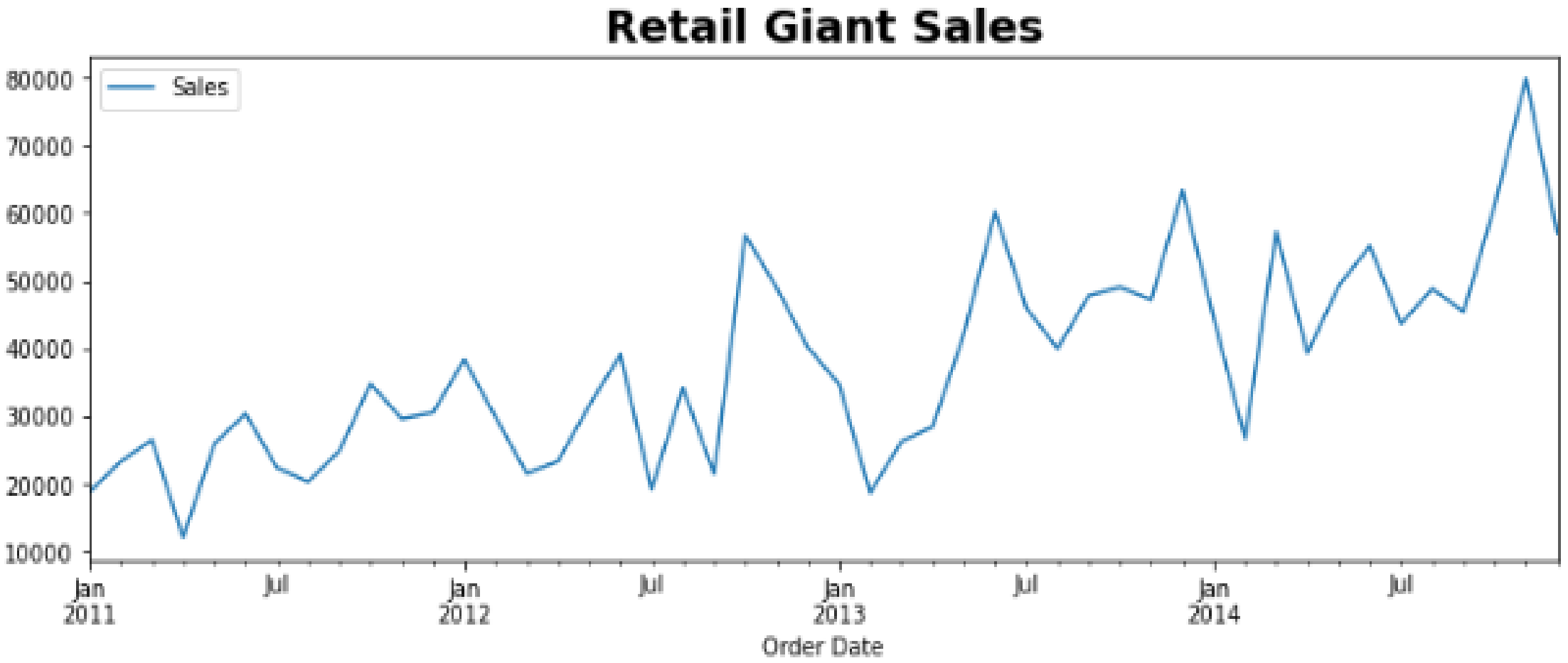
Now, We compare the variance between the market segments using the coefficient of variation so that we could identify the most profitable market segment.

We want to forecast the sales where the market segment is reliable or in other words, there is less variation in the profits.



Time Series Analysis

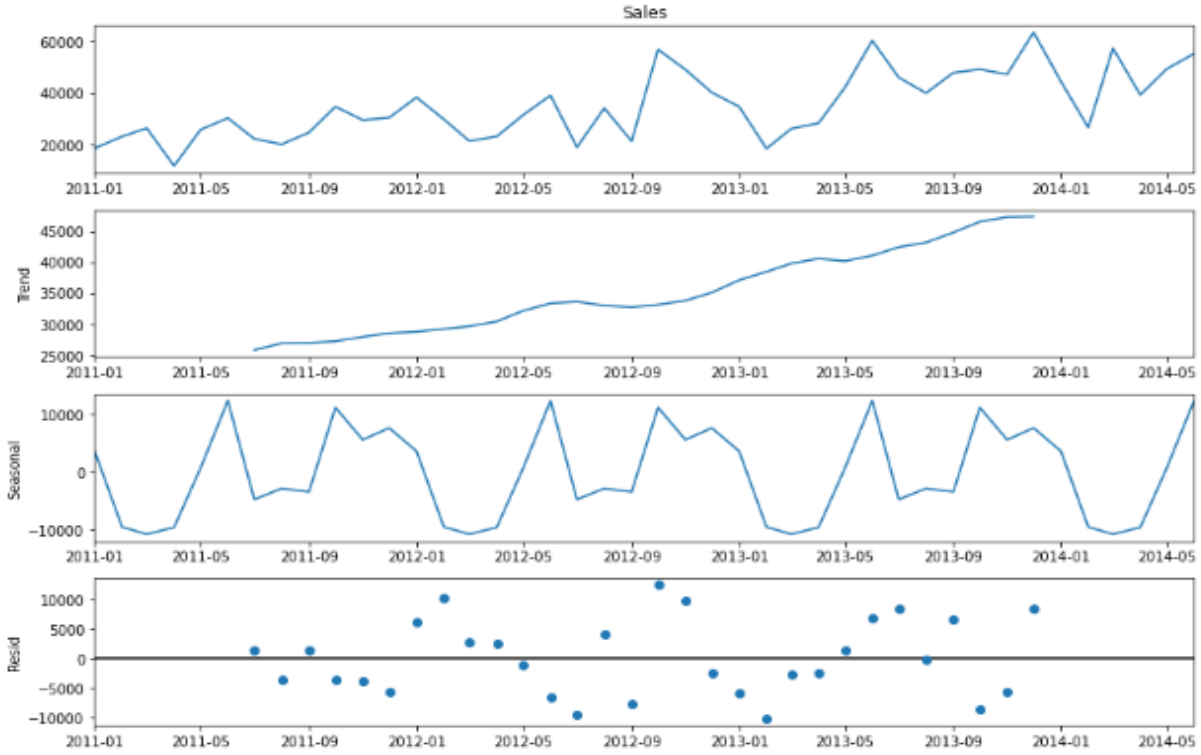
Plot time series data



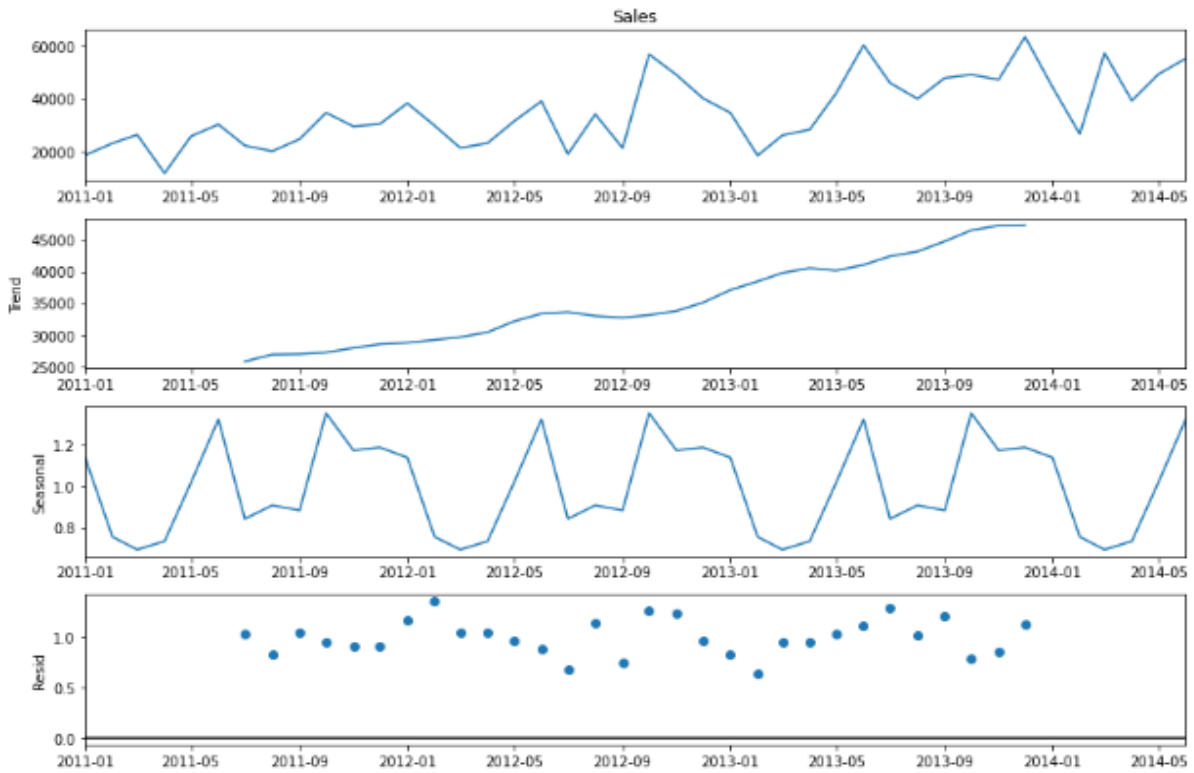
Time series Decomposition

Let’s understand how a time series can be split into its various components that is the Trend, Seasonality, and residuals

Additive seasonal decomposition



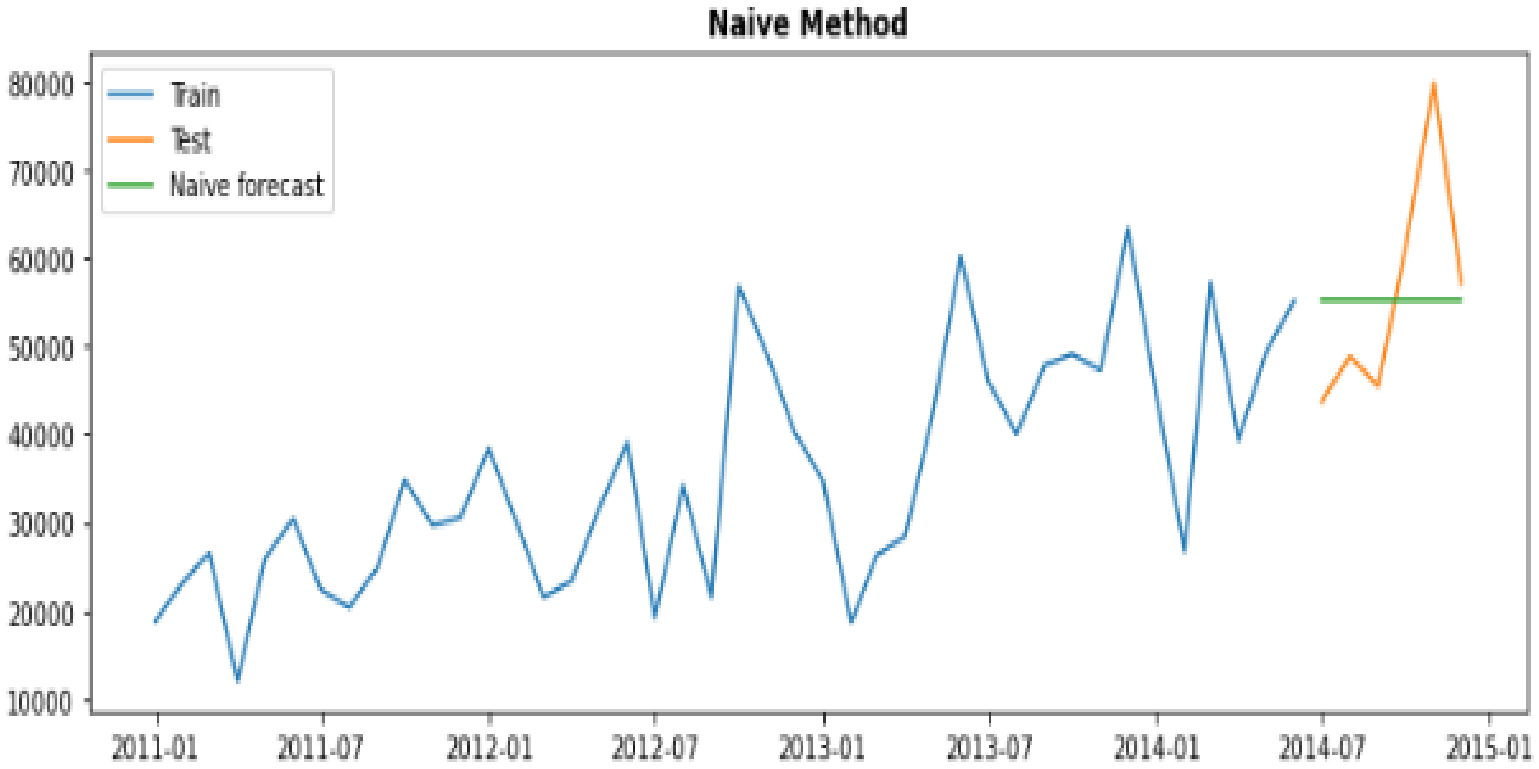
Multiplicative seasonal decomposition



Building models to forecast sales and quantity for the most profitable market segment

Naive method

RMSE = 12355.97 and MAPE =17.47

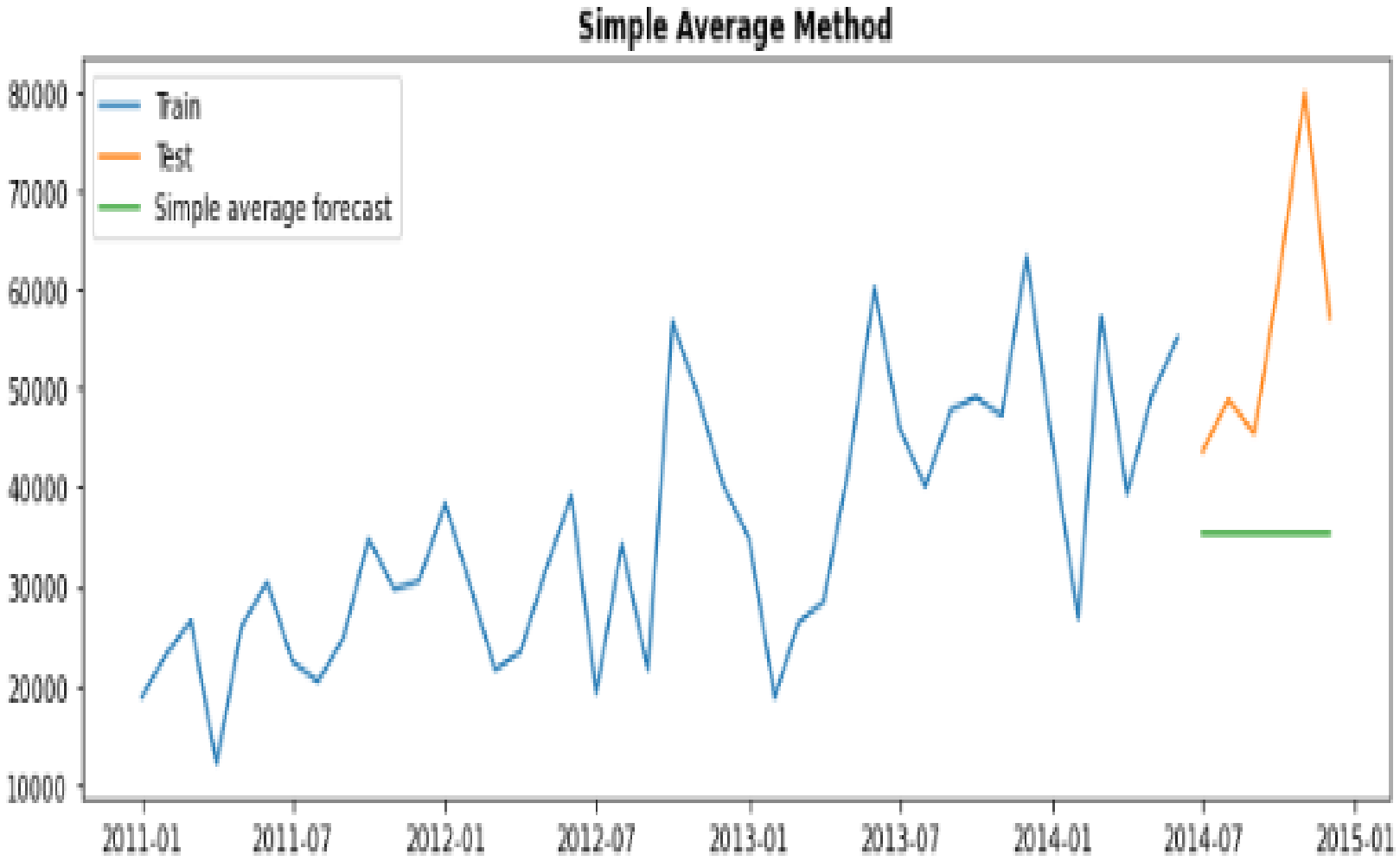


We can see that the forecast for the next six months is the same value(green line) as the last observation of the blue line

i.e. we used the last month's data which is 2014-06

Simple average method

RMSE=24146.06 and MAPE= 34.34



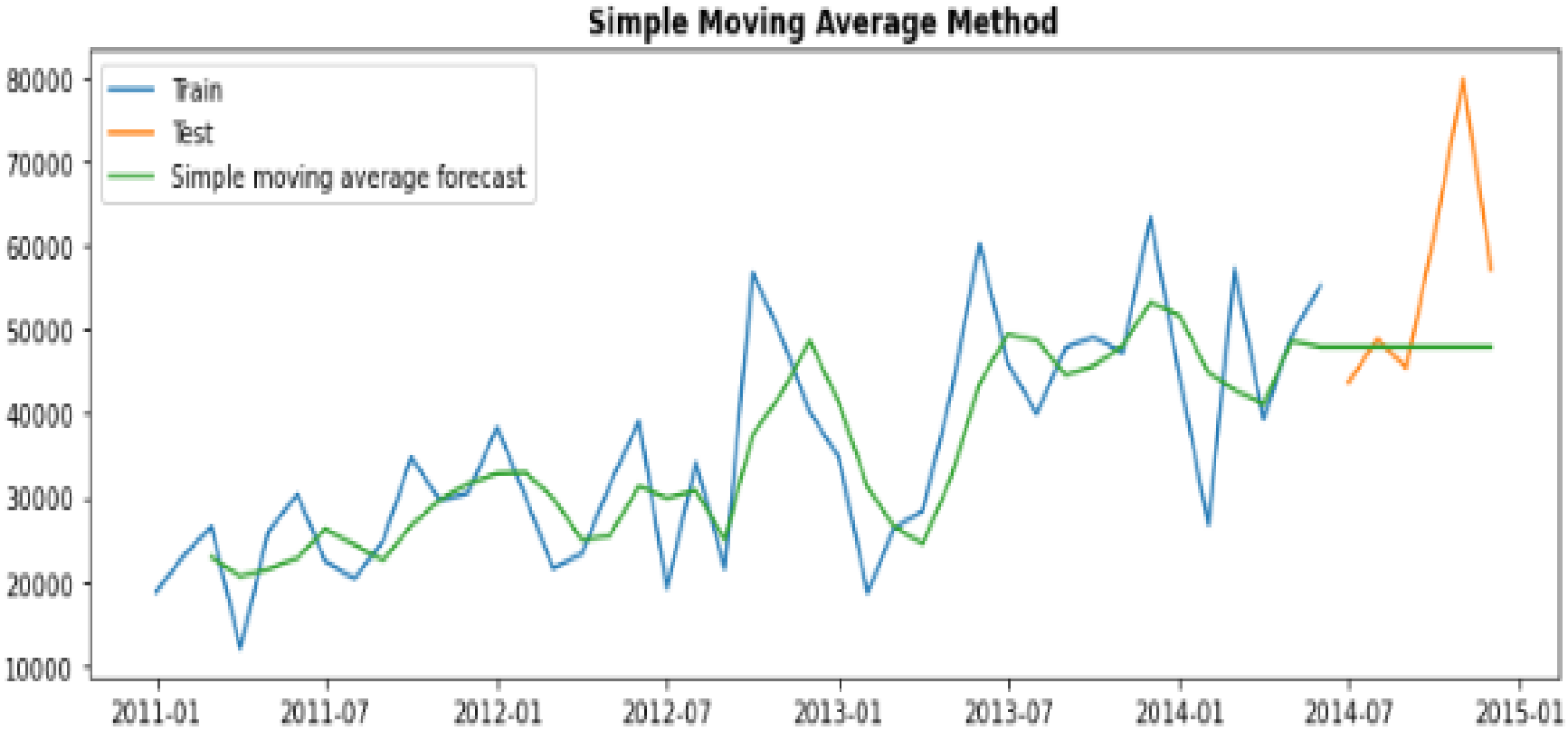
Forecast of months from 2014-07 to 2015-01 = Average of all past months' sales

The green line is the average of all the 42 months of sales data

The green line we forecasted is not showing any trend or seasonality while our train and test data had both trend and seasonality

Simple Moving Average Method

RMSE=14756.73 and MAPE= 15.82

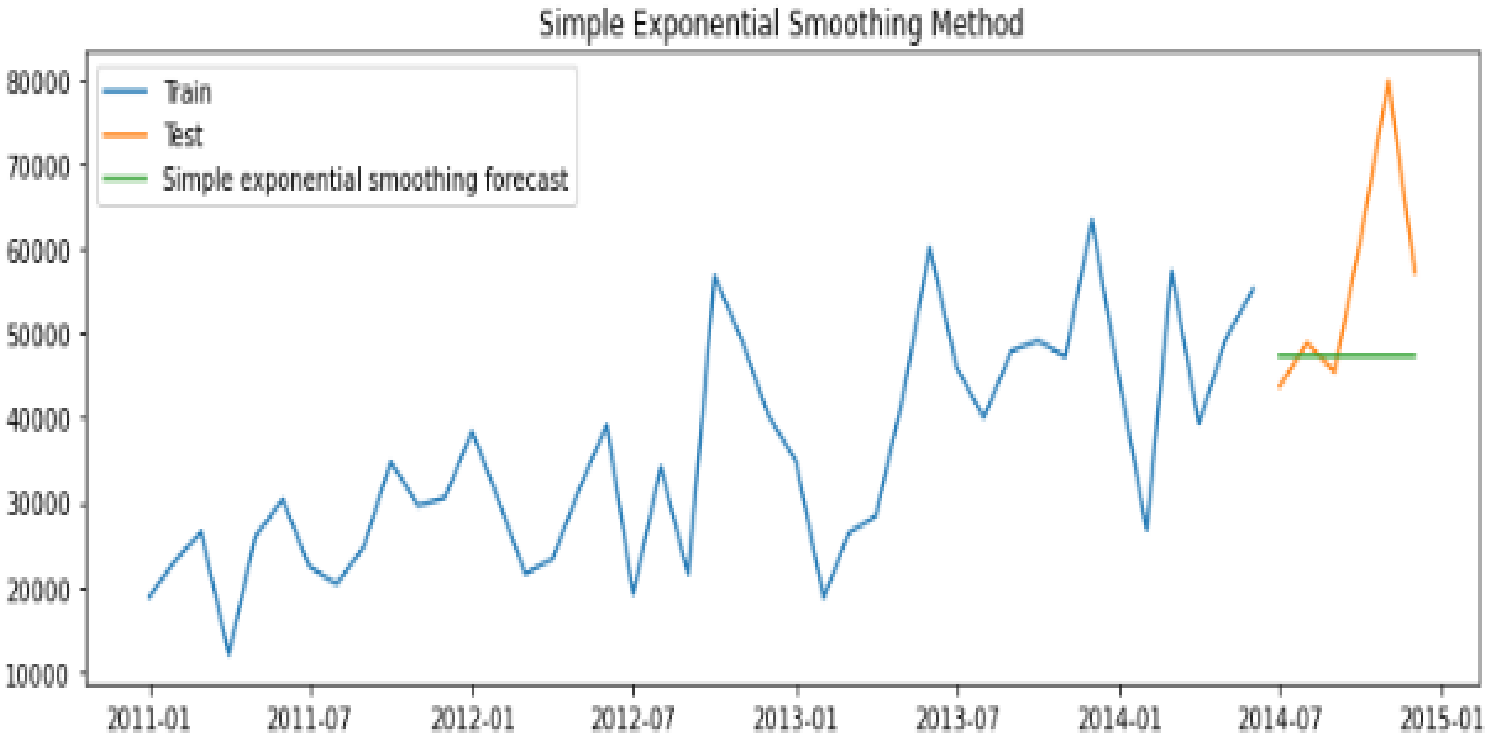


Average of only the last few observations to forecast the future
Reduces unsystematic noise in the data

Exponential smoothing methods

Simple Exponential Smoothing Method

RMSE=15011.49 and MAPE= 15.99



The most recent period's demand was multiplied by the smoothing factor.
It is a time series forecasting method for univariate data without a trend or seasonality.

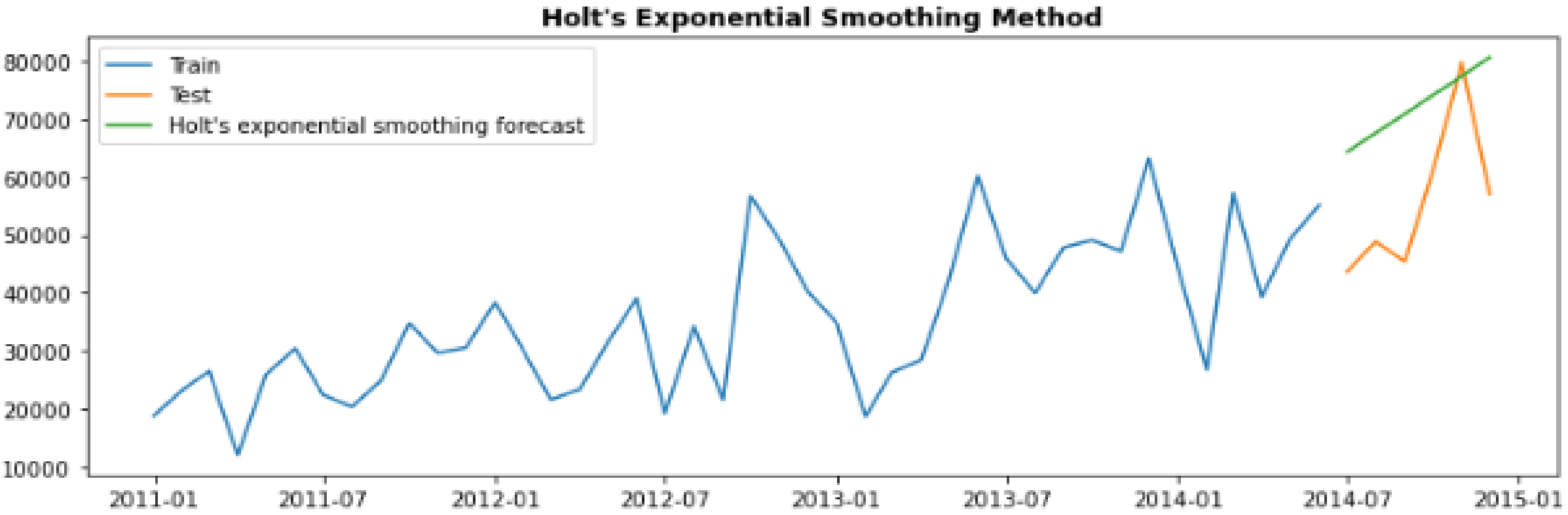
The simple exponential model captures the level of a time series.

Holt's Exponential Smoothing Method

Holt's Exponential Smoothing Captures both the level and trend of a time series in the forecast.

We can see that the forecast is a straight line, sloping upwards as Holt's method captured both level and trend

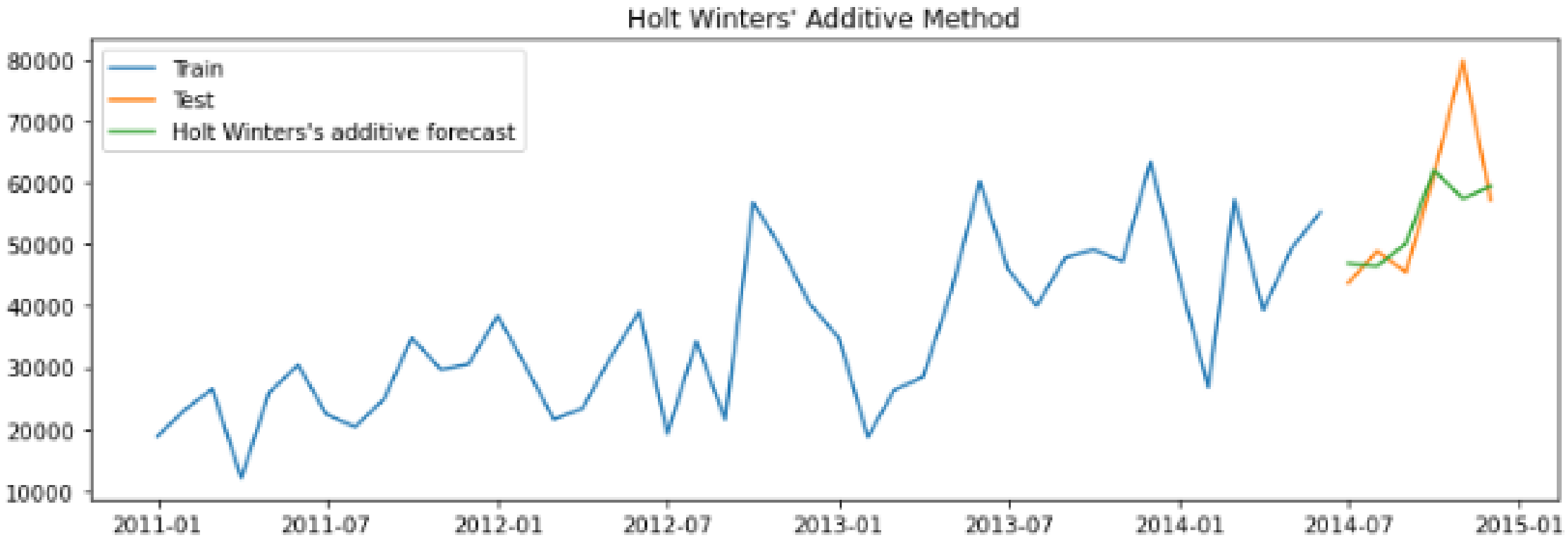
RMSE=18976.37 and MAPE= 34.57



Holt Winters Additive Method

Holt Winters' additive method with trend and seasonality Forecasts based on the level, trend, and seasonality of a time series. This method has the lowest RMSE and MAPE values and that means error measures are very less.

RMSE=9555.61 and MAPE=9.33

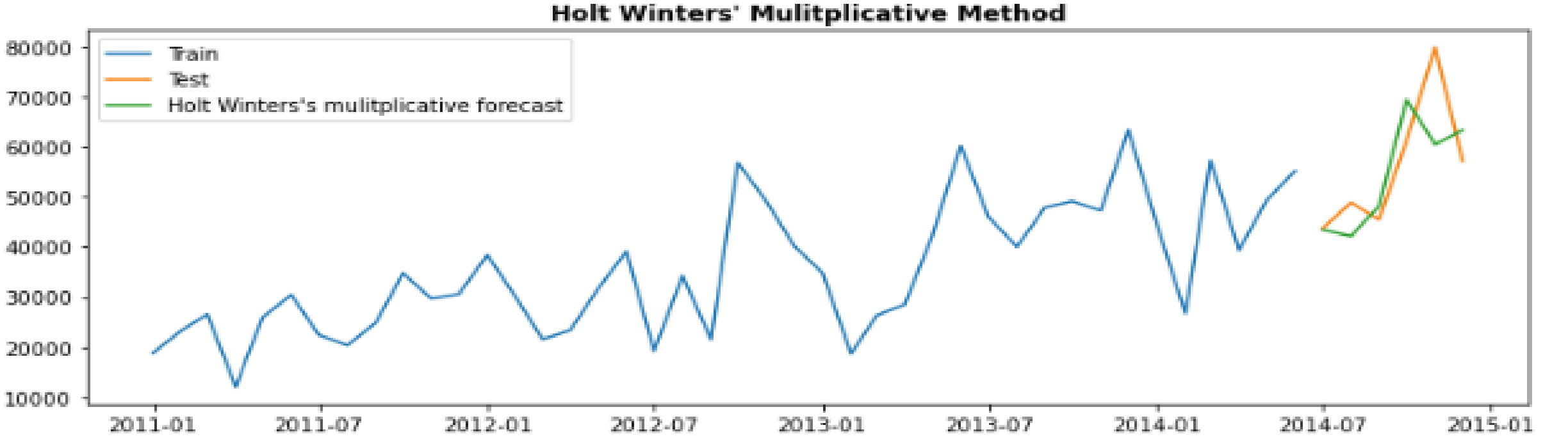


Holt Winter's multiplicative method

Holt Winter's multiplicative method with trend and seasonality Forecast= Multiplies the trended forecast by the seasonality

From the Smoothing Techniques performed we can conclude that Holt Winter's Additive Method is giving a better forecast to Holt Winter's multiplicative method.

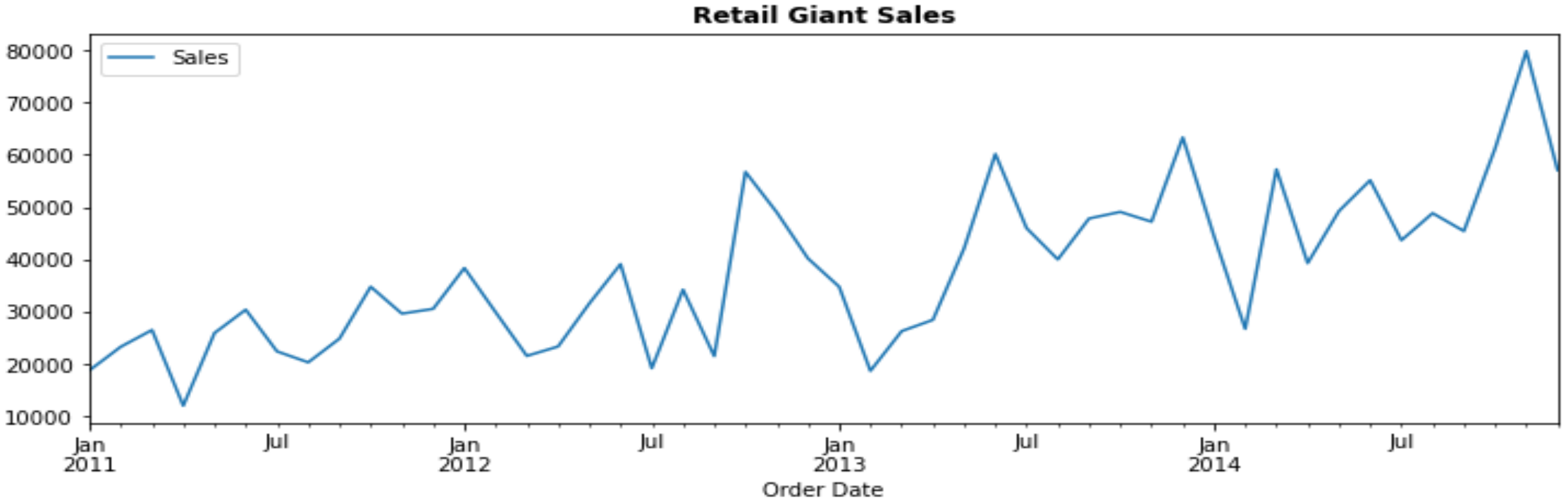
RMSE=9423.23 and MAPE=11.43



Auto-Regressive methods

In an autoregressive model, the regression technique is used to formulate a time series problem. To implement autoregressive models, we forecast future observations using a linear combination of past observations of the same variable

Stationarity vs non-stationary time series, to identify this we use Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test



Augmented Dickey-Fuller (ADF) test

Null Hypothesis (H0) : The series is not stationary p−value>0.05 Alternate Hypothesis: (H1) The series is stationary p−value<=0.05

Result:-
ADF Statistic: -3.376024
Critical Values @ 0.05: -2.93
p-value: 0.011804

From results, We can see that the p-value is 0.011, which is less than 0.05, So The series is stationary. And Reject the null hypothesis (H0)

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

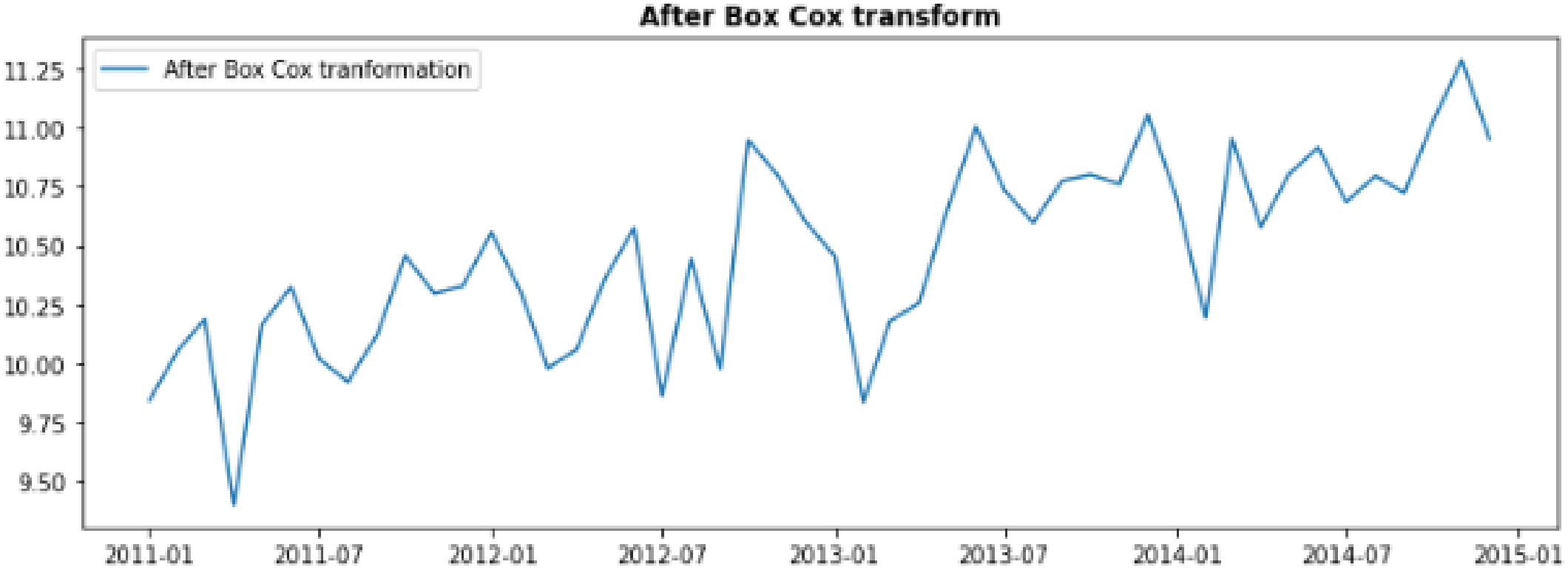
Null Hypothesis (H0) : The series is stationary p−value>0.05 Alternate Hypothesis: (H1) The series is not stationary p−value<=0.05

KPSS Statistic: 0.577076
Critical Values @ 0.05: 0.46
p-value: 0.024720

p-value is 0.024 < 0.05, Which means the series is not Stationary

Differencing Transformation

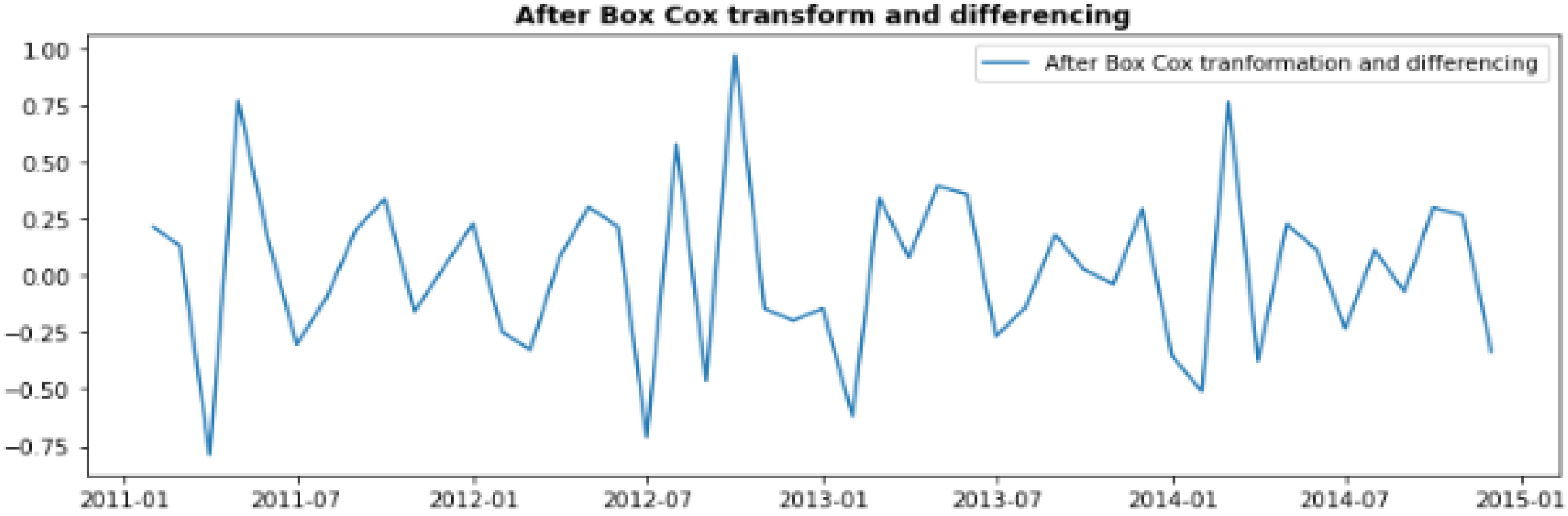
Box-Cox Transformation makes the variance constant in a Time series.



Differencing Transformation

Differencing to remove the trend ,
Differencing is performed by subtracting the previous observation from the current observation.
Differencing can remove both Trend and seasonality in a Time series.

The series looks Stationary The fluctuations are under constant limits The mean is also centered a



Thus after performing the Stationarity Tests we can see that we are able to convert a non-stationary series into a stationary series to build an Auto Regressive model.

Augmented Dickey-Fuller (ADF) test after Differencing

ADF Statistic: -4.535011
Critical Values @ 0.05: -2.95
p-value: 0.000170

p-value is 0.0001 < 0.05, Reject the null hypothesis (H0) , The series is stationary

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test after Differencing

KPSS Statistic: 0.577076
Critical Values @ 0.05: 0.46
p-value: 0.024720

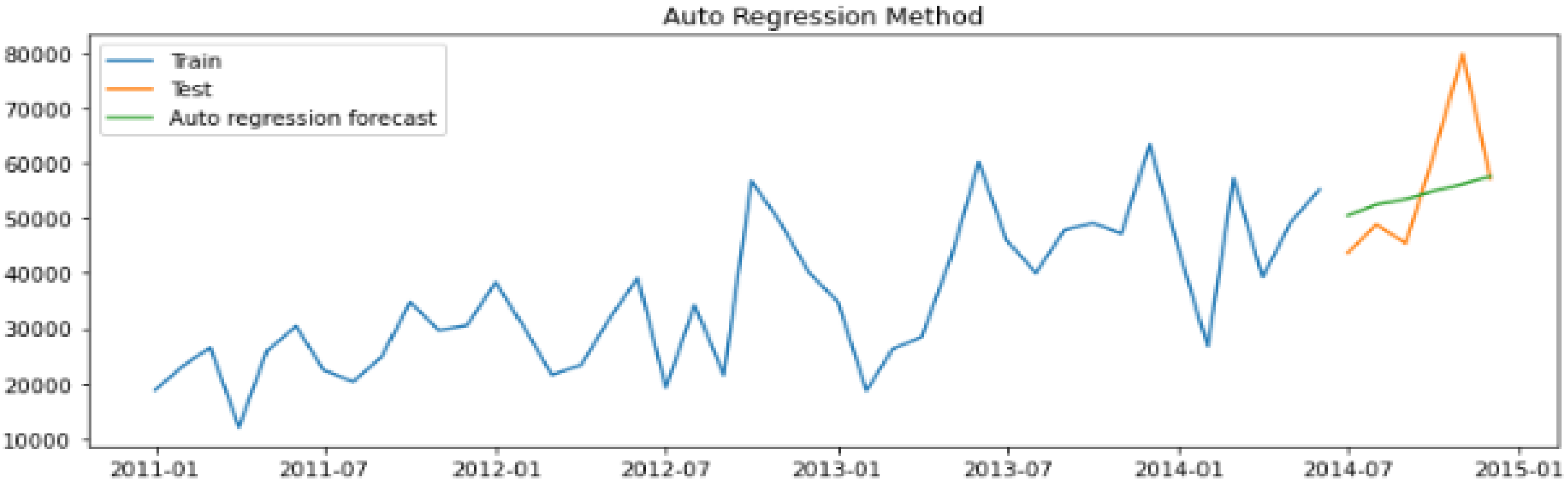
p-value is 0.100 >= 0.05, Fail to reject the null hypothesis (H0) The series is stationary

Auto regression method (AR)

It models the future observation as a linear regression of one or more past observations. We will directly use p=1, q=1, and d=1 as the forecasts are relatively better for these values.

From the plot, we can see that we can capture the trend in the forecast but could not capture the seasonality

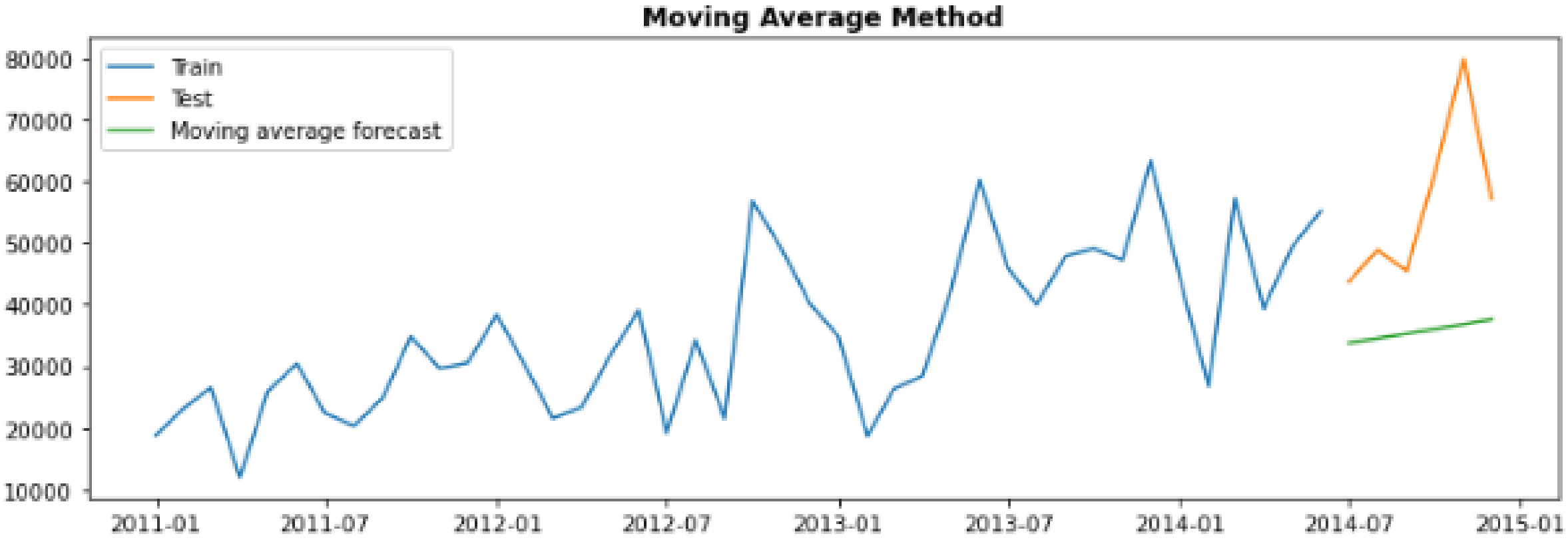
The RMSE =10985.28 and MAPE = 13.56 values are slightly high again



Moving average method (MA)

The Moving Average Model models the future forecasts using past forecast errors in a regression-like model. We are able to capture trends but not seasonality in the forecast.

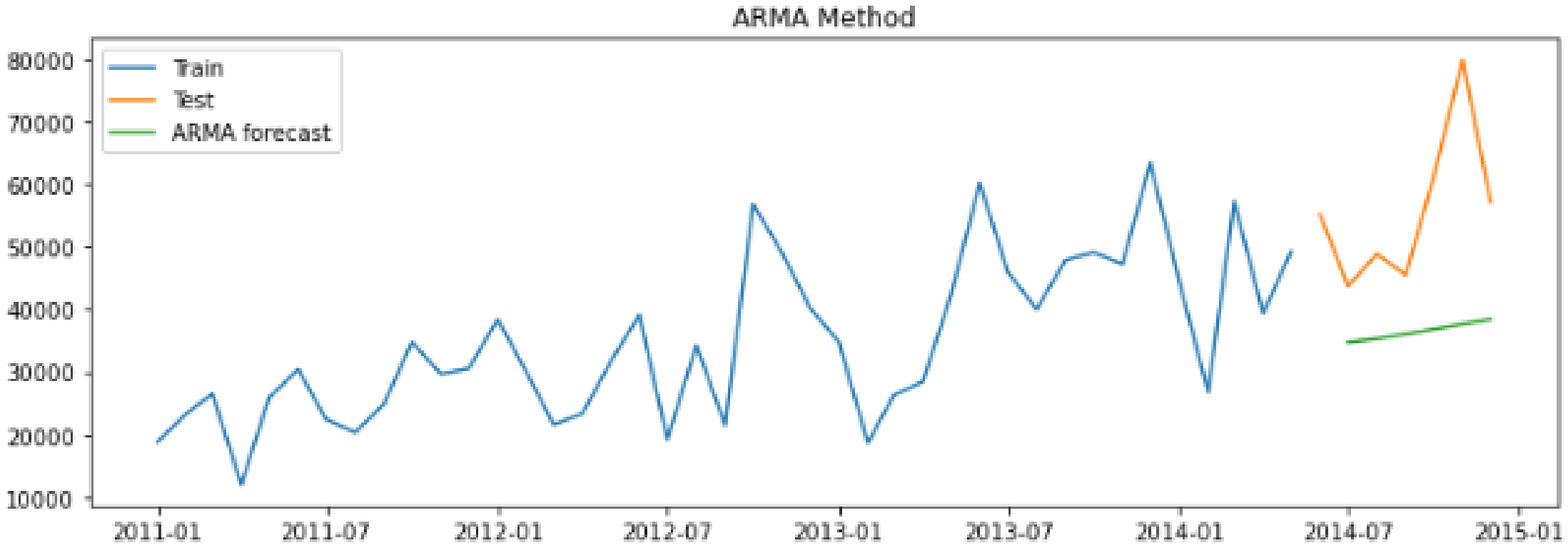
RMSE=23360.02 and MAPE=33.93



Auto regression moving average method (ARMA)

A time series that exhibits the characteristics of an AR(p) and/or MA(q) process can be modeled using an ARMA(p,q) model It models the future observation as the linear regression of one or more past observations and past forecast errors.

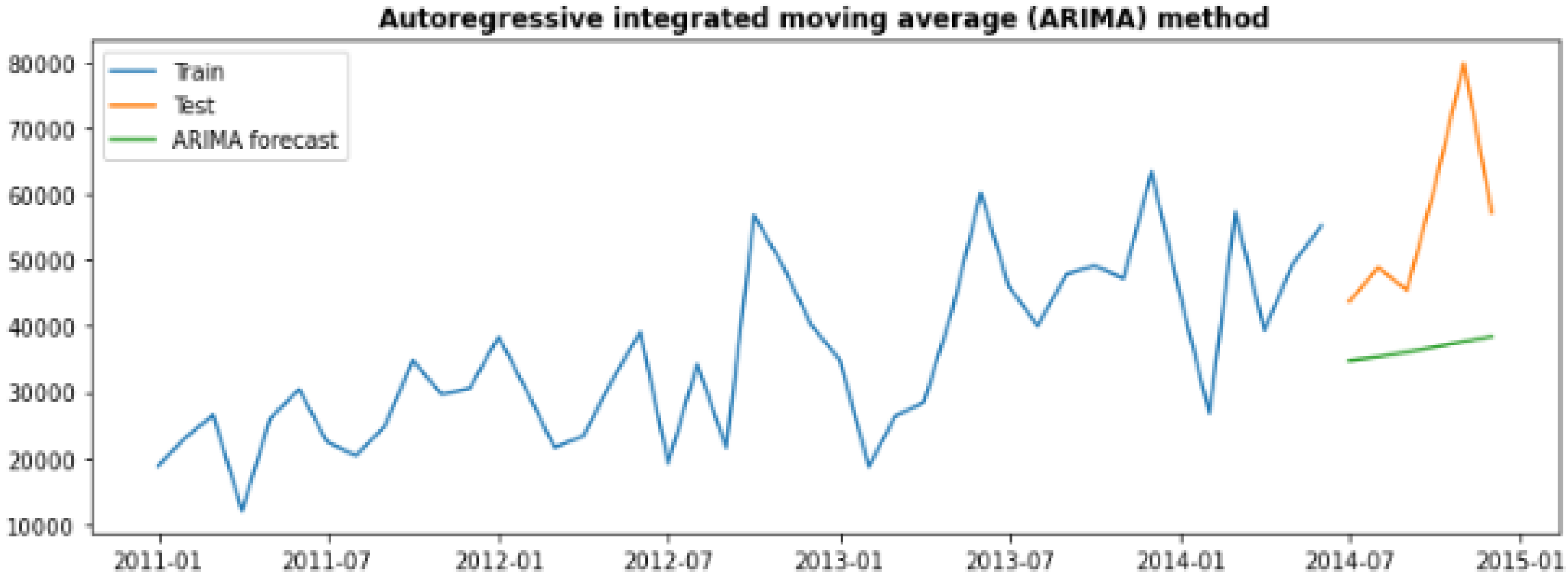
ARMA model captured Trend but no seasonality, We can see high values of MAPE = 32.40 and RMSE=22654.32



Autoregressive integrated moving average (ARIMA)

ARIMA model has three parameters p: Highest lag included in the regression model d: Degree of differencing to make the series stationary q: Number of past error terms included in the regression model ‘d’ is the differencing parameter.

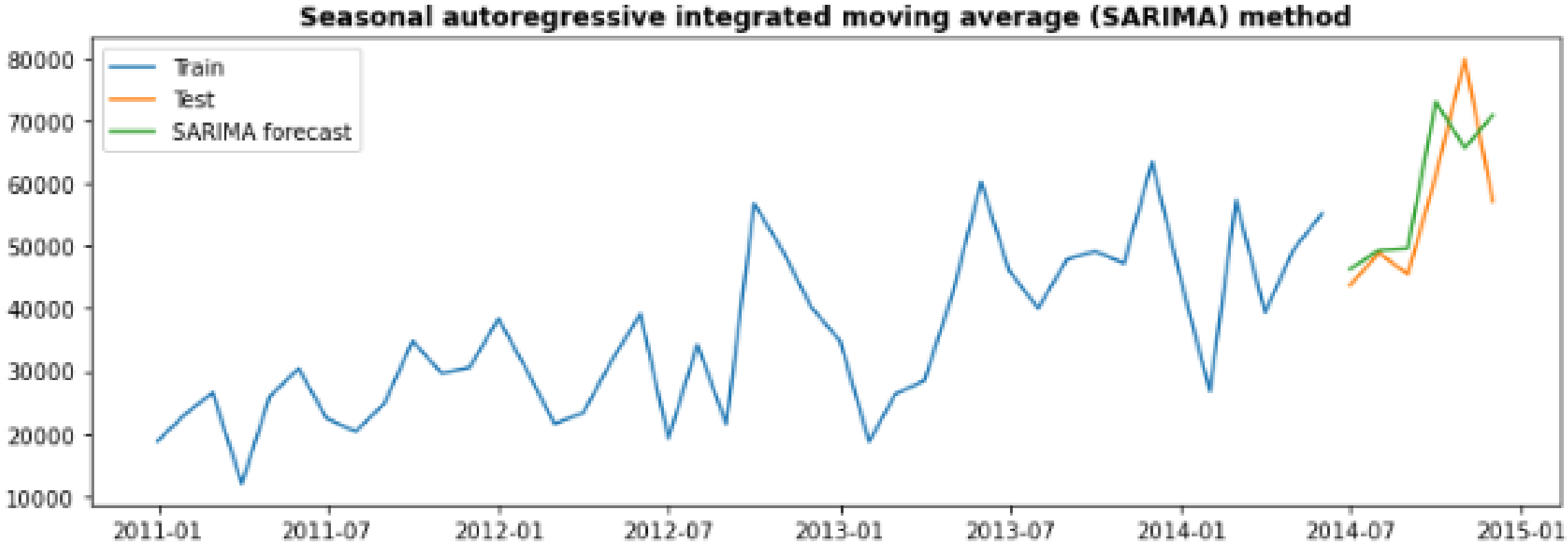
RMSE=22654.32 and MAPE=32.40



Seasonal auto regressive integrated moving average (SARIMA)

SARIMA Model has both non seasonal elements and seasonal elements. SARIMA brings all the features of an ARIMA model with an extra feature - seasonality. SARIMA has six parameters along with seasonality. The forecast captured both trend and seasonality

RMSE=9618.18 and MAPE=12.88



Model wise MAPE comparison

	Method	RMSE	MAPE
0	Naive method	12355.97	17.47
0	Simple average method	24146.06	34.34
0	Simple moving average forecast	14756.73	15.82
0	Simple exponential smoothing forecast	15011.49	15.99
0	Holt's exponential smoothing method	18976.37	34.57
0	Holt Winters' additive method	9555.61	9.33
0	Holt Winters' multiplicative method	9423.23	11.43
0	Autoregressive (AR) method	10985.28	13.56
0	Moving Average (MA) method	23360.02	33.93
0	Autoregressive moving average (ARMA) method	22654.32	32.40
0	Autoregressive integrated moving average (ARIM...	22654.32	32.40
0	(SARIMA) Seasonal autoregressive integrated mo...	9618.18	12.88

Among all the methods done in the ARIMA above, we can conclude that forecast done using the SARIMA method is able to predict the sales closer to the actual values

RMSE and MAPE values for this method are the least among all the methods done

Conclusions

Thus we can conclude that **the Holt-Winters additive method is the best forecasting method in the smoothing technique And SARIMA - Seasonal Autoregressive Integrated moving average is the best method in the ARIMA set of techniques.**

Method	RMSE	MAPE
Holt Winters' additive method	9555.61	9.33
(SARIMA) Seasonal autoregressive integrated mo...	9618.18	12.88

Thank You

Yathestha Siddh

IIIT-B Roll Number - EDS21090334
Batch:- upGrad & IIITB | Data Science Program - November 2021
Contact details :-
Email:- yathestha@gmail.com /yathesthasiddh.eds36@iiitb.net
Phone:- 7506933629
LinkedIn:- https://www.linkedin.com/in/yathestha-siddh-