RETAIL GIANT SALES FORECAST

Shubham Kokate

srkokate.297@gmail.com

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.

DATA DESCRIPTION

- Data has 5 attributes
- There are no missing values

```
#   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
```

Store caters to 7 different geographical market segments and 3 major customer segments

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

ANALYSIS STEPS

Data Preparation

- Data Preparation involves converting column into date, aggregate data on month basis, creating market segment from existing columns
- Splitting data into Train and Test

Profitable Market-Segment

- We need to use Coefficient of Variance (CoV) to identify most profitable market-segment
- The market-segment with least CoV value is the most consistently profitable.

Time Series Decomposition

- Analyse the trend, seasonality and noise component of the time series data.

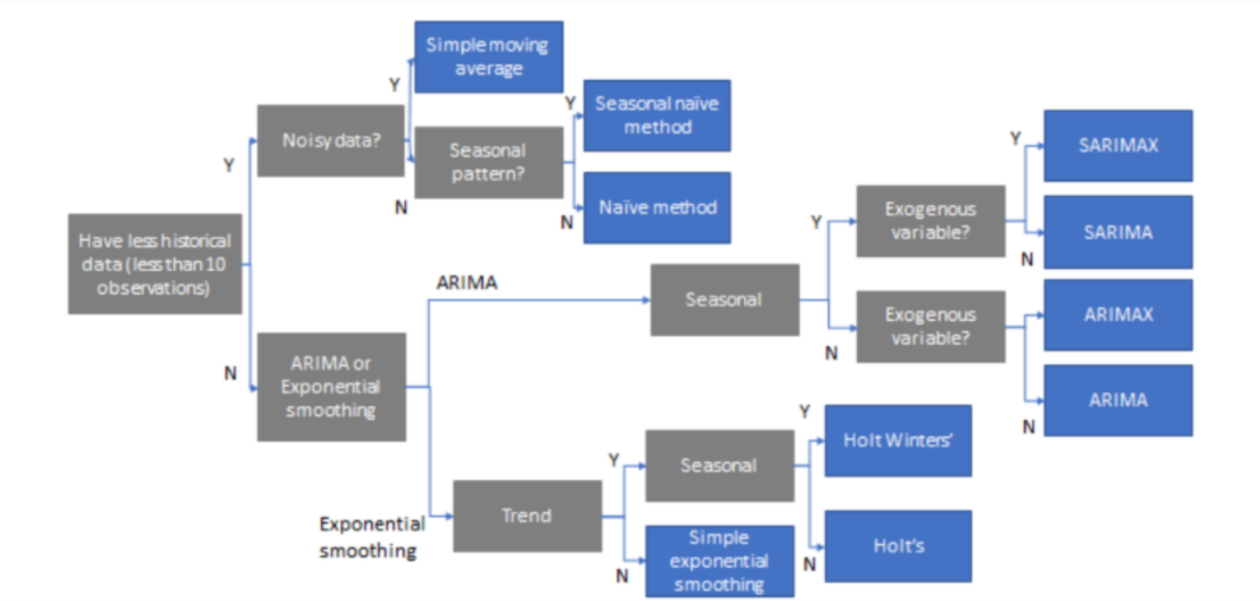
Exponential Smoothing

- Exponential forecast are equal to a weighted average of past observations and the corresponding weights decrease exponentially as we go back in time

ARIMA Model

- The AR and MA models capture the level and trend.
- SARIMA model captures the level , trend and seasonality

CHOOSING TIME SERIES METHOD



1. We have more than 10 records
2. In the Exponential Method:
   There is trend and Seasonality present thus we can see that Holt Winter's Method will perform the best.
3. In the ARIMA Method:
   There is Seasonality and No Exogenous variable present so SARIMA will perform the best

DATA PREPARATION

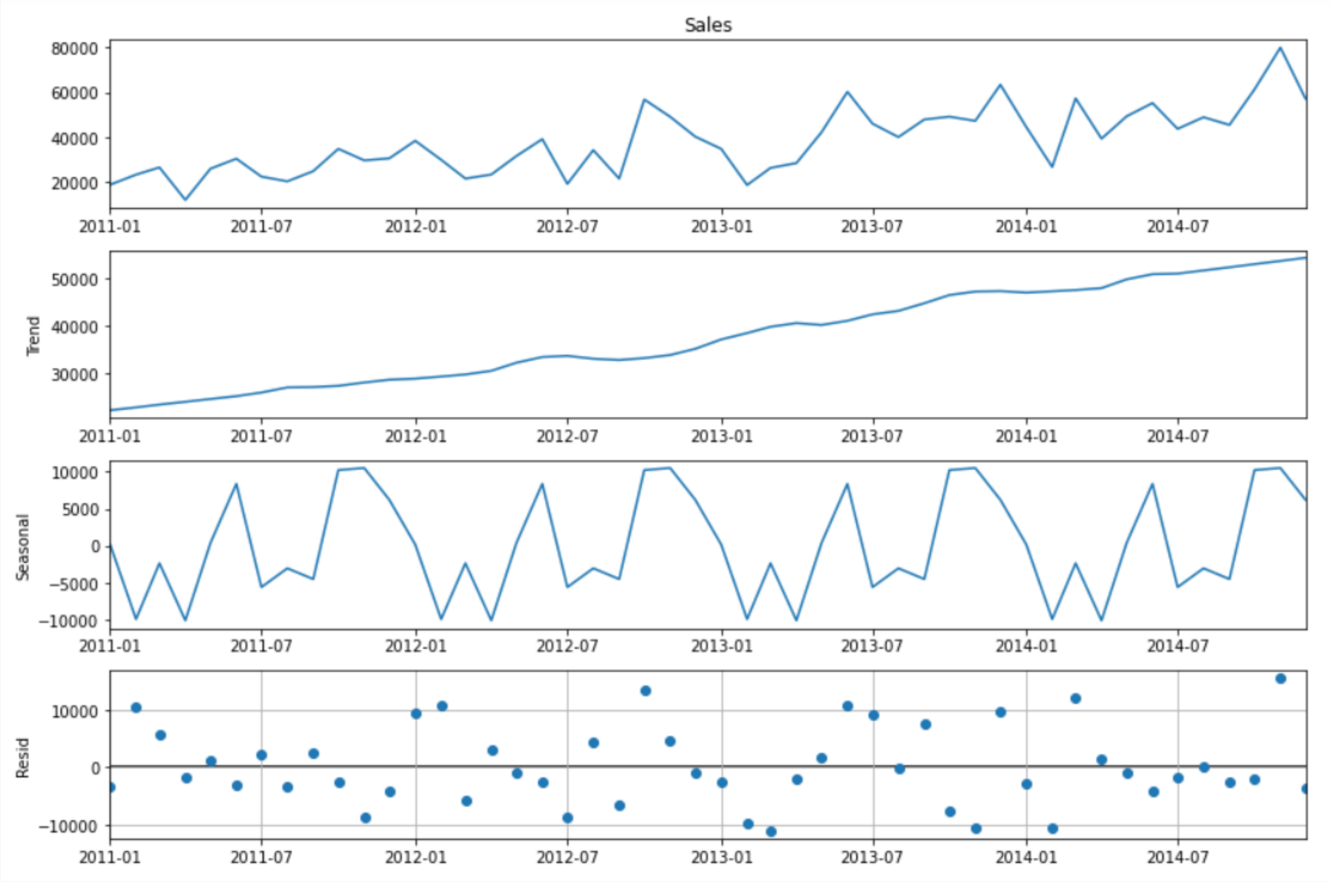
1. Convert Order Date to Year-Month
2. Create Market_Segment columns from Market and Segment columns
3. Aggregate profit data by month
4. Split aggregated data into train and test
5. Use train data to find the CoV value of all the market-segments
6. Filter the data for the market-segment with the least CoV as it is the most profitable market-segment

	Market_Segment	cov
0	APAC_Consumer	0.522725
1	APAC_Corporate	0.530051
12	EU_Consumer	0.595215
15	LATAM_Consumer	0.683770
13	EU_Corporate	0.722076
16	LATAM_Corporate	0.882177
14	EU_Home Office	0.938072
2	APAC_Home Office	1.008219
18	US_Consumer	1.010530
19	US_Corporate	1.071829
20	US_Home Office	1.124030
17	LATAM_Home Office	1.169693
6	Canada_Consumer	1.250315
3	Africa_Consumer	1.310351
7	Canada_Corporate	1.786025
4	Africa_Corporate	1.891744
5	Africa_Home Office	2.012937
8	Canada_Home Office	2.369695
9	EMEA_Consumer	2.652495
10	EMEA_Corporate	6.355024
11	EMEA_Home Office	7.732073

MOST PROFITABLE MARKET SEGMENT

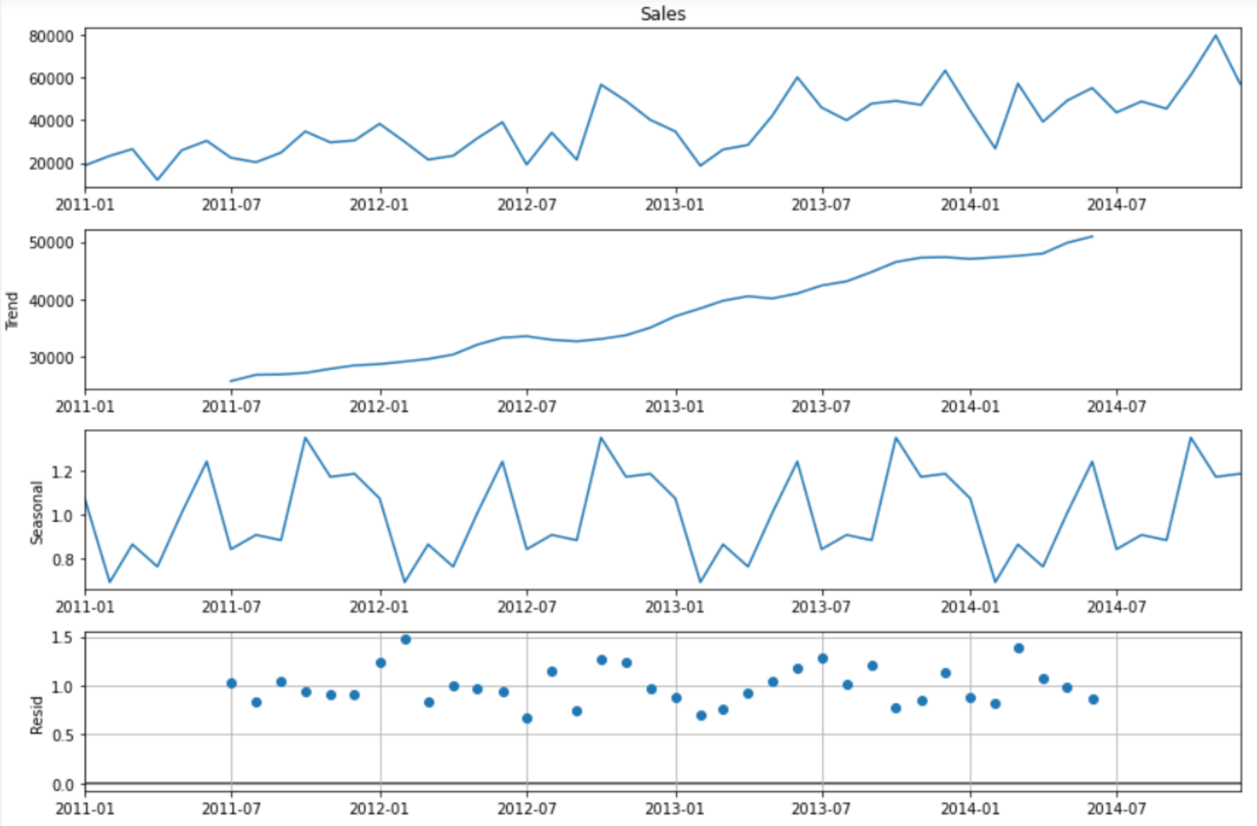
- We can use Coefficient of Variance (CoV) to identify the most profitable market-segment
- We checked the CoV value of the 21 market-segments.
- After performing CoV analysis we can conclude that **APAC_Consumer** is the most consistently performing market-segment

TIME SERIES DECOMPOSITION - ADDITIVE



- Additive decomposition argues that time series data is a function of the sum of its components trend-cycle component, seasonal component, and the remainder.
- There is a visible trend although not completely linear.
- We can observe presence of seasonality in the data.
- There is no visible pattern in Residual graph.

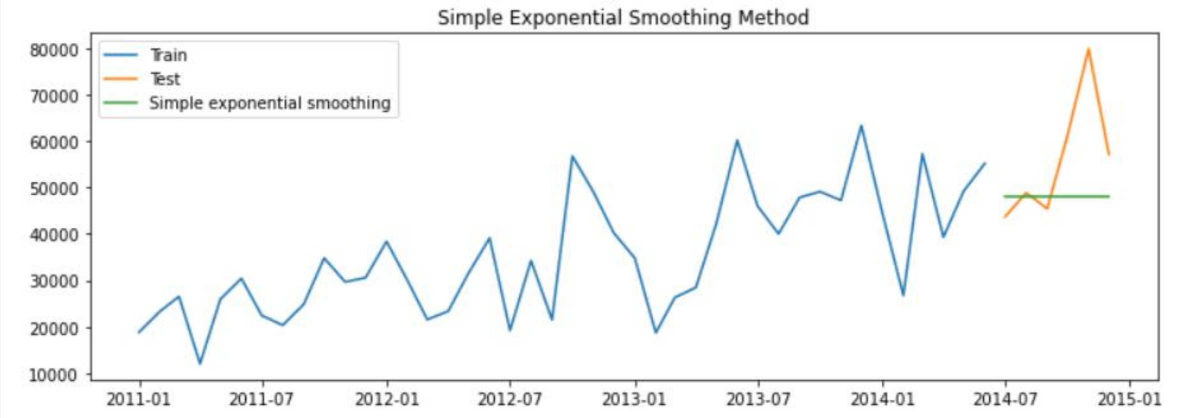
TIME SERIES DECOMPOSITION - MULTIPLICATIVE



- Additive decomposition argues that time series data is a function of the product of its components trend-cycle component, seasonal component, and the remainder.
- There is a visible trend although not completely linear
- We can observe presence of seasonality in the data
- There is no visible pattern in Residual graph

EXPONENTIAL MODELS

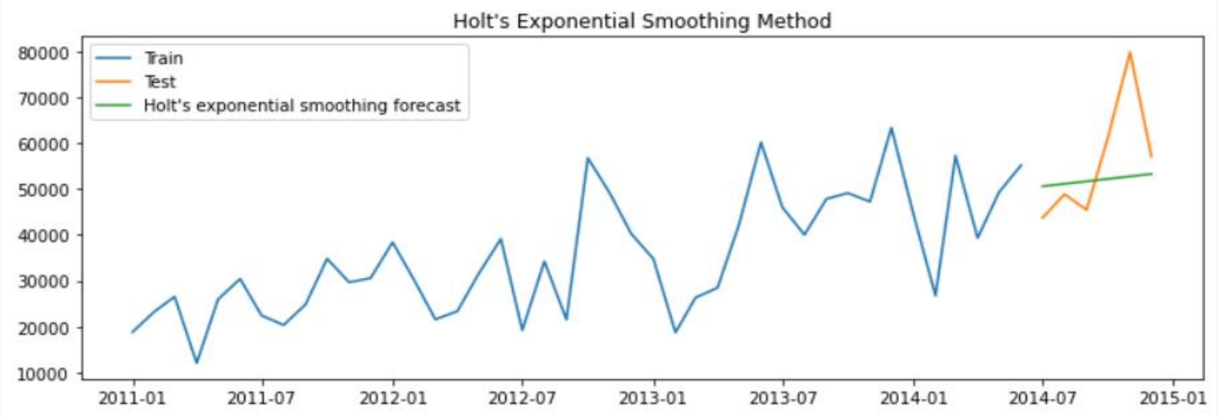
SIMPLE EXPONENTIAL SMOOTHING



RMSE	MAPE
14627.34	15.74

- Using the Simple Exponential Model we can only capture the level of the test data. Trend and Seasonality will not be captured by Simple Exponential Model.
- This will cause maximum errors in the forecast

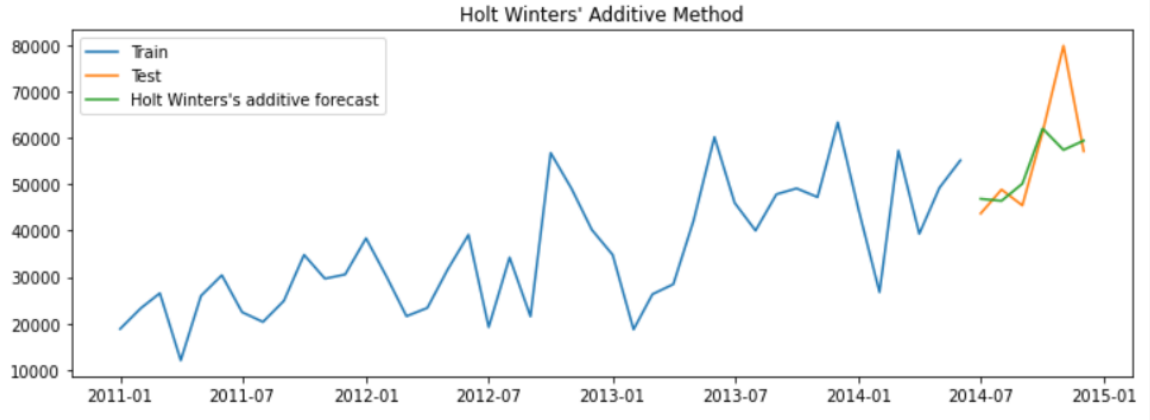
HOLT'S EXPONENTIAL



RMSE	MAPE
12403.84	14.93

- Using Holt's exponential model we will be able to caputre the level and the trend but not the seaonality for the test data.
- This will reduce the errors in the forecast compared to Simple Exponential Smoothing

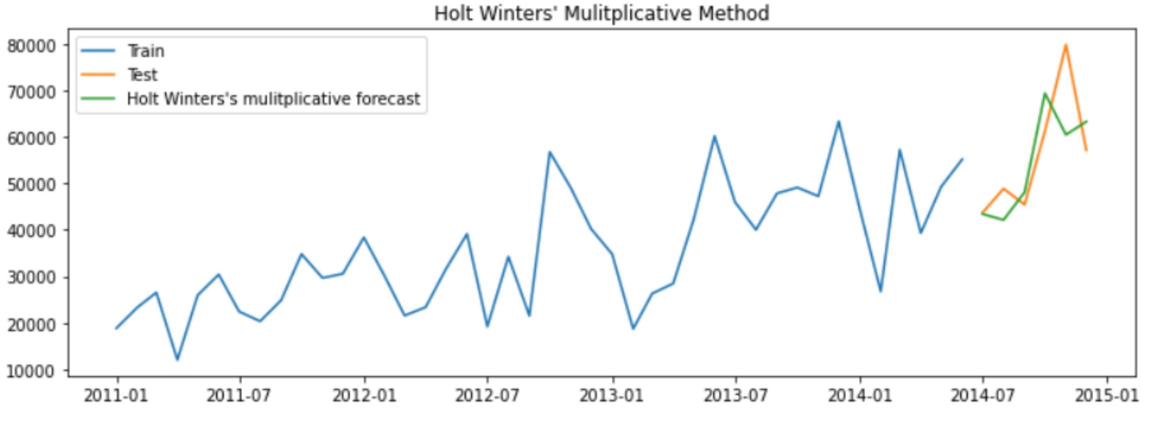
HOLT WINTER'S ADDITIVE



RMSE	MAPE
9555.63	9.33

Holt Winter's Additive Method captures the Level, the Trend and the Seasonality component of the Test data thus reducing the errors in the forecast

HOLT WINTER'S MULTIPLICATIVE



RMSE	MAPE
9423.23	11.43

Holt Winter's Multiplicative Method captures the Level, the Trend and the Seasonality component of the Test data thus reducing the errors in the forecast

TEST FOR STATIONARITY

There are two tests for stationarity

1. **Augment Dickey-Fuller (ADF) test :**

```
ADF Statistic: -3.376024
Critical Values @ 0.05: -2.93
p-value: 0.012
```

The series is stationary as p-value is less than 0.05.

2. **Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test**

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

The series is not stationary as p-value is less than 0.05.

TYPES OF STATIONARITY

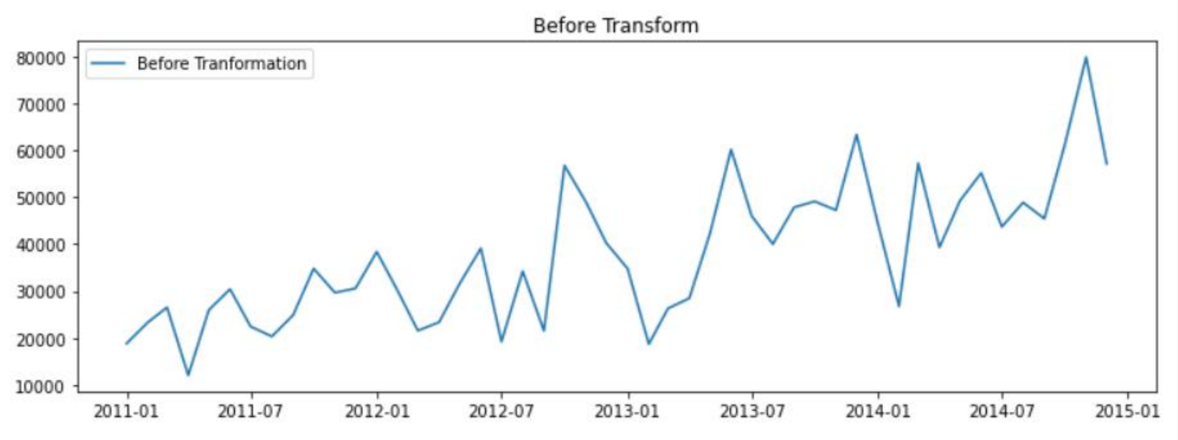
There are three types of Stationarity:

1. Strictly Stationary
2. Trend Stationary
3. Difference Stationary

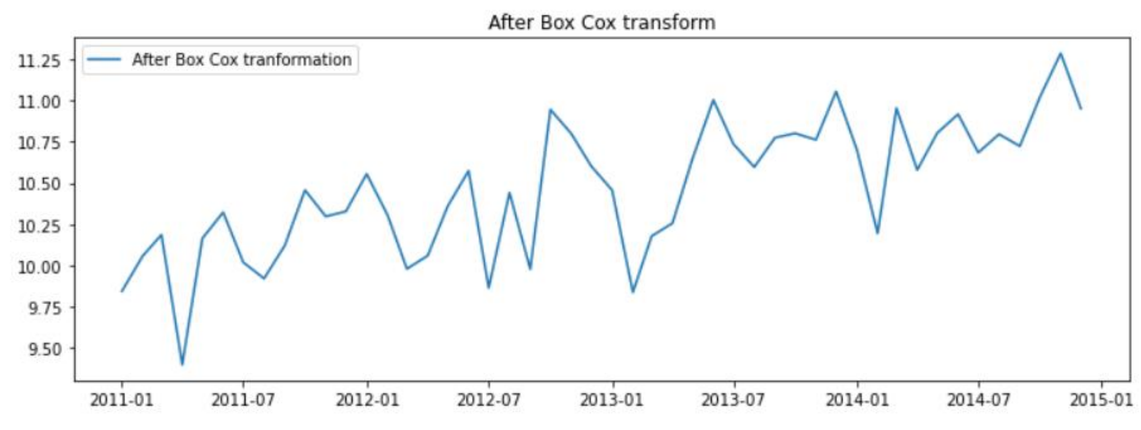
When :

1. Both tests conclude that the series is not stationary -> series is not stationary
2. Both tests conclude that the series is stationary -> series is stationary
3. ADF - not stationary and KPSS - stationary -> trend stationary, remove the trend to make series strict stationary
4. ADF - stationary and KPSS - not stationary -> difference stationary, use differencing to make series strict stationary

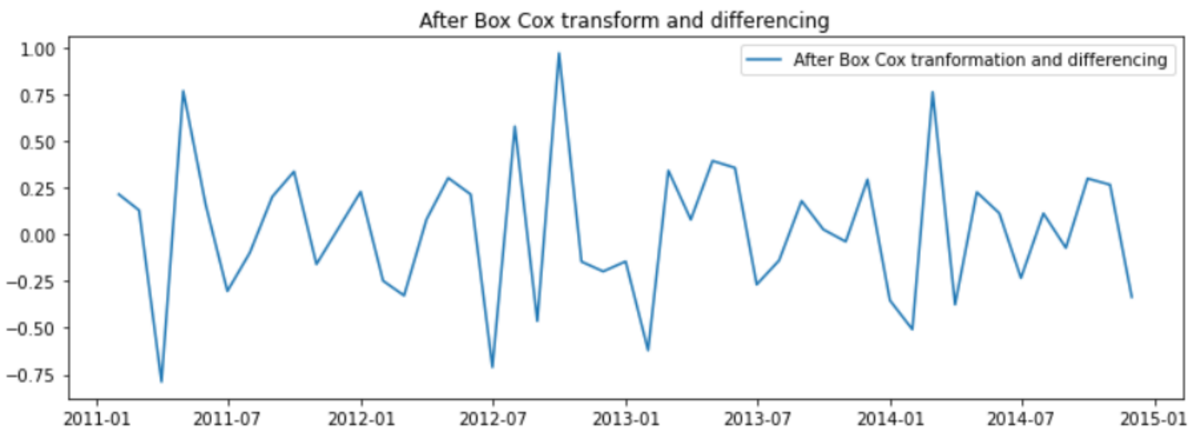
TRANSFORMATION



Before any transformation



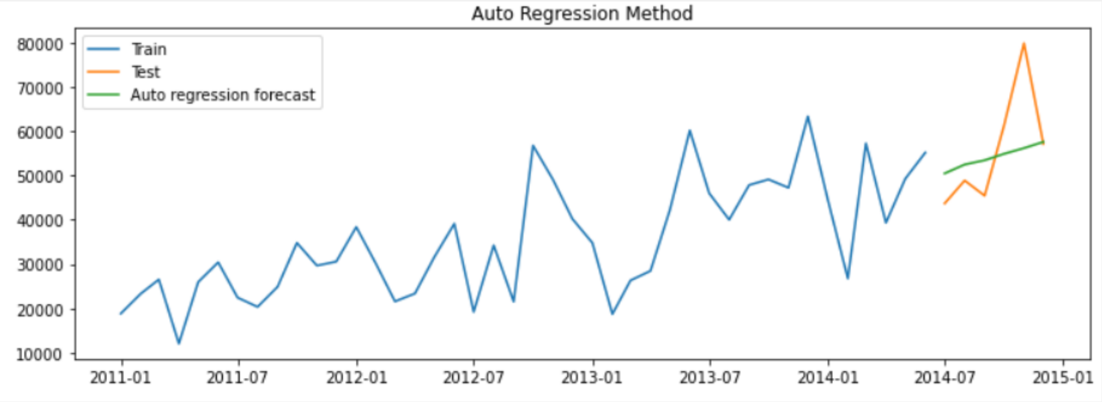
After Box Cox transformation



After Box Cox and Differencing

ARIMA MODELS

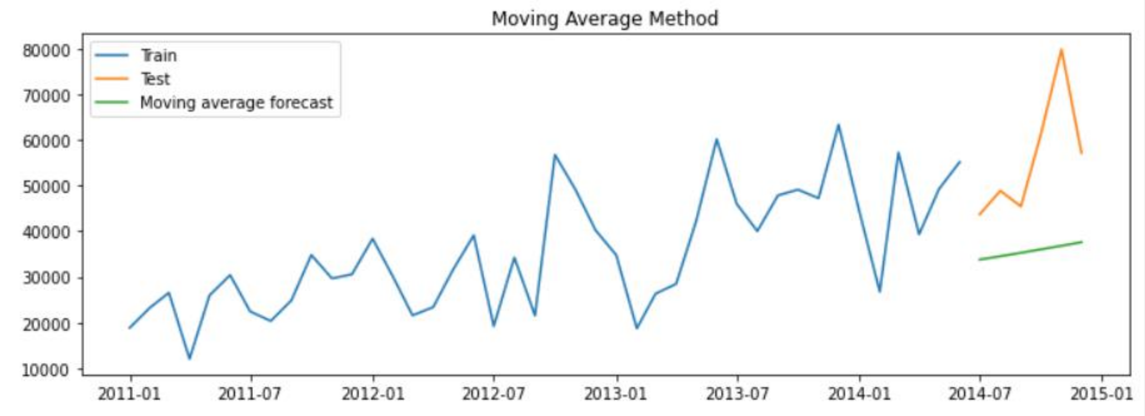
AUTO REGRESSION



RMSE	MAPE
10985.28	13.56

Using the AR component of ARIMA model we can only capture the trend and level of the data

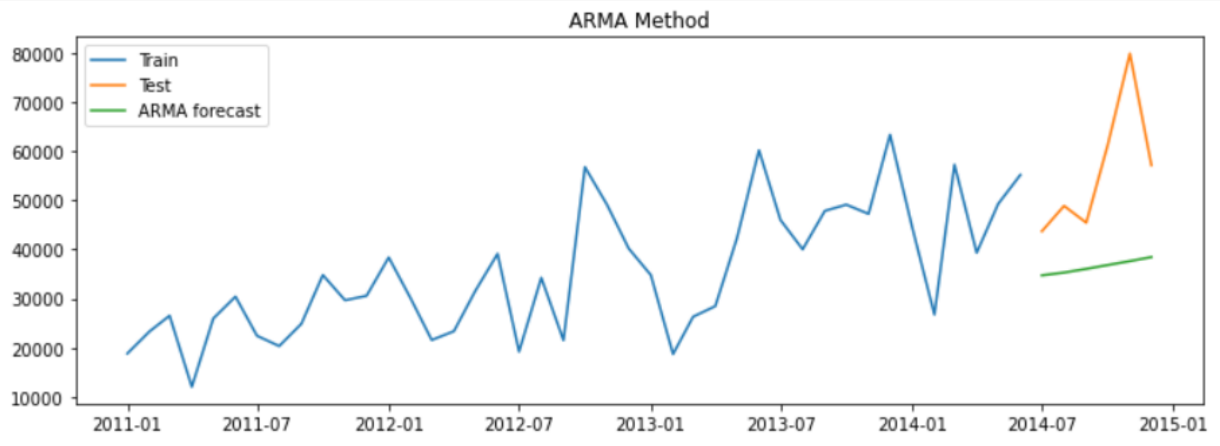
MOVING AVERAGE



RMSE	MAPE
23360.02	33.93

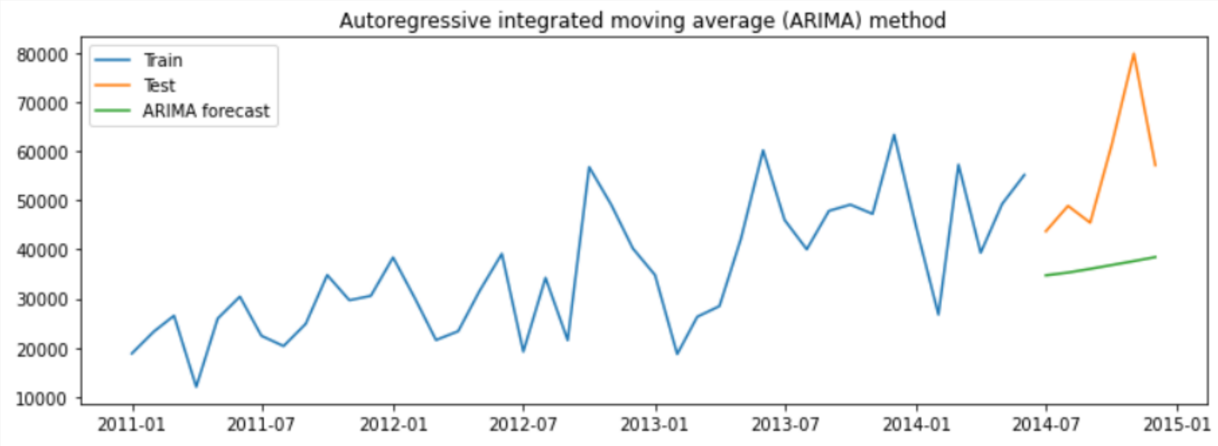
Using the MA component of ARIMA model we can only capture the trend and level of the data

ARMA



RMSE	MAPE
22654.32	32.40

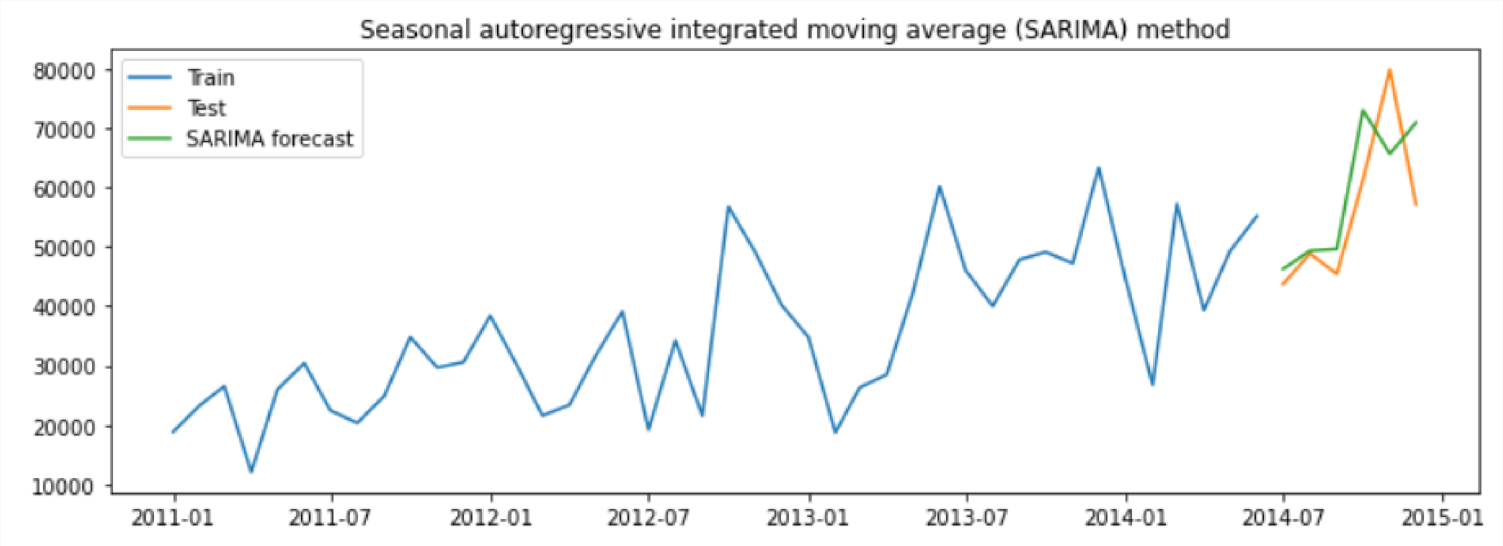
ARIMA



RMSE	MAPE
22654.32	32.40

ARMA and ARIMA have the same error values. The only difference between these methods is that ARIMA method doesn't require transformed data whereas ARMA does

SARIMA



RMSE	MAPE
9616.66	12.87

- SARIMA is the best performing model as it captures Seasonality as well.
- It uses the SARIMAX model

CONCLUSION

- We observed that the given data is timeseries data as it contains date component
- According to CoV, **APAC_Consumer** is the most consistently profitable market segment
- After decomposing data we could see an upward trend although not linear
- There is a week seasonality component in the data
- The data is difference stationary
- Among the Exponential models:
  - **Simple Exponential Smoothing** has maximum errors (Only captures Level)
  - **Holt Winter's Additive** method has the least errors (Captures Level, Trend and Seasonality)
- Among the Auto Regression Models:
  - **Moving Average** has maximum errors (Captures only Level and Trend)
  - **SARIMA** has least errors (Captures Level, Trend and Seasonality)