



Giant_Super_Store_Sales Forecasting

By Jyoti Shukla



Objective: To Forecast Sales and Demand For Next 6 Months

Data Highlights:

- Sales transaction data is available for four(4) years
- Each data point is a transaction
- Attributes of Interest:
 - Segment
 - Market
 - Quantity sold
 - Value of the sale
 - Profit made on the sale
- We have multiple markets and segments so we would focus on only top 5
 - Criteria for selection
 - Maximum profit
 - Consistent profit (based on profit %age month on month)



Helping to manage the revenue and inventory of the Giant retail store.





Business Understanding:

- An online supermarket with a worldwide presence that deals with product orders across categories like consumer, corporate and home office.
- Store caters to products across 7 different segments in 3 major categories

Business Objective:

- Identify two most profitable (and consistent) Market S egments
- Forecast sales and demand for next 6 months in the identified two segments
- Helping to manage the revenue and inventory accordingly

Success criteria:

- Forecast the sales and the demand for the next 6 months, that would help Giant Super Store to manage the revenue and inventory accordingly.





Data Understanding:

- The data received for the analysis had 24 attributes. The important ones were :

Attribute	Description
Order Date	Date on which the order was placed
Segment	The market segment to which the customer belongs
Market	Geographical market sector to which the customer belongs
Sales	Total sales value of the transaction
Quantity	Quantity of the product ordered
Profit	Profit made on the transaction





Data Understanding – Segmentation of Data

- The “Market” attribute has 7-Geographical market sector that the customer belongs to.
- The “Segment” attribute tells the 3 segments that customer belongs to.

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

Segment
Consumer
Corporate
Home Office

- Obtained data from
- <https://www.kaggle.com/datasets>





Data Understanding - Data Cleaning Methods Used

- Checking for missing column names: No Column name is missing
- Check for duplicate rows: No rows are duplicate in the dataset
- Looking for Missing values
- Check for cells with Null values: Only “Postal Code” column contains null values
- Removing the Row Id column from the dataset as it is not required for analysis
- Checking for blanks"": No blanks
- Checking for Lower and uppercase issues: No upper and lowercase issues
- Checking for columns with 1 unique value: No columns with 1 unique value
- Changing Order Date & Ship Date to default R date format: Ex: 03-01-2011 to 2011-01-03





Data Preparation:

- The entire customer population of “Giant_Super_Store” can be divided into 21(7*3) market segments, such as APAC Consumer, US Home Office, EU Corporate etc.
 - Combining 2 columns (Market, Segment) to create a new column Market_Segment in the dataset
- So, segmenting the whole dataset into the 21 subsets based on the market and the customer segment level
- Extract the Month-Year from the Order.Date and create a new column Year.Month for the aggregation of data
- Extract the Year from the Order.Date and create a new column Year for the aggregation of data
- Calculate Coefficient of Variation(CV) and Pick Top 2 Segments based on CV & Profit
- Aggregating data for sales, profit and quantity for all dataset year wise





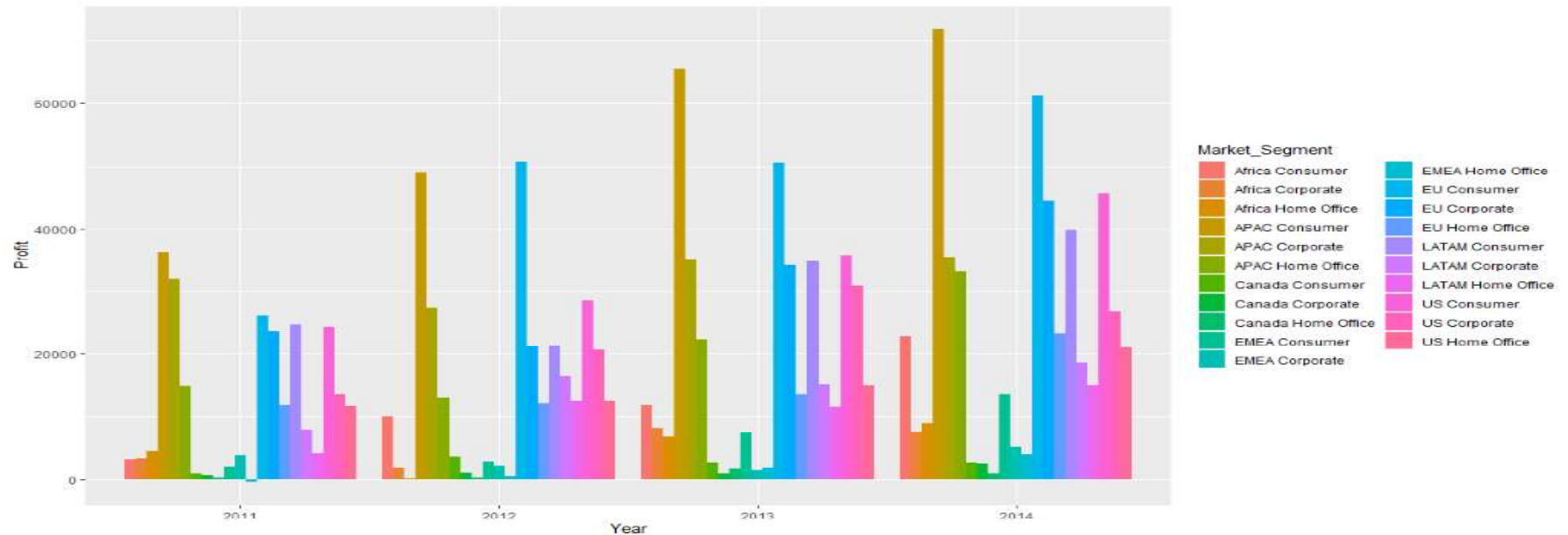
Data Preparation & EDA

- Created 21 data subset buckets based on Market & Segment they belong.
- Aggregated data in each bucket by Sales, Quantity & Profit.
- Calculated Coefficient of Variation(CV) using aggregated Profit for each Market-Segment using below: $CV = \text{sd}(\text{Profit}) * 100 / \text{mean}(\text{Profit})$
- Using CV & Profit found Top 2 most profitable Market-Segments as APAC_Consumer & EU_Consumer with below values:

Market	Segment	Sales	Profit	CV
APAC	Consumer	1816753.7	222817.56	420.6702
EU	Consumer	1529716.24	188687.707	471.8084

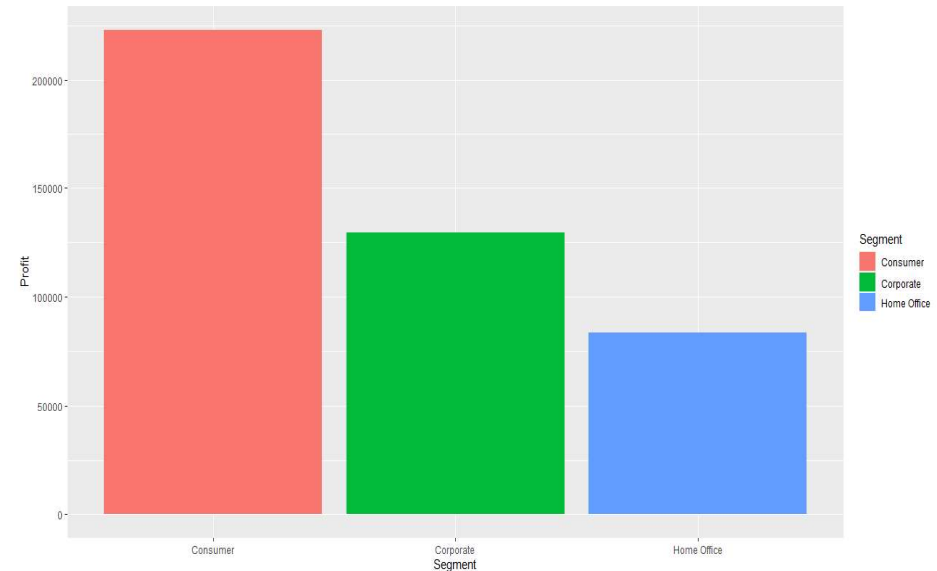
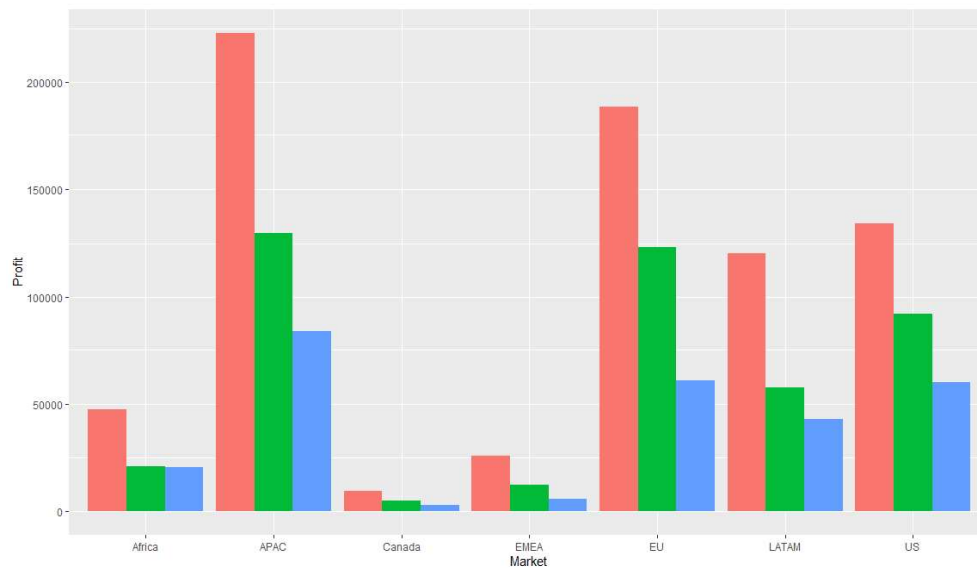


Data Preparation: Individual Profit & Consistent Market Segment View year wise



- So, in all the 4 years of data the Market: 'APAC' and 'EU' in the customer segment level: 'Consumer' are the 2 most profitable and consistently profitable segments



Data Preparation: Plotted Graphs between Market, Segment and Profit



- On plotting Market segments with respect to year and profit found that Canada doesn't have full data so ruled out Canada
- Consumer segment has highest profit across three segments
- Next need to Forecasting Sales and Demand – APAC consumer Sales and EU consumer Sales



Modeling:

- Created time series for aggregated data of APAC_Consumer and EU_Consumer subsets for first 48 months:
`ts(APAC_Consumer_Agg$Sales,frequency=12,start=c(2011,1),end=c(2014,12))`
`ts(EU_Consumer_Agg$Sales,frequency=12,start=c(2011,1),end=c(2014,12))`
 - Smoothened time series using Moving Average method, also tested Holt Winters smoothing.
 - Time series data was divided into train(1-42 month), validation(43-48 month) & test sets(49-54 month).
 - Model 1: Linear model : Created using `tslm()` function from forecast package in R on train data.
 - Model 2: Auto Arima Model: Created using `auto.arima()` function from forecast package in R train data
 - Both models were evaluated using Mean Absolute Percentage Error(MAPE).
 - ACF plots of residuals were used check that it resembles white noise.
- 
- 

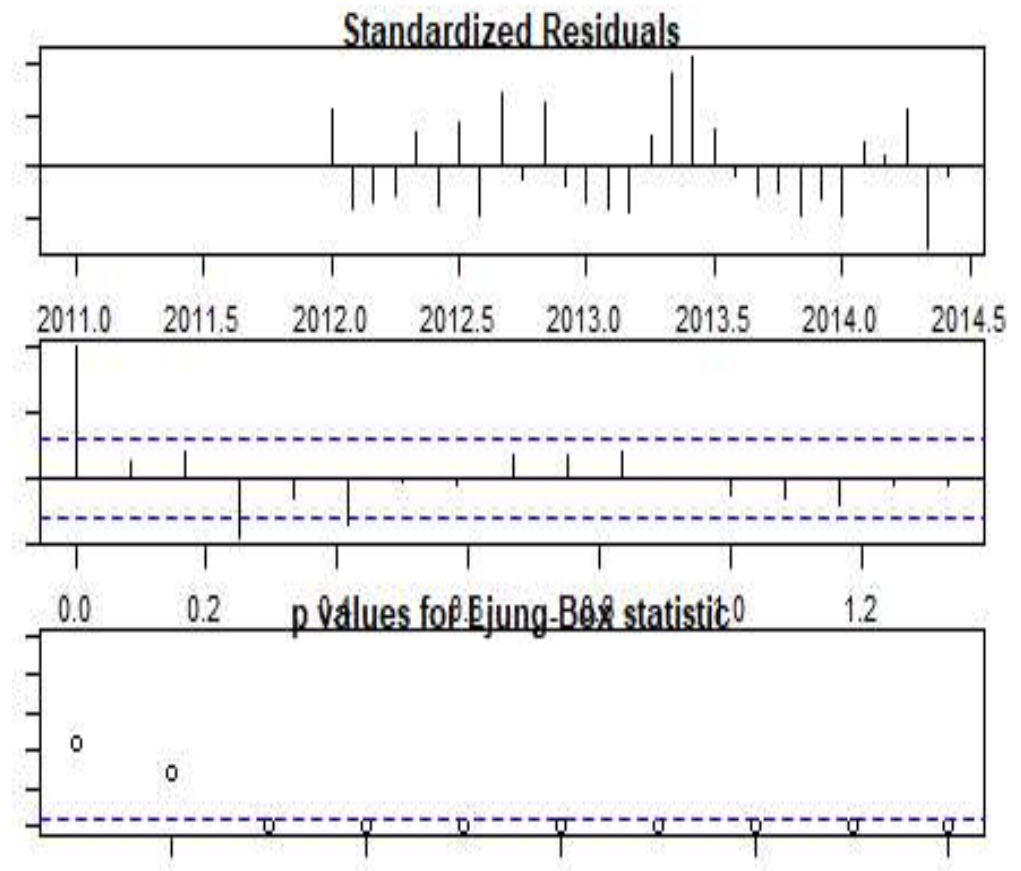
Analysis: Model Evaluation

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	7450.556	10395.58	7450.556	10.2114	10.2114	0.3845775	1.135307

- a)MAPE and other measures for APAC Consumer time series Linear Model

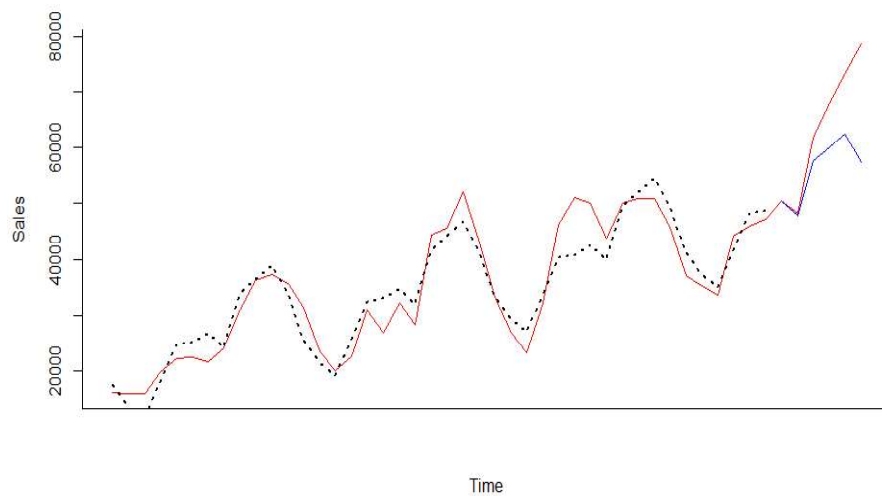
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	9.31539	3626.009	2674.287	-0.8725545	7.351146	0.3253097	0.1151259	NA
Test set	7137.96351	10283.426	7313.76	9.768693	10.116737	0.8896714	0.3134169	1.111616

- b)MAPE and other measures for APAC Consumer time series Auto ARIMA Model

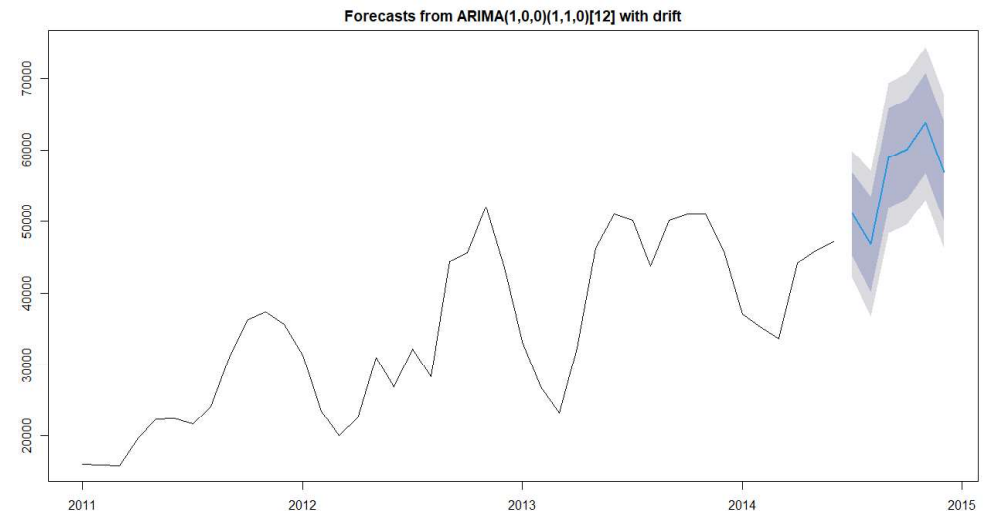




Result 1a. APAC Consumer Sales Forecast on validation set



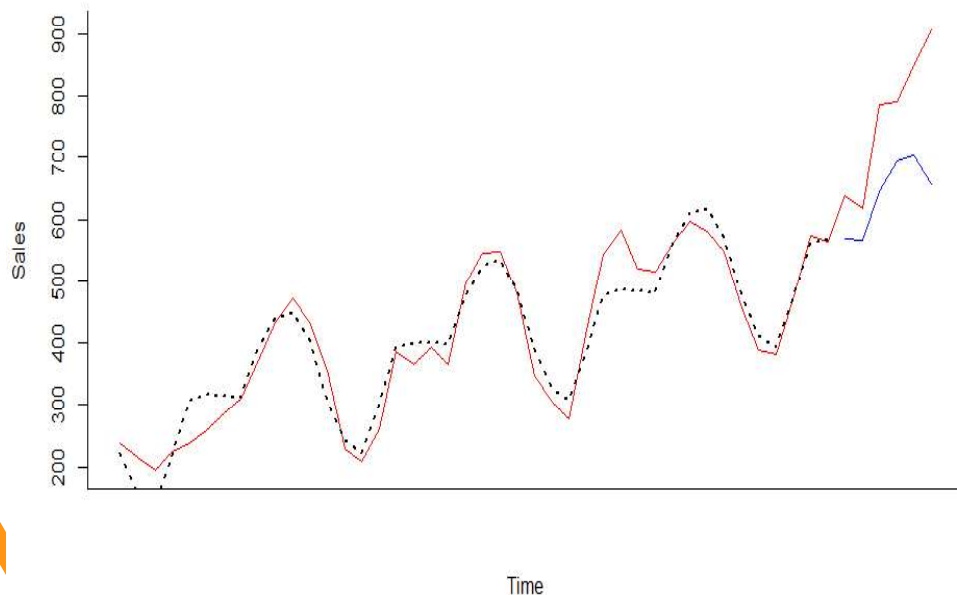
a) Linear Model Forecast



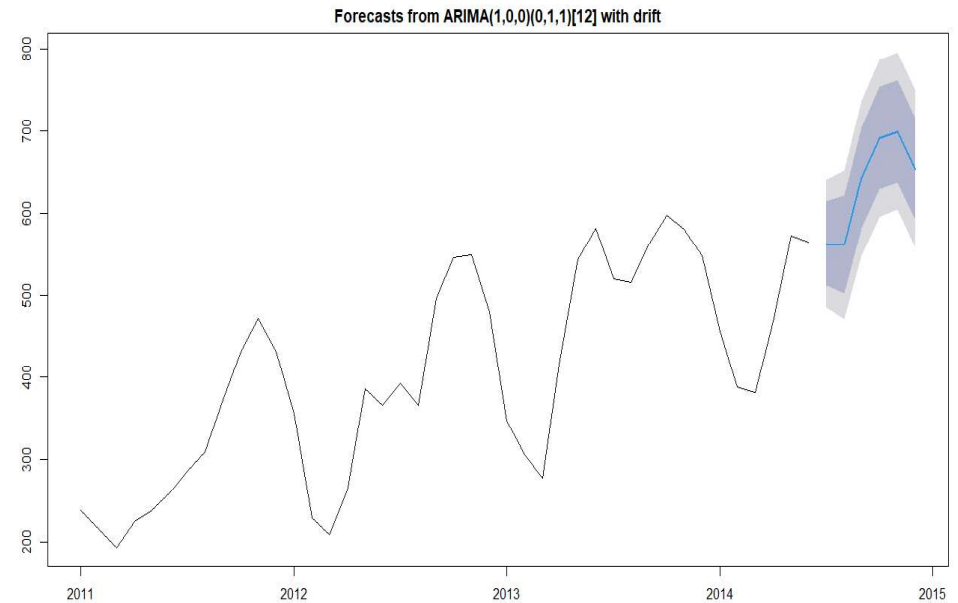
b) Auto ARIMA Model Forecast

- Time Series of APAC Sales for the 48 months shows an increasing trend and it also shows the Sales forecasting for the next 6 months
- Classical Decomposition method shows that there is a seasonality trend present in the time series with a forecast of drop in Sales in the next 6 months and is expected to increase after 6 months

Result 1b. APAC Consumer Quantity Forecast on validation set



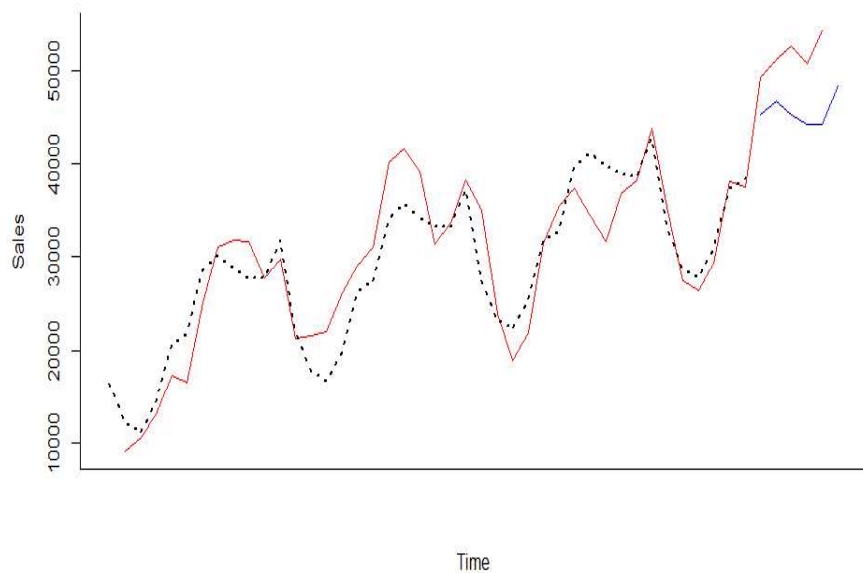
a) Linear Model Forecast



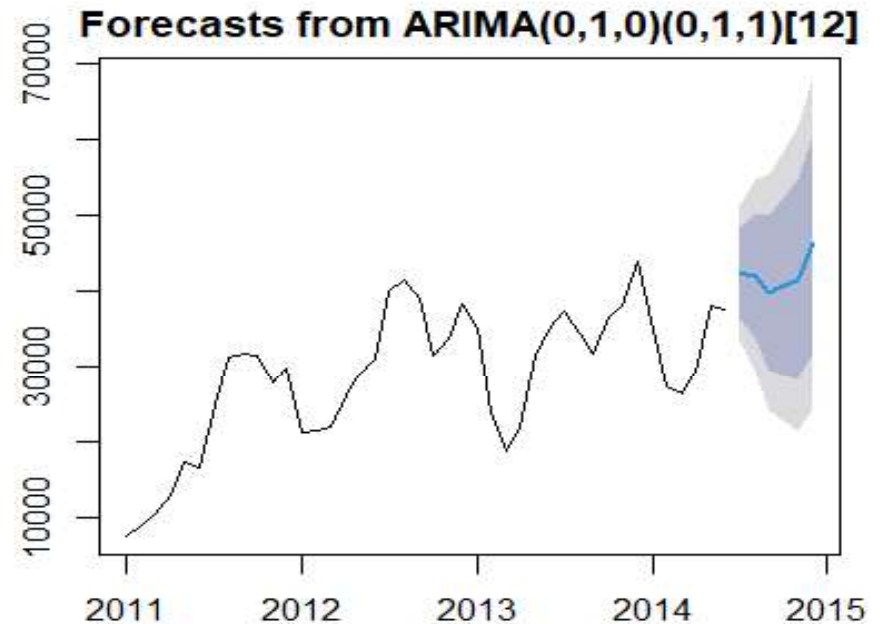
b) Auto ARIMA Model Forecast

- Time Series of APAC Quantity for the 48 months shows an increasing trend and it also shows the Quantity forecasting for the next 6months.
- Classical Decomposition method shows that there is a upward trend in the time series with a forecast of initial decrease and then increase in Quantity in the next 6months

Result 2a. EU Consumer Sales Forecast on validation set



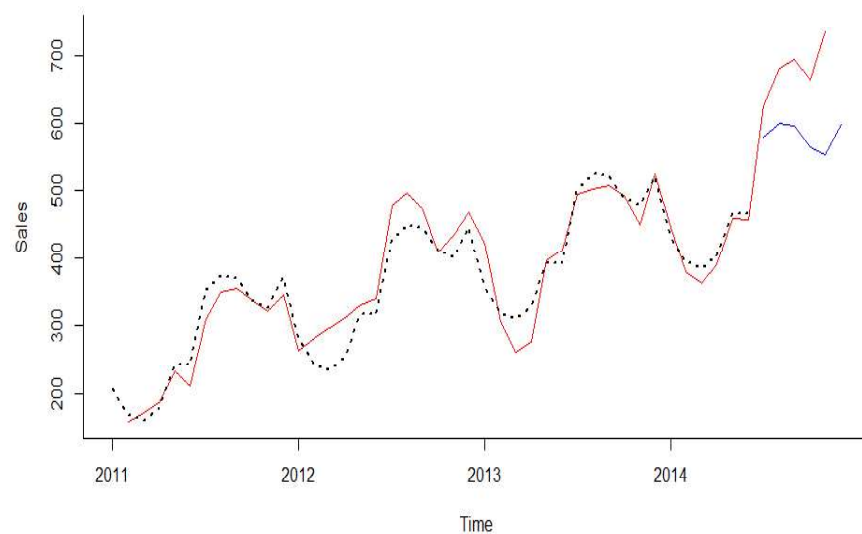
a) Linear Model Forecast



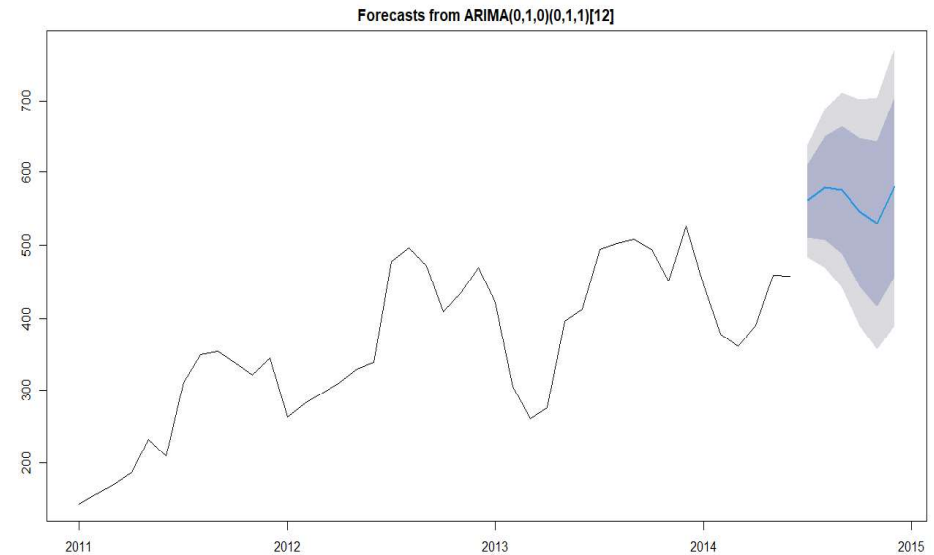
b) Auto ARIMA Model Forecast

- Time Series of Europe Sales for the 48 months indicates it is stabilizing and it also shows the Sales forecasting for the next 6 months
- Classical Decomposition method shows that there is a stabilizing trend in the time series with a forecast of drop in Sales in the next 6 months

Result 2b. EU Consumer Quantity Forecast on validation set



a) Linear Model Forecast



b) Auto ARIMA Model Forecast

- Time Series of Europe Quantity for the 48 months shows it is stabilizing and it also shows the Quantity forecasting for the next 6 months
- Classical Decomposition method shows that there is a stabilizing trend in the time series with a forecast of drop in Quantity in the next 6 months



Evaluation:

- Evaluate model on validation set using MAPE
- Choose best model out of Model 1 & Model 2 using MAPE
- For Auto ARIMA plot ACF of residuals to check it resembles white noise





Recommendations for Inventory Management

Most profitable segments:

- APAC and EU consumer segments seem to be the most profitable ones. Items corresponding to these segments should be kept more in stock.
- **Inventory Levels:**
 - Inventory levels should be kept as predicted by the ARIMA model for the case of EU consumer segment, since the ARIMA model's predictions had a low MAPE value.
 - However, the regression model should be used for predicting inventory requirement for the APAC consumer segment, as it is the only one of the two that is able to capture the seasonal behavior of sales and demand for this segment.
 - In general, a buffer of at least 25% should be kept on inventory levels, as none of the models used was extremely accurate.





Deployment and Maintenance

- For deployment, it needs only R and not any other tools as the data is a .csv file.
- When performing maintenance, need to consider the portions of the original process should be re-executed. As the attribute and market segments may change over time.





Conclusions

1. Based on data provided we helped “Giant Super Store” in identifying 2 most profitable market segments as APAC Consumer and EU Consumer.
2. We created total 8 forecasting models for top 2 segments out of which 4 best were selected for forecasting future 6 months sales & quantity for months January 2015 to June 2015.
3. Below is summary of 4 key forecasts on test data(Jan – June 2015):
 - a) APAC Consumer Sales is likely to rise in next 6 months with small fluctuations.
 - b) APAC Consumer is also likely to rise steeply in coming 6 months.
 - c) EU Consumer Sales may show slow rise in coming months.
 - d) EU Consumer Quantity is likely to drop during initial 1 or 2 months & then rise rapidly in next 3 months, eventually reaching a plateau.

