Time Series : Retail-Giant Sales Forecasting

Submitted by :

1. Prakash Babu
2. Asher Peter Babu
3. Jithin Thomas
4. Ajay Joshy

Date: 29/10/2017

Introduction

Problem Statement

- “Global mart” who is an online store super giant wishes to forecast and predict the sales for the next 6 months in order to manage their inventory for their 7 different market segments and in 3 major categories (7x3 buckets)

Goal

- To identify the top 2 most profitable and consistent market buckets catered to by the store and forecast the its demand and sales

Analysis approach

- Convert the 3 transaction-level attributes Sales, Quantity and Profit into time series by aggregating them to their monthly values for each of the 21 “Market_Segments” based on their order dates.
- Find the COV (coefficient of variation) for each “Market_Segment” and identify the top 2 segments based on total profit
- Build model for forecasting the next 6 months “sales” and “quantity” attributes using classical decomposition and auto ARIMA
- Model evaluation is done using MAPE on last 6 months sales/quantity data aggregated on monthly basis

Data Understanding and Preparation

Understanding the data

1. The data contains transactional level data with 51290 unique records with 24 attributes
2. There are 7 markets - "Africa“, "APAC“, "Canada“, "EMEA“, "EU“, "LATAM“, "US" and 3 product segments - "Consumer“, "Corporate“, "Home Office“
3. There is 4 years of transactional data from year 2011 to 2014

Preparing the data

1. **Missing values imputation:**
   There are 41296 NA values present in postal code which are imputed with 0
2. **Standardizing date format:**
   Transform dates stored in character and numeric vectors to Date or POSIXlt objects in "%d-%m-%Y“ format
3. **Derived variables:**
   Derive a new attribute named “Market_Segment” by concatenating market and segment attributes
4. **Outlier treatment for attributes:**
   The values are capped at 95th percentile for sales, quantity, discount attributes and at both 5th and 95th percentile for profit and shipping cost
5. **Aggregating data and ordering data:**
   The data is first ordered based on the year and month, then it is aggregated based on “Market_Segment” for further analysis

EDA-Multivariate Analysis

- Correlation between attributes – sales, quantity, discount and shipping.Cost and profit are checked

	Sales	Quantity	Discount	Profit	Shipping.Cost
Sales	1	0.358	-0.110	0.577	0.880
Quantity	0.358	1	-0.018	0.200	0.3223
Discount	-0.11	-0.018	1	-0.473	-0.1005
Profit	0.577	0.200	-0.473	1	0.5182
Shipping.Cost	0.72	0.272	-0.076	0.443	1

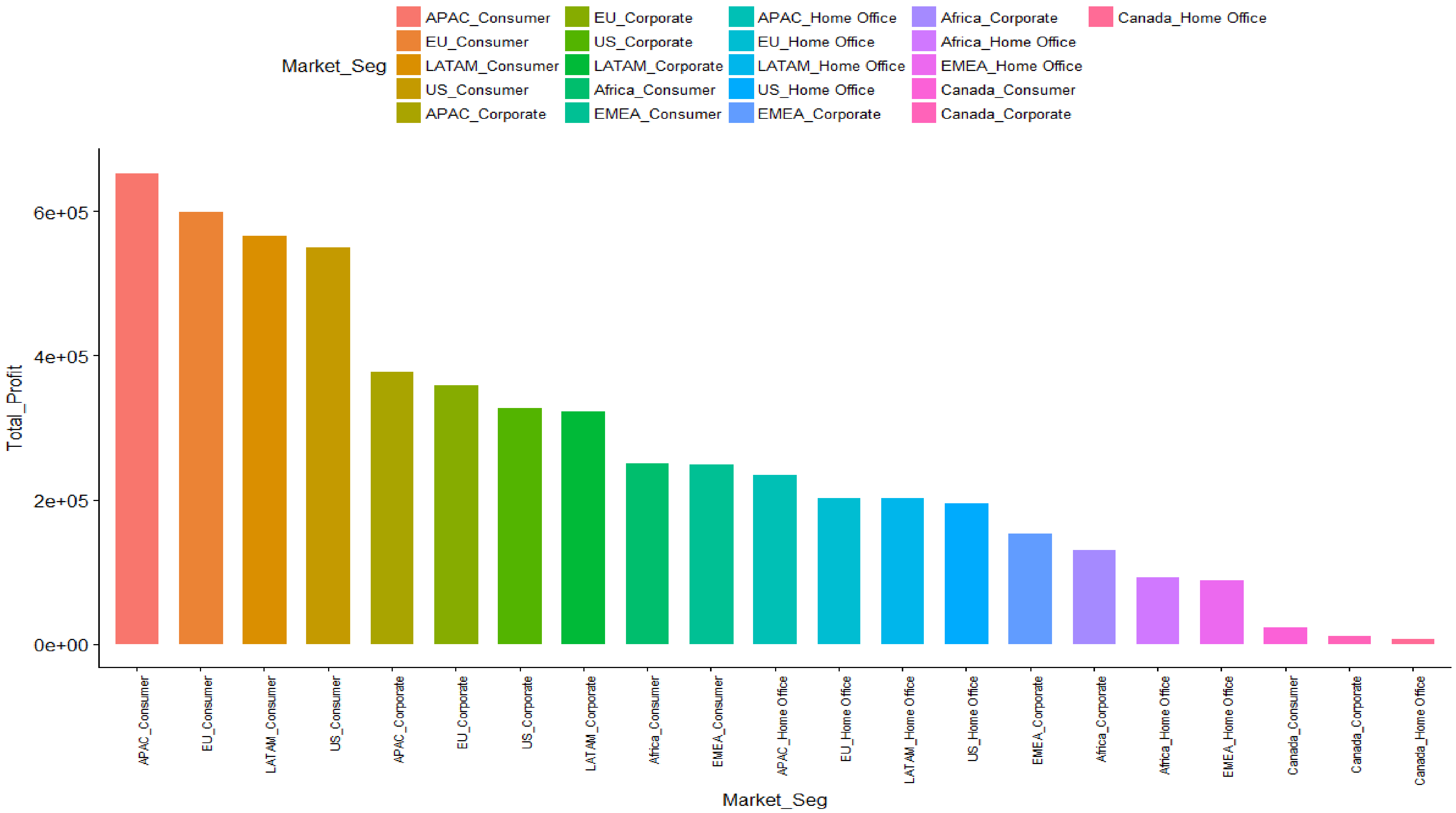
- Each Market_Segment data is ordered by year and month and then aggregated for Total Profit and COV to identify the top 2 Market_Segments. The top 2 segments are below:

1. **APAC_Consumer** with total profit of 654616.148 and COV of 0.4430291

2. **EU_Consumer with** total profit of 599659.444 and COV of 0.4881849

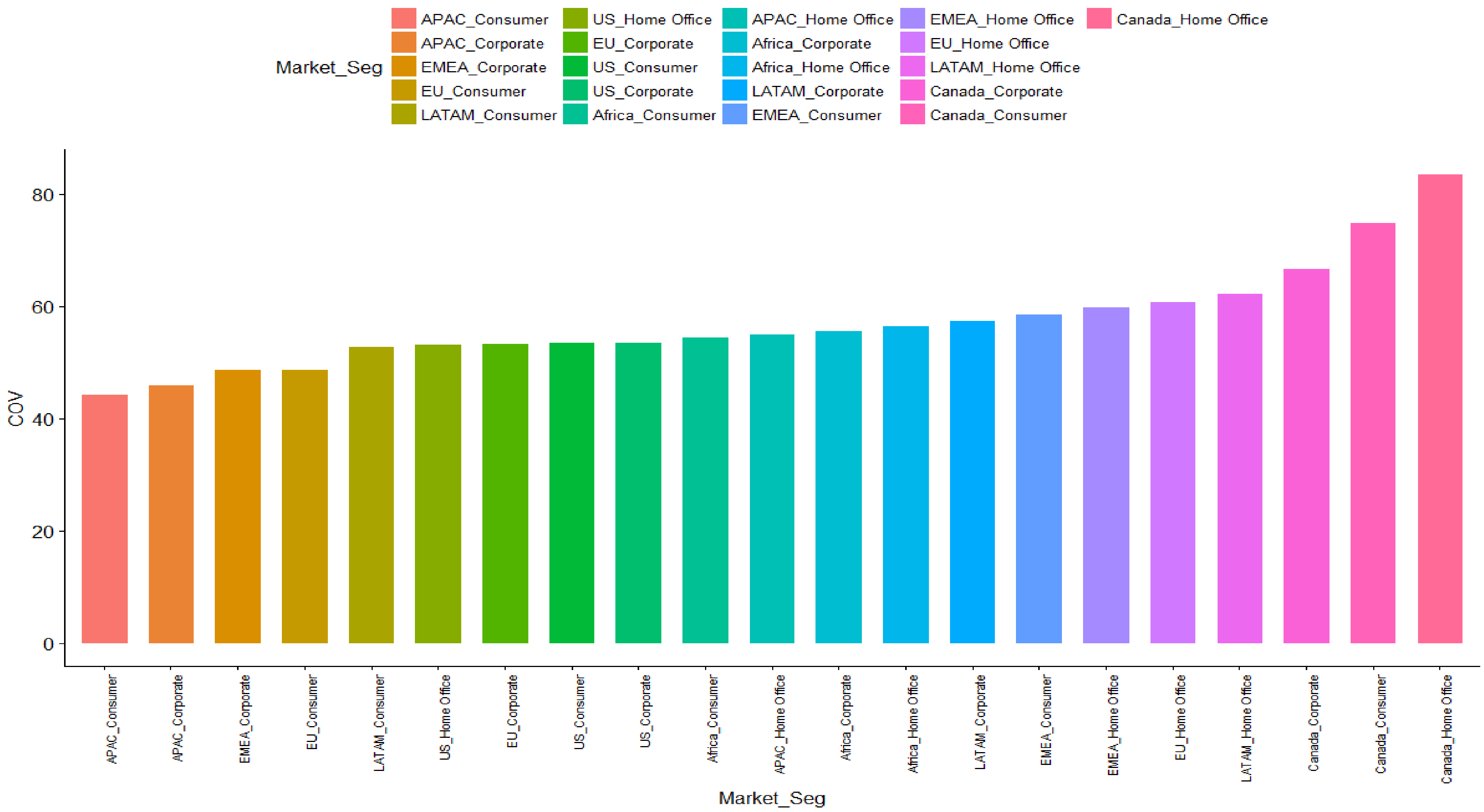
EDA Cont..

Total Profit aggregated for each Market_Segment



EDA Cont..

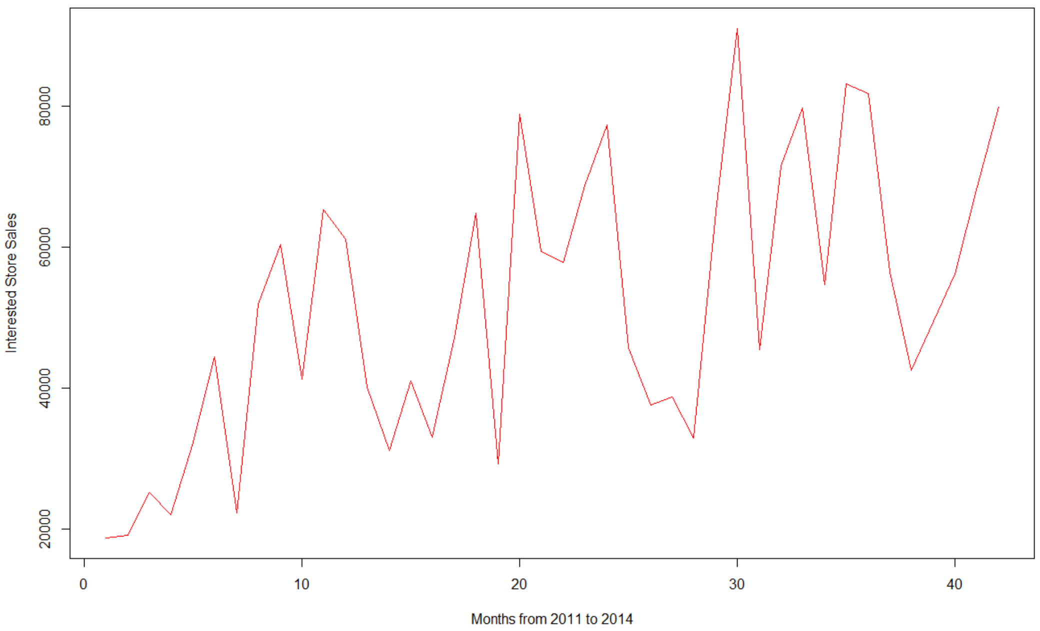
COV of Total_Profit for each Market_Segment



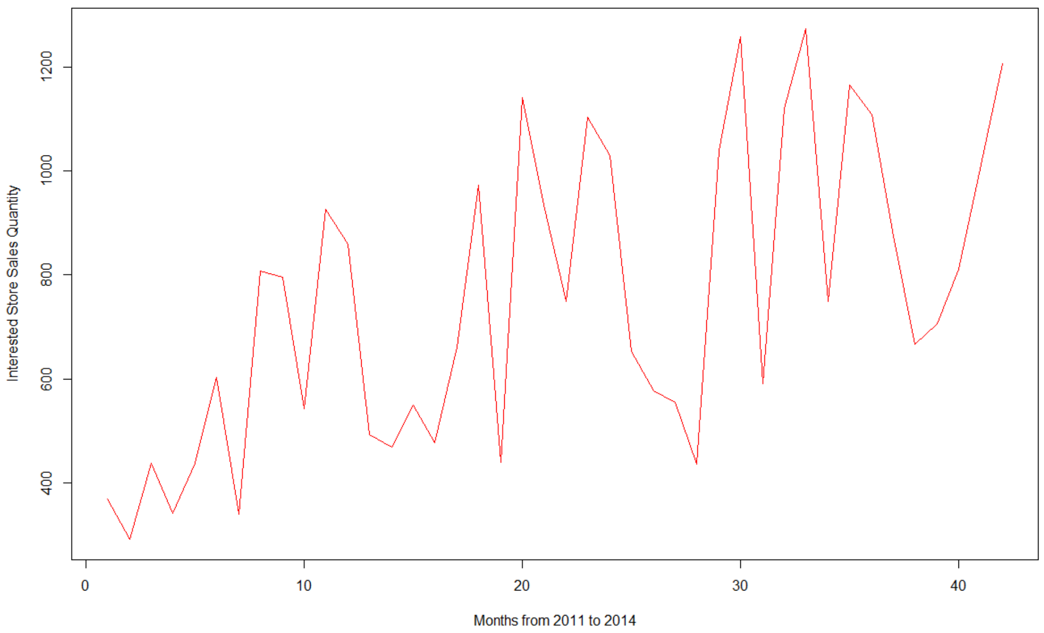
Model Creation-Smoothening

- The 2 most profitable Market_Segments are **APAC_Consumer, EU_Consumer**
- The Sales, Quantity and Profit attributes are aggregated and grouped based on year and month
- There are 48 records signifying 4 years of monthly aggregate data
- The first 42 rows are used for model building and the remaining 6 used for testing the model
- The monthly **sales** and **quantity** plot for the APAC_Consumer and EU_Consumer is given below

Interested Store Sales: 2011 to 2014

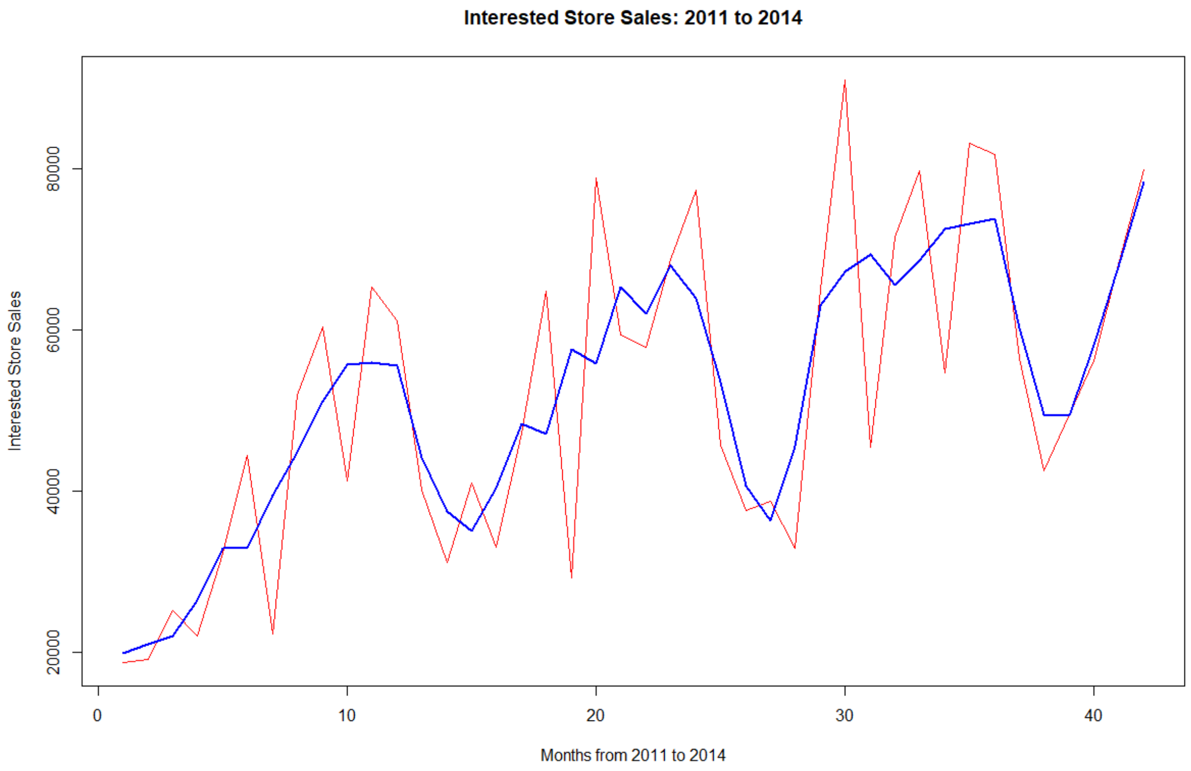


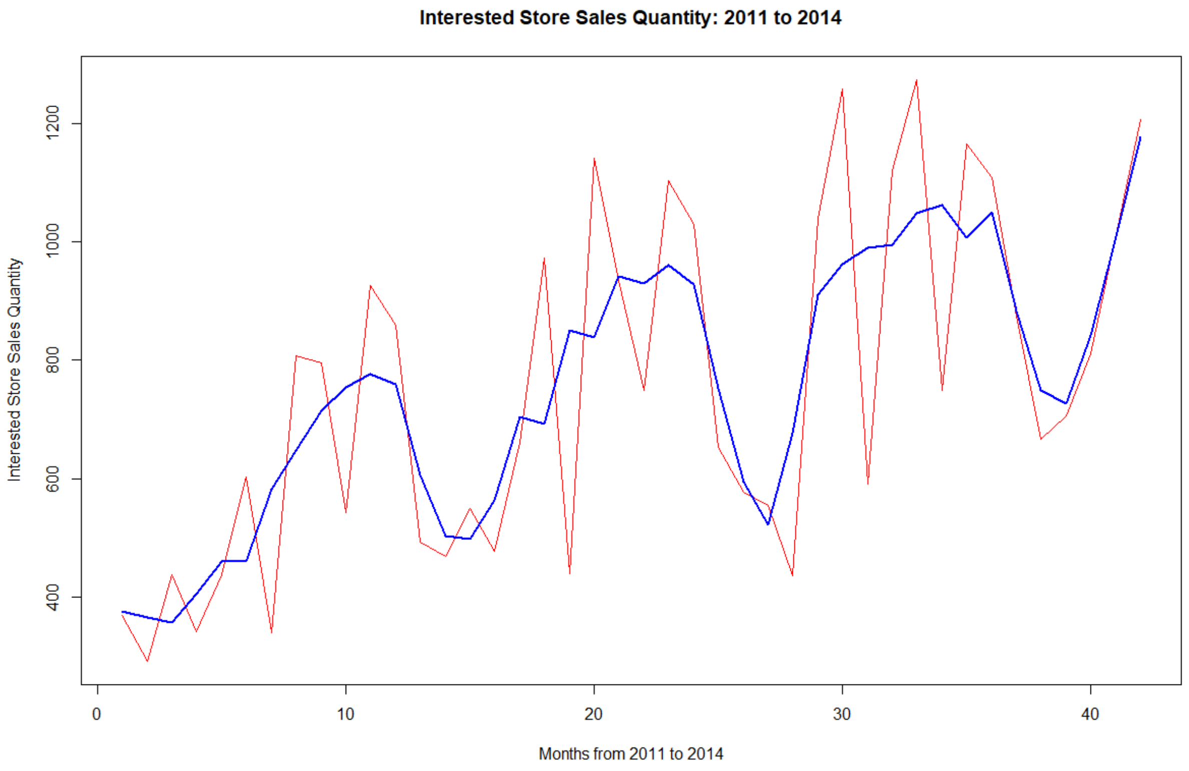
Interested Store Sales Quantity: 2011 to 2014



Model Creation Cont…

- Plots after applying Moving Average Smoothing

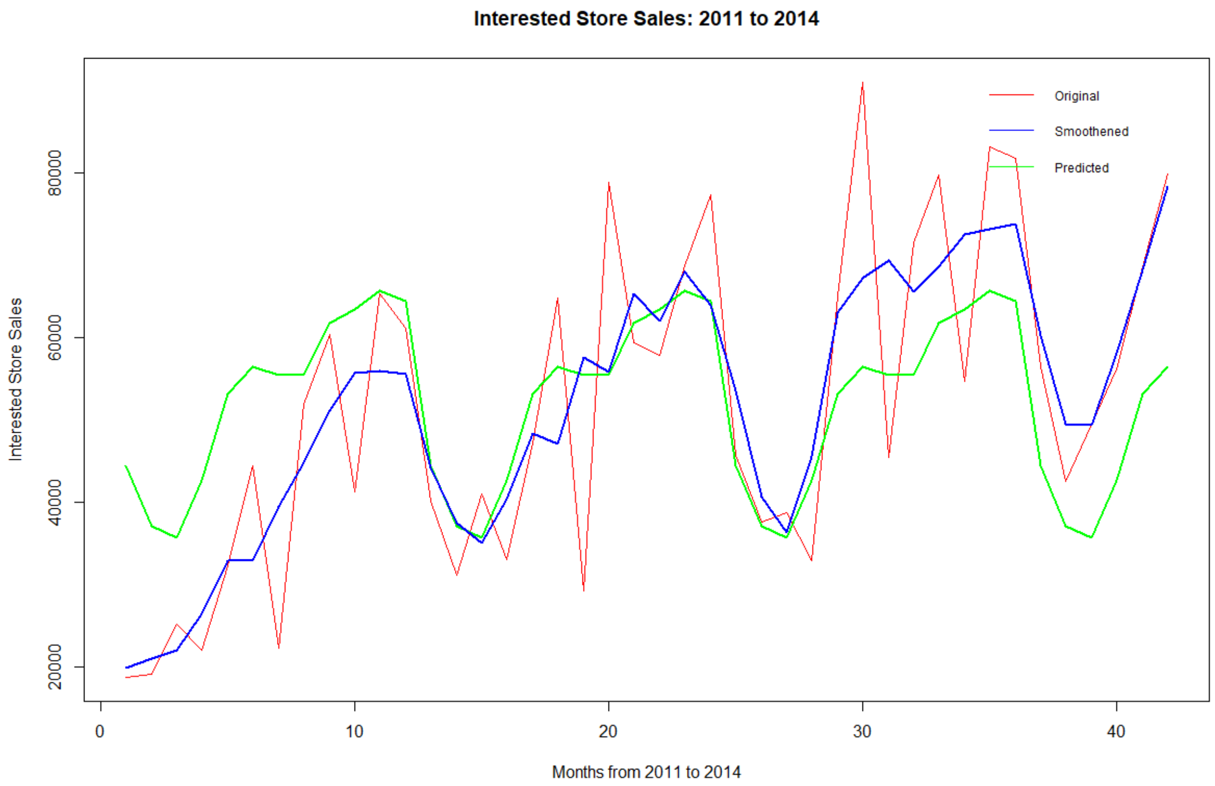


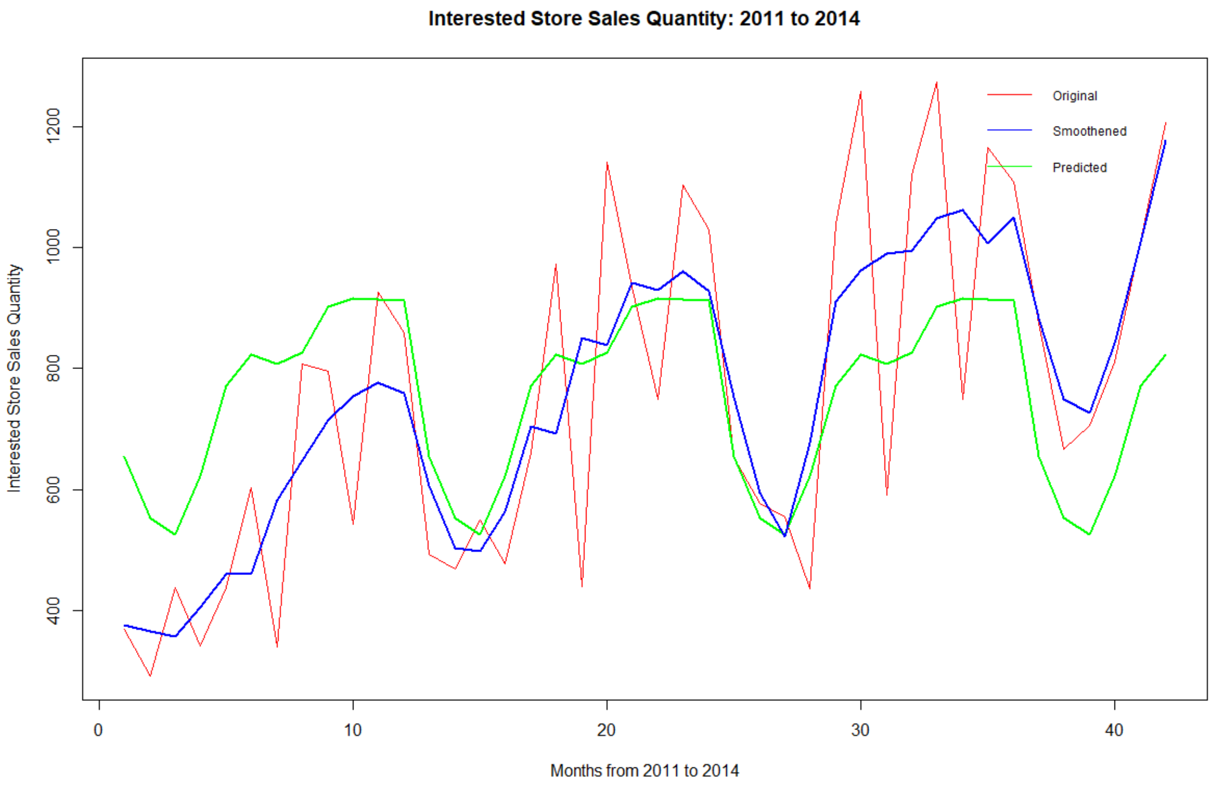


Model Creation – Classical Decomposition

modelling global predictors

- Fitting a model with trend and seasonality to the smoothened sales and quantity data.
- Trend and Seasonality is modelled as multiplicative model using sinusoid function

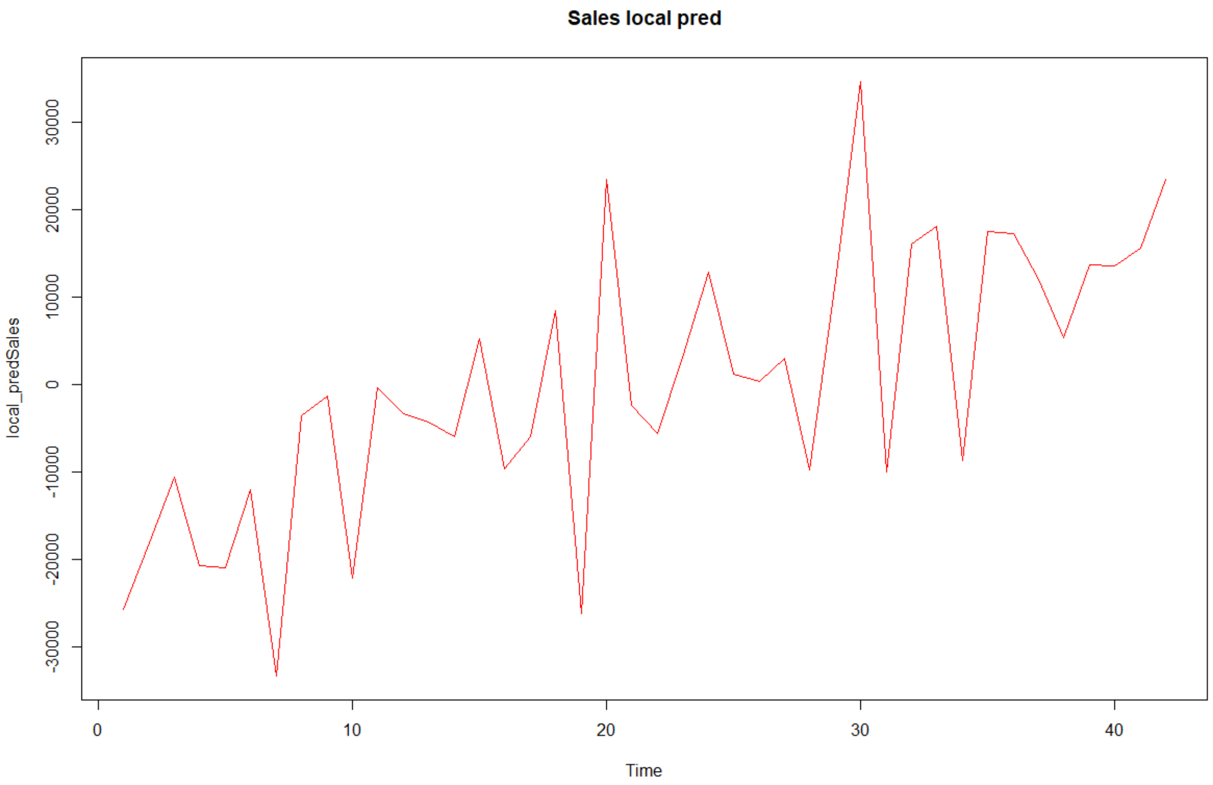


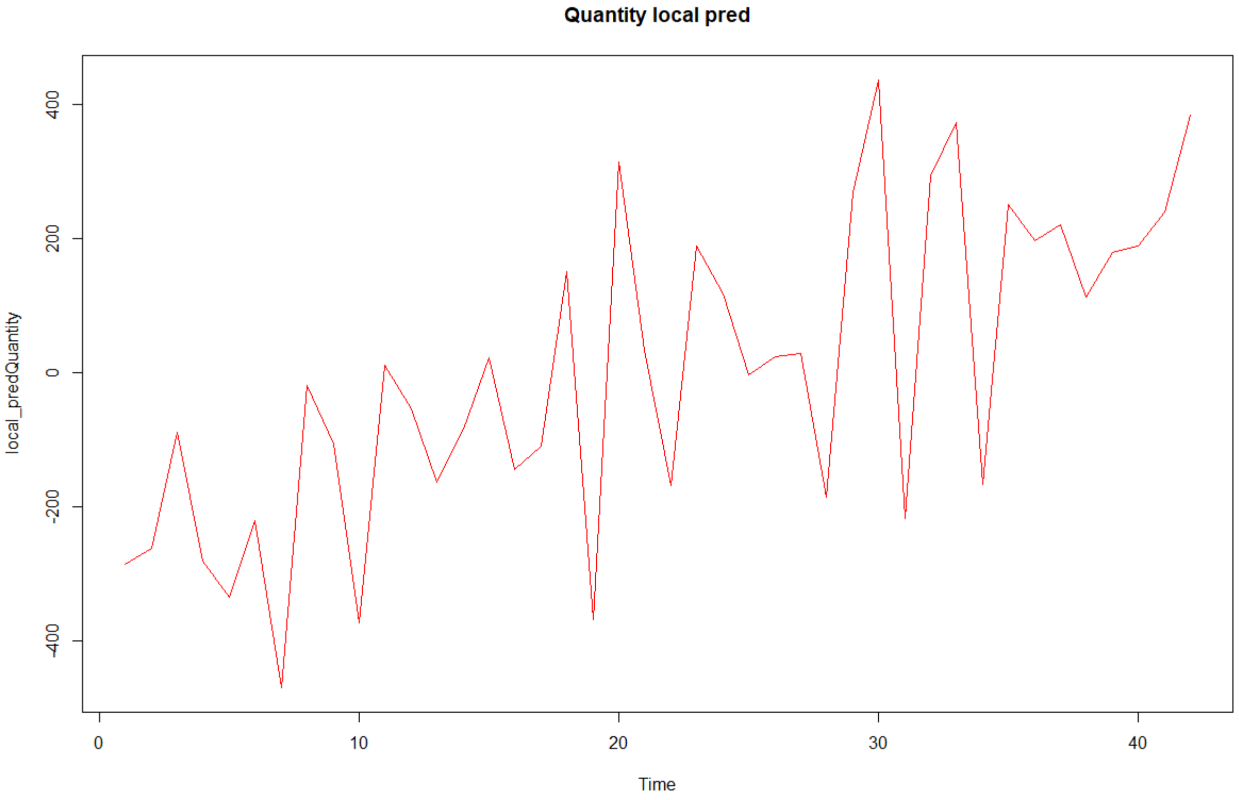


Model Creation Cont…

modelling local predictors

- Local predictors are calculated by subtracting global predictors from the time series data

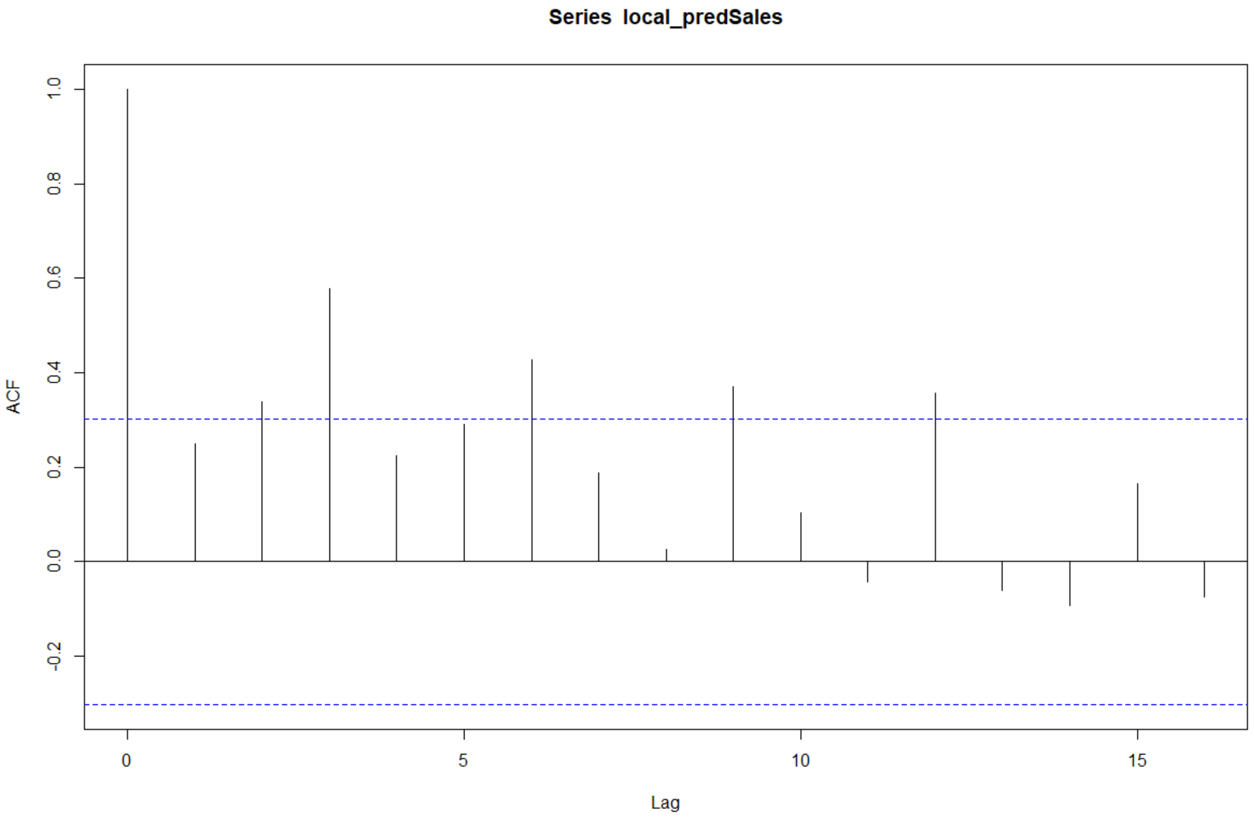


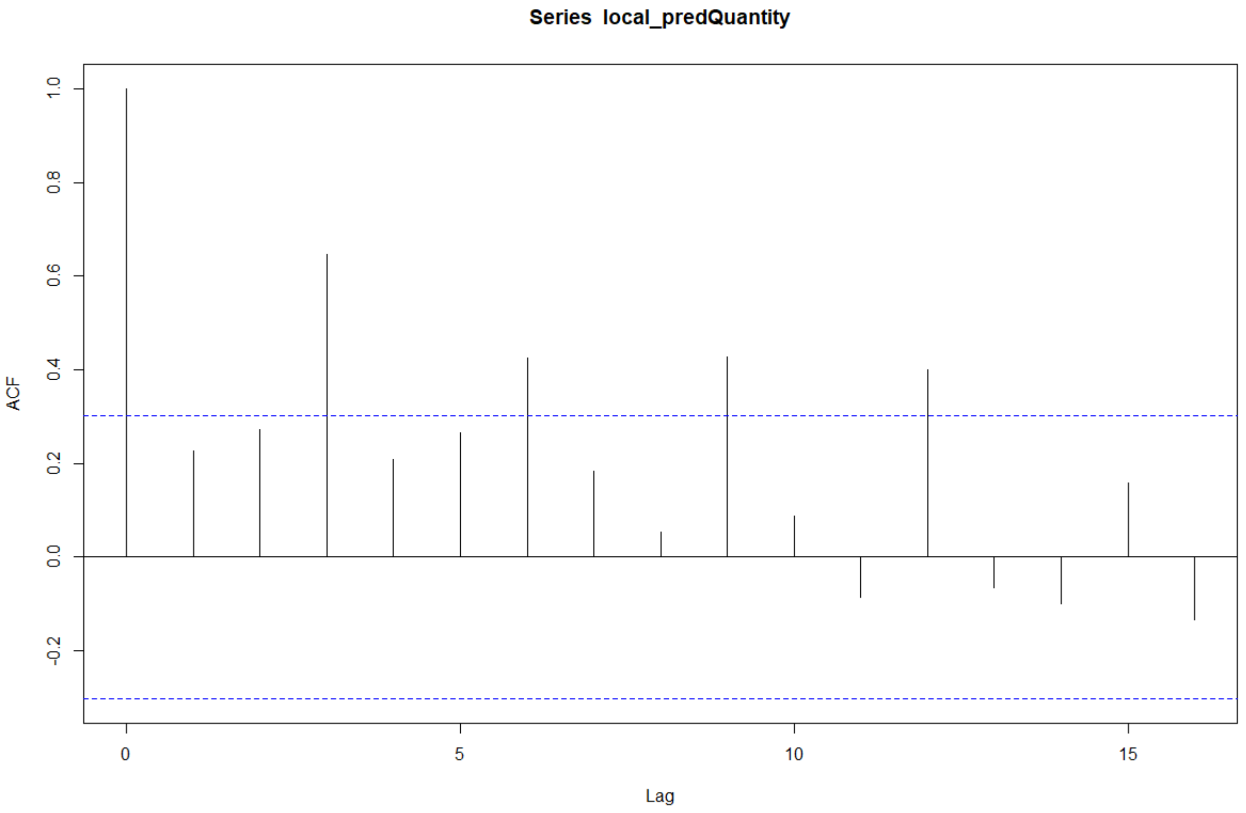


Model Creation Cont…

Residual series

- The residual series is calculated from the local predictors (time series – global trend) using the auto-correlation function





Model Creation Cont…

Model summary

- **Sales**
  To check for auto-regressive behavior in the local predictors we do an ARIMA fit on it

- Checking for autoregressive behavior using ARIMA

```
Series: local_predSales ARIMA(2,0,2) with non-zero mean
Coefficients: ar1 ar2 ma1 ma2 mean 1.4610 -0.4756 -1.7986 0.9999 -
1879.683 s.e. 0.1535 0.1561 0.2416 0.2661 13213.775
sigma^2 estimated as 117287356: log likelihood=-450.61 AIC=913.21
AICc=915.61 BIC=923.64
```

- Checking if Residual series is white Noise

1. Adf test :p-value = 0.01- Null hypothesis Rejected, Stationary series, ie white noise
2. Kpss test: p-value = 0.1 – Null Hypothesis accepted, stationary series ie white noise

- **Quantity**
  To check for auto-regressive behavior in the local predictors we do an ARIMA fit on it

- Checking for autoregressive behavior using ARIMA

```
Series: local_predQuantity ARIMA(2,0,2) with non-zero mean
Coefficients: ar1 ar2 ma1 ma2 mean 1.042 -0.0601 -1.3296
0.5643 1.2223 s.e. 0.480 0.4634 0.4612 0.3133 181.1414
sigma^2 estimated as 31166: log likelihood=-275.43
AIC=562.85 AICc=565.25 BIC=573.28
```

- Checking if Residual series is white Noise

1. Adf test :p-value = 0.01- Null hypothesis Rejected, Stationary series, ie white noise
2. Kpss test: p-value = 0.1 – Null Hypothesis accepted, stationary series ie white noise

Model Creation Cont…

last 6 months predicted values: Classical Decomposition

Month	Sales: Predicted	Sales: Actual
43	55470.47	56136.07
44	55396.24	112600.15
45	61713.17	87491.74
46	63395.26	93519.86
47	65682.31	112961.79
48	64406.44	105288.76

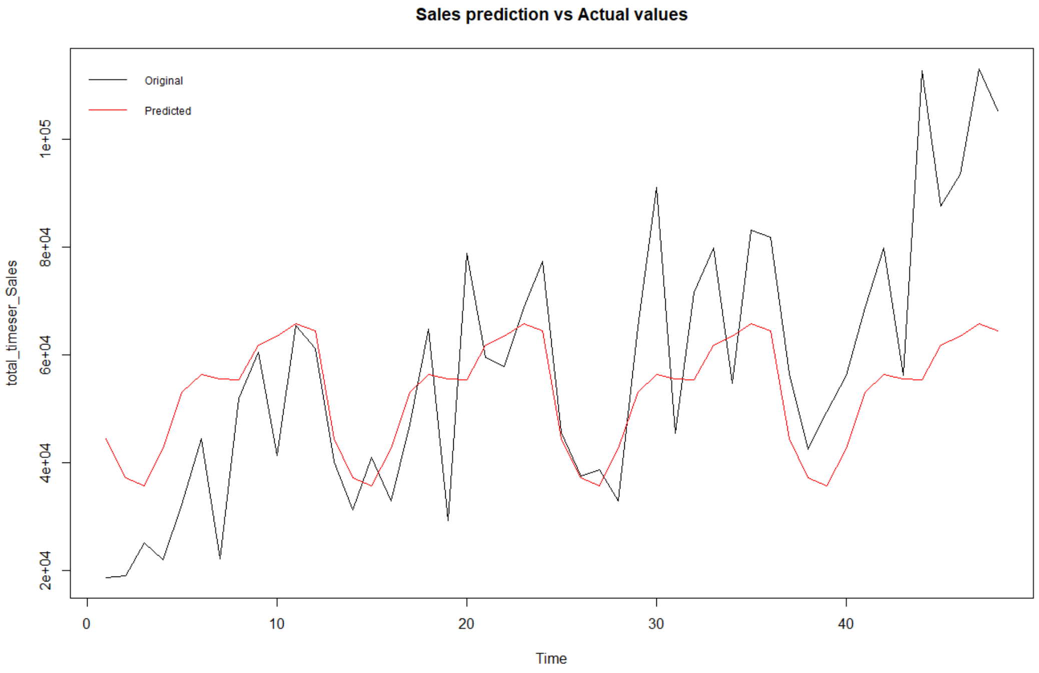
Month	Quantity: Predicted	Quantity: Actual
43	807.6667	793
44	826.5556	1715
45	901.1111	1313
46	915.2222	1322
47	914.4444	1631
48	912.1111	1697

Model Evaluation-MAPE

MAPE is calculated for last 6 months (ie month 43 to 48)

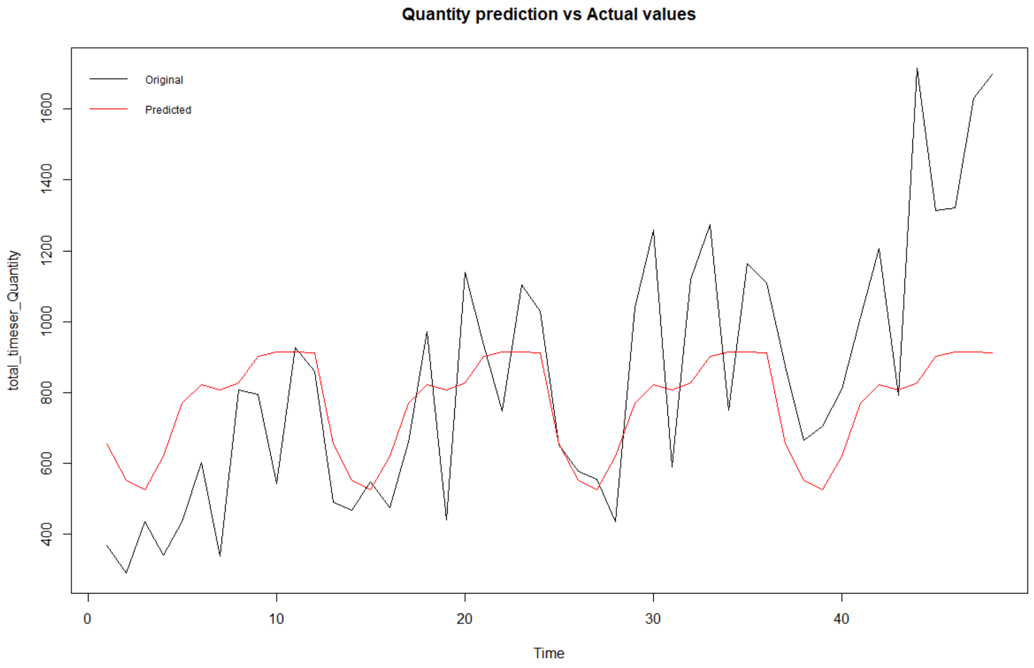
Sales

- MAPE accuracy for sales: 32.391%



Quantity

- MAPE accuracy for quantity: 34.329%



Model Creation - ARIMA

Sales

-
  ```
  Series: timeserSales ARIMA(2,0,2) with non-zero mean
  Coefficients: ar1 ar2 ma1 ma2 mean 0.3819 0.4890 0.0501 -0.5105
  49971.661 s.e. 0.4707 0.2962 0.5054 0.2718 9416.777
  sigma^2 estimated as 314118548: log likelihood=-468.12 AIC=948.25
  AICc=950.65 BIC=958.67
  ```

- Level of differencing : 0
- AIC values are higher than those for classical decomposition (lower the better)
- Loglikelihood values are lower than that for classical decomposition (higher the better)
- Checking if Residual series is white Noise

1. Adf test :p-value = 0.02819- Null hypothesis Rejected, Stationary series, ie white noise
2. Kpss test: p-value = 0.1 – Null Hypothesis accepted, stationary series ie white noise

Quantity

-
  ```
  Series: timeserQuantity ARIMA(2,1,0)
  Coefficients: ar1 ar2 -0.5850 -0.4884 s.e. 0.1355 0.1327
  sigma^2 estimated as 61720: log likelihood=-283.63
  AIC=573.26 AICc=573.91 BIC=578.4
  ```

- Level of differencing : 1
- AIC values are higher than those for classical decomposition (lower the better)
- Loglikelihood values are lower than that for classical decomposition (higher the better)
- Checking if Residual series is white Noise

1. Adf test :p-value = 0.05674 – Since not significantly greater than 0.05 rejecting the Null hypothesis , hence its Stationary series, ie white noise
2. Kpss test: p-value = 0.1 – Null Hypothesis accepted, stationary series ie white noise

Model Creation Cont…

last 6 months predicted values: ARIMA

Month	Sales: Predicted	Sales: Actual
43	64408.98	56136.07
44	60180.12	112600.15
45	60930.29	87491.74
46	59148.99	93519.86
47	58835.44	112961.79
48	57844.67	105288.76

Month	Quantity: Predicted	Quantity: Actual
43	946.3679	793
44	877.0119	1715
45	900.4517	1313
46	874.6234	1322
47	876.0513	1631
48	863.4559	1697

Model Evaluation-MAP

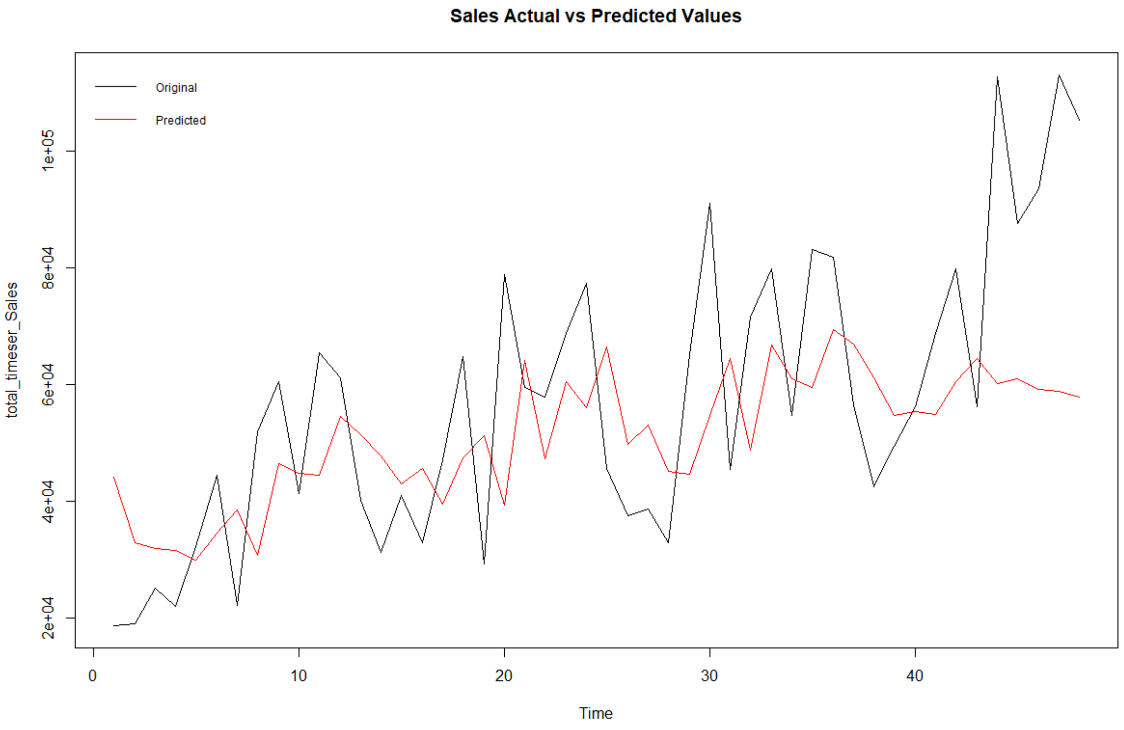
MAPE is calculated for last 6 months (ie month 43 to 48)

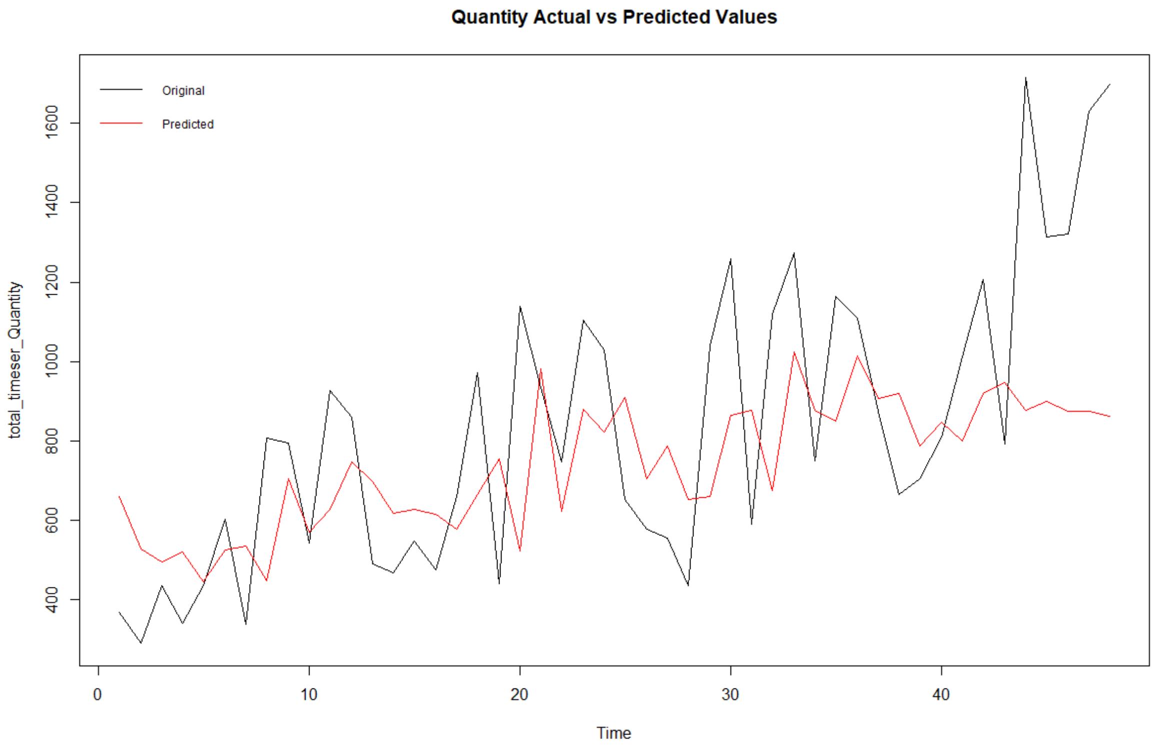
Sales

- MAPE Accuracy: 36.896%

Quantity

- MAPE Accuracy: 38.144%


Sales Actual vs Predicted Values
Original
Predicted
total_timeser_Sales
1e+05
8e+04
6e+04
4e+04
2e+04
Time


Quantity Actual vs Predicted Values
Original
Predicted
total_timeser_Quantity
Time