Retail-Giant Sales Forecasting - Case Study

SUBMISSION

Group Name:

1. Ashish Chandan
2. Kasinath Kalava
3. Ramashankar Nayak
4. Ravi Bhavsar

Business Objective

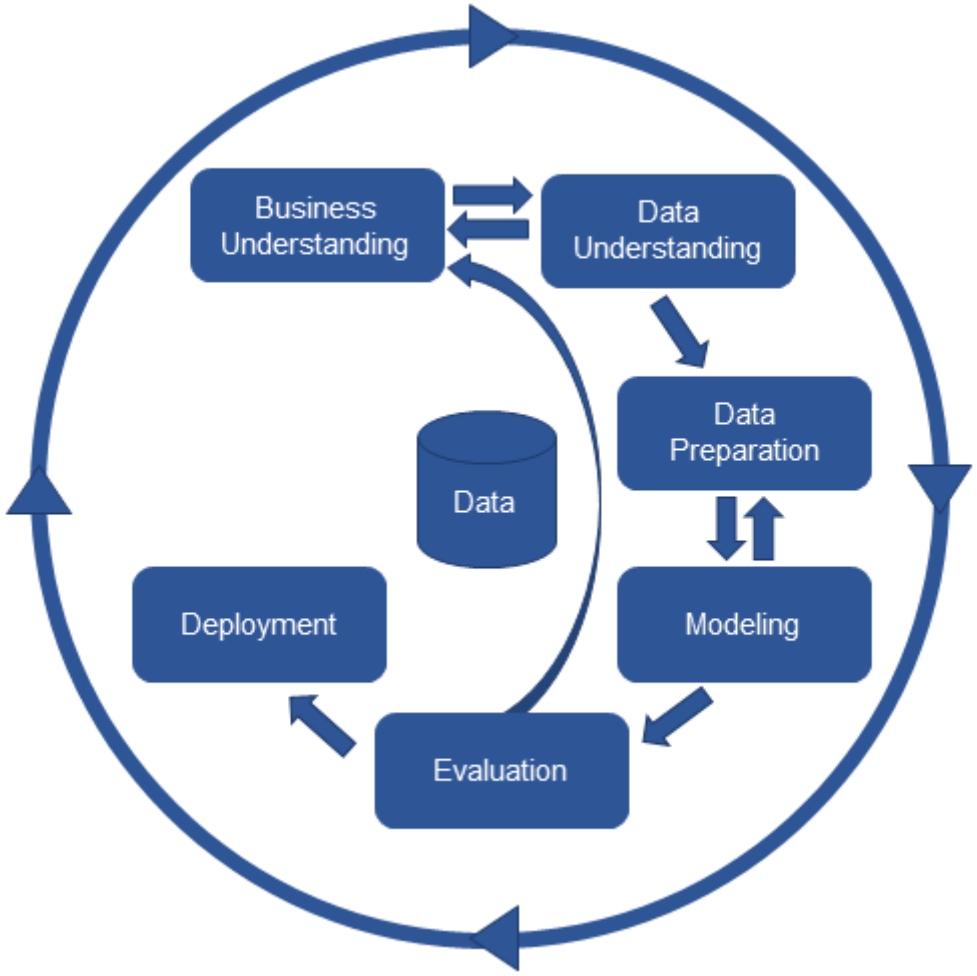
"Global Mart" is an online store super giant having worldwide operations. It takes orders and delivers across the globe and deals with all the major product categories - consumer, corporate & home office.

The Sales/ Operations manager objective is to finalize the plan for the next 6 months, that would help to manage the revenue and inventory accordingly.

Goals:

- To forecast the sales and the demand for the next 6 months that would help you manage the revenue and inventory accordingly.
- To subset the data into 21 (7*3) buckets before analyzing these data (As the store caters to 7 different market segments and in 3 major categories.)
- To find out 2 most profitable (and consistent) segments from these 21 buckets.
  - Criteria for selection
    - Maximum profit
    - Consistent profit (based on profit %age month on month)
- To forecast the sales and demand for these 2 most profitable segments.

Problem Solving Methodology - CRISP DM Framework

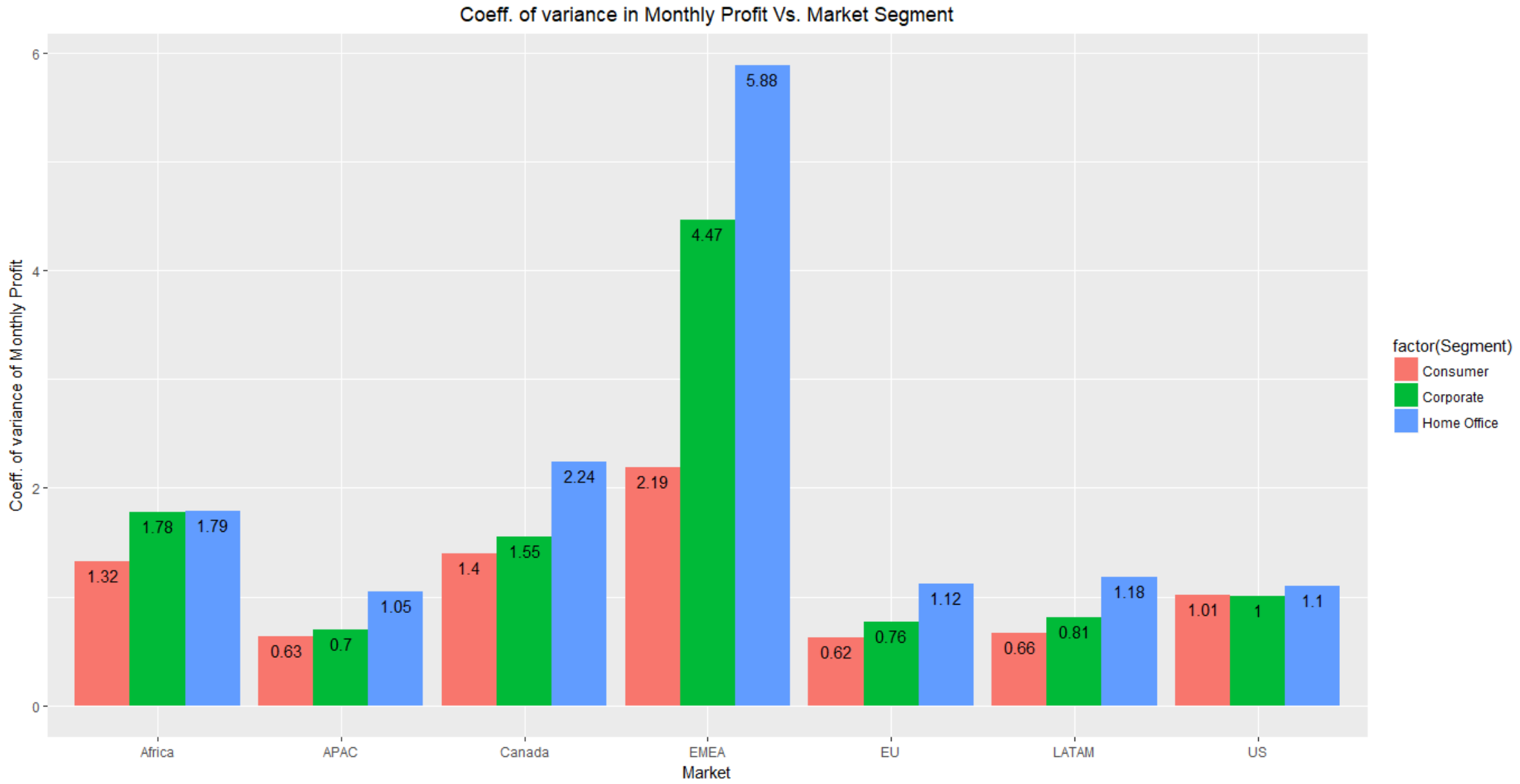


- **Data Understanding**: Superstore Sales transaction data has been provided for different market segments for 4 years. Each data point is a transaction.
- **Data Preparation**: Involves handling the missing values, formatting date and removing unwanted data for proper analysis.
  - To get Best market segment i.e. most Profitable and Consistent
  - Preparing Time Series data against Sales and Quantity
- **Modeling**: involves building model that closely represents trend and seasonality. To check for residue using ARIMA whether residue is pure noise.
  - Decompose and see the shape of trend and seasonality
  - Smoothen the Curve Using Simple avg. or Weighted
- **Evaluation**: steps involve to run the model on test data and perform other tests. Once the test goes good and model is accepted use the model to predict future Quantity and Profit.

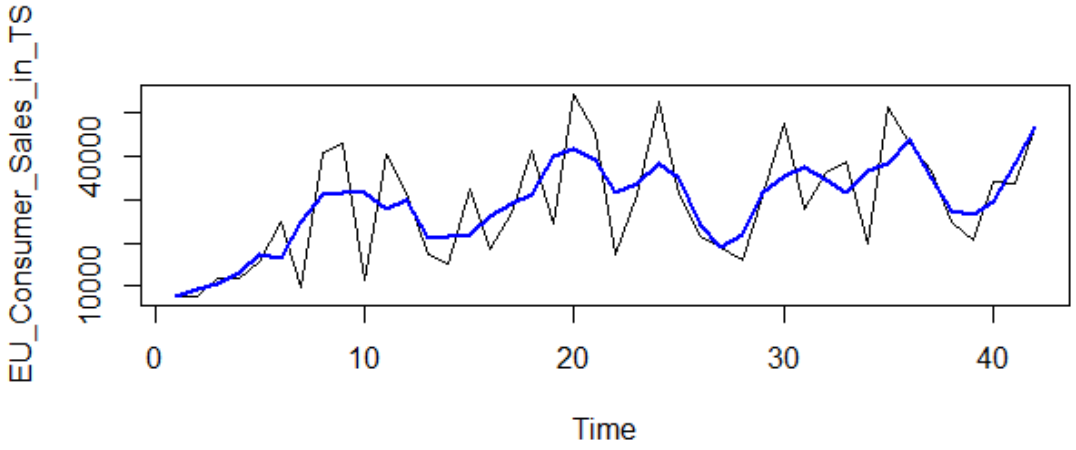
Analysis : Data Cleaning and Preparation

- Global Superstore Dataset:
  - Number of data points in complete dataset : 51,290
  - Number of Attributes : 24
- Removed unwanted attributes from the dataset and selected attributes of interest for data analysis:
  - Segment
  - Market
  - Order Date
  - Aggregate Sales
  - Aggregate Quantity
  - Aggregate Profit on Sales
- Handled missing values: we do have missing values in complete dataset. However after taking required attributes for data analysis no missing values were found.
- We have formatted the Order Date as Date and replaced each order date value with first day of corresponding month. Computed Coefficient of Variable (CV) for aggregated Profit.
- Based on the maximum profits and consistent profit month on month, we have chosen two Market Segments
  - EU Consumer
  - APAC Consumer

Analysis : Selection of Top two Market Segments

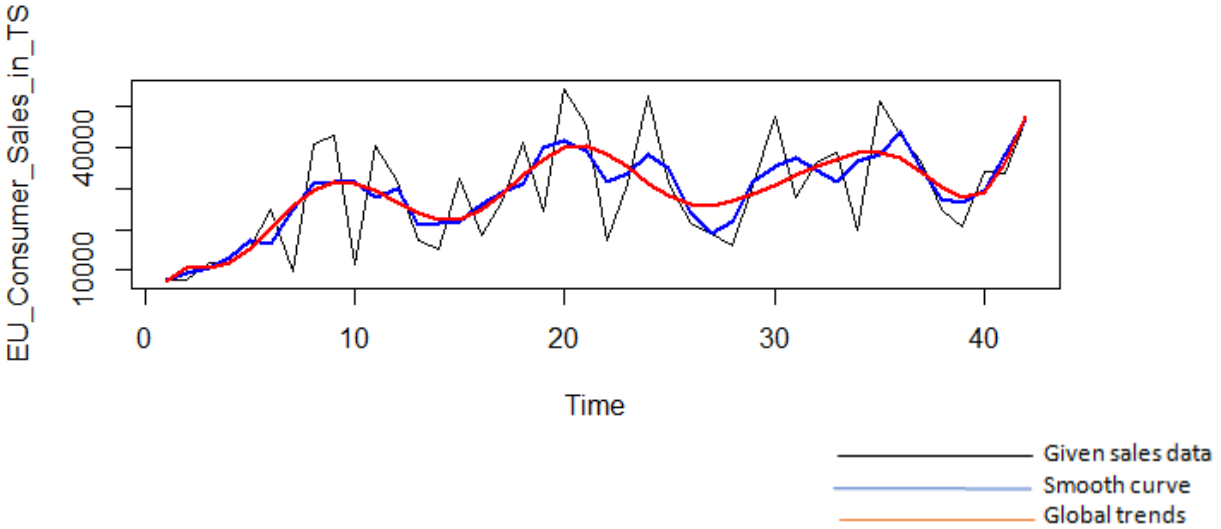


Analysis: Time Series for EU Consumer Sales (ARMA)



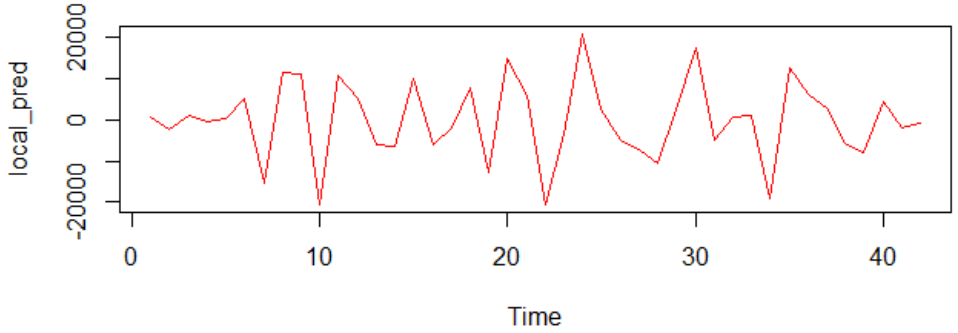
- Amplitude of the seasonal curve doesn't seem increasing with time.
- So, will try fitting additive model for the case.

To fit an additive model with trend and seasonality (Global)

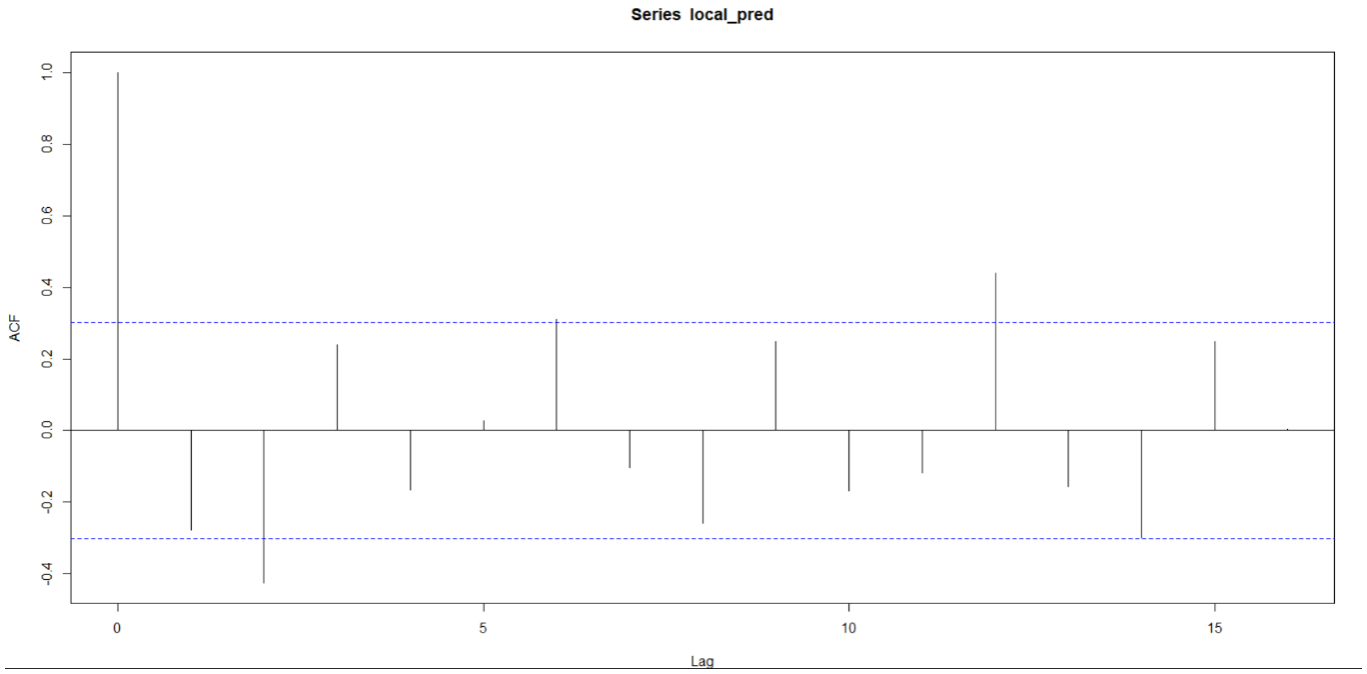


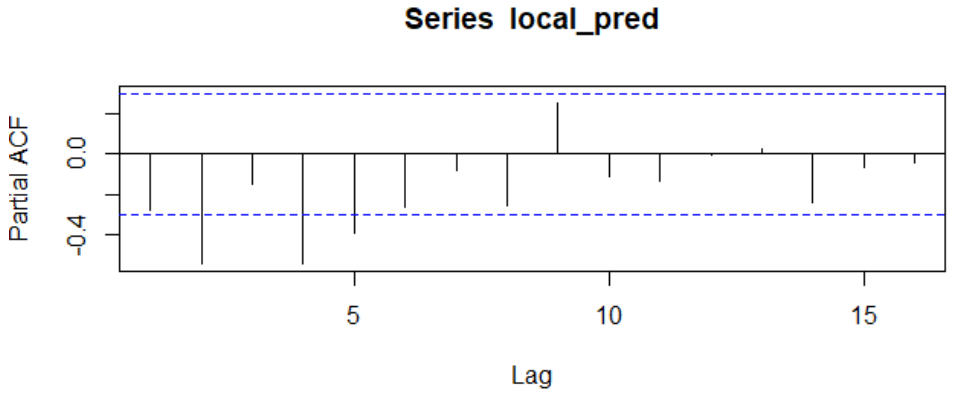
- Different degree (1,2,3,4) of polynomial curves were considered
- Adjusted R-squared for poly degree 3 looks best. So, considered lmfit3 for determining global component

Analysis: Locally predictable series using ARMA for EU consumer Sales



- Graph shows local component of time series left after removing global component from the given sales data





- There is no significant ACF or P-ACF values for local component

Evaluate the model using MAPE for EU Consumer sales

- Output of Auto Arima on Local component

  Series: local pred
  ARIMA(0,0,0) with zero mean
  Sigma^2 estimated as 92551568: log likelihood=-444.8
  AIC=891.61 AICc=891.71 BIC=893.35

- From above its clear that there is no predictable part in the derived local component for sales, To confirm further
- Below are the results of Dickey-Fuller and KPSS test which further confirms that the residual is only NOISE

  Dickey-Fuller : p-value = 0.01(<0.05) , kpss.test(resi) : p-value = 0.1 (>0.05)

- So there is only Global component which can be predicted, After making prediction for next six months (For Validation), MAPE value comes up to be **92.96**

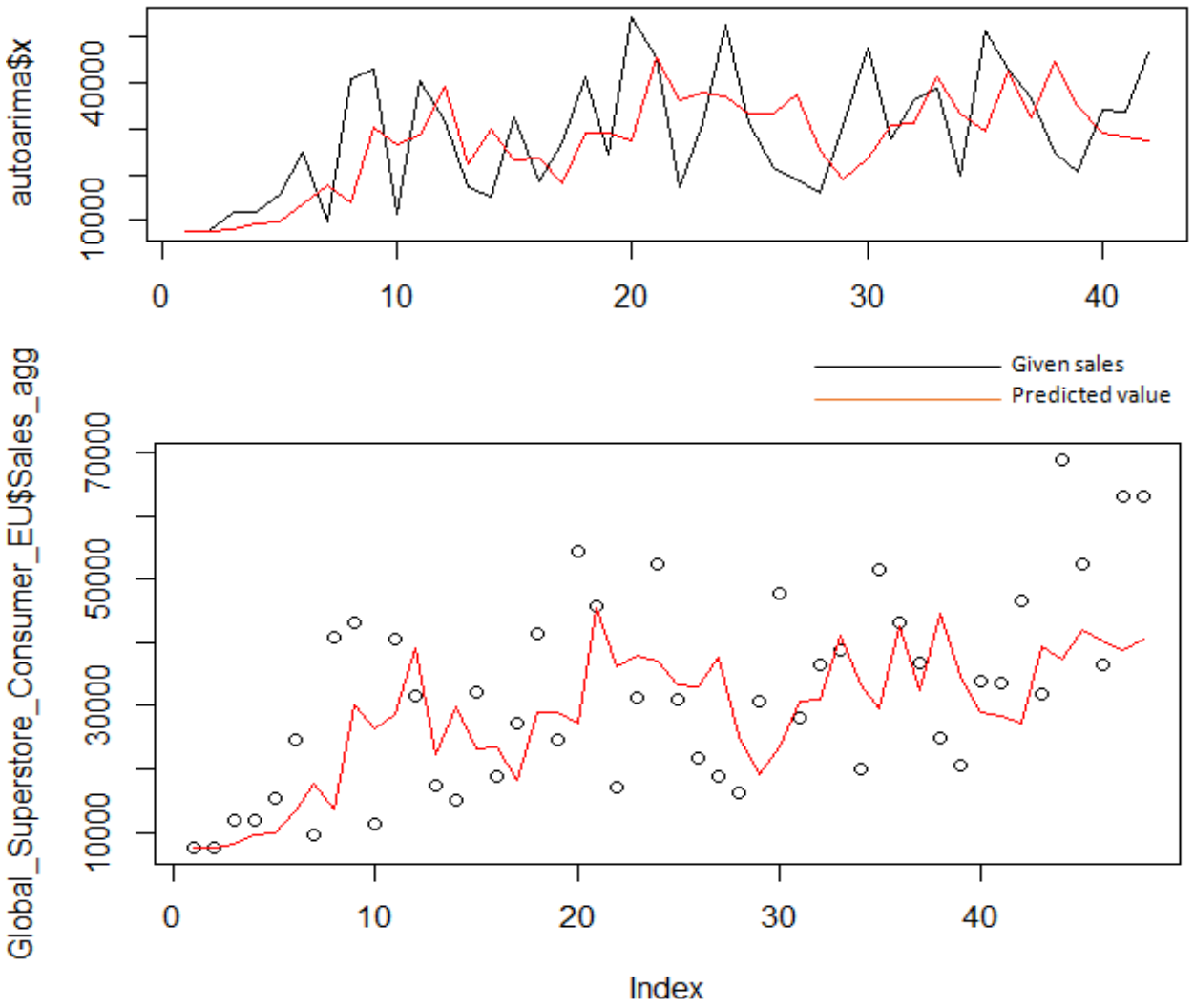
OUTPUT of AUTO ARIMA on EU Consumer Sales

- Series:
  ARIMA(2,1,0)
  Coefficients:

	ar1	ar2
	-0.5796	-0.4906
s.e.	0.1346	0.1310

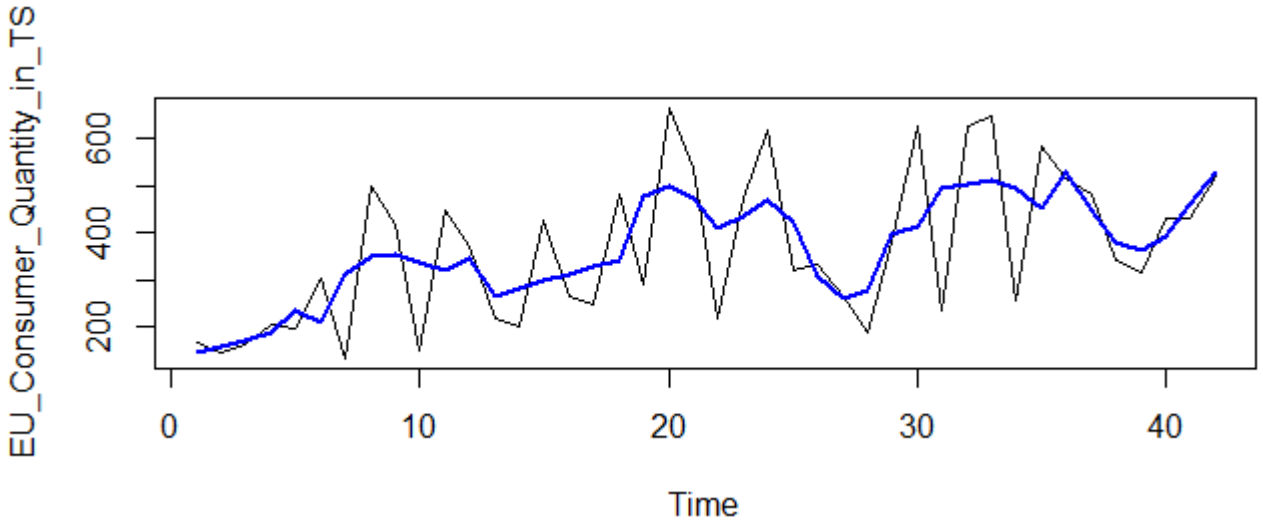
  sigma^2 estimated as 168564623: log likelihood=-445.84
  AIC=897.67 AICc=898.32 BIC=902.81

Analysis: Time Series for EU Consumer Sales (auto ARIMA)



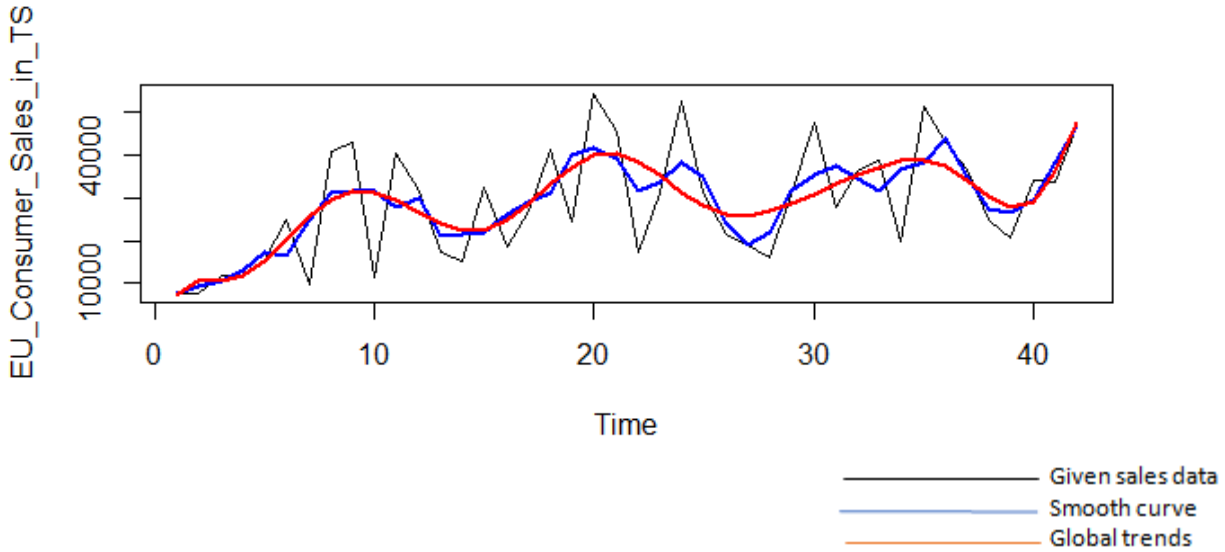
- Now after making prediction for next six months (For Validation), MAPE value comes up to be **28.92**
- Comparing MAPE values of ARMA(Manual) and AUTO ARIMA models , We can see that AUTO ARIMA model is better based on the MAPE value
- So we will use AUTO ARIMA (ARIMA(2,1,0)) for forecasting the next six months sales
- Below is the out come of Forecast
  49358.71 58063.62 59714.33 54191.79 56811.55 58010.84

Analysis: Time Series for EU Consumer Quantity (ARMA)



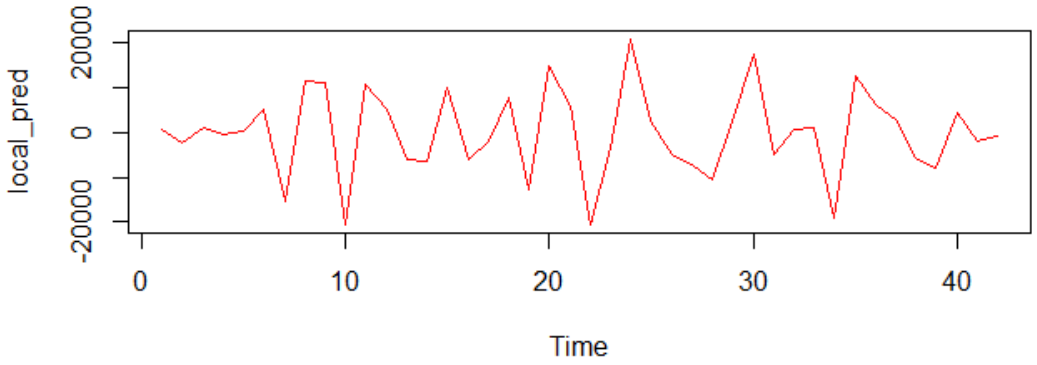
- Amplitude of the seasonal curve doesn't seem increasing with time.
- So, will try fitting additive model for the case.

To fit an additive model with trend and seasonality (Global)

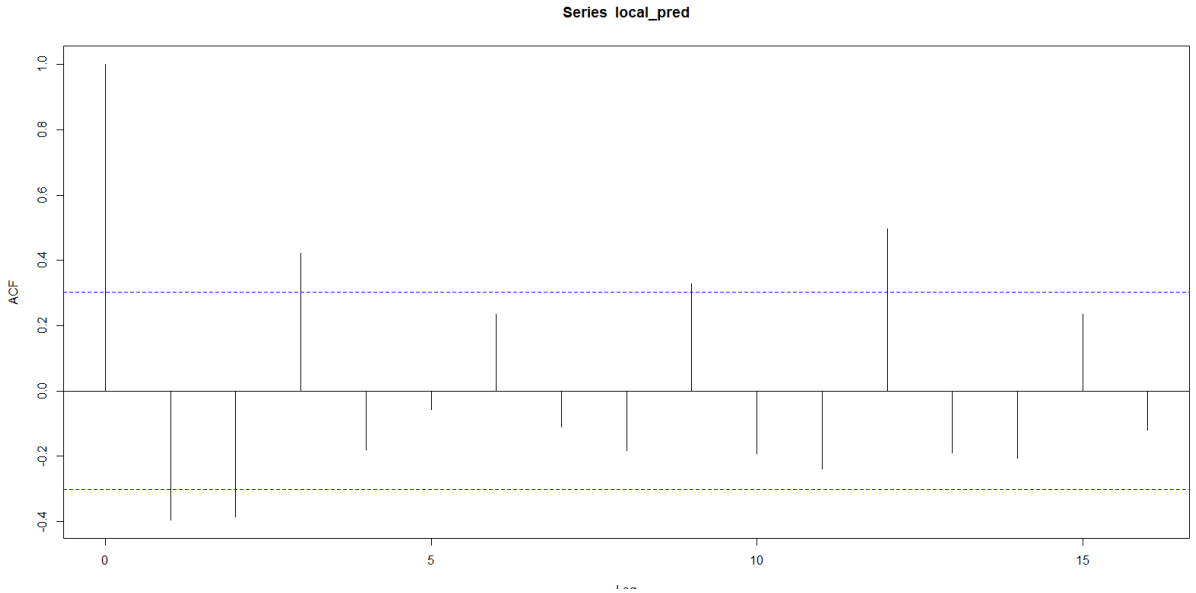


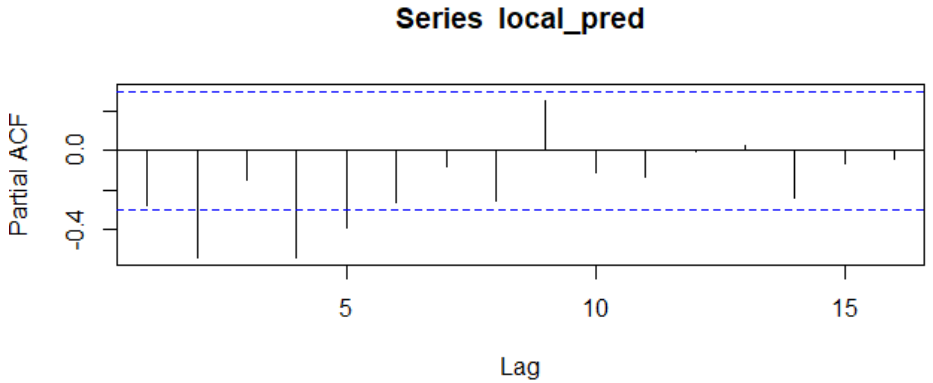
- Different degree (1,2,3,4) of polynomial curves were considered
- Adjusted R-squared for poly degree 3 looks best. So, considered lmfit3 for determining global component

Analysis: Locally predictable series using ARMA for EU consumer (Qty)



- Graph shows local component of time series left after removing global component from the given sales data





Evaluate the model using MAPE for EU Consumer Quantity

- Output of Auto Arima on Local component

Series: Series: local_pred
ARIMA(2,0,0) with zero mean
Coefficients: ar1 ar2
-0.6341 -0.6158
s.e. 0.1173 0.1131
sigma^2 estimated as 7284: log likelihood=-245.89 AIC=497.79 AICc=498.42 BIC=503

- Below are the results of Dickey-Fuller and KPSS test which further confirms that the residual(after ARMA) is only NOISE

Dickey-Fuller : p-value = 0.01(<0.05) , kpss.test(resi) : p-value = 0.1 (>0.05)

- So there is only Global component which can be predicted, After making prediction for next six months (For Validation), MAPE value comes up to be **31.45**

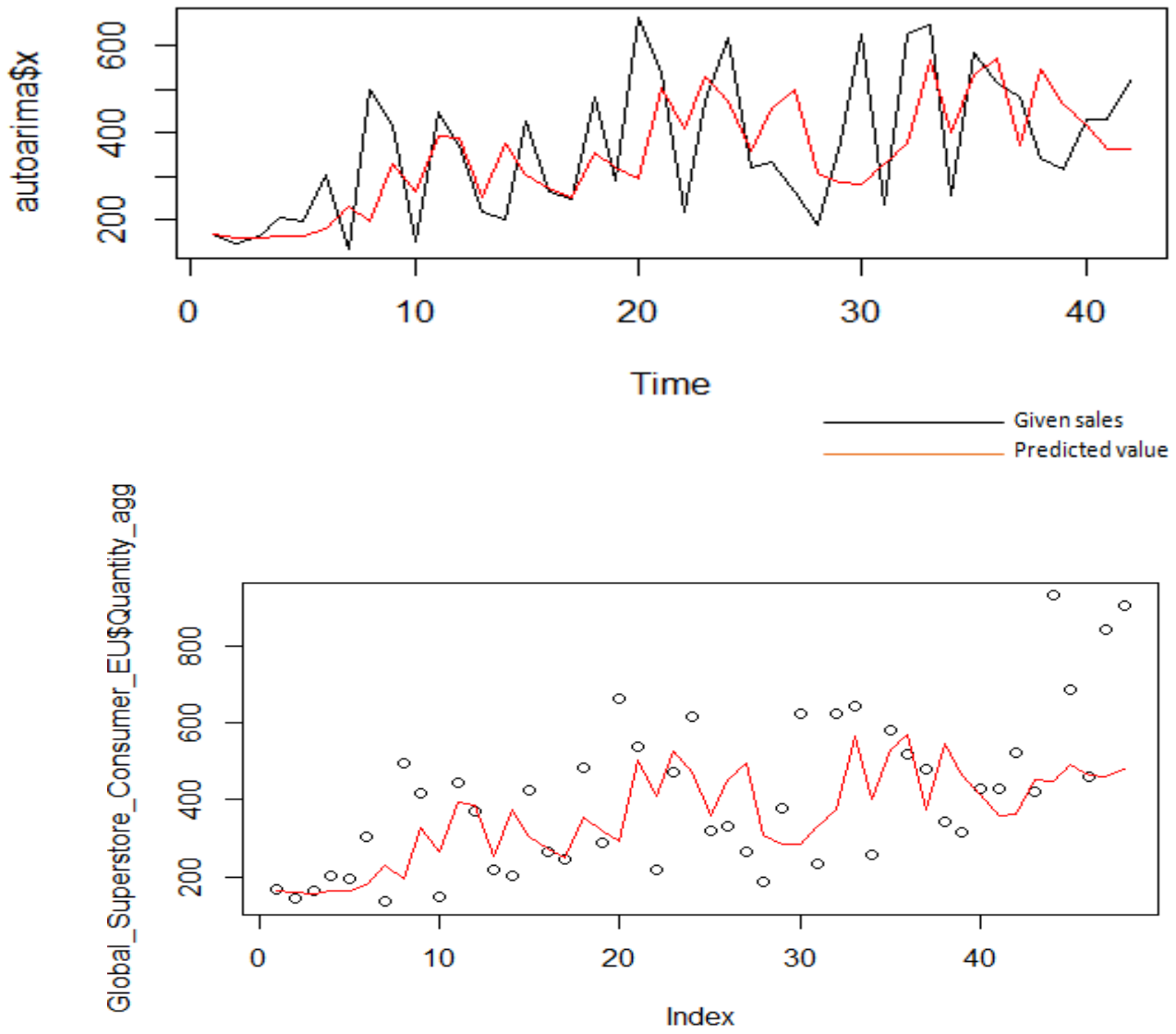
OUTPUT of AUTO ARIMA on EU Consumer Quantity

- Series:

local pred
ARIMA(2,0,0) with zero mean
Coefficients : ar1 ar2
0.6341 -0.6158
s.e. 0.1173 0.1131

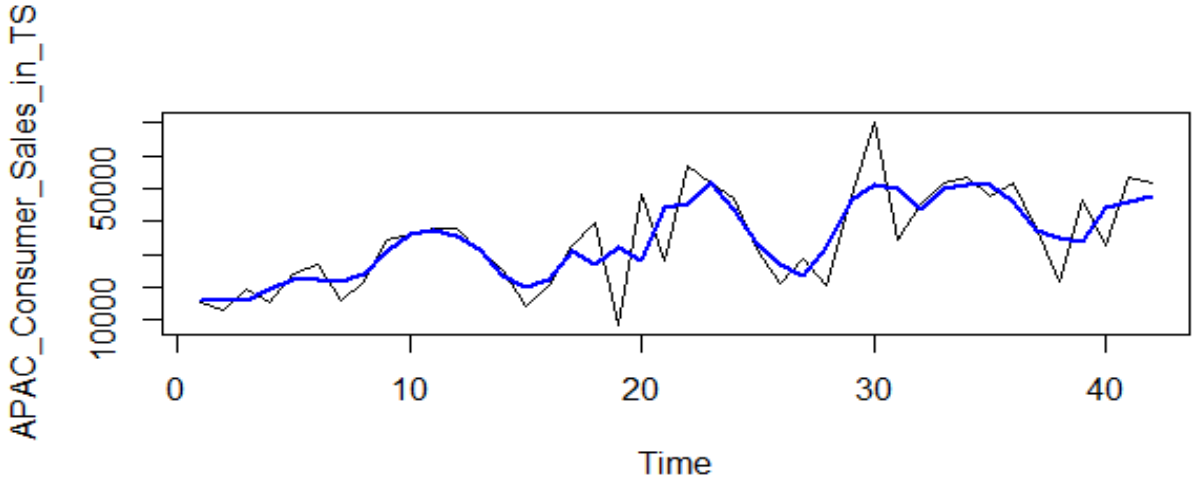
sigma^2 estimated as 7284: log likelihood=-245.89
AIC=497.79 AICc=498.42 BIC=503

Analysis: Time Series for EU Consumer Qty (auto ARIMA)



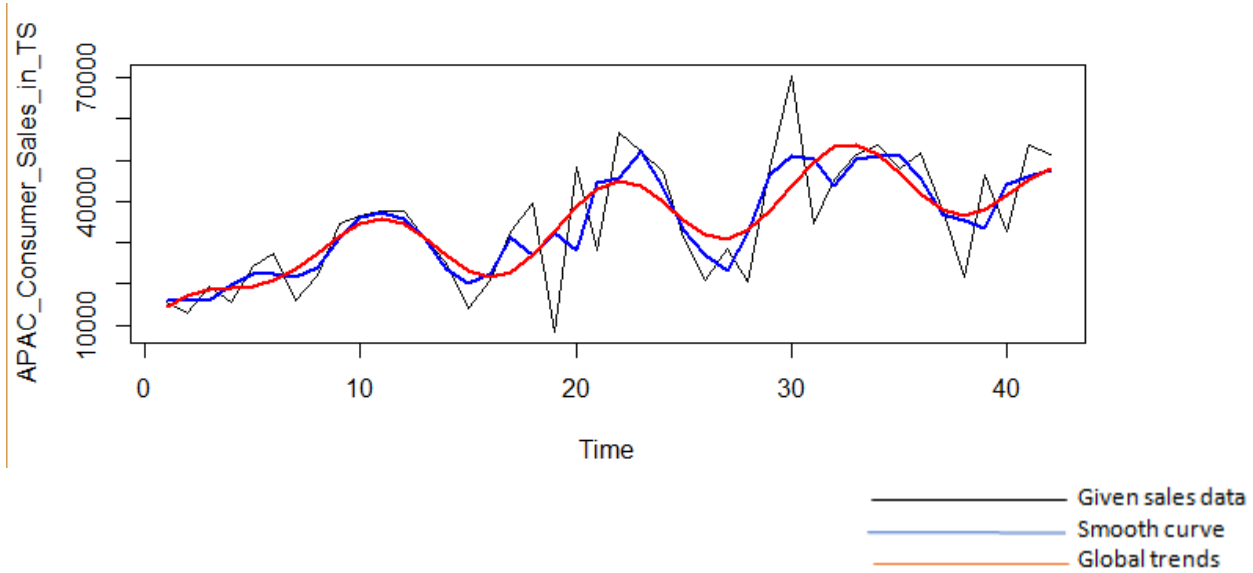
- Now after making prediction for next six months (For Validation), MAPE value comes up to be **30.13**
- Comparing MAPE values of ARMA(Manual) and AUTO ARIMA models , We can see that AUTO ARIMA model is better based on the MAPE value
- So we will use AUTO ARIMA (ARIMA(2,1,0)) for forecasting the next six months sales
- Below is the out come of Forecast
  626.2009 786.6056 842.9179 704.8258 768.6274 807.6497

Analysis: Time Series for APAC Consumer Sales (ARMA)



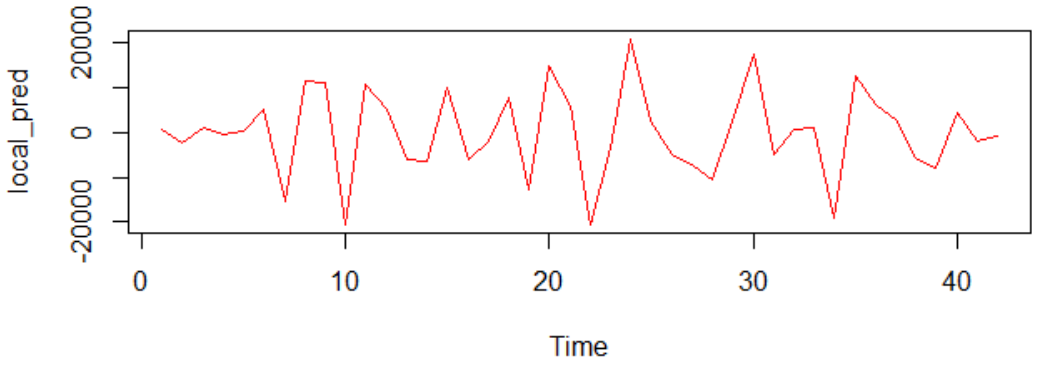
- Amplitude of the seasonal curve doesn't seem increasing with time.
- So, will try fitting additive model for the case.

To fit an additive model with trend and seasonality (Global)

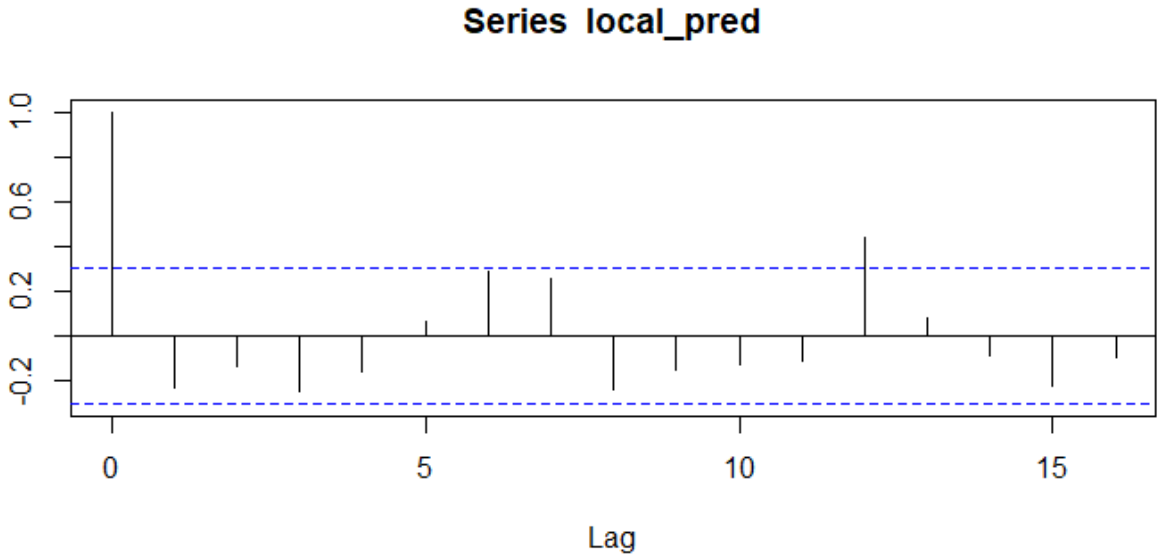


- Different degree (1,2,3,4) of polynomial curves were considered
- Adjusted R-squared for poly degree 3 looks best. So, considered lmfit3 for determining global component

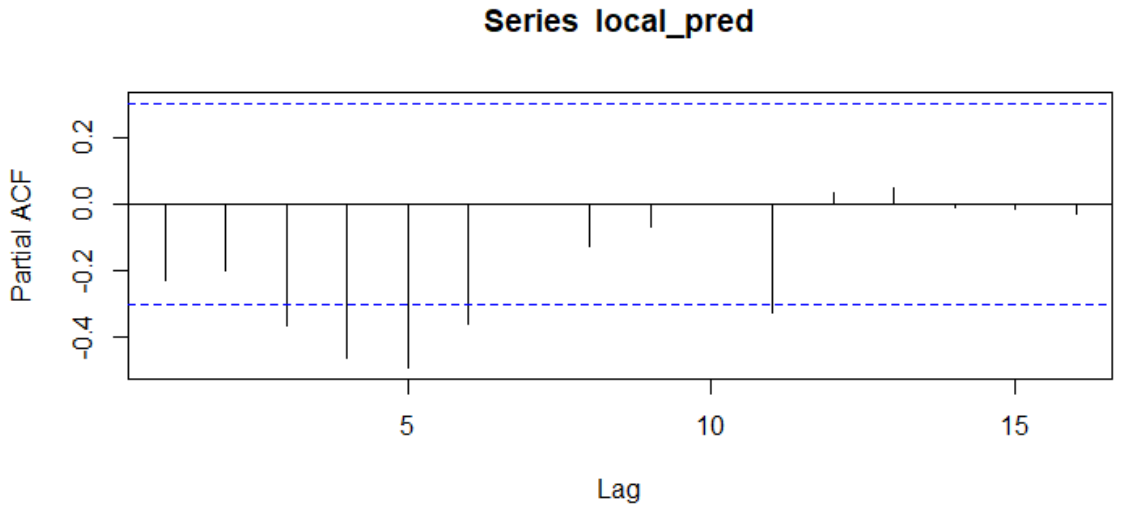
Analysis: Locally predictable series using ARMA for APAC consumer Sales



- Graph shows local component of time series left after removing global component from the given sales data



ACF



P ACF

Evaluate the model using MAPE for APAC Consumer Sales

- Output of Auto Arima on Local component

  Series: local_pred

  ARIMA(0,0,0) with zero mean

  sigma^2 estimated as 8.8e+07: log likelihood=-443.75 AIC=889.49 AICc=889.59 BIC=891.23

- Below are the results of Dickey-Fuller and KPSS test which further confirms that the residual(after ARMA) is only NOISE

  Dickey-Fuller : p-value = 0.01(<0.05) , kpss.test(resi) : p-value = 0.1 (>0.05)

- So there is only Global component which can be predicted, After making prediction for next six months (For Validation), MAPE value comes up to be **31.07**

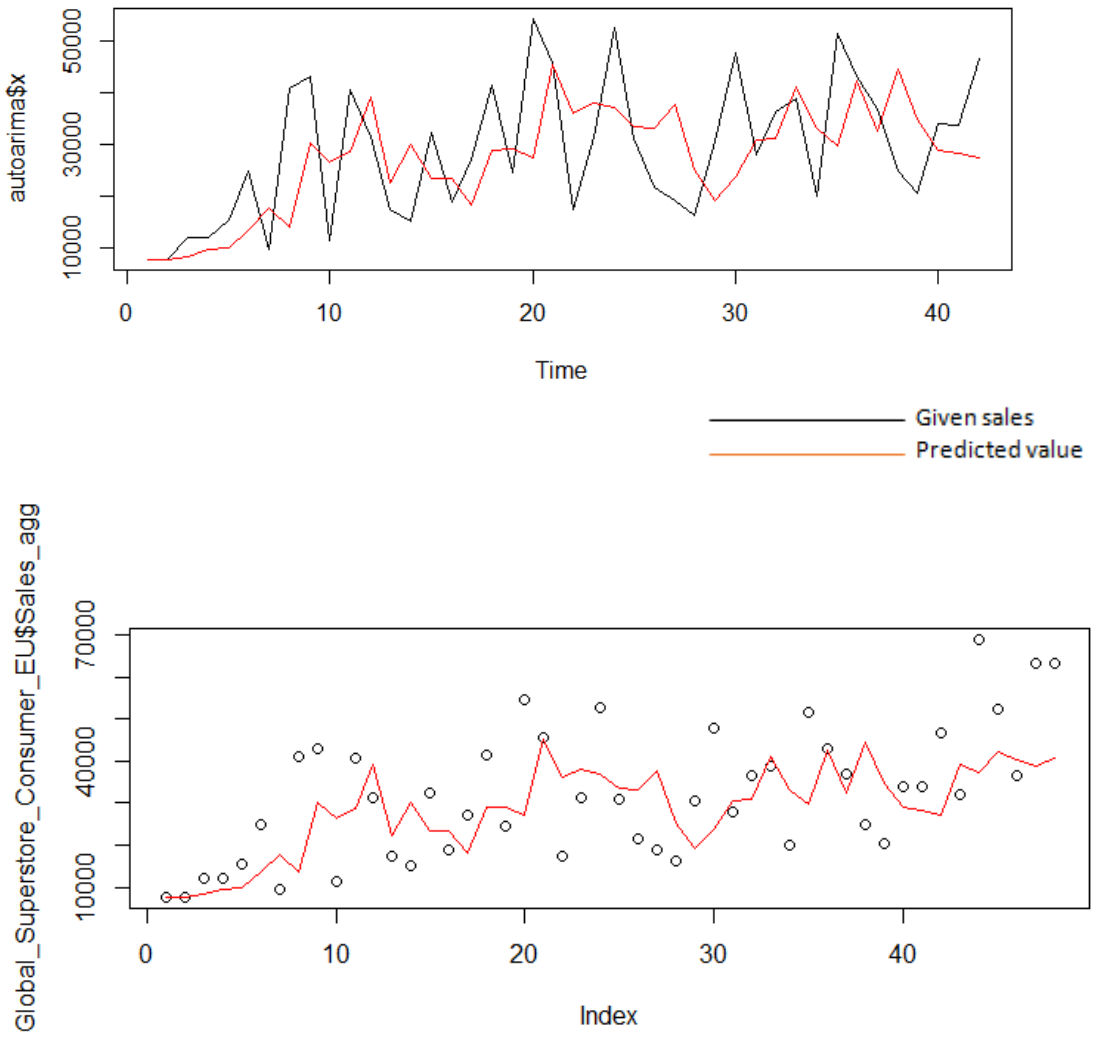
OUTPUT of AUTO ARIMA on APAC Consumer Sales

Series: . ARIMA(2,1,0)

Coefficients:	ar1	ar2
	-0.5796	-0.4906
s.e.	0.1346	0.1310

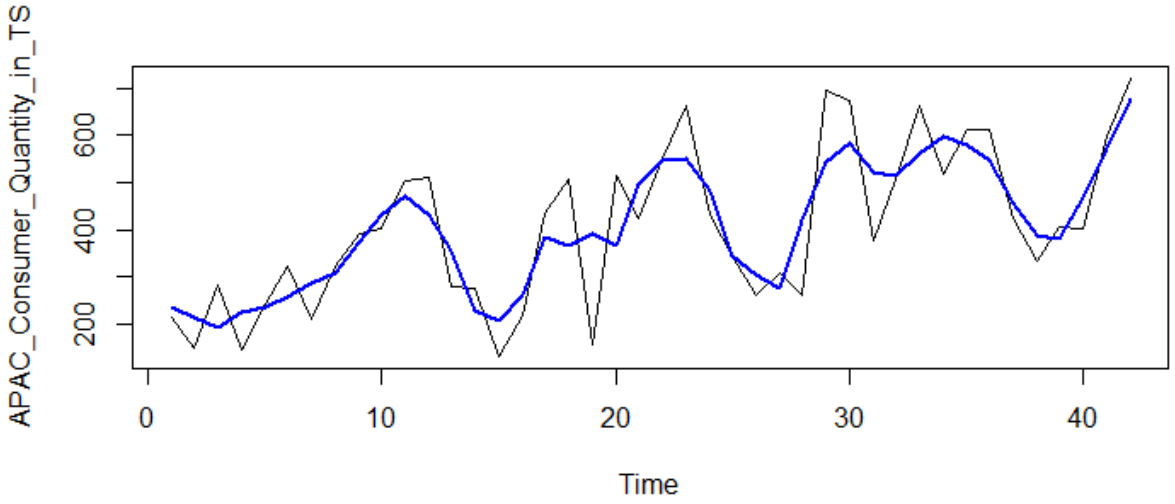
sigma^2 estimated as 168564623: log likelihood=-445.84 AIC=897.67 AICc=898.32 BIC=902.81

Analysis: Time Series for APAC Consumer Sales (auto ARIMA)



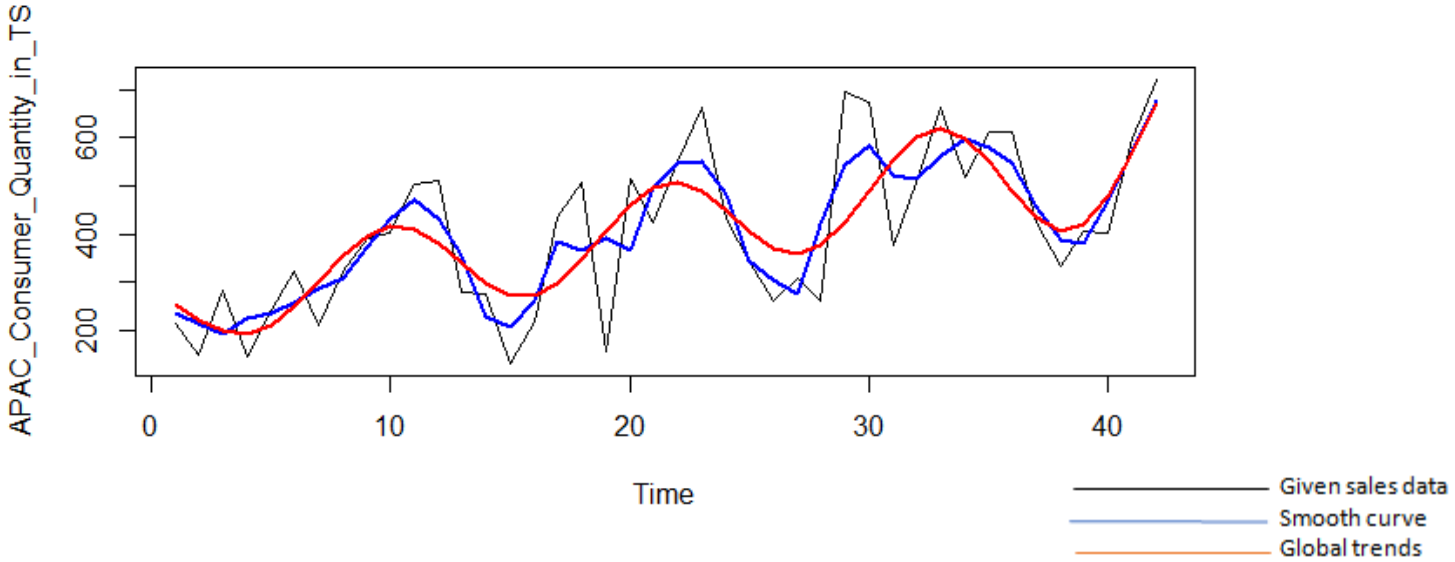
- Now after making prediction for next six months (For Validation), MAPE value comes up to be **28.92**
- Comparing MAPE values of ARMA(Manual) and AUTO ARIMA models , We can see that AUTO ARIMA model is better based on the MAPE value
- So we will use AUTO ARIMA (ARIMA(2,1,0)) for forecasting the next six months sales
- Below is the out come of Forecast
  71649.02 69216.51 68565.57 69356.25 69031.66 69072.60

Analysis: Time Series for APAC Consumer Qty (ARMA)



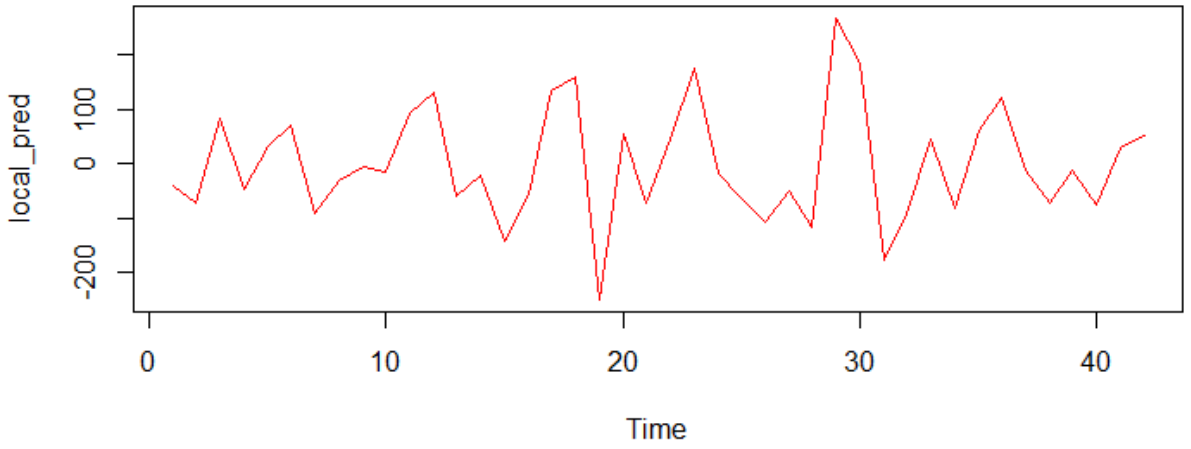
- Amplitude of the seasonal curve doesn't seem increasing with time.
- So, will try fitting additive model for the case.

To fit an additive model with trend and seasonality (Global)

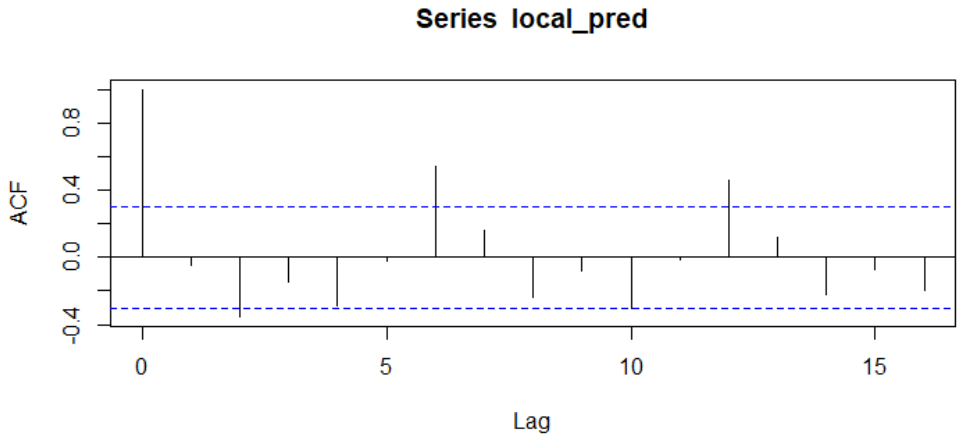


- Different degree (1,2,3,4) of polynomial curves were considered
- Adjusted R-squared for poly degree 2 looks best. So, considered lmfit2 for determining global component

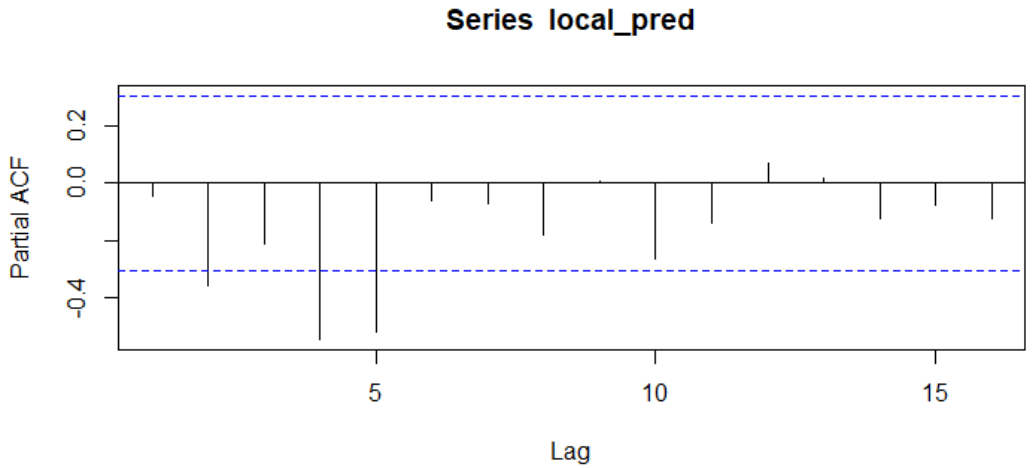
Analysis: Locally predictable series using ARMA for APAC consumer Qty



- Graph shows local component of time series left after removing global component from the given sales data



ACF



P ACF

Evaluate the model using MAPE for APAC Consumer Quantity

- Output of Auto Arima on Local component

  Series: local_pred
  ARIMA(0,0,0) with zero mean
  sigma^2 estimated as 8.8e+07: log likelihood=-443.75 AIC=889.49 AICc=889.59 BIC=891.23

- Below are the results of Dickey-Fuller and KPSS test which further confirms that the residual(after ARMA) is only NOISE

  Dickey-Fuller : p-value = 0.01(<0.05) , kpss.test(resi) : p-value = 0.1 (>0.05)

- So there is only Global component which can be predicted, After making prediction for next six months (For Validation), MAPE value comes up to be **62.10**

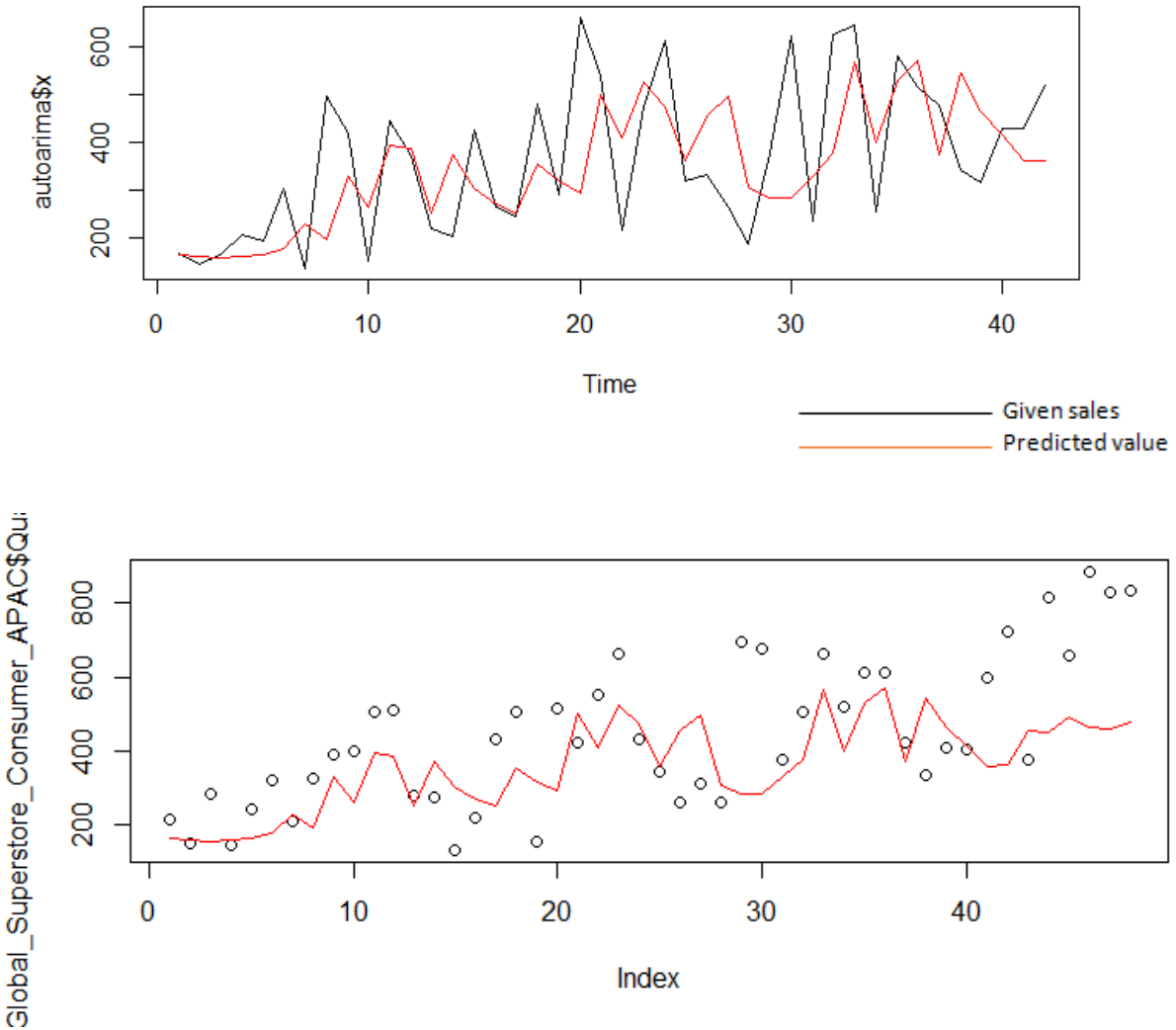
OUTPUT of AUTO ARIMA on APAC Consumer Quantity

Series: . Series: . ARIMA(2,1,0)

Coefficients:	ar1	ar2
	-0.7359	-0.5879
s.e.	0.1224	0.1185

sigma^2 estimated as 21185: log likelihood=-261.9 AIC=529.8 AICc=530.44 BIC=534.94

Analysis: Time Series for APAC Consumer Sales (auto ARIMA)



- Now after making prediction for next six months (For Validation), MAPE value comes up to be **37.54**
- Comparing MAPE values of ARMA(Manual) and AUTO ARIMA models , We can see that AUTO ARIMA model is better based on the MAPE value
- So we will use AUTO ARIMA (ARIMA(2,1,0)) for forecasting the next six months sales
- Below is the out come of Forecast
  842.6532 837.2847 837.8234 838.6757 838.1635 838.2291

Conclusions

- The two most profitable buckets out of 21 are EU, Consumer and APAC, Consumer segment.
- The MAPE values of all the predictions are low.
- The multiplicative model has been followed for modelling out the globally predictable part for both the segments
- The ACF,PACF plots for all the segments shows that the locally predictable part is itself weakly stationary and residuals for all the segments came out to be pure white noise.