ECOMMERCE SALES FORECAST

Business understanding:

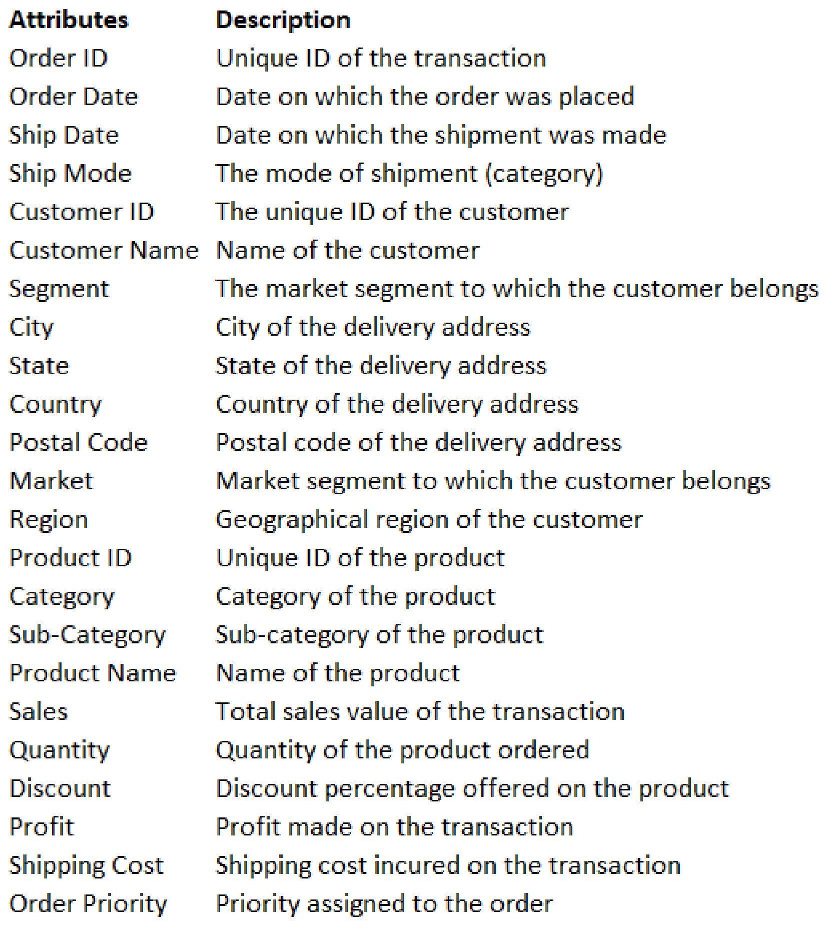
Sales and Operations manager needs to finalize the plan for their e-commerce business. They seek to forecast the sales and demand for next 6 months, that would help them manage revenue and inventory accordingly. The store caters to 7 different market segments and in 3 major categories. We need to forecast at this granular level, so we would need to retrieve this data as 21 (7x3) buckets before analyzing.

Not all 21 market buckets are important, so we will find out 5 most profitable (and consistent) segments from these 21 and forecast the sales and demand for these segments.

Data Understanding:

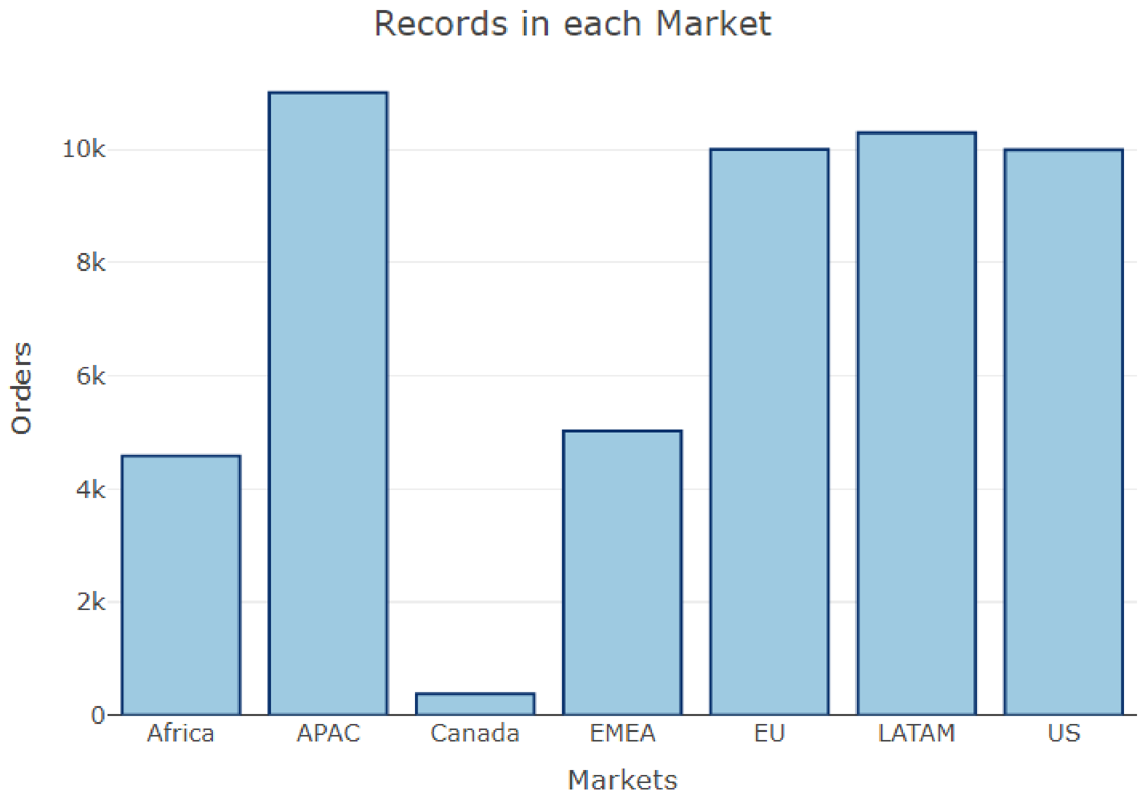
1. Dataset contains transactional-level data of customer orders of an e-commerce store
2. Each row represents a customer order
3. Total Records: 51,290
4. Total Attributes: 24

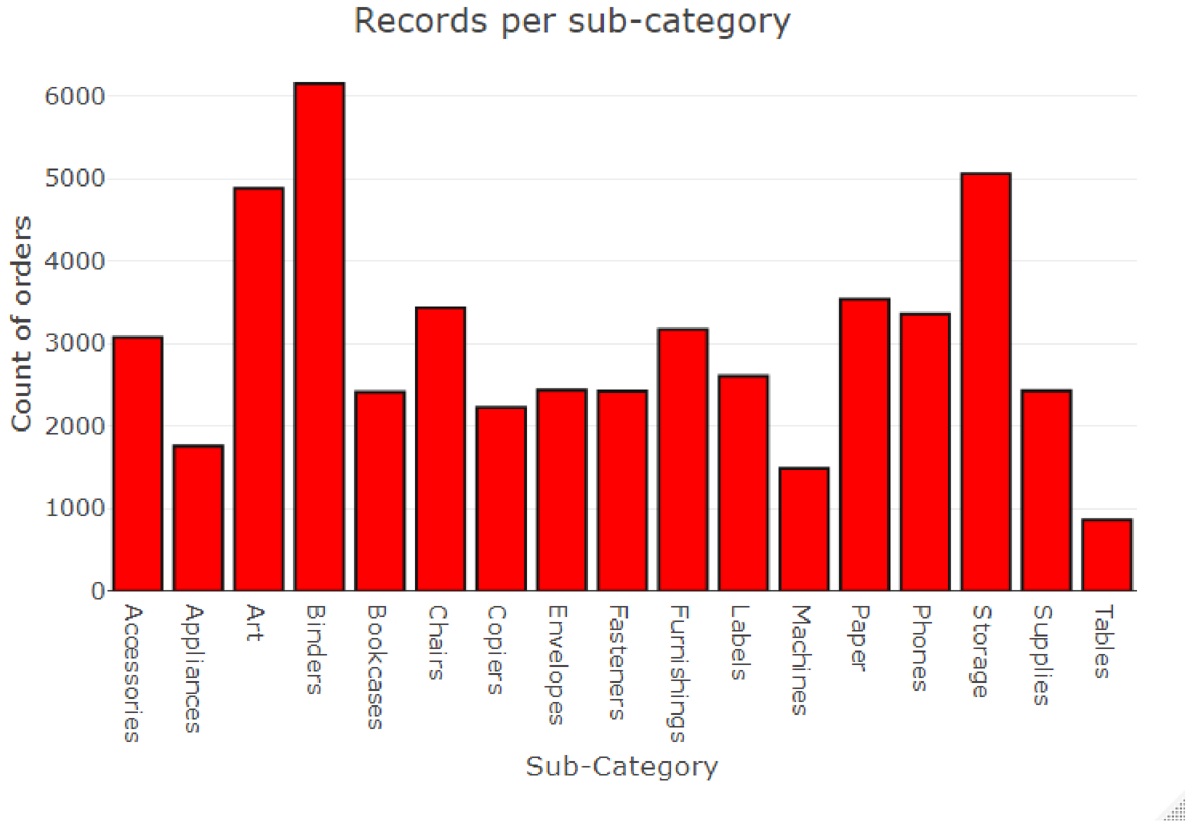
Data Dictionary:



Attributes	**Description**
Order ID	Unique ID of the transaction
Order Date	Date on which the order was placed
Ship Date	Date on which the shipment was made
Ship Mode	The mode of shipment (category)
Customer ID	The unique ID of the customer
Customer Name	Name of the customer
Segment	The market segment to which the customer belongs
City	City of the delivery address
State	State of the delivery address
Country	Country of the delivery address
Postal Code	Postal code of the delivery address
Market	Market segment to which the customer belongs
Region	Geographical region of the customer
Product ID	Unique ID of the product
Category	Category of the product
Sub-Category	Sub-category of the product
Product Name	Name of the product
Sales	Total sales value of the transaction
Quantity	Quantity of the product ordered
Discount	Discount percentage offered on the product
Profit	Profit made on the transaction
Shipping Cost	Shipping cost incured on the transaction
Order Priority	Priority assigned to the order

Exploratory Data Analysis:





Data preparation:

1. First the dataset was segmented into 21 subsets based on the market and the customer segment level
2. Transaction-level data was converted into a time series
3. The 3 attributes - Sales, Quantity & Profit were aggregated over the Order Date to arrive at monthly values for these attributes
4. When the 3 timeseries for each of the 21 segments were derived, 5 most profitable and consistently profitable segments were chosen based on Total Profit and

coefficient of variation of the Profit for all 21 market segments

Coefficient of Variation in Corporate Finance:

Formula: **CV = [STD DEV / MEAN] x 100**

Example: If the SPDR S&P 500 ETF has an average annual return of 5.47% and a standard deviation of 14.68%, the SPDR S&P 500 ETF's coefficient of variation is 2.68

Reference: https://www.investopedia.com/terms/c/coefficientofvariation.asp

Top 5 Segments balancing Profit and Coefficient of Variation:

# APAC Consumer
# EU Consumer
# APAC Corporate
# EU Corporate
# LATAM Consumer

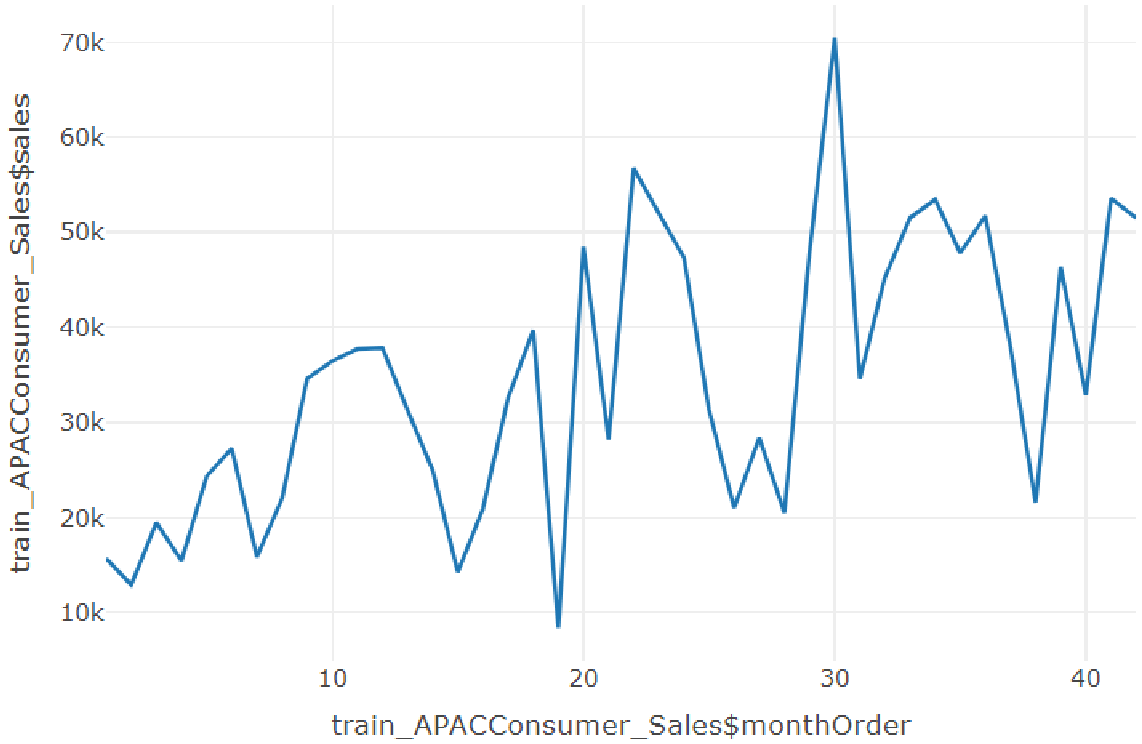
Model Building:

Defined Task: Forecast the sales and quantity for the next 6 months

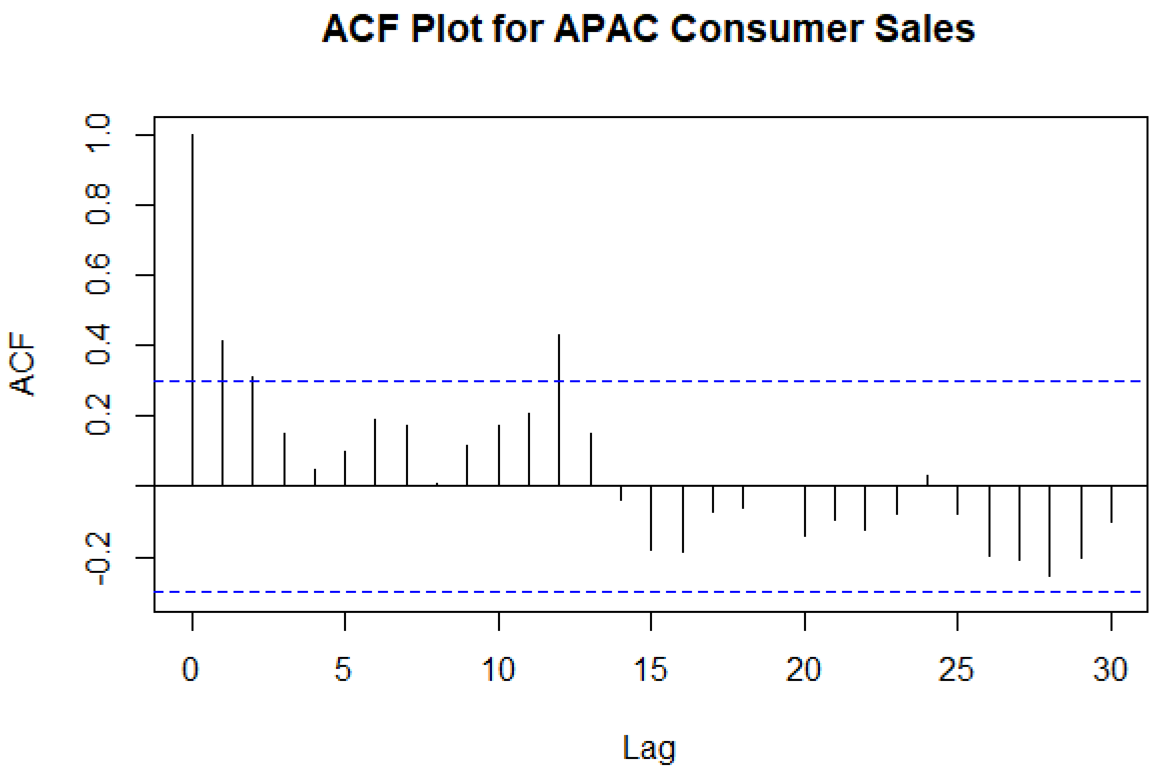
Shown below is the **Time Series Analysis** of **Sales** for **CONSUMER** Segment in **APAC** region. Similar analysis was derived for the rest of the subsets (refer code).

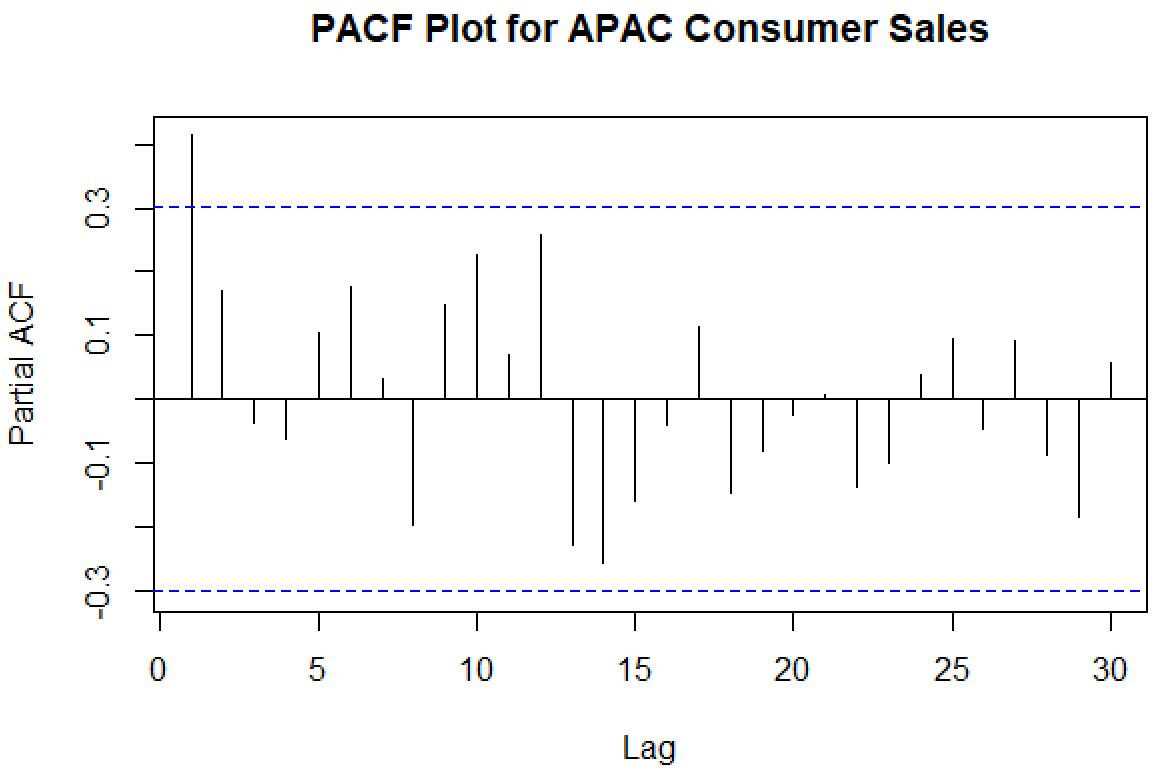
Analysis Steps:

1. Plot the time series for sales/quantity



Plot shows us a slow upward moving trend with lots of ups and downs





2. Smoothen the series using any of the smoothing techniques. This would help identify the trend/seasonality component

Analysis of time series - Manual decomposition and Auto-ARIMA

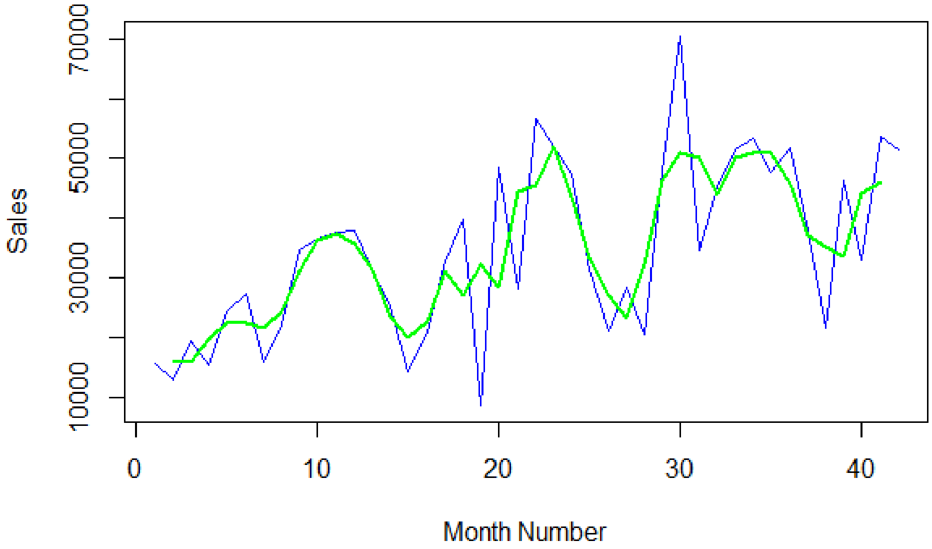
Manual Decomposition was performed first:

Smoothing technique used: Convolution (Moving Average)
Filter(window) = 2*w+1 is fitted in this method

Overlayed plot of smoothed series

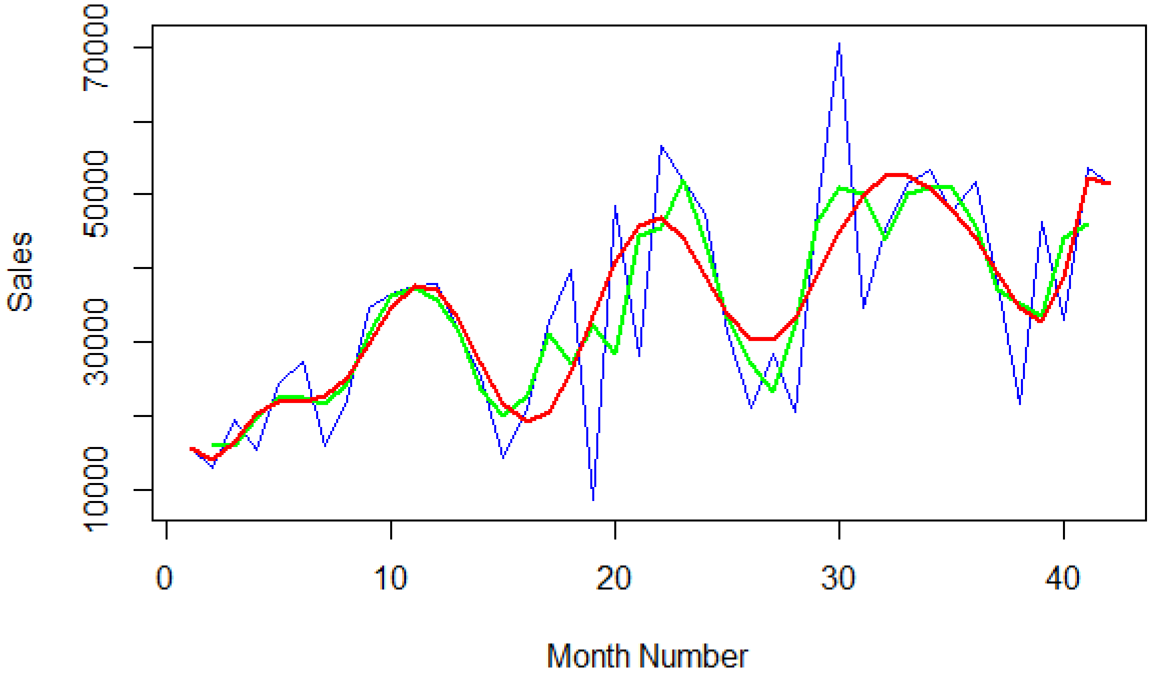
Blue: original TS

Green: smoothed TS



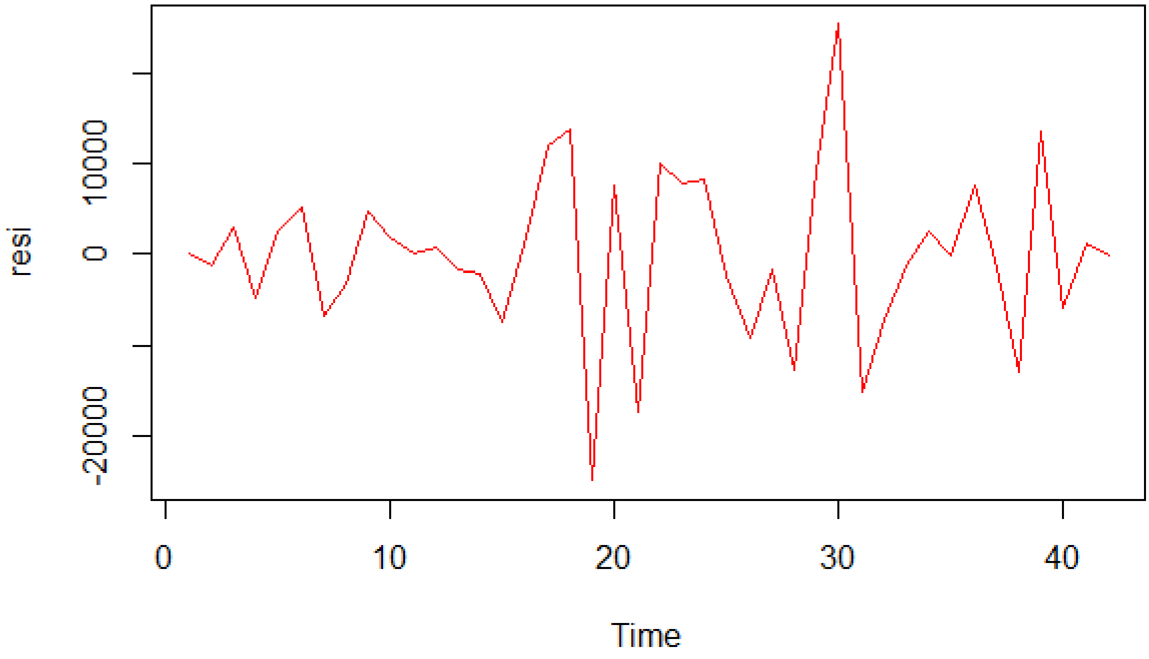
3. Used feature engineering to come up with the best regression fit

Fitting the TREND LINE / Regression line

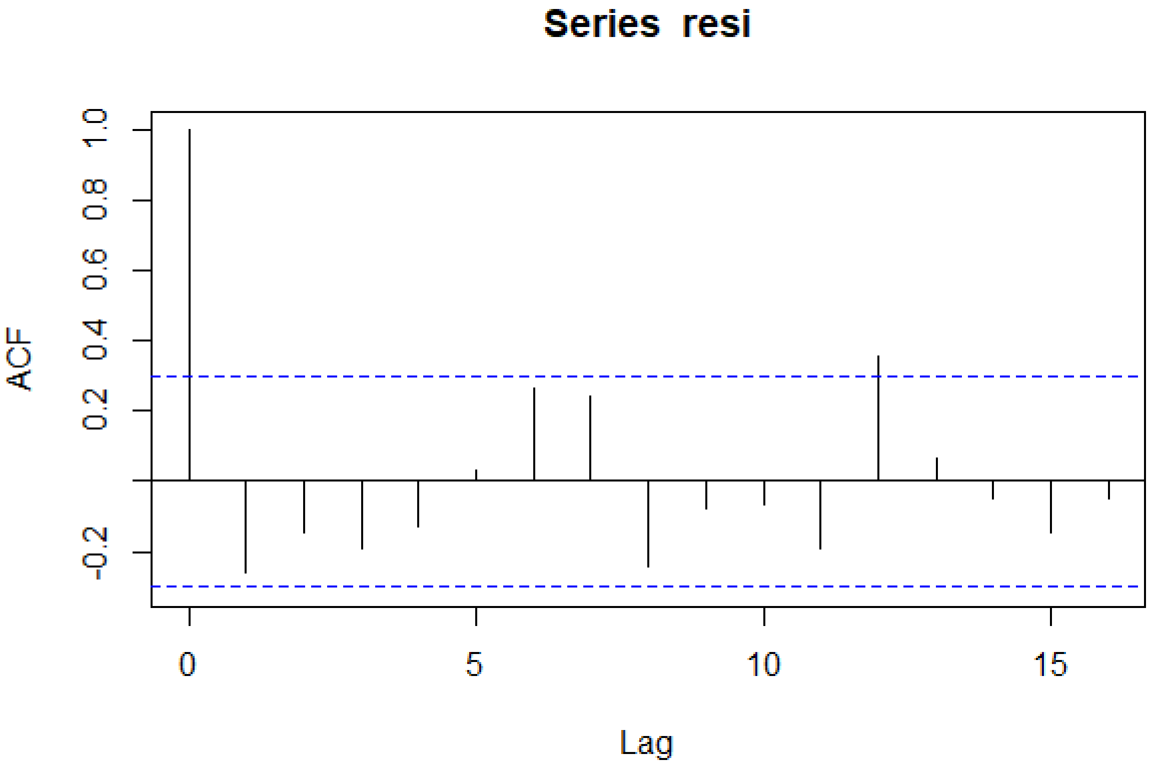


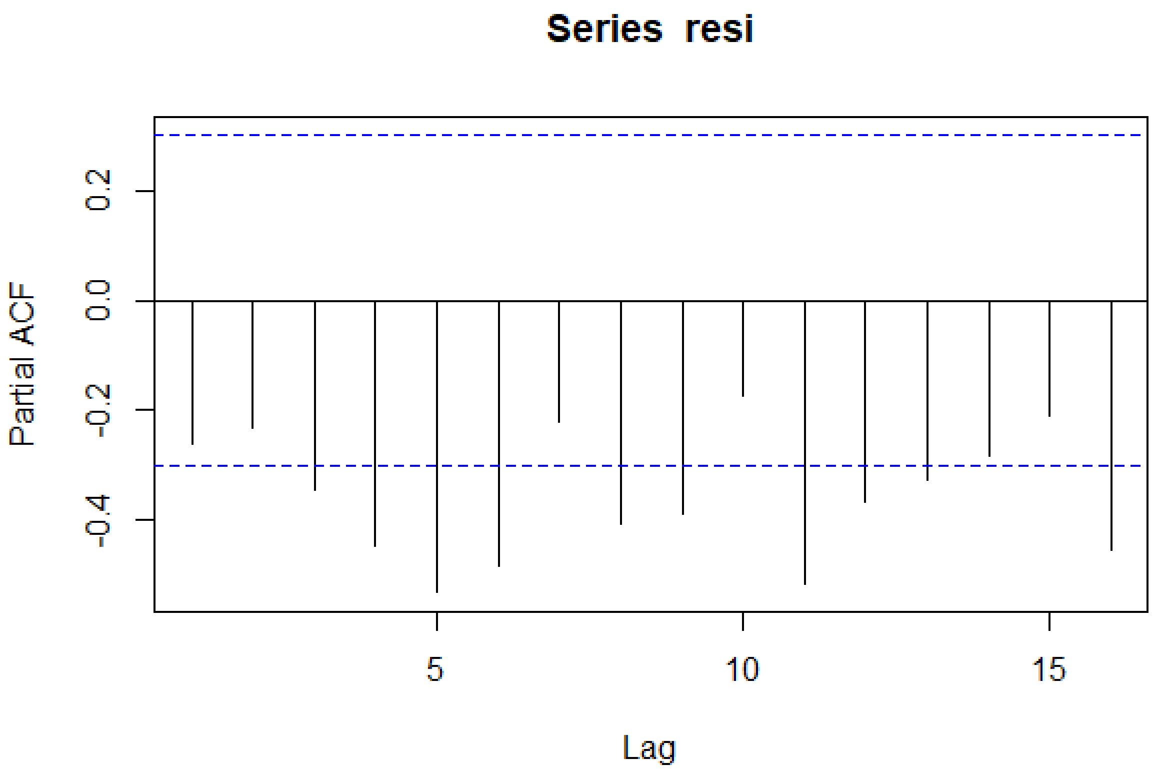
4. Checking the residual series for White noise

Residue - Taking the predictable part out of the Time Series leaves us with the residual TS



ACF and PACF (same as ACF with type = partial) plots show many lines nearly or completely out of the confidence interval (dotted blue lines) indicating noise

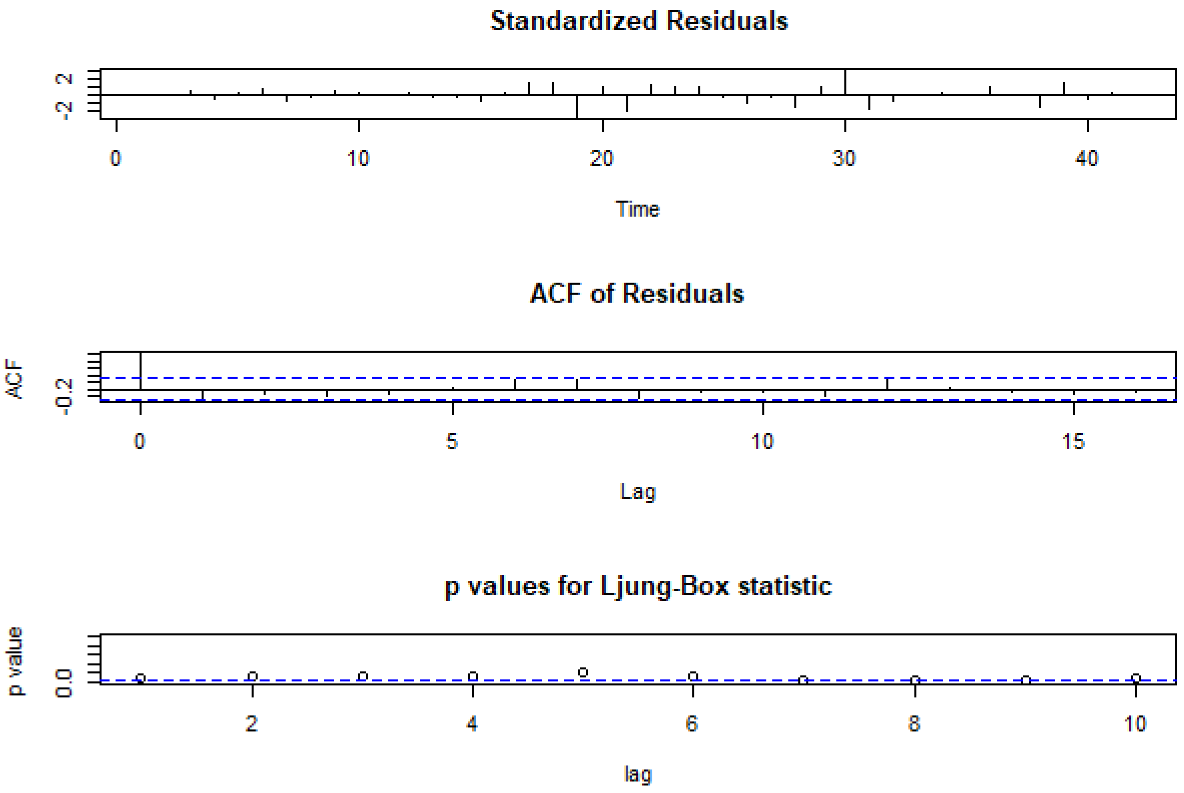




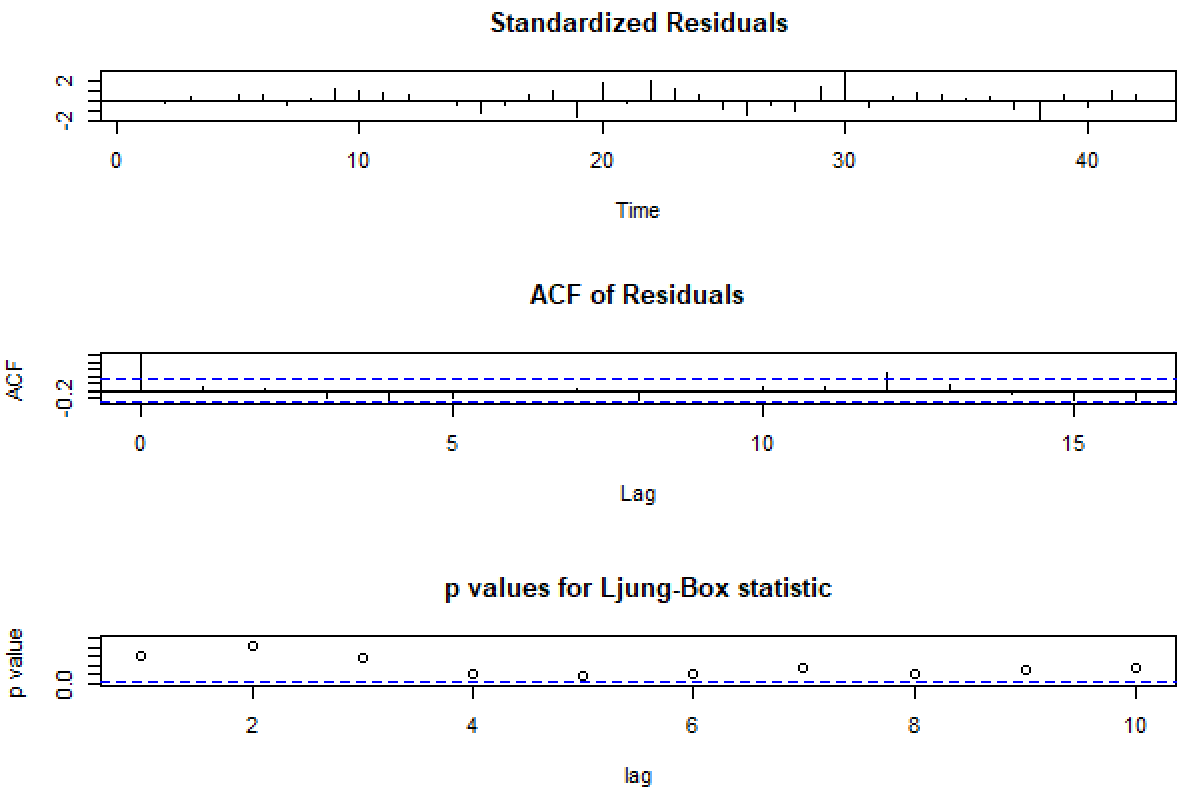
5. Find the optimal value of p,d,q for ARIMA modelling

ARMA fit

```
# ARIMA(0,0,0) with zero mean
# sigma^2 estimated as 83420737:  log likelihood=-442.62
# AIC=887.25   AICc=887.35   BIC=888.98
# p,d,q : (0,0,0)
```

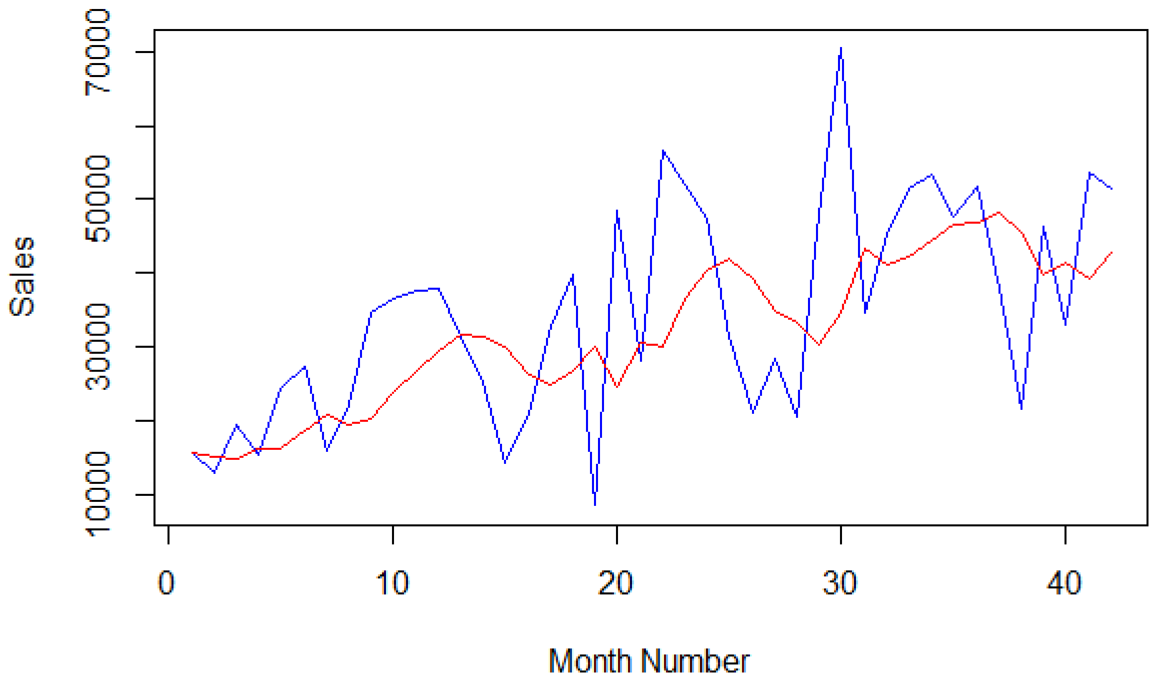


Auto ARIMA on original time series



```
# ARIMA(0,1,1)
# sigma^2 estimated as 174361555:  log likelihood=-447.11
# AIC=898.23   AICc=898.55   BIC=901.66
# p,d,q : (0,1,1)
```

Autoarima plot overlayed on original time series



# **Log likelihood, AIC, AICc, BIC** values indicate that manual model is clearly better than the Auto generated one in terms of all the parameters

Model Evaluation:

MAPE values (forecasted values using out-of-sample data). **Mean absolute percentage error** is commonly used as a loss function for regression problems and in model evaluation, because of its very intuitive interpretation in terms of relative error. **MAPE** expresses accuracy as a percentage of the error. Because the MAPE is a percentage, it can be easier to understand than the other accuracy measure statistics. For example, if the MAPE is 5, on average, the forecast is off by 5%.

NOTE: We know for sure that there are no data points for which there are zero sales, so we are safe to use MAPE. Remember that we must interpret it in terms of percentage points.

READ MORE AT:
https://www.dataquest.io/blog/understanding-regression-error-metrics/
https://stats.stackexchange.com/questions/327464/mape-vs-r-squared-in-regression-models

Time Series Analysis of Sales for CONSUMER Segment in APAC marketplace
# MAPE for Regression: 28.68488
# MAPE for Auto-Arima: 27.68952

Time Series Analysis of Quantity for CONSUMER Segment in APAC marketplace
# MAPE for Regression: 37.13402
# MAPE for Auto-Arima: 26.24458

Time Series Analysis of Sales for Consumer Segment in EU marketplace
# MAPE for Regression: 22.71581
# MAPE for Auto-Arima: 28.9226

Time Series Analysis of Quantity for Consumer Segment in EU marketplace
# MAPE for Regression: 29.37811
# MAPE for Auto-Arima: 30.13319

Time Series Analysis of Sales for CORPORATE Segment in APAC marketplace
# MAPE for Regression: 26.55771
# MAPE for Auto-Arima: 27.97408

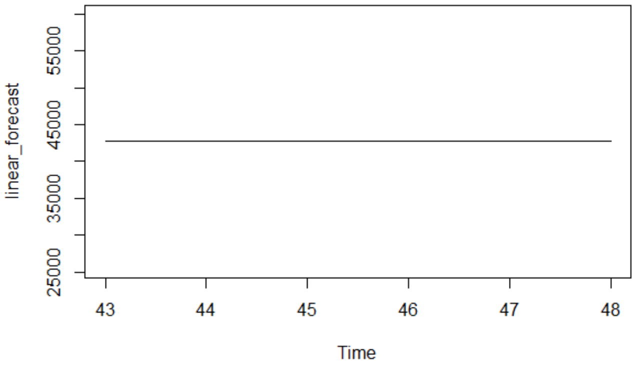
Time Series Analysis of Quantity for CORPORATE Segment in APAC marketplace
# MAPE for Regression: 27.67883
# MAPE for Auto-Arima: 24.13219

Time Series Analysis of Sales for CORPORATE Segment in EU marketplace
# MAPE for Regression: 79.79463
# MAPE for Auto-Arima: 36.35092

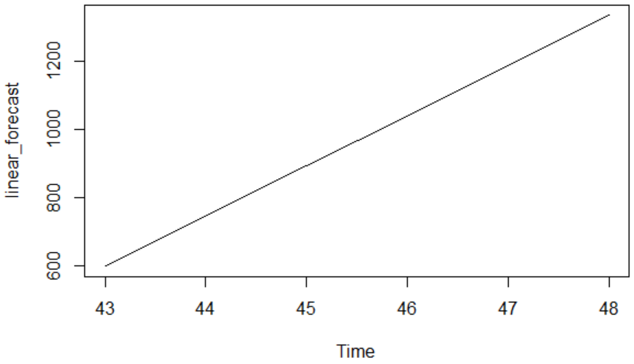
Time Series Analysis of Sales for Consumer Segment in LATAM marketplace
# MAPE for Regression: 31.66988
# MAPE for Auto-Arima: 33.96611

Forecasts:

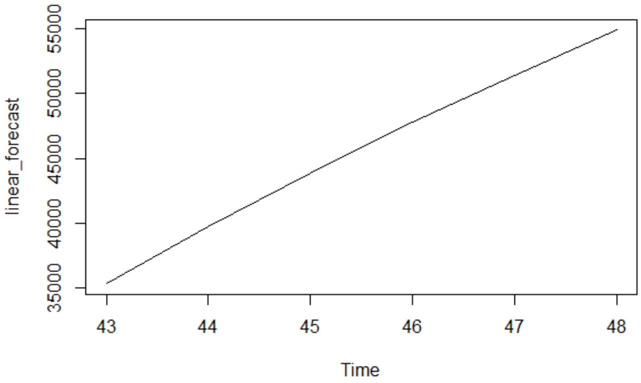
Time Series Analysis of Sales for CONSUMER Segment in APAC marketplace



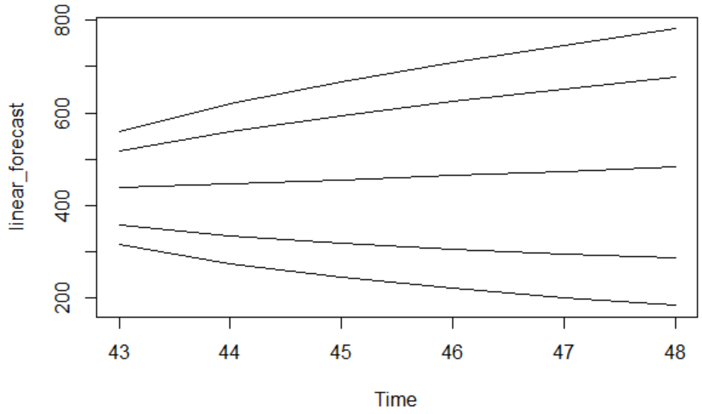
Time Series Analysis of Quantity for CONSUMER Segment in APAC marketplace



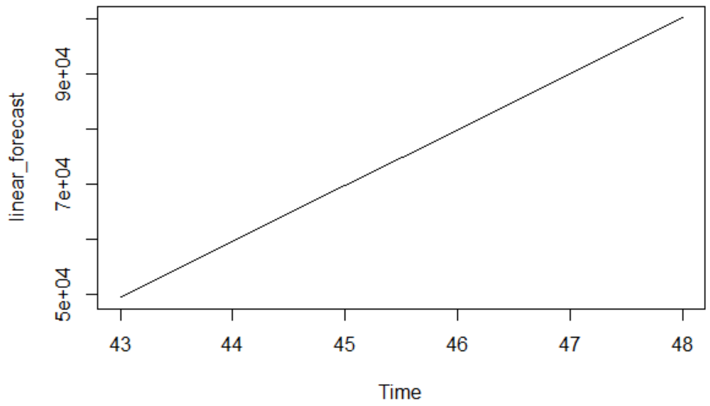
Time Series Analysis of Sales for Consumer Segment in EU marketplace



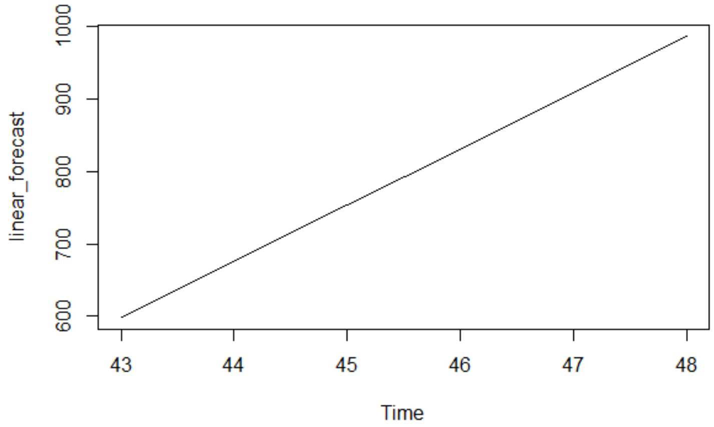
Time Series Analysis of Quantity for Consumer Segment in EU marketplace



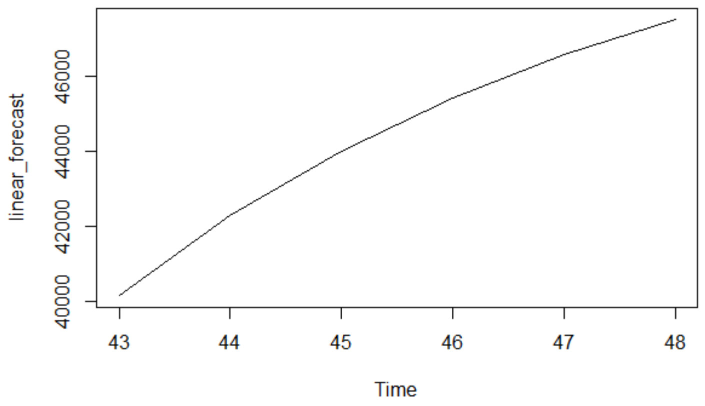
Time Series Analysis of Sales for CORPORATE Segment in APAC marketplace



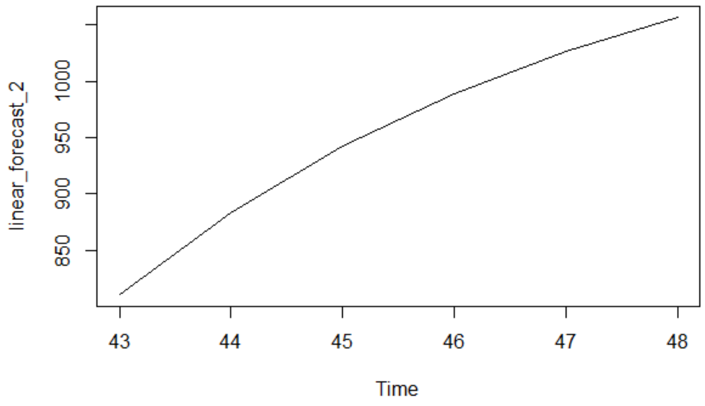
Time Series Analysis of Quantity for CORPORATE Segment in APAC marketplace



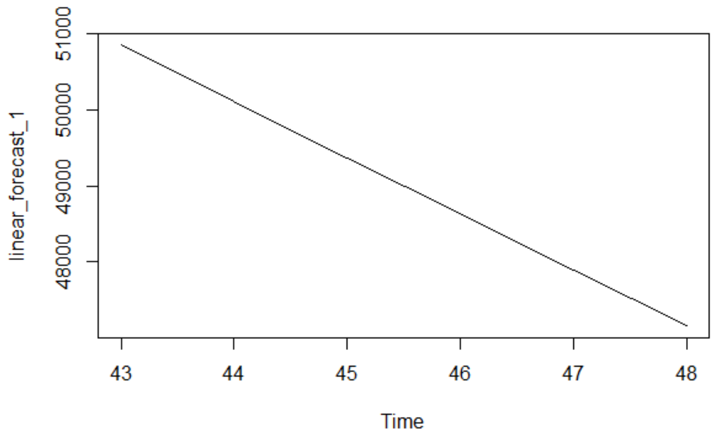
Time Series Analysis of Sales for CORPORATE Segment in EU marketplace



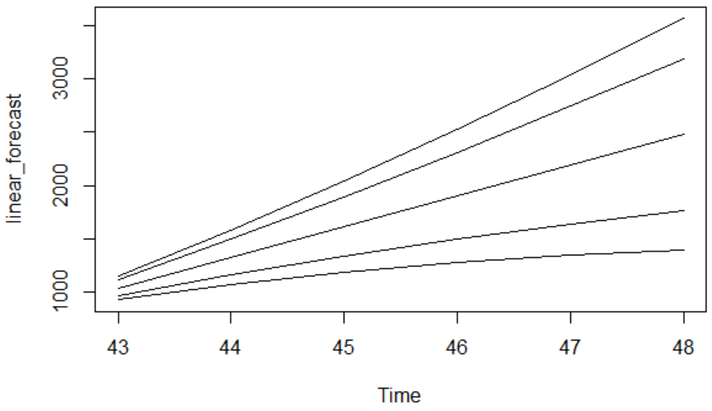
Time Series Analysis of Quantity for CORPORATE Segment in EU marketplace



Time Series Analysis of Sales for Consumer Segment in LATAM marketplace



Time Series Analysis of Quantity for Consumer Segment in LATAM marketplace



Conclusions:

Highlighted segments were selected to calculate the average MAPE value for model evaluation. MAPE states that our model's predictions are, on average, 27.77% off from actual value in Regression model and 29.38% off from actual value in case of Auto-Arima model.

According to the next 6 months' forecast, below are the insights for revenue and resource allocation:

1. Total Sales for CONSUMER Segment in LATAM marketplace may decline from 51K USD to 47K USD
2. Total Sales for CONSUMER Segment in APAC remain steady ~45K USD
3. Total Sales for CONSUMER Segment in EU will rise from 35K USD to 55K USD
4. Total Sales figure for CORPORATE Segment in EU shows a slow rise from 40K USD to 46K USD
5. There seems to be a proportional increase in Total Quantity Sold with the Total Sales for each split, except for LATAM where Total Quantity Sold increased 3-fold with about 5.8% decrease in Total Sales in the next 6 months. Possible reason behind this could be an increase in discounts to clear the inventory in LATAM
6. Company also needs to stock up on the inventory where quantity ordered shows big jump in the forecast. Detailed analysis at sub-category level data is needed to identify the product lines showing such increase.