RETAIL GIANT SALES FORECASTING

USING TIMER SERIES ANALYSIS

Group Members:

1. Vijayanand Narayanan
2. Arunachalam Meenakshisundaram
3. Akash Ashokan
4. Dharamarajan Thiyagarajan

Business Problem

Global Mart is an online retail store that takes orders and delivers worldwide. The company caters to 7 Markets (Africa, APAC, Canada, EMEA, EU, LATAM and US) and 3 Segments (Consumer, Corporate and Home Office).

The company wishes to finalise the operations plan for the next 6 months. So, it is required to forecast the sales and the demand for the next 6 months, that would help the company manage the revenue and inventory accordingly.

Objectives

Carry out research and conduct analysis using the store's past data to fulfil the following objectives,

- Identify top 2 consistently profitable segments using Coefficient of Variation of profit
- Forecast Sales and Demand for next 6 months in future using Classical Decomposition and Auto ARIMA methods

Problem Solving Methodology

- Understand the business problem
- Define project objectives and expected outcomes
- Use dataset that has been provided
- Clean data and address all data quality issues
- Perform Exploratory Data Analysis by creating buckets based on Market and Segment
- Select the top 2 consistently profitable segments using Coefficient of Variation of profit
- Create Training and Test datasets
- Develop a Time Series model using Classical Decomposition and then using ARIMA to compare and contrast the sales and demand predictions
- Fine tune model to best fit
- Select the appropriate model based on MAPE performance metrics
- Predict Sales and Demand for the last 6 months using test data
- Verify quality of prediction
- Predict values for next 6 months in future

Data

Data from Global Superstore.csv dataset was used to carry out the time series analysis.

1. There were 51290 Observations and 25 Variables
2. Variables predominantly used in the analysis were Market, Segment, OrderDate, Sales, Quantity and Profit
3. A new variable YearMonth was created by extracting value from OrderDate in order to generate monthly summary of Sales, Demand and Profit

Data Cleaning

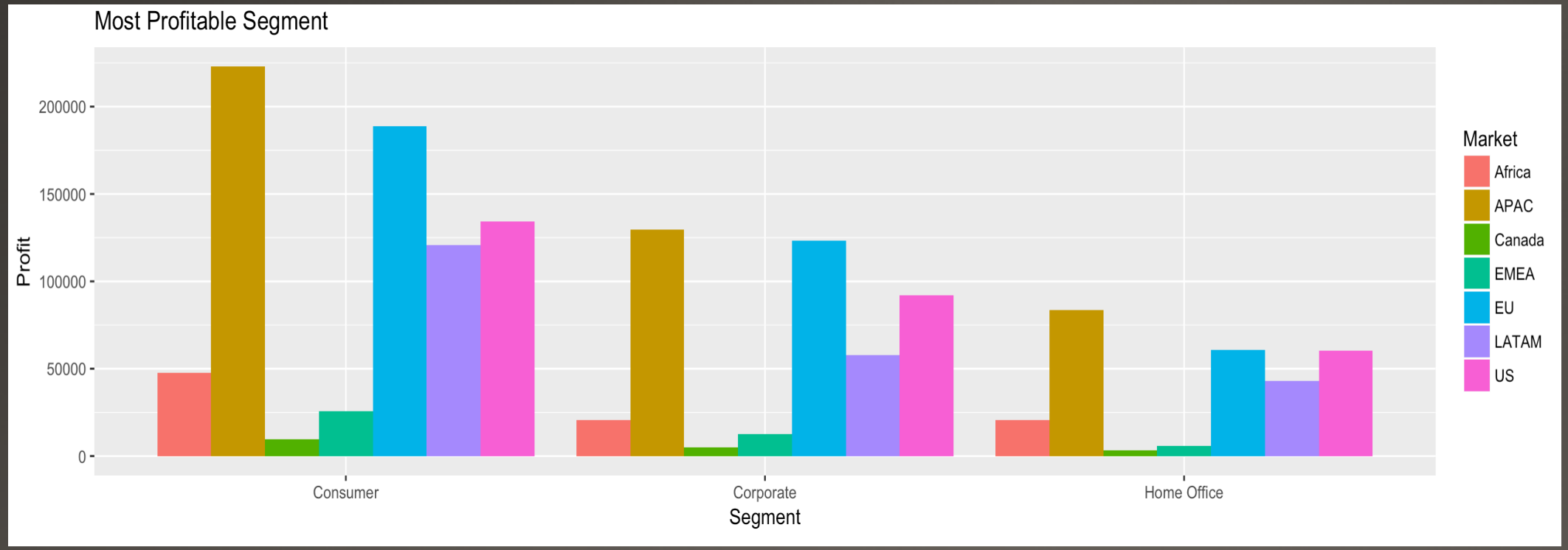
Following steps were taken to clean up the super store data

1. Checked for duplicate records
2. Handled missing values. Postal.Code variable had 41296 missing values. It was confirmed with additional checks that postal code was populate only for all US orders and there was a missing value for other countries
3. Converted character to date format. Both Order.Date and Ship.Date were converted to Date format

Data Preparation

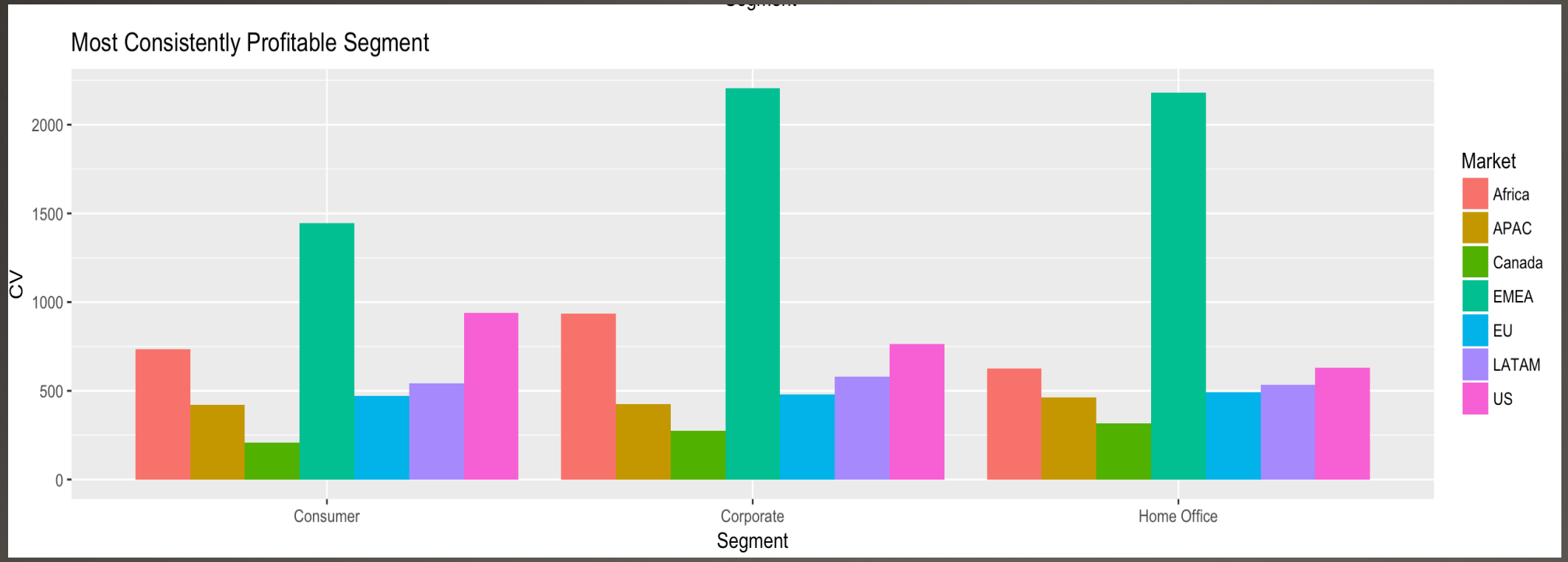
1. There were 7 Markets and 3 Segments. Created 21 buckets for Market and Segments
2. Aggregated Profit for each market segment over all dates to to identify top 2 most profitable segments
3. Calculated CV of Profit i.e Std Dev(Profit)*100/ Mean(Profit) and aggregated CV of Profit for each market segment over all dates to identify top 2 most consistently profitable segments
4. Aggregated Sales, Quantity & Profit over the Order Date to arrive at monthly values for these attributes to use for Time Series Analysis

Top 2 Profitable Segments



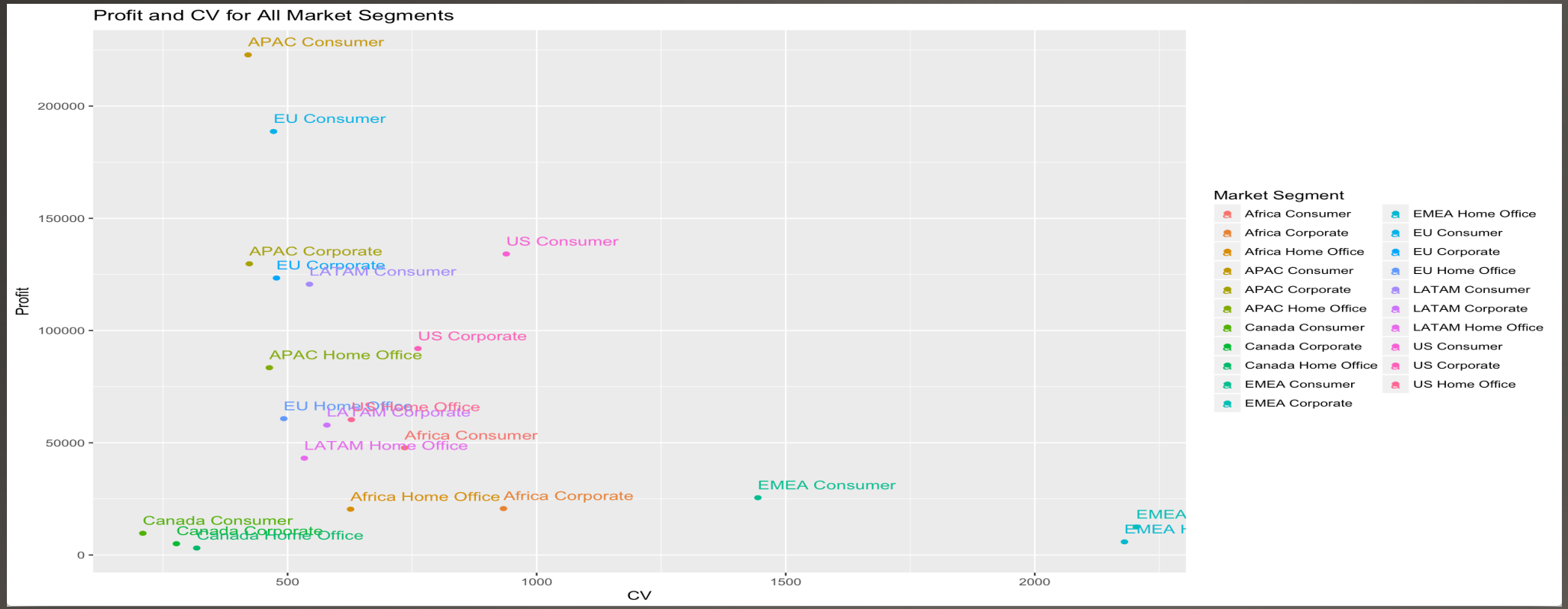
- APAC Consumer and EU Consumer segments were the top 2 profitable segments based on net profit

CV of Segments



- Canada Consumer and Canada Corporate segments have the least Coefficient of Variation of profit
- APAC and EU Consumer segments have low CV values

Top 2 Consistently Profitable Segments



- APAC Consumer and EU Consumer segments were the top 2 consistently profitable segments based on a combination of maximum profits and least Coefficient of Variation of profit

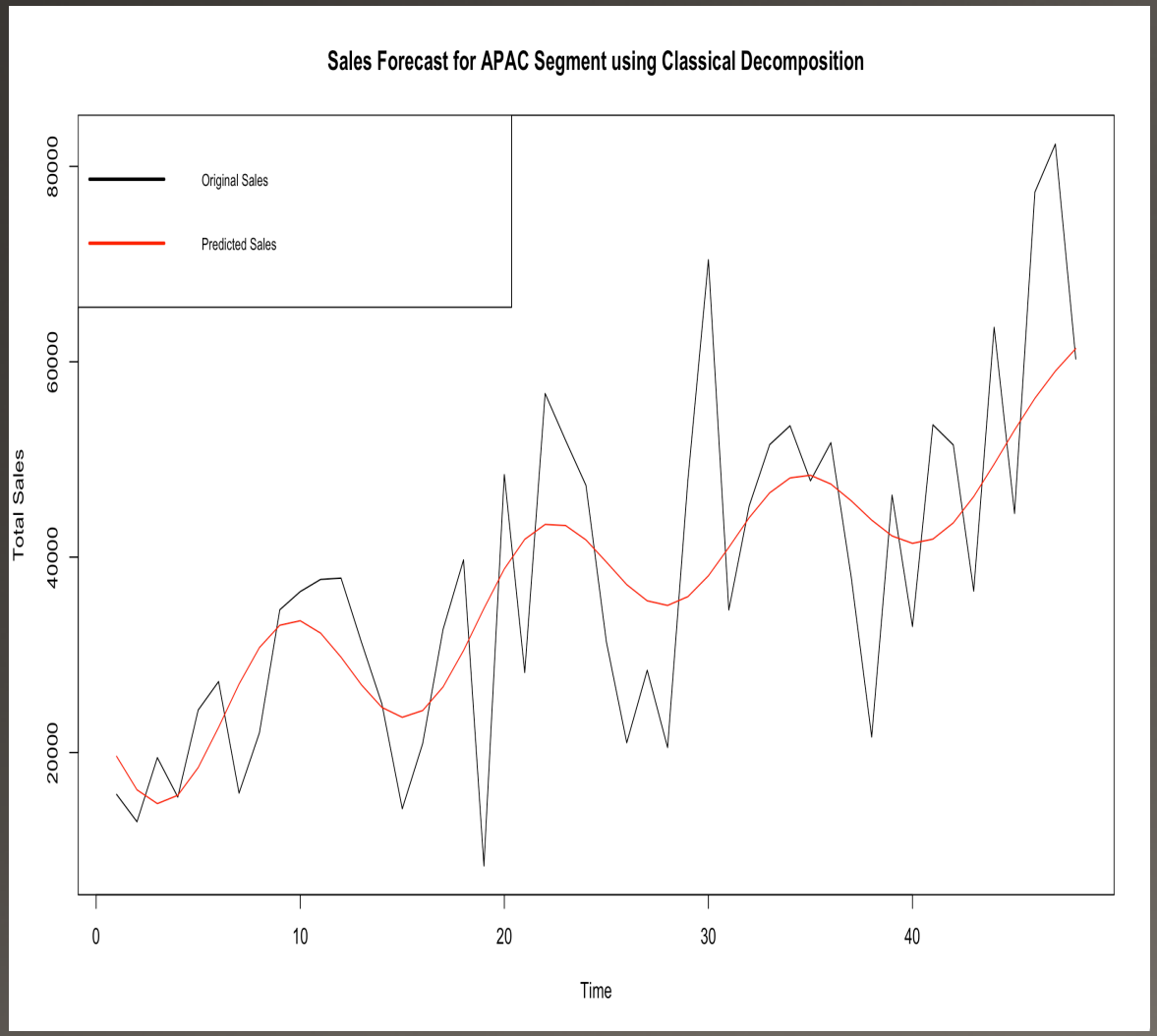
Model Building – Part 1

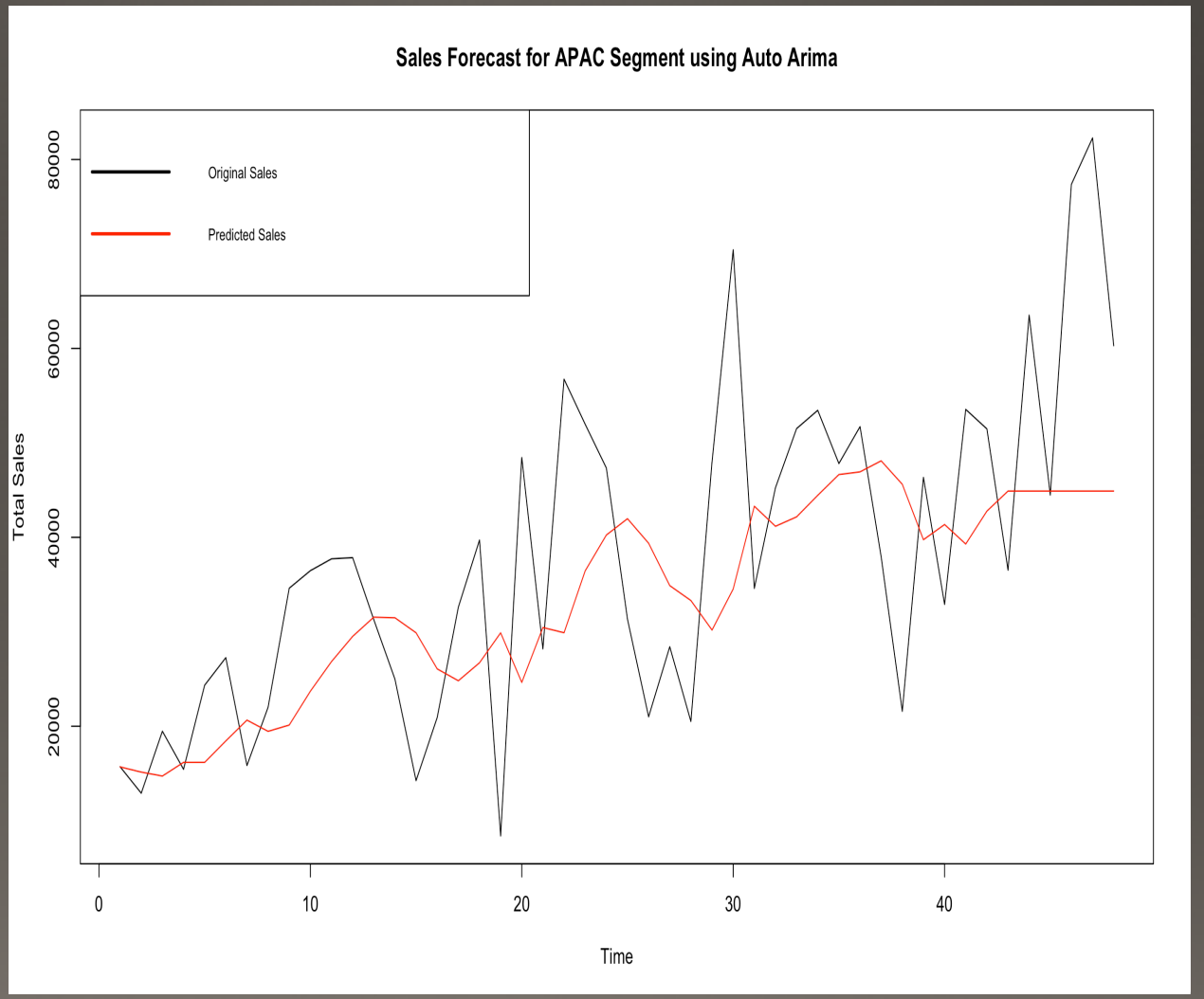
1. Filtered out observations for APAC Consumer and EU Consumer segments
2. Created Time Series for Sales and Quantity (Demand)
3. Set aside last 6 months data for validation
4. Smoothened series using Moving Average with window size of 3
5. Plotted Time Series Sales and Demand to identify any Global Trends and Seasonality
6. For Classical Decomposition approach, used linear regression to capture trend and seasonality and used it predict values for Sales and Demand
7. Removed the trend and seasonality to create a residual time series
8. Carried out stationarity tests (ACF, PACF and Auto ARIMA) on residual time series to confirm that there is no local predictable behaviour
9. Used Quartile Quartile plot, Dickey-Fuller and KPSS tests to check for white noise

Model Building – Part 2

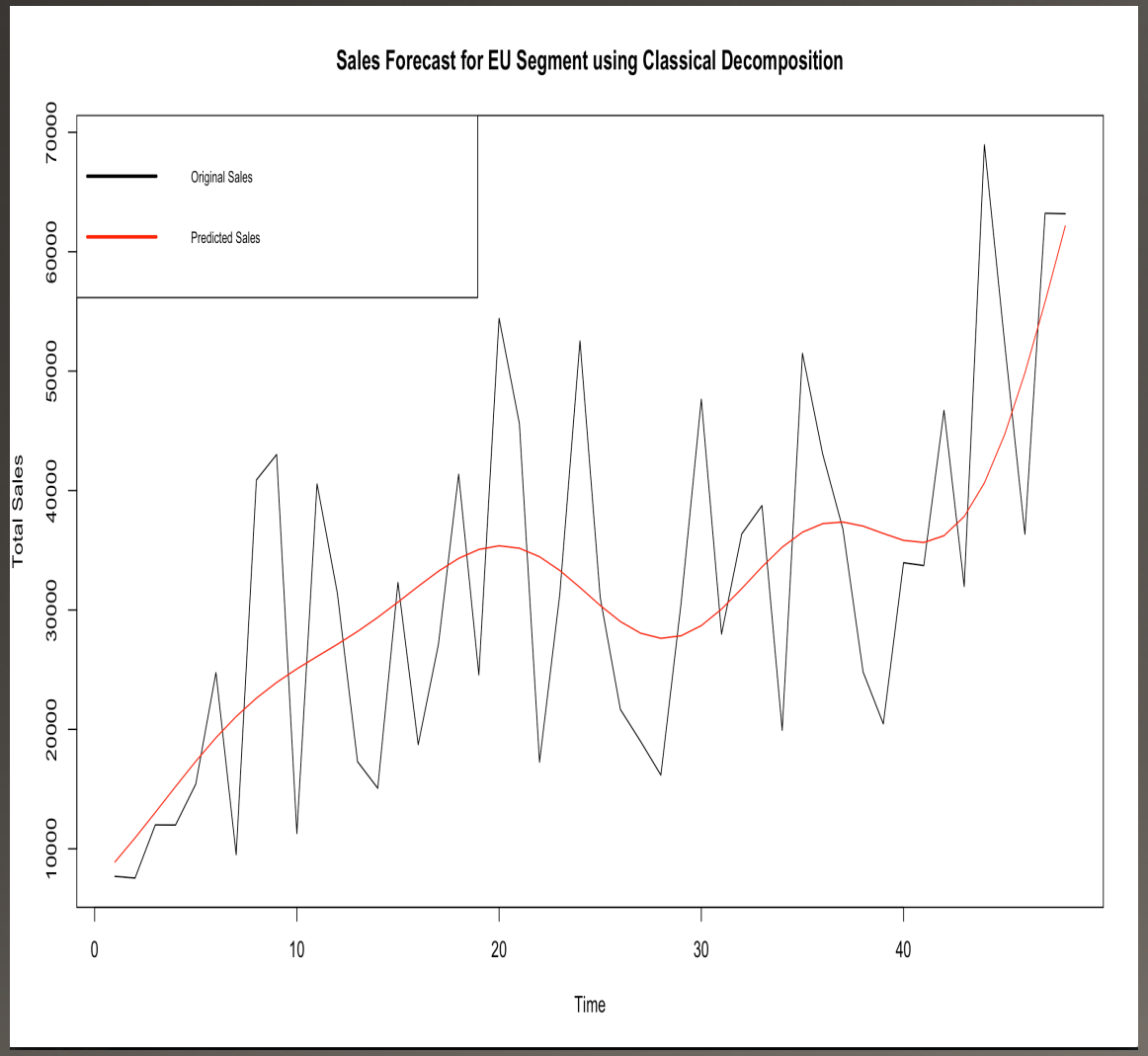
10. Predicted last 6 months Sales and Demand values using a linear regression model
11. Plotted predictions against original values
12. In Auto ARIMA approach called auto.arima function by passing in the time series for Sales and Demand
13. Checked for stationarity in residual time series using ACF, PACF
14. Checked for white noise using Dickey-Fuller and KPSS
15. Predicted last 6 months Sales and Demand values using the Auto ARIMA model

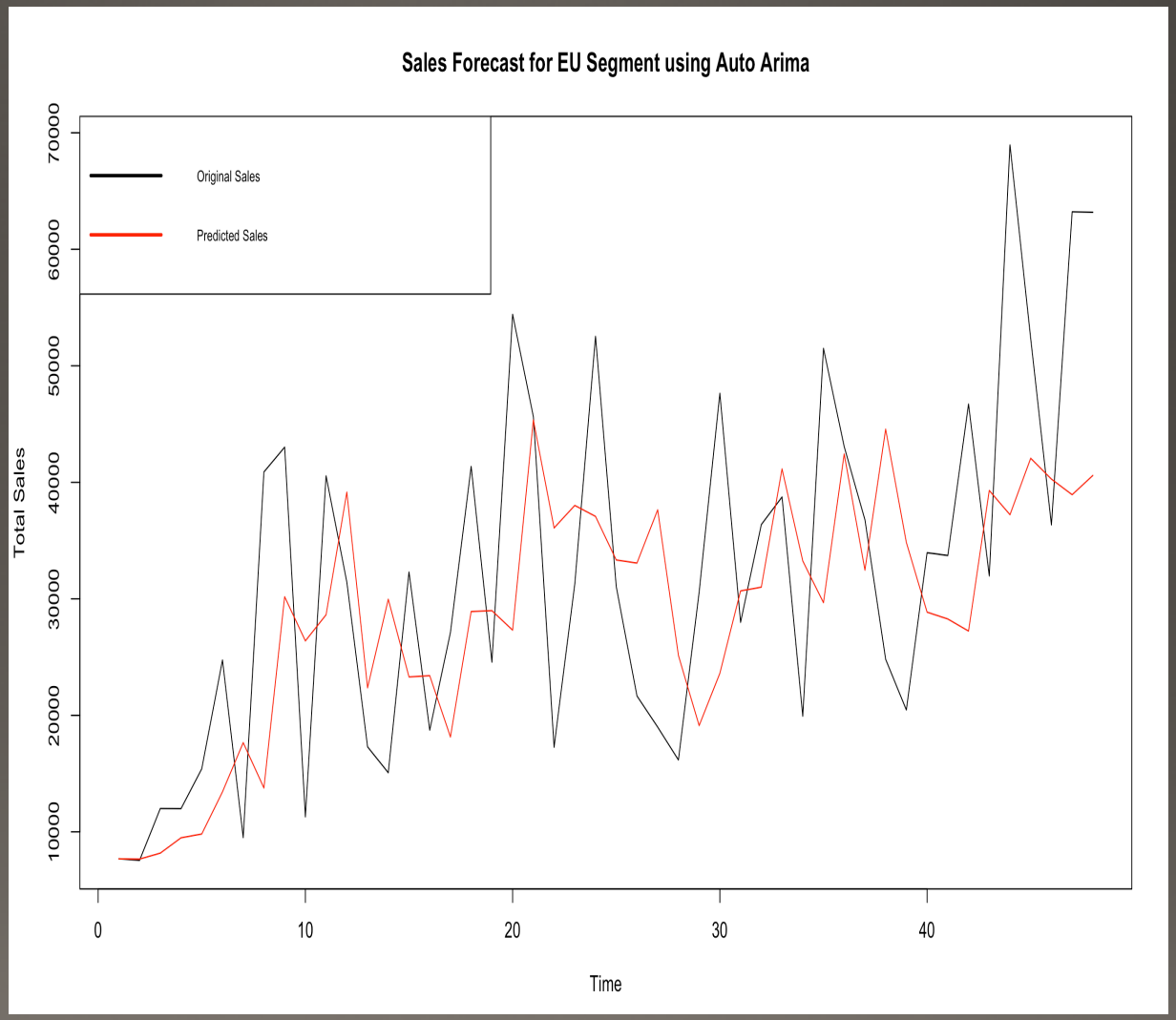
APAC Sales Forecast Comparison



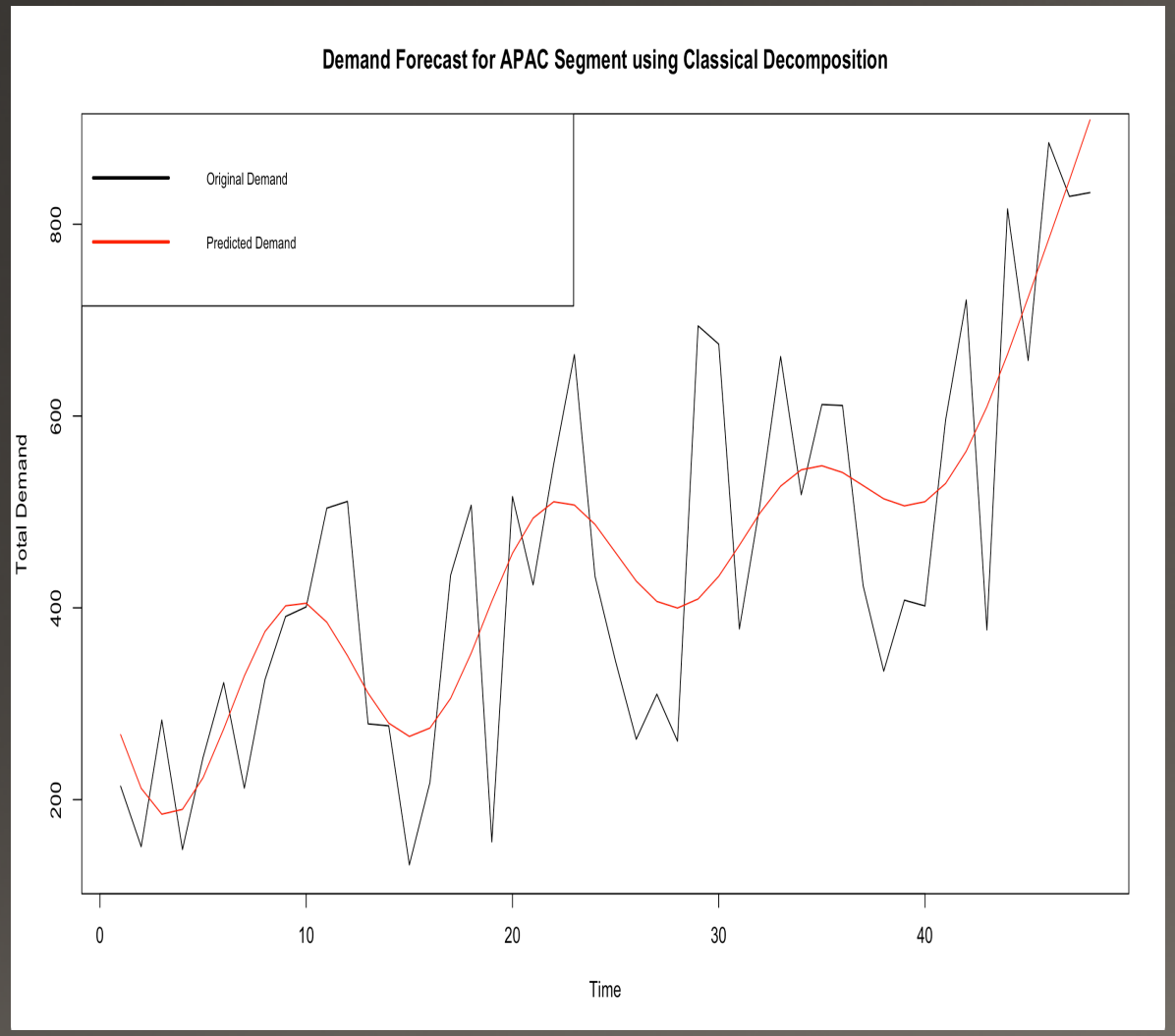


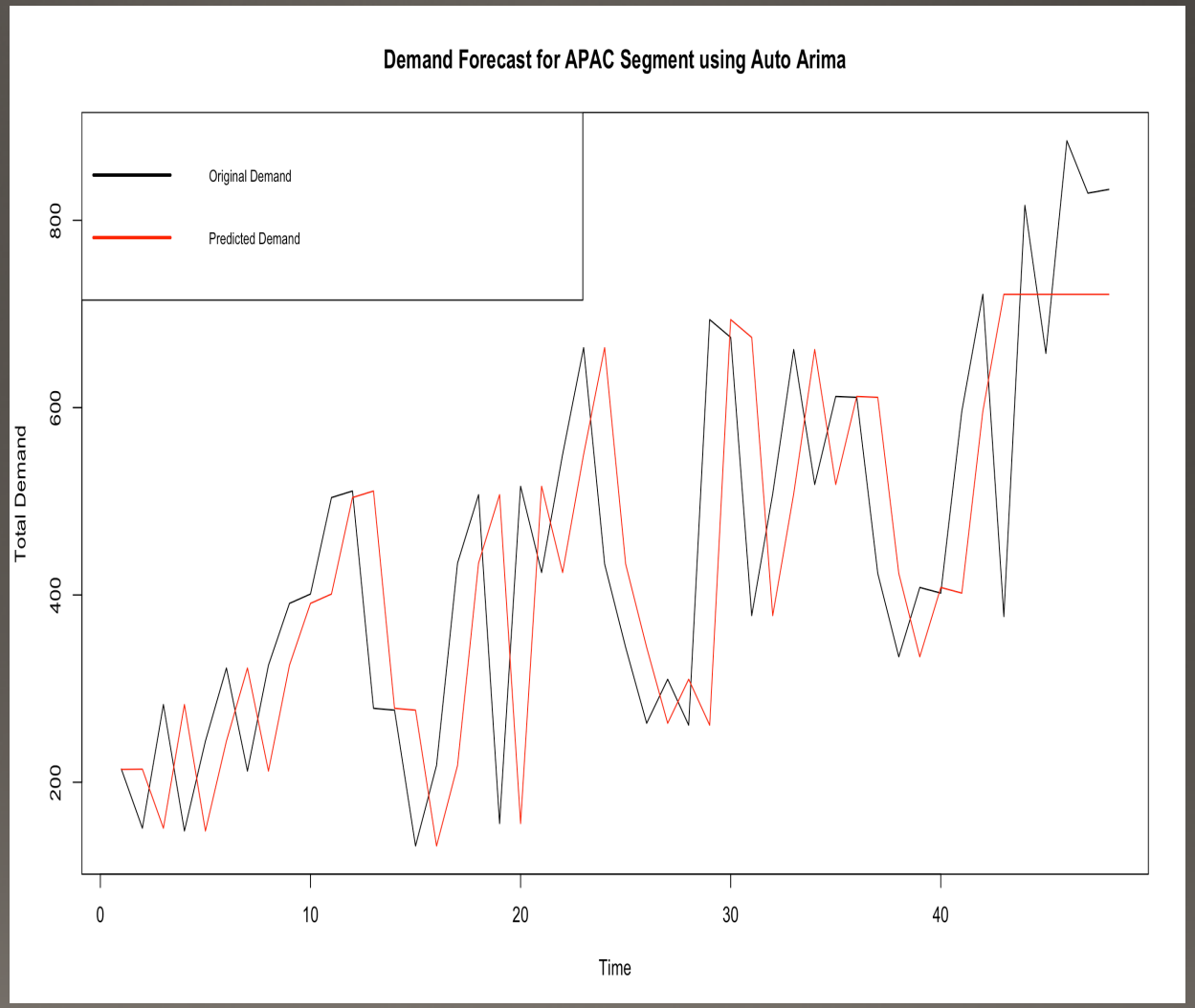
EU Sales Forecast Comparison



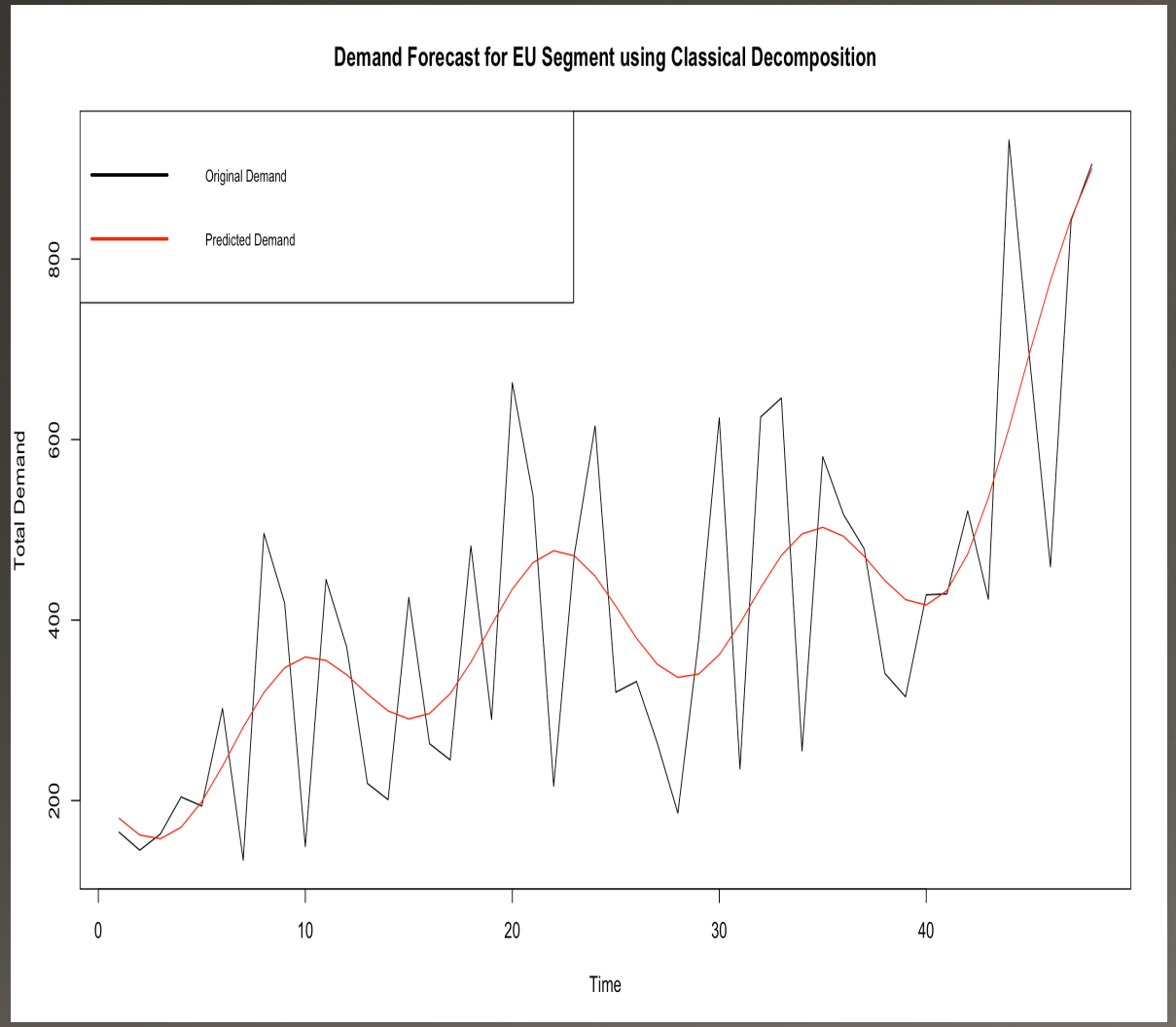


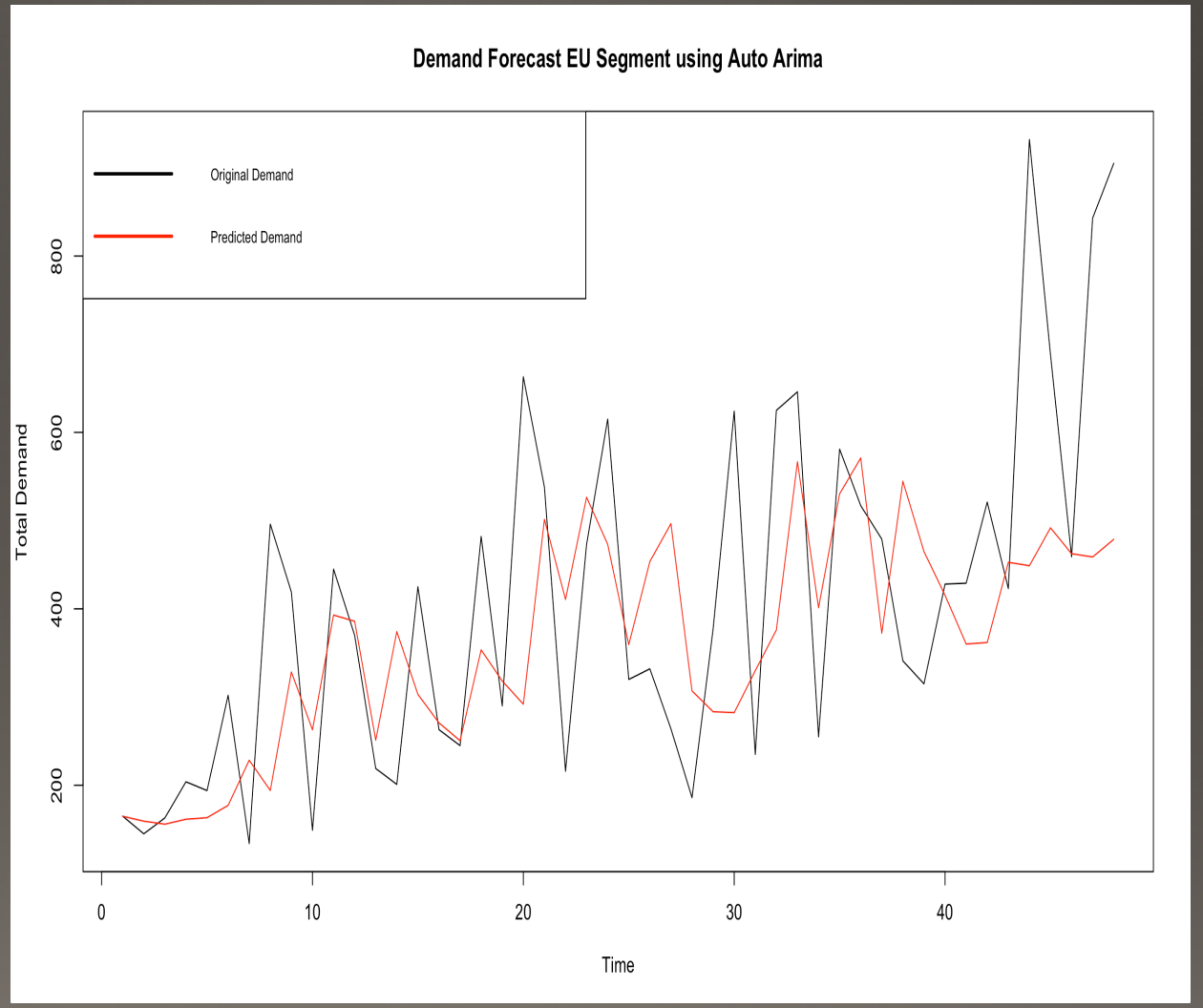
APAC Demand Forecast Comparison





EU Demand Forecast Comparison





APAC Consumer Sales Model Evaluation

Market Segment	Evaluation Parameter	Classical Decomposition model	Auto ARIMA model
APAC Consumer Sales	Global behaviour	lm(Sales ~ sin(0.5*Month) * poly(Month,1) + cos(0.05*Month) * poly(Month,1) + tan(0.02*Month), Month, data=apac_sales_smootheddf)	ARIMA(0,1,1). This means that 1 stage differencing was performed and the resulting time series was modeled as MA(1).
	Local behaviour	ARIMA(0,0,0)	N/A
	Residual component	Stationary	Stationary
	ADF test on stationary	Augmented Dickey-Fuller Test shows a p-value of 0.01 which is < 0.05	Augmented Dickey-Fuller Test shows a p-value < 0.05
	KPSS test on stationary	KPSS test shows a p-value of 0.1 with is > 0.05	KPSS test shows a p-value > 0.05
	Forecast MAPE (%)	20.83	27.68

APAC Consumer Demand Model Evaluation

Market Segment	Evaluation Parameter	Classical Decomposition model	Auto ARIMA model
APAC Consumer Demand	Global behaviour	lm(Demand ~ sin(0.5*Month) * poly(Month,1) + cos(0.1*Month) * poly(Month,1) + tan(0.02*Month), Month, data=apac_demand_smootheddf)	ARIMA(0,1,0). This means that 1 stage differencing was performed.
	Local behaviour	ARIMA(0,0,0)	N/A
	Residual component	Stationary	Stationary
	ADF test on stationary	Augmented Dickey-Fuller Test shows a p-value of 0.01 which is < 0.05	Augmented Dickey-Fuller Test shows a p-value < 0.05
	KPSS test on stationary	KPSS test shows a p-value of 0.1 with is > 0.05	KPSS test shows a p-value > 0.05
	Forecast MAPE (%)	18.79	26.24

EU Consumer Sales Model Evaluation

Market Segment	Evaluation Parameter	Classical Decomposition model	Auto ARIMA model
EU Consumer Sales	Global behaviour	lm(Sales ~ sin(0.4*Month) * poly(Month,1) + cos(0.09*Month) * poly(Month,1), Month, data=eu_sales_smootheddf)	ARIMA(2,1,0). This means that 1 stage differencing was performed and the resulting time series was modeled as AR(2).
	Local behaviour	ARIMA(0,0,0)	N/A
	Residual component	Stationary	Stationary
	ADF test on stationary	Augmented Dickey-Fuller Test shows a p-value of 0.01 which is < 0.05	Augmented Dickey-Fuller Test shows a p-value < 0.05
	KPSS test on stationary	KPSS test shows a p-value of 0.1 with is > 0.05	KPSS test shows a p-value > 0.05
	Forecast MAPE (%)	20.73	28.92

EU Consumer Demand Model Evaluation

Market Segment	Evaluation Parameter	Classical Decomposition model	Auto ARIMA model
EU Consumer Demand	Global behaviour	lm(Demand ~ sin(0.5*Month) * poly(Month,1) + cos(0.09*Month) * poly(Month,1) + tan(0.02*Month), Month, data=eu_demand_smootheddf)	ARIMA(2,1,0). This means that 1 stage differencing was performed and the resulting time series was modeled as AR(2).
	Local behaviour	ARIMA(0,0,0)	N/A
	Residual component	Stationary	Stationary
	ADF test on stationary	Augmented Dickey-Fuller Test shows a p-value of 0.02 which is < 0.05	Augmented Dickey-Fuller Test shows a p-value < 0.05
	KPSS test on stationary	KPSS test shows a p-value of 0.1 with is > 0.05	KPSS test shows a p-value > 0.05
	Forecast MAPE (%)	21.98	30.13

Results

- APAC Consumer and EU Consumer market segments were the top 2 performing segments

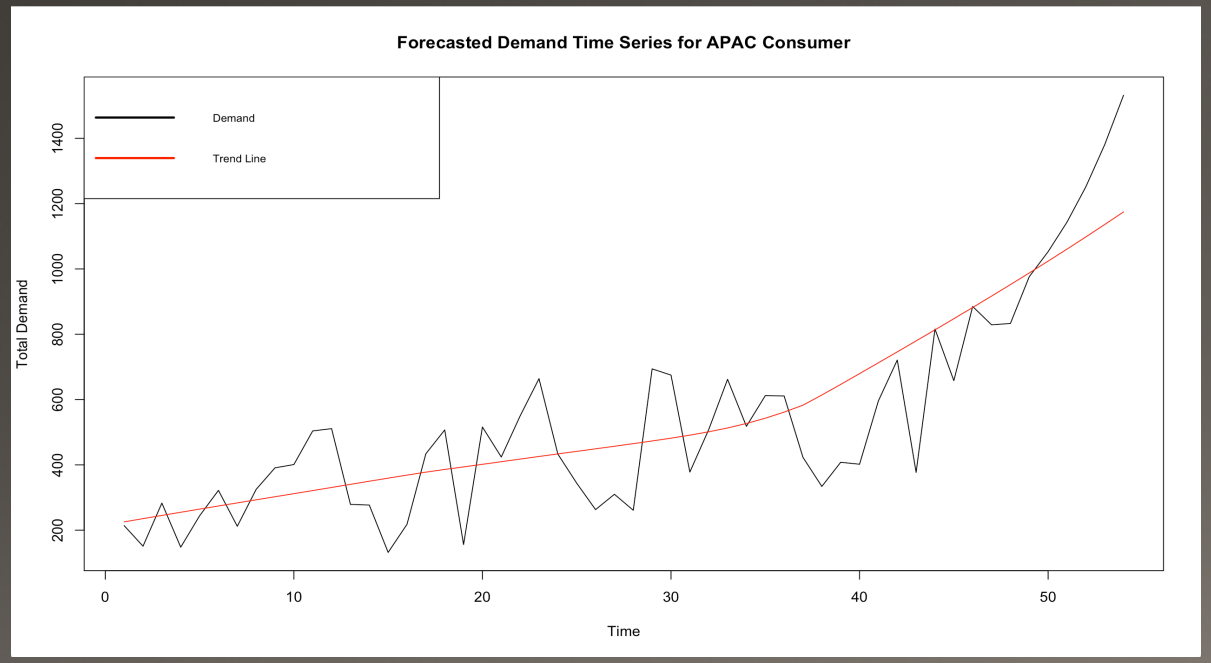
Market Segment	Type	Model	MAPE	Chosen Model
APAC Consumer	Sales	Classical Decomposition	20.83	√
APAC Consumer	Sales	Auto ARIMA	27.68	
EU Consumer	Sales	Classical Decomposition	20.73	√
EU Consumer	Sales	Auto ARIMA	28.92	
APAC Consumer	Demand	Classical Decomposition	18.79	√
APAC Consumer	Demand	Auto ARIMA	26.24	
EU Consumer	Demand	Classical Decomposition	21.98	√
EU Consumer	Demand	Auto ARIMA	30.13	

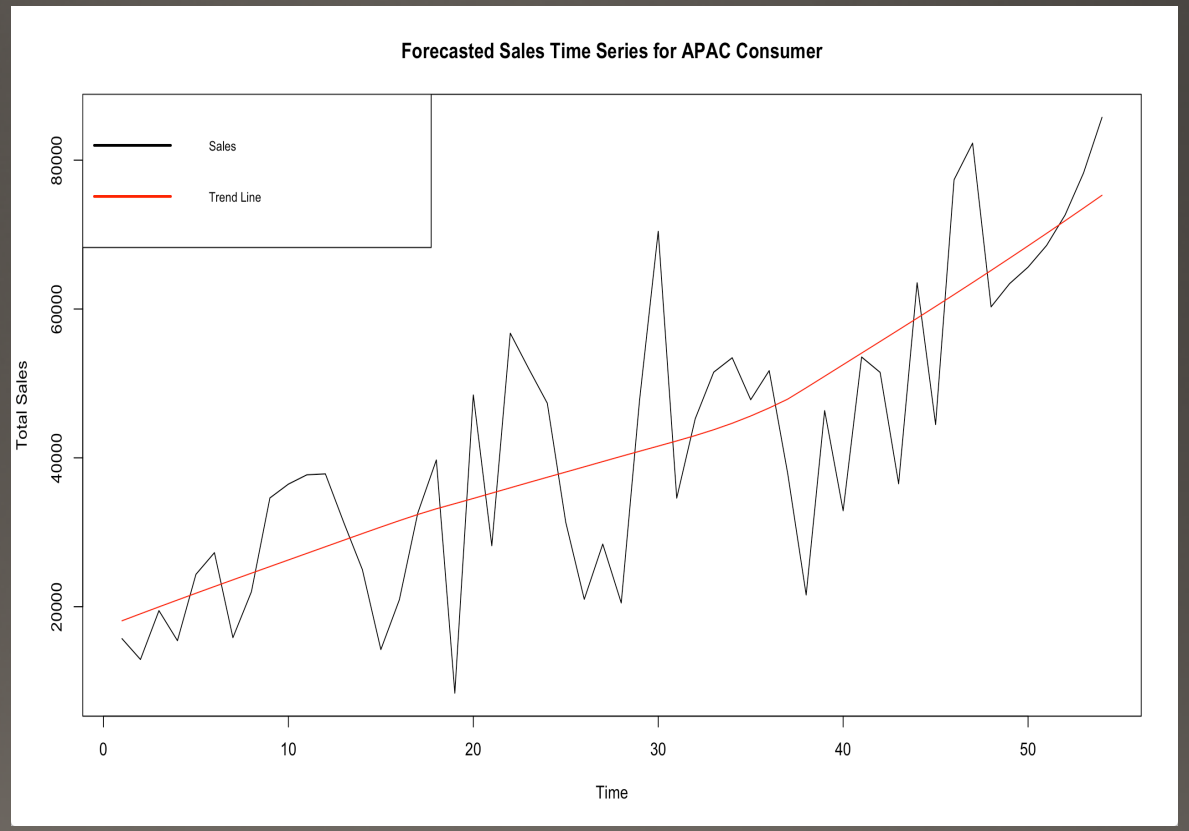
- On comparing the MAPE values of Classical Decomposition and Auto ARIMA for both APAC Consumer and EU Consumer sales and demand, it can be concluded that Classical Decomposition model performs better that the Auto ARIMA model

APAC Future Sales and Demand

- Forecasted Sales and Demand values for the next 6 months in future

Segment	Date	Sales	Demand
APAC Consumer	2015-01	63405.051	976.3252
APAC Consumer	2015-02	65629.256	1052.9020
APAC Consumer	2015-03	68534.041	1143.2185
APAC Consumer	2015-04	72620.157	1251.3994
APAC Consumer	2015-05	78283.062	1380.3938
APAC Consumer	2015-06	85754.444	1531.8233





EU Future Sales and Demand

- Forecasted Sales and Demand values for the next 6 months in future

Segment	Date	Sales	Demand
EU Consumer	2015-01	68506.395	942.8787
EU Consumer	2015-02	74397.705	981.0561
EU Consumer	2015-03	79502.235	1024.5725
EU Consumer	2015-04	83628.334	1084.6744
EU Consumer	2015-05	86762.668	1171.0478
EU Consumer	2015-06	897076.069	1289.6338