

**YDS 2018 APPROACH Document**

1. ***Quality checks performed / Errors found***:

* There are a few errors in the problem /statement and training dataset which can be identified easily and corrected them
* Promotion expenses were not provided for Argentina & Columbia x Product ID 3

1. ***Data preprocessing steps:***

* The Mekktronix training data used to forecast was of weekly data which has the monthly details as well. Data contains the sales details of 6 countries with 3 product categories. Along with this, the promotional expenses and holidays details were provided in separate files.
* Stated with installing required packages like dplyr, tseries, forecast, graphics, lubridate etc.
* Merged the data set and converted them to monthly data point as we need to predict monthly forecast
* Divided the entire data into 11 different data sets based on 6 different countries and 3 product type

1. **Key observations / Trends:**

* With visualization and descriptive statistics, finalized that promotion expenses is consistent for each product type and country. Also, the holidays details are not significant as it is monthly data point

1. **Model choice explanation:**

* Since no strong visibility on affecting the external factors with available data, finalized to build ARIMA model for this time series forecast and followed the below steps for all datasets
* Plotted the sales data as time series
* Differenced the data to make data stationary on mean
* Performed log transformation and then differenced to make data stationary on variance and mean for required dataset
* Plotted ACF and PACF to identify potential AR and MA model.
* Autocorrelation factor (ACF) and Partial autocorrelation factor (PACF) plots are used to identify patterns in the data obtained
* The idea is to identify presence of AR and MA components in the residuals by getting ARIMA (p,d,q) values.
* Carried out both Arima and Auto arima functions for best fit ARIMA model and selected the model which has a better value
* The best fit model is selected based on Akaike Information Criterion (AIC) values. The idea is to choose a model with minimum AIC value
* Forecasted sales using the best fit ARIMA model and plotted ACF and PACF for residuals of ARIMA model to ensure no more information is left for extraction and verified that p is insignificant using the Box-Ljung test

1. **Expected error for submission**:

- The expected error SMAPE came as 1.739 which looks good for this model

1. **Top 5 most significant variables in model:** please provide the most significant feature in the dataset.

* Increasing sales trend for Product Category for all the 4 countries – Argentina, Belgium, Columbia, Denmark
* Product Category 5 too in increasing revenue trend
* Product category 1 shows in decreasing as well as consistent trend
* Denmark has highest sales record for Product category 2
* England has less market for Product Category 4 or it has recently introduced