ELECKART
MARKET MIX MODELLING

Upgrad – IIITB PGDDS - Ecommerce Capstone Project

By
Amrita Bhadrannavar
Shafeeq Rahiman

CONTENTS

- Business Objective
- Data Understanding
- Data Preparation
- Exploratory Data Analysis
- Model Building
- Model Results and Analysis
- Recommendations

BUSINESS OBJECTIVE

Problem Statement

EleсKart is an e-commerce firm specializing in electronic products. To enhance their revenues they have done significant investment in their marketing efforts, like promotions over last one year. They are about to create a marketing budget for the next year which includes spending on commercials, online campaigns, and pricing & promotion strategies. They want to reallocate their budget optimally across different marketing levers to improve the revenue response using Market Mix modelling.

Objective

- To develop a market mix model for 3 product sub-categories:
  - Camera accessory
  - Gaming accessory
  - Home Audio
- To observe the actual impact of different marketing levers over sale of last one year (July 2015 -June 2016)
- To recommend the optimal budget allocation for different marketing levers for the next year.

DATA UNDERSTANDING

ConsumerElectronics.csv

FIELD	DESCRIPTION
fsn_id	The unique identification of each SKU
order_date	Date on which the order was placed
Year	Year of the order
Month	Month of order
order_id	The unique identification number of each order
order_item_id	Suppose you order 2 different products under the same order, it generates 2 different order Item IDs under the same order ID; orders are tracked by the Order Item ID.
GMV	Gross Merchandise Value or Revenue
Units	Number of units of the specific product sold

DATA UNDERSTANDING (Cont.)

ConsumerElectronics.csv

FIELD	DESCRIPTION
Deliverybdays	Dispatch delay from Warehouse
Deliverycdays	Delivery delay to customer
s1_fact.order_payment_type	How the order was paid – prepaid or cash on delivery
Sla	Number of days it typically takes to deliver the product
cust_id	Unique identification of a customer
Pincode	Zip code
product_analytic_sub_category	Product sub category
product_mrp	Maximum retail price of the product
product_procurement_sla	Time typically taken to procure the product

DATA UNDERSTANDING (Cont.)

- **Media data and other information.xlsx**
  - Product List - List of all products sold and their sales frequency
  - Product Category, Frequency of sale, % of Total Sale
  - Media Investment - Media investment details for prior periods, by channel.
    - Year, Month, Total Investment, TV, Digital, Sponsorship, Content Marketing, Online marketing, Affiliates, SEM, Radio, Other
  - Special Sale Calendar - Holidays and other special events on calendar
  - Monthly NPS Score - Monthly Net Promoter Score

- **Product details.docx**
  - Product information with super category, category, sub category and vertical details

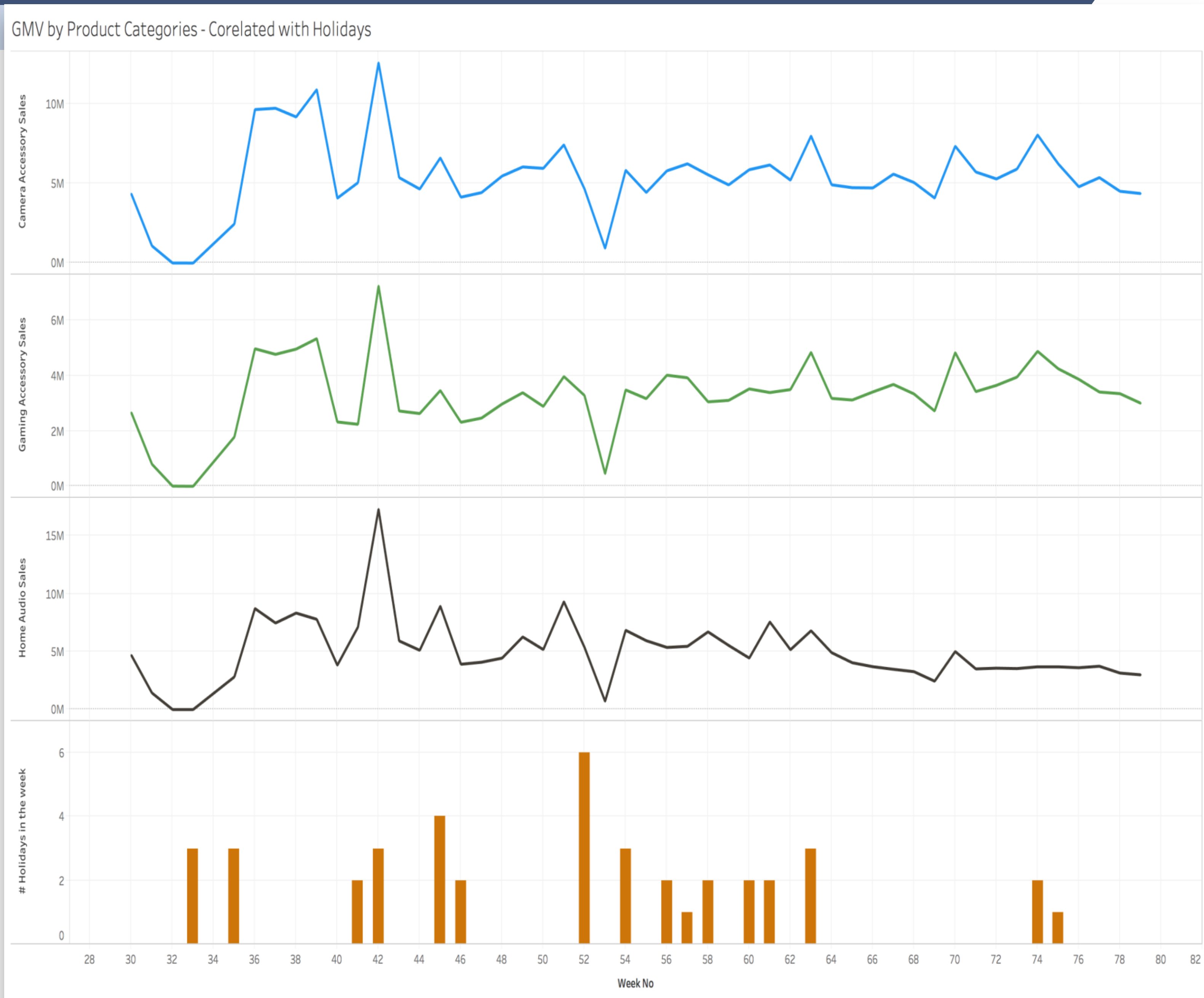
DATA PREPARATION

- Order data has 1.6M records
- Field Observations :
  - GMV : 1349 values were zero, few negligible values like 1.
  - Date range: 906 data points were outside of July-2015 to June-2016. These were eliminated.
  - deliverycdays and deliverybdays columns had negative and “\\N” values, which were eliminated.
  - Sla: some very high numbers were present(100+ and also 100+ days)
  - pin_code: 4904 NA values were found
  - cust_id: 1.2M unique customers , but 4904 NAs and negative values were found.
  - mrp: 5308 records with zero MRP found. These were assumed to be freebies and eliminated.
  - product_procurement_sla: Negative values found and updated to zero.
- Conversion of data to week level of granularity
- Convert to numeric fields where applicable.
- NPS data is at month level and is assumed to be same across all weeks of the same month.

DATA PREPARATION (Cont.)

- Adstock calculation :
  - Data for 9 channels of investment are provided (In Cr. INR) at month level
  - Two channels, ‘Radio’ and ‘Other’ have data only for 3 months - so these are eliminated
  - Remaining 7 channels of investment are used to create adstock data.
  - Number of weeks are calculated in each month and monthly investment is averaged across each week of the month.
  - Adstock rate is assumed to be 0.5
  - Final calculation is based on the weekly investment, prior week investment and adstock rates.
- Holiday data:
  - Derived manually from “Special Sale Calendar” sheet of “Media data and other information.xlsx” file.
  - Week Number of holiday is identified
  - Number of special days or holidays in each week is identified and
  - saved as new column.
- Data is divided into three different data frames for analysis and modelling (Camera Accessories, Gaming Accessories and Home Audio)

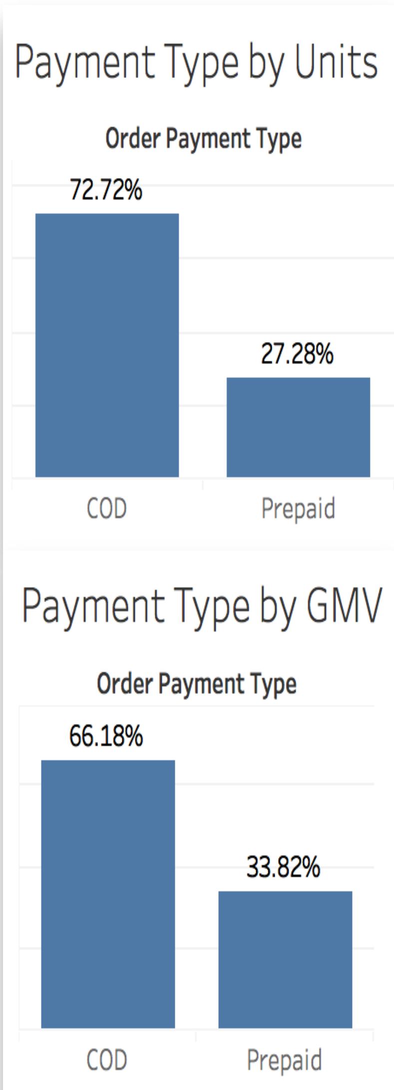
EDA PLOTS



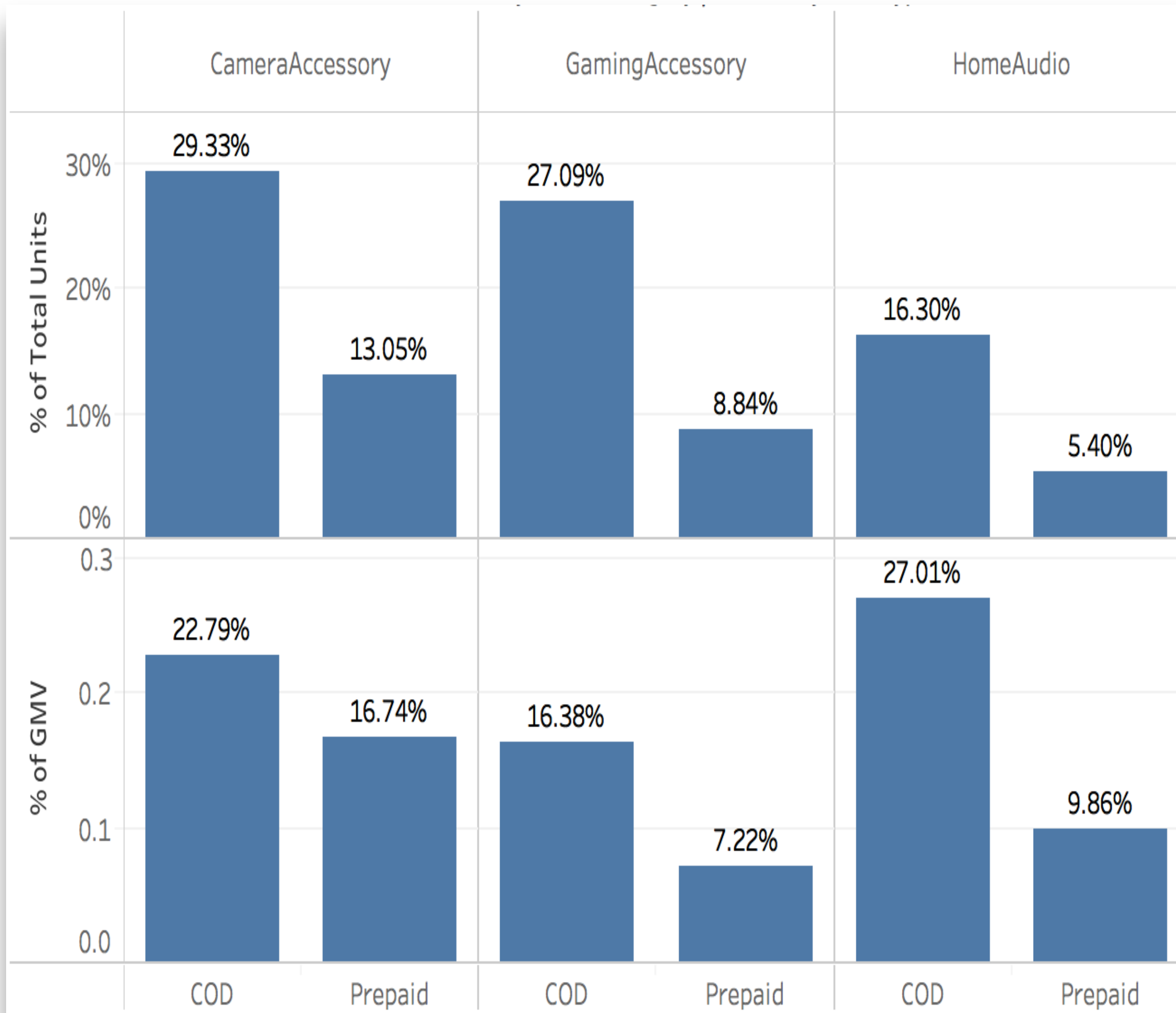
Week over Week GMV by Category

EDA PLOTS – Sales share by Order Type

Overall

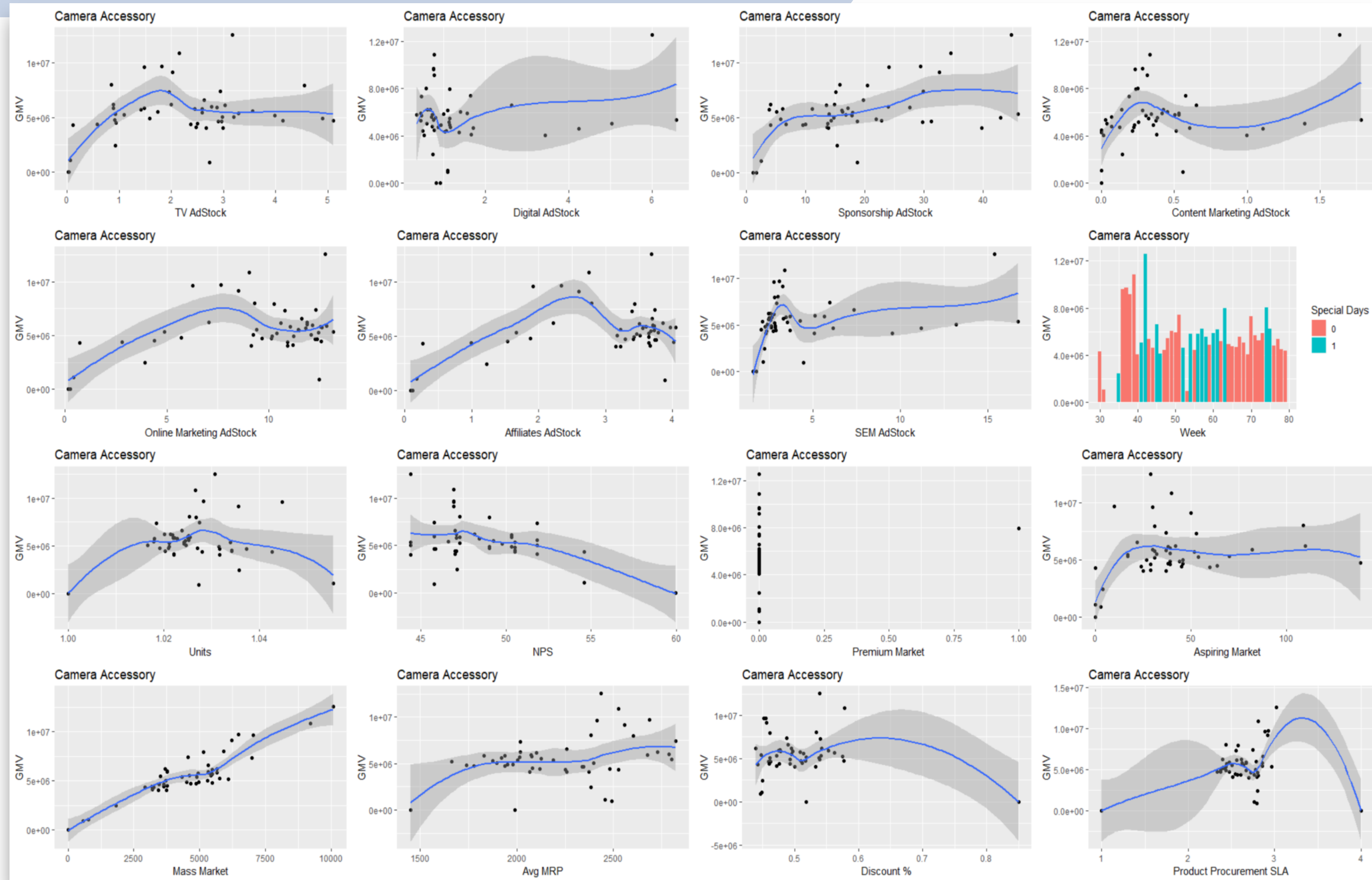


By categories



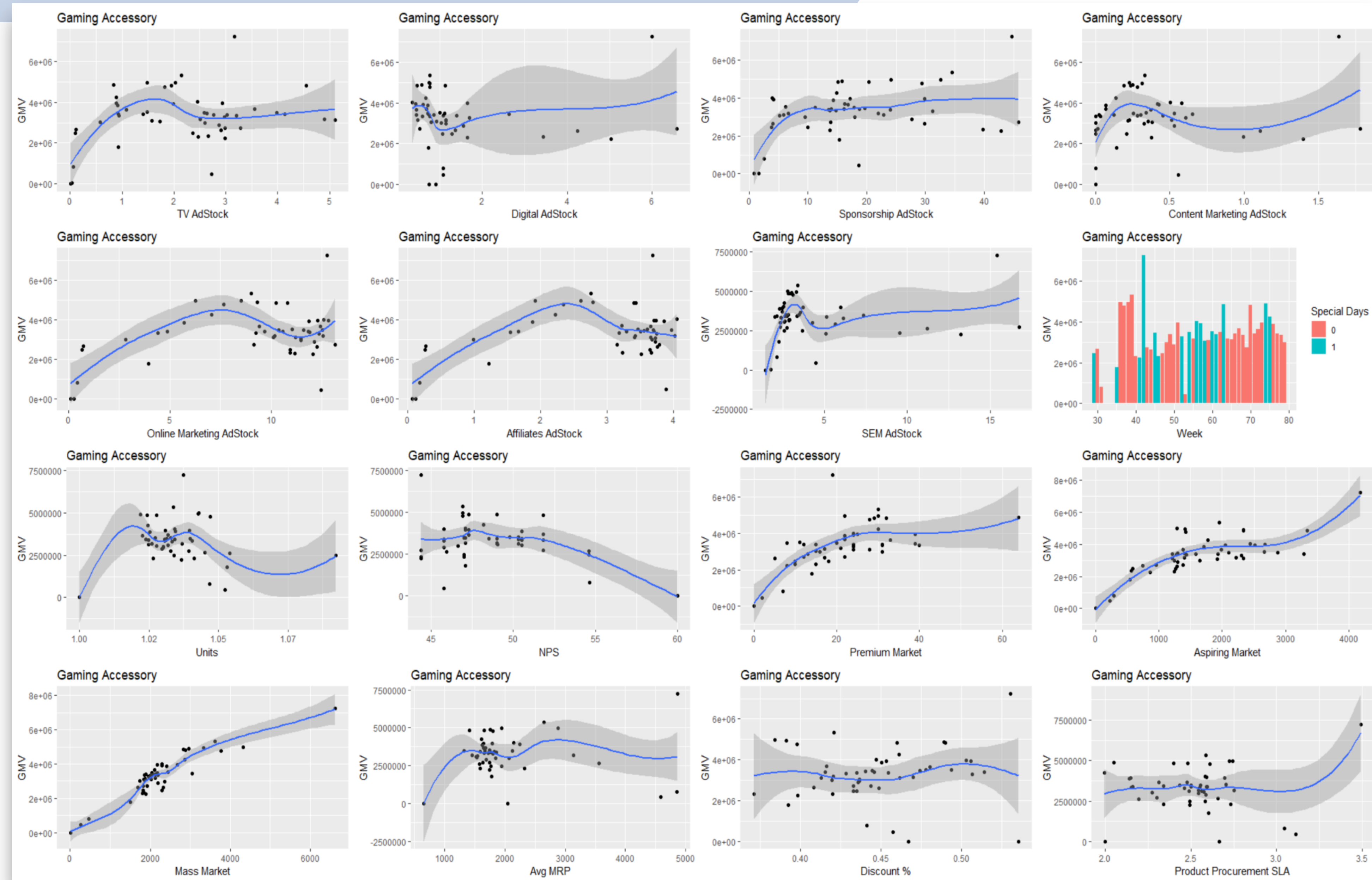
FEATURE ENGINEERING EDA PLOTS

Camera Accessory – GMV vs other variables



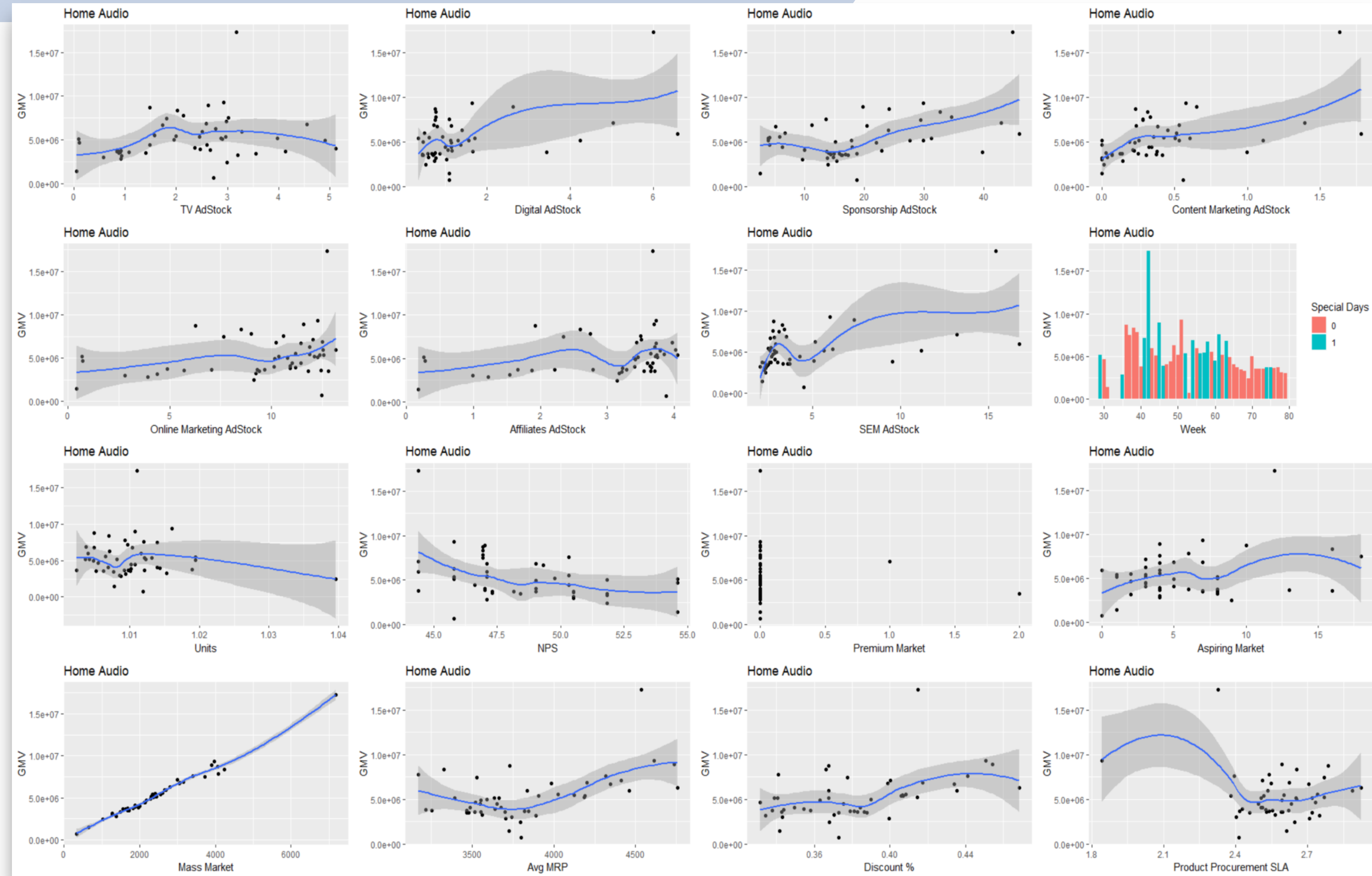
FEATURE ENGINEERING EDA PLOTS

Gaming Accessory – GMV vs other variables



FEATURE ENGINEERING EDA PLOTS

Home Audio – GMV vs other variables



EDA CONCLUSIONS

- Home Audio is most expensive product bringing in maximum revenue amongst 3 sub categories.
- Camera Accessory have maximum no of units sold followed by Gaming Accessory
- A large portion of sales are with cash on delivery payment type. The Cash on Delivery mode should be made more convenient to enhance sales. But introspection on the Prepaid mode is also needed to plan for better and smooth processes to increase sales as it is beneficial to the company from both operational and financial p.o.v.
- The increase in GMV is observed for most investment channels, but only up to a certain limit. After that, the increase in spend is not impacting GMV to a high degree.
- We also notice that there could be a carry-over effect on certain marketing channels where the increase in GMV is noted even when there is no substantial investment.
- During holiday season, there are peaks in both sales and items sold across all product categories.
- Heavy discount is seen during holiday season but discounts over 50% is not observed to be effective in increasing the sales.

DERIVED METRICS

- List Price of the product
- Discount Percentage
- Clustering of products based on price of product for brand perception
  - Mass Market
  - Aspiring Market
  - Premium Market
- Prepaid Orders % (based on Payment Type)
- Number of Orders by Delivery Status
  - Delayed delivery
  - On Time delivery
  - Early delivery
- Number of holidays or special days in each week
- Moving averages of list price and discount percentage for inflation analysis up to 3 weeks
- Lag of list price, discount percentage

MODEL BUILDING

- Models are built for the three product categories separately:
  - Camera Accessories
  - Gaming Accessories
  - Home Audio

- Models considered
  - Linear Model
  - Multiplicative Model
  - Koyck Model
  - Distributed Lag Model
  - Multiplicative + Distributed Lag Model

MODEL BUILDING (Cont.)

- Modelling Approach
  - Build the Basic Linear Model with all the KPIs
  - Build the multiplicative model using the log of the individual KPIs
  - Build the Koyck model using the lag of the dependent variable
  - Build the distributed lag model using the past lags of both the dependent and the independent variables
  - Build the multiplicative + distributed lag model using log of the past lags of both the independent variables
  - Tabulate Adj-R squared , MSE value and significant variables for each model. MSE figures are based on the 10 fold cross validation again on the training data
  - Choose the best performing model based on Adj-R squared value, MSE value and variables which are business actionable

MODELLING RESULTS & ANALYSIS
Camera Accessory

Model	Adj - R2	MSE	Variables
Linear Model	0.669	0.432	total_investment, discount_per, aspiring_market, Sponsorship, inc_MA_LP1, inc_MA_DP1
Multiplicative Model	0.882	0.564	avg_units, prepaid_per, aspiring_market
Koyck Model	0.237	0.889	Sponsorship
Distributed Lag Model	0.223	0.825	product_procurement_sla
Multiplicative + Distributed Lag Model	0.971	0.070	**product_procurement_sla, aspiring_market, NPS.3**

- Koyck & Distributed Lag Model have low Adj-R2 & high MSE values and hence are rejected.
- Multiplicative & Multiplicative + Distributed Lag Model have high R2 values which can be chosen as best model. But Multiplicative model has higher MSE value.
- So **Multiplicative + Distributed Lag Model** with **high Adj-R2** value and **low MSE** value would be best choice for **Camera Accessory** sub category.

MODELLING RESULTS & ANALYSIS
Gaming Accessory

Model	Adj_R2	MSE	Variables
Linear Model	0.840	0.198	discount_per, aspiring_market, premium_market, inc_MA_DP1
Multiplicative Model	0.988	0.069	sla, discount_per, aspiring_market, premium_market
Koyck Model	0.871	0.169	sla, total_investment, list_price, aspiring_market, Sponsorship
Distributed Lag Model	0.911	0.154	sla, total_investment, list_price, aspiring_market, Sponsorship, inc_MA_LP1, Lag_DP3_per
Multiplicative + Distributed Lag Model	0.988	0.079	sla, discount_per, aspiring_market, premium_market

- Linear model is ruled out due to lowest Adj-R2 and high MSE values
- Multiplicative & Multiplicative + Distributed Lag Model are both best options as Adj-R2 has high value and MSE has low values. Also the significant variables are same both Models. **Multiplicative Model** is chosen finally on the basis of selection of simpler model
- Distributed Lag Model also is decent option to consider as it has more business variables to work with.

MODELLING RESULTS & ANALYSIS
Home Audio

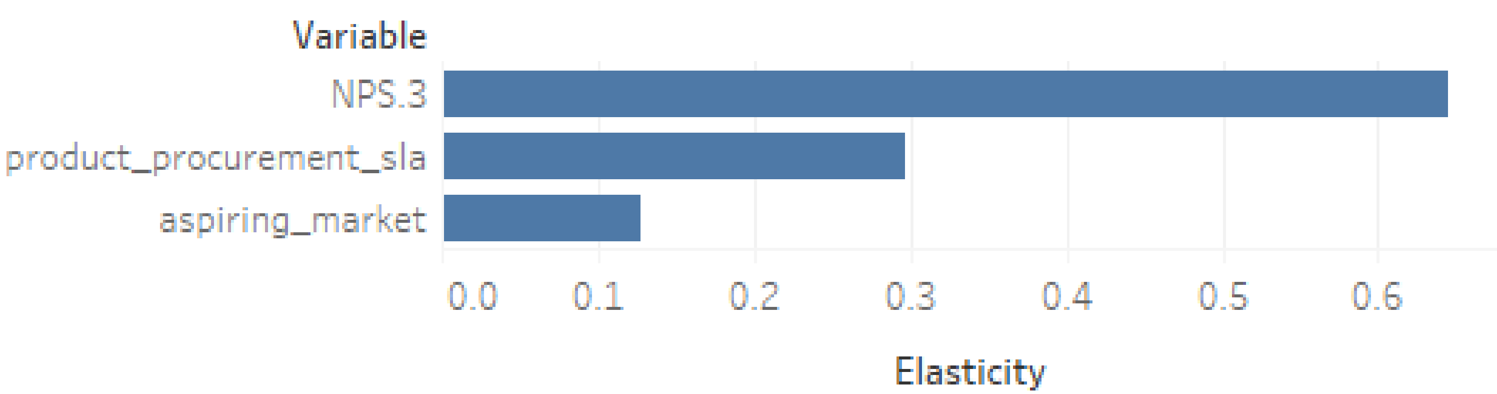
Model	Adj_R2	MSE	Variables
Linear Model	0.561	0.603	avg_mrp, aspiring_market, Digital
Multiplicative Model	0.379	0.185	aspiring_market, Digital
Koyck Model	0.560	0.617	avg_mrp, aspiring_market, Digital
Distributed Lag Model	0.557	0.648	avg_mrp, aspiring_market, Digital
Multiplicative + Distributed Lag Model	0.525	0.174	**discount_per, aspiring_market, Digital**

- Significant variables aspiring_market and Digital are common across all models.
- Multiplicative Model has least Adj-R2 and hence is ruled out
- Linear and Koyck Model have good Adj-R2 but MSE values are high
- **Multiplicative + Distributed Lag** Model has Adj-R2 in mid-range and MSE value is least which makes this model best option for Home Audio Accessory

RECOMMENDATIONS
Camera Accessory

- Increase in promotion of aspiring market products like:
  - Camera Film Rolls, Reflector Umbrella
- Discount offers will also have an positive impact on sales
- Allocate resources for channel of communication and feedback with customers to improve customer experience who become brand advocates leading to increase in sales
- The below figure represents the elasticity of different variables w.r.t the overall sales for Gaming Accessory sub category. Positive elasticity means increasing the value of the KPI would lead to increase in the sale.

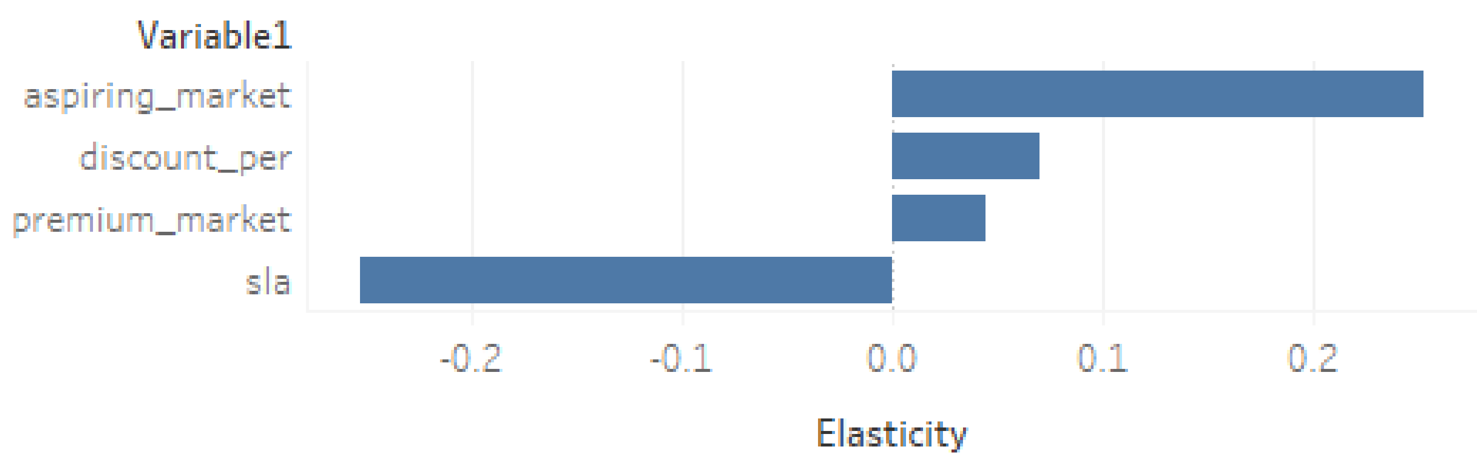
Multiplicative + Distributed Lag Model



RECOMMENDATIONS
Gaming Accessory

- Increase in promotion of aspiring market products like:
  - Cooling Pad, Gaming Headset, Gaming Adapter, Gaming Charging Station, Gaming Keyboard, Gaming Mouse Pad, Gaming Memory Card & Gaming Speakers
- Increase in promotion of premium market products like:
  - Motion Controller, Game Control Mount
- Discount offers also play a good role in increasing sales and hence more resources can be channelized here.

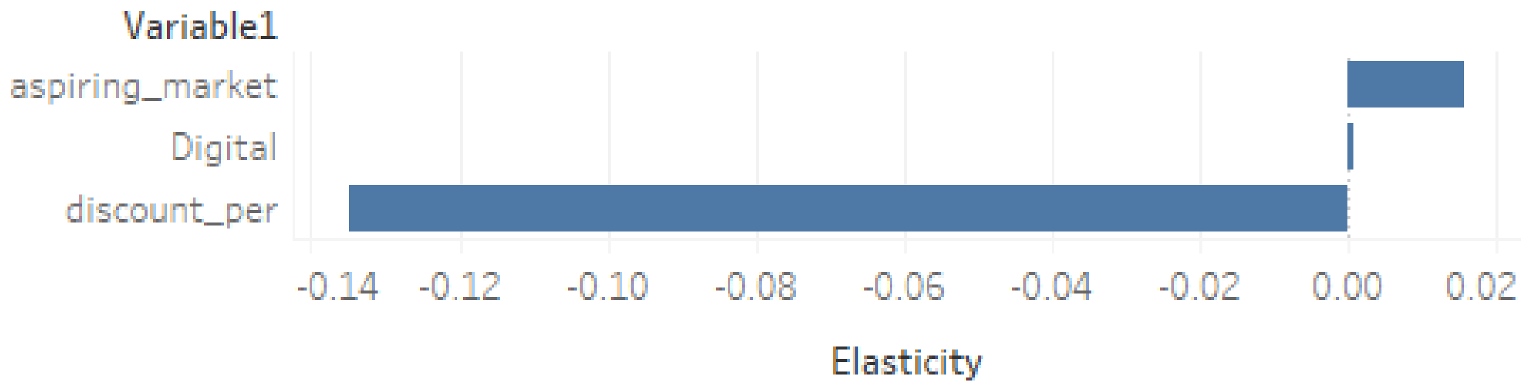
Multiplicative Model



RECOMMENDATIONS
Home Audio

- Increase in promotion of aspiring market products like:
  - Sound Mixer, DJ Controller, Karaoke Player
- Increase expenditure in Digital Marketing to see positive sales results
- Decrease in discount offers should also be adopted due to negative elasticity being observed.

Multiplicative + Distributed Lag Model



CHALLENGES FACED

- Deciding on number of derived KPI needed for modelling which are important as well logical from business and marketing p.o.v.
- Many iterations were performed to arrive at best dataset for each model type and sub category as the model results were leading to perfectly fitting model
- Creation of adstock data, lag variables, identifying the functions/ packages to perform these.
- Limited availability of the sample implementation Kyock, Lag + Multiplicative models in R on internet.
- Decision on removal of variables from analysis. Removal was based on correlation concept to remove correlated variables,
- Selection of significant variables which are finally decided by correlation matrix and non zero variance.
- Setup issues faced with read.xlsx function with errors w.r.t Java and Perl

Thank you!