Capstone Project
Market Mix Modelling

Shubham Kokate

Business Objective

To create a market mix model for ElecKart (an e-commerce firm based out of Ontario, Canada) for 3 product sub-categories - Camera Accessory, Gaming Accessory and Home Audio - to observe the actual impact of various marketing variables over one year

Performance driver analysis: Which KPIs drive the top-line performance

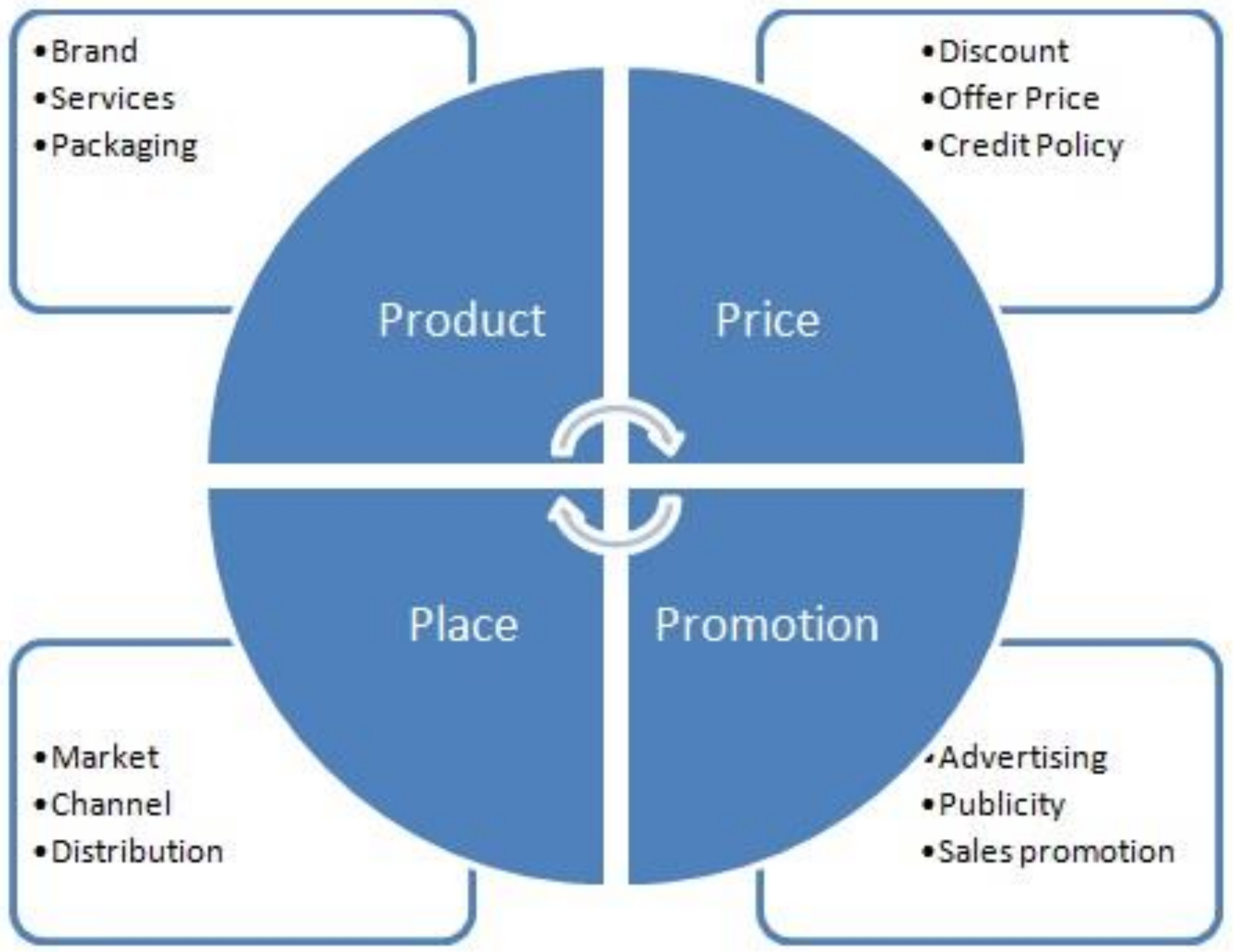
Impact analysis on marketing ROI: What is the quantitative impact of each commercial lever on revenue?

Optimizing marketing spends: How to best allocate the marketing budget to gain the highest outcome?

Understanding Data

The following data files were available to us for analysis of budget optimization:

- Main Consumer file with order details at a daily basis
- Media Investment file with amount invested in each advertising medium for the past year
- Sale Calendar file showing dates from past year when there was a promotional offer
- NPS file showing net promotion score and company stock value for last year
- Weather file having detail weather reports from last year in the state of Ontario, Canada



Data Preparation and Cleanup

Handling Incorrect values

- Imputing "\N" value in deliverybdays & deliverycdays by 0
- Treating incorrect GMV values (where gmv > product_mrp * units) by imputing the faulty MRP values with GMV/units
- Handling Negative values for product_procurement_sla, deliverybdays & deliverycdays by dropping them
- Handling large values(0.3%) for product_procurement_sla by dropping them

De-Duplication of Data

- After converting all column values to lower case, we see that there are around 6.33% rows that are duplicates. We went De-Duplication ahead and dropped them

Treating Null values and whitespaces

- Initially there weren't any NULL values in the dataframe. However, there were quite a few Whitespaces present in some of the columns in the dataframe
- We first converted these whitespaces to NaNs and the dropped these values

Dropping insignificant columns

- Dropping Columns with Single Unique Value (as it doesn't add any information to the analysis)
- Dropping some of the 'Id' Columns which are insignificant to the analysis

Outlier Treatment

- Since we have already deleted some records on erroneous grounds, in order that we don't lose any further data, we chose not to delete outlier values
- For the variables - 'SLA', 'deliverybdays', 'deliverybdays', 'gmv', 'product_mrp', 'list_price' where outliers are present, we CAPPED the values above 99 percentile to the value corresponding to 99 percentile
- Thus the outliers couldn't affect the predictive model while at the same time there was enough data to build a generalizable model

Converting Categorical Attributes to Numerical Form

- Binary encoding for categorical variable with 2 levels
- One Hot Encoding for categorical variable with multiple levels by creating dummy variables

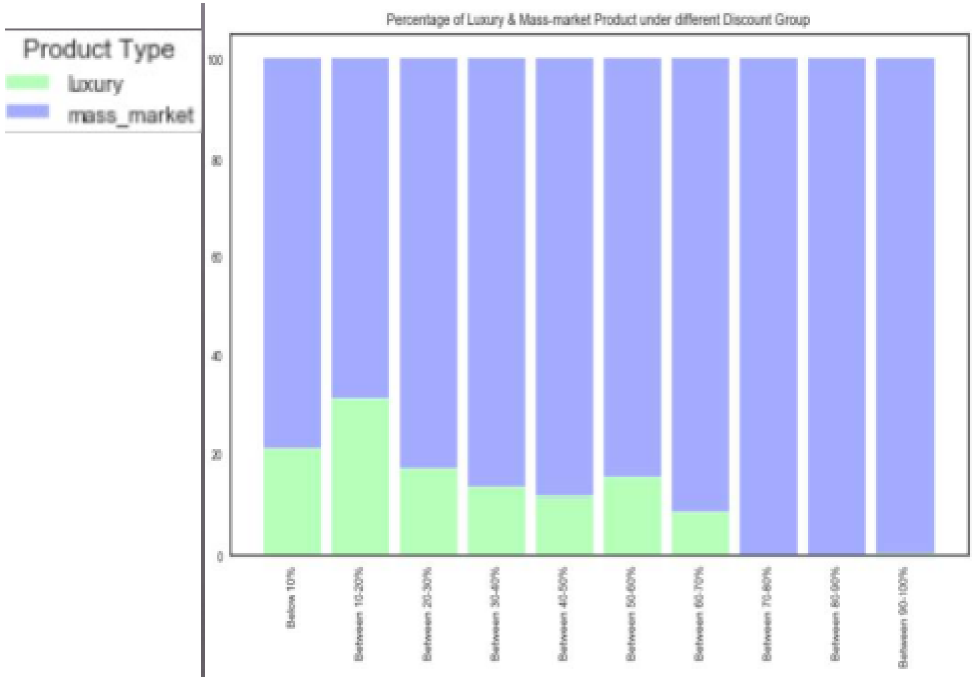
Additional Data Preparation for Model Building

- Merging Order dataset with all other secondary dataframes
- Extracting 3 separate dataframes for 3 product subcategories – camera accessory, home audio and gaming accessory
- Roll Up daily Order Data to Weekly Level by aggregating the numeric variables based on Week#
- Scaling and dividing the master dataframes into train and test datasets for all 3 product subcategories

Feature Engineering

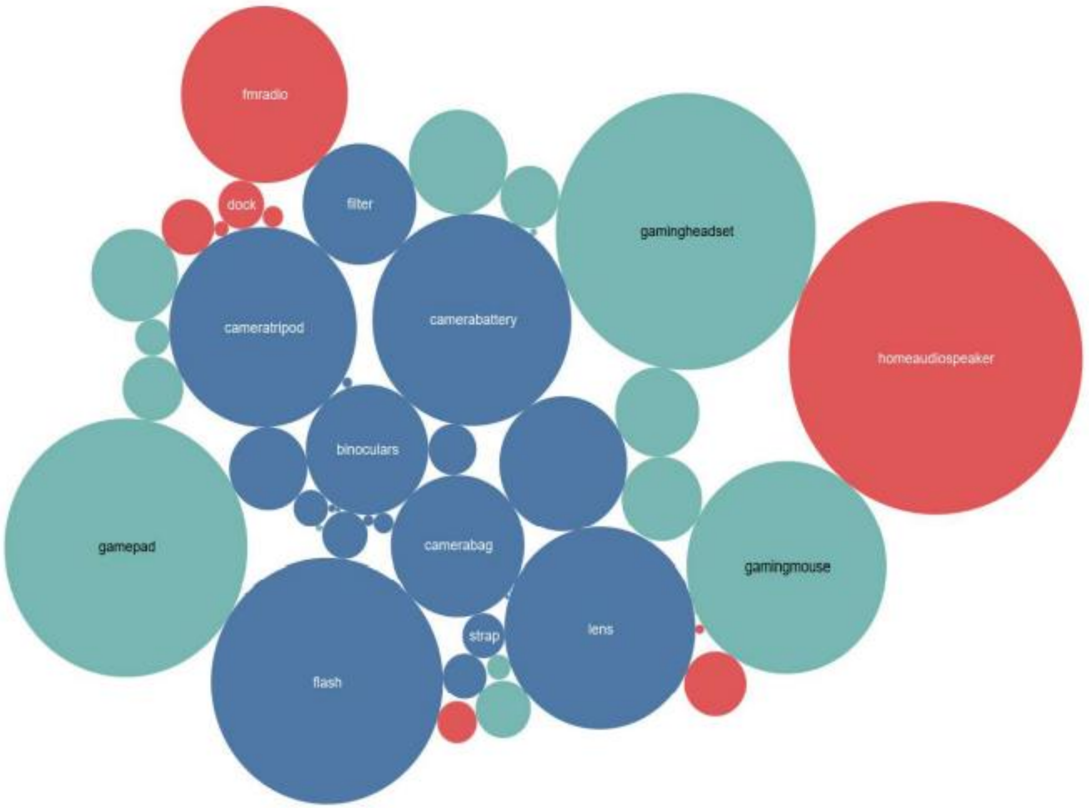
Attribute	Description
Week#	Generating Week# column from the order date
List Price	List Price = GMV * Units
Payday Week	If Payday falls within the week, then payday week = 1, else 0
Holiday Week	If Holiday falls within the week, then payday week = 1, else 0
Product Type	If GMV value is greater than 80 percentile, then luxury, else mass-market
Discount%	Discount% = 100*(product_mrp – list_price)/product_mrp
SMA#	3 & 5-weeks Simple Moving Average for all Advertising media channels, NPS and Stock Index
EMA#	8-weeks Exponential Moving Average for all Advertising media channels
Lag Variables	Lag variables(lag by 1, 2 & 3 days) for all KPIs were taken for Distributive Lag Models
Adstock Values	Calculating Ad Stock values for all Advertising media(assuming ad stock rate as 60%)

Visualizations

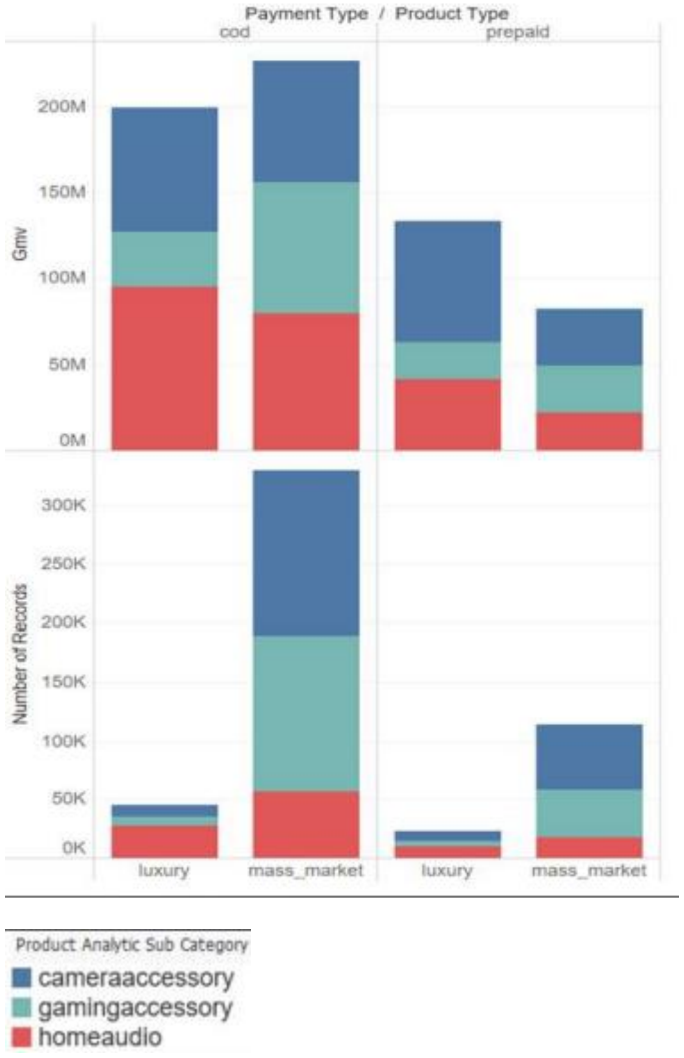


The median discount percentage offered for luxury items is less compared to that of Mass Market Products. This is a known trend among luxury products or luxury brands, to offer limited discounts, to retain the exclusivity of their products.

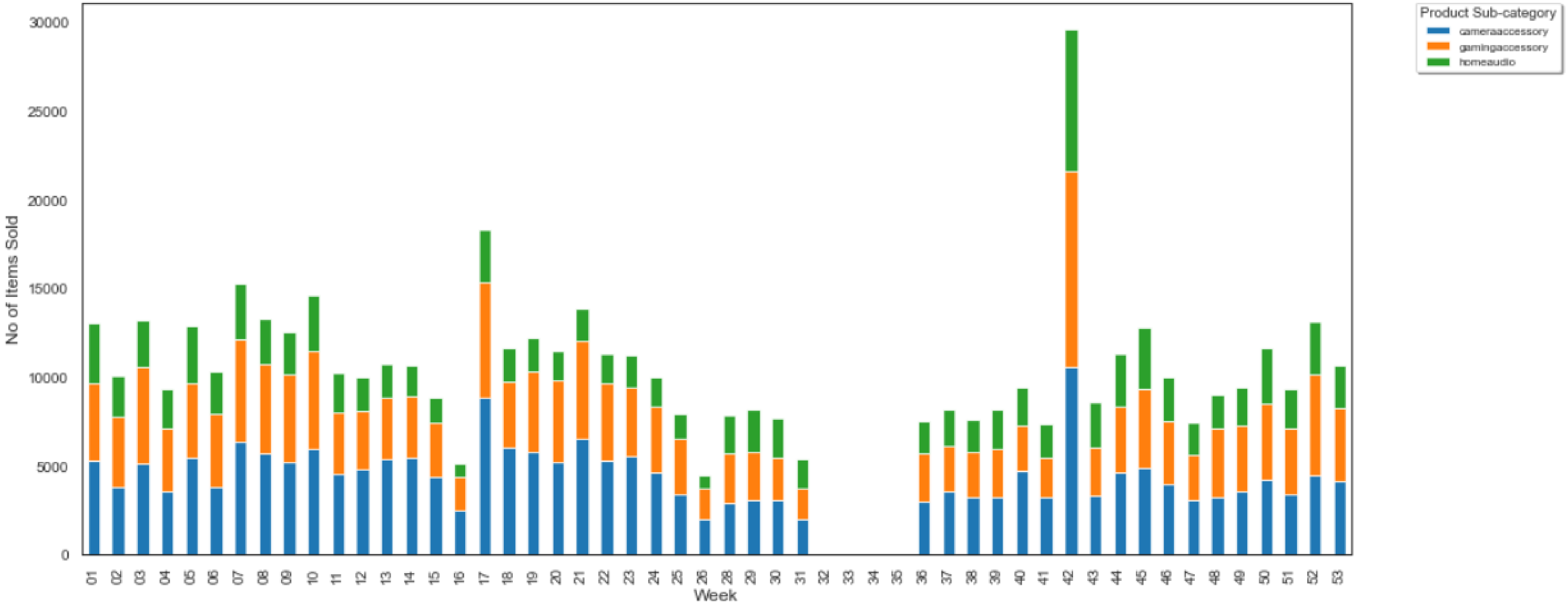
Product Verticals with Most Number of Sales



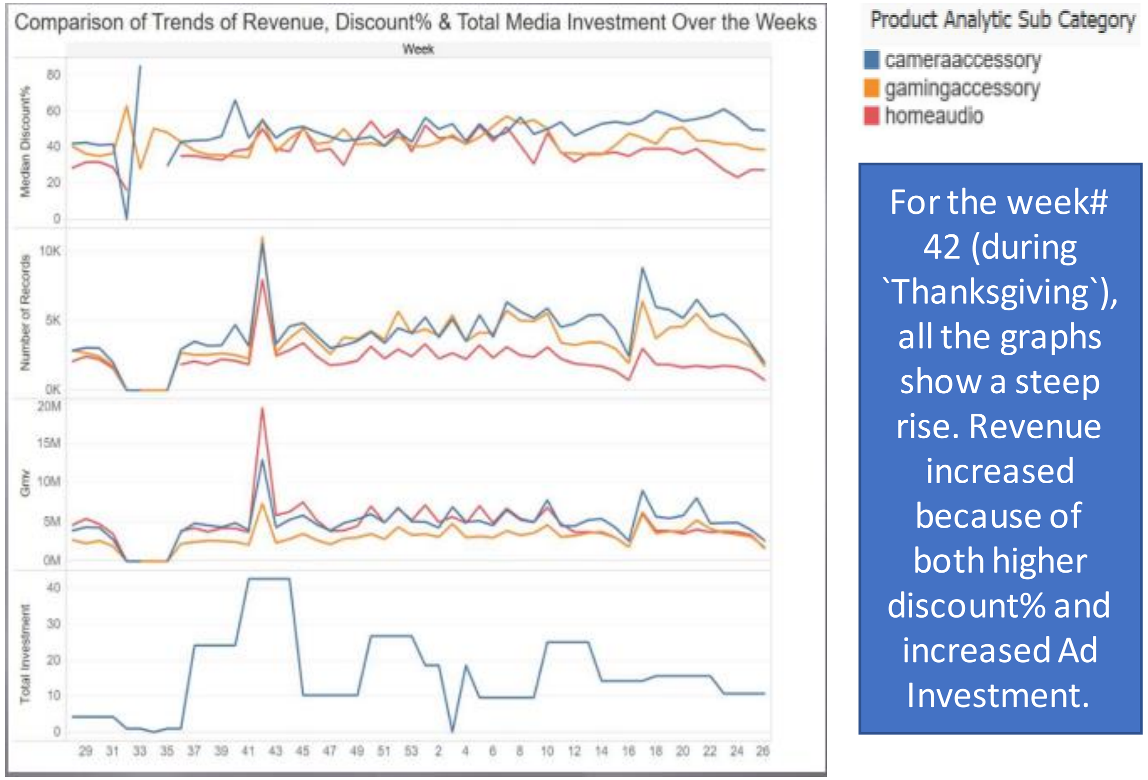
Analyzing how Sales Amount and Revenue vary based on Payment Types & Product Types



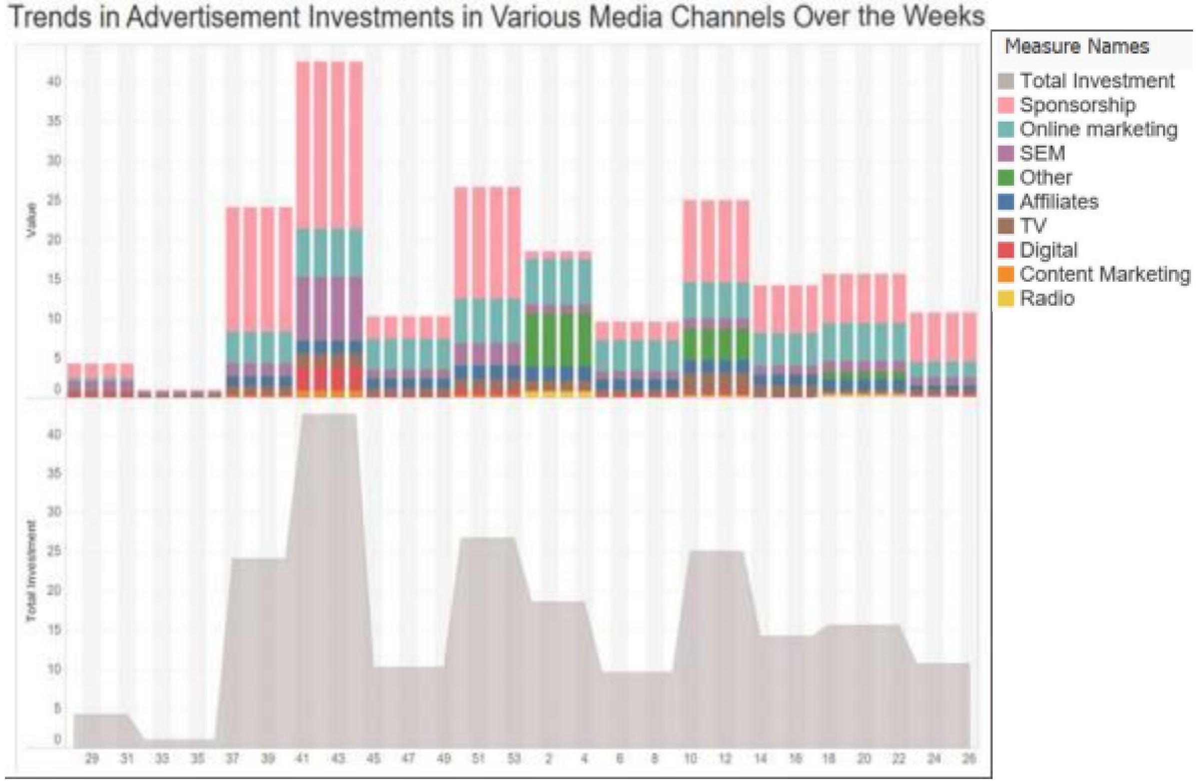
- Home Audio fetches more revenue both for prepaid and COD products even though they are sold to a lesser extent
- Audio Speaker contributes mostly to the revenue fetched by the category
- COD products in general sell more and bring in more revenue



The sale on the 42nd week (`Thanksgiving week`) is maximum. Overall, October has seen most no of items being sold.

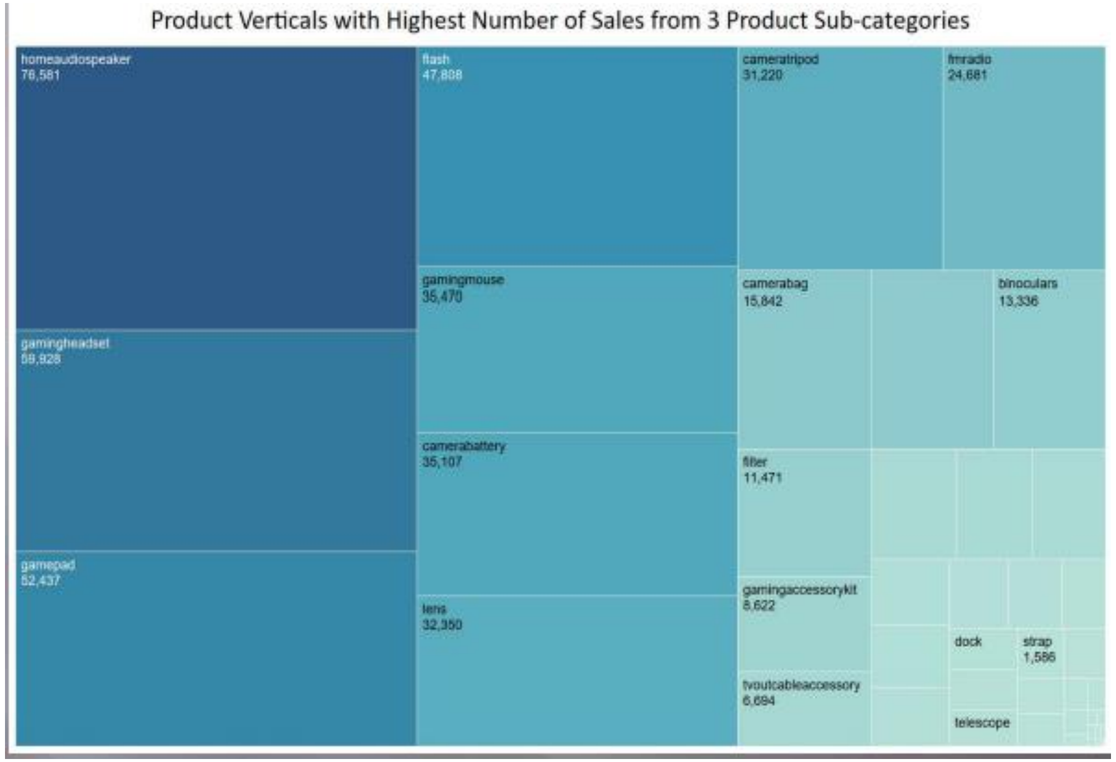


For the week# 42 (during `Thanksgiving`), all the graphs show a steep rise. Revenue increased because of both higher discount% and increased Ad Investment.

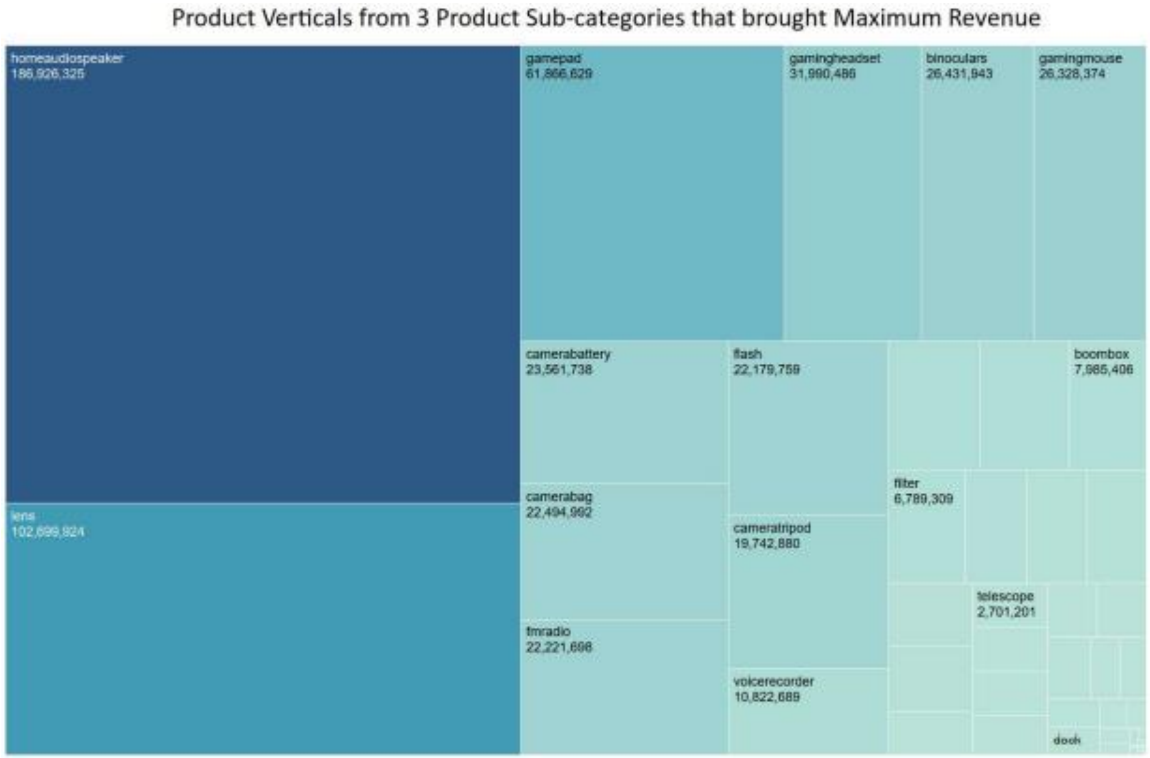


For the weeks 32 - 35(August), Revenue generated was the lowest from all 3 product subcategories. This can be observed as a direct relation to minimum amount of total investment in Ads. Discount was also lowest for all products apart from camera accessories. Post this dip in revenue, discount% was increased to bring about higher sales. This increase in Discount% was observed most in the case of gaming accessories

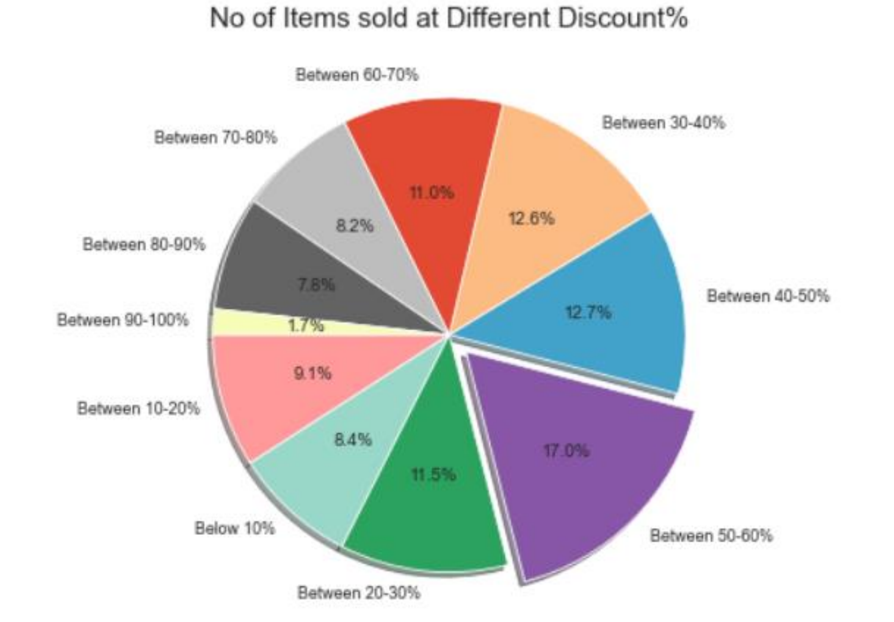
In general the average discount% offered for home audio products is lesser compared to that of the other product subcategories.
Over the past year, bulk of the Ad Investment has been made in Sponsorships followed by Online Marketing & Search Engine Marketing(specially during Thanksgiving).



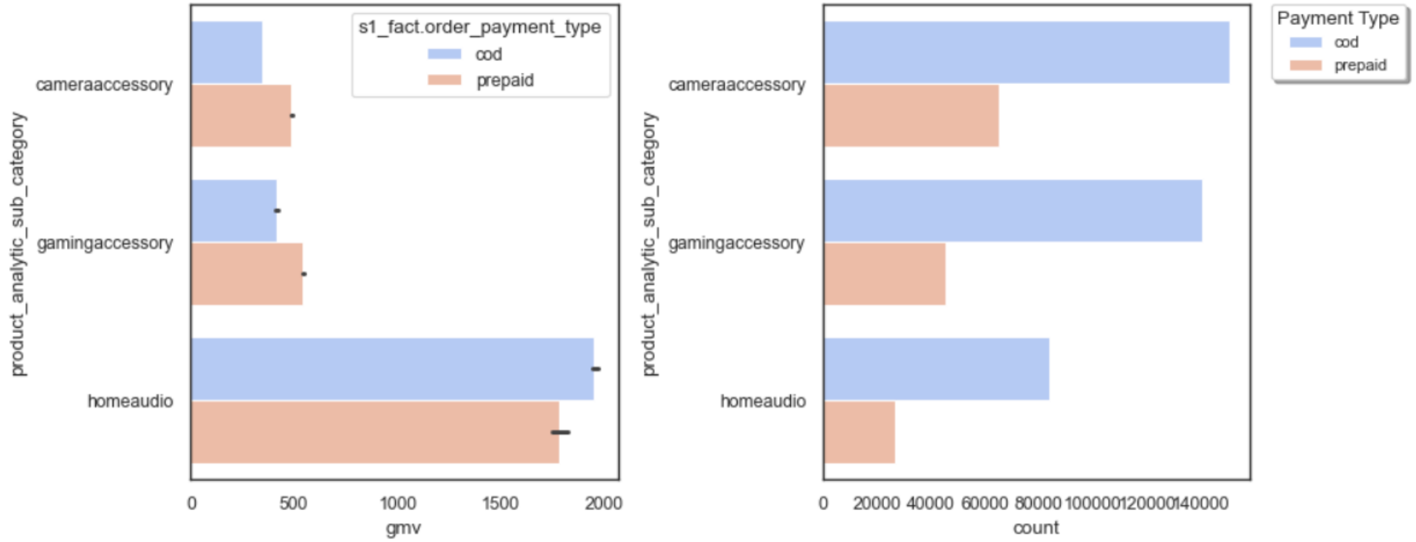
Home Audio Speaker under Home Audio segment brought the largest revenue followed by Camera Lens under Camera Accessory & Gamepad under Gaming Accessory.



Home Audio Speaker under Home Audio segment had the most no of sales followed by Gaming Headset & Gamepad under Gaming Accessory



Most of the sales take place when Discount% is between 50-60%.



Except for Home Audio Products, for the other 2 product sub categories, we observe that the median Revenue from Prepaid orders is more than that from COD products even though the no of products sold is way higher in case of COD products for all categories.

Description of models

Additive

- Linear model is used to capture the current effect of several KPIs. This model assumes an additive relationship between the different KPIs. Hence their impacts are also additive towards the dependent Y variable.
- The equation: $Y = \alpha + \beta 1At + \beta 2Pt + \beta 3Dt + \beta 4Qt + \beta 5Tt + \epsilon$

Multiplicative

- Multiplicative model is used when there are interactions between the KPIs. To fit a multiplicative model, take logarithms of the data, then analyse the log data as before
- $Y = e^\alpha . X1^{\beta 1} . X2^{\beta 2} . X3^{\beta 3} . X4^{\beta 4} . X5^{\beta 5} + \epsilon$
- $\ln Y = \alpha + \beta 1\ln(X1) + \beta 2\ln(X2) + \beta 3\ln(X3) + \beta 4\ln(X4) + \beta 5\ln(X5) + \epsilon'$

Koyck Model

- Koyck model is used to capture the carry-over effect of different KPIs, ie.to model the current revenue figures based on the past figures of the KPIs. The Koyck tells us that the current revenue generated is not just influenced by the different independent attributes, but also because of the revenue generated over the last periods
- $Yt = \alpha + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \epsilon$
- $Yt = \alpha + \mu Yt\text{-}1 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \epsilon$

Distributive Lag Model (Additive)

- In the distributed lag model, not only is the dependent variable entered in its lagged version, but the independent variables are as well. This is a more generalizable model and captures the carry-over effect of all the variables:
- $Yt = \alpha + \mu 1Yt\text{-}1 + \mu 2Yt\text{-}2 + \mu 3Yt\text{-}3 + \ldots. + \beta 1X1t + \beta 1X1t\text{-}1 + \beta 1X1t\text{-}2 + \ldots. + \beta 2X2t + \beta 2X2t\text{-}1 + \beta 2X2t\text{-}2 + \ldots. + \beta 3X3t + \beta 3X3t\text{-}1 + \beta 3X3t\text{-}2 + \ldots. + \beta 4X4t + \beta 4X4t\text{-}1 + \beta 4X4t\text{-}2 + \ldots. + \beta 5X5t + \beta 5X5t\text{-}1 + \beta 5X5t\text{-}2 + \ldots. + \epsilon$

Distributive Lag Model (Multiplicative)

- Distributive Lag Model(Multiplicative) will help us capture the interactions between current and carry over effects of the KPIs
- $Yt = \alpha + \mu 1\ln(Yt\text{-}1) + \mu 2\ln(Yt\text{-}2) + \mu 3\ln(Yt\text{-}3) + \ldots. + \beta 1\ln(X1t) + \beta 1\ln(X1t\text{-}1) + \beta 1\ln(X1t\text{-}2) + \ldots. + \beta 2\ln(X2t) + \beta 2\ln(X2t\text{-}1) + \beta 2\ln(X2t\text{-}2) + \ldots. + \beta 3\ln(X3t) + \beta 3\ln(X3t\text{-}1) + \beta 3\ln(X3t\text{-}2) + \ldots. + \beta 4\ln(X4t) + \beta 4\ln(X4t\text{-}1) + \beta 4\ln(X4t\text{-}2) + \ldots. + \beta 5\ln(X5t) + \beta 5\ln(X5t\text{-}1) + \beta 5\ln(X5t\text{-}2) + \ldots. + \epsilon'$

Top 5 KPIs in each product sub-category

Product Sub-category	Linear Regression Model	R-square on Test Dataset	Mean Square Error	Top 5 KPIs
cameraaccessory	Multiplicative with CV	0.91	0.09	product_vertical_lens (**0.181**)
				product_vertical_camerabattery (**0.160**)
				is_mass_market (**0.149**)
				product_vertical_camerabatterycharger (**0.121**)
				TV (**0.105**)
gamingaccessory	Multiplicative with CV	0.94	0.06	product_vertical_gamingheadset (**0.250**)
				is_mass_market (**0.234**)
				product_vertical_gamingmouse (**0.224**)
				product_vertical_gamepad (**0.211**)
				Online marketing_SMA_3 (**0.157**)
cameraaccessory	Multiplicative with CV	0.86	0.14	product_vertical_homeaudiospeaker (**0.469**)
				is_mass_market (**0.289**)
				product_vertical_fmradio (**0.224**)
				Radio_Ad_Stock (**0.147**)
				Sponsorship (**0.121**)

Model Equation

Considering the top 5 KPIs from the models for our 3 product subcategories, we can see that the equation of our best fitted lines as follows:

Camera Accessory:

Revenue = 0.0 + (0.181 × **product_vertical_lens**) + (0.160 × **product_vertical_camerabattery**) + (0.149 × **is_mass_market**) + (0.121 × **product_vertical_camerabatterycharger**) + (0.105 × **TV**)

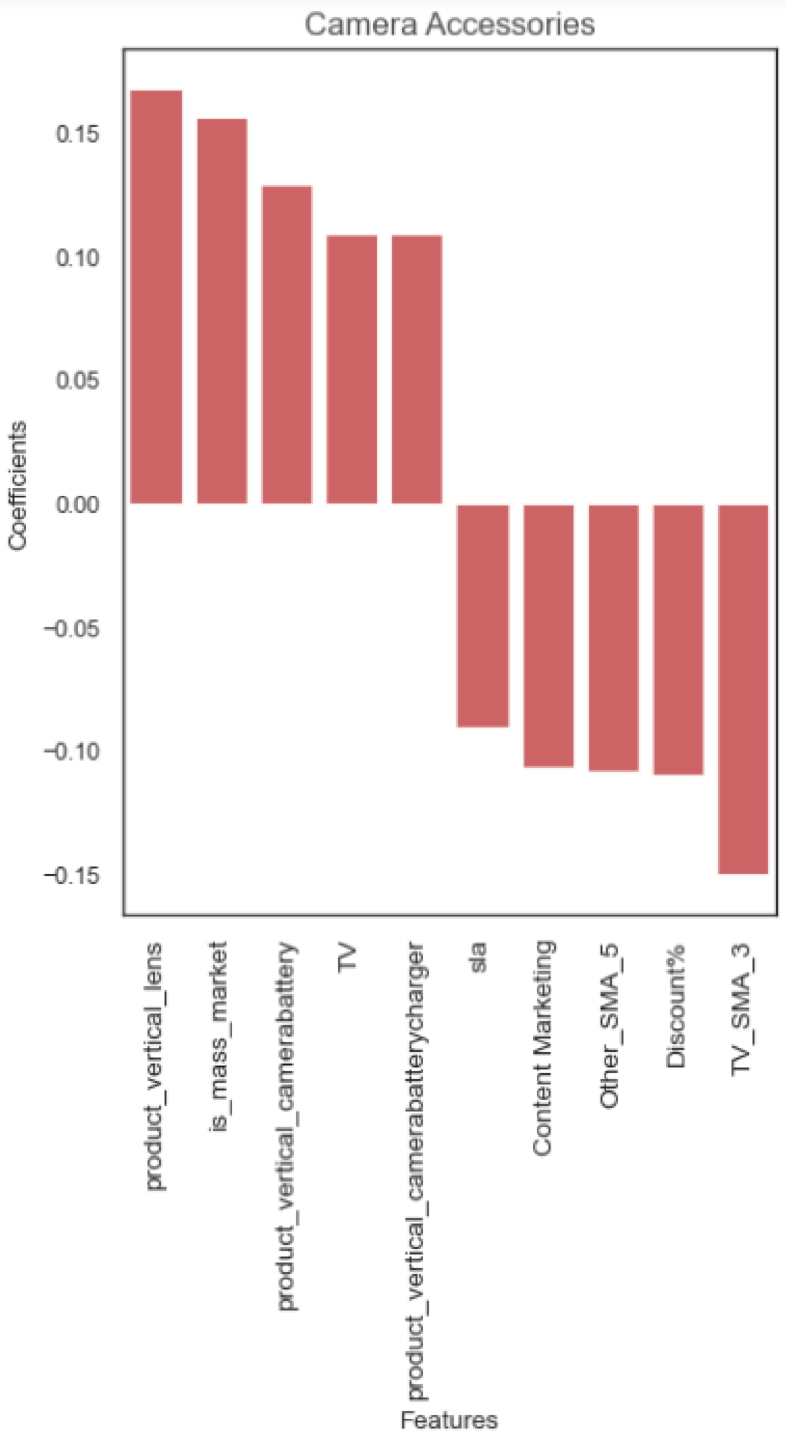
Gaming Accessory:

Revenue = 0.0 + (0.250 × **product_vertical_gamingheadset**) + (0.234 × **is_mass_market**) + (0.224 × **product_vertical_gamingmouse**) + (0.211 × **product_vertical_gamepad**) + (0.157 × **Online marketing_SMA_3**)

Home Audio:

Revenue = 0.0 + (0.469 × **product_vertical_homeaudiospeaker**) + (0.289 × **is_mass_market**) + (0.224 × **product_vertical_fmradio**) + (0.147 × **Radio_Ad_Stock**) + (0.121 × **Sponsorship**)

This equation implies how much the `Revenue` will grow with a unit growth in any of these independent KPIs with all other KPIs held constant.



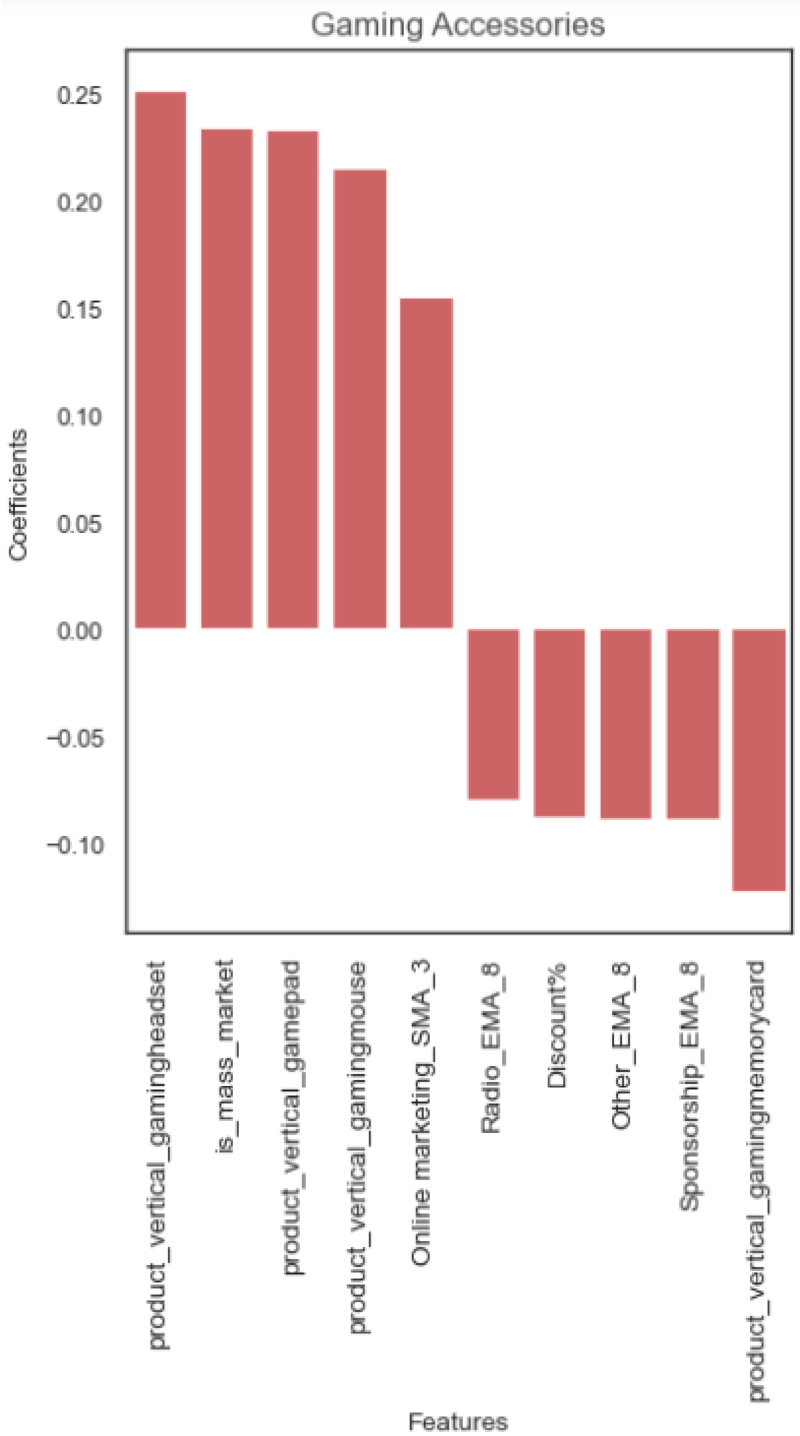
Recommendation

• Company should promote `Lens`, `Camera Batteries` & `Camera Battery Chargers` as they fetch the highest revenue.

•Advertisement spends on TV has a positive impact on revenue. One unit of TV spend can boost the revenue by 0.105 units. Content Marketing spends on the other hand impacts negatively.

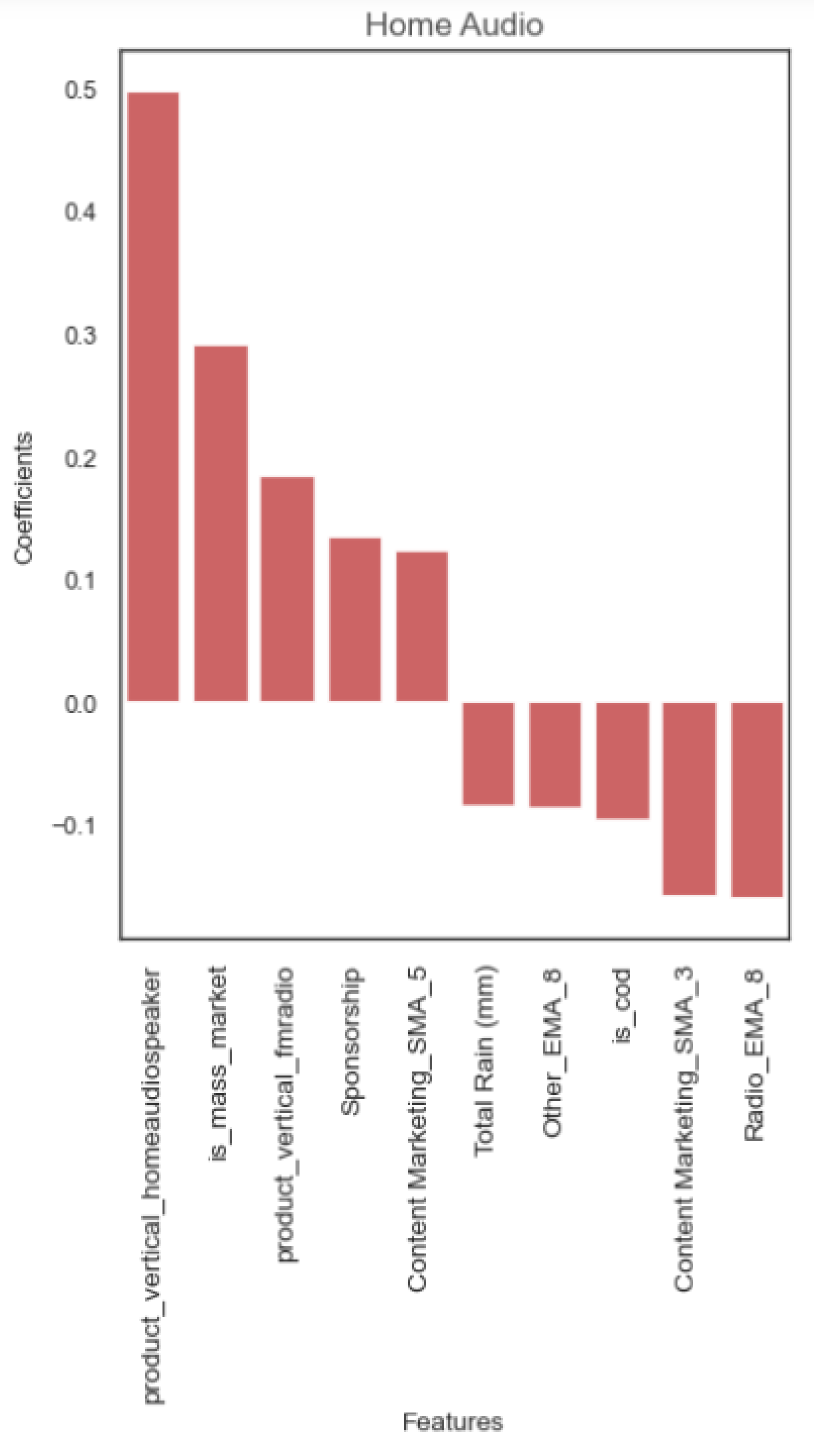
•`Mass-market` products are better contributors to the increased revenue in comparison to the Luxury products.

•Higher percentage of Discounts in general given for this sub category works adversely towards bringing down the revenue.



Recommendation

- Company should promote `Gaming Headset`, `Gaming Mouse` & `Gamepad` as they fetch the highest revenue. On the contrary, `Gaming Memory Cards` results in loss.
- Advertisement spends on Online Marketing, Radio & Others have a positive cumulative impact on revenue. Sponsorship spends on the other hand has a negative cumulative effect.
- `Mass-market` products are better contributors to the increased revenue in comparison to the Luxury products.
- Higher percentage of Discounts in general given for this sub category works adversely towards bringing down the revenue



Recommendation

- Company should promote `Home Audio Speakers` & `FM Radios as they fetch the highest revenue.
- `Mass-market` products are better contributors to the increased revenue in comparison to the Luxury products.
- Radio Adstock (carry over effect of Radio Advertisement) spends helps to boost the revenue to a significant extent.
- Advertisement spends on Sponsorship has a positive impact on revenue. Content Marketing spends on the other hand impacts negatively.
- COD payments in general for this sub category are bad in bringing down the revenue.

General Observation

- Most of the sales take place when Discount% is between 50-60%. However, that doesnt necessarily help in boosting the revenue. EDA shows that an average discount% between 10-20% is the most profitable for the company specially among luxury items.
- In general most of the Home Audio items sold are luxury items and hence, customers prefer to use COD instead of paying upfront.
- During festive time(eg. Thanksgiving) more investment is made on Advertisement and good promotional offers were rolled out. This usually boosts the revenue. However just providing discounts without properly adertising for it on several media channels doesn't help. We have seen that for the weeks 32 - 35(August), revenue generated was the lowest from all 3 product subcategories even though median discount% was raised after the initial drought. In fact, this dip in revenue can be observed as a direct relation to minimum amount of total investment in Ads during the given timeframe.