CAPSTONE PROJECT ON E-COMMERCE

By Abhijeet Srivastava EDS C32 BA

CONTENTS

BUSINESS OBJECTIVE

- To create a market mix model for ElecKart (a Canadian e-commerce firm based out of Ontario) for 3 product sub-categories - Camera Accessory, Gaming Accessory and Home Audio Accessory - to observe the actual impact of various marketing variables over one year (July 2015 to June 2016) and recommend the optimal budget allocation for different marketing platforms for the next year.
- The business objective can be divided into 3 sub parts:
  1. Performance driver analysis: What KPIs drive the performance?
  2. Impact analysis on ROI on marketing spends: What is the return on invest across all marketing platforms?
  3. Optimize marketing spends: How to best allocate marketing budget to maximize revenue across all 3 product sub categories
- This project was conducted using the CRISP DM framework: Deployment is out of scope for this project

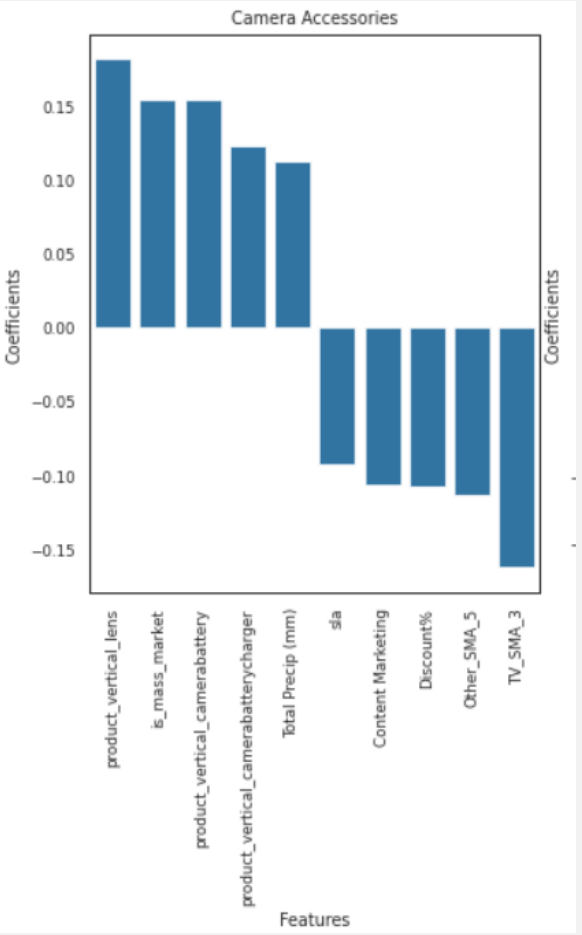
RECOMMENDATIONS I

- Most of the sales take place **when Discount% is between 50-60%.** However, that doesn't 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 holiday season (eg. Thanksgiving) more investment is made on **Ads and good promotional offers were rolled out. This boosts the revenue.**
- However, just providing discounts without properly advertising for it on several media channels doesn't help. 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.**

RECOMMENDATIONS 2

Camera Accessory:

- ElecKart should promote **Lens, Camera Batteries & Camera Battery Chargers** as they bring the highest revenue.
- Content Marketing spends 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 work adversely towards bringing down the revenue.**



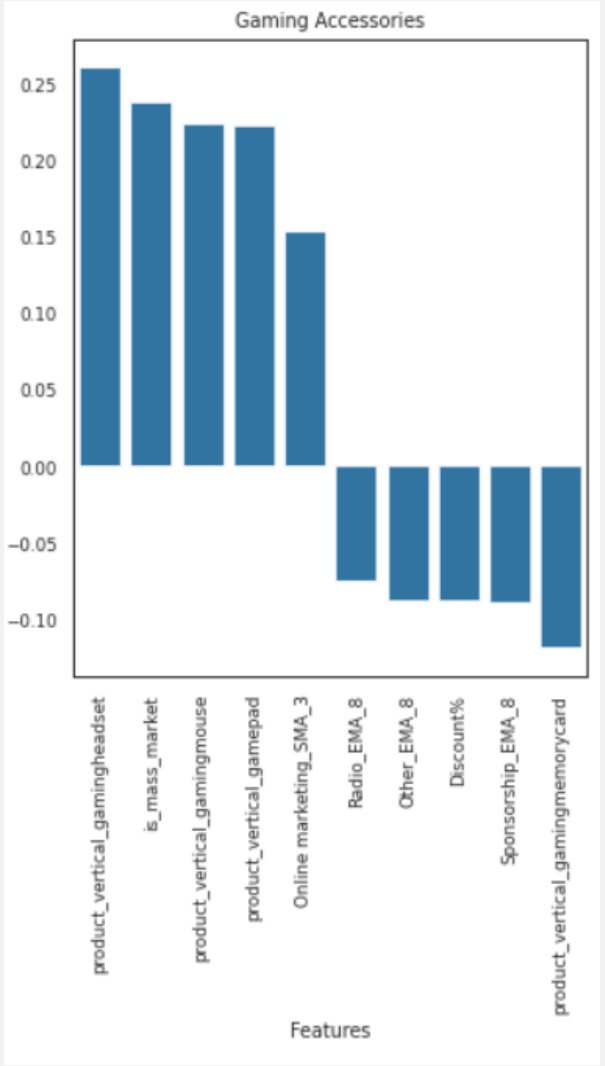
Camera Accessory:

Revenue = 0.0 + (0.181 × **product_vertical_lens**) + (0.154 × **is_mass_market**) + (0.154 × **product_vertical_camerabattery**) + (0.122 × **product_vertical_camerabatterycharger**) + (0.112 × **Total Precip (mm)**

RECOMMENDATIONS 3

Gaming Accessory:

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



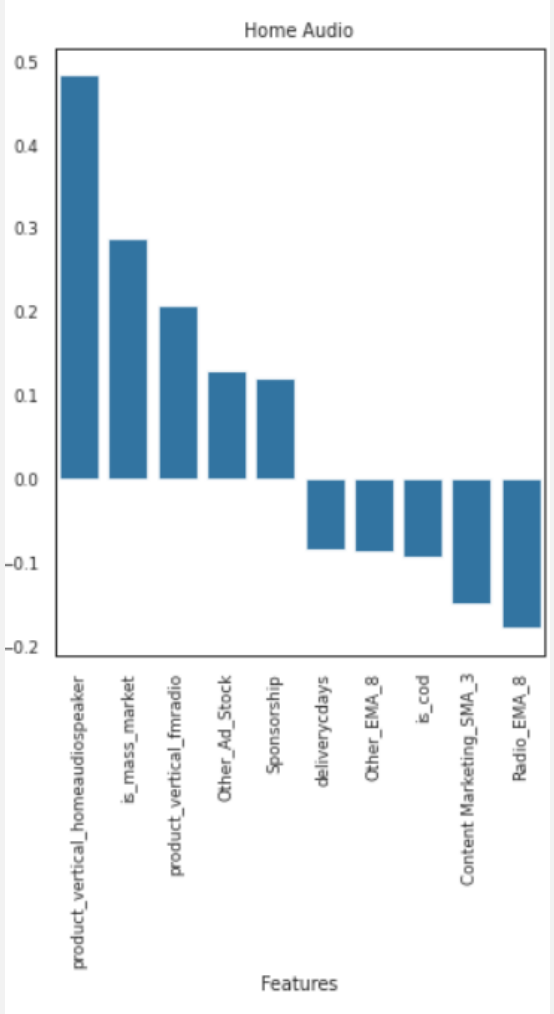
Gaming Accessory:

Revenue = 0.0 + (0.260 × **product_vertical_gamingheadset**) + (0.237 × **is_mass_market**) + (0.223 × **product_vertical_gamingmouse**) + (0.222 × **product_vertical_gamepad**) + (0.154 × **Online marketing_SMA_3**)

RECOMMENDATIONS CONTINUED

Home Audio:

- ElecKart 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.
- Ad 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.**



Home Audio:

Revenue = 0.0 + (0.482 × **product_vertical_homeaudiospeaker**) + (0.288 × **is_mass_market**) + (0.207 × **product_vertical_fmradio**) + (0.130 × **Other_Ad_Stock**) + (0.121 × **Sponsorship**)

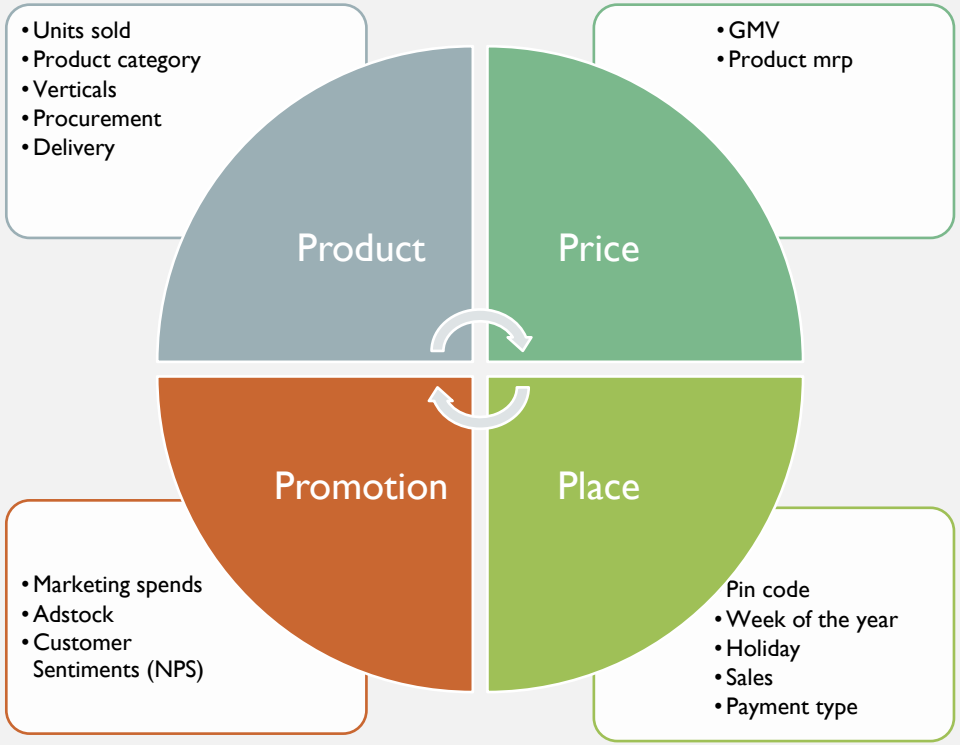
APPENDIX

UNDERSTANDING THE DATA

- A brief summary of all the files available:

1. Main consumer file with order details for 2015 and 2016
2. Media investment file with budget allocations across all platforms for the year
3. Sales calendar giving details for promotional/discount offers
4. NPS showing net promotion score and ElecKart's stock index for the year
5. Climate file(s) with weather details for the state of Ontario

4 P's of Marketing



DATA PREP AND CLEANING

Handling Incorrect values	• Impute "\N" value in deliverybdays & deliverycdays by 0 • Treat incorrect GMV values (where gmv > product_mrp * units) by replacing the faulty MRP values with GMV/units • Handle Negative values for product_procurement_sla, deliverybdays & deliverycdays by dropping them • Handle large values(0.3%) for product_procurement_sla by dropping 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 • converted these whitespaces to NaNs and then dropped these values
Handling Duplicate values	• To handle duplicate data, converted all columns to lower case, and checked for duplicates using .duplicated() found 99283 (6.33%) rows that are duplicates. • Dropped them
Dropping Insignificant columns	• Drop Columns with Single Unique Value (as it doesn't add any information to the analysis) • Drop some of the 'Id' Columns which are insignificant to the analysis

DATA PREP AND CLEANING CONTINUED

Outlier Treatment	• Since some records have been dropped, to not lose more data, outlier values were not deleted • For variables - 'SLA', 'deliverybdays', 'deliverybdays', 'gmv', 'product_mrp', 'list_price' where outliers are present, CAPPED the values above 99 percentile to the value corresponding to 99 percentile • SO, outliers couldn't affect the predictive model while at the same time there was enough data to build a generalizable model
Handling data outside date range	• For this project, data has to be from July 2015 till June 2016. • All other rows (592) were dropped
Encoding for Categorical Columns	• Binary encoding was used for categorical columns with 2 levels • One Hot Encoding for categorical variable with multiple levels by creating dummy variables
Merging and Aggregation	• Merge Consumer dataset with all other secondary dataframes (nps, climate, media investment) • Extract 3 separate dataframes for the 3 product subcategories - camera accessory, gaming accessory and home audio • Aggregate daily Order Data to Weekly Level by aggregating the numeric variables based on Week# • Scale and divide the master dataframes into train and test datasets for all 3 product subcategories

FEATURE ENGINEERING

Week #
Generated 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 > 80 percentile, then luxury, else mass-market

Discount:
Discount% = 100*(product_mrp–list price) / product_mrp

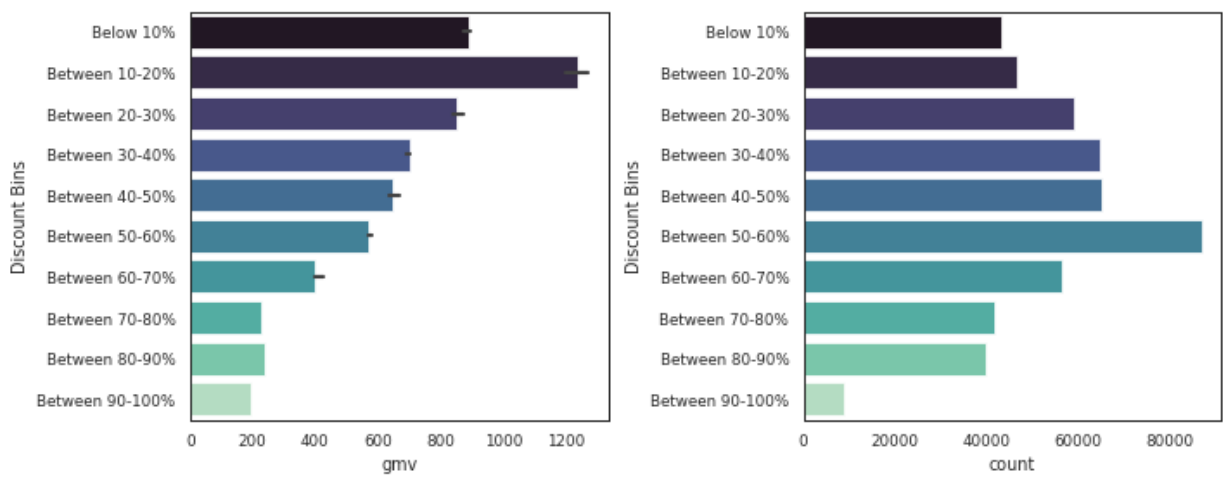
Lag Variables:
Lag variables(lag by 1, 2 & 3 days) for all KPIs can be taken for Distributive Lag Models

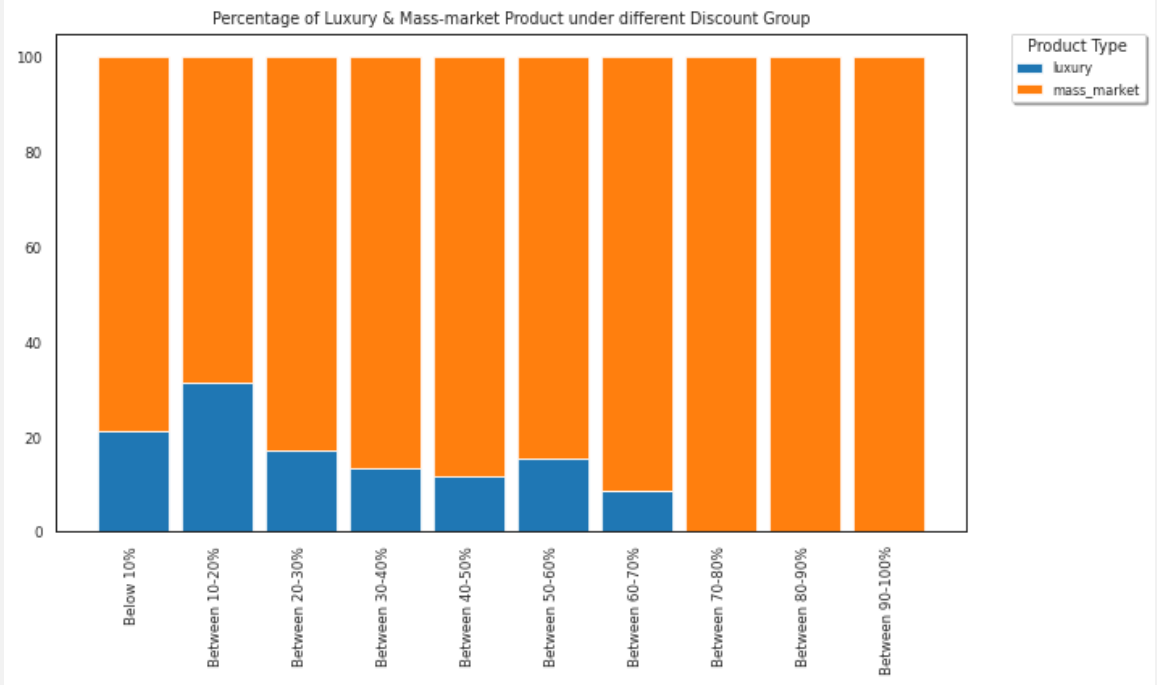
Adstock Values:
Calculate Ad Stock values for all Advertising media(assuming ad stock rate as 60%)

VISUALIZATIONS

Discount% across product sub categories and product type

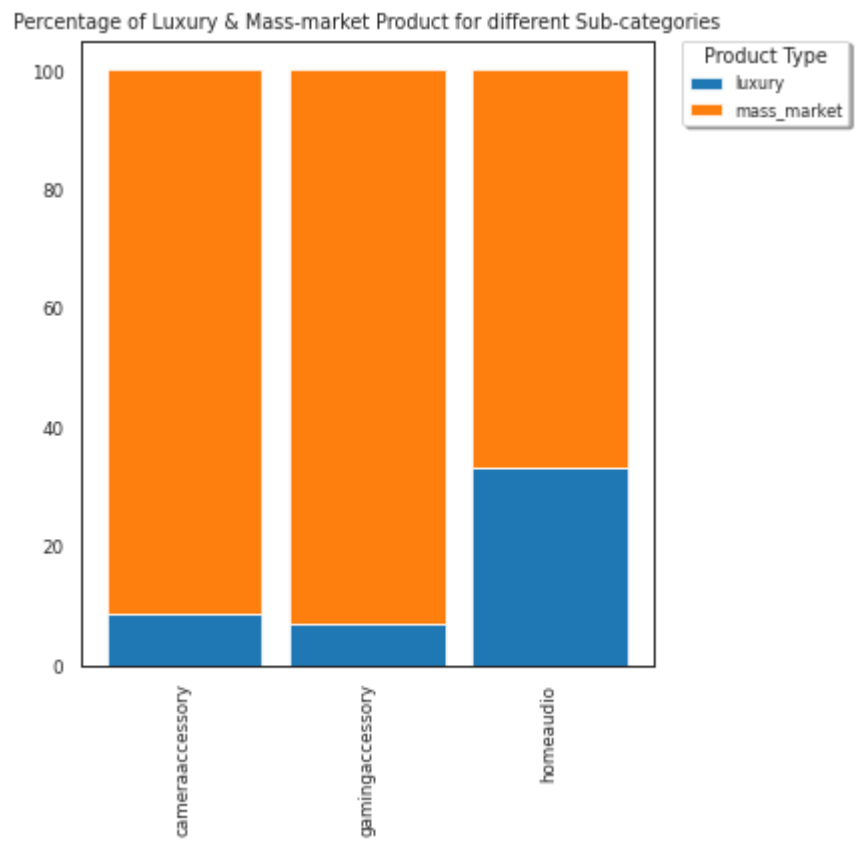
- Median Revenue is maximum when Average discount% is between 10-20%. But beyond that, average revenue slowly starts to decline.
- The sales on the other hand shows a steady increase with increase in Discount % till it peaks at 50-60% after which it starts to fall
- Max number of luxury products were offered a discount between 10-20%. This shows that at higher discount, although the sales are good, the revenue collapses signifying a loss for the company. **An average discount of 10-20% is the most profitable for the company.**
- The median discount percentage offered for luxury items is less compared to that of Mass Market Products. This s a known trend among luxury products or luxury brands to offer limited or no discounts to retain the exclusivity of their products.



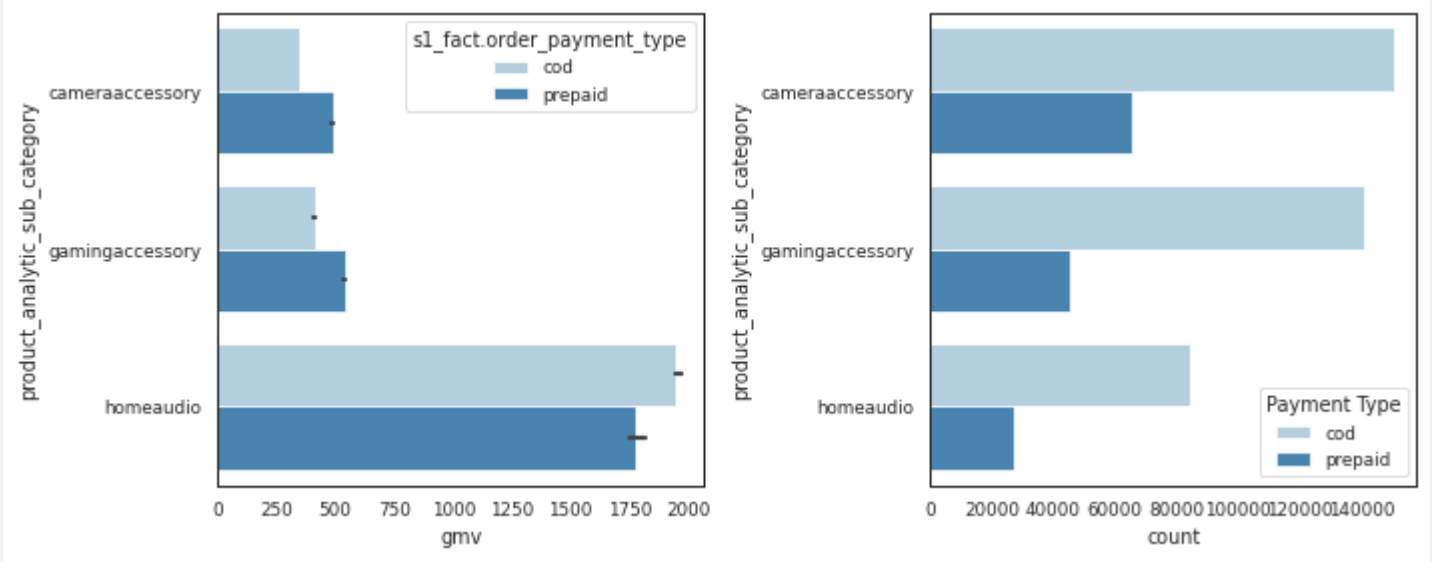


The median discount percentage offered for luxury items is less compared to that of Mass Market Products. This is a known trend among luxury products or luxury brands, to offer limited discounts, to retain the exclusivity of their products.

VISUALIZATIONS CONTINUED

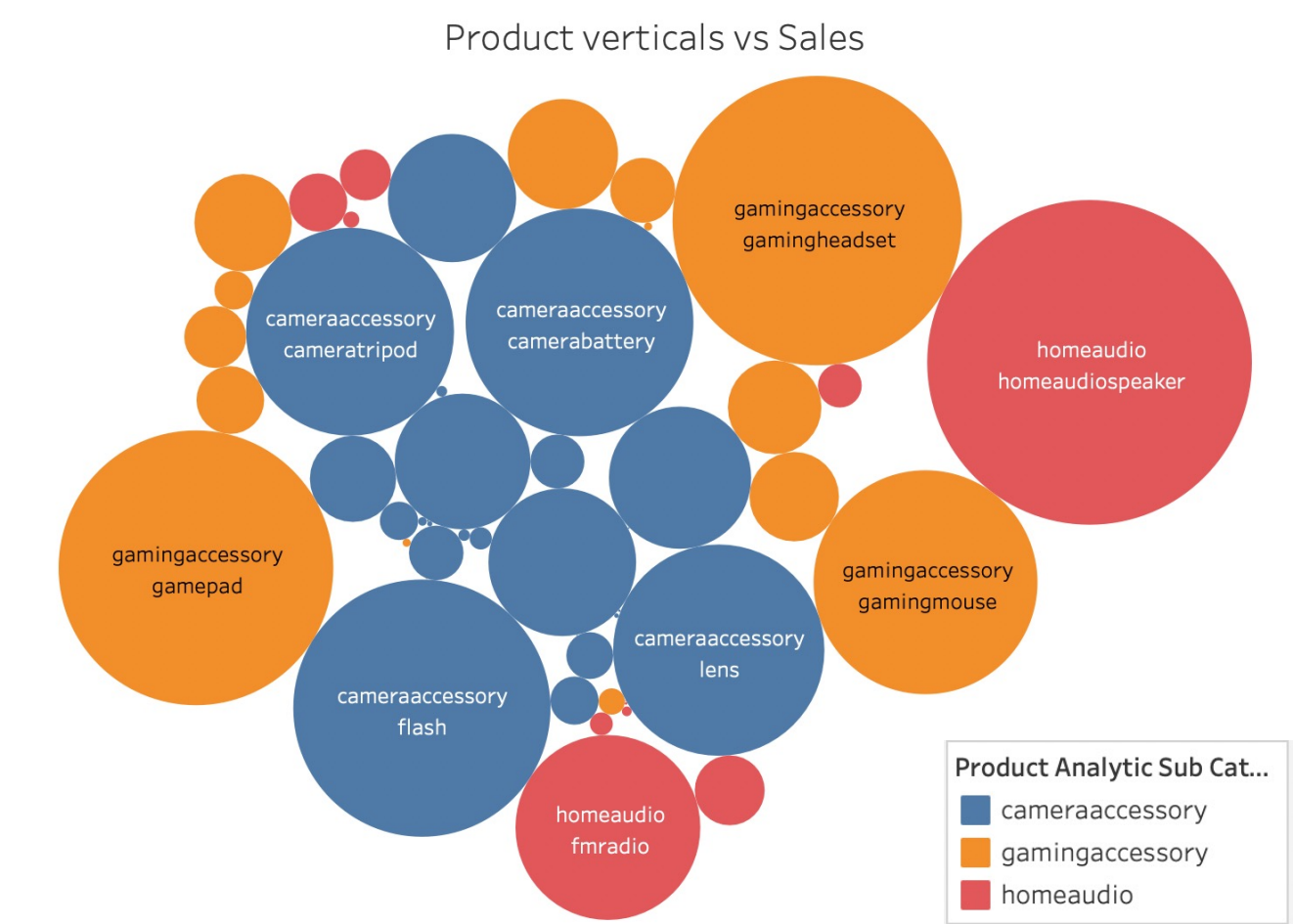


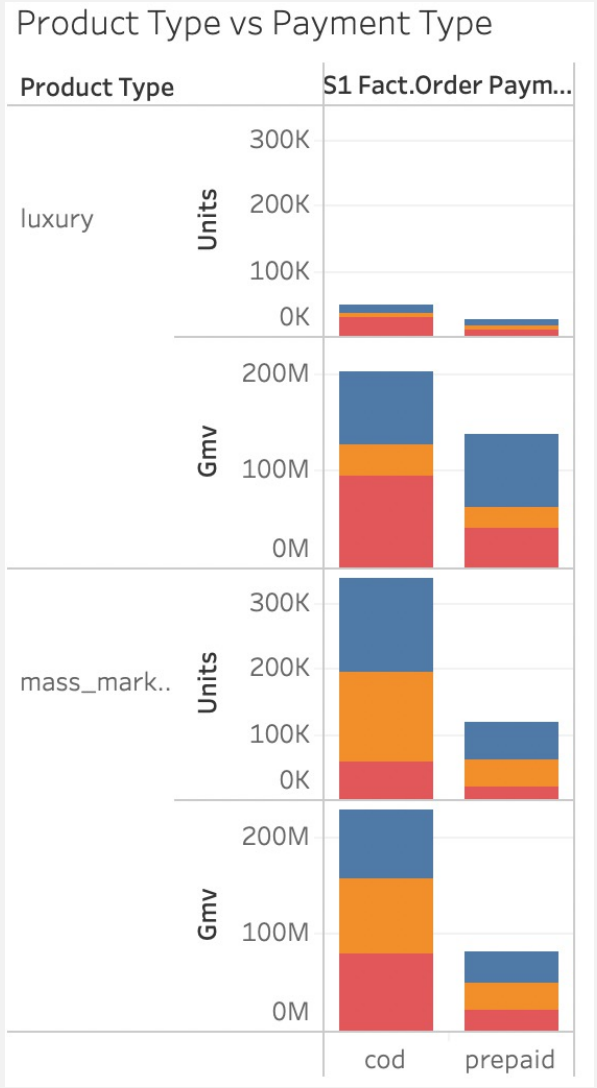
Percentage of luxury products under Home Audio is much more compared to the other sub categories.



- Median gmv for camera accessory, gaming accessory is more for prepaid than cod
- It is the opposite for home audio
- It is interesting to note that no of units sold is significantly higher for cod than prepaid across all 3 sub categories

VISUALIZATIONS CONTINUED



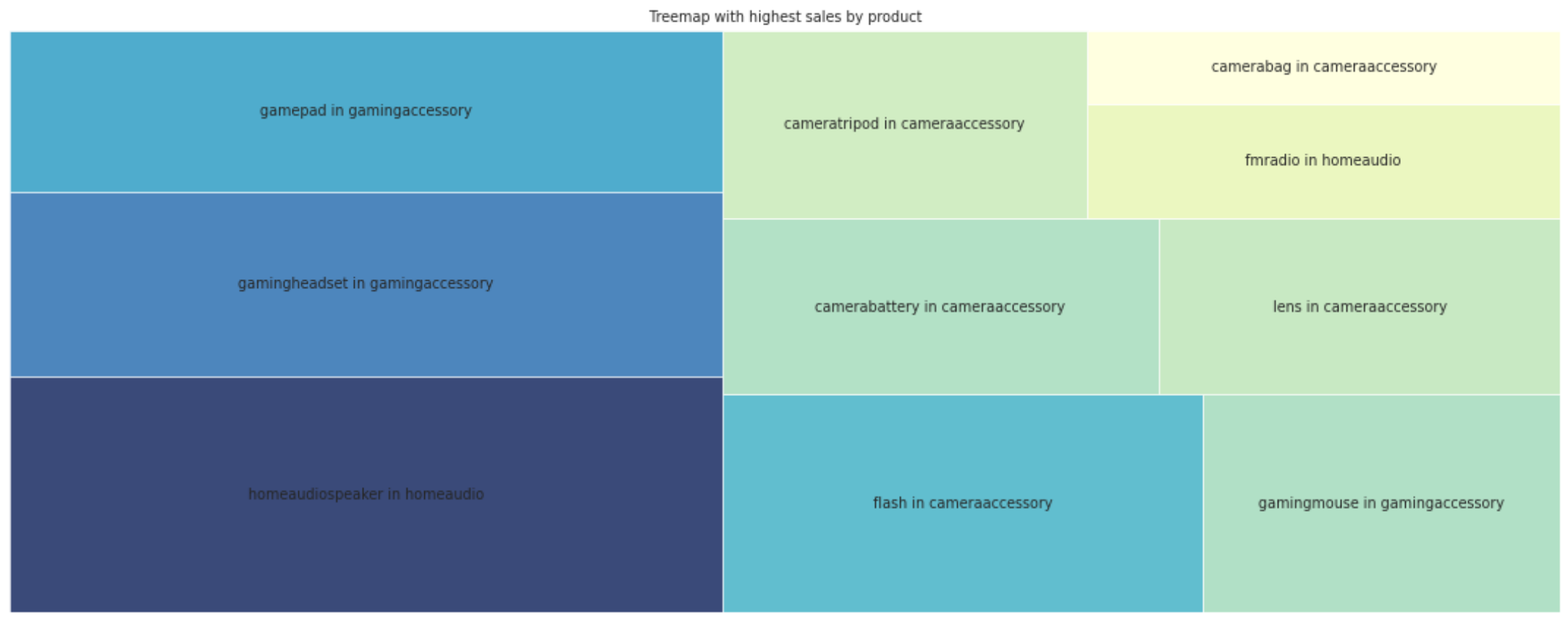


- Home Audio fetches more revenue both for prepaid and COD products even though they are sold to a lesser extent
- Audio Speaker contributes mostly to the revenue fetched by the category
- COD products in general sell more and bring in more revenue

VISUALIZATIONS CONTINUED

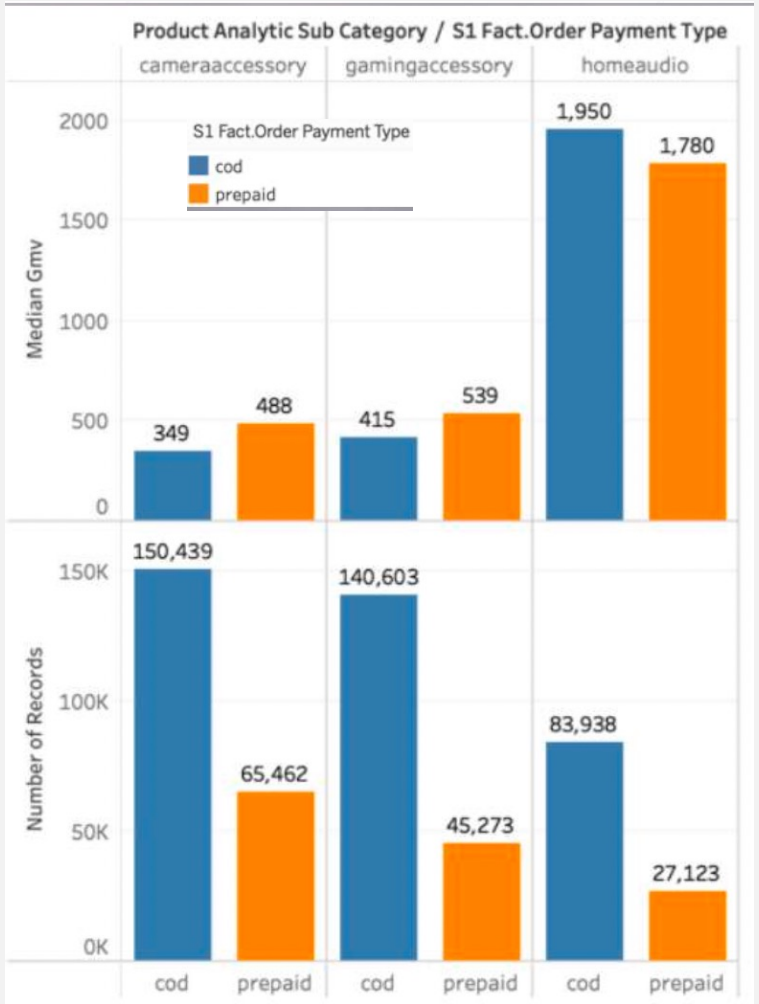
Home Audio Speaker under Home Audio segment brought the largest revenue followed by Camera Lens under Camera Accessory & Gamepad under Gaming Accessory.

VISUALIZATIONS CONTINUED



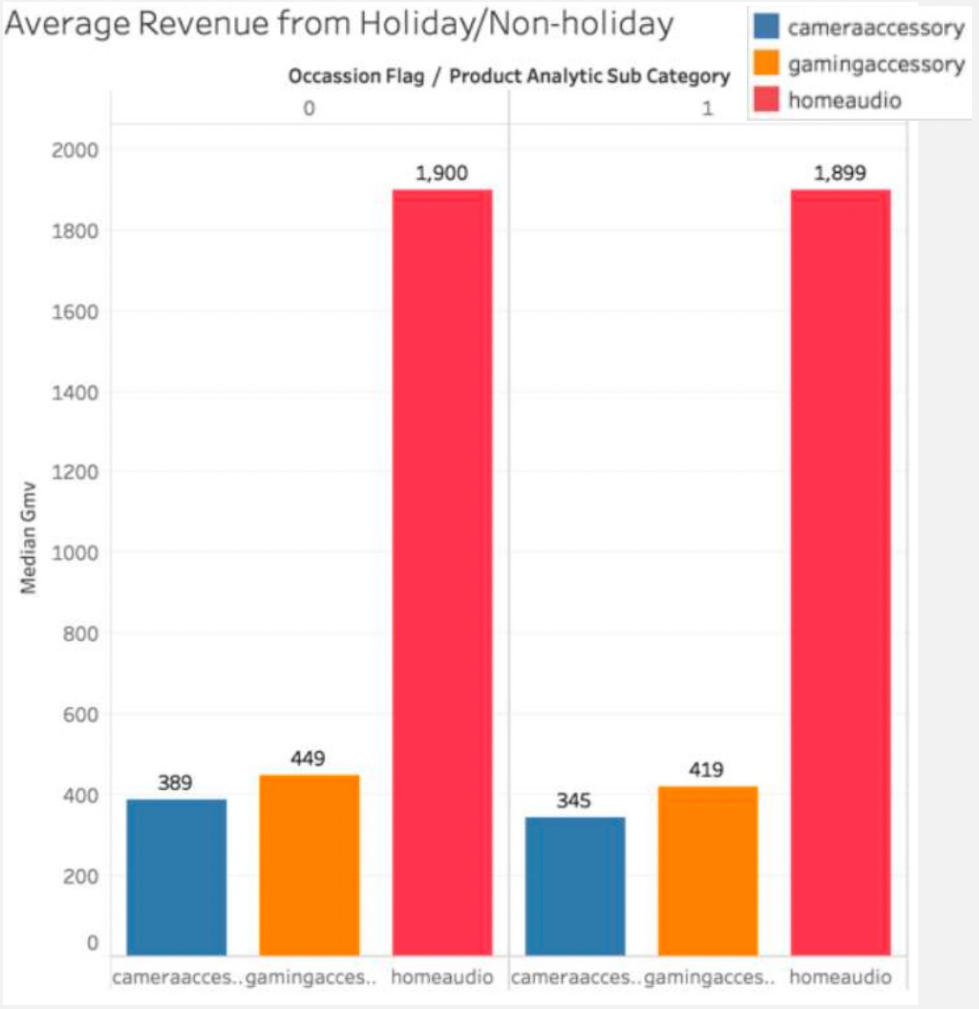
Home Audio Speaker under Home Audio segment had the most no of sales followed by Gaming Headset & Gamepad under Gaming Accessory.

VISUALIZATIONS CONTINUED

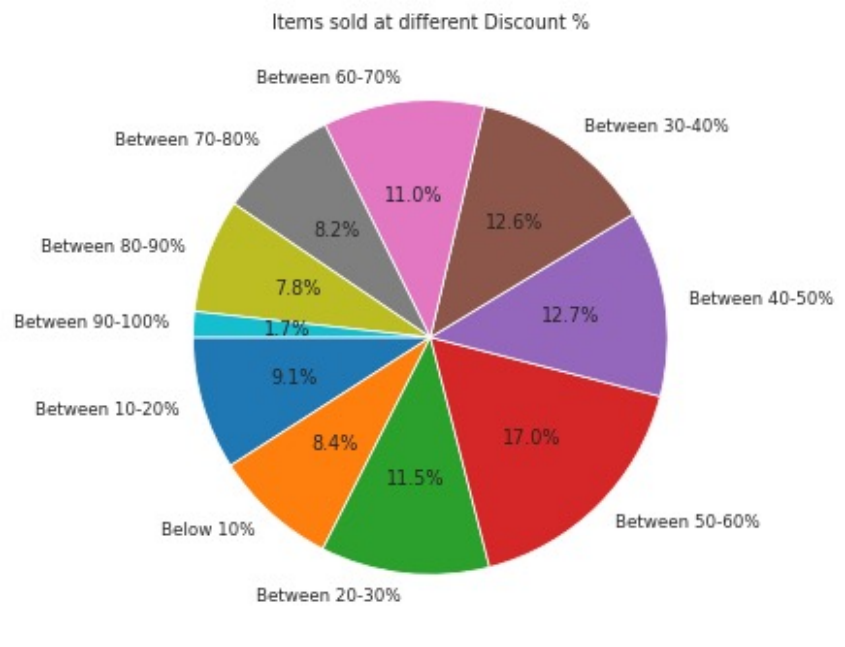


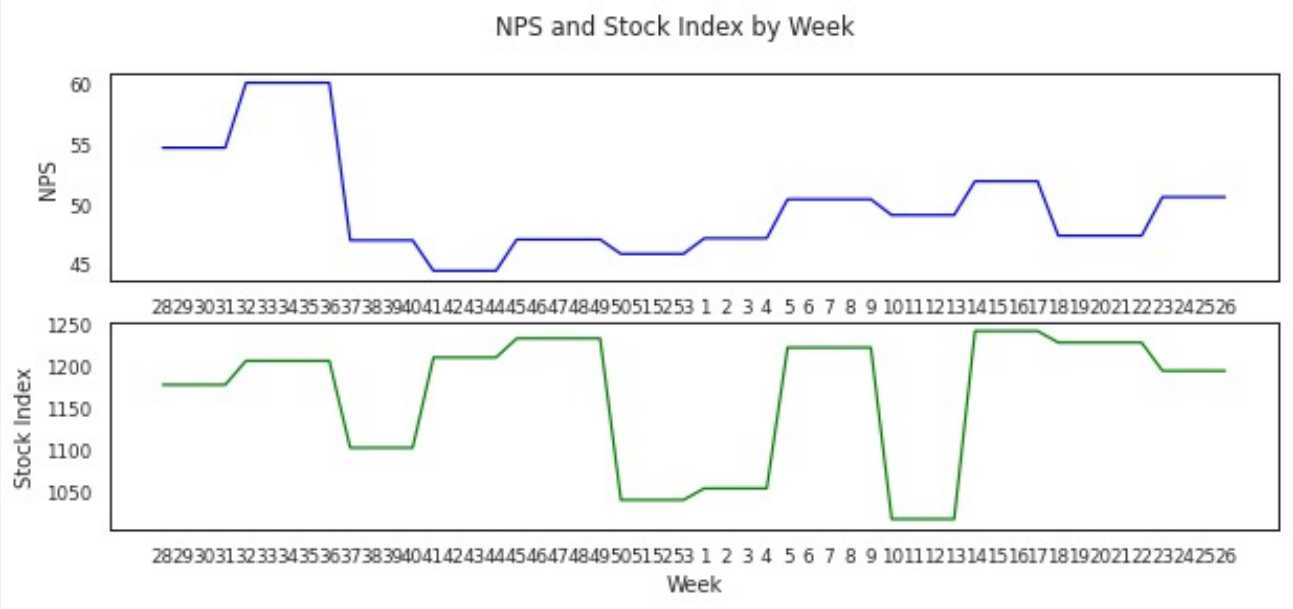
- Prepaid payments generate almost as much revenue as COD payments across the 3 sub categories.
- However, products using COD payment type have much larger number of units

Revenue generated from holidays is almost at par with that from non-holiday days across all sub-categories



VISUALIZATIONS CONTINUED



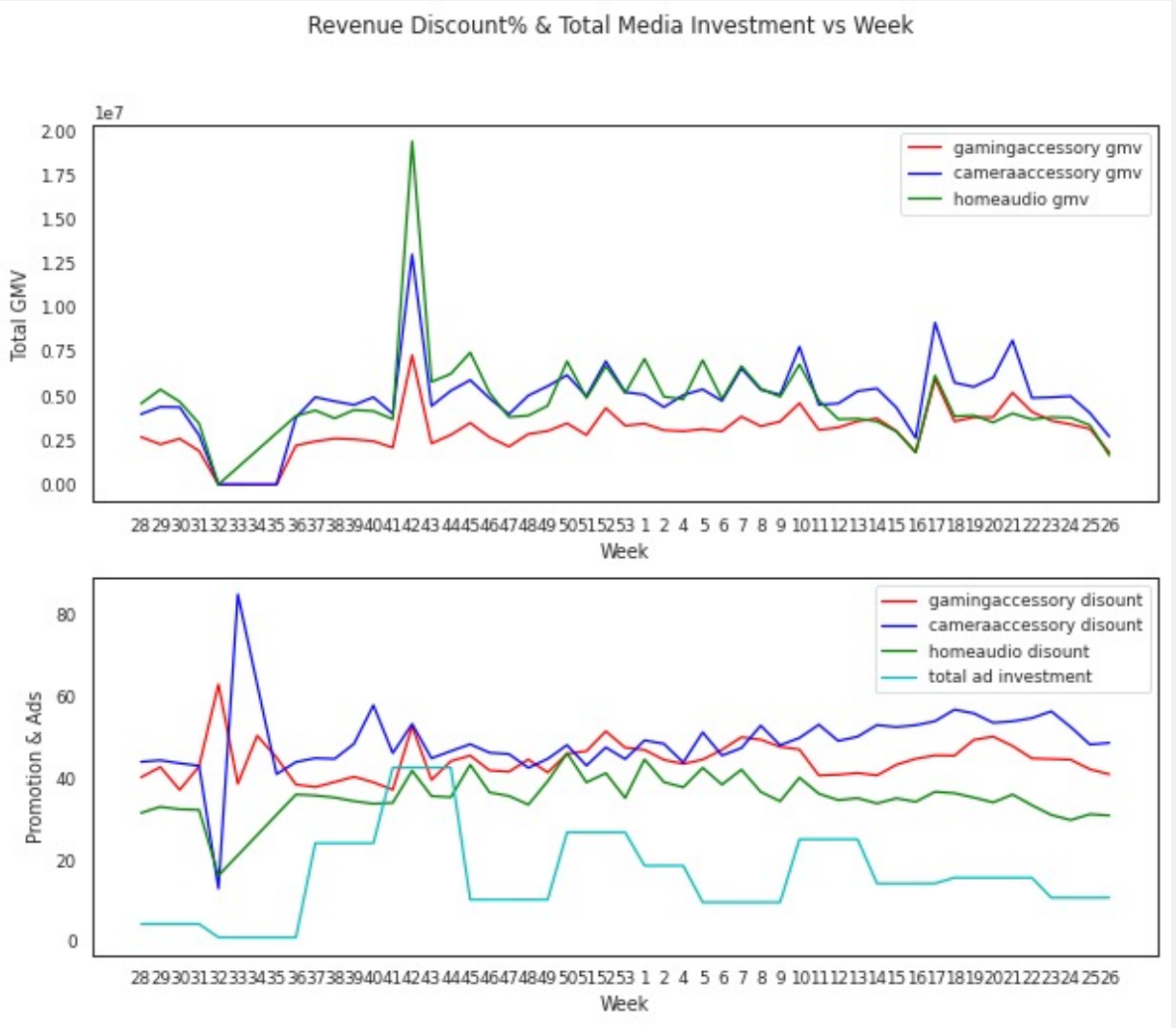


Most of the sales take place when Discount% is between 50-60%

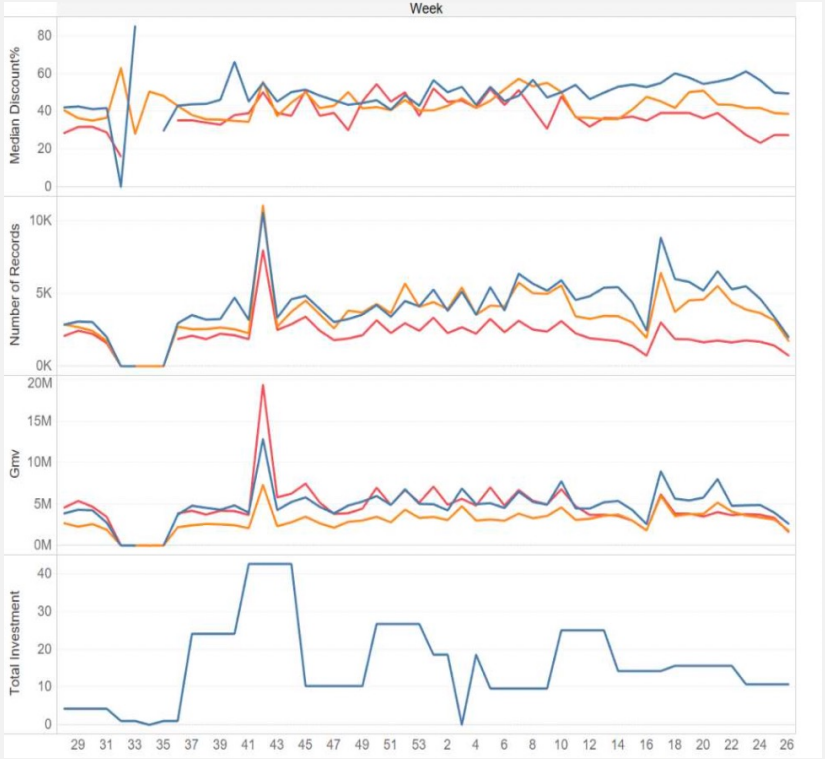
Consumer NPS score is highest in weeks 32 – 35 , which coincides with the time when maximum discounts were being offered.
Company Stock Index has seasonal ups and downs over the span of the target year.

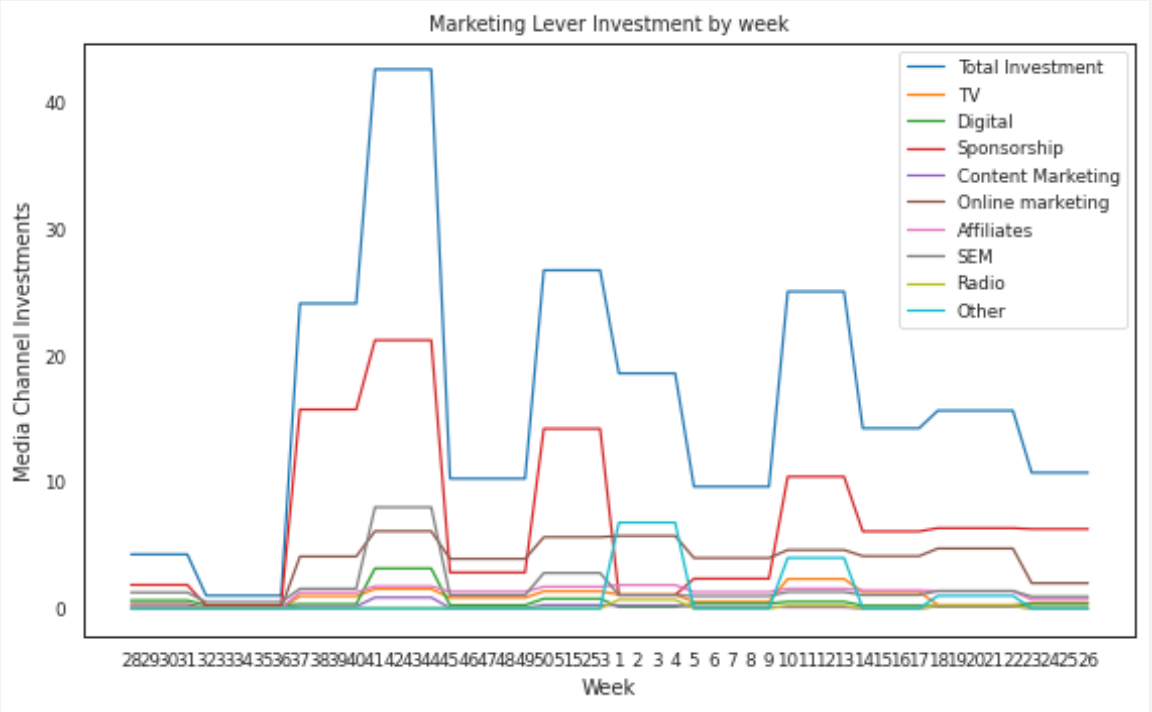
VISUALIZATIONS CONTINUED

INSIGHT:
- For week 42 (during `Thanksgiving`), all the graphs show a steep rise. Revenue increased because of both higher discount% and increased Ad Investment.
- For week 32(August), Revenue generated was the lowest from all 3 product subcategories. This is 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. However, barring home audio products, the revenue from other products was seen to be constant for the next 3 weeks after which, the revenue started to pick up.
- In general the average disount% offered for home audio products is lesser compared to that of the other product subcategories.



VISUALIZATIONS CONTINUED



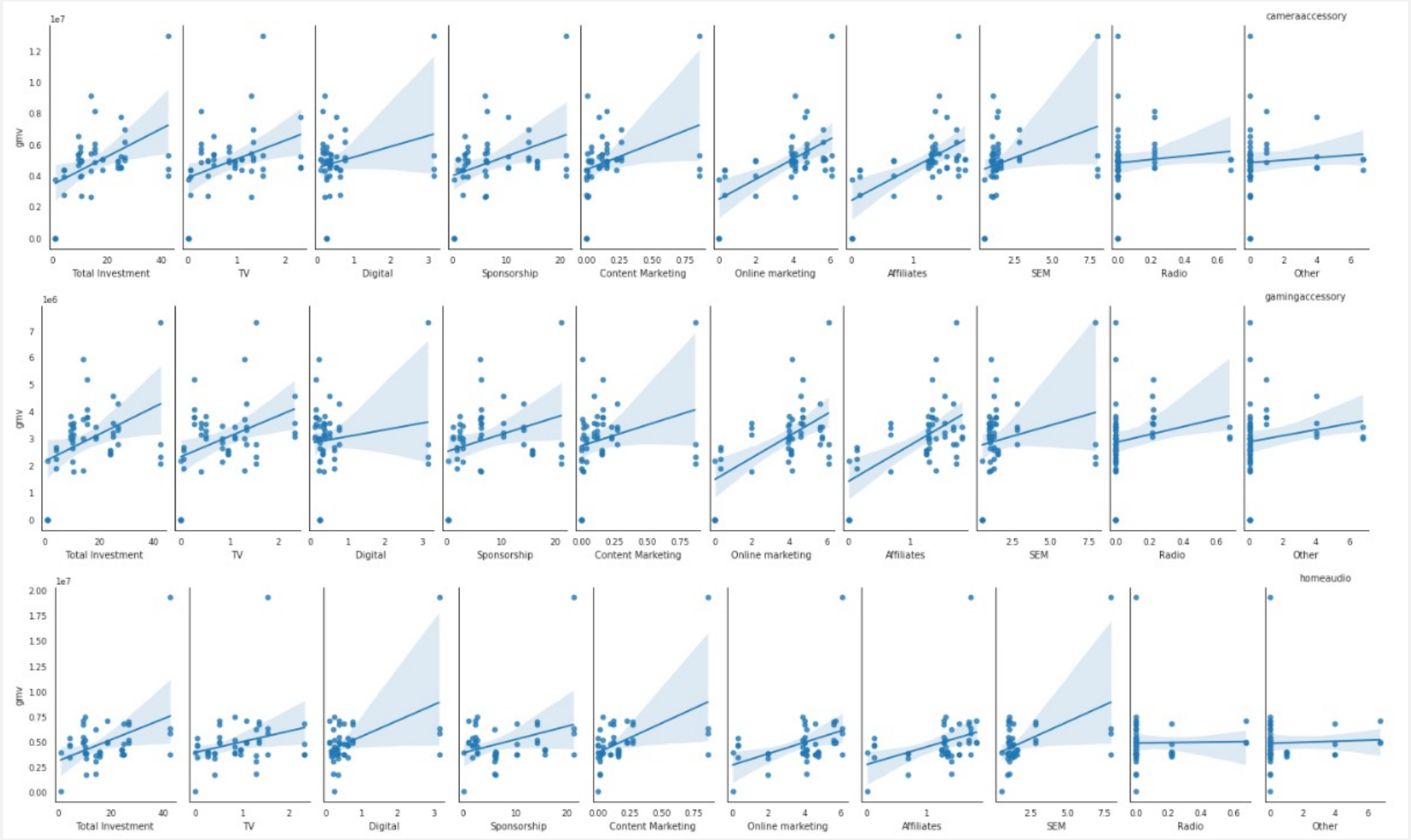


- For weeks 32 - 35(August), Revenue generated was the lowest from all 3 product sub categories. This is a direct relation to minimum amount of total investment in Ads. Discount was also lowest for all products apart from camera accessories. Post this dip in revenue, discount% was increased to bring about higher sales. This increase in Discount% was observed most in the case of gaming accessories.
- For the week# 42 (during Thanksgiving), all the graphs show a steep rise. Revenue increased because of higher discount% and increased Ad Investment.
- Over the past year, bulk of the Ad Investment has been made in Sponsorships followed by Online Marketing & Search Engine Marketing(specially during Thanksgiving).
- Barring home audio products, the revenue from other products was seen to be constant for the next 3 weeks after which, the revenue started to pick up.

VISUALIZATIONS CONTINUED

Relationship b/w revenue and Ad Spends:

- TV, Online Marketing & Affiliates seem to have a moderately Positive correlation with Revenue.
- Radio and Other seem to have a very low correlation with revenue



MODEL BUILDING

Additive Model

- 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 can be represented as:
- $Y = \alpha + \beta 1At + \beta 2Pt + \beta 3Dt + \beta 4Qt + \beta 5Tt + \epsilon$

Multiplicative Model

- Linear model is used to capture the current effect of several KPIs. This model assumes an additive relationship between Multiplicative model is used when there are interactions between the KPIs. To fit a multiplicative model, take logarithms of the data(on both sides of the model), 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.
- $Yt = \alpha + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \epsilon$
- $Yt = \alpha + \mu Yt\text{-}1 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \epsilon$

MODEL BUILDING CONTD

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:
- $Y_t = \alpha + \mu_1 Y_{t-1} + \mu_2 Y_{t-2} + \mu_3 Y_{t-3} + \ldots.$

$+ \beta_1 X1_t + \beta_1 X1_{t-1} + \beta_1 X1_{t-2} + \ldots.$
$+ \beta_2 X2_t + \beta_2 X2_{t-1} + \beta_2 X2_{t-2} + \ldots.$
$+ \beta_3 X3_t + \beta_3 X3_{t-1} + \beta_3 X3_{t-2} + \ldots.$
$+ \beta_4 X4_t + \beta_4 X4_{t-1} + \beta_4 X4_{t-2} + \ldots.$
$+ \beta_5 X5_t + \beta_5 X5_{t-1} + \beta_5 X5_{t-2} + \ldots.$
$+ \epsilon$

Distributive Lag Model (Multiplicative)

- Distributive Lag Model(Multiplicative) will help us capture the interactions between current and carry over effects of the KPIs.
- $Y_t = \alpha + \mu_1 \ln(Y_{t-1}) + \mu_2 \ln(Y_{t-2}) + \mu_3 \ln(Y_{t-3}) + \ldots.$

$+ \beta_1 \ln(X1_t) + \beta_1 \ln(X1_{t-1}) + \beta_1 \ln(X1_{t-2}) + \ldots.$
$+ \beta_2 \ln(X2_t) + \beta_2 \ln(X2_{t-1}) + \beta_2 \ln(X2_{t-2}) + \ldots.$
$+ \beta_3 \ln(X3_t) + \beta_3 \ln(X3_{t-1}) + \beta_3 \ln(X3_{t-2}) + \ldots.$
$+ \beta_4 \ln(X4_t) + \beta_4 \ln(X4_{t-1}) + \beta_4 \ln(X4_{t-2}) + \ldots.$
$+ \beta_5 \ln(X5_t) + \beta_5 \ln(X5_{t-1}) + \beta_5 \ln(X5_{t-2}) + \ldots.$
$+ \epsilon'$

MODEL SUMMARIES

The following table contains the details of all models built, their accuracy scores color encoded based on values

Model	R2	MSE	Model	R2	MSE
lr_camera_accessory	0.83	0.17	koy_CV_gaming_accessory	0.49	0.51
lr_CV_camera_accessory	0.32	0.68	koy_homeaudio_accessory	0.96	0.09
lr_gaming_accessory	0.93	0.05	koy_CV_homeaudio_accessory	0.74	0.26
lr_CV_gaming_accessory	0.53	0.47	dladd_camera_accessory	0.87	0.12
lr_homeaudio_accessory	0.96	0.09	dladd_CV_camera_accessory	0.82	0.18
lr_CV_homeaudio_accessory	0.73	0.27	dladd_gaming_accessory	0.87	0.10
mul_camera_accessory	0.84	0.36	dladd_CV_gaming_accessory	0.92	0.08
mul_CV_camera_accessory	0.92	0.08	dladd_homeaudio_accessory	0.42	1.39
mul_gaming_accessory	0.94	0.10	dladd_CV_homeaudio_accessory	0.55	0.45
mul_CV_gaming_accessory	0.94	0.06	dlmul_camera_accessory	0.78	0.50
mul_homeaudio_accessory	-0.77	0.37	dlmul_CV_camera_accessory	0.81	0.19
mul_CV_homeaudio_accessory	0.82	0.18	dlmul_gaming_accessory	0.93	0.10
koy_camera_accessory	0.84	0.16	dlmul_CV_camera_accessory	0.90	0.10
koy_CV_camera_accessory	0.27	0.73	dlmul_homeaudio_accessory	-0.21	0.25
koy_gaming_accessory	0.93	0.05	dlmul_CV_homeaudio_accessory	0.51	0.49

Notes:

1. Models in bold are the final chosen models
2. For R2 column, 3 color scale was used, green being best and red being worst
3. For MSE, 3 color scale was sued, red being worst, green being best

MODEL SELECTION

- The criteria of choosing the model is based on the accuracy parameters -- R2 score & MSE score -- and the business relevance of the important attributes chosen by the model. Models with CV were chosen as they dependable & generalizable, owing to CV.
- By referring to the model summary, the following models have been chosen for Camera Accessory, Gaming Accessory & Home Audio:

Product Sub-category	Linear Regression Model	R-square on Test Dataset	Mean Square Error	Top 5 KPIs
cameraaccessory	Multiplicative with CV	0.92	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	0.82	0.18	product_vertical_homeaudiospeaker (**0.469**)
				is_mass_market (**0.289**)
				product_vertical_fmradio (**0.224**)
				Radio_Ad_Stock (**0.147**)
				Sponsorship (**0.121**)

Insights:

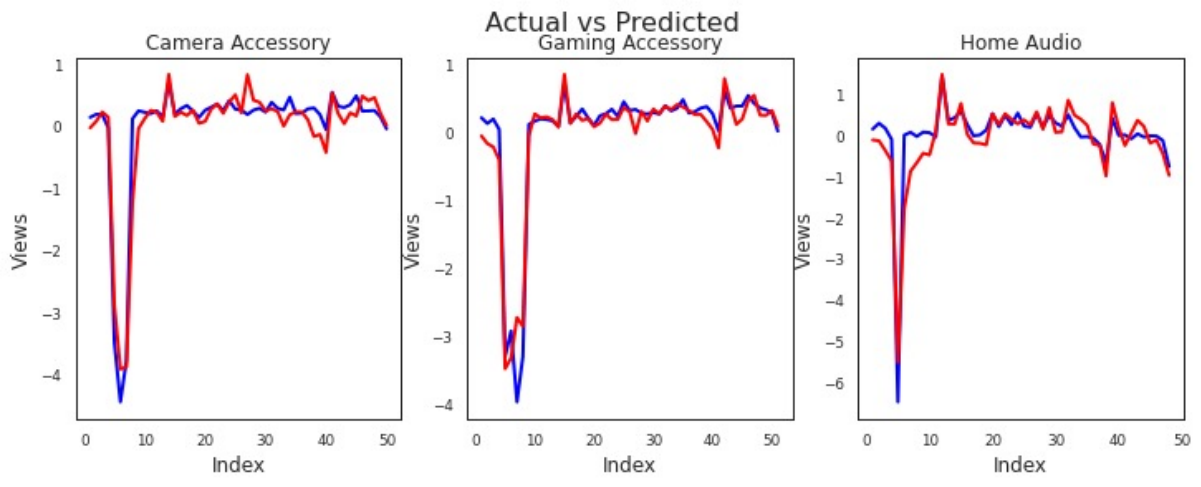
Notice that all the 3 chosen models for the 3 sub-categories are Multiplicative models.

This implies there exists some interaction between the KPIs for all the 3 models

These models show the growth of revenue vs the interactive growth of the KPIs.

MODEL EVALUATION

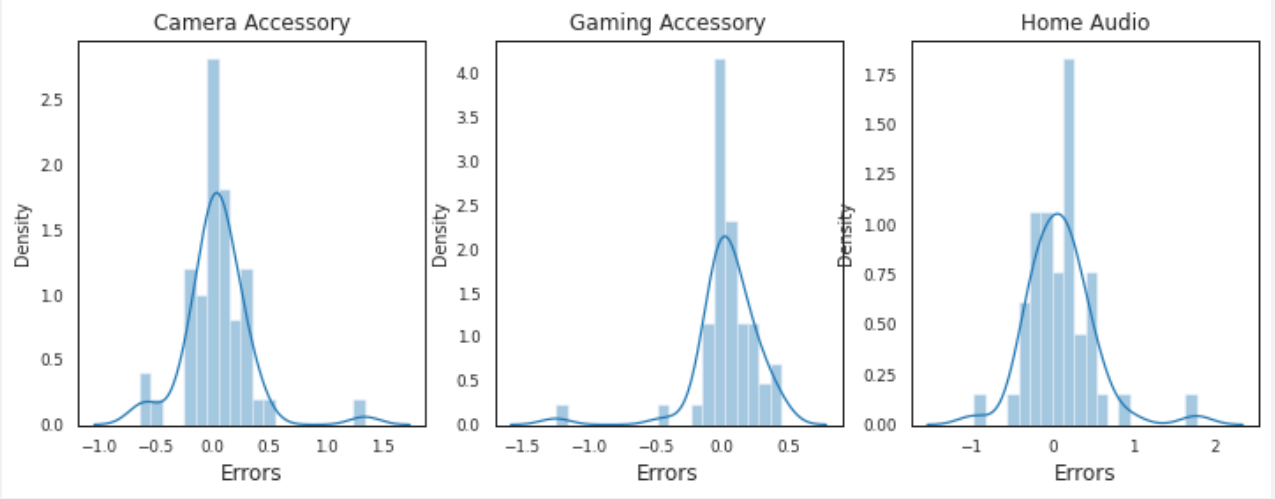
Plotting the actual and predicted values from the dataset for comparison



Scatter plots of Error Terms show that the error terms have homoscedasticity
The variance doesn't increase or decrease or follow a pattern as the error values change.

MODEL EVALUATION

The error terms follow a normal distribution with mean at 0 barring some outlier values.



Plotting a scatter plot with actual and predicted values from the dataset to check the spread and adding the best fit line