Capstone Project:
Market Mix Modeling
Final Submission Report

- *To model the impact of different levers on the sales figure of Eleckart*

Group Members : Rajarshi Palit | Ram Dittakavi | Sai Preetham | Sanchit Thareja

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 (July 2015 to June 2016) and **recommend the optimal budget allocation** for different marketing levers for the next year.*

The objective is thus classified into the following sub-goals:

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?

Problem Solving Methodology

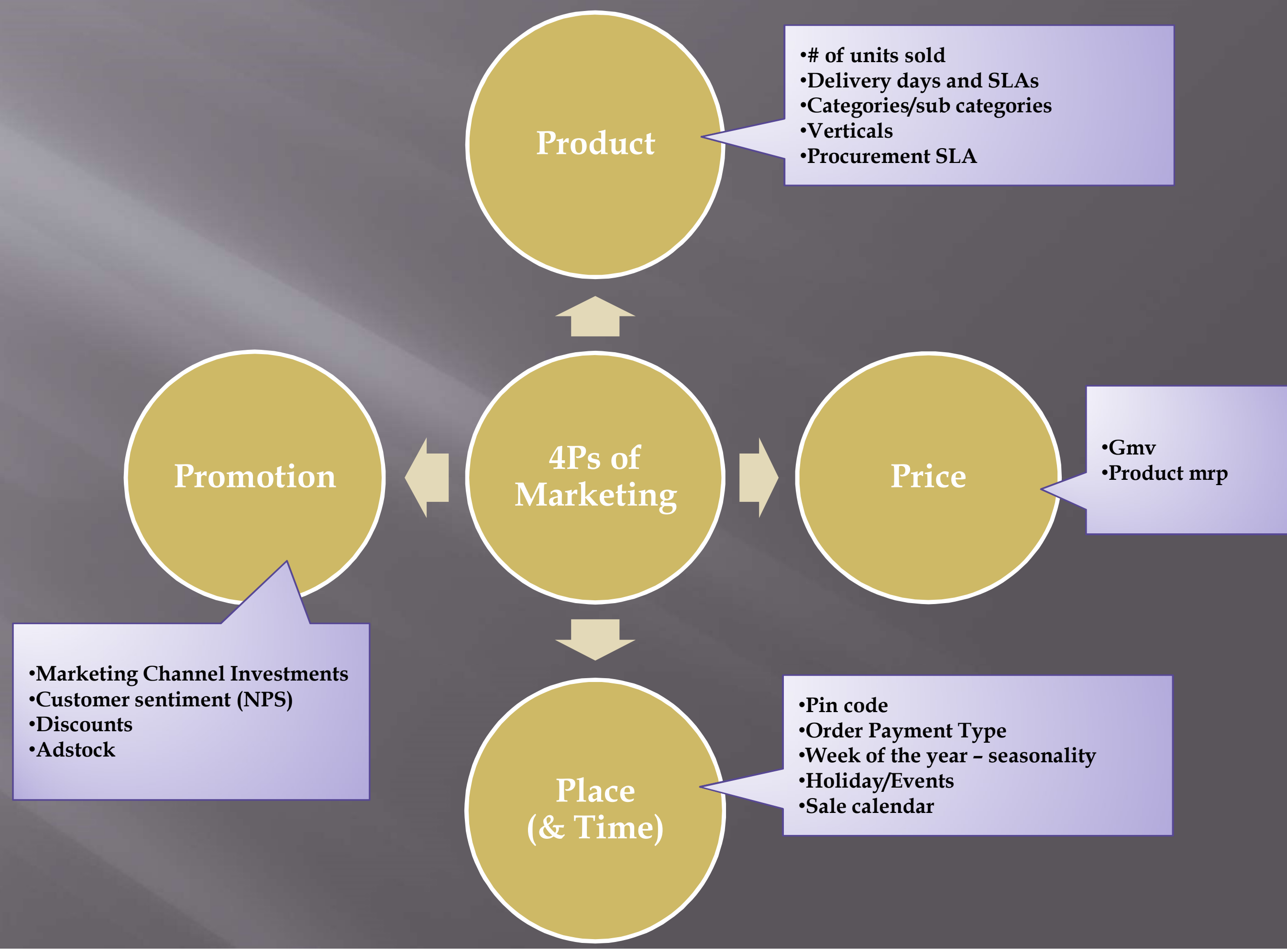
- The approach for this project has been designed to follow the **CRISP DM Framework**. The various stages of the framework are represented below in a sequential flow:

Understanding the Business Data → Data Preparation & Feature Engineering → Exploratory Data Analytics & Visualization → Preparing Regression Models for Prediction & Determining Important KPIs → Model Evaluation and Model Selection → Model Deployment (out-of-scope for this project)

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 & Cleanup

Handling Incorrect values in some columns

- **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 **99283 (6.33%) rows that are duplicates.** We went 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

Data Preparation & Cleanup contd…

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

Selecting One Year Data

- **Selecting1 Year Dat**a from July, 2015 – June, 2016. In the process, 592 records were dropped

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: Creation of new KPIs

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 - Luxury / Mass-market:
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

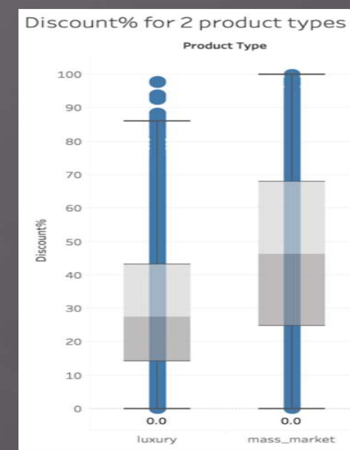
Lag Variables:
Lag variables(lag by 1, 2 & 3 days) for all KPIs were taken for Distributive Lag Models

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

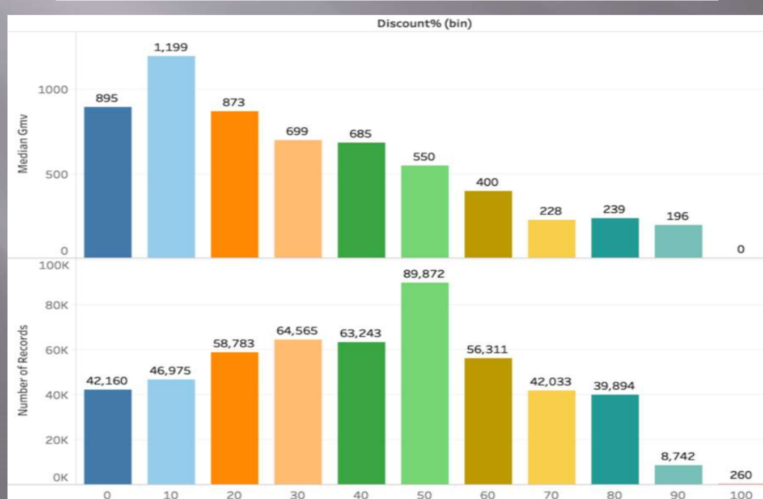
Visualization: An Insight into the Data

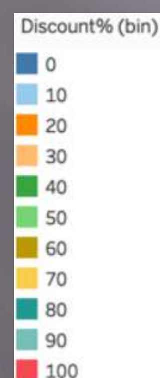
Discount% for Various Product Types

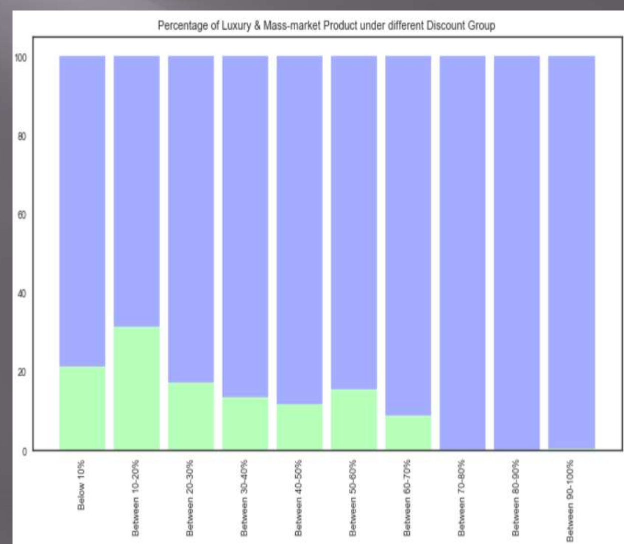
- **Median Revenue is maximum when Average discount% is between 10-20%.** But beyond that, average revenue slowly starts to decline.
- The sales on the other hand shows a steady increase with increase in Discount percentage till it **peaks at 50-60%** after which it starts to fall again.
- Maximum number 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 is a known trend among luxury products or luxury brands to offer limited or no discounts to **retain the exclusivity of their products.**



Analyzing how Sales Amount and Revenue vary based on Discount%

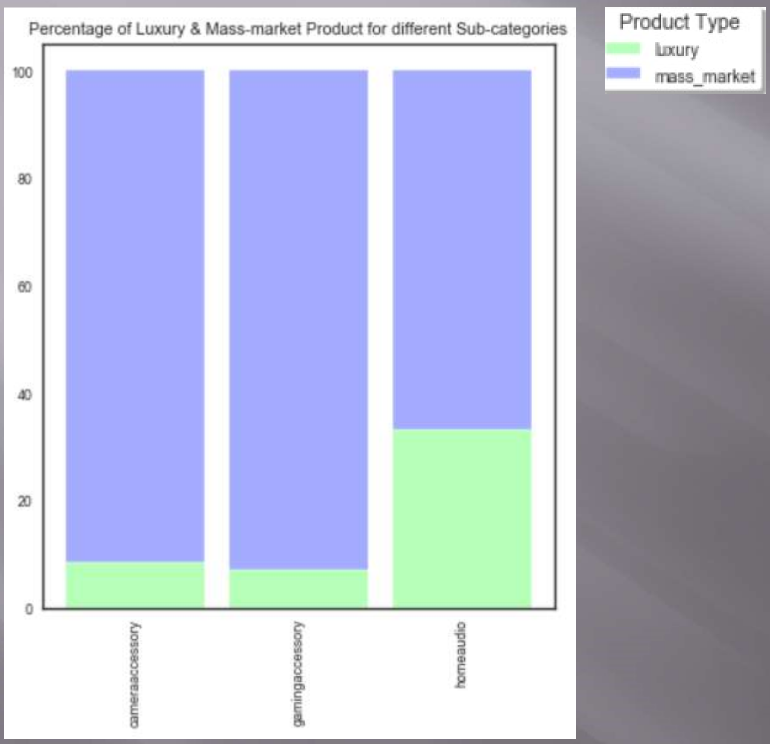




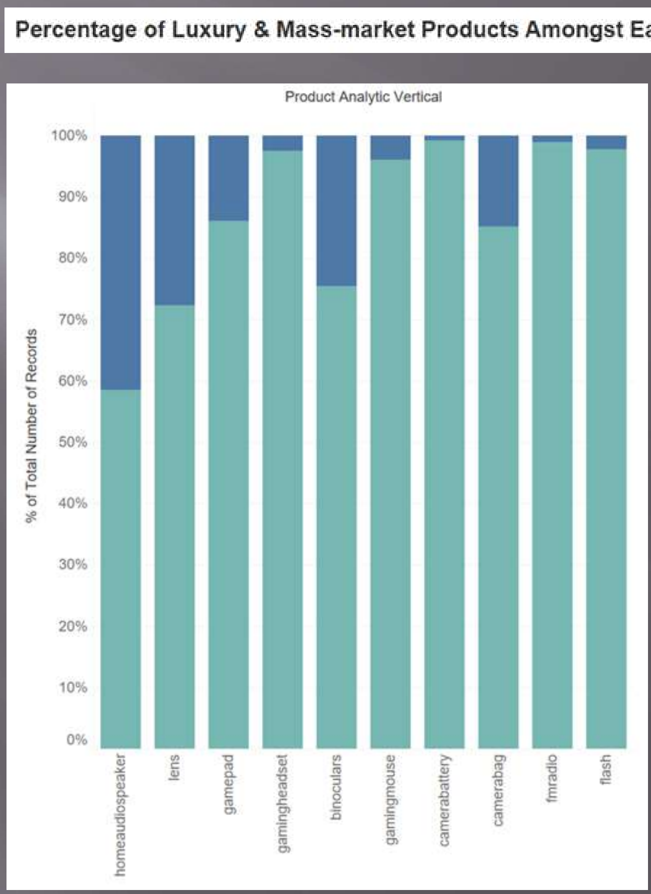


The median discount percentage offered for luxury items is less compared to that of Mass Market Products. This is a known trend among luxury products or luxury brands,
to offer limited discounts,
to retain the exclusivity of their products.

Visualization (contd)



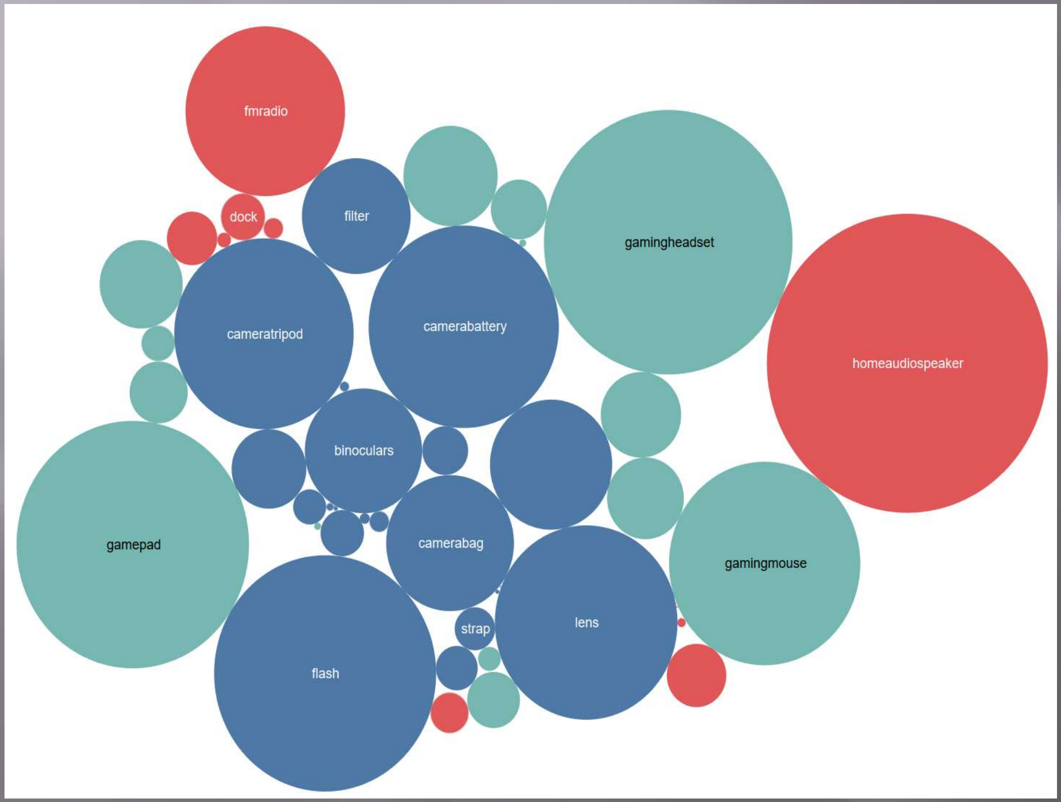
Percentage of luxury products under Home Audio is much more compared to the other sub categories.



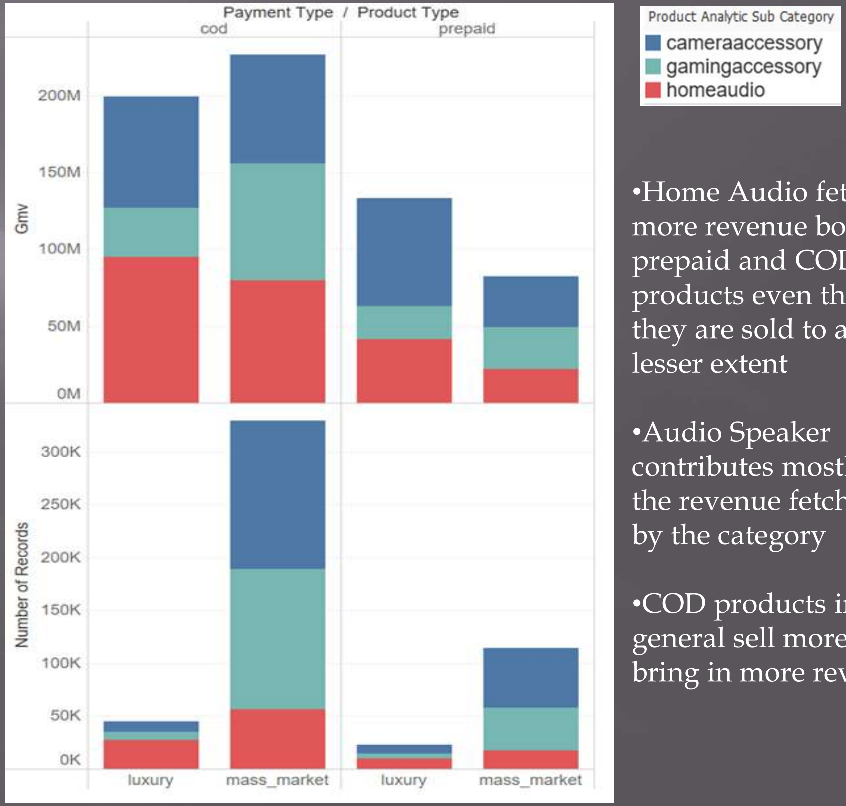
Most of the luxury products belong to the product verticals – Home Audio Speaker, Camera Lens & Binoculars.

Visualization (contd)

Product Verticals with Most Number of Sales



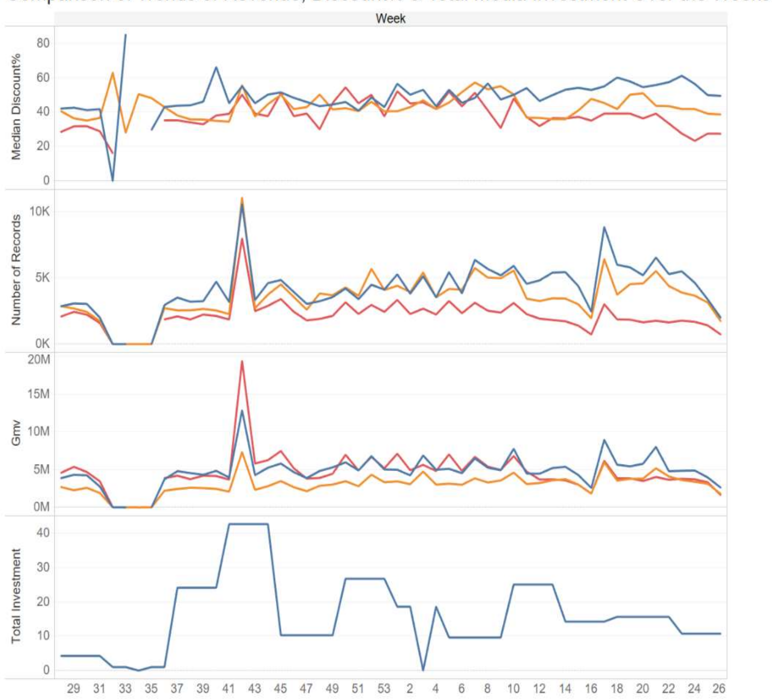
Analyzing how Sales Amount and Revenue vary based on Payment Types & Product Types



- 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

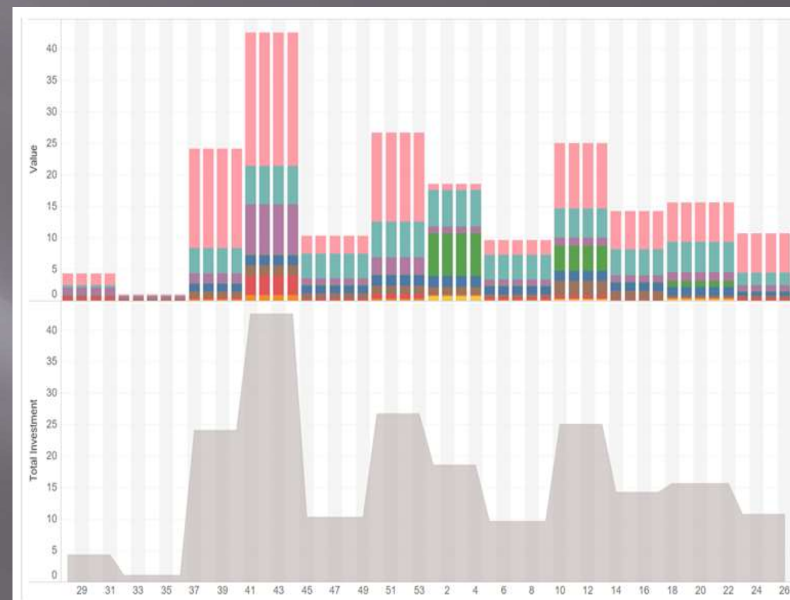
Visualization (contd)

Comparison of Trends of Revenue, Discount% & Total Media Investment Over the Weeks



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

Trends in Advertisement Investments in Various Media Channels Over the Weeks



- 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.

- 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.

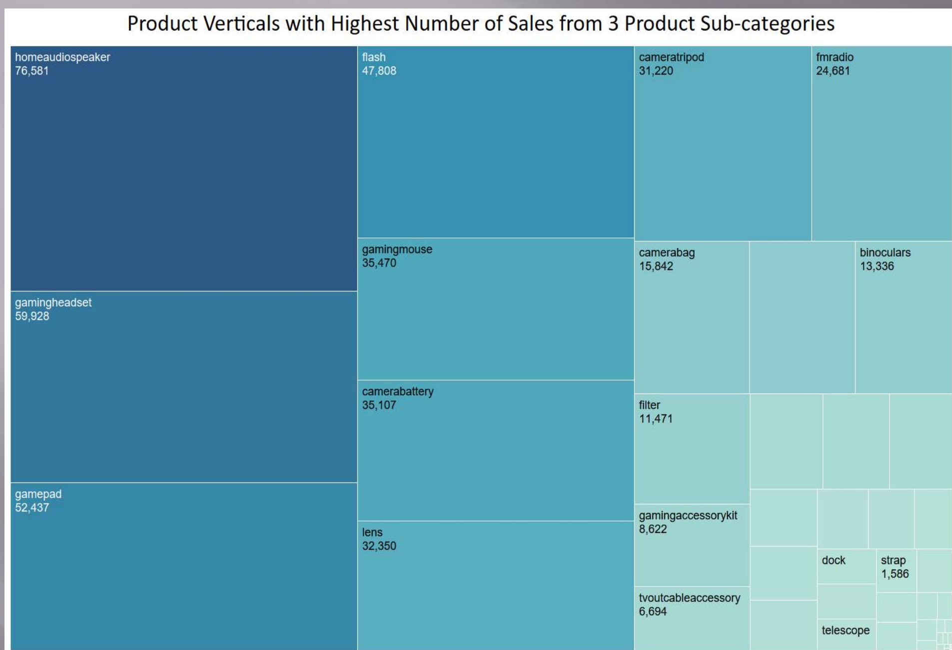
-In general the average discount% offered for home audio products is lesser compared to that of the other product subcategories.

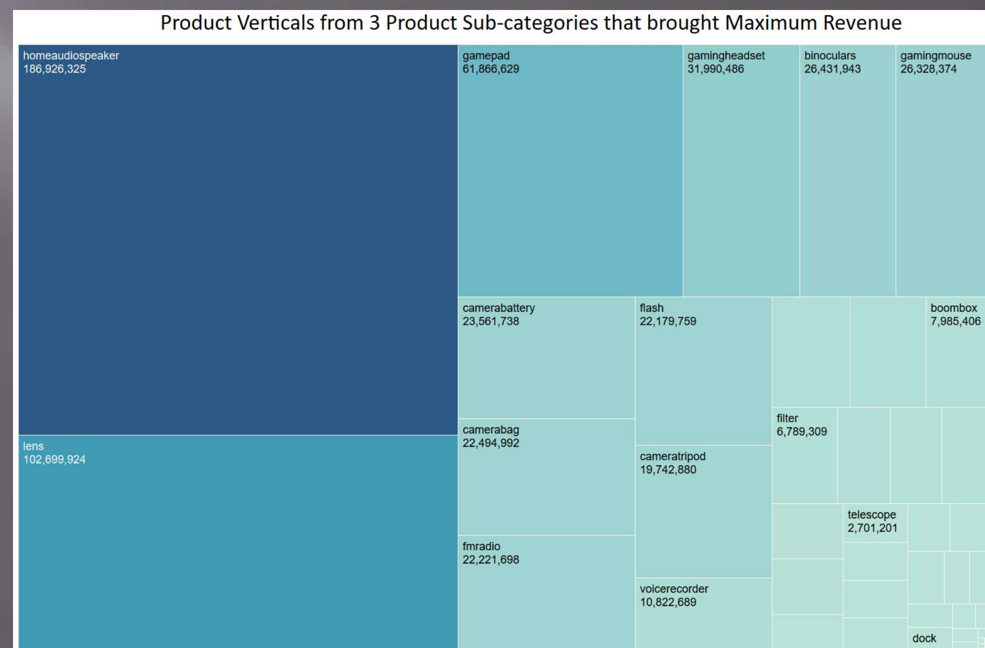
-Over the past year, bulk of the Ad Investment has been made in Sponsorships followed by Online Marketing & Search Engine Marketing(specially during Thanksgiving).

Visualization (contd)

More visualizations are included in the Appendix section.

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.

A Brief Description of the Models Built

The primary objective of the case study being Revenue prediction and determination of important KPIs that influence the revenue growth, we have build the following Linear Regression 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 can be represented as:
  - **$Y = \alpha + \beta 1At + \beta 2Pt + \beta 3Dt + \beta 4Qt + \beta 5Tt + \epsilon$**

Multiplicative

- Multiplicative model is used when there are interactions between the KPIs. To fit a multiplicative model, take logarithms of the data(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.
  - $Yt = \alpha + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \epsilon$
  - **$Yt = \alpha + \mu Yt\text{-}1 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \epsilon$**

A Brief Description of the Models Built (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:

$$\begin{aligned}Yt = \alpha &+ \mu1Yt\text{-}1 + \mu2Yt\text{-}2 + \mu3Yt\text{-}3 + \ldots.\\ &+ \beta1X1t + \beta1X1t\text{-}1 + \beta1X1t\text{-}2 + \ldots.\\ &+ \beta2X2t + \beta2X2t\text{-}1 + \beta2X2t\text{-}2 + \ldots.\\ &+ \beta3X3t + \beta3X3t\text{-}1 + \beta3X3t\text{-}2 + \ldots.\\ &+ \beta4X4t + \beta4X4t\text{-}1 + \beta4X4t\text{-}2 + \ldots.\\ &+ \beta5X5t + \beta5X5t\text{-}1 + \beta5X5t\text{-}2 + \ldots.\\ &+ \epsilon\end{aligned}$$

Distributive Lag Model (Multiplicative)

- *Distributive Lag Model(Multiplicative) will help us capture the interactions between current and carry over effects of the KPIs.*

$$\begin{aligned}Yt = \alpha &+ \mu1\ln(Yt\text{-}1) + \mu2\ln(Yt\text{-}2) + \mu3\ln(Yt\text{-}3) + \ldots.\\ &+ \beta1\ln(X1t) + \beta1\ln(X1t\text{-}1) + \beta1\ln(X1t\text{-}2) + \ldots.\\ &+ \beta2\ln(X2t) + \beta2\ln(X2t\text{-}1) + \beta2\ln(X2t\text{-}2) + \ldots.\\ &+ \beta3\ln(X3t) + \beta3\ln(X3t\text{-}1) + \beta3\ln(X3t\text{-}2) + \ldots.\\ &+ \beta4\ln(X4t) + \beta4\ln(X4t\text{-}1) + \beta4\ln(X4t\text{-}2) + \ldots.\\ &+ \beta5\ln(X5t) + \beta5\ln(X5t\text{-}1) + \beta5\ln(X5t\text{-}2) + \ldots.\\ &+ \epsilon'\end{aligned}$$

Model Dashboard

The following table contains the details of all models built, their accuracy scores and the top 5 KPIs returned by them:

Product Sub-category	Linear Regression Model	Cross Validation	R2 Score	MSE Score	Top 5 KPIs
cameraaccessory	Additive	No	0.83	0.17	product_vertical_lens, product_vertical_camerabattery, product_vertical_camerabag, product_vertical_camerahousing, Online marketing
		Yes	-0.8	1.08	
	Multiplicative	No	0.84	0.36	**product_vertical_lens, product_vertical_camerabattery, is_mass_market, product_vertical_camerabatterycharger, TV**
		Yes	**0.91**	**0.09**	
	Koyck	No	0.84	0.16	product_vertical_lens, product_vertical_camerabag, product_vertical_camerahousing, product_vertical_camerabattery, Online marketing
		Yes	0.27	0.73	
	Distributive Lag Model (Additive)	No	0.87	0.12	product_vertical_lens, product_vertical_filter, product_vertical_camerabag, product_vertical_cameraremotecontrol, is_mass_market
		Yes	0.82	0.17	
	Distributive Lag Model (Multiplicaitive)	No	0.77	0.5	is_mass_market, product_vertical_lens, product_vertical_cameraaccessory, product_vertical_camerabattery, product_vertical_cameratripod
		Yes	0.82	0.18	
gamingaccessory	Additive	No	0.93	0.05	product_vertical_gamepad, product_vertical_gamingheadset, is_mass_market, product_vertical_gamingaccessorykit, product_vertical_gamingmouse
		Yes	0.51	0.49	
	Multiplicative	No	0.94	0.09	**product_vertical_gamingheadset, is_mass_market, product_vertical_gamingmouse, product_vertical_gamepad, Online marketing_SMA_3**
		Yes	**0.94**	**0.06**	
	Koyck	No	0.93	0.05	product_vertical_gamepad, product_vertical_gamingheadset, is_mass_market, product_vertical_gamingaccessorykit, product_vertical_gamingmouse
		Yes	0.49	0.51	
	Distributive Lag Model (Additive)	No	0.87	0.1	product_vertical_gamepad, product_vertical_gamingaccessorykit, is_mass_market, product_vertical_motioncontroller, product_vertical_gamingkeyboard
		Yes	0.92	0.08	
	Distributive Lag Model (Multiplicaitive)	No	0.93	0.11	product_vertical_gamepad, product_vertical_gamingmouse, is_mass_market, product_vertical_gamingkeyboard, is_cod
		Yes	0.89	0.11	
homeaudio	Additive	No	0.96	0.09	product_vertical_homeaudiospeaker, is_mass_market, Digital_SMA_3, product_vertical_fmradio, is_cod
		Yes	0.73	0.27	
	Multiplicative	No	-0.63	0.34	**product_vertical_homeaudiospeaker, is_mass_market, product_vertical_fmradio, Radio_Ad_Stock, Sponsorship**
		Yes	**0.86**	**0.14**	
	Koyck	No	0.96	0.09	product_vertical_homeaudiospeaker, is_mass_market, is_cod, NPS, Mean Temp
		Yes	0.7	0.3	
	Distributive Lag Model (Additive)	No	0.42	1.39	product_vertical_homeaudiospeaker, product_vertical_karaokeplayer, is_mass_market, is_cod, product_vertical_fmradio
		Yes	0.53	0.47	
	Distributive Lag Model (Multiplicaitive)	No	-0.23	0.26	product_vertical_homeaudiospeaker, is_mass_market, product_vertical_fmradio, is_cod, product_vertical_voicerecorder
		Yes	0.57	0.43	

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.

Also we tried to choose models with cross validation because even though the ones without, sometimes give us good scores, they are not very dependable & generalizable, owing to limited dataset.

By referring to the model dashboard, we finalize the following models for the 3 mentioned product subcategories - Camera Accessory, Gaming Accessory & Home Audio:

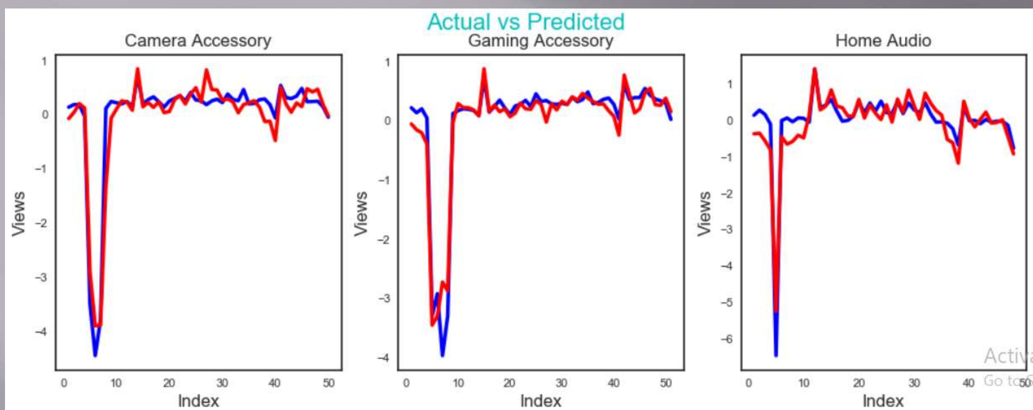
Product Sub-category	Linear Regression Model	R-square on Test Dataset	Mean Square Error	Top 5 KPIs
cameraaccessory	Multiplicative with CV	0.91	0.09	product_vertical_lens (0.181)
				product_vertical_camerabattery (0.160)
				is_mass_market (0.149)
				product_vertical_camerabatterycharger (0.121)
				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)

•We notice that all the 3 chosen models for the 3 sub-categories are **Multiplicative models.**

•This fact tells us that there exists some **interaction between the KPIs** for all the 3 model.

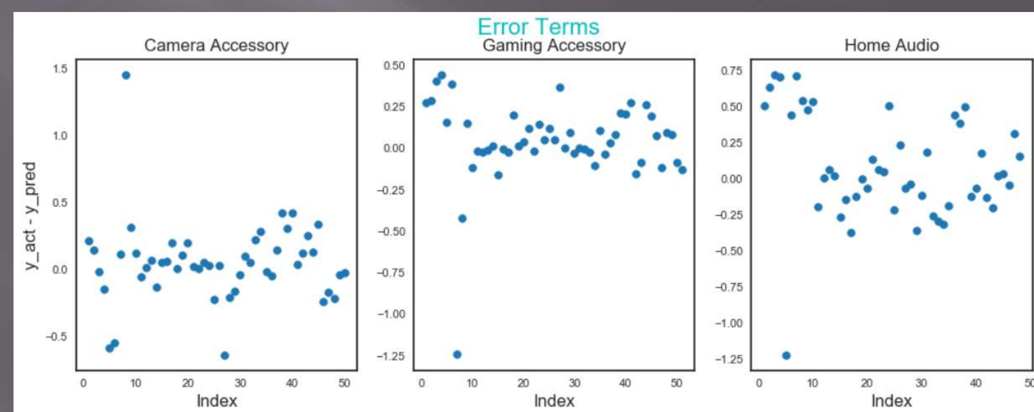
•These models tell us about the **growth of revenue vs the interactive growth of the KPIs.**

Model Validation

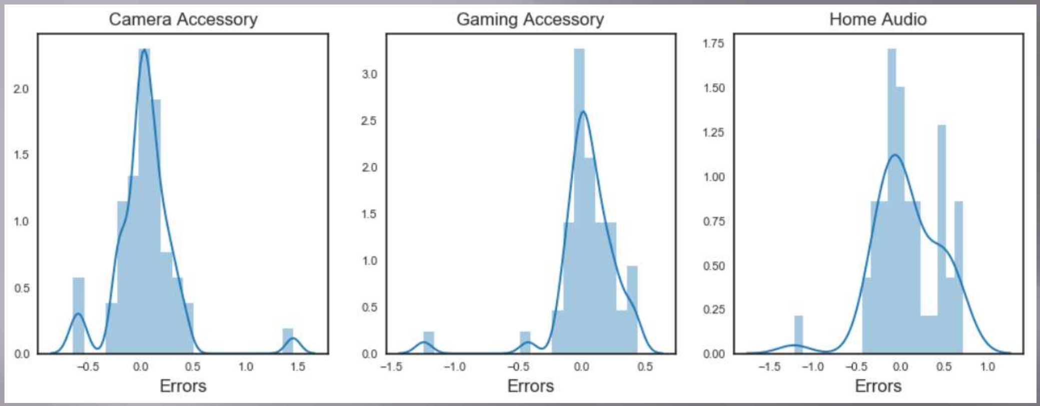


Plotting the actual and predicted price values from the dataset to check the likeness.

Drawing a scatter plot of the Error Terms to check the spread to ensure that the error terms have constant variance (homoscedasticity).
The variance doesn't increase or decrease or follow a pattern as the error values change.

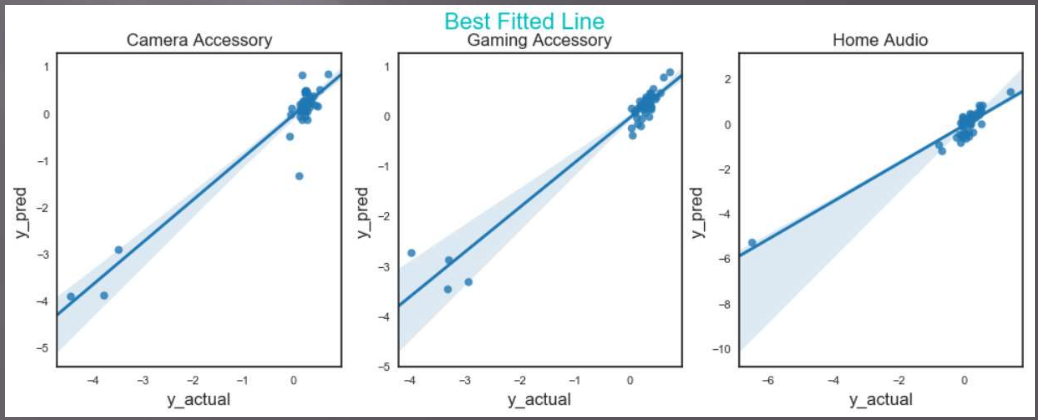


Model Validation (contd)



Plotting the distribution of the error terms.
The error terms follow a normal distribution with mean at 0 barring a few outlier values.

Plotting a scatter plot with actual and predicted price values from the dataset to check the spread and drawing the best fitted line through it.



Equation for the Best Fitted Line

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

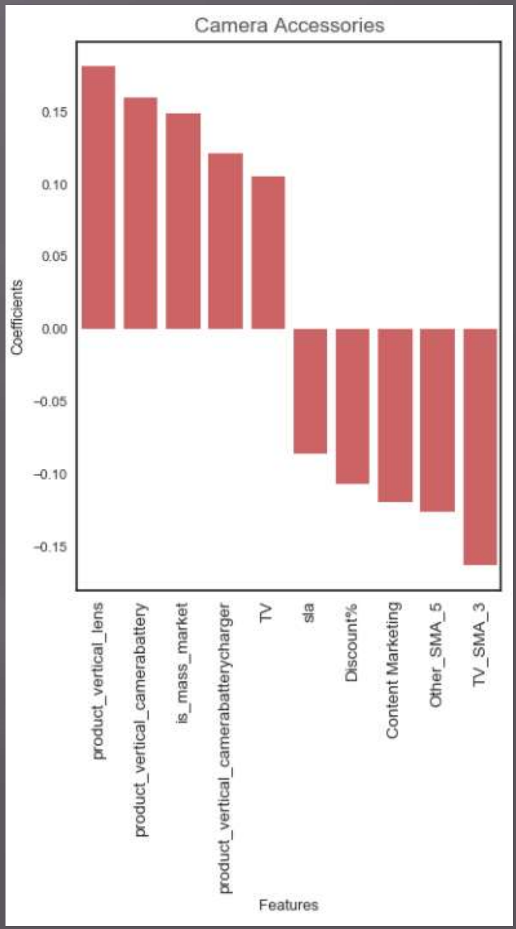
- **Revenue** = 0.0 + (0.469 × **product_vertical_homeaudiospeaker**) + (0.289 × **is_mass_market**) + (0.224 × **product_vertical_fmradio**) + (0.147 × **Radio_Ad_Stock**) + (0.121 × **Sponsorship**) + …

This equation implies how the revenue can grow with a unit growth in any of these independent KPIs with all other KPIs held constant.

Recommendation

Camera Accessory

- 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.

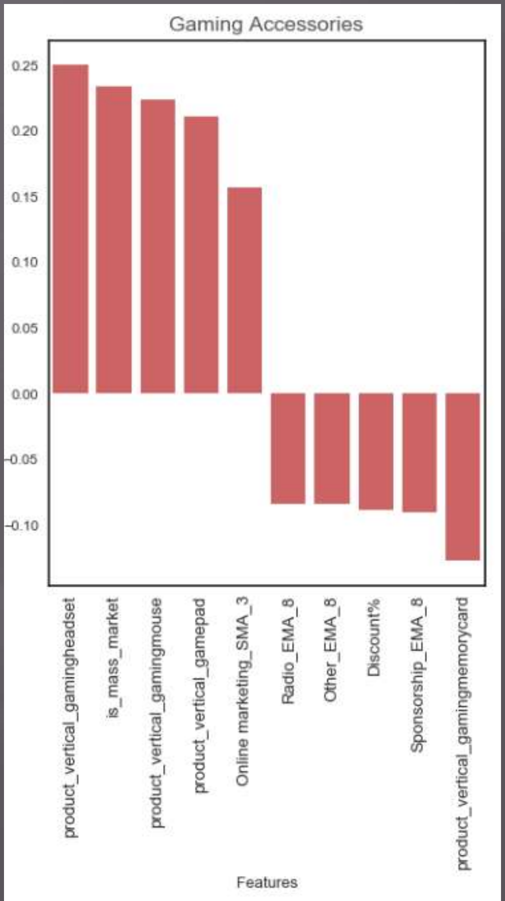


This figures on the right describes the Elasticity of different KPIs w.r.t. the Product Revenue.

Recommendation (contd)

Gaming Accessory

- 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.

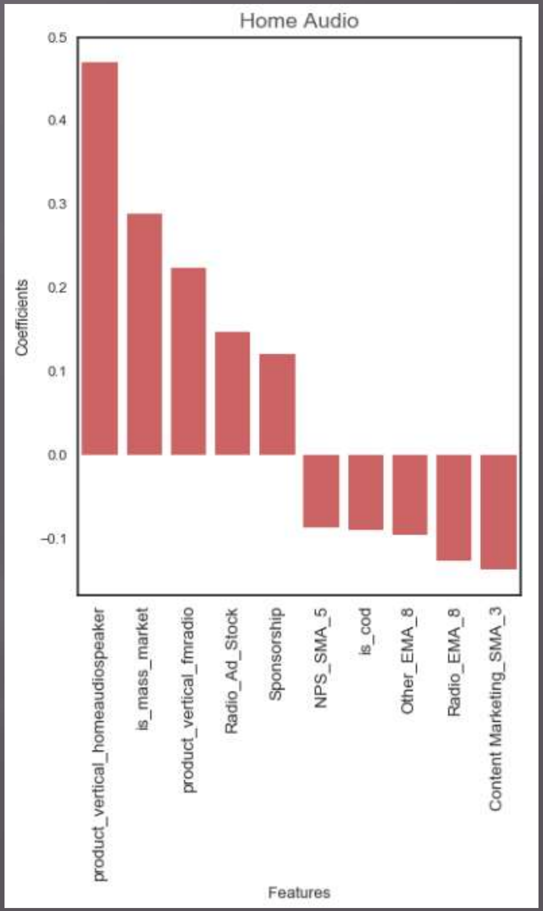


This figures on the right describes the Elasticity of different KPIs w.r.t. the Product Revenue.

Recommendation (contd)

Home Audio

- 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.



This figures on the right describes the Elasticity of different KPIs w.r.t. the Product Revenue.

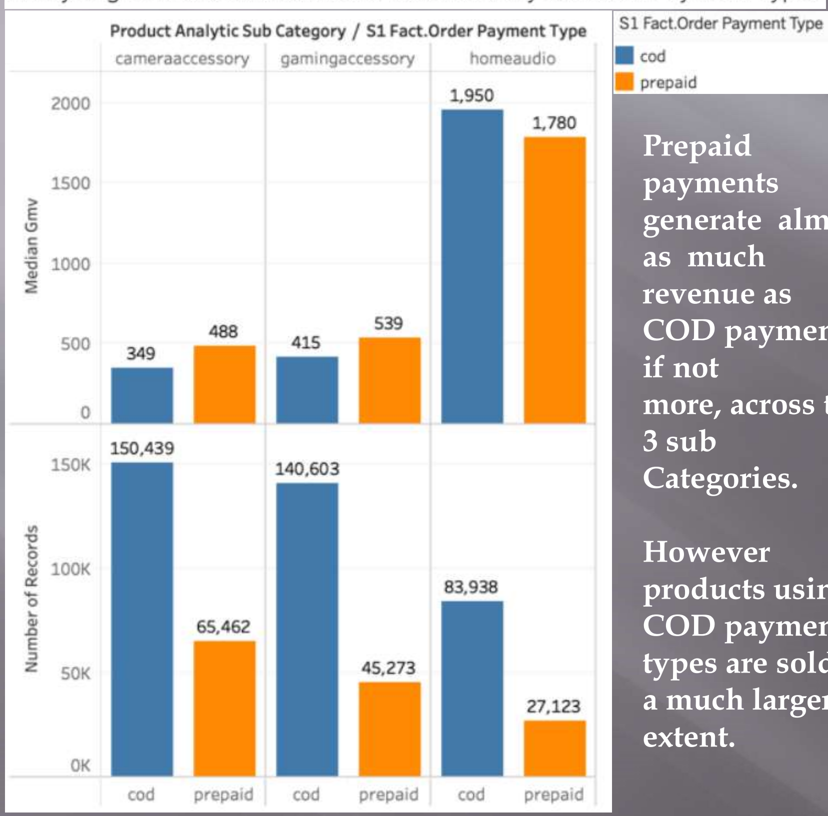
Recommendation (contd)

In General

- 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 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.

Appendix: Some More Visualizations

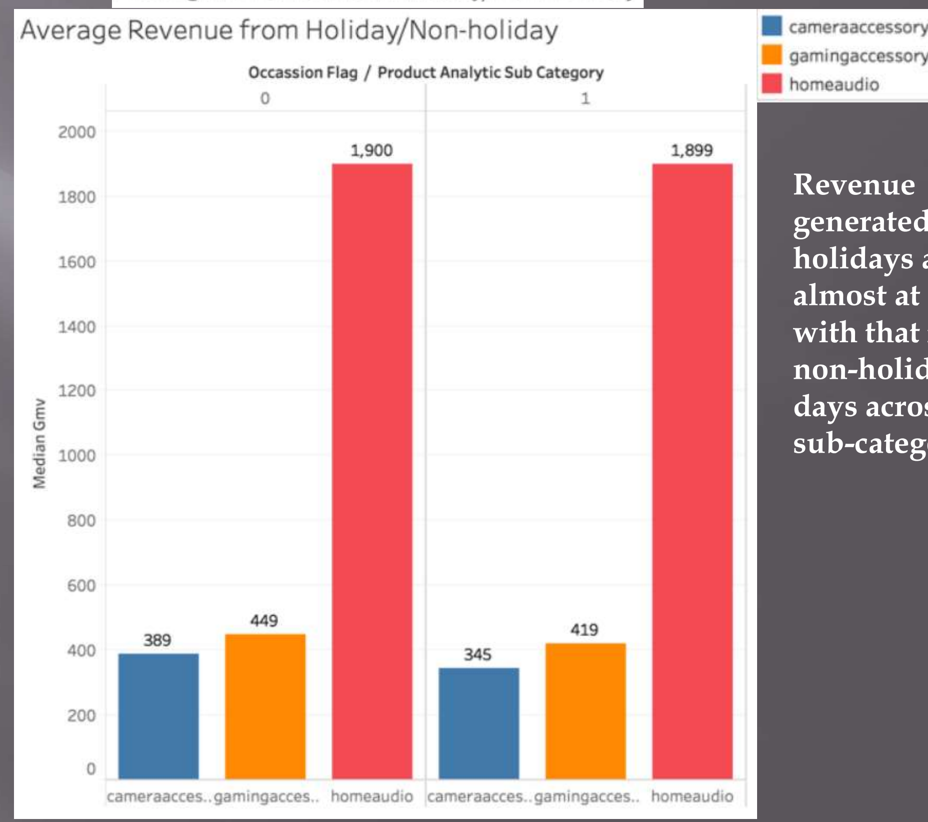
Analyzing how Sales Amount and Revenue vary based on Payment Types



Prepaid payments generate almost as much revenue as COD payments if not more, across the 3 sub Categories.

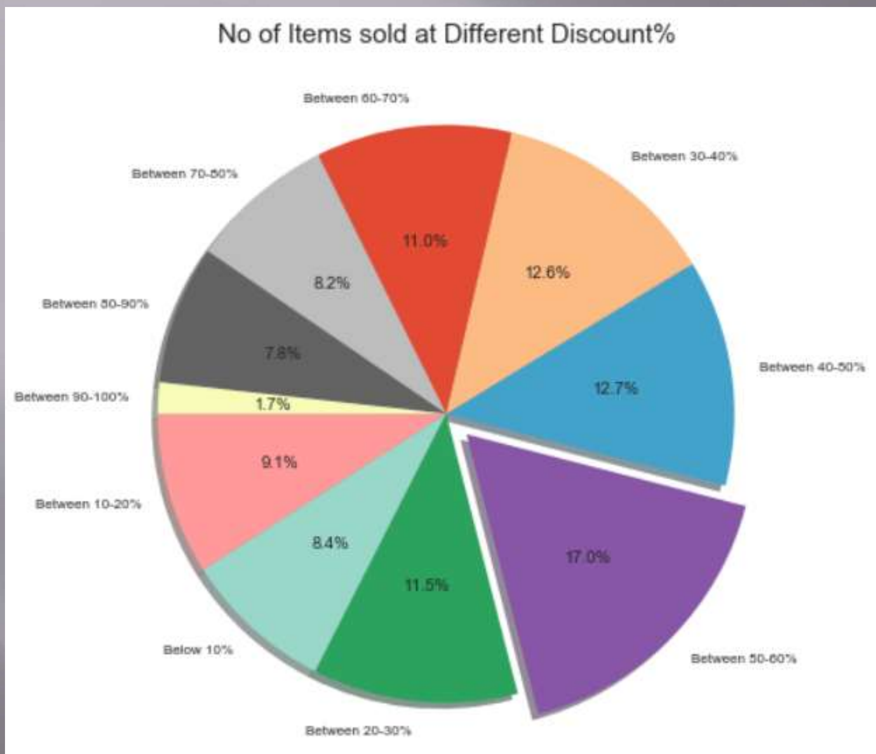
However products using COD payment types are sold to a much larger extent.

Average Revenue from Holiday/Non-holiday



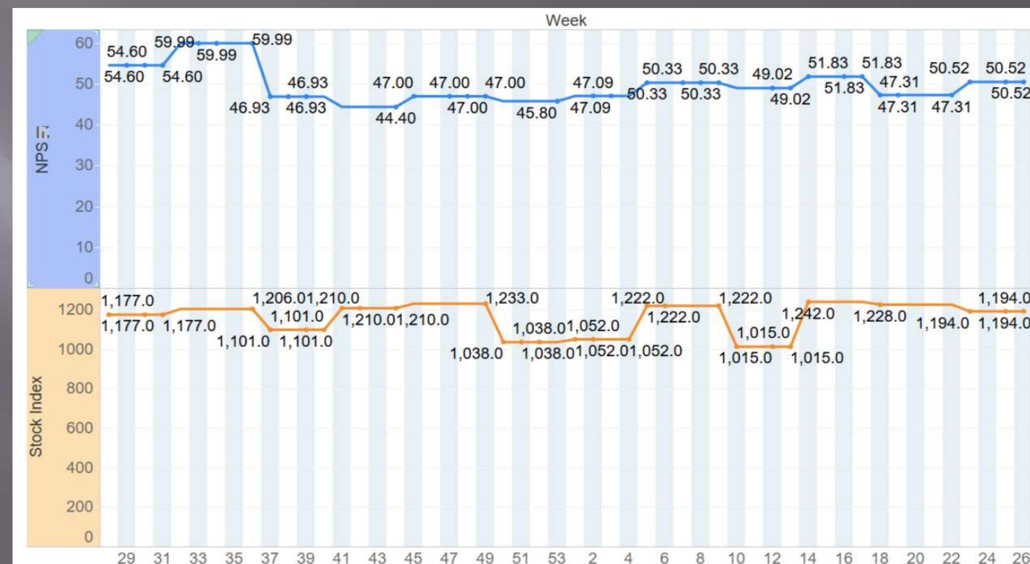
Revenue generated from holidays are almost at par with that from non-holiday days across 3 sub-categories.

Appendix: More Visualizations



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

Trends in NPS & Stock Index Over the Weeks

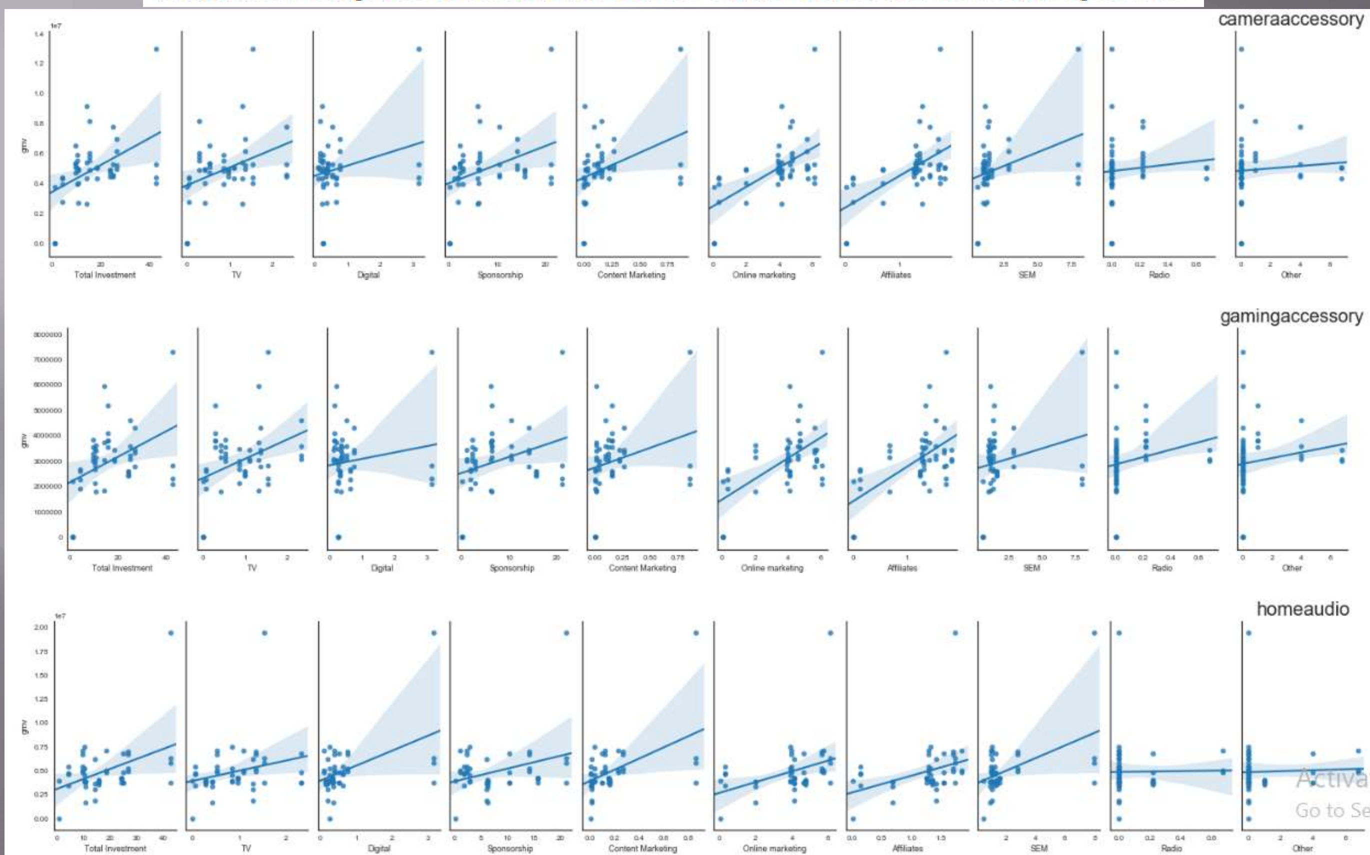


Consumer NPS score is highest in weeks 32 – 35 , which coincides with the time when maximum discounts were being offered.

Company Stock Index has seasonal ups and downs over the span of 1 year.

Appendix: More Visualizations

Relationship between Revenue and Advertisement Spends



TV, Online Marketing & Affiliates seem to have a moderately positive correlation with Revenue.