ELECKART CAPSTONE

SUBMITTED BY:

ADITYA CHOPRA

SAHIL BANSAL

RICHA NARAYAN

KIRAN LVS

Business Objectives

ElecKart is an e-commerce firm based out of Ontario, Canada specializing in electronic products.

Objectives:

- Analyze the previous one year marketing spends & returns on the same
- Develop various market mix models to observe the actual impact of different marketing variables over the last year
- Recommend significant areas which would result into better results
- Recommend the optimal budget allocation for different marketing levers for the next year

Methodology

Data Preparation & Analysis

- Gather & review metadata
- Inspect for quality and consistency
- Fix data quality issues, impute missing values etc.
- Create various plots to visually depict the relationship between different dimensions

Pre-Processing

- Standardize Data
- Create Features for analysis
- Identify and eliminate insignificant and multi-collinear features

Model Creation and Evaluation

- Build model on the data
  - Basic linear model
  - Multiplicative model
  - Kyock's Model
  - Distributed Lag Model
- Choose the best model for each category and it's related important features

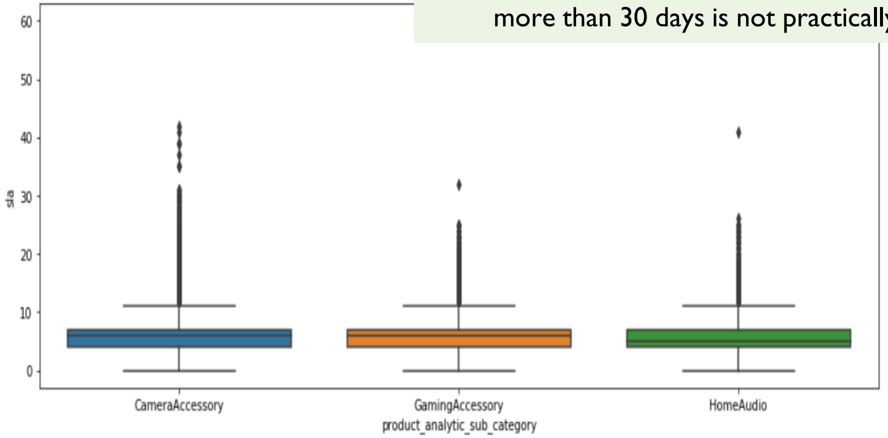
Data Exploration and Cleaning

ConsumerElectronics (1648824 rows, 20 features)		
Data Quality Issues	**Features**	**Resolution**
Duplicate Rows/Transaction	Combination of ["order_date", "order_id", "order_item_id" ,"units"]	Dropped duplicate rows (Duplicate Transactions were dropped) Number of duplicate records removed: 112528
Transactions outside the period of analysis Transaction start date: 2015-05-19 Transaction end date: 2016-07-25	order_datetime	Filtered records for within the period of analysis Transaction start date: 2015-07-01 Transaction end date: 2016-06-30
Very high proportion of null values(78%)	deliverybdays, deliverycdays	Dropped the features
Null Values	product_analytic_vertical	Dropped rows containing null values(the number is quite low as compared to the entire dataset)
Values are 0, which is not logical	product_mrp	Dropped rows containing 0 values(the number is quite low as compared to the entire dataset)
Values are 0, which is not logical	Gmv(target variable)	Dropped rows containing 0 values • Imputing it with 1 (One) won't add any additional value to the target variable • Also, imputing it with MRP * Units might mislead the model from the actual scenario
GMV greater than product_mrp*units	Product_mrp, gmv, units	Dropped such rows as it is not logical that products were sold above MRP
Negative values for product_procurement_sla	product_procurement_sla	Dropped rows containing negative values(the number is quite low as compared to the entire dataset)
Scientific Notation for some features	cust_id, order_id, order_item_id, pincode	Suppressed Scientific Notations

Outlier Analysis – sla/product_procument_sla

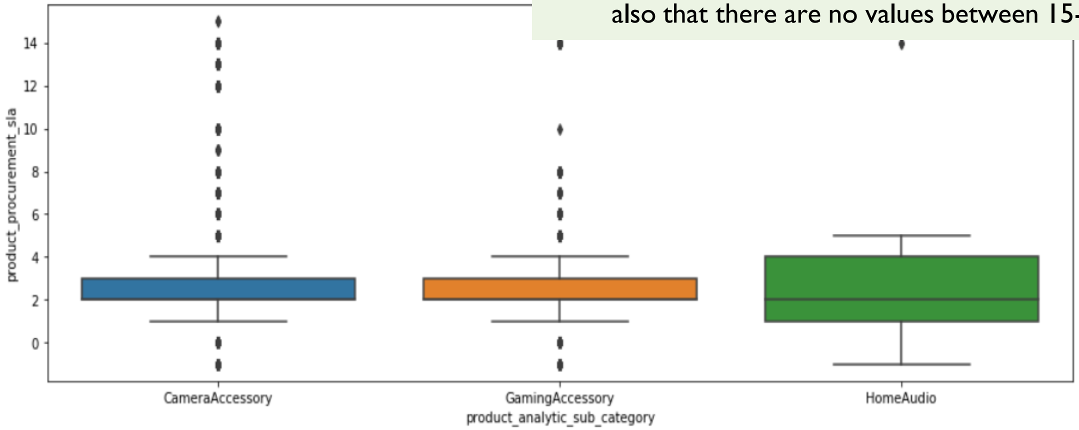
Box Plot for sla

- Only 0.005% of the data has sla more than 30 days
- These sla values have been capped to 30 days, as more than 30 days is not practically acceptable



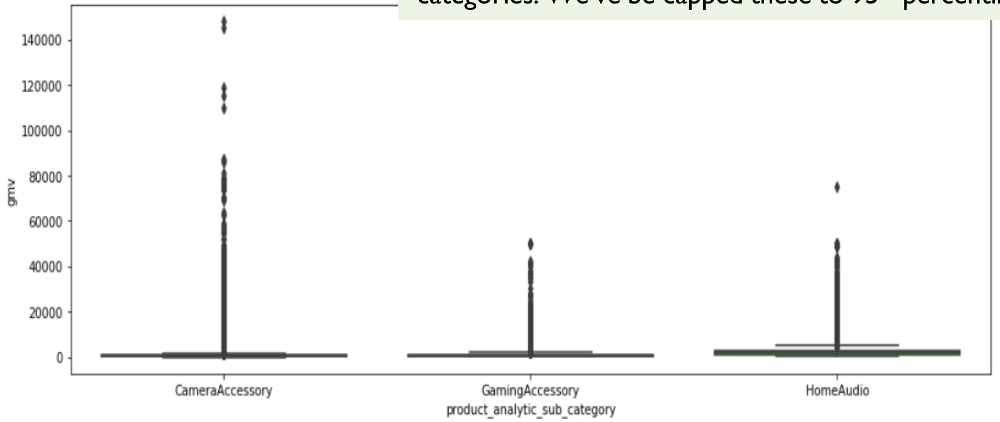
product_procurement_sla

- There are 4.2% of negative values, with 1.9% for the 3 relevant categories, these rows are dropped
- The sla's greater than 60 were capped to 15, as value greater than 60 does not seem practically acceptable & also that there are no values between 15-60

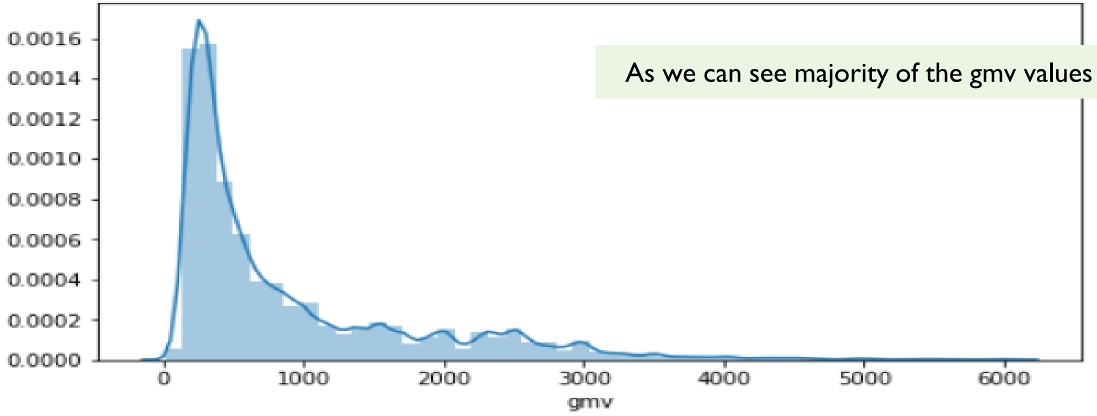


Box Plot for GMV

We see a humongous amount of outliers in the category all 3 categories. We've be capped these to 95th percentile.



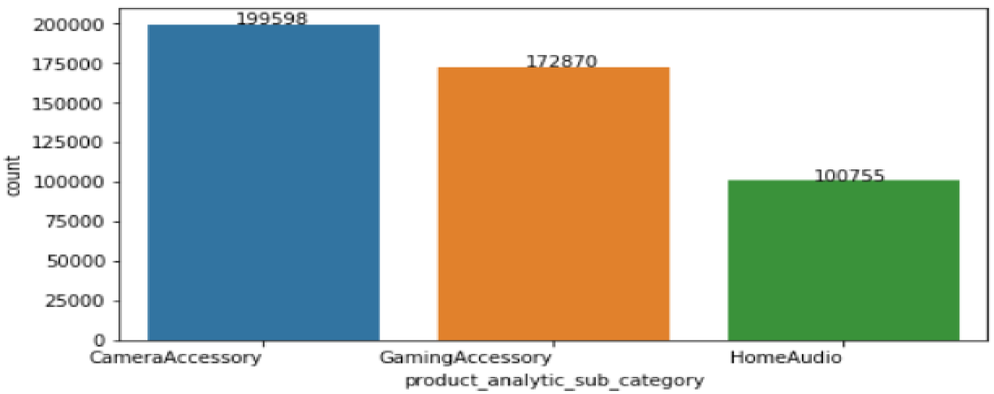
Distribution of GMV after outlier analysis



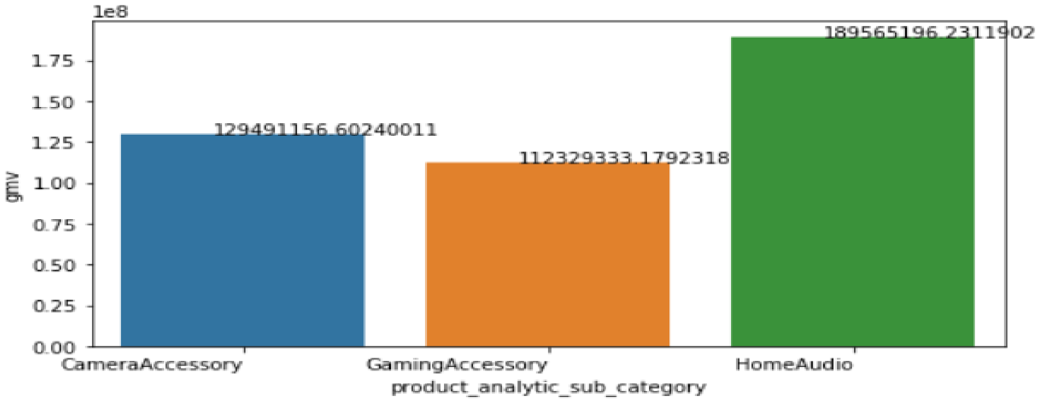
As we can see majority of the gmv values are under 1000

EDA: Part I

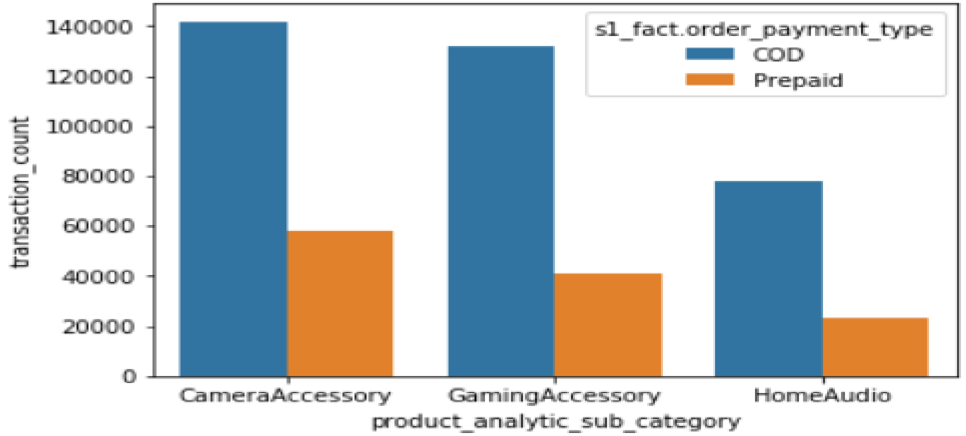
No of transactions by Sub-Category



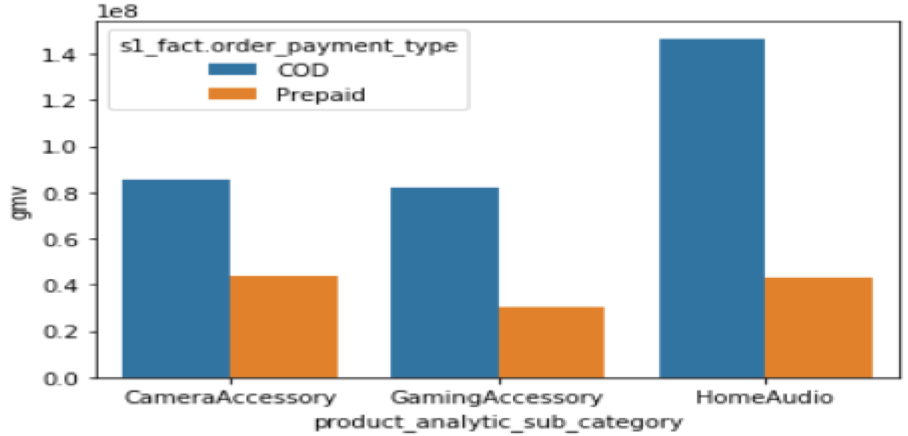
Total GMV by Sub-Category



No of transactions by Sub-Category and payment mode



Total GMV by Sub-Category and payment mode



- As we can the number of transactions are highest for Camera Accessory however gmv is largest for Home Audio
- Another observation is consumers prefer COD over Prepaid payment mode

Engineered KPIs

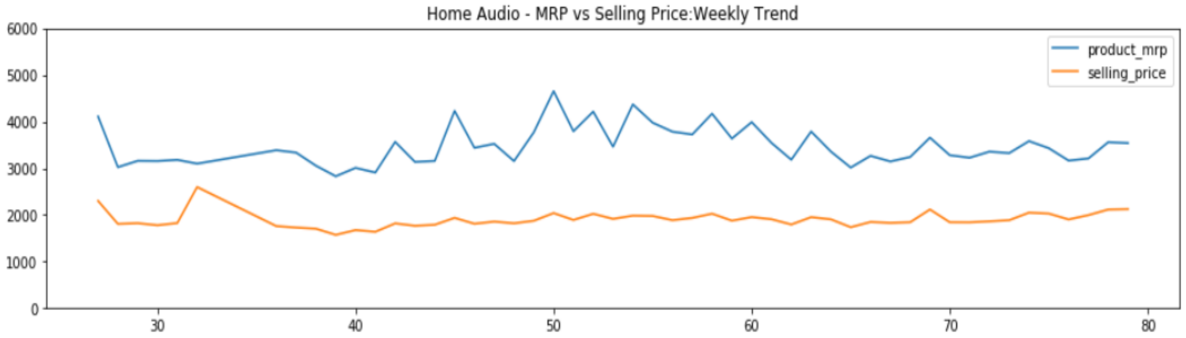
KPI	Description
order_week	Order_week column was created from order_date to enable weekly analysis of the dataset
payment_mode_indicator	Feature indicating payment_mode was created from the original feature so that the algorithm can understand it
selling_price, discount	Created Selling price i.e. price per unit from GMV & Units, Discount offered on products was created from GMV & selling price But selling_price wasn't included in the model building
total_holiday, is_holiday	Curated the list of public holidays in Ontario, Canada & mapped to the data, total_holiday captures total holidays in a week, is_holiday captures whether a week has a holiday
total_special_sale_day, is_special_sale_day	total_special_sale_day captures special sale days in a week, is_special_sale_day captures whether a week has a special sale day
total_pay_days, is_pay_days	total_pay_days captures special sale days in a week, is_pay_days captures whether a week has a pay day
COD_count, Prepaid_count, pct_online_transactions	COD_count and Prepaid_count was created to capture the number of Cod transactions and Prepaid transactions for each subcategory pct_online_transactions capturing the percentage of prepaid orders vs the total orders in each week. However, only pct_online_transaction was passed on to the Model building as it captures the information from both COD_count & Prepaid_count
product_analytic_vertical	Product analytical verticals KPIs were created for each subcategory to capture the number of units sold in each product analytical vertical for the respective subcategories in each week
Adstocks	Created adstocks for each of the advertisement investment at 0.5
NPS	Integrated NPS & Stock Index to the dataset
Climate Data	Included Temperature column to analyze (not included in the final model as its not a factor which can be controlled by the company) But still we can have discounts based upon the climatic conditions as well. Also, we observed some trends b/w the GMV & climate using plots
Advanced KPIs	Lag values were created for Week 1, 2 and 3 for gmv, adstocks, sla(s), NPS, Percentage online transactions for Distributed lag modls & lag value for gmv for week 1 only for Kyock's Model

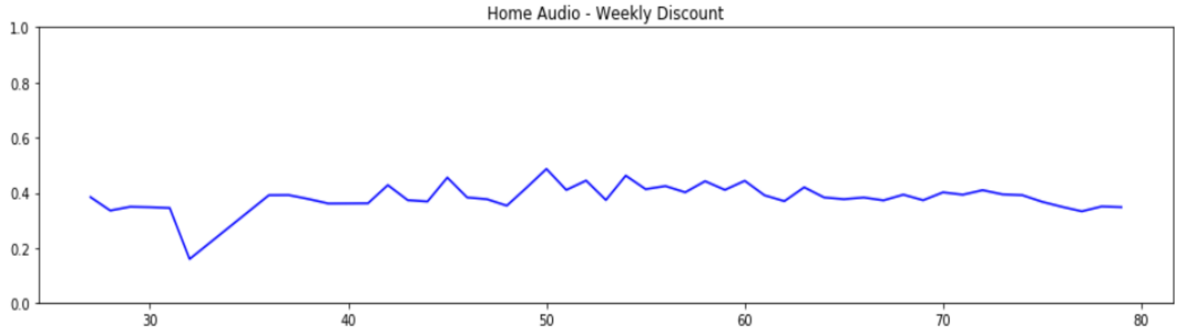
EDA: Weekly Trends – GMV and Number of Transactions

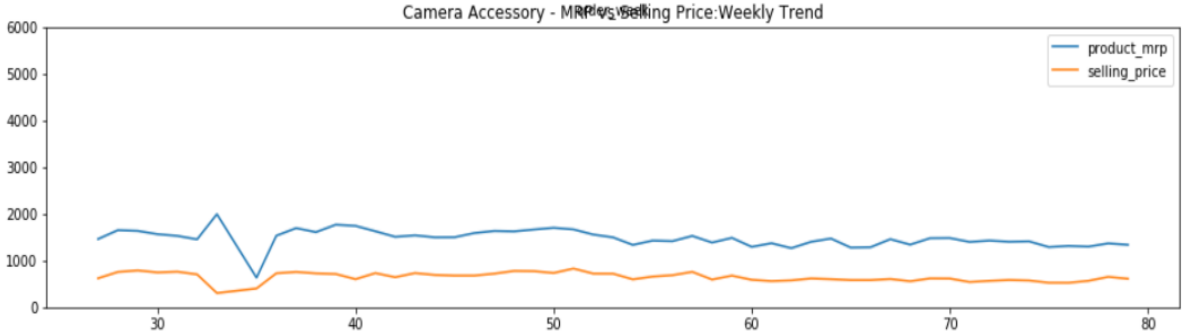
- GMV and number of transactions seem to be higher during weeks that have a holiday or special sale day (Plots added to annexure)
- Observing the trends of Total GMV & Number of transactions across the weeks
- Highest GMV seems to be around week 42 for all categories with the maximum gmv for HomeAudio
- Highest number of transactions are around week 42 with maximum transactions for Gaming Accessory

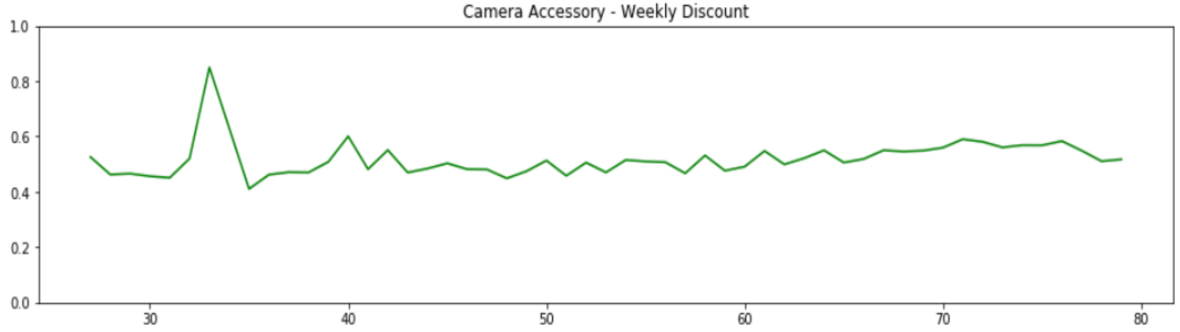
EDA: Weekly Trend – Selling Price, MRP and Discount

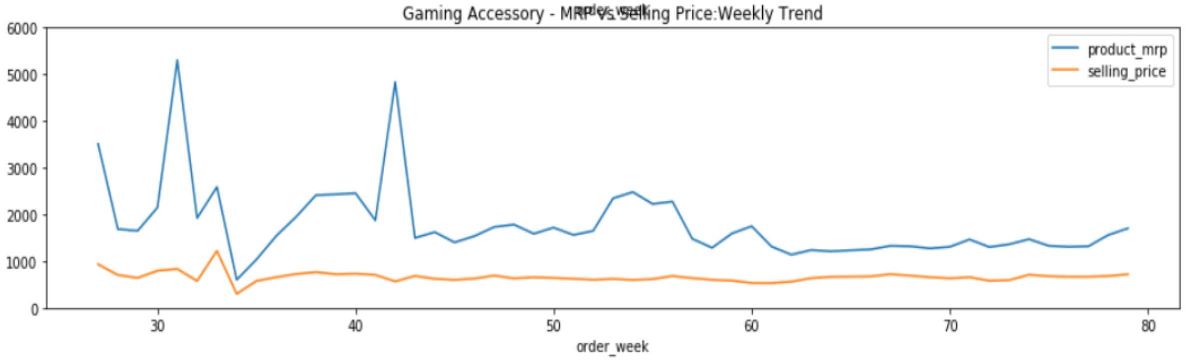
- Product_mrp and selling price seem to be highest for HomeAudio. We can also see huge fluctuations in the product_mrp for GamingAccessory
- Discounts seem to be highest for Camera Accessory

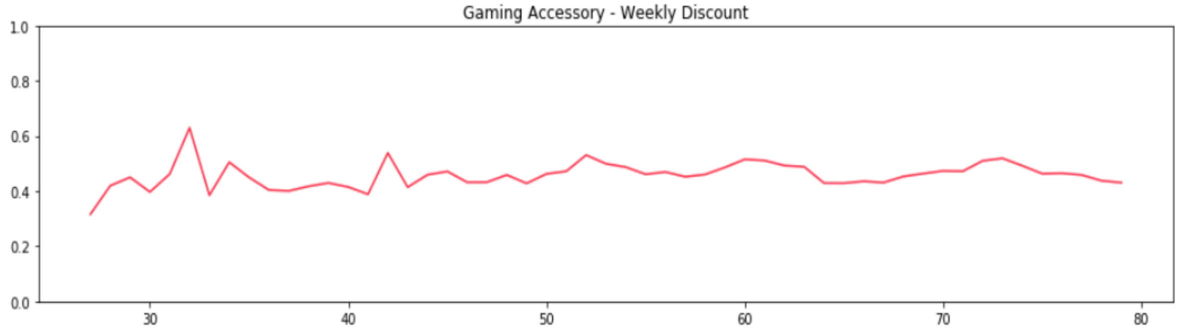








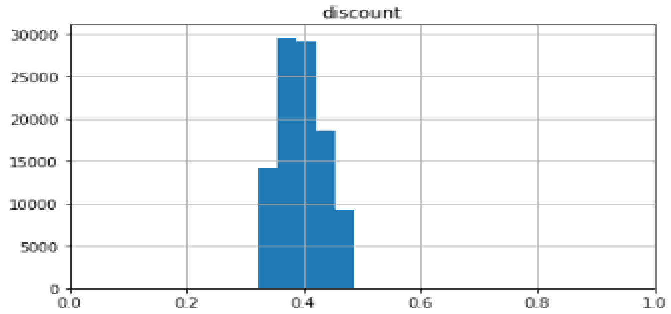




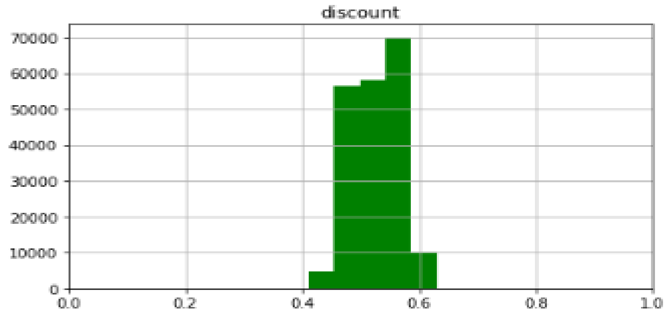
EDA: Discount Analysis

Hist Plot for Number of transactions vs Discount

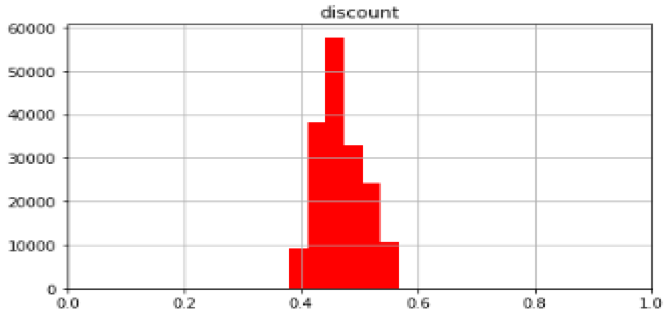
Home Audio



Camera Accessory

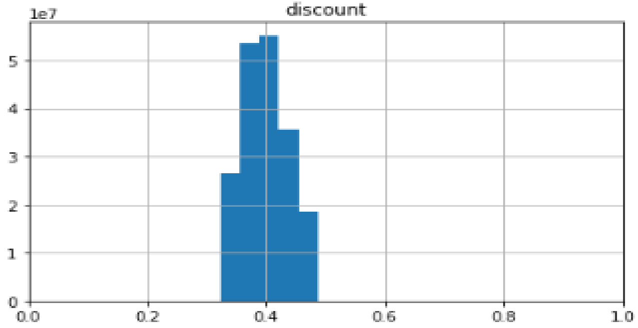


Gaming Accessory

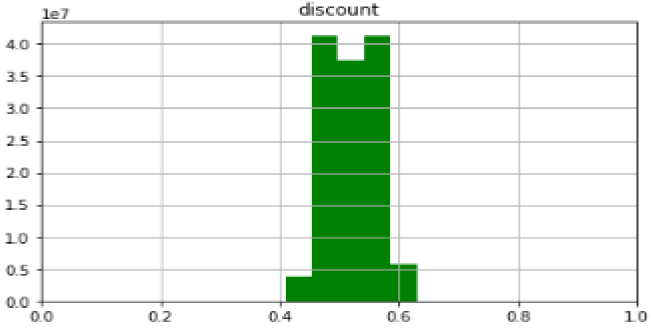


Hist Plot for GMV vs Discount

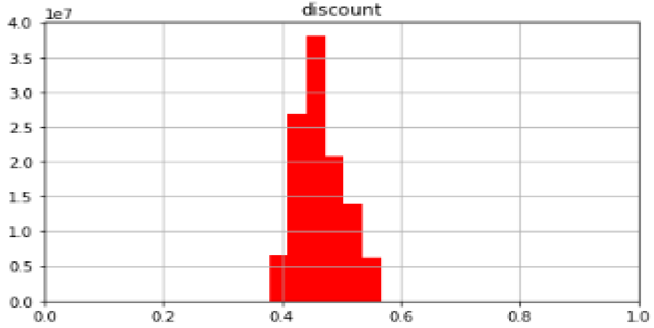
Home Audio



Camera Accessory



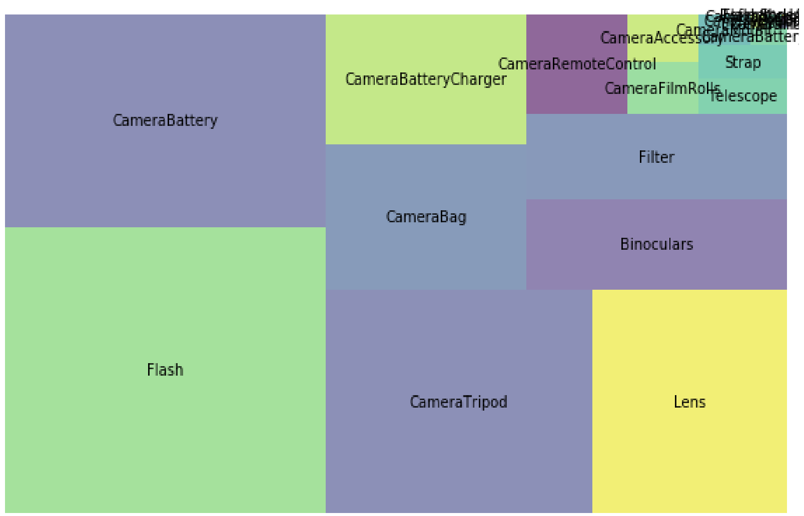
Gaming Accessory



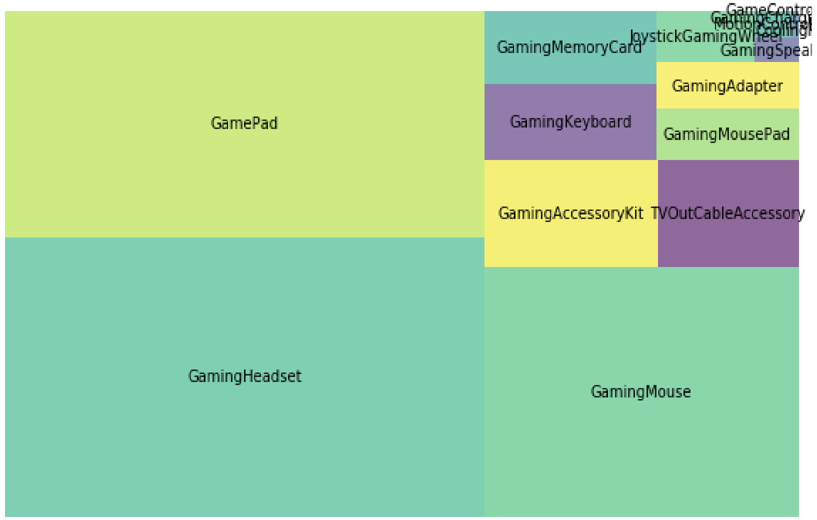
- In case of HomeAudio and Gaming Accessory, discounts over 50% is not very useful
- In case of Camera Accessory, discounts could go upto approx. 58-59%.

EDA: Product Analytic Vertical

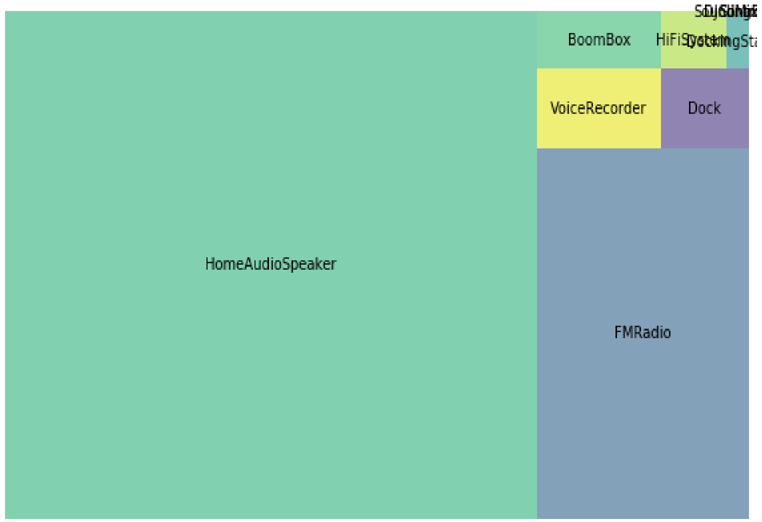
Camera Accessory



Gaming Accessory



Home Audio



Highest selling product is Flash followed by Camera battery and Camera tripod

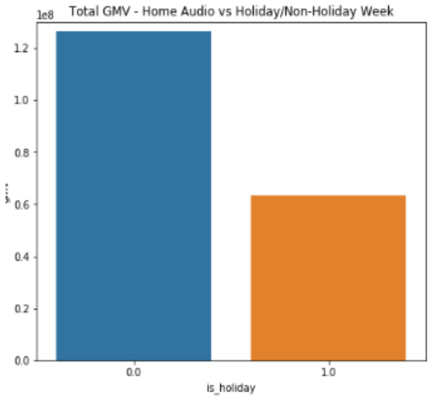
Highest selling product is Gaming headset followed by Gaming pad and Gaming mouse

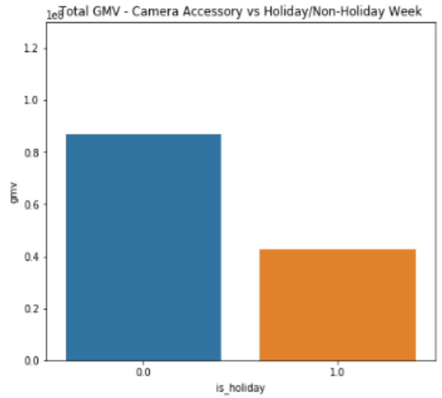
Highest selling product is HomeAudio Speaker followed by FMRadio

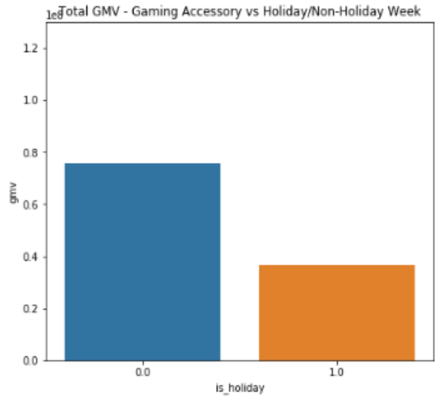
EDA: Analysis of Holiday vs GMV

- Overall, total gmv on weeks without holidays is more, however mean_gmv (i.e. Total GMV/ Number of Transactions) for weeks with holiday is higher.

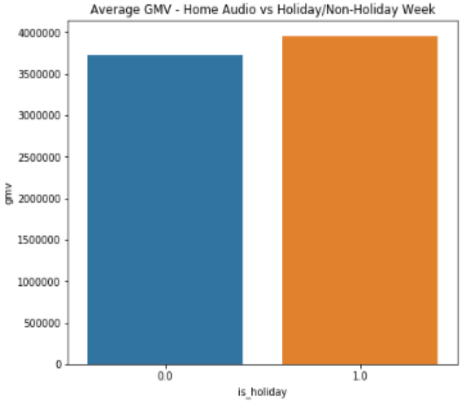
Total GMV

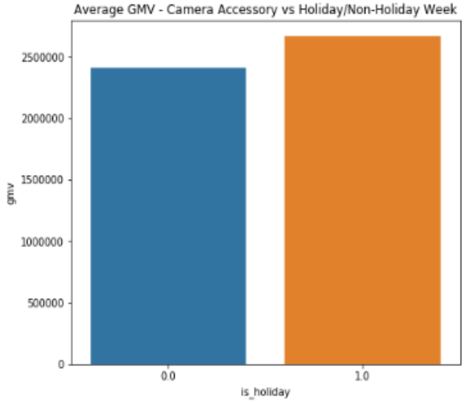


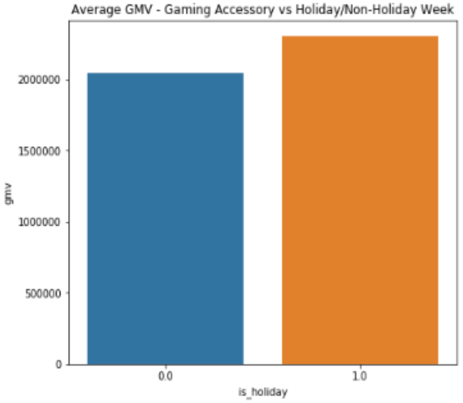




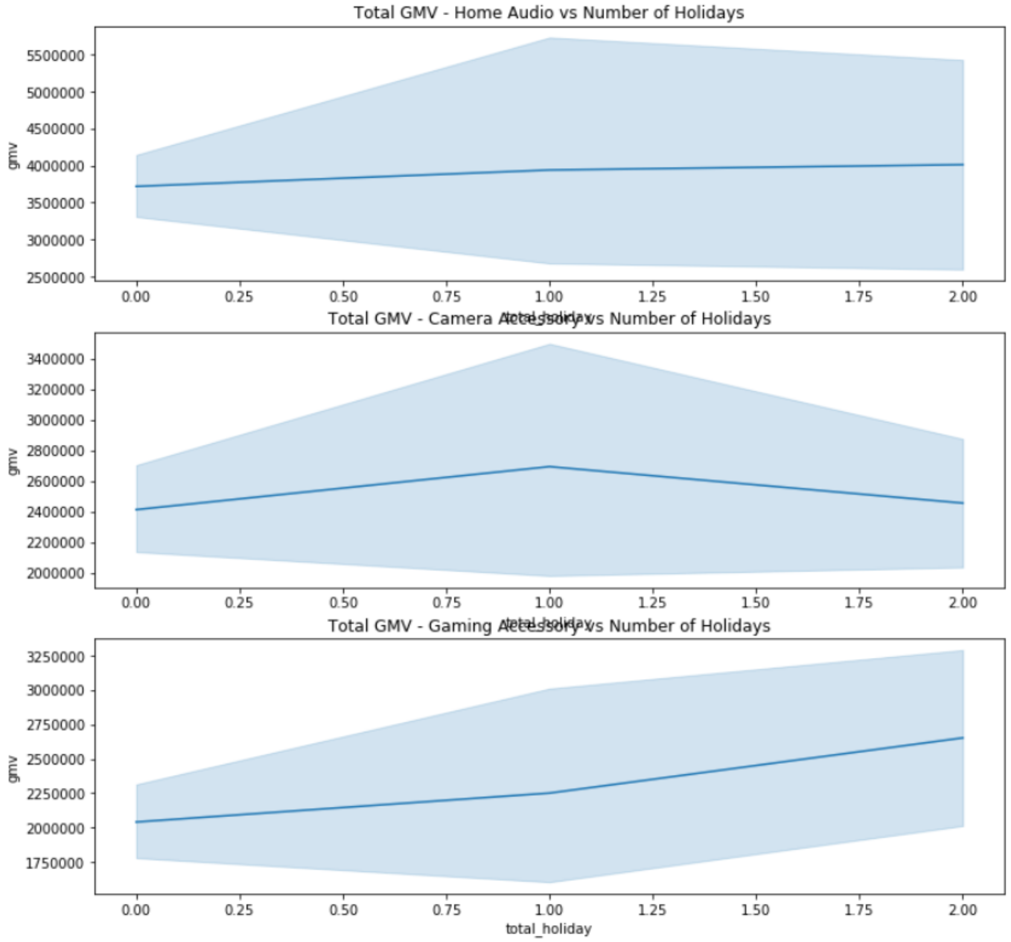
Mean GMV







- Total number of holidays shows a great effect in the Gaming Accessory category.
- For Camera Accessory, the gmv can be seen increasing in a linear fashion but then suddenly dropping a bit at number of holidays = 2
- For Home Audio, the sales are pretty constant across the holiday count

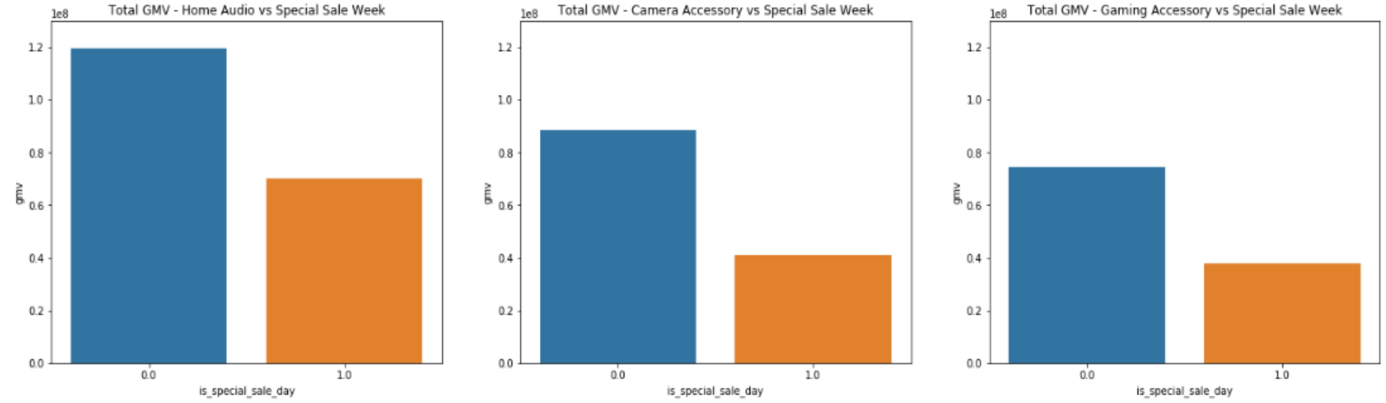


EDA: Analysis of Special Sale Day vs GMV

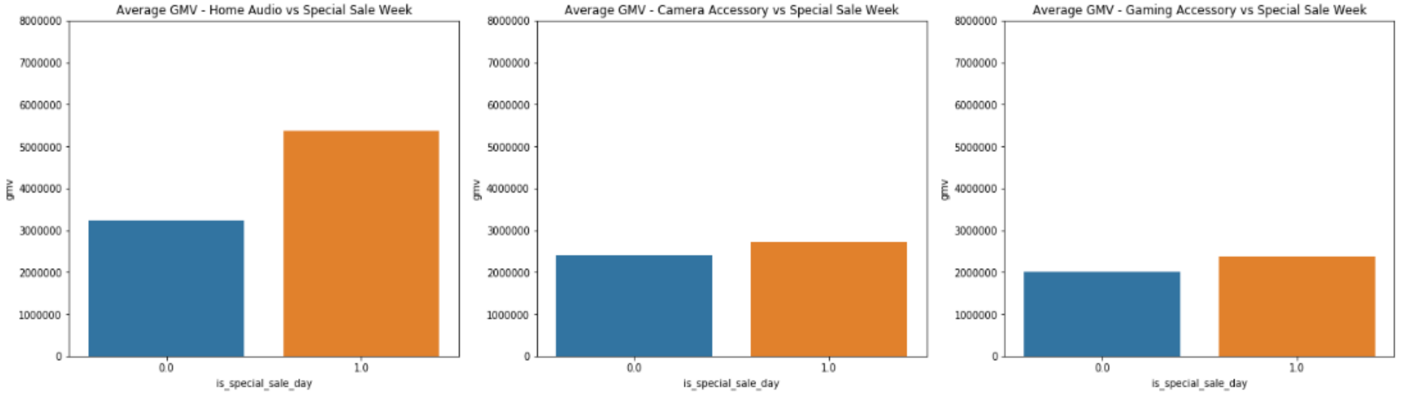
- Overall, total gmv on weeks without special sale days is more, however mean_gmv for weeks with special sale days is higher.
- We can also see the special sale have more effect on Home Audio than other categories

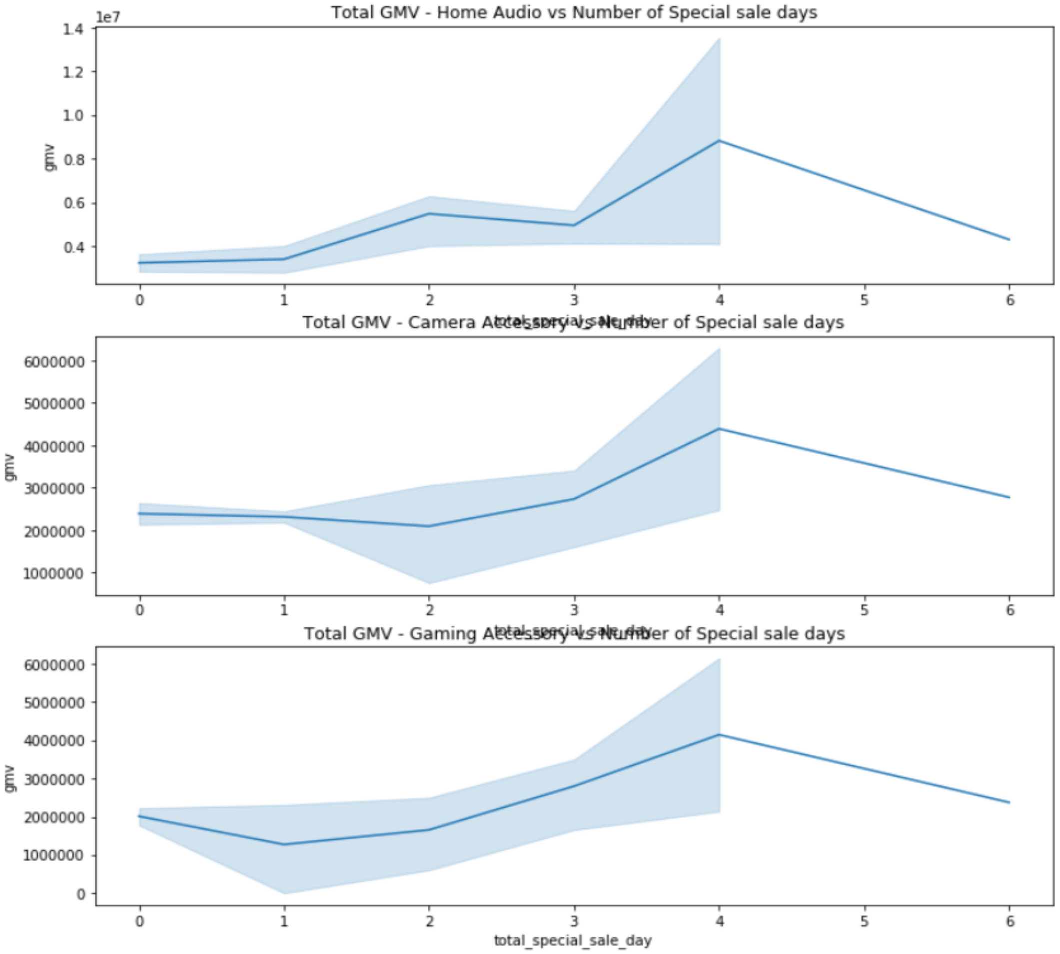
- Total_special_sale_days again seems an important feature.
- Among all the three categories 2-4 number of special sale days in a week, the sale is showing an increasing trend.
- Also we can see that if the number of special sale days increases 4 in a week the sales come down drastically among all the three categories

Total GMV

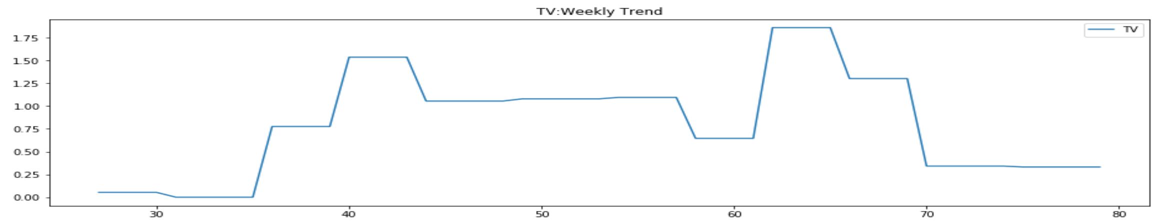


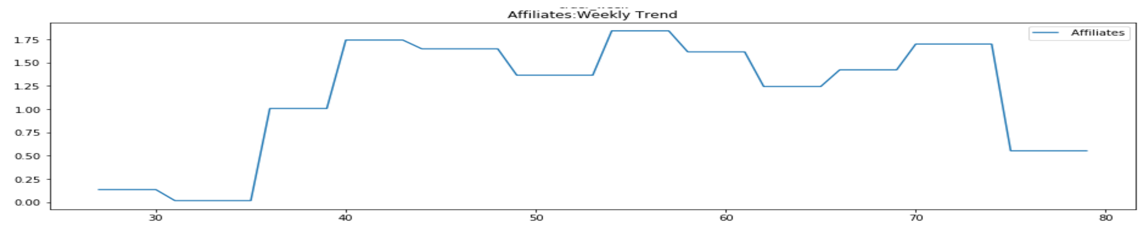
Mean GMV

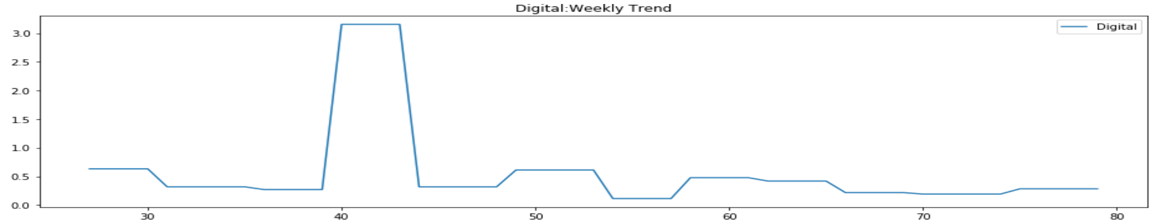


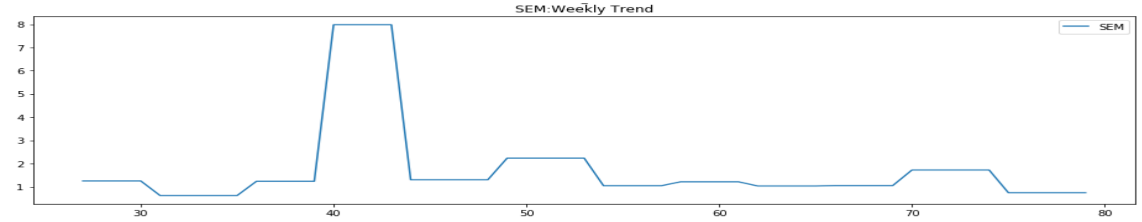


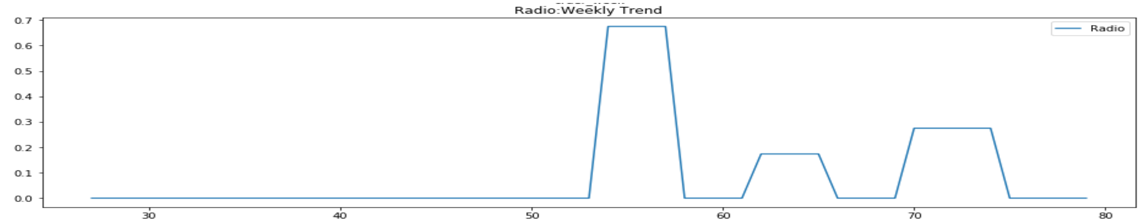
EDA: Media Investments – Weekly Trend

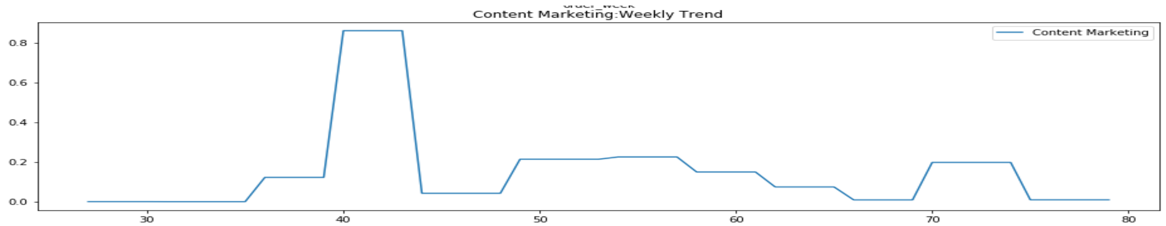


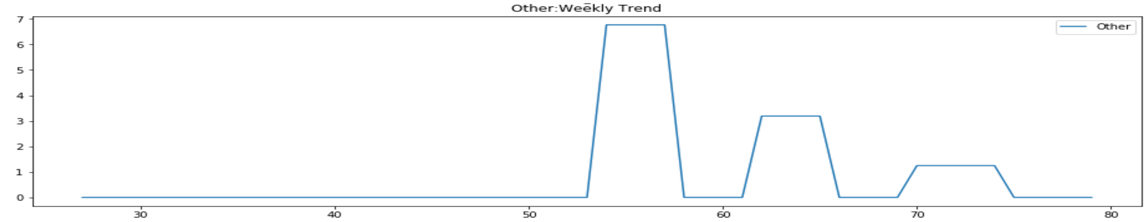


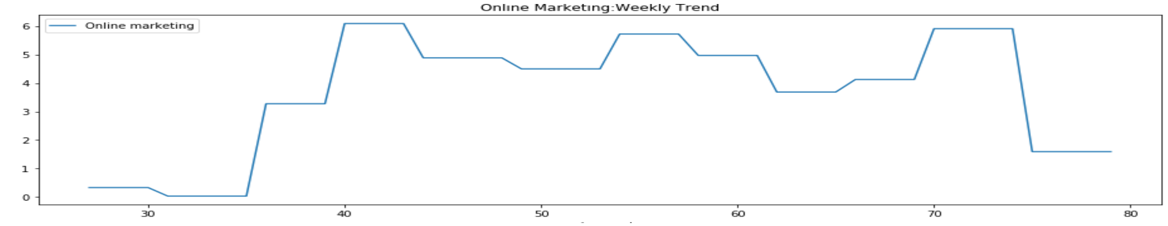










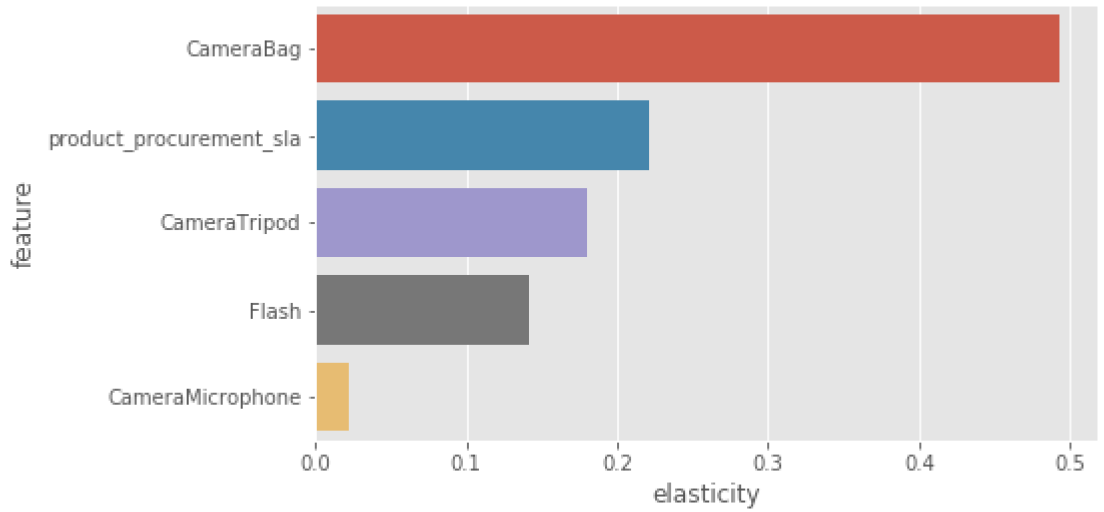


- The maximum investments occurred between week 39-45 for most channels
- Maximum investments was done through Sponsorship media

Camera Accessory – Comparison of 4 models

Model	Significant Variables	Adjusted R-square (with cross validation)	MSE
Simple Linear Model	discount, NPS, pct_online_transactions, CameraBag, CameraBattery, CameraTripod, Flash	0.571	0.00081
Multiplicative Model	adStock_Radio, sla, CameraTripod	0.451	0.00722
Kyock's Model	product_procurement_sla, CameraBag, CameraMicrophone, CameraTripod, Flash	0.904	0.00187
Distributed Lag Model	NPS, CameraBattery, CameraTripod, Filter, Flash	0.891	0.001889

Kyock's Model is selected as the best model based on the highest Adjusted R-square and low MSE values. Also, contains the features that the company can act upon

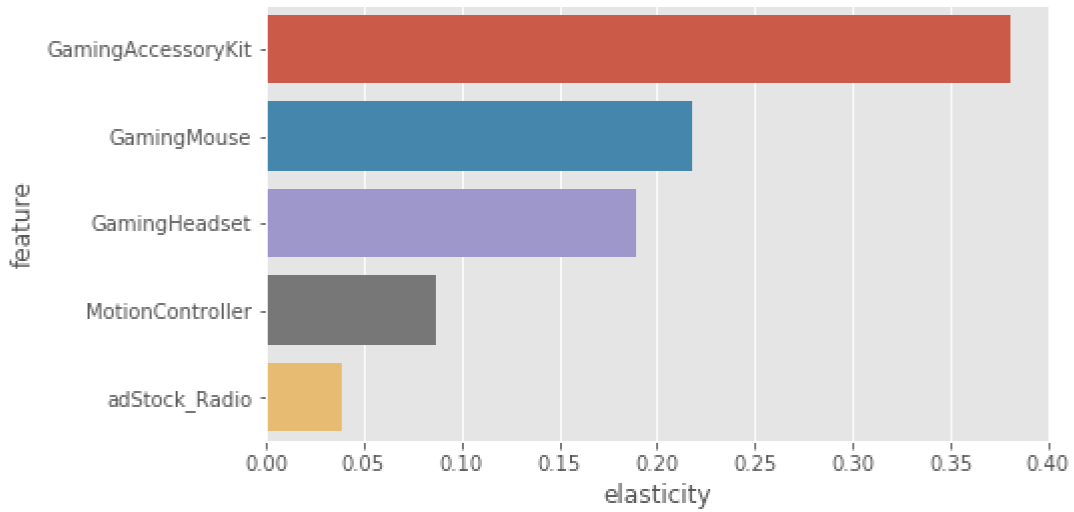


- **The adjoining figure represents the elasticity of different variables w.r.t. GMV**
- **Positive KPI means positive change in the KPI will lead to positive change in target variable**
- **From the graph, we can see that, the sale of Camera Bag, Product procurement SLA, CameraTripod, Flash & Camera Microphone have positive impact on the GMV value. Hence, the company should promote these products or pitch in more products in these categories**

Gaming Accessory – Comparison of 4 models

Model	Significant Variables	R-square(with cross validation)	MSE
Simple Linear Model	adStock_Radio, GamingAccessoryKit, GamingHeadset, GamingMouse	0.888	0.00308
Multiplicative Model	discount, GamingHeadset, GamingMouse	0.592	0.01761
Kyock's Model	adStock_Radio, GamingAccessoryKit, GamingHeadset, GamingMouse, MotionController	0.911	0.00235
Distributed Lag Model	GamingAccessoryKit, GamingHeadset, GamingKeyboard, GamingMouse, GamingSpeaker	0.910	0.00207

Kyock's Model is selected as the best model based on high R-square and low MSE values & the features which the company can act upon

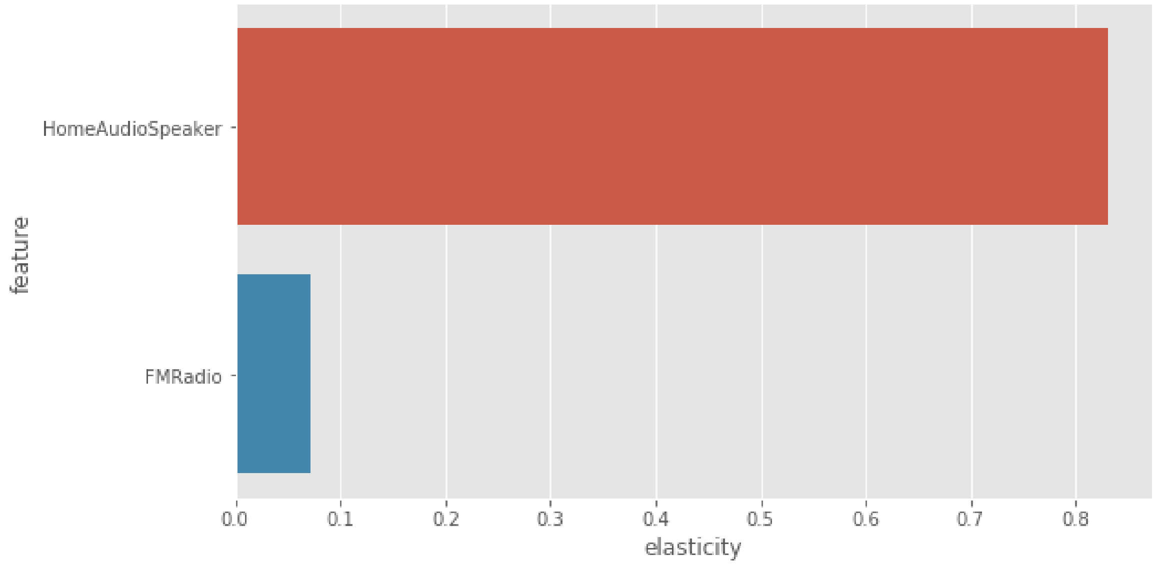


- **The adjoining figure represents the elasticity of different variables w.r.t. GMV**
- **Positive KPI means positive change in the KPI will lead to positive change in target variable**
- **From the graph, we can see that, the sale of GameAccessoryKit, GamingMouse, Gaming Headset, Motion Controller has positive impact on the GMV value**
- **Hence, the company should promote these products or pitch in more products in these categories**
- **We also see that adstock Radio has a positive impact on GMV, the company should look into the optimum spend on this channel**

Home Audio – Comparison of 4 models

Model	Significant Variables	R-square(with cross validation)	MSE
Simple Linear Model	FMRadio, HomeAudioSpeaker, VoiceRecorder	0.985	8.162359e-05
Multiplicative Model	sla, HomeAudioSpeaker, SoundMixer	0.904	0.0002496
Kyock's Model	adStock_Other, HomeAudioSpeaker	0.969	0.00017
Distributed Lag Model	FMRadio, HomeAudioSpeaker	0.978	0.00012

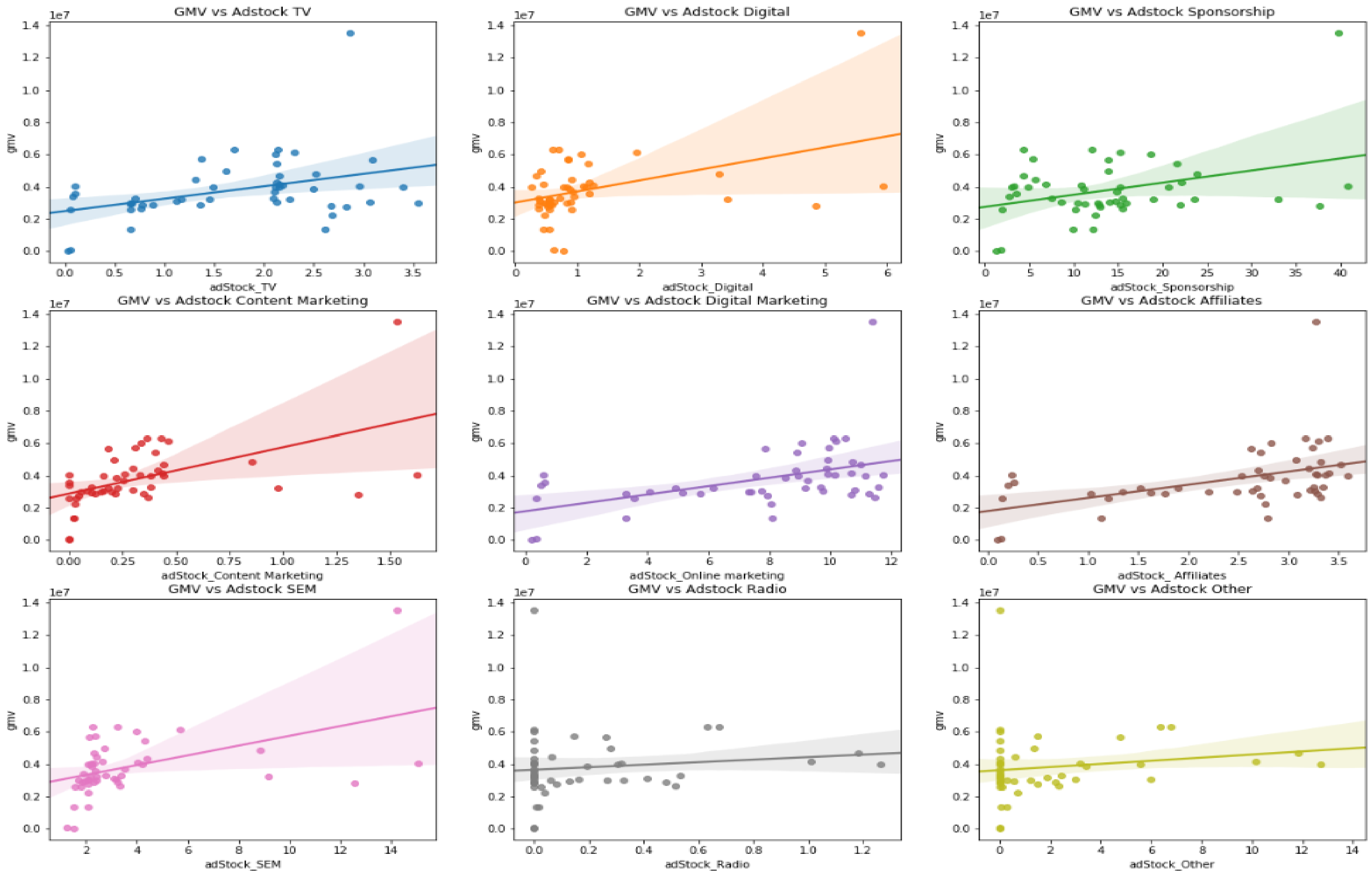
Distributed Lag Model is selected as the best model based on lesser number of features being able to explain almost equal variance as explained by the feature with the highest R-square and low MSE values so that the company has one less feature to focus upon & grow almost equivalently.



- **The adjoining figure represents the elasticity of different variables w.r.t. GMV**
- **Positive KPI means positive change in the KPI will lead to positive change in target variable**
- **From the graph, we can see that, the sale of HomeAudio Speaker, and FMRadio has positive impact on the GMV value**
- **Hence, the company should promote these products or pitch in more products in these categories**

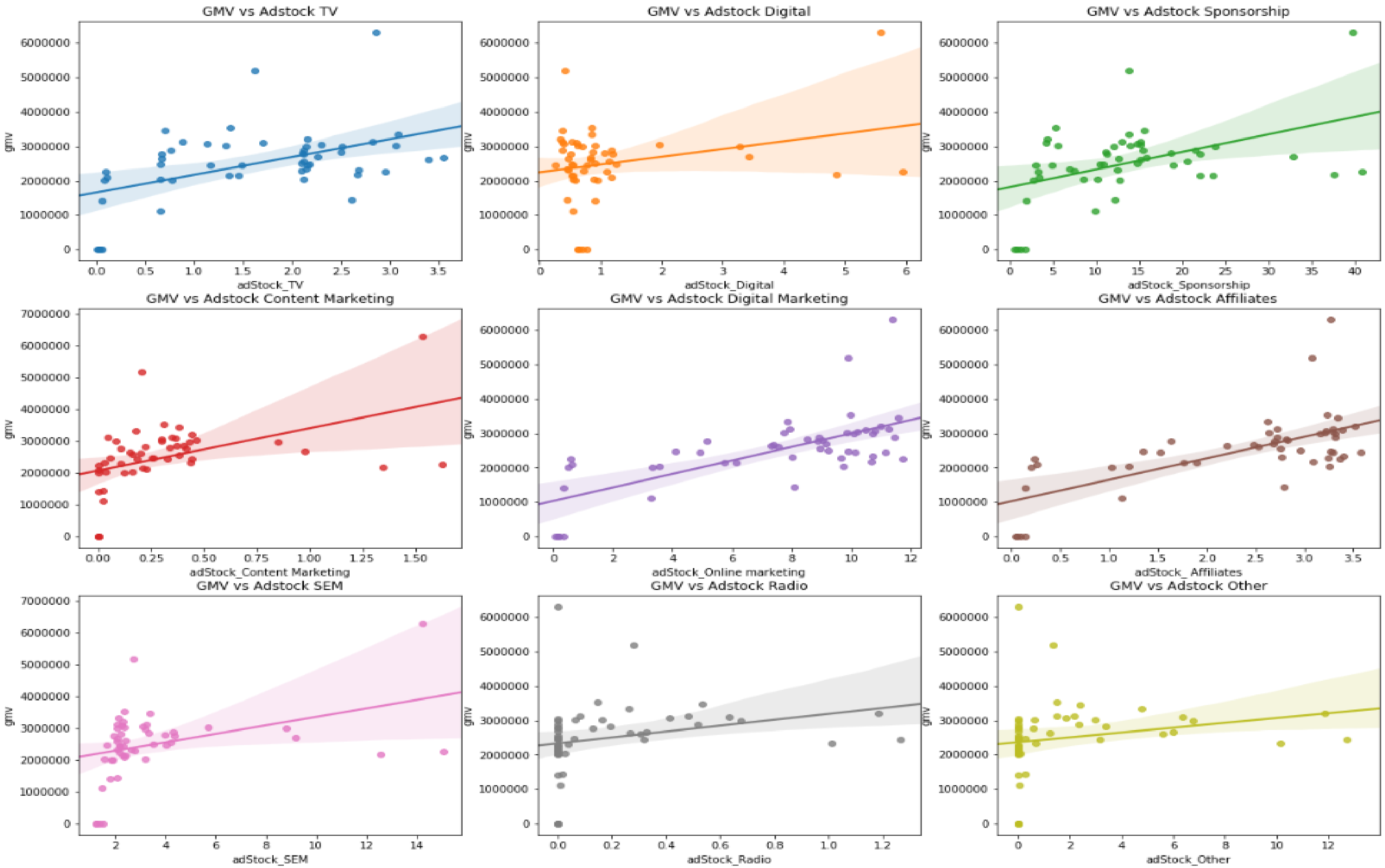
Thank You

Appendix - EDA: Adstock – Home Audio



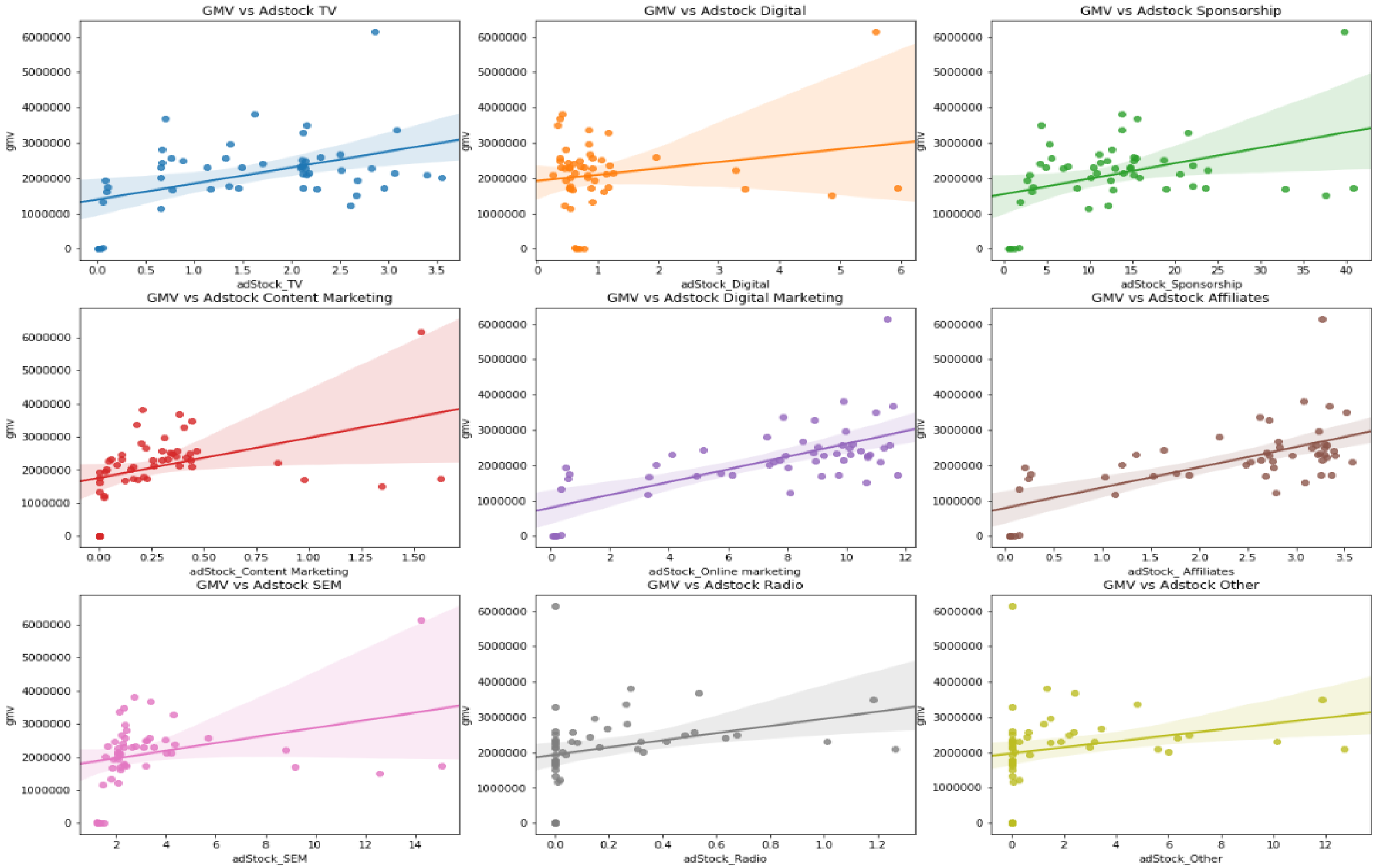
- There seems to be slightly positive correlation between Home Audio GMV and Adstock for different channels

Appendix - EDA: Adstock – Camera Accessory



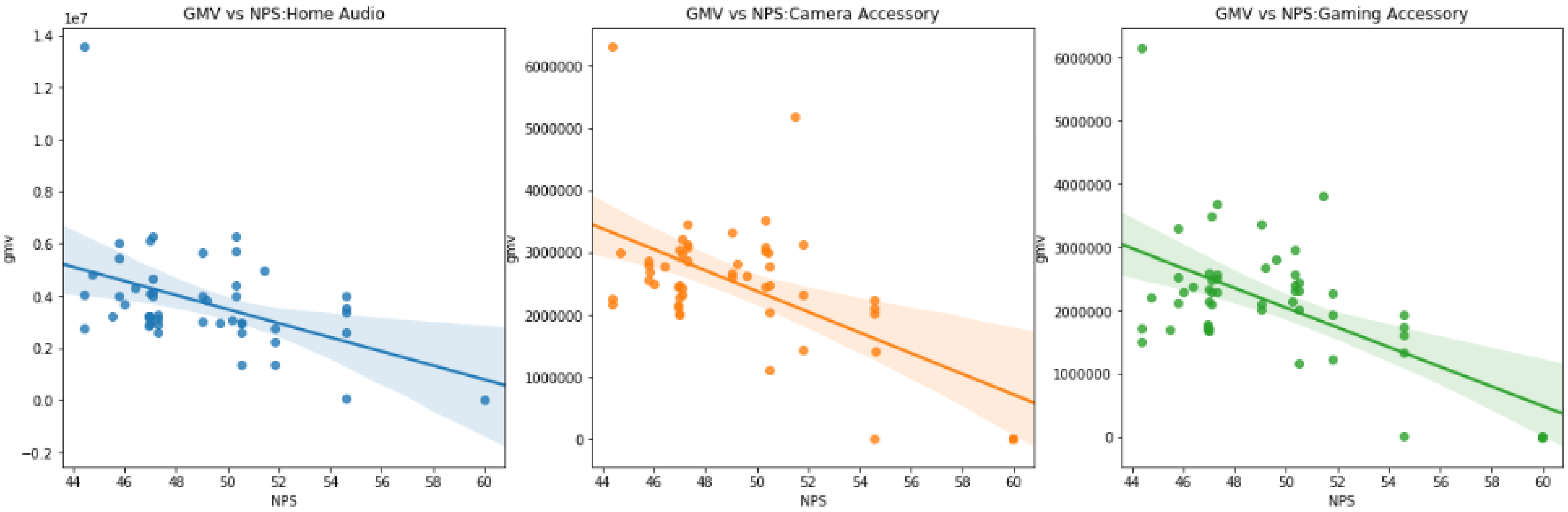
- There seems to be slightly positive correlation between Camera Accessory GMV and Adstock for different channels

Appendix - EDA: Adstock – Gaming Accessory



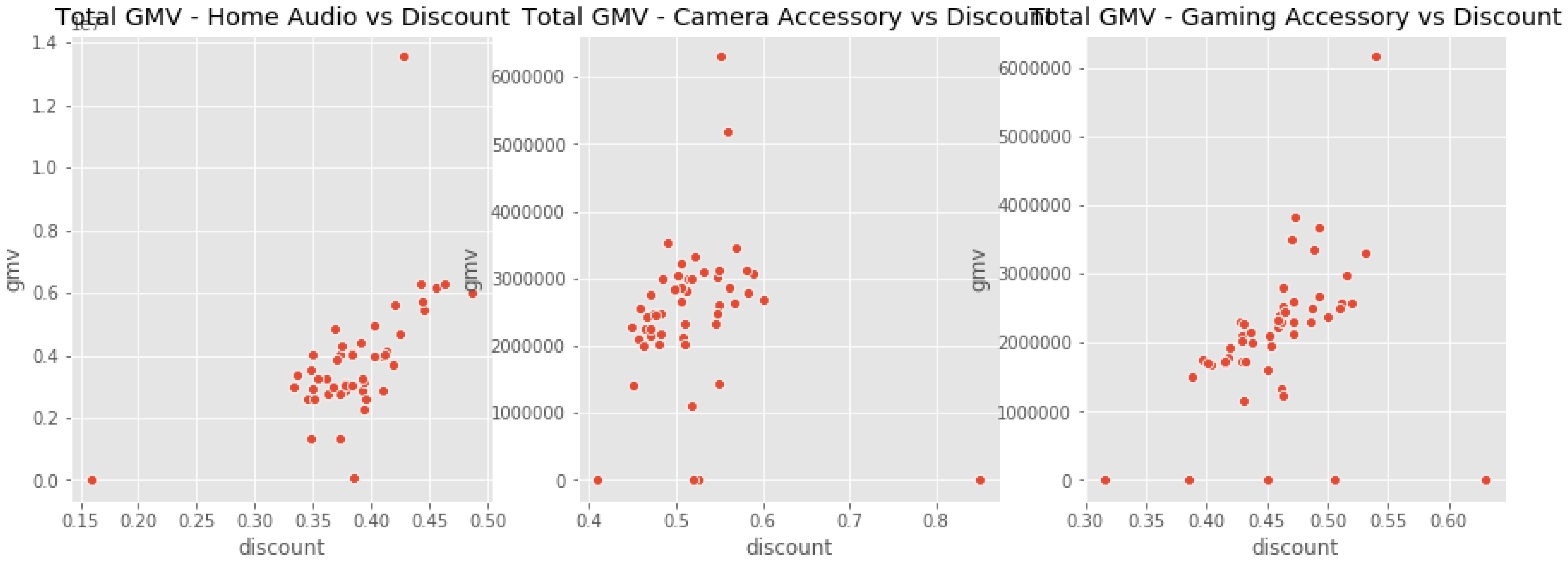
- There seems to be slightly positive correlation between Gaming Accessory GMV and Adstock for different channels

Appendix - EDA: GMV vs NPS



There seems to be a negative correlation between GMV and NPS for all 3 subcategories

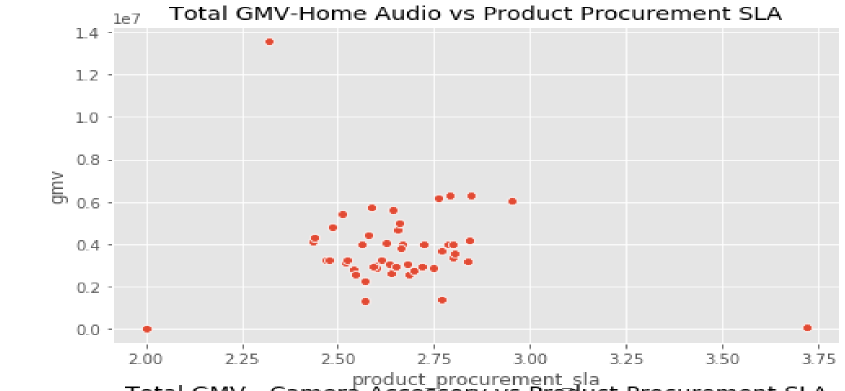
Appendix – GMV vs Discount

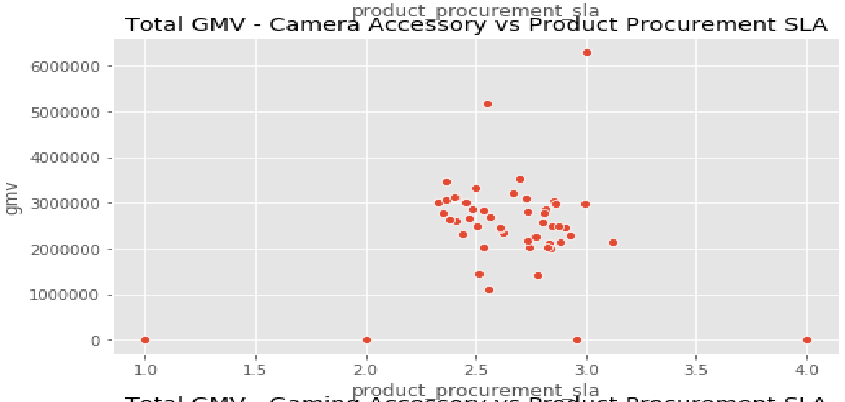


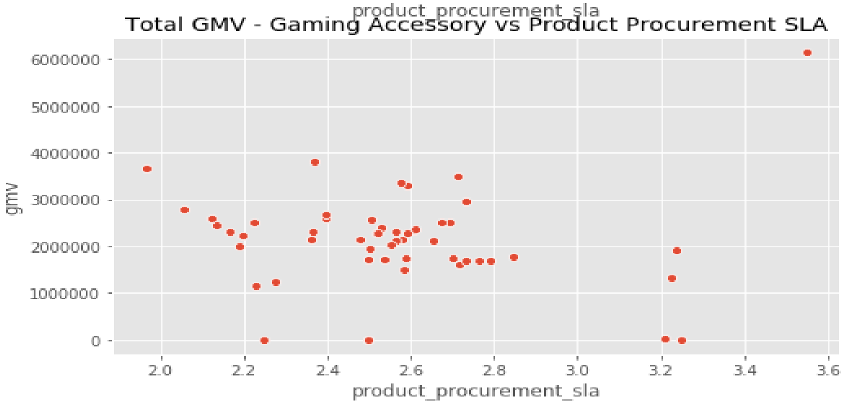
Clearly, we can see a positive relation b/w the discount & gmv for Gaming Accessory & Home Audio

However, GMV for Camera Accessory seems to be pretty ineffective of discount offered as we see both high & low values at the same discount percentage

Appendix – GMV vs product_procurement_sla

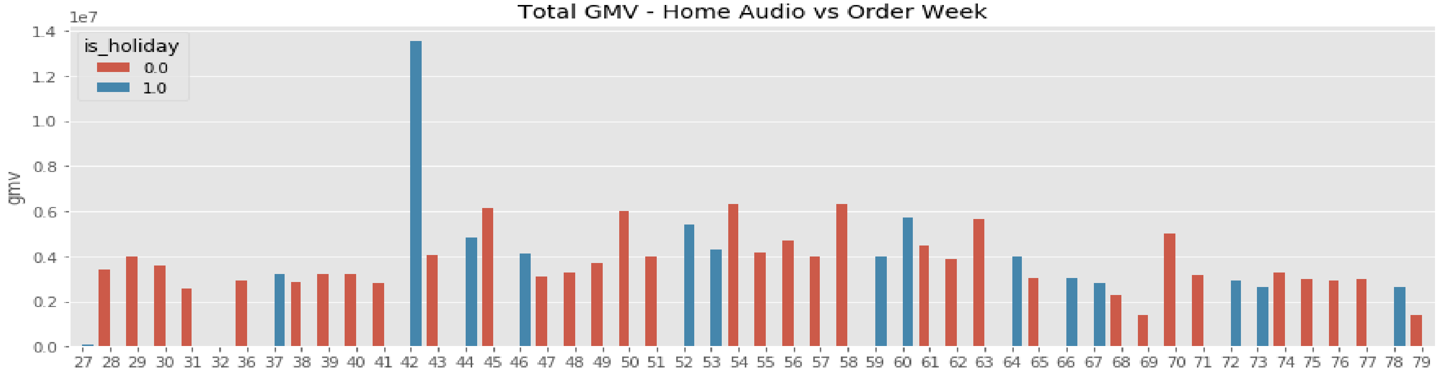


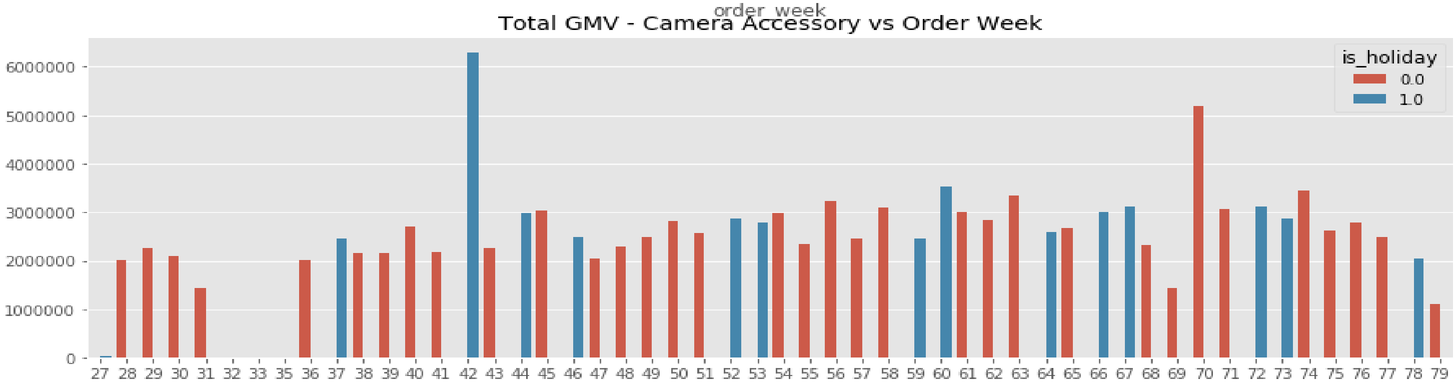


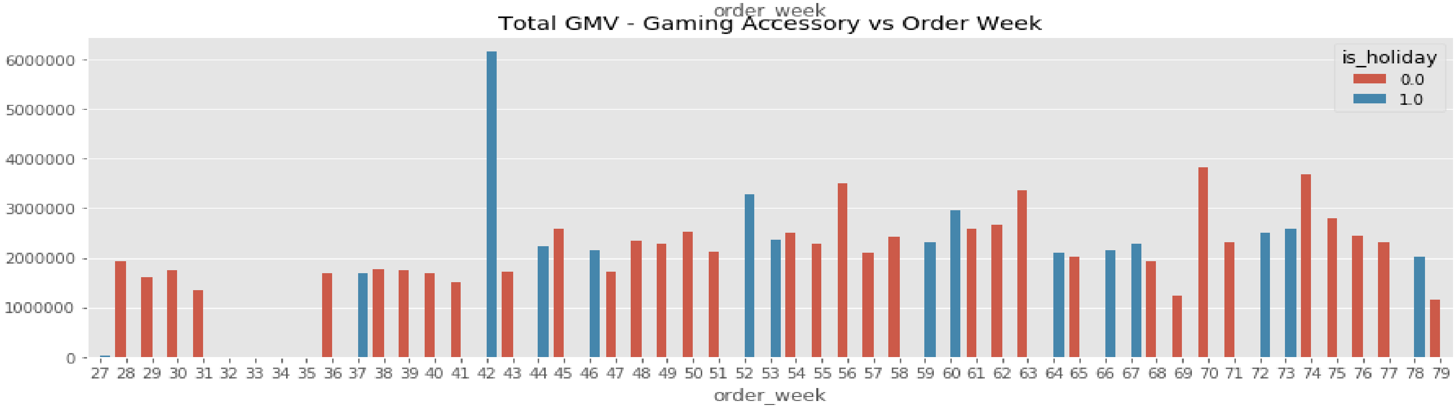


For Gaming accessory, we can see that there is a negative relation b/w the product_procurement_sla & gmv which seems practical, as there would be higher number of orders if the delivery time is less & the product procurment sla definitely affects the total delivery time

Appendix - Total GMV vs Holiday : Weekly Trend



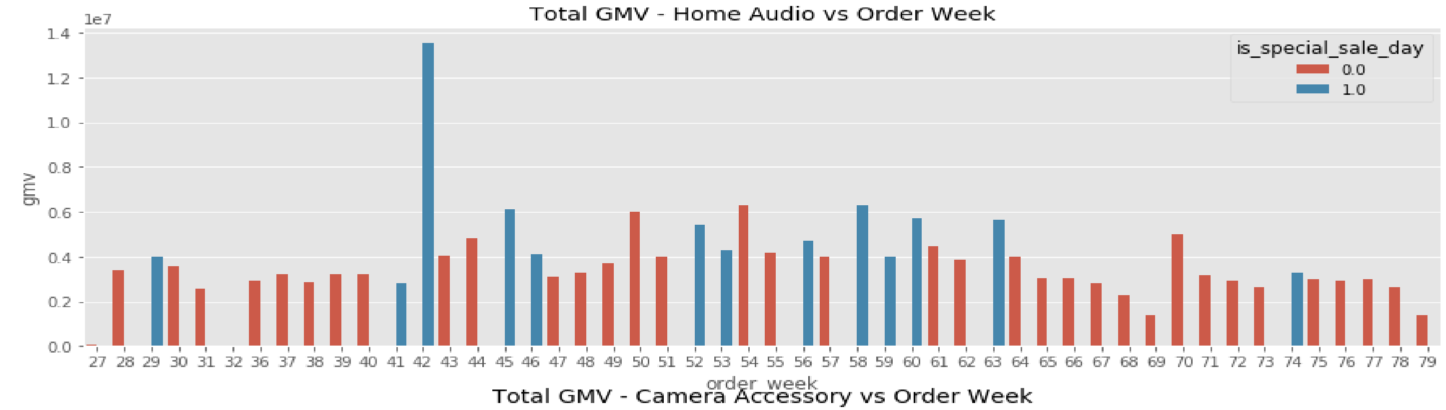


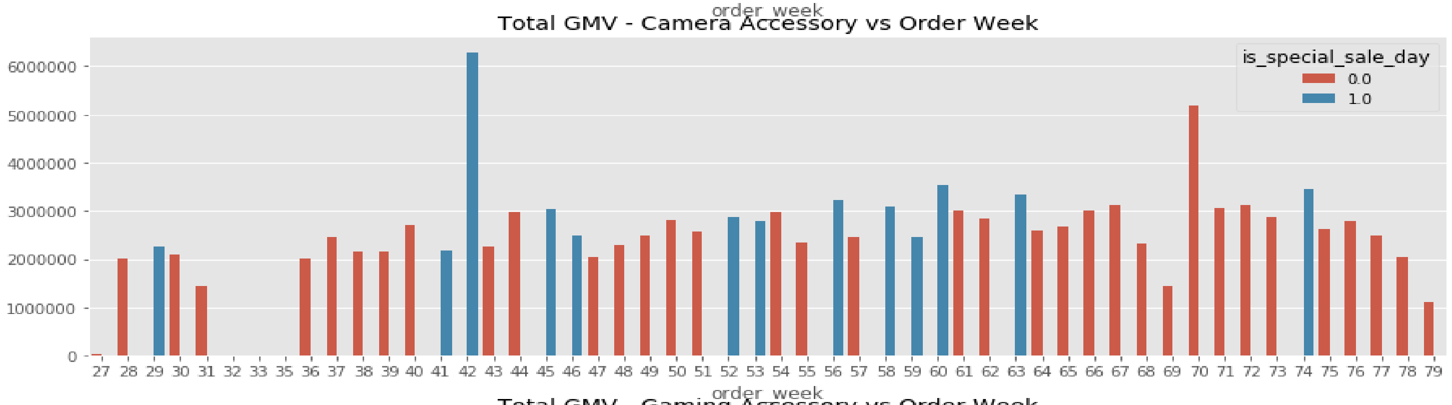


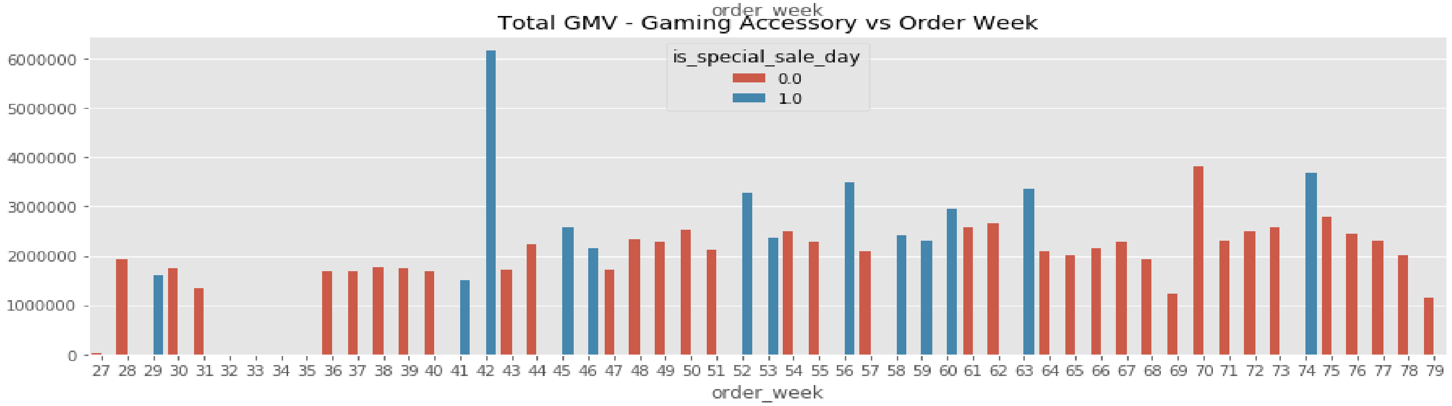
Highest sales can be seen in the 42nd week across all the 3 categories.

However, not much trend can be seen i.e. there are times when it was a holiday in the week but the sales are pretty low compared to the adjacent week when there was no hoilday

Appendix - Total GMV vs Special Sale Day : Weekly Trend

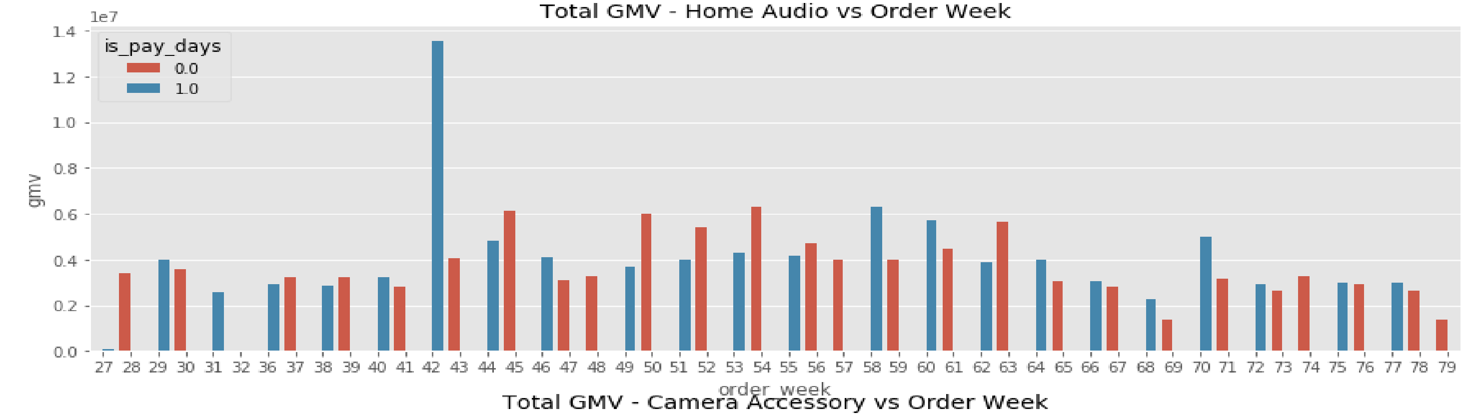


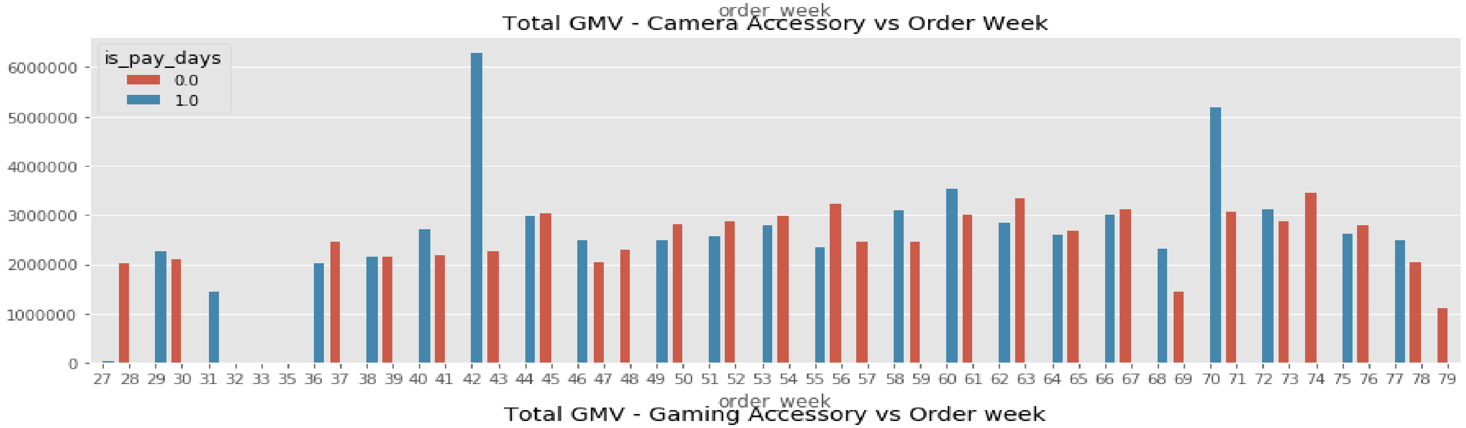


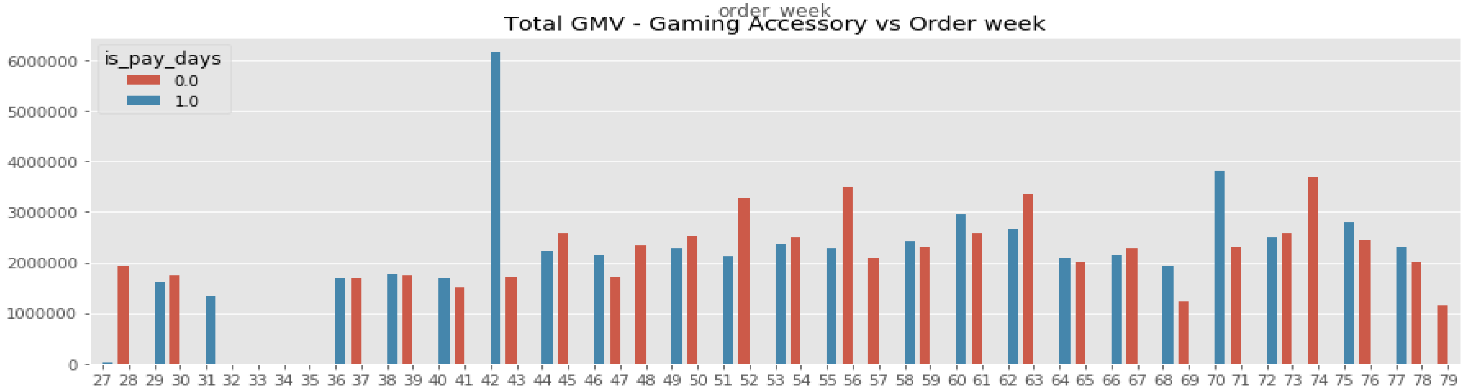


Special Sale day seems to be an important feature across all the three categories as the sales in the week when there was a sale day are pretty higher as compared to the adjacent weeks in which sale was not there

Appendix - Total GMV vs Pay Day : Weekly Trend







Pay day also seems to be an important feature as a clear trend can be seen across all the three categories i.e. the week with a Pay day is a having lesser total sales as compared to the adjacent weeks not having any pay day

Which is quite practical as people must be meeting their regular expenses in the pay day week & then going out for shopping in the following week