Approach document – ElecKart's MMM

Ecommerce Capstone Project

Aman Rai
Deva Sharma
Mukesh Tibrewala
Utkarsh Kant

Document Description

Approach document (also includes the **future roadmap**) contains:

- Steps followed for solving the business problem
- List of engineered KPIs
- First iteration of the basic linear model

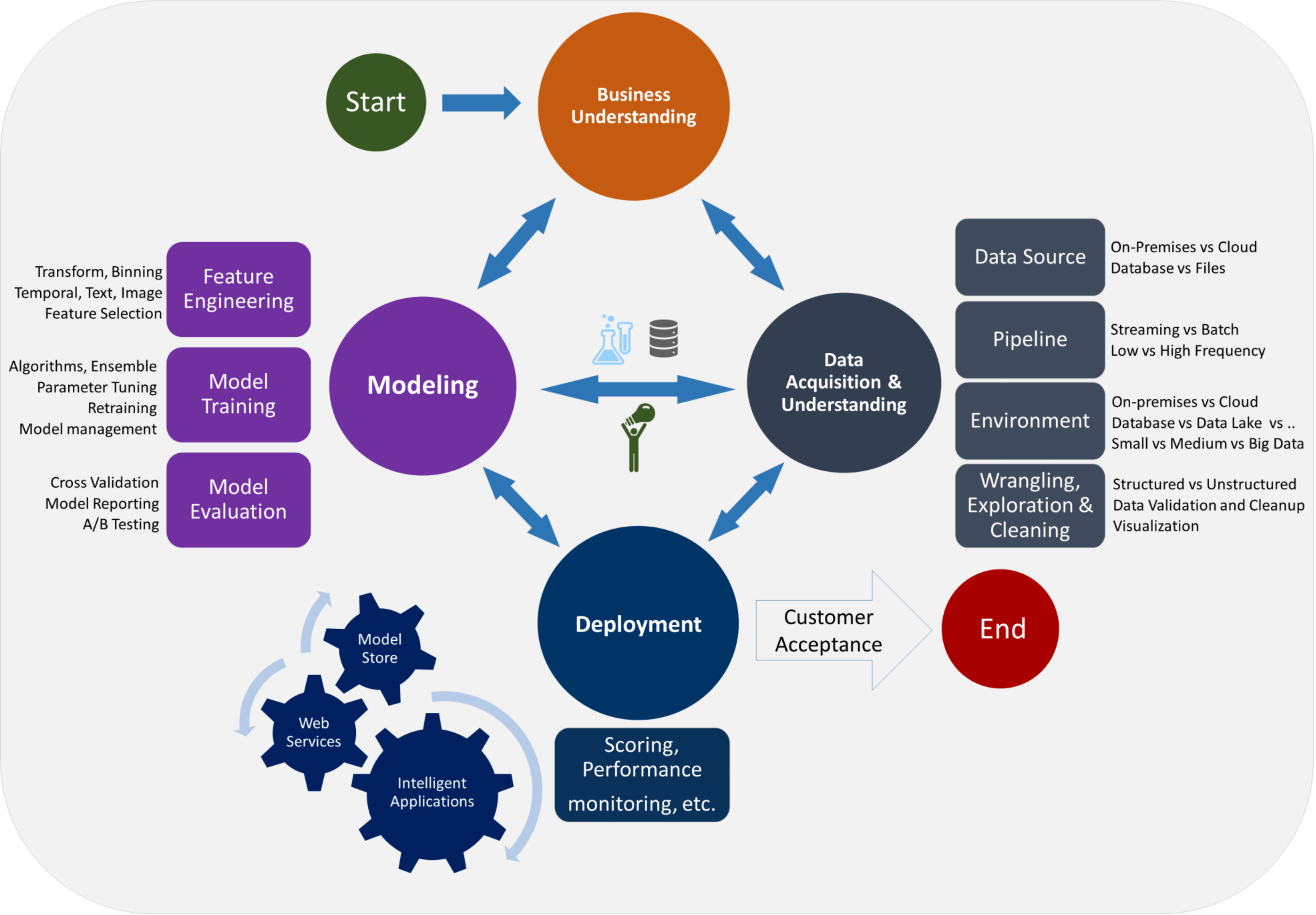
Background and Problem Statement

- ElecKart is an e-commerce firm based out of Ontario, Canada specializing in electronic products.
- They spend a significant amount of money on marketing. Occasionally, they also offer big-ticket promotions (similar to the Big Billion Day).
- They are about to create a marketing budget for the next year, which includes spending on commercials, online campaigns, and pricing & promotion strategies.
- The CFO feels that the money spent over the last 12 months on marketing was not impactful, and they can either cut on the budget or reallocate it optimally across marketing levers to improve the revenue response.
- We are a part of the marketing team working on budget optimization. We need to develop a market mix model to observe the actual impact of different marketing variables over the last year. Using our understanding of the model, we have to recommend the optimal budget allocation for different marketing levers for the next year.

Data Science Process

The overall process is divided into:

- Business Understanding
- Data Acquisition, Preparation & Understanding
- Feature Engineering (Adding KPIs)
- Exploratory Data Analysis
- Model Building (Simple & Complex Models)
- Model Deployment
- Customer Acceptance



Scope of Project

Develop understanding of EllecKart's business Model, Understand the dataset and Information given

Preparing the dataset after cleaning and merging extra information given in excel

Create additional variables, merging data week wise, making assumptions

Undertake EDA, Perform Univariates, Bivariate, Correlation, Cross-tabs and visualizations on the dataset

Modelling: Build basic, logarithmic and multiplicative models etc

Presentation of Results: After choosing best models for each of the three product sub-categories, we need to explain them to the CMO, CFO etc

Overall Project Roadmap (Gantt Chart)

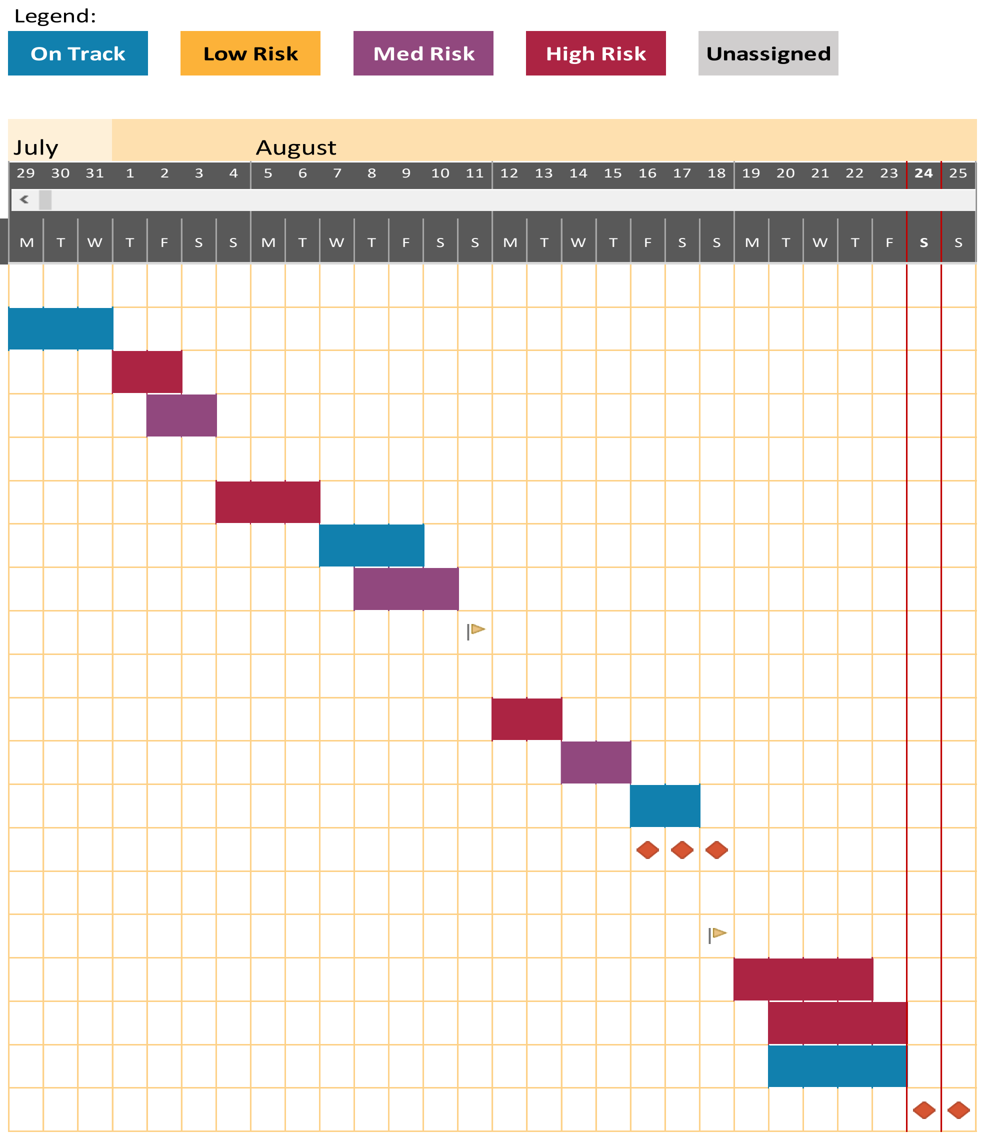
ElecKart's MMM

ElecKart
Marketing Team

Start Date: 7/29/2019

Increment: 0

Milestone Description	Category	Progress	Start	No. Days
Overview				
Understanding of case study, MMM & ElecKart's business model	On Track	100%	7/29/2019	3
Reading the dataset and drawing inferences	High Risk	100%	8/1/2019	2
Understanding information given in excel	Med Risk	100%	8/2/2019	2
Data Preparation				
Cleaning of Dataset (Column by Column) and removing null values	High Risk	100%	8/4/2019	3
Merging info given in excel with DS: Sale calendar, Payday	On Track	100%	8/7/2019	3
Reading additional info like Climate data, Media Investments	Med Risk	100%	8/8/2019	3
Call 1 with Mentor (Case study, MMM and clearing of doubts)	Milestone	100%	8/11/2019	1
Pre Analysis				
Grouping data weekwise	High Risk	100%	8/12/2019	2
Feature engineering i.e. creation of Discount column etc	Med Risk	100%	8/14/2019	2
Making of Adstock	On Track	100%	8/16/2019	2
Segregating final dataset into 3 parts basis product category	Goal	100%	8/16/2019	3
EDA				
Call 2 with Mentor (Clearing of doubts)	Milestone	100%	8/18/2019	1
Undertaking EDA, Perform Univariates, Bivariates analysis	High Risk	100%	8/19/2019	4
Correlation, Cross-tabs and visualizations on the dataset	High Risk	100%	8/20/2019	4
Creation of Basic Linear Model	On Track	100%	8/20/2019	4
Creation & Submission of Approach Note	Goal	100%	8/24/2019	3



Overall Project Roadmap (Gantt Chart)

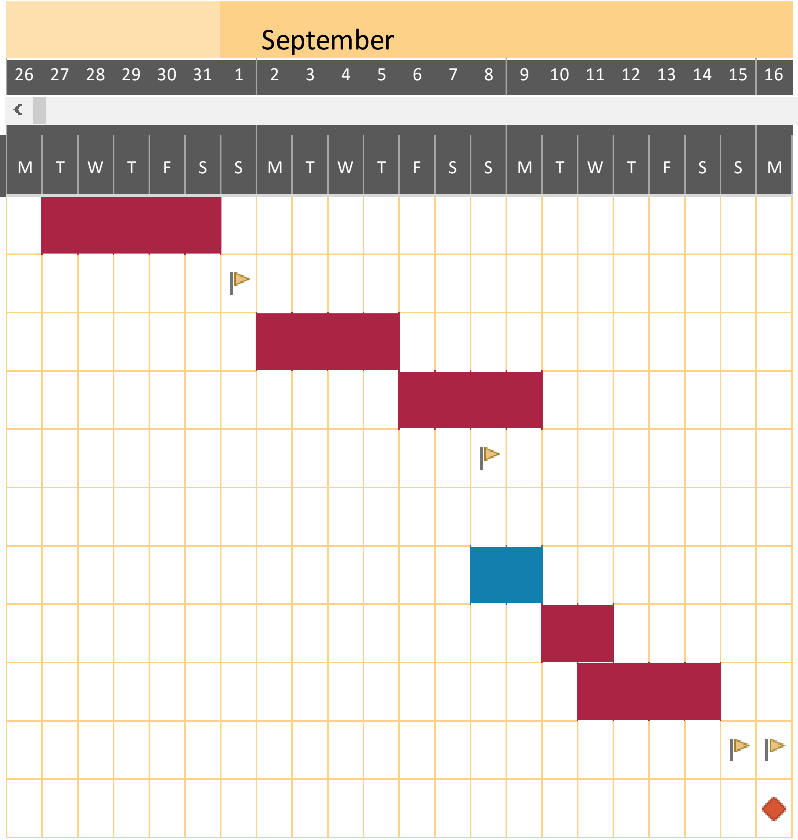
ElecKart's MMM

ElecKart
Marketing Team

Start Date: 7/29/2019

Increment: 0

Milestone Description	Category	Progress	Start	No. Days
Build Basic Model	High Risk	10%	8/27/2019	5
Call 3 with Mentor	Milestone	0%	9/1/2019	1
Build logarithmic Model	High Risk	0%	9/2/2019	4
Build Multiplicative Model	High Risk	0%	9/6/2019	4
Call 4 with Mentor	Milestone	0%	9/8/2019	1
Model Evaluation				
Evaluating Various Models and then Finalising	On Track	0%	9/8/2019	2
Analyse impact of attributes on the target variable through	High Risk	0%	9/10/2019	2
Preparation of PPT with graphs showing imp of each variable	High Risk	0%	9/11/2019	4
Finalising the Coding and PPT	Milestone	0%	9/15/2019	2
Submission of Project	Goal	0%	9/16/2019	1



Data Wrangling (Data Cleaning)

Data Wrangling

The overall process is divided into:

- **Treating Duplicates**
- **Treating wrong classes**
- **Threating out of scope data**
- **Validating Business Logics**
- **Removal of outliers**
- **Formation of sub category based dataset**
- **Creation of aggregated data, i.e. weekly aggregation**
- **Standardizing data prior to modelling**

1. Removed duplicate values
2. Checked column wise unique value
3. Removed non numeric data from gmv & pin code column ('space')
4. Removed '\\N' from various columns
5. Filtered out data which does not fall within the timelines of this analysis i.e. outside of 1st July 2015 to 30th June 2016
6. Converted order_date and gvm, pin code to proper data format
7. Removed rows with negative product MRP; GMV and units
8. Removed rows where (product_mrp*unit) < GMV
9. Removed deliverybdays and deliverycdays (more than 70% null)
10. Computed discount_percentage for each transaction
11. Categorized items into Luxury (priced more the 80 percentile) and Mass Market (priced less than 20 percentile)
12. Removed Columns which will not be used in analysis
13. Removed outliers less than Q1-1.5*IQR or greater than Q3+1.5*IQR
14. Stored the total GMV proportion for each of the 3 categories w.r.t. the total GMV for all items
15. Filtered and keeping only the 3 required categories
16. Created weeks from the date data.
17. Retained around 87.616 % of our initial data rows in data cleaning process (lost 12.384%)

Feature Engineering (Adding KPIs)

Fine-tuning Existing Features

- Dummying ***Payment Type***
  - Considering 0 for COD and 1 for Prepaid payment
- Deriving ***luxury_products***
  - Considering products with top 20% MRP as luxury products
- Deriving ***mass_products***
  - Considering products with bottom 20% MRP as mass products
- Deriving ***discount_percentage***
  - Considering 100*(MRP*UNIT – GMV)/(MRP*UNIT) as discount percentage

Adding NPS & Stock Index

- Adding ***nps***
  - Adding Net Promoter Score from the given additional data
  - Merging it using year and month to our original dataset
- Adding ***stock_index***
  - Adding Stock Index from the given additional data
  - Merging it using year and month to our original dataset

Adding Pay Day, Sales Day Information

- Adding ***is_pay_day***
  - Adding Pay Day information from the given additional data
  - Merging it using date of each month to our original dataset
- Adding ***is_sale_day***
  - Adding Sale Day information from the given additional data
  - Merging it using specific date to our original dataset

Adding Weather Information

- Weather Data is linearly interpolated to account for missing values.
- Adding ***is_rainy***
  - Adding Rainy Day information from the given additional data
  - 0 for no rain and 1 for any rain on given date
- Adding ***is_hot***
  - Adding Hot Day information from the given additional data
  - 0 for normal days and 1 if mean temperature is above 25-degree Celsius
- Adding ***is_snow_on_ground***
  - Adding Snow on Ground Day information from the given additional data
  - 0 for normal days and 1 if any snow present on ground

Adding Investment Information (Ad-Stock)

- Prepared Ad-Stock by converting monthly data to daily data and then aggregating on weekly basis.
- We have excluded Radio and Other from our Ad-Stock computation since the values were comparatively very low and rare. Also less / no actions could be taken considering them.
- Adding ***'Total Investment', 'TV', 'Digital', 'Sponsorship', 'Content Marketing', 'Online marketing', ' Affiliates', 'SEM'***
  - Ad-Stock computation is done directly in python at weekly level
  - Ad-Stock rate is taken as **0.5**
  - All the individual investments are added to sub category specific dataset based on the contribution of the sub category in GMV.
    Example: Camera Accessory contributes 0.142119 or 14.2119% share in GMV
    Thus we would consider 14.2119% of all investments for Camera Accessory.

	CameraAccessory	HomeAudio	GamingAccessory
Record Share	0.142119	0.073572	0.123721
GMV Share	0.067691	0.064318	0.041553

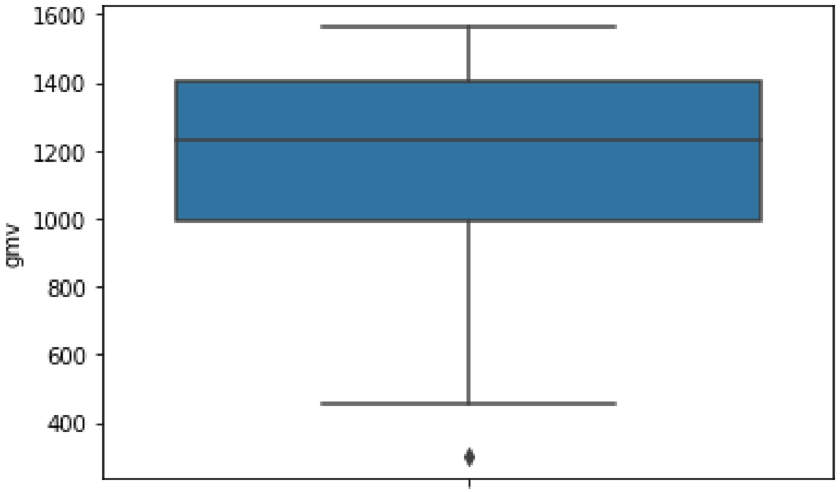
Actual Computed Ad-Stock (Weekly level)

week_of_year	Total Investment	TV	Digital	Sponsorship	Content Marketing	Online marketing	Affiliates	SEM
27	2.751899	0.034731	0.408551	1.19585	0.00015	0.214077	0.088267	0.810274
28	3.852659	0.048623	0.571971	1.67419	0.000211	0.299708	0.123573	1.134383
29	3.852659	0.048623	0.571971	1.67419	0.000211	0.299708	0.123573	1.134383
30	3.852659	0.048623	0.571971	1.67419	0.000211	0.299708	0.123573	1.134383
31	3.078629	0.035146	0.491007	1.264452	0.000151	0.222415	0.093021	0.972437
32	1.143553	0.001454	0.288597	0.240107	0.000001	0.029184	0.016638	0.567571
33	1.143553	0.001454	0.288597	0.240107	0.000001	0.029184	0.016638	0.567571
34	1.143553	0.001454	0.288597	0.240107	0.000001	0.029184	0.016638	0.567571
35	1.143553	0.001454	0.288597	0.240107	0.000001	0.029184	0.016638	0.567571
36	19.414241	0.776108	0.312534	12.591831	0.122059	3.280167	1.01003	1.321511
37	22.459355	0.905218	0.316523	14.650452	0.142401	3.821998	1.175595	1.447168
38	22.459355	0.905218	0.316523	14.650452	0.142401	3.821998	1.175595	1.447168
39	22.459355	0.905218	0.316523	14.650452	0.142401	3.821998	1.175595	1.447168
40	31.581089	1.180816	1.76436	17.204253	0.505426	4.782745	1.40366	4.739829
41	38.42239	1.387515	2.850237	19.119604	0.777694	5.503305	1.574709	7.209325
42	38.42239	1.387515	2.850237	19.119604	0.777694	5.503305	1.574709	7.209325
43	38.42239	1.387515	2.850237	19.119604	0.777694	5.503305	1.574709	7.209325
44	34.640684	1.329987	2.485576	16.860636	0.672216	5.369171	1.569609	6.353489
45	11.950451	0.984814	0.297609	3.306827	0.039348	4.564367	1.539012	1.218474
46	11.950451	0.984814	0.297609	3.306827	0.039348	4.564367	1.539012	1.218474
47	11.950451	0.984814	0.297609	3.306827	0.039348	4.564367	1.539012	1.218474
48	11.950451	0.984814	0.297609	3.306827	0.039348	4.564367	1.539012	1.218474
49	22.36759	1.185366	0.635424	11.447646	0.212197	5.007618	1.541202	2.338138
50	24.10378	1.218791	0.691726	12.804449	0.241005	5.081493	1.541567	2.524749
51	24.10378	1.218791	0.691726	12.804449	0.241005	5.081493	1.541567	2.524749
52	24.10378	1.218791	0.691726	12.804449	0.241005	5.081493	1.541567	2.524749
53	20.953847	1.120323	0.439401	7.72328	0.224814	5.119839	1.594121	1.849165
54	16.753935	0.989032	0.102968	0.948387	0.203226	5.170968	1.664194	0.948387
55	16.753935	0.989032	0.102968	0.948387	0.203226	5.170968	1.664194	0.948387
56	16.753935	0.989032	0.102968	0.948387	0.203226	5.170968	1.664194	0.948387

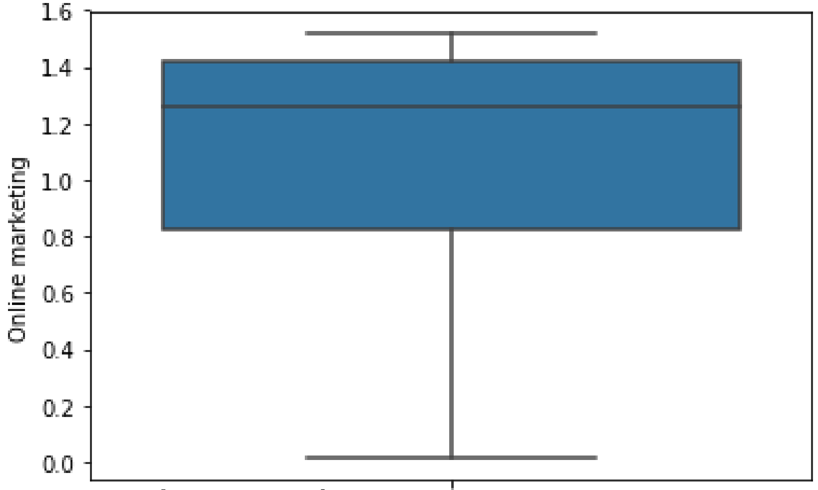
EDA

Camera Accessory

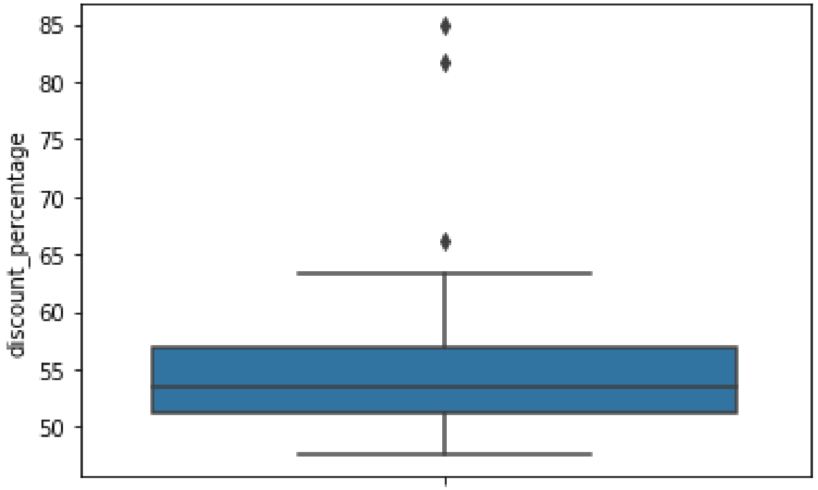
Univariate Analysis



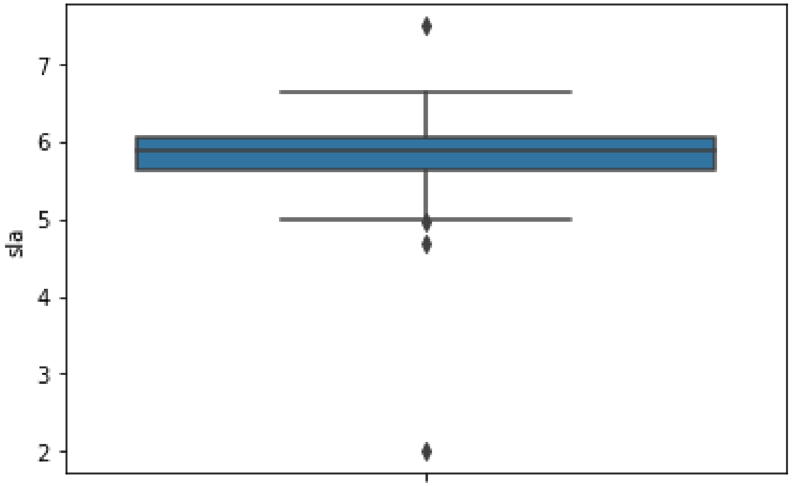
Average GMV is ~ 1225 dollars



Online Marketing investment is quite high when compared to other medias



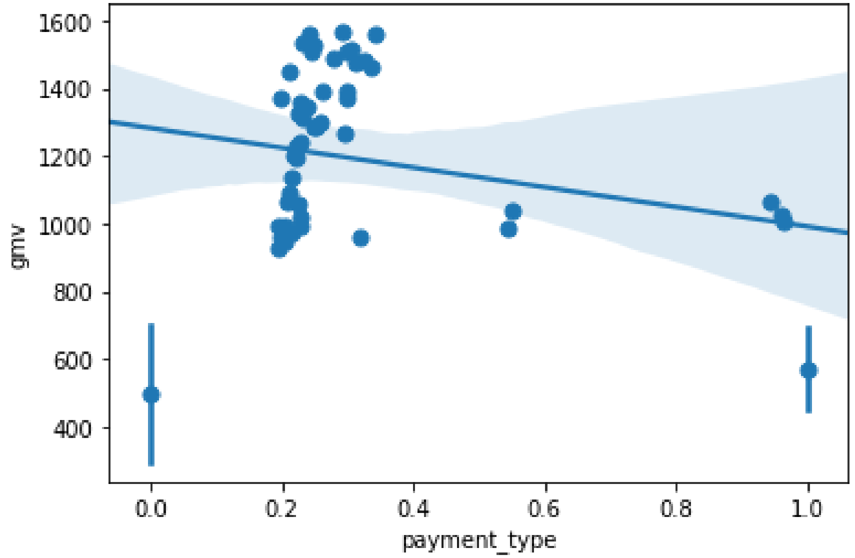
Avg discount %age is ~54%



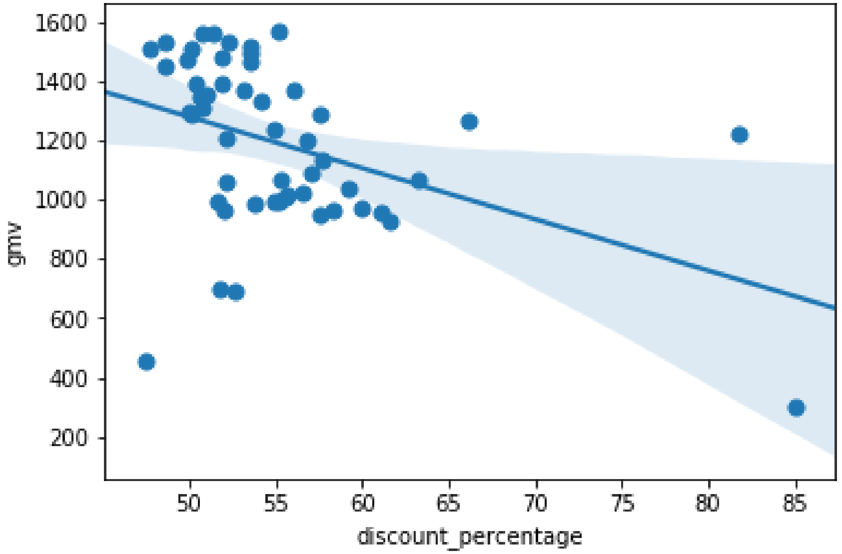
Product Delivery SLA is ~ 6 days on avg

Camera Accessory

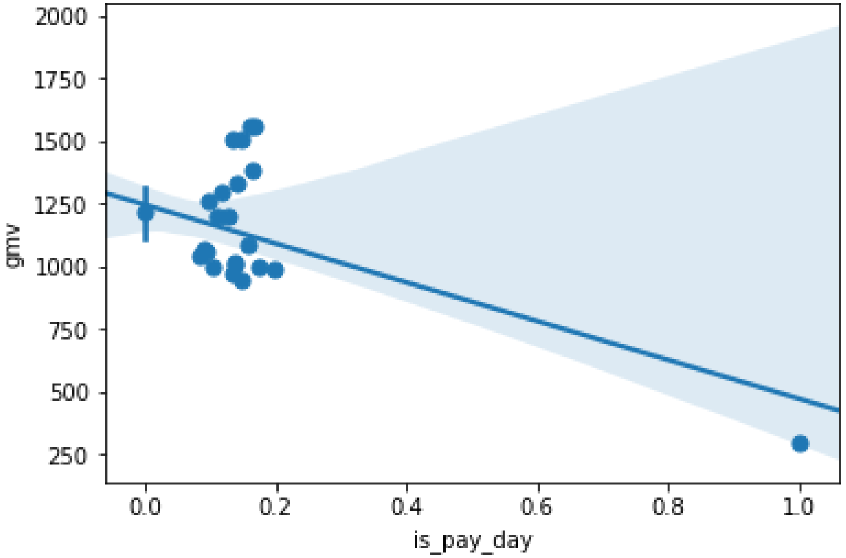
Bi Variate Analysis



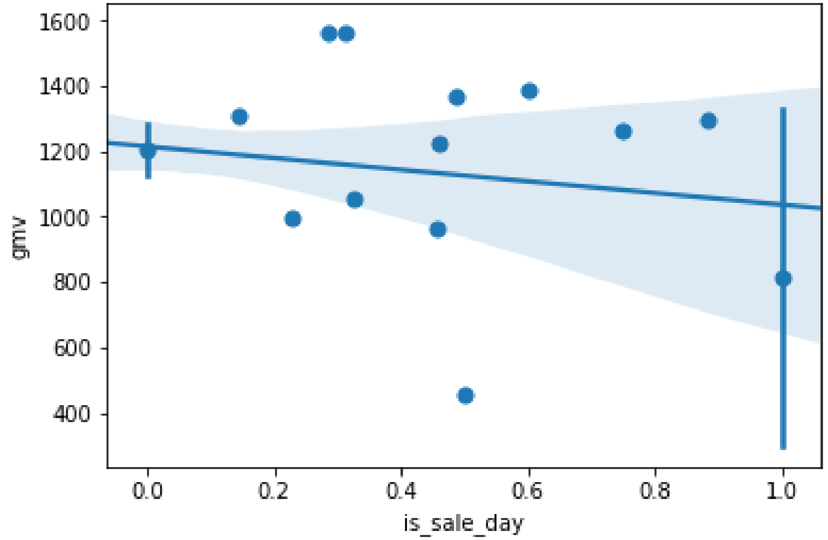
As share of Prepaid orders increase, there is slight drop in GMV



Extremely higher discounts (65%+) is leading to lower GMV



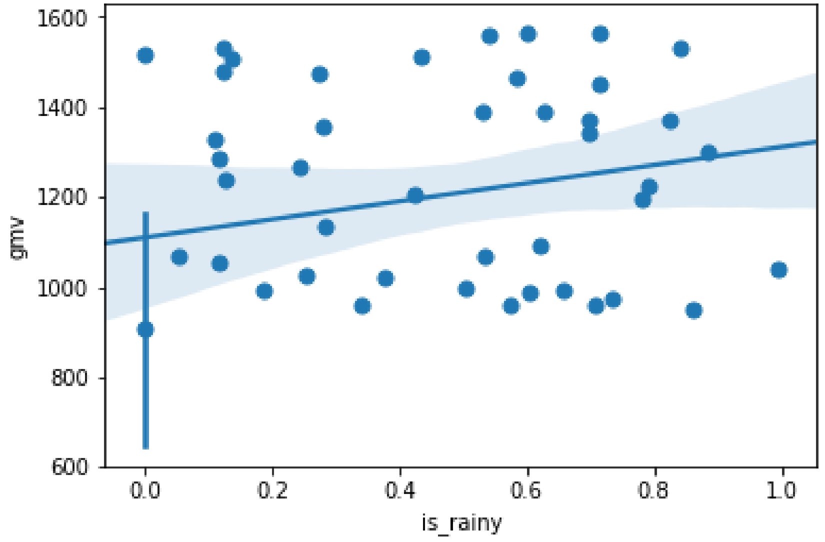
With payday approaching, the GMV is dropping, however less evident



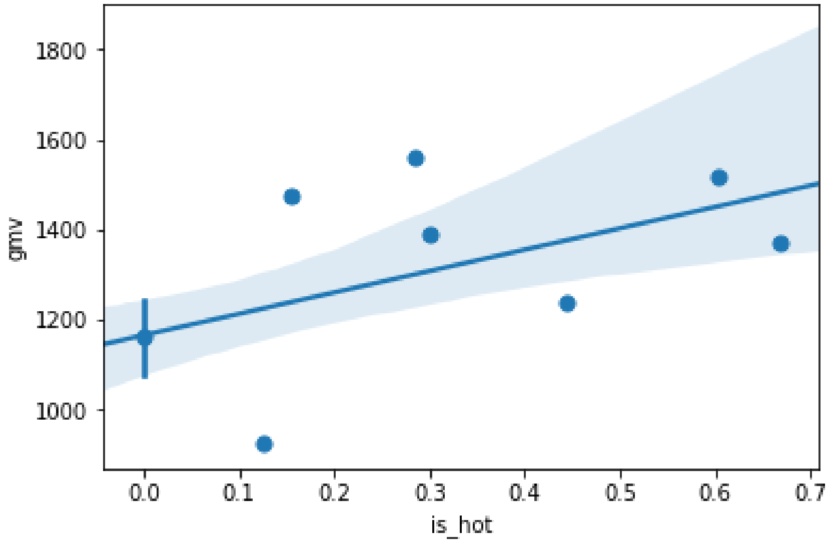
Sale Day leads to slight drop in GMV, may be due to discounting

Camera Accessory

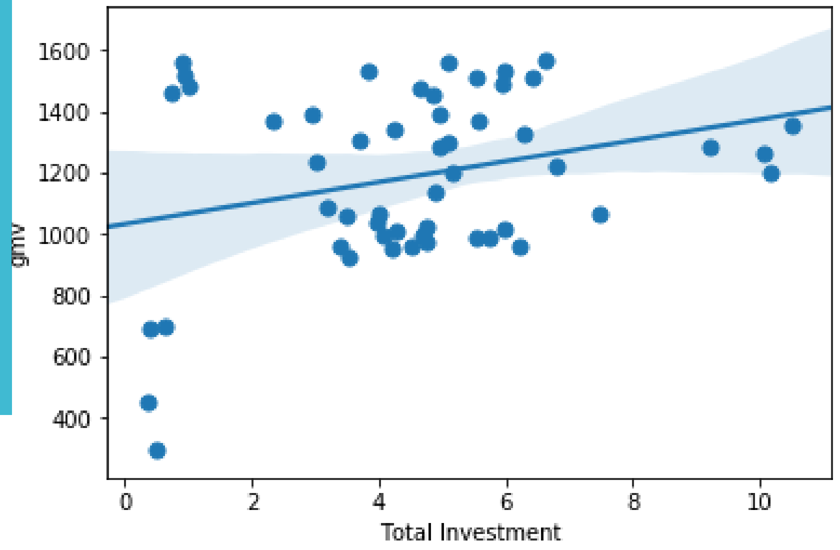
Bi Variate Analysis



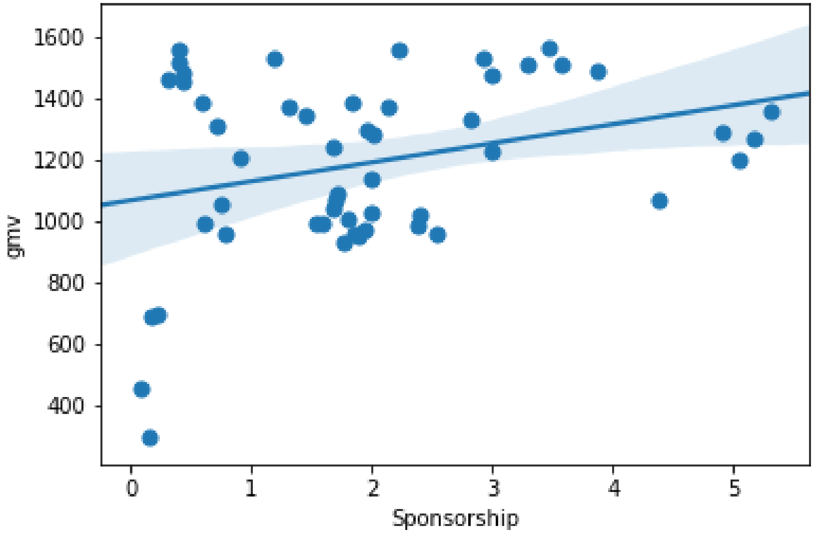
On Rainy days, there is positive impact on GMV



Canada is cold country; hot days leads to customers look for a good camera accessory hence higher GMV



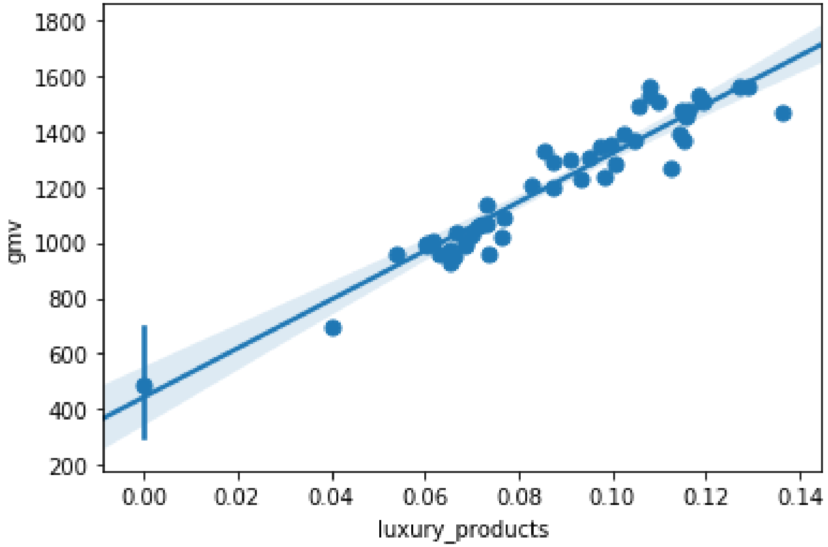
Media investments have positive correlation with GMV



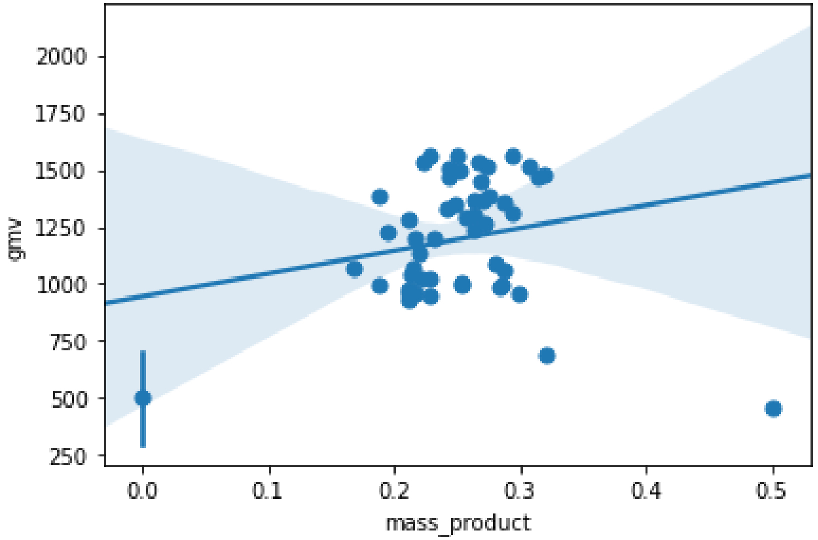
Higher sponsorship leading to higher GMV

Camera Accessory

Bi Variate Analysis



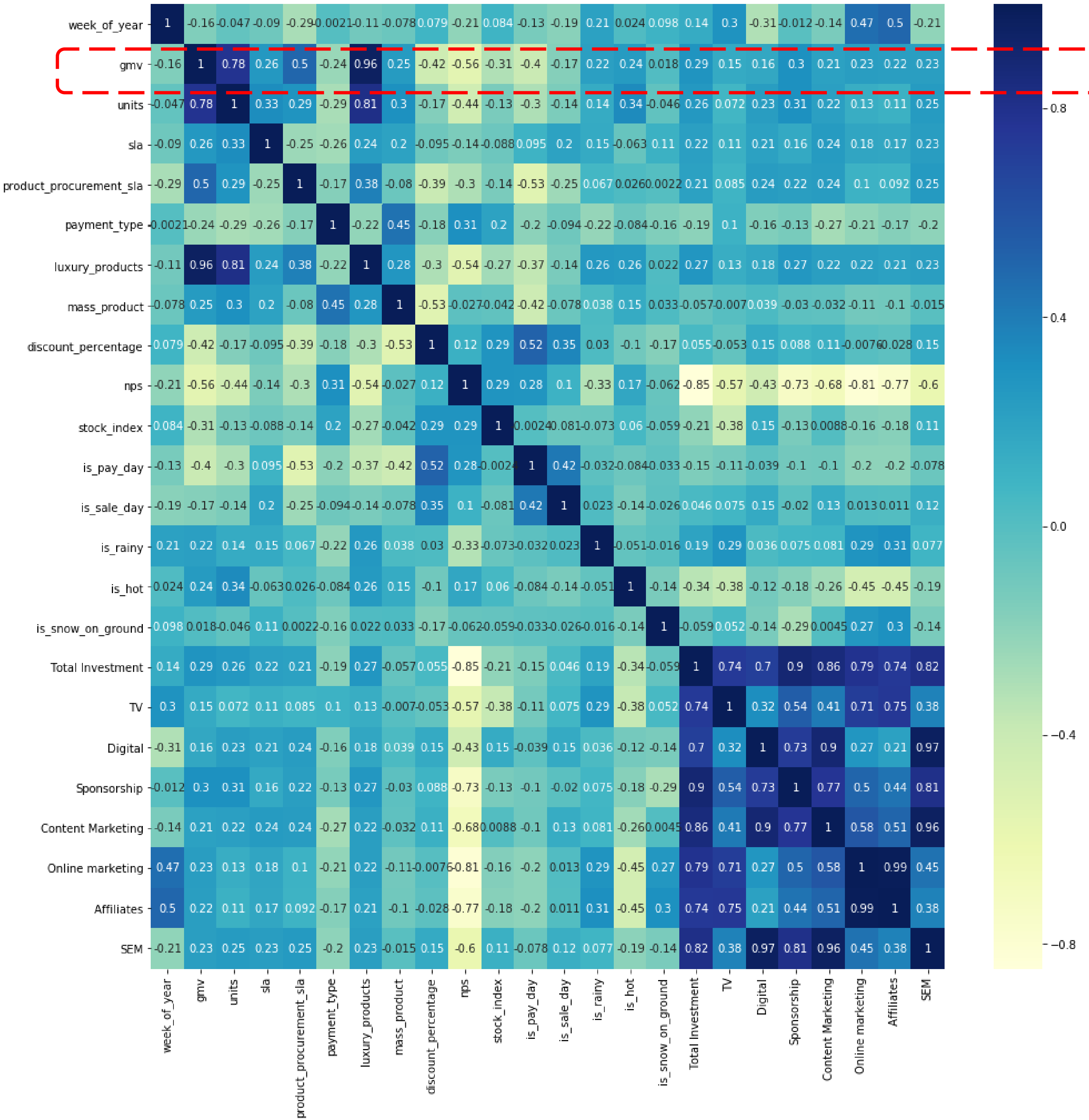
Luxury Tagged Camera accessories are leading to higher GMV



Mass Market tagged products are leading to growth in GMV

Camera Accessory

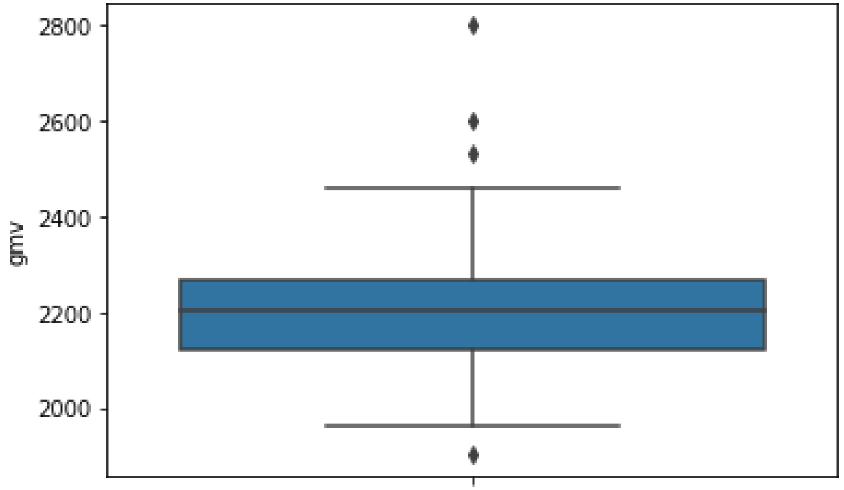
Correlation Matrix



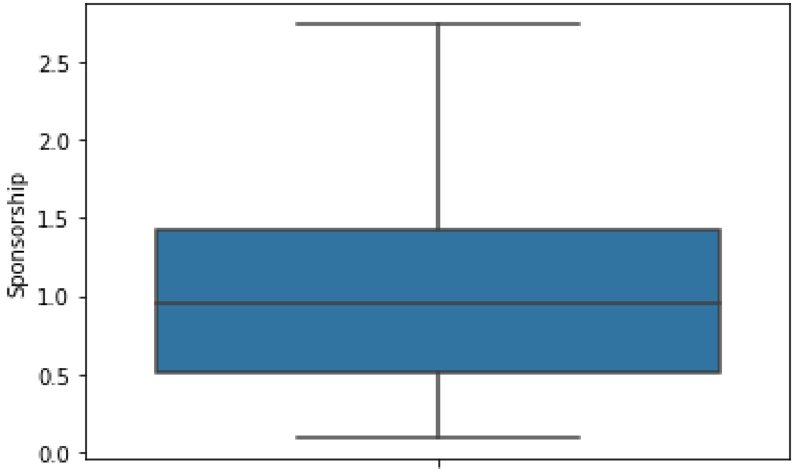
GMV has positive correlation with SLA, Luxury tagged products, weather and investments variables and negative correlation with discount and Payday

Home Audio

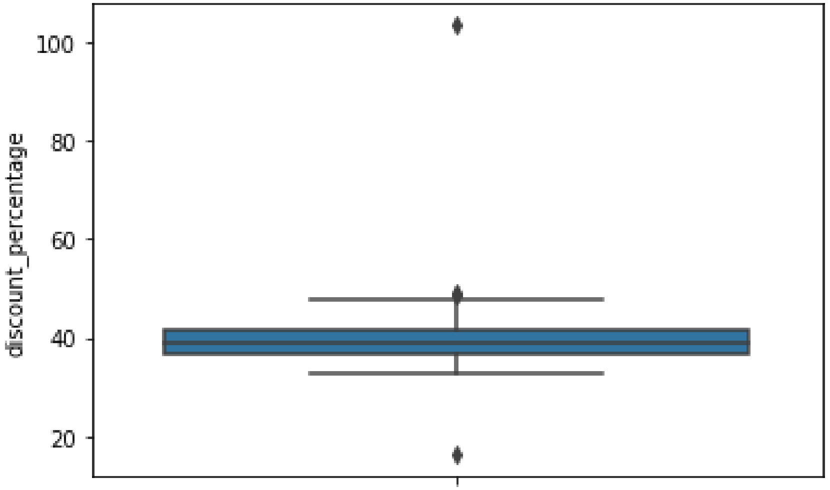
Univariate Analysis



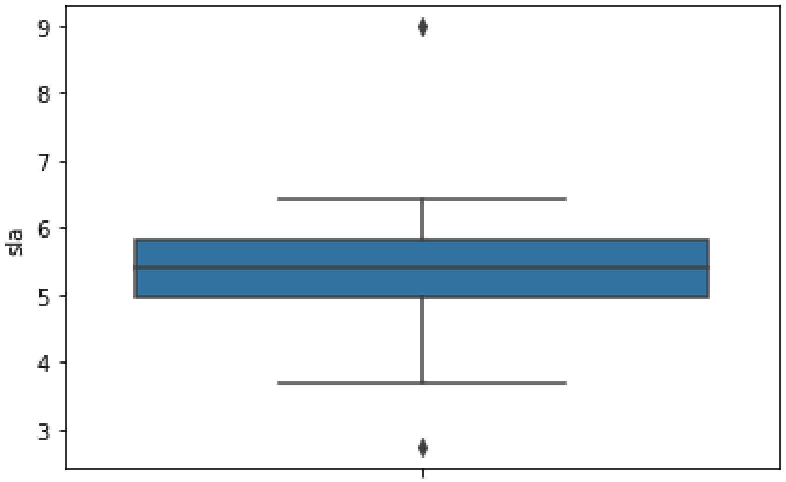
Average GMV is ~ 2200 dollars, much higher than camera accessory



Sponsorship as a media investment is quite large for Home Audio category



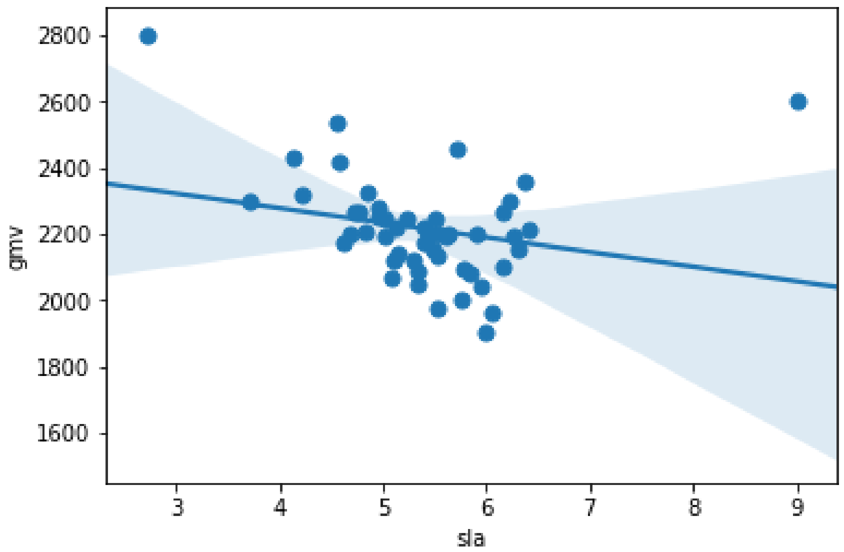
Avg discount %age is ~40%, lesser then camera accessory



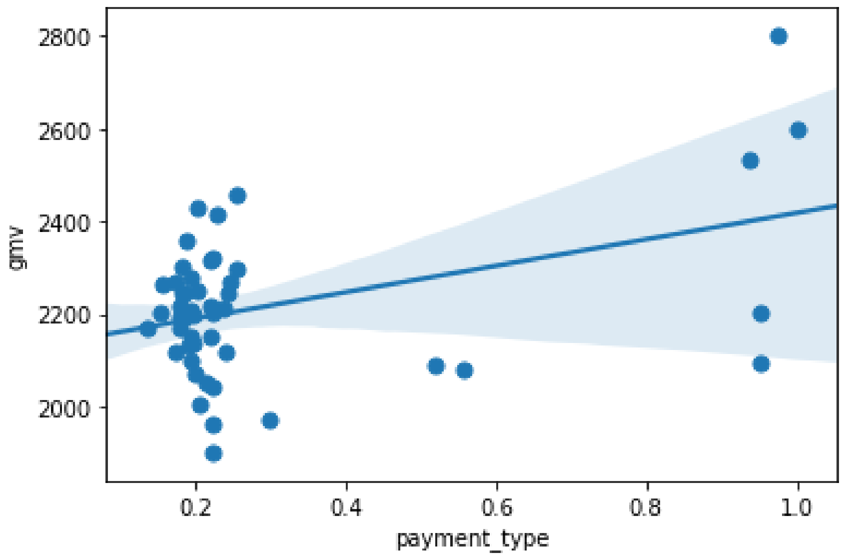
Product Delivery SLA is ~ 5.5 days on avg

Home Audio

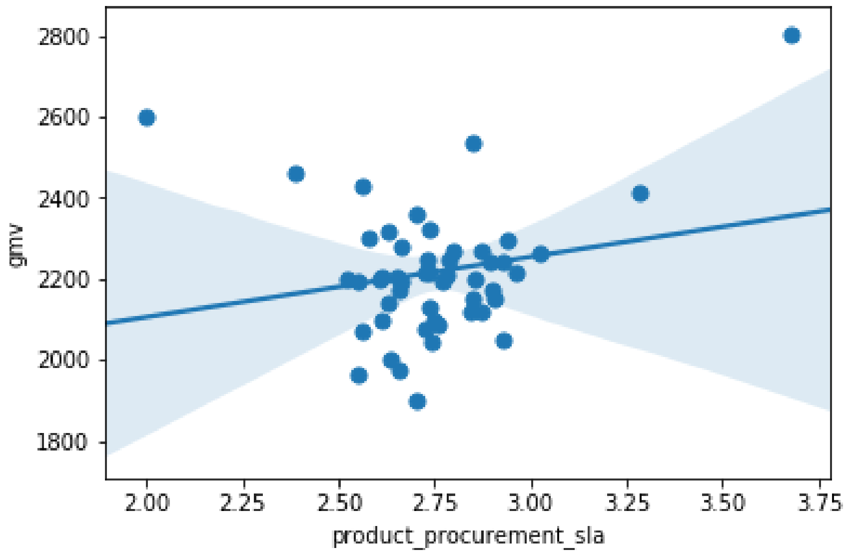
Bi Variate Analysis



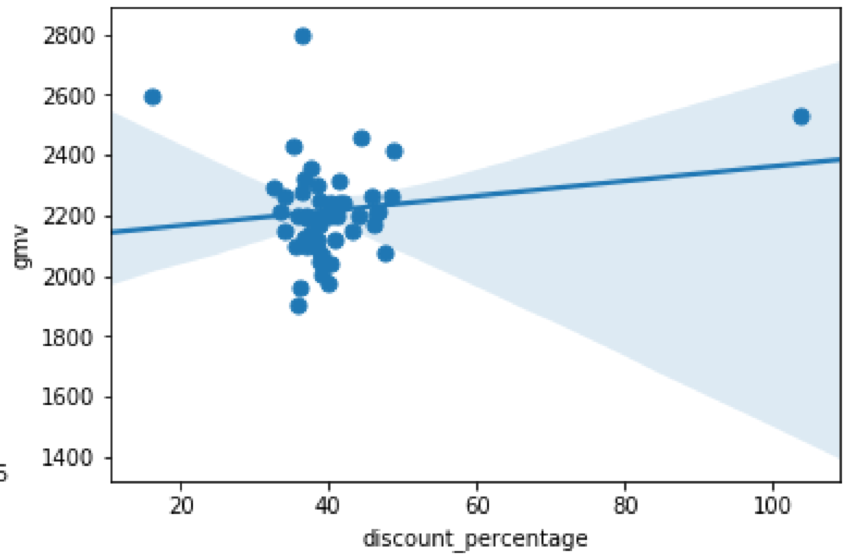
As delivery SLA increases the GMV drops



As share of prepaid orders increase, GMV rises however opposite trend was seen in camera accessory



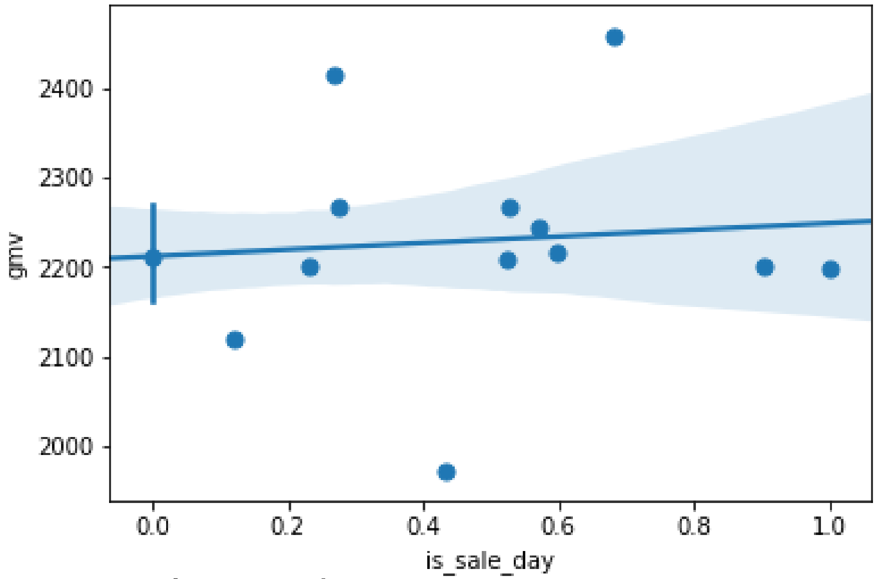
Products having higher procurement SLA have higher GMV, as higher priced products may not be stocked



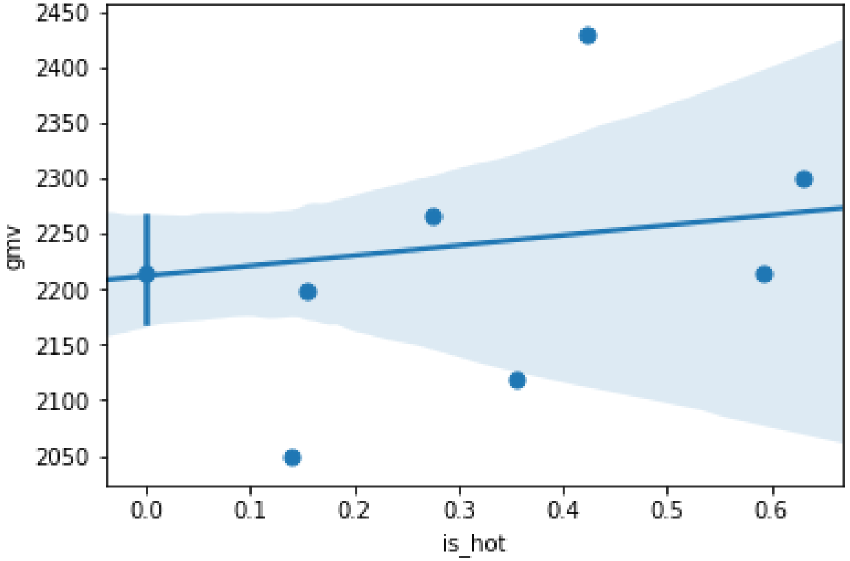
Higher discount %age is leading to higher GMV, perhaps making products lucrative to customers

Home Audio

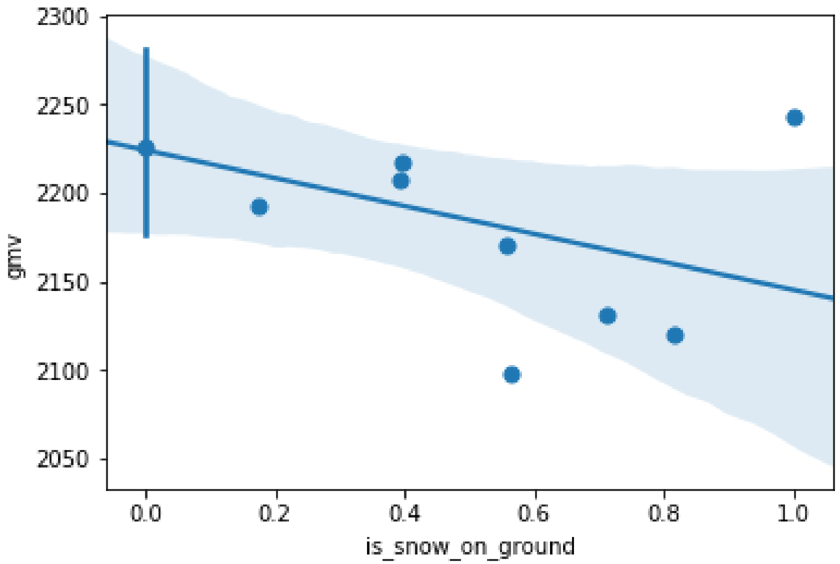
Bi Variate Analysis



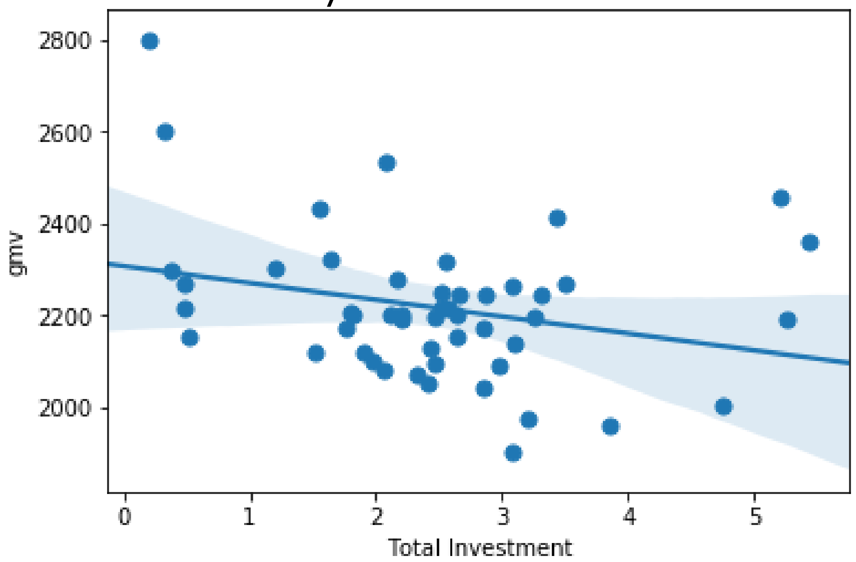
Sale Day does not impact GMV much



Again in hot days, GMV is increasing similar trend noticed in camera accessory



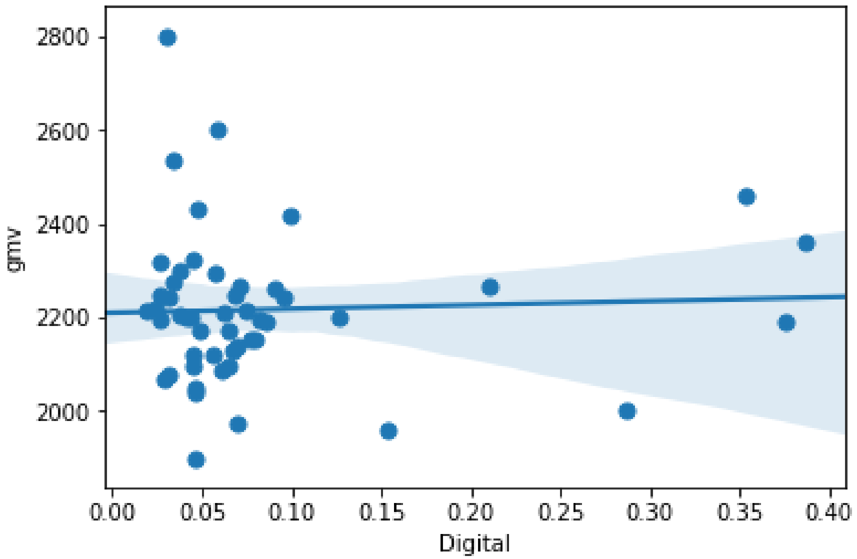
Snow is not good for Home Audio business



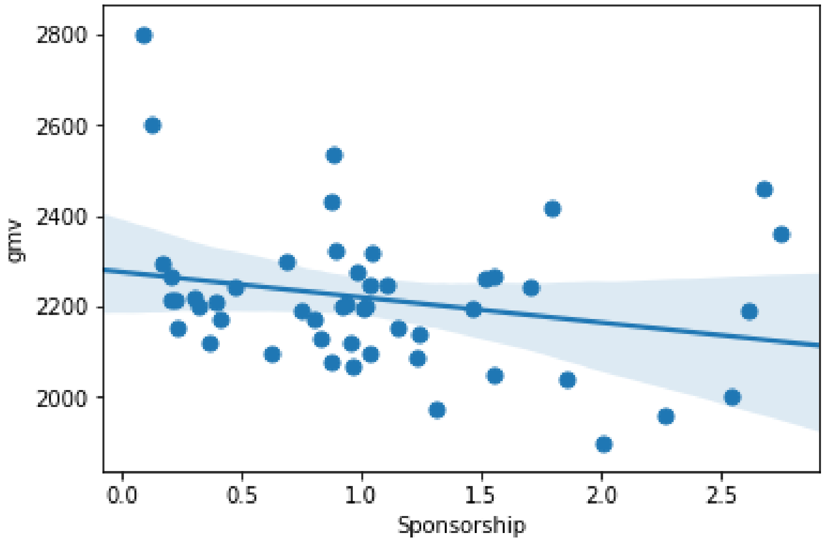
Surprising, as media investment is increasing the GMV is dropping

Home Audio

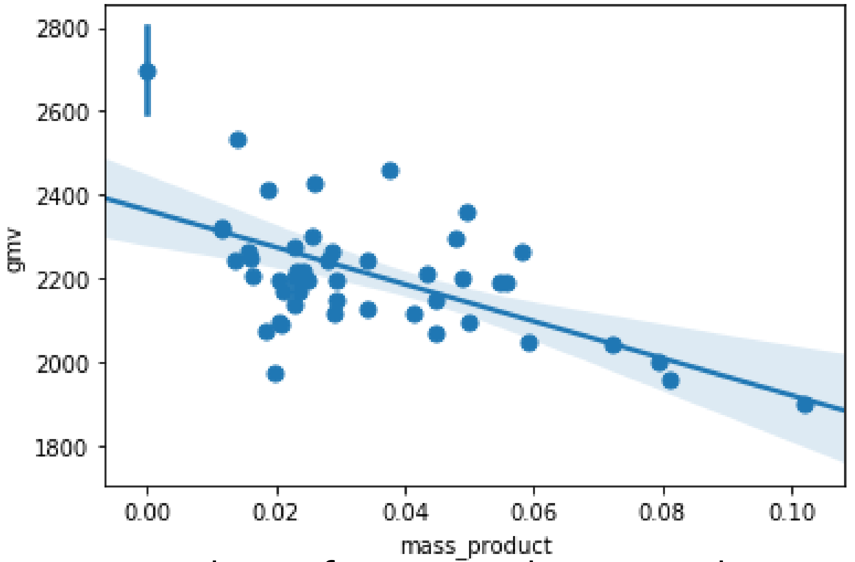
Bi Variate Analysis



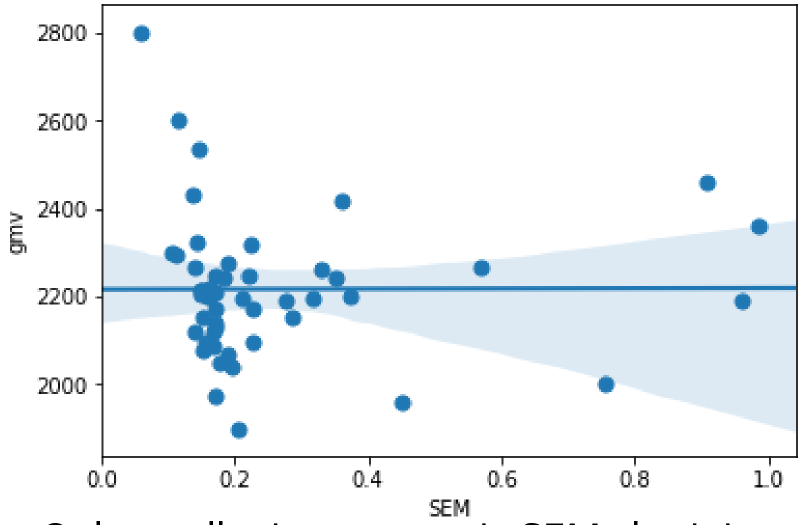
As digital investment increase the GMV is stagnant at least not dropping



Large investments in sponsorships is also not leading to growth in GMV, may be choice of properties to be rechecked



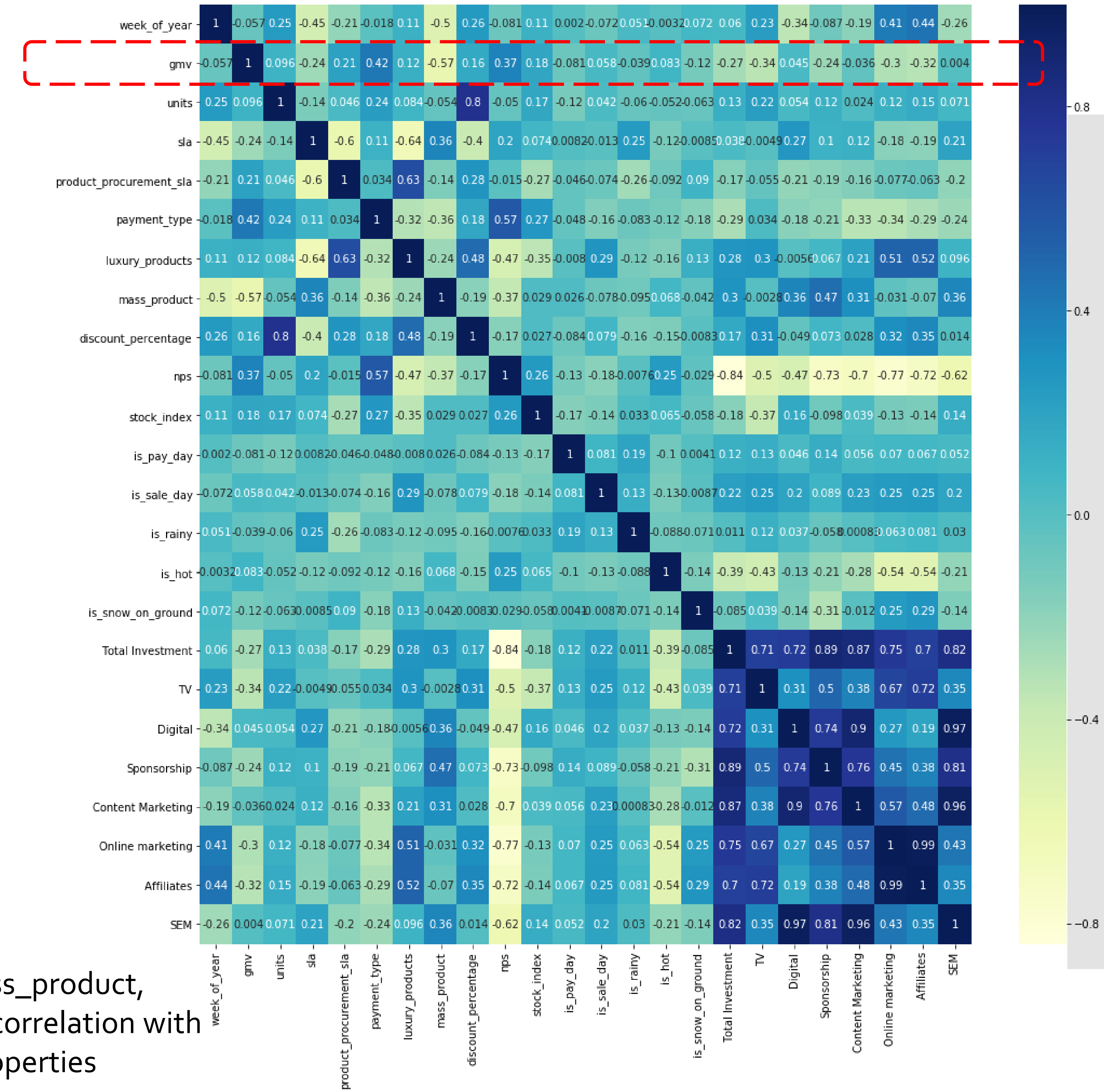
As share of mass_product tagged products increase, the GMV drops



Only smaller investment in SEM also it is not leading to drop in GMV as compared to other medias

Home Audio

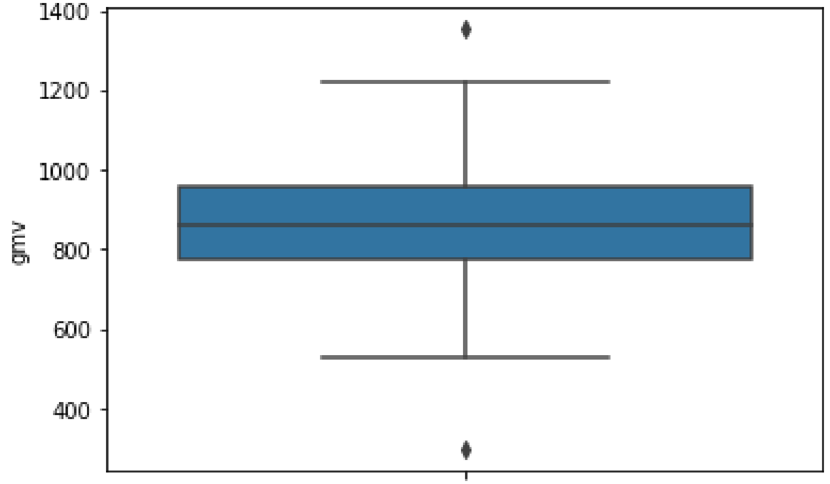
Correlation Matrix



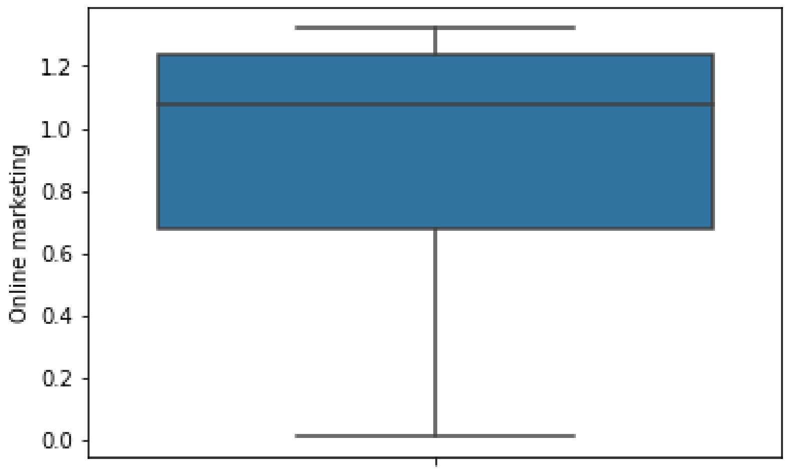
Payment type & NPS have positive correlation whereas mass_product, delivery SLA, along with media investments have negative correlation with GMV. May be we need to check campaign, media plan & properties

Gaming Accessory

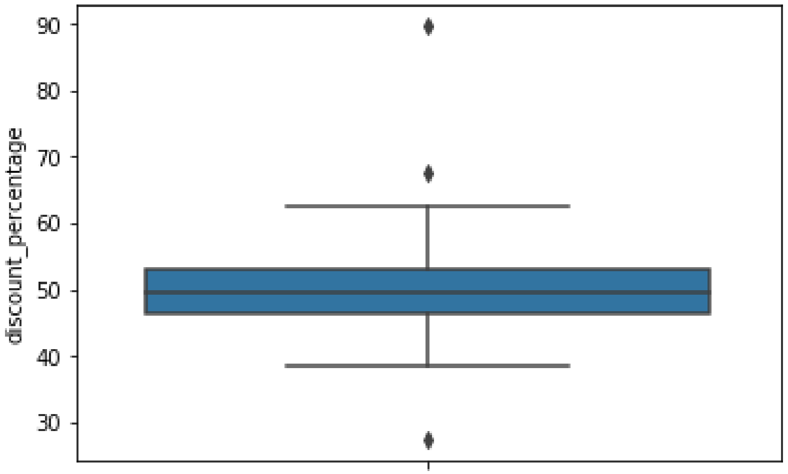
Univariate Analysis



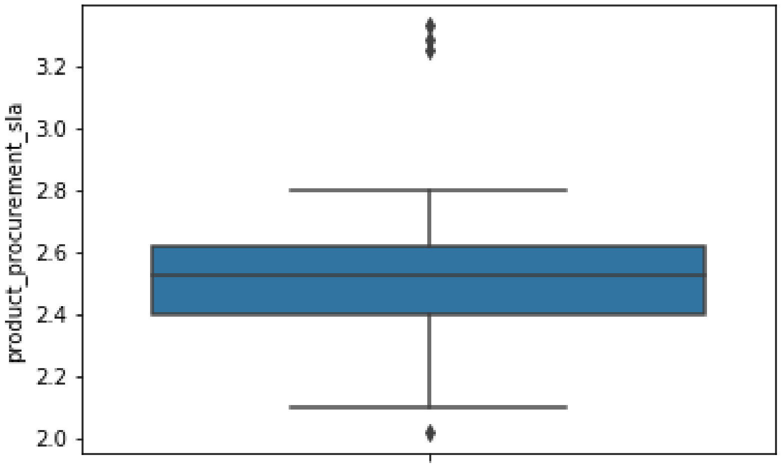
Average GMV is ~ 875 dollars, lowest as compared to other 2 categories



Online Marketing investment is quite high



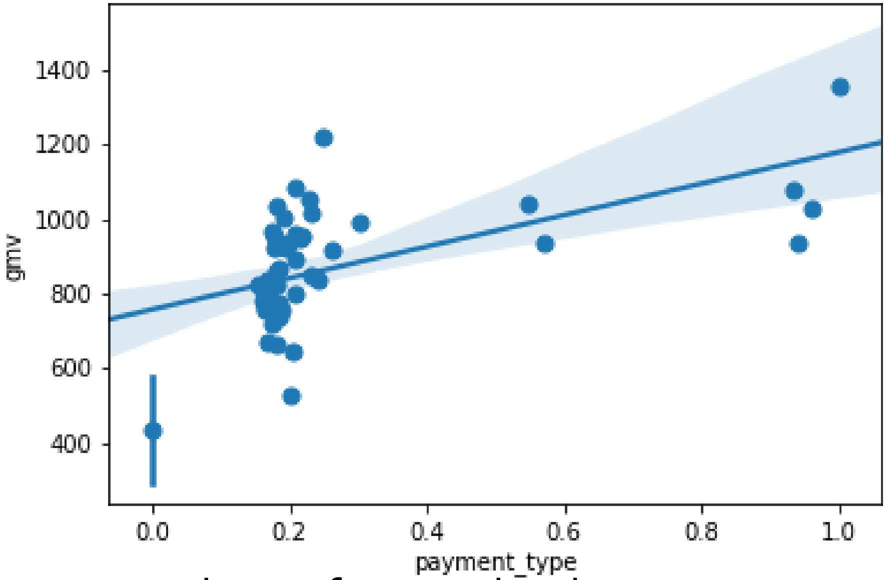
Avg discount %age is ~50%



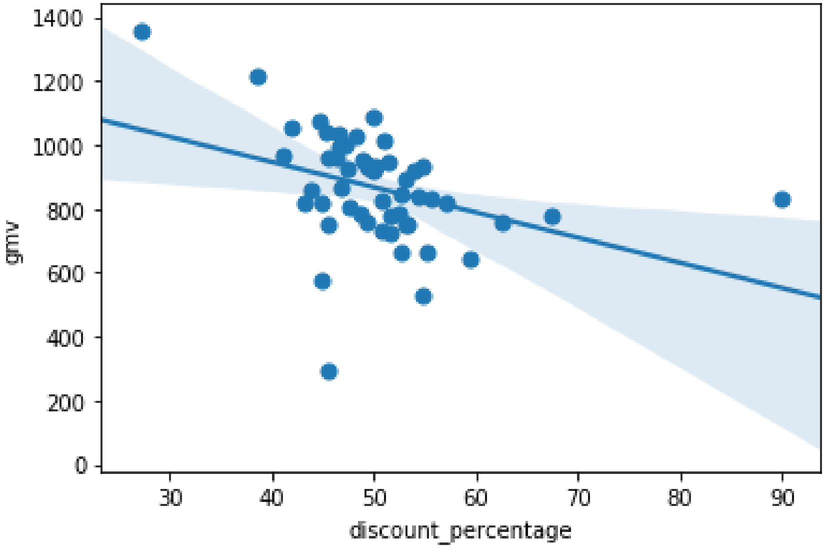
Product procurement SLA is ~ 2.5 days on avg

Gaming Accessory

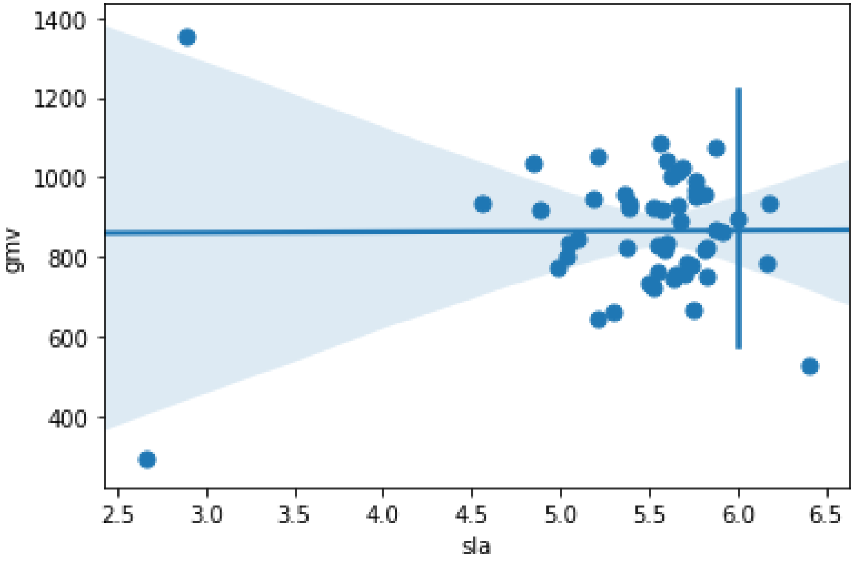
Bi Variate Analysis



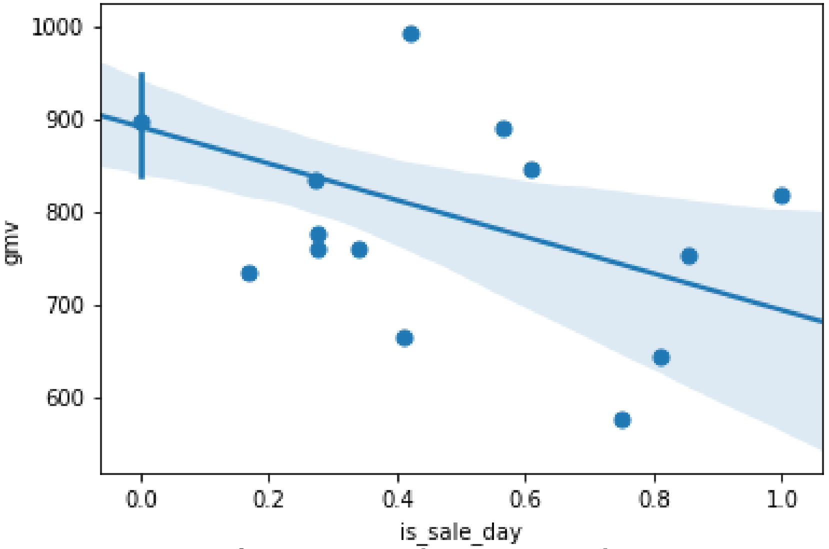
As share of Prepaid orders increase, there is gain in GMV, opposite trend noticed in camera accessory



Extremely higher discounts (50%+) is leading to lower GMV



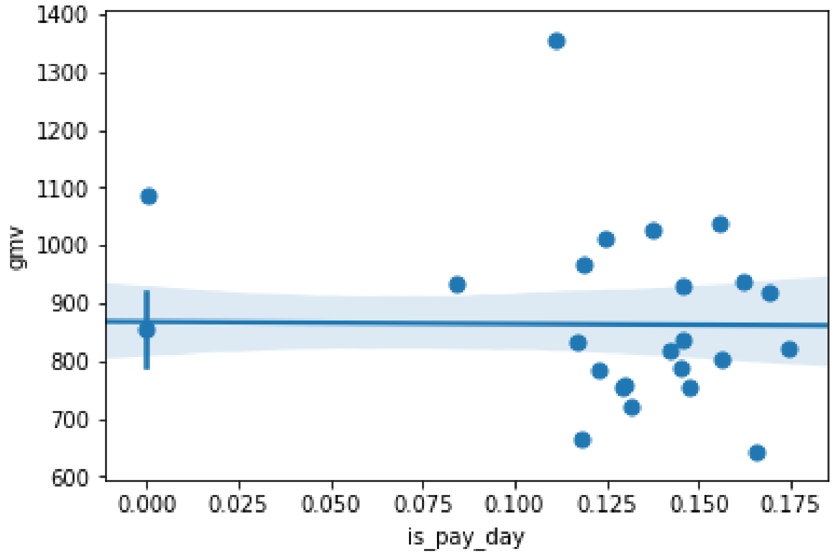
Delivery SLA has no impact on GMV



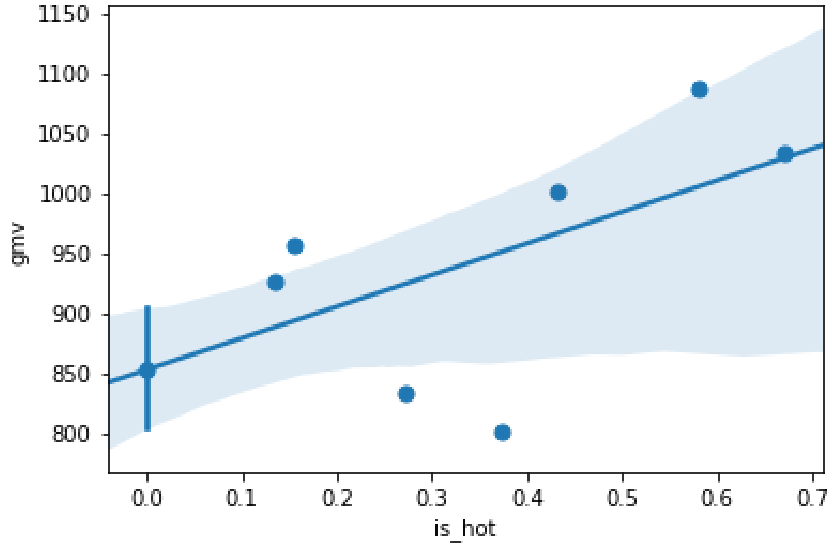
On Sale Days, the GMV drops considerably maybe due to discounting

Gaming Accessory

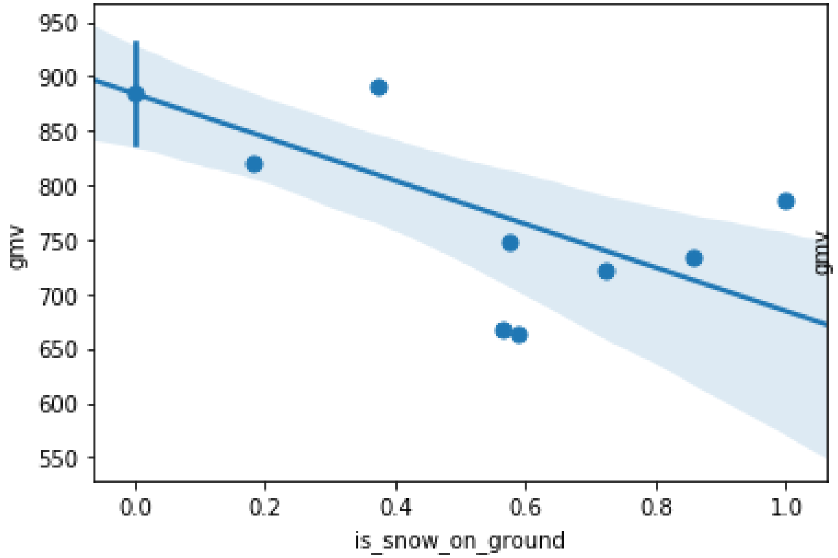
Bi Variate Analysis



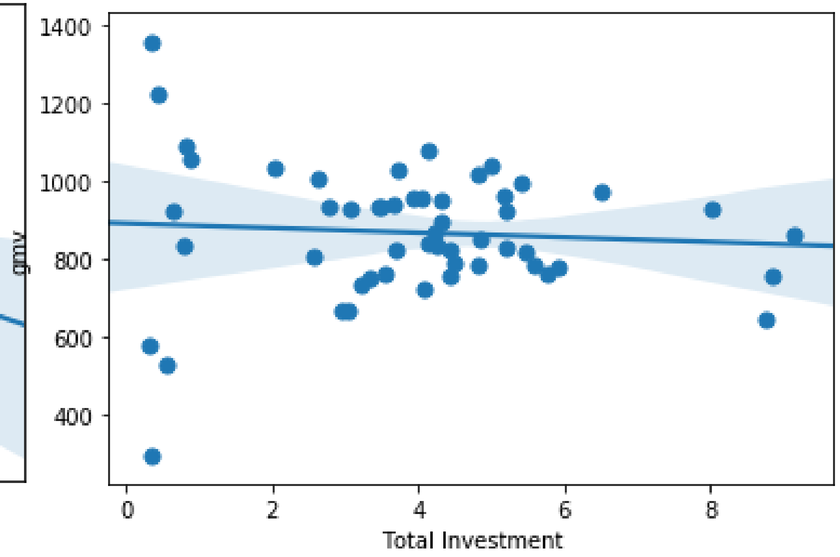
No impact of pay_day on GMV



On Hot days, again positive trend in GMV, similar trend seen in other 2 product categories too



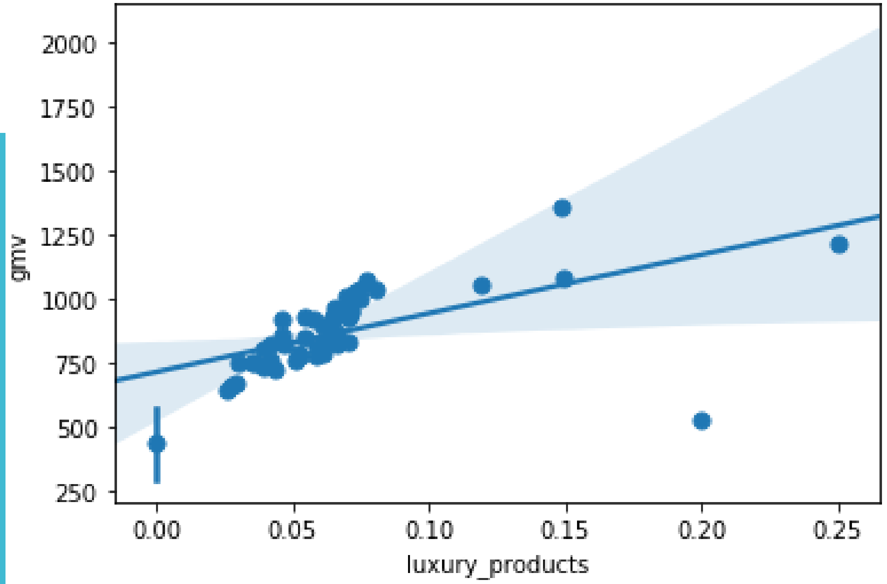
Snow is leading to drop in GMV



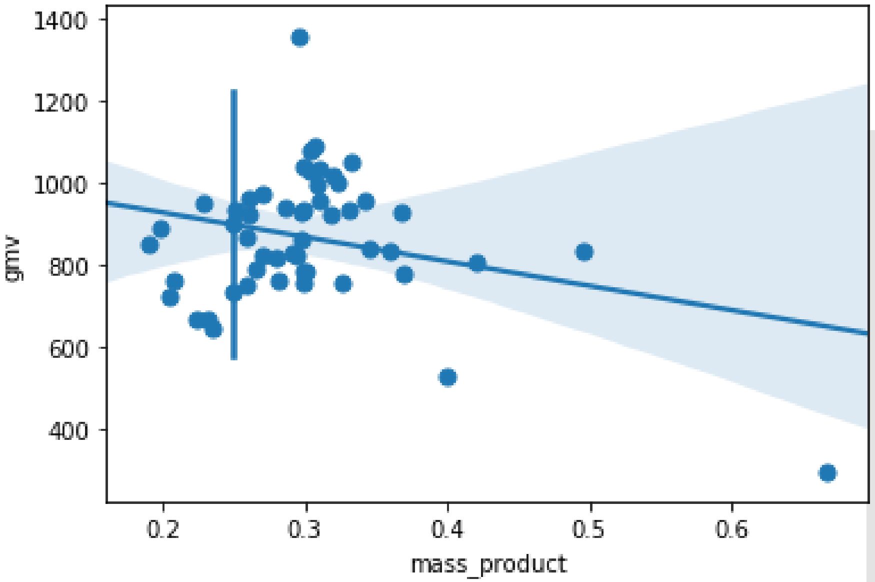
Media Investments are not showing any significant impact on GMV

Gaming Accessory

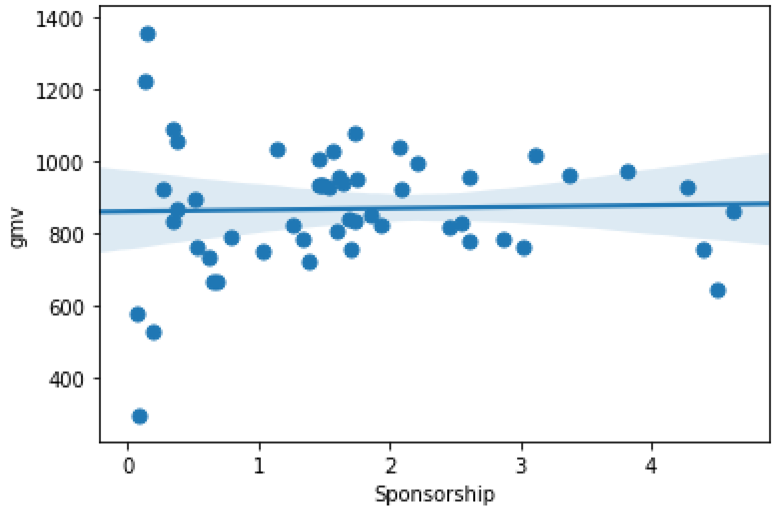
Bi Variate Analysis



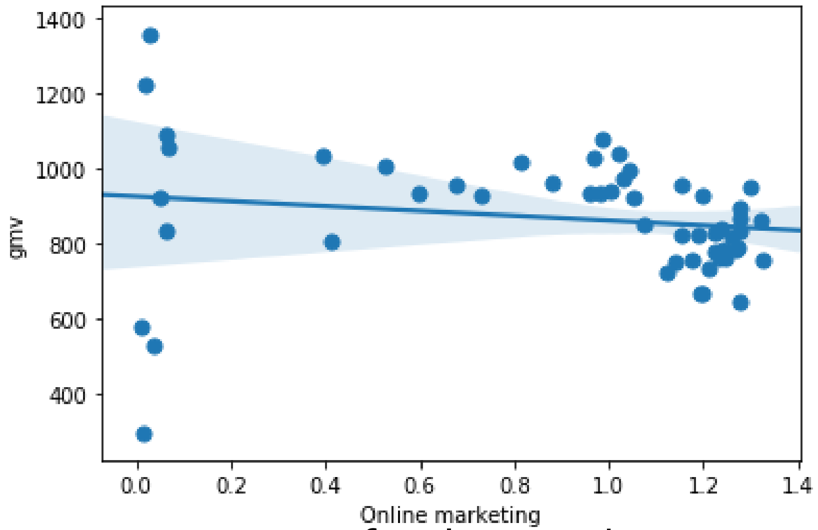
Luxury Tagged gaming accessories are leading to higher GMV



Increase in share of mass_product tagged products leading to drop in GMV



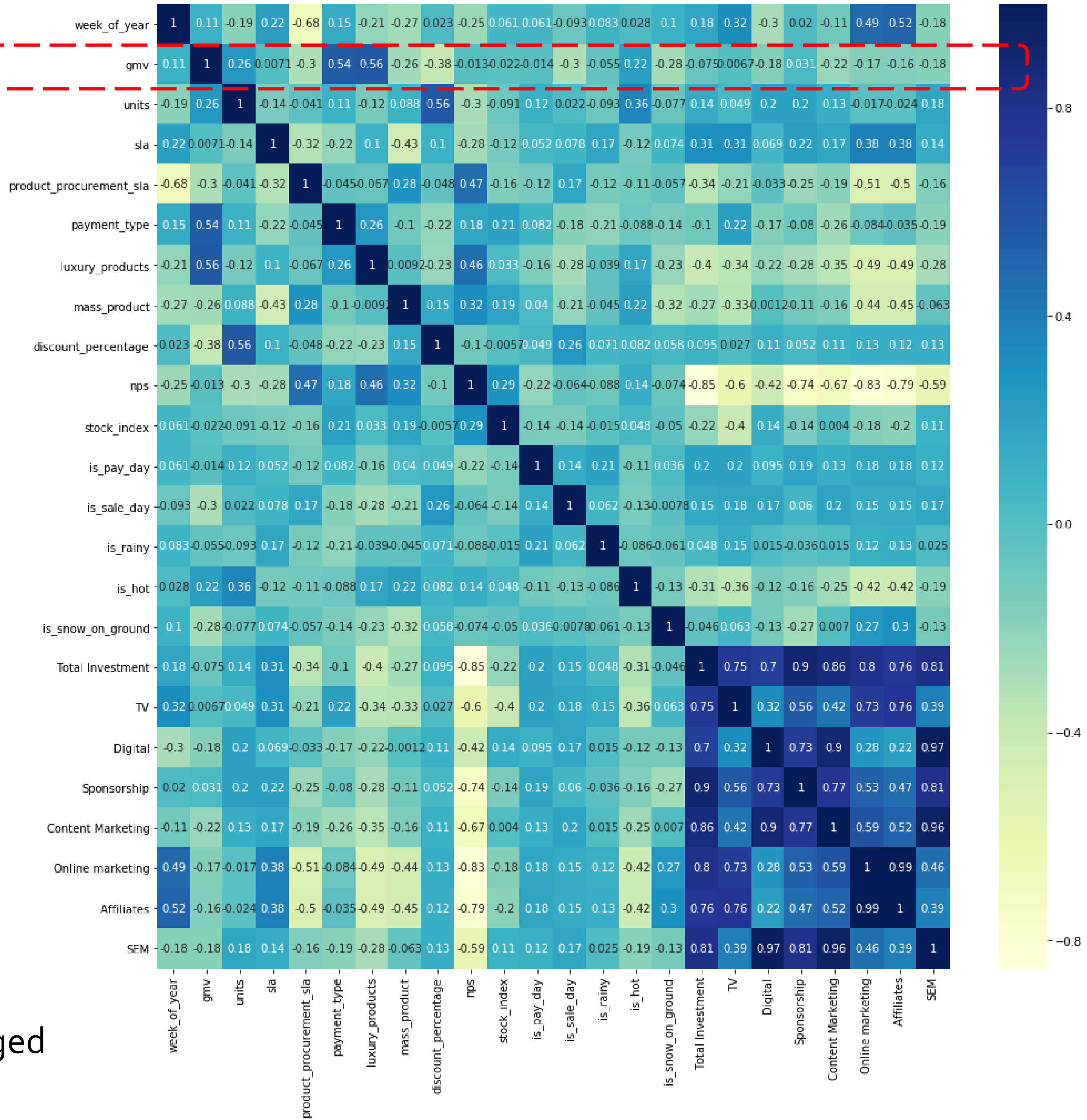
No major impact of Sponsorships



No impact of Online Marketing too

Gaming Accessory

Correlation Matrix



GMV has positive correlation with Payment type, hot weather and Luxury tagged product variables and negative correlation with procurement SLA, mass tagged products, Discount %age and content marketing etc

First Iteration
Basic Linear Model

Camera Accessory

Basic Linear Model

```
PREDICTORS:      product_procurement_sla, luxury_products, discount_percentage, SEM

R-Square:                0.9298096812604782
Adjusted R-Square:       0.9269447702915181
```

- We are able to get a decent metrics score with our basic linear model
- Adjusted R Square figures are based on the performance of the model on the training data
- As per our model, the most crucial factors for determination of GMV are:
  - product_procurement_sla
  - luxury_products
  - discount_percentage
  - SEM

Home Audio

Basic Linear Model

```
PREDICTORS:      payment_type, luxury_products, mass_product, TV, Digital

R-Square:                 0.6647786019057769
Adjusted R-Square:        0.6429163368126753
```

- We are able to get somewhat decent metrics score with our basic linear model
- Adjusted R Square figures are based on the performance of the model on the training data
- As per our model, the most crucial factors for determination of GMV are:
  - payment_type
  - luxury_products
  - mass_product
  - TV
  - Digital

Gaming Accessory

Basic Linear Model

```
PREDICTORS:     units, product_procurement_sla, payment_type, luxury_products, discount_percentage

R-Square:                0.7639817798513394
Adjusted R-Square:       0.7495316847401969
```

- We are able to get a decent metrics score with our basic linear model
- Adjusted R Square figures are based on the performance of the model on the training data
- As per our model, the most crucial factors for determination of GMV are:
  - units
  - product_procurement_sla
  - payment_type
  - luxury_products
  - discount_percentage

Future Plans (Roadmap)

Future Plans (Roadmap)

- Our model is still a basic linear model, which could be significantly improved in order to better understand the factors affecting GMV
- We will try to build the following types of model in our future attempts to improve the model:
  - ***Multiplicative Model***
  - ***Koyck Model***
  - ***Distributed Lag (additive) Model***
  - ***Distributed Lag (multiplicative) Model***
- We would try to perform rigorous ***hyperparameter tuning and cross validations*** to further improve each of these models ***for each of the given sub categories***.

Thank you