

# Ecommerce Capstone Project

## (Mid Submission)

Group Members:

1. Merin Jose
2. Suresh Balla
3. Deepak Aneja
4. Fayiz Mayam Veettil

21 Oct 2018

# Table of Contents

- **Business Understanding**
- **Data Understanding**
- **Data Preparation**
- **Modelling**
- **Recommendations**

## Business Understanding

“ElecKart” is an e-commerce firm specializing in electronic products. They spent a significant amount of money in marketing and occasionally they offer big-ticket promotions as well. It takes orders and delivers at door steps and deal with all major electronics product categories - Camera, Home Audio & Gaming Accessories

The money ElecKart spent over last 12 months on marketing was not sufficiently impactful. Now they want to create a marketing budget for the next year which includes spending on commercials, online campaigns, and pricing & promotion strategies and maximum revenue out of it.

The key objectives of the analysis are:

- Develop a market mix model to observe the actual impact of different marketing variables over the last year
- Recommend the optimal budget allocation for different marketing levers for the next year
- Improve the revenue response either by cutting down the marketing budget or reallocating it optimally across marketing levers

# Data Understanding

- Dataset contains around 1.6 M records of Consumer Electronics data  
**from July 2015 to June 2016**
- Other available datasets are
  - Media Investments
  - Special Sale Calendar
  - Monthly NPS Score
- Product sub-categories used for analysis are
  - Camera Accessory
  - Home Audio
  - Gaming Accessory

## Data Understanding – Major Attributes

The data received for the analysis have 20 attributes. The important ones are -

Attribute	Description
FSN ID	The unique identification of each SKU
Order Date	Date on which the order was placed
Order ID	The unique identification number of each order
Order Item ID	Unique Product/Item IDs under the same order ID;
GMV	Gross Merchandise Value or Revenue
Units	Number of units of the specific product sold
Order Payment Type	How the order was paid – prepaid or cash on delivery
SLA	Number of days it typically takes to deliver the product
Cust ID	Unique identification of a customer
Product MRP	Maximum retail price of the product
Product Procurement SLA	Time typically taken to procure the product

## Data Understanding – Additional Information

apart from consumer data following addition information received for model building

- Monthly spends on various advertising channels –  
Monthly spends split by various medias types such as TV, Internet, Radio and Others
- Days when there was any special sale –  
Information about the special sale calendar
- Monthly NPS score –  
Net promote score by month

## Step by Step Approach for Analysis

- Cleaning and formatting data
- Derive useful metrics out of data
- Check the data density and correlations
- Data Scaling
- Splitting the data between train and test
- Data Modeling
- Conclusion

# Exploratory Data Analysis – Data Cleaning

Check for possible data inconsistencies

- Check for duplicate records
- Check for NA values in the columns used for analysis
- Check the period of the provided data
- Status of data cleaning
  - Around 100K records found duplicated and removed from the dataset
  - Around 4K removed where Gross Merchandise Value is Null
  - Removed the data which are not in the period used for analysis
  - Below columns were removed after applying Near Zero Variances
    - units, product\_analytics\_super\_category
  - Records with orders with revenue per unit more than MRP, negative SLA or product\_procurement\_sla – 26186 records.
    - Capped the GMV accordingly (MRP\*units) so we do not lose revenue as we have investment for all days in a month

## Exploratory Data Analysis – Derived Metrics

Below the new columns derived for further analysis

- **Offer Percentage** =  $((\text{Product MRP} * \text{Unit Sold}) - \text{Gross Merchandise Value or Revenue})$  divided by  $(\text{Product MRP} * \text{Unit Sold})$
- **Product MRP Class** = classifying products into three categories – **premium**, **medium** and **cheap**
  - **Premium** – price range more than 75 percentile in same product\_analytic\_vertical
  - **Medium** – price within 25 to 75 percentile in same product\_analytic\_vertical
  - **Cheap** – price less than 25 percentile in same product\_analytic\_vertical
- **Week Number**, week number of the year ranging from 0-53 with Monday as first day of the week
- **Day**, day of the month ranging from 0-31

# Categorical Vs Numerical Variables

Below the list of categorical and numerical variables considered for further analysis

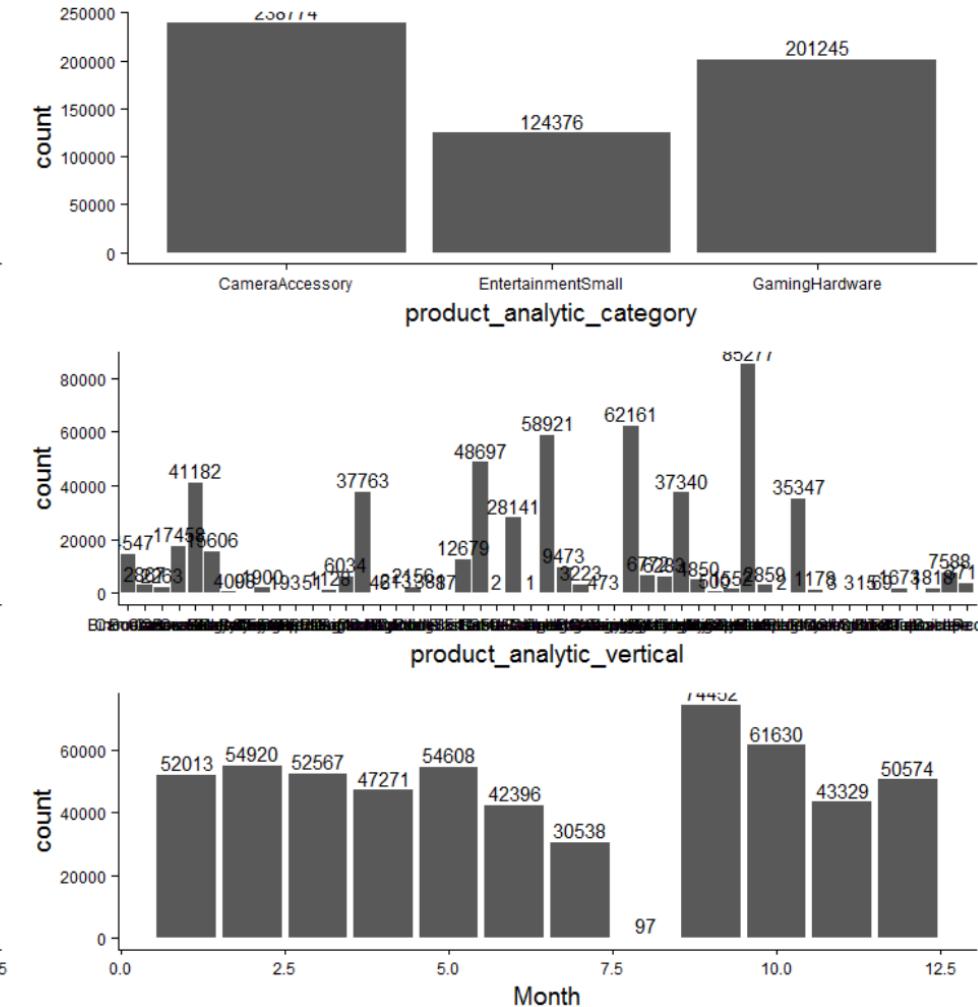
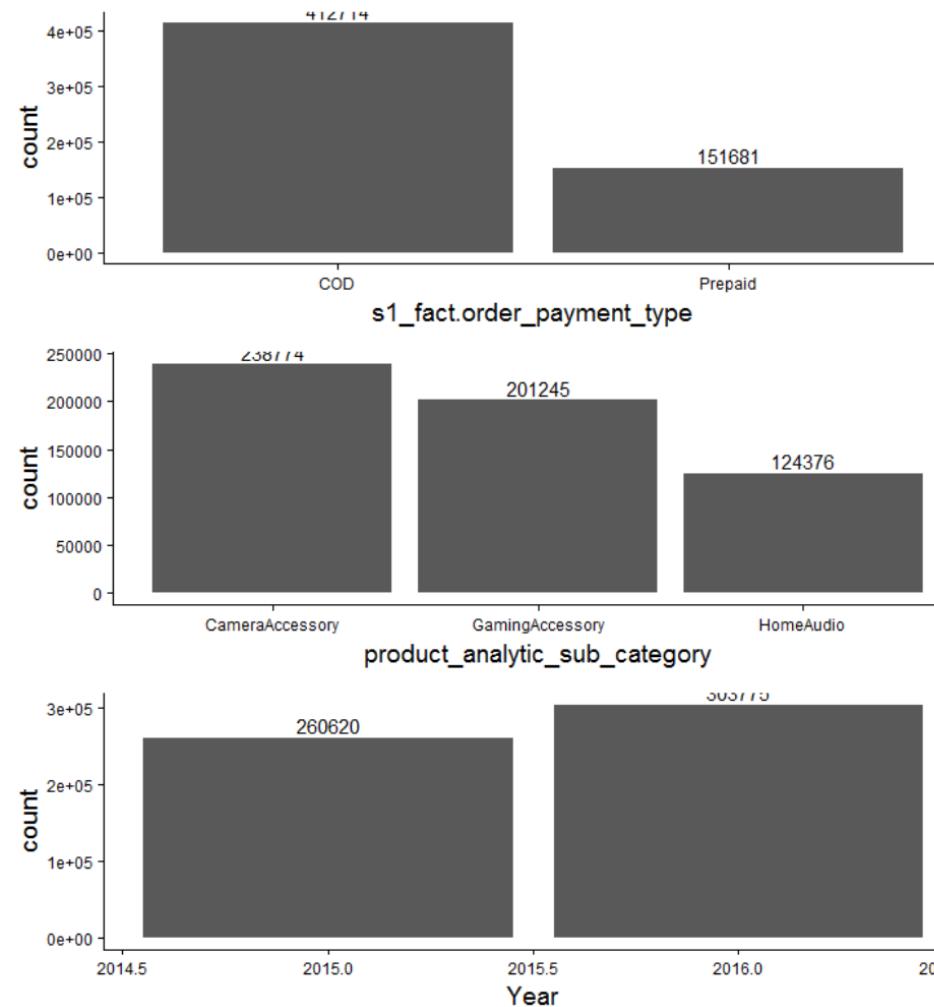
**Categorical Variables**

s1\_fact.order\_payment\_type  
product\_analytic\_category  
product\_analytic\_sub\_category  
product\_analytic\_vertical  
product\_mrp\_class  
Year  
Month

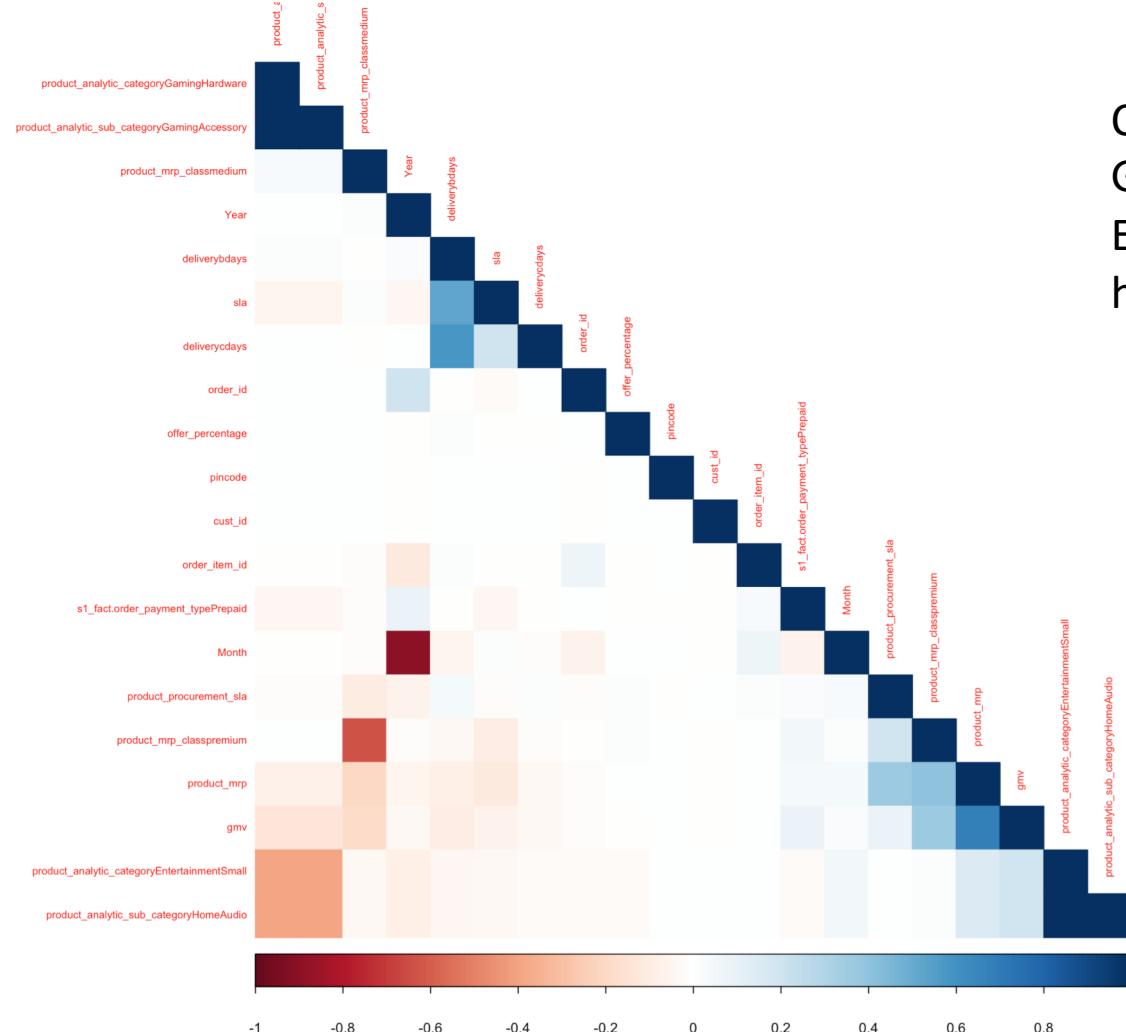
**Numerical Variables**

gmv  
deliverybdays  
deliverycdays  
sla  
pincode  
product\_mrp  
product\_procurement\_sla  
offer\_percentage

# Univariate Analysis – Categorical Variables

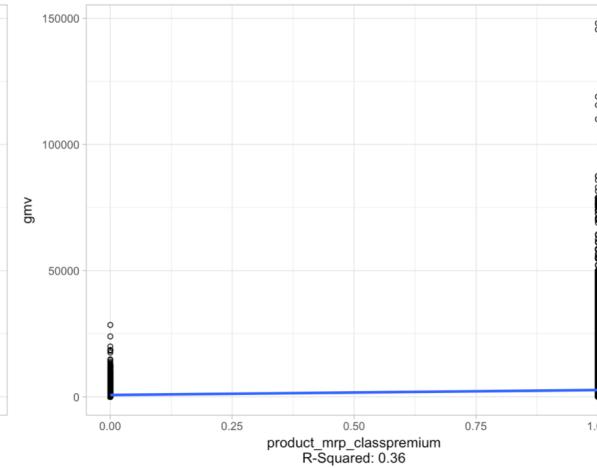
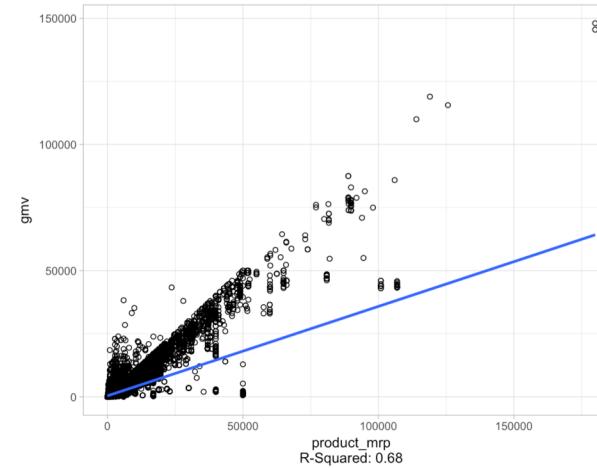
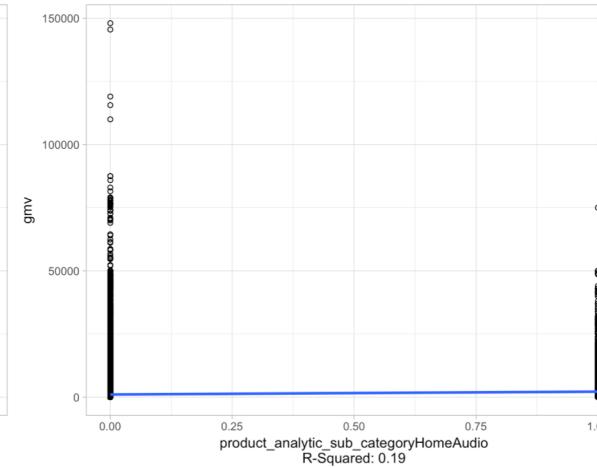
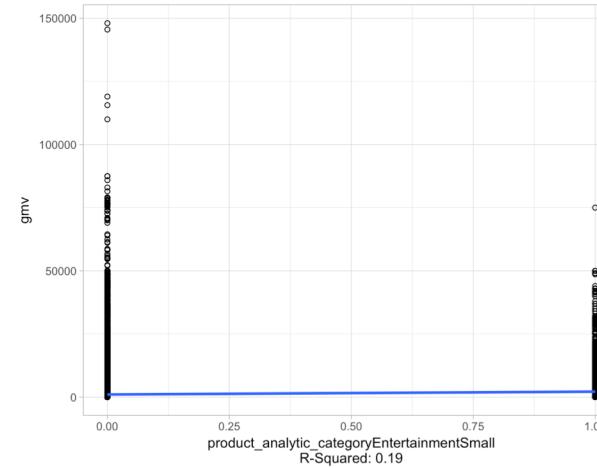


# Bivariate Analysis – Correlation Plot



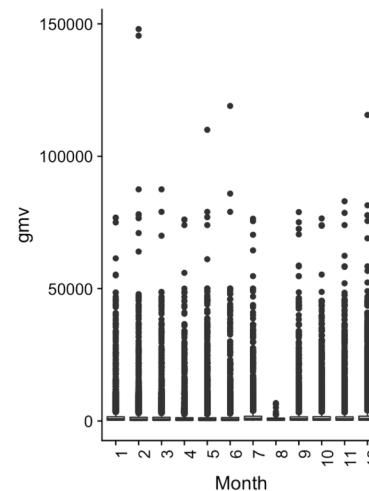
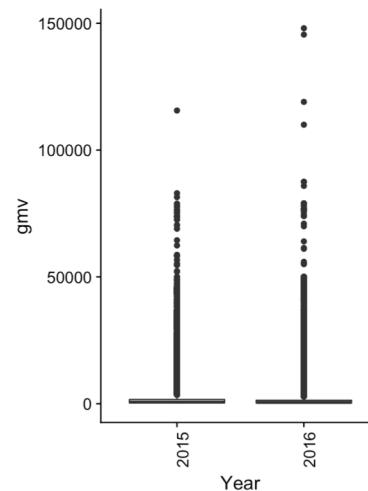
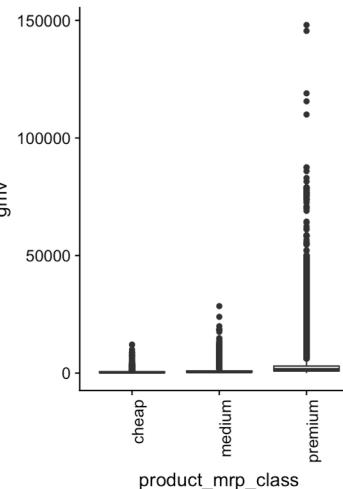
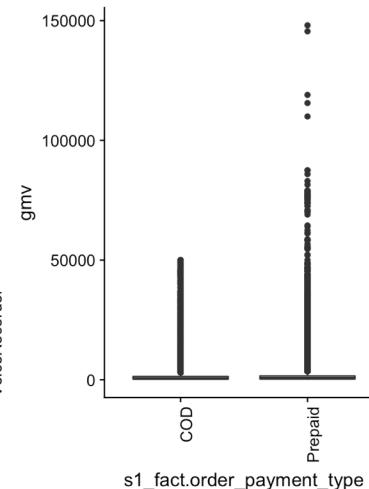
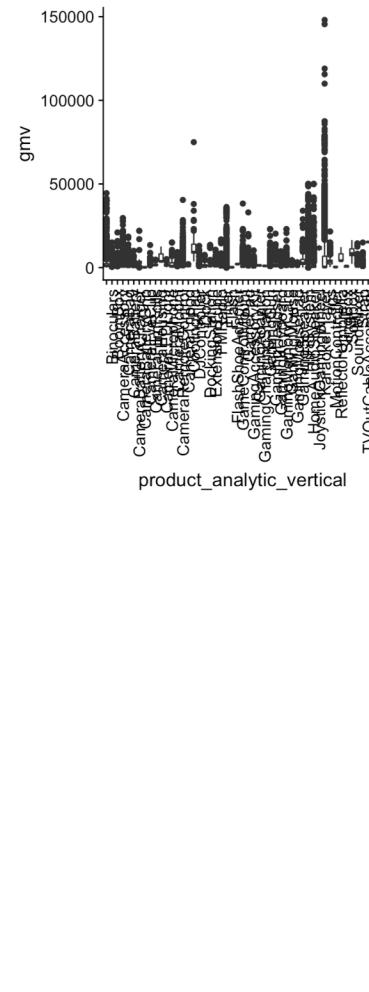
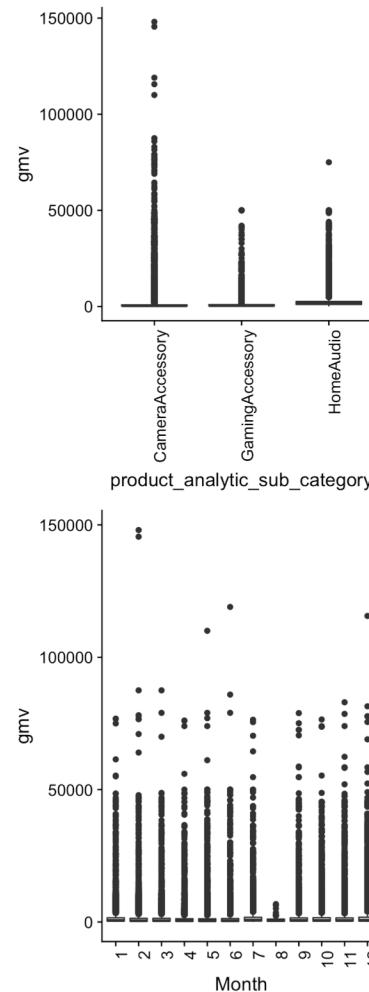
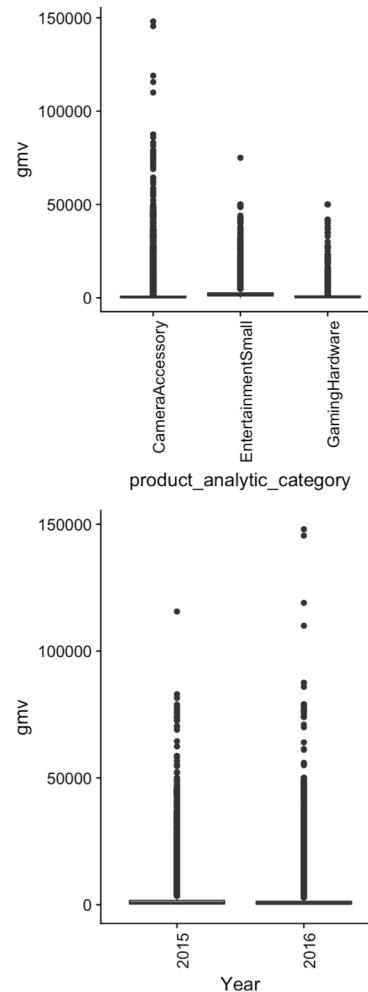
Categories such as  
Gaming Hardware, Gaming Accessory,  
Entertainment Small and Home Audio are  
highly correlated

# Bivariate Analysis – High Correlated Variable

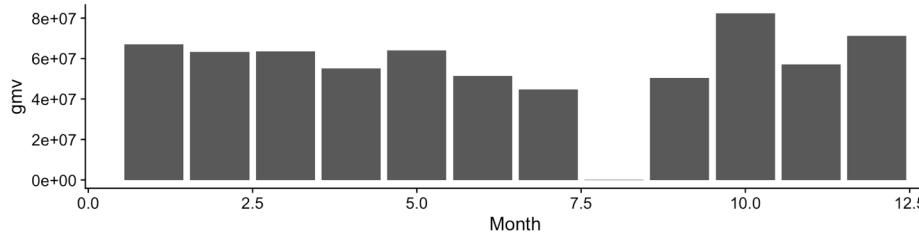
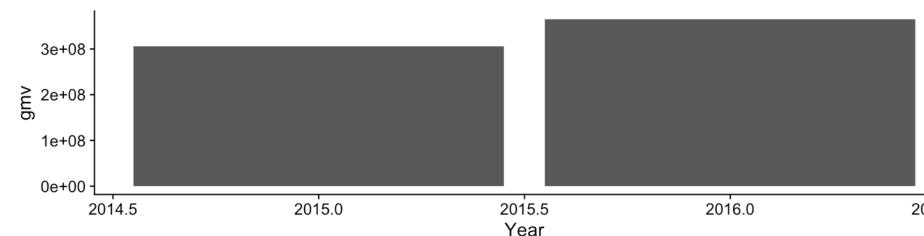
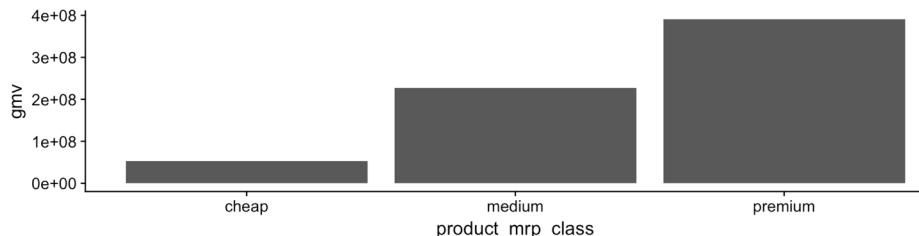
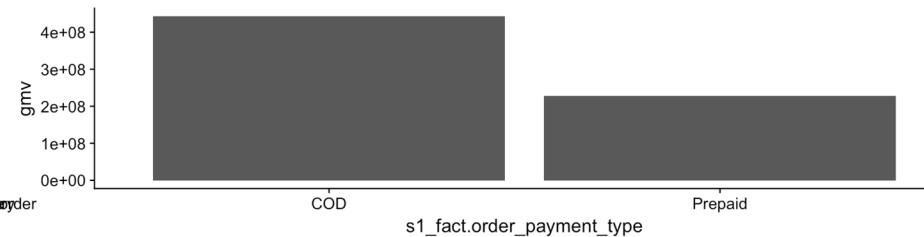
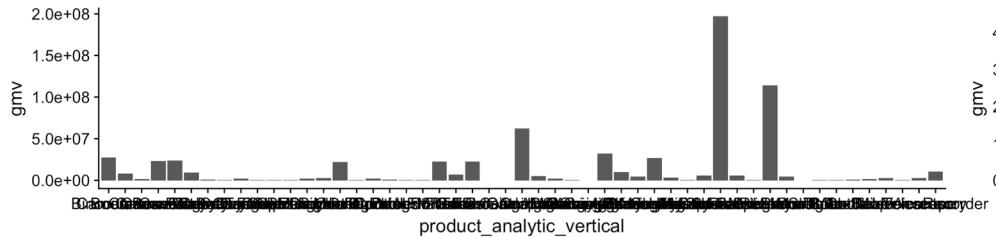
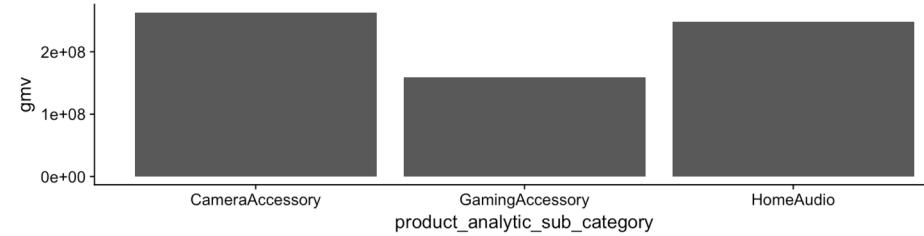
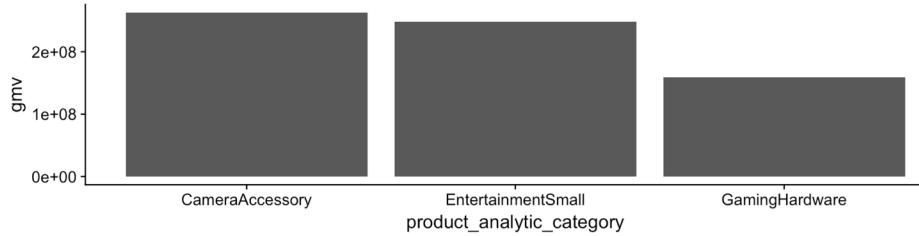


**Product\_MRP** showed high correlated to **GMV**,  
followed by sales of premium products

## Boxplots of numeric variables relative to gmv



## Aggregated Plots



Plots explains the aggregation of different values of the Categorical Variables

# Aggregation Strategy

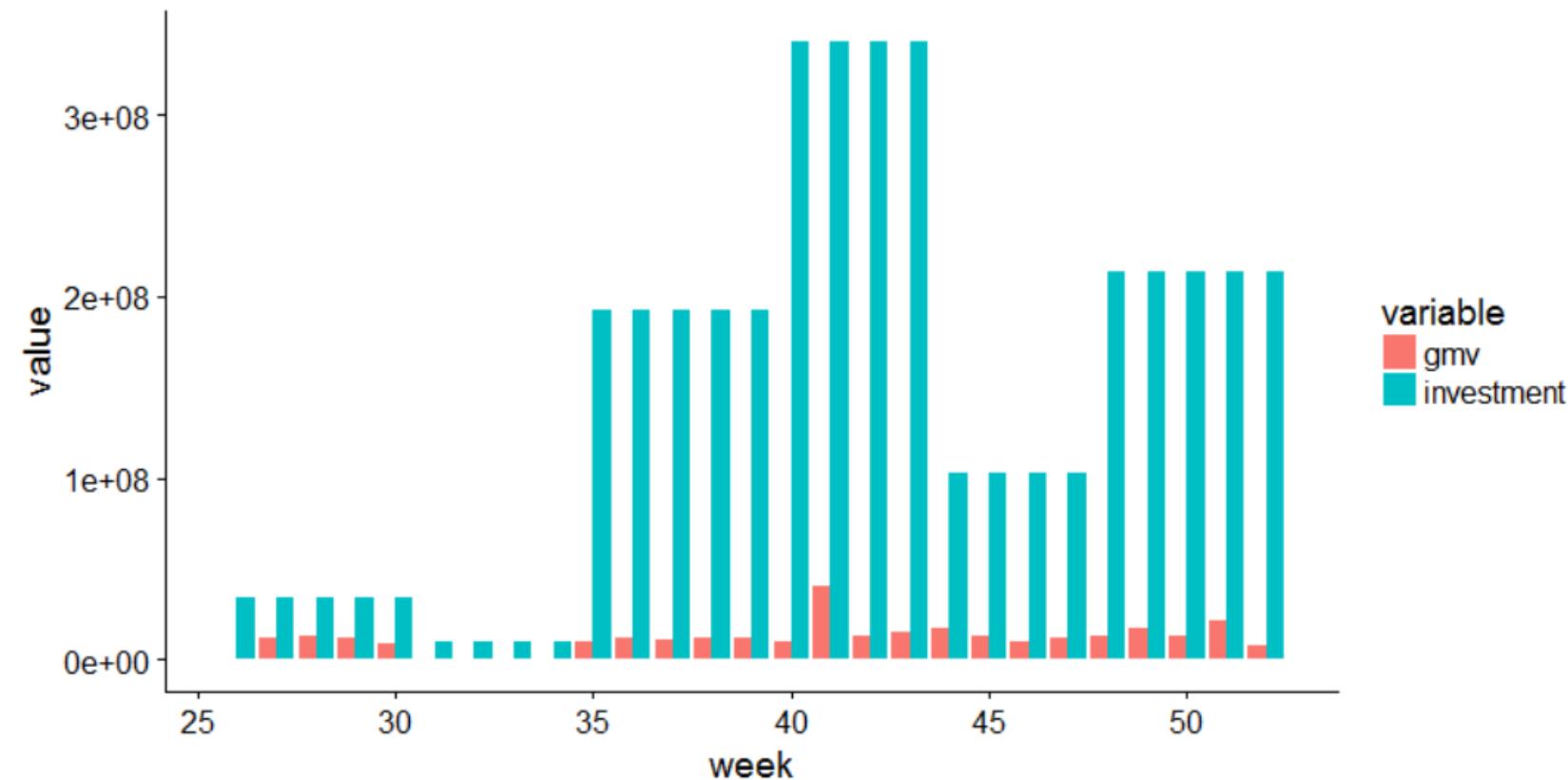
Strategies applied to aggregate the data are

- **Aggregate By Week**
  - **Year and Week** variables used to **GROUP** the dataset
  - Applied **SUM** function for **GMV, Product MRP, Offer Price** and other numerical variables
  - Applied **MEAN\*** function for deliverybdays, deliverycdays, product\_procurement\_sla variables
  - Created new columns for each possible value of category (repeated across all category columns) that specifies percentage of categorical value among other values in same columns. For example, a week with 2 prepaid and 3 COD orders, aggregated columns are payment\_prepaid = 2/5 and payment\_cod = 3/5

\* since these variables are represents the number of days applying SUM function is not logical hence used MEAN function

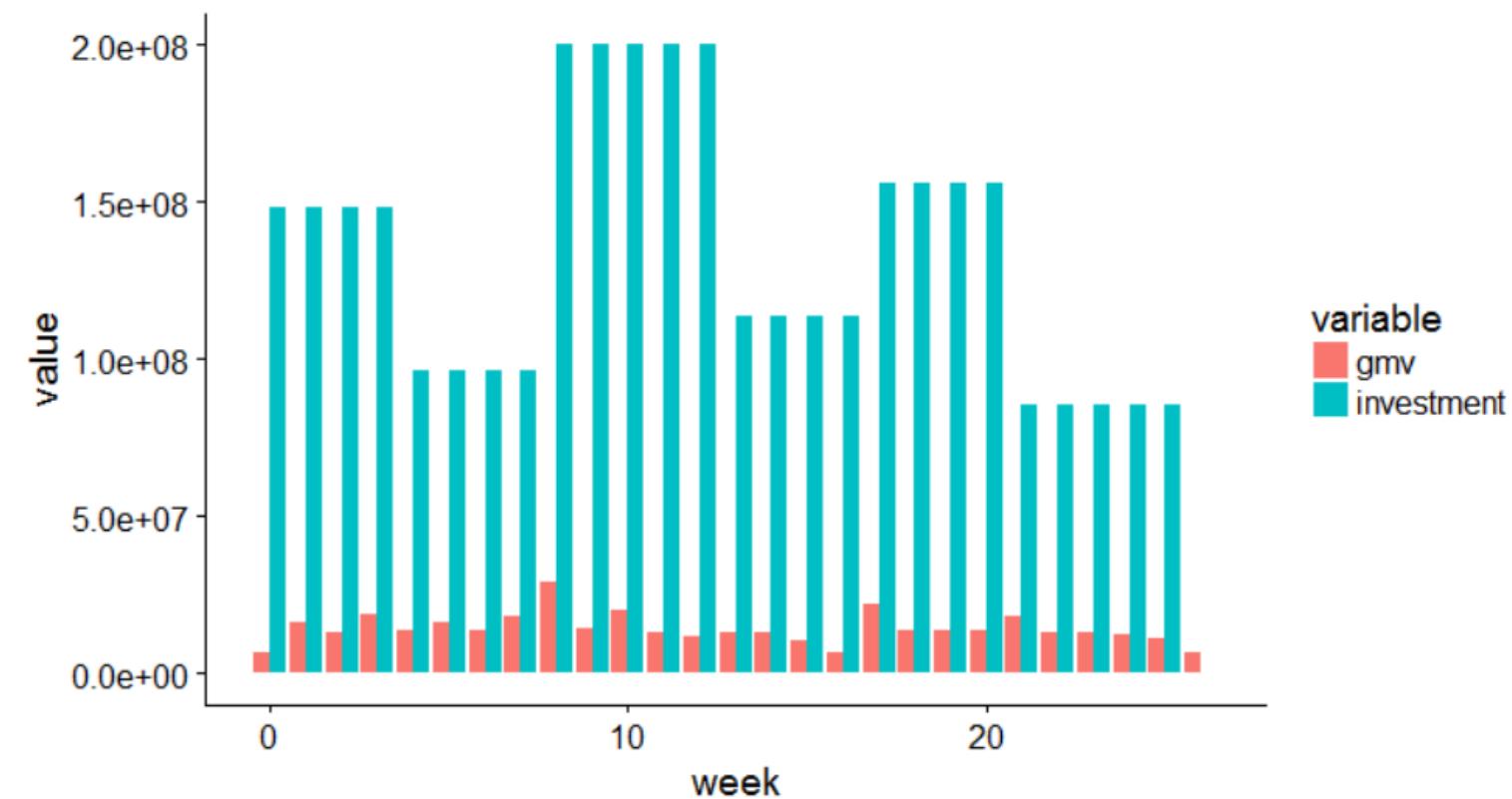
# Weekly Analysis

## Weekly Analysis, 2015 – Revenue Vs Investments



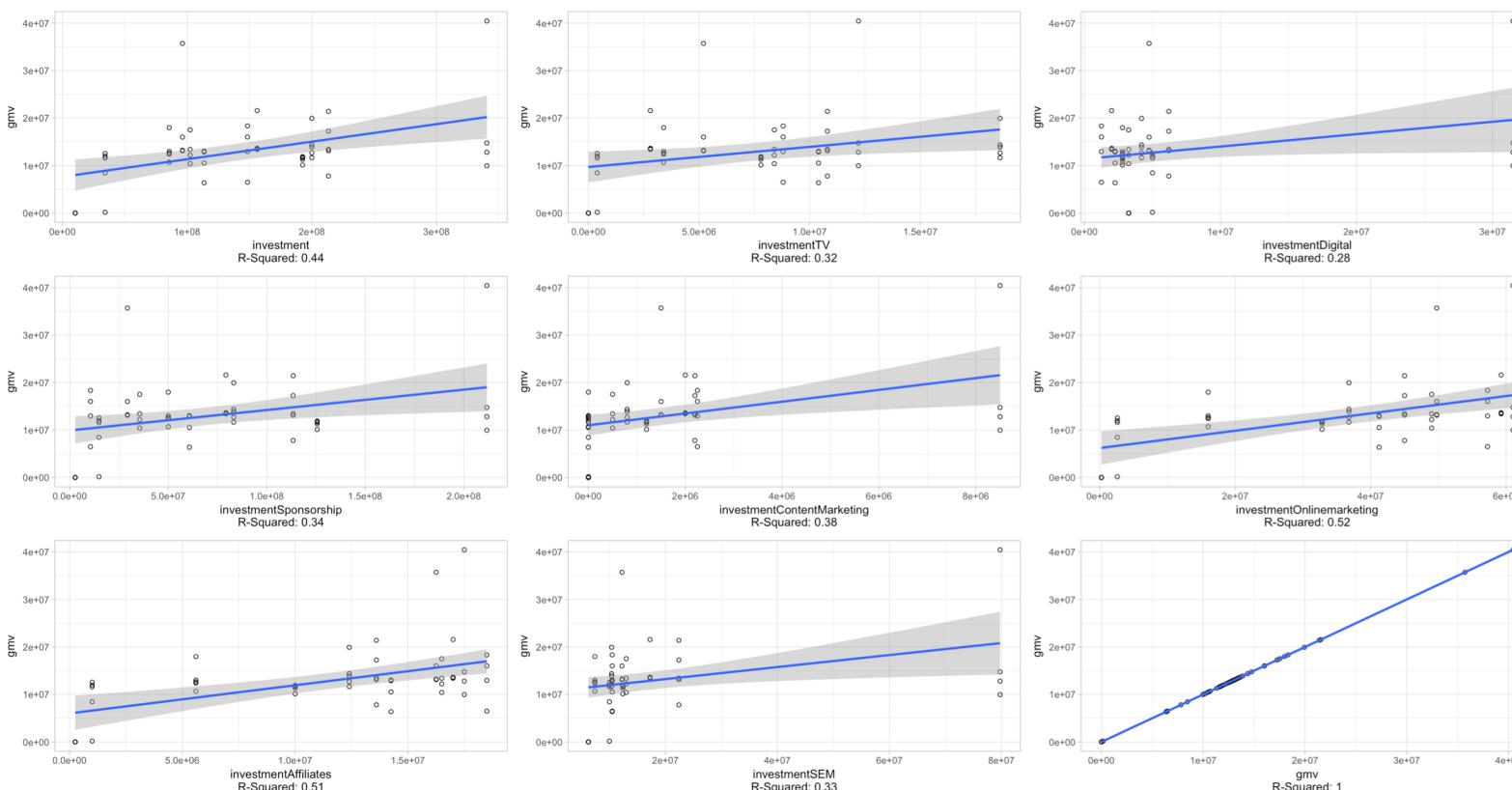
Weekly Revenue vs. Investments Analysis from July – Dec 2016 (Week# 26-52)

## Weekly Analysis, 2016 – Revenue Vs Investments

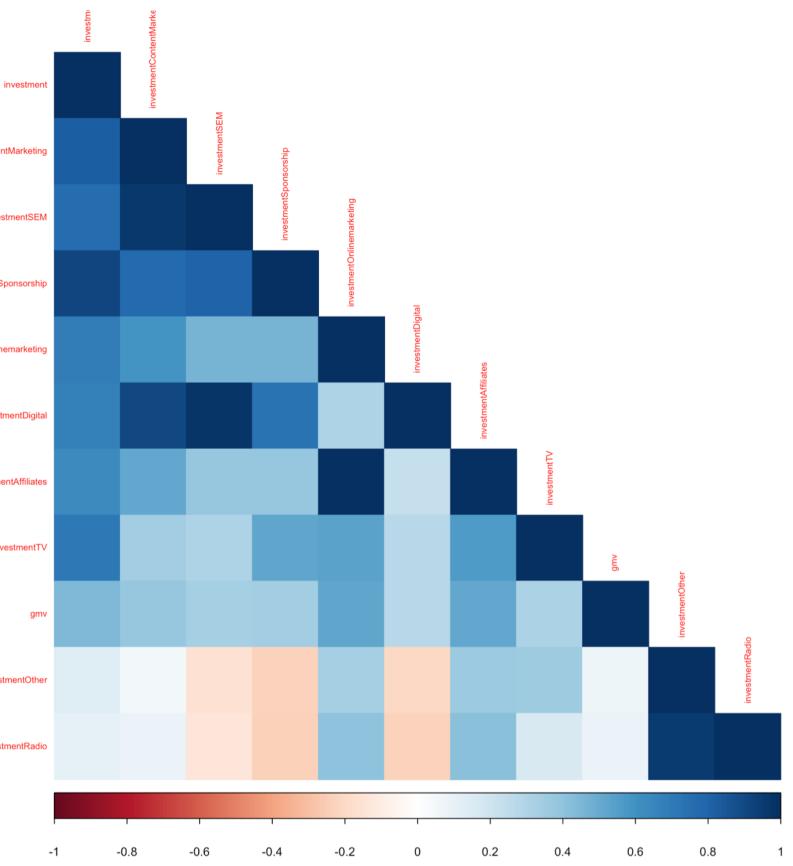


Weekly Revenue vs. Investments Analysis from Jan – Jun 2015 (Week # 0-25)

# Weekly Analysis, High Correlated Variables



Seems TV, Digital, sponsorship, Content Marketing, Affiliates and SEM are highly correlated with Gmv



Highly correlated variables with Investments

# Modelling

## Models Attempted

1. Linear Regression
2. Multiplicative – Log of gmv to be modelled with Log of independent variables
3. Kyock – Linear model but with one previous week of Marketing variables
4. Distributed Lag - Linear model but with last 3 weeks of Marketing variables (3 additional variable for each channel ex -1 week of TV channel investment, -2 week and -3 week)
5. Multi distributed Lag – Hybrid Model – Hybrid of 2nd and 4th model. Take last weeks variable but with log

## Home Audio - Models

Model	Variables	Adj. R Square	
Linear			
Multiplicative			
Kyock			
Distributed Lag			
Multi distributed Lag			

## Camera Accessories - Models

Model	Variables	Adj. R Square	
Linear			
Multiplicative			
Kyock			
Distributed Lag			
Multi distributed Lag			

## Gaming Accessories - Models

Model	Variables	Adj. R Square	
Linear			
Multiplicative			
Kyock			
Distributed Lag			
Multi distributed Lag			

## Recommendations



# Thanks

Merin Jose

Suresh Balla

Deepak Aneja

Fayiz Mayam Veettil