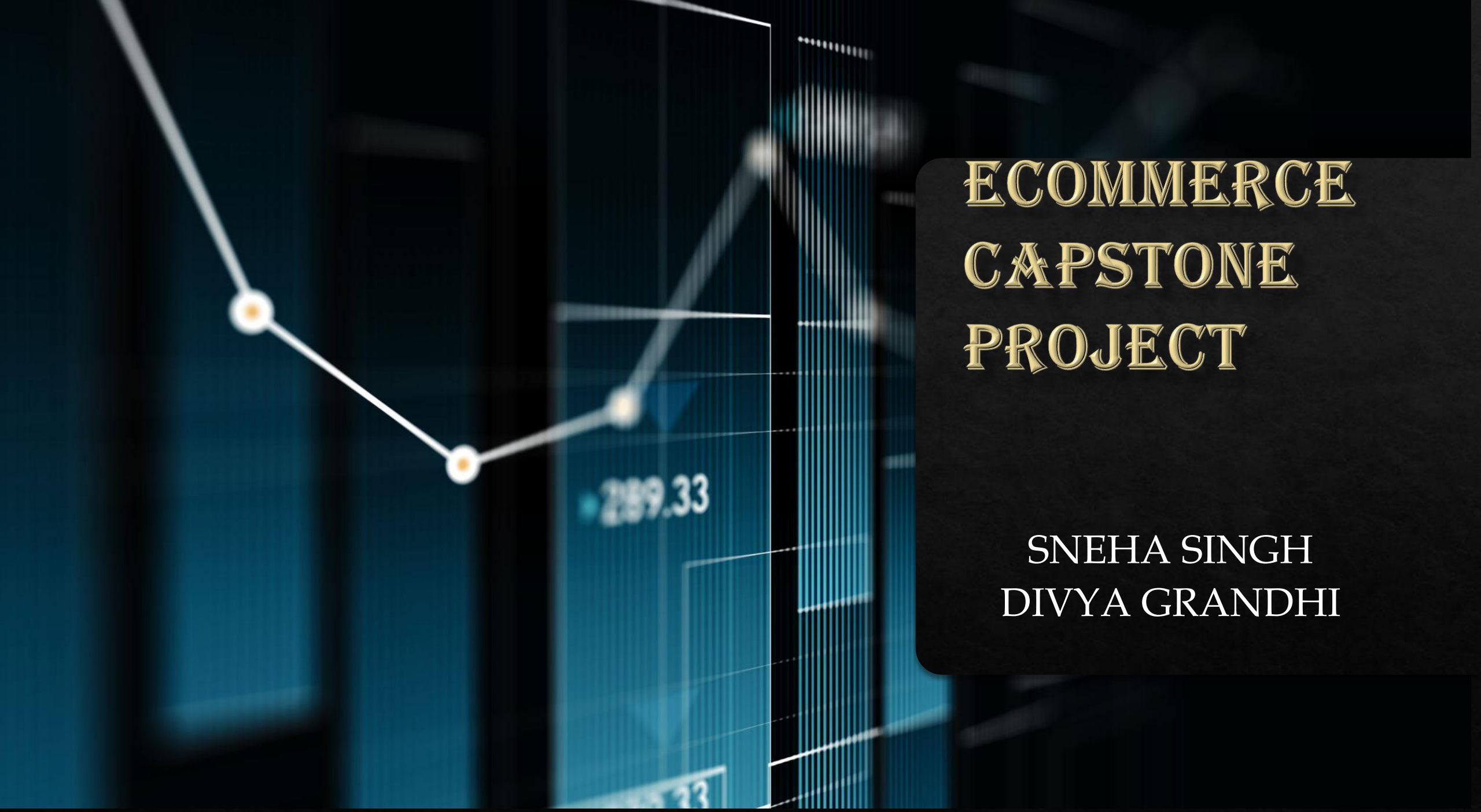


ECOMMERCE CAPSTONE PROJECT

SNEHA SINGH
DIVYA GRANDHI

MARKET MIX MODELLING

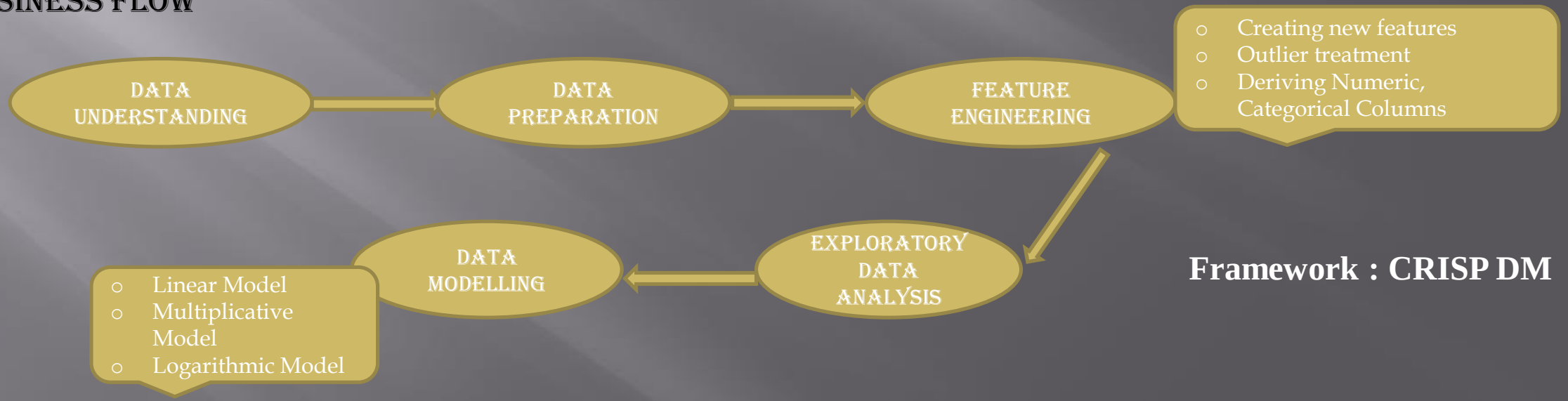
- Market mix modeling involves the use of multiple regression techniques to help predict the optimal mix of marketing variables. Regression is based on a number of inputs (or independent variables) and how these relate to an outcome (or dependent variable) such as sales or profits. Once the model is built and validated, the input variables (advertising, promotion, etc.) can be manipulated to determine the net effect on a company's sales or profits.

- Market mix modeling brings accountability to the spending and decision making in marketing. It uses various types of statistical models to model the relationship between the different categories of spending and their impact on the sales and revenue

BUSINESS OBJECTIVE

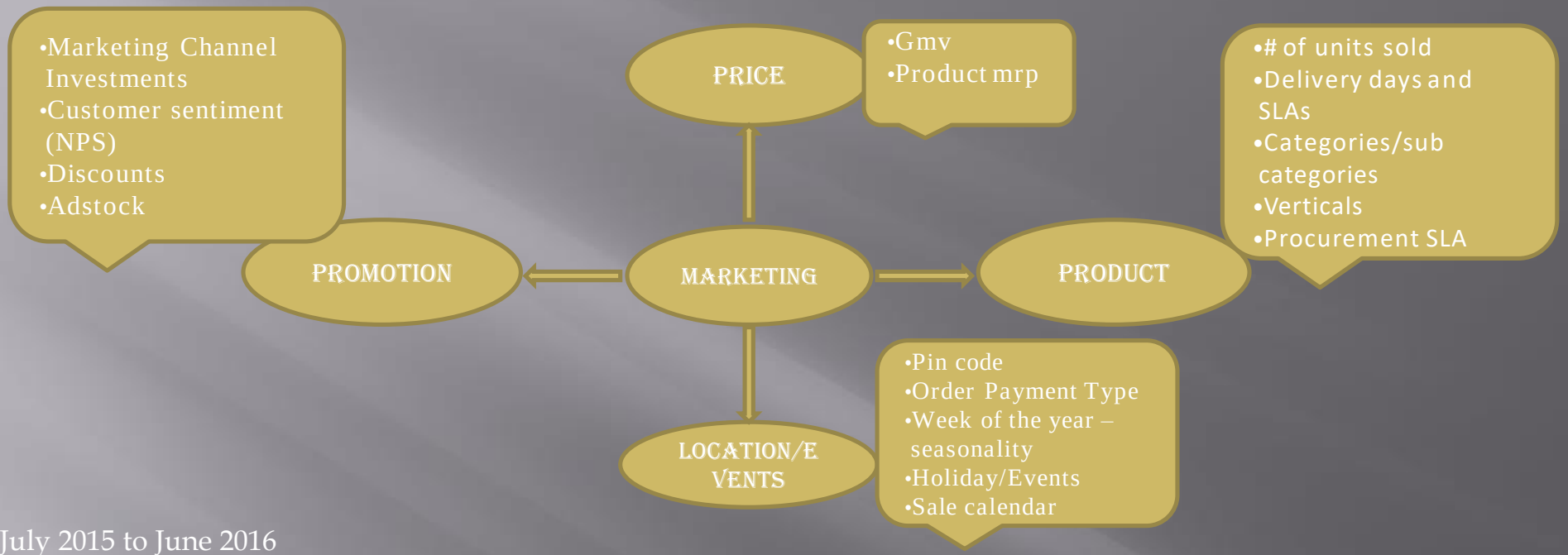
CFO of ElecKart (an e-commerce firm based out of Ontario, Canada) has decided to cut on the budget or reallocate it optimally across marketing levers to improve the revenue. So, now we have to create a Market Mix model using One Year Data (July 2015 to June 2016) for the three specified categories Camera Accessory, Gaming Accessory and Home Audio to Recommend the optimal budget allocation for marketing to the next year.

BUSINESS FLOW



Framework : CRISP DM

DATA UNDERSTANDING



Data : July 2015 to June 2016

Columns : FSN ID, Order Date, Order ID, Order item ID, GMV, Units, Order payment type, SLA, Cust id, Product MR, Product procurement SLA

Apart from above, the following information is also available:

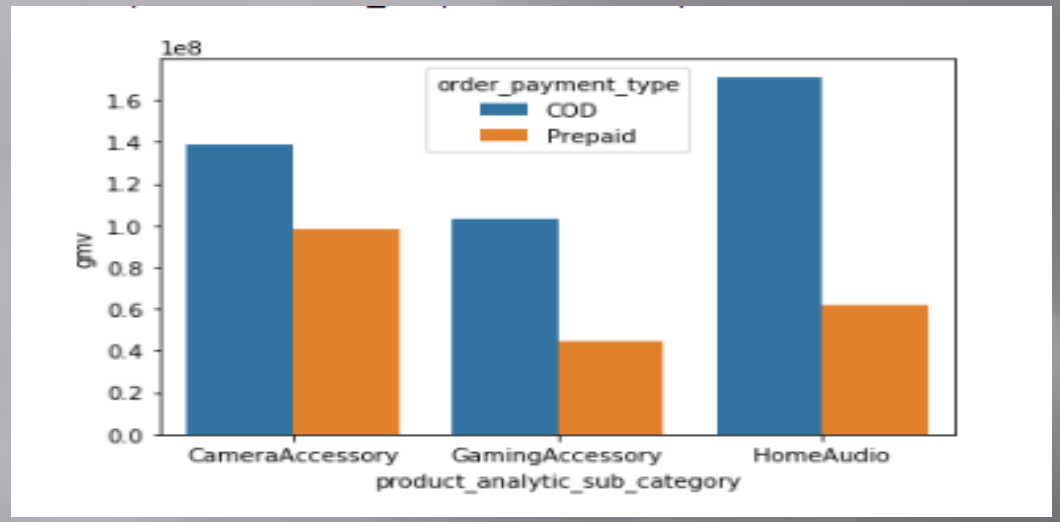
- Monthly spend on various advertising channels
- Days when there was any special sale
- Monthly NPS score – this may work as a proxy to 'voice of the customer'
- Stock Index of the company on a monthly basis
- Order level data, Holidays in Ontario State, SKU- Stock Keeping Unit
- deliverybdays- days to get item or order from warehouse for shipping
- deliverycdays - days to deliver item to customer

DATA PREPARATION

- Importing Libraries, Reading Data
- Converting order_date to datetime
- Data type conversion for columns order_id and order_item_id, GMV into suitable format
- Fixing Data types for deliverybdays, deliverybdays columns having "\N" value with 0
- Treating incorrect GMV values w.r.t product_mrp * units
- Impute the faulty mrp values with gmv/units
- Handling values with less than or equal to 0 in gmv, deliverybdays & deliverycdays, MRP, Units, product_procurement_sla,
- Check for various payment types
- Dropping product_analytic_super_category column as all the values are same(CE)
- Dropping columns like pincode, cust_id, fsn_id as they do not contribute much to the analysis
- Dropping Duplicates in the Data frame and dropping duplicates in the column order_item_id
- Filtering the data for the 3 categories with Camera Accessory, Gaming Accessory, Home Audio
- Generate Week Column

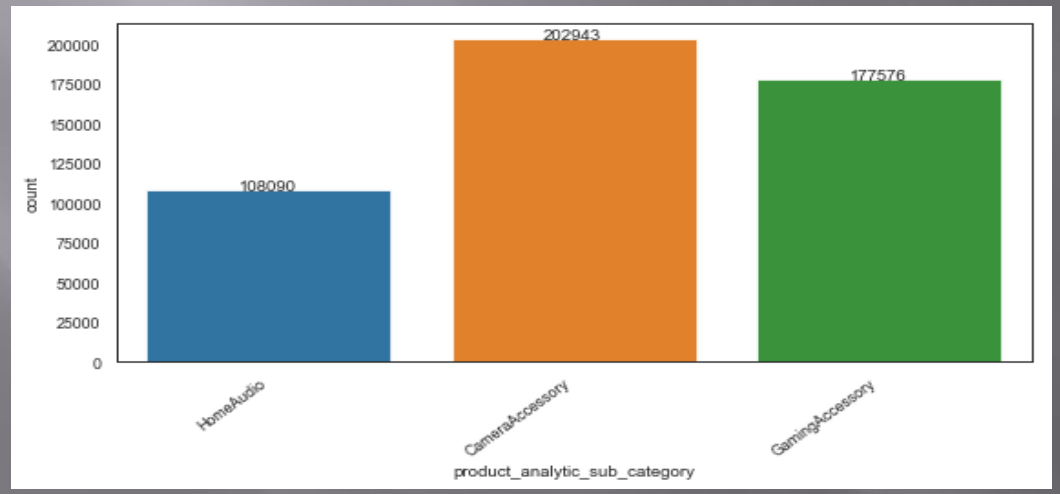
EDA

GMV is the Target variable



Insights Observed

- Max revenue for COD order is from the class Home Audio followed by Camera Accessory, then Gaming Accessory
- For prepaid orders, the maximum revenue is from Camera Accessory, followed by Home Audio and a slight decrease in the category of Gaming Accessories



No of transactions by Sub-Category

Based on our analysis , maximum number of transaction were from Camera Accessory category, then Gaming Accessory , followed by Home Audio category

FEATURE ENGINEERING

- Generate Week Column
- Updating the year as 2015 for the week of 53 but belonging to 2016 in accordance with the other data sets
- Updating the month to 12 for consistency in the above rows
- Dropping rows (less in number) with week# 27 as it belongs to the weeks in June 2015

```
Consumer_df.loc[(Consumer_df.Year == '2016') & (Consumer_df.Week == '53'), 'Year'] = '2015'
Consumer_df.loc[(Consumer_df.Year == '2015') & (Consumer_df.Week == '53'), 'Month'] = 12
Consumer_df.drop(Consumer_df[Consumer_df['Week'] == '27'].index, inplace = True)
```

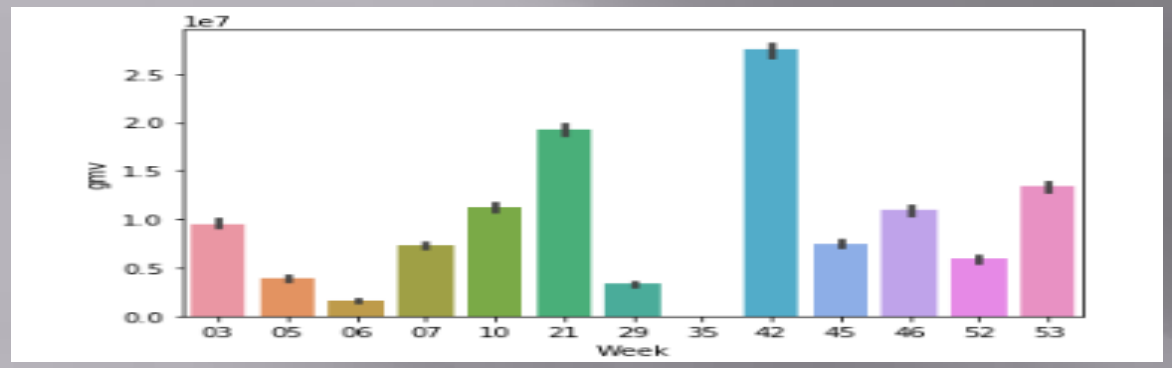
- Dropping Columns with Single Value or all Different Values
- Creating a new feature List Price by dividing GMV by Units
- Creating Payday column, If it is nearer to the salary day(1st and 15th of every month), we flag the column value as 1, else as 0
- Creating holiday_flag if there is holiday or occassion in Ontario, we're flagging the line as 1 else 0
- Creating a new column Product Type, If GMV value > 80 % then considering it as Premium_product, else mass_market
- Discount Created by substraction of product_mrp and list price
- % Discount by dividing discount and product_mrp
- Creating dummy variables for order_payment_typeFiltering data from July 2015 to June 2016
- Calculating GMV % to derive the marketing spend

FEATURE ENGINEERING

- Reading product list using the file “Media data and other information”
- Making holiday information usable using the 2nd sheet from the above file and cleaning the dataset
- Extracting start and End dates as per our given requirement
- Creating column ‘NumDays’ as the count of the sales
- Creating special_sales as 0 in the place of Nulls or 1 in the Event dates
- Calculating Average Sales based on these special_sales
- Mapping months to a month's number of weeks and getting a count of 52 weeks
- Importing the file with Media Investment and cleaning the dataset
- Creating Temp Data Frame and the monthly values are divided by the number of weeks in that month and taken as weekly data to fill the DF
- Calculate 8-weeks Exponential Moving Average for all Advertising media channels
- Calculate 5-weeks Simple Moving Average for advertising media channels, NPS and Stock_Index
- Calculate 3-weeks Simple Moving Average for Advertising media channels, NPS and Stock_Index
- Calculating Ad Stock values
- Assuming the value of Adstock rate(engagement factor)
- Creating (Net Promoter score)NPS - works as a proxy to ‘voice of the customer’ and Stock Index

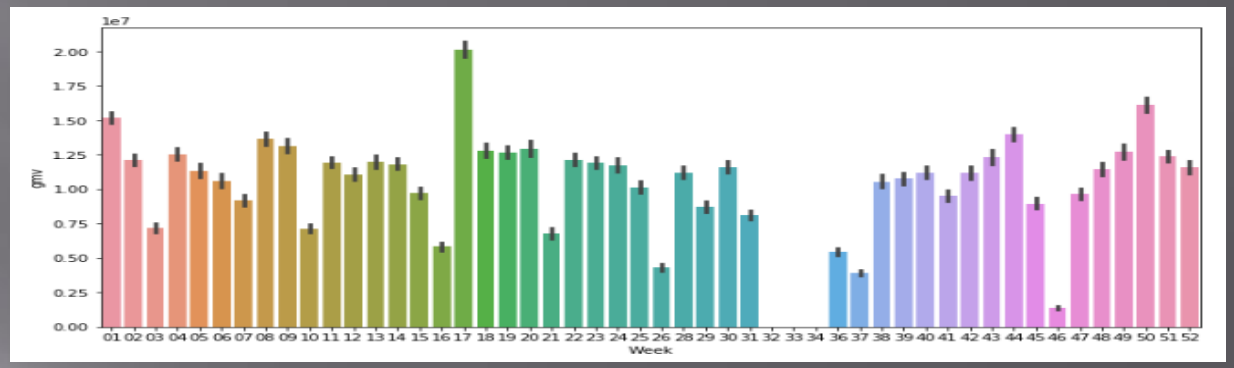
EDA

Plot for special sales flag =1



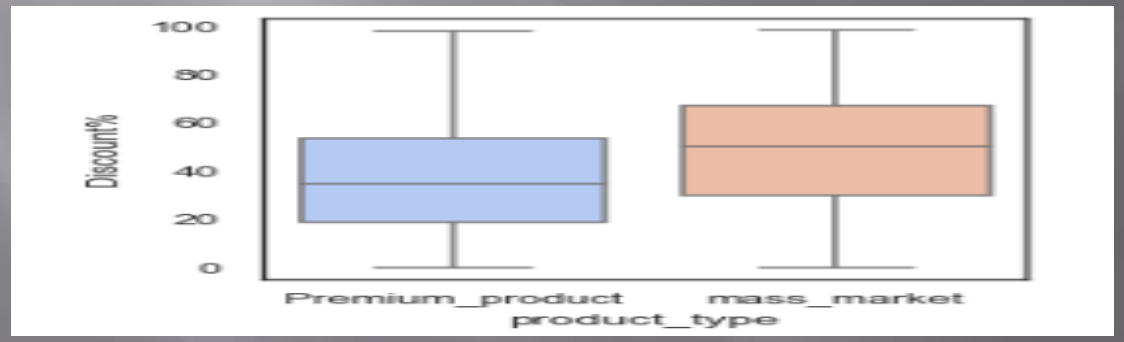
Insights : Observed highest Sales on week 42 for the special sale day

Plot for special sales flag =0



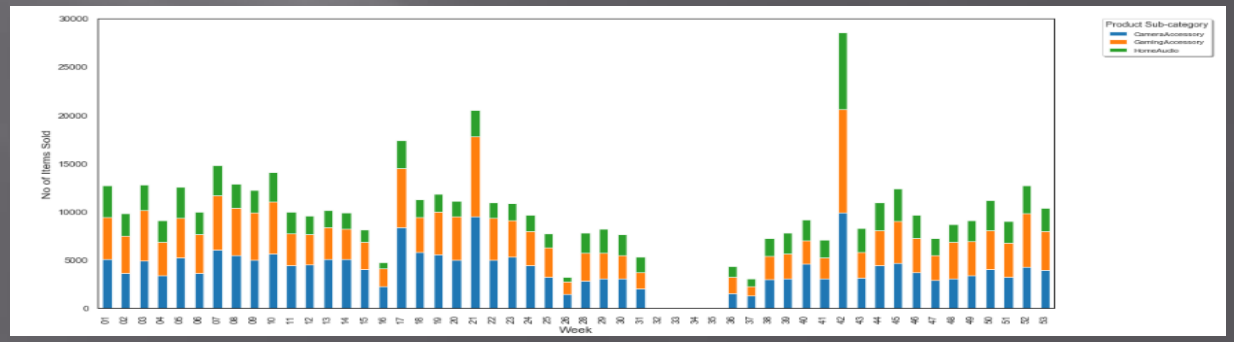
Insights : Observed highest Sales on week 17 for the special sales flag 0

Comparing Distribution of Discount% for product types



Insights : Relative to Mass Market Items, the average discount rate offered for premium_product is lower. This is a well-known phenomenon among premium_product or premium brands to offer limited or no discounts to maintain their products exclusivity.

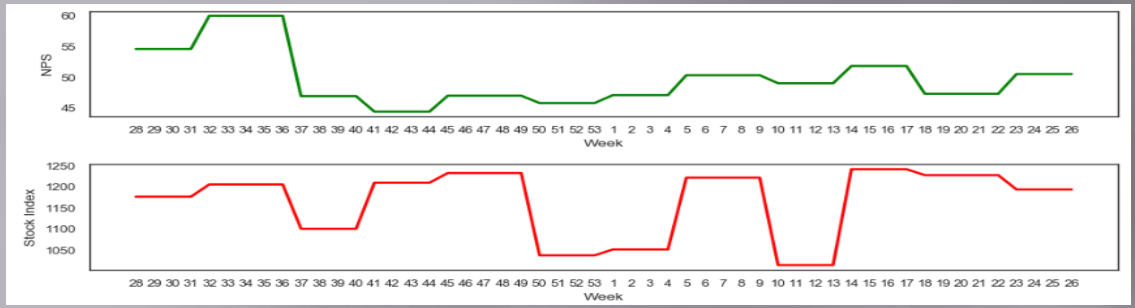
Total items sold per 3 product subcategories per Week



Insights : The sale on the 42nd week is maximum. Overall, October has seen most no of items being sold

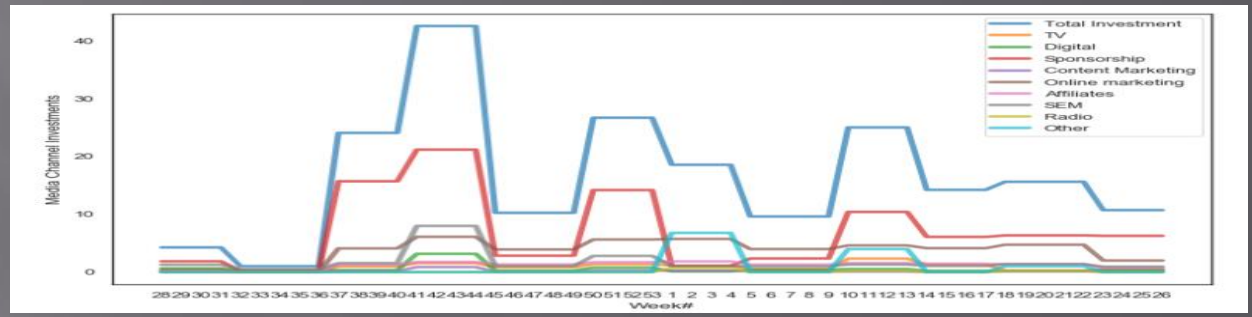
EDA

Displaying trend of NPS and Stock Index



Insight : In weeks 32 – 35 the product NPS rating was highest, which corresponds with the period when peak discounts are given.

Display trend of Media Channel Investments by week

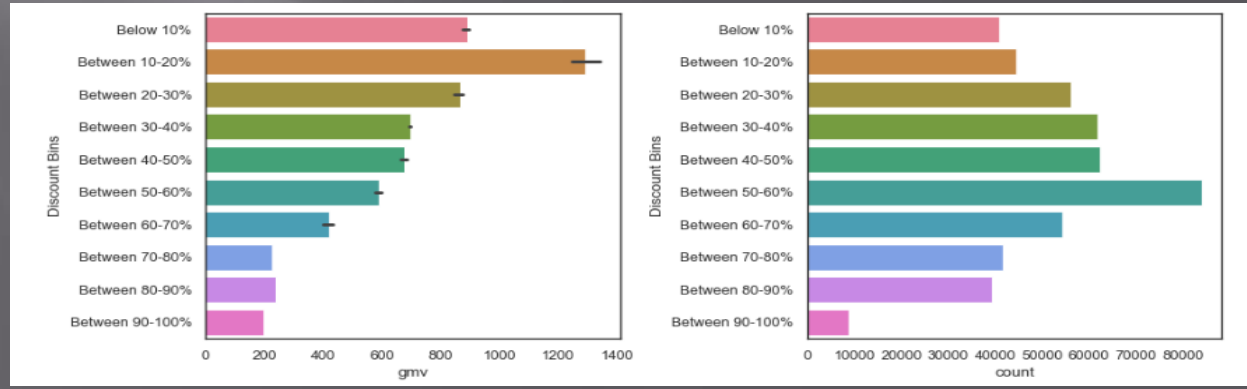


Insight : Most of the Ad Investment was made in sponsorships over the past year, followed by Online Marketing & Search Engine Marketing

Tree map for Revenue

Insights : The homeaudio speaker created the largest revenue followed by the camera accessory and gamepad lens in gaming accessories

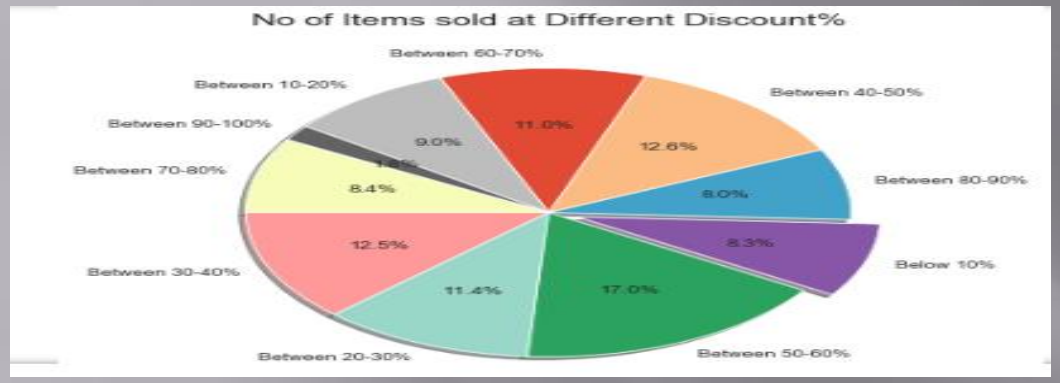
Plots for Discount Vs gmv/count



Insights : It shows that the revenue falls at a higher discount, even though the profits are strong, which means a loss to the company. An average 10-20 percent discount is the company's most profitable

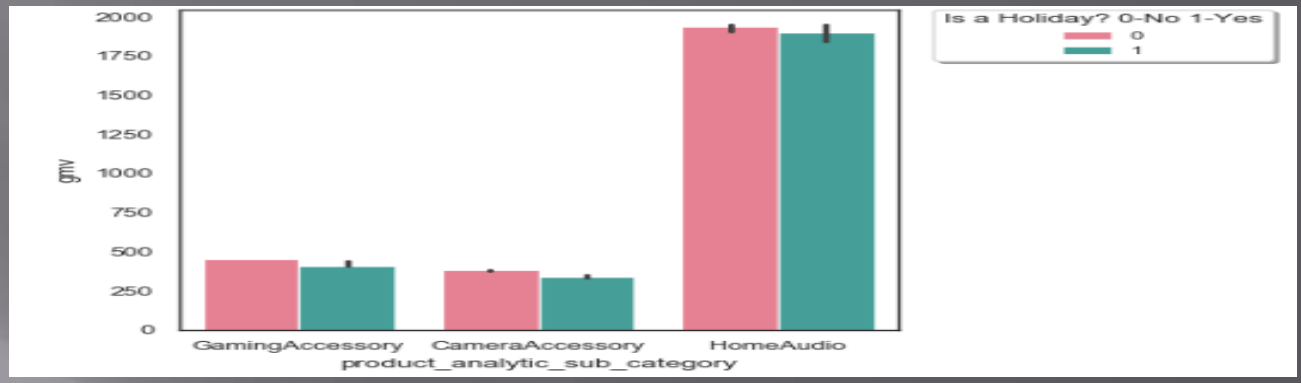
EDA

Percentage of items sold at different Discount% segments



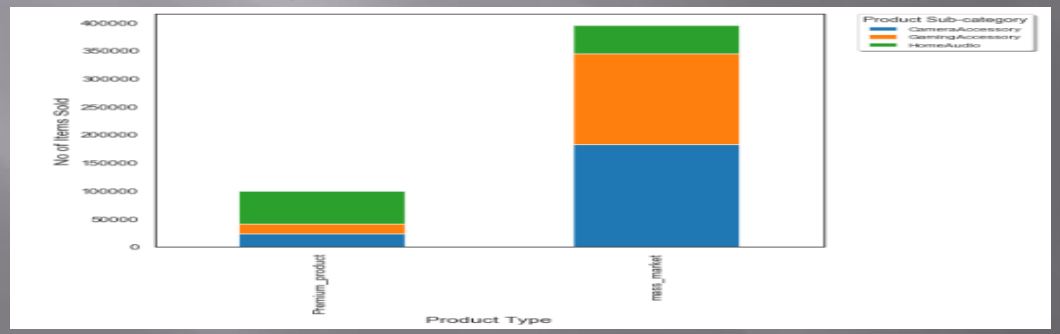
Insight : Majority of sales b/w 50-60%

Average Revenue from Holiday/Non-holiday days for product subcategories



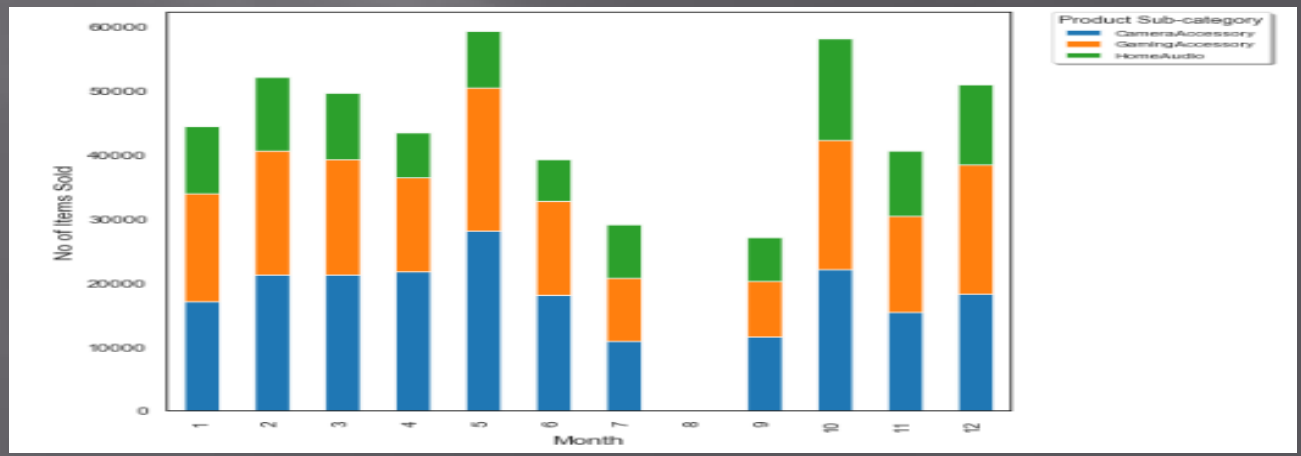
Insight : The average revenue from holiday and non-holiday days for 3 brand sub-categories is more or less equivalent.

Items(Premium_product/Mass-market) sold per product subcategories



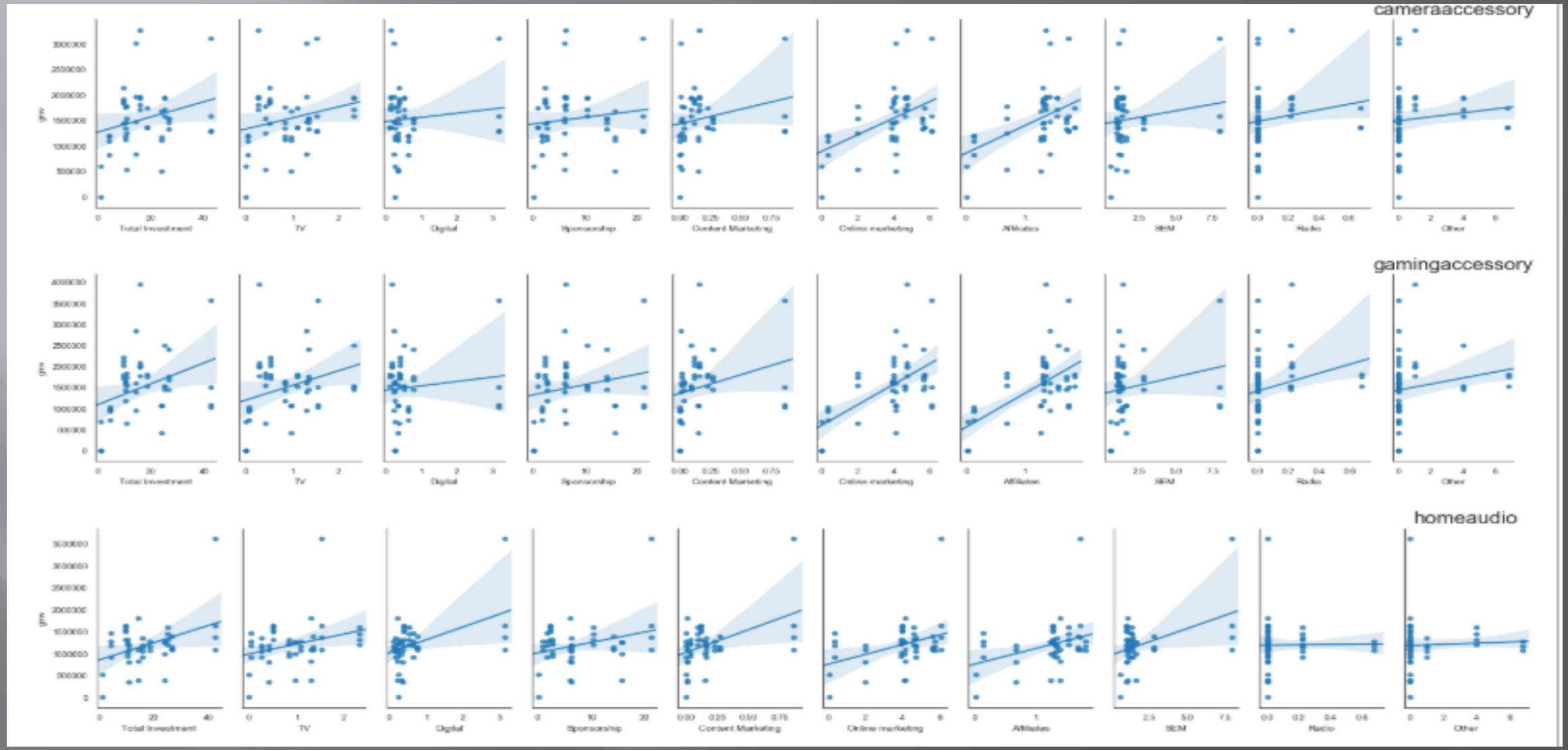
Insights : Most of the units sold belonged to the class of the mass market and among them camera and gaming accessories were sold on the mass market for highest….,Home audio products were among the best-selling products in Premium_product

Total items sold per 3 product subcategories per Month



Insights : Months 5, 10 has more sales

RELATIONSHIP BETWEEN SALES AND AD-STOCK OF DIFFERENT MEDIA CHANNELS



MODEL BUILDING

Simple linear model : The basic linear model can capture the Current effect. This model can be estimated using Multivariate linear regression method.

$$Y = \alpha + \beta 1At + \beta 2Pt + \beta 3Dt + \beta 4Qt + \beta 5Tt + \epsilon$$

Multiplicative model : Linear model assumes an additive relationship between the different KPIs. In such a case we use multiplicative model. To fit a multiplicative model, take logarithms of the data(on both sides of the model), then analyse the log data as before.

$$Y = e^{\alpha} * At^{\beta 1} * P_t^{\beta 2} * D_t^{\beta 3} * Q_t^{\beta 4} * T_t^{\beta 5} * \epsilon$$

After applying log → $$\ln Y = \alpha + \beta_1 \ln(At) + \beta_2 \ln(Pt) + \beta_3 \ln(Dt) + \beta_4 \ln(Qt) + \beta_5 \ln(Tt) + \epsilon'$$

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.

$$Y = \alpha + \mu Y\ t\text{-}1 + \beta 1At + \beta 2Pt + \beta 3Dt + \beta 4Qt + \beta 5Tt + \epsilon$$

Distributed Lag Model (Additive) : The Distributive Lag Model(Additive) helped us capture the not only the current, but also the carry-over effect of all the variables(depedent and independent).

$$Yt = \alpha + \mu 1Yt\text{-}1 + \mu 2Yt\text{-}2 + \mu 3Yt\text{-}3 + \ldots + \beta 1X1t + \beta 1X1t\text{-}1 + \beta 1X1t\text{-}2 + \ldots + \beta 2X2t + \beta 2X2t\text{-}1 + \beta 2X2t\text{-}2 + \ldots$$
$$+ \beta 3X3t + \beta 3X3t\text{-}1 + \beta 3X3t\text{-}2 + \ldots + \beta 4X4t + \beta 4X4t\text{-}1 + \beta 4X4t\text{-}2 + \ldots + \beta 5X5t + \beta 5X5t\text{-}1 + \beta 5X5t\text{-}2 + \ldots + \epsilon$$

Distributed Lag Model (Multiplicative) : The Distributive Lag Model(Multiplicative) will now help us capture current-to-current interactions and carry over KPIs effects. Take data logarithms (on both sides of the model) to fit a multiplicative model, then analyze the log data as before.

$$Yt = \alpha + \mu 1\ln(Yt\text{-}1) + \mu 2\ln(Yt\text{-}2) + \mu 3\ln(Yt\text{-}3) + \ldots + \beta 1\ln(X1t) + \beta 1\ln(X1t\text{-}1) + \beta 1\ln(X1t\text{-}2) + \ldots$$
$$+ \beta 2\ln(X2t) + \beta 2\ln(X2t\text{-}1) + \beta 2\ln(X2t\text{-}2) + \ldots + \beta 3\ln(X3t) + \beta 3\ln(X3t\text{-}1) + \beta 3\ln(X3t\text{-}2) + \ldots$$
$$+ \beta 4\ln(X4t) + \beta 4\ln(X4t\text{-}1) + \beta 4\ln(X4t\text{-}2) + \ldots + \beta 5\ln(X5t) + \beta 5\ln(X5t\text{-}1) + \beta 5\ln(X5t\text{-}2) + \ldots + \epsilon'$$

MODEL BUILDING

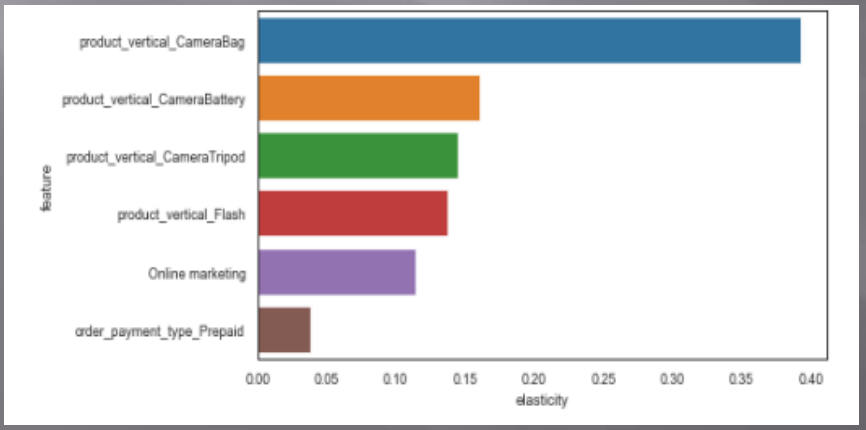
Flow for Building a model

- **Split the Dataframe into Training and Testing data sets**
- **Rescaling the Features using MinMax scaling**
- **For building our model, we will be using the LinearRegression function from SciKit Learn for its compatibility with RFE (which is a utility from sklearn)**
- **Building model using statsmodel, for the detailed statistics**
- **Calculate the VIFs for the new model**
- **Dropping features with high VIF**
- **Calculate the VIFs for the new model and repeat the flow untill the suitable VIF Value comes**
- **Check for Elasticity for finding the Top KPI's**
- **Finding the** Mean_Squared_Error Value

MODEL BUILDING CAMERA ACCESSORY AND RECOMMENDATIONS BASED ON ELASTICITY OF KPI'S

Model	Top KPI's	R Square	Adj. R Square	MSE
Simple Linear Model	product_vertical_CameraBattery+product_vertical_Strap+product_vertical_CameraTripod+order_payment_type_Prepaid + product_vertical_CameraHousing	0.956	0.941	0.0037
Koyck Model	product_vertical_CameraBag+product_vertical_CameraBattery+product_vertical_CameraTripod+product_vertical_Flash+Online marketing + order_payment_type_Prepaid	0.981	0.977	0.0011
Multiplicative Model	product_vertical_Lens+Discount%+Affiliates_SMA_5	0.975	0.973	0.0048
Distributed Lag Model(Additive)	product_vertical_CameraRemoteControl + product_vertical_Lens + product_vertical_Flash	0.963	0.950	0.0023
Distributed Lag Model(Multiplicative)	product_vertical_Filter	0.965	0.964	0.0050

Kyoc'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

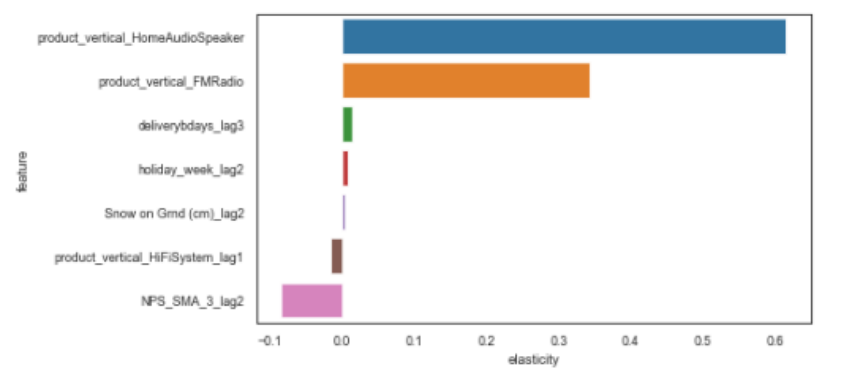


- The above figure represents the elasticity of different variables w.r.t. GMV
- Positive KPI means positive change in the KPI which leads to positive change in target variable
- From the graph, we can see that, the sales of CameraBag, CameraBattery, CameraTripod and Flash by order payment type Prepaid through Online marketing have positive impact on the GMV value.
- Hence, company can promote these products or pitch in more products in these categories

MODEL BUILDING HOME AUDIO AND RECOMMENDATIONS BASED ON ELASTICITY OF KPI'S

Model	Top KPI's	RSquare	Adj. R Square	MSE
Simple Linear Model	product_vertical_HomeAudioSpeaker+ product_vertical_VoiceRecorder+order_ payment_type_Prepaid+ Content Marketing_SMA_5+Sponsorship	0.971	0.963	0.0018
Koyck Model	product_vertical_HomeAudioSpeaker + product_vertical_FMRadio + product_vertical_VoiceRecorder	0.988	0.987	0.0006
Multiplicative Model	product_mrp+product_vertical_HomeA udioSpeaker+sla	0.993	0.993	0.0028
Distributed Lag Model(Additive)	product_vertical_HomeAudioSpeaker + product_vertical_FMRadio	0.995	0.994	0.0002
Distributed Lag Model(Multiplicative)	product_vertical_HomeAudioSpeaker	0.992	0.991	0.0033

Distributed Lag Model(Additive) is selected as the best model based highest R- square and low MSE values so that the company has one less feature to focus upon & grow almost equivalently.

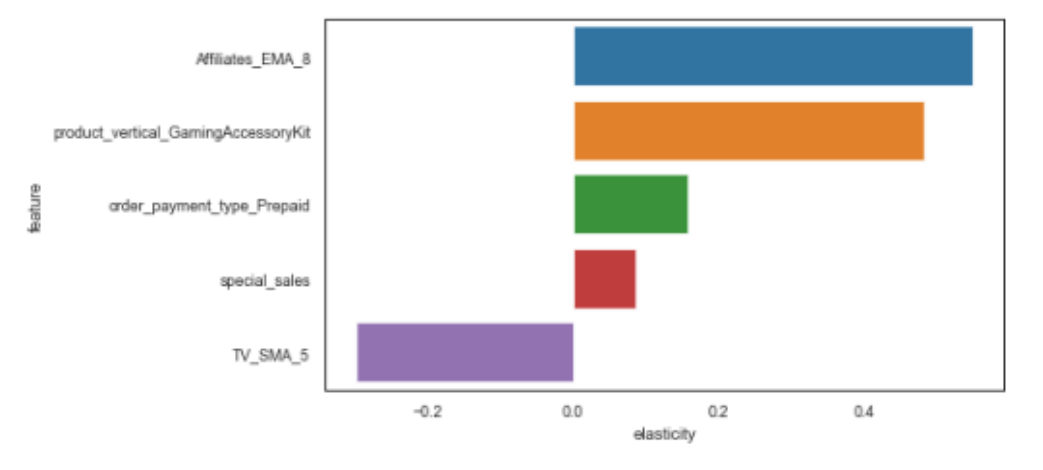


- The adjoining figure represents the elasticity of different variables w.r.t. GMV
- Positive KPI means positive change in the KPI will lead to positive change in target variable
- From the graph, we can see that the sales for HomeAudioSpeaker and FMRadio has positive impact on the GMV value.
- Hence,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

MODEL BUILDING GAMING ACCESSORY AND RECOMMENDATIONS BASED ON ELASTICITY OF KPI'S

Model	Top KPI's	R Square	Adj. R Square	MSE
Simple Linear Model	product_vertical_GamingAccessoryKit+Affiliates_EMA_8+special_sales+product_vertical_JoystickGamingWheel+ order_payment_type_prepaid	0.961	0.954	0.0026
Koyck Model	Affiliates_EMA_8+ product_vertical_GamingAccessoryKit+ order_payment_type_Prepaid + special_sales	0.980	0.976	0.0012
Multiplicative Model	product_mrp+product_vertical_TVOutCableAccessory+Content Marketing_EMA_8	0.966	0.963	0.0019
Distributed Lag Model(Additive)	product_vertical_GamingAccessoryKit + product_vertical_JoystickGamingWheel+ special_sales + order_payment_type_Prepaid	0.967	0.961	0.0025
Distributed Lag Model(Multiplicative)	product_mrp + product_vertical_TVOutCableAccessory	0.964	0.961	0.0021

Kyock 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 above figure represents the elasticity of different variables w.r.t. GMV
- Positive KPI means positive change in the KPI which leads to positive change in target variable
- From the graph, we can see that the sale of GamingAccessoryKit by order payment type Prepaid on special sales through Affiliates , has positive impact on the GMV value
- Hence, company should promote these products or pitch in more products in these categories

THANK YOU