iiit·b International Institute of Information Technology Bangalore

ज्ञानमुत्तमम्

UpGrad

CAPSTONE PROJECT

E-COMMERCE

Presented By-
Avanni Gudimetla
Vaishali Ramachandran

OBJECTIVE OF THE STUDY

- The objective here is to develop a market mix model for ElecKart to observe the actual impact of different marketing variables over the last year and recommend the optimal budget allocation for different marketing levers for the next year.

- We are required to create market mix models for three different product sub-categories -
  a)camera accessory
  b) home audio
  c) gaming accessories

- We need to observe the actual impact of different marketing variables for the year (2015 -2016) and recommend the optimal budget allocation for the various marketing levers for the next year to ensure that the budget is utilized effectively.

Below are the Steps we are following to achieve this objective :
1.Business Understanding and Data Cleaning
2.Exploratory Data Analysis
3.Feature Engineering
4.Model Building
5.Deriving insights based on the results
6.Recommendations to the business

BUSINESS UNDERSTANDING – AN OVERVIEW

- We are presented with Consumer data which indicates all the purchases made during the specified time period in addition to user specific details such as Order ID, zip code etc.

- Climate data in the Ontario region for the years 2015 and 2016

- media investment data which gives us a fair idea of how much has been invested in each of the channels such as TV, Radio etc

- NPS score data which gives us an idea of customer satisfaction which in turn determines the stock market price.

- We are also given the holiday list in that region and we also know that the pay dates are on the 1st and 15th of every month.

DATA CLEANING STEPS

The sole purpose of data understanding and data cleaning is to deal with inconsistencies in the dataset such as missing values, exponential values present in certain fields etc. We also convert the various datatypes to a common datatype which will help us in analysis.

- We convert the relevant fields to the datetime datatype.
- For the exponential values in the dataset, we remove them using encoding or by changing their display format to 3 digits after the decimal using the pandas set function.
- We will be dropping the columns like deliverybdays and deliverycdays which have more than 80% missing values, since they are of least significance.
- We do some preliminary analysis and calculations on the dataset.
- We impute all the missing values in the different datasets with appropriate values.
- Aggregate data on a weekly basis.
- Repeat the same steps for the climate dataset and the media investment datasets as well to bring it all into a common structure and merge the data frames together into one.
- Dropped duplicates and negative values present in the dataset for fields like gmv, product_mrp etc.

SIGNIFICANCE OF ATTRIBUTES IN THE DATASET

- 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: Suppose you order 2 different products under the same order, it generates 2 different order Item IDs under the same order ID; orders are tracked by the Order Item ID.
- GMV: Gross Merchandise Value or Revenue and this is the target variable
- Units: Number of units of the specific product sold
- Order payment type – Indicates the payment mode : prepaid or cash on delivery
- SLA: Number of days it typically takes to deliver the product
- Customer ID : Unique identification of a customer
- Product MRP: Maximum retail price of the product
- Product procurement SLA: Time typically taken to procure the product
- Monthly spend on various advertising channels
- Holidays and Paydays – Days when there was any special sale
- Monthly NPS score – Customer satisfaction index

KPIs AND EXPECTED RESULTS

- The target variable (independent variable) is the Gross Merchandise Value (GMV) which can otherwise be termed as revenue and we are required to predict the features that will be responsible for an increase in revenue.
- The dependent variables are the variables which will directly contribute to the increase or decrease in the GMV such as discount and discount percentage offered on a product, NPS score etc.
- We also calculate listing price which is the GMV / no.of units.
- We observe from our codes that variables like NPS score, holidays and paydates(1st and 15th of each month) positively affect sales.
- We need to find out which subcategory the company needs to invest in to improve its revenue in the next financial year under the 3 subcategories i.e, gaming, audio and camera.

FEATURE ENGINEERING

Feature Engineering is the process where we derive new features out of the existing data to add relevance and improve the quality of data analysis. In this study, we do the below as part of feature engineering :

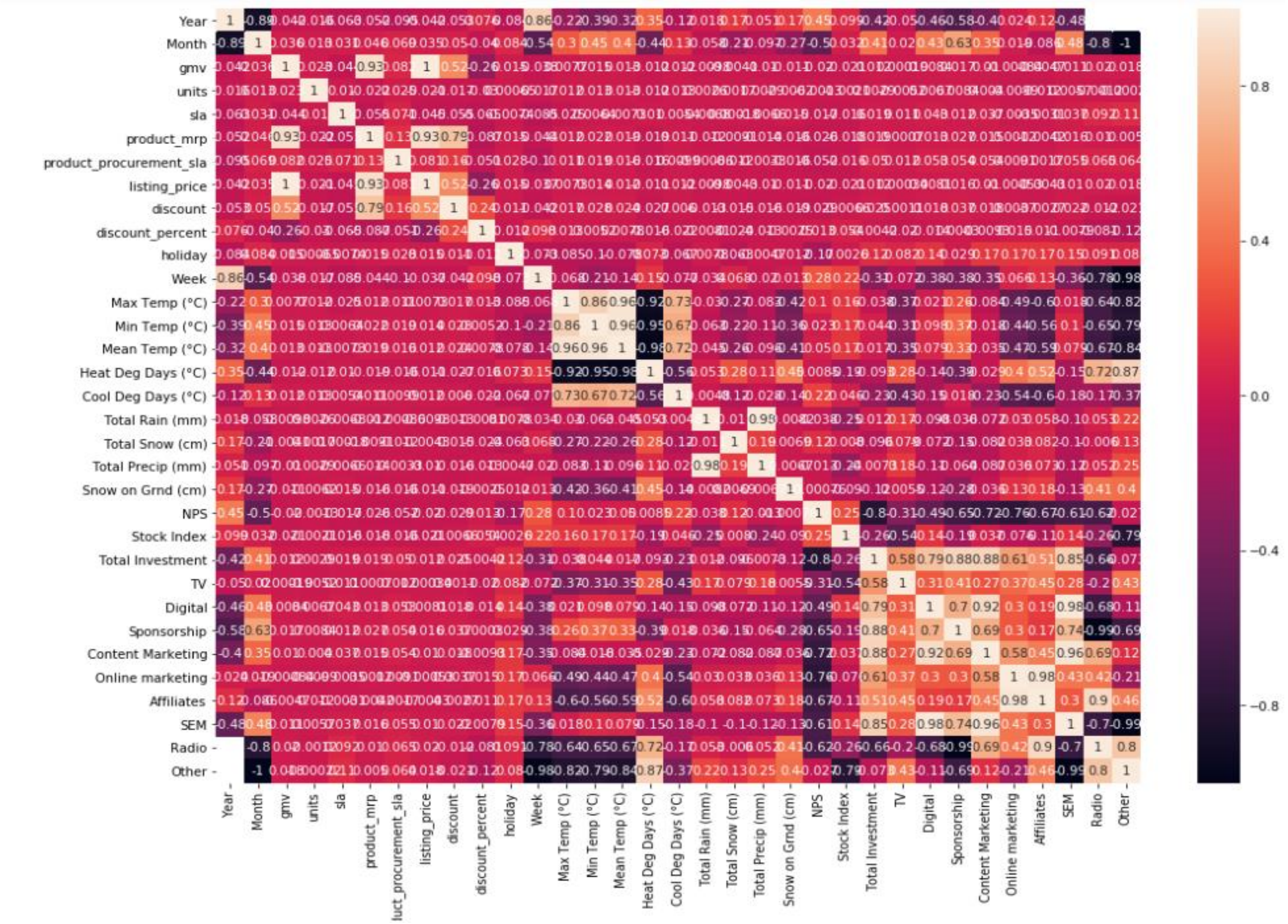
- We start with creating 3 dataframes by dividing them according to the relevant subcategories- one for audio, gaming and camera.
- Created a field for Adstock and lag based on the NPS Score.
- Created 2 new flags for holidays and paydays and indicated them as 1 on the 1st and 15th of the month and the rest as 0.
- Reduced the data and aggregated it at a weekly level after merging the different dataframes together.

LIST OF ENGINEERED KPIs

Below are some of the engineered KPIs that we use in this project :

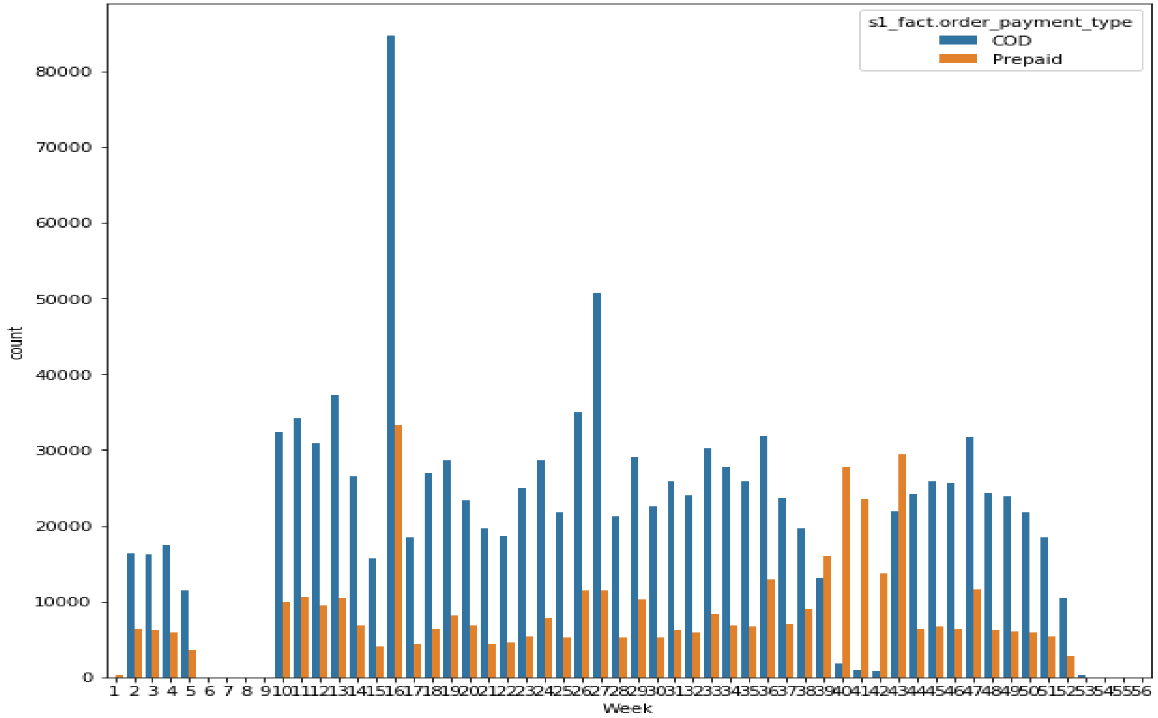
1. **NPS** : Net Promoter Score or NPS is a customer satisfaction benchmark that indicates the growth potential of the company.
2. **Listing Price** : This metric is calculated by dividing the GMV with the number of units.
3. **Discount** : Discount is calculated as the difference between the product MRP and the list price and then we calculate the discount % as the discount offered*100/Product MRP or discounted price*100/Listing price.
4. **Holidays** and **Paydays** : The trend tells us that people tend to shop a lot during holidays and paydays . This leads to an increase in sales and revenue on those days. We create a new column for both holidays and paydays and from the given holiday and payday dataset and we use a flag to identify and mark ‘1’ for holidays and paydays and ‘0’ for the other days.
5. **AdStock** : Ad Stock is the lagged effect of advertising on consumer purchase behaviour. It measures the decaying effect of advertising through the different weeks. Each time the consumers are exposed to the advertisement of the product on various platforms, awareness is created and awareness is higher for recent exposures and lower for the not-so-frequent ones. Adstock value diminishes to negligible levels beyond a certain span of no exposures and that is when the company needs to invest on advertising.
6. **ROI** : Return of Investment is the most common profitability ratio and it is calculated as net profit / total assets.
7. **Product Premiumness**: We have interpreted product premiumness in terms of 3 categories i.e, whether the product is a premium product, aspiring product or a mass product. This is achieved by performing k-means clustering on the productMRP and the no.of units sold and then based on the highest values of MRP in each cluster, we categorize it accordingly into one of the above mentioned groups.

EXPLORATORY DATA ANALYSIS



CORRELATION MATRIX :Checking the correlation only for the consumer data set. Varibles like Discount, listing price, product_mrp correlate with the target variable which is GMV.

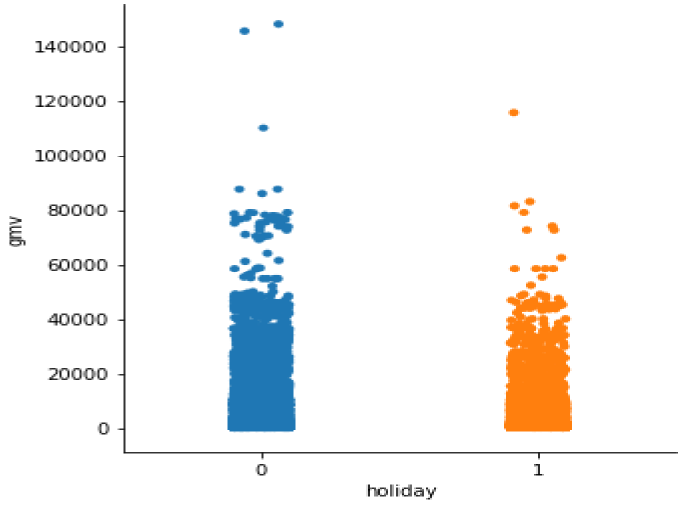
Displaying all graphs w.r.t Camera subcategory. Similar graphs have been displayed in the Python code file for the other two subcategories.



Graph indicating weekly revenue in COD vs. prepaid channels :

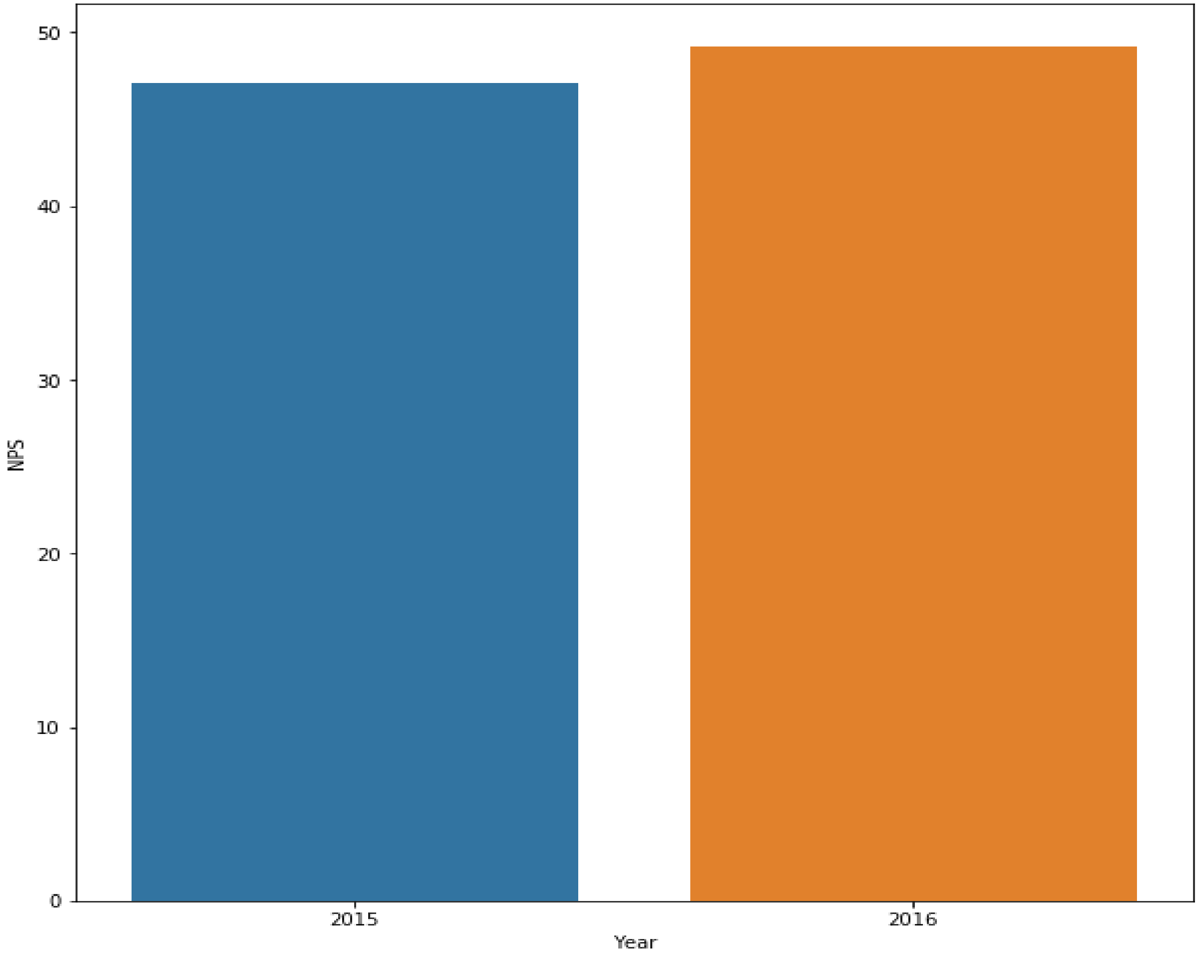
Observations:
COD is the most frequent option used by costumers on weekly basis.

week 6,7,8,9 have very less orders of COD and Prepaid and y-axis scale is huge and that is the reason it appears as if there are no results on the graph.

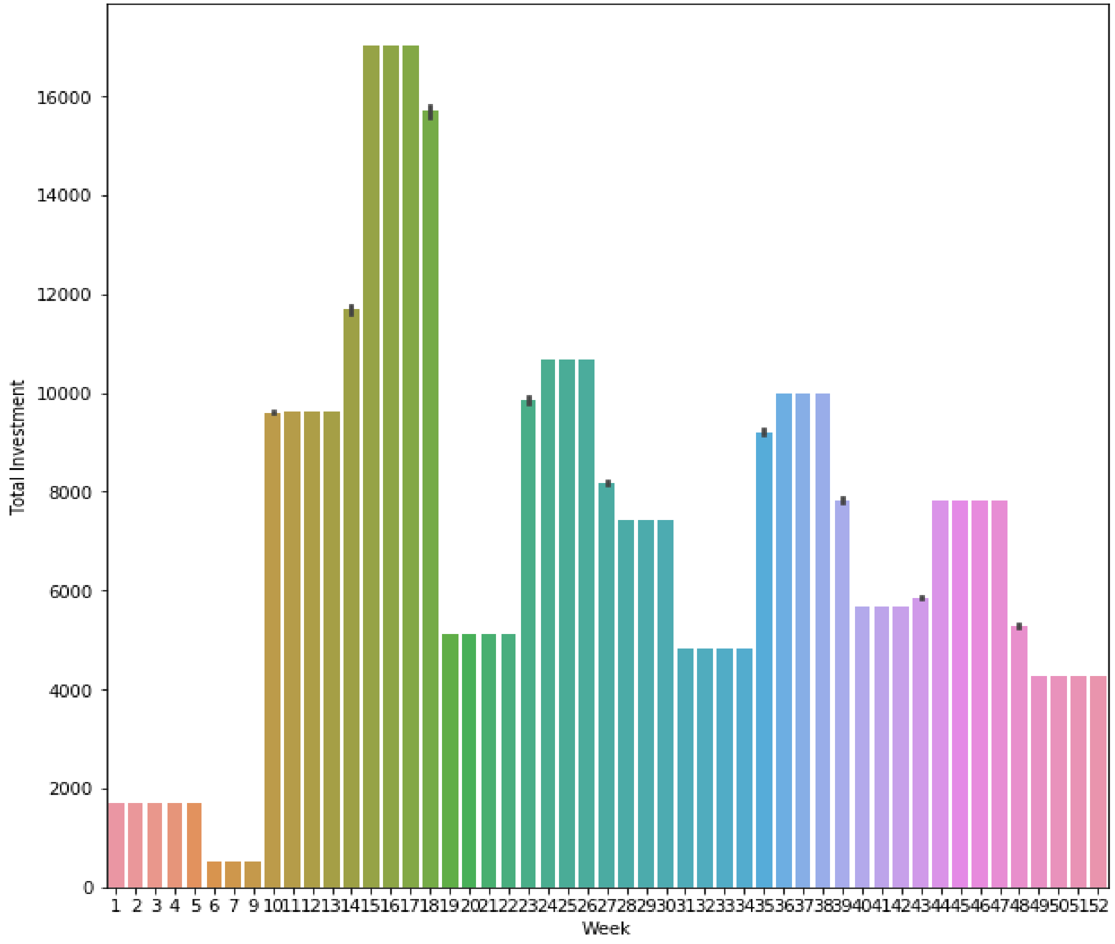


Graph which shows how holidays and Paydays affect GMV for camera subcategory

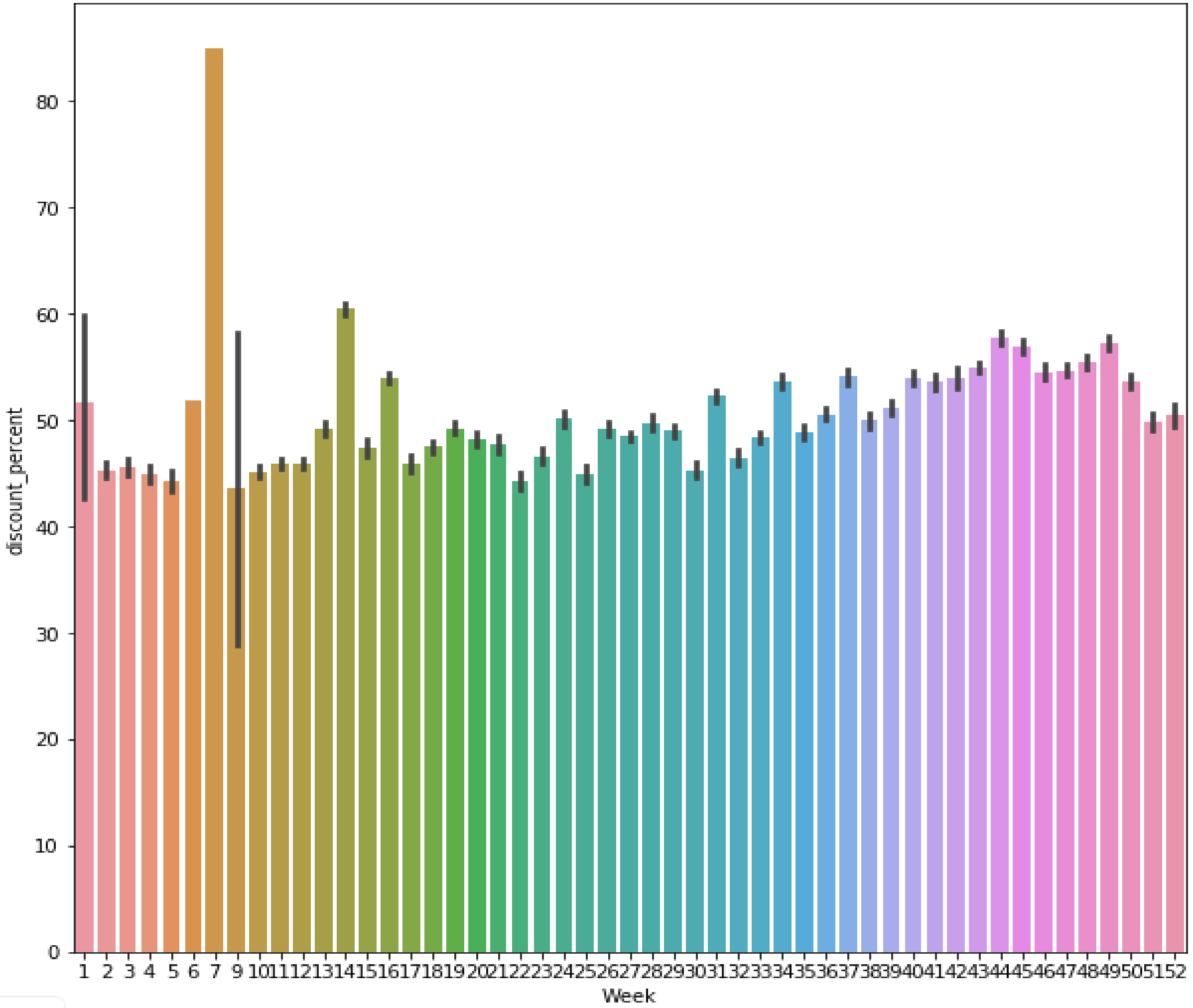
Graph displaying the NPS variation for camera subcategory
For both 2015 and 2016 (Years).



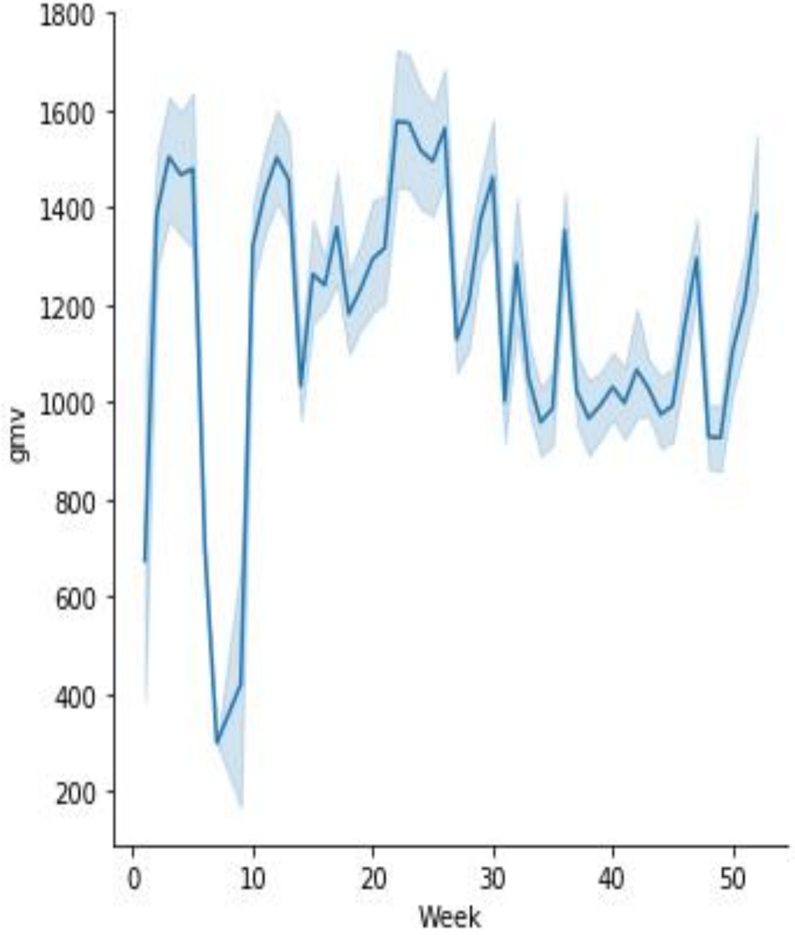
This diagram indicates:
During which week the total investment is the highest i.e analyzing total investment on weekly basis for camera
Observations: Week 15,16,17,18 has the highest investments in camera accessory.



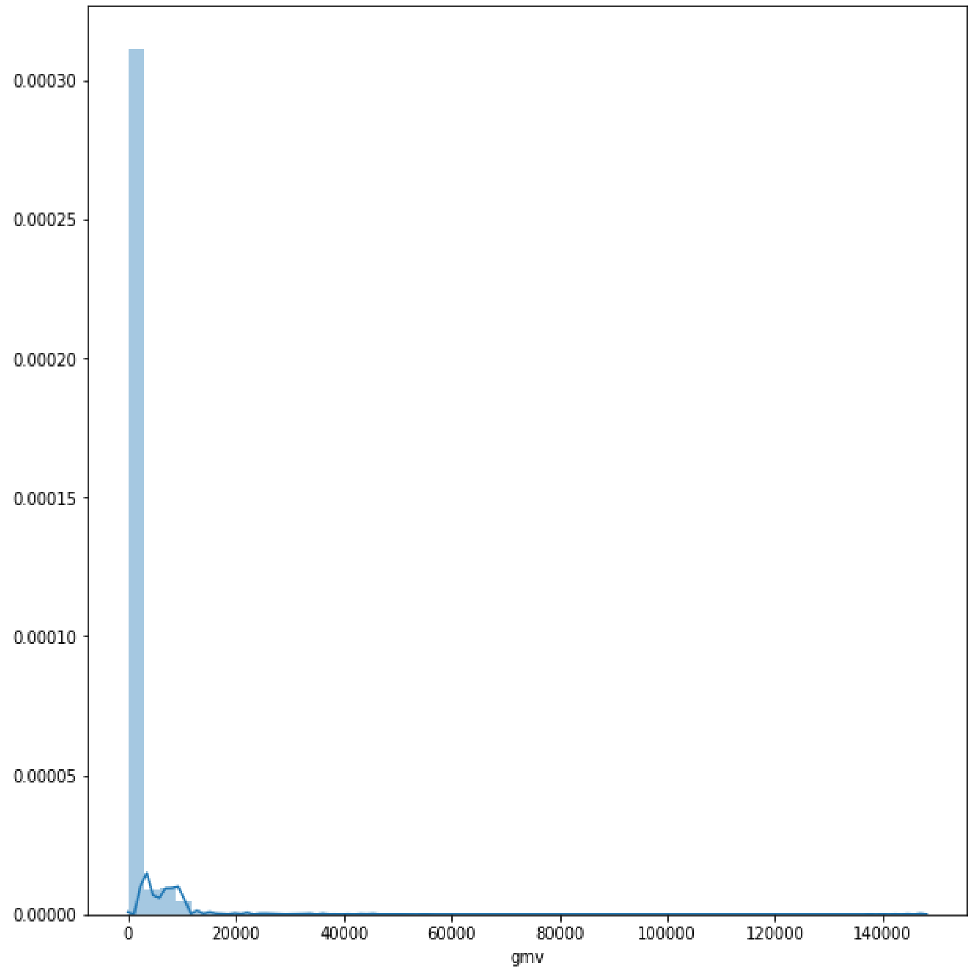
This graph indicates the discount percentage offered on a weekly basis for the camera subcategory



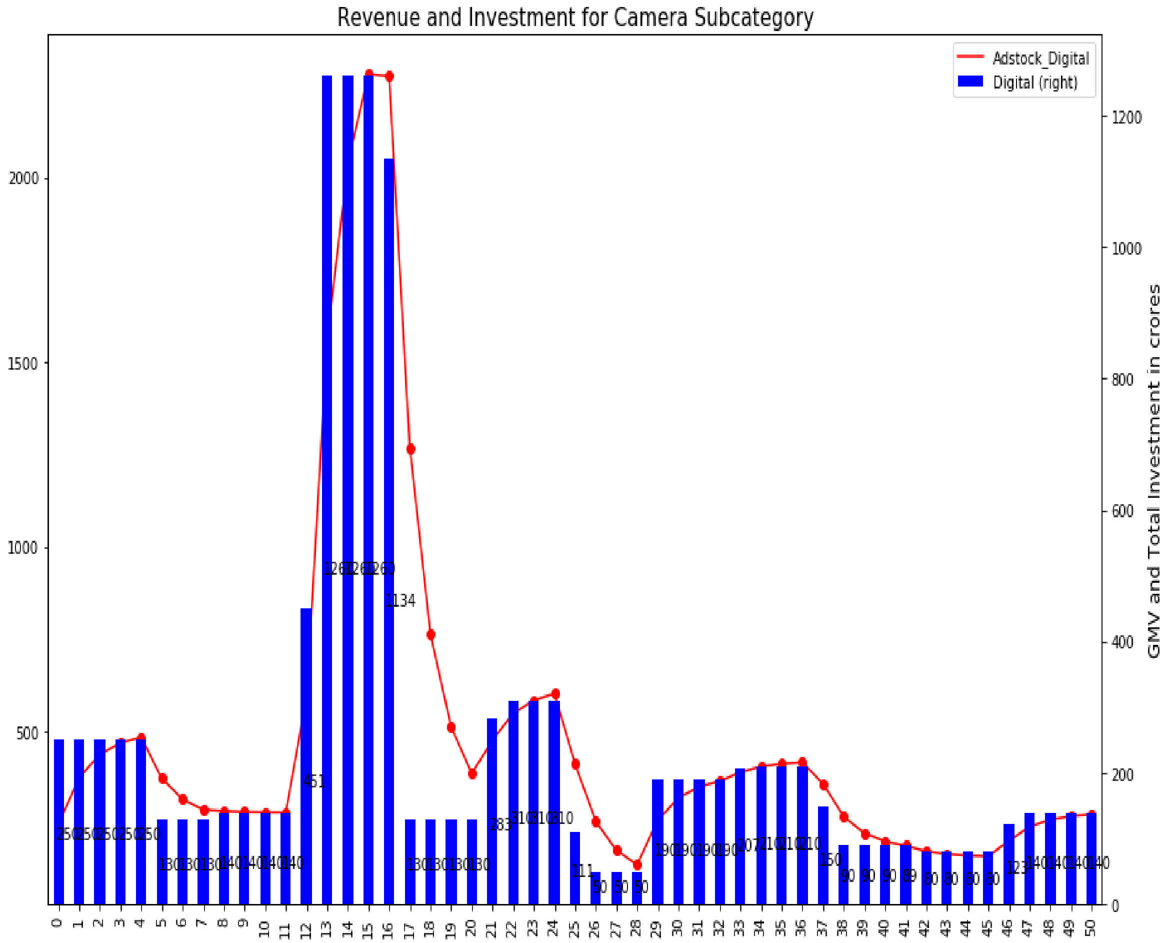
This diagram indicates the increase in revenue every week (indicates the GMV variation on a weekly basis)

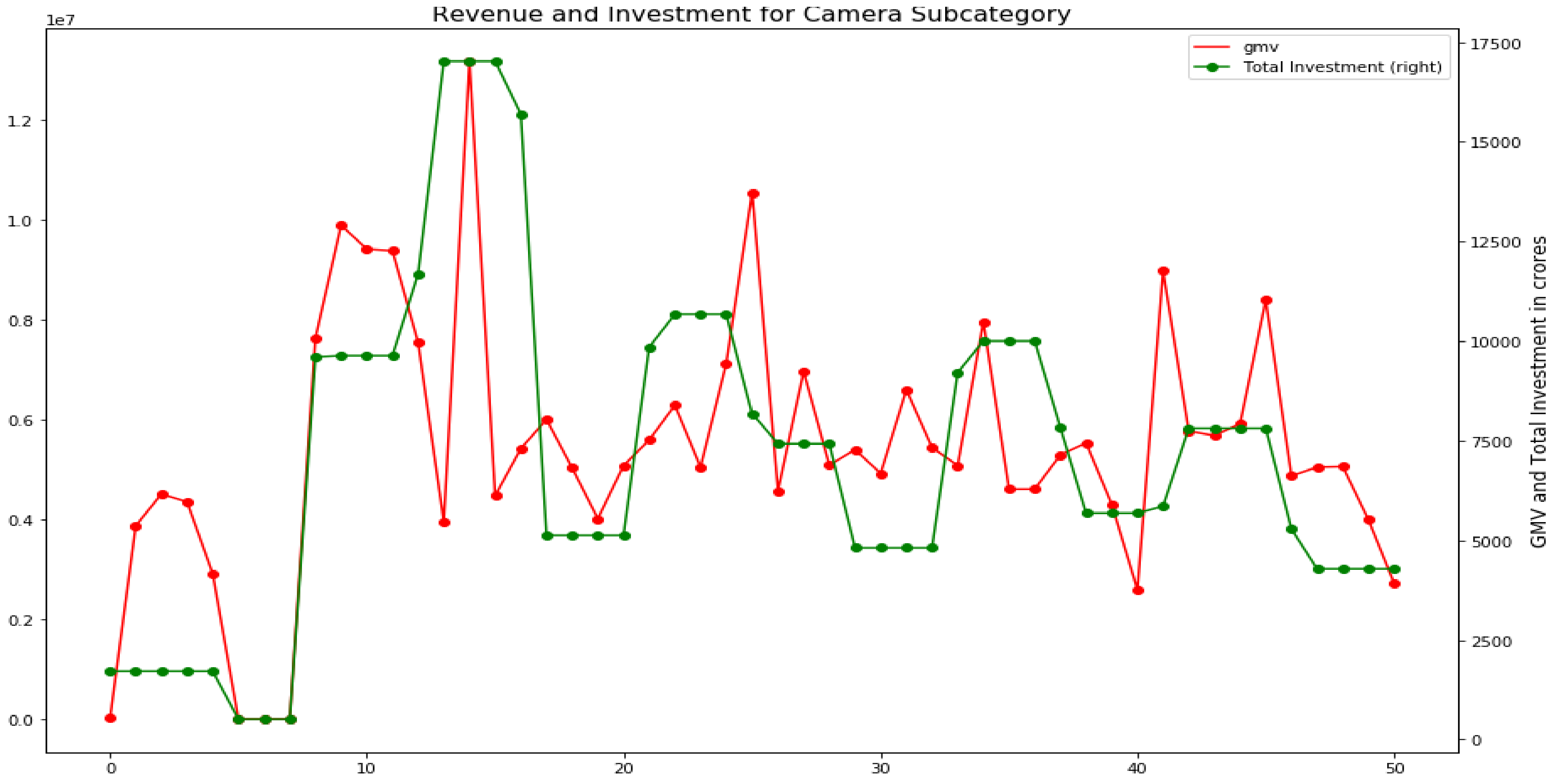


This is a density plot for camera subcategory which illustrates the distribution of data over a continuous time period. Target variable in camera sub category.

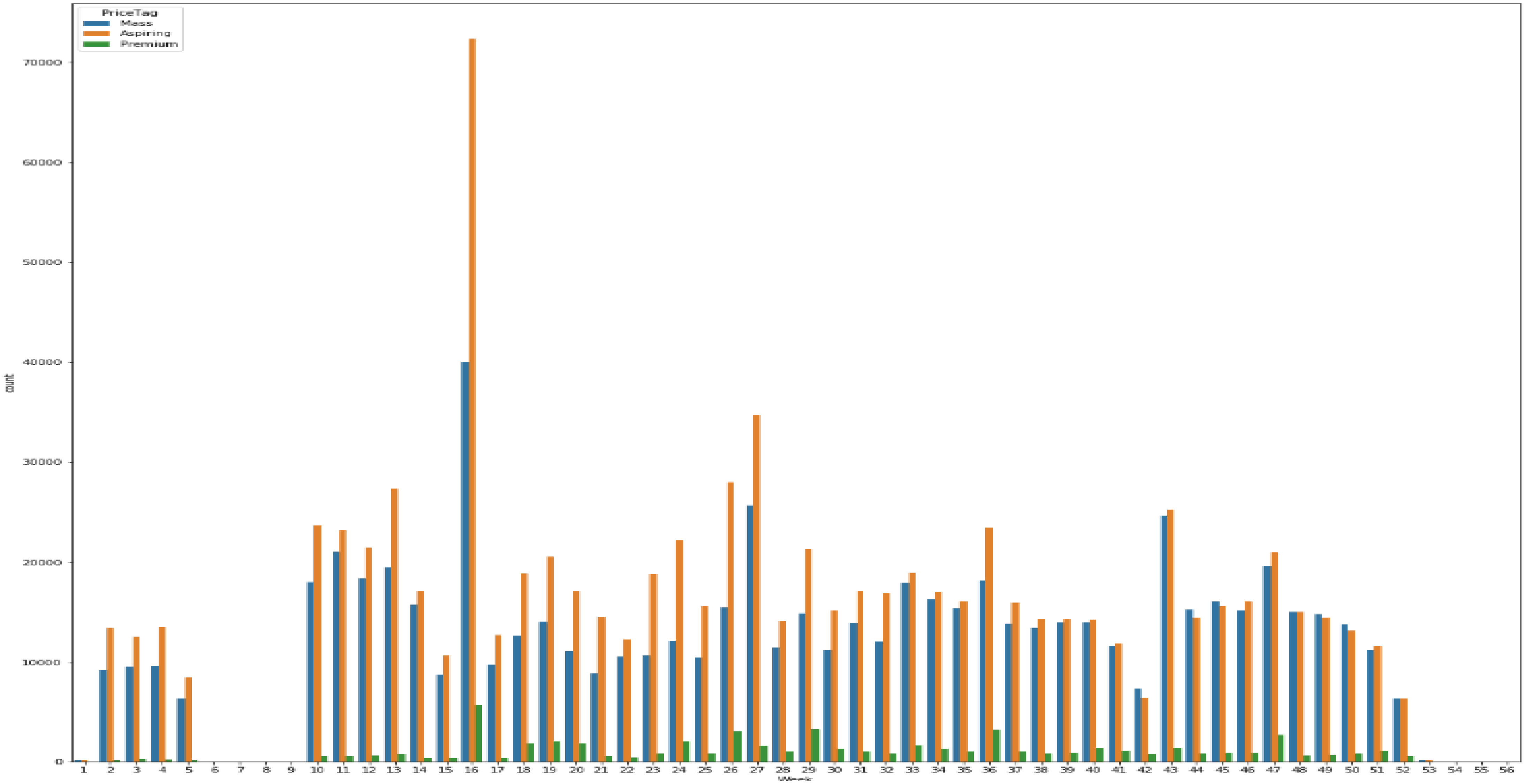


This graph is the adstock digital plot for camera which shows the rise and fall in revenue w.r.t the advertising on the digital platforms.





This graph illustrates how the revenue (GMV) increases with a change in total investment.



This graph illustrates the product premiumness feature and shows us that sales for Aspiring products is the highest across majority of the weeks and week 6,7,8,9 have very less orders of COD and Prepaid and y-axis scale is huge and that is the reason it appears as if there are no results on the graph.

MODELS BUILT

- **Simple Linear Models:**

linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

- **Multiplicative models :**

In this model, the independent factors are multiplied together to get the marketing mix. It includes the interaction of the independent variables through multiplication. In order to simplify the equation, we use logarithmic transformation to make it a linear model again. After converting it to the linear model, a multivariate linear regression can be used to estimate the alpha and beta values.

- **Distributed lag models:**

In the distributed lag model, both the dependent and the independent variables are entered in their lagged form.This is one of the models that help us capture the carry over effects of advertising and helps us model the revenue on the past figures of advertising and revenue spends.

- **Koyck models:**

This model is an extension of the basic linear model with time lagged version of the dependent variable is also considered as an independent variable in estimating Sales,Revenue and Traffic. It is a combination of AutoRegressive and Multivariate linear regression models.

MODELLING RESULTS - CAMERA

For the CAMERA subcategory, the below table collates the outcomes of all 4 of our models :

Model Name	Some Significant Variables	Rsq,Adj R-sq and R2 score	RMSE	MSE
Basic Linear model	Listing_price,discount_price,NPS,holiday_lag1	Rsq = 0.986 Adj Rsq = 0.984 R2 score = 0.63	0.132	0.017

Multiplicative Model	Stock_index,Adstock_TV,sla,discount_percent	R2score=0.62
		Rsq =0.905
		Adj Rsq =0.875
			1.171	1.372

Koyck Model	product_analytic_vertical_Filter,holiday_lag1,discount_percent	R2 score = 0.710 R sq = 0.954 Adj R sq. = 0.938	0.567	0.321
Distributed lag model	Units,discount_percent,sla,moving_avg_3	R2 score = 0.92 R sq = 0.982 Adj Rsq = 0.968	0.208	0.043

From the above table , For Camera subcategory we choose Basic Linear Model because of high R squared and low MSE.

MODELLING RESULTS - GAMING

For the GAMING subcategory, the below table collates the outcomes of all 4 of our models :

Model Name	Some Significant Variables	Rsq, Adj Rsq and R2 score	RMSE	MSE
Basic Linear Model	Holiday,moving_avg_2,nps_lag_2, product_analytic_vertical_TVOutCableAccessory	Rsq =0.975 Adj Rsq =0.965 R2 score =0.764	0.121	0.014
Multiplicative Model	Discount_percent,listing_price,moving_avg_2	R2 score=0.841 R sq=0.926 Adj Rsq=0.903	76.46	5.847
Koyck Model	Discount_percent,Adstock_TV,holiday,moving_average_3	Rsq = 0.985 Adj Rsq = 0.979 R2 score = 0.905	0.303	0.091
Distributed Lag Model	Holiday_lag1,Adstock_digital,Total investment	R2 score = 0.601 Rsq = 0.966 Adj Rsq =0.945	0.390	0.152

For the gaming subcategory, we choose Basic Linear Model because of high R squared and low MSE.

MODELLING RESULTS - AUDIO

For the AUDIO subcategory, the below table collates the outcomes of all 4 of our models :

Model Name	Some Significant Variables	Rsq, Adj Rsq and R2 score	RMSE	MSE
Basic Linear Model	NPS,disount_percent,holiday_lag_1,listing_price	R sq = 0.980 Adj R sq = 0.975 R2 score = 0.796	0.056	0.003
Multiplicative Model	sla,adstock_digital,discount_percent	R2 score=0.830 R sq = 0.934 Adj Rsq=0.906	8.756	0.766
Koyck Model	Listing_price,discount,discount_percentage,sla	R2 score = 0.956 R sq = 0.999 Adj Rsq = 0.998	0.127	0.016
Distributed Lag model	Disount,sla,listing_price,gmv1	R2 score=0.989 Rsq = 0.998 Adj Rsq = 0.997	0.042	0.001

For the audio subcategory, we choose the Distributed lag model because of the high R squared and low MSE.

RECOMMENDATIONS TO THE BUSINESS AND CONCLUSIONS OF THE STUDY

From the above study, we can draw a few conclusions and make some recommendations to the business :

1. On the whole the Distributed lag model, performs better in comparison to the other three models used in this study. This is closely followed by the Basic Linear model.
2. We also observe that the multiplicative model does not perform well in comparison with the other three models.
3. We consider a good model to have low MSE value and good R squared value and a high R2 score(1-RSS/TSS)

RECOMMENDATIONS:

1. For Gaming subcategory, NPS score and total investment have strong positive impact on the revenue and on the other hand, we see that Adstock_Digital has a negative impact on the revenue and our recommendation would be to invest the money spent on advertising on digital platforms into other channels.
2. For Camera subcategory, we see that discount,NPS and holiday have a positive impact on revenue and we also see that holiday_lag and sla have a negative impact.Our recommendation would be to not invest in these segments.
3. For Audio subcategory,we see that product_MRP and holiday have the strongest positive impact on the revenue.This means that customers are most likely to shop products of the audio subcategory during holidays and based on the product MRP. We also see that discount and NPS score have a negative impact on GMV.Hence, our recommendation would be to cut down on offering product discounts and instead market the products well during holidays or paydays to improve sales.