Customer buying Patterns in E-Commerce – An Emperical Analysis on Online Cart Abondonment

Sai Krishna Pedapudi SaiKrishnaPedpaudi@my.unt.edu
Sai kumar Reddy Reddymalla SaiKumarReddyReddymalla@my.unt.edu
Sribala Putcha SribalaPutcha@my.unt.edu
Jyothika Pagilla JyothikaPagilla@my.unt.edu

UNIVERSITY OF NORTH TEXAS

***Abstract*— Online cart abandonment is a key challenge facing e-commerce retailers leading to lost sales. This research aims to analyze customer behavior data from an e-commerce site data to uncover factors driving abandonment. Exploratory data analysis and statistical analysis are conducted with over 380,000 customer events. The analysis identifies product price, categories, brand, time of day, and customer demographics as key drivers of abandonment rates. based on the analysis, strategies can be formulated to predict users prone to abandoning and proactively convert them to customers. By characterizing abandonment behavior, the study aims to help e-commerce sites minimize lost sales and improve purchase conversion rates. The analytical approach developed can be extended to other sales funnels beyond shopping carts.**

Keywords—cart abandonment, customers,products,pricces

I. INTRODUCTION

A. MOTIVATION

In today's digital age, the world of e-commerce has transformed the way consumers shop. While online platforms offer unparalleled convenience and variety, they also face a unique challenge: cart abandonment. This phenomenon, where potential customers leave items in their cart without completing the purchase, has become one of the most significant pain points for online retailers. The underlying reasons can range from price concerns to brand preferences, decision fatigue, or even the sheer complexity of choices. For e-commerce businesses, understanding these reasons can unlock the potential to recover lost sales and optimize the shopping experience.

B. SIGNIFICANCE

Our Project deals on the Customer buying patterns which aims to understand the root causes of the Online cart abandonment that cause a huge impact on the e-commerce platforms. To solve this problem, we analyze Real customer data and we hope we find the root cause which is driving the Cart-abandonment. Our conclusions can help the platforms to increase their sale by knowing the customer's way of shopping. Even a small improvement in the conversion of abandonment into sale gives major boost to the companies. The conclusions drawn in this project enables the companies to optimize the inventory and sales techniques to acquire more customers and also increase their sales. We can also provide a way to improve the customer satisfaction by proving targeted measures. Our final result is aimed to provide value to both Companies and to Customers as with the improved Strategies enable the users to turn the views into sales.

C. OBJECTIVES

In this Project, Our main Objective is to identify the Buying patterns of the customers on e-commerce sites on the basis of the cart abandonment. To achieve this, we have secondary objectives which are crucial in determining the result of the project. They are as follows: -

1. Identifying the customer behavior by analyzing the buying pattern and shopping process.
2. Calculating the rate and instances of Cart abandonment which gives us a clear picture of the reason behind the cart abandonment.
3. Correlating some product features such as category, price, and brand etc. with the likelihood of a sale, where we can identify which attributes contribute to higher conversion rates.
4. Identifying shopping trends and sales peak allows us to understand how the customer shopping habits changes.
5. Understanding how pricing plays a key role in the sale of the product.
6. Identifying the customer behavior by conducting different tests and incorporating various techniques.

D. FEATURES

Our project uses multiple features to identify the customer buying patterns. They are: -

1. Analyzing customer online activity and pattern of purchasing.
2. Evaluating the Price sensitivity of the customers.
3. Evaluating the performance of the product.
4. Time series Analysis to analyze consumer activity over time.
5. Data Visualization to easily understand the scope and essence of the data.
6. Using Predictive models to analyze the buying patterns and changes in future.
7. Segmentation of Users based on the purchasing patterns and habits.
8. Using machine learning techniques to find in depth cause of the cart abandonment.

II. PRE-IMPLEMENTAION WORK

A. *RELATED WORK*

We are doing our project based on the works of some authors who contributed their time and effort in topics allied and related to the Customer buying patterns. We baselined their approach and implemented some features based on their work. Some of the works are as follows: -

[1]We examined the study of Mounika Veeragandham and her team members who examined how people change their online shopping habits in Covid pandemic and also analyzing the satisfaction of people in online shopping experiences. They found out that the people are keener in what they are buying. Our goal is to find out why people are leaving without buying.[1]

[2]We took inspiration from the work done by Thomas W. Dillon and Harry L. Reif who worked on the study "Identifying purchase perceptions that promote frequent e-commerce buying.". Their research surveys 190 adults and provide insights on the customers purchase perceptron's affect the e-commerce buying and its frequency. They discovered that younger audience are more inclined to shop online. Their research also deals in understanding the e-commerce behavior. Our project mainly takes inspiration from their method and focus on shopping experience role in the consumer behavior. we aim to build upon their work by analyzing how these factors make the customer abandon their carts.

[3]We also had a look into the research on predicting repeat purchase patterns in e-commerce by Ye Tian and others. Their work focuses on forecasting future purchases based on timing and frequency data. This provides a valuable perspective in our cart abandonment analysis as a repeated customer, or a regular customer can give us more data on why he is shopping repeatedly and how he is different from others who just views and doesn't buy. We feel that their study provides a practical model for predicting future purchases and also provides an insight on repeated customers. Our goal is to learn from their approach and findings to develop relevant models to understand the shopping behavior more accurately[3].

[4]We also been motivated from the research in the topic "A Study on E-commerce and Online Shopping: Issues and Influences." by Dr.Anukrati Sharma. This study deals with the change in buying patterns and preferences of the customers over time. This study also provides the analysis of customer trends and preferences. This work is particularly relevant to our project as it provides the analysis of trends and changes in consumer buying patterns. The authors suggestions for improving online shopping might serve as foundational recommendations in our project, particularly in the reducing the cart abandonment.

B. *DATASET*

Source: The dataset is procured from Kaggle and focuses on e-commerce events within an electronics store.

Here is the link of the dataset and it has 9 columns and 400000+ rows.

https://www.kaggle.com/datasets/mkechinov/ecommerce-events-history-in-electronics-store

Preliminary preprocessing steps are undertaken, which include addressing any missing values, categorizing certain variables, and normalizing the data for consistent analysis. i.e. We divided category code into category, sub-category and item name for better understanding and also removed the rows which has no data in any of the columns. The link below shows the dataset which is saved and uploaded after the data preprocessing steps mentioned above. After making the changes we are now having 12 columns and 300000+ rows.
https://www.kaggle.com/datasets/pssk12/customer-data

III. DETAILED DESIGN OF THE FEATURES

The following features provide good Insights into the behavior of customer, Product performances. These features play a keen role in the formulating the way we handle the project and drive us towards the project goal. Some of the Features in the project are :

1. LOADING DATASET AND PERFORMING BASIC PRE-PROCESSING STEPS:

We loaded the database into the workspace and started some basic preprocessing steps which includes extracting content of a column into multiple meaningful columns and also we removed all the rows which contains any null values in any of the columns. Also we dropped the whole rows which have some missing or improper data.

2. BASIC ANALYSIS

To know the structure of the data we displayed initial rows and columns and rows and also summary of the numerical columns. This helps us to ensure that the data is correctly structured and usable for further tasks.

3. DATA VISUALIZATION

Various plots such as bar charts, pie charts and histograms are used to visualize different variables of the data. The distribution of prices and the count of unique products by category. These visualizations help in understanding the essence and key variable and their distributions.

4. DATA TRANSFORMATION AND FEATURE ENGINEERING

The event time column is converted into date-time which allows us to perform more time-based analysis. Also, additional features are created to facilitate the analysis of user interactions. This enables us to do time-series analysis more effectively in a detailed and logical way and also prepare the dataset for advanced and complex algorithms.

5. EXPLORATORY DATA ANALYSIS

We have done Exploratory data analysis where we examined all the variables to derive meaningful insights. These analysis are essential to understand user behavior and performance of different categories.

6. TIME SERIES ANALYSIS:

We conducted Time series Analysis to examine event counts that are done on daily basis. Doing this helps us to identify the trends and patterns that are important in further analysis.

7. EVENT CATEGORY ANALYSIS

We organized the events by user actions such as views, carts and purchases. This categorization is used to calculate the Cart abandonment rate.

8. PRICE ANALYSIS

The products are compared with price categories which is then used in analyzing the abandonment rate among different prices. This will help us develop the conclusion on cart abandonment w.r.t Prices.

9. CORRELATION ANALYSIS

We computed a Correlation matrix and Visualized to identify relationships between different features. This can show some significant factors that influence user behaviour and decisions of purchasing.

10. CUSTOMER JOURNEY ANALYSIS

We have done the customer journey analysis by tracking the transition from viewing to adding items in the cart and making purchase at last. This analysis is important to understand the conversion of events.

11. PRODUCT AND BRAND ANALYSIS

We have done the product and brand analysis to analyze the popularity of the brand and conversion rates of products. This helps us to identify the products and brands with high sales. Ultimately this helps us to understand us how customers tend to buy popular brands vs less popular brands.

IV. IMPLEMENTATION

A. *DATA PRE-PROCESSING*

We imported the dataset into workspace and performed some data pre-processing steps.

First, We displayed the first few rows of the dataset to understand how the data is in the dataset.

Then, we printed the Dataset shape using df.shape.

Next we have printed the columns w.r.t the number of missing values

a) Removing Empty Lines: We dropped the lines which are empty from the dataset.

b) Filtering dataset: We filtered the dataset to include only rows with "category_code" containing three variables remaing all are dropped.

c) Now, we saved the updated data into a new dataset file.and from here onwards we will updating the changes to this dataset directly.

d) Data Splitting: We splitted the column named "category_code" into three different columns namely "main_category", "sub_category", "Item_Name" and saved.

B. *EDA*

xii. We have imported the updated dataset and loaded the file path.

xiii. We displayed the first few rows of the dataframe to understand the data.

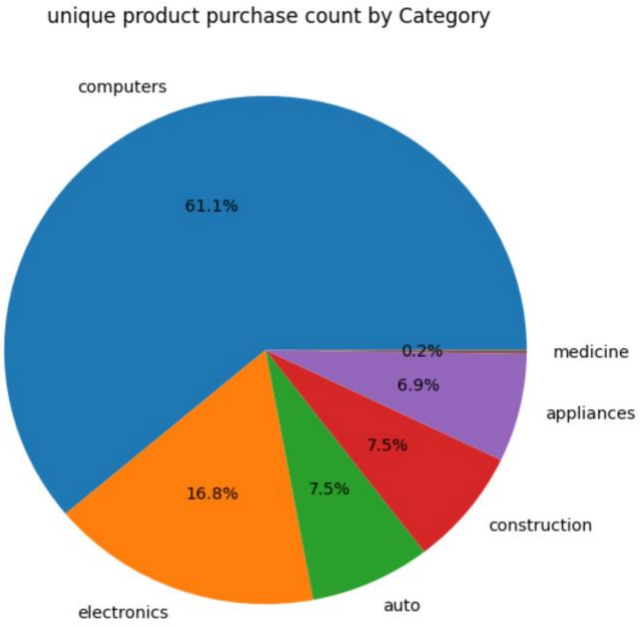
xiv. We displayed the basic information about the dataset also we made sure there are no null values in the dataset.

xv. We printed the Summary statistics of the Numerical data.

xvi. We printed the main category values which are unique and counted them.

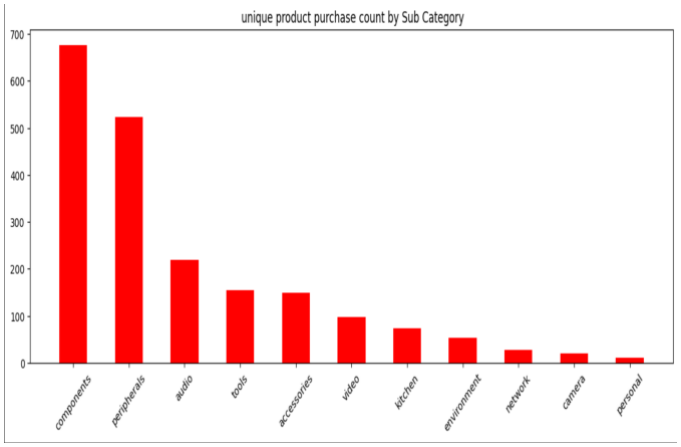
xvii. We also found the number of purchases done in main category.

xviii. We visualized the purchase count by each category in a pie chart.



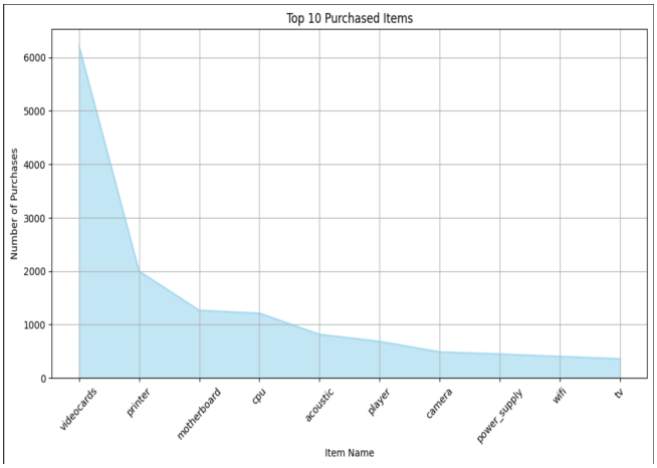
viii. We printed the sub category values which are unique and counted them.

ix. We then found the Unique product purchase count by Sub-Category by using a bar plot

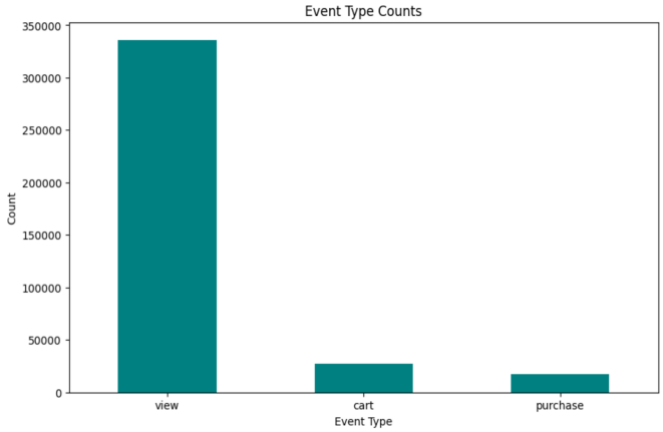


x. We grouped the purchased in Item_Names which are unique and also we found out the top ten unique purchases.

xi. We drawn a Area Chart for the top 10 purchased Items



xii. We then drawn a bar graph to see the distribution of event types.



C. *DATA TRANSFORMATION*

i. We created a copy of our dataset.

ii. We then divided the events type into their 3 different variables as columns.

iii. Now, we assigned the numerical values as 0,1 for every positive outcome among the 3 possibilities.

iv. We also counted and printed the number of events to identify the number of events.

v. Then, we took the 'event time' into date and time separately and printed it.

vi. We then found out the purchases by date and found out the number of unique purchases on that date.

D. *FEATURE EXTRACTIONS:*

xi. We calculated the number of cart events and purchase events.

xii. Then we found out the cart abandonment rate by using the formula.

Cart Abandonment Rate = (Cart - Purchase) / Cart

iii. We got the output as 0.37 which means 37% of the items in the cart are abandoned.

iv. Then we counted the dates from the starting date as 1 to 7 and assigned separately in the dataset.

v. Then, we assigned the numbers with the day of the week.

vi. We also calculated the purchases by hour of the day

vii. We also printed the prices of the purchased products.

viii. Now, we created a data frame for Cart events and another for purchase events and compared the average price of the products in carts and purchased events.

ix. We then compared the top categories and brands in cart and purchased events.

x. From the above comparisons we got the following outputs:

Average Price in Cart: 234.20,
Average Price Purchased: 217.81
Top Categories in Cart:
computers.components.videocards:11192 computers.peripherals.printer:2521 computers.components.cpu:2241 computers.components.motherboard:2113 electronics.audio.acoustic: 1074
Top Categories Purchased: computers.components.videocards: 6220 computers.peripherals.printer : 1999 computers.components.motherboard : 1266 computers.components.cpu : 1213 electronics.audio.acoustic : 817
Top Brands in Cart: gigabyte : 3152
msi : 3135
palit : 2241
asus : 2110
amd : 1896
Top Brands Purchased: msi - 1788
gigabyte : 1741
asus : 1250
palit : 1077
amd : 1020

xii. We counted the number of cart events and purchase events per user.

xiii. We also added a column for the abandonment rate per user by adding user_behaviour in both ascending order and descending order.

USER ID cart, purchase, view abandonment_rate
1515915625567798594 1.0 0.0 2.0 1.0
1515915625572789431 1.0 0.0 4.0 1.0
1515915625572589989 1.0 0.0 3.0 1.0
1515915625572638732 1.0 0.0 2.0 1.0
1515915625572657643 1.0 0.0 4.0 1.0
User id. cart purchase view abandonment_rate
1515915625523761189 1.0 11.0 5.0 -10.0
1515915625520679716 1.0 10.0 12.0 -9.0
1515915625561999052 1.0 10.0 8.0 -9.0
1515915625592612760 1.0 8.0 6.0 -7.0
1515915625537630874 1.0 8.0 19.0 -7.0
1515915625562163238 1.0 7.0 7.0 - 6.0
1515915625603460163 1.0 7.0 108.0 -6.0

1515915625537032857 1.0 7.0 4.0	-6.0
1515915625379210214 1.0 7.0 8.0	-6.0
1515915625547773774 1.0 7.0 9.0	-6.0

xiii. Then, we grouped by category calculated abandonment rate and displayed the categories with highest abandonment rates

cart	Purchase	view	Abandonment rate
1	0	59	1.0
2	0	24	1.0
1	0	32	1.0
1	0	75	1.0
1	0	418	1.0
2	0	96	1.0
16	4	802	0.75
54	16	1802	0.70
6	2	192	0.66

xiv. We categorized the prices into low,medium,high and very high and then we calculated the abandonment for each category and displayed it.

price category	Low	Medium	High	Very High
abandonment rate	0.290371	0.300756	0.387698	0.436476

xv. We used group.by for hour of day, price category and event type and calculated the time price abandonment rate.

hour_of_day	price category	abandonment rate
0	Low	0.431373
	Medium	0.441176
	High	0.383333
	Very High	0.392308
1	Low	0.523810
	Medium	0.516667
	High	0.475728
	Very High	0.523256
2	Low	0.510204
	Medium	0.525424

xvi. We used group.by for category code brand and event type for brand category abandonment rate

brand	NaN
ballu	NaN
edisson	NaN
electrolux	NaN
engy	NaN
first	NaN
hyundai	- 2.0
irit	NaN
magnit	NaN
neoclima	NaN
noirot	1.0

xvii. We defined session length by grouping user session and event time and calculated the abandonment rate for it and printed the output.

session_length	abandonment_rate
(-0.001, 156.0]	0.579832
(156.0, 2021.0]	0.299715
(2021.0, 86396.0]	0.280296

xviii. Now, we defined customer journey by grouping user session and event type and calculated the journey of the customer from viewing to adding to cart and purchasing. Finally, we calculated the mean of the journey.

view_to_cart	0.000042
cart_to_purchase	0.018343

xix. We then defined the Item popularity by grouping price and event type and then we calculated the conversion rate from view to purchase and printed the output.

brand	cart	purchase	view	conversion
marvo	1.0	1.0	3.0	0.333333
motorola	6.0	8.0	25.0	0.320000
foxline	2.0	3.0	10.0	0.300000
continent	3.0	6.0	27.0	0.222222
atrix	1.0	1.0	5.0	0.200000
alphachem	4.0	3.0	17.0	0.176471
rus	2.0	4.0	27.0	0.148148
content	29.0	24.0	166.0	0.144578
hipro	4.0	4.0	28.0	0.142857
hgst	8.0	8.0	56.0	0.142857

xx. We then defined the Item popularity by grouping price and main category and then we got the output

main category	cart	purchase	view	Result
computers	22113.0	13491.0	209686.0	0.064339
auto	1262.0	977.0	27750.0	0.035207
construction	840.0	615.0	17509.0	0.035125
electronics	2291.0	1629.0	54881.0	0.029682
appliances	781.0	607.0	25736.0	0.023586
medicine	9.0	7.0	305.0	0.023586
furniture	3.0	0.0	144.0	0.000000

xxi. We defined the event count and created a series from views and carts and also we defined product event story and user event story and from that we got the following output.

max purchases by product	564
max carts by product	1033
max views by product	5721
max purchases by user	56
max carts by user	68
max views by user	572

FURTHER METHODS:

i. Confidence Interval:

Here the confidence interval is used to evaluate the Average Purchase price were 135.517 is lower interval and Upper Interval is 138.964 were it lies in between them and also evaluated the Conversion rate interval which is 0.0417 and upper is 0.0426

Confidence interval for avg_purchase_price:
(135.51727562664067, 138.96436203201088)
Confidence interval for conversion_rate:
(0.041773923825841074, 0.042611516036597054)

ii. Boot Strapping:

The Boot Strapping is one of the resampling technique used to statically analyze the price distribution and constructed the confidence interval were the average price is 135.5167

iii. SVM:

Support vector Machine (Svm) is one of the classifier used to find the accuracy,F-value,precision of categorical variables item name which comes on main category and depending on price, user id were this model gives the 89.6% accuracy and it gives precision values of view 0.90 and Cart, Purchase are zero and recall,F1 score are also same states that the model performs best on categorical variable View. It means the 89.6% user are not purchasing and not adding to the cart.

iv. Random Forest:

Random Forest is also one of the ML Model used to reduce the over fitting and making less noise which is efficient for large dataset and extensive tuning. Here we the model is used to evaluate the Accuracy of users buying the product were the model gives the accuracy of 89.3% and precison of 'View ' is 0.90 and recall is 1.00 and f1 score is 0.94.

v. Decision tree:

Decision tree is used a classifier to evaluate the user accuracy to know the accuracy of purchasing the product were the model gives the 89.3 accuracy and the model performed well on View variable.

vi. Neural Networks:

Lstm (Long-Short Term Memory): This model performs well compared to the other model which gives the 95.8% accuracy were performing on 10 epochs were evaluated the cart is abandoned .The model states that 5% of users are purchasing the products added to the cart.

vii. Resampling:

Resampling is used to find the accuracy of users purchasing and not purchasing so I have used logistic regression model to evaluate the accuracy which is 52% were the precision and f1 scores are balanced and the model struggled to perform to distinguish the purchased events on the resampled dataset

viii. PCA:

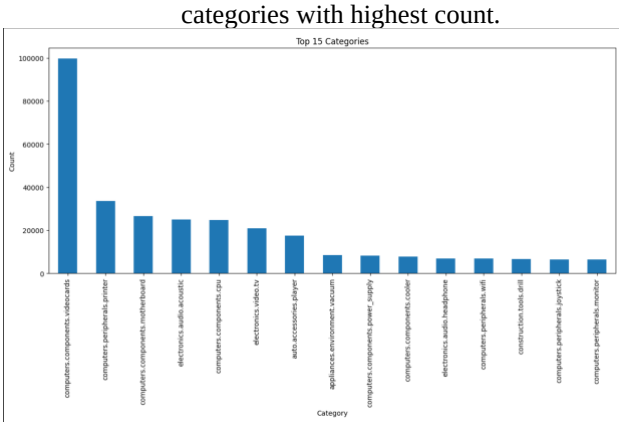
The principal component analysis is used to extract the main features on numerical data which is price ,user id ,category id and product id in this we analyzed the explained variance ratio and visualized data points it by using Scree plot and Pca features on Biplot.Finally analysed the original features impacting on features extracted.

E. *DATA VISUALIZATION*

i. We have drawn a hist plot to plot the price distribution of the data.

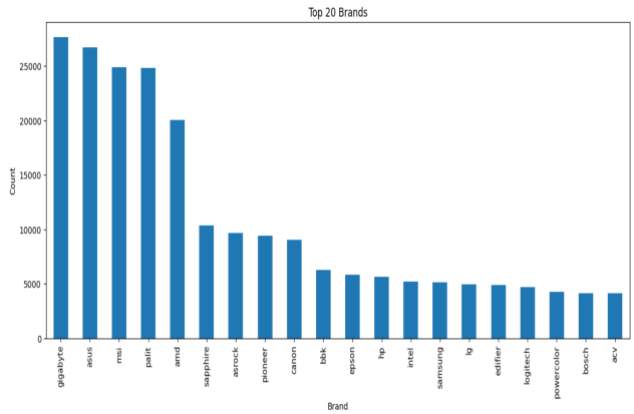
From the above graph we can see that the price in the dataset varies from 0 $ to 6400 $

ii. We have drawn a bar plot to plot the top 15 categories with highest count.



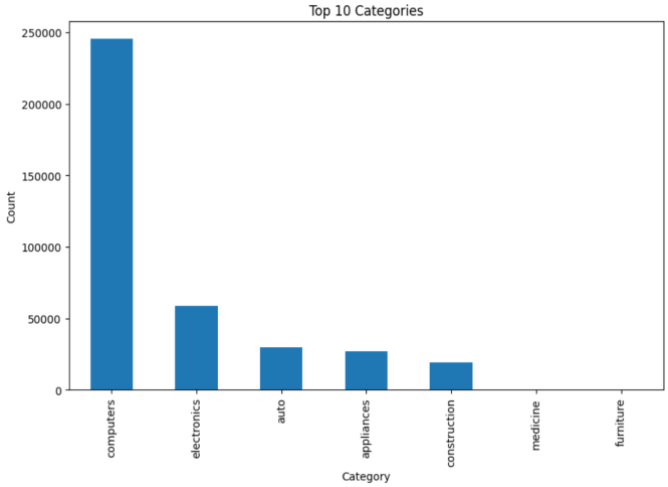
In this graph we have visualized the top 15 most popular categories

iii. We have drawn a bar graph for top 20 brands with highest count.

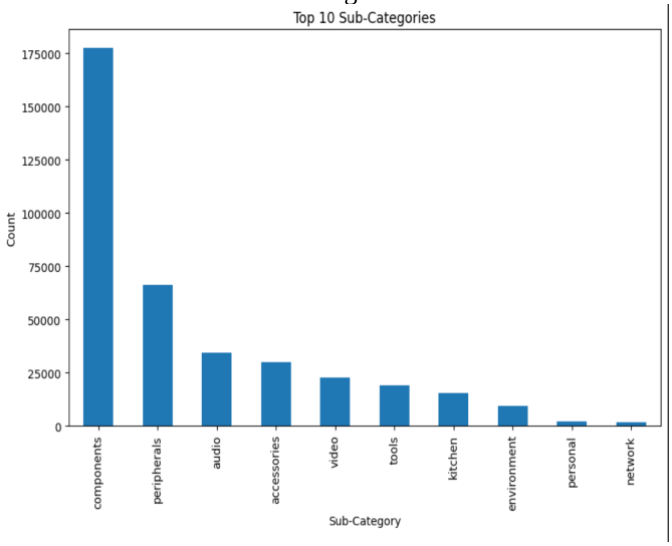


In this graph we have visualized the top 20 most popular brands.

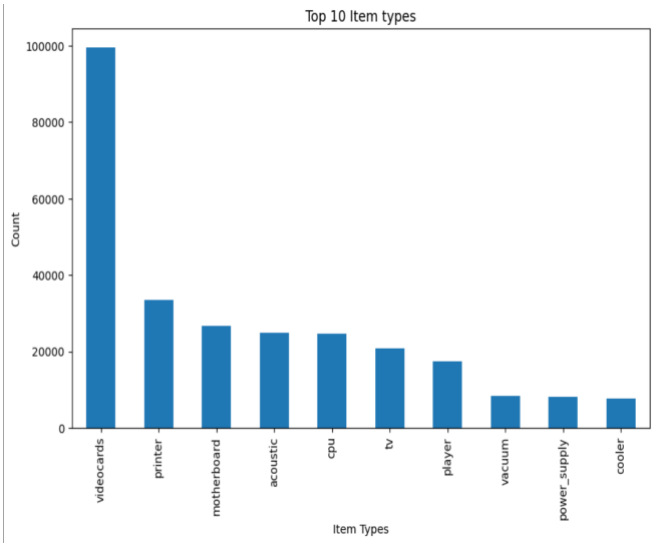
iv. We have drawn a bar graph for plooting the Top 10 main categories



v. We have drawn a bar graph for plotting the top 10 Subcategories.

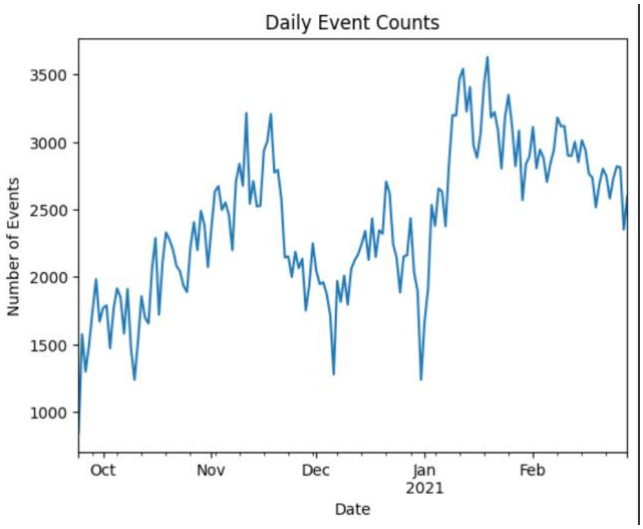


vi. We have drawn a bar graph for plotting the top 10 Item types.



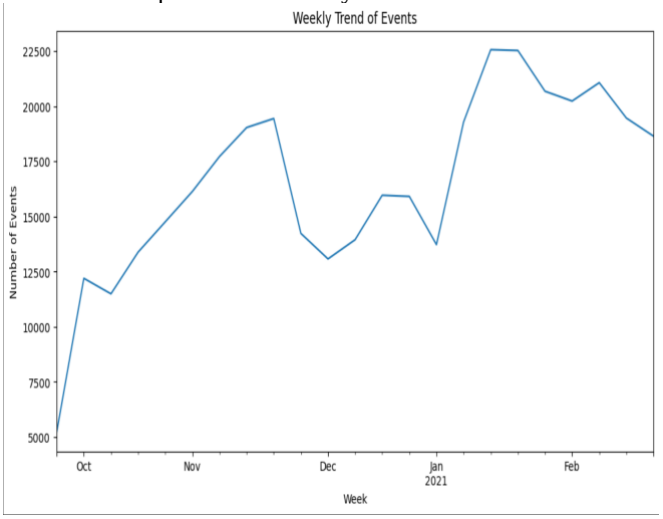
From above three graphs we have found Top 10 most popular Main categories, Subcategories and Item Names. All the above tasks allow us to understand the data properly.

vii. We have plotted a time series plot to visualize the daily events count from the dataset. We do this by plotting both the date and number of events in total on both the axis.

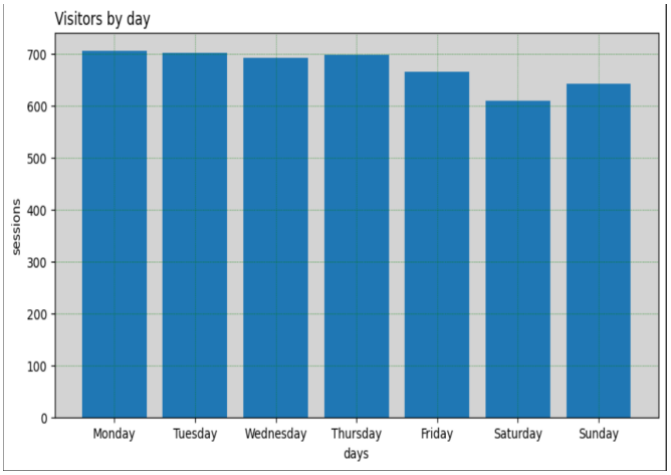


viii. Similarly, we have also plotted the number of purchases by date which is very helpful in understanding the sales of the products.

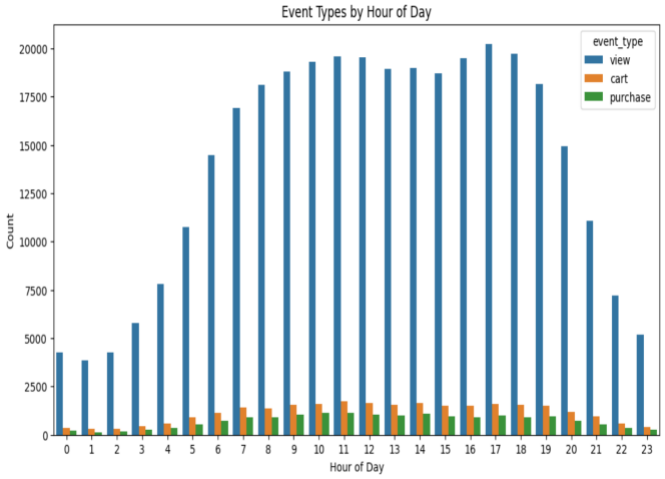
ix. We also plotted the weekly trend of events.



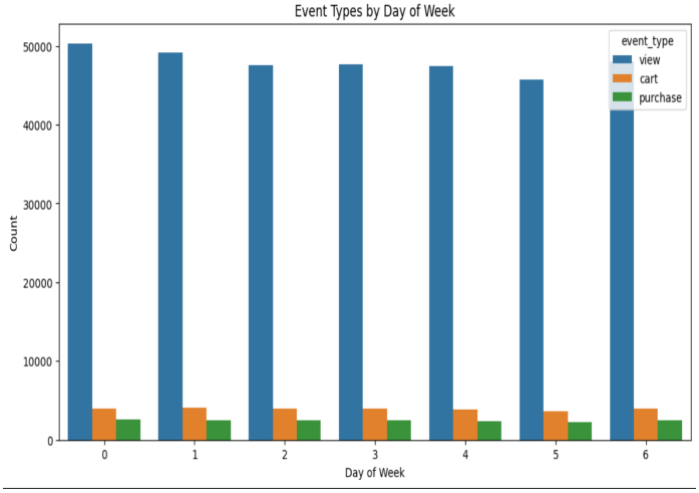
x. We have also plotted a bar graph which shows the number of unique vistors per day in a week.This chart helps us to identify the customer traffic day wise in a week. This is a very important step in determining the future of our project. This also helps us to determine the customer shopping trends and patterns that changes with the day of the week.



xi. We have plotted a Count plot which helps us visualize the Event types by hour of the day. We have done this by extracting the hour of the day and day of the week from event time and then we have plotted event types by hour of the day

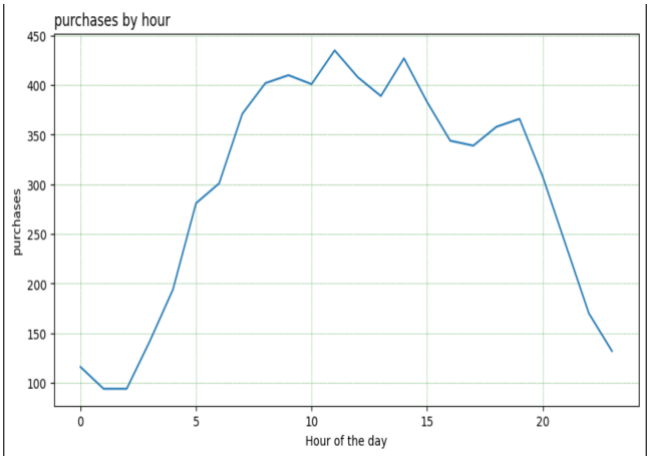


xii. Similarly, We have plotted the Event types by day of the week.

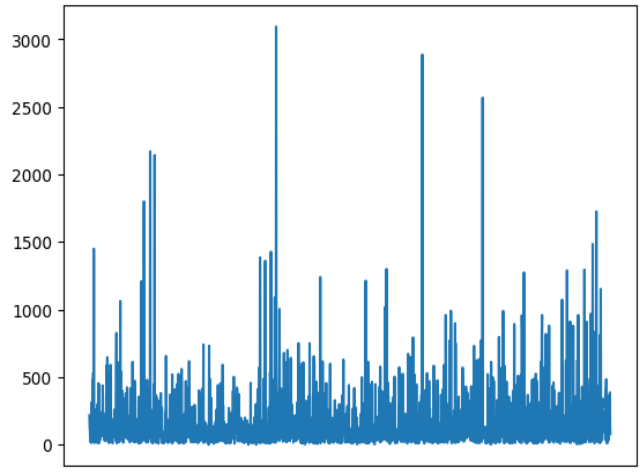


Both these graphs help us to understand the type of events that are occurred throughout the day and thorough the week respectively.
From both the graphs we can clearly see that the viewing the products by the customer is significantly much higher when compared to other two events.

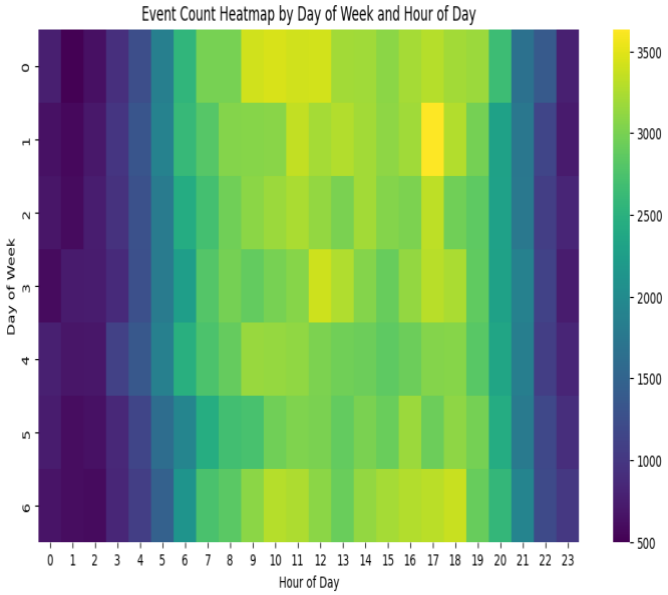
xiii. We have also plotted a graph which shows the unique purchases by hour of the day.



xiv. We then plotted a graph which shows the vertical distribution of the prices of the products.

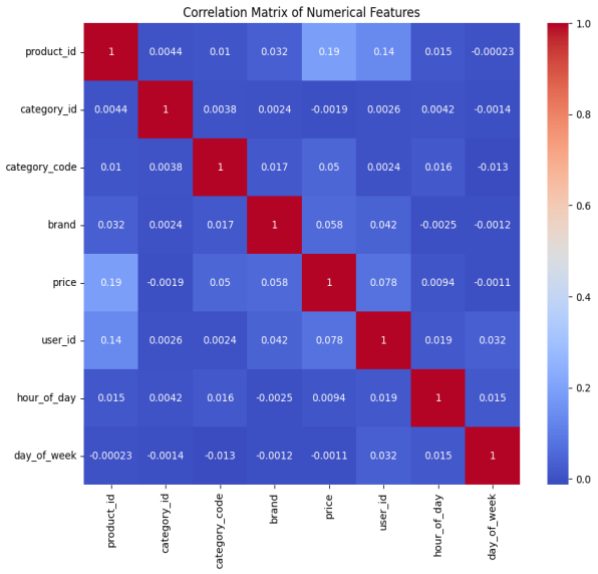


xv. Then we plotted a heat map which shows the event count by day of week and Hour of day.

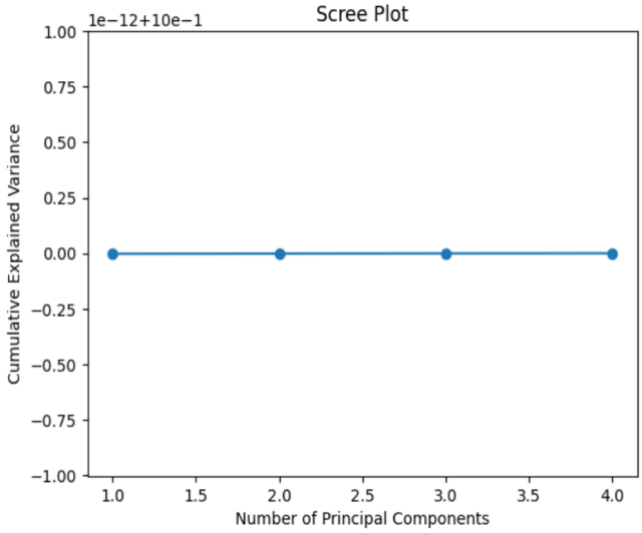


This heatmap clearly show the times of the day in which the customers foot fall is more and when its less.
From the output, we can clearly see that there is no much difference between the days of the week but if w observe clearly we can identify clear increase and decrease of the events at the starting and ending of the day. Also with careful observation we can also tell much more insights on this.

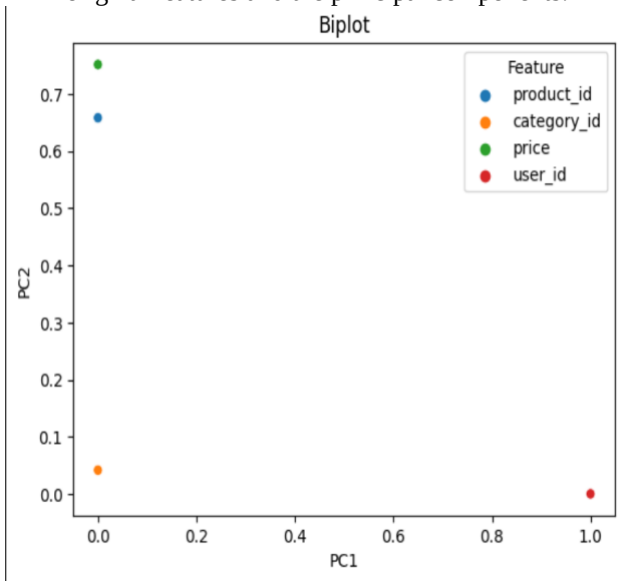
xvi. We also visualized a heat map for correlation matrix of Numerical features where we converted categorical variables into numerical values for the sake of correlation analysis.



xvii. We have visualized the explained variance using a scree plot by comparing the number of principal components and cumulative explained variance.



xviii. We have visualized the relationship between the original features and the principal components.



F. *TESTINGS*

i. HYPOTHESIS TESTING:

We have conducted a two-sample t-test to determine whether there is a significant difference among the purchased products prices and prices of those abandoned in the carts. We performed the test between purchased items and just carted items. After the test we got an output of t-statistic – 11.678 and p-value of 2.195e-30. Here as the p-value is less than the threshold, we rejected the null hypothesis and provide support to the statement that there is a significant difference between the two compared group's prices. The results finally conclude that the customers are not willing to complete the transactions for expensive items. This test helps to provide an in depth and important analysis on the price sensitivity in purchases.

ii. ANNOVA TEST:

We have conducted the one way Annova test to determine if the price differs between the categories. From the annova test the f-statistic is large were p-value is small showcases that the statistical evidence of the average prices are not identical among the categories. We can therefore reject the idea that the means are the same, and conclude there is a significant difference in price between product categories. The analysis clearly shows that the prices of items bought online tend to vary substantially based on the category. This could reflect differences in inherent value, market demand, competitive pricing, or other factors. This task demonstrates how a simple ANOVA test can extract powerful category-level insights from the data. The same technique could be applied to assess differences across brands, time periods, or other dimensions.

TESTS IN SPSS:

The following are outputs of the tests done in SPSS by using the pre processed dataset.

Dataset :

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean	Std. Deviation
product_id	380636	1245	4183866	845503975689	2221292.72	1538726.551
price	380636	.90	64771.06	82656037.46	217.1524	380.45007
Valid N (listwise)	380636					

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
product_id	2221292.72	1538726.551	380636
price	217.1524	380.45007	380636
user_id	1.52E+18	36041004.523	380636

Correlations

		product_id	price	user_id
Pearson Correlation	product_id	1.000	.190	.141
	price	.190	1.000	.081
	user_id	.141	.081	1.000
Sig. (1-tailed)	product_id	.	<.001	<.001
	price	.000	.	.000
	user_id	.000	.000	.
N	product_id	380636	380636	380636
	price	380636	380636	380636
	user_id	380636	380636	380636

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	user_id, price[b]	.	Enter

a. Dependent Variable: product_id

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	Change Statistics: F Change
1	.227[a]	.052	.052	1498401.763	.052	10382.435

Model Summary

Model	Change Statistics: df1	Change Statistics: df2	Change Statistics: Sig. F Change
1	2	380633	<.001

a. Predictors: (Constant), user_id, price

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.662E+16	2	2.331E+16	10382.435	<.001[b]
	Residual	8.546E+17	380633	2.245E+12		
	Total	9.012E+17	380635			

a. Dependent Variable: product_id

b. Predictors: (Constant), user_id, price

Coefficients[a]

Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.
1	(Constant)	-8.167E+15	1.025E+14		-79.683	<.001
	price	725.035	6.405	.179	113.200	<.001
	user_id	.005	.000	.126	79.683	<.001

a. Dependent Variable: product_id

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
product_id * price	0	0.0%	380636	100.0%	380636	100.0%
product_id * user_id	0	0.0%	380636	100.0%	380636	100.0%
category_id * price	0	0.0%	380636	100.0%	380636	100.0%
category_id * user_id	0	0.0%	380636	100.0%	380636	100.0%

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
product_id	2221292.72	1538726.551	380636
price	217.1524	380.45007	380636

Correlations

		product_id	price
Pearson Correlation	product_id	1.000	.190
	price	.190	1.000
Sig. (1-tailed)	product_id	.	<.001
	price	.000	.
N	product_id	380636	380636
	price	380636	380636

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	price[b]	.	Enter

a. Dependent Variable: product_id

b. All requested variables entered.

Model Summary

					Change Statistics	
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change
1	.190[a]	.036	.036	1510845.459	.036	14179.076

Model Summary

	Change Statistics		
Model	df1	df2	Sig. F Change
1	1	380634	<.001

a. Predictors: (Constant), price

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.237E+16	1	3.237E+16	14179.076	<.001[b]
	Residual	8.689E+17	380634	2.283E+12		
	Total	9.012E+17	380635			

a. Dependent Variable: product_id

b. Predictors: (Constant), price

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2054853.105	2819.695		728.750	<.001
	price	766.464	6.437	.190	119.076	<.001

a. Dependent Variable: product_id

Data Visualization Analysis

Descriptive Statistics[a]

	Mean	Std. Deviation	Analysis N
product_id	.	.	0
category_id	.	.	0

a. Only cases for which price = 1 are used in theanalysis phase.

T-Test

Notes

Output Created		18-NOV-2023 21:40:38
Comments		
Input	Data	C:\Users\sriba\Downloads\Hypothesis Testing • ANOVA (Analysis of Variance) • Principal Component Analysis (PCA).sav
	Active Dataset	DataSet2
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	380636
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	Statistics for each analysis are based on the cases with no missing or out-of-range data for any variable in the analysis.
Syntax		T-TEST GROUPS=event_type('cart' 'purchase') /MISSING=ANALYSIS /VARIABLES=price /ES DISPLAY(TRUE) /CRITERIA=CI(.95).
Resources	Processor Time	00:00:00.06
	Elapsed Time	00:00:00.11

Group Statistics

	event_type	N	Mean	Std. Deviation	Std. Error Mean
price	cart	27299	234.1986	198.82267	1.20335
	purchase	17326	217.8122	189.27609	1.43796

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
price	Equal variances assumed	35.297	<.001	8.644	44623
	Equal variances not assumed			8.739	38198.055

Independent Samples Test

		t-test for Equality of Means			
		Significance			
		One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference
price	Equal variances assumed	<.001	<.001	16.38633	1.89576
	Equal variances not assumed	<.001	<.001	16.38633	1.87504

Independent Samples Test

		t-test for Equality of Means	
		95% Confidence Interval of the Difference	
		Lower	Upper
price	Equal variances assumed	12.67061	20.10206
	Equal variances not assumed	12.71120	20.06146

Independent Samples Effect Sizes

				95% Confidence Interval	
		Standardizer[a]	Point Estimate	Lower	Upper
price	Cohen's d	195.17165	.084	.065	.103
	Hedges' correction	195.17493	.084	.065	.103
	Glass's delta	189.27609	.087	.068	.106

a. The denominator used in estimating the effect sizes. Cohen's d uses the pooled standard deviation.

Notes

Output Created		18-NOV-2023 21:48:01
Comments		
Input	Data	C:\Users\sriba\Downloads\Hypothesis Testing • ANOVA (Analysis of Variance) • Principal Component Analysis (PCA).sav
	Active Dataset	DataSet2
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	380636
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics for each analysis are based on cases with no missing data for any variable in the analysis.
Syntax		ONEWAY price BY category_id /ES=OVERALL /STATISTICS DESCRIPTIVES HOMOGENEITY /MISSING ANALYSIS /CRITERIA=CILEVEL (0.95) /POSTHOC=TUKEY ALPHA(0.05).
Resources	Processor Time	00:00:00.27
	Elapsed Time	00:00:00.62

Warnings

Post hoc tests are not performed for price because there are more than 50 groups.

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
price	Based on Mean	148.695	156	380477	<.001
	Based on Median	108.909	156	380477	<.001
	Based on Median and with adjusted df	108.909	156	25032.886	<.001
	Based on trimmed mean	119.329	156	380477	<.001

ANOVA

price

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7583234915.1	158	47995157.691	384.356	<.001
Within Groups	47510733614	380477	124871.500		
Total	55093968529	380635			

ANOVA Effect Sizes[a]

		Point Estimate	95% Confidence Interval Lower	Upper
price	Eta-squared	.138	.135	.139
	Epsilon-squared	.137	.135	.139
	Omega-squared Fixed-effect	.137	.135	.139
	Omega-squared Random-effect	.001	.001	.001

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.544
Bartlett's Test of Sphericity	Approx. Chi-Square	22508.105
	df	6
	Sig.	<.001

Communalities

	Initial	Extraction
price	1.000	.436
category_id	1.000	.999
product_id	1.000	.533
user_id	1.000	.310

Extraction Method: Principal Component Analysis.

Factor Analysis

Notes

Output Created		18-NOV-2023 22:01:29
Comments		
Input	Data	C:\Users\sriba\Downloads\Hypothesis Testing • ANOVA (Analysis of Variance) • Principal Component Analysis (PCA).sav
	Active Dataset	DataSet2
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	380636
Missing Value Handling	Definition of Missing	MISSING=EXCLUDE: User-defined missing values are treated as missing.
	Cases Used	LISTWISE: Statistics are based on cases with no missing values for any variable used.

Component Matrix[a]

	Component 1	2
price	.659	-.037
category_id	.012	.999
product_id	.730	.002
user_id	.556	.018

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix[a]

	Component 1	2
price	.660	-.031
category_id	.003	.999
product_id	.730	.009
user_id	.556	.024

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Component Transformation Matrix

Component	1	2
1	1.000	.009
2	-.009	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Principal Component in SPSS :

Factor Analysis :

[DataSet1]

Communalities

	Initial	Extraction
product_id	1.000	.589
price	1.000	.589

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.177	58.867	58.867	1.177	58.867	58.867
2	.823	41.133	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component 1
product_id	.767
price	.767

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

Rotated Component Matrix[a]

a. Only one component was extracted. The solution cannot be rotated.

Resampling Data :

Dataset-1

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
product_id	841002	102	4183880	1906570.01	1458547.560
price	841002	.22	64771.06	146.3449	300.34209
Valid N (listwise)	841002				

Bootstrap

Bootstrap Specifications

Sampling Method	Simple
Number of Samples	1000
Confidence Interval Level	95.0%
Confidence Interval Type	Percentile

V. PRELIMINARY RESULTS

The analysis of the customer dataset revealed several key insights:

i. The data contained over 380k records of user events like product views, cart adds and purchases. This provided a rich dataset to work with.

ii. The computers category had the most unique products while components led for purchases, showing the most popular segments.

iii. Visualizations highlighted computers as the top category by share of purchases. Components ranked highest by sub-category purchase count.

iv. Video cards turned out to be the most frequently purchased product among users.

v. When looking at events, product views dominated, with cart adds and purchases significantly lower. This pointed to drop-offs inthe purchase funnel.

vi. The average price was higher for carted items versus purchased ones, indicating potential price sensitivity.

vii. An area chart of top purchased products illustrated changing purchase frequencies over time.

viii. A key metric - the cart abandonment rate - stood at 37%, presenting a major gap.

ix. Daily and seasonal trends were uncovered by converting the data to a time series format.

x. Conversion rates were calculated from product views to purchases to identify popular and engaging items.

xi. A correlation matrix revealed associations between variables like price and likelihood of purchase.

xii. The customer journey from initial view to final purchase highlighted poor conversion once items were carted.

In summary, the analysis provided a data-backed understanding of customer behavior - their preferences, paths and pain points. Key problem areas like high abandonment rates presented opportunities for optimization. The results established a baseline to guide business decisions to better serve and convert customers.

EVALUATION METRICS:

In our project we have done some evaluation metric tests which includes RMSE which tells us how much our model's price predictions differ from real prices. In the linear regression model, the RSME is 285.76 and we have also done the Random Forest Model in which it is lower with 129.61 which shows that the Random Forest model was better in getting closer predictions of the prices.

We have also Logistic regression model to see if we can predict if customer would buy anything or not. we found out the model’s accuracy and f1 score i.e 52% & 0.52 respectively. We got these results after training the bought and left over sets.

At the end, we used PCA which helps use to minimalize the huge dataset in a simpler way which in turn makes our models work easily. All the above steps helped us to understand our models in a better way which also clearly show how good our approach in our project is.

VI. ANALYSIS

To Perform EDA, we have imported required libraries needed for data exploration and visualization, uses Pandas to load the data set events.csv, and displays the few rows of data. Further analysis and research should understand the characteristics of the data set, such as data types, statistical data, and different distributions. In the next step we extracted main-category, sub-category, and item name from the 'category code' column using string splitting, creating three new columns from it.

Secondly, to know key insights such as central tendencies, Numerical data, categorical data we have done Summary statistics. From this we have identified the unique values in the main category and analyzed the average purchase from each category and visualized them by using the pie plot. Furthermore, from the subcategory we have represented the unique purchase count and visualized the Top 10 purchased items.

Next , The main agenda of the of project is to analyze the cart abandon rate so we have counted the event types which are view,cart,purchase were the activities are number of views is 393471,Items added to cart are 23824 and the number of items purchased are 17072.By using the plotting techniques we have done data distribution over price, top categories, top brands and listed top main categories, Top Subcategories, Top Item types.

To know the daily event count we have converted event time to date time by time series plot and analysis the number of purchases by date in summary we have calculated the Cart Abandonment Rate where number of cart events and purchase events is 0.28. From this analysis the weekly trend of the events.

Here in the next step to know the number of users visiting the page per day we have used bar plot to represent it and extracted features which view, items adding to the cart and number of items purchased in an hour and a day in a week. Finally Analyzed the average price in the cart 112.28 and the average price purchased is 99.52 were top categories in the cart are electronics which are telephones and second is computers but the top categories purchases are electronics 2108 in next place stationery.cartrige 1612. And analysis the user behavior pattern by using highest abandonment rate and listed the group by category and calculate abandonment rate.

From this we came to know that by analyzing the price factor the abandonment rate Low price is 0.211595, Medium price is 0.217132, High price 0.320156 and Very High price 0.378667.

To continue this based of certain brands we calculated the abandonment rate and analyzed based on user session length which event type session length (-0.001, 55.0) abandonment rate is 0.569356, 2. (-0.001, 55.0) where ab rate is 0.569356, (55.0, 910.0) ab rate is 0.316684 and 4. (910.0, 86393.0) ab rate id 0.158337. visualized it with the heat map, Convert categorical variables to numeric for correlation analysis. Summarized the customer journey view_to_cart 0.000044 cart_to_purchase 0.014327. and then calculated the view to purchase conversion.

By using product id calculated the Max purchases, max carts product, max views by product and identified the max cart by user.

In summary, we have tested by using multiple testing methods were first one is hypothesis testing were the probability value 0.03 show that there is significance difference in price and other models One-way Annova, logistic regression failed to reject the null hypothesis.

To the extension of above I have performed machine learning models to evaluate the accuracy of the cart purchase rate and user activity initially I have used logistic regression based on the price and event type were it give the 96% accuracy it shows that 4% of users were purchasing products which are added to cart.

To know the users interest in purchasing product were depends on price so I evaluated the average price of product were 135 is the lower and 138 is the upper interval by using confidence interval and conversion rate which is 0.0417 is the lower and 0.0426 is is upper and bootstrapping also gives the same mean price of the product .

The regression model support vector machine is one of the model to extract the categorical variables to find the user activities were it gives the 89.6% accuracy show that the users are not purchasing products and the activity is 'VIEW'.

The next model is Random Forest is used to give better accuracy on large dataset but the model give 86.4% accuracy which is almost equal to the svm and end of the output states same as compared to svm were users activity is viewing the products.

The Third model is decision tree which is used to split the data and focus on individual points and it is used to evaluate the user purchasing the products were the accuracy of the model is 89.3% on viewing .The model also states that user are not purchasing the products and their activity is viewing the items .

The final model is neural networks which is lstm which is long short term memory is one of the efficient model suitable for numerical data and categorical data and focusing on dependencies . we analyzed the cart abandoned on 10 epochs and the accuracy is 95.8% users are not purchasing the products.

The cross validation is used to find the key point depending on the price were we used linear regression and random forest performed k-fold cross validation RMSE of the linear regression is 285.76 and random forest is 129.61 were random forest model indicates better predictive performance and accuracy compared to Linear Regression.

Resampling is also one method used to find the cart abandonment accuracy of the users where the model used is logistic regression which gives 52% were the model struggled to perform well because the data set which it generated is balanced.

Principal component analysis is used on numerical variables to extract the main features of the given data set we computed covariance matrix, eigen values and eigen vectors .were the two features identifies 100% variance and user id feature significantly influences the pc1.

Finally, by performing various techniques, classification method and regression methods we analyzed that 95% of user are not purchasing the products and the most the activity of them is viewing the products and identified that the dependent reason is price which is affecting more. The

regression model LSTM gives the best accuracy, which is 95%. And performed hypothesis testing ,t-test and null hypothesis testing also states to reject the null hypothesis .

In conclusion, Customers show distinct preferences for specific product categories and brands. The prominence of categories of products Computer which Asus's brand are highly significantly more attractive to customers. So, we visualized the top products and brands having high purchase rates. Price sensitivity is one of the major impact effects on sales where highly priced items have less purchase rate. The next is Cart abandonment rate 31% is substantial loss of potential sales based on the factor price, quality, and popularity of the product. And the weekday sales peeks on Monday, Friday, Saturday. For the user buying pattern many views compared to purchases highlighted the customer journey from interest to decision-making to purchasing the product. And event type analysis is the final part of this where it is divided into 3 parts adding to cart, Purchasing, Viewing based on the regression models we performed it can show the behavior of user who is purchasing the product.

VII. PROJECT MANAGEMENT

A. *WORK COMPLETED*

i. *DESCRIPTION:*

Ecommerce customer data is loaded from CSV to pandas Data Frame. Basic data frame information such as rows, columns, and data types are printed. - Calculate aggregate statistics such as average, median, min, and max for statistical columns such as price and product id. - Costs are analyzed in categorical columns such as category and brand to understand cardinality and distribution. - Create groups and aggregations on columns such as category and event type to analyze distributions and find top categories, common events, and so on. - Provides visualizations such as histograms, bar charts, and pie charts to visualize categories such as prices, categories, and more. Purchase surveys - For the purposes of the analysis, only 'purchase' events have been selected for consideration. - Using aggregates such as unique and count to understand the top product categories, products and brands. - Examine the acquisition by date to see trends over time. - Analysis of the conversion funnel from product management to cart addition to purchase. - Metrics such as cart abandonment rate and purchase count per hour are calculated. Statistical analysis - Hypothesis testing is carried out using t-test and ANOVA to compare the prices of items in different categories and items purchased and items abandoned. - Performing correlation analysis to determine the relationship between statistical variables. - Group analytics track user journeys from product management to purchase to cart addition. - Time series analysis looks for trends and timing of events in time. - Confidence intervals are calculated for metrics such as average purchase price. - Bootstrapping is used to estimate confidence intervals for average purchase prices.

As for the final submission we have also done some other testing's and techniques to ensure our model is performing well and also make our conclusions easier. We have calculated the confidence Interval, Boot strapping, SVM, Random forest, Decision tree, LSTM, Cross Validation, Linear regression, Resampling PCA and Explained variance and also in the spss part we have further done the PCA and Resampling. At the end we have done the analysis part very detailed and clear manner and we have also dawn the conclusions of our project.

ii. *RESPONSIBILITY*

NAME OF THE TASK	**CONTRIBUTED BY**
DATA PRE-PROCCESSING	SAI KUMAR REDDY
RESEARCHING ON TOPIC	SRIBALA
EDA	SAI KRISHNA
DESIGNING FEATURES	SRIBALA
DATA TRANSFORMATION	SAI KRISHNA
FEATURE EXTRACTION	SAI KRISHNA
DATA VISUALIZATION	SAI KRISHNA
SPSS PART -1 & its Analysis	JYOTHIKA
SPSS PART -2 & its Analysis	SRIBALA
INITIAL TESTING	SAI KUMAR REDDY
CODE ANALYSIS	SAI KUMAR REDDY
REPORT INTRODUCTION	JYOTHIKA
RELATEDWORK INREPORT	JYOTHIKA
IMPLEMENTATION STATUS	SRIBALA
WORK PENDING REPORT	JYOTHIKA
REFERENCES IN REPORT	SAI KUMAR REDDY
CONFIDENCE INTERVAL	SAI KUMAR REDDY
BOOTSTRAPPING	SAI KUMAR REDDY
SVM, RANDOM FOREST(1)	SAI KUMAR REDDY
DECISION TREE, LSTM	SRIBALA
CROSS VALIDATION	SRIBALA
LINEAR REGRESSION, RANDOM FOREST(2)	SRIBALA
RESAMPLING,PCA, EXPALINED VARIANCE,	JYOTHIKA
PCA AND RESAMPLING IN SPSS	JYOTHIKA
FINAL ANALYSIS	SRIBALA,SAI KUMAR
UPDATING FINAL DOC FOR SUBMISSION	SAI KRISHNA

iii. *CONTRIBUTIONS*

MEMBER NAME	**CONTRIBUTION %**
SAI KRISHNA PEDAPUDI	28 %
SAI KUMAR REDDY	28 %
SRIBALA PUTCHA	24%
JYOTHIKA PAGILLA	20%

GITHUB LINK:

We have uploaded the code and project documentation in GitHub, which can be accessed from the link given below.

https://github.com/saikumar2903/E.A-Project/tree/main

VIII. REFERENCES

[1] M. Veeragandham, N. Patnaik, R. Tiruvaipati, and M. Guruprasad, “Consumer Buying Behaviour towards E-Commerce during COVID-19”, IJRESM, vol. 3, no. 9, pp. 78–82, Sep. 2020.

[2] Ifeoma Adaji and Julita Vassileva. 2017. A Gamified System for Influencing Healthy E-commerce Shopping Habits. In Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17). Association for Computing Machinery, New York, NY, USA, 398–401. https://doi.org/10.1145/3099023.3099110

[3] Tian, Y., Ye, Z., Yan, Y. et al. A practical model to predict the repeat purchasing pattern of consumers in the C2C e-commerce. Electron Commer Res 15, 571–583 (2015). https://doi.org/10.1007/s10660-015-9201-8

[4] Dr.Anukrati Sharma “A Study on E-commerce and Online Shopping: Issues and Influences.” International Journal of Computer Engineering and Technology. Volume 4, Issue 1, January- February 2013, https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_4_ISSUE_1/IJCET_04_01_035.pdf