eCommerce behaviour data from multi-category store

Lekha Ajit Kumar, Sushmitha Dandu, Dauren Omarov, Navyasree Sriramoju
Department of Information Systems, California State University Los Angeles
Los Angeles, USA
lajitku@calstatela.edu, sdandu3@calstatela.edu, domarov@calstatela.edu, nsriram@calstatela.edu

Abstract This paper explains about customer behavior patterns for an online eCommerce muti-category store. We can interpret data by seeing the customer buying and viewing behavior for a particular brand and category. The paper explains the method and process used for data manipulation and further analysis. The major goal of the project is to give a clear flow of handling big data files and data cleaning processes using Hadoop and Beeline. The data interpretation and analysis of this data was done using Excel, Tableau and Power BI. Charts and visuals like as bar diagrams, timeline and charts on customer behavior patterns for an online eCommerce muti-category store.

1. Introduction

This dataset contains customer behavior data for October and November 2019. The size of the dataset is 5GB. There are 10 columns in this set, which are, as follows: time, type, product id, category id, category code, brand, price, user id, and user session. In this data set our goal is to analyze customer buying behavior. We have Data from 2 different months, hence we also compare the buying and viewing patterns of customers. We have chosen this dataset because it will help companies understand requirements of the users and help maximize their sales and production of demanding goods.

2. Related Work

There are plenty of studies regarding e-commerce shops behavior. One study is based on the data set from Kaggle.com [**?**]. In this research, the authors study the behavior of an online e-commerce store and its habits. In their methodology, the authors removed the irrelevant features of the study. The authors used bar charts and pie charts. For example, the authors disregarded null values, and used only Event type, product_id, category_id, brand, price columns from the data set. Also, in the study duplicate entries are ignored. Moreover, one column 'purchased' was added to the dataset to indicate if there was a purchase action. In the study results, more people are viewing the merchandise than buying it. Hence, most customers simply looked at the goods and left it there. The authors concluded that so few people purchased the item because users are knowledgeable enough to recognize its significance in their lives, and as a result, they must have looked it up elsewhere before adding it to their basket. In our study we are using dataset from the same source but implementing different approach as we are using Big Data analysis using Hive. Another study made in 2020 focuses on eCommerce behavior data from multi category store [**?**]. The dataset was derived from [4][5][6]. Obviously, traffic department of Louisville currently have a sustained flow of data and use it on a daily basis, whereas in our work, even if the dataset has the same structure, we were limited to a small portion of data. However our work, apart from analysis of traffics, also Kaggle.com. The author of the study established several primary goals. One of them is to calculate the daily traffic in November. The second is to find out which product categories and brands are most popular. The third is to find out whether the buyer will buy goods by adding them to the cart. In the research the author is using Python 3 with squarify and xgboost libraries, and basic prediction model through XGBoost. The author calculated that in November there were 3,696,117 website visitors. Electronics.smartphone, Electronics.video.tv, Computers.Notebook, Electronics.Clocks, and Apparel.Shoes are the leading 5 categories that people view, as the author says. The most popular brands include Samsung, Apple, Xiaomi, and other brands. The author comes to the conclusion that only a tiny percentage of users actually add the item to the basket, which results in a purchase conversion of about 30% (1.36% divided by 4.49%) once the item was added to the cart by buyers. In contrast to our study, which uses a Big Data model using Hive, the author of this study used Python 3 libraries and a simple prediction model through XGBoost.

3. Specifications

The dataset consists of behavior data for 2 months (from October and November 2019) from a large multi-category online store. This dataset is publicly open and stored in a CSV file format. The dataset is of the size 14.68 GB and covers several days in October and November, the file consists of irregularities majority in the month of November. Although the data is of a big size, we assume that the same data processes can be applied to a similar dataset (as large as 100 GB+ or as small as 25 MB). Table 1 shows the files and size of the files from the dataset.

4. Implementation Flowchart

Initially, the first version of the dataset known as raw dataset, which comprises all detail customer behavior patterns for an online eCommerce muti-category store, was downloaded from Kaggle. The whole process of data manipulation

and analysis is shown in the below flowchart (Figure 1).There are two data logs in csv format were uploaded to the Hadoop File System. After that, we created the tables' schema, cleaned the data, create a cleaned data table and summary table, then exported the results. Once the output file has been downloaded we opened in Excel format, we use Excel's in built charts, Power BI and tableau to obtain the visualizations.

Data Set	Size (Total 14.68BG)
2019-Nov	9.01 GB
2019-Oct	5.67 GB

TABLE I
DATA SPECIFICATION

5. DATA CLEANING

Firstly the uncleaned or raw files were uploaded and stored in HDFS in the tmp directory. Then we shifted it to our directores ecommerce1 and ecommerce2 which we created in our login. Then created tables and loaded the raw data into tables using Beeline. Since we had two files, each was separately cleaned and then exported for further analysis. Data cleaning was conducted using different techniques such as regular expressions, select statements, like "NULL" and exporting tables. Below is one peace of the code used for cleaning data. SELECT * from octuncleaned where category_code not like "NULL" AND brand not like "NULL" AND user_session not like "NULL";

In addition to that we created a cleaned table to portray all cleaned data in a seprate table so we can permorm qeuire on it freely without always not removing null values.

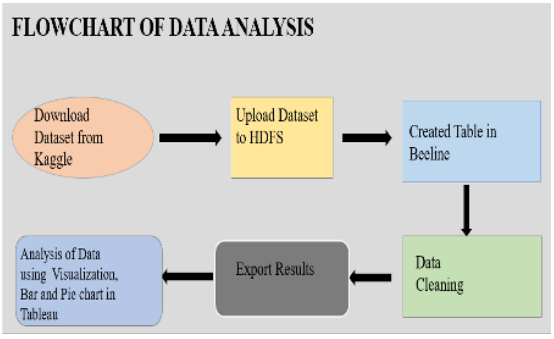

Fig. 1. Figure 1 - Implementation Flowchart

6. ANALYSIS AND VISUALIZATION

We extracted the files into excel, Power BI and tableau for further analysis. We used Pie charts, Bar graphs, line graphs and dotted graphs to represent the sales, brands, count, categories, user details and purchases. This helps to interpret the customer behavior. Here are few visualizations from our analysis:

6.1 Graph in Excel

From the line graphs shown in Figure 2 and Figure 3 we can see the customer's engagement on different brands in

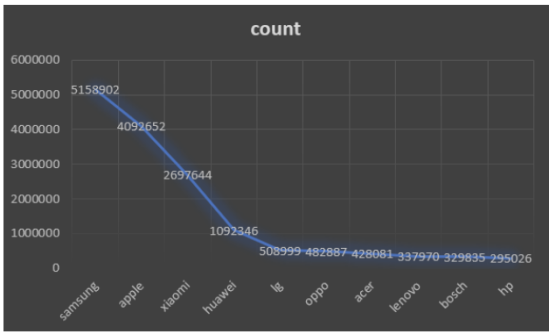

Fig. 2. Most popular brands in October

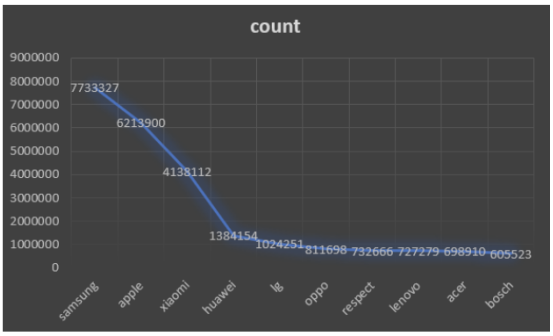

Fig. 3. Most popular brands in November

the month of October and November [**?**]. The data includes customer's who viewed, added the items to the cart and purchased products. Customer's pattern of interest is almost the same for the two months. Samsung being in the top followed by apple, Xiaomi, Huawei, lg and oppo. Companies which are in the bottom can promote and offer deals in order to attract customers to increase their customer engagement with the brand.

6.2 Bar Graph in Tableau

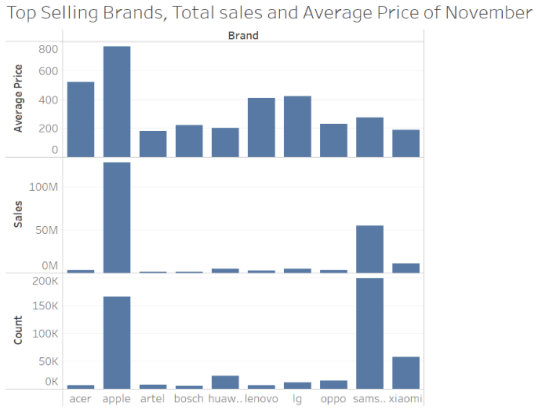

Fig. 4. Most purchased brand and average price of November

This visualization is done in tableau to understand the customer behaviour precisely. From the bar graphs Figure

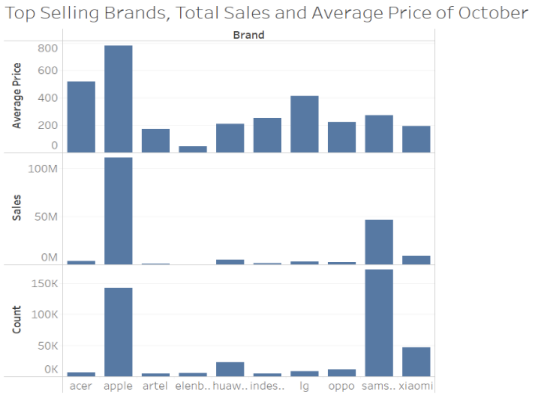

Fig. 5. Most purchased brand and average price of October

4 and Figure 5 we can conclude that though Samsung is leading on number of sales for the months of October and November the average selling price of apple is comparatively more resulting in more revenue for apple. The data indicates reputation of both companies and gives a glimpse of choice for customers based on their budget.

6.3 Pie Chart in Excel

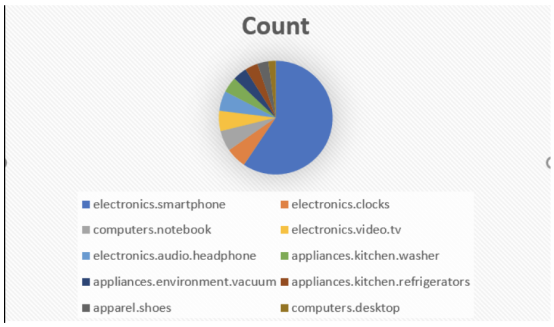

Fig. 6. Most popular categories in October

Through this visualization from Figure 5 we can conclude that electronics contributed to the most popular category than the appliances. We can see that the most popular categories are electronics. Businesses can use this data in order to improve their performance in these categories. They also can analyse this information to calculate revenue, EBITDA, etc. Businesses can also use this data to generate forecasts for future months. It also shows that electronics and appliances categories are the driving force for the e-commerce stores. We also made Pie charts visualization for better interpretation.

6.4 Bar Graph in Excel

Here are top 10 categories which have been least popular in November from Figure 6. We can see that most unpopular categories are from apparel. Businesses can use this data in order to improve their performance in these categories. It also shows that people prefer to use in-person stores for apparel rather than online e-commerce stores.

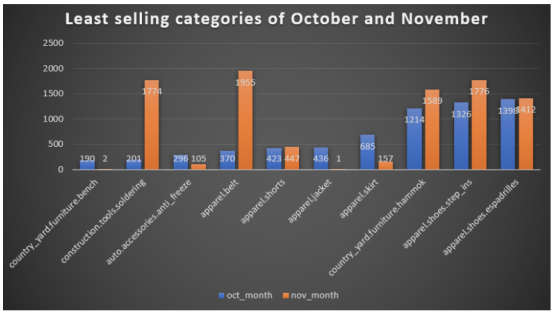

Fig. 7. Least selling categories of October and November

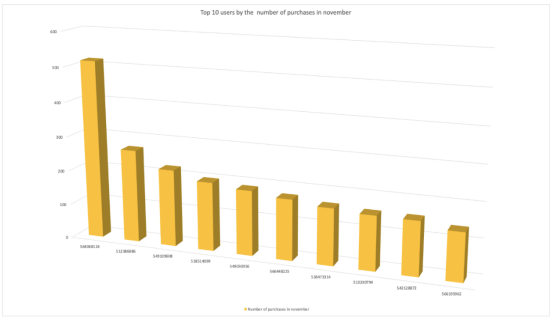

Fig. 8. Users who made the greatest number of Purchases

From Figure 6 we get the user data who did the greatest number of purchases, using this information these users should be given special discounts and some benefit to gain their trust and maintain customer relationship as they would recommend few others. This helps the business to gain customers without promotion.

7. Conclusion

a) As per the above analysis we can conclude the following: b) Smartphones are the most popular & purchased category in both the Months. c) Furniture Bench and Jackets are the least purchased categories in October and November, respectively. d) Samsung is the most popular & purchased brand of October and November. e) Besafe and Ava are the least purchased brands in October and November respectively. f) Users viewed, added the Products in Cart and Purchased mostly in November than in October. g) Sales in November are more than Sales in October. h) Most Viewed but not purchased brand is Casio. i) Most of the purchases happened around 9'o clock. j) Most of the purchases happened around 16th of October. k) User id – 564068124 has made most of the Purchases.

This report contains the customer behaviour pattern for two months, we were able to provide few insights which can be used by the business owners for the improving their revenue and introduce new strategies for higher profits.

References

[1] https://stackoverflow.com/questions/51097895/hive-sql-find-most-popular-value-across-multiple-columns

[2] https://powersync.biz/blog/types-of-data-analysis-used-in-ecommerce-to-get-key-business-insights

[3] https://sanyasachdeva1.github.io/Portfolio/files/Analysis%20of%20e-commerce%20behavior%20in%20Multi-Category%20Store.pdf