Zewail City of Science and Technology

CIE 427

Big Data Analytics

Final Project - Analyzing eCommerce Data

Fatma Moanes NourEl-Din	201700346
Mohamed Elshaffei	201700030
Mahmoud Ashraf	201700989

Special Thanks to Eng. Ahmed Sameh/Elsayed and Eng. Muhammad AlAref

Contents

1 Introduction

This report analyzes and describe different patterns in eCommerce data. The provided analytics provide data-driven insights into how shoppers interact with the site. These insights take the guesswork and subjectivity out of website optimization, and uncover opportunities for improvement and innovation.

2 Data

2.1 Background

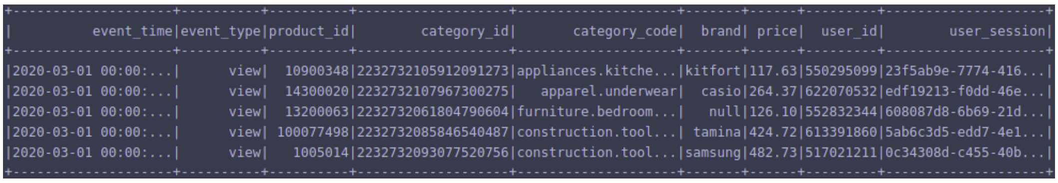
The data used in this report belongs to a large multi-category online store. The data contains behaviour for seven months from October 2019 to April 2020. The dataset was collected by REES46 Marketing Platform and can be downloaded from Kaggle.

2.2 Data Description

The data is split into seven different files, where each file contains customer behaviour for one month. The first line of each file includes the column headers. The dataset has the following columns:

1. event_time: the time when the event happened at (in UTC).
2. event_type: can take only one of four different values: "view", "cart", "remove_from_cart", and "purchase".
3. product_id: ID of the product involved in this event.
4. category_id: ID of the category of the product.
5. category_code: product's category taxonomy, ex: "appliances.kitchen.oven".
6. brand: brand name of the product.
7. price: float price of the product.
8. user_session: temporary user's session ID. Same for each user's session. It changes every time users return to the online store after a long pause.

Figure 1 shows a snapshot of the dataset.



```
+--------------------+----------+----------+-------------------+--------------------+-------+------+---------+--------------------+
|          event_time|event_type|product_id|        category_id|       category_code|  brand| price|  user_id|        user_session|
+--------------------+----------+----------+-------------------+--------------------+-------+------+---------+--------------------+
|2020-03-01 00:00:...|      view|  10900348|2232732105912091273|appliances.kitche...|kitfort|117.63|550295099|23f5ab9e-7774-416...|
|2020-03-01 00:00:...|      view|  14300020|2232732107967300275|   apparel.underwear|  casio|264.37|622070532|edf19213-f0dd-46e...|
|2020-03-01 00:00:...|      view|  13200063|2232732061804790604|furniture.bedroom...|   null|126.10|552832344|608087d8-6b69-21d...|
|2020-03-01 00:00:...|      view| 100077498|2232732085846540487|construction.tool...| tamina|424.72|613391860|5ab6c3d5-edd7-4e1...|
|2020-03-01 00:00:...|      view|   1005014|2232732093077520756|construction.tool...|samsung|482.73|517021211|0c34308d-c455-40b...|
+--------------------+----------+----------+-------------------+--------------------+-------+------+---------+--------------------+
```

Figure 1: Snapshot of the dataset

The total number of events for all months is 412 million events. Figure 2 shows the distribution of the dataset based on the count of each event.

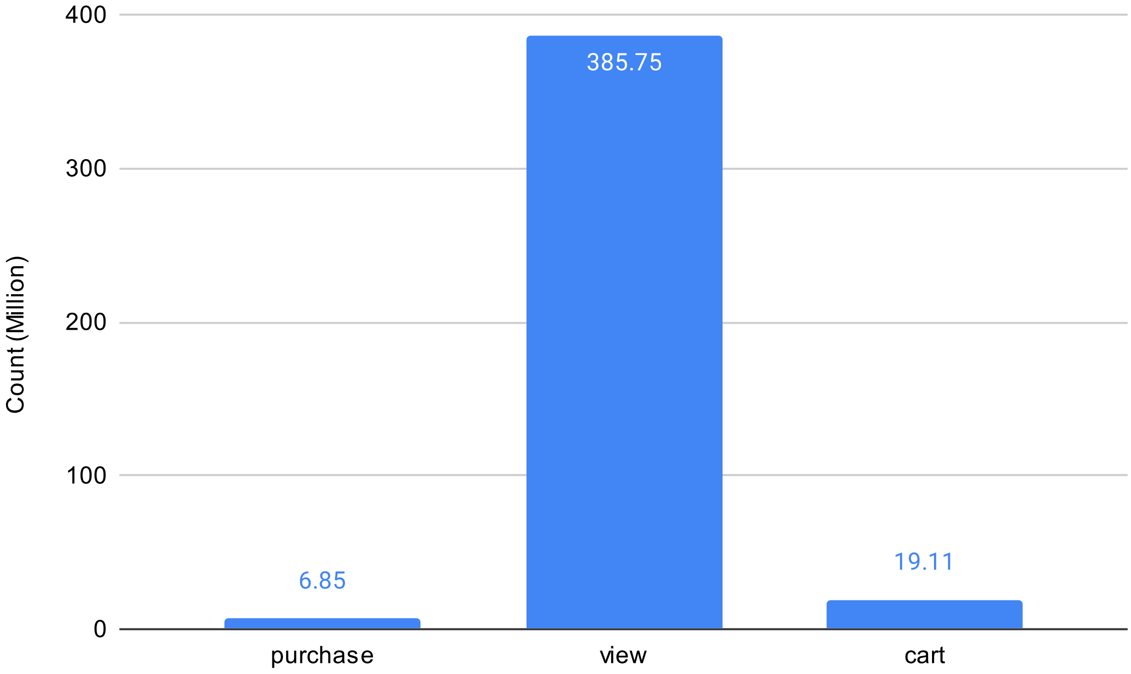


Figure 2: Event type vs its count

As expected, the above graph shows that the view product event dominates the dataset. The chart also shows that the event “remove_from_cart” is missing from the dataset, although the data source stated it should be present.

2.3 Data Problems

The dataset exploration revealed several problems. Firstly, some rows have null ”category_code” and/or brand. Furthermore, the ”category_code” is invalid for a

large number of rows. For example, while exploring, the team discovered that "construction.tools.light" mainly contains brands like Samsung, Apple, Xiaomi, Huawei, Oppo, and Meizu. Such data does not make sense because brands like Apple, Oppo, and Meizu do not manufacture these products.
Further investigation revealed that this category is the best-selling category across the whole dataset, which confirms that the data is wrong because Apple sales of screws and drivers simply can not be more than those of iPhones. The team replaced all occurrences of the category "construction.tools.light" with the smartphones category to solve this problem. This solution is justified because 99% of the brands in this category are smartphones brands.

Another potential issue is that no purchases events are present on the 2nd of January while customers made tens of thousands of purchases on the 1st and 3rd of the same month, which again does not make sense. Similarly, on the 20th and 21st of April, only 22 and 29 purchases were made, while on the 19th and 22nd of the same month, consumers made tens of thousands of purchases. The team believes that those are not standard consumption patterns but instead a data problem. Therefore, those values were replaced by the average number of purchases in their corresponding months to ensure that any related analysis is not flawed.

3 Customer Behavior Analysis

3.1 Best-Selling

The analysis begins by first understanding the main patterns in the data. Figure 3 shows the distribution of the best-selling categories. The graph shows that the smartphones category dominates the sales by a far margin.

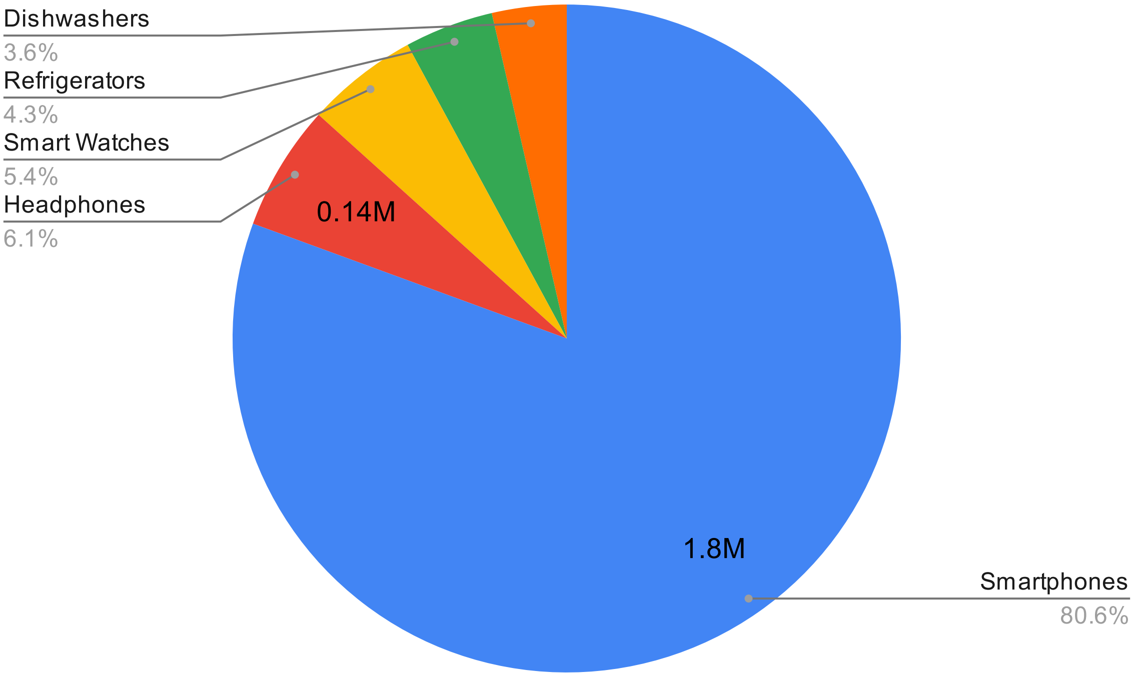


Figure 3: Count of purchases per category

Now, let's explore the best-selling brands in the smartphones category. Figure 4 shows that consumers seem to be willing to buy Samsung and Apple smartphones more than others.

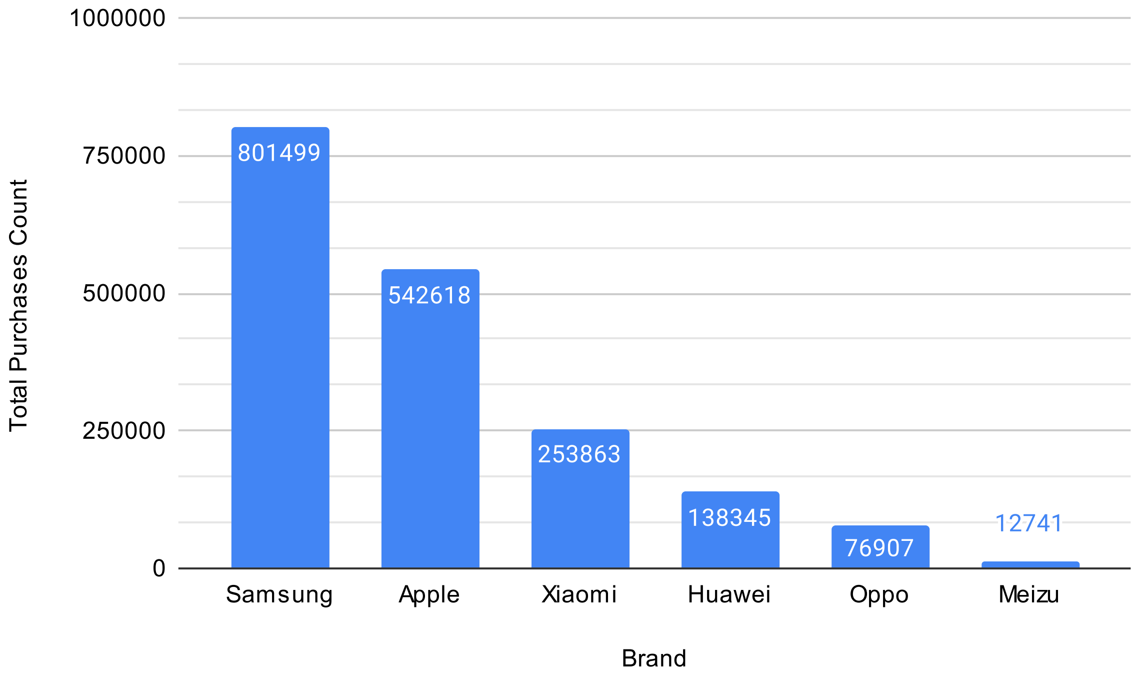


Figure 4: Count of purchases per smartphone brand

3.2 Customer Conversion Analysis

Customer conversion rate is the percentage of potential customers who take a specific desired action. Customer Conversion Rate is one of the key metrics for any business, because optimizing it the business is able to lower its customer acquisition costs by getting more value from the visitors and users it already have. After reviewing different online sources such as databox[1], salecycle[2], and statista[3] it appears that on average customer conversion rate for electronic devices is anywhere from 3% to 5% while add to cart conversion rate is on average from 4% to 8%.

Figure 5 shows the funneling ratios of the smartphones category. The conversion rate and the add to cart rate of smartphones both fall within the upper ranges of the standard conversion and add to cart rates.

[1] https://databox.com/improve-your-funnel-conversion-rate

[2] salecycle.com/blog/strategies/mobile-conversion-rates-lower-desktop/

[3] statista.com/statistics/439558/us-online-shopper-conversion-rate/

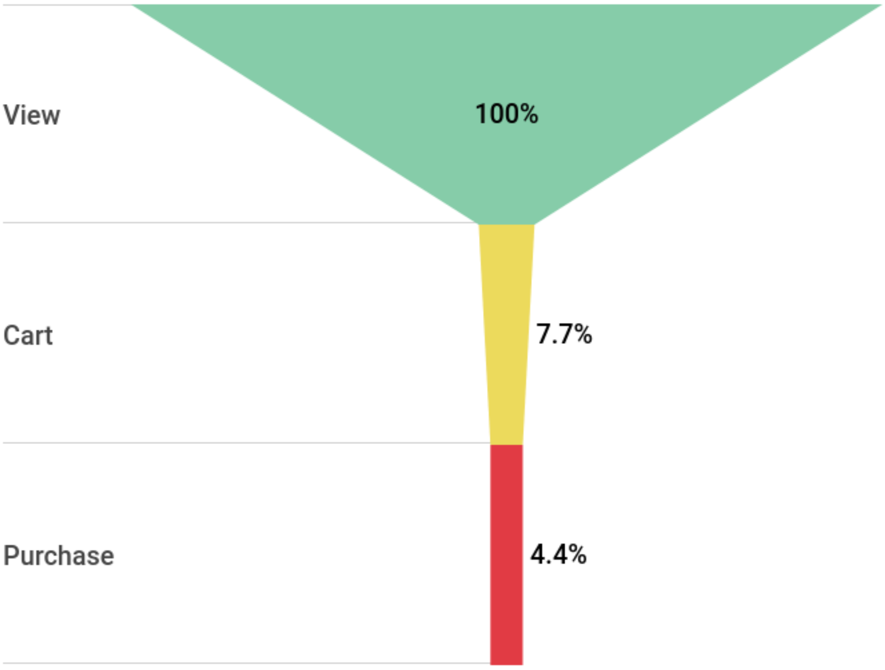


Figure 5: Funneling pipeline for smartphones

Figure 6 shows the funneling ratios of the printers category. The conversion rate and the add to cart rate of smartphones both are lower then the standards.

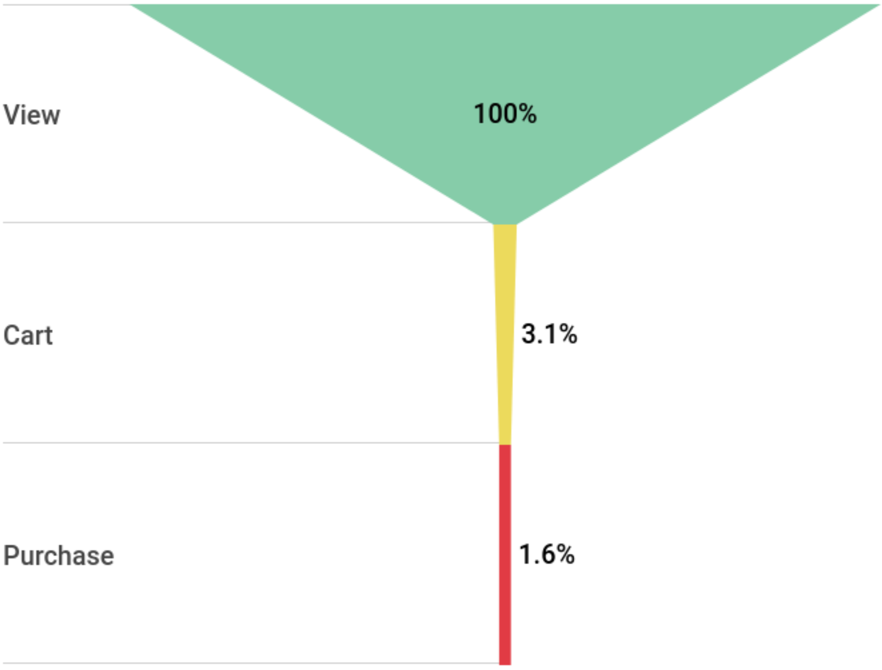


Figure 6: Funneling pipeline for printers

Figure 7 shows the funneling ratios of the refrigerators category. The standard conversion rate of home appliances as per irpcommerce[4] is around 2.38%. Therefore, the conversion rate shown in figure 7 is lower than it should be.

[4] https://www.irpcommerce.com/en/gb/ecommercemarketdata.aspx?Market=15

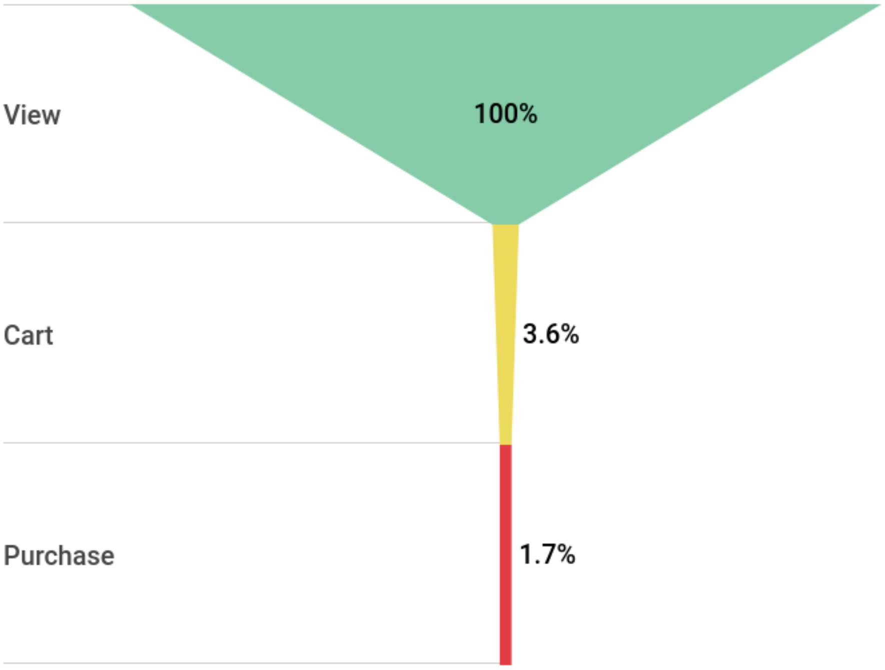


Figure 7: Funneling pipeline for refrigerators

The above figures show that smartphones have the best conversion rates among other top categories. Figure 8 investigate the conversion rates of different smartphones brands to get a closer look. The results are consistent with the results shown in figure 4 with Samsung and Apple once again coming on top achieving above average conversion and add to cart rates.

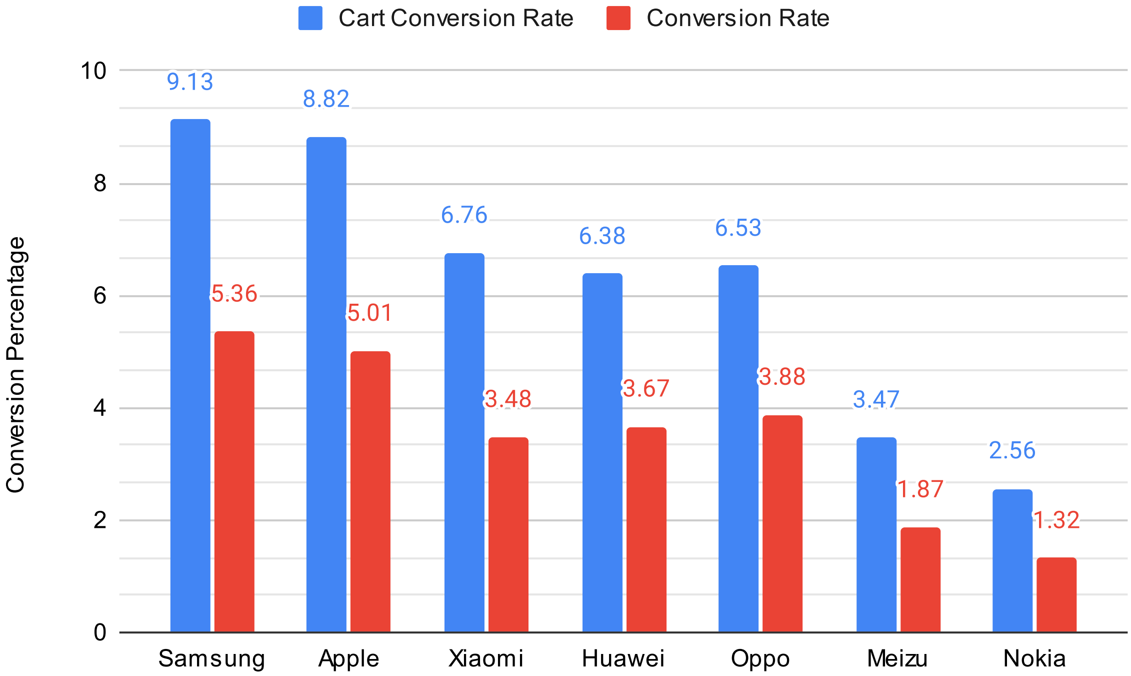


Figure 8: Conversion rates for popular smartphones brands

3.3 Time Analysis of Customer Behavior

In this section we investigate the customer behavior patterns and how they vary with time. Firstly, we investigate how the conversion rate vary with different months. Figure 9 shows the results. The results show that the worst conversion rate occurs in the month of November. This figure may be due to the fact that during the Black Friday season in November many people may be curious to just skim the prices of different products looking for sweet deals and promotions. Therefore, more people than usual are viewing different products looking for promotions without the strong intention of buying if no suitable promotion is found.

The peak during December can also be explained by the fact that many customers are buying Christmas gifts, so customers are more oriented and not just wandering looking for promotions like in November. The peak in February can be explained by referring to b2bmarketing[5] which states that the conversion rate for February is usually 10% above average due to companies and shops beginning the new year with new marketing campaigns that attract more customers.

[5] b2bmarketing.net/en-gb/resources/blog/3-reasons-why-months-matter-sales

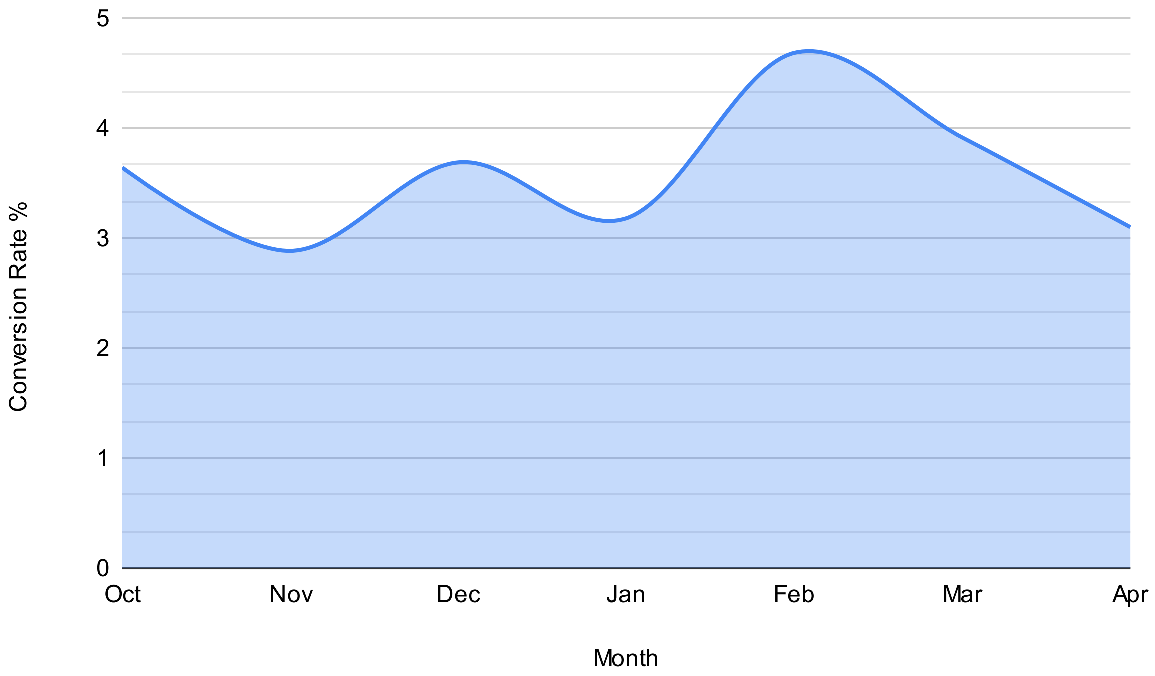


Figure 9: Average Conversion rate for all categories per month

To confirm this hypothesis we can inspect figure 10 which investigates the total number of views for each month. We can clearly see the large peak of views during the month of November which supports our hypothesis.

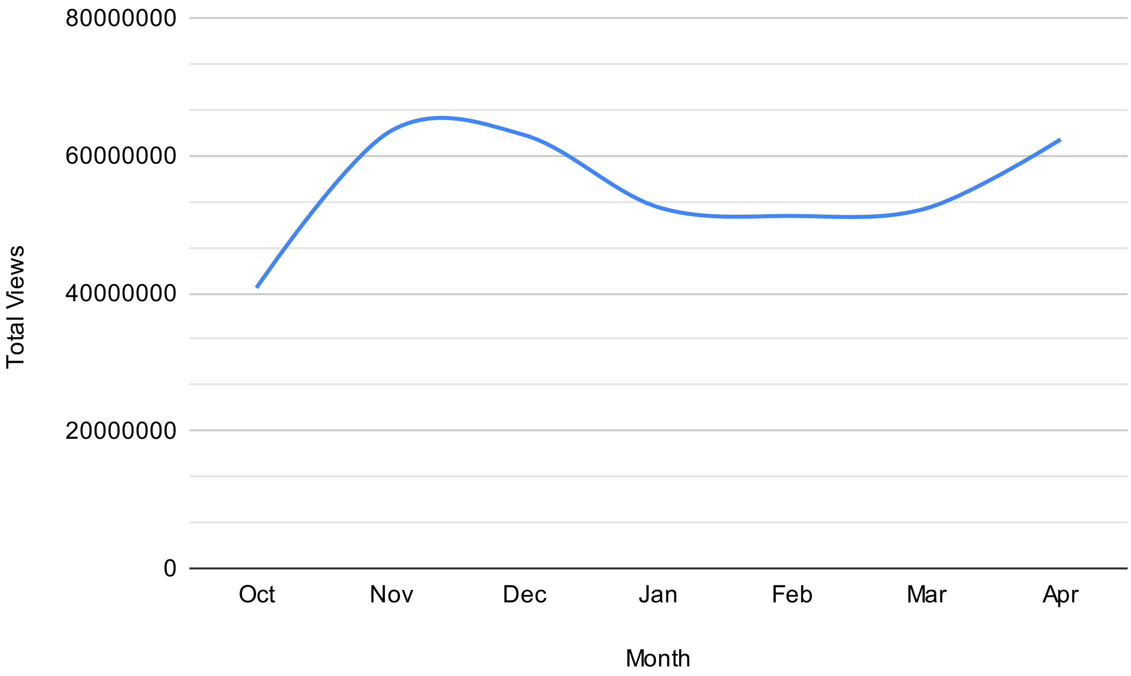


Figure 10: Total views per month

Figure 11 shows the total sales per month. The top three months according to sales are November, December, and April. November contains the Black Friday season, so it's only natural to see consumers cashing more in it than in other months. During December, customers also pay more than usual to buy Christmas gifts. Concerning the month of April, we believe that the large amount of sales in this month is due to the COVID-19 outbreak in the US and the lockdown. Many people during this period did not want to go to local shops to avoid being infected, so they relied on online shopping.

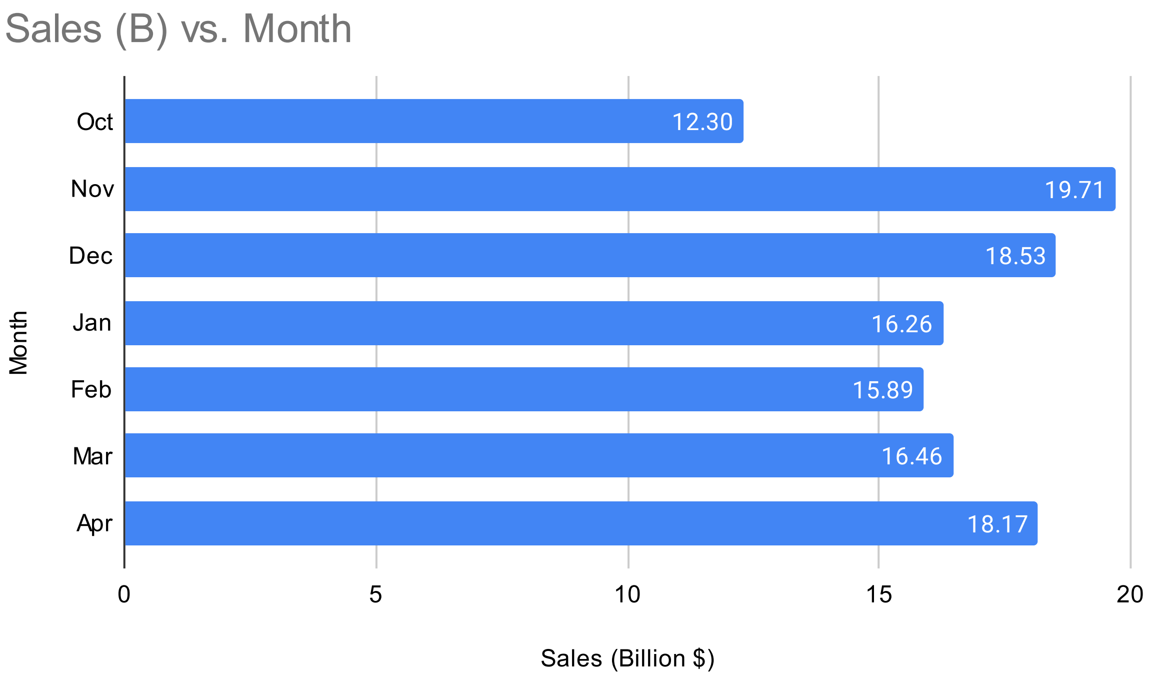


Figure 11: Total sales per month

Next, we want to analyze the consumer spending patterns within each month. In other words, we would like to understand on which days of the month are customers buying more. Figures 12, 13, and 14 shows the customer behaviour patterns daily averaged across data from all seven months. All the results show a clear surge in number of views, add to cart, and purchases by the midpoint of the month. However, we can also notice that the surge in views and add to cart is noticeably smoother than the surge of the purchases. This may be because customers spend a couple of days considering different options and looking at different products but only spend a single day checking out and finalizing their purchase. Now, the question is why customers tend to cash more at the middle of the month? There are many explanations for this behavior. One possible reason could be, because most of the products in the data are not necessities. Therefore, we can assume that customers tend to pay their most critical bills, rent, insurance, etc... at the beginning of the month shortly after they receive their salaries. After this, customers start thinking about the less necessary products which is by the middle of the month. Another explanation could be because the store that produced this data usually offers promotions at the middle of the months invoking more customers to buy.

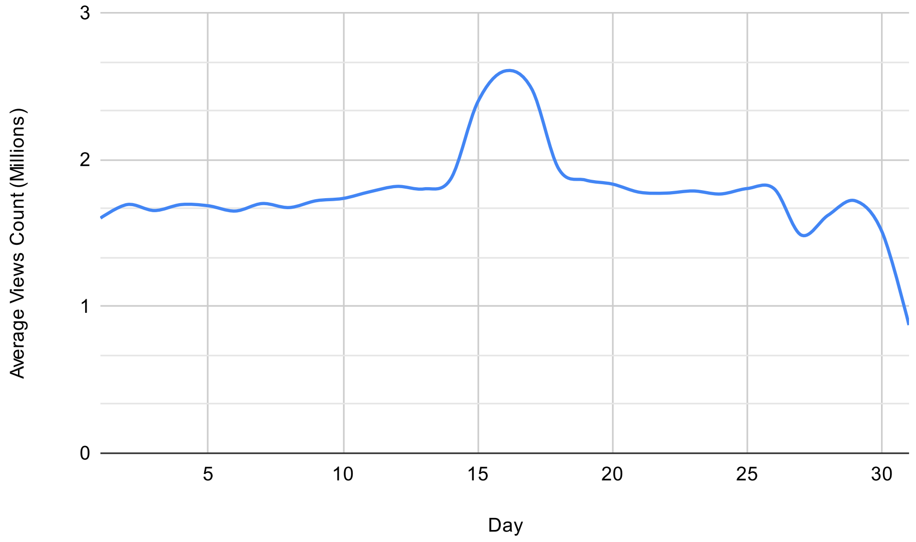


Figure 12: Average daily views

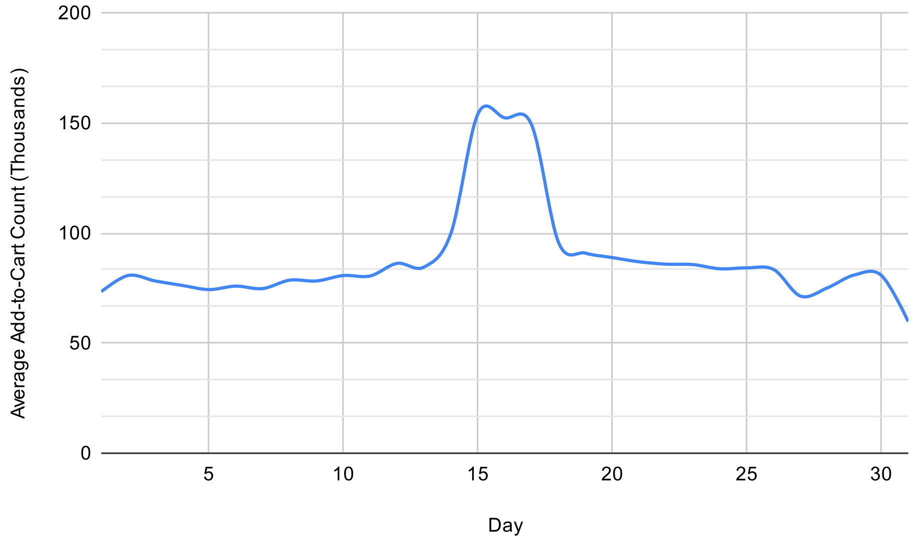


Figure 13: Average daily add to cart

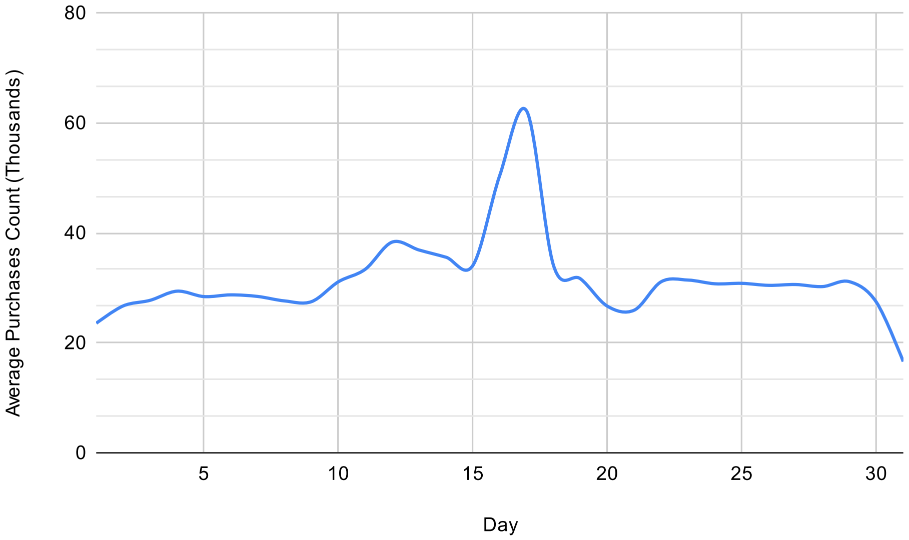


Figure 14: Average daily purchases

Lastly, to take a closer look at the customer purchases variance with time we use figure 15. The results of most months again show that more purchases are made at the middle of the month. The results also showed that among all months, December and April are the months where purchases remain high during second half of the months. For December this may be due to people buying more gifts as Christmas gets closer. For April we believe this is because COVID-19 lockdown altered the customer behavior patterns, because customers were using online shopping a lot more frequent than before due to the lockdown and fear of infection.

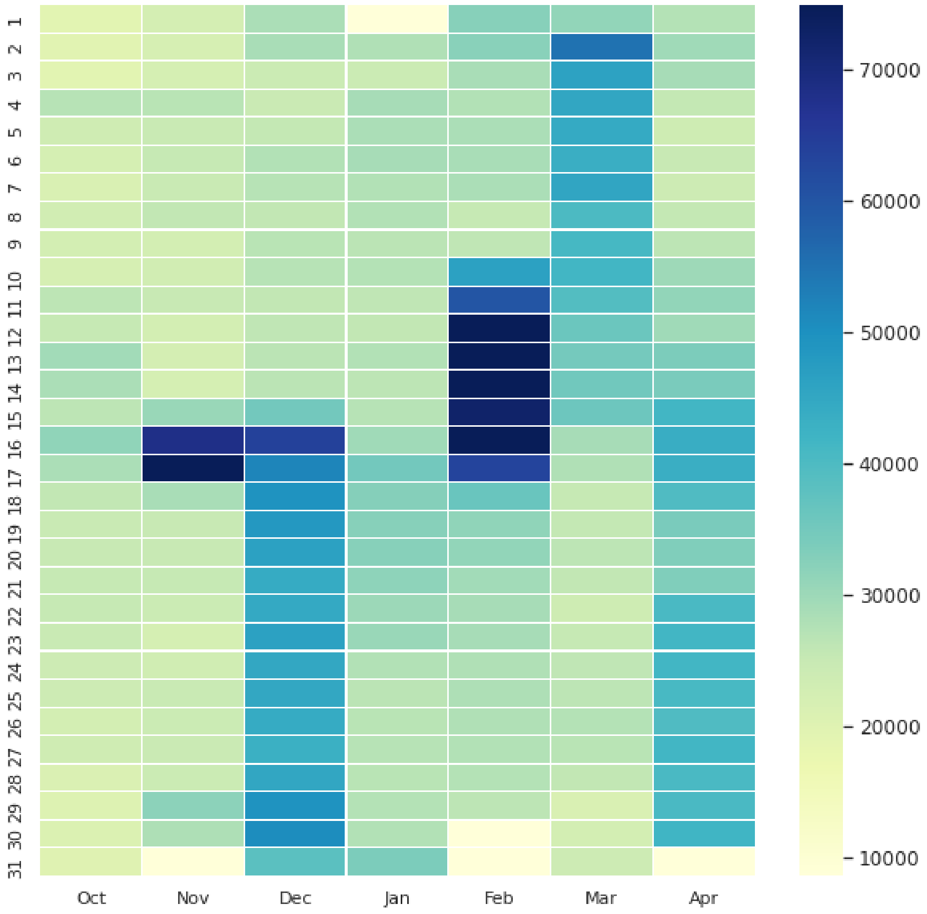


Figure 15: Detailed analysis of customer purchases per day for each month

4 Predicting Customer Behavior

4.1 Business Value

Building on the previous discussion about customer conversion rate and its importance to the business performance, in this section we use AI-based model to predict whether a customer would buy a product he/she added to cart or not.

The business value of such a model is huge. When the model predicts that a customer may not buy the product just added to cart, this piece of information can be used to create unique promotions specific to this customer or to recommend other products that have better conversion rates. Consequently, the business can improve its conversion rates, as well as reducing the expenses paid for unnecessary promotions to customers who were going to by the product anyway.

4.2 Technical Details

A key trend over the past few years has been session-based recommendation algorithms that provide recommendations solely based on a user's interactions in an ongoing session, and which do not require the existence of user profiles or their entire historical preferences. SparkML was used to implement a machine learning model that predicts whether a user will buy a product given the user's current session history of cart events. The problem was formulated as a binary classification problem as there are only 2 possible outcomes.

Starting from raw data, all events other than 'cart' and 'purchase' were dropped as they were assumed to be irrelevant to the buyer decision and all the events in a session are counted to obtain the "activity_count" column. Then the 'purchase' event is transformed into a boolean column named "is_purchase", this column will be used as the label in training after further processing. The dataframe is then grouped by "user_session" and "product_id" and the maximum value of "is_purchase" is calculated among each group, this value then replaces the original "is_purchase" so that the session's final result is assigned to all session's actions. The "event_type" column is then dropped due to redundancy as the purchase event is represented in the "is_purchase" column and all the remaining events are 'cart' events. "event_time" is then transformed to obtain the "week_day" column as we believe that it has an effect on the buyer's decision. "category_code" is split into two columns indicating category and subcategory.
The following columns are then selected to proceed to the pipeline stage:

1. brand
2. price
3. is_purchase
4. event_weekday
5. category_code_level1
6. category_code_level2
7. activity_count

The pipeline stage consists of three indexers for each of the following categorical columns:

1. brand
2. category_code_level1

3. category_code_level2

This transforms the categorical features into numerical features which is the VectorAssembler which is responsible for combining all the feature columns into a single vector. The last stage in the pipeline is a Gradient Boosted Tree Classifier.
The resulting accuracy was 60% which is not high but acceptable and could be further improved by feeding the data as a time series. Linear SVM and MLP networks were tried but yielded lower accuracies, also one hot encoding was tested along with MinMax Scaling but the result was worse.