Recommendation of E-commerce Commodity

A study on E-commerce behavior from a large online store

Bei Wang

2020.02

1. Introduction of E-commerce in the world

E-commerce on the Internet has now been recognized as the development direction of modern commerce. This is a market with great potential for development and has attractive development prospects. According to statistics from the International Internet Association, in 1997, there were 2 million global e-commerce turnovers of US $ 2.6 billion; by the end of 1999, there were more than 100 million Internet users, and transaction activities through the Internet had reached US $ 43 billion. E-commerce, as a new business transaction method, has become the most critical driving force for future economic growth. Therefore, a huge e-commerce market has been formed around the world. In Asia, e-commerce transactions exceeded $ 3 billion by the end of 1999 and are growing rapidly. The complete two-way information communication, flexible transaction methods and fast delivery methods in the e-commerce market will bring huge economic benefits to people and promote a large increase in social productivity. The widespread implementation of e-commerce has greatly accelerated the circulation of goods throughout the society, especially enabling small and medium-sized enterprises to enter the international market and participate in competition at a lower cost. E-commerce also provides consumers with more consumption choices, so that consumers get more benefits.

2. The problem of e-commerce business and why recommendation

E-commerce provides a huge market to the enterprises, compared to traditional trade, E-commerce is easier and lower investment, so it attracted countless people to join the business, but meanwhile, it increases the competition to each other.

To customers, e-commerce is the most convenient way to shop. It is not just limited to the purchase of clothing, accessories, electronics, cosmetics, etc., but also to purchase something like large furniture, vegetables, fruits, a freshly prepared meal, drinks and print service etc. In areas where the epidemic has occurred, online shopping is also the only way to avoid infection through contact.

Although e-commerce is so convenient, after all, it has the limitations of online browsing after all:

1. Miss your favorite product.
2. Miss products which you may be interested in.
2. Missed the upgraded version of the purchased products, or reduced the price.
3. Missing the surrounding goods of the purchased/viewed products.
2. Lack of personal service experience.
3. Lack of loyalty.

In summary, the recommendation system came into being. It provides solutions to all the above problems and gives users a better shopping experience. It can be applied to large shopping mall websites, daily supermarket/Grocery stores websites, restaurant websites, takeaway websites and more. Its role is similar not only to the traditional VIP merchandise brochure specially designed for each advanced user, but also adds the unique interactivity of online shopping and the convenience of direct purchase, which makes users feel at the highest level. Be paid attention to and be treated as a VIP, then a sense of loyalty has been built.

3. Date and description

I used a data set that describes the on-line products been view or sale by E-commerce events in a large multi category store:

1. eCommerce behavior data from a large multi-category store

(**https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store**)

This file contains behavior data for two months (October and November 2019) from a large multi-category online store.

Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relations between products and users. It provides variable events：event-time, event-type, product-id, category-id, category-code, brand, price, user-id and user-session.

Although the data is only based on 2 months of on-line store's record, it contains all events when customers do online shopping. We will find out the most attractive factors of online store products after analyzing the data.

4. Date Wrangling

This data set contains 2 csv files: Oct. and Nov.. The size of files are around 9G and 5G, 60 millions rows in each file. To make it fit the capacity of the storage platform, I smaller both of them to 1 million rows each. Those two files contain exactly the same columns, so first of all: I will do data wrangling for two files, and then merge two clean csv files in one which is based on the same column.

- Deal with missing data

  There is some empty data in category_code and brand. At beginning I supposed to fill the data as below :

  1) In a row, if there is category_code data, but missing brand: search the same category_code with non-empty brand in the dataset, backward fill missing brand.
  2) In a row, if there is brand data, but missing category_code: search the same brand with non-empty category_code in the dataset, backward fill missing brand.
  3) In a row, if both category_code and brand name are missing, drop the row.
  4) If there is not the reference value from brand or category_code column by the same product_it to fill the missing value, then fill the value by creating a new column name: Not-available.
  5) Merge two clean files together.

- Select the data in columns

  To get the result of recommendation, I will select data from some columns to analysis as below:

  1. Event_type: contains 3 kinds of customer behaviors: 'view', 'purchase', 'cart'
  2. Product_id: the number of product
  3. Category_id: the number of categord
  4. Category_code: category name

5. Brand: brand of product
6. Price: price of product
7. User_id: the permanent user's id number

How does the recommendation work? It will show recommendation products according by customer's behavior:

1. Most frequalice view or ordered categories
2. Category which customer has been view or ordered
3. Relative products of view or ordered
4. New version of bought products
5. Top sale product list

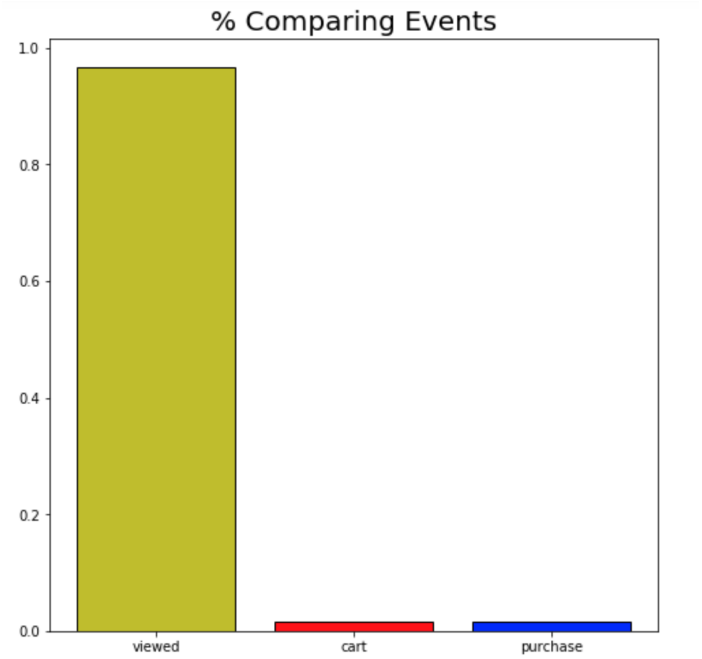
5. Exploratory Data Analysis

The data will display the answers of our concerned, and it also will bring more questions after that. So let's see what the data tells us from a macro perspective. And then analyze it in detail at the micro level.

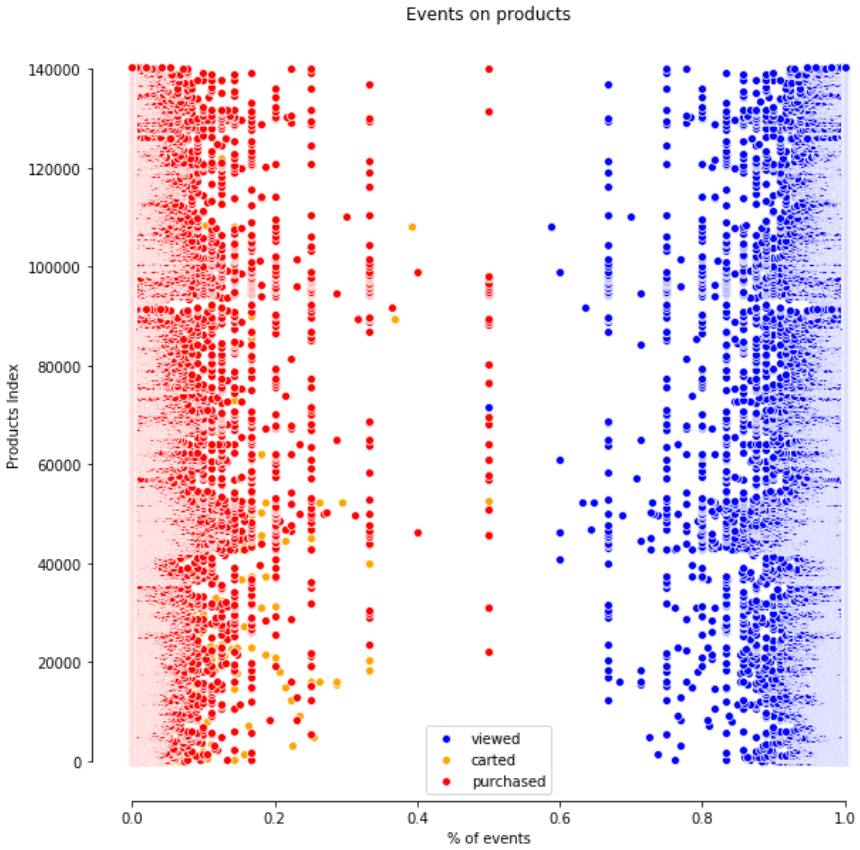
1. Customers behaviors analysis

A. Customer's behaviors comparing:

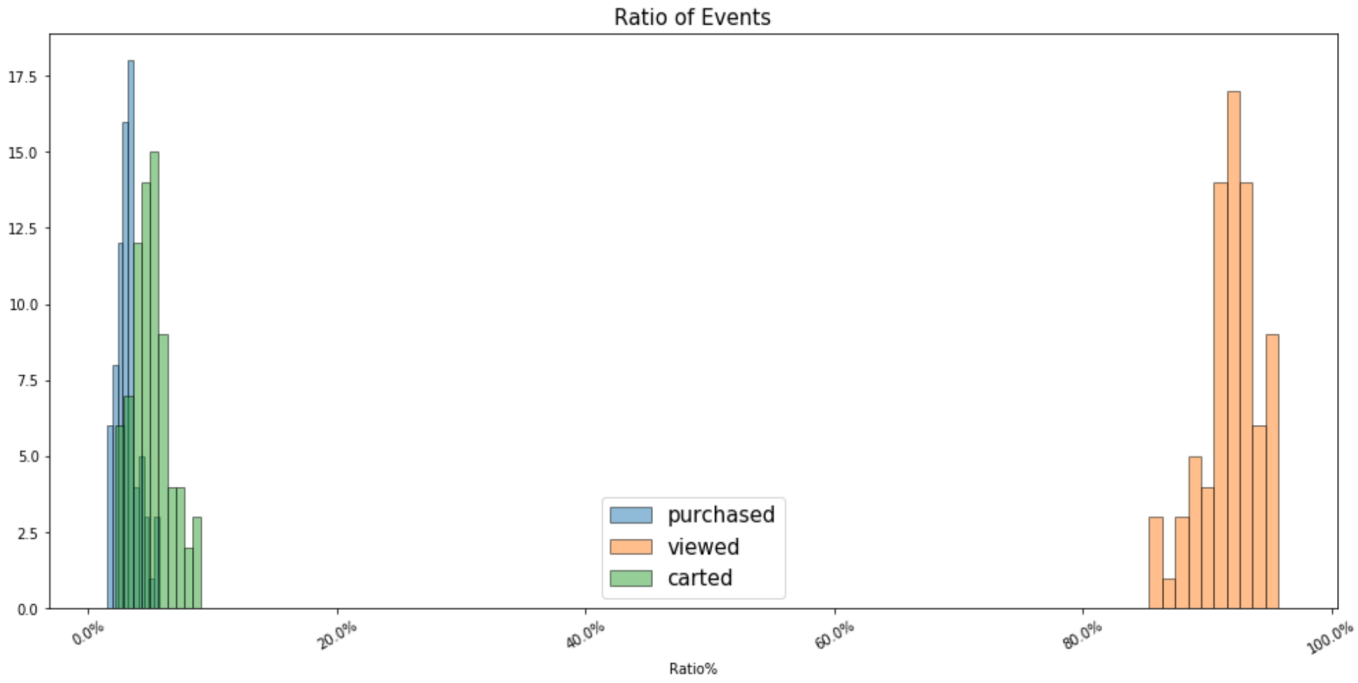
We focus on what customers do when they visit the online shopping website: they are viewing; if they are interested in the product, will put it in the cart; maybe purchase it or give up later. Then my first question is: what is the radio of three kinds of behaviors? Let us see the comparing plot as below:

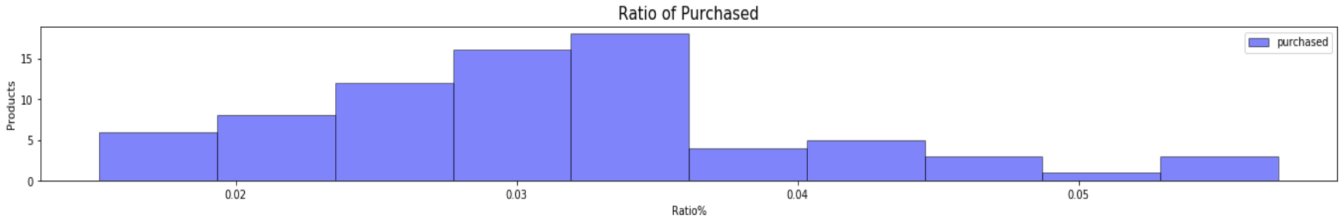


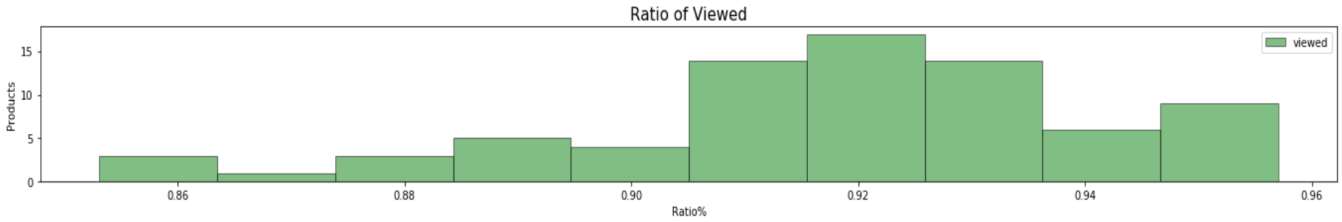
We can see the view behavior is about 97%, and put in cart and purchase are only 1~2% each. The next scatter plot will more intuitively list the proportions of the three behaviors:

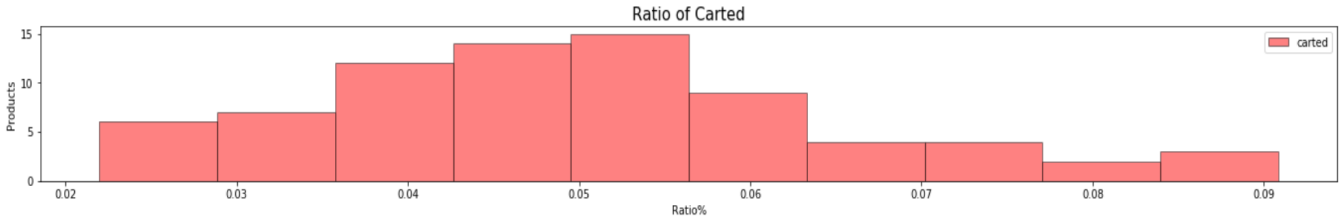


Let's verify whether this ratio is reasonable:





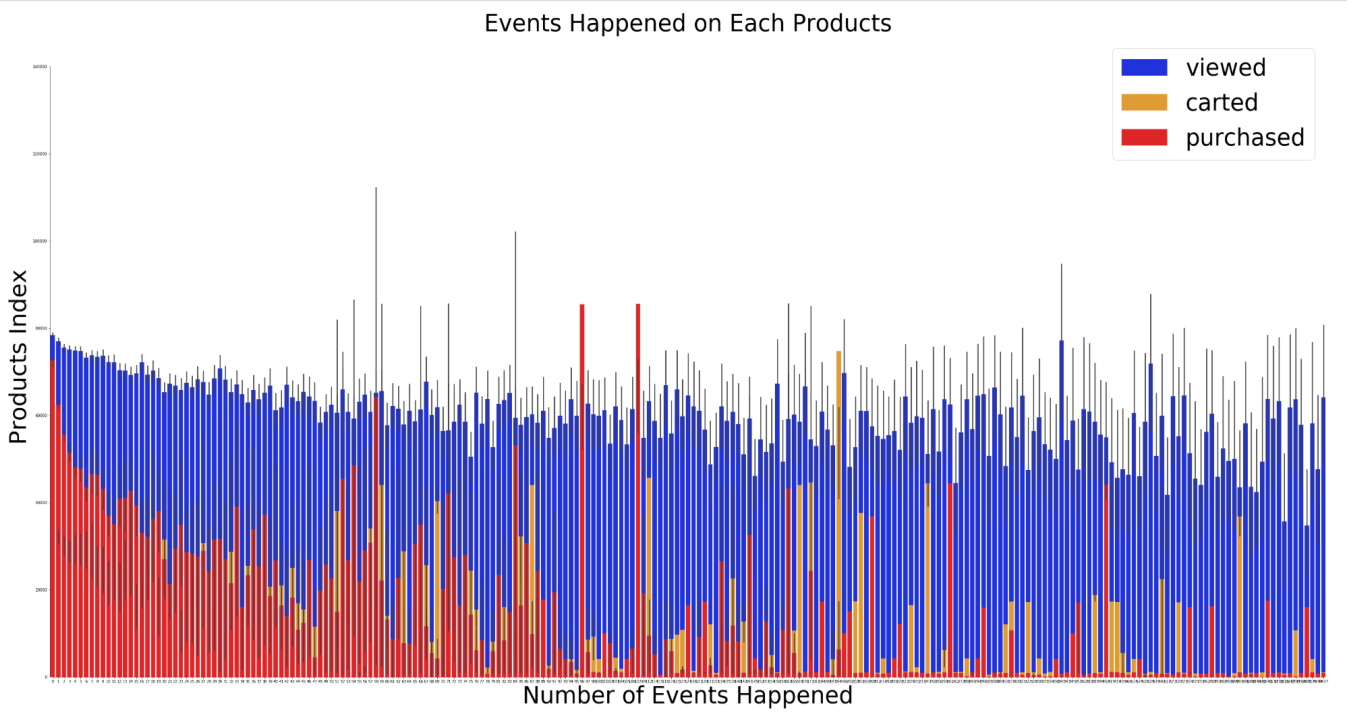




The shape of three behaviors are similar to normal distribution, the ratios are reasonable.

B. Connection of behaviors

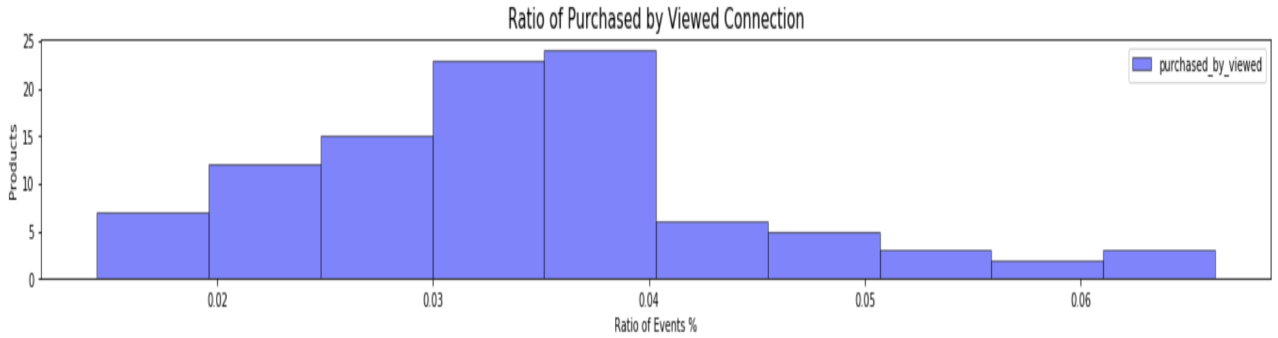
Think about the relative importance of large view events and small purchase events, does that mean view multiple times will cause the second step: put in the cart or directly to purchase? I only took 100 data to do those kinds of analysis. Let's see the stacked plot of three behaviors under the same product:



In this plot, we will find most purchases are based on multiple views, only a few purchases without view. A few cases of large put in cart events only cause a few purchases.

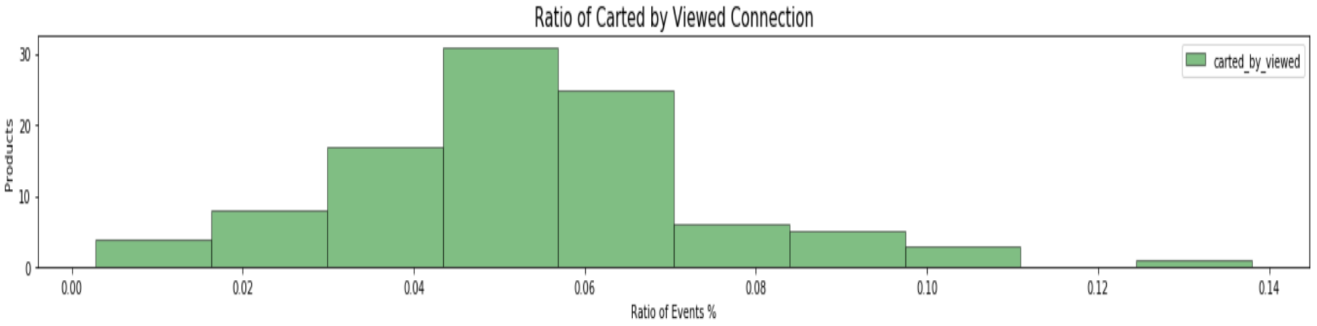
Let's continue analyzing customers' behaviors. Generally, when users make online purchases, they will be accustomed to browsing first, putting them into the shopping cart, and finally paying for the product. But is this hypothesis true? Is there a necessary connection between the three behaviors? I was interested in this, so I made an analysis chart as follows:

01. Ratio of purchased after viewing.



Most products were purchased after viewing, this kind of ratio is about 4.1%. After that, this kind of ratio goes up but the amount of purchased products goes down, maybe because some customers already clear what to buy before browsing the e-shopping website.

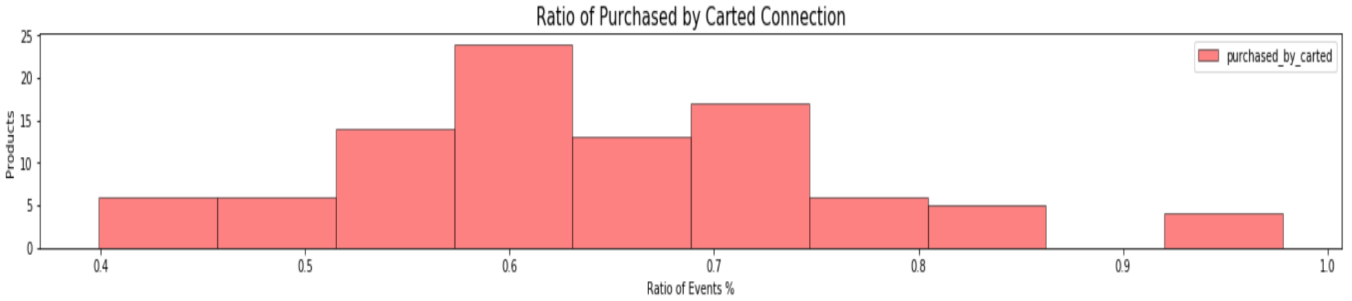
02. Ratio of carted before viewing.



The ratio starts from 0.5%, top ratio is about 5%. There is a gap in 12%.

Based on 100 data sizes, Xiaomi's product with a price $29.56 is 0% ratio that was not being put in the cart.

03. Ratio of Purchased before putting in cart.



What's the toppest product being purchased after putting it in the cart based on 100 data sizes? There are Apple smartphones.

What is the 0% ratio of products at the beginning?

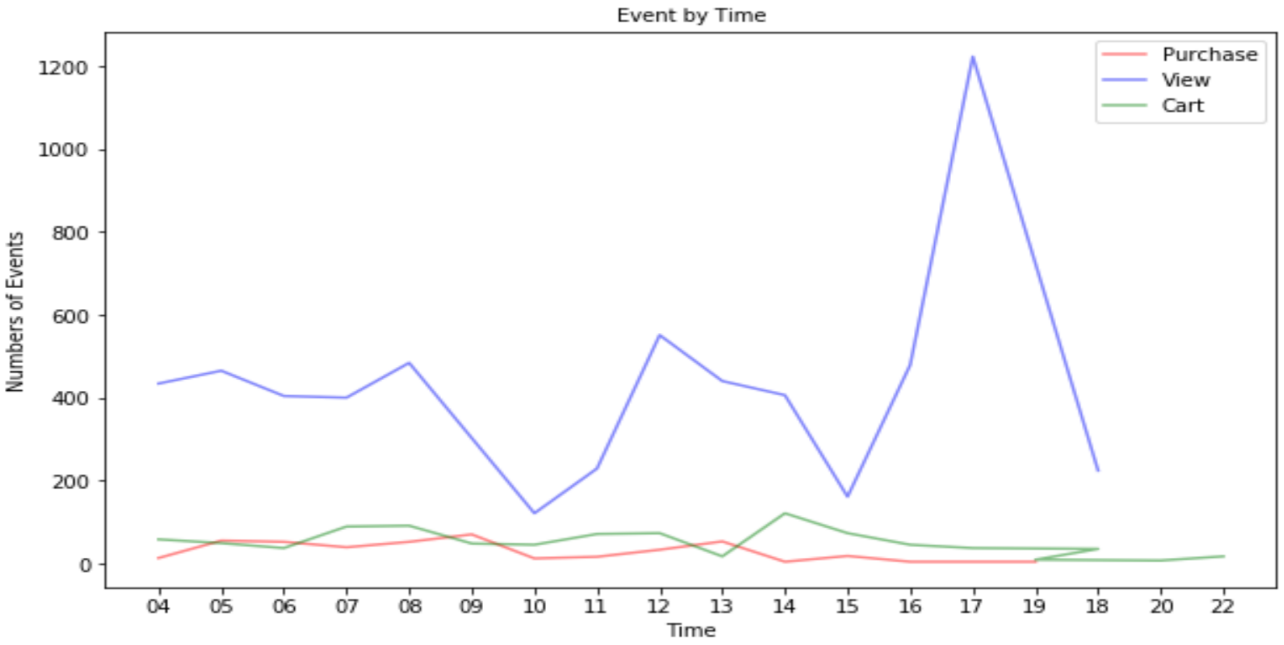
It's Xiaomi's product with a price $29.56 is 0% ratio that was not being purchased. It may be caused by competitiveness if it is much lower than other similar products.

The product in the ratio gap is Samsung's smartphone. Higher price ($771.94) may be the reason why customers cannot be determined to buy.

Three shapes of ratio of events is similar to a normal distribution,therefore, we can conclude that most online shopping users operate in accordance with general shopping habits.

C. Customer's behaviors happened in time of the day

In 24 hours a day, I am interested to know when customers have the highest rate of purchase, when they browse the most products, and when customers have the most desire to buy: put the products in the shopping cart? So I merged the all events by time of the day, did the analysis and displayed the results as shown below:



Top view happens at 17:00, that's the end of the working day. Some people will leave the office laterer to avoid the rush hour, some people just finished the work and want to relax a while before leaving the office, online view products are a good way to spend the time. Second top time is 12:00-13:00, lunchtime is another good time to do online window shopping.

Top cart happens at 14:00, after viewing the products online, most people finished the lunch and made a half decision to purchase something, but have to go back to work, so put the products in the cart.

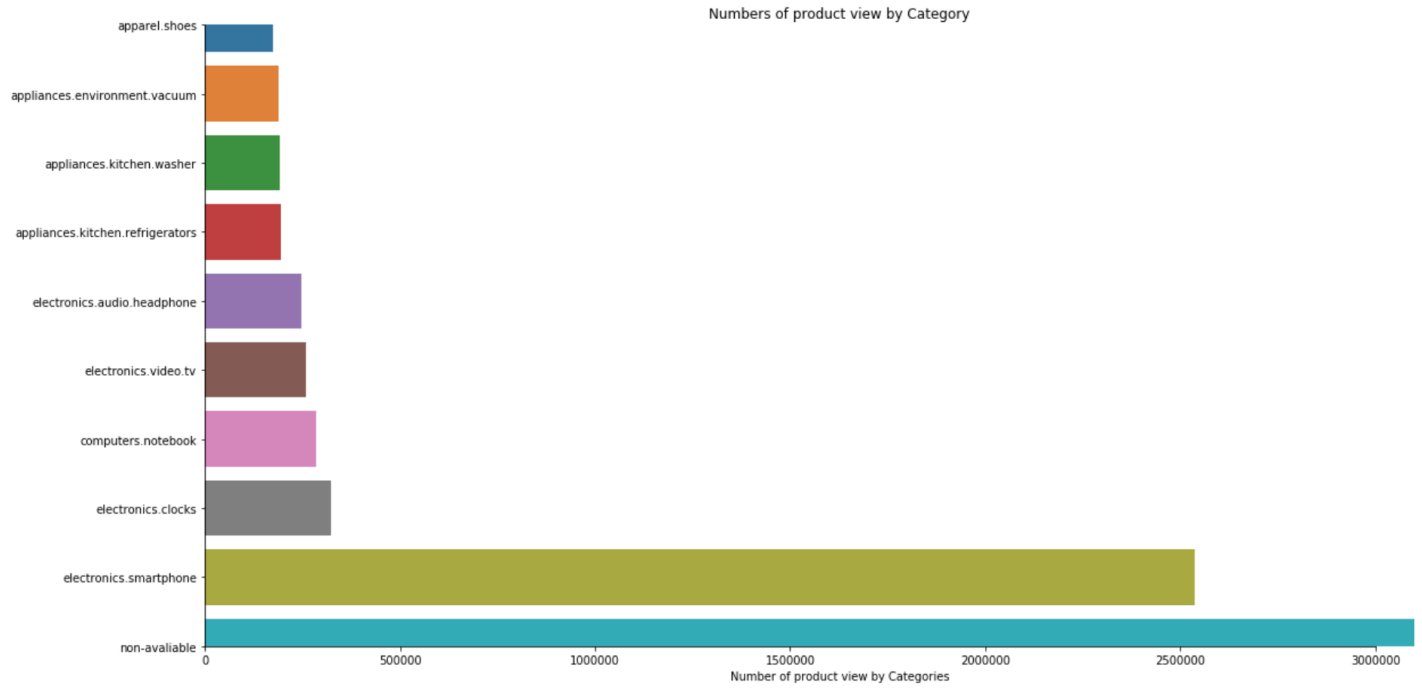
Top purchase happens at 9:00, it might be the housewives buy house need time, 8:00 is the third top view, and then purchase happens at 9:00, it can explain after 1 hours viewing, purchases happen.

Analyzing a customer's behavior based on the time of the day is very meaningful. It will help us to plan the strategy of product promotions, new products advertising, and focus on different customers to recommend products at different times of the day.

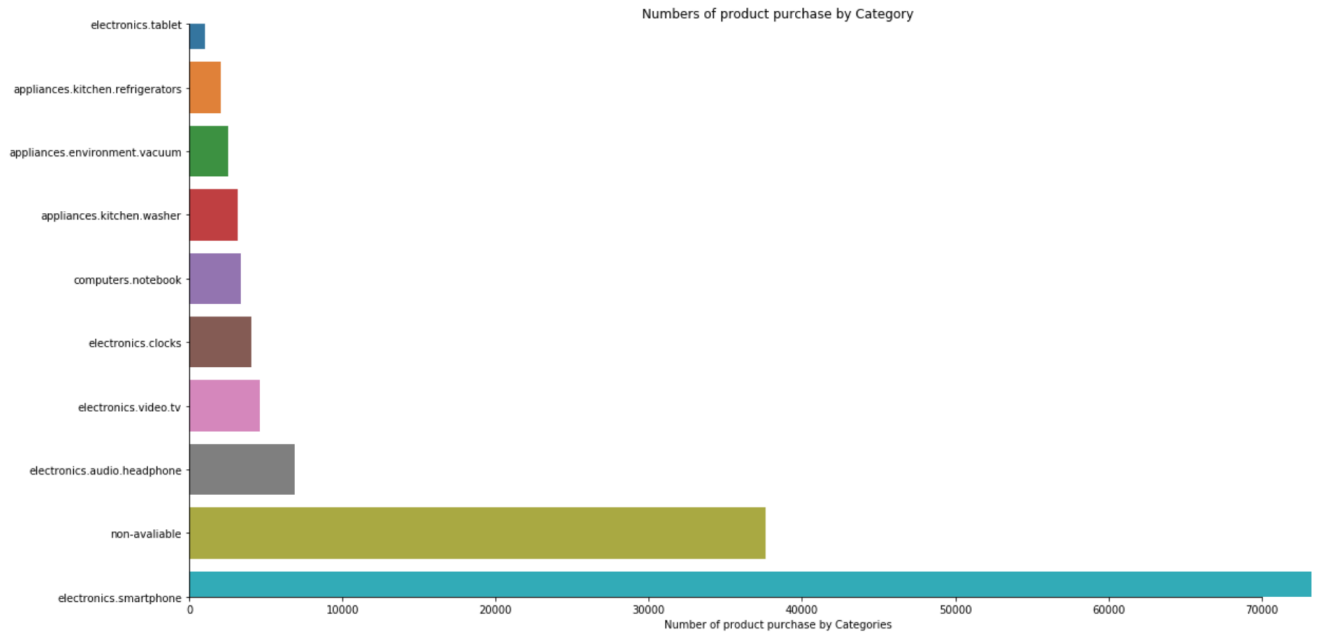
2. Top selling analysis

Now I am using whole clear data to continue analysis.

So what's the most popular category for view, cart and purchase?

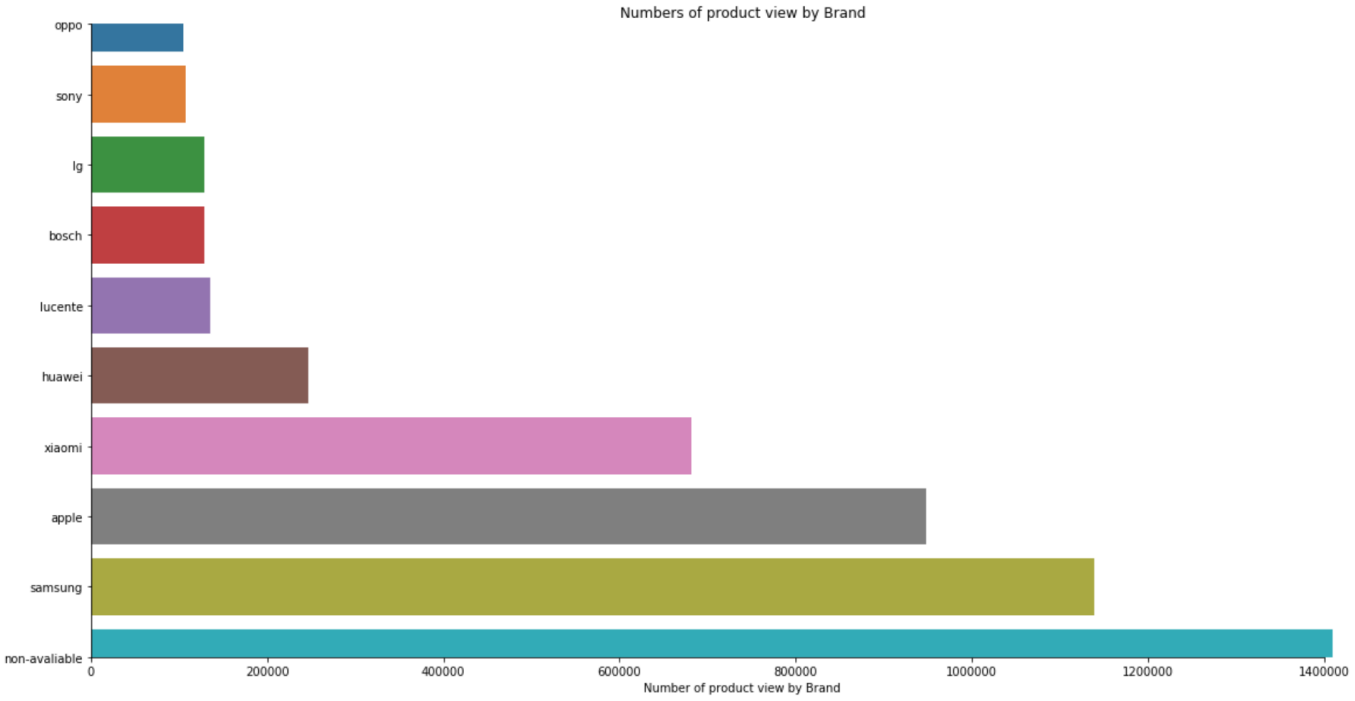

Numbers of product view by Category
apparel.shoes
appliances.environment.vacuum
appliances.kitchen.washer
appliances.kitchen.refrigerators
electronics.audio.headphone
electronics.video.tv
computers.notebook
electronics.clocks
electronics.smartphone
non-avaliable
0
500000
1000000
1500000
2000000
2500000
3000000
Number of product view by Categories

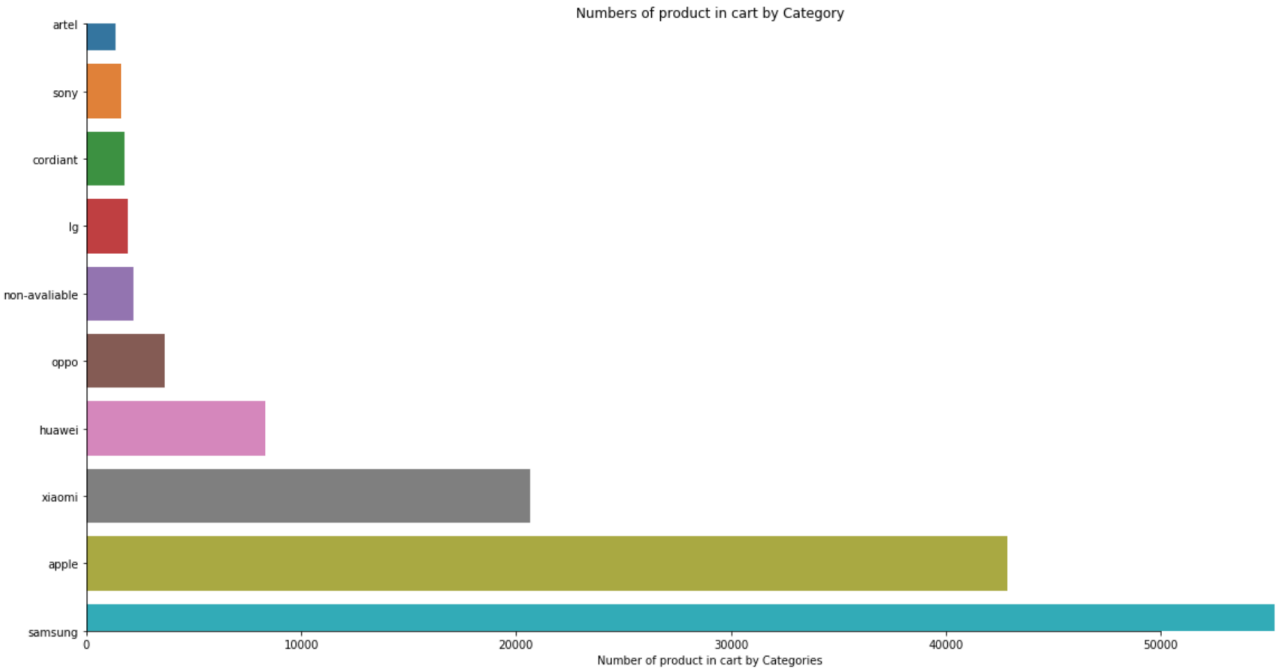

Numbers of product in cart by Category
appliances.kitchen.refrigerators
electronics.tablet
appliances.kitchen.washer
computers.notebook
appliances.environment.vacuum
electronics.clocks
electronics.video.tv
electronics.audio.headphone
non-avaliable
electronics.smartphone
0
20000
40000
60000
80000
100000
Number of product in cart by Categories

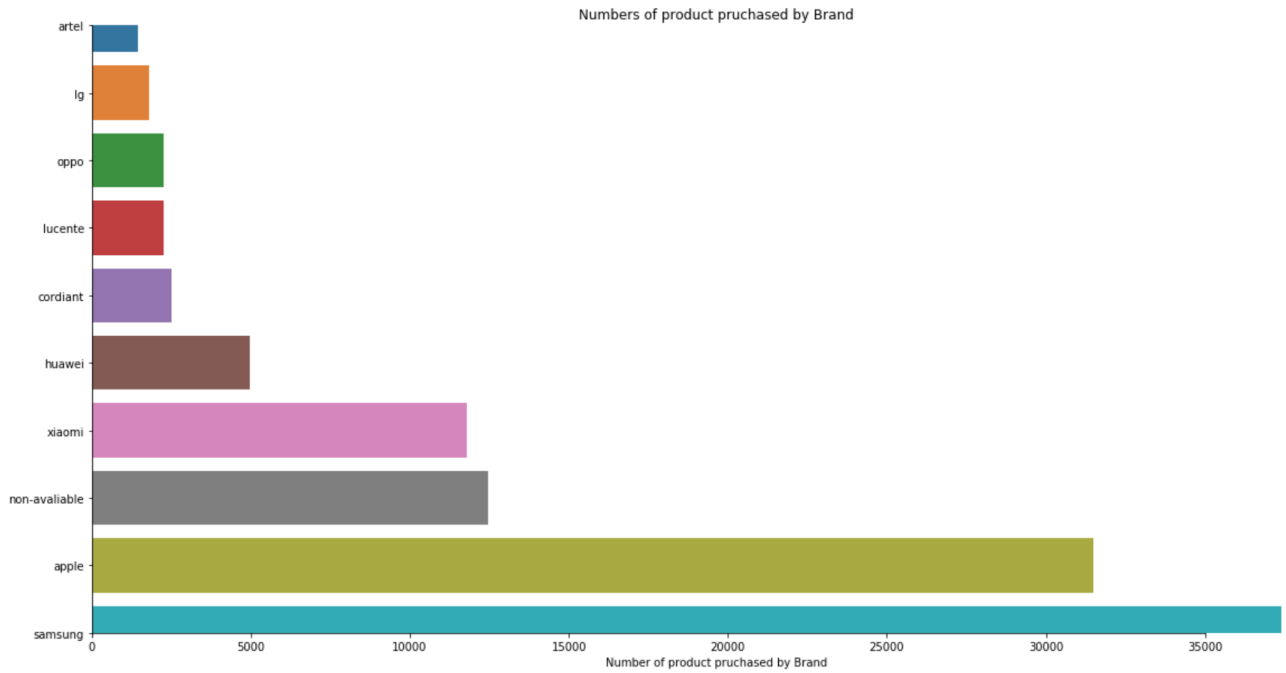


No matter which kind of events, ‘electronics smartphone’ alway occupies the top position of the category being viewed, put in cart and purchased.

Then which brand is the top in three events?







Very interesting, if the top category is smartphones, the most popular brand is not Apple but Samsung, and the numbers of Samsung being viewed, in cart or purchased are far more than Apple. Xiaomi and Huawei keep following the top two brands, taking No.3 and No.4 positions. However, there are all brands of smartphone and electrical brands. This result is consistent with the results of the previous category analysis.

Now we know the most popular category of this large multiple E-commerce is electrical smartphone, brands are Samsung, Apple, Xiaomi

and Huawei. Customers like to view multiple times before buying, if they familiarize themselves with the product, they may directly buy it without view. However, people spend a lot of time viewing products, compared to the lower purchase ratio, we need to concerent how to increase purchase numbers and shorter the customer view time on irrelevant products.

Recommendation may help customers stop wasting time on irrelevant products viewing, but spend time on system recommended products that he/she may like categories or brands, have bought or viewed many times.

6. Recommendation Result

We already know that some behaviors have occurred when customers browse products, from which we can know the categories of products that customers have used or preferred. Next, we hope to use this analysis to predict out the categories of products that customers will be interested in, and finally generate recommendations for each user.

Recommendation system implemented by collaborative filtering from two ways: Item based and User based. Regarding our dataset situation, I select item based collaborative filtering to implement.

In the three customer behaviors, ‘View’ collects most data in 3 events, so I will focus on 'view' to do analysis. The steps to do recommendations as below:

1. Collect and organize information on users and products; focus on products which have been ‘view’.
2. Compare items that have been viewed by user A and other users.
3. Create a function that finds products that user A has not viewed, but which similar users have.
4. Rank and recommend.
5. Evaluate and test.

After organized the dataset, I using SVD to reduce dimension of the dataset, finally got a sparse matrix:

product_id user_id	1002098	1002099	1002101	1002102	1002398	1002484	1002524	1002528	1002531	1002532	...	54900012	54900
183503497	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
184265397	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
208669541	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
214470341	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
216064734	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
483782965	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
483785194	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
483806238	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
483823174	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	
483869744	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	

Followed the steps, we finally get the users and product's rank to recommend like this:

```
array([[  10, 1471,   14,   11,  756]])
```

It shows if a user A has viewed any product by index name in the array, it is probably to view others in the array. To clear this result, I post the relative information as below:

```
product_id                    1001588
category_code    electronics.smartphone
brand                           meizu
price                          128.25
Name: 10, dtype: object

product_id                    1002098
category_code    electronics.smartphone
brand                          samsung
price                           370.64
Name: 756, dtype: object

product_id                    1001588
category_code    electronics.smartphone
brand                           meizu
price                          128.28
Name: 14, dtype: object
```

```
product_id                      1002099
category_code    electronics.smartphone
brand                            samsung
price                             370.41
Name: 1471, dtype: object
```