Taobao-User Behavior Analysis

1. About the Data and Analysis Background

1.1 Analysis Background

After the initial rough development, China's e-commerce industry has gradually transformed from the mode of selling what is available to the mode of refined operation. Through in-depth analysis of a large amount of data, it is found that the user needs behind the data gradually accompany the operation of e-commerce. With the development of the e-commerce industry becoming more and more mature, coupled with the emphasis on data and the improvement of the data based platform, the database and the collected data are more complete, providing strong support for analysis. Decision-making is becoming more and more important, and in this context, this paper conducts analysis based on e-commerce user behavior data.

1.2 Introduction to the Data

"UserBehavior.csv" is a Taobao-user behavior dataset provided by Alibaba for the study of implicit problems. This dataset contains all behaviors (including clicks, purchases, add-ons, likes) of about one million random users between November 25, 2017 and December 3, 2017. The organization of the dataset is similar to MovieLens-20M, that is, each row of the dataset represents a user behavior, which consists of user ID, product ID, product category ID, behavior type, and timestamp, separated by commas. A detailed description of each column in the dataset is as follows:

File name	Introduction	Contained features
UserBehavior.csv	It contains all user behavior data	User ID, Product ID, Product Category ID, Behavior Type, Timestamp

Columns	Introduction
User ID	Integer, serialized user ID

Columns	Introduction
Product ID	Integer, serialized product ID
Category ID	Integer, serialized product category ID
Behavior Type	String, enum type, including ('pv', 'buy', 'cart', 'fav')
Timestamp	The timestamp when the behavior occurred

There are four types of user behaviors, they are described as follows:

Behavior Type	Introduction
pv	Click the product details page
buy	Buy products
cart	Add product to the cart
fav	Add product to the list of the favorite products

Some notes on the dataset size are as follows:

Dimensions	Quantity
Amount of users	987,994
Amount of products	4,162,024
Amount of product categories	9,439
Amount of behaviors	100,150,807

2. Prepare and clean the data

2.1 Prepare the data

I used Navicat to import the data. Because the dataset is too big, so I just used 1000,000 records to do the analysis.

2.2 Clean the data using SQL

To get better analysis, I still used Navicat and MySQl to explore the data' integrity.

2.2.1 Import the data in the BigQuery and check the structure of the data.

Let's have a look first, how does the dataset look like:

f1	UserID	ProductID	CategoryID	BehaviorType	Timestamps
57312886	4	1762079	1686309	pv	1511668276
70436363	9	5014921	1729993	pv	1512031361
32298437	11	5015102	4244487	pv	1512053747
1426411	17	2581080	3645362	pv	1512225175
16841765	18	1755153	982926	fav	1511753961
74656714	25	1039428	4217906	pv	1512143179
59569833	27	4955960	1986754	pv	1512301516
32843519	29	1337505	2321441	pv	1511843568
42211013	29	255259	3579754	pv	1511875625
76660199	41	3618449	4217906	pv	1512284708
98224218	43	5068248	3695050	fav	1512308007
74910711	43	2947452	3607361	pv	1511607946

We notice that there are in total 6 columns. The first column is the row information from the last step, which we can delete.

```
ALTER TABLE UserBehavior_train
DROP COLUMN f1;
```

Now the table looks like this:

UserID	ProductID	CategoryID	BehaviorType	Timestamps
4	1762079	1686309	pv	1511668276
9	5014921	1729993	pv	1512031361
11	5015102	4244487	pv	1512053747
17	2581080	3645362	pv	1512225175
18	1755153	982926	fav	1511753961
25	1039428	4217906	pv	1512143179
27	4955960	1986754	pv	1512301516
29	1337505	2321441	pv	1511843568
29	255259	3579754	pv	1511875625
41	3618449	4217906	pv	1512284708
43	5068248	3695050	fav	1512308007
43	2947452	3607361	pv	1511607946

2.2 Check the duplications

After analyzing the field content, I set the UserID, ProductID and TimeStamp as the joint primary key. After verification, there is no data duplication.

```
SELECT UserID, ProductID, Timestamps
FROM UserBehavior_train
GROUP BY UserID, ProductID, Timestamps
HAVING COUNT(1) > 1;
```

2.3 Check the missing values

```
SELECT COUNT(UserID), COUNT(ProductID), COUNT(CategoryID),
COUNT(BehaviorType), COUNT(Timestamps)
FROM UserBehavior_train;
```

From the following results we can see that there are no missing values.

COUNT(UserID)	COUNT(ProductID)	COUNT(CategoryID)	COUNT(BehaviorType)	COUNT(Timestamps)
500754	500754	500754	500754	500754

2.4 Check the format

The time data here is a timestamp, we need to convert this data to our daily used time format. First of all, two columns need to be added: Datee and Timee. Then the timestamp will be converted into data and varchar type.

```
ALTER TABLE UserBehavior_train
ADD Datee date,
ADD Timee varchar(10);

UPDATE UserBehavior_train
SET Datee = FROM_UNIXTIME(Timestamps, '%Y-%m-%d'), Timee = FROM_UNIXTIME(Timestamps,'%k');
```

Here is the result:

UserID	ProductID	CategoryID	BehaviorType	Timestamps	Datee	Timee
4	1762079	1686309	pv	1511668276	2017-11-26	4
9	5014921	1729993	pv	1512031361	2017-11-30	9
11	5015102	4244487	pv	1512053747	2017-11-30	15
17	2581080	3645362	pv	1512225175	2017-12-02	15
18	1755153	982926	fav	1511753961	2017-11-27	4
25	1039428	4217906	pv	1512143179	2017-12-01	16
27	4955960	1986754	pv	1512301516	2017-12-03	12
29	1337505	2321441	pv	1511843568	2017-11-28	5
29	255259	3579754	pv	1511875625	2017-11-28	14
41	3618449	4217906	pv	1512284708	2017-12-03	8

2.5 Check outliers

Check if the date is within the specified range of the dataset, which is 2017-11-25 to 2017-12-3.

```
SELECT MAX(Datee), MIN(Datee)
FROM UserBehavior_train;
```

There are outliers, which we will delete.

MAX(Datee)	MIN(Datee)
2025-10-27	2010-01-02

```sql
SELECT COUNT(1)
FROM UserBehavior_train
WHERE Datee < '2017-11-25' OR Datee > '2017-12-3'

DELETE FROM UserBehavior_train
WHERE Datee < '2017-11-25' OR Datee > '2017-12-3'
```

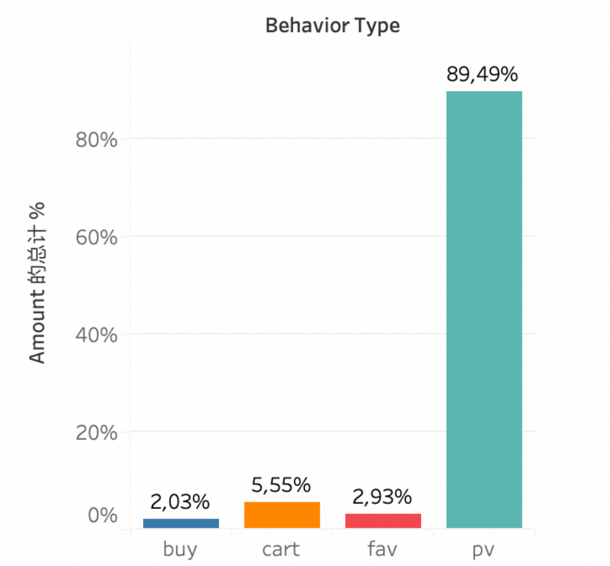
After that, we got about 4500 records less.

3. Data Analysis

3.1 Customer Behavior Analysis

```sql
SELECT BehaviorType, COUNT(1) AS 'Amount'
FROM UserBehavior_train
GROUP BY BehaviorType;
```

Amount of different user behaviors



It is found that user clicks account for 89.63%, while purchases only account for 2.03% of all data, and the conversion rate from browsing to purchasing is only 2.27%. What is the reason for the low conversion rate?

UserID	ProductID	Click	Add to cart	Add to favorite	Buy
4	1762079	1	0	0	0
9	5014921	1	0	0	0
11	5015102	1	0	0	0
17	2581080	1	0	0	0
18	1755153	0	0	1	0
25	1039428	1	0	0	0
27	4955960	1	0	0	0
29	255259	1	0	0	0
29	1337505	1	0	0	0
41	3618449	1	0	0	0
43	1243304	1	0	0	0
43	2922853	1	0	0	0
43	2947452	1	0	0	0
43	5068248	0	0	1	0
44	2067676	1	0	0	0
44	2755138	0	0	1	0

When users come to the website, they first need to browse the homepage, check the recommended products or enter keywords to find the products they like, then add to the cart or to the favorite, and finally click on such as purchase and payment. We only notice that the conversion rate is low from browsing to shopping. Where does problems occur in this process needs further analysis. The funnel analysis method is used to analyze where is the low conversion rate, which leads to the overall low conversion rate.

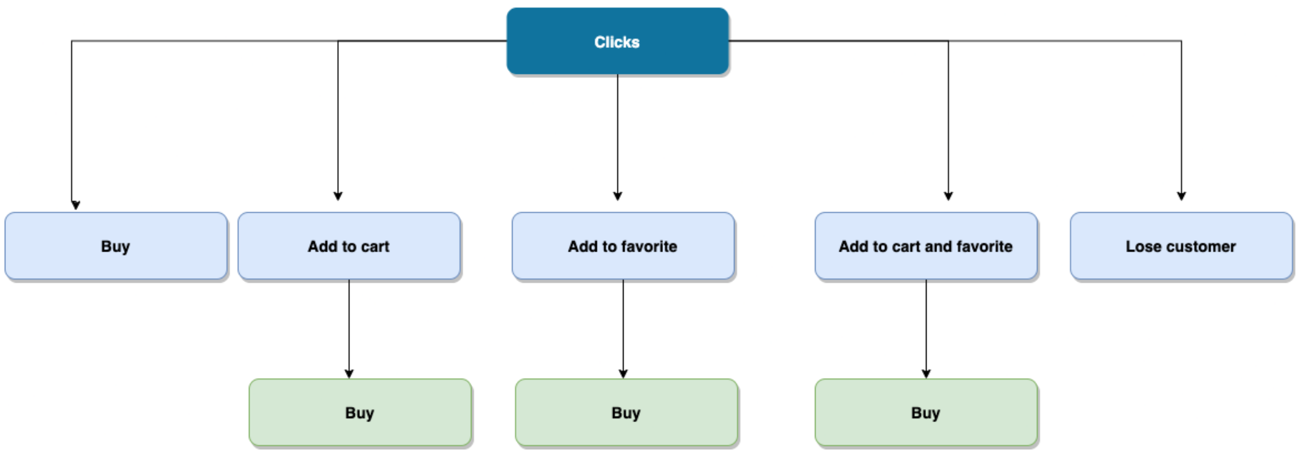
To answer this question, I will analyze it according to different user behavior path.

First of all, the user's behavior record on each product, including the number of “clicks”, the number of “add to cart”, the number of “add to favorite” and the number of “purchases”.

As the next step, I will calculate the sum of each of the following behavior and calculate the conversion rate:

The amount of “clicks”;

The amount of “clicks” and “buy”;



```sql
CREATE TABLE BehaviorPath
AS
SELECT UserID, ProductID,
SUM(CASE WHEN BehaviorType = 'pv' THEN 1 ELSE 0 END) as 'Click',
SUM(CASE WHEN BehaviorType = 'cart' THEN 1 ELSE 0 END) as 'Add to cart',
SUM(CASE WHEN BehaviorType = 'fav' THEN 1 ELSE 0 END) as 'Add to favorite',
SUM(CASE WHEN BehaviorType = 'buy' THEN 1 ELSE 0 END) as 'Buy'
FROM UserBehavior_train
GROUP BY UserID, ProductID;
```

The amount of “clicks” and “add to cart”;

The amount of “clicks” and “add to cart” and “buy”;

The amount of “clicks” and “add to favorite”;

The amount of “clicks” and “add to favorite” and “buy”;

The amount of “clicks” and “add to cart” and “add to favorite”;

The amount of “clicks” and “add to cart” and “add to favorite” and “buy”;

The amount of “clicks” and “lose customer” .

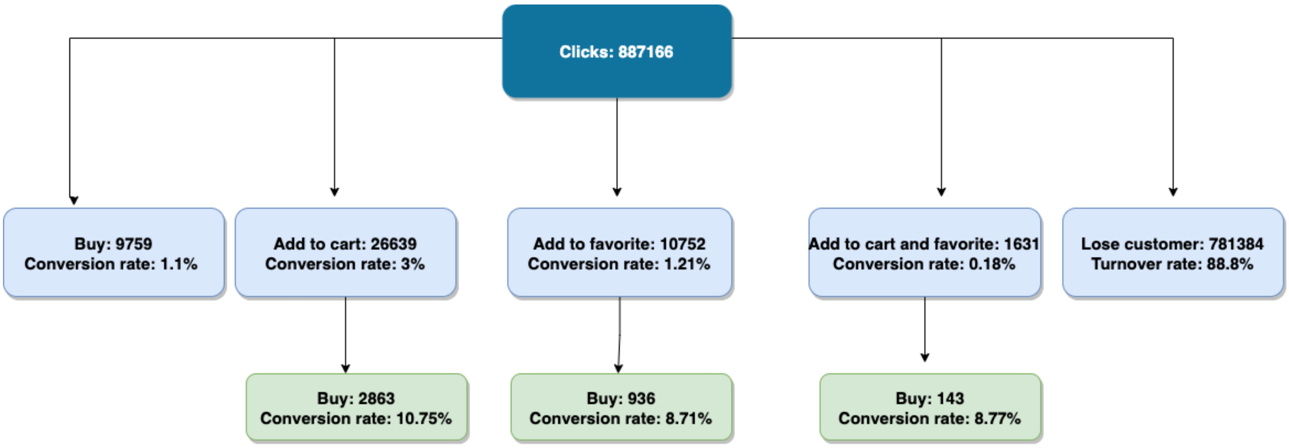
```sql
-- Amount of clicks: 887166
SELECT SUM(Click) FROM BehaviorPath;

-- Amount of click-buy: 9759
SELECT SUM(Buy) FROM BehaviorPath
WHERE Click > 0 AND Add_to_cart = 0 AND Add_to_favorite = 0 AND Buy > 0;
```

```sql
-- Amount of click-Add_to_cart: 26639
SELECT SUM(Add_to_cart) FROM BehaviorPath
WHERE Click > 0 AND Add_to_favorite = 0 AND Add_to_cart > 0;

-- Amount of click-Add_to_cart-buy: 2863
SELECT SUM(Buy) FROM BehaviorPath
WHERE Click > 0 AND Add_to_cart > 0 AND Add_to_favorite = 0 AND Buy > 0;
```

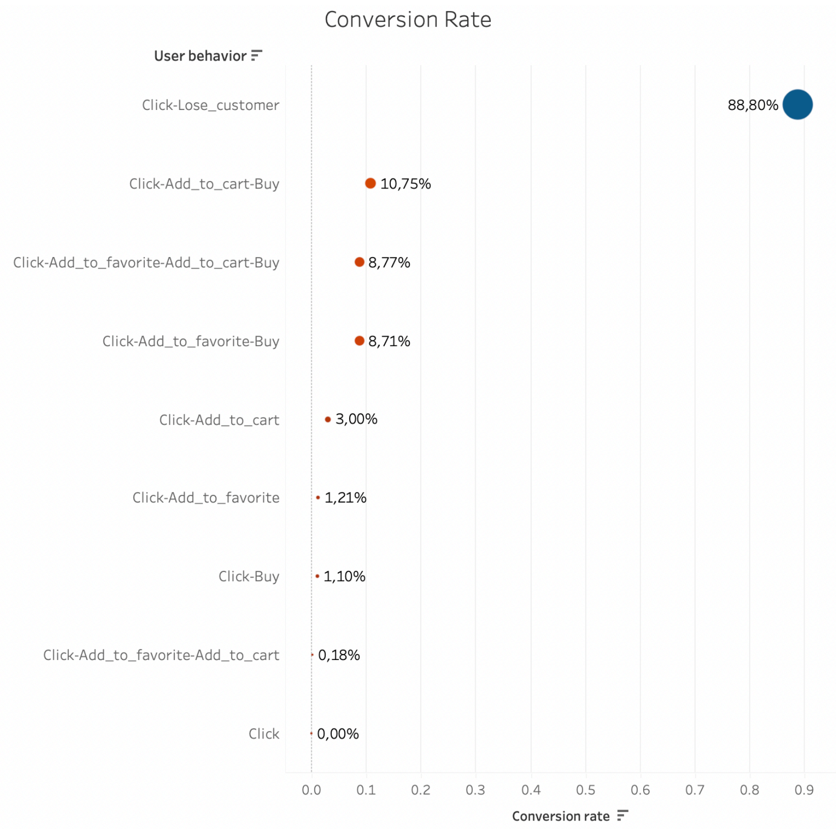
```sql
-- Amount of click-Add_to_favorite: 10752
SELECT SUM(Add_to_favorite) FROM BehaviorPath
WHERE Click > 0 AND Add_to_cart = 0 AND Add_to_favorite > 0;

-- Amount of click-Add_to_favorite-buy: 936
SELECT SUM(Buy) FROM BehaviorPath
WHERE Click > 0 AND Add_to_cart = 0 AND Add_to_favorite > 0 AND Buy > 0;
```

```sql
-- Amount of click-Add_to_favorite-Add_to_cart: 1631
SELECT SUM(Add_to_favorite) + SUM(Add_to_cart) FROM BehaviorPath
WHERE Click > 0 AND Add_to_cart > 0 AND Add_to_favorite > 0;

-- Amount of click-Add_to_favorite-Add_to_cart-buy: 143
SELECT SUM(Buy) FROM BehaviorPath
WHERE Click > 0 AND Add_to_cart > 0 AND Add_to_favorite > 0 AND Buy > 0;

-- Amount of click-lose_customer: 781384
SELECT SUM(Click) FROM BehaviorPath
WHERE Click > 0 AND Add_to_cart = 0 AND Add_to_favorite = 0 AND Buy = 0;
```

To make it better to understand, I made the following flow chart.



And then I used Excel and created the table called "Conversion Rate" and calculated the conversion rate of each behavior. Tableau is used to visualize the result.



From this result we can see that the conversion rate from "click" direct to "buy" is only 1.1%, while the conversion rate of "click – add_to_cart – buy" is 10.7%, and the conversion rate of

“click – add_to_cart – buy” is 8.7%. That means the conversion rate will increase if customers add the product to the cart or their favorite list.
According to this result, as could develop such marketing strategies: we need to improve the product interaction interface and marketing mechanism so that users are more willing to add the products to the cart or the favorite list. At the same time, we found that the conversion rate of “click – add_to_cart” is 3.1%, and the conversion rate of “click – add_to_cart – buy” is the highest at 10.7%, so compared to encourage the customer add the products into the favorite list, it is easier and more efficient to encourage the customer to add the products to the cart.

From the results above, we could also find another huge problem: that is, the conversion rate from clicks to the next level of operation is very low, indicating that users spend a lot of time browsing products, but few actually place orders. We all know that most of the users of Taobao and Tmall are women, who like to go shopping. They could go shopping for a whole afternoon or even a whole day. Why? This is because women go shopping not to buy what they need, but what they like, which is the so-called idle time. Therefore, for a shopping platform such as Taobao and Tmall, the recommendation function of products is particularly important. If they can recommend users to exactly what they like, they will buy those recommended products.

3.2 Low Conversion Rate Analysis

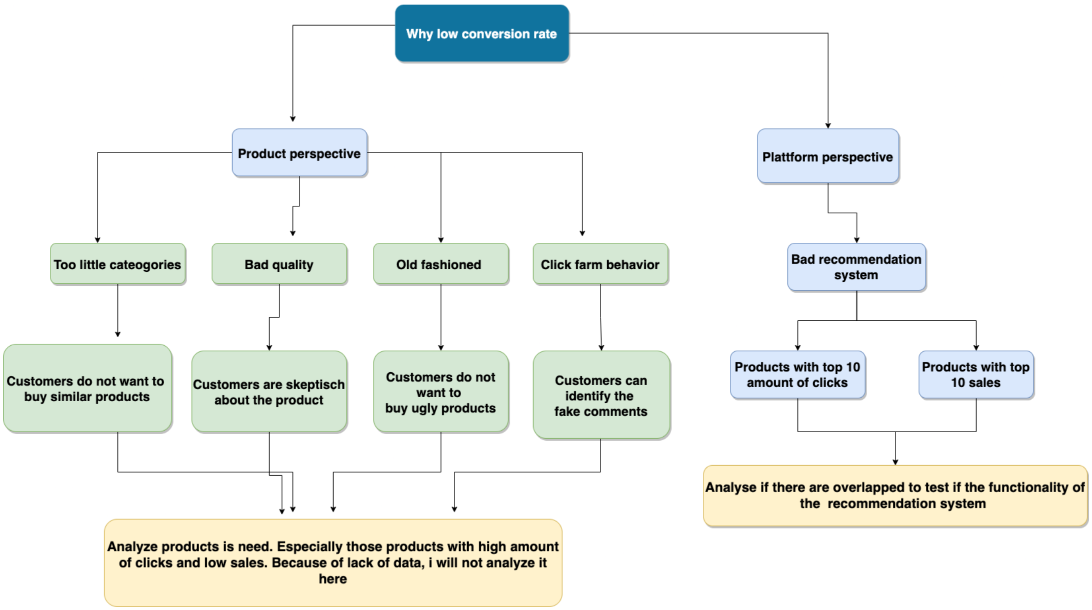
There are two types of problems, which could lead to low conversion rate: product problems and shopping platform problems.

Regarding to the product perspective, there might exist problems like low quality, little product categories, old fashioned products and click farming. To solve this kind of problem, we need to analyze those products with little sales. Because we don’t have such data, this aspect will not be analyzed.

Regarding to the platform perspective, the hypothesis is that the recommendation mechanism is unreasonable. The recommended products do not meet customers' needs, which results in a low conversion rate.

We can analyze whether there is a high degree of overlap between the high-viewed products and the high-purchased products. If yes, it means that the recommended products are liked by users and the hypothesis is not true. If not, the hypothesis is proved to be true.

The analysis above can be presented with the following chart:



The top ten products regarding to the amount of clicks.

```
CREATE TABLE Clicks
AS
SELECT CategoryID, COUNT(CategoryID) AS 'Amount_of_Clicks'
FROM UserBehavior_train
WHERE BehaviorType = 'pv'
GROUP BY CategoryID
ORDER BY Amount_of_Clicks DESC LIMIT 10;
```

The top ten products, which get the most sales.

```
CREATE TABLE Sales
AS
SELECT CategoryID, COUNT(CategoryID) AS 'Amount_of_Sales'
FROM UserBehavior_train
WHERE BehaviorType = 'buy'
GROUP BY CategoryID
ORDER BY Amount_of_Sales DESC LIMIT 10;
```

Let us combine the products above and have look if the products with high amount of clicks also have high amount of sales.

```
SELECT *
FROM Clicks
left JOIN Sales
ON Clicks.CategoryID = Sales.CategoryID
ORDER BY Amount_of_Clicks DESC;
```

CategoryID	Amount_of_Clicks	CategoryID(1)	Amount_of_Sales
4756105	47416	4756105	260
4145813	30362	4145813	325
2355072	30171	(NULL)	(NULL)
3607361	29091	(NULL)	(NULL)
982926	27780	982926	236
2520377	18606	(NULL)	(NULL)
4801426	18459	4801426	280
1320293	17411	1320293	170
2465336	15374	(NULL)	(NULL)
3002561	13650	3002561	168

From the table above, we can see that the products with top 10 sales do not definitely have top 10 amount of the clicks.

Next we break down to each product.

```
-- The top 10 products with the most clicks
CREATE TABLE ProductsClicks
AS
SELECT ProductID, COUNT(ProductID) AS 'Amount_of_Clicks'
FROM UserBehavior_train
WHERE BehaviorType = 'pv'
GROUP BY ProductID
ORDER BY Amount_of_Clicks DESC LIMIT 10;

-- The top 10 products with the most sales
CREATE TABLE ProductsSales
AS
SELECT ProductID, COUNT(ProductID) AS 'Amount_of_Sales'
FROM UserBehavior_train
WHERE BehaviorType = 'buy'
GROUP BY ProductID
ORDER BY Amount_of_Sales DESC LIMIT 10;
```

```sql
-- Calculate the amount of clicks with the top 10 products, which have the most sales
CREATE TABLE ClicksofMostSales
AS
SELECT * FROM
(SELECT ProductID,
COUNT(BehaviorType) AS 'Amount_of_Clicks'
FROM UserBehavior_train
WHERE BehaviorType = 'pv'
GROUP BY ProductID) AS A
WHERE ProductID IN('3122135', '3237415', '2124040', '2964774', '4401268', '1004046',
'1910706', '3991727', '3147410', '1595279')


SELECT *
FROM ProductsClicks AS P
LEFT JOIN SalesofMostClicks AS S ON P.ProductID = S.ProductID;

SELECT *
FROM ProductsSales AS P
LEFT JOIN ClicksofMostSales AS S ON P.ProductID = S.ProductID;
```

At the end, I used Excel to combine the two tables and I got the following results:

Sales of Most Clicks			Clicks of Most Sales		
ProductID	**Amount_of_Clicks**	**Amount_of_Sales**	**ProductID**	**Amount_of_Sales**	**Amount_of_Clicks**
2032668	191	2	3122135	17	16
1535294	169	7	3237415	11	42
3708121	193	1	2124040	11	4
812879	284	1	2964774	11	81
2338453	180	3	4401268	10	29
138964	221	1	1004046	8	13
4211339	165	0	1910706	8	12
3371523	161	0	3991727	8	18
3845720	216	0	3147410	7	31
2331370	185	0	1595279	7	33

On one hand, from the table "Sales of Most Clicks" we can see that there are almost no purchases corresponding to the top 10 products. That means, for the products which get high traffic from the platform can also get a lot of clicks, but the sales is pretty bad. Because e-commerce business is sales-oriented, so these products should not be given too much traffic support.

On the other hand, from the table "Clicks of Most Sales" we can see that, there are not so many clicks for the products with good sales and the products with a lot of clicks are not the same with the products with good sales.

Therefore, it can be concluded that the recommended product does not meet customers' needs, and the conversion rate is low because high amount of clicks do not lead to good sales.

We found that 2,735,466, 1,464,116, and 4,145,813 have relatively high purchase volume. On the basis of high demand, we consider Taobao to adjust the recommend mechanism and increase the traffic of these types of products to meet user needs.

Summarize:

1. Optimize the recommendation system to give more traffic to the products, which customers are willing to buy.
2. Reduce the turnover rate through better product recommendation, page interaction, loyalty points and other functions.
3. To increase additional purchases, sellers can improve the marketing strategy to encourage customers to purchase more, such as encourage customers to add products to shopping cart and to contact customer service to receive coupons.

3.3 User Hierarchy Analysis Based on FRM Model

RFM is the abbreviation of 3 indicators, the most recent consumption interval (Recency), consumption frequency (Frequency) and consumption amount (Monetary). Users are classified by these 3 indicators.

Next, I used SQL statements to calculate the last consumption time interval and consumption frequency of each user. The consumption amount is not considered for the analysis due to the lack of data.

Calculate the last consumption interval (R)

R = data recorded time – user's latest consumption time

F = sum of user purchases

```
SELECT UserID, DATEDIFF('2017-12-3',max(Datee))+1 AS R,
COUNT(BehaviorType) AS F
FROM UserBehavior_train
WHERE BehaviorType = 'buy'
GROUP BY UserID;
```

Give scores according to the recency and frequency.

Score	R	F
1	8 - 9 days	1 - 6 times
2	5 - 7 days	7 - 12 times
3	3 - 4 days	13 - 18 times
4	Within 2 days	More than 19 times

```sql
SELECT *,
(CASE WHEN R <= 2 THEN 4
WHEN R BETWEEN 3 AND 4 THEN 3
WHEN R BETWEEN 5 AND 7 THEN 2
WHEN R BETWEEN 8 AND 9 THEN 1 END) AS Rscore,
(CASE WHEN F BETWEEN 1 AND 6 THEN 1
WHEN F BETWEEN 7 AND 12 THEN 2
WHEN F BETWEEN 13 AND 18 THEN 3
WHEN F >= 19 THEN 4 END) AS Fscore
FROM
(SELECT UserID, DATEDIFF('2017-12-3',max(Datee))+1 AS R,
COUNT(BehaviorType) AS F
FROM UserBehavior_train
WHERE BehaviorType = 'buy'
GROUP BY UserID) AS M;
```

UserID	R	F	Rscore	Fscore
100	6	8	2	2
1000001	1	1	4	1
1000011	9	2	1	1
100002	4	1	3	1
1000027	1	2	4	1
1000028	1	4	4	1
1000037	3	5	3	1
1000054	2	4	4	1
1000060	3	1	3	1
1000061	1	3	4	1
1000070	6	1	2	1
1000084	1	7	4	2

Calculate the average value of the two scores.

```sql
SELECT AVG(Rscore), AVG(Fscore)
FROM Score;
```

AVG(Rscore)	AVG(Fscore)
3.0366	1.1028

Customer segmentation

Customer segmentation	Recency	Frequency	Monetary
Important value customers	High	High	High
Important development customers	High	Low	High
Important keep customers	Low	High	High
Important save customers	Low	Low	High
Normal value customers	High	High	High
Normal development customers	High	Low	Low
Normal keep customers	Low	High	Low
Normal save customers	Low	High	Low

```sql
SELECT CustomerSegmentation, COUNT(UserID) AS Amount_of_People
FROM
(SELECT UserID,
(CASE WHEN Rscore > '3.0366' AND Fscore > '1.1028' THEN 'Important value customers'
WHEN Rscore > '3.0366' AND Fscore < '1.1028' THEN 'Important development customers'
WHEN Rscore < '3.0366' AND Fscore > '1.1028' THEN 'Important keep customers'
WHEN Rscore < '3.0366' AND Fscore < '1.1028' THEN 'Important save customers'
ELSE 0 END) AS 'CustomerSegmentation'
FROM Score) AS N
GROUP BY CustomerSegmentation
ORDER BY Amount_of_People DESC;
```

CustomerSegmentation	Amount_of_People
Important save customers	3545
Important development customers	2572
Important value customers	394
Important keep customers	155

It can be found that most users are important development users and important save users.

There is a large proportion of important development users. Their consumption frequency is low, but the recent consumption time is short, so we must find a way to increase their consumption frequency.

For important save users, the recency is high and the consumption frequency is low. This kind of users can be lost easily. It is necessary to actively contact users and find out the problems. Then those problems can be analyzed and solved.

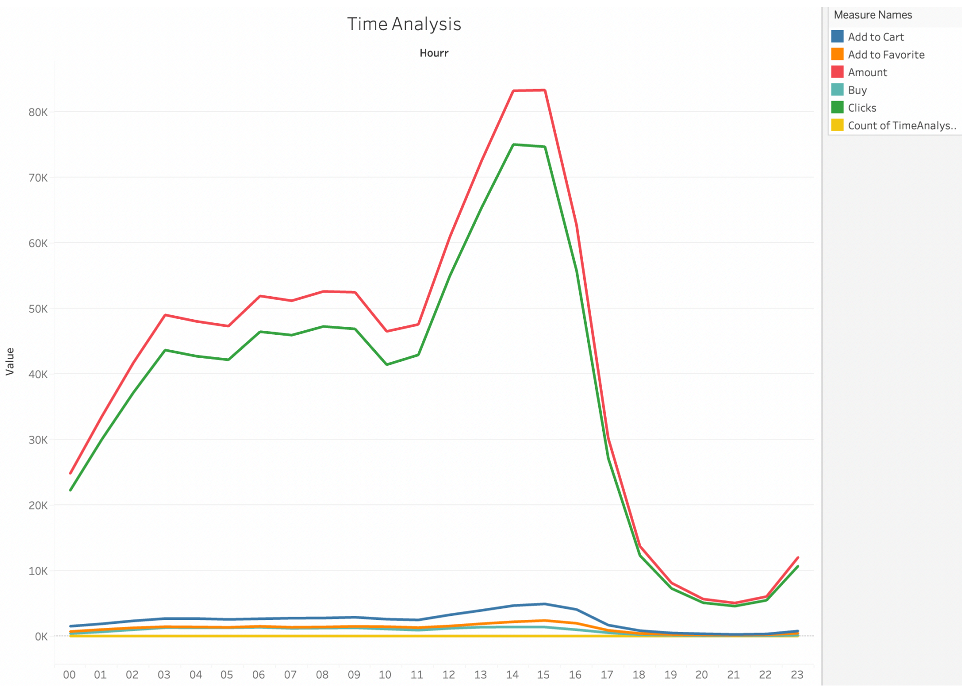
For important value users, the consumption frequency is high and the recency is low, so it is necessary to provide them with VIP services.

For important keep users, the recency is high and the consumption frequency is high. Such users are loyal customers who haven't shop for a while. We can send emails, APP reminder or SMS reminder during promotional activities, so that the repurchase rate can be improved.

3.4 Time Analysis

Let us have a look, when will customers shop more or click more in a day.

```sql
ALTER TABLE UserBehavior_train
ADD Hourr varchar(10);

UPDATE UserBehavior_train
SET Hourr = FROM_UNIXTIME(Timestamps,'%H');
```



We can notice that, during a day, customers are the most active during 14 – 16 PM.

So we can consider doing some promotions and product recommendations during 14-16 PM to increase conversion rates.

4. Conclusion and Suggestion

4.1 Conclusion

1. Products with high traffic are not products with high purchase volume. Low purchase volume of products with high traffic results in a low overall traffic conversion rate. That means, the recommended system from the platform is not sales-oriented.
2. It can be seen from the user behavior path that the conversion rate of direct purchases after clicks is low, while the conversion rate of purchases is high after customers add products to the cart or their favorite list. So it is necessary to guide customers to actively add products to the cart or the favorite list. After comparing the conversion rate, it is found that the conversion rate of purchases is higher after customers adding products to the shopping cart than adding products to the favorite list.
3. Most users are important development users and important save users. These two kinds of customers account for 91% of the all the customers.

4.2 Suggestion

1. It is recommended that the **algorithm department show those products with TOP 10 sales to customers first,** such as 2735466, 1464116, 4145813, etc. If the products with high clicks are new products or products that have been promoted recently, we can consider to sell the products with TOP 10 sales together with those kind of new products according to the categories of these products, so that the conversion rate and joint rate can be improved.
2. It is necessary to actively guide customers to add products to shopping carts or their favorite lists. Interface design department could consider how to interact with customers

so that the amount of clicks can be improved. Operation department could set up some guidance strategies, for example, customers can contact customer service for coupons or a small gift if they add products to the cart.

3. For important development users, their consumption frequency is low, but the recency is low, so we need to find ways to increase his consumption frequency. For example, member rewards or SMS reminder discounts can be used to improve the consumption frequency.
4. For important save users, the recency is high and the consumption frequency is low. Such users are in danger of being lost. It is recommended to issue paid questionnaires through APP, SMS or email, etc. We need to actively contact users, investigate clearly what went wrong, and formulate corresponding recovery strategies.
5. We can consider doing some promotions and product recommendations during 14-16 PM to increase conversion rates, because customers are most active in this time period.