Data Set <Mum Baby Goods>
By Jia Ziyi, Zhang Yakai, Tang Huimin

AS&DA Project Report

1. Introduction

➢ Data set description

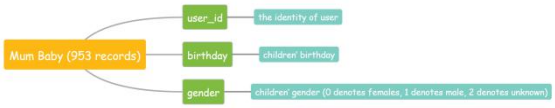


Figure 1.1

Figure 1.2

➢ Our intention

- ✓ Understand the overall situation of the maternal and infant products market;
- ✓ Understand the sales situation of various products;
- ✓ Through the analysis of the user's purchase preference, we can predict the user's purchase behavior;
- ✓ Provide effective suggestions for the sales business of e-commerce platform.

➢ Questions

User's purchase preference: Taking the age and gender of infants as the characteristics of users, the purchase behavior preference of users is analyzed.

Commodities characteristics: Analyze the influence of commodity category on commodity sales according to commodity purchase quantity. Guess the product attributes according to the purchase situation.

Transaction characteristics: Based on historical transaction, the relationship between overall market sales volume and time is analyzed.

2. Data processing

➢ Summary of two data set

Descriptive statistical analysis of users

Gender: of the 953 users in the <mum_baby> data set, 489 are girls, 438 are boys, and 26 have unknown gender. It should be noted that some users have unknown gender.

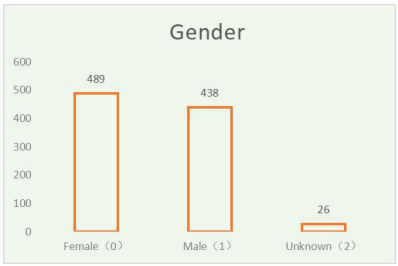


Figure 2.1

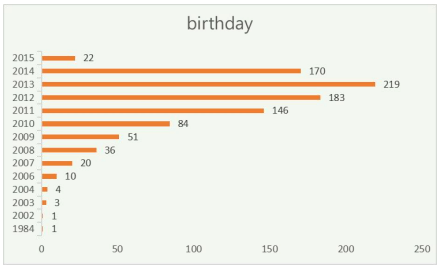


Figure 2.2

Birthday: most infants were born in 2010~2014.

Descriptive statistical analysis of trade history

Root category (Cat1): 6 types

Sub category (Cat_id): 662 types

Transaction time:20120702~20150205

buy_mount: $sum = 76250, mean = 2.5, var = 4094.3, sd = 63.986$. The standard deviation is too large. We should consider abnormal data in the following data

process.

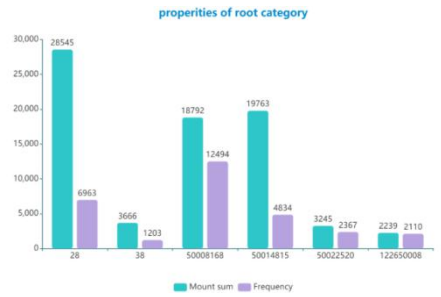


Figure 2.3

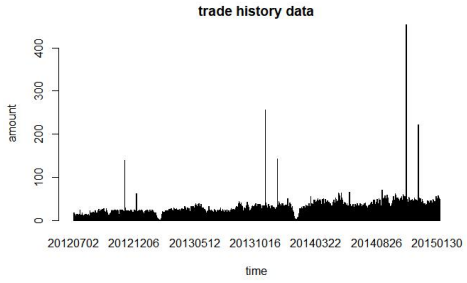


Figure 2.4

We find that there is significantly correlation between auction_id and day, this might because auction_id is based on auction time.

If we want to analysis auction_id and the property data, we need a data dictionary to identify the meaning of each. Therefore, in the following discussion, we completely drop auction_id and property data.

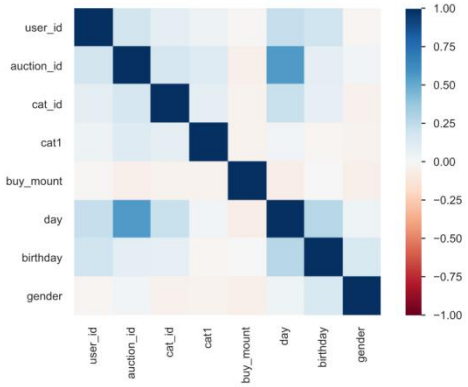


Figure 2.5

➢ Data cleaning

In order to analyze the user transaction data with known information, we used a LEFT JOIN to merge < mum_baby> and < mum_baby_trade_history > into <mydata3>.

Gender outliers: In the analysis of gender influence, the unknown gender (gender = 2) was extracted separately as variable analysis.

Processing of time data: In the age analysis, we need to make a difference between the birthday and the transaction time to find the baby's age at the time of purchase. The age was divided into six groups: before pregnancy (< - 0.8 years old); fetal period (- 0.8, 0); infant period (0, 1); (1, 2); (2, 3); (3, inf). We can see that there is an outlier data with the date of birth of 1984 and the age of 28 years old, which should be eliminated.

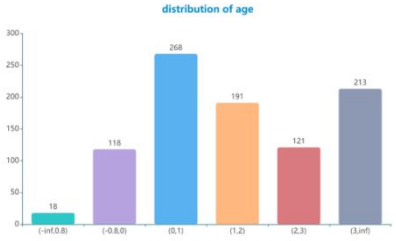


Figure 2.6

Data Set <Mum Baby Goods>
By Jia Ziyi, Zhang Yakai, Tang Huimin

Deal with the quantity of goods in a single transaction: After the first step of data overview, we found that the variance is huge and the purchase volume of some transaction records is significantly larger than that of the group. The reason for the huge number of commodities in a single transaction may be that the user is a commodity dealer, or the transaction is a malicious order from the store. In any case, the outlier data will affect our research on user preferences and behavior, so it needs to be special treatment. Our processing method is to take the maximum amount of single transaction data of known user information 160 as the upper limit, and directly delete the data with the number of single transaction products more than 160 in dataset <mom_baby_trade_history>. At this time, data set <mom_baby_trade_history> is updated and recorded as data set <2a>. Now, $sum = 49410, mean = 1.65, var = 24.255, sd = 4.9249$. Compared with the data before data cleaning, standard derivation is significantly smaller.

- Data analysis

User's purchase preference:

Because there are too many kinds of sub-directories (cat_id), in this step, we focus on analyzing the purchase behavior preference under the root directory products (cat1).

A study on the preferences of different age users for root products.

H_0: no obvious difference among different age groups.

H_1: have different preferences.

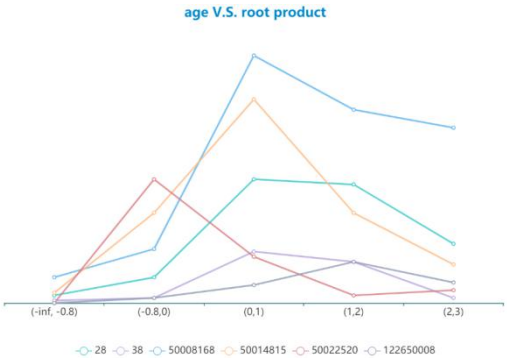


Figure 2.7

$$\chi^2 = 297.1564 \ \& \ \text{df} = (6-1)\times(6-1) = 25 \ \& \ \text{p} - \text{value} = 1.034712\text{e} - 48$$

We reject H_0, as p-value is too small.

A study on the preferences of different gender users for root category products.

H_0: Different genders have the same preference for root category products.

H_1: Different genders have different preference for root category products.

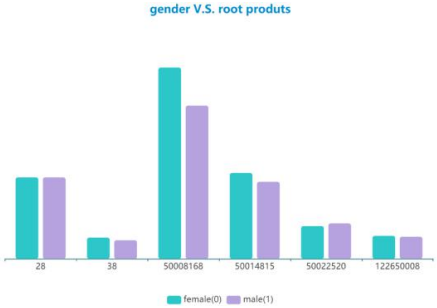


Figure 2.8

Pearson's Chi-squared test: $X - squared = 2.5061, df = 5, p - value = 0.7756$

Therefore, we don't reject H_0, which means that whether female or male do not influence sales of root products.

H_0: Users with unknown gender (2 in the table) and users with known gender have the same preference for root category products;

H_1: Users with unknown gender and users with known gender have different preference for root category products.

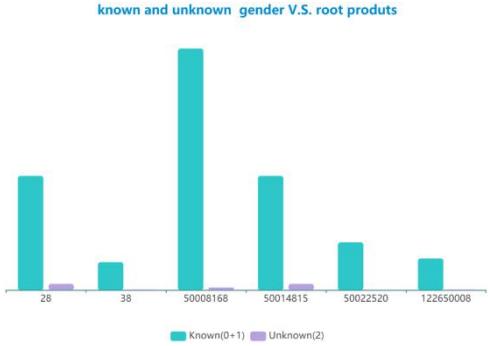


Figure 2.9

$df = 5, p - value = 1.08 \times 10^{-107}$

In this set of data, we found that the number of users with Unknown Gender is too small and the transaction data is not enough. Therefore, we have met the problem when calculating chi square statistics. The final conclusion is not reliable.

H_0: The behavior of users with known user information is consistent with that of users with unknown user information.

H_1: The behavior of users with known user information is inconsistent with that of users with unknown user information.

	28	38	50008168	50014815	50022520	122650008
Whole user	6941	1203	12489	4830	2366	2110
Known user	194	46	393	194	77	52

$df = 5, P - value = 0.00271984$

In the same gender user group, the proportion of sales volume of large categories (cat1) of goods:

Sum of mount	28	38	50008168	50014815	50022520	122650008
Gender=0(female)	167	103	247	352	38	28
Gender=1(male)	119	44	200	117	54	25
Gender=2(unknown)	11	1	4	27	0	1

We can find that for root category 38 and 50014815, The purchase amount of female infant families was significantly larger than that of male infant families.

Commodities characteristics:

After processing the outlier data, the sales volume under the root directory products is shown in the figure below:

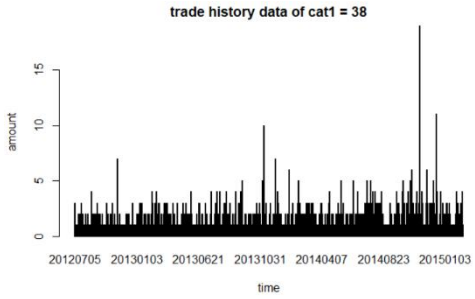
Cat1	28	38	50008168	50014815	50022520	122650008

Data Set <Mum Baby Goods>
By Jia Ziyi, Zhang Yakai, Tang Huimin

Frequency	6963	1203	12494	4834	2367	2110
Mount sum	16321	3666	14944	9195	3045	2239
AVG (buy_mount)	2.35	3.04	1.19	1.90	1.28	1.06

Time analysis of popular categories 50008168 and 38:

38: 50008168:



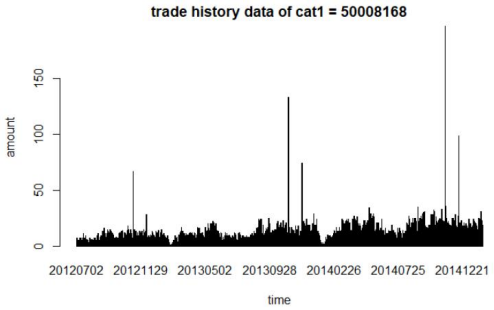


Figure 2.10 cat1 = 38 figure 2.11 cat1 = 50008168

The single purchase volume of 38 categories is large, which may be high consumables. The number of transactions of 38 categories is small, but the per capita purchase volume is large, and the number of subcategories under this major category is small, which means that users have less choice and strong demand, so subcategories can be added to provide users with more choices.

The sales volume and per capita purchase volume of 122650008 categories are both small, so e-commerce should reduce inventory.

For the popular Root Category 50014815, we should increase the promotion means to keep the sales volume.

Use SQL to analyze the purchase of known users:

transaction number	Once	Twice	Four times
Sum of number	29887	24	1

Among the 29887 transactions data, only 24 users have two transaction histories, one user has four transaction histories, and the remaining 29835 transactions are the only transaction data of other users. This shows that the re purchase rate of the shopping platform is very low, there are basically no return customers, and the sales volume in the transaction is all generated by new customers.

From the user data of four buybacks, we can see that the root category (cat1) of the commodity purchased by the user is 28, and the sub category (cat_id) is 50015146.

user_id	auction_id	cat_id	cat 1	property	buy_mount	day
814316568	17742071206	50015146	28		1	20130314
814316568	17741967845	50015146	28		1	20131028
814316568	20778228751	50015146	28		1	20130313
814316568	14509614014	50015146	28	11666049:62519847	1	20130507

Data Set <Mum Baby Goods>
By Jia Ziyi, Zhang Yakai, Tang Huimin

It is speculated that the product with Root Category cat1 = 28 may be consumable, such as milk powder. The user is very satisfied with the product, so he purchases the product again after the product is consumed, which is also consistent with the purchase interval.

The figure below shows the proportion of times of purchasing different root categories of goods among 24 users who buy back twice. It can be seen from the figure that half of all users who have repurchase behavior to buy 50008168.

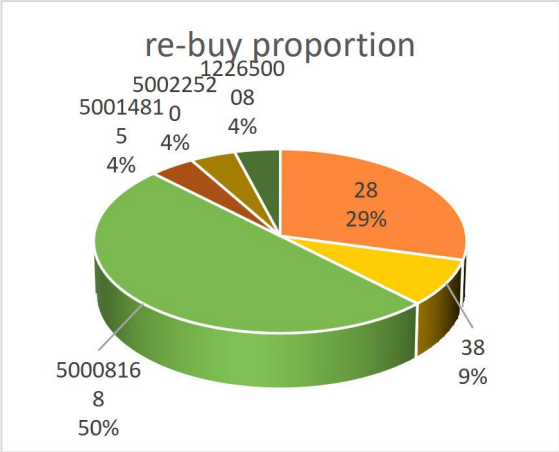


Figure 2.12

Transaction characteristics:

Figure 2.13

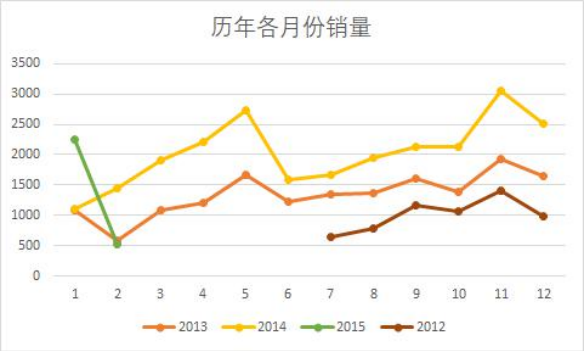


Figure 2.14

There are only six months of data in 2012 and two months of data in 2015, so it has no reference value. Only from the data of 2013 and 2014, we can see that both the number of sales and the sales of goods are showing a rapid growth trend.

The specific month analysis: The sales volume of January and February in each year decreased sharply, while the sales volume in May and November increased significantly. In February 2015, only the data of the first five days are available, so no analysis is made here.

Possible causes of increased sales in May and November hypothesis:

1. Merchants launch new products;
2. New sales channels have been added;
3. Holidays and vacations;
4. Shops have promotional activities.

The first two assumptions require specific contact with the shop owner to reach a conclusion, so do not study them here.

Data Set <Mum Baby Goods>
By Jia Ziyi, Zhang Yakai, Tang Huimin

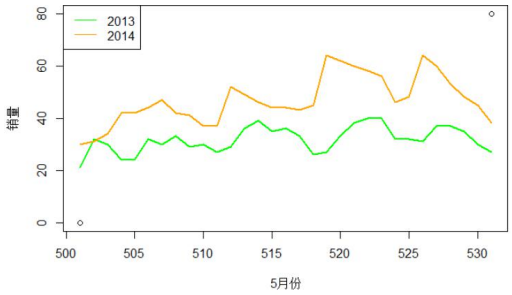


Figure 2.15

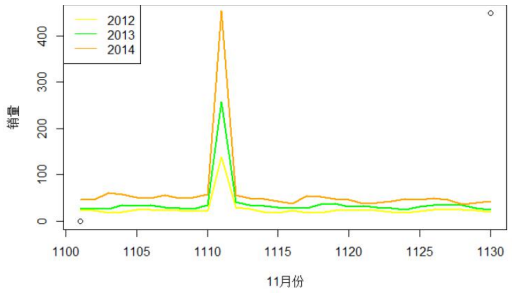


Figure 2.16

1) As can be seen from Figure 2.15, the sales volume of time nodes such as May 1, May 12 and May 21 in 2013 are relatively large, which are Labor Day, Mother's Day and online Valentine's Day in 2013 respectively; On May 11 and May 21, 2014, the sales volume of time nodes increased sharply, which were Mother's Day and online Valentine's Day. The time of the festival coincides with the time of sales increase, so the hypothesis is true.

2) From Figure 2.16, it can be seen that the sales volume of the three years on November 11 has increased dramatically. It is speculated that the promotion of "double eleven" has made a large number of users choose to buy and stock goods.

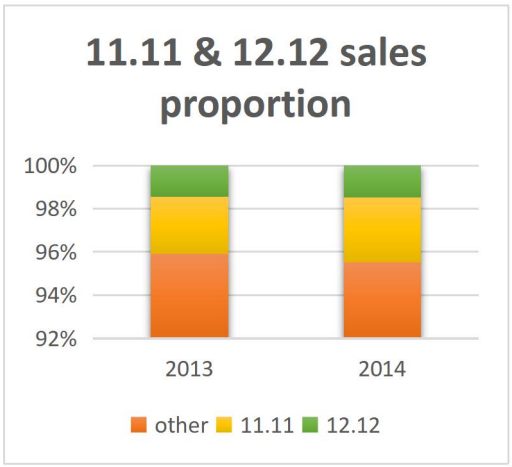


Figure 2.17

The reason for the sharp drop in sales in January and February may be the outage of express delivery during the Spring Festival. Since February 2015 has only data for the first five days, it will not be considered here.

As can be seen from figures 2.18 and 2.19, the sales volume from January 28 to February 5 in 2014 and from February 5 to February 14 in 2013 all had a long period of low ebb. Considering that the Spring Festival in 2014 is on January 30 and the Spring Festival in 2013 is on February 10, the low sales volume coincides with the time of Spring Festival, so the assumption is reasonable.

Data Set <Mum Baby Goods>
By Jia Ziyi, Zhang Yakai, Tang Huimin

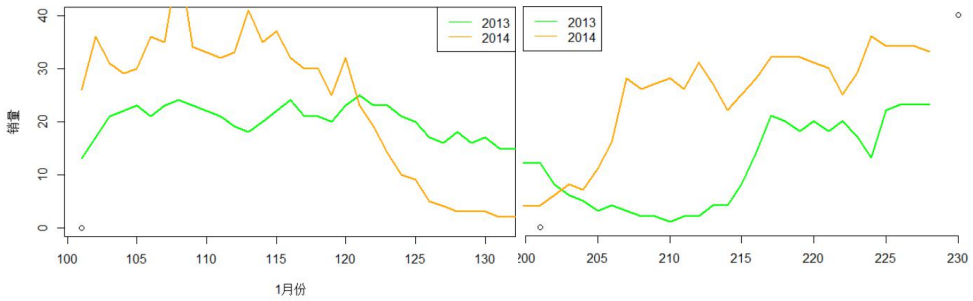


Figure 2.18 figure 2.19

3. Conclusions

✓ In the cleaned data, users with large single sales volume have a greater impact on the overall maternal and child products market. If it is malicious order behavior, it needs to be investigated and dealt with by the e-commerce platform; If it is a normal transaction, e-commerce should do a good job of corresponding services for such large customers.

✓ The real name registration of user information is of great significance to e-commerce. From this data set, the proportion of users with known information is very small.

✓ This data set shows that the repurchase rate of all maternal and child products is at a low level, and the user has only one transaction record, which means that the user's stickiness is insufficient. It is urgent to find out the reasons for the low repurchase rate. In the face of new users, we also need to pay attention to the means of publicity, improve single consumption.

✓ For all kinds of commodities, the consumption of female infant families is generally higher than that of male infant families.

✓ The highest demand for mum baby goods is between 0 and 2 years old. And when the shops have promotional activities like double-eleven, the maternal and child products also face a peak. In Spring Festival, sales are down. Therefore, shops should pay special attention to special days.

✓ For a few special categories (e. g. cat1=38), there still be a huge market.

4. Limitations

One of the problems reflected in the whole data analysis is that the official data collection is only a simple record.

As all the data is digital and we don't have a data dictionary to index, we can't understand completely what they represent. After trying quite a few methods, we had to drop these two. Our processing of "cat_id" and "property" is difficult, even unable to start.

At the beginning, we intended to use our data to draw a portrait for users, using the methods of regression to predict the behaviors. But then after dropping two columns, we found that our independent variables are small. There is no obvious regression relationship between variable. When we analyzed whole trade history, it was sad to discover that almost all the users just shopped once in the data set, so any prediction on user's gender or age is useless (just based on one known behavior, randomness and unreliable!)

We changed our analysis directions many times in the process and tried quite a few analysis methods. Limited by our ability, we tried our best to find information behind the huge data.

Data Set <Mum Baby Goods>
By Jia Ziyi, Zhang Yakai, Tang Huimin

Appendix

Working list

Project parts	Jia Ziyi	Zhang YaKai	Tang Huimin
Statistical description (summary of two dataset)	50%	30%	20%
Processing data (data cleaning)	50%	40%	10%
Data analysis	40%	40%	20%
Visualizing data	25%	35%	40%
Report writing	50%	40%	10%
Poster showing	30%	5%	65%