JD User Consumer Behavior Analysis

Background: JingDong has hundreds of millions of loyal users and accumulated a large amount of real data while keeping the rapid development. To find out the rules from the historical data, solve the customer's actual problems and enhance the customer's shopping experience effectively is the key for big data to act in precision marketing, and also the core technology for all e-commerce platforms to do intelligent upgrade. This project is mainly about the users' purchase behavior analysis and prediction, to attract more customers and improve company's business by efficient precious sales.

Define the problem: Find out the rules from the historical data, solve the customer's actual problems and enhance the customer's shopping experience effectively, provide users with the most suitable products and services, and contribute the most benefits to the company with the least cost.

Detail the problem: 1. Analyze users' purchase trend: Study the relationship between the users' action type as 'order' and time

2. Factor analysis: Besides seasonal changes, marketing strategies, etc., focus on ①demographic variables (age range, gender), ②merchant variables (number of fans (segmentation), store rating (segmentation)), ③account variable (member level , city level), ④product variables (base categories of shops, detailed categories of products, brands and listing dates)

3. User purchasing propensity prediction: Cluster analysis method can divide customer into several different groups first, and then find high-value user groups (users who are most likely to use coupons to purchase behavior) by classification analysis, also need to establish a logistic regression model (score by 0-1). For different customer groups develop precise marketing solutions and advertising channels, and use regression analysis to decide the most effective advertising marketing activities, at last inter-group difference analysis to find the most popular marketing tools.

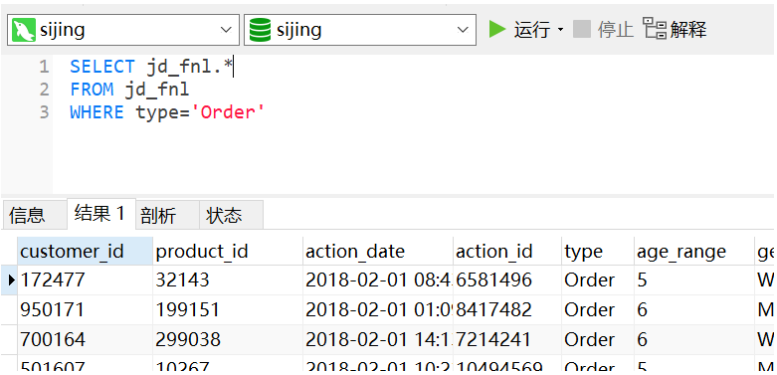
General Steps: 1. SQL to extract and evaluate data (integrity and timeliness consistency), data diagnosis and cleaning (to exclude abnormal value, here mainly involves missing data, exceptions), and one is data manipulation (that is, to convert data into the shape which can facilitate subsequent analysis, such as converting data into business-meaning groupings), because the project data contains only one table, no need to joint, then integration is required if needed.

2. Tableau to visualize the data, analyze the major factors from different perspectives, explain and summarize the main factors, create dashboards and make some suggestions.

3. List and describe a range of analytical methods and details used for predicting user purchase propensity, such as how to conduct hypothesis testing.

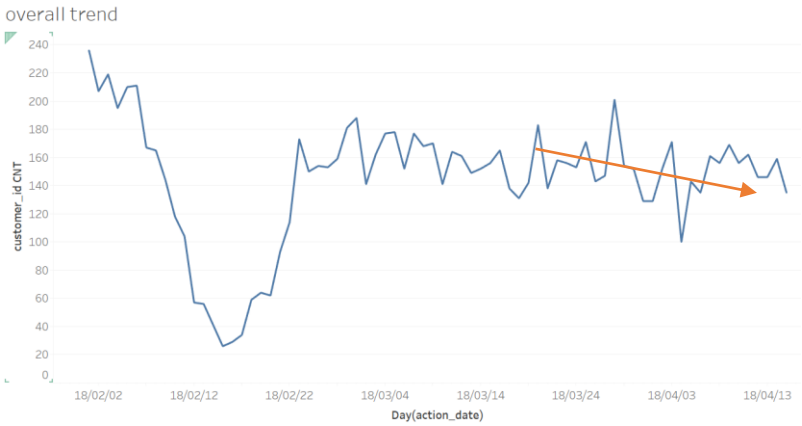
4. Implement the marketing plan and set a standard to verify the success of the program.

1) For excel file checking, firstly, observe the table information, and find that there is no metric information such as sales amount/volume, so consider the purchase number of each product in each unit of time (like day/month/year) as a reference target (so no need to DISTINCT); secondly, in this data appears 0, -1 and NULL values, query the data dictionary or check with the relevant person: whether these values are truly abnormal and consider whether need to remove these outliers or do other actions. For example, under the existing conditions, observing that all the stores which registration date is NULL belong to electronics, and the corresponding number of fans and shop scores are 0, but not all electronic stores. We can make a hypothesis that some electronic sellers have made mistakes in register info or missing data collection, but there is little impact to sales trends analysis, so it can be remained; but these outliers need to be excluded in Tableau when we check the factor of shop score and registration date.
Check what kind of data info we need, extract the useful data by SQL, so we only need the users whose action type is order.



customer_id	product_id	action_date	action_id	type	age_range	ge
172477	32143	2018-02-01 08:4	6581496	Order	5	W
950171	199151	2018-02-01 01:0	8417482	Order	6	M
700164	299038	2018-02-01 14:1	7214241	Order	6	W
501607	10267	2018-02-01 10:2	10494569	Order	5	M

The general sales status during the Feb/Mar/Apr 2018

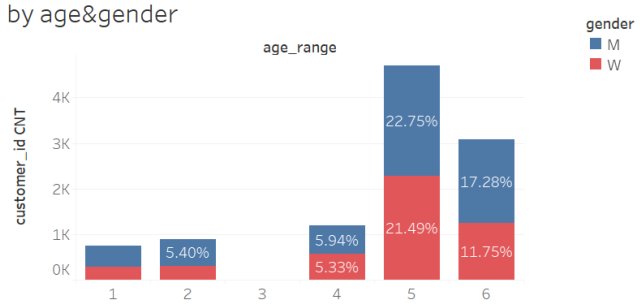


It can be found that 2018.2.15. is Chinese New Year's Eve, so analysis can be made according to before and after new year. It can be seen that before 2/7 the order volume is obviously higher than the average level after the New Year. Since 2/7, the purchase speed has dropped rapidly, and the purchase volume during the New Year period was the lowest during the three months, but after 2/20 (after new year), the order volume began to rise rapidly until it was stable. The reason may be that a large amount of purchases of new year's goods were needed before the festival, so the order quantity is higher than usual. During the Chinese New Year, some shops were on vacation and the express delivery also was out of service, resulting in low purchase volume. After the new year, people return back to work as normal, so the purchase status is also back to baseline. The average level was flat, but sales trend down after mid-March, so the number of orders lower in early April compared to March.

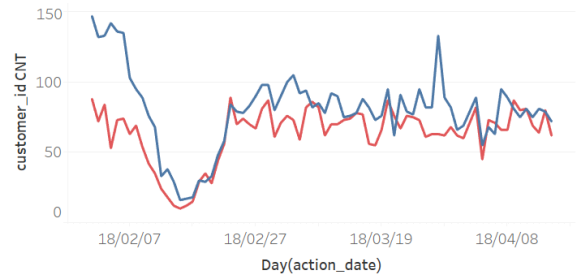
2) Effect factors analysis

①demographic variables

by age&gender



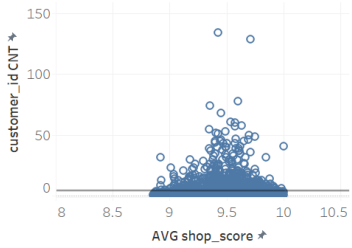
user trend



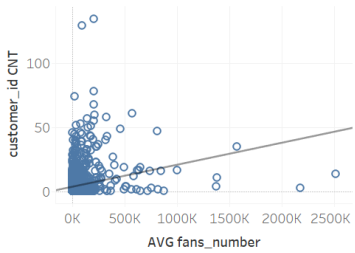
Draw stacked bar find that the customer with the age_range of 5 has the most order volume This may due to the higher customer income in the fifth stage, so the purchasing power is stronger, and the male customer purchase ratio is slightly more than the female; according to the time trend chart some marketing methods can be used to increase the purchase amount of female customers before Chinese new year in the future. The purchasing level of females is relatively stable after the new year, while the male customers are slightly lower at the end of March and early April, so if possible need to consider increasing the marketing plan for male customers.

②merchant variables

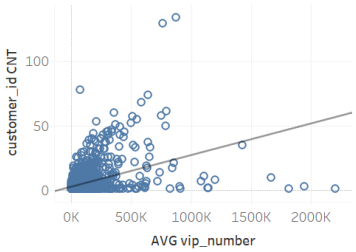
by shop_score



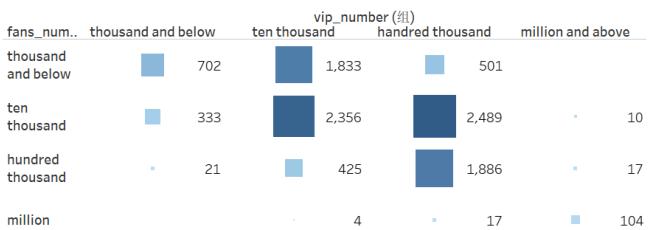
by fans_number



by vip_number

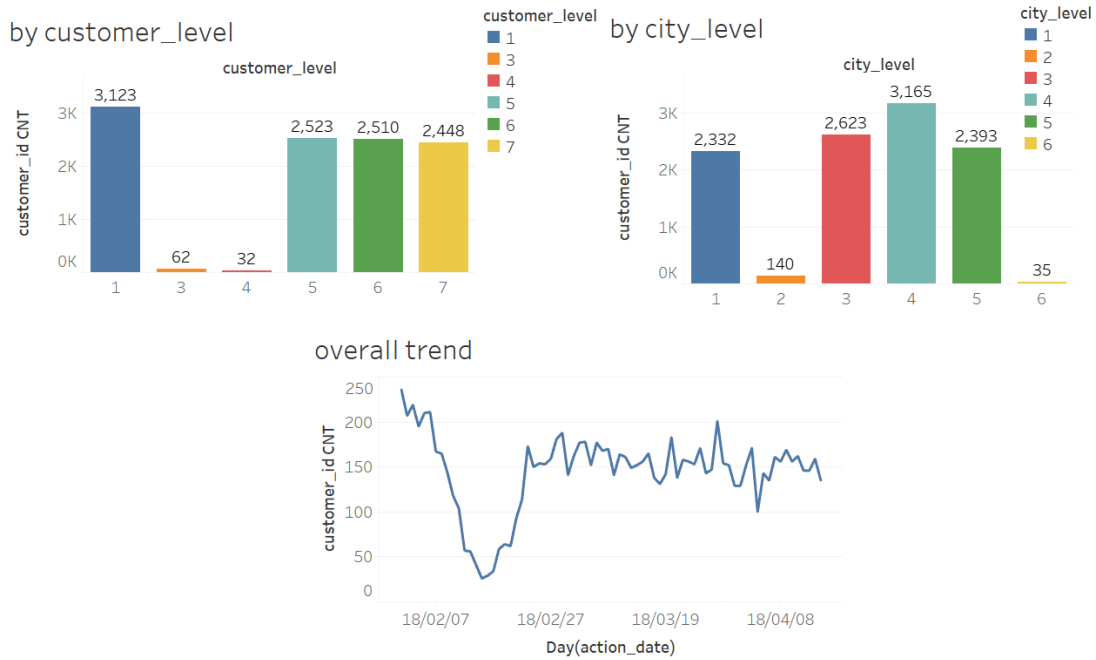


by vip&fans

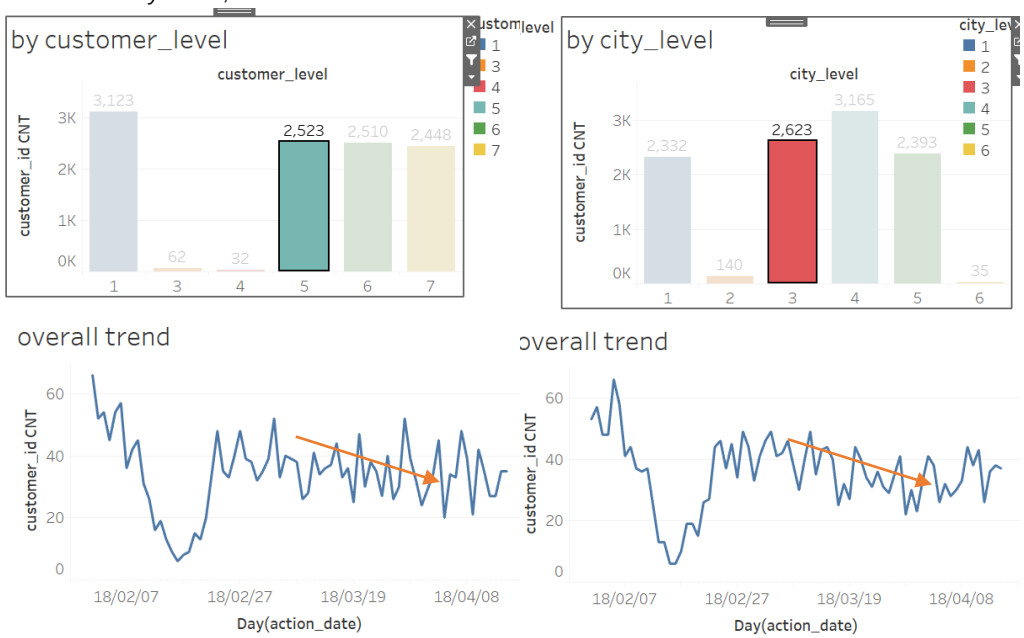


According to the scatter plots, the analysis of the shop scores, number of fans/members on sales shows that: 1. There is no obvious correlation between the score and order number. The distribution is in the range of 9.2-9.7, so the shops below this score should try to make corresponding actions to raise the score to this range; 2. In general, the more fans/members makes better sales, but the actual situation is not strictly in accordance with the trend. By setting the linkage filter, can make the proper scheme to increase the number of fans/members according to each shop with different levels.

③account variable

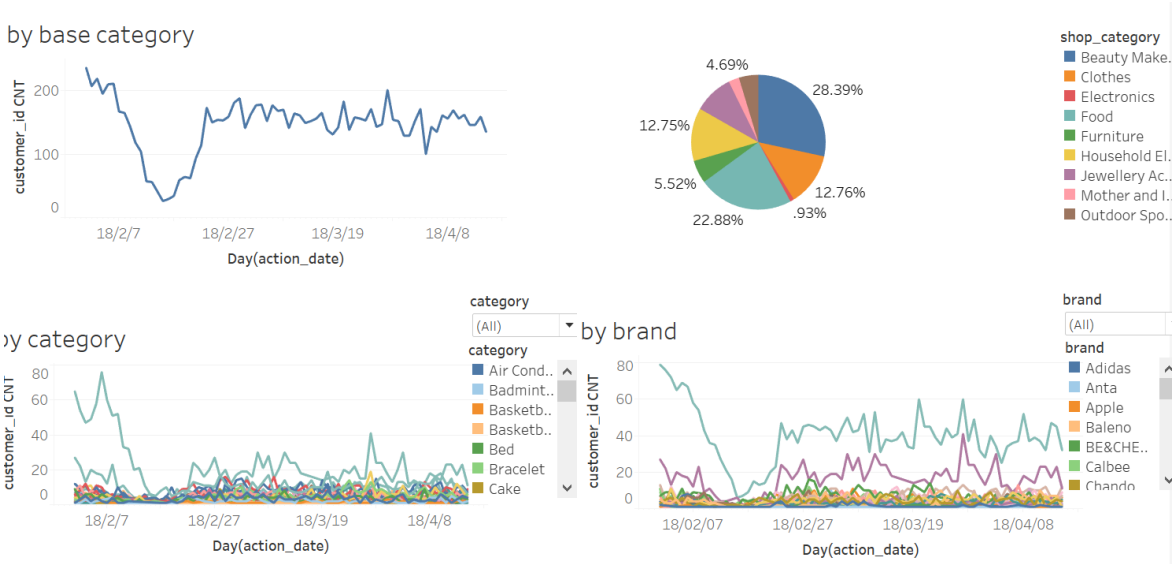


For membership level, the customer with the level 1 has the largest purchase amount, and the difference among the 5/6/7 level is quite small. From the city level, the level 4 city is the highest, and the remaining 1/3/5 is slightly lower with little difference, there is no obvious relationship between the sales and the member&city level;

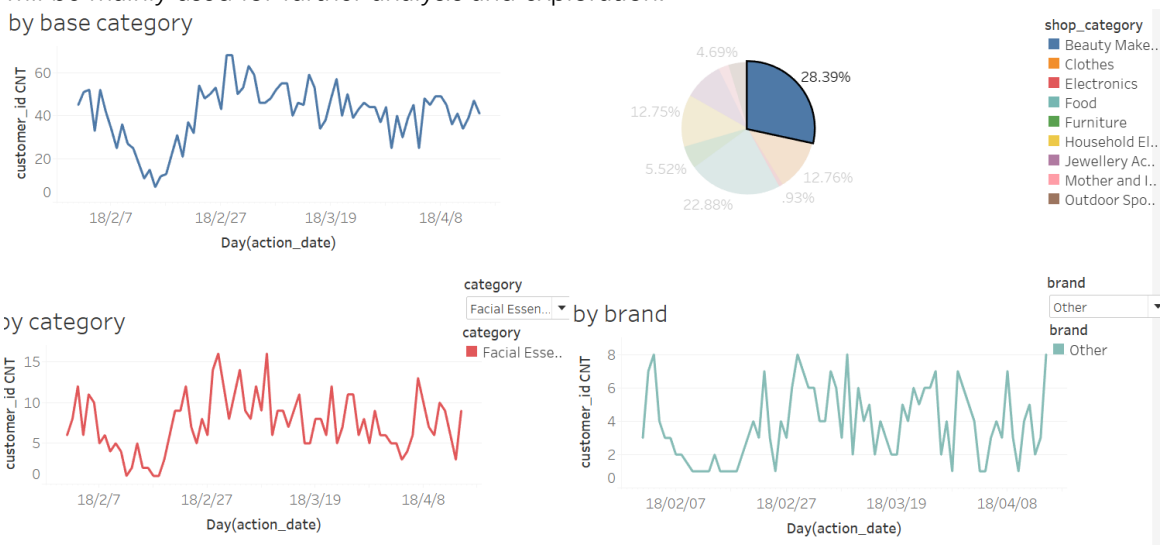


By setting up the linkage filter, it was found that users with the rating of 5 and the city rating of 3 sales volume began to reduce at the end of March, and can carry out special marketing programs or increase marketing advertising for this group.

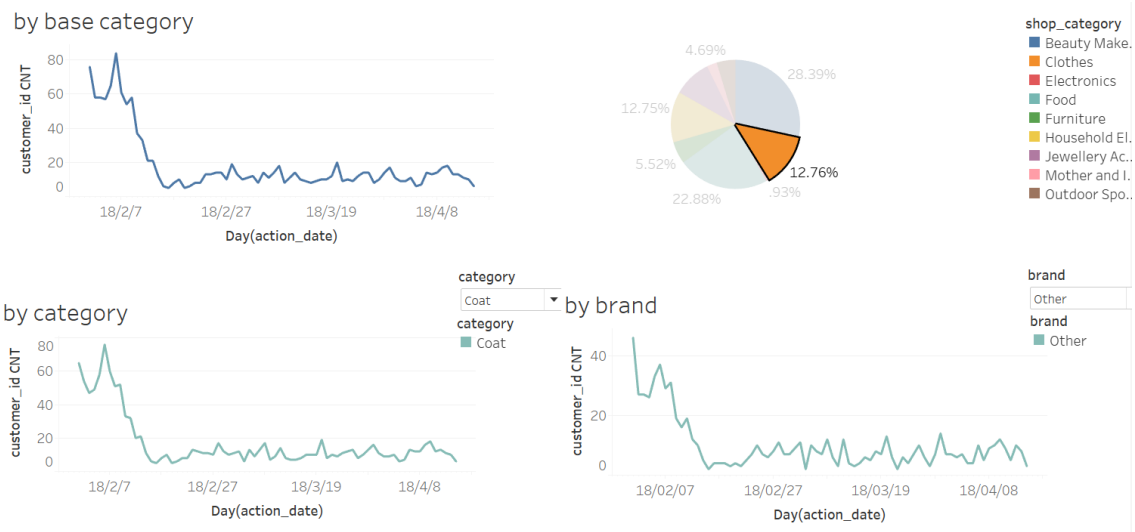
④product variables



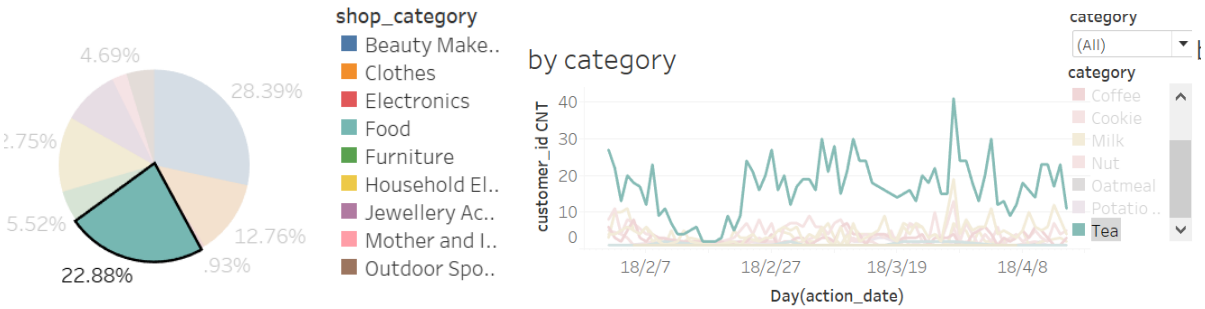
Through the dashboard and linkage filter, we can see from pie chart product categories with the highest sales ratio are Beauty Makeup 28.39% and Food 22.88%. also the three categories (base/shop_category, category and brand) of products can be chained together to analyze the sales trend of the three months; observed that the Beauty Makeup and Clothes got obvious trend, however Jewellery Accessories has increased at the end of March, the rest kinds of goods are relatively stable, so these three kinds of product will be mainly used for further analysis and exploration.



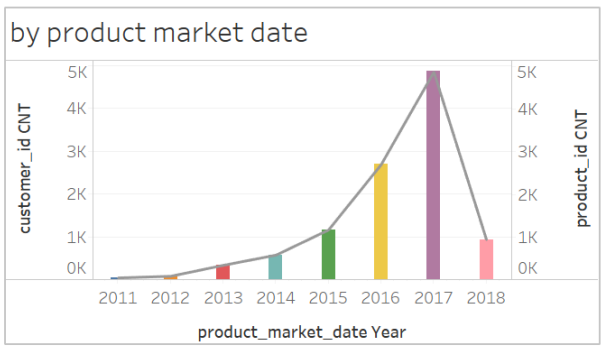
Beauty Makeup decreased from the end of March, mainly due to Facial Essence reduced after checking, there's little difference among the brand, but still can see other brand (non-famous brand) is the major cause compared to big/famous brand. This kind of trend may be caused by season, quantity demanded of facial essence reduced because of warmer weather. So we can make the marketing scheme according to this situation.

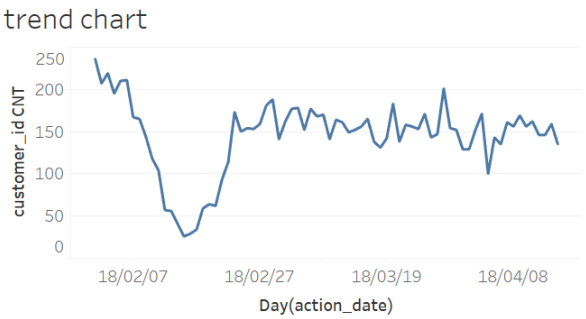


Clothes never comes back rather than keep low and stable, check category found that coat is the key factor, because weather got warmer the requirement of coat get down, also the other brand of coat is more popular.

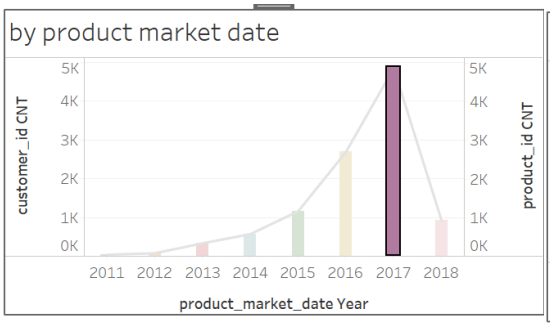


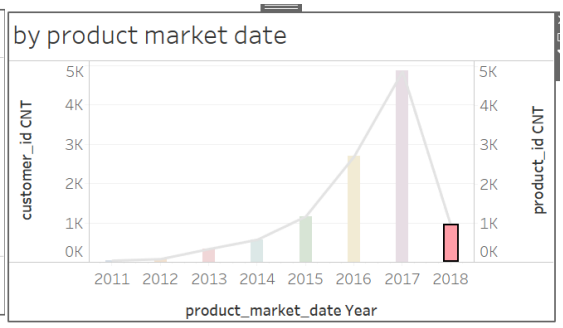
The second largest quantity is Food, also can see tea is the most popular, so we can keep this part of customer for considering further marketing scheme about tea (Lipyon).

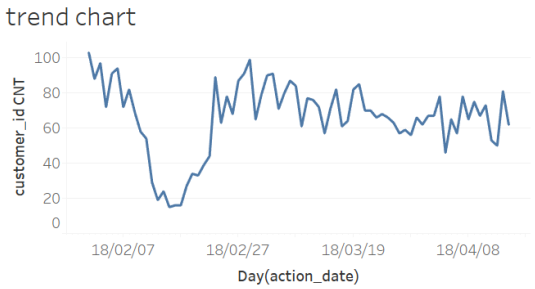


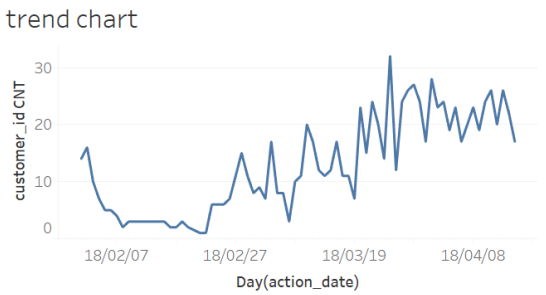


For effect of product market date on sales quantity, conclude that the date is closer to current day the sales would be better (2018 only got data till April, so can ignore first)









We can see the trend chart clearly that down trend in 2017 and before, increasing trend in 2018.

Conclusion:

- Male customers have stronger purchase capability than female users.
- The member and city level impact little on order quantity
- The sales have positive correlation with the shop score (<9.7) and the number of fans/member
- Beauty Makeup and Food are the main part of orders
- Product market date is an important factor for customers' selection, the newer is better.

3) Prediction of users purchase trend

Clustering analysis: Customer Segmentation according to different data dimensions, select male customer with age range 5 from users' personal info, member level 1 and city level 4 from account info, spend more on electrical appliance/furniture/electronics from product info to compose one group, select female customer with age range 5 or 6, member level 1 city level 4 prefer to buy beauty makeup or mother and infant in shop with high score or more fans' number to join one group. K-means clustering analysis to keep difference as large as possible among the groups, while smaller distinction within the group. Also need to consider if the users in group got the enough quantity and these groups can be contacted.

Classification analysis model:1. According to the characteristics of users or products in the existing data, convert abstract to model, and predict for classification to achieve precise marketing for users. Organize some activities or establish product response model to convert users those who are not the buyer to buyers, mainly about the part of users' action type is not order, analyze their behavior

2. For customers who have the potential to purchase or who got a single purchase structure, can build cross-sell or up-sell models to optimize the value of existing customers and increase customers' loyalty. For example, for female customers, member level 1, city level 4 with a stable and flat sales status in clothes and food, we need to consider expanding the various selections of product for this kind of group

3. At the stage when the customer has a risk of loss, such as Beauty Makeup, facial essence, we can see there's a decreasing trend recently, as for this situation, we can build an early warning model of customer loss or a customer win back model.