Part 2 - 3

a. How many:

- **Store shopping trips are recorded in your database?**

  Answer: there are 7,596,145 store shopping trips in total.

- **Households appear in your database?**

  Answer: there are 39,577 households in total.

- **Stores of different retailers appear in our database?**

  Answer: there are 863 retailers in total.

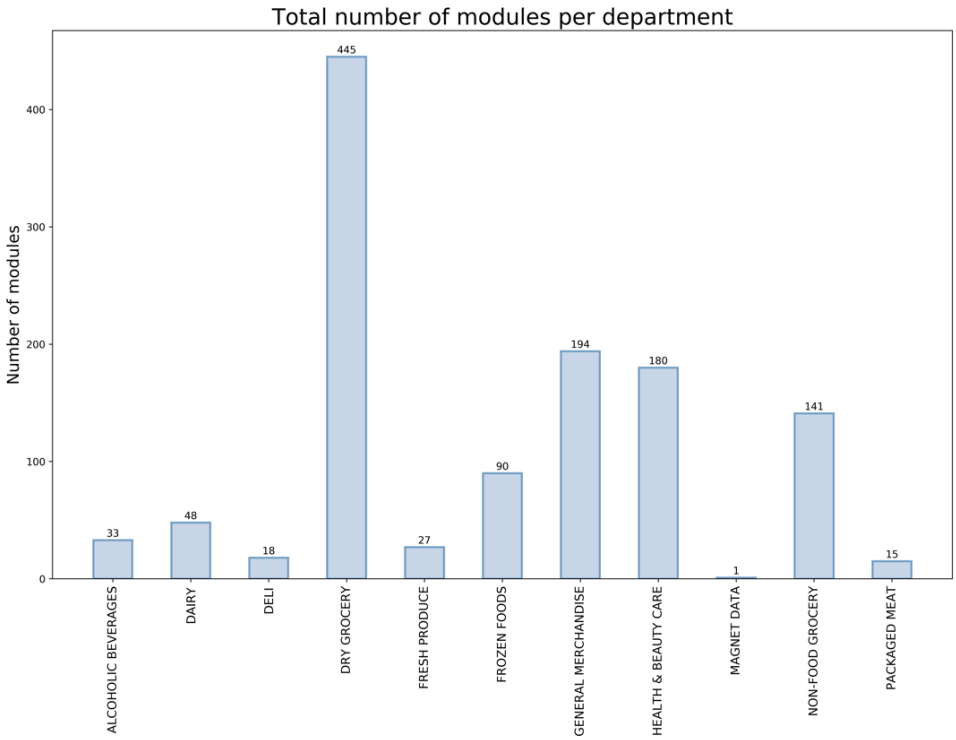
- **Different products are recorded?**

  Answer: there are 4,231,283 different products are recorded in total.

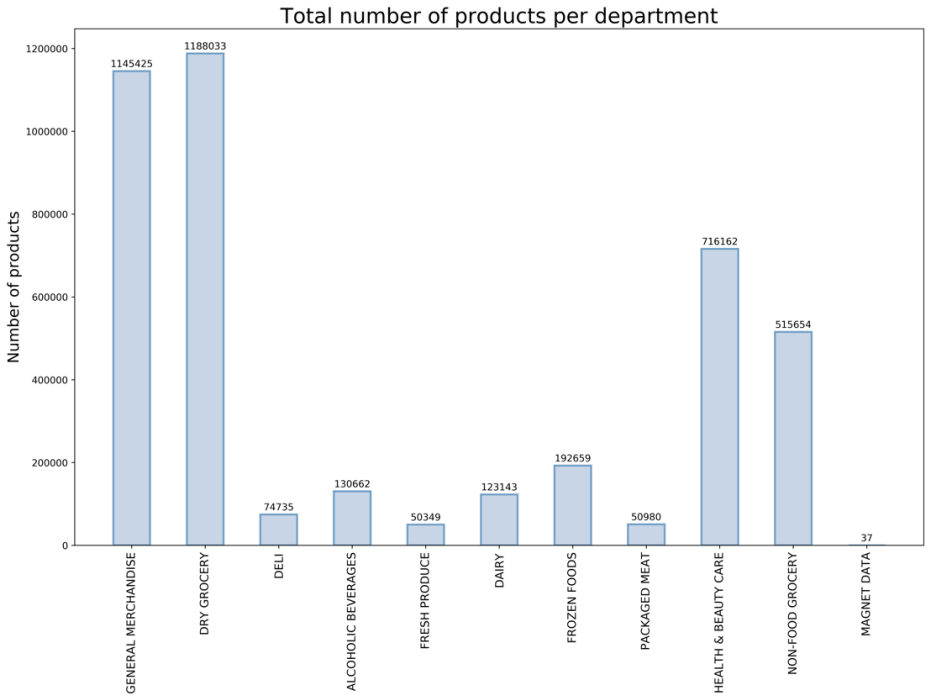
i. Products per category and products per module

Answer: there are 118 different products per category, and 1,224 different products per module.

ii. Plot the distribution of products and modules per department



In the graph *Total number of modules per department*, module 'Dry grocery' has the greatest value, followed by 'General merchandise', 'Health & beauty care' and 'Non-food grocery'; the number of module 'Magnet data' is the smallest.



In the graph *Total number of products per department*, the number of 'Dry grocery' is the greatest, followed by 'General merchandise', 'Health & beauty care' and 'Non-food grocery'; the number of module 'Magnet data' is the smallest.

- **Transactions?**
  - **i. Total transactions and transactions realized under some kind of promotion.**

    Answer: there are 7,596,145 transactions in total, there are 11,384,077 transactions realized under some kind of promotion.

b. Aggregate the data at the household-monthly level to answer the following questions:

i. How many households do not shop at least once on a 3 month periods.

i. Is it reasonable?

Answer: this is not reasonable because for a normal household, it's almost impossible to not shop in 3 months period.

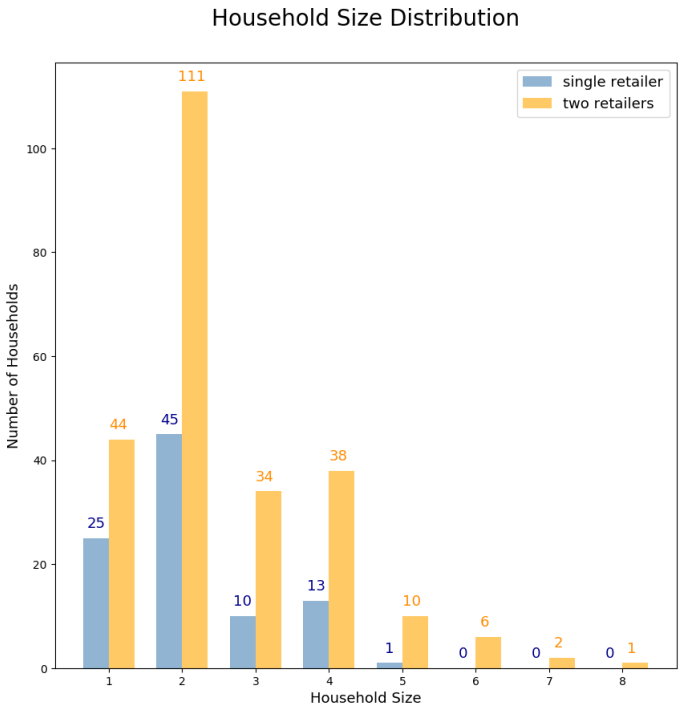
ii. Why do you think this is occurring?

Answer: The reason why this is occurring is that there might be someone who seldomly go shopping, instead, they usually recieve donations from others(the charity, neighbors). Those data is kind of weird as a result, we decided to exclude those data!

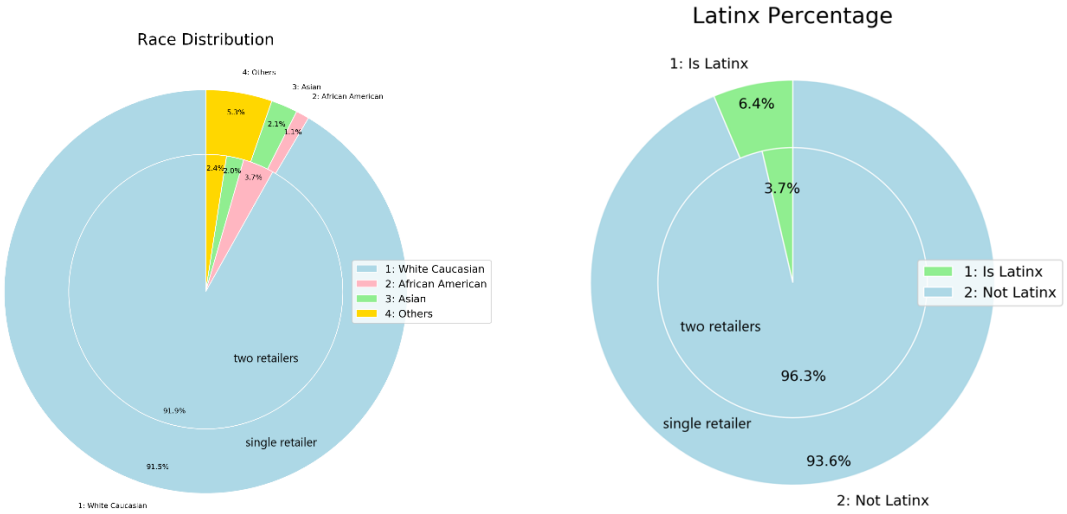
ii. Loyalism: Among the households who shop at least once a month, which % of them concentrate at least 80% of their grocery expenditure (on average) on single retailer? And among 2 retailers?

Answer: there are 94 households who are loyal to single retailer, which accounts for 0.24 % of the total(39577) households. And there are 246 households who are loyal to two retailers, which accounts for 0.62 % of the total(39577) households.

i. Are their demographics remarkably different? Are these people richer? Poorer?



When analyzing household demographics, firstly, we focus on the household size, namely, the number of family members. The bar chart shows similar household size distribution. For both kinds of households loyal to single retailer and those loyal to two retailers, the small families, with no more than 5 members, account for the majority. Households with only one or two people weighs more than 50%, and they are more likely to be loyal to single retailer than two retailers. These phenomena all indicate that small families have higher brand loyalty, probably because tastes are more various when family member increasing.

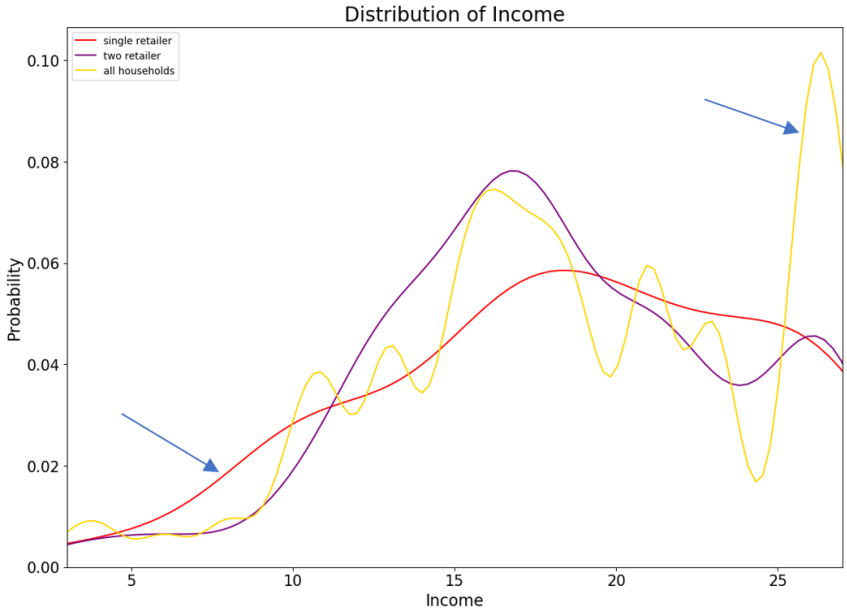


Then we look at the race distribution. Both two kinds of households are mainly White Caucasian, having a portion of over 90%, and the portion of Asian households is almost the same. The main difference between households loyal to single retailer and those loyal to two retailers is that the latter have a larger portion of African American, which is 3.7% while 1.1% for the former.

We also consider whether latinx household is related to loyalism. The pie chart shows the difference. Among households loyal to single retailer, latinx has a larger portion (6.4%), compared with those loyal to two retailers (3.7%).

Compared to American race demographics in 2000 (White Caucasian: 75.1%, African American: 12.3%, Asian: 3.8%, Latinx: 12.5%), we can conclude from the pie charts above that, among the loyalists, there are more white Caucasians than other races.

Overall, demographics of households loyal to single retailer and those loyal to two retailers has some small different but maybe not remarkable.



According to the kernel distribution of household income, we can find that households loyal to single retailer have a larger probability to have lower income, and those loyal to two retailers are more likely to have a medium income level. When the household income is higher than $18,000 a year, the probability to be loyal goes down. And we can see remarkable gap in probability between the overall and those with loyalty over the income of $25,000. These all indicate that loyal households tend to have lower income; namely, they are poorer.

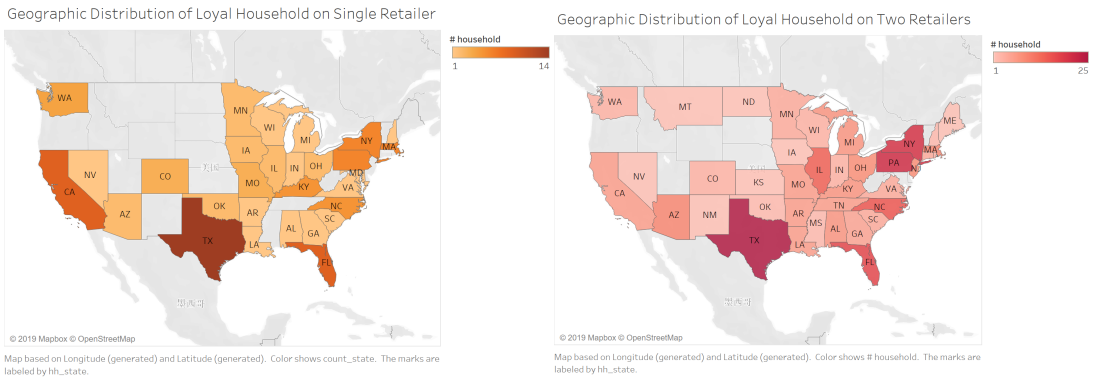
ii. What is the retailer that has more loyalists?

Answer: 6920 is the retailer with the most loyalists.

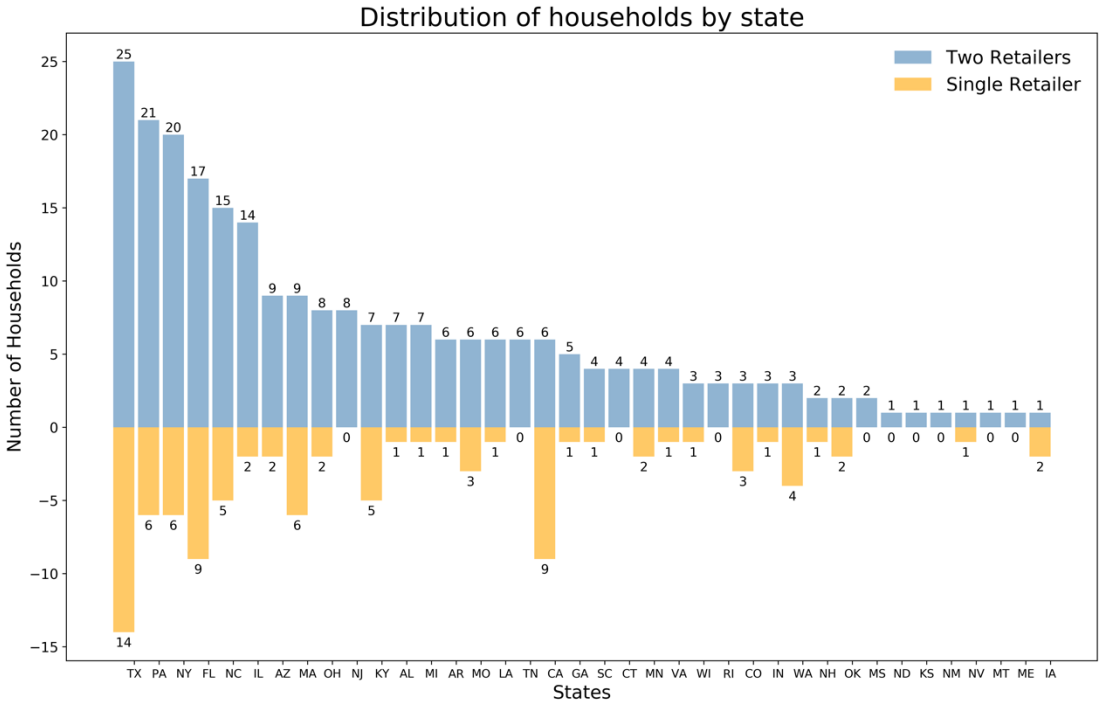
Retailer Code	Number of Loyalists
6920	32
6901	27
4904	26
9999	20
7003	18
6905	15
5850	14
5853	13
6904	11
6999	10

Table 1. Top ten retailers that have the most loyalists.

iii. Where do they live? Plot the distribution by state.



Among the households loyal to one retailer, households from TX account for the greatest proportion, followed by those from FL and CA, then come PA, NY, MA, and KY.
Among the households loyal to two retailers, the state accounts for the greatest proportion is also TX, then PA, NY, FL, NC, IL.



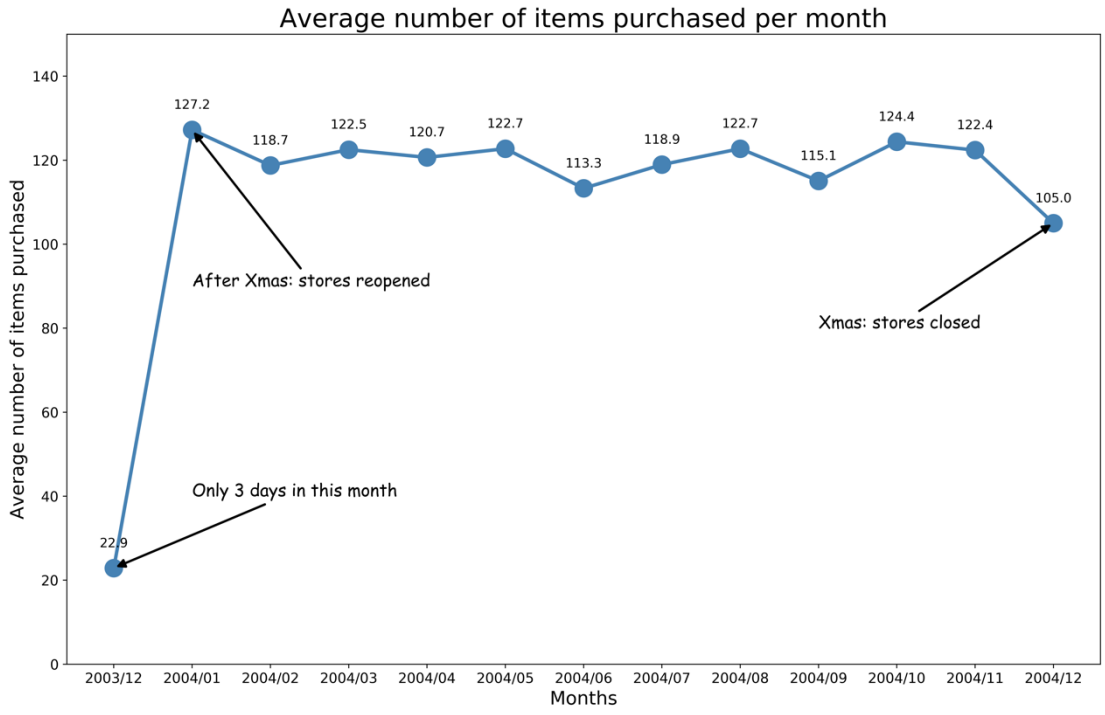
As for loyalism, we think conclusions cannot be drawn that people from TX, FL, PA, NY are more loyal to the retailers they shop from.

1) The y-axis of the graph shows the number of household rather than the percentage, comparing the absolute value cannot yield valid inference about loyalism of people from different states because we didn't take the customer base in different areas in to account.
2) The data doesn't take the available number of retailers in different states into account. We cannot say the households are loyal to one or two retailers when those retailers are the only choices for them.

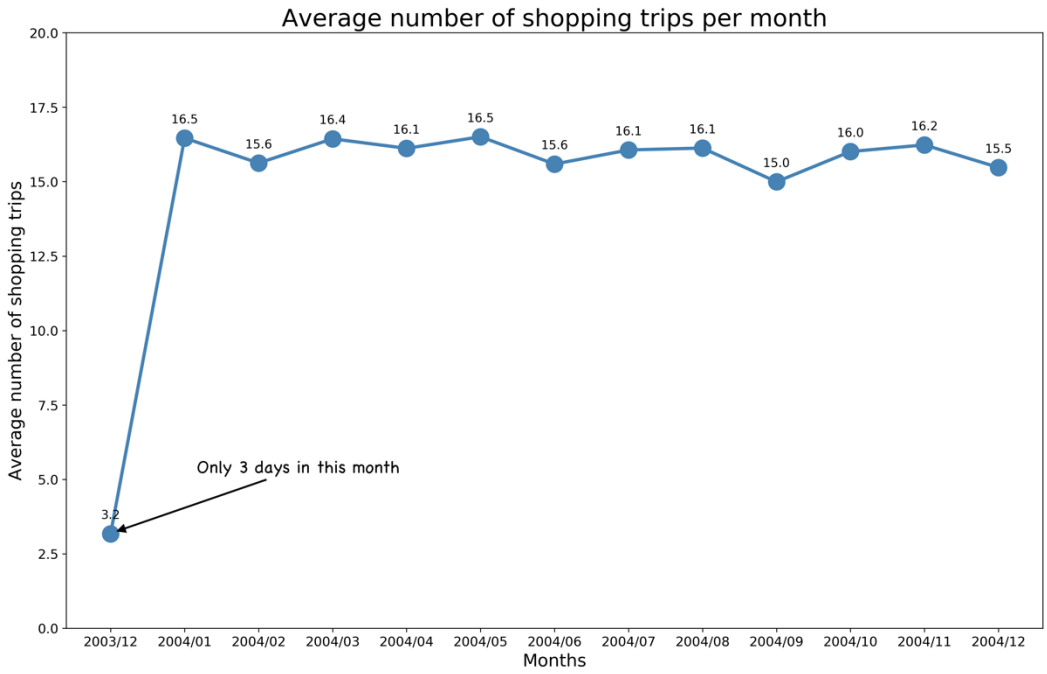
- **Plot with the distribution:**

i. Average number of items purchased on a given month.

The average number of items purchased per month is shown in the graph above. We can see the value is the lowest in Dec. 2003, it does not mean people purchase less in this month, and we believe this is due to the data we have only includes 3 days of Dec. 2003, that is, the value (22.9) is not a valid value for average number of items purchased per month in this case. Removing the value of Dec.2003, we can see the item purchasing level keeps fluctuate a little bit every month throughout the year 2004, there are 2 significate points: 127.2 at Jan. 2004 and Dec. 2004. In explaining this, we believe the purchasing increase in Jan. 2004 is due to stores closed during Christmas reopened, thus the stores available for customers increased and this may contribute to this growth; likewise, the decreasing purchasing level in Dec. 2004 is due to the fewer available stores for customers because of Christmas.

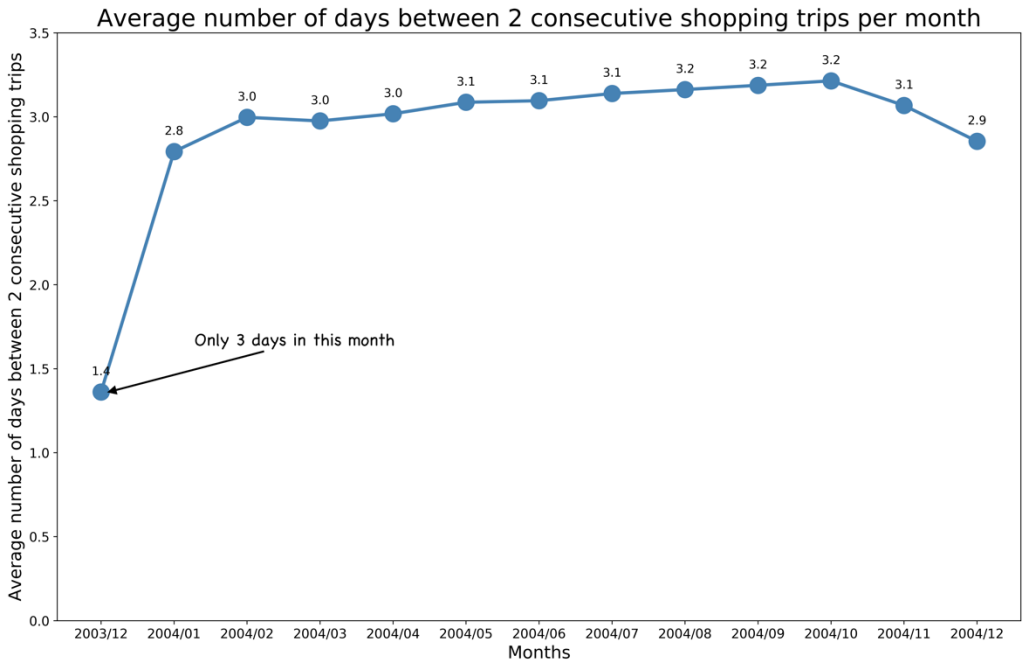


ii. Average number of shopping trips per month.



The average number of shopping trips per month is shown in the graph above. We can see the value is the lowest in Dec. 2003, it does not mean people go shopping less in this month, and we believe this is due to the data we have only includes 3 days of Dec. 2003, that is, the value (3.2) is not a valid value for average number of shopping trips per month in this case. Removing the value of Dec.2003, we can see the number of average monthly shopping trips keeps fluctuate a little bit every month throughout the year 2004.

iii. Average number of days between 2 consecutive shopping trips.

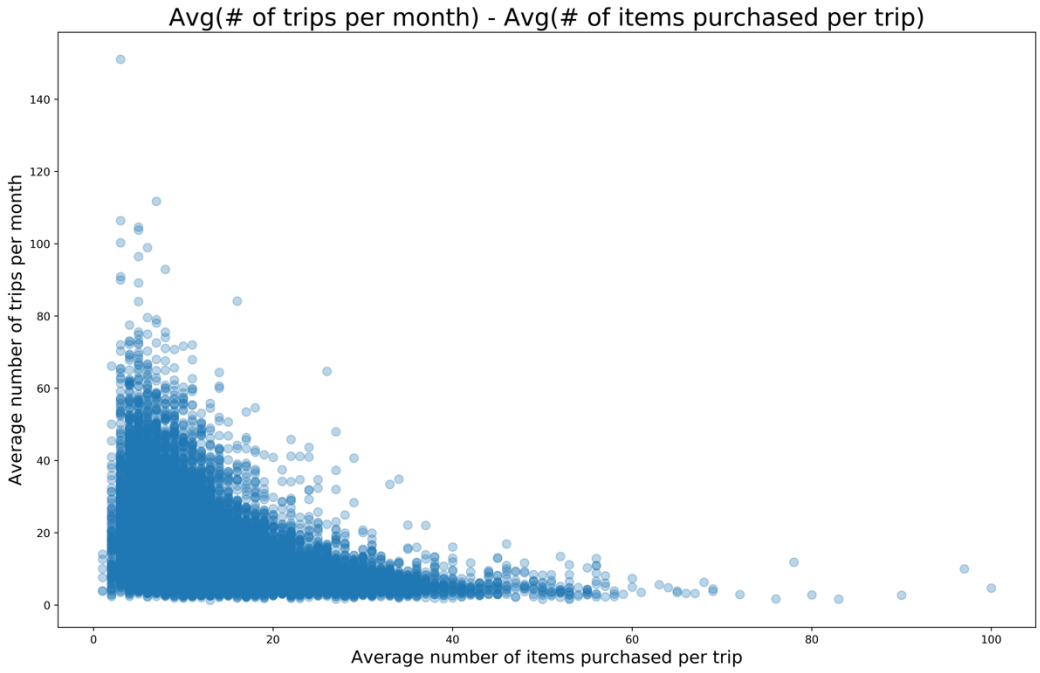


The average number of days between 2 consecutive shopping trips is shown in the graph above. We can see the value is the lowest in Dec. 2003, it does not mean people go shopping less frequently in this month, and we believe this is due to the data we have only includes 3 days of Dec. 2003, that is, the value (1.4) is not a valid value for the average number of days between 2 consecutive shopping trips in this case. Removing the value of Dec.2003, we can see the number of average number of days between 2 consecutive shopping trips keeps fluctuate a little bit every month throughout the year 2004.

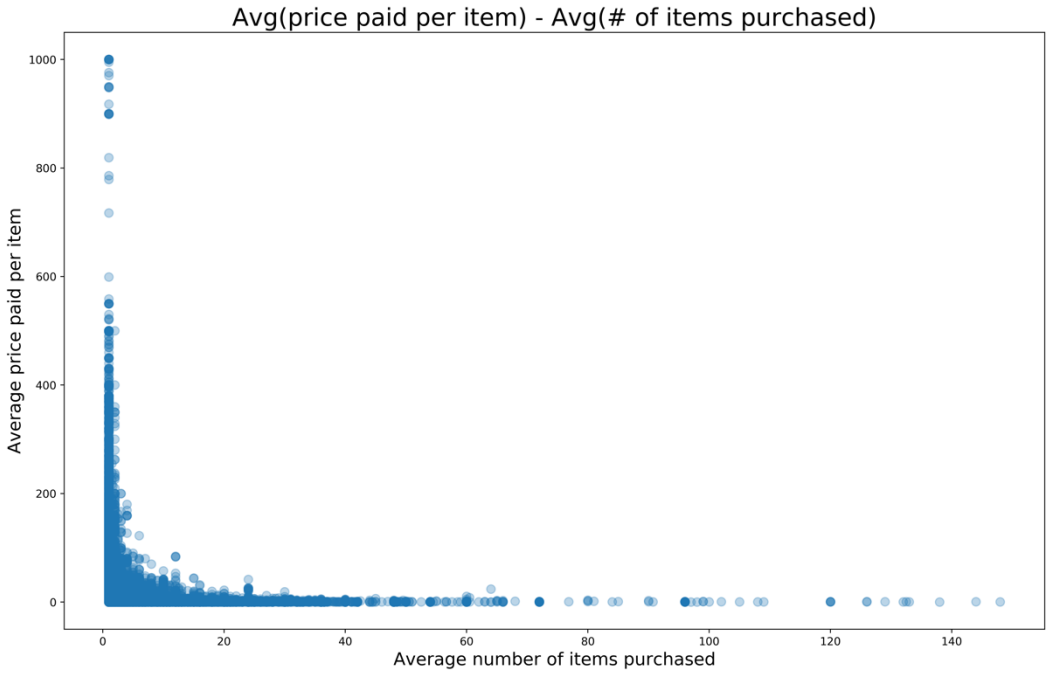
c. Answer and reason the following questions: (Make informative visualizations)

- **Is the number of shopping trips per month correlated with the average number of items purchased?**

According to the graph, we believe the number of shopping trips per month is negatively correlated with the average number of items purchased. The trend of this graph is pretty clear that, the average number of trips per month is higher when average number of items purchased per trip is lower, and vice versa, which is reasonable because people tend to shop for more times when they shop fewer items during every shopping trip. For example, some households live far away from big cities and supermarkets tend to shop a great many items during their every shopping trips because it takes time to go to shopping places. Besides, using the data we have in our hands, we can see a cluster of points close the original, this may because these households normally don't shop much, due to low-income or some other factors.

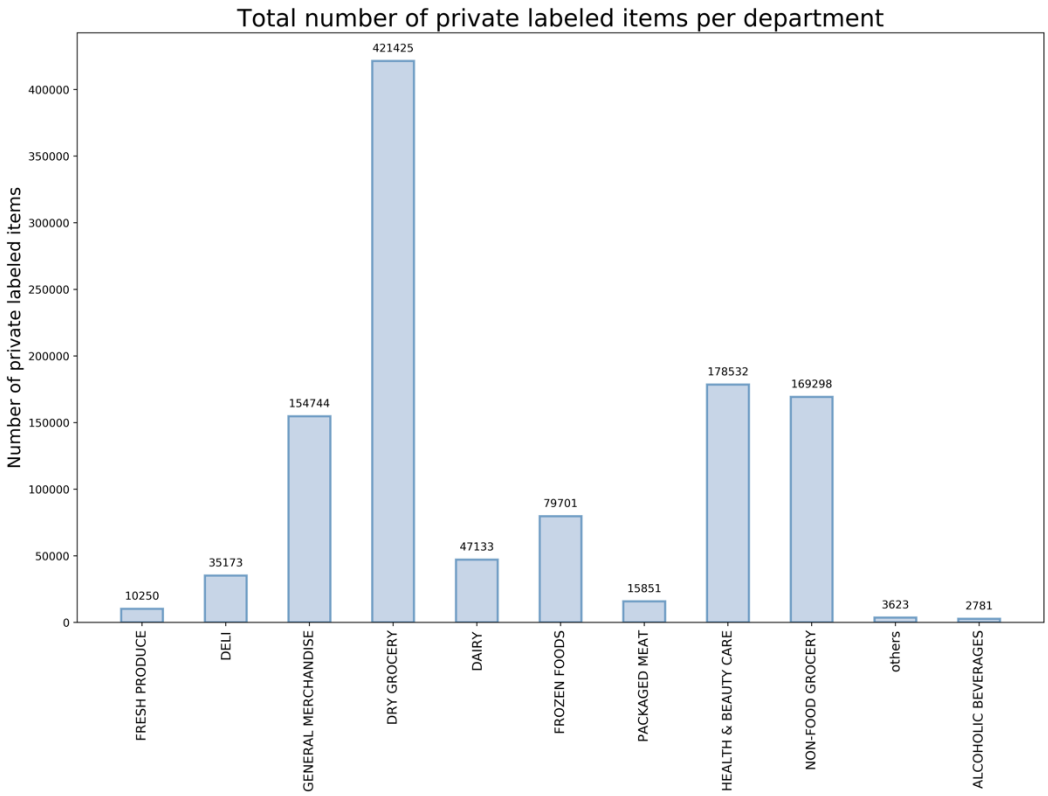


- **Is the average price paid per item correlated with the number of items purchased?**



According to the graph, the average price paid per item is negatively correlated with the number of items purchased. We can see clearly that when the number of items purchased gets higher, the average price paid per item, from which we can confer that people buy more cheap products and fewer expensive products, which is reasonable and common in reality. Besides, we can still see a cluster if points located near the original, representing those household who purchase fewer cheap items, thus we infer that these households may be of small family size (they require less) or they earn little (they only can afford less).

- **Private Labeled products are the products with the same brand as the supermarket. In the data set they appear labeled as 'CTL BR'**
    - **i. What are the product categories that have proven to be more "Private labelled"**

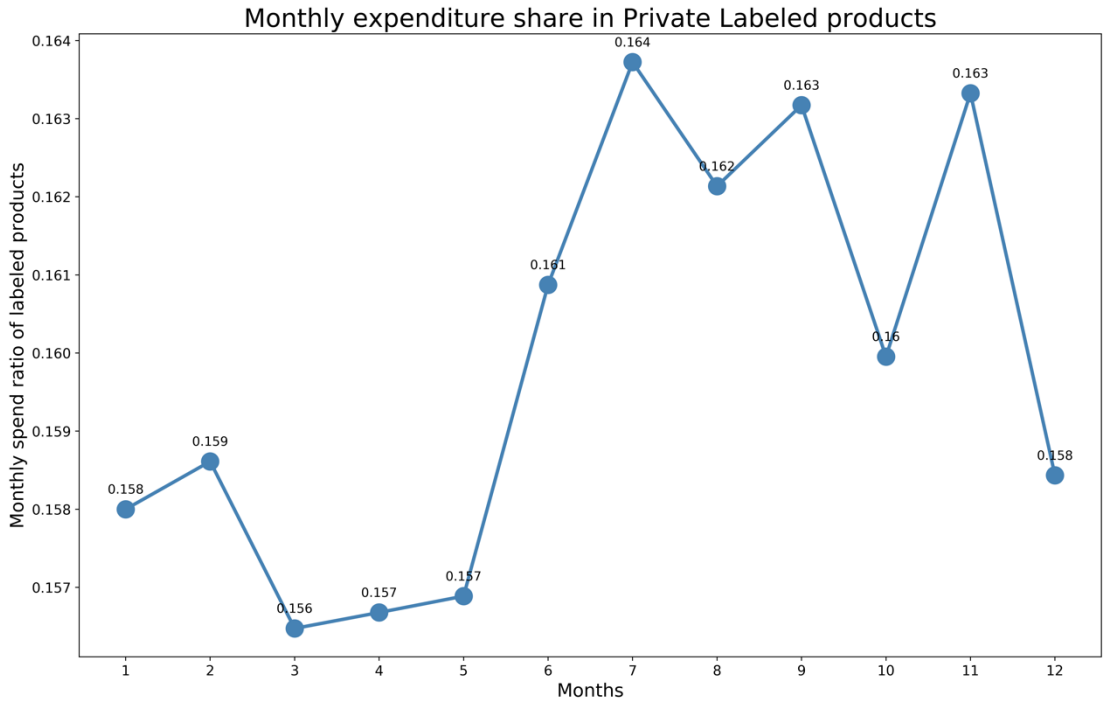


As shown in the graph above, we can see among all the categories of products, the greatest number belongs to 'Dry grocery', followed by 'Health & beauty care', 'Non-food grocery' and 'General merchandise', so these four categories, especially 'Dry grocery' has proven to be more private labelled.

- **ii. Is the expenditure share in Private Labeled products constant across months?**

According to the graph *Monthly expenditure share in Private Labeled products*, we can easily see that it is not constant across months; rather, there is significant fluctuation among the months. The monthly expenditure share in Private Labeled products keeps lower than 16%(0.160) from January to May 2004, then it boomed to 16.1% in June, keeps high around 16.3% from July to September, drops at October, and increases again at November, and then decreases at December. In explaining this, we believe that every year from December to May next year, people only shop private labeled products when they need them; but as summer approaching at June, it's getting warm outside and households would go out to go on small trips or have picnics more and buy something eatable and easy to take out with (products belong to 'Dry grocery'). And as weather gets colder, people tend to shop less these products thus the share of private labeled drops; but November is an exclusion, we infer that maybe

the Black Friday promotion in every November encourage households to shop more Dry groceries for storage, as well as products belong to 'Health & beauty care', 'Non-food grocery' and 'General merchandise', which contributes to the increase in the share.



iii. Cluster households in three income groups, Low, Medium and High. Report the average monthly expenditure on grocery. Study the % of private label share in their monthly expenditures. Use visuals to represent the intuition you are suggesting.

From the plot *% of private label share in average monthly grocery expenditures*, we can see that households with higher income level spend more on grocery. The households of high-income level spend $161,000 on grocery monthly, while households of low-income level spend $108,000. However, the expenditure on private-labeled products among all the grocery products seems to have little difference over different income level. Looking at the percentage of private label share, we can find that households with lower income level tend to buy a larger portion of private-labeled products. The reasons behind this phenomenon may include that the private-labeled products usually have lower price, which may be the main factor to consider when households with lower income buy products. This indicates that low-income households are the target of private label.

% of private label share in average monthly grocery expenditures