

BUS 211 FINAL PROJECT ANSWER

a. How many:

i. Store shopping trips are recorded in your database?

7596145

ii. Households appear in your database?

39577

iii. Stores of different retailers appear in our database?

863

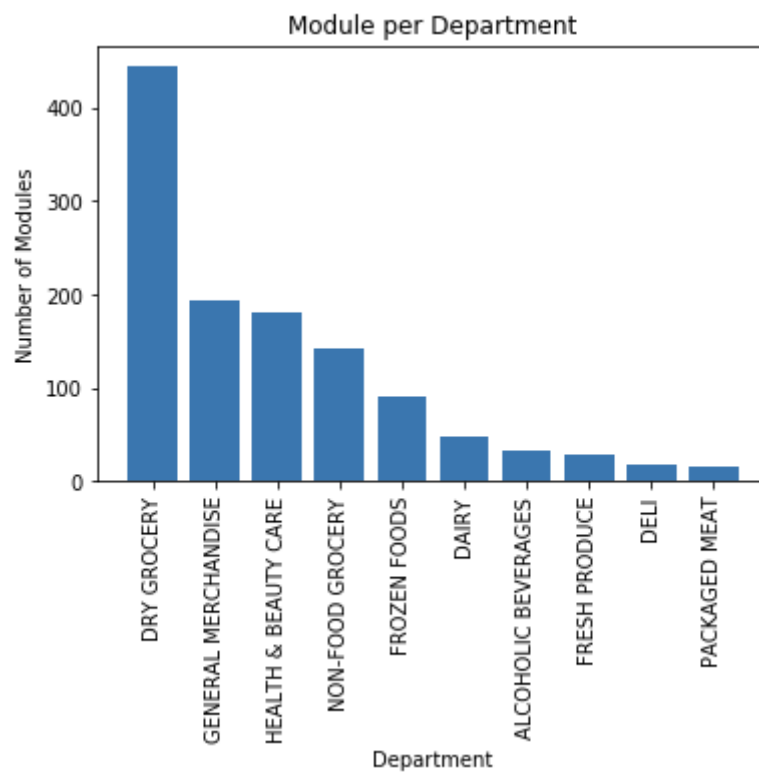
iv. different products are recorded?

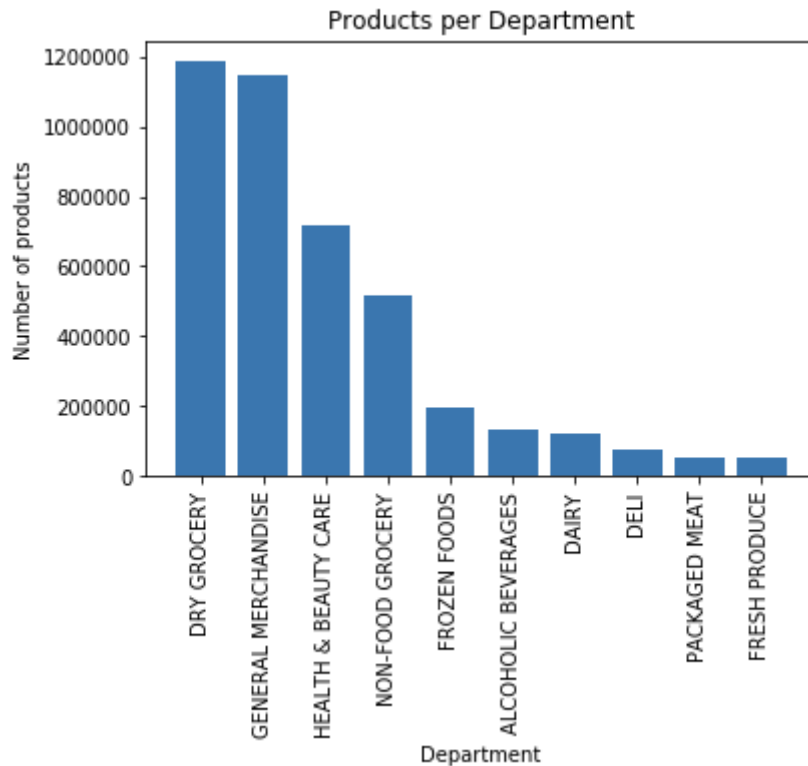
1. Products per category and products per module

category: 11

module: 1224

2. plot the distribution of products and modules per department





Note:

For deeper analysis, I have uploaded the pie charts of products per department, which illustrate the distribution of products in one specific department. However, since there are hundreds of modules in on department, it is hard to tell the distribution from the pie chart, therefore, I decided to give up the pie chart analysis on modules per department. Both the tables of modules and products per department are included in the package, please refer to those if necessary.

v. Transactions?

1. Total transactions and transactions realized under some kind of promotion

Total Transactions: 5651255

Under Promotion: 2603946

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

- count the number of households that do not go shopping in three months

There are 37 households that did not go shopping at least once in three months.

Note:

1. We regard 91 days as three months. If the days between two consecutive shopping trips is no less than 91, then it means that a household does not shop in three months.
2. We define that the days between two consecutive dates is ONE day. (eg. 2004.1.1 and 2004.1.2).

i. Is it reasonable?

From the data that we extracted, we can find that the households that do not go shopping for over 3 months are just very less. So it is reasonable.

ii. Why do you think this is occurring?

It is possible that those house households that did not shop for three months might go out for their summer vacation or for work or study.

- 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 a single retailer? And among 2 retailers?

2.16% of households concentrate at least 80% of their grocery expenditure (on average) on a single retailer.

8.7% of households concentrate at least 80% of their grocery expenditure (on average) on two retailers.

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

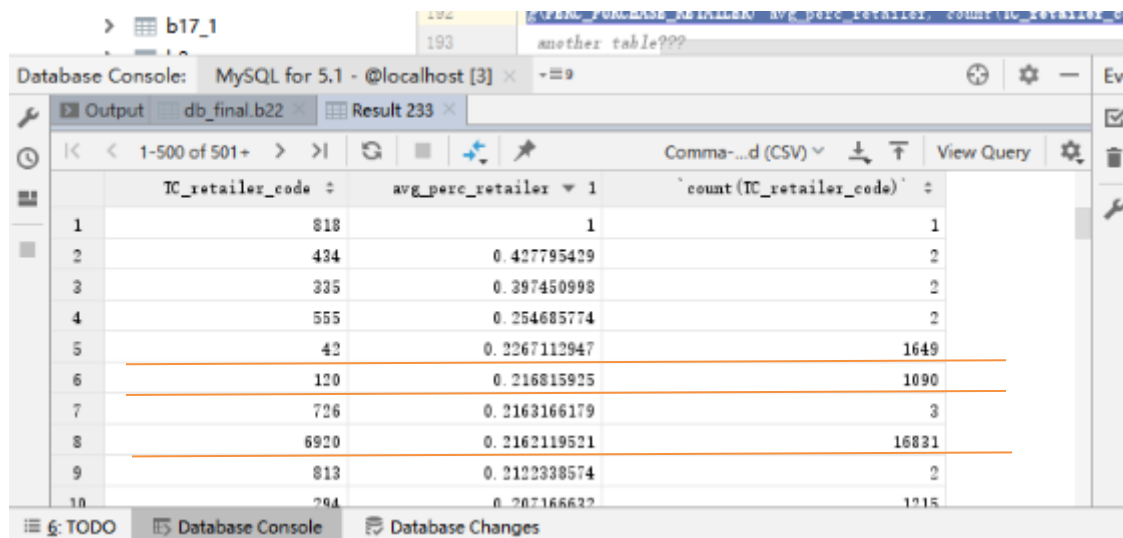
The average income level for all the households is 18.717, while that for the households shopping at least once a month is 18.685.

Thus, we think these households shopping frequently are relatively poorer ones.

ii. What is the retailer that has more loyalists?

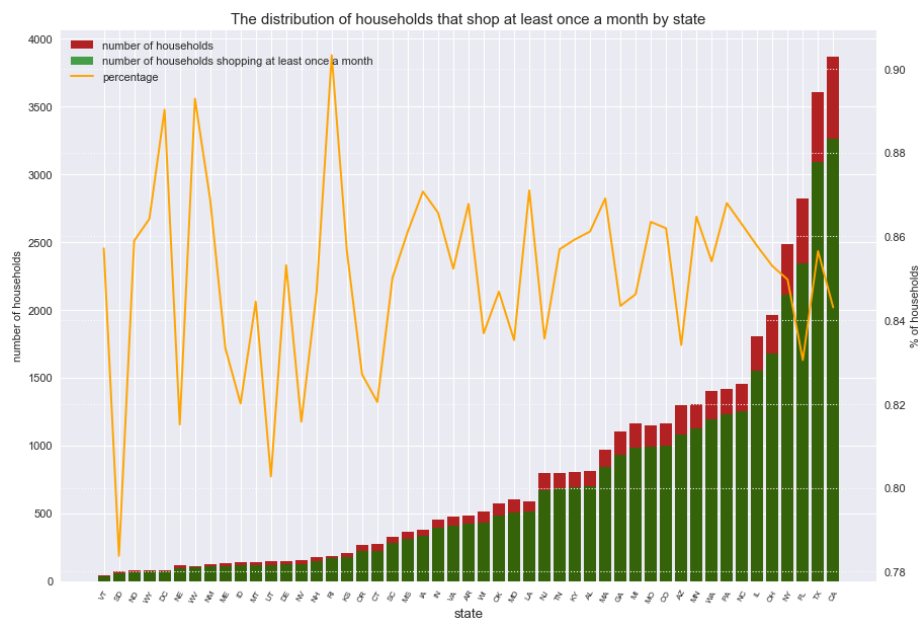
6920 had more loyalists than other retailers. Though 6920 does not have the highest shopping expenditure ratio if considering the number of households that concentrated 80% of their expenditure in a single retailer, 6920 has more loyalists than the others.

Retailer 42 and retailer 120 also had relative more loyal customers.



	TC_retailer_code	avg_perc_retailer	count(TC_retailer_code)
1	818	1	1
2	434	0.427795429	2
3	335	0.397450998	2
4	555	0.254685774	2
5	42	0.2267112947	1649
6	120	0.216815925	1090
7	726	0.2163166179	3
8	6920	0.2162119521	16831
9	813	0.2122338574	2
10	794	0.207166622	1215

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



- Plot with the distribution:

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

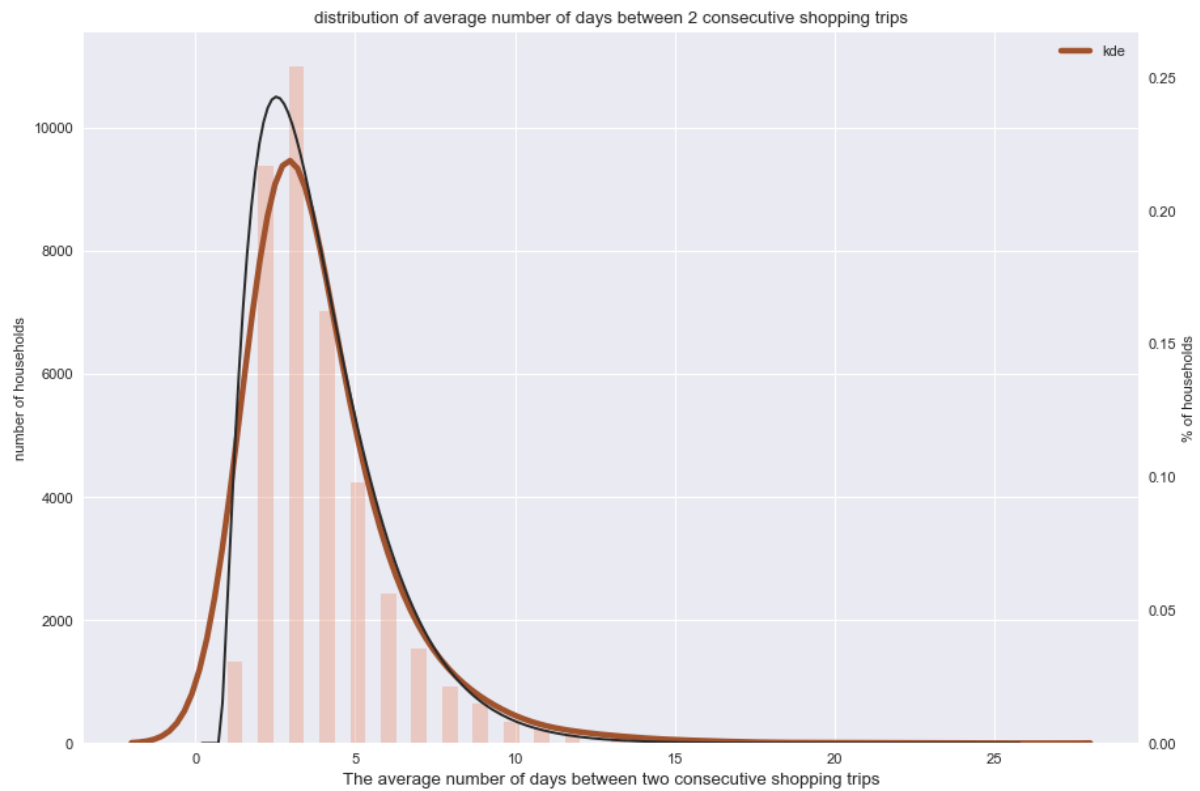


ii. Average number of shopping trips per month.



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

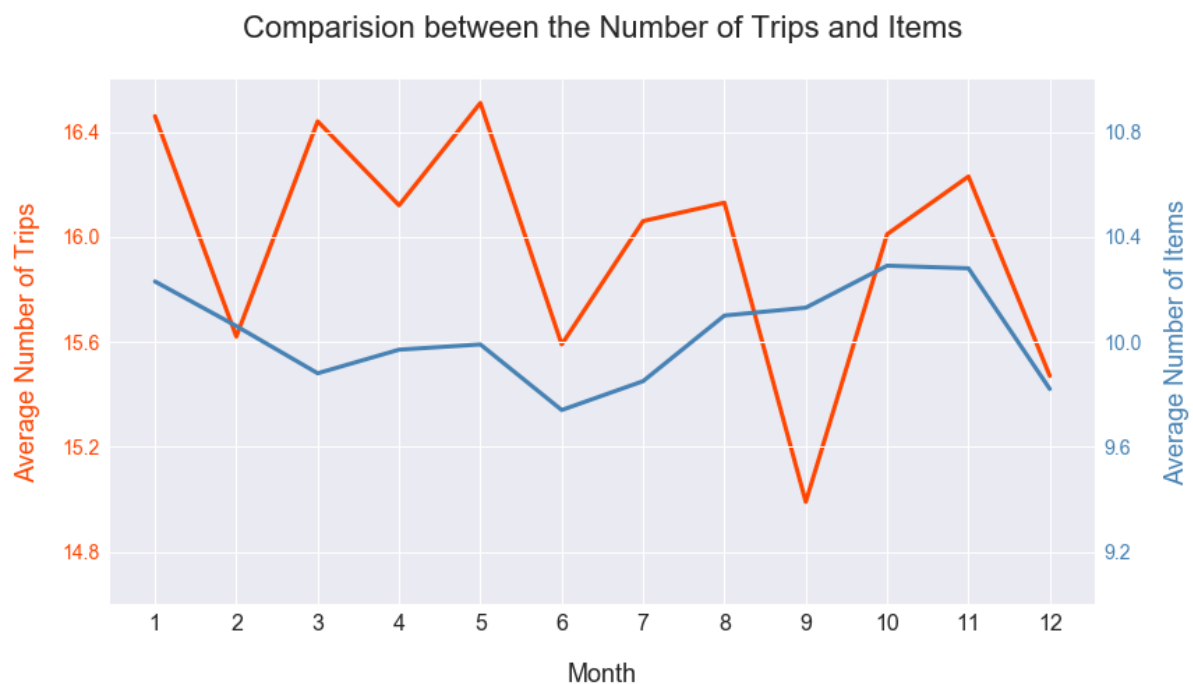
We counted the households that shared the same average number of days between two consecutive shopping trips. From the distribution plot, we can easily find that the kernel density estimation of the distribution is close to that of gamma distribution (black line).



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

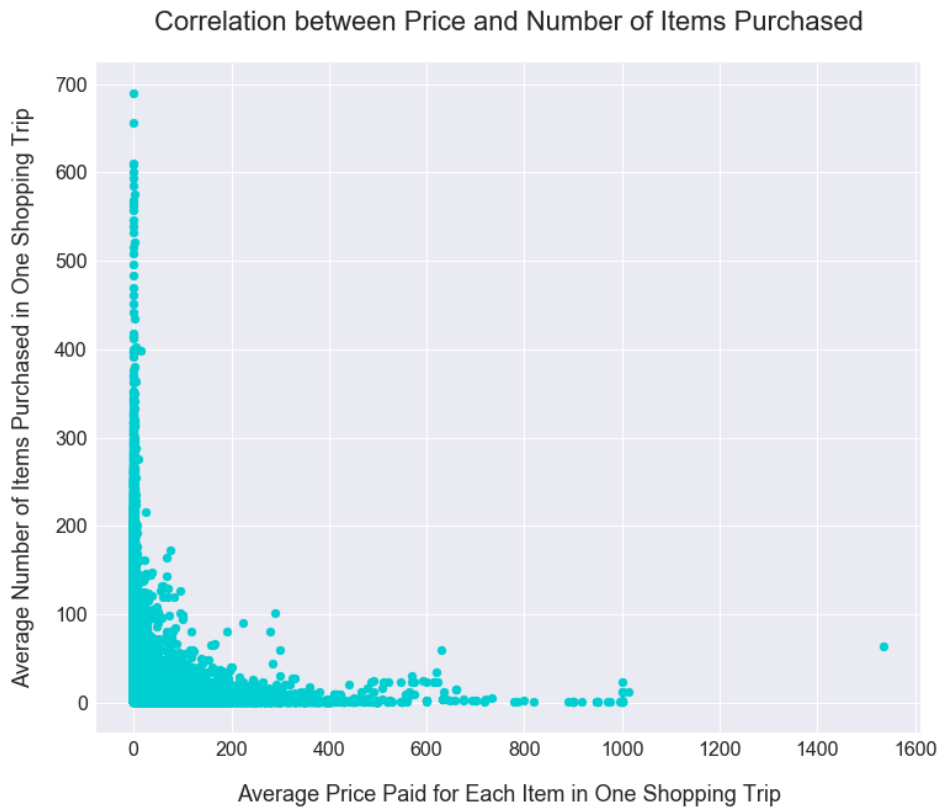
- **Is the number of shopping trips per month correlated with the average number of items purchased?**

In most months, the number of shopping trips and items both have the same trends. For example, from June to August, there is a rising trend in the average number of trips and the same trend also shows in the average number of items. However, in March and September, there is an opposite trend in trips and items. In March, the average number of trips increases, but the average number of items decreases. Conversely, the average number of trips decreases, but the average number of items increases.



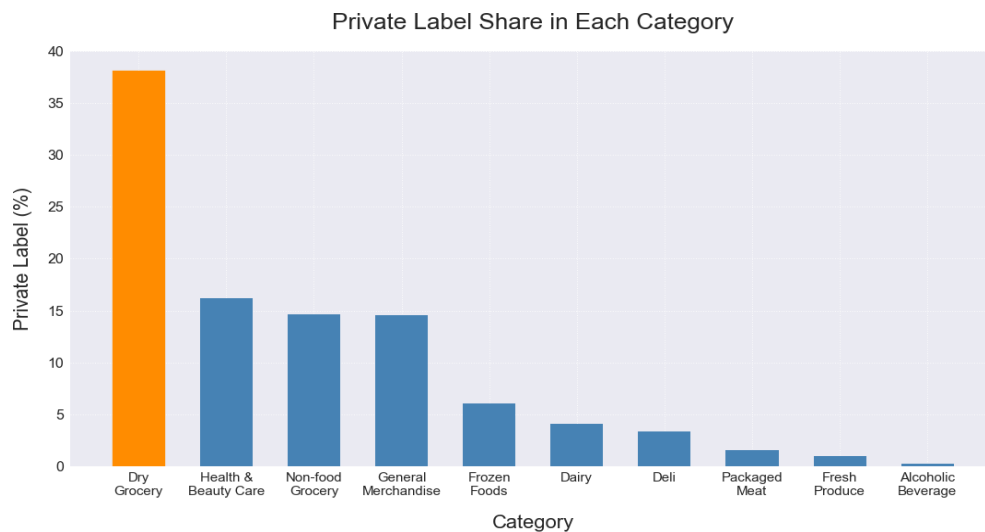
- **Is the average price paid per item correlated with the number of items purchased?**

Yes, the average price paid per item is negatively correlated with the number of items purchased. In the scatterplot, the distribution of points shows “L-shape”. It implies that when an item has a lower unit price (about less than 20 dollars), the range of numbers that people are willing to purchase will be large (from 0 to 700). That is to say, people are willing to purchase more products when the unit price is low. On the other hand, when the unit price goes up more than 100, people tend to purchase few units of products. In fact, lots of points centralize in the bottom left corner. It represents that when an item charges a medium price (about 30~300), people tend to buy less than 100 units.



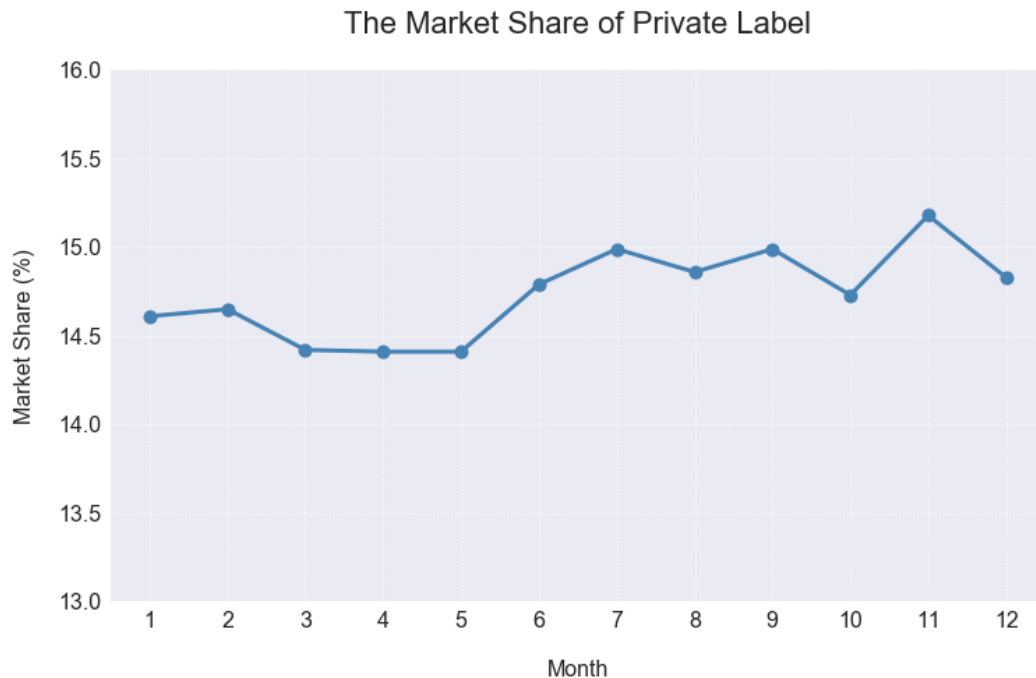
- **Private Labeled products are the products with the same brand as the supermarket. In the data set, they appear labeled as ‘CTL BR’**
 - i. **What are the product categories that have proven to be more “Private labeled”?**

Dry grocery has more “private labeled” products, which accounts for about 38% in this category.



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

Yes, the expenditure share in Private Labeled products is generally constant across months, although there is a slightly upward trend from about 14.5% to 15.3%.



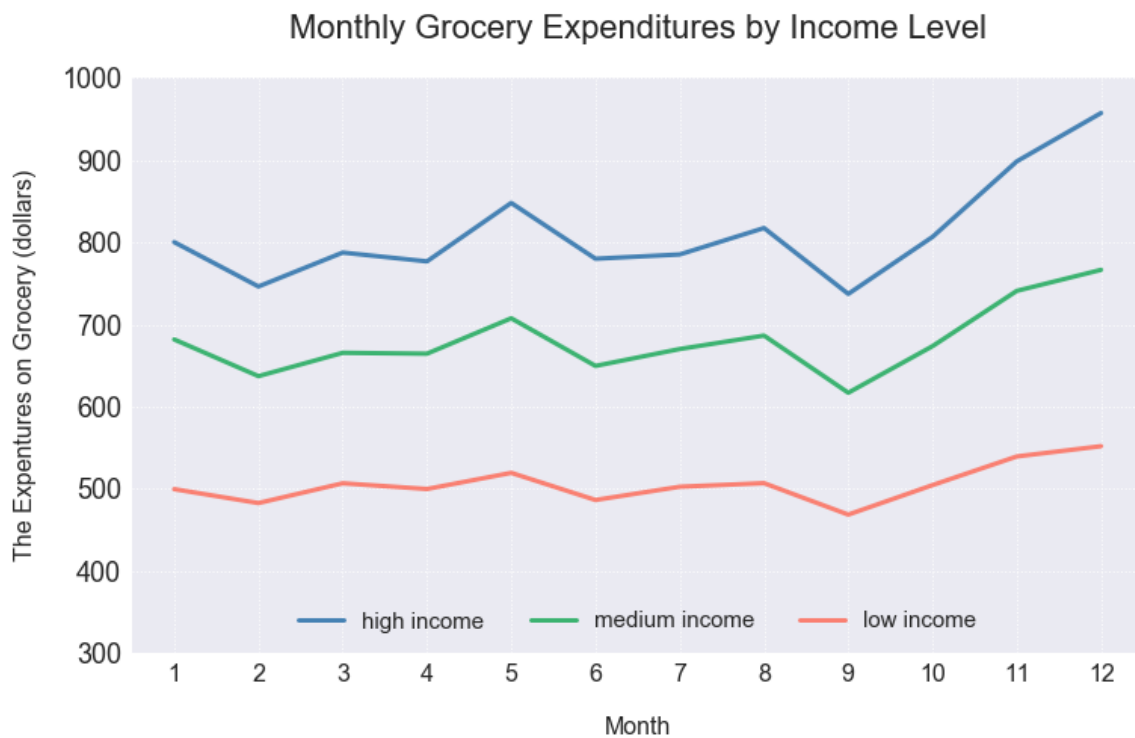
iii. Cluster households in three income groups, Low, Medium and High. Report the average monthly expenditure on the grocery. Study the % of private label share in their monthly expenditures. Use visuals to represent the intuition you are suggesting.

At first, we can define three kinds of income level as the following:

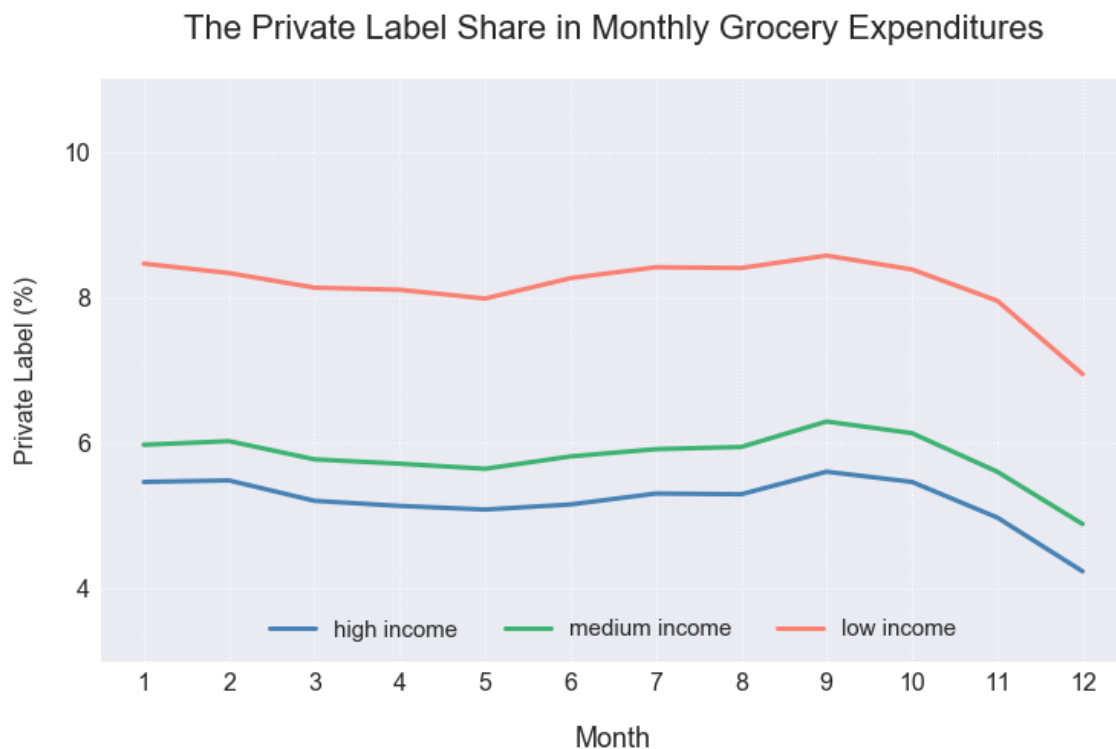
- Low income: < \$49,999
- Medium income: \$50,000 - \$99,999
- High income: > \$100,000

In monthly grocery expenditures, high-income households spend the most, medium-income is the next, and low-income households is the last. Overall, All income levels have a similar trend on monthly grocery expenditures. Before September, the trend maintains

constant, but significantly goes up after September.



However, in the private label share, low-income households spend the most on the private label products, medium-income households are the next, and the last one is low-income households. Overall, there is a consistent trend in three kinds of households. Before October, the trend keeps constant, but dramatically goes down after September.



When comparing expenditures between all groceries and private label in all income-level households, we can see the total cost of private label products are similar and constant in three categories of households across months, even if the total expenditures on groceries are different.

