Final Project Report

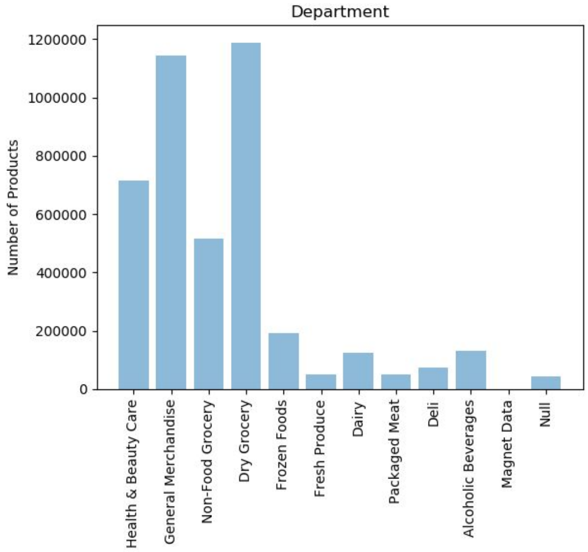
After importing data into Mysql through Python, we analyzed the given consumer panel data and summarized our findings into the following three parts.

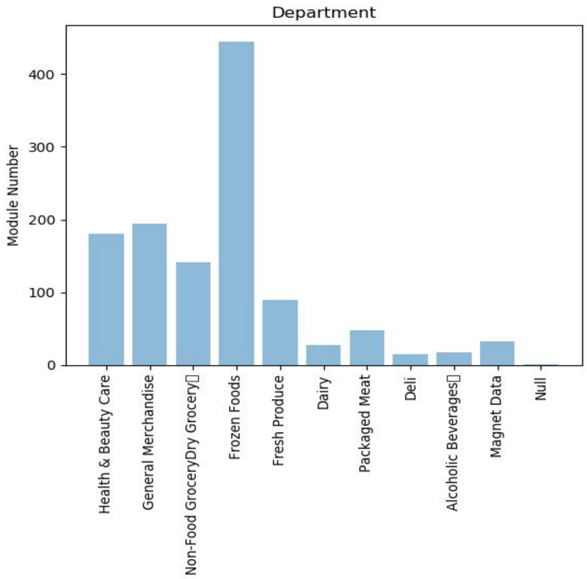
Part A:

For part a, we did some basic analysis on the database. We found that there are 7,596,145 total shopping trips, 39,577 households, 863 retailers and 26,402 stores recorded in the database. There are 4,231,283 different products in total. They belong to the following departments:

Department:	Number of Products
Health & Beauty Care	716,162
General Merchandise	1,145,425
Non-Food Grocery	515,654
Dry Grocery	1,188,033
Frozen Foods	192,659
Fresh Produce	50,349
Dairy	123,143
Packaged Meat	50,980
Deli	74,735
Alcoholic Beverages	130,662
Magnet Data	37
Null	43,444

There are 118 groups and 1224 modules. For number of products in each group and module, please see attached excel file.



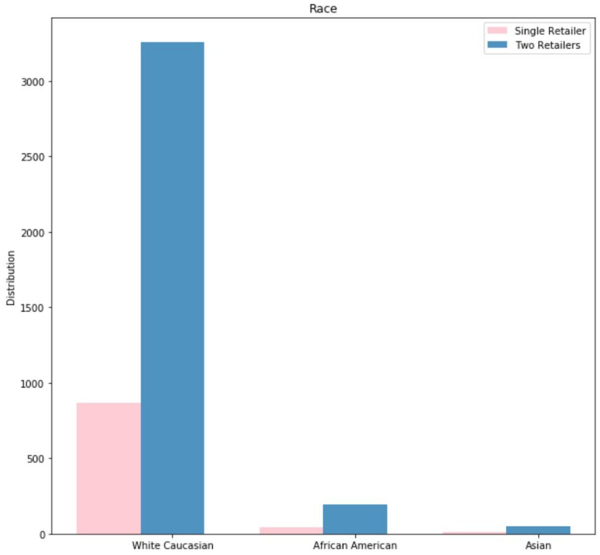


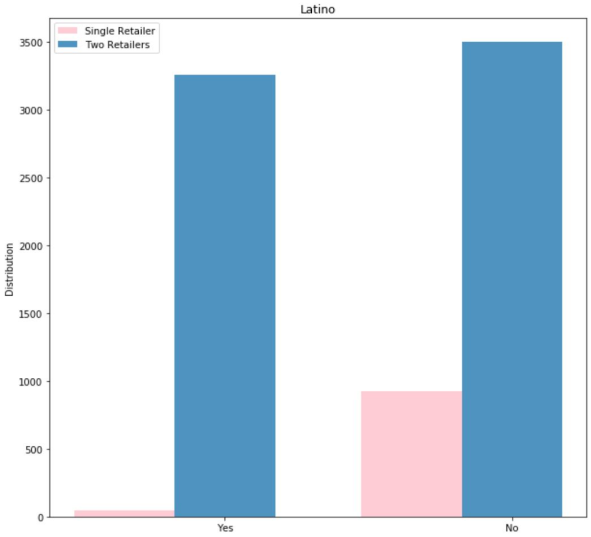
From the products per department and module per department graphs above, we found that General Merchandise and Dry Grocery have largest number of products (over 1,000,000 products per department). In terms of modules per department, Frozen Foods has the largest number of modules -- over 400 while other departments have less than 200.
There are 38,587,942 total transactions. Since this database is too big to be fully loaded into Mysql, we randomly sampled half of the data(19,293,971 transactions) into Mysql. Among those 19,293,971 transactions, 5,823,393 are realized under some kind of promotion(coupon or dealflag). In other words, roughly 30% of transactions are realized under promotion.

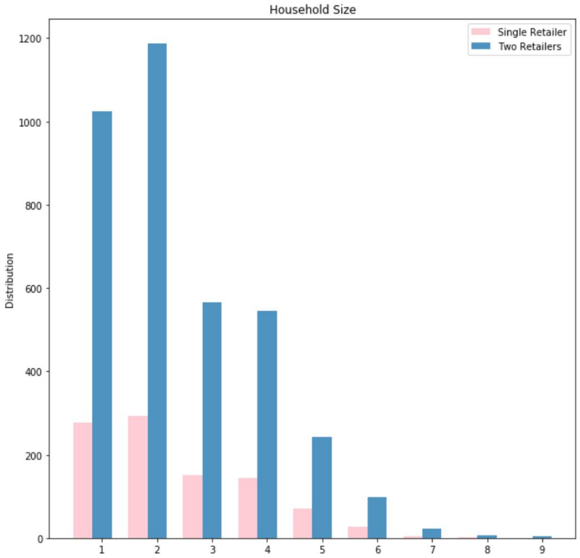
Part B:
For part b, we aggregate data at households on monthly level. We find that 4371 households, which is 12% of total households, don't shop at least once on a 3 month period. This is not reasonable because the percentage of households who don't shop for their daily life is a little bit too high. This may be because some of these households who shop at other places (not in the database) more frequently or who are busy working and don't really shop and cook at home.
For loyalism, 2.45% of households concentrate at least 80% of their grocery expenditure (on average) on a single retailer. 9.34% of households concentrate at least 80% of their grocery expenditure on 2 retailers.

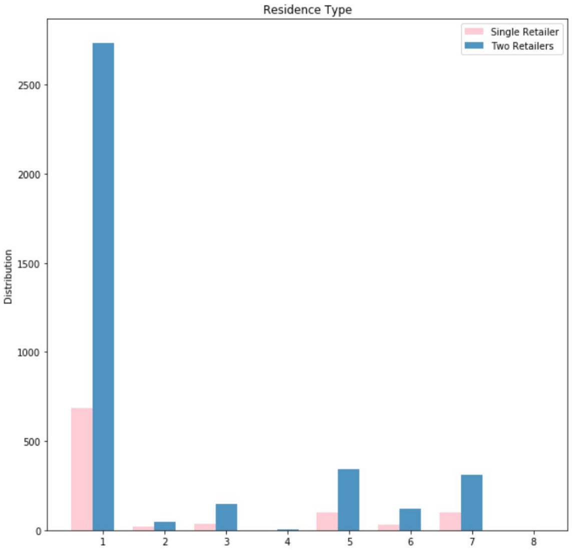
Retailer has more loyalist is the retailer with code of 6920.

Then, we have done some demographic analysis on these loyal customers:









.

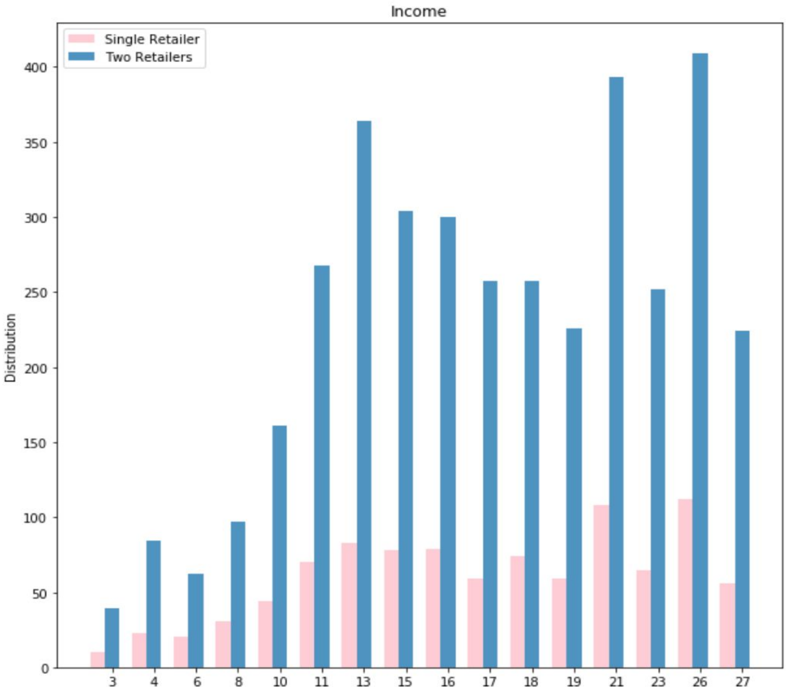
The graph of race shows that for each race, percentage of grocery expenditure concentration on single retailer is remarkably different from that on two retailers. Among these loyal customers, a large percentage of them are white.

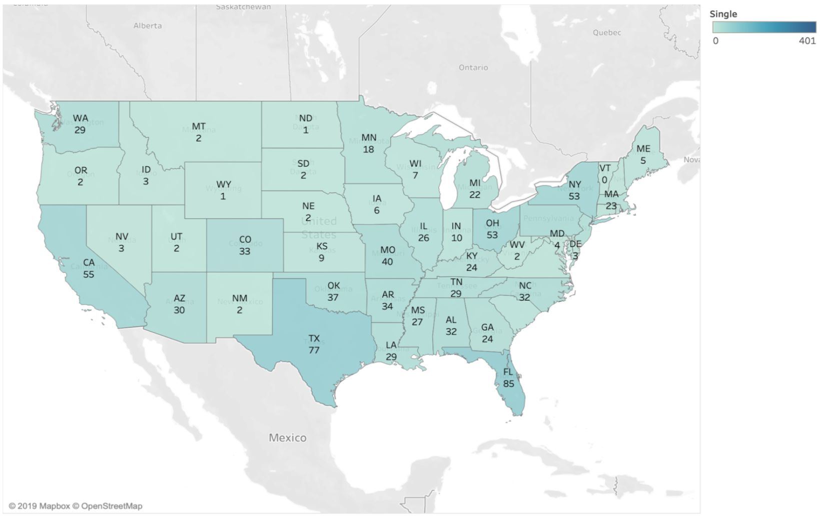
The graph of latino indicates that latino people generally don’t concentrate shopping on a single retailer while non-latino people will concentrate their grocery expenditure on a single retailer more often.

The graph of household size shows that for each size, more households concentrate on two retailers rather than 1. Most of the loyal customers in the database are single member or two-member family.

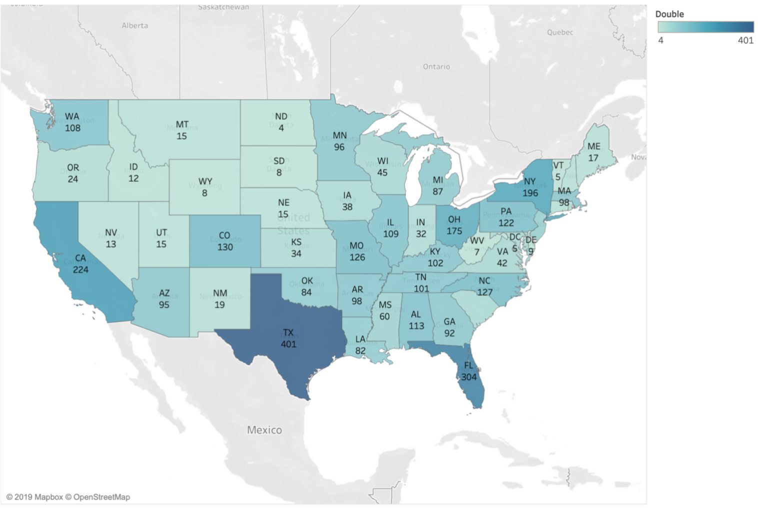
The graph of residence type shows that for all different kinds of residence type, more customers tend to concentrate on two retailers rather than a single retailer. Most of the loyal customers live in one family house, some live in three family house and trailer.

The graph of income (below) shows that ricker people tend to concentrate their shopping on a single or two retailers.





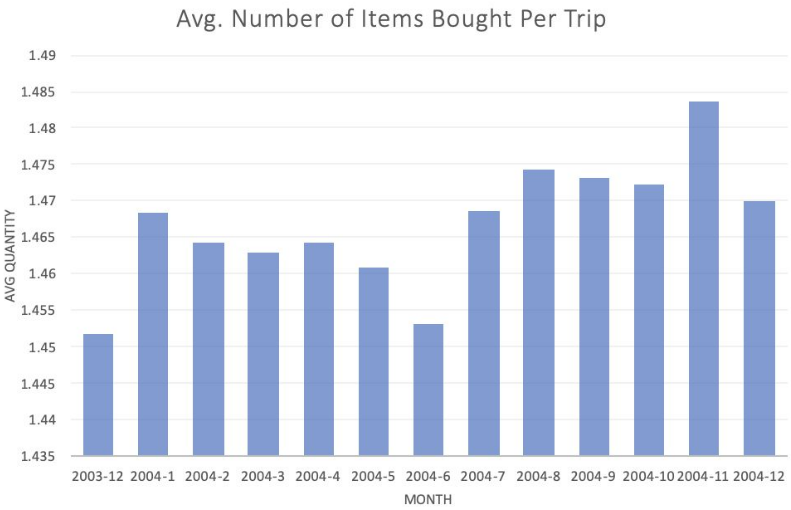
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Single. The marks are labeled by F1 and sum of Single. Details are shown for F1. The view is filtered on F1, which keeps 49 of 49 members.



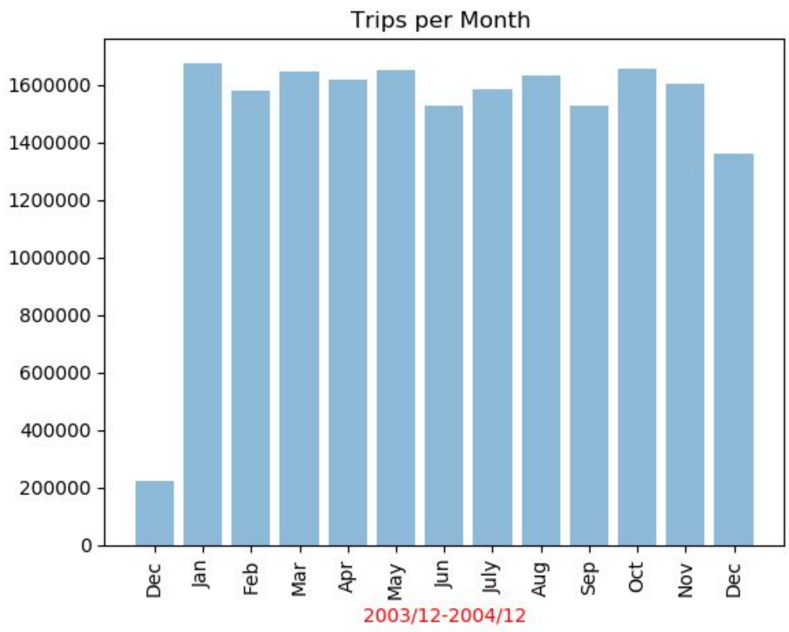
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Double. The marks are labeled by F1 and sum of Double. Details are shown for F1. The view is filtered on F1, which keeps 49 of 49 members.

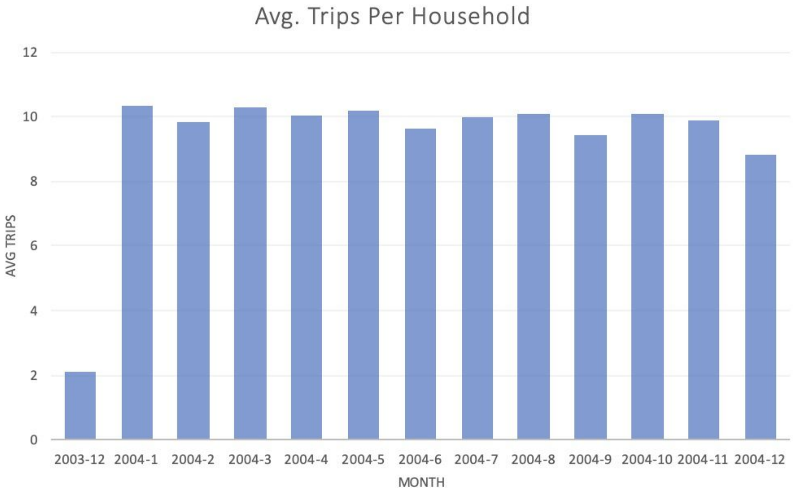
From the two maps above we can see, most of the loyal customers who concentrate on either single or two retailers live in TX, CA and FL.

Then we plot with several distributions below trying to find customers' shopping patterns:

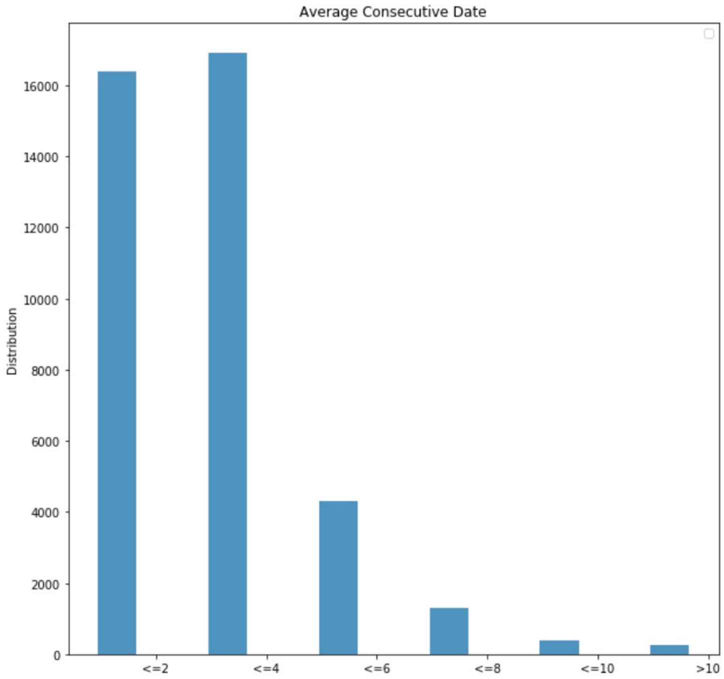


First, we plot the average number of items purchased per trip in each month, we found that people shop less in December 2003 and June 2004. Customers clearly tend to buy more items per trip in November since they may be preparing for the coming holidays.





Secondly, we plot the number of total shopping trips per month and average trips per household per month. It shows that there are significantly less number of trips in December 2003. We went back and checked the database, the database have only started recording trips since December 28, 2003, and this is why the trips number is significantly lower than other months. Disregard this outlier, number of trips hold even throughout the year, except that there are fewer trips in December 2004.

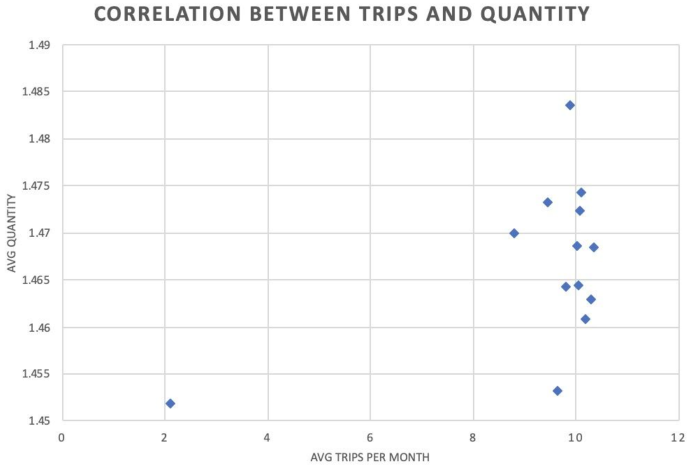


Thirdly, we try to find out the average number of days between 2 consecutive shopping trips, and plot the graph above. It shows that the most common gap between consecutive trips are between 2 days and 4 days, the secondary most common gap is less than or equal to 2 days. This tells us that most households shop frequently -- at least once or twice a week.

Part C:

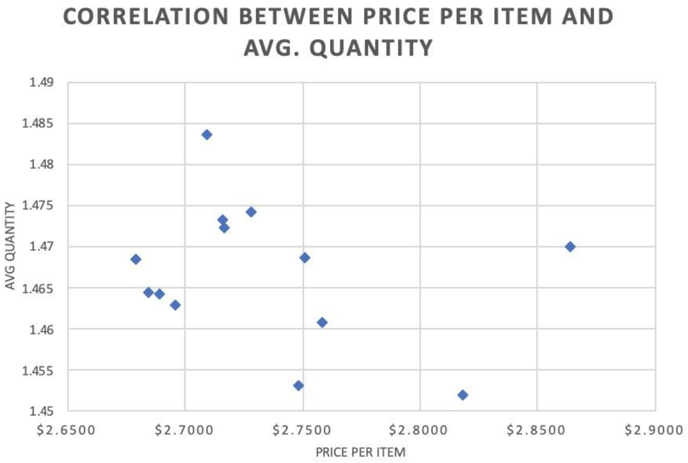
For part c, we plot more visualizations and looked into the potential correlations between data variables.

First we explored the possible correlation between the shopping trips per month with the average number of items purchased by looking at the plot below:



Since the average trips per month barely changes, there’s not really any correlation between trips per month and the average number of items purchased. After running a regression, we found the adjusted R-square is only 0.176, which means only 17.6% changes in average number of items purchased can be explained by trips per month.

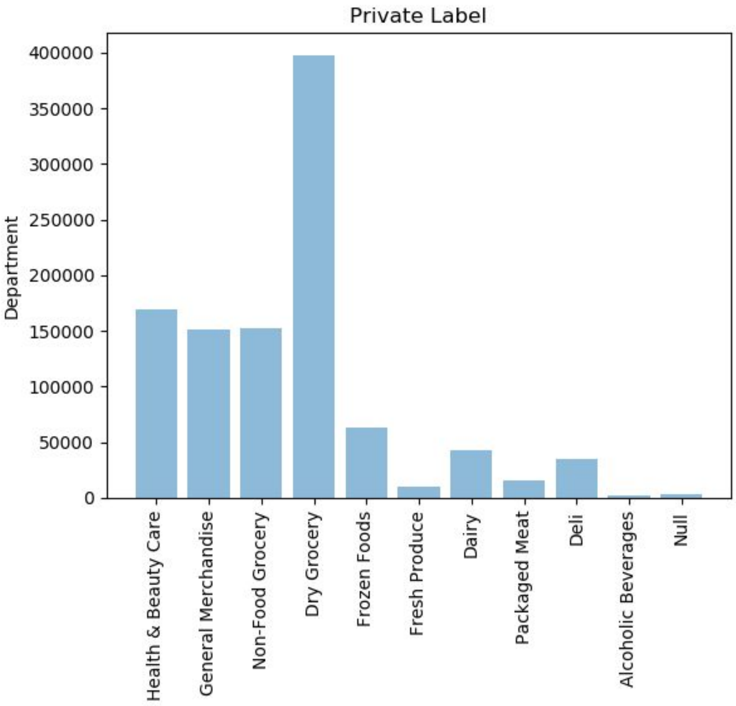
Next, we looked into the possible correlation between average price paid per item and the number of items purchased:



Again, we find that there's no clear correlation between these two variables. Regression test shows that the adjusted R-square is -0.008, which is very close to 0. This means there’s basically no relation between average price per item and number of items purchased.

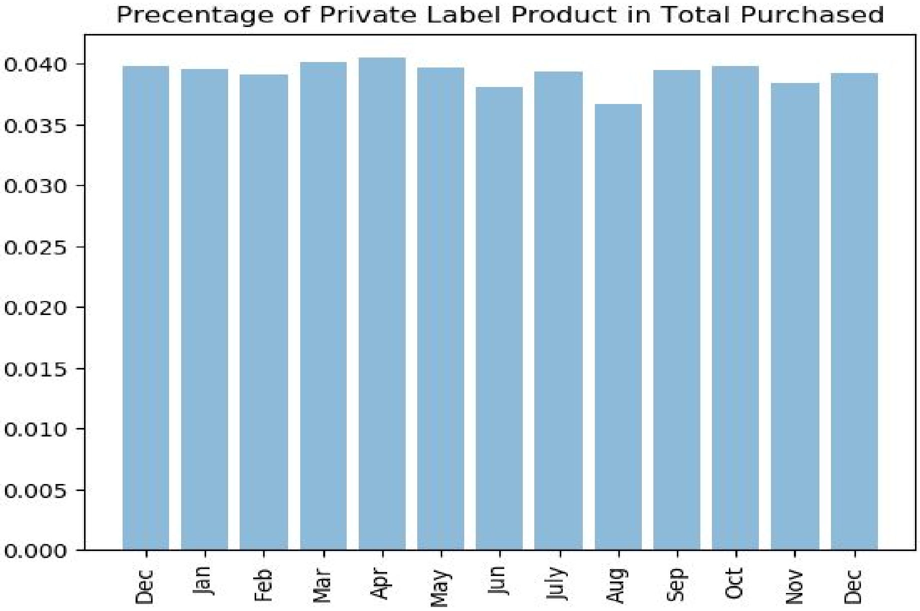
Then, we look into private labeled products in each category:

Department:	Number of Private Labeled Products
Health & Beauty Care	169,236
General Merchandise	151,670
Non-Food Grocery	152,747
Dry Grocery	397,527
Frozen Foods	63,623
Fresh Produce	10,250
Dairy	42,434
Packaged Meat	15,851
Deli	34,623
Alcoholic Beverages	2,600
Null	3,620

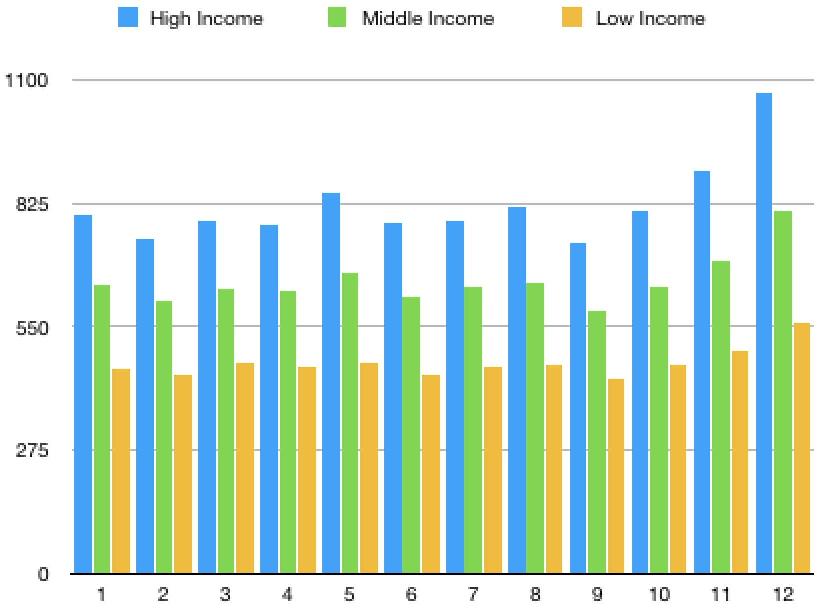


From the chart and graph above, we can easily see that Day Grocery category tends to be more private labelled -- it has over 350,000 private labeled products while other departments have less than 200,000.

Meanwhile, the expenditure share in Private Labeled products is not constant across months as the graph shown below:

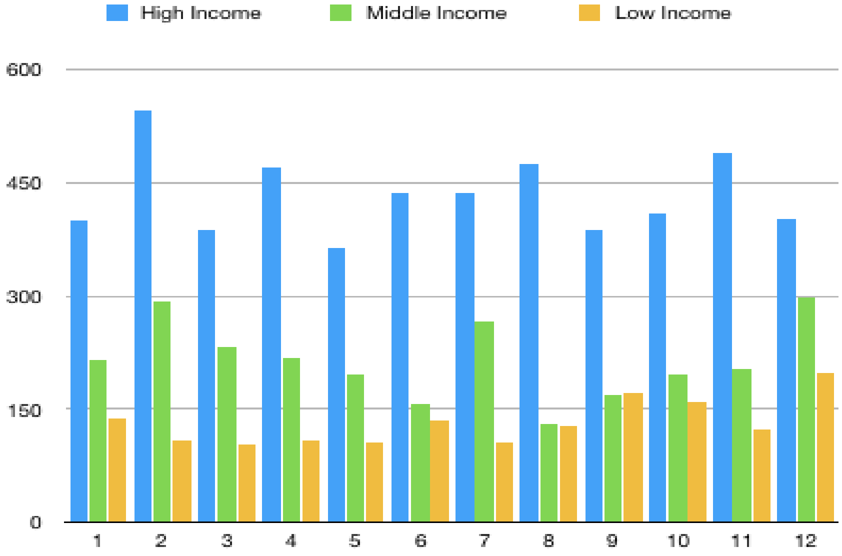


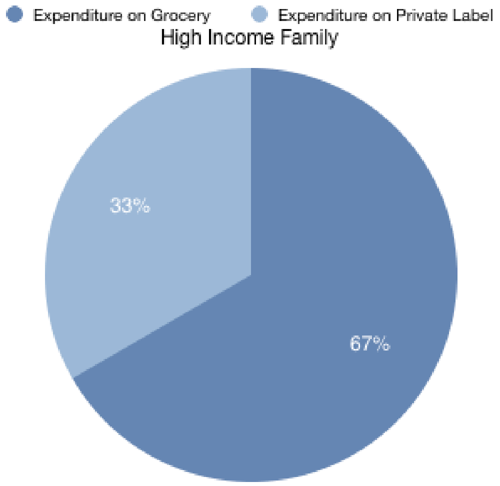
Based on the research, we seperate the households into three groups based on the annual income: lower than $35,000 per year as low income, $35,000 to $99,900 as middle income and $100,000 as High income.

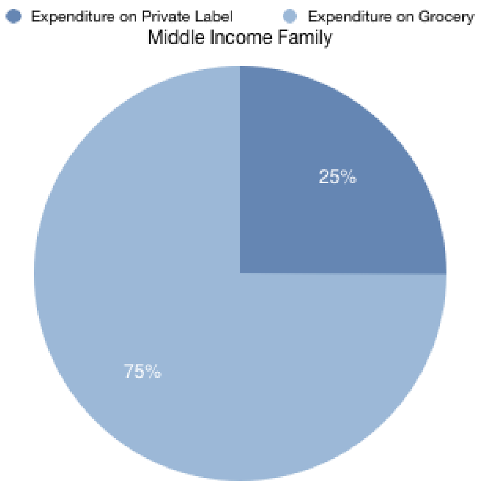


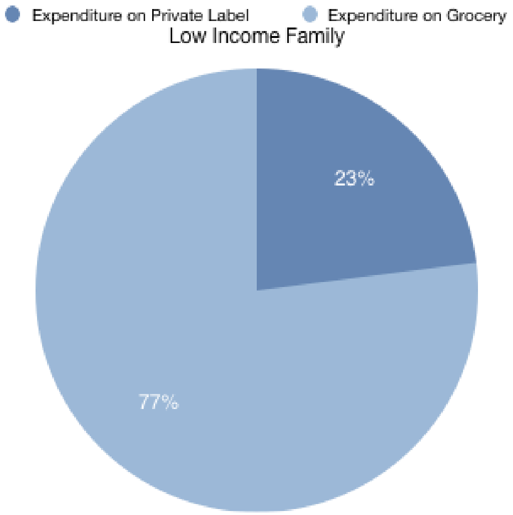
The high income family's monthly expense is between $737-1031 per household.
The middle income family's monthly expense is between $585-809 per household.
The low income family's monthly expense is between $431-556 per household.

As the charts show the high income family spent the most on private label products, on average 33% of their monthly expenses are spent on private labels compared with 25% for middle income family and 23% for Low income families.









Therefore, we suggest all the private label products should have more clear marketing towards high income family for boosting future sales.