Groceries shopping behavior project

Big Data Final Project

November 2019

ISSUED BY

Chuyue Wu

Part I Database Construction

First Step: Import Data Via Python

I use python to import data to MySQL, which could save a lot of time to import big data files. Packages of pymysql, pandas, sqlalchemy are used to connect python to db_consumer_panel database. The codes for importing data to MySQL through Python are listed in the file part1_database_construction.py.

Second Step: Adjust the Variable Format

The default variable types of column hh_id are different in the Table Households and Table Trips. For instance, hh_id is bigint (20) in the Households table, but is int (11) in the Trips table. In order to match Primary Key and Foreign Key successfully during the process of database construction, I change the variable hh_id in different tables to the same type, int (11).

Third Step: Create Primary Key and Foreign Key

I create three primary keys and foreign keys in python to improve the speed. The primary key hh_id, prod_id and TC_id in Households, Products and Trips are created through python using codes listed in the file part1_database_construction.py.

Part II Groceries data analysis

a. Data Overview

(1) How many store shopping trips are recorded in the database

By counting TC_id in the table Trips, there are 7,596,145 store shopping trips recorded in the database.

(2) How many households appear in the database

By counting hh_id in the table Households, there are 39,577 households in the database.

(3) How many stores of different retailers appear in the database

By counting distinct numbers of TC_retailer_code_store_code in the table Trips, there are 26,402 stores of different retailers appear in the database.

(4) How many different products are recorded

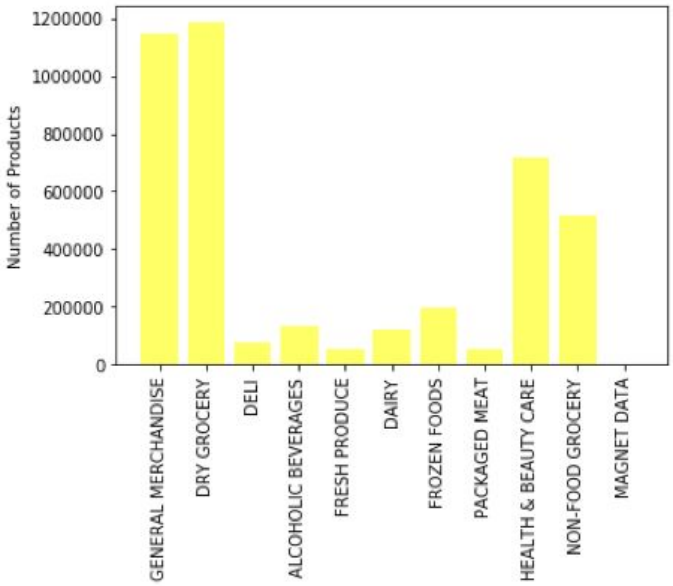
i. Products per category and products per module

By counting distinct numbers of group_of_prod_id in the table Product, there are 118 products per category; and by counting distinct numbers of module_of_prod_id in the table product, there are 1224 different products per module.

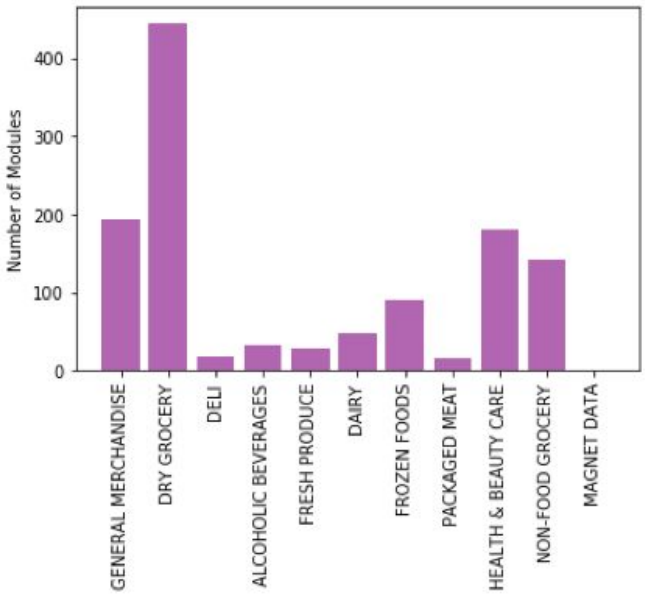
ii. Plot the distribution of products and modules per department

I used bar charts to display the distribution of products per department and modules per department.

Distribution of products per department



Distribution of modules per department



(5) Transactions

Total transactions and transactions realized under some kind of promotion.

By counting distinct numbers of deal_flag_TC_prod_id in the table Purchase, there are 38587942 different transactions appear in the database. Out of these transactions,

11384077 of them are under some kind of promotion (WHERE deal_flag_at_TC_prod_id =1).

b. Household Analysis

(1) Absent households

By counting distinct hh_id of TC_date greater than 92, I found that there are 6 households that do not shop at least once a 3 month period.

It is reasonable that 6 households do not shop at least once in a 3 month period. There are about 40,000 households in the database, this is only about 0.015% of the families, which means most households still go shopping at least every three month .

Most households go shopping at least every three month. As for the households that do not shop for longer than 90 days, this is occurring because some households would like to shop once for all the stuff they need in a long time, and some households may be travelling, or at their summer/winter place that's in other places. In addition, if the couple in the household both have traveling jobs, it can also result in no shopping in 3 month. Also, considering the development of e-commerce, there is a likelihood that people would rather shop online instead of a store.

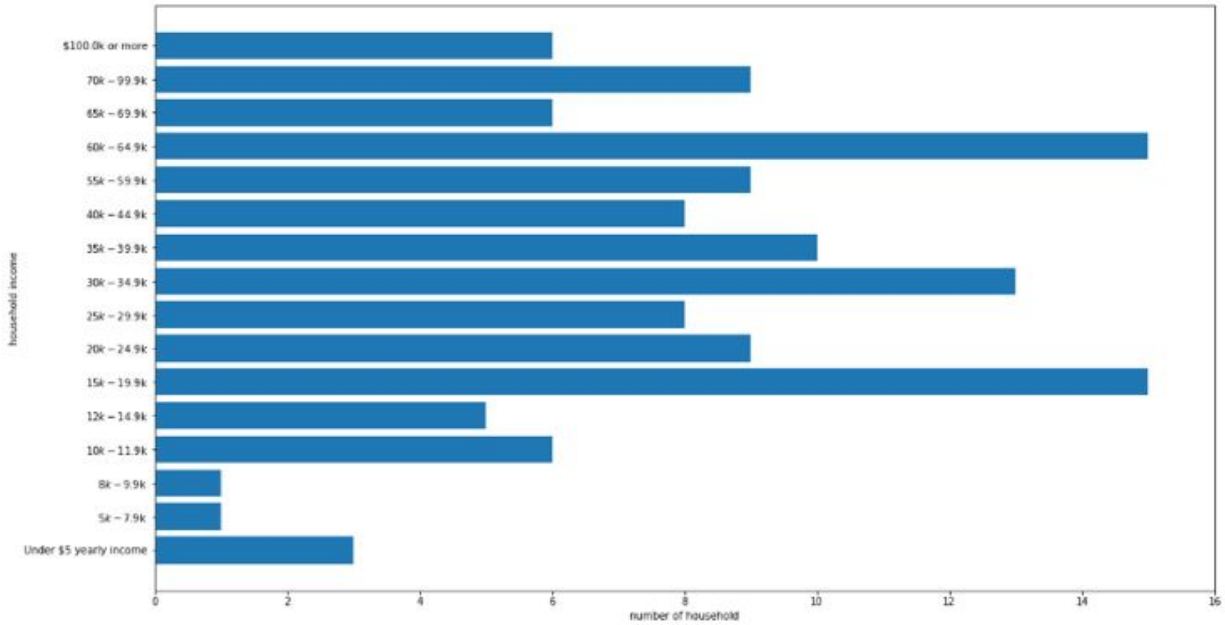
(2) Loyalism

0.33% of households (124 households among 37211 households that go shopping at least every month) concentrate at least 80% of their grocery expenditure (on average) on a single retailer. 0.44% of them (165 households among 37211 households) concentrate at least 80% of their grocery expenditure among 2 retailers.

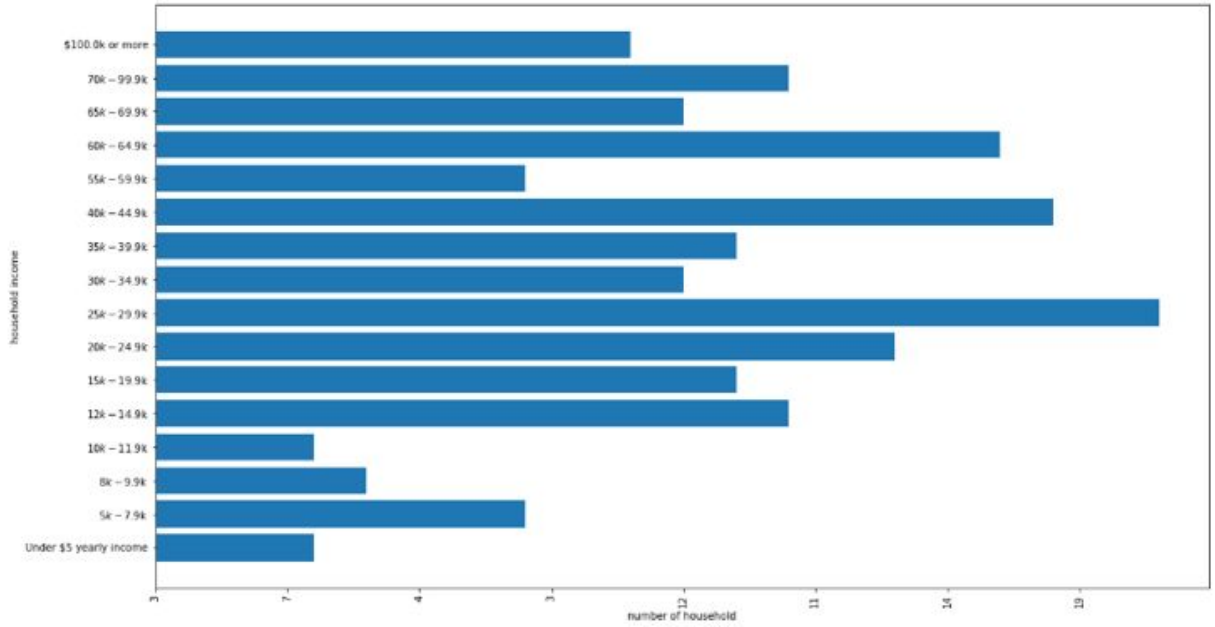
i. Are their demographics remarkably different? Are these people richer? Poorer?

Considering the graph above, the households concentrating at least 80% of their grocery expenditure (on average) on single retailers and the ones concentrating at least 80% of their grocery expenditure among 2 retailers are not remarkably different.

Income distribution for household making at least 80% of their grocery expenditure on a single retailer



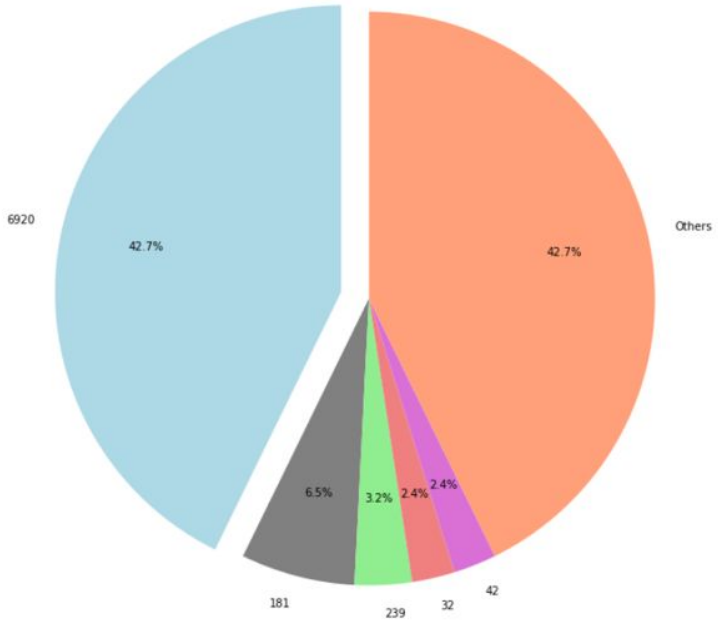
Income distribution for household making at least 80% of their grocery expenditure among two retailers



We can see from the plot that there is no significant difference based on income. Income below 20k is relevant low but that is because households with an income below that is relatively less nationwide.

ii. What is the retailer that has more loyalists?

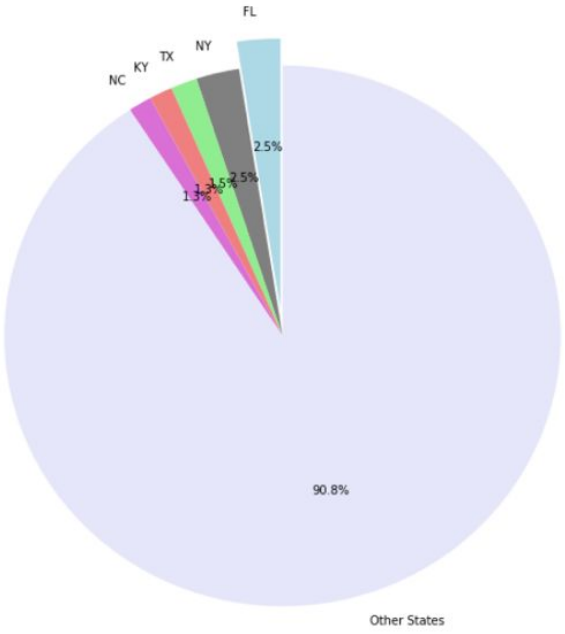
Five retailers with more loyalists



The five retailers with more loyalists are: 6920, 181, 239, 32 and 42 (In this project, we don't have the names of retailers. So we just use ID to represent them).

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

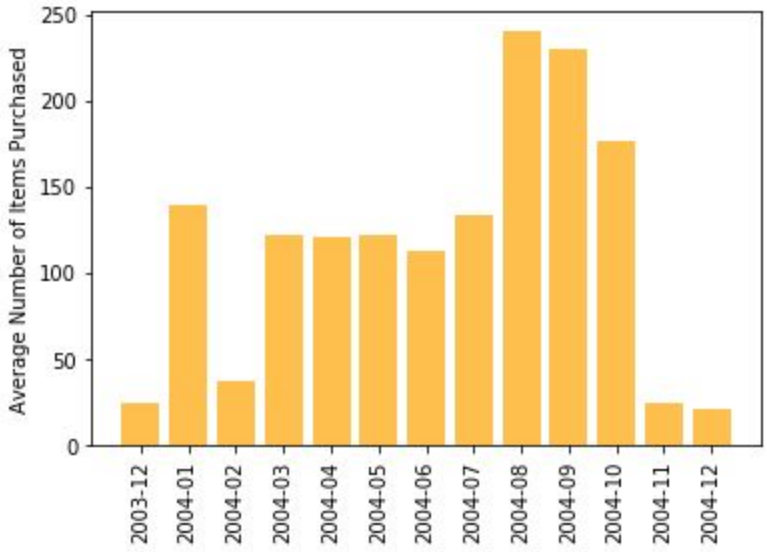
Household distribution around the country



Because there are too many states, I plot the states with more households that go shopping at least every month and group all other states into the 'Other States' group. Florida is the state that has most households. Overall, households are fairly evenly distributed around the country.

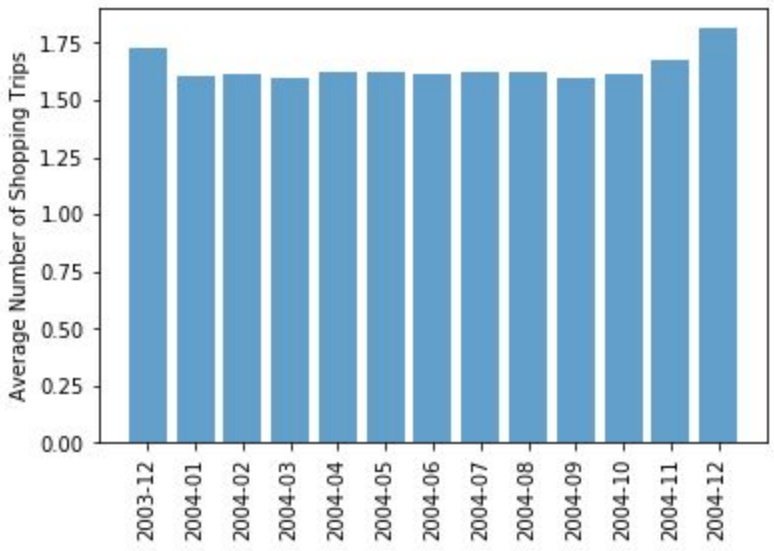
(3) Household shopping behavior distribution

Average number of items purchased in a given month



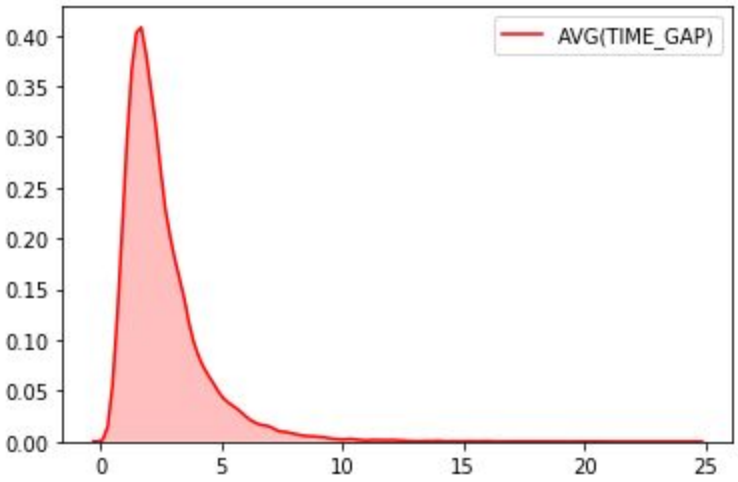
I use a bar chart to show the distribution of the average number of items purchased in each month from 2003-12 to 2004-12. We can see that during the winter months, people tend to buy less items when shopping except in January if 2014; people shopped the most items in August and September of 2014.

Average number of shopping trips per month



I use a bar chart to show the distribution of the average number of shopping trips in each month from 2003-12 to 2004-12. We can see that every month people go shopping pretty much equally frequently, about 1.75 times.

Average number of days between 2 consecutive shopping trips

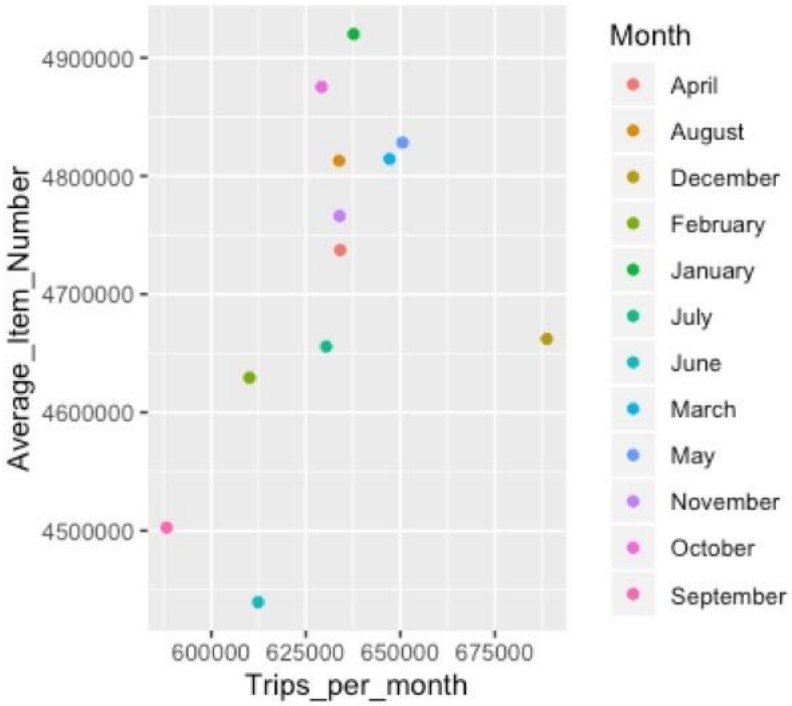


I use seaborn visualization to plot the average number of days between 2 consecutive shopping trips. We found that 40% of the household in our data take 2.5 days in between consecutive shopping trips, and usually that number is between 0-5 days.

c. Groceries Analysis

(1) Is the average number of items each time correlated with the number of shopping trips per month?

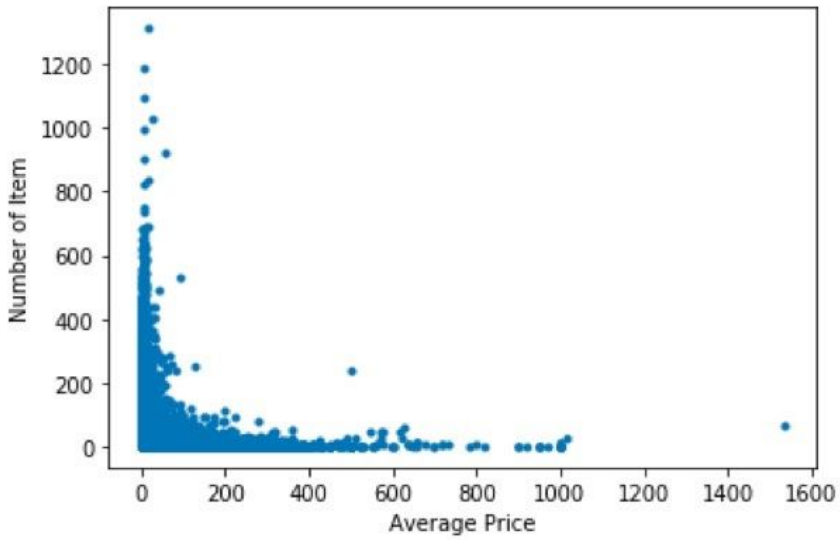
The relationship between number of items and the number of shopping trips



There is a positive linear relationship between the number of shopping trips per month and the average number of items purchased. It is reasonable that the more times people trip, the more items they will purchase on average. However, there is an abnormal data point in December. This abnormal data point is caused by missing data in December. Therefore, we can omit this point in analysis.

(2) Is the number of items purchased correlated with the average price paid per item?

The relationship between number of items and average price

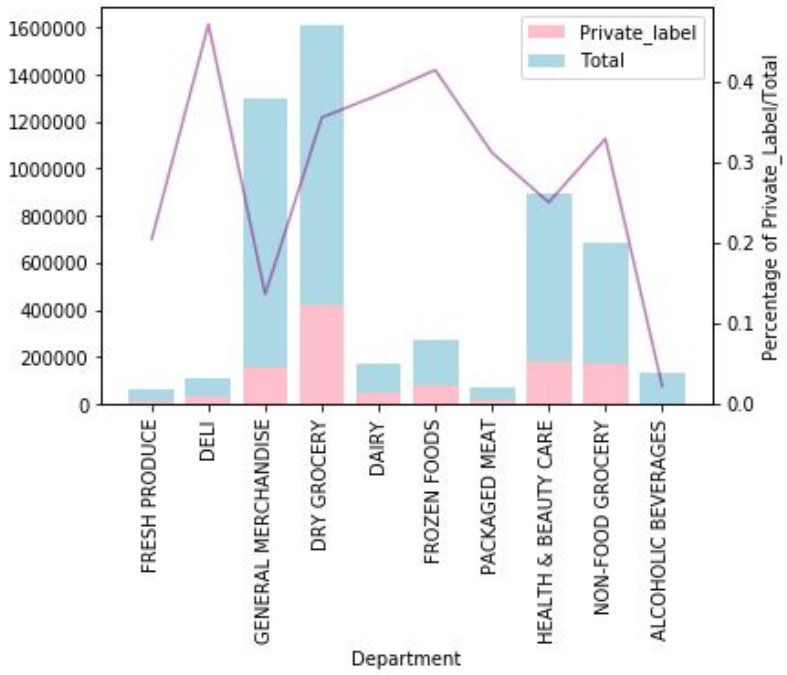


The relationship between number of items and average price is a curve that convex towards the origin. The number of items purchased decreases as average price increases. People tend to buy more items if the price ranges from 0 to 200 dollars. Therefore, the price range, $0 to $200, is the most acceptable price range for the majority.

(3) Private labeled products sale analysis

i.What are the product categories that have proven to be more “Private labelled”?

Private labelled product in each category

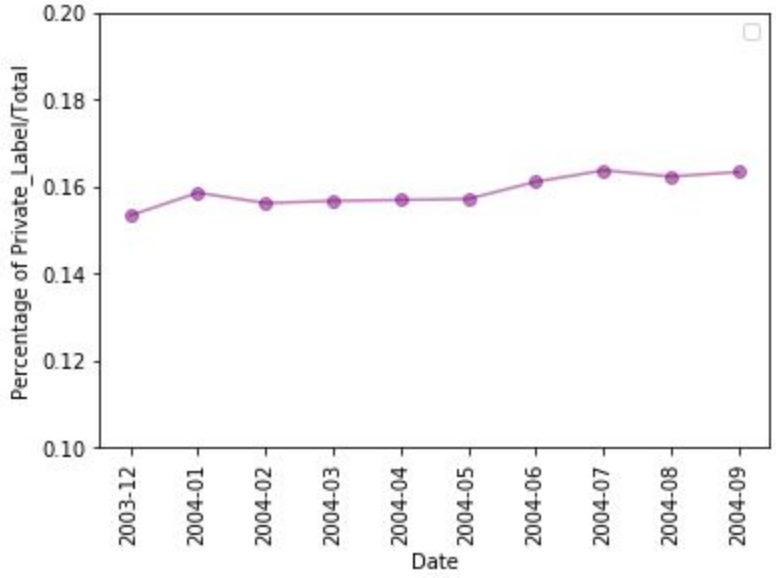


We can see from the picture above that DELI category has the highest percentage of "Private labelled". It is reasonable because most supermarkets provide their own deli. It's a good way to make food fresh. People can buy them immediately after the food has been produced by the supermarket. Otherwise, if deli food is produced by other suppliers, it costs time to ship it to supermarkets, which could make food not that fresh.

In addition, FROZEN FOODS, DAIRY and DRY GROCERY also have a high percentage of "Private Labelled". Those categories are all foods, which is consistent with our explanations above.

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

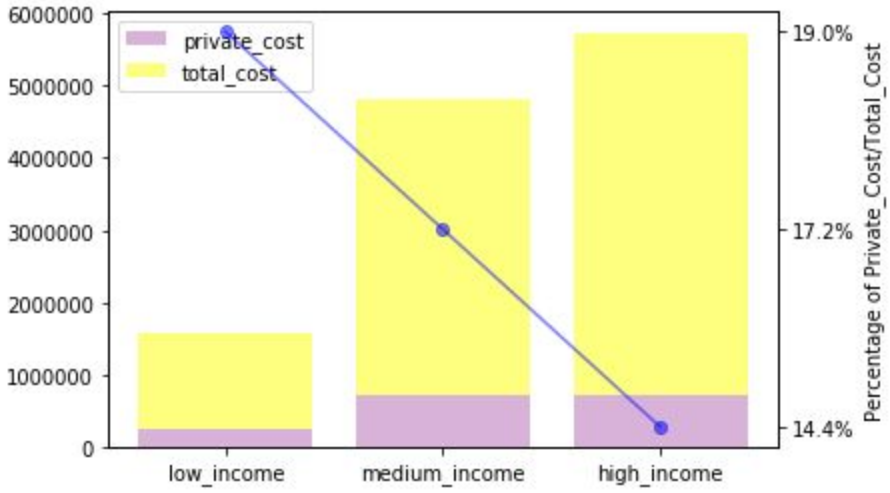
Percentage of the sales from private Labeled products by month



The expenditure share in Private Labeled products varies little across months. It ranges from 15% to 17%. So we can draw the conclusion that people buy private labeled products all the time and there is no strong relationship between months and labeled products purchasing behavior.

iii. Cluster households in three income groups, Low, Medium and High. Report the average monthly expenditure on grocery.

Monthly expenditure on grocery for households in different income level



In this question, I assume that families whose income below 20k are low income groups, families whose income between 20k and 50k are medium income groups, and families whose income above 50k are high income groups.

Therefore, from the figure, we can tell that although low_income people spend lowest money on average, they are more likely to buy private labeled products. That makes sense. Because private labeled products are cheaper in general which could attract low income people. In addition, low income people are more likely to buy food in supermarkets instead of eating outside in restaurants. Private labeled products have many food related categories as we found in C3 (1), so they buy more private labeled products.