[Date]

Big Data I Final Project

—— Research on Customer Behavior

qingyue su
MSBA

BIG DATA I FINAL PROJECT

November 3rd, 2019

Part 1 - Introduction

We use five Python modules to:
1, create a database named “db_consumer_panel” in mysql server from Python;
2, import data from csv files into Python pandas dataframes;
3, populate the “db_consumer_panel” tables in mysql server by importing data from the pandas dataframes.

The five modules are: mysql.connector, csv, os, sqlalchemy and pandas.
In the beginning, we create two variables in Python. They are DB_NAME, which is a string and equals to the name of the database we would like to name, and TABLES, which is a dictionary. TABLES takes names of tables as keys and mysql queries to create variables in one table as values.

Then we enable connect and connect functionality from mysql.connector to connect to mysql server from Python. After that, we define a function create_database which takes cursor as variable to create a database named using the variable “DB_NAME.” Then we run a for loop over TABLES, and in each loop, cursor.execute function takes keys from TABLES to create tables in mysql server.

After that, we import data from csv files to Python and store them as pandas dataframes by using pandas.read_csv(file_path) function and drop unnecessary columns and duplicated data entries in the dataframes.

Finally, we use engine function from sqlalchemy module to create an engine to import data from pandas to mysql server. pandas.to_sql function takes name of the table in mysql database, a corresponding dataframe in pandas, con=engine, index=False to populate the table in mysql database using data from pandas. In addition, we define chunksize = 1000 for each importation so that the importation process can be faster.

Part 2 - Researches

A. Basic Description of Data Set

A1. Total Number of trips

There is a total number of 7596145 trips recorded in database.

A2. Total Number of households

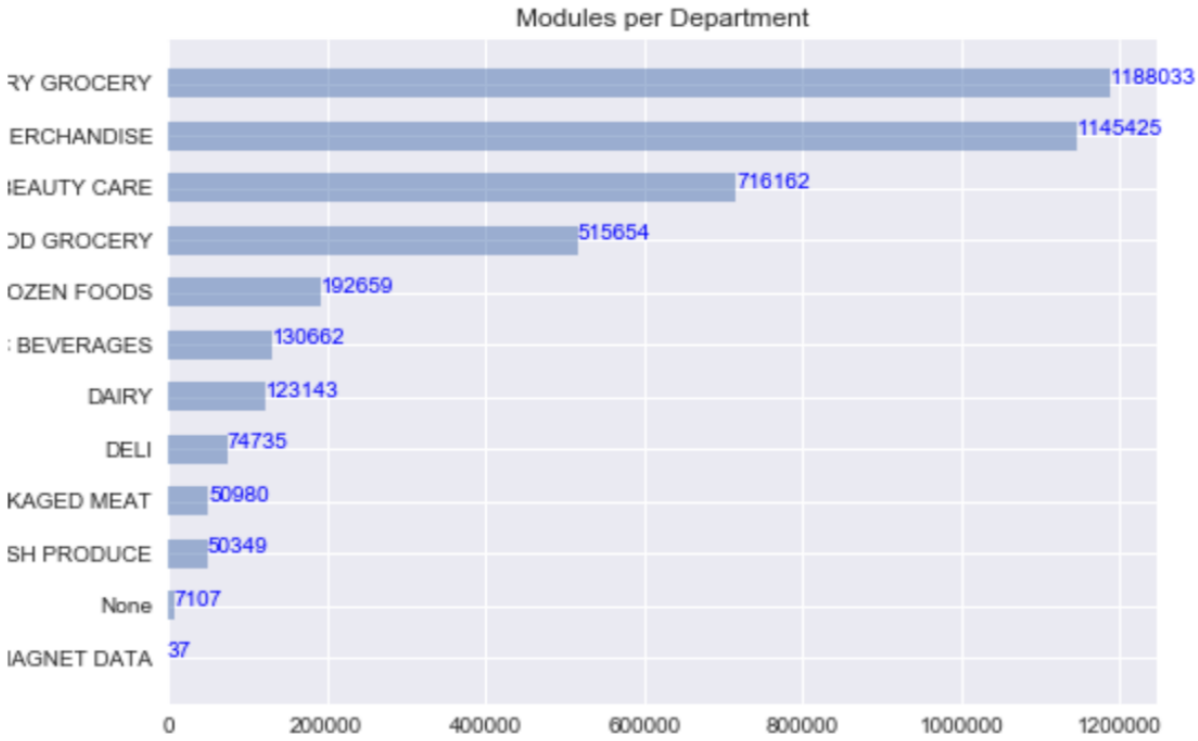
There is a total number of 39577 households recorded in database.

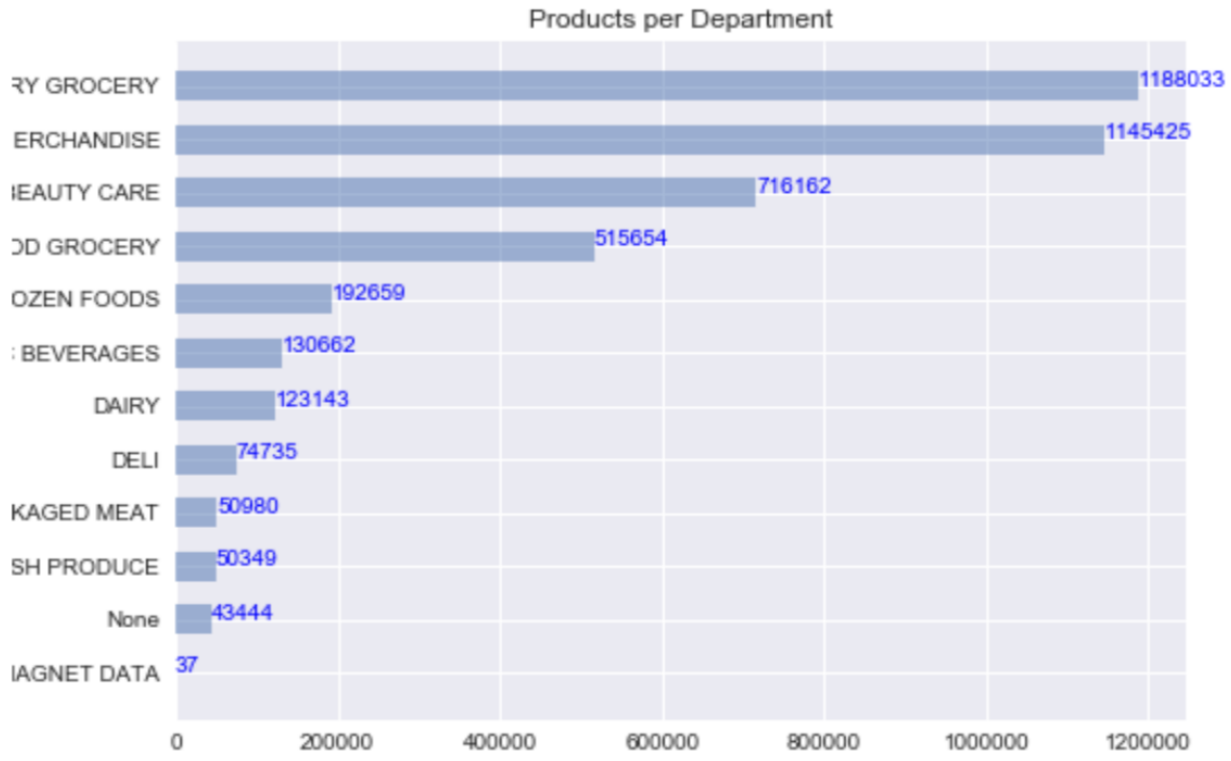
A3. Total Number of retailers

There is a total number of 863 different retailers in database.

A4. Main Modules / Products per Department

Please see tables in appendix: products per module; products per category. A4ii.





A5. Total Transactions with / without promotions

Total transactions without promotions: 35988296; total transactions with promotions: 2599646; total transactions: 38587942.

B. Deep Exploration of Data

B1. Number of households do not shop at least once on a 3 months period.

Assumptions:

Our team made two different assumption on this question: by 90 days or by monthly-level.

(1) First assumption- by 90 days:

We assume that 3 months are 90 days. There are 37 households do not shop once on a 3 months periods.

(2) Second assumption- by monthly-level:

We assume that the value of month extracted from the date represents the month level, no matter whether the actual interval between two dates are larger than 90 days or not, as long as the difference between the month level is no more than 3 months, it means that households do shopping again within three months. Based on this assumption, there are 6 households do not shop once on a 3 months periods.

Conclusion:

The reason that the answer would be significantly different is because if we use monthly-level to count day periods, sometimes we will have larger bias. For example: If there is a household shop on Jan20,2004 and

April28,2004, the period between two date is more than 90 days but no more than 3 months. As a result, the result will be different depended on assumption.

*Is it reasonable? **Why do you think this is occurring?***

The result is reasonable because of the following reasons:
First, as online shopping becomes more popular, some busy people tend to rely on online shopping, they may not go to retailer stores really often. Online shopping will not be included in this dataset, as a result, it will result in bias.
Second, some households may go out somewhere to travel or work for a long time. During the time, they will not shop at retailer stores. For example, if most of the households living in this community are students. During the summer, students may back to their countries or home. As a result, they will not shop at the retailer stores around this area.

B2. 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?

Assumption:

Our team made two different assumption on this question: by 30 days or by monthly-level.

(1) First Assumption: by 30 days: 971/39577=0.2%

    We assume that a month is 30 days and that 80% of their grocery expenditure means their 80% total spent money during the year at retailers. There are 0.2% of households spend their total expense on single retailer.

(2) Second Assumption: by monthly-level
    971/39577=0.2%

Conclusion:

Based on the second assumption, we can get the number of households who shop at least once a month is 39577, which is the same as the result according to the first assumption. The reason why we get the same answer is not a coincidence, but is the logic that as long as one household who shopped once in any month during the whole period, this household will be counted in. So, no matter whether we accept the first assumption or the second one, we will get the same answer.
Besides, under the second assumption, we also get the same number of the households who as well have at least 80% of their grocery expenditure on single retailer. That is also not just by accident, but because we measure their expenditure in a gross way, which means we calculate this value on the scale of household, but not household-monthly level. This is a more scientific approach to this question, because if you calculate

based on the household-monthly level, as long as one household spent once at least 80% of their expenditure on single retailer, it will be counted in, which makes no sense. Therefore, the answers to these questions have no difference under these two assumptions, and we just discuss it without considering the problem of assumption.

In conclusion, no matter which assumption we take, we can get the outcome that among all the households who shopped at least once within 30 days, only 971 households spend 80% of their money at single retailer during the data year.

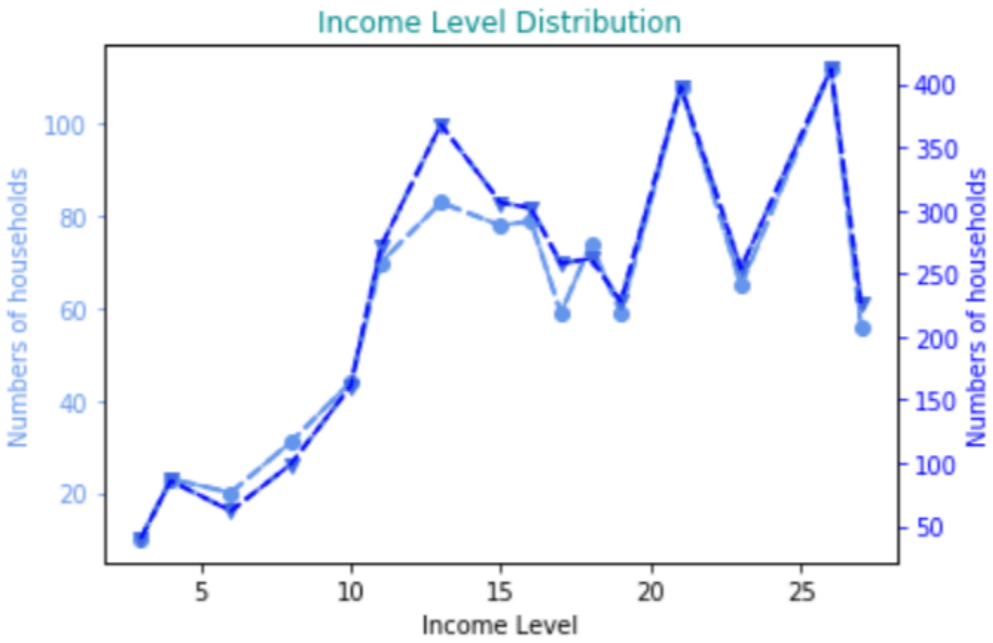
And among 2 retailers?

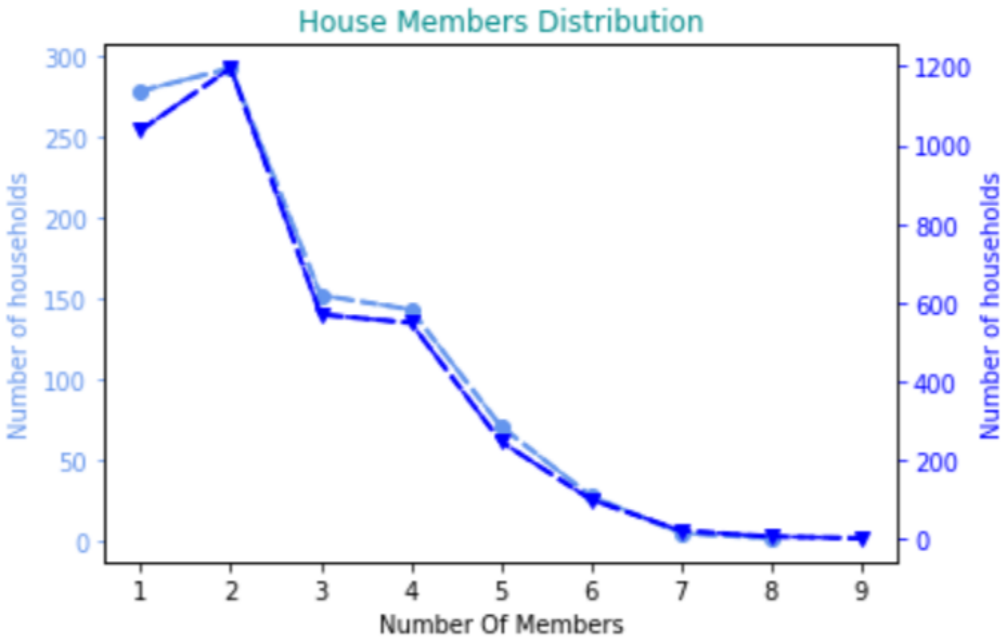
3733/39577=9.4%

Assumption:

We assume that 80% total spent money at among two retailers means that households spend 80% of their spent money on two retailers. As a result, we rank the total spent money of each households and order by each retailer. Afterward, we choose the first two retailers that each household spend their money on.

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





As we can see from the above graphs, the Income Level Distribution and Houses Members Distribution of households spend 80% of their total spent money at single retailer and among two retailers are similar to each other.

The left graph shows Income Level Distribution, most of family in two groups are in middle to high income level. As a result, we can assume that because the samples of data should meet specific requirement when they were collected, these samples have similar income level no matter which retailers these households go.

The right graph shows House Members Distribution is also similar between two groups. Most of households have less than four members. The mode of Number of Members is two members. Moreover, more than half of households are less than two members. We assume that this community may be located in a city, in which most members of this community have jobs and are young people. Therefore, they haven't formed a family and are not ready to have a child yet.

Most importantly, the samples of households among two retailers include all the samples of single retailer, no wonder that distribution would be so similar to each other.

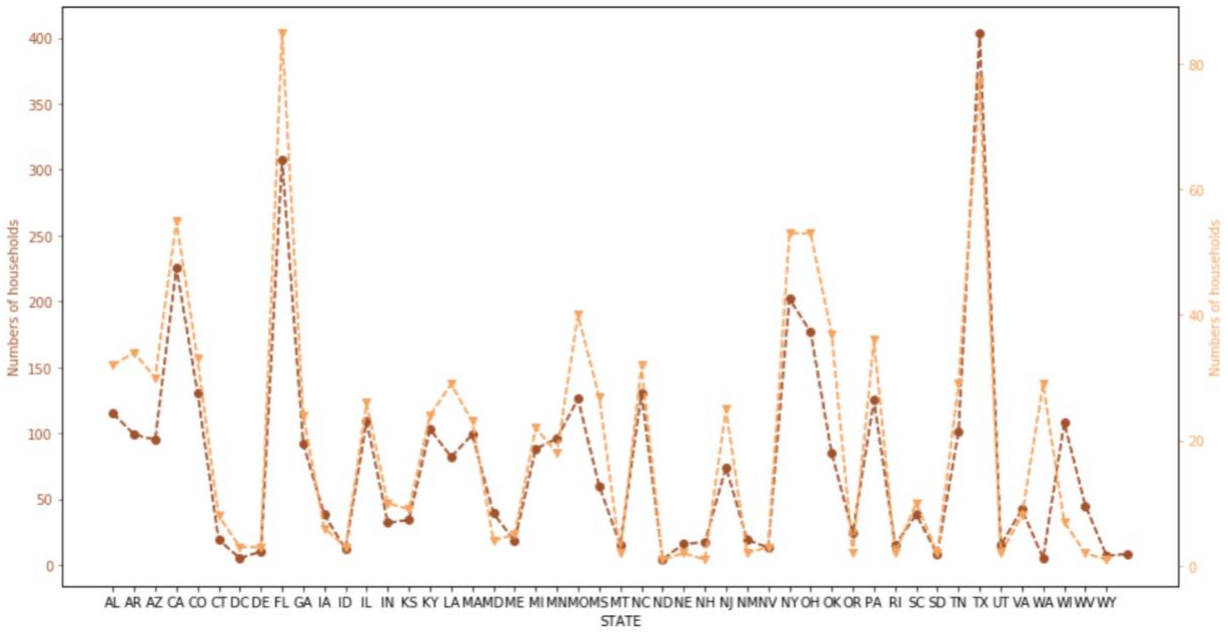
ii. What is the retailer that has more loyalists?

TC_retailer_code	count(H.hh_id)
6920	418
181	51
130	38
32	35
120	27
151	26
42	23
817	18
9103	17
239	15
294	14
111	13
9	12
248	12
6905	11
221	10
79	9
128	9
129	9
219	8
3999	8

TC_retailer_code	count(P.hh_id)
6920	10144
6905	5516
9101	3335
130	3135
9103	3118
6901	1891
181	1819
32	1747
9	1604
3997	1547
9999	1354
151	1265
9099	1115
42	1059
7003	955
248	930
4904	922
3999	908
79	843
817	807
294	759

As shown in above graph, the retailer 6920 has more loyalists.

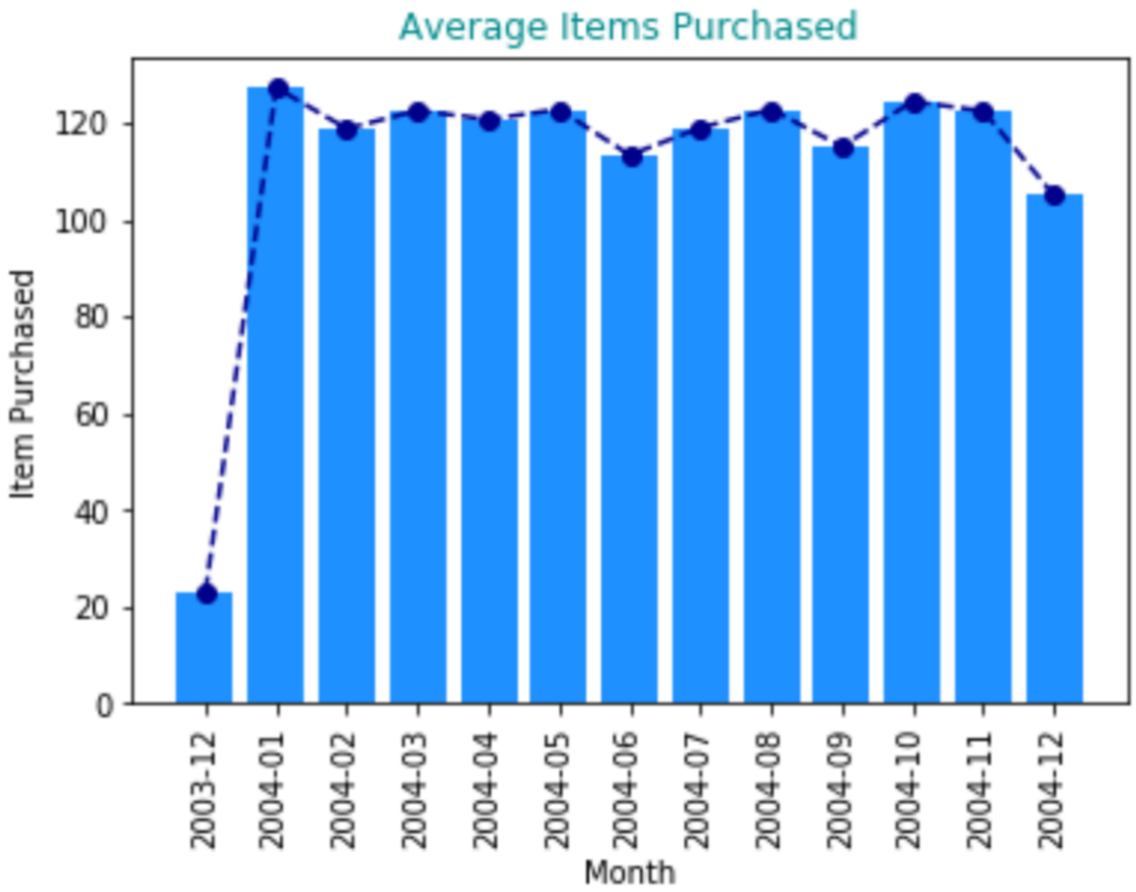
iii. Where do they live? Plot the distribution by state.



After making the line chart between the state and the number of households who put 80% of their grocery expenditure on single retailer and who put these money on two retailers, we can tell that people from Texas and Florida have the top two highest number of loyalist to the certain retailor, which are both on the south part of USA and distinctly higher than that of other states. However, if seeing it as a whole except these two states, we can tell that the overall number of households are around 100, which is an average level. On opposite, people from North Dakota, Washington and some states like these have the lowest loyalty to retailors, which are all northern states of America.

B3. Plot with the distribution:

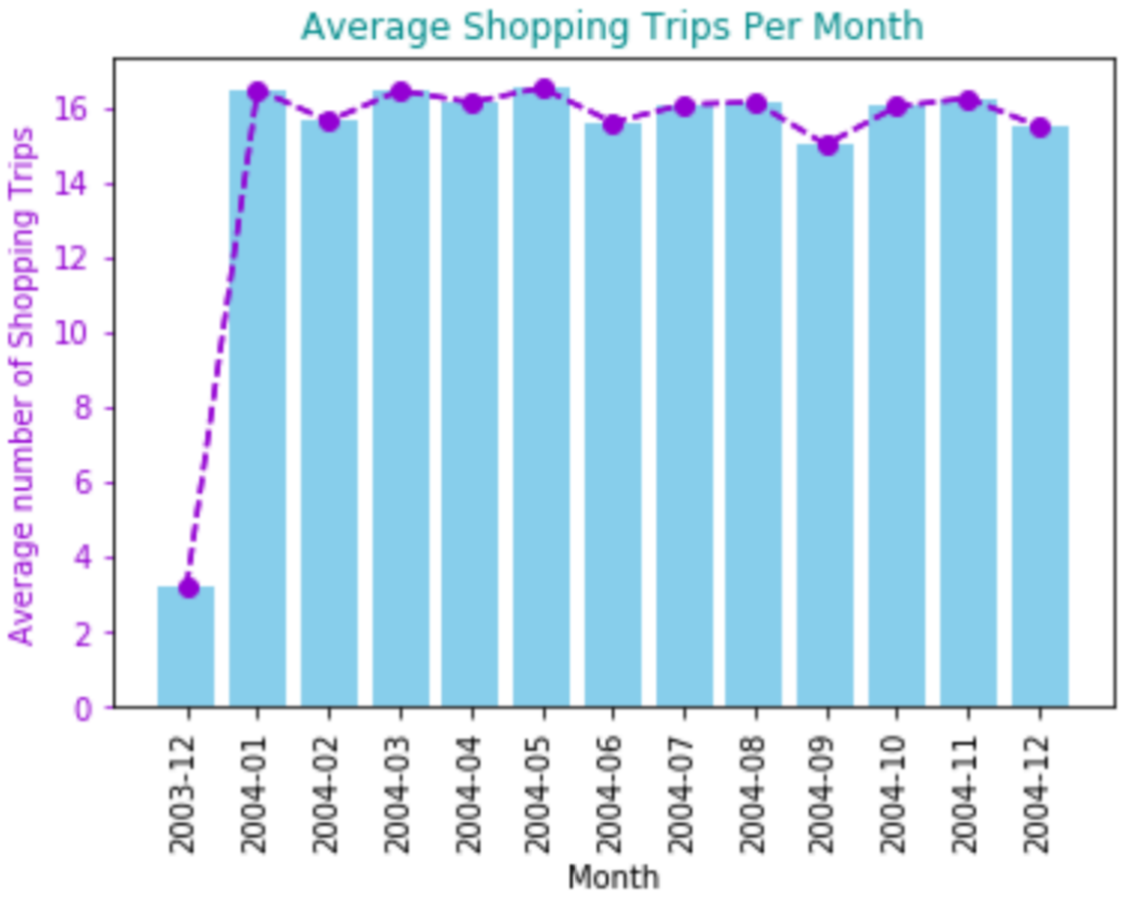
Average number of items purchased on a given month.



As 2013-12 only has few days, the Item Purchased will be much smaller than any other months. However, if we only see the statistics during 2014, the overall trend is stable around 115, which means that the average number of items purchased on a given month do not change too much.

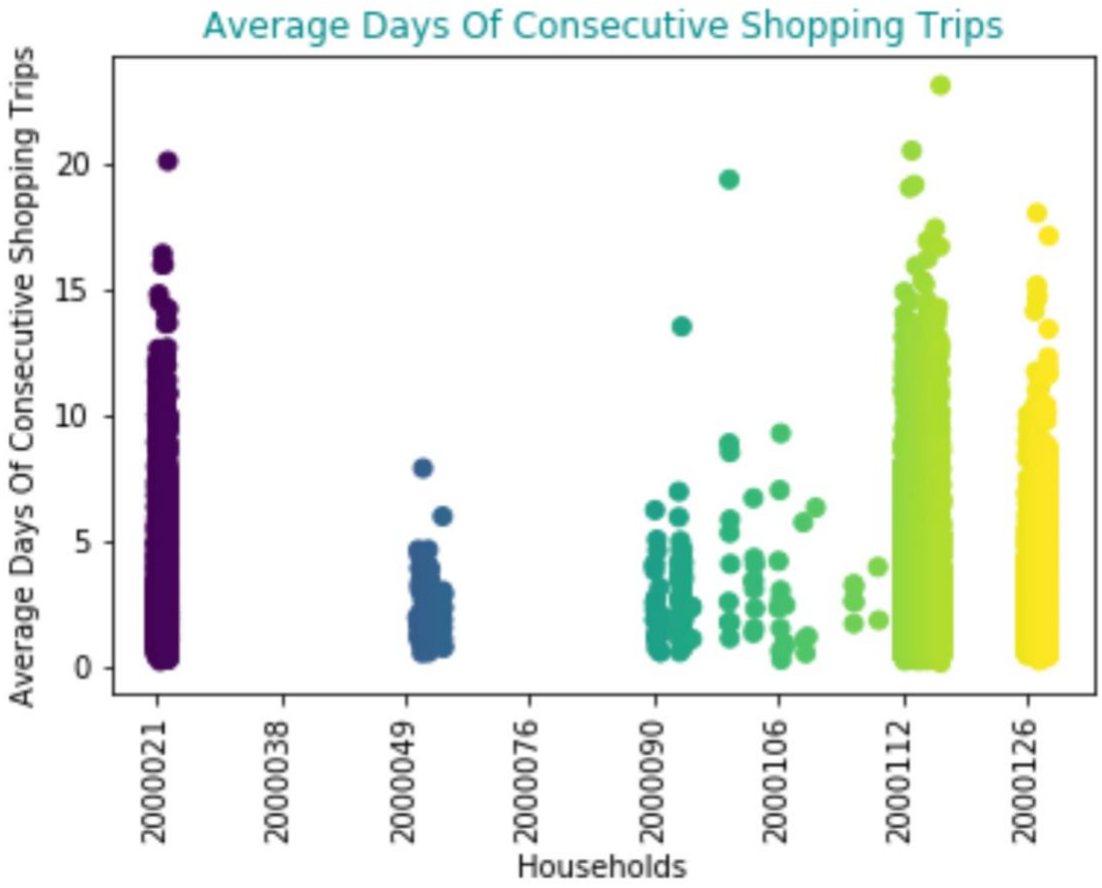
Average number of shopping trips per month.

As what said before, since 2013-12 only has few days, the average shopping trips of that month is much smaller than that of any other months. Besides, the average number of shopping trips are around 15 in the other 12 months, which means that people go to groceries for almost 3 or 4 times a week, which is satisfied as what we anticipated.



As what said before, since 2013-12 only has few days, the average shopping trips of that month is much smaller than that of any other months. Besides, the average number of shopping trips are around 15 in the other 12 months, which means that people go to groceries for almost 3 or 4 times a week, which is satisfied as what we anticipated.

Average number of days between 2 consecutive shopping trips.



As the graph above, the average days of consecutive shopping trips varies from zero to nearly 25 among these households, which means that almost every family will go to grocery at least once a month, which makes sense. Besides, because the total number of households is very large, so many blank places in the graph above is some relative small number (close to zero) of the time window of those households.

C. Make Information Visualization

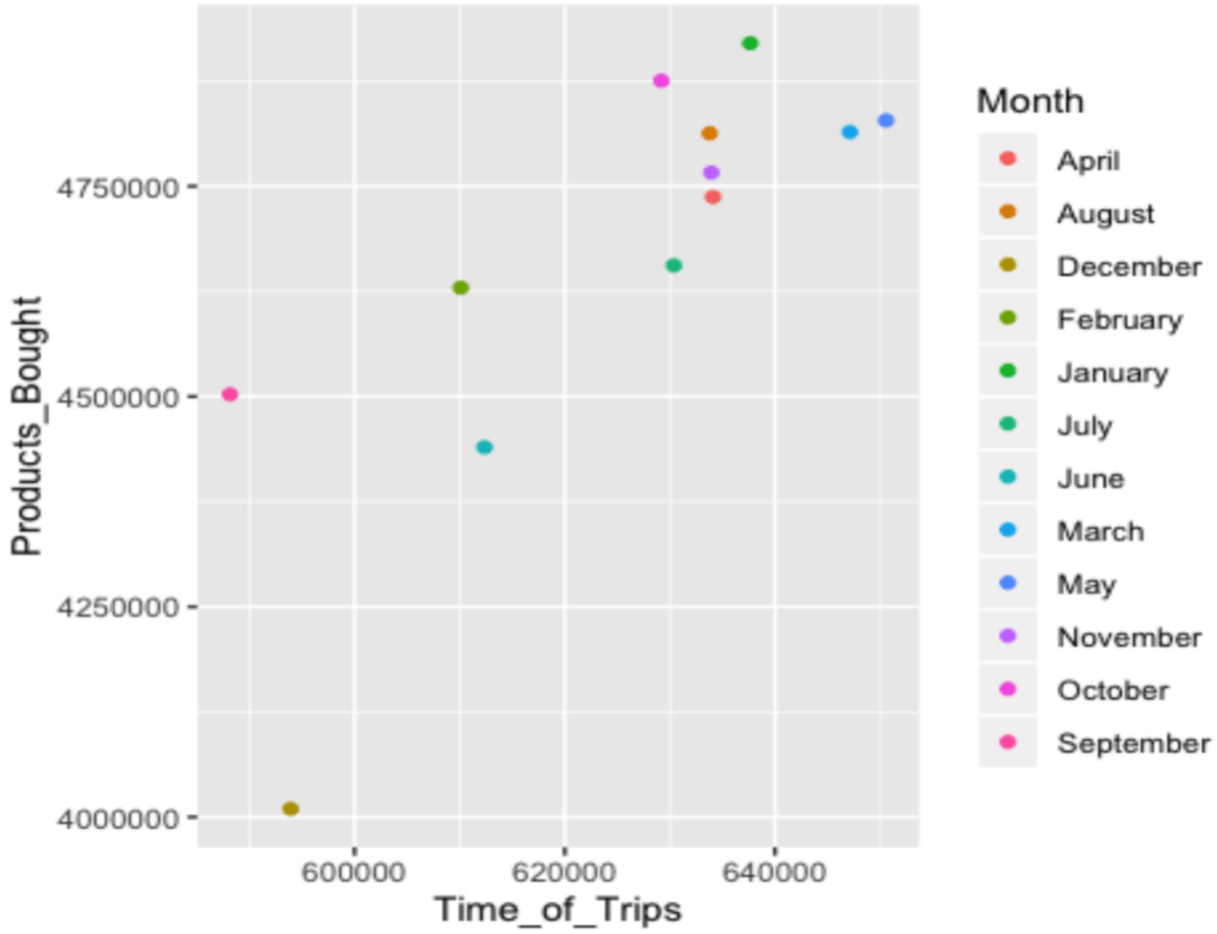
C1. Is the number of shopping trips per month correlated with the average number of items purchased?

Assumption:

The data sets of purchase and trips contain the data of last few days in December 2003. Because the dates in December 2003 is very few and have little influence, we decide to simply exclude those data from data sets and use whole 2004 year to solve this problem.

Process:

After clearing the data from 2003, we count the times of trips for each month and save it to table "month_trip_times". Then, we combine Trip and Product tables according using "TC_id" as key. Afterwards, we count the total number of purchases per month, and save the table as "month_prod_buy". Using another merge function, we are able to combine two table "month_trip_times" and "month_prod_buy" according to month. At last, we plot the graph that shows the relationship between number of shopping trips and average number of items purchased for each month.



Conclusion:

From the graph, we see that the average number of items purchased is positively correlated with the number of shopping trips per month. Also, the point of December seems low for both times of trips and products bought. The explanation for this situation is people don't like to go outside and buy stuffs.

C2. Is the average price paid per item correlated with the number of items purchased?

We divide total_price_paid_at_TC_Prod_id by quantity_at_TC_prod_id to get average price per product. We then calculate the correlation between average price and quantity using pandas.corr function. The correlation

is -0.091, which indicates there is a slightly negative relationship between the two, as shown in the diagram below.

C3. 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"

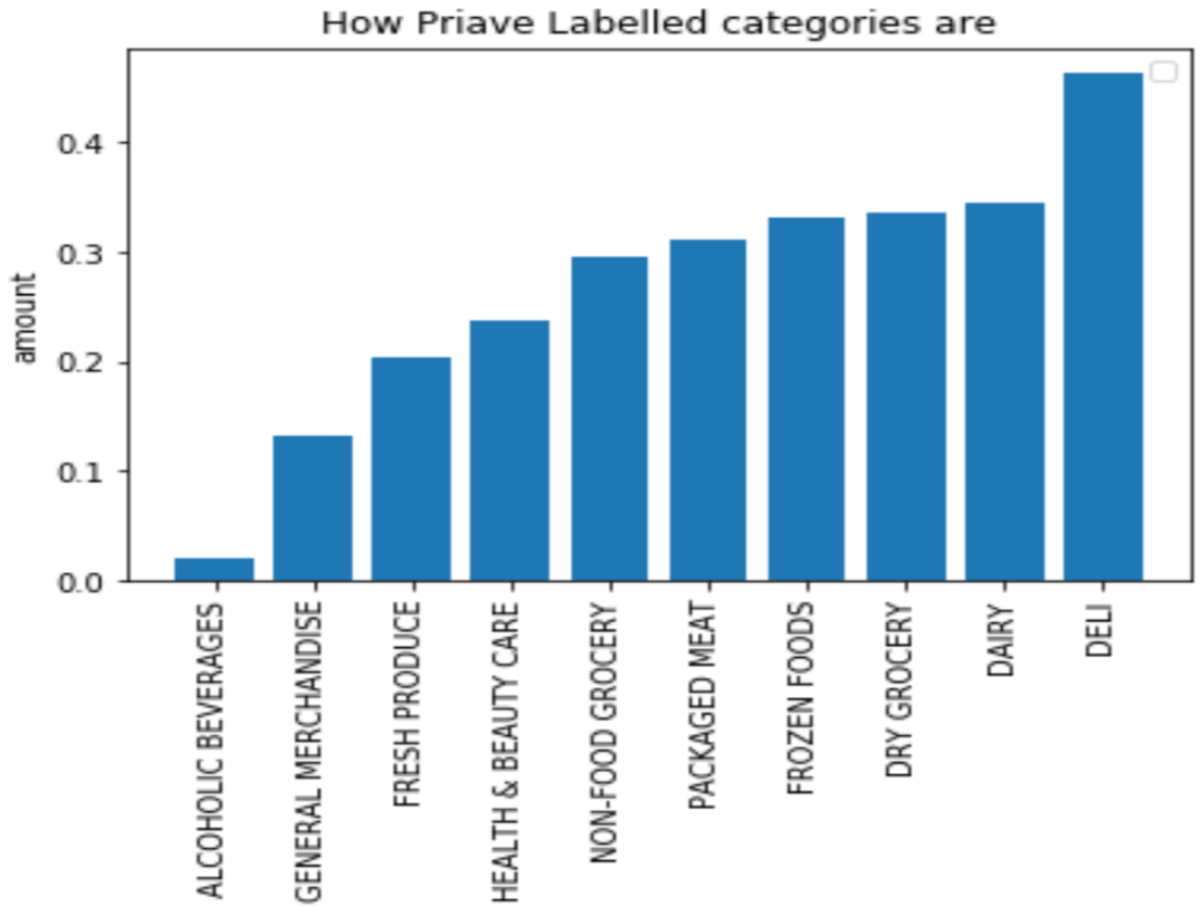
Private Labeled products are the products with the same brand as the supermarket. In the data set they appear labeled as 'CTL BR': What are the product categories that have proven to be more "Private labeled"?

Assumption:

We define 'more Private Labeled as: In one certain category, "the number of private labeled products/ the number of whole products in that category " is higher. So we calculate that portion of every category and compare them in bar graph.

Process:

First, we count number of products of each category, then we select products with 'Private Label', then count the number of 'CTL_BR' products of each category, then calculate the percentage of private labeled products, as shown below:



Conclusion:

From graph we can see that, 'DELI' Category is the most 'Private Labeled category, because it has the highest portion of private labeled products. While 'ALCOHOLIC BEVERAGES' Category is the least, since the percentage of private labeled products is the lowest.

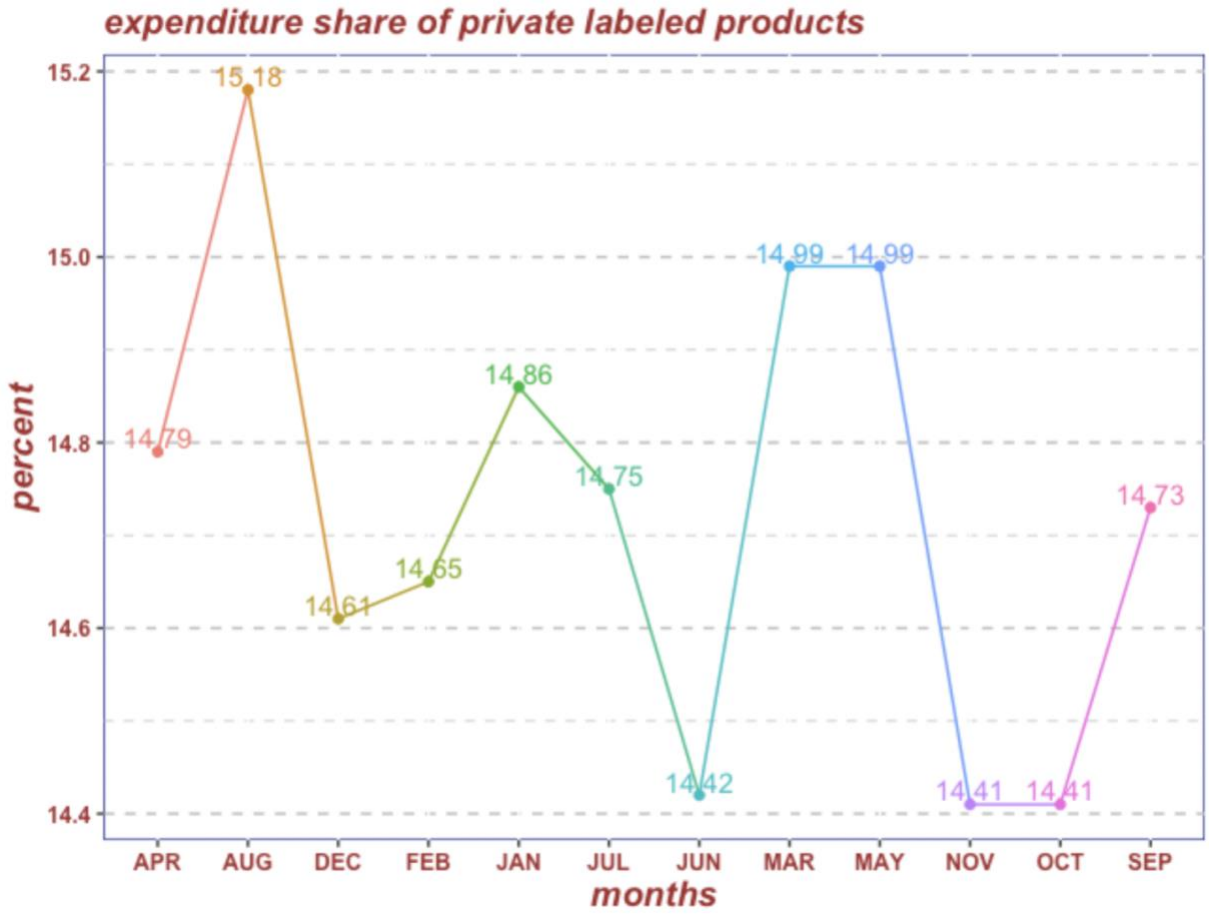
ii. Is the expenditure share in Private Labeled products constant across months?

Assumption:

In this case, we assume data from 2003-12 belong to 2004-12, because the point is to show changes in year 2004, and it's easier to see result by doing so.

Process:

First, we select products of each month and sum the 'total_price_paid_at_prod_id' of them group by months. Then we select 'Private labeled' products and follow the previous process. Use Monthly expenditure of CTL_BR products/ Total monthly expenditure of all products, then we get expenditure share(percentage) of CTL_BR products.



Conclusion:

graph shows that expenditure share of CTL_BR products fluctuates a little bit across months but basically stay constant at about 14 to 15 percent.

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.

We use sklearn.KMeans function to cluster income into three groups, and calculate the average monthly expenditure for the three groups: high income households spend roughly 735.06 per month for grocery; middle income 572.95 and low income 426.46.

High income households on average spend 6.47% on private labelled products, while middle income households spend 8.12% on private labelled products, and low income households 9.8%. The result makes

sense, since private labelled products tend to cheaper, and thus they are more welcomed by low income households. The diagram below confirms our intuition. The regression line indicates there is a slightly negative relationship between income level and percentage spent on private labelled products.