Big Data Project
Group Id:12
Yuan Tian, Jiajie Yuan, Peihan Tian, Salma Mohammed

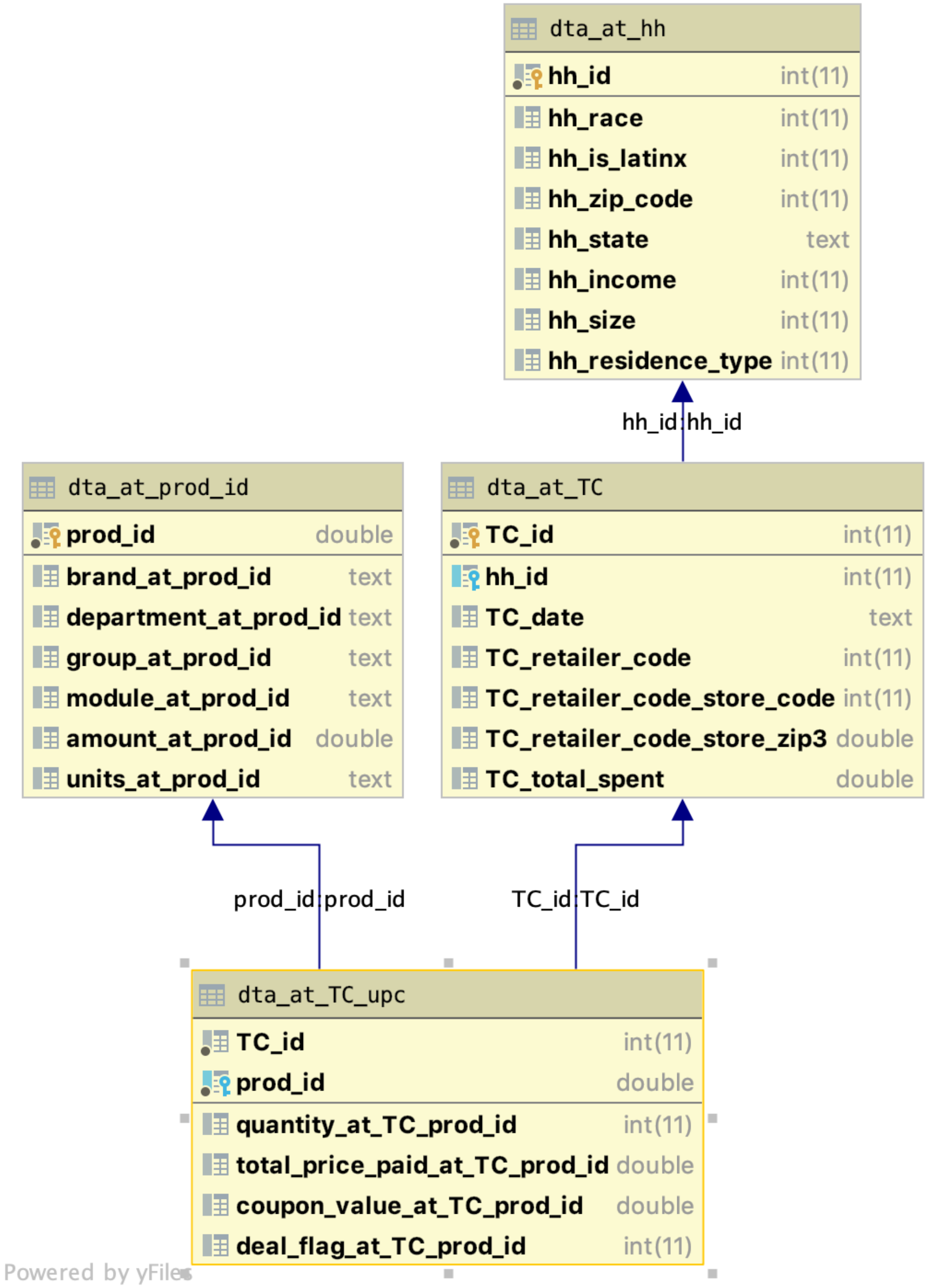
I. Instructions:

1) To create the database, we used three different methods:
   a) **Data-Grip:** we first imported Data-Grip, which is a software that allows building a connection to MySQL dashboard, but process information and commands much faster. We established the connection between the two platforms and then began writing our MYSQL code. The steps we followed to create this connection are: In the database tab, we created a new connection by clicking on the `+' sign. Then, we chose MySQL as our database.
   b) **Python:** we imported and processed the data fully in python.
   c) **MySQL workbench :** we can directly import dataset into MySQL using load infile,which is very fast and efficient. Firstly, we created the four tables.After that,we used load infile to import data into these four tables separately.

After importing the data in Data-Grip, we set primary keys and foriegn keys as follows:

Table	**Primary Key**	**Foriegn Key**
Households	hh_id	
Products	prod_id	
Trips	TC_id	
Purchases	The combination of (TC_id & prod_id)	

For our tables and build 1 to many, and many to many relationships as detailed in the database dictionary and in the Schema shown below.



dta_at_hh	
hh_id	int(11)
hh_race	int(11)
hh_is_latinx	int(11)
hh_zip_code	int(11)
hh_state	text
hh_income	int(11)
hh_size	int(11)
hh_residence_type	int(11)

dta_at_prod_id	
prod_id	double
brand_at_prod_id	text
department_at_prod_id	text
group_at_prod_id	text
module_at_prod_id	text
amount_at_prod_id	double
units_at_prod_id	text

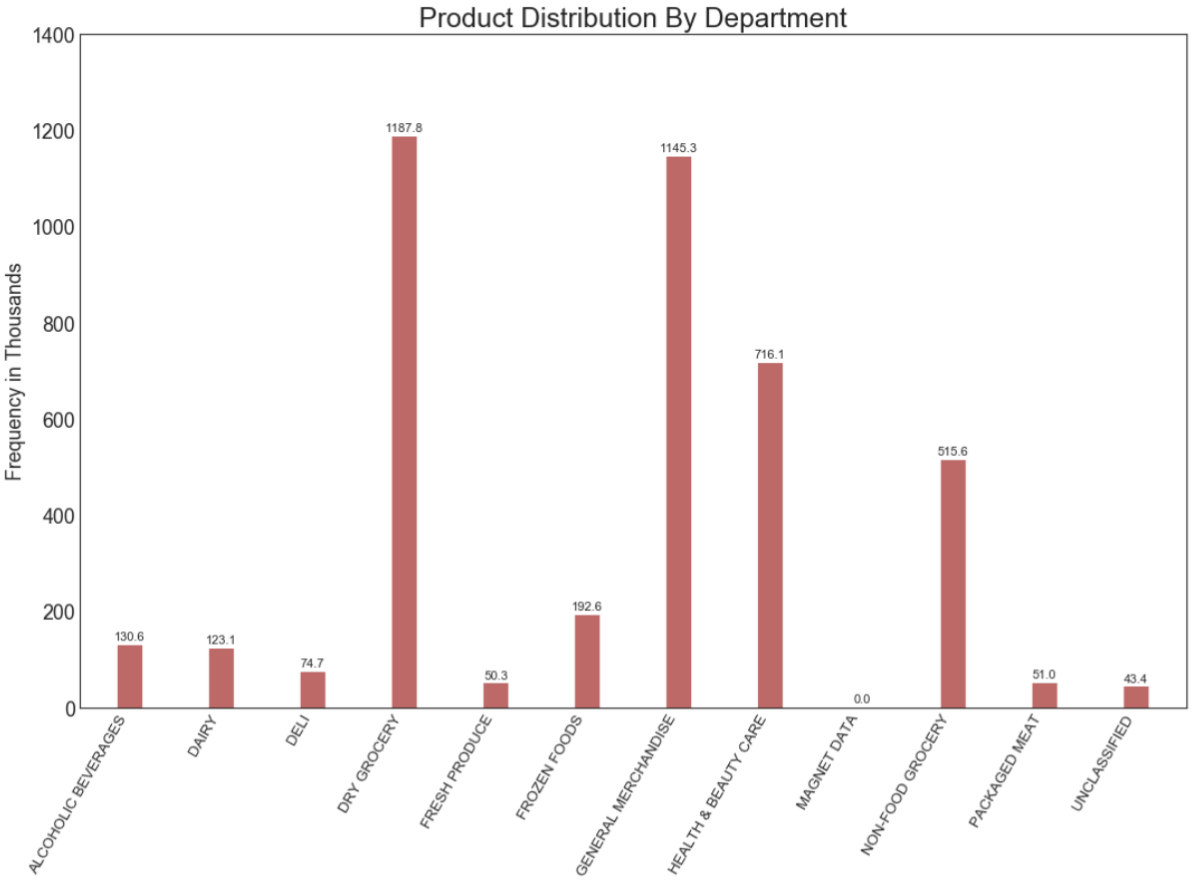
dta_at_TC	
TC_id	int(11)
hh_id	int(11)
TC_date	text
TC_retailer_code	int(11)
TC_retailer_code_store_code	int(11)
TC_retailer_code_store_zip3	double
TC_total_spent	double

dta_at_TC_upc	
TC_id	int(11)
prod_id	double
quantity_at_TC_prod_id	int(11)
total_price_paid_at_TC_prod_id	double
coupon_value_at_TC_prod_id	double
deal_flag_at_TC_prod_id	int(11)

2) After that, we were able to create the necessary queries to answer the questions.
3) Once we had the written queries, we exported those tables, imported them in Python, and using pandas and matplotlib, created the desired graphs.

II. Big Picture:

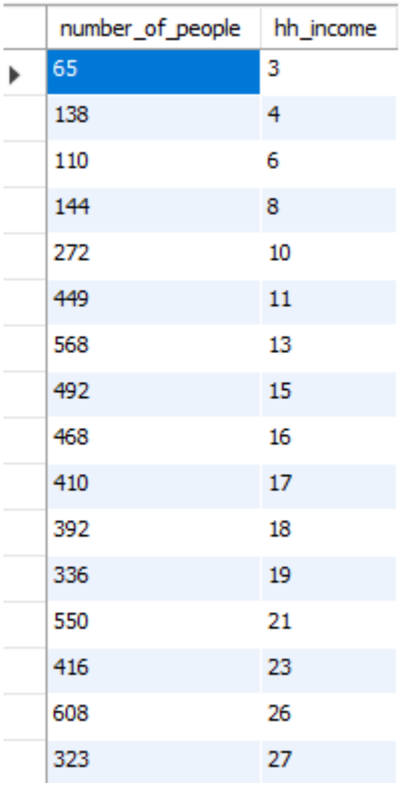
1) Number of store shopping trips recorded in the database: *7596145*
2) Number of households in the database: *39577*
3) Number of stores of different retailers that appear in the database: *863*
4) Number of different products recorded: *4230759*



5) Number of Transactions:
    a) Total transactions: *5,651,255*
    b) Total transactions realized under some kind of promotion: *2,670,312*

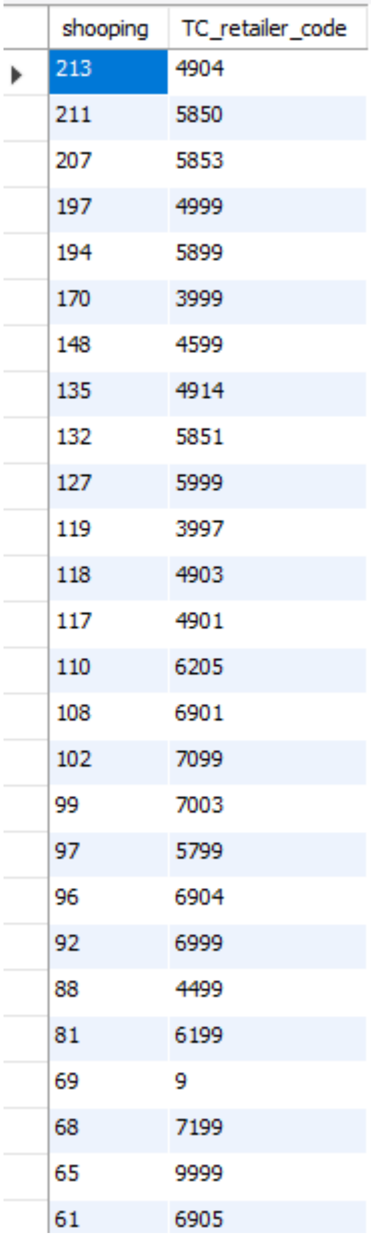
III. Household-Monthly Level Data:

1) Number of households that do not shop at least once on a 3 months periods: 84
    a) Is it reasonable?
        i) The fact that the number is small is reasonable. However, it is unreasonable that households spend three months without shopping given the need to get produce, and other life necessities at least once a month.
    b) Why do you think this is occuring?
        i) Our guess is that this number is the result of missing data or inaccurately recorded responses.
2) **Loyalism:** Among the households who shop at least once a month (32,953), the % of them which spends at least 80% of their grocery expenditure (on average) on single retailer is 6.7% (2,219) , and 17.4% (5,741) on 2 retailers. If we regard people whose income is below 10 as poor, we can find that people concentrate on 2 retailers, most of whom are richer.



number_of_people	hh_income
65	3
138	4
110	6
144	8
272	10
449	11
568	13
492	15
468	16
410	17
392	18
336	19
550	21
416	23
608	26
323	27

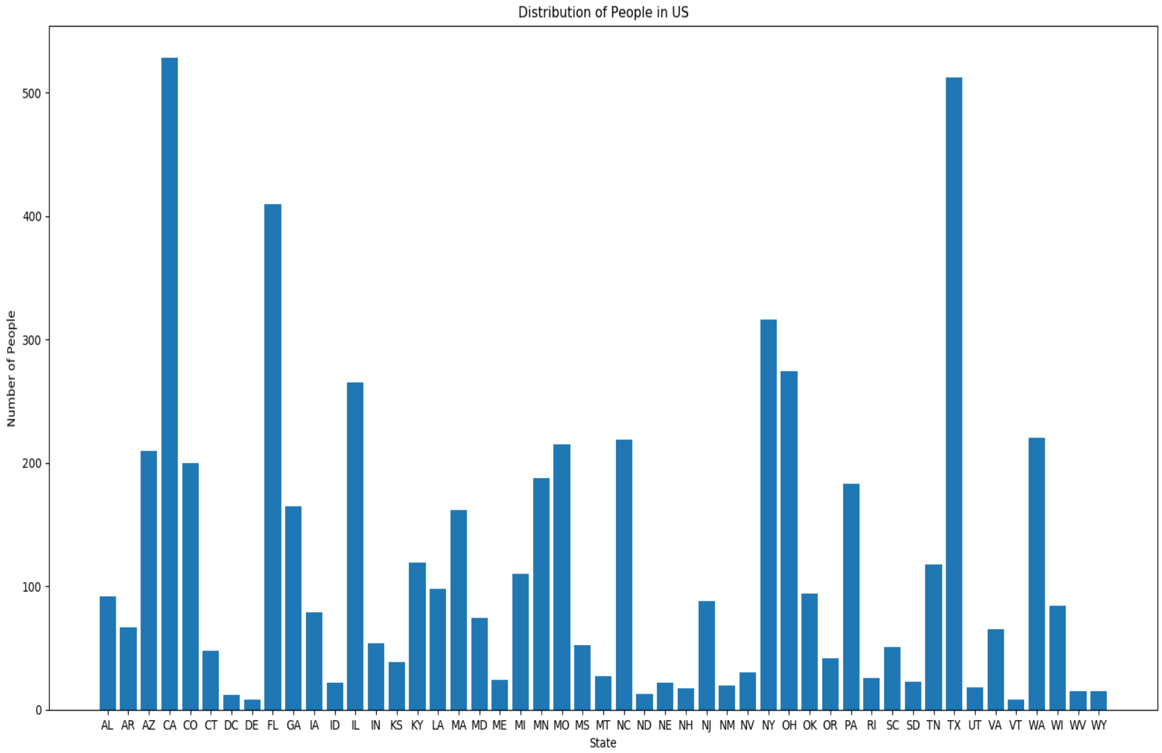
a) What is the retailer that has more loyalists?



shooping	TC_retailer_code
213	4904
211	5850
207	5853
197	4999
194	5899
170	3999
148	4599
135	4914
132	5851
127	5999
119	3997
118	4903
117	4901
110	6205
108	6901
102	7099
99	7003
97	5799
96	6904
92	6999
88	4499
81	6199
69	9
68	7199
65	9999
61	6905

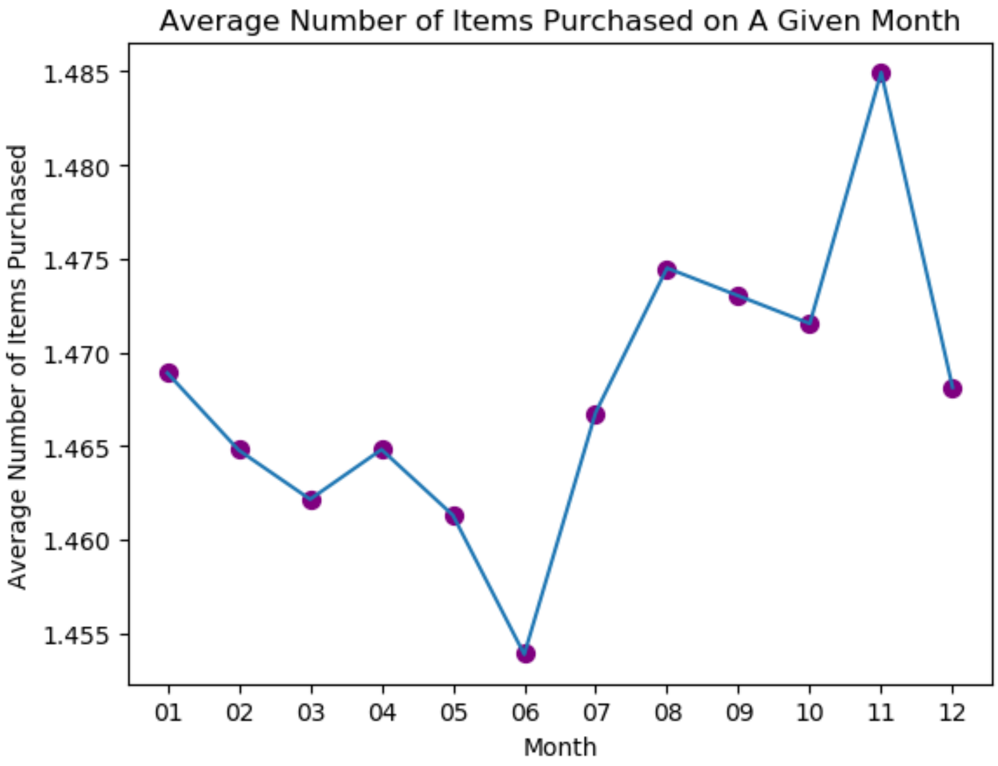
For people who shop at least once a month and spend at least 80% of their expenditure on single retailer, they are more likely to go to grocery with retailer code 4904 for shopping.

b) Where do they live? Plot the distribution by state.

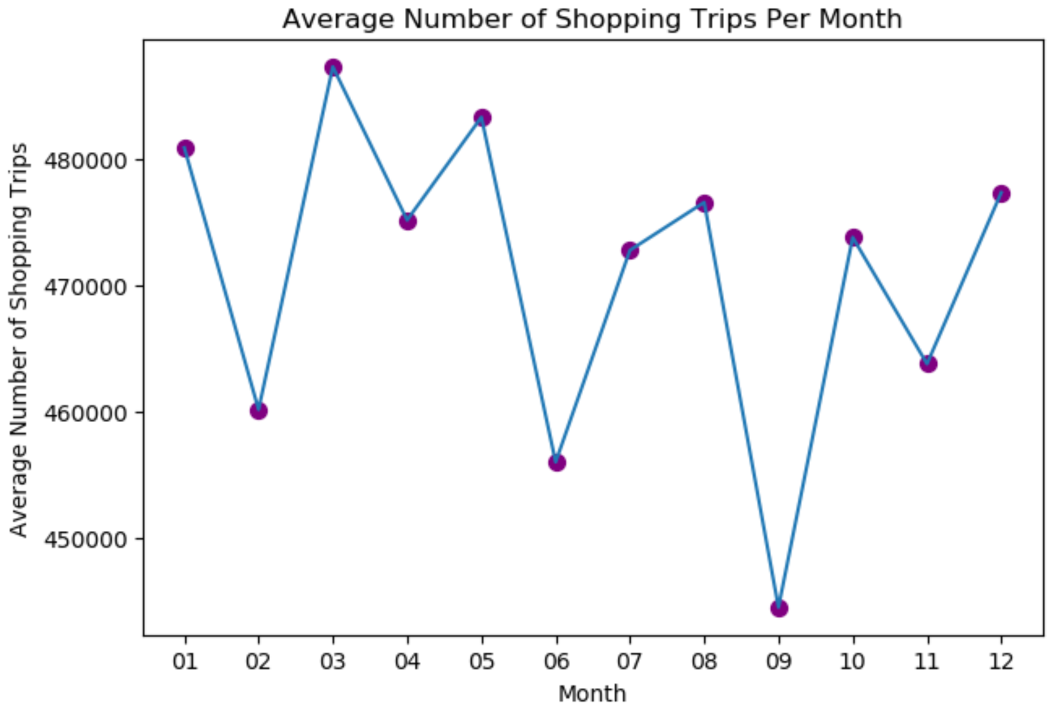


3) Plot with the distribution:

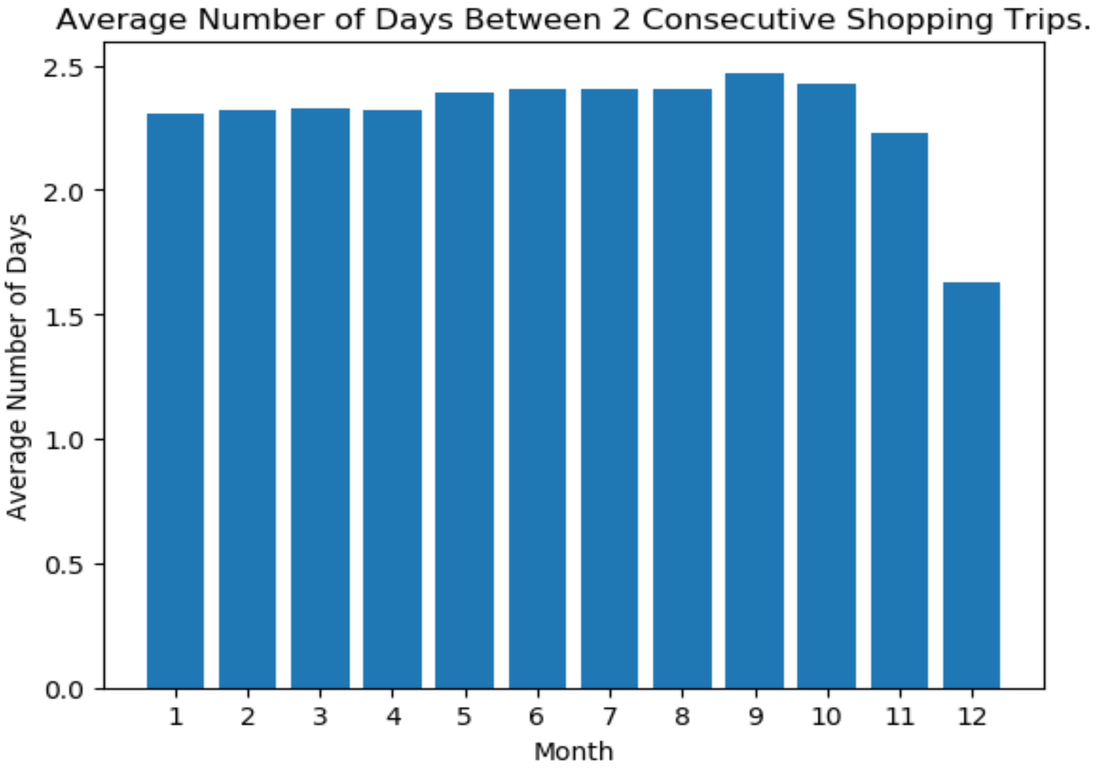
a) Average number of items purchased on a given month.



b) Average number of shopping trips per month.

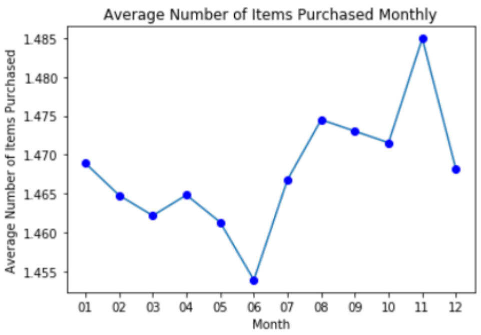


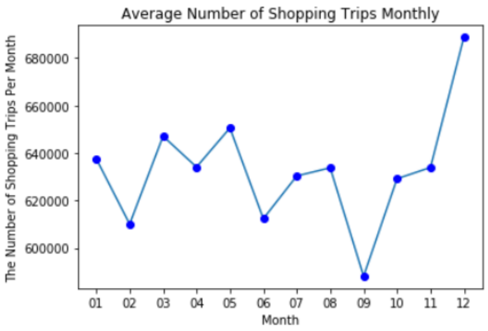
c) Average number of days between 2 consecutive shopping trips.

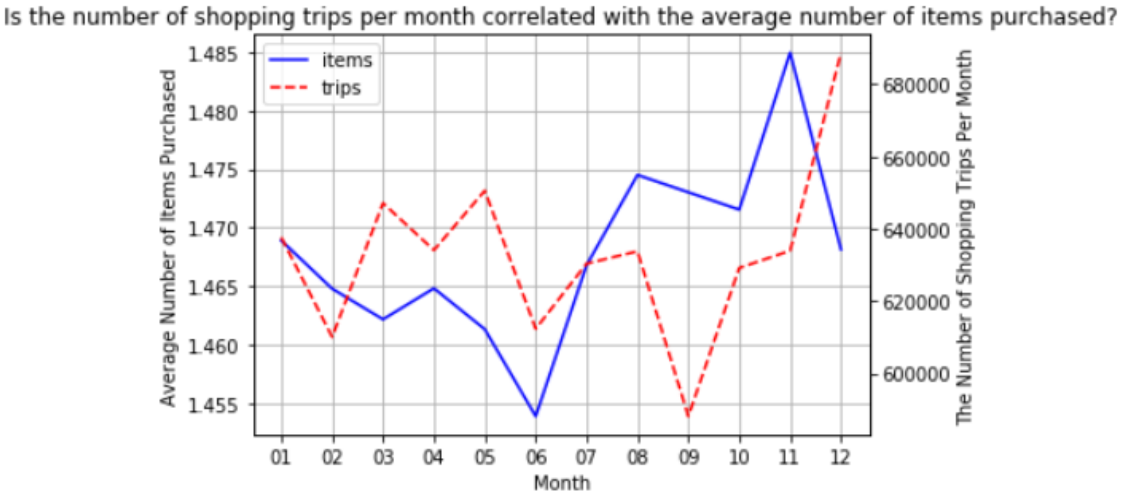


IV. Trends and Relationships between Variables:

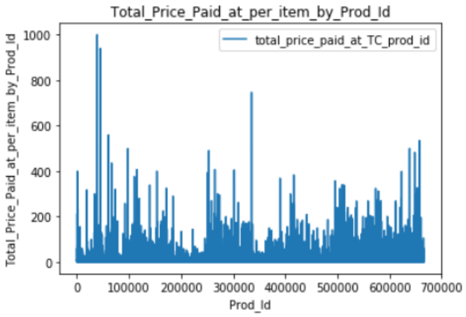
1) **Number of Shopping Trips & Average Number of Items Purchased:** The number of shopping trips per month are not correlated with the average number of items purchased. We reached this conclusion by first calculating the average number of shopping trips per month and plotting that trend throughout the year. Then, we calculated the average number of items purchased per month and also plotted that in a different graph. Finally, we combined both graphs to compare the two trends and we saw that the two trends are very different from each other.

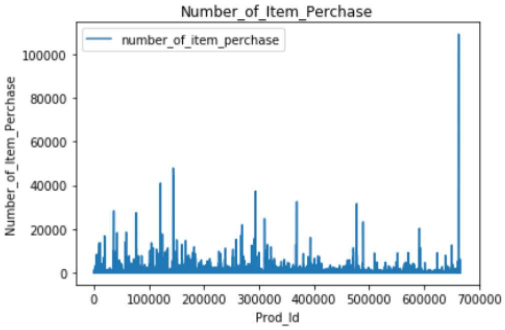


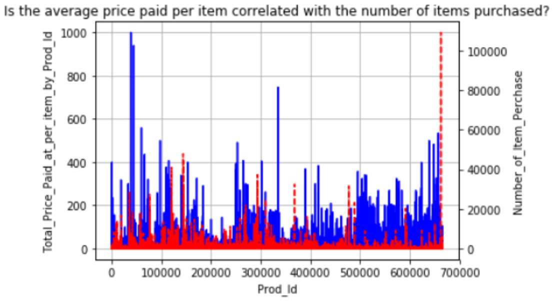


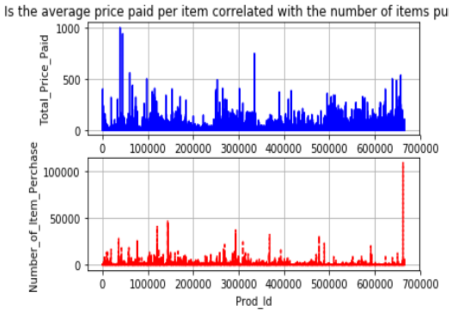


2) **Price Paid Per Item & Number of Items Purchased:** The average price paid per item is not correlated with the number of items purchased on a single trip. We reached this conclusion by first calculating the price paid per item and plotting that. Then, we calculated the number of times a product was purchased and also plotted that in a different graph. Similar to what we did in the previous questions, we combined both graphs to compare the data and we saw that the two are very different from each other. In addition, and just to confirm, we also created a correlation heat map. The heatmap of the average price paid per item and the number of items purchased shows that the correlation between two variables is close to 0.

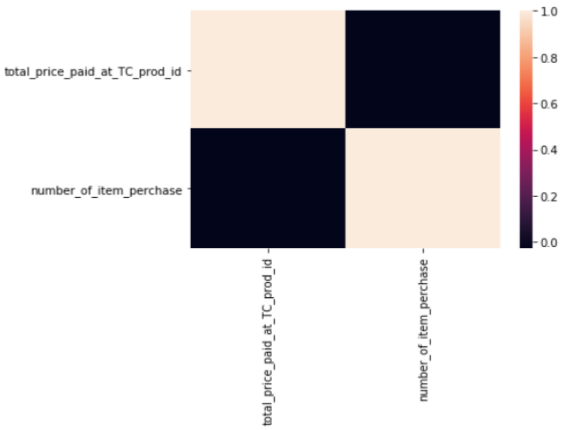






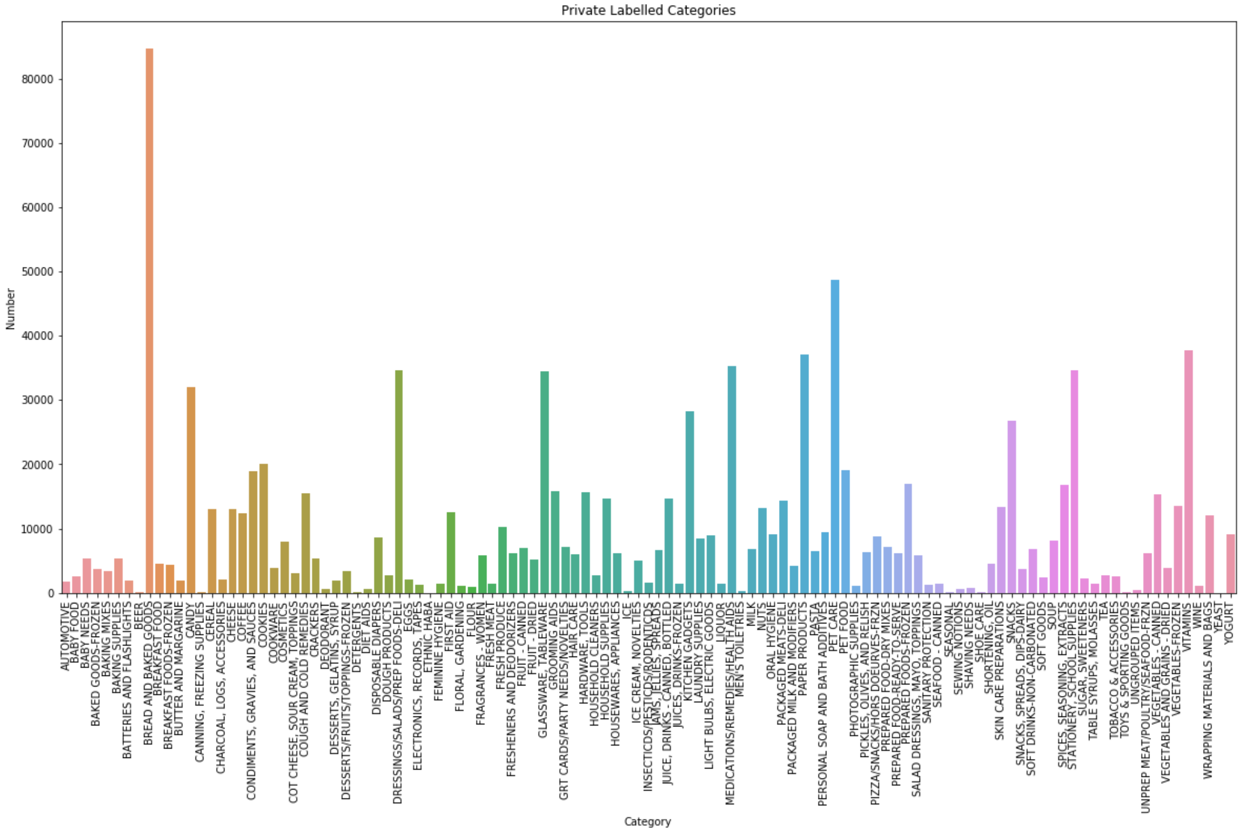


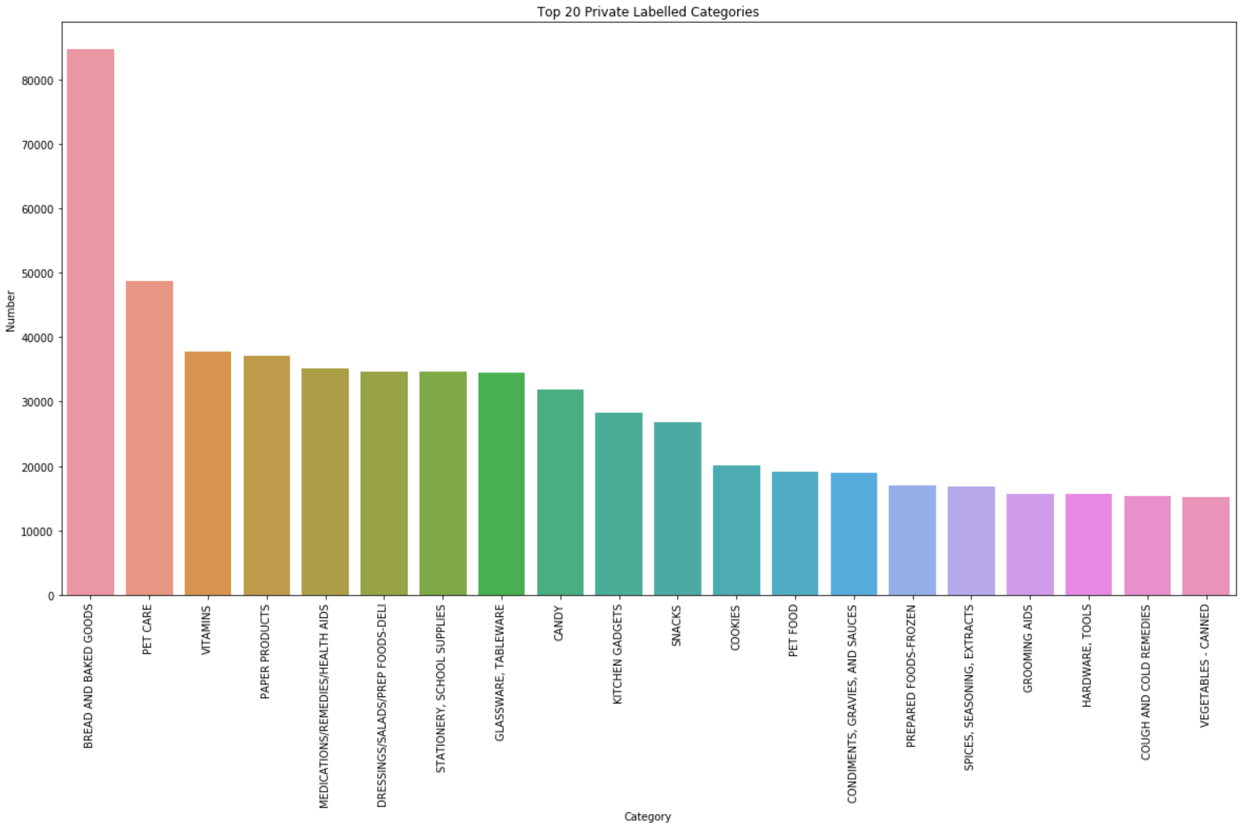
```
ıtplotlib.axes._subplots.AxesSubplot at 0x1cf5926dcc0>
```



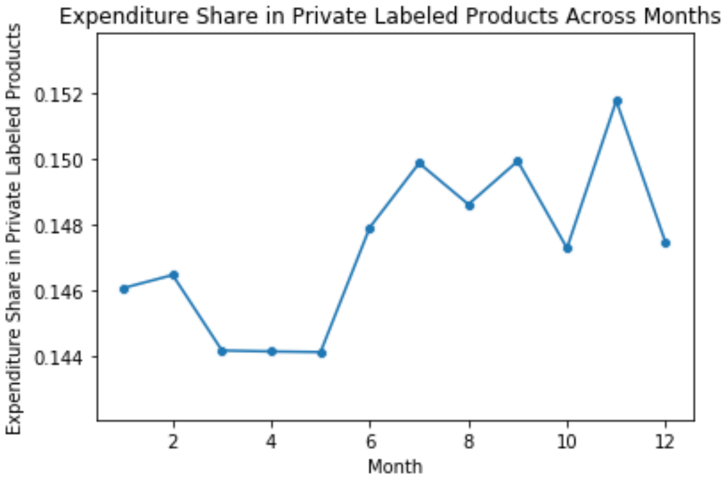
3) **Private Labels:**

a) **Product Categories and Private Labeling:** To understand which product categories tend to have the most private labeled products, we first plotted the number of private labeled products in all 114 categories. Then, we chose the top 20 categories, and determined that those were the categories with the highest number of private labeled products as shown in the second graph. From the result, we can see the ‘Bread and Baked Goods’ category has the most private labeled products and ‘Pet Care’, ‘Vitamins’ are followed.

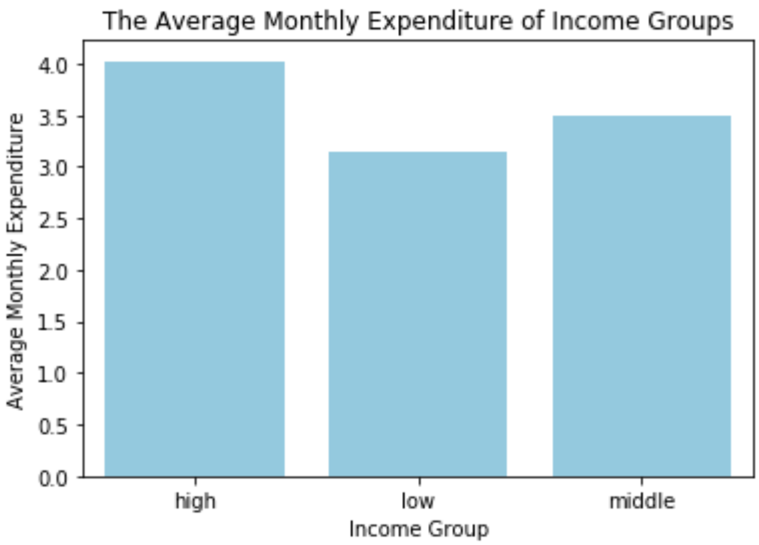




b) **Expenditure Share in Private Labeled Products Across Months:** We calculated both the total expenditure and private labeled products expenditure across months and then divide them to get the ` private_share’. From the result we can see the expenditure share in Private Labeled products is almost constant across months. The share number ranges between 14.4% and15.2%.



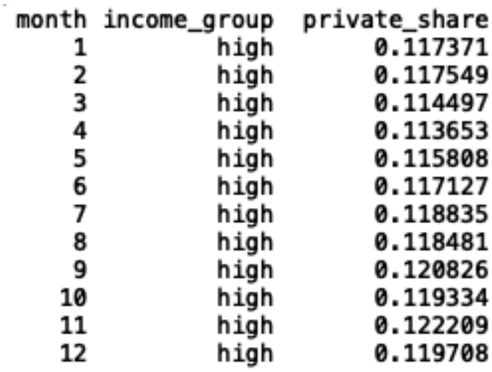
c) **Average Monthly Expenditure on Grocery by Income Group (Low, Medium and High) & the % of Private Label Share in their Monthly Expenditures:** We first clustered the households in three income groups based on their yearly income: low income group: 3-16(‘Under $5k yearly income’ to ‘$30k – $34.9k yearly income’); middle income group :17-26(‘$35k – $39.9k yearly income’ to’$70k – $99.9kyearly income’); high income group: 27($100.0k or more yearly income). Then we analyze the average monthly expenditure among the income groups:

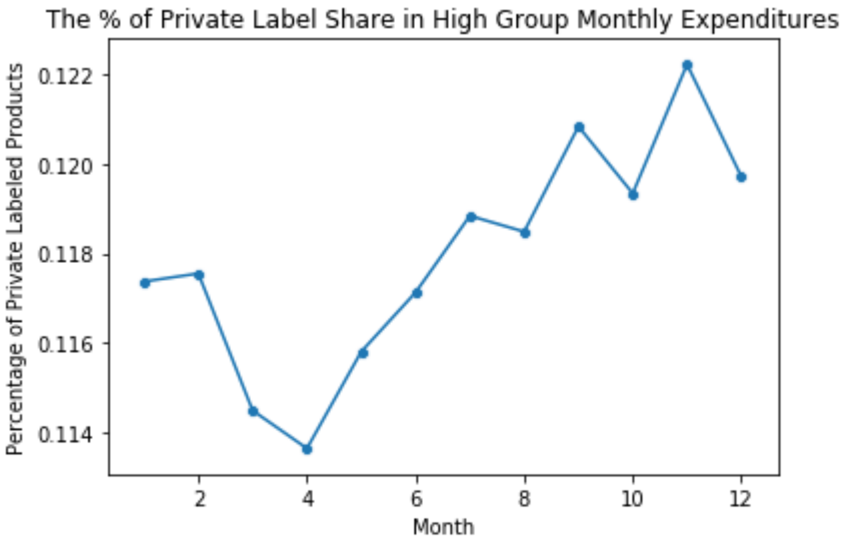


We can see the high income group has the highest monthly expenditure, while the low income group has the lowest one. The expenditure is adequate to their income statues.

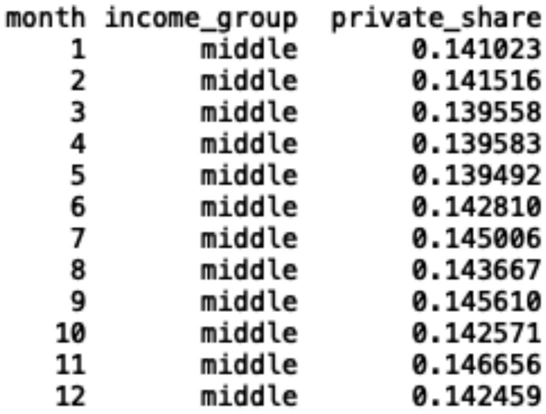
Then we analyze the % of private label share in the monthly expenditures of each group.

High income group:

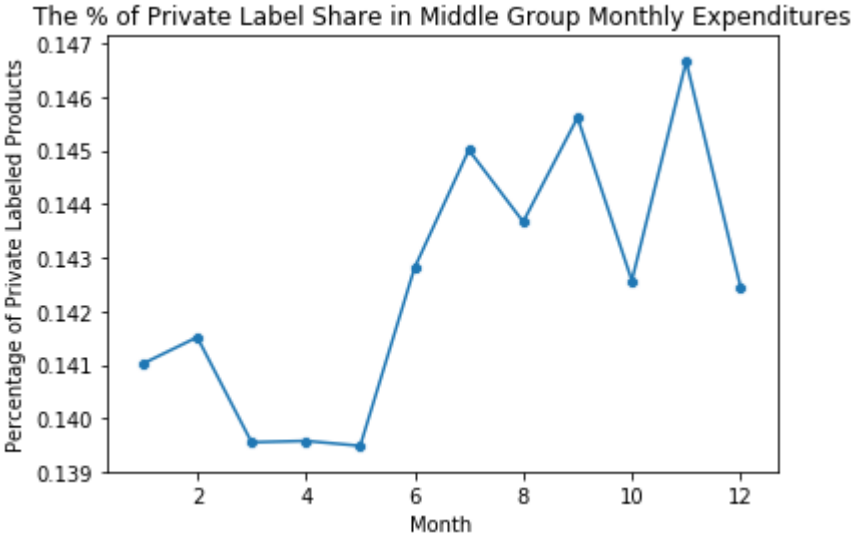

```
month income_group  private_share
    1         high       0.117371
    2         high       0.117549
    3         high       0.114497
    4         high       0.113653
    5         high       0.115808
    6         high       0.117127
    7         high       0.118835
    8         high       0.118481
    9         high       0.120826
   10         high       0.119334
   11         high       0.122209
   12         high       0.119708
```



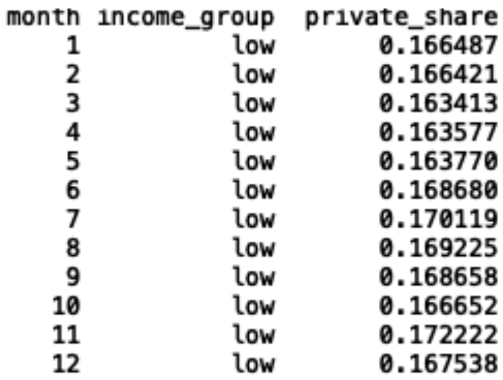
Middle income group:



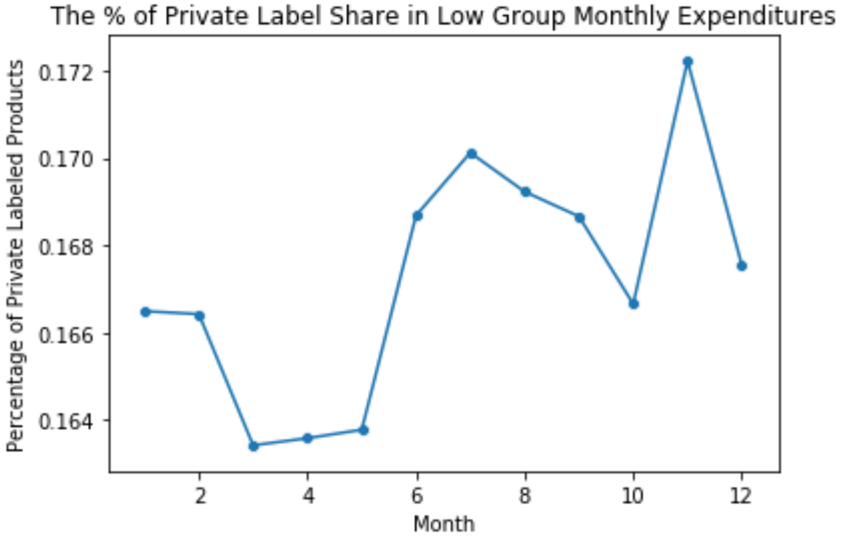
```
month income_group  private_share
    1       middle       0.141023
    2       middle       0.141516
    3       middle       0.139558
    4       middle       0.139583
    5       middle       0.139492
    6       middle       0.142810
    7       middle       0.145006
    8       middle       0.143667
    9       middle       0.145610
   10       middle       0.142571
   11       middle       0.146656
   12       middle       0.142459
```



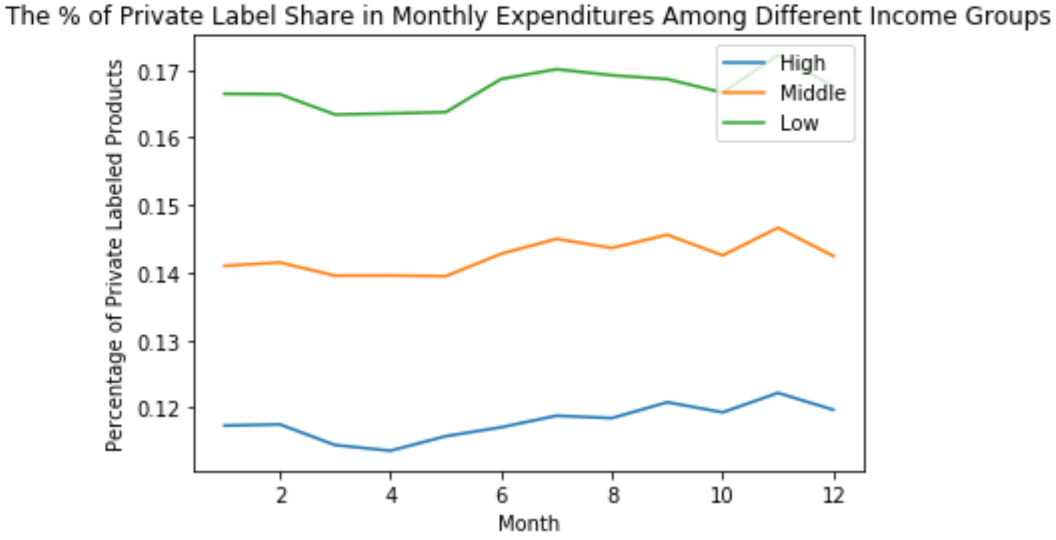
Low income group:



```
month income_group  private_share
    1          low       0.166487
    2          low       0.166421
    3          low       0.163413
    4          low       0.163577
    5          low       0.163770
    6          low       0.168680
    7          low       0.170119
    8          low       0.169225
    9          low       0.168658
   10          low       0.166652
   11          low       0.172222
   12          low       0.167538
```



Finally, we combine all the three groups together,



From the result we can see that for different income groups, the private purchase share changes in trends are similar across months. Low-income group has the highest private expenditure share across months, while high income group has the lowest one. In addition, all income groups have lower private share during month 3-5,and have highest private share in month 11.