Final_Project_Group_22_Part_2

Answers for Questions

Part a. How many:

(1) Store shopping trips are recorded in your database?

There are 7596145 shopping trips in the database.

(2) Households appear in your database?

There are 39577 households appear in the database.

(3) Stores of different retailers appear in our database?

There are 863 different retailers appear in the database.

(4) Different products are recorded?

There are 4231283 different products recorded in the database.

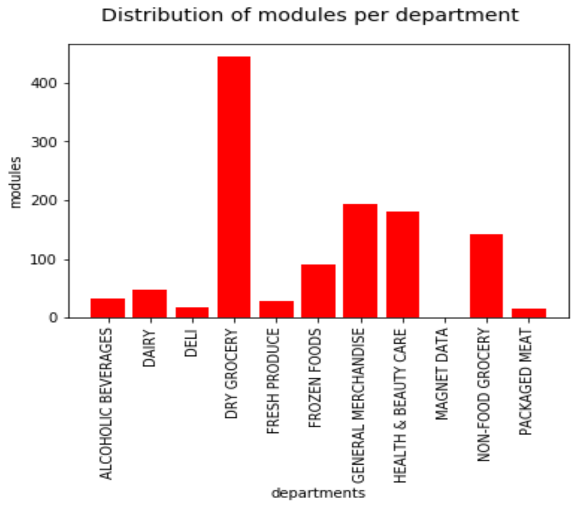
i. Products per category and products per module

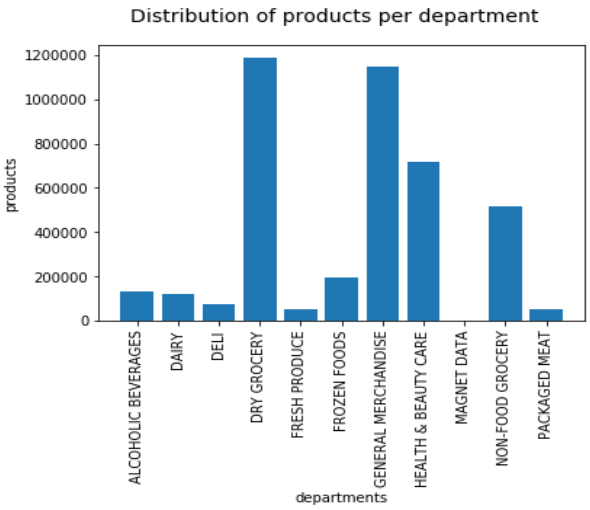
Assumption: The category and module which show the value: ‘NULL’ are not counted.

There are 1224 products per category recorded in the database.

There are 118 products per module recorded in the database.

ii. Plot the distribution of products and modules per department





(5) Transactions?

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

Assumption: In our understanding, coupon is a kind of promotion. So we filter the transactions using coupon to analyze.

The total transactions are 38587942.

Transactions realized under some kind of promotion are 2603946.

Part b.

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

(1) How many households do not shop at least once on a 3-month periods.

Assumption: We use 90 days as 3-month period.

There are 32 households who don't shop at least once in 90 days.

i. Is it reasonable?

It is reasonable.

ii. Why do you think this is occurring?

There are several reasons accounting for this situation.

Firstly, the loss of data and imprecise could rise because of inevitable human factors.

Secondly, the base number is huge, which is 39577, 32 only accounts for 0.08%. Everything has an exception. There should be some people who seldom go shopping. It is a normal social phenomenon.

Thirdly, there could be some extreme examples. For example, some people are so rich or so busy that they do not need to go shopping or they cannot go shopping. Additionally, there will be some extremely poor people who cannot shop at all.

In general, these households who don't shop in a three-month-period cause tiny influence to the overall conclusion.

(2) 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? And among 2 retailers?

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

Assumption: Use race to illustrate demographics.

There demographics are remarkably different. Most of them are White Caucasian.

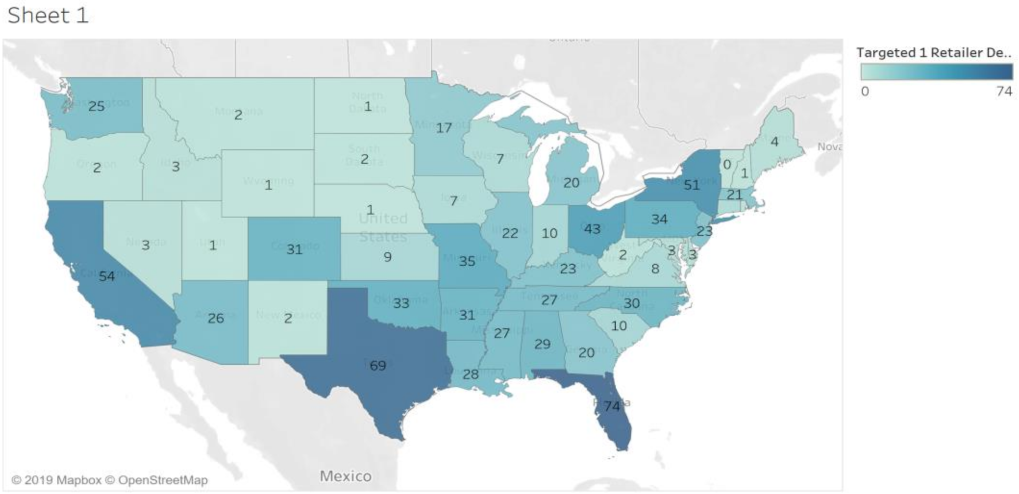
These people are poorer.

ii. What is the retailer that has more loyalists?

For those concentrate at least 80% on single or two retailers, the retailer whose retailer number is 6920 has more loyalists.

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

The distribution of households who concentrate at least 80% on single retailer is as follows:

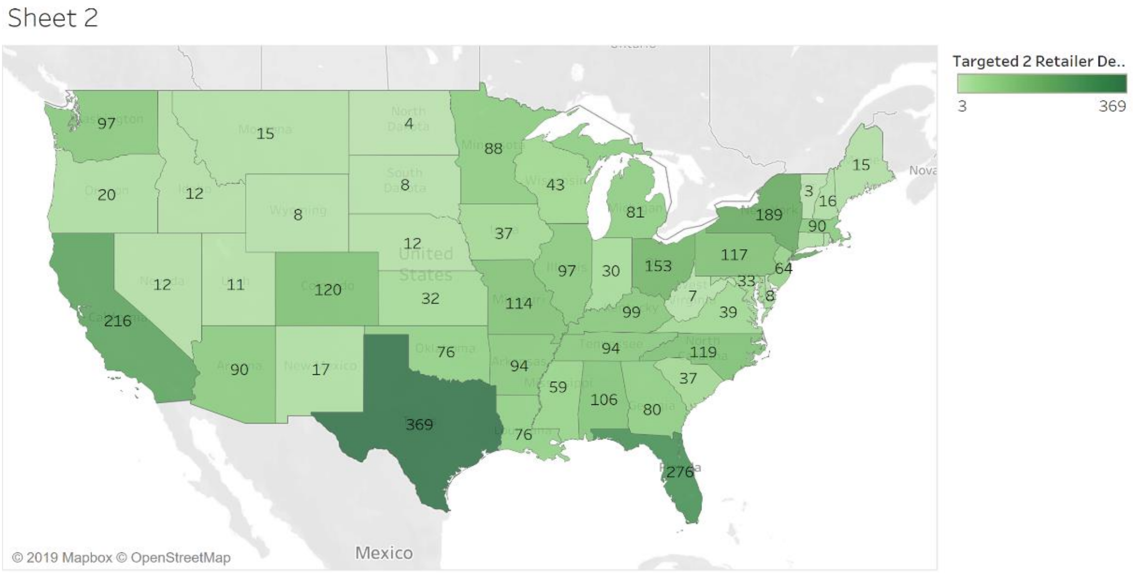


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Targeted 1 Retailer Demo. The marks are labeled by sum of Targeted 1 Retailer Demo. Details are shown for Hh State.

The darker the state on the map, the more household distribution there is in that state. We can see the state with the most households is Florida. States in the east and south have more households who concentrate at least 80% on single retailer.

The distribution of households who concentrate at least 80% on two retailers is as follows:

2 retailers



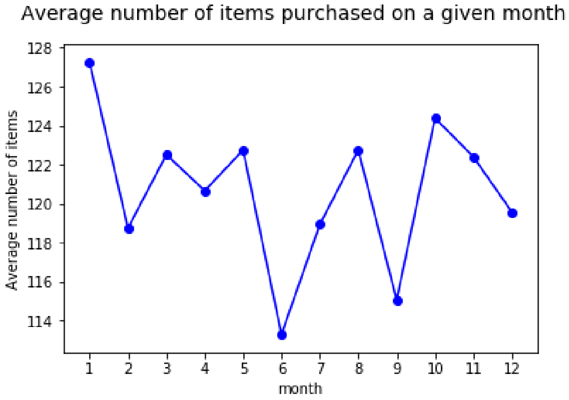
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Targeted 2 Retailer Demo. The marks are labeled by sum of Targeted 2 Retailer Demo. Details are shown for Hh State.

The darker the state on the map, the more household distribution there is in that state. We can see the state with the most households is Texas. States in the east and south have more households who concentrate at least 80% on two retailers.

(3) Plot with the distribution:

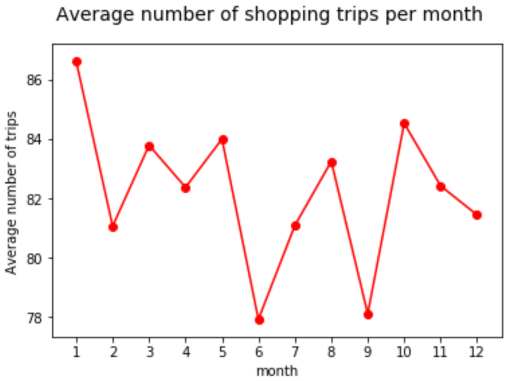
i. Average number of items purchased on a given month. (per household)

Firstly, we get the total number of items every household purchased every month. Then, we calculate the average of all households purchased every month. According to the results, we get the graph below.



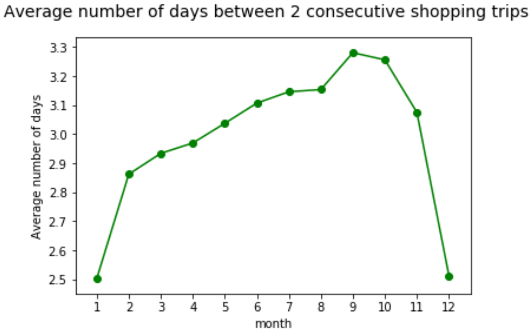
ii. Average number of shopping trips per month. (per household)

Just like the progress above, the total trips every household make every month are calculated, and then the average is available.



iii. Average number of days between 2 consecutive shopping trips. (per household)

The situation of every household is calculated first, and then the average interval every month is calculated.

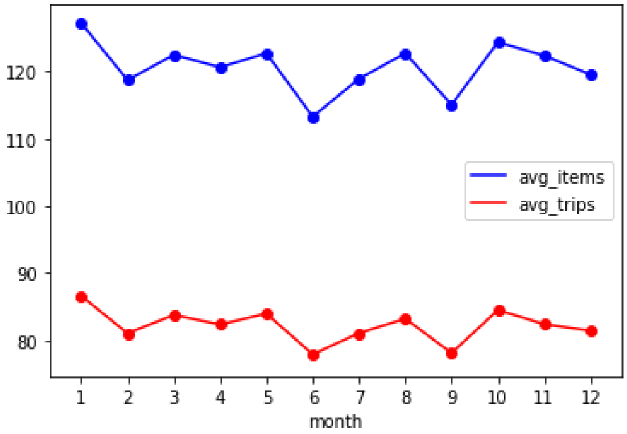


Part c. Make informative visualizations

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

Assumption: We conduct the analysis based on the shopping situation per household per month.

First, we plot with the number of shopping trips per month and the average number of items purchased per month in one graph.



The blue line is the number of items purchased per month and the red line is the number of average shopping trips per month. From the graph we can see that the two lines have similar trends.

Then, we use R to build a linear regression model with the response variable avg_trips and the predictor avg_items. The result of the regression is as follows:

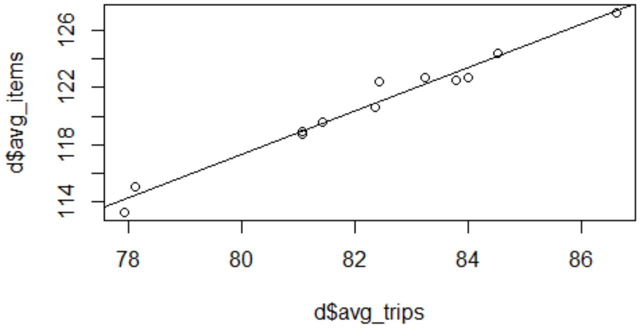
```
Call:
lm(formula = d$avg_items ~ 1 + d$avg_trips)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.22309    6.39958   -0.66    0.524
d$avg_trips  1.51930    0.07781   19.53 2.71e-09 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.65 on 10 degrees of freedom
Multiple R-squared:  0.9744,      Adjusted R-squared:  0.9719
F-statistic: 381.3 on 1 and 10 DF,  p-value: 2.712e-09
```

The multiple R-squared is 0.9744, which means the variance of avg_items can be well explained by avg_trips (97.44%).

Finally, use scatter plot to visualize the relationship:

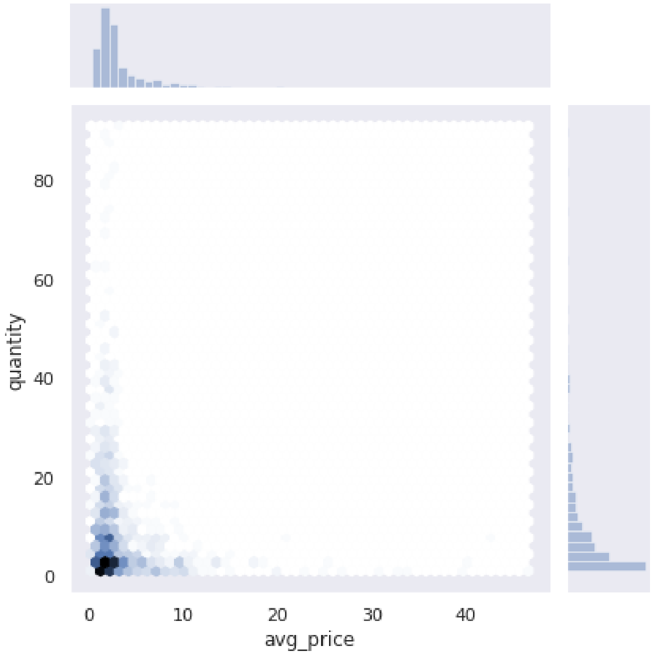


We can see from the scatter plot that the line is straight and smooth, which means the number of shopping trips per month and the average number of items purchased per month has a linear and positive relationship. If a household make monthly shopping trip more frequently, more items will be purchased on average. This conclusion is also reasonable in reality.

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

Assumption: This result is based on the condition of every trip. We calculated the number of items and the average of price

First, we draw a joint plot of the average price paid per item and the average number of items purchased per month.



From this graph, it is clear that most scatters are accumulated at the left bottom, which means lower price the item is, more will be purchased by people.

Then, we use R to build a linear regression model with average price paid per item and the number of items purchased. The result of the regression is as follows:

```
Call:
lm(formula = d1$quantity ~ 1 + d1$avg_price)

Residuals:
Min      1Q  Median      3Q     Max
-10.007  -7.130  -3.907   2.632  81.223

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.26621    0.48298  23.327  < 2e-16 ***
d1$avg_price -0.51873    0.08794  -5.898 5.02e-09 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.85 on 998 degrees of freedom
Multiple R-squared:  0.03369,   Adjusted R-squared:  0.03272
F-statistic: 34.79 on 1 and 998 DF,  p-value: 5.018e-09
```

The multiple R-squared is 0.03369, which means the variance of num_items cannot be well explained by avg_trips (3.369%), the two variables don’t have a significant linear correlation.

(3) 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”

Assumption: We filter all brands which start with ‘CTL BR’, which include their subclasses. Additionally, we use the ‘group_at_prod_id’ as categories.

We define ‘more Private Labelled’ products as those with a higher percentage calculated using ‘CTL BR’ products/all products within a grocery type.

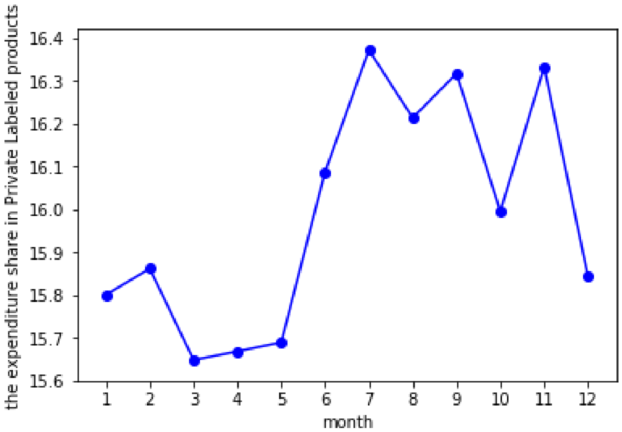
According to the calculation, these categories are proven to be more private labelled, since all of them have over 50% of products labeled as ‘CTL BR’.

Grocery Type	**Percentage(%)**
COUGH AND COLD REMEDIES	65.5
DISPOSABLE DIAPERS	63.3
VEGETABLES-FROZEN	63.1
DOUGH PRODUCTS	60.8
JUICES, DRINKS-FROZEN	59.2
WRAPPING MATERIALS AND BAGS	59
FRUIT - CANNED	56.6
FIRST AID	55.6

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

Assumption: Monthly expenditure share here means monthly expenditure spent on ‘CTL BR’/sum of monthly expenditure spent on all products.

Plot with the distribution of expenditure share in Private Labeled products across months:



Because the y-axis in the graph has a small interval, we may find a high fluctuation. However, the total range of y is from 15.6% to 16.4%, which is an absolutely small range in the whole scale. Therefore, we can conclude that the expenditure share in Private Labeled products is constant across months.

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.

Income level segment

Step1: adjustment of household income

Household income is adjusted for household size. It is reasonable because for the same level of income, the budget of households with more members may be tighter than that of households with fewer members.

We make adjustments using the “equivalence scales” (Garner, Ruiz-Castillo and Sastre, 2003, and Short, Garner, Johnson and Doyle, 1999).

The equation is shown below:

Adjusted household income = Household income / (Household size) ^N

Here we denote N=0.5, following other researchers.

Step2: calculation of medium income

We used SQL to get the median adjust_income is 13.000.

Step3: calculation of range for each class

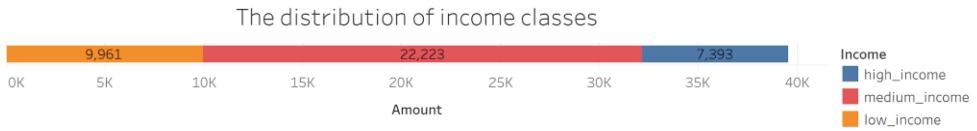
We use the definition of Pew[i] to divide middle class: the category of middle-income is made up of people earning between two-thirds and double the amount of the median household income.

Step4: modification of range for each class

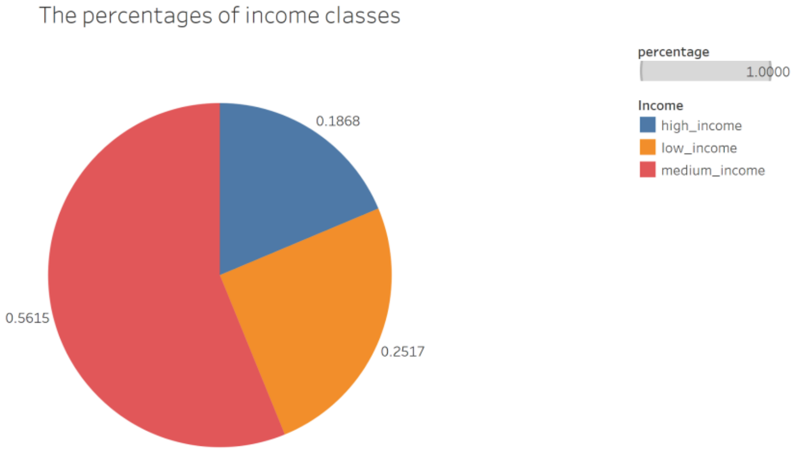
After calculating, we modify the range according to the statistics around 2000. Eventually, the adjusted income range is defined below:

Low income: [0, 10]
Medium income: [10,18]
High income: [18, +]

The distribution of income classes is as follows:


The distribution of income classes

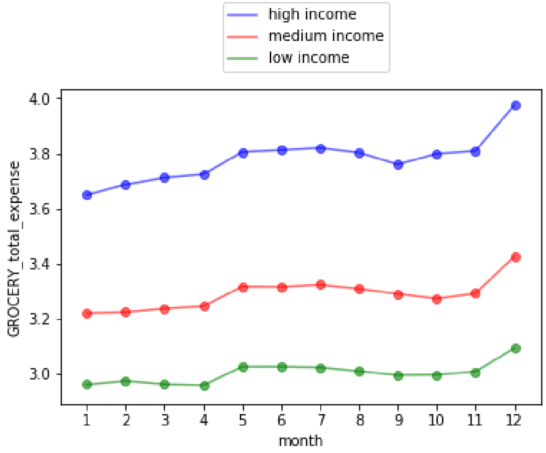
Sum of Amount. Color shows details about Income.


The percentages of income classes

Income (color) and sum of percentage (size).

Report of the average monthly expenditure on grocery:

We draw a graph of average monthly expenditure on grocery of high, medium and low income households.

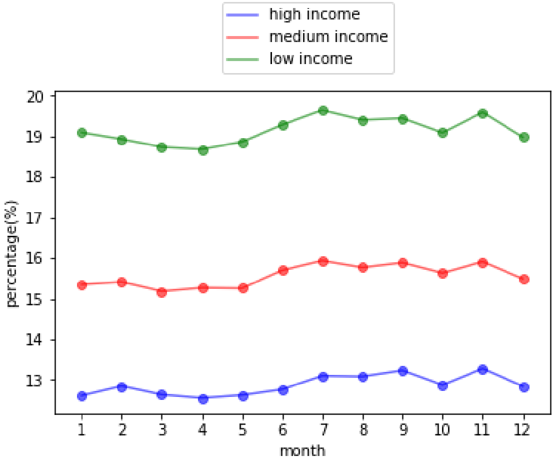


From the graph we can see that high-income class spends the most, medium income class goes with second, and low-income class spends the least money on total grocery expenses. This is because high income class households may purchase more expensive products in all categories, including the grocery.

And in monthly expenditure, we can see the expenditure is significantly higher in December. The reason might be that people need to buy more food for Christmas!!

Study the % of private label share in their monthly expenditures

We draw a graph of percentages of private label share in monthly expenditures of high, medium and low-income households.



From this graph, low income class spends the highest percentage of income in private label products, medium income class goes with second, and high income class spends the least. It can be explained that low income class will spend most of their income on necessities of life, however for high income class, they will have much more disposable money left every month.

[i] https://www.pewsocialtrends.org/2016/05/11/americas-shrinking-middle-class-a-close-look-at-changes-within-metropolitan-areas/