All of visualization codes and separate results are in the python file Visualization.ipynb, and SQL queries to gain data for each question are in All_questions_SQLquery.sql. Extension csv files are in the same folder.

a. How many:

- **Store shopping trips are recorded in your database?**
  7596145 shopping trips.

- **Households appear in your database?**
  39577 households.

- **Stores of different retailers appear in our database?**
  863 different retailers.

- **Different products are recorded?**
  4231283 products

  i. Products per category and products per module
  See extension file final_a4i_products_per_category.csv and final_a4i_products_per_module.csv

  ii. Plot the distribution of products and modules per department

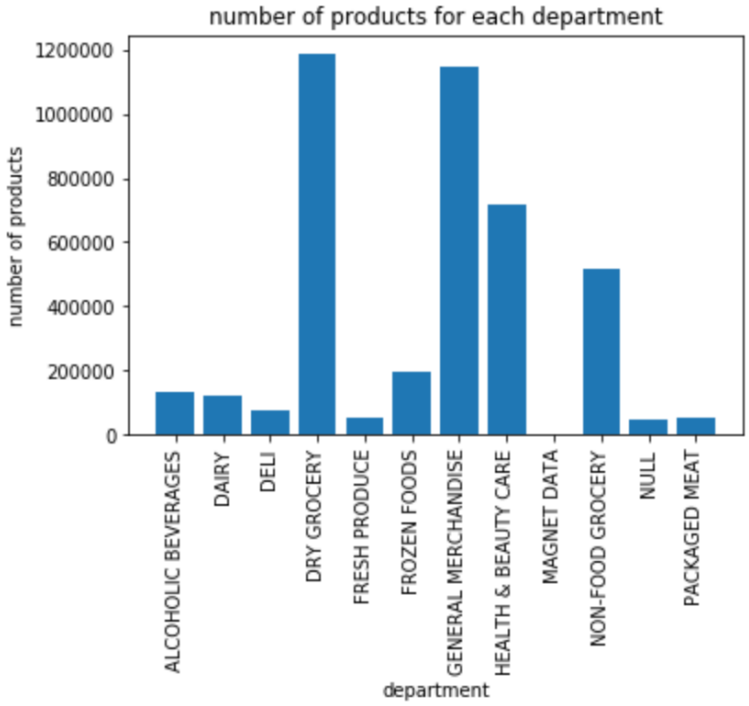

Figure a.1 Number of products for each department

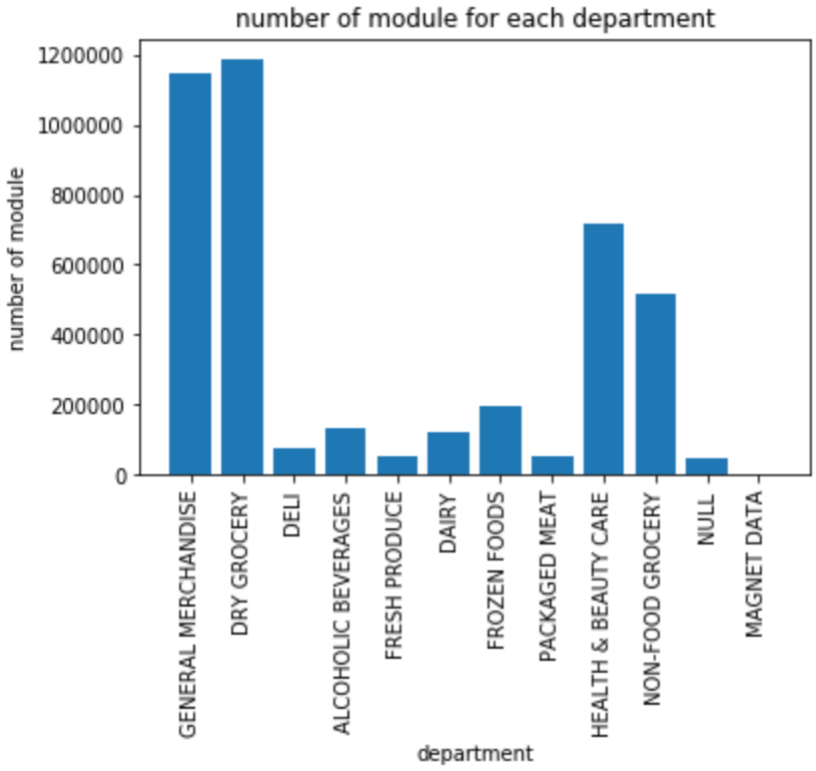

Figure a.2 Number of modules for each department

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

- **How many households do not shop at least once on a 3 month periods?**
  32 households

  i. Is it reasonable?
  Yes, it is reasonable.

  ii. Why do you think this is occurring?
  Under the assumption that household do not shop at least once on a 3 month periods is because of household being out of country for more than 3 month, we believe only a small percentage of household will take a vacation more than 3 months.

- **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?**
  2.45% of them concentrate at least 80% of their grocery expenditure (on average) on single retailer. 9.43% among 2 retailers.

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

NO Remarkably different, both rich and poor household who shop at least once a month will spend 80% of their grocery expenditure in 1 retail store. No Remarkably different, both rich and poor household who shop at least once a month will spend 80% of their grocery expenditure in 2 retail store.

ii. What is the retailer that has more loyalists?

The retailer (TC_retailer_code=6920) has more loyalists (418 household).

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

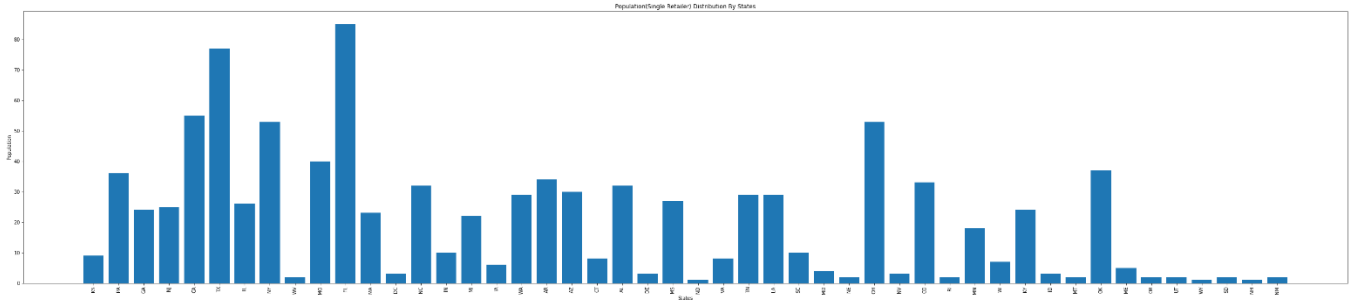

Figure b.1 Population (Single retailer) distribution by states

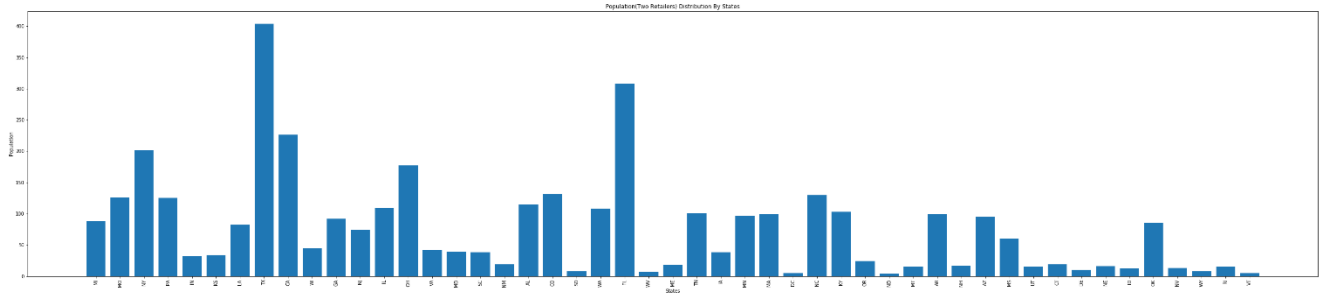

Figure b.2 Population (Two retailers) distribution by states

- **Plot with the distribution:**
  - **i. Average number of items purchased on a given month.**

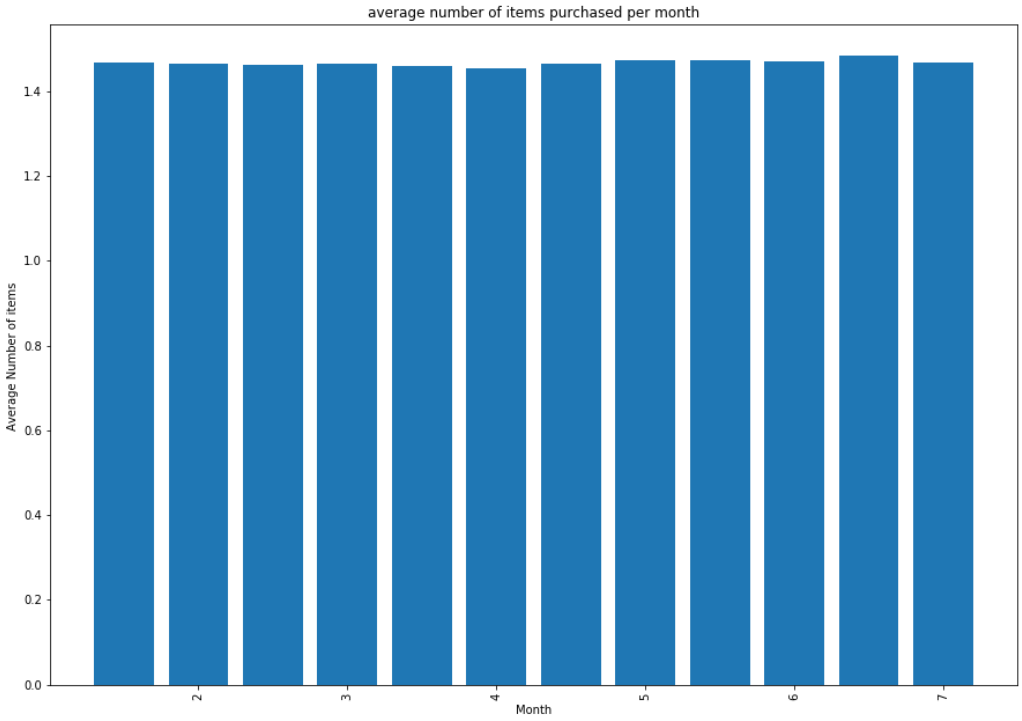

Figure b.3 Average number of items purchased per month

ii. Average number of shopping trips per month.

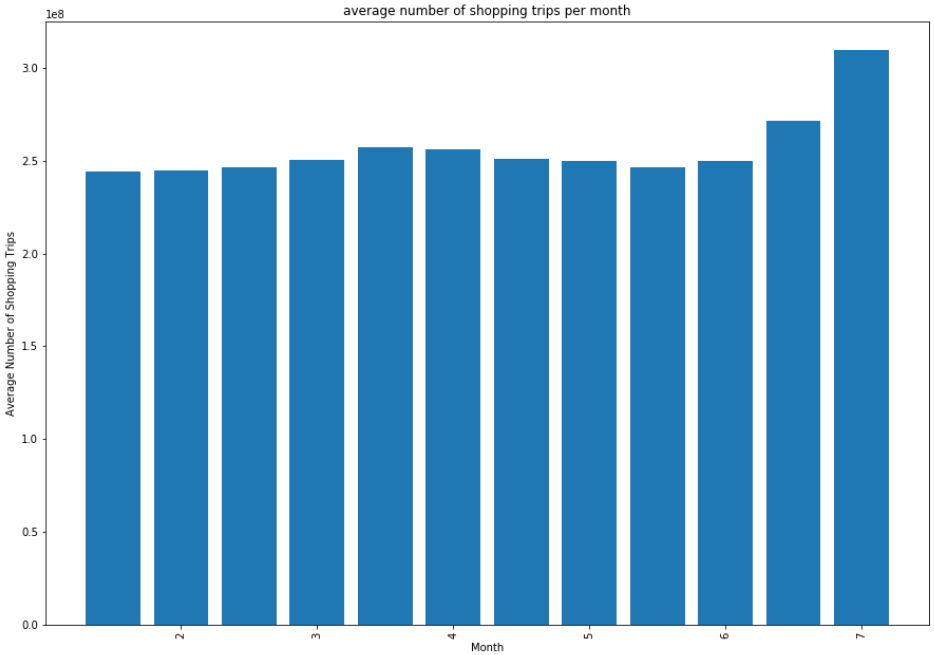

Figure b.4 Average number of shopping trips per month

iii. Average number of days between 2 consecutive shopping trips.

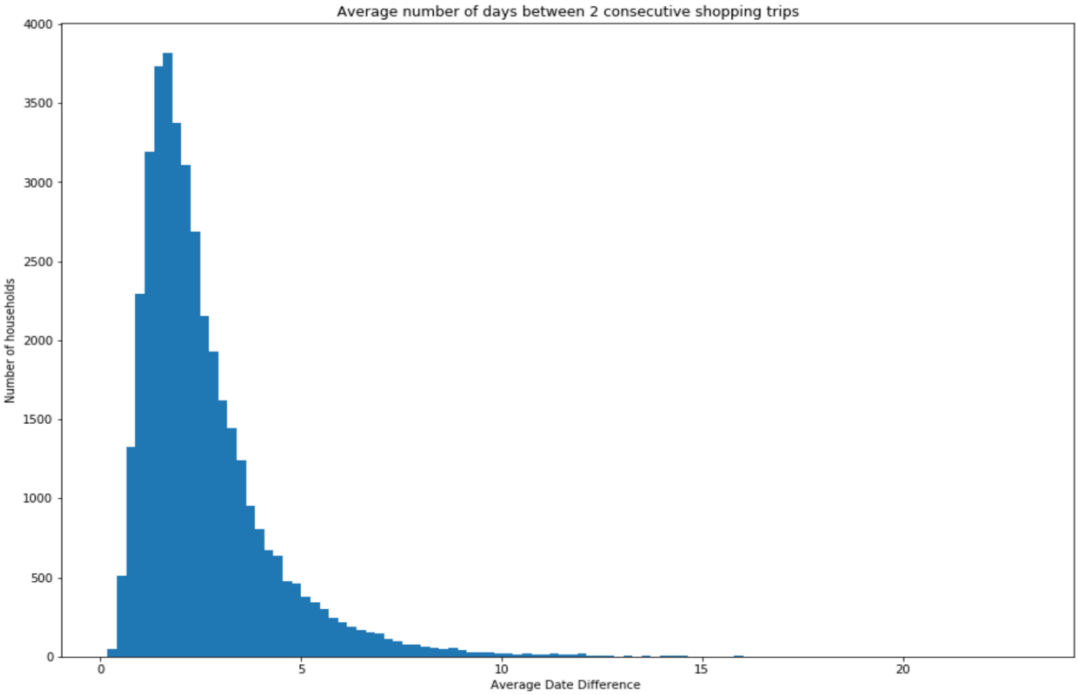

Figure b.5 Average number of days between 2 consecutive shopping trips

c. Answer and reason the following questions: (Make informative visualizations)

- **Is the number of shopping trips per month correlated with the average number of items pur chased?**

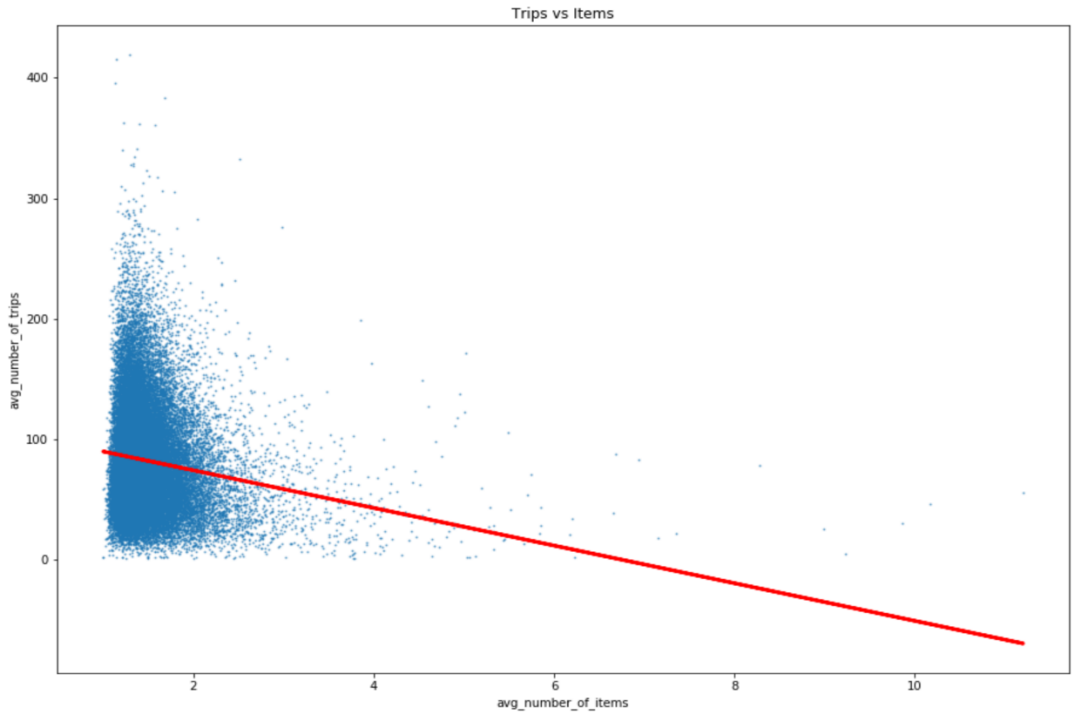

Figure c.1 number of shopping trips per month to average number of items purchased

Because R-square is 0.01919, there is NOT significant correlation between number of shopping trips per month and average number of items purchased.

- **Is the average price paid per item correlated with the number of items purchased?**

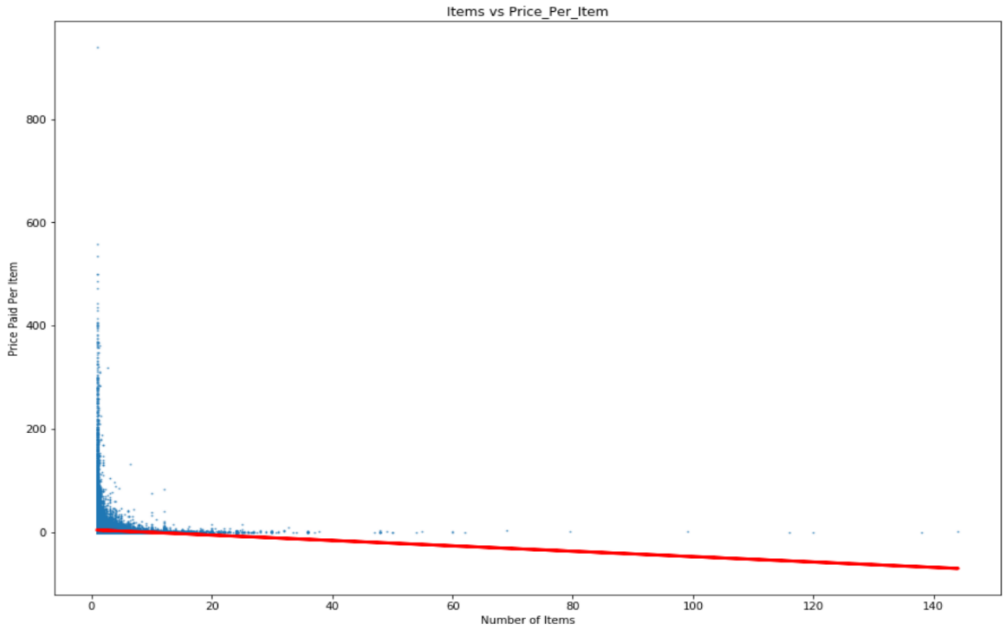

Figure c.2 average price paid per item to number of items purchased

Because R-square is 0.01919, there is NOT significant correlation between average price paid per item and number of items purchased.

- **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"?**

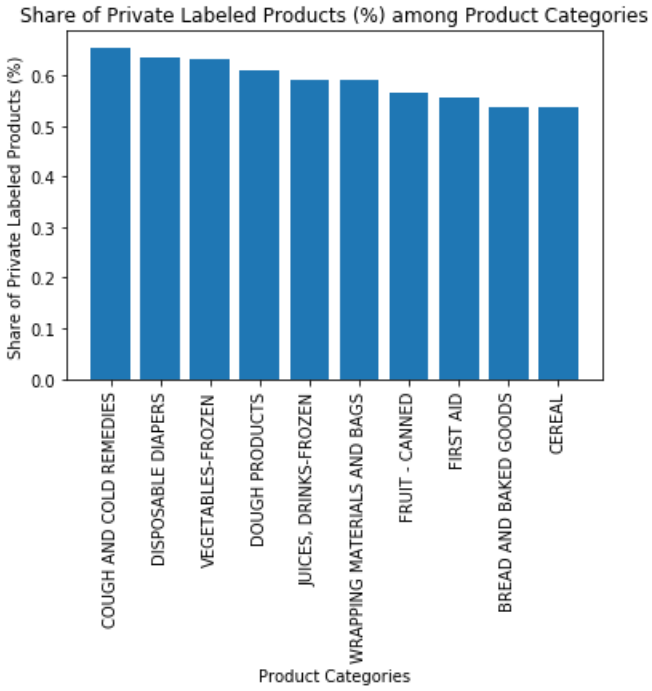

Figure c.3 Share of private labeled products among product categories

We only plotted top 10 product categories, and it is obvious that cough and cold remedies is more "Private labelled".

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

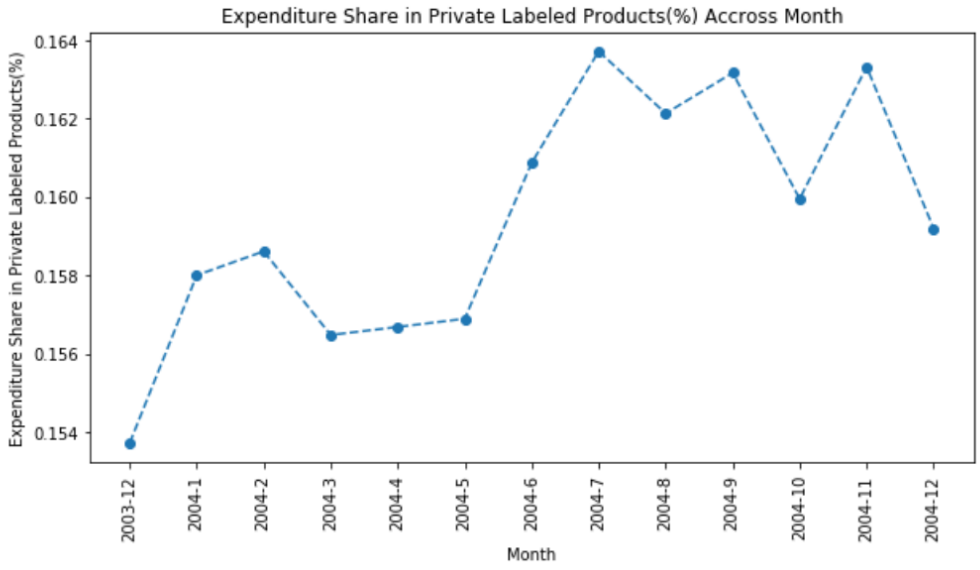

Figure c.4 Expenditure share in private labeled products across month

The plot interprets that the expenditure share in Private Labeled products is NOT constant across month.

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

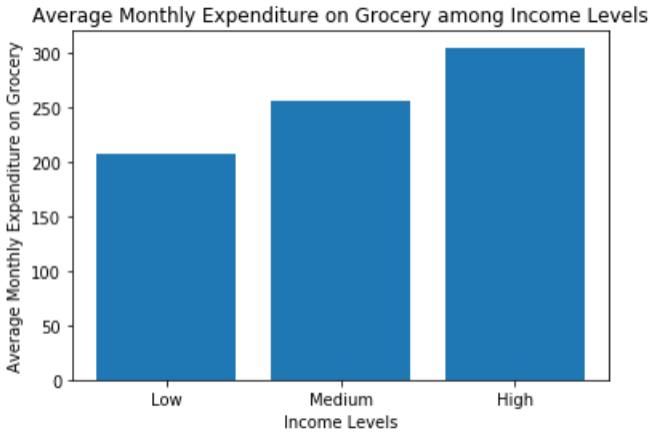

Figure c.4 The average monthly expenditure on grocery

We assumed low income was less than 25k, medium was greater than 25k and less than 50k, and high was greater than 50k. At the meantime, all the products were belong to grocery.

Study the % of private label share in their monthly expenditures. Use visuals to represent the intuition you are suggesting.

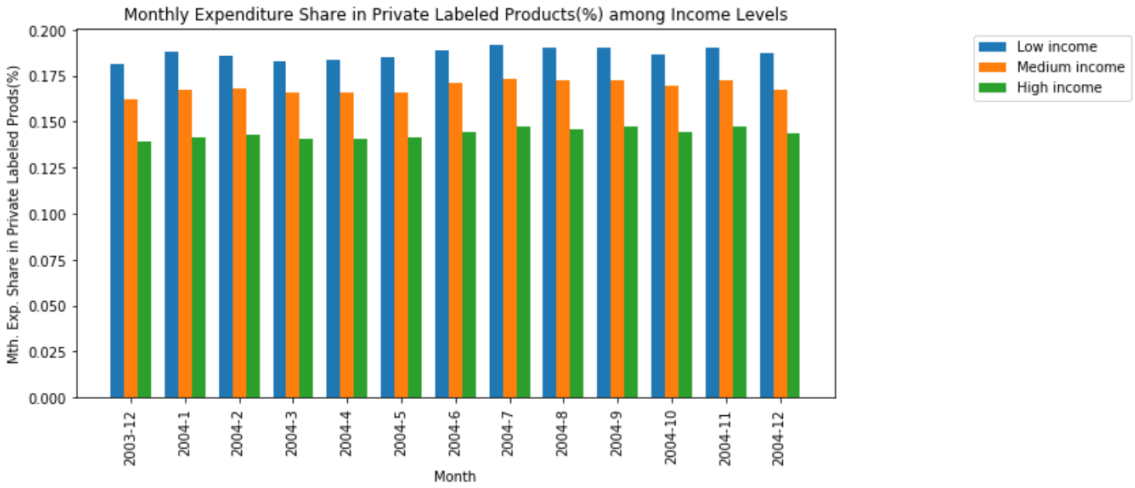

Figure c.5 Monthly expenditure share in Private Labeled products among income levels

According to the plot, we conclude that belonged to the lower level of income, families tend to spend more percentage of money on private labeled products.