Grocery Shopping Behavior Analysis Project

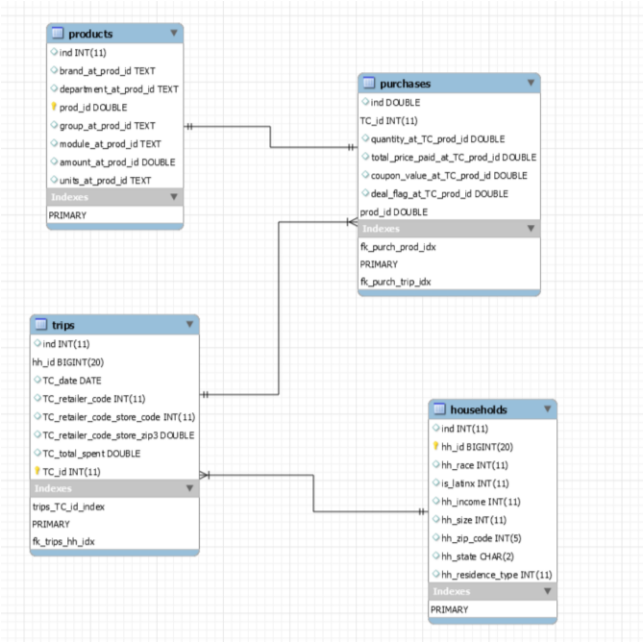
This is the final project for Big Data I (a graduate stage course).

The **Grocery Shopping Behavior Analysis Project** is based on a real-life retailing dataset, which contains 39296405 shopping records for 37916 different households in 7395060 shopping trips of one year.

We designed a MySQL schema for four separated tables:

1. **products** (table about product information, such as brand, amount, unit, …) (3844900 rows)
2. **purchases** (table about what are bought in each purchase, such as product ID, amount that are bought for each item, coupon value, …) (39296405 rows)
3. **trips** (table about purchasing trip, such as which household, when, go to which retailer, …) (7395060 rows)
4. **households** (table about household, such as race, income level, location, …) (37916 rows)

Then import the datasets into this well-organized database. Below is the ER diagram we got.



Based on this database, we further performed queries to extract the shopping patterns displayed by these about 37000 households. Those aggregated results gave us meaningful business insights to the purchasing behavior. We learned that there is always a story behind customers' shopping habit or behavior.

The questions leading to interesting patterns and conclusions mainly included:

(We write sql code for these specific problems)

1. At the household-monthly level, how many households did not shop at least once in a 3-month -periods? Why? Is that reasonable?

2. How many households concentrate most of their shops (over 80% expenditure) on a certain retailer; and among two certain retailers?
3. Which is the retailer that has the most loyalists? (percentage and total number)
4. Where did these loyal customers live? (how many in each state)
5. Average number of days between 2 consecutive shopping trips.
6. Is the number of shopping trips per month correlated with the average number of items purchased?
7. Is the average price paid per item correlated with the number of items purchased?
8. What are the product categories that have proven to be more “Private labelled”?
9. Is the expenditure share in Private Labeled products constant across months?
10. Compare the % of private label share in monthly expenditures among high-level income, median-level income, and low-level income households’ groups.

The sample query code is attached below:

the number of shopping trips & average number of items purchased per month

```
DROP TABLE IF EXISTS tb1;
CREATE TABLE tb1
    SELECT *, SUBSTR(T1.TC_date,3,5) as month
    FROM (SELECT hh_id,
                            TC_id as TC_id_1,
                            TC_date
                FROM trips) AS T1
    LEFT JOIN (SELECT TC_id as TC_id_2,
                                SUM(quantity_at_
                        FROM purchases
        GROUP BY TC_id) AS T2
    ON T1.TC_id_1 = T2.TC_id_2;

DROP TABLE IF EXISTS tb2;
CREATE TEMPORARY TABLE tb2
    SELECT hh_id,
                month,
                COUNT(TC_id_1) as num_trips
    FROM tb1
    GROUP BY hh_id, month;

DROP TABLE IF EXISTS tb3;
CREATE TEMPORARY TABLE tb3
    SELECT month,
                round(AVG(num_trips),2) as avg_trips
    FROM tb2
    GROUP BY month
    ORDER BY month;

DROP TABLE IF EXISTS tb4;
CREATE TEMPORARY TABLE tb4
    SELECT month,
            round(AVG(num_item),2) as avg_items
    FROM tb1
    GROUP BY month
    ORDER BY month;

DROP TABLE IF EXISTS tb5;
CREATE TEMPORARY TABLE tb5
    SELECT T1.month, avg_trips, avg_items
    FROM tb3 as T1
        INNER JOIN tb4 as T2
        ON T1.month = T2.month;

SELECT * FROM tb5;
```

More than that, we applied Python to visualize some of those interesting patterns. The packages that we used include: matplotlib, seaborn.
You can see the answers for the above questions and other interesting findings in the attached PDF. Since the dataset of this project is confidential (belonging to Brandeis University), the data cannot be shared.