Business Challenge: EDA and SQL

- Project: Business Challenge: EDA and SQL Analysis
- Jarian Del Valle, Dylan Rodriguez, Dara Guadalupe, Natalia Torres
- 19-20 of December, 2024

---

1. Introduction

- **Objective**: Analyzing supermarket sales data to answer 10 business questions.
- **Dataset Description**:
  - Source of the dataset: Kaggle
  - Link: https://www.kaggle.com/datasets/chadwambles/supermarket-sales?resource=download)
  - Author: Chad Wambles
  - Key fields in the dataset:
    - Unique sales id.
    - Branch of the supermarket (New York, Chicago, and Los Angeles).
    - City of the supermarket (New York, Chicago, and Los Angeles).
    - Customer Type (Member or Normal). Members receive reward points.
    - Gender (Male or Female)
    - Product name of the product sold.
    - Product category of the product sold.
    - Unit price of each product sold.
    - Quantity of the product sold.
    - 7% sales tax of each product.
    - Total price of the product after tax.
    - Reward points for only members customer type.
- **Tools Used**: SQL, Python, Tableau, Google Docs, Power Point

---

2. Data Cleaning and Assumptions

- **Cleaning Steps:**
  The **Pandas** library was employed during the data exploration and cleaning process to ensure the integrity of the **supermarket sales dataset**. A designated `.py` file in the project repository contains the full Exploratory Data Analysis (EDA) process for reference.
  Notably, the dataset did not include any `NULL` values, which streamlined database, table, and query creation for subsequent data analysis.

- **Assumptions:**
  The dataset was found to be fairly clean and ready for use, with minimal preprocessing required.

---

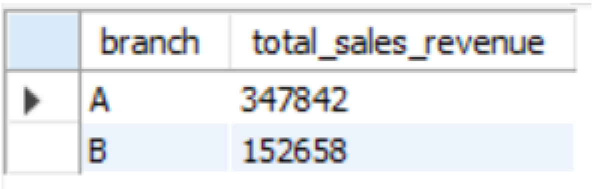
4. Business Questions and SQL Insights

1. Which branch has the highest total sales revenue?

   Query:

```
SELECT
        branch,
        round(sum(total_price),0) AS total_revenue_per_branch
FROM sales
GROUP BY branch;
```

   Result:



	branch	total_sales_revenue
▶	A	347842
	B	152658

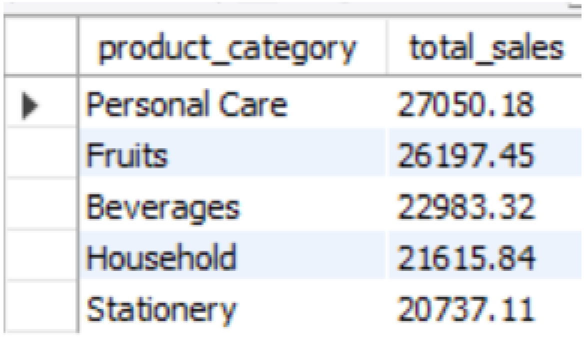
Data shows us that branch A more than doubles the sales revenue of branch B.

2. What is the most purchased product category across all branches?

   Query:

```
SELECT product_category, SUM(total_price) AS total_sales
FROM Sales
GROUP BY product_category
ORDER BY total_sales
DESC LIMIT 5;
```

Result:



product_category	total_sales
Personal Care	27050.18
Fruits	26197.45
Beverages	22983.32
Household	21615.84
Stationery	20737.11

Based on the data retrieved, the product category with more sales across the branches is Personal care.

3. Do members or normal customers spend more on average per transaction?

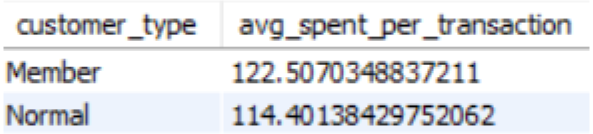
Query:

SELECT customer_type, AVG(total_price) AS avg_spent_per_transaction

FROM supermarket_sales.sales

GROUP BY customer_type;

Result:



customer_type	avg_spent_per_transaction
Member	122.5070348837211
Normal	114.40138429752062

The results indicate that members spend on average $12 more than normal customers.

4. How does the quantity purchased vary by product category?

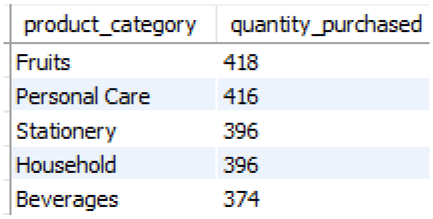
Query:

SELECT product_category, count(quantity) AS quantity_purchased

FROM supermarket_sales.sales

GROUP BY product_category;

Result:



product_category	quantity_purchased
Fruits	418
Personal Care	416
Stationery	396
Household	396
Beverages	374

Fruits: 418, Personal care: 416, Stationary: 396, Household: 396 and Beverages: 374.

The highest quantity of purchased products seem to be products of everyday necessities, prioritizing on food, then personal care.

5. Are there differences in total sales revenue between male and female customers?

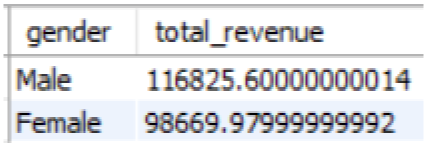
Query:

SELECT gender, SUM(total_price - unit_price) AS total_revenue

FROM supermarket_sales.sales

GROUP BY gender;

Result:



gender	total_revenue
Male	116825.60000000014
Female	98669.97999999992

The results indicate that males contribute significantly more to the total revenue than females. Male: $116,825, Female: $98,669

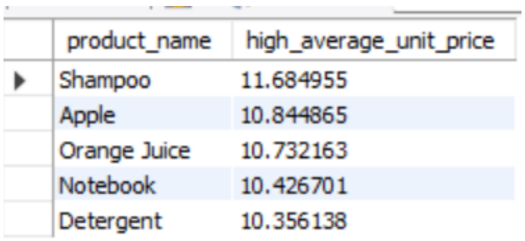
6. Which product has the highest average unit price?

Query:

SELECT product_name, AVG(unit_price) AS high_average_unit_price

```
FROM supermarket_sales.sales

GROUP BY product_name

ORDER BY high_average_unit_price DESC

LIMIT 5;
```

Result:



	product_name	high_average_unit_price
▶	Shampoo	11.684955
	Apple	10.844865
	Orange Juice	10.732163
	Notebook	10.426701
	Detergent	10.356138

The data suggests that the product with the higher average unit price is shampoo at an average of $11.68.

7. How do reward points earned correlate with total price spent?

Query:

```
SELECT reward_points, total_price

FROM supermarket_sales.sales;
```

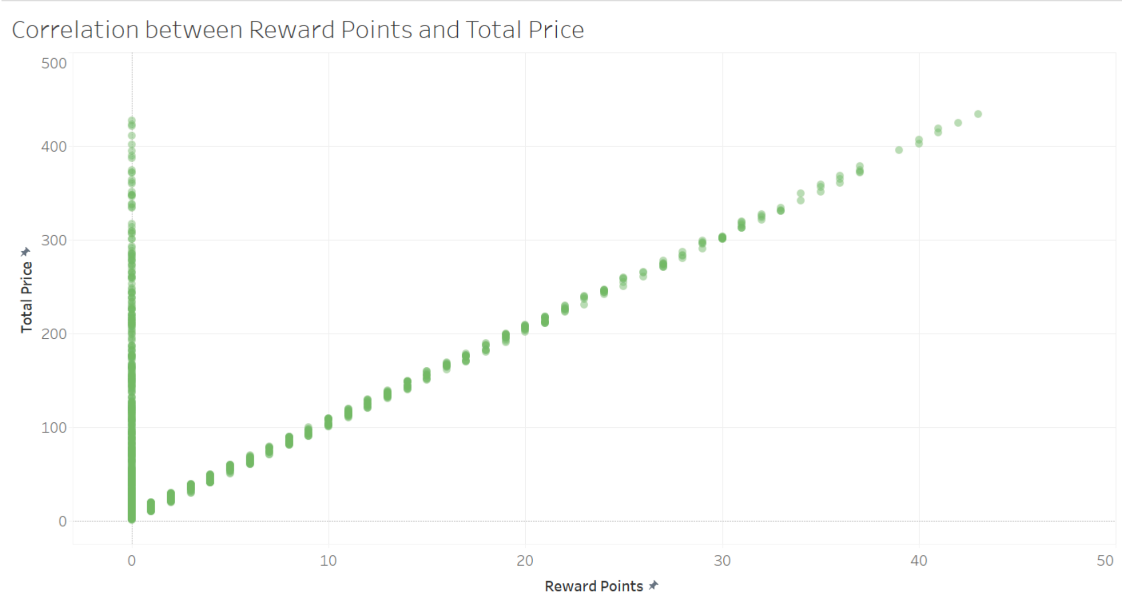
This SQL query retrieves the `reward_points` and `total_price` columns from the `sales` table. The data is then used to calculate the correlation between these two variables in Python.

Python code used:

```
# Import necessary libraries

import pandas as pd

# Define the SQL query to fetch reward points and total price

query = "SELECT reward_points, total_price FROM supermarket_sales.sales;"

# Execute the SQL query and load the results into a pandas DataFrame
```

```
df = pd.read_sql(query, conn)  # `conn` is the connection object to the SQL database

# Calculate the correlation coefficient between reward points and total price

# Using the .corr() method to find the Pearson correlation coefficient

correlation = df['reward_points'].corr(df['total_price'])

# Print the result with a rounded value for clarity

print(f"Correlation between reward points and total price: {round(correlation, 1)}")
```

Result:



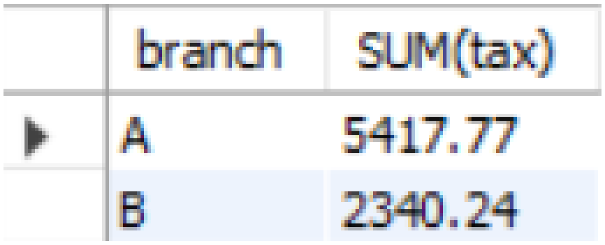
When evaluating correlation, there seems to be a moderate correlation between the reward points earned and the total price spent on a grocery trip of about 0.6 (rounded). The scatter plot generated suggests that for every $100 spent on groceries the rewards points earned duplicate (i.e. $100 -> 10 reward points, $200 -> 20 reward points, etc.)

8. What is the total tax collected per branch?

Query:

```
SELECT branch, SUM(tax) AS total_tax_collected

FROM sales

GROUP BY branch

ORDER BY branch;

-- Query #8
```

Result:



	branch	SUM(tax)
▶	A	5417.77
	B	2340.24

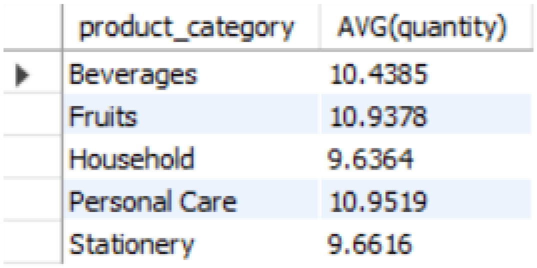
The total tax collected for branch A: $10,835.54, and for branch B: $4,680.48.

9. What is the average quantity purchased per transaction for each product category?

Query:

```
SELECT product_category,

AVG(quantity) AS avg_quantity_purchased

FROM sales

GROUP BY product_category

ORDER BY product_category;
```

Result:



	product_category	AVG(quantity)
▶	Beverages	10.4385
	Fruits	10.9378
	Household	9.6364
	Personal Care	10.9519
	Stationery	9.6616

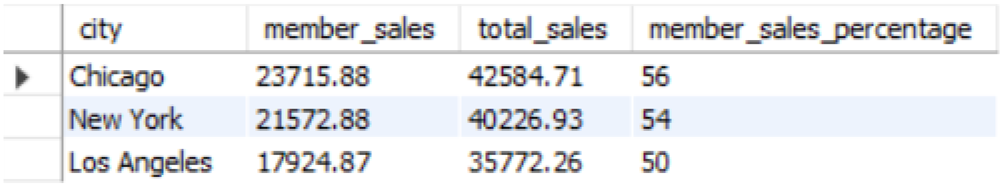
We see that the average quantity purchased per transaction was higher in the beverages category.

10. Which city has the highest percentage of sales contributed by members?

Query:

```
SELECT city, SUM(CASE WHEN customer_type = 'Member' THEN total_price

ELSE 0 END) AS member_sales,  SUM(total_price) AS total_sales,

ROUND((SUM(CASE WHEN customer_type = 'Member' THEN total_price ELSE

0 END) / SUM(total_price)) * 100, 0) AS member_sales_percentage

FROM sales

GROUP BY city

ORDER BY member_sales_percentage DESC;
```

Result:



	city	member_sales	total_sales	member_sales_percentage
▶	Chicago	23715.88	42584.71	56
	New York	21572.88	40226.93	54
	Los Angeles	17924.87	35772.26	50

The results show that Chicago has the highest percentage of sales contributed by members.

---

5. Analysis and Insights

- **Key Findings**:
    - Branch A had a higher total sale.
    - Personal care products were the most frequently purchased.
    - Members spent more on average than normal customers.
    - Customers have shown a higher average purchase rate for beverages.
    - Chicago had more purchases made by members.
    - Branch A had a higher tax collection.
    - Highest average price is in shampoo.
    -
- **Actionable Recommendations**:
    - Suggested actionable steps: Introduce loyalty programs targeting female customers to increase engagement.
    - Continue sharing the benefits of having a membership in the company for higher utilization.
    - Boost promotions in branch B locations for higher sales amounts.
    - Market specials surrounding the most purchased items which are the beverages.
    - Include special discount in higher prices items, like shampoo, for members only to promote the member’s program and improve the sales profit.

---

6. Conclusion

The analysis of supermarket sales data provided valuable insights into branch performance, customer behavior, and product trends. Key findings revealed significant revenue contributions from Branch A and male customers, as well as the popularity of categories like Personal Care and Fruits. The moderate correlation between reward points and spending highlights an opportunity to enhance customer loyalty programs.

This project aligns with business goals by:

- Identifying areas of strength (e.g., high-performing branches and product categories).
- Highlighting opportunities for growth (e.g., improving membership incentives, targeting female customers, and boosting engagement in underperforming branches).
- Providing actionable recommendations to guide future business strategies.

By implementing the suggested strategies, the business can enhance revenue, optimize operations, and foster stronger customer loyalty. This data-driven approach ensures decisions are informed and aligned with market demands.