Supermarket Sale Analysis

Nicole Huang

The analysis will be carried with Excel, SQL and PowerBI. Excel will serve as first repository for the data, SQL will look deep into the data and PowerBI will visualize trends and results.

Steps of the project:

1. **Objective**

   Analyse sales data to:

   1) Identify top-performing branches and product lines.
   2) Understand customer behaviour (e.g., by gender, membership type, city).
   3) Evaluate revenue contributions from product lines and branches.
   4) Assess tax contributions and overall revenue growth patterns.
   5) Fucus on the peak time of shopping.

2. **Data preparation**

   Data overview:

   1) This data is publicly available through Kaggle under https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales?resource=download.
   2) It comes with 1000 rows with 100% pure data. It contains data recorded from January 2019 to March 2019.
   3) It contains 17 columns namely: Invoice ID, Branch, City, Customer type, Gender, Product line, Unit price, Quantity, Tax 5%, Total, Date, Time, Payment, cogs, gross margin percentage, gross income, Rating.
      Here are the details for the columns:

   - Invoice ID: A unique identifier for each transaction.
   - Branch: The branch of the supermarket where the sale occurred. 3 branches are available identified by A, B and C.
   - City: The location of the branch.
   - Customer type: Indicates whether the customer is a "Member" or "Normal" (non-member).
   - Gender: Gender of the customer.
   - Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel.
   - Unit price: Price of each product in $.
   - Quantity: Number of units purchased.
   - Tax 5%: 5% tax fee for customer buying.

- Total: Total price including tax.
- Date: Date of purchase (Record available from January 2019 to March 2019).
- Time: Purchase time (10am to 9pm).
- Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet).
- cogs: Cost of goods sold.
- gross margin percentage: Gross margin percentage.
- gross income: Gross income.
- Rating: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10).

3. Data Exploration and Cleaning

1) Check for null values and duplicates.

- **Clean**: No missing or incomplete data (e.g., every transaction has a valid Invoice ID, Branch, etc.).This is checked by Excel.
- **Free of Errors**: There are no incorrect values (e.g., Quantity is always positive, and total is accurately calculated as Unit price * quantity + tax_5%).

  Query :

```
SELECT COUNT(*) AS wrong_values
FROM supermarket_sales
WHERE ABS(`Unit price` * Quantity + `Tax 5%`- Total) > 0.01;
```

  Result:

wrong_values
0

  The result shows that no discrepancies were found, indicating accurate calculations in the dataset.

2) Check data types.

Query:

```
DESCRIBE supermarket_sales;
```

Result:

Field	Type	Null	Key	Default	Extra
Invoice ID	text	YES		NULL	
Branch	text	YES		NULL	
City	text	YES		NULL	
Customer type	text	YES		NULL	
Gender	text	YES		NULL	
Product line	text	YES		NULL	
Unit price	double	YES		NULL	
Quantity	int	YES		NULL	
Tax 5%	double	YES		NULL	
Total	double	YES		NULL	
Date	text	YES		NULL	
Time	text	YES		NULL	
Payment	text	YES		NULL	
cogs	double	YES		NULL	
gross margin…	double	YES		NULL	
gross income	double	YES		NULL	
Rating	double	YES		NULL	

Column `Date` and `Time` should be Date and Time data type. To address the problem, create two new columns for related data and time.

```
ALTER TABLE supermarket_sales
ADD COLUMN formatted_date DATE;

SET SQL_SAFE_UPDATES = 0;
UPDATE supermarket_sales
SET formatted_date = STR_TO_DATE(Date, '%m/%d/%Y');
SET SQL_SAFE_UPDATEs =  1;

ALTER TABLE supermarket_sales
ADD COLUMN formatted_time TIME;
SET SQL_SAFE_UPDATES = 0;
UPDATE supermarket_sales
SET formatted_time = STR_TO_DATE(Time, '%H:%i');
SET SQL_SAFE_UPDATEs =  1;
```

3) Check duplicates for `Invoice ID`.
   Query:

```sql
SELECT  COUNT(DISTINCT `Invoice ID`)
FROM supermarket_sales;
```

Result:

COUNT(DISTINCT `Invoice ID`)
1000

The result shows that there are no duplicates in `Invoice ID`.

4) Check negative or unrealistic values.

Query:

```sql
SELECT *
FROM supermarket_sales
WHERE `Unit price` <0
OR Quantity <0
OR `Tax 5%`<0
OR total <0;
```

Result:

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date

The result shows that there are no negative or unrealistic values in related columns.

4. Key Business questions

1) Branch and City performance

Query 1:

Identified the branch with the highest total sales.

```sql
SELECT City, Branch, SUM(Total) as Branch_sales
FROM supermarket_sales
GROUP BY City, Branch
ORDER BY Branch_sales DESC;
```

Result:

City	Branch	Branch_sales
Naypyitaw	C	110568.70649999994
Yangon	A	106200.3705000001
Mandalay	B	106197.67199999996

The result shows that branch C (Naypyitaw) outperformed all others, contributing the highest sales. Branch B (Mandalay) had the lowest sales.

Query 2:

Analyzed monthly sales trends across cities.

```
SELECT City, DATE_FORMAT(formatted_date, '%Y-%m') AS month_date,
SUM(Total) as monthly_sales
FROM supermarket_sales
GROUP BY City, month_date
ORDER BY City, month_date;
```

Result:

City	month_date	monthly_sales
Mandalay	2019-01	37176.058500000014
Mandalay	2019-02	34424.27099999999
Mandalay	2019-03	34597.3425
Naypyitaw	2019-01	40434.68099999999
Naypyitaw	2019-02	32934.982500000006
Naypyitaw	2019-03	37199.043
Yangon	2019-01	38681.1285
Yangon	2019-02	29860.120500000005
Yangon	2019-03	37659.12150000001

The result shows All cities had their highest sales in January and the lowest in February, indicating potential seasonal factors.

2) Customer Behavior Analysis

Query 1:

Analyzed sales by customer type (Members vs. Non-Members).

```sql
SELECT
    `Customer type`,
    COUNT(*) AS transaction_count,
    SUM(Total) AS total_sales,
    ROUND(SUM(Total) / (SELECT
                    SUM(Total)
                FROM
                    supermarket_sales),
            2) AS percentage_contribution
FROM
    supermarket_sales
GROUP BY `Customer type`
ORDER BY total_sales DESC;
```

Result:

Customer type	transaction_count	total_sales	percentage_contribution
Member	501	164223.44400000002	0.51
Normal	499	158743.30500000005	0.49

The result shows that Members contributed slightly more transactions and revenue than Non-Members. This highlights the importance of membership programs for driving sales.

Query 2:

Examined sales by gender.

```sql
SELECT
    Gender,
    COUNT(*) AS transaction_count,
    AVG(Total) AS avg_sales,
    SUM(Total) AS total_sales
FROM
    supermarket_sales
GROUP BY Gender;
```

Result:

Gender	transaction_count	avg_sales	total_sales
Female	501	335.0956586826348	167882.92500000002
Male	499	310.78922645290606	155083.82400000014

The result shows that Female customers contributed marginally higher transactions and revenue compared to male customers.

Query 3:

Combined customer type and gender for detailed insights.

```sql
SELECT
    `Customer type`,
    Gender,
    COUNT(*) AS transaction_count,
    AVG(Total) AS avg_sales,
    SUM(Total) AS total_sales
FROM
    supermarket_sales
GROUP BY `Customer type`, Gender
ORDER BY total_sales DESC;
```

Result:

Customer type	Gender	transaction_count	avg_sales	total_sales
Member	Female	261	337.7277528735631	88146.94349999996
Normal	Female	240	332.2332562499999	79735.98149999998
Normal	Male	259	305.0475810810811	79007.3235
Member	Male	240	316.9854187499998	76076.50049999995

The result shows that Female members contributed the most revenue, while male non-members contributed the least.

3) Product Line Insights

Query1:

Ranked product lines by total revenue.

```sql
SELECT
    `Product line`, SUM(Total) AS total_sales
FROM
    supermarket_sales
GROUP BY `Product line`
ORDER BY total_sales DESC
LIMIT 3;
```

Result:

Product line	total_sales
Food and beverages	56144.844000000005
Sports and travel	55122.826499999996
Electronic accessories	54337.531500000005

The result shows that "Food and Beverages", “Sports and travel”, “Electronic accessories ” were the top 3 revenue-generating categories.

Query 2:

Analyzed average unit price and quantity sold per product line.

```
SELECT
    `Product line`,
    AVG(`Unit price`) AS avg_unitprice,
    AVG(Quantity) AS avg_quantity
FROM
    supermarket_sales
GROUP BY `Product line`
ORDER BY avg_unitprice DESC;
```

Result:

Product line	avg_quantity	avg_unitprice
Fashion accessories	5.0674	57.15365168539324
Sports and travel	5.5422	56.993253012048164
Food and beverages	5.4713	56.00885057471268
Home and lifestyle	5.6938	55.31693749999997
Health and beauty	5.6184	54.85447368421053
Electronic accessories	5.7118	53.551588235294155

The result shows that higher unit prices did not always correlate with higher quantities sold, indicating varied customer preferences.

Query 3:

```sql
SELECT
    `Product line`,
    SUM(Total) AS total_sales,
    AVG(`Unit price`) AS avg_unitprice,
    AVG(Quantity) AS avg_quantity
FROM
    supermarket_sales
GROUP BY `Product line`
ORDER BY total_sales DESC;
```

Result:

Product line	total_sales	avg_unitprice	avg_quantity
Food and beverages	56144.844000000005	56.00885057471268	5.4713
Sports and travel	55122.826499999996	56.993253012048164	5.5422
Electronic accessories	54337.531500000005	53.551588235294155	5.7118
Fashion accessories	54305.895	57.15365168539324	5.0674
Home and lifestyle	53861.91300000001	55.31693749999997	5.6938
Health and beauty	49193.739000000016	54.85447368421053	5.6184

The results indicate that the "Food and Beverages" product line generates the highest total revenue, while "Fashion Accessories" has the highest average unit price.

4) Tax contributions

Query:

Calculated total tax contributions by branch.

```sql
SELECT
    Branch,City, SUM(`Tax 5%`) AS total_tax
FROM
    supermarket_sales
GROUP BY Branch, City
ORDER BY total_tax DESC;
```

Result:

Branch	City	total_tax
C	Naypyitaw	5265.176500000002
A	Yangon	5057.160500000002
B	Mandalay	5057.032000000003

The result shows that Branch C generated the highest tax revenue, consistent with its sales performance.

5) Seasonal trends

Query:

Analyzed monthly sales to identify peak periods.

```
SELECT
    DATE_FORMAT(formatted_date, '%Y-%m') AS month_date,
    SUM(Total) AS total_sales
FROM
    supermarket_sales
GROUP BY month_date
ORDER BY total_sales DESC;
```

Result:

month_date	total_sales
2019-01	116291.86800000005
2019-03	109455.50700000004
2019-02	97219.37399999997

The result shows that January had the highest total sales across all branches, while February saw a dip in performance.

5. Advanced Insights

Query 1:

Calculated gross income and profit margins.

```
SELECT
    `Product line`,
    SUM(Total - `Unit price` * Quantity) AS profit,
    ROUND(SUM(Total - `Unit price` * Quantity) / SUM(Total),
            2) AS profit_margin
FROM
    supermarket_sales
GROUP BY `Product line`
ORDER BY profit DESC;
```

Result:

Product line	profit	profit_margin
Food and beverages	2673.5640000000003	0.05
Sports and travel	2624.896499999997	0.05
Electronic accessories	2587.5015	0.05
Fashion accessories	2585.994999999999	0.05
Home and lifestyle	2564.853000000002	0.05
Health and beauty	2342.559	0.05

The result shows that "Food and Beverages" had the highest profit, and all product lines shared the same margin of 5%.

Query 2:

Average revenue per transaction across branches

```
SELECT
    branch,
    COUNT(`Invoice ID`) AS total_transactions,
    SUM(Total) AS total_revenue,
    ROUND(SUM(Total) / COUNT(`Invoice ID`), 2) AS avg_revenue_per_transaction
FROM
    supermarket_sales
GROUP BY Branch
ORDER BY avg_revenue_per_transaction DESC;
```

Result:

branch	total_transactio...	total_revenue	avg_revenue_per_transact...
C	328	110568.70649999994	337.1
B	332	106197.67199999996	319.87
A	340	106200.3705000001	312.35

The result shows that branch C has the highest average revenue per transaction while branch A has the lowest.

Query 3:

Grouped orders into "Morning," "Afternoon," and "Night".

```sql
SELECT
    CASE
        WHEN formatted_time BETWEEN '6:00:00' AND '11:59:59' THEN 'Morning'
        WHEN formatted_time BETWEEN '12:00:00' AND '17:59:59' THEN 'Afternoon'
        ELSE 'Night'
    END AS time_period,
    COUNT(`Invoice ID`) AS order_count,
    ROUND(COUNT(`Invoice ID`) / (SELECT
                    COUNT(*)
                FROM
                    supermarket_sales),
            2) AS order_percent
FROM
    supermarket_sales
GROUP BY time_period
ORDER BY order_count;
```

Result:

time_period	order_count	order_perce...
Morning	191	0.19
Night	281	0.28
Afternoon	528	0.53

The result shows that afternoon had the highest order count, while morning had the least.

Query 4:

Grouped orders into "Morning," "Afternoon," and "Night" for each city.

```sql
SELECT City,
    CASE
        WHEN formatted_time BETWEEN '6:00:00' AND '11:59:59' THEN 'Morning'
        WHEN formatted_time BETWEEN '12:00:00' AND '17:59:59' THEN 'Afternoon'
        ELSE 'Night'
    END AS time_period,
    COUNT(`Invoice ID`) AS order_count,
    ROUND((COUNT(`Invoice ID`) * 100.0) / SUM(COUNT(`Invoice ID`)) OVER (PARTITION BY City), 2) AS percentage_within_city
FROM
    supermarket_sales
GROUP BY City, time_period
ORDER BY City, order_count DESC;
```

Or

```sql
SELECT
    City,
    CASE
        WHEN formatted_time BETWEEN '6:00:00' AND '11:59:59' THEN 'Morning'
        WHEN formatted_time BETWEEN '12:00:00' AND '17:59:59' THEN 'Afternoon'
        ELSE 'Night'
    END AS time_period,
    COUNT(`Invoice ID`) AS order_count,
    ROUND((COUNT(`Invoice ID`) * 100.0) / (SELECT COUNT(`Invoice ID`)
                                           FROM supermarket_sales AS total_orders
                                           WHERE total_orders.City = supermarket_sales.City), 2)
                                           AS percentage_within_city
FROM
    supermarket_sales
GROUP BY
    City, time_period
ORDER BY
    City, order_count DESC;
```

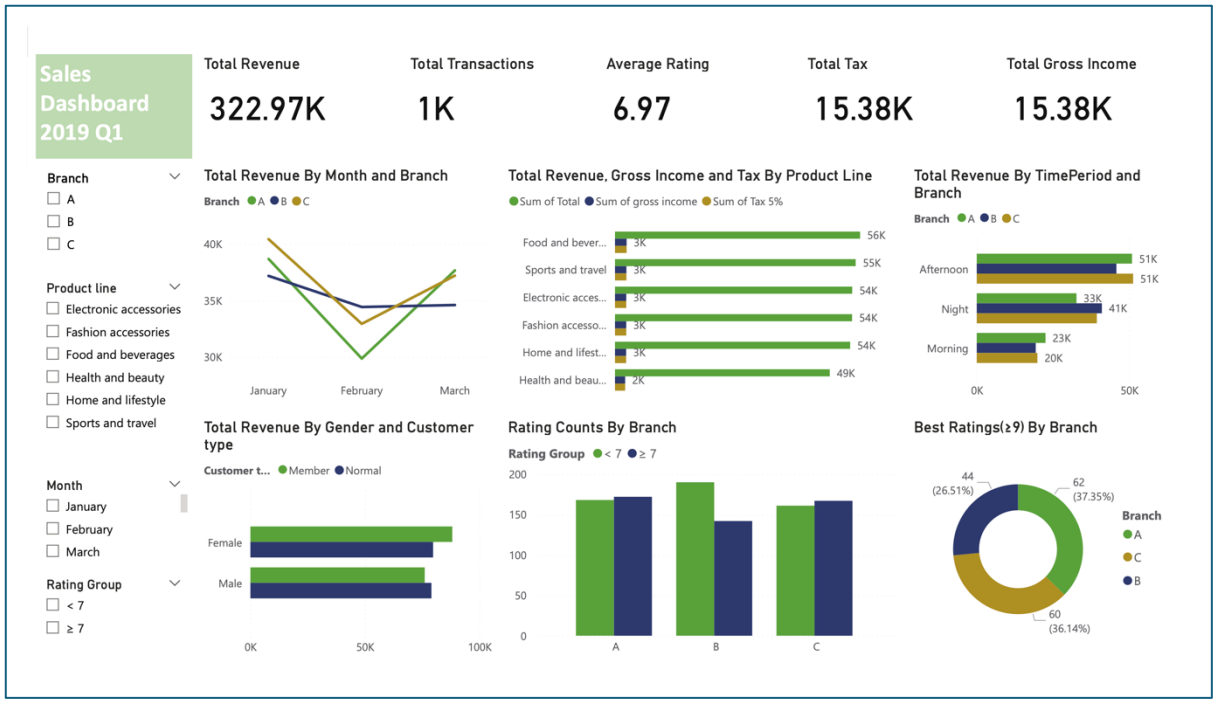
Result:

City	time_period	order_count	percentage_within_cit
Mandalay	Afternoon	162	48.80
Mandalay	Night	111	33.43
Mandalay	Morning	59	17.77
Naypyitaw	Afternoon	181	55.18
Naypyitaw	Night	88	26.83
Naypyitaw	Morning	59	17.99
Yangon	Afternoon	185	54.41
Yangon	Night	82	24.12
Yangon	Morning	73	21.47

The result shows that afternoon had the highest order count across all cities, while morning had the least.

6. Dashboard

Below is a picture of the dashboard that represents the main information of the data.



This is an in-depth analysis of the supermarket sales dataset. The dashboard focuses on the monthly sales from January 2019 to March 2019.

7. Conclusion and Recommendations

Based on the analysis of the supermarket sales data, the following conclusions and actionable recommendations are proposed to enhance performance and drive growth:

1) Branch and City Performance

- Focus on Top-Performing Branches:
  Branch C (Naypyitaw) is the top-performing branch, contributing the highest sales. To further capitalize on its success, allocate additional resources such as marketing budgets and inventory to this location. Additionally, analyze the strategies employed by Branch C and replicate them in underperforming branches, particularly Branch B (Mandalay), which has the lowest sales.
- Seasonal Promotions:
  January recorded the highest sales across all branches, indicating a peak shopping period. To maximize revenue, run targeted promotions and marketing campaigns during this time. For February, which experienced a dip in sales, introduce discounts or loyalty programs to boost customer engagement during this slower period.

2) Customer Behavior Analysis

- Membership Programs:
  Members contributed slightly more revenue than non-members. To further

enhance this, consider offering exclusive discounts, early access to sales, or additional perks for members. This will encourage more customers to join the membership program and increase customer loyalty.

- Target Female Customers:
  Female customers contributed more transactions and revenue compared to male customers. Tailor marketing campaigns to appeal more to female shoppers by promoting products that align with their preferences or offering female-centric promotions.
- Personalized Marketing:
  Female members contributed the most revenue. Create personalized marketing campaigns for this demographic by sending targeted offers or recommendations based on their purchase history.

3) Product Line Insights

- Focus on Top-Performing Product Lines:
  The top three revenue-generating categories are "Food and Beverages," "Sports and Travel," and "Electronic Accessories." Ensure these product lines are well-stocked and prominently displayed in stores. Consider expanding the range of products in these categories to attract more customers.
- Pricing Strategy:
  While "Fashion Accessories" has the highest average unit price, it does not correlate with higher quantities sold. Review the pricing strategy for this category and consider offering bundle deals or discounts to increase sales volume.
- Promote High-Margin Products:
  Since all product lines have the same profit margin (5%), focus on promoting high-volume products like "Food and Beverages" to maximize overall profit.

---

4) Tax Contributions

- Monitor Tax Contributions:
  Branch C generated the highest tax revenue, consistent with its sales performance. Ensure that all branches are compliant with tax regulations and consider optimizing pricing strategies to maintain healthy tax contributions across all locations.

---

5) Seasonal Trends

- Peak Shopping Times:
  Afternoon is the peak shopping time across all cities. Schedule more staff during this period to handle increased customer traffic and ensure a smooth shopping experience. Additionally, consider running afternoon-specific promotions to attract more customers during this time.

6) Advanced Insights

- Gross Income and Profit Margins:
  "Food and Beverages" had the highest profit. Focus on expanding this category by introducing new products or improving the quality of existing ones to attract more customers.
- Average Revenue per Transaction:
  Branch C has the highest average revenue per transaction. Analyze the factors contributing to this success, such as product mix, customer service, or store layout, and implement similar strategies in other branches.
- Time-Based Promotions:
  Since afternoon is the busiest time, consider offering time-based promotions, such as "Happy Hour" discounts, to further boost sales during this period.

7) Customer Experience

- Improve Customer Ratings:
  Customer ratings indicate that Branch B receives the highest number of low ratings (less than 7) and the fewest good ratings (≥7), suggesting potential issues with product quality, service, or store experience. In contrast, Branch A outperforms others in positive customer feedback. To improve overall satisfaction, implement targeted strategies at Branch B, such as staff training, service enhancements, or product quality checks. Enhancing the customer experience can drive repeat business and ultimately boost sales.

8) Marketing and Promotions

- Leverage Data for Targeted Campaigns:
  Use insights from customer behavior, product performance, and shopping patterns to create targeted marketing campaigns. For example, send personalized offers to female members or promote "Food and Beverages" during peak shopping hours.
- Seasonal Campaigns:
  Since January is the peak sales month, plan major marketing campaigns and promotions during this period. For slower months like February, introduce loyalty programs or special discounts to maintain customer engagement.

9) Inventory Management

- Optimize Stock Levels:
  Ensure that high-performing product lines like "Food and Beverages" and "Sports and Travel" are always well-stocked, especially during peak seasons. Use sales data to forecast demand and avoid stockouts.

- Reduce Slow-Moving Inventory:
  For product lines with lower sales, consider reducing inventory levels or introducing promotions to clear out excess stock.

10) Employee Training and Incentives

- Train Staff on Upselling:
  Since Branch C has the highest average revenue per transaction, train staff in other branches on upselling techniques to increase the average transaction value.
- Incentivize Performance:
  Introduce performance-based incentives for employees in underperforming branches to motivate them to improve sales and customer service.

By focusing on the top-performing branches, customer segments, and product lines, the supermarket can maximize revenue and profitability. Targeted marketing campaigns, optimized inventory management, and improved customer experience will further drive growth and ensure long-term success. Regular monitoring of sales trends and customer behavior will allow for continuous improvement and adaptation to changing market conditions.