Supermarket Sales Analysis Project Report

1. Project Overview

The Supermarket Sales Analysis project aims to analyze transactional sales data from a supermarket to uncover meaningful insights that support business decision-making. The project integrates Python (Jupyter Notebook) for data preprocessing and exploratory analysis, SQL for structured querying and KPI calculation, and Power BI for interactive data visualization. By combining these tools, the project demonstrates an end-to-end data analysis workflow, from raw data to actionable insights..

2. Dataset Summary

- Rows: 1000
- Columns: 17
- Key Features:
    - Transaction Details: Invoice ID, Date, Time
    - Location Information: Branch, City
    - Customer Information: Customer Type (Member/Normal), Gender, Rating
    - Product Information: Product Line, Unit Price, Quantity
    - Financial Metrics: Sales (Revenue), Tax (5%), COGS, Gross Income, Gross Margin Percentage
    - Payment Information: Payment Method
    - This rich combination of customer, product, and financial attributes enables comprehensive sales, profitability, and customer behavior analysis.

3. Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

- **Data Loading:** Imported the dataset using Pandas.
- **Initial Exploration:** Used df.info() to check structure and .describe() for summary statistics.

	Unit price	Quantity	Tax 5%	Sales	cogs	gross margin percentage	gross income	Rating
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1.000000e+03	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	307.58738	4.761905e+00	15.379369	6.97270
std	26.494628	2.923431	11.708825	245.885335	234.17651	6.131498e-14	11.708825	1.71858
min	10.080000	1.000000	0.508500	10.678500	10.17000	4.761905e+00	0.508500	4.00000
25%	32.875000	3.000000	5.924875	124.422375	118.49750	4.761905e+00	5.924875	5.50000
50%	55.230000	5.000000	12.088000	253.848000	241.76000	4.761905e+00	12.088000	7.00000
75%	77.935000	8.000000	22.445250	471.350250	448.90500	4.761905e+00	22.445250	8.50000
max	99.960000	10.000000	49.650000	1042.650000	993.00000	4.761905e+00	49.650000	10.00000

- **Column Standardization:** Renamed columns to **snake case** for better readability and documentation.

Specifically renamed the column tax_5% to tax_5_percent to avoid special characters in column names ● **Feature Engineering:**

○ Split the original date column into year, month, and day to support time-based analysis.

- **Data Consistency Check**

Checked for missing values using .isnull() and confirmed that the dataset contains no null values.

```
df.isnull().sum()
```

```
Invoice ID                0
Branch                    0
City                      0
Customer type             0
Gender                    0
Product line              0
Unit price                0
Quantity                  0
Tax 5%                    0
Sales                     0
Date                      0
Time                      0
Payment                   0
cogs                      0
gross margin percentage   0
gross income              0
Rating                    0
dtype: int64
```

Ensured date and time fields were correctly formatted for analysis.

- **Database Integration:** Imported the clean and feature-engineered dataset into the SQL database for SQL analysis.

4. Data Analysis using SQL (Business Transactions)

I performed structured analysis in PostgreSQL to answer key business questions:

1. **Revenue by Product Line** – Analyzed total sales revenue across different product lines to identify the most profitable product categories.

	product_line character varying (100)	revenue numeric
1	Food and beverages	56144.87
2	Sports and travel	55122.89
3	Electronic accessories	54337.55
4	Fashion accessories	54305.89
5	Home and lifestyle	53861.87
6	Health and beauty	49193.81

Total rows: 6 Query complete 00:00:00.230

2. **Revenue by Branch** – Evaluated sales performance across branches to determine which locations contribute the highest revenue.

	branch character varying (10)	revenue numeric
1	Cairo	106197.79
2	Alex	106200.37
3	Giza	110568.72

3. **Monthly Sales Trend & Profit Trend** – Analyzed monthly revenue trends to identify seasonal patterns and sales fluctuations over time.

	month_name character varying (20)	monthly_sales numeric	monthly_profit numeric
1	January	116291.90	5537.74
2	March	109455.57	5212.23
3	February	97219.41	4629.53

4. **Profit by Product Line** – Compared total gross income across product lines to assess profitability beyond just sales volume.

	product_line character varying (100)	profit numeric
1	Food and beverages	2673.59
2	Sports and travel	2624.96
3	Electronic accessories	2587.52
4	Fashion accessories	2585.99
5	Home and lifestyle	2564.81
6	Health and beauty	2342.63

5. **Payment Method Distribution** – Examined transaction counts by payment method to understand customer payment preferences.

	payment character varying (20)	total_sales numeric
1	Credit card	100767.12
2	Ewallet	109993.18
3	Cash	112206.58

6. **Customer Type Revenue Analysis** – Compared total revenue generated by member vs. normal customers to evaluate customer loyalty impact.

	customer_type character varying (20)	gender character varying (10)	total_sales numeric
1	Member	Female	125206.18
2	Normal	Male	63806.30
3	Normal	Female	69465.70
4	Member	Male	64488.70

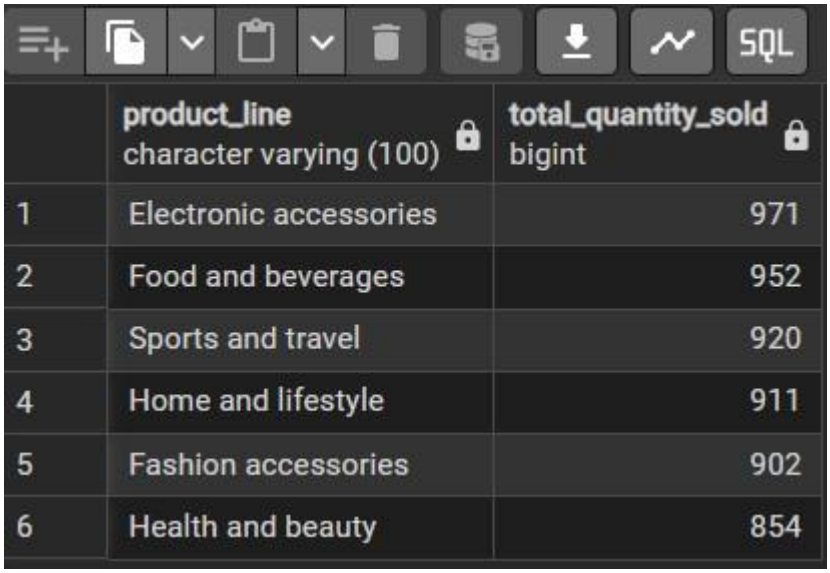
7. **Average Sales per Transaction** – Calculated the average sales value per transaction to understand customer spending behavior.

	avg_sales numeric
1	0.32

8. **Sales by Product Line and Gender** - Analyzed sales distribution across product lines segmented by gender to identify targeted marketing opportunities.

	product_line character varying (100)	gender character varying (10)	total_sales numeric
1	Electronic accessories	Female	31226.72
2	Electronic accessories	Male	23110.83
3	Fashion accessories	Female	33919.65
4	Fashion accessories	Male	20386.24
5	Food and beverages	Male	19679.06
6	Food and beverages	Female	36465.81
7	Health and beauty	Female	23418.76
8	Health and beauty	Male	25775.05
9	Home and lifestyle	Male	20141.32
10	Home and lifestyle	Female	33720.55
11	Sports and travel	Male	19202.50
12	Sports and travel	Female	35920.39

9. **Top-Selling Product Line by Quantity**– Identified product lines with the highest total quantity sold, highlighting high-demand categories.



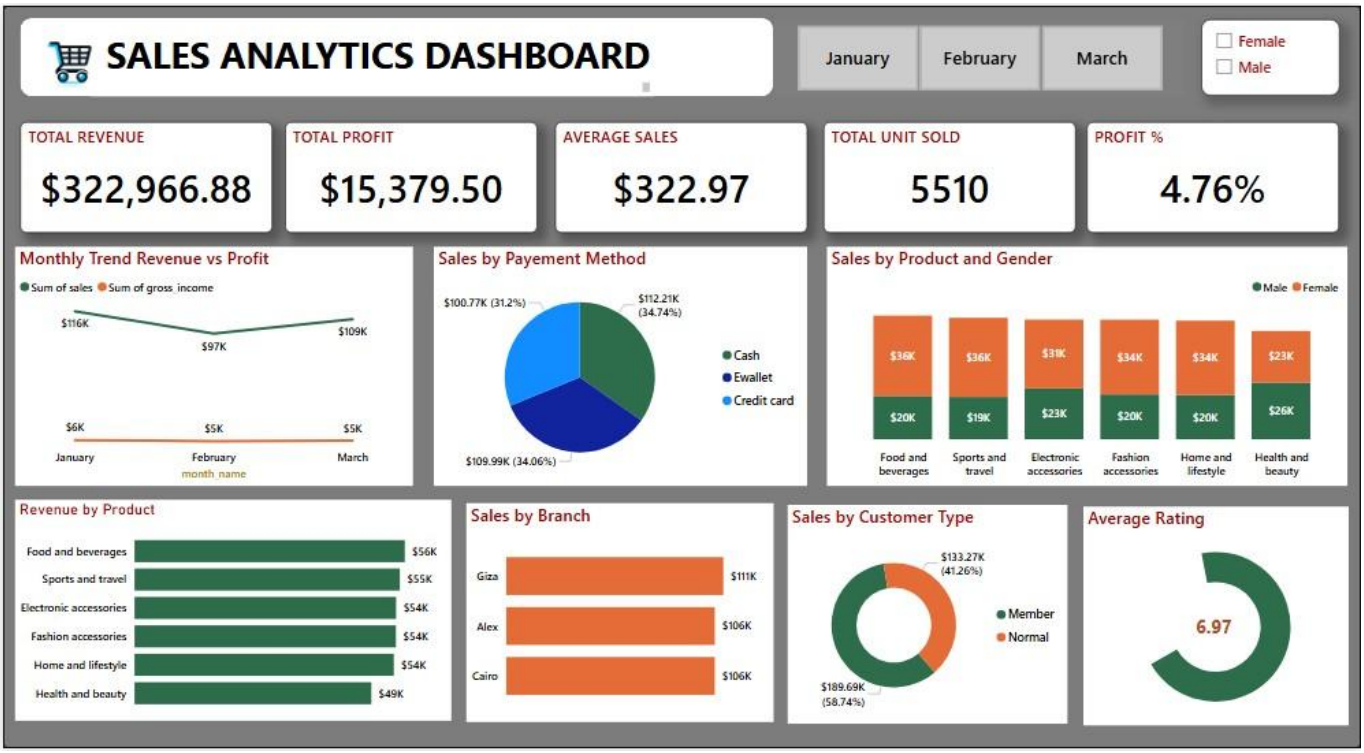
	product_line character varying (100)	total_quantity_sold bigint
1	Electronic accessories	971
2	Food and beverages	952
3	Sports and travel	920
4	Home and lifestyle	911
5	Fashion accessories	902
6	Health and beauty	854

10. **Customer Summary View**.

	customer_type character varying (20)	gender character varying (10)	total_transactions bigint	total_sales numeric	avg_rating numeric
1	Member	Female	356	125206.18	6.89
2	Normal	Male	220	63806.30	7.01
3	Normal	Female	215	69465.70	7.08
4	Member	Male	209	64488.70	6.97

5. Dashboard in Power BI

Finally, I built an interactive dashboard in **Power BI** to present insights visually.



6. Business Recommendations

- Optimize Product Line Strategy – Focus marketing and shelf placement on top-selling and high-profit product lines identified through sales and quantity analysis.

- Branch-Level Performance Improvement – Replicate successful sales strategies from high-performing branches in lower-performing locations to balance overall revenue contribution.

- Enhance Payment Method Offers – Promote popular payment methods through targeted offers to improve transaction convenience and checkout efficiency.

- Seasonal Sales Planning – Use monthly sales trend analysis to plan promotions and inventory for peak demand periods, reducing stock shortages and overstocking.

- Gender-Based Marketing Insights – Tailor marketing campaigns based on observed purchasing patterns between male and female customers to improve conversion rates.

- Profit Margin Monitoring – Regularly monitor gross income and profit rate KPIs to ensure discount strategies do not negatively impact overall profitability.