Project Title: Amazon Sales Analysis Using SQL

Name: Nikhilesh Varasala

Department: AI & Data Science

Project Type: Self-driven

Introduction

The objective of this self-driven project is to analyze Amazon sales data from three branches Mandalay, Yangon, and Naypyitaw to gain insights into product performance, customer behavior, and sales trends. SQL was used to clean the data, engineer new features, and perform exploratory data analysis, enabling actionable insights for strategic business decisions.

Dataset Description

The dataset contains **1000 rows** and **17 columns**, representing sales transactions from January to March 2019. Key columns include:

- invoice_id – unique sales identifier
- branch & city – branch location
- customer_type & gender – customer demographics
- product_line – type of product sold
- unit_price, quantity, total, cogs, VAT, gross_income – sales metrics
- date & time – transaction timing
- rating – customer feedback score

The dataset was initially in Excel format and converted to CSV for import into MySQL Workbench.

Tools Used

- **MySQL Workbench** – for database creation, SQL analysis, and feature engineering
- **Microsoft Excel** – for dataset formatting and CSV conversion
- **Microsoft Word / Google Docs** – for writing this report

Methodology

Data Wrangling

- Imported dataset from CSV into MySQL Workbench.
- Fixed date formatting issues using `clean_date` to ensure SQL functions work correctly.
- Verified row counts and ensured no NULL values in the dataset.

Feature Engineering

- Created `time_of_day` to categorize sales into Morning, Afternoon, and Evening.
- Created `day_name` to identify the weekday of each transaction.
- Created `month_name` to extract the month from the transaction date.

Exploratory Data Analysis (EDA)

EDA was performed to answer the 25 business questions, divided into three areas:

1. **Product Analysis** – understanding top-performing and underperforming products
2. **Sales Analysis** – identifying monthly trends, peak COGS, branch and city performance, time-of-day trends
3. **Customer Analysis** – exploring customer type, gender trends, and ratings

Product Analysis

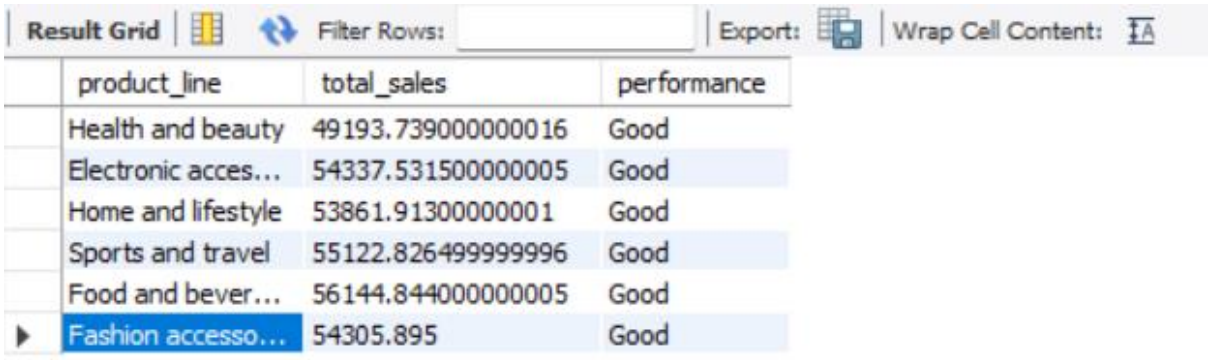
- Product lines with the highest revenue: The product line with the highest revenue is Food and Beverages, generating a total revenue of **56,144.84**

- Product lines classified as Good/Bad based on average sales:

Product Line	Total Sales	Performance
Health and beauty	49,193.73	Good
Food and beverages	56,144.84	Good
Electronic Accessories	54,337.53	Good
Home and Lifestyle	53,861.91	Good
Sports and Travel	55,122.82	Good
Fashion Accessories	54,305.89	Good

All product lines recorded total sales above the dataset's average sales value; hence, they were classified as 'Good'.

- Product line with the highest VAT: The product line that incurred the highest Value Added Tax (VAT) is Food and Beverages, with a total VAT contribution of **2,673.56**



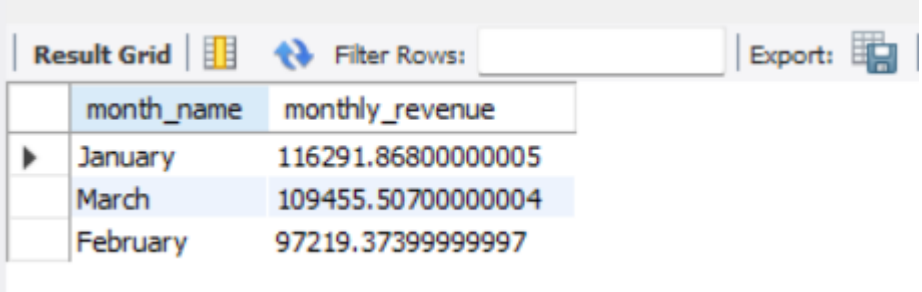
product_line	total_sales	performance
Health and beauty	49193.739000000016	Good
Electronic acces...	54337.531500000005	Good
Home and lifestyle	53861.91300000001	Good
Sports and travel	55122.826499999996	Good
Food and bever...	56144.844000000005	Good
Fashion accesso...	54305.895	Good

Fig1: SQL output showing revenue generated by each product line along with Good/Bad performance classification based on average sales

Sales Analysis

- **Highest sales month**: The highest sales were recorded in January, with a total revenue of **1,16,291.86**.
- **Branch or city with highest revenue**: Naypyitaw emerged as the top-performing city, generating the highest revenue of **1,10,568.70** among all branches.
- **Peak COGS month**: The cost of goods sold (COGS) peaked in January, reaching a total value of **1,10,754.16**.
- **Time of day with most sales**: Most sales transactions occurred during the afternoon, indicating it as the busiest time of the day.
- **Monthly revenue trends**:

Month	**Revenue**
January	1,16,291.86
February	1,09,455.50
March	97,219.37



Result Grid | Filter Rows: | Export:

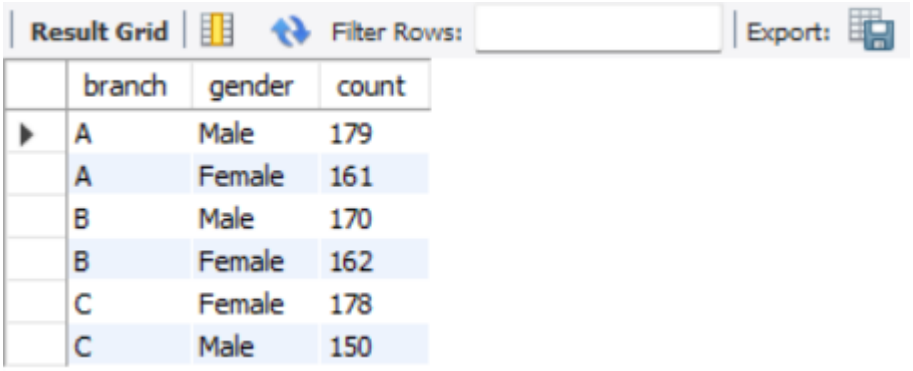
month_name	monthly_revenue
January	116291.86800000005
March	109455.50700000004
February	97219.37399999997

Fig2: SQL output showing monthly revenue trends across the three-month period.

Customer Analysis

- **Customer type contributing most revenue**: Member customers contributed the highest revenue, with total sales amounting to **1,64,223.44**.
- **Predominant gender among customers**: Female customers formed the predominant gender group, accounting for **501** transactions.
- **Customer type with highest purchase frequency**: Member customers also recorded the highest purchase frequency with **501** transactions, indicating stronger engagement compared to normal customers.
- **Gender distribution per branch**:

Branch	Male	Female
Mandalay	170	162
Yangon	179	161
Naypyitaw	150	178



Result Grid | Filter Rows: | Export:

branch	gender	count
A	Male	179
A	Female	161
B	Male	170
B	Female	162
C	Female	178
C	Male	150

Fig3: Gender-wise distribution of customers across branches

Branch codes A, B, and C represent Yangon, Mandalay, and Naypyitaw respectively

Key Insights

- Certain product lines consistently generate higher revenue and VAT.
- Sales peak during specific times of day and days of the week.
- Member customers contribute higher overall revenue.
- One city outperforms the others in total revenue generation.
- Customer ratings vary significantly by branch and time-of-day.
- Branches can optimize staffing and inventory based on peak sales periods.

Conclusion

This self-driven project demonstrates the ability to clean and analyze real-world sales data using SQL. The findings provide actionable insights on product performance, customer behavior, and sales trends across Amazon branches, which can help in making informed business decisions.

Future Scope

- Integrate visualization tools like **Power BI** or **Tableau** for interactive dashboards.
- Extend analysis using **predictive modeling** to forecast sales trends.
- Explore **sentiment analysis** on customer reviews to enhance product strategies.