Amazon Sales Data Analysis

1. Introduction

Purpose of the Capstone Project

The major aim of this project is to gain insights into Amazon's sales data to understand the factors affecting sales across different branches. This analysis will help evaluate performance, uncover trends, and provide actionable recommendations.

2. Dataset Description

About the Data

The dataset contains sales transactions from three Amazon branches located in Mandalay, Yangon, and Naypyitaw. It includes 17 columns and 1000 rows of data.

Column	Description	Data Type
invoice_id	Invoice of the sales made	VARCHAR(30)
branch	Branch where the sales were made	VARCHAR(5)
city	Location of the branch	VARCHAR(30)
customer_type	Type of customer	VARCHAR(30)
gender	Gender of the customer making the purchase	VARCHAR(10)
product_line	Product line of the product sold	VARCHAR(100)
unit_price	Price of each product	DECIMAL(10, 2)
quantity	Amount of the product sold	INT
VAT	Amount of tax on the purchase	FLOAT(6, 4)
total	Total cost of the purchase	DECIMAL(10, 2)
date	Date of the purchase	DATE
time	Time of the purchase	TIMESTAMP
payment_method	Payment method used	DECIMAL(10, 2)

Column	Description	Data Type
cogs	Cost of Goods Sold	DECIMAL(10, 2)
gross_margin_percentage	Gross margin percentage	FLOAT(11, 9)
gross_income	Gross Income	DECIMAL(10, 2)
rating	Rating given by the customer	FLOAT(2, 1)

3. Analysis List

Product Analysis

- Understand different product lines and identify the best-performing and underperforming product lines.

Sales Analysis

- Analyze sales trends to measure the effectiveness of sales strategies and identify areas for improvement.

Customer Analysis

- Uncover customer segments, purchase trends, and the profitability of each customer segment.

4. Approach Used

Data Wrangling

1. **Build a Database:** Create a database to store the data.
2. **Create Tables and Insert Data:** Define table schema and insert the dataset into the database.
3. **Check for Null Values:** Ensure that there are no null values as NULL constraints are applied.

Feature Engineering

1. **timeofday Column:** Categorize sales into Morning, Afternoon, and Evening to analyze sales patterns throughout the day.
2. **dayname Column:** Extract and add day names (e.g., Mon, Tue) to analyze weekday sales trends.

3. **monthname Column:** Extract and add month names (e.g., Jan, Feb) to identify monthly sales patterns.

Exploratory Data Analysis (EDA)

- Analyze data to answer the following business questions:

5. Business Questions to Answer

1. City Analysis

- What is the count of distinct cities in the dataset?
- For each branch, what is the corresponding city?

2. Product Analysis

- What is the count of distinct product lines in the dataset?
- Which product line has the highest sales?
- Which product line generated the highest revenue?
- Which product line incurred the highest VAT?
- For each product line, add a column indicating "Good" if its sales are above average, otherwise "Bad."
- Which product line is most frequently associated with each gender?

3. Sales Analysis

- Which payment method occurs most frequently?
- How much revenue is generated each month?
- In which month did the cost of goods sold reach its peak?
- Identify the branch that exceeded the average number of products sold.
- Count the sales occurrences for each time of day on every weekday.

4. Customer Analysis

- Identify the customer type contributing the highest revenue.
- Determine the city with the highest VAT percentage.
- Identify the customer type with the highest VAT payments.

- What is the count of distinct customer types in the dataset?
- What is the count of distinct payment methods in the dataset?
- Which customer type occurs most frequently?
- Identify the customer type with the highest purchase frequency.
- Determine the predominant gender among customers.
- Examine the distribution of genders within each branch.
- Identify the time of day when customers provide the most ratings.
- Determine the time of day with the highest customer ratings for each branch.
- Identify the day of the week with the highest average ratings.
- Determine the day of the week with the highest average ratings for each branch.

6. Conclusion

Summarize the insights gained from the analysis, including key findings, trends, and recommendations. Mention any limitations and suggest areas for further analysis if applicable.

7. Appendices

Additional Data

- Include detailed SQL queries or code snippets used in the analysis.

References

- Cite any sources or references used in the project.