SQL CAPSTONE PROJECT

Amazon Sales Data Analysis

1. Introduction

This document provides a comprehensive analysis of Amazon sales data, focusing on product performance, sales trends, and customer behaviour. The analysis aims to uncover insights that can guide business decisions, improve sales strategies, and enhance customer satisfaction.

The dataset includes transactions from three Amazon branches located in Mandalay, Yangon, and Naypyitaw.

The data contains 17 columns and 1000 rows

2. Dataset Overview

The Amazon sales dataset includes the following columns:

- **Invoice ID**
- **Branch**
- **City**
- **Customer type**
- **Gender**
- **Product line**
- **Unit price**
- **Quantity**
- **Tax 5%**
- **Total**
- **Date**
- **Time**
- **Payment**
- **cogs (Cost of Goods Sold)**
- **gross margin percentage**
- **gross income**
- **Rating**

3. Data Preparation

Before performing analysis, several data preparation steps were taken:

Conversion of Columns to Appropriate Data Types:

- Converted the Date column to DATE.
- Converted the Time column to TIME.
- Converted numerical columns like Unit price, Quantity, Tax 5%, Total, gross margin percentage, gross income, and Rating to appropriate decimal formats.

Feature Engineering:

- Added a new column time_ of_ day to categorize the time of purchase into "Morning", "Afternoon", and "Evening".
- Added columns day _name to extract the day of the week from the Date column.
- Added columns month_ name to extract the month_ name from the Date column.

4. Exploratory Data Analysis (EDA)

The following questions I explored to understand the dataset better:

1. Count of Distinct Cities in the Dataset

```
SELECT COUNT (DISTINCT city) FROM amazon;
```

There are 3 distinct cities in the dataset: Yangon, Naypyitaw, and Mandalay.

2. For each branch, what is the corresponding city

```
SELECT Branch, City
FROM amazon
GROUP BY Branch, City;
```

Branch A is located in Yangon.

Branch B is located in Mandalay.

Branch C is located in Naypyitaw.

3. Count of Distinct Product Lines in the Dataset

```
SELECT COUNT (DISTINCT `Product line`) AS distinct_ product_ lines
FROM amazon;
```

The dataset contains 6 distinct product lines

4. most frequent payment method

```
SELECT `Payment`, COUNT (*) AS frequency
FROM amazon
GROUP BY `Payment`
limit 1;
```

The most frequent payment method is Ewallet, with a frequency of 345.

5.Product Line with the Highest Sales

```
select `Product line`,
SUM(Quantity) AS total_sales
FROM amazon
GROUP BY `Product line`
ORDER BY total_sales DESC
LIMIT 1;
```

Electronic Accessories has the highest sales, with total sales being the highest among all product lines.

6.Revenue Generated Each Month & year

```
SELECT
    MONTH(Date) AS month,
    YEAR(Date) AS year,
    SUM(Total) AS total_revenue
FROM amazon
GROUP BY month, year
ORDER BY year, month;
```

January generated the highest revenue: 116,292.11.

February generated a revenue of 97,219.58.

March generated a revenue of 109,455.74.

7.Month When the Cost of Goods Sold (COGS) Reached Its Peak

```
SELECT
    MONTH(Date) AS month,
    YEAR(Date) AS year,
    SUM(cogs) AS total_cogs
FROM amazon
GROUP BY month, year
ORDER BY total_cogs DESC
LIMIT 1;
```

The COGS reached its peak in January

8.Product Line that Generated the Highest Revenue

```
SELECT `Product line`, SUM(Total) AS total_revenue
FROM amazon
GROUP BY `Product line`
ORDER BY total_revenue DESC
LIMIT 1;
```

Food and Beverages generated the highest revenue: 56,144.96.

9.City with the Highest Revenue

```
SELECT City, SUM(Total) AS total_revenue
FROM amazon
GROUP BY City
ORDER BY total_revenue DESC
LIMIT 1;
```

Naypyitaw is the city with the highest revenue: 110,568.86.

10.Product Line with the Highest Value Added Tax (VAT)

```
SELECT `Product line`, SUM(`Tax 5%`) AS total_vat
FROM amazon
GROUP BY `Product line`
ORDER BY total_vat DESC
LIMIT 1;
```

Food and Beverages has the highest VAT: 2,673.68.

11.Label Each Product Line as "Good" or "Bad" Based on Sales

```
SELECT `Product line`,
      CASE
         WHEN SUM(Quantity) > (SELECT AVG(Quantity) FROM amazon) THEN 'Good'
         ELSE 'Bad'
      END AS performance
FROM amazon
GROUP BY `Product line`;
```

All product lines are labeled as "Good" based on sales.

12.Branch That Exceeded the Average Number of Products Sold

```
SELECT Branch, SUM(Quantity) AS total_quantity
FROM amazon
GROUP BY Branch
HAVING total_quantity > (SELECT AVG(Quantity) FROM amazon);
```

Branches A, B, and C all exceeded the average number of products sold, with Branch A leading with 1,859 units sold.

13.Most Frequent Product Line Associated with Each Gender

```
SELECT Gender, `Product line`, COUNT (*) AS frequency
FROM amazon
GROUP BY Gender, `Product line`
ORDER BY Gender, frequency DESC;
```

Females most frequently purchased Fashion Accessories.
Males most frequently purchased Health and Beauty.

14.Average Rating for Each Product Line

```
SELECT `Product line`, AVG(Rating) AS average_rating
FROM amazon
GROUP BY `Product line`;
```

Food and Beverages and Fashion Accessories have the highest average rating of 7.11322.

15.Count Sales Occurrences for Each Time of Day on Every Weekday

```
SELECT `day_name`, `time_of_day`, COUNT(*) AS sales_count
FROM amazon
GROUP BY `day_name`, `time_of_day`
ORDER BY `day_name`, `time_of_day`;
```

Saturday evening has the highest sales count: 136.

16.Customer Type Contributing the Highest Revenue

```
SELECT `Customer type`, SUM(Total) AS total_revenue
FROM amazon
GROUP BY `Customer type`
ORDER BY total_revenue DESC
LIMIT 1;
```

Members contribute the highest revenue: 164,223.81.

17.City with the Highest VAT Percentage

```
SELECT City, AVG(`Tax 5%`) AS average_vat
FROM amazon
GROUP BY City
ORDER BY average_vat DESC
LIMIT 1;
```

Naypyitaw has the highest average VAT percentage: 16.052835.

18.Customer Type with the Highest value added tax(VAT) Payments

```
SELECT `Customer type`, SUM(`Tax 5%`) AS total_vat
FROM amazon
GROUP BY `Customer type`
ORDER BY total_vat DESC
LIMIT 1;
```

Members have the highest VAT payments: 7,820.53.

19. Count of Distinct Customer Types in the Dataset

```
SELECT COUNT (DISTINCT `Customer type`) AS distinct_customer_types
FROM amazon;
select distinct `Customer type` from amazon;
```

There are 2 distinct customer types in the dataset: Member and Normal.

20.Count of Distinct Payment Methods in the Dataset

```
SELECT COUNT (DISTINCT `Payment`) AS distinct_payment_methods
FROM amazon;
```

there are 2 distinct payment methods in the dataset.

```
select distinct `Payment` from amazon;
```

There are 3 distinct payment methods: Ewallet, Cash, and Credit Card.

21.Customer Type with the Highest Purchase Frequency

```
SELECT `Customer type`, COUNT (*) AS purchase_count
FROM amazon
GROUP BY `Customer type`
ORDER BY purchase_count DESC
LIMIT 1;
```

Members have the highest purchase frequency with 501 purchases.

22.Predominant Gender Among Customers

```
SELECT Gender, COUNT(*) AS purchasing_count
FROM amazon
GROUP BY Gender
```

```
ORDER BY purchasing_count DESC
LIMIT 1;
```

Female customers are the predominant gender with 501 purchases.

```
SELECT Gender, COUNT(*) AS purchasing_count
FROM amazon
GROUP BY Gender
ORDER BY purchasing_count DESC
LIMIT 2;
```

female purchasing_counts-501

male purchasing_counts-499

23.Distribution of Genders Within Each Branch

```
SELECT Branch, Gender, COUNT(*) AS frequency
FROM amazon
GROUP BY Branch, Gender
ORDER BY Branch, frequency DESC;
```

Branch A: More Males (179) than Females (161).

Branch B: More Males (170) than Females (162).

Branch C: More Females (178) than Males (150).

24.Time of Day When Customers Provide the Most Ratings

```
SELECT `time_of_day`, AVG(Rating) AS average_rating
FROM amazon
GROUP BY `time_of_day`
ORDER BY average_rating DESC
limit 1;
```

The evening has the highest customer ratings with an average rating of 6.97553.

```
SELECT `time_of_day`, AVG(Rating) AS average_rating
FROM amazon
GROUP BY `time_of_day`
ORDER BY average_rating DESC;
```

The evening has been getting most rating-6.97553.

The morning has been getting most rating-6.96073.

26.Time of Day with the Highest Customer Ratings for Each Branch

```
SELECT Branch, `time_of_day`, AVG(Rating) AS average_rating
FROM amazon
GROUP BY Branch, `time_of_day`
ORDER BY Branch, average_rating DESC;
```

Branch C (Naypyitaw) has the highest customer ratings in the evening.

27.Day of the Week with the Highest Average Ratings

```
SELECT `day_name`, AVG(Rating) AS average_rating
FROM amazon
GROUP BY `day_name`
ORDER BY average_rating DESC;
```

Monday has the highest average ratings.

28.Day of the Week with the Highest Average Ratings for Each Branch

```
SELECT Branch, `day_name`, AVG(Rating) AS average_rating
FROM amazon
GROUP BY Branch, `day_name`
ORDER BY Branch, average_rating DESC;
```

Product Analysis

This section delves into the performance of different product lines:

1. **Distribution of Product Lines:**

   Query: SELECT Product line, COUNT(*) AS sales_count FROM amazon GROUP BY Product line ORDER BY sales_count DESC;

Food and beverages	174
Electronic accessories	170
Sports and travel	166
Home and lifestyle	160
Health and beauty	152

2. **Revenue by Product Line:**

   Query: SELECT Product line, SUM(Total) AS total_revenue FROM amazon GROUP BY Product line ORDER BY total_revenue DESC;

   'Food and Beverages' generates the highest revenue.

3. **Product Lines Needing Improvement:**

   Query: SELECT Product line, SUM(Total) AS total_revenue, COUNT(*) AS sales_count FROM amazon GROUP BY Product lineHAVING total_revenue < (SELECT AVG(Total) FROM amazon) OR sales_count < (SELECT AVG(sales_count) FROM (SELECT COUNT(*) AS sales_count FROM amazon GROUP BY Product line) AS subquery);

   Health and beauty

   Home and lifestyle

   Sports and travel

Sales Analysis

This analysis provides insights into the sales performance:

1. **Monthly Sales Trends:**

   Query: SELECT MONTH(Date) AS month, YEAR(Date) AS year, SUM(Total) AS total_revenue FROM amazon GROUP BY year, month ORDER BY year, month;

   In January 2019, the total revenue generated was 116,292.11.

   In February 2019, the total revenue generated was 97,219.58.

   In March 2019, the total revenue generated was 109,455.74.

2. **Sales by Payment Method:**

   Query: SELECT Payment, SUM(Total) AS total_revenue FROM amazon GROUP BY Payment ORDER BY total_revenue DESC;

   'Ewallet' is the most frequently used payment method.

3. **Peak Sales Times:**

   Query: SELECT Time, SUM(Total) AS total_revenue FROM amazon GROUP BY Time ORDER BY total_revenue DESC;

   The highest transaction occurred at 14:42 on 2024-08-25, with a value of 2,534.65

Customer Analysis

This section uncovers insights related to customer behaviour:

Customer Segment Analysis:

Query: SELECT Customer type, COUNT(*) AS sales_count, SUM(Total) AS total_revenue FROM amazon GROUP BY Customer type ORDER BY total_revenue DESC;

'Members' contribute the highest revenue.

Customer Gender Analysis:

Query: SELECT Gender, COUNT(*) AS sales_count, SUM(Total) AS total_revenue FROM amazon GROUP BY Gender ORDER BY total_revenue DESC;

Female customers make more purchases than male customers.

Most Profitable Customer Segment:

Query: SELECT Customer type, SUM(Total) AS total_revenue FROM amazon GROUP BY Customer type ORDER BY total_revenue DESC LIMIT 1;

The 'Member' customer type is the most profitable.

Conclusion

Through comprehensive data exploration and analysis, several key findings have been identified:

1. **City and Branch Performance:**

   Naypyitaw emerged as the city with the highest revenue, while Branch C located in Naypyitaw was particularly strong in terms of both revenue generation and customer satisfaction, especially during the evening.

   All branches exceeded the average number of products sold, with Branch A leading in total units sold.

2. **Product Line Analysis:**

   Food and Beverages not only generated the highest revenue but also had the highest VAT(value added tax) contributions, indicating its strong market presence.

   Electronic Accessories had the highest total sales, reflecting its popularity among customers.

   Health and Beauty, Home and Lifestyle, and Sports and Travel product lines showed room for improvement in both revenue and sales counts, highlighting potential areas for strategic enhancement.

3. **Customer Behaviour:**

   Members contributed the highest revenue and had the highest purchase frequency, indicating that loyal customers are key revenue drivers.

   Female customers slightly outnumbered male customers in terms of purchase frequency, though both genders showed strong engagement across different product lines.

4. **Sales Trends:**

   The sales were highest in January 2019, with a notable peak in revenue generation during this period. However, February saw a dip, followed by a recovery in March.

   The most frequently used payment method was Ewallet, emphasizing the importance of digital payment options for customers.

   Peak sales occurred during the early afternoon, particularly around 14:42, which had the highest transaction value recorded in the dataset.

5. **Customer Ratings:**

   Evening hours were associated with the highest average customer ratings, suggesting that customers were more satisfied with their purchases during this time of day.

   Monday had the highest average ratings, which could indicate a positive customer experience at the beginning of the week.

Recommendations

Based on the insights derived from the analysis, the following recommendations are proposed:

1. **Strengthen Underperforming Product Lines:**

   Focus on improving the performance of Health and Beauty, Home and Lifestyle, and Sports and Travel product lines through targeted marketing campaigns, promotional offers, or product diversification.

2. **Enhance Customer Experience in Naypyitaw:**

   Given that Naypyitaw generates the highest revenue, consider enhancing customer experience in this city by introducing loyalty programs, personalized offers, and optimizing inventory based on customer preferences.

3. **Capitalize on Digital Payments:**

   Continue to promote and possibly expand digital payment options like Ewallet, as it is the most frequently used method. This could include offering incentives for customers who use digital payments.

4. **Targeted Marketing for Peak Sales Hours:**

   Leverage the insights around peak sales times by running time-specific promotions during early afternoons to maximize sales and customer engagement.

5. **Improve Customer Satisfaction in Branch B:**

   Focus on enhancing customer satisfaction and ratings in Branch B (Mandalay), particularly during the evening hours, to bring it on par with the high ratings seen in Branch C.

6. **Focus on Member Retention and Growth:**

   Since Members are the most profitable customer segment, invest in strategies to retain existing members and convert new customers into members. This could include exclusive deals, membership benefits, and personalized shopping experiences.

7. **Monitor and Boost Sales During Low Revenue Months:**

   Analyse the factors contributing to lower revenues in February and develop strategies to boost sales during this month, such as special offers or discounts.

Reference file

Author

k. Sai Madhu Sri