**Project Report: Amazon Sales Data Analysis**

**1. Introduction**

**1.1 Project Overview**

The objective of this project is to gain insights into the sales performance of Amazon across its branches in Mandalay, Yangon, and Naypyitaw. By analyzing the dataset, we aim to understand factors influencing sales, customer behavior, and product performance. The key aspects to analyze include sales trends, customer segmentation, and product performance.

**1.2 Purpose of the Project**

The primary goal of this project is to:

* Assess product sales performance across branches.
* Identify customer trends and segmentation.
* Understand the impact of various factors like time, gender, customer type, and payment methods on sales.
* Make data-driven recommendations for sales strategy improvements.

**1.3 Dataset Description**

The dataset consists of 17 columns and 1000 rows of sales transactions, including the following information:

* Invoice ID, branch, city, customer type, gender, product line, unit price, quantity, VAT, total cost, payment method, and more.
* The data is structured to allow for in-depth analysis of sales, customer behaviors, and product performance.

**2. Data Preparation**

**2.1 Database and Table Creation**

To start the analysis, a database was created, and a sales table was set up to store the data:

sql

Copy code

CREATE DATABASE amazon\_sales;

USE amazon\_sales;

CREATE TABLE sales (

invoice\_id VARCHAR(30),

branch VARCHAR(5),

city VARCHAR(30),

customer\_type VARCHAR(30),

gender VARCHAR(10),

product\_line VARCHAR(100),

unit\_price DECIMAL(10, 2),

quantity INT,

VAT FLOAT(6, 4),

total DECIMAL(10, 2),

date DATE,

time TIMESTAMP,

payment DECIMAL(10, 2),

cogs DECIMAL(10, 2),

gross\_margin\_percentage FLOAT(11, 9),

gross\_income DECIMAL(10, 2),

rating FLOAT(2, 1)

);

**2.2 Data Cleaning**

* **Null values** were inspected and handled appropriately during table creation by using NOT NULL constraints.
* **Data Wrangling** was conducted to clean and prepare the dataset for analysis. Missing values were handled using methods like data replacement.

**3. Feature Engineering**

**3.1 Time of Day Analysis**

A new column, timeofday, was added to categorize sales based on the time of day (Morning, Afternoon, Evening):

ALTER TABLE amazon\_sales1

ADD timeofday VARCHAR(50);

UPDATE amazon\_sales

SET timeofday =

CASE

WHEN TIME(Time) BETWEEN '00:00:01' AND '10:59:59' THEN 'Morning'

WHEN TIME(Time) BETWEEN '11:00:00' AND '15:59:59' THEN 'Afternoon'

ELSE 'Evening'

END;

**3.2 Day of the Week and Month Analysis**

Additional columns, dayname and monthname, were added to help analyze trends based on the day of the week and month:

**Weekday**

alter table amazon\_sales

add dayname varchar(30);

UPDATE amazon\_sales

SET dayname = DAYNAME(Date);

**Month Name**

alter table amazon\_sales

add monthname varchar(30);

UPDATE amazon\_sales

SET monthname = MONTHNAME(Date);

**4. Exploratory Data Analysis (EDA)**

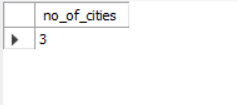
**4.1 Business Questions and SQL Queries**

Several business questions were addressed through SQL queries. Below are the key questions and findings:

**4.1.1 What is the count of distinct cities in the dataset?**

select count(distinct city) as no\_of\_cities

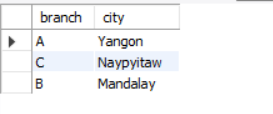
from amazon\_sales;



**4.1.2 For each branch, what is the corresponding city?**

select distinct(branch), city

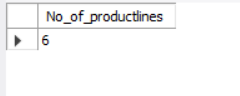
from amazon\_sales;



**4.1.3 What is the count of distinct product lines in the dataset?**

select count(distinct product\_line) as No\_of\_productlines

from amazon\_sales;



**4.1.4 Which payment method occurs most frequently?**

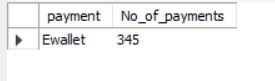
select payment, count(\*) as No\_of\_payments

from amazon\_sales

group by payment

order by No\_of\_payments desc

limit 1 ;



**4.1.5 Which product line has the highest sales?**

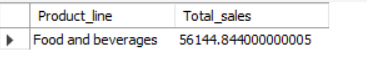
select Product\_line, sum(Total) as Total\_sales

from amazon\_sales

group by Product\_line

order by Total\_sales desc

limit 1 ;

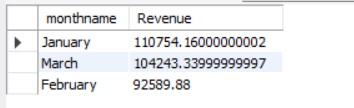
****

**4.1.6 How much revenue is generated each month?**

select monthname, sum(cogs) as Revenue

from amazon\_sales

group by monthname;



**4.1.7 In which month did the cost of goods sold reach its peak?**

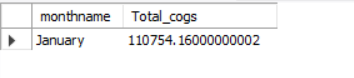
select monthname, sum(cogs) as Total\_cogs

from amazon\_sales

group by monthname

order by Total\_cogs desc

limit 1;



**4.1.8 Which product line generated the highest revenue?**

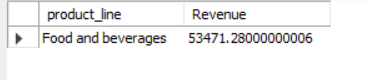
select product\_line, sum(cogs) as Revenue

from amazon\_sales

group by product\_line

order by Revenue desc

limit 1;



**4.1.9 In which city was the highest revenue recorded?**

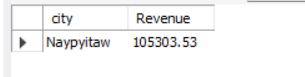
select city, sum(cogs) as Revenue

from amazon\_sales

group by city

order by Revenue desc

limit 1;



**4.1.10 Which product line incurred the highest Value Added Tax?**

select product\_line, sum(Vat) as Total\_vat

from amazon\_sales

group by product\_line

order by Total\_vat desc

limit 1;



**4.1.11 For each product line, add a column indicating "Good" if its sales are above average, otherwise "Bad."**

with cte1 as(

select product\_line, avg(Total) as avg\_sales

from amazon\_sales

group by Product\_line)

select amazon\_sales.\*, avg\_sales,

case

when Total > avg\_sales then "Good"

else "Bad"

end as Sales\_performance

from amazon\_sales inner join cte1

on amazon\_sales.Product\_line = cte1.Product\_line;

**4.1.12 Identify the branch that exceeded the average number of products sold.**

with cte1 as(

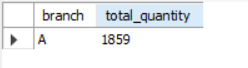
select branch, sum(quantity) as total\_quantity

from amazon\_sales

group by branch)

select \* from cte1

where total\_quantity > (select avg(total\_quantity) from cte1);



**4.1.13 Which product line is most frequently associated with each gender?**

with cte1 as(

select Gender, product\_line, count(\*) as frequency,

dense\_rank() over(partition by Gender order by count(\*) desc) as rn

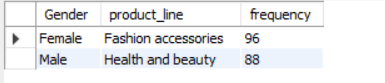
from amazon\_sales

group by Gender, product\_line)

select Gender, product\_line, frequency

from cte1

where rn = 1;

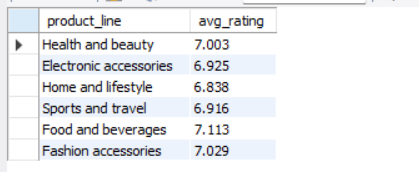


**4.1.14 Calculate the average rating for each product line.**

select product\_line, Round(avg(rating),3) as avg\_rating

from amazon\_sales

group by product\_line;



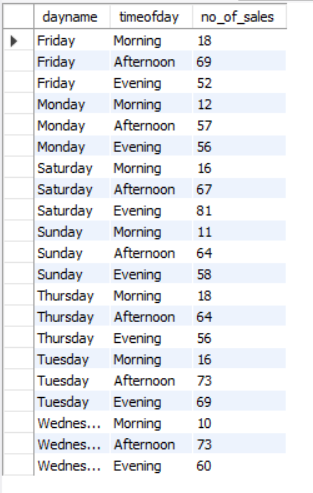
**4.1.15 Count the sales occurrences for each time of day on every weekday.**

select dayname, timeofday, count(\*) no\_of\_sales

from amazon\_sales

group by dayname, timeofday

order by dayname, field(timeofday, 'Morning', 'Afternoon', 'Evening');



**4.1.16 Identify the customer type contributing the highest revenue.**

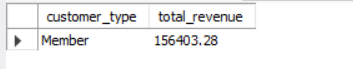
select customer\_type, Round(sum(cogs), 2) as total\_revenue

from amazon\_sales

group by Customer\_type

order by total\_revenue desc

limit 1;



**4.1.17 Determine the city with the highest VAT percentage.**

with cte1 as (

select city, sum(unit\_price \* quantity) as Product\_cost, sum(vat) as tax

from amazon\_sales

group by city),

cte2 as(

select city, ((tax/Product\_cost)\*100) as tax\_percentage,

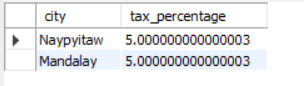
dense\_rank() over(order by (tax/Product\_cost)\*100 desc) as rn

from cte1)

select city, tax\_percentage

from cte2

where rn = 1;



**4.1.18 Identify the customer type with the highest VAT payments.**

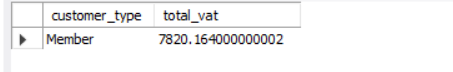
select customer\_type, sum(vat) as total\_vat

from amazon\_sales

group by customer\_type

order by total\_vat desc

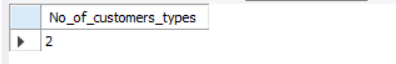
limit 1;



**4.1.19 What is the count of distinct customer types in the dataset?**

select count(distinct customer\_type) as No\_of\_customers\_types

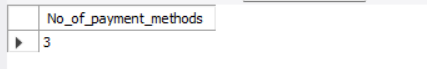
from amazon\_sales;



**4.1.20 What is the count of distinct payment methods in the dataset?**

select count(distinct payment) as No\_of\_payment\_methods

from amazon\_sales;



**4.1.21 Which customer type occurs most frequently?**

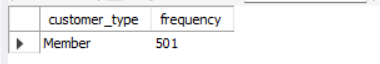
select customer\_type, count(\*) as frequency

from amazon\_sales

group by customer\_type

order by frequency desc

limit 1;



**4.1.22 Identify the customer type with the highest purchase frequency.**

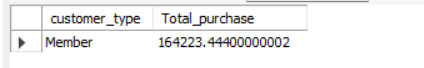
select customer\_type, sum(total) as Total\_purchase

from amazon\_sales

group by customer\_type

order by Total\_purchase desc

limit 1;



**4.1.23 Determine the predominant gender among customers.**

select gender, count(\*) as no\_of\_customers

from amazon\_sales

group by gender

order by no\_of\_customers desc

limit 1;



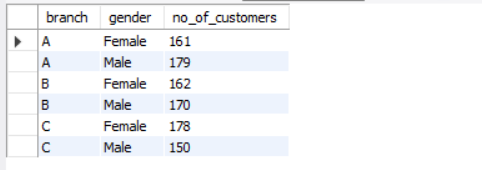
**4.1.24 Examine the distribution of genders within each branch.**

select branch, gender, count(\*) as no\_of\_customers

from amazon\_sales

group by branch, gender

order by branch, gender;



**4.1.25 Identify the time of day when customers provide the most ratings.**

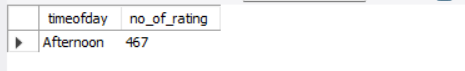
select timeofday, count(rating) as no\_of\_rating

from amazon\_sales

group by timeofday

order by no\_of\_rating desc

limit 1;



**4.1.26 Determine the time of day with the highest customer ratings for each branch.**

select branch, timeofday, no\_of\_rating

from(

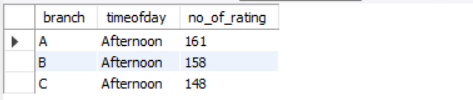
select branch, timeofday, count(rating) as no\_of\_rating,

dense\_rank() over(partition by branch order by count(rating) desc) as rn

from amazon\_sales

group by branch, timeofday) x1

where rn = 1;



**4.1.27 Identify the day of the week with the highest average ratings.**

select dayname, round(avg(rating),3) as avg\_rating

from amazon\_sales

group by dayname

order by avg\_rating desc

limit 1;



**4.1.28 Determine the day of the week with the highest average ratings for each branch.**

select branch, dayname, avg\_rating

from

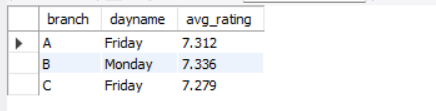
(select branch, dayname, round(avg(rating),3) as avg\_rating,

dense\_rank() over(partition by branch order by round(avg(rating),3) desc) as rn

from amazon\_sales

group by branch, dayname) x1

where rn = 1;



**5. Key Findings**

**5.1 Product Analysis**

* The Food and Beverages product line generated the highest sales across all Products.
* Food and Beverages product line is generating highest VAT.

**5.2 Sales Analysis**

* **Ewallet** is the most frequent payment method used by customers.
* **Members** contribute to highest sales.
* The majority of sales occur in the **Afternoon** and **Evening**.
* Cost of goods sold reached peak in January followed by March.
* **Naypyitaw** city generates highest revenue.

**5.3 Customer Analysis**

* **Female** customers are more interested in Fashion accessories.
* **Male** customers are more interested in Health and beauty.
* Most frequent customers are Female then male with slight difference.

**6. Recommendations**

**6.1 For Product Line Improvement**

* Continue to invest in Food and Beverages, as it consistently shows strong performance.
* Introduce new products in Food and Beverages product line.

**6.2 For Sales Strategy**

* **Naypyitaw** consistently generates the highest revenue. Expand promotional activities in other cities.
* The peak times for sales are in the **Afternoon** and **Evening**; therefore, optimize store operations and staffing during these hours.

**6.3 For Customer Engagement**

* **Female** customersare doing more sales, increace promotional activities more focused on female customers.
* Increase promotional activities to attract more customers to join as members.

**7. Conclusion**

This project successfully analyzed Amazon’s sales data to uncover valuable insights into sales trends, customer behavior, and product performance. By using SQL queries to explore the data, we were able to identify key trends that can help shape future sales strategies and improve customer engagement across different branches. Based on the analysis, we provided recommendations for improving product lines, optimizing customer engagement, and refining sales strategies.