Amazon Sales Analysis

amazon

Purposes Of The Project:

The major aim of this project is to gain insight into the sales data of Amazon to understand the different factors that affect sales of the different cities and their corrosponding branches of Amazon located in Myanmar, using MySQL. The project aims to uncover insights into sales trends, customer behavior, and product performance. By leveraging the power of MySQL queries, we extract meaningful statistics and patterns from complex datasets.

amazon

About Data:

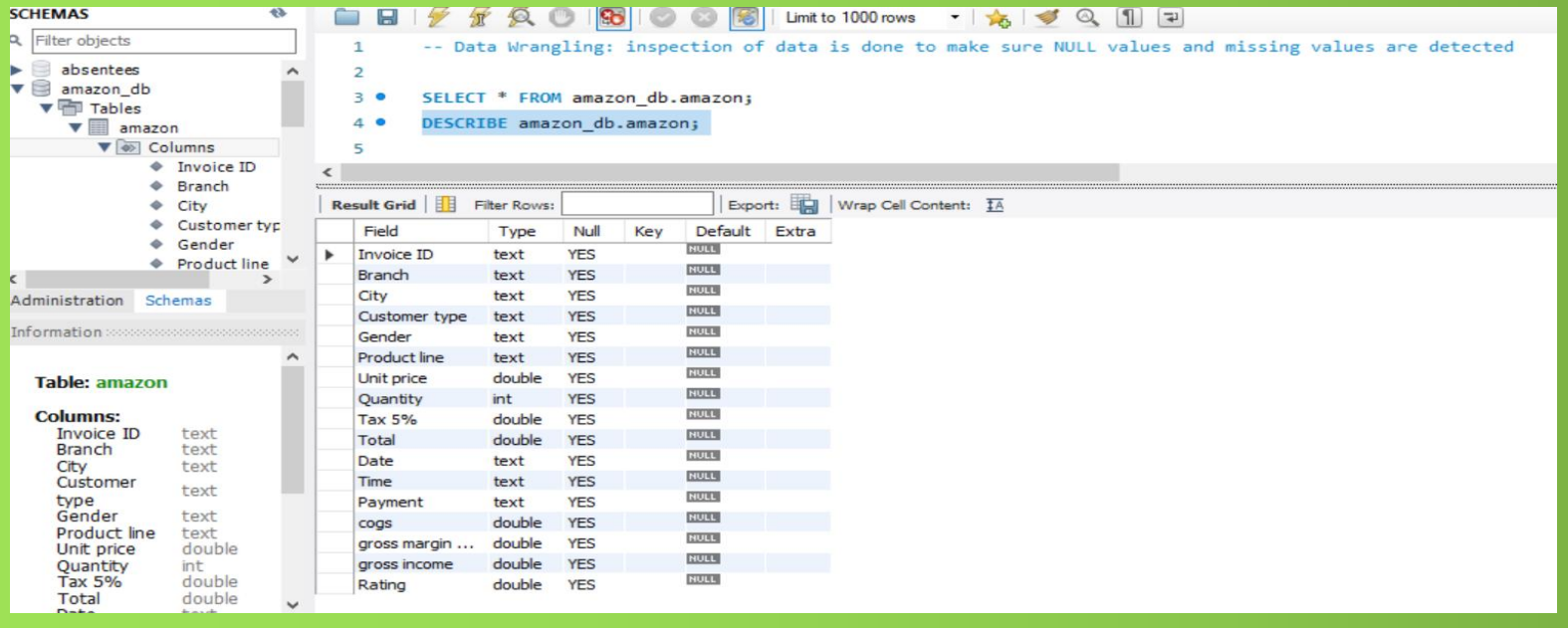
This dataset contains sales transactions from three different branches of Amazon, respectively located in Mandalay, Yangon and Naypyitaw. The data contains 17 columns and 1000 rows:

Column	Description	Data Type
invoice_id	Invoice of the sales made	VARCHAR(30)
branch	Branch at which sales were made	VARCHAR(5)
city	The location of the branch	VARCHAR(30)
customer_type	The type of the customer	VARCHAR(30)
gender	Gender of the customer making purchase	VARCHAR(10)
product_line	Product line of the product sold	VARCHAR(100)
unit_price	The price of each product	DECIMAL(10, 2)
quantity	The amount of the product sold	INT
VAT	The amount of tax on the purchase	FLOAT(6, 4)
total	The total cost of the purchase	DECIMAL(10, 2)
date	The date on which the purchase was made	DATE
time	The time at which the purchase was made	TIMESTAMP
payment_method	The total amount paid	DECIMAL(10, 2)
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	FLOAT(2, 1)

Approach Used:

Data Wrangling: **This is the first step where inspection of data is done to make sure NULL values and missing values are detected and data replacement methods are used to replace missing or NULL values.**

- **Build a database**
- **Create a table and insert the data.**
- **Select columns with null values in them. There are no null values in the database as in creating the tables, as NOT NULL is set for each field, hence null values are filtered out.**

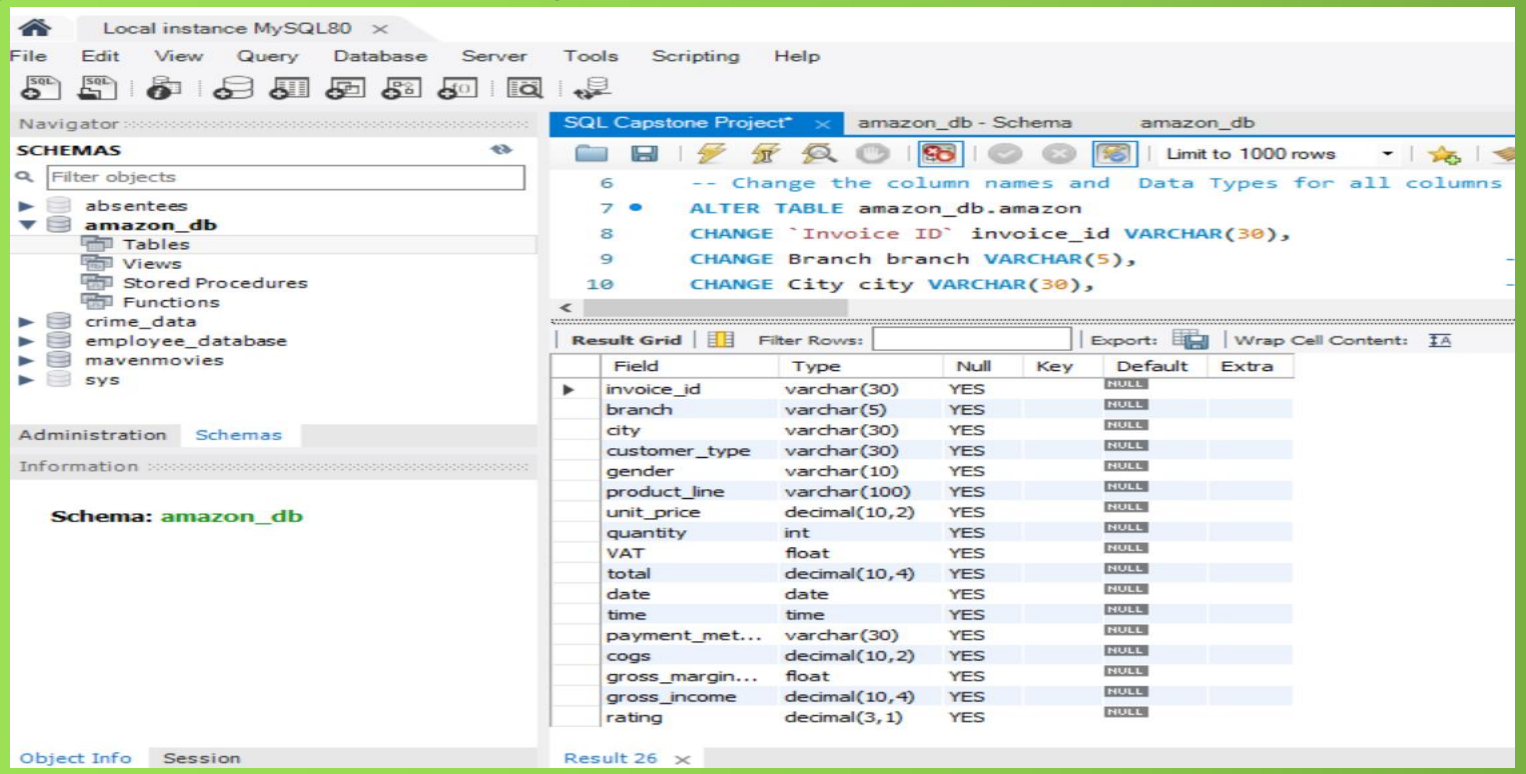


Observation: The dataset does not contain any NULL values in any of the columns. Each column has zero NULL values

Approach Used:

Data Wrangling:

- Modify the column names and Data Types for all columns



amazon

Approach Used:

Feature Engineering: This will help us generate some new columns from existing ones.

- Add a new column named **timeofday** to give insight of sales in the Morning, Afternoon and Evening. This will help answer the question on which part of the day most sales are made.
- Add a new column named **dayname** that contains the extracted days of the week on which the given transaction took place (Mon, Tue, Wed, Thur, Fri). This will help answer the question on which week of the day each branch is busiest.
- Add a new column named **monthname** that contains the extracted months of the year on which the given transaction took place (Jan, Feb, Mar). Help determine which month of the year has the most sales and profit.

VAT	total	date	time	payment_method	cogs	gross_margin_percentage	gross_income	rating	timeofday	dayname	monthname
26.1415	548.9715	2019-01-05	13:08:00	Ewallet	522.83	4.7619	26.1415	9.1	Afternoon	Saturday	January
3.82	80.2200	2019-03-08	10:29:00	Cash	76.40	4.7619	3.8200	9.6	Morning	Friday	March
16.2155	340.5255	2019-03-03	13:23:00	Credit card	324.31	4.7619	16.2155	7.4	Afternoon	Sunday	March
23.288	489.0480	2019-01-27	20:33:00	Ewallet	465.76	4.7619	23.2880	8.4	Evening	Sunday	January
30.2085	634.3785	2019-02-08	10:37:00	Ewallet	604.17	4.7619	30.2085	5.3	Morning	Friday	February
29.8865	627.6165	2019-03-25	18:30:00	Ewallet	597.73	4.7619	29.8865	4.1	Evening	Monday	March
20.652	433.6920	2019-02-25	14:36:00	Ewallet	413.04	4.7619	20.6520	5.8	Afternoon	Monday	February
36.78	772.3800	2019-02-24	11:38:00	Ewallet	735.60	4.7619	36.7800	8.0	Morning	Sunday	February
3.626	76.1460	2019-01-10	17:15:00	Credit card	72.52	4.7619	3.6260	7.2	Afternoon	Thursday	January
8.226	172.7460	2019-02-20	13:27:00	Credit card	164.52	4.7619	8.2260	5.9	Afternoon	Wednesday	February
2.896	60.8160	2019-02-06	18:07:00	Ewallet	57.92	4.7619	2.8960	4.5	Evening	Wednesday	February
5.102	107.1420	2019-03-09	17:03:00	Cash	102.04	4.7619	5.1020	6.8	Afternoon	Saturday	March
11.7375	246.4875	2019-02-12	10:25:00	Ewallet	234.75	4.7619	11.7375	7.1	Morning	Tuesday	February

Business Questions To Answer:

1. What is the count of distinct cities in the dataset?

```
61
62      -- Business Questions To Answer:
63          -- 1. What is the count of distinct cities in the dataset?
64 •            SELECT
65              COUNT(DISTINCT city) AS distinct_city_count,
66              GROUP_CONCAT(DISTINCT city ORDER BY city SEPARATOR ', ') AS city_names
67              FROM amazon_db.amazon;
68
69
70
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

distinct_city_count	city_names
3	Mandalay, Naypyitaw, Yangon

2. For each branch, what is the corresponding city?

```
68          -- 2. For each branch, what is the corresponding city?
69 •            SELECT branch, city
70              FROM amazon_db.amazon
71              GROUP BY branch, city;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

branch	city
A	Yangon
C	Naypyitaw
B	Mandalay

Business Questions To Answer:

3. What is the count of distinct product lines in the dataset?

```
73      -- 3.What is the count of distinct product lines in the dataset?
74      SELECT COUNT(DISTINCT product_line) AS distinct_product_line_count
75      FROM amazon_db.amazon;
76
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

distinct_product_line_count
6

4. Which payment method occurs most frequently?

```
77      -- 4. Which payment method occurs most frequently?
78      SELECT payment_method, COUNT(*) AS frequency
79      FROM amazon_db.amazon
80      GROUP BY payment_method
81      ORDER BY frequency DESC
82      LIMIT 1;
83
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

payment_method	frequency
Ewallet	345

Business Questions To Answer:

5. Which product line has the highest sales?

```
84        -- 5. Which product line has the highest sales?
85 •          SELECT product_line, SUM(total) AS total_sales
86            FROM amazon_db.amazon
87            GROUP BY product_line
88            ORDER BY total_sales DESC
89            ;
90
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_line	total_sales
Food and beverages	56144.8440
Sports and travel	55122.8265
Electronic accessories	54337.5315
Fashion accessories	54305.8950
Home and lifestyle	53861.9130
Health and beauty	49193.7390

6. How much revenue is generated each month?

```
91        -- 6. How much revenue is generated each month?
92 •          SELECT
93            monthname AS month,
94            SUM(total) AS total_revenue
95            FROM amazon_db.amazon
96            GROUP BY month
97            ORDER BY total_revenue desc;
98
99        -- 7.
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

month	total_revenue
January	116291.8680
March	109455.5070
February	97219.3740

Business Questions To Answer:

7. In which month did the cost of goods sold reach its peak?

```
98
99        -- 7. In which month did the cogs(cost of goods sold) reach its peak?
100       SELECT
101       monthname AS month,
102       SUM(cogs) AS total_cogs
103       FROM amazon_db.amazon
104       GROUP BY month
105       ORDER BY total_cogs DESC
106       LIMIT 1;
107
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

month	total_cogs
January	110754.16

8. Which product line generated the highest revenue?

```
108       -- 8. Which product line generated the highest revenue?
109       SELECT
110       product_line,
111       SUM(total) AS total_revenue
112       FROM amazon_db.amazon
113       GROUP BY product_line
114       ORDER BY total_revenue DESC
115       limit 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_line	total_revenue
Food and beverages	56144.8440

Business Questions To Answer:

9. In which city was the highest revenue recorded?

```
117        -- 9. In which city was the highest revenue recorded?
118        SELECT
119        city,
120        SUM(total) AS total_revenue
121        FROM amazon_db.amazon
122        GROUP BY city
123        ORDER BY total_revenue DESC
124        LIMIT 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

city	total_revenue
Naypyitaw	110568.7065

10. Which product line incurred the highest Value Added Tax?

```
126        -- 10. Which product line incurred the highest Value Added Tax?
127        SELECT
128        product_line,
129        SUM(vat) AS total_vat
130        FROM amazon_db.amazon
131        GROUP BY product_line
132        ORDER BY total_vat DESC
133        LIMIT 1;
134
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

product_line	total_vat
Food and beverages	2673.563990712166

Business Questions To Answer:

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

```sql
-- 11. For each product line, add a column indicating "Good" if its sales are above average, otherwise "Bad."
    SELECT
    product_line,
    SUM(total) AS total_sales,
    CASE
        WHEN SUM(total) > (SELECT AVG(total_sales)
                           FROM (SELECT SUM(total) AS total_sales
                                 FROM amazon_db.amazon
                                 GROUP BY product_line) AS subquery)
        THEN 'Good'
        ELSE 'Bad'
    END AS sales_status
FROM amazon_db.amazon
GROUP BY product_line;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_line	total_sales	sales_status
Health and beauty	49193.7390	Bad
Electronic accessories	54337.5315	Good
Home and lifestyle	53861.9130	Good
Sports and travel	55122.8265	Good
Food and beverages	56144.8440	Good
Fashion accessories	54305.8950	Good

12. Identify the branch that exceeded the average number of products sold.

```sql
-- 12. Identify the branch that exceeded the average number of products sold.
    WITH BranchSales AS (
    SELECT
        branch,
        SUM(quantity) AS total_products_sold
    FROM amazon_db.amazon
    GROUP BY branch
),
AverageSales AS (
    SELECT
        AVG(total_products_sold) AS avg_products_sold
    FROM BranchSales
)
SELECT
    branch,
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

branch	total_products_sold
A	1859

Business Questions To Answer:

13. Which product line is most frequently associated with each gender?

```sql
-- 13. Which product line is most frequently associated with each gender?
WITH ProductLineCounts AS (
    SELECT
        gender,
        product_line,
        COUNT(*) AS count_per_line
    FROM amazon_db.amazon
    GROUP BY gender, product_line
),
RankedLines AS (
    SELECT
        gender,
        product_line,
        count_per_line,
        RANK() OVER (PARTITION BY gender ORDER BY count_per_line DESC) AS ranks
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

gender	product_line	count_per_line
Female	Fashion accessories	96
Male	Health and beauty	88

14. Calculate the average rating for each product line.

```sql
-- 14. Calculate the average rating for each product line.
SELECT
product_line,
AVG(rating) AS average_rating
FROM amazon_db.amazon
GROUP BY product_line
ORDER BY average_rating DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_line	average_rating
Food and beverages	7.11322
Fashion accessories	7.02921
Health and beauty	7.00329
Electronic accessories	6.92471
Sports and travel	6.91627
Home and lifestyle	6.83750

Business Questions To Answer:

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

```
201     -- 15. Count the sales occurrences for each time of day on every weekday.
202 •       SELECT
203         dayname AS weekday,
204         timeofday AS time_of_day,
205         COUNT(*) AS sales_occurrences
206         FROM amazon_db.amazon
207         GROUP BY weekday, time_of_day
208         ORDER BY weekday, time_of_day;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

weekday	time_of_day	sales_occurrences
Friday	Afternoon	74
Friday	Evening	36
Friday	Morning	29
Monday	Afternoon	75
Monday	Evening	29
Monday	Morning	21
Saturday	Afternoon	81
Saturday	Evening	55
Saturday	Morning	28
Sunday	Afternoon	70
Sunday	Evening	41
Sunday	Morning	22
Thursday	Afternoon	76
Thursday	Evening	29

Result 57

16. Identify the customer type contributing the highest revenue.

```
209
210     -- 16. Identify the customer type contributing the highest revenue.
211 •       SELECT
212         customer_type,
213         SUM(total) AS total_revenue
214         FROM amazon_db.amazon
215         GROUP BY customer_type
216         ORDER BY total_revenue DESC
217         LIMIT 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

customer_type	total_revenue
Member	164223.4440

Business Questions To Answer:

17. Determine the city with the highest VAT percentage.

```sql
-- 17. Determine the city with the highest VAT percentage.
SELECT
city,
AVG((vat / total) * 100) AS avg_vat_percentage
FROM amazon_db.amazon
GROUP BY city
ORDER BY avg_vat_percentage DESC
LIMIT 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

city	avg_vat_percentage
Mandalay	4.761904769004359

18. Identify the customer type with the highest VAT payments.

```sql
-- 18. Identify the customer type with the highest VAT payments.
SELECT
customer_type,
SUM(vat) AS total_vat_payments
FROM amazon_db.amazon
GROUP BY customer_type
ORDER BY total_vat_payments DESC
LIMIT 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

customer_type	total_vat_payments
Member	7820.163996100426

Business Questions To Answer:

19. What is the count of distinct customer types in the dataset?

Member, Normal

```
236
237     -- 19. What is the count of distinct customer types in the dataset?
238 •       SELECT
239         COUNT(DISTINCT customer_type) AS distinct_customer_type_count
240         FROM amazon_db.amazon;
241
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

distinct_customer_type_count
2

20. What is the count of distinct payment methods in the dataset?

Cash, credit card, ewallet

```
241
242     -- 20. What is the count of distinct payment methods in the dataset?
243 •       SELECT
244         COUNT(DISTINCT payment_method) AS distinct_payment_method_count
245         FROM amazon_db.amazon;
246
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

distinct_payment_method_count
3

Business Questions To Answer:

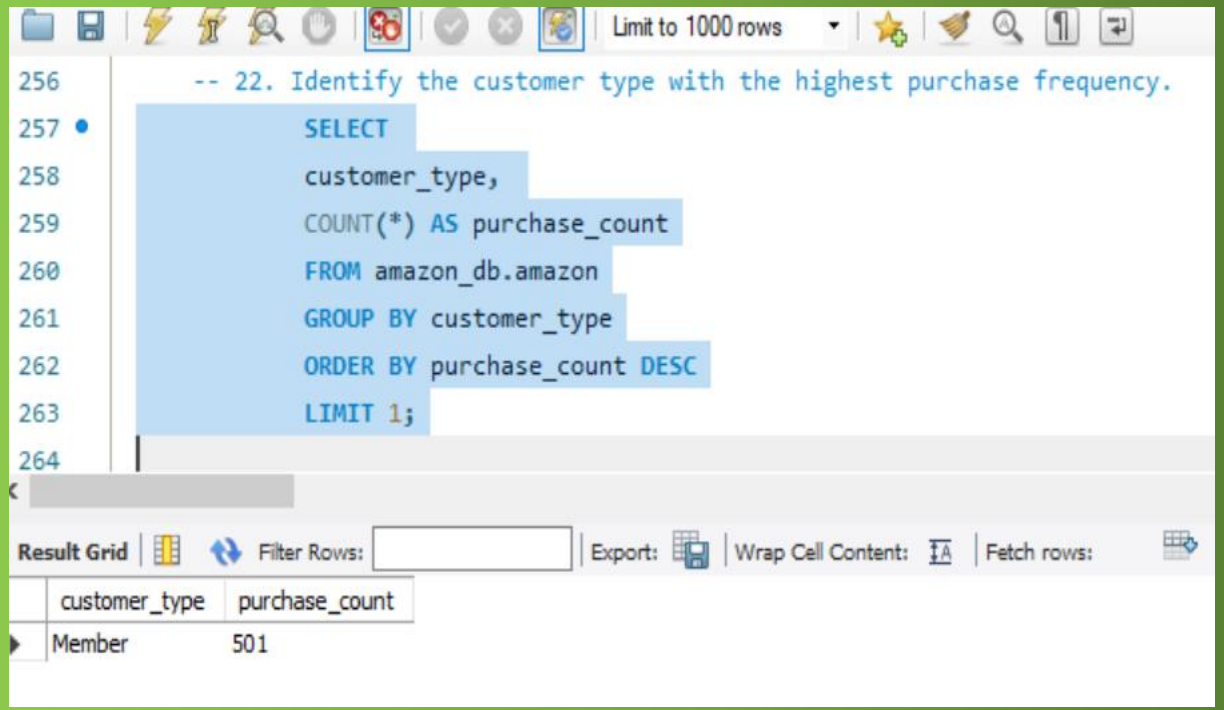
21. Which customer type occurs most frequently?

```
247     -- 21. Which customer type occurs most frequently?
248 •       SELECT
249         customer_type,
250         COUNT(*) AS occurrence_count
251         FROM amazon_db.amazon
252         GROUP BY customer_type
253         ORDER BY occurrence_count DESC
254         LIMIT 1;
255
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

customer_type	occurrence_count
Member	501

22. Identify the customer type with the highest purchase frequency.



Business Questions To Answer:

23. Determine the predominant gender among customers.

```
265    -- 23. Determine the predominant gender among customers.
266    SELECT
267    gender,
268    COUNT(*) AS gender_count
269    FROM amazon_db.amazon
270    GROUP BY gender
271    ORDER BY gender_count DESC
272    LIMIT 1;
273
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

gender	gender_count
Female	501

24. Examine the distribution of genders within each branch.

```
274    -- 24. Examine the distribution of genders within each branch.
275    SELECT
276    branch, gender,
277    COUNT(*) AS gender_count
278    FROM amazon_db.amazon
279    GROUP BY branch, gender
280    ORDER BY branch, gender;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

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

Business Questions To Answer:

25. Identify the time of day when customers provide the most ratings.

```sql
-- 25. Identify the time of day when customers provide the most ratings.

SELECT
timeofday AS hour_of_day,
COUNT(rating) AS rating_count
FROM amazon_db.amazon
GROUP BY hour_of_day
ORDER BY rating_count DESC
LIMIT 1;
```

hour_of_day	rating_count
Afternoon	528

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

```sql
-- 26. Determine the time of day with the highest customer ratings for each branch.
WITH RatingsByHour AS (
SELECT
    branch,
    timeofday AS hour_of_day,
    COUNT(rating) AS rating_count
FROM amazon_db.amazon
GROUP BY branch, hour_of_day
)
SELECT
    branch,
    hour_of_day,
    rating_count
FROM RatingsByHour
WHERE (branch, rating_count) IN (
```

branch	hour_of_day	rating_count
A	Afternoon	185
B	Afternoon	162
C	Afternoon	181

Business Questions To Answer:

27. Identify the day of the week with the highest average ratings.

```
313    -- 27. Identify the day of the week with the highest average ratings.
314    SELECT
315    dayname AS day_of_week,
316    AVG(rating) AS average_rating
317    FROM amazon_db.amazon
318    GROUP BY day_of_week
319    ORDER BY average_rating DESC
320    LIMIT 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

day_of_week	average_rating
Monday	7.15360

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

```
322    -- 28. Determine the day of the week with the highest average ratings for each branch.
323
324    WITH AverageRatings AS (
325    SELECT
326        branch,
327        dayname AS day_of_week,
328        AVG(rating) AS average_rating
329    FROM amazon_db.amazon
330    GROUP BY branch, day_of_week
331    ),
332    RankedRatings AS (
333    SELECT
334        branch,
335        day_of_week,
336        average_rating,
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

branch	day_of_week	average_rating
A	Friday	7.31200
B	Monday	7.33590
C	Friday	7.27895

Business Questions To Answer:

1. What is the count of distinct cities in the dataset?

```
61
62      -- Business Questions To Answer:
63          -- 1. What is the count of distinct cities in the dataset?
64 •            SELECT
65              COUNT(DISTINCT city) AS distinct_city_count,
66              GROUP_CONCAT(DISTINCT city ORDER BY city SEPARATOR ', ') AS city_names
67              FROM amazon_db.amazon;
68
69
70
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

distinct_city_count	city_names
3	Mandalay, Naypyitaw, Yangon

2. For each branch, what is the corresponding city?

```
68          -- 2. For each branch, what is the corresponding city?
69 •            SELECT branch, city
70              FROM amazon_db.amazon
71              GROUP BY branch, city;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

branch	city
A	Yangon
C	Naypyitaw
B	Mandalay

Analysis List:

Product Analysis

Conduct analysis on the data to understand the different product lines, the products lines performing best and the product lines that need to be improved.

Sales Analysis

This analysis aims to answer the question of the sales trends of product. The result of this can help us measure the effectiveness of each sales strategy the business applies and what modifications are needed to gain more sales.

Customer Analysis

This analysis aims to uncover the different customer segments, purchase trends and the profitability of each customer segment.

amazon

Product Analysis :

Total Sales: Food and Beverages is the top performing product line. Health and Beauty product line needs improovment being the lowest performing product line.

- product line with highest sales-['Food and beverages', '56144.8440']/ product line with lowest sales- ['Health and beauty', '49193.7390']
- product line that generated the highest revenue- ['Food and beverages', '56144.8440']
- product line incurred the highest Value Added Tax-Food and Beverages. city-['Mandalay', '4.761904769004359']
- product line is most frequently associated with each gender-
  'Female', 'Fashion accessories', '96'
  'Male', 'Health and beauty', '88'

Customer Analysis :

Customer Segmentation- Member/Non-member

- *Customer type*- Members generate the most revenue['Member', '164223.4440'] and shop more frequently than the non- members[158743.9].
- Customer type with the highest purchase frequency- Member[501]

Customer Gender Distribution:

- Predominant Gender - Females are predominant [501] than males
- Branch-specific gender preference- Males dominating in branch A and B , whereas females in branch C
- Female customers porefers Fashion Accessories and Males prefer health and beauty products.

Customer Feedback:

- Time of day for Rating: Most customers provided their ratings in Afternoon[528]
- Highest average rating- Monday[7.33590] has highest average rating . Branch- B, Month-['January', '116291.8680']

Customer Payment Mode:

- Popular payment mode is- ['Ewallet', '345']

Branch that exceeded the average number of products sold- ['A', '1859']

amazon

Sales Analysis :

Highest Monthly Revenue generated- Month-['January', '116291.8680'],
Lowest Monthly Revenue generated- ['February', '97219.3740']

Highest Day Revenue: ['Saturday', 'Morning', '28', '164']
Lowest Day Revenue: ['Monday', '125']

Highest selling time of the day - Afternoon

city was the highest revenue recorded-'Naypyitaw', '110568.7065'

Sales performance (Above/Below average): 'Health and beauty', '49193.7390', 'Bad' categoty with below average performance.Other product lines are above average sale(Good)

Key Insights and Strategies:

- ‘Food and Beverages’ is the top performng product line , leading highest Sales, Revenue and VAT
- 'Health and beauty’ being the lowest performing product line ,needs- Reevaluation, introduce new products, Price adjustments, targeted promotions to improve sales
- Promotion of highly rated products like -’Food and Beverages’
- Enhancing loyality programs and personalised offers to Members Customers to retain them which generates more revenue .
- Gender specific Marketing in the branches having maximum males/females customers
- Promotion on Peak Sell periods like - afternoon(peak time of sell) , weekends(Specially SATURDAY which has highest daily sell) and festive months to maximize revenue
- Discounts on slower periods like February month
- Replicate the strategies of successful branches like branch A to underperforming branches by analyzing their performance metrics, identify key success factors, and implement those in other branches.