SQL Project Report – Amazon Analysis

Author: *SAHIL SHAIKH*
Date: *AUGUST 2025*

---

1. Project Overview

The objective of this project is to analyze Amazon product data to identify top-performing brands, their ratings, and sales trends across different categories.
This analysis also focuses on determining the highest and lowest priced products in the dataset and studying patterns in discounts, ratings, and product popularity.
The project demonstrates data cleaning, transformation, and advanced SQL queries to extract meaningful insights that can help businesses in pricing strategies, category performance evaluation, and customer preference analysis.Dataset Details:

- **Source:** https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset
- **Rows:** *339*
- **Columns:** *19*
- **Key Fields:** brand, product_name, rating, rating_count, actual_price_num, discounted_price_num

---

2. Tools & Technologies Used

- Oracle SQL Developer
- SQL (SELECT, GROUP BY, ORDER BY, UNION, Aggregate Functions)
- CSV dataset

---

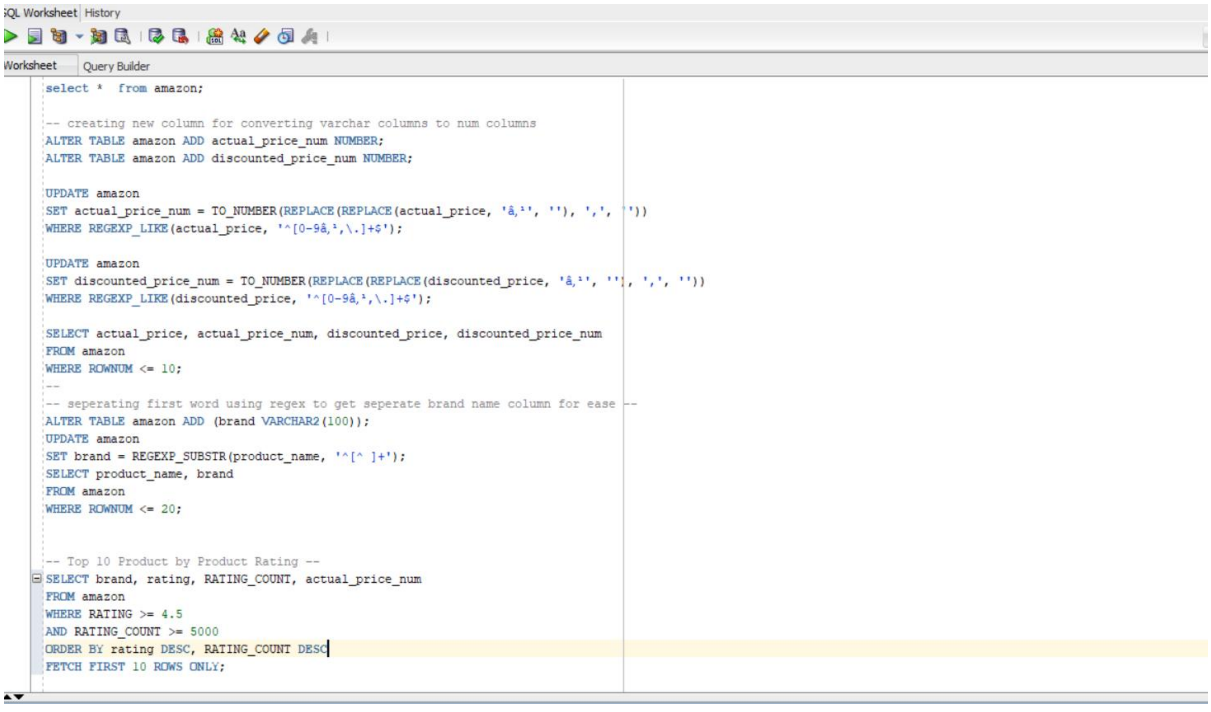
3. Queries & Insights

Converting actual price column from varchar to number and create new column brand

```
-- creating new column for converting varchar columns to num columns

ALTER TABLE amazon ADD actual_price_num NUMBER;

ALTER TABLE amazon ADD discounted_price_num NUMBER;


UPDATE amazon

SET actual_price_num = TO_NUMBER(REPLACE(REPLACE(actual_price, 'â,¹', ''), ',', ''))

WHERE REGEXP_LIKE(actual_price, '^[0-9â,¹,\.]+$');


UPDATE amazon
```

```
SET discounted_price_num = TO_NUMBER(REPLACE(REPLACE(discounted_price, 'â,¹', ''), ',', ''))
WHERE REGEXP_LIKE(discounted_price, '^[0-9â,¹,\.]+$');

SELECT actual_price, actual_price_num, discounted_price, discounted_price_num
FROM amazon
WHERE ROWNUM <= 10;
--
-- seperating first word using regex to get seperate brand name column for ease --
ALTER TABLE amazon ADD (brand VARCHAR2(100));
UPDATE amazon
SET brand = REGEXP_SUBSTR(product_name, '^[^ ]+');
SELECT product_name, brand
FROM amazon
WHERE ROWNUM <= 20;
```

--Result Screenshot:



Insight:

This is done to ensure the arithmetic operation such as discount should be calculated and creating a new column brand to get top brand

---

Query 1 – Top 10 Brand by Product Rating

SELECT brand, rating, RATING_COUNT, actual_price_num

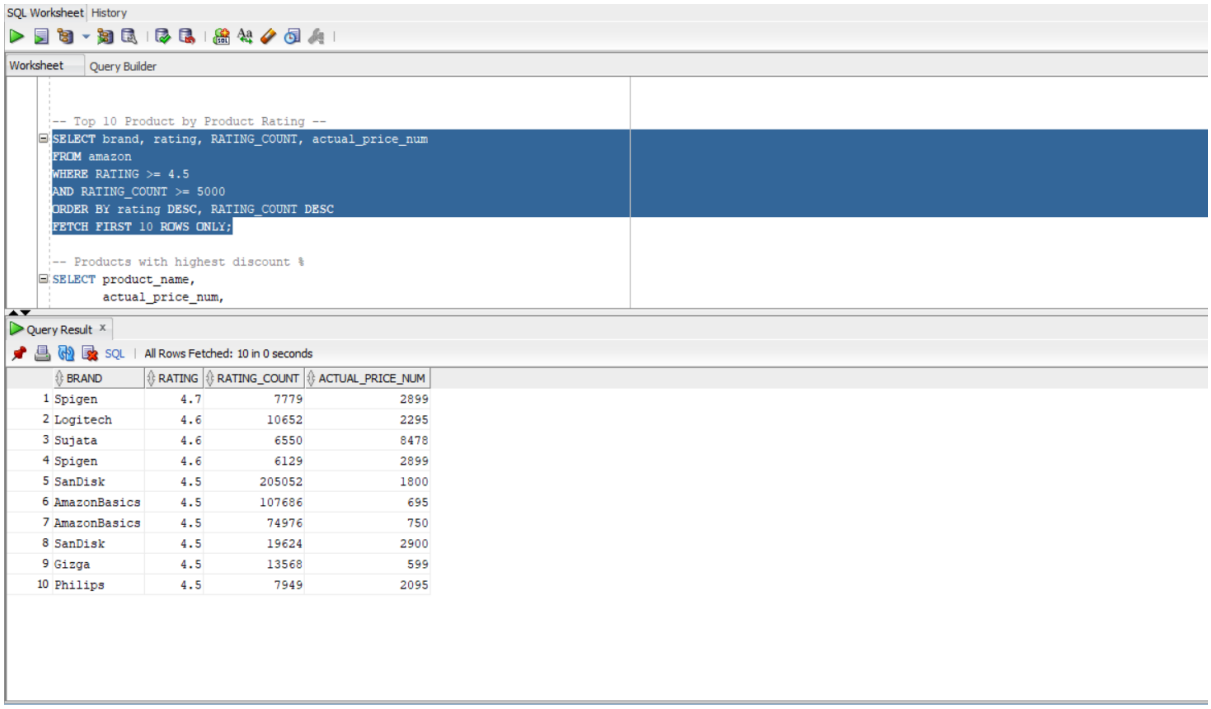
FROM amazon

WHERE RATING >= 4.5

AND RATING_COUNT >= 5000

ORDER BY rating DESC, RATING_COUNT DESC

FETCH FIRST 10 ROWS ONLY;**Result Screenshot:**



Insight:
Top 10 Brand with product rating, their price and rating count

---

Query 2 – Products with highest discount %

SELECT PRODUCT_NAME,

actual_price_num,

discounted_price_num,

ROUND(((actual_price_num - discounted_price_num) * 100) / actual_price_num, 2) AS discount_pct

FROM amazon

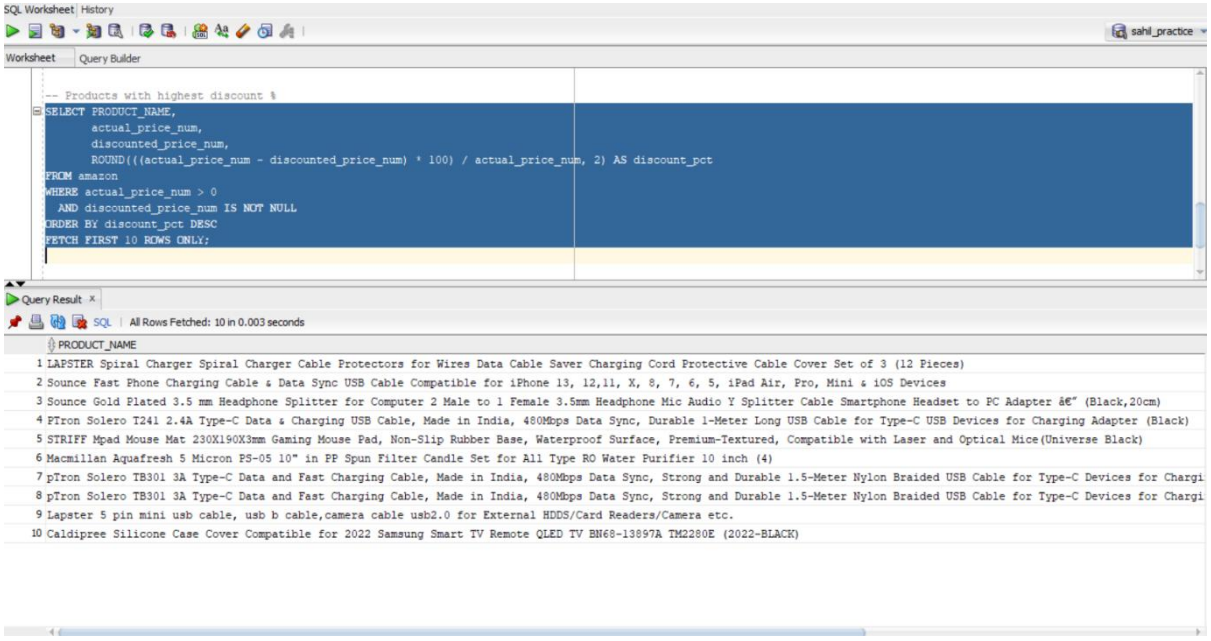
WHERE actual_price_num > 0

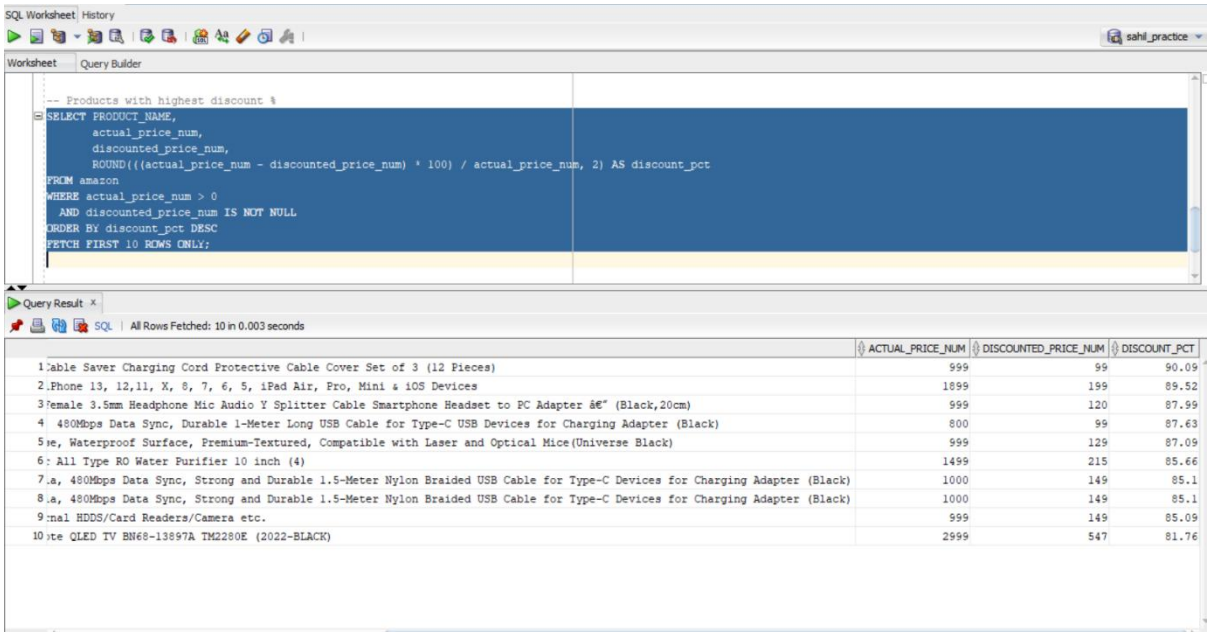
AND discounted_price_num IS NOT NULL

ORDER BY discount_pct DESC

FETCH FIRST 10 ROWS ONLY;ORDER BY distance DESC;

Result Screenshot:





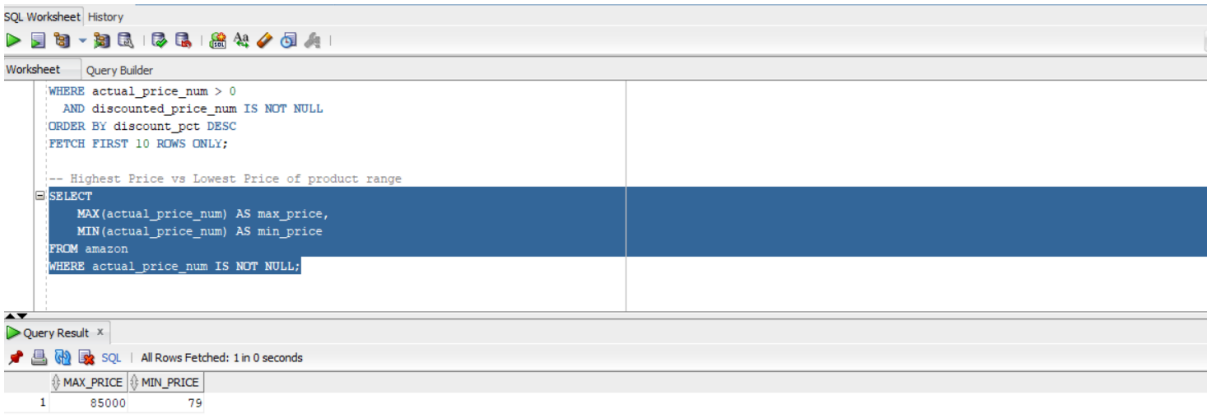
Insight:
query is used to get the most discounted products on amazon with actual price and discounted price

---

Query 3 – Highest Price vs Lowest Price of product range

```
SELECT
    MAX(actual_price_num) AS max_price,
    MIN(actual_price_num) AS min_price
FROM amazon
WHERE actual_price_num IS NOT NULL;
```

Result Screenshot:



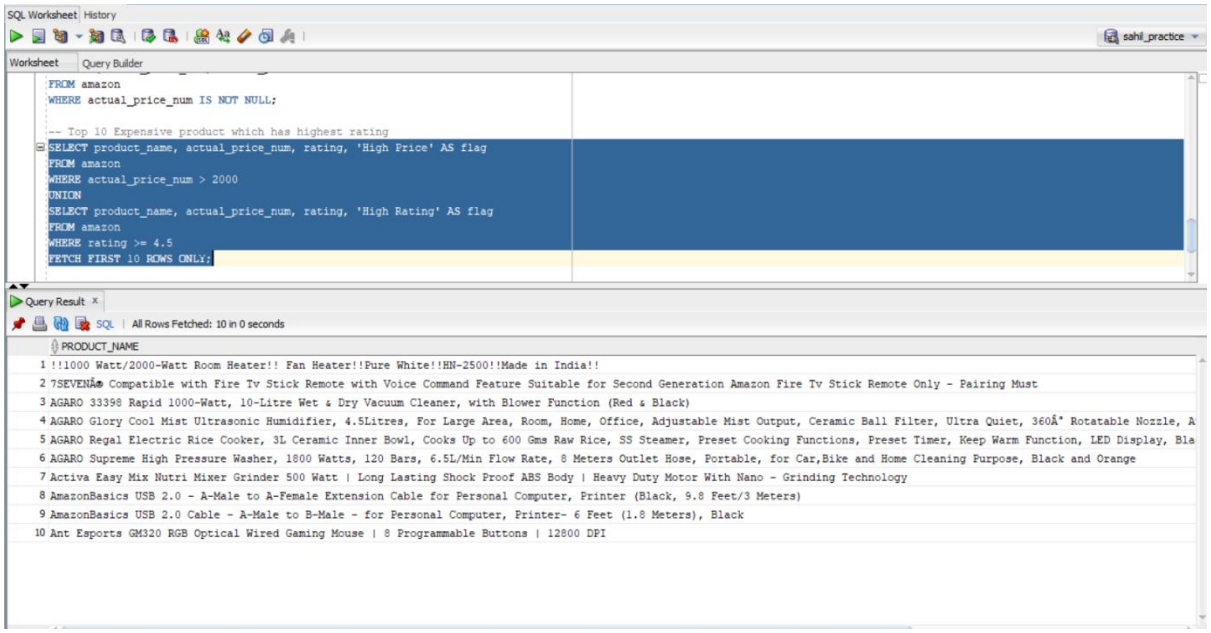
Insight:
This Query is used to get highest price and lowest price range in amazon

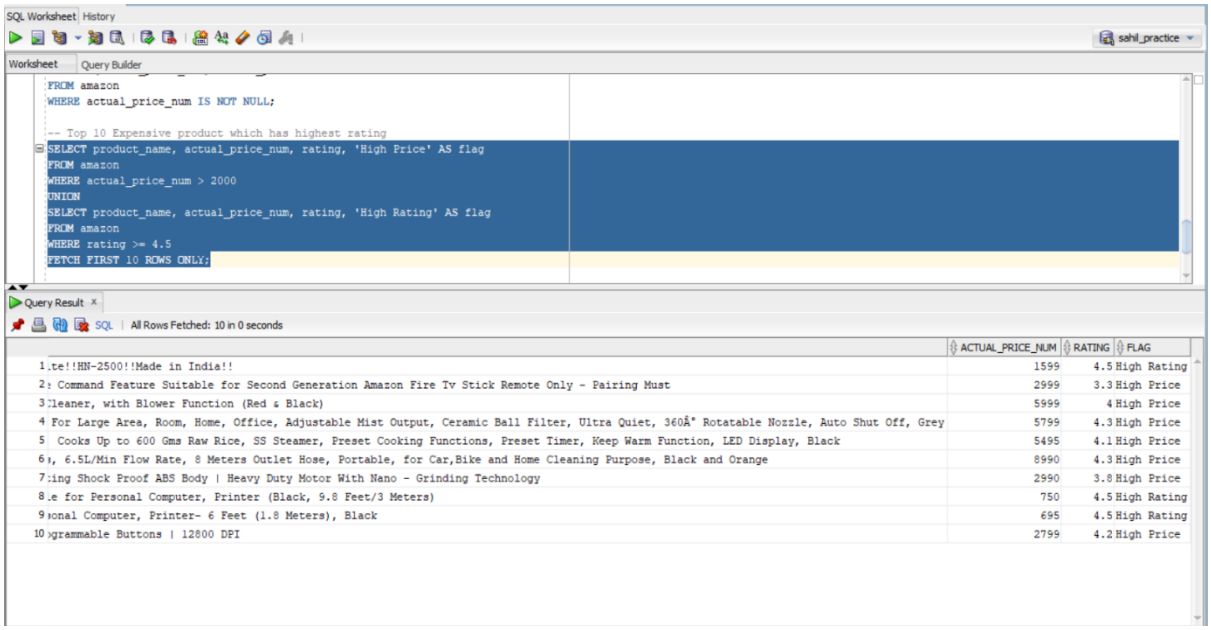
---

Query 4 – Top 10 Expensive product which has highest rating

```
SELECT product_name, actual_price_num, rating, 'High Price' AS flag
FROM amazon
WHERE actual_price_num > 2000
UNION
SELECT product_name, actual_price_num, rating, 'High Rating' AS flag
FROM amazon
WHERE rating >= 4.5
FETCH FIRST 10 ROWS ONLY;
```

Result Screenshot:





Insight:

Checking the most expensive products and their price and rating to see whether they are worth it or not.

Query 5 – Brand with highest rating and rating count on their products

```
SELECT brand,
    SUM(rating_count) AS total_ratings,
    ROUND(AVG(rating),2) AS avg_rating
FROM amazon
```

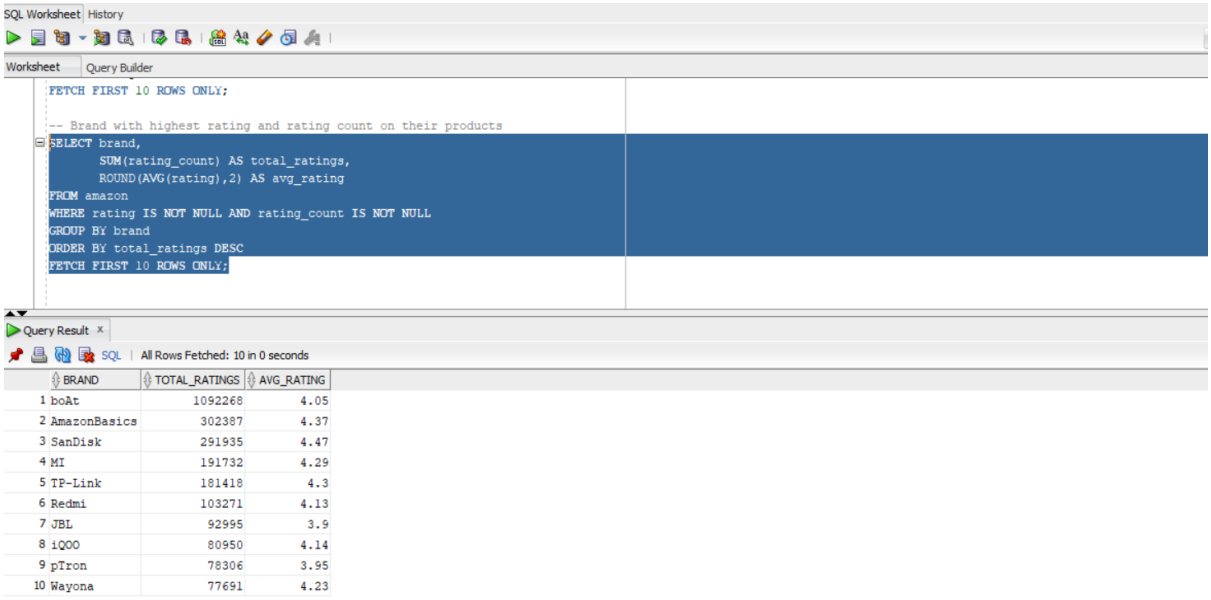
WHERE rating IS NOT NULL AND rating_count IS NOT NULL

GROUP BY brand

ORDER BY total_ratings DESC

FETCH FIRST 10 ROWS ONLY;

Result Screenshots:-



	BRAND	TOTAL_RATINGS	AVG_RATING
1	boAt	1092268	4.05
2	AmazonBasics	302387	4.37
3	SanDisk	291935	4.47
4	MI	191732	4.29
5	TP-Link	181418	4.3
6	Redmi	103271	4.13
7	JBL	92995	3.9
8	iQOO	80950	4.14
9	pTron	78306	3.95
10	Wayona	77691	4.23

Insight:
Top 10 Brands with highest number of Rating counts and avg number of rating on their every products

---

Query 6 – Category with most rating count

SELECT CATEGORY,

SUM(rating_count) AS total_ratings,

ROUND(AVG(rating),2) AS avg_rating

FROM amazon

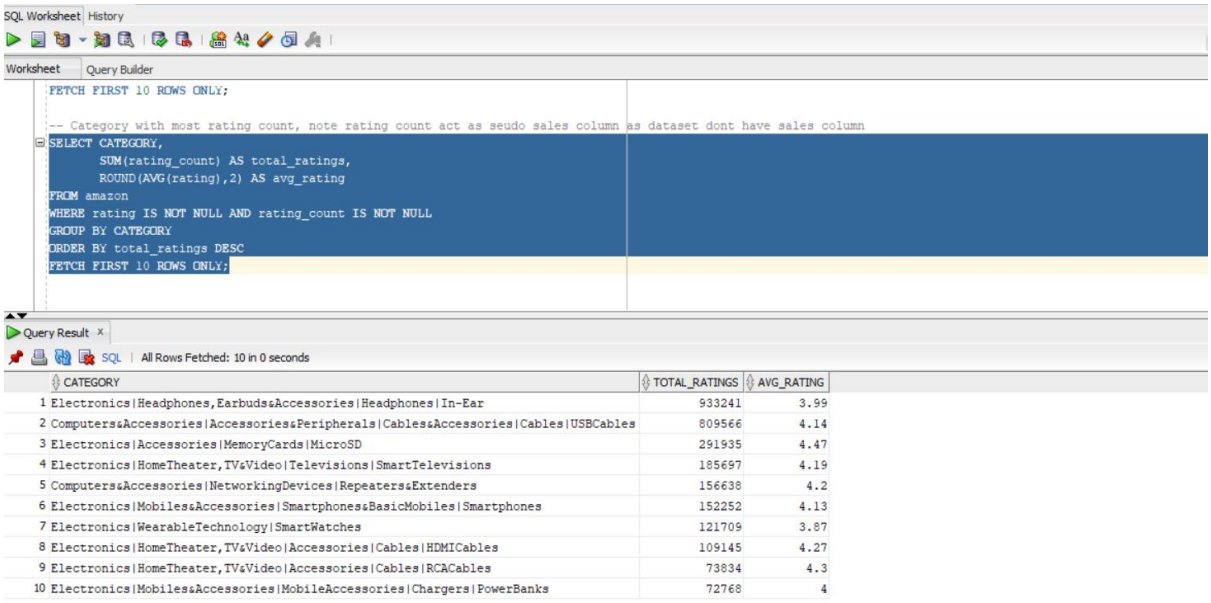
WHERE rating IS NOT NULL AND rating_count IS NOT NULL

GROUP BY CATEGORY

ORDER BY total_ratings DESC

FETCH FIRST 10 ROWS ONLY;

Result Screenshots:-



```
FETCH FIRST 10 ROWS ONLY;

-- Category with most rating count, note rating count act as seudo sales column as dataset dont have sales column
SELECT CATEGORY,
       SUM(rating_count) AS total_ratings,
       ROUND(AVG(rating),2) AS avg_rating
FROM amazon
WHERE rating IS NOT NULL AND rating_count IS NOT NULL
GROUP BY CATEGORY
ORDER BY total_ratings DESC
FETCH FIRST 10 ROWS ONLY;
```

Query Result — All Rows Fetched: 10 in 0 seconds

	CATEGORY	TOTAL_RATINGS	AVG_RATING
1	Electronics|Headphones,Earbuds&Accessories|Headphones|In-Ear	933241	3.99
2	Computers&Accessories|Accessories&Peripherals|Cables&Accessories|Cables|USBCables	809566	4.14
3	Electronics|Accessories|MemoryCards|MicroSD	291935	4.47
4	Electronics|HomeTheater,TV&Video|Televisions|SmartTelevisions	185697	4.19
5	Computers&Accessories|NetworkingDevices|Repeaters&Extenders	156638	4.2
6	Electronics|Mobiles&Accessories|Smartphones&BasicMobiles|Smartphones	152252	4.13
7	Electronics|WearableTechnology|SmartWatches	121709	3.87
8	Electronics|HomeTheater,TV&Video|Accessories|Cables|HDMICables	109145	4.27
9	Electronics|HomeTheater,TV&Video|Accessories|Cables|RCACables	73834	4.3
10	Electronics|Mobiles&Accessories|MobileAccessories|Chargers|PowerBanks	72768	4

Insight:
Analysing which category type of products get the most rating count and average rating, note rating count act as seudo sales column as dataset dont have sales column

---

Query 7 – Most expensive vs cheapest products in Electronics,Headphones Category

```
SELECT
   MAX(actual_price_num) AS max_price,
   MIN(actual_price_num) AS min_price
FROM amazon
WHERE actual_price_num IS NOT NULL
 AND CATEGORY LIKE '%Electronics|Headphones%';
```

Result Screenshots:-

Insight:
after analysing top category we analyze its most expensive to least expensive price range

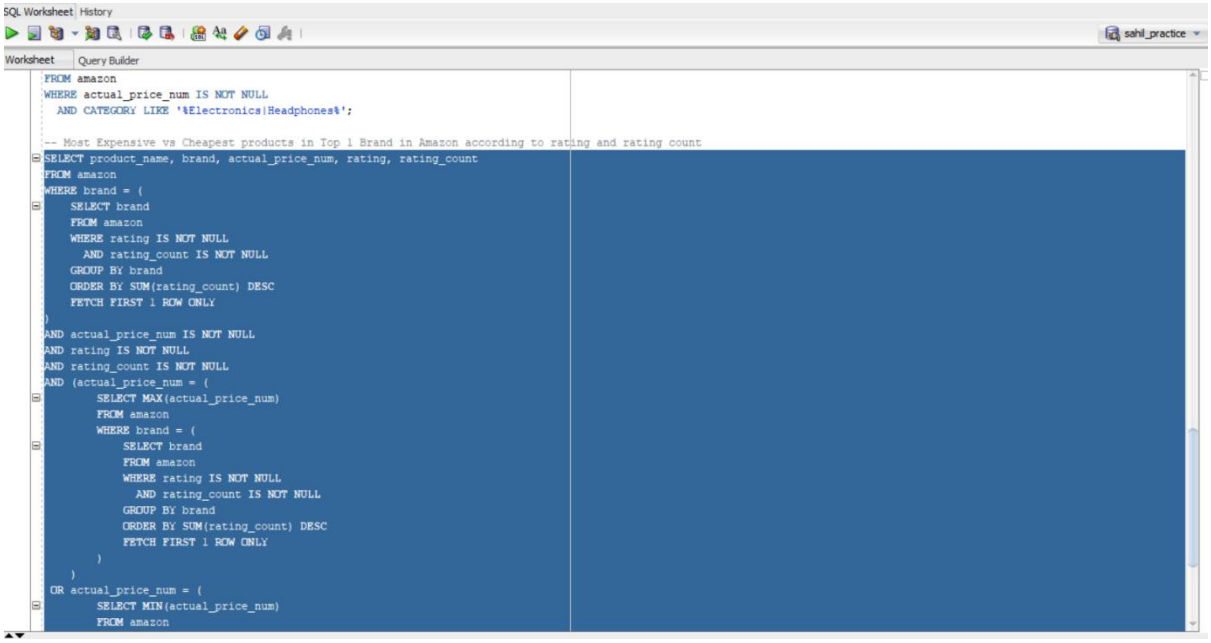
---

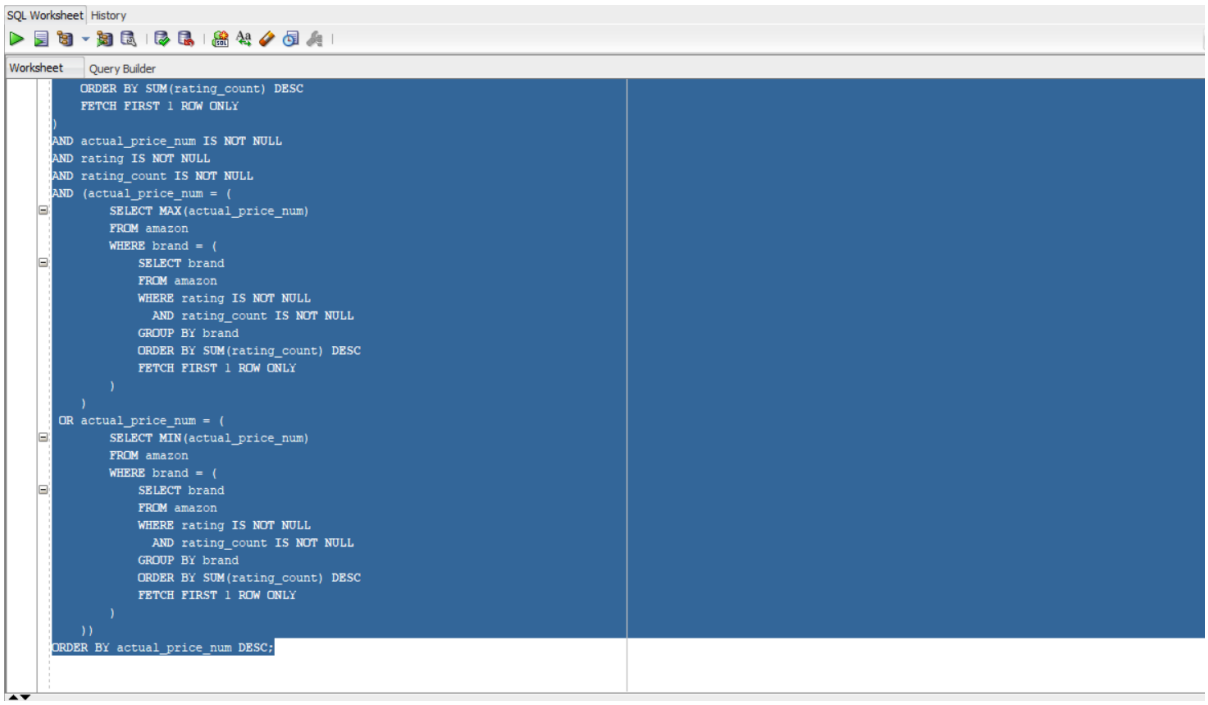
Query 7 – Most Expensive vs Cheapest products in Top 1 Brand in Amazon according

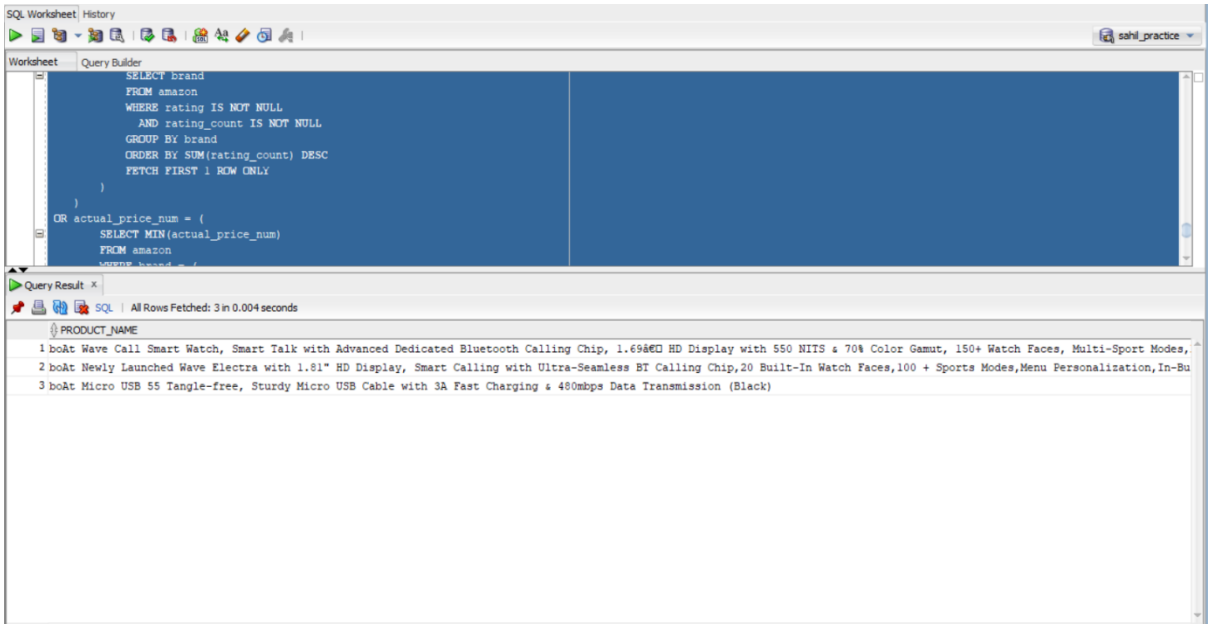
```
SELECT product_name, brand, actual_price_num, rating, rating_count
FROM amazon
WHERE brand = (
    SELECT brand
    FROM amazon
    WHERE rating IS NOT NULL
      AND rating_count IS NOT NULL
    GROUP BY brand
    ORDER BY SUM(rating_count) DESC
    FETCH FIRST 1 ROW ONLY
)
AND actual_price_num IS NOT NULL
AND rating IS NOT NULL
```

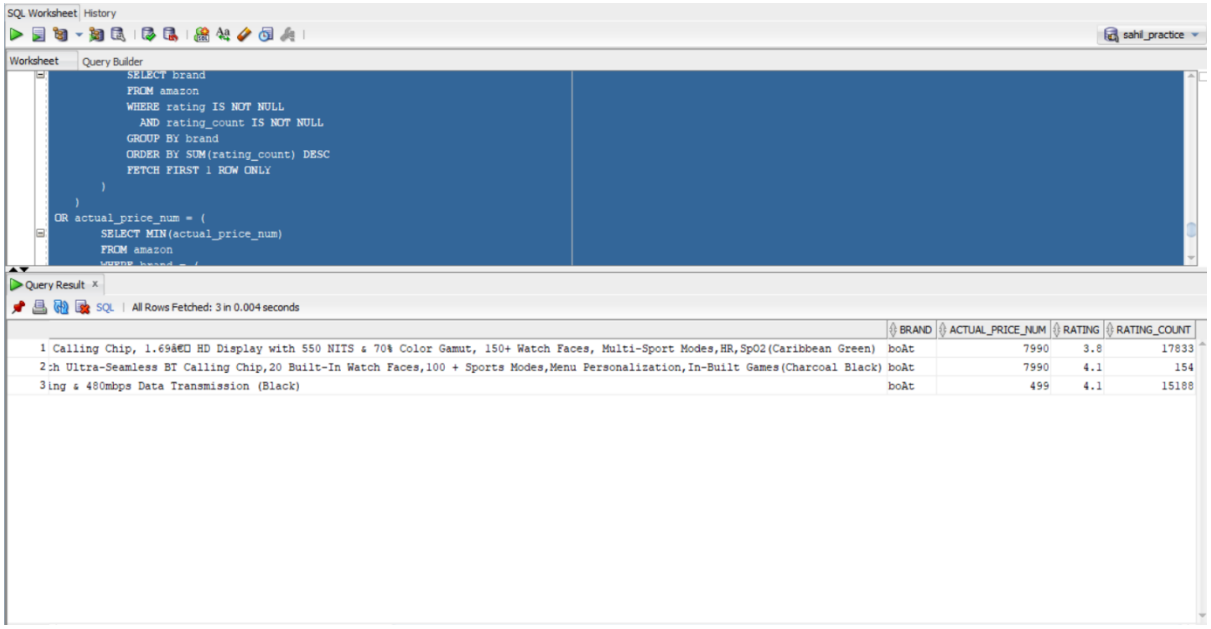
```
AND rating_count IS NOT NULL
AND (actual_price_num = (
      SELECT MAX(actual_price_num)
      FROM amazon
      WHERE brand = (
         SELECT brand
         FROM amazon
         WHERE rating IS NOT NULL
          AND rating_count IS NOT NULL
         GROUP BY brand
         ORDER BY SUM(rating_count) DESC
         FETCH FIRST 1 ROW ONLY
      )
   )
 OR actual_price_num = (
      SELECT MIN(actual_price_num)
      FROM amazon
      WHERE brand = (
         SELECT brand
         FROM amazon
         WHERE rating IS NOT NULL
          AND rating_count IS NOT NULL
         GROUP BY brand
         ORDER BY SUM(rating_count) DESC
         FETCH FIRST 1 ROW ONLY
      )
   ))
ORDER BY actual_price_num DESC;
```

Result Screenshots:-









Insight:
Analysing top 1 brand i.e Boat with its most expensive product to least expensive product according to actual price and rating_count

---

4. Key Findings

Data Cleaning and Transformation

- Converted price columns from VARCHAR2 to numeric (NUMBER) for analysis.
- Extracted **brand names** from product titles using regex for grouping insights.

Top Performing Products

- Identified products with **ratings ≥ 4.5** and **5,000+ reviews**, highlighting the most trusted items.

Highest Discounts

- Found products offering **>50% discount**, useful for price strategy evaluations.

Price Distribution

- Calculated **minimum and maximum product prices** to understand the marketplace spread.
- Highlighted **high-priced products with high ratings** to show premium items well received by customers.

Brand Insights

- Determined **top brands by total rating count and average rating**, identifying leaders in customer engagement.

Category Insights

- Used **rating count as a proxy for sales** to determine **categories with the highest customer activity**.

Special Focus Queries

- Analyzed **price range within the Electronics & Headphones category**.
- Found **most expensive vs cheapest products** for **top brand by rating count** (executed as an advanced nested query).

---

5. Conclusion

This project demonstrates **end-to-end SQL skills**:

- **Schema transformation** (altering tables, adding numeric columns).
- **Data parsing using regex** to derive brand information.
- **Analytical SQL queries** with filtering, grouping, and subqueries.

The insights are directly **actionable for e-commerce companies**:

- Identify **best-performing brands** and **products worth promoting**.
- Understand **customer behavior via rating count and discount impact**.
- Spot **premium vs budget product positioning** to optimize sales strategies.

This project also shows you can handle **real-world messy datasets** and convert them into **clean, structured insights** for decision-making.