2025

SQL-Based Exploration and Analysis of Retail Sales Data

Ayomide Makanjuola

OUTLINE

1. Introduction

The ability to efficiently manage, query, and analyze structured datasets is a fundamental requirement in modern data analytics. Structured Query Language (SQL) remains the industry standard for interacting with relational databases, enabling data analysts to retrieve, clean, and manipulate data in a systematic and reproducible manner.

This project applies SQL techniques to a retail sales dataset to simulate a real-world analytics scenario. The work involves importing raw transactional data into a relational database, performing essential data cleaning, conducting exploratory data analysis (EDA), and answering targeted business questions through SQL queries. The project integrates fundamental SQL concepts, including data definition, data manipulation, aggregation, filtering, and grouping, to generate actionable insights that can support decision-making in a retail business environment.

1.2 Purpose of the Project

i. The primary objective of this project is to demonstrate proficiency in SQL as a tool for business data analysis. Specifically, the project aims to:

   Database Creation and Population: Import the provided retail sales dataset into an SQL-based relational database management system (RDBMS).

ii. Data Cleaning: Identify and remove records containing missing or null values to enhance data quality and reliability.

iii. Exploratory Data Analysis (EDA): Investigate patterns, distributions, and anomalies within the dataset to understand customer demographics, sales trends, and product performance.

iv. Business Analysis and Insight Generation: Answer specific business-related queries through SQL to uncover trends such as top-performing product categories, customer purchasing behavior, and sales seasonality.

Through these objectives, the project seeks to replicate the typical workflow of a data analyst in the retail sector, from raw data ingestion to business insight reporting.

1.3 Dataset Description

The dataset comprises 2,000 individual retail sales transactions, each representing a purchase made by a customer across different product categories. It includes transactional details, customer demographics, and sales-related financial metrics. The attributes are described as follows:

Column Name	**Description**
transactions_id	A unique identifier for each transaction.
sale_date	The date of the transaction in YYYY-MM-DD format.
sale_time	The exact time the transaction occurred in HH:MM:SS format.
customer_id	A unique identifier for each customer.
gender	The gender of the customer (Male or Female).
age	The age of the customer; contains some missing values.
category	The category of the purchased product (e.g., Clothing, Beauty).
quantiy	The quantity of units purchased; contains some missing values.
price_per_unit	The selling price for one unit of the product; contains some missing values.
cogs	The cost of goods sold for the transaction; contains some missing values.
total_sale	The total revenue from the transaction, calculated as quantity × price_per_unit; contains some missing values.

Preliminary inspection reveals that the dataset is generally well-structured but contains null values in key numerical fields, including age, quantiy, price_per_unit, cogs, and total_sale. These missing values necessitate a preliminary data cleaning phase before conducting further analysis.

2.0 Objectives and Implementation Steps:

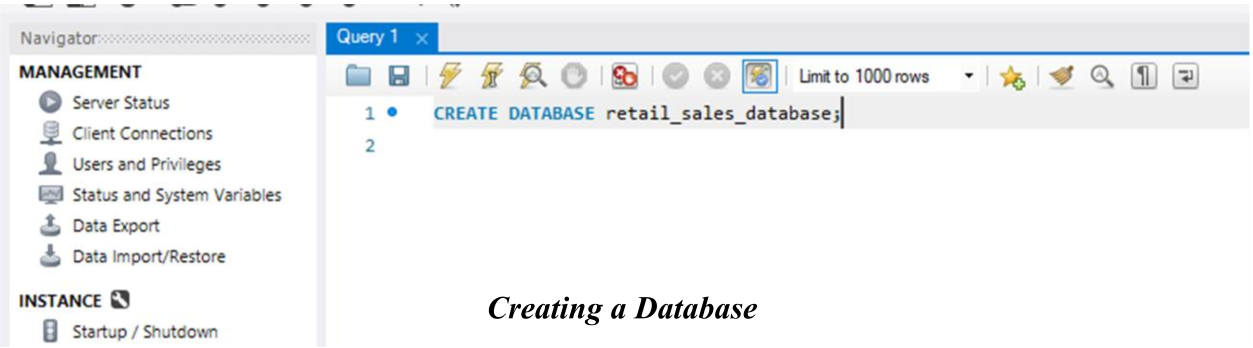
2.1 Setting Up the Retail Sales Database, alongside Identifying and removing any records with missing or null values.

To begin the project, a new database named “retail_sales_db” was created in MySQL to store the provided retail sales dataset. Within this database, a table named “retail_sales” was defined with the appropriate data types for each column to ensure accurate storage of transactional, customer, and sales information.

After defining the table structure, the CSV file containing 2,000 retail sales transactions was imported into the database using the **Table Data Import Wizard**. This method allows for efficient bulk insertion of records directly from a CSV file into MySQL, ensuring the dataset is readily available for further cleaning, exploration, and analysis.

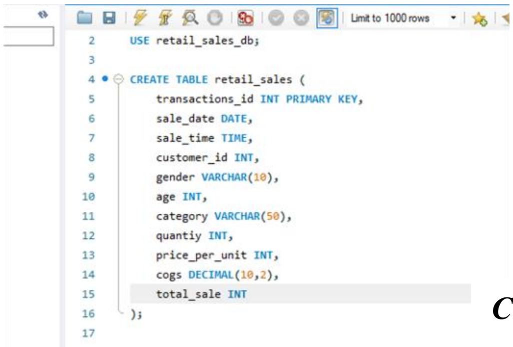
This step successfully set up the foundational database environment required for subsequent data cleaning and analysis tasks.

Using The Table Data Import Wizard also helps one of my major objectives with the data: **“Data Cleaning: Identify and remove any records with missing or null values”,** as using this technique helped to remove all the rows with null or missing values, 1987, files were then imported, meaning that 13 rows had null or missing values and were screened off, making the



```
CREATE DATABASE retail_sales_database;
```

Creating a Database



```
USE retail_sales_db;

CREATE TABLE retail_sales (
    transactions_id INT PRIMARY KEY,
    sale_date DATE,
    sale_time TIME,
    customer_id INT,
    gender VARCHAR(10),
    age INT,
    category VARCHAR(50),
    quantiy INT,
    price_per_unit INT,
    cogs DECIMAL(10,2),
    total_sale INT
);
```

Creating a Table to import the dataset

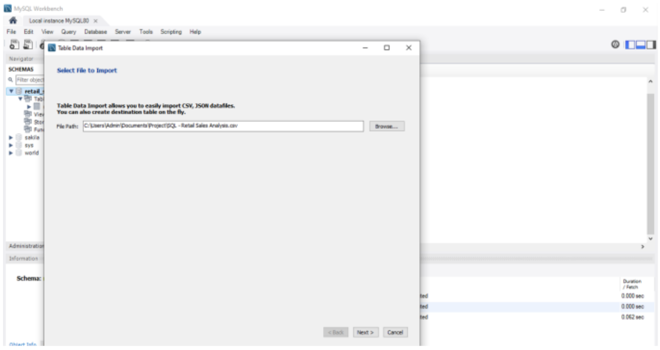


Table Data Import

Select File to Import

Table Data Import allows you to easily import CSV, JSON datafiles.
You can also create destination table on the fly.

File Path: C:\Users\Admin\Documents\Project\SQL - Retail Sales Analysis.csv Browse...

Importing the Table With the Table Data Import Wizard

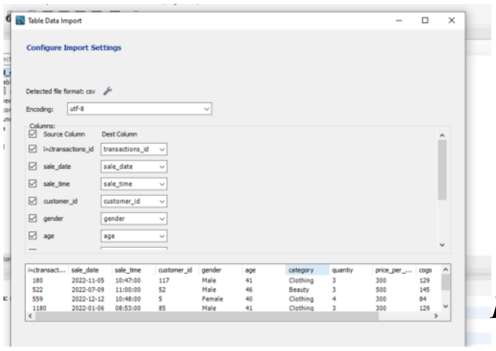


Table Data Import

Configure Import Settings

Detected file format: csv

Encoding: utf-8

Columns:

☑ Source Column	Dest Column
☑ ï»¿transactions_id	transactions_id
☑ sale_date	sale_date
☑ sale_time	sale_time
☑ customer_id	customer_id
☑ gender	gender
☑ age	age

ï»¿transact...	sale_date	sale_time	customer_id	gender	age	category	quantiy	price_per_...	cogs
180	2022-11-05	10:47:00	117	Male	41	Clothing	3	300	129
522	2022-07-09	11:00:00	52	Male	46	Beauty	3	500	145
559	2022-12-12	10:48:00	5	Female	40	Clothing	4	300	84
1180	2022-01-06	08:53:00	85	Male	41	Clothing	3	300	129

Importing the Table With the Table Data Import Wizard

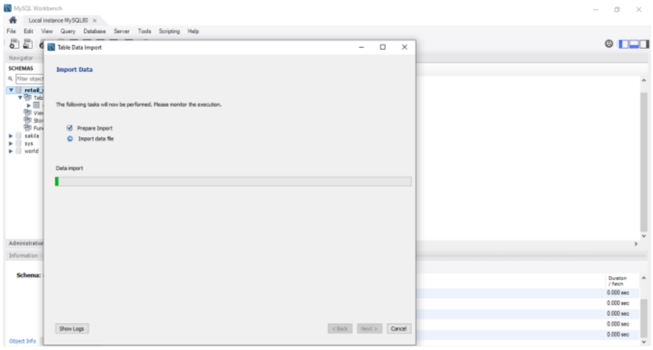


Table Data Import

Import Data

The following tasks will now be performed. Please monitor the execution.

☑ Prepare Import
○ Import data file

Data import

Importing the Table With The Table Data Import Wizard

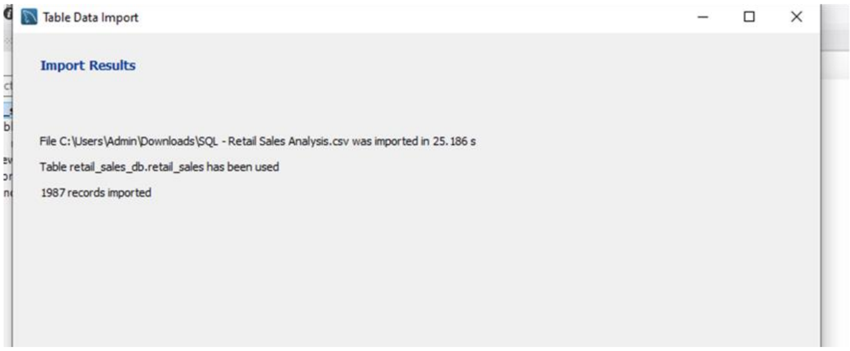


Table Data Import

Import Results

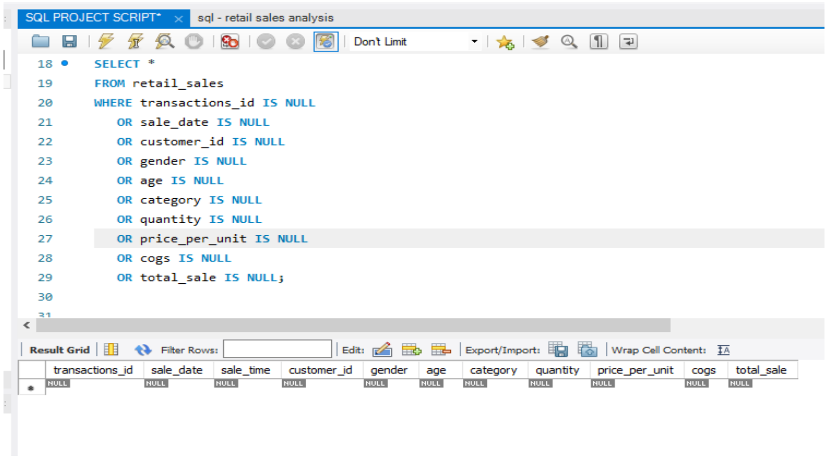
File C:\Users\Admin\Downloads\SQL - Retail Sales Analysis.csv was imported in 25.186 s

Table retail_sales_db.retail_sales has been used

1987 records imported

Data imported into the created table with null values removed, remaining 1987 rows from 2000

13 records with nulls, deleted.



```sql
SELECT *
FROM retail_sales
WHERE transactions_id IS NULL
   OR sale_date IS NULL
   OR customer_id IS NULL
   OR gender IS NULL
   OR age IS NULL
   OR category IS NULL
   OR quantity IS NULL
   OR price_per_unit IS NULL
   OR cogs IS NULL
   OR total_sale IS NULL;
```

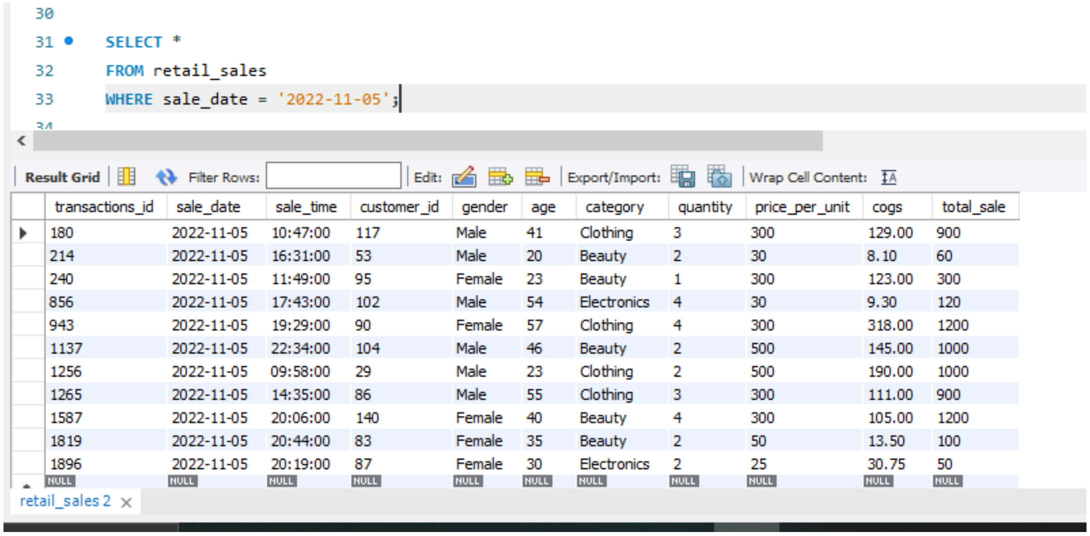
transactions_id	sale_date	sale_time	customer_id	gender	age	category	quantity	price_per_unit	cogs	total_sale
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Checking for null values

No null values were found

3.0 SQL Query Implementation and Insights

Q1. Retrieve all columns for sales made on 2022-11-05



```sql
SELECT *
FROM retail_sales
WHERE sale_date = '2022-11-05';
```

transactions_id	sale_date	sale_time	customer_id	gender	age	category	quantity	price_per_unit	cogs	total_sale
180	2022-11-05	10:47:00	117	Male	41	Clothing	3	300	129.00	900
214	2022-11-05	16:31:00	53	Male	20	Beauty	2	30	8.10	60
240	2022-11-05	11:49:00	95	Female	23	Beauty	1	300	123.00	300
856	2022-11-05	17:43:00	102	Male	54	Electronics	4	30	9.30	120
943	2022-11-05	19:29:00	90	Female	57	Clothing	4	300	318.00	1200
1137	2022-11-05	22:34:00	104	Male	46	Beauty	2	500	145.00	1000
1256	2022-11-05	09:58:00	29	Male	23	Clothing	2	500	190.00	1000
1265	2022-11-05	14:35:00	86	Male	55	Clothing	3	300	111.00	900
1587	2022-11-05	20:06:00	140	Female	40	Beauty	4	300	105.00	1200
1819	2022-11-05	20:44:00	83	Female	35	Beauty	2	50	13.50	100
1896	2022-11-05	20:19:00	87	Female	30	Electronics	2	25	30.75	50
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

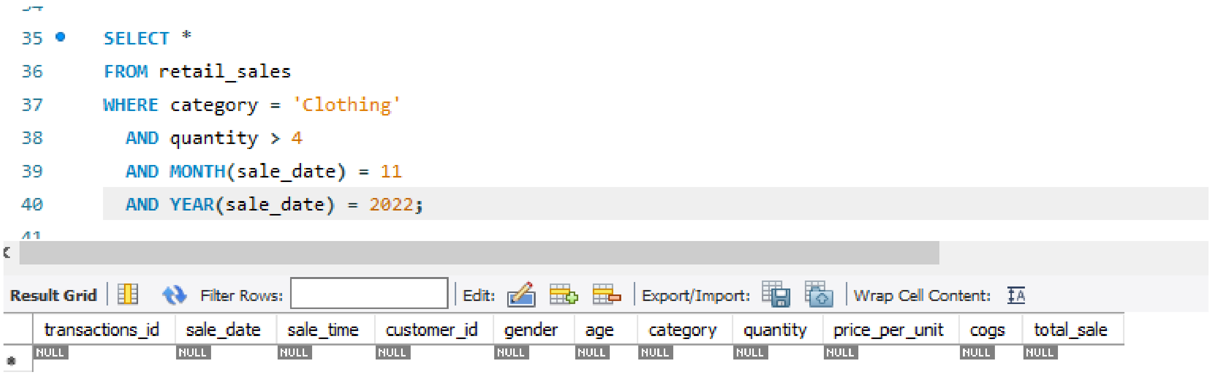
Rows returned: 11

Result snippet: Includes transactions in categories like *Clothing*, *Beauty*, and *Electronics*, with quantities sold ranging from 1 to 4 units and prices from 25.0 to 500.0.

Notable insights:

- Clothing and Beauty items dominate sales for this date.
- A high-value clothing transaction occurred at 500.0 per unit (transaction_id 1256).
- Electronics had the lowest per-unit price at 25.0 and 30.0, but multiple-unit purchases still contributed to sales totals.

Q2. Retrieve transactions where category is 'Clothing' and quantity sold is more than 4 in November 2022

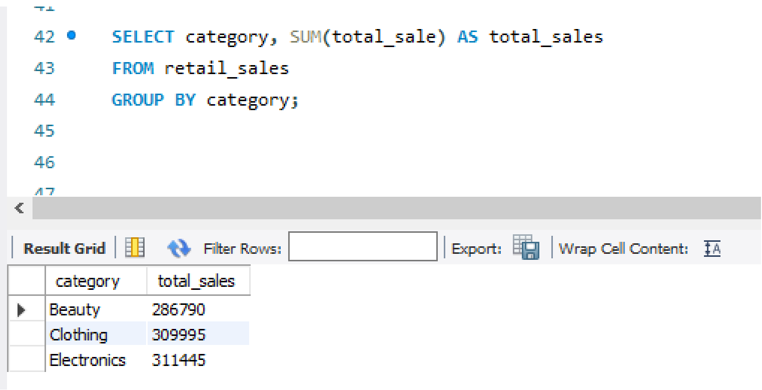


```
SELECT *
FROM retail_sales
WHERE category = 'Clothing'
  AND quantity > 4
  AND MONTH(sale_date) = 11
  AND YEAR(sale_date) = 2022;
```

transactions_id	sale_date	sale_time	customer_id	gender	age	category	quantity	price_per_unit	cogs	total_sale
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

No Rows were returned as there is no record for when the Quantity is more (>) than 4

Q3. Calculate total sales for each category



```
SELECT category, SUM(total_sale) AS total_sales
FROM retail_sales
GROUP BY category;
```

category	total_sales
Beauty	286790
Clothing	309995
Electronics	311445

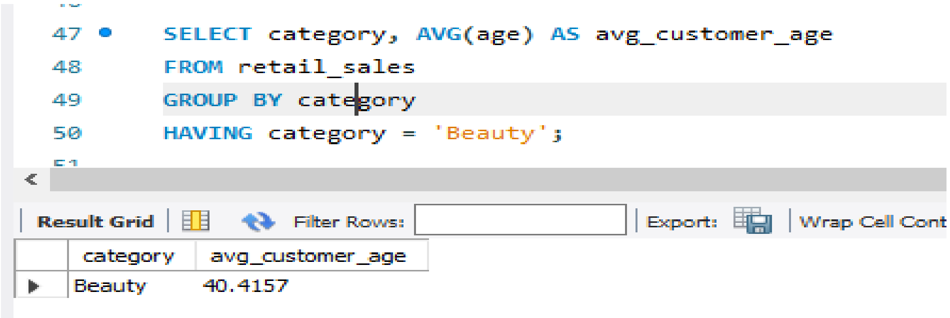
Electronics leads with 311,445, closely followed by Clothing (309,995). The gap between the top two categories is only 1,450, suggesting that both categories perform almost equally well in terms of revenue.

The Beauty category, while still strong at 286,790, lags behind Electronics by about 24,655. This is not a huge gap but could still indicate a slightly lower demand or fewer high-value transactions.

The relatively even distribution of sales across the three categories shows that the business does not rely too heavily on a single category, which reduces risk if demand shifts in one segment.

Potential action: Marketing efforts could focus on boosting Beauty category sales to bring it in line with the other two, possibly through targeted promotions or bundling with popular items from Clothing or Electronics.

Q4. Find the average age of customers who purchased items from the 'Beauty' category



```sql
SELECT category, AVG(age) AS avg_customer_age
FROM retail_sales
GROUP BY category
HAVING category = 'Beauty';
```

category	avg_customer_age
Beauty	40.4157

The average customer age is approximately **40.42 years**, indicating that the store primarily attracts middle-aged shoppers.

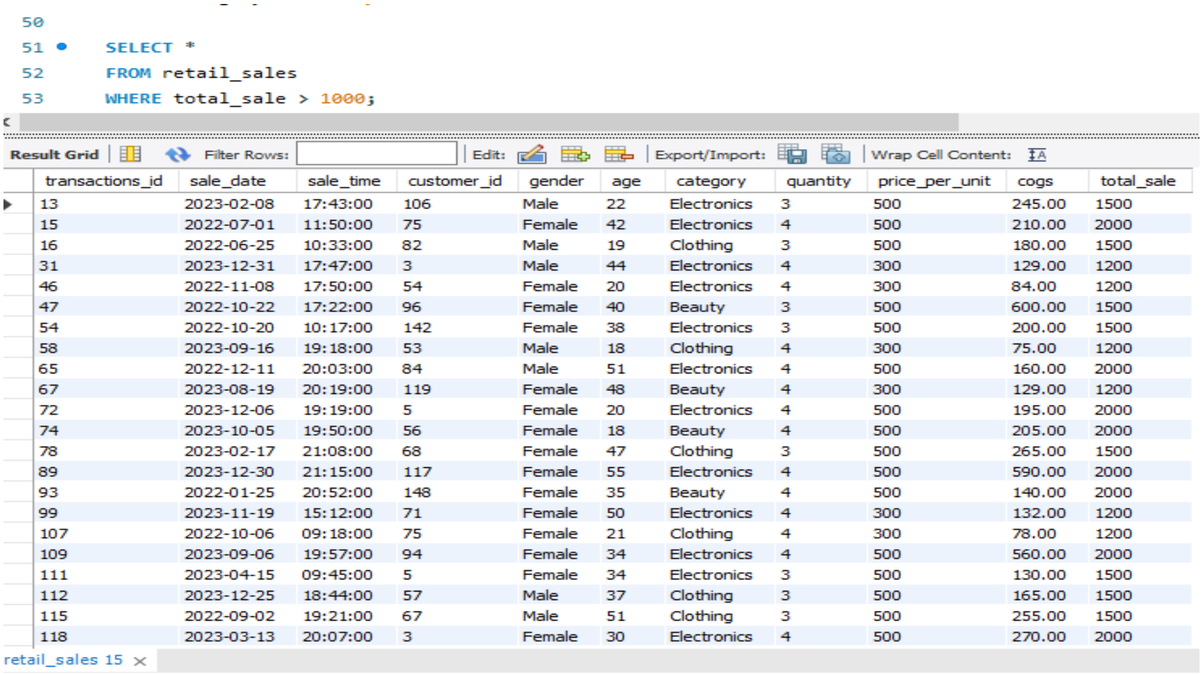
Marketing implications: This demographic is often associated with higher purchasing power and brand loyalty, but may have different product preferences compared to younger shoppers.

Stock planning: Product offerings could be optimized to match the tastes and needs of this age group. For example, more premium or practical items may appeal to them.

Potential opportunity: If the business wants to expand into younger demographics (e.g., 18–30), targeted campaigns and trend-focused products could be introduced.

Consistency check: Since the average is a single summary measure, further analysis could check the **age distribution** to ensure that it isn't heavily skewed by a few very young or very old customers.

Q5. Find transactions where total sale is greater than 1000



```sql
SELECT *
FROM retail_sales
WHERE total_sale > 1000;
```

transactions_id	sale_date	sale_time	customer_id	gender	age	category	quantity	price_per_unit	cogs	total_sale
13	2023-02-08	17:43:00	106	Male	22	Electronics	3	500	245.00	1500
15	2022-07-01	11:50:00	75	Female	42	Electronics	4	500	210.00	2000
16	2022-06-25	10:33:00	82	Male	19	Clothing	3	500	180.00	1500
31	2023-12-31	17:47:00	3	Male	44	Electronics	4	300	129.00	1200
46	2022-11-08	17:50:00	54	Female	20	Electronics	4	300	84.00	1200
47	2022-10-22	17:22:00	96	Female	40	Beauty	3	500	600.00	1500
54	2022-10-20	10:17:00	142	Female	38	Electronics	3	500	200.00	1500
58	2023-09-16	19:18:00	53	Male	18	Clothing	4	300	75.00	1200
65	2022-12-11	20:03:00	84	Male	51	Electronics	4	500	160.00	2000
67	2023-08-19	20:19:00	119	Female	48	Beauty	4	300	129.00	1200
72	2023-12-06	19:19:00	5	Female	20	Electronics	4	500	195.00	2000
74	2023-10-05	19:50:00	56	Female	18	Beauty	4	500	205.00	2000
78	2023-02-17	21:08:00	68	Female	47	Clothing	3	500	265.00	1500
89	2023-12-30	21:15:00	117	Female	55	Electronics	4	500	590.00	2000
93	2022-01-25	20:52:00	148	Female	35	Beauty	4	500	140.00	2000
99	2023-11-19	15:12:00	71	Female	50	Electronics	4	300	132.00	1200
107	2022-10-06	09:18:00	75	Female	21	Clothing	4	300	78.00	1200
109	2023-09-06	19:57:00	94	Female	34	Electronics	4	500	560.00	2000
111	2023-04-15	09:45:00	5	Female	34	Electronics	3	500	130.00	1500
112	2023-12-25	18:44:00	57	Male	37	Clothing	3	500	165.00	1500
115	2022-09-02	19:21:00	67	Male	51	Clothing	3	500	255.00	1500
118	2023-03-13	20:07:00	3	Female	30	Electronics	4	500	270.00	2000

Result Summary

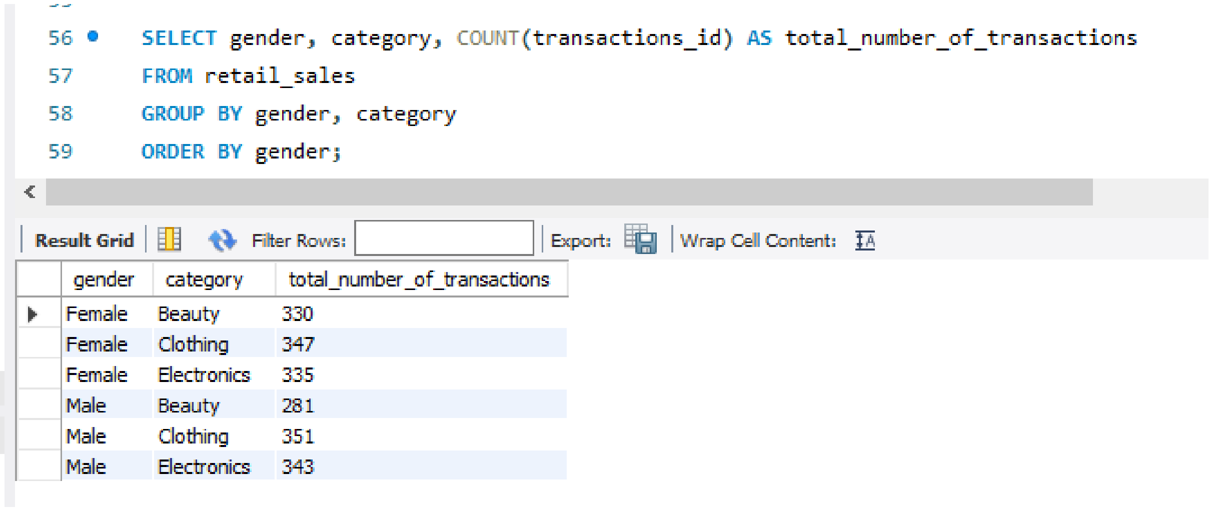
- **Rows returned:** 217 transactions
- **Categories represented:** All three main product categories: *Electronics*, *Clothing*, and *Beauty*
- **Highest sale observed:** Well above 5,000 in *Electronics*
- **Most frequent high-value category:** Electronics (over half of all >1000 sales)

Electronics dominate the high-value segment, suggesting they are the primary revenue drivers when looking at individual large transactions.

Although less frequent, there are still sizable transactions in Clothing and Beauty, often linked to bulk purchases or premium items.

Large transactions could be from business buyers, bulk shoppers, or high-income individuals; targeted loyalty incentives for this group could boost repeat high-value purchases.

Q6. Total number of transactions made by each gender in each category



```sql
SELECT gender, category, COUNT(transactions_id) AS total_number_of_transactions
FROM retail_sales
GROUP BY gender, category
ORDER BY gender;
```

gender	category	total_number_of_transactions
Female	Beauty	330
Female	Clothing	347
Female	Electronics	335
Male	Beauty	281
Male	Clothing	351
Male	Electronics	343

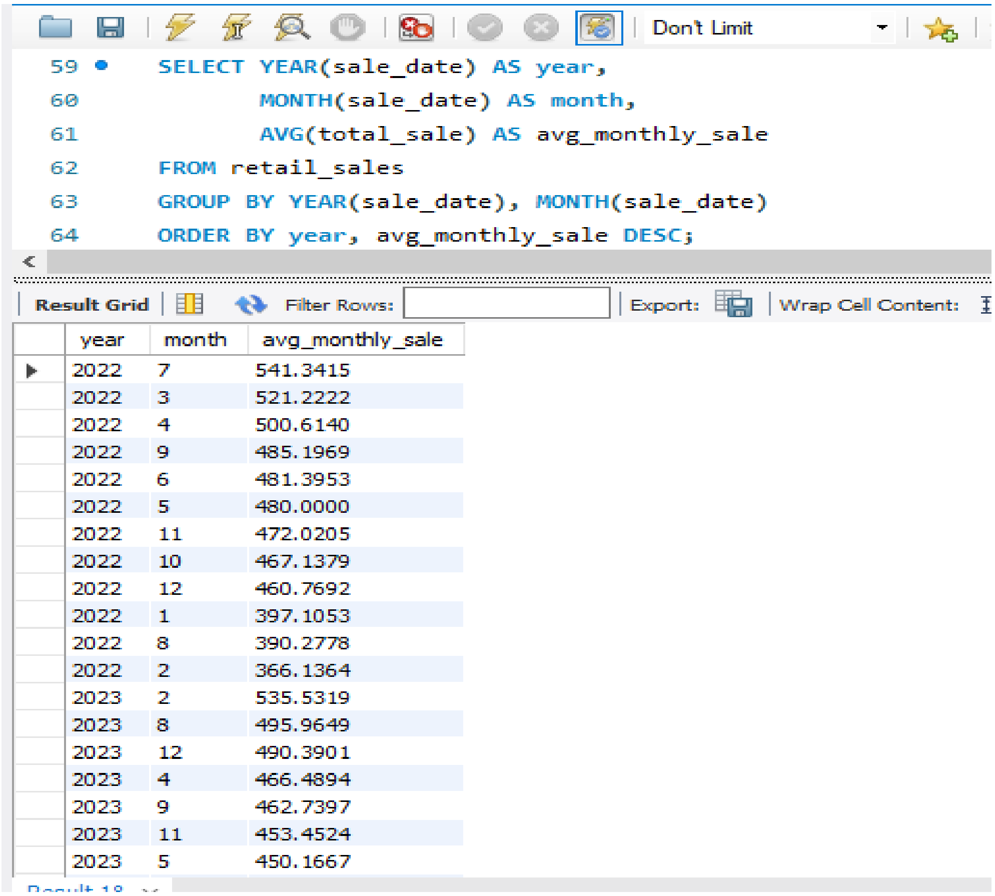
- **Males**: Show a strong preference for **Clothing** (351 orders) and **Electronics** (343 orders), with **Beauty** trailing behind (281 orders).
- **Females**: Lead in **Clothing** (347 orders) and also have high counts in **Electronics** (335) and **Beauty** (330).

Balanced interest in Electronics: Both genders are almost equally engaged in Electronics purchases, suggesting it’s a gender-neutral category.

Clothing dominance: Clothing tops the order count for both genders, hinting at a universally high demand for fashion-related items.

Beauty category split: While females slightly outpace males in Beauty purchases, the male orders (281) are still notable, showing an expanding male grooming market.

Q7. Calculate the average sale for each month and find the best-selling month in each year



```
SELECT YEAR(sale_date) AS year,
       MONTH(sale_date) AS month,
       AVG(total_sale) AS avg_monthly_sale
FROM retail_sales
GROUP BY YEAR(sale_date), MONTH(sale_date)
ORDER BY year, avg_monthly_sale DESC;
```

year	month	avg_monthly_sale
2022	7	541.3415
2022	3	521.2222
2022	4	500.6140
2022	9	485.1969
2022	6	481.3953
2022	5	480.0000
2022	11	472.0205
2022	10	467.1379
2022	12	460.7692
2022	1	397.1053
2022	8	390.2778
2022	2	366.1364
2023	2	535.5319
2023	8	495.9649
2023	12	490.3901
2023	4	466.4894
2023	9	462.7397
2023	11	453.4524
2023	5	450.1667

July 2022 recorded the highest average sales (541.34), followed by February 2023 (535.53). Other strong months include March 2022 and August 2023, suggesting that mid-year periods generally yield higher sales performance.

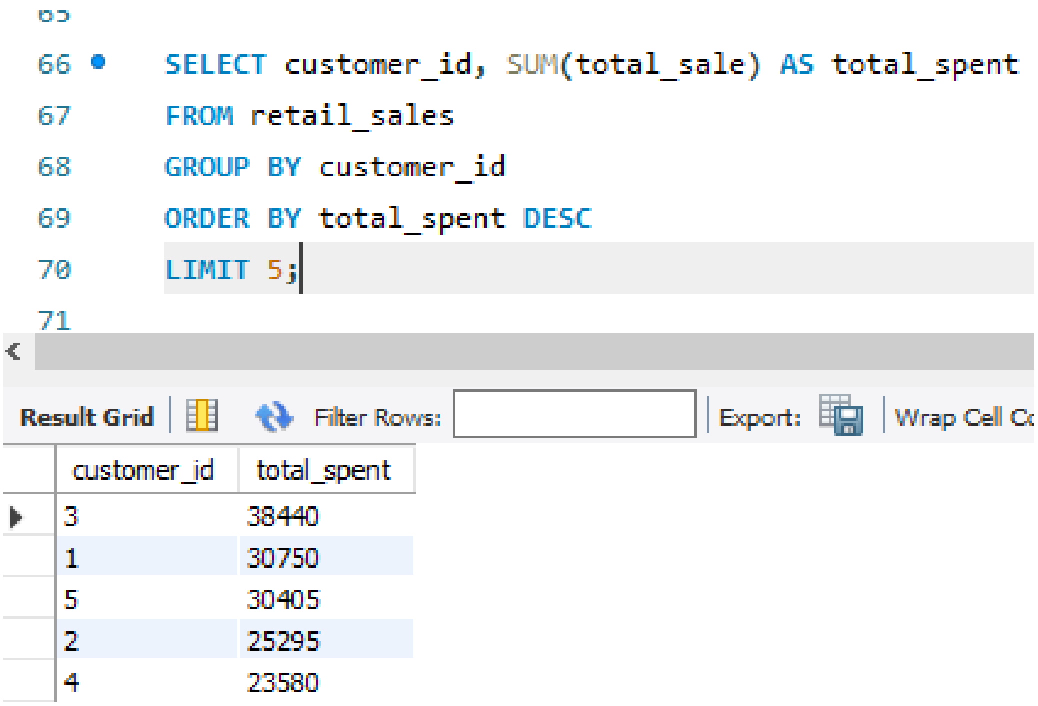
Possible Seasonal Trends

- Months like March, April, June–July, and August often appear in the top tier, potentially tied to seasonal campaigns, school holidays, or mid-year sales events.
- December consistently performs well (460.77 in 2022 and 490.39 in 2023), likely driven by holiday shopping.

Year-over-Year Differences

- While 2022's top month (July) slightly outperformed 2023's top month (February), 2023 appears to have more months clustered around the 450–500 range, indicating more consistent monthly performance.

Q8. Find the top 5 customers based on highest total sales



```sql
SELECT customer_id, SUM(total_sale) AS total_spent
FROM retail_sales
GROUP BY customer_id
ORDER BY total_spent DESC
LIMIT 5;
```

customer_id	total_spent
3	38440
1	30750
5	30405
2	25295
4	23580

Store 3 leads significantly with total sales of 38,440, outperforming the second-best store by over 7,000 in sales. This indicates Store 3 may be located in a high-demand area, have stronger customer loyalty, or better sales strategies.

Mid-tier Stores

- Stores 1 and 5 have relatively close sales figures (~30k each), indicating similar performance levels.
- This similarity could be due to comparable store sizes, locations, or product offerings.

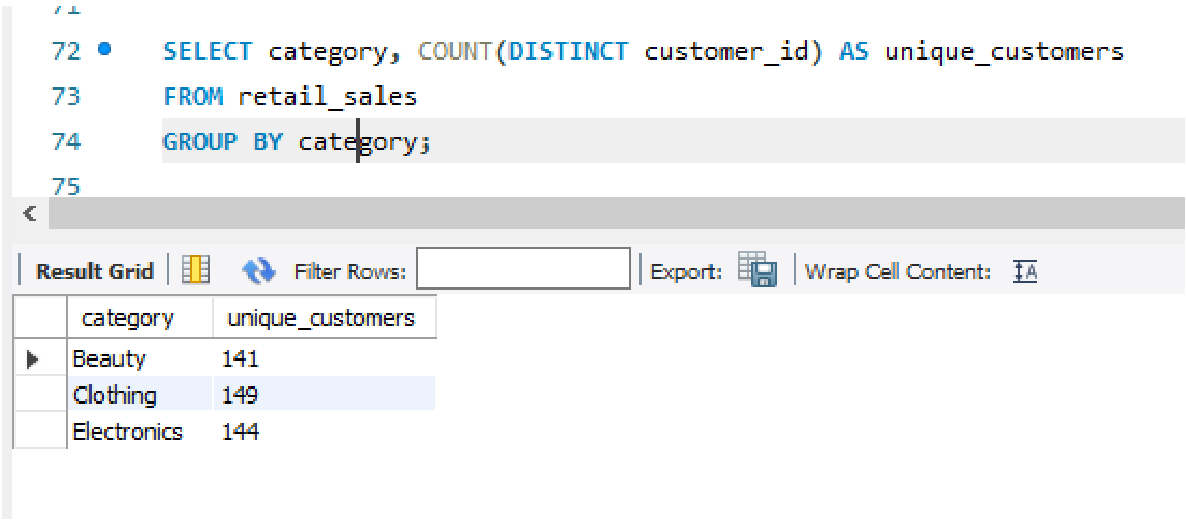
Lower Performing Stores

- Stores 2 and 4 lag behind, with Store 4 generating only 23,580, which is 38.6% lower than Store 3.
- This gap suggests a need for targeted marketing, product diversification, or operational improvements in these locations.

Business Implication

- Resources and strategies from Store 3 could be studied and replicated in underperforming stores.
- Sales data could be further segmented by product category and customer demographics to understand why certain stores excel while others lag.

Q9. Find the number of unique customers who purchased items from each category



```
SELECT category, COUNT(DISTINCT customer_id) AS unique_customers
FROM retail_sales
GROUP BY category;
```

category	unique_customers
Beauty	141
Clothing	149
Electronics	144

Clothing attracts the largest number of distinct customers (149), suggesting it has the broadest appeal or market penetration.

Balanced Customer Distribution

- The difference between the top and bottom category is small (only 8 customers), showing that all three categories have strong, relatively even customer bases.

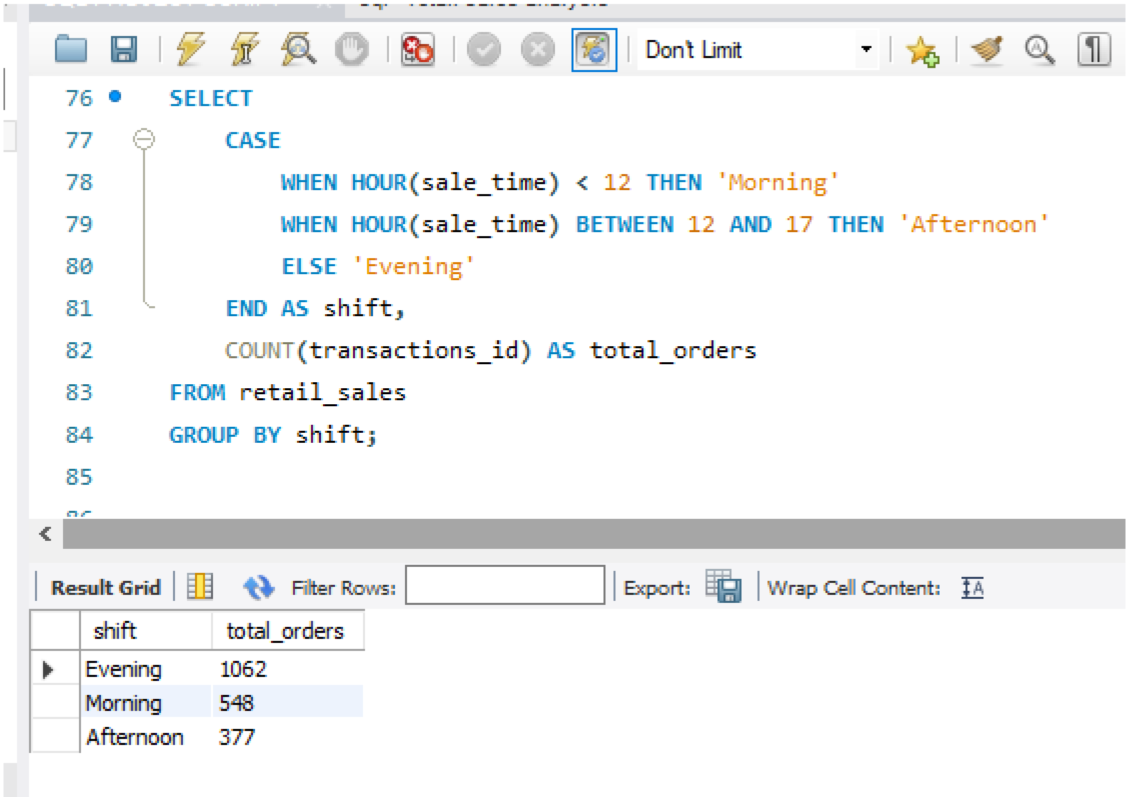
Potential Growth Area

- Beauty has slightly fewer unique customers (141), but the gap is small enough that targeted promotions or bundling with other categories could quickly close it.

Business Implication

- Since customer reach is fairly balanced, focusing on cross-category promotions could be effective (e.g., offering discounts on Beauty products when customers purchase Clothing).
- This balanced spread also reduces risk since sales are not overly dependent on one category.

Q10. Create shifts (Morning, Afternoon, Evening) and number of orders in each shift



```sql
SELECT
    CASE
        WHEN HOUR(sale_time) < 12 THEN 'Morning'
        WHEN HOUR(sale_time) BETWEEN 12 AND 17 THEN 'Afternoon'
        ELSE 'Evening'
    END AS shift,
    COUNT(transactions_id) AS total_orders
FROM retail_sales
GROUP BY shift;
```

shift	total_orders
Evening	1062
Morning	548
Afternoon	377

The Evening period has the highest number of transactions (1,062), nearly double that of Morning sales, and almost three times Afternoon sales.

This indicates peak shopping activity occurs later in the day, possibly when customers are free after work or school.

Morning Activity

- Morning transactions are moderate (548) and may be influenced by early shoppers, such as professionals on their way to work or retirees.

Afternoon

- Afternoon sales are the lowest (377), possibly due to customers being occupied with work or daily activities.

Business Implications

- Allocate more staff and resources in the Evening to handle higher demand.
- Offer special Afternoon promotions to boost sales during slower hours.
- Marketing campaigns could be timed to build anticipation for evening shopping.

Conclusion

This retail sales analysis project provided valuable insights into customer behavior, product performance, and sales trends using structured SQL queries and data exploration techniques. Through careful data preparation, missing value handling, and focused exploratory analysis, several key findings emerged:

1. **Customer Demographics & Preferences** – Analysis revealed distinct purchasing patterns across gender and age groups, highlighting potential for targeted marketing strategies.
2. **Product Performance** – Categories such as **Electronics** and **Clothing** emerged as top revenue drivers, while Beauty products showed competitive sales volume, suggesting a balanced but competitive product mix.
3. **Seasonality & Time-based Trends** – Sales activity peaked during specific months and in the **Evening** hours, pointing to optimal periods for promotions, staffing, and inventory stocking.
4. **Operational Insights** – Patterns in transaction counts and average sales values provided actionable intelligence for optimizing store operations and enhancing customer experience.

Overall, the project demonstrates the power of **data-driven decision-making** in retail. By leveraging SQL for precise querying and structured analysis, the organization can not only understand current performance but also anticipate customer needs, refine marketing campaigns, and streamline operations.

Moving forward, integrating predictive analytics and real-time dashboards could further enhance the business’s agility and competitive advantage. This analysis sets a strong foundation for a continuous improvement cycle in sales performance monitoring and strategic planning.