📊 Retail Sales Analysis using SQL

📌 Project Overview

This project focuses on analyzing retail sales data using **SQL** to extract meaningful business insights. The analysis answers common business questions related to sales performance, customer behavior, time-based trends, and operational patterns.

This project is designed as a **Data Analyst portfolio project**, showcasing strong SQL fundamentals and analytical thinking.

🕐 Objectives

- Analyze sales performance across different **categories** and **time periods**
- Identify **top customers** and **high-performing months**
- Understand customer behavior based on **gender, age, and purchase quantity**
- Perform **data cleaning** by handling NULLs and duplicates
- Use **advanced SQL concepts** like CTEs and window functions

---

🗂️ Dataset Description

The dataset contains transactional retail sales data with the following columns:

- `transactions_id` – Unique transaction identifier
- `sale_date` – Date of transaction
- `sale_time` – Time of transaction
- `customer_id` – Unique customer identifier
- `gender` – Gender of the customer
- `age` – Age of the customer
- `category` – Product category (Clothing, Beauty, etc.)
- `quantity` – Quantity sold
- `price_per_unit` – Price per unit
- `cogs` – Cost of goods sold
- `total_sale` – Total sales value per transaction

🛠️ Tools & Technologies

- **MySQL**
- **SQL** (DDL, DML, Aggregations)
- **MySQL Workbench**

🔑 SQL Concepts Used

- `SELECT`, `WHERE`, `GROUP BY`, `ORDER BY`
- Aggregate functions (`SUM`, `AVG`, `COUNT`, `MIN`, `MAX`)
- **Common Table Expressions (CTEs)**
- **Window Functions** (`RANK()`)
- Date & time functions (`YEAR`, `MONTH`, `MONTHNAME`, `HOUR`)
- Data cleaning (NULL checks, duplicate removal)

📊 Key Analysis Performed

✂️ Data Cleaning

- Identified and removed **NULL values**
- Checked and removed **duplicate transactions**
- Prepared data for primary key constraints

✒️ Sales Analysis

- Average monthly sales
- Best-selling month in each year
- Category-wise and gender-wise transaction counts
- Top 5 customers by total sales

🛏️ Time-Based Analysis

- Sales analysis by **year and month**
- Shift-based order analysis:
- Morning (< 12 PM)
- Afternoon (12–5 PM)
- Evening (> 5 PM)

🌂 Customer Insights

- Average age per category
- Gender-wise purchase behavior
- High-value and low-value transactions

Sample Business Questions Answered

- Which month performs best in each year?
- Who are the top 5 customers by total sales?
- Which product category has the highest number of transactions?
- How do sales vary across different times of the day?

- Are there duplicate or invalid transactions in the dataset?

📈 Key Insights (Example)

- Certain months consistently outperform others in terms of average sales
- Evening shifts tend to have higher order volumes
- A small group of customers contributes significantly to total revenue
- Clothing and Beauty categories show different customer demographics

---

🗜️ How to Run This Project

1. Import the dataset into **MySQL**
2. Create the `Retail_sales` table
3. Run the SQL script: `Retail_sales_analysis.sql`
4. Execute queries step by step to view insights

📎 Project Structure

```
├── Retail_sales_analysis.sql
├── README.md
└── Dataset (CSV file)
```

🔮 Future Improvements

- Add **indexes** for performance optimization
- Perform **category-level profitability analysis**
- Integrate with **Power BI / Tableau** for visualization
- Automate data loading using Python

Author

Harsh Kumar
Aspiring Data Analyst | SQL | Data Analytics

⭐ If you found this project useful, feel free to star the repository!