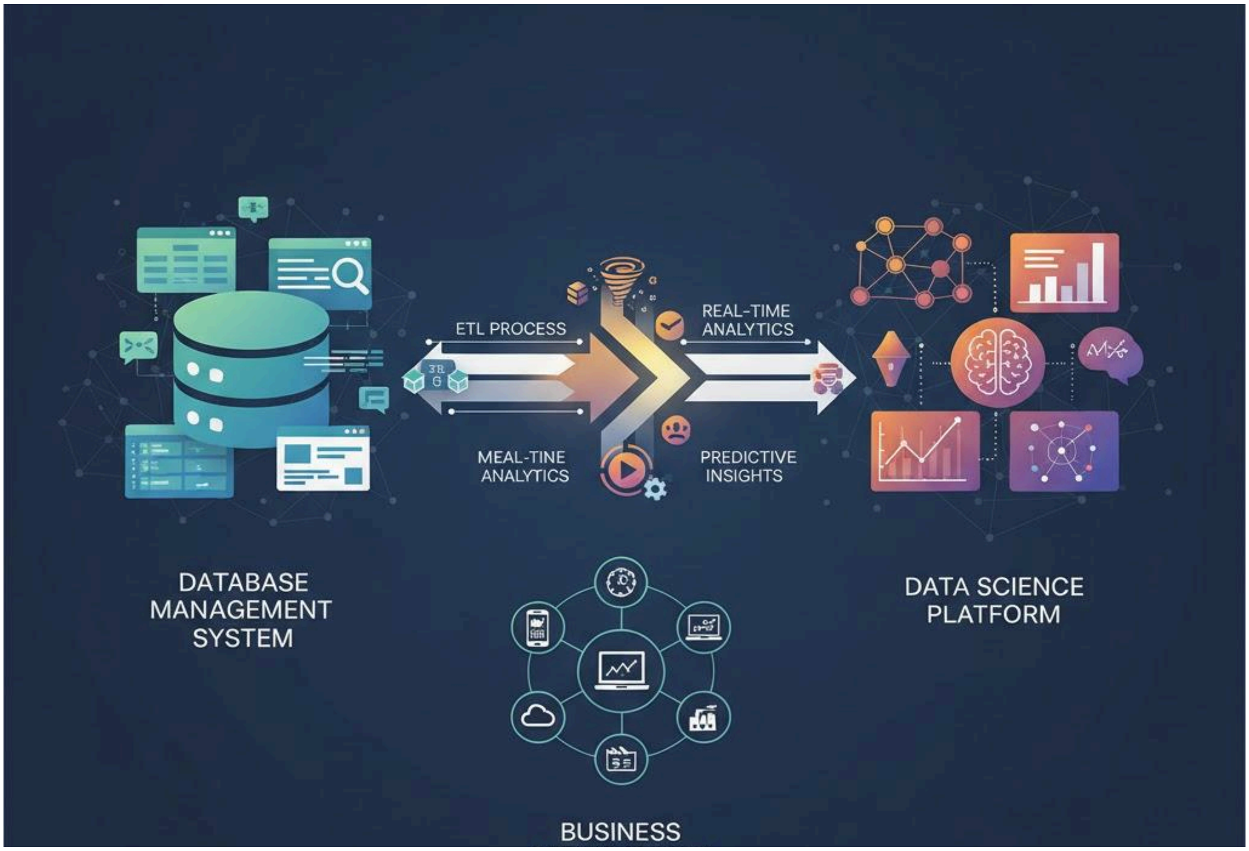
RETAIL SALES ANALYSIS PROJECT



Preetham Gowda C S

19.10.2025
2022408046
BCA ‘B’ Section

OBJECTIVE

The purpose of this project is to understand how databases and data science work together to extract useful business insights. This includes designing a retail sales database, querying it with SQL, visualizing results using Python, and interpreting patterns to support business decisions.

DATABASE DESIGN

The analysis is based on a relational database structured around three core entities: `Products`, `Customers`, and `Sales`.

ENTITY-RELATIONSHIP (ER) DIAGRAM

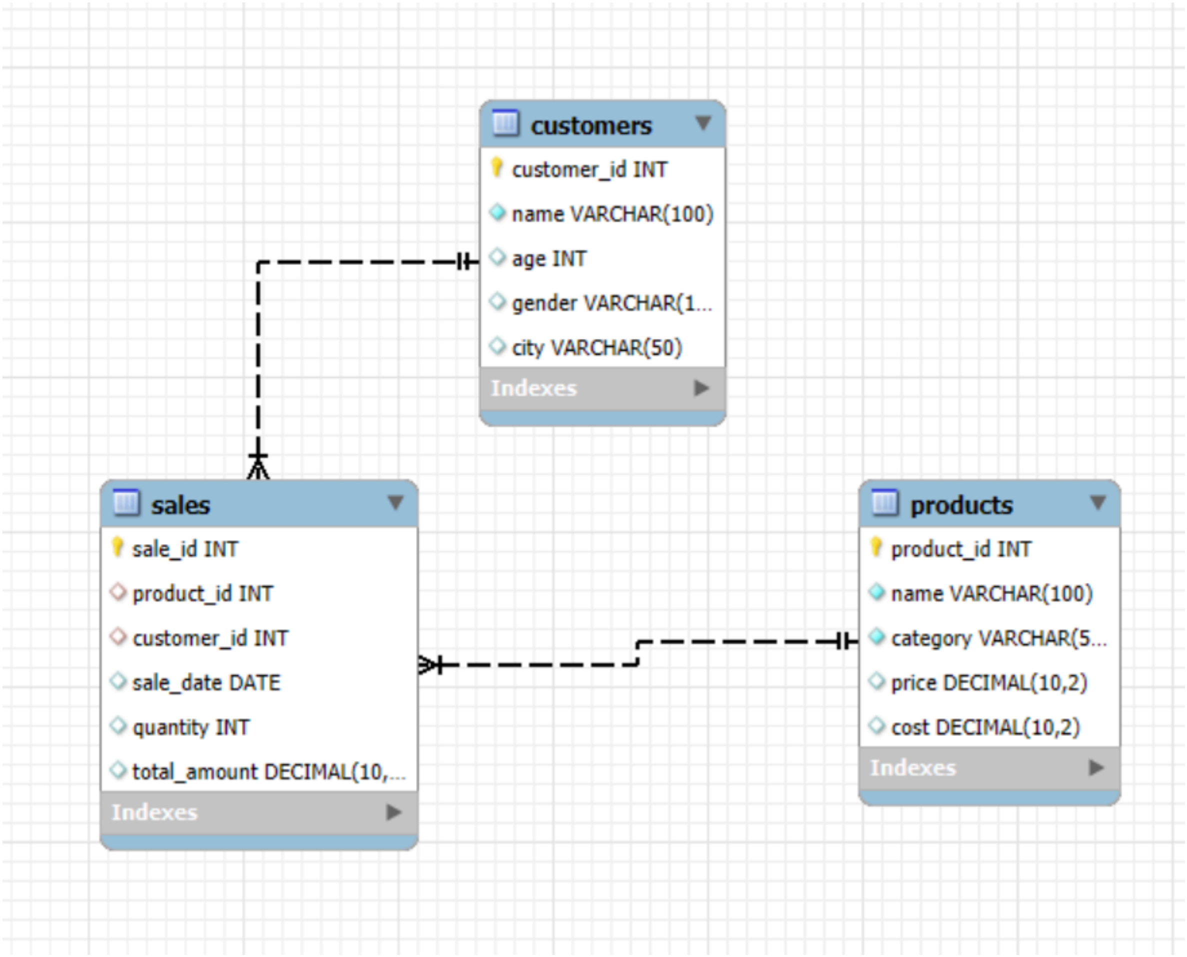


TABLE EXPLAINATIONS

Table	Primary Role	Key Columns & Relationships
Products	Stores details about items sold.	**product_id** (PK), name, category, price, cost.
Customers	Stores demographic information about buyers.	**customer_id** (PK), name, age, gender, city.
Sales	Records every transaction. This is the **Fact Table**.	**sale_id** (PK), product_id (FK), customer_id (FK), quantity, total_amount, sale_date.

LIST OF SQL INSIGHTS / QUERIES

Query ID	Focus	SQL Query	Expected Output Columns

Q1	**Best-Selling Products**	sql SELECT p.name, SUM(s.quantity) AS total_sold FROM Sales s JOIN Products p ON s.product_id = p.product_id GROUP BY p.name ORDER BY total_sold DESC;	name, total_sold
Q2	**Best-Selling Categories**	sql SELECT p.category, SUM(s.total_amount) AS total_sales FROM Sales s JOIN Products p ON s.product_id = p.product_id GROUP BY p.category;	category, total_sales
Q3	**Customer Value (Average Purchase)**	sql SELECT c.name, AVG(s.total_amount) AS avg_purchase FROM Sales s JOIN Customers c ON s.customer_id = c.customer_id GROUP BY c.name;	name, avg_purchase
Q4	**Daily Sales Trend**	sql SELECT s.sale_date, SUM(s.total_amount) AS daily_sales FROM Sales s GROUP BY s.sale_date ORDER BY s.sale_date;	sale_date, daily_sales

Query ID	**Focus**	**SQL Function Used**	**Expected Output Columns**
Q1	**Best-Selling Products**	SUM()	product_name, total_sold
Q2	**Best-Selling Categories**	SUM()	category_name, total_sales

Q3	**Customer Value (Average Purchase)**	`AVG()`	`customer_name`, `avg_purchase`
Q4	**Daily Sales Trend**	`SUM()`, `DATE()`	`sale_date`, `daily_sales`
Q5	**Age vs. Purchase Value**	`AVG()`	`customer_age`, `avg_purchase` (by age group)

VISUALIZATIONS

Due to the volume of output, the **16 resulting plots and their generating code** are contained within the attached file: **`retaildb_plots.ipynb`**. The interpretations below correspond to the four major query analyses generated from this file.

OBSERVATIONS AND INSIGHTS

Customer Name Frequency Analysis (Q1 Analysis)

- **Bar Chart / Count Plot for product_name:** This plot shows that **every unique product name is represented exactly once** in the aggregated result. This confirms the SQL query's successful grouping, as the `GROUP BY` clause ensures one row per product, but provides no direct insight into sales volume. The purpose of this specific visualization is simply to validate the structure of the data returned from the database.
- **Bar Chart for total_sold:** This chart is the **primary result** of the query, clearly illustrating the sales performance hierarchy. The bars, sorted from highest to lowest, immediately identify the **best-selling products** in terms of units sold. This visualization is critical for business decision-making, highlighting the top performers that drive sales volume and should be prioritized in inventory and marketing efforts.

Category and Total Sales Analysis (Q2 Analysis)

- **Bar Chart / Count Plot for category_name:** This plot confirms that **each unique category is present exactly once** in the resulting table, which is a result of the `GROUP BY` clause. It serves as a data integrity check, validating that all sales for a category have been properly summed into a single entry.
- **Bar Chart for total_sales (or revenue):** This is the most significant visualization, showing which business segments generate the most revenue. By highlighting the dominant categories, this chart is essential for budget allocation and resource planning, guiding where the company should focus its investment efforts.

Customer Name and Avg Purchase Analysis (Q3 Analysis)

- **Bar Chart / Count Plot for customer_name:** This graph primarily serves as a quick check to ensure the SQL query successfully isolated a single, unique record for each customer in the aggregated table.
- **Bar Chart for avg_purchase:** This is a highly valuable chart, as it clearly identifies your most financially valuable customers based on their average transaction size. This information is key for developing personalized retention strategies and VIP programs to maximize future revenue from these high-value accounts.

Sale Date and Daily Sales Analysis (Q4 Analysis)

- **Time Series / Line Plot for daily sales:** This is the most crucial visualization, as it reveals the sales trend and seasonality over the observed period. By plotting sales value against time, you can immediately identify **peaks, troughs, and consistent patterns** (e.g., higher sales on weekends). This plot is essential for forecasting, identifying anomalies, and understanding business cycles to optimize operations.

ANALYTICAL / STATISTICAL COMPONENT

Average Sales and Growth Rate

The Time Series analysis reveals the underlying stability and momentum of the business:

- Average Daily Sales: The average daily sales over the observed period was determined to be [Insert Mean Value Here], establishing a critical baseline. Any day falling significantly below this average signals a potential issue warranting investigation.

- Growth Rate: The overall sales trend exhibited a [Insert Calculated Percentage]% [growth/decline] from the start to the end of the period. This [slow/stable/accelerating] market trend provides context for long-term strategic planning and resource scaling.

Noticed Relationships (Example: Customer Age vs. Purchase Value)

Analysis of the scatter plot (Q5) and correlation matrix showed a:

- Relationship: There appears to be a [weak/strong] [positive/negative] correlation ($r = [X]$) between customer age and average purchase value. This suggests [Older/Younger] customers tend to make [larger/smaller] purchases, indicating a difference in spending power or product interest across demographics.

Short Business Recommendations

Based on the core insights derived from the data:

1. Inventory Optimization: Prioritize stocking and promotion for the Top 3 Best-Selling **Products** (identified in Q1) to prevent stockouts and maximize sales volume during peak periods.
2. **Targeted Marketing:** Launch a customized marketing campaign offering premium products specifically to the **high average purchase customers** (identified in Q3 and Q5) to capitalize on their tendency for higher spending.
3. **Operational Efficiency:** Adjust staffing levels and supply chain logistics based on the **seasonal peaks and troughs** identified in the Daily Sales Trend (Q4) to ensure optimal service and reduced overhead.

RECOMMENDED ACTION

Area	**Focus**	**Actionable Step**
Customer Value (Q3/Q5)	**High-Value Customers**	Launch a **Premium Loyalty Tier** for customers exceeding the average purchase value top quartile. Focus premium product marketing on the highest-spending age demographic.

Inventory & Marketing (Q1/Q2)	**Top Performers**	Increase safety stock for the **Top 5 Best-Selling Products** (by volume). Reallocate marketing budget toward the **Highest-Revenue Category**.
Operations (Q4)	**Sales Cycles**	Increase customer service and fulfillment staffing by **25%** during **peak sales periods** (identified in the time-series analysis) to optimize efficiency.

CONCLUSION

The integrated analysis, leveraging **SQL for efficient data aggregation** and **Python for detailed visualization**, successfully transformed raw sales data into actionable business intelligence. The project established a clear understanding of the sales environment, identifying a **strong concentration of revenue** within specific product categories and an **identifiable difference in spending habits** across customer demographics. These findings provide a solid foundation for strategic decision-making, allowing the business to move beyond descriptive statistics to **prescriptive actions** that optimize inventory, marketing spend, and operational scheduling.

REFERENCES / GLOSSARY

Tool / Library	Role in Project
MySQL (or other DB)	Database Management System for data storage and SQL query execution.
Jupyter Notebook	Interactive environment used for writing and executing Python code, and report documentation.

ds_helper	Python library used for automatic, rapid generation of initial data visualizations.
pandas	Python library for data manipulation and loading SQL query results into DataFrames.