**Summer Internship Project**

on

**“Sales Data ETL & Analytics on Databricks”**

by

**Shaik Mohammed Adnan**

**Reg.No. : 23005A0501**

**Under the guidance of**

**Ravi Sigicharla**

**Director, Varseno solutions**

****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY**

**ANANTAPUR**

**COLLEGE OF ENGINEERING (*Autonomous*)**

**ANANTHAPURAM - 515002**

**19/05/2025 -18/07/2025**

INDEX

1. Abstract
2. Objectives
3. Tools and Technologies
4. Project Overview
5. Methodologies
6. Key Findings
7. Conclusion
8. Future Scope

**Abstract**

This project explores end-to-end sales data analysis using Databricks and Apache Spark, focusing on understanding customer behavior, regional sales performance, and overall revenue trends. The dataset, sourced in CSV and JSON formats, was ingested using Databricks Volumes to ensure seamless and persistent access across collaborative sessions. Comprehensive data cleaning and transformation steps were carried out, including handling missing values, correcting inconsistent data types, and removing invalid or empty entries. Processed data was stored as Delta Lake tables, which provided robust support for ACID transactions, schema evolution, and time travel, enabling accurate historical comparisons and safe iterative development.

Key performance indicators (KPIs) were computed using Spark SQL and DataFrame operations, such as monthly revenue analysis, top-performing products, sales contribution by region, and a detailed churn analysis distinguishing between new and returning customers. Particular emphasis was placed on identifying retention patterns and understanding their impact on annual revenue. In addition, performance optimization techniques like partitioning and caching were applied to enhance query efficiency. The insights were visualized through dynamic, interactive dashboards built within Databricks, tailored to serve both technical and non-technical stakeholders.

This project not only illustrates the scalability and flexibility of the Databricks environment for big data analytics but also delivers actionable insights to support strategic business decisions.

**Objectives**

The primary objectives of this project were defined to ensure a structured approach to sales data analysis using the Databricks platform. The key goals included:

* **Data Ingestion**: Import sales records provided in CSV and JSON formats into Databricks using Volumes for persistent and scalable access.
* **Data Cleaning & Transformation**: Address missing values, correct data type inconsistencies, and perform necessary aggregations to prepare the data for analysis.
* **Delta Lake Integration**: Store cleaned and transformed datasets in Delta Lake tables, leveraging features such as ACID transactions and time travel for reliable data versioning and updates.
* **Analytical Querying**: Develop Spark SQL queries to derive key performance indicators (KPIs), such as monthly revenue trends, top-selling products, and customer retention metrics.
* **Visualization**: Create interactive dashboards within Databricks to present insights in a visually intuitive format for both technical and non-technical stakeholders.

**Tools & Technologies**

* **Databricks**: Served as the unified analytics platform, providing an environment for collaborative development, scalable data processing, and integrated visualization capabilities.
* **Apache Spark (PySpark)**: Used for distributed data processing and transformation. PySpark's ability to handle large datasets efficiently enabled fast execution of both DataFrame operations and SQL-based queries.
* **Delta Lake**: Implemented as the storage layer for transformed data, offering benefits such as ACID transaction support, schema validation, and time travel for versioned data management.
* **Databricks Volumes**: Utilized to ingest raw sales data in CSV and JSON formats, allowing consistent and persistent access to source files within the Databricks workspace.
* **Spark SQL**: Applied for querying structured datasets, enabling the calculation of key performance indicators (KPIs) such as revenue trends and customer segmentation.
* **Databricks SQL Dashboards**: Used to create interactive visualizations, providing business insights and enabling effective communication of trends to both technical and non-technical stakeholders.
* **Pandas**: Employed during initial stages for lightweight data inspection and sanity checks before large-scale transformations.

**Project Review**

This project provided valuable hands-on experience in building a structured data analytics workflow using the Databricks platform and Apache Spark. It offered insight into real-world data challenges, particularly when working with multi-format files and inconsistent records. While the data transformation processes were performed manually, the project maintained a clear and logical sequence, mimicking a standard data pipeline framework.

The use of Delta Lake for data storage proved highly beneficial, especially with features like schema enforcement and time travel, which allowed safe and trackable data updates across different transformation stages. Spark SQL was effectively used to compute business-relevant metrics such as monthly sales trends, customer churn behavior, and top-performing regions and products. The ability to switch between PySpark code and SQL queries provided flexibility throughout the analysis phase.

One of the most rewarding aspects of the project was creating dashboards using Databricks SQL. These visual reports added significant value by making the findings accessible to non-technical audiences, aiding in data-driven decision-making.

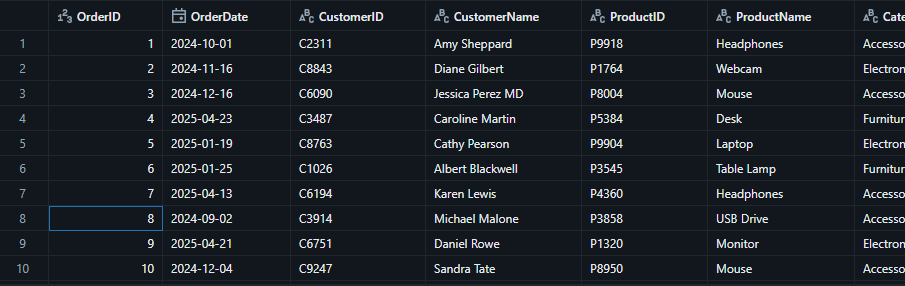
Although the pipeline was not fully automated, the structured approach ensured reproducibility and clarity. Overall, the project successfully demonstrated how a modern analytics platform like Databricks can be leveraged to derive insights from raw data and support strategic business goals.

**Methodology**

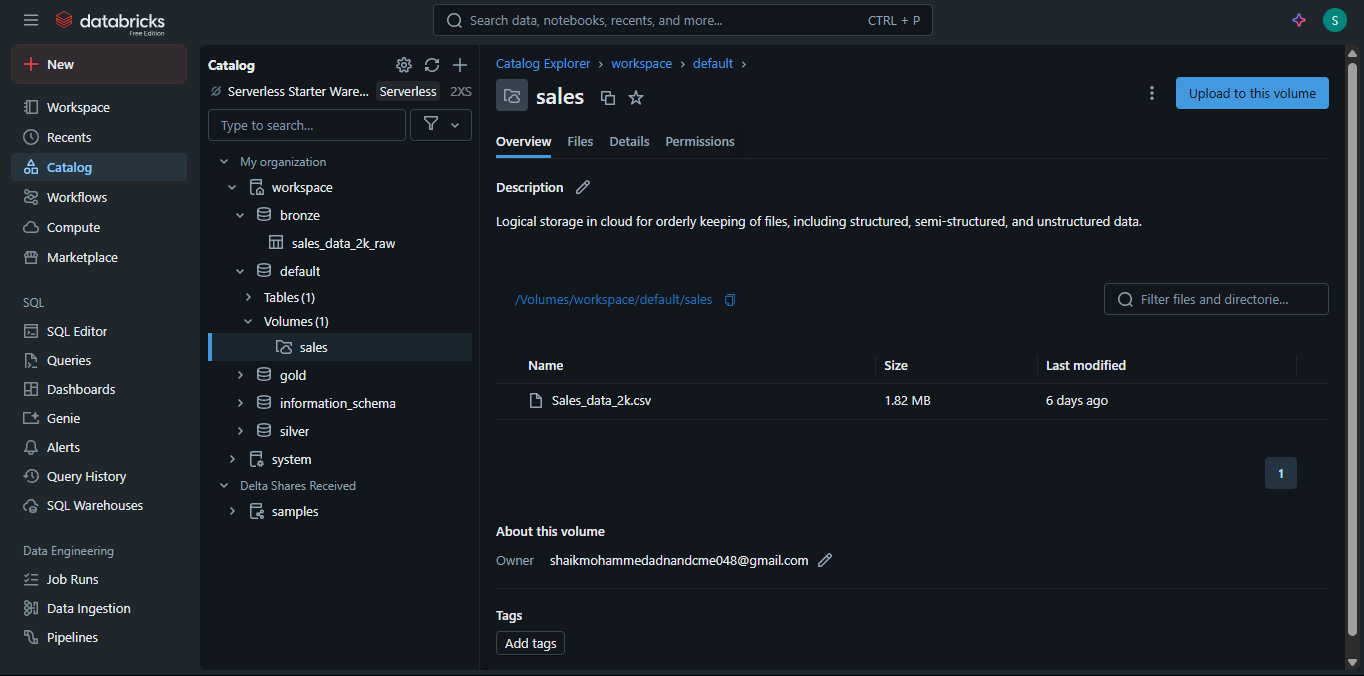
**Raw Data:**

The dataset used in this project consisted of sales transaction records in both **CSV** and **JSON** formats. These files included fields such as:

* OrderID, OrderDate, CustomerID, CustomerName
* ProductID, ProductName, Quantity, Unit Price
* Category, Sub-Category, Total Sales, Region

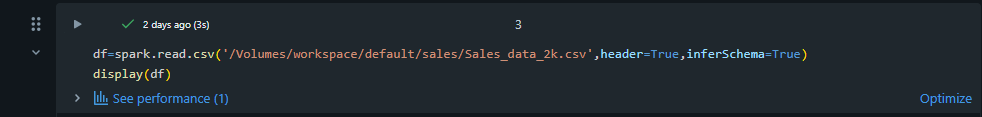


The files were stored in the Volumes feature of Databricks *(/Volumes/default/sales/raw-data),* to ensure consistent and persistent access to the raw data during development and analysis.



**Data Ingestion:**

Using Pyspark, the raw data was read from Volumes using:



Initial inspection showed:

* Missing values in ProductName and Total Sales
* Empty strings in CustomerName
* Mixed data types in OrderDate (string vs. date)

This step ensured that the raw dataset was available as a data frame for downstreaming process.



**Bronze Table (Raw Layer):**

In the Lakehouse architecture (commonly used in Databricks), data is organized into layers to manage its quality and purpose. The Bronze layer is the first layer where raw data is stored after ingestion.

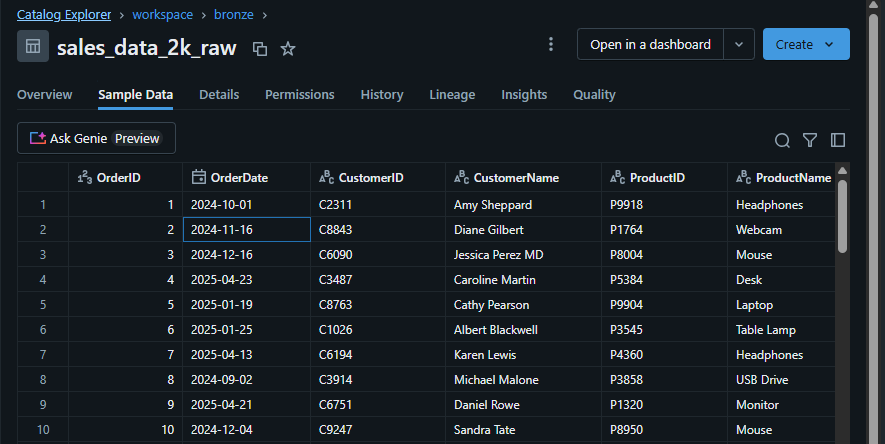
The raw data was written into a Bronze Delta Table without major transformation:

Purpose of Lake house architecture:

* Preserve original data for auditing
* Enable Time Travel (data versioning)
* Enforce schema structure

Table Schema**:**

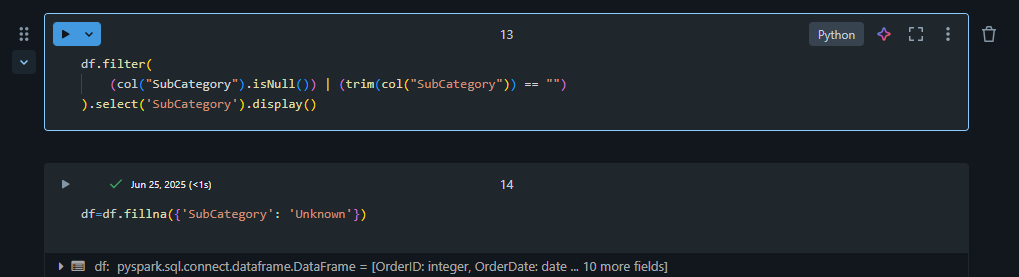
* OrderID: Integer
* OrderDate: String
* CustomerID: String
* ProductName: String
* Quantity: Integer
* Unit Price: Float
* Total Sales: Float
* Region: String



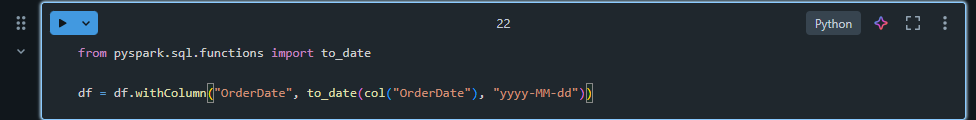
**Silver Table (Clean Layer):**

In order to make the bronze layer i.e, raw data into clean data without any litigations, we perform transformations to prepare data for analysis and the key transformations are:

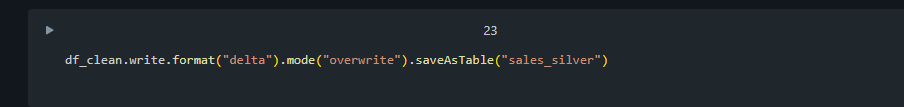
Missing values in ProductName, Total Sales handled using fillna() and logic:



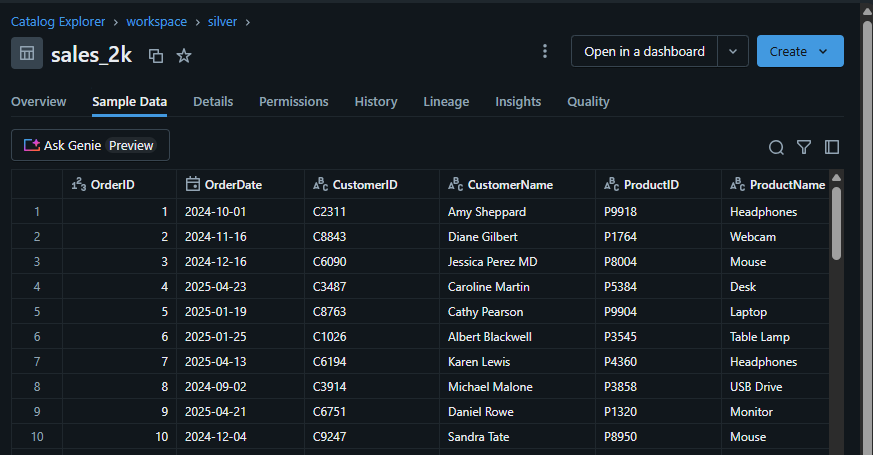
Data Parsing:



Data Type Consistency and deduplication also applied:



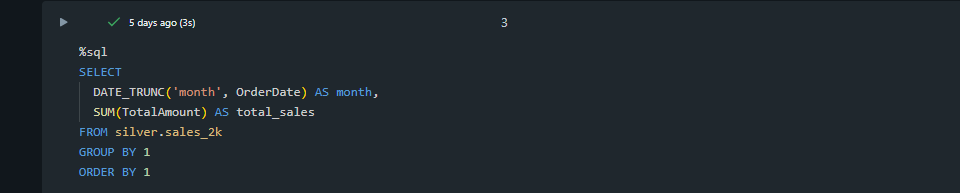
Once the deduplication is applied then the silver table is ready for data analysis



**Gold Table (Aggregated KPIs):**

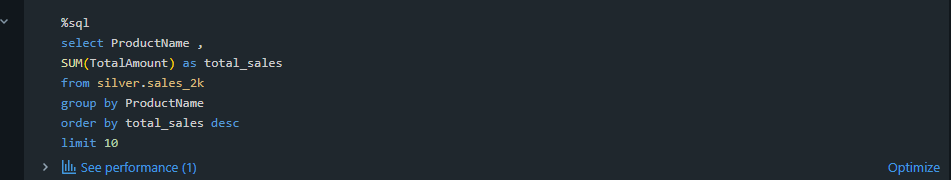
This layer focused on aggregating data to extract key business insights. Metrics were calculated using Spark SQL queries on the cleaned Silver Table.

Monthly Revenue:

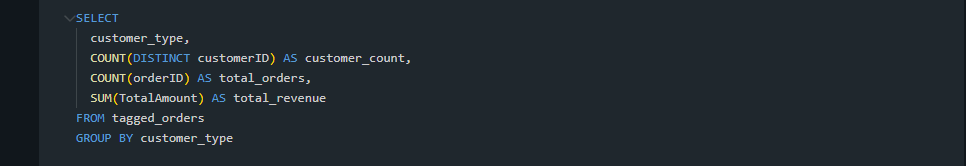


This query calculates the total revenue generated each month by truncating order dates to the month level. It helps identify revenue trends and seasonal sales patterns over time.

Top Products:

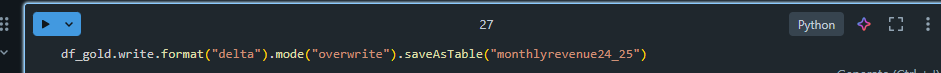


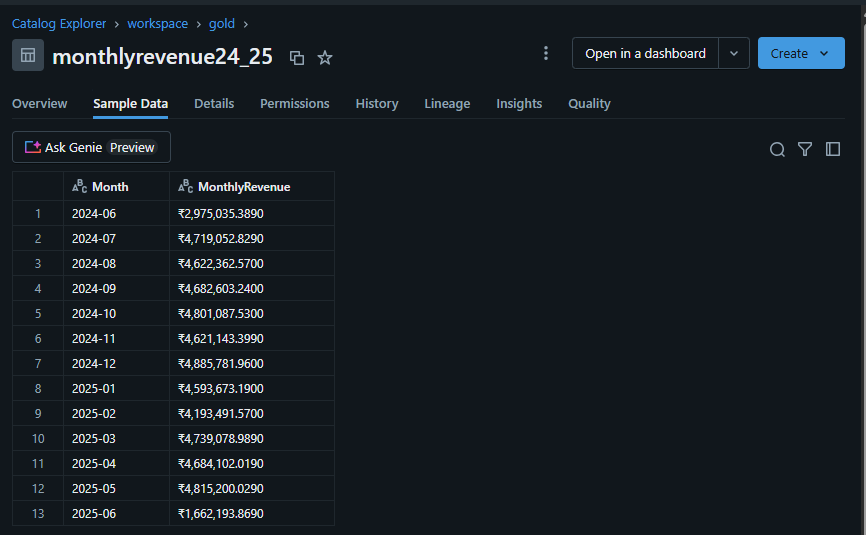
This query ranks products by total units sold, highlighting the top 10 best-selling items. It helps stakeholders understand customer demand and optimize inventory or promotions.

Churn Rate based on New vs Returning Customer Revenue:  


This query aggregates revenue based on customer type—either new or returning. It provides insight into customer retention and the contribution of loyal customers to overall sales.

Results from these queries were stored in the gold.monthlyrevenue24\_25 Delta Table for efficient access and visualization:





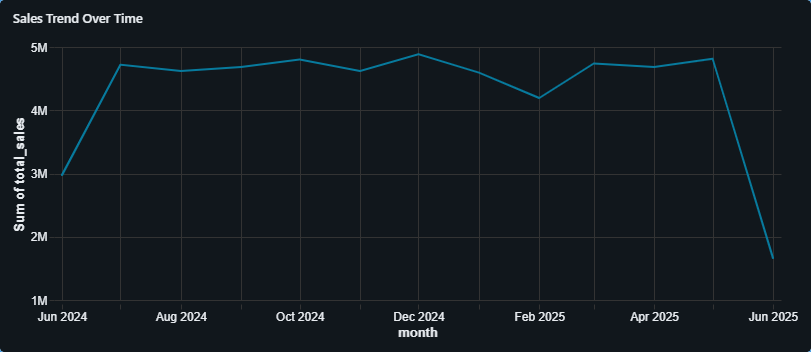
**Dashboards and Insights:**

The final output of the project involved building a set of interactive dashboards using Databricks SQL. These dashboards translated aggregated KPIs into visual insights to assist stakeholders in interpreting sales performance, customer behavior, and market trends.

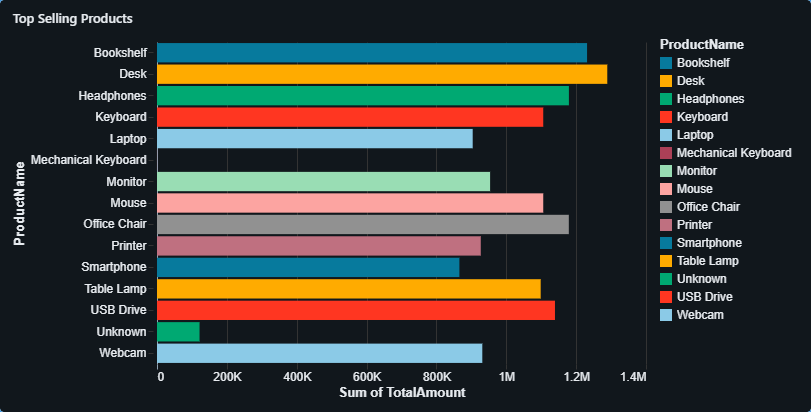
Each visualization was backed by queries on the gold Delta Table, offering real-time, queryable analytics within the Databricks environment.

Monthly Revenue Trend Chart:

A trend chart was used to visualize the monthly revenue generated over the reporting period. This helped in identifying peak seasons, sales dips, and revenue growth patterns.

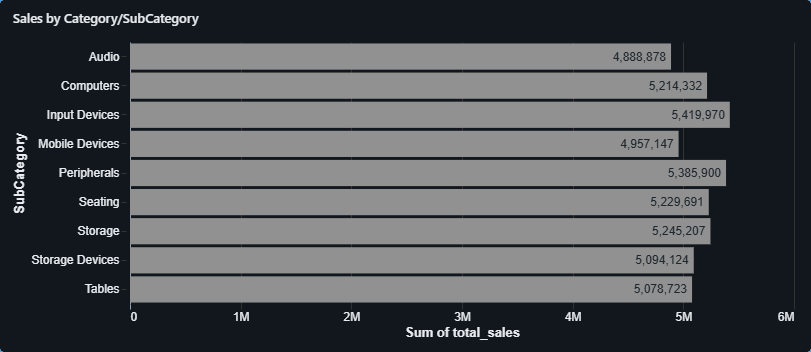


Top selling Products:

A chart displayed the top 10 products based on total units sold. This insight supported product demand analysis and inventory planning.  


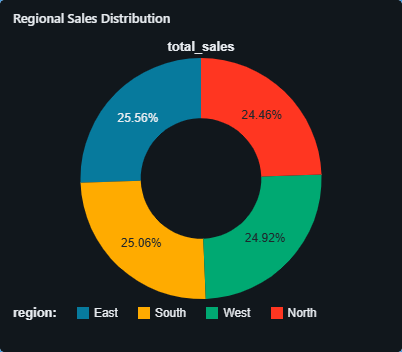
Category-Wise Sales Breakdown:

This chart showed sales distribution across product categories and sub-categories. It helped evaluate which categories contributed most to overall revenue and highlighted underperforming segments.



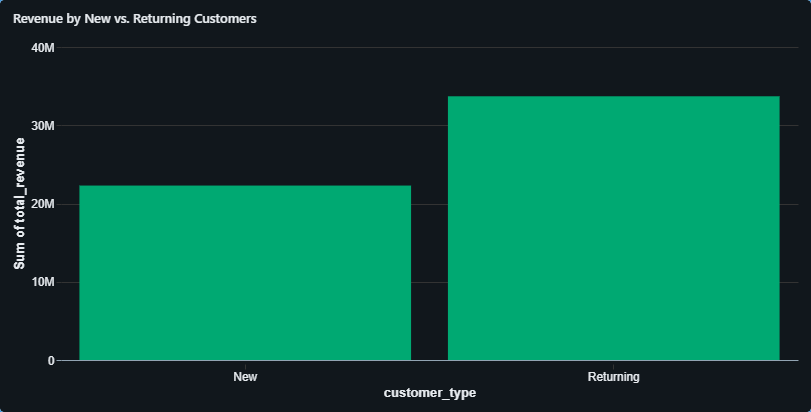
Region-Wise Sales Distribution:

Regional revenue contribution was visualized using a pie chart or geographical map. This helped identify high-performing zones and areas needing business attention.



Churn Rate based on New vs Returning Customer Revenue:

This chart compared revenue contributions from new versus returning customers. It provided a quick snapshot of customer retention effectiveness and acquisition success.



Discount Impact Analysis:

This chart analyzes the correlation between price and quantity sold across different price buckets. It helps assess whether price adjustments (such as discounts) are positively or negatively influencing customer purchasing behavior.



**Key Findings**

**Consistent Monthly Revenue with Sharp Drop in June 2025**The "Sales Trend Over Time" line chart indicates strong and relatively stable revenue between July 2024 and May 2025, consistently staying above ₹4M. However, June 2025 shows a sudden drop below ₹2M, suggesting either incomplete data or a decline in orders.

**High Demand for Tech Accessories and Office Items**The "Top Selling Products" bar chart shows that items like Desks, Headphones, Keyboards, and USB Drives recorded the highest sales volumes. This suggests a high demand for tech-related and ergonomic office products.

**Top Customers Contribute Significantly to Revenue**The "Top Customers" chart shows several customers (e.g., C9957, C5150) spending above ₹35K individually. This implies that customer loyalty and high-value clients play a crucial role in overall revenue generation.

**Strong Regional Sales in South and East Zones**From the "Product Performance by Region" matrix, most products, especially Bookshelves, Headphones, Keyboards, and Desks, performed well in the South and East zones, suggesting better reach, marketing, or delivery efficiency in these areas.

**Returning Customers Drive Majority Revenue**The "Revenue by New vs. Returning Customers" chart shows that Returning customers contributed nearly 60% of total revenue. This highlights strong customer retention and reinforces the value of loyalty programs.

**Balanced Sales Across Sub-Categories**The "Sales by Category/SubCategory" chart reveals that sub-categories like Input Devices, Storage Devices, and Computers have nearly equal contributions (around ₹5M each), indicating balanced performance across product types.

**Fluctuating AOV with Minor Seasonal Peaks**The "Average Order Value (AOV)" chart shows variability between ₹2600–₹2900, with noticeable peaks in November 2024 and May 2025, likely due to high-ticket purchases during promotional seasons.

**Mixed Correlation Between Price Buckets and Quantity Sold**The "Discount Impact Analysis" bar chart shows positive correlation in some price buckets (e.g., ₹30–₹40) and negative in others. This implies that discounts are effective only in selective pricing tiers and must be carefully targeted.

**Conclusion**

This project successfully demonstrated how a structured data pipeline from raw ingestion to interactive dashboards can uncover valuable business insights using Databricks and Apache Spark. Through progressive transformations across Bronze, Silver, and Gold layers, sales data was cleaned, standardized, and enriched for analytical use.

Key metrics like monthly revenue, top-selling products, regional performance, and customer segmentation were extracted using Spark SQL, enabling meaningful and actionable insights. Visual dashboards further enhanced stakeholder communication, allowing both technical and non-technical audiences to interpret trends with ease.

One of the most important conclusions drawn was the impact of customer retention, as returning customers contributed a majority of the revenue. Regional performance also varied significantly, offering direction for location-based strategy. Meanwhile, discount analysis revealed that price sensitivity isn't uniform, encouraging more targeted pricing strategies in future campaigns.

Overall, the project not only delivered on its technical objectives but also added business value by making the data exploration interactive, scalable, and stakeholder-friendly. The use of Delta Lake features like time travel and ACID compliance ensured reliable data handling throughout.

This project reinforces how well-structured data pipelines can translate complex datasets into insights that drive strategic decision-making.

**Future Scope**

This project successfully demonstrated the use of Databricks and Apache Spark for analyzing sales data. However, there is potential for several enhancements that can increase both the scale and impact of the solution in future phases:

### 1. **Real-Time Data Processing**

The current approach is based on static files. Future work could involve integrating real-time data pipelines using tools such as Apache Kafka or cloud-based stream services. This would allow for up-to-the-minute insights on sales performance and customer behavior.

### 2. **Machine Learning Integration**

Once the data pipeline is stable, predictive models can be introduced to forecast trends like sales volume, customer churn, or product demand. Databricks offers built-in support for ML lifecycle management through MLflow, which can simplify development and deployment.

### 3. **Automated Data Pipelines**

Manual data cleaning and transformations can be automated using Databricks Workflows or Delta Live Tables. This would reduce manual effort, minimize errors, and make the solution production-ready for continuous updates.

### 4. **Advanced Visualizations and BI Tools**

Although Databricks dashboards provide useful insights, connecting the data to external BI platforms like Power BI, Tableau, or Google Data Studio can offer richer visualizations, better user interactivity, and support for larger audiences.

### 

### 

### 

### 5. **Incorporating Additional Data Sources**

Adding new data streams, such as marketing performance, customer feedback, or competitor pricing can broaden the analysis and support better business decision-making through a more complete picture.

### 6. **Data Governance and Access Control**

To prepare for enterprise deployment, access control mechanisms can be implemented using Databricks Unity Catalog or workspace-level permissions. This ensures sensitive data is shared only with the appropriate stakeholders.