

International Journal Research Publication Analysis

Page: 01-09

RETAIL SALES AND CUSTOMER ANALYTICS USING PYTHON, SQL, POWER BI, AND WEB-BASED VISUALIZATION

^{*1}M. Gopi, ²Dr. M. Ramchander, ³Dr. B. Indira

¹Department of Master of Computer Applications, Chaitanya Bharathi Institute of Technology, Hyderabad, India.

²Assistant Professor, Department of MCA.

³Head of the Department: MCA, Ph.D., Associate Professor, Department of MCA.

Article Received: 05 September 2026

Article Revised: 25 September 2026

Published on: 15 October 2026

***Corresponding Author: M. Gopi**

Department of Master of Computer Applications, Chaitanya Bharathi Institute of Technology, Hyderabad, India.

DOI: <https://doi-doi.org/101555/ijrpa.8435>

ABSTRACT

Retail organizations generate large volumes of transactional data, but raw records are not directly useful for business decision-making without systematic cleaning, aggregation, and visualization. This paper presents an end-to-end retail sales and customer analytics workflow implemented using Python, SQL, Microsoft Power BI, and a lightweight web reporting layer. The study uses a synthetic practice dataset containing 1,000 retail orders recorded during 2025, 310 unique customers, 20 products, 5 categories, and 8 cities. Python with Pandas is used for duplicate removal, missing-value handling, derived time features, KPI calculation, aggregation, and exploratory visualization. SQL queries provide reusable business analysis for sales, profit, categories, cities, products, customers, payment methods, monthly performance, and profit margins. Power BI is specified as the interactive business-intelligence layer, while a Flask/HTML dashboard provides a lightweight web interface with filtering and runtime dataset support. The analysis reports total sales of ₹11,354,750.00, total profit of ₹3,659,302.06, 1,000 orders, 310 customers, 1,847 units, and an average order value of ₹11,354.75. Electronics contributes the largest category sales, and Laptop is the highest-selling product by sales value. The paper demonstrates how multiple analytics tools can be combined into a reproducible workflow for descriptive retail intelligence.

KEYWORDS: Retail Analytics, Sales Analysis, Customer Analytics, Python, Pandas, Sql, Power Bi, Business Intelligence, Data Visualization.

I. INTRODUCTION

Retail transaction systems capture order dates, customer identifiers, products, quantities, prices, discounts, payment methods, locations, and customer segments. Converting these records into actionable information requires data-quality checks, aggregation, visualization, and interpretation. Customer segmentation and transaction-based analysis are established approaches for understanding purchasing behavior and identifying meaningful groups within retail data [1]–[4].

This work develops a compact, reproducible analytics pipeline for retail sales and customer analysis. The contribution is not a predictive model; instead, it is an integrated descriptive analytics workflow that connects data preparation, relational querying, dashboard design, and web presentation.

The objectives are to: 1) validate and clean retail transaction records; 2) calculate core sales and customer KPIs; 3) analyze sales across time, category, city, and product; 4) provide SQL queries for repeatable business questions; 5) define a Power BI dashboard structure; and 6) expose the analysis through a lightweight web application.

II. RELATED WORK

Transaction-based customer analytics commonly uses Recency, Frequency, and Monetary (RFM) variables to segment customers and characterize purchasing behavior [1]–[4]. Recent retail studies have combined RFM with clustering and machine learning for more detailed segmentation and prediction [1], [4]. These approaches motivate future extensions of the present descriptive workflow, but the current study intentionally remains focused on transparent KPI and aggregation-based analysis.

Business-intelligence dashboards are used to convert transactional information into visual summaries for decision support. Recent work has examined Power BI for retail business analytics and decision-making [5]. An IEEE study on supermarket sales visualization also demonstrates the role of visual analytics in communicating sales patterns to decision makers [6].

The present project combines these themes at a practical implementation level: Pandas for data preparation, SQL for structured business questions, Power BI for interactive reporting, and Flask/HTML for lightweight web presentation.

III. DATASET AND SYSTEM OVERVIEW

The supplied dataset is a synthetic practice dataset created for learning and portfolio

purposes. It contains 1,000 orders from 2025 and 14 columns: Order_ID, Order_Date, Customer_ID, Customer_Name, City, Category, Product, Unit_Price, Quantity, Discount, Sales, Profit, Payment_Method, and Customer_Segment.

Because the dataset is synthetic, the numerical findings in this paper should be interpreted as experimental project results rather than real company performance.

Table I: Dataset Profile.

Property	Value
Orders	1,000
Unique customers	310
Products	20
Categories	5
Cities	8
Date range	2025
Total quantity	1,847



Fig. 1. End-to-end architecture of the proposed retail analytics workflow.

IV. METHODOLOGY

A. Data Preparation

The Python analysis reads the CSV using Pandas, removes duplicate rows, fills missing discount values with zero, derives missing sales from $\text{Unit_Price} \times \text{Quantity} \times (1 - \text{Discount})$, and fills missing profit values with zero. Order_Date is parsed as a date and month labels and month numbers are derived for time-based aggregation.

B. KPI Computation

For each transaction i , net sales are represented as:

$$\text{Sales}_i = \text{UnitPrice}_i \times \text{Quantity}_i \times (1 - \text{Discount}_i)$$

$$\text{Average Order Value} = \text{Total Sales} / \text{Distinct Orders}$$

$$\text{Profit Margin (\%)} = (\text{Total Profit} / \text{Total Sales}) \times 100$$

C. SQL Analysis

The SQL layer creates a retail_analytics database and sales table and provides queries for overall KPIs, category performance, city performance, top products, monthly sales, customer value, payment-method usage, and category-level profit margin. These queries provide a database-oriented equivalent of the Pandas group-by analysis.

D. Dashboard and Web Layer

The Power BI design uses KPI cards for sales, profit, orders, customers, and average order

value, together with time, category, city, product, and customer-segment visuals. The Flask application serves the HTML dashboard and exposes API endpoints for datasets, filters, KPI summaries, monthly sales, category sales, city sales, and top products. Uploaded CSV files are validated against required columns before being added to the runtime dataset.

V. IMPLEMENTATION

The implementation is organized into four complementary layers. Python generates analytical summaries and chart exports. SQL supports relational analysis. Power BI is specified as an interactive reporting environment. The web layer provides browser-based presentation and runtime dataset management.

Table II Implementation Layers.

Layer	Technology	Main Function
Data analysis	Python + Pandas	Cleaning, KPI calculation, aggregation
Database analysis	MySQL 8+	Business queries and grouped analysis
BI reporting	Power BI	Interactive dashboard and slicers
Web reporting	Flask + HTML/CSS/JS	Browser dashboard and API filtering

Pandas is the primary DataFrame processing layer. Pandas was developed to provide data structures and building blocks for practical statistical computing and data analysis [7]. The project therefore uses standard tabular operations such as grouping, aggregation, sorting, and CSV export.

VI. RESULTS

A. Overall KPIs

Table IV Category Performance.

Category	Sales	Profit	Units
Electronics	₹8,002,650	₹2,589,123	370
Home & Kitchen	₹1,551,960	₹494,635	407
Clothing	₹719,370	₹232,601	358
Sports	₹647,440	₹206,307	364
Beauty	₹433,330	₹136,636	348

Electronics is the dominant category with ₹8,002,650.00 in sales, above Home & Kitchen at ₹1,551,960.00. The concentration is driven largely by high-value electronics products such as laptops and smartphones.

D. City Performance

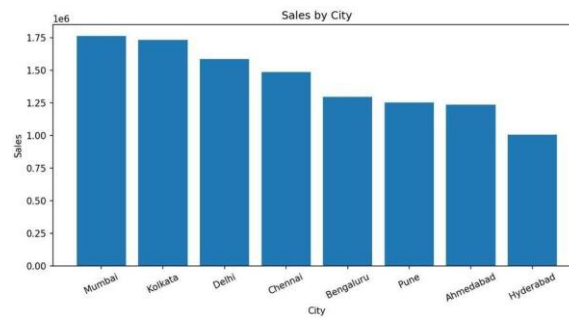


Fig. 3. Sales by product category.

Table III: Overall Key Performance Indicators.

Metric	Result
Total Sales	₹11,354,750.00
Total Profit	₹3,659,302.06
Total Orders	1,000
Total Customers	310
Total Quantity	1,847
Average Order Value	₹11,354.75
Profit Margin	32.23%

B. Monthly Sales

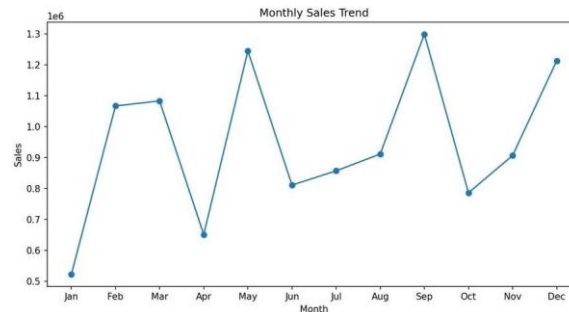


Fig. 2. Monthly sales trend for 2025.

September records the highest monthly sales at ₹1,299,347.50, followed by May at ₹1,245,080.00 and December at ₹1,213,492.50. January records the lowest monthly sales at ₹522,362.50. This variation shows why annual totals should be complemented by time-based analysis.

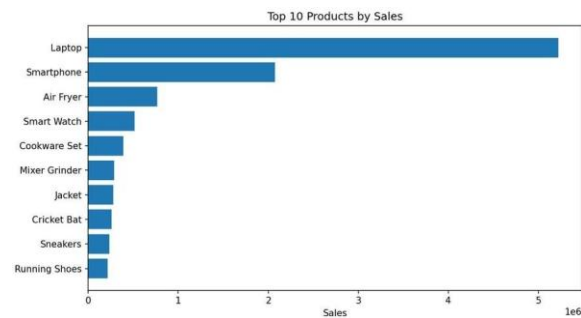
c. Category Performance

Table V City Performance.

City	Sales	Orders
Mumbai	₹1,760,628	124
Kolkata	₹1,733,638	142
Delhi	₹1,587,108	110
Chennai	₹1,484,130	132
Bengaluru	₹1,294,860	120
Pune	₹1,251,565	121
Ahmedabad	₹1,235,068	137
Hyderabad	₹1,007,755	114

Mumbai records the highest city-level sales at ₹1,760,627.50, followed by Kolkata at ₹1,733,637.50 and Delhi at ₹1,587,107.50. Hyderabad records ₹1,007,755.00.

E. Product Performance

**Fig. 4. Sales by city.****Table VI: Top 10 Products by Sales.**

Product	Sales	Units
Laptop	₹5,222,250	106
Smartphone	₹2,073,400	83
Air Fryer	₹768,625	133
Smart Watch	₹513,000	95
Cookware Set	₹390,425	125
Mixer Grinder	₹290,430	78
Jacket	₹279,680	98
Cricket Bat	₹257,320	104
Sneakers	₹235,480	95
Running Shoes	₹216,480	74

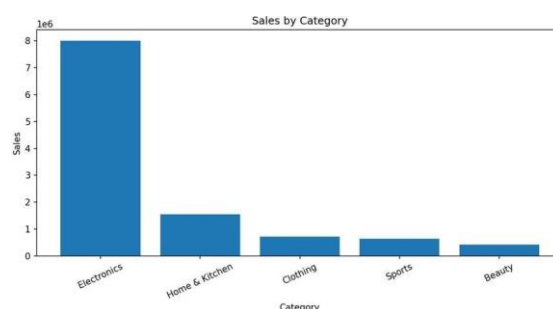


Fig. 5. Top 10 products by sales.

Laptop is the highest-selling product with ₹5,222,250.00 in sales, • Connect Power BI directly to MySQL for automated refresh. followed by Smartphone with ₹2,073,400.00. These results • Add role-specific dashboards for management, marketing, and illustrate the value of product-level analysis in retail reporting.

VII. CUSTOMER AND BUSINESS ANALYSIS

The dataset contains three customer segments: New, Returning, and Loyal. The segment field can be used in Power BI and the web dashboard to compare revenue and order behavior across customer groups. The SQL layer also provides a customer-value query that ranks customers by distinct orders and total sales.

Payment-method analysis can identify transaction volume by payment type. Discount values can be compared against sales and profit in the dashboard. These dimensions support descriptive monitoring, although the current project does not claim causal relationships between discounts and sales.

VIII. DISCUSSION

The results demonstrate the value of combining several analytical layers around a single cleaned dataset. Python provides exploratory analysis and chart generation; SQL provides a database-oriented query interface; Power BI provides interactive reporting; and Flask/HTML provides a lightweight browser interface. Maintaining the same KPI definitions across layers reduces the risk of disconnected calculations.

The project also distinguishes descriptive analytics from predictive analytics. The current system summarizes the supplied transaction data and does not forecast future sales, estimate customer lifetime value, or infer causal effects. Research on RFM and machine-learning approaches provides suitable directions for future extensions [1]–[4].

IX. LIMITATIONS

- The dataset is synthetic and intended for learning and portfolio use; results are not evidence of real retailer performance.
- The analysis is descriptive and does not include forecasting or predictive modeling.
- Customer segment labels are supplied in the dataset and are not independently learned from behavior.
- The workflow does not implement RFM scoring, clustering, recommendation models, or customer lifetime value estimation.
- The Power BI file is not embedded in the source archive; the project provides a dashboard construction guide and DAX measures.
- The web application is designed for local project use and can be extended for production deployment.

X. CONCLUSION

This paper presented an end-to-end retail sales and customer analytics workflow using Python, SQL, Power BI, and a Flask/HTML reporting layer. Using 1,000 synthetic retail orders, the system calculated ₹11,354,750.00 in sales and ₹3,659,302.06 in profit, with 310 customers and an average order value of ₹11,354.75. The analysis identified strong variation across months, categories, cities, and products, with Electronics and Laptop contributing the largest sales values in their respective dimensions. The principal contribution is an integrated and reproducible descriptive analytics workflow rather than a new predictive algorithm. Future work can incorporate RFM scoring, clustering, forecasting, customer lifetime value estimation, automated refresh, and direct SQL-to-Power-BI connectivity.

XI. FUTURE WORK

- Implement RFM-based customer segmentation and compare it with the supplied customer-segment labels.
- Apply clustering methods such as K-Means to discover behavior-based customer groups.
- Develop sales forecasting models using monthly or daily transaction history.
- Add customer lifetime value and churn-risk estimation. operations.

ACKNOWLEDGMENT

The author thanks the Department of MCA, Chaitanya Bharathi Institute of Technology, Hyderabad, for academic guidance and support. The author also acknowledges

Dr.M.Ramchander, Assistant Professor, Department of MCA, and Dr. B. Indira, Associate Professor and Head of the Department of MCA, for their academic guidance.

REFERENCES

1. A. Joy Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, “RFM ranking – An effective approach to customer segmentation,” *Journal of King Saud University–Computer and Information Sciences*, vol. 33, no. 10, pp. 1251–1257, 2021, doi: 10.1016/j.jksuci.2018.09.004.
2. T. Ho, S. Nguyen, H. Nguyen, N. Nguyen, D. Man, and T. Le, “An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry,” *Business Systems Research*, vol. 14, no. 1, pp. 26–53, 2023, doi: 10.2478/bsrj-2023-0002.
3. M. Abdul Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, “RFM-based repurchase behavior for customer classification and segmentation,” *Journal of Retailing and Consumer Services*, vol. 61, Art. no. 102566, 2021, doi: 10.1016/j.jretconser.2021.102566.
4. B. Rajeshwari, M. R. Doodipala, K. K. Varma, P. P. Ram, and J. R. N., “Retail Industry Analytics: Unraveling Consumer Behavior through RFM Segmentation and Machine Learning,” in *Proc. 2024 IEEE Int. Conf. on Electro Information Technology (eIT)*, 2024, doi: 10.1109/eIT60633.2024.10609927.
5. H. Dang Quoc, “Applying Power BI for improved retail business analytics and decision-making,” *Applied Computer Science*, vol. 21, no. 2, pp. 154–163, 2025, doi: 10.35784/acs_7130.
6. “Application of Tableau in Visual Analysis Data of a US Supermarket Sales,” in *Proc. 2021 Int. Conf. on Signal Processing and Machine Learning*, 2021, doi: 10.1109/CONF-SPML54095.2021.00044.
7. W. McKinney, “Data Structures for Statistical Computing in Python,” in *Proc. 9th Python in Science Conf.*, pp. 56–61, 2010, doi: 10.25080/Majora-92bf1922-00a.
8. Microsoft, “Power BI documentation,” Microsoft Learn, 2026.
9. MySQL, “MySQL 8.0 Reference Manual,” Oracle, 2026.
10. Pandas Development Team, “pandas documentation,” 2026.