

# Comprehensive Data Analytics Project Report (Phases 1–6)

## **E-Commerce Data Analysis Report**

### **Phase 1: Data Collection & Understanding**

- The project began with the acquisition of a transactional e-commerce dataset containing 499 rows and 16 initial columns. This data included User\_ID, Product\_ID, Product Category, Price, Discount, Final Price, Payment Method, and Purchase\_Date. The primary objective in Phase 1 was to understand the business domain—an online retail environment focusing on user purchase behavior.
- Key work done in this phase included: identifying data sources, validating schema definitions, performing data profiling, noting missing values, detecting anomalies, and documenting assumptions such as currency usage (INR), date formatting irregularities, and duplicate transaction risks.
- The output of Phase 1 was a detailed requirements understanding and a data dictionary describing every column, its purpose, and its expected analytical role in subsequent phases

## Phase 2: Data Cleaning & Transformation

This phase involved extensive preprocessing. All null rows, inconsistent values, and duplicate entries were handled. Price columns were converted into numerical format. Discount and Final Price fields were validated to ensure logical consistency.

Work completed includes:

- Removing unwanted columns (e.g., automatically generated unnamed columns).
- Converting Purchase\_Date into a datetime datatype.
- Extracting Year, Month, and Day from Purchase\_Date.
- Creating DayOfWeek for weekly pattern analysis.
- Label Encoding categorical columns such as Category and Payment Method.
- Validating transformed data and exporting the clean dataset as clean\_data.csv.

All cleaned data was saved and used for visualization, ML modeling, and dashboard creation in later phases.

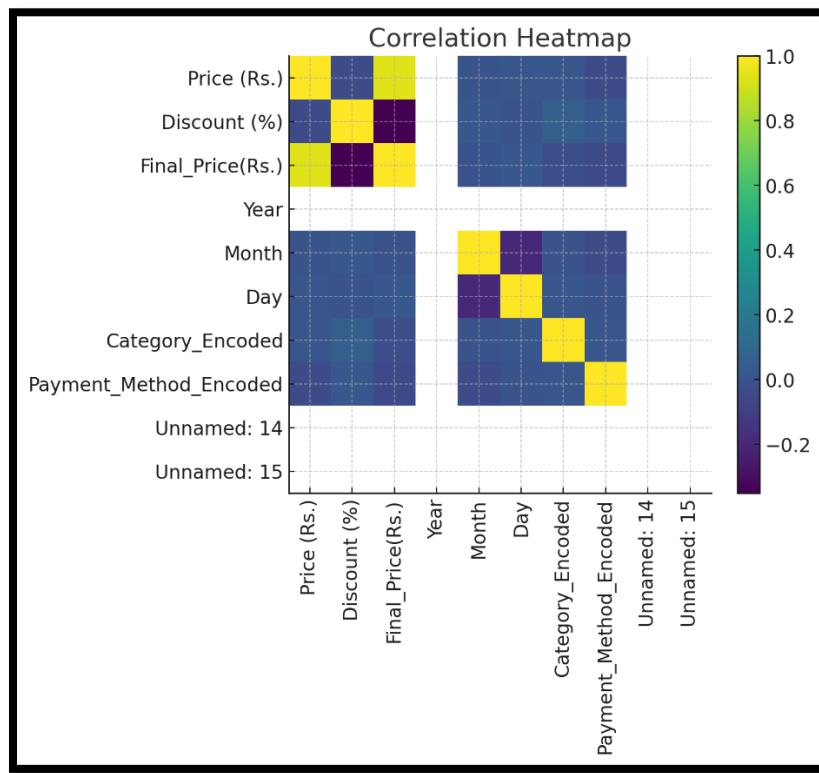
## Phase 3: Statistical Analysis

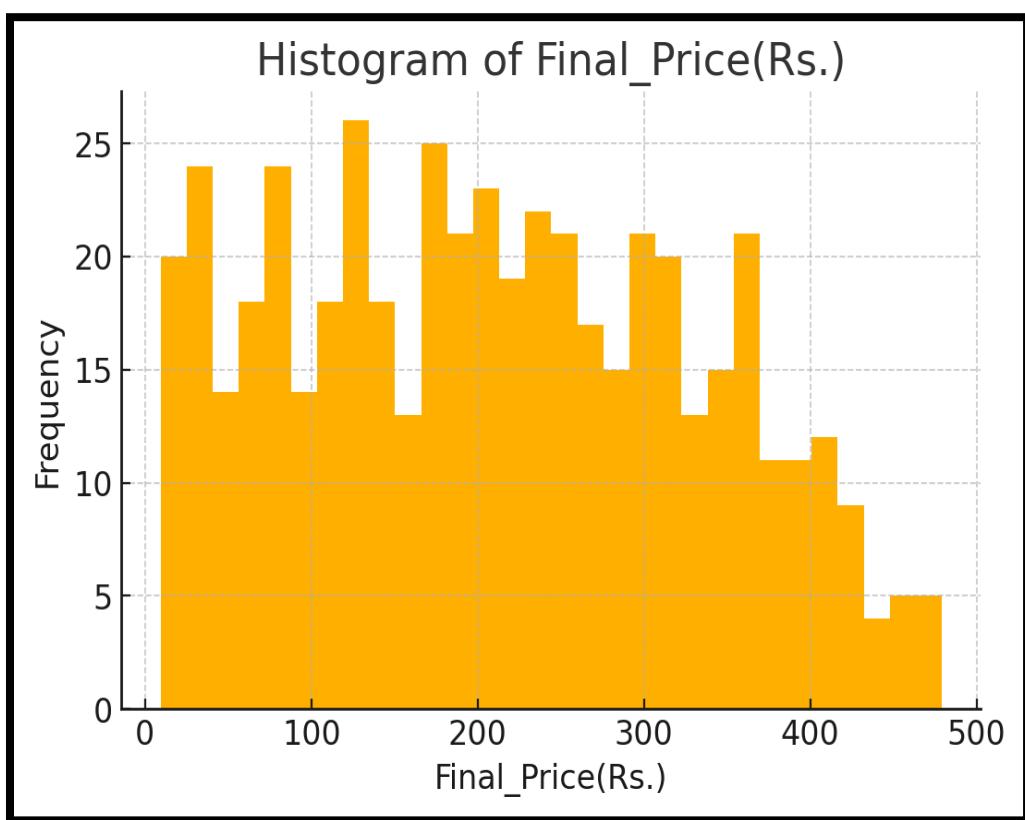
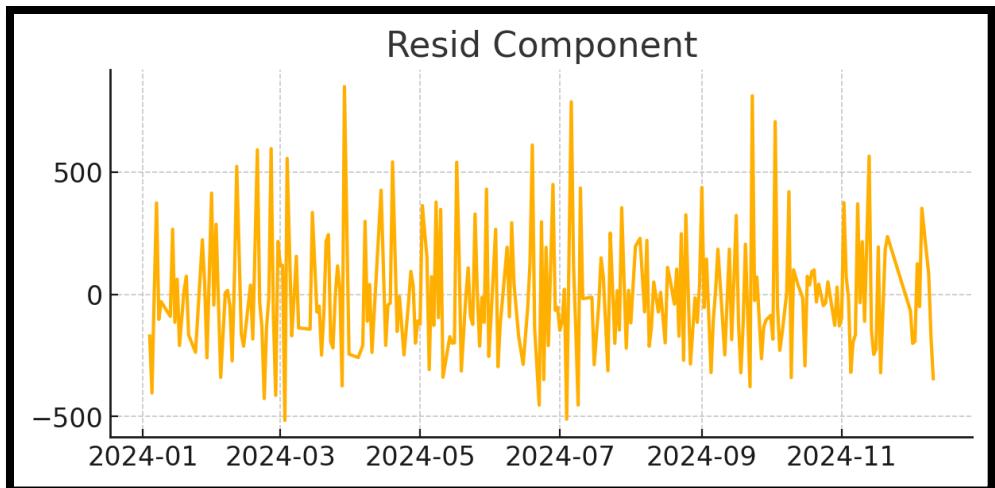
In Phase 3, foundational statistical techniques were applied to uncover descriptive insights and relationships among variables.

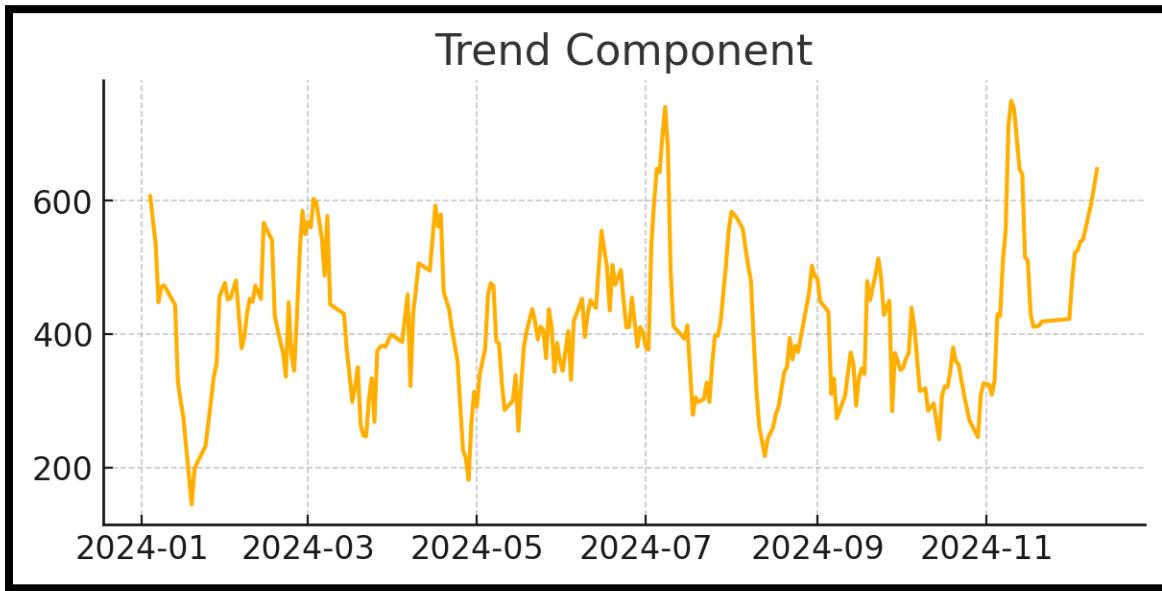
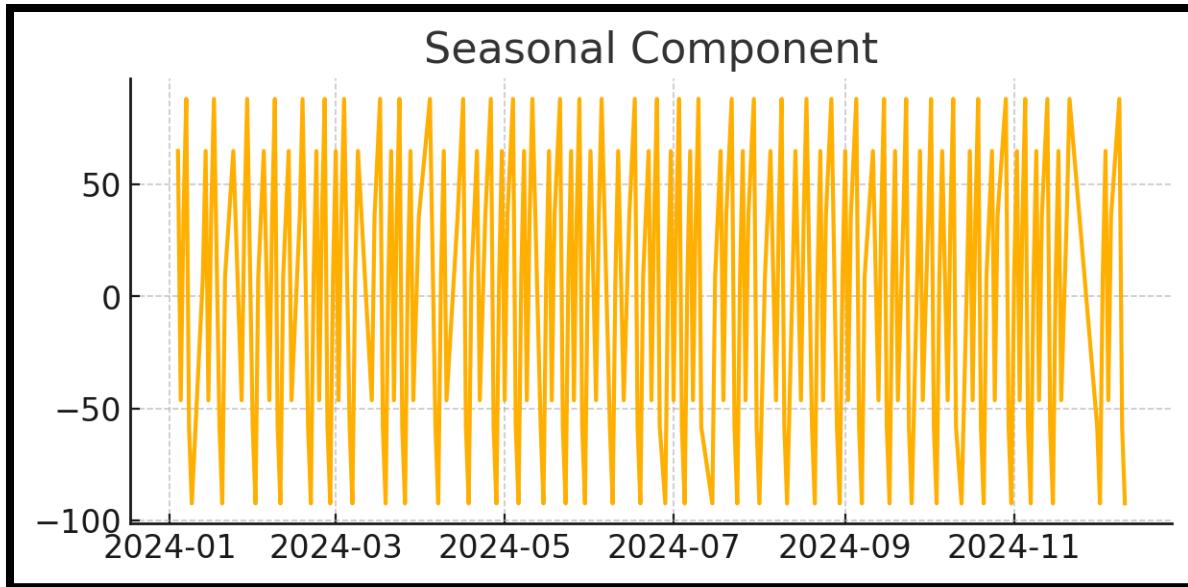
### Work completed includes:

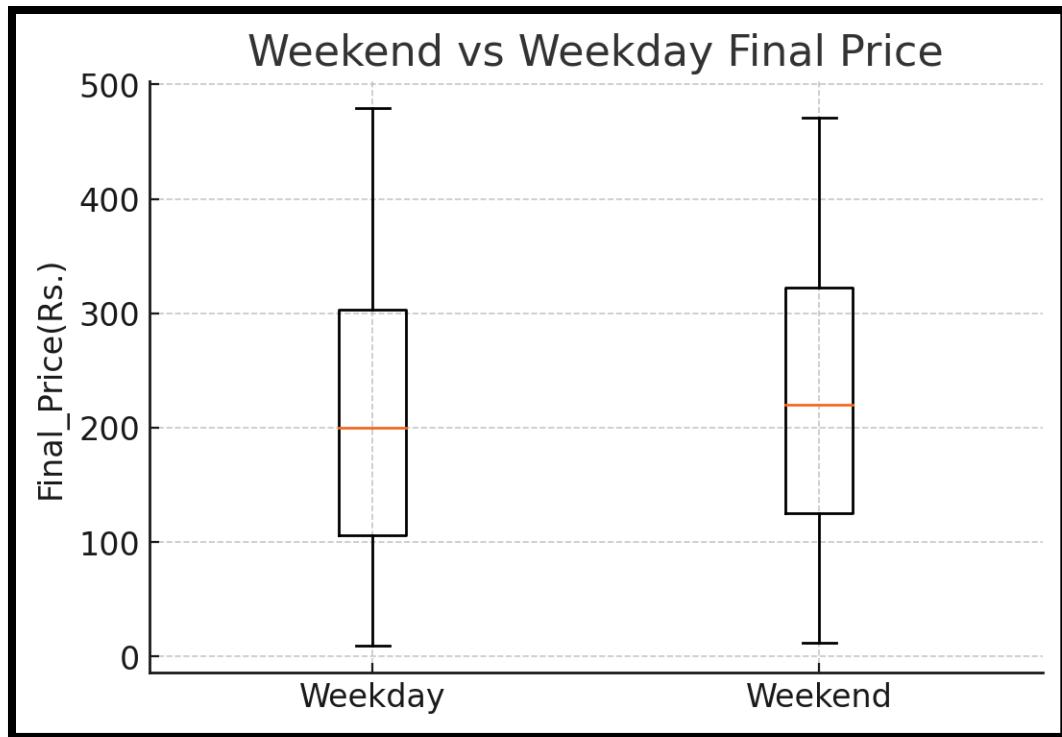
- Descriptive statistics: mean, median, standard deviation for prices and discounts.
  - Distribution analysis for understanding skewness and variability.
  - Correlation matrix to determine relationships between Price, Discount, Final Price, temporal features.
  - Hypothesis testing:
    - Weekend vs Weekday sales using T-test.
    - Category impact on sales via ANOVA.
    - Country relevance (if applicable) using Chi-square.

Outputs included visuals and statistical summaries used in the final reporting.







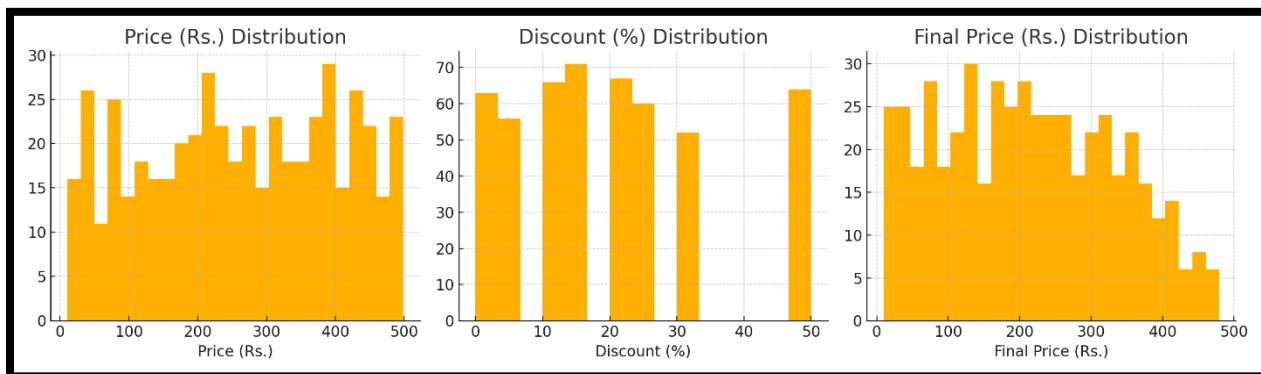
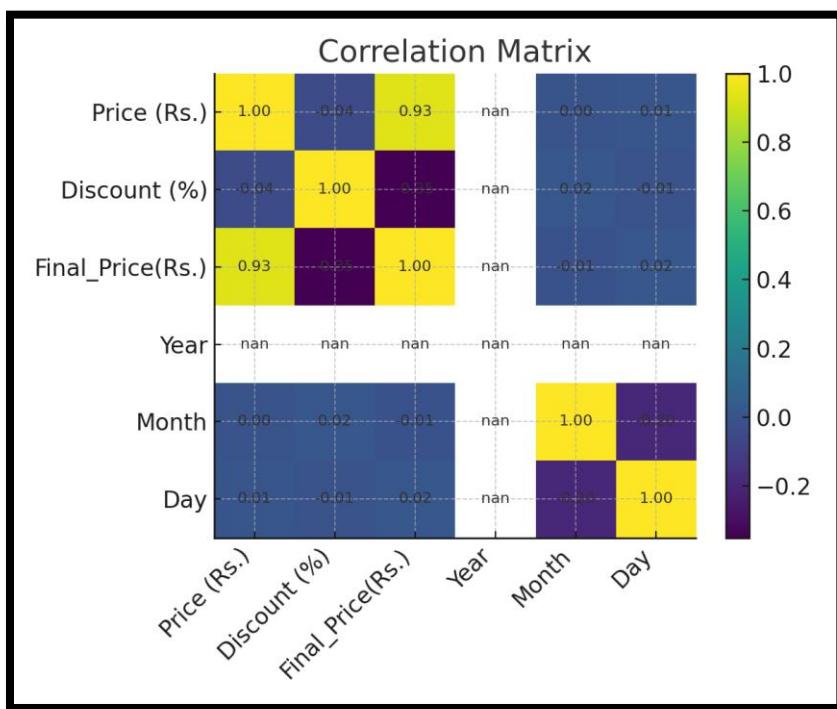


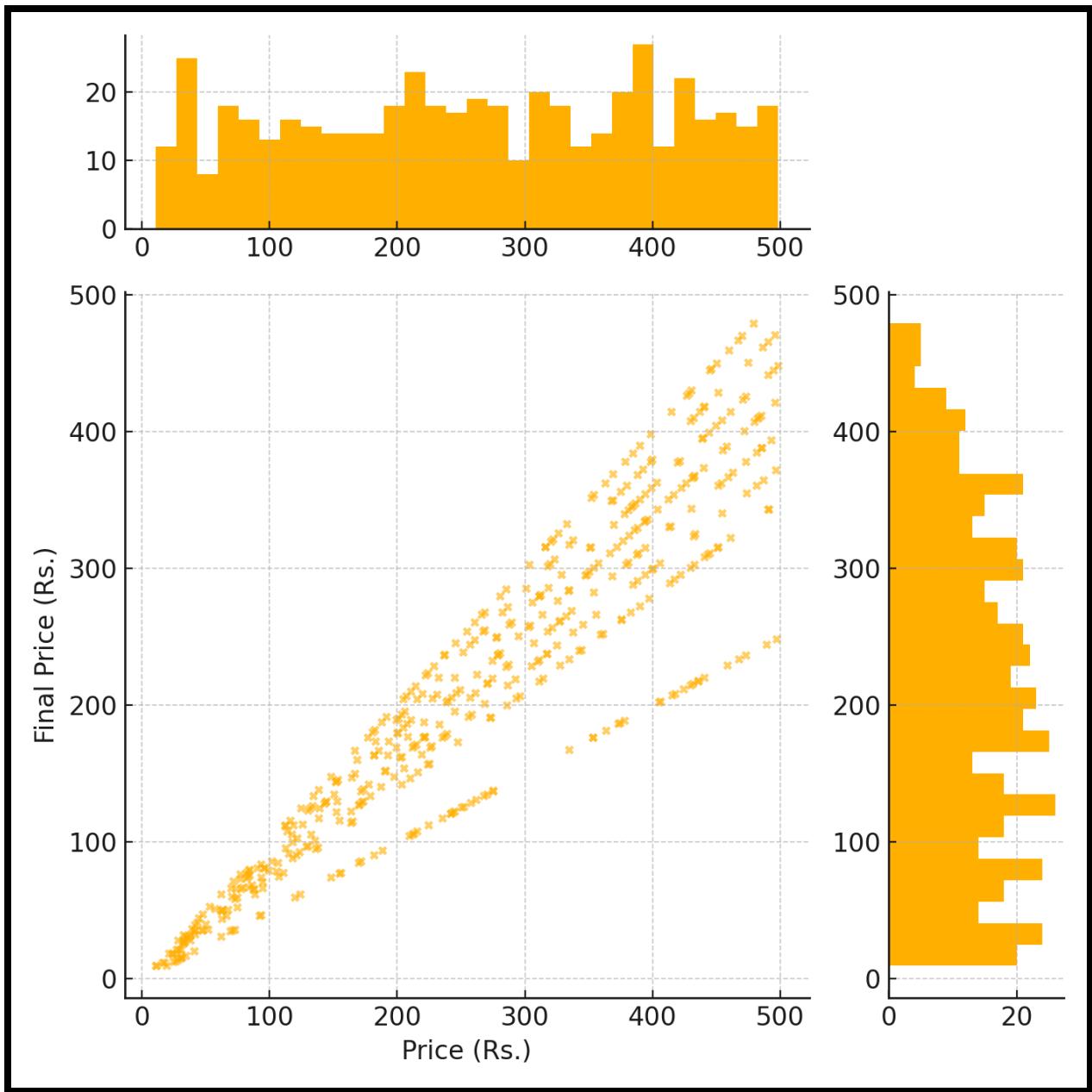
## Phase 4: Data Visualization

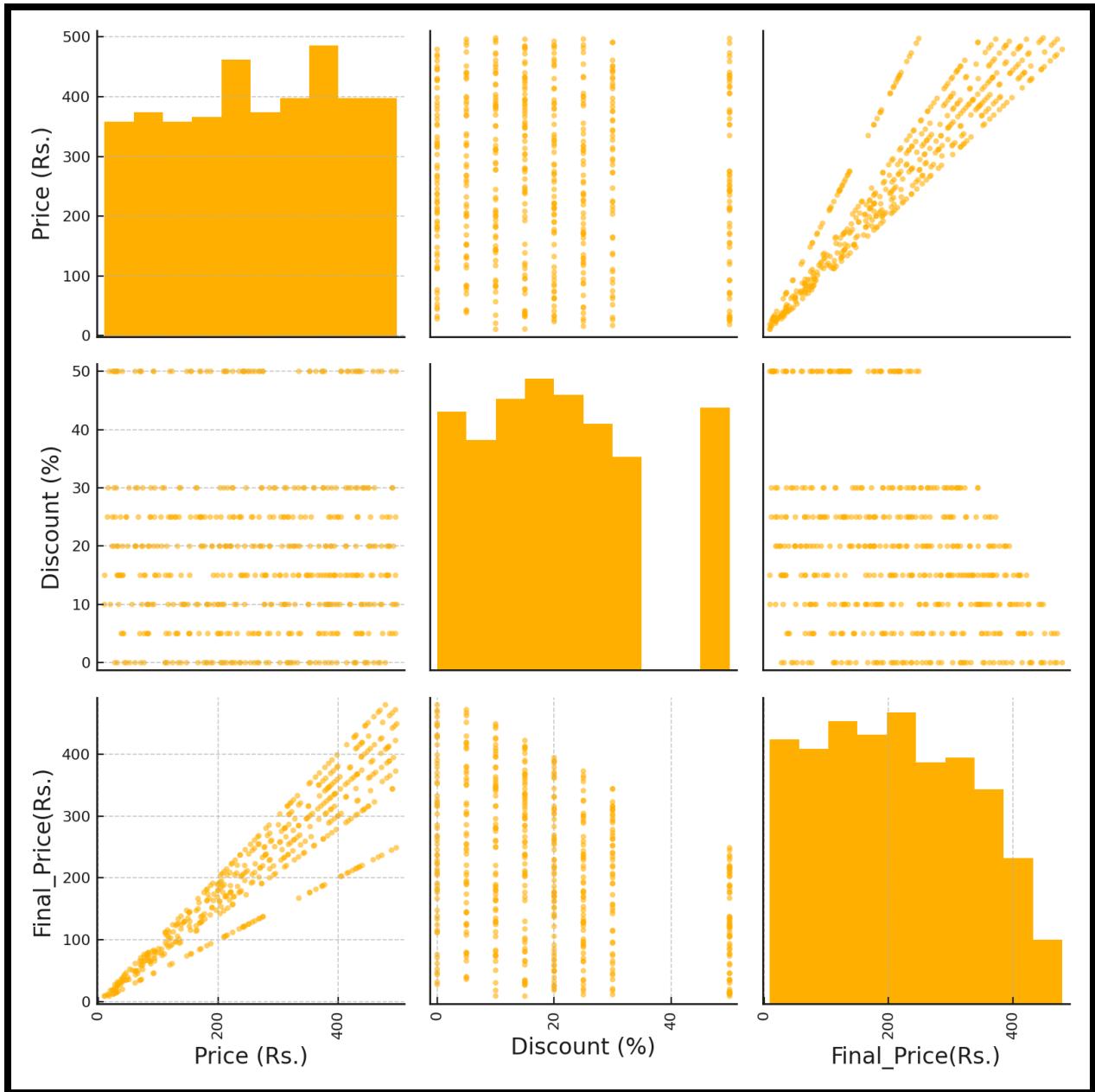
This phase produced multiple Python-based Matplotlib visualizations and business-ready dashboards.

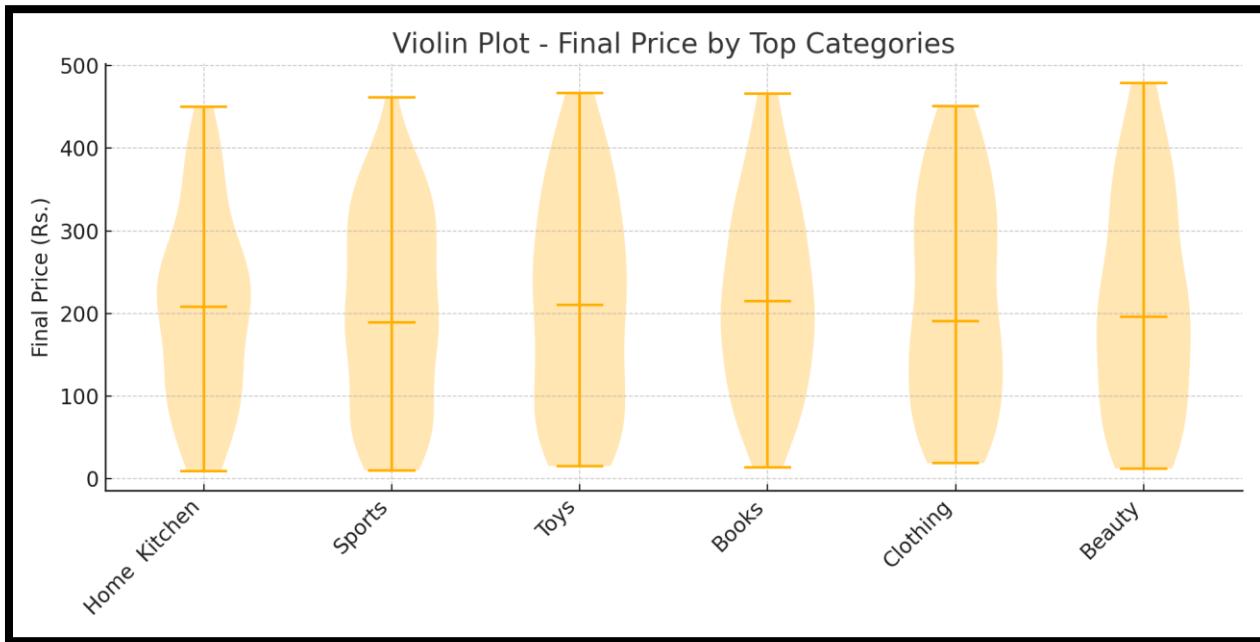
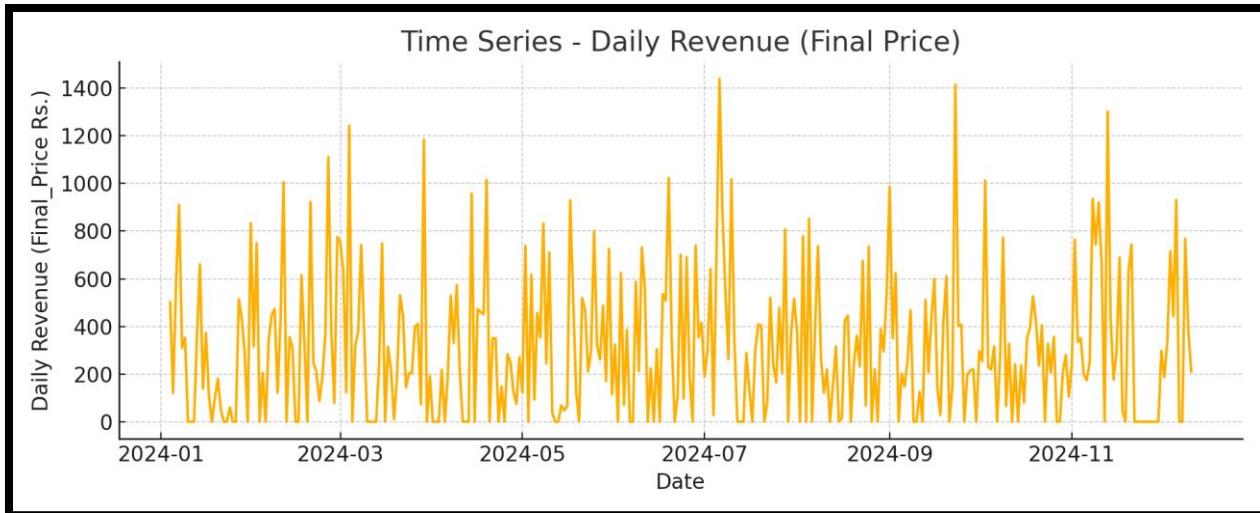
i) Python visualizations created:

- Time-series plot for daily revenue trends.
- Distribution plots for Price, Discount, and Final Price.
- Correlation heatmap with annotated values.
- Pairwise scatter matrix to observe multivariate relationships.
- Violin plot to compare pricing distributions across top categories.
- Joint plot for analyzing Price–Final Price interaction.









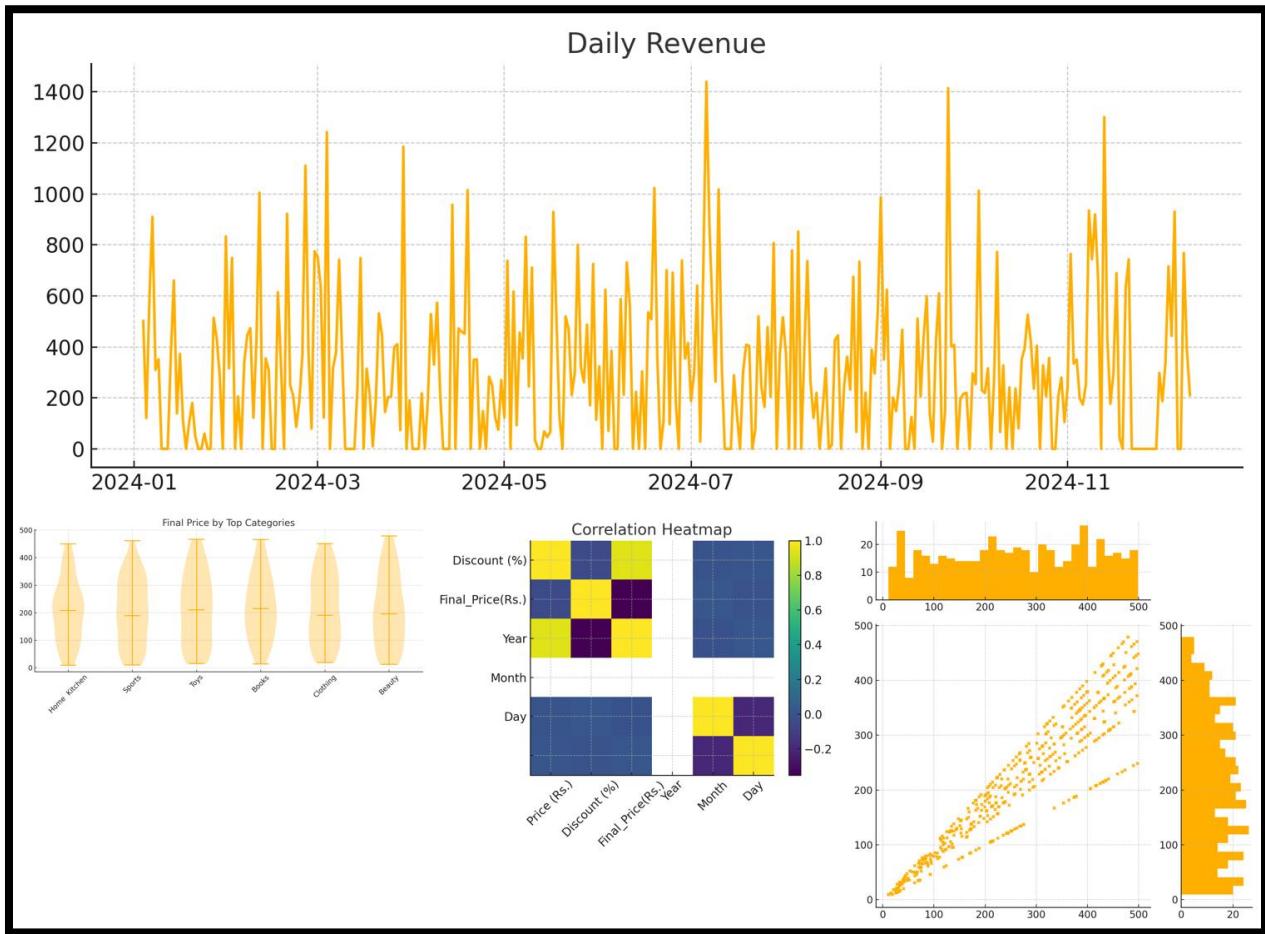
ii) Power BI Dashboard included:

- KPI Cards: Total Revenue, Total Orders, AOV, Discounts.
- Time-series revenue line chart.
- Category-wise sales bar chart.
- Payment method donut chart.
- Interactive slicers: Category, Payment Method, Year, Month, DayOfWeek.



### iii) Tableau Mock Dashboard:

Due to the absence of Tableau Desktop/Public, a stitched Tableau-styled dashboard image was generated reflecting a typical layout with top-level time-series visualization and multiple supporting analytics charts.



## Phase 5: Advanced Analytics & Machine Learning

This phase applied predictive and segmentation models to uncover deeper business insights.

Work completed includes:

- K-Means Clustering (RFM Analysis): Recency, Frequency, and Monetary value were computed per user to create segments based on spending and engagement behavior. Clusters were visualized and interpreted.
- Linear Regression: Built to predict Final Price based on Price, Discount, Category, and Payment Method. Evaluation metrics (RMSE, R<sup>2</sup>) were computed to measure predictive strength.
- Logistic Regression: A churn label was constructed based on recency > 90 days. The model predicted customer churn likelihood with accuracy, precision, and recall metrics.
- Decision Tree Classifier: Built to predict product categories based on purchase behavior. Feature importances were extracted to understand key drivers.

.pkl model files were generated for reproducibility and all outputs were compiled in a Jupyter Notebook.

## Phase 6: Excel Analysis & Final Deliverables

Phase 6 focused on advanced Excel analytics and final project packaging.

Excel work completed:

- Pivot Tables for aggregating sales by category, payment method, and date.
- VLOOKUP and INDEX-MATCH for dynamic lookup operations.
- SUMIFS for multi-criteria filtering.
- Dynamic Excel Dashboard with slicers and conditional formatting.
- VBA script for automating repetitive tasks like refreshing pivot tables.
- What-If Analysis: Goal Seek for reaching revenue targets and Scenario Manager for comparing pricing/discount strategies.

Final Deliverables:

- Detailed project report.
- PowerPoint presentation summarizing insights .
- Complete package containing datasets, Python scripts, ML models, dashboards, and Excel files.