Exploratory Data Analysis – Findings Report

Dataset: Supermarket Sales
Prepared by: Sakshi

1. Dataset Overview

- The dataset contains **no missing values**, so no cleaning or imputation was required.
- It includes both **numerical** (e.g., Unit Price, Quantity, Total, cogs, gross income, Rating) and **categorical** (e.g., Branch, Product line, Gender, Payment) variables.

2. Distribution Insights

- Monetary variables like **Total, cogs, gross income** show **right skewness**, meaning a small number of high-value purchases pull the distribution upward.
- Outliers exist in these numerical columns, but they appear to be **valid business outliers**, not errors (high-value bills).

3. Categorical Insights

- Some product lines generate more transactions than others.
- Different branches have different sales volumes, showing variation in store performance.
- Popular payment modes dominate revenue, indicating customer preference patterns.

4. Correlation Analysis

- No strong correlations (above 0.5) were found between numerical columns.
- Rating does not strongly correlate with revenue-based variables, meaning customer rating doesn't directly predict purchase amount.

5. Business Insights (Groupby Findings)

- Some **product lines contribute significantly more revenue** than others.
- **Branch-level revenue** varies, indicating opportunities to improve performance in weaker branches.
- **Payment methods** show clear popularity differences — certain modes are used far more frequently.

6. Outliers & Skewness

- Numerical columns show moderate skewness.
- Log transformation is recommended *only* if building predictive models later (not needed for this assignment).

7. Key Conclusions

- Dataset is clean and ready for analysis.

- Revenue is uneven across product lines and branches.
- No major correlation patterns — variables are mostly independent.
- Customer rating does not significantly influence spending.
- Some variables are skewed but acceptable for descriptive analysis.