Data Cleaning and Exploratory Data Analysis Report

Submitted by- Vardayinee Sandoo

Enrollment number- 23116105

Branch- ECE (EC4)

1. Introduction

This report presents an analysis of the Grocery Inventory and Sales Dataset. The objective is to clean the dataset, perform exploratory data analysis (EDA), and derive insights from the data. The steps include handling missing values, removing duplicates, identifying and treating outliers, and conducting univariate, bivariate, and multivariate analyses.

2. Data Cleaning

2.1 Loading and Inspecting Data

- The dataset was loaded using Pandas and its structure was inspected using .info() and .describe().

2.2 Handling Missing Values

- Missing numerical values were imputed with the mean of their respective columns.
- Categorical missing values were replaced using mode imputation.

2.3 Removing Duplicates

- Duplicate records were identified and removed to ensure data integrity.

2.4 Detecting and Treating Outliers

- The Z-score method was applied to numerical columns, removing data points beyond three standard deviations.
- This reduced the dataset size, improving data reliability.

2.5 Standardizing Categorical Values

- Categorical string values were standardized to lowercase and stripped of extra spaces.

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

- Summary statistics (mean, median, mode) were computed for numerical variables.
- Histograms and box plots were used to visualize distributions and identify skewness.
- Frequency distributions were analyzed for categorical variables.

3.2 Bivariate Analysis

- A correlation matrix and heatmap were generated to explore relationships among numerical features.
- Scatter plots were used to visualize relationships between numerical variables.
- Box plots and violin plots compared distributions across categorical variables.

3.3 Multivariate Analysis

- Pair plots were created to analyze relationships between multiple numerical variables.
- Heatmaps were used to explore deeper correlations and patterns.
- Grouped comparisons helped identify the combined effects of multiple variables.

4. Key Findings and Inferences

4.1 Data Cleaning Insights

- The dataset initially contained missing values, which were successfully imputed. Numerical missing values were replaced with mean values, while categorical values were replaced with mode, ensuring minimal data loss.
- Duplicate entries were found and removed, eliminating redundant data that could distort analysis.
- The presence of outliers was addressed using the Z-score method. This step helped in making statistical summaries more robust and prevented extreme values from skewing results.
- Standardizing categorical values improved consistency, reducing variations caused by formatting inconsistencies (e.g., different capitalizations of the same category).

4.2 Univariate Analysis Findings

- The distribution of sales figures showed positive skewness, indicating that most items have low to moderate sales, while a few have very high sales.
- Inventory levels were widely spread, with some products being overstocked while others had minimal stock.
- Certain categorical variables (such as product categories and supplier names) had unbalanced distributions, suggesting that a few categories dominated the dataset.

4.3 Bivariate Analysis Findings

- A strong positive correlation was observed between inventory levels and sales. This suggests that well-stocked items tend to have higher sales, which aligns with expectations.
- Price and sales showed a moderate negative correlation, indicating that lower-priced items tend to have higher sales volumes, while expensive items may sell less frequently.
- Some product categories exhibited significantly different sales patterns, as shown in box plots and violin plots, highlighting the influence of product type on sales performance.

4.4 Multivariate Analysis Findings

- Pair plots indicated that sales, inventory, and pricing interact in complex ways, where certain product segments show strong clustering in high or low sales zones.
- Heatmaps confirmed that some variables (such as supplier reliability and inventory restocking rates) had significant correlations with sales trends.
- Grouped comparisons revealed that seasonal factors played a role in sales trends, with specific product categories showing spikes in certain time frames, suggesting the need for seasonal inventory adjustments.

5. Conclusion

The analysis successfully cleaned the dataset and extracted meaningful insights through EDA. Key findings revealed important relationships between inventory levels, pricing, and sales. Addressing outliers and standardizing data improved the reliability of statistical analyses. Future steps could include predictive modeling to further analyze sales trends and optimize inventory levels.