# Data Wrangling Report

## Introduction

This report provides an overview of the data wrangling process applied to the supermarket sales dataset. The objective is to clean and transform the dataset to ensure it is suitable for analysis and visualization.

## Data Loading and Overview

- The dataset was loaded using pandas.read_csv().
- The first few rows were examined using df.head().
- Dataset structure was analyzed using df.info().
- Summary statistics were generated using df.describe().

## Handling Missing Values

- Missing values were identified using df.isnull().sum().
- A visualization of missing data was created using the missingno library.
- Missing values were handled by removing rows with null values using df.dropna(inplace=True).

## Handling Duplicates

- Duplicate rows were checked using df.duplicated().sum().
- A total of 6 duplicate rows were found.
- These duplicate rows were removed to ensure data integrity.

## Data Transformation

- Categorical variables were converted to the category dtype for optimized storage and processing.

## Exploratory Data Analysis (EDA)

- Distribution of Numerical Features
    - Histograms were plotted for numerical columns to observe distributions.
- Sales Trends Over Time
    - A time-series line plot was created to analyze daily sales trends.

- Customer Demographics Analysis
    - Gender distribution was visualized using a bar chart.
    - Gender-based purchasing trends were analyzed:
    - Females tend to buy more Accessories, Fashion, Sports, and Travel products.
    - Males dominate purchases in Electronics, Food, Travel, Health & Beauty, and Home & Lifestyle categories.
- Data Cleaning and Final Dataset
    - The dataset was cleaned, with missing values handled, duplicate rows removed, and data types optimized.
    - The final cleaned dataset was saved as cleaned_supermarket_sales.csv.

# Conclusion

- The dataset underwent extensive cleaning, including handling missing values, removing duplicates, and optimizing data types.
- Key insights into gender-based purchasing behavior were identified.
- The cleaned dataset is now ready for further analysis and predictive modeling.