Data Wrangling Report

1. Introduction

This report summarizes the data wrangling process applied to the supermarket sales dataset. It includes data exploration, cleaning, and transformation steps to prepare the data for analysis.

2. Data Loading

- The dataset was loaded using pandas from a CSV file.
- Initial inspection was performed using .head(), .info(), and .describe() to understand the structure of the data.
- The dataset contains **1000 rows** and **17 columns**, including attributes like Invoice ID, Branch, City, Customer type, Gender, Product line, Unit price, Quantity, Tax 5%, Total, Date, Time, Payment method, COGS, Gross margin percentage, Gross income, and Rating.

3. Data Exploration

- Checked for missing values using .isnull().sum(), finding **no missing values** in the dataset.
- Identified duplicate rows using .duplicated().sum(), confirming **no duplicates**.
- Examined data types of each column to ensure consistency.

4. Data Cleaning

- **Handling Missing Values:** No missing values were found in this dataset.
- **Removing Duplicates:** Since no duplicates were detected, no further action was needed.
- **Data Type Conversion:**
  - The Date column was converted to datetime format using pd.to_datetime(df['Date']).
  - The Time column was converted to a time format.

    - Numeric columns such as Unit price, Quantity, Tax 5%, Total, COGS, and Gross income were verified to have appropriate numeric data types.

- **Standardizing Categorical Variables:**
    - The Customer type column was cleaned to ensure it contained only Member and Normal values.
    - The Payment method was checked for consistency across all entries.

5. Data Transformation

- **One-Hot Encoding:**
    - Categorical variables such as Branch, City, Customer type, Gender, Product line, and Payment method were converted into numerical format using one-hot encoding.
    - This process ensured that categorical variables could be used in machine learning models and statistical analysis.

6. Summary

The dataset was successfully cleaned and transformed, ensuring it was ready for further analysis. The key steps included:

- Confirming the dataset contained **1000 transactions** with **no missing or duplicate values**.
- Converting date and time fields for better analysis.
- Encoding categorical variables to facilitate data modeling.
- Saving the cleaned dataset for visualization and business insights extraction.

The cleaned dataset is now structured and formatted appropriately for further business analysis, including identifying sales trends, customer behavior, and performance insights.