Data Wrangling Report

Overview of the Data:

The dataset contains supermarket sales records, including branch information, product categories, prices, payment methods, and customer ratings.
Our goal during the wrangling process was to clean, organize, and prepare the data to ensure it is reliable and ready for deep analysis.

---

1. What was the goal of Our data wrangling process?

• Our goal was to deliver a clean and consistent dataset to support accurate analysis and meaningful insights.

2. What challenges did we face during the wrangling process?

• we encountered several issues, such as missing values, duplicate records, inconsistent data types, and the presence of outliers.
• Each challenge required careful strategies to resolve without losing important information.

3. How did we handle missing values?

• we first checked for missing values using isnull().sum().
• we decided to remove rows with missing entries using dropna() after evaluating their impact on the overall dataset.
• Finally, we verified that no missing values remained after cleaning.

4. How did we manage duplicated data?

• we detected duplicate rows using duplicated().sum() and removed them with drop_duplicates().
• After removal, we double-checked to ensure that the dataset was completely free of duplicates.

5. How did we deal with inconsistent data types?

• we converted the 'Date' column into datetime format to enable time-based analysis.
• we also transformed categorical features like 'Payment Method' into numerical codes using .astype('category').cat.codes.

6. How did we analyze the data distribution and detect outliers?

• During the cleaning phase, we analyzed the values using descriptive statistics like describe() to identify potential outliers numerically. we also reviewed maximum and minimum values manually, but we chose not to remove or modify any outliers to preserve important data.

7. Did we standardize or adjust the data distribution to ensure accuracy?

• In this cleaning process, we did not perform standardization or distribution adjustment. The main goal was to clean missing values, correct data types, and prepare the raw data for more advanced analysis later.

8. How did we verify data quality after cleaning?

• we re-checked for missing values, duplicates, and outliers.
• we visualized some key columns before and after cleaning to confirm improvements.
• Finally, we ran basic descriptive statistics to ensure that the data distribution looked logical and clean.

9. What additional enhancements did we apply during wrangling?

- We renamed columns to more readable and understandable names.
- Some columns contained unclear or meaningless values, so we replaced them with more meaningful values.
- We removed unnecessary columns like links, images, and other irrelevant fields.
- We adjusted the date columns properly to ensure correct handling for time-based analysis.
- We added new columns that could enhance and support deeper analysis later on.

10. What tools and libraries did we rely on during this process?

• we mainly used Pandas for data manipulation and Seaborn/Matplotlib for data visualization during the wrangling phase.