**1. Data Loading and Exploration**

• **Data Loading**: You started by reading in the sales data using pandas.read\_csv().

• **Initial Exploration**: You explored the dataset using methods like head(), dtypes, and describe() to understand its structure and data types.

• **Column Exploration**: Specifically, you explored the sales\_method column, checking for unique values and ensuring consistency by converting the strings to lowercase and standardizing variations (e.g., “em + call” to “email + call”).

**2. Data Preprocessing**

• **Sales Method Standardization**:

• You standardized the values in the sales\_method column by converting them to lowercase and replacing any variations (such as “em + call” to “email + call”) to ensure consistency.

• You also converted the sales\_method column to a categorical data type for better memory management and ease of analysis.

**3. Outlier Handling**

• **Outlier Detection**:

• Using a boxplot, you visually inspected the distribution of revenue for each sales method and checked for outliers.

• You then used the number\_of\_outliers function to quantify the number of outliers in the revenue column for each sales method.

• **Outlier Removal**:

• After identifying outliers, you used the remove\_outliers function (imported from an external module) to remove outliers for each sales method (call, email, and email + call).

• After cleaning the data, you concatenated the three cleaned subsets (sales\_call\_no\_outliers, sales\_email\_no\_outliers, and sales\_email\_call\_no\_outliers) back together into a single DataFrame (sales\_no\_outliers).

• **Verification**:

• You verified the number of rows removed by comparing the length of the original DataFrame with the cleaned one (sales\_no\_outliers).

**4. Missing Value Handling**

• **Missing Values Analysis**:

• You checked for missing values in the revenue column and summarized the missingness by sales method.

• You created a missing\_df DataFrame to track the percentage of missing revenue for each sales method and calculated the missing value percentage.

• **Handling Missing Values for** call **and** email:

• For the call and email sales methods (which had relatively low missingness), you decided to **drop** the rows with missing revenue. This was done using the dropna() method to ensure that you weren’t distorting the results with imputation.

• **Imputation for** email + call:

• For the email + call sales method (which had a higher rate of missing values), you decided to **impute** the missing values using the **median revenue** for each value of nb\_sold. This was done because revenue showed a strong correlation with nb\_sold, and using the median allowed for filling in the missing values without introducing bias.

• You calculated the median revenue per nb\_sold for the email + call method and used this value to fill in the missing revenue entries for that sales method.

• **Concatenation**:

• After handling missing values (either by dropping or imputing), you concatenated the cleaned subsets (missing\_call\_clean, missing\_email\_clean, and missing\_email\_call) back into the final cleaned DataFrame (sales\_final).