Data Science Assignment

Data Cleaning and Data Wrangling

1. You are working with a dataset of employee salaries. Some salary values are missing, and the experience column contains values in different formats (e.g., "5 years", "Six Years", "7 YRS"). You need to:

   a. Fill the missing salary values with the median salary.

   b. Standardize the experience column so that all values are numerical.

   Input:

   data = {'Employee': ['Alice', 'Bob', 'Charlie', 'David', 'Eve'],

   'Salary': [50000, np.nan, 55000, np.nan, 60000],

   'Experience': ['5 years', 'Six Years', '7 YRS', '3 years', 'Ten Years']}

2. You have a dataset containing customer reviews where the **"Feedback"** column contains text comments like "Great Service!", "Very Poor Support!", etc. You also have a **"Spending"** column. Your tasks are:

   a. Create a new column **"Sentiment"** by classifying reviews that contain the word "Poor" as "Negative" and others as "Positive".

   b. Bin the Spending column into three categories: "Low", "Medium", "High".

   Input:

   data = {'Customer': ['John', 'Sarah', 'Mike', 'Lisa', 'Tom'],

   'Feedback': ['Great Service!', 'Very Poor Support!', 'Excellent Product!', 'Poor Experience', 'Loved it!'],

   'Spending': [150, 700, 1200, 500, 300]}

3. Given the two datasets:
    a. A **sales dataset** containing sales representatives and their respective sales.
    b. A **target dataset** with the target sales assigned to each representative.

Write python code to:

i. Merge both datasets based on the representative name.

ii. Detect and remove outliers in the sales column using the Interquartile Range (IQR) method.

Input:

```
sales_data = {'Rep': ['Alice', 'Bob', 'Charlie', 'David', 'Eve'],
              'Sales': [50000, 75000, 1200000, 65000, 90000]}
target_data = {'Rep': ['Alice', 'Bob', 'Charlie', 'David', 'Eve'],
               'Target': [55000, 80000, 95000, 70000, 85000]}
```