### **Lab 1: Basic Data Understanding**

### **Questions:**

#### **1. Data Understanding:**

**Task:** Print the data types of each column and use descriptive statistics to understand the data.

**Questions:**

* Identify and justify the appropriateness of the data types for each attribute. Suggest changes if necessary.
* What does the statistical summary tell you about potential issues with data quality, such as range problems, missing values, or format inconsistencies?
* If there is a typo issue, fix it.

#### 

#### 

#### **2. Identifying and Handling Missing Values:**

**Task:** Identify missing values in the dataset and propose methods to handle them.

**Questions:**

* What patterns of missing data did you observe in the dataset?
* How will you handle the missing values? Justify your approach.

#### **3. Detecting and Correcting Invalid Entries:**

**Task:** Identify and correct invalid entries in the dataset (e.g., negative values in columns where only positive values are appropriate, unrealistic dates, or other logical inconsistencies). Also, If there is a typo issue, fix it.

**Questions:**

* What invalid entries did you find in the dataset? Provide examples.
* Explain the steps you took to correct these invalid entries, and justify your methods.

#### **4. Addressing Duplicate Records:**

**Task:** Identify and remove duplicate records from the dataset.

**Questions:**

* How did you identify duplicate records in the dataset?
* What criteria did you use to decide which duplicates to remove, if any? Justify your approach.

#### **5. Data Range Issues:**

**Task:** Identify and address any data range issues (e.g., values outside expected ranges, negative sizes, or dates in the future).

**Questions:**

* What range issues did you find in the dataset? Provide specific examples.
* How did you address these range issues? Explain and justify your approach.

#### 

#### **6. Format Inconsistencies:**

**Task:** Identify and correct any format inconsistencies in the dataset (e.g., inconsistent date formats, units, or text formats).

**Questions:**

* What format inconsistencies did you find in the dataset? Provide examples.
* Explain how you standardized these formats and why it is important to do so.

#### **7. Misclassified Data:**

**Task:** Detect and correct any misclassified data within the dataset (e.g., numeric data in text fields or vice versa).

**Questions:**

* What misclassified data did you identify in the dataset?
* How did you correct these misclassifications, and why did you choose these methods?

#### **8. Data Visualization:**

**Task:** Use visualizations to identify patterns, inconsistencies, or outliers in the dataset.

**Questions:**

* Which visualizations did you use to explore the data, and what insights did they provide?
* How did these visualizations help in identifying inconsistencies, outliers, or other data quality issues?