Task 1

1. What are missing values and how do you handle them?

- **Meaning:** Missing values are gaps in the dataset where no value is recorded for a variable.
- **Handling methods:**
    1. **Remove** rows/columns with too many missing values (using dropna() in Pandas).
    2. **Impute** missing values with:
        - Mean/median/mode for numerical columns.
        - Most frequent value or "Unknown" for categorical columns.
        - Interpolation for time series data.

---

2. How do you treat duplicate records?

- **Meaning:** Duplicates are repeated rows in the dataset.
- **Handling:**
    - Identify: df.duplicated()
    - Remove: df.drop_duplicates()
    - Keep only the first or last occurrence depending on business rules.

---

3. Difference between dropna() and fillna() in Pandas?

Function	Purpose	Example
dropna()	Removes rows/columns containing NaN values	df.dropna()
fillna()	Replaces NaN values with a given value or strategy	df.fillna(0)

---

4. What is outlier treatment and why is it important?

- **Meaning:** Outliers are values that deviate significantly from the rest of the data.
- **Importance:** They can skew mean, affect model performance, and give misleading results.
- **Treatment:**
    - Remove using statistical methods like **IQR** or **Z-score**.
    - Cap values to a threshold (Winsorization).
    - Transform data (e.g., log transformation).

---

5. Explain the process of standardizing data.

- **Meaning:** Standardization ensures data is in a consistent format, scale, and unit.
- **Steps:**
    1. Convert text to a consistent case (e.g., lowercase for city names).
    2. Use consistent naming for categories (e.g., "Male" vs "M").
    3. Standardize units (e.g., kg instead of grams, consistent date formats).
    4. Scale numeric features (e.g., z-score standardization) if needed for modeling.

---

6. How do you handle inconsistent data formats (e.g., date/time)?

- Convert all date/time columns to **datetime objects** using Pandas pd.to_datetime().
- Apply a consistent format (e.g., YYYY-MM-DD).
- Ensure timezone consistency.
- For strings, trim spaces and correct typos.

---

7. What are common data cleaning challenges?

- Missing or incomplete data.

- Duplicates and redundancy.
- Inconsistent formatting (dates, text case, units).
- Outliers and extreme values.
- Mixed data types in a single column.
- Human errors during data entry.

---

8. How can you check data quality?

- **Completeness:** Check for missing values.
- **Uniqueness:** Look for duplicate rows.
- **Validity:** Verify data matches expected formats/ranges.
- **Consistency:** Ensure similar data matches across sources.
- **Accuracy:** Compare with reliable references.