Task 1 - Data Cleaning and Preprocessing (Theory Answers)

Q1. What are missing values and how do you handle them?

- Missing values are blank or null entries in data.
- Handling methods:
  - Drop rows/columns with missing data (`dropna()`).
  - Fill with defaults (`fillna()`), e.g., mean/median/mode for numbers, 'Unknown' for text.
  - Advanced imputation using ML models (e.g., KNN, regression).

Q2. How do you treat duplicate records?

- Duplicates are repeated rows.
- Detect using `.duplicated()`.
- Remove using `.drop_duplicates()`.
- Sometimes keep the first/last copy depending on business rules.

Q3. Difference between dropna() and fillna() in Pandas?

- dropna(): removes rows/columns containing null values.
- fillna(): replaces null values with a given value (like mean, median, 0, or 'Unknown').

Q4. What is outlier treatment and why is it important?

- Outliers are extreme values that differ greatly from most data.
- Treatment methods:
  - Remove them if due to errors.
  - Cap/floor them using IQR or z-score.
  - Apply transformations (log, sqrt) to reduce effect.
- Important because outliers can skew averages, correlations, and ML models.

Q5. Explain the process of standardizing data.

- Making data consistent and uniform across the dataset.
- Examples:
  - Convert column names to lowercase with underscores.
  - Gender → 'Male/Female' instead of 'M/F/male/female'.
  - Dates into a single format (YYYY-MM-DD).
  - Units into the same scale (e.g., cm instead of cm/inches mixed).

Q6. How do you handle inconsistent data formats (e.g., date/time)?

- Use `pd.to_datetime()` to parse and unify formats.
- Decide on one standard format (e.g., ISO YYYY-MM-DD).
- For time zones, convert using `tz_localize` or `tz_convert`.

Q7. What are common data cleaning challenges?

- Missing data.
- Duplicate rows.
- Inconsistent formats (dates, text case, units).
- Outliers.
- Wrong data types.
- Human errors during entry.

Q8. How can you check data quality?

- Use `.info()` and `.describe()` in Pandas.
- Check null counts with `.isnull().sum()`.
- Look for duplicates.
- Validate ranges (e.g., Age > 0).
- Check for invalid categories (e.g., gender not in {Male, Female}).
- Use profiling libraries like `pandas-profiling` or `ydata-profiling`.