Task 1 – Interview Questions and Answers (Data Cleaning and Preprocessing)

---

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

Answer:

Missing values (NaN, null, or empty fields) are entries where no data is recorded. Handling them depends heavily on the reason for their absence:

- **Removal (Listwise Deletion):** Dropping rows or columns. This is suitable when data is **Missing Completely at Random (MCAR)** and the percentage of missingness is very small (e.g., under 5%).
- **Imputation:** Replacing missing values with estimates. Key techniques include:
    - **Statistical:** Using the **mean** (for normally distributed numeric data), **median** (for skewed numeric data), or **mode** (for categorical data).
    - **Model-Based:** Using advanced techniques like **K-Nearest Neighbors (KNN) Imputer** or regression models to predict the missing values based on other features.

2. How do you treat duplicate records?

Answer:

Duplicate records are observations that are identical across all, or a key subset of, features. They must be identified and removed to ensure the statistical integrity of the dataset and prevent models from being over-represented by certain instances.

- **Identification:** In Pandas, we use **df.duplicated()** to find them.
- **Treatment:** They are removed using **df.drop_duplicates()**. It's crucial to first define what constitutes a duplicate (e.g., matching across all columns vs. matching only on Customer_ID and Date).

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

Answer:

These are two fundamental functions for managing missing data in a DataFrame:

In short, dropna() **removes**, while fillna() **replaces**.

Feature	**df.dropna()**	**df.fillna(value)**
Function	**Deletes** rows or columns containing NaN.	**Substitutes** (imputes) NaN values with a specified value.
Impact on Size	Reduces the size of the dataset.	Preserves the original dataset size.
Primary Use	When data loss is acceptable and you need clean rows.	When data must be preserved and an estimation is needed.

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

Answer:
Outliers are extreme data points that lie far outside the expected range of observations.
Importance: Outliers can severely **distort summary statistics** (like the mean) and negatively impact the training of sensitive machine learning algorithms (e.g., Linear Regression, K-Means) by widening the error margin.
Common Methods:

- **Detection:** Using the **IQR (Inter-Quartile Range)** method or the **Z-score** method (values beyond ± 3 standard deviations).
- **Treatment: Capping/Winsorizing** (replacing outliers with the nearest non-outlier boundary value) or **transforming** the data (e.g., using a log transformation to normalize the distribution).

5. Explain the process of standardizing data.
Answer:
Standardization (or Z-Score Normalization) is a scaling technique that transforms numeric features so the resulting distribution has a mean (μ) of 0 and a standard deviation (σ) of 1.
The formula for the Z-score is:
$$Z = \frac{x - \mu}{\sigma}$$
This process is vital because it ensures that all features contribute equally to the model, preventing features with larger scales (e.g., Salary) from disproportionately dominating algorithms that rely on distance calculations, such as **K-Means** or **PCA**.

6. How do you handle inconsistent data formats (e.g., date/time)?
Answer:
Inconsistent data (e.g., mixed date formats like "01-05-2023" vs. "May 1, 2023" or varying text casing) must be converted to a single, common format.

- **Date/Time:** Use powerful functions like **pd.to_datetime()** in Pandas, explicitly specifying the format string (e.g., %m/%d/%Y).
- **Text:** Standardize categorical text features using string methods like **.str.lower()** to ensure the same entity ("USA" and "usa") is recognized as a single, consistent category.

7. What are common data cleaning challenges?
Answer:
Data cleaning regularly involves overcoming several key challenges:

1. **Missingness:** Incomplete or sparse data.
2. **Structural Errors:** Inconsistent formatting, typos, and non-standardized units or names.
3. **Anomalies:** Outliers and values that are statistically implausible.
4. **Redundancy:** Duplicate observations and irrelevant features (high-cardinality IDs).
5. **Type Mismatches:** Data stored with an incorrect type (e.g., numeric IDs stored as strings), hindering mathematical operations.

8. How can you check data quality?
Answer:
Data quality is verified through a process known as Data Profiling, which validates against five key dimensions:

1. **Completeness:** Are all essential values present? (Low missingness).
2. **Consistency:** Is the data uniform in format and structure across all records?
3. **Accuracy:** Does the data reflect the true value in the real world? (Requires domain knowledge and external checks).
4. **Uniqueness:** Are there no duplicate records?
5. **Validity:** Does the data adhere to predefined business rules (e.g., age must be positive, prices cannot be zero)?

Tools like **summary statistics** (df.describe()), **data type inspection**, and **visualizations** (histograms, box plots) are critical for auditing these dimensions.