Data Cleaning Interview Questions & Answers

1. How do you treat duplicate records?

Duplicate records are handled using drop_duplicates() in Pandas. You can drop full duplicate rows or specify subset columns. Example:
df.drop_duplicates(inplace=True)

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

dropna() removes missing values, while fillna() replaces them.
- dropna(): Removes rows with NaN values.
- fillna(): Fills NaNs with a specified value or method.
Example: df['col'].fillna(0)

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

Outlier treatment helps remove or reduce the impact of extreme values.
Methods include IQR, Z-score, or capping values. It improves model accuracy.

4. Explain the process of standardizing data.

Standardization converts data to a mean of 0 and std of 1. Useful for ML models.
Example using sklearn:
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df[['col']] = scaler.fit_transform(df[['col']])

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

Use pandas to parse and convert formats. Example:
df['date'] = pd.to_datetime(df['date'], errors='coerce')

6. What are common data cleaning challenges?

Missing data, duplicates, inconsistent formatting, incorrect types, outliers, and typos are common.

7. How can you check data quality?

Use:
- df.isnull().sum()
- df.duplicated().sum()
- df.info()
- df.describe()
- Value counts

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

Data Cleaning Interview Questions & Answers

Missing values are absent entries in a dataset.

Handled by:

- Removing (dropna)
- Filling (fillna, ffill, bfill)
- Imputing (mean, median)