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

Missing values are data points that are not recorded or are unknown. To handle them, we can either drop the rows or columns containing missing values, or impute them using mean, median, or mode for numerical data, and mode or a specific value for categorical data. We can also use interpolation or regression imputation for more complex cases.

2. How do you treat duplicate records?

To treat duplicate records, we first identify them using techniques like grouping or sorting. Then, we can either drop duplicates or keep the first or last occurrence, depending on the context. It's essential to analyze the data to determine the best approach.

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

dropna() removes rows or columns containing missing values, whereas fillna() replaces missing values with a specified value. dropna() is used when the missing values are not significant, while fillna() is used when we want to preserve the data and impute the missing values.

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

Outlier treatment involves identifying and handling data points that are significantly different from the rest of the data. Outliers can skew analysis and models, so it's essential to detect and treat them. We can use techniques like Winsorization, transformation, or removal to handle outliers.

5. Explain the process of standardizing data.

Standardizing data involves transforming it to have a common scale, usually with a mean of 0 and a standard deviation of 1. This is done to prevent features with large ranges from dominating the analysis. We can use techniques like Min-Max Scaler or Standard Scaler to standardize data.

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

To handle inconsistent data formats, we can use techniques like data normalization, where we convert data to a standard format. For date/time data, we can use libraries like Pandas to parse and standardize the format.

7. What are common data cleaning challenges?

Common data cleaning challenges include handling missing values, outliers, and inconsistent data formats. Other challenges include dealing with duplicate records, data entry errors, and data quality issues.

8. How can you check data quality?

To check data quality, we can perform exploratory data analysis (EDA), summary statistics, and data visualization. We can also use data validation techniques, like checking for missing values, data types, and ranges, to ensure the data is accurate and consistent