# Data Preprocessing - Questions and Answers

## 29. What is data preprocessing, and why is it important in data science?

Data preprocessing is the process of cleaning and transforming raw data into a usable format. It is important because real-world data is often incomplete, inconsistent, or noisy. Preprocessing improves data quality and helps machine learning models perform better.

## 30. Explain missing data imputation techniques.

Missing data imputation techniques include mean/median/mode substitution, forward or backward fill, interpolation, and model-based methods like k-NN or regression. The choice depends on the nature of the data and the reason for missingness.

## 31. What is one-hot encoding, and when is it used?

One-hot encoding is a method to convert categorical variables into a binary matrix. It is used when categorical variables are nominal (no inherent order), allowing machine learning algorithms to interpret them numerically.

## 32. How do you handle categorical data in machine learning?

Categorical data can be handled using techniques like label encoding, one-hot encoding, or embedding. The choice depends on the algorithm and whether the categories are ordinal or nominal.

## 33. Describe the process of data normalization and standardization.

Normalization scales data to a range of [0, 1], while standardization transforms data to have a mean of 0 and standard deviation of 1. These techniques help ensure that features contribute equally to model training.

## 34. What is feature scaling, and why is it necessary?

Feature scaling is the process of adjusting the range of features in a dataset. It is necessary because many machine learning algorithms are sensitive to the scale of input data, especially those based on distance metrics.

## 35. What is outlier detection, and how can you identify outliers in a dataset?

Outlier detection involves identifying data points that deviate significantly from other observations. Techniques include statistical methods (e.g., Z-score, IQR), visualization (e.g., box plots), and machine learning methods (e.g., isolation forest).