1. Why is exploratory data analysis significant?

Exploratory Data Analysis (EDA) is significant because it helps in understanding data, identifying patterns, detecting outliers, and guiding subsequent analysis and decision-making.

1. What is data wrangling?

Data wrangling involves cleaning, transforming, and enriching raw data to make it suitable for analysis or modeling.

1. How can you handle missing values?

Missing values can be handled by imputation methods such as mean, median, or mode imputation, or using advanced techniques like K-nearest neighbors (KNN) imputation or predictive models.

1. What is univariate analysis?

Univariate analysis involves the analysis of a single variable at a time to understand its distribution, summary statistics, and characteristics.

1. What is a spurious correlation?

A spurious correlation is a statistical relationship between two variables that is actually due to a third variable, rather than a direct causal relationship between the two variables.

1. Does causation imply correlation?

Yes, causation does not imply correlation. While a causal relationship implies a correlation between variables, correlation does not necessarily imply causation.

1. How do you treat outliers in a dataset?

Outliers in a dataset can be treated by removing them, transforming the data, or using robust statistical methods that are less sensitive to outliers.

1. What are the benefits of data cleaning?

The benefits of data cleaning include improved data quality, increased accuracy of analyses, reduced bias, and better reliability of results.

1. What is multivariate analysis?

Multivariate analysis involves the analysis of two or more variables simultaneously to understand their relationships, dependencies, and interactions.

1. What is an interquartile range?

The interquartile range (IQR) is a measure of statistical dispersion, representing the range between the first quartile (25th percentile) and the third quartile (75th percentile) of a dataset. It is used to identify the spread of the middle 50% of the data and to detect outliers.