1. What are the key tasks involved in getting ready to work with machine learning modeling?

Key tasks include data collection, data preprocessing (cleaning, handling missing values, normalization), exploratory data analysis (EDA), feature engineering, selecting an appropriate model, splitting data into training and testing sets, training the model, evaluating performance using metrics, and tuning hyperparameters. These steps ensure the model is built on reliable data and performs effectively.

2. What are the different forms of data used in machine learning? Give a specific example for each of them.

Forms of data include:

- Numerical data: Continuous values, e.g., temperature readings.

- Categorical data: Distinct categories, e.g., colors like red, blue, green.

- Ordinal data: Categorical data with a meaningful order, e.g., customer satisfaction ratings (low, medium, high).

- Time series data: Data points indexed in time order, e.g., stock prices over time.

3. Distinguish:

1. Numeric vs. categorical attributes:

- Numeric attributes: Represented by numbers (e.g., height, weight).

- Categorical attributes: Represent distinct groups or categories (e.g., gender, colors).

2. Feature selection vs. dimensionality reduction:

- Feature selection: Process of selecting the most relevant features for the model.

- Dimensionality reduction: Techniques like PCA reduce the number of features by transforming the original features into a lower-dimensional space.

4. Make quick notes on any two of the following:

1. The histogram:

A histogram displays the distribution of a dataset by showing the frequency of data points within specified ranges (bins). It helps in understanding the data's underlying distribution, identifying skewness, and detecting outliers.

2. Use a scatter plot:

A scatter plot visualizes the relationship between two variables. Each point represents an observation, plotted based on its values for the two variables. Scatter plots help identify patterns, trends, correlations, and outliers.

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

Investigating data helps uncover patterns, relationships, and anomalies, ensuring the data is suitable for modeling. Qualitative data is explored through thematic analysis and categorization, while quantitative data is analyzed using statistical methods and visualizations. The approach varies due to the different nature of these data types.

6. What are the various histogram shapes? What exactly are ‘bins'?

Histogram shapes include:

- Uniform: Evenly distributed data.

- Normal: Bell-shaped curve.

- Skewed: Data skewed left or right.

- Bimodal: Two peaks.

Bins are intervals that represent the range of data points in a histogram, allowing visualization of data distribution within these intervals.

7. How do we deal with data outliers?

Outliers can be handled by:

- Removing them: If they are errors or irrelevant.

- Transforming data: Using techniques like log transformation.

- Using robust statistical methods: That are less sensitive to outliers.

- Capping/flooring: Setting upper and lower bounds.

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

Central inclination measures include:

- Mean: Average of all data points.

- Median: Middle value of ordered data.

- Mode: Most frequent value.

The mean can vary significantly from the median in skewed distributions due to extreme values affecting the average more than the middle value.

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

A scatter plot investigates bivariate relationships by plotting data points based on two variables, revealing correlations, patterns, and trends. Outliers appear as points distant from the main cluster, indicating deviations from the typical relationship.

10. Describe how cross-tabs can be used to figure out how two variables are related.

Cross-tabs (contingency tables) display the frequency distribution of two categorical variables. By analyzing the table, one can observe the interaction between variables, identify patterns, and measure the strength of their relationship using metrics like Chi-square.