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

Answer 1: Key tasks in preparing for machine learning modeling include:

* Data Collection: Gathering relevant data for the problem at hand.
* Data Cleaning: Addressing missing values, handling outliers, and ensuring data quality.
* Feature Engineering: Creating meaningful features to enhance model performance.
* Data Splitting: Dividing data into training and testing sets for model evaluation.
* Data Scaling: Normalizing or standardizing features to ensure fair comparison.
* Exploratory Data Analysis (EDA): Understanding the data's characteristics and patterns.

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

Answer 2:

* Numeric Data: Represented by numerical values. Example: Temperature readings.
* Categorical Data: Consists of categories or labels. Example: Types of fruits (apple, orange, banana).
* Ordinal Data: Has a meaningful order. Example: Education levels (high school, bachelor's, master's).
* Text Data: Involves textual information. Example: Customer reviews.

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

**2. Feature selection vs. dimensionality reduction**

Answer 3:

**Numeric vs. Categorical Attributes:**

* Numeric Attributes: Represent measurable quantities (e.g., height, weight).
* Categorical Attributes: Represent categories or labels (e.g., color, gender).

**Feature Selection vs. Dimensionality Reduction:**

* Feature Selection: Involves choosing relevant features to include in the model.
* Dimensionality Reduction: Aims to reduce the number of features by transforming them into a lower-dimensional space.

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

**1. The histogram**

**2. Use a scatter plot**

**3. PCA (Personal Computer Aid)**

Answer 4:

**The Histogram:**

* Visual representation of data distribution.
* Helps identify patterns, central tendency, and spread.

**Use a Scatter Plot:**

* Displays the relationship between two variables.
* Identifies patterns, correlations, and outliers.

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

Answer 5: Investigating data is crucial to understand its characteristics, detect anomalies, and inform modeling decisions. The approach may differ for qualitative and quantitative data, but the goal remains the same: to gain insights and make informed decisions.

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

Answer 6: Histogram shapes include:

* Normal Distribution: Symmetrical bell curve.
* Skewed Distribution: Skewed left or right.
* Bimodal Distribution: Two distinct peaks.

'Bins' in a histogram are intervals that group data points. The number of bins affects the granularity of the histogram.

**7. How do we deal with data outliers?**

Answer 7: Outliers can be addressed by:

* Transformation: Applying mathematical operations to normalize the data.
* Removal: Excluding extreme values, if justified.
* Imputation: Replacing outliers with more typical values.

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

Answer 8: Central inclination measures include mean, median, and mode. The mean is sensitive to outliers, causing it to vary significantly from the median in datasets with extreme values.

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

Answer 9: A scatter plot visualizes relationships between two variables, helping identify patterns and correlations. Outliers can be spotted as data points that deviate significantly from the overall trend or cluster in unexpected ways.

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

Answer 10: Cross-tabs, or contingency tables, display the distribution of two categorical variables. They help reveal relationships and dependencies between variables by showing how the frequencies or proportions are distributed across different categories.