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

**Ans:** Key Tasks in Getting Ready for Machine Learning Modeling:

Data Collection: Gather relevant data from various sources.

Data Preprocessing: Clean and preprocess data to handle missing values, outliers, and inconsistencies.

Feature Engineering: Select, create, or transform features to make them suitable for modeling.

Data Splitting: Divide the dataset into training, validation, and test sets for model evaluation.

Feature Scaling: Normalize or standardize features to ensure they have a similar scale.

Exploratory Data Analysis (EDA): Analyze and visualize the data to gain insights and understand its characteristics.

Model Selection: Choose an appropriate machine learning algorithm or model.

Model Training: Train the selected model on the training dataset.

Model Evaluation: Assess model performance using appropriate metrics.

Hyperparameter Tuning: Optimize model hyperparameters for better results.

Model Deployment: Deploy the trained model for making predictions or decisions.

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

**Ans**: Different Forms of Data in Machine Learning:

Numeric Data: Numeric data consists of numerical values and can be continuous or discrete. Example: Temperatures recorded in degrees Celsius.

Categorical Data: Categorical data represents discrete categories or labels. Example: Vehicle types (e.g., sedan, SUV).

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

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

**Ans:** Distinctions:

Numeric vs. Categorical Attributes:

Numeric attributes contain numerical values that can be measured or counted (e.g., height, age).

Categorical attributes represent categories or labels, often with no inherent order (e.g., color, vehicle make).

Feature Selection vs. Dimensionality Reduction:

Feature selection involves selecting a subset of the most relevant features from the original set of attributes while retaining their original meaning.

Dimensionality reduction reduces the number of features by creating new features (e.g., principal components) that capture most of the original data's variance.

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

**1. The histogram**

**2. Use a scatter plot**

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

**Ans:** Quick Notes:

The Histogram: A histogram is a graphical representation of the distribution of data. It displays data points in bins or intervals on the x-axis and their frequency (or count) on the y-axis.

Scatter Plot: A scatter plot is a graphical representation used to visualize the relationship between two continuous variables. Each data point is plotted as a point on the graph, and it helps identify patterns, trends, or outliers.

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

**Ans:** Importance of Investigating Data:

Data investigation is essential to understand the dataset's characteristics, identify patterns, and assess data quality.

There may be some differences in how qualitative (categorical) and quantitative (numeric) data are explored. For qualitative data, methods like frequency tables and bar charts are more common, while for quantitative data, histograms and scatter plots are frequently used.

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

**Ans:** Histogram Shapes and 'Bins':

Histogram shapes can include:

Normal Distribution: Bell-shaped, with a symmetric peak in the center.

Skewed Distribution: Asymmetric, with a long tail on one side.

Bimodal Distribution: Two distinct peaks.

'Bins' in a histogram are the intervals into which the data range is divided. They determine how data is grouped and displayed on the x-axis.

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

**Ans:** Dealing with Data Outliers:

Outliers can be addressed by:

Truncation: Capping extreme values at a certain threshold.

Transformation: Applying mathematical transformations (e.g., log transformation) to make the data more normally distributed.

Robust Statistics: Using robust statistical measures that are less affected by outliers.

Removing Outliers: In some cases, outliers may be removed, but this should be done cautiously and with proper justification.

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

**Ans:** Central Inclination Measures:

Measures of central inclination include the mean (average), median (middle value), and mode (most frequent value).

Mean can vary significantly from median in datasets with outliers or skewed distributions. The median is more robust to outliers.

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

**Ans:** Scatter Plot for Investigating Bivariate Relationships:

Scatter plots display the relationship between two variables by plotting data points on a two-dimensional graph.

Outliers can be identified visually as data points far from the main cluster.

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

**Ans:** Cross-Tabulation for Understanding Relationships:

Cross-tabs (contingency tables) are used to analyze the relationships between two categorical variables.

They show the frequency or count of data points that fall into each combination of categories for the two variables. This helps identify associations and dependencies between the variables.