## Q1

Key Tasks in Preparing for Machine Learning Modeling:

Data Collection: Gathering relevant data from various sources.

Data Cleaning: Removing or handling missing data, dealing with outliers, and correcting errors.

Feature Engineering: Creating new features or transforming existing ones to improve model performance.

Data Splitting: Dividing data into training, validation, and test sets.

Model Selection: Choosing an appropriate machine learning algorithm.

Model Training: Fitting the chosen model on the training data.

Hyperparameter Tuning: Optimizing model hyperparameters.

Model Evaluation: Assessing model performance using suitable metrics.

Model Deployment: Implementing the model in a production environment.

Forms of Data in Machine Learning:

## Q2

Numeric Data: This includes data that represents measurable quantities. Example: Temperature readings.

Categorical Data: This includes data with discrete categories or labels. Example: Car brands (Toyota, Ford, Honda).

Text Data: Unstructured textual information, like articles or reviews.

Image Data: Pixels of an image, often used in computer vision tasks.

Time Series Data: Sequential data points recorded over time, like stock prices.

Audio Data: Waveform data used in speech recognition or audio analysis.

Distinguishing Concepts:

## Q3

Numeric vs. Categorical Attributes:

Numeric attributes are continuous and represent measurable quantities (e.g., age, temperature).

Categorical attributes are discrete and represent categories or labels (e.g., color, city).

Feature Selection vs. Dimensionality Reduction:

Feature Selection involves choosing a subset of relevant features from the original set to improve model performance.

Dimensionality Reduction aims to reduce the number of features by transforming them into a lower-dimensional space (e.g., Principal Component Analysis - PCA).

Quick Notes:

## Q4

Histogram: A graphical representation of the distribution of numerical data. It consists of bins (intervals) on the x-axis and the frequency of data points in each bin on the y-axis.

Scatter Plot: A plot of data points in a two-dimensional space, where each point represents a pair of values from two variables. It is used to visualize relationships between variables.

Why Investigate Data:

## Q5

It helps in understanding the characteristics of the dataset.

Identifying data quality issues like missing values or outliers.

Revealing patterns or relationships that can guide modeling.

Ensuring data is suitable for the chosen machine learning task.

Histogram Shapes and Bins:

## Q6

Common shapes: Normal (bell-shaped), Skewed (positively or negatively), Uniform, Bimodal (two peaks).

Bins are intervals used to group data in a histogram. The number of bins affects the granularity of the representation.

Dealing with Data Outliers:

## Q7

Options include removing outliers, transforming data, or using robust statistical methods.

Deciding depends on the impact of outliers on the analysis and the specific problem.

Central Inclination Measures:

## Q8

Mean: The average of data values.

Median: The middle value when data is sorted.

They can differ significantly in datasets with outliers because mean is sensitive to extreme values.

Scatter Plot for Bivariate Relationships:

## Q9

It shows how two variables relate by displaying data points in a two-dimensional space.

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

Cross-Tabulation (Cross-Tab):

## Q10

It's used to explore relationships between two categorical variables.

It creates a table showing the frequency of combinations of categories from both variables, helping to identify associations.