## Q1

Key Tasks in Machine Learning and Data Pre-processing:

Data Collection: Gathering relevant data from various sources.

Data Cleaning: Handling missing values, outliers, and errors.

Feature Engineering: Creating new features or transforming existing ones.

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

Model Selection: Choosing an appropriate machine learning algorithm.

Model Training: Fitting the model to the training data.

Hyperparameter Tuning: Optimizing model settings.

Model Evaluation: Assessing model performance with metrics.

Model Deployment: Implementing the model in a real-world application.

Quantitative and Qualitative Data:

## Q2

Quantitative Data: Numeric and measurable. Examples include age, temperature, and income.

Qualitative Data: Categorical or non-numeric. Divided into nominal (categories with no inherent order, like colors) and ordinal (categories with a meaningful order, like education levels).

Basic Data Collection Example:

## Q3

Suppose you're collecting data about cars:

Numeric: Car Price, Engine Size.

Categorical (Nominal): Car Color, Manufacturer.

Categorical (Ordinal): Car Condition (Excellent, Good, Fair).

Text: Car Description.

Causes of Data Issues:

## Q4

Missing Data: Data points with no recorded values.

Outliers: Extreme values that deviate significantly from the majority.

Inconsistent Data: Errors, discrepancies, or conflicting information.

Imbalanced Data: Skewed distribution of target classes.

Approaches to Categorical Data Exploration:

## Q5

Frequency Counts: Count occurrences of each category.

Bar Charts: Visualize categorical data distribution.

Cross-Tabulation (Cross-Tab): Explore relationships between two categorical variables.

## Q6

Impact of Missing Values:

Can lead to biased or inaccurate model predictions.

May reduce the amount of usable data.

## Q7

Methods for Dealing with Missing Data:

Removing Rows: Remove data points with missing values.

Imputation: Fill missing values with mean, median, mode, or predicted values.

Advanced Imputation: Use techniques like regression or k-nearest neighbors.

## Q8

Data Pre-processing Techniques:

Dimensionality Reduction: Reduce the number of features while preserving important information (e.g., PCA).

Feature Selection: Choose a subset of relevant features for modeling.

## Q9

IQR and Box Plots:

IQR (Interquartile Range): Measure of statistical dispersion between the first quartile (Q1) and third quartile (Q3).

Box Plot Components: Box (IQR), Whiskers (data range), Outliers (individual data points).

Length of Whiskers: Depends on data distribution.

Box Plots for Outliers: Outliers are typically shown as individual points beyond the whiskers.

## Q10

Comparisons:

Data with Nominal and Ordinal Values:

Nominal: No inherent order (e.g., colors).

Ordinal: Meaningful order (e.g., education levels).

Histogram vs. Box Plot:

Histogram: Visualizes data distribution using bins.

Box Plot: Shows quartiles, IQR, and identifies outliers.

Average vs. Median:

Average: Sum of values divided by count.

Median: Middle value when data is ordered. Resistant to outliers.