Q1. What are the key tasks involved in getting ready to work with machine learning modeling?

**Key Tasks for Machine Learning Modeling:**

* **Data Collection and Preprocessing:** Gather relevant data, clean, and handle missing values and outliers.
* **Feature Engineering:** Select and create relevant features from the data.
* **Data Splitting:** Divide the data into training, validation, and test sets.
* **Model Selection:** Choose appropriate algorithms and architectures.
* **Model Training:** Train the model on the training set.
* **Model Evaluation:** Assess model performance on validation and test sets.
* **Hyperparameter Tuning:** Optimize model hyperparameters for better results.
* **Model Interpretation:** Analyze model predictions and feature importance.
* **Deployment:** Deploy the model in real-world applications.
* **Monitoring and Maintenance:** Continuously monitor and update the model as needed.

Q2. What are the different forms of data used in machine learning? Give a specific example for each of them.

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

* **Numerical Data:** Includes continuous or discrete numeric values. Example: Age, temperature.
* **Categorical Data:** Represents categories or labels. Example: Gender, color.
* **Ordinal Data:** Categorical data with a defined order. Example: Education levels (High School, College, PhD).
* **Text Data:** Unstructured textual information. Example: Customer reviews.
* **Image Data:** Pixel values of images. Example: Digital photographs.
* **Time Series Data:** Sequential data points over time. Example: Stock prices over days.

Q3. Distinguish:

1. Numeric vs. categorical attributes

2. Feature selection vs. dimensionality reduction

**Distinctions:**

* **Numeric vs. Categorical Attributes:** Numeric attributes represent quantities, while categorical attributes represent categories. Numeric attributes can undergo mathematical operations, while categorical attributes may need encoding.
* **Feature Selection vs. Dimensionality Reduction:** Feature selection selects a subset of relevant features, while dimensionality reduction reduces the number of features while preserving important information.

Q4. Make quick notes on any two of the following:

1. The histogram

2. Use a scatter plot

3.PCA (Personal Computer Aid)

**Quick Notes:**

* **Histogram:** A graphical representation of data distribution showing frequencies of data within predefined intervals (bins).
* **Scatter Plot:** A graph displaying individual data points as dots on a two-dimensional plane.
* **PCA (Principal Component Analysis):** A dimensionality reduction technique that transforms data into a new coordinate system to maximize variance and reduce redundancy.

Q5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

**Investigating Data:**

* Data investigation helps understand data patterns, relationships, and potential issues.
* Qualitative data exploration involves visualization and understanding categorical distributions.
* Quantitative data exploration involves summary statistics, distributions, and relationships.

Q6. What are the various histogram shapes? What exactly are ‘bins'?

**Histogram Shapes and Bins:**

* Histogram shapes include symmetric (normal), skewed (positively or negatively), bimodal, and uniform.
* Bins are intervals used to group data in a histogram.

Q7. How do we deal with data outliers?

**Dealing with Data Outliers:** Outliers can be treated by removing, transforming, or imputing them based on domain knowledge and impact on analysis.

Q8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

**Central Inclination Measures:**

* Mean, median, and mode are measures of central tendency.
* Mean varies from median in skewed distributions or when outliers are present.

Q9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

**Scatter Plot for Bivariate Relationships:**

* Scatter plots visualize relationships between two variables.
* Outliers can be identified as data points far from the main cluster.

Q10. Describe how cross-tabs can be used to figure out how two variables are related.

**Cross-Tabs for Variable Relationships:** Cross-tabs (contingency tables) display the relationship between two categorical variables, showing frequencies or percentages in each cell.