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

The key tasks involved in preparing to work with machine learning modeling include:

1. **Problem Definition**: Clearly define the problem you are trying to solve and determine whether it is a classification, regression, clustering, or other type of problem.
2. **Data Collection**: Gather the data necessary for training the model. This could include datasets from different sources (databases, APIs, etc.).
3. **Data Cleaning and Preprocessing**: Handle missing data, remove duplicates, outliers, and inconsistencies, and perform necessary transformations like encoding categorical variables, scaling numeric data, etc.
4. **Feature Engineering**: Create new features from the existing ones that may help improve the performance of the model.
5. **Splitting the Data**: Divide the data into training and testing sets (usually 80/20 or 70/30 split).
6. **Model Selection**: Choose an appropriate algorithm or model based on the type of problem and the dataset.
7. **Model Training**: Train the model using the training data.
8. **Model Evaluation**: Evaluate the model's performance using metrics like accuracy, precision, recall, or others appropriate for the problem.
9. **Model Tuning**: Fine-tune the model’s hyperparameters to improve performance.
10. **Model Deployment**: Deploy the model into production for real-world use.

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

* **Numerical data**: Data that is quantifiable and represented by numbers.
  + **Example**: Age, salary, temperature, height.
* **Categorical data**: Data that represents categories or labels and is typically non-numeric.
  + **Example**: Gender (Male/Female), color (Red/Blue/Green), country (USA/Canada).
* **Text data**: Data that represents words, sentences, or documents.
  + **Example**: Product reviews, social media posts, emails.
* **Image data**: Data represented by pixel values in images.
  + **Example**: Photographs, scanned documents, facial recognition images.
* **Time series data**: Data that is indexed by time, where observations are recorded at different time intervals.
  + **Example**: Stock prices, weather data, sensor data.

**3. Distinguish:**

1. **Numeric vs. categorical attributes**:
   * **Numeric attributes**: These are variables that represent measurable quantities and are represented as numbers (e.g., age, income, weight).
   * **Categorical attributes**: These are variables that represent categories or classes, and the values are typically labels or names (e.g., gender, country, product type).
2. **Feature selection vs. dimensionality reduction**:
   * **Feature selection**: The process of selecting a subset of relevant features from the original set of features based on their importance for model performance.
   * **Dimensionality reduction**: The process of reducing the number of input features in a dataset while maintaining its relevant characteristics. This is done through techniques like Principal Component Analysis (PCA).

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

1. **The histogram**:
   * A histogram is a graphical representation of the distribution of a dataset. It displays the frequency of data points within certain ranges or "bins." It is used to understand the underlying distribution of data (e.g., normal, skewed).
2. **Use a scatter plot**:
   * A scatter plot is a visual representation of the relationship between two continuous variables. Each point represents an observation in the dataset, and the axes correspond to the values of the variables. Scatter plots help identify correlations, trends, or outliers.
3. **PCA (Principal Component Analysis)**:
   * PCA is a technique used for dimensionality reduction. It transforms a large set of correlated variables into a smaller set of uncorrelated variables (principal components), while retaining as much variance as possible. PCA is widely used for simplifying complex datasets and making them easier to visualize or analyze.

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

It is necessary to investigate data to:

* Understand its structure and ensure it is appropriate for modeling.
* Identify and handle any missing values, outliers, or noise in the data.
* Discover relationships or patterns in the data that may help improve model performance.
* Ensure that data is consistent and relevant for the problem at hand.

**Discrepancy between qualitative and quantitative data exploration**:

* **Quantitative data**: Explored through statistical methods like mean, median, standard deviation, correlations, and visual tools like histograms, box plots, and scatter plots.
* **Qualitative data**: Explored using techniques like frequency counts, cross-tabulations, and visualizations like bar charts, word clouds, or pie charts.

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

* **Histogram shapes**:
  + **Normal distribution**: A bell-shaped curve with most of the data points around the mean.
  + **Skewed distribution**: The distribution is skewed to the left or right, meaning the tail on one side is longer than the other.
  + **Uniform distribution**: Data points are evenly distributed across the range of values.
  + **Bimodal distribution**: Data has two peaks, indicating two dominant values or groups.
* **Bins**: Bins in a histogram are the intervals or ranges into which data is grouped. Each bin represents a range of values, and the height of the bin indicates how many data points fall into that range.

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

To deal with data outliers, we can:

* **Remove outliers**: If the outliers are due to errors or irrelevant data, they can be removed.
* **Transform data**: Apply transformations (like log or square root) to reduce the impact of outliers.
* **Cap or clip values**: Limit the extreme values by setting upper or lower thresholds (e.g., replacing outliers with the nearest acceptable value).
* **Use robust algorithms**: Some machine learning algorithms (like tree-based models) are less sensitive to outliers and can handle them more effectively.

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

**Central inclination measures**:

* **Mean**: The average of all the values in a dataset.
* **Median**: The middle value when the data is ordered from least to greatest.
* **Mode**: The value that appears most frequently in the dataset.

**Mean vs. Median**:

* The mean can vary significantly from the median in datasets with skewed distributions or extreme outliers because the mean is sensitive to all values in the dataset, including outliers. The median is less affected by extreme values, making it more representative of the central tendency in skewed data.

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

A scatter plot can be used to investigate bivariate relationships by plotting two continuous variables along the x and y axes. This helps to visualize any correlation or trends (positive, negative, or none). It also allows for the identification of clusters, linear/non-linear patterns, and potential outliers, which appear as points that fall far from the general pattern of data.

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

Cross-tabulations (or contingency tables) are used to examine the relationship between two categorical variables by displaying the frequency distribution of their values in a tabular form. This can help identify patterns and dependencies between the variables. For example, it can show the relationship between gender and product preference in a marketing survey. Statistical tests like Chi-square can be used to assess whether there is a significant relationship between the variables.