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

- Data collection: Gathering relevant data from various sources.
- Data preprocessing: Cleaning the data, handling missing values, and dealing with outliers.
- Data exploration: Analyzing and visualizing the data to gain insights and identify patterns.
- Feature engineering: Selecting or creating appropriate features that will be used as inputs for the machine learning model.
- Data splitting: Dividing the data into training, validation, and testing sets.
- Model selection: Choosing the appropriate machine learning algorithm or model for the task.
- Model training: Fitting the model to the training data and adjusting its parameters.
- Model evaluation: Assessing the performance of the trained model using appropriate evaluation metrics.
- Model optimization: Tuning the model parameters to improve its performance.
- Deployment: Integrating the trained model into a production environment for making predictions.

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

- Numerical data: Data represented by numbers, such as age, temperature, or income.
- Categorical data: Data represented by categories or labels, such as gender (male/female), color (red/blue/green), or product categories (electronics/clothing/books).
- Text data: Unstructured data represented by text, such as customer reviews, tweets, or articles.
- Image data: Data represented by images or pixels, commonly used in computer vision tasks.
- Time series data: Data collected over a sequence of time intervals, such as stock prices, weather data, or sensor readings.
- Audio data: Data represented by audio waveforms, used in tasks like speech recognition or music classification.

For example, in a dataset of customer records, the age of customers would be numerical data, the gender would be categorical data, and customer reviews would represent text data.

3. Distinguish:

1. Numeric vs. categorical attributes:

- Numeric attributes represent quantities or measurements that can be expressed as numbers, such as height, weight, or temperature. They can be further categorized as continuous (infinite possible values) or discrete (finite possible values).
- Categorical attributes represent qualitative characteristics or labels that are not numerical in nature. Examples include gender, color, or product categories. Categorical attributes can be further divided into nominal (unordered categories) or ordinal (ordered categories).

2. Feature selection vs. dimensionality reduction:

- Feature selection refers to the process of selecting a subset of relevant features from the original set of features. It aims to remove irrelevant or redundant features to improve model performance and reduce computational complexity.
- Dimensionality reduction aims to transform the original high-dimensional feature space into a lower-dimensional space while preserving the essential information. It helps to overcome the curse of dimensionality and can be achieved through techniques like Principal Component Analysis (PCA) or t-SNE (t-Distributed Stochastic Neighbor Embedding).

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 consists of a series of bars where the height of each bar represents the frequency or count of data points falling within a specific range or bin.
- Histograms help visualize the shape of the data distribution, identify outliers, and understand the central tendency and spread of the data.

2. Use a scatter plot:
- A scatter plot is a two-dimensional plot that displays individual data points as dots.
- It is used to investigate the relationship or correlation between two continuous variables.
- Scatter plots help visualize patterns, trends, and the strength of the relationship between variables. They can also reveal the presence of outliers or clusters in the data.

3. PCA (Principal Component Analysis):
- PCA is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional representation.
- It identifies the principal components, which are orthogonal directions that capture the maximum variance in the data.

- PCA helps in visualizing and exploring the structure of complex datasets, reducing noise, and improving computational efficiency in machine learning tasks.

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

- Investigating data is necessary to understand its characteristics, identify patterns, relationships, and outliers, and make informed decisions throughout the machine learning process.
- Data exploration helps in selecting appropriate features, identifying potential issues like missing data or outliers, and gaining insights to guide the modeling process.
- The exploration of qualitative and quantitative data may differ. Qualitative data is explored through techniques such as frequency counts, cross-tabulations, and visualizations specific to categorical data. Quantitative data is explored using statistical measures, histograms, scatter plots, and correlation analysis.

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

- Histograms can have various shapes, indicating different patterns in the data distribution. Some common shapes include:
  - Normal (Gaussian) distribution: Bell-shaped curve with a peak at the mean.
  - Skewed distribution: Either positively skewed (tail on the right) or negatively skewed (tail on the left).
  - Bimodal distribution: Two distinct peaks in the histogram.
  - Uniform distribution: Constant frequency across all bins.

- Bins in a histogram represent intervals or ranges into which the data is divided. The x-axis represents the range of values, and the y-axis represents the frequency or count of data points falling within each bin. The number of bins determines the granularity of the histogram. Choosing an appropriate number of bins is important to accurately represent the underlying data distribution.

7. How do we deal with data outliers?

- Data outliers are extreme values that deviate significantly from the majority of the data points. They can impact the performance and validity of machine learning models. Some approaches to deal with outliers include:
  - Visual inspection: Identify outliers by plotting the data and visually examining points that are far from the main cluster.
  - Statistical methods: Use statistical measures like z-scores or modified z-scores to detect outliers based on their deviation from the mean or median.
  - Trimming or Winsorizing: Replace outliers with a predetermined cutoff value or limit their impact by assigning them values at a specific percentile.
  - Removing outliers: In some cases, outliers can be removed if they are determined to be errors or if they significantly affect the analysis. However, caution must be exercised as outliers may contain valuable information or represent rare events.

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

- Central inclination measures provide information about the center or typical value of a dataset. The two commonly used measures are:
  - Mean: The arithmetic average of a set of values. It is calculated by summing all values and dividing by the number of values.
  - Median: The middle value of a dataset when it is sorted in ascending or descending order. If there is an even number of values, the median is the average of the two middle values.

- The mean and median can vary significantly in certain data sets due to the presence of outliers or skewed distributions. Outliers have a stronger influence on the mean, as it takes into account all values, while the median is more robust to extreme values. Skewed distributions, particularly with long tails, can cause the mean to be pulled towards the tail, resulting in a deviation from the median.

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 is a useful tool to

investigate the relationship between two continuous variables. Each data point is plotted as a dot on the graph, with one variable represented on the x-axis and the other variable on the y-axis.
- By visualizing the scatter plot, patterns or trends in the data can be identified. For example, a positive linear relationship between the variables would be indicated by the dots forming a roughly upward-sloping line.

- Outliers can be detected in a scatter plot as data points that significantly deviate from the overall pattern or trend of the data. Outliers may appear as points that are far away from the main cluster or that fall outside the general pattern of the data points.

However, the identification of outliers in a scatter plot is subjective and requires human judgment, especially in complex datasets.

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

- Cross-tabs, or contingency tables, are used to analyze the relationship between two categorical variables.
- A cross-tabulation presents the frequency or count of observations for each combination of values between the two variables.
- By examining the cross-tabulation, patterns or dependencies between the variables can be observed. For example, it can reveal whether certain categories of one variable are more likely to co-occur with specific categories of the other variable.
- Additional statistical measures, such as chi-squared tests, can be performed on the cross-tabulation to assess the significance of the relationship between the variables.