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

- Understanding the business problem and defining the goal of the model

- Gathering and preprocessing data, including cleaning, transforming, and normalizing the data

- Choosing a suitable algorithm and model architecture

- Training and validating the model on a subset of the data

- Tuning the model hyperparameters to optimize its performance

- Testing the final model on a separate test set to evaluate its generalization performance

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

- Numeric data: data represented by numbers, such as age, height, or temperature

- Categorical data: data that can be divided into distinct categories, such as gender, color, or type of car

- Text data: unstructured data represented by text, such as tweets, reviews, or news articles

- Image data: data represented by images or videos, such as medical images, satellite images, or surveillance footage

- Time-series data: data that changes over time, such as stock prices, weather data, or sensor data

3. Distinguish:

1. Numeric vs. categorical attributes:

Numeric attributes represent values that are continuous, such as temperature or age, while categorical attributes represent values that are categorical, such as gender or color.

2. Feature selection vs. dimensionality reduction:

Feature selection involves selecting a subset of the original features that are relevant to the model, while dimensionality reduction involves transforming the original features into a lower-dimensional space while preserving as much of the original information as possible.

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

1. The histogram:

A histogram is a graph that displays the distribution of a numerical variable. It groups the data into bins and shows the frequency of each bin.

2. Use a scatter plot:

A scatter plot is a graph that displays the relationship between two variables. It shows the values of the variables on the x and y-axis and plots a dot for each data point. It is useful for identifying patterns and outliers in the data.

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 characteristics, such as distribution, outliers, and correlations, and to identify any issues with the data that may affect the model's performance. There may be a discrepancy in how qualitative and quantitative data are explored because qualitative data is often subjective and requires a different approach, such as text analysis or sentiment analysis.

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

The various histogram shapes are:

- Normal distribution: bell-shaped curve with a symmetric distribution

- Skewed left distribution: long tail on the left side of the curve

- Skewed right distribution: long tail on the right side of the curve

- Bimodal distribution: two peaks in the distribution

Bins are intervals that group the data together and are used to create the histogram.

7. How do we deal with data outliers?

We can deal with data outliers by removing them, replacing them with a more appropriate value, or transforming the data using a scaling or normalization technique.

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

The various central inclination measures are mean, median, and mode. Mean is the average of the data and is sensitive to outliers, while median is the middle value in the data and is more robust to outliers. Mode is the most frequent value in the data. Mean varies too much from median in certain data sets because of the presence of outliers, which can significantly affect the mean.

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is