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

The key tasks involved in getting ready to work with machine learning modeling are:

- **Data Collection**: Gathering relevant data for the problem at hand. This can involve acquiring data from various sources, such as databases, APIs, or external datasets.
- **Data Cleaning and Preprocessing**: Handling missing values, dealing with outliers, and addressing data inconsistencies or errors. This step ensures that the data is in a suitable format for further analysis.
- **Feature Engineering**: Creating new features or transforming existing features to improve the predictive power of the model. This can involve techniques like scaling, normalization, one-hot encoding, or creating interaction variables.
- **Data Splitting**: Splitting the dataset into training, validation, and test sets. The training set is used to train the model, the validation set is used for hyperparameter tuning and model selection, and the test set is used to evaluate the final model's performance.
- **Model Selection and Training**: Choosing an appropriate machine learning algorithm or model based on the problem at hand and the available data. The selected model is trained using the training set to learn patterns and relationships.
- **Model Evaluation and Validation**: Assessing the performance of the trained model using suitable evaluation metrics and validating its performance on unseen data. This step helps in understanding how well the model generalizes to new, unseen instances.

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

- **Numeric Data**: Numeric data consists of numerical values that can be measured or counted. Examples include age, height, temperature, or sales revenue.
- **Categorical Data**: Categorical data represents discrete, non-numeric values that belong to specific categories. Examples include gender (male/female), color (red/blue/green), or product category (electronics/clothing/furniture).
- **Text Data**: Text data consists of unstructured textual information. It can include documents, customer reviews, social media posts, or email content.
- **Time-Series Data**: Time-series data is collected over time at regular intervals. It includes data such as stock prices, temperature recordings, or website traffic over time.
- **Image Data:** Image data consists of visual information represented in the form of pixels. Examples include photographs, satellite images, or medical scans.

3.Distinguish between:

Numeric vs. Categorical Attributes:

- Numeric attributes represent measurable quantities with a continuous range of values. They can be used for mathematical calculations and statistical analysis.

- Categorical attributes represent discrete values or categories and cannot be used for mathematical calculations directly. They are often represented as labels or codes.

Feature Selection vs. Dimensionality Reduction:

- Feature selection involves selecting a subset of relevant features from the original set of features. It aims to improve model performance, reduce complexity, and enhance interpretability.

- Dimensionality reduction aims to reduce the number of features by transforming the data into a lower-dimensional space. It helps to alleviate the curse of dimensionality and capture the most important information while minimizing redundancy.

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

- **The Histogram**: A histogram is a graphical representation of the distribution of a dataset. It consists of a set of bins along the x-axis and the frequency or count of data points falling into each bin on the y-axis. Histograms provide insights into the shape, central tendency, and spread of data.

- **Scatter Plot:** A scatter plot is a two-dimensional plot that represents the relationship between two continuous variables. It displays individual data points as dots on the plot, with one variable on the x-axis and the other on the y-axis. Scatter plots help visualize patterns, trends, and correlations between variables.

- **PCA :** PCA, or Principal Component Analysis, is a dimensionality reduction technique used to transform a high-dimensional dataset into a lower-dimensional space while retaining the most important information. It achieves this by identifying the principal components, which are new orthogonal variables that capture the maximum variance in the data.

Here are some key points about PCA:

- Dimensionality Reduction: PCA reduces the dimensionality of the data by projecting it onto a lower-dimensional subspace. This helps in simplifying the data representation and can improve computational efficiency in machine learning tasks.

- Variance Maximization: PCA selects the principal components in such a way that they capture as much variance in the original data as possible. The first principal component explains the most variance, followed by the second, third, and so on.
- Orthogonal Transformation: The principal components obtained through PCA are orthogonal to each other, meaning they are uncorrelated. This ensures that each principal component provides independent information about the data.
- Feature Interpretability: The principal components are linear combinations of the original features. They can be interpreted as new features that represent patterns or combinations of the original features.
- Data Compression: By selecting a subset of the principal components that capture a significant amount of variance, PCA can compress the data while preserving the most important information. This can be useful for storage, visualization, or analysis purposes.
- Applications: PCA is widely used in various fields, including image and signal processing, pattern recognition, data visualization, and feature extraction. It can help in identifying important features, reducing noise, visualizing high-dimensional data, and improving model performance.
- Implementation: PCA involves computing the covariance matrix of the data, performing eigendecomposition or singular value decomposition, and selecting the desired number of principal components based on the explained variance or other criteria.

Overall, PCA is a powerful technique for dimensionality reduction and data exploration, allowing for a simplified representation of complex datasets while preserving important patterns and relationships.

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 gain insights, understand patterns, and make informed decisions in the machine learning process. Both qualitative and quantitative data require exploration, but the methods and techniques used may differ.

- Qualitative data exploration often involves techniques such as content analysis, thematic analysis, or sentiment analysis. It focuses on understanding the meaning, context, and themes present in textual or narrative data.
- Quantitative data exploration typically involves statistical analysis, data visualization, and pattern recognition techniques. It aims to uncover relationships, trends, and patterns in numerical data.

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

Histogram shapes can vary and provide information about the underlying distribution of the data. Some common shapes include:

- Normal Distribution: A symmetric bell-shaped histogram indicating a normal distribution of the data.
- Skewed Distribution: A histogram with a longer tail on one side, indicating a skewed distribution towards higher or lower values.
- Bimodal Distribution: A histogram with two distinct peaks, indicating the presence of two separate groups or modes in the data.

Bins in a histogram represent intervals or ranges along the x-axis where data points are grouped and counted. They define the granularity of the histogram and influence the visual representation of the data distribution.

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

Dealing with data outliers can involve various approaches:

- Removing Outliers: Outliers can be removed from the dataset if they are considered as data errors or noise. However, caution must be exercised as outliers may contain valuable information or indicate anomalies in the data.
- Transforming Outliers: Outliers can be transformed using techniques such as winsorization or log transformation. These methods modify extreme values to bring them closer to the rest of the data.
- Treating Outliers Separately: Outliers can be analyzed separately or assigned special treatment, such as being considered as a separate class or group in the modeling process.

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

Central inclination measures, also known as measures of central tendency, provide information about the typical or central value of a dataset. The main measures include:

- Mean: The arithmetic average of all the values in a dataset. It is calculated by summing all the values and dividing by the total number of observations. The mean is influenced by extreme values and may not be representative if the data has outliers.

- Median: The middle value in a dataset when it is sorted in ascending or descending order. It is less affected by extreme values and provides a better representation of the central tendency, especially for skewed distributions.

The mean and median can vary significantly when there are outliers or the data distribution is skewed. The mean is sensitive to extreme values, whereas the median is more robust to outliers.

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

   Scatter plots are used to investigate bivariate relationships between two variables. They plot individual data points as dots on a graph, with one variable represented on the x-axis and the other on the y-axis. Scatter plots allow visual examination of the relationship, patterns, or trends between the variables. Outliers can be identified as data points that deviate significantly from the general pattern observed in the scatter plot.

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

    Cross-tabs, short for cross-tabulation or contingency tables, are used to understand the relationship between two categorical variables. They provide a tabular representation of the joint distribution of the variables, showing the frequency or count of occurrences for each combination of categories. Cross-tabs help identify associations, dependencies, or patterns between the variables and can provide insights into their relationship.