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

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

them.

3. Distinguish:

1. Numeric vs. categorical attributes

2. Feature selection vs. dimensionality reduction

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

1. The histogram

2. Use a scatter plot

3.PCA (Personal Computer Aid)

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

data are explored?

6. What are the various histogram shapes? What exactly are ‘bins&#39;?

7. How do we deal with data outliers?

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

certain data sets?

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

outliers using a scatter plot?

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

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

1. **Define the Problem:** Clearly articulate the problem you want to solve and determine the objectives of the machine learning model.
2. **Collect Data:** Gather relevant data from various sources, ensuring it is sufficient and representative of the problem domain.
3. **Data Preprocessing:** Clean the data by handling missing values, removing duplicates, and correcting errors. Transform and normalize data as needed.
4. **Feature Engineering:** Identify and create relevant features that will improve model performance. This includes encoding categorical variables and scaling numerical features.
5. **Split Data:** Divide the dataset into training, validation, and test sets to evaluate model performance and prevent overfitting.
6. **Select a Model:** Choose an appropriate machine learning algorithm based on the problem type (classification, regression, etc.).
7. **Train the Model:** Use the training data to fit the model and adjust its parameters.
8. **Evaluate the Model:** Assess the model’s performance using evaluation metrics and the validation set.
9. **Tune Hyperparameters:** Optimize the model by adjusting hyperparameters to improve performance.
10. **Deploy the Model:** Implement the trained model in a real-world application where it can make predictions on new data.
11. **Monitor and Maintain:** Continuously monitor the model’s performance and update it as necessary to adapt to new data or changes in the environment.

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

1. **Structured Data:** Data that is organized into rows and columns, often found in relational databases or spreadsheets.
   * **Example:** A table of customer information with columns for customer ID, name, age, and purchase history.
2. **Unstructured Data:** Data that does not have a predefined format or structure, such as text, images, or audio.
   * **Example:** Social media posts, images of products, or audio recordings of customer service calls.
3. **Semi-Structured Data:** Data that has some organizational properties but does not fit into a traditional database format.
   * **Example:** JSON or XML files containing hierarchical data, such as configuration settings or product catalogs.
4. **Time-Series Data:** Data that is collected at successive points in time, often used for forecasting and trend analysis.
   * **Example:** Stock prices recorded daily over several years or temperature readings taken hourly.
5. **Transactional Data:** Data that captures transactions or events, often used in systems like e-commerce or financial services.
   * **Example:** Records of online purchases, including transaction IDs, timestamps, and item details.

### **3. Distinguish:**

**1. Numeric vs. Categorical Attributes**

* **Numeric Attributes:** Quantitative data that represent measurable quantities. They can be continuous (e.g., height, weight) or discrete (e.g., count of items).
  + **Example:** Age (continuous), Number of purchases (discrete).
* **Categorical Attributes:** Qualitative data that represent categories or groups. They can be nominal (no inherent order) or ordinal (with a meaningful order).
  + **Example:** Gender (nominal), Education level (ordinal).

**2. Feature Selection vs. Dimensionality Reduction**

* **Feature Selection:** The process of selecting a subset of relevant features from the original set of features, aiming to improve model performance and reduce overfitting.
  + **Example:** Choosing the most informative predictors for a regression model, such as selecting only the most influential variables.
* **Dimensionality Reduction:** The process of reducing the number of features by transforming the data into a lower-dimensional space, while preserving as much information as possible.
  + **Example:** Using Principal Component Analysis (PCA) to reduce the number of features in a dataset from hundreds to a smaller set of principal components.

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

**1. The Histogram:**

* **Definition:** A histogram is a graphical representation of the distribution of numerical data. It consists of bars that represent the frequency of data points within specified ranges (bins).
* **Use:** Histograms help visualize the distribution, central tendency, and variability of data, allowing for easy identification of patterns and outliers.

**2. Use a Scatter Plot:**

* **Definition:** A scatter plot is a graph that displays the relationship between two continuous variables. Each point on the plot represents an observation in the dataset.
* **Use:** Scatter plots are used to identify correlations, trends, and potential outliers between variables. They help in understanding the relationship and dependencies between two variables.

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

**Investigation of Data:**

* **Necessity:** Data investigation is crucial for understanding its quality, structure, and patterns. It helps in identifying issues like missing values, outliers, and biases, and guides the preprocessing steps.
* **Qualitative vs. Quantitative Data:**
  + **Qualitative Data:** Often explored through methods like content analysis, thematic analysis, and qualitative data coding. It requires understanding the context and categories within the data.
  + **Quantitative Data:** Explored through statistical analysis, descriptive statistics, and visualizations like histograms and scatter plots. It involves numerical summaries and relationships between variables.

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

**Histogram Shapes:**

* **Normal Distribution:** Bell-shaped curve, indicating a symmetric distribution around the mean.
* **Skewed Distribution:** Can be positively (right-skewed) or negatively (left-skewed) skewed, indicating an asymmetrical distribution.
* **Uniform Distribution:** Flat shape, indicating that data is evenly distributed across the range.
* **Bimodal Distribution:** Two peaks, indicating two prevalent values or groups within the data.

**Bins:**

* **Definition:** Bins are intervals into which data is grouped in a histogram. Each bin represents a range of values and the height of the bar indicates the frequency of data points within that range.

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

**Dealing with Outliers:**

* **Detection:** Identify outliers using statistical methods (e.g., Z-scores, IQR) or visualization tools (e.g., box plots, scatter plots).
* **Handling:**
  + **Removal:** Exclude outliers if they are errors or irrelevant.
  + **Transformation:** Apply transformations (e.g., logarithmic) to reduce the impact of outliers.
  + **Imputation:** Replace outliers with median or mean values, depending on the context.

### **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 values in a dataset, calculated as the sum of values divided by the number of values.
* **Median:** The middle value when the data is ordered, or the average of the two middle values if the dataset has an even number of values.
* **Mode:** The most frequently occurring value in the dataset.

**Variation Between Mean and Median:**

* **Skewed Distributions:** In skewed datasets, the mean can be significantly different from the median. For example, in a positively skewed dataset, the mean is higher than the median because the mean is affected by extreme values.

### **9. 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:**

* **Investigation:** A scatter plot shows the relationship between two continuous variables. By plotting data points on the x and y axes, you can observe patterns, correlations, and trends between the variables.
* **Outliers:** Outliers can be identified as data points that fall far away from the general pattern or trend in the scatter plot. They appear as isolated points that do not follow the overall distribution.

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

**Cross-Tabs (Cross-Tabulations):**

* **Definition:** Cross-tabs are tables that display the frequency distribution of variables. They show how two categorical variables relate to each other by summarizing the counts or proportions of observations within each combination of categories.
* **Use:** They help in identifying relationships and dependencies between variables. For example, a cross-tab might show the relationship between customer age groups and purchase categories, revealing patterns in how different age groups buy different types of products.