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'?

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.

Answers

**1. Key tasks before ML modeling**

* Define the problem clearly
* Collect relevant data
* Clean and preprocess data (missing values, outliers, normalization)
* Feature selection/engineering
* Split data into training, validation, and test sets

**2. Forms of data in ML (with examples)**

* **Numeric (quantitative)** → Continuous numbers (e.g., temperature, salary)
* **Categorical (qualitative)** → Discrete categories (e.g., gender, product type)
* **Ordinal** → Ordered categories (e.g., customer satisfaction: low, medium, high)
* **Time-series** → Sequential data over time (e.g., stock prices, weather data)
* **Text** → Words or documents (e.g., customer reviews)
* **Image** → Pixel data (e.g., medical scans)

**3. Distinctions**

1. **Numeric vs Categorical Attributes**
   * Numeric → Values have arithmetic meaning (age = 25)
   * Categorical → Values represent categories (color = red)
2. **Feature Selection vs Dimensionality Reduction**
   * Feature selection → Choosing most important original features
   * Dimensionality reduction → Transforming features into fewer dimensions (e.g., PCA)

**4. Quick notes (examples)**

1. **Histogram**: Shows frequency distribution of numeric data; bins group continuous values.
2. **Scatter plot**: Plots two numeric variables; reveals correlation, trends, and outliers.
3. **PCA (Principal Component Analysis)**: Reduces dimensionality while preserving variance; simplifies data structure.

**5. Why investigate data**

* Identify patterns, missing values, anomalies, and distributions
* **Qualitative data** → Summarized via counts, proportions, or bar charts
* **Quantitative data** → Summarized via mean, median, standard deviation, histograms

**6. Histogram shapes & bins**

* **Shapes**:
  + Symmetric
  + Skewed right/left
  + Uniform
  + Bimodal
* **Bins**: Intervals that group continuous data to display frequency in a histogram

**7. Dealing with outliers**

* Remove them if erroneous
* Transform variables (log, sqrt)
* Use robust statistics (median, IQR)
* Cap/floor extreme values

**8. Central tendency measures**

* **Mean**: Average
* **Median**: Middle value
* **Mode**: Most frequent value  
  **Why mean varies from median**: Skewed data or extreme outliers distort mean.

**9. Scatter plot for bivariate relationships**

* Plots two numeric variables on x- and y-axes
* Helps detect correlation (positive, negative, none)
* Outliers appear as points far from general trend

**10. Cross-tabs for variable relationships**

* Tables showing frequency of combinations of two categorical variables
* Helps detect patterns, associations, or dependencies (e.g., gender vs. product preference)