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

Key task for getting ready to work with machine learning modelling are:

* Data collection: Collecting data from various sources like api, database, local, etc.
* Data preprocessing: Data should be cleaned. Removing nan values, handling missing values, etc.
* Feature selection and engineering: Select relevant features
* Splitting data: Splitting data into Training set, testing set, and validation/development set.
* Model selection: Selecting the appropriate machine learning algorithm.
* Model training: Training the model
* Model evaluation: Evaluating the model
* Hyperparameter tuning: Hypertune the parameter to achieve the goal
* Model Deployment: Deploy the model
* Maintenance: Maintain the model by continuously monitoring it.

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

Various types of data used in machine learning are:

* Numerical data: Stock prices, temperature readings, etc. for regression problems
* Categorical data: Gender, yes/no, 0/1, malignant/benign etc. for classification problems.
* Text data: Emails, social media posts, tweets, etc for NLP techniques
* Image data: photos, drawings, etc. CNN is used to process these data.
* Time series data: Weather data, stock data, etc.
* Audio data: speech data, music, etc

**3. Distinguish:**

**A) Numeric vs. categorical attributes**

Numerical data represents numerical continuous data for example temperature readings where as categorical data consist of category data or grouped data.

**B) Feature selection vs. dimensionality reduction**

Selecting features that are more relevant to target variable that contribute the most to predict. Whereas dimensionality aim to transform the dataset from high dimensional data to lower dimensional data

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

**1. The histogram:** It is the graphical representation of the data. It represents a visual summary of the frequency or count of data. Mostly used to represent continuous or discrete data

**2. Use a scatter plot:** Plotting individual data points on a graph. It represents by dot or symbols. It is useful for detecting relationships between varaibles. It is effective in visualizing in continuous or quantitative data.

**3. PCA (Principal component analysis)** It is a dimensionality reduction technique that transforms high dimensional data into low dimensional data. It is also used for data visualization, feature selection, noise reduction, and preprocessing too.

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

Investigating data allows us to understand the characteristics, patterns, and relationships within the data. While investigating it gives insight of data that can help in decision-making. It uncovers hidden pieces of information

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

Some common histogram shapes are:

* Normal Distribution
* Skewed Distribution
* Uniform Distribution
* Bimodal Distribution

Bins represent intervals or ranges into which data is divided.

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

Some approaches to handle outliers are:

* Removal of the data
* Transform the data
* Imputation of the data where we can remove with the estimates values

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

The three commonly used central inclination measures are Mean, Median and Mode. Because of outliers it affects a lot.

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

To investigate bivariate relationships using scatter plot we can check few things

* Pattern
* Direction
* Strength
* Outliers

Yes it is possible to find outliers

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

We can create contingency table, or calculate row and column percentage, analyze pattern..