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

Ans: **data gathering, data pre processing, model training, model evaluation. Apart from these, the knowledge of language, statistics and ML algorithms are also required.**

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

Ans**: in machine learning we use numerical data such as age, house price,**

**Categorical variable such as good vs bad/smoker vs non smoker/male vs female**

**Structured vs unstructured data: ML can handle both types of data. The data which is In a tabular format and the data which is in the form of images, audio and video files.**

3. Distinguish:

1. Numeric vs. categorical attributes: **numerical attributes are numbers like age, height and categorical attributes are the ones which can be divided into different classes like gender, color or smoker or non smoker**

2. Feature selection vs. dimensionality reduction: **these both concepts are used to reduce the number of features in a given dataset but the only difference is feature selection only selects the feature by excluding the features which are not going to be used and it doesn’t transform the features but dimensionality reduction basically reduce the dimensionality by transforming the features.**

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

1. The histogram: **the histogram is used to see the occurrences of a particular observation. It also helps to see the type of distribution of a dataset whether it is a normal distribution or standard normal distribution or log distribution. This is achieved by smoothing the graph with the use of probability density function.**

2. Use a scatter plot: **scatter plot is used to find the comparison of two or more features in a data set. We can determine the relationship of two or more features and how are affecting each other.**

3.PCA (Personal Computer Aid): **I think PCA here stands for Principal Component Analysis which is an unsupervised Machine learning and it is used for dimensionality reduction.**

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

Ans: **it is of paramount importance to investigate and understand the data. A better understanding of data helps you perform data pre-processing steps and feature engineering and select the right ML algorithm. Some of the aspects of data exploration could be different when it comes to qualitative and quantitative data. You may have to calculate mean or median for the quantitative data and mode for the qualitative data.**

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

Ans: **the different types of histogram shapes are uniform histogram, symmetric histogram, bimodal histogram.**

**bins are categories or groups you want to divide your data into and show them on a graph. Bins are used to divide data into a series of intervals.**

7. How do we deal with data outliers?

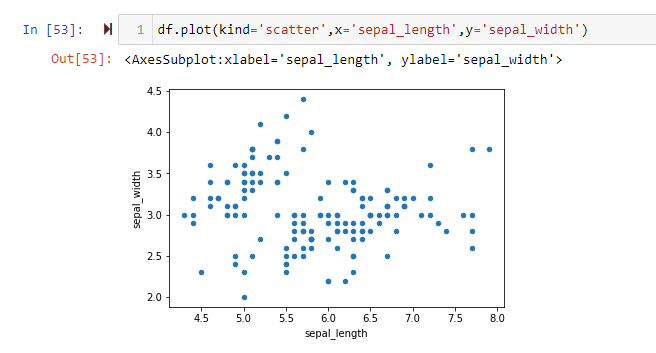
Ans: **different techniques can be employed to deal with outliers. They can be removed or they can be imputed with mean/median.**

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

Ans: **there are 3 measures of central tendency including mean, median and mode. Mean would vary a lot from median when there are outliers in a dataset as they would cause a significant shift in the mean value depending upon the magnitude of the outliers.**

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

Ans: **below is an example of a scatter plot showing a bivariate relationship using df.plot() method. Here you can see how sepal\_length and sepal\_width are related to each other. In linear regression when we use the best fit line then we can determine outliers by choosing the data points which are far from the best fit line.**



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

Ans: **cross tabulation is used to compare two categorical variables. One categorical variable is used to determine the rows of the table and the other categorical variable is used to determine the columns of the table.**