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

First is to define a problem statement that we wish to solve using machine learning.

Second is to collect data for our machine learning problem statement.

Third is to tabulate that data.

Fourth is to clean the data. Cleaning the data may include imputing missing values, removing outliers, standardizing the keywords or headers, conversions to proper datatypes etc.

Fifth step would be to perform an exploratory data analysis to find the key details that would later help in our modeling.

Last step before machine learning modelling can begin would be to standardize data and split the cleaned data into training and test data.

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

Qualitative data -> One that does not necessarily have an order to it and is categorical in nature. For example, in IRIS dataset, the qualitative data would be the variety having class labels or categories Virginicolor, Setosa and Versicolor

Quantitative data -> One that has an order to it and is numerical in nature. For example, Petal length in IRIS dataset is a quantitative data.

Time series data -> Time series data is one that has time based data. For example, Any covid-19 based data will have date time against infections, death etc. because time plays an important role in Covid analysis in finding the spread rate etc.

Text based data -> String based data that are generally remarks, comments or reviews, used for Natural Language Processing.

**3. Distinguish:**

1. **Numeric vs. categorical attributes**

Numerical attributes are those data types that are numerical in nature. More importantly, they have an order to themselves, are mathematical and can be of integer or float datatype

Categorical attributes are those data types that provide characteristics of something. It can be numeric but without any order or mathematical meaning to it.

1. **Feature selection vs. dimensionality reduction**

Feature selection is a technique using which, the most important of features or those features that bring most information with them to machine learning modeling process in predicting or clustering are included and the rest are removed

Dimensionality reduction is a process by which the explanatory features are transformed and elevated to a lower dimension. It is also the process by which the irrelevant features are removed. Several machine learning algorithms are highly negatively affected by high dimensionality, due to which this process becomes very important.

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

1. **The histogram:** Histograms are graphs used to visualize continous and discrete quantitative variables and distribution of data. Histogram requires bin size to make changes in granularity.

**2. Use a scatter plot:** Scatter plots are used to plot 2-D or 3-D plots using 2 or 3 variables of data points. They are important in visually finding trends in data with low dimensionality.

**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?**

Investigation of data is important to find trends, note interesting points, find errors and missing values and outliers that may require remedy. Quantitative data exploration requires numbers recieved from observation and experiment. Qualitative data exploration is different as most of it hinges on frequency table to formulate observations.

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

Histogram shapes can be left skewed, right skewed, approximately normal, approximately uniform, bimodal and more. Bins are classes in which several qualitative values are clubbed together based on required granularity along with their corresponding frequencies

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

Outliers can be removed along with the entire data point if they are in low numbers and their removal does not result in heavy reductions of data. Else, outliers can be replaced with imputed values if highly extreme and low in numbers against a high dimensional data.

Outliers might also be bringing some meaning in the data in case of skewed distributions. In that case, outliers must be kept.

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

Central tendency measures are of primarily 2 types

numeric averages - Arithmetic mean, geometric mean, harmonic mean

and positional averages - median and mode

Mean can vary too much from median if the data has extremities or outliers.

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

Scatterplots can be a great way to find linear relationships between two variables having highly positive or negative correlation coefficient as most of the data points would be in a straight line. Other non linear relationships like quadratic or trigonometric relations can also be approximated and identified visually between two variables. Outliers can also be found using scatter plots.

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

Cross tabs or cross tabulation are only used to find relations between two categorical variables where one categorical variable value would be giving row data and the other variable would give column data. Several tests like the chi square test can help find relations based on the cross tab.