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

- There are few regularly utilised tasks involved while working with machine learning modelling –

task 1: Collecting Data

task 2: Preparing the data ie analysis

task 3: Choose the model or algorithm

task 4 Train your machine model

task 5: Evaluation of that trained model

task 6: Parameter Tuning for better accuracy if required

task 7: Predicting the outcome.

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

- Different forms of data used in machine learning are :-

**a. Numerical data** – A kind of data which consists numbers like even or odd , whole numbers etc , which consists of continuous values in them. Example – Stock price ,house Price.

**b. Categorical data** - A form of data which is present in categories or classes. These types of data can be identified with help of labels or names assigned to them. Example – Flowers classes , Genders.

**c. Time series data** – Data which includes time related quantity in it. Example – whether, Activity tracker.

**d. Text** – Data which is in the form of text and is converted into numbers or labels form before providing to train the model. Example – Text analysis .

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

- **Numerical attributes** – it is a form of data which is present in numbers not in any language or in descriptive form. It may contain whole numbers , even odd etc in it. All types of arithematic operations can be performed using this kind of data. This consists of two types distinct Numerical data and continuous numerical data.

- **Categorical attributes** – It is a form of data which is present in categories or classes. These types of data can be identified with help of labels or names assigned to them. These kind data consists like gender ie male or female , binary data such as 0 or 1 etc. This consists of two types Nominal categorical data and ordinal categorical data.

**2. Feature selection vs. dimensionality reduction**

- Both of the methods can be used to get or extract a proper data which could be provided to the model. In Feature selection , we drop the irrelevant features from the dataset and select only those which are required ie subset of dataset for model training but in case of dimensionality reduction we are transforming the data or features to the lower dimension. PCA ie Principle of component analysis could be used to reduce the dimensionality of the data.

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

**1. The histogram**

A Graphical technique which can be used to display the data in the form bars of different heights which include bins in them. The data can be displayed in such a manner that each bar consists of a range of values in it. So, it the bar height is more then the bar consists of more values in it. A histogram displays the shape and spread of continuous sample data

**2. Use a scatter plot**

A Graphical representation of data in the forms scatters. This is plotted with the help of co-ordinates ie x and y. The more the scattered data lies in a particular region the more the values it contains. This plot helps us to find out the relationship between two variables ie x and y.

**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 helps in creating a well balanced and structured data for training a model. Unbalanced or unstructured data can create a huge problem in training a model as well as it creates a big huddle to achieve expected accuracy from the model. There could be even a chance where the model will not proceed further for training due to incorrect data. Therefore , Investigating data which basically includes data cleaning, feature extracting or selection, removal of missing values , not much outliers and well formed data is must.

Qualitative data are the measures of types or classes or groups which are basically into the form of labelled ones ie name , symbol etc. But if we consider Quantitative data it is all about numerical values so either they may be distinct or they may be continuous.

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

- A histogram when plotted do look like few of shapes listed below:-

**a. Bell shaped Histogram** – a kind of plot which is bell in shape and basically looks like normal distribution.

**b. Skewed right or skewed left** – A plot which is skewed either from right side or from left side. So basically most the data range values are present either at right side or left side.

**c. Uniform** – A Plot where data is distributed uniformly

**d. Random** - A plot where data is distributed randomnly.

**Bins** – Bins are nothing but called as intervals or classes present in the range of data. A histogram displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin.

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

- Outliers are such values which are mostly present in the dataset and they are more far away from the values inside the dataset. For detecting whether the dataset has outliers in it or not there are few methods such as calculating Z-Score from looping into the data, Plotting the whole dataset using Boxplot to locate where exactly the outliers are and how many. These outliers bring a major change in the accuracy when they are trained. As well as for a model it becomes difficult to establish the relationship between the features and outliers. Removing them or dropping the outliers would be great method to improve the model accuracy.

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

- Central inclination or central tendency measures have three types in them.

**1. Mean** – Average of all the values present in the data.

**2. Median** – Dividing the data into two parts and finding out the middle values.

**3. Mode** – A value that appears most frequently in that data set.

- Mean is a simple calculation of data in which we take out averages of all values but in some cases median comes into picture and it is complex part as we need to arrange the data into either ascending order and descending order for the computation purpose. So therefore , there is huge chance whether the output of mean data differs a lot with median data.

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

- Scatter plots are plotted with the help of co-ordinates ie x and y. These plots help us to identify the bivariate relationships between x and y. So, In some case we come across few data where we find out correlation between those two variables such as the amount of time spent on reading books of two people etc. and these cases the scatter plot helps to determine whether such a correlation exists by mapping out the points on the graph.

Yes, it is possible to find outliers using scatter plot but not in all cases. So, If the problem statement is of regression, then there could be a chance where we can find out few outliers easily with plot and remove them. But in some cases the scatter plot wont work to find out the outliers and Boxplot gets used to find out these data.

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

- Cross tabs is a process of finding out the correlation between two variables or features.

If two variables are dependent on each other for example if feature 1 is f1 and feature 2 is f2 then if f1 = 2\*f2 then f2 is dependent on feature 1 by twice of it.