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

**Ans:** Data selection -> Data description -> Data analysis(both statistical & graphical) -> Data transformation & derivation of new attributes if any -> Selection of ML algo on basis of EDA -> Data standardization & normalization -> Split Train & test data set -> Prepare ML Model.

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

**Ans:** Numerical data, Categorical data, Time Series data, Text data.

3. Distinguish:

1. Numeric vs. categorical attributes – Numerical attributes are basically which are numbers in nature(e.g. age), and categorical are which basically have categories(e.g. sex – male, female)

2. Feature selection vs. dimensionality reduction - Feature selection yields a subset of features from the original set of features, which are best representatives of the data.

Dimensionality reduction is usually performed either by selecting a subset of the original dimensions or/and by constructing new dimensions.

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

1. The histogram - A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data.

2. Use a scatter plot – Scatter plot is used to identify collinearity between 2 variables. It is a dotted kind of representation of data.

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?

**Ans:** Investigation of data is required because provided data cannot be in the form to feed directly in model, it can have some missing values, multicollinear attributes or non standardized data.

Quantitative data analysis: It is numerical measurements expressed in terms of numbers

Qualitative data analysis: It is categorical measurements expressed in terms of natural language descriptions

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

**Ans:** Different shapes of histogram are –

Normal distribution – bell shaped curve, skewed distribution(left or right), double peak, comb distribution(alternate tall & short).

In a histogram, the total range of data set (i.e from minimum value to maximum value) is divided into 8 to 15 equal parts. These equal parts are known as bins or class intervals.

7. How do we deal with data outliers?

**Ans:** Below are methods to deal with outliers –

Transformation – transform variables can remove outliers, e.g. scaling, log transformation, cube root normalization.

Imputation - Like imputation of missing values, we can also impute outliers. We can use mean, median, zero value in this method.

Separately treating - If there are significant number of outliers and dataset is small, we should treat them separately in the statistical model. One of the approaches is to treat both groups as two different groups and build individual model for both groups and then combine the output. But this technique is tedious when the dataset is large.

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

**Ans:** Mean, Median and Mode are various central inclination measures. Mean vary too much from median because mean is computed by adding up all the values and dividing that score by the number of values, and median is the number found at the exact middle of the set of values.

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

**Ans:** Scatter plot is best to identify relationship between bivariate data. So in a certain dataset, it displays bivariate plot for every 2 variables, so it becomes easy to identify collinearity between every 2 variables.

If there is a regression line on a scatter plot, you can identify outliers. An outlier for a scatter plot is the point or points that are farthest from the regression line.

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

**Ans:** A contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. They are heavily used in survey research, business intelligence, engineering, and scientific research. They provide a basic picture of the interrelation between two variables and can help find interactions between them.