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

A: Data collection, data preparation, model training, evaluation, model tuning

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

A: Labelled data: Contains a target variable. Eg: Call data with spam or not spam labels.

Unlabelled data: Does not contain target variables. Eg: Youtube view history data.

3. Distinguish:

1. Numeric vs. categorical attributes

A: Attributes can take continuous data mostly in the form of numbers while categorical attributes contain discrete data it may contain both number or string values

2. Feature selection vs. dimensionality reduction

A: Feature selection uses the original columns and selects the attributes that have high values based on business and statistical criteria while dimensionality reduction treat all the attributes/columns as vectors and reduces the number of vectors so that all the numerical relations between these vectors are preserved.

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

1. The histogram

A: Histogram is used to represent the frequency of a value.

2. Use a scatter plot

A: Scatter plot is just a collection of points plotted between two attributes on a graph.

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?

A: Sometimes there might be null values, outliers or spelling mistakes that affect the fit of a model. So it is important to investigate data and remove or treat these discrepancies. Quantitative data is numeric and quantitative data might be in the form of strings. So, steps like checking spellings or cases will not be applicable for quantitative data. In the same way checking mean, median values is not applicable for qualitative data.

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

A: Histograms are made up of rectangles. Bin is an individual rectangle in a histogram and it represents a range for which height of the bin represents the frequency.

7. How do we deal with data outliers?

A: Sometimes they are removed from the data, while some other times they are treated as different buckets within the data. Defining limits is also a method to deal with outliers.

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

A: Mean, median and mode are central inclination measures. If the mean varies too much from the median, it might be due to the existence of outliers.

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

A: Yes it is possible to find outliers using a scatter plot. Scatter plots show trends within data, sometimes showing clusters, or shapes like straight lines.

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

A: Row, column and total percentages are calculated using cross-tabs. These percentages can be used to analyse the relation between two variables.