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

- Step 1: Collect Data. ...
- Step 2: Prepare the data.
- Step 3: Choose the model.
- Step 4 Train your machine model.
- Step 5: Evaluation.
- Step 6: Parameter Tuning.
- Step 7: Prediction

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

Most data can be categorized into 4 basic types from a Machine Learning perspective: numerical data, categorical data, time-series data, and text.

3. Distinguish:

1. Numeric vs. categorical attributes

Categorical data refers to a data type that can be stored and identified based on the names or labels given to them. Numerical data refers to the data that is in the form of numbers, and not in any language or descriptive form. Also known as qualitative data as it qualifies data before classifying it

2. Feature selection vs. dimensionality reduction:

Feature selection is simply selecting and excluding given features without changing them whereas Dimensionality reduction transforms features into a lower dimension.

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

1. **The histogram**

   A histogram is a display of statistical information that uses rectangles to show the frequency of data items in successive numerical intervals of equal size

2. **Use a scatter plot:**
   Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.

3.PCA (Personal Computer Aid) :

Principal Component Analysis instead of Personal Computer Aid.

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

Data analysis is important in business to understand problems facing an organisation, and to explore data in meaningful ways. Data in itself is merely facts and figures. Data analysis organises, interprets, structures and presents the data into useful information that provides context for the data.

There exists a fundamental distinction between two types of data: Quantitative data is information about quantities, and therefore numbers, and qualitative data is descriptive, and regards phenomenon which can be observed but not measured, such as language.

6. What are the various histogram shapes?

Histogram shapes are :

Bell-shaped: A bell-shaped picture, shown below, usually presents a normal distribution.

Bimodal: A bimodal shape, has two peaks. This shape may show that the data has come from two different systems. If this shape occurs, the two sources should be separated and analyzed separately.

Skewed right: A skewed distribution can result when data is gathered from a system with has a boundary such as zero. In other words, all the collected data has values greater than zero.

Skewed left: A skewed distribution can result when data is gathered from a system with a boundary such as 100. In other words, all the collected data has values less than 100.

Uniform: uniform distribution often means that the number of classes is too small.

Random: A random distribution often means there are too many classes.

What exactly are 'bins'?

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. Bins are also sometimes called "intervals", "classes", or "buckets".

7. How do we deal with data outliers?

There are many approach to deal with outliers:

1. Drop the outlier records.
2. Cap your outliers data.
3. Assign a new value.
4. Try a transformation.

8. What are the various central inclination measures?

Three of the many ways to measure central tendency are the **mean, median and mode.**

Why does mean vary too much from median in certain data sets?

This problem occurs because outliers have a substantial impact on the mean. Extreme values in an extended tail pull the mean away from the center. As the distribution becomes more skewed, the mean is drawn further away from the center.

9. Describe how a scatter plot can be used to investigate bivariate relationships.

A bivariate data set shows a linear relationship if the scatterplot shows points bunched randomly around a straight line. The points might be tightly bunched and fall almost exactly on a line, or they might be wildly scattered, forming a cloud of points.

Is it possible to find outliers using a scatter plot?

Yes. 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.

In a cross-tabulation, the categories of one variable determine the rows of the table, and the categories of the other variable determine the columns.