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

Solution - Key tasks for machine learning are given below as:

Gathering Data - Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.

Data preparation - After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.

Data Wrangling - Data wrangling is the process of cleaning and converting raw data into a useable format.

Analyse Data - The aim of this step is to build a machine learning model to analyse the data using various analytical techniques and review the outcome.

Train the model - In this step we train our model to improve its performance for better outcome of the problem.

Test the model - In this step, we check for the accuracy of our model by providing a test dataset to it.

Deployment - The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system

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

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

Numerical Data - Numerical data is any data where data points are exact numbers. Statisticians also might call numerical data, quantitative data.

Categorical Data - Categorical data represents characteristics, such as a hockey player's position, team, hometown. Categorical data can take numerical values.

Time Series Data - Time series data is a sequence of numbers collected at regular intervals over some period.

Text - Text data is basically just words. A lot of the time the first thing that you do with text is you turn it into numbers using some interesting functions like the bag of words formulation.

3. Distinguish:

1. **Numeric vs. categorical attributes:**

Solution - A categorical variable is a variable with a set number of groups (gender, colors of the rainbow, brands of cereal), while a numeric variable is generally something that can be measured (height, weight, miles per hour).

3. **Feature selection vs. dimensionality reduction**

Solution - Feature selection is simply selecting and excluding given features without changing them. 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 graphical representation of data points organized into user-specified ranges.

2. Use a scatter plot - Scatter plots are the graphs that present the relationship between two variables in a dataset. It represents data points on a two-dimensional plane or on a Cartesian system.

3.PCA - Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

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

Solution - By looking at data and data trends, we can make decisions based upon research rather than taking a guess and hoping for the best.

Quantitative data is numbers-based, countable, or measurable. Qualitative data is interpretation-based, descriptive, and relating to language.

Quantitative data tells us how many, how much, or how often in calculations. Qualitative data can help us to understand why, how, or what happened behind certain behaviours.

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

Normal Distribution – In a normal or "typical" distribution, points are as likely to occur on one side of the average as on the other

Skewed Distribution – The skewed distribution is asymmetrical because a natural limit prevents outcomes on one side

Double-Peaked or Bimodal – The outcomes of two processes with different distributions are combined in one set of data.

Plateau or Multimodal Distribution – Several processes with normal distributions are combined. Because there are many peaks close together, the top of the distribution resembles a plateau.

Edge Peak Distribution – The edge peak distribution looks like the normal distribution except that it has a large peak at one tail

Comb Distribution – In a comb distribution, the bars are alternately tall and short. This distribution often results from rounded-off data and/or an incorrectly constructed histogram.

Truncated or Heart-Cut Distribution – The truncated distribution looks like a normal distribution with the tails cut off.

Dog Food Distribution - The dog food distribution is missing something—results near the average.

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?

Solution - There are some techniques used to deal with outliers.

Deleting observations - We delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers.

Transforming values - Transforming variables can also eliminate outliers. These transformed values reduces the variation caused by extreme values.

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

Separately treating - One of the approach is to treat both groups as two different groups and build individual model for both groups and then combine the output.

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

Solution - There are three main measures of central tendency: mode. median. mean.

The reason that mean cannot be applied to all distributions is because it gets unduly impacted by values in the sample that are too small to too large. The disadvantage of median is that it is difficult to handle theoretically. There is no easy mathematical formula to calculate the median.

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

Solution - A scatterplot shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point on the graph.

Scatter plots often have a pattern. We call a data point an outlier if it doesn't fit the pattern.

10. Describe how crosstabs can be used to figure out how two variables are related.

Solution - To describe the relationship between two categorical variables, we use a special type of table called a cross-tabulation (or "crosstab" for short). 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