1. **What are the key tasks that machine learning entails?**
   A machine learning task is the type of prediction or inference being made, based on the problem or question that is being asked, and the available data. For example, the classification task assigns data to categories, and the clustering task groups data according to similarity.

   What does data pre-processing imply?
   Data pre-processing in machine learning is the process of preparing the raw data to make it ready for model making. It is the first and the most crucial step in any machine learning model process
2. **Describe quantitative and qualitative data in depth. Make a distinction between the two.**

   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 behaviors3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

4. What are the various causes of machine learning data issues? What are the ramifications?

Data quality can be considered as a major common problem while processing machine learning algorithms.

Noisy data, incomplete data, inaccurate data, and unclean data lead to less accuracy in classification and low-quality results.

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

For categorical data, typically only **graphical and descriptive** methods are used.

6. **How would the learning activity be affected if certain variables have missing values?**

Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions.

Having said that, what can be done about it?

One can :
Delete or impute them.

7. Describe the various methods for dealing with missing data values in depth.

When dealing with missing data, data scientists can use two primary methods to solve the error: **imputation or the removal of data**. The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low.

8. What are the various data pre-processing techniques?

Important Data Preprocessing Techniques are :Data Cleaning , Dimensionality Reduction,Feature Engineering, Sampling Data,Data Transformation, Imbalanced Data.

Explain dimensionality reduction and function selection in a few words.

Dimensionality reduction is a machine learning (ML) or statistical technique of reducing the amount of random variables in a problem by obtaining a set of principal variables.

Function selection is the process to remove the irrelevant data improves learning accuracy

9.

i. **What is the IQR?**
Interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data.

What criteria are used to assess it?
The interquartile range is calculated in much the same way as the range. All you do to find it is subtract the first quartile from the third quartile:
IQR = Q_3 – Q_1. The interquartile range shows how the data is spread about the median

ii. Describe the various components of a box plot in detail?

The median (middle quartile) marks the mid-point of the data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less. The middle "box" represents the middle 50% of scores for the group.

When will the lower whisker surpass the upper whisker in length?

How can box plots be used to identify outliers?
Each whisker extends to the furthest data point in each wing that is **within 1.5 times the IQR**. Any data point further than that distance is considered an outlier, and is marked with a dot.

10. Make brief notes on any two of the following:

1. **Data collected at regular intervals**
An example of interval data is the data collected on a thermometer—its gradation or markings are equidistant.

2. The gap between the quartiles

The **interquartile range** is the difference between upper and lower quartiles. The semi-interquartile range is half the interquartile range. When the data set is small, it is simple to identify the values of quartiles.

3. **Use a cross-tab**

With cross tabulation, people do not need statistical programming to correlate categorical variables.

11 - Make a comparison between:

1. Data with nominal and ordinal values

Nominal data is classified without a natural order or rank, whereas ordinal data has a predetermined or natural order.

2. Histogram and box plot

Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data

3. The average and median

The average is calculated by adding up all of the individual values and dividing this total by the number of observations whereas the median is calculated by taking the “middle” value, the value for which half of the observations are larger and half are smaller.