1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Data preprocessing is a process of **preparing the raw data and making it suitable for a machine learning model**. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two.

3. 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?

**Here are five common machine learning problems and how you can overcome them.**

* 1) Understanding Which Processes Need Automation. ...
* 2) Lack of Quality Data. ...
* 3) Inadequate Infrastructure. ...
* 4) Implementation. ...
* 5) Lack of Skilled Resources.

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

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

that, what can be done about it?

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

Deleting Rows with missing values

Impute missing values for continuous variable

Impute missing values for categorical variable

Other Imputation Methods

Using Algorithms that support missing values

Prediction of missing values

Imputation using Deep Learning Library — Datawig

8. What are the various data pre-processing techniques? Explain dimensionality reduction and

function selection in a few words.

Data Preprocessing can be done in four different ways. **Data cleaning/cleaning, data integration, data transformation, and data reduction** are the four categories.

Dimensionality reduction technique can be defined as, "**It is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information**."

9. i.What is the IQR? What criteria are used to assess it?

The IQR **describes the middle 50% of values when ordered from lowest to highest**. To find the interquartile range (IQR), ​first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

We can use the IQR method of identifying outliers **to set up a “fence” outside of Q1 and Q3**. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3.

ii. Describe the various components of a box plot in detail? When will the lower whisker

surpass the upper whisker in length? How can box plots be used to identify outliers?

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

1. Data collected at regular intervals

1. The gap between the quartiles

A quartile is a [statistical](https://www.investopedia.com/terms/s/statistics.asp) term that describes a division of observations into four defined intervals based on the values of the data and how they compare to the entire set of observations.

* The quartiles data into three points—a lower quartile, median, and upper quartile—to form four groups of the dataset.
* Along with the minimum and maximum values of the data set, the quartiles divide a set of observations into four sections, each representing 25% of the observations.
* Quartiles are used to calculate the interquartile range, which is a measure of variability around the median.

3. Use a cross-tab

1. Make a comparison between:

1. Data with nominal and ordinal values

2. Histogram and box plot

3. The average and median