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

Key tasks may be include the below:-

- Data gathering.
- Data pre-processing.
- Exploratory data analysis (EDA)
- Feature engineering.
- Training machine learning models of the following kinds: Regression. Classification. Clustering.
- Multivariate querying.
- Density estimation.
- Dimensionality reduction.

Steps in Data Preprocessing in Machine Learning

1. Acquire the dataset

2. Import all the crucial libraries

3. Import the dataset

4. Identifying and handling the missing values

5. Encoding the categorical data

6. Splitting the dataset

7. Feature scaling

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

Qualitative analysis	Quantitative Analysis
Subjective analysis that is more concerned with statistical data that cannot be computed.	Object analysis that qualifies data
Data include measurable quantities such as gender, color, nationality etc	Data include measurable quantities such as length, size, weight etc
sample is small and representative of entire population	sample is large and can be generalized to cover entire population
research method is exploratory	research method is often conclusive.

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?

1) Understanding Which Processes Need Automation

It's becoming increasingly difficult to separate fact from fiction in terms of Machine Learning today. Before you decide on which AI platform to use, you need to evaluate which problems you're seeking

to solve. The easiest processes to automate are the ones that are done manually every day with no variable output. Complicated processes require further inspection before automation.

2) Lack of Quality Data

The number one problem facing Machine Learning is the lack of good data. While enhancing algorithms often consumes most of the time of developers in AI, data quality is essential for the algorithms to function as intended. Noisy data, dirty data, and incomplete data are the quintessential enemies of ideal Machine Learning.

3) Inadequate Infrastructure

Machine Learning requires vast amounts of data churning capabilities. Legacy systems often can't handle the workload and buckle under pressure. You should check if your infrastructure can handle Machine Learning. If it can't, you should look to upgrade, complete with hardware acceleration and flexible storage.

4) Implementation

Organizations often have analytics engines working with them by the time they choose to upgrade to Machine Learning. Integrating newer Machine Learning methodologies into existing methodologies is a complicated task. Maintaining proper interpretation and documentation goes a long way to easing implementation.

5) Lack of Skilled Resources

Deep analytics and Machine Learning in their current forms are still new technologies. Thus, there is a shortage of skilled employees available to manage and develop analytical content for Machine Learning. Data scientists often need a combination of domain experience as well as in-depth knowledge of science, technology, and mathematics.

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

1. Unique value count - One of the first things which can be useful during data exploration is to see how many unique values are there in categorical columns. This gives an idea of what is the data about.

2. Frequency Count - Frequency count is finding how frequent individual values occur in column.

3. Variance - When it comes to analysing numeric values, some basic information such as minimum, maximum and variance are very useful. Variance gives a good indication how the values are spread.

4. Pareto Analysis - Pareto analysis is a creative way of focusing on what is important. Pareto 80–20 rule can be effectively used in data exploration.

5. Histogram - Histogram are one of the data scientists favourite data exploration techniques. It gives information on the range of values in which most of the values fall. It also gives information on whether there is any skew in data.

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.

- Listwise or case deletion. ...
- Pairwise deletion.
- Mean substitution.
- Regression imputation.
- Last observation carried forward.
- Multiple imputation.

8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

1. Data Cleaning/Cleansing - Data Cleaning/Cleansing routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. “Dirty” data can cause confusion for the mining procedure. Although most mining routines have some procedures, they deal incomplete or noisy data, which are not always robust.

2. Data Integration - Data Integration is involved in data analysis task which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. The issue to be considered in Data Integration is schema integration.

3. Data Transformation - Data are transformed into appropriate forms of mining. Data Transformation involves the following:

In Normalisation, where the attribute data are scaled to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0.

Smoothing works to remove the noise from the data. Such techniques include binning, clustering, and regression.

4. Data Reduction - Complex data analysis and mining on huge amounts of data may take a very long time, making such analysis impractical or infeasible. Data Reduction techniques are helpful in analysing the reduced representation of the data set without compromising the integrity of the original data and yet producing the qualitative knowledge

9.

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

The interquartile range is a measure of where the “middle fifty” is in a data set. Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie.

The IQR is used to measure how spread out the data points in a set are from the mean of the data set. The higher the IQR, the more spread out the data points; in contrast, the smaller the IQR, the more bunched up the data points are around the mean.

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

2. The gap between the quartiles

3. Use a cross-tab

Cross tabulation is a method to quantitatively analyse the relationship between multiple variables. It also shows how correlations change from one variable grouping to another. It is usually used in statistical analysis to find patterns, trends, and probabilities within raw data.

1. Make a comparison between:

1. Data with nominal and ordinal values

A purely nominal variable is one that simply allows you to assign categories but you cannot clearly order the categories. An ordinal variable is similar to a categorical variable. The difference between the two is that there is a clear ordering of the categories. For example, suppose you have a variable, economic status, with three categories (low, medium and high). In addition to being able to classify people into these three categories, you can order the categories as low, medium and high.

2. Histogram and box plot

A histogram is a type of bar chart that graphically displays the frequencies of a data set. Similar to a bar chart, a histogram plots the frequency, or raw count, on the Y-axis (vertical) and the variable being measured on the X-axis (horizontal).

A box plot, also called a box-and-whisker plot, is a chart that graphically represents the five most important descriptive values for a data set. These values include the minimum value, the first quartile, the median, the third quartile, and the maximum value.

2. The average and median
   A mean is a mathematical term, that describes the average of a sample. A mean can be defined as an average of the set of values in a sample of data. An Average can be defined as the sum of all numbers divided by the total number of values.