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

Ans : 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.

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.

Ans :

Quantitative data refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it’s quantitative data. Quantitative data can tell you “how many,” “how much,” or “how often”—for example, how many people attended last week’s webinar? How much revenue did the company make in 2019? How often does a certain customer group use online banking? To analyze and make sense of quantitative data, you’ll conduct statistical analyses.

Unlike quantitative data, qualitative data cannot be measured or counted. It’s descriptive, expressed in terms of language rather than numerical values.

Researchers will often turn to qualitative data to answer “Why?” or “How?” questions. For example, if your quantitative data tells you that a certain website visitor abandoned their shopping cart three times in one week, you’d probably want to investigate why—and this might involve collecting some form of qualitative data from the user. Perhaps you want to know how a user feels about a particular product; again, qualitative data can provide such insights. In this case, you’re not just looking at numbers; you’re asking the user to tell you, using language, why they did something or how they feel.

Qualitative data also refers to the words or labels used to describe certain characteristics or traits—for example, describing the sky as blue or labeling a particular ice cream flavor as vanilla.

* Quantitative data is countable or measurable, relating to numbers. Qualitative data is descriptive, relating to language.
* Quantitative data tells us how many, how much, or how often (e.g. “20 people signed up to our email newsletter last week”). Qualitative data can help us to understand the “why” or “how” behind certain behaviors, or it can simply describe a certain attribute—for example, “The postbox is red” or “I signed up to the email newsletter because I’m really interested in hearing about local events.”
* Quantitative data is fixed and “universal,” while qualitative data is subjective and dynamic. For example, if something weighs 20 kilograms, that can be considered an objective fact. However, two people may have very different qualitative accounts of how they experience a particular event.
* Quantitative data is gathered by measuring and counting. Qualitative data is collected by interviewing and observing.
* Quantitative data is analyzed using statistical analysis, while qualitative data is analyzed by grouping it in terms of meaningful categories or themes.

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

Ans :

| **Country** | **Age** | **Salary** | **Purchased** |
| --- | --- | --- | --- |
| India | 38 | 48000 | No |
| France | 43 | 45000 | Yes |
| Germany | 30 | 54000 | No |
| France | 48 | 65000 | No |
| Germany | 40 | 70000 | Yes |
| India | 35 | 58000 | Ye |

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

Ans :

One of the significant issues that machine learning professionals face is **the absence of good quality data**. Unclean and noisy data can make the whole process extremely exhausting. We don't want our algorithm to make inaccurate or faulty predictions. Hence the quality of data is essential to enhance the output.

The ramifications of **a decision, plan, or event are all its consequences and effects**, especially ones which are not obvious at first.

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?

Ans :

Even in a well-designed and controlled study, missing data occurs in almost all research. **Missing data can reduce the statistical power of a study and can produce biased estimates**, leading to invalid conclusions.

Handle missing values by :

1. Delete the observations

2. Replacing them with mean/mode/median as per data.

3. Using classifier models to predict them like KNN.

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

Ans : Imputation vs. Removing Data. 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.

1. Replacing them with mean/mode/median as per data.
2. Replacing them with a constant say -1.
3. Using classifier models to predict them like KNN.

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

Ans :

## 1. Data Cleaning/Cleansing

Real-world data tend to be incomplete, noisy, and inconsistent. Data Cleaning/Cleansing routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

## 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. It is tricky.

How can real-world entities from multiple data sources be ‘matched up’? This is referred as entity identification problem. For example, how can a data analyst be sure that customer\_id in one database and cust\_number in another refer to the same entity? The answer is metadata. Databases and data warehouses typically have metadata. Simply, metadata is data about data.

## 3. Data Transformation

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

1. 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.
2. Smoothing works to remove the noise from the data. Such techniques include binning, clustering, and regression.
3. In Aggregation, summary or aggregation operations are applied to the data. For example, daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.
4. In Generalisation of the Data, low level or primitive/raw data are replaced by higher level concepts through the use of concept hierarchies. For example, categorical attributes are generalised to higher level concepts street into city or country. Similarly, the values for numeric attributes may be mapped to higher level concepts like, age into young, middle-aged, or senior.

## 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?

Ans :

The Interquartile range formula helps in finding the difference between the third quartile and the first quartile. The Interquartile range formula measures the variability, based on dividing an ordered set of data into quartiles. Quartiles are three values or cuts that divide each respective part as the first, second, and third quartiles, denoted by Q1, Q2, Q3 respectively.

Q1 consists of the first 25 % quantile.

Q2 consists of the first 50 % quantile.

Q3 consists of the first 75 % quantile.

### IQR Formula

Interquartile range(IQR) = Upper Quartile – Lower Quartile

Upper Quartile = Q3 + 1.5\*IQR

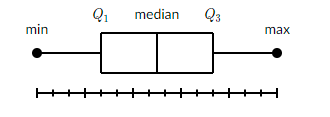
Lower Quartile = Q1 - 1.5\*IQR

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?

Ans :

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.

In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.



So from the above Box plot we can easily detect the outlier. Q1 and Q3 contain a major amount of data so here from the graph itself we can get to know that outliers are present or not.

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

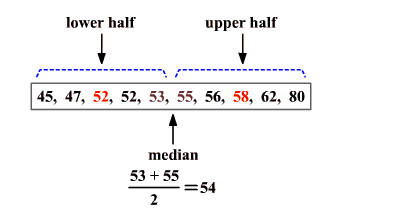
1. Data collected at regular intervals

2. The gap between the quartiles

Ans :

Interquartile, Semi-Interquartile and Mid-quartile Ranges

In a set of data, the quartiles are the values that divide the data into four equal parts. The median of a set of data separates the set in half.



The median of the lower half of a set of data is the lower quartile ( LQ ) or Q1 .

The median of the upper half of a set of data is the upper quartile ( UQ ) or Q3 .

The upper and lower quartiles can be used to find another measure of variation call the interquartile range .

The interquartile range or IQR is the range of the middle half of a set of data. It is the difference between the upper quartile and the lower quartile.

Interquartile range = Q3−Q1

In the above example, the lower quartile is 52 and the upper quartile is 58 .

The interquartile range is 58−52 or 6 .

Data that is more than 1.5 times the value of the interquartile range beyond the quartiles are called outliers .

Statisticians sometimes also use the terms semi-interquartile range and mid-quartile range .

The semi-interquartile range is one-half the difference between the first and third quartiles. It is half the distance needed to cover half the scores. The semi-interquartile range is affected very little by extreme scores. This makes it a good measure of spread for skewed distributions. It is obtained by evaluating Q3−Q12 .

The mid-quartile range is the numerical value midway between the first and third quartile. It is one-half the sum of the first and third quartiles. It is obtained by evaluating Q3+Q12 .

3. Use a cross-tab

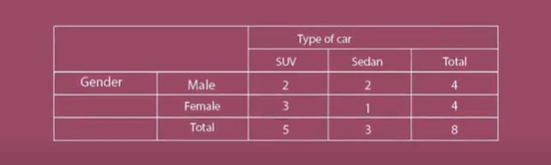
Ans : Cross tabulation is used to find Quantitative Analysis of relationships between multiple variables in a table. It is also known as contingency table or crosstab.

Cross tabulation shows how correlation changes from one variable grouping to another; this is used in Statistical analysis to find patterns and trends within raw data.

eg : We conduct Survey on Car type with respect to Gender.

| Gender | Car Type |
| --- | --- |
| M | SUV |
| F | Sedan |
| M | SUV |
| F | SUV |
| M | Sedan |
| F | SUV |
| M | Sedan |
| M | Sedan |

This is very difficult to analyse the data, so we use cross tab table :



1. Make a comparison between:

1. Data with nominal and ordinal values

Ans :

## Nominal Data

Nominal data simply names something without assigning it to an order in relation to other numbered objects or pieces of data. An example of nominal data might be a "pass" or "fail" classification for each student's test result. Nominal data provides some information about a group or set of events, even if that information is limited to mere counts.

For example, if you want to know how many people were born in Florida each year for the past five years, find those figures and plot your results on a bar graph. The data represented on the graph have no natural ranking or ordering; the numbers simply illustrate a fact, not necessarily a preference, and are just labels that answer the question "how many?" These are nominal data.

## Ordinal Data

Ordinal data, unlike nominal data, involves some order; ordinal numbers stand in relation to each other in a ranked fashion. For example, suppose you receive a survey from your favorite restaurant that asks you to provide feedback on the service you received. You can rank the quality of service as "1" for poor, "2" for below average, "3" for average, "4" for very good and "5" for excellent. The data collected by this survey are examples of ordinal data. Here the numbers assigned have an order or rank; that is, a ranking of "4” is better than a ranking of “2.”

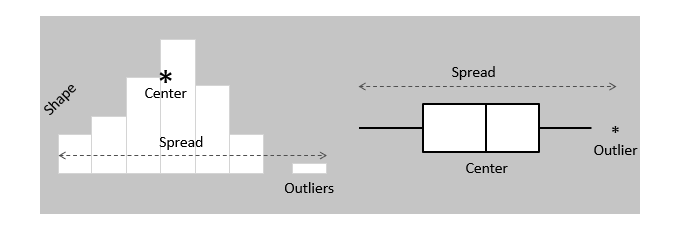
However, even though you have assigned a number to your opinion, this number is not a quantitative measure: Although a ranking of “4” is clearly better than a ranking of “2,” it is not necessarily twice as good. The numbers are not mathematically measured or determined but are merely assigned as labels for opinions.

2. Histogram and box plot

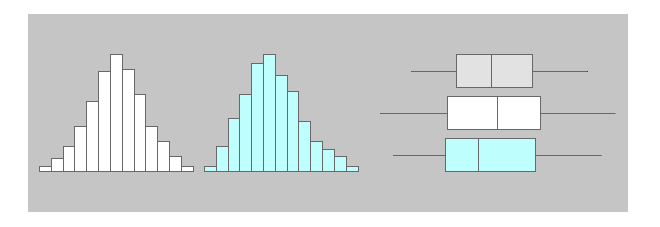
Ans :

Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.

Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.



Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.



3. The average and median

Ans : Average and median are both measures of “central tendency,” in that they are intended to provide some indication of a typical or middle value of a set of data. The average is calculated by adding up all of the individual values and dividing this total by the number of observations. The median is calculated by taking the “middle” value, the value for which half of the observations are larger and half are smaller.

When there is a possibility of extreme values, the median is generally the better measure to use. To see this, suppose that five homes sold in a market with the following prices: $80,000, $90,000, $100,000, $110,000 and $500,000. The median price is $100,000, while the average price is (80,000 + 90,000 + 100,000 + 110,000 + 500,000) / 5 = $176,000. In this instance, the single high-priced home pulled up the average price well above the prices of the more typical homes in the market. Thus, the median price provides a better measure of the typical value of a home.