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

Answer : Get Data. The first step in the Machine Learning process is getting data. Clean, Prepare & Manipulate Data.

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

Answer : Different Types Of Data Types
The Data Type Is Broadly Classified Into

1. Quantitative
2. Qualitative

1. Quantitative Data Type: –
This Type Of Data Type Consists Of Numerical Values. Anything Which Is Measured By Numbers.

E.G., Profit, Quantity Sold, Height, Weight, Temperature, Etc.

This Is Again Of Two Types

A.) Discrete Data Type: –
The Numeric Data Which Have Discrete Values Or Whole Numbers. This Type Of Variable Value If Expressed In Decimal Format Will Have No Proper Meaning. Their Values Can Be Counted.

E.G.: – No. Of Cars You Have, No. Of Marbles In Containers, Students In A Class, Etc.

B.) Continuous Data Type: –
The Numerical Measures Which Can Take The Value Within A Certain Range. This Type Of Variable Value If Expressed In Decimal Format Has True Meaning. Their Values Can Not Be Counted But Measured. The Value Can Be Infinite

E.G.: – Height, Weight, Time, Area, Distance, Measurement Of Rainfall, Etc.

2. Qualitative Data Type: –
These Are The Data Types That Cannot Be Expressed In Numbers. This Describes Categories Or Groups And Is Hence Known As The Categorical Data Type.

This Can Be Divided Into:-
A. Structured Data:
This Type Of Data Is Either Number Or Words. This Can Take Numerical Values But Mathematical Operations Cannot Be Performed On It. This Type Of Data Is Expressed In Tabular Format.

E.G.) Sunny=1, Cloudy=2, Windy=3 Or Binary Form Data Like 0 Or1, Good Or Bad, Etc.

B. Unstructured Data:
This Type Of Data Does Not Have The Proper Format And Therefore Known As Unstructured Data.This Comprises Textual Data, Sounds, Images, Videos, Etc.

Besides This, There Are Also Other Types Refer As Data Types Preliminaries Or Data Measures:-

1. Nominal
2. Ordinal
3. Interval
4. Ratio

These Can Also Be Refer Different Scales Of Measurements.

I. Nominal Data Type:
This Is In Use To Express Names Or Labels Which Are Not Order Or Measurable.

E.G., Male Or Female (Gender), Race, Country, Etc.

II. Ordinal Data Type:
This Is Also A Categorical Data Type Like Nominal Data But Has Some Natural Ordering Associated With It.

E.G., Likert Rating Scale, Shirt Sizes, Ranks, Grades, Etc.

III. Interval Data Type:
This Is Numeric Data Which Has Proper Order And The Exact Zero Means The True Absence Of A Value Attached. Here Zero Means Not A Complete Absence But Has Some Value. This Is The Local Scale.

E.G., Temperature Measured In Degree Celsius, Time, Sat Score, Credit Score, PH, Etc. Difference Between Values Is Familiar. In This Case, There Is No Absolute Zero. Absolute

IV. Ratio Data Type:
This Quantitative Data Type Is The Same As The Interval Data Type But Has The Absolute Zero. Here Zero Means Complete Absence And The Scale Starts From Zero. This Is The Global Scale.

E.G., Temperature In Kelvin, Height, Weight, Etc.

3. Distinguish:

1. Numeric vs. categorical attributes

2. Feature selection vs. dimensionality reduction

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

Numeric attributes consists Of Numerical Values, anything Which Is Measured By Numbers.

Categorical Attributes : These Are The Data Types That Cannot Be Expressed In Numbers. This Describes Categories Or Groups And Is Hence Known As The Categorical Data Type.

2. Feature selection vs. dimensionality reduction

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

2. Use a scatter plot

3.PCA (Personal Computer Aid)

Answer : **The histogram :**

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

2. Use a scatter plot

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

3.PCA (Personal Computer Aid):

Principal Component Analysis or PCA is a widely used technique for dimensionality reduction of the large data set. Reducing the number of components or features costs some accuracy and on the other hand, it makes the large data set simpler, easy to explore and visualize.

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

Answer : If your data set is messy, building models will not help you to solve your problem. What will happen is “garbage in, garbage out.” In order to build a powerful machine learning algorithm. We need to explore and understand our data set before we define a predictive task and solve it.

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

Answer : Various Histogram Shape :

- Skewed Distribution. The skewed distribution is asymmetrical because a natural limit prevents outcomes on one side. ...
- Double-Peaked or Bimodal. ...
- Plateau or Multimodal Distribution. ...
- Edge Peak Distribution. ...
- Comb Distribution. ...
- Truncated or Heart-Cut Distribution. ...
- Dog Food Distribution

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?

Answer : **Z-Score:**

This can be done with just one line code as we have already calculated the Z-score.

```
boston_df_o = boston_df_o[(z < 3).all(axis=1)]
```

IQR Score –

Calculate IQR score to filter out the outliers by keeping only valid values.

```
boston_df_out = boston_df_o1[~((boston_df_o1 < (Q1 - 1.5 * IQR)) |(boston_df_o1 > (Q3 + 1.5 * IQR))).any(axis=1)]boston_df_out.shape
```

Quantile function :

Use quantile() to remove amount of data.

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

Answer : Mean , median and mode are central inclination measure.

Mean varies more than median due to presence of outliers, as mean is averaging all points while median in like finding a middle number.

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

Answer : A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. So this visualization gives us the idea of bivariate relationship.

Scatter plot can also help finding outliers as outliers can be visualized at farther distance than regular data.

10. Describe how cross-tabs can be used to figure out how two variables are related.

Answer : Cross tabulation is a method to quantitatively analyze the relationship between multiple variables. Also known as contingency tables or cross tabs, cross tabulation groups variables to understand the correlation between different variables. It also shows how correlations change from one variable grouping to another.