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

* All the Machine Learning algorithms take data as input, but what they want to achieve is different. They can broadly be classified in a few groups based on the task they are designed to solve.
* These tasks are classification, regression, and clustering.

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

* Almost anything can be turned into DATA.
* Building a deep understanding of the different data types is a crucial prerequisite for doing Exploratory Data Analysis (EDA) and Feature Engineering for Machine Learning models. You also need to convert data types of some variables in order to make appropriate choices for visual encodings in data visualization and storytelling.
* Most data can be categorized into 4 basic types from a Machine Learning perspective: numerical data, categorical data, time-series data, and text.
* Numerical Data
* Numerical data is any data where data points are exact numbers. Statisticians also might call numerical data, quantitative data. This data has meaning as a measurement such as house prices or as a count, such as a number of residential properties in Los Angeles or how many houses sold in the past year.
* Numerical data can be characterized by continuous or discrete data. Continuous data can assume any value within a range whereas discrete data has distinct values.
* Numerical Data
* For example, the number of students taking Python class would be a discrete data set. You can only have discrete whole number values like 10, 25, or 33. A class cannot have 12.75 students enrolled. A student either join a class or he doesn’t. On the other hand, continuous data are numbers that can fall anywhere within a range. Like a student could have an average score of 88.25 which falls between 0 and 100.
* The takeaway here is that numerical data is not ordered in time. They are just numbers that we have collected.
* Categorical Data
* Categorical data represents characteristics, such as a hockey player’s position, team, hometown. Categorical data can take numerical values. For example, maybe we would use 1 for the colour red and 2 for blue. But these numbers don’t have a mathematical meaning. That is, we can’t add them together or take the average.
* In the context of super classification, categorical data would be the class label. This would also be something like if a person is a man or woman, or property is residential or commercial.
* There is also something called ordinal data, which in some sense is a mix of numerical and categorical data.
* In ordinal data, the data still falls into categories, but those categories are ordered or ranked in some way.
* An example would be class difficulty, such as beginner, intermediate, and advanced. Those three types of classes would be a way that we could label the classes, and they have a natural order in increasing difficulty.
* Another example is that we just take quantitative data, and splitting it into groups, so we have bins or categories of other types of data Ordinal Data.
* For plotting purposes, ordinal data is treated much in the same way as categorical data.
* But groups are usually ordered from lowest to highest so that we can preserve this ordering.
* Time Series Data
* Time series data is a sequence of numbers collected at regular intervals over some period. It is very important, especially in particular fields like finance. Time series data has a temporal value attached to it, so this would be something like a date or a timestamp that you can look for trends in time.
* For example, we might measure the average number of home sales for many years.
* The difference of time series data and numerical data is that rather than having a bunch of numerical values that don’t have any time ordering, time-series data does have some implied ordering. There is a first data point collected and the last data point collected.
* Text
* Text data is basically just words. A lot of the time the first thing that you do with text is you turn it into numbers using some interesting functions like the bag of words formulation.

3. Distinguish:

1. Numeric vs. categorical attributes

* A categorical variable is a variable with a set number of groups (gender, colours of the rainbow, brands of cereal),
* while a numeric variable is generally something that can be measured (height, weight, miles per hour).
* Categorical data refers to a data type that can be stored and identified based on the names or labels given to them.
* Numerical data refers to the data that is in the form of numbers, and not in any language or descriptive form.
* Also known as qualitative data as it qualifies data before classifying it.

1. Feature selection vs. dimensionality reduction

* Feature Selection vs Dimensionality Reduction
* While both methods are used for reducing the number of features in a dataset, there is an important difference. Feature selection is simply selecting and excluding given features without changing them.
* Dimensionality reduction transforms features into a lower dimension.
* Feature Selection:
* Remove features with missing values
* Remove features with low variance
* Remove highly correlated features
* Univariate feature selection
* Recursive feature elimination
* Feature selection using SelectFromModel
* Dimensionality Reduction:
* PCA(Principal component analysis)

4.Make quick notes on any two of the following:

1. The histogram:

A histogram is a display of statistical information that uses rectangles to show the frequency of data items in successive numerical intervals of equal size. In the most common form of histogram,

the independent variable is plotted along the horizontal axis and the dependent variable is plotted along the vertical axis.

1. Use a scatter plot:

Scatter plots' primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

Identification of correlational relationships are common with scatter plots.

1. PCA (Personal Computer Aid)

PCA is a tool for identifying the main axes of variance within a data set and allows for easy data exploration to understand the key variables in the data and spot outliers.

Properly applied, it is one of the most powerful tools in the data analysis tool kit

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

* Simply put, quantitative data gets you the numbers to prove the broad general points of your research. Qualitative data brings you the details and the depth to understand their full implications.
* To get the best results from these methods in your surveys, it's important that you understand the differences between them.
* Using a combination of qualitative and quantitative data can improve an evaluation by ensuring that the limitations of one type of data are balanced by the strengths of another.
* This will ensure that understanding is improved by integrating different ways of knowing.

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

* A histogram is a chart that plots the distribution of a numeric variable's values as a series of bars. Each bar typically covers a range of numeric values called a bin or class.
* a bar's height indicates the frequency of data points with a value within the corresponding bin.
* Types of Histograms
* There are different types of distributions, such as normal distribution, skewed distribution, bimodal distribution, multimodal distribution, comb distribution, edge peak distribution, dog food distribution, heart cut distribution
* The towers or bars of a histogram are called bins. The height of each bin shows how many values from that data fall into that range. Width of each bin is = (max value of data – min value of data) / total number of bins.
* The default value of the number of bins to be created in a histogram is 10.

1. How do we deal with data outliers?

* 5 ways to deal with outliers in data
* Set up a filter in your testing tool. Even though this has a little cost, filtering out outliers is worth it.
* Remove or change outliers during post-test analysis.
* Change the value of outliers.
* Consider the underlying distribution.
* Consider the value of mild outliers.

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

* There are three main measures of central tendency: the mode, the median and the mean.
* Each of these measures describes a different indication of the typical or central value in the distribution.
* The 3 most common measures of central tendency are the mean, median and mode. The mode is the most frequent value. The median is the middle number in an ordered data set.
* The mean is the sum of all values divided by the total number of values.

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

* Scatter plots are a tool you can use to display bivariate data.
* If there is a regression line on a scatter plot, you can identify outliers. An outlier for a scatter plot is the point or points that are farthest from the regression line.
* There is at least one outlier on a scatter plot in most cases, and there is usually only one outlier.
* The distance of the points to the line is called "scatter". A large amount of scatter around the line indicates a weak relationship. Little scatter represents a strong relationship.
* If all points fall directly on a straight line, we have a perfect linear relationship between our two variables.
* The scatter plot is a fundamental tool for looking at bivariate data.
* It shows the important characteristics of the data and can be used to decide what model may describe the relationship between the variables.
* If there is a regression line on a scatter plot, you can identify outliers. An outlier for a scatter plot is the point or points that are farthest from the regression line. There is at least one outlier on a scatter plot in most cases, and there is usually only one outlier.

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

* To describe the relationship between two categorical variables, we use a special type of table called a cross-tabulation (or "crosstab" for short). In a cross-tabulation, the categories of one variable determine the rows of the table, and the categories of the other variable determine the columns.
* Cross tabulation is used to quantitatively analyse the relationship between multiple variables.
* Cross tabulations also referred to as contingency tables or crosstabs group variables together and enable researchers to understand the correlation between the different variables.
* Cross-tabulation analysis, also known as contingency table analysis, is most often used to analyse categorical (nominal measurement scale) data. At their core, cross-tabulations are simply data tables that present the results of the entire group of respondents, as well as results from subgroups of survey respondents.