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

Answer: 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.we will discuss the workflow of a Machine learning project this includes all the steps required to build the proper machine learning project from scratch.

We will also go over data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine Learning Model Performance. We will also cover a couple of the pre-modelling steps that can help to improve the model performance.

Python Libraries that would be need to achieve the task:

1. Numpy
2. Pandas
3. Sci-kit Learn
4. Matplotlib

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

Answer:Data can come in many forms, but machine learning models rely on four primary data types. These include numerical data, categorical data, time series data, and text data.

Numerical data

Numerical data, or quantitative data, is any form of measurable data such as your height, weight, or the cost of your phone bill. You can determine if a set of data is numerical by attempting to average out the numbers or sort them in ascending or descending order. Exact or whole numbers (ie. 26 students in a class) are considered discrete numbers, while those which fall into a given range (ie. 3.6 percent interest rate) are considered continuous numbers. While learning this type of data, keep in mind that numerical data is not tied to any specific point in time, they are simply raw numbers.

Categorical data

Categorical data is sorted by defining characteristics. This can include gender, social class, ethnicity, hometown, the industry you work in, or a variety of other labels. While learning this data type, keep in mind that it is non-numerical, meaning you are unable to add them together, average them out, or sort them in any chronological order. Categorical data is great for grouping individuals or ideas that share similar attributes, helping your machine learning model streamline its data analysis.

Time series data

Time series data consists of data points that are indexed at specific points in time. More often than not, this data is collected at consistent intervals. Learning and utilizing time series data makes it easy to compare data from week to week, month to month, year to year, or according to any other time-based metric you desire. The distinct difference between time series data and numerical data is that time series data has established starting and ending points, while numerical data is simply a collection of numbers that aren't rooted in particular time periods.

Text data

Text data is simply words, sentences, or paragraphs that can provide some level of insight to your machine learning models. Since these words can be difficult for models to interpret on their own, they are most often grouped together or analyzed using various methods such as word frequency, text classification, or sentiment analysis.

```
# 3. Distinguish:

1. Numeric vs. categorical attributes

2. Feature selection vs. dimensionality reduction


Answer: 1... A categorical variable is a category or type. For example, hair
color is a categorical value or hometown is a categorical variable. Species,
treatment type, and gender are all categorical variables.

A numerical variable is a variable where the measurement or number has a
numerical meaning. For example, total rainfall measured in inches is a
numerical value, heart rate is a numerical value, number of cheeseburgers
consumed in an hour is a numerical value.

A categorical variable can be expressed as a number for the purpose of
statistics, but these numbers do not have the same meaning as a numerical value
. For example, if I am studying the effects of three different medications on
an illness, I may name the three different medicines, medicine 1, medicine 2,
and medicine 3. However, medicine three is not greater, or stronger, or faster
than medicine one. These numbers are not meaningful.

2.Feature selection vs dimensionality reduction

Often, feature selection and dimensionality reduction are grouped together
(like here in this article). 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.

In this article we will explore the following feature selection and
dimensionality reduction techniques:

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

One shouldn't just throw everything at your machine learning model and rely on
your training process to determine which features are actually useful – I have
discussed this in my posts before. Thus, it is imperative to carry out feature
selection and | or dimensionality reduction to reduce the number of features in
a dataset. Whilst both ‘feature selection’ and ‘dimensionality reduction’ 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
Whereas Dimensionally Reduction transforms the features into a lower dimension
 Feature selection identifies the features that best represent the relationship
amongst all in the feature space as well as the target that the model will try
to predict. Feature selection methods remove the features that do not influence
the outcome. This reduces the size of the feature space, hence reducing the
resource requirements for processing the data and model complexity too.
```

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

1. The histogram
2. Use a scatter plot
3. PCA (Personal Computer Aid)

Anser 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. The data appears as colored or shaded rectangles of variable area.

The illustration, below, is a histogram showing the results of a final exam given to a hypothetical class of students. Each score range is denoted by a bar of a certain color. If this histogram were compared with those of classes from other years that received the same test from the same professor, conclusions might be drawn about intelligence changes among students over the years. Conclusions might also be drawn concerning the improvement or decline of the professor's teaching ability with the passage of time. If this histogram were compared with those of other classes in the same semester who had received the same final exam but who had taken the course from different professors, one might draw conclusions about the relative competence of the professors.A histogram is an approximate representation of the distribution of numerical data. The term was first introduced by Karl Pearson.[1] To construct a histogram, the first step is to "bin" (or "bucket") the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent and are often (but not required to be) of equal size.

If the bins are of equal size, a bar is drawn over the bin with height proportional to the frequency—the number of cases in each bin. A histogram may also be normalized to display "relative" frequencies showing the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1.

However, bins need not be of equal width; in that case, the erected rectangle is defined to have its area proportional to the frequency of cases in the bin.[3] The vertical axis is then not the frequency but frequency density—the number of cases per unit of the variable on the horizontal axis. Examples of variable bin width are displayed on Census bureau data below.

2. Use a scatter Plot....Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another. Each row in the data table is represented by a marker whose position depends on its values in the columns set on the X and Y axes.

A third variable can be set to correspond to the color or size of the markers, thus adding yet another dimension to the plot.

The relationship between two variables is called their correlation. If the markers are close to making a straight line in the scatter plot, the two variables have a high correlation. If the markers are equally distributed in the scatter plot, the correlation is low, or zero. However, even though a correlation may seem to be present, this might not always be the case. Both variables could be related to some third variable, thus explaining their variation, or, pure coincidence might cause an apparent correlation.Scatter plots are used to show relationships. For correlation, scatter plots help show the strength of the linear relationship between two variables. For regression, scatter plots often add a fitted line. In quality control, scatter plots can often include specification limits or reference lines.

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

Answer: This helps them to build accurate models and check assumptions required for fitting models. Create meaningful data visualizations, predict future trends from the data. If you are good at understanding data preparation almost 80% of the work is completed.Real statistical data investigations involve a number of components: formulating a problem so that it can be tackled statistically; planning, collecting, organising and validating data; exploring and analysing data; and interpreting and presenting information from data in context.

Qualitative data analysis works a little differently from quantitative data, primarily because qualitative data is made up of words, observations, images, and even symbols. Deriving absolute meaning from such data is nearly impossible; hence, it is mostly used for exploratory research.Both qualitative and quantitative research methods have their flaws. However, it is imperative to note that quantitative research method deals with a larger population and quantifiable data and will, therefore, produce a more reliable result than qualitative research.

6. What are the various histogram shapes? What

exactly are ‘bins'?

Answer: 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. Bell-shaped: A bell-shaped picture, shown below, usually presents a normal distribution.

Bimodal: A bimodal shape, shown below, has two peaks. ...

Skewed left: Some histograms will show a skewed distribution to the left, as shown below.

Skewed left: Some histograms will show a skewed distribution to the left, as shown below. A distribution skewed to the left is said to be negatively skewed. This kind of distribution has a large number of occurrences in the upper value cells (right side) and few in the lower value cells (left side). A skewed distribution can result when data is gathered from a system with a boundary such as 100. In other words, all the collected data has values less than 100

Uniform: A uniform distribution, as shown below, provides little information about the system. An example would be a state lottery, in which each class has about the same number of elements. It may describe a distribution which has several modes (peaks). If your histogram has this shape, check to see if several sources of variation have been combined. If so, analyze them separately. If multiple sources of variation do not seem to be the cause of this pattern, different groupings can be tried to see if a more useful pattern results. This could be as simple as changing the starting and ending points of the cells, or changing the number of cells. A uniform distribution often means that the number of classes is too small..

Random: A random distribution, as shown below, has no apparent pattern. Like the uniform distribution, it may describe a distribution that has several modes (peaks). If your histogram has this shape, check to see if several sources of variation have been combined. If so, analyze them separately. If multiple sources of variation do not seem to be the cause of this pattern, different groupings can be tried to see if a more useful pattern results. This could be as simple as changing the starting and ending points of the cells, or changing the number of cells. A random distribution often means there are too many classes

7. How do we deal with data outliers?

Answer: We all have heard of the idiom ‘odd one out which means something unusual in comparison to the others in a group.

Similarly, an Outlier is an observation in a given dataset that lies far from the rest of the observations. That means an outlier is vastly larger or smaller than the remaining values in the set.In statistics, we have three measures of central tendency namely Mean, Median, and Mode. They help us describe the data.

Mean is the accurate measure to describe the data when we do not have any outliers present.

Median is used if there is an outlier in the dataset.

Mode is used if there is an outlier AND about ½ or more of the data is the same.

'Mean' is the only measure of central tendency that is affected by the outliers which in turn impacts Standard deviation.If our dataset is small, we can detect the outlier by just looking at the dataset. But what if we have a huge dataset, how do we identify the outliers then? We need to use visualization and mathematical techniques.

Below are some of the techniques of detecting outliers

Boxplots Z-score

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.

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

Answer: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 median is the most informative measure of central tendency for skewed distributions or distributions with outliers. For example, the median is often used as a measure of central tendency for income distributions, which are generally highly skewed.Mean is the average value of set of given data and median is the middle value when the data set is arranged in an order either ascending or descending.

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 scatterplot shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point on the graph.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.Scatter plots are a tool you can use to display bivariate data. You can make coordinate pairs using the data.

Yes, 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.

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. 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.A cross tabulation (or crosstab) report is used to analyze the relationship between two or more variables. The report has the x-axis as one variable (or question) and the y-axis as another variable. This type of analysis is crucial in finding underlying relationships within your survey results.

In [ ]: