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

- Step 1: Collect Data.
- Step 2: Prepare the data.
- Step 3: Choose the model.
- Step 4 Train your machine model.
- Step 5: Evaluation.
- Step 6: Parameter Tuning.
- Step 7: Prediction or Inference.

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

Most data can be categorized into 4 basic types from a Machine Learning perspective: numerical data, categorical data, time-series data, and text.

3. Distinguish:

1. Numeric vs. categorical attributes

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.

2. 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

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

1. 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

Use a scatter plot to determine whether or not two variables have a relationship or correlation. Are you trying to see if your two variables might mean something when put together? Plotting a

scattergram with your data points can help you to determine whether there's a potential relationship between them.

3. PCA (Personal Computer Aid)

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

Data analysis is important in research because it makes studying data a lot simpler and more accurate. It helps the researchers straightforwardly interpret the data so that researchers don't leave anything out that could help them derive insights from it.

Qualitative and differ in their approach and the type of data they collect.Quantitative data refers to any information that can be quantified — that is, numbers. If it can be counted or measured, and given a numerical value, it's quantitative in nature. Think of it as a measuring stick.Quantitative variables can tell you "how many," "how much," or "how often." Some examples of quantitative data:

- How many people attended last week's webinar?
- How much revenue did our company make last year?
- How often does a customer rage click on this app?

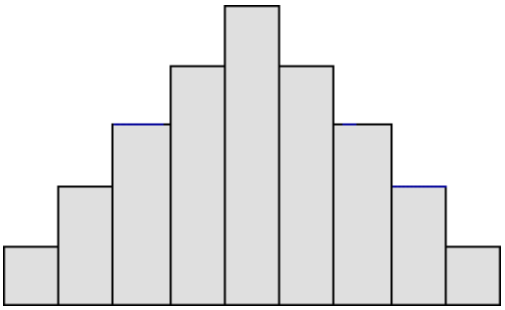
To analyze these research questions and make sense of this quantitative data, you'd normally use a form of statistical analysis—collecting, evaluating, and presenting large amounts of data to discover patterns and trends. Quantitative data is conducive to this type of analysis because it's numeric and easier to analyze mathematically.

Quantitative data is all about the numbers. Quantitative research is based on the collection and interpretation of numeric data. It focuses on measuring (using inferential statistics) and generalizing results.

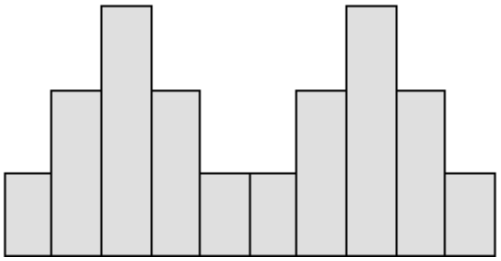
In terms of digital experience data, it puts everything in terms of numbers (or discrete data)—like the number of users clicking a button, bounce rates, time on site, and more.

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

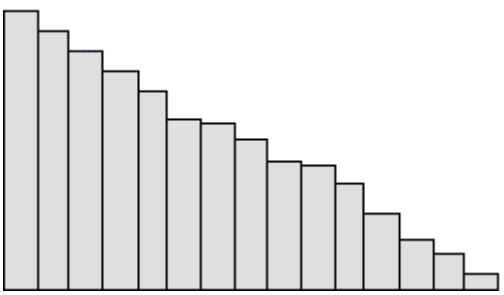
Bell-shaped: A bell-shaped picture, shown below, usually presents a normal distribution.



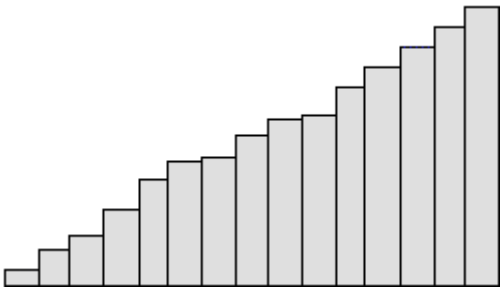
Bimodal: A bimodal shape, shown below, has two peaks. This shape may show that the data has come from two different systems. If this shape occurs, the two sources should be separated and analyzed separately.



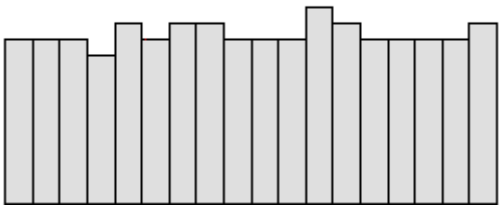
Skewed right: Some histograms will show a skewed distribution to the right, as shown below. A distribution skewed to the right is said to be positively skewed. This kind of distribution has a large number of occurrences in the lower value cells (left side) and few in the upper value cells (right side).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



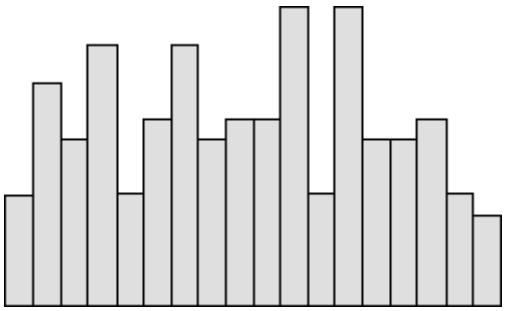
Skewed left: Some histograms will show a skewed distribution to the left, as shown below. A distribution skewed to the left is said to be negatively skewed. This kind of distribution has a large number of occurrences in the upper value cells (right side) and few in the lower value cells (left side). 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?

- 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 the median in certain data sets?

Mean, median, and mode are the most important measures of central tendency.

Mean is simple to use and can be applied to any data array set, whether even or odd. Median is slightly complex to use, and the data set needs to be arranged in the ascending or descending order first before calculation. The mean is generally used for normal distributions.

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

Scatterplots are useful for identifying relationships in bivariate data. In considering the relationship between two quantitative variables, we can sometimes identify one of the variables as the explanatory variable, or independent variable, and the other as the response variable, or dependent variable**.**

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.