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.**

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**.

**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.

1. 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 bar graph-like representation of data that buckets a range of classes into columns along the horizontal x-axis**. The vertical y-axis represents the number count or percentage of occurrences in the data for each column. Columns can be used to visualize patterns of data distributions.

1. 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)

Principal component analysis (PCA) is **a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss**. It does so by creating new uncorrelated variables that successively maximize variance.

Principal components analysis (PCA) is an ordination technique used primarily to display patterns in multivariate data. It aims to display the relative positions of data points in fewer dimensions while retaining as much information as possible, and explore relationships between dependent variables.

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

data are explored?

This **helps them to build accurate models and check assumptions required for fitting models**. Create meaningful data visualizations, predict future trends from the data.

**How to Analyze Data in 5 Steps**

Step 1: Define your goals.

Step 2: Decide how to measure goals.

Step 3: Collect your data.

Step 4: Analyze your data.

Step 5: Visualize and interpret results.

**Quantitative research deals with numbers and statistics, while qualitative research deals with words and meanings**. **Quantitative methods allow you to systematically measure variables and test hypotheses. Qualitative methods allow you to explore concepts and experiences in more detail.**

**Mixed methods are commonly used in health services research; however, data are not often integrated to explore complementarity of findings. A triangulation protocol is one approach to integrating such data. A retrospective triangulation protocol was carried out on mixed methods data collected as part of a process evaluation of a trial. The multi-country randomised controlled trial found that a web-based training in communication skills (including use of a patient booklet) and the use of a C-reactive protein (CRP) point-of-care test decreased antibiotic prescribing by general practitioners (GPs) for acute cough. The process evaluation investigated GPs’ and patients’ experiences of taking part in the trial.**

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

The histogram can be classified into different types based on the frequency distribution of the data.  
...  
**The different types of a histogram are:**

* Uniform histogram.
* Symmetric histogram.
* Bimodal histogram.
* Probability histogram.

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.

7. How do we deal with data outliers?

**5 ways to deal with outliers in data**

1. Set up a filter in your testing tool. Even though this has a little cost, filtering out outliers is worth it. ...
2. Remove or change outliers during post-test analysis. ...
3. Change the value of outliers. ...
4. Consider the underlying distribution. ...
5. 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?

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.

One of the basic tenets of statistics that every student learns in about the second week of intro stats is that **in a skewed distribution, the mean is closer to the tail in a skewed distribution**. So in a right skewed distribution (the tail points right on the number line), the mean is higher than the median.

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

outliers using a scatter plot?

We often see patterns or relationships in scatterplots. **When the y variable tends to increase as the x variable increases, we say there is a positive correlation between the variables**. When the y variable tends to decrease as the x variable increases, we say there is a negative correlation 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.

A scatter plot is **a graphical tool**. It has been designed to ensure that it provides a convenient view of the process to the manager at a single glance. The scatter plot studies the correlation between the important variables. When it studies the correlation between two variables, it is called a bivariate scatter plot.

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.**

By showing how correlations change from one group of variables to another, cross tabulation **allows for the identification of patterns, trends, and probabilities within data sets**. When it comes to analyzing survey response data, cross tabulation reports depict the relationship between two or more survey questions.