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

Ans: The key tasks involved ingetting ready to work with machine learning modeling are-

* Data Ingestion
* Data Cleaning
* Data Analysis
* Data Preprocessing
* Feature Engineering
* Feature Selection

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

Ans: In Machine Learning we mainly use Structured data which can be of two types –   
(i) Labled Data and (ii) Unlabled Data.

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

Ans: Numeric attributes are those columns that contains only numerical data with data type as Integer and Decimal values.  
 Where as Categorical features contain string values.

**2. Feature selection vs. dimensionality reduction**

Ans: In Feature Selection we find out which features are most useful to predict the target featrure and then we remove the less important or useless or irrelevant features. But in dimentionality reduction we dont drop any column directly, instead we transformsfeatures into a lower dimension.

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

**1. The histogram**

Ans: A histogram is a graph used to represent the frequency distribution of a few data points of one variable. Histograms often classify data into various “bins” or “range groups” and count how many data points belong to each of those bins.

**2. Use a scatter plot**

Ans: A scatter plot is a set of points plotted on a horizontal and vertical axes. Scatter plots are important in statistics because they can show the extent of correlation, if any, between the values of observed quantities or phenomena (called variables).

**3.PCA (Personal Computer Aid)**

Ans: Principal Component Analysis (PCA) is a way to bring out strong patterns from large and complex datasets. The essence of the data is captured in a few principal components, which themselves convey the most variation in the dataset. PCA reduces the number of dimensions without selecting or discarding them.

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

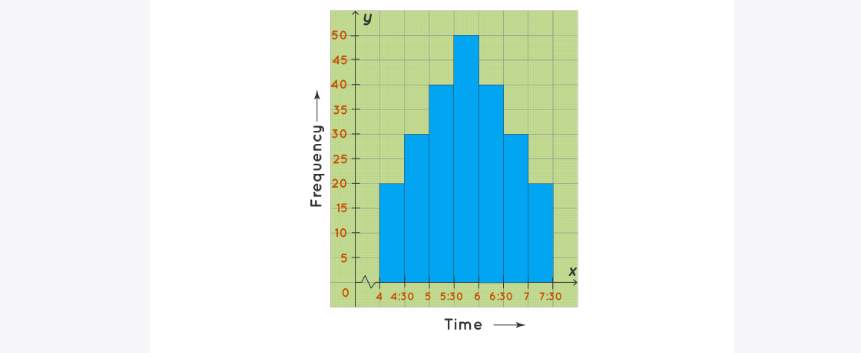
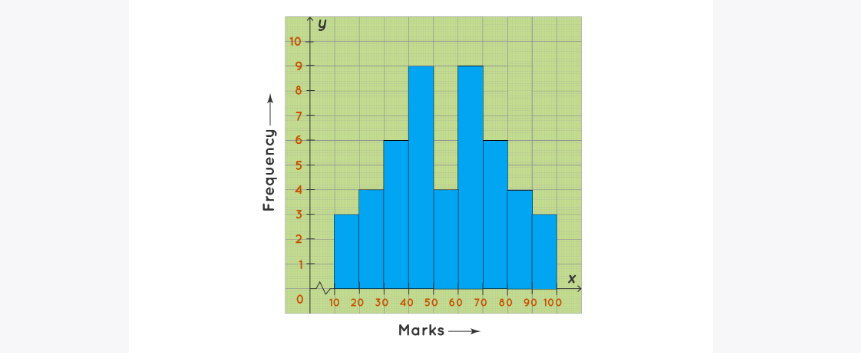
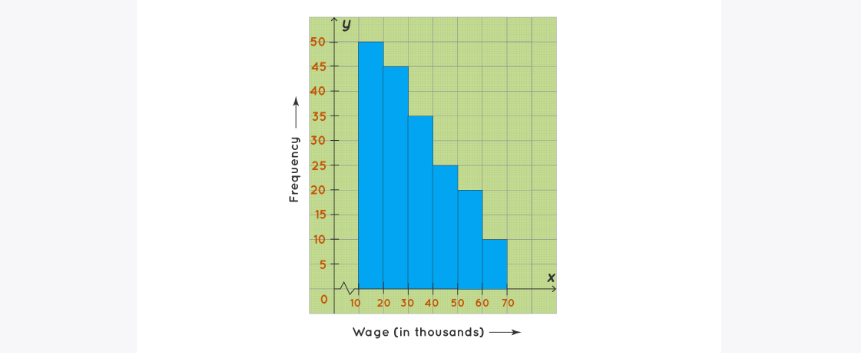
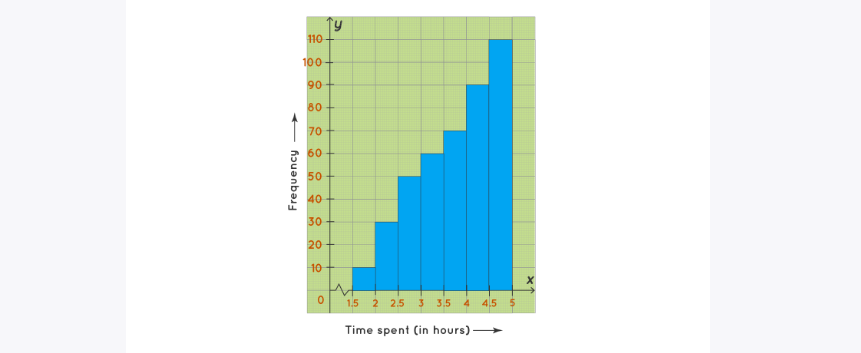
Ans: Data Investigation involves identifying its variables, its subjects (that is, on what or who are our observations going to be collected) and how to collect or access relevant and representative data. Machine learning data analysis uses algorithms to continuously improve itself over time, but quality data is necessary for these models to operate efficiently.

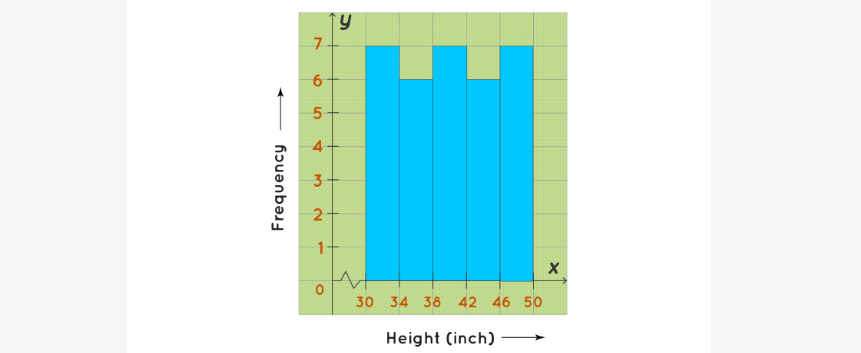
To explore Quantitative data we can use inbuild statistical and visualisation tools and it is very easy to train model on numeric data. On the other hand Qualitative data is interpretation-based, descriptive, and relating to language and not that easy to explore and use them directly to train a model.

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

Ans: **Bins** - A bin is a single range of continuous values used to group values in a chart. 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".

We have mainly 5 types of histogram shapes. They are listed below:

1. Bell Shaped Histogram
2. Bimodal Histogram
3. Skewed Right Histogram
4. Skewed Left Histogram
5. Uniform Histogram

****

**7. How do we deal with data outliers?**

Ans: It might be tempting to just remove the records where there are outliers in the data set but it’s not always the best approach. The outlier treatment method can vary from case to case and should be discussed with the business before finalizing the method. There are different approaches such as *replacing the outlier with the mean value*, or *median value* or in some cases *dropping the observation with the suspected outlier* so as to *avoid any bias in them*. We tend to delete the outlier if they are due to data entry errors caused due to human error, data processing errors.

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

Ans: There are three main measures of central tendency: the mode, the median and the mean.

Mean is a Value based Central Tendency i.e. it depends on the value of each observation where as Median is a Position dependent Central Tendency. If a dataset has few outliers then for those outliers the Mean value will get affected but the Median will not. For example, consider the following dataset

Dataset = {2, 3, 5, 3, 6, 7, 5, 8, 9, 6, 7, 4, 5, 4, 6}

Sorted Data = {2, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9}

Mean = 80/15 = 5.33 and Median = 5

So, both Mean and Median are nearly same.

But if we introduce two outliers, say 61 and 71 in the dataset then The mean will vary too much but Median will not.

Updated Dataset = {2, 3, 5, 3, 6, 7, 5, 8, 9, 6, 7, 4, 5, 4, 6, 61, 71}  
Sorted Data = {2, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9, 61, 71}  
Mean = 168/17 = 12.47 and Median = 6

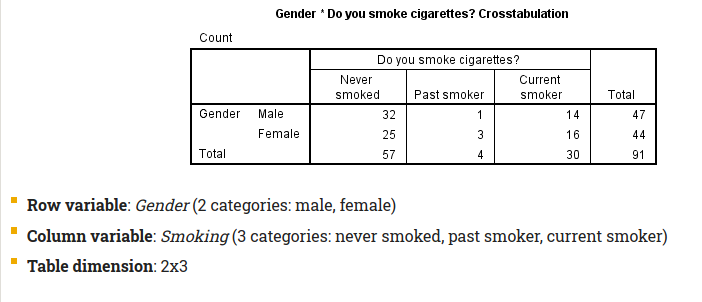
**9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?**

Ans: With the help of following observation in a scatter plot of two variables we can understand the relationship between them.  
(i) A large amount of scatter around the line indicates a weak relationship.   
(ii) Little scatter represents a strong relationship.   
(iii) If all points fall directly on a straight line, we have a perfect linear relationship between our two 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.

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

Ans: 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.

****