**Machine Learning Assignment 4**

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

Machine Learning Steps: There are seven major steps to start working with Machine Learning.

(a)Collecting Data:

(b)Preparing Data:

(c)Choosing a Model:

(d)Training the Model:

(e)Evaluating the Model:

(f)Parameter Tuning:

(g)Making Predictions:

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

**Types of Data**

**Structured Data**

This type of data is usually composed of numbers or words. They are usually stored in Relational databases and can be easily searched using SQL queries.

**Qualitative or Categorical Data**

Qualitative or Categorical Data is data that can’t be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That’s why it is also known as Categorical Data. These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data.

#### **The Qualitative data are further classified into two parts :**

### ****Nominal Data****

Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can’t be compared with another color.

**Examples of Nominal Data :**

* Colour of hair (Blonde, red, Brown, Black, etc.)
* Marital status (Single, Widowed, Married)
* Nationality (Indian, German, American)
* Gender (Male, Female, Others)
* Eye Color (Black, Brown, etc.)

### ****Ordinal Data****

Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can’t do any arithmetical tasks on them.

#### ****Examples of Ordinal Data :****

* When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
* Letter grades in the exam (A, B, C, D, etc.)
* Ranking of people in a competition (First, Second, Third, etc.)
* Economic Status (High, Medium, and Low)
* Education Level (Higher, Secondary, Primary)

### ****Numeric or Quantitative Data****

Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis. These kinds of data are also known as Numerical data. It answers the questions like “how much,” “how many,” and “how often.” For example, the price of a phone, the computer’s ram, the height or weight of a person, etc., falls under quantitative data.

Quantitative data can be used for statistical manipulation. These data can be represented on a wide variety of graphs and charts, such as bar graphs, histograms, scatter plots, boxplots, pie charts, line graphs, etc.

#### ****Examples of Quantitative Data :****

* Height or weight of a person or object
* Room Temperature
* Scores and Marks (Ex: 59, 80, 60, etc.)
* Time

#### **The Quantitative data are further classified into two parts :**

### ****Discrete Data****

The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can’t be broken into decimal or fraction values.

The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table.

#### ****Examples of Discrete Data :****

* Total numbers of students present in a class
* Cost of a cell phone
* Numbers of employees in a company
* The total number of players who participated in a competition
* Days in a week

### ****Continuous Data****

Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.

The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers to record different types of data such as temperature, height, width, time, speed, etc.

#### **Examples of Continuous Data :**

* Height of a person
* Speed of a vehicle
* “Time-taken” to finish the work
* Wi-Fi Frequency
* Market share price

## Unique

This type of data has a unique value for each sample and the number of categories is usually large. Sometimes it is so large that it cannot be called categorical data, but it still consists of alphabets and numbers. Examples are product id of all items in a store, student numbers of all individuals in a college, postal code of individuals birthplace, and so on.

# Tabular data

This usually refers to the collection of data from multiple different data types as shown in Figure 1. The tabular data consists of multiple features/columns with each of them having a particular data type.

# Unstructured data

This type of data is usually composed of everything else including texts, images, videos, speech/audio, time series, and so on. They are usually stored in non-relational databases and cannot be searched easily.

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

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

**Numerical vs. Categorical Data**

• Numerical data are values obtained for quantitative variable, and carries a sense of magnitude related to the context of the variable (hence, they are always numbers or symbols carrying a numerical value). Categorical data are values obtained for a qualitative variable; categorical data numbers do not carry a sense of magnitude.

• Numerical data always belong to either ordinal, ratio, or interval type, whereas categorical data belong to nominal type.

• Methods used to analyse quantitative data are different from the methods used for categorical data, even if the principles are the same at least the application has significant differences.

• Numerical data are analysed using statistical methods in descriptive statistics, regression, time series and many more.

• For categorical data usually descriptive methods and graphical methods are employed. Some non-parametric tests are also used.

**Feature selection vs. dimensionality reduction**

# Dimensionality reduction:

# Dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction.

* **Feature selection:** you select a subset of the original feature set; while
* **Feature extraction:** you build a new set of features from the original feature set.  
  Examples of feature extraction: extraction of contours in images, extraction of digrams from a text, extraction of phonemes from recording of spoken text, etc.
* **Feature extraction** involves a transformation of the features, which often is not reversible because some information is lost in the process of dimensionality reduction.

# Feature Selection

* Feature selection yields a subset of features from the original set of features, which are best representatives of the data. It is an exhaustive search.  
  -In text data, features might be size of characters or some global features of the text. Feature selection will keep only certain features of those.  
  -Feature selection is done in the context of an optimization problem.

# Dimension Reduction

* Dimensionality reduction is generic and only depends on the data and not on what you plan to do with it.
* Assuming a classification problem you select the features that will help you classify your data better, while a dimensionality reduction algorithm is unaware of this and just projects the data into a lower dimensionality space. That in turn can work quite well or not for your classification algorithm.

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

**(i). The histogram**

**(ii) Use a scatter plot**

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

**(i) 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 (also known as bins). Basically, it summarizes discrete or continuous data. In the most common form of histogram, the [independent variable](https://www.techtarget.com/whatis/definition/independent-variable) is plotted along the horizontal axis and the [dependent variable](https://www.techtarget.com/whatis/definition/dependent-variable) is plotted along the vertical axis. The data appears as colored or shaded rectangles of variable area.

Histogram definition can be put forward as a tool that visualizes the distribution of data over a continuous interval or a certain [time](https://www.vedantu.com/maths/time) period. It helps us to get an estimate of where the values are concentrated, what are the extremes if there is any gap or unusual values. To some extent, a Histogram also gives us a brief view of a [probability distribution](https://www.vedantu.com/maths/probability-distribution). A Histogram is quite similar to a vertical bar graph but the difference that lies between them is that there is no gap between the bars in the Histogram, unlike a bar graph.

**(ii) Use a scatter plot**

* A scatter plot is a chart type that is normally used to observe and visually display the relationship between variables. It is also known as a scattergram, scatter graph, or scatter chart.
* The data points or dots, which appear on a scatter plot, represent the individual values of each of the data points and also allow pattern identification when looking at the data holistically.
* The most common use of the scatter plot is to display the relationship between two variables and observe the nature of such a relationship. The relationships observed can either be positive or negative, non-linear or linear, and/or, strong or weak.

### Scatter Plot Applications and Uses

#### 1. Demonstration of the relationship between two variables

The most common use of the scatter plot is to display the relationship between two variables and observe the nature of the relationship. The relationships observed can either be positive or negative, non-linear or linear, and/or, strong or weak.

The data points or dots, which appear on a scatter plot, represent the individual values of each of those data points and also allow pattern identification when looking at the data holistically.

#### 2. Identification of correlational relationships

Another common use of scatter plots is that they enable the identification of correlational relationships. Scatter plots tend to have [independent variables](https://corporatefinanceinstitute.com/resources/knowledge/modeling/independent-variable/) on the horizontal axis and dependent variables on the vertical axis. It allows the observer to know or get an idea of what the possible vertical value may be, provided there is information on the horizontal value.

#### 3. Identification of data patterns

Data pattern identification is also possible with scatter plots. Data points can be grouped together based on how close their values are, and this also makes it easy to identify any outlier points when there are data gaps.

Seeing as scatter plots aid in the identification of correlations between variables, the nature of the correlations can also be estimated based on a specific confidence level.

* Positive correlation depicts a rise, and it is seen on the diagram as data points slope upwards from the lower-left corner of the chart towards the upper-right.
* Negative correlation depicts a fall, and this is seen on the chart as data points slope downwards from the upper-left corner of the chart towards the lower-right.
* Data that is neither positively nor negatively correlated is considered uncorrelated (null).

**(iii) PCA (Personal Computer Aid)**

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D. PCA helps in finding a sequence of linear combinations of variables.

**Applications of PCA in Machine Learning**

* PCA is used to visualize multidimensional data.
* It is used to reduce the number of dimensions in healthcare data.
* PCA can help resize an image.
* It can be used in finance to analyze stock data and forecast returns.
* PCA helps to find patterns in the high-dimensional datasets.

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

The machine learning algorithms need training data in a single view i.e. a flat structure. As most organizations maintain multiple sources of data, the data preparation by combining multiple data sources to bring all necessary attributes in a single flat file is a time and resource (domain expertise) expensive process.

The data gets exposed to multiple sources of error at this step and requires strict peer review to ensure that the domain-established logic has been communicated, understood, programmed, and implemented well.

Since data warehouses integrate data from multiple sources, quality issues related to data acquisition, cleaning, transformations, linking, and integration become critical.

A very popular notion among most the data scientists is that the data preparation, cleaning, and transformation take up the majority of the model building time – and it is an absolute truth. Hence, it is advised not to rush through the data to feed into the model and perform extensive data quality checks.

* **Quantitative data:** The data collected on the grounds of the numerical variables are quantitative data. Quantitative data are more objective and conclusive in nature. It measures the values and is expressed in numbers. The data collection is based on “how much” is the quantity. The data in quantitative analysis is expressed in numbers so it can be counted or measured. The data is extracted from experiments, surveys, market reports, matrices, etc.
* **Qualitative data:** The data collected on grounds of categorical variables are qualitative data. Qualitative data are more descriptive and conceptual in nature. It measures the data on basis of the type of data, collection, or category. The data collection is based on what type of quality is given. Qualitative data is categorized into different groups based on characteristics. The data obtained from these kinds of analysis or research is used in theorization, perceptions, and developing hypothetical theories. These data are collected from texts, documents, transcripts, audio and video recordings, etc.

Comparative table of qualitative and quantitative data,

| Qualitative Data | Quantitative Data |
| --- | --- |
| 1. Qualitative data uses methods like interviews, participant observation, focus on a grouping to gain collective information. | 1. Quantitative data uses methods as questionnaires, surveys, and structural observations to gain collective information. |
| 2. Data format used in it is textual. Datasheets are contained of audio or video recordings and notes. | 2. Data format used in it is numerical. Datasheets are obtained in the form of numerical values. |
| 3. Qualitative data talks about the experience or quality and explains the questions like ‘why’ and ‘how’. | 3. Quantitative data talks about the quantity and explains the questions like ‘how much’, ‘how many . |
| 4. The data is analyzed by grouping it into different categories. | 4. The data is analyzed by statistical methods. |
| 5.Qualitative data are subjective and can be further open for interpretation. | 5. Quantitative data are fixed and universal. |

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

A **histogram** is a type of chart that allows us to visualize the distribution of values in a dataset.

The x-axis displays the values in the dataset and the y-axis shows the frequency of each value.

Depending on the values in the dataset, a histogram can take on many different shape

The following examples show how to describe a variety of different histograms.

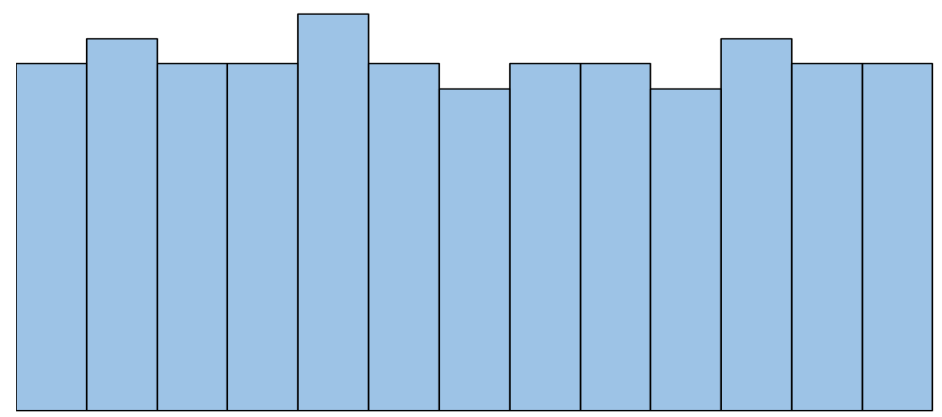
### 1. Bell-Shaped

A histogram is bell-shaped if it resembles a “bell” curve and has one single peak in the middle of the distribution. The most common real-life example of this type of distribution is the [normal distribution](https://www.statology.org/the-normal-distribution/).

### https://www.statology.org/wp-content/uploads/2021/08/hist_shape2.png

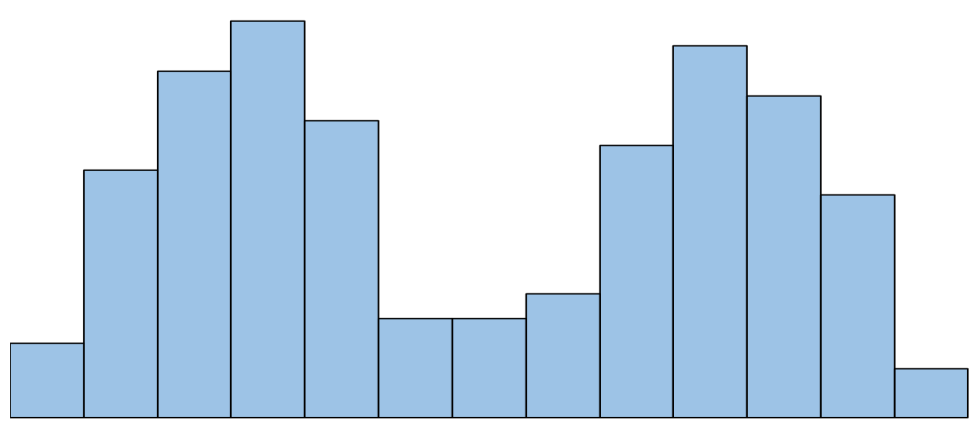
### 2. Uniform

A histogram is described as “uniform” if every value in a dataset occurs roughly the same number of times. This type of histogram often looks like a rectangle with no clear peaks.



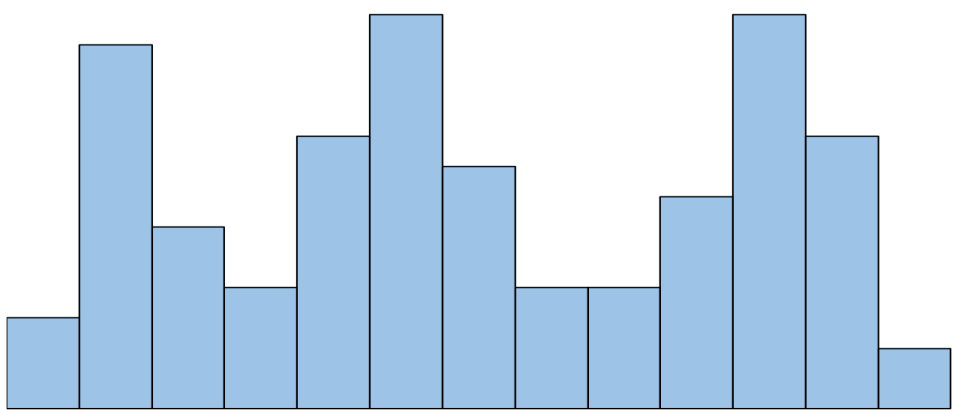
### 3. Bimodal

A histogram is described as “bimodal” if it has two distinct peaks. We often say that this type of distribution has multiple modes – that is, multiple values occur most frequently in the dataset.



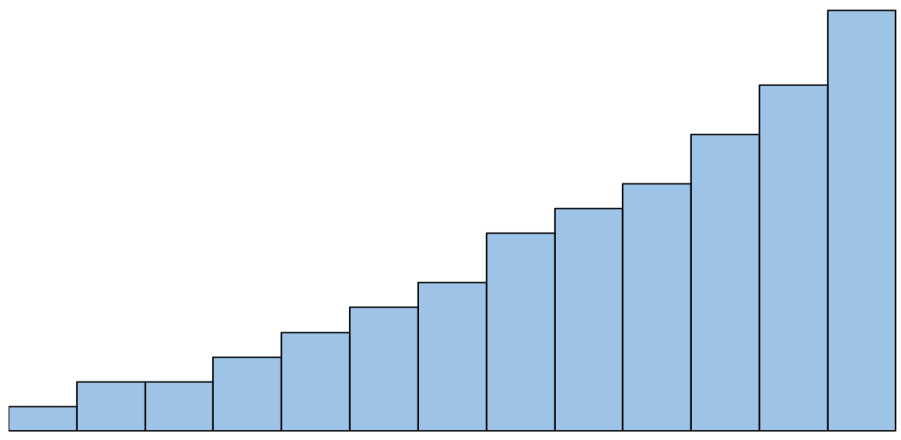
### 4. Multimodal

A histogram is described as “multimodal” if it has more than two distinct peaks.



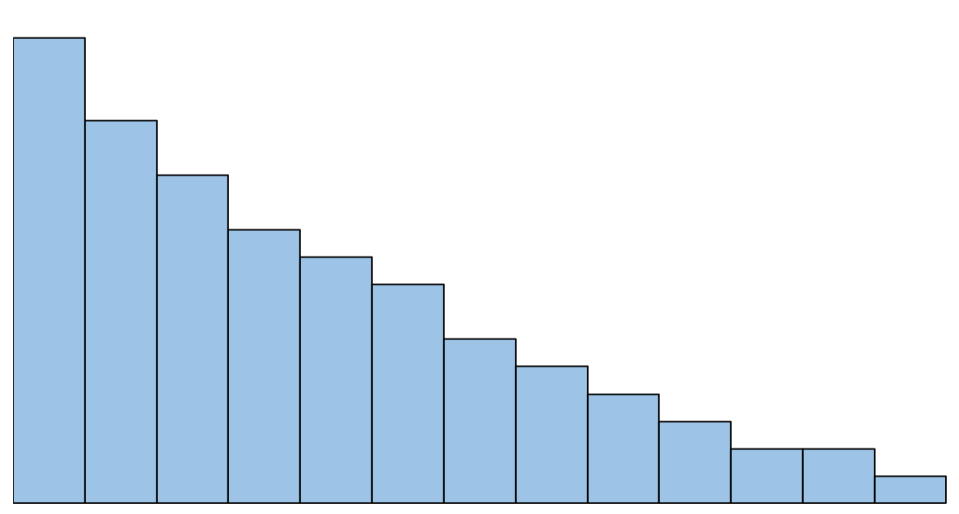
### 5. Left Skewed

A histogram is left skewed if it has a “tail” on the left side of the distribution. Sometimes this type of distribution is also called “negatively” skewed.



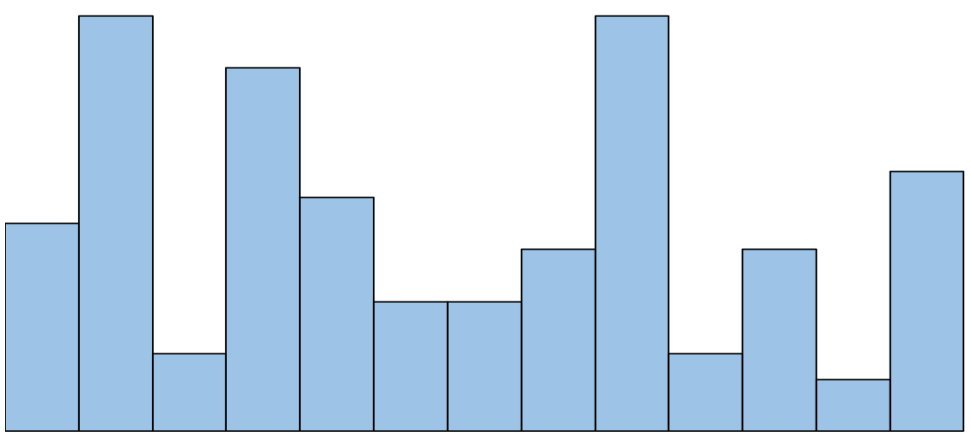
### 6. Right Skewed

A histogram is right skewed if it has a “tail” on the right side of the distribution. Sometimes this type of distribution is also called “positively” skewed.



### 7. Random

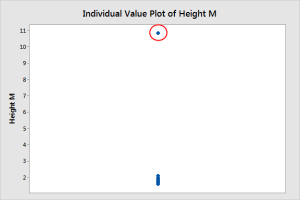
The shape of a distribution can be described as “random” if there is no clear pattern in the data at all.



### A bin is a single range of continuous values used to group values in a chart. Binning data helps simplify data visualizations, so people can get a sense of their data’s distribution and easily spot outliers

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

[Outliers](https://statisticsbyjim.com/glossary/outliers/) are unusual values in your dataset, and they can distort statistical analyses and violate their assumptions. Unfortunately, all analysts will confront outliers and be forced to make decisions about what to do with them. Given the problems they can cause, you might think that it’s best to remove them from your data. But, that’s not always the case. Removing outliers is legitimate only for specific reasons.



Outliers can be very informative about the subject-area and data collection process. It’s essential to understand how outliers occur and whether they might happen again as a normal part of the process or study area. Unfortunately, resisting the temptation to remove outliers inappropriately can be difficult. Outliers increase the variability in your data, which decreases statistical power. Consequently, excluding outliers can cause your results to become statistically significant.

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

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.  
  
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.

**What is the mean?**  
  
The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

**What is the median?**  
  
The median is the middle value in distribution when the values are arranged in ascending or descending order.  
  
The median divides the distribution in half (there are 50% of observations on either side of the median value). In a distribution with an odd number of observations, the median value is the middle value.

**What is the mode?**  
  
The mode is the most commonly occurring value in a distribution.  
  
Consider this dataset showing the retirement age of 11 people, in whole years:  
  
54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60  
  
This table shows a simple frequency distribution of the retirement age data.

|  |  |
| --- | --- |
| Age | Frequency |
| 54 | 3 |
| 55 | 1 |
| 56 | 1 |
| 57 | 2 |
| 58 | 2 |
| 60 | 2 |

The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.  
  
Advantage of the mode:  
  
The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.  
  
Limitations of the mode:  
  
The are some limitations to using the mode. In some distributions, the mode may not reflect the centre of the distribution very well. When the distribution of retirement age is ordered from lowest to highest value, it is easy to see that the centre of the distribution is 57 years, but the mode is lower, at 54 years.  
  
54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60  
  
It is also possible for there to be more than one mode for the same distribution of data, (bi-modal, or multi-modal). The presence of more than one mode can limit the ability of the mode in describing the centre or typical value of the distribution because a single value to describe the centre cannot be identified.  
  
In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).  
  
In cases such as these, it may be better to consider using the median or mean, or group the data in to appropriate intervals, and find the modal class.

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

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