# Reading Data to R

- data.frame()
- Scan()
- readline()
- read.table()
- read.csv()
- data()  :-to see a list of built-in datasets
  - Command **library** loads the package MASS (Modern Applied Statistics with S)
  - Library("MASS")
  - Command **data()** will list all the datasets in loaded packages.

Sample code- read-mass-lib.r

- uci machine learning repository
- kaggle

# Importing Data from Files

- If you are using R, you will likely need to read in data at some point.
- While R can read excel .xls and .xlsx files these file types often cause problems.
- Comma separated files (.csv) are much easier to work with. It's best to save these files as csv before reading them into R.
- If you need to read in a csv with R the best way to do it is with the command read.csv.
- Here is an example of how to read CSV in R:

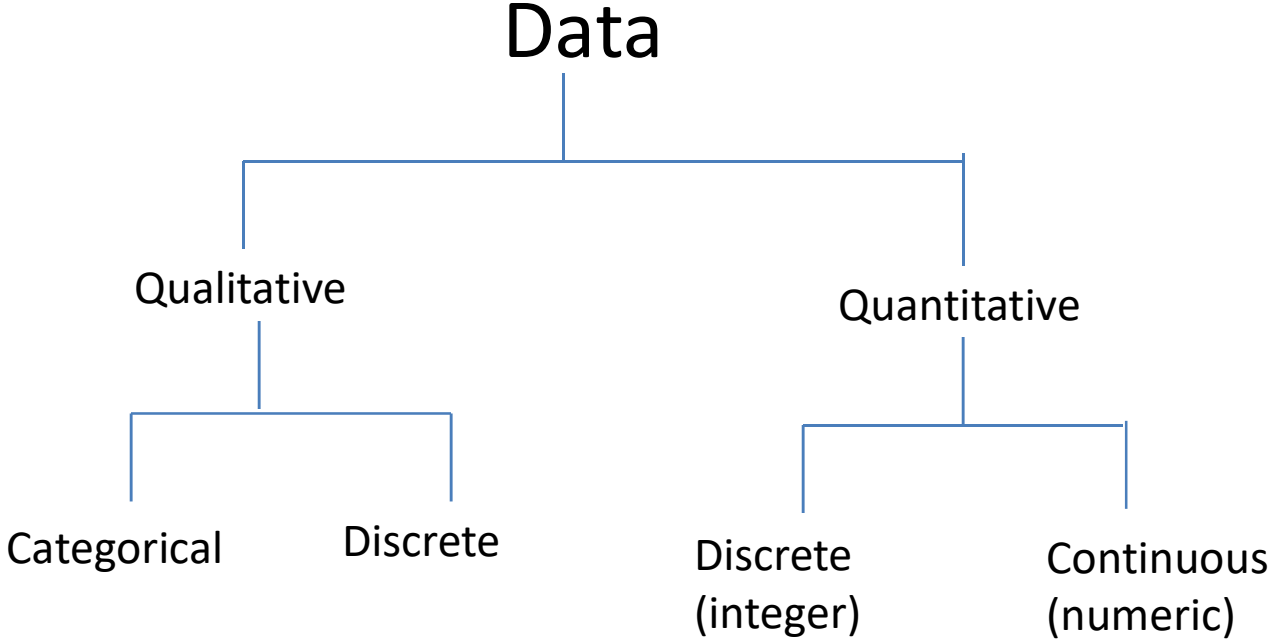
- read.csv(“filename.csv”,header=TRUE, sep=“,”)
- read.csv(“filename.csv”)
- Then use the functions
  - dim() :- find the dimension as nxp (number of rows x number of attributes)
  - head():- display first 6 rows
  - tail():- display last 6 rows
- After this exploring the structure of each attribute using the function str().

- The attributes or coloumns are numeric (non-integer type) ,integer type , character type etc.
- The integer type attributes may be qualitative one (categorical) or discrete-quantitative.
- The numeric type data are continuous or discrete quantitative.

**Sample code- read-txt-example.r**

**Sample code- read-txt-example2.r**

- In a data set the attributes are



- The summary() function displays several common summary statistics
- Measuring the mean and median of our data provides one way to quickly summarize the values.
- But these measures of center tell us little about whether or not there is diversity in the measurements.

- The **five-number summary** is a set of five statistics that roughly depict the spread of a dataset.
- All five of the statistics are included in the output of the summary() function.
- Written in order, they are:

1. Minimum (Min.)
2. First quartile, or Q1 (1st Qu.)
3. Median, or Q2 (Median)
4. Third quartile, or Q3 (3rd Qu.)
5. Maximum (Max.)

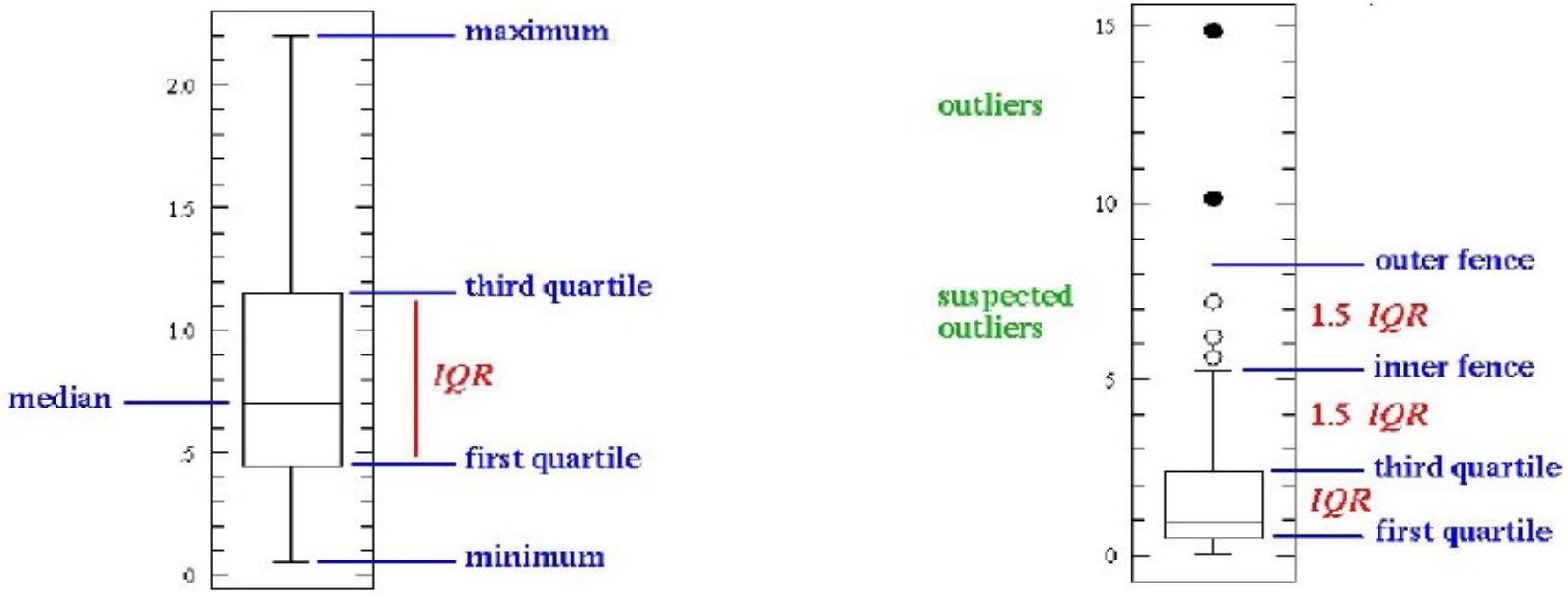
- The minimum and maximum are the most extreme values found in the dataset, indicating the smallest and largest values respectively.
- The first and third quartiles, Q1 and Q3, refer to the value below or above which one quarter of the values are found.
- Along with the median (Q2), the quartiles divide a dataset into four portions, each with the same number of values.

- The difference between Q1 and Q3 is known as the **interquartile range** (**IQR**)

# Impact of outliers in the data sets?

- Outliers can drastically change the results of the data analysis and statistical modelling
- Very common in data science / big data
- Unfavourable impacts of outliers in the data set:
  - It increases the error variance
  - Reduces the power of statistical tests
  - Biased estimates
  - Impact the basic linear assumptions (Linear regression, ANOVA, t-test and other statistical model assumptions)

# Box plot




- The first **quartile** ($Q_1$) is defined as the middle number between the smallest number and the median of the data set
- The second **quartile** ($Q_2$) is the median of the data
- The third **quartile** ($Q_3$) is the middle value between the median and the highest value of the data set

- The horizontal lines forming the box in the middle of each figure represent Q1, Q2 (the median), and Q3 when reading the plot from bottom-to-top.
- The median is denoted by the dark line,

## Visualizing numeric variables – boxplots

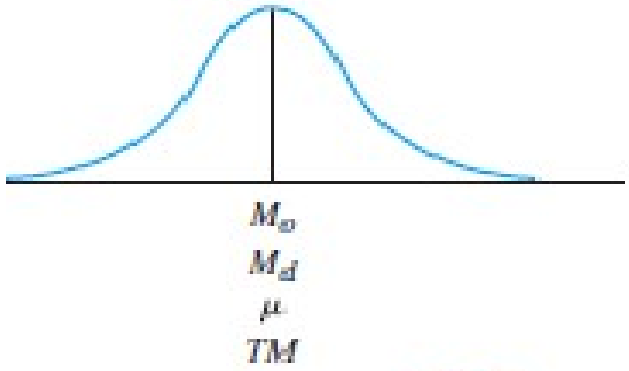
- Visualizing numeric variables can be helpful for diagnosing many problems with data.
- A common visualization of the five-number summary is a **boxplot**.
- The boxplot displays the center and spread of a numeric variable in a format that allows you to quickly obtain a sense of the range and skew of a variable, or compare it to other variables.
- boxplot(attributename,main="name ")

## Visualizing numeric variables – histograms

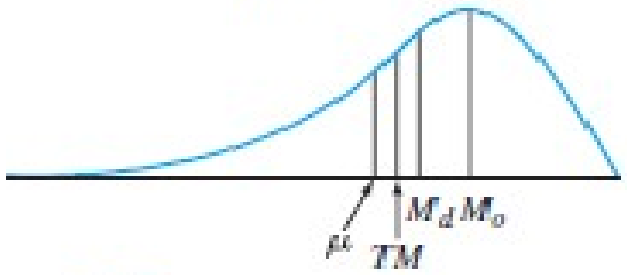
- A **histogram** is another way to graphically depict the spread of a numeric variable.
- hist(attribute-name, main=“ ”)
- The histogram is composed of a series of bars with heights indicating the count, or f**requency**

- A histogram is **symmetric** in shape if the right and left sides have essentially the same shape.
- When the right side of the histogram, containing the larger half of the observations in the data, extends a greater distance than the left side, the histogram is referred to as **skewed to the right.**
- The histogram is **skewed to the left** when its left side extends a much larger distance than the right side.

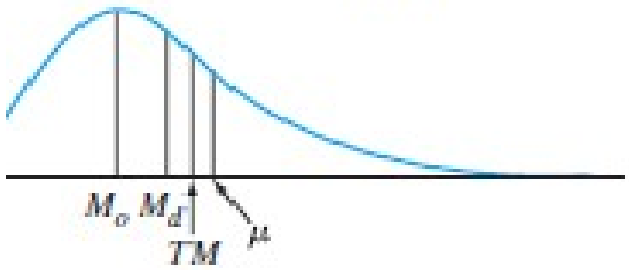
- The measures of central tendency related for a given set of measurements depends on the **skewness** of the data.
- If the distribution is mound-shaped and symmetrical about a single peak, the mode (*Mo*), median (*Md*), mean (m), and Trimmed mean(*TM*) will all be the same.
- This is shown using a smooth curve and population quantities.
- If the distribution is skewed, having a long tail in one direction and a single peak, the mean is pulled in the direction of the tail; the median falls between the mode and the mean; and depending on the degree of trimming,
- The trimmed mean usually falls between the median and the mean.
- The following figures illustrate this for distributions skewed to the left and to the right.
- If mean value is greater than median this implies that the distribution of the attribute is right skewed.




(a) A mound-shaped distribution




(b) A distribution skewed to the left




(c) A distribution skewed to the right

- Another method is the uses of apply()
  - apply(X, MARGIN, FUN)

    Example :-Return the sum of each of the columns of the matrix m

    apply(m,2,sum)

We can compare the mean and median of each attribute from the apply() function result.