Applied Statistics Using R

Unit-2

Data Frame's Row And Column Names

```
names(mtcars)      # see the column names
mtcars1=mtcars   #Copy to other object
mtcars1
names(mtcars1) <- c("MilesPerGallon","NumOfCylinders",
                    "Displacement", "HorsePower",
                    "RearAxleRatio","Weight","MileTime",
                    "EngineType", "Transmission",
                    "NumOfGears", "NumOfCarburetors")
View(mtcars1)
dim(mtcars1)         # check number of rows and columns
length(mtcars1)      #check number of columns
mtcars1[1:2,3]
attach(mtcars1)     #attach data frame
```

- **DataFrames – used to store data table with different type**

```
nrow(mtcars)
ncol(mtcars)
tail(mtcars)
head(mtcars)
```

- Column vector

```
mtcars[[9]]
mtcars$am
mtcars[,"am"]
mtcars[,9]
```

- Row Slice

```
mtcars[24,]
mtcars[c(3, 24),]
mtcars[c(3:14),]
```

- Column Slice

```
mtcars[9]
mtcars$am
mtcars["am"]
mtcars[c("mpg", "am")]
```

- mtcars[mtcars$am == 0,]

Subset Data

- **Using subset function**
  - subset() will subset the dataframe
  - gear3=subset(mtcars,mtcars$gear==3, select = c(wt, qsec))
  - gearG3=subset(mtcars,mtcars$gear>3, select = -wt)
- **Subscripting from data frames**
  - mtcars[,1] gives first column of mtcars
- **Specifying a vector**
  - mtcars[1:5] gives first 5 columns of data
- **Sorting data frame columns**
  - newdata <- mtcars[order(mpg),] # sort by mpg
  - newdata <- mtcars[order(mpg, cyl),] # sort by mpg and cyl
  - newdata <- mtcars[order(mpg, -cyl),] #sort by mpg (ascending) and cyl (descending)

Exercises-4

- Import “Titanic” dataset
- Make two new dataframes :as subset of male survivors, and as subset off female survivors. Use either ***the square brackets,or subset*** to make the subsets.
- Find the name of the oldest surviving male and youngest female surviving name. Use ***which.max & which.min***
- Take random names of passengers from the Titanic, and sort them alphabetically. Hint: ***use sort***

- mydata2<-read.csv("titanic3.csv")
- nrow(mydata2)

- set1=subset(mydata2,(sex=="male")&(survived==1))
- nrow(set1)
- View(set1)

- mydata2$name[which.max(mydata2$age)]

- sortData=sort(mydata2$name)
- View(sortData)

Install Packages in R

- R packages provide a powerful mechanism for extending the functionality of R
- R packages are obtained from CRAN or other repositories
- The install.packages() can be used to install packages at the R console

install.packages("rmarkdown")
~~***install.packages(c("slidify", "ggplot2", "devtools"))***~~

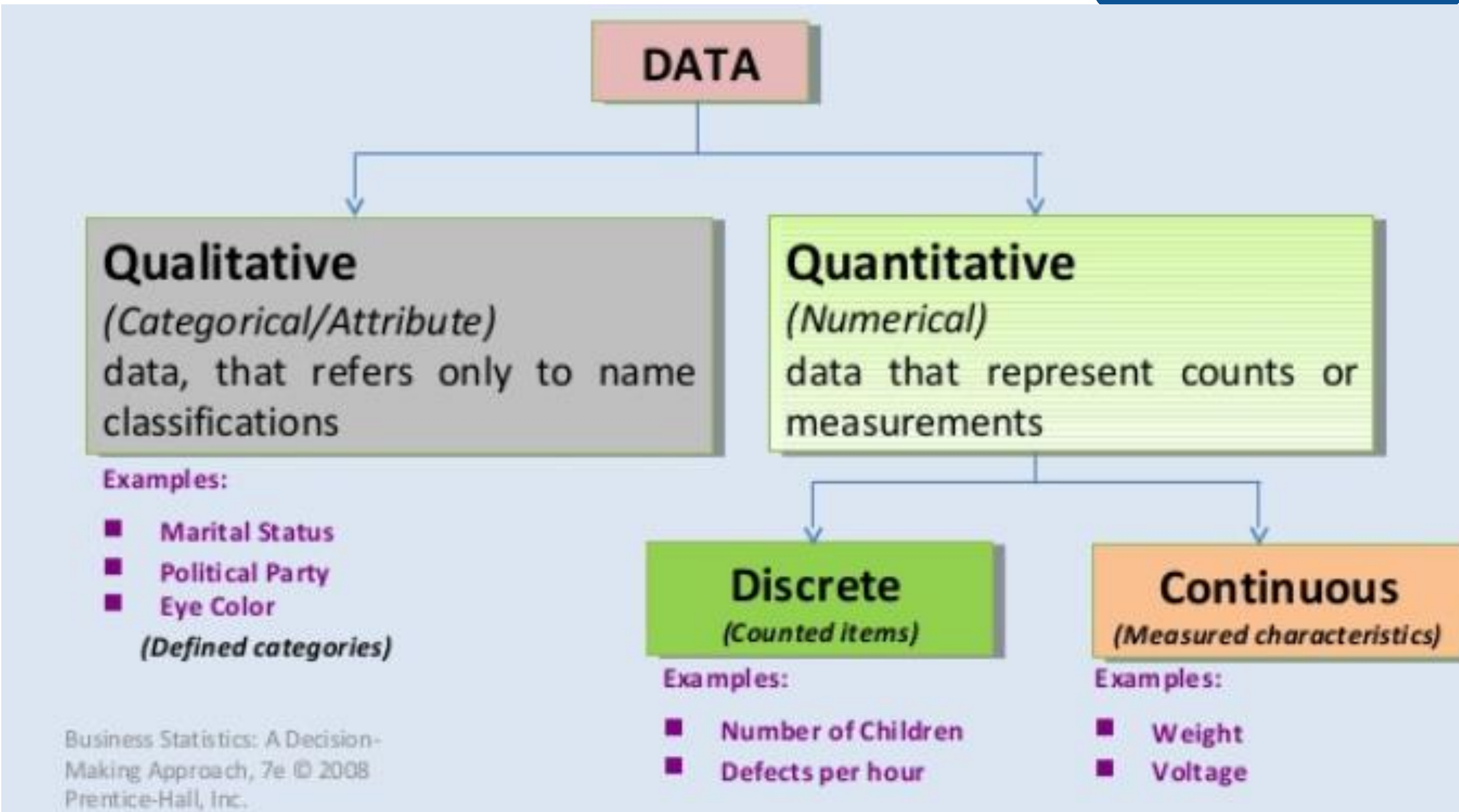
- The library() function loads the intalled packages to access the functionality of the package

library(rmarkdown)

- install.packages('rmarkdown')
- library(rmarkdown)

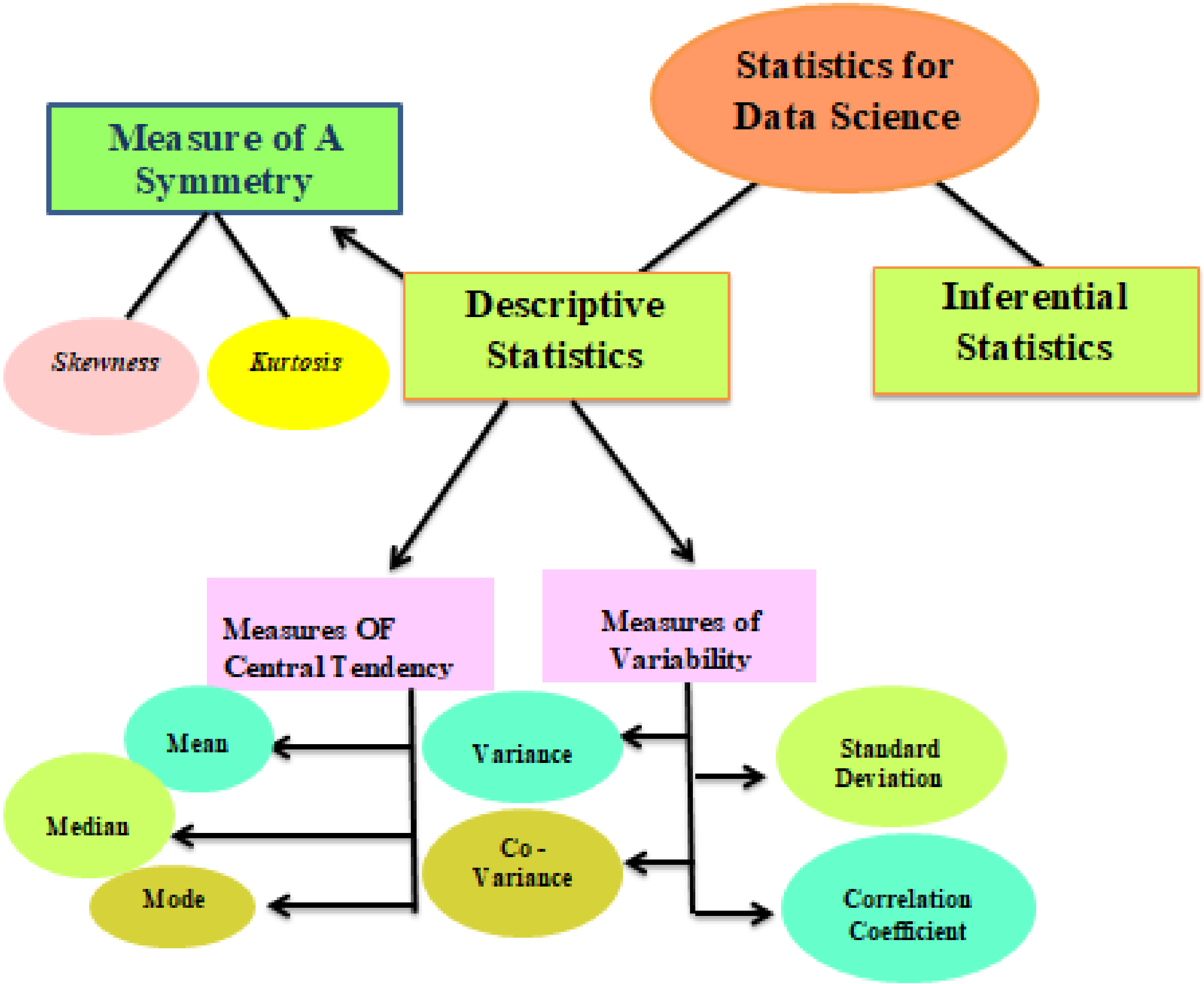
- ~~if (!requireNamespace("devtools"))~~
- ~~install.packages('devtools')~~
- ~~devtools::install_github('rstudio/rmarkdown')~~

- install.packages("tinytex")
- tinytex::install_tinytex(force = TRUE) # install TinyTeX


DATA
Qualitative
(Categorical/Attribute)
data, that refers only to name classifications
Examples:
Marital Status
Political Party
Eye Color
(Defined categories)
Quantitative
(Numerical)
data that represent counts or measurements
Discrete
(Counted items)
Examples:
Number of Children
Defects per hour
Continuous
(Measured characteristics)
Examples:
Weight
Voltage
Business Statistics: A Decision-Making Approach, 7e © 2008 Prentice-Hall, Inc.

Examples

Quantitative Data ("Numerical")	Qualitative Data ("Categorical")
• Height of 1st graders • Weight of sumo wrestlers • Duration of red lights • Age of Olympians • Distance of planets • Money in 401k plans • Temperature of coffee (200 F)	• Happiness rating • Gender • Pass/Fail • Eye Color • Interview transcript • Categories of plants • Descriptive temperature of coffee ("very hot")


Statistics for Data Science
Measure of A Symmetry
Descriptive Statistics
Inferential Statistics
Skewness
Kurtosis
Measures OF Central Tendency
Measures of Variability
Mean
Median
Mode
Variance
Co - Variance
Standard Deviation
Correlation Coefficient

CENTRAL TENDENCY

1. **Mean** = Sum of scores divided by the number of scores (often referred to as the statistical average)

Pronounced "x-bar"

Capital Sigma for "Sum of"

$$\bar{X} = \frac{\sum X}{N}$$

N represents the number of scores

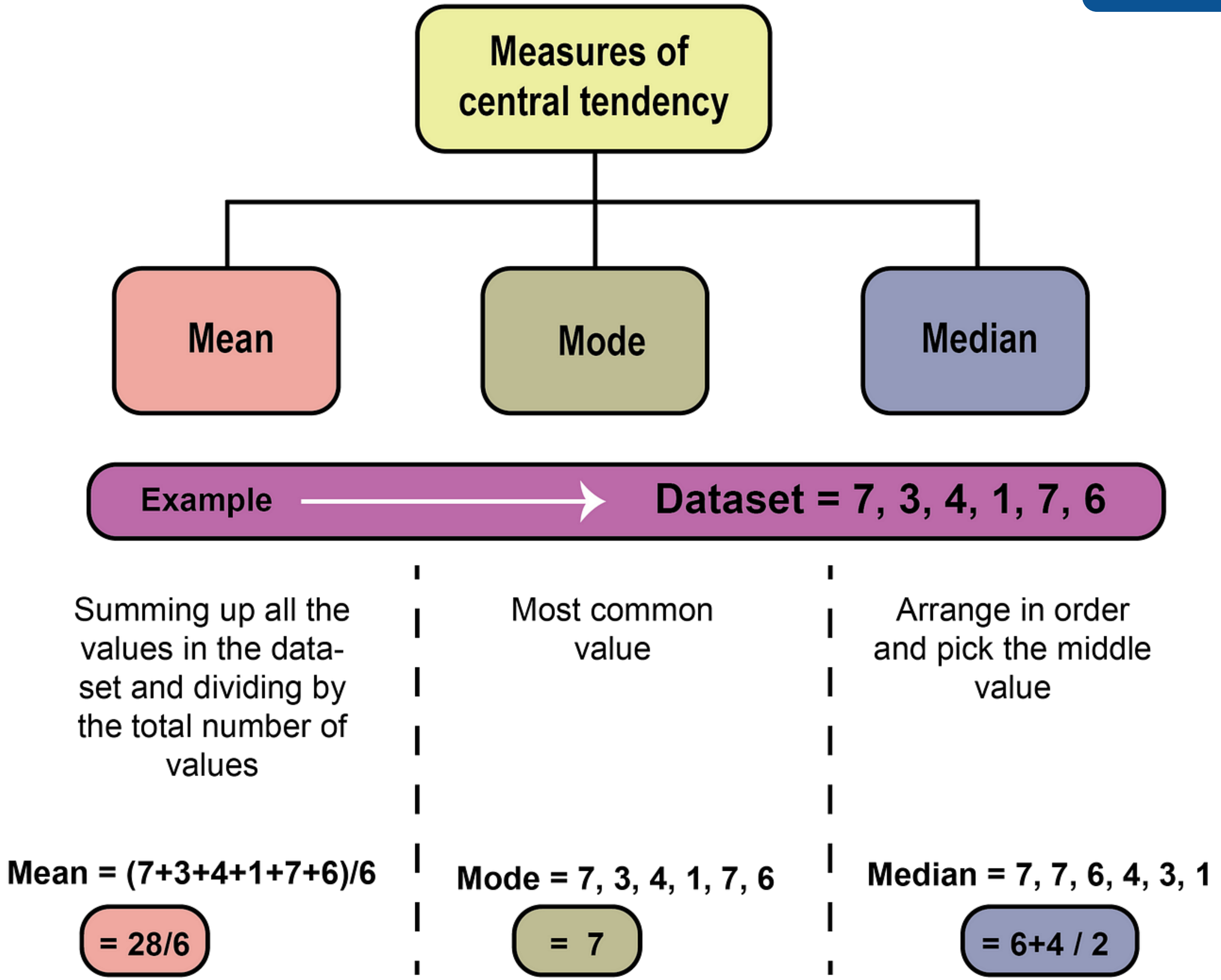
"x" represents each score

2. **Median** = Middle Most Number

$$M_d$$

3. **Mode** = Most Frequently Occurring Number

$$M_o$$


Measures of central tendency
Mean
Mode
Median
Example
Dataset = 7, 3, 4, 1, 7, 6
Summing up all the values in the data-set and dividing by the total number of values
Most common value
Arrange in order and pick the middle value
Mean = (7+3+4+1+7+6)/6
= 28/6
Mode = 7, 3, 4, 1, 7, 6
= 7
Median = 7, 7, 6, 4, 3, 1
= 6+4 / 2

MODE

- most frequent data point
- mode exists as a data point
- unaffected by extreme values
- useful for qualitative data
- may have more than 1 value

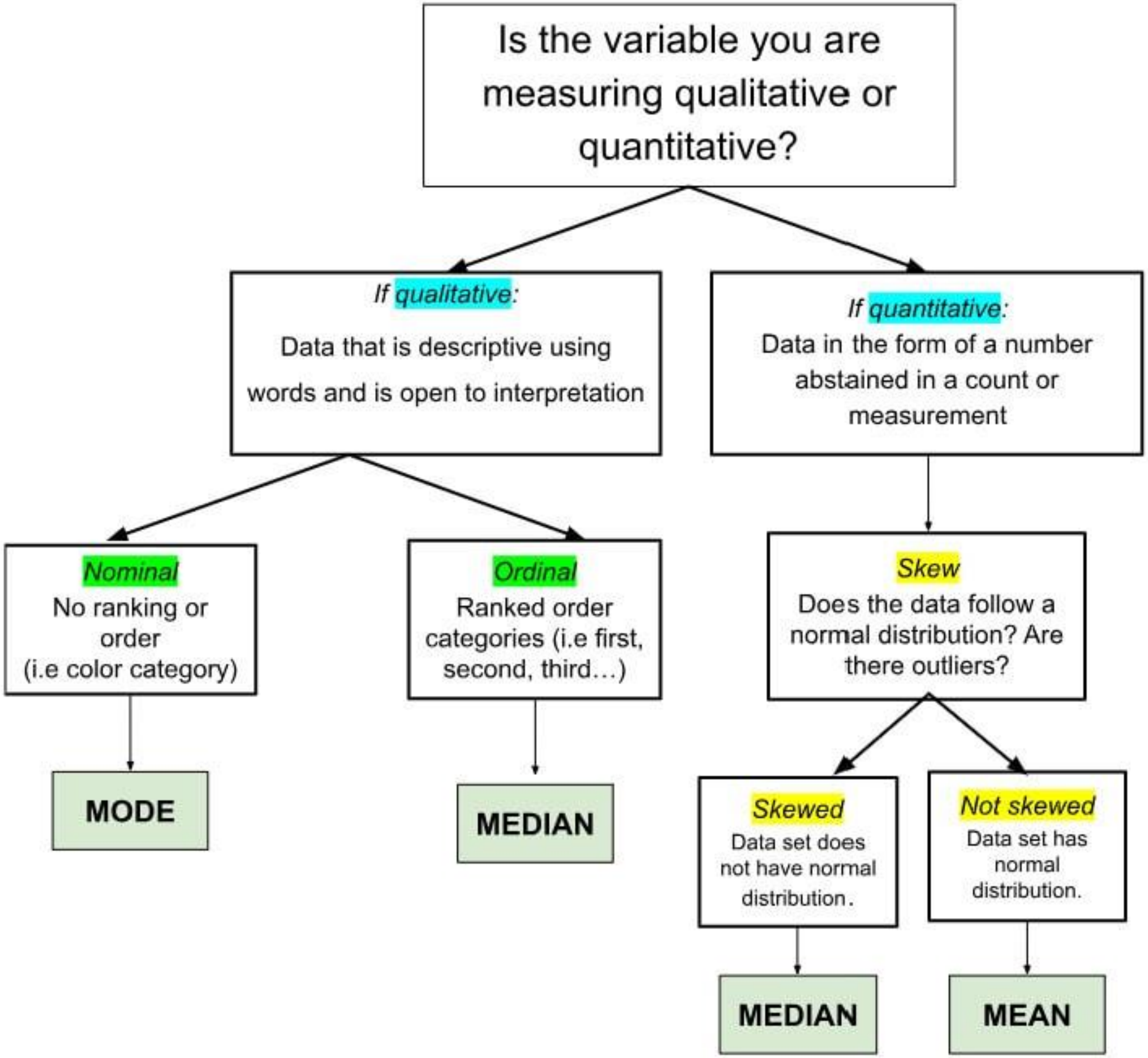
MEDIAN

- value that divides ranked data points into halves: 50% larger than it, 50% smaller
- may not exist as a data point in the set
- influenced by position of items, but not their values

MEAN

$$\bar{x} = \frac{\sum x}{N}$$

- most stable measure
- affected by extreme values
- may not exist as a data point in the set


Is the variable you are measuring qualitative or quantitative?
If qualitative:
Data that is descriptive using words and is open to interpretation
If quantitative:
Data in the form of a number abstained in a count or measurement
Nominal
No ranking or order
(i.e color category)
Ordinal
Ranked order categories (i.e first, second, third…)
Skew
Does the data follow a normal distribution? Are there outliers?
MODE
MEDIAN
Skewed
Data set does not have normal distribution.
Not skewed
Data set has normal distribution.
MEDIAN
MEAN

Measures Variability

- Range
- Standard Deviation
- Variance

B18 | =SQRT(D16/(10-1))

Using Standard Deviation Formula

	A			D	E
4		**A**	**B**	**C = B^2**	
5	**No.**	**Returns**	**Return - Mean**	**(Return - Mean)^2**	
6	**1**	72	12	144	
7	**2**	45	-15	225	
8	**3**	58	-2	4	
9	**4**	84	24	576	
10	**5**	60	0	0	
11	**6**	10	-50	2500	
12	**7**	91	31	961	
13	**8**	65	5	25	
14	**9**	55	-5	25	
15	**10**	60	0	0	
16	**Total**	600		4460	
17	**Mean**	60			
18	**Standard Deviation**	22.26			
19					

Variance, Standard Deviation, Range

Dispersion	Population
Variance	$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
Standard Deviation	$\sigma = \sqrt{\frac{\Sigma(X_i - \mu)^2}{N}}$
Range	Max – Min

Dispersion	Sample
Variance	$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$
Standard Deviation	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$
Range	Max – Min

- μ =Mean
- X =Random Variable
- N =no. of data types

- $\bar{x}$ =Mean
- x =Random Variable
- n =no. of data types

Example

Find the standard deviation and variance

x	$x-\bar{x}$	$(x-\bar{x})^2$
30	4	16
26	0	0
22	-4	16
78		32

Mean = 26

Sum = 0

The variance

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = 32 \,\square\, 2 = 16$$

The standard deviation

$$s = \sqrt{16} = 4$$

Skewness

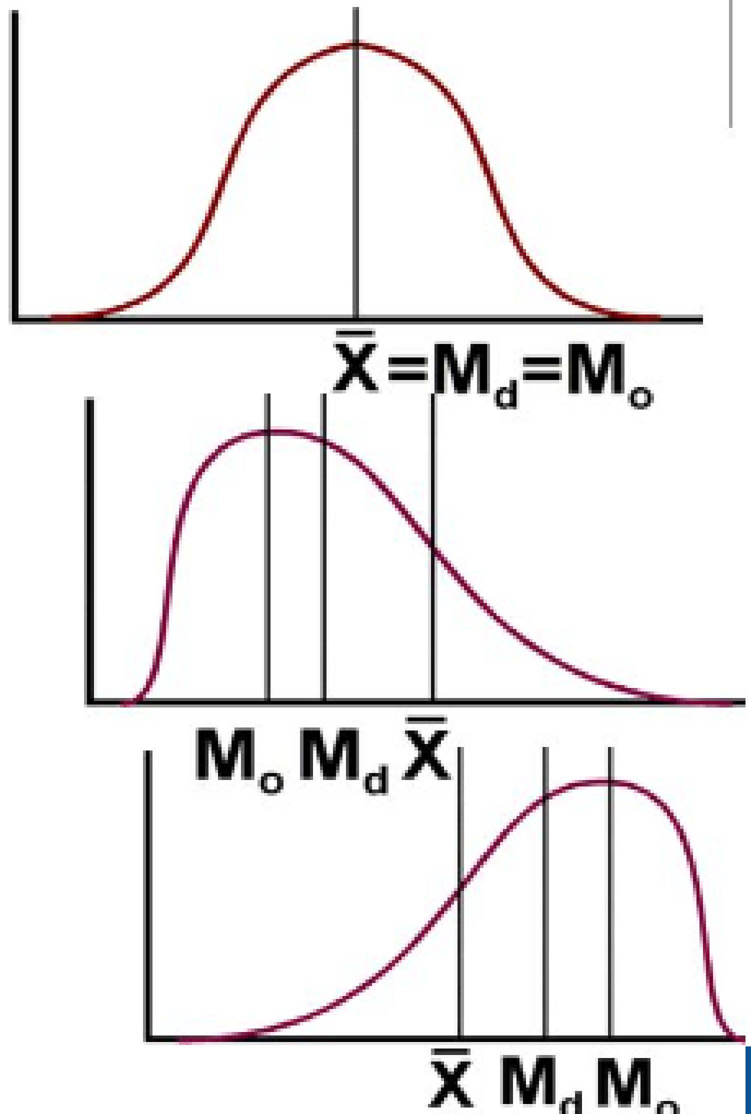
- If the mean exceeds the mode and median (Mode < Median < Mean) then the distribution is positively skewed. In other words, if the coefficient of skewness is positive then the distribution is skewed to the right.
- If the mode exceeds the median and mean (Mean < Median < Mode) then the distribution is negatively skewed. Thus, the coefficient of skewness will be negative and the distribution will be skewed to the left.
- If the value of the mean, median, and mode are equal then the distribution is a normal distribution and the coefficient of skewness will be 0.

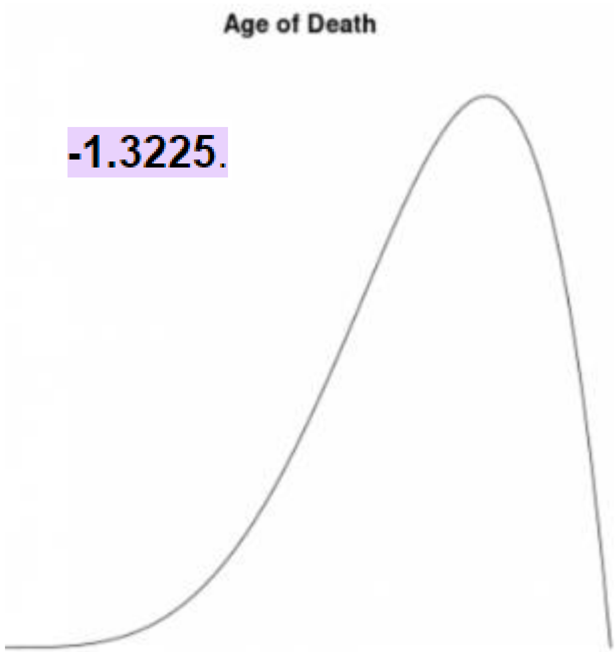
Using Mode: $\frac{\bar{x} - \text{Mode}}{s}$

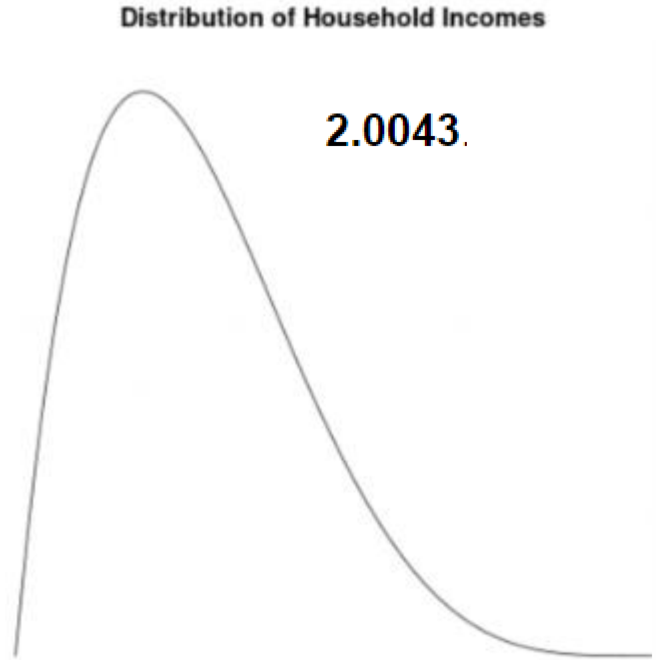
Using Median: $\frac{3(\bar{x} - \text{Median})}{s}$

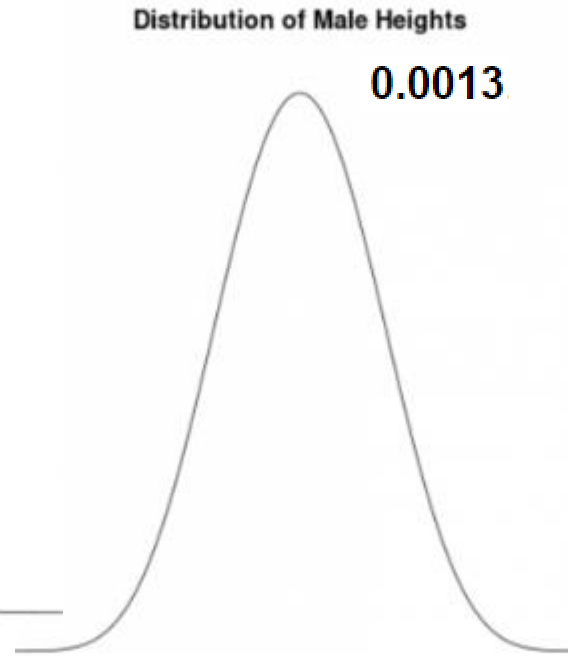
Measures of Central Tendency
The Shape of Distributions

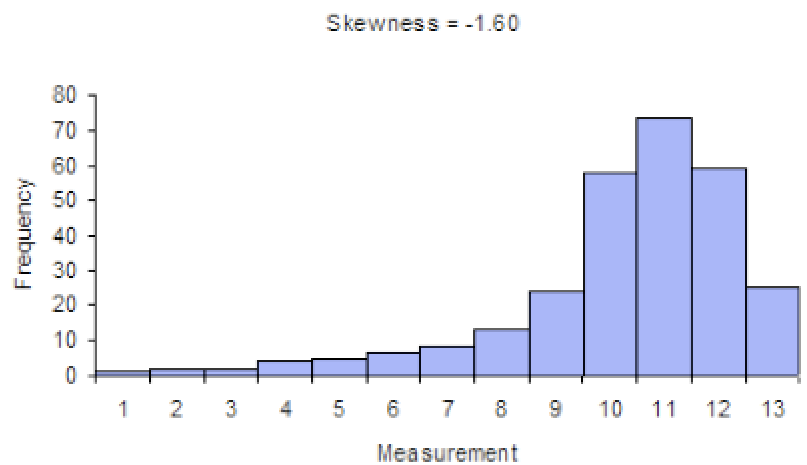
- With perfectly bell shaped distributions, the mean, median, and mode are identical.
- With positively skewed data, the mode is lowest, followed by the median and mean.
- With negatively skewed data, the mean is lowest, followed by the median and mode.

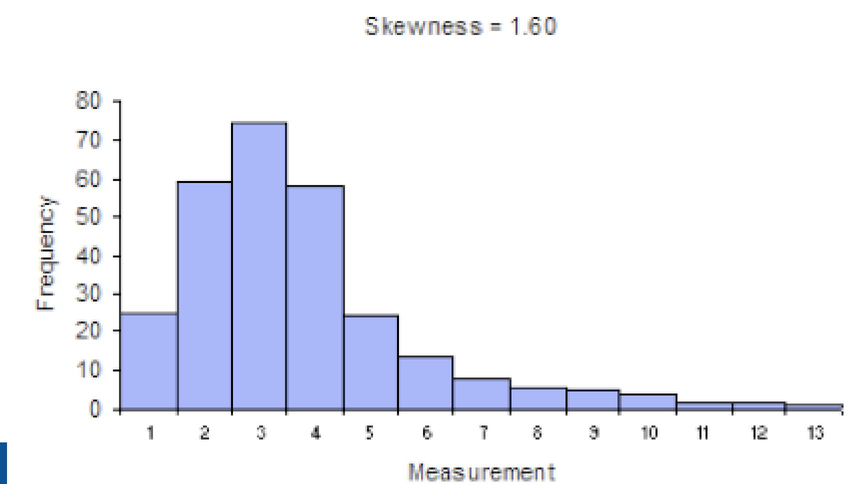



Age of Death
-1.3225.


Distribution of Household Incomes
2.0043.


Distribution of Male Heights
0.0013.


Skewness = -1.60
Frequency
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
Measurement


Skewness = 1.60
Frequency
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
Measurement

Measures of Relative Position: Z-Score

- Z-score to Compare the Variation in Different Populations
- Charlie got a mark of 85 on a math test which had a mean of 75 and a standard deviation of 5. Daisy got a mark of 75 on an English test which had a mean of 69 and a standard deviation of 2. Relative to their respective mean and standard deviation, who got the better grade?

$$Z_X = \frac{x - \bar{x}}{s}$$

Where:

Z_x = *Z score*

x = *to the data value*

$\bar{x}$ = *mean of the data set*

s = *standard deviation of the data set (*

$$Z_{Ch} = \frac{x-\mu}{\sigma} = \frac{85-75}{5} = \frac{10}{5} = 2$$

$$Z_D = \frac{x-\mu}{\sigma} = \frac{75-69}{2} = \frac{6}{2} = 3$$

Charlie got a test mark 2 standard deviations higher than the mean of the class, while Daisy got a mark that is 3 standard deviations higher than the mean in her class. Therefore, proportionally speaking, Daisy did better within her class in comparison to Charlie.

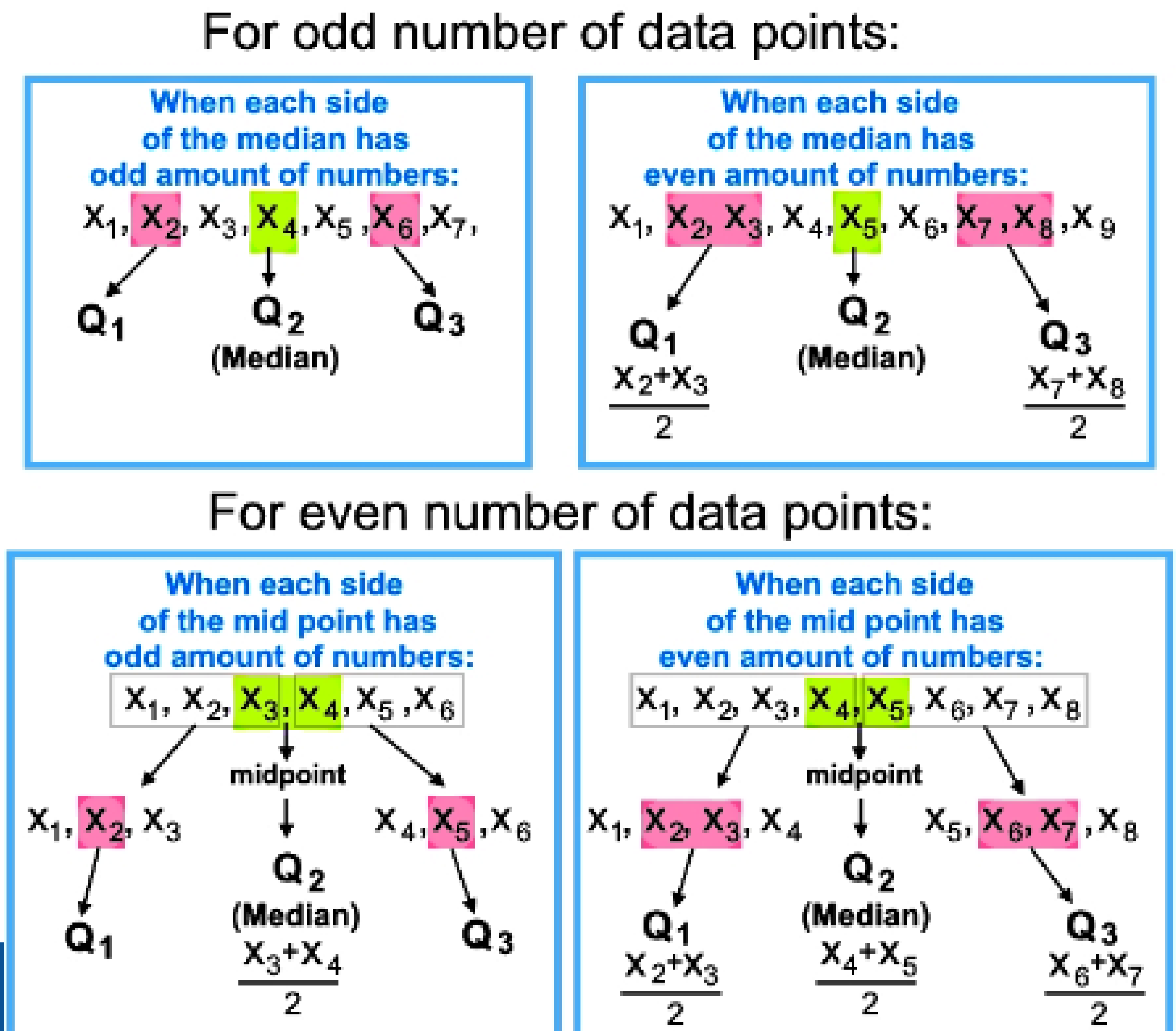
Measures of Relative Position: Quartiles.

- dividing the data distribution into four parts, where each quartile is the specific point marking the division between the first quarter and the second, the second quarter and the third or the third quarter and the fourth.
- **Where:**
  Q1 = splits the lowest 25% of the sorted data
  Q2 = Median=splits the lowest 50% of the sorted data
  Q3 = splits the lowest 75% of the sorted data
- Find the quartiles for each data set: **{9, 3, 7, 5, 2, 8, 12}**

{2, 3, 5, **7**, 8, 9, 12} = {2, **3**, 5, **7**, 8, **9**, 12}

*Q*2=7, *Q*1=3 and *Q*3=9.

Process to calculate quartiles



Measures of Relative Position: Percentiles

- Percentiles divide the whole data set into a hundred equal parts

$$\text{Percentile of } X = \frac{\textit{number of data points less than } X}{\textit{total number of data points}} \times 100$$

- Sidney is taking a biology course in university. She got a mark of 78% and the list of all marks from her class (including her mark) is given by {56, 83, 74, 67, 47, 54, 82, 78, 86, 90}.What percentile did she score in?

First we order the scores from lowest to highest: {47, 54, 56, 67, 74, **78**, 82, 83, 86, 90}.

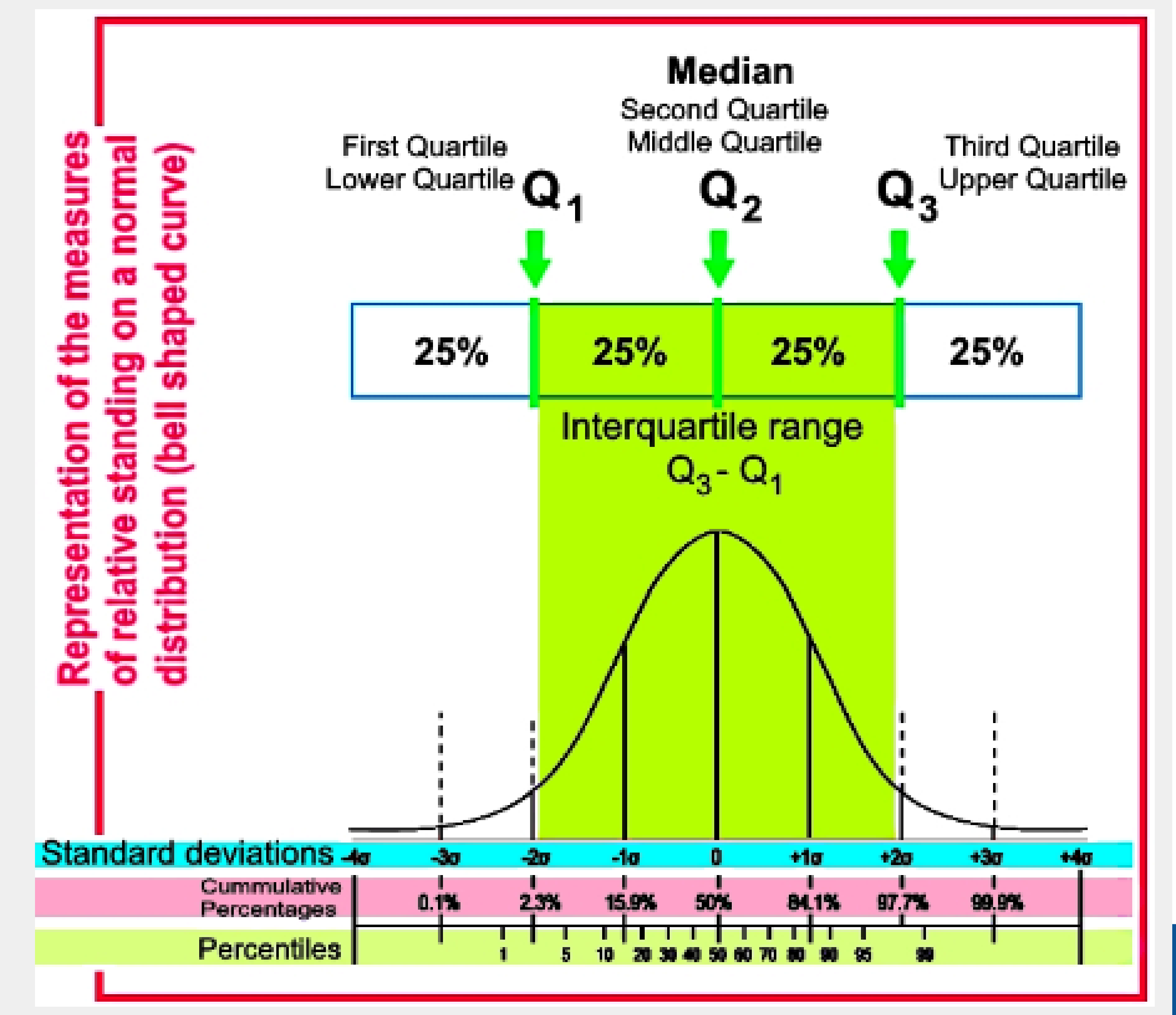
$$\text{Percentile for Sidney's score} = \frac{5}{10} \times 100 = 50$$

- Sidneys friend Billy knows he got in the 70% percentile, what was his mark?

$$\frac{\textit{Percentile of } X \times \textit{total number of data points}}{100} = \text{number of data points less than } X$$

$$\frac{70 \times 10}{100} = 7$$

Measures of Relative Position: Quartiles, Percentiles.



Box Plot

Example……

$\frac{11}{2} = 5.5$

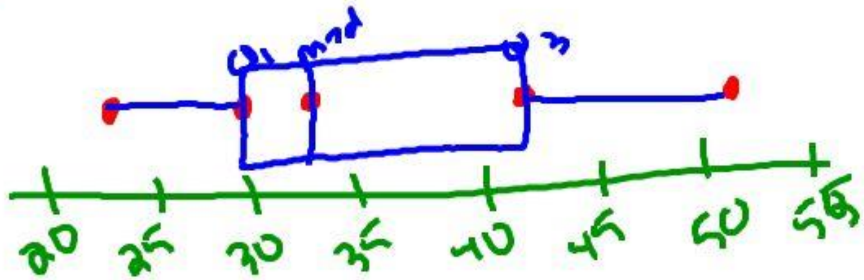
- Draw a box plot of the following data.

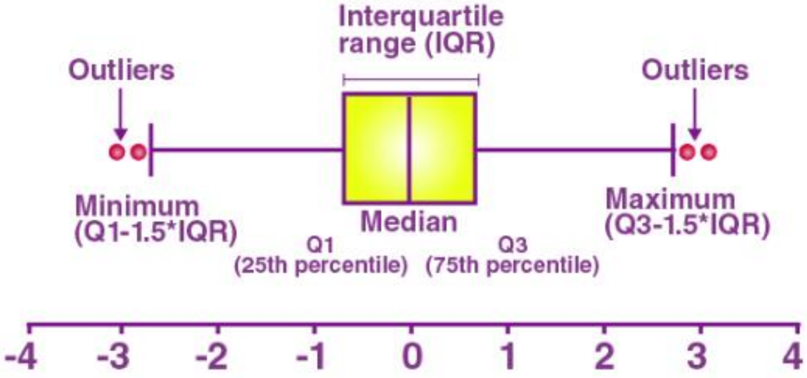
33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31

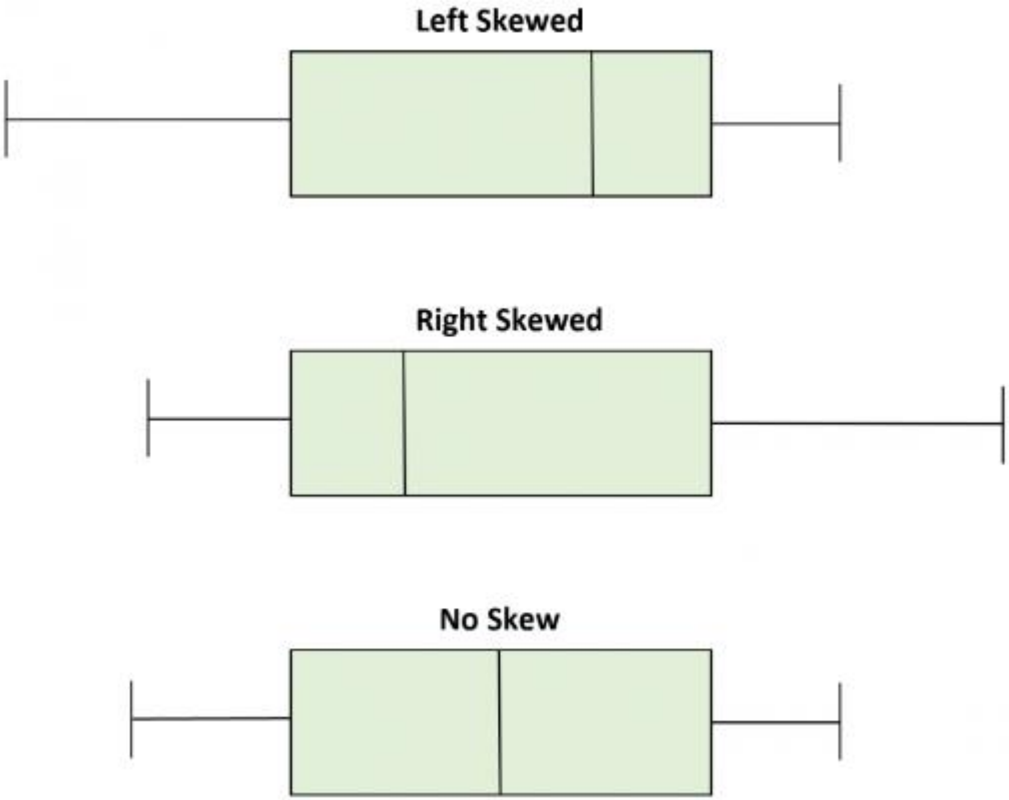
23, 27, 29, 30, 31, $\boxed{33}$ 38, 40, 42, 43, 51

min Q1 med Q3 max

min = 23
Q1 = 29
median = 33
Q3 = 42
Max = 51

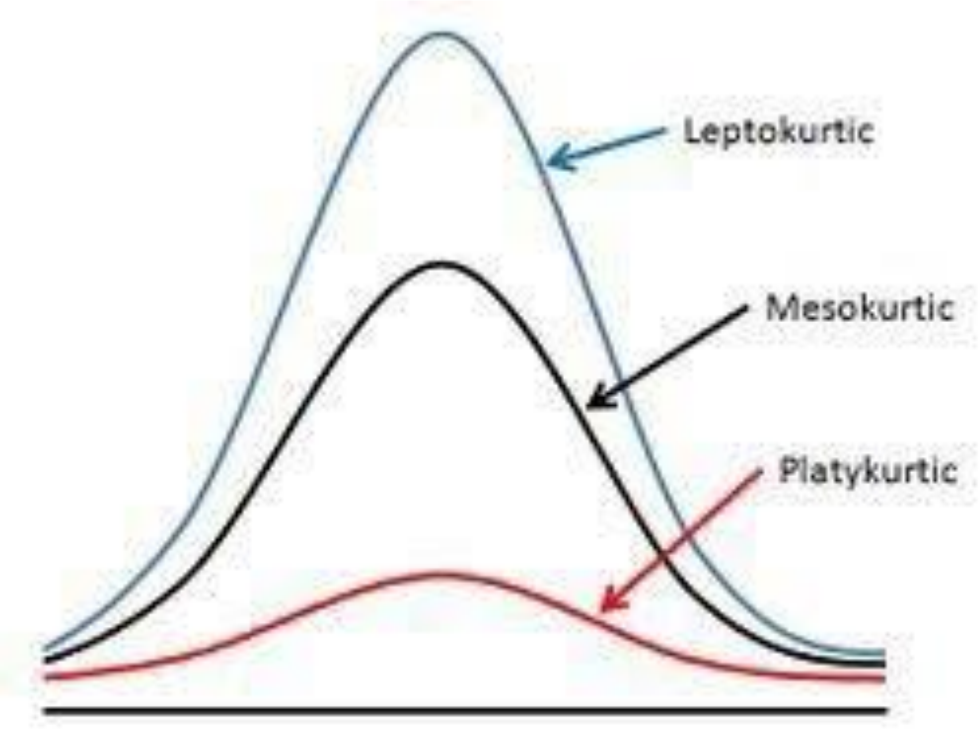


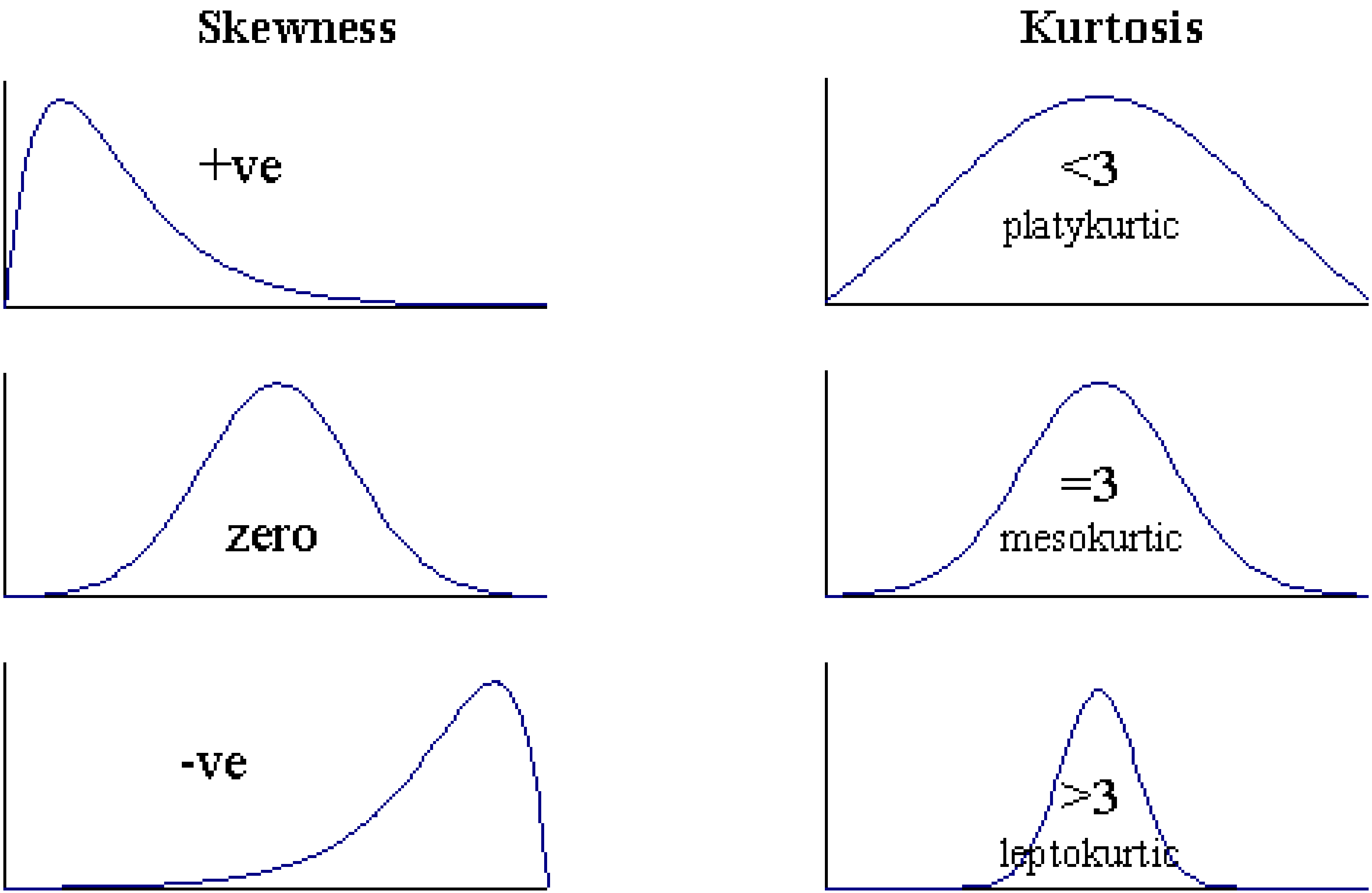

Interquartile
range (IQR)
Outliers
Outliers
Minimum
(Q1-1.5*IQR)
Median
Maximum
(Q3-1.5*IQR)
Q1
(25th percentile)
Q3
(75th percentile)
-4
-3
-2
-1
0
1
2
3
4


Left Skewed
Right Skewed
No Skew

Kurtosis

- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.
- Leptokurtic **(Kurtosis > 3),** Platykurtic (Kurtosis < 3), Mesokurtic (Kurtosis = 3)




Skewness
Kurtosis
+ve
<3
platykurtic
zero
=3
mesokurtic
-ve
>3
leptokurtic

Stem-and-Leaf Plot

- A stem-and-leaf plot is a way of organizing data values from least to greatest using place value.
- This type of graph uses a “stem” as the leading part of a data value and a “leaf” as the remaining part of the value.

20	12	39	38	18	58	49	59	66	50
23	32	43	53	67	35	29	13	42	55
37	19	38	22	46	71	9	65	15	38

Stem	Leaf
0	9
1	2, 8, 3, 9, 5
2	0, 3, 9, 2
3	9, 8, 2, 5, 7, 8, 8
4	9, 3, 2, 6
5	8, 9, 0, 3, 5
6	6, 7, 5
7	1

Stem	Leaf
0	9
1	2, 3, 5, 8, 9
2	0, 2, 3, 9
3	2, 5, 7, 8, 8, 8, 9
4	2, 3, 6, 9
5	0, 3, 5, 8, 9
6	5, 6, 7
7	1

Median - 38
Mode - 38

- Construct a stem-and-leaf plot to represent the data, and list 3 facts that you know about the growth of the plants.
-

18	10	37	36	61
39	41	49	50	52
57	53	51	57	39
48	56	33	36	19
30	41	51	38	60

Stem	Leaf
1	0, 8, 9
2	
3	0, 3, 6, 6, 7, 8, 9, 9
4	1, 1, 8, 9
5	0, 1, 1, 2, 3, 6, 7, 7
6	0, 1

- From the stem-and-leaf plot, the growth of the plants ranged from a minimum of 10 cm to a maximum of 61 cm.
- The median of the data set is the value in the 13th position, which is 41 cm.
- There was no growth recorded in the class of 20 cm, so there is no number in the leaf row.
- The data set is multimodal.

- Use the stem-and-leaf plot below to answer the following questions:
  - What is the mode of the data set?
  - What is the median of the data set?
  - How many of the data values are greater than 40?
  - What percentage of the data values are less than 40?

Stem	Leaf
2	3, 4, 5, 6, 7
3	1, 1, 5, 6, 8, 9, 9
4	0, 0, 0, 0, 9, 9, 9
5	5, 5, 7, 8, 8, 9
6	0, 1, 2, 3
7	2, 2, 3, 3, 4, 5, 5, 6
8	0, 1, 2

Two-Sided Stem-and-Leaf Plots

- The girls and boys in one of BDF High School's AP English classes are having a contest. They want to see which group can read the most number of books. Mrs. Stubbard, their English teacher, says that the class will tally the number of books each group has read, and the highest mode will be the winner. The following data was collected for the first semester of AP English:

Girls	11	12	12	17	18	23	23	23	24	33	34	35	44	45	47	50	51	51
Boys	15	18	22	22	23	26	34	35	35	35	40	40	42	47	49	50	50	51

- Draw a two-sided stem-and-leaf plot for the data.
- Determine the mode for each group.
- Help Mrs. Stubbard decide which group won the contest.

Girls		Boys
8, 7, 2, 2, 1	1	5, 8
3, 3, 3, 2	2	2, 2, 3, 6
5, 4, 3	3	4, 5, 5, 5
7, 5, 4	4	0, 0, 2, 7, 9
1, 1, 0	5	0, 0, 1

Types of Analysis

- Univariate analysis:

Univariate analysis is to simply describe the data to find patterns within the data

- Bivariate Analysis

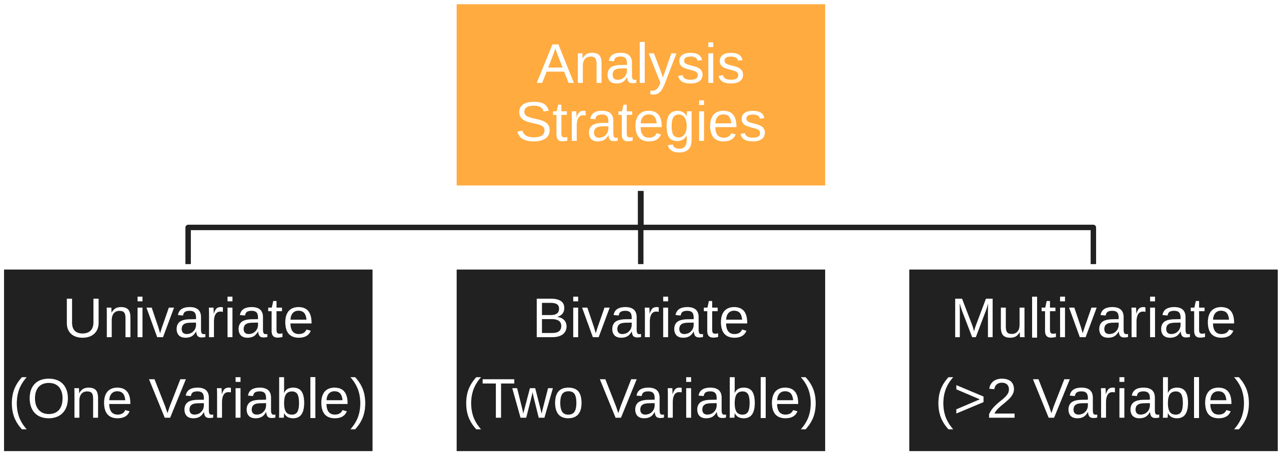
In a survey of a classroom, the researcher may be looking to analysis the ratio of students who scored above 85% corresponding to their genders. In this case, there are two variables – gender = X (independent variable) and result = Y (dependent variable).

- Multivariate Analysis

A doctor has collected data on cholesterol, blood pressure, and weight. She also collected data on the eating habits of the subjects (e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed per week). She wants to investigate the relationship between the three measures of health and eating habits?

- Univariate analysis:
  - Frequency Distribution Tables
  - Histograms
  - Frequency Polygons
  - Pie Charts
  - Bar Charts
- Bivariate Analysis
  - Correlation coefficients
  - Regression analysis
- Multivariate Analysis
  - Factor Analysis
  - •Cluster Analysis
  - •Variance Analysis
  - •Discriminant Analysis
  - •Multidimensional Scaling
  - •Principal Component Analysis
  - Redundancy Analysis

Analysis Strategies



Univariate - Categorical Data

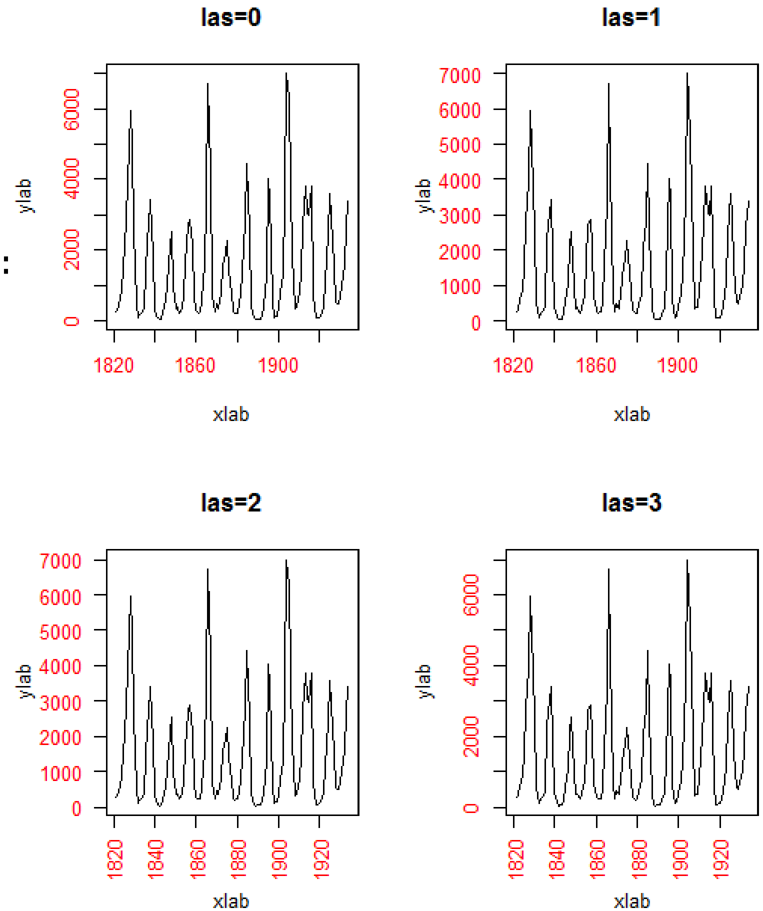
- *table()* command allows us to look at data in table form
- factor() command classify data into various levels or factors.

```
> x=c("Yes","No","No","Yes","Yes")
> table(x)
x
 No Yes
  2   3
> factor(x)
[1] Yes No  No  Yes Yes
Levels: No Yes
```

Plot a Graph

- plot(lynx, main=“GraphTitle", col.main: el", ylab="ylabel",pch=12, col=75)

  - main:Graph title
  - col.main:color for title
  - cex.main:title letter size
  - las: Axis label orientation
  - xlab: X axis label
  - ylab: Y axis label
  - Pch:plot character
  - col:color for datapoint



- parallel to the axis (the default, 0)
- horizontal (1)
- perpendicular to the axis (2)
- vertical (3)

Colors

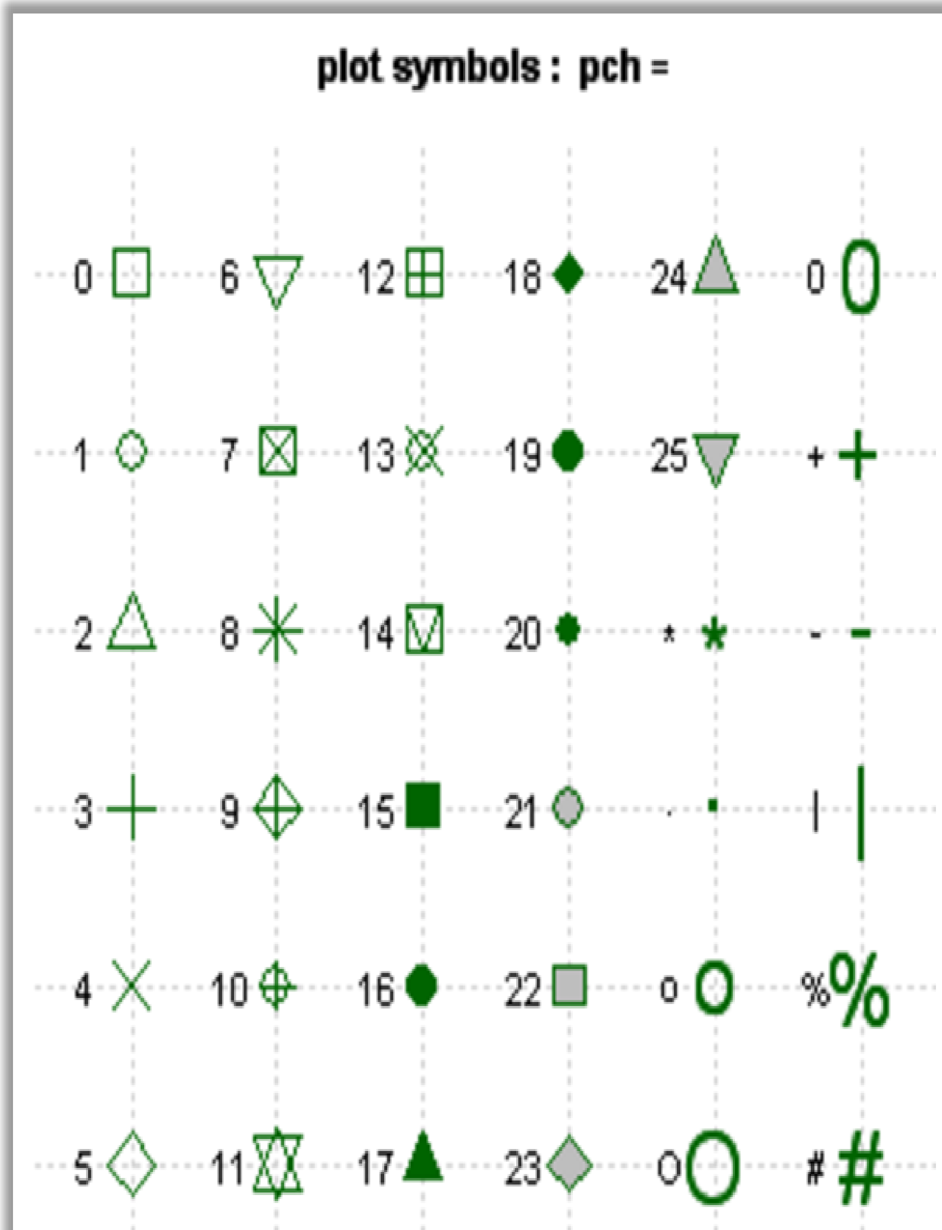
- colors( )
- Total number of available colors: 657

- Option I: Type in the “code” of the color
  - col=50
- Option II: Type in the “name” of the color
  - col="red"
- Option III: Type in the “alpha numeric code” of the color
  - col="seagreen3"

```
> colors( )
  [1] "white"             "aliceblue"         "antiquewhite"      "antiquewhite1"
  [5] "antiquewhite2"     "antiquewhite3"     "antiquewhite4"     "aquamarine"
  [9] "aquamarine1"       "aquamarine2"       "aquamarine3"       "aquamarine4"
 [13] "azure"             "azure1"            "azure2"            "azure3"
 [17] "azure4"            "beige"             "bisque"            "bisque1"
 [21] "bisque2"           "bisque3"           "bisque4"           "black"
 [25] "blanchedalmond"    "blue"              "blue1"             "blue2"
 [29] "blue3"             "blue4"             "blueviolet"        "brown"
 [33] "brown1"            "brown2"            "brown3"            "brown4"
 [37] "burlywood"         "burlywood1"        "burlywood2"        "burlywood3"
 [41] "burlywood4"        "cadetblue"         "cadetblue1"        "cadetblue2"
 [45] "cadetblue3"        "cadetblue4"        "chartreuse"        "chartreuse1"
 [49] "chartreuse2"       "chartreuse3"       "chartreuse4"       "chocolate"
 [53] "chocolate1"        "chocolate2"        "chocolate3"        "chocolate4"
 [57] "coral"             "coral1"            "coral2"            "coral3"
 [61] "coral4"            "cornflowerblue"    "cornsilk"          "cornsilk1"
 [65] "cornsilk2"         "cornsilk3"         "cornsilk4"         "cyan"
 [69] "cyan1"             "cyan2"             "cyan3"             "cyan4"
 [73] "darkblue"          "darkcyan"          "darkgoldenrod"     "darkgoldenrod1"
 [77] "darkgoldenrod2"    "darkgoldenrod3"    "darkgoldenrod4"    "darkgray"
 [81] "darkgreen"         "darkgrey"          "darkkhaki"         "darkmagenta"
 [85] "darkolivegreen"    "darkolivegreen1"   "darkolivegreen2"   "darkolivegreen3"
 [89] "darkolivegreen4"   "darkorange"        "darkorange1"       "darkorange2"
 [93] "darkorange3"       "darkorange4"       "darkorchid"        "darkorchid1"
 [97] "darkorchid2"       "darkorchid3"       "darkorchid4"       "darkred"
[101] "darksalmon"        "darkseagreen"      "darkseagreen1"     "darkseagreen2"
[105] "darkseagreen3"     "darkseagreen4"     "darkslateblue"     "darkslategray"

[621] "tan1"              "tan2"              "tan3"              "tan4"
[625] "thistle"           "thistle1"          "thistle2"          "thistle3"
[629] "thistle4"          "tomato"            "tomato1"           "tomato2"
[633] "tomato3"           "tomato4"           "turquoise"         "turquoise1"
[637] "turquoise2"        "turquoise3"        "turquoise4"        "violet"
[641] "violetred"         "violetred1"        "violetred2"        "violetred3"
[645] "violetred4"        "wheat"             "wheat1"            "wheat2"
[649] "wheat3"            "wheat4"            "whitesmoke"        "yellow"
[653] "yellow1"           "yellow2"           "yellow3"           "yellow4"
[657] "yellowgreen"
>
```

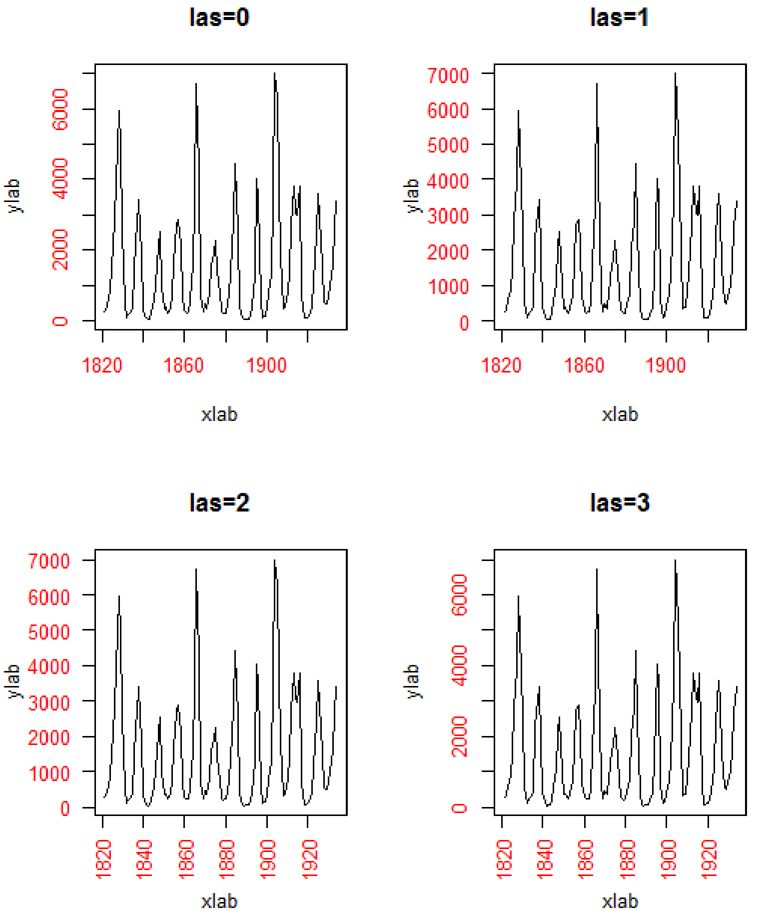
pch Symbols Used in R

- pch: plotting ‘character’
- To get help on this
  - ?pch



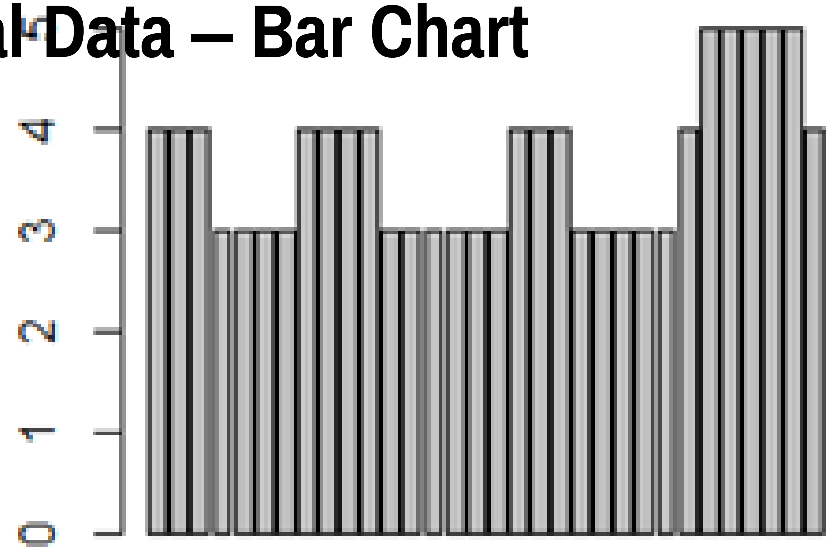
- Generate plots
  - x=2:6
  - plot(x, pch="d")
  - plot(x, pch=14)

Four Graphs in a Window

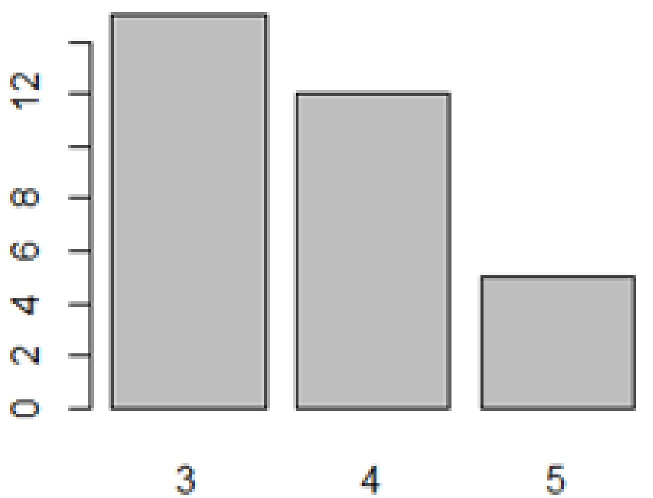


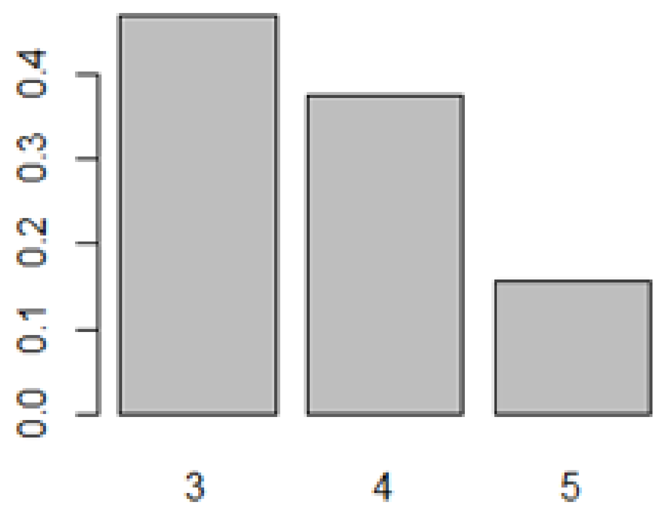
- Four Graphs in a Window
  - par(mfrow=c(2,2), col.axis="red")
  - plot(lynx, las=0, xlab="xlab", ylab="ylab", main="las=0")
  - plot(lynx, las=1, xlab="xlab", ylab="ylab", main="las=1")
  - plot(lynx, las=2, xlab="xlab", ylab="ylab", main="las=2")
  - plot(lynx, las=3, xlab="xlab", ylab="ylab", main="las=3")
- One Graph in a Window
  - par(mfrow=c(1,1), col.axis="black")

Univariate - Categorical Data – Bar Chart



data=mtcars$gear
#Barplot
barplot(data)
#Barplotwith Catagory
barplot(table(data))





Univariate - Categorical Data – Pie Chart

```
data.counts = table(data)
pie(data.counts)
names(data.counts) = c("Three Gear","Four Gear",
                       "Five Gear")
pie(data.counts,col=c("purple","green2","cyan","white"))
```

Univariate - Numerical Data

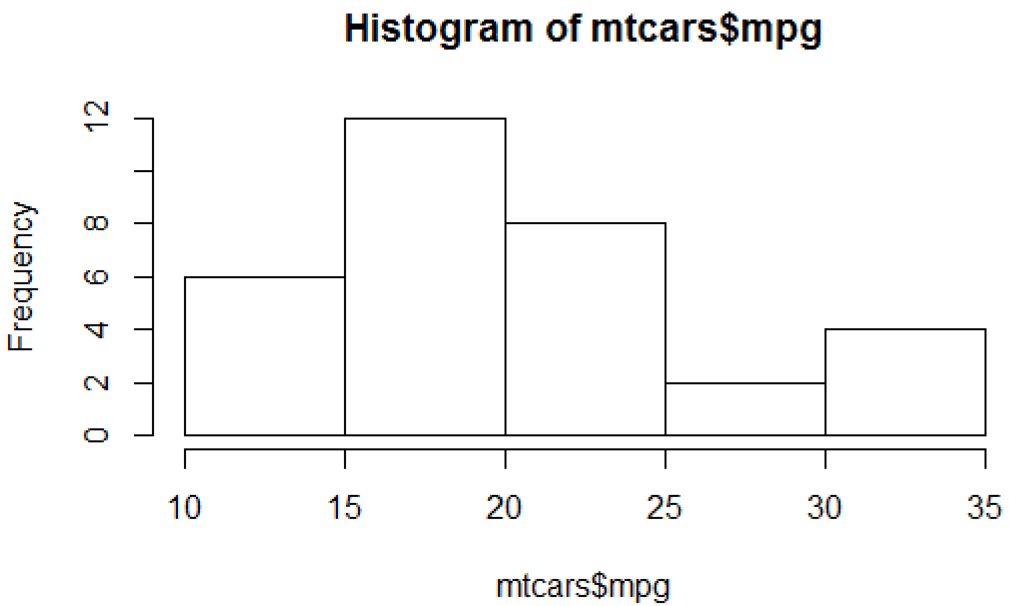
```
> mtcars$mpg
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4
[13] 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3
[25] 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4
> summary(mtcars$mpg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  10.40   15.43   19.20   20.09   22.80   33.90 
```

Univariate - Numerical Data to Category Data

```
> cats = cut(mtcars$mpg,breaks=c(10,15,20,max(mtcars$mpg)))
> cats
 [1] (20,33.9] (20,33.9] (20,33.9] (20,33.9] (15,20]   (15,20]
 [7] (10,15]   (20,33.9] (20,33.9] (15,20]   (15,20]   (15,20]
[13] (15,20]   (15,20]   (10,15]   (10,15]   (10,15]   (20,33.9]
[19] (20,33.9] (20,33.9] (20,33.9] (15,20]   (15,20]   (10,15]
[25] (15,20]   (20,33.9] (20,33.9] (20,33.9] (15,20]   (15,20]
[31] (10,15]   (20,33.9]
Levels: (10,15] (15,20] (20,33.9]
> table(cats)
cats
  (10,15]   (15,20] (20,33.9]
        6        12        14
> levels(cats)
[1] "(10,15]"   "(15,20]"   "(20,33.9]"
```

Univariate - Numerical Data – Histogram

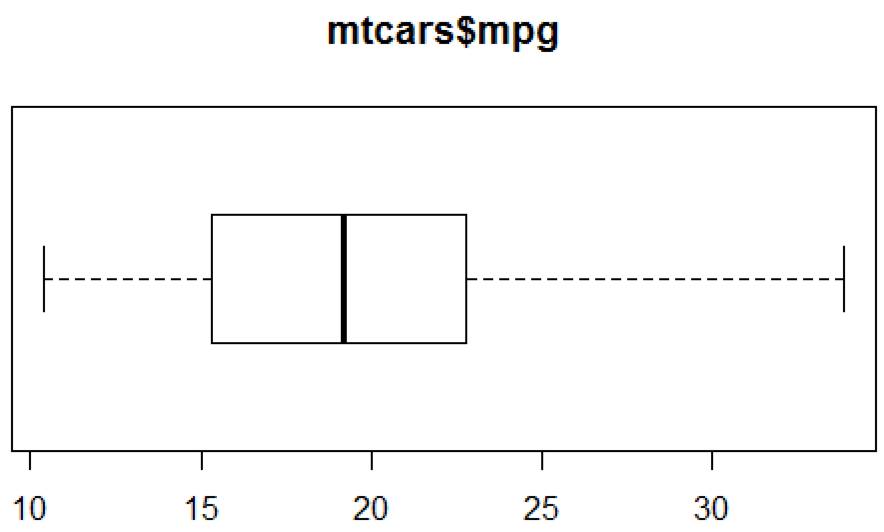
```
hist(mtcars$mpg)
hist(mtcars$mpg,probability=TRUE)
hist(mtcars$mpg,breaks=c(10,15,20,max(mtcars$mpg),probability=T))
```



Univariate - Numerical Data – Boxplot

- 5-number summary -the lower hinge (basically Q1), the Median, the upper hinge (basically Q3) and whiskers which extend to the min and max.

```
boxplot(mtcars$mpg, main="mtcars$mpg",horizontal=TRUE)
summary(mtcars$mpg)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
10.40   15.43   19.20   20.09   22.80   33.90
```



Random Number Generation

```
> rnorm(10, mean = 0, sd = 1)
 [1]  0.6630510 -0.6494815 -1.5608564  0.7608942  0.2414662
 [6]  0.8645266  1.9450196  1.1489713 -0.8341809 -0.5650081
> rnorm(10, mean = 2, sd = 1)
 [1] 2.0149087 3.3166223 3.0229688 2.0763079 2.3702358 2.2968198
 [7] 1.5225275 0.6924751 4.4802255 2.8191071
> rnorm(10, mean = 2, sd = 2)
 [1]  6.4804102  0.8276148  0.8038300  1.3979884 -0.3607730
 [6]  1.3823101 -1.0899623  5.6021054  2.5985916  2.5312193
> rnorm(10, mean = 2, sd = 0)
 [1] 2 2 2 2 2 2 2 2 2 2
```

Exercises

- Create two variables X1 and X2 with 100 random numbers using normal distributions.
- Create two different histograms for two variables X1 & X2. Do you get the same histogram?
- Create Box plot for two variables X1 & X2. Do you get the same plot?

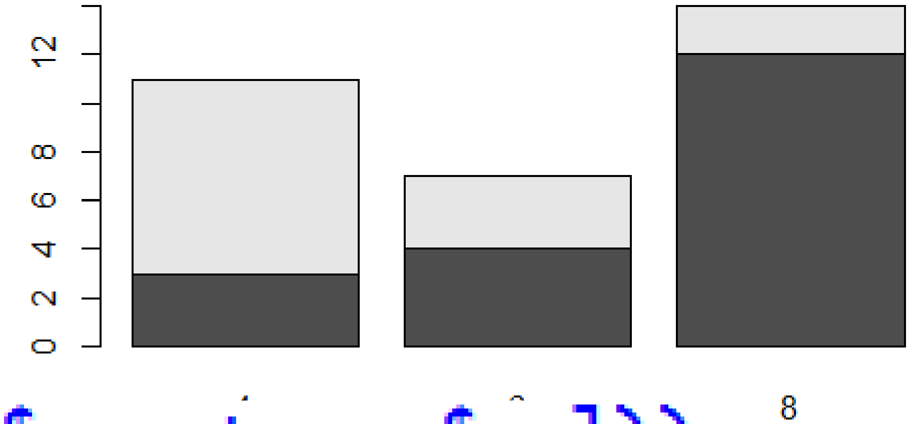
Bivariate - Categorical Data

- The relationship between 2 variables

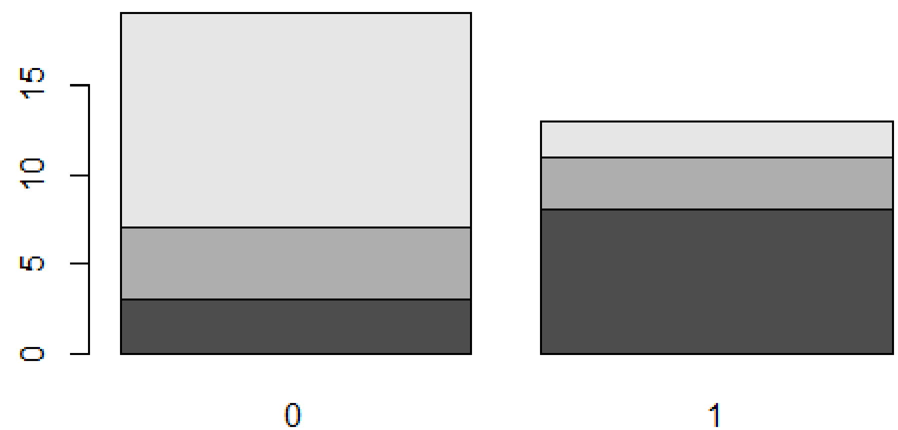
```
> compare=table(mtcars$am,mtcars$cyl)
> old.digits = options("digits")
> options(digits=3)
> compare

     4  6  8
  0  3  4 12
  1  8  3  2
> prop.table(compare)

         4      6      8
  0 0.0938 0.1250 0.3750
  1 0.2500 0.0938 0.0625
> prop.table(compare,1)

        4     6     8
  0 0.158 0.211 0.632
  1 0.615 0.231 0.154
> prop.table(compare,2)

        4     6     8
  0 0.273 0.571 0.857
  1 0.727 0.429 0.143
```

Bivariate - Categorical Data –Bar Plot

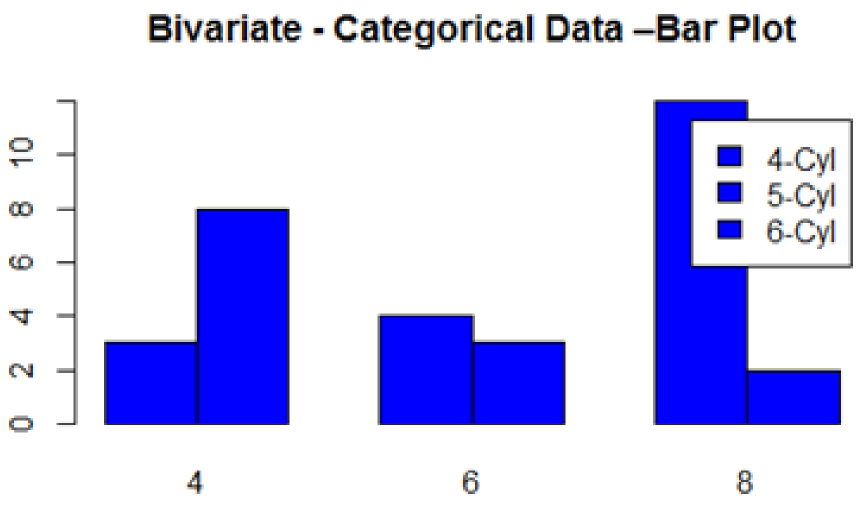
```
> compare

     4  6  8
  0  3  4 12
  1  8  3  2
```

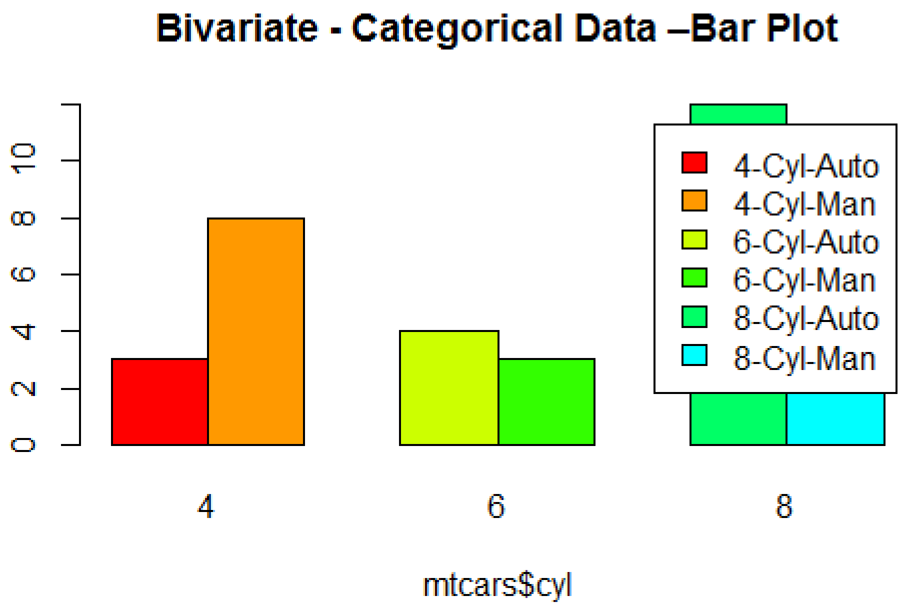



```
> barplot(table(mtcars$am,mtcars$cyl))
> barplot(table(mtcars$cyl,mtcars$am))
```



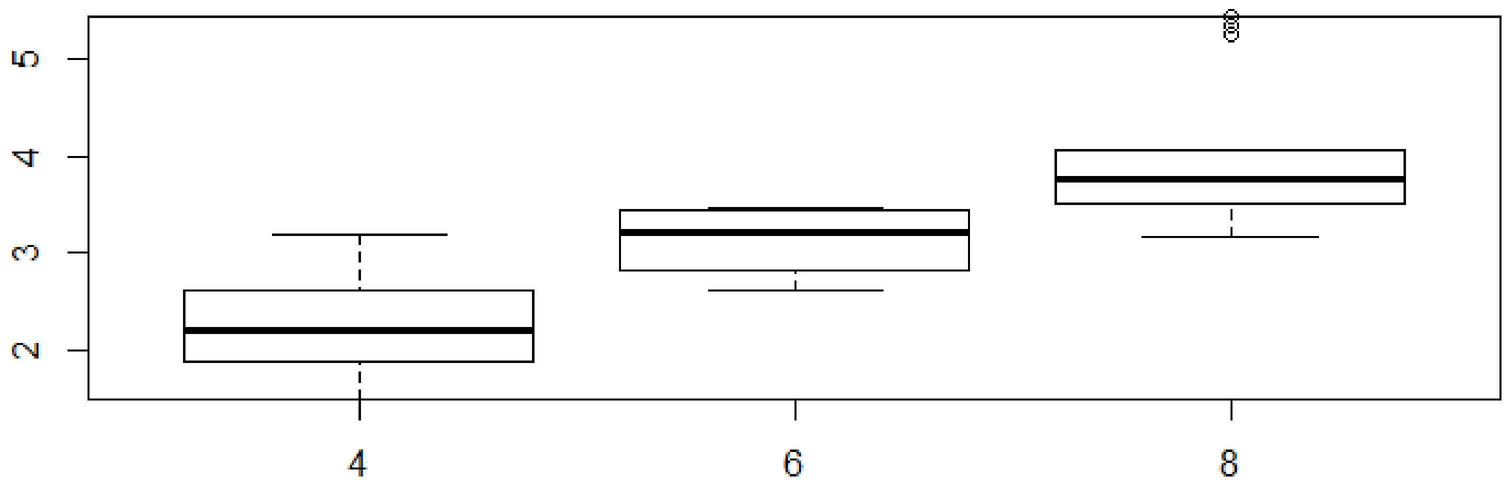
Try This....


Bivariate - Categorical Data –Bar Plot
0
2
4
6
8
10
4-Cyl
5-Cyl
6-Cyl
4
6
8

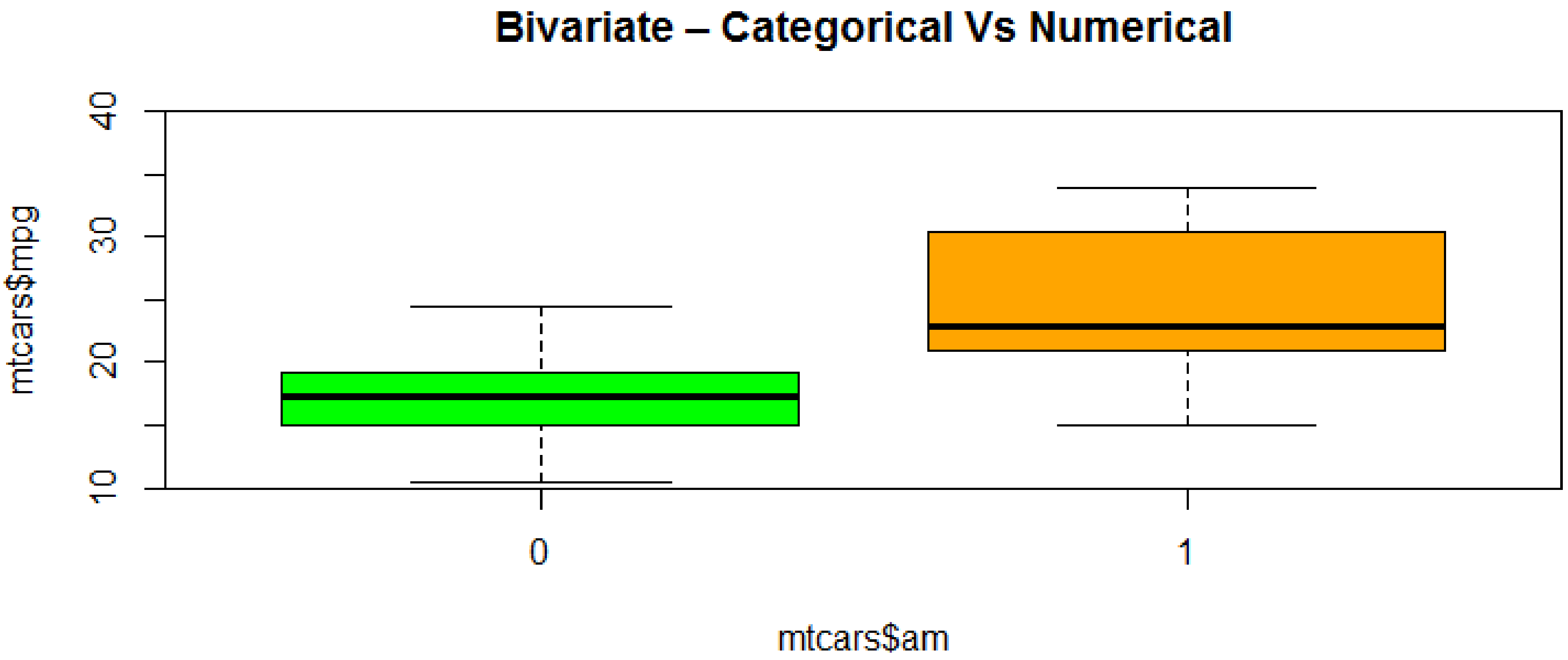

Bivariate - Categorical Data –Bar Plot
0
2
4
6
8
10
4-Cyl-Auto
4-Cyl-Man
6-Cyl-Auto
6-Cyl-Man
8-Cyl-Auto
8-Cyl-Man
4
6
8
mtcars$cyl

Bivariate – Categorical Vs Numerical

- boxplot(mtcars$wt ~ mtcars$cyl)

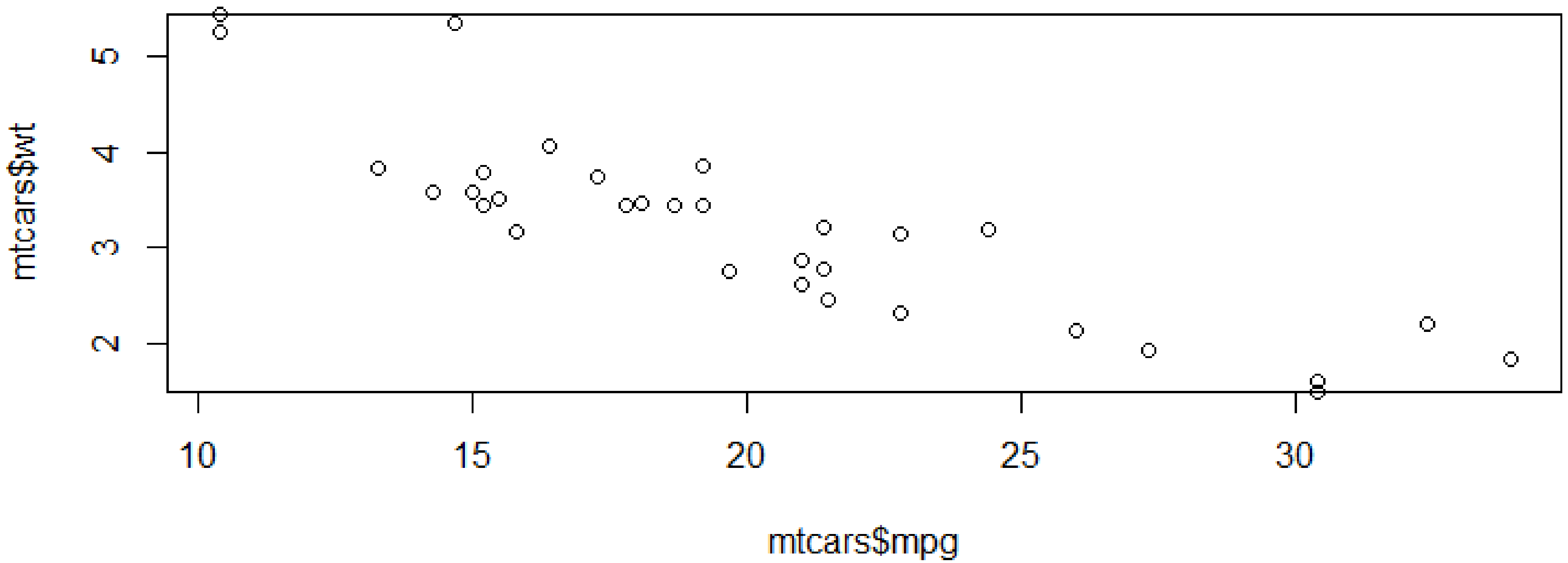


Try This.....

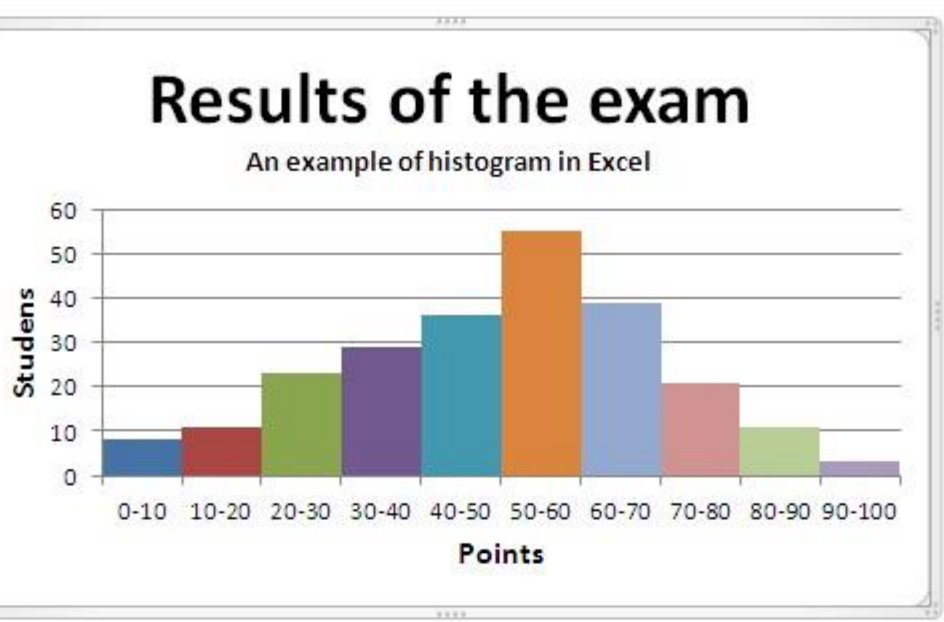

Bivariate – Categorical Vs Numerical
mtcars$mpg
10
20
30
40
0
1
mtcars$am

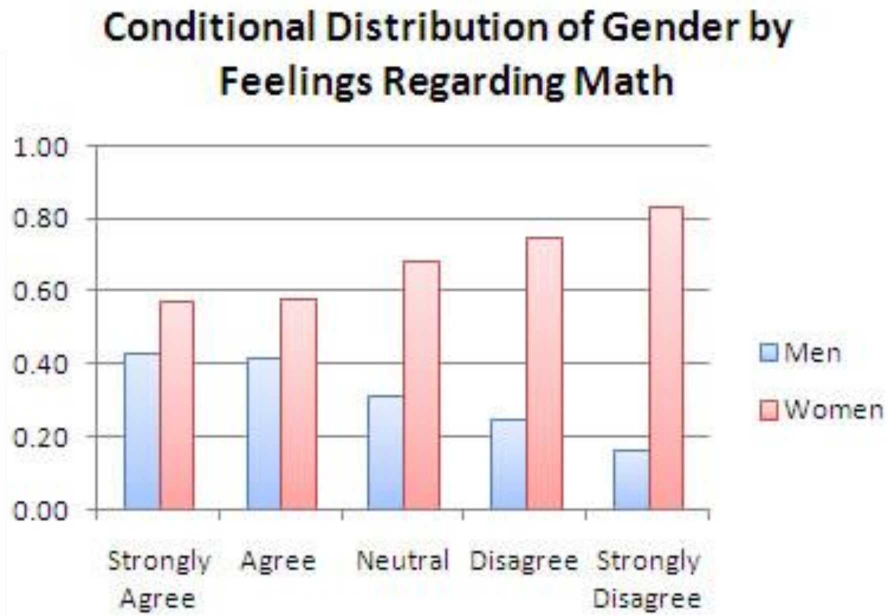
Bivariate – Numerical Vs Numerical

- plot(mtcars$mpg,mtcars$wt)



- Find a graph of Univariate and Bivariate data from the newspaper or other media source. Use R to generate a similar graphs.
  - Histogram, Piechart, Barchart, Boxplot(Univariate)
  - Barchart, Boxplot, Scatterplot (Bivariate)





Exercises

This exercise involves the ***Boston*** housing data set. The Boston data set is part of the MASS library in R.

- How many rows are in this data set? How many columns? What do the rows and columns represent?
- Make some pair wise scatter plots of the predictors (columns) in this data set. Describe your findings.
- Are any of the predictors associated with per capita crime rate? If so, explain the relationship.
- Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.
- How many of the suburbs in this data set bound the Charles river?

- library("MASS")
- ?Boston
- Boston[Boston$crim==max(Boston$crim),]
- boxplot(Boston$crim)
- boxplot(Boston$tax)
- boxplot(Boston$ptratio)
- nrow(Boston[Boston$chas==1,])

Multivariate Data - Dataframe

```
> weight = c(150, 135, 210, 140)
> height = c(65, 61, 70, 65)
> gender = c("Fe","Fe","M","Fe")
> study = data.frame(weight,height,gender) # make the data frame
> study
  weight height gender
1    150     65     Fe
2    135     61     Fe
3    210     70      M
4    140     65     Fe
> study = data.frame(w=weight,h=height,g=gender)
> study
    w  h  g
1 150 65 Fe
2 135 61 Fe
3 210 70  M
4 140 65 Fe
> row.names(study)<-c("Mary","Alice","Bob","Judy")
> study
        w  h  g
Mary  150 65 Fe
Alice 135 61 Fe
Bob   210 70  M
Judy  140 65 Fe
```

Install Packages in R

- R packages provide a powerful mechanism for extending the functionality of R
- R packages are obtained from CRAN or other repositories
- The install.packages() can be used to install packages at the R console

install.packages("MASS")
~~***install.packages(c("slidify", "ggplot2", "devtools"))***~~

- The library() function loads the intalled packages to access the functionality of the package

library(MASS)

Multivariate Data

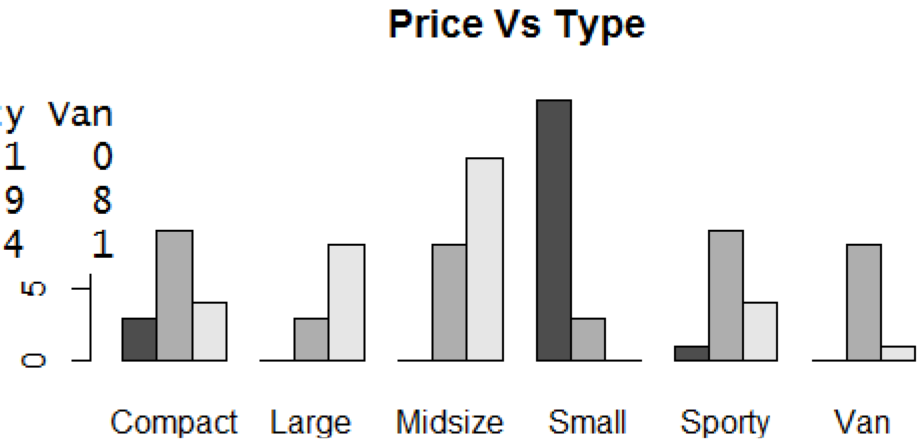
- ?Cars93
- attach(Cars93)

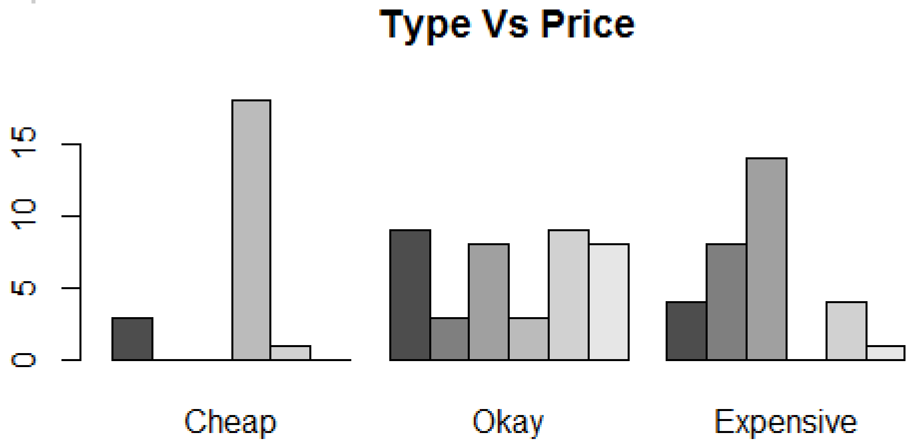
```
> Newprice=cut(Price,c(0,12,20,max(Price)))
> levels(Newprice)=c("Cheap","Okay", "Expensive")
> Newmpg=cut(MPG.highway,c(0,20,30,max(MPG.highway)))
> levels(Newmpg)=c("Gas Guzzler","Okay", "Miser")
> table(Type)
Type
Compact   Large Midsize   Small  Sporty     Van
     16      11      22      21      14       9
> table(Newprice,Type)
           Type
Newprice    Compact Large Midsize Small Sporty Van
  Cheap           3     0       0    18      1   0
  Okay            9     3       8     3      9   8
  Expensive       4     8      14     0      4   1
```

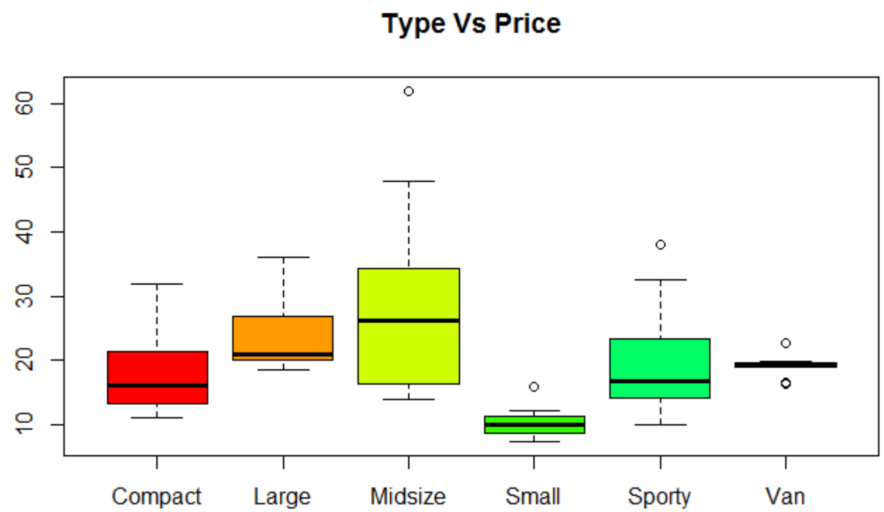
```
> table(Newprice,Type,Newmpg)
, , Newmpg = Gas Guzzler

           Type
Newprice    Compact Large Midsize Small Sporty Van
  Cheap           0     0       0     0      0   0
  Okay            0     0       0     0      0   2
  Expensive       0     0       0     0      0   0

, , Newmpg = Okay

           Type
Newprice    Compact Large Midsize Small Sporty Van
  Cheap           1     0       0     4      0   0
  Okay            5     3       6     0      6   6
  Expensive       4     8      14     0      4   1

, , Newmpg = Miser

           Type
Newprice    Compact Large Midsize Small Sporty Van
  Cheap           2     0       0    14      1   0
  Okay            4     0       2     3      3   0
  Expensive       0     0       0     0      0   0
```

```
> #Barplot-Price Vs Type
> barplot(table(Newprice,Type),beside=T,main="Price Vs Type")
> #Barplot-Type Vs Price
> barplot(table(Type,Newprice),beside=T,main="Type Vs Price")
> #Boxplot - Type Vs Price
> boxplot(Price~Type,data=Cars93,col=rainbow(10),main="Type Vs Price")
```

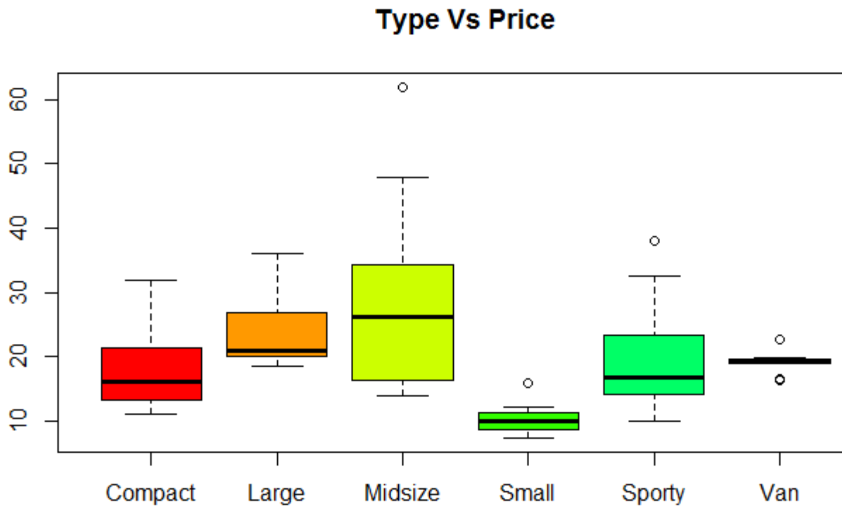
```
> table(Newprice,Type)
           Type
Newprice     Compact Large Midsize Small Sporty Van
  Cheap            3     0       0    18      1   0
  Okay             9     3       8     3      9   8
  Expensive        4     8      14     0      4   1
```



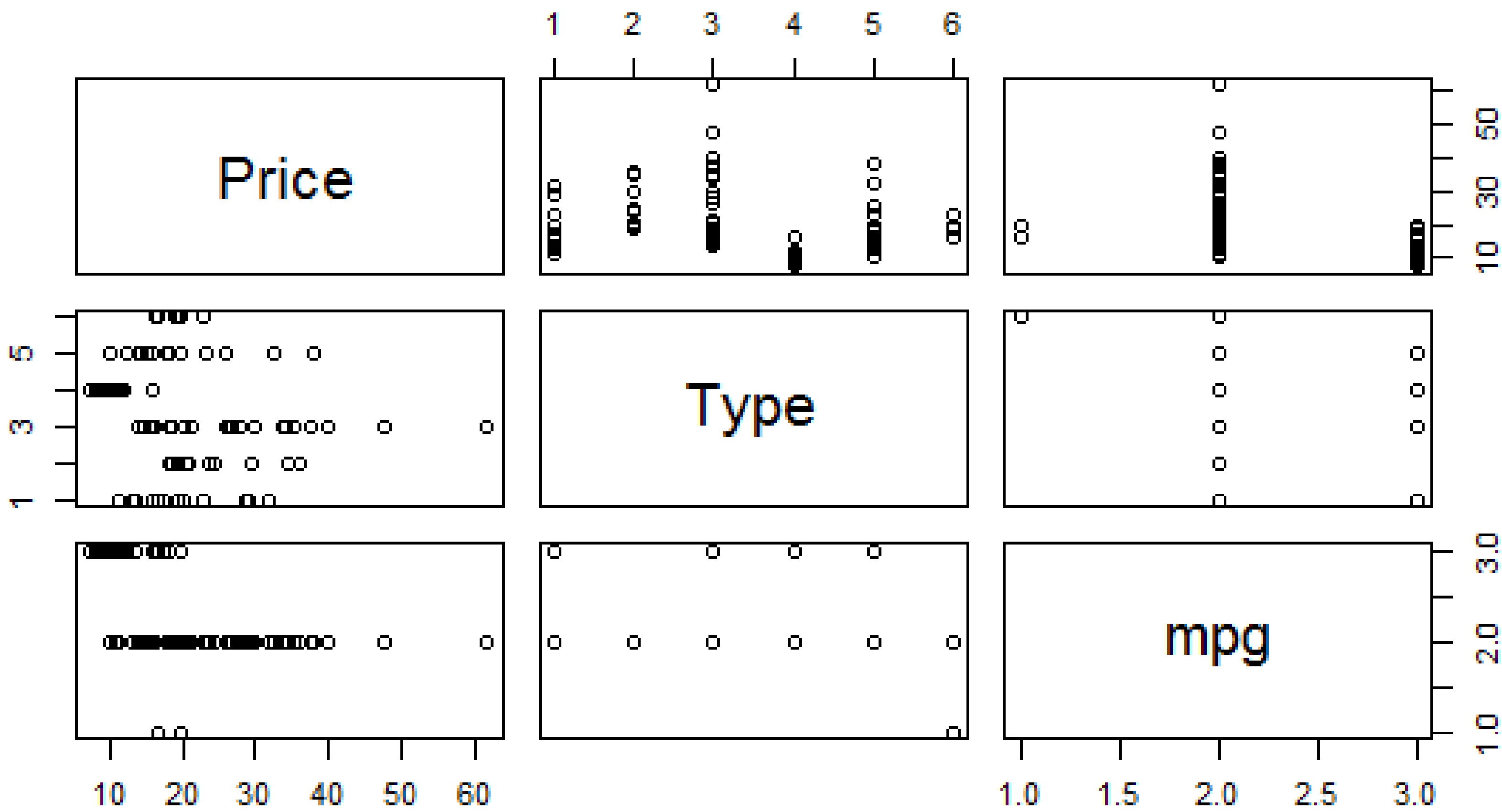




```
> tapply(Price, Type, summary)
$`Compact`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.10   13.38   16.15   18.21   20.68   31.90

$Large
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.40   20.00   20.90   24.30   26.95   36.10

$Midsize
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.90   16.77   26.20   27.22   34.20   61.90

$Small
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.40    8.60   10.00   10.17   11.30   15.90

$Sporty
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   14.18   16.80   19.39   22.43   38.00

$Van
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   16.3    19.0    19.1    19.1    19.7    22.7
```



- pairs(data.frame(Price,Type,mpg))

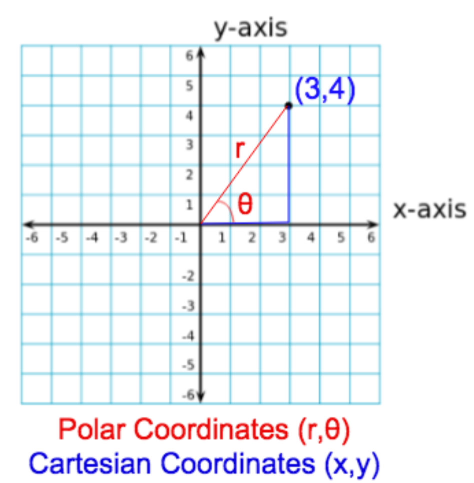


ggplot2

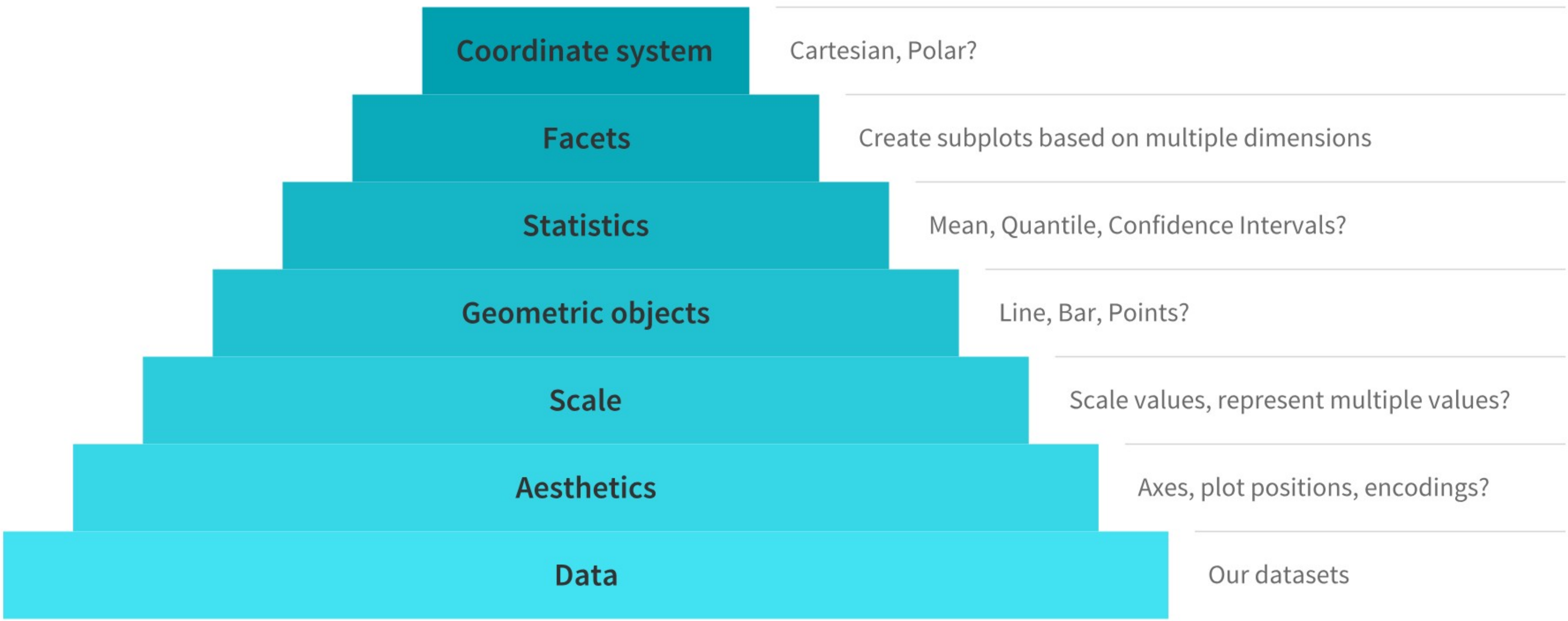
- ggplot2 provides two ways to produce plot objects:
  - **qplot() # quick plot**
    - uses some concepts of The Grammar of Graphics, but doesn't provide full capability
    - designed to be very similar to plot() and simple to use
    - may make it easy to produce basic graphs
  - **ggplot() # grammar of graphics plot**
    - provides fuller implementation of The Grammar of Graphics
    - may have steeper learning curve but allows much more flexibility when building graphs

Grammar Defines Components of Graphics

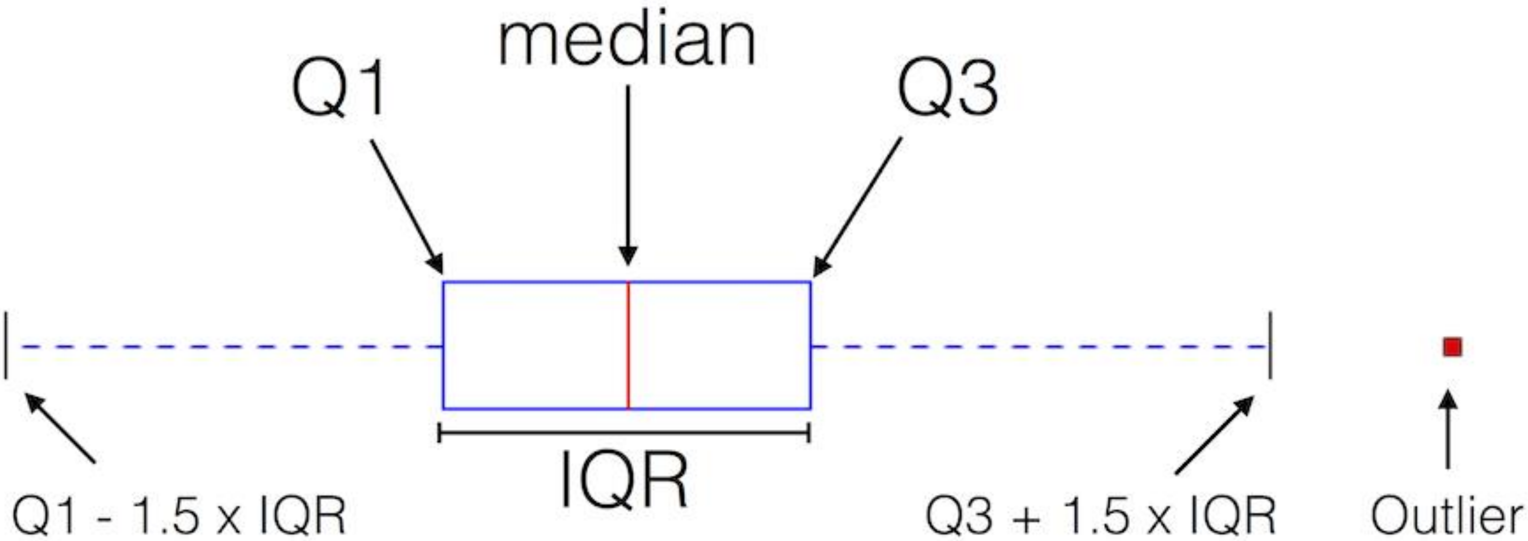
- **data**: in ggplot2, data must be stored as an R data frame
- **coordinate system**: describes 2-D space that data is projected onto - for example, Cartesian coordinates, polar coordinates, map projections, ...
- **geoms**: describe type of geometric objects that represent data - for example, points, lines, polygons,
  - help.search("^geom_", package = "ggplot2")
- **aesthetics**: describe visual characteristics that represent data - for example, position, size, color, shape, transparency, fill
- **scales**: for each aesthetic, describe how visual characteristic is converted to display values - for example, log scales, color scales, size scales, shape scales, ...
- **stats** : describe statistical transformations that typically summarize data - fo means, medians, regression lines, ...
  - help.search("^stat_", package= "ggplot2")
- **facets**: describe how data is split into subsets and displayed as multiple sm



Major Components of the Grammar of Graphics



- https://data.library.virginia.edu/setting-up-color-palettes-in-r/
- https://www.google.com/search?q=rcolorbrewer+palettes&tbm=isch&source=iu&ictx=1&fir=NHgmpgl5sIOqJM%253A%252Ccwchcbz5KdbGWM%252C_&vet=1&usg=AI4_-kTRIwvjUhIEZdKPwVgkOpFqToJR6A&sa=X&ved=2ahUKEwiw96W1ruLnAhU6yDgGHQEdCd4Q9QEwBXoECAoQJg#imgrc=Uv4h8V9I4WnMLM

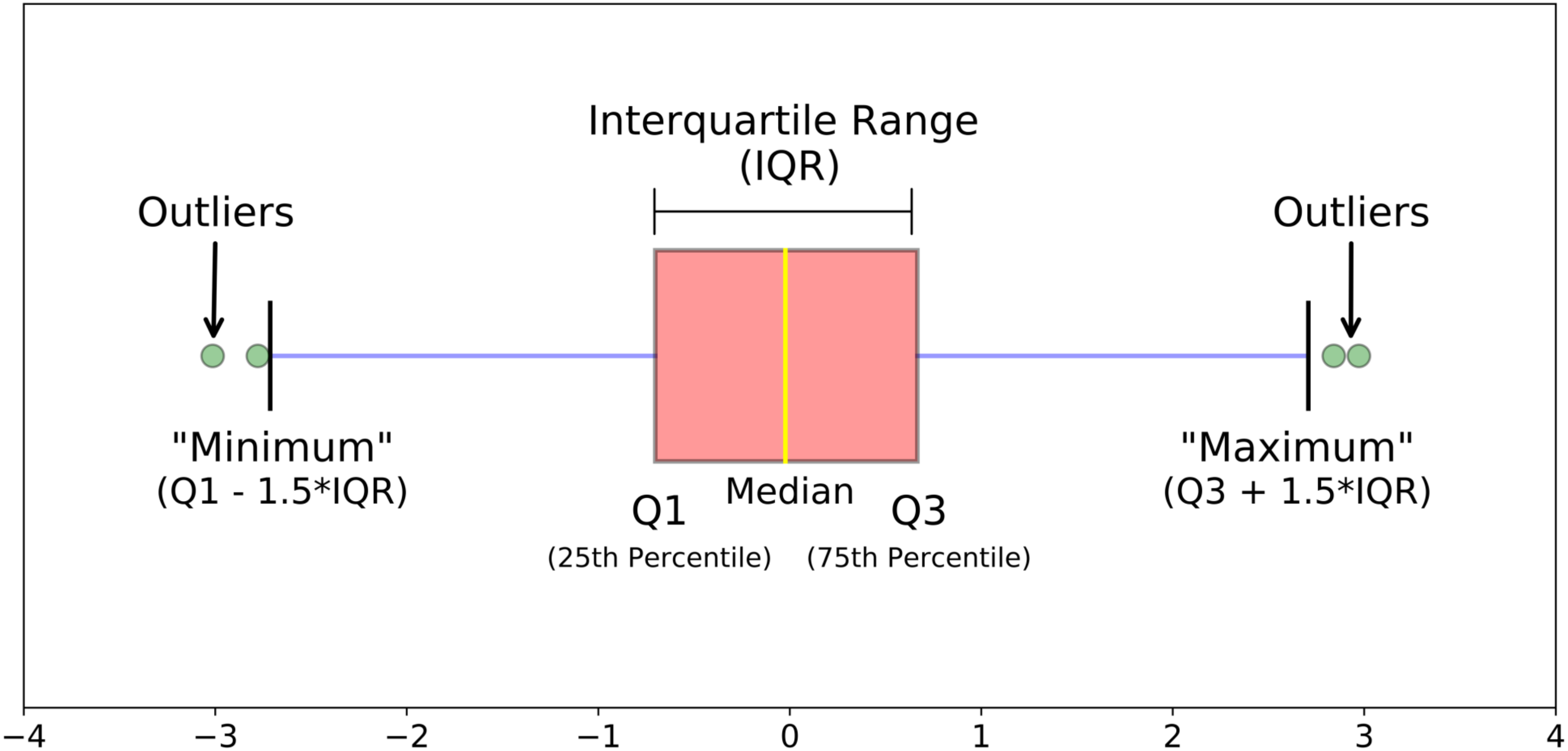


Q1: *Quartile 1*, or median of the *left* data subset
after dividing the original data set into 2 subsets via the median
(25% of the data points fall below this threshold)

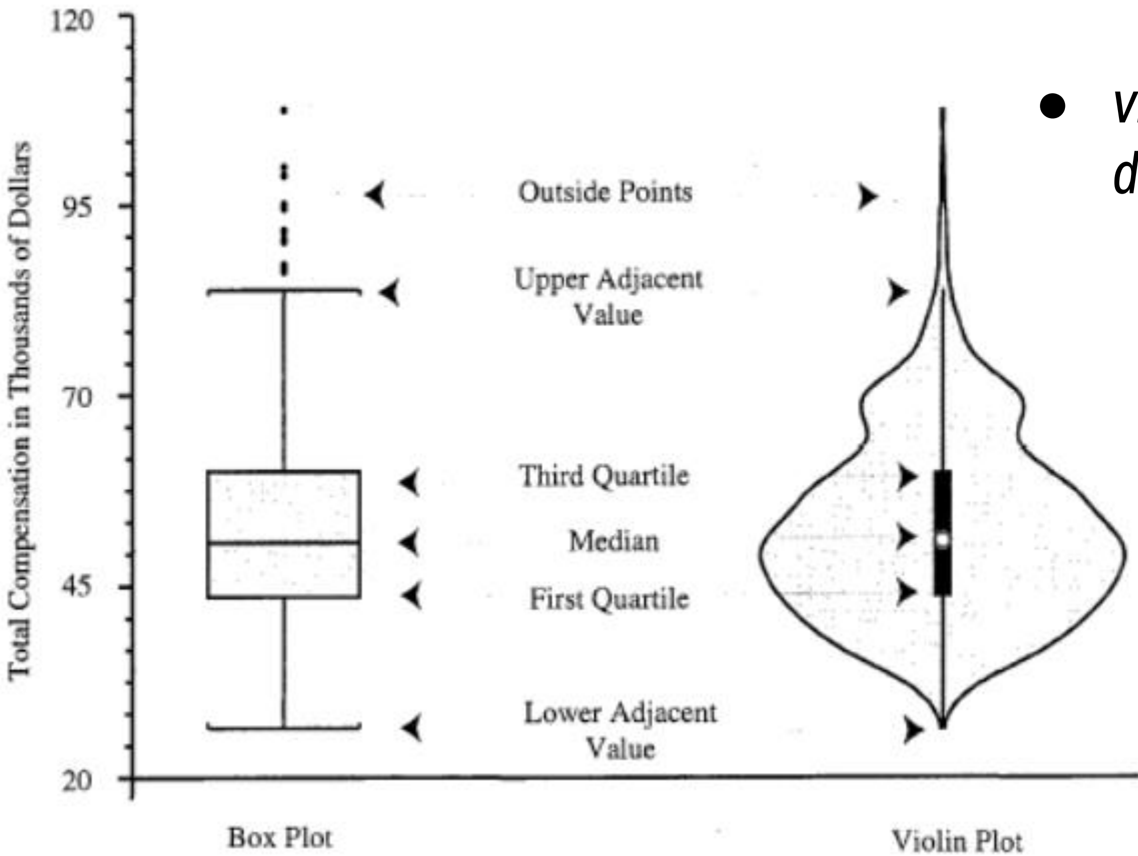
Q3: *Quartile 3*, median of the *right* data subset
(75% of the data points fall below this threshold)

IQR: *Interquartile-range*, Q3 - Q1

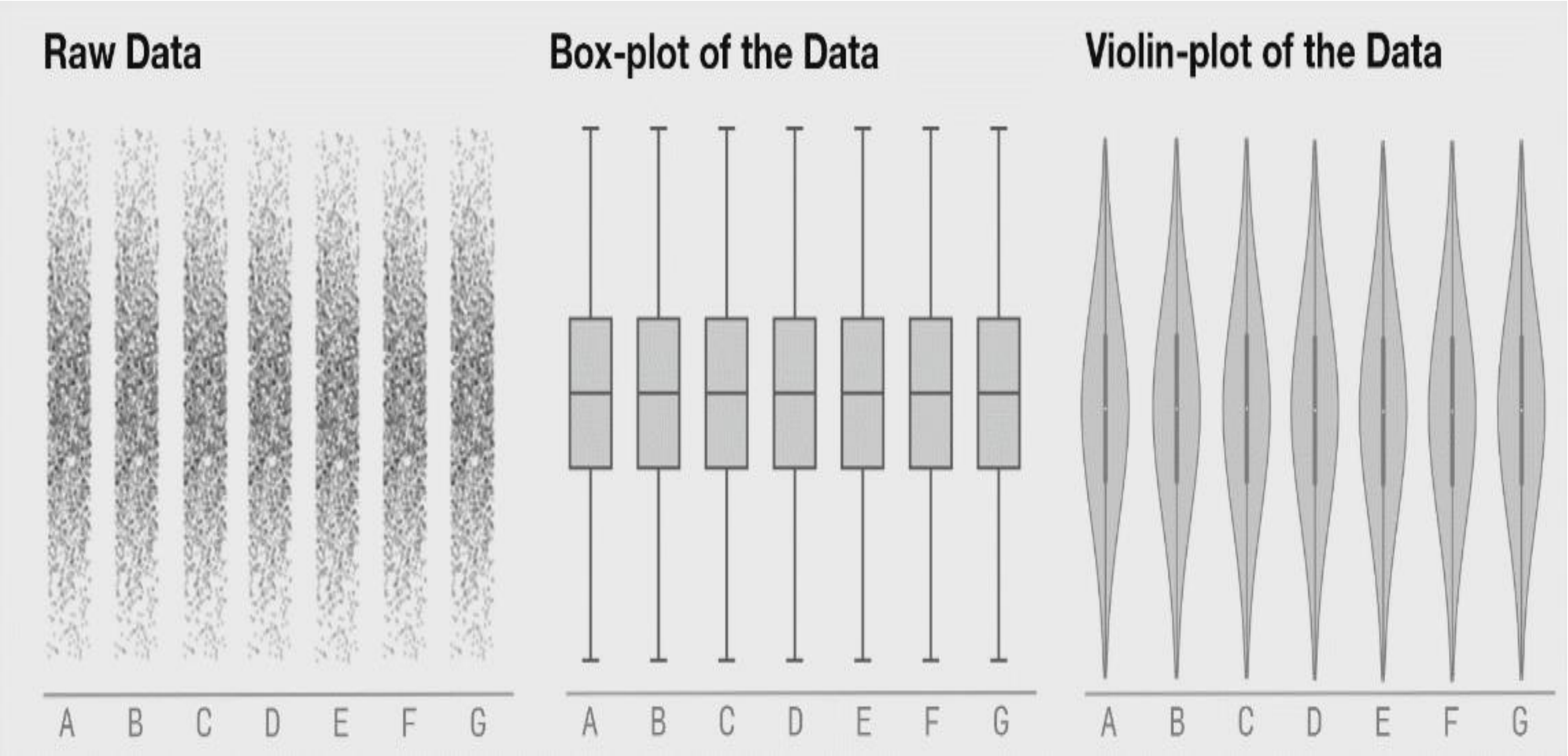
Outliers: Data points are considered to be outliers if
value < *Q1 - 1.5 x IQR* or
value > Q3 + 1.5 x IQR


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

Violin Plot

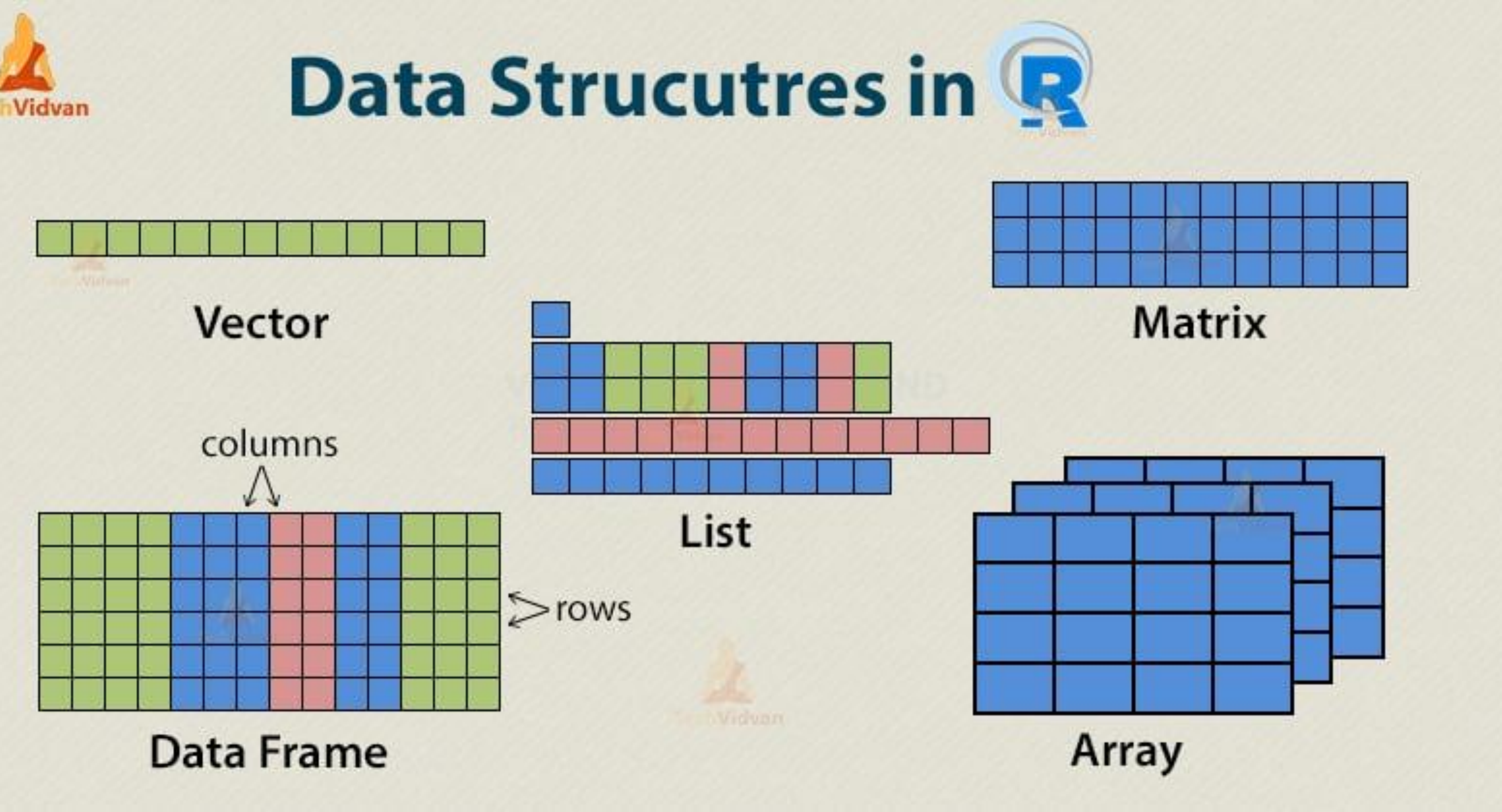


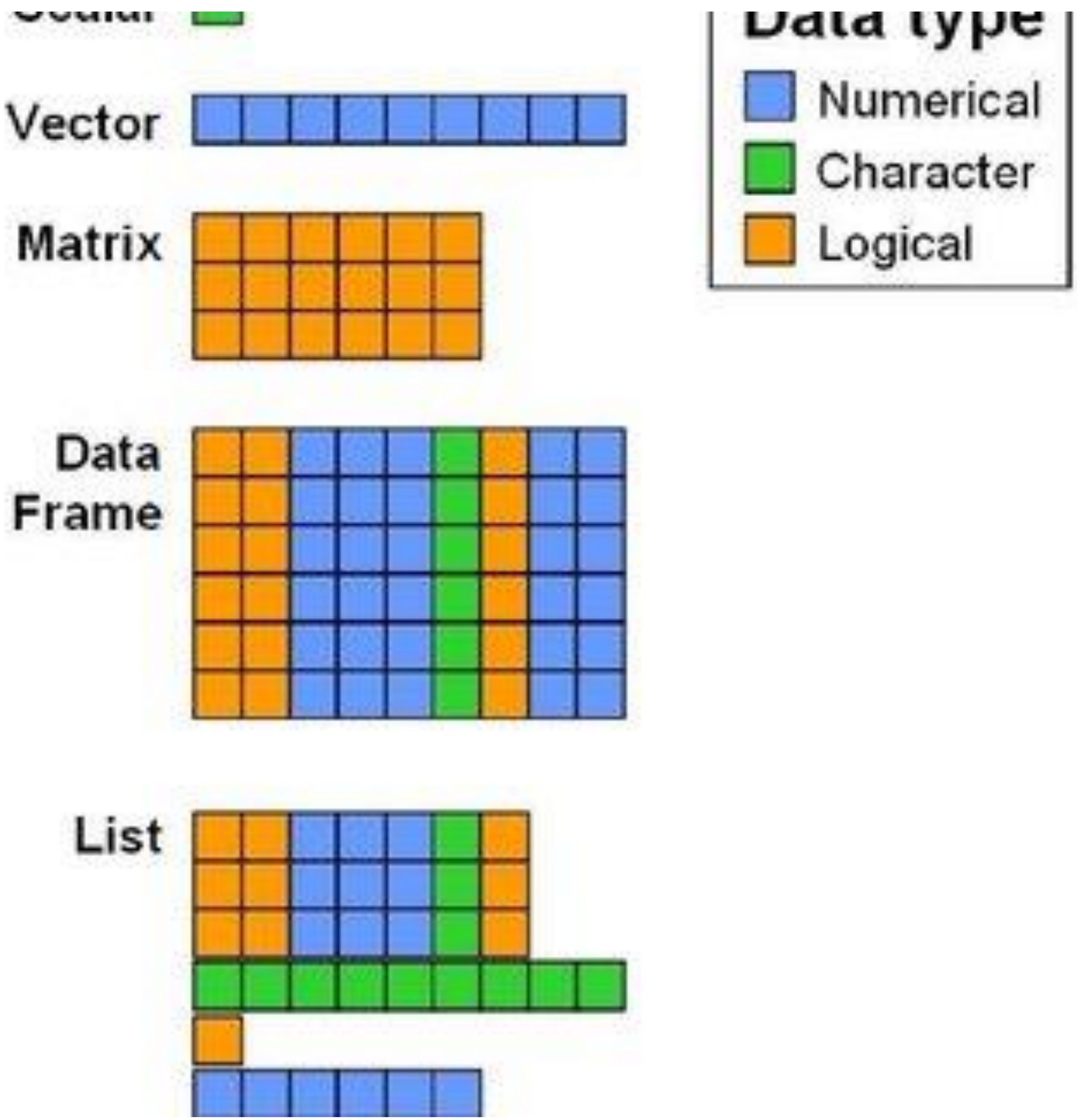
- *visualise the distribution of the data and its probability density*


Raw Data
Box-plot of the Data
Violin-plot of the Data
A B C D E F G
A B C D E F G
A B C D E F G

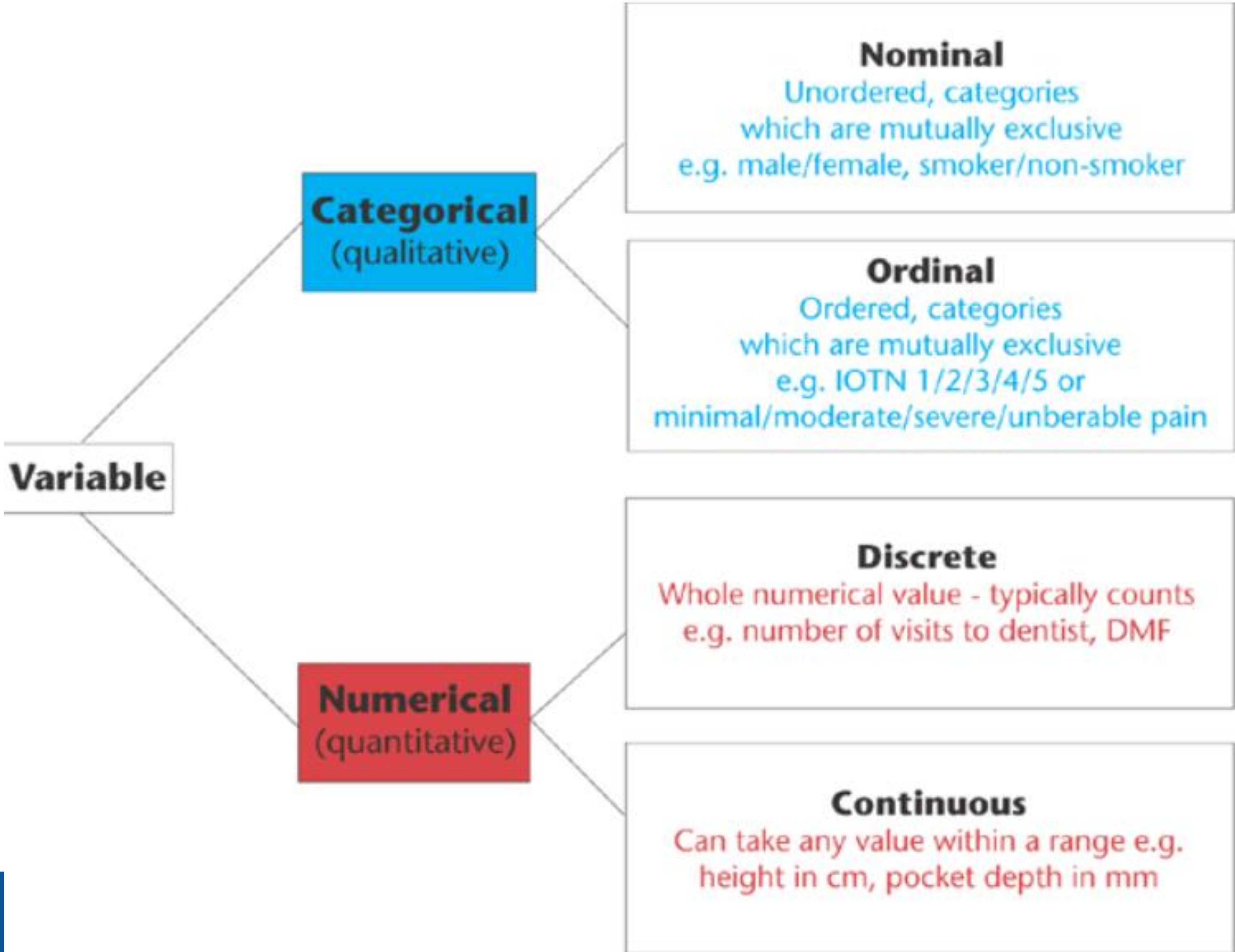
- Regression
- http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/
- https://www.tutorialspoint.com/r/r_linear_regression.htm


Thank you


Data Strucutres in R
Vector
Matrix
List
columns
rows
Data Frame
Array


Vector
Matrix
Data
Frame
List
Numerical
Character
Logical

Data Types


Variable
Categorical
(qualitative)
Numerical
(quantitative)
Nominal
Unordered, categories
which are mutually exclusive
e.g. male/female, smoker/non-smoker
Ordinal
Ordered, categories
which are mutually exclusive
e.g. IOTN 1/2/3/4/5 or
minimal/moderate/severe/unberable pain
Discrete
Whole numerical value - typically counts
e.g. number of visits to dentist, DMF
Continuous
Can take any value within a range e.g.
height in cm, pocket depth in mm

Employee Attitude Survey

Q.1 What is your Gender?

- ☐ Male
- ☐ Female

Q.3 In which department do you work?

- ☐ Marketing
- ☐ R&D
- ☐ Accounting
- ☐ Manufacturing

Q.4 On a scale where 5 represents strongly agree and 1 represents strongly disagree how would you rate each of the following statements?

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Manager offers constructive criticism	☐	☐	☐	☐	☐
Manager praises me for good work	☐	☐	☐	☐	☐
Manager considers my suggestions	☐	☐	☐	☐	☐
Company has good employee benefits	☐	☐	☐	☐	☐
I am paid a fair salary	☐	☐	☐	☐	☐