Descriptive Statistics (with R)

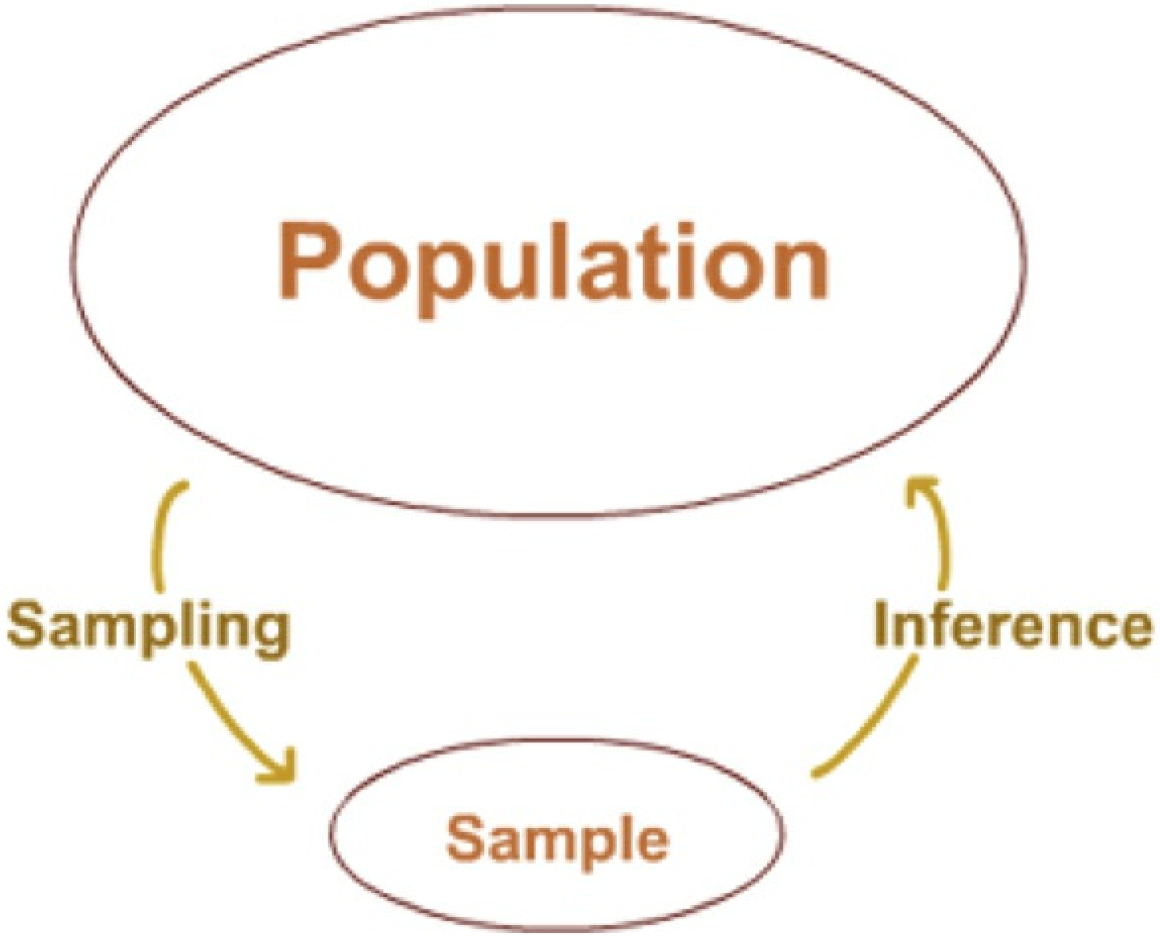
IS 665 Data Mining, Data Warehousing and Visualization

Types of Statistics

- **Descriptive Statistics:** Simplify (summarize) the data to make it easier to understand / compare

- **Inferential Statistics:** Use the data we have to make informed conjectures about larger questions for which we do not have full information.

Inferential Process



Use information about the sample to infer about the population

Data Basics

- **Data** are recorded measurements
- A **variable** is a characteristic of any entity being studied that is capable of taking on different values
    - return on investment, advertising dollars, labor productivity, stock price...

Data

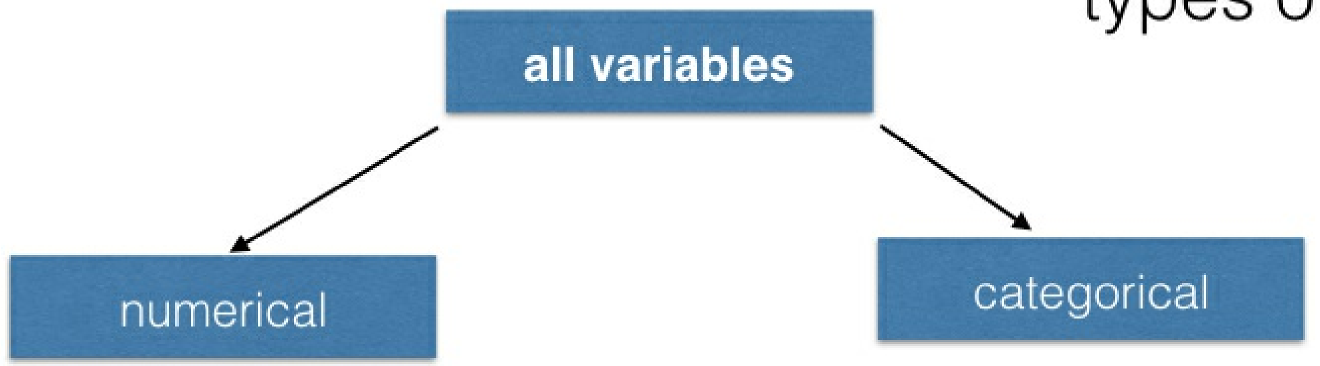
country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...	...	...	...	...	...	...	...
United States	92	63	5950	93	...	northern	very high

observation (case)

variable

http://www.google.com/transparencyreport/?hl=en_US

types of variables



(quantitative)

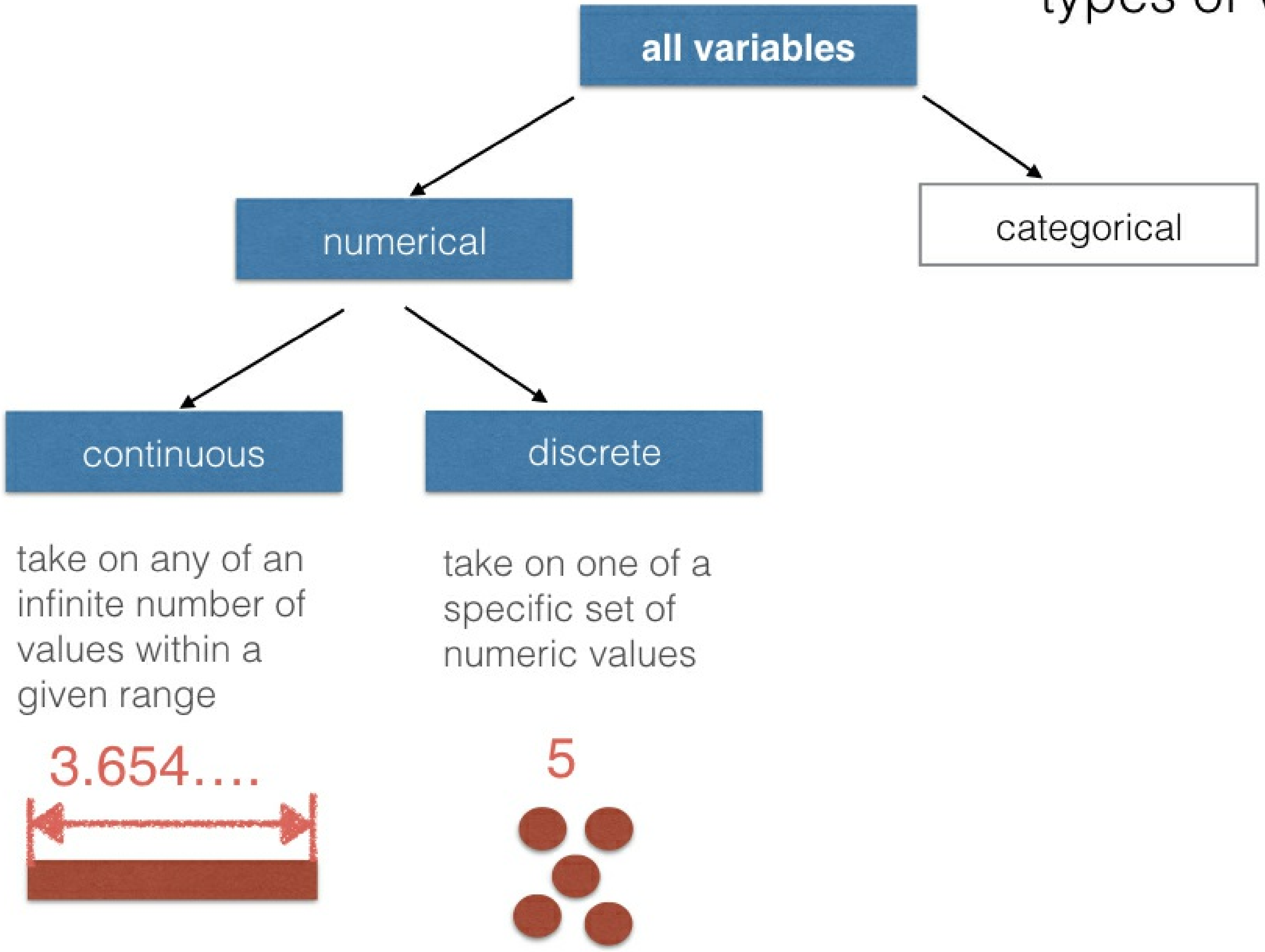
take on numerical values sensible to add, subtract, take averages, etc. with these values

(qualitative)

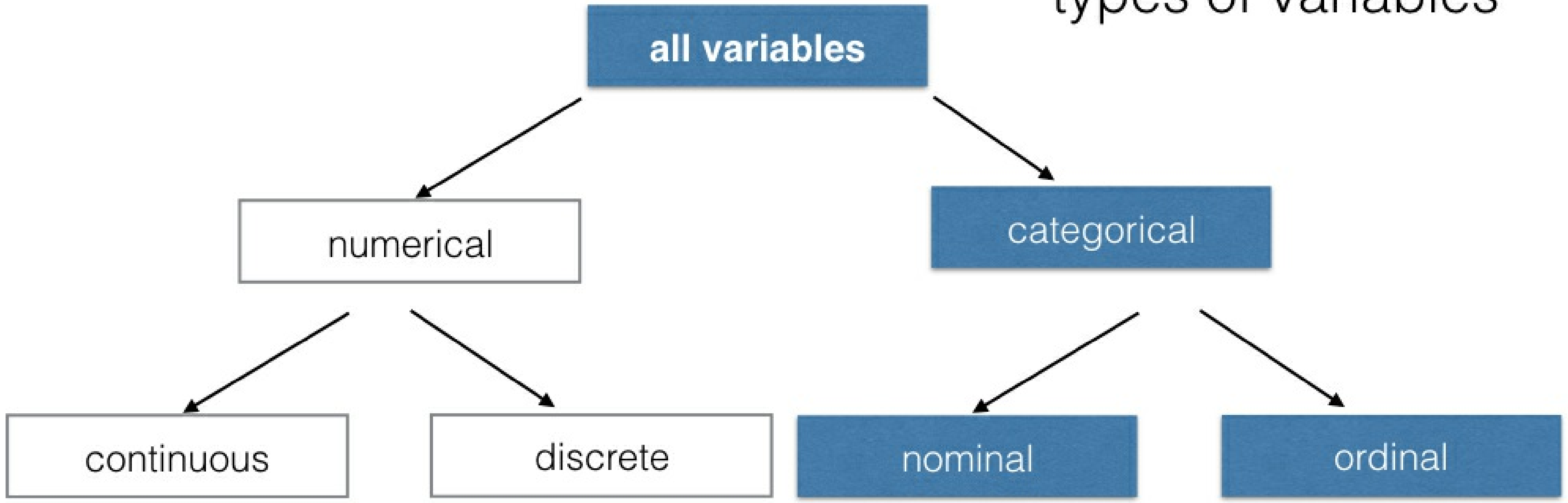
take on a limited number of distinct categories

categories can be identified with numbers, but not sensible to do arithmetic operations

types of variables



types of variables



nominal: only assigns to a category

- Male
- Female

ordinal: levels have an inherent ordering

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

Working with data:

- Each row is an observation (case)
- Each column is a variable

```
# install.packages('ggplot2')
require("ggplot2")
head(mpg)
```

```
# A tibble: 6 × 11
  manufacturer model displ  year   cyl      trans   drv   cty   hwy    fl
         <chr> <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>
1         audi    a4   1.8  1999     4   auto(l5)     f    18    29     p
2         audi    a4   1.8  1999     4 manual(m5)     f    21    29     p
3         audi    a4   2.0  2008     4 manual(m6)     f    20    31     p
4         audi    a4   2.0  2008     4   auto(av)     f    21    30     p
5         audi    a4   2.8  1999     6   auto(l5)     f    16    26     p
6         audi    a4   2.8  1999     6 manual(m5)     f    18    26     p
# ... with 1 more variables: class <chr>
```

When working with data the first thing you will look at is:

- some measure of the **middle** of the data (or **central tendency**).
- ex. what is a **typical** highway gas mileage in your data set.

Then you look at:

- some measure of the **variance** of the data around the middle (or **dispersion**).
- ex. how close are the cars in my data set to the typical highway gas mileage

Lastly check if data is normally distributed:

- some measure of **shape** of the data

Descriptive Statistics

Centrality and Spread measures for Quantitative Variables

Measures of Central Tendancy	**Measures of Spread (Dispersion)**
Mean	Range
Median	IQR (Interquartile Range)
Mode	Variance
Percentiles	Standard Deviation
Quartiles	

Measures Central Tendency

Average :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n)$$

```
mean(mpg$hwy)
```

```
[1] 23.44017
```

Median :

Middle value in an ordered array of numbers. It’s the (n+1)/2 th ordered observation

```
median(mpg$hwy)
```

```
[1] 24
```

Central Tendency and Outliers

When do we use median instead of average ?

- Mean is affected by each value in the dataset, including **extreme outliers**

```
head(x <- rexp(1000, 0.01))
```

```
[1]  42.483285 380.922452 107.185322 183.066265 174.040195   5.911086
```

```
mean(x)
```

```
[1] 98.23088
```

```
median(x)
```

```
[1] 68.23764
```

```
mean(x, trim = 0.2)
```

```
[1] 75.29664
```

Measures of Central Tendency (cnt'd.)

Quartile

```
quantile(mpg$hwy)
```

```
  0%  25%  50%  75% 100% 
  12   18   24   27   44 
```

```
max(mpg$hwy)
```

```
[1] 44
```

Percentile

```
quantile(mpg$hwy, 0.9)
```

```
90% 
 30 
```

```
quantile(mpg$hwy, c(0.1, 0.2, 0.3, 0.4, 0.5))
```

```
 10%  20%  30%  40%  50% 
16.3 17.0 19.0 22.0 24.0 
```

Measures of Spread

Range

```r
range(mpg$hwy)
```

```
[1] 12 44
```

```r
diff(range(mpg$hwy))
```

```
[1] 32
```

```r
max(mpg$hwy) - min(mpg$hwy)
```

```
[1] 32
```

IQR (interquartile range)

```r
IQR(mpg$hwy)
```

```
[1] 9
```

if you can't remember IQR

```r
quantile(mpg$hwy, 0.75) - quantile(mpg$hwy, 0.25)
```

```
75% 
  9 
```

Measures of Spread (ctn'd.)

MAD

(Mean Absolute Deviation)

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

```r
# install.packages('lsr')
require("lsr")
aad(mpg$hwy)
```

```
[1] 4.959128
```

If you can't remember aad():

```r
mean(abs(mpg$hwy - mean(mpg$hwy)))
```

```
[1] 4.959128
```

Standard Deviation(s)

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

$$\bar{x} = \text{sample average}$$

```r
sd(mpg$hwy)
```

```
[1] 5.954643
```

Variance(s^2)

```r
var(mpg$hwy)
```

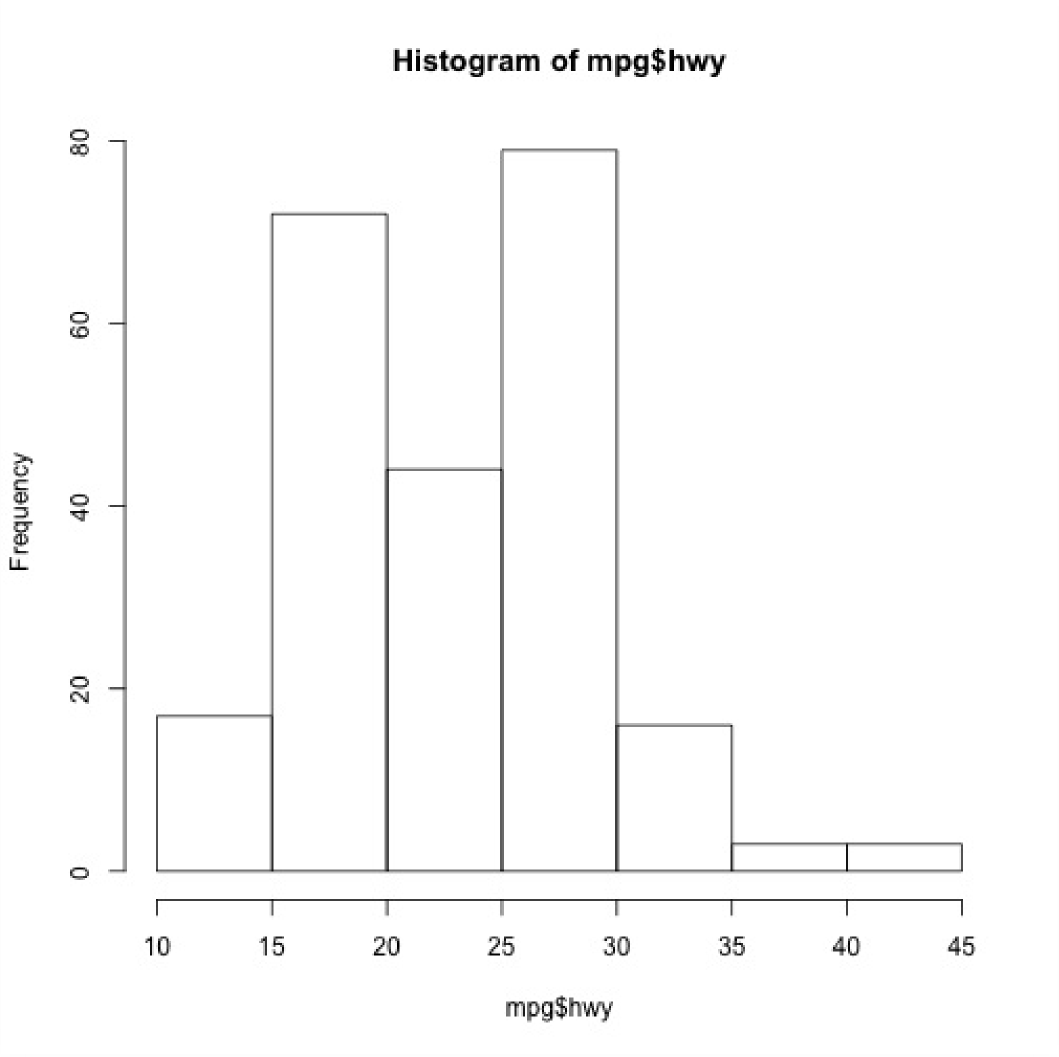
```
[1] 35.45778
```

In R default is sample *corrected*

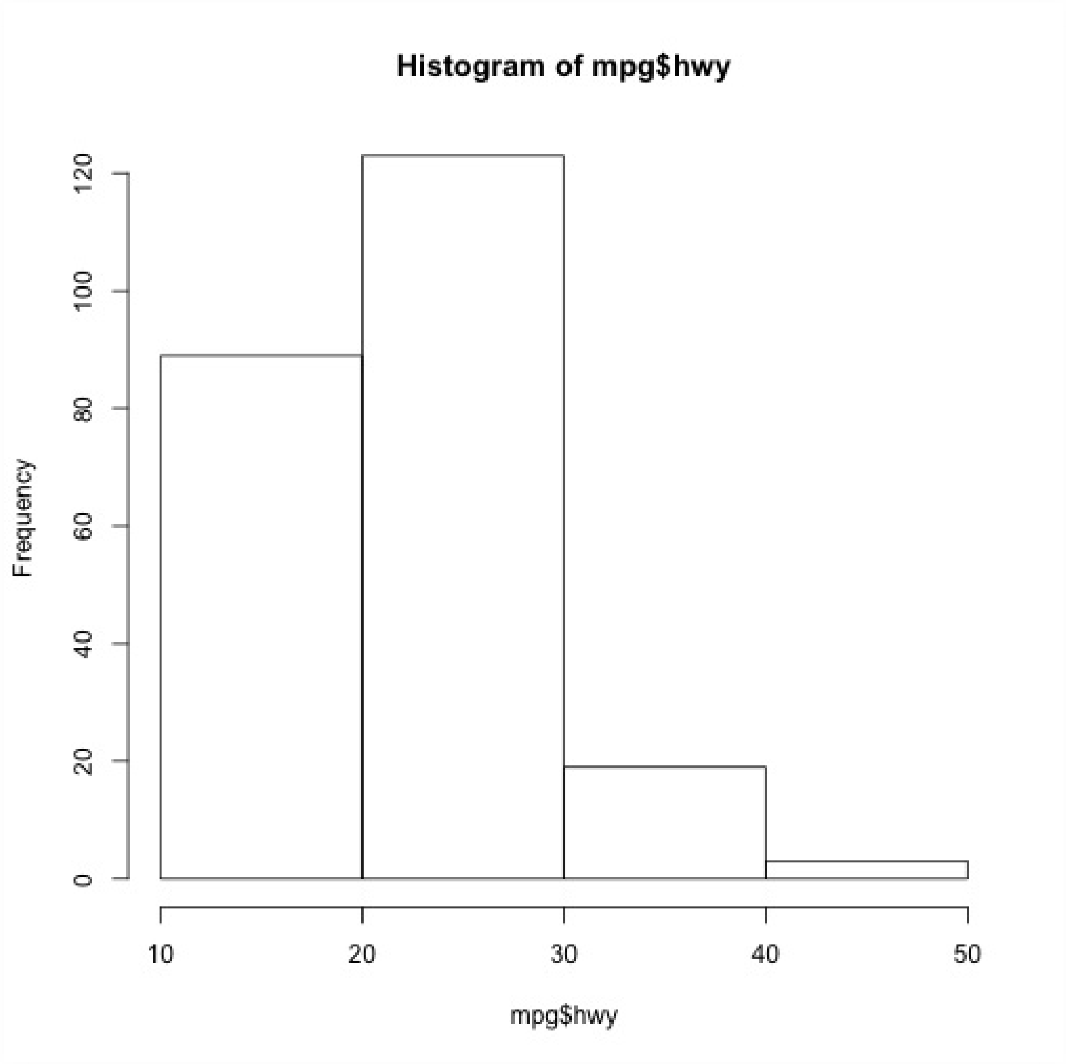
Measurements of shapes: Graphs and Charts

Histograms and Frequency Distributions

```
hist(mpg$hwy)
```




```
hist(mpg$hwy, breaks = 4)
```

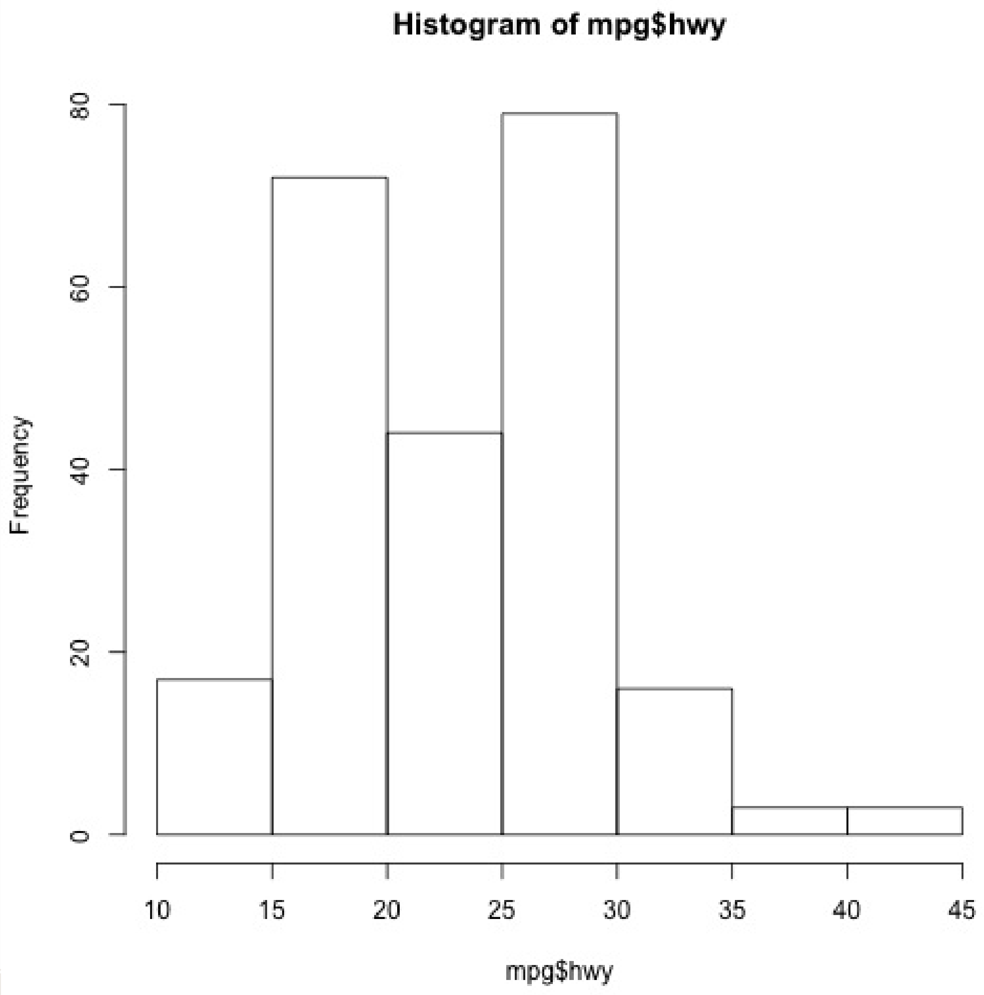


You can also identify the number of breaks

```
range(mpg$hwy)
```

```
[1] 12 44
```

```
my.breaks = seq(10, 45, 5)
hist(mpg$hwy, breaks = my.breaks)
```



Frequencies

```
my.hist = hist(mpg$hwy, breaks = my.breaks, plot =
F)
my.hist$breaks
```

```
[1] 10 15 20 25 30 35 40 45
```

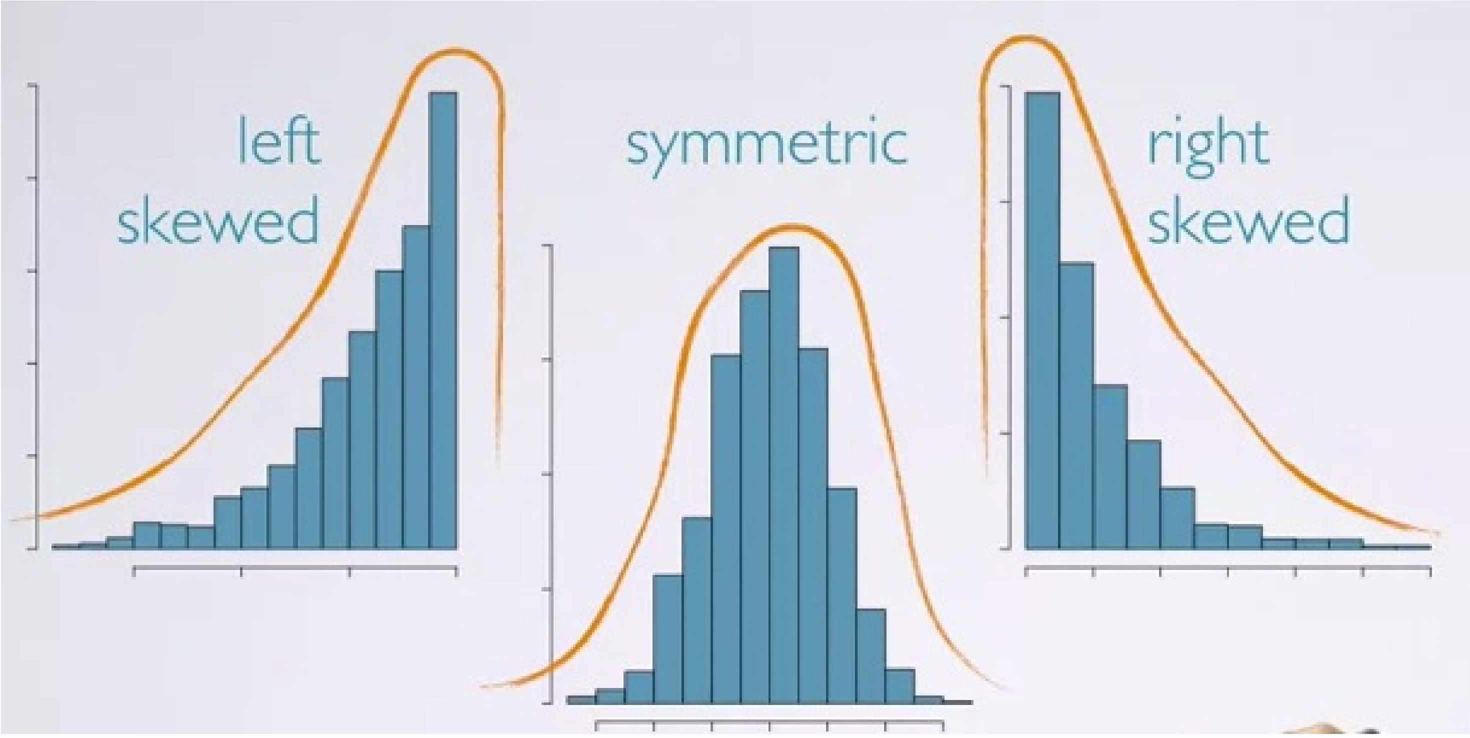
```
my.hist$counts
```

```
[1] 17 72 44 79 16  3  3
```

```
freq.dist = cbind(bin.end = my.hist$breaks[1:7], freq
= my.hist$counts)
(freq.dist = data.frame(freq.dist))
```

```
  bin.end freq
1      10   17
2      15   72
3      20   44
4      25   79
5      30   16
6      35    3
7      40    3
```

Measurements of shape - Skeweness



- Data is usually expected to be normally distributed in nature.
- So the shape of the distribution is expected to be symmetric.
- If the it is not, the data is considered to be **skewed**

Skewness

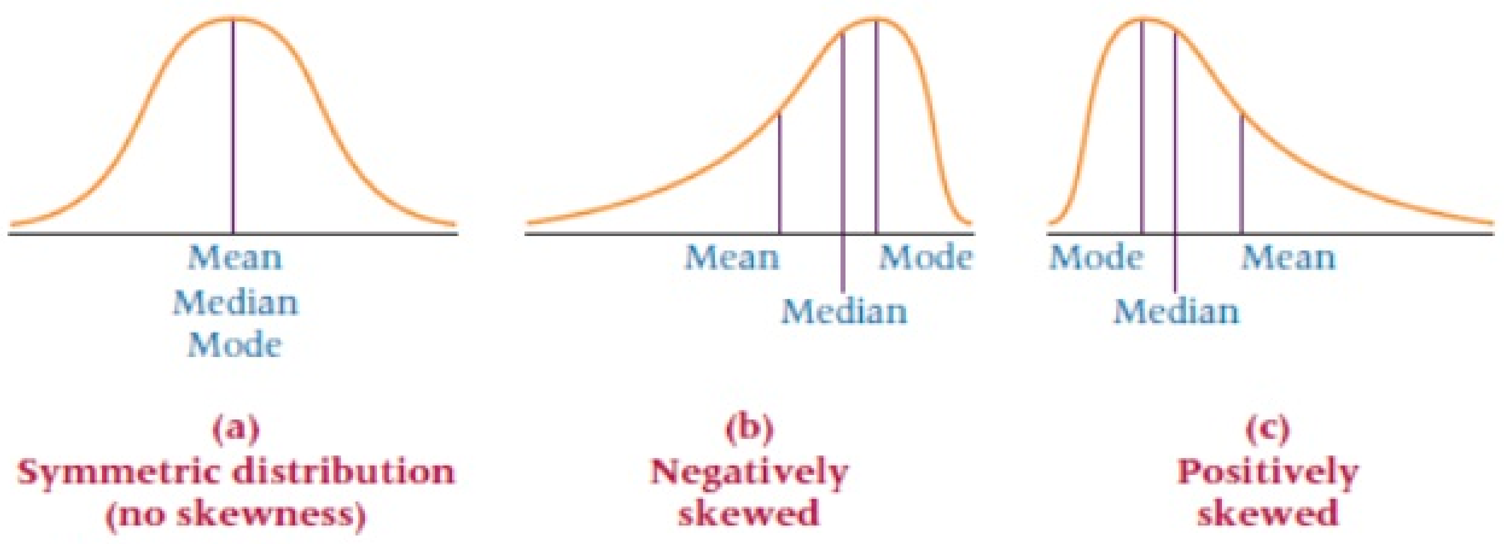


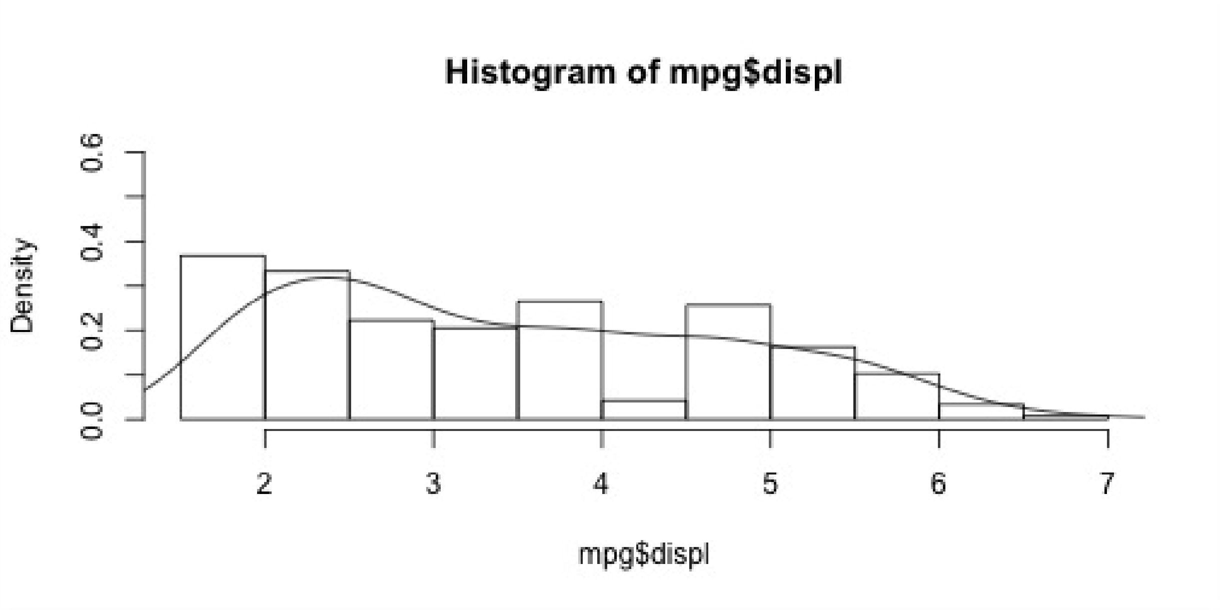
Figure 3.11

Relationship of Mean, Median, and Mode

$$Skewness = \frac{3(mean - median)}{standard\ deviation}$$

Skewness

```r
require(lattice)
hist(mpg$displ, prob = TRUE, ylim = c(0, 0.6))
lines(density(mpg$displ))
```




```r
mean.displ = mean(mpg$displ)
median.displ = median(mpg$displ)
skewness_displ = 3 * (mean.displ - median.displ)/sd(mpg$displ)
skewness_displ
```

```
[1] 0.3989172
```

Descriptive Stats. for Qualitative Variables

- Qualitative (Categorical) variables are often used to classify data into various levels or factors.

```
head(mpg)
```

```
# A tibble: 6 × 11
  manufacturer model displ  year   cyl      trans  drv   cty   hwy    fl
         <chr> <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>
1         audi    a4   1.8  1999     4   auto(l5)     f    18    29     p
2         audi    a4   1.8  1999     4 manual(m5)     f    21    29     p
3         audi    a4   2.0  2008     4 manual(m6)     f    20    31     p
4         audi    a4   2.0  2008     4   auto(av)     f    21    30     p
5         audi    a4   2.8  1999     6   auto(l5)     f    16    26     p
6         audi    a4   2.8  1999     6 manual(m5)     f    18    26     p
# ... with 1 more variables: class <chr>
```

What are the categorical variables in this data set ?

Tabulation

```
table(mpg$year)
```

```

1999 2008
 117  117
```

```
table(mpg$manufacturer)
```

```

      audi  chevrolet      dodge       ford      honda    hyundai
        18         19         37         25          9         14
      jeep land rover    lincoln    mercury     nissan    pontiac
         8          4          3          4         13          5
    subaru     toyota volkswagen
        14         34         27
```

```
table(mpg$cyl)
```

```

 4  5  6  8
81  4 79 70
```

From counts to percentages

```
prop.table(table(mpg$year))
```

```

1999 2008
 0.5  0.5
```

```
prop.table(table(mpg$manufacturer))
```

```

      audi  chevrolet      dodge       ford      honda    hyundai
0.07692308 0.08119658 0.15811966 0.10683761 0.03846154 0.05982906
      jeep land rover    lincoln    mercury     nissan    pontiac
0.03418803 0.01709402 0.01282051 0.01709402 0.05555556 0.02136752
    subaru     toyota volkswagen
0.05982906 0.14529915 0.11538462
```

```
prop.table(table(mpg$cyl))
```

```

         4          5          6          8
0.34615385 0.01709402 0.33760684 0.29914530
```

Cross-Tabulation

```
table(mpg$manufacturer, mpg$cyl)
```

```
             4  5  6  8
  audi        8  0  9  1
  chevrolet   2  0  3 14
  dodge       1  0 15 21
  ford        0  0 10 15
  honda       9  0  0  0
  hyundai     8  0  6  0
  jeep        0  0  3  5
  land rover  0  0  0  4
  lincoln     0  0  0  3
  mercury     0  0  2  2
  nissan      4  0  8  1
  pontiac     0  0  4  1
  subaru     14  0  0  0
  toyota     18  0 13  3
  volkswagen 17  4  6  0
```

```
table(mpg$manufacturer, mpg$cyl, mpg$year)
```

```
, ,  = 1999


             4  5  6  8
  audi        4  0  5  0
  chevrolet   1  0  1  5
  dodge       1  0  8  7
  ford        0  0  7  8
  honda       5  0  0  0
  hyundai     4  0  2  0
  jeep        0  0  1  1
  land rover  0  0  0  2
  lincoln     0  0  0  2
  mercury     0  0  1  1
  nissan      2  0  4  0
  pontiac     0  0  3  0
  subaru      6  0  0  0
  toyota     11  0  8  1
  volkswagen 11  0  5  0

, ,  = 2008


             4  5  6  8
  audi        4  0  4  1
  chevrolet   1  0  2  9
  dodge       0  0  7 14
  ford        0  0  3  7
  honda       4  0  0  0
```

Percentages in Cross-Tabulations

```r
(man.by.cyl = table(mpg$manufacturer, mpg$cyl))
```

```
             4  5  6  8
  audi        8  0  9  1
  chevrolet   2  0  3 14
  dodge       1  0 15 21
  ford        0  0 10 15
  honda       9  0  0  0
  hyundai     8  0  6  0
  jeep        0  0  3  5
  land rover  0  0  0  4
  lincoln     0  0  0  3
  mercury     0  0  2  2
  nissan      4  0  8  1
  pontiac     0  0  4  1
  subaru     14  0  0  0
  toyota     18  0 13  3
  volkswagen 17  4  6  0
```

```r
prop.man.by.cyl = prop.table(man.by.cyl)
round(prop.man.by.cyl, digits = 2)
```

```
                4    5    6    8
  audi       0.03 0.00 0.04 0.00
  chevrolet  0.01 0.00 0.01 0.06
  dodge      0.00 0.00 0.06 0.09
  ford       0.00 0.00 0.04 0.06
  honda      0.04 0.00 0.00 0.00
  hyundai    0.03 0.00 0.03 0.00
  jeep       0.00 0.00 0.01 0.02
  land rover 0.00 0.00 0.00 0.02
  lincoln    0.00 0.00 0.00 0.01
  mercury    0.00 0.00 0.01 0.01
  nissan     0.02 0.00 0.03 0.00
  pontiac    0.00 0.00 0.02 0.00
  subaru     0.06 0.00 0.00 0.00
  toyota     0.08 0.00 0.06 0.01
  volkswagen 0.07 0.02 0.03 0.00
```

Prop.cell = cell count / N. of observations

Percentages

Percentages in row

```r
prop.by.row = prop.table(man.by.cyl, margin = 1)
head(round(prop.by.row, digits = 2))
```

```
              4    5    6    8
audi       0.44 0.00 0.50 0.06
chevrolet  0.11 0.00 0.16 0.74
dodge      0.03 0.00 0.41 0.57
ford       0.00 0.00 0.40 0.60
honda      1.00 0.00 0.00 0.00
hyundai    0.57 0.00 0.43 0.00
```

```r
rowSums(prop.by.row)[1:4]
```

```
     audi chevrolet     dodge      ford 
        1         1         1         1 
```

Percentages in column

```r
prop.by.column = prop.table(man.by.cyl, margin = 2)
head(round(prop.by.column, digits = 2))
```

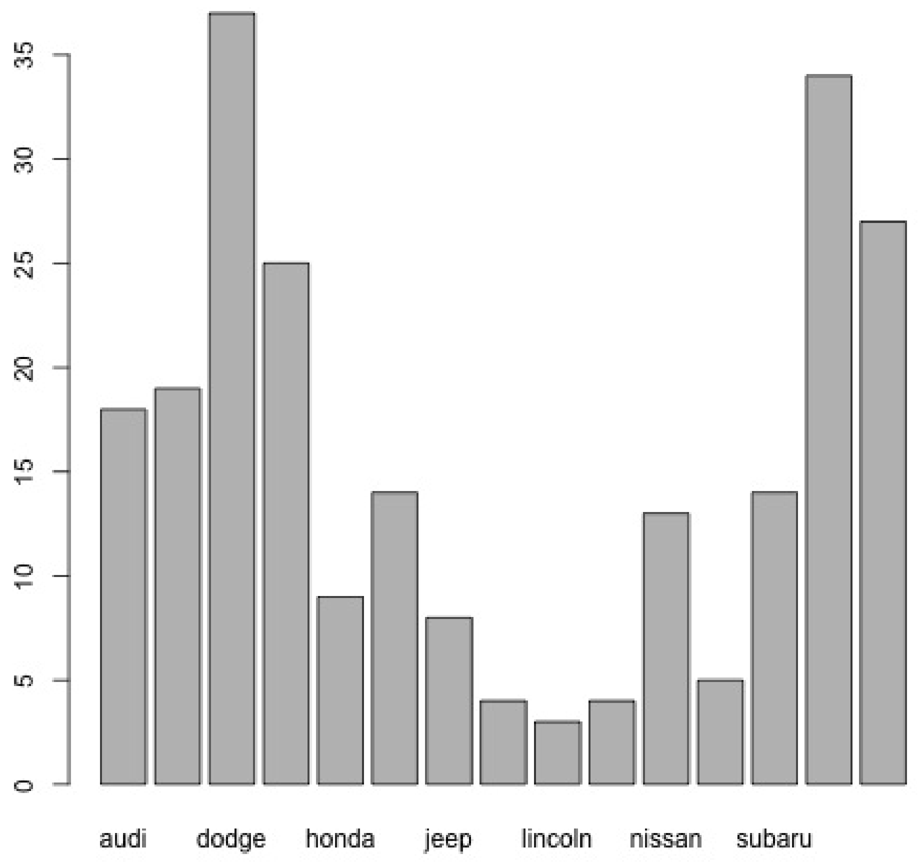
```
              4    5    6    8
audi       0.10 0.00 0.11 0.01
chevrolet  0.02 0.00 0.04 0.20
dodge      0.01 0.00 0.19 0.30
ford       0.00 0.00 0.13 0.21
honda      0.11 0.00 0.00 0.00
hyundai    0.10 0.00 0.08 0.00
```

```r
colSums(prop.by.column)[1:4]
```

```
4 5 6 8 
1 1 1 1 
```

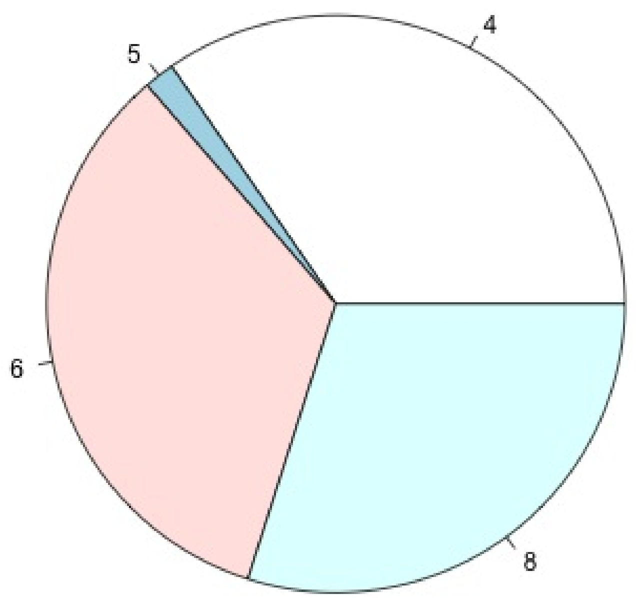
Bar charts

```
barplot(table(mpg$manufacturer))
```



Pie Charts

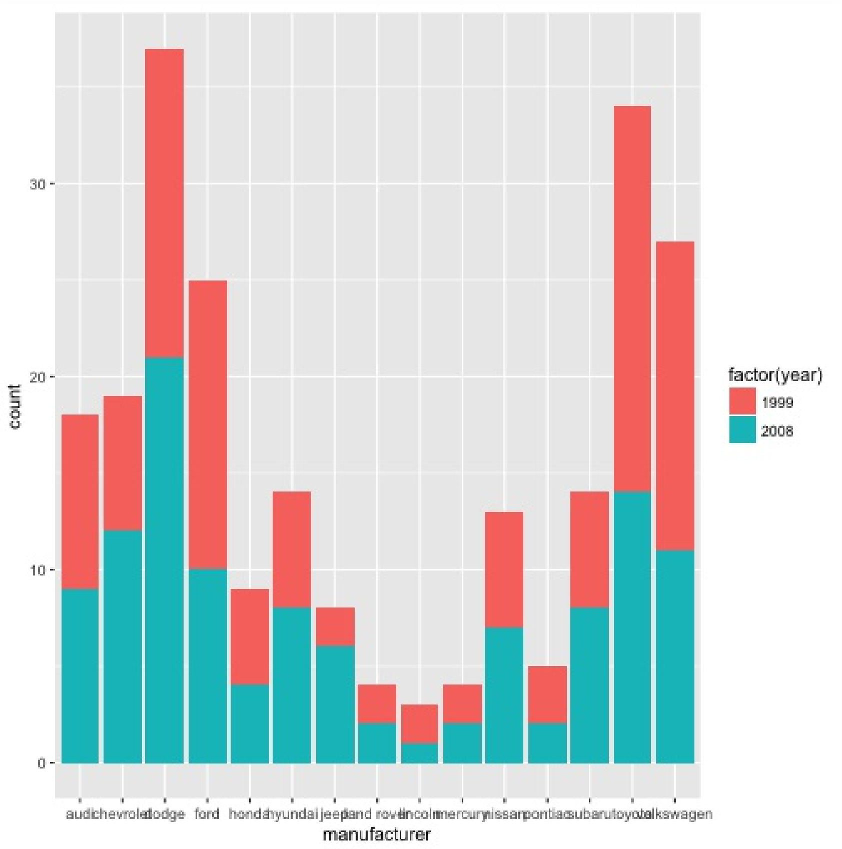
```
pie(table(mpg$cyl))
```



GGplot for visualization

We usually use libraries that can generate nicer looking graphs.
The syntax is a little more complicated tough

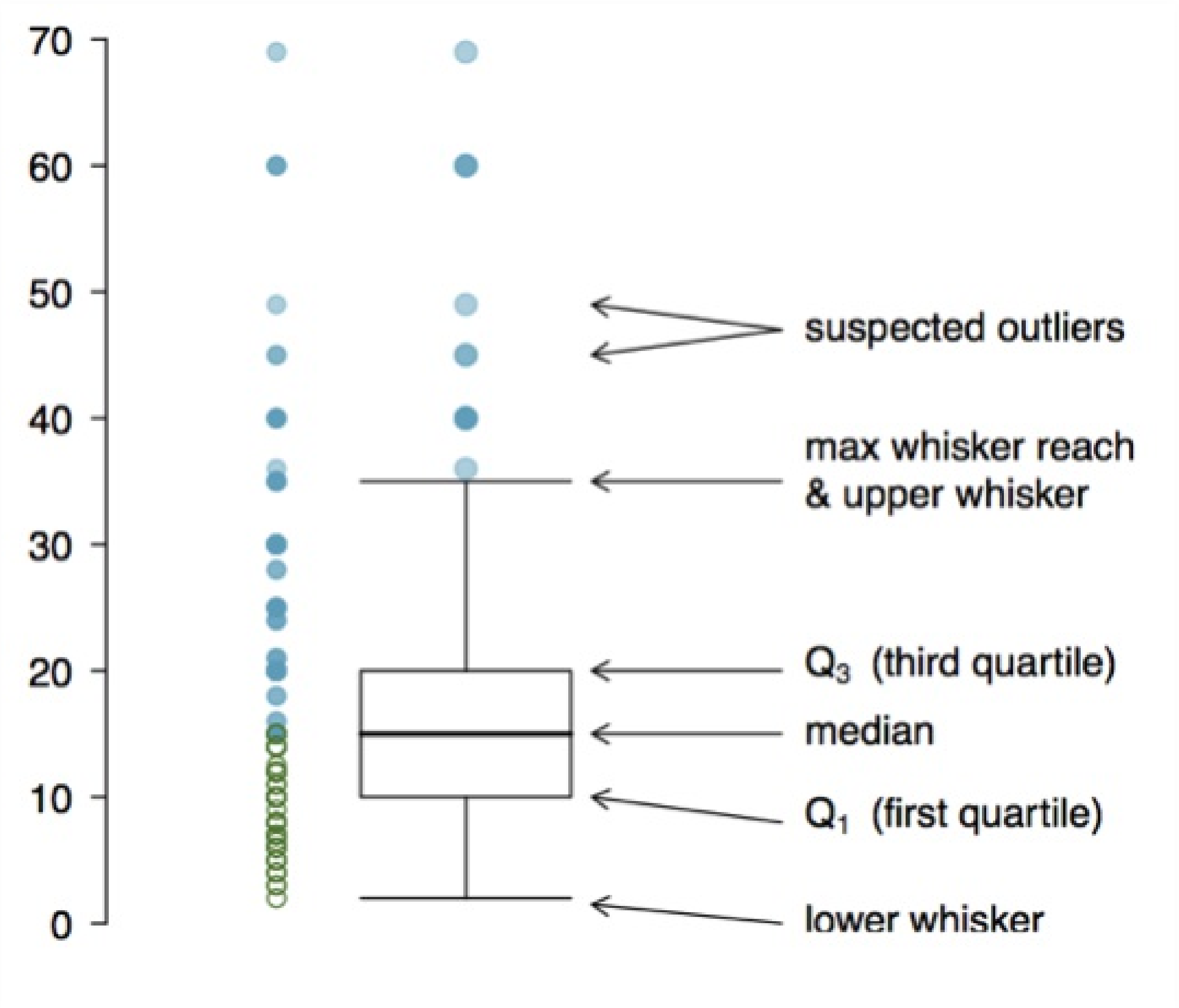
```r
# install.packages(ggplot2)
require(ggplot2)
ggplot(data = mpg) + geom_bar(aes(x = manufacturer, fill = factor(year)))
```



Box Plot

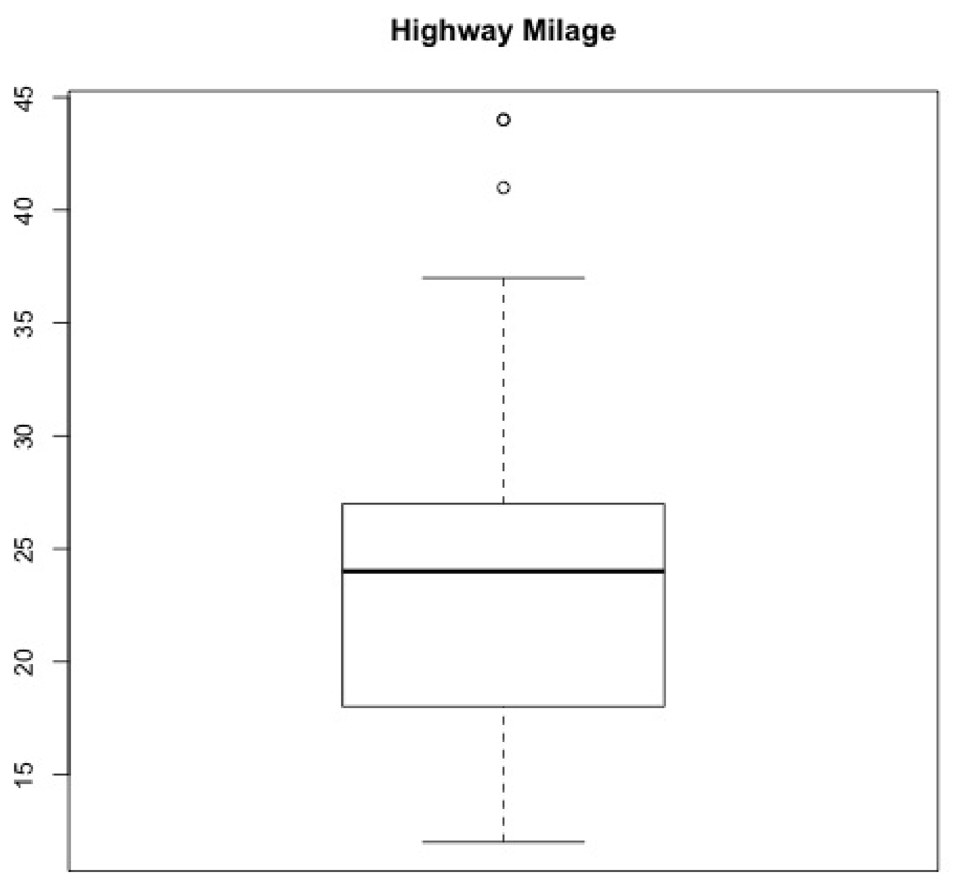
The box in a **box plot** represents the middle 50% of the data, and the thick line in the box is the median.

- Anatomy of a box plot



Box plots in R

```
boxplot(mpg$hwy, main = "Highway Milage")
```

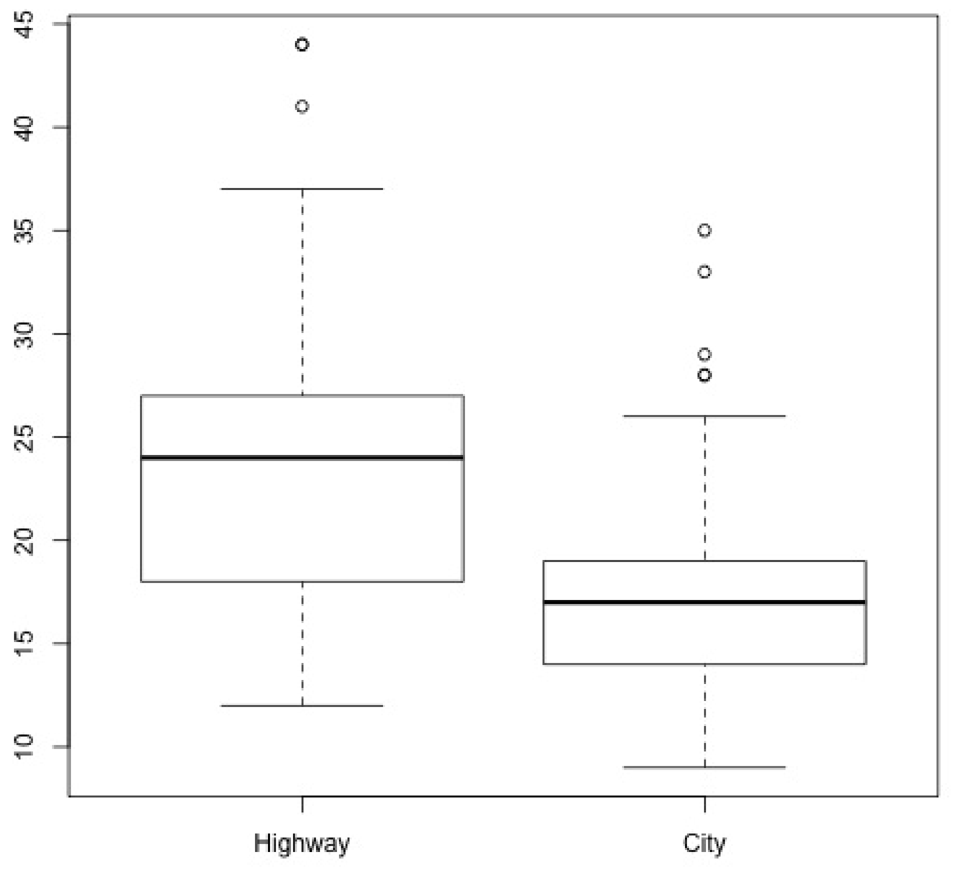



```
summary(mpg[c("cty")])
```

```
      cty
 Min.   : 9.00
 1st Qu.:14.00
 Median :17.00
 Mean   :16.86
 3rd Qu.:19.00
 Max.   :35.00
```

Box plots

```
boxplot(mpg$hwy, mpg$cty, names = c("Highway",
"City"))
```

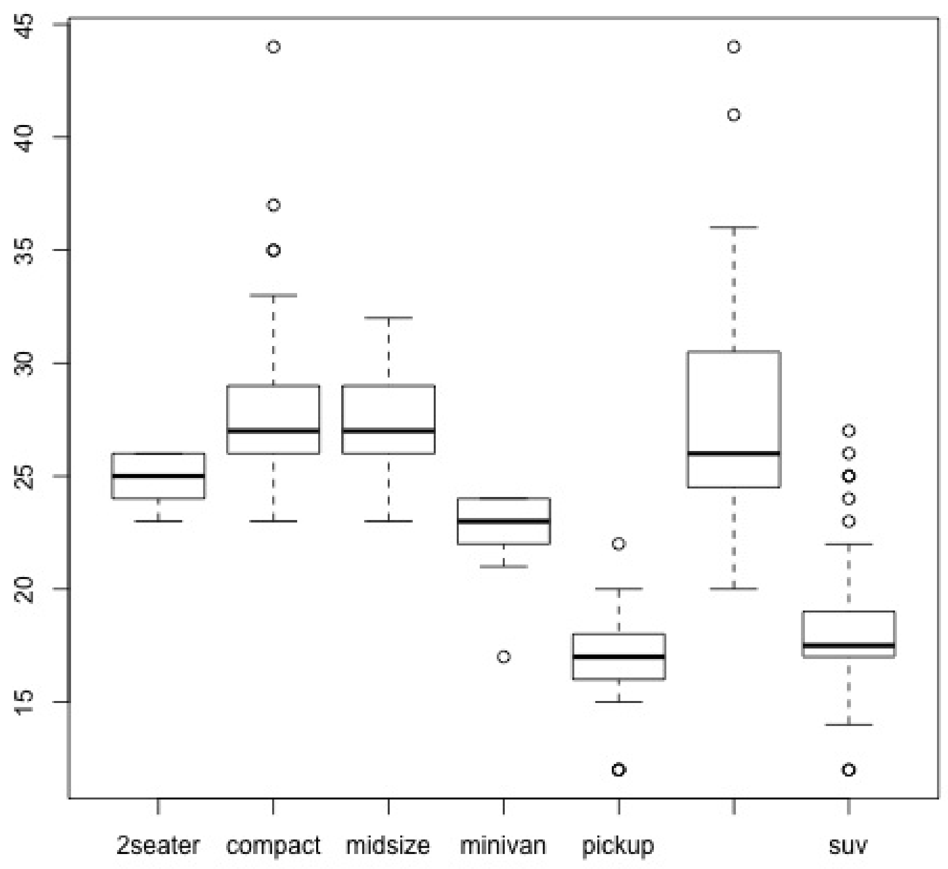



```
summary(mpg[c("cty", "hwy")])
```

```
     cty             hwy
 Min.   : 9.00   Min.   :12.00
 1st Qu.:14.00   1st Qu.:18.00
 Median :17.00   Median :24.00
 Mean   :16.86   Mean   :23.44
 3rd Qu.:19.00   3rd Qu.:27.00
 Max.   :35.00   Max.   :44.00
```

Box plots

```
boxplot(mpg$hwy ~ mpg$class)
```




```
# install.packages('doBy')
require("doBy")
summaryBy(hwy ~ class, data =
as.data.frame(mpg), FUN = c(median))
```

```
       class hwy.median
1    2seater       25.0
2    compact       27.0
3    midsize       27.0
4    minivan       23.0
5     pickup       17.0
6 subcompact       26.0
7        suv       17.5
```