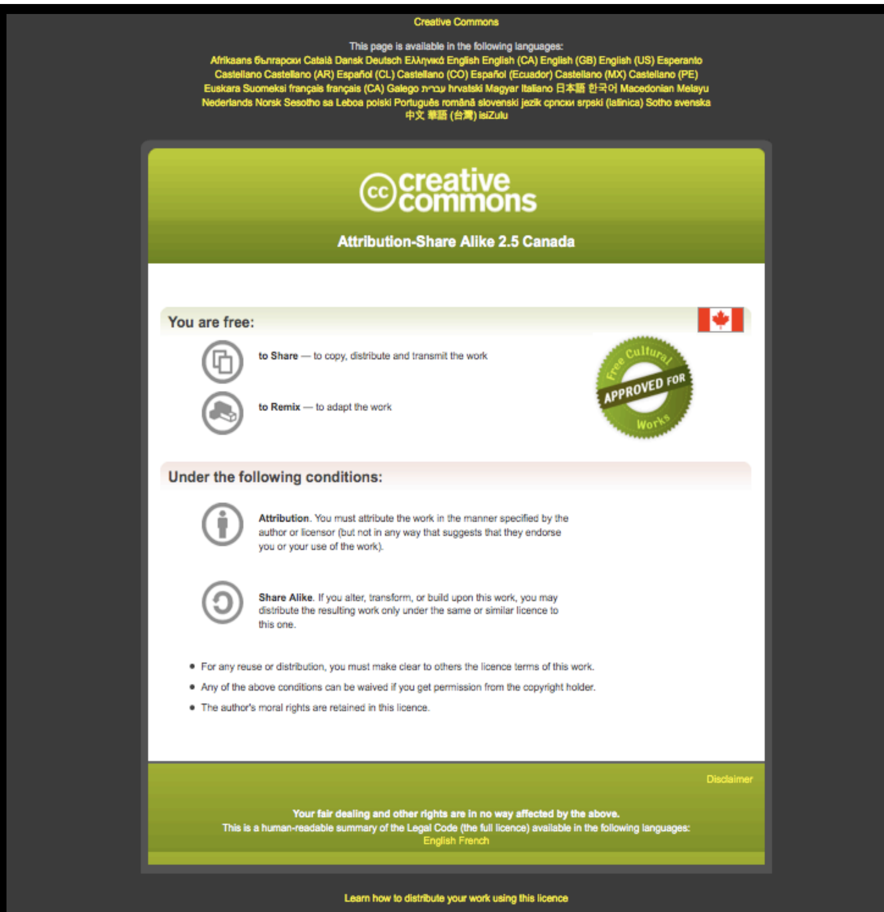
Canadian Bioinformatics Workshops

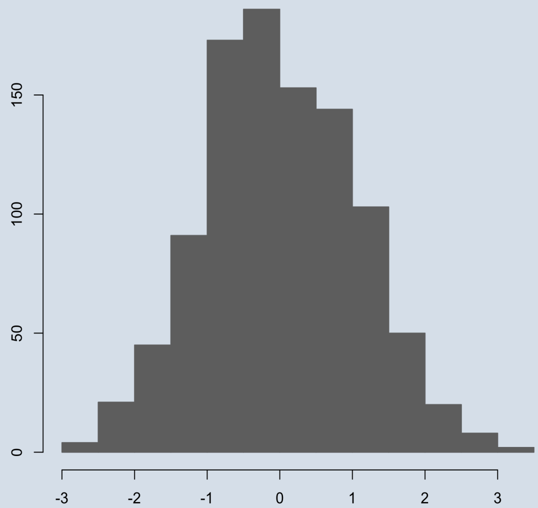
www.bioinformatics.ca



Module 2
Exploratory Data Analysis

Daniele Merico
Exploratory Data Analysis and Essential Statistics using R
January 24-25, 2011

bioinformatics.ca



UNIVERSITY OF TORONTO

Donnelly Centre for Cellular + Biomolecular Research

Post-doctoral Fellow
Donnelly Centre
University of Toronto

http://baderlab.org/DanieleMerico

Goals of Exploratory Data Analysis

- Suggest hypotheses about observed phenomena
  - these can then be tested using *statistical inference* methods
- Identify highly related variables
- Assess assumptions,

  support the selection of appropriate statistical techniques
  - statistical inference will require assumptions on distributions (e.g. normality) and correlation (e.g. independence)
- Identify outliers
  - outliers are sparse anomalies with a strong effect
- Suggest the presence of systematic errors or biases

Philosophy of Exploratory Data Analysis

- Mostly graphical

  it's important to produce clear pictures to help natural pattern recognition
  - Usually avoid 3D graphics (unless you have 3D vision glasses)
  - Don't show too much information

- Understand how the data globally looks like

  But also explore specific features that may be anomalies or interesting patterns

Case Study: Forbes 2004 Data

- 2000 leading companies in 2004 according to Forbes magazine

  adapted from:

  B. S. Everitt, T. Hothorn.
  A Handbook of Statistical Analysis Using R.
  Chapman and Hall (2006)

data originally obtained from HSAUR package

```
library (HSAUR)
data ("Forbes2000", package = "HSAUR")
Forbes.df <- Forbes2000
```

- Read from tab-sep file:

```
Forbes.df <- read.table (
          file = "Forbes_2004.txt",
          sep = "\t", header = T,
          stringsAsFactors = T
          )
```

Inspecting the Structure

```
str (Forbes.df)
```

```
'data.frame':        2000 obs. of  8 variables:
 $ rank       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ name       : Factor w/ 2000 levels "Aareal Bank",..: 438 747 100 659 311 219 870 1827 663 ...
 $ country    : Factor w/ 61 levels "Africa","Australia",..: 60 60 60 60 56 60 56 28 60 60 ...
 $ category   : Factor w/ 27 levels "Aerospace & defense",..: 2 6 16 19 19 2 2 8 9 20 ...
 $ sales      : num  94.7 134.2 76.7 222.9 232.6 ...
 $ profits    : num  17.85 15.59 6.46 20.96 10.27 ...
 $ assets     : num  1264 627 648 167 178 ...
 $ marketvalue: num  255 329 195 277 174 ...
```


Statistical Summary

```
summary (Forbes.df)
```

```
rank                              name                  country                      category
 Min.   :   1.0   Aareal Bank                :   1   United States :751   Banking                  : 313
 1st Qu.: 500.8   ABB Group                  :   1   Japan         :316   Diversified financials: 158
 Median :1000.5   Abbey National             :   1   United Kingdom:137   Insurance                : 112
 Mean   :1000.5   Abbott Laboratories        :   1   Germany       : 65   Utilities                : 110
 3rd Qu.:1500.2   Abercrombie & Fitch        :   1   France        : 63   Materials                :  97
 Max.   :2000.0   Abertis Infraestructuras:   1   Canada        : 56   Oil & gas operations     :  90
                  (Other)                    :1994   (Other)       :612   (Other)                  :1120

     sales              profits              assets             marketvalue
 Min.   :  0.010   Min.   :-25.8300   Min.   :   0.270   Min.   :  0.02
 1st Qu.:  2.018   1st Qu.:  0.0800   1st Qu.:   4.025   1st Qu.:  2.72
 Median :  4.365   Median :  0.2000   Median :   9.345   Median :  5.15
 Mean   :  9.697   Mean   :  0.3811   Mean   :  34.042   Mean   : 11.88
 3rd Qu.:  9.547   3rd Qu.:  0.4400   3rd Qu.:  22.793   3rd Qu.: 10.60
 Max.   :256.330   Max.   : 20.9600   Max.   :1264.030   Max.   :328.54
                   NA's   :  5.0000
```

Statistical Summary

```
summary (Forbes.df)
```

Counts (categorical variables)

```
rank                               name                country            category
 Min.   :   1.0  Aareal Bank                :   1  United States :751  Banking                 : 313
 1st Qu.: 500.8  ABB Group                  :   1  Japan         :316  Diversified financials: 158
 Median :1000.5  Abbey National             :   1  United Kingdom:137  Insurance               : 112
 Mean   :1000.5  Abbott Laboratories        :   1  Germany       : 65  Utilities               : 110
 3rd Qu.:1500.2  Abercrombie & Fitch        :   1  France        : 63  Materials               :  97
 Max.   :2000.0  Abertis Infraestructuras:  1  Canada        : 56  Oil & gas operations  :  90
                 (Other)                    :1994  (Other)       :612  (Other)                 :1120

     sales             profits             assets           marketvalue
 Min.   :  0.010   Min.   :-25.8300   Min.   :   0.270   Min.   :  0.02
 1st Qu.:  2.018   1st Qu.:  0.0800   1st Qu.:   4.025   1st Qu.:  2.72
 Median :  4.365   Median :  0.2000   Median :   9.345   Median :  5.15
 Mean   :  9.697   Mean   :  0.3811   Mean   :  34.042   Mean   : 11.88
 3rd Qu.:  9.547   3rd Qu.:  0.4400   3rd Qu.:  22.793   3rd Qu.: 10.60
 Max.   :256.330   Max.   : 20.9600   Max.   :1264.030   Max.   :328.54
                   NA's   :  5.0000
```

Summary statistics (quantitative variables)

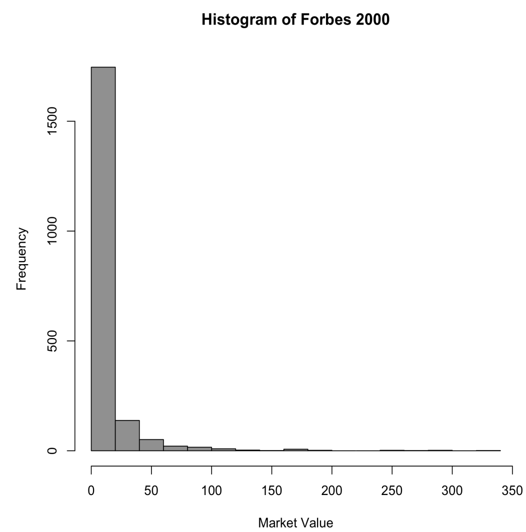
Statistical Summaries Quantitative or Categorical?

- **Quantitative Variables**
  can assume any value in a given numeric range
  (e.g. height, glucose levels, etc…)

- **Categorical Variables**
  can assume only a finite number of values
  are qualitative rather than strictly quantitative
  - **Ordinal** (e.g. severe, mild, absent)
  - **Not ordinal** (e.g. democrat, republican, independent)

Histogram

- Can be used to graphically inspect the distribution of *quantitative* variables
  - Frequencies are determined for regular value intervals



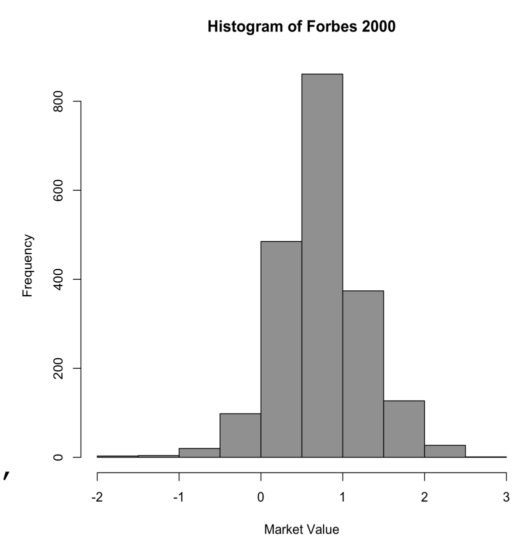

```
hist (
   Forbes.df$marketvalue,
   col = "gray50",
   xlab = "Market Value",
   main = "Histogram of Forbes 2000"
   )
```


Histogram: Scale Change?

- When values are so much accumulated on the low end of the scale,

  with a small number of very high values,

  it's useful to have a look at the **log-transformed** data distribution

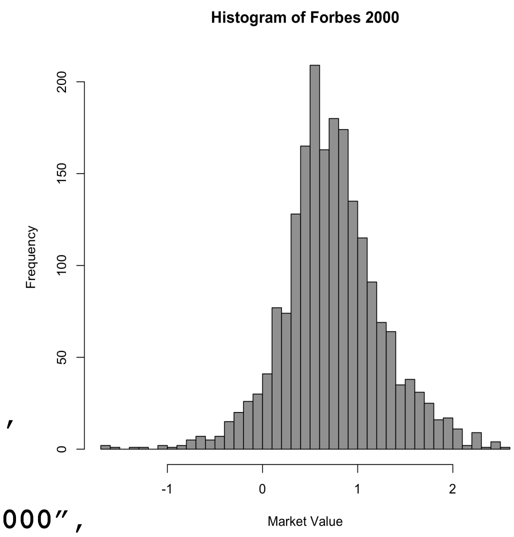



```
hist (
  log10 (Forbes.df$marketvalue),
  col = "gray50",
  xlab = "Market Value",
  main = "Histogram of Forbes 2000"
  )
```

<u>Now the curve is bell-shaped!</u>

Histogram: Modulate Resolution

- To use smaller intervals for the bars use **`breaks`**




```
hist (
 log10 (Forbes.df$marketvalue),
 col = "gray50",
 xlab = "Market Value",
 main = "Histogram of Forbes 2000",
 breaks = 50
 )
```


What's a histogram useful for?

- Is the order of magnitude the same? Do I have to change the scale? (e.g. log-transform)
- Are there positive as well as negative values?
- Are the values all clustered in the same area? Are there values that are more extreme?
- What's the shape of the distribution? What inferential statistics technique can I use? (see next lesson)

Save plot to file

```
dev.copy (
   device = x11,
   file = filename,
   type = "pdf"
   )
dev.off()
```


Statistical Summaries: Central Value

- Mean $M(x) = \frac{1}{N}\sum_{N}^{i=1} x_i$

  sum of all values divided by the number of values

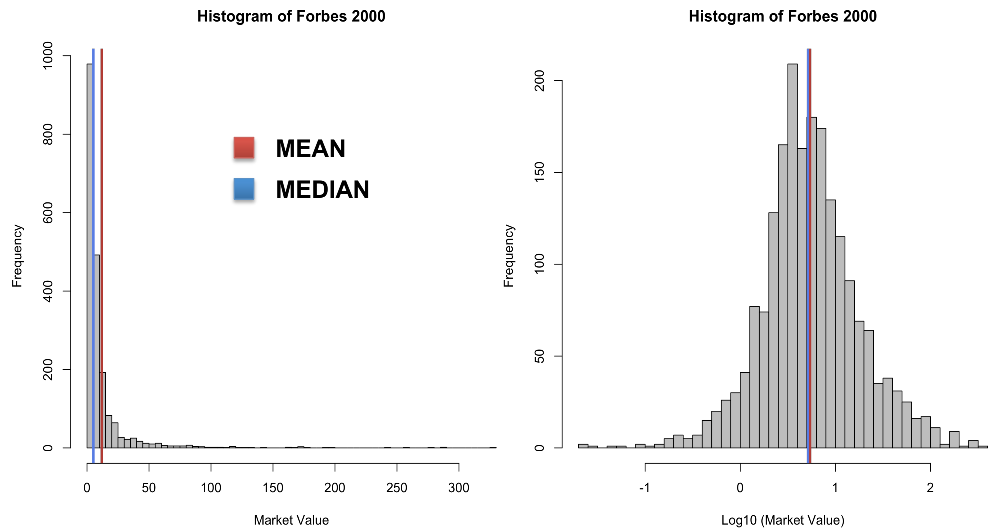
```
mean (Forbes.df$marketvalue)
mean (log10 (Forbes.df$marketvalue))
```

- Median $P(x \geq Mdn(x)) \geq \frac{1}{2} \;\; AND \;\; P(x \geq Mdn(x)) \leq \frac{1}{2}$

  the value separating all other values in two halves

```
median (Forbes.df$marketvalue)
median (log10 (Forbes.df$marketvalue))
```

Mean vs Median



The mean and median differ when:

- The data distribution is *skewed* (not simmetric)
- There are *outliers* (i.e. few anomalies with high values)


```
# Linear Scale

hist ( log10 (Forbes.df$marketvalue),
     col = "gray70",
     xlab = "Log10 (Market Value)",
     main = "Histogram of Forbes 2000",
     breaks = 50)

abline (v = mean (log10 (Forbes.df$marketvalue)),
     col = "brown", lwd = 3)
abline (v = median (log10 (Forbes.df$marketvalue)),
     col = "royalblue", lwd = 3)

dev.copy (
     device = x11,
     file = "Forbes_HistMarketvalue_log10_MeanMdn.pdf",
     type = "pdf")
dev.off ()
```

```
# Log10 Scale

hist (  Forbes.df$marketvalue,
     col = "gray70",
     xlab = "Market Value",
     main = "Histogram of Forbes 2000",
     breaks = 50)

abline (v = mean (Forbes.df$marketvalue),
     col = "brown", lwd = 3)
abline (v = median (Forbes.df$marketvalue),
     col = "royalblue", lwd = 3)

dev.copy (
     device = x11,
     file = "Forbes_HistMarketvalue_MeanMdn.pdf",
     type = "pdf")
dev.off ()
```


Mean vs Median

- Use the mean if
  - The distribution is (almost) symmetric
  - There are lots of tied values
- Use the median if
  - There may be outliers
  - The distribution is markedly asymmetric

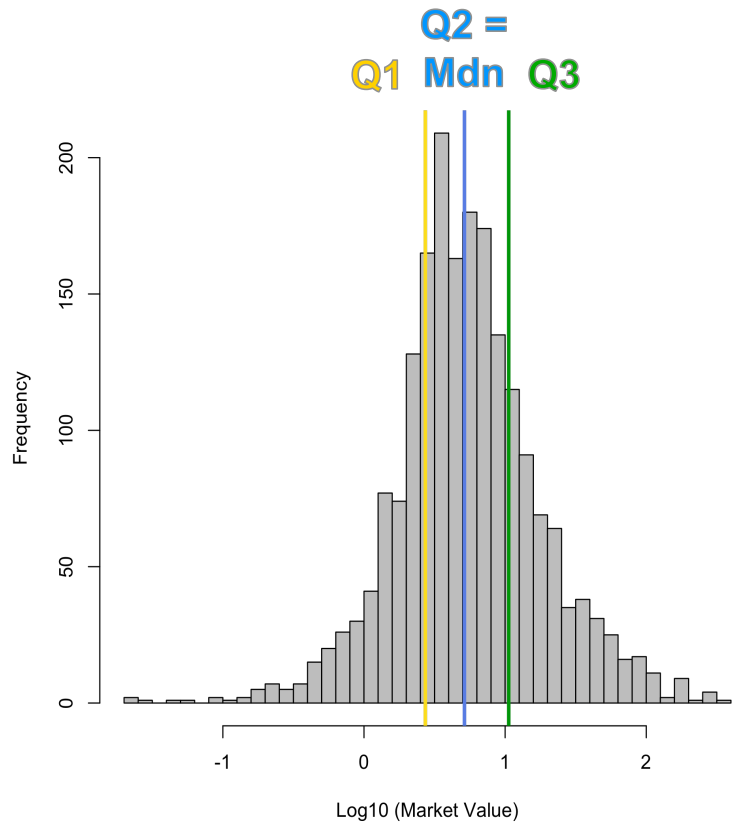
What are central value statistics useful for?

- Where are values clustered?
- Is the distribution symmetric?
  (by comparing mean and median)

Statistical Summaries: Quantiles

$$P(x \le Q_{k/q}(x)) \le k/q$$

- Quantiles are an extension of the median
  0% quantile (i.e. min)
  25% quantile (i.e. 1st quartile): the value separating the data in 25% and 75%
  50% quantile (i.e. median)
  75% quantile (i.e. 3rd quartile): the value separating the data in 75% and 25%
  100% quantile (i.e. max)
- We will see how to use them with
  - the IQR, a measure of spread
  - boxplots

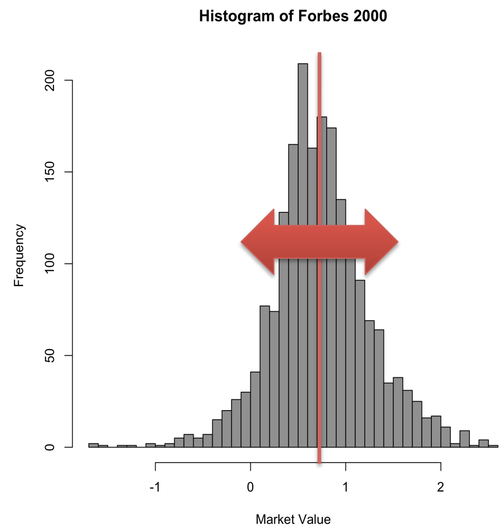


Statistical Summaries: Quantiles

```
# 0-25-50-75-100 quantiles
quantile (log10 (Forbes.df$marketvalue))

# quantile(s) at selected probability(ies)
quantile (log10 (Forbes.df$marketvalue), prob = 0.25)
```

Statistical Summaries: Spread

- Spread statistics give an idea of how much the data differ from the central value



Statistical Summaries: Spread

- Standard Deviation (sd) $$SD(x) = \sqrt{\frac{1}{N}\sum_{N}^{i=1}\left(M(x) - x_i\right)^2}$$

  root square of the mean quadratic difference from the mean

```
sd (log10 (Forbes.df$marketvalue))
```

- Inter-Quartile Range (IQR) $$IQR(x) = Q_{75/100}(x) - Q_{25/100}(x)$$

  difference between the 3rd (75%) and 1st (25%) quartile

```
IQR (log10 (Forbes.df$marketvalue))
```

Statistical Summary

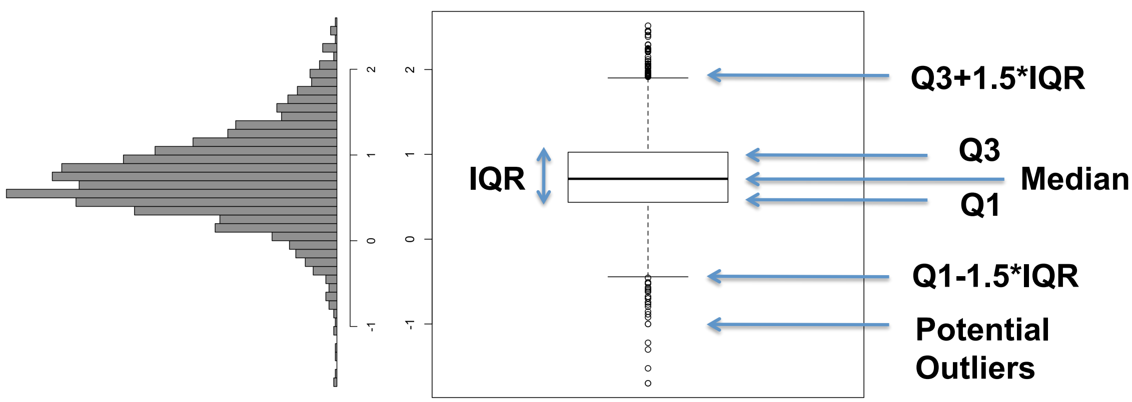
```
summary (Forbes.df)
```

```
rank                            name                  country                   category
 Min.   :   1.0  Aareal Bank                 :   1  United States :751   Banking                  : 313
 1st Qu.: 500.8  ABB Group                   :   1  Japan         :316   Diversified financials: 158
 Median :1000.5  Abbey National              :   1  United Kingdom:137   Insurance                : 112
 Mean   :1000.5  Abbott Laboratories         :   1  Germany       : 65   Utilities                : 110
 3rd Qu.:1500.2  Abercrombie & Fitch         :   1  France        : 63   Materials                :  97
 Max.   :2000.0  Abertis Infraestructuras:   1  Canada        : 56   Oil & gas operations    :  90
                 (Other)                     :1994  (Other)       :612   (Other)                  :1120

     sales            profits              assets           marketvalue
 Min.   :  0.010   Min.   :-25.8300   Min.   :   0.270   Min.   :  0.02
 1st Qu.:  2.018   1st Qu.:  0.0800   1st Qu.:   4.025   1st Qu.:  2.72
 Median :  4.365   Median :  0.2000   Median :   9.345   Median :  5.15
 Mean   :  9.697   Mean   :  0.3811   Mean   :  34.042   Mean   : 11.88
 3rd Qu.:  9.547   3rd Qu.:  0.4400   3rd Qu.:  22.793   3rd Qu.: 10.60
 Max.   :256.330   Max.   : 20.9600   Max.   :1264.030   Max.   :328.54
                   NA's   :  5.0000
```

Summary statistics
(quantitative variables)

Statistical Summaries: Boxplots

- Boxplots are are used to inspect distributions
  - *Histograms* are more informative for *single* distributions
  - *Boxplots* are more practical to *compare* distributions



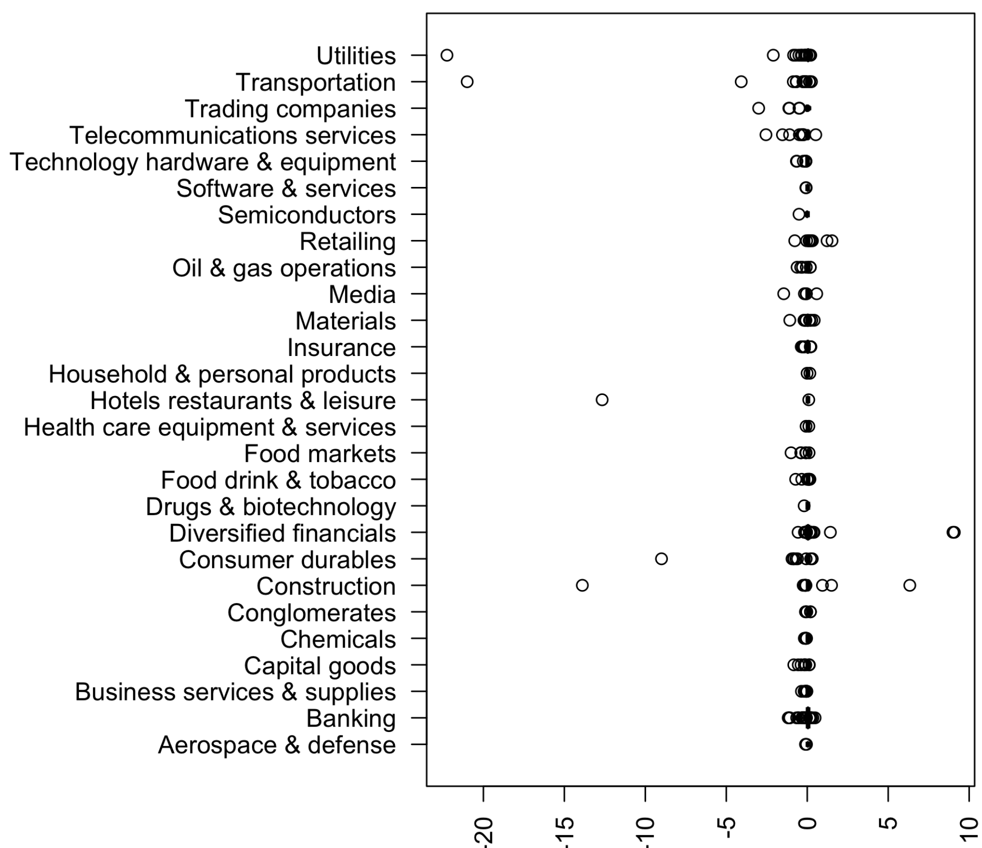

```
boxplot (log10 (Forbes.df$marketvalue))
```

Introduction to R

bioinformatics.ca

Statistical Summaries: Boxplots

- Inspecting the profitability distributions of different business categories
  1. Define *profitability* as profits / marketvalue
  2. Draw boxplots by category



```
Forbes.df$profitability <-
   Forbes.df$profits / Forbes.df$marketvalue

par (omd = c (0.3, 1, 0, 1))
```

Creates some space on the left vertical border for the very long category labels

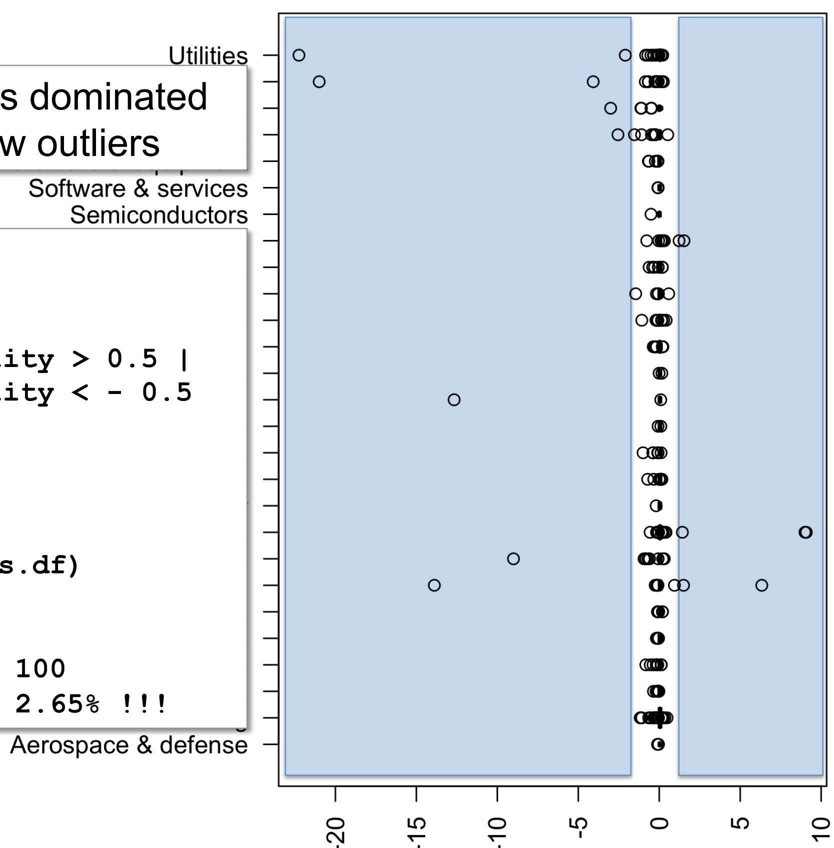
The *formula* means:
plot *profitability* split into different groups by *category*

```
boxplot (
   formula = profitability ~ category,
   data = Forbes.df,
   varwidth = T,
   las = 2,
   horizontal = T)
```

Varwidth = T makes widths proportional to number of values

las = 2 makes labels perpendicular to axis

Horizontal = T makes the boxplots horizontal instead of vertical



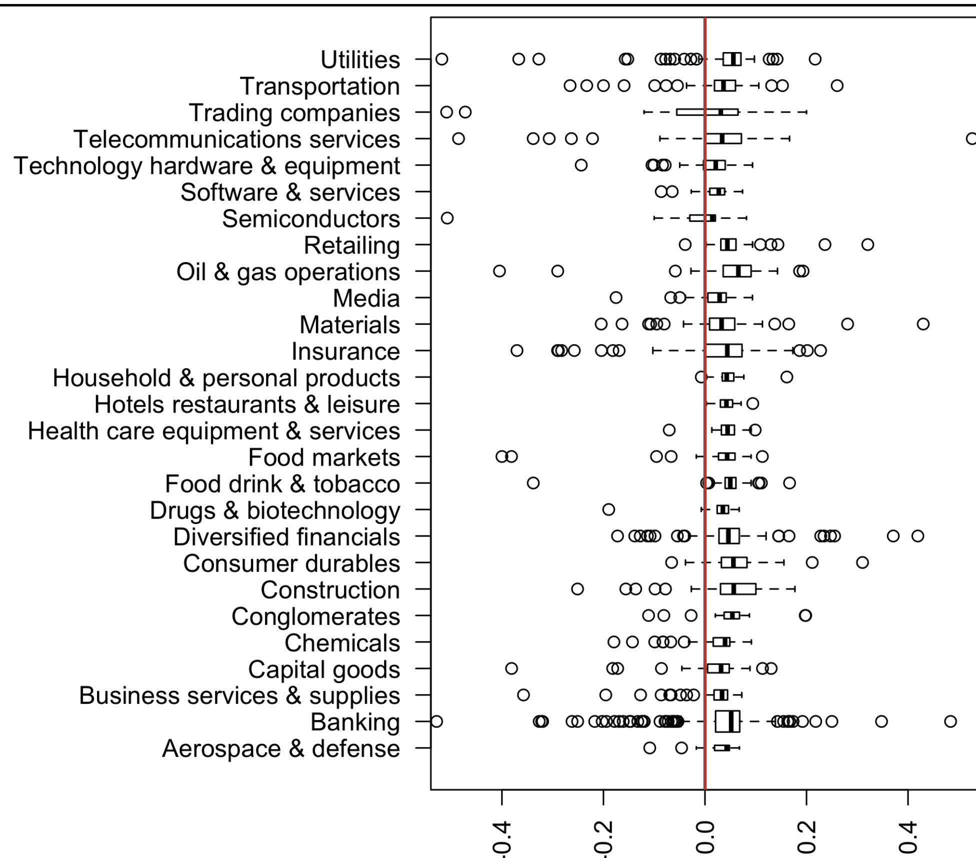
This plot is dominated by a few outliers

```
nrow (
 subset (
 Forbes.df,
 profitability > 0.5 |
 profitability < - 0.5
 )
)
# 53

nrow (Forbes.df)
# 2000

53 / 2000 * 100
# it's only 2.65% !!!
```

- Let's plot only the smaller area between 0.5 and -0.5

```
par (omd = c (0.3, 1, 0, 1))

boxplot (
   formula = profitability ~ category,
   data = Forbes.df,
   varwidth = T,
   ylim = c (-0.5, 0.5),
   las = 2,
   horizontal = T)

abline (v = 0, col = "brown", lwd = 2)
```

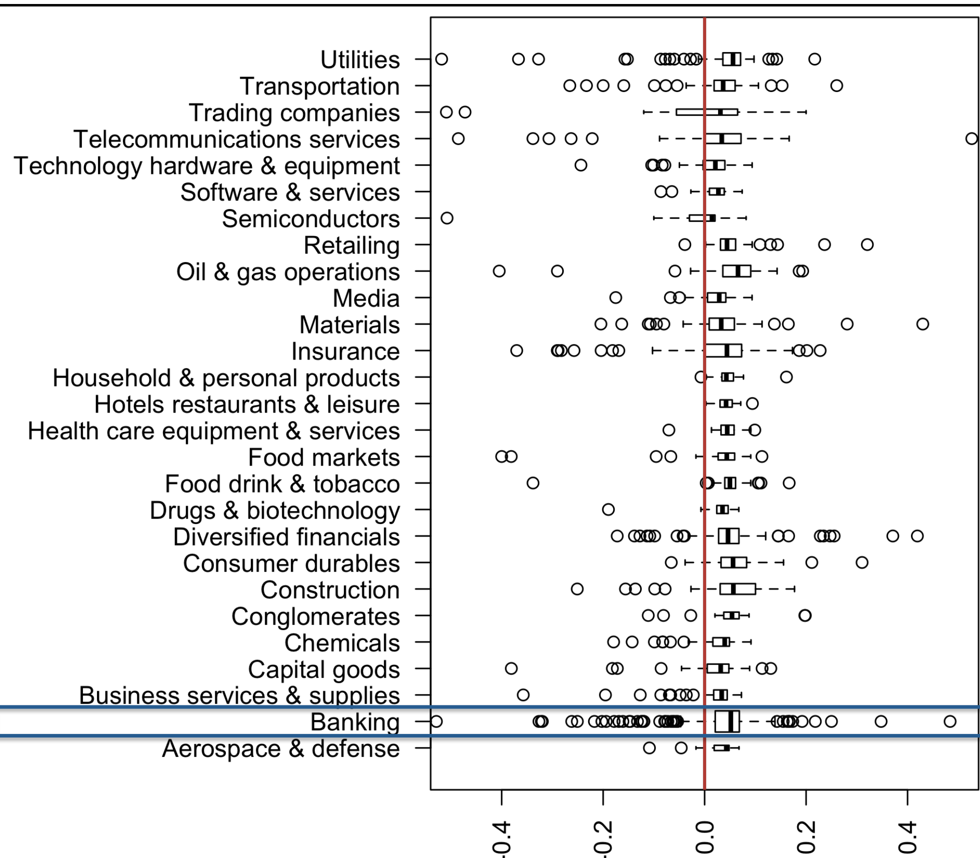
ylim = c (-0.5, 0.5) restrict the plot area to the corresponding interval on the y-axis



Forbes 2000
Profitability Distributions

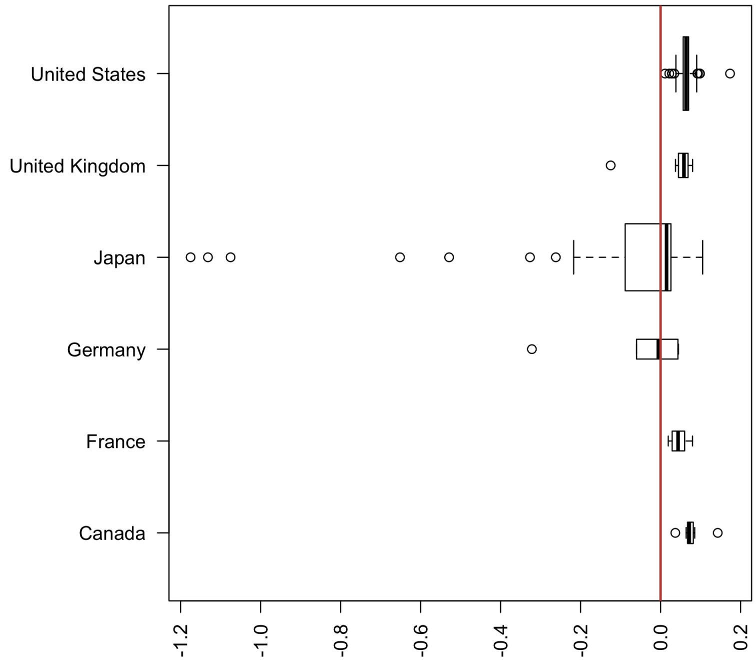
- All sectors are on average profitable (median > 0)
- Some sectors are more profitable on average
- Different sectors have different spreads
- Different sectors have different amount of potential outliers

- We can follow up on these observations, by making more detailed breakdowns
  - E.g.: is there a different distribution by country within specific sectors?



- Some trouble in Japan and Germany…
  (remember this was 2004 and not 2010)

Top 6 countries (in 2004)



This is an example of hypothesis generation:
Banks are more troubled in Japan and Germany than in other developed countries


```
top6countries.chv <- c ("United States", "Canada",
      "United Kingdom", "Germany", "France", "Japan")

Forbes_bktopc.df <- subset (
   Forbes.df,
   category == "Banking" & country %in% top6countries.chv)

Forbes_bktopc.df$country <- factor (Forbes_bktopc.df$country)

par (omd = c (0.1, 1, 0, 1))
boxplot (
   formula = profitability ~ country,
   data = Forbes_bktopc.df,
   las = 2, horizontal = T, varwidth = T)

abline (v = 0, col = "brown", lwd = 2)
```

Statistical Summary

```
summary (Forbes.df)
```

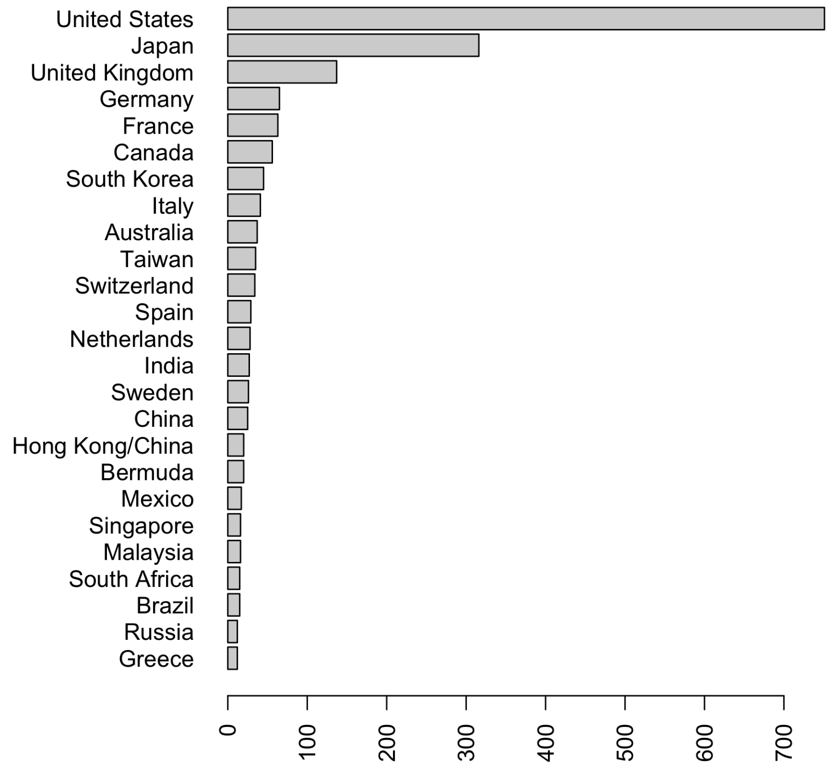
Counts
(categorical variables)

```
rank                              name                country               category
 Min.   :   1.0  Aareal Bank             :   1  United States :751  Banking               : 313
 1st Qu.: 500.8  ABB Group               :   1  Japan         :316  Diversified financials: 158
 Median :1000.5  Abbey National          :   1  United Kingdom:137  Insurance             : 112
 Mean   :1000.5  Abbott Laboratories     :   1  Germany       : 65  Utilities             : 110
 3rd Qu.:1500.2  Abercrombie & Fitch     :   1  France        : 63  Materials             :  97
 Max.   :2000.0  Abertis Infraestructuras:   1  Canada        : 56  Oil & gas operations  :  90
                 (Other)                 :1994  (Other)       :612  (Other)               :1120

     sales            profits             assets            marketvalue
 Min.   :  0.010  Min.   :-25.8300  Min.   :   0.270  Min.   :  0.02
 1st Qu.:  2.018  1st Qu.:  0.0800  1st Qu.:   4.025  1st Qu.:  2.72
 Median :  4.365  Median :  0.2000  Median :   9.345  Median :  5.15
 Mean   :  9.697  Mean   :  0.3811  Mean   :  34.042  Mean   : 11.88
 3rd Qu.:  9.547  3rd Qu.:  0.4400  3rd Qu.:  22.793  3rd Qu.: 10.60
 Max.   :256.330  Max.   : 20.9600  Max.   :1264.030  Max.   :328.54
                  NA's   :  5.0000
```


Barplot

- Can be used to graphically inspect the distribution of *categorical* variables
  - For example, let's plot the number of companies in the Forbes2000 for the top 25 countries

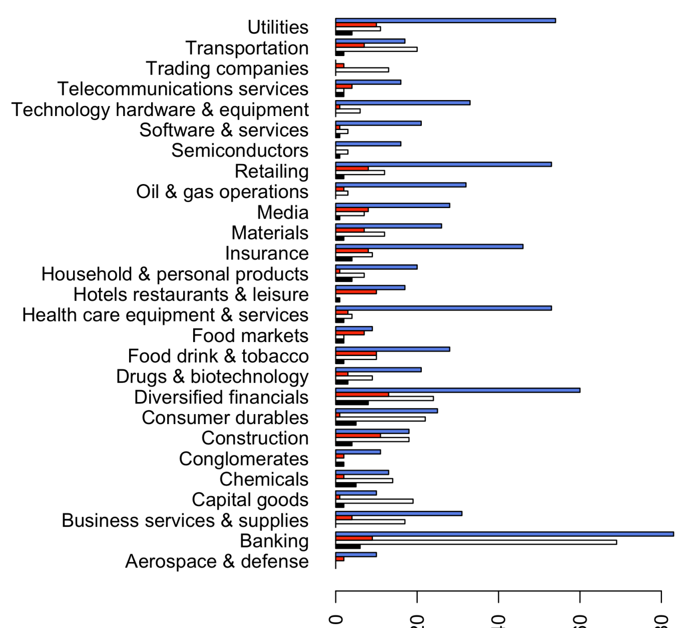
```
Top25countries.tab <-
    sort (table (Forbes.df$country), decreasing = T)[1: 25]
par (omd = c (0.2, 1, 0, 1))
barplot (
    sort (Top25countries.tab, decreasing = F),
    las = 2, horiz = T)
```



Relations between Categorical Variables

- As we did with boxplots, we can break down distributions by categories

U.S.
U.K.
Japan
Germany



```
top4countries.chv <- c (
   "United States", "United Kingdom", "Germany", "Japan")

Forbes_top4c.df <- subset (
   Forbes.df,
   country %in% top4countries.chv)

Forbes_top4c.df$country <- factor (Forbes_top4c.df$country)

Forbes_top4c.tab <- table (
   Forbes_top4c.df[, c ("country", "category")])

par (omd = c (0.3, 1, 0, 1))

barplot (
   Forbes_top4c.tab,
   beside = T,
   horiz = T, las = 2,
   col = c ("black", "white", "red", "royalblue"))
```


Relations between Quantitative Variables

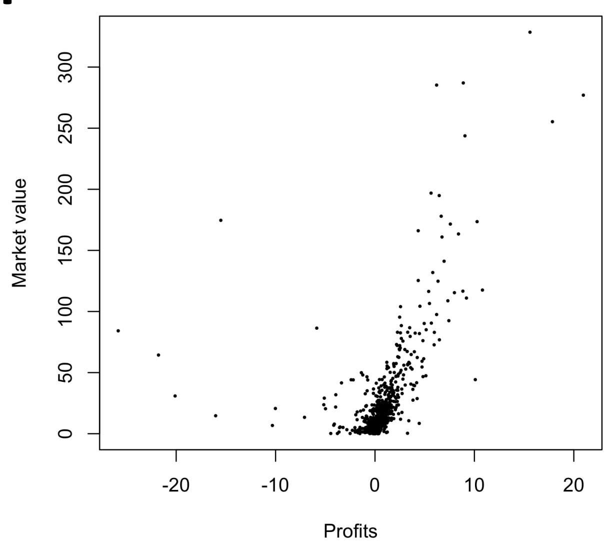
Scatterplots and Correlation

- We now explore relations between quantitative variables
- We start by looking at pairs of variables at a time
  1. Profits and Market Value
  2. Sales and Market Value

- Before proceeding we check how many NA values we have and we get rid of the corresponding rows (i.e. companies)

```
notna.ix <- which (!is.na (Forbes.df$profits))

length (notna.ix) / nrow (Forbes.df)
# 0.9975

Forbes_nna.df <- Forbes.df[notna.ix, ]
```


Scatterplot: Profits and Market Value




```
plot ( data = Forbes_nna.df,
       marketvalue ~ profits,
       main = "All Forbes Companies",
       xlab = "Profits",
       ylab = "Market value",
       pch = 19, cex = 0.25)
```

Linear Dependence: Pearson and Spearman Correlation

- **Pearson Correlation** $$\rho_{\mathrm{Pr}}(x) = \frac{\overbrace{M\big((M(x)-x)\cdot(M(y)-y)\big)}^{\text{covariance}}}{SD(x)\cdot SD(y)}$$

  better when variables are in the same scale and there are no outliers

- **Spearman Correlation** $$\rho_{Sp}(x) = 1 - \frac{6\sum_{N}^{i=1} d_i^2}{n\left(n^2-1\right)}$$ d = rank difference

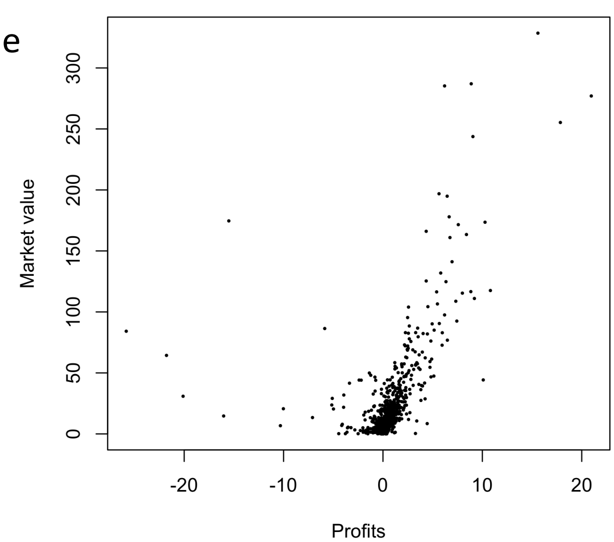
  works on ranks
  → better when variables are in different scales or there are outliers

Profits and Market Value

- The variables seem to be in the same scale
- There are outliers

→ Prefer Spearman

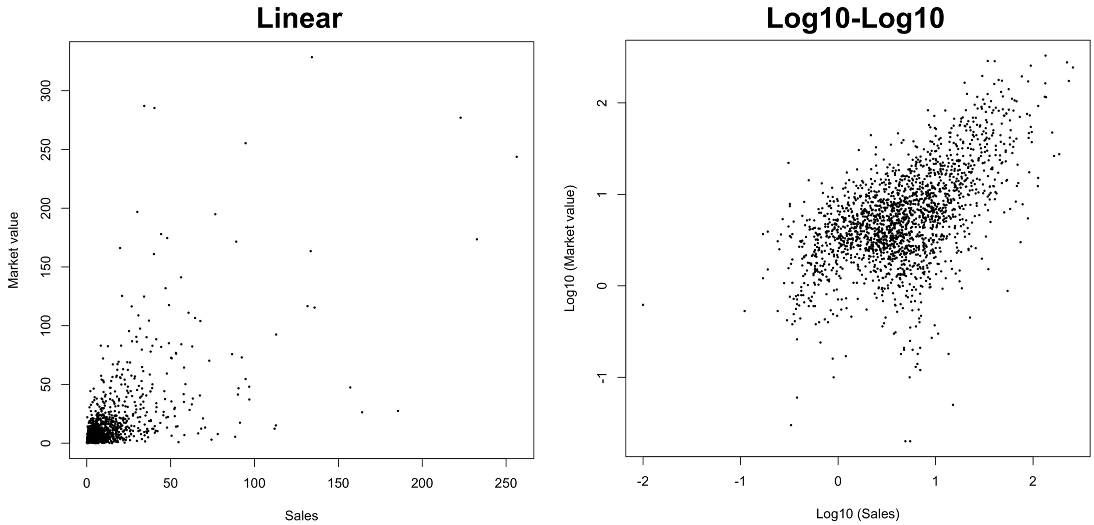
- Pearson: 0.55
- **Spearman: 0.63**




```
cor (Forbes_nna.df$marketvalue, Forbes_nna.df$profits, method = "spearman")
cor (Forbes_nna.df$marketvalue, Forbes_nna.df$profits, method = "pearson")
```

Sales and Market Value

- The scatterplot is more insightful when both variables are in log-scale



Pearson: 0.64
Spearman: 0.53

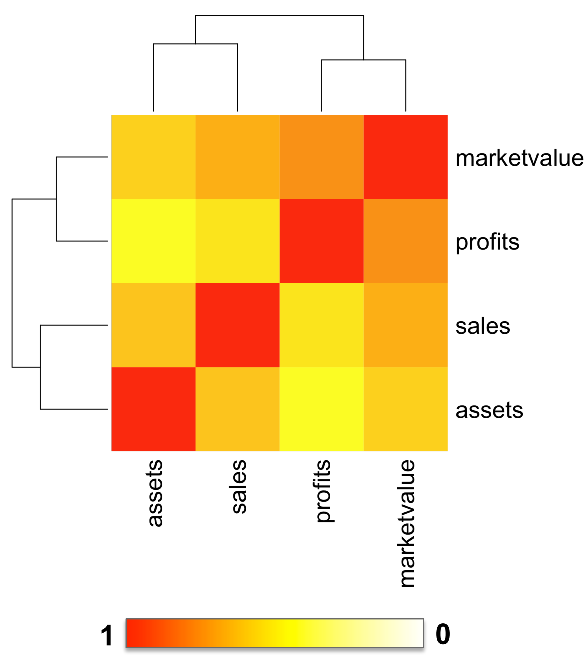
Pearson: 0.56
Spearman: 0.53

Visualizing Correlation Matrices

- If we want to visualize all correlation patterns

  we can generate a correlation matrix

  and visualize it using a **heat-map**



Observations:

- *Market value is correlated to Profits, Sales and Assets (in the order)*
- *Sales and Assets are more correlated to each other and to Market value than to Profits*

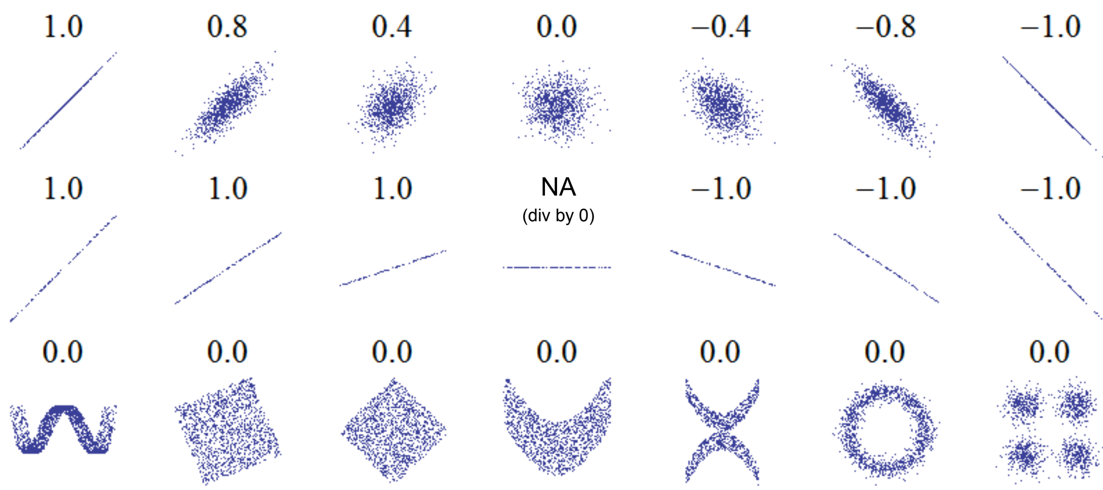
```
cor_sp.mx <- cor (
   Forbes_nna.df[, c ("marketvalue", "assets", "sales", "profits")],
   method = "spearman")

par (omd = c (0.1, 0.9, 0.1, 0.9))

# set color range for heat.colors ()
# (color generation function)
# * 1     = red
# * tot.n = white
#   (the larger tot.n the higher the number of intermediate hues)
tot.n <- 50
max.n <- round (tot.n - max (cor_sp.mx) * (tot.n - 1))
min.n <- round (tot.n - min (cor_sp.mx) * (tot.n - 1))

# col: from min to max
# scale: not needed when handling corr values (1-0 range)
heatmap (
   cor_sp.mx,
   col = heat.colors (n = tot.n)[min.n: max.n],
   scale = "none")
```


Correlation *is Not* Causation

- Whenever we look at correlation between variables we must remember to be cautious in inferring *causal* relations
- This holds true for *association* between categorical variables as well
  - Example:
    Poliomyelitis incidence is correlated to ice cream and soda consumption…
    …but only because poliomyelitis outbreaks are more common in summer time, when ice cream and soda consumption are high
    (this is an example of a ***lurking variable***)

Dependence Beyond Linearity

- Pearson correlation for linear and non-linear dependence

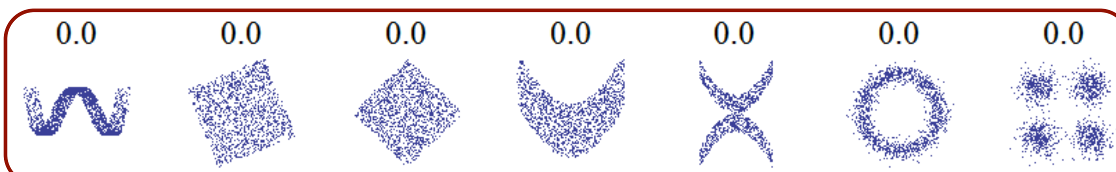


http://en.wikipedia.org/wiki/Correlation_and_dependence

Dependence Beyond Linearity

- Correlation is incapable of detecting non-linear patterns of dependence

- Other statistics should be used in such cases



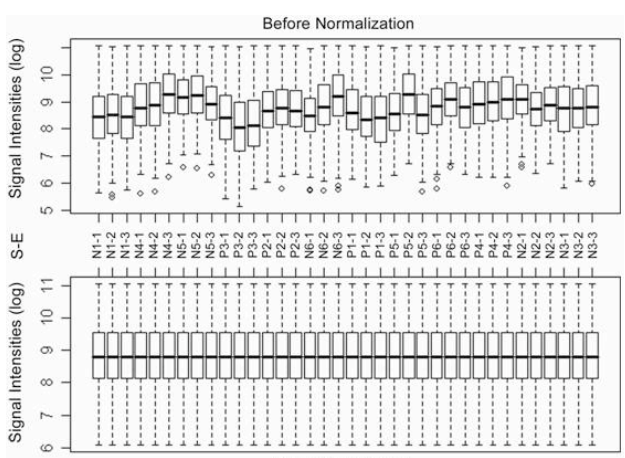
http://en.wikipedia.org/wiki/Correlation_and_dependence

Advanced Exploratory Analysis Techniques

- **Clustering** and **Principal Components Analysis** (lessons tomorrow) can be used as more advanced tools for exploratory analysis
  - How are country similar/dissimilar from each other on the basis of:
    - the number of companies in each sector?
    - the total market value of each sector?

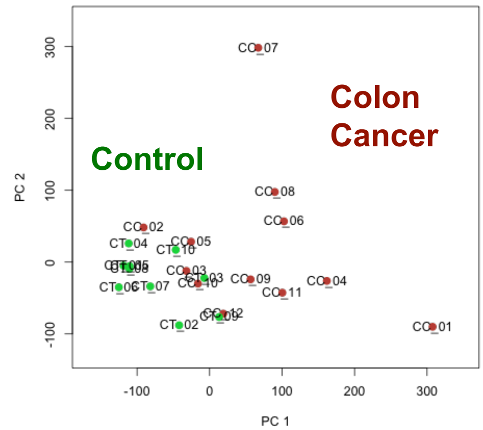
Exploratory Analysis For Microarray Data

- **Boxplots** are typically used to evaluate the signal distribution for different samples and apply correction techniques (***normalization***) if these differ



Exploratory Analysis For Microarray Data

- **Clustering** and **Principal Components Analysis** are typically used to evaluate the similarity/dissimilarity among samples and check if they are compatible with the ***experimental design***



Lab Assignments

- Reproduce the plots in these slides using the code provided
- Tweak the parameters and see what happens
- Use the help pages for the commands you don't understand

References

- Books
  - Tukey, John Wilder. *Exploratory Data Analysis*. Addison-Wesley (1977)
- Online material
  - Free online course hosted by Carnagie-Mellon
    http://oli.web.cmu.edu/openlearning/forstudents/freecourses/statistics
  - Exploratory data analysis for microarray data
    http://baderlab.org/DanieleMerico#Educational

We are on a Coffee Break & Networking Session