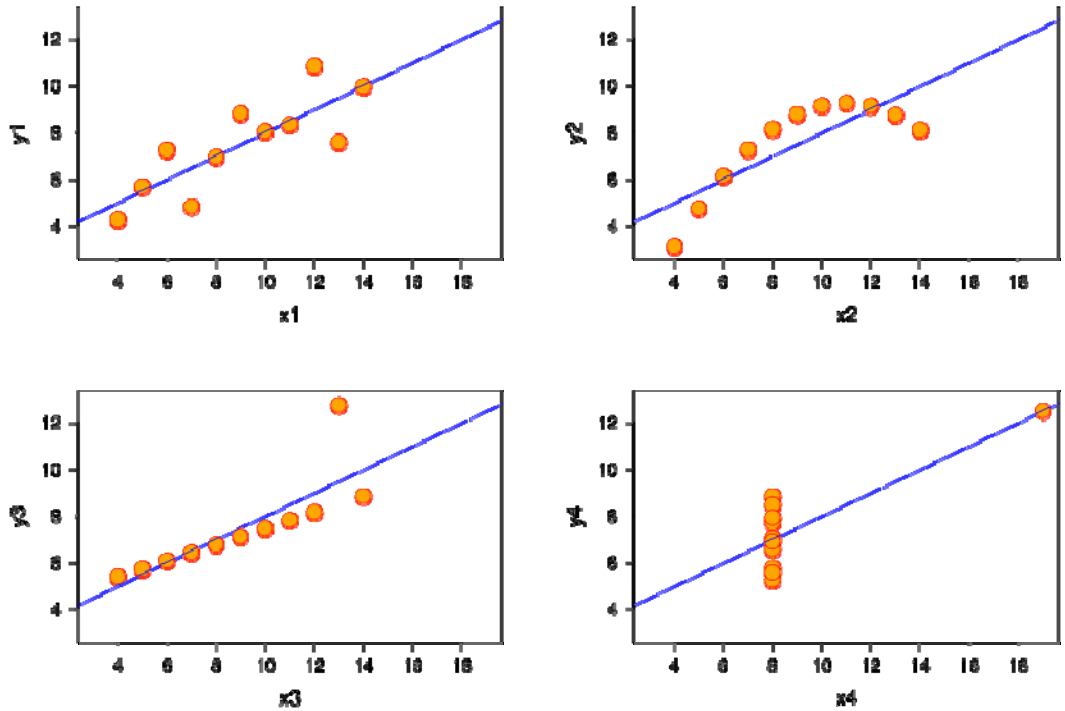
…on the importance of looking at your data

- Look at your data!
- Introduce the plot() and xyplot() functions for graphing data

Anscombe's Quartet

- All 4 datasets have:
  - $Mean_x$ 9.0
  - $Variance_x$ 10.0
  - $Mean_y$ 7.5
  - $Variance_y$ 3.75
  - Cor(x,y) 0.816
  - Regression
    - y= 3 + 0.5x



Check Wikipedia for more information…

Files to use

See http://leonardlab.biology.dal.ca/Bob

- ans.txt
  - Raw data of the quartet, as comma-delimited text
- anscombe.R
  - Script for looking at the data's summary statistics and graphs
- makeAns.R (optional)
  - "Longhand" file used to create "ans.txt"
    - (could have also been done in Excel – save as .csv)

(see anscombe.R for the rest!)

Shortcut! by()

- See anscombe2.R for (optional) shortcut code which uses the by() function

R Console

File Edit Misc Packages Help

```
> rm(list=ls())
> ans<-read.csv(file="ans.txt")
> summary(ans)
       X               Y           Set
 Min.   : 4   Min.   : 3.100   I  :11
 1st Qu.: 7   1st Qu.: 6.117   II :11
 Median : 8   Median : 7.520   III:11
 Mean   : 9   Mean   : 7.501   IV :11
 3rd Qu.:11   3rd Qu.: 8.748
 Max.   :19   Max.   :12.740
>
> ans[c(1:5,12:16,23:27),]
    X     Y Set
1  10  8.04   I
2   8  6.95   I
3  13  7.58   I
4   9  8.81   I
5  11  8.33   I
12 10  9.14  II
13  8  8.14  II
14 13  8.74  II
15  9  8.77  II
16 11  9.26  II
23 10  7.46 III
24  8  6.77 III
25 13 12.74 III
26  9  7.11 III
27 11  7.81 III
>
```

Subsample of ans for illustration

by() applies a function to subgroups of the dataset (specified by "INDICES")

```
> by(ans$X, INDICES = ans$Set, mean)
ans$Set: I
[1] 9
------------------------------------------------------------
ans$Set: II
[1] 9
------------------------------------------------------------
ans$Set: III
[1] 9
------------------------------------------------------------
ans$Set: IV
[1] 9
>
```