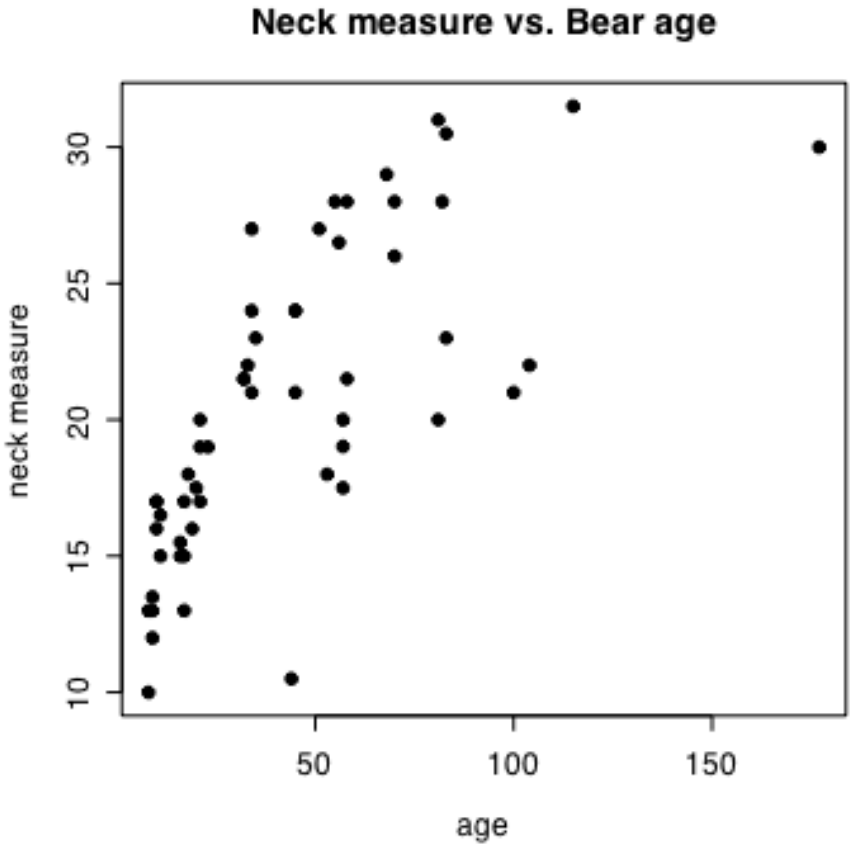
Chapter 7

- Scatterplots,
  Association,
  and Correlation

Scatterplots & Correlation

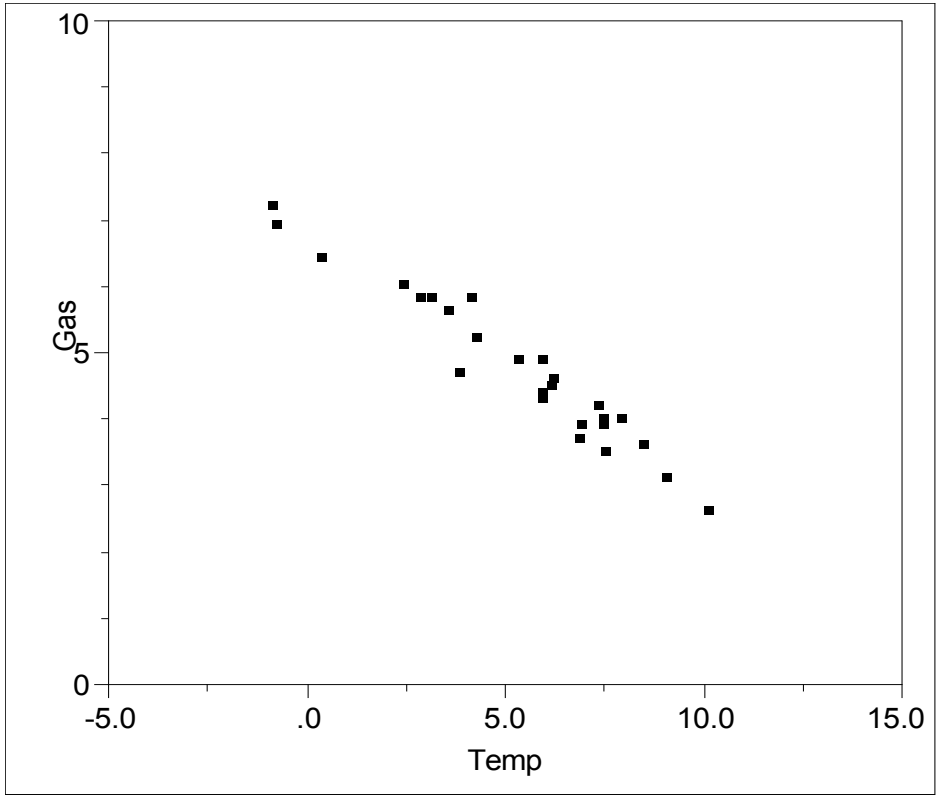
- Here, we see a **positive** relationship between a bear's age and its neck diameter.



As a bear gets older, it tends to have a larger neck.

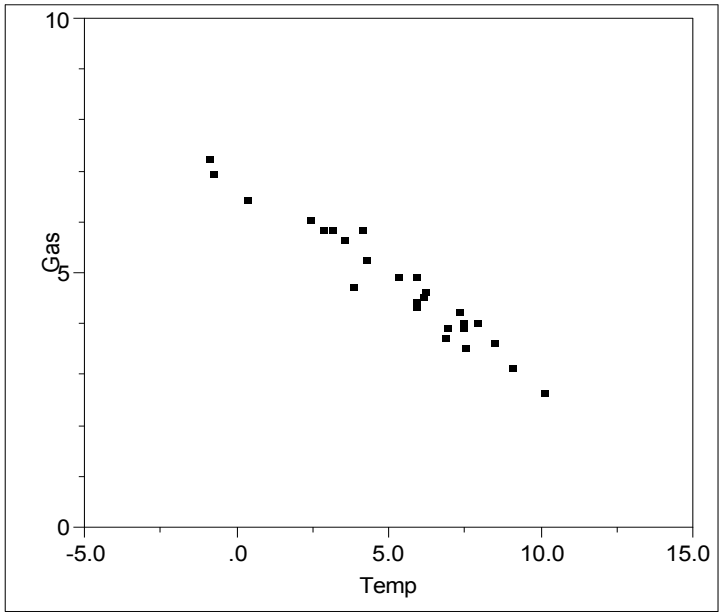
Negative Association

- Outside temperature and amount of natural gas used.
- These variables have a **negative** correlation…
  - Days with higher temperature tend to use less natural gas.
  - Higher temperature→ Less gas used



Scatterplots & Correlation

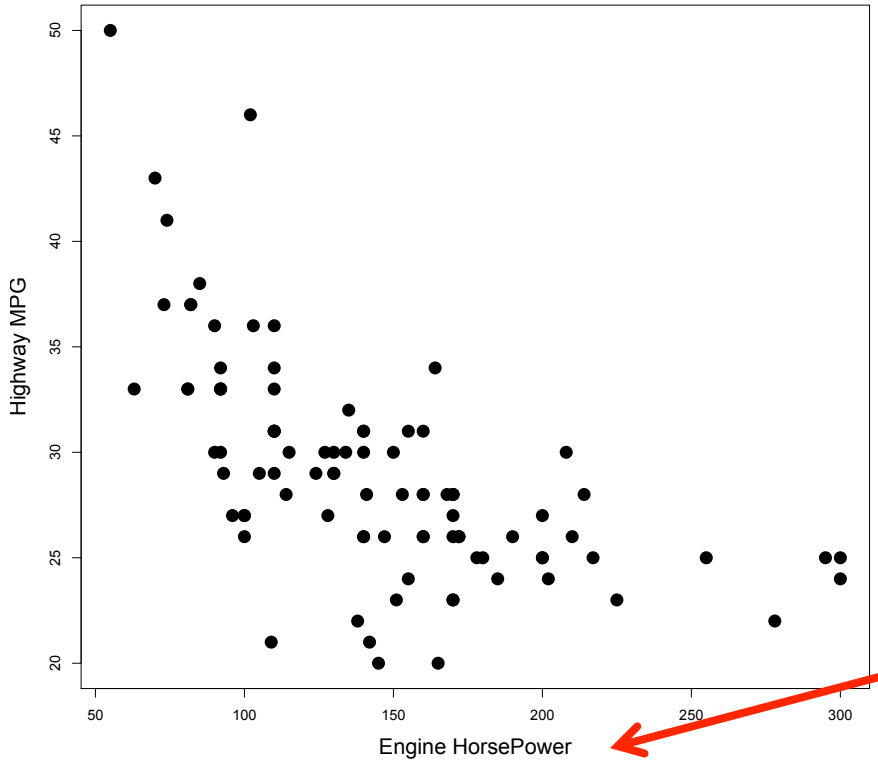
- When the two variables of interest are **continuous variables**, we can plot their relationship with a **scatterplot** (or scatter diagram).



- A scatterplot gives you a quick look at the general relationship between the variables.
- Each observation provides one point on the plot.

- Response variable – plotted on the vertical axis.
  - Also called the dependent variable.
- Explanatory variable – plotted on the horizontal axis.
  - Used to try to explain variation in the response variable.
  - Also called the independent variable.

HWY-mpg is the response variable



Here, we use Engine HPW to explain the variability in HWY-mpg.

Engine Horsepower is the explanatory variable

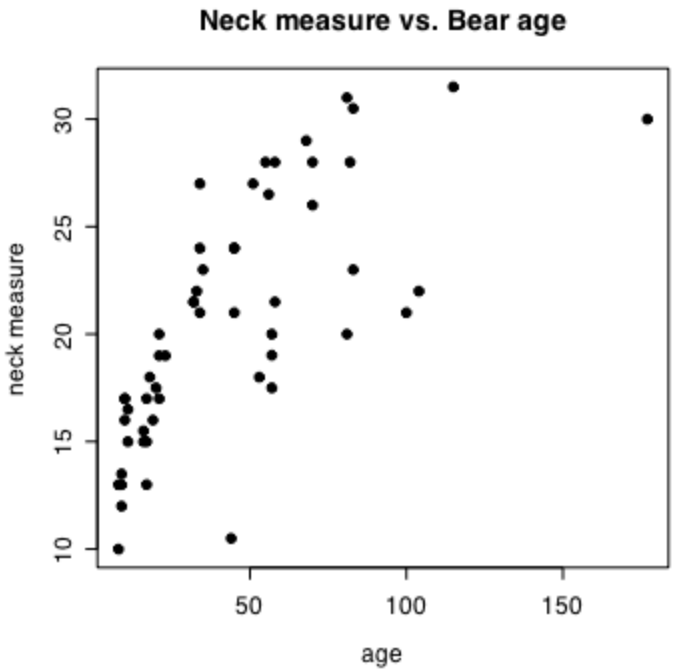
Correlation and Association

Definition

A **correlation** exists between two variables when higher values of one variable consistently go with higher values of another variable or when higher values of one variable consistently go with lower values of another variable.

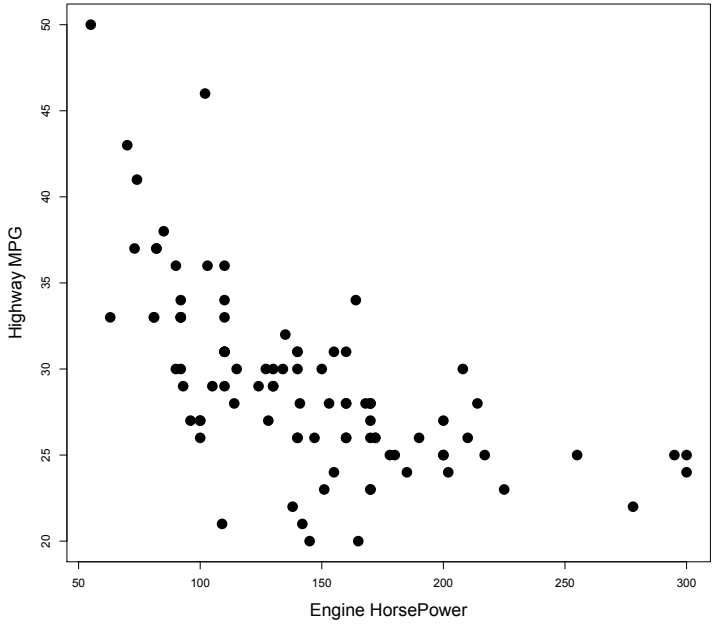
- When describing relationships, we use the terms correlation and association interchangeably. If variables are **correlated**, we say they are **associated**.

Positive Association
(correlation)



- Positive Association
  - Above average values of Age are associated with above average values of Neck Measure (age-high goes with neck-high)
  - Below average values of Age are associated with below average values of Neck Measure(age-low goes with neck-low)

Negative Association
(correlation)



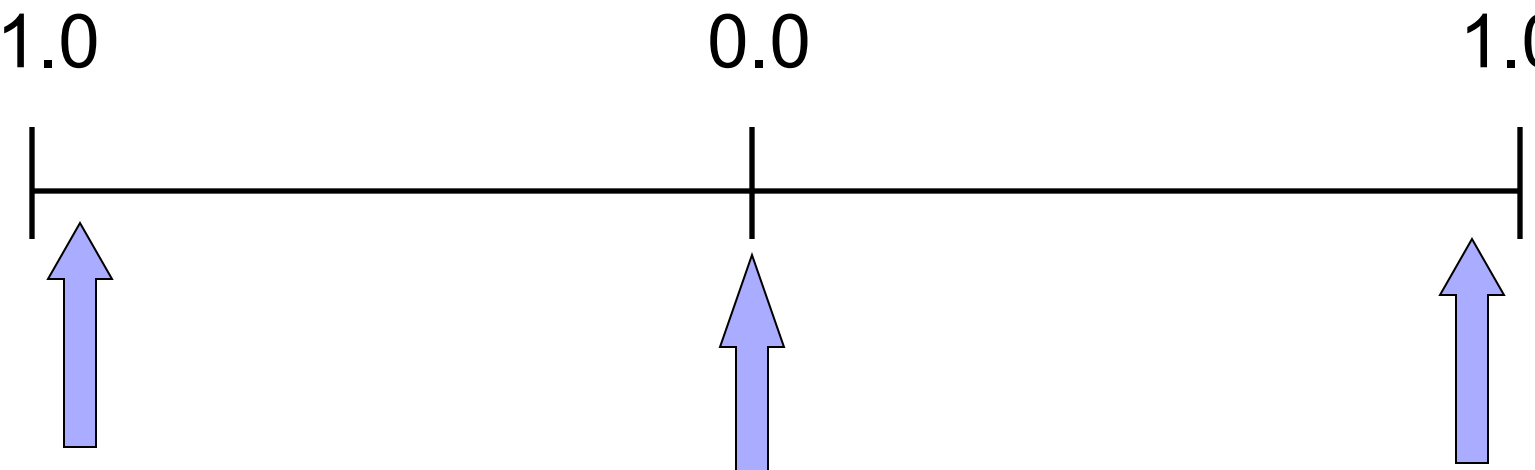
- Negative Association
  - Below average values of Engine HPW are associated with above average values of HWY-mpg (HPW-low goes with MPG-high).
  - Above average values of Engine HPW are associated with below average values of HWY-mpg (HPW-high goes with MPG-low).

Strength of Association

- Correlation applies only to **quantitative (continuous) variables**.
- Correlation measures the strength of **linear** association.
- The **correlation coefficient** (r) gives the direction of the linear association and quantifies the strength of the linear association between two quantitative variables.
- Correlation is a `unitless' quantity (not in 'feet' or 'inches'… no units)

Strength of Association

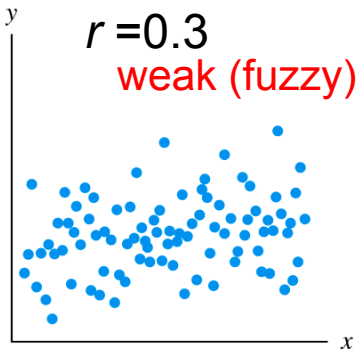
Correlation Coefficient (*r*) will be between -1 and 1.



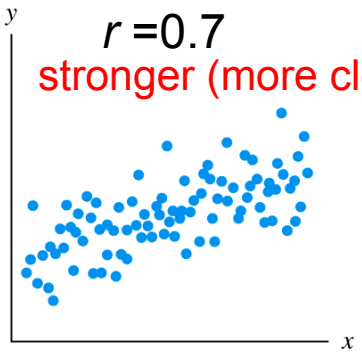
Strong Negative Linear Relationship

Very Weak or No Linear Relationship

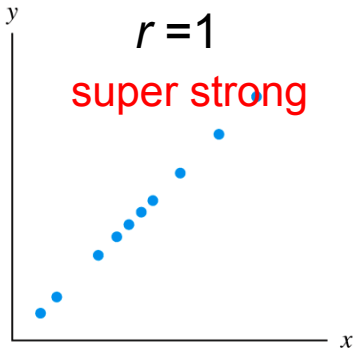
Strong Positive Linear Relationship



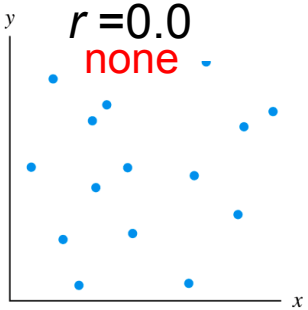
(a) Weak positive correlation between x and y



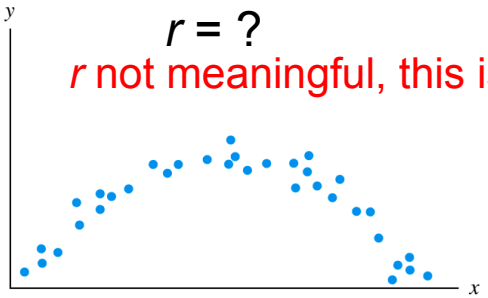
(b) Strong positive correlation between x and y



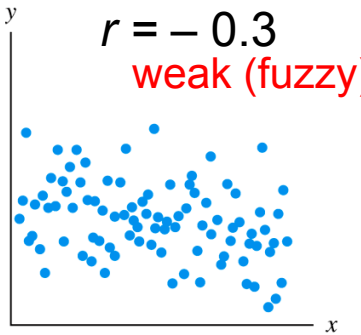
(c) Perfect positive correlation between x and y



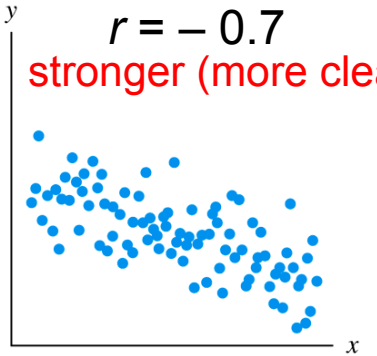
(g) No correlation between x and y



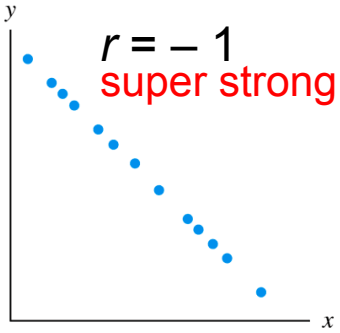
(h) Nonlinear correlation between x and y



(d) Weak negative correlation between x and y

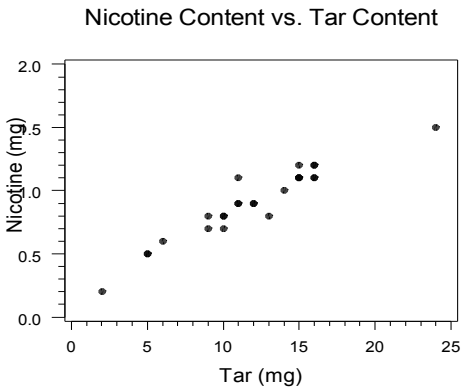


(e) Strong negative correlation between x and y



(f) Perfect negative correlation between x and y

Things to look for in a scatterplot



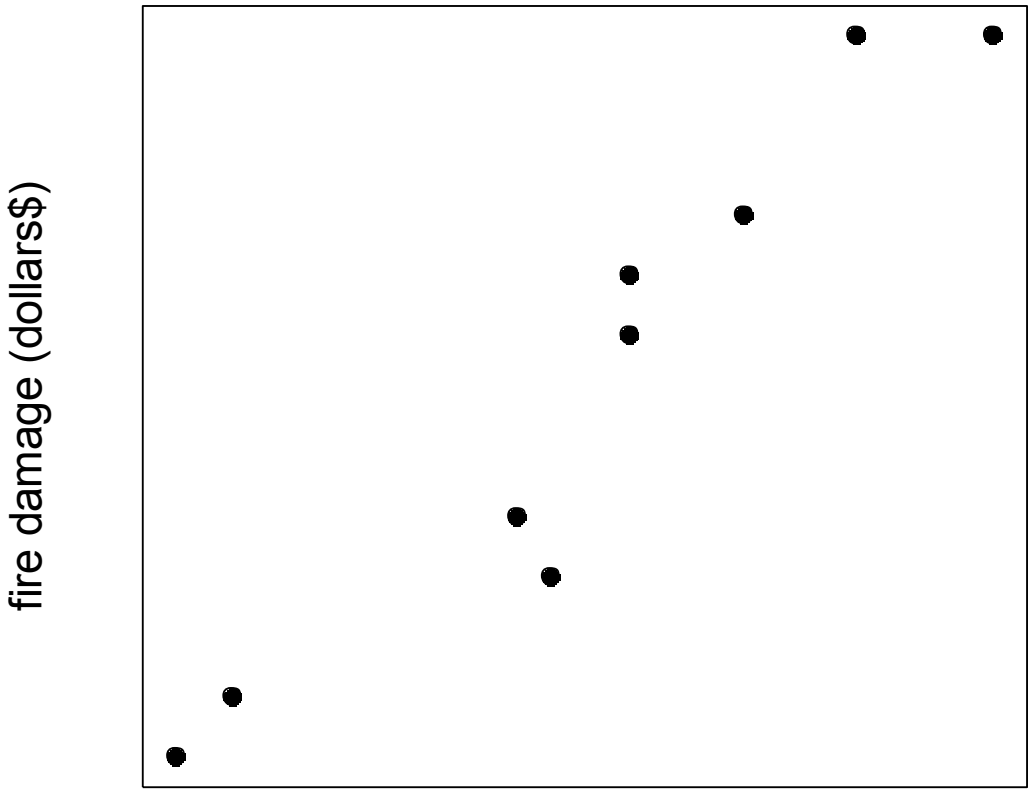
- 1. Direction of association
  - Positive or negative.
- 2. Form of association
  - Linear, curved, clustered, scattered (no relationship).
- 3. Strength of association
  - How closely the points follow a clear form.
- 4. Outliers
  - A point that lies outside of the general pattern.

Association vs. Causation

- The existence of an association does not equate to causation.
- To imply that a change in one variable **causes** a change in another is a *very* strong statement – use ‘association’ for our relationships in this class.

Beware of lurking variables

- **Lurking variable** – a hidden variable that stands behind a relationship and affects the other two variables.



Size of fire?

Association vs. Causation

- Increasing the size of the fire will cause greater damage.
- Increasing the number of firefighters at the fire will not cause greater damage, but we do tend to see more firefighters at larger fires.
- Correlation does NOT imply causality.

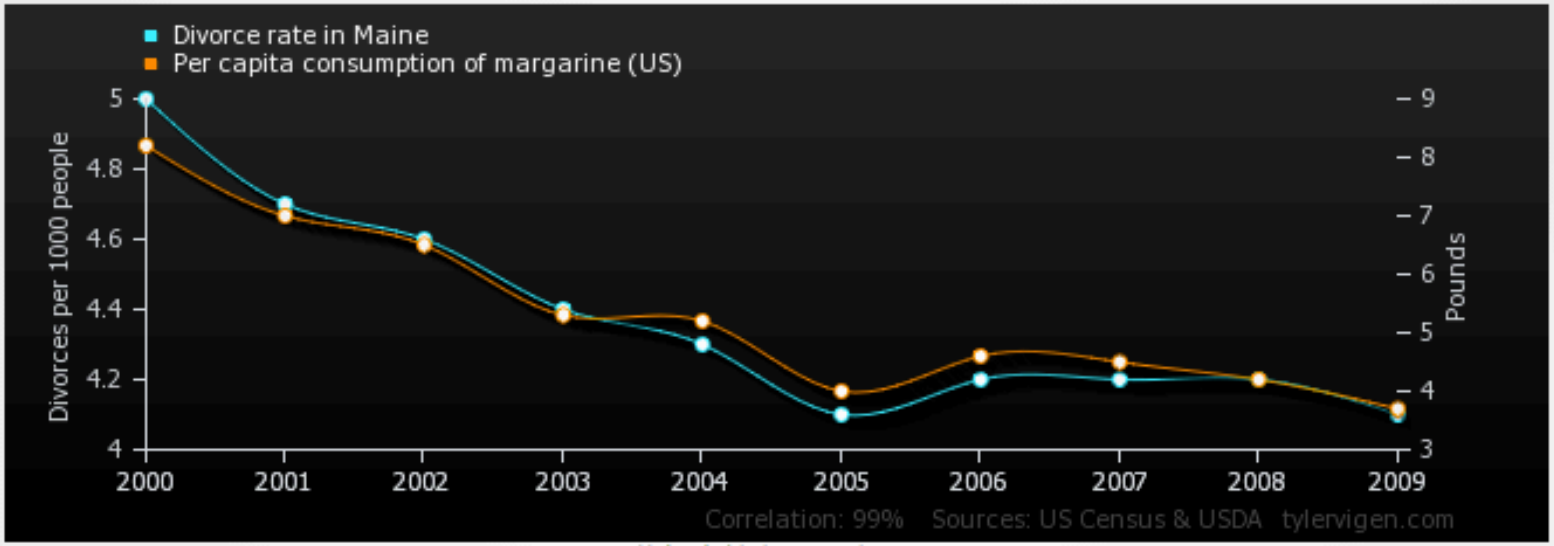
Correlation Cautions

- Don't confuse correlation with causation.
  - There is a strong positive correlation between shoe size and intelligence.
- Beware of lurking variables.
- Beware of totally coincidental associations… (next slide)

Divorce rate in Maine

correlates with

Per capita consumption of margarine (US)



Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

http://io9.gizmodo.com/our-new-favorite-website-spurious-correlations-1574464459

Simpson's Paradox

- A statistical relationship between two variables can be reversed by including additional factors in the analysis.

- Sometimes a simple Y vs. X plot can give a false impression (be careful).

Example:
SAT score vs. public $$$ spent on education

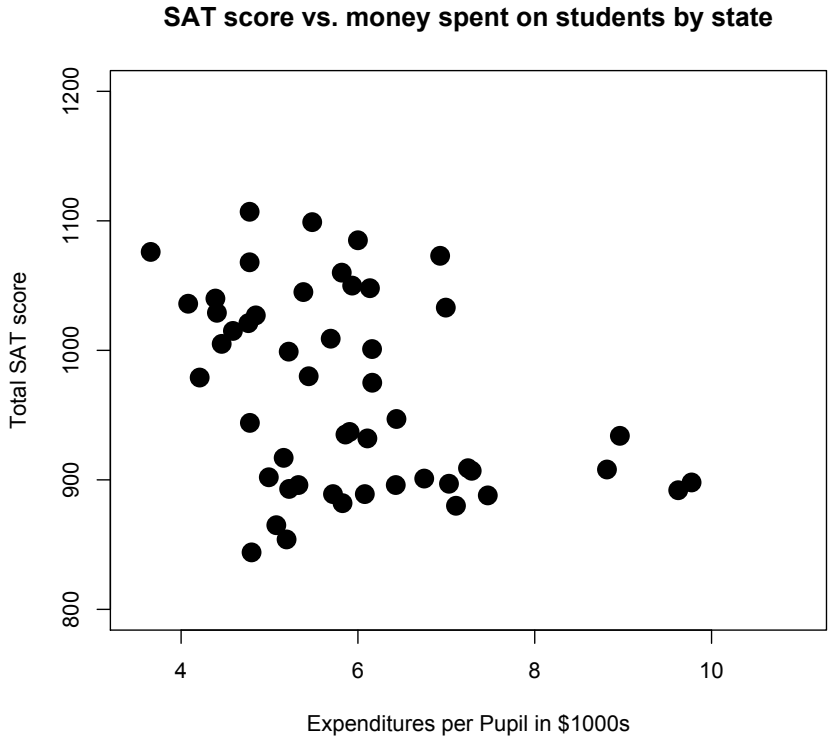
- In setting public policy, we may often hear something like…

> "State spending on education is positively correlated with SAT scores and therefore we should increase our state's spending on education."

Example:

SAT score vs. public $$$ spent on education

- Data was taken from the 1997 Digest of Education Statistics, an annual publication of the U.S. Department of Education.



$r = -0.3805$

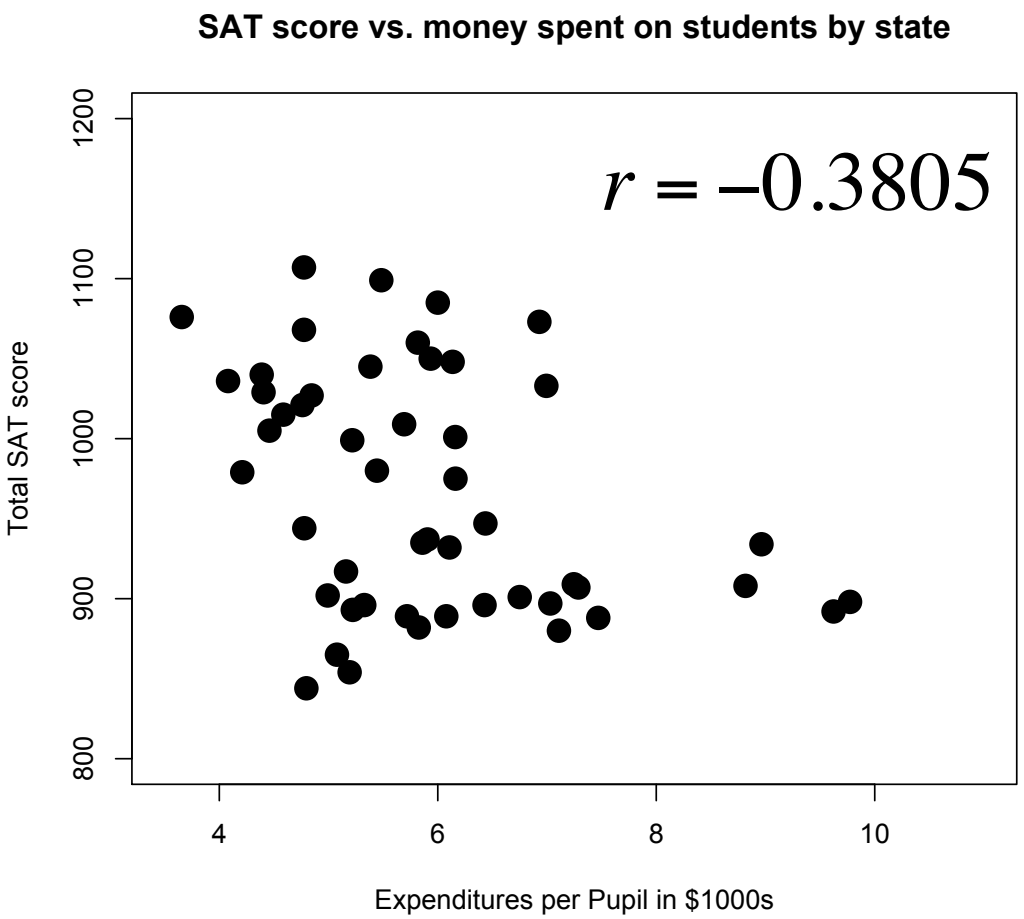
Are you surprised by the relationship in this plot?

What could be going on here?

Example:

SAT score vs. public $$$ spent on education

- First, just looking at this scatterplot, would we interpret it as "Increasing expenditures causes a decrease in SAT scores?"

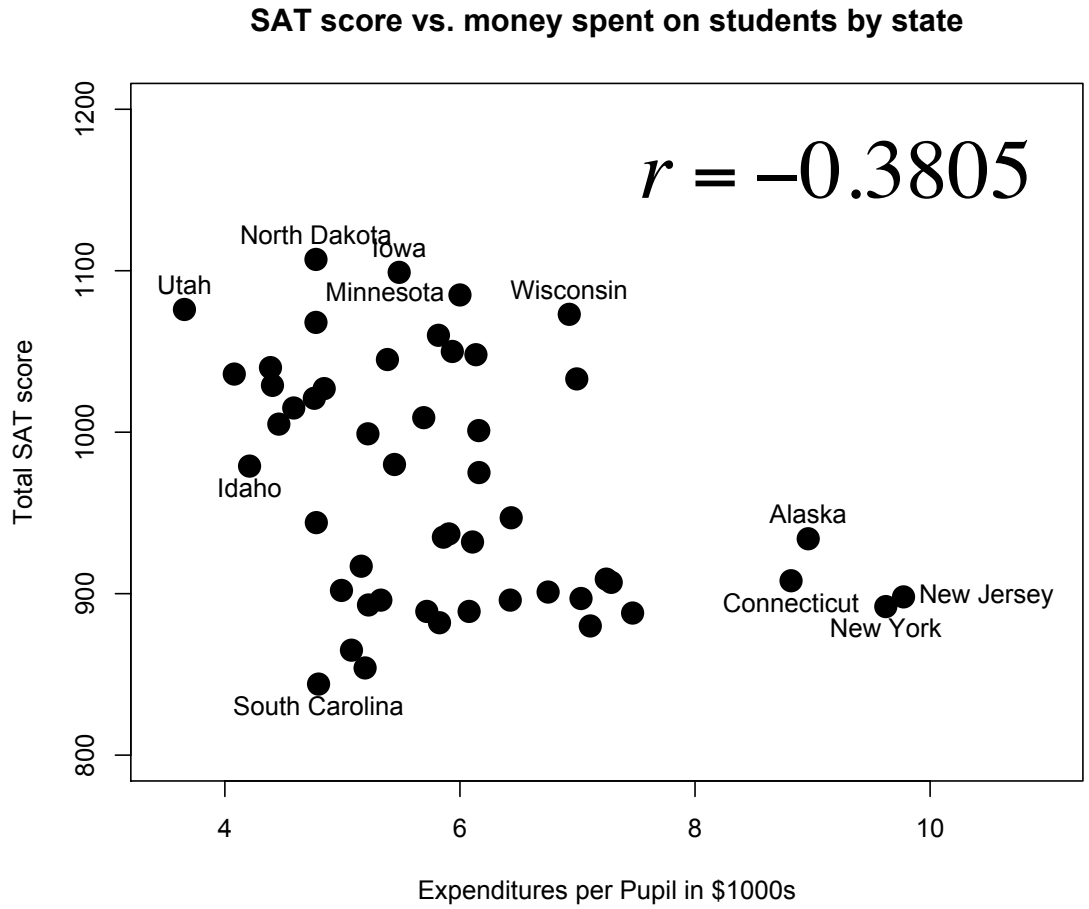


NO.

Example:

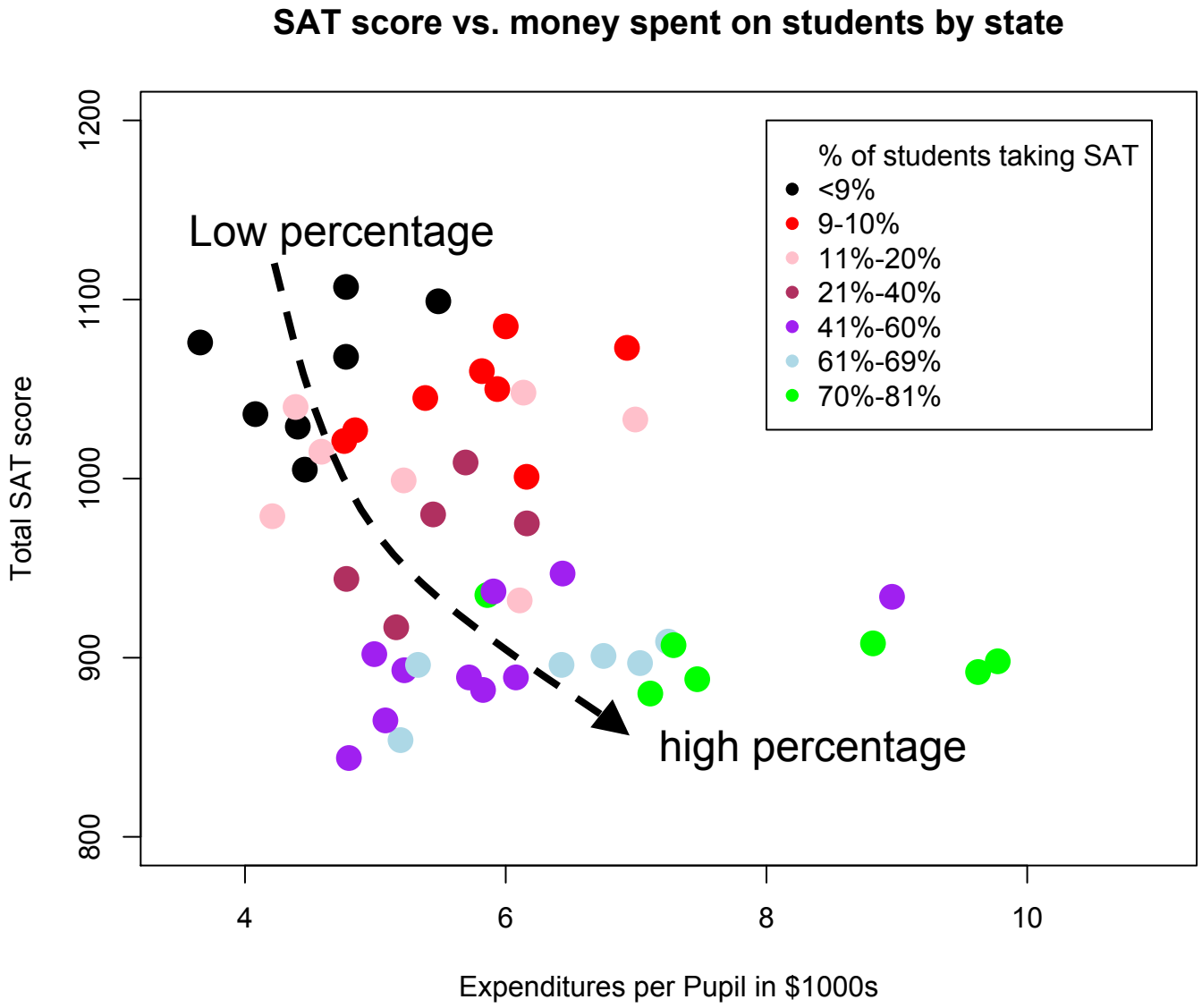
SAT score vs. public $$$ spent on education

- Let's ask… do ALL students in these states take the SAT?



It turns out that the answer is 'no' and this REALLY matters…

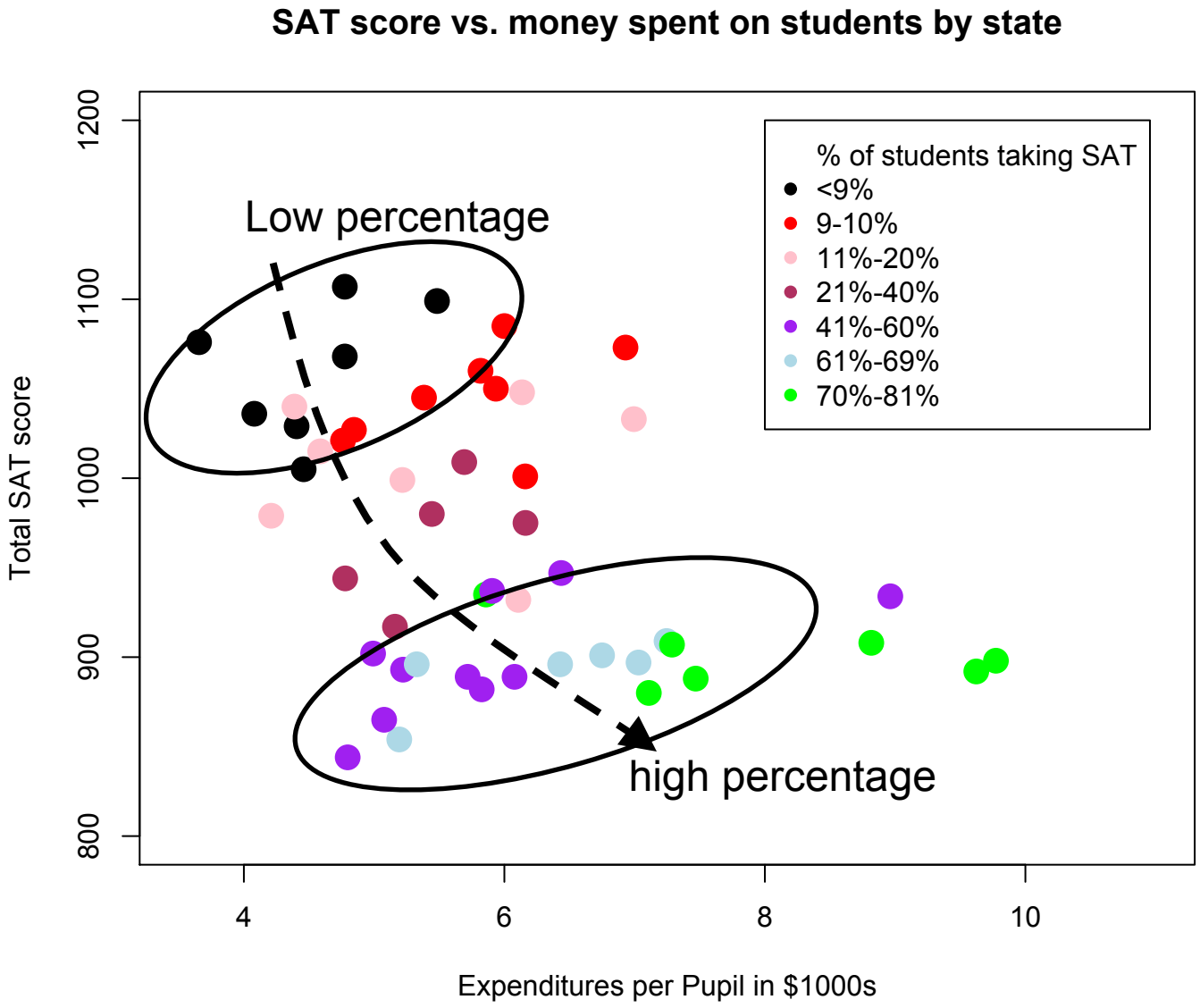
Example: SAT score vs. public $$$ spent on education



Does the percent of students taking the SAT in a state help explain the paradox?

YES!

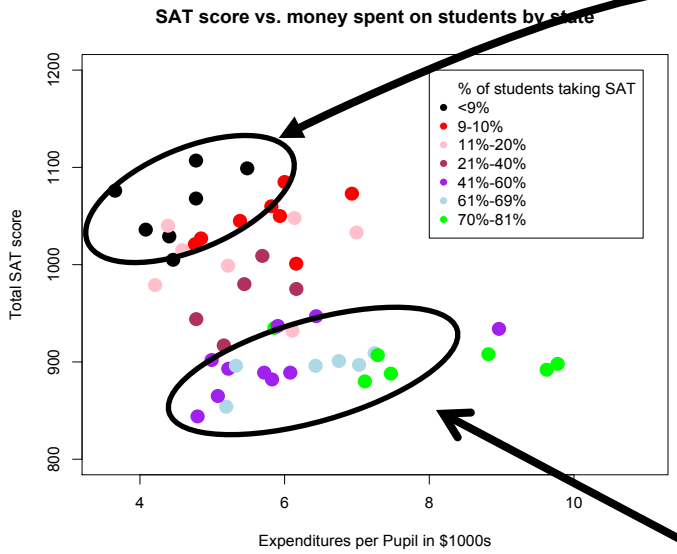
Example: SAT score vs. public $$$ spent on education



Within states with the same % taking the SAT, we actually see a positive relationship!!

Example:

SAT score vs. public $$$ spent on education

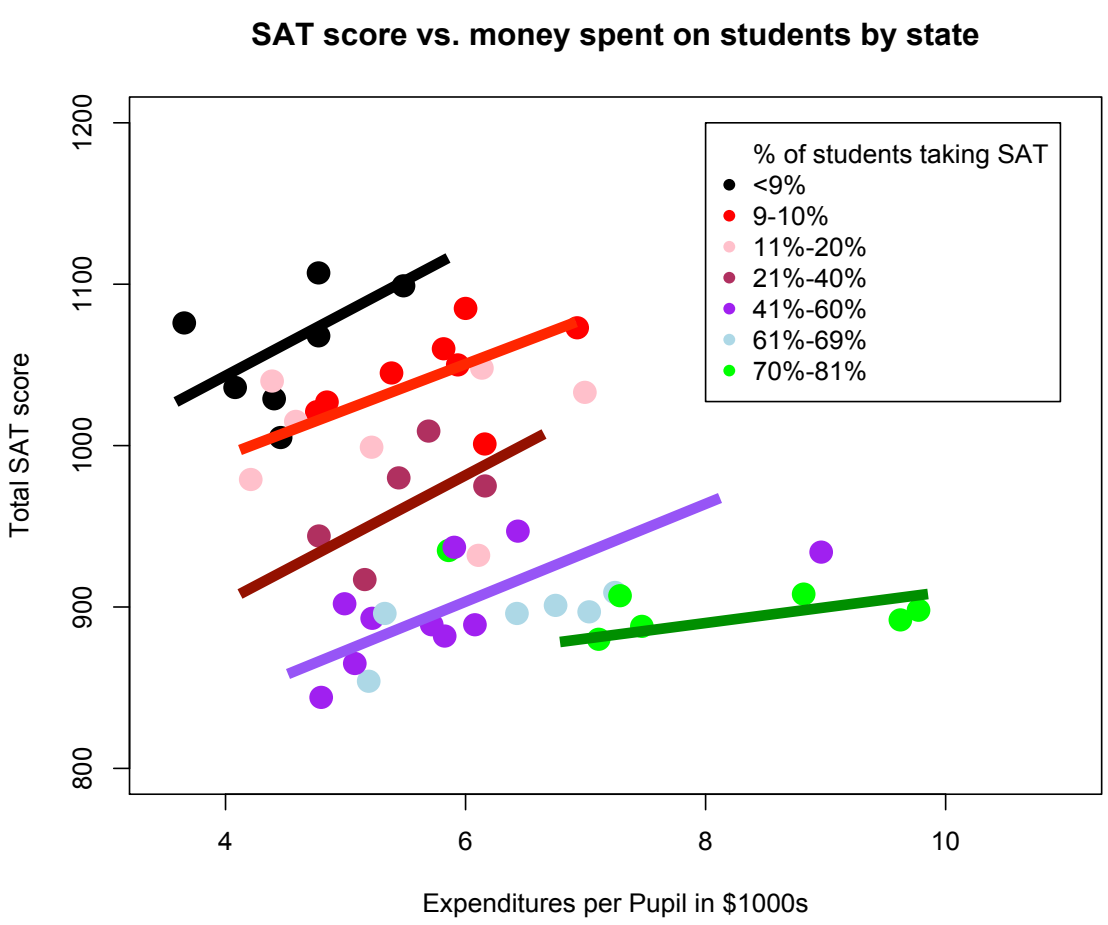


When only a few students take the SAT in a state, who are these students? (best students)

If you have ALL students in a state taking the SAT, then you won't be grabbing just the 'good students'… and the average SAT will be lower compared to states with a small percentage of their 'best' students taking the SAT (is this a fair state-to-state comparison?)

Example:

SAT score vs. public $$$ spent on education



Within similar states (i.e. similar % taking the SAT) we see that spending more $$$ is associated with higher SAT scores.

Simpson's Paradox

- A statistical relationship between two variables can be reversed by including additional factors in the analysis.
  - By including the variable called "% of students taking the SAT", we saw a reversal of the relationship shown in the original scatterplot between SAT score and $$$ spent per student.

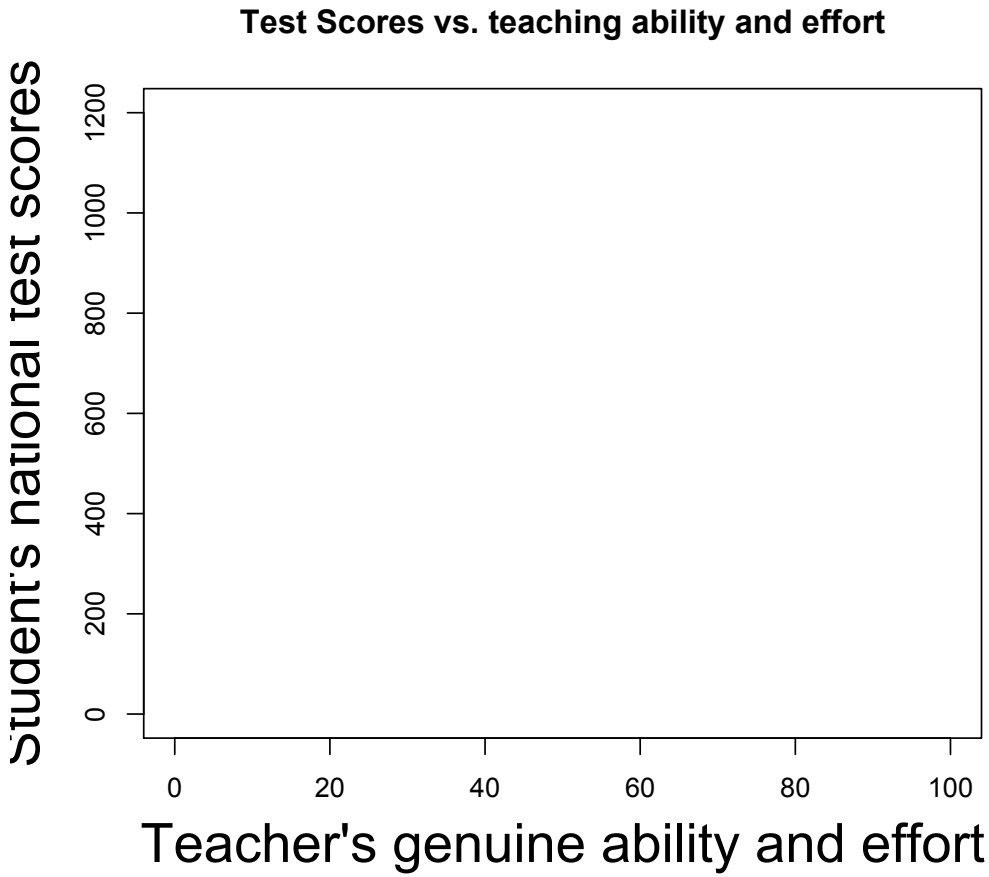
Interpret correlation with caution

- Remember that correlation is a simple summary of a sometimes complex situation.
- Scatterplots are useful, but they do have limitations when many variables impact each other in a complex manner.

The SAT score and expenditure information was a modification of material available at:

www.stat.ucla.edu/labs/pdflabs/sat.pdf

Food for thought…
Should we reward school teachers based on student's standardized test scores?



What might this scatterplot look like?

If students score high, was the teacher good?

If students score low, was the teacher bad?