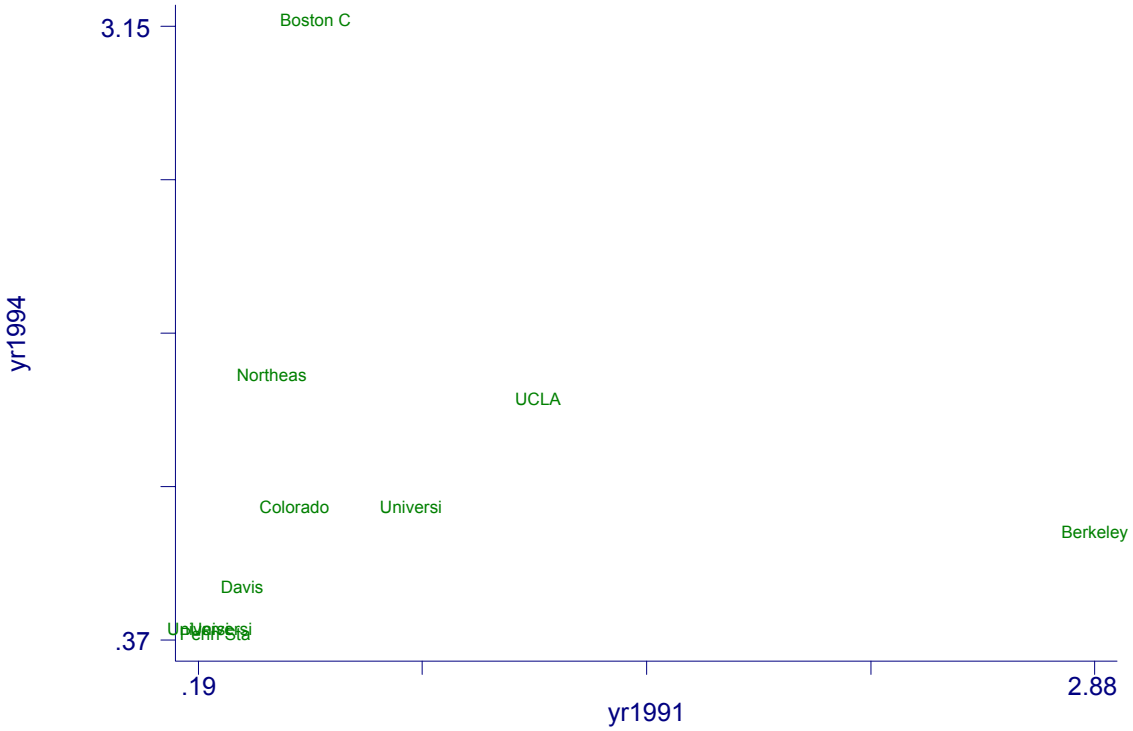
1. Reviewing the Correlation Coefficient

The following data related the number of violent crimes (defined as murder, rape, robbery, and assault) at a sampling of college campuses in 1991 and 1994. The data listed in terms of the number of crimes per 1,000 students.

SCHOOL	1991	1994
Berkeley	2.88	0.83
UCLA	1.21	1.43
Davis	0.32	0.58
Colorado State	0.48	0.94
University of Florida	0.83	0.94
University of Minnesota	0.26	0.39
University of Iowa	0.19	0.39
Penn State	0.24	0.37
Northeastern	0.41	1.54
Boston College	0.54	3.15
Average	0.736	1.056
SD	0.7736	.8013



A. Compute the sample correlation coefficient r for the data.

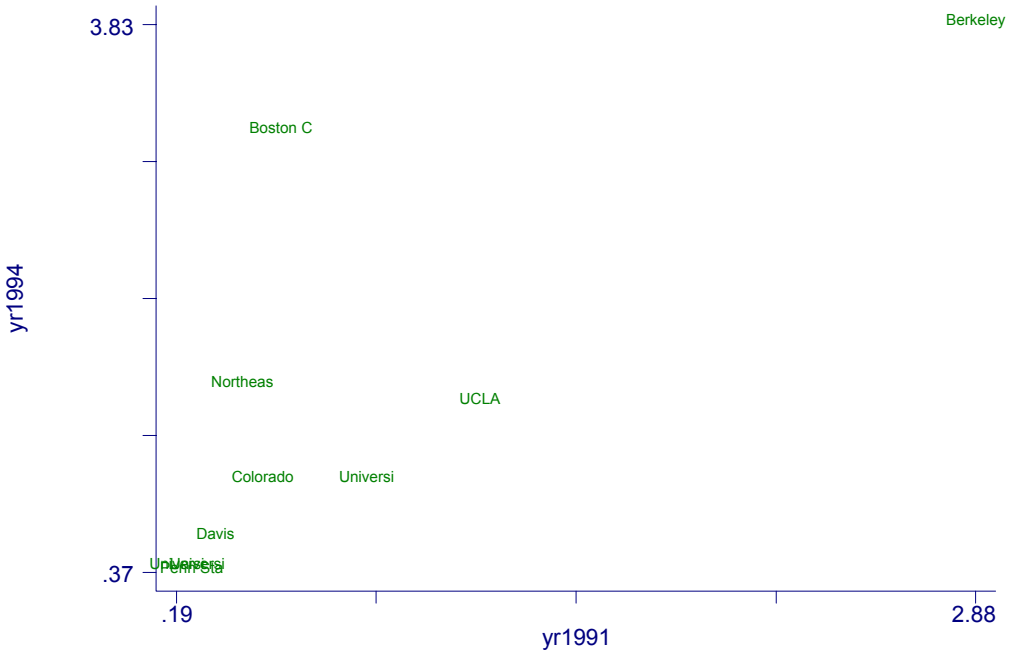
1991 Average is .7360 and the SD is .7736 and 1994 Average is 1.056 and the SD is .8013

The average of the products of each pair in 1991 and 1994 is .81344. The correlation coefficient is:

$$\frac{.81344 - (.7360 \times 1.056)}{.7736 \times .8013} = .0584$$

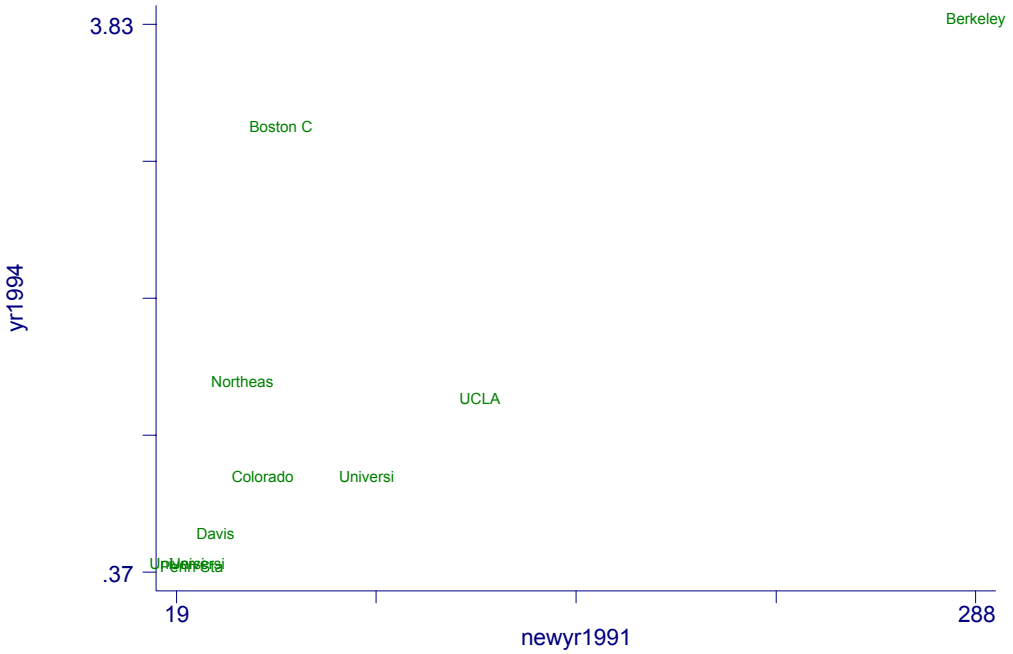
B. Suppose the Berkeley pair is wrong and it should have been 2.88 and 3.83 instead of 2.88 and 0.83. Can a single error change the correlation coefficient?

Absolutely. The new coefficient is .7655...a big change. Note that the correlation coefficient is sensitive to a change in a single number. Here is the graph of the corrected data.



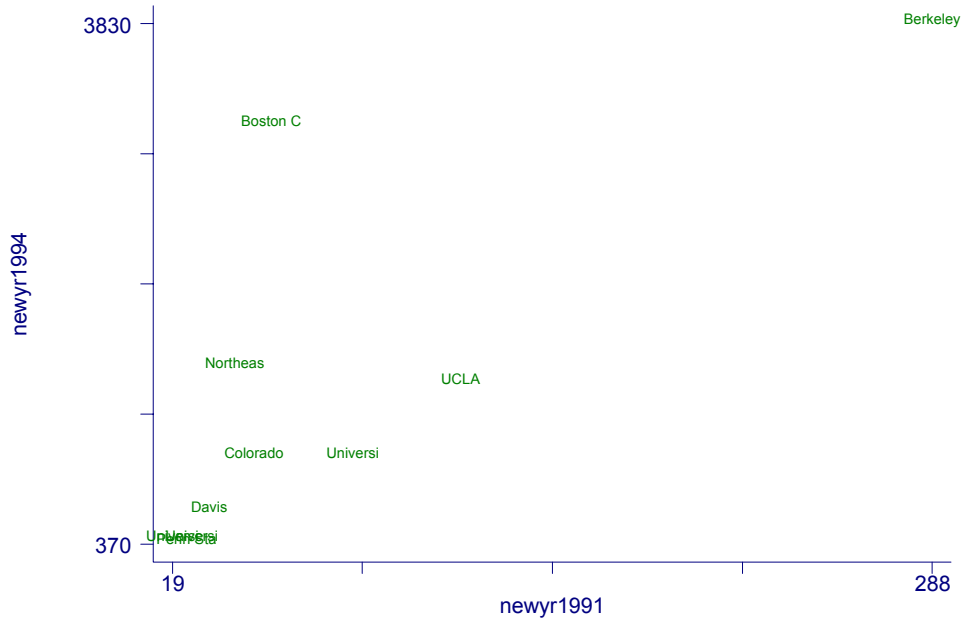
C. Suppose I didn't want to work with decimal points so I multiplied all of the numbers in the first column (1991) by 100. Would the correlation coefficient change?

No ,the correlation will still be .7655 (after the Berkeley change) all multiplication does is shift the data along the number line here's a graph:



D. Multiplying the numbers in the second column by 1000. Would the correlation coefficient change?

No. It would still be .7655 (after the Berkeley change). The average for 1991 and 1994 would change in predictable ways (100 times larger for the first and 1000 times larger for the second) but as long as the relationship between the two years remains unchanged, the correlation is unchanged.

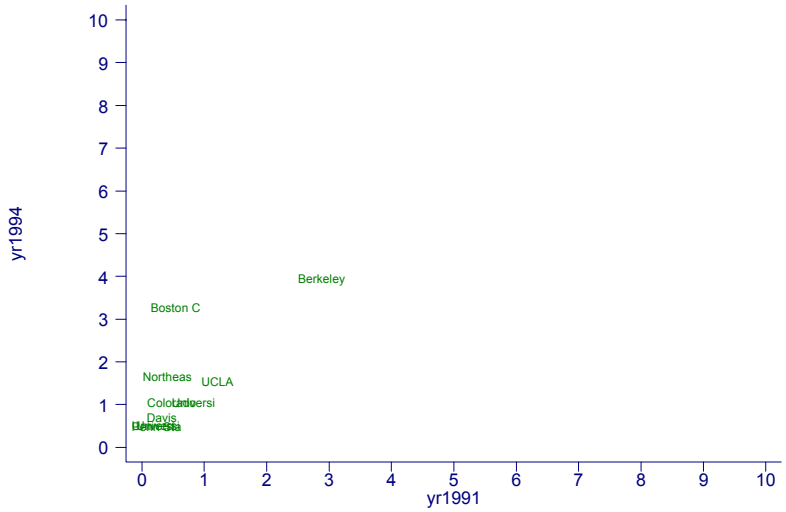


E. Why might knowing this be useful?
Suppose the government was going to change the way it measured crimes, suppose from per thousand as above to per 100,000? So the Berkeley number would go from 2.88 to 288, the UCLA number would go from 1.21 to 12, etc1. Would that change the correlations? No. It will save you time to know the behavior of a correlation coefficient rather than recalculating all of the numbers. And remember, a correlation has no units, it is a unit-less measure, a pure number.

3. Where the correlation can fail you or deceive you

Different SDs
Appearances can be deceptive, the overall appearance of a scatter diagram depends on the Standard Deviations of the individual X and Y variables. Smaller standard deviations make the scatter diagram look "tighter" or more closely placed together. This happens because r is defined by how far individual points deviate from their means divided by their SDs. Clustering is relative. So beware, your eye can be fooled. Be sure to:

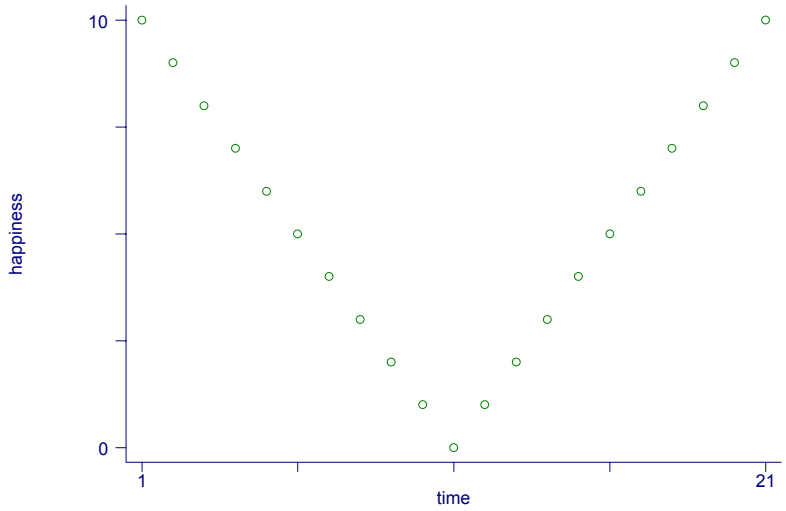
- examine the ranges of the X and Y variables when comparing two sets of data
- look at the standard deviation for the X variable and the Y variable. Are most points within a standard deviation or not?
- calculate the correlation coefficient (if it hasn't been done yet) or review them (if they have been calculated for you)



Compare the above graphic (small SD) with the graphic just above it (large SD). The one with a small SD looks more "tight" and therefore you might think it to be more highly correlated (when it is not). The principle you want to learn is this: look at a graphic AND its numerical summary before coming to any conclusions about a variable or pair of variables.

Outliers --The correlation is useful when your data points are football shaped. But sometimes r will mislead. A single outlier can counterbalance a strong linear relationship. Consider Boston College in the previous graphics. If I were to remove that observation from the dataset, the correlation would jump from .7655 to about .952 so from something football shaped to virtually a straight line.

Non-Linear relationships -- Two variables can be related, but their relationship is not well described by a straight line. Correlations are good for straight-line relationships, they are terrible for curved or other non-linear relationships. The correlation coefficient for a curved relationship will be near zero even though there is clearly a relationship.



Solutions? Always take a look at the scatter diagram if possible.

Ecological Correlations

Correlations for aggregated information (e.g. information by state) are almost always stronger than correlations for individuals. Often people calculate correlations based on rates or averages that are based on many individuals combined. This tends to make a relationship look stronger than it is. Beware when people do this.

Association is NOT causation

Remember: a correlation tells you how strongly two variables are related in a linear fashion. It doesn't mean one variable causes another to happen. The problem here is one of CONFOUNDING, that is, a third variable may be present that has not be taken into account.