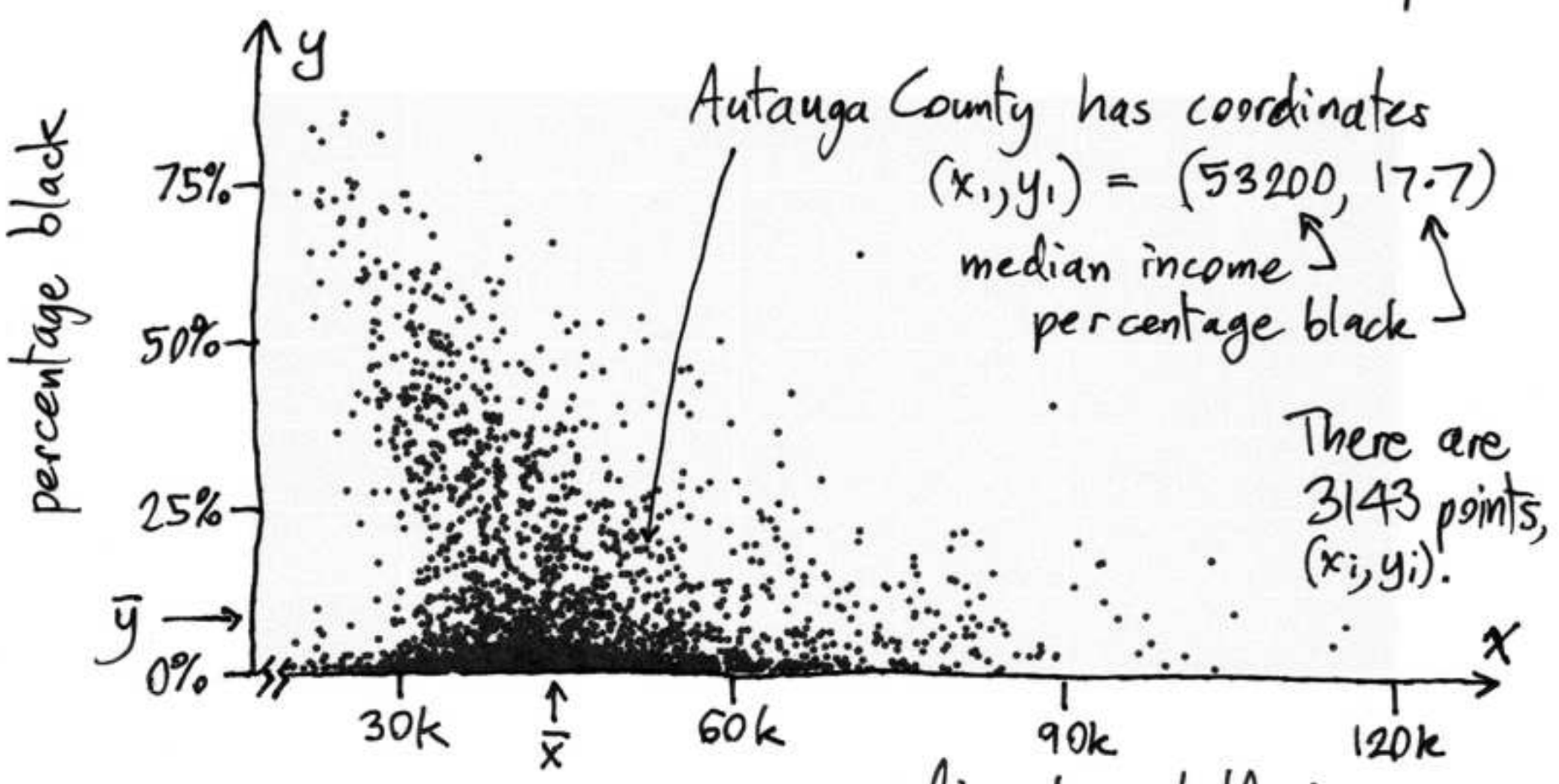
CORRELATION

There are n=3143 counties in the US, and lots of publicly available data about them. (Here we use the "countyComplete" data in the "openintro" package for the free statistical software "R". Most data is from 2010.) Counties are indexed i=1 to 3143. Eg, i=1 is Autauga County, AL. Let's plot on the x-axis median household income, vs the y-axis the percentage of the county population that is black. This is a "scatter plot":



What does this show?

- A correlation between race & poverty: the points lean leftwards as one moves up. (≈4k less per 10% increase)
- Counties with income > 70k are almost all < 20% black. Thus income can be a surrogate for race,
- Poor counties are segregated: for incomes < 30k, the distribution is "bimodal", very white (<3%) or black (>30%).

So, a scatter plot can tell many stories. However, often only Pearson's "correlation coefficient" is given, which mathematically is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

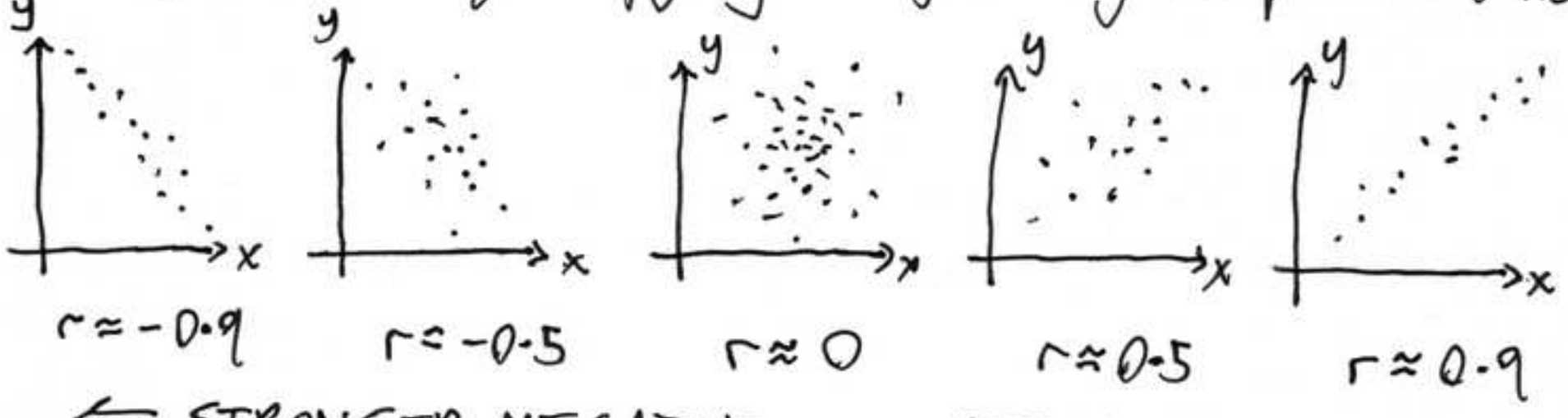
This (perhaps scary) formula involves two familiar quantities:

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the <u>mean</u> of the income over counties

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean of the percentages.

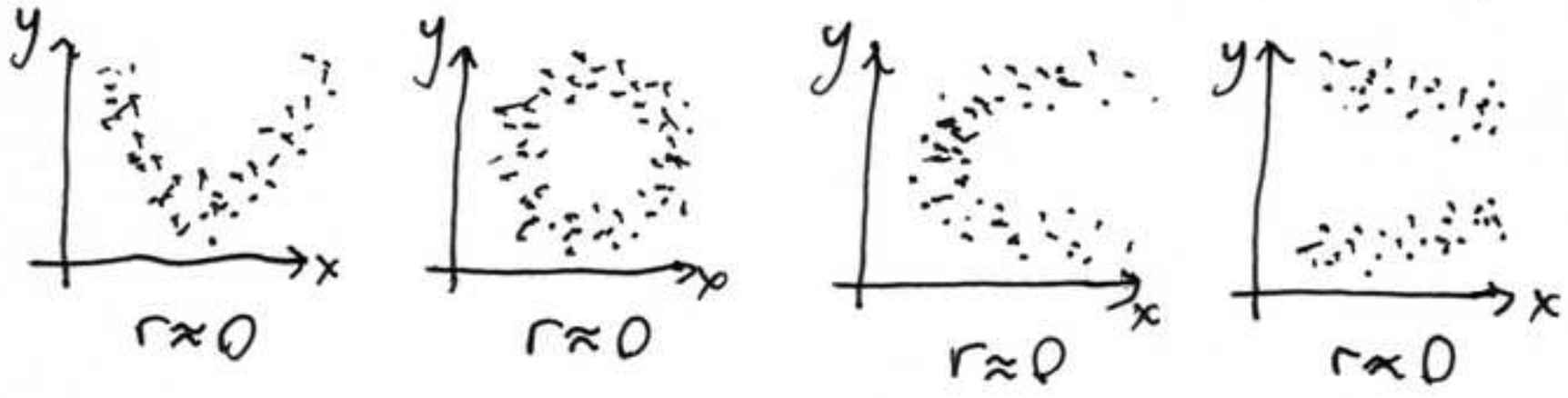
For our data $\bar{x} \approx 44k$, $\bar{y} \approx 9\%$, and these are shown on the plot.

r is good at quantifying the following example correlations:



⇐ STRONGER NEGATIVE STRONGER POSITIVE ⇒

However there are many interesting & informative "nonlinear" correlations that r is oblivious to:



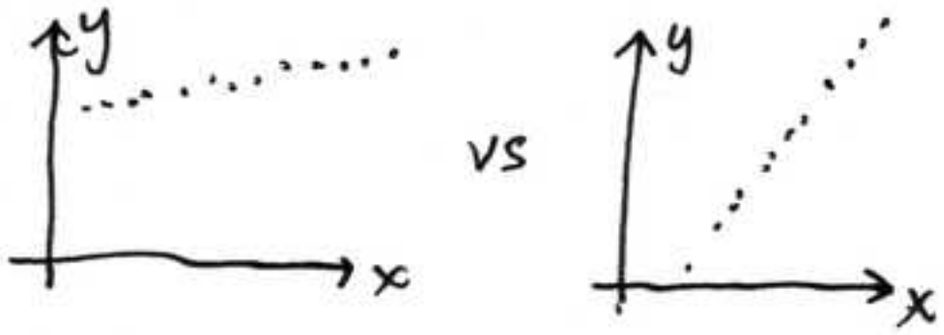
in each case there <u>are</u> correlations, but r cannot tell you this fact! It is insensitive to bimodality (the indicator of segregation earlier).

Returning to our county income and percentage black data, what is r? It turns out to be $r \approx -0.22$, which is negative (as expected from the overall downwards slope), but would be interpreted as very weak. This shows the <u>limitation</u> of the correlation coefficient: it fails to capture the many aspects that a glance at the full scatter plot can show. One must <u>look at</u> the data rather than trust r.

Notes:

- you do not need to handle the formula for r: all statistical software has it built in.
- r lies between -1 and +1, and tells you the strength of the linear correlation, not to be confused with the strength of the <u>effect</u> (which r does <u>not</u> tell you).

Eg.



Both have $r \approx 1$, but in the 2nd case y changes much faster with x.

- scatter plots can be 3D too with (x_i, y_i, z_i) data, or even higher dimension, but it is hard to picture!
- a better analysis of county data might "weight" each point by the county population.
- nonlinear correlations (bimodality, etc) <u>can</u> be found by using, eg, powers of variables, x^2, x^3, etc.