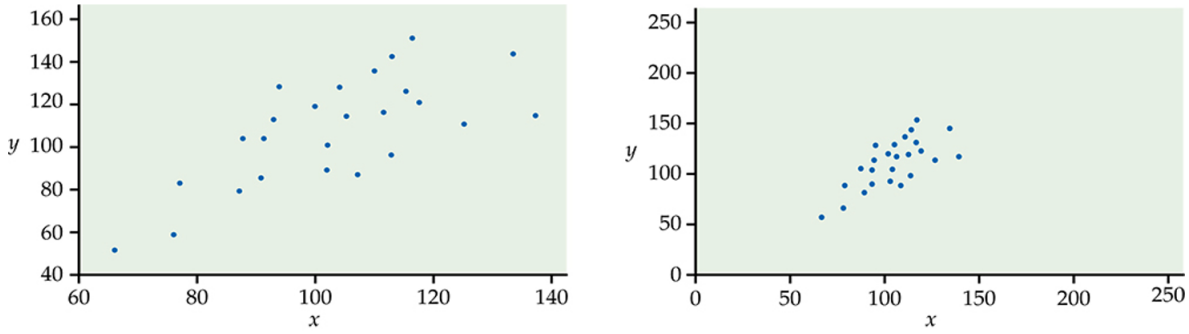
Topic 6: Correlation

Using Correlation to Measure Strength of Association

Which scatter plot below displays a stronger linear association?



Correlation, which we will denote with r, gives us a more mathematical way to measure the strength of the linear association between two numerical variables:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Practice: Find correlation between X and Y given the observations below:

X	Y
3	4
5	5
7	3

Properties of Correlation

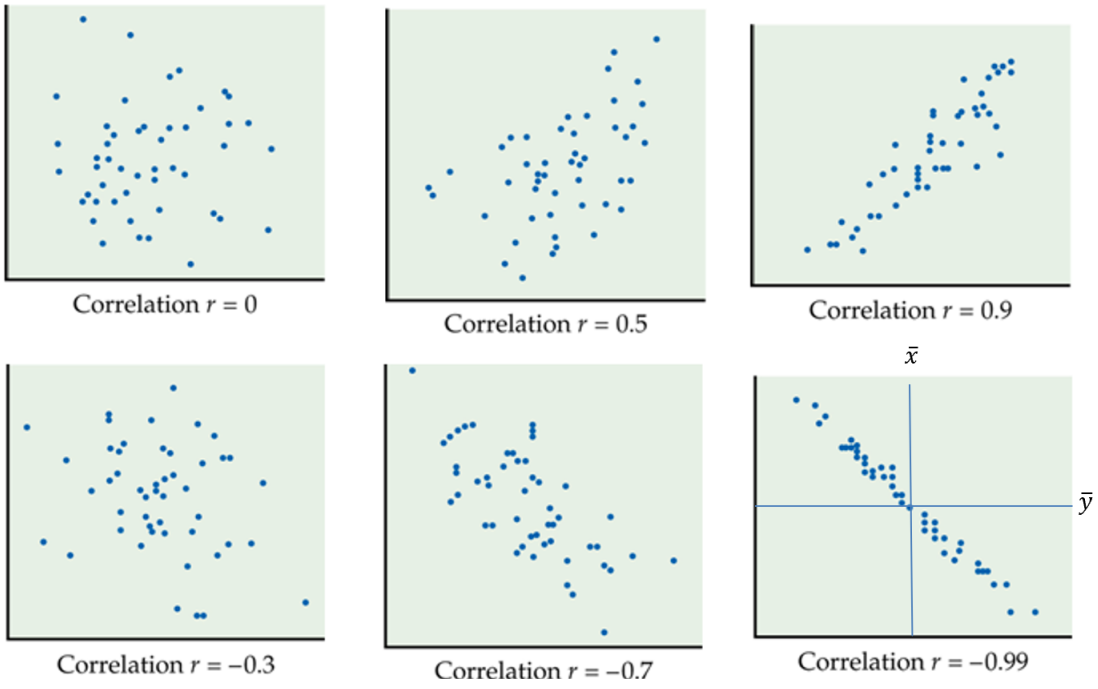
- Possible values

- Sign meaning

- r does not change even if we...

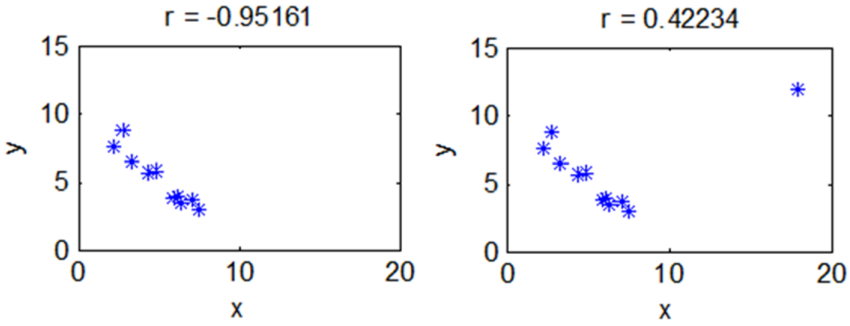
- Note on outliers

Examples:



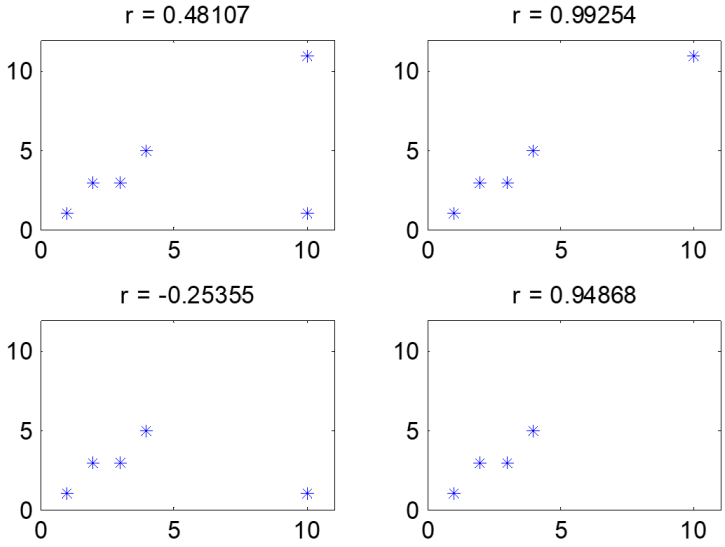
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

How an outlier can distort an association:

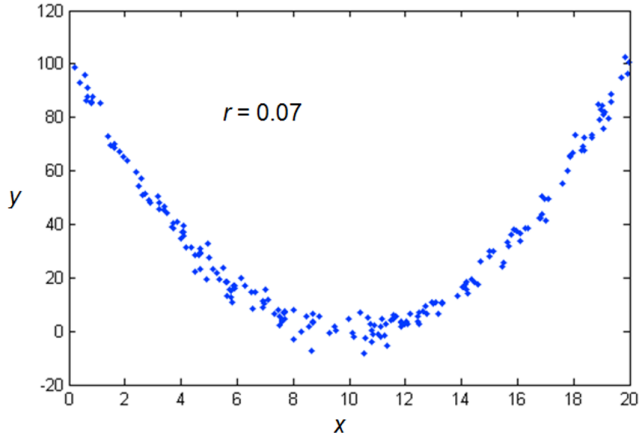


Another example:

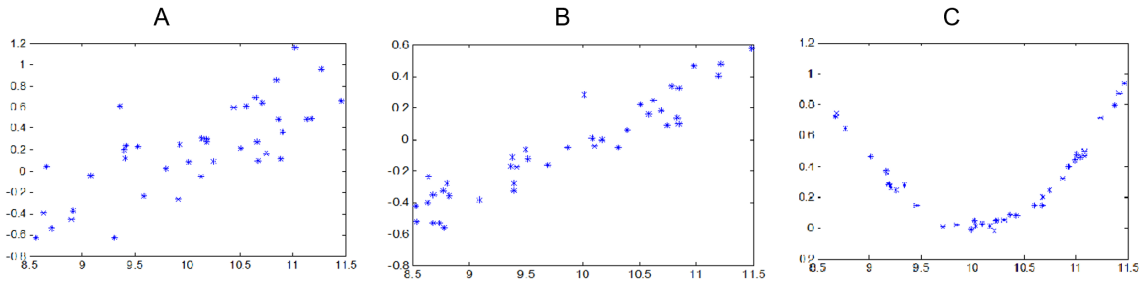
x	y
1	1
2	3
3	3
4	5
10	1
10	11



Correlation does not describe nonlinear associations!

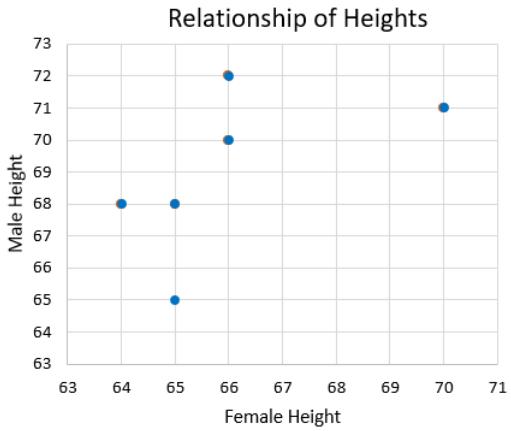


Practice: Which scatterplot has the largest correlation (closest to +1)?



Exercise: Heights of couples

This scatterplot (derived from the Excel sheet `heights.xlsx`) displays the paired heights of 6 heterosexual couples. On this plot, the value of r is:



1. If all the men were 6 inches shorter, would correlation change? Does the correlation tell us about whether women tend to date men taller than themselves?

2. If heights were in centimeters, would correlation change?

3. If each woman dated a man exactly 3 in. taller than herself, what would be the correlation?