Lecture 05
Scatterplots & Correlation

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Review: Key Points

Chapter 5: Comparing distributions

- Outliers are context dependent
- Timeplots
- Moving-averages, smoothing
- Re-expressing data (log, sqrt)
  - to improve symmetry
  - to equalize spread

Chapter 6: Normal model

- z-score, shifting and rescaling
- Normal model,
- Normality assumption; unimodal, symmetric
- 68-95-99.7 Rule
- z-to-p, p-to-z
- Normal probability plots

Scatterplots

- Dots to show the relationship between two variables

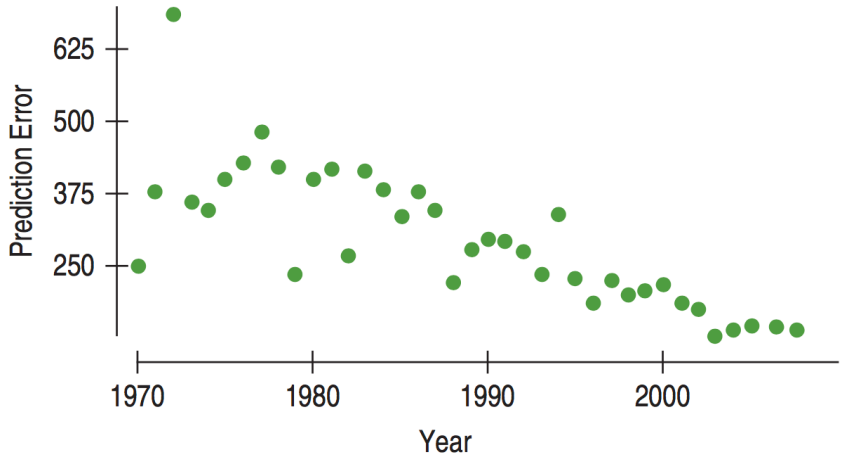


FIGURE 7.1
A scatterplot of the average error in nautical miles of the predicted position of Atlantic hurricanes for predictions made by the National Hurricane Center of NOAA, plotted against the *Year* in which the predictions were made.

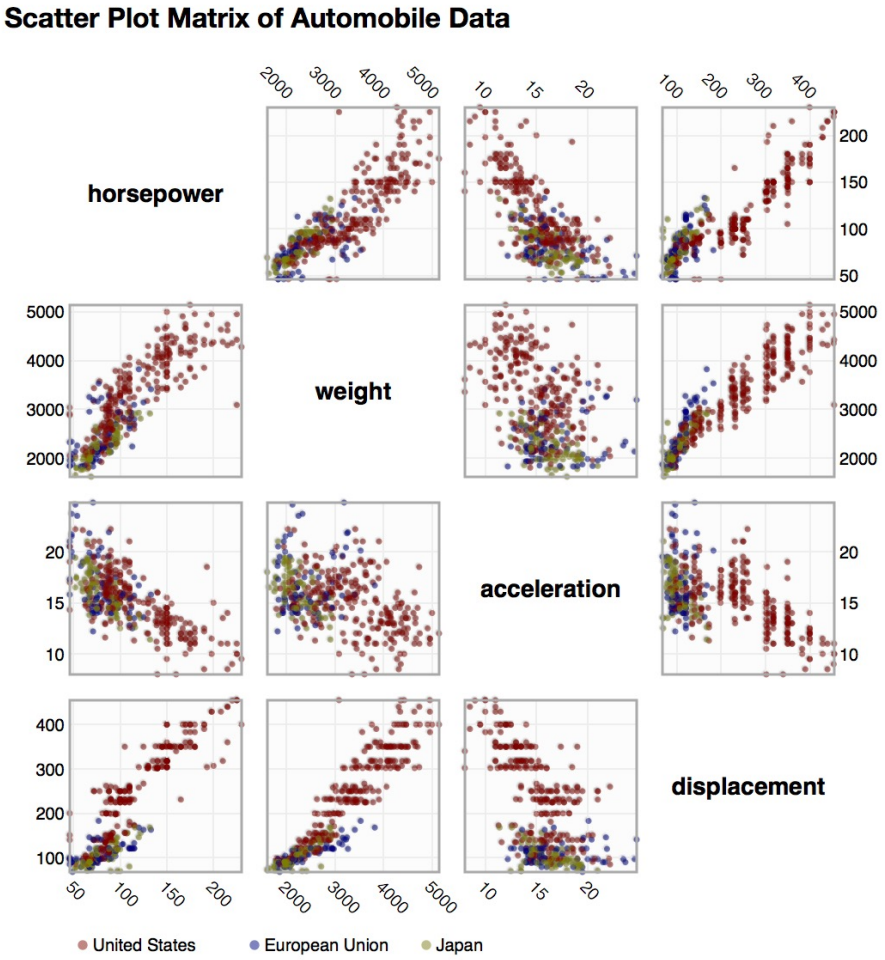
- Most common display for data
- "You can observe a lot just by watching", Yogi Berra

Scatterplots

- Displaying multiple attributes
- https://homes.cs.washington.edu/~jheer/files/zoo/ex/stats/splom.html

✓ Cool webpage for data visualization:

✓ https://homes.cs.washington.edu/~jheer/files/zoo/



Four dimensions of a database of cars plotted in a scatter plot matrix, with different colors to indicate the country of origin. Each pair of variables is represented in two (transposed) plots. Dragging a rectangle on any of the graphs highlights the selected points in all the graphs, a technique called *brushing and linking*.

Source: GGobi

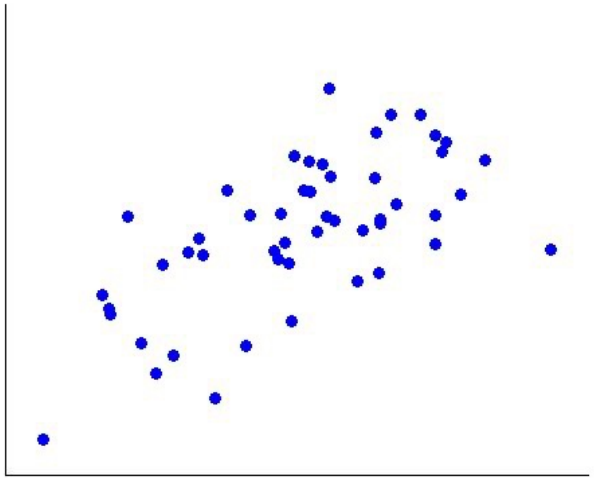
Scatterplots

- When examining the scatter plot, be sure to study:
  - Direction: "What's my sign—positive, negative, or neither?"
  - Form: "Straight, curved, something exotic, or no pattern at all?"
  - Strength: "How much scatter?"
  - Outliers: "Are there outliers or subgroups?"

Scatterplots: Direction

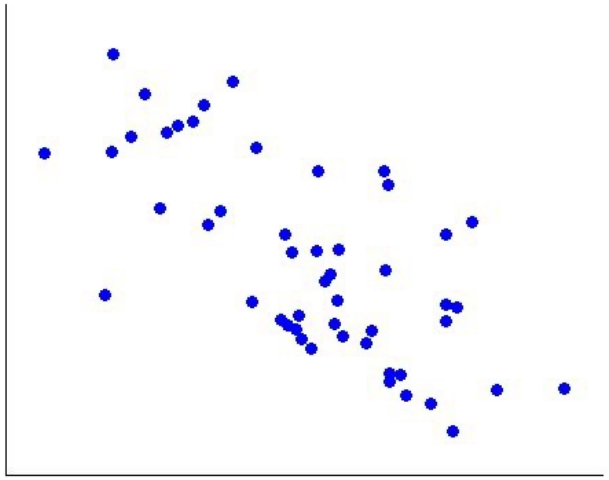
(a) Positive association

As x increases, y increases.



(b) Negative association

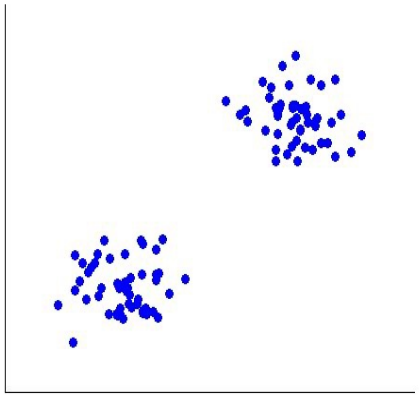
As x increases, y decreases.



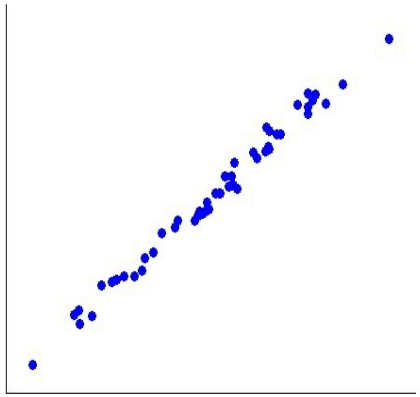
Slide credit: Martin Lindquist

Scatterplots: Form

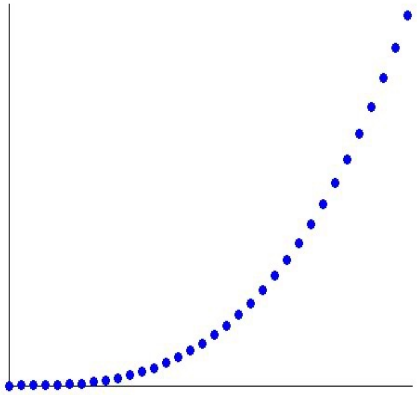
- What does the data look like? Is the data clustered together, does the data have a linear shape or a curved shape.



Clustered



Linear

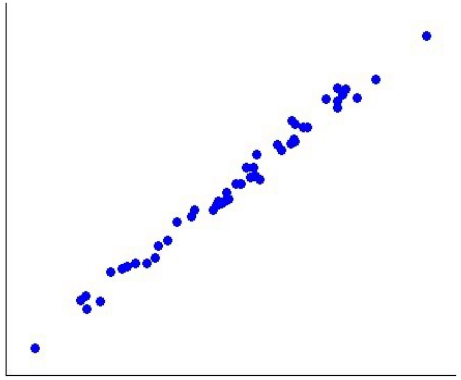


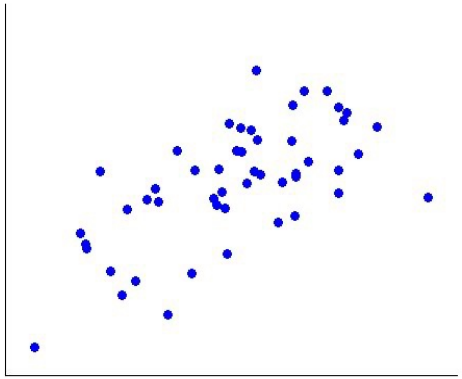
Curved

Slide credit: Martin Lindquist

Scatterplots: Strength

- How closely do the points follow a clear form?

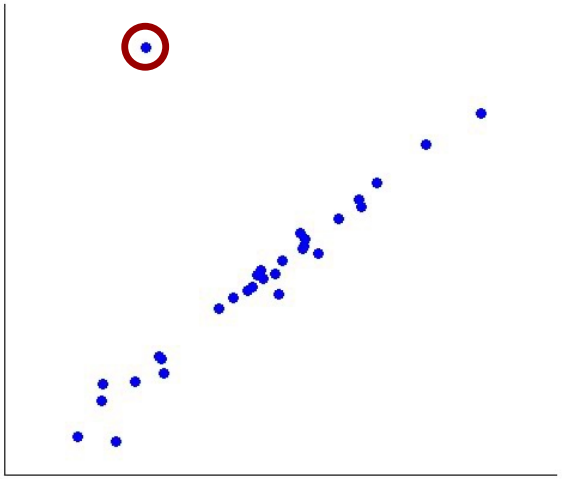




Slide credit: Martin Lindquist

Scatterplots: Outliers

- Outliers can be determined from a graph by looking for points that are not within the overall pattern of the data.



Slide credit: Martin Lindquist

Variables in different questions

- Do baseball teams that score more runs sell more tickets to their games?
  - ✓ Each data point: team, x: score, y: tickets they sell
- Do older houses sell for less than newer ones of comparable size and quality?
  - ✓ Each data point: houses with similar size and quality, x: age, y: the price of the house
- Do students who score higher on their SAT tests have higher grade point averages in college?
  - ✓ Each data point: students, x: SAT scores, y: GPA in college
- Can we estimate a person's percent body fat more simply by just measuring waist or wrist size?
  - ✓ Each data point: person, x: percent body fat, y: waist (or wrist size)

Roles for variables

- x: explanatory variable, predictor variable, independent variable
- y: the variable of interest, response variable, dependent variable

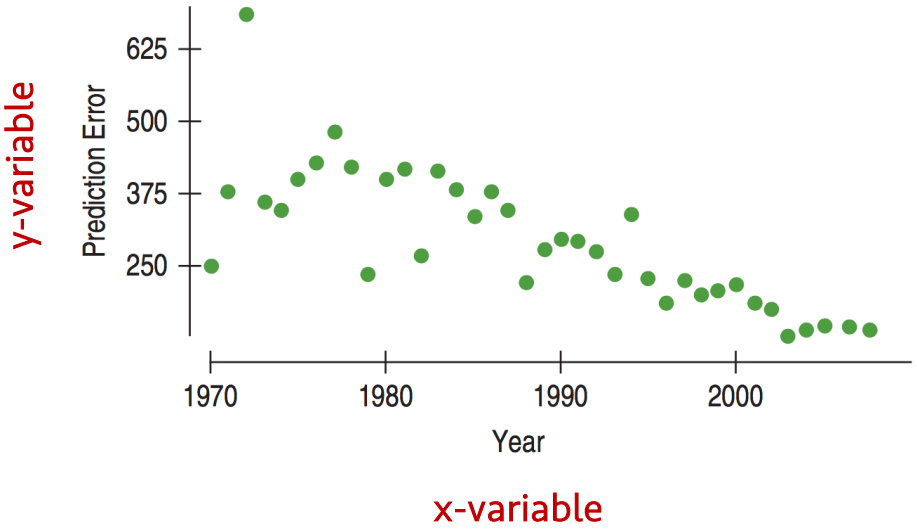


FIGURE 7.1
A scatterplot of the average error in nautical miles of the predicted position of Atlantic hurricanes for predictions made by the National Hurricane Center of NOAA, plotted against the *Year* in which the predictions were made.

Linear relationships

- If the form of the plot looks like a line, this indicates there may be a linear relationship between the two variables.
- The relationship is strong if all the data points approximately make up a straight line.
- It is weak if the points are widely scattered about the line.

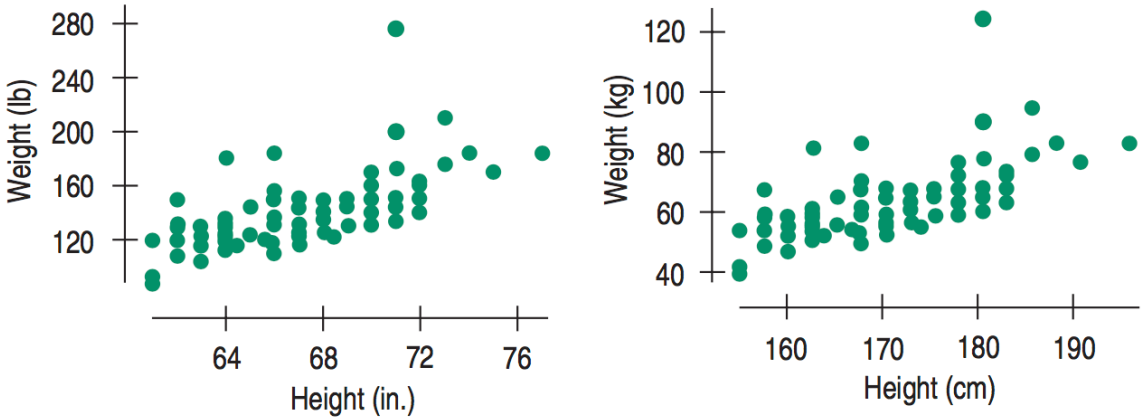
Slide credit: Martin Lindquist

Correlation

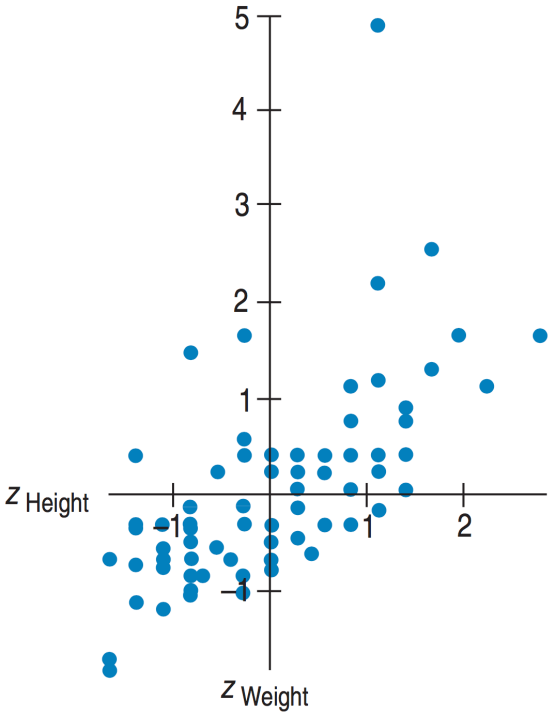
- We want a numerical summary that can be used to measure the strength of a linear relationship.
- The correlation is a measure of strength and direction of a linear relationship between two quantitative variables.
- Correlations are usually denoted by *r*.

Slide credit: Martin Lindquist

Correlation



- The units should have no effects on our measure of strength of a linear relationship.
- *z*-scores can be used: $(z_x, z_y) = \left(\frac{x - \bar{x}}{s_x}, \frac{y - \bar{y}}{s_y}\right)$



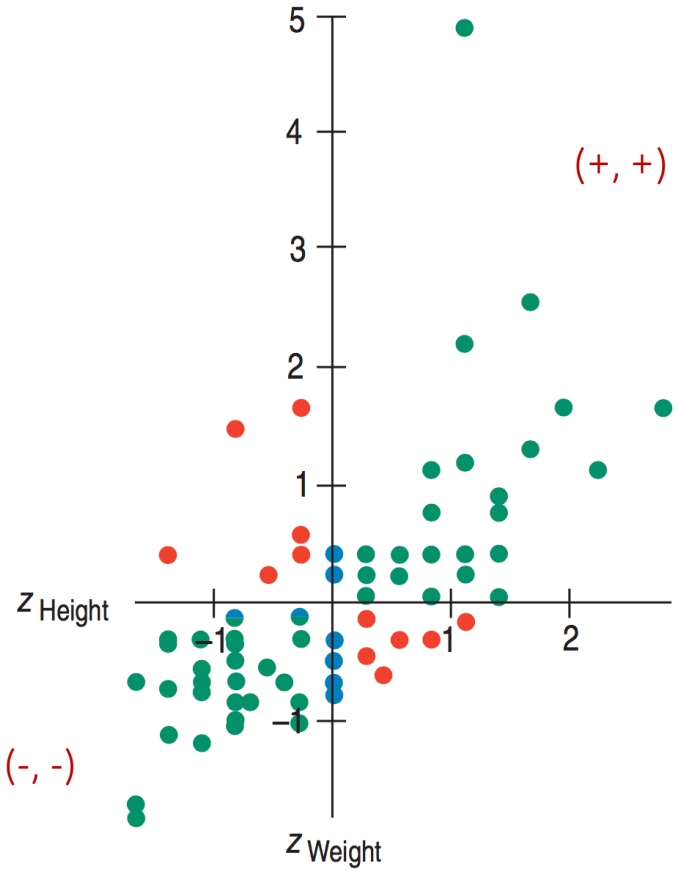
Correlation

- Strength of linear relationship

  should be proportional to $\sum z_x z_y$

- Correlation coefficient: $r = \frac{\sum z_x z_y}{n - 1}$

  , where $(z_x, z_y) = \left(\frac{x - \bar{x}}{s_x}, \frac{y - \bar{y}}{s_y} \right)$



Correlation

- Strength of linear relationship

  should be proportional to $\sum z_x z_y$

- Correlation coefficient: $r = \frac{\sum z_x z_y}{n-1}$

  , where $(z_x, z_y) = \left(\frac{x - \bar{x}}{s_x}, \frac{y - \bar{y}}{s_y} \right)$

- Different expressions, but mathematically same:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1) s_x s_y}$$

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

Assumptions and Conditions for Correlation

- Quantitative variables condition
- Straight enough condition
- No outliers condition
  - ✓ Outliers can distort the correlation dramatically.

Properties of Correlation

- **Sign of correlation:** the direction of the association (e.g., positive, negative)
- **Range:** *r* is always between -1 and 1.
  - When *r* = 1 all of the points lie on a straight line with a positive slope.
  - *r* < 0 indicates a negative association.
  - When *r* = -1 all points lie on a straight line with negative slope.
  - If *r* is close to 0, this indicates a very weak linear relationship.
- **Symmetry:** The correlation of x with y is the same as the correlation of y with x.
- **No units**
  - The value of *r* does not change even if units of measure are changed.
  - The correlation has no unit of measurement.
- **Only linear:** Correlation measures only the strength of a *linear* relationship.
- **Sensitive to outliers:** The correlation is sensitive to outliers.

Other correlation measures (non-parametric):

- Kendall's Tau (τ)
  - can be used for Likert-type scale data (*ordinal* variable)
  - Likert-type scale: e.g., 0 = not at all, 1 = a little, 2 = moderately, 3 = very much
  - commonly used in questionnaires or survey

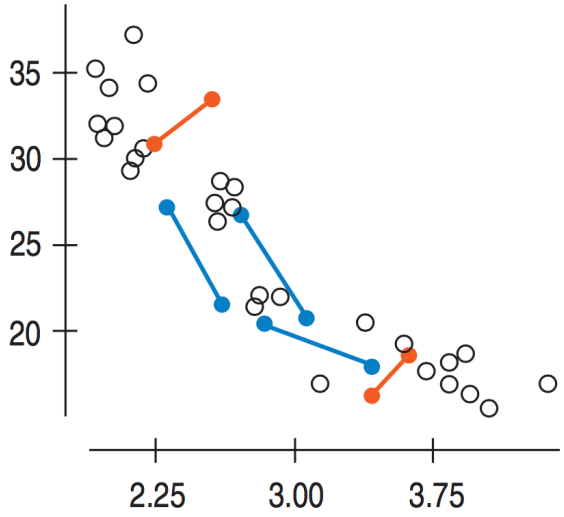
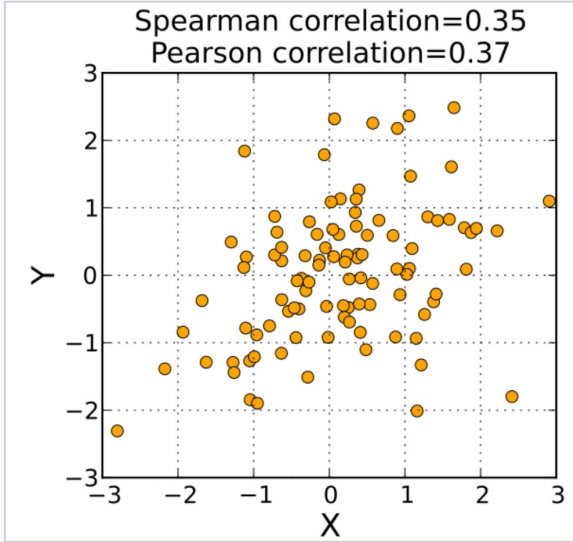


FIGURE 7.6
For each pair of points, Kendall's tau records whether the slope between them is positive (red), negative (blue), or zero.

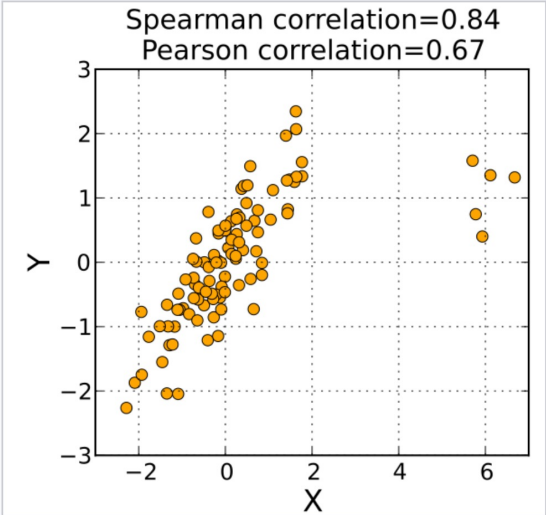
$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

Other correlation measures (non-parametric):

- Spearman's Rho (ρ)
  - replaces the original data with their ranks within each variable
  - then, calculate the correlation



When the data are roughly elliptically distributed and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.



The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's rho limits the outlier to the value of its rank.

https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Correlation ≠ Causation

- We, as human, tend to see causes and effects in everything.
- Storks → more babies!

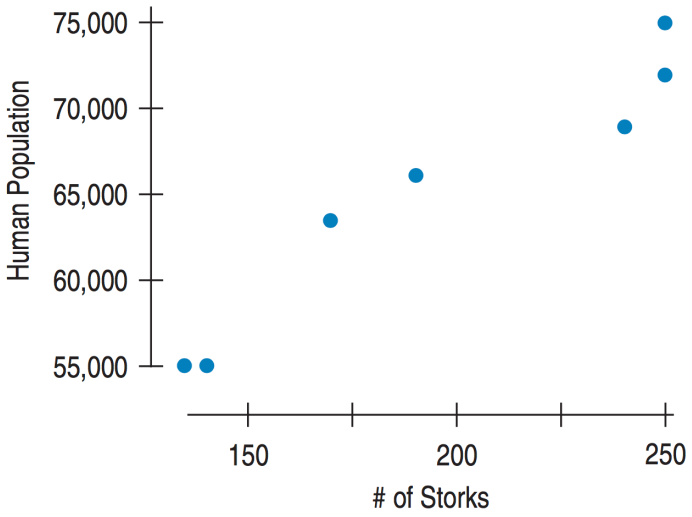


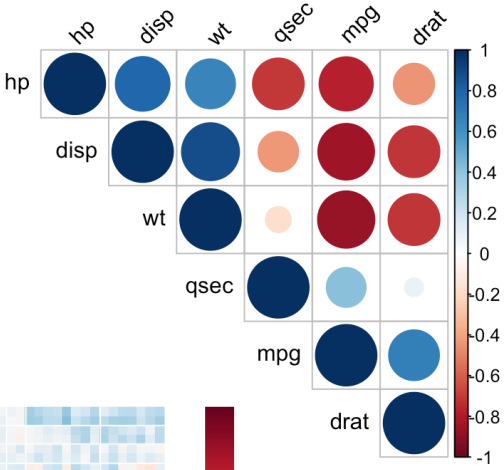
FIGURE 7.5
The number of storks in Oldenburg, Germany, plotted against the population of the town for 7 years in the 1930s. The association is clear. How about the causation? (*Ornithologishe Monatsberichte*, 44, no. 2)

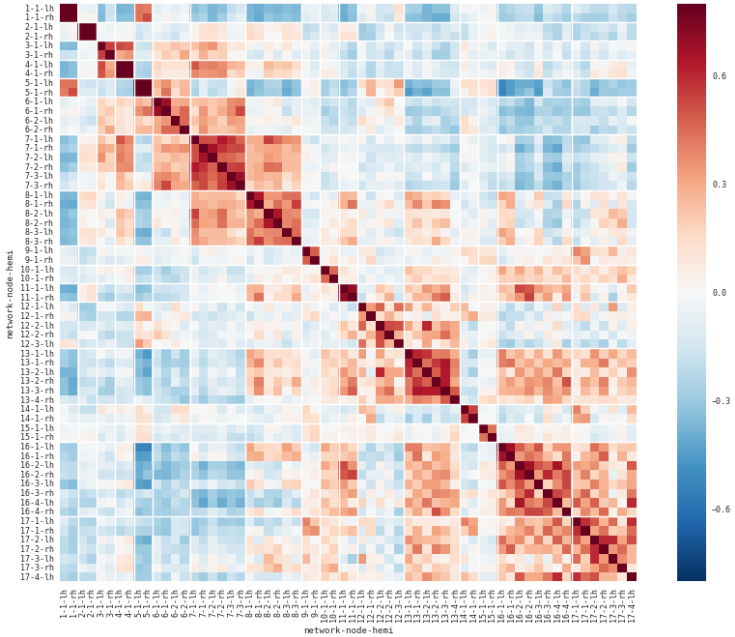
- In fact, more people → more houses → more nesting sites → more storks (the *opposite* direction)
- Lurking variable: a *hidden* variable that simultaneously influences the other two variables
- Scatterplots and correlation *never* prove causation.

Correlation table

	Assets	Sales	Market Value	Profits	Cash Flow	Employees
Assets	1.000					
Sales	0.746	1.000				
Market Value	0.682	0.879	1.000			
Profits	0.602	0.814	0.968	1.000		
Cash Flow	0.641	0.855	0.970	0.989	1.000	
Employees	0.594	0.924	0.818	0.762	0.787	1.000

TABLE 7.1
A correlation table of data reported by *Forbes* magazine for large companies. From this table, can you be sure that the variables are linearly associated and free from outliers?





Assumptions and Conditions for Correlation (again)

- Quantitative variables condition
- Straight enough condition
- No outliers condition
  - ✓ Outliers can distort the correlation dramatically.

- **What if the relationship is not straight enough (i.e., non-linear)**
  - ✓ If the scatterplot shows a bent shape that consistently increases or decreases, we can straighten the relationship by re-expressing, or transforming, one or both variables.

Key Points

Chapter 7: Scatterplots, Correlation

- Scatterplots (direction, form, strength, outliers)
- x- and y-variables: explanatory/independent vs. response/dependent variables
- Correlation: strength and direction
- Assumptions and conditions:
  - ✓ Quantitative variables condition
  - ✓ Straight enough condition
  - ✓ No outliers condition
- Non-parametric correlations: Kendall's tau, Spearman's rho
- Correlation ≠ Causation
- Correlation table/matrix