

Chapter 12

Fundamentals of Data Visualization

May 8, 2023

Associations among two or more quantitative variables

- height, weight, length, daily energy demands
- pH, alkalinity, nitrate/nitrite

Associations among two or more quantitative variables

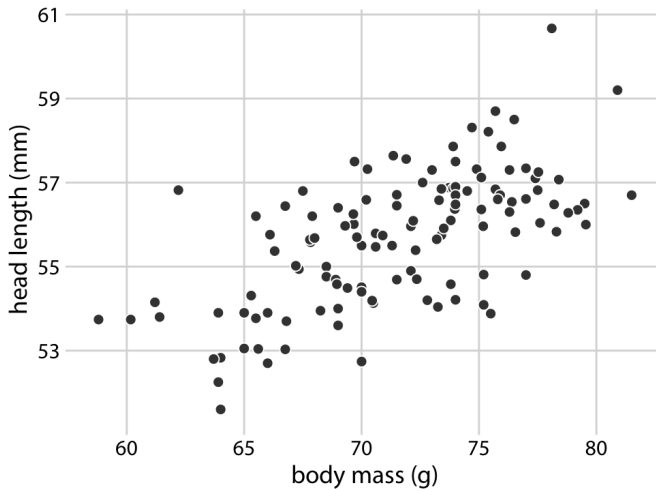
- scatter plot
- bubble plot
- scatter plot matrix
- correlogram
- dimensionality reduction
 - principal components

Bluejays dataset

```
> library(Stat2Data)

> glimpse(BlueJays)
Rows: 123
Columns: 9
$ BirdID      <fct> 0000-00000, 1142-05901, 114...
$ KnownSex    <fct> M, M, M, F, M, F, M, M, F, ...
$ BillDepth   <dbl> 8.26, 8.54, 8.39, 7.78, 8.7...
$ BillWidth   <dbl> 9.21, 8.76, 8.78, 9.30, 9.8...
$ BillLength  <dbl> 25.92, 24.99, 26.07, 23.48,...
$ Head        <dbl> 56.58, 56.36, 57.32, 53.77,...
$ Mass        <dbl> 73.30, 75.10, 70.25, 65.50,...
$ Skull       <dbl> 30.66, 31.38, 31.25, 30.29,...
$ Sex         <int> 1, 1, 1, 0, 1, 0, 1, 1, 0, ...
>
```

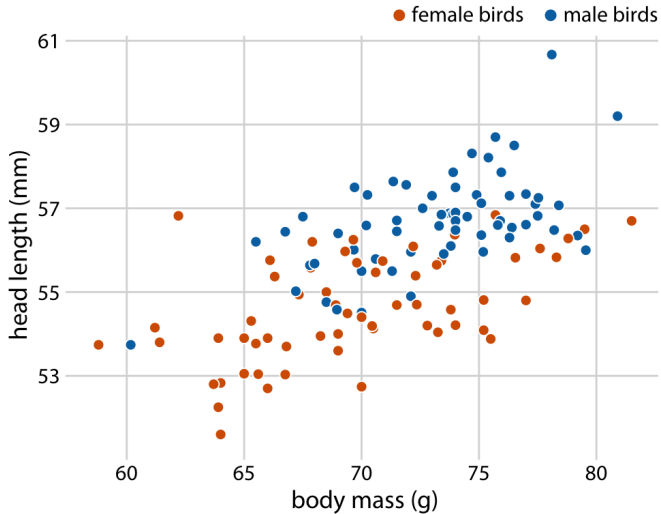
Bluejays scatterplot



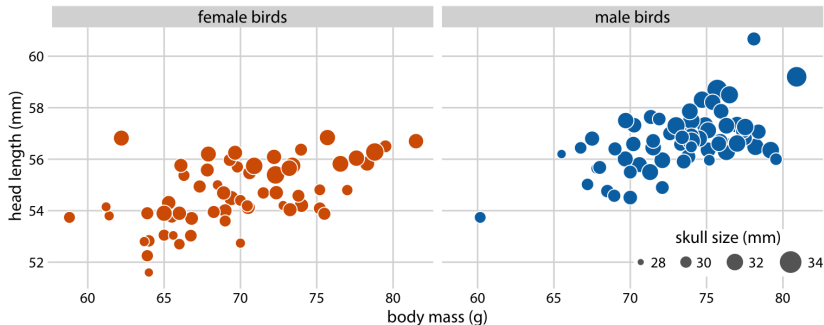
John Snow's 1854 cholera map



Bluejays scatterplot by sex



Adding skull size with bubbles

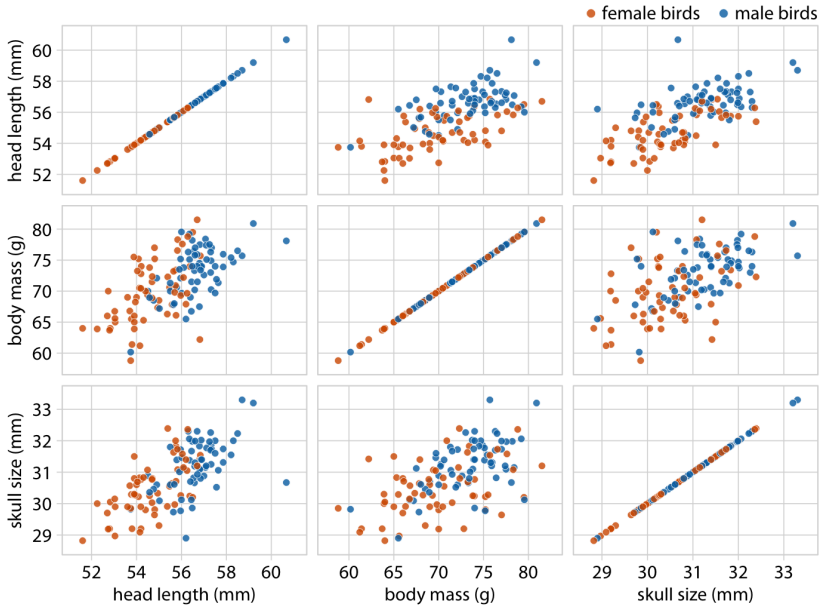


- Head length is from tip of bill to back of head.
- Skull size does not include the bill.
- Plotting a fourth variable with bubble size

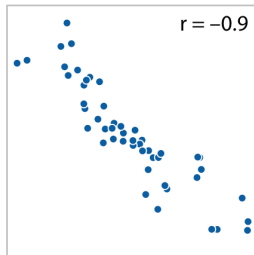
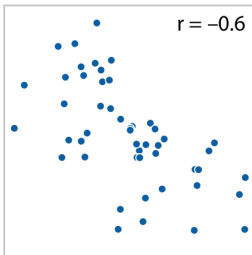
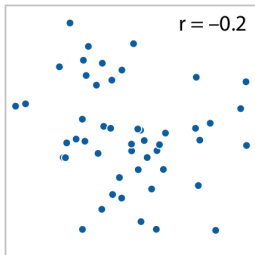
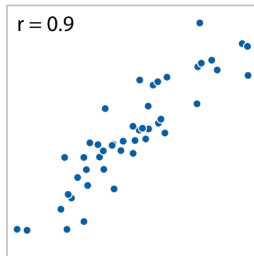
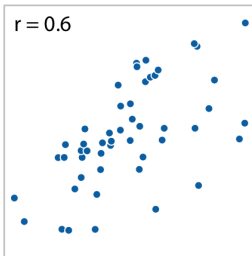
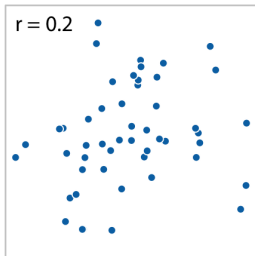
Bubble charts

- Bubble charts show quantitative variables with two aesthetics:
 - position
 - size
- Difficult to visualize the strength of association.
- Bubble size is harder to perceive than position

All-against-all scatterplot matrix



Correlation coefficients



Correlation coefficients

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

where x_i and y_i are two sets of observations and \bar{x} and \bar{y} are the sample means.

- Symmetric in x and y
- Only depend on the differences from the mean, so independent of shifting.
- Independent of scaling, too, since a constant C will appear in both numerator and denominator

Root mean square

Correlation coefficient:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

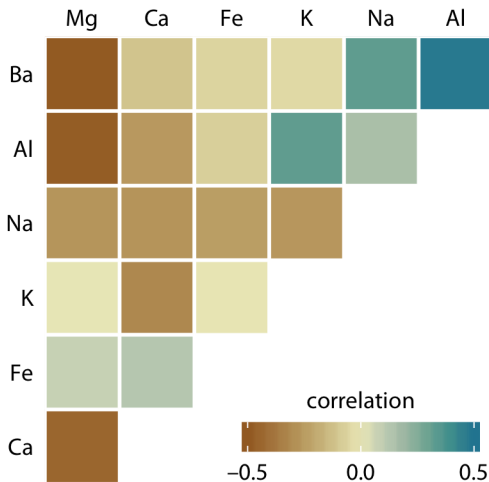
Root mean square:

$$\sqrt{\frac{\sum_i a_i^2}{n}}$$

Standard deviation:

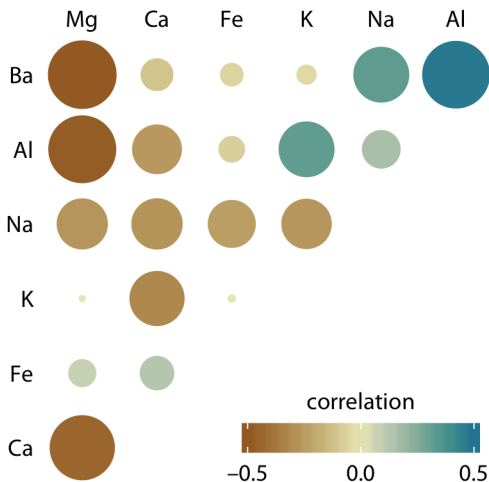
$$\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$$

Correlegram maps the correlations to a visualization



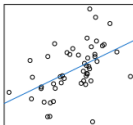
- Forensic data on glass samples

Size and color together visualize small values better

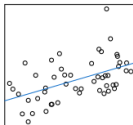


All correlations: $r(50) = 0.5$

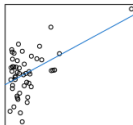
(1) Normal x, normal residuals



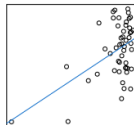
(2) Uniform x, normal residuals



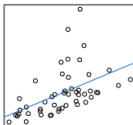
(3) ++skewed x, normal residuals



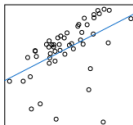
(4) --skewed x, normal residuals



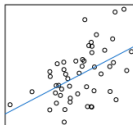
(5) Normal x, ++skewed residuals



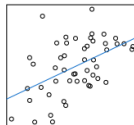
(6) Normal x, --skewed residuals



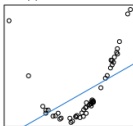
(7) Increasing spread



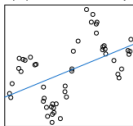
(8) Decreasing spread



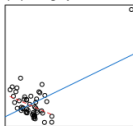
(9) Quadratic trend



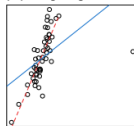
(10) Sinusoid relationship



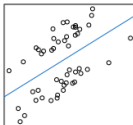
(11) A single positive outlier



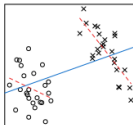
(12) A single negative outlier



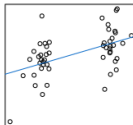
(13) Bimodal residuals



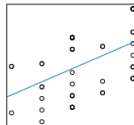
(14) Two groups



(15) Sampling at the extremes



(16) Coarse data



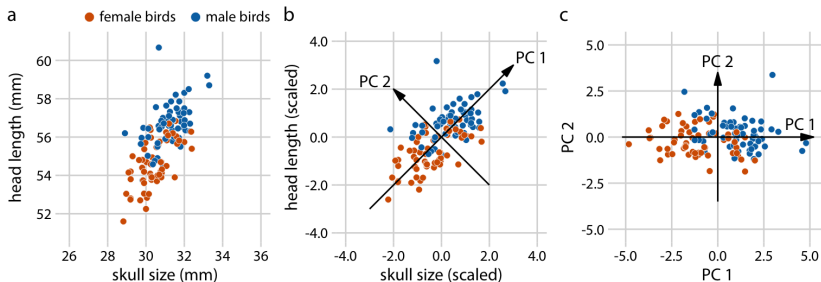
Principal components

[https://builtin.com/data-science/
step-step-explanation-principal-component-analysis](https://builtin.com/data-science/step-step-explanation-principal-component-analysis)

Dimensionality reduction

- Many variables are correlated:
 - height
 - weight
 - arm length, leg length
 - chest, waist, leg circumference
- If we combine these into a single super-variable we may be able to see other, more subtle factors.

Principal components

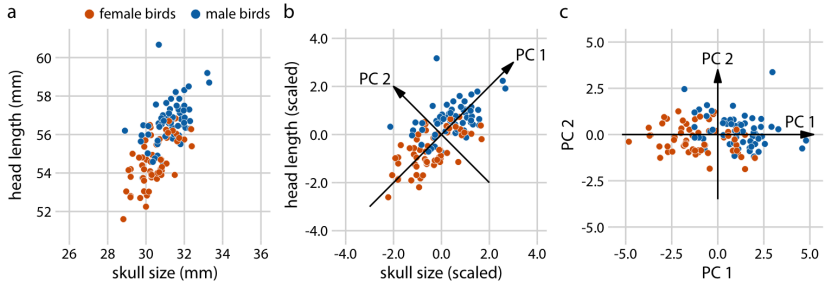


- Scale each variable to zero mean and unit variance
- A rigid rotation of the data around the origin
- Each component has zero correlation with the others
- The first component has the most variance, the second the second most, *etc.*

Correlation with normalized variables

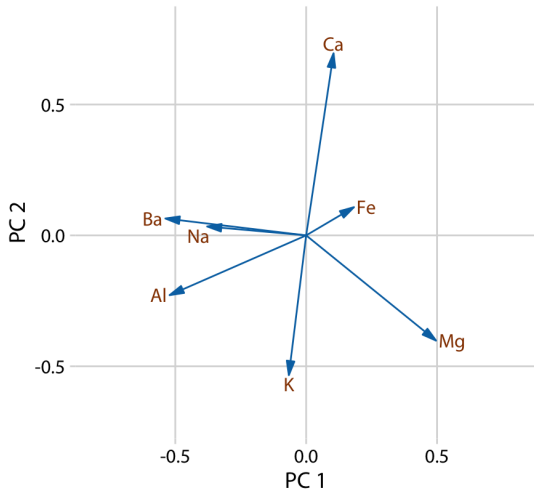
$$\begin{aligned} r &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \\ &= \frac{\sum_i (x_i)(y_i)}{\sqrt{\sum_i (x_i)^2} \sqrt{\sum_i (y_i)^2}} \\ &= \frac{\sum_i (x_i)(y_i)}{1 \cdot 1} \\ &= \sum_i (x_i)(y_i) \\ &= \vec{x} \cdot \vec{y} \\ &= \vec{x}^T \vec{y} \end{aligned}$$

Principal components



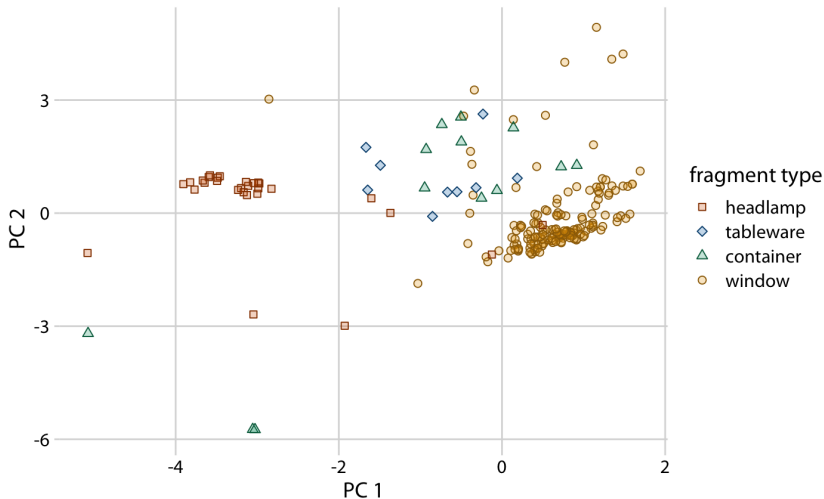
- Usually key features of the data can be seen in the first two or three components.
- Two things to look at:
 - the composition of the components
 - the positions of the points on the components

Composition of the components



- The components are linear combinations of the variables

Project points onto the components



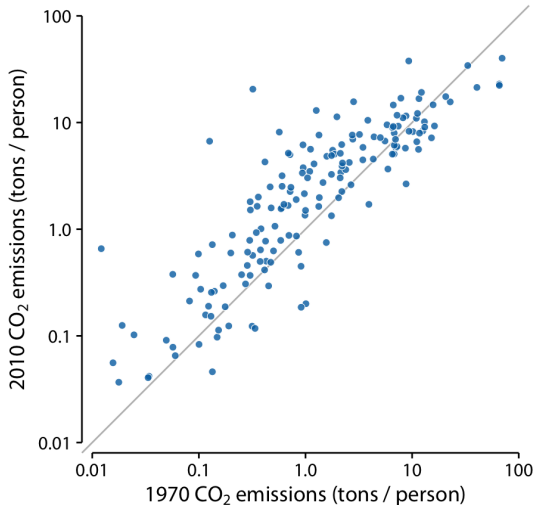
Principal components

- Sometimes the dominant trend is known:
 - Summer-winter
 - Time
 - Treatment
- Verify that PC1 captures this trend.
- Look for other trends in PC2, PC3, etc.

Paired data

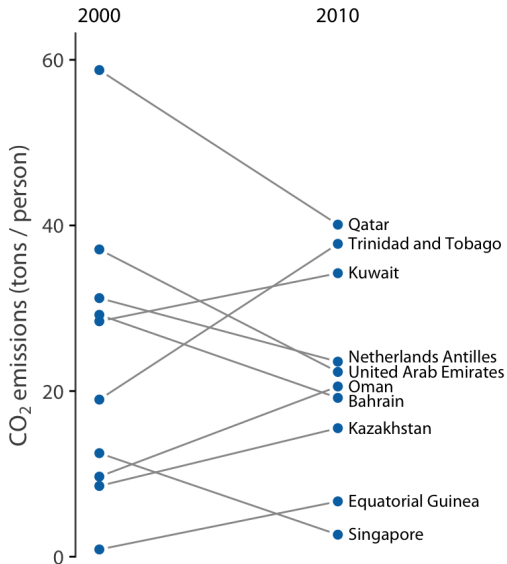
- Data where there are two or more measurements of the same quantity under slightly different conditions.
- Examples include:
 - two comparable measurements on each subject (e.g., the length of the right and the left arm of a person),
 - repeat measurements on the same subject at different time points (e.g., a person's weight at two different times during the year),
 - or measurements on two closely related subjects (e.g., the heights of two identical twins).

Scatter plot with diagonal line where $x = y$



- Countries are consistent over the decades
- Systematic shift to higher emissions

Slopegraph, for smaller numbers of subjects



Slopegraph can be used for multiple years

