CP321 Data Visualization

- Visualize Associations & Image file format

Jiashu (Jessie) Zhao

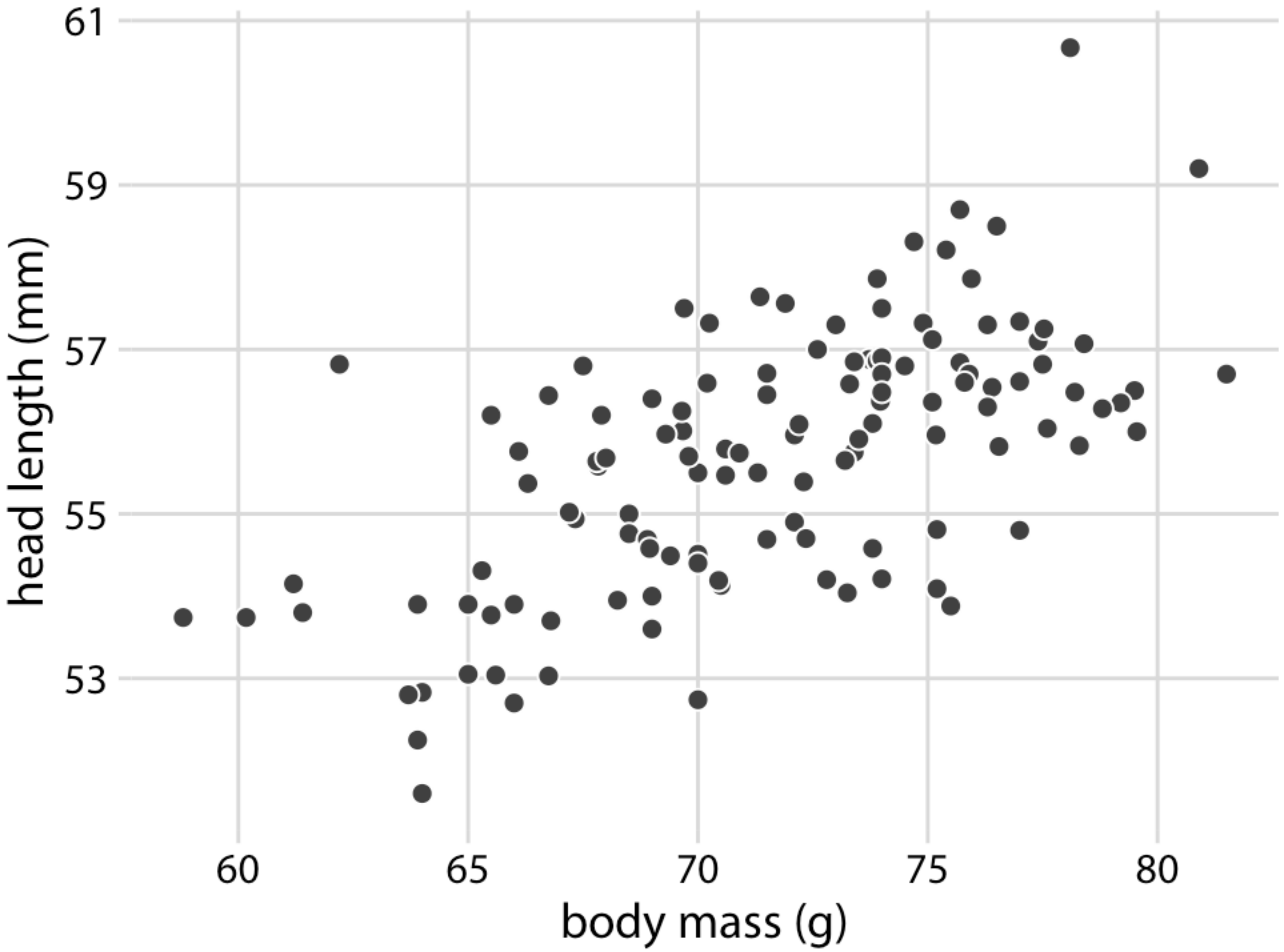
Visualizing associations

- How the quantitative variables relate to each other
- To plot the relationship of just *two* such variables, e.g. the height and weight, we will normally use a *scatter plot*.
- To show *more than two variables* at once, we may plot a *bubble chart*, a *scatter plot matrix*, or a *correlogram*

Scatter plot

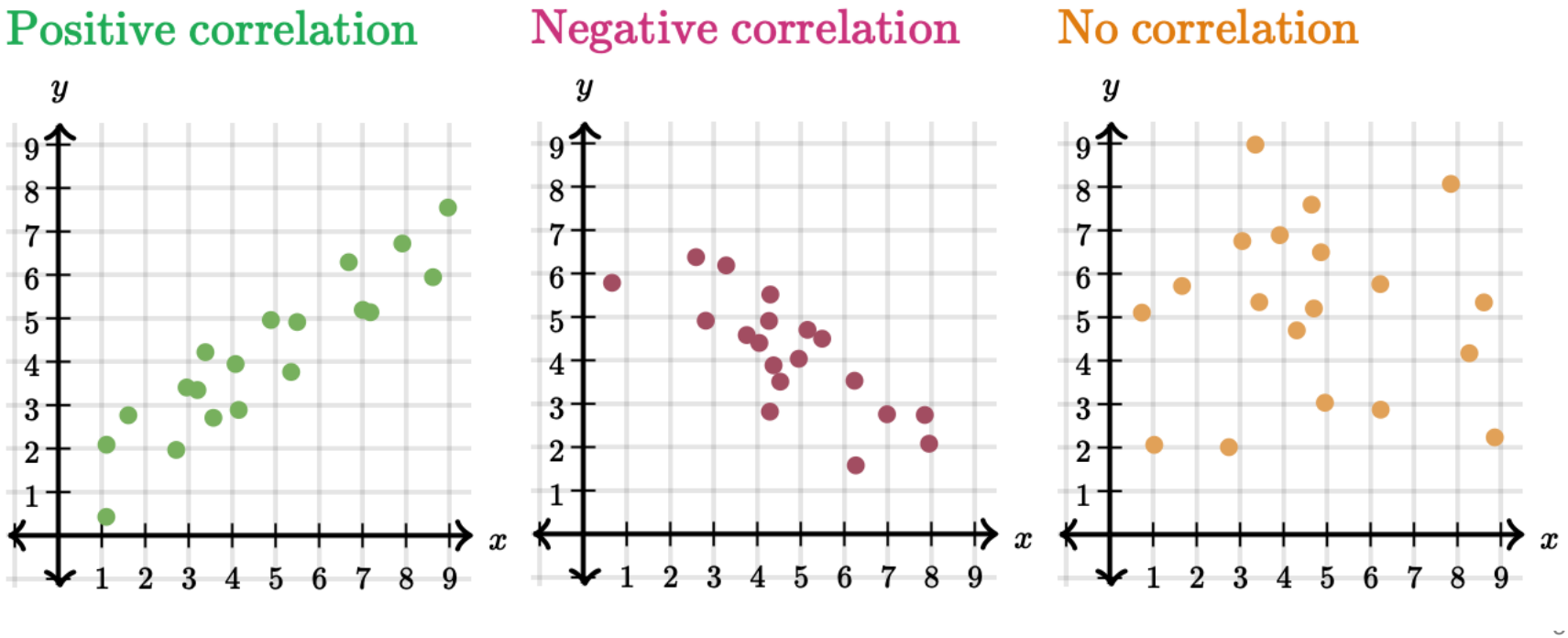
- Each instance of the dataset gets plotted as a point whose (x,y) coordinates relates to its values for the two variables.

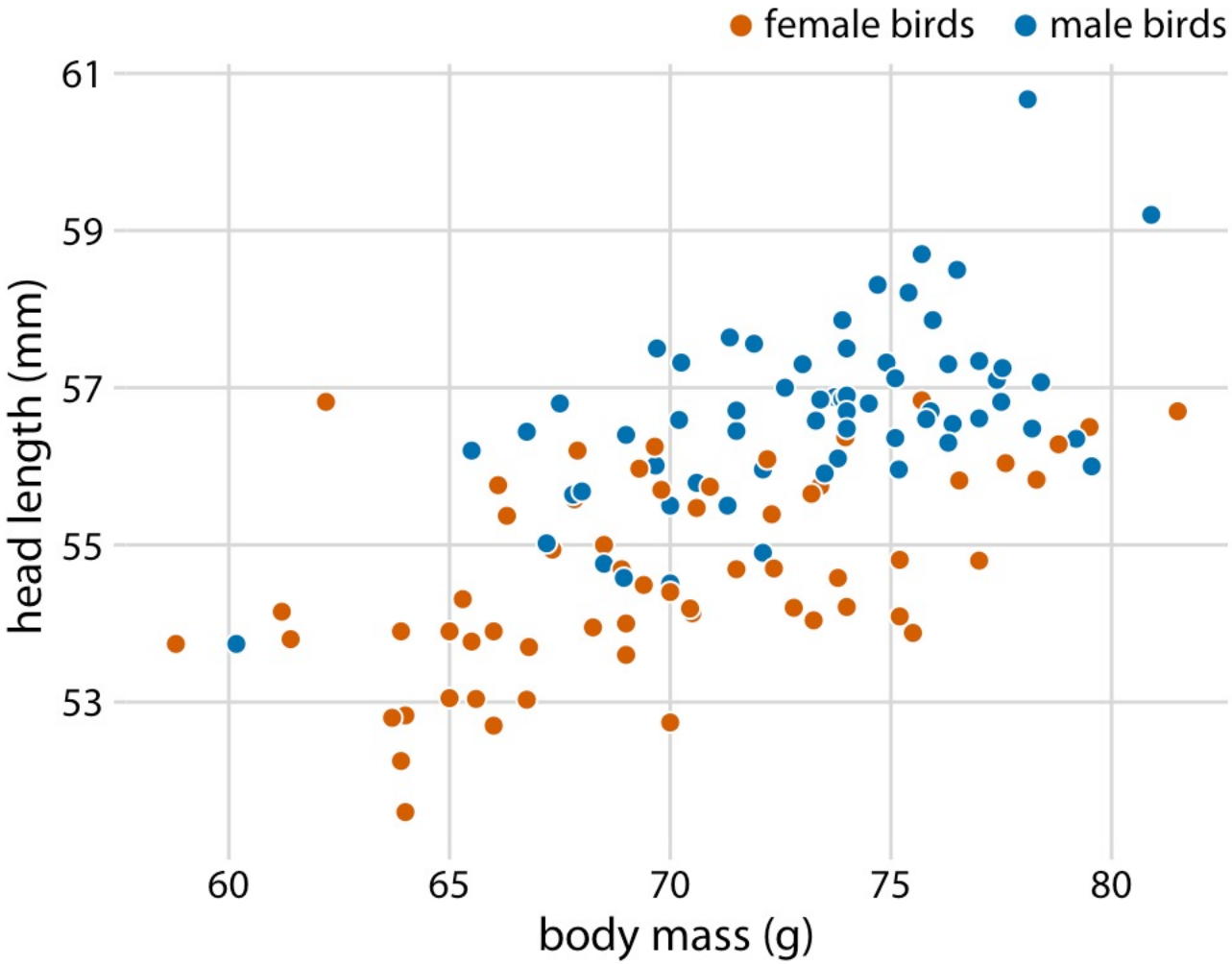
- Patterns or relationships in scatterplots represent correlation between the variables.



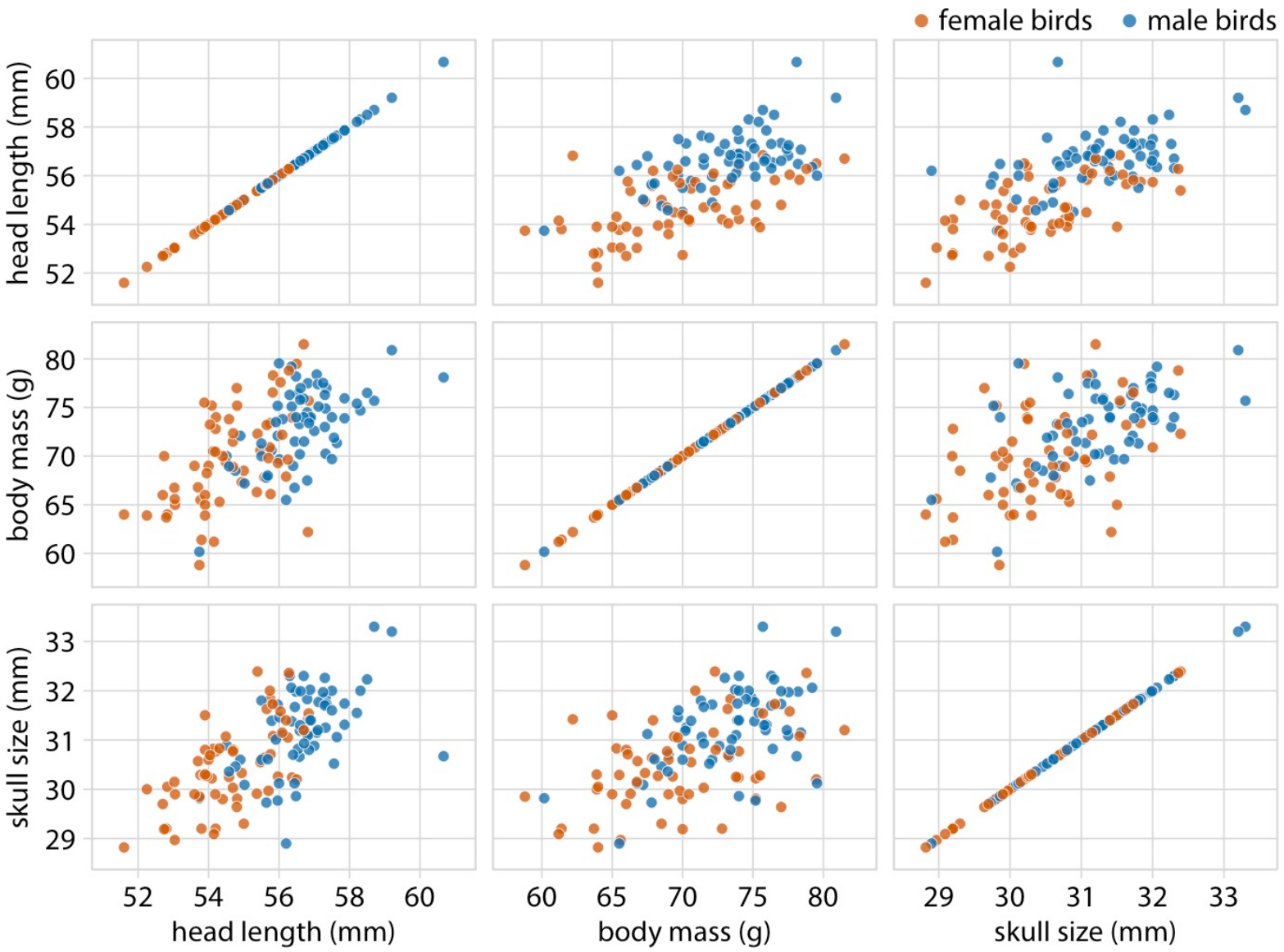
Information of 123 blue jay birds: head length against body mass. Each dot corresponds to one bird.

- In a scatter plot, when the y variable tends to increase as the x variable increases, we say there is a **positive correlation** between the variables.





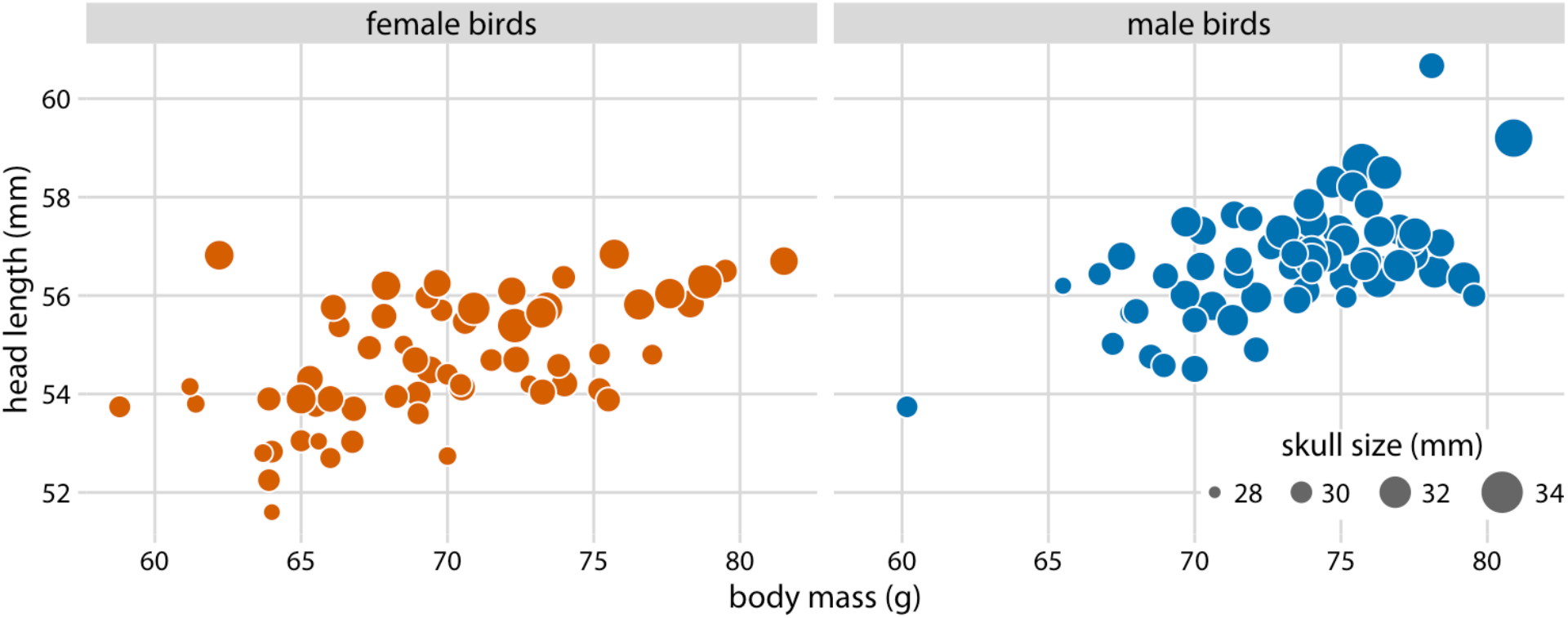
Head length versus body mass for 123 blue jays. The birds' sex is indicated by color.



Scatter Plot Matrix

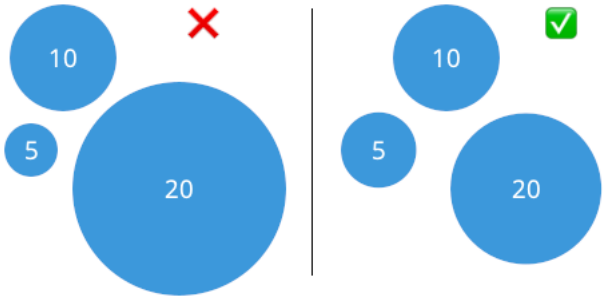
Bubble chart

- An extension of the scatter plot used to look at relationships between three numeric variables.

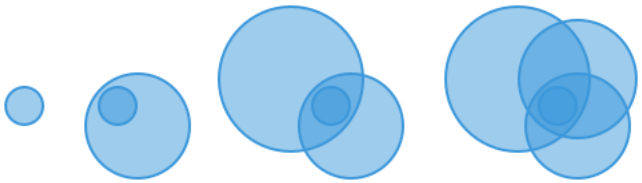


Key points for plotting a bubble chart.

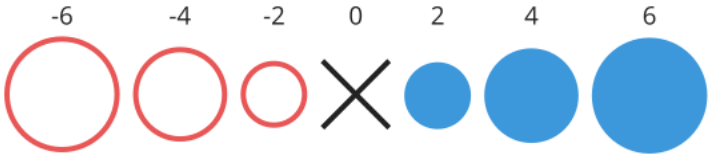
- Scale bubble area by value



- Limit number of points to plot



- Present a clear trend
- Include a legend
- Incorporating negative values (when the negative numbers make sense)

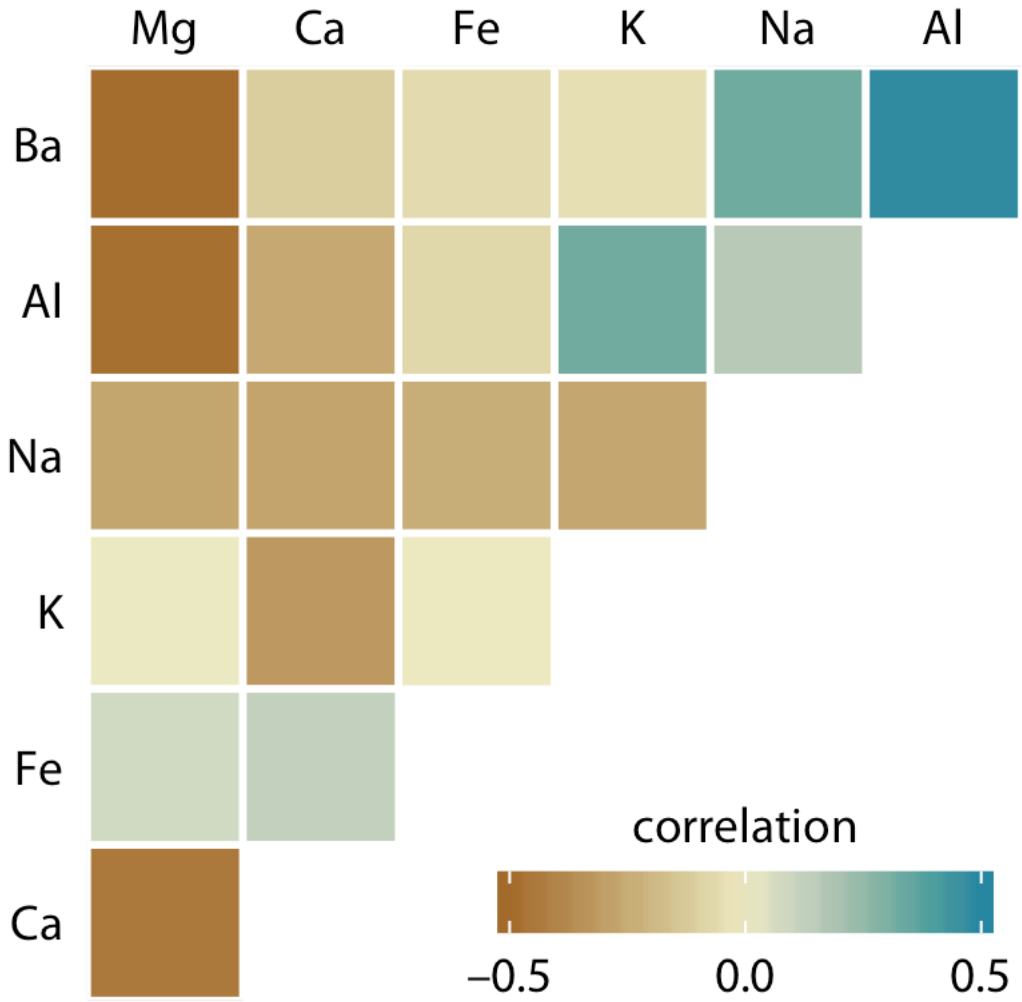


Correlograms

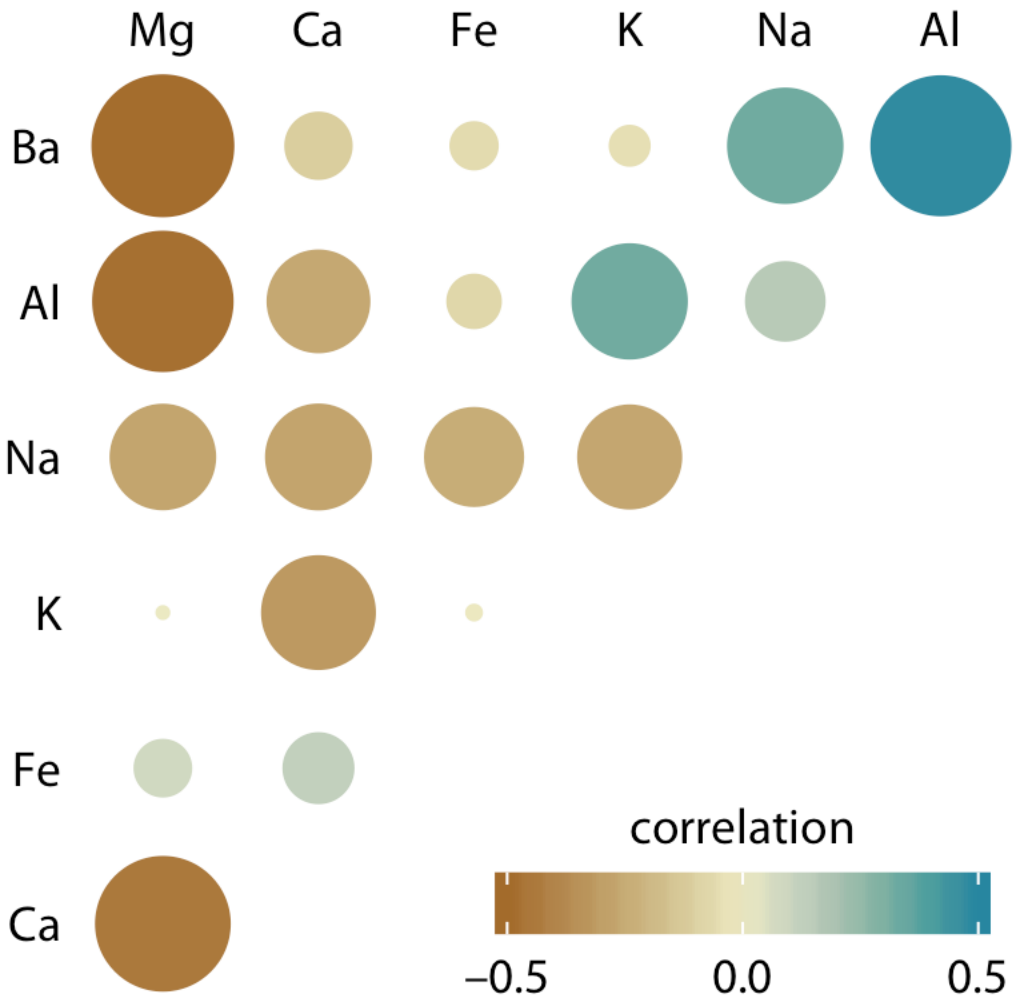
- Visualizations of correlation coefficients are called *correlograms*.

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

- An abstract way for visualize associations.



Correlations in mineral content for 214 samples of glass fragments obtained during forensic work. The dataset contains seven variables measuring the amounts of magnesium (Mg), calcium (Ca), iron (Fe), potassium (K), sodium (Na), aluminum (Al), and barium (Ba) found in each glass fragment.



the magnitude of each correlation is also encoded in the size of the colored circles

Paired data

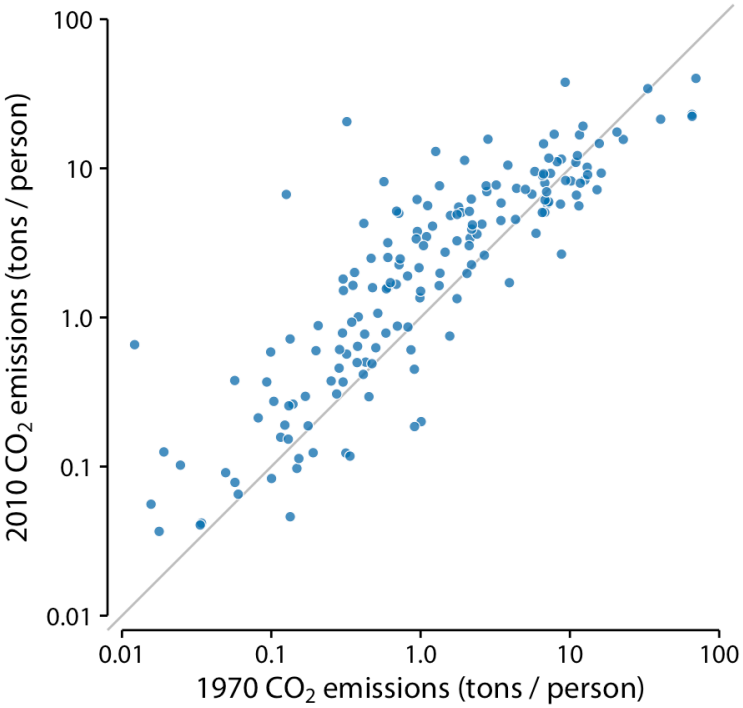
- A special case: Data where there are two or more measurements of the same quantity under slightly different conditions.

- The two measurements belonging to a pair are more similar to each other than to the measurements belonging to other pairs

- For example: the length of the right and the left arm of a person

Visualize Paired Data

- We need to choose visualizations that highlight any differences between the paired measurements.



Scatter Plot for Carbon dioxide (CO_2) emissions per person in 1970 and 2010, for 166 countries.

Slopegraph

- For a small amount of paired observations
- We draw individual measurements as dots arranged into two columns and indicate pairings by connecting the paired dots with a line.
- The slope of each line highlights the magnitude and direction of change.
- Can be used to compare *more than two measurements* at a time

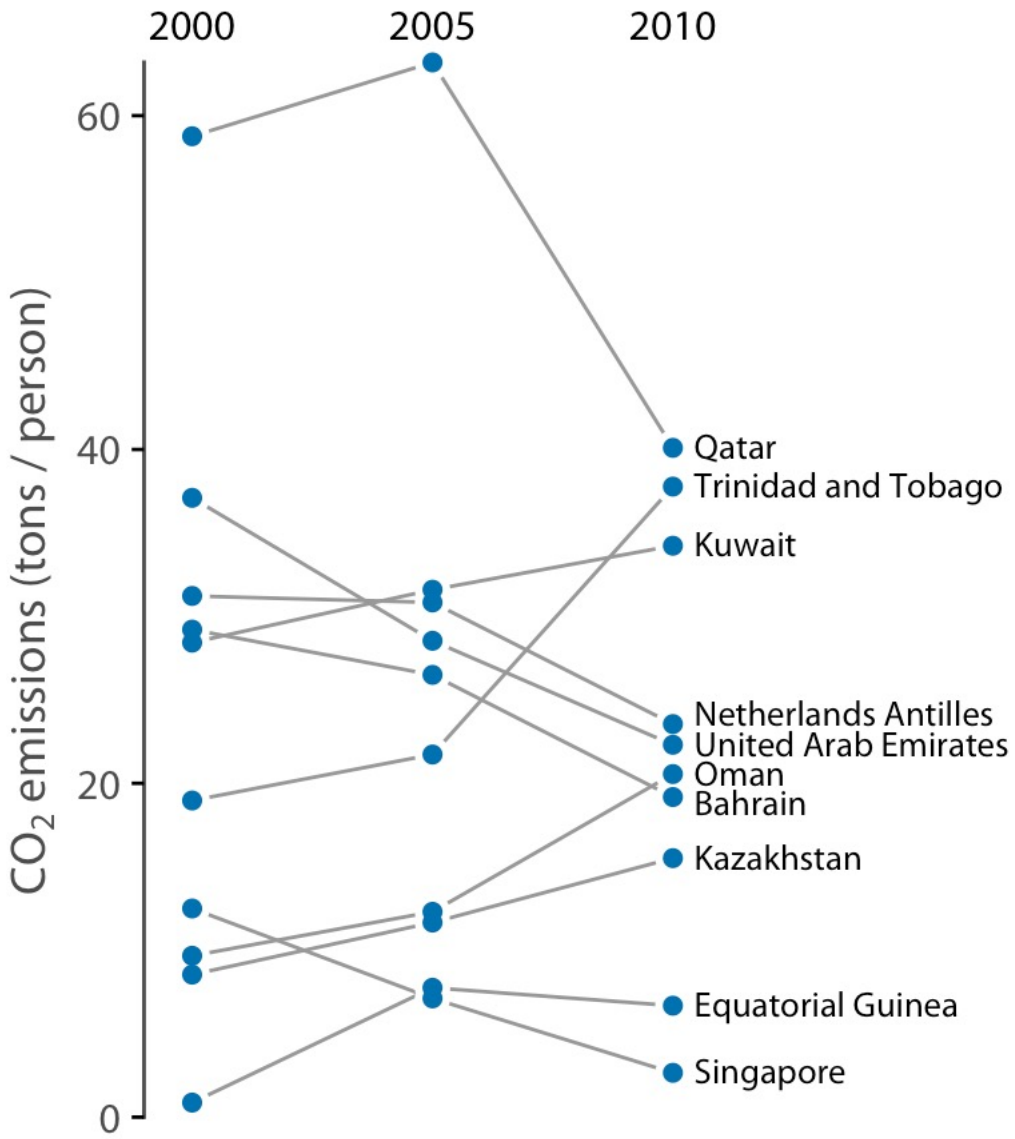
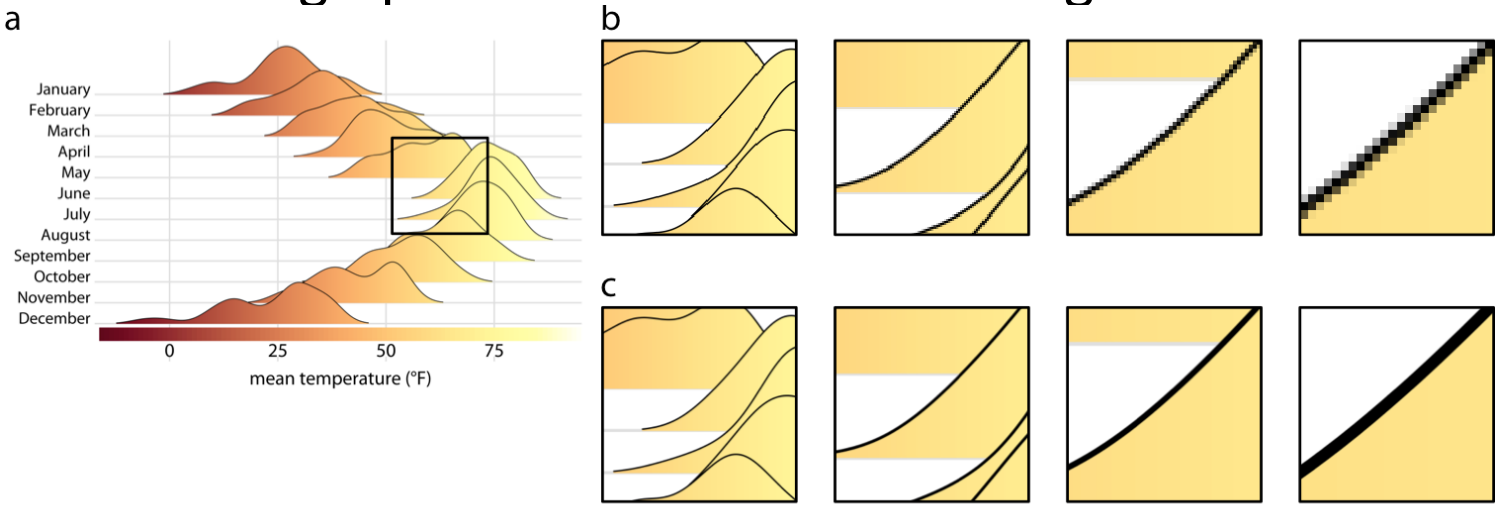

2000
2005
2010
60
40
20
0
CO_2 emissions (tons / person)
Qatar
Trinidad and Tobago
Kuwait
Netherlands Antilles
United Arab Emirates
Oman
Bahrain
Kazakhstan
Equatorial Guinea
Singapore

Image file formats

- Bitmap vs vector
  - Bitmaps graphics store the image as a grid of individual points (called pixels)
  - Vector graphics store the geometric arrangement of individual graphical elements in the image



(a) Original image
(b) Increasing magnification for a bitmap graphic
(c) Increasing magnification for a vector graphic

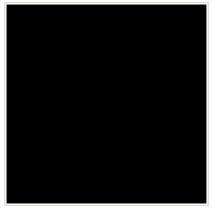
- Vector graphics
  - Resolution-independent
  - May have differences in how the same graphic looks
  - Large in size for complex figures

- Commonly used image file formats

Acronym	Name	Type	Application
pdf	Portable Document Format	vector	general purpose
eps	Encapsulated PostScript	vector	general purpose, outdated; use pdf
svg	Scalable Vector Graphics	vector	online use
png	Portable Network Graphics	bitmap	optimized for line drawings
jpeg	Joint Photographic Experts Group	bitmap	optimized for photographic images
tiff	Tagged Image File Format	bitmap	print production, accurate color reproduction
raw	Raw Image File	bitmap	digital photography, needs post-processing
gif	Graphics Interchange Format	bitmap	outdated for static figures, Ok for animations

- Compression of bitmap graphics
  - Most bitmap file formats employ some form of data compression to keep file sizes manageable.
  - Two types of compression
    - Lossless compression guarantees that the compressed image is pixel-for-pixel identical to the original image (e.g. png)
    - Lossy compression accepts some image degradation in return for smaller file sizes (e.g. jpeg)

How to achieve lossless compression?



10,000 pixels

RGB(0,0,0)

(0 0 0)
(0 0 0)
…
(0 0 0)

Need to store
30,000 numbers

Store two
numbers
30,000 0

Which kind of images could the lossless compression algorithms perform the best?

- How to achieve lossy compression?
  - Some details in an image are too subtle for the human eye, and those can be discarded without obvious degradation in the image quality.

Which kind of images could the lossy compression algorithms perform the best?

1000 pixels, each with a slightly different color value

200 different colors and coloring every five adjacent pixels in the exact same color