CSCI 491: Data Visualization

12- Visualizing Associations Among Two or More Quantitative Variables

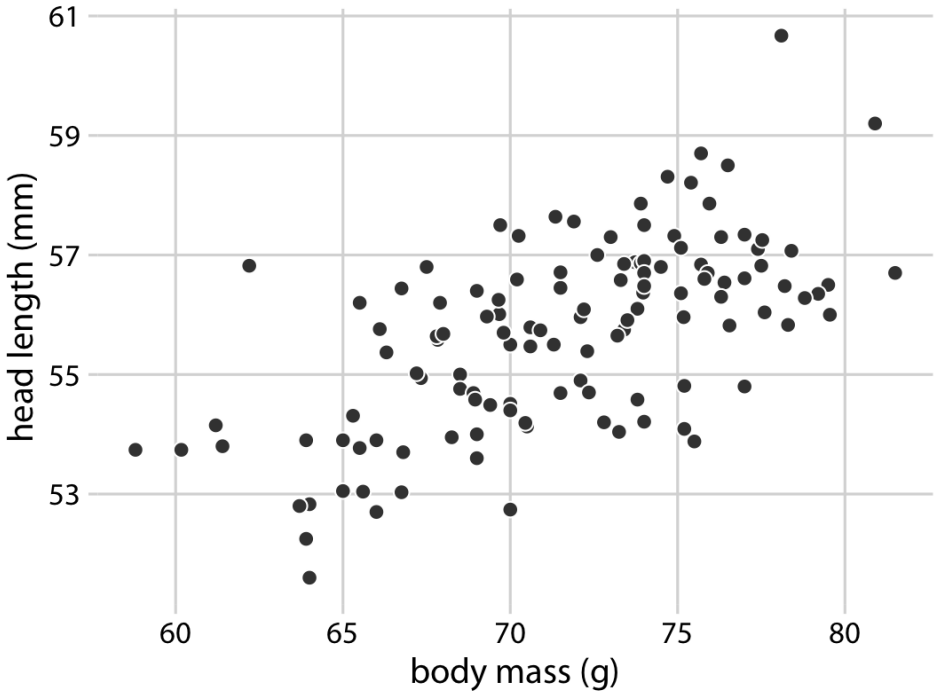
Visualizing Associations Among Two or More Quantitative Variables

- To plot the relationship of just two such variables, such as the height and weight, we will normally use a scatterplot
- If we want to show more than two variables at once, we may opt for a bubble chart, a scatterplot matrix, or a correlogram
- For very high-dimensional datasets, it may be useful to perform dimension reduction, for example in the form of principal components analysis.

Scatterplot (two variables)

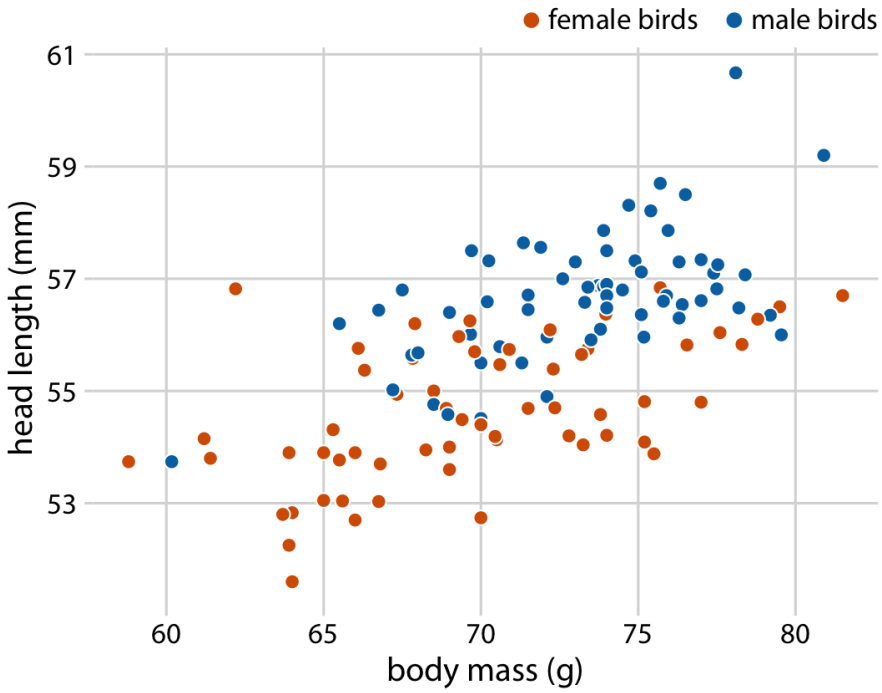
- The dataset contains measurements performed on 123 blue jay birds. It incudes information such as the head length, the skull size, and the body mass of each bird. We expect that there are relationships between these variables.

There is a moderate tendency for heavier birds to have longer heads. Data source: Keith Tarvin, Oberlin College.



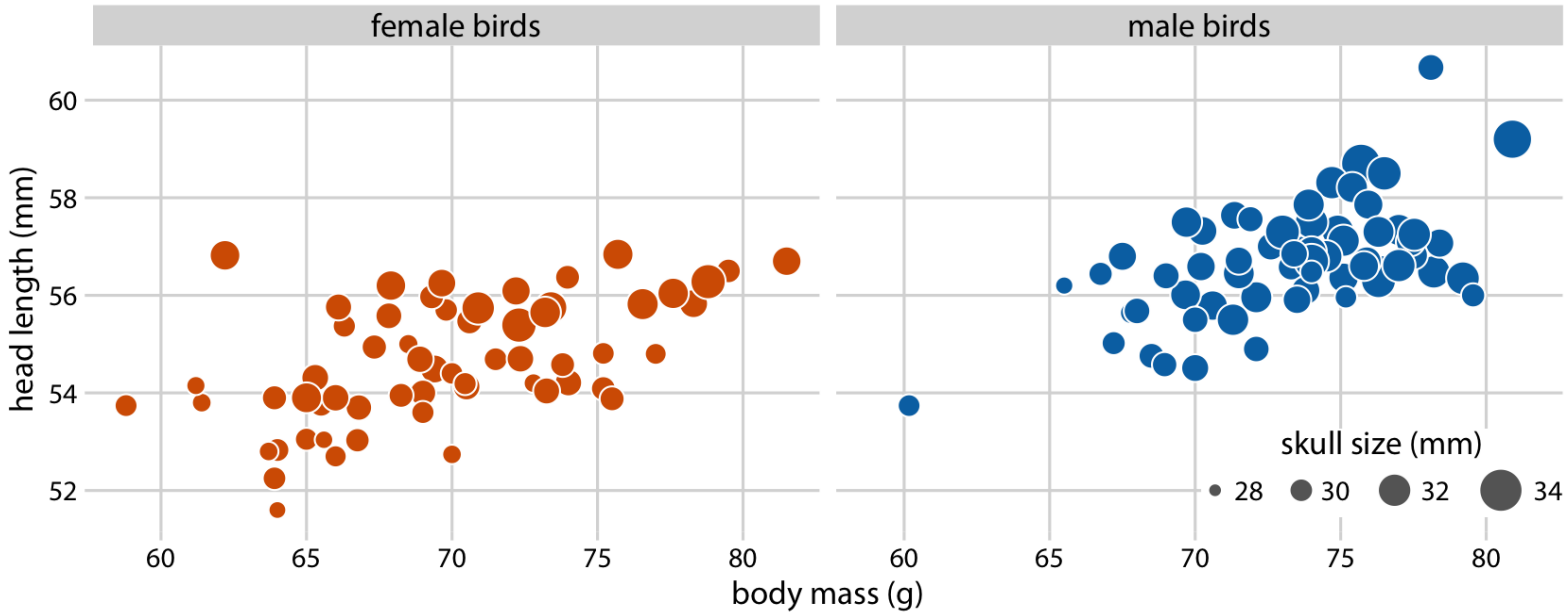
Scatterplot (three vars)

- At the same body mass, females tend to have shorter heads than males.
- At the same time, females tend to be lighter than males on average.



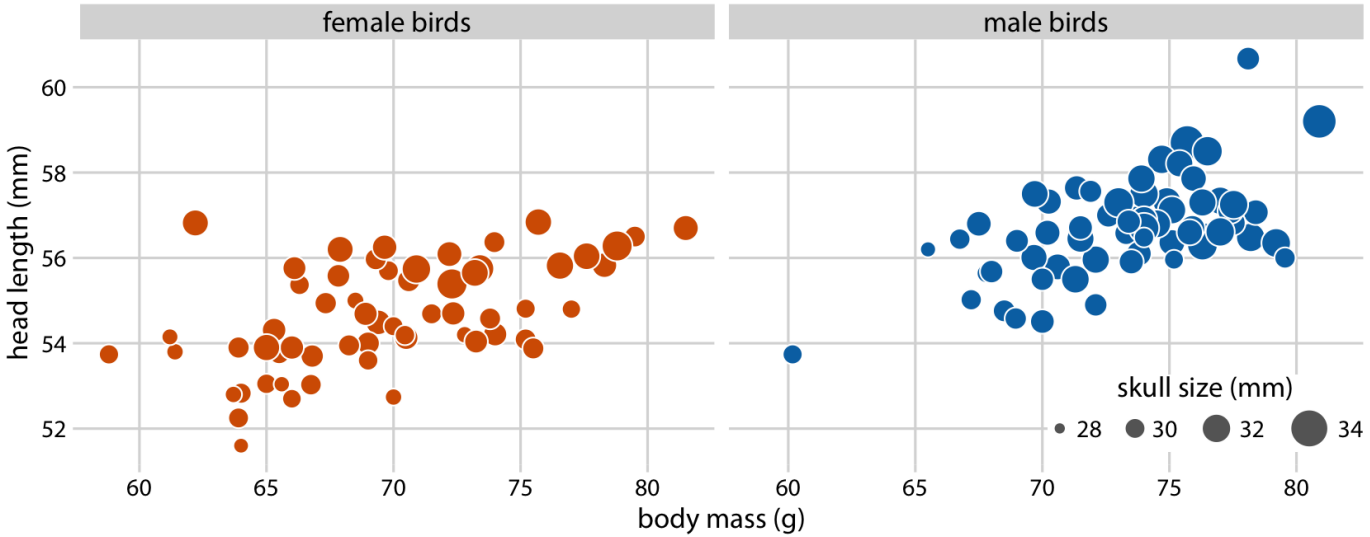
Scatterplot (four vars) or Bubble Chart

- Bubble charts have the disadvantage that they show the same types of variables—quantitative variables—with two different types of scales, position and size.

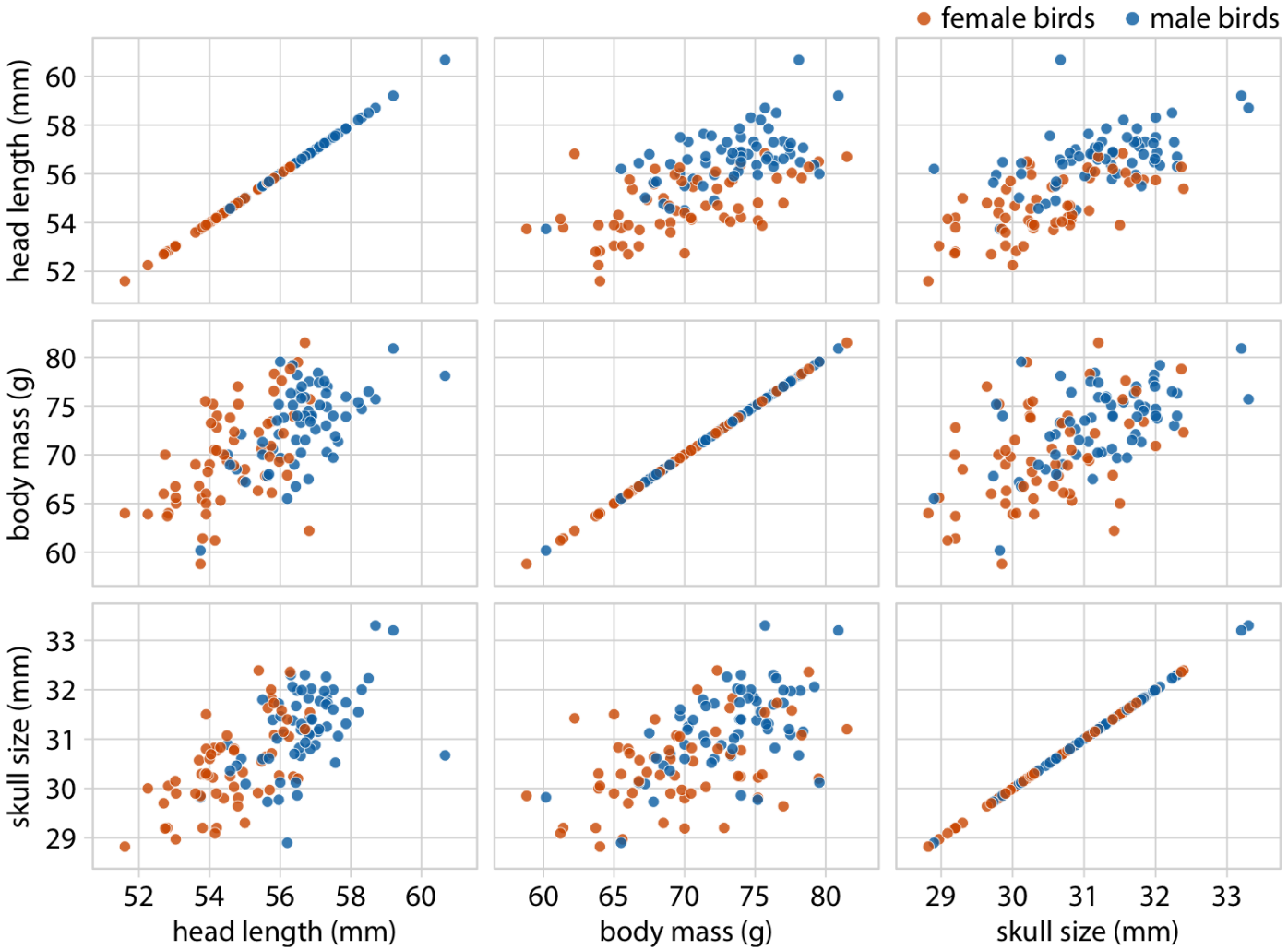


Bubble Chart Limitation

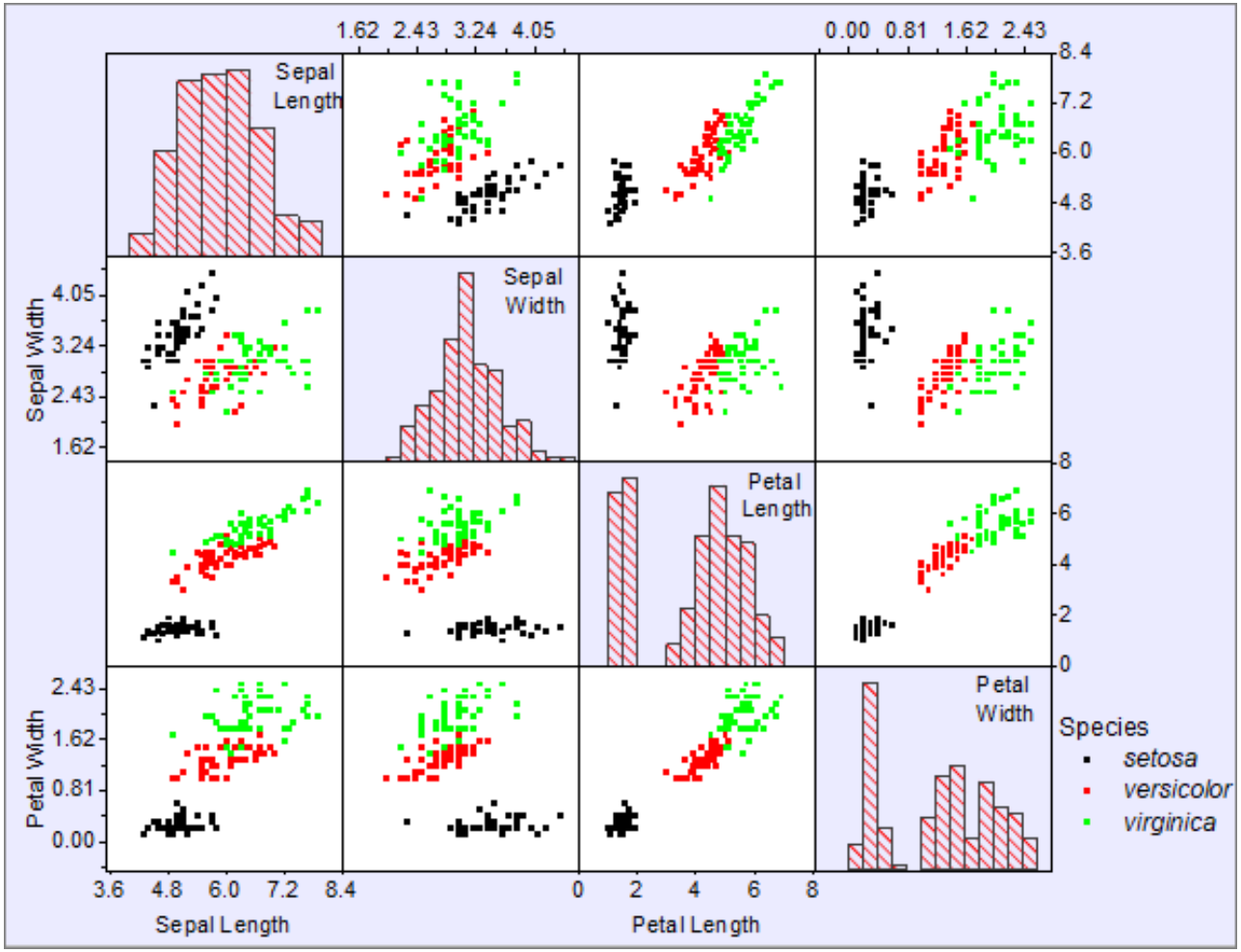
- This makes it difficult to visually identify the strengths of associations between the various variables.
- Differences between data values encoded as bubble size are harder to perceive than differences between data values encoded as position.



Pairwise Scatterplots



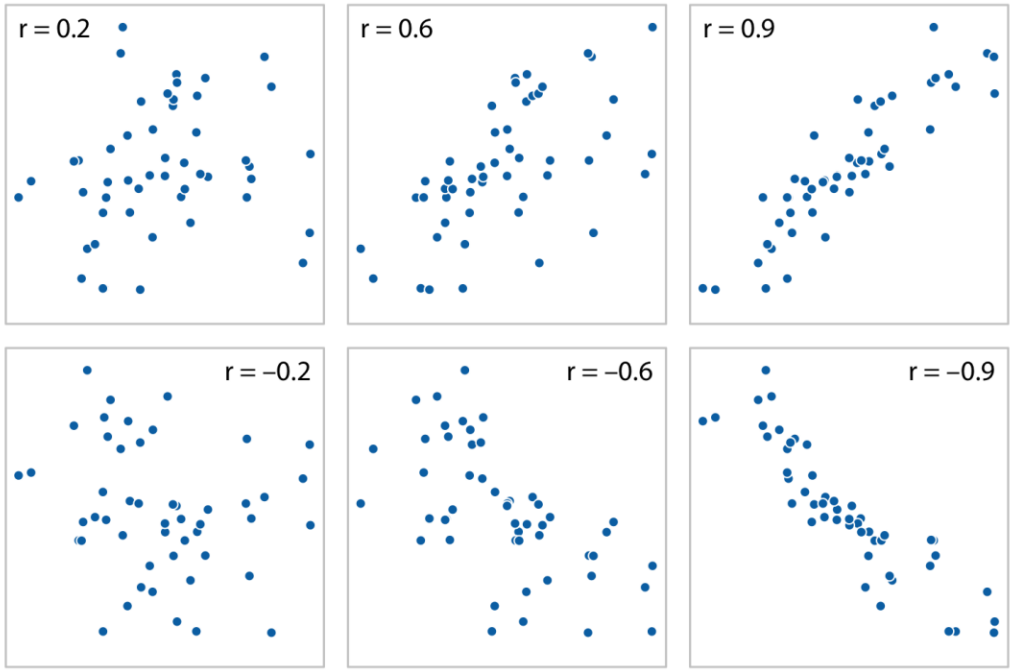
Pairwise Scatterplots with Diagonal Histograms



Correlograms

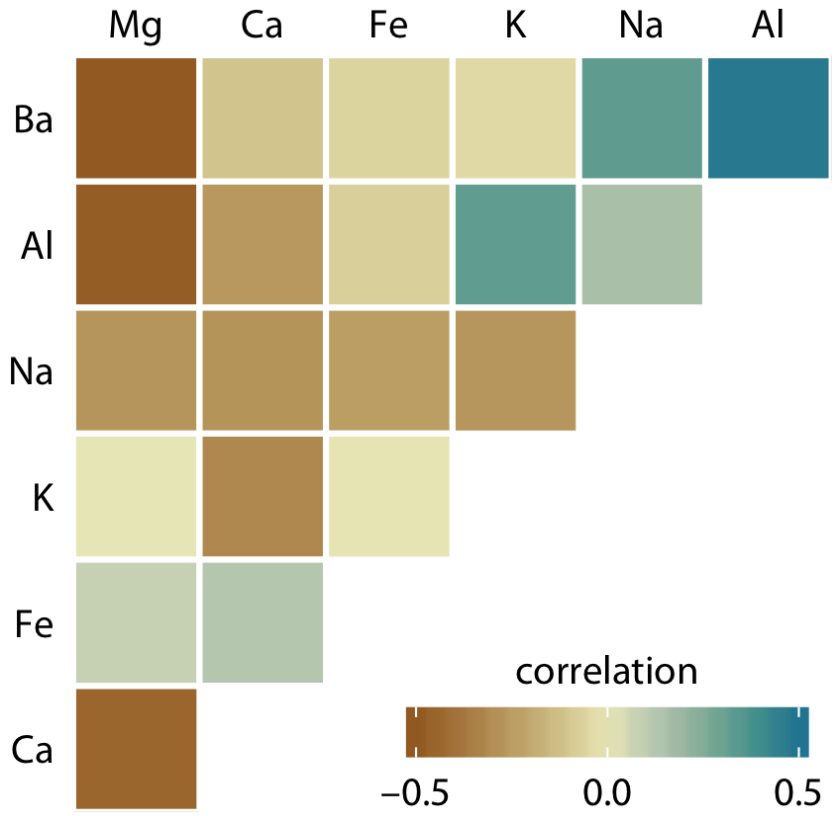
A value of r = 0 means there is no association whatsoever, and a value of either 1 or –1 indicates a perfect association. The sign of the correlation coefficient indicates whether the variables are correlated or anticorrelated.

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



Correlograms

Correlations in mineral content for 214 samples of glass fragments obtained during forensic work.



Correlograms Limitation

To overcome this limitation, we can display the correlations as colored circles and scale the circle size with the absolute value of the correlation coefficient