Introduction to Data Visualization

Visualizing Nested Proportions
&
Associations
+
Introduction to ggplot (R)

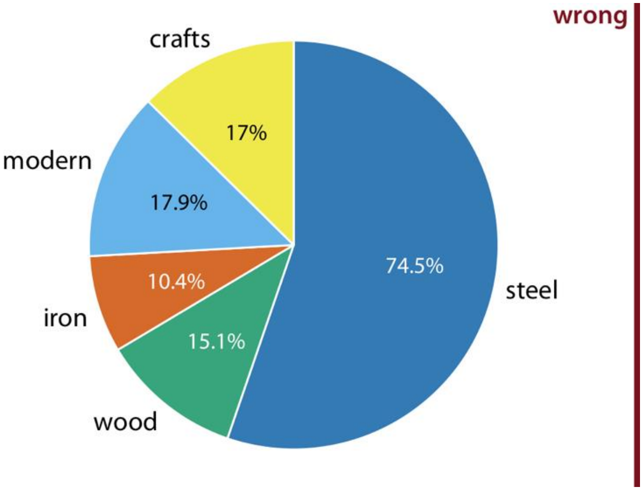
Halil Bisgin, Ph.D.

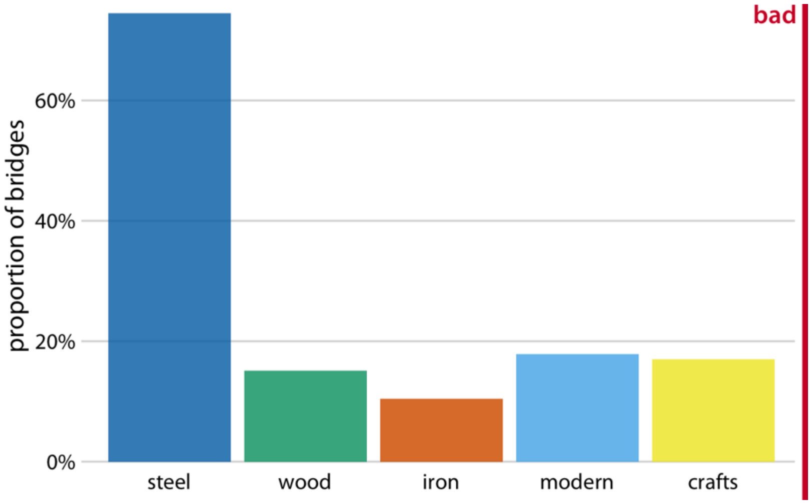
Nested Proportions

- We may want to drill down further and break down a dataset by multiple categorical variables at once.
  - *We could be interested in the proportions of seats by party and by the gender of the representatives.*

Nested Proportions Gone Wrong

- A dataset of 106 bridges in Pittsburgh.
  - *material from which they are constructed (steel, iron, or wood),*
  - *based on the year of erection, bridges are grouped into distinct categories, such as crafts and modern.*
  - *On which river?*



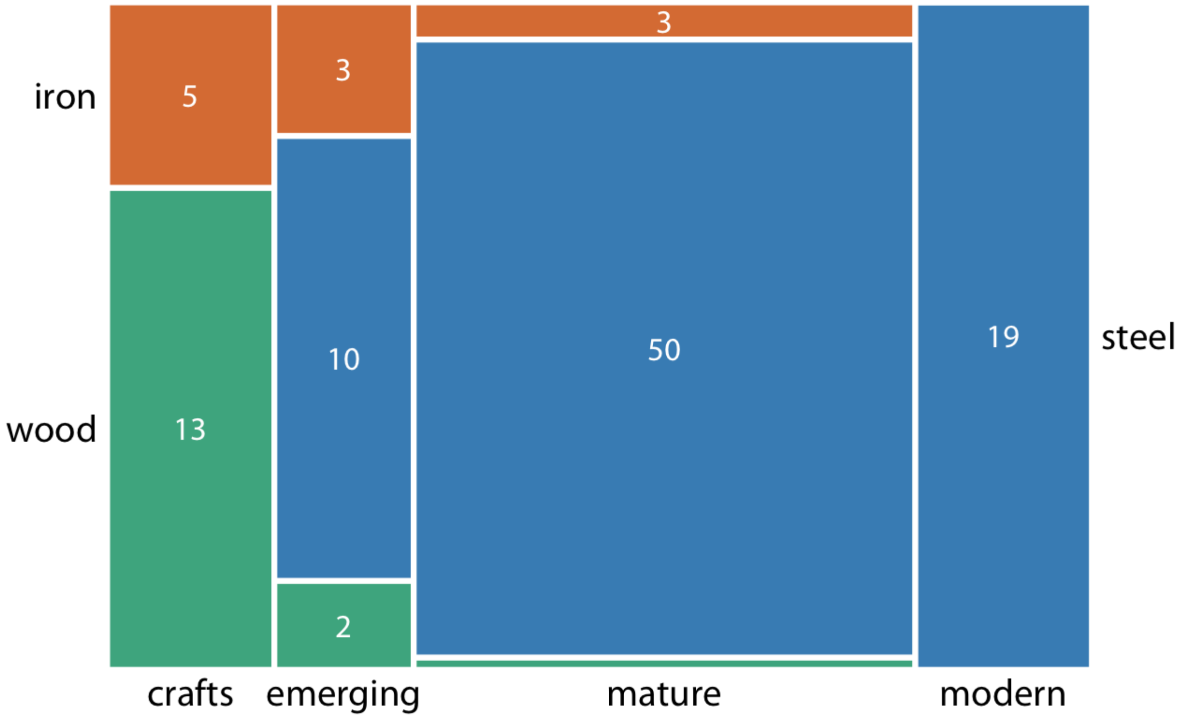


Sum exceeds 100% due to double count

It does not clearly indicate the overlap

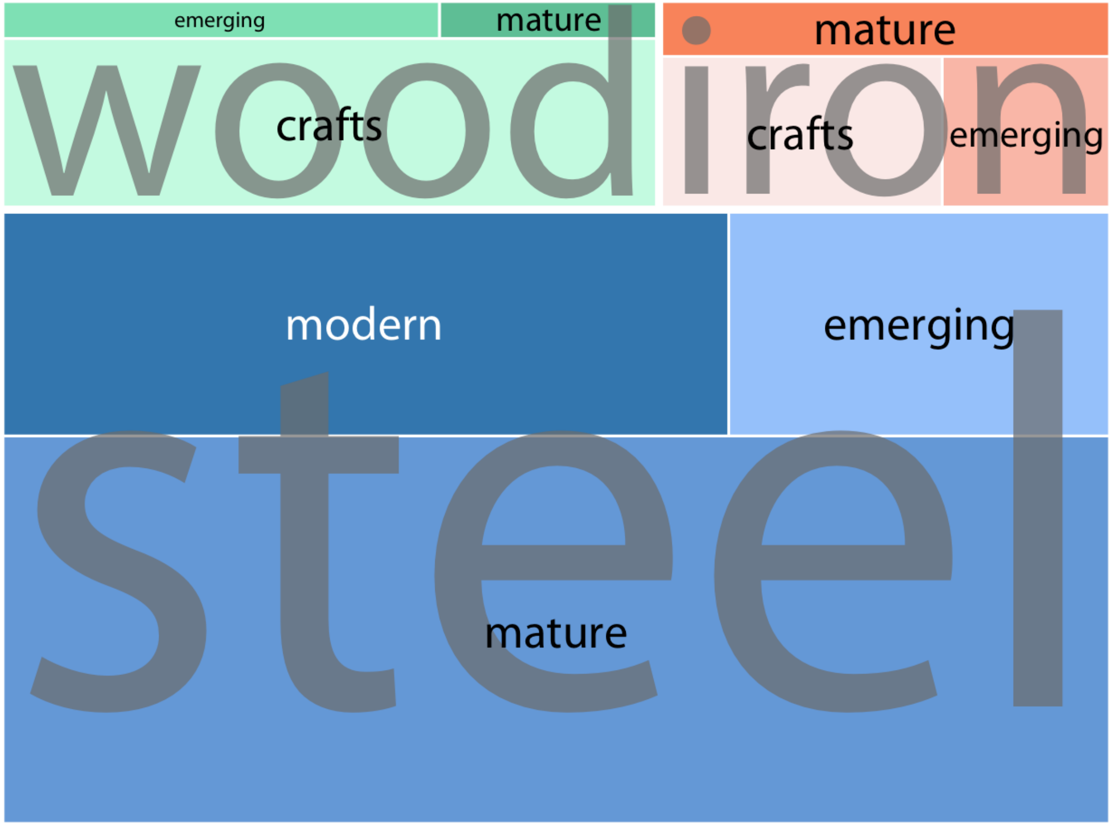
Nested Proportions-Mosaic Plot

- When there are overlapping categories, it is best to show explicitly how they relate to each other.

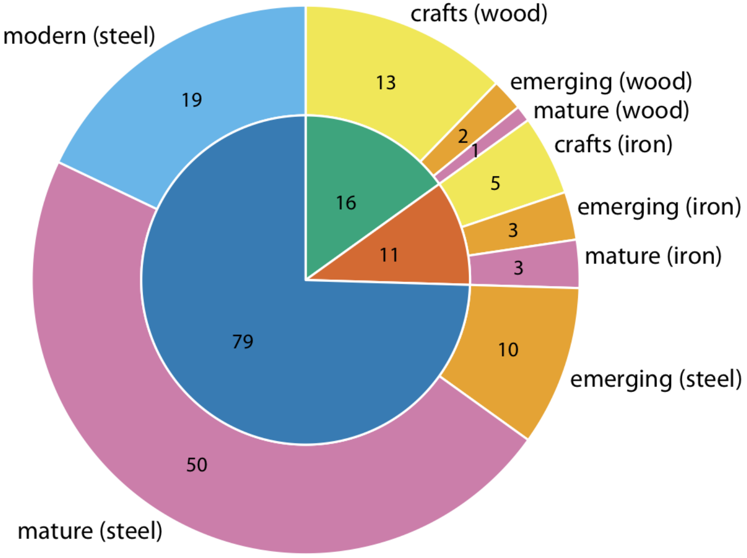


Nested Proportions-Treemap Plot

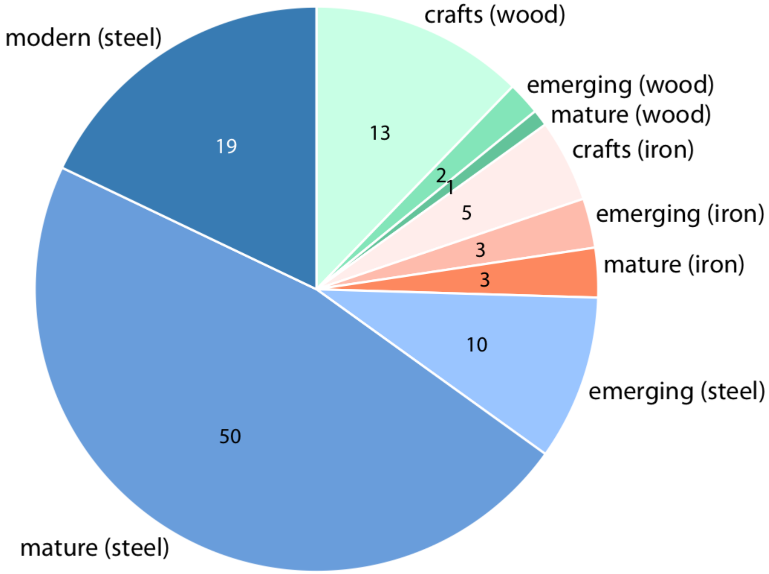
- Shows the counts for every possible combination



Nested Proportions-Nested Pies

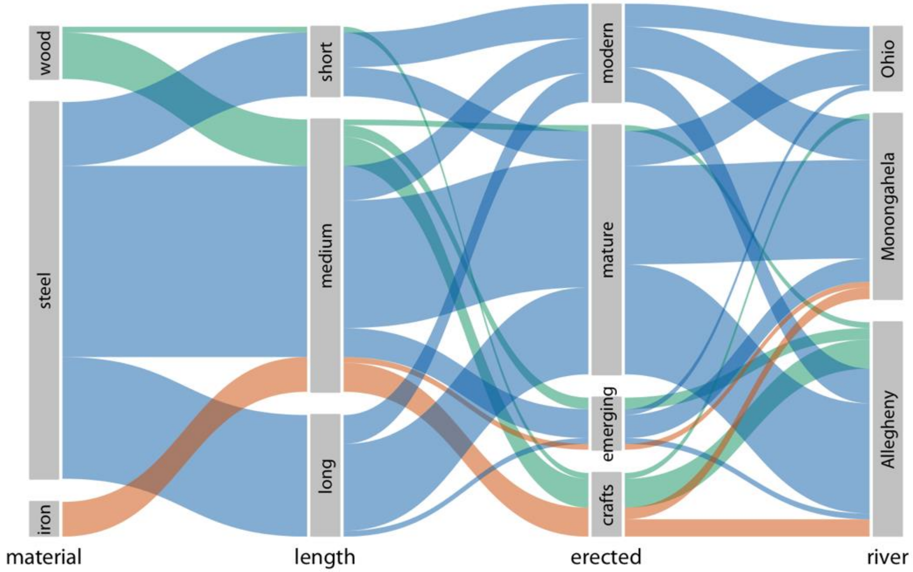


ugly



Nested Proportions-Parallel Sets

- When more than two categorical variables, parallel sets plot can offer a less crowded view.
- It shows how the total dataset breaks down by each individual categorical variable by using shaded bands.

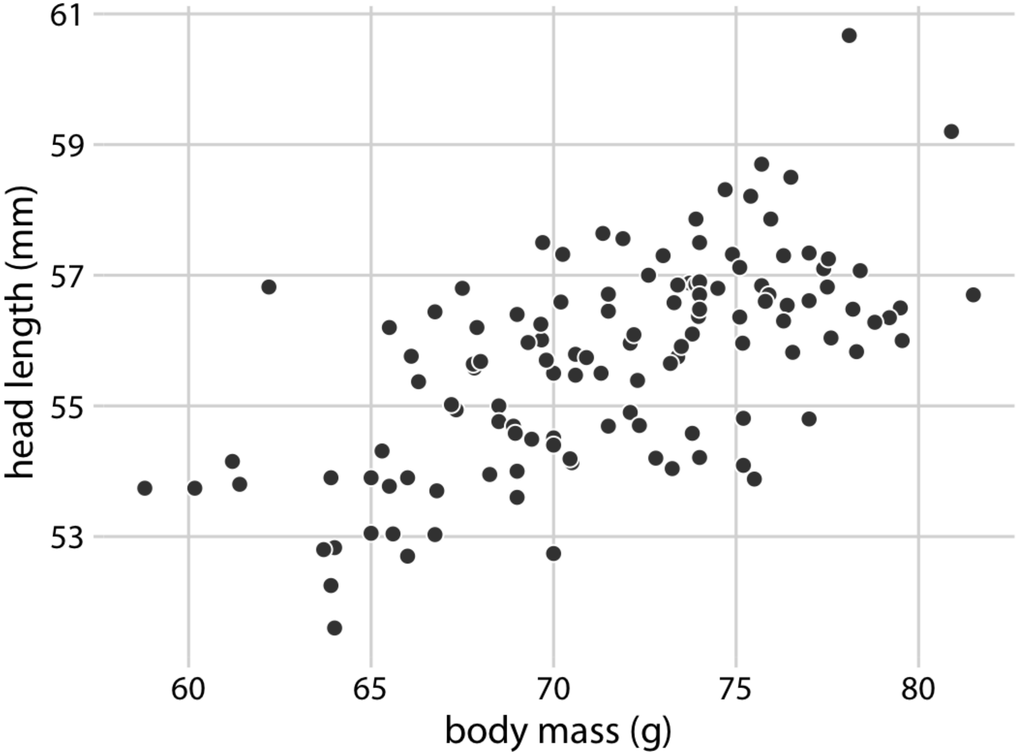


x-y Relationships

- Many datasets contain two or more quantitative variables, and we may be interested in how these variables relate to each other.
    - *animals' height, weight, length, and daily energy demands.*
- Two variables → scatterplot.
- More than two variables → bubble chart, scatterplot matrix, correlogram.
- For very high-dimensional datasets, it may be useful to perform dimension reduction,
    - *Principal components analysis.*

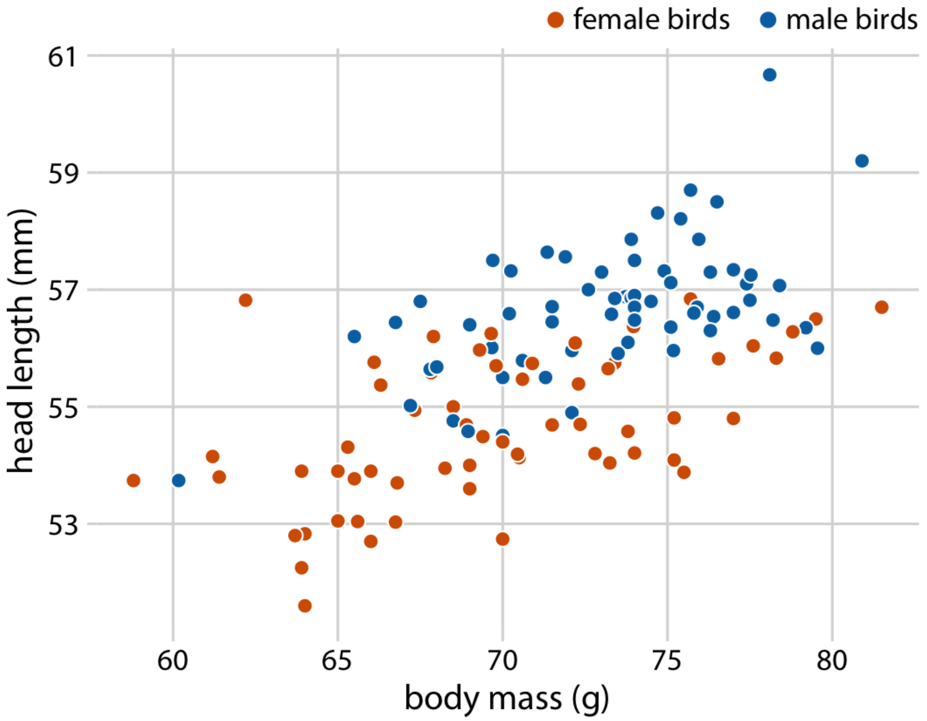
Scatterplots

- The dataset of 123 blue jay birds
  - *head length (measured from the tip of the bill to the back of the head)*
  - *the skull size (head length-bill length)*
  - *the body mass*



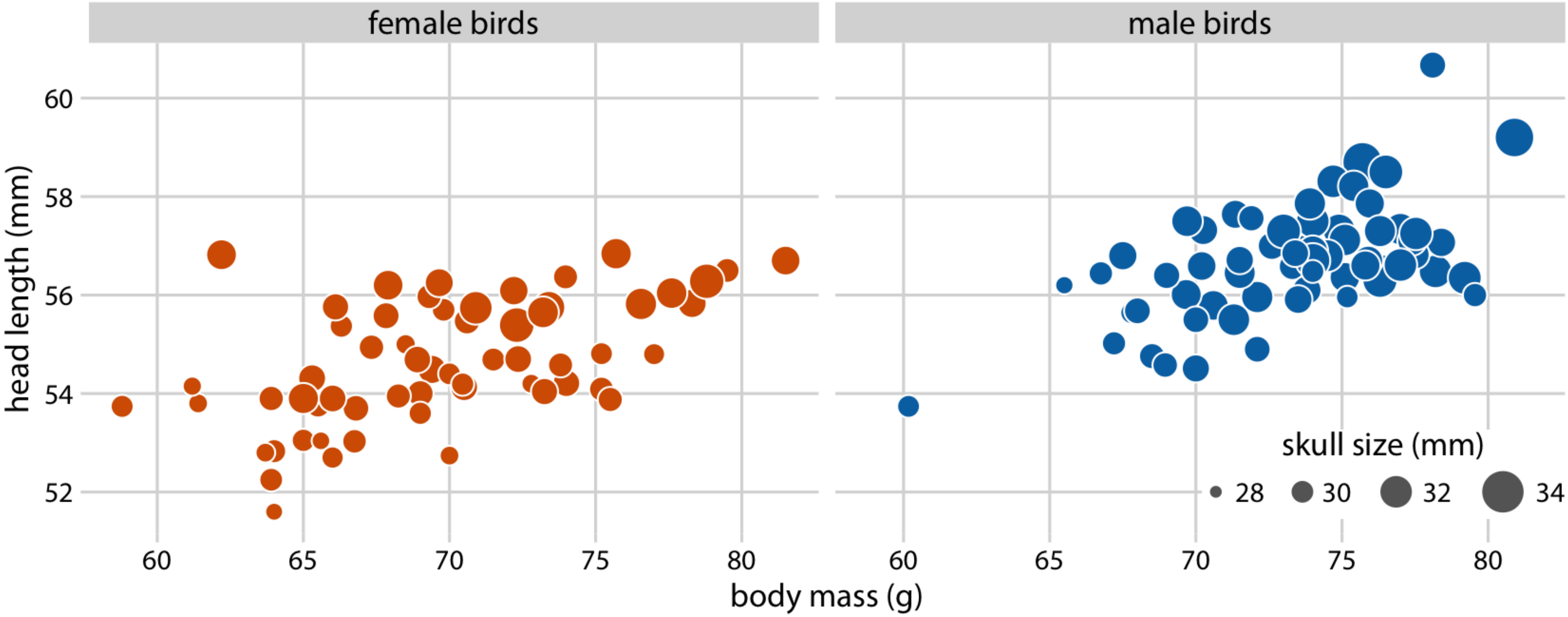
Scatterplots

- The blue jay dataset contains both male and female birds
  - *The relationship between head length and body mass for each sex.*



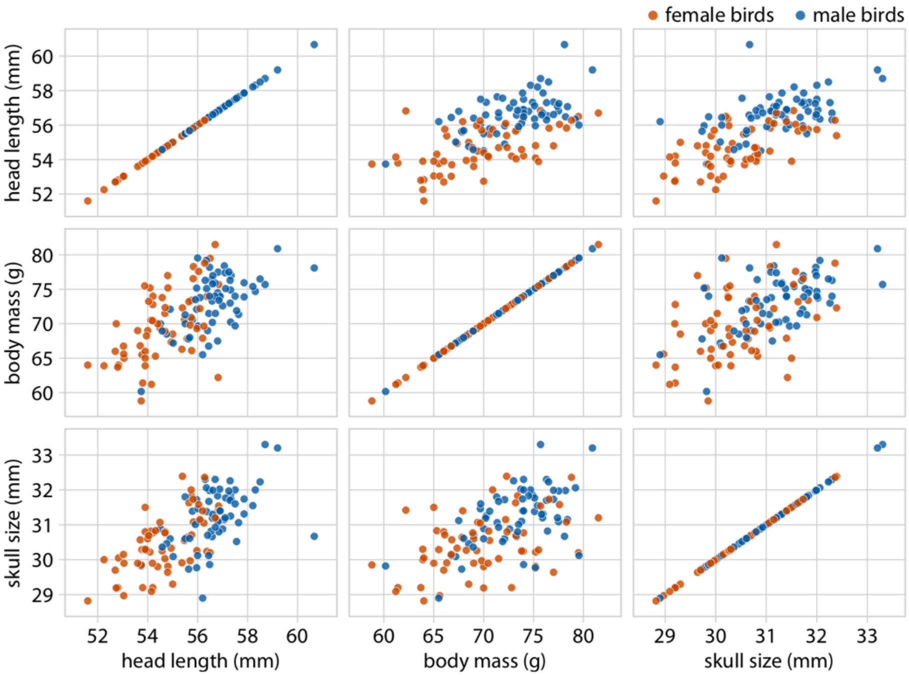
Bubble Chart

- We can disentangle bill length and skull size by looking at another variable in the dataset, the skull size, which is similar to the head length but excludes the bill.



Pairwise plots

- It may be preferable to show an *all-against-all* matrix of scatterplots.



Correlograms

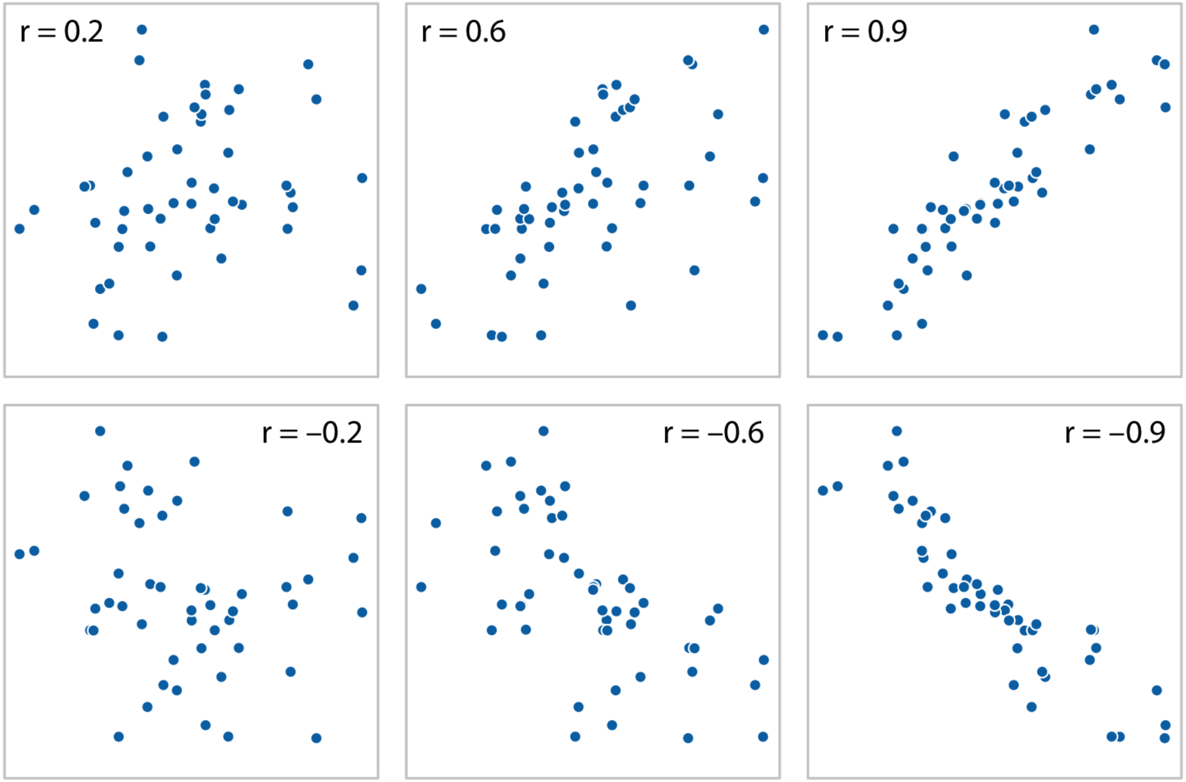
- When there are more variables, it is more useful to quantify the amount of association between pairs of variables and visualize these quantities rather than the raw data.
- Calculate correlation coefficients (r).

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

  - *Between –1 and 1 that measures to what extent two variables covary.*
  - *r = 0 means there is no association whatsoever, 1 or –1 indicates a perfect association.*
  - *r indicates whether the variables are correlated (larger values in one variable coincide with larger values in the other) or anticorrelated (larger values in one variable coincide with smaller values in the other)*

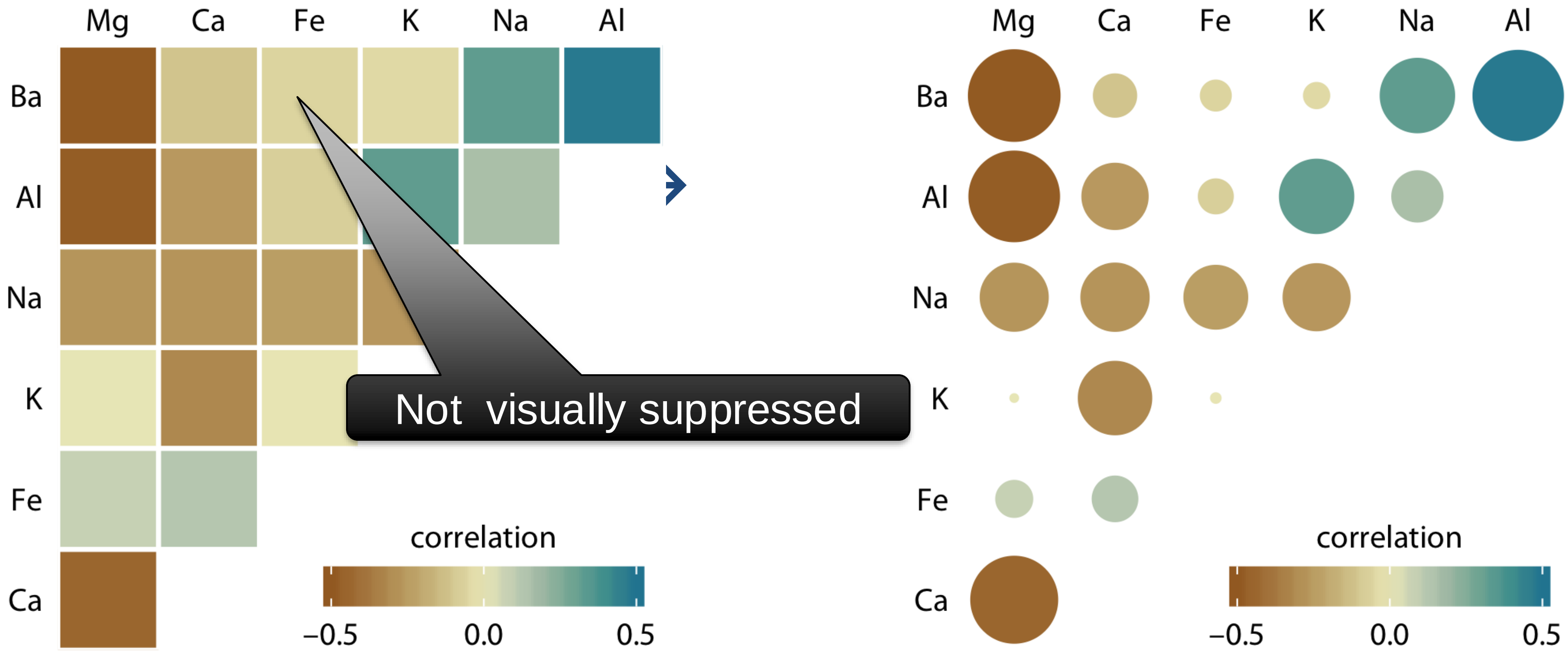
Correlograms

- Examples of correlations of different magnitude and direction, with associated correlation coefficient r



Correlograms

- Dataset of over 200 glass fragments obtained during forensic work.
- Correlations between minerals.



Not visually suppressed

Dimension Reduction

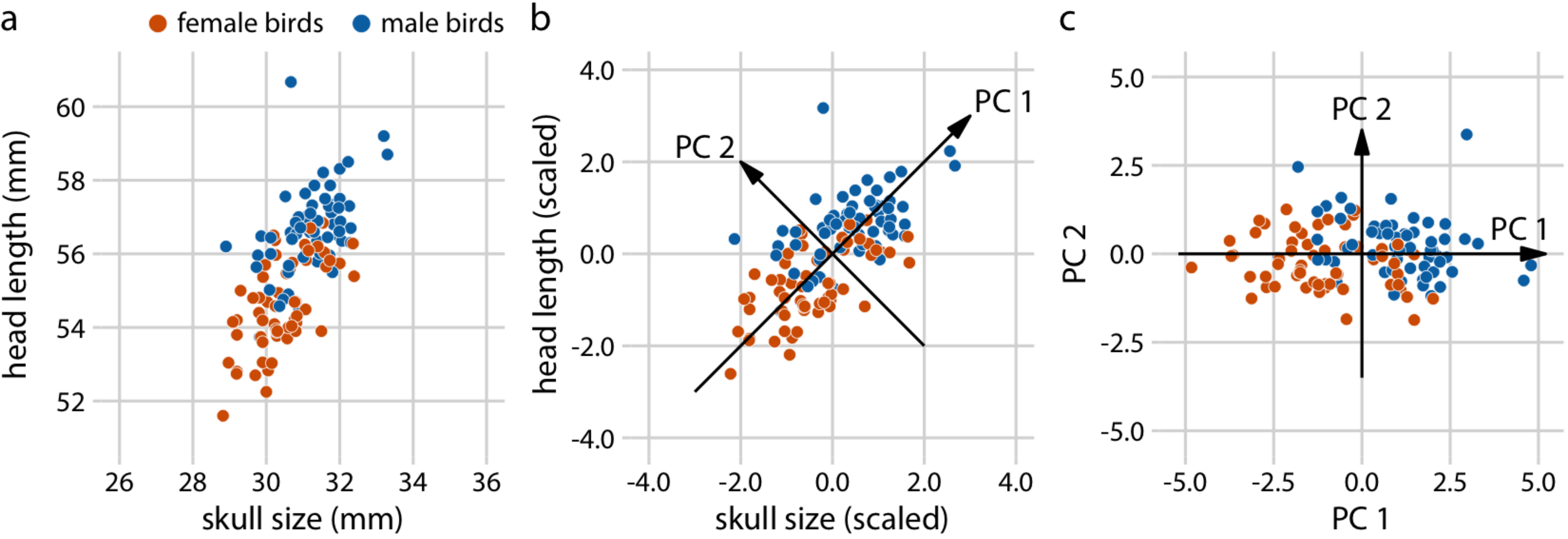
- Most high-dimensional datasets have multiple correlated variables that convey overlapping information.
- Such datasets can be reduced to a smaller number of key dimensions without loss of much critical information.
  - *dataset of multiple physical traits of people, including quantities such as each person's height and weight, the lengths of their arms and legs, the circumferences of their waist, hips, and chest, etc.*

Dimension Reduction-PCA

- Principal components analysis (PCA)
- PCA introduces a new set of variables, called principal components (PCs), by linear combination of the original variables in the data, standardized to zero mean and unit variance
- The PCs are uncorrelated, and ordered such that the first component captures the largest possible amount of variation and subsequent components capture increasingly less.
- Usually, key features in the data can be seen from only the first two or three PCs.

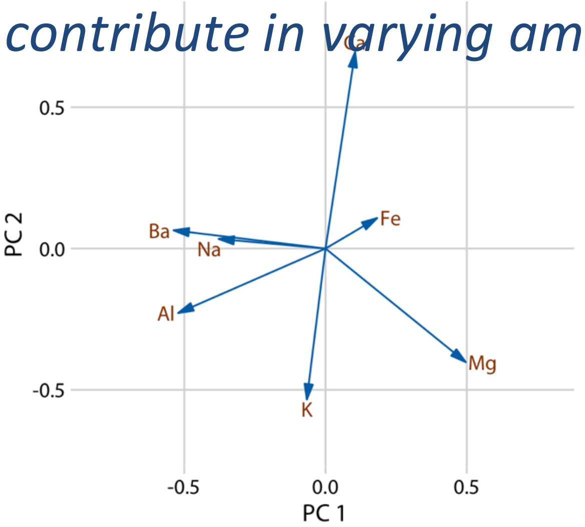
Dimension Reduction-PCA

a. The original data
b. Values scaled to zero mean and unit variance
c. Data projected into the new coordinates (equivalent to a rotation of the data points around the origin.)



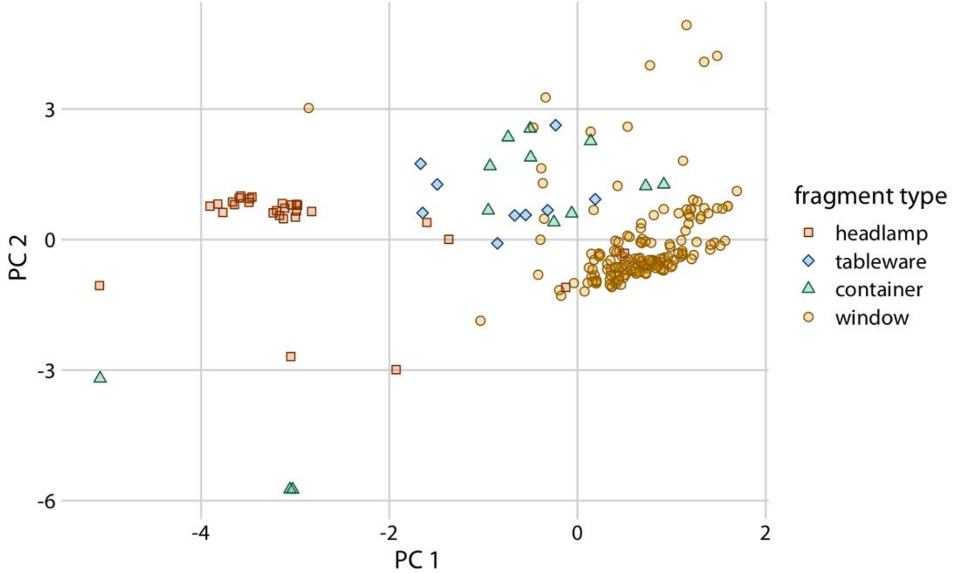
Dimension Reduction-PCA

- PCs are linear combinations of the original variables.
- We can represent the original variables as arrows indicating to what extent they contribute to the PCs.
  - *Barium and sodium contribute primarily to PC 1 and not to PC 2*
  - *Calcium and potassium contribute primarily to PC 2 and not to PC 1,*
  - *Other variables contribute in varying amounts to both components.*



Dimension Reduction-PCA

- We project the original data into the PC space
  - *We see a defined clustering of distinct types of glass fragments.*
  - *Fragments from both headlamps and windows fall into clearly delineated regions in the PC plot*
  - *Fragments from tableware and from containers are a little more spread out*

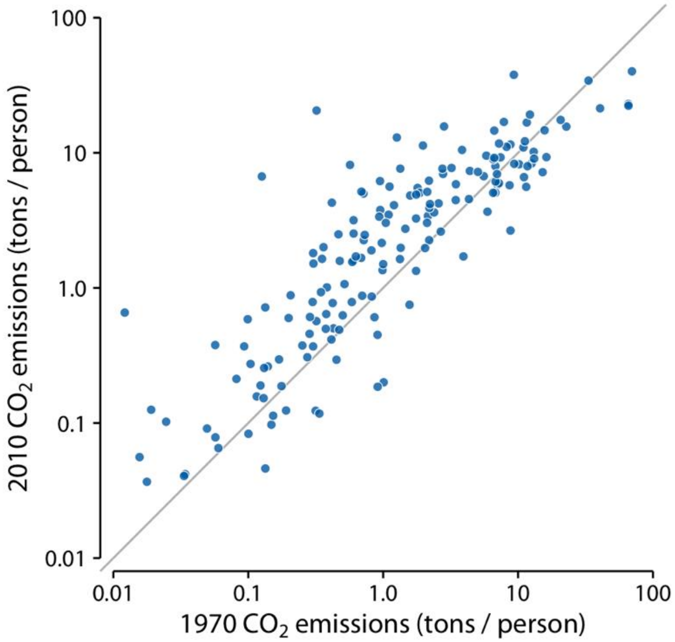


Paired Data

- A special case of multivariate quantitative data is paired data: data where there are two or more measurements of the same quantity under slightly different conditions.
  - *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)*
  - *measurements on two closely related subjects (e.g., the heights of two identical twins).*
- An excellent choice in this case is a simple scatterplot on top of a diagonal line marking x = y.

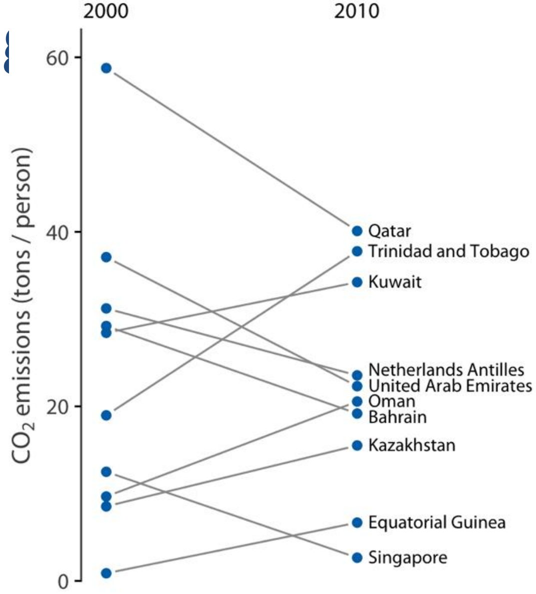
Paired Data

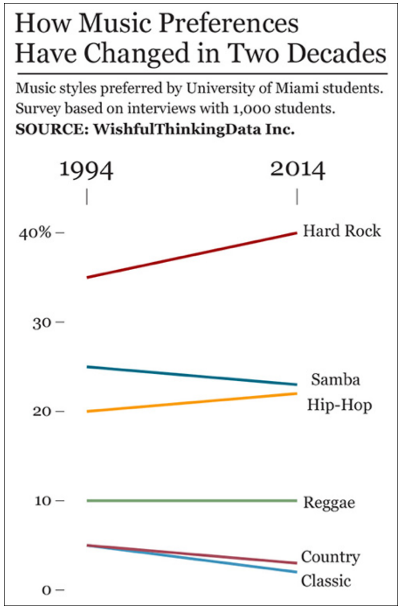
- The carbon dioxide (CO2) emissions per person, measured for 166 countries both in 1970 and in 2010
  - *Most points are relatively close to the diagonal line.*
  - *Points are systematically shifted upwards relative to the diagonal line.*
  - *The majority ase in CO2 emissions.*



slopegraph

- If we have only a small number of observations, we can draw individual measurements as dots arranged into two columns.
- Paired dots connected with a line.
- Slopes highlight the magnitude & direction of chan

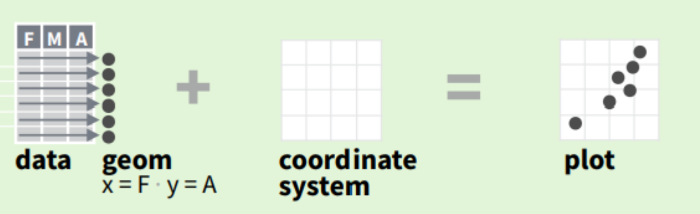



How Music Preferences Have Changed in Two Decades

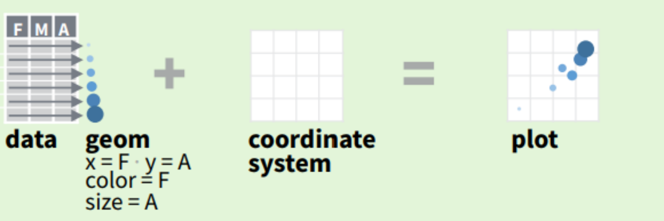
Music styles preferred by University of Miami students. Survey based on interviews with 1,000 students.
SOURCE: WishfulThinkingData Inc.

ggplot2

- ggplot2 is based on the grammar of graphics,
  - *the idea that you can build every graph from the same components: a data set, a coordinate system, and geoms—visual marks that represent data points*

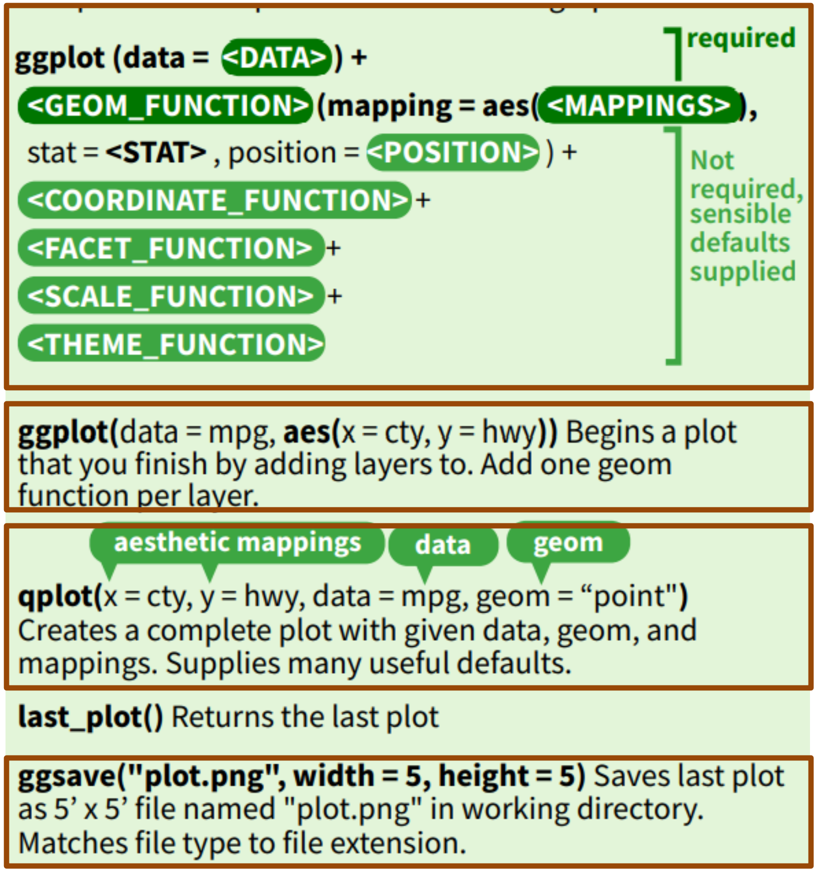


- To display values, map variables in the data to visual properties of the geom (aesthetics) like size, color, and x and y locations.



ggplot2

- Complete the template below to build a graph.



ggplot (data = <DATA>) +
<GEOM_FUNCTION>(mapping = aes(<MAPPINGS>),
stat = **<STAT>** , position = **<POSITION>**) +
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>

required

Not required, sensible defaults supplied

ggplot(data = mpg, **aes**(x = cty, y = hwy)**)** Begins a plot that you finish by adding layers to. Add one geom function per layer.

aesthetic mappings
data
geom

qplot(x = cty, y = hwy, data = mpg, geom = “point"**)** Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

last_plot() Returns the last plot

ggsave("plot.png", width = 5, height = 5) Saves last plot as 5’ x 5’ file named "plot.png" in working directory. Matches file type to file extension.