National Institutes of Health
Office of Management

Visualizing Relationships in ggplot

Doug Joubert
Pamela Katzen Burrows

2023-07-11

NIH Library | nihlibrary.nih.gov

Office of Research Services
National Institutes of Health
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

Class Description

- Provides a basic overview of the methods used to visualize the association among two or more quantitative variables
- Focus on scatterplots, scatterplot matrix, and visualizing paired data
- Participants are expected to have taken the Introduction to Data Visualization in R: ggplot class

Class Assumptions

- This class makes a few assumptions about your understanding of R and RStudio:
  - You have already installed R and RStudio
  - You have experience with R
  - You have experience working in RStudio and creating scripts and/or markdown files

Class Objectives: In Class

- Upon completion of this class students should be able to:
  - **Define bivariate data**
  - **Create a scatterplot using ggplot**
  - **Define linear regression**
  - **Demonstrate how to perform a simple linear regression in R**
  - Identify positive and negative associations from a scatter plot
  - Describe what Pearson’s correlation measures
  - State the possible range for Pearson’s correlation

Class Objectives: In Handout

- Upon completion of this class students should be able to:
  - Define bivariate data
  - Create a scatterplot using ggplot
  - Define linear regression
  - Demonstrate how to perform a simple linear regression in R
  - Identify positive and negative associations from a scatter plot
  - Describe what Pearson’s correlation measures
  - State the possible range for Pearson’s correlation

Resource: Theory

- Science-based approach:
  - Visual system
  - Cognition and perception
- 3rd edition is available electronically from the NIH Library
- Substantial changes in 4th Edition, and a completely new chapter

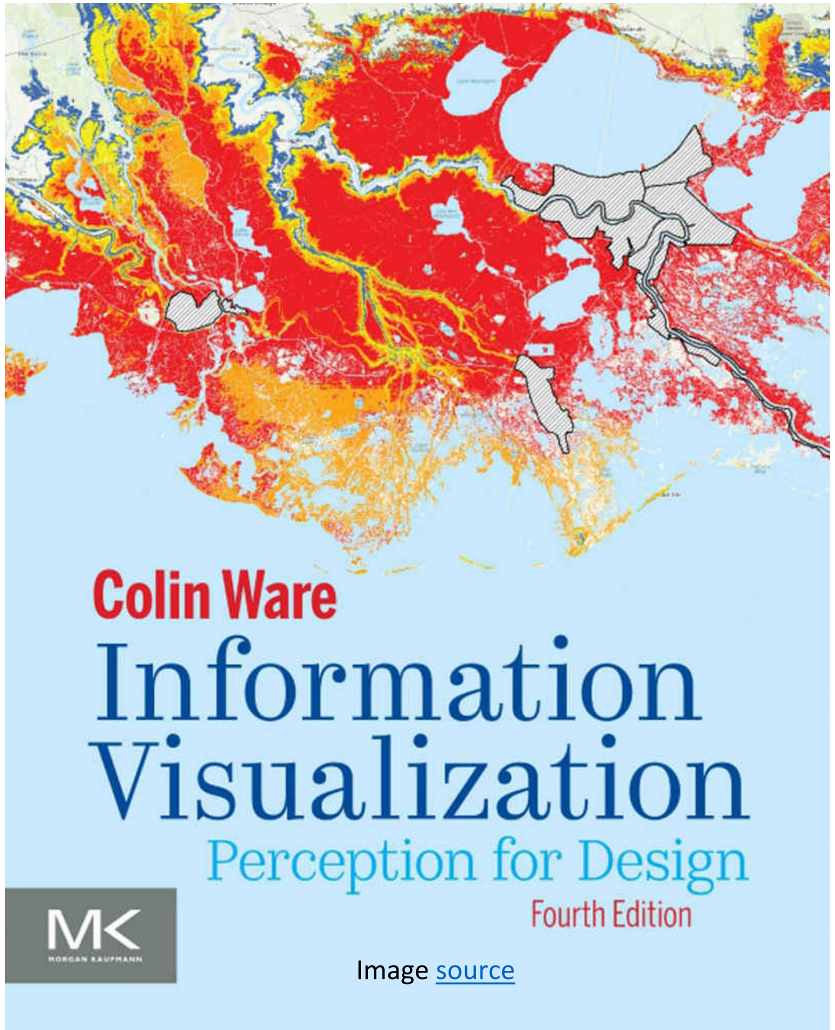
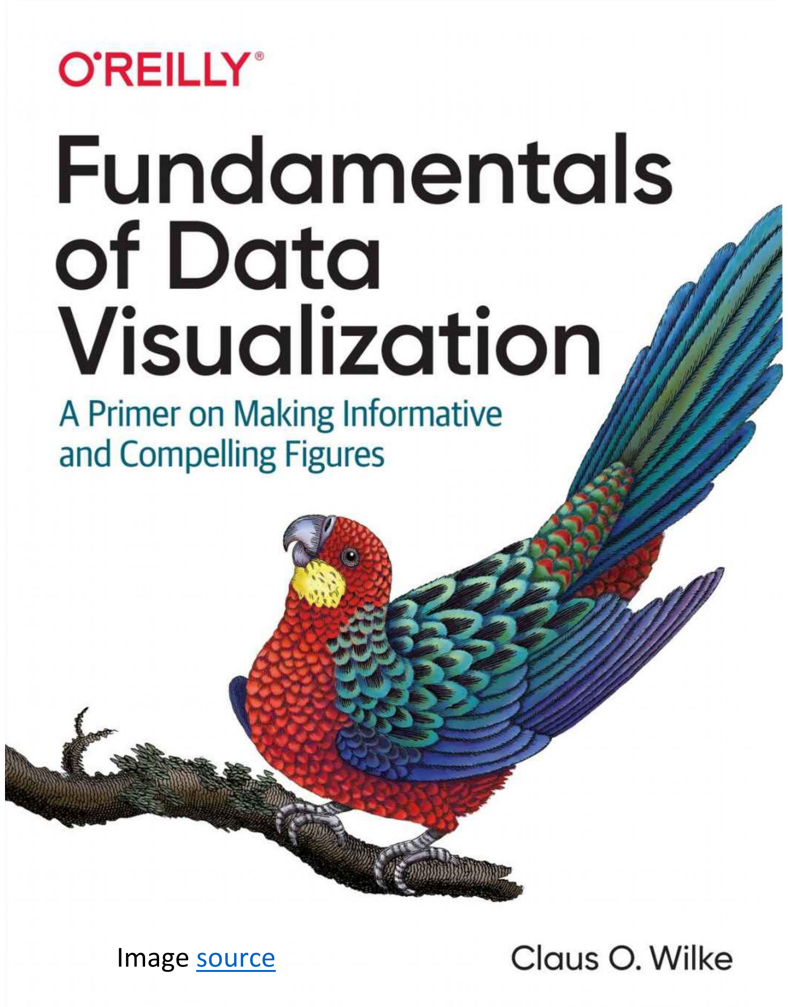


Image source

Resource: Theory/Practical

- Combines theory and practical application of design principles
- Code agnostics but a lot of the graphics were produced in R
- Thanks Claus, for making your book available online (for free)!!



O'REILLY

Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures

Claus O. Wilke

Image source

Resource: Theory/Practical

- Work-in-progress 3rd edition is available online for free
- Primary focus is explaining the Grammar of Graphics that ggplot2 uses
- Not a cookbook
- Will help you understand the details of the underlying theory

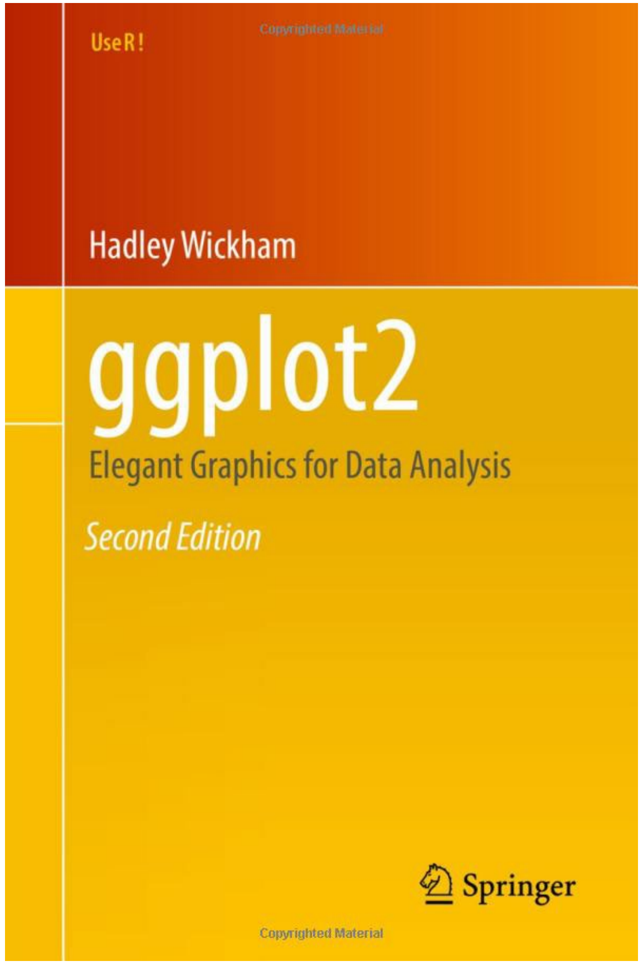


Image source

Configuration for Exercises

R and RStudio

- R is a programming language that is especially powerful for data exploration, visualization
- RStudio is an integrated development environment (IDE) that makes using R easier
- R and RStudio are two separate pieces of software
- **Must install R before you install RStudio**

R and RStudio: Windows

1. Download R from the CRAN website
2. Run the .exe file that was just downloaded

R and RStudio: Windows

1. Go to the RStudio download page
2. Under Installers select RStudio x.yy.zzz - Windows Vista/7/8/10 (where x, y, and z represent version numbers)
3. Double click the file to install it

R and RStudio: Mac

1. Download R from the CRAN website
2. Select the .pkg file for the latest R version
3. Double click on the downloaded file to install R
4. It is also a good idea to install XQuartz (needed by some packages)

R and RStudio: Mac

1. Go to the RStudio download page
2. Under Installers select RStudio x.yy.zzz - Mac OS X 10.6+ (64-bit) (where x, y, and z represent version numbers)
3. Double click the file to install RStudio

Configuration for Exercises

- GGally adds several functions to reduce the complexity of combining geoms with transformed data
- OpenIntro package includes supplemental functions and data for open-source textbooks and resources for introductory statistics

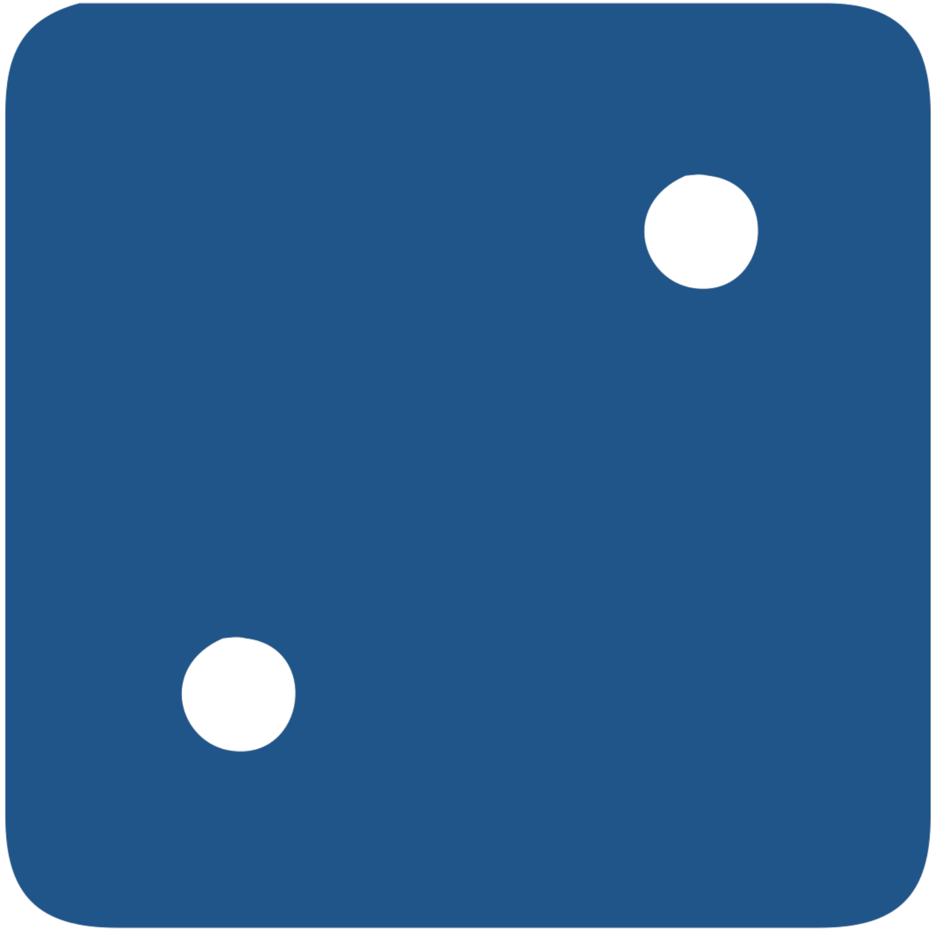
Configuration for Exercises

- Tidyverse: collection of R packages designed for data science

Scatterplots and Correlation

Bivariate Data

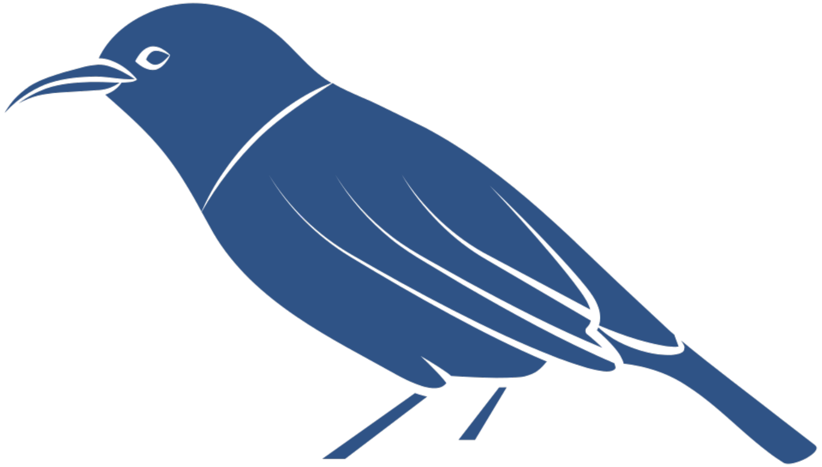
- Measures of central tendency, variability, and spread summarize a single variable
- Often, more than one variable is collected on each individual in a study
- Two quantitative variables for each individual
- Example: relationship between the height and weight



Lane, D. (2007)

Data: Blue Jay

- Blue jay data
  - 123 rows of data
  - Head length
  - Skull size
  - Body mass of each bird
- Import data into an object called `blue_jays`



Blue Jay Data: Histogram (Body Mass)

- Create a histogram of body_mass_g
- What can we say about this distribution?
- Add your thoughts to the Google Doc

Blue Jay Data: Histogram (Body Mass)

```
bj_body_mass_hist <- blue_jays %>%
  ggplot(mapping = aes(x = body_mass_g)) +
  geom_histogram(color = "black", fill = "white") +
  geom_vline(mapping = aes(xintercept = mean(body_mass_g,
na.rm = TRUE)), color = "red", linetype = "dashed", size =
1)


bj_body_mass_hist
```

Blue Jay Data: Histogram (Head Length)

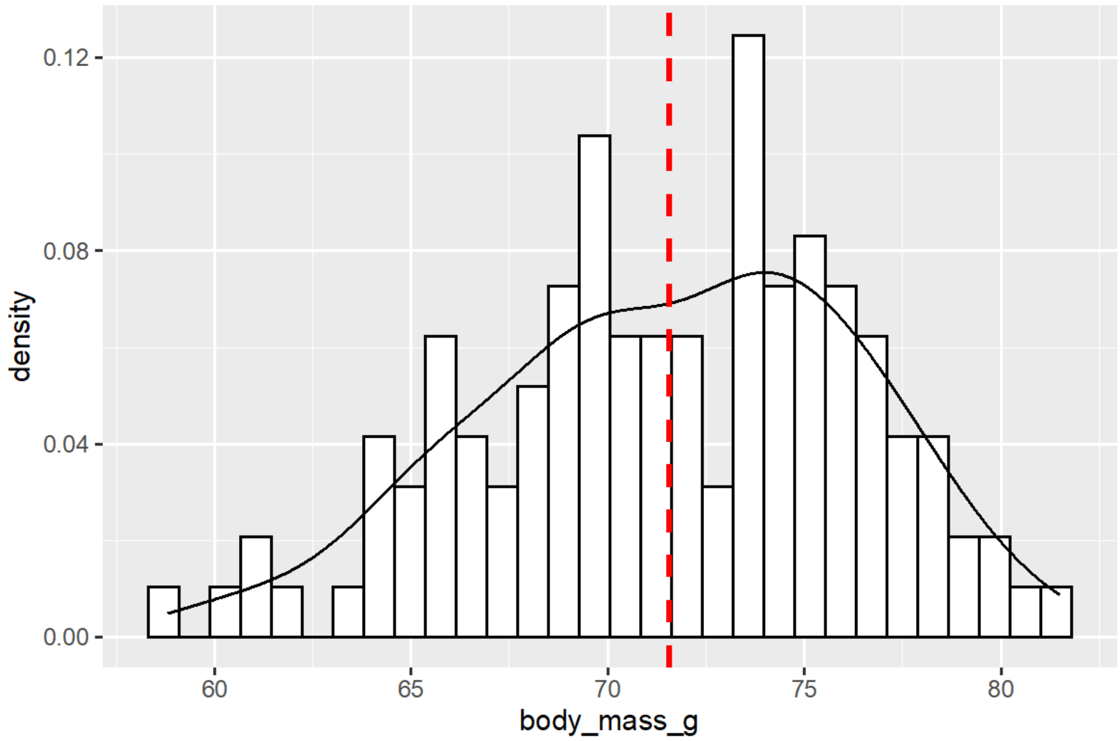
- Create a histogram of head_length_mm
- What can we say about this distribution?
- Add your thoughts to the Google Doc

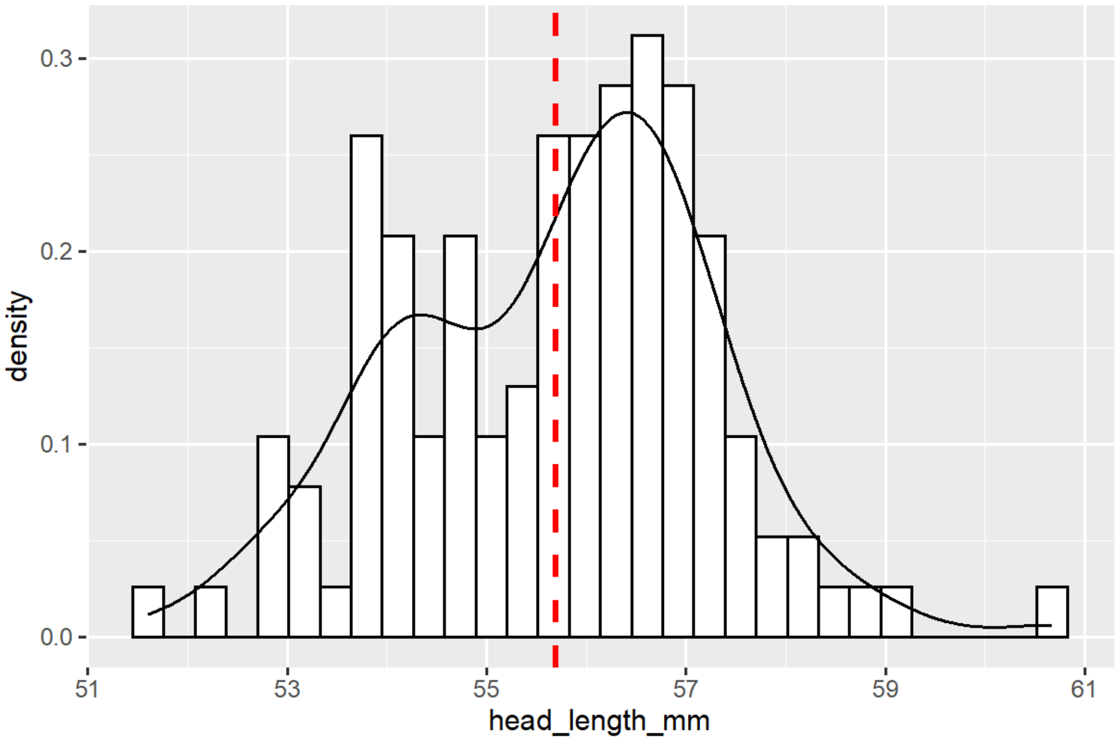
Blue Jay Data: Histogram (Head Length)

```
bj_head_length_hist <- blue_jays %>%
  ggplot(mapping = aes(x = head_length_mm)) +
  geom_histogram(color = "black", fill = "white") +
  geom_vline(mapping = aes(xintercept = mean(head_length_mm,
na.rm = TRUE)), color = "red", linetype = "dashed", size = 1)

bj_head_length_hist
```

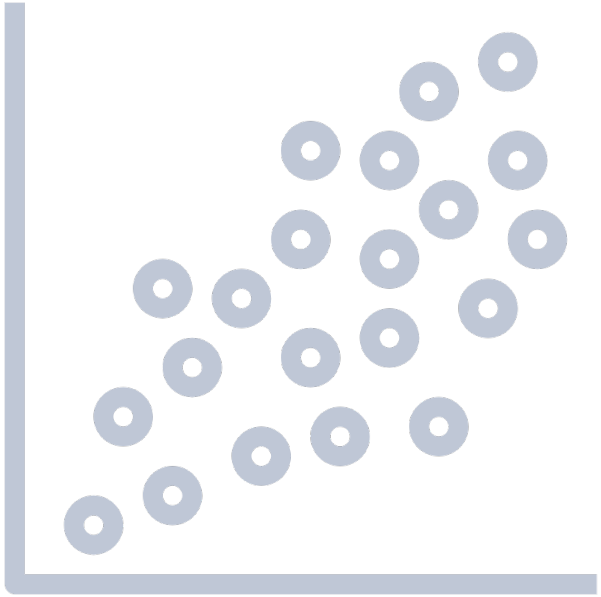
What if We Wanted to Compare the Data?





Scatterplots

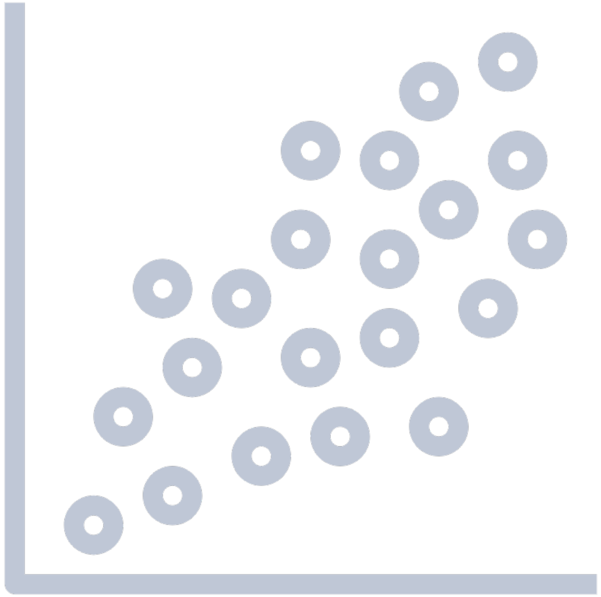
- Plot of paired (x, y) quantitative data
- Horizontal axis is used for the first variable (x)
- Vertical axis is used for the second variable (y)
- Implemented in ggplot using the `geom_point()`



Triola & Lossi, (2018)

Scatterplots: Attributes

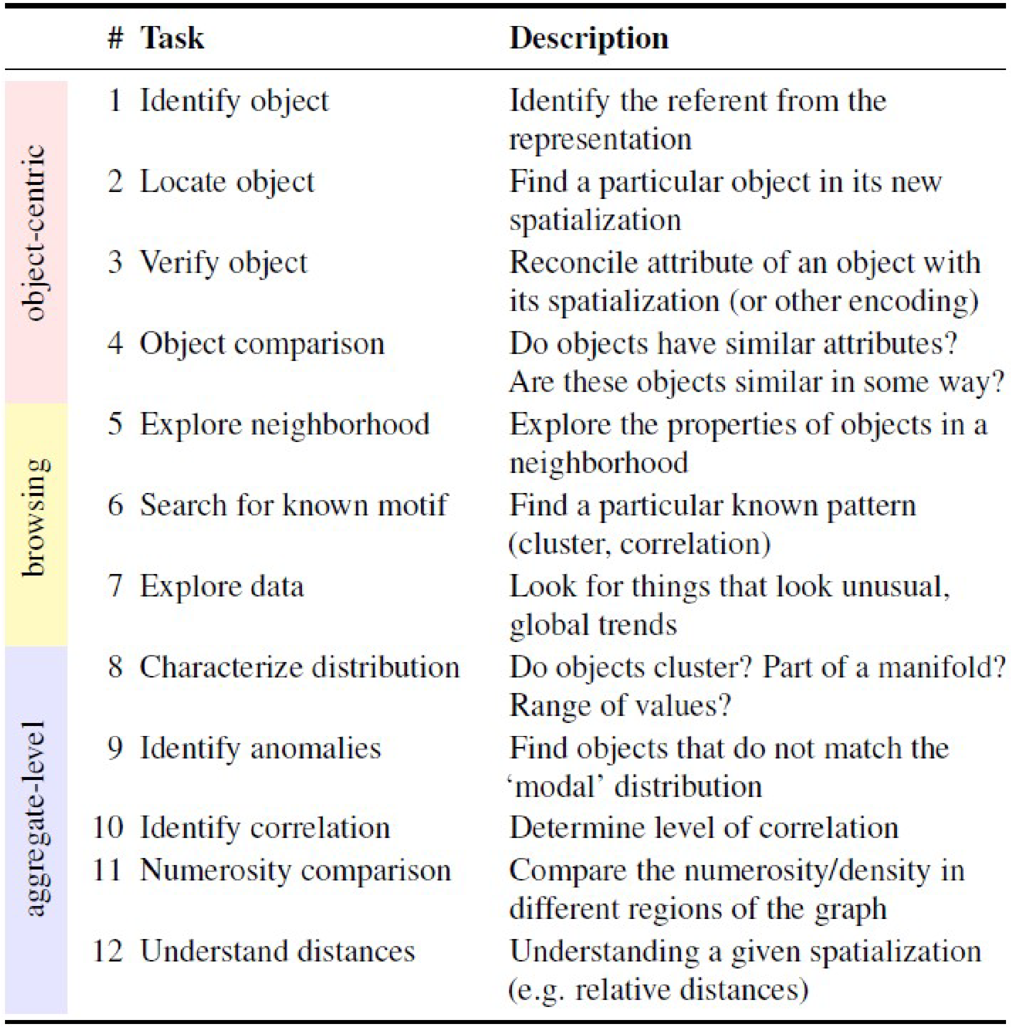
- Designed to emphasize the spatial distribution of data plotted in two-dimensions:
  - Marks or points designed with preattentive features
  - Designed with the detection of individual objects
  - Distances between objects represent a notion of similarity



Cleveland, W. S., & McGill, R. (1984)

Scatterplots: Tasks

- Abstracted analysis tasks that are performed with scatterplots

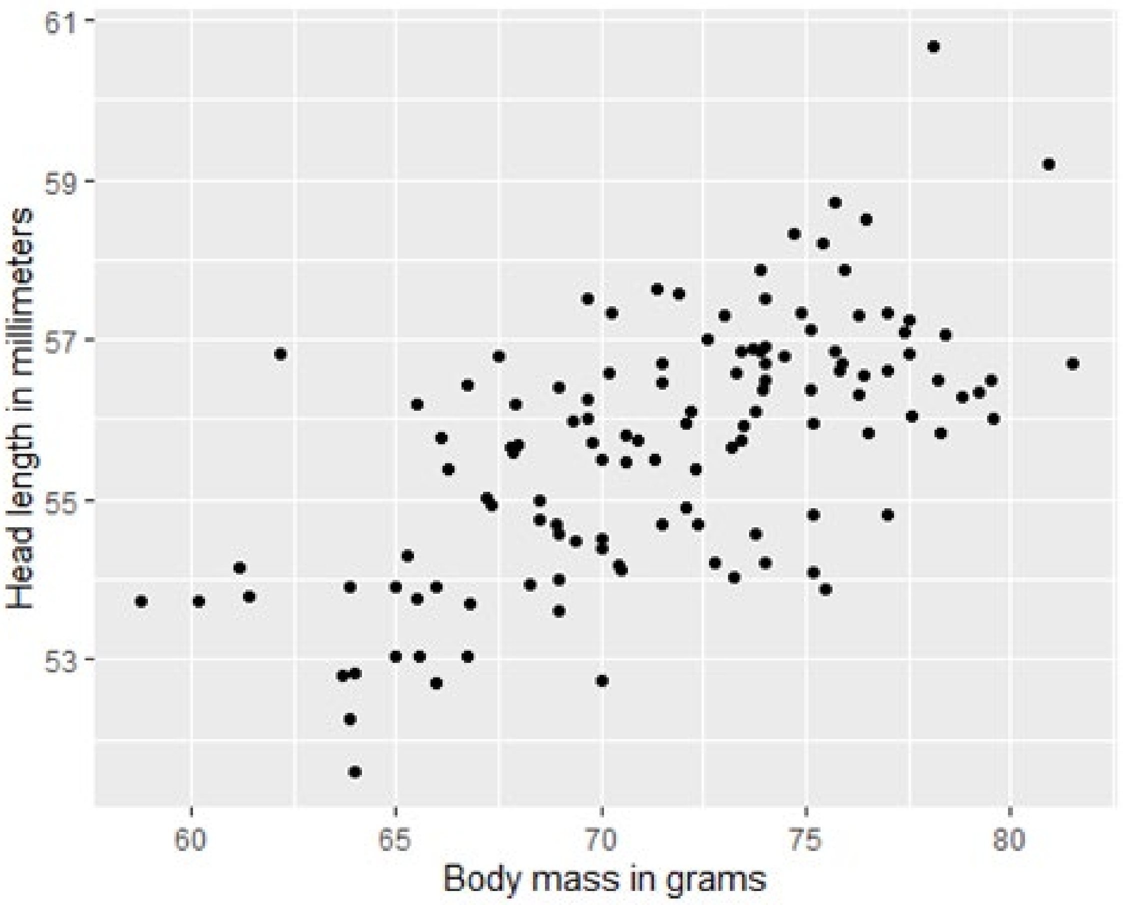


	#	Task	Description
object-centric	1	Identify object	Identify the referent from the representation
	2	Locate object	Find a particular object in its new spatialization
	3	Verify object	Reconcile attribute of an object with its spatialization (or other encoding)
	4	Object comparison	Do objects have similar attributes? Are these objects similar in some way?
browsing	5	Explore neighborhood	Explore the properties of objects in a neighborhood
	6	Search for known motif	Find a particular known pattern (cluster, correlation)
	7	Explore data	Look for things that look unusual, global trends
aggregate-level	8	Characterize distribution	Do objects cluster? Part of a manifold? Range of values?
	9	Identify anomalies	Find objects that do not match the ‘modal’ distribution
	10	Identify correlation	Determine level of correlation
	11	Numerosity comparison	Compare the numerosity/density in different regions of the graph
	12	Understand distances	Understanding a given spatialization (e.g. relative distances)

Sarikaya, A., & Gleicher, M. (2018)

Scatterplots Example: Blue Jay Data 1

- Head length on y-axis and body mass on x axis
- We “say” that we plot the variable shown along the y-axis against the variable shown along the x-axis.
- So, what does this tell us?

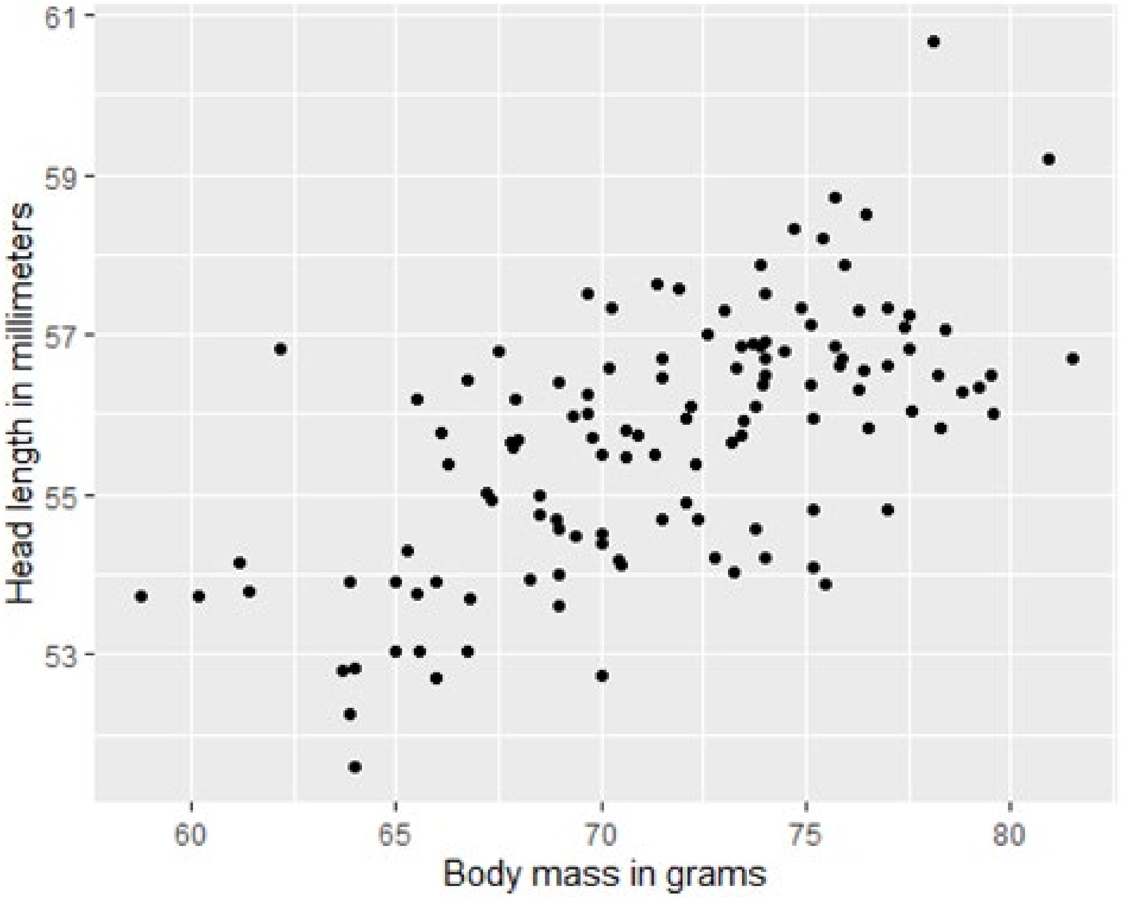


Blue Jay Data 1: Code

```
blue_jays %>%
  ggplot(mapping = aes(x = body_mass_g, y = head_length_mm)) +
  geom_point(size = 1.5) +
  labs(y = "Head length in millimeters",
       x = "Body mass in grams")
```

Scatterplots Example: Blue Jay Data 1

- Moderate tendency for heavier birds to have longer heads
- How can we also look at the "sex" of each bird?

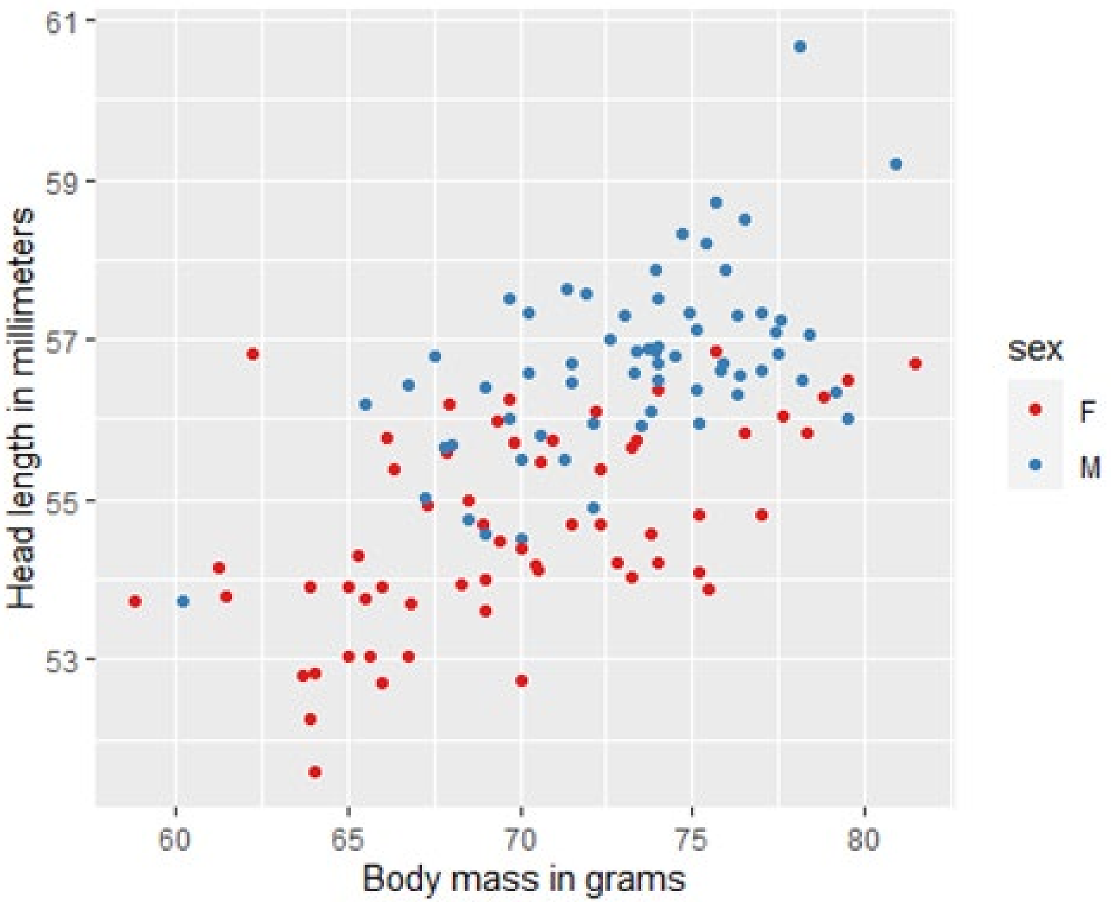


Scatterplots Example: Blue Jay Data 2

- geom_point() aesthetics:
  - x
  - y
  - alpha
  - color
  - fill
  - group
  - shape
  - size
  - stroke

Blue Jay Data 2: Interpretation

- Birds’ sex is indicated by color
- Overall trend in head length and body mass is at least in part driven by the sex of the birds
- Meaning that at the same body mass, male birds tend to have longer heads than female birds.



Blue Jay Data 2: Code

```
blue_jays %>%
  ggplot(mapping = aes(x = body_mass_g, y = head_length_mm,
color = sex)) +
  geom_point(size = 1.5) +
  scale_color_brewer(palette="Set1") +
  labs(y = "Head length in millimeters",
       x = "Body mass in grams")
```

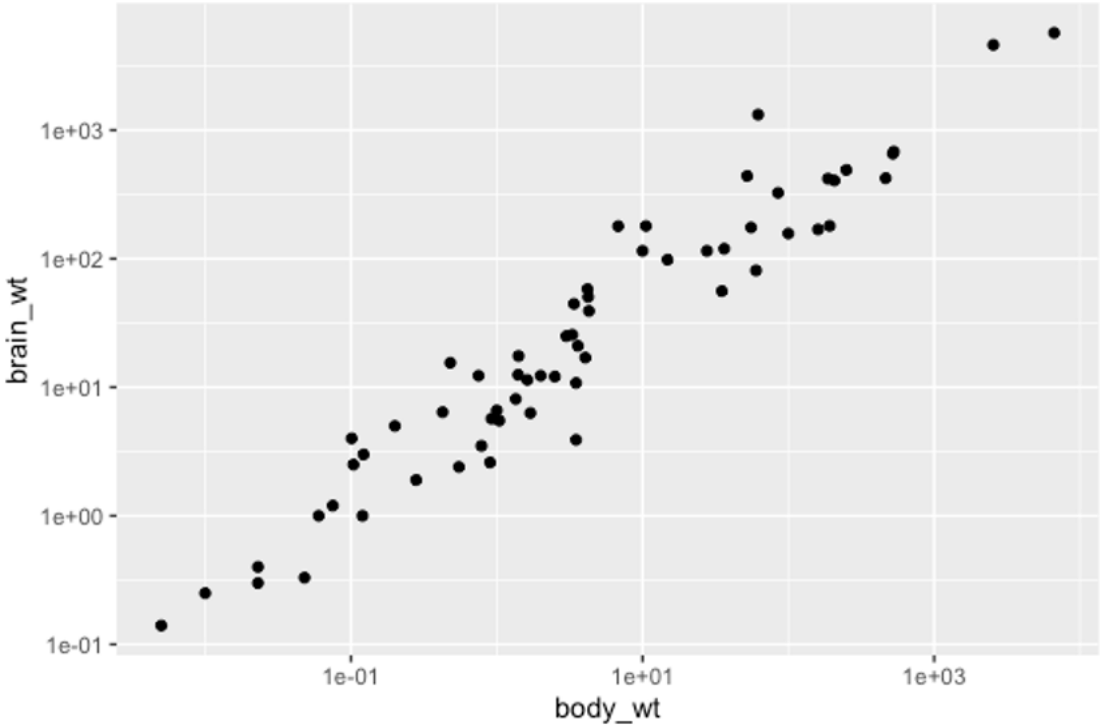
Scatterplots: Transformations

- Sometimes transformation is necessary
- Relationship between two variables may not be linear
- Sometimes there is no meaningful relationship between the two variables

Wilke, C. (2019)

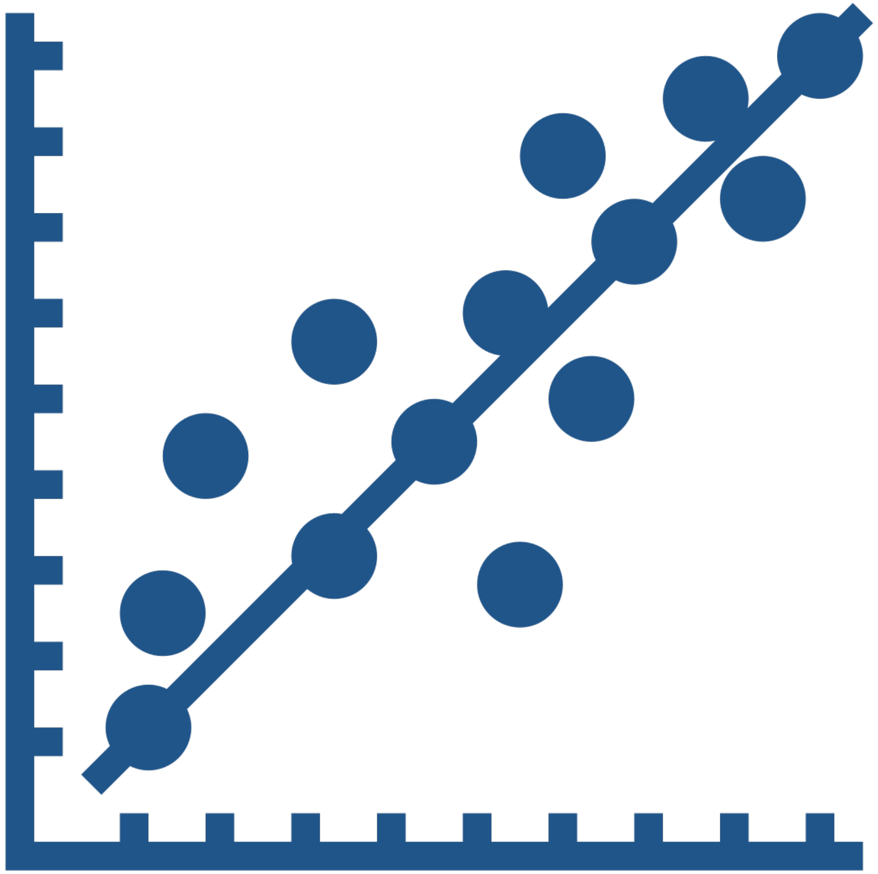
Scatterplots: Transformations Example

- ggplot has several methods for transforming a plot:
  - coord_trans() transforms the coordinates of the plot
- scale_x_log10() and scale_y_log10() perform a base-10 log transformation of each axis
- This graph uses base-10 log transformation



Correlations

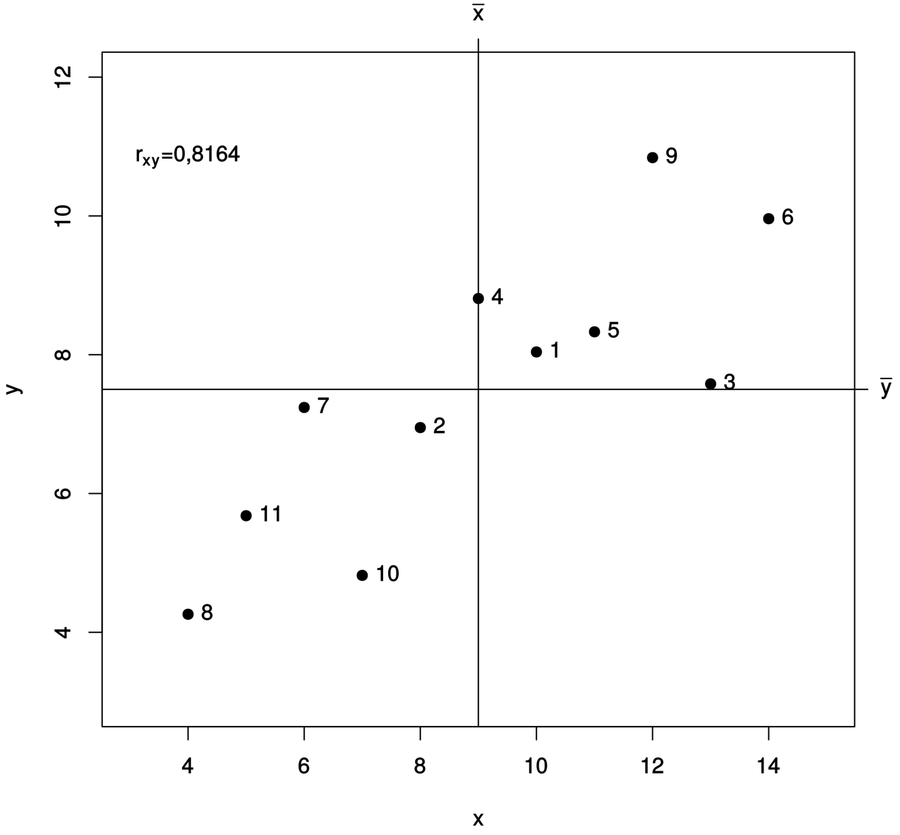
- Method for quantifying the strength of bivariate relationships
- Exists when the values of one variable are somehow associated with values of the other variable
- Correlation between two variables is not evidence that one of the variables causes the other



Triola & Lossi. (2018)

Pearson Correlation

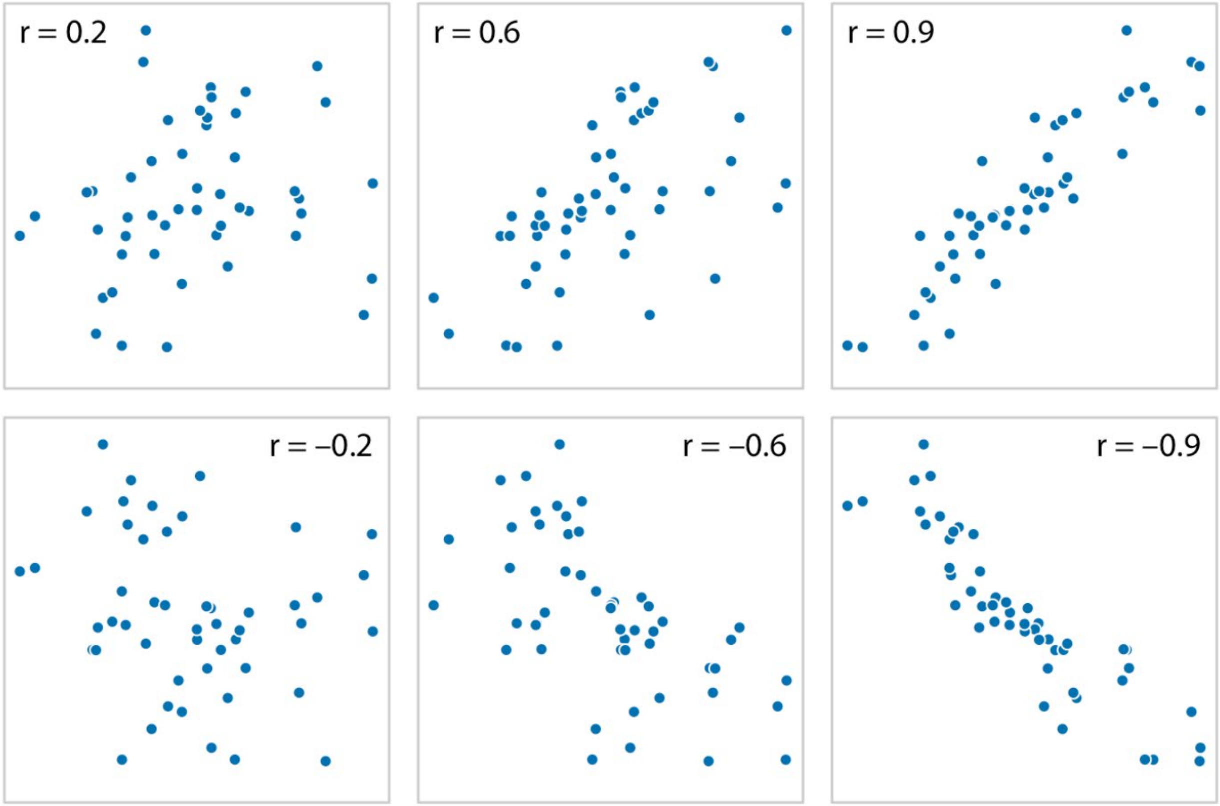
- Measure of the strength of the linear relationship between two variables
- If relationship is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables
- Correlation says nothing about how much Y changes when X changes



Triola & Lossi. (2018)

Correlations: Interpreting

Randomly generated sets of points to illustrate different correlations, in both rows, from left to right correlations go from weak to strong



r	Rough meaning
± 0.1–0.3	Modest
± 0.3–0.5	Moderate
± 0.5–0.8	Strong
± 0.8–0.9	Very strong

Wilke, C. (2019) & Heiss, A. (2021)

Correlations in R: cor()

- Compute the Pearson correlation
- Very conservative when it encounters missing data (e.g., NAs)
- use argument allows you to override the default behavior of returning NA

```
cor(x = blue_jays$body_mass_g,
y = blue_jays$head_length_mm,
use ="pairwise.complete.obs",
method = "pearson")
[1] 0.6294447
```

Using the guide from Heiss, this is a strong positive correlation

Correlations in R: cor.test()

- Provides access to the values returned by the correlation
- Returns:
  - p.value: the p-value of the test
  - estimate: the correlation coefficient

```
cor.test(blue_jays$body_mass_g,

blue_jays$head_length_mm)
    Pearson's product-moment correlation
data:  blue_jays$body_mass_g and blue_jays$he
ad_length_mm
t = 8.9105, df = 121, p-value = 6.302e-15
alternative hypothesis: true correlation is n
ot equal to 0
95 percent confidence interval:
 0.5091462, 0.7256207
sample estimates: cor
```

Linear Relationships and Regression

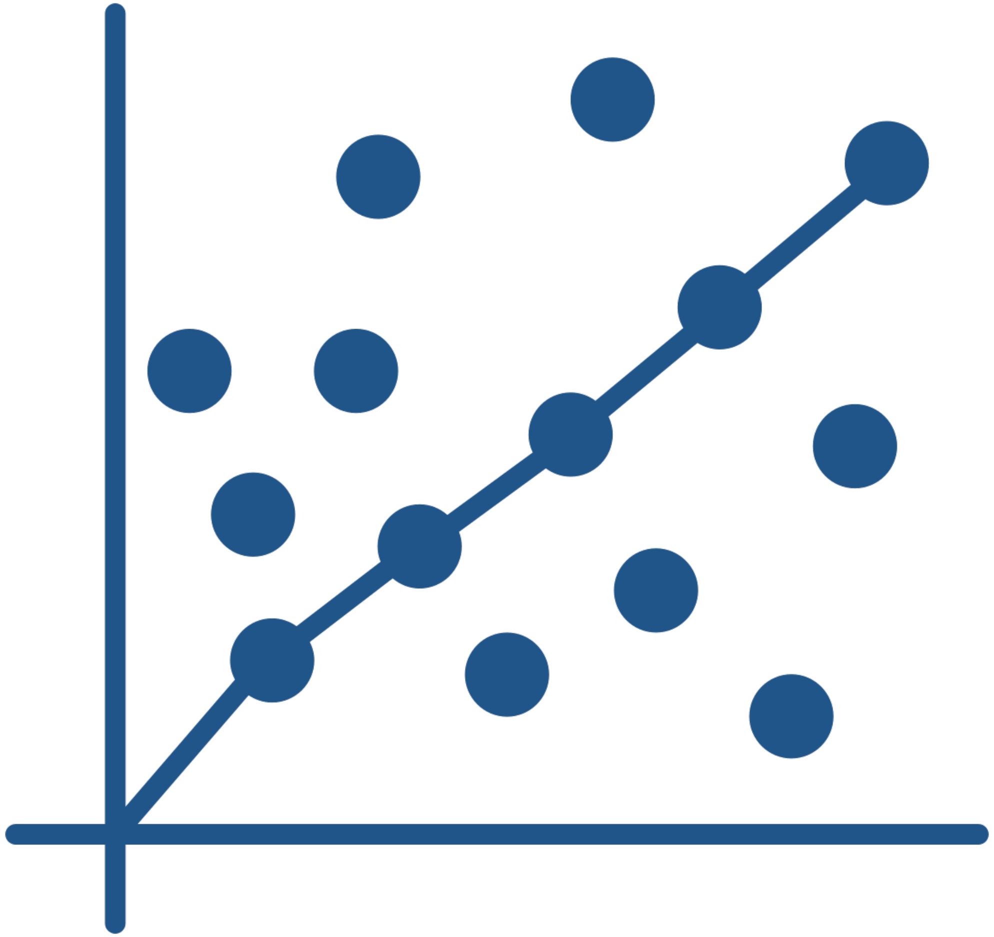
Simple Linear Regression

- Scatterplots are a common method for visualizing the relationship between two numeric variables
- Simple linear regression can be visualized on a scatterplot by a straight line
- “Best fit” line cuts through the data in a way that minimizes the distance between the line and the data points
- We will define “best-fitting line”

Lane, D. (2007)

Simple Linear Regression

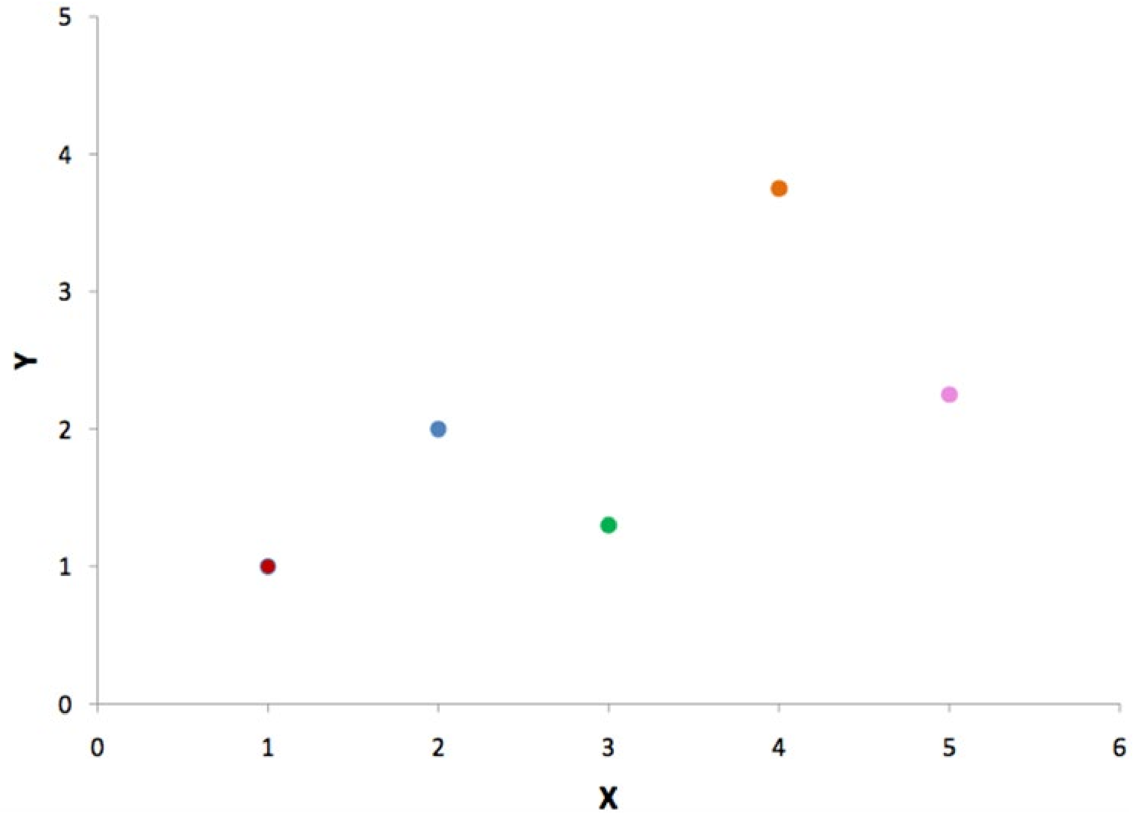
- Predict values on one variable from the values on a second variable:
  - Variable we are predicting is Y
  - Variable we are basing our predictions on is X
  - When there is only one predictor variable, the prediction method is called simple regression



Lane, D. (2007)

Simple Linear Regression

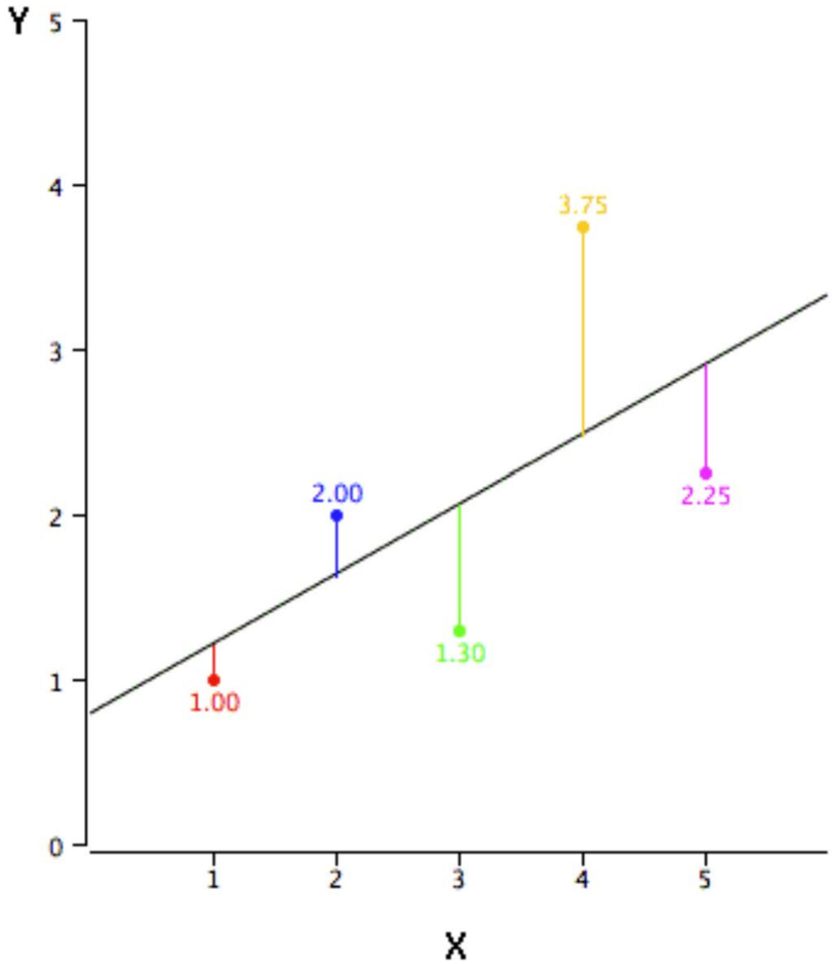
X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25



Lane, D. (2007)

Regression Line

- Vertical lines from the points to the regression line represent the errors of prediction
  - Red point is very near the regression line; its error of prediction is small
  - Yellow point is much higher than the regression line and therefore its error of prediction is large.

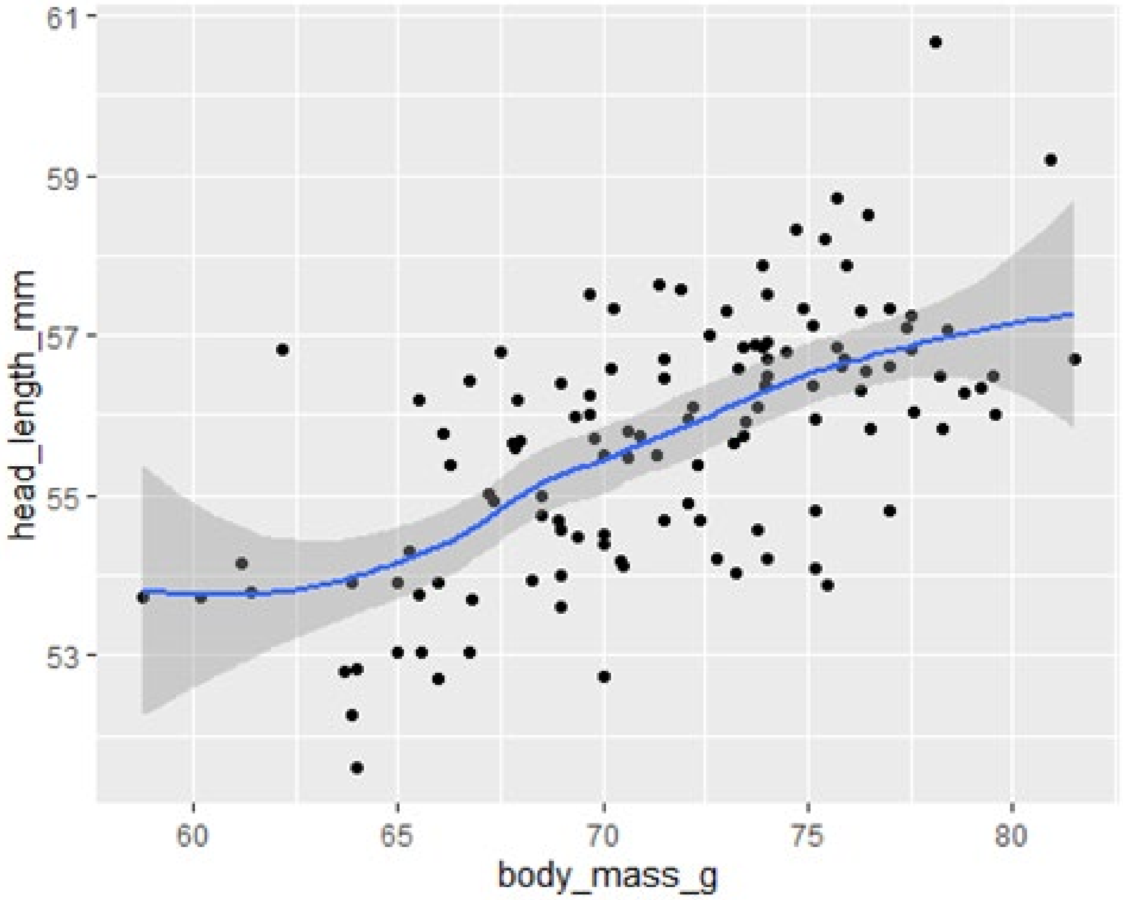


Lane, D. (2007)

Simple Linear Regression: Example 1

```
blue_jays %>%
  ggplot(mapping = aes(x = body_mass_g,
y = head_length_mm)) +
  geom_point(size = 1.5) +
  geom_smooth()
```

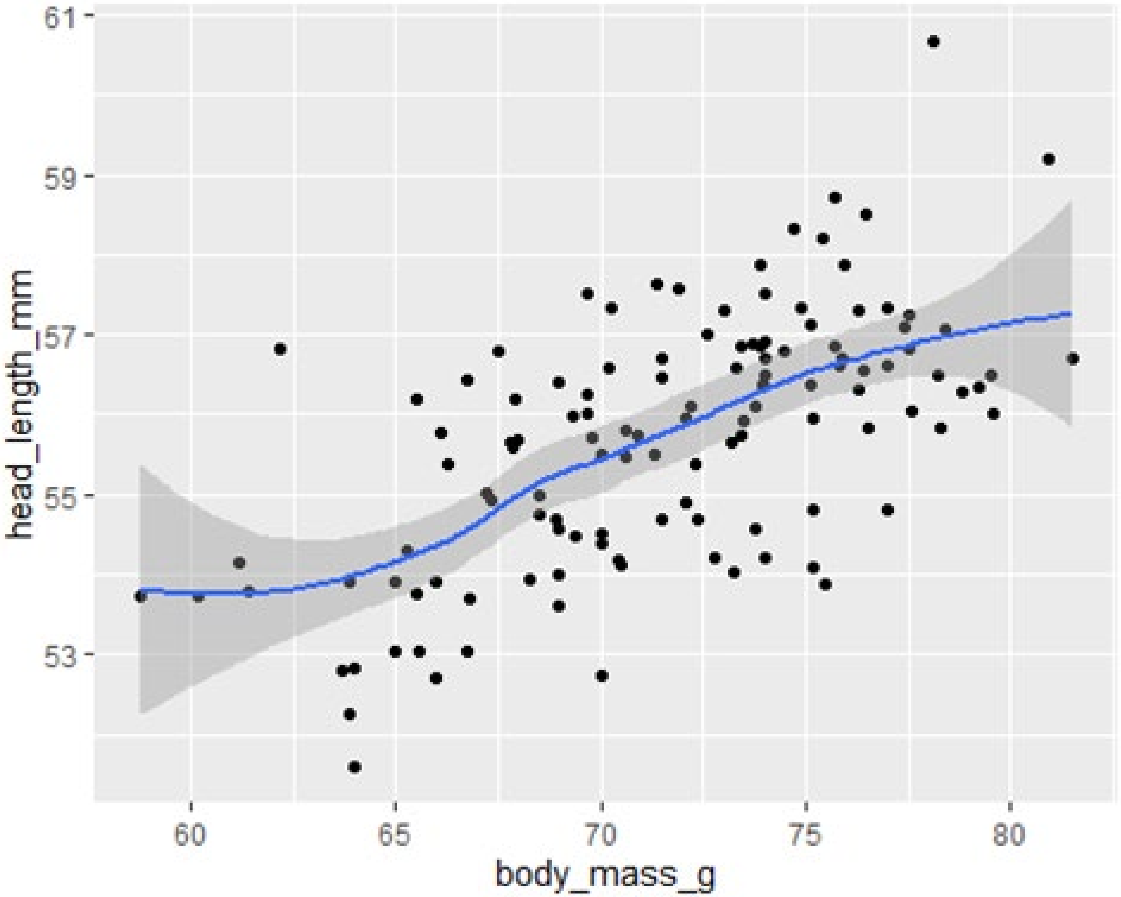
```
## `geom_smooth()` using method = 'loes
s' and formula 'y ~ x'
```



Wilke, C. (2019)

Example 1: Interpretation

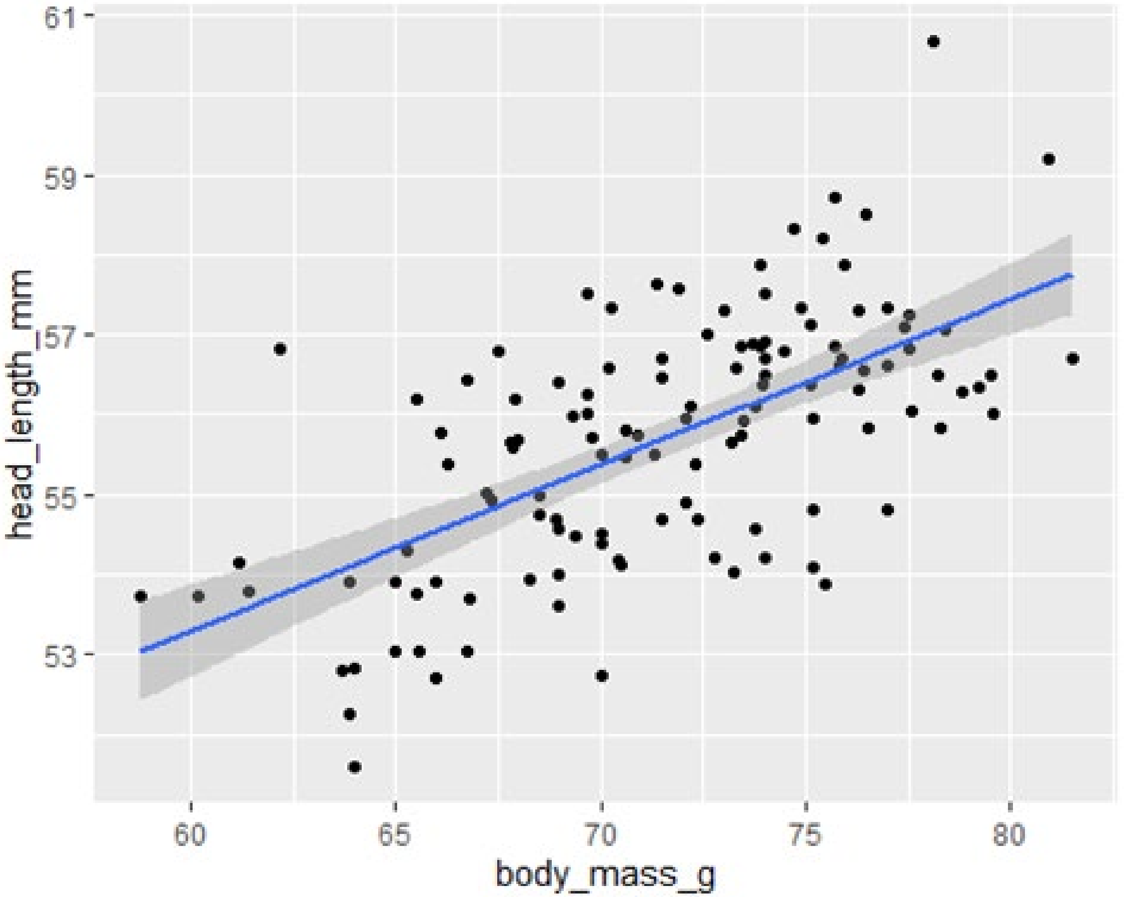
- Head length versus body mass (in grams), for 123 blue jays
- Each dot corresponds to one bird
- `geom_smooth` overlays the scatterplot with a smooth curve
- Confidence intervals (CI) shown in grey
- CI turned off: `geom_smooth(se = FALSE)`



Wilke, C. (2019); Wickham & Sievert. (2016)

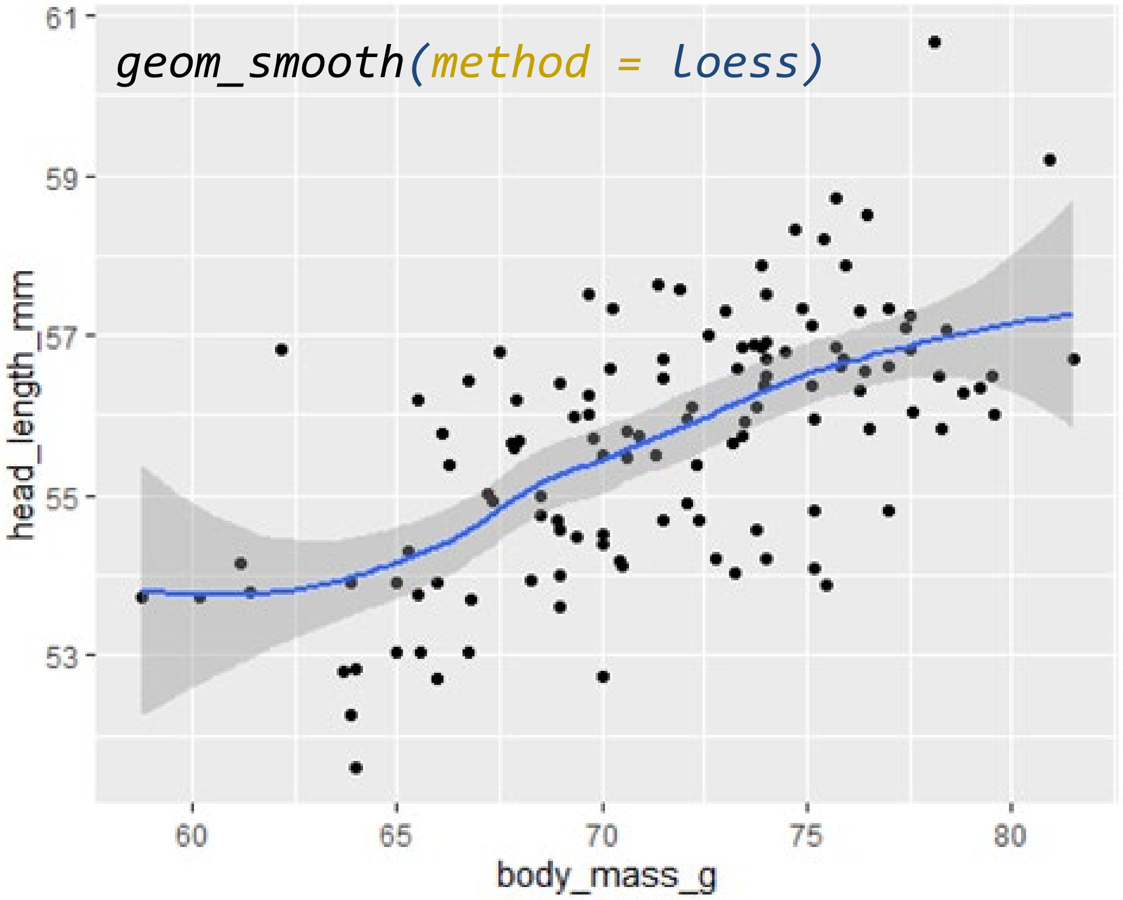
Simple Linear Regression: Modifications

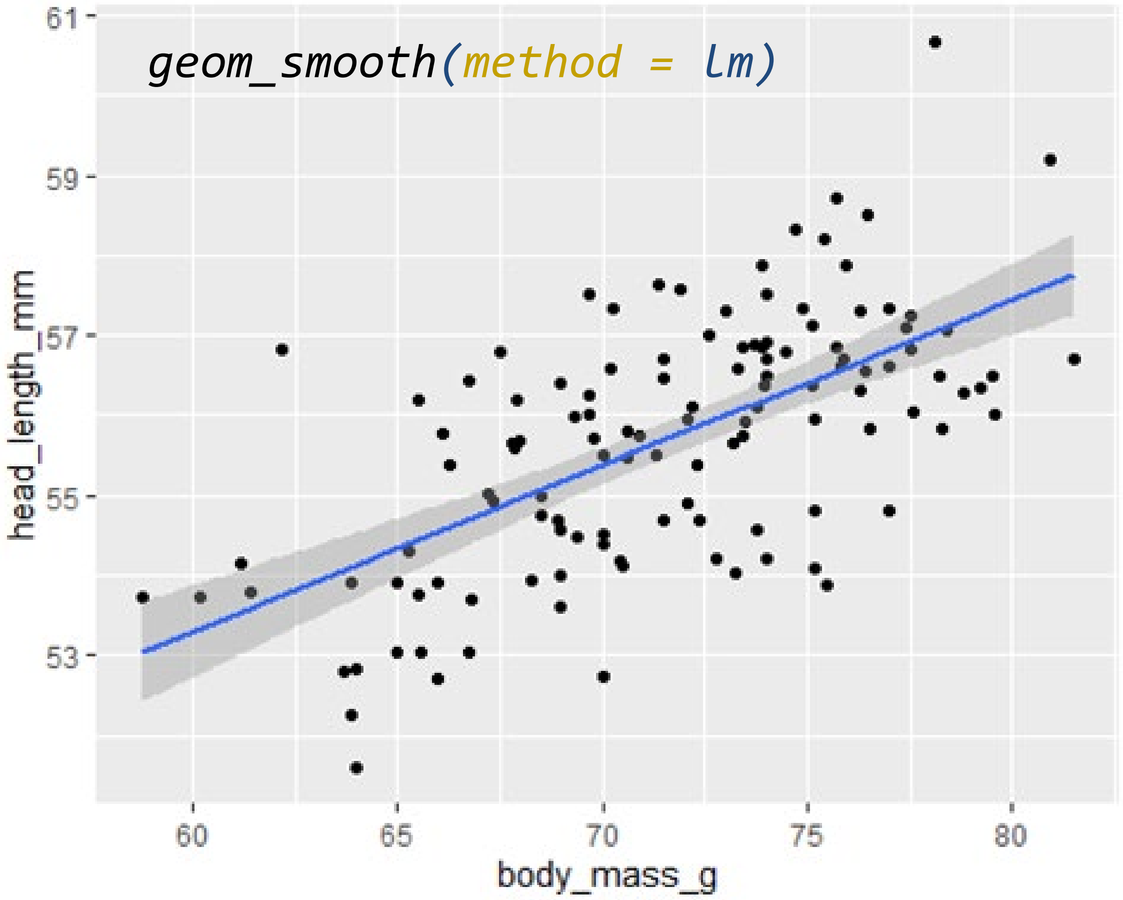
- `geom_smooth(method = ) for` choosing type of model:
  - `method = "loess,"` the default for small n
  - Loess does not work well for large datasets
  - n is greater than 1,000. `method = "lm"` fits a linear model, gives the line of best fit



Wilke, C. (2019); Wickham & Sievert. (2016)

Simple Linear Regression: Comparison





Wilke, C. (2019); Wickham & Sievert. (2016)

Fitting Simple Linear Models: `lm()`

- `geom_smooth(method = “lm”)` is useful for drawing linear models on a scatterplot
  - However, it does not return the characteristics of the model
- `lm()` function takes two arguments:
  - Formula that specifies the model
  - Data argument for the data frame
  - StatQuest has a great overview of performing this in R

```r
fit_blue_jays <- lm(data = blue_jays, head_length_mm ~ body_mass_g)
# Saves the output from the lm() function
```

lm(): Viewing Output

- summary() function includes:
  - Standard error
  - p-value for each coefficient
  - R2, and adjusted R2
  - Residual standard error
- Handout has information on interpreting these values

`lm()`: Residuals

- Distance from the data to the fitted line
- Should be symmetrically distributed around the line (which is equal to 0)
- Want the min value and max value to be approximately the same distance from 0
- Likewise, you would like the 1Q and 3Q to be equidistant from 0

```
summary(fit_blue_jays)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6308 -0.9115  0.0271  0.7307  3.6204 
```

`lm()`: Viewing Output

- The `coef()` function displays only the values of the coefficients

```
coef(fit_blue_jays)

(Intercept) body_mass_g
 40.8621453   0.2072662
```

lm(): Coefficients - Intercept

- Intercept is 40.86215
- Indicates that the head length of Blue Jays was 40.8625 millimeters when the body weight was 0 grams
- Please note that the importance or relevance of the intercept value is dependent on the nature of the biological systems which are being examined

```
Coefficients:


Estimate Std.

(Intercept) 40.86215
body_mass_g  0.20727
```

`lm()`: Coefficients - Slope

- Slope is 0.20727
- Slope indicates the change in Y (or dependent variable) for every one unit increase in X (or independent variable)
- Slope value (0.20727) indicates that the head length of Blue Jays increased 0.20727 millimeters per every 1 gram increase in body weight

```
Coefficients:


Estimate Std.
(Intercept) 40.86215
body_mass_g  0.20727
```

lm(): Coefficients – Regression Equation

- Estimates can be used to write the following regression equation: head_length_mm = (0.20727 x body_mass_g) + 40.86215
- The structure of a regression equation is often times described as: y = mx + b. Such that:
  - y = response variable (body mass (grams))
  - m = slope x = independent variable (head length (mm))
  - b = intercept

`lm()`: Coefficients – R^2

- Multiple R-squared `head_length_mm` can explain 40% of the variation in `body_mass_g`
- Adjusted R-squared is the R-squared scaled by the number of parameters in the model

```
Multiple R-squared:  0.3962,
Adjusted R-squared:  0.3912
```

`lm()`: Coefficients – p-value

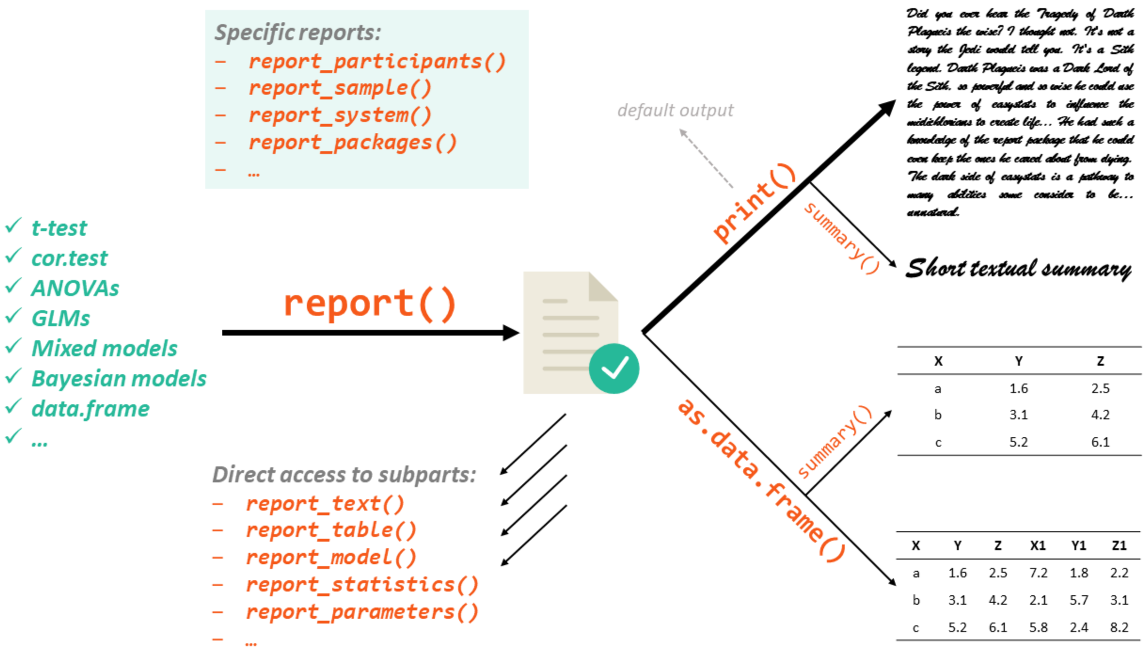
- P = 6.302e-15
- Based on this value and an established alpha level of 0.05, we can conclude that there is a significant effect of body mass (in grams) on the head length (in mm) of Blue Jay birds

```
F-statistic:
79.4 on 1 and 121 DF,
p-value: 6.302e-15
```

Data Reporting with Report Package

Report Package

- Works in a two-step fashion:
  - Create a report object with the [report()](https://) function
  - Report object can be displayed either textually (the default output) or as a table, using [as.data.frame()](https://)
- Can also access a compact version of the report using [summary()](https://) on the report object



Specific reports:
- `report_participants()`
- `report_sample()`
- `report_system()`
- `report_packages()`
- …

✓ t-test
✓ cor.test
✓ ANOVAs
✓ GLMs
✓ Mixed models
✓ Bayesian models
✓ data.frame
✓ …

Direct access to subparts:
- `report_text()`
- `report_table()`
- `report_model()`
- `report_statistics()`
- `report_parameters()`
- …

Did you ever hear the Tragedy of Darth Plagueis the wise? I thought not. It's not a story the Jedi would tell you. It's a Sith legend. Darth Plagueis was a Dark Lord of the Sith, so powerful and so wise he could use the power of easystats to influence the midichlorians to create life… He had such a knowledge of the report package that he could even keep the ones he cared about from dying. The dark side of easystats is a pathway to many abilities some consider to be… unnatural.

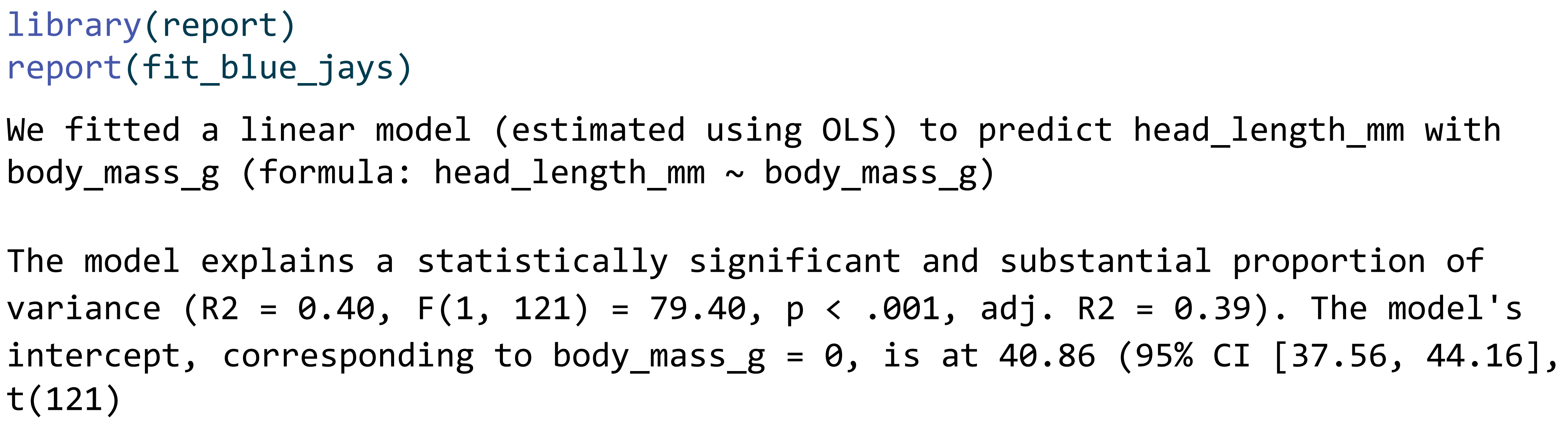
Short textual summary

X	Y	Z
a	1.6	2.5
b	3.1	4.2
c	5.2	6.1

X	Y	Z	X1	Y1	Z1
a	1.6	2.5	7.2	1.8	2.2
b	3.1	4.2	2.1	5.7	3.1
c	5.2	6.1	5.8	2.4	8.2

report() Function

- Nice and easy way to report results of a regression analysis in R is with the `report()` function from the `{report}` package:



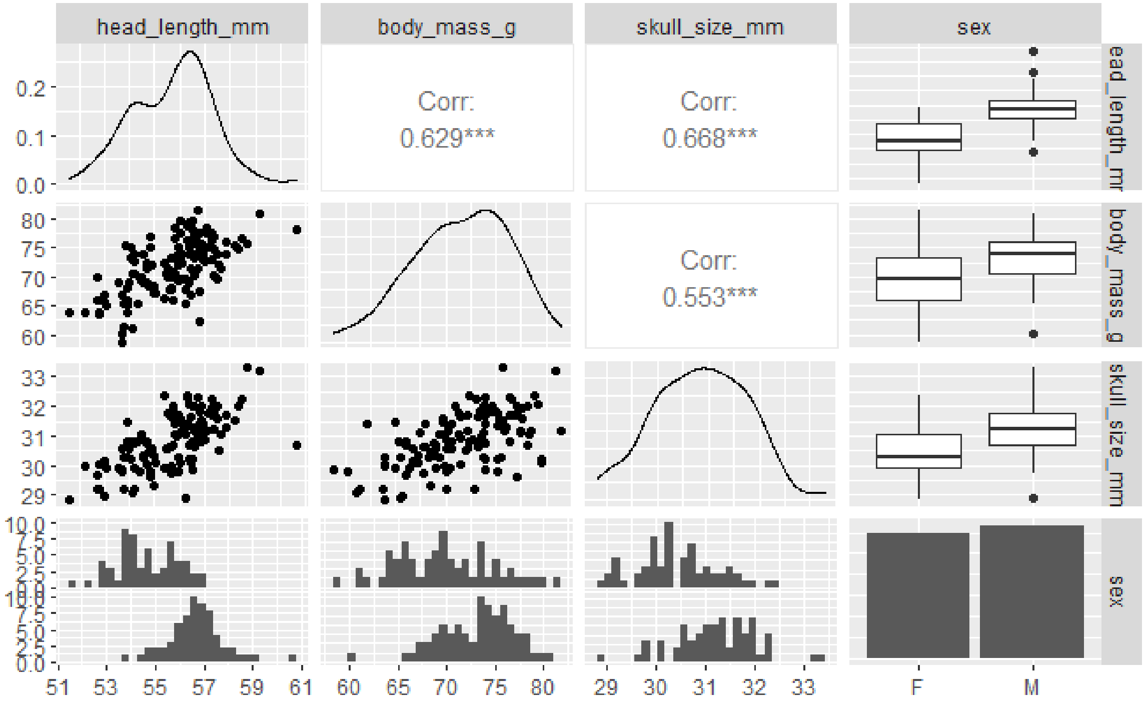
```r
library(report)
report(fit_blue_jays)
```

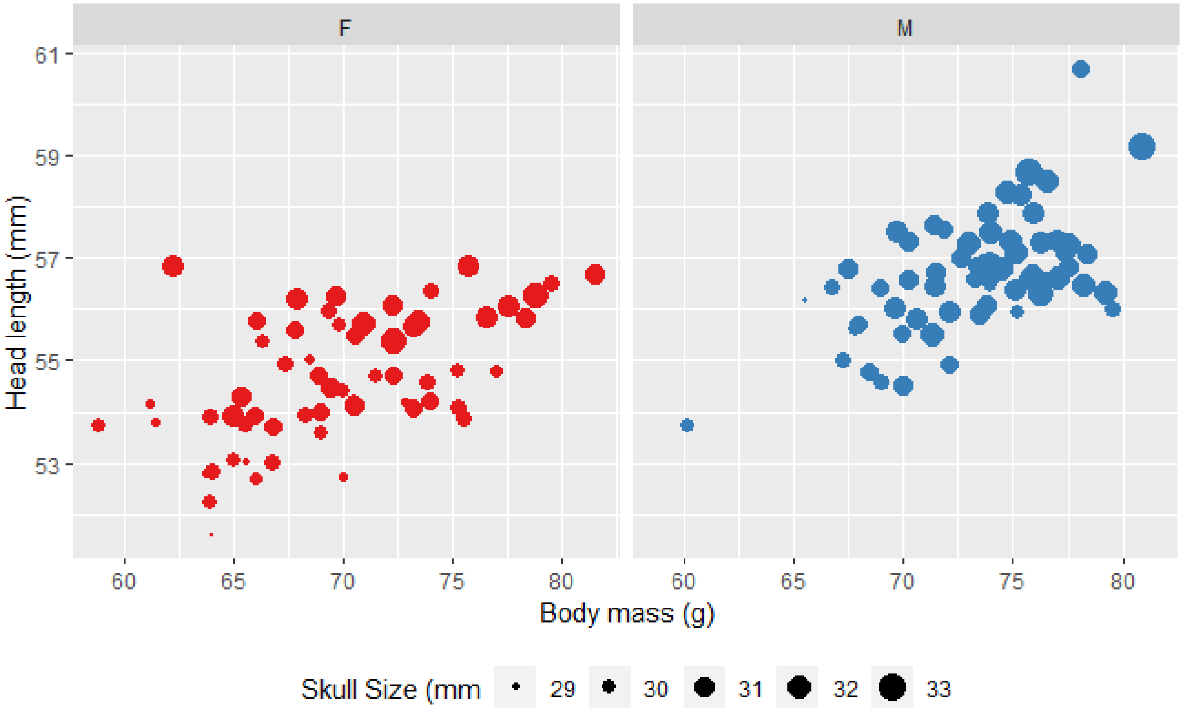
```
We fitted a linear model (estimated using OLS) to predict head_length_mm with
body_mass_g (formula: head_length_mm ~ body_mass_g)

The model explains a statistically significant and substantial proportion of
variance (R2 = 0.40, F(1, 121) = 79.40, p < .001, adj. R2 = 0.39). The model's
intercept, corresponding to body_mass_g = 0, is at 40.86 (95% CI [37.56, 44.16],
t(121)
```

Other Methods for Visualizing Associations

Correlation Matrix and Bubble Plot





Correlation Matrix

- Used with three or more quantitative variables
- In this case, it is more useful to quantify the amount of association between pairs of variables and visualize these quantities
- Common method is using correlation coefficients

Wilke, C. (2019)

Correlation Matrix

- Correlogram or correlation matrix allows to analyze the relationship between each pair of numeric variables in a dataset
- Gives a quick overview of the whole dataset. It is more used for exploratory purpose than explanatory

Wilke, C. (2019)

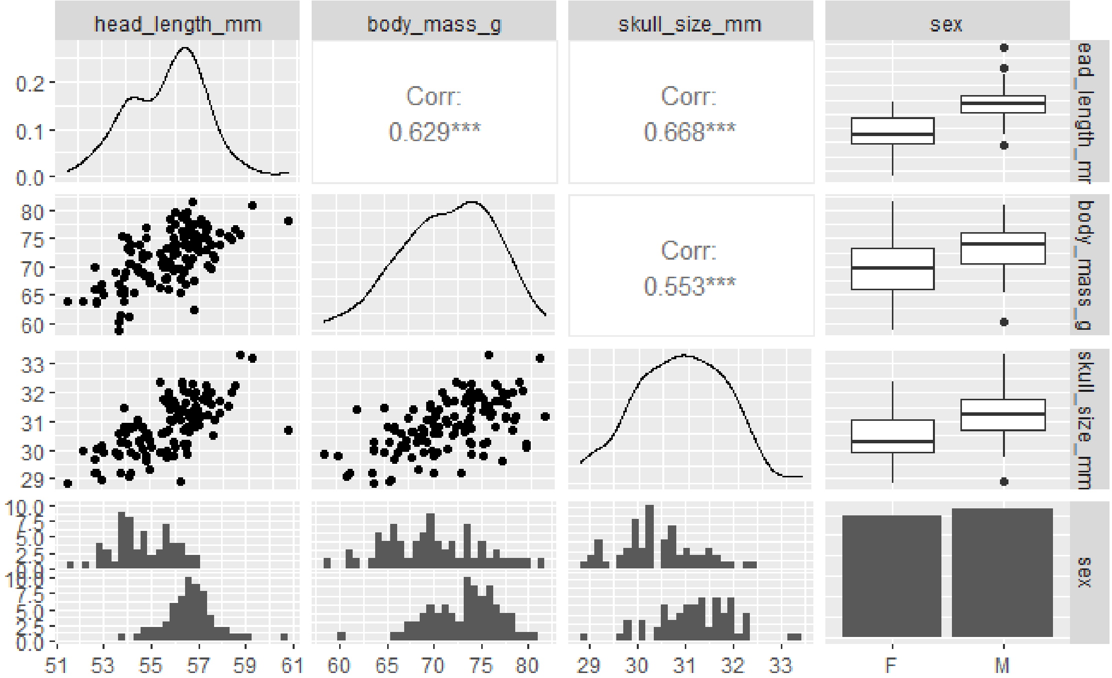
Correlation Matrix: GGalley

- Ggally options to build correlograms:
  - pairwise plot matrix
  - scatterplot plot matrix
  - parallel coordinates plot
  - survival plot
- ggpairs() function build a classic correlogram with scatterplot, correlation coefficient and variable distribution

Wilke, C. (2019)

Correlation Matrix: Example

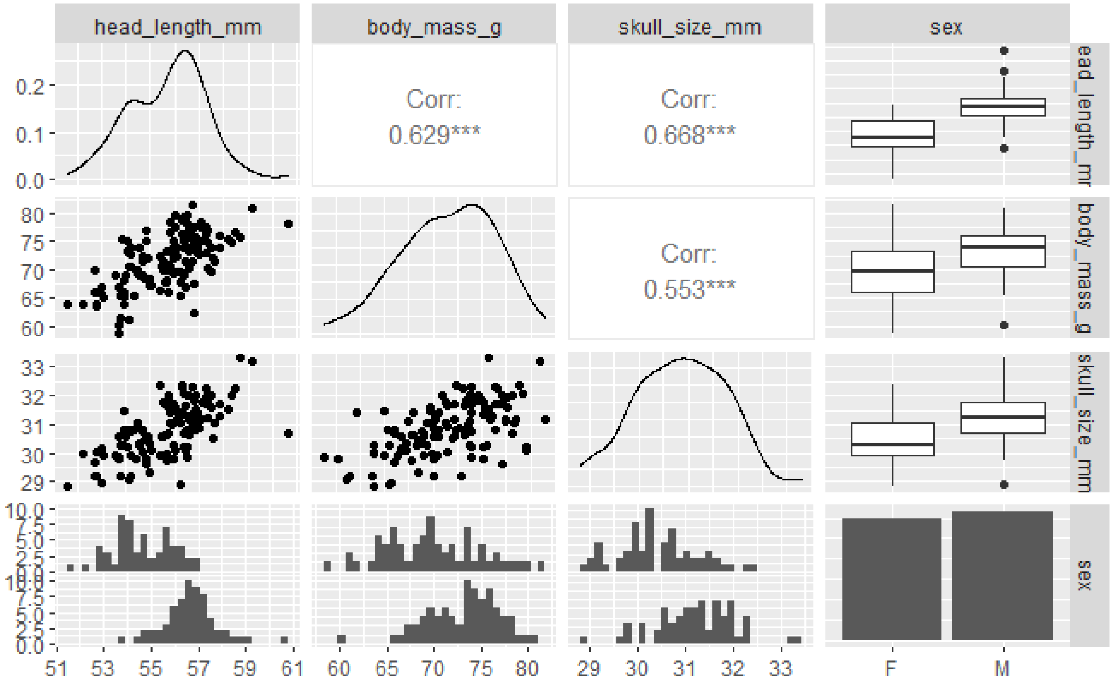
- ggpairs plots each variable against the other
- Scatterplots for quantitative, quantitative pairs
- Top half, is the correlation for each of the quantitative, quantitative pairs



Wilke, C. (2019)

Correlation Matrix: Example

- On the diagonal, we have the density functions for each of the variables
- Boxplots and histograms for the qualitative variable, sex
- All variables are correlated, with significance

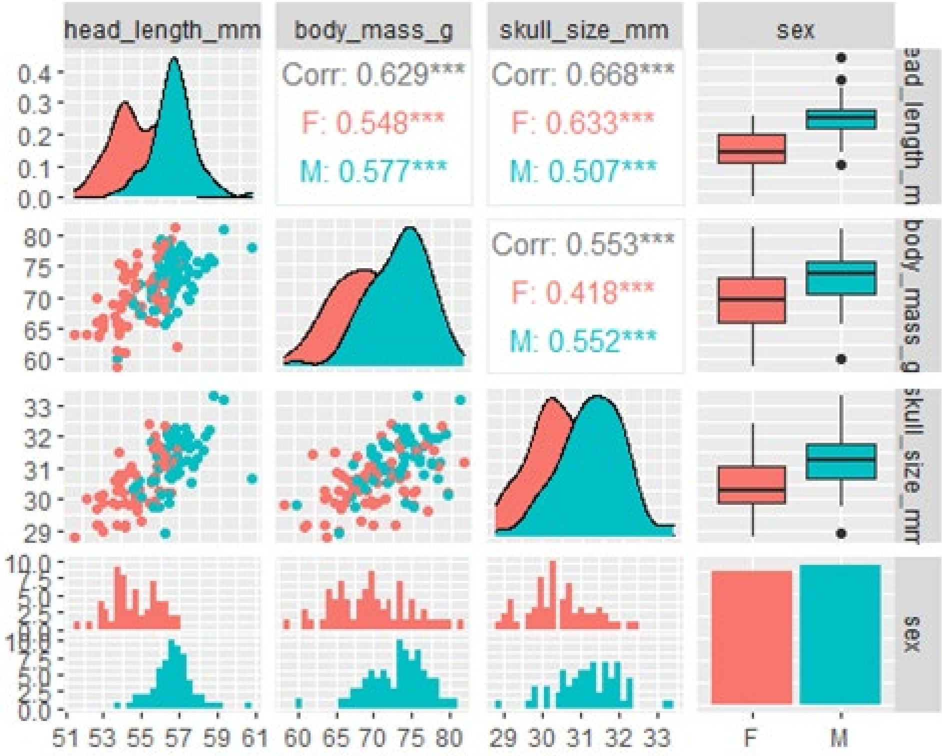


Wilke, C. (2019)

Correlation Matrix: Example – ggplot

- It is also possible to inject ggplot2 code into a ggcor statement. For example, you can add color categories

```
ggpairs(blue_jays_matrix,
ggplot2::aes(colour=sex))
```

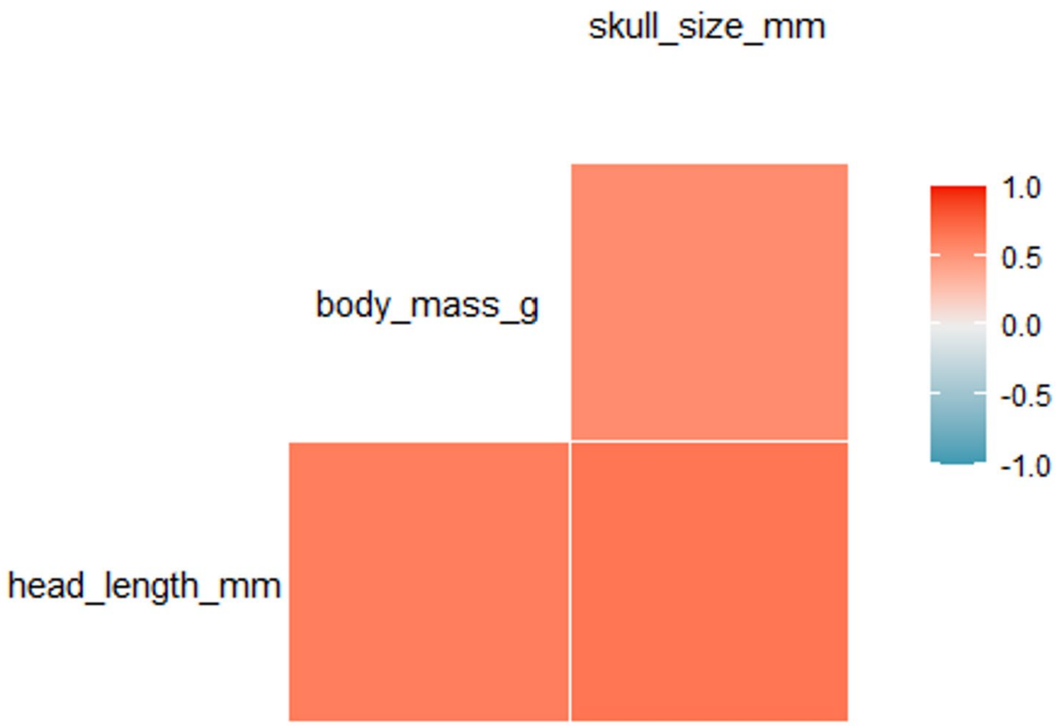


Wilke, C. (2019)

Correlation Matrix: Example – ggcorr()

- Another option is the ggcorr() function
- Visualize the correlation of each pair of variable as tiles
- Method sets the correlation type

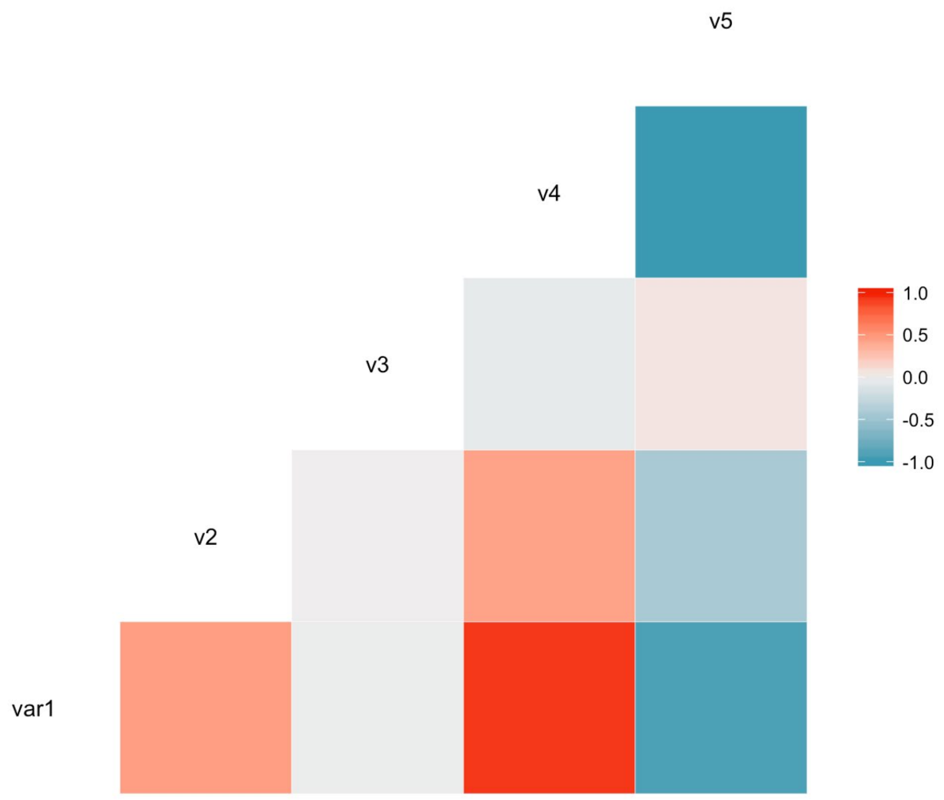
```
blue_jays_matrix %>%
  ggcorr(method = c("everything",
"pearson"))
```



Wilke, C. (2019)

Correlation Matrix: Example – ggcorr()

- Blue jar data only has positive correlations
- This figure is demonstrating what this would look like if we had both positive and negative correlations between the variables

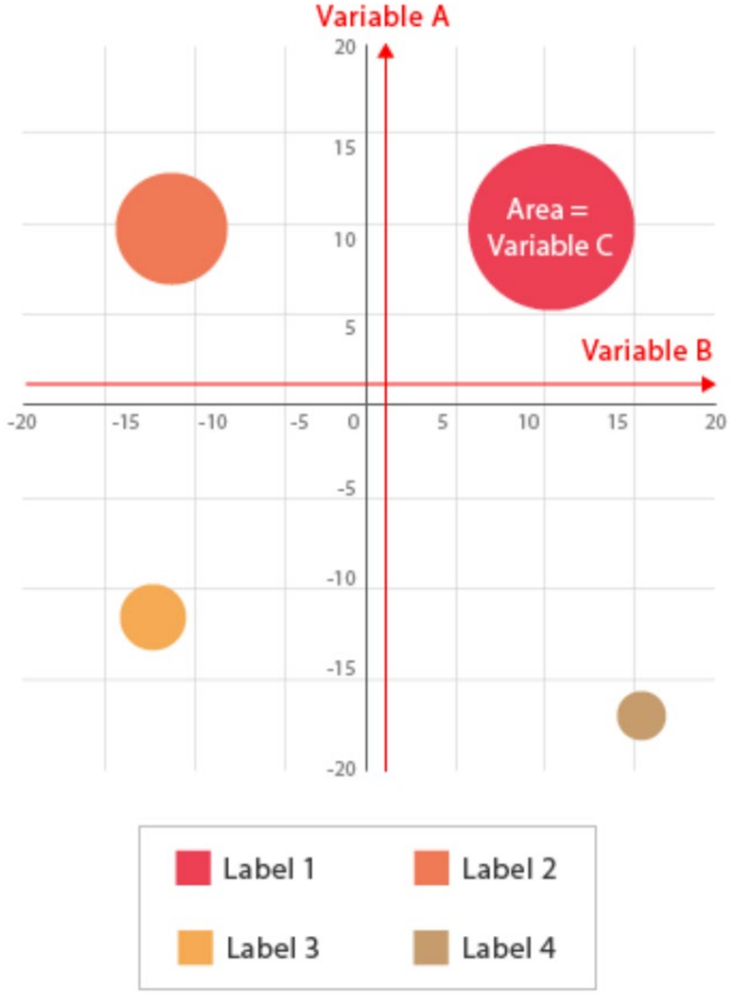


Wilke, C. (2019)

Bubble Graph

Bubble Graph

- Multi-variable graph
- Cross between a scatterplot and a proportional area chart
- Compares and show the relationships between categorized circles
- Uses positioning and proportions
- Can be used to analyze for patterns/correlations



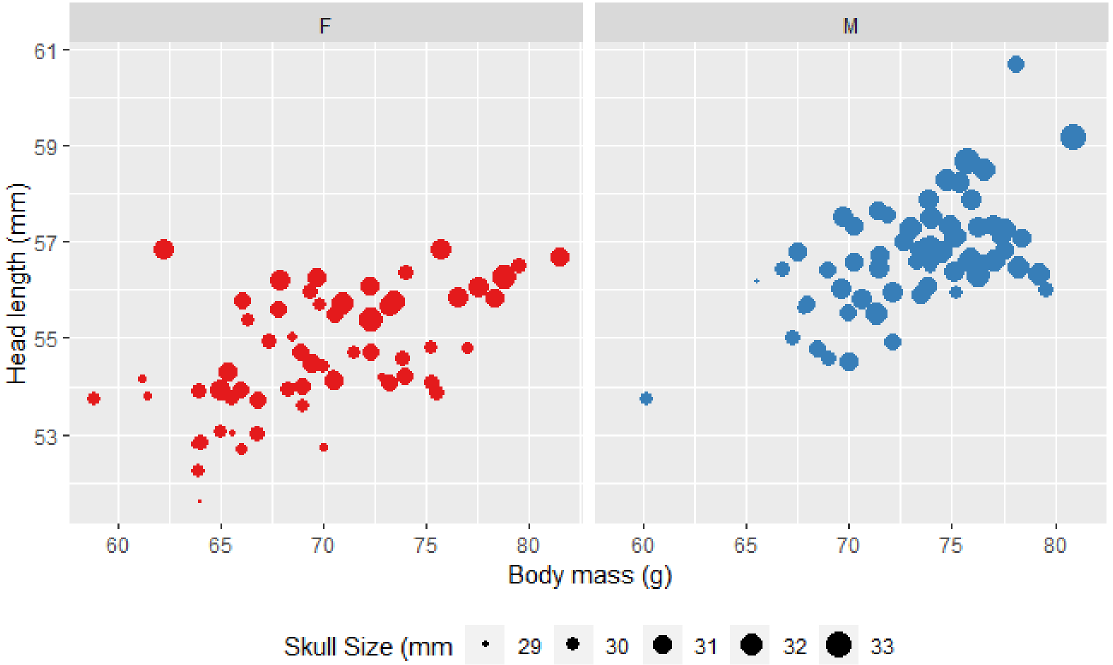
Ribecca, S. (2019)

Bubble Graph: Example

- If we wanted to look at head size and bill size, we can do that too
- Need another aesthetic to which we can map skull size, what could it be?

Bubble Graph: Example

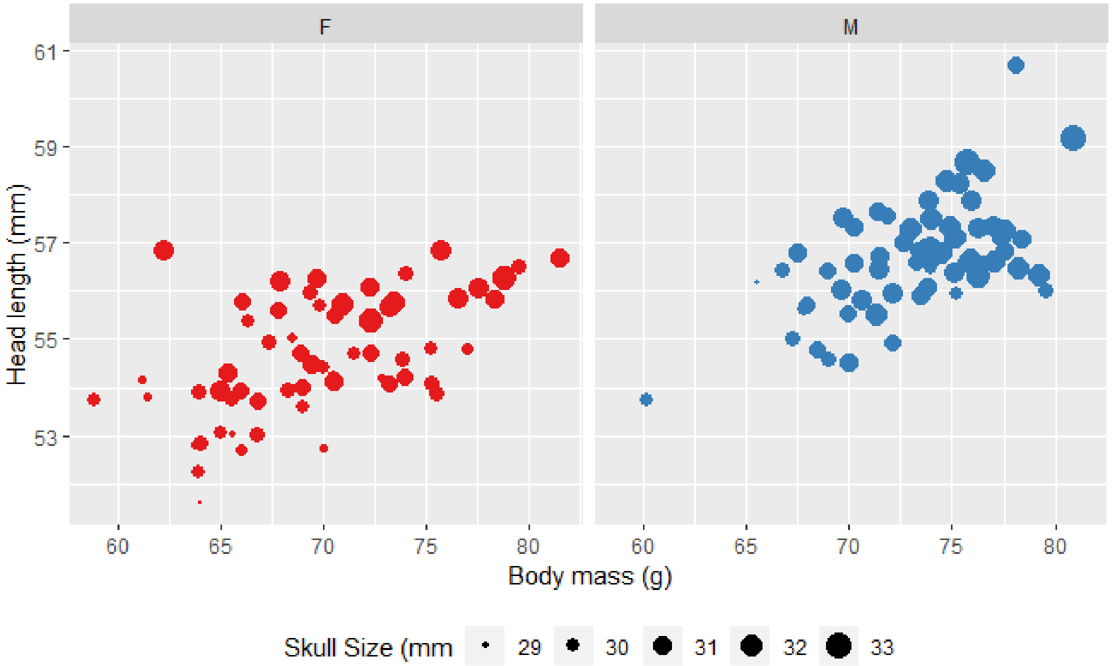
- Already using the x position for body mass
- Position for head length
- Dot color for bird sex
- Birds' skull size by symbol size



Wilke, C. (2019)

Bubble Graph: Interpretation

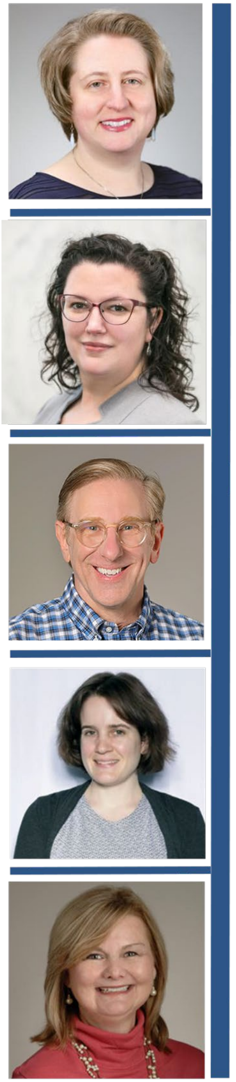
- Head length and skull size appear to be correlated
- Some birds with unusually long or short bills given their skull size
- Guidance on how to only display certain legends
- Change the position of the legend using this resource and this resource



Bubble Graph: Code

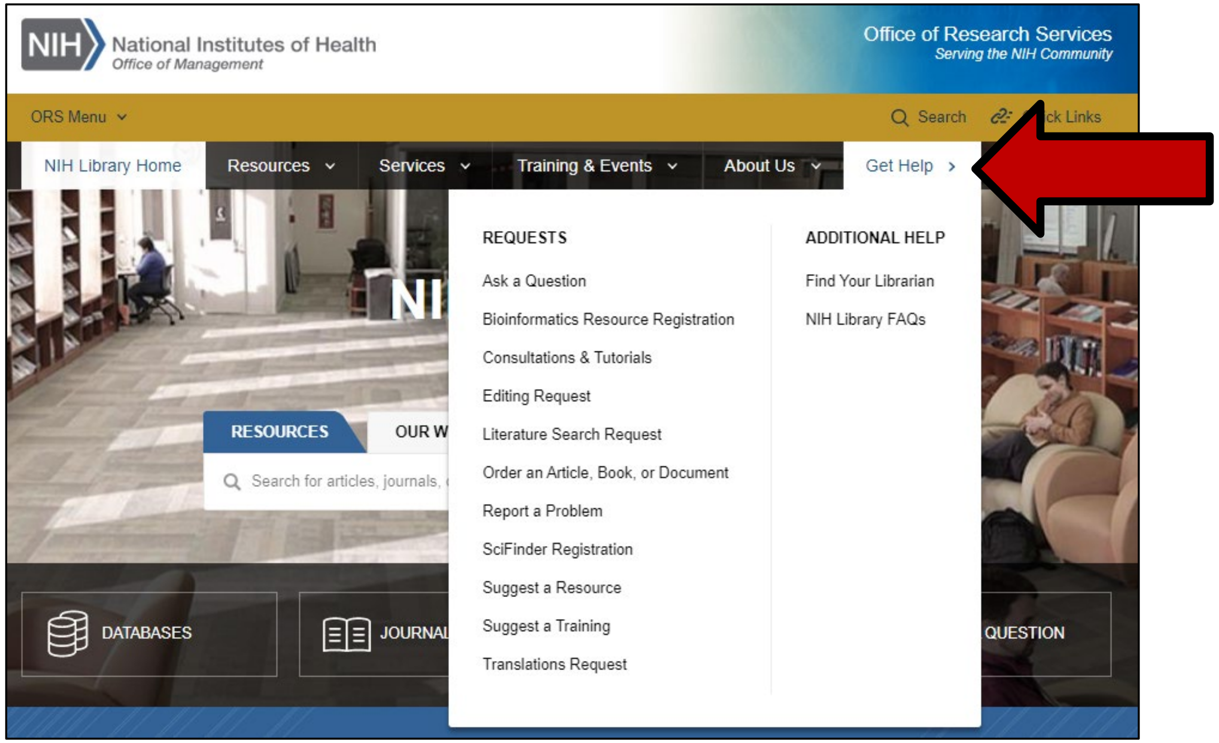
```
blue_jays %>%
  ggplot(mapping = aes(x = body_mass_g, y = head_length_mm,
  size = skull_size_mm, color = sex)) +
  geom_point() +
  scale_size(range = c(.1, 5), name= "Skull Size (mm)") +
  facet_wrap(vars(sex)) +
  scale_color_brewer(palette="Set1") +
  labs(y = "Head length (mm)", x = "Body mass (g)") +
  guides(col = FALSE) +
  theme(legend.position="bottom")
```

We Can Help



- Classes on a variety of data-related topics, including:
  - Data management
  - Data visualization
  - Data analysis
  - R and RStudio
- Computers which offers a suite of tools for data analysis, processing, and visualization

Contact Us for Ongoing Support



Doug Joubert

Bioinformatics Support Program

301-827-3829

douglas.joubert@nih.gov

NIH Library Help Desk

(301) 496-1080

- **Ask a Question**: https://www.nihlibrary.nih.gov/get-help/ask-question
- **Request a Tutorial**: https://www.nihlibrary.nih.gov/get-help/consultations-tutorials
- **Sign up for Additional Classes**: https://www.nihlibrary.nih.gov/training/calendar

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Institutes of Health
Office of Management