Data viz and the grammar of graphics

Data Science for Biologists, Fall 2021

Reminder: The "Types of Plots" App is here for you!!

- Access from anywhere: https://sjspielman.shinyapps.io/types_of_plots
- Access from your RStudio Cloud project:

```
library(ds4b.materials) # Load the library if not already loaded
launch_app("types_of_plots") # Launch the app once library is loaded
```

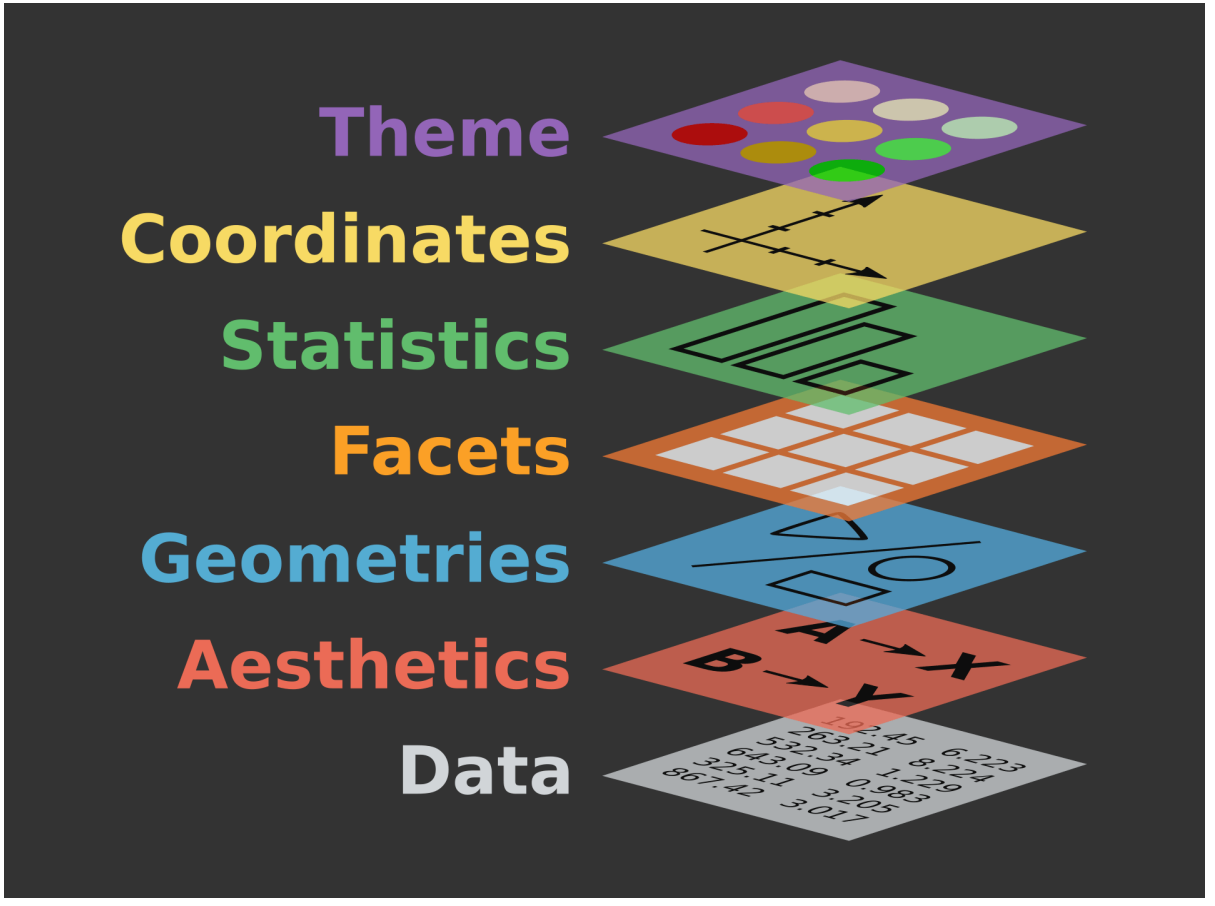
Grammar

The dog runs in a park.

The runs in park dog a.

Runs dog park in a the.

In park a the runs dog.



Aesthetics --> *aesthetic mappings*

The dataset

```
msleep_subvore
```

```
## # A tibble: 46 × 5
##    name                          vore  awake brainwt  bodywt
##    <chr>                         <fct> <dbl>   <dbl>   <dbl>
##  1 Owl monkey                    omni    7   0.0155    0.48
##  2 Greater short-tailed shrew    omni    9.1 0.00029   0.019
##  3 Cow                           herbi  20   0.423   600
##  4 Dog                           carni  13.9 0.07     14
##  5 Roe deer                      herbi  21   0.0982   14.8
##  6 Goat                          herbi  18.7 0.115    33.5
##  7 Guinea pig                    herbi  14.6 0.0055    0.728
##  8 Chinchilla                    herbi  11.5 0.0064    0.42
##  9 Star-nosed mole               omni   13.7 0.001     0.06
## 10 African giant pouched rat     omni   15.7 0.0066    1
## # … with 36 more rows
```

The dataset

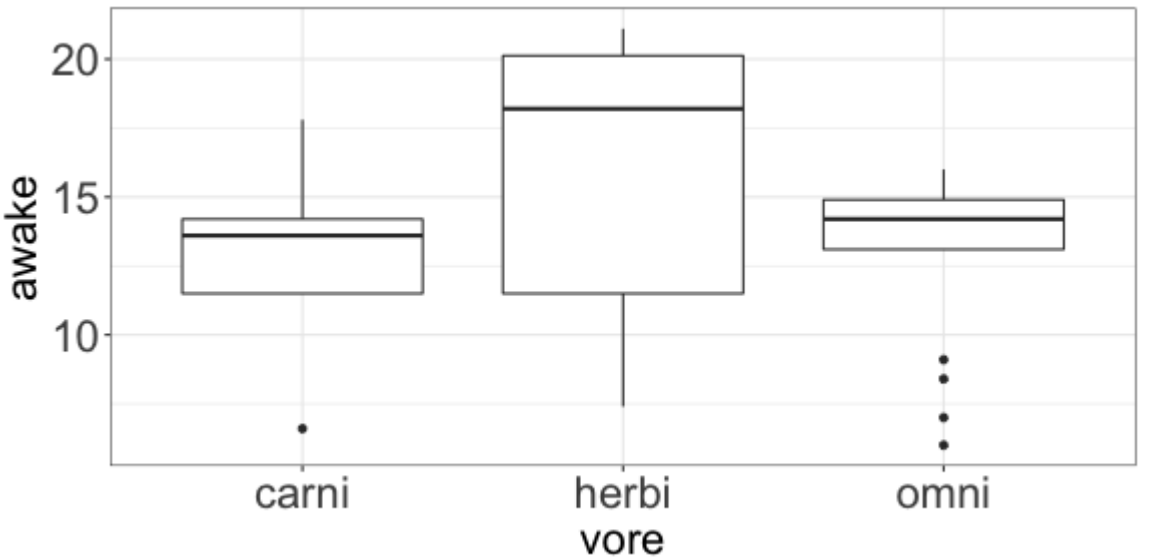
```
summary(msleep_subvore)
```

```
##      name                vore        awake          brainwt
##  Length:46          carni: 9   Min.   : 6.00   Min.   :0.000140
##  Class :character   herbi:20   1st Qu.:11.50   1st Qu.:0.005125
##  Mode  :character   omni :17   Median :14.25   Median :0.016500
##                                Mean   :14.39   Mean   :0.339623
##                                3rd Qu.:17.70   3rd Qu.:0.173500
##                                Max.   :21.10   Max.   :5.712000
##      bodywt
##  Min.   :   0.005
##  1st Qu.:   0.542
##  Median :   2.788
##  Mean   : 245.575
##  3rd Qu.:  47.525
##  Max.   :6654.000
```

```
unique(msleep_subvore$name)
```

```
##  [1] "Owl monkey"                       "Greater short-tailed shrew"
##  [3] "Cow"                              "Dog"
##  [5] "Roe deer"                         "Goat"
##  [7] "Guinea pig"                       "Chinchilla"
##  [9] "Star-nosed mole"                  "African giant pouched rat"
## [11] "Lesser short-tailed shrew"        "Long-nosed armadillo"
## [13] "Tree hyrax"                       "North American Opossum"
## [15] "Asian elephant"                   "Horse"
## [17] "Donkey"                           "European hedgehog"
## [19] "Patas monkey"                     "Domestic cat"
## [21] "Galago"                           "Gray seal"
## [23] "Gray hyrax"                       "Human"
## [25] "African elephant"                 "Macaque"
## [27] "Golden hamster"                   "House mouse"
## [29] "Slow loris"                       "Rabbit"
## [31] "Sheep"                            "Chimpanzee"
## [33] "Jaguar"                           "Baboon"
## [35] "Laboratory rat"                   "Squirrel monkey"
## [37] "Cotton rat"                       "Arctic ground squirrel"
## [39] "Thirteen-lined ground squirrel" "Pig"
## [41] "Brazilian tapir"                  "Tenrec"
## [43] "Tree shrew"                       "Genet"
## [45] "Arctic fox"                       "Red fox"
```

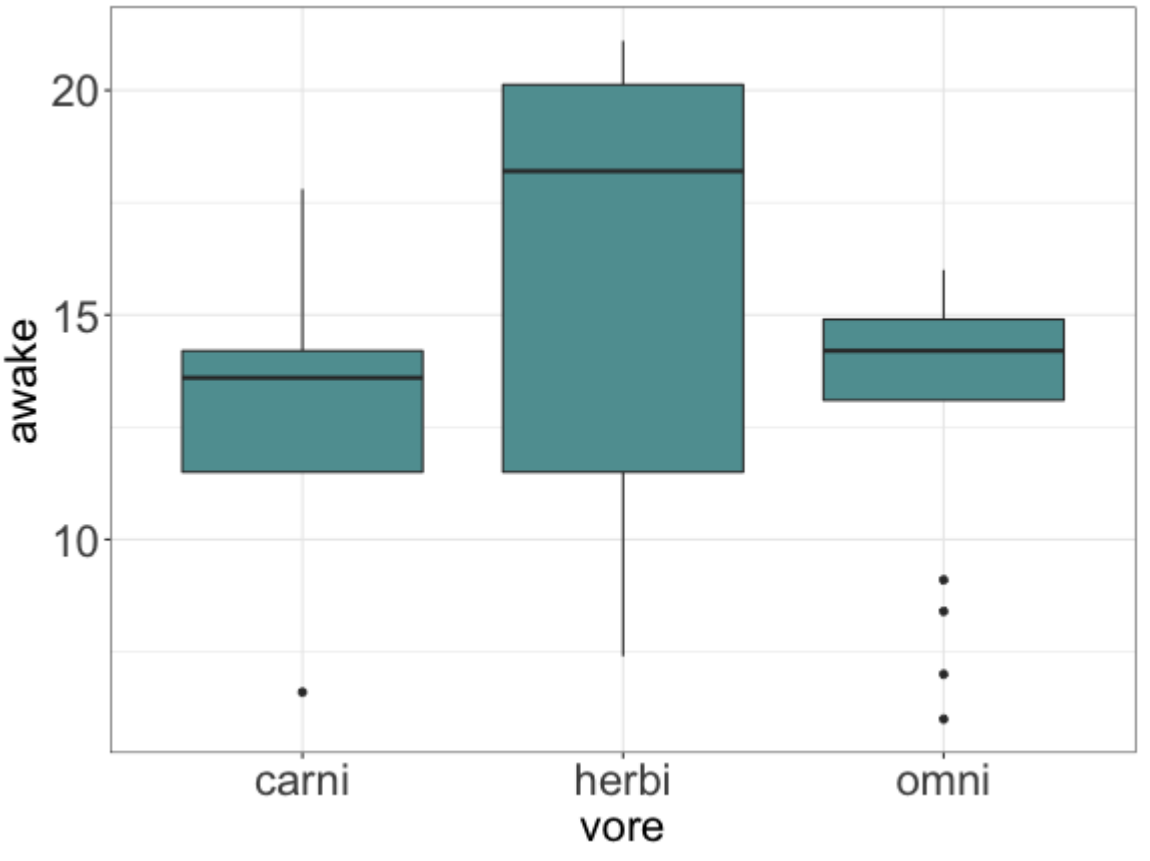
Identifying components of a plot

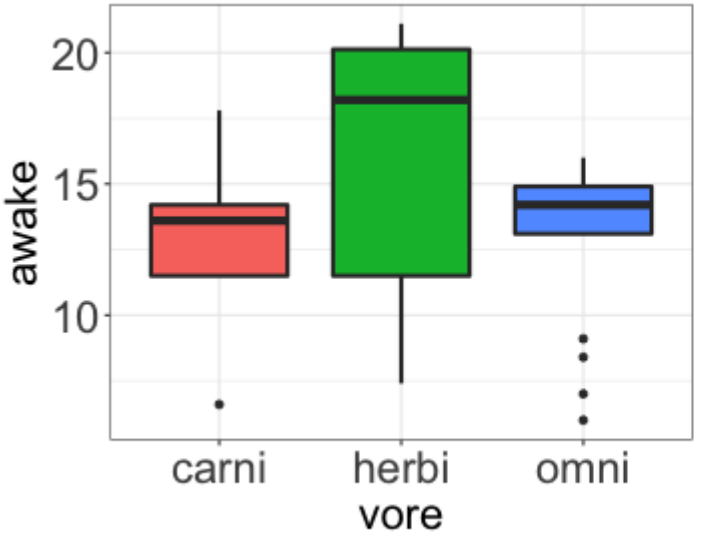


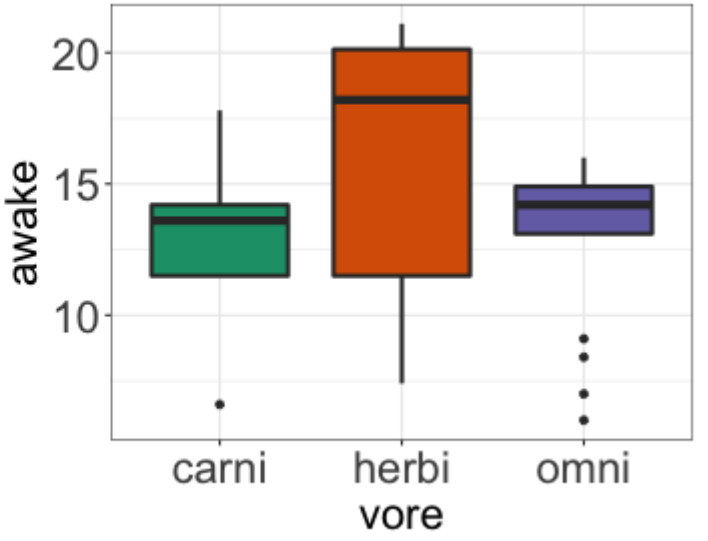
Aesthetics: How is the data *mapped onto* visual components of the plot?

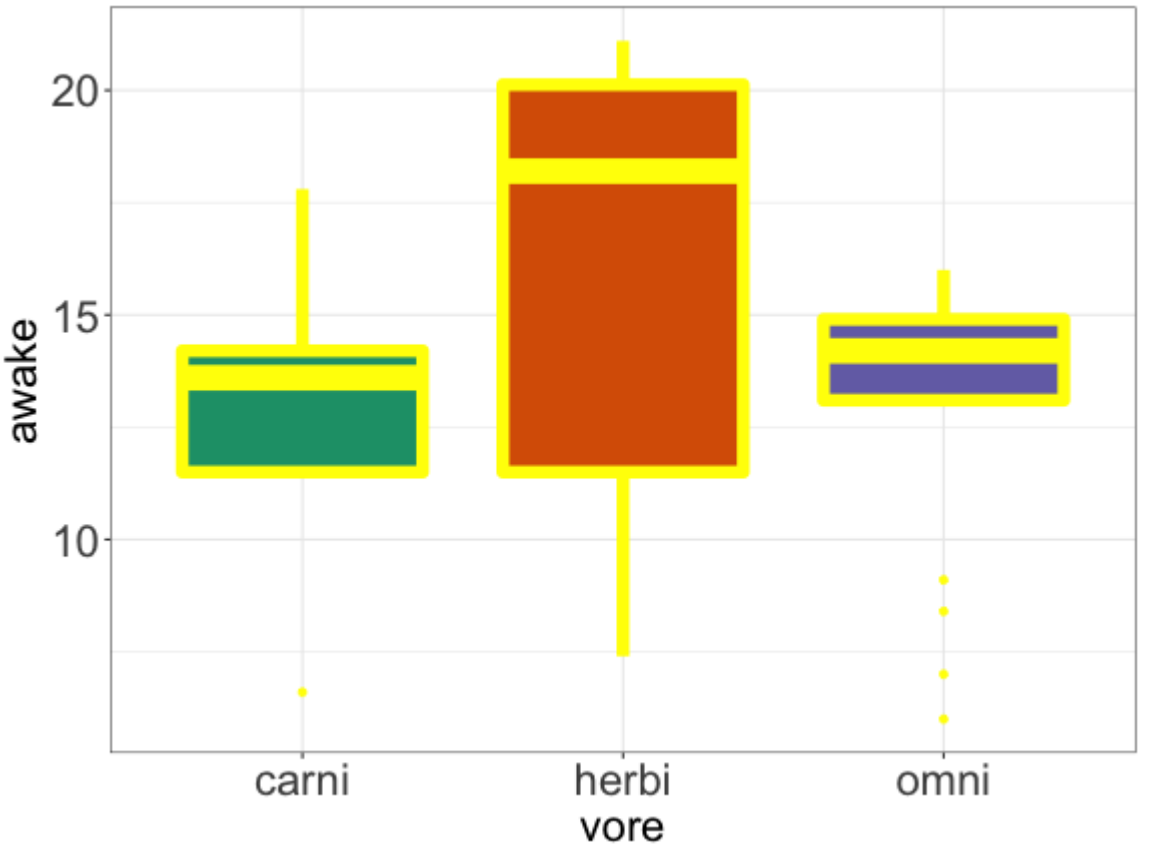
- X-axis?
- Y-axis?
- Colors? Shapes? Sizes?

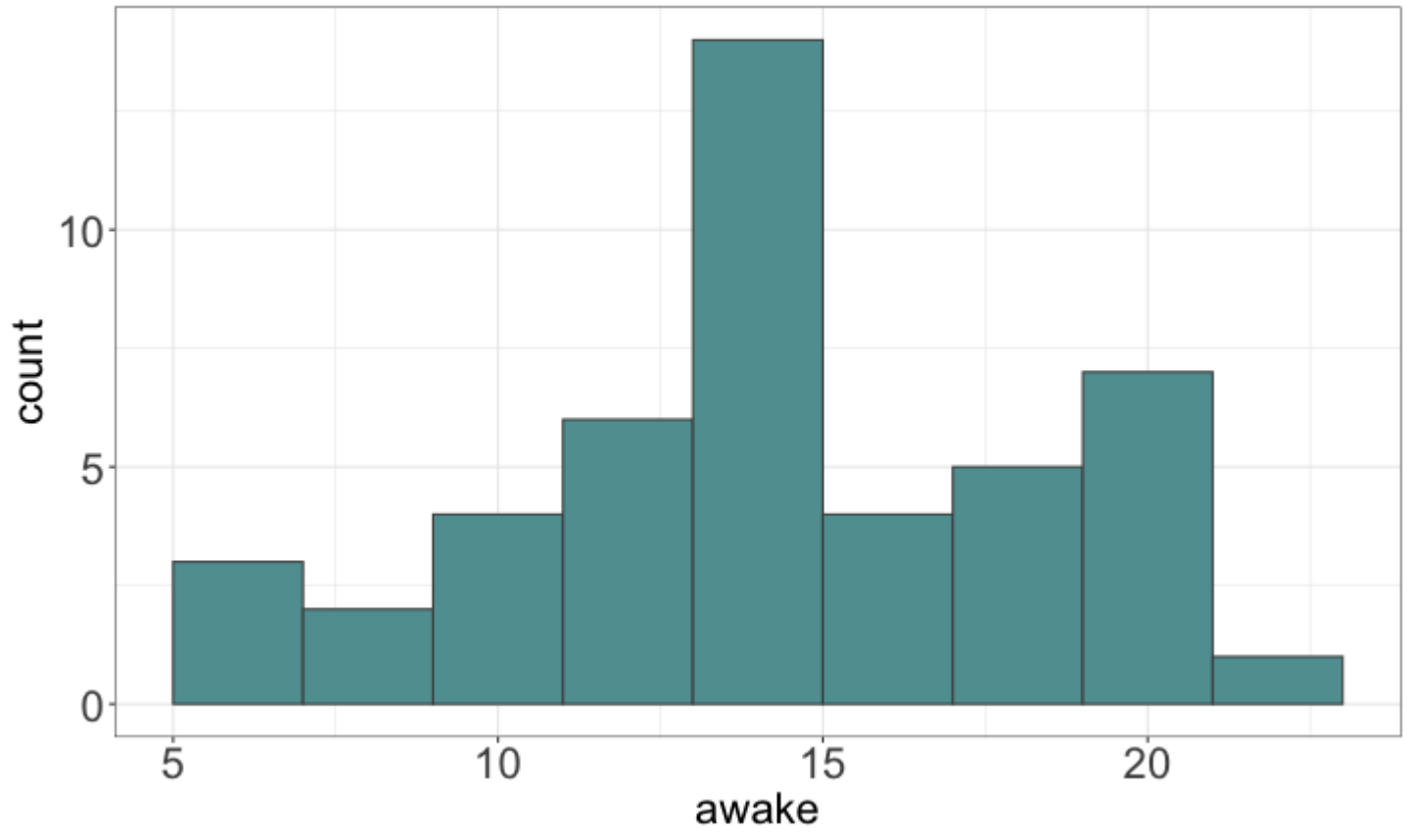
Geometries: What *shapes* aka *geometric objects* are displayed in the plot? (Often AKA: What type of plot?)


20
15
10
awake
carni
herbi
omni
vore

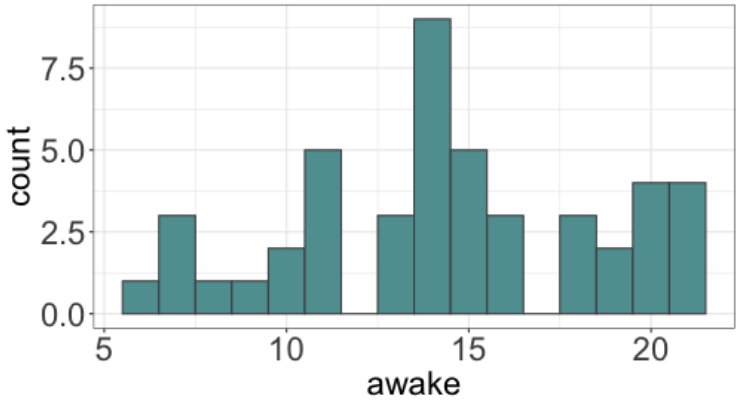

awake
20
15
10
carni
herbi
omni
vore

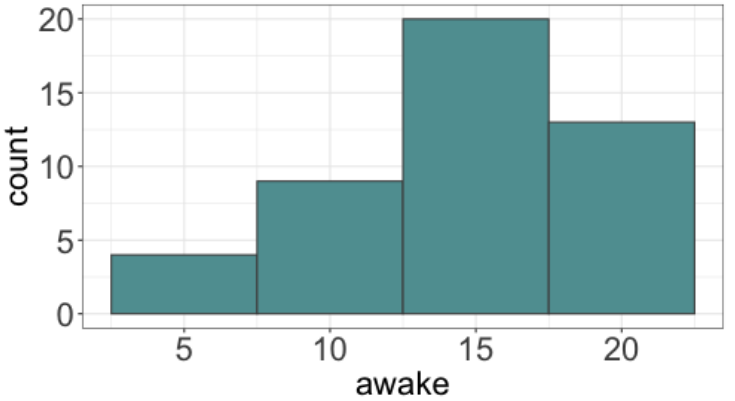

awake
20
15
10
carni
herbi
omni
vore

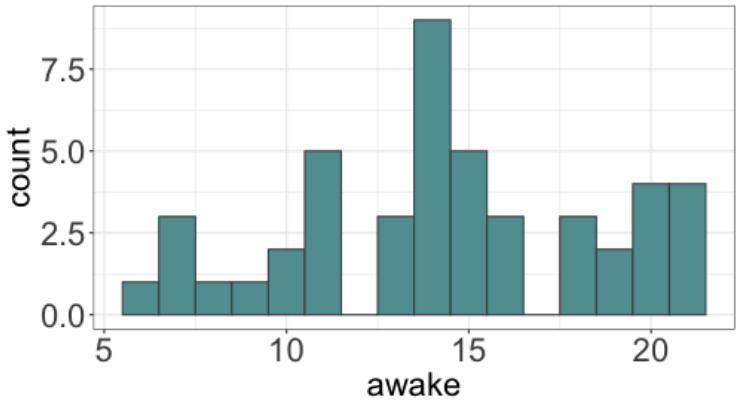

20
15
10
awake
carni
herbi
omni
vore

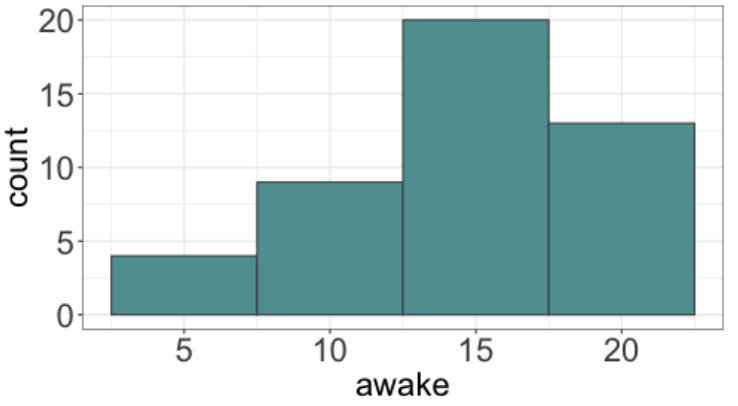


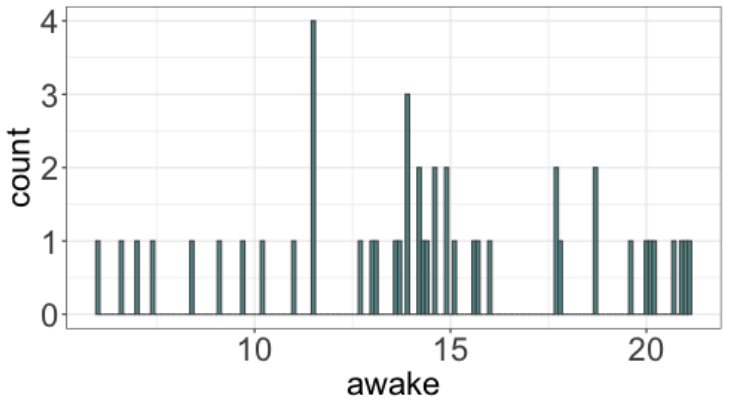

```
##  [1]  6.0  6.6  7.0  7.4  8.4  9.1  9.7 10.2 11.0 11.5 11.5 11.5 11.5 12.7
## [16] 13.1 13.6 13.7 13.9 13.9 13.9 14.2 14.2 14.3 14.4 14.6 14.6 14.9 14.9
## [31] 15.6 15.7 16.0 17.7 17.7 17.8 18.7 18.7 19.6 20.0 20.1 20.2 20.7 20.9
## [46] 21.1
```

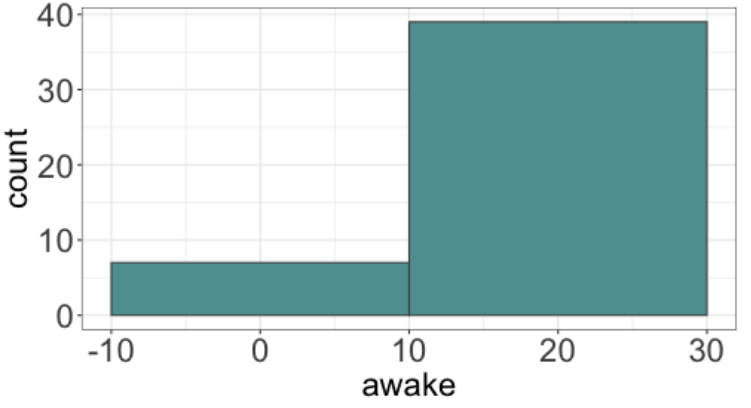

count
7.5
5.0
2.5
0.0
5
10
15
20
awake

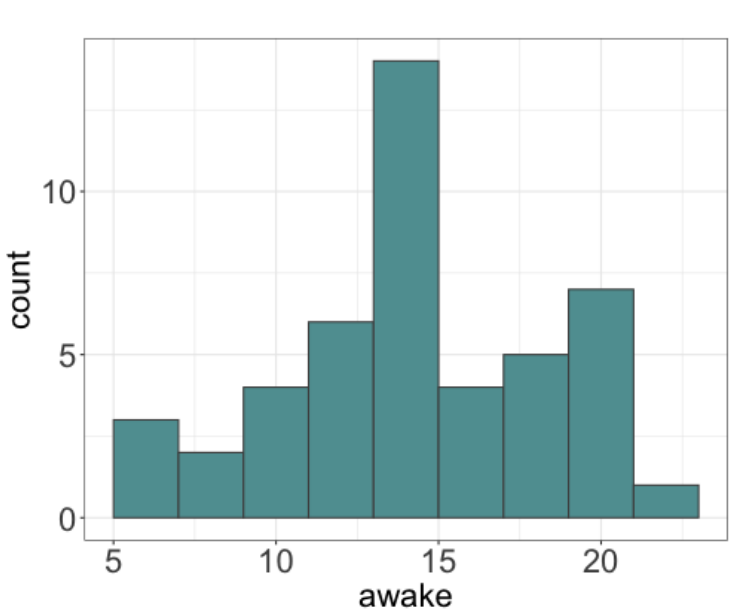

count
20
15
10
5
0
5
10
15
20
awake

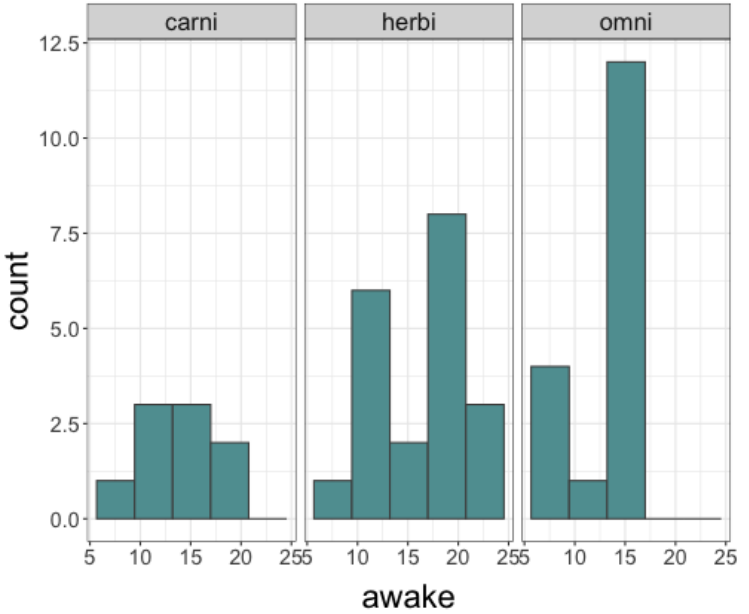

count
7.5
5.0
2.5
0.0
5
10
15
20
awake

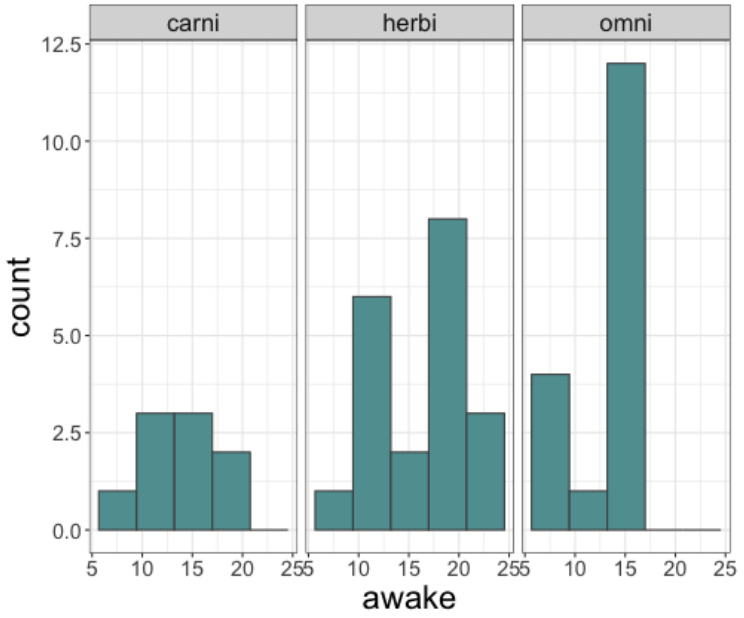

count
20
15
10
5
0
5
10
15
20
awake

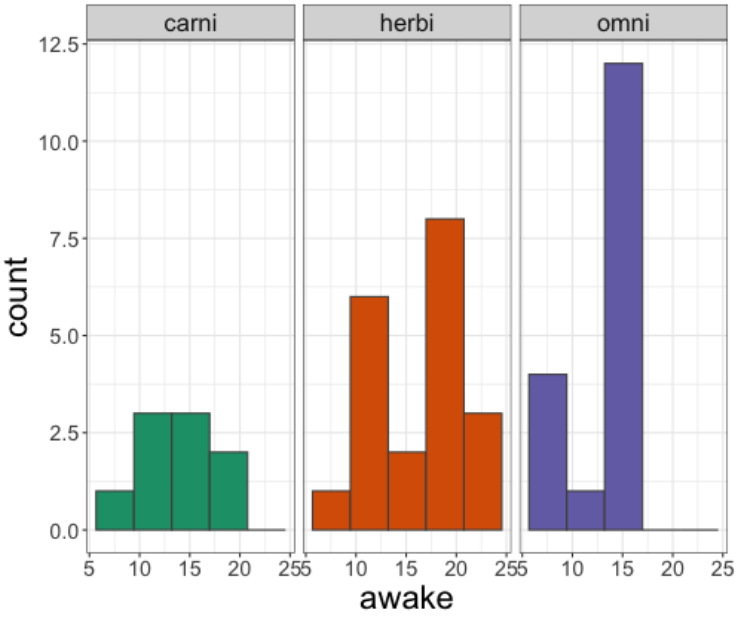

count
4
3
2
1
0
10
15
20
awake

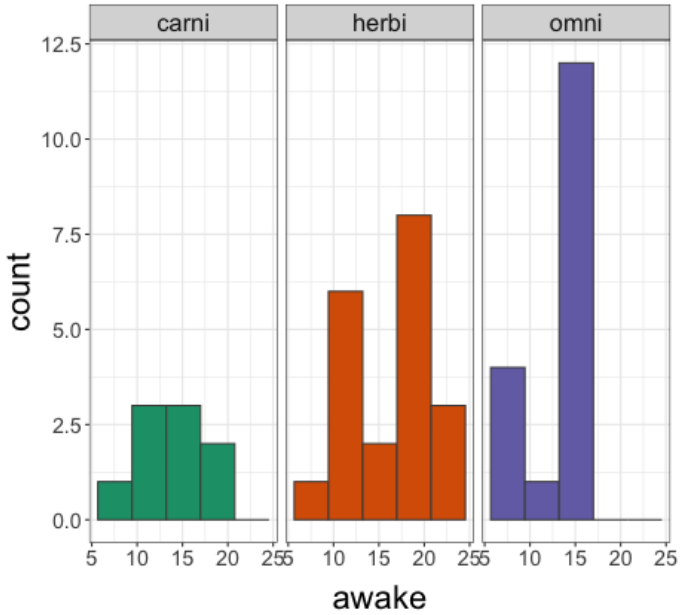

count
40
30
20
10
0
-10
0
10
20
30
awake

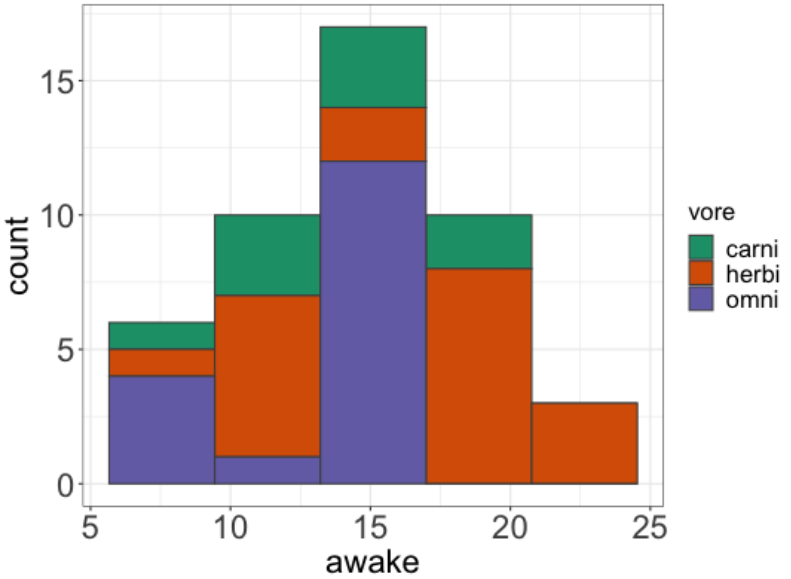

count
10
5
0
5
10
15
20
awake

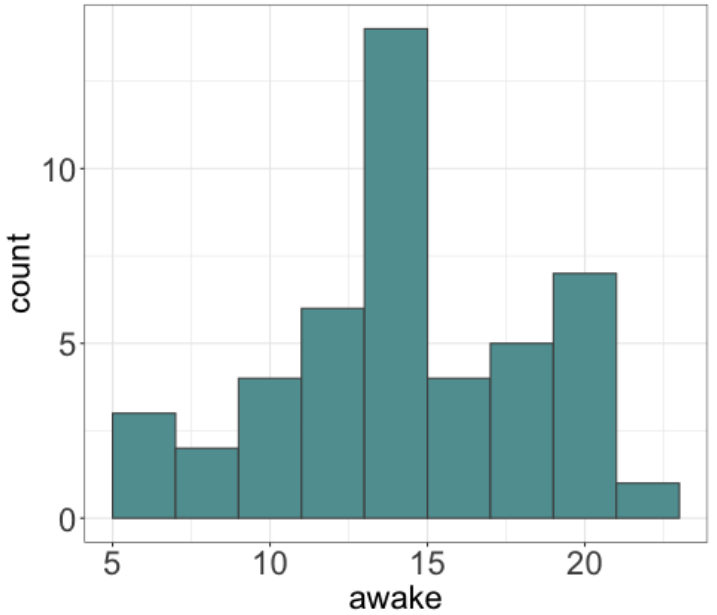

carni
herbi
omni
count
12.5
10.0
7.5
5.0
2.5
0.0
5
10
15
20
25
awake

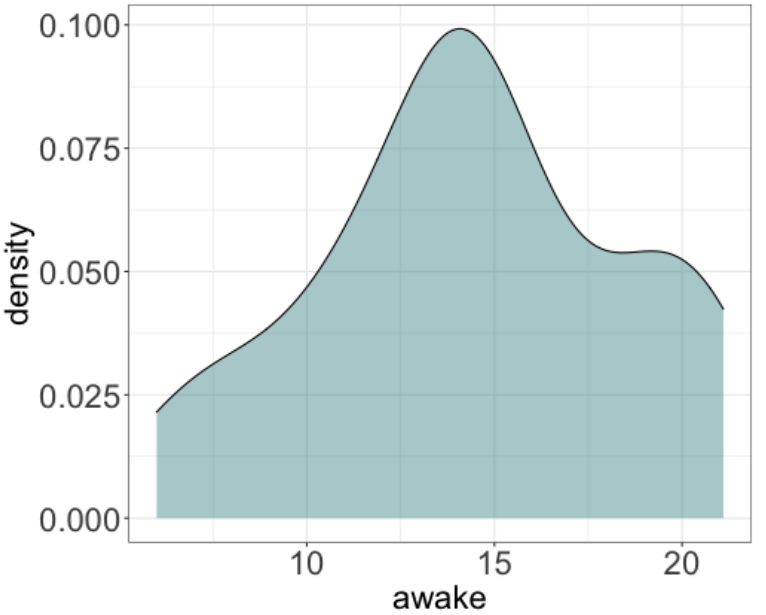

carni
herbi
omni
count
awake
0.0
2.5
5.0
7.5
10.0
12.5
5
10
15
20
25

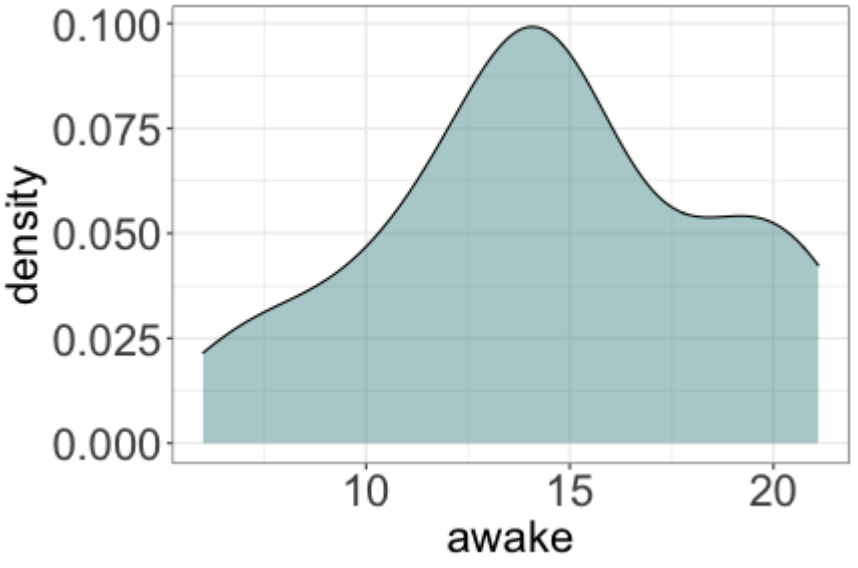

carni
herbi
omni
count
awake
0.0
2.5
5.0
7.5
10.0
12.5
5
10
15
20
25

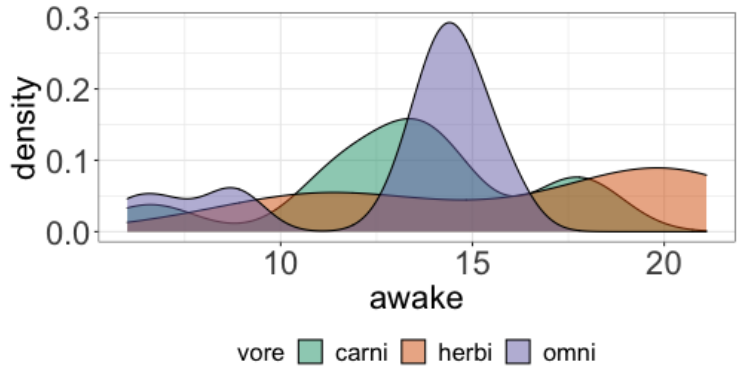

carni
herbi
omni
count
awake
0.0
2.5
5.0
7.5
10.0
12.5
5
10
15
20
25

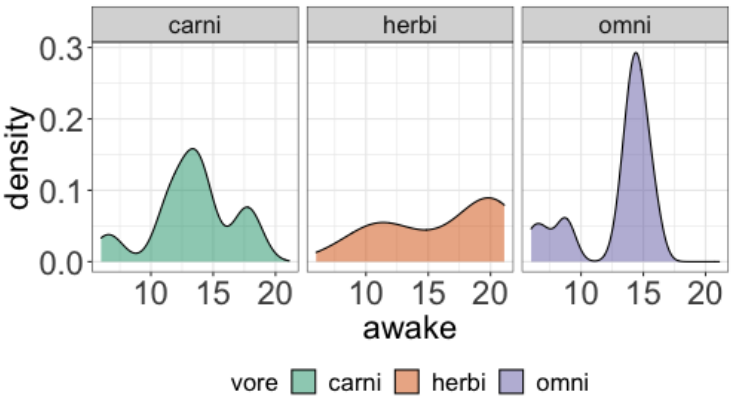

count
awake
vore
carni
herbi
omni
0
5
10
15
5
10
15
20
25

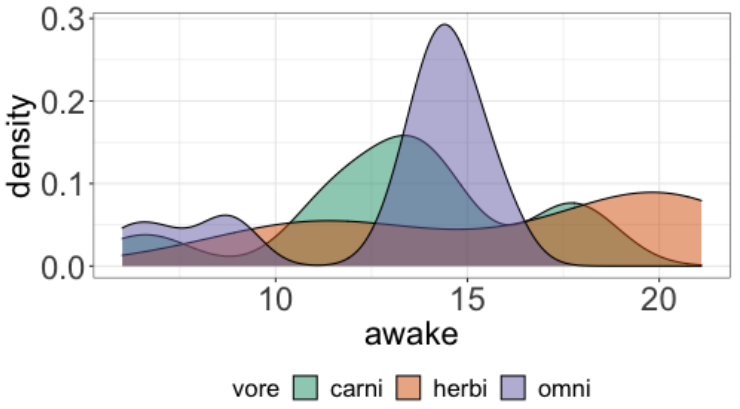

count
10
5
0
5
10
15
20
awake

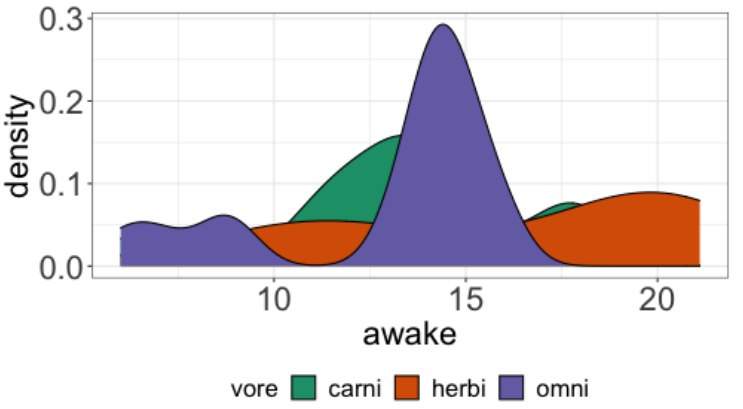

density
0.100
0.075
0.050
0.025
0.000
10
15
20
awake

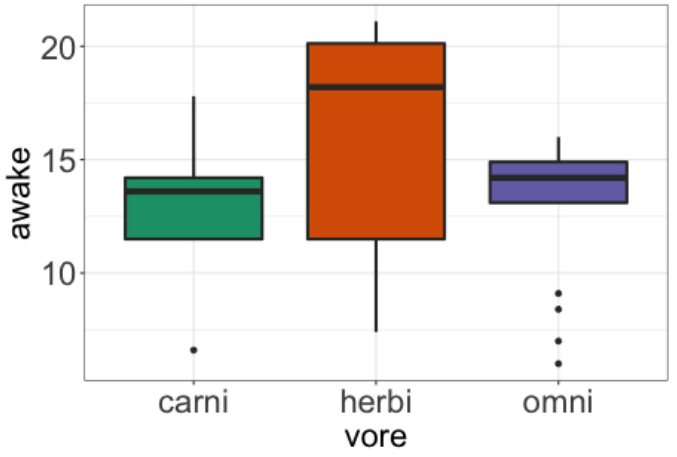

0.100
0.075
0.050
0.025
0.000
density
10
15
20
awake

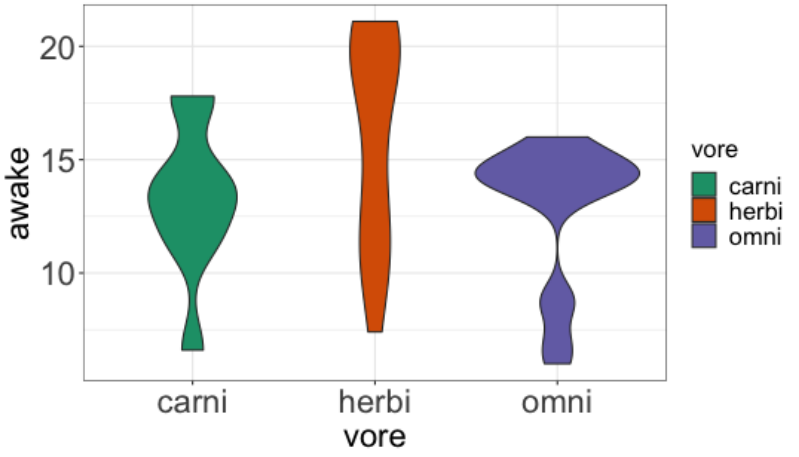

0.3
0.2
0.1
0.0
density
10
15
20
awake
vore
carni
herbi
omni

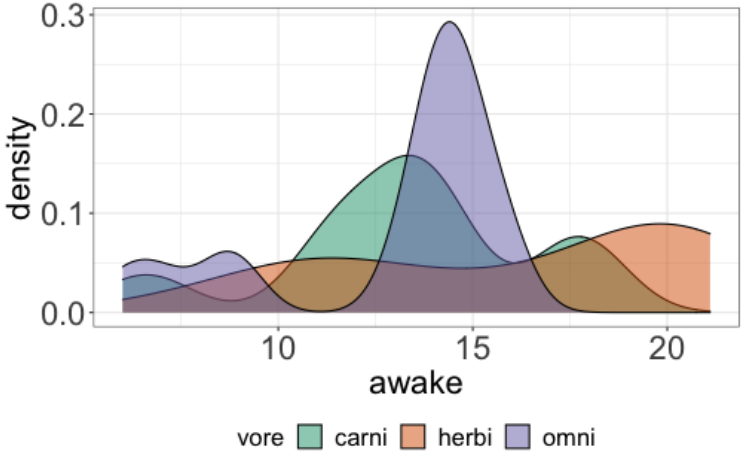

carni
herbi
omni
0.3
0.2
0.1
0.0
density
10
15
20
awake
vore
carni
herbi
omni

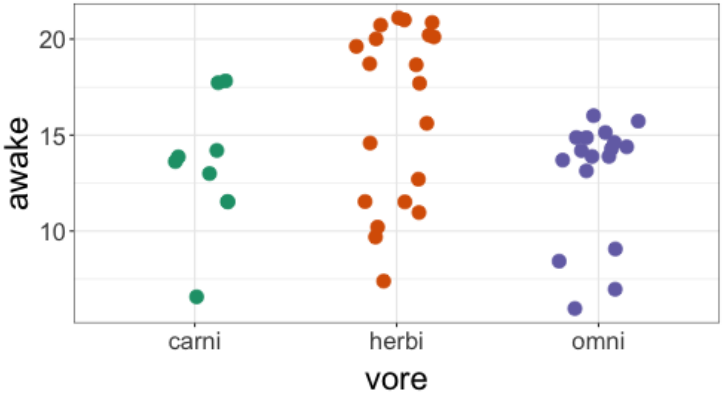

density
0.3
0.2
0.1
0.0
10
15
20
awake
vore
carni
herbi
omni

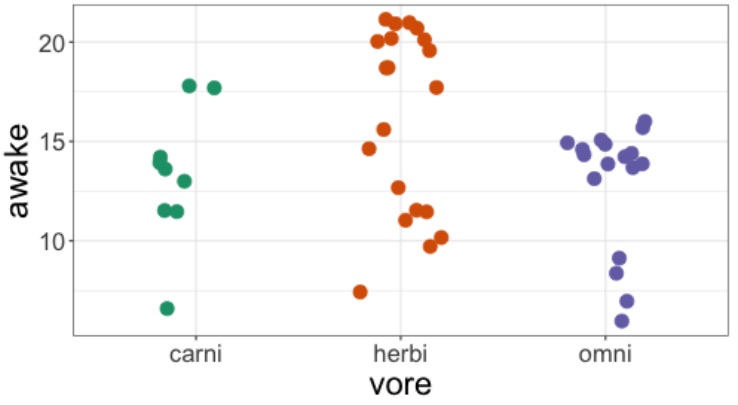

density
0.3
0.2
0.1
0.0
10
15
20
awake
vore
carni
herbi
omni

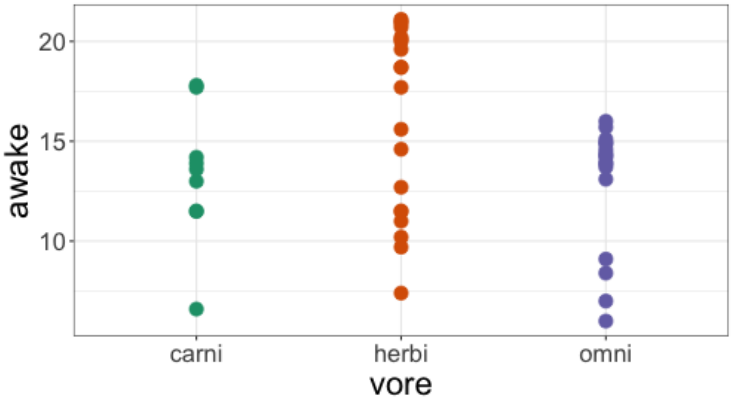

awake
20
15
10
carni
herbi
omni
vore

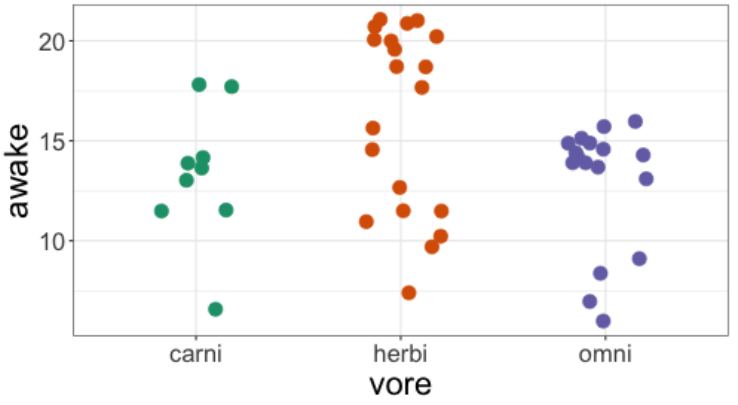

awake
20
15
10
carni
herbi
omni
vore
vore
carni
herbi
omni

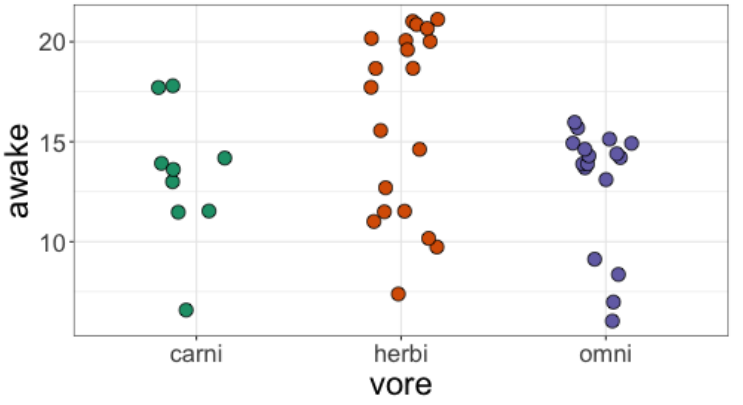

density
0.3
0.2
0.1
0.0
10
15
20
awake
vore
carni
herbi
omni


awake
20
15
10
carni
herbi
omni
vore

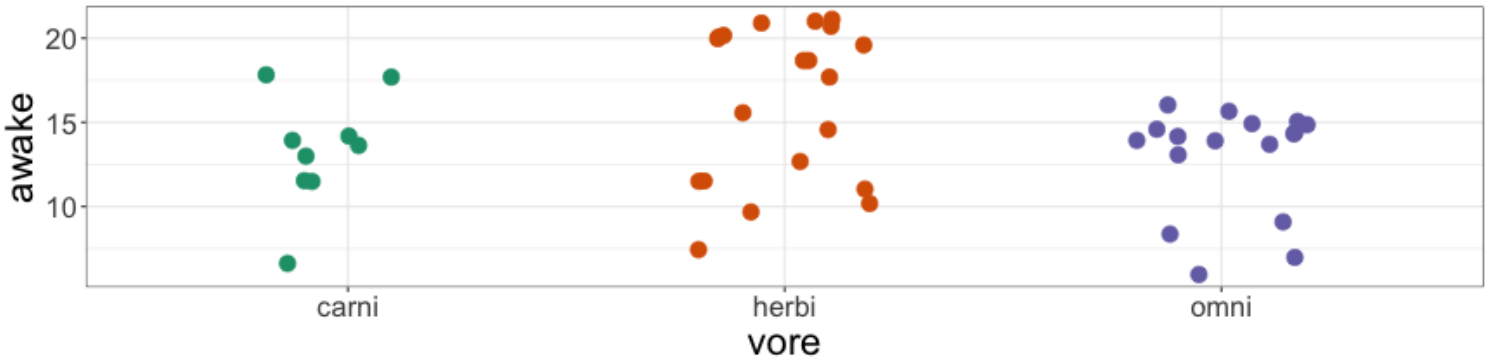

awake
20
15
10
carni
herbi
omni
vore

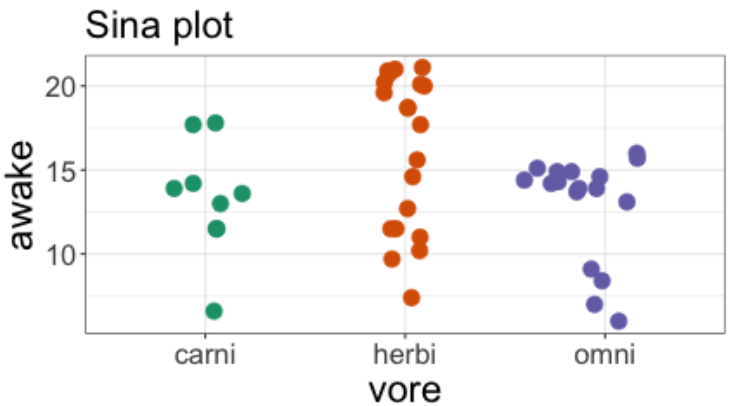

awake
20
15
10
carni
herbi
omni
vore

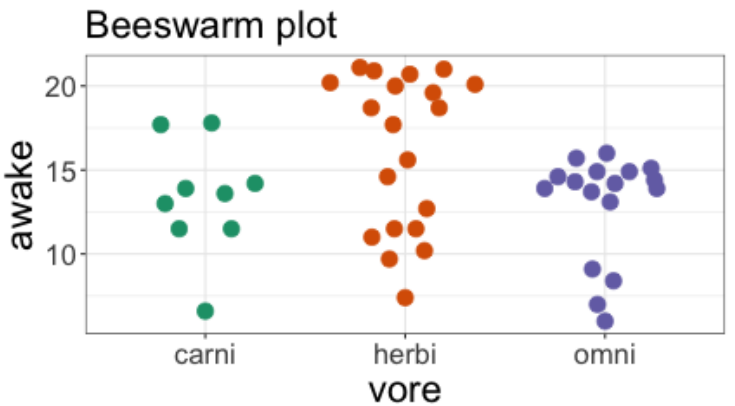

awake
20
15
10
carni
herbi
omni
vore


awake
20
15
10
carni
herbi
omni
vore

The world is your oyster

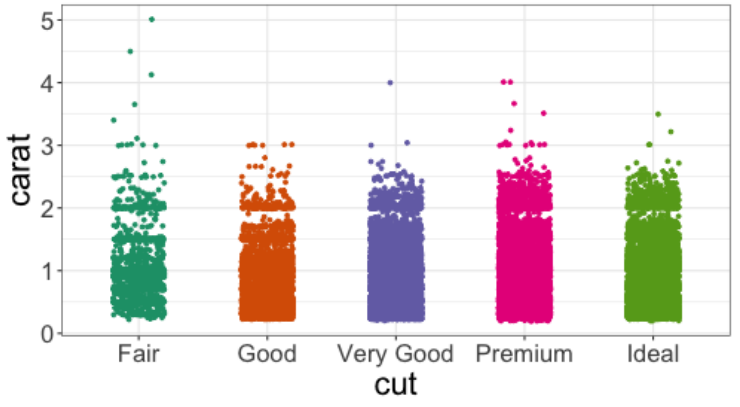


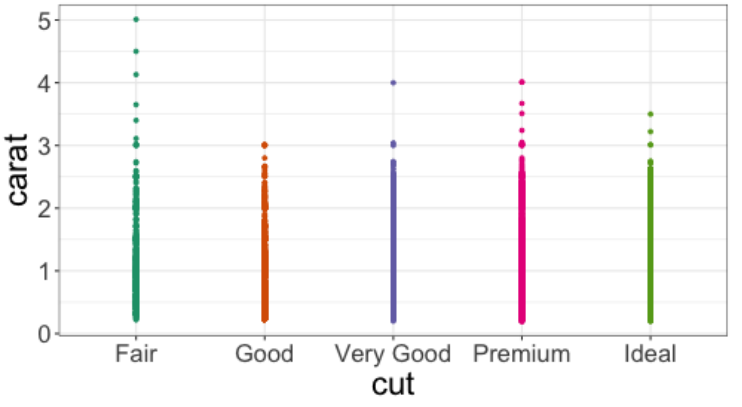


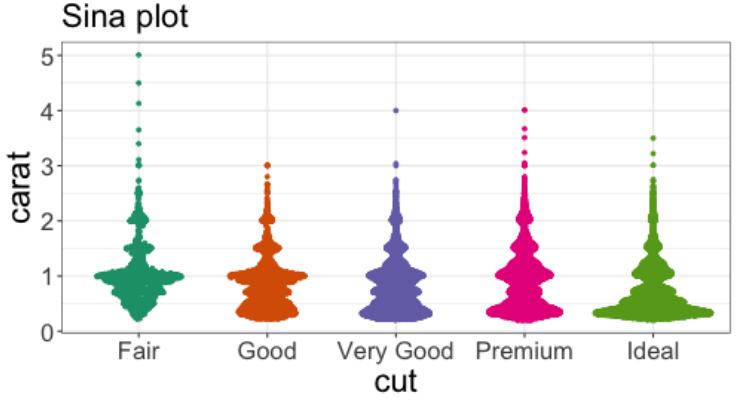


Ok, that wasn't super compelling...

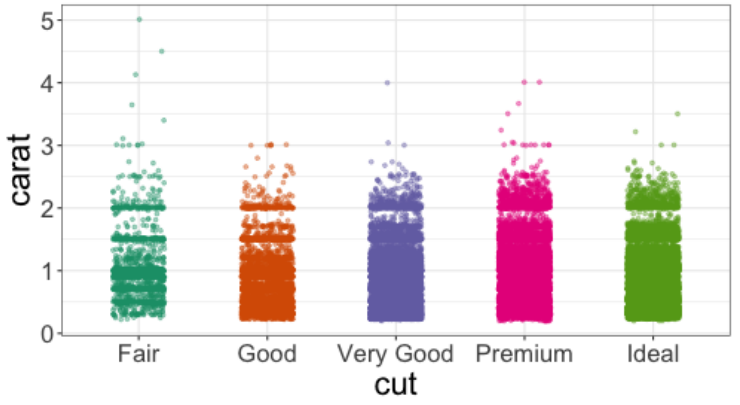
```
## # A tibble: 53,940 × 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
##  2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
##  3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
##  4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
##  5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
##  6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
##  7  0.24 Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
##  8  0.26 Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
##  9  0.22 Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
## 10  0.23 Very Good H     VS1      59.4    61   338  4     4.05  2.39
## # … with 53,930 more rows
```

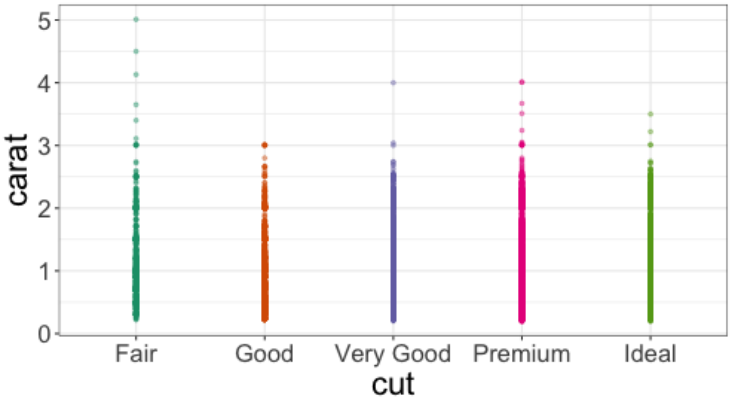

carat
0
1
2
3
4
5
Fair
Good
Very Good
Premium
Ideal
cut

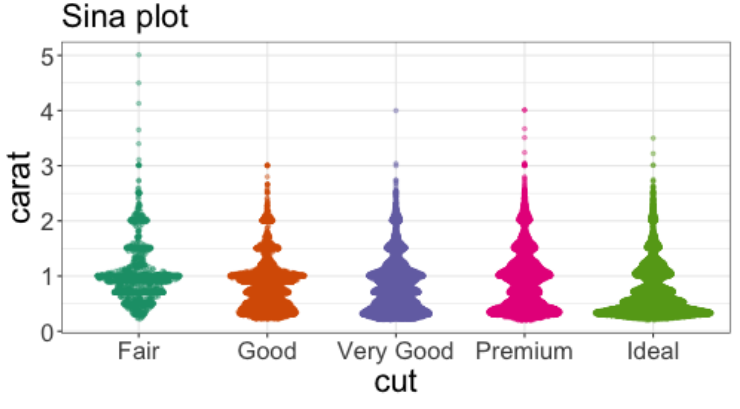

carat
0
1
2
3
4
5
Fair
Good
Very Good
Premium
Ideal
cut


Sina plot
carat
0
1
2
3
4
5
Fair
Good
Very Good
Premium
Ideal
cut

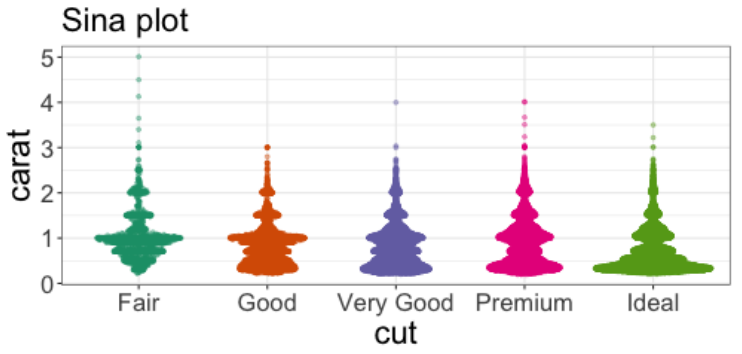

Beeswarm plot
carat
0
1
2
3
4
5
Fair
Good
Very Good
Premium
Ideal
cut

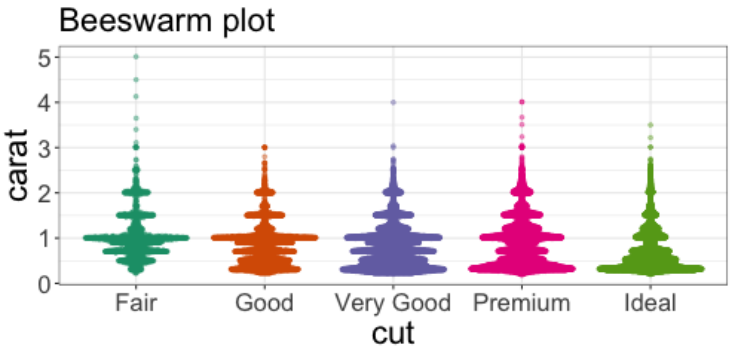

carat
5
4
3
2
1
0
Fair
Good
Very Good
Premium
Ideal
cut

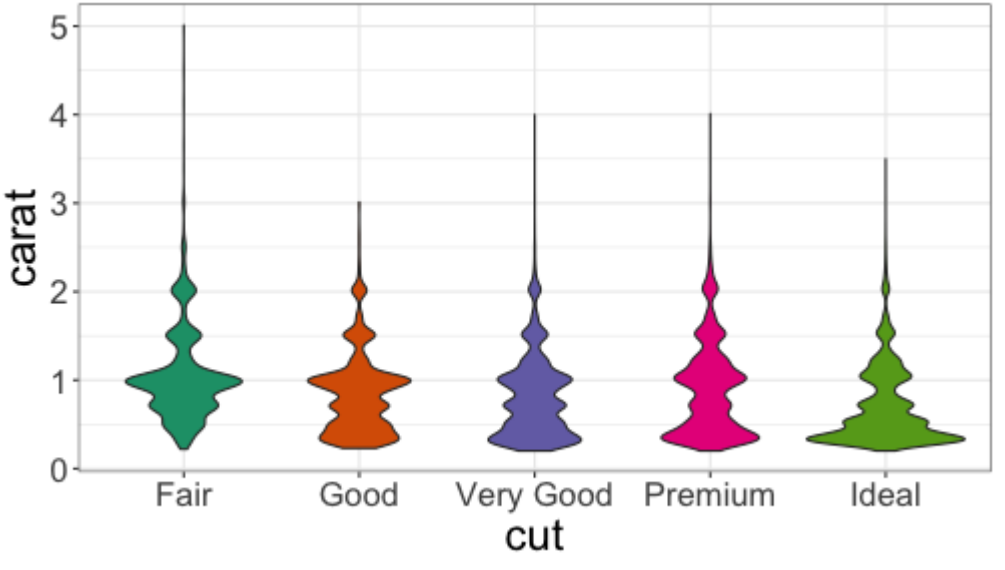

carat
5
4
3
2
1
0
Fair
Good
Very Good
Premium
Ideal
cut


Sina plot
carat
5
4
3
2
1
0
Fair
Good
Very Good
Premium
Ideal
cut

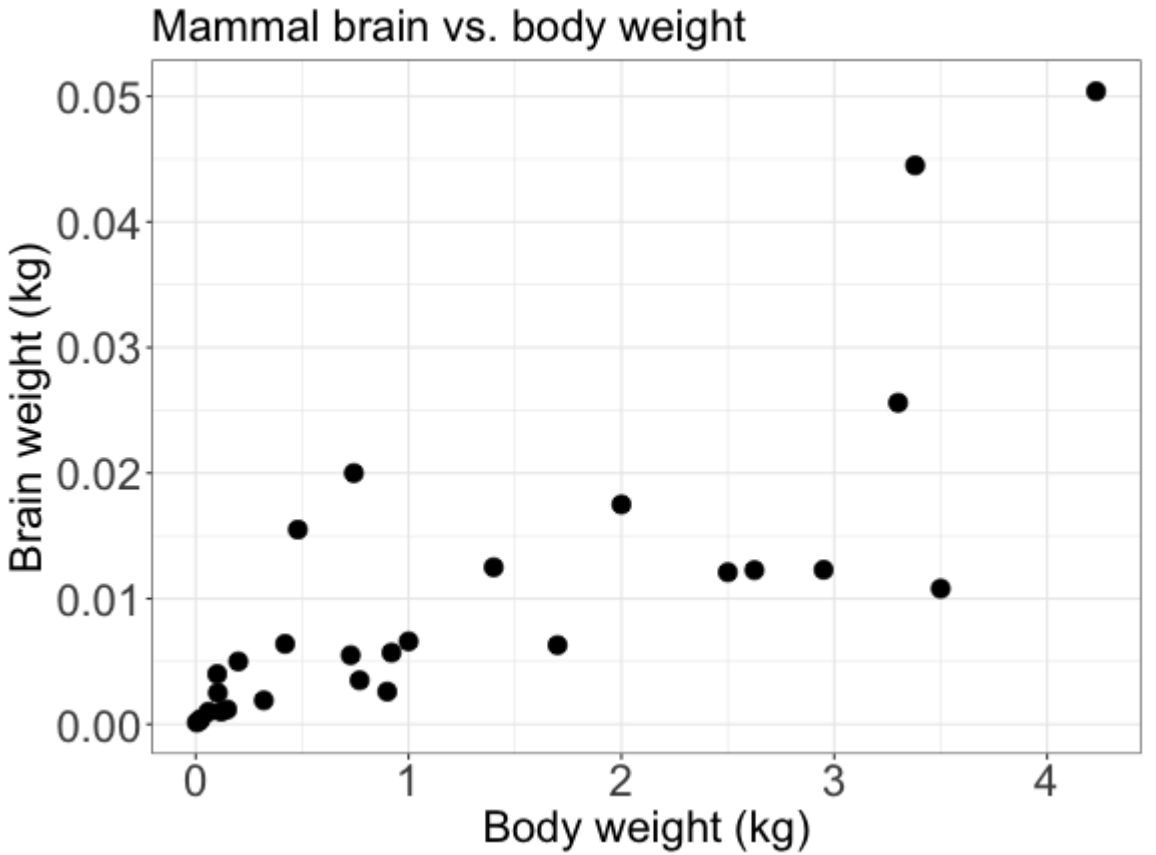

Beeswarm plot
carat
5
4
3
2
1
0
Fair
Good
Very Good
Premium
Ideal
cut

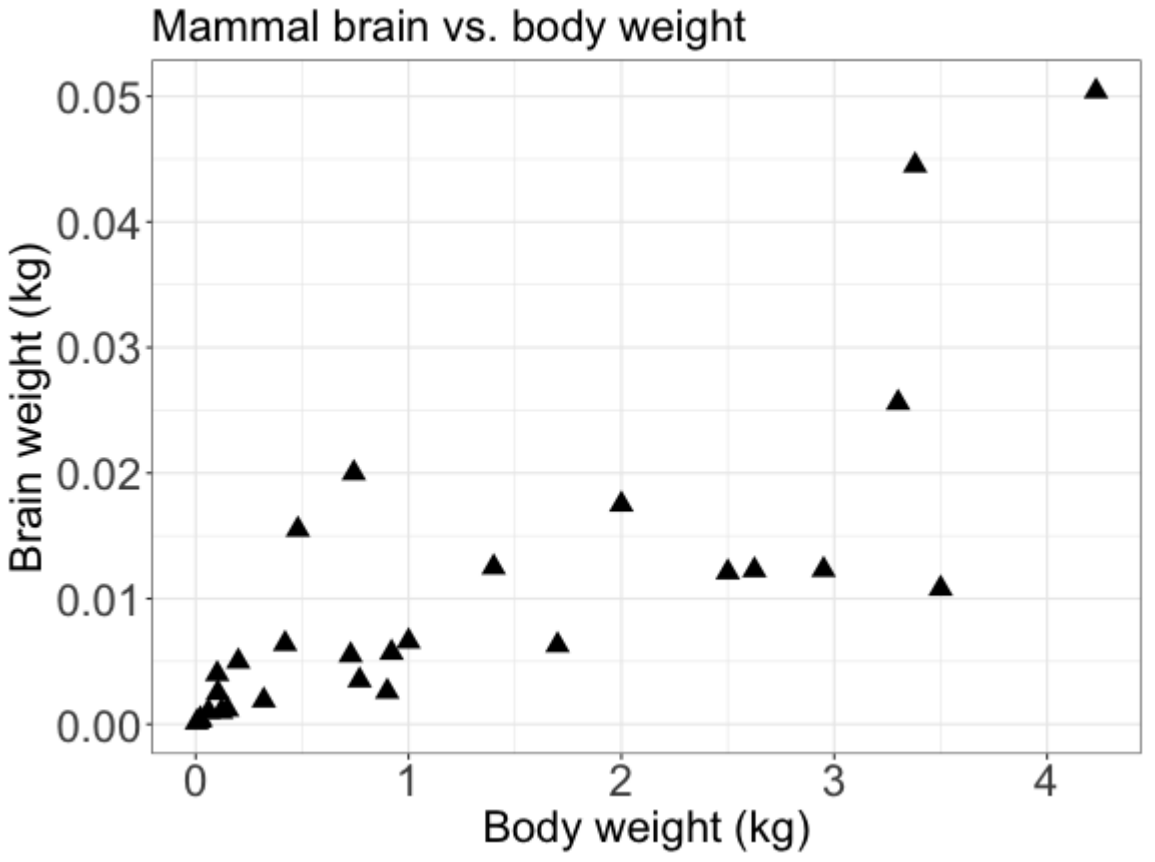

Sina plot
carat
5
4
3
2
1
0
Fair
Good
Very Good
Premium
Ideal
cut

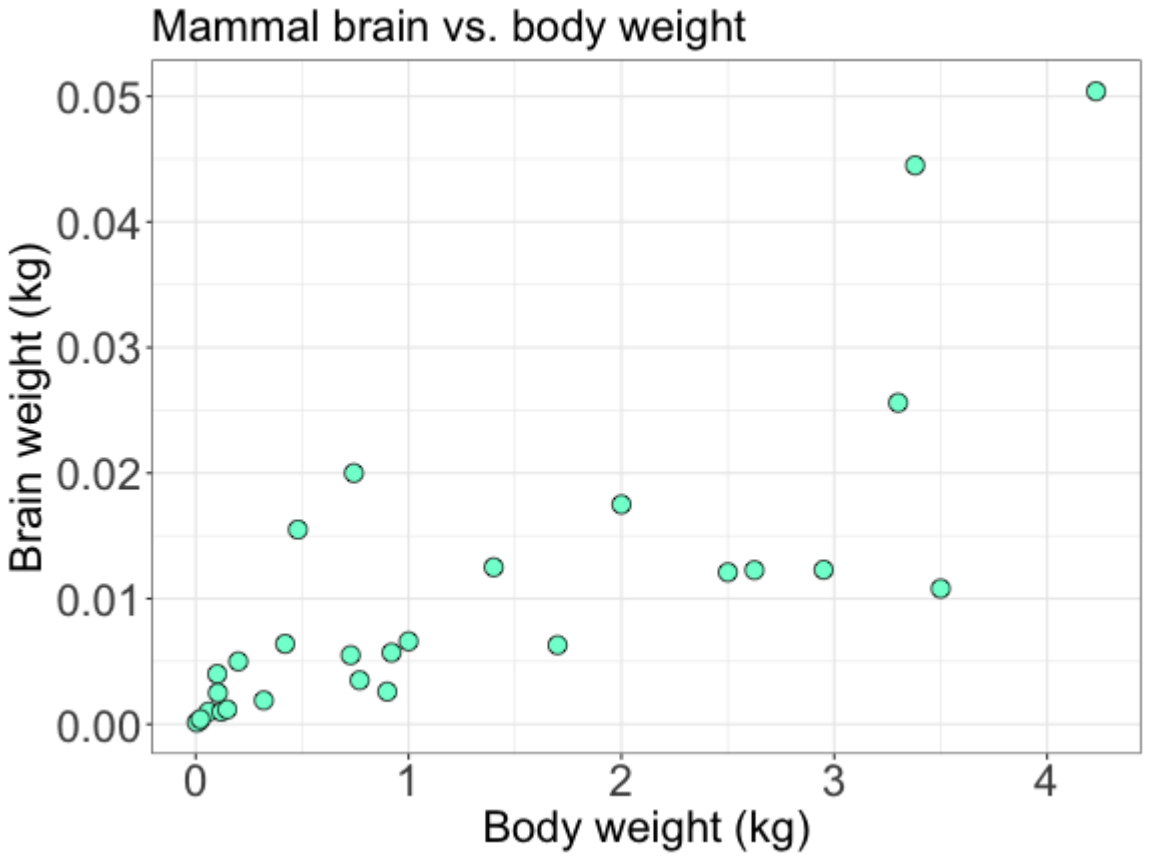

Beeswarm plot
carat
5
4
3
2
1
0
Fair
Good
Very Good
Premium
Ideal
cut

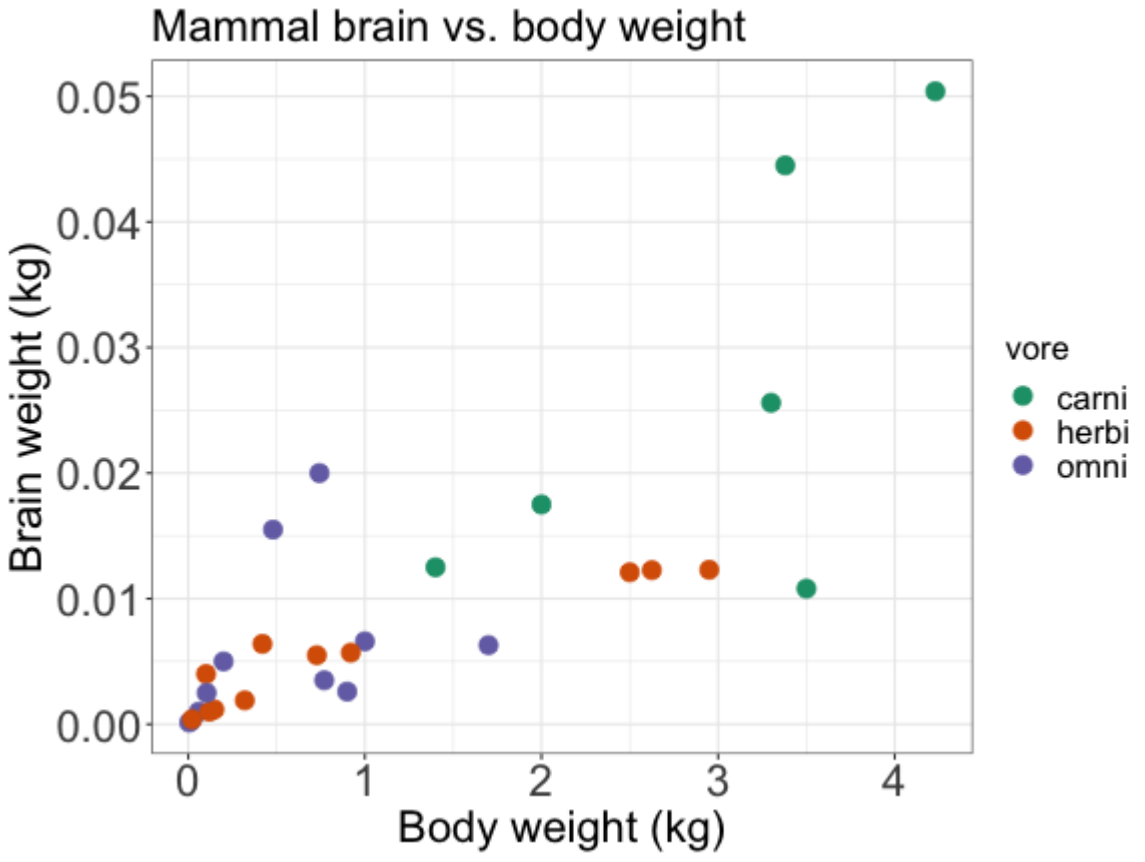

carat
5
4
3
2
1
0
Fair
Good
Very Good
Premium
Ideal
cut

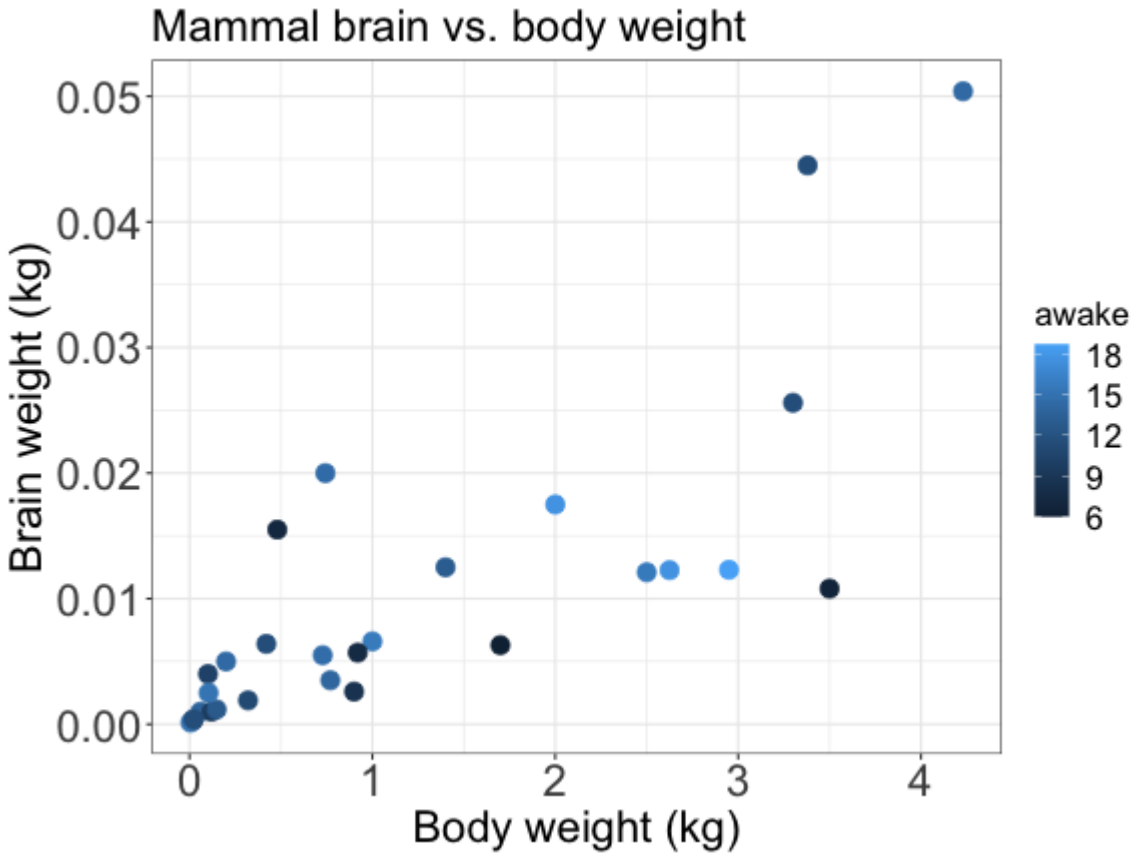
Onto scatterplots!

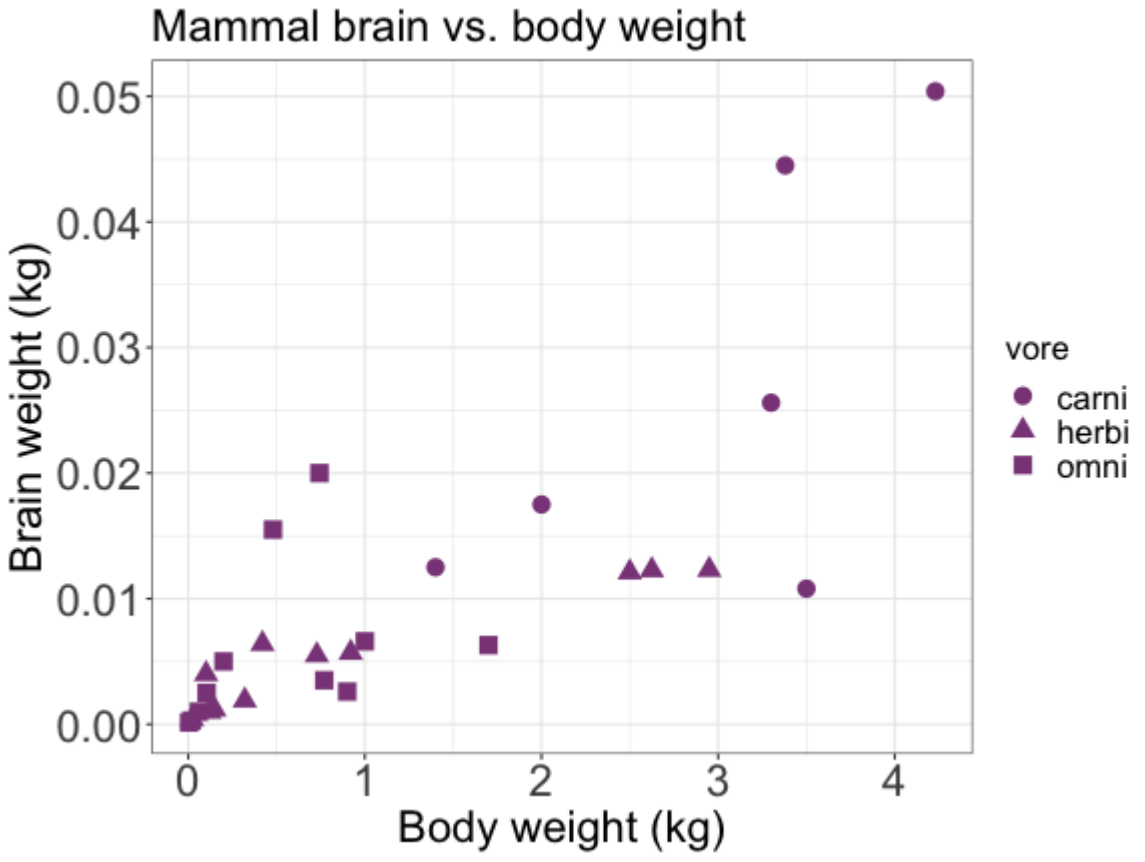


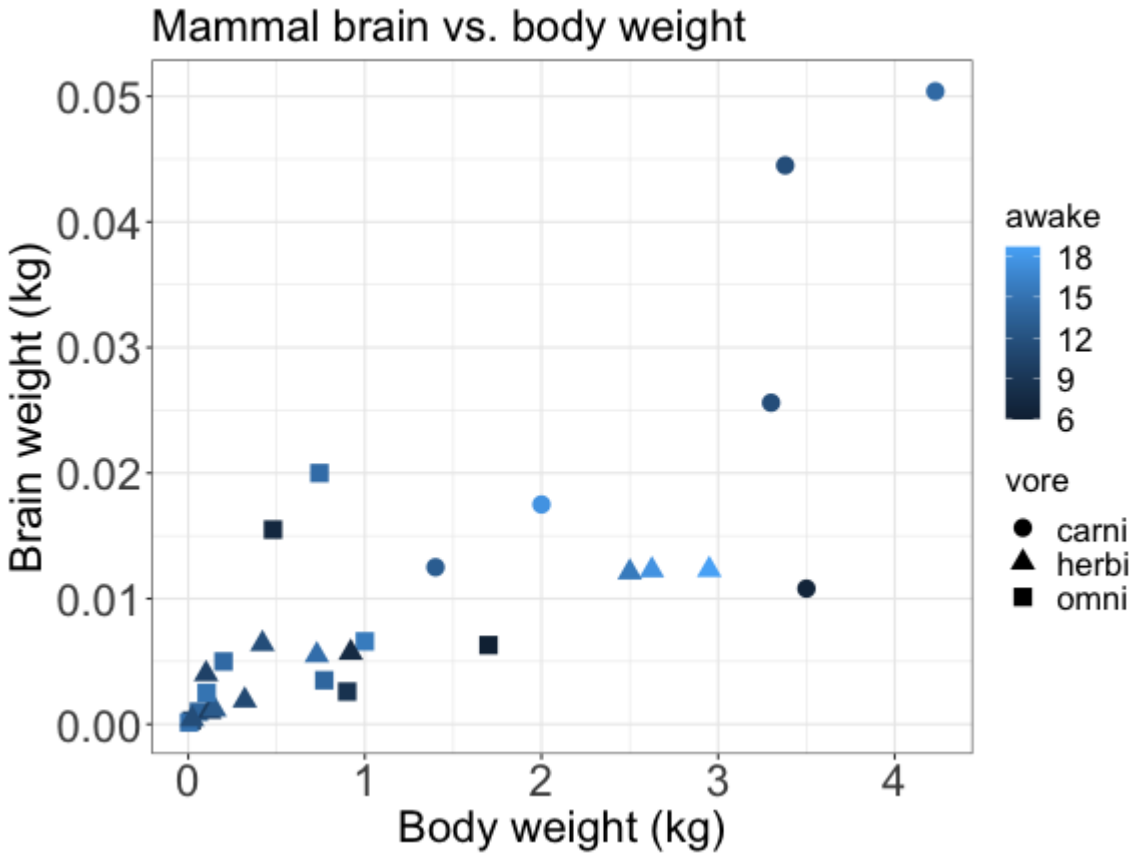

Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)


Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)

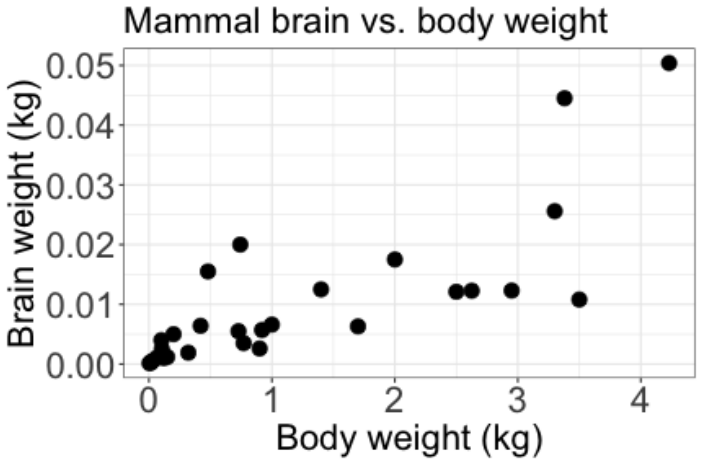

Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)
vore
carni
herbi
omni


Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)
awake
18
15
12
9
6

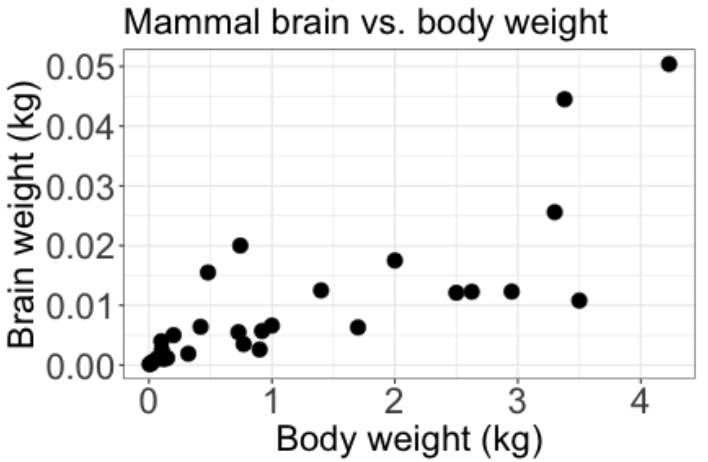

Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)
vore
carni
herbi
omni

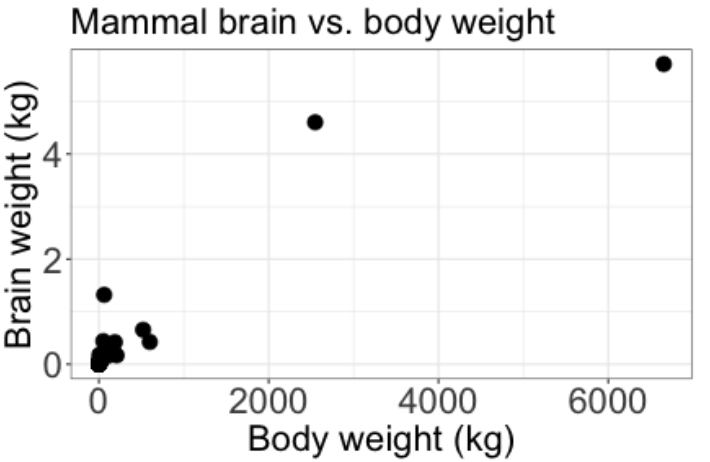

Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)
awake
18
15
12
9
6
vore
carni
herbi
omni

Do the axes look at all "strange" to you?

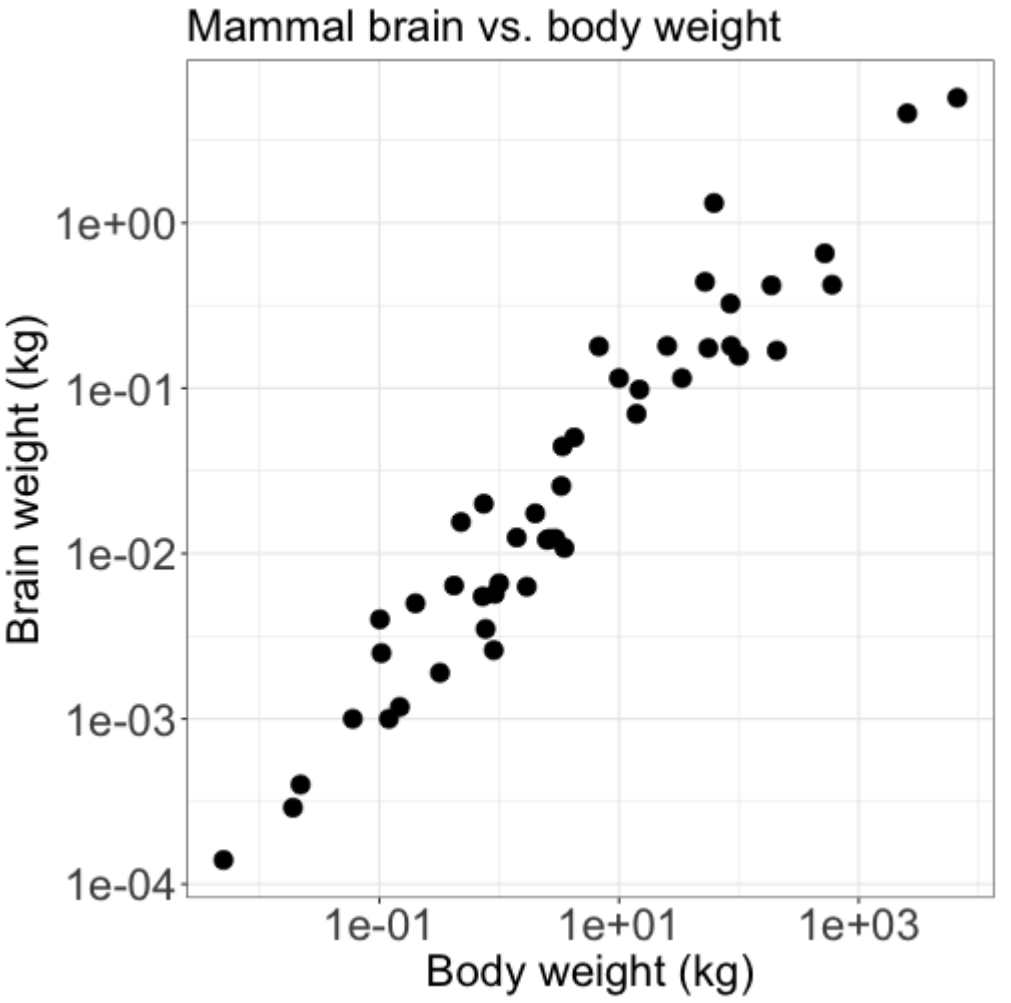


Do the axes look at all "strange" to you?





Use log scales for data with extreme ranges



Let's practice

- *Always start with axes*:
  - What variable is on the X-axis? What *type* of data is it? < br>
  - What variable is on the Y-axis? What *type* of data is it?

- Are there colors or fills? Are they "just colors" or are they *aesthetics*?

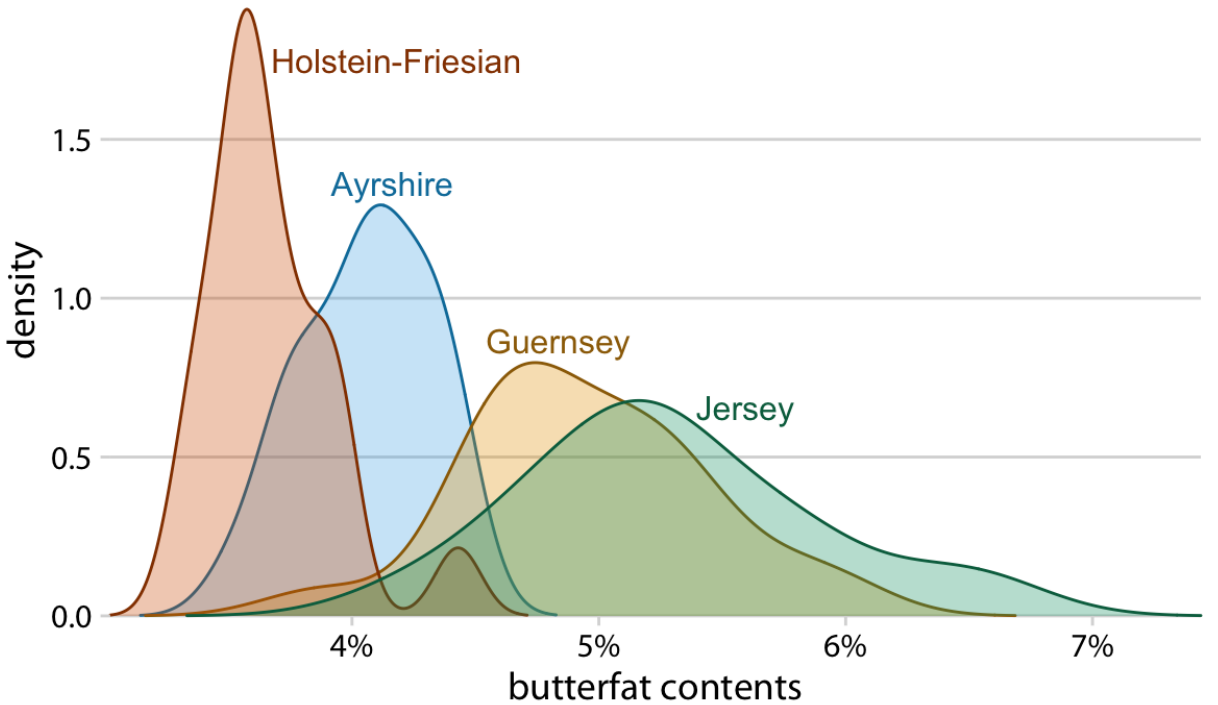
- What are the geometries in the plot?

- What *interpretations* can we make about the plot? What question(s) does the plot address or not address? (there are MANY right answers here!).

- What might the underlying dataset actually look like? *What variables (columns) are likely present?*

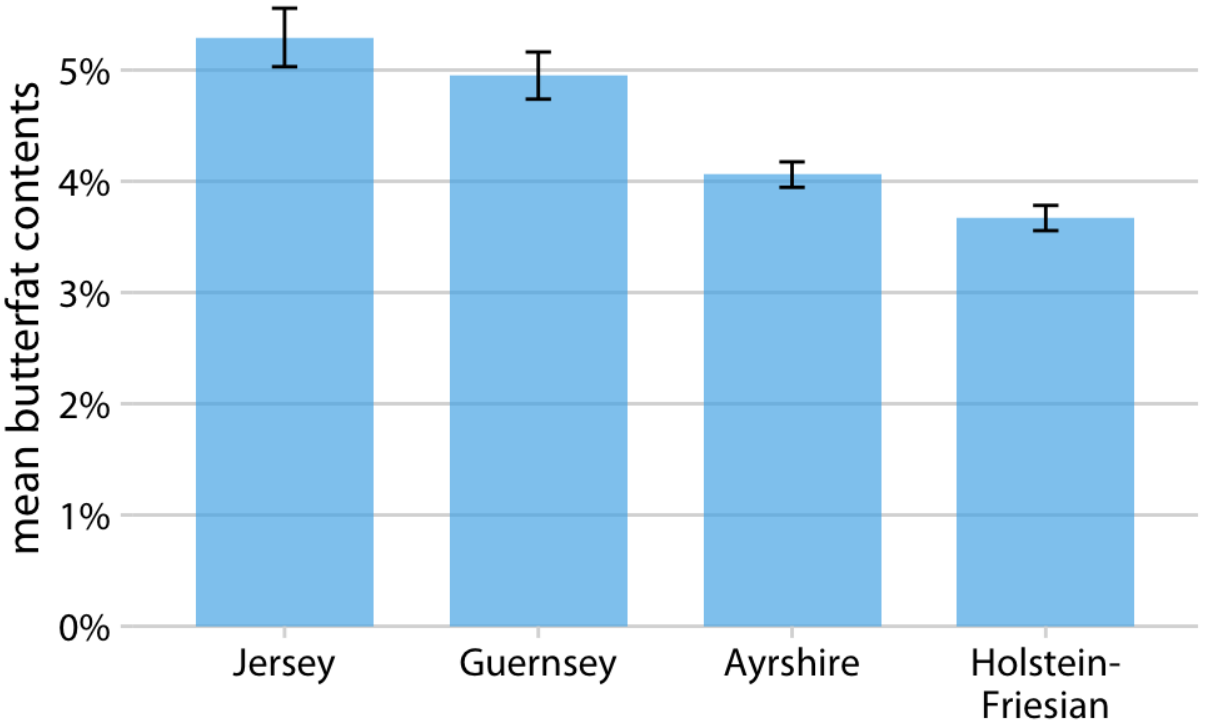
All figures in the following slides are from Fundamentals of Data Visualization.

Butterfat from different cows

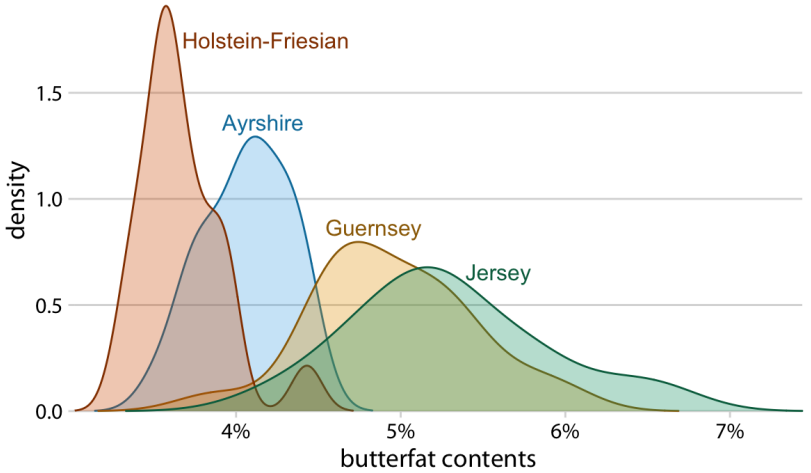


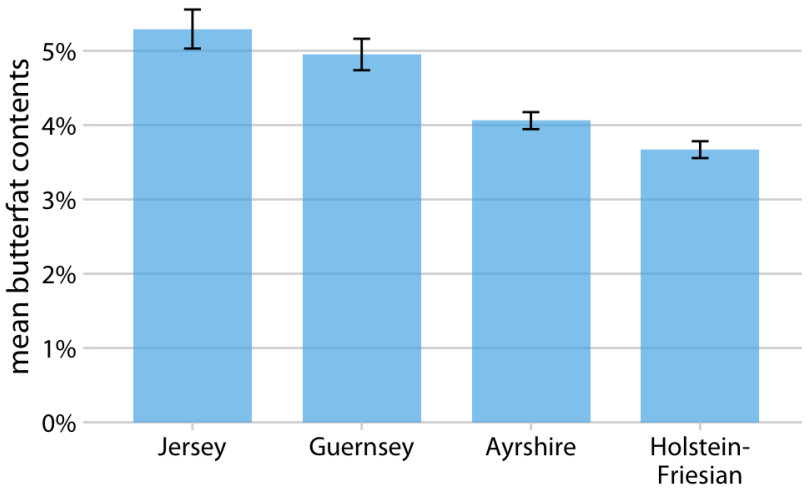
Density estimates of the butterfat percentage in the milk of four cattle breeds.
Data Source: Canadian Record of Performance for Purebred Dairy Cattle.

Butterfat from different cows, as bars

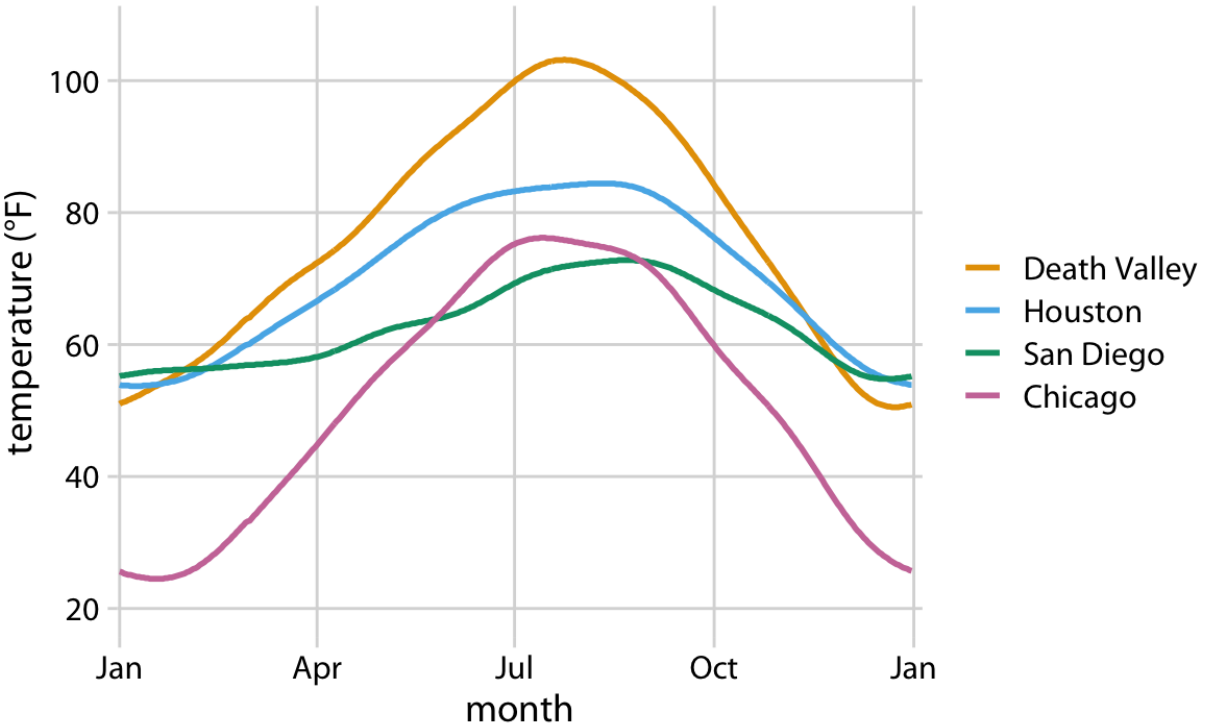


Let's compare:



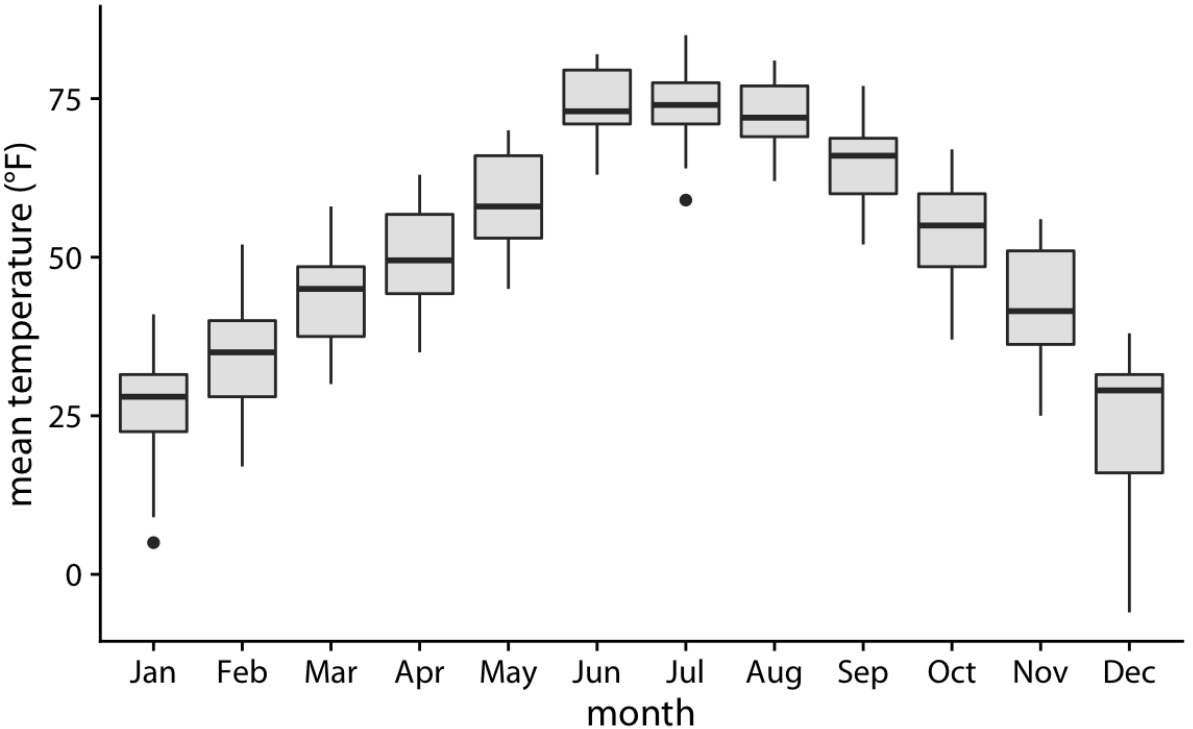


Average daily temperatures

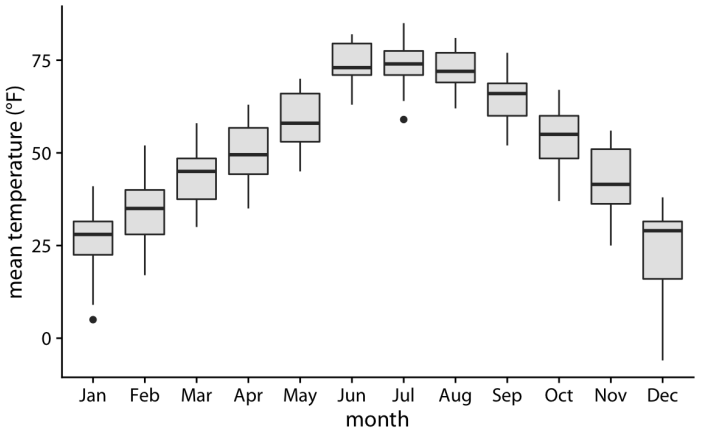


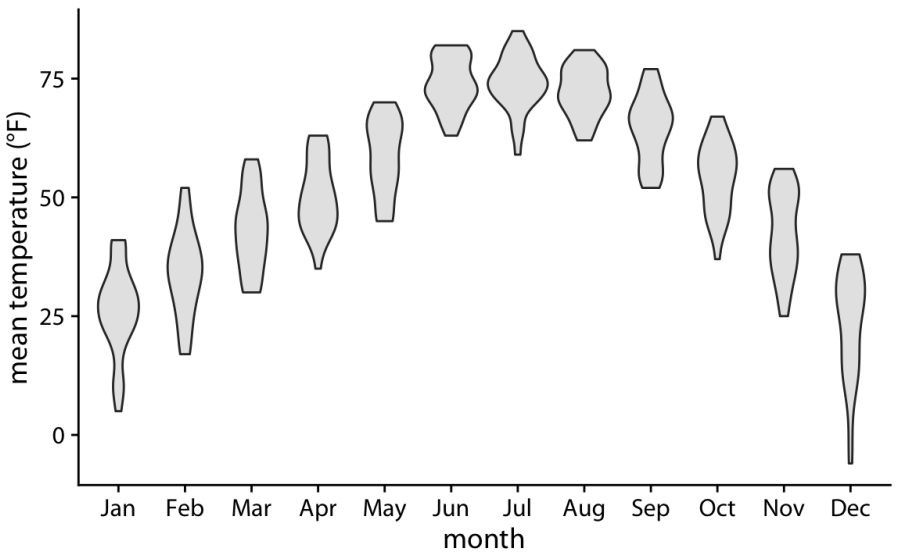
Daily temperature normals for four selected locations in the U.S. Temperature is mapped to the y axis, day of the year to the x axis, and location to line color. Data source: NOAA.

Temperatures in Lincoln, NE in 2016

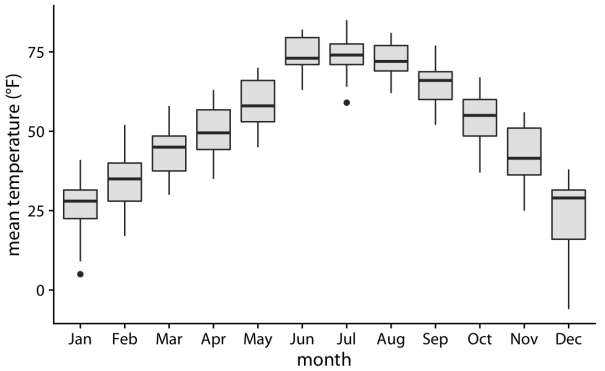


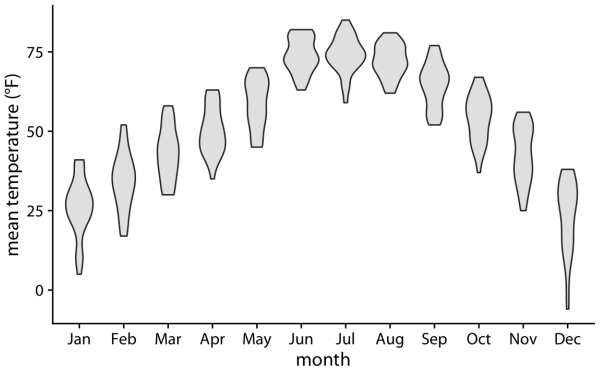
Temperatures in Lincoln, NE in 2016

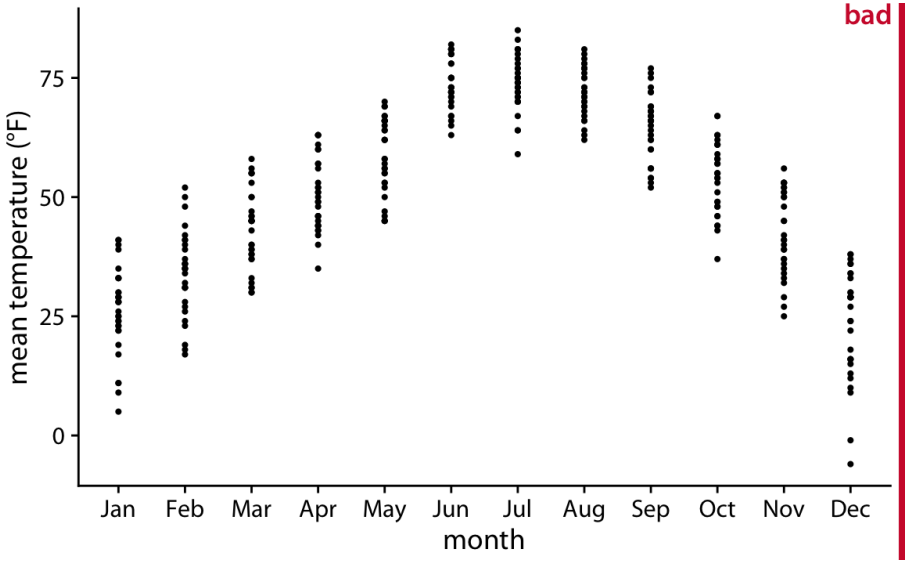




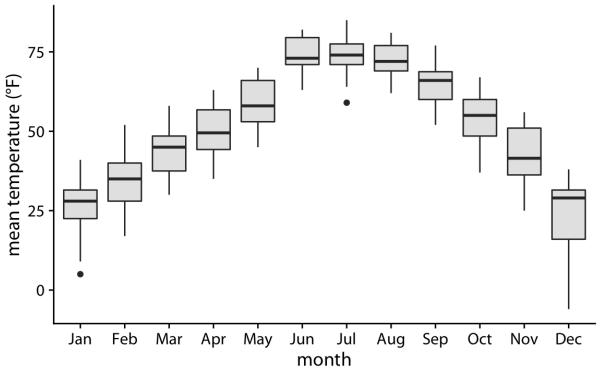
Temperatures in Lincoln, NE in 2016

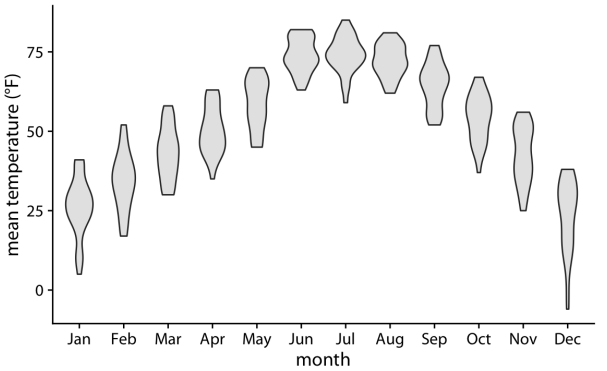


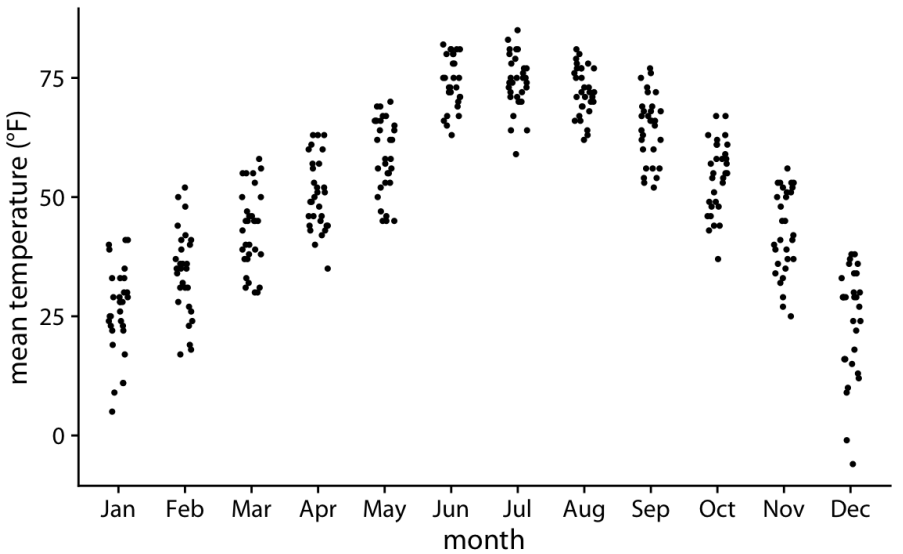




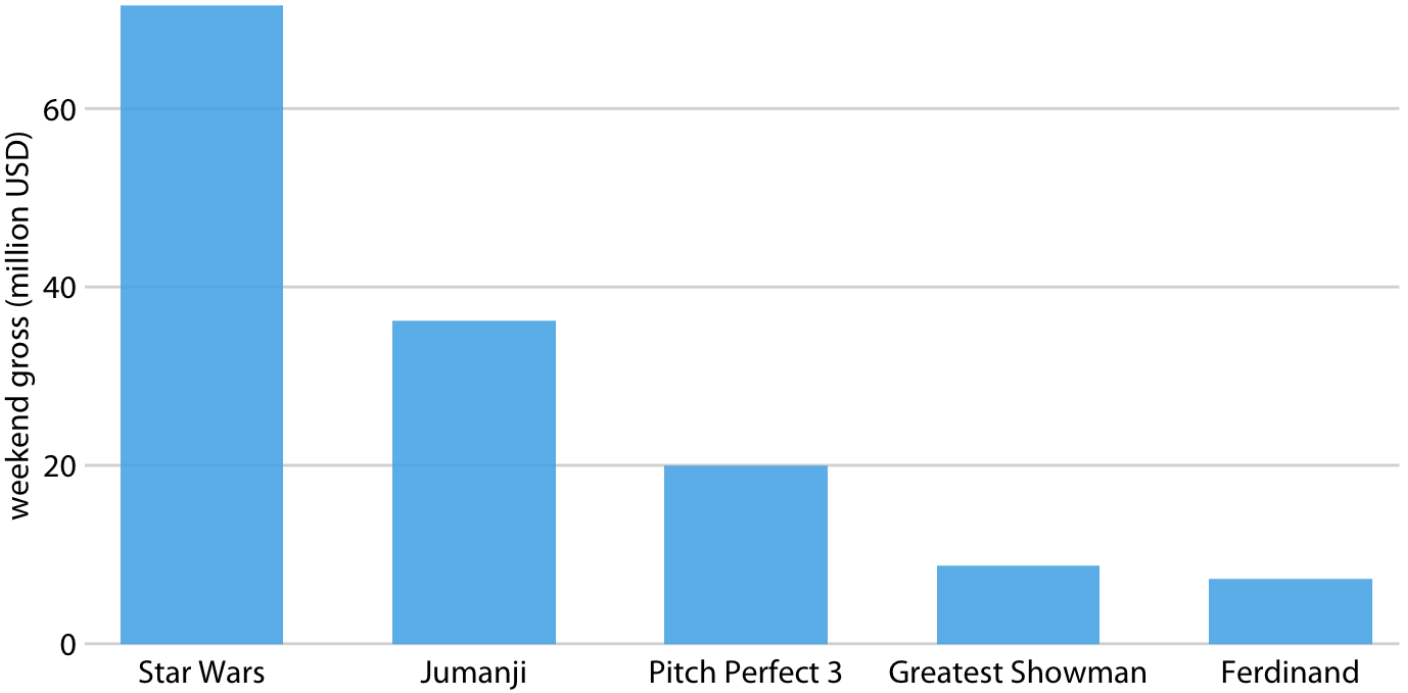
Temperatures in Lincoln, NE in 2016





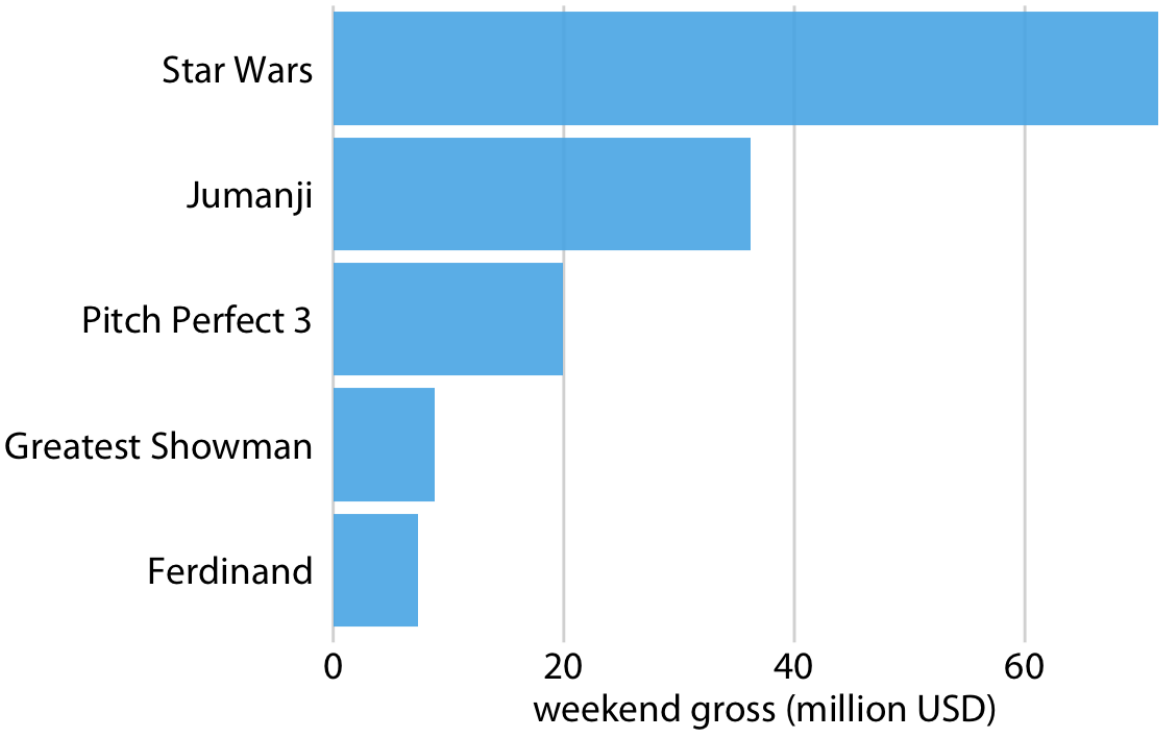


Box office income

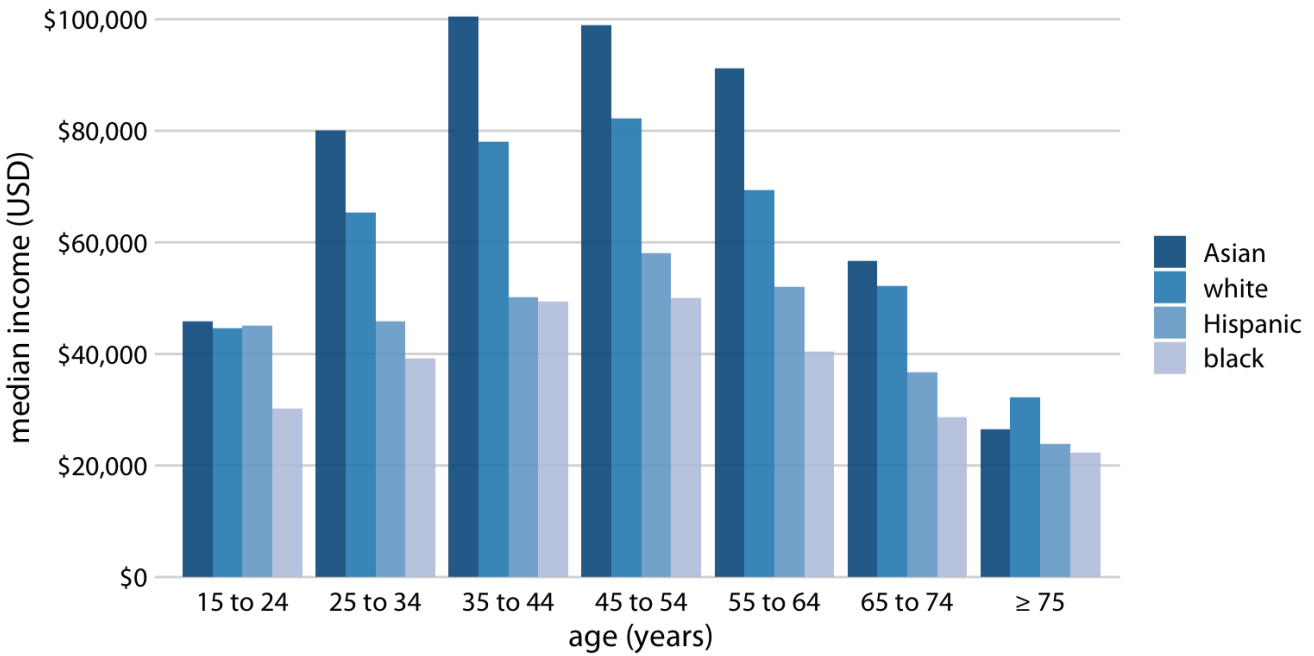


Highest grossing movies for the weekend of December 22-24, 2017. Data source: Box Office Mojo.

Box office income - what's different?

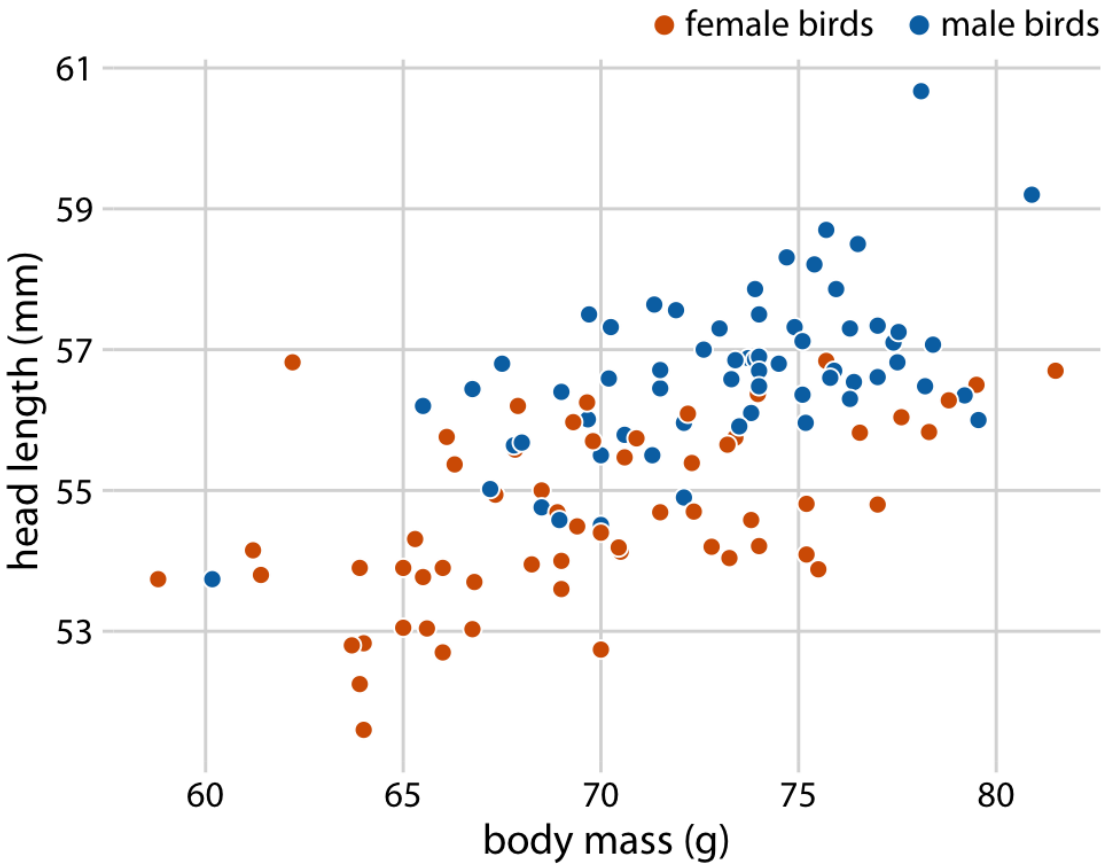


Median household income



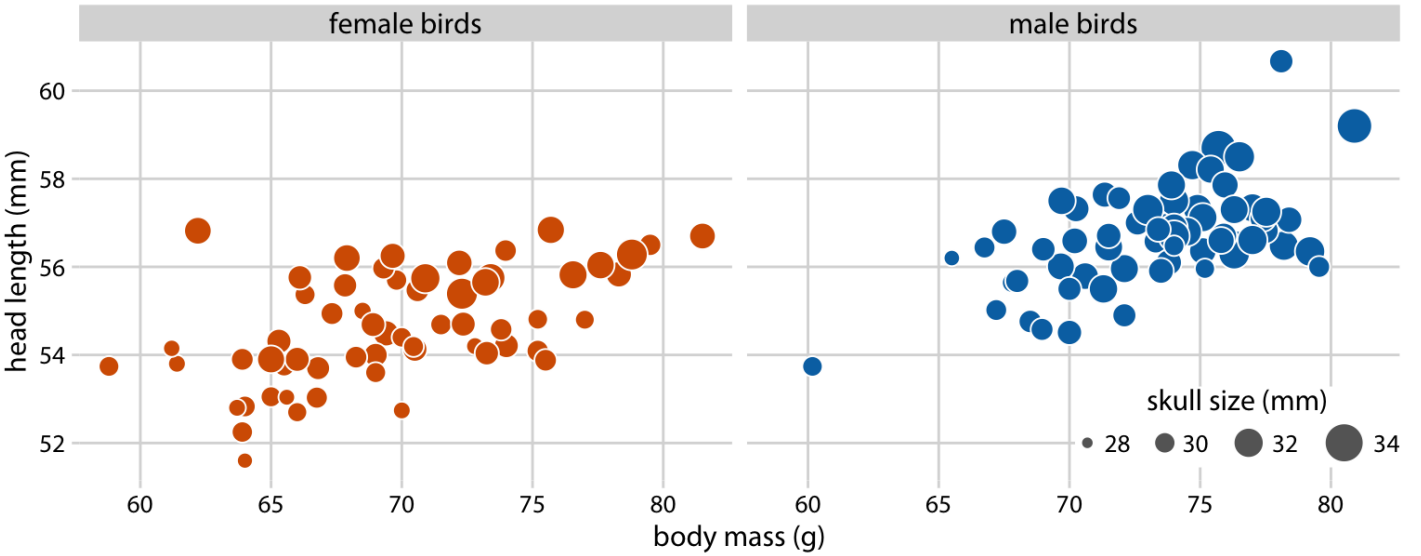
2016 median U.S. annual household income versus age group and race. For each age group there are four bars, corresponding to the median income of Asian, white, Hispanic, and black people, respectively. Data source: United States Census Bureau.

Bluejays

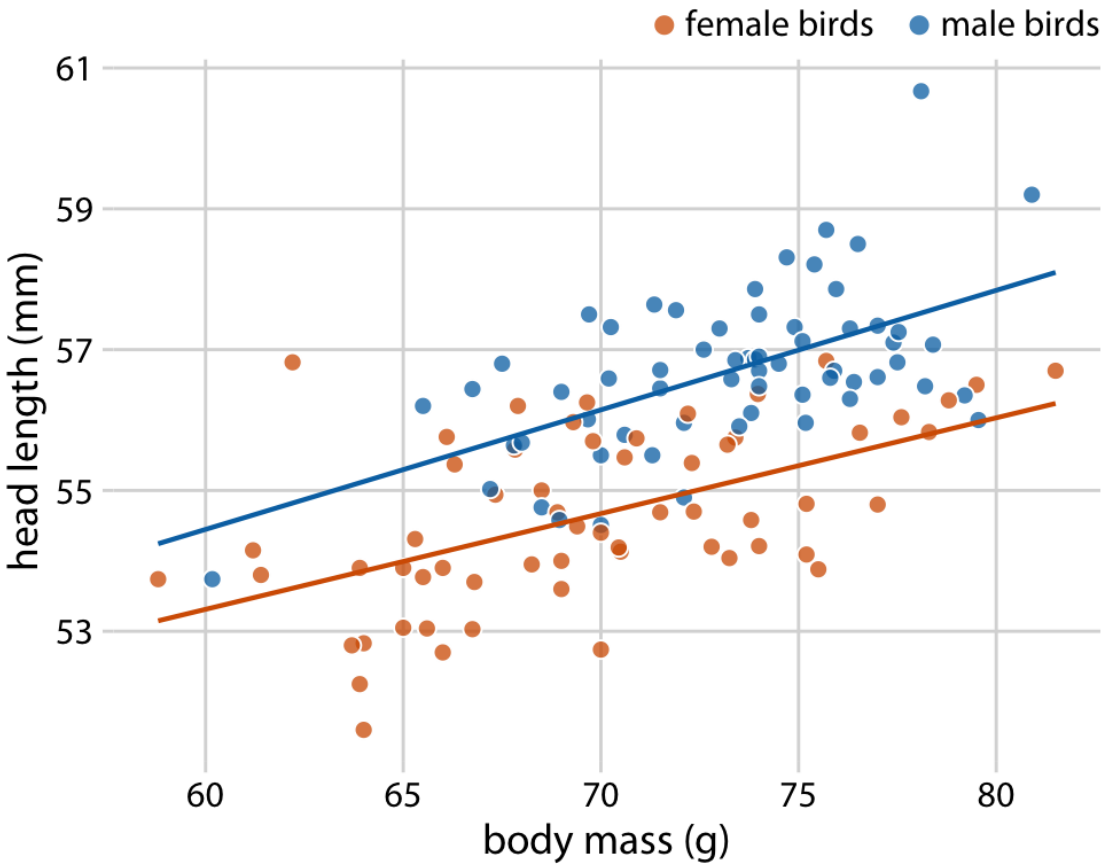


Head length versus body mass for 123 blue jays. The birds' sex is indicated by color. Data source: Keith Tarvin, Oberlin College.

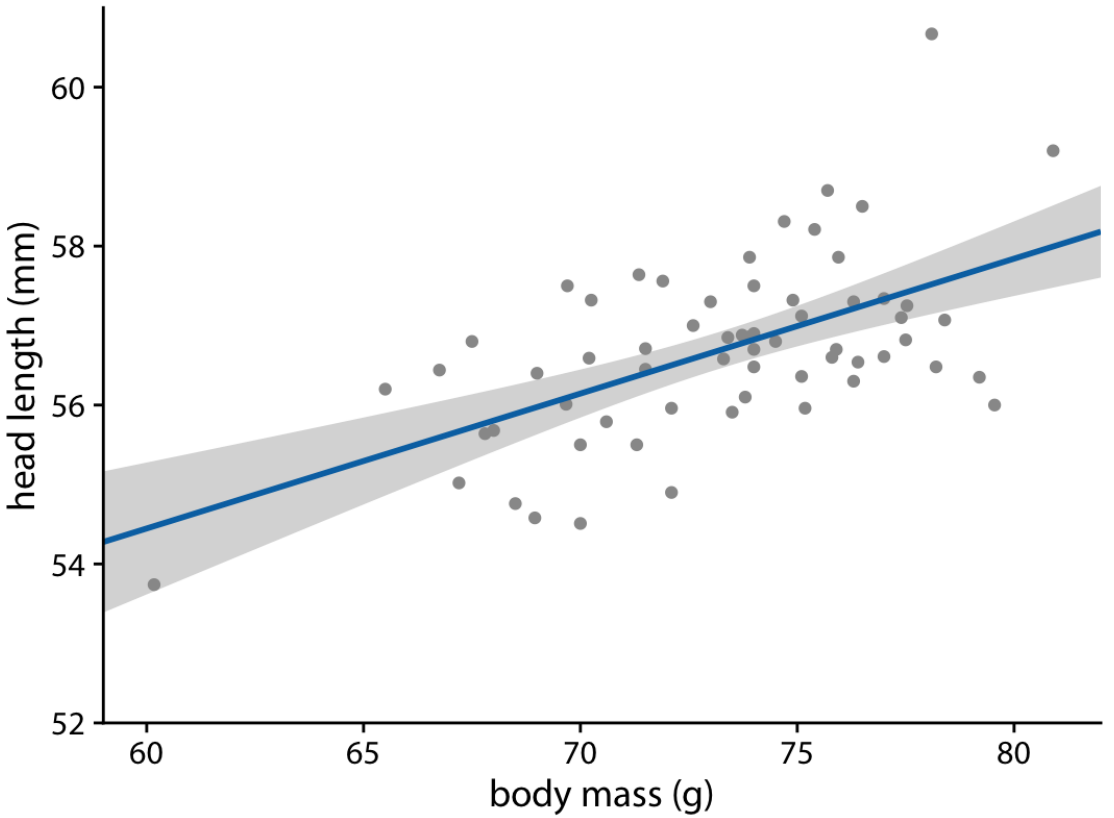
Bluejays, redux 1



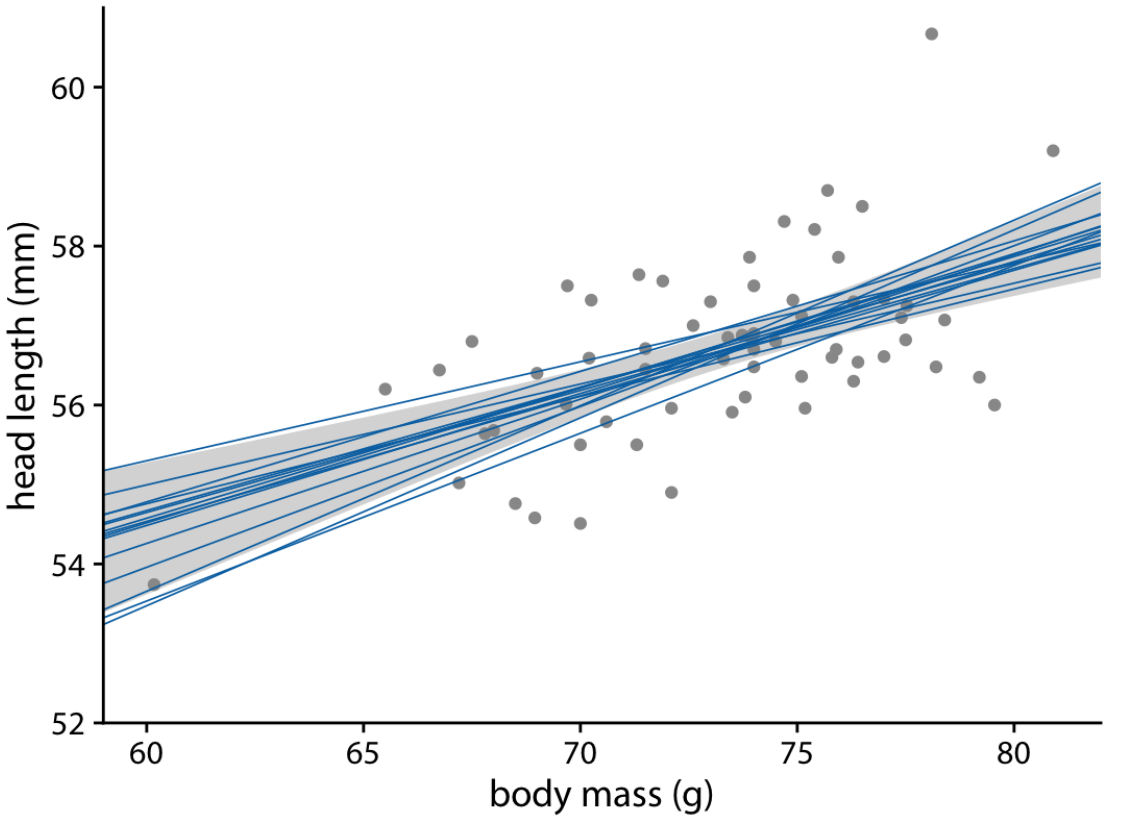
Bluejays, redux 2



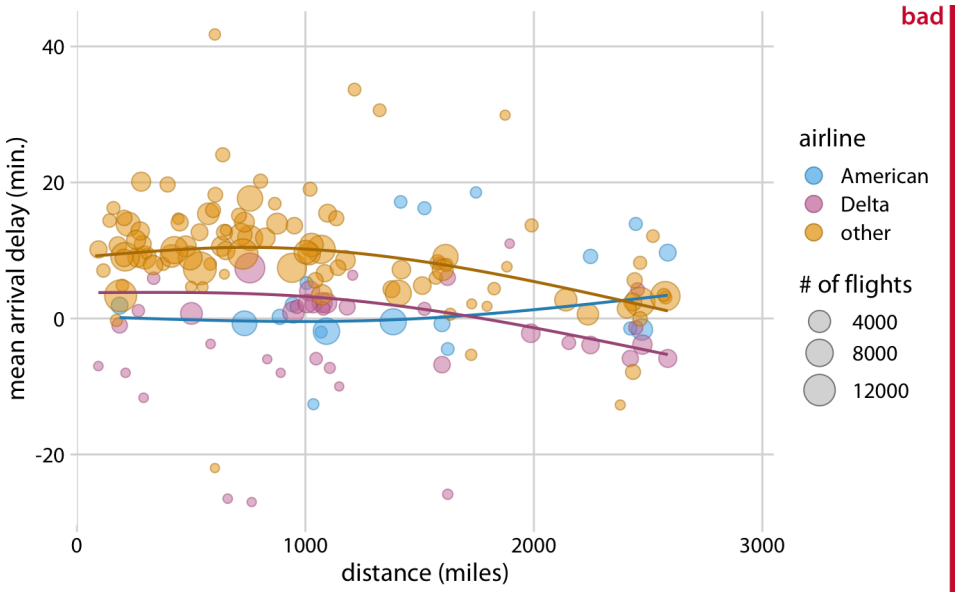
Male bluejays only



Getting an intuition for the *confidence interval*



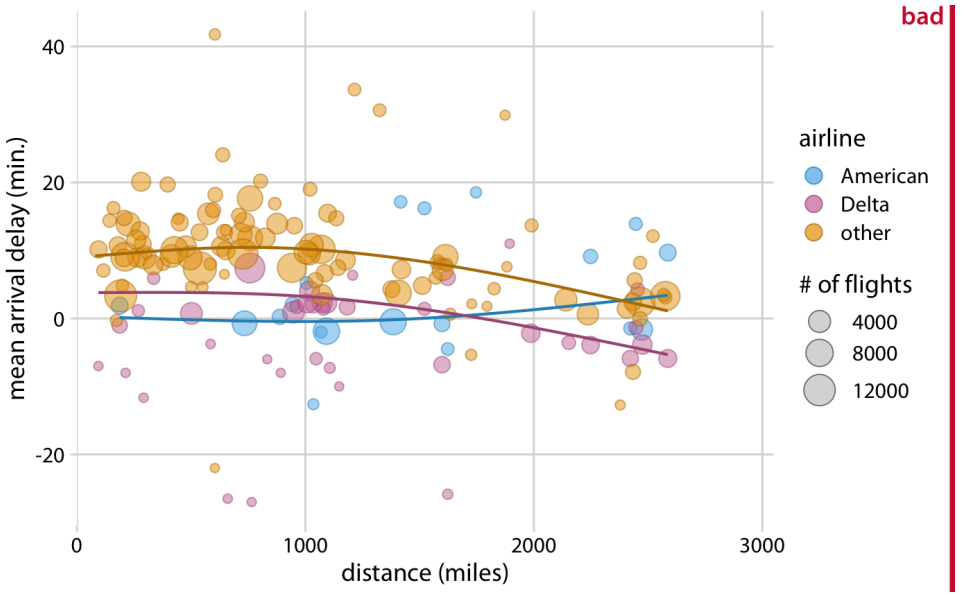
Airplane delays



Mean arrival delay versus distance from New York City. Data source: U.S. Dept. of Transportation, Bureau of Transportation Statistics.

This figure is labeled as “bad” because it is overly complex. Most readers will find it confusing and will not intuitively grasp what it is the figure is showing.

Airplane delays



Mean arrival delay versus distance from New York City. Data source: U.S. Dept. of Transportation, Bureau of Transportation Statistics.

"Looking cool/smart" is NOT the same as effectively communicating. We'll talk more about data viz style and best practices after we start learning how to plot in R next week!