The grammar of graphics

Data Science for Biologists

Image collage by Lalita Martin

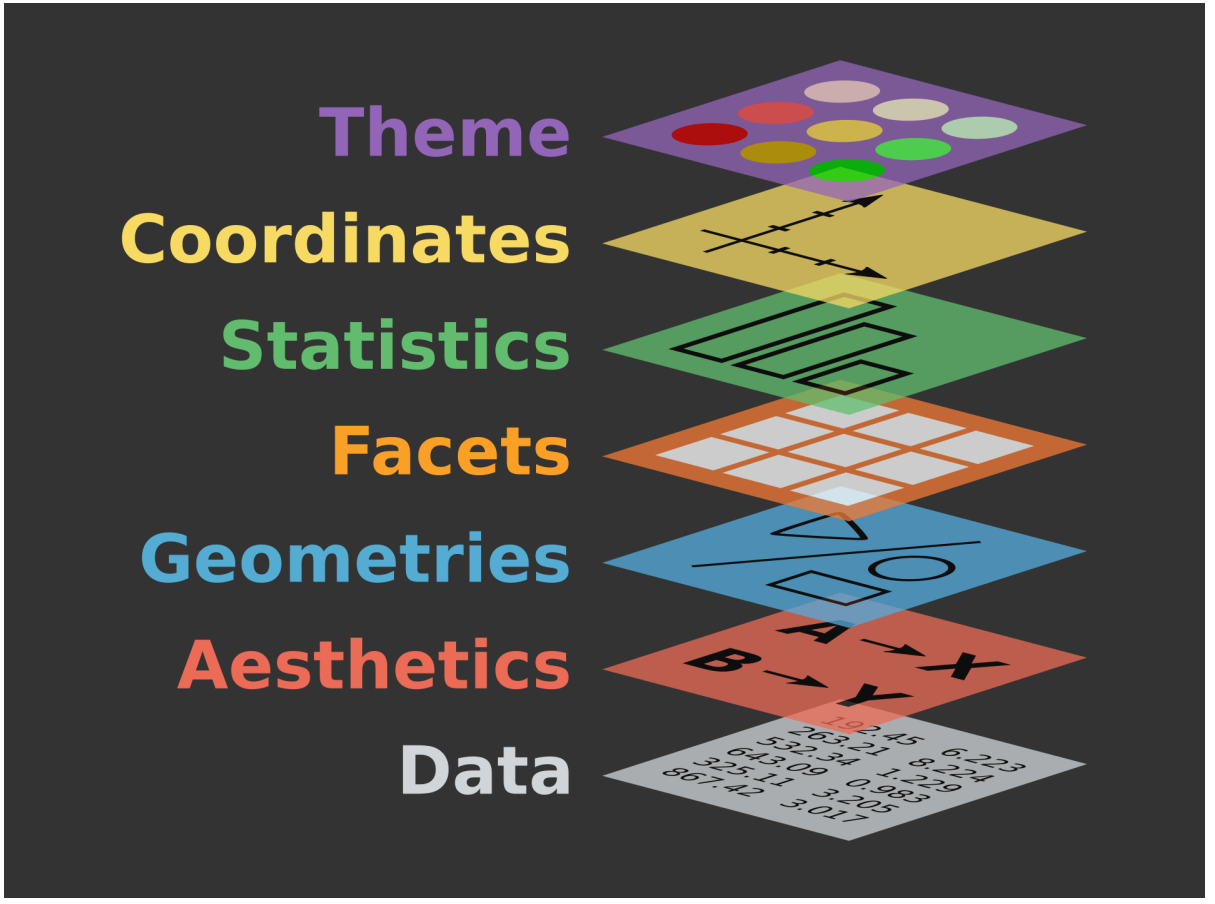
Grammar

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 x 5
##    name                vore  awake brainwt  bodywt
##    <chr>               <fct> <dbl>   <dbl>   <dbl>
##  1 Owl monkey          omni    7   0.0155    0.48
##  2 Greater short-tai… 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 pou… omni   15.7 0.0066    1
## # … with 36 more rows
```

The dataset

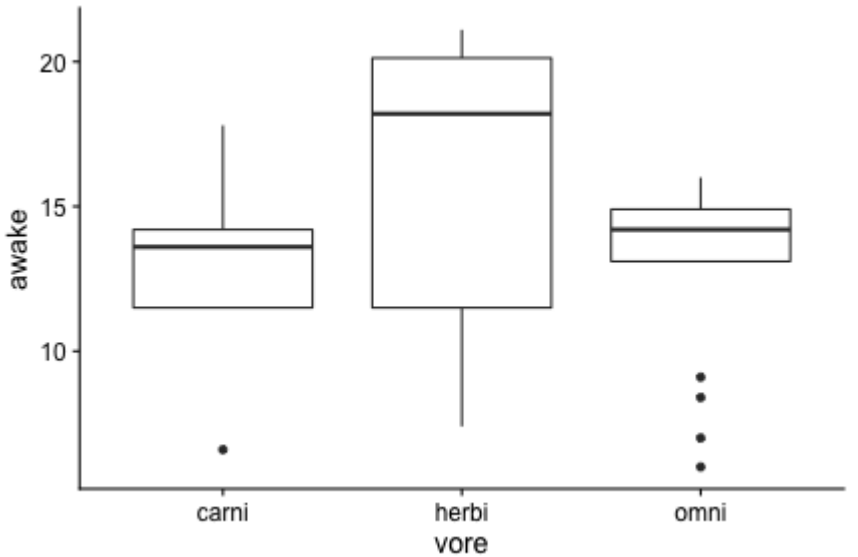
```
summary(msleep_subvore)
```

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

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

```
##  [1] "Owl monkey"
##  [2] "Greater short-tailed shrew"
##  [3] "Cow"
##  [4] "Dog"
##  [5] "Roe deer"
##  [6] "Goat"
##  [7] "Guinea pig"
##  [8] "Chinchilla"
##  [9] "Star-nosed mole"
## [10] "African giant pouched rat"
## [11] "Lesser short-tailed shrew"
## [12] "Long-nosed armadillo"
## [13] "Tree hyrax"
## [14] "North American Opossum"
## [15] "Asian elephant"
## [16] "Horse"
## [17] "Donkey"
## [18] "European hedgehog"
## [19] "Patas monkey"
## [20] "Domestic cat"
## [21] "Galago"
## [22] "Gray seal"
## [23] "Gray hyrax"
## [24] "Human"
## [25] "African elephant"
## [26] "Macaque"
## [27] "Golden hamster"
## [28] "House mouse"
## [29] "Slow loris"
## [30] "Rabbit"
## [31] "Sheep"
## [32] "Chimpanzee"
```

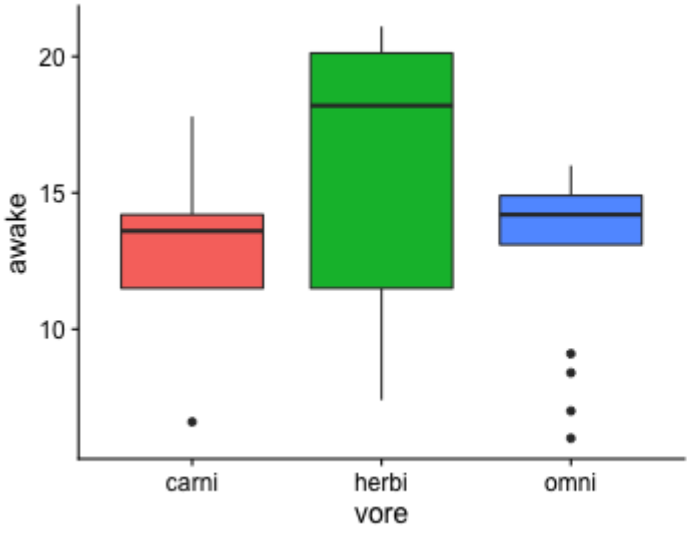
Identifying components of a plot

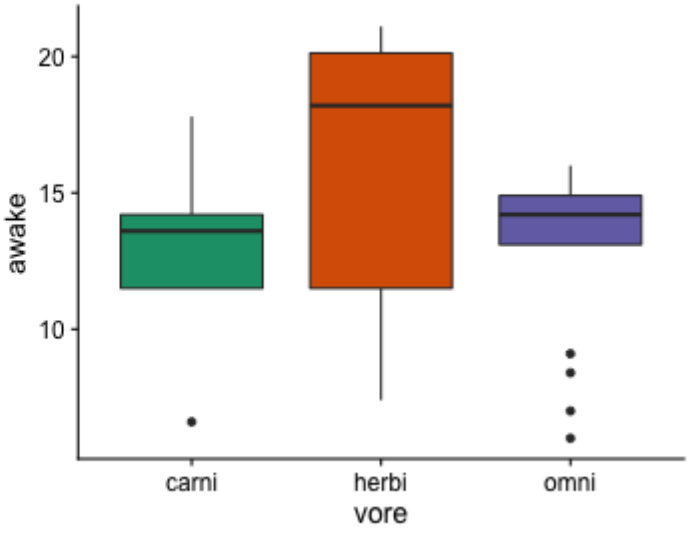


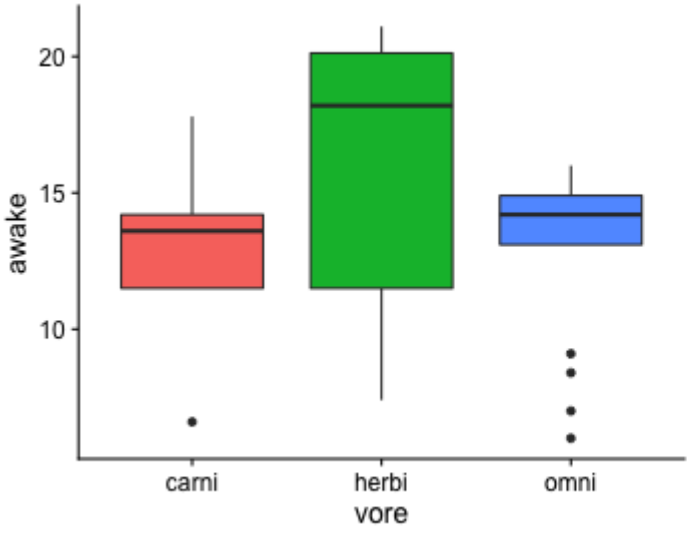
Aesthetics: How is the data *mapped onto* visual components of the plot?

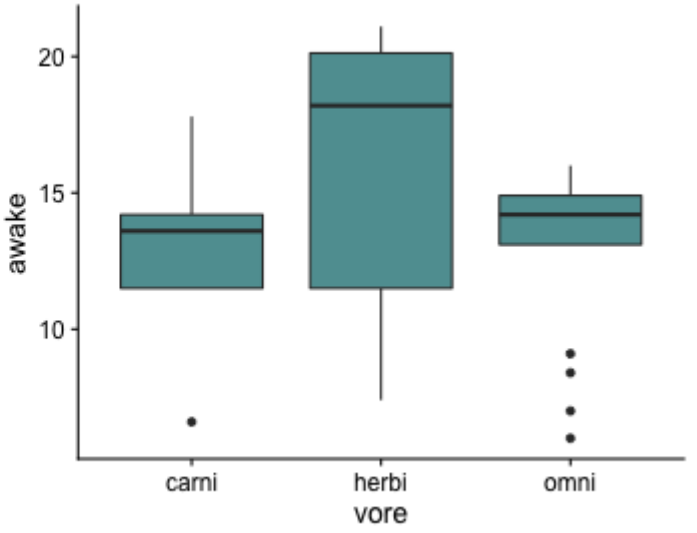
- X-axis?
- Y-axis?
- Colors? Shapes? Sizes?

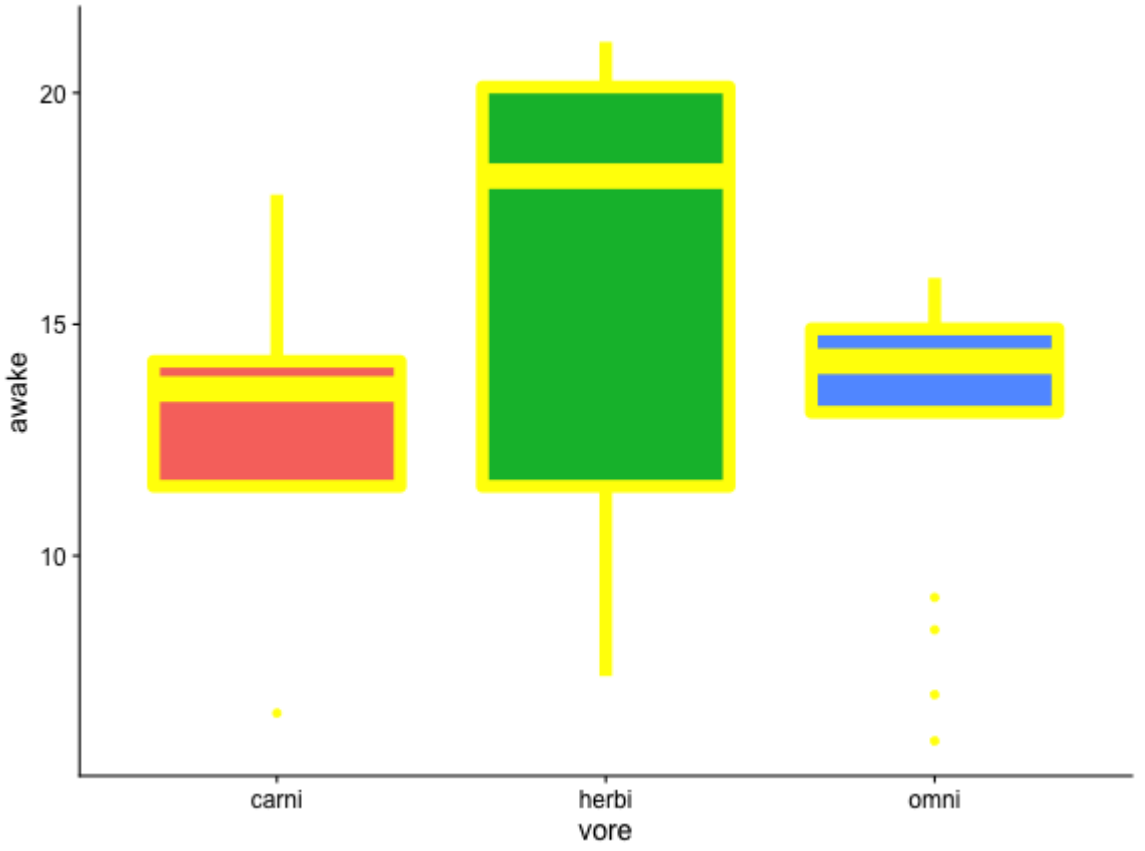
Geometries: What *shapes* aka *geometric objects* are displayed in the plot?

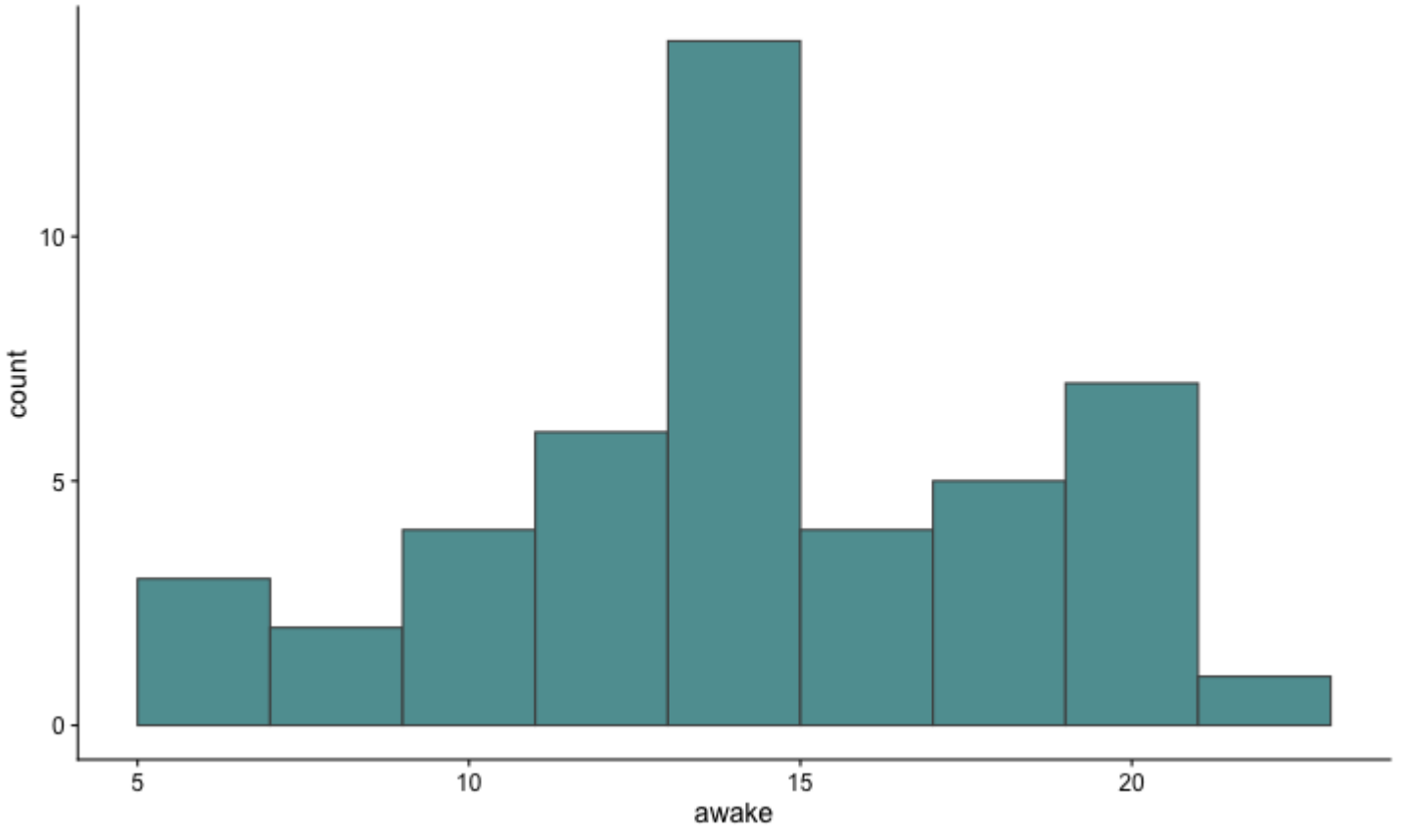

awake
20
15
10
carni
herbi
omni
vore

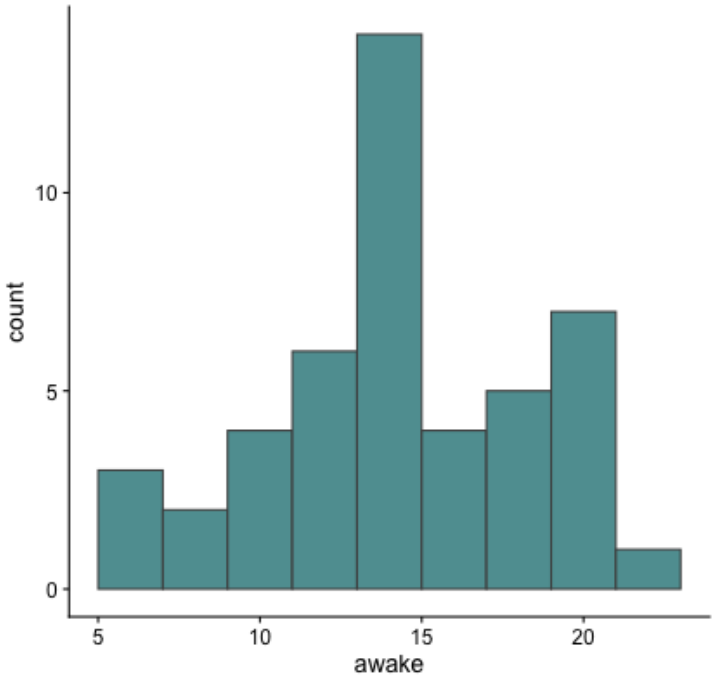

awake
20
15
10
carni
herbi
omni
vore

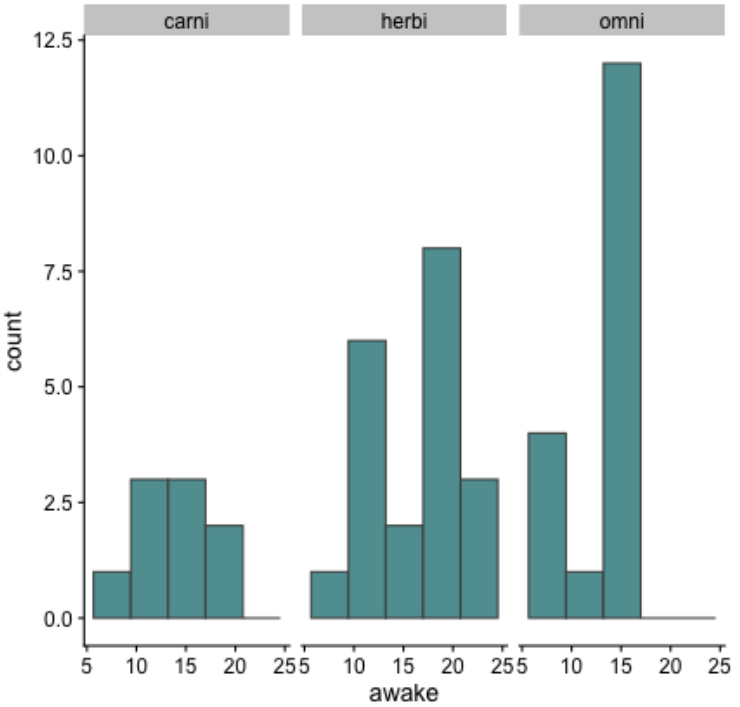

awake
20
15
10
carni
herbi
omni
vore

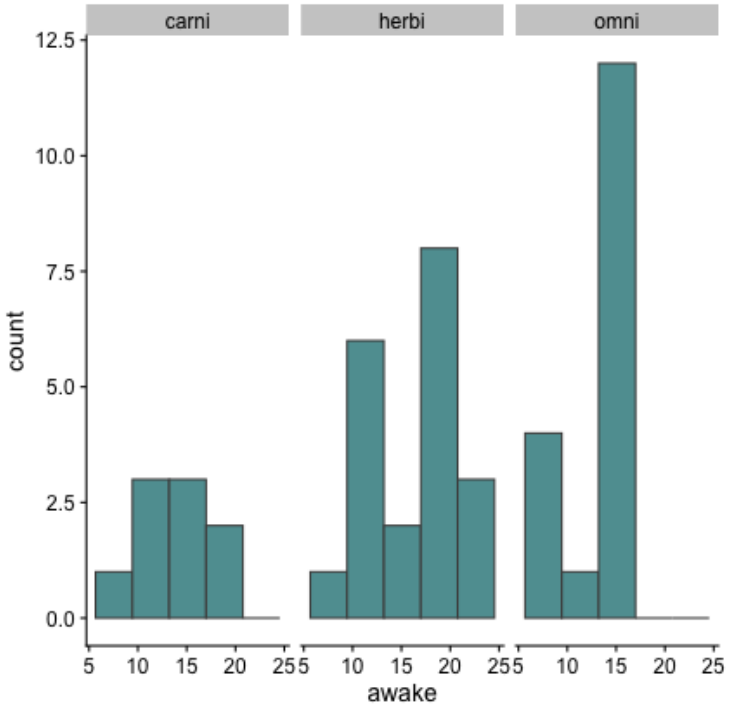

awake
20
15
10
carni
herbi
omni
vore

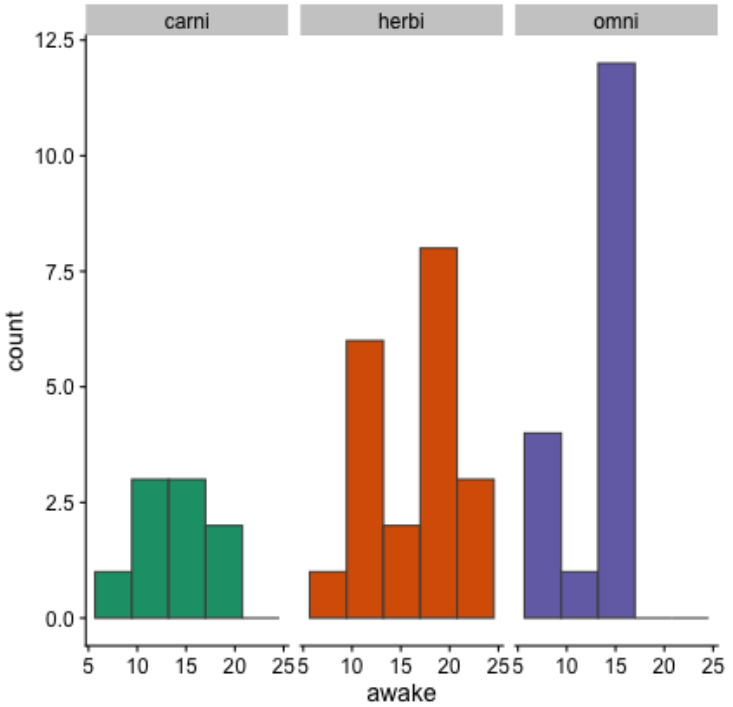

20
15
10
awake
carni
herbi
omni
vore

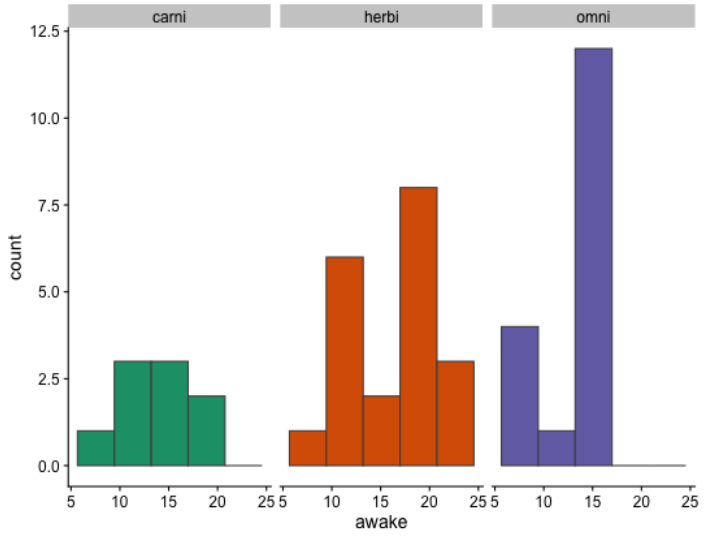

count
10
5
0
5
10
15
20
awake

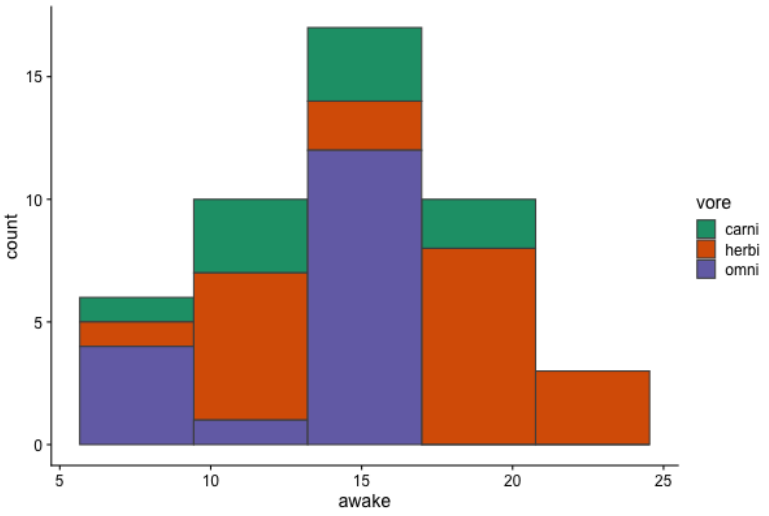

count
10
5
0
5
10
15
20
awake

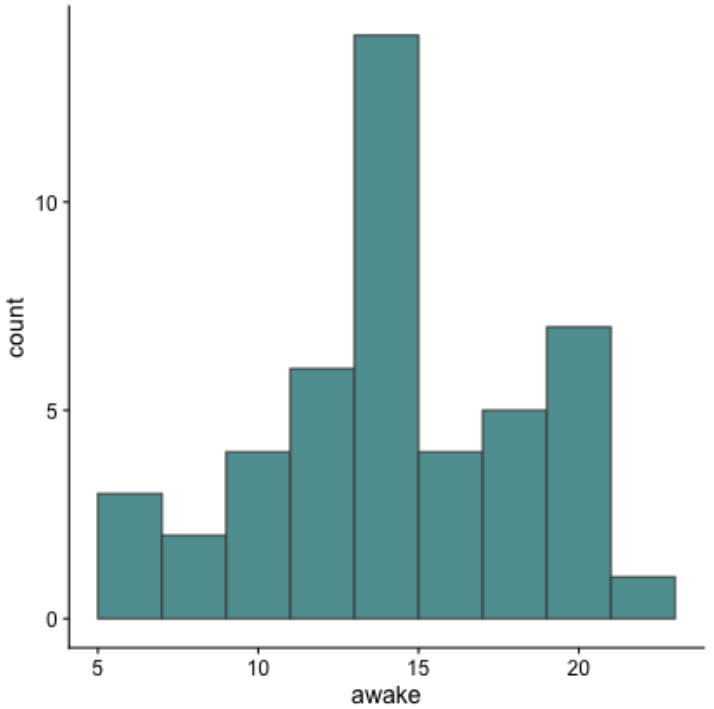

carni
herbi
omni
count
12.5
10.0
7.5
5.0
2.5
0.0
5
10
15
20
25
awake

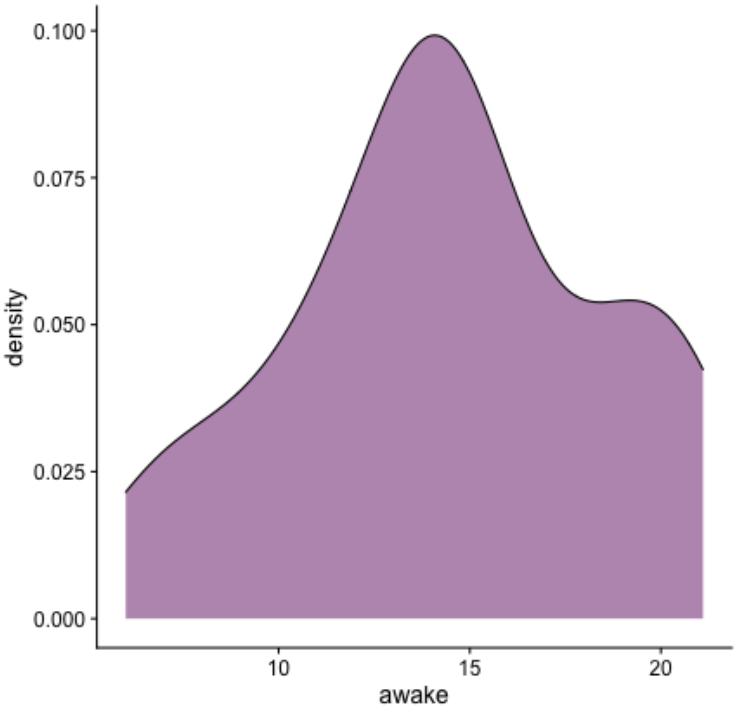

carni
herbi
omni
12.5
10.0
7.5
5.0
2.5
0.0
count
5
10
15
20
25
awake

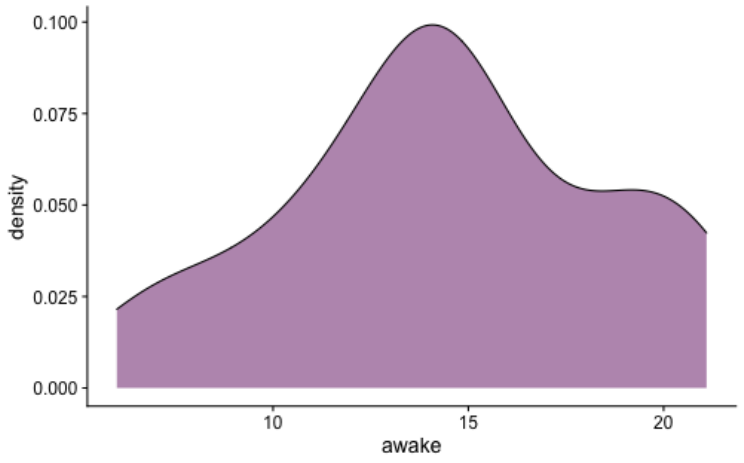

carni
herbi
omni
12.5
10.0
7.5
5.0
2.5
0.0
count
5
10
15
20
25
awake

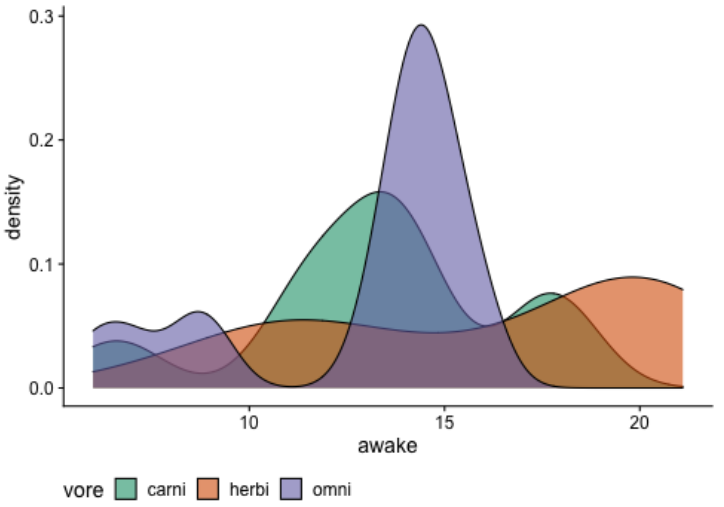

carni
herbi
omni
12.5
10.0
7.5
5.0
2.5
0.0
count
5
10
15
20
25
awake

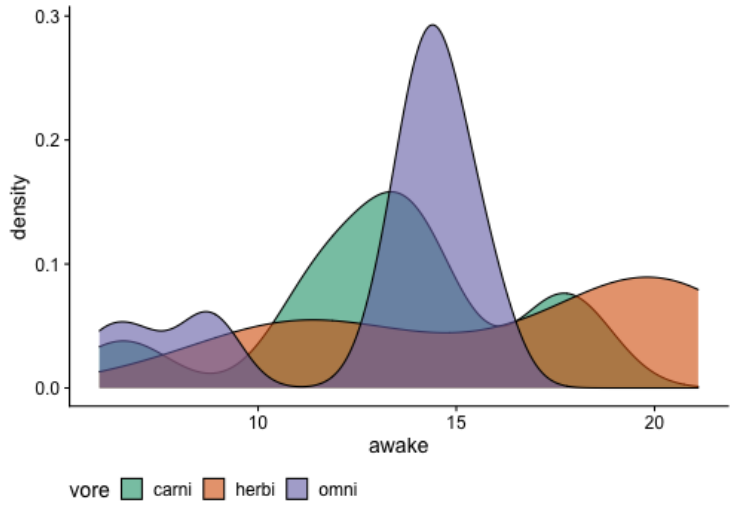

15
10
5
0
count
5
10
15
20
25
awake
vore
carni
herbi
omni

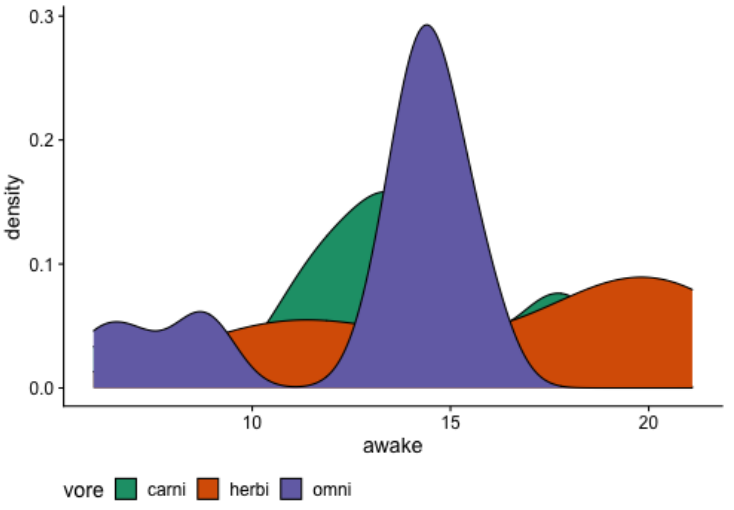

count
10
5
0
5
10
15
20
awake

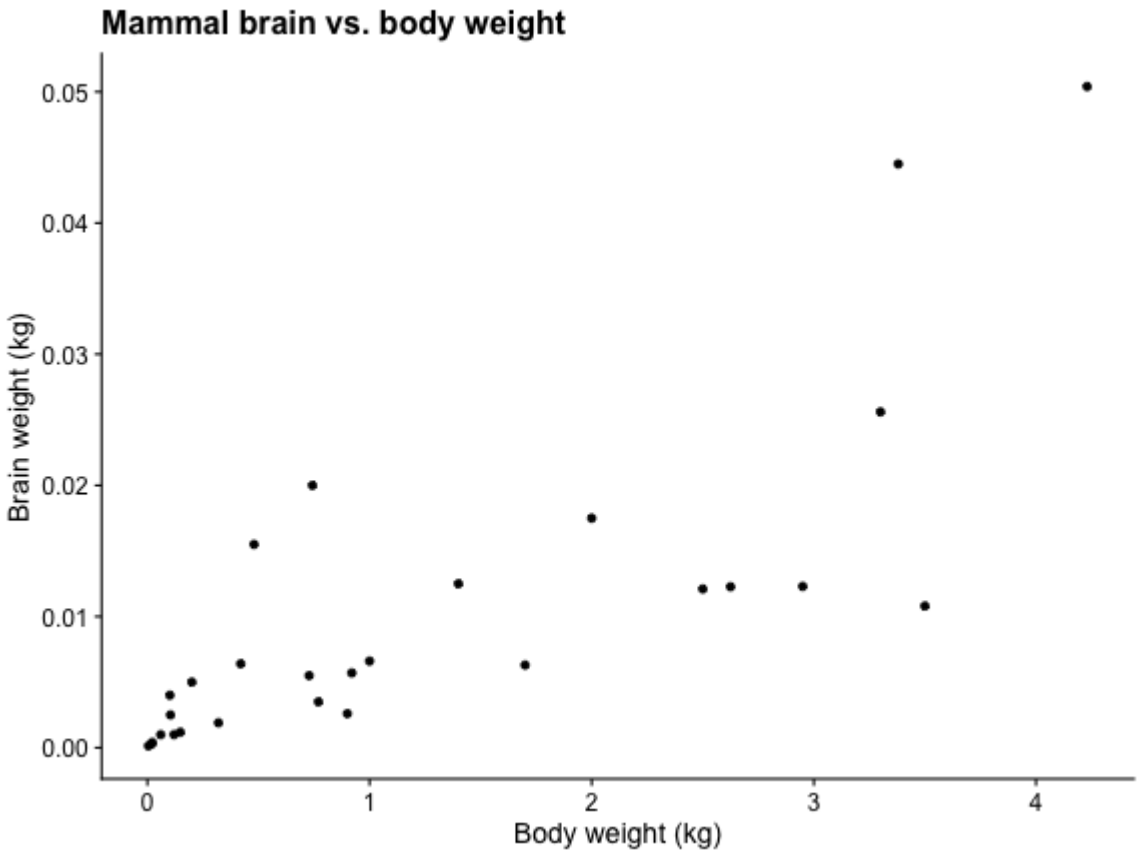

density
0.100
0.075
0.050
0.025
0.000
10
15
20
awake

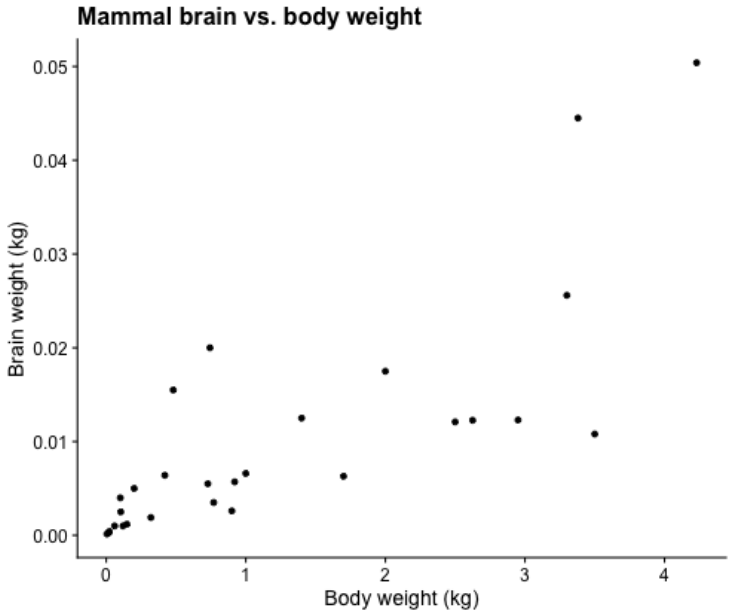

0.100
0.075
0.050
0.025
0.000
density
10
15
20
awake

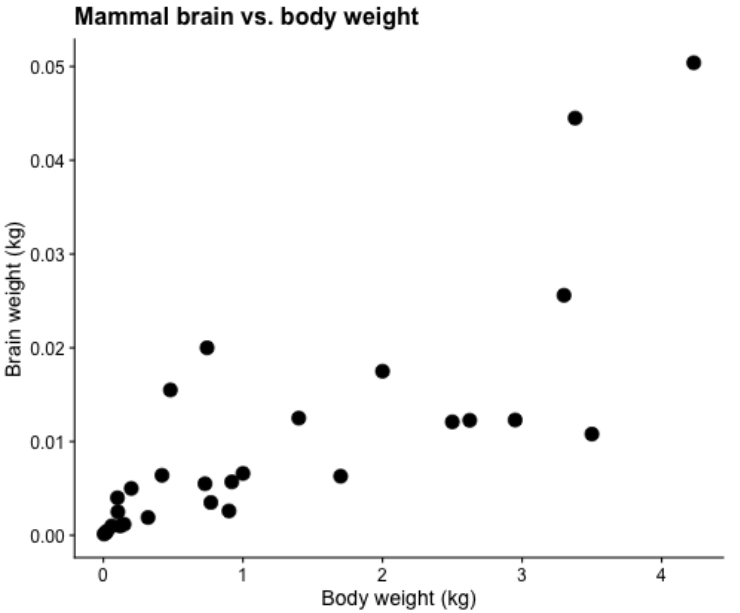

0.3
0.2
0.1
0.0
density
10
15
20
awake
vore
carni
herbi
omni

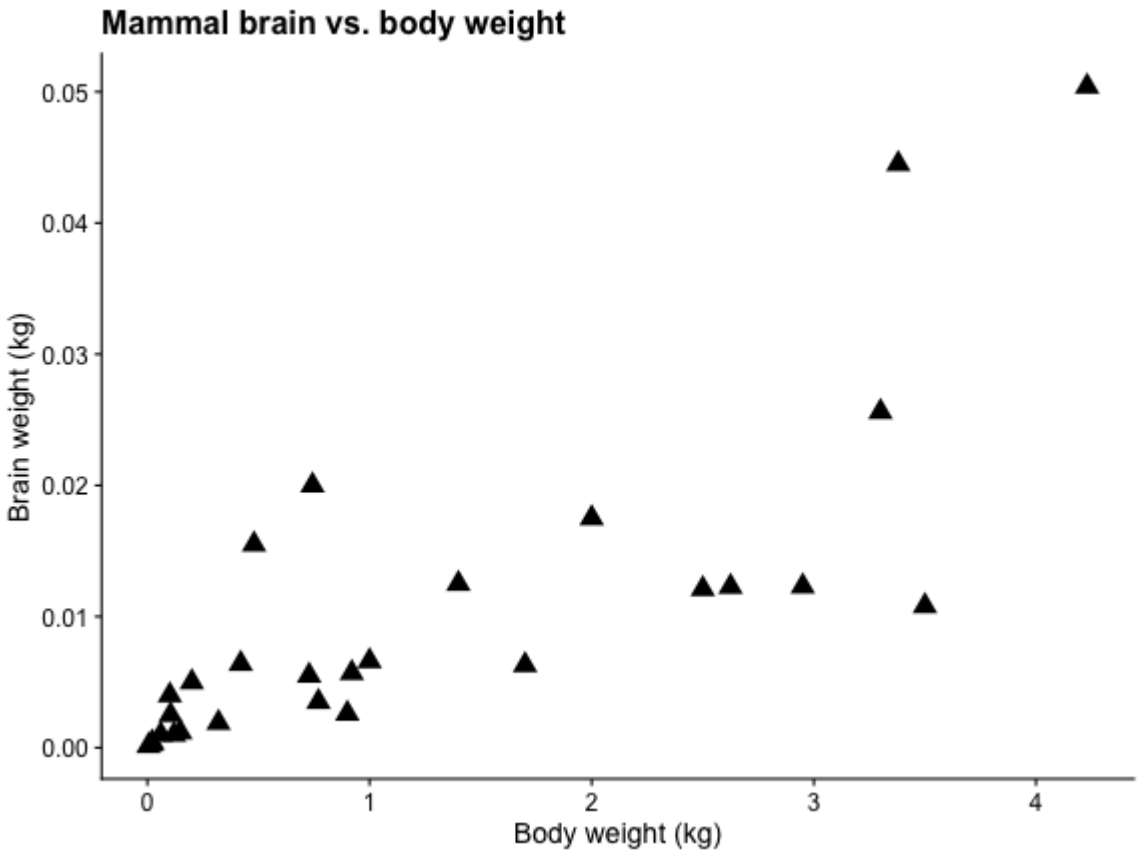

0.3
0.2
0.1
0.0
density
10
15
20
awake
vore
carni
herbi
omni

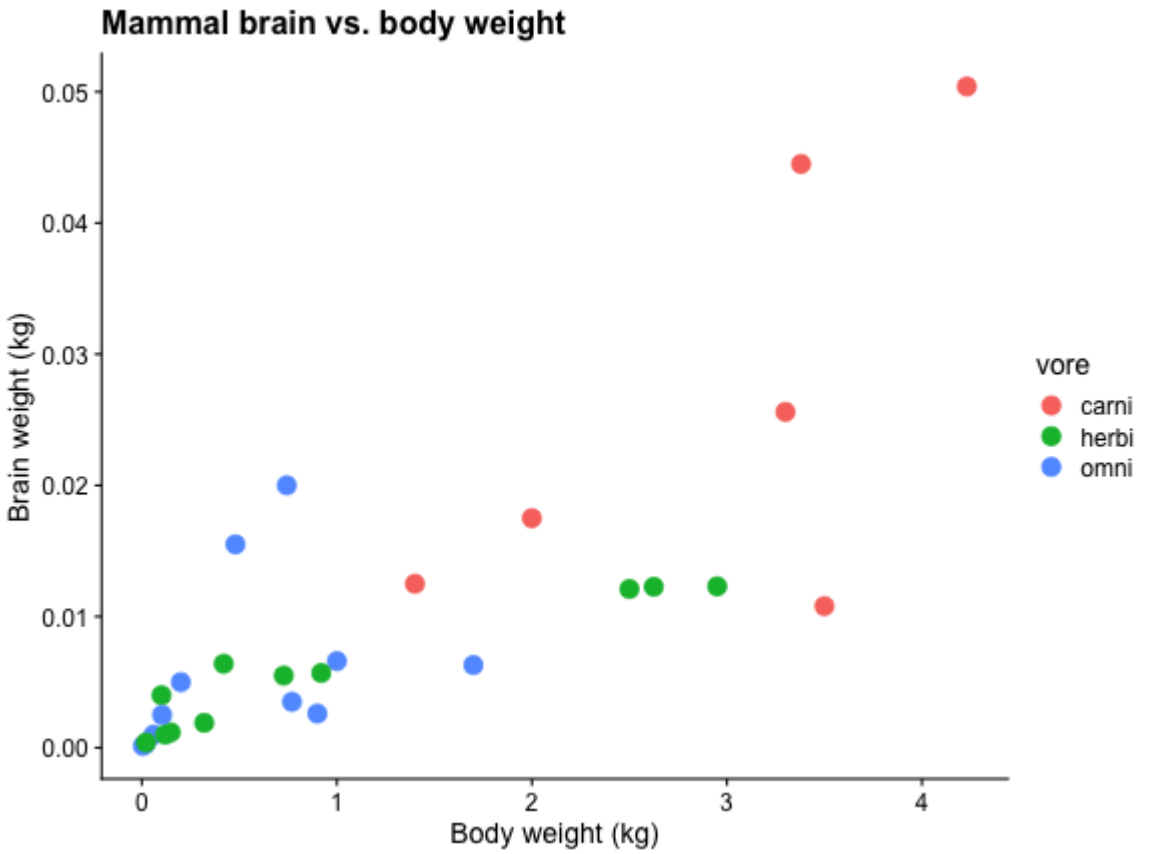

0.3
0.2
0.1
0.0
density
10
15
20
awake
vore
carni
herbi
omni

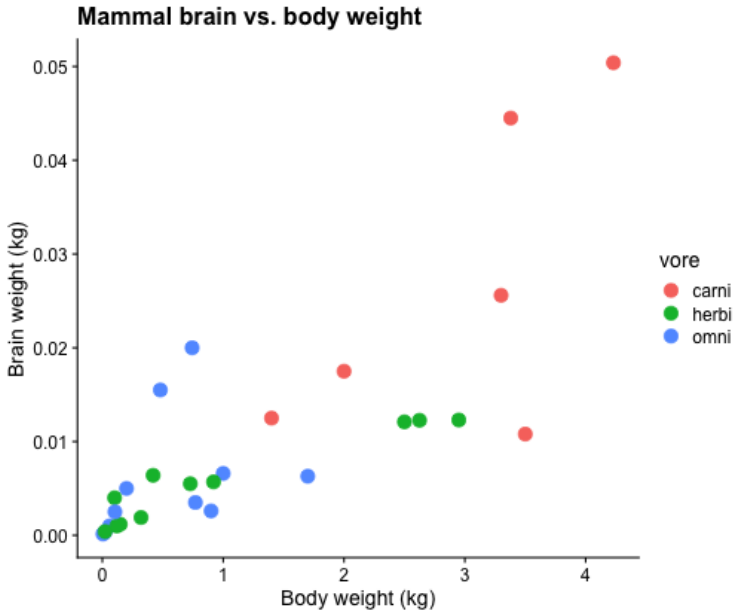

Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)

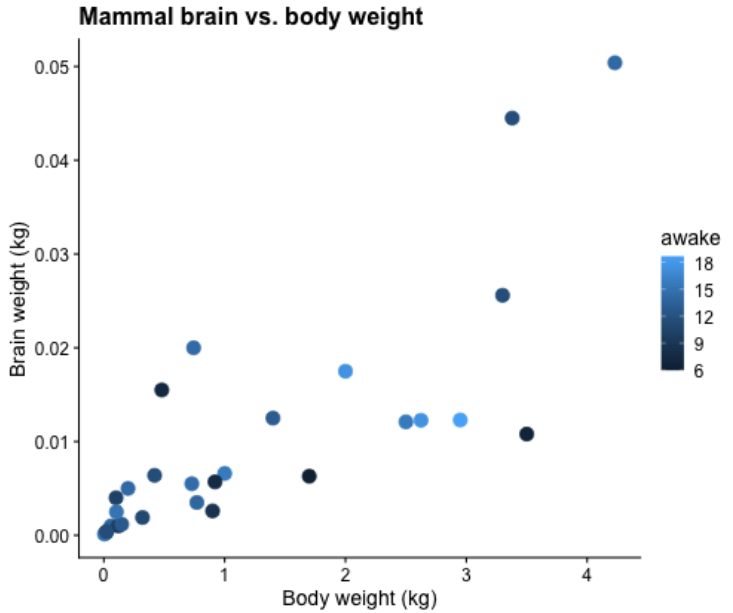

Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)

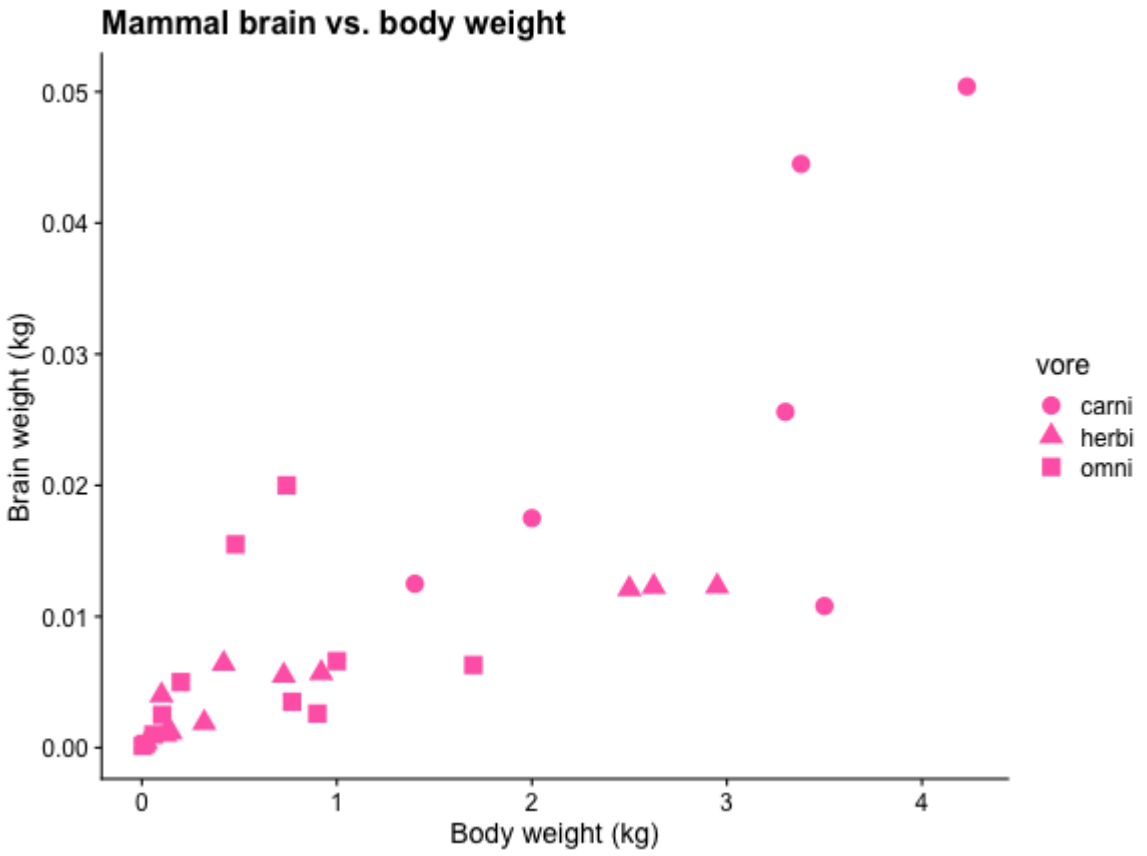

Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)

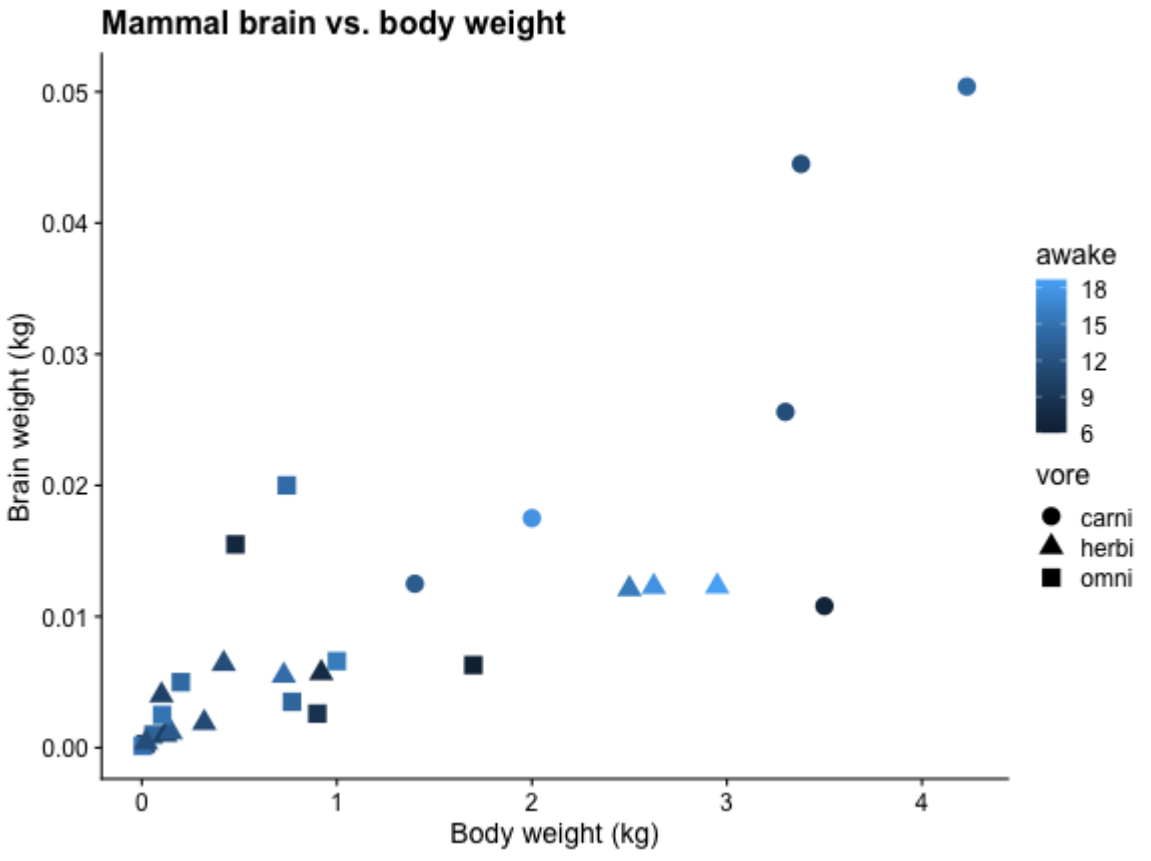

Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)


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

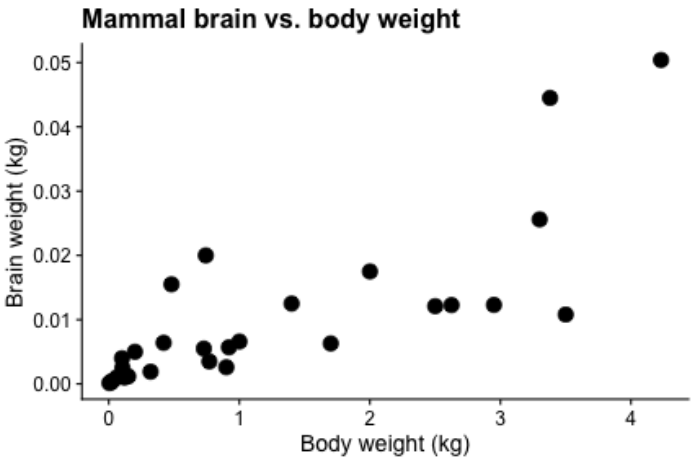

Mammal brain vs. body weight
Brain weight (kg)
Body weight (kg)
0.00
0.01
0.02
0.03
0.04
0.05
0
1
2
3
4
vore
carni
herbi
omni

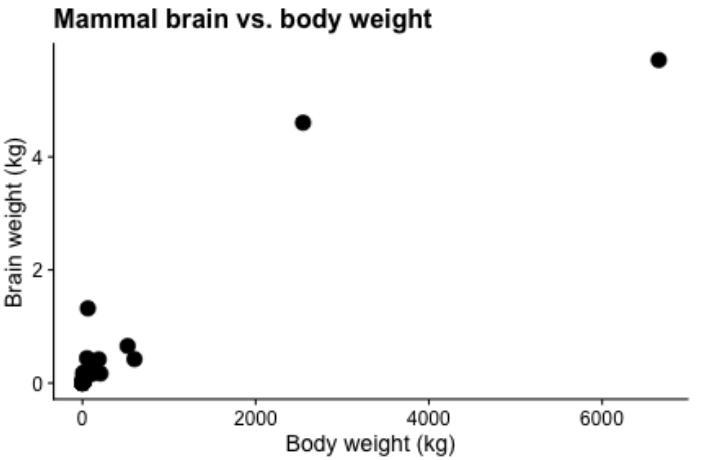

Mammal brain vs. body weight
Brain weight (kg)
Body weight (kg)
0.00
0.01
0.02
0.03
0.04
0.05
0
1
2
3
4
awake
18
15
12
9
6


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

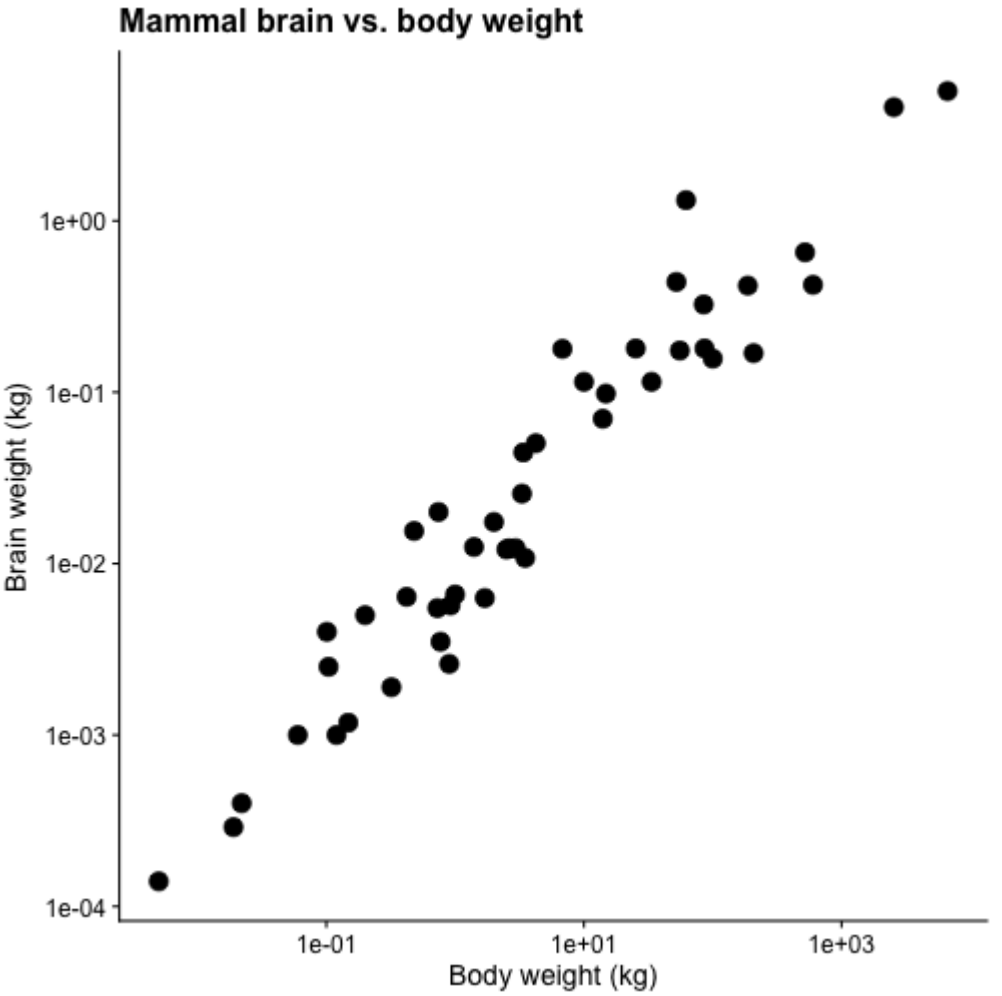

Mammal brain vs. body weight
Brain weight (kg)
0.05
0.04
0.03
0.02
0.01
0.00
0
1
2
3
4
Body weight (kg)
awake
18
15
12
9
6
vore
carni
herbi
omni

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





Use log scales for data with extreme ranges

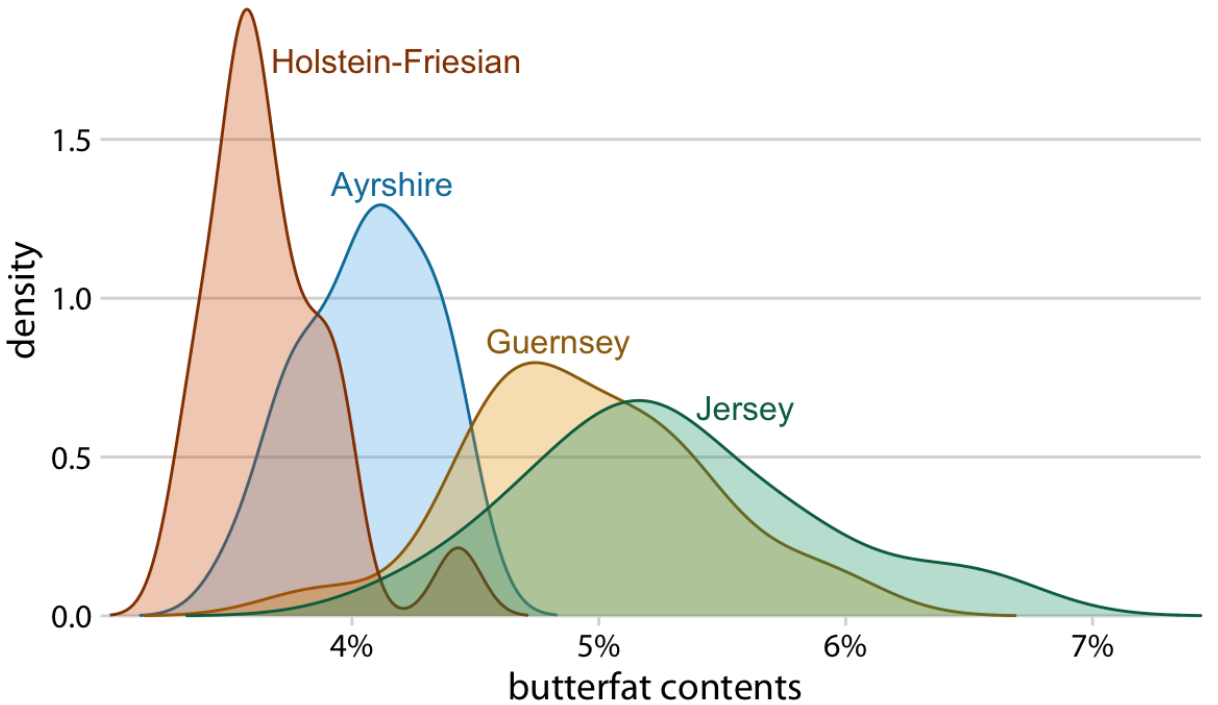


Let's practice

- What variable is on the X-axis? What *type* of data is it?
- 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 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?*

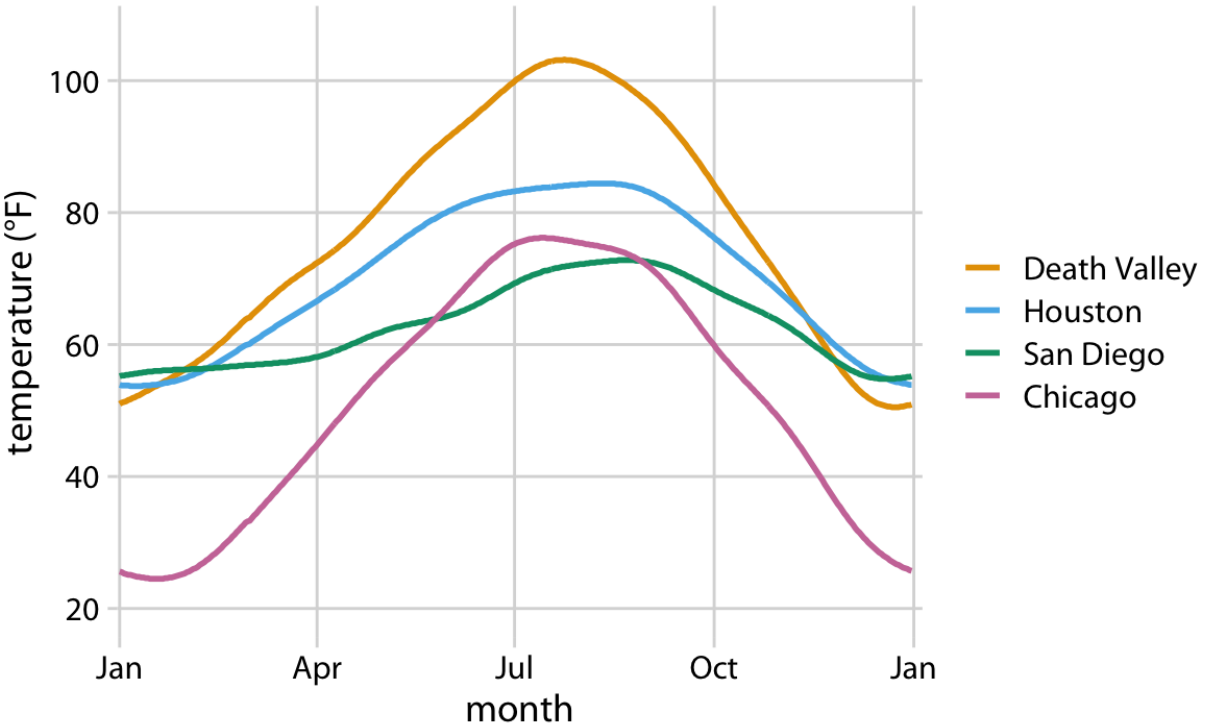
All figures in the following slides are from Fundamentals of Data Visualization.

Butterfat from different cows



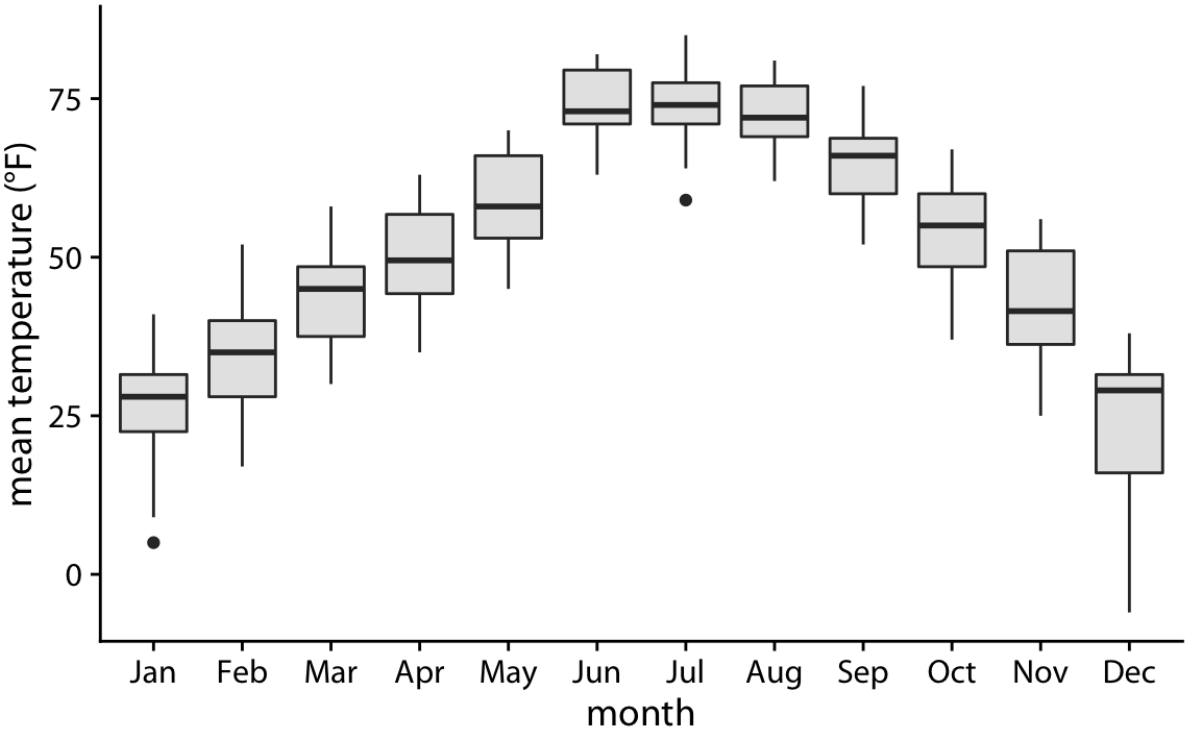
Density estimates of the butterfat percentage in the milk of four cattle breeds. Data Source: Canadian Record of Performance for Purebred Dairy Cattle.

Average daily temperatures

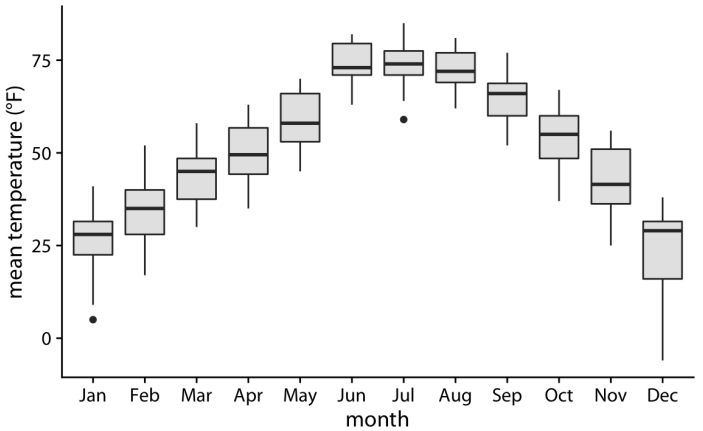


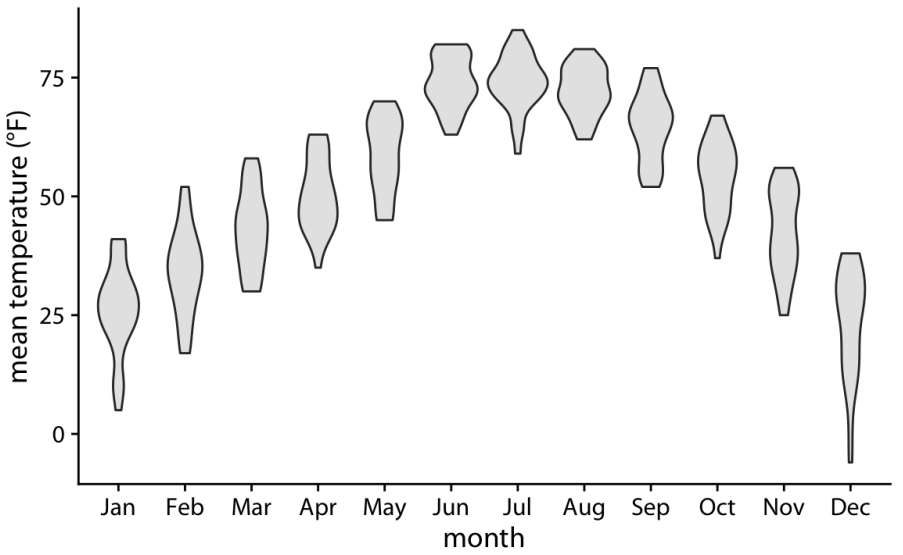
Daily temperature normals for four selected locations in the U.S. Temperature is mapped to the y axis, day of the year to the x axis, and location to line color. Data source: NOAA.

Temperatures in Lincoln, NE in 2016

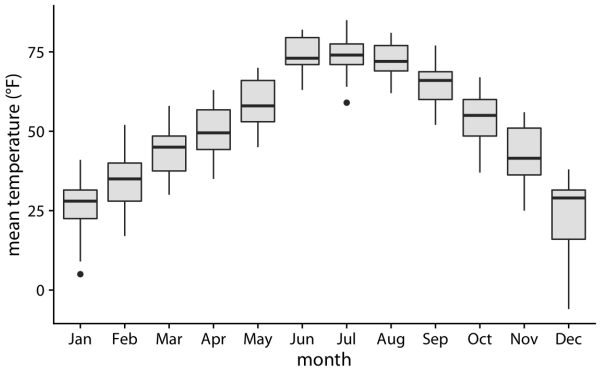


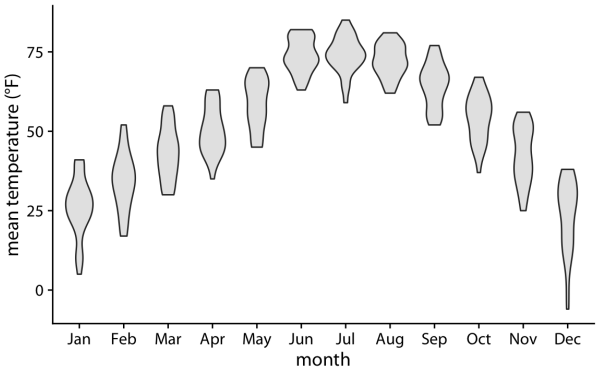
Temperatures in Lincoln, NE in 2016

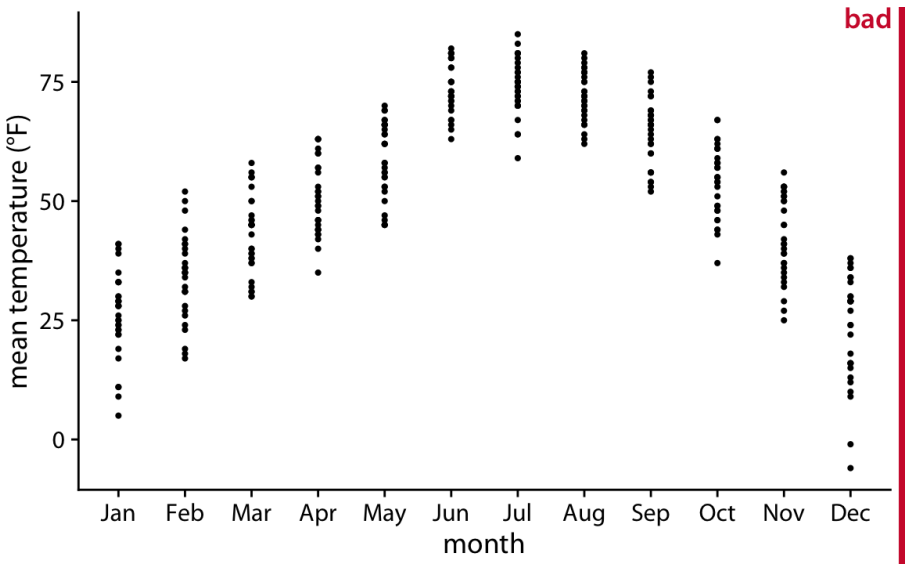




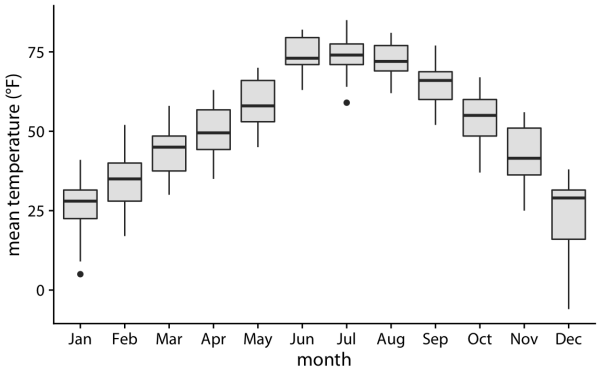
Temperatures in Lincoln, NE in 2016

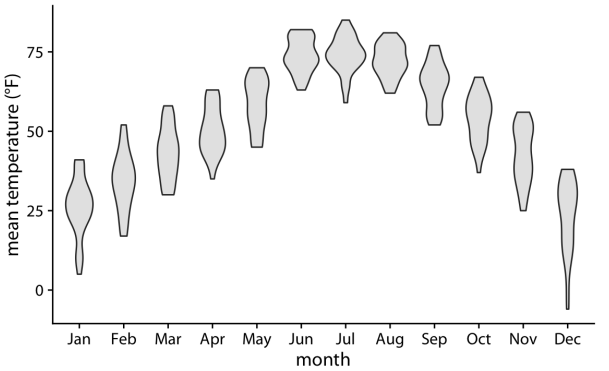


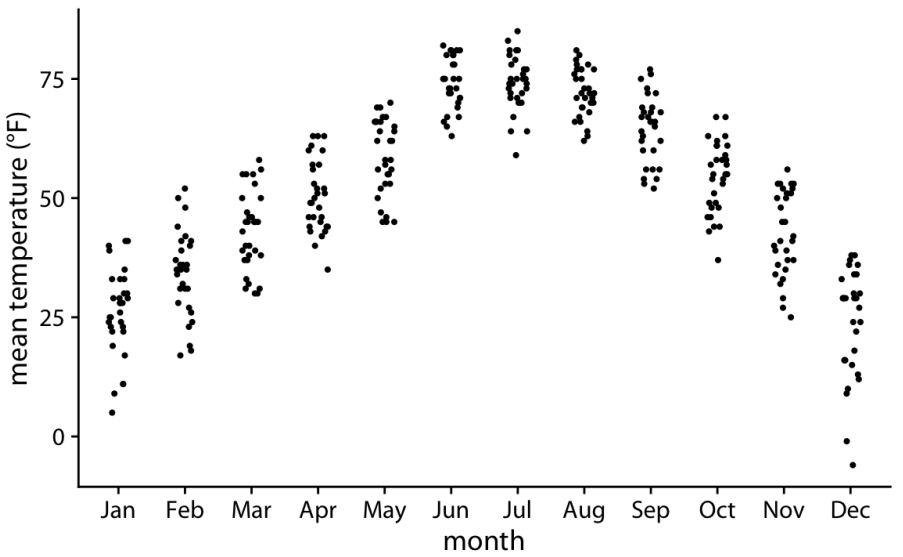




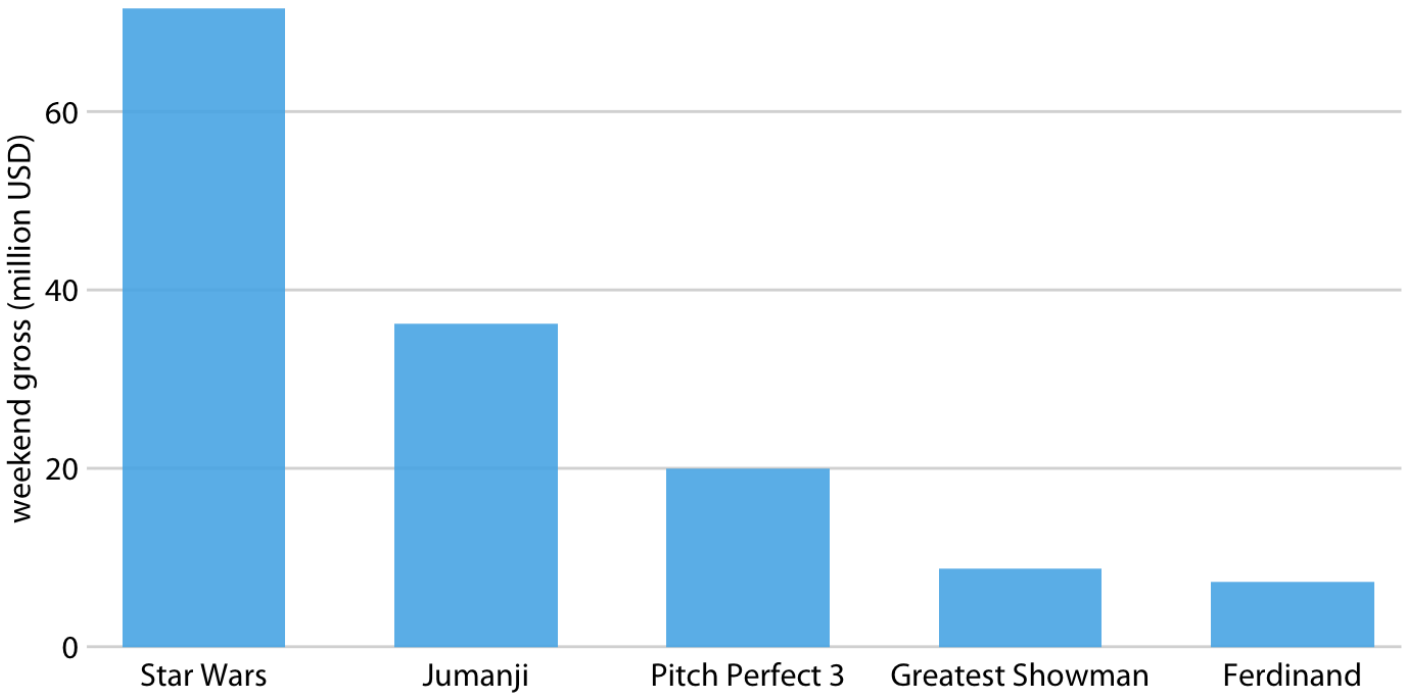
Temperatures in Lincoln, NE in 2016





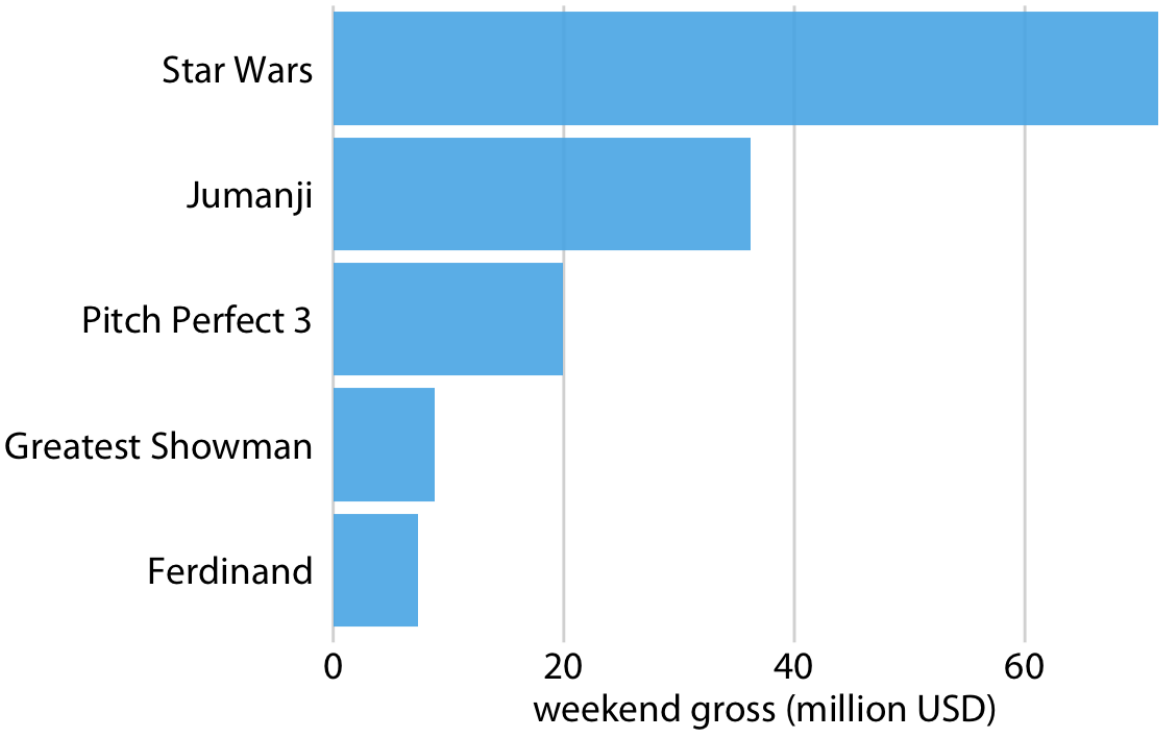


Box office income

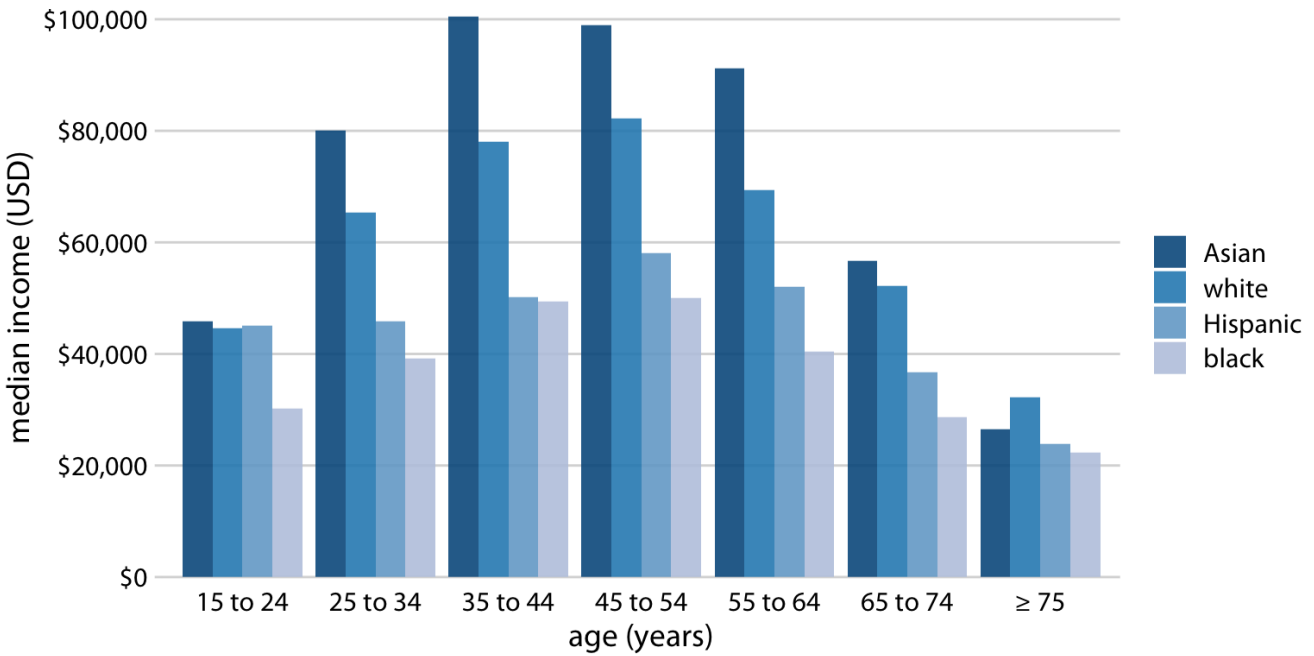


Highest grossing movies for the weekend of December 22-24, 2017. Data source: Box Office Mojo.

Box office income - what's different?

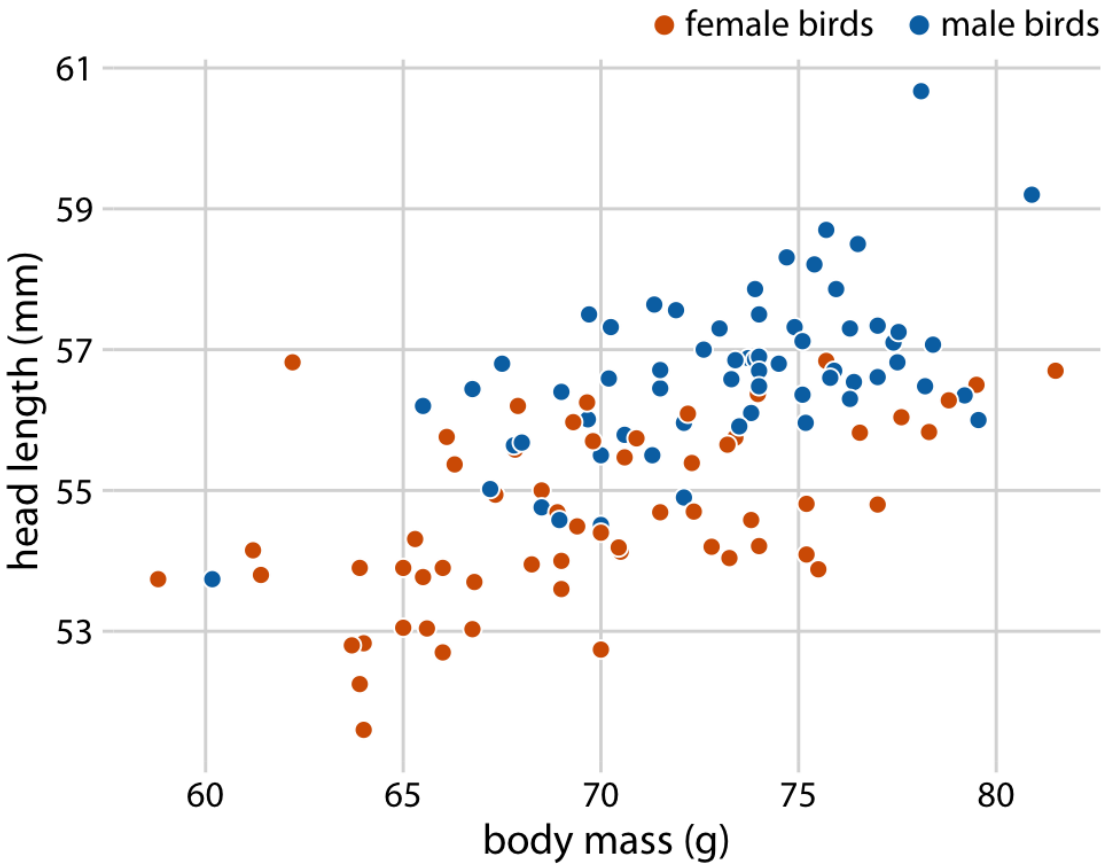


Median household income



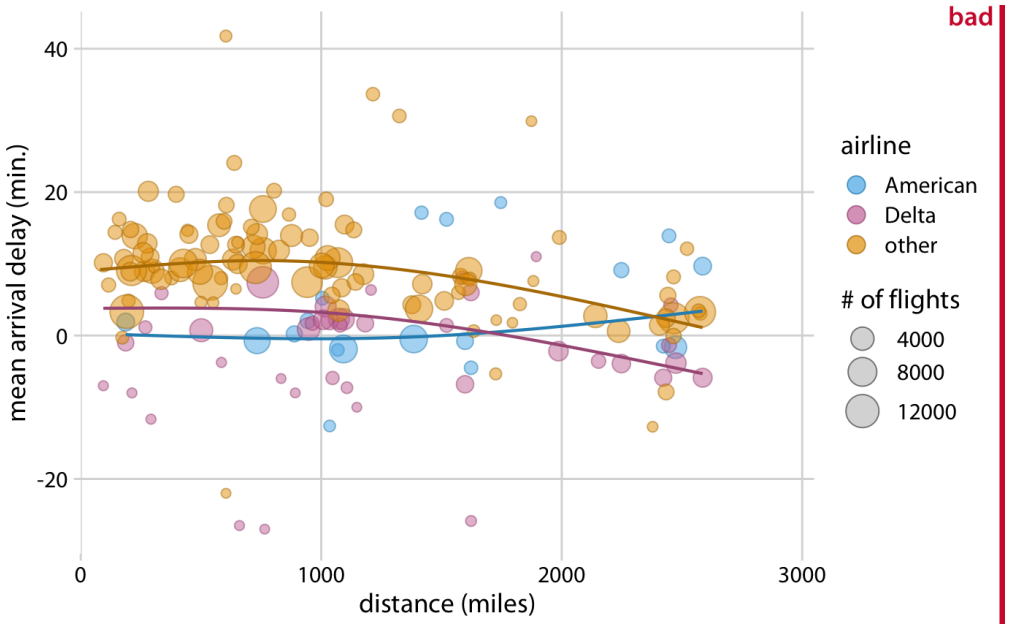
2016 median U.S. annual household income versus age group and race. For each age group there are four bars, corresponding to the median income of Asian, white, Hispanic, and black people, respectively. Data source: United States Census Bureau.

Bluejays



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

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.

"Looking cool/smart" is NOT the same as effectively communicating."