



Best practices for data visualization

Data Science for Biologists

Good doggos from Caitlyn (L) and Kevin (R)

Riley



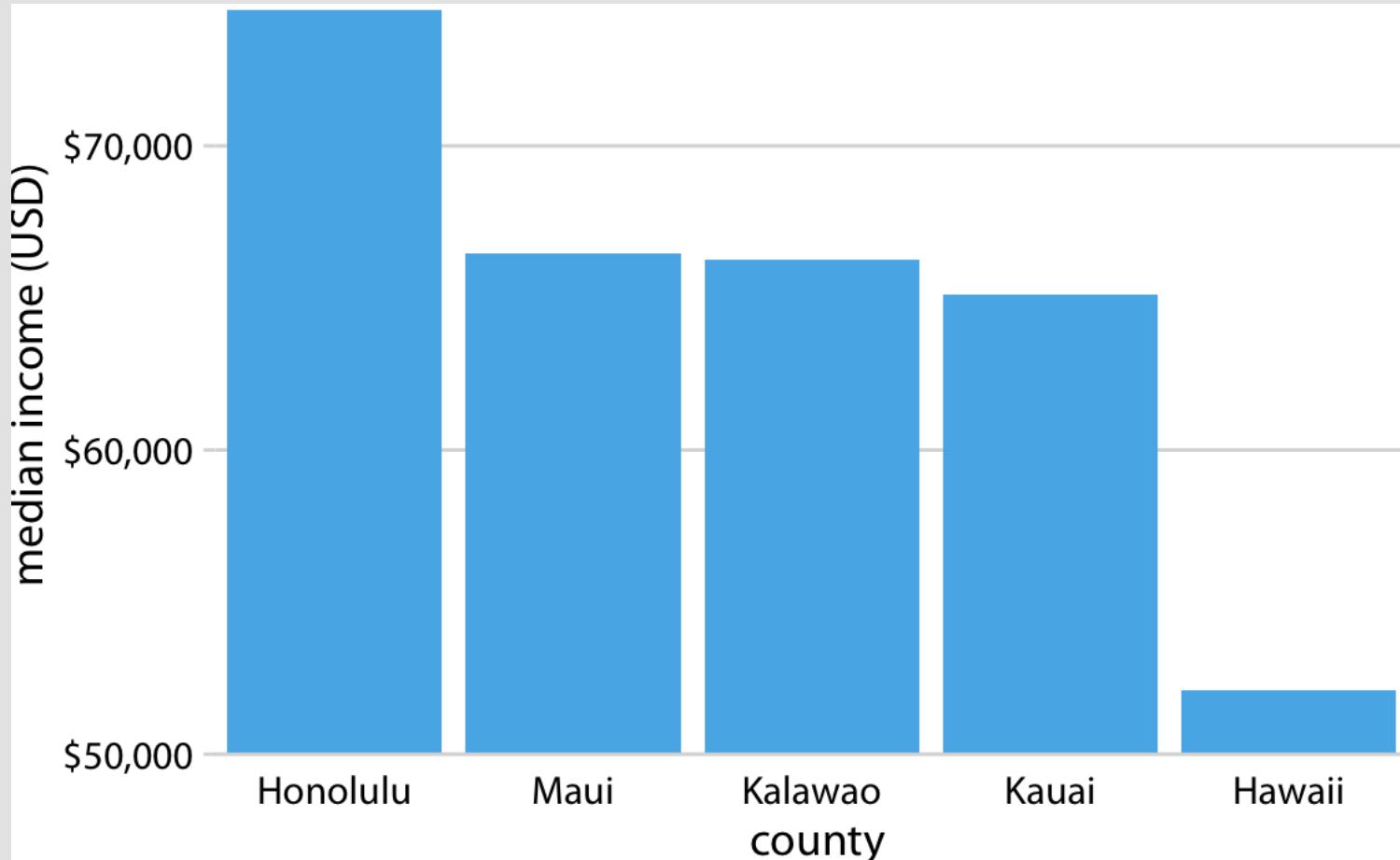
Guiding principle

The goal of data visualization is to communicate.

We do not want to create plots that are...

- factually misleading
- factually incorrect
- impossible or overly difficult to interpret
- so ugly that you don't even want to try interpreting them

All figures (except the `msleep` figures) in the following slides are from [Fundamentals of Data Visualization](#).



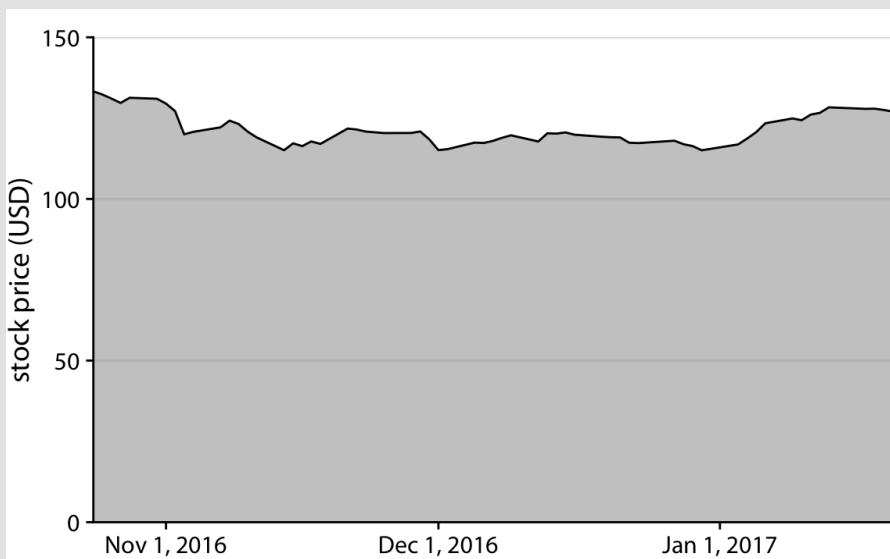
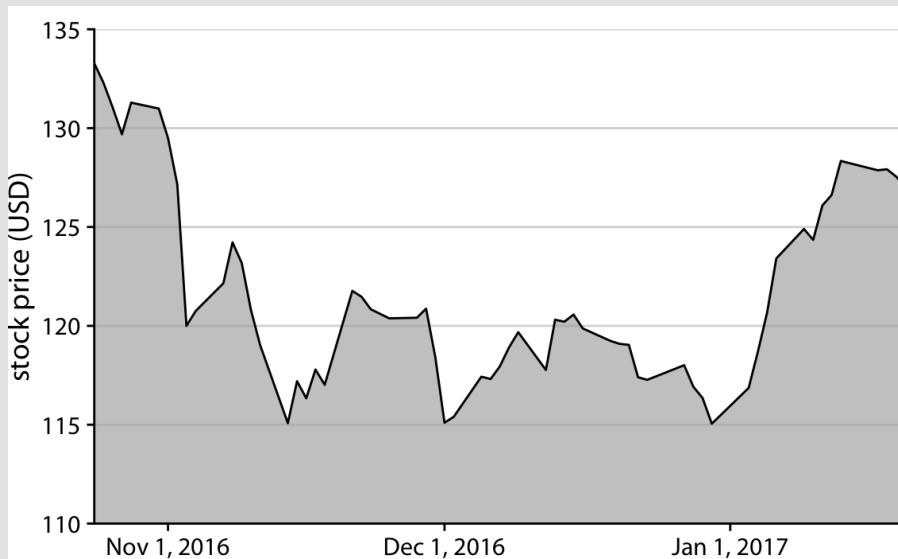
Median income in the five counties of the state of Hawaii. Data source: 2015 Five-Year American Community Survey.

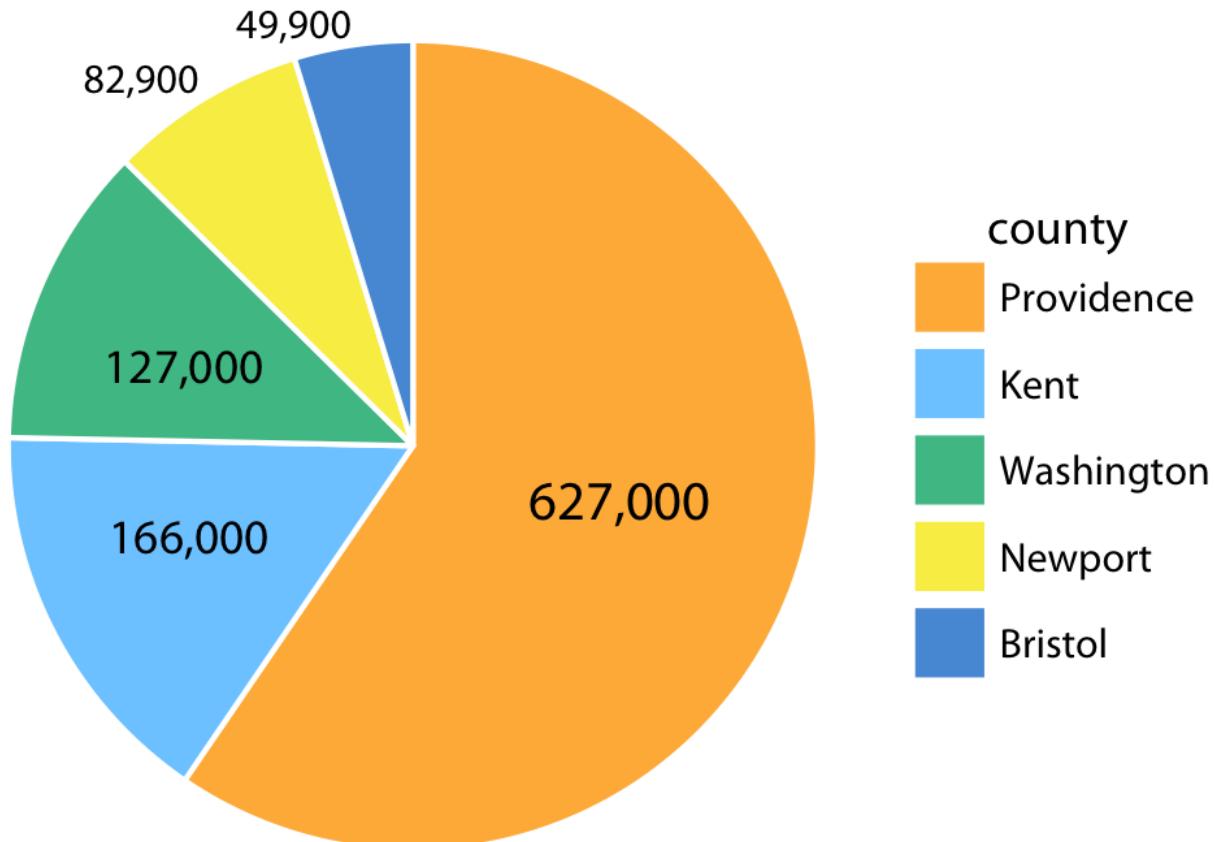
Best Principle: The principle of proportional ink

The sizes of shaded areas in a visualization need to be proportional to the data values they represent.

Truncating the y-axis is highly misleading.

Stock price of
Facebook (FB)
from Oct. 22,
2016 to Jan. 21,
2017.

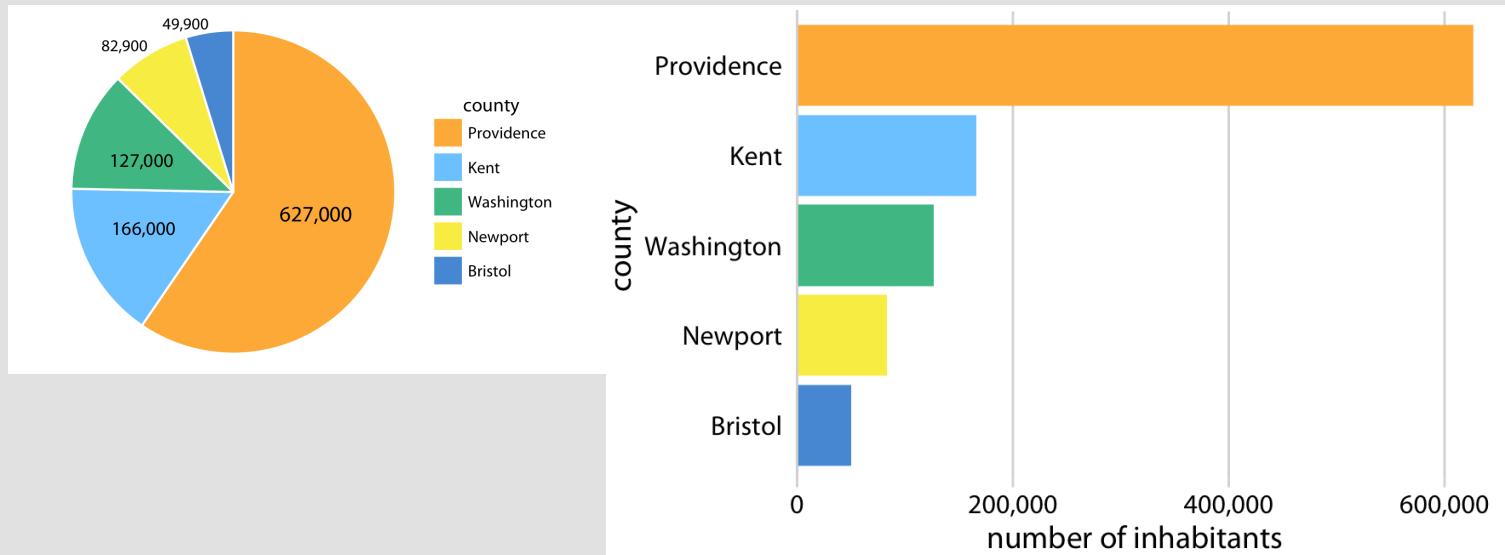


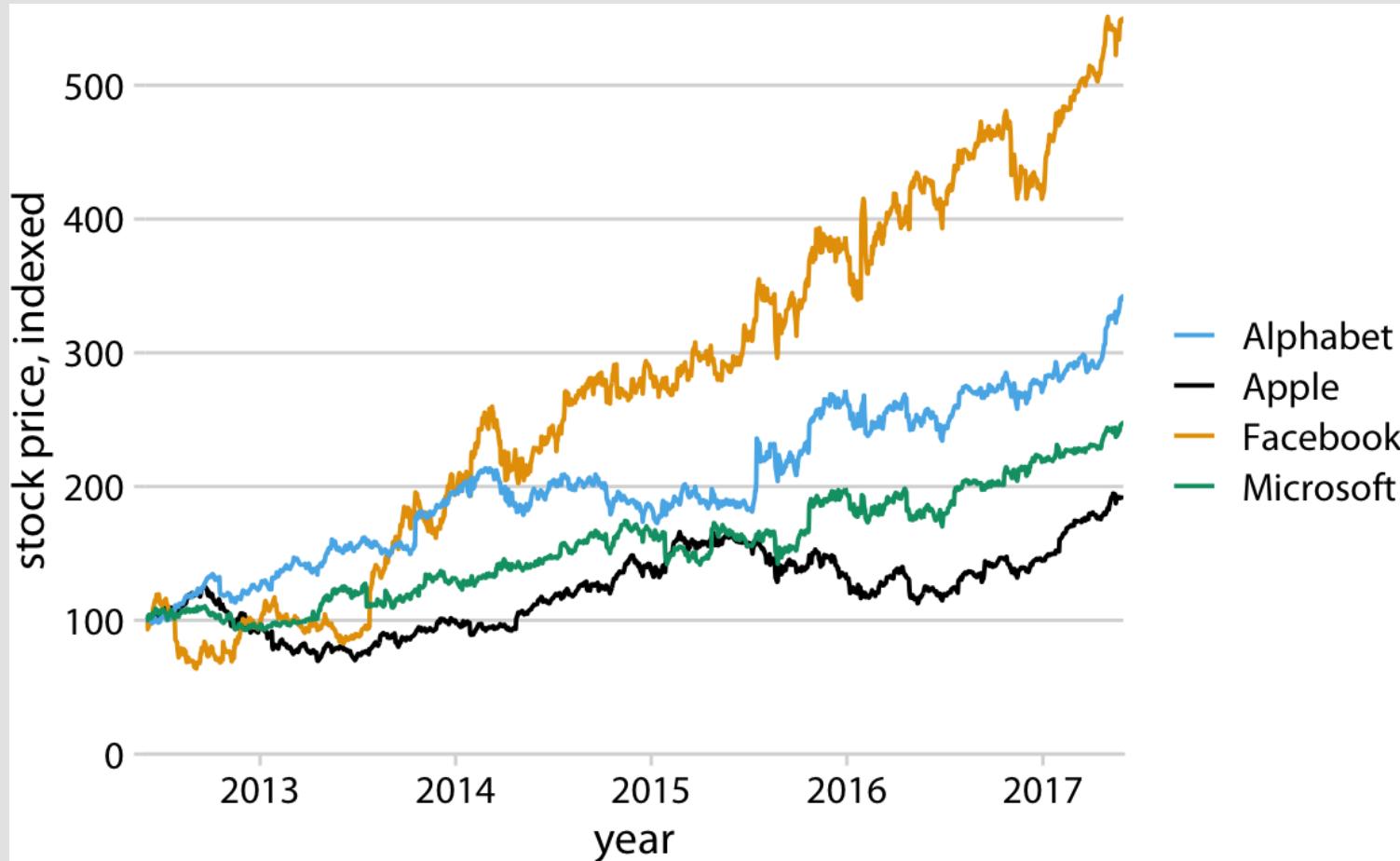


Number of inhabitants in Rhode Island counties. Data source: 2010 Decennial U.S. Census.

Best Principle: Don't use pie charts.

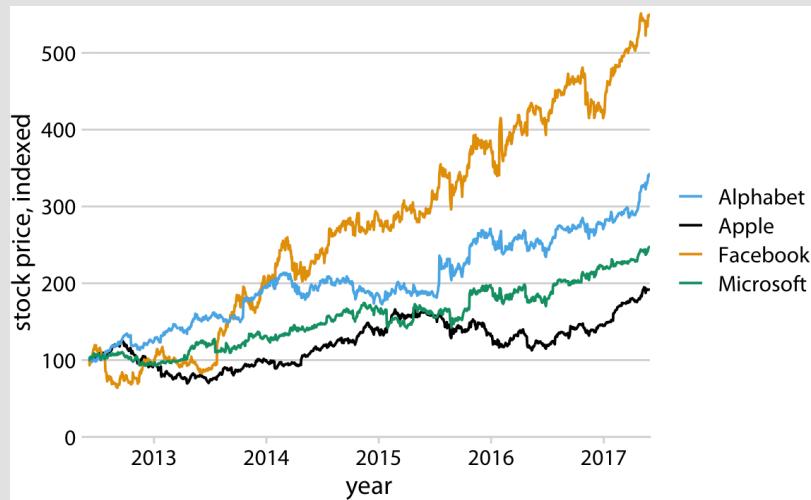
Pies are sized based on relative *area*. Human eyes are really bad at properly perceiving area.

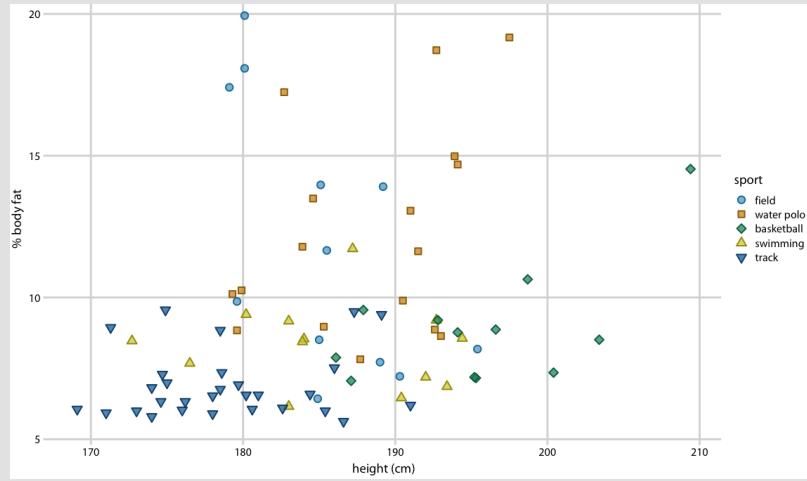




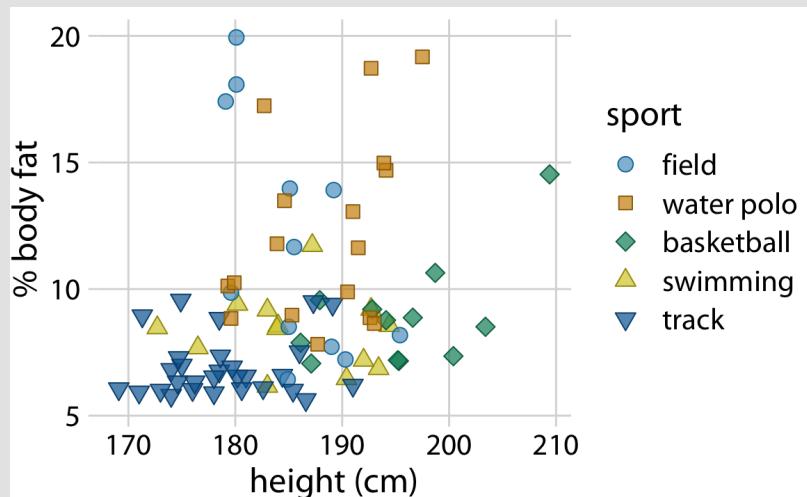
Stock price over time for four major tech companies. The stock price for each company has been normalized to equal 100 in June 2012. Data source: Yahoo Finance.

Best Principle: Make your viewer's life easy



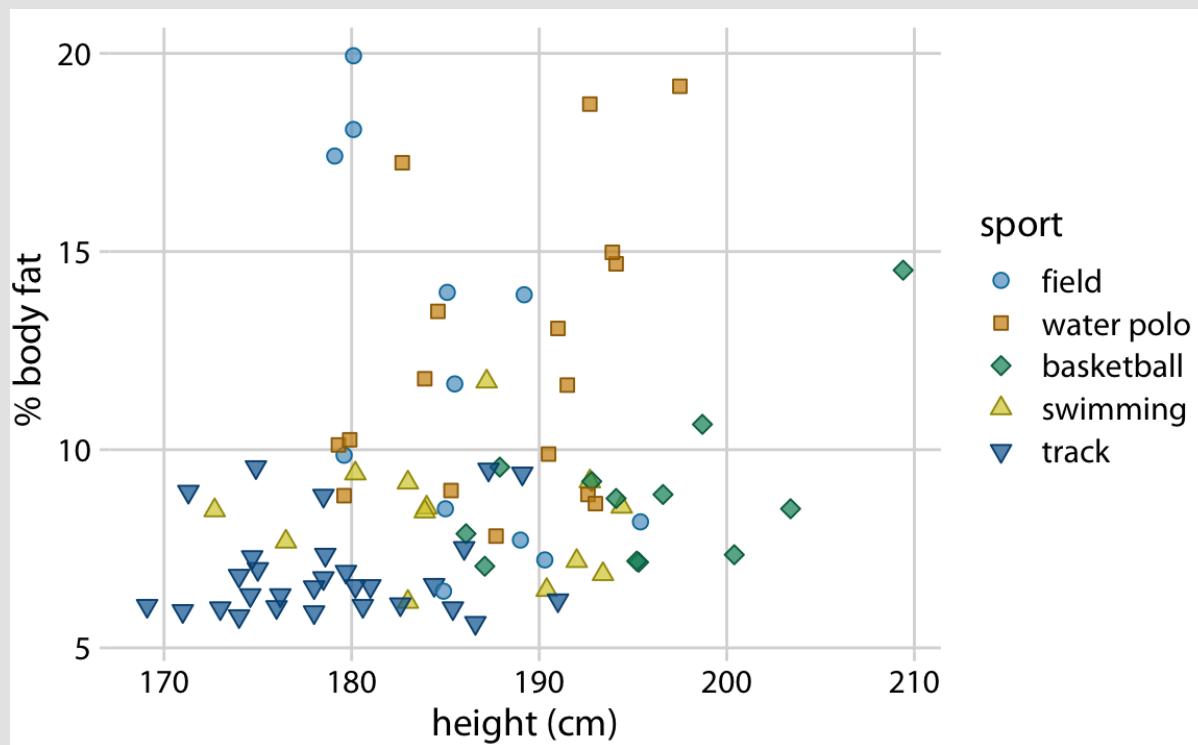


Percent body fat versus height in professional male Australian athletes. (Each point represents one athlete.) Data source: Telford and Cunningham (1991)



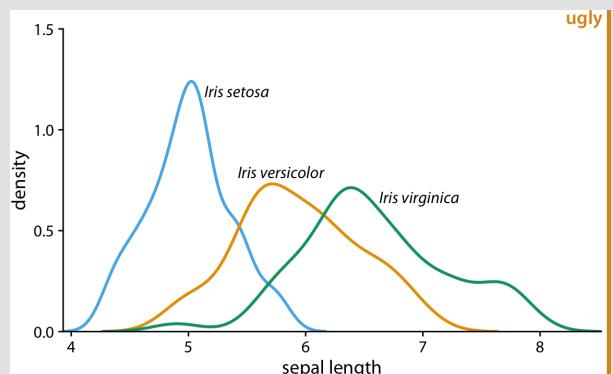
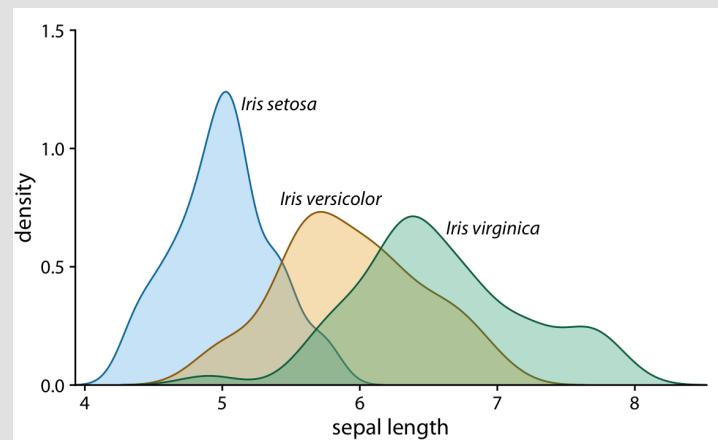
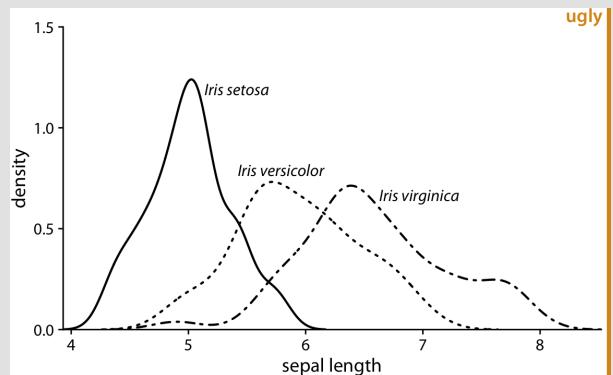
Use trial and error to get the right style

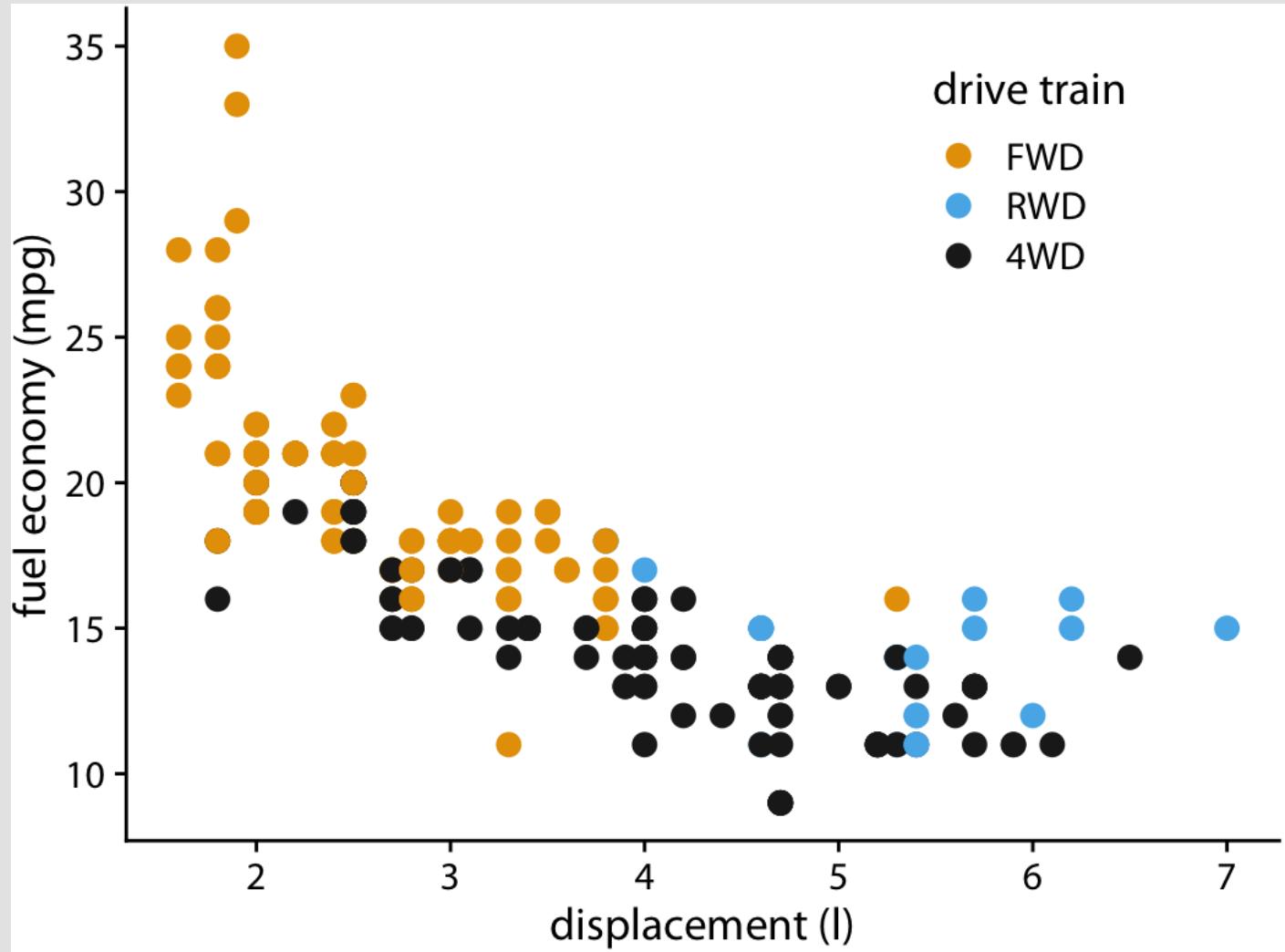
"Goldilocks" text size:



Don't annoy your viewers, either

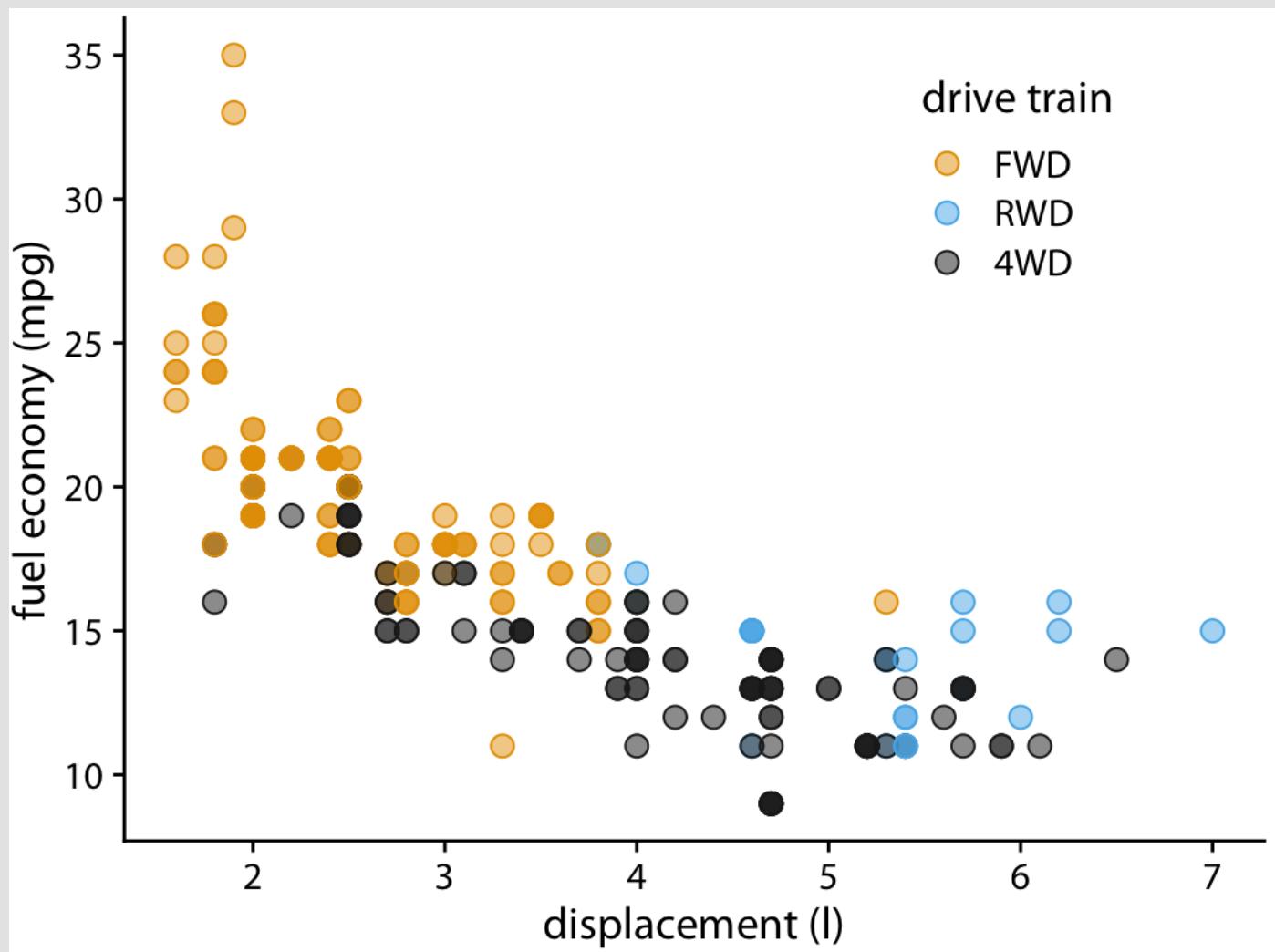
People are more likely to *want* to interpret aesthetically pleasing figures

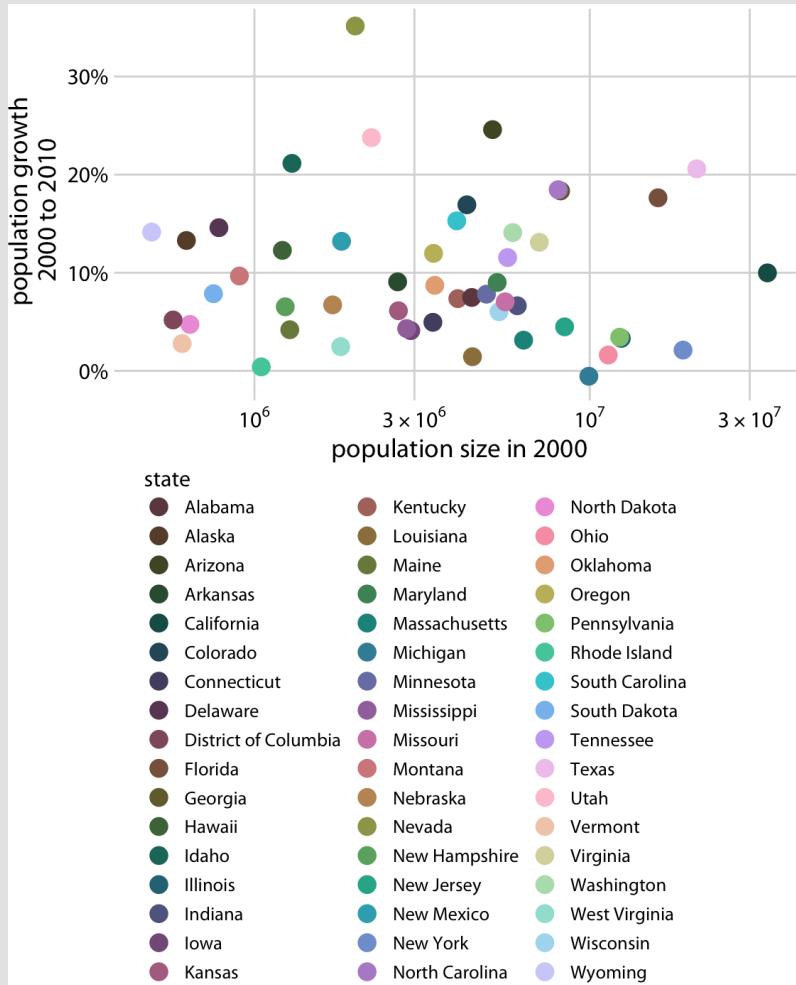




City fuel economy versus engine displacement, for popular cars released between 1999 and 2008. Each point represents one car. The point color encodes the drive train: front-wheel drive (FWD), rear-wheel drive (RWD), or four-wheel drive (4WD).

Best Principle: Make sure your data is visible

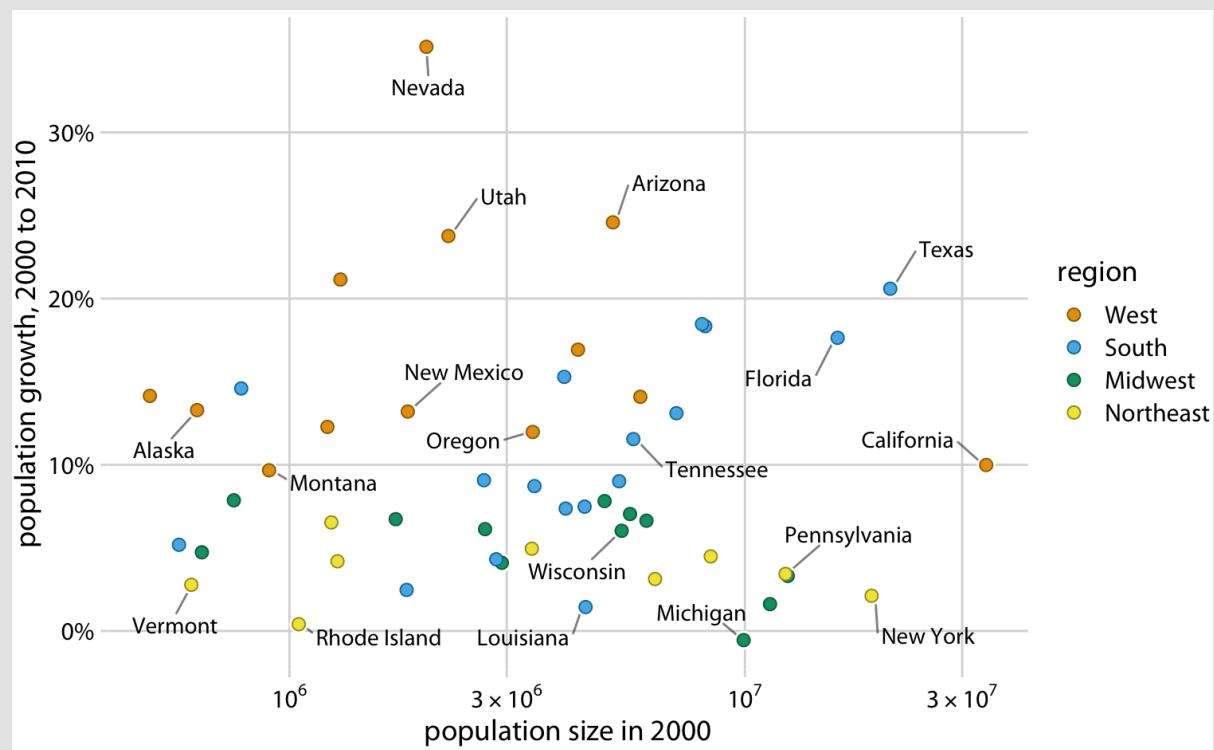




Population growth from 2000 to 2010 versus population size in 2000, for all 50 U.S. states and the District of Columbia. Every state is marked in a different color. Data source: U.S. Census Bureau

Best principle: More is not better (less is more?)

One possible solution to the "busy-ness":

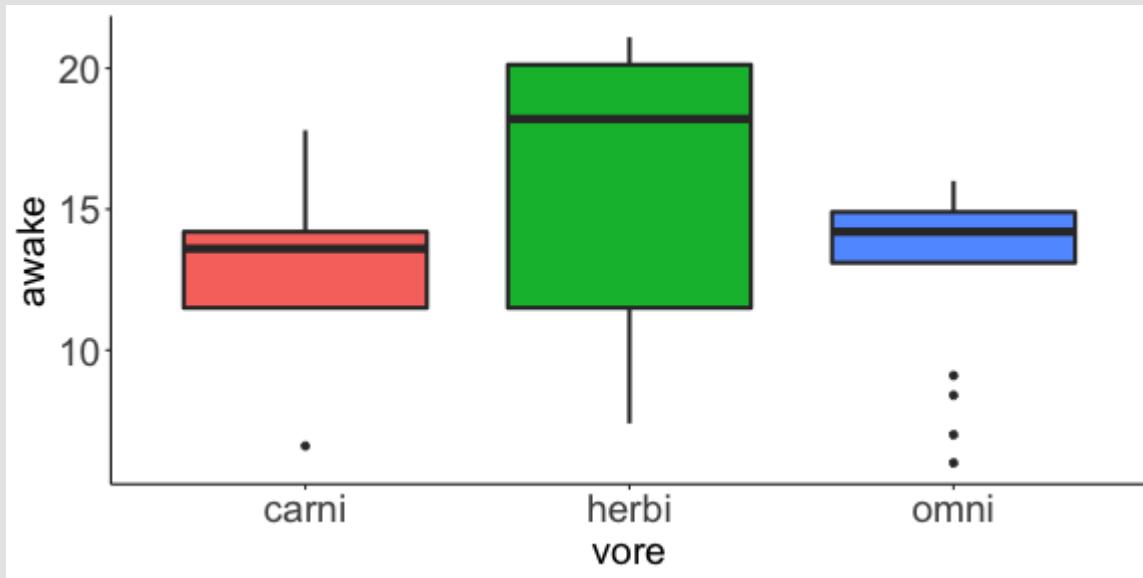


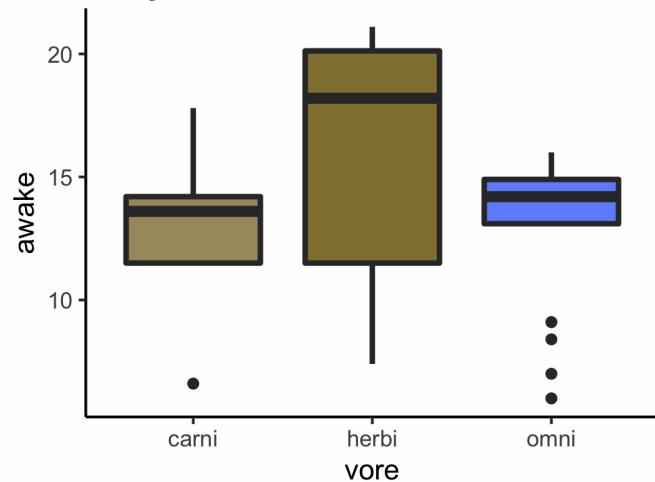
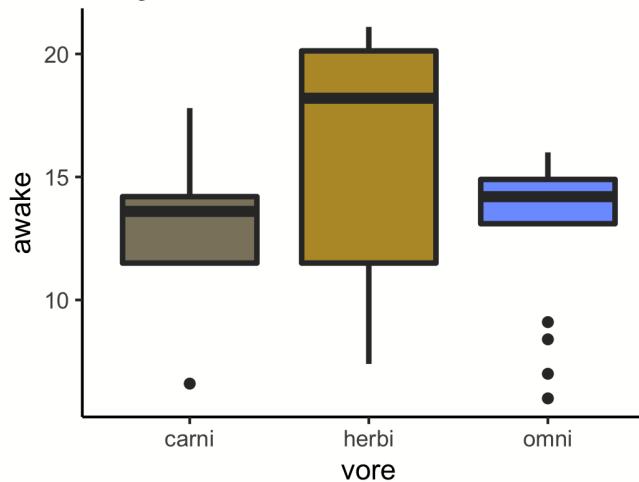
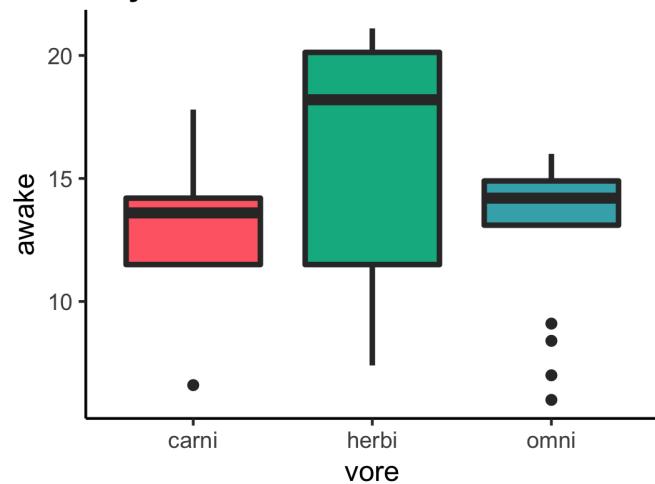
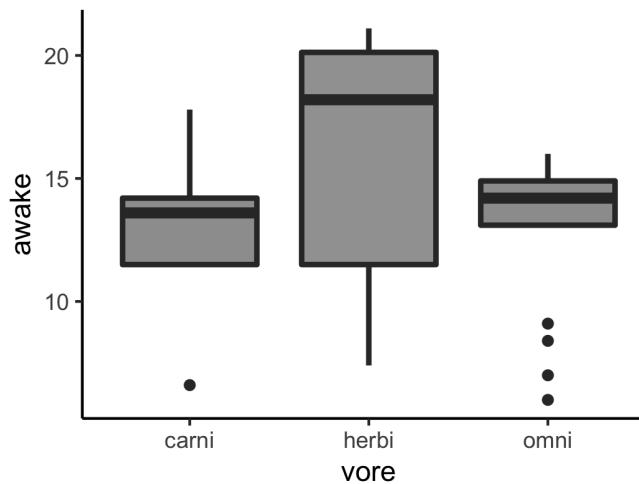
Recall this dataset:

```
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-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
```

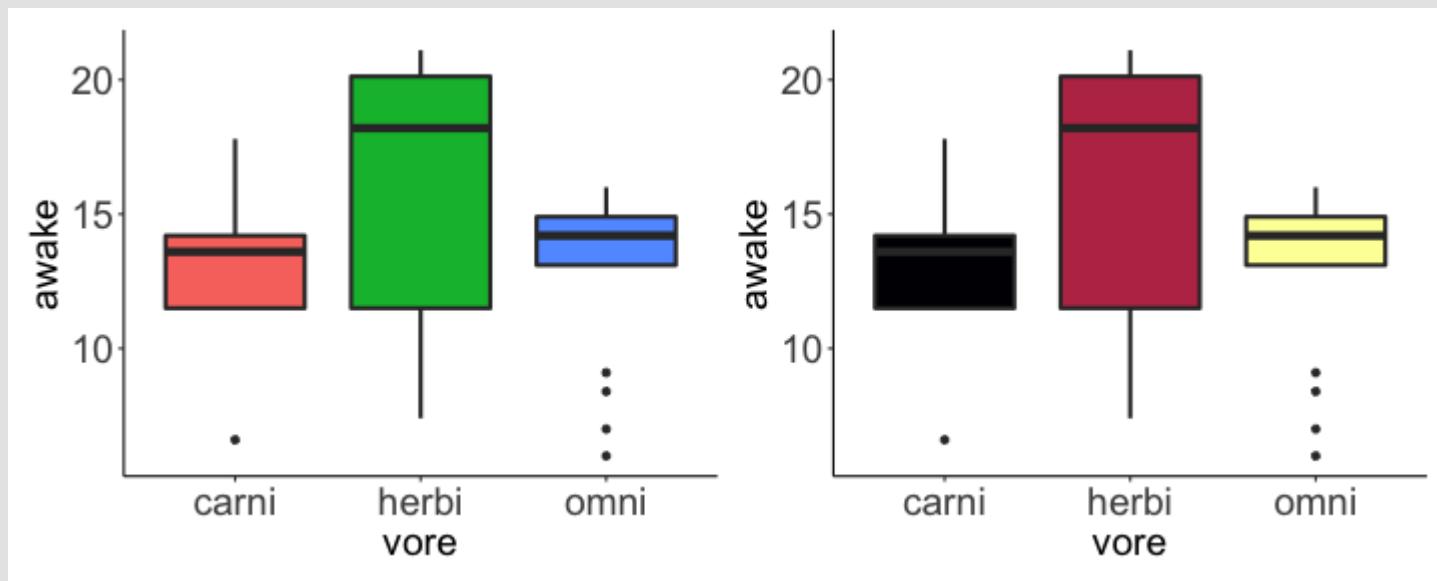
How many hours do the different "vores" spend awake?

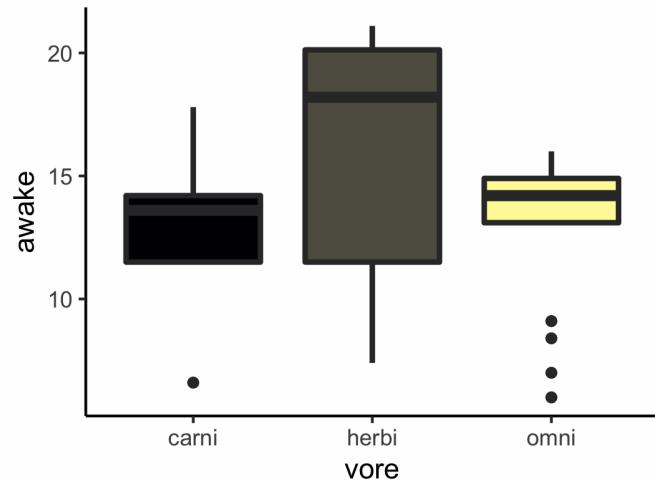
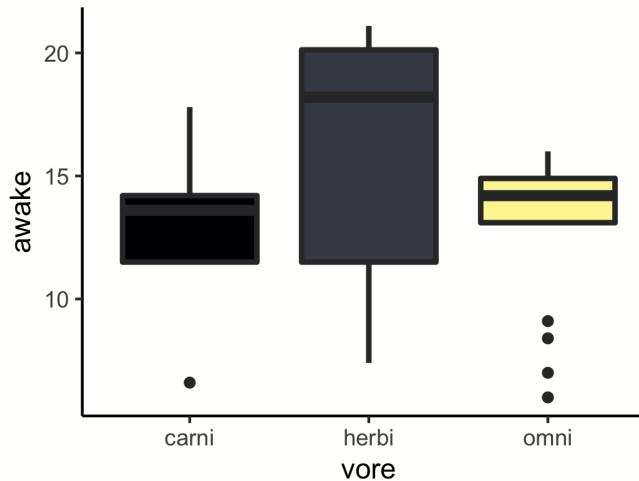
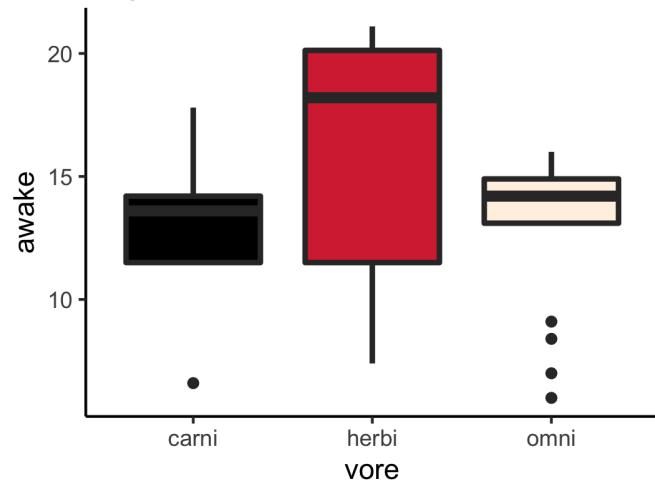
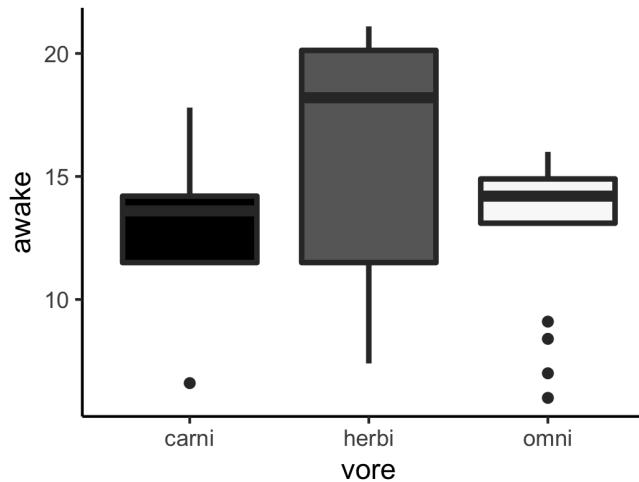


Deutanomaly**Protanomaly****Tritanomaly****Desaturated**

Best Principle: Ensure a colorblind-friendly palette

In this example, color differentiation is not *needed* to interpret the plot. In other cases when colors are needed, always check that your plot is accessible.



Deutanomaly**Protanomaly****Tritanomaly****Desaturated**

Want manual fills (colors)?

```
myplot <- ggplot(msleep_subvore,  
                  aes(x = vore, y = awake, fill = vore)) +  
  geom_violin() +  
  scale_fill_manual(name = "Legend name",  
                    values = c("cornflowerblue",  
                              "coral3",  
                              "blueviolet"))  
  
print(myplot)
```

How to check CVD friendliness

```
library(colorblindr) # installed for you our RStudio Cloud  
cvd_grid(myplot)
```

