DSCI 100 - Introduction to Data Science

Lecture 4 - Data visualization in R

2019-01-23

Housekeeping

- Grades are coming! Thanks for your patience!!!
- Quiz next week!
  - 45 min
  - open book (but not collaborative!)
  - in class, but on Canvas
  - you will get some practice quiz questions by the end of the week

Reminder

Where are we? Where are we going?

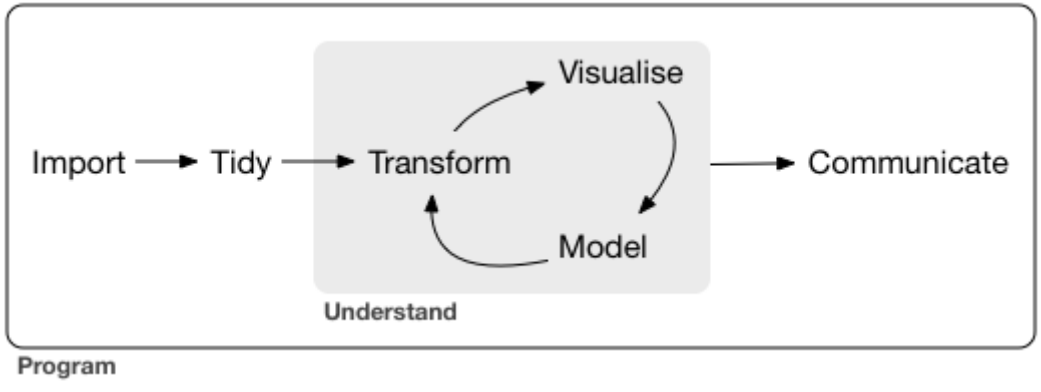


image source: R for Data Science (https://r4ds.had.co.nz/) by Grolemund & Wickham

The basic `ggplot` call:

```
plot_object <- ggplot(dataframe, aes(x = a_column, y = another_column)) +
    geom_something()

plot_object

...

plot_object <- plot_object +
    ...
```

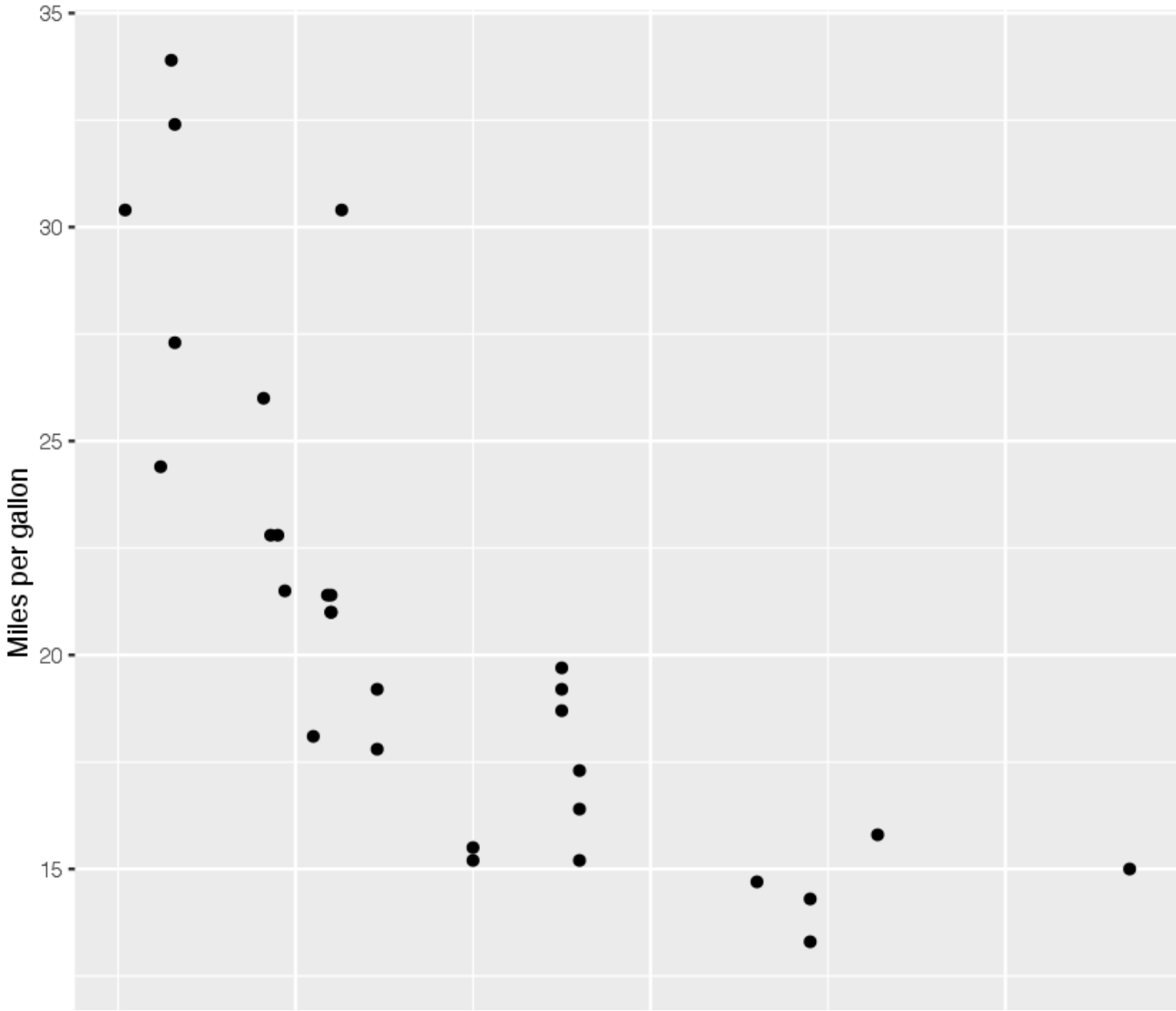
Where to get help (and ideas) for creating `ggplot2` visualizations?

- https://www.r-graph-gallery.com/portfolio/ggplot2-package/ (https://www.r-graph-gallery.com/portfolio/ggplot2-package/)

Only make the plot area as big as needed!

- the default size is ridiculous!

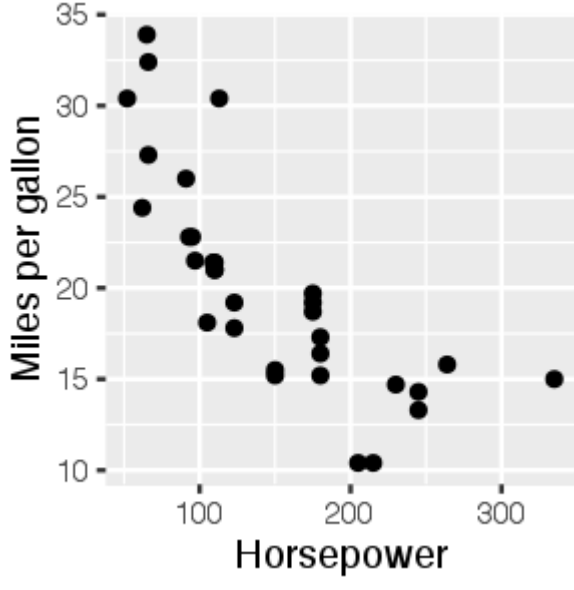
In [2]:
```
too_big
```



- use the `repr` package to set your plot size with R in Jupyter

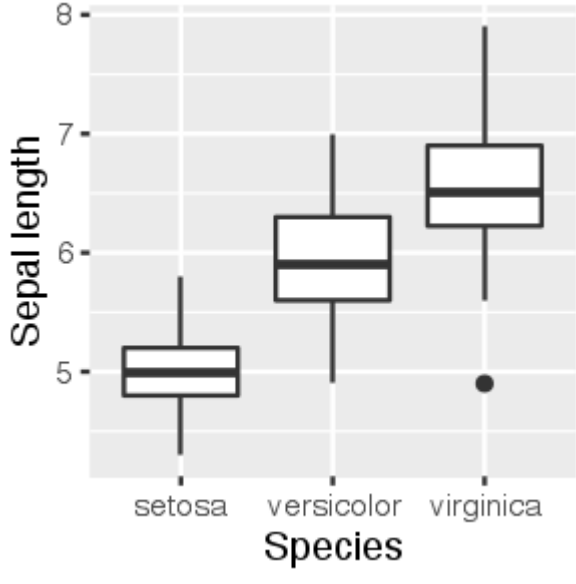
In [3]:
```
library(repr)
options(repr.plot.width = 2.5, repr.plot.height = 2.5)
```

In [4]:
```
too_big
```



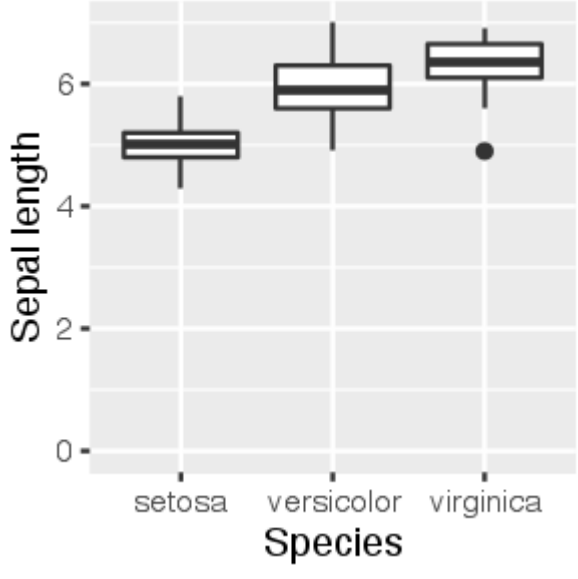
Don’t adjust the axes to zoom in on small differences (if the difference is small, show that its small!)

In [6]:
```
not_a_big_deal
```



In [8]:
```
not_a_big_deal
```

```
Warning message:
“Removed 12 rows containing non-finite values (stat_boxplot).”
```

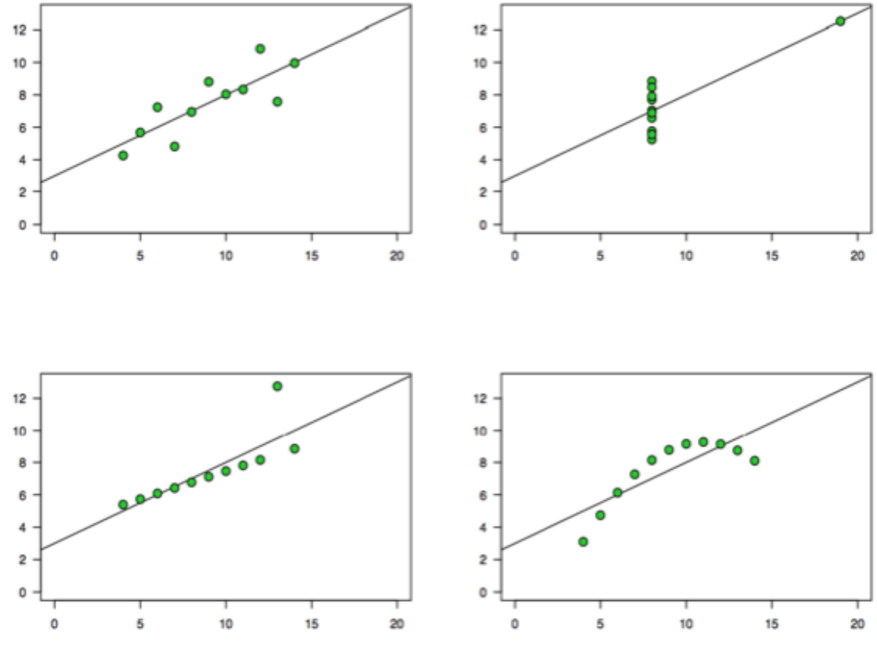


Show the data (don't hide the shape/distribution of the data behind a bar)

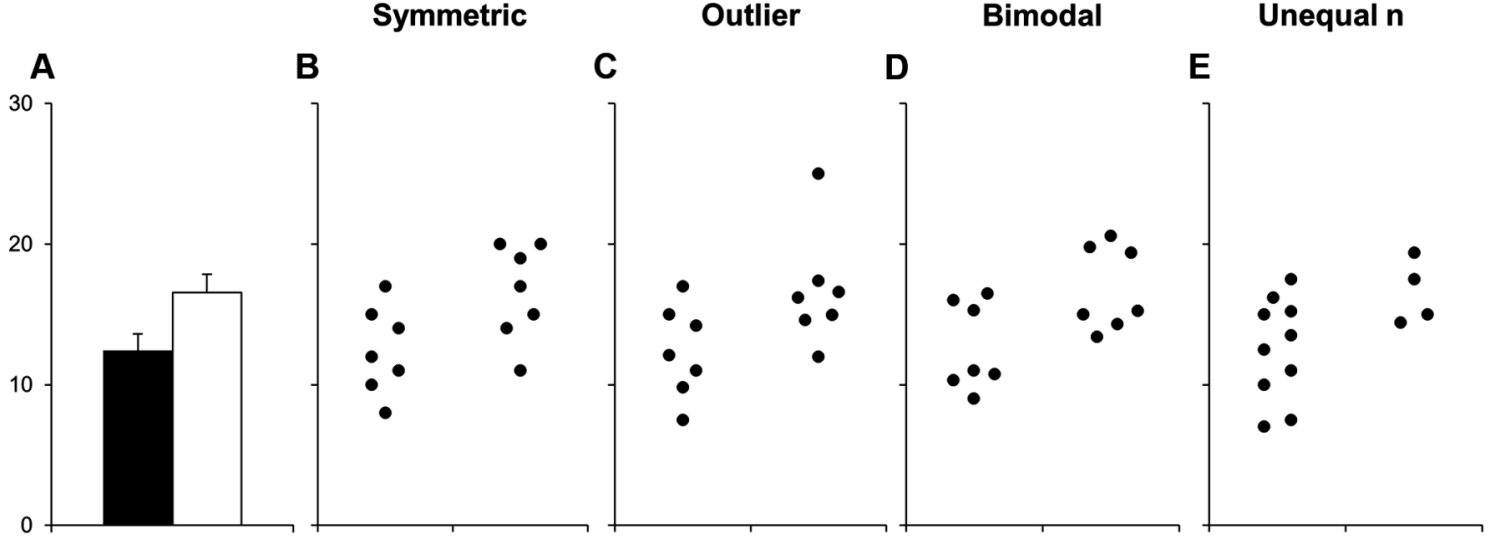
next two slides borrowed from Jeff Leek (http://jtleek.com/genstats_site/lecture_notes/01_09_Exploratory_Analysis.pdf)

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

$\hat{\beta}_0 = 3.0, \quad \hat{\beta}_1 = 0.5, \quad$ p-value (slope) = 0.002, $\quad R^2 = 0.67.$



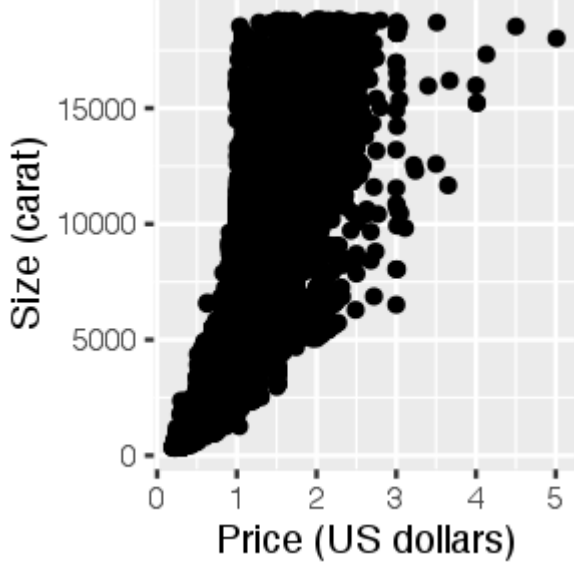
http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128



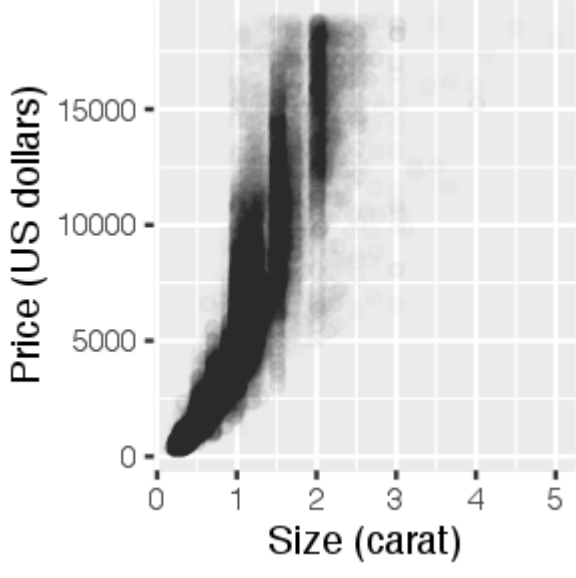
Test				
	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

Be wary of overplotting...

In [10]:
```
too_much
```

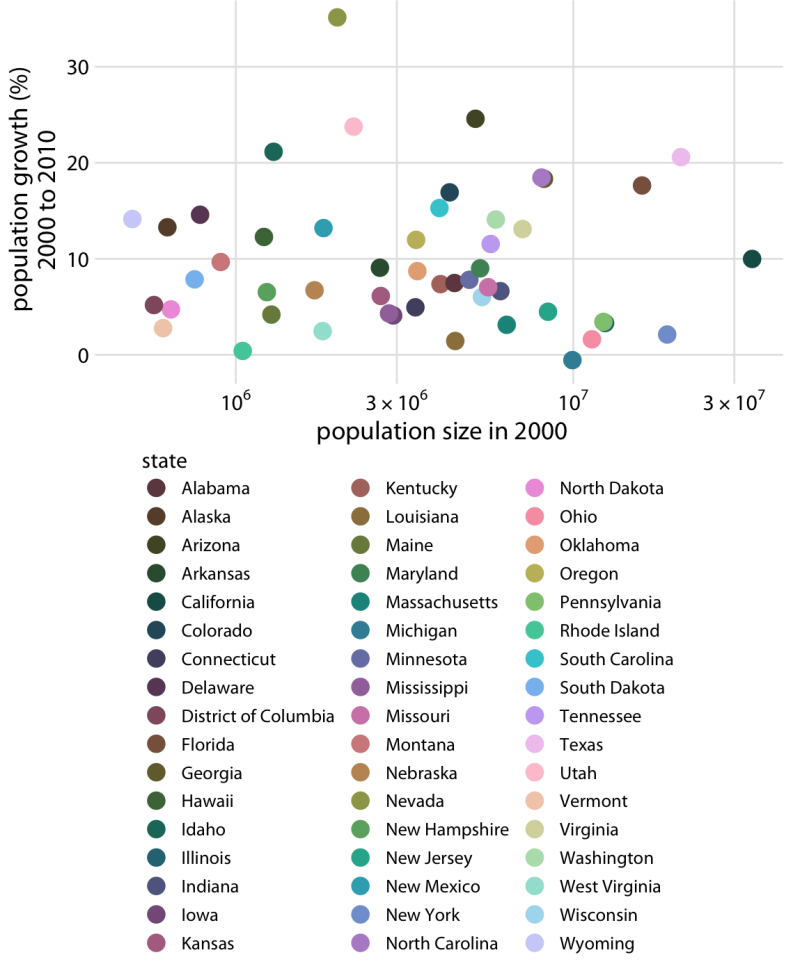


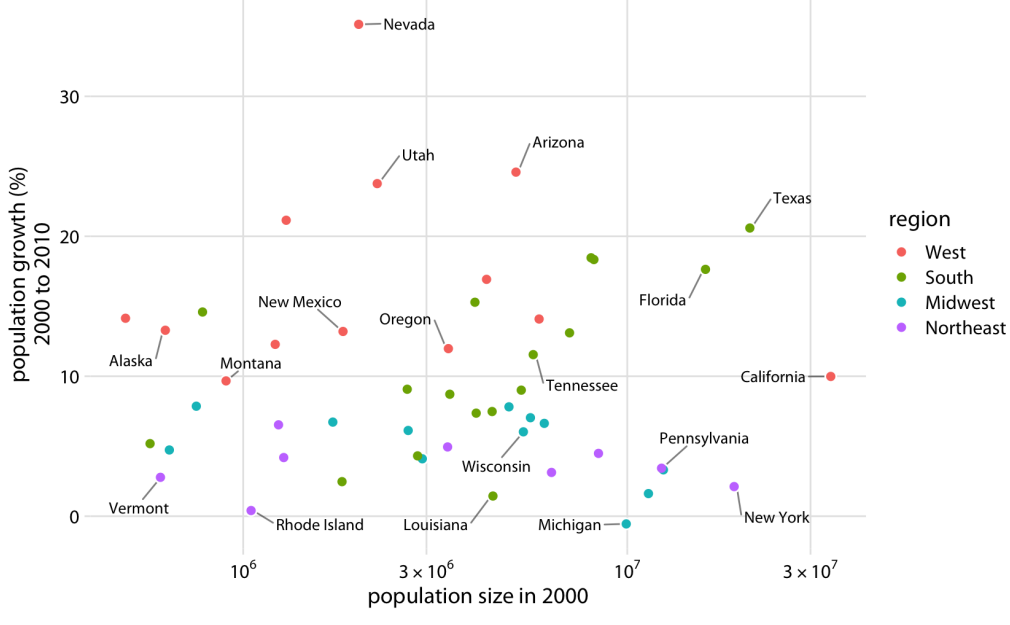
In [12]:
```
an_improvement
```



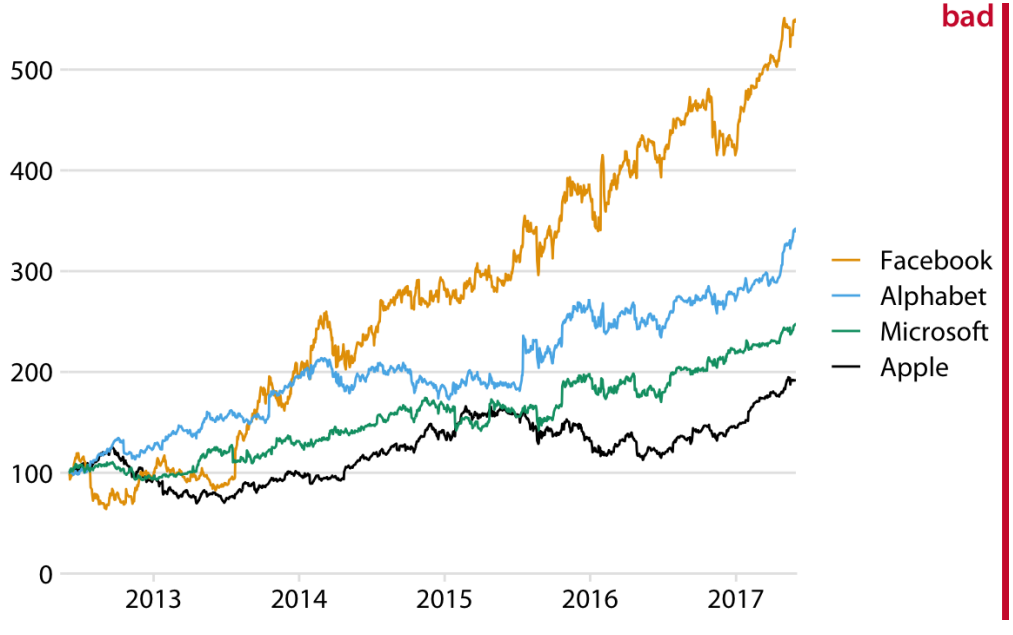
Use colors sparingly

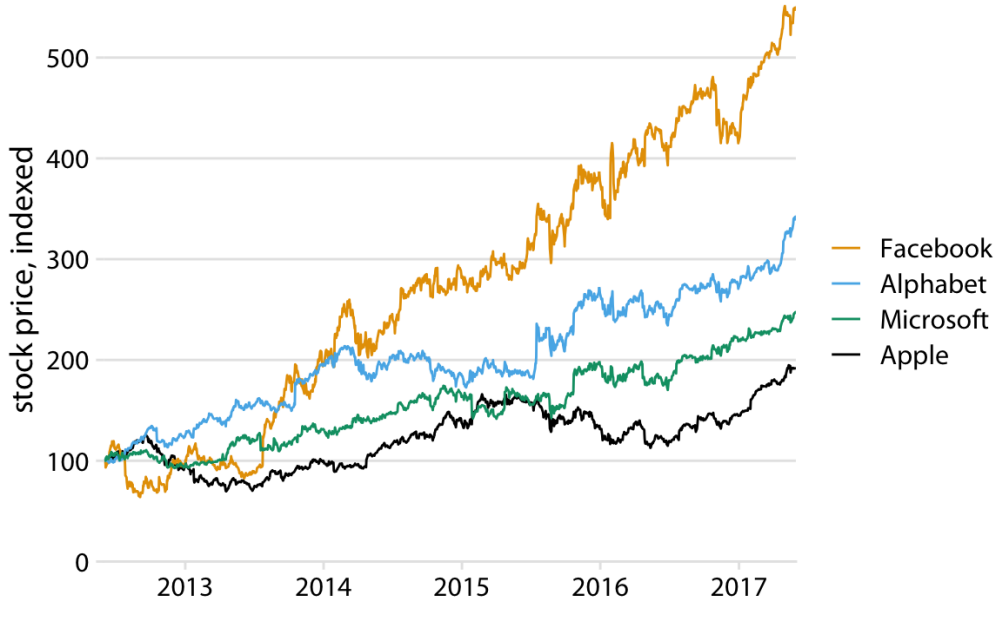
bad


population growth (%)
2000 to 2010
30
20
10
0
10^6
3×10^6
10^7
3×10^7
population size in 2000
state
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

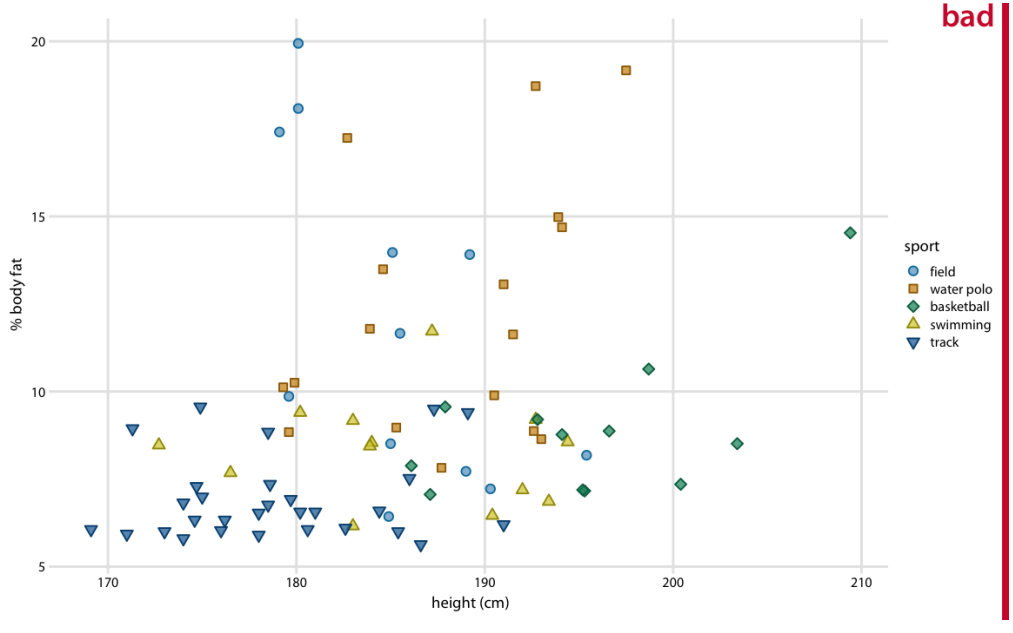

Nevada
Utah
Arizona
Texas
Florida
New Mexico
Oregon
Alaska
Montana
Tennessee
California
Pennsylvania
Wisconsin
Vermont
Rhode Island
Louisiana
Michigan
New York
30
20
10
0
population growth (%)
2000 to 2010
10^6
3×10^6
10^7
3×10^7
population size in 2000
region
West
South
Midwest
Northeast

Use legends and labels so that your visualization is understandable without reading the surrounding text

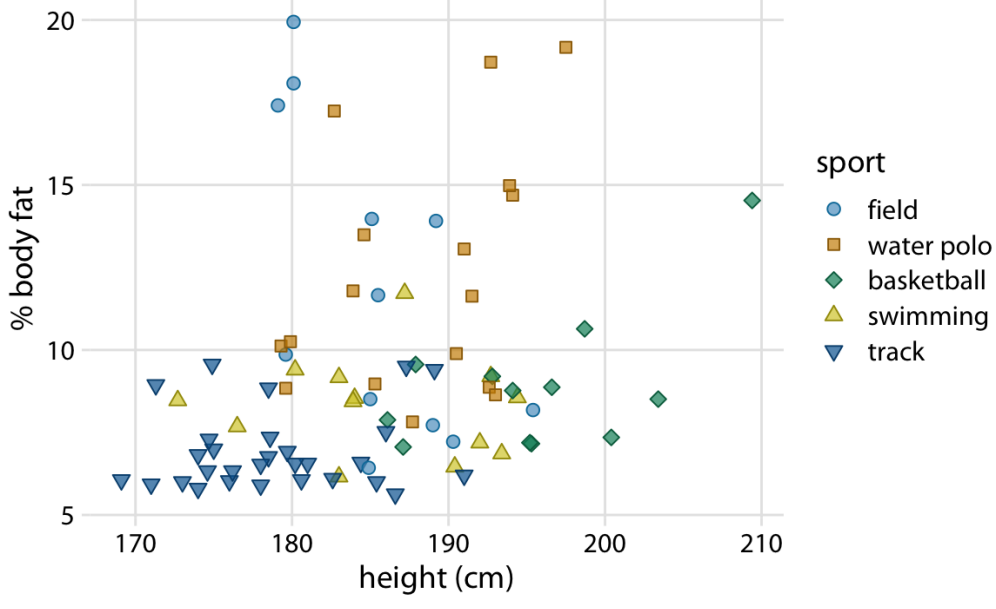

bad
500
400
300
200
100
0
2013
2014
2015
2016
2017
Facebook
Alphabet
Microsoft
Apple


stock price, indexed
500
400
300
200
100
0
2013
2014
2015
2016
2017
Facebook
Alphabet
Microsoft
Apple

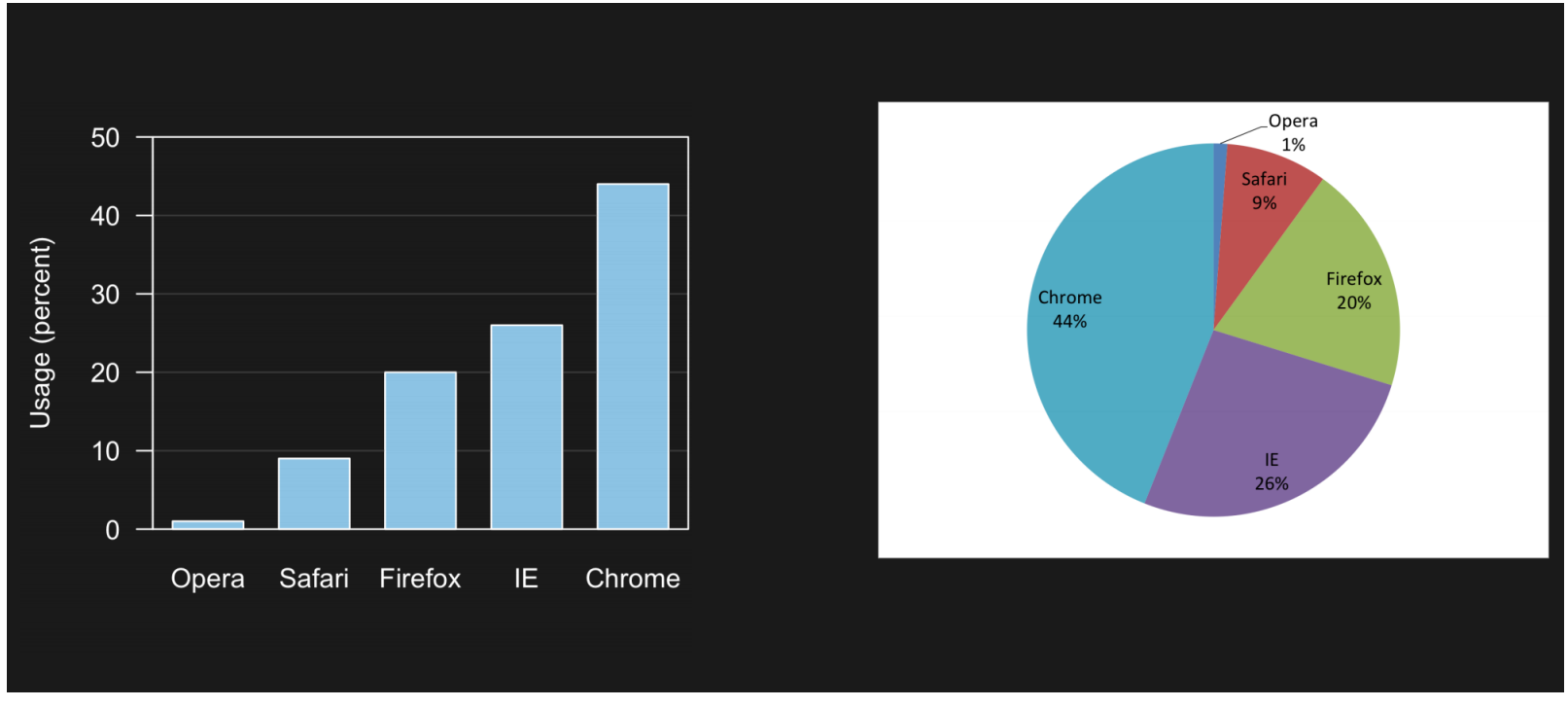
Ensure the text on your visualization is big enough to be easily read


20
15
10
5
% body fat
170
180
190
200
210
height (cm)
sport
field
water polo
basketball
swimming
track

bad

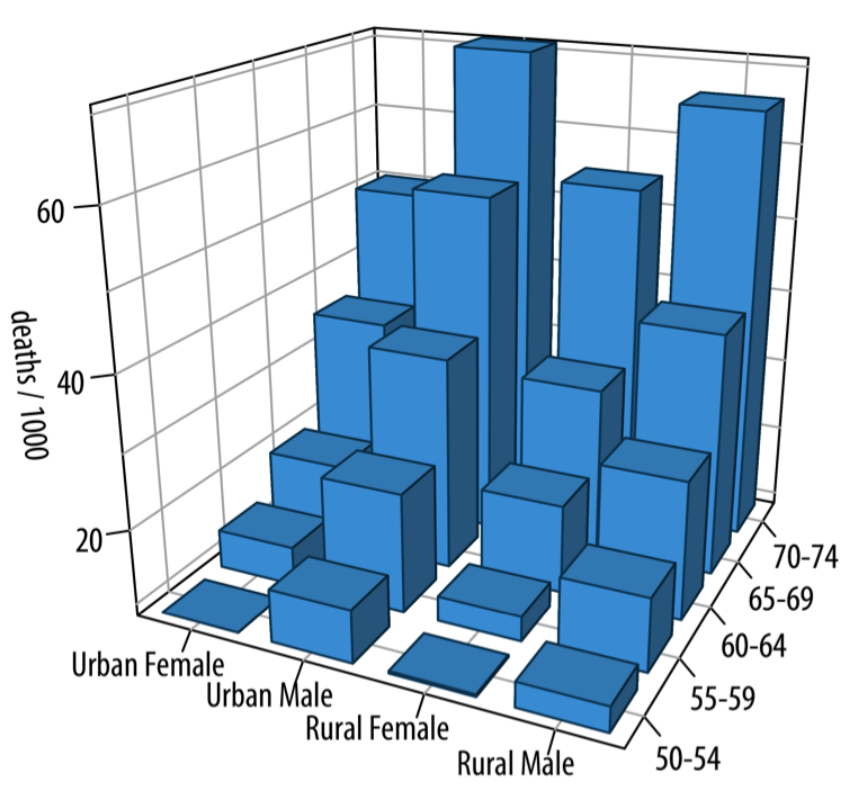

20
15
10
5
% body fat
170
180
190
200
210
height (cm)
sport
field
water polo
basketball
swimming
track

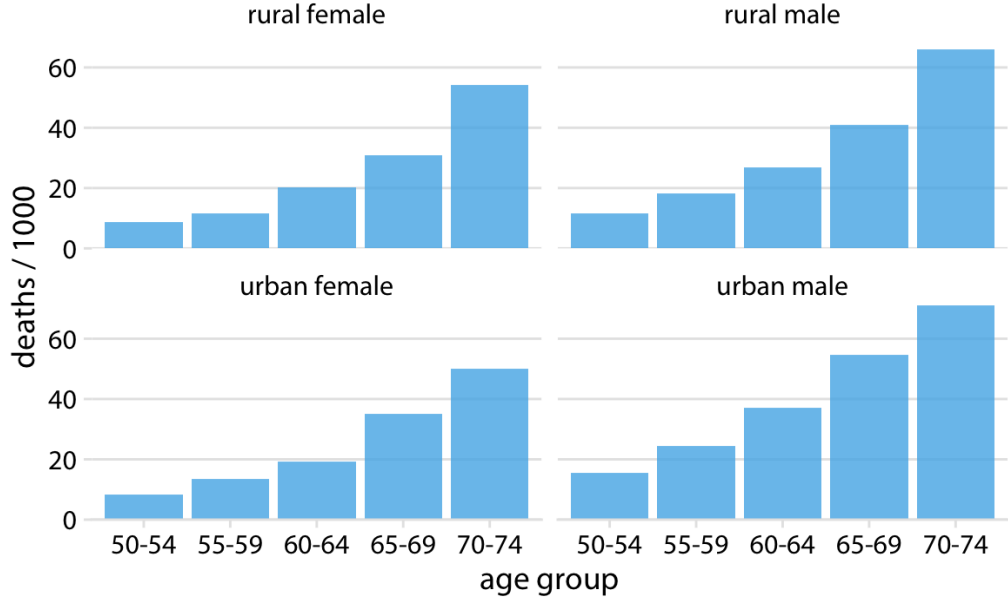
Do not use pie charts!


Usage (percent)
50
40
30
20
10
0
Opera
Safari
Firefox
IE
Chrome
Opera
1%
Safari
9%
Firefox
20%
IE
26%
Chrome
44%

Do not use 3D!

bad


60
40
20
deaths / 1000
Urban Female
Urban Male
Rural Female
Rural Male
50-54
55-59
60-64
65-69
70-74


rural female
rural male
urban female
urban male
deaths / 1000
0
20
40
60
50-54
55-59
60-64
65-69
70-74
age group

Go and create!

Make an effective plot!

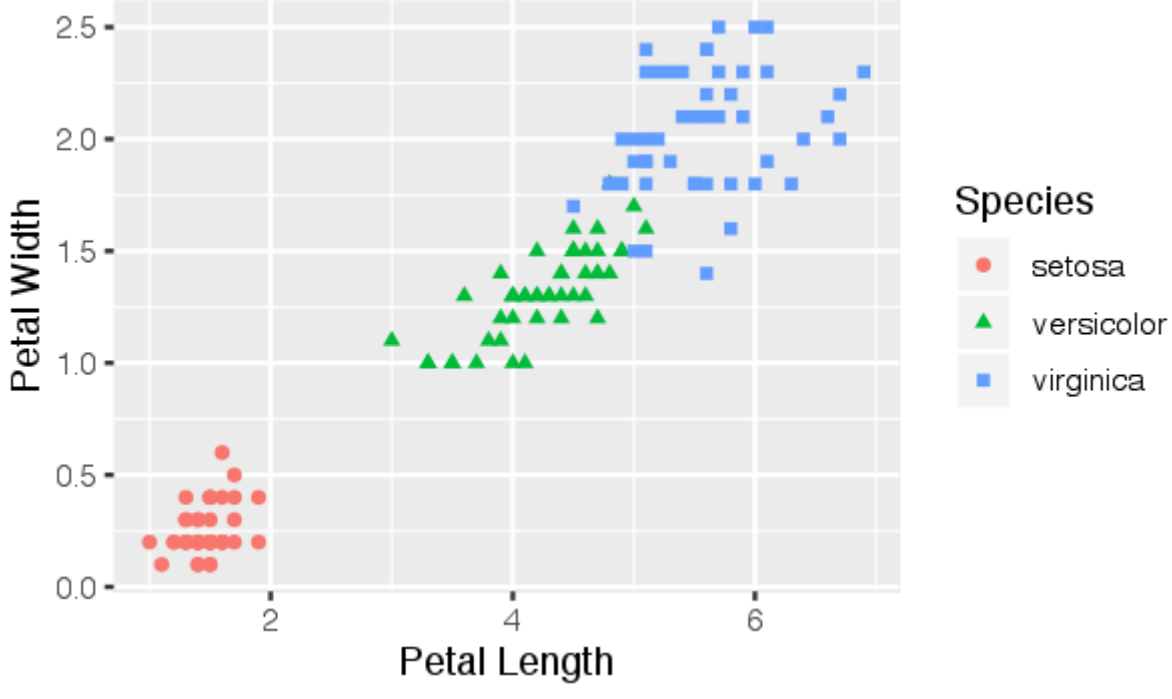
Can petal length and petal width be used to separate the Iris flower species? Create a plot to answer this question!

In [14]:
```
head(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

In [17]:

```
# solution
library(tidyverse)
options(repr.plot.width = 5, repr.plot.height = 3)
plot <- ggplot(iris,aes(x = Petal.Length, y = Petal.Width)) +
    geom_point(aes(color = Species, shape = Species)) +
    xlab("Petal Length") +
    ylab("Petal Width")
plot
```



What did we learn today

- Combine color and shape for points to separate groups in scatter plots
- Rarely use 3D or pie charts, there are other, more effective ways of communicating the information
- Some rules of thumb/guidelines for making visualizations
- How to negatively filter using `!`
- How to use `options` from the `repr` packages to set plot size (and that we should set plot size)!