Dataviz principles

Rafael A. Irizarry

Introduction

These slides are inspired by:

"Creating effective figures and tables" by Karl Broman. Code available on GitHub repository. Introduction to Data Visualization course by Peter Aldhous

It is a chapter in a data visualization chapter in this book

Introduction

- We show some examples of plot styles we should avoid, explain how to improve them, and use these as motivation for a list of principles.
- We compare and contrast plots that follow these principles to those that don’t.

Introduction

- The principles are mostly based on research related to how humans detect patterns and make visual comparisons.
- The preferred approaches are those that best fit the way our brains process visual information.
- When deciding on a visualization approach it is also important to keep our goal in mind.
- We may be comparing a viewable number of quantities, describing distribution for categories or numeric values, comparing the data from two groups, or describing the relationship between two variables.

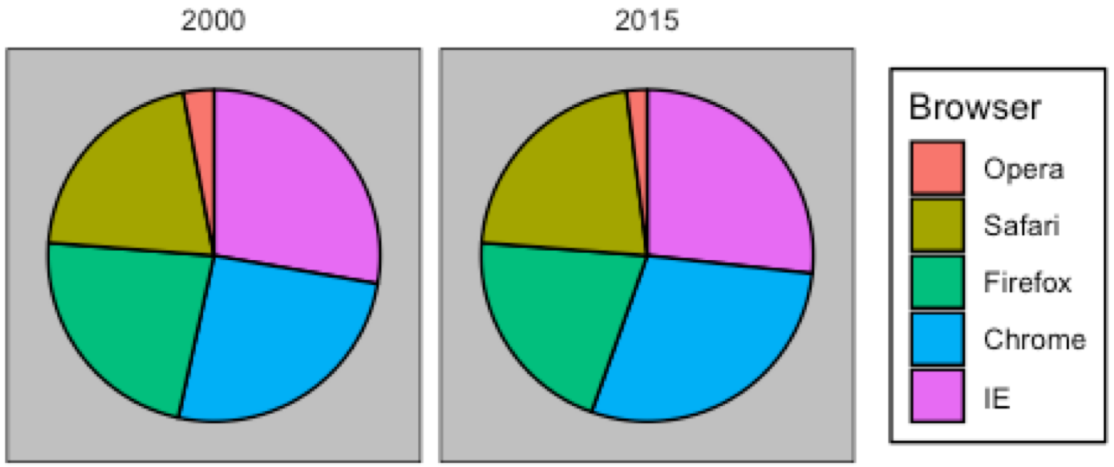
Encoding data using visual cues

- We start by describing some principles for encoding data.
- There are several approaches at our disposal including:
  - position
  - aligned lengths
  - angles
  - area
  - brightness
  - color hue.

First example

- To illustrate how some of these strategies compare let's suppose we want to report the results from two hypothetical polls asking regarding browser preference taken in 2000 and then 2015.
- Here, for each year, we are simply comparing four quantities, four percentages.
- A widely used graphical representation of percentages, popularized by Microsoft Excel, is the pie chart:

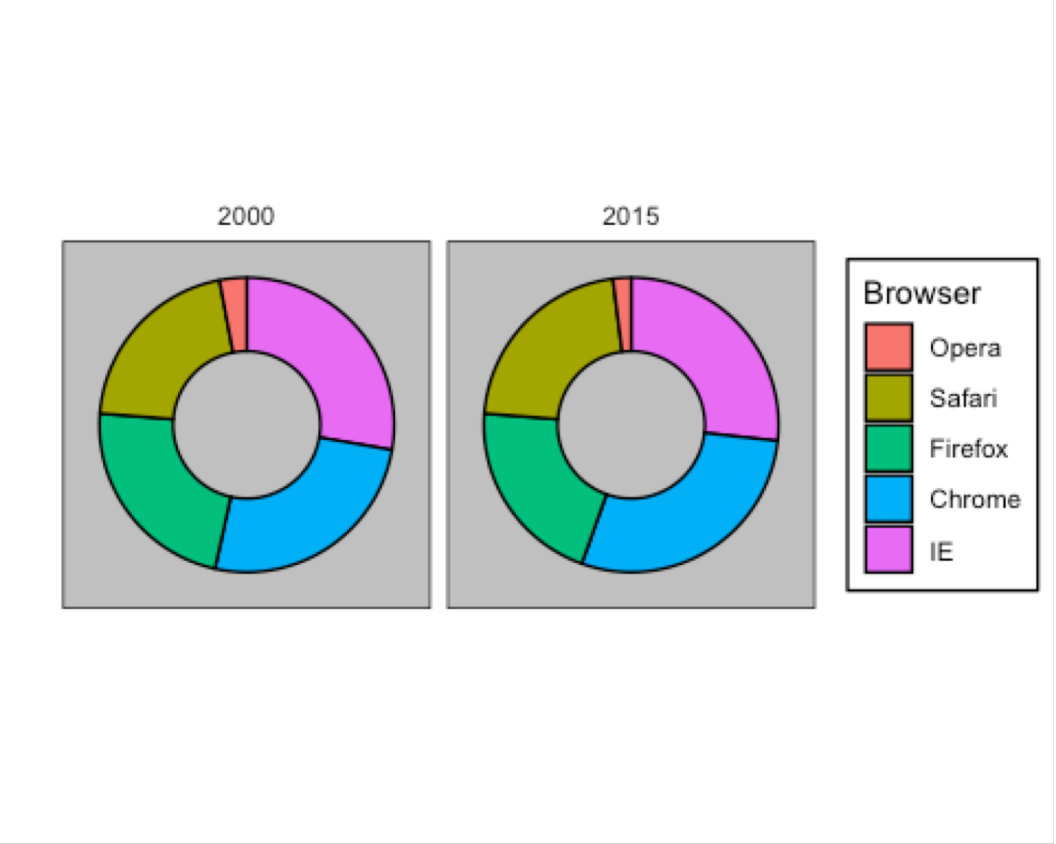
First example



Encoding data with angles and areas: not recommended

- Here we are representing quantities with both areas and angles since both the angle and area of each pie slice is proportional to the quantity it represents.
- This turns out to be a sub optimal choice since, as demonstrated by perception studies, humans are not good at precisely quantifying angles and are even worse when only are is available.

Encoding data with just area: even less recommended



Pie chart of browser usage.

Pie charts

- To see how hard it is to quantify angles and are note that the rankings and all the percentages in the plots above changed fro 2000 to 2015.
- Can you determine the actual percentages and rank the browsers' popularity? Can you see how the percentages changed from 2000 to 2015? It is not easy to tell from the plot.

Pie charts

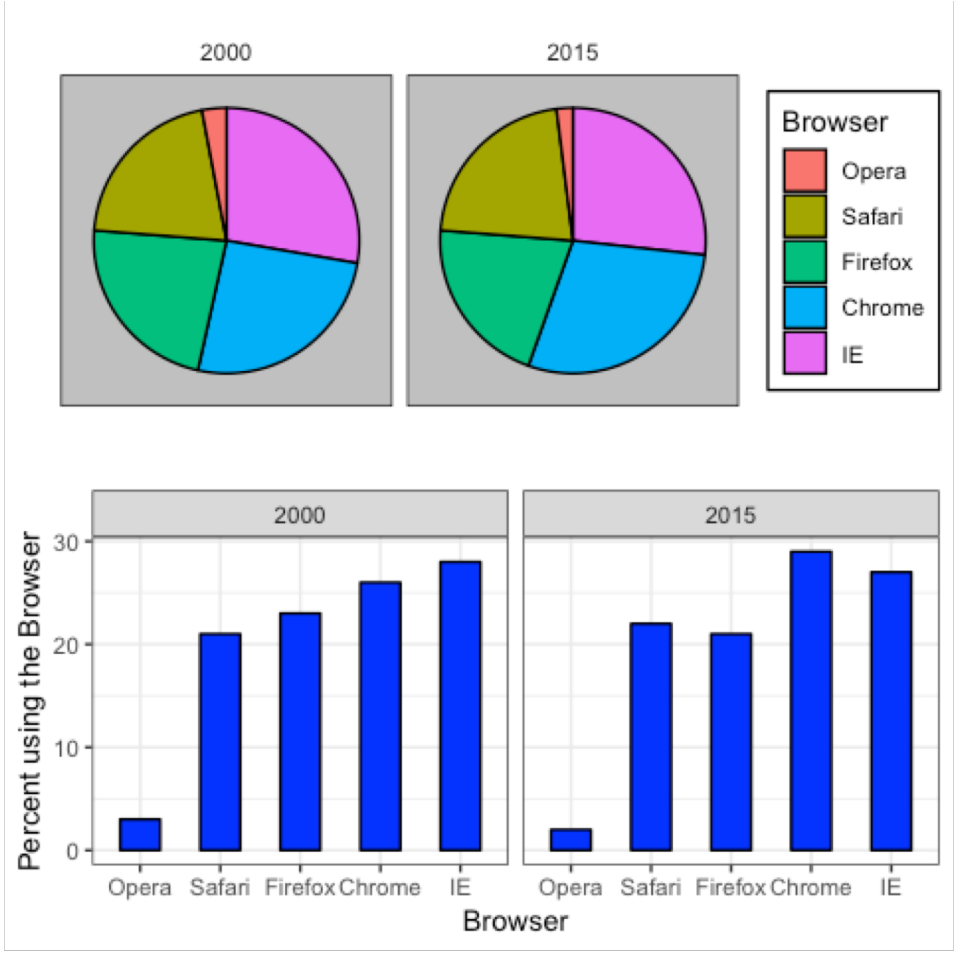
- In this case, simply showing the numbers is not only clearer, but it would saves on print cost if making a paper version.

Browser	2000	2015
Opera	3	2
Safari	21	22
Firefox	23	21
Chrome	26	29
IE	28	27

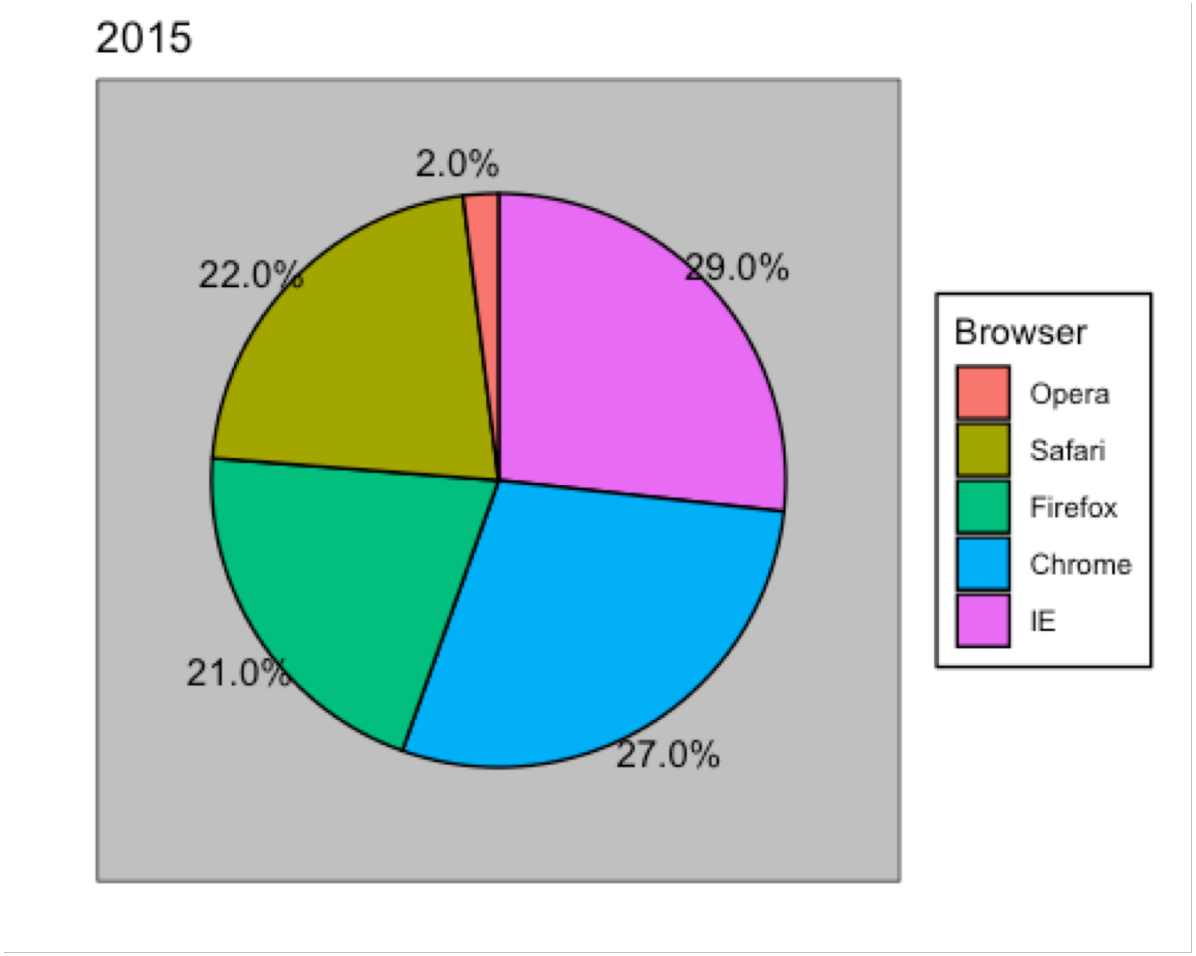
barplots

- The preferred way to plot quantities is to use length and position since humans are much better at judging linear measure.
- The barplot uses bars use this approach by using bars of length proportional to the quantities of interest.
- By adding horizontal lines at strategically chosen values, in this case at every multiple of 10, we ease the quantifying through the position of the top of the bars.

pie chart vs barplots



If forced to make a pie chart at percentages



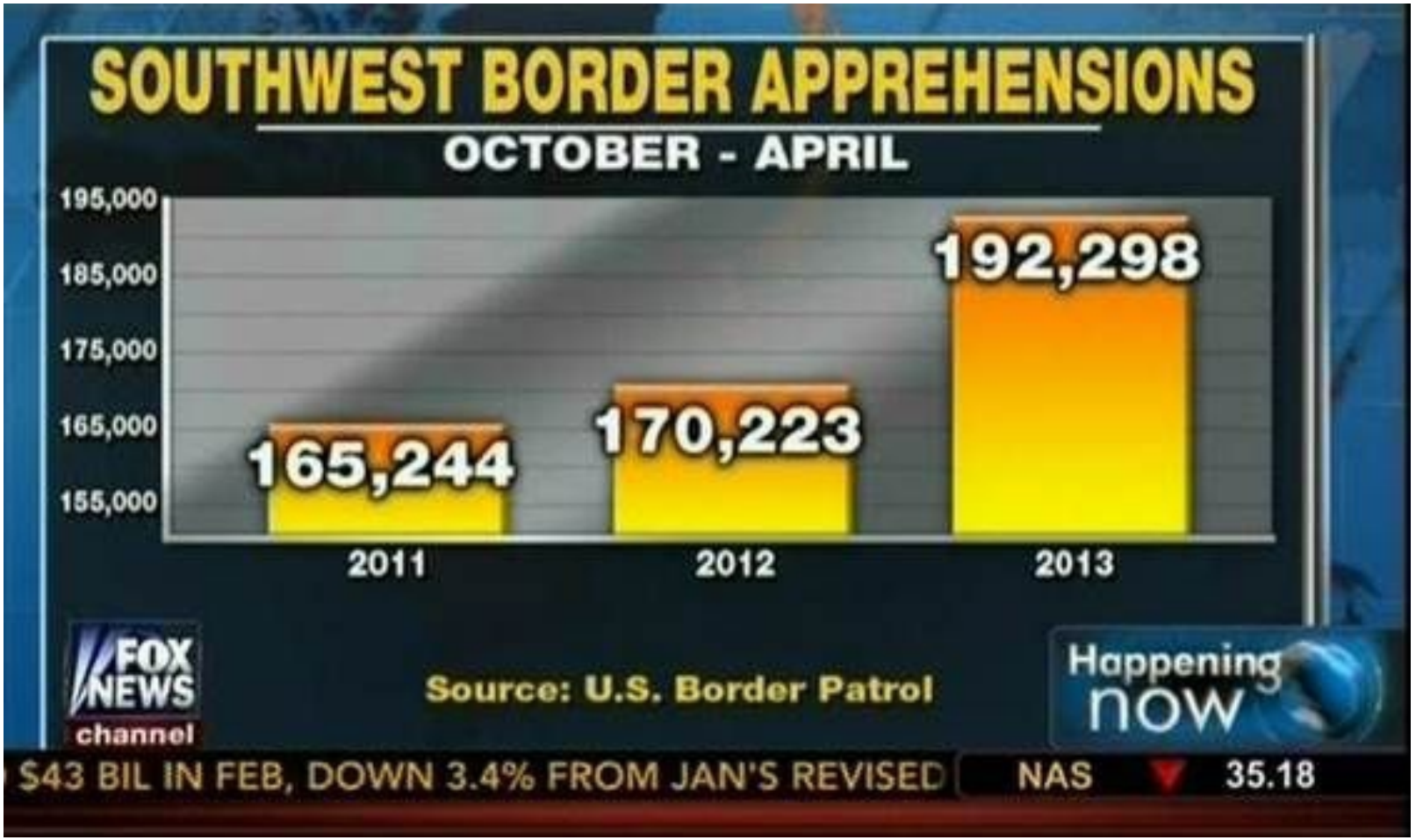
Visual cues

- Position and length are the preferred ways to display quantities over angles which are preferred to area.
- Brightness and color are even harder to quantifying that angles and area but, as we will see later, they are sometimes useful when more than two dimensions are being displayed.

When to include 0

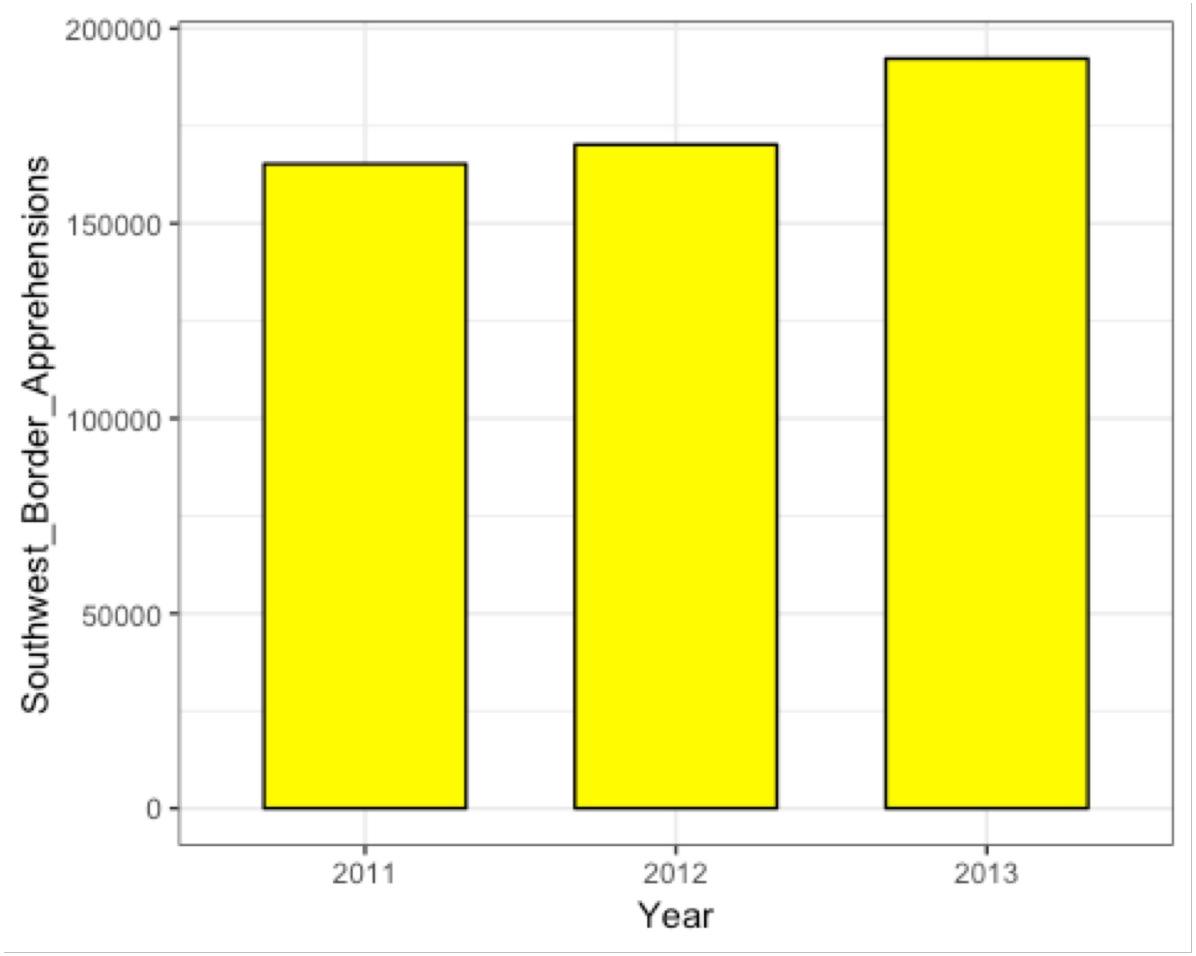
- When using length (e.g. barplots) it is misleading not to start the bars at 0.
- This is because, by using a barplot, we are implying the length is proportional to the quantities being displayed.
- By avoiding 0, relatively small difference can be made to look much bigger than they actually are.
- This approach is often used by politicians or media organizations trying to exaggerate a difference.

Example

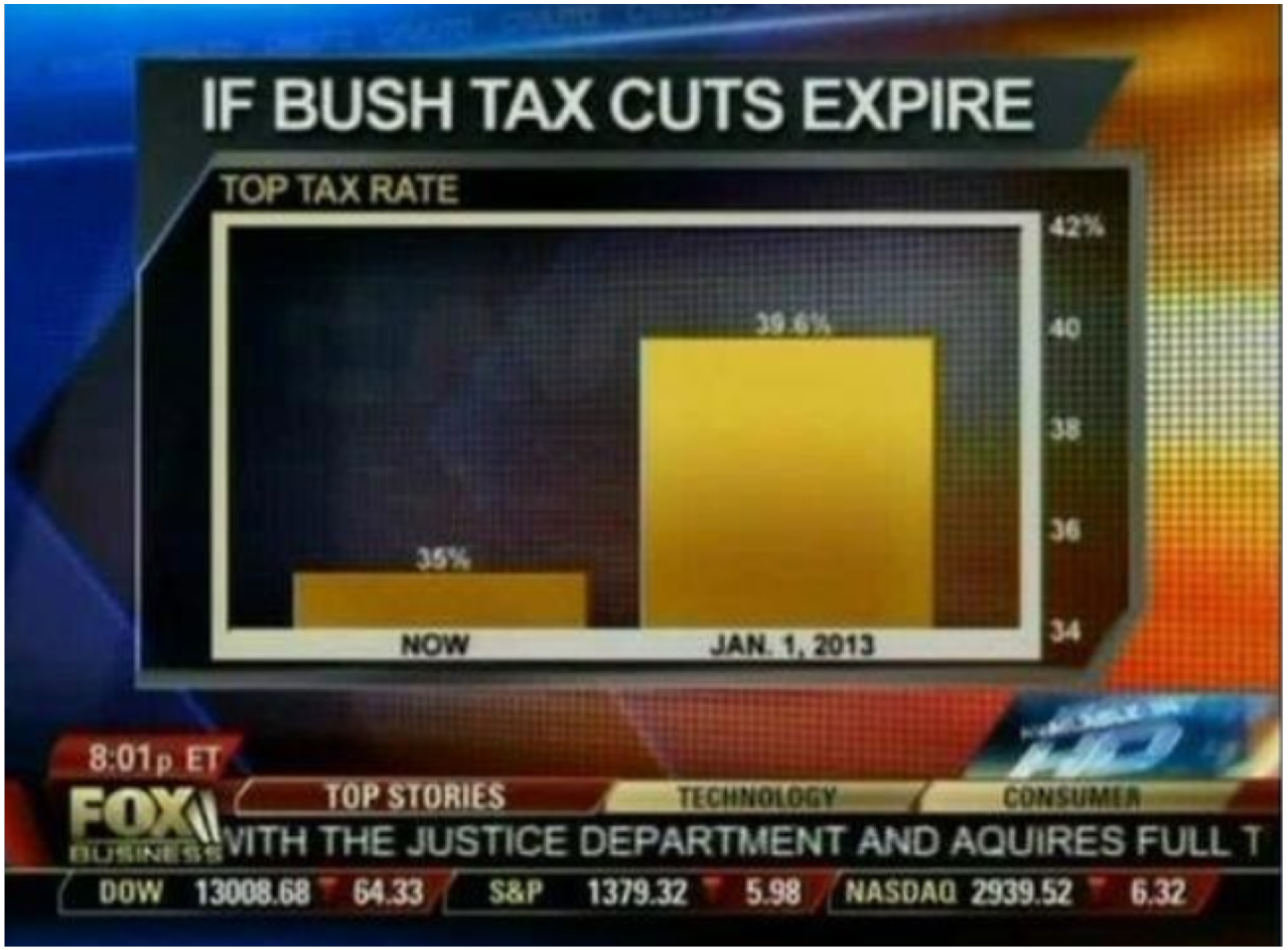

SOUTHWEST BORDER APPREHENSIONS
OCTOBER - APRIL
195,000
185,000
175,000
165,000
155,000
165,244
170,223
192,298
2011
2012
2013
FOX NEWS channel
Source: U.S. Border Patrol
Happening now
$43 BIL IN FEB, DOWN 3.4% FROM JAN'S REVISED
NAS
35.18

– (Source: Fox News, via Media Matters via Fox News via Peter Aldhous

Same data with plot that includes 0

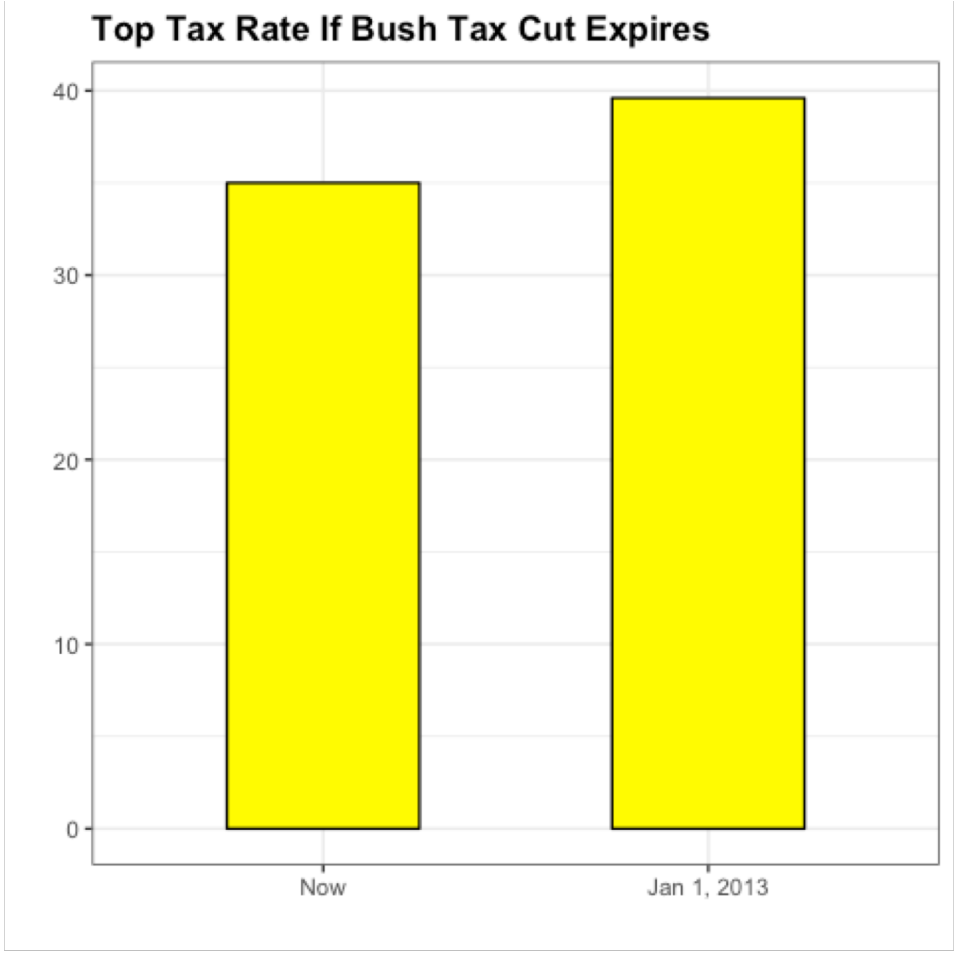


Another example



– Via Fox News via Flowingdata

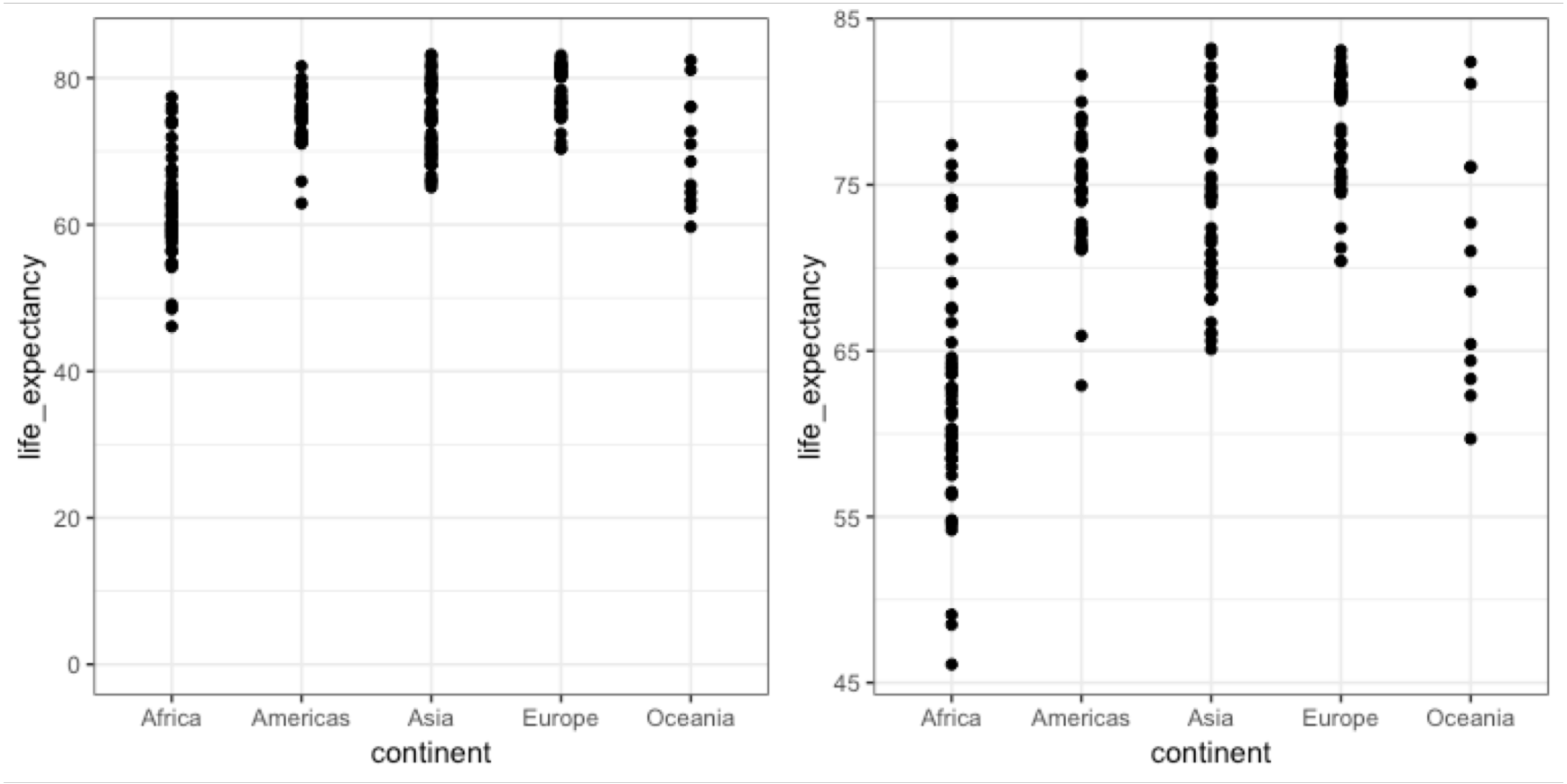
Same data with plot the includes 0



When not to include 0

- When using position rather than length, it is **not** necessary to include 0.
- This is particularly the case when we want to compare differences between groups relative the variability seen within the groups.

Example: Life expectancy by continent in 2012



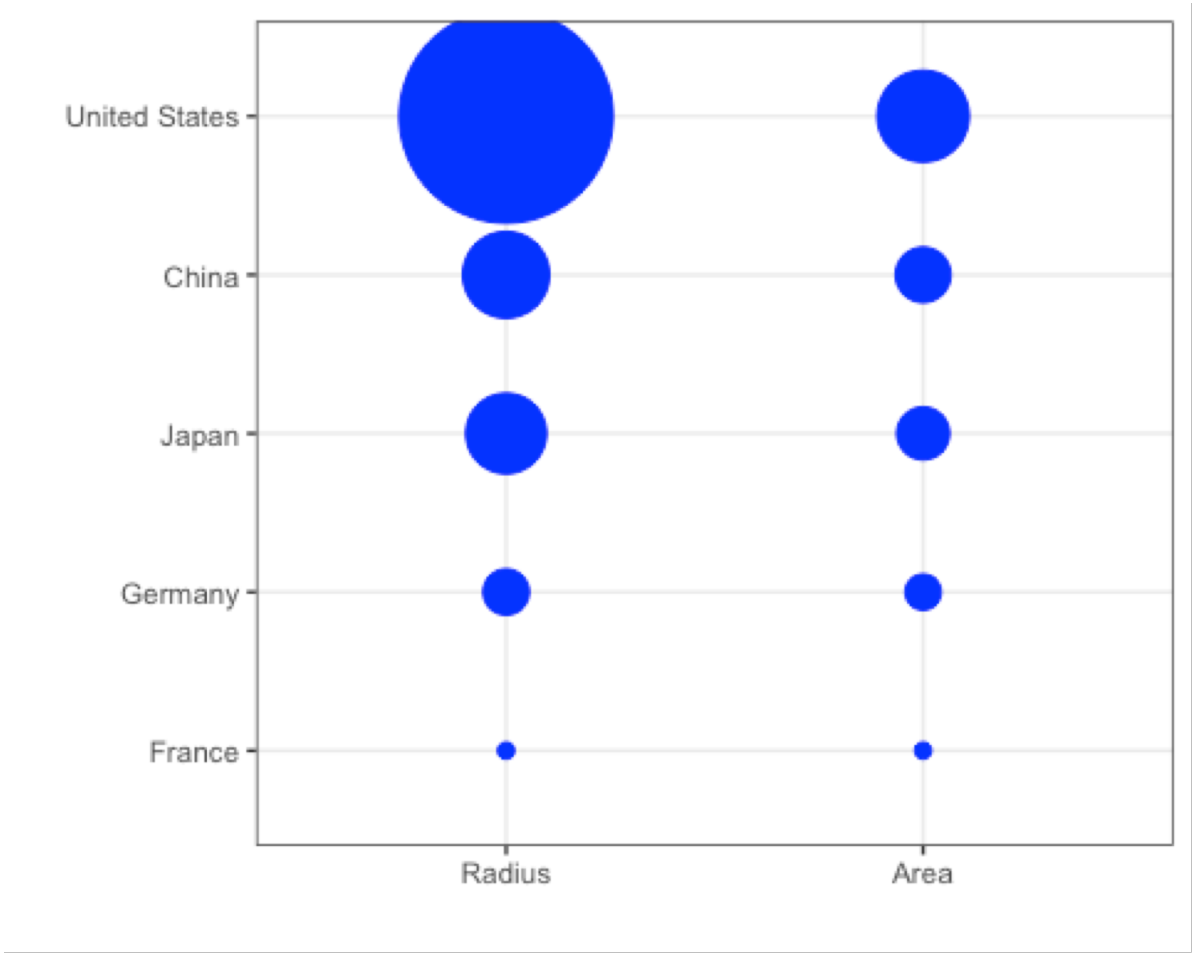
Do not distrort quantities

(Source: President Barack Obama's 2011 State of the Union Address via [Peter Aldhous] (http://paldhous.github.io/ucb/2016/dataviz/index.html))

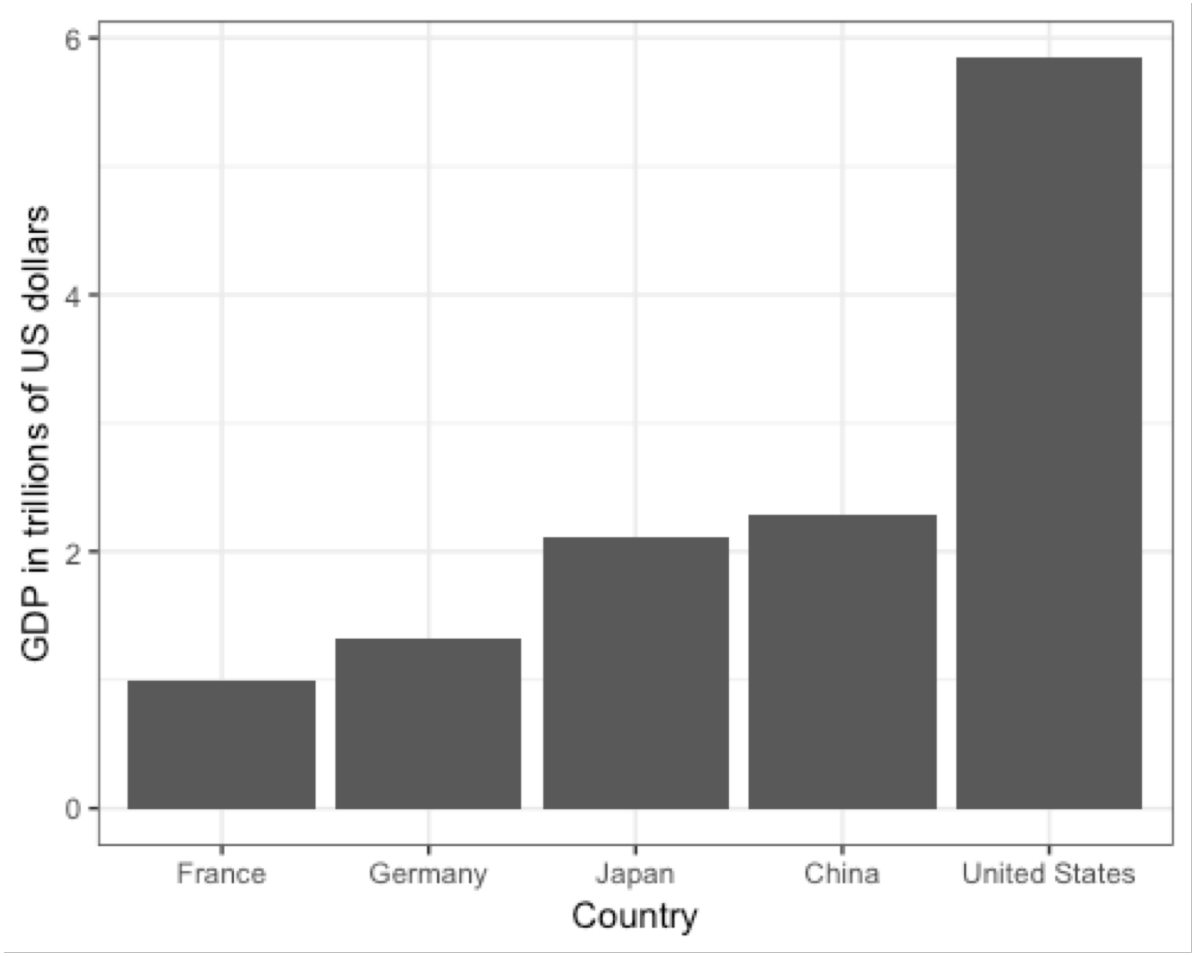
Do not distrort quantities

- The reason for this distortion is that the radius, rather than the area, was made to be proportional to the quantity.
- This implies that the proportion between the areas is squared: 2.6 turns into 6.5 and 5.8 turns into 34.1.

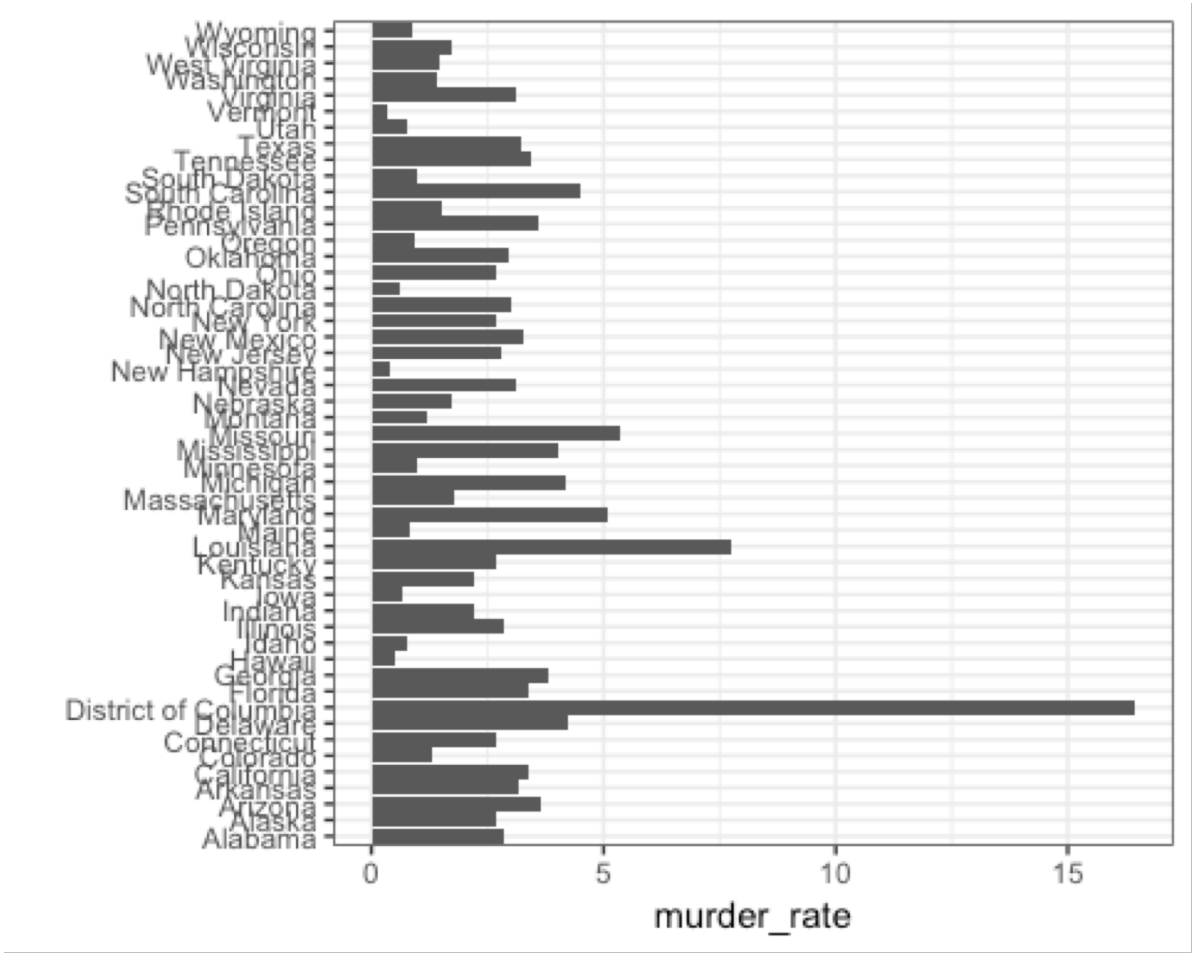
Use area not radius



But we should not be using area or radius



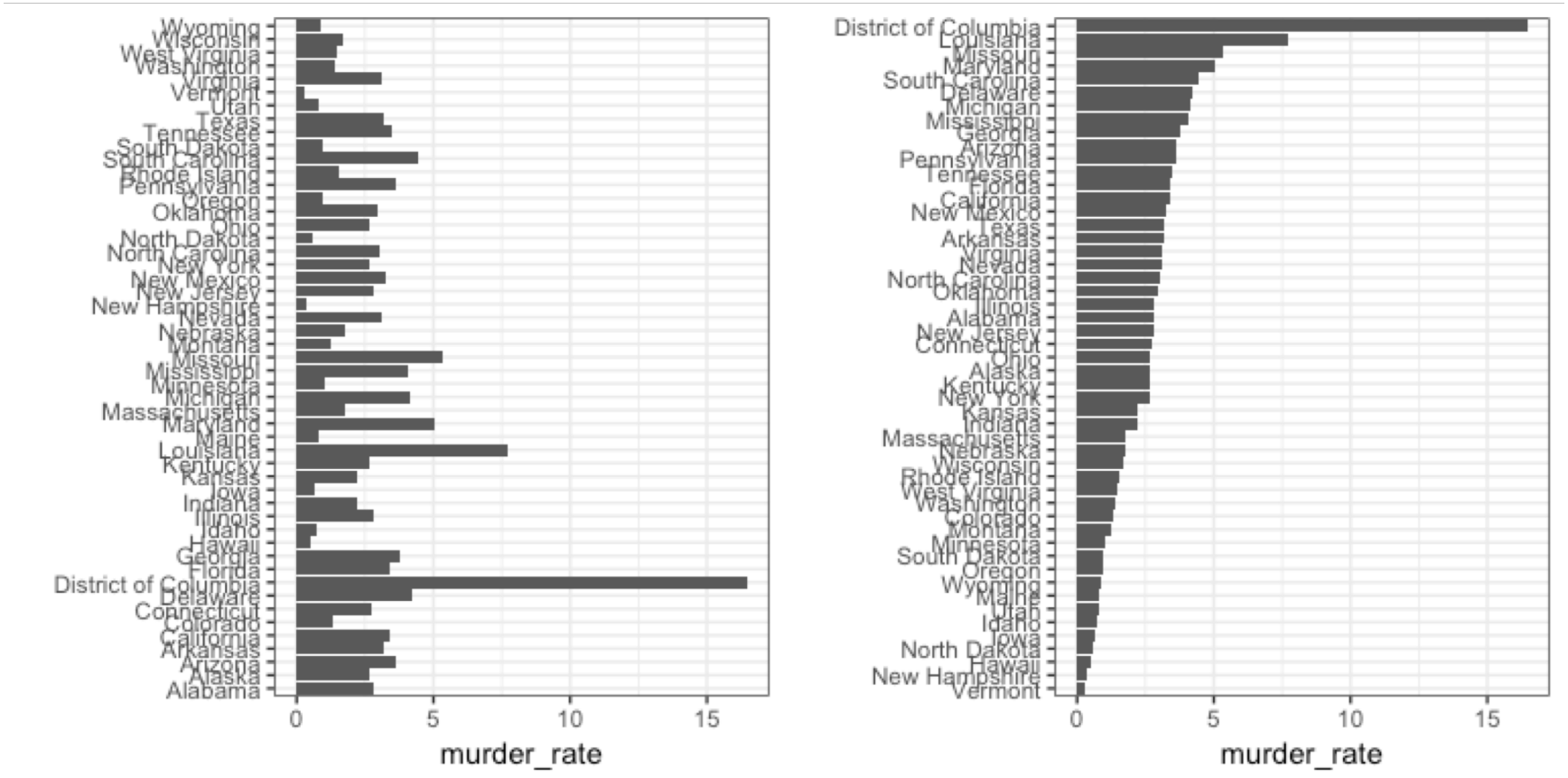
Order by a meaningful value



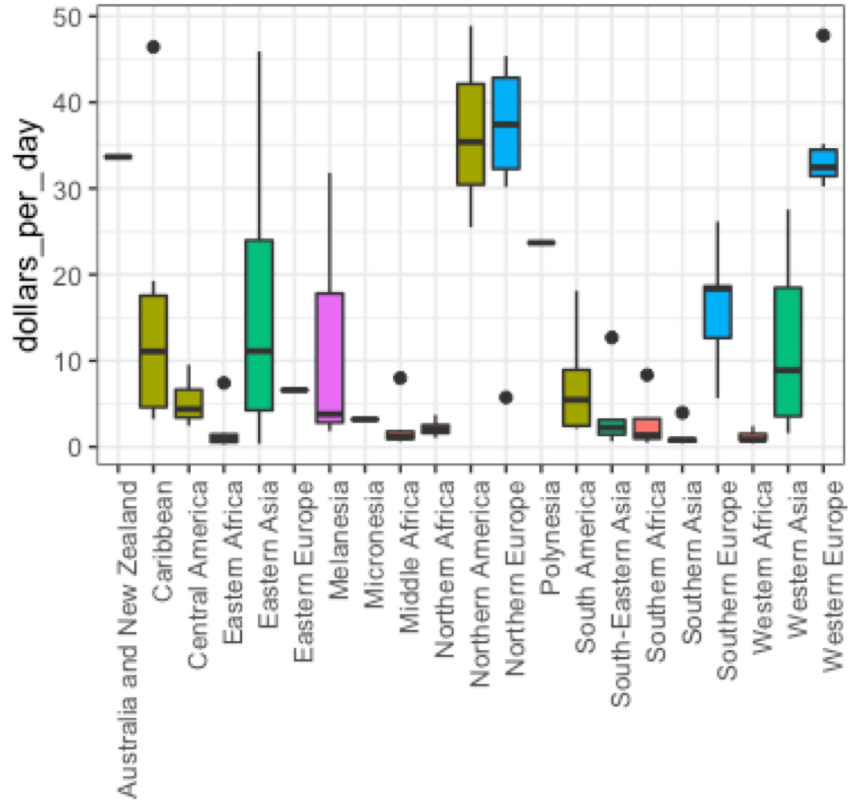
Order by a meaningful value

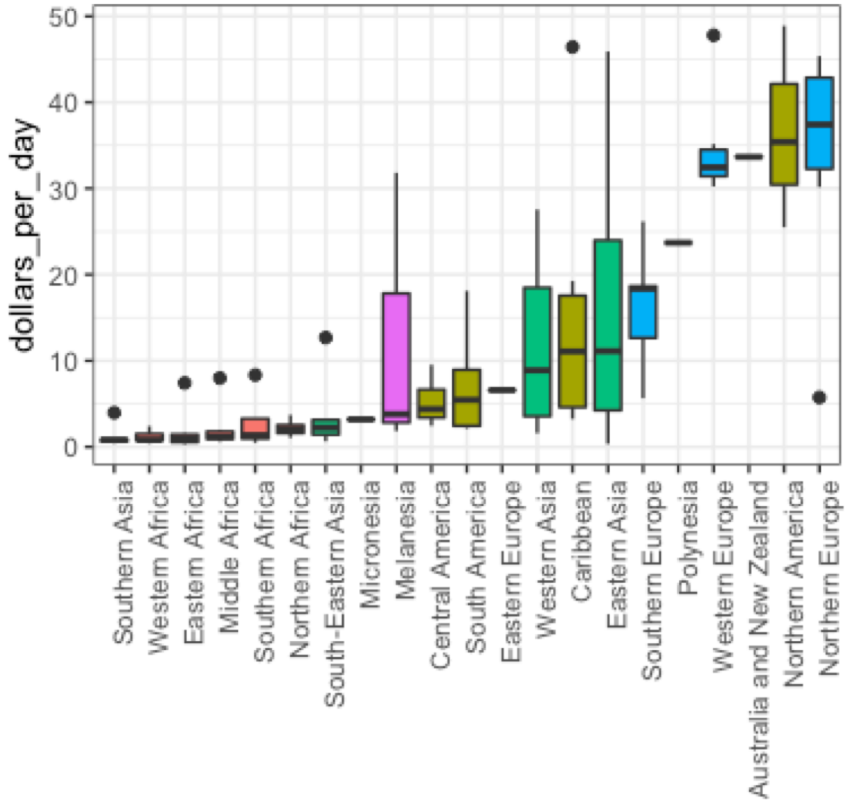
- When one of the axes is used to show categories, as is done in barplots, the default ggplot behavior is to order the categories alphabetically when they are defined by character strings.
- If they are defined by factors, they are ordered by the factor levels.
- We rarely want to use alphabetical order.
- Instead we should order by a meaningful quantity.

Order by a meaningful value



Another example: Average income by region in 1970 (colors = continent)

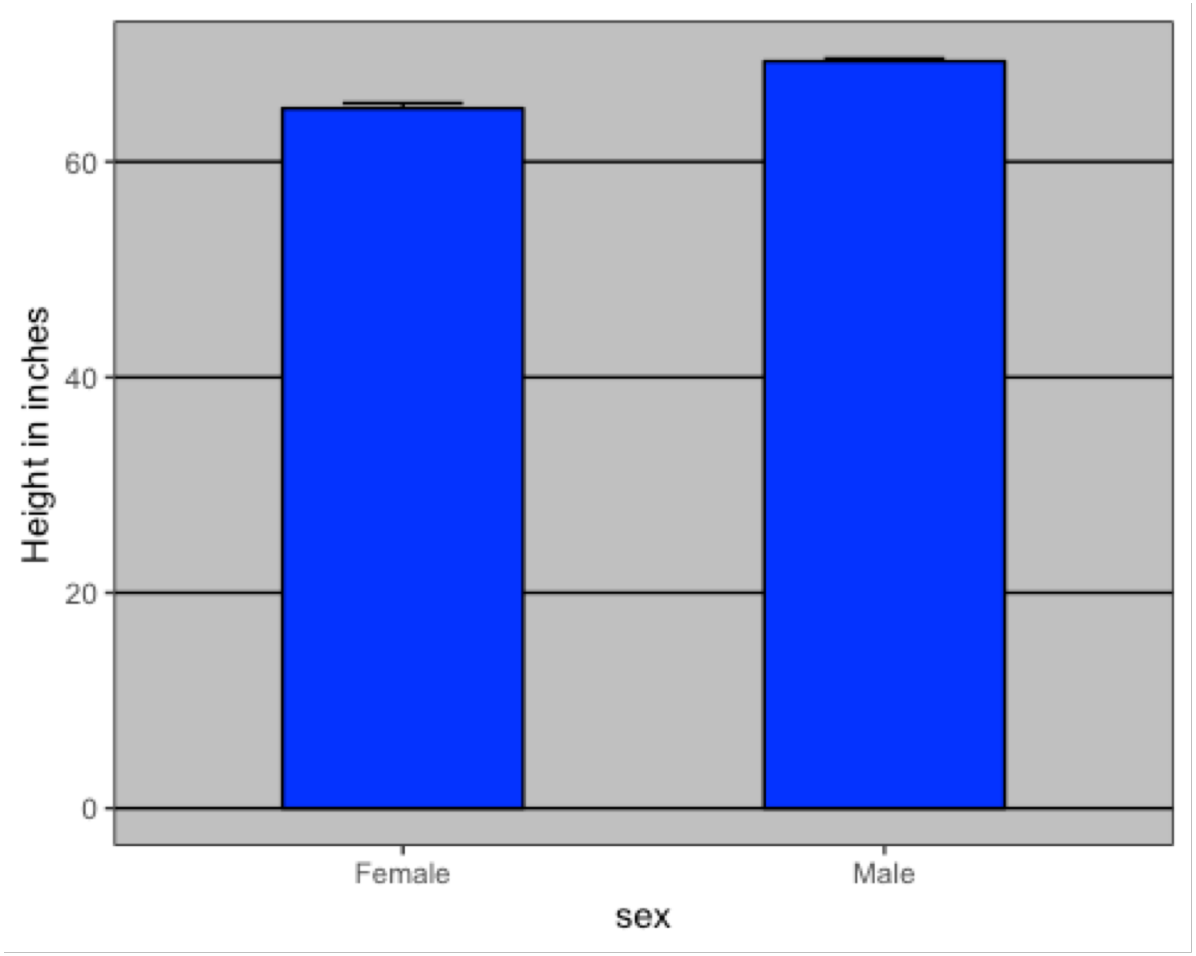




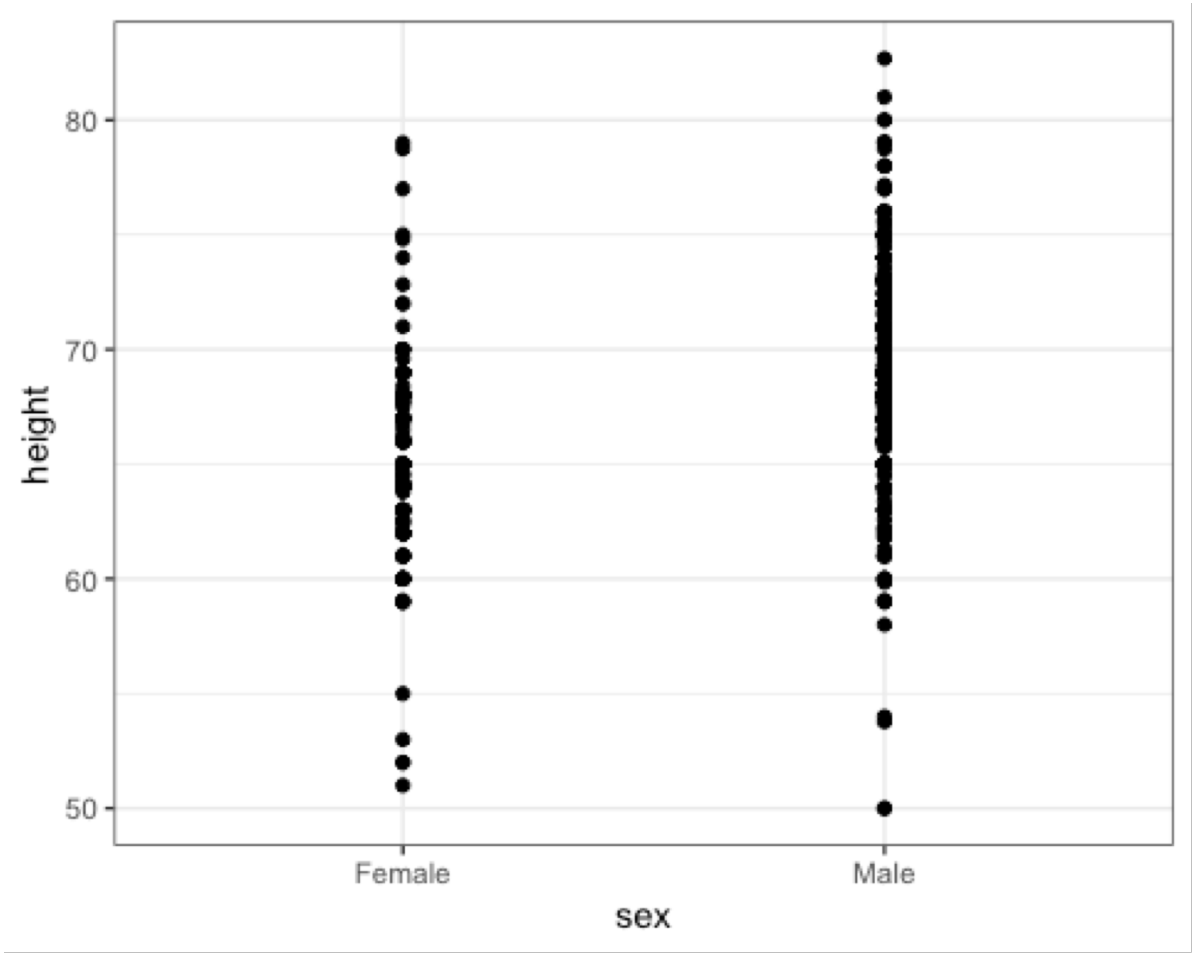
Show the data

- To motivate this principle let's asssume an extraterrestrial is interested is the difference in heights between males and females.
- A commonly seen plot used for comparisons between groups, popularized by software such as Microsoft Excel, shows the average and standard errors (standard errors are defined in a later chapter, but don't confuse them with the standard deviation of the data).

Show the data



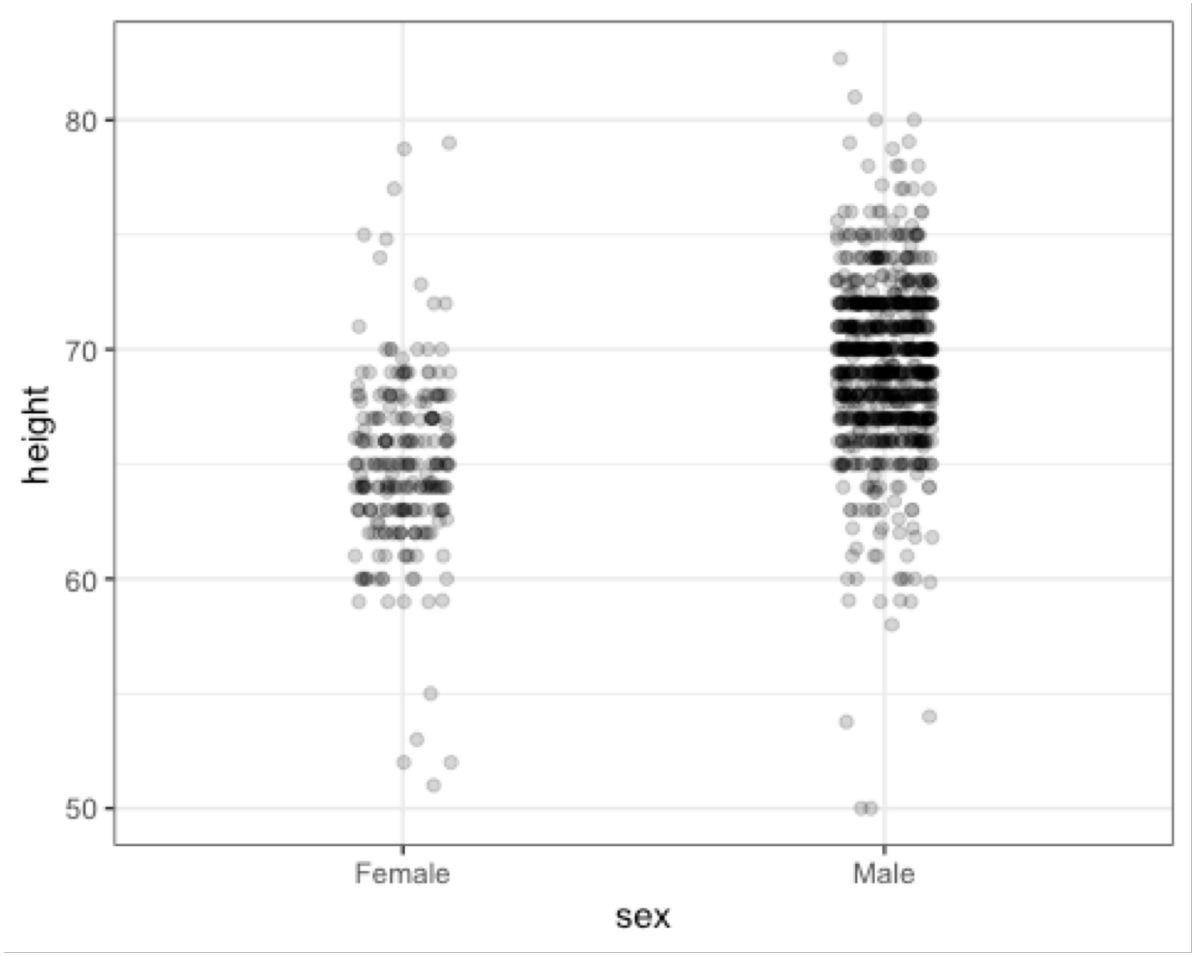
Show the data



Show the data

- What do we learn? We get an idea of the range of the data.
- However this plot has limitations as well since we can't really see all the 238 and 812 points plotted for females and males respectively, and many points are plotted above each other.

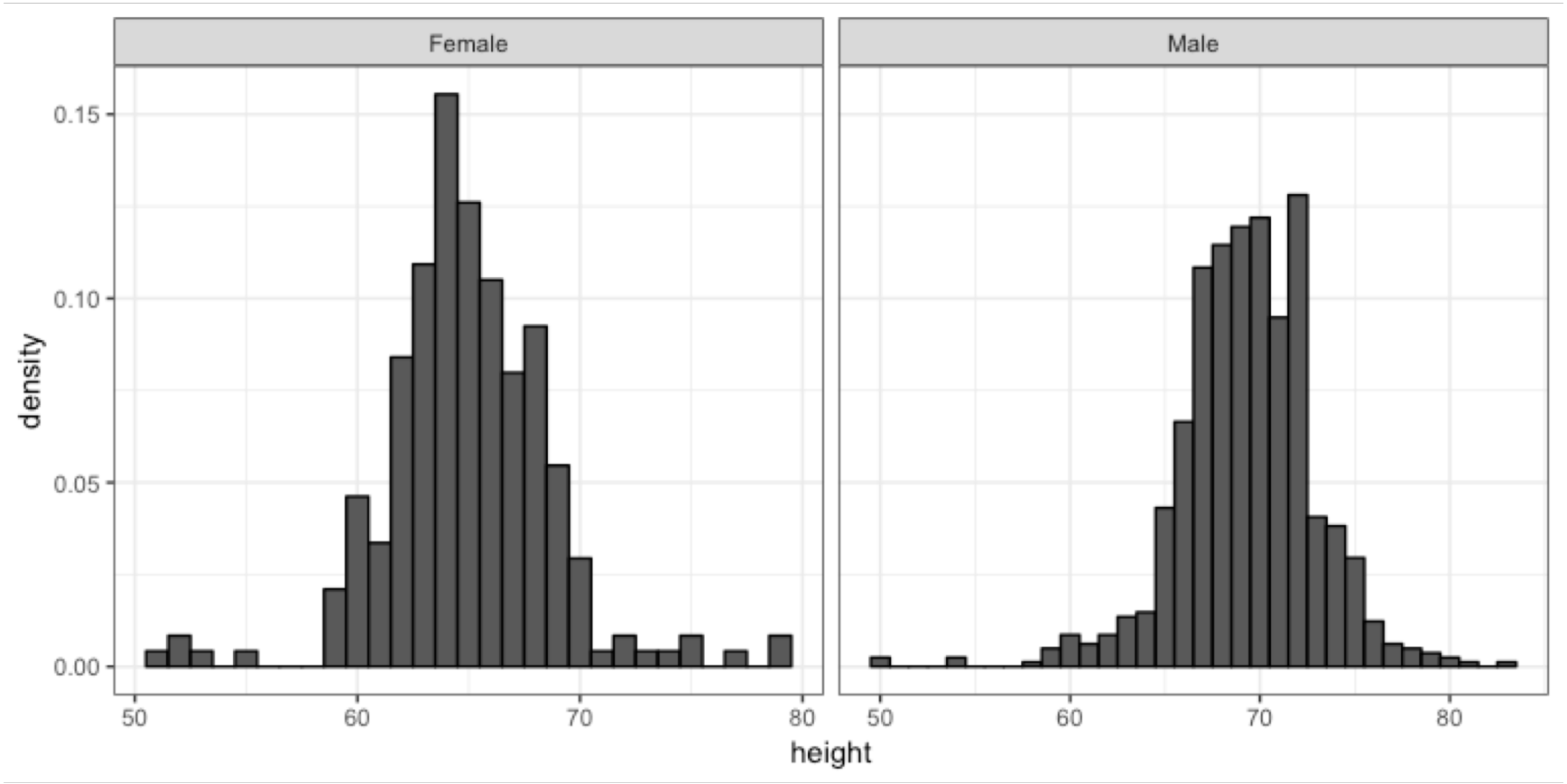
jitter and alpha blending



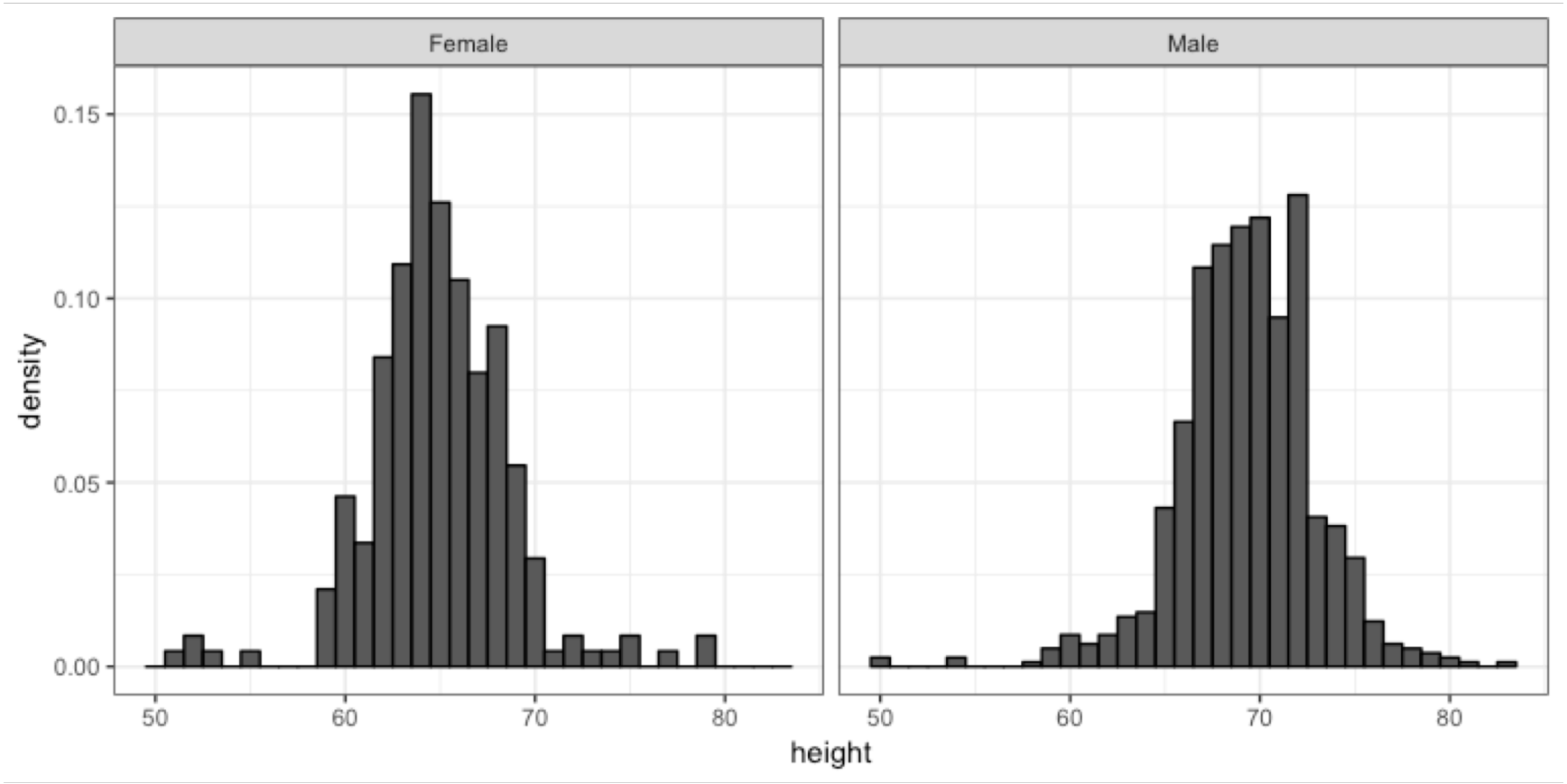
Show the data

- What do we learn? On average, males are taller than females.
- We also note dark horizontal demonstrating that many report values are rounded to the nearest integer.
- Since there are so many points it is more effective to show distributions, rather than show individual points.

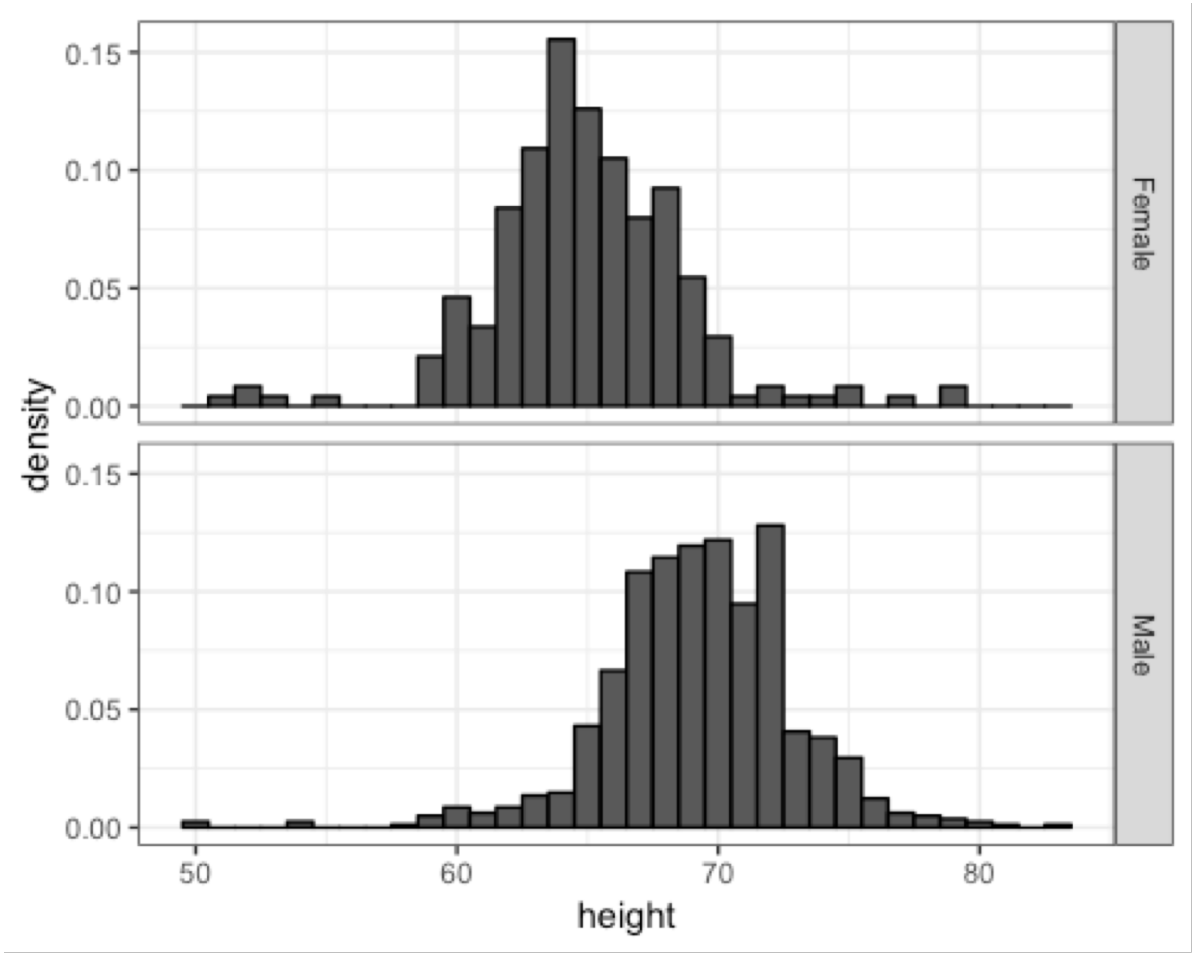
Compare distributions if too many points



Ease comparisons: Use common axes



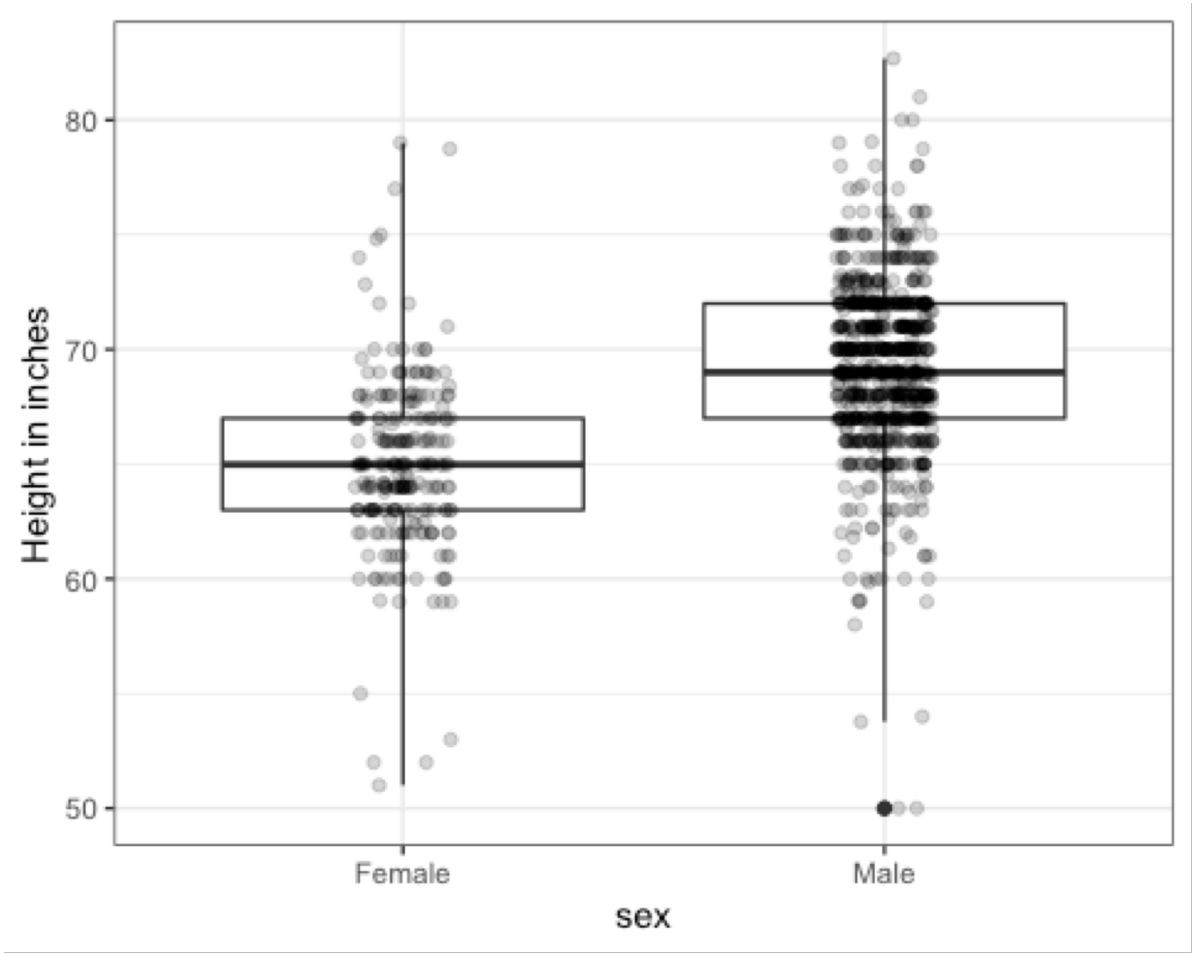
Ease comparisons: align vertically



Ease comparisons

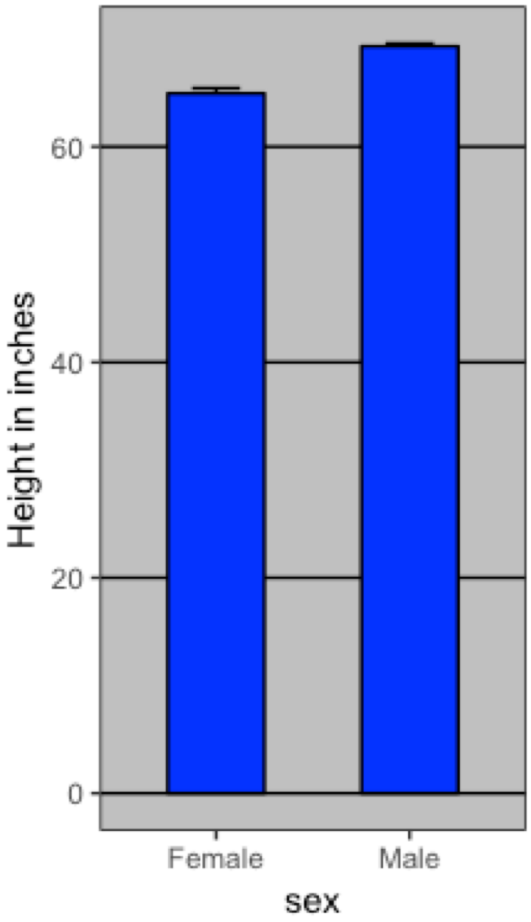
- This plot makes it much easier to notice that men are, on average, taller.
- If instead of histograms we want the more compact summary provided by boxplot, then we align the horizontally, since, by default, boxplots move up and down with changes in height.

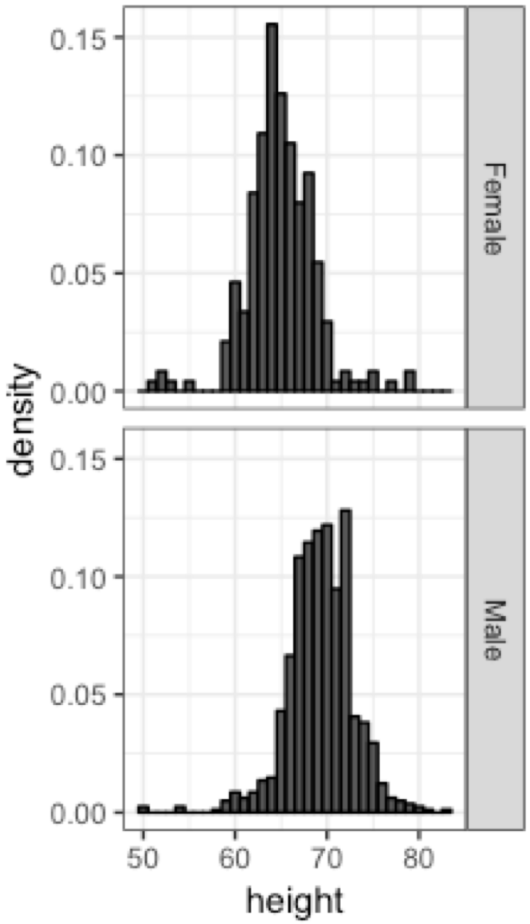
Ease comparisons: align horizontally

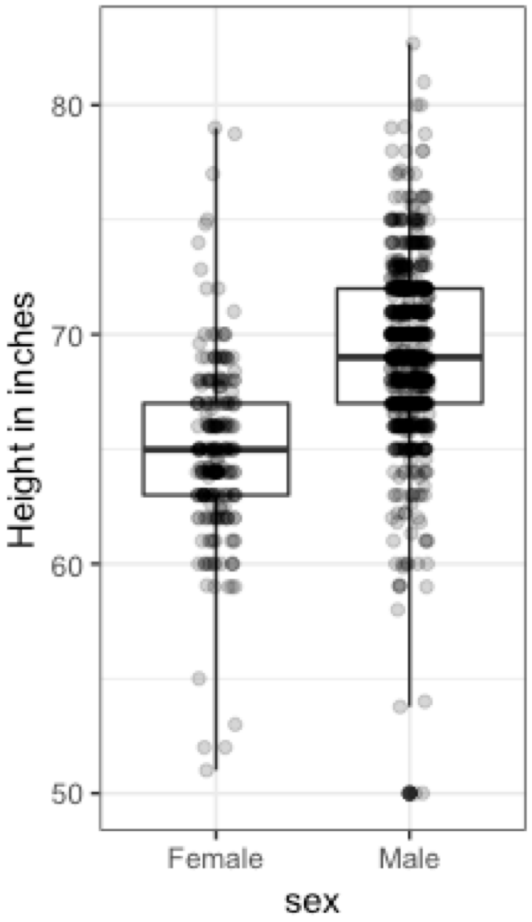


Comparison

```
grid.arrange(p1, p2, p3, ncol = 3)
```


Height in inches
60
40
20
0
Female
Male
sex

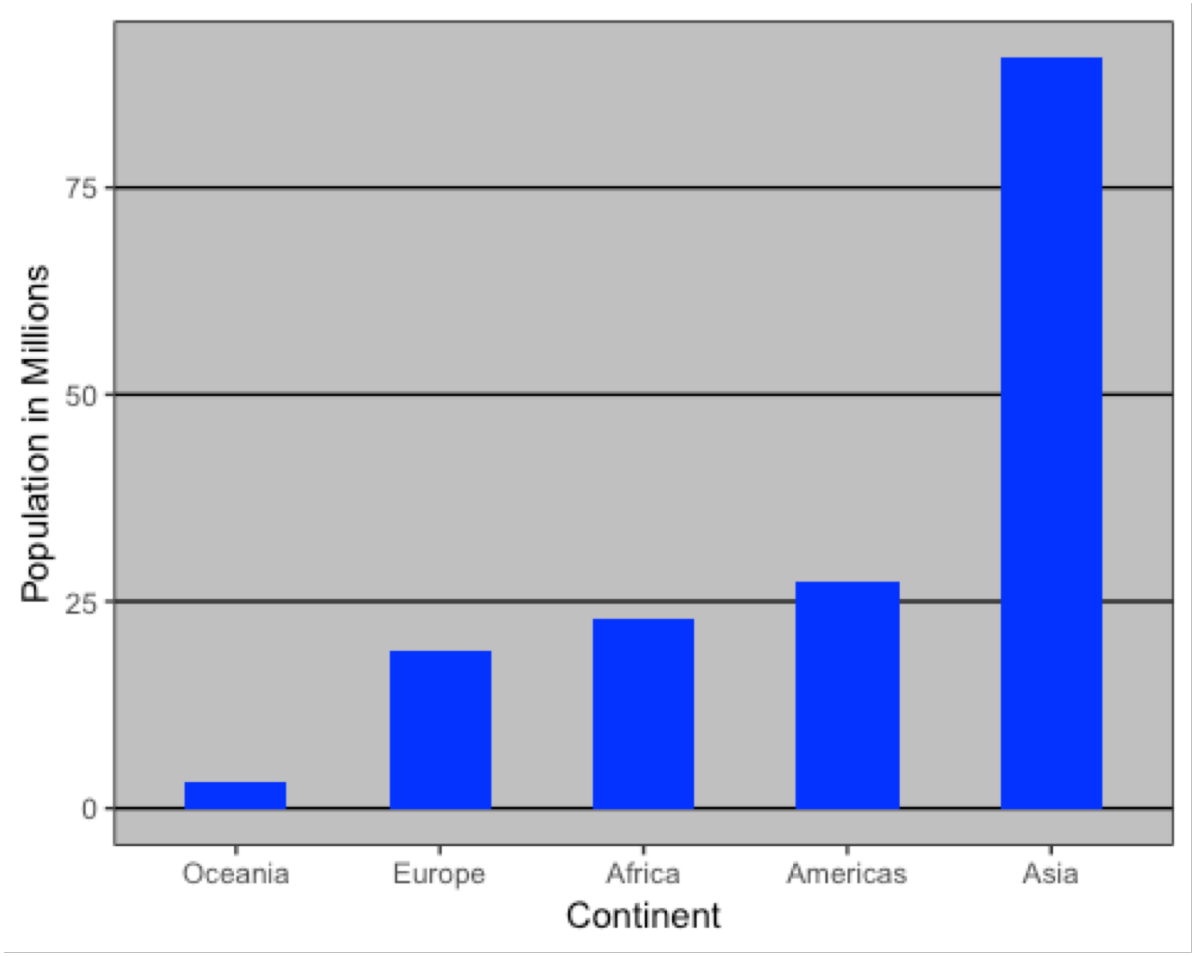

density
0.15
0.10
0.05
0.00
Female
Male
50
60
70
80
height


Height in inches
80
70
60
50
Female
Male
sex

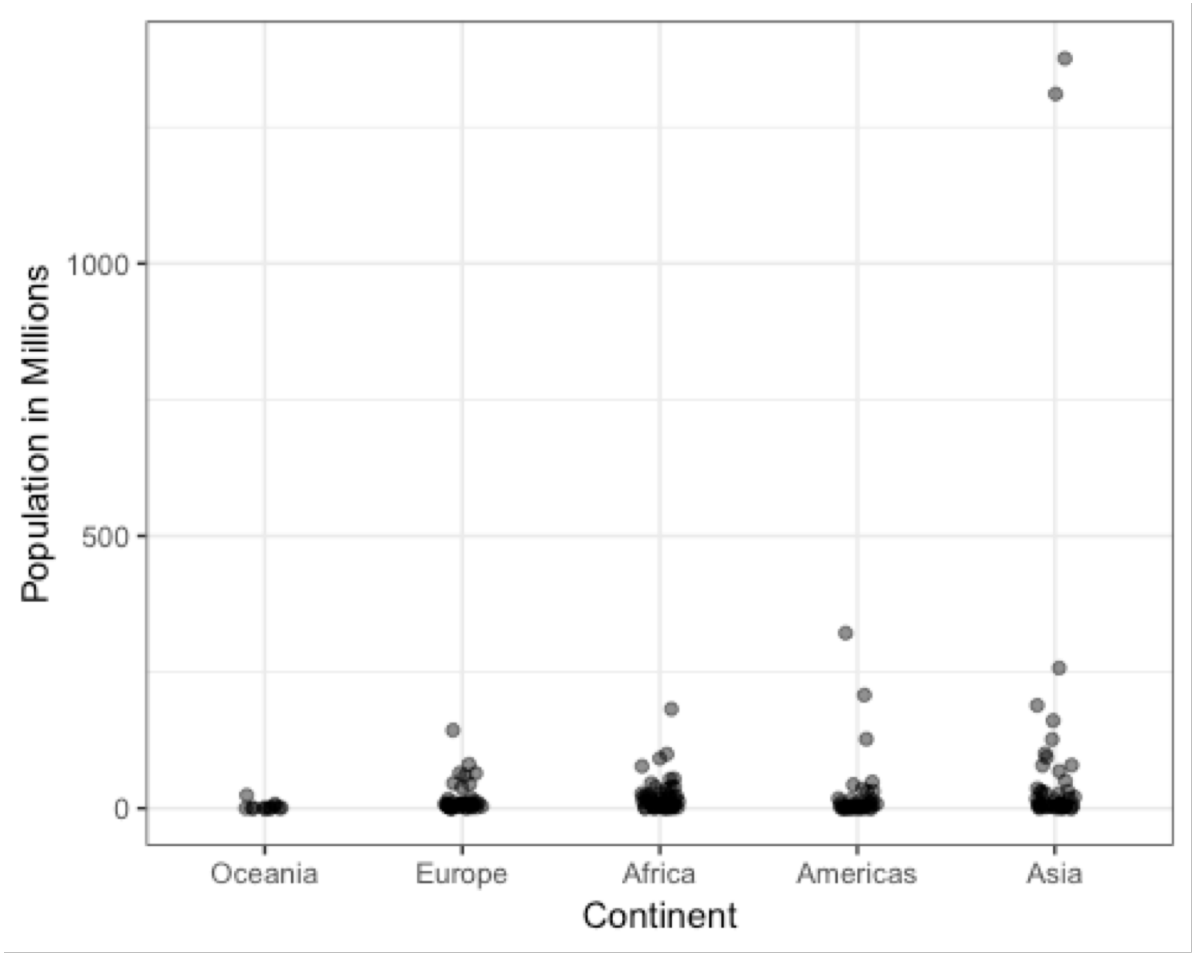
Consider transformations

– As an example consider this barplot showing the average population sizes for each continent in 2015:

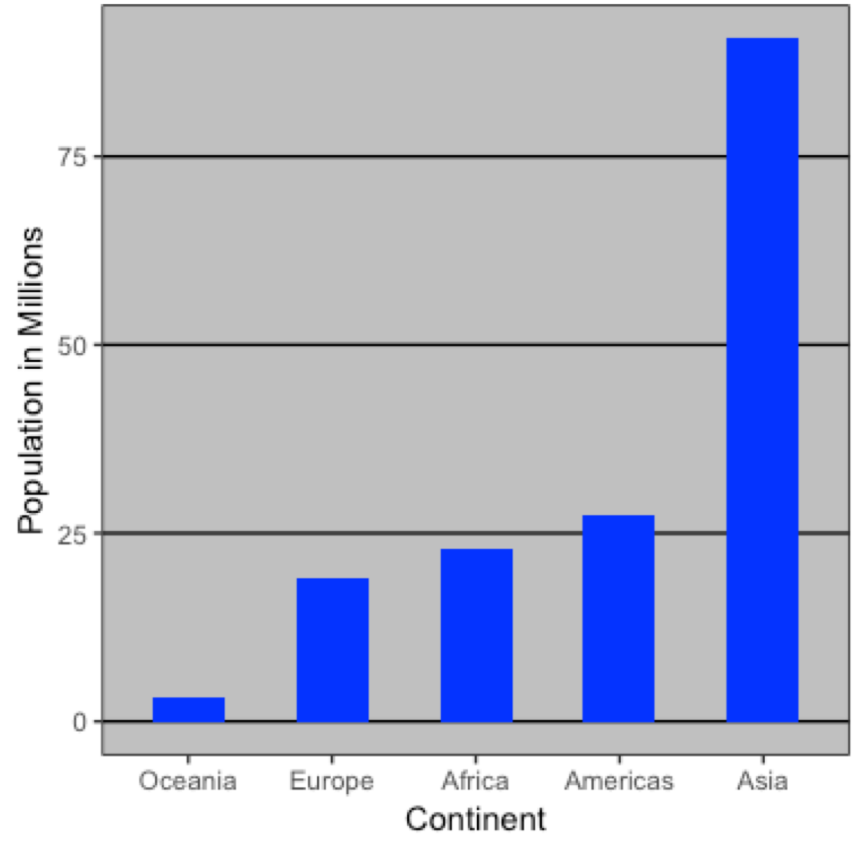
Consider transformations

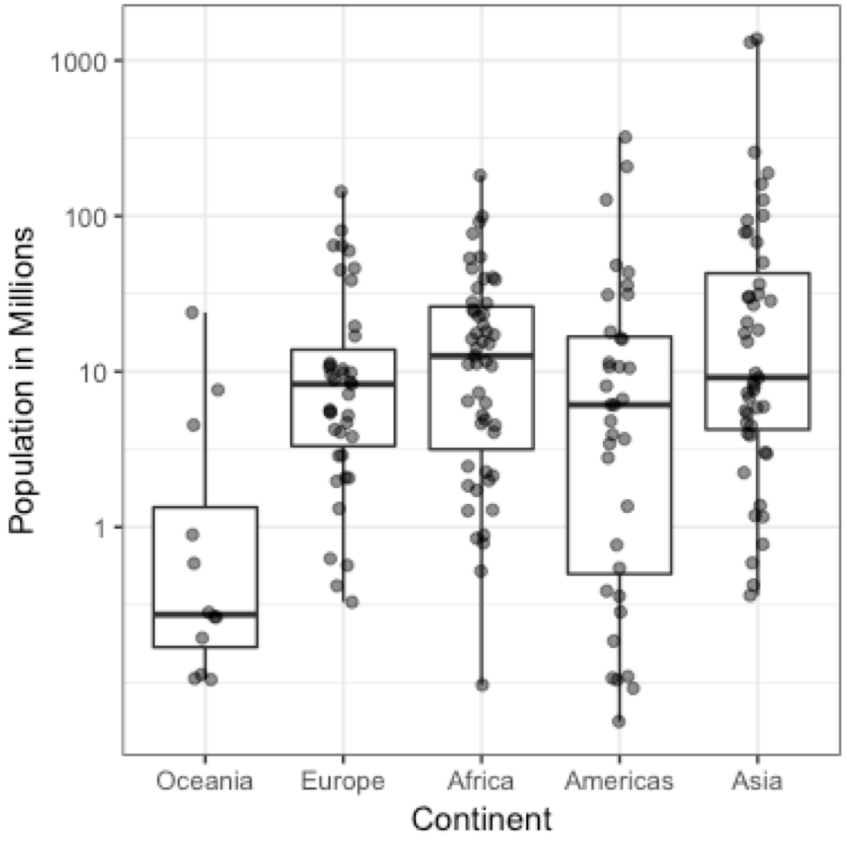


Show the data



Consider transformations

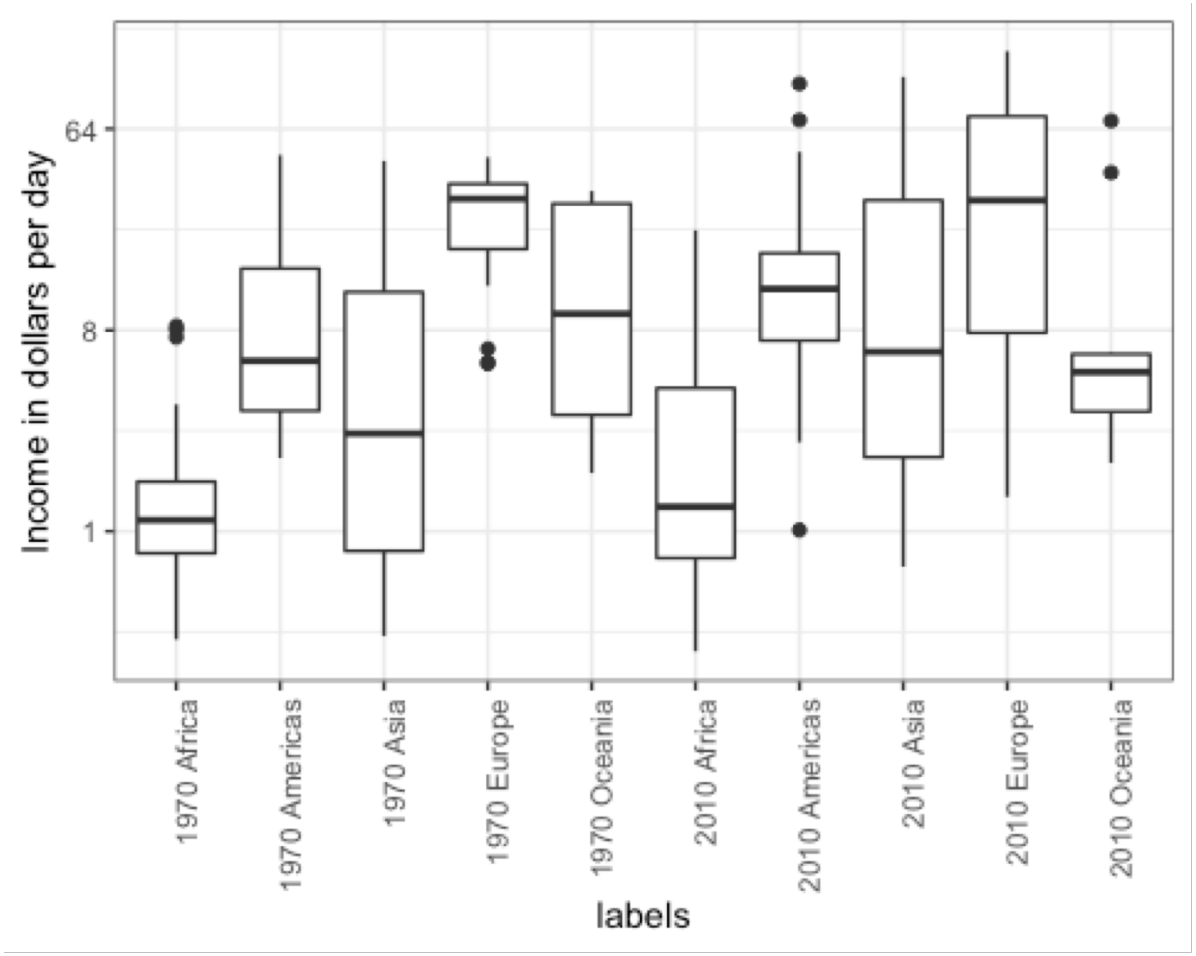




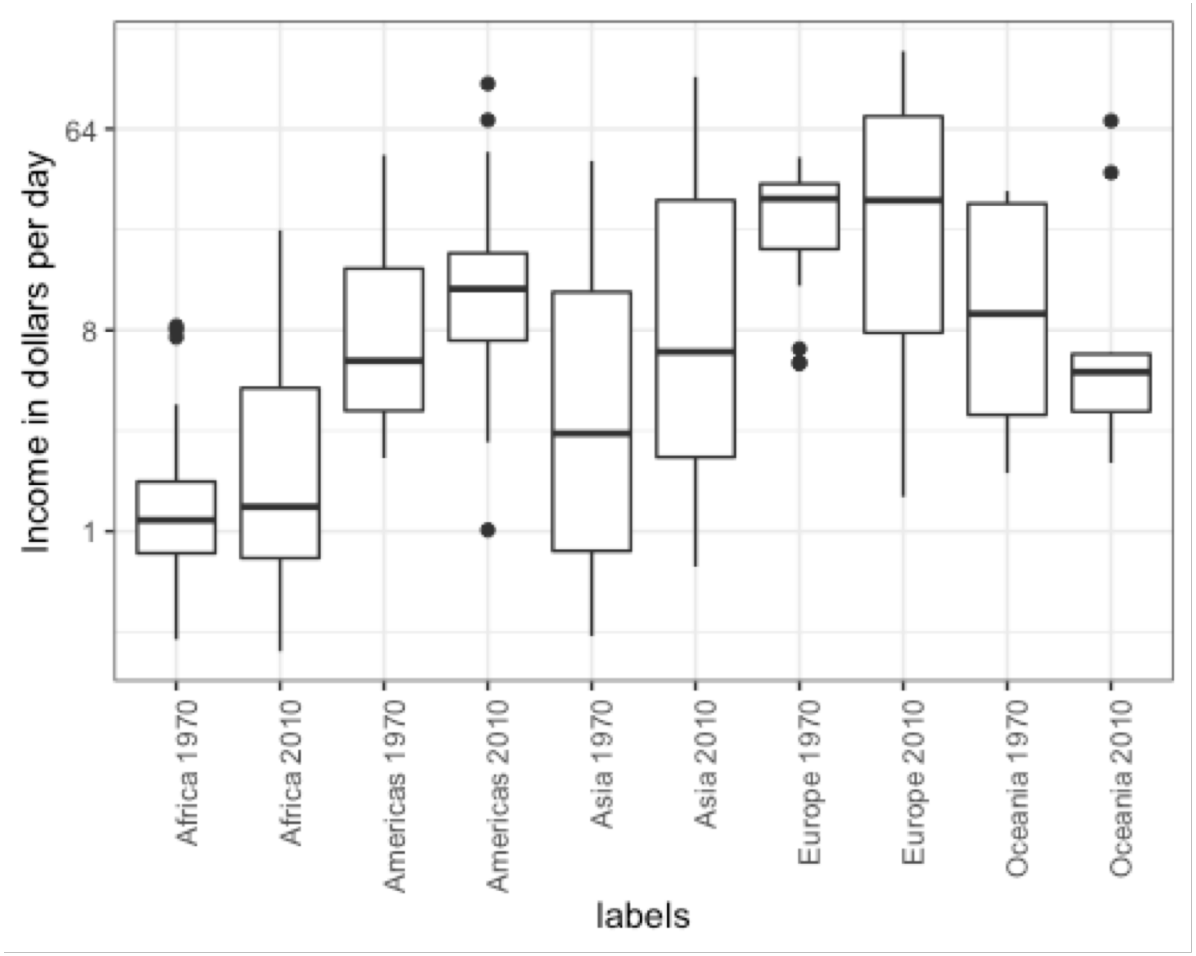
Ease comparisons

- Visual cues to be compared should be adjacent
- When comparing income data between 1970 and 2010 across region we made a figure similar to the one below.
- A difference is that here we look at continents instead of regions, but this is not relevant to the point we are making.

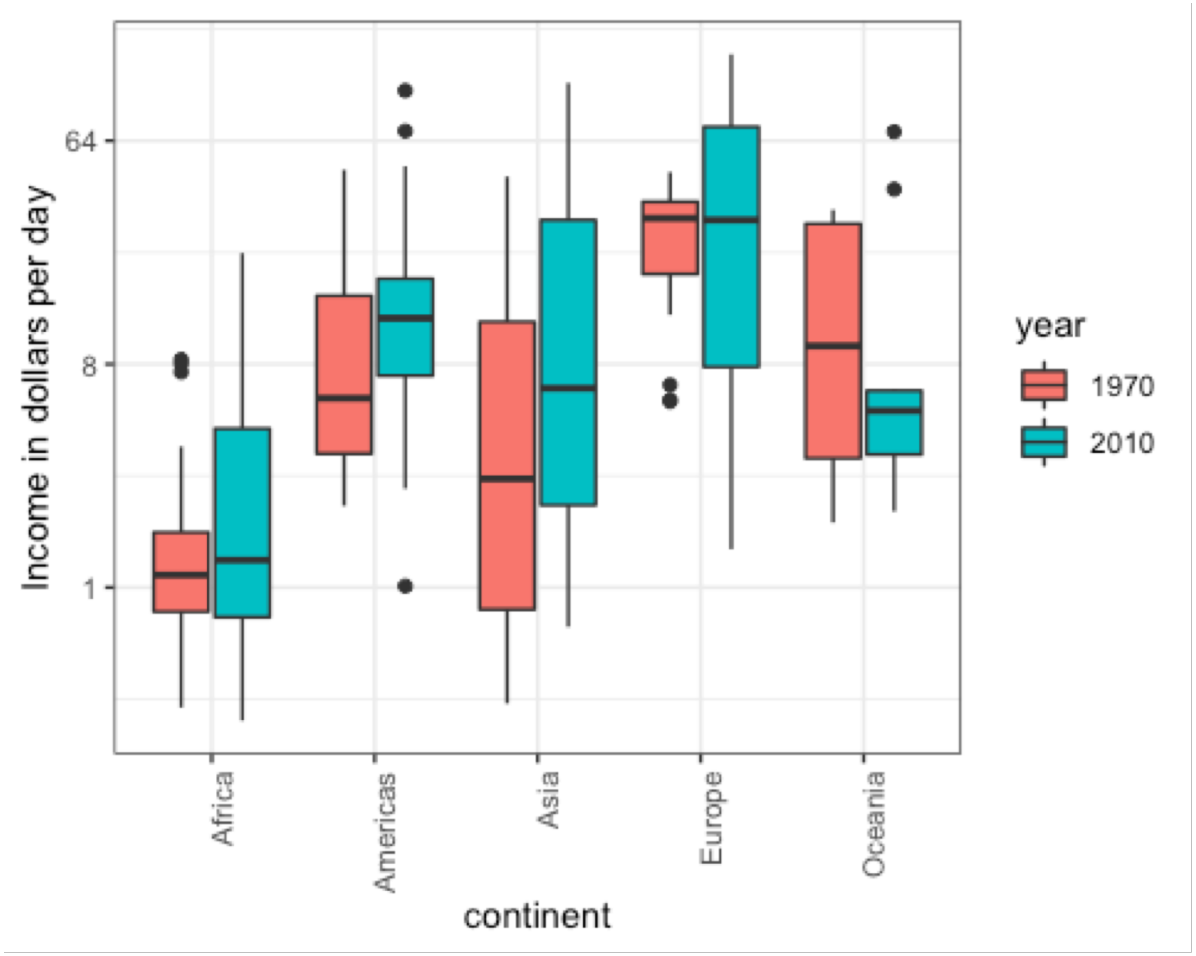
Ease comparisons



Comparisons should be adjacent



Use color to highlight comparison



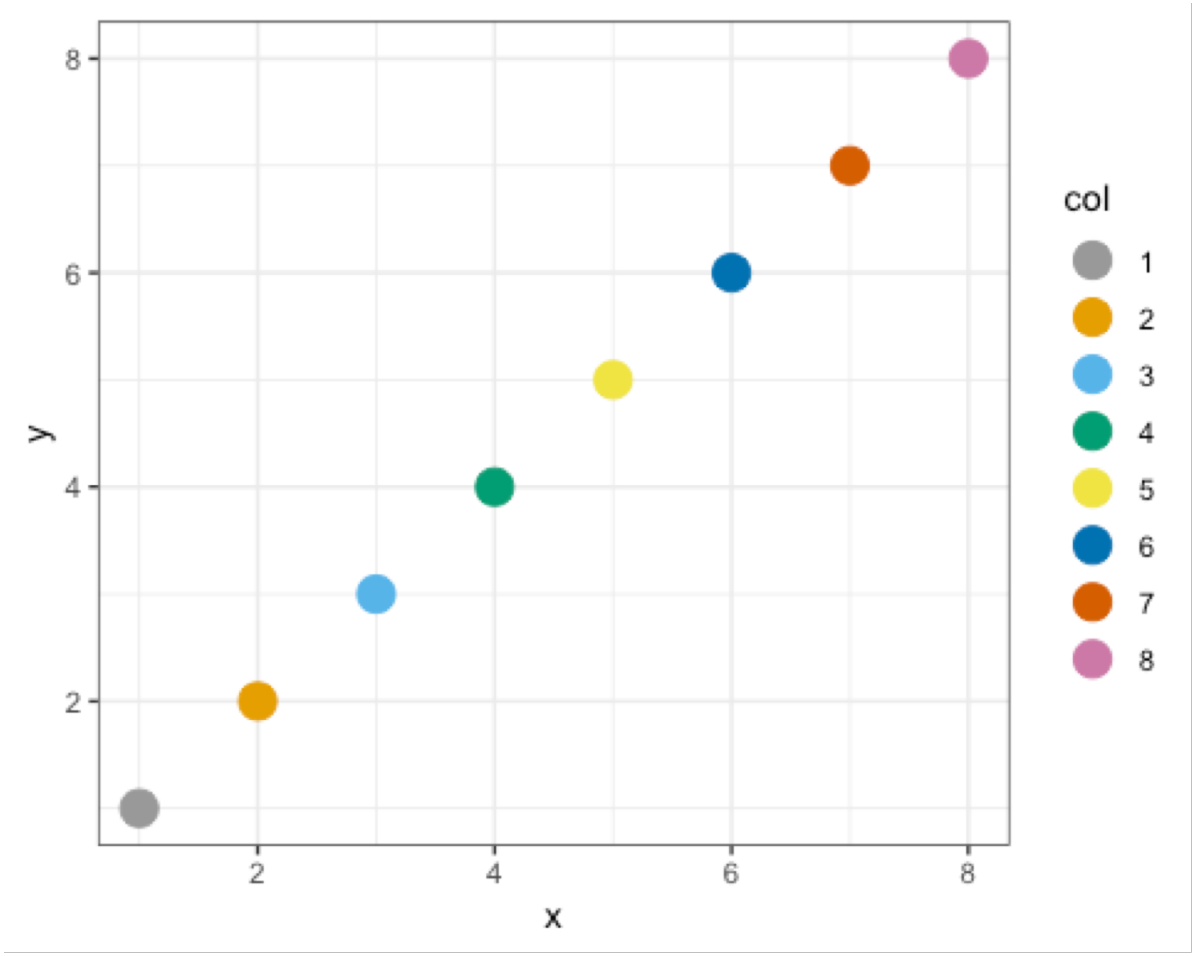
Think of the color blind

- About 10% of the population is color blind.
- Unfortunately, the default colors used in ggplot are not optimal for this group.
- However, ggplot does it make it easy to change the color palette used in the plots.

Think of the color blind

– Here is an example of how we can use color blind friendly pallet described here:

Think of the color blind



Think of the color blind

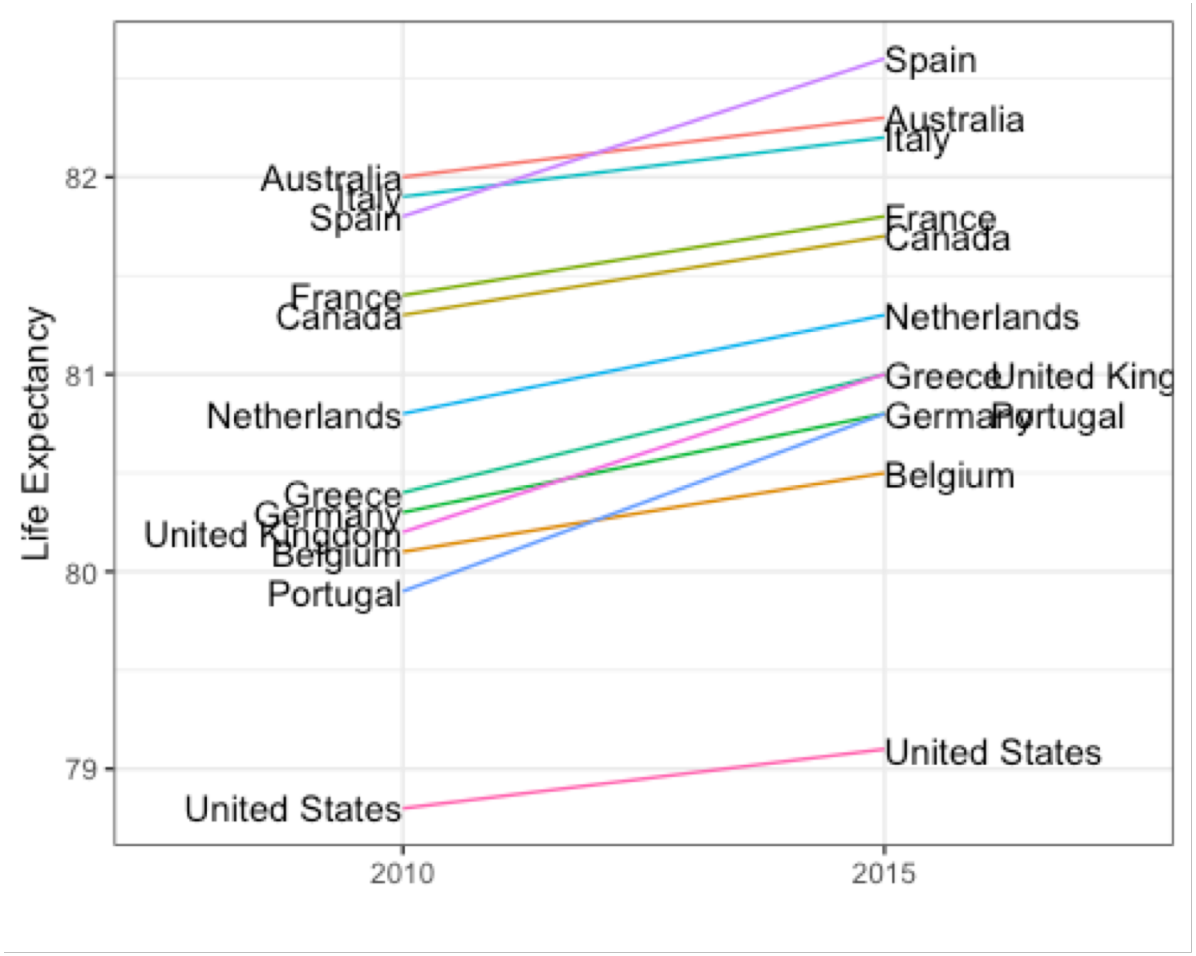
- There are several resources that help you select colors, for example this one.

Use scatter-plots to examine the relationship between two variables

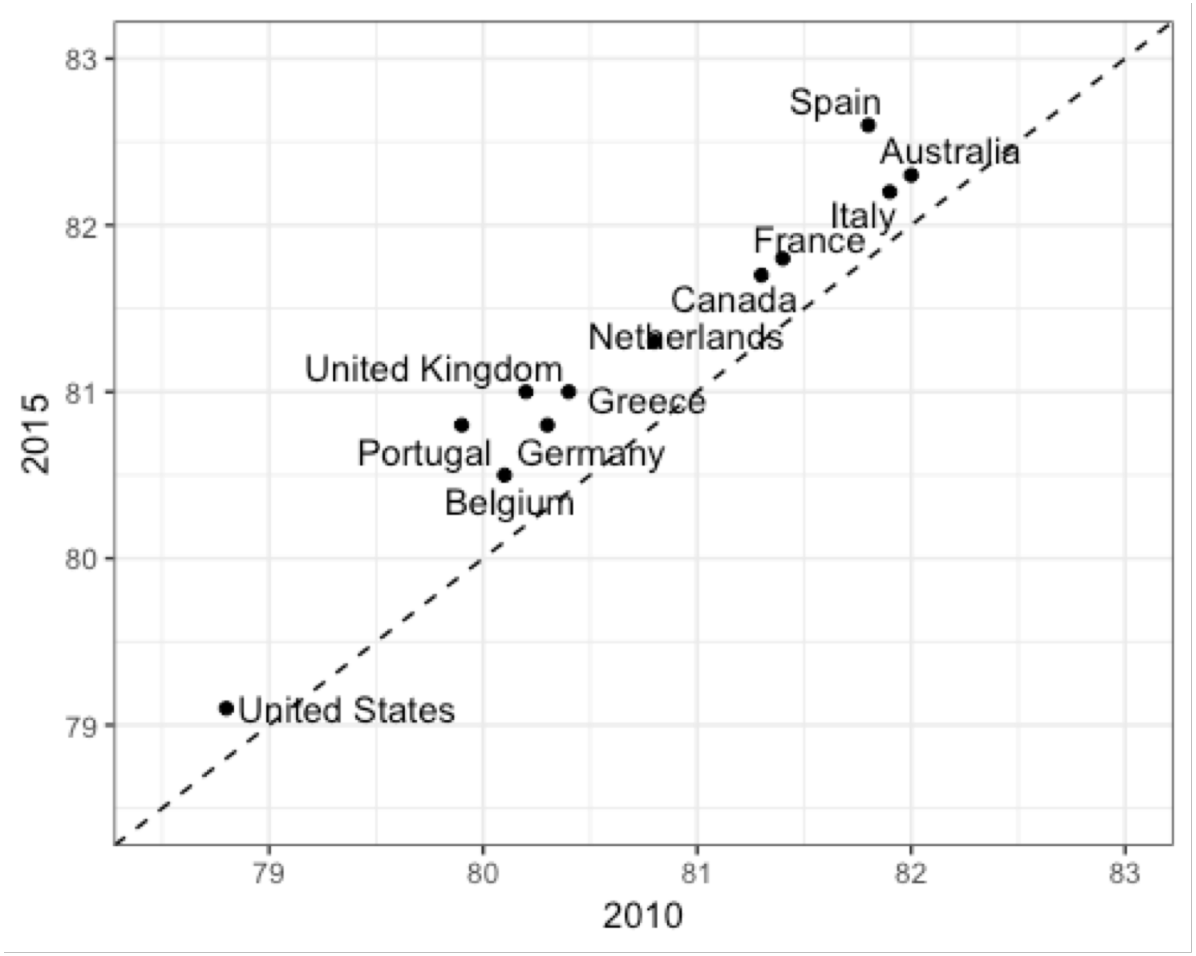
Exceptions are

- Slope charts
- Bland-Altman plots

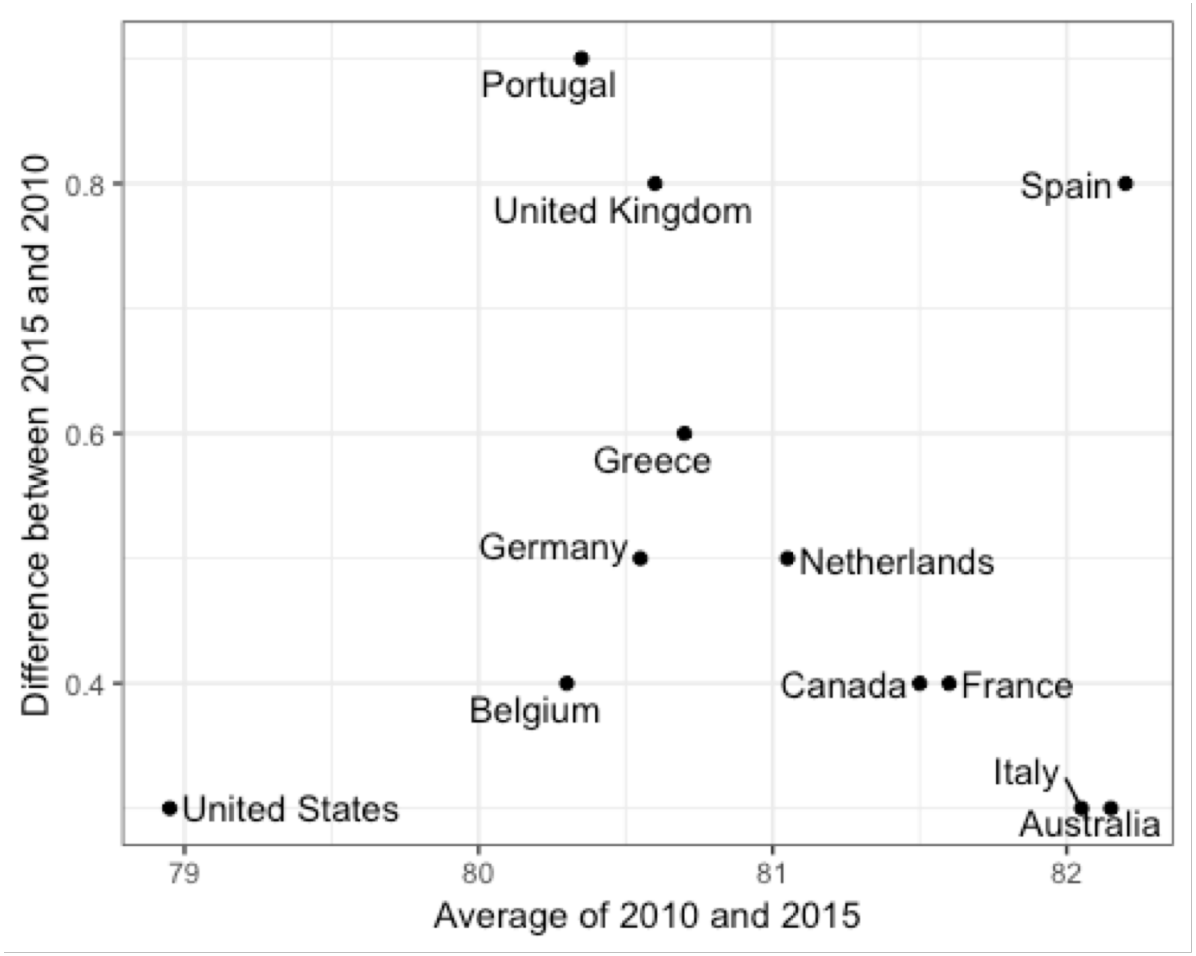
Slope charts



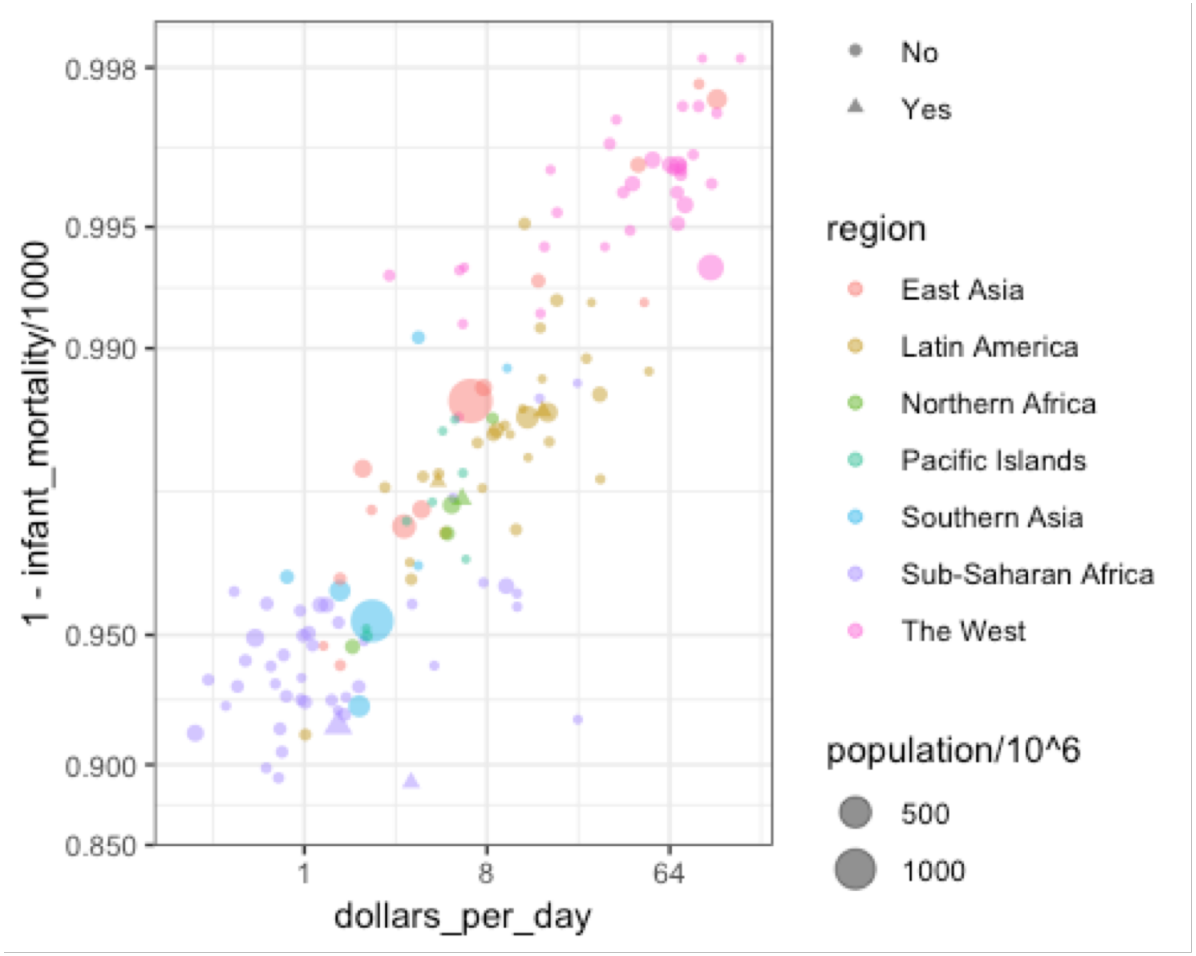
Scatter plot (with common axes)



Bland-Altman plot



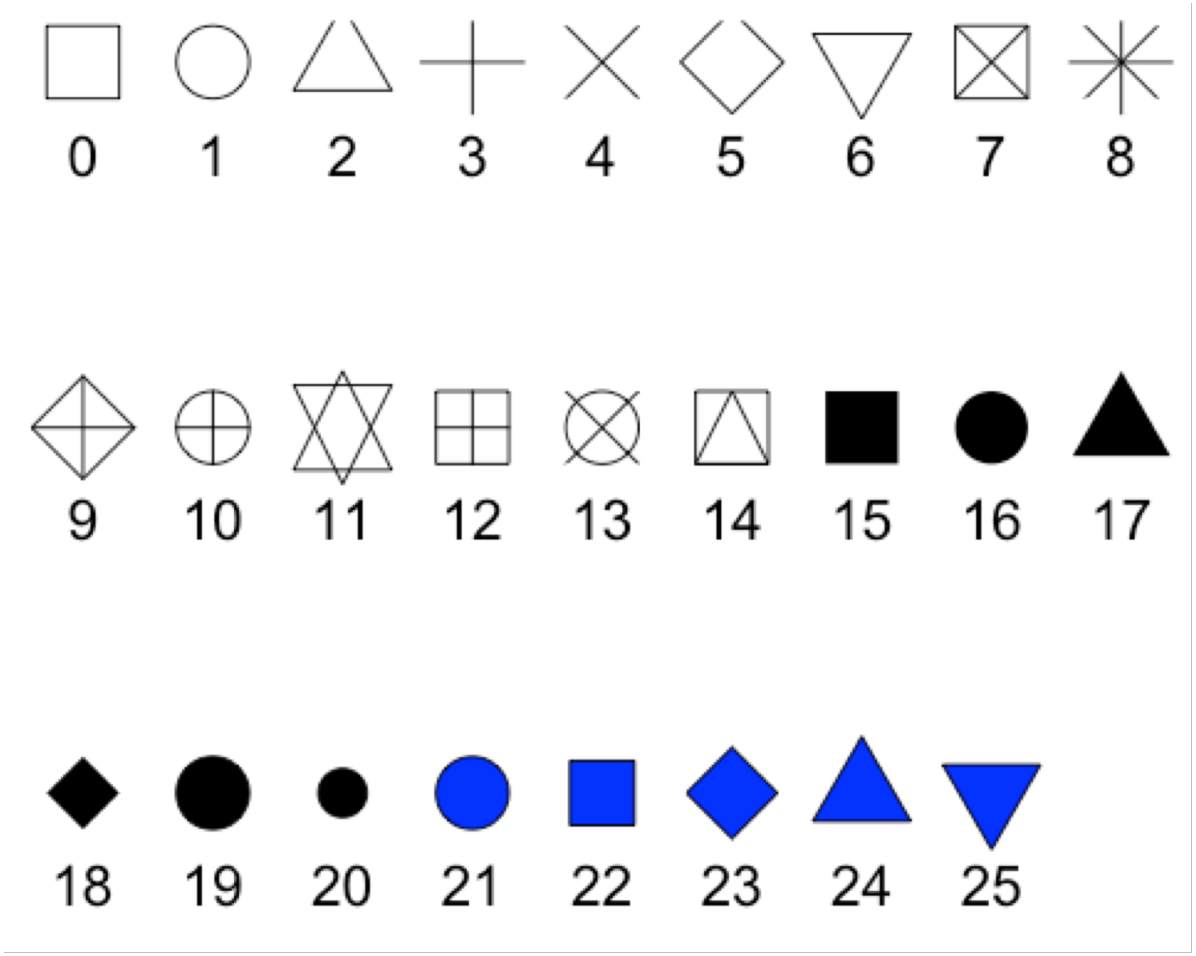
Encoding a third variable



Encoding a third variable

- Note that we encode categorical variables with color hue and shape.
- These shape can be controlled with `shape` argument.
- Below are the shapes available for use in R.
- Note that for the last five, the color goes inside.

Encoding a third variable



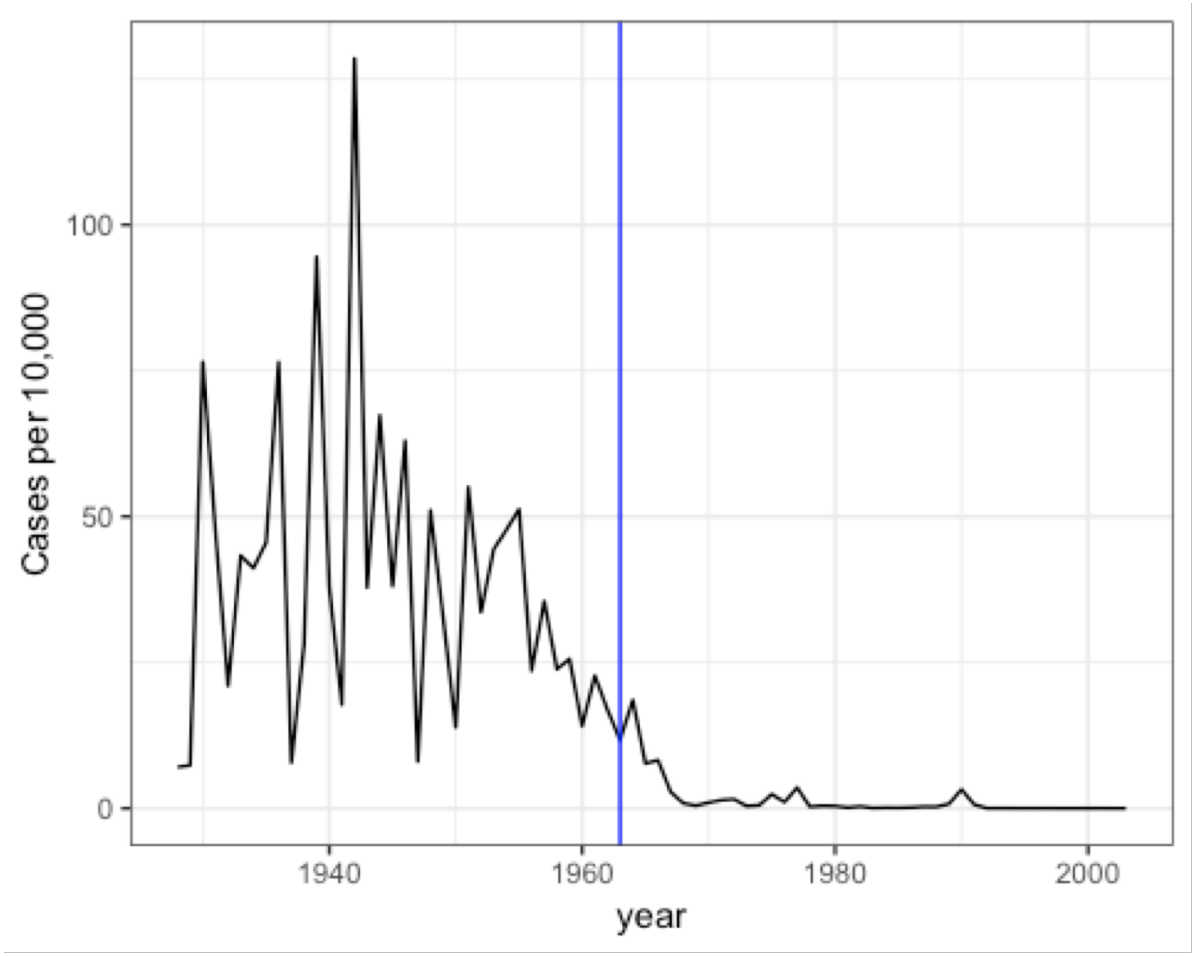
Encoding a third variable

- For continuous variables we can use color, intensity or size.
- We now show an example of how we do this with a case study.

Example

- The data used for these plots were collected, organized and distributed by the Tycho Project.
- They include weekly reported counts data for seven diseases from 1928 to 2011, from all fifty states.

One state is easy



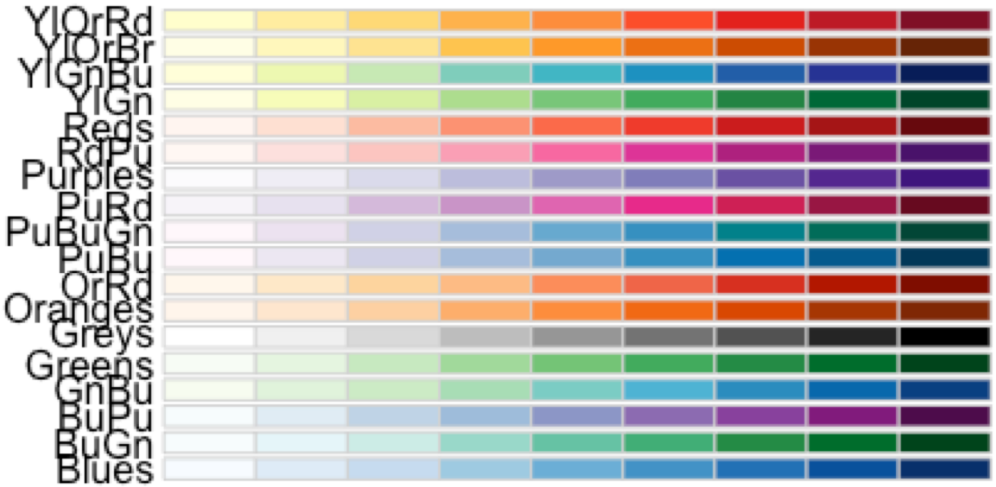
We need other cues for third dimenion

- When choosing colors to quantify a numeric variable we chose between two options sequential and diverging.
- Sequential colors are suited for data that goes from high to low.
- High values are clearly distinguished from low values.
- Here are some examples offered by the package `RColorBrewer`

Paletts

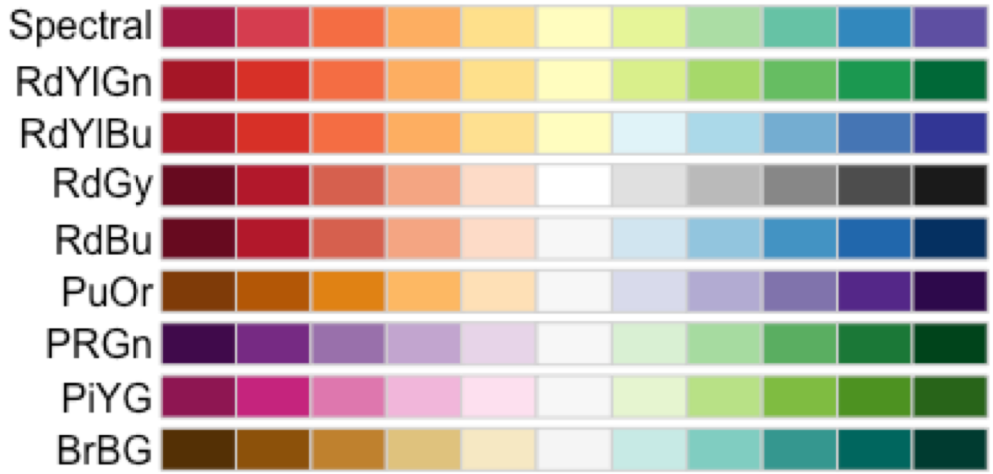
- Diverging colors are used to represent values that diverge from a center.
- We put equal emphasis on both ends of the data range: higher than the center and lower than the center.
- An example of when we would use a divergent pattern would be if we were to show height in standard deviations away from the average.
- Here are some examples of divergent patterns:

Sequential Paletts

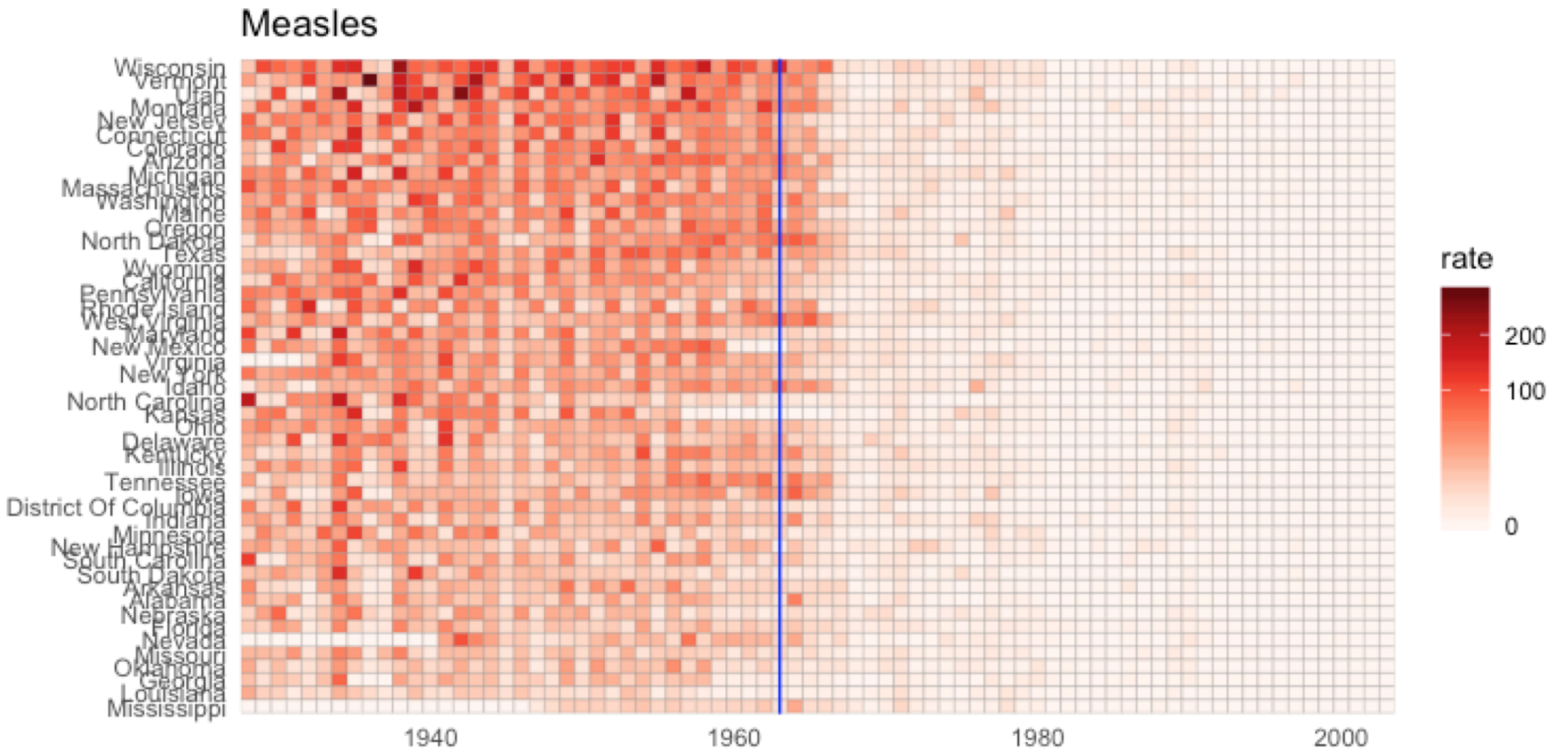


Divergent Paletts

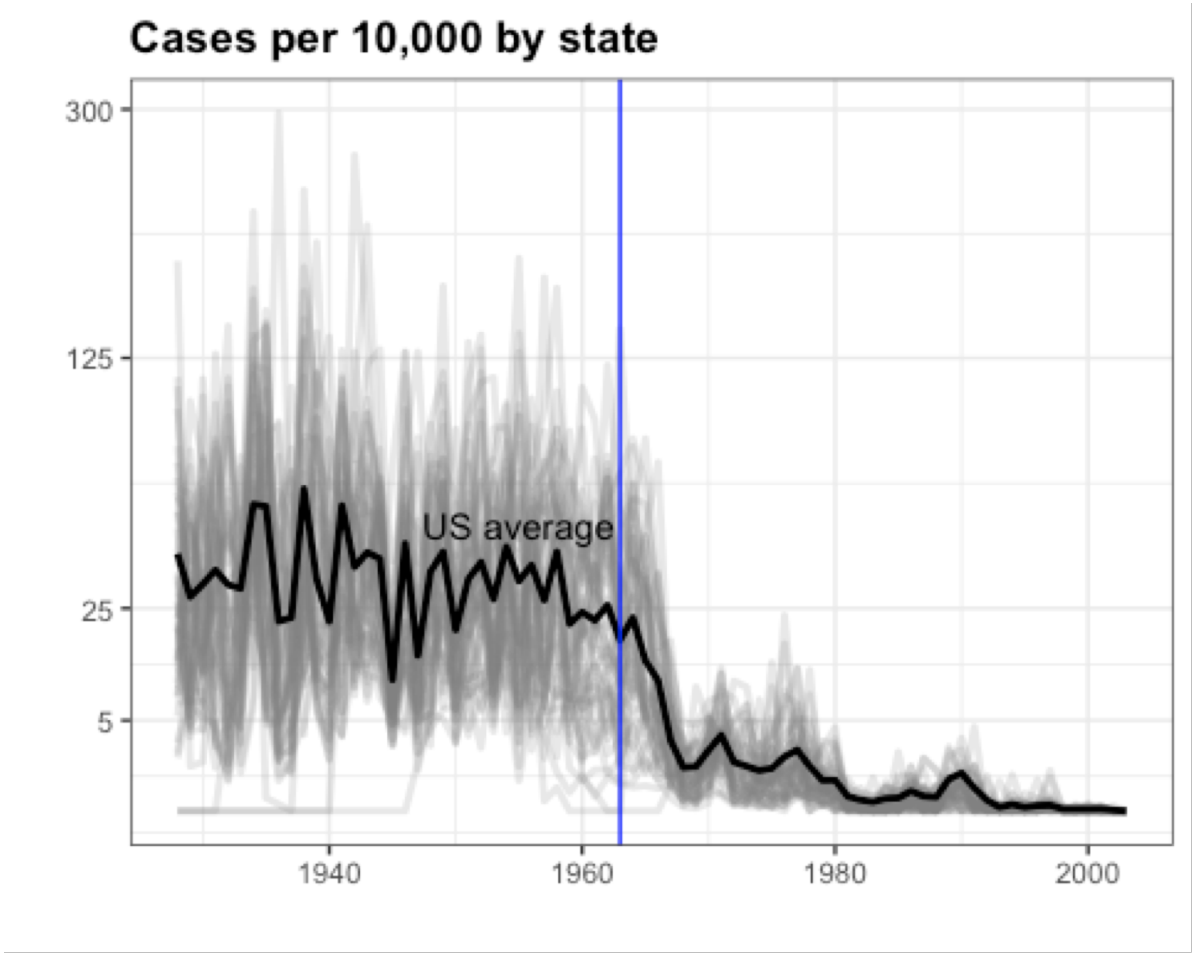
```
library(RColorBrewer)
display.brewer.all(type="div")
```


Spectral
RdYlGn
RdYlBu
RdGy
RdBu
PuOr
PRGn
PiYG
BrBG

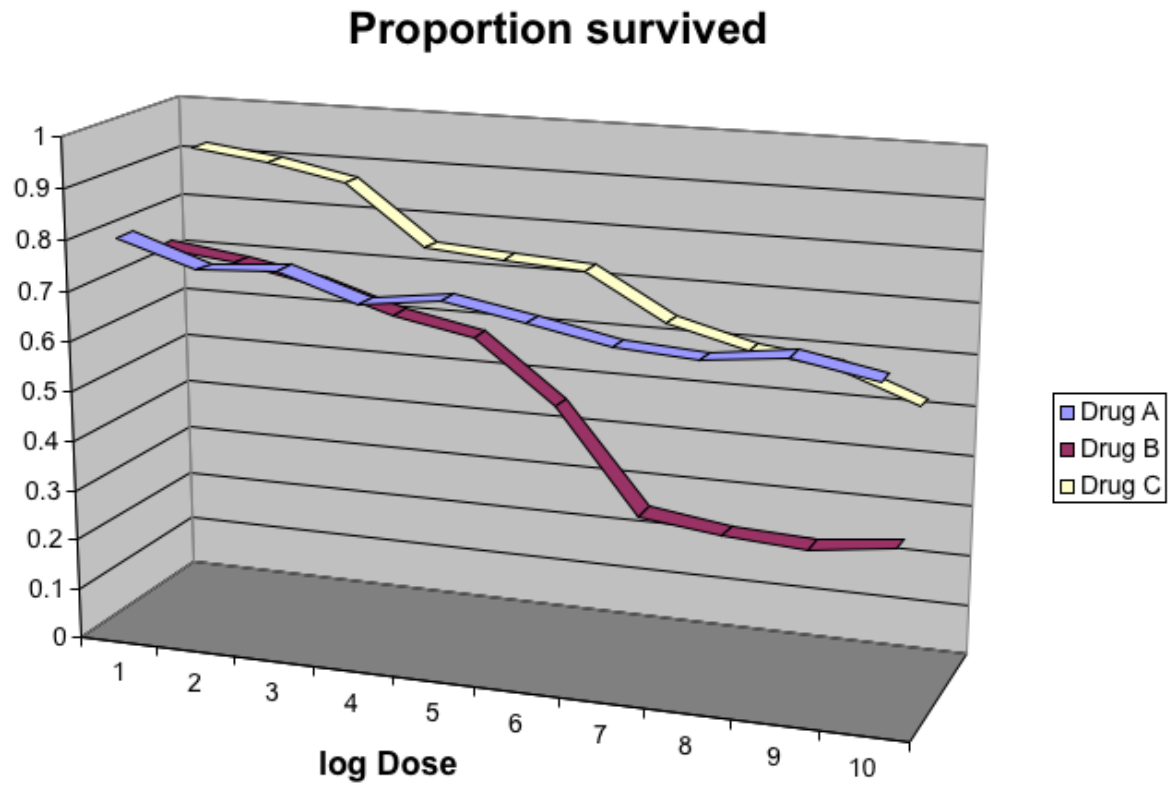
Example


Measles
Wisconsin
Vermont
Utah
Montana
New Jersey
Connecticut
Colorado
Arizona
Michigan
Massachusetts
Washington
Maine
Oregon
North Dakota
Texas
Wyoming
California
Pennsylvania
Rhode Island
West Virginia
Maryland
New Mexico
Virginia
New York
Idaho
North Carolina
Kansas
Ohio
Delaware
Kentucky
Illinois
Tennessee
Iowa
District Of Columbia
Indiana
Minnesota
New Hampshire
South Carolina
South Dakota
Arkansas
Alabama
Nebraska
Florida
Nevada
Missouri
Oklahoma
Georgia
Louisiana
Mississippi
1940
1960
1980
2000
rate
200
100
0

Alternative: eliminate one variable



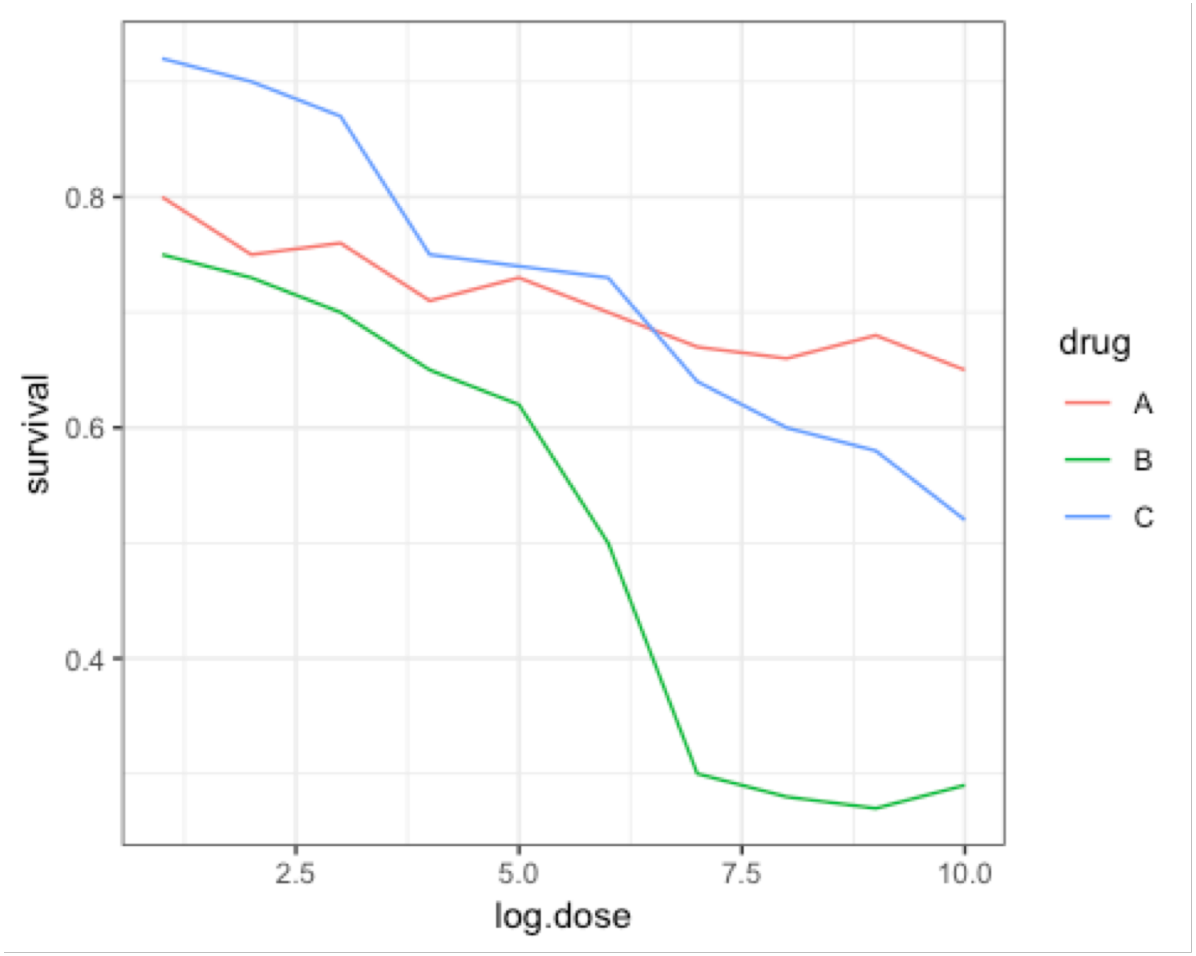
Avoid pseudo three dimensional plots



Pseudo 3-D.

(Source: Karl Broman)

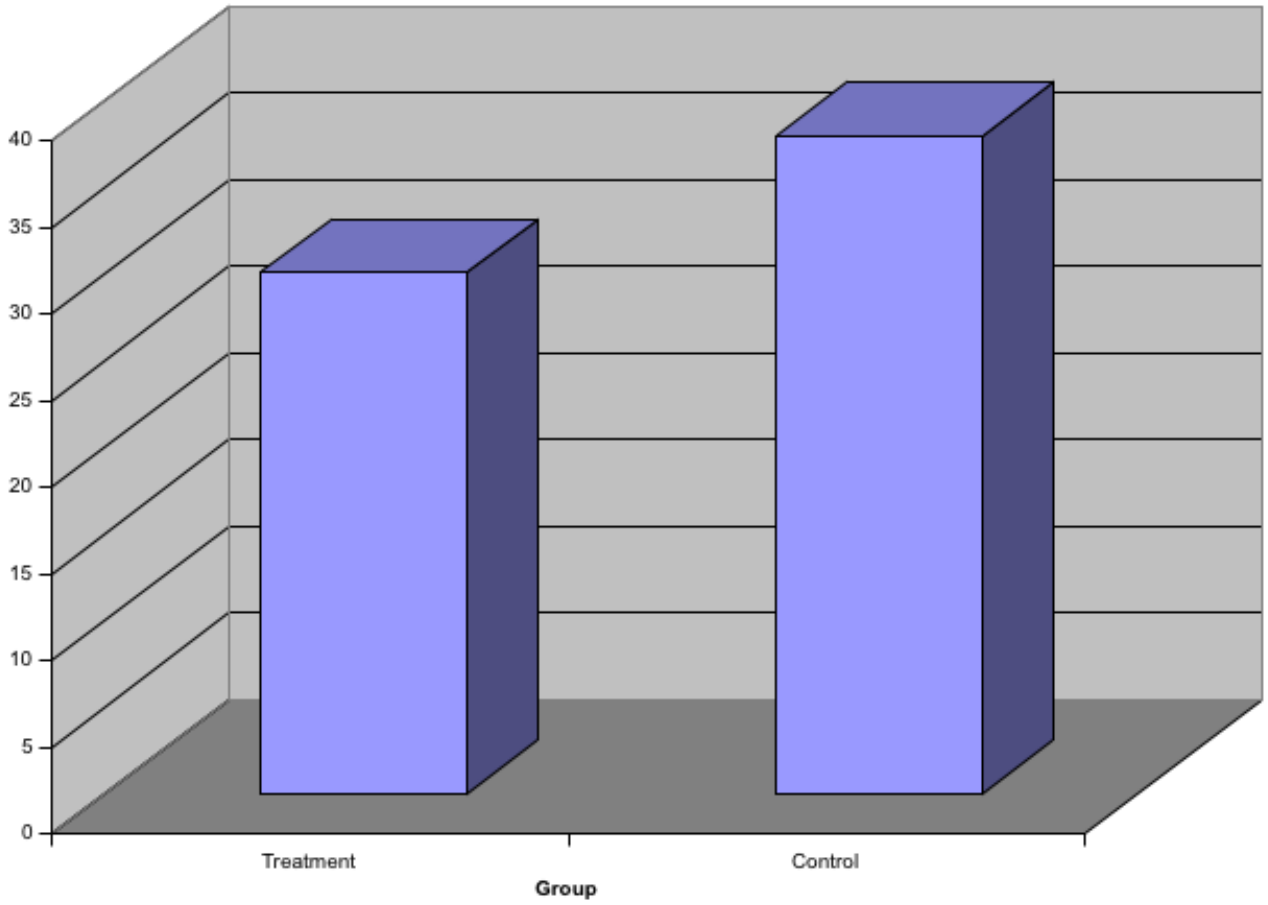
Avoid pseudo three dimensional plots



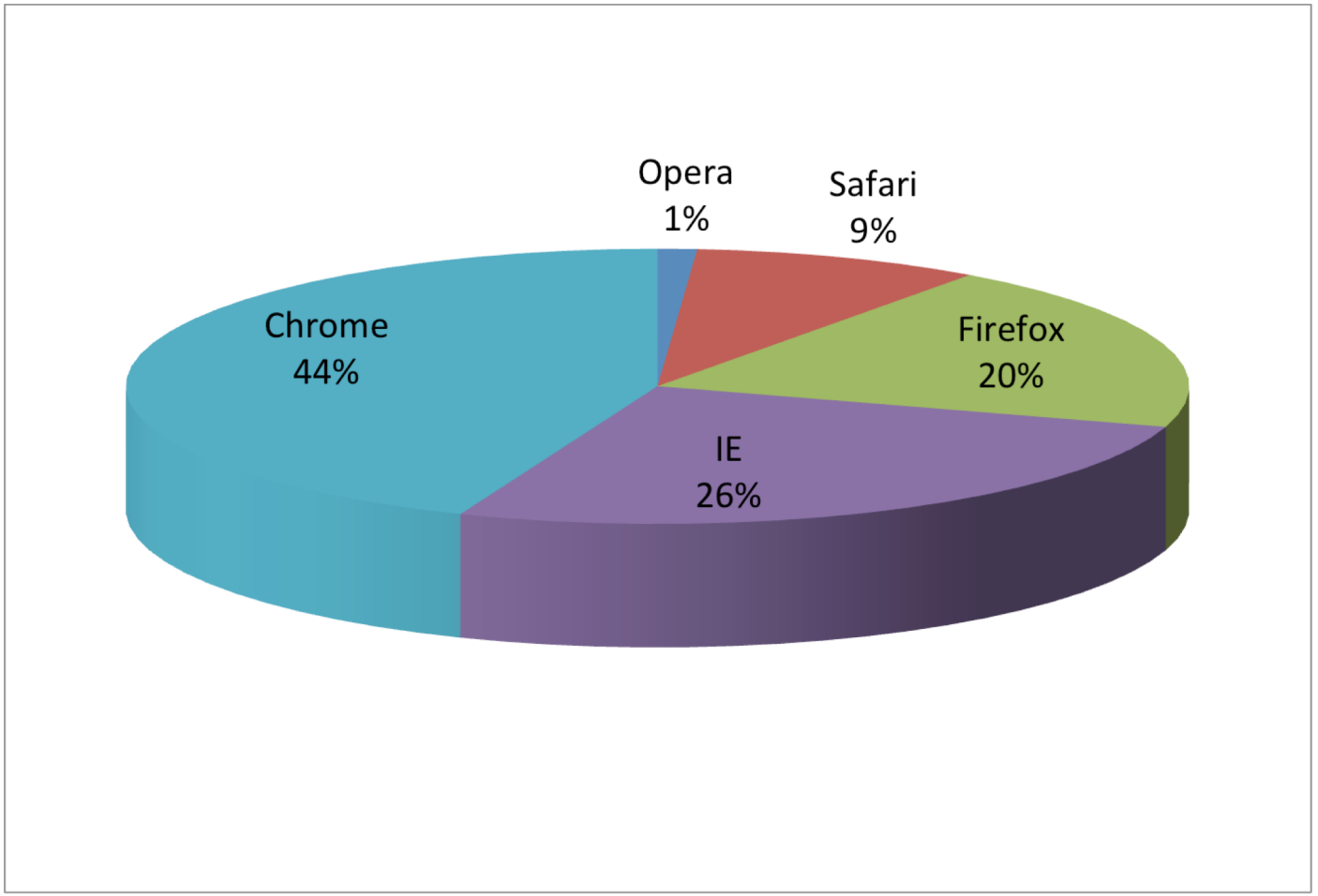
Avoid gratuitousthree dimensional plots

- Pseudo 3D is sometimes used completely gratuitously: plots are made to look 3D even when the 3rd dimension does not represent a quantity.
- This only adds confusion and makes it harder to relay your message.

Avoid gratuitousthree dimensional plots



Avoid gratuitousthree dimensional plots



Avoid too any significant digits

- By default, statistical software like R returns many significant digits.
- The default behavior in R is to show 7 significant digits.
- So many digits often adds no information and the visual clutter than can makes it hard for the consumer of your table to understand the message.
- As an example here are the per 10,000 disease rates for California across the five decades

Avoid too any significant digits

state	year	Measles	Pertussis	Polio
California	1940	37.8826320	18.3397861	18.3397861
California	1950	13.9124205	4.7467350	4.7467350
California	1960	14.1386471	0.0000000	0.0000000
California	1970	0.9767889	0.0000000	0.0000000
California	1980	0.3743467	0.0515466	0.0515466

Avoid too any significant digits

- We are reporting precision up to 0.00001 cases per 10,000, a very small value in the context the changes that are occurring across the dates.
- In this case 2 significant figure is more than enough and makes the point that rates are decreasing clearly:

Avoid too any significant digits

```
## Warning: funs() is soft deprecated as of dplyr
0.8.0
## please use list() instead
##
## # Before:
## funs(name = f(.)
##
## # After:
## list(name = ~f(.))
## This warning is displayed once per session.
```

state	year	Measles	Pertussis	Polio
California	1940	37.9	18.3	18.3
California	1950	13.9	4.7	4.7
California	1960	14.1	0.0	0.0
California	1970	1.0	0.0	0.0
California	1980	0.4	0.1	0.1

Compare vertically

– Another principle, related to displaying tables, is to place values being compared on columns rather than rows.

Compare vertically

state	year	Measles	Pertussis	Polio
California	1940	37.9	18.3	18.3
California	1950	13.9	4.7	4.7
California	1960	14.1	0.0	0.0
California	1970	1.0	0.0	0.0
California	1980	0.4	0.1	0.1

Do not compare horizontally

state	disease	1940	1950	1960	1970	1980
California	Measles	37.9	13.9	14.1	1	0.4
California	Pertussis	18.3	4.7	0.0	0	0.1
California	Polio	18.3	4.7	0.0	0	0.1

Further reading:

- ER Tufte (1983) The visual display of quantitative information. Graphics Press.
- ER Tufte (1990) Envisioning information. Graphics Press.
- ER Tufte (1997) Visual explanations. Graphics Press.
- WS Cleveland (1993) Visualizing data. Hobart Press.
- WS Cleveland (1994) The elements of graphing data. CRC Press.
- A Gelman, C Pasarica, R Dodhia (2002) Let's