STU SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

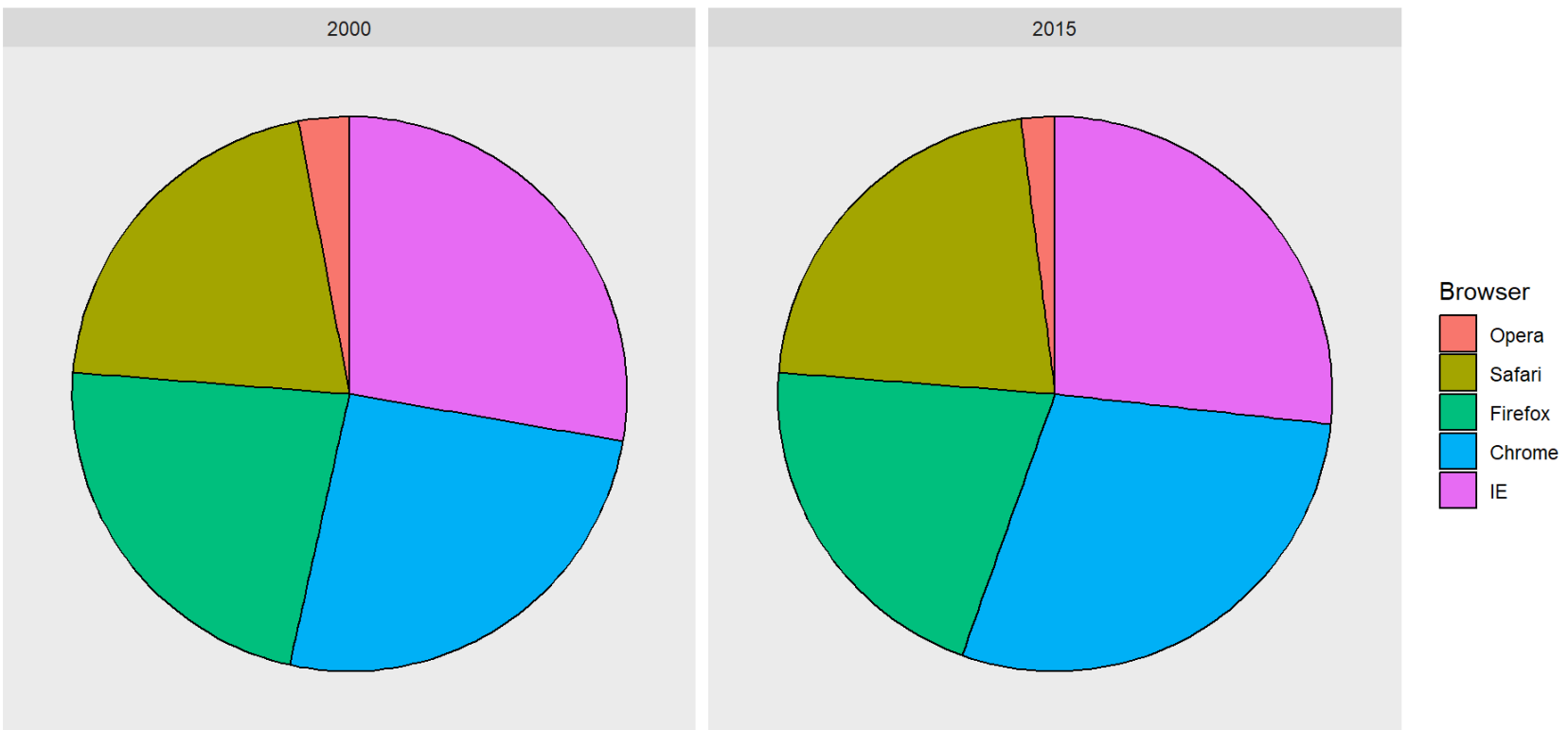
Data visualization principles

Dr.Sc. Oleksii Yehorchenkov
Department of Spatial Planning

Encoding data using visual cues

We start by describing some principles for encoding data. There are several visual cues at our disposal including position, aligned lengths, angles, area, brightness, and color hue.

A widely used graphical representation of percentages, popularized by Microsoft Excel, is the pie chart:



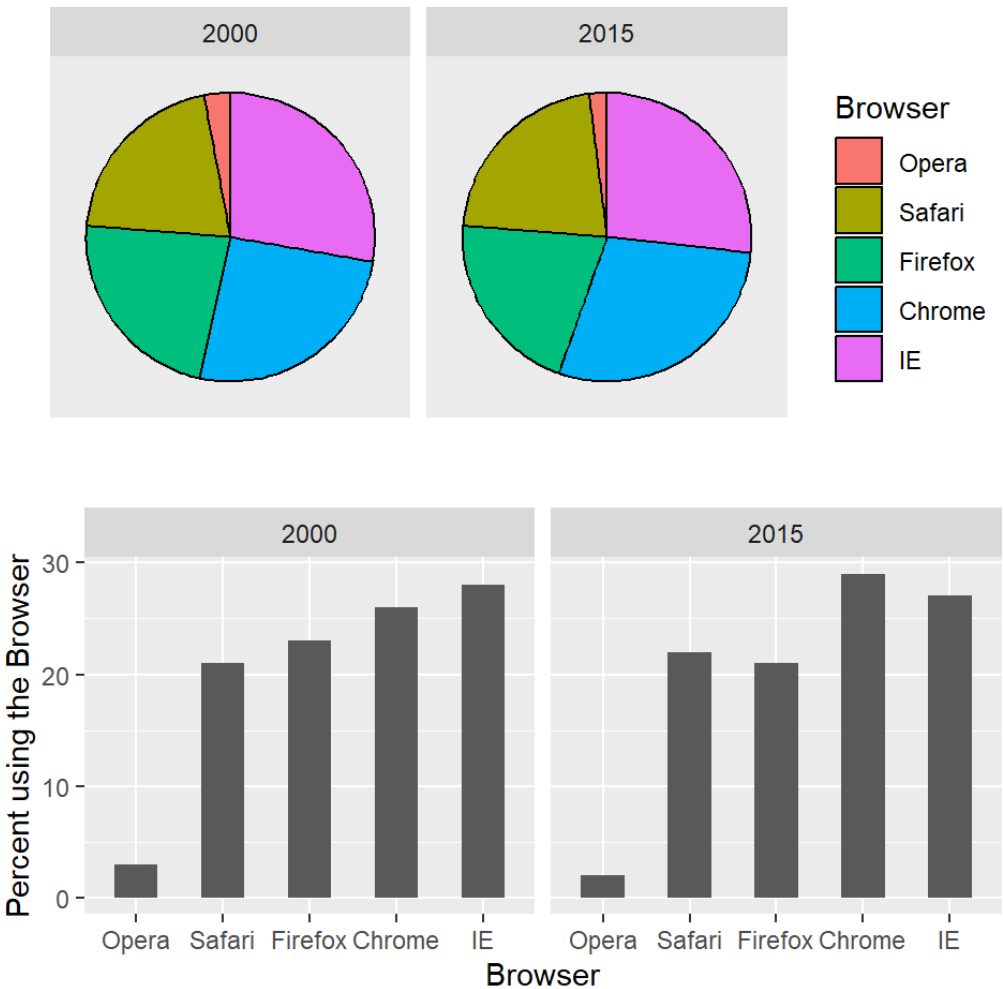
Encoding data using visual cues

To see how hard it is to quantify angles and area, note that the rankings and all the percentages in the plots above changed from 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.

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

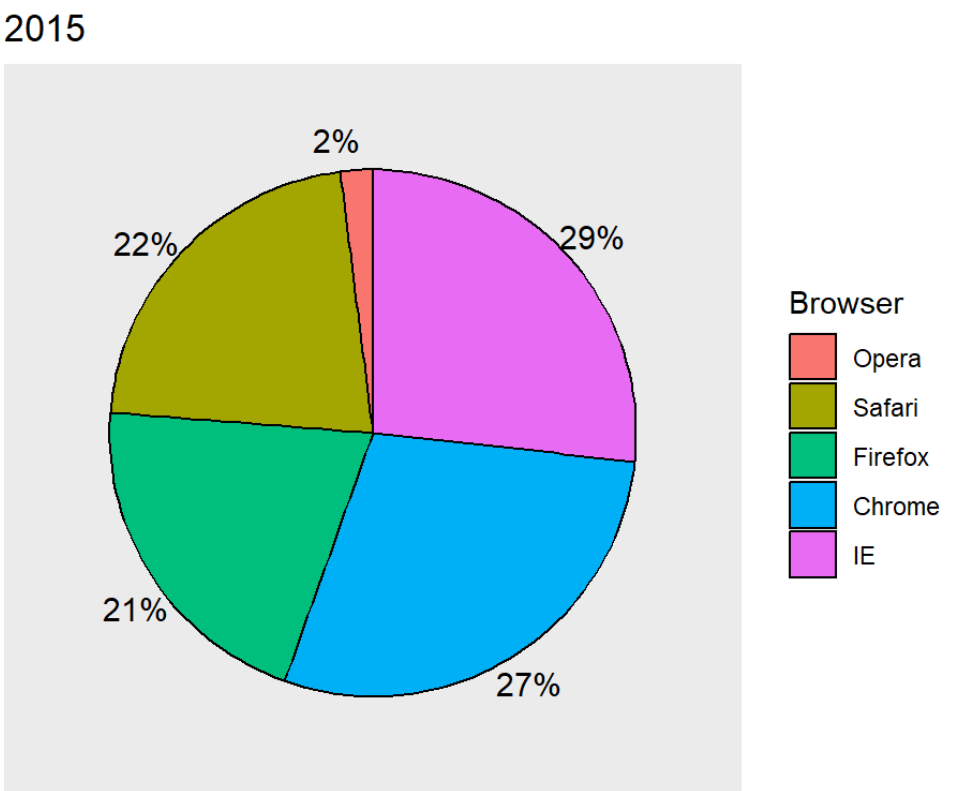
Encoding data using visual cues

The preferred way to plot these quantities is to use length and position as visual cues, since humans are much better at judging linear measures.



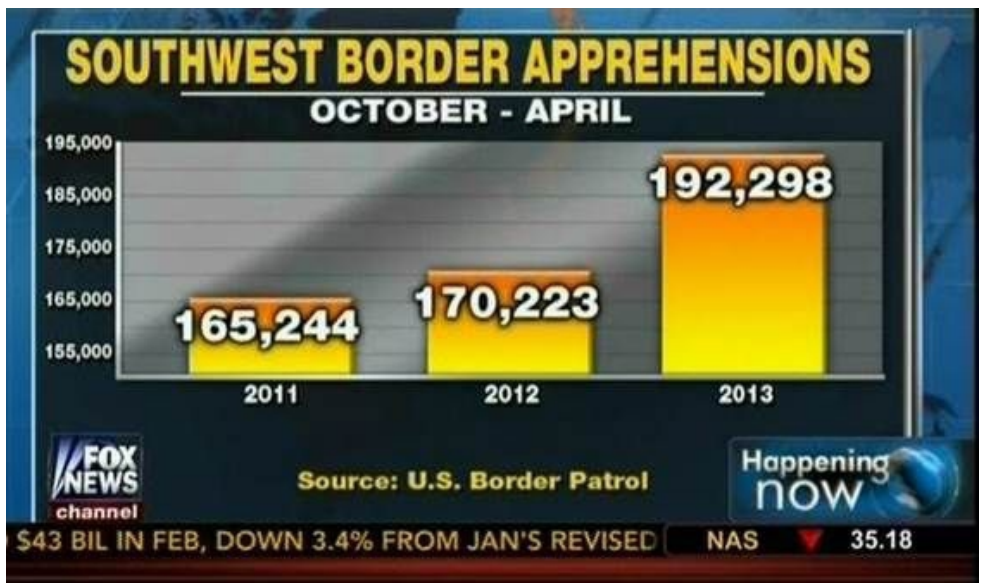
Encoding data using visual cues

If for some reason you need to make a pie chart, label each pie slice with its respective percentage so viewers do not have to infer them from the angles or area:

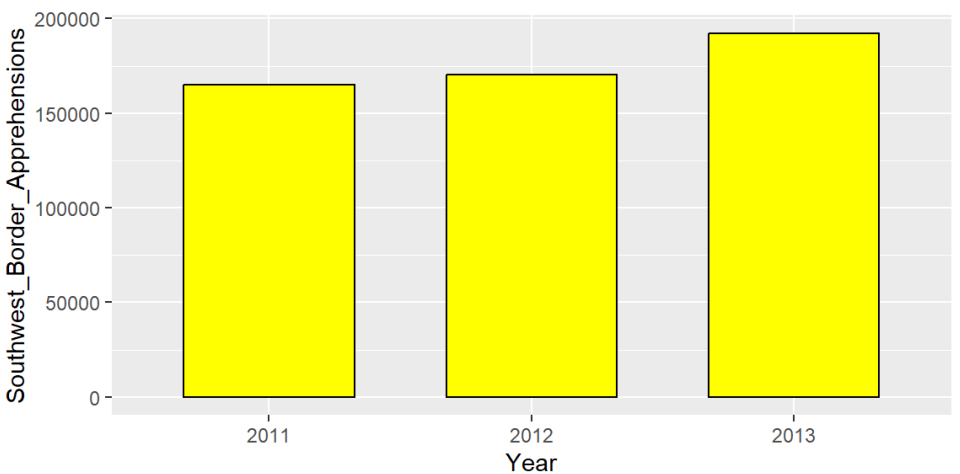


Know when to include 0

By avoiding 0, relatively small differences 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.

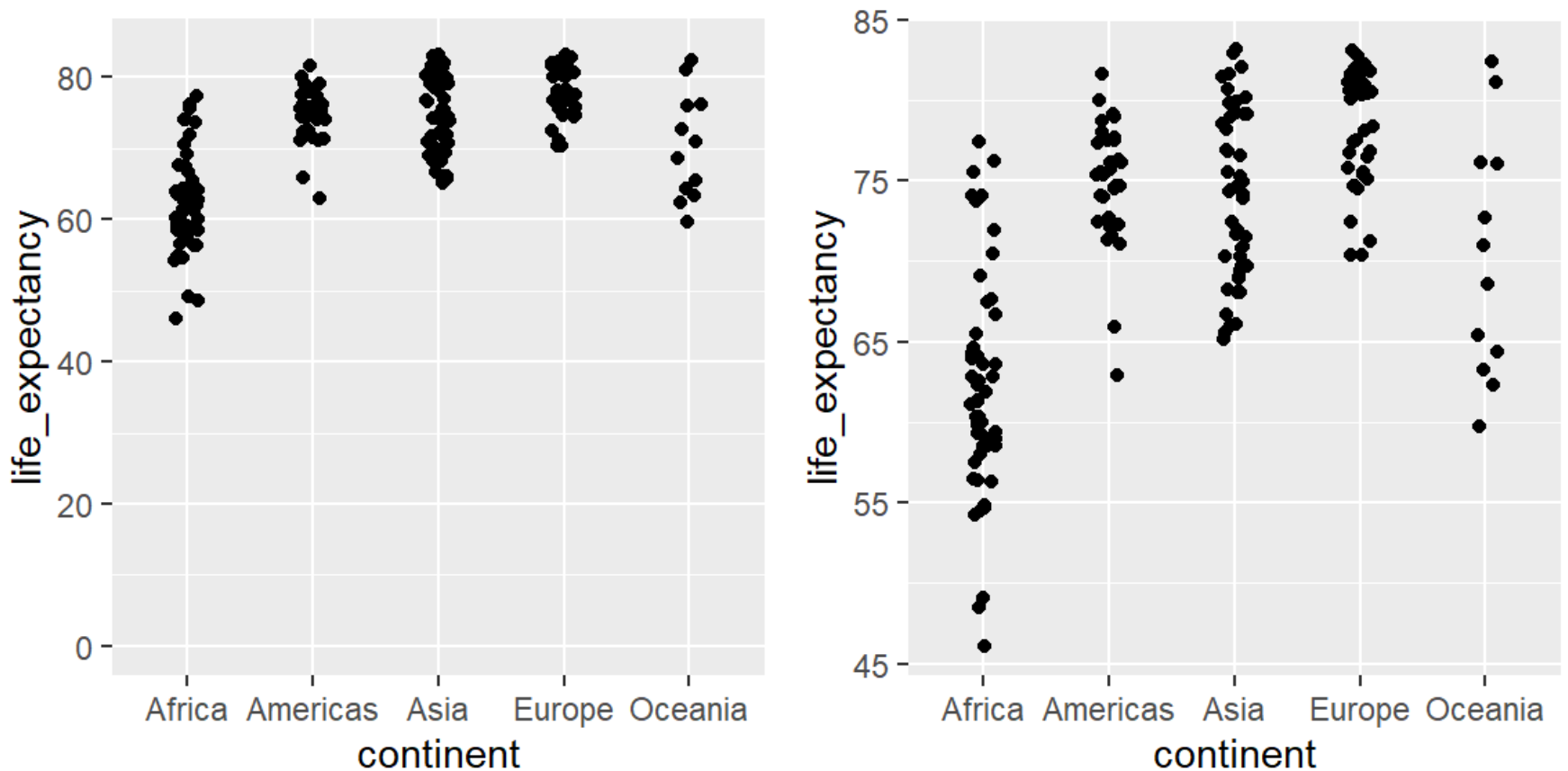


From the left plot, it appears that apprehensions have almost tripled when, in fact, they have only increased by about 16%. Starting the graph at 0 illustrates this clearly:



Know when to include 0 (cont'd)

When using position rather than length, it is then not necessary to include 0. This is particularly the case when we want to compare differences between groups relative to the within-group variability. Here is an illustrative example showing country average life expectancy stratified across continents in 2012:

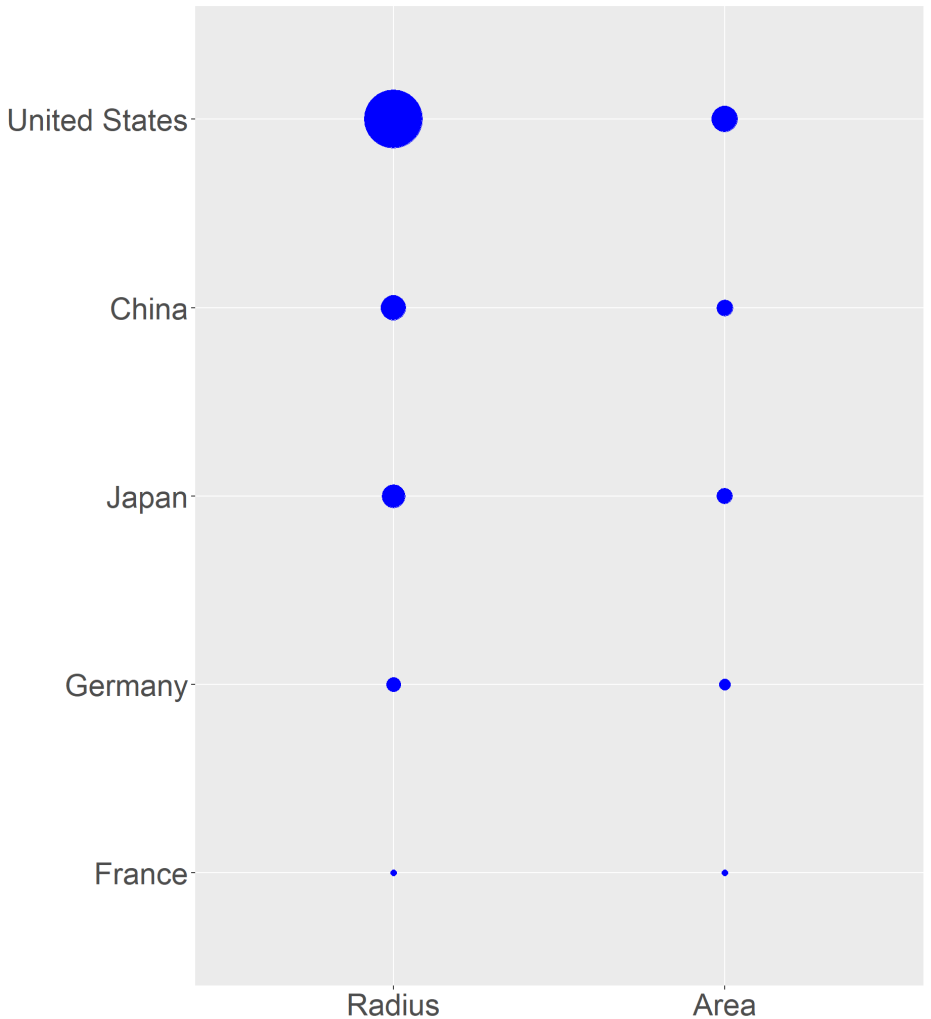


Do not distort quantities

(Source: The 2011 State of the Union Address)

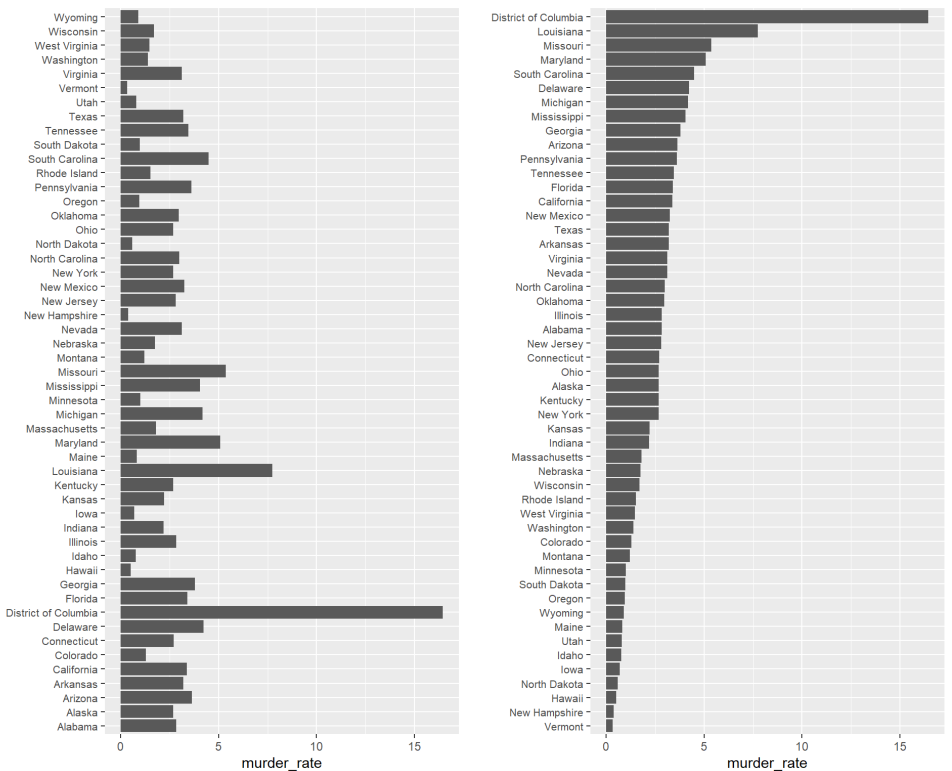
Do not distort quantities (cont'd)

Judging by the area of the circles, the US appears to have an economy over five times larger than China's and over 30 times larger than France's. However, if we look at the actual numbers, we see that this is not the case. The actual ratios are 2.6 and 5.8 times bigger than China and France, respectively. The reason for this distortion is that the radius, rather than the area, was made to be proportional to the quantity, which implies that the proportion between the areas is squared: 2.6 turns into 6.5 and 5.8 turns into 34.1. Here is a comparison of the circles we get if we make the value proportional to the radius and to the area:



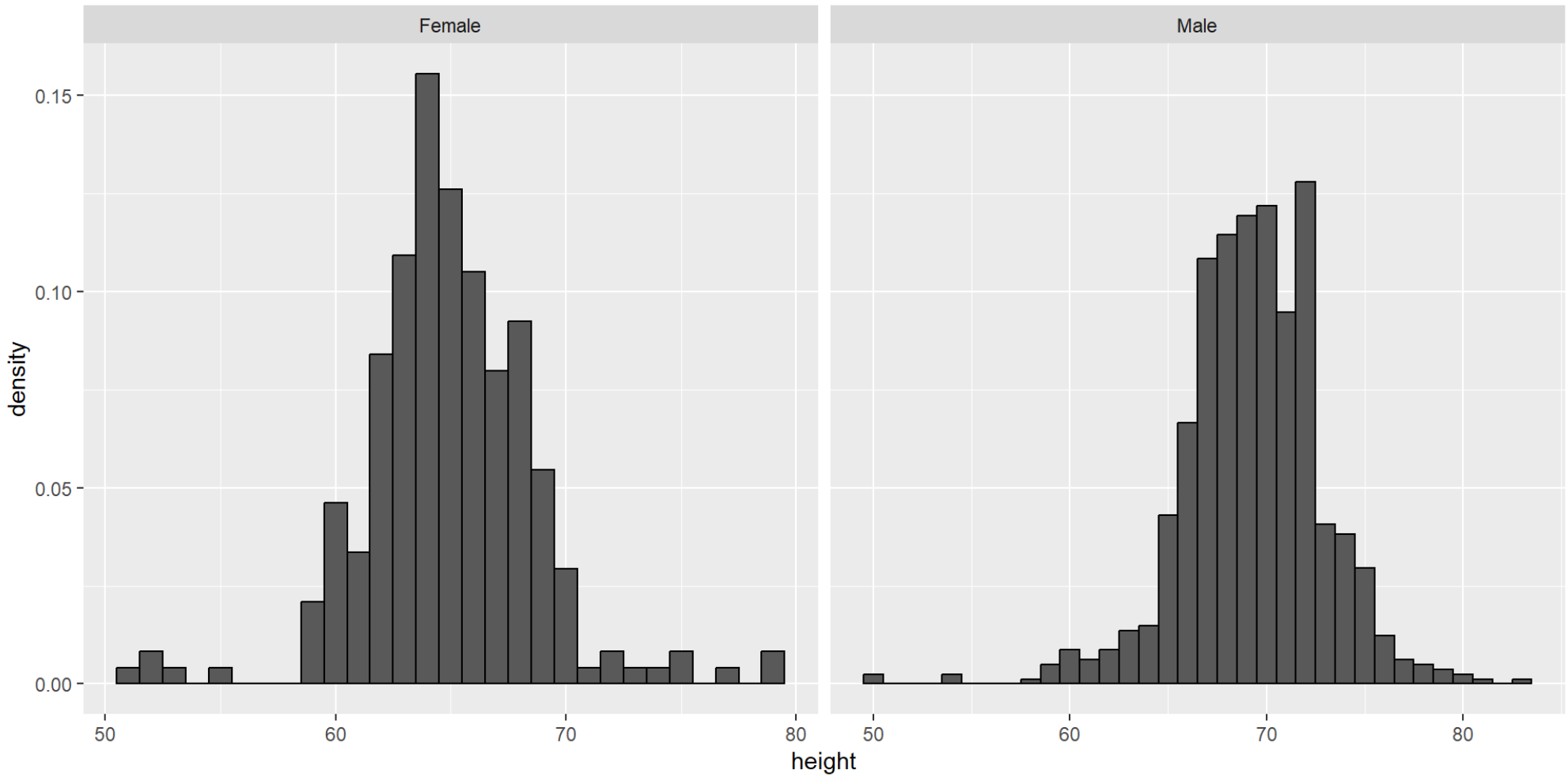
Order categories by a meaningful value

When one of the axes is used to show categories, as is done in barplots and boxplots, the default ggplot2 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.



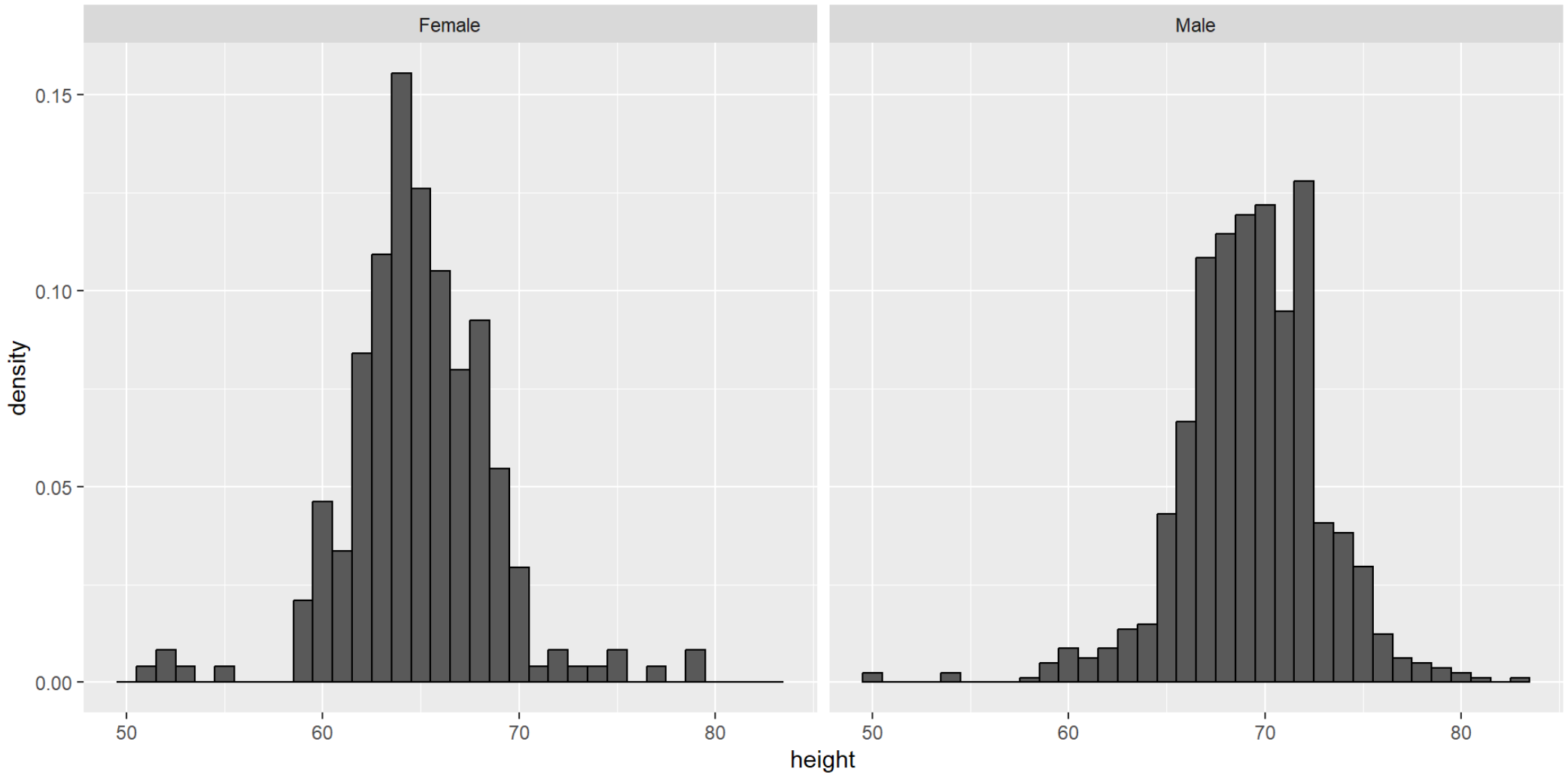
Ease comparison

Use common axes



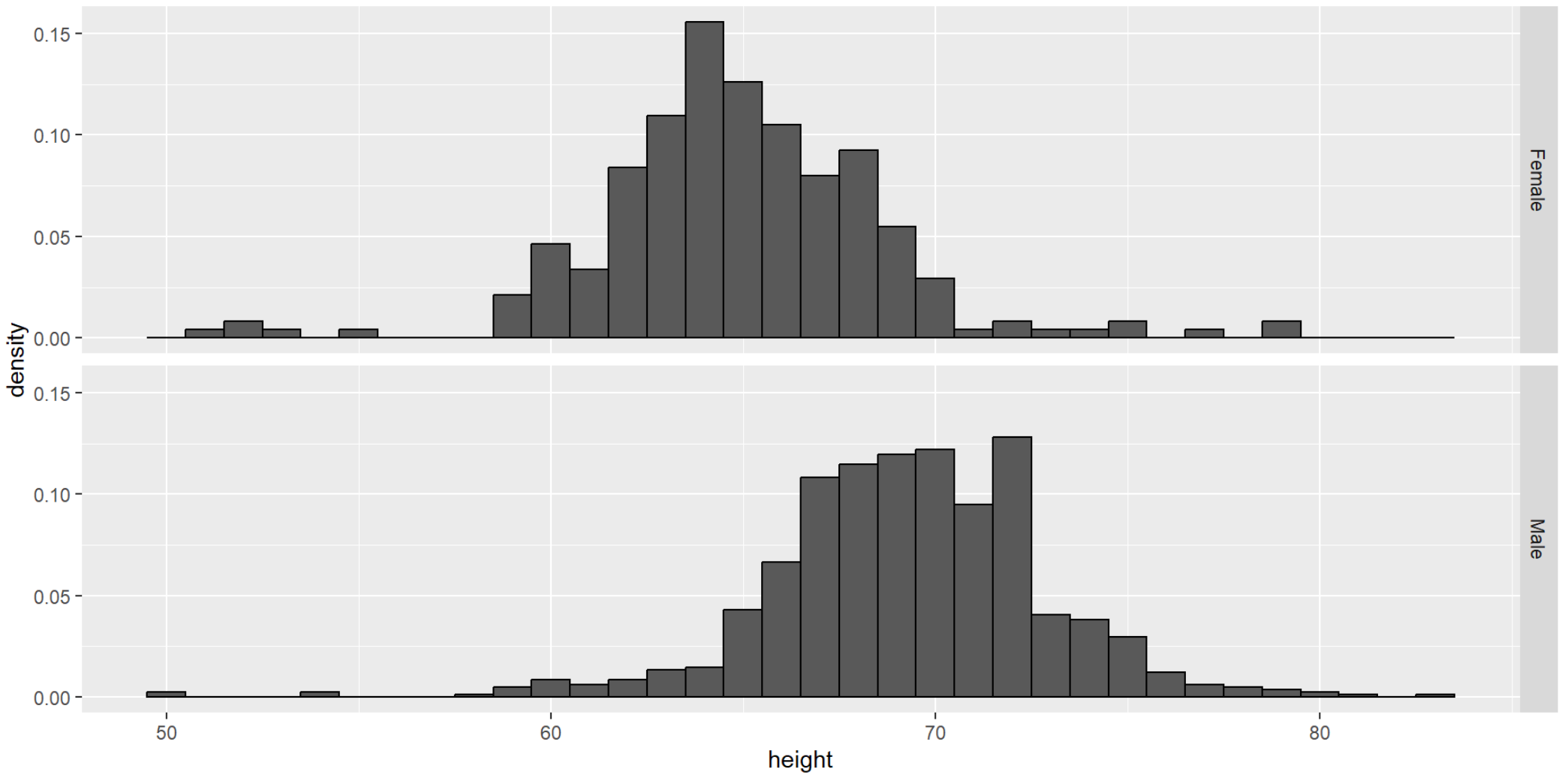
Use common axes (cont'd)

For better comparison keep the axes the same when comparing data across two plots.



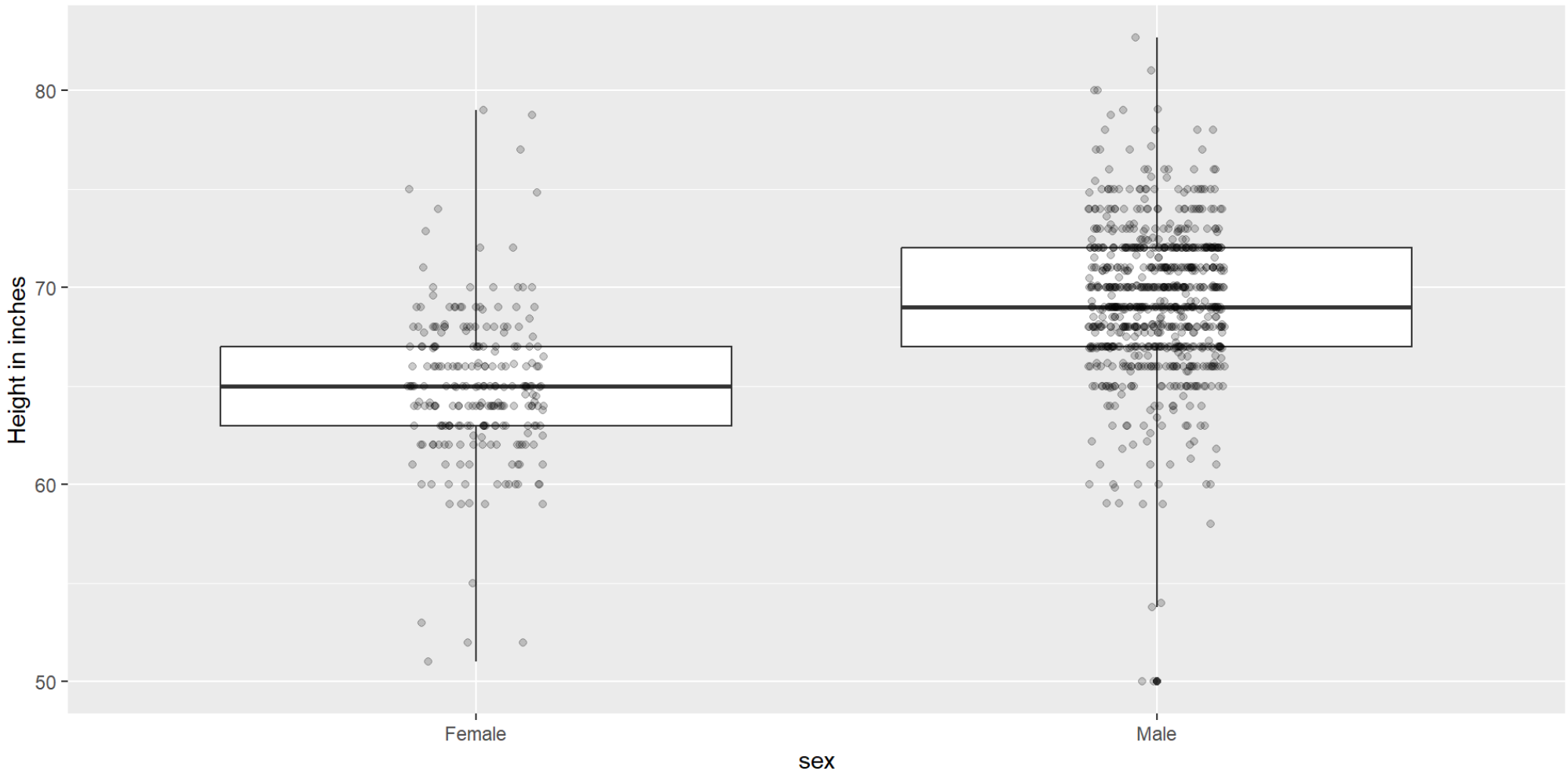
Align plots vertically to see horizontal changes

This plot makes it much easier to notice that men are, on average, taller.



and horizontally to see vertical changes

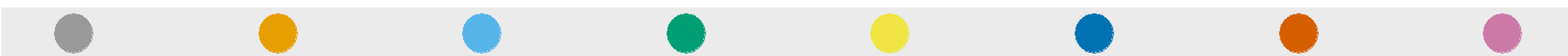
If we want the more compact summary provided by boxplots, we then align them horizontally since, by default, boxplots move up and down with changes in height.



Think of the color blind

About 10% of the population is color blind. Unfortunately, the default colors used in **ggplot2** are not optimal for this group. However, **ggplot2** does make it easy to change the color palette used in the plots. An example of how we can use a color blind friendly palette is described here: http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/#a-colorblind-friendly-palette:

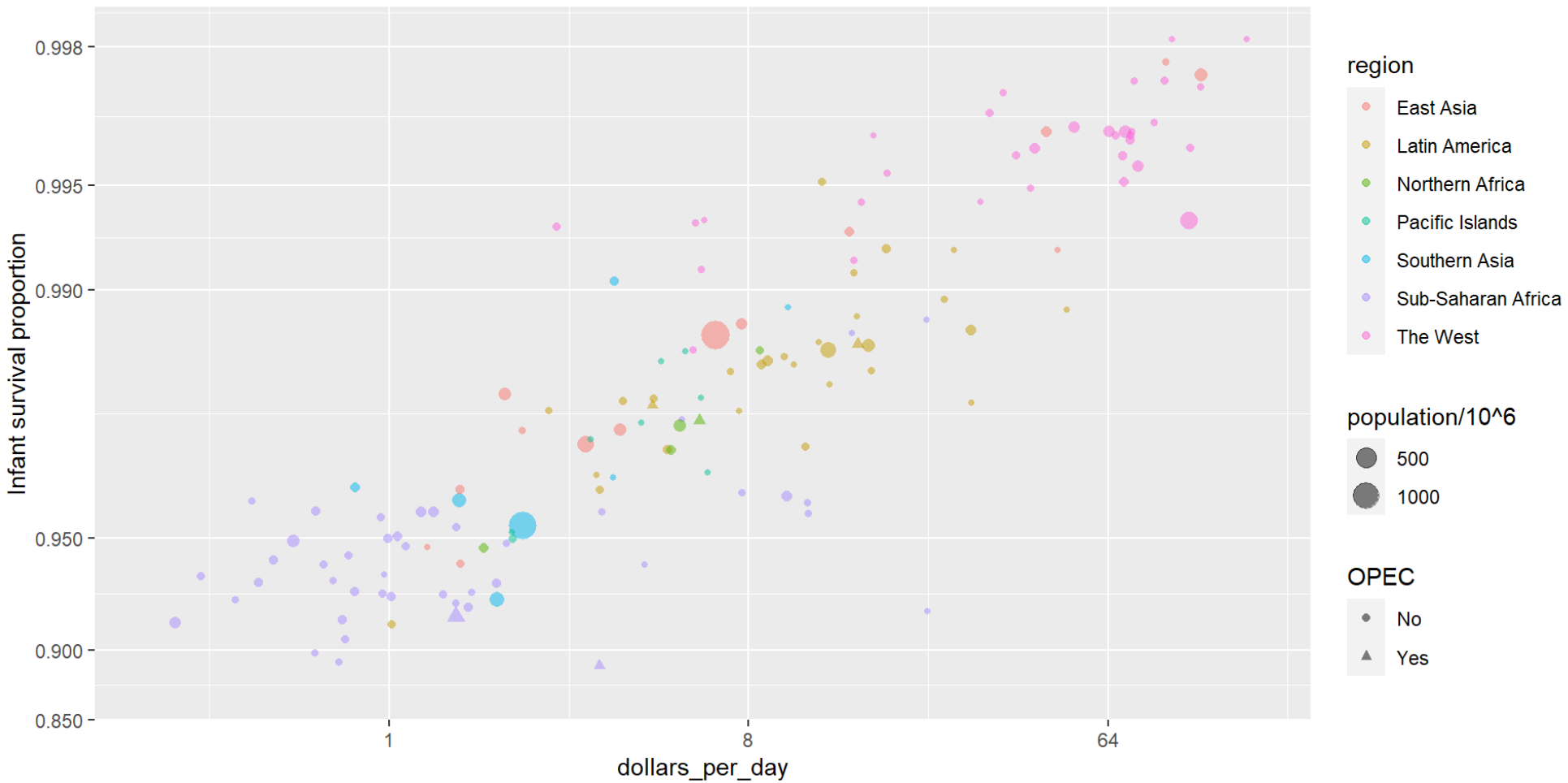
Here are the colors



There are several resources that can help you select colors, for example this one: http://bconnelly.net/2013/10/creating-colorblind-friendly-figures/.

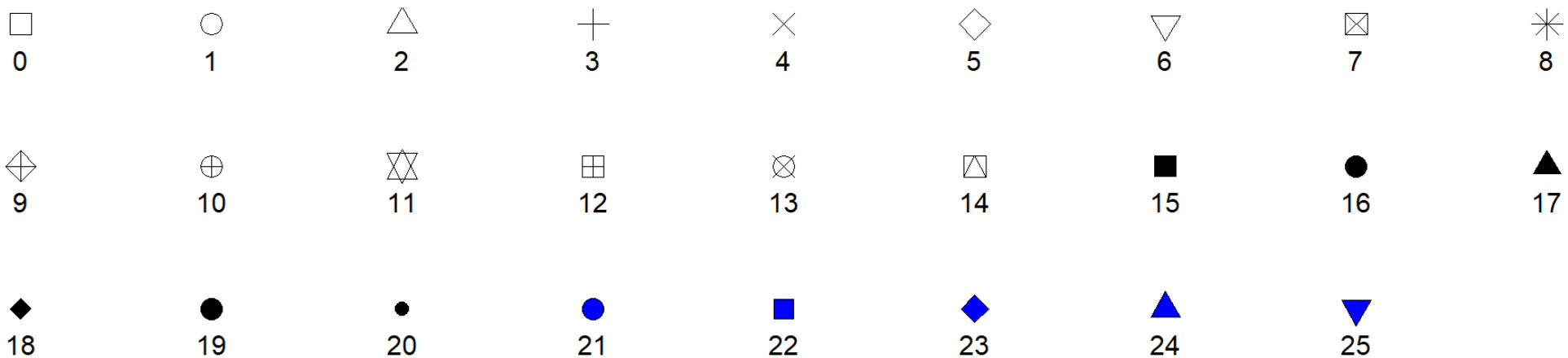
Encoding a third variable

This scatterplot showed the relationship between infant survival and average income. Below is a version of this plot that encodes three variables: OPEC membership, region, and population.

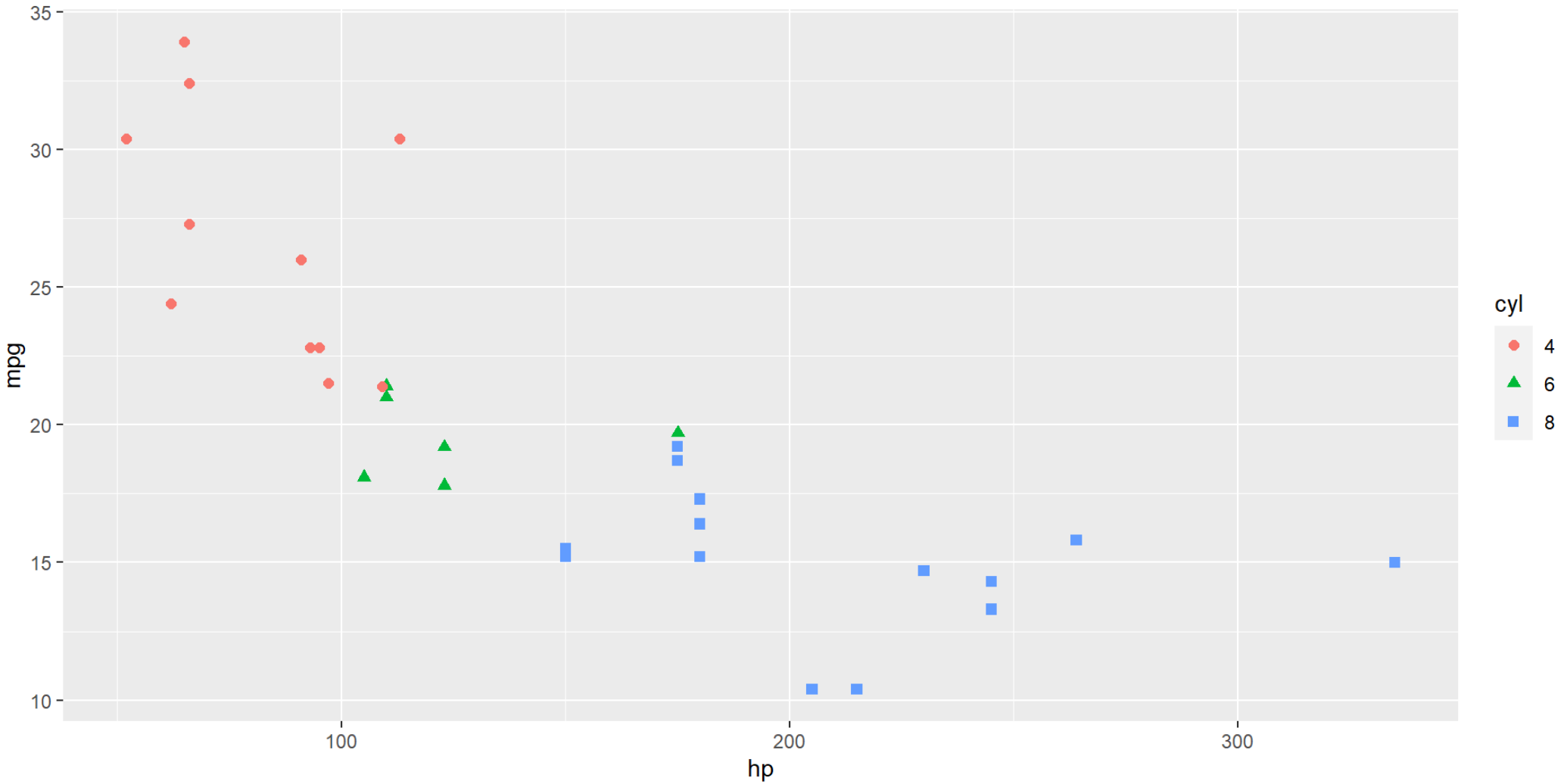


Encoding a third variable (cont'd)

We encode categorical variables with color and shape. These shapes can be controlled with shape argument. Below are the shapes available for use in R. For the last five, the color goes inside.

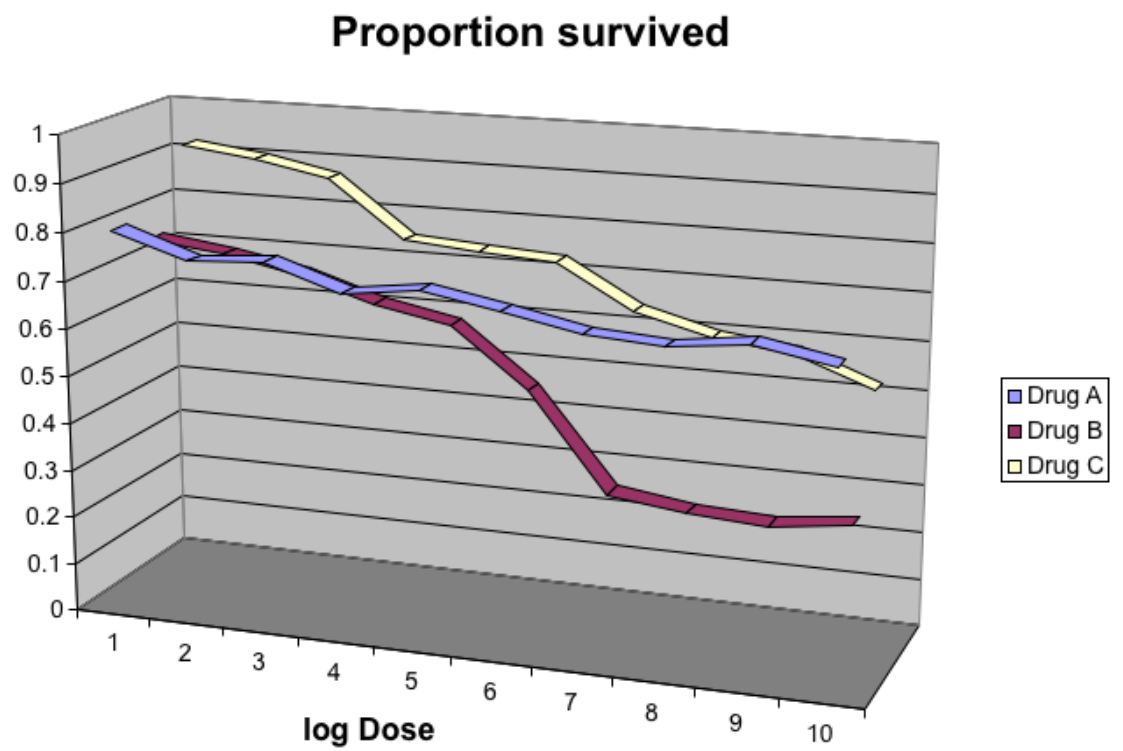


Encoding a third variable (cont'd)

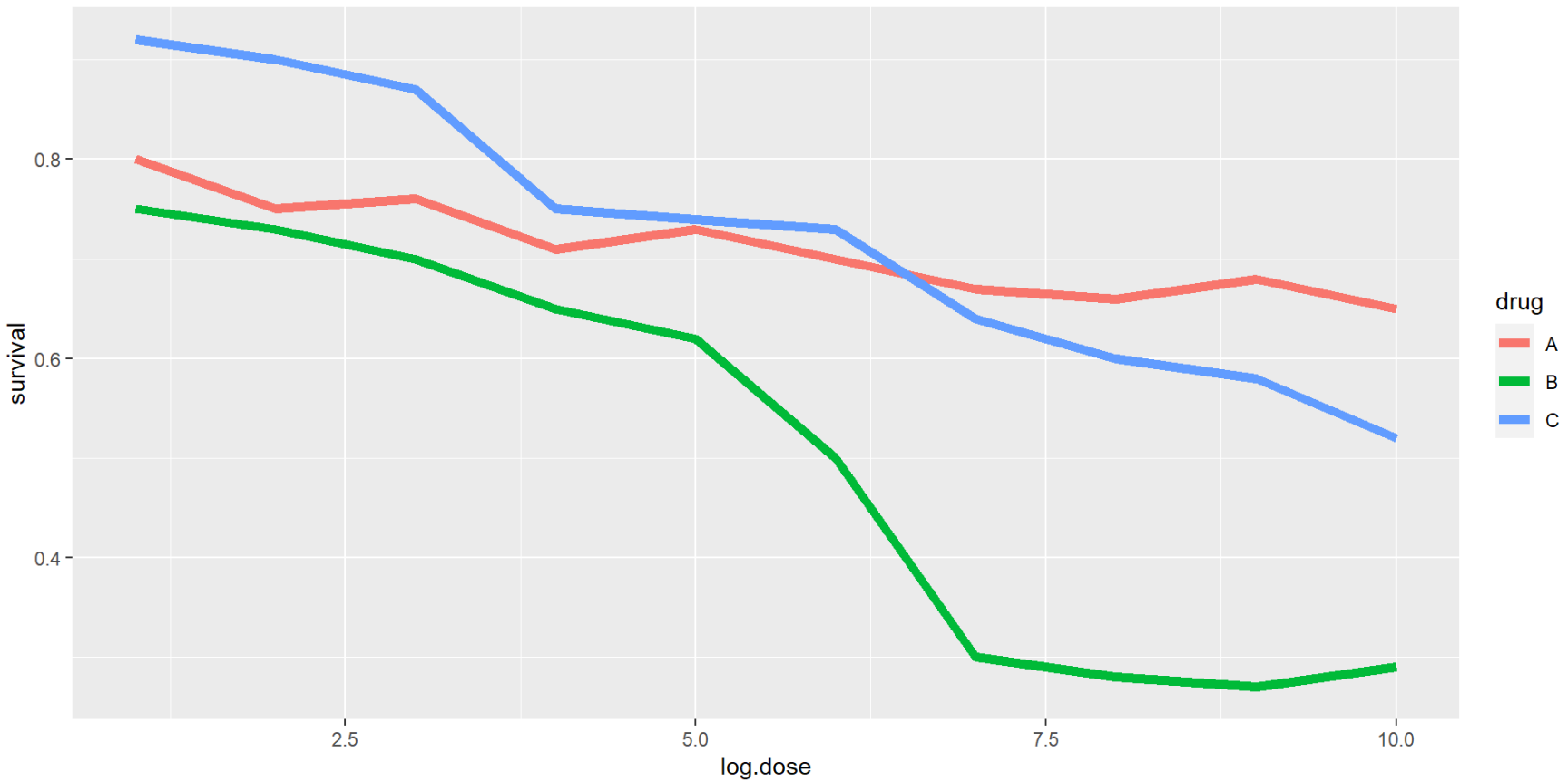


Avoid pseudo-three-dimensional plots

The figure below shows three variables: dose, drug type and survival. The plot tries to imitate three dimensions and assigned a dimension to each variable.



Avoid pseudo-three-dimensional plots (cont'd)



Notice how much easier it is to determine the survival values.

Avoid too many significant digits

state	year	Measles	Pertussis	Polio
California	1940	37.8826320	18.3397861	0.8266512
California	1950	13.9124205	4.7467350	1.9742639
California	1960	14.1386471	NA	0.2640419
California	1970	0.9767889	NA	NA
California	1980	0.3743467	0.0515466	NA

Two significant figures is more than enough and clearly makes the point that rates are decreasing:

state	year	Measles	Pertussis	Polio
California	1940	37.9	18.3	0.8
California	1950	13.9	4.7	2.0
California	1960	14.1	NA	0.3
California	1970	1.0	NA	NA
California	1980	0.4	0.1	NA

Avoid too many significant digits (cont'd)

Another principle related to displaying tables is to place values being compared on columns rather than rows. Note that our table above is easier to read than this one:

state	disease	1940	1950	1960	1970	1980
California	Measles	37.9	13.9	14.1	1	0.4
California	Pertussis	18.3	4.7	NA	NA	0.1
California	Polio	0.8	2.0	0.3	NA	NA

Know your audience

Graphs can be used for

1. our own exploratory data analysis,
2. to convey a message to experts, or
3. to help convery a message to a general audience.

Make sure that the intended audience understands each element of the plot.

References

Rafael A. Irizarry. **Introduction to Data Science**. Data Wrangling and Visualization with R

http://rafalab.dfci.harvard.edu/dsbook-part-1/