Visualization Principles

Max Turgeon

DATA 2010—Tools and Techniques for Data Science

Lecture Objectives

- Choose the right visual cues for your data visualization
- Understand when the axis should include 0
- Identify which visual cues can distort quantities
- Emphasize important comparisons

Motivation

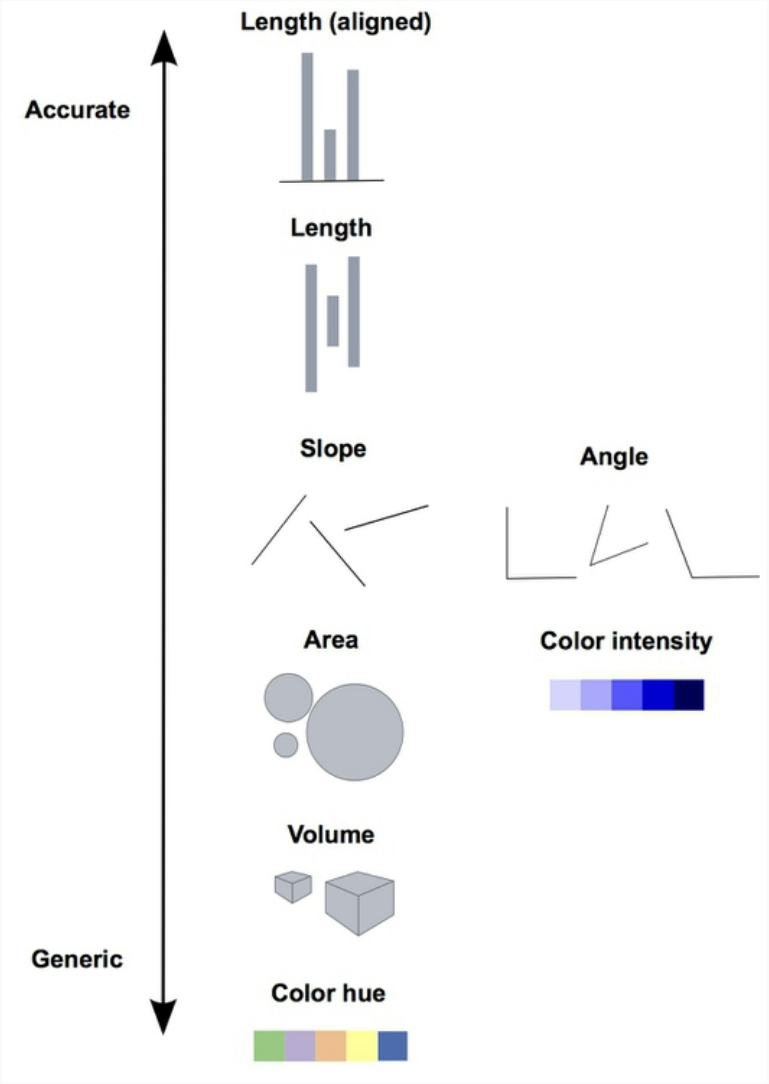
- In the previous lecture, we discussed different types of data visualizations.
- We also discussed their pros and cons.
- But we need general principles of effective data visualizations.
    - Better visualization = better communication

Visual toolbox

- Let's say we want to create a data visualization based on a single continuous variable
    - E.g. age, income, blood pressure.
- What is the best way to highlight *differences* between subgroups?
- As data analysts, we have several visual cues we can use:
    - Position, angle, length, area, colour, angle
- Not all visual cues are created equal!
- It turns out that humans are much better at understanding some visual cues (e.g. lengths) than others (e.g. volumes).

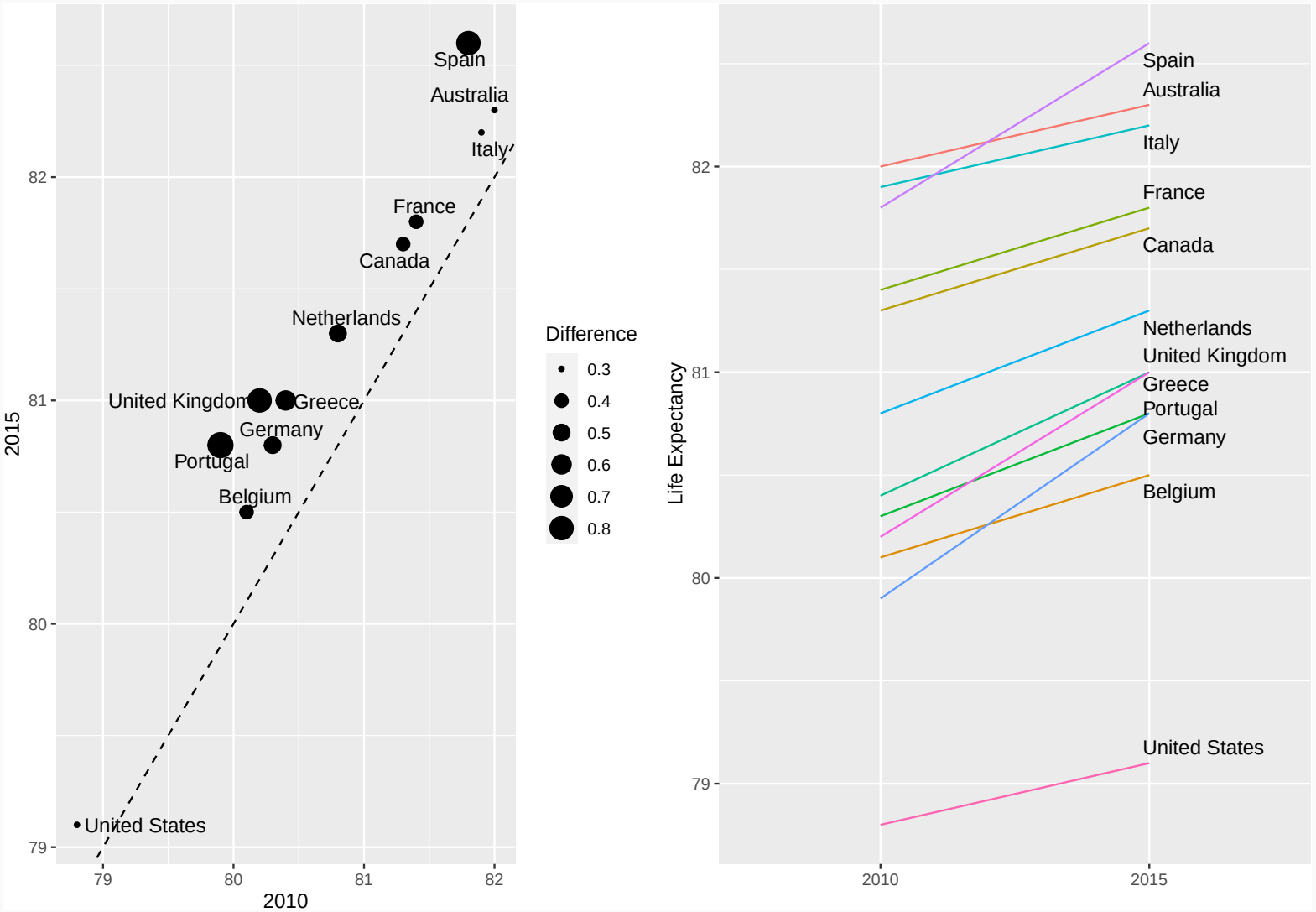
Hierachy of visual cues

Researchers have found that there is a hierarchy of visual cues, from ”accurate” to ”generic”, and we should aim to be as accurate as possible.



Exercise

What are the visual cues?



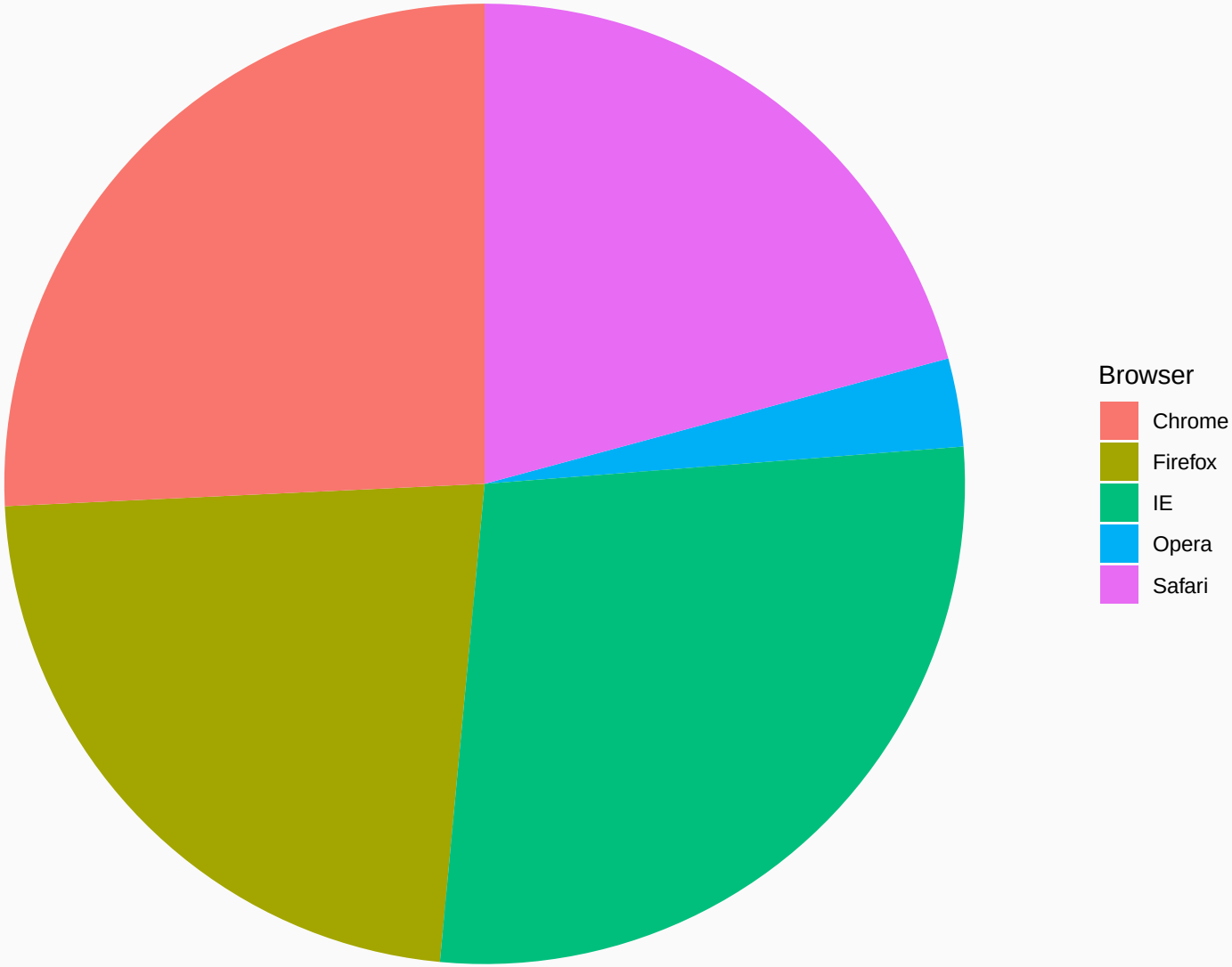
Solution

- Size and slope are used to highlight differences over time.
- Position is used to highlight differences across countries.
- Colour and text are used to identify countries.
- Length/position (with respect to diagonal line) is also used to highlight differences.

Pie charts i

- **Pie charts** form somewhat of a paradox:
  - they are very popular;
  - they are considered poor data visualizations.
- Let's look at an example:

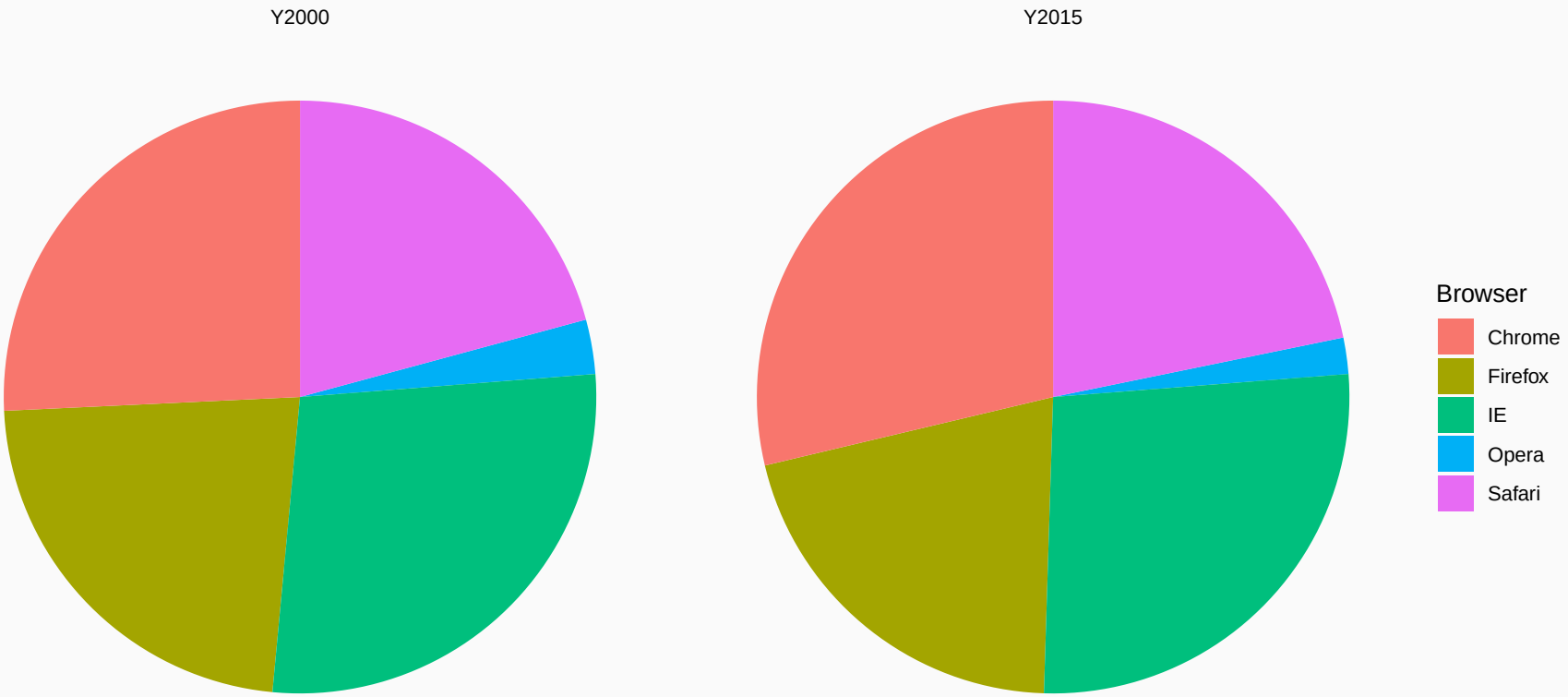
Pie charts ii



Pie charts iii

- Can you really tell which browser is the most popular?
- Pie charts are difficult to read because *angles* can be hard to understand.
    - But they are also very hard to compare.

Pie charts iv

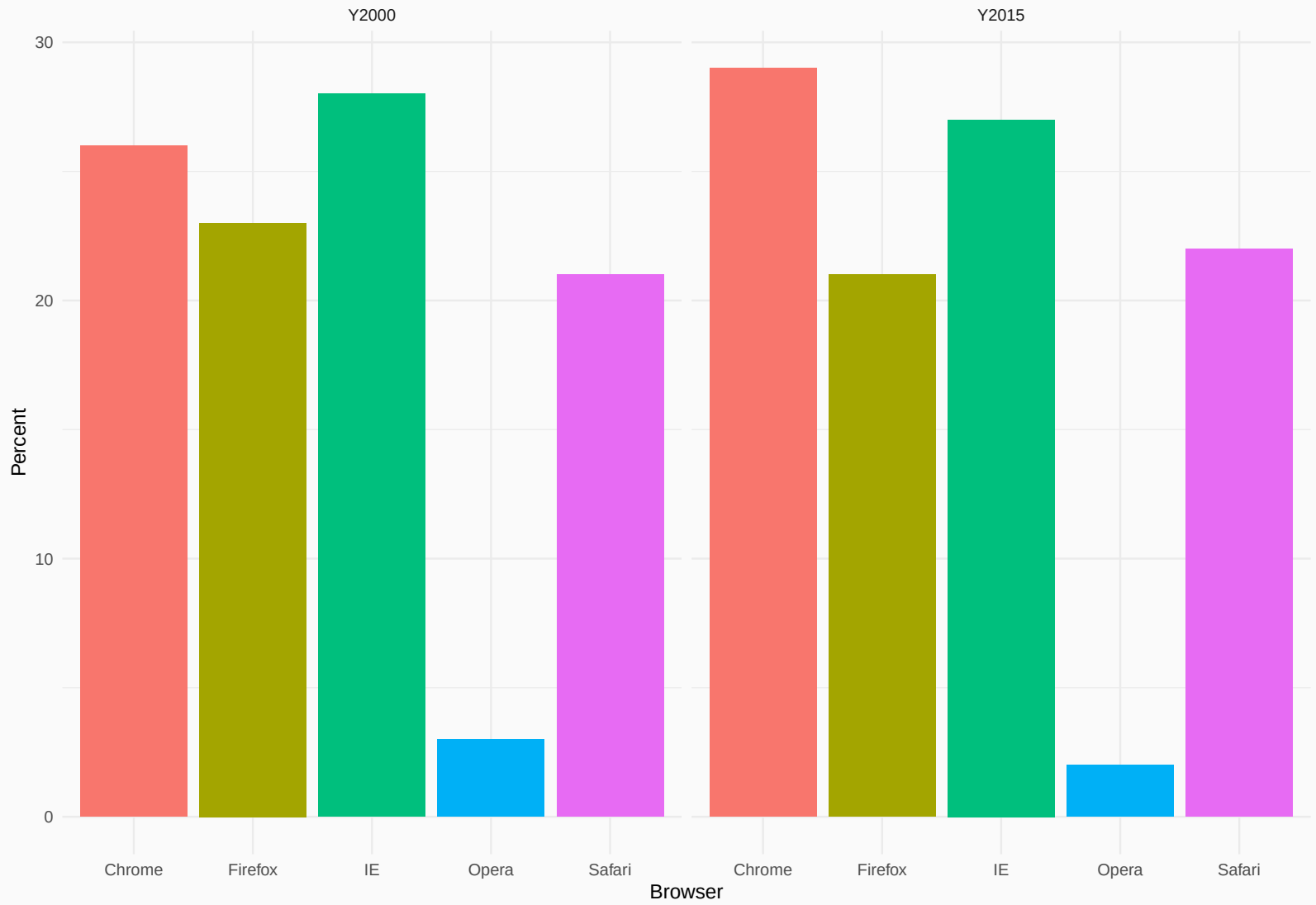


- Can you tell if the preference ranking changed between 2000 and 2015?

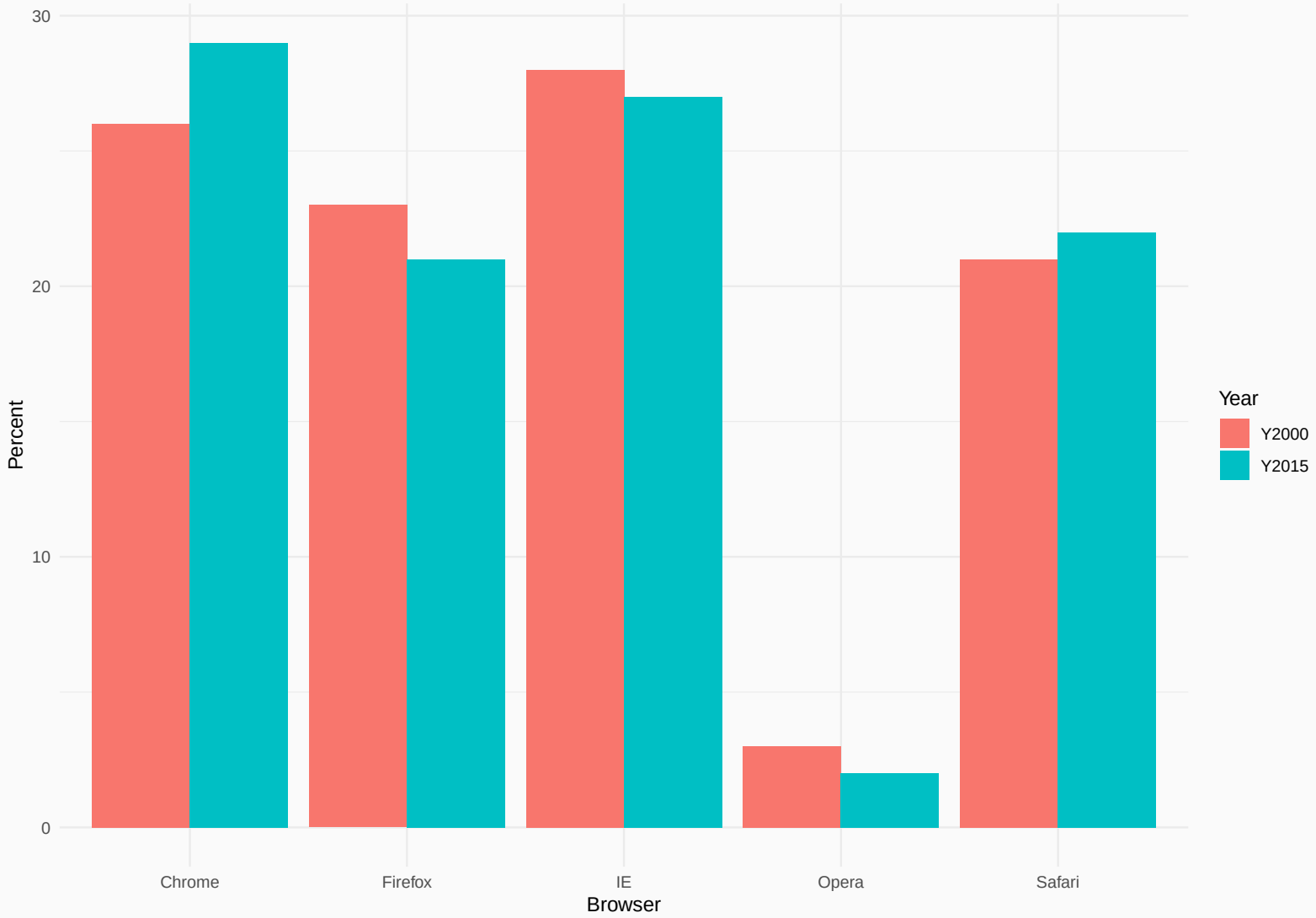
Pie charts v

- **Pie charts can almost always be replaced by a more useful chart** (or even a table).
- In this case, a bar chart can more easily show us changes.

Pie charts vi



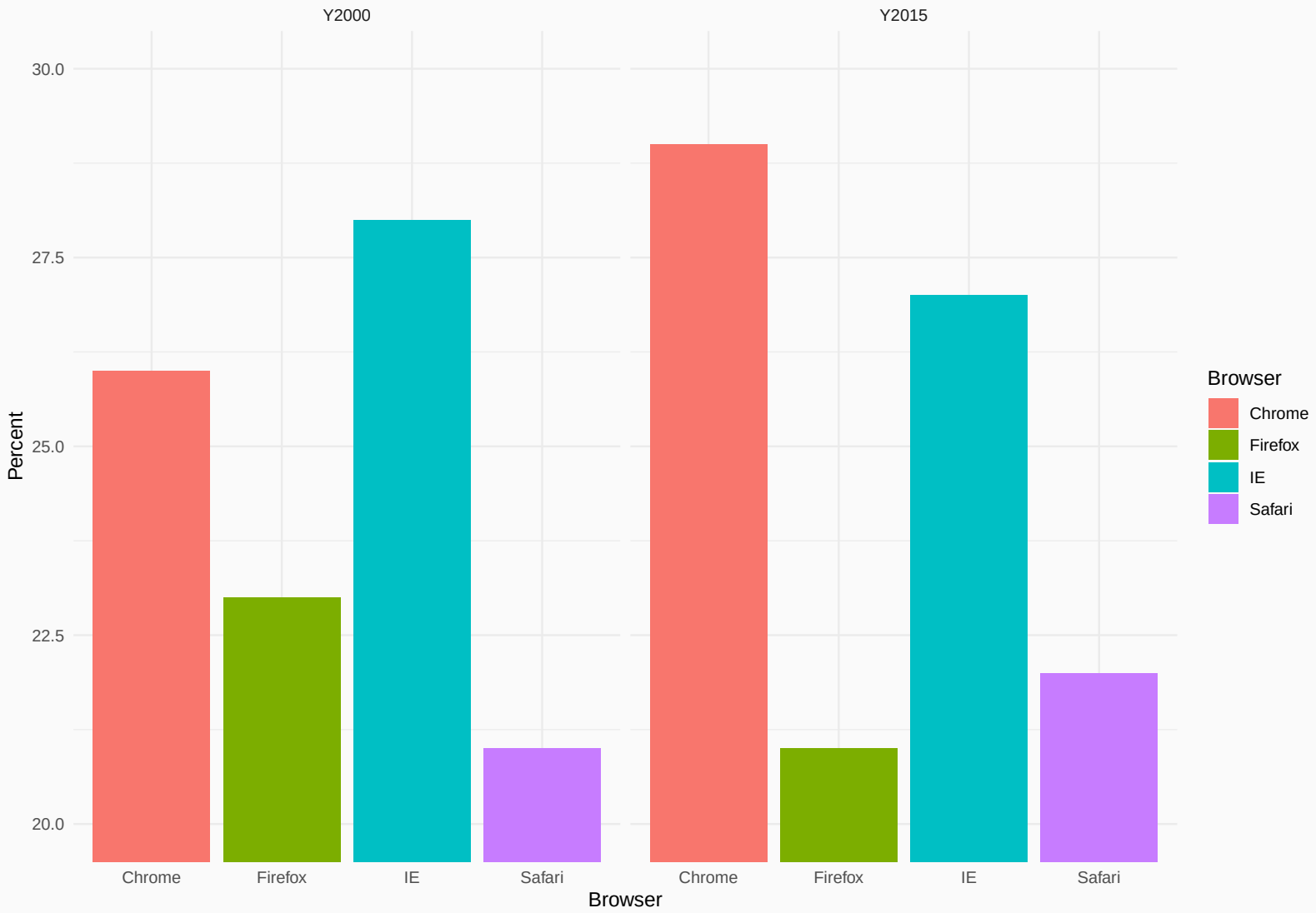
Pie charts vii



Include Zero? i

- When constructing the barplots above, we made an important assumption: the length of the bar is proportional to the quantity being displayed.
- This assumption is important, because it was what allows to compare across browsers and years.
    - Also, it allows us to assess the *relative* difference between them.
- When we don't include zero in a barplot, the lengths are no longer proportional to the quantities.

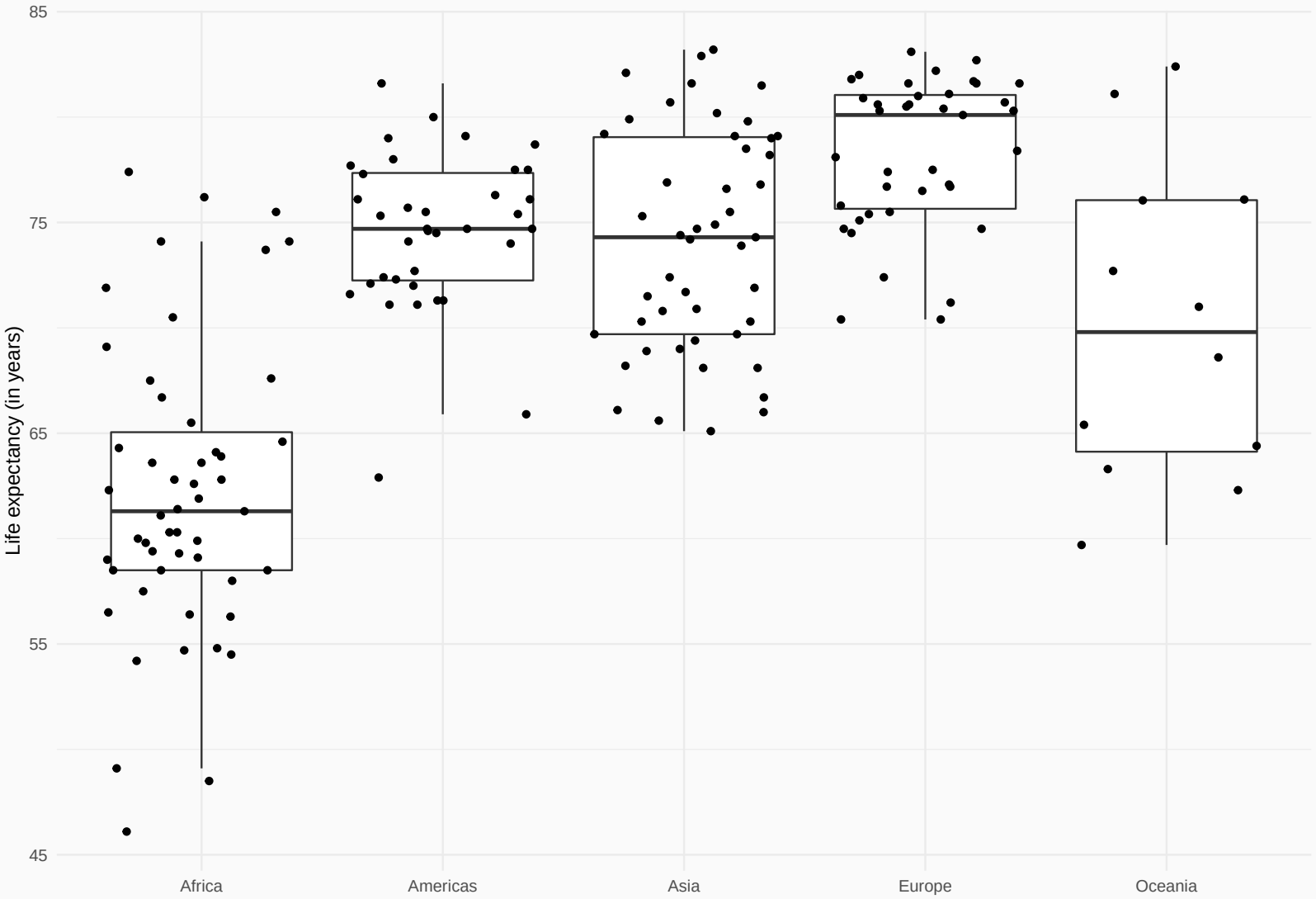
Include Zero? ii



Include Zero? iii

- Differences are exaggerated when we remove the bottom of the bars.
    - Looks like the popularity of Firefox has shrunk by 50%, but the difference is only 2%!
- **When should we include zero?**
    - When the visual cue we use is **length** (e.g. barplots).
- When the visual cue is **position**, then excluding zero does not distort the quantity of interest.

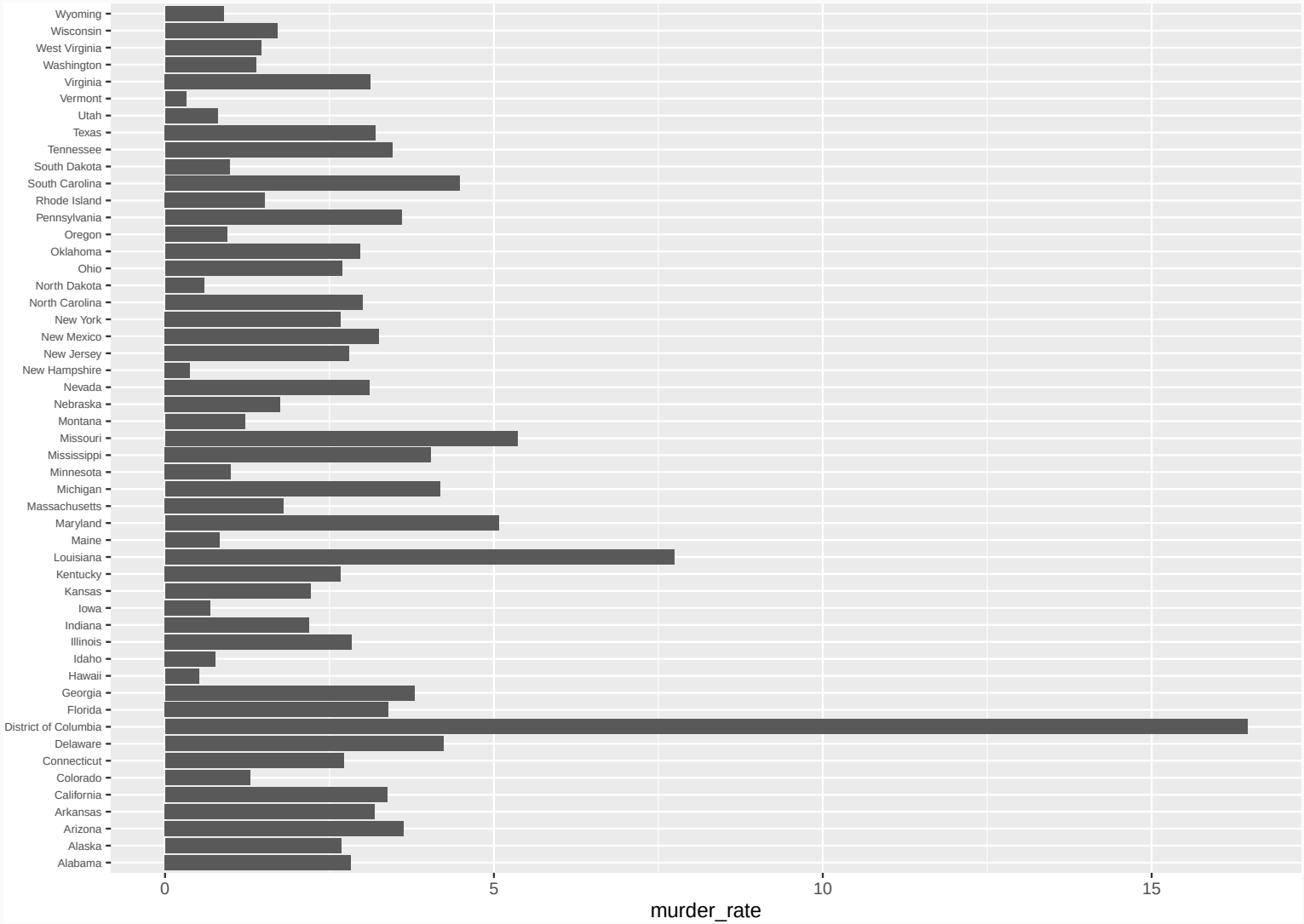
Include Zero? iv



Ordering categories i

- By default, R orders strings in alphabetical order.
  - When is alphabetical order the best way to order information?
  - Wouldn't you rather order Canadian provinces from West to East?
- Reordering categories according to a more meaningful criterion can often improve your data visualization by making it easier to compare.

Ordering categories ii



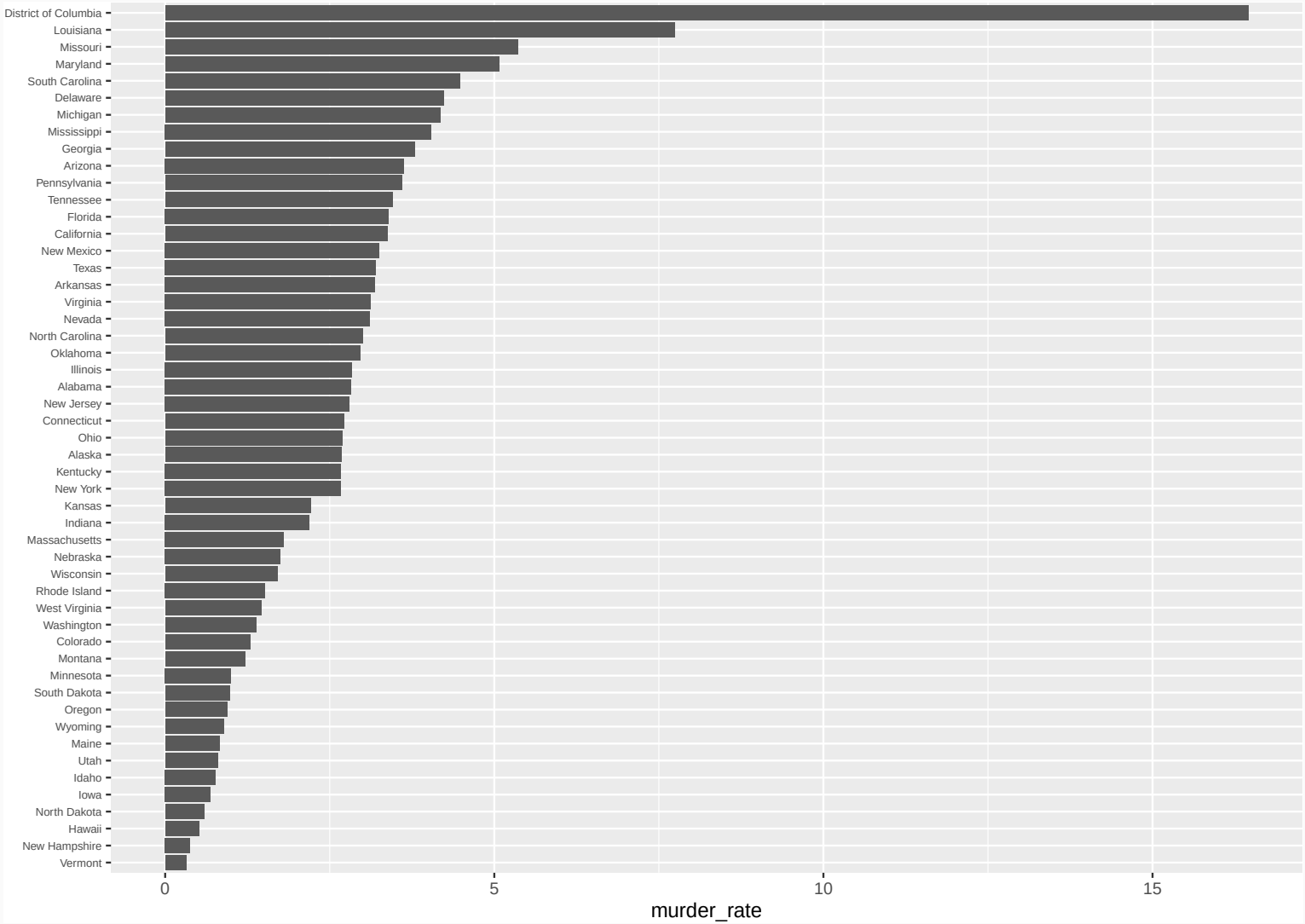
Ordering categories iii

- We can easily identify the state with the highest rate (DC), but it's more difficult to identify the state with the smallest rate.
    - Is it HI, NH, or VT?
- More crucially, it is hard to identify states that have similar rates.
- Instead, we can reorder the states according to the rate themselves.
    - We can use the function `reorder` to arrange US states according to murder rate

Ordering categories iv

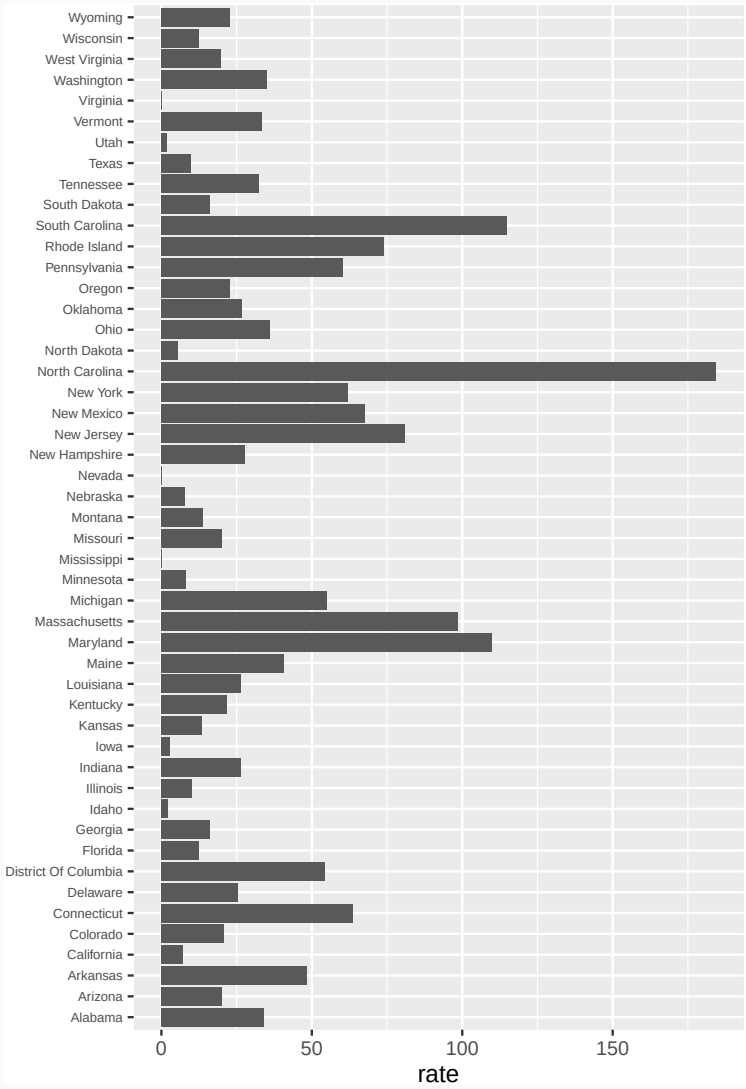
```
library(tidyverse)
library(dslabs)

data_murders %>%
  mutate(murder_rate = total / population * 100000) %>%
  mutate(state = reorder(state, murder_rate)) %>%
  ggplot(aes(state, murder_rate)) +
  geom_bar(stat = "identity") +
  coord_flip() + # For horizontal bars
  theme(axis.text.y = element_text(size = 6)) +
  xlab("")
```

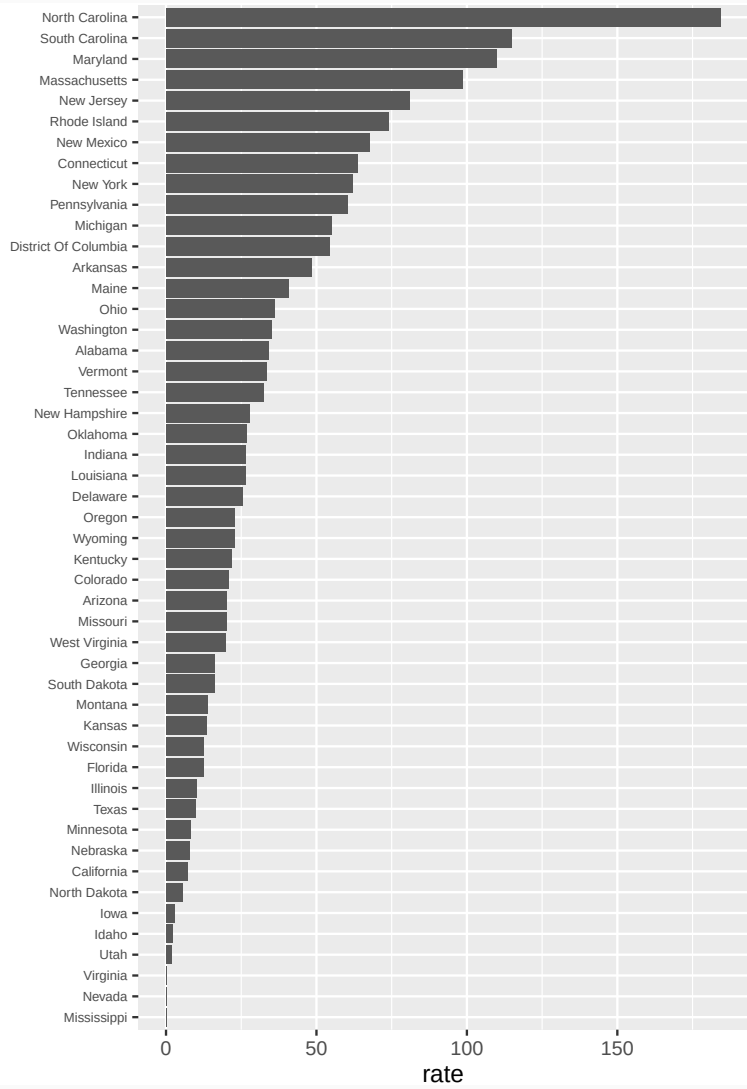
Ordering categories v



Exercise

Which is easier to interpret?





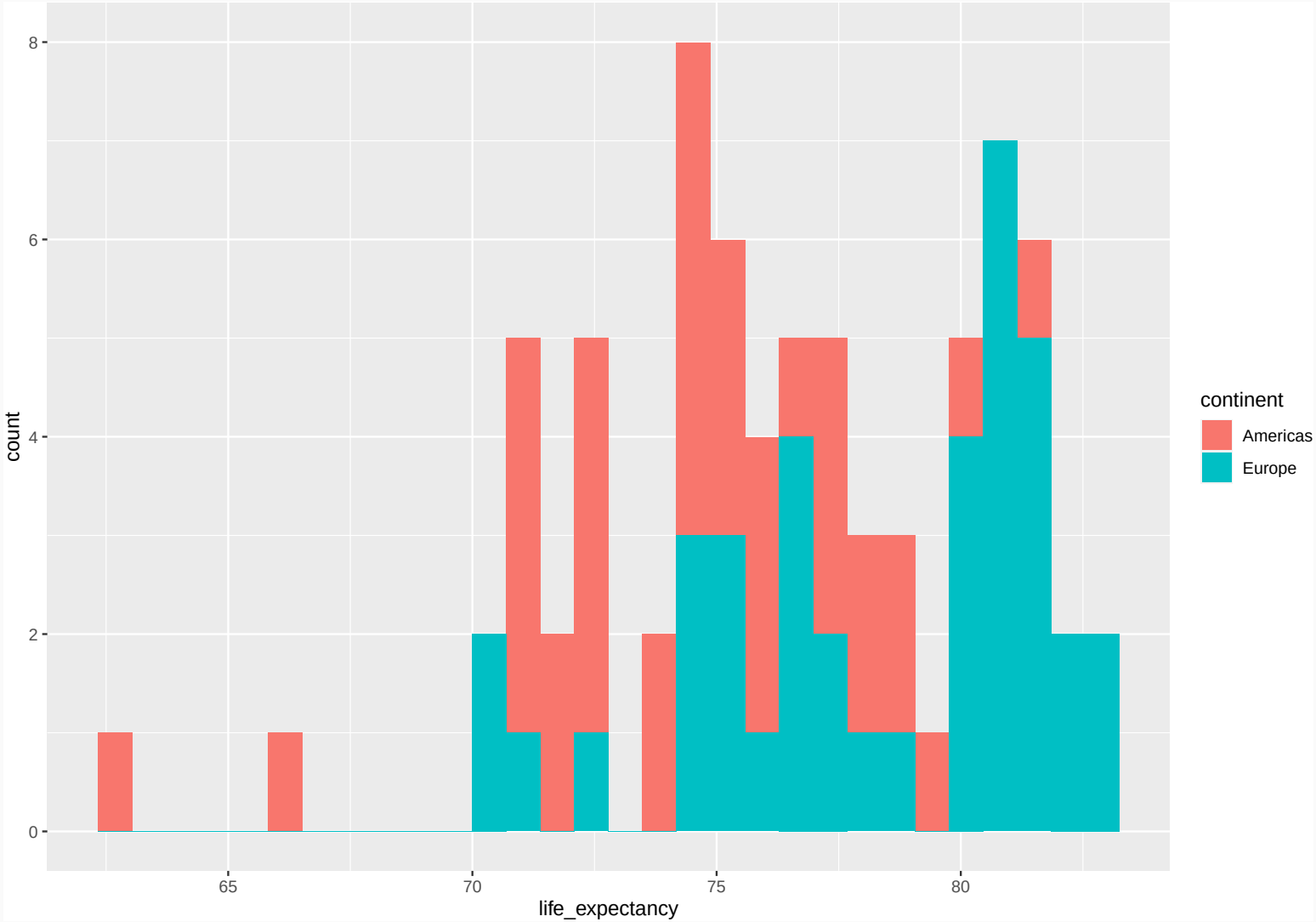
Easier comparisons i

- How you arrange your visual cues can also make comparisons easier.
- Let's say we want to compare life expectancy between Americas and Europe.
    - E.g. We can use histograms.

Easier comparisons ii

```
library(tidyverse)
library(dslabs)

gapminder %>%
  filter(year == 2012,
         continent %in% c("Americas", "Europe")) %>%
  ggplot(aes(life_expectancy,
             fill = continent)) +
  geom_histogram()
```

Easier comparisons iii

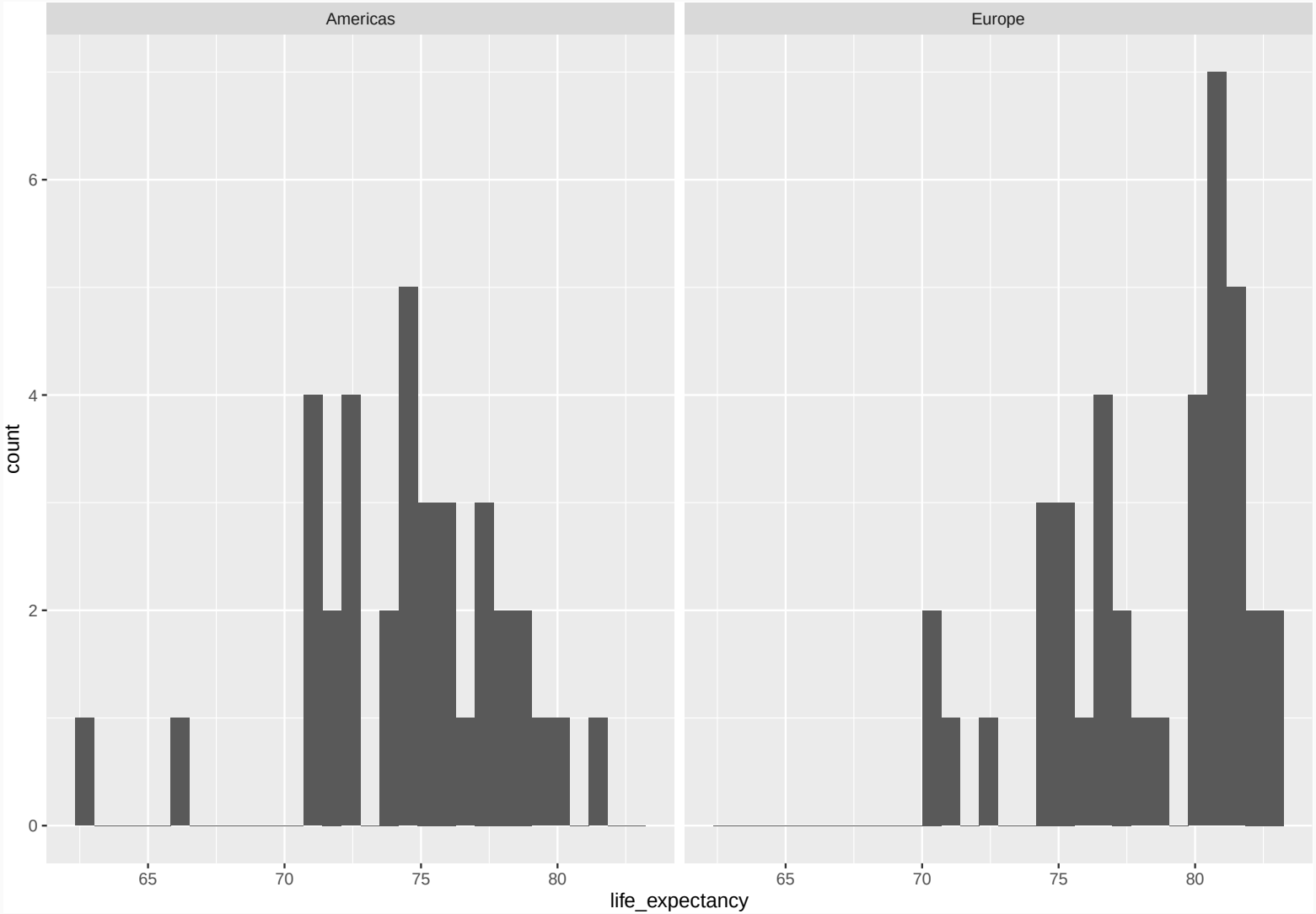


Easier comparisons iv

- Can you determine which continent has the highest average?
- Let's plot two separate histograms.

```
gapminder %>%
  filter(year == 2012,
         continent %in% c("Americas", "Europe")) %>%
  ggplot(aes(life_expectancy)) +
  geom_histogram() +
  facet_grid(~ continent)
```

Easier comparisons v

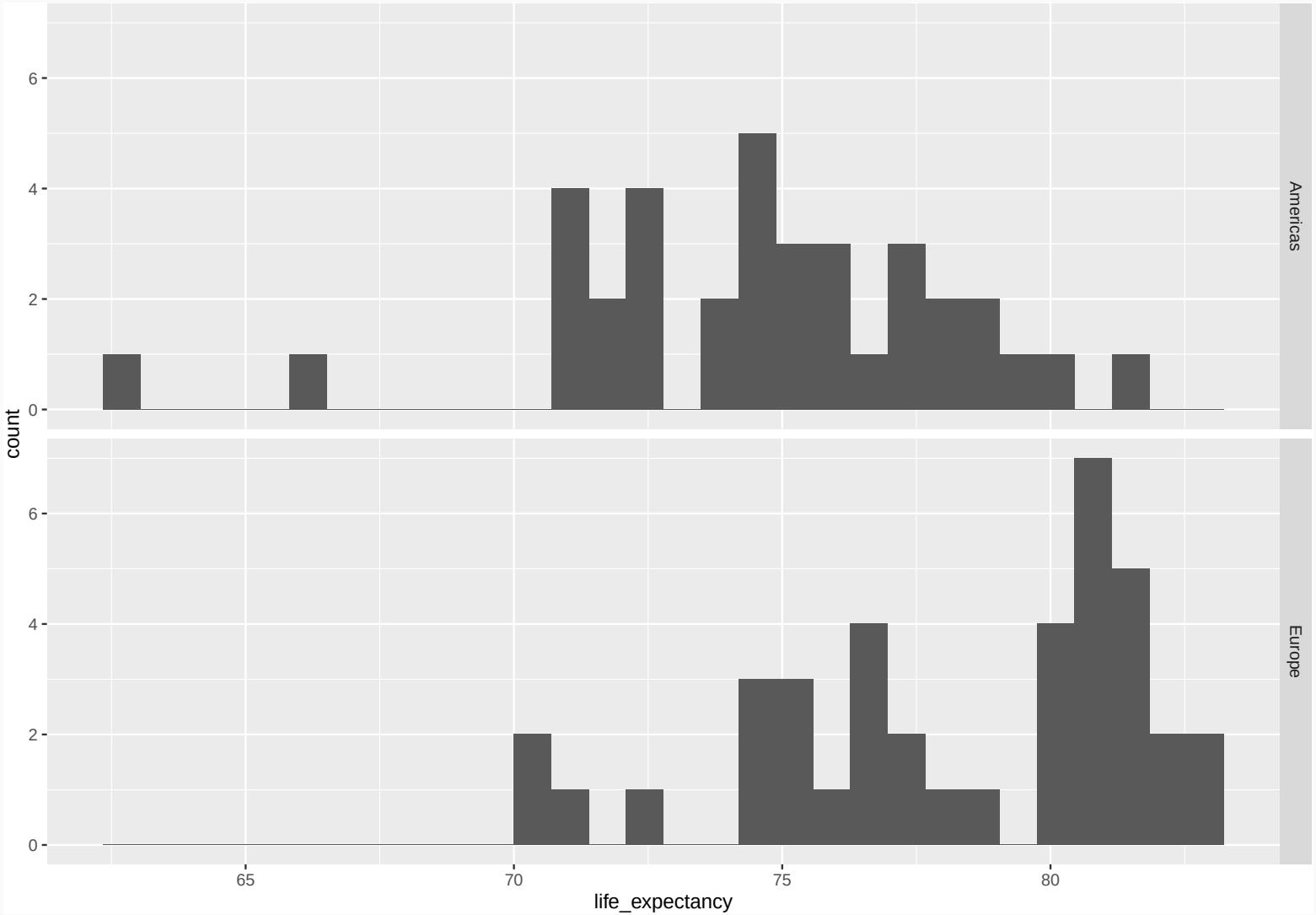


Easier comparisons vi

- Note: by default, `ggplot2` uses the same range for the axes to prevent distortions.
- But it is still difficult to determine which continent has the highest average...
- A better solution is to align the two histograms *vertically*.

```
gapminder %>%
  filter(year == 2012,
         continent %in% c("Americas", "Europe")) %>%
  ggplot(aes(life_expectancy)) +
  geom_histogram() +
  facet_grid(continent ~ .) # Note: which side of ~ is
```

Easier comparisons vii



Exercise

Actually, it may even be easier to see the difference if we used boxplots.

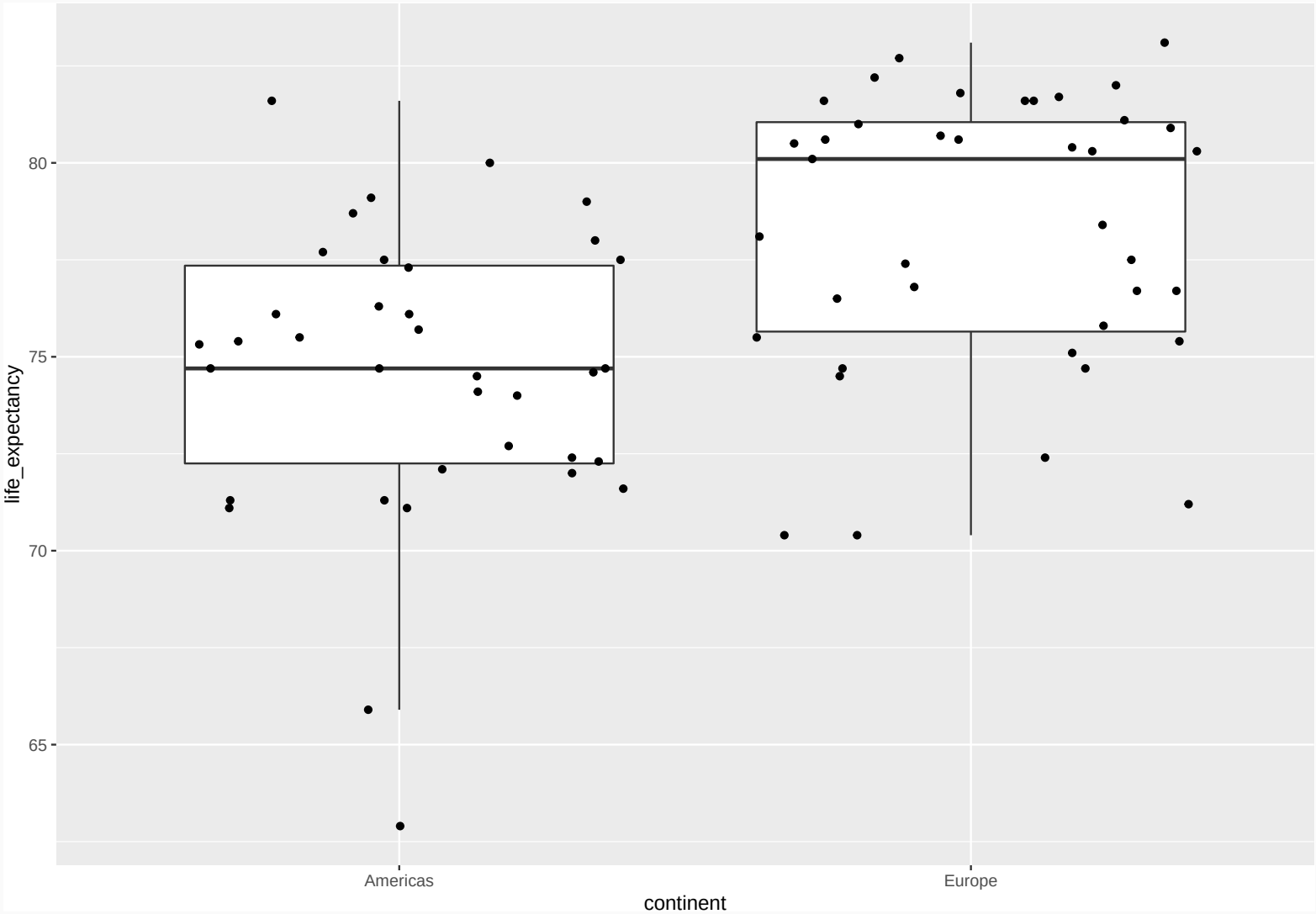
Create a plot with two side-by-side boxplots (one for Americas, one for Europe).

Bonus: add the data points on top!

Solution i

```
gapminder %>%
  filter(year == 2012,
         continent %in% c("Americas", "Europe")) %>%
  ggplot(aes(continent, life_expectancy)) +
  geom_boxplot(outlier.colour = NA) +
  geom_jitter(height = 0)
```

Solution ii



Summary

- Not all visual cues are created equal!
- Pie charts are almost always a bad idea.
- Include zero when length is an important visual cue.
- Order your categories in a meaningful way.
- When deciding how to arrange your visual cues on the graph, try to find the arrangement that makes the **important comparisons** as clear as possible.