Distributions

Week 7

AEM 2850: R for Business Analytics
Cornell Dyson
Spring 2022

Acknowledgements: Andrew Heiss, Claus Wilke

Announcements

This week we are back to our old routine:

- Today: slides
- Wednesday: reflection is due
- Thursday: work through an example
- Monday: next lab is due

First: discuss survey resonses and mini project 1

Questions before we get started?

Plan for today

Prologue

Mini Project 1

Proportions: a brief interlude

Distributions

Prologue

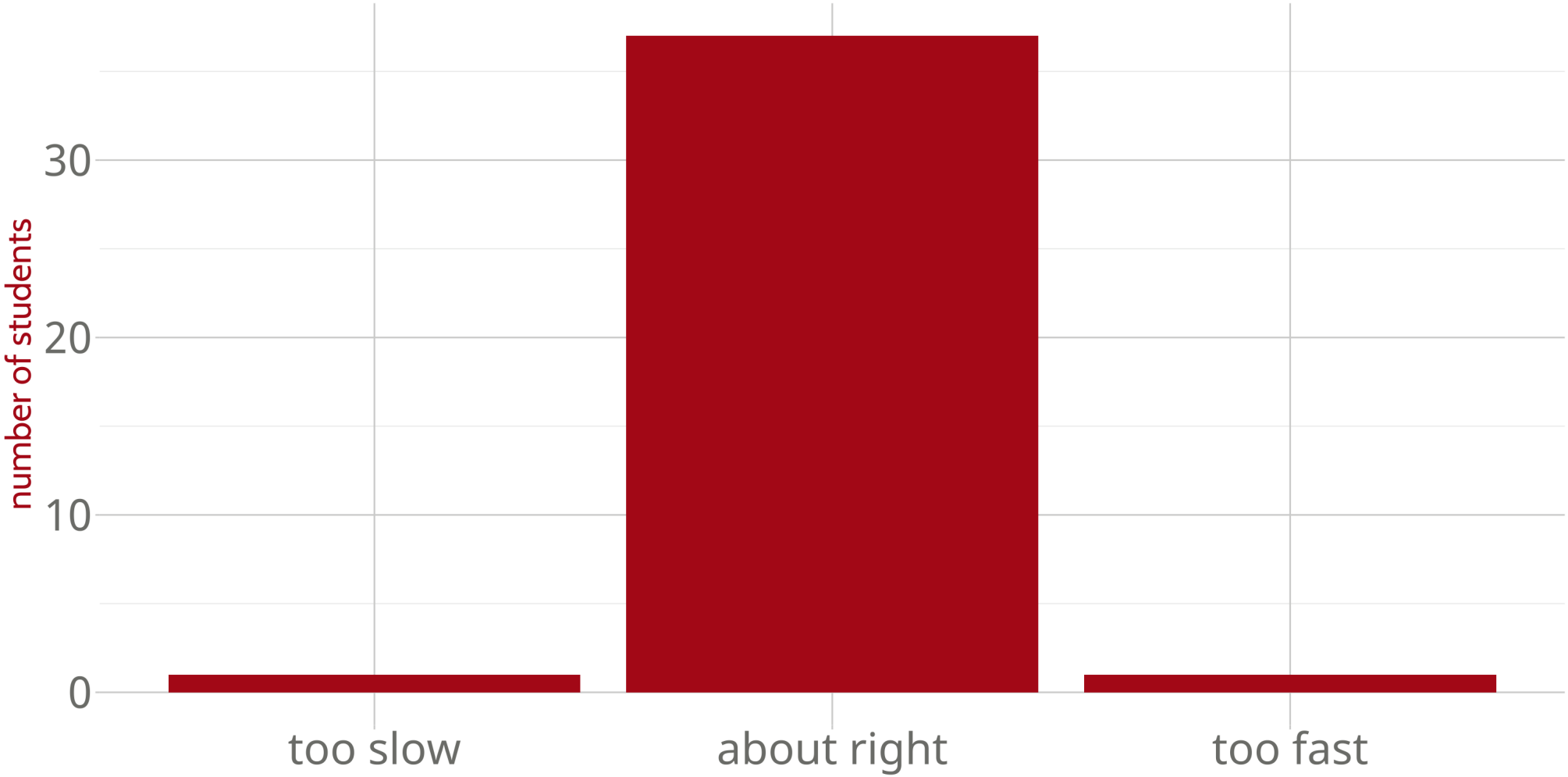
Thank you!

I read all the Reflection - Week 6 survey responses

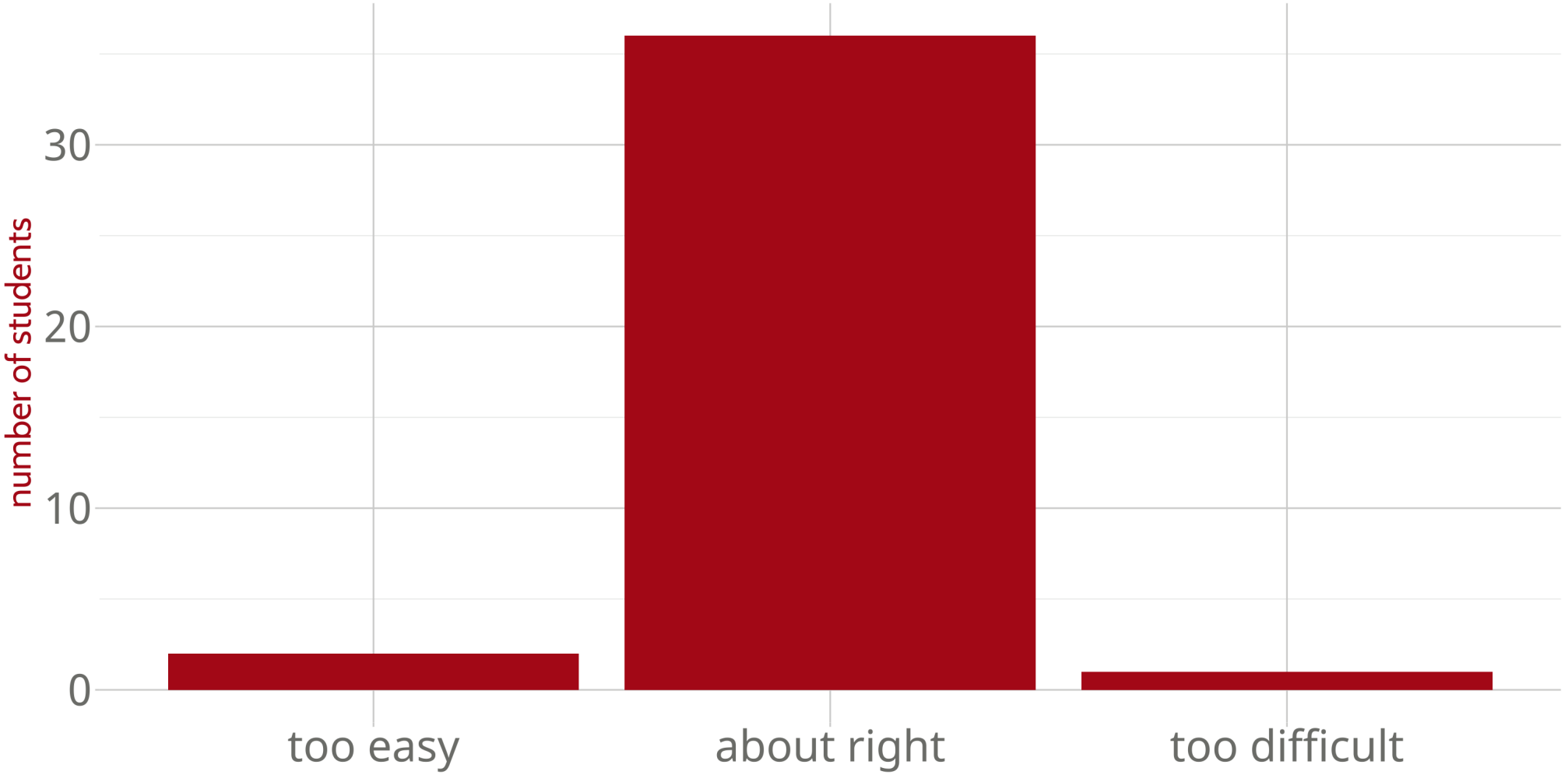
I will briefly summarize some of the responses and share some reactions

Please don't hesitate to provide feedback on how the course progresses!

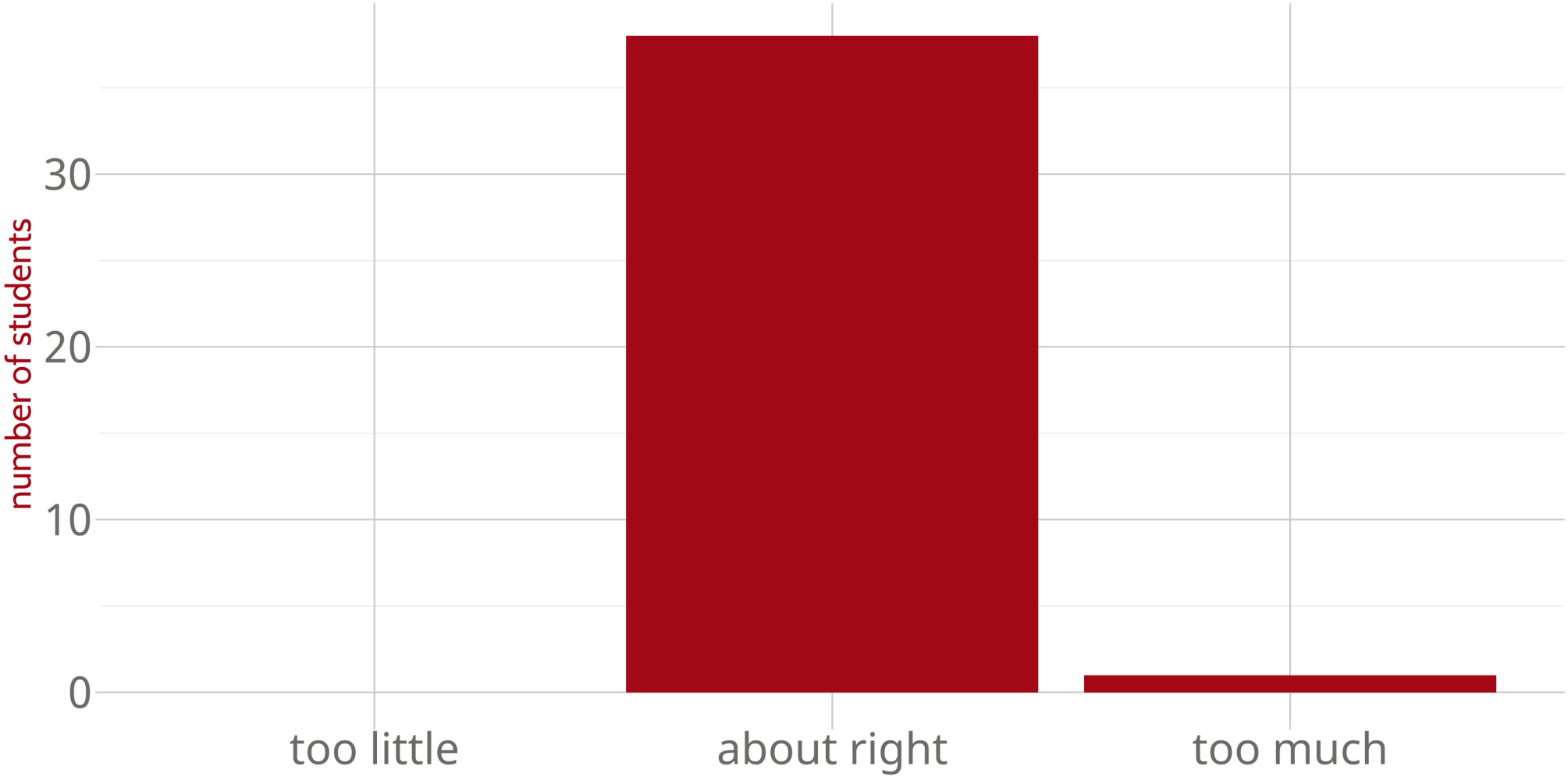
Please rate the pace overall:



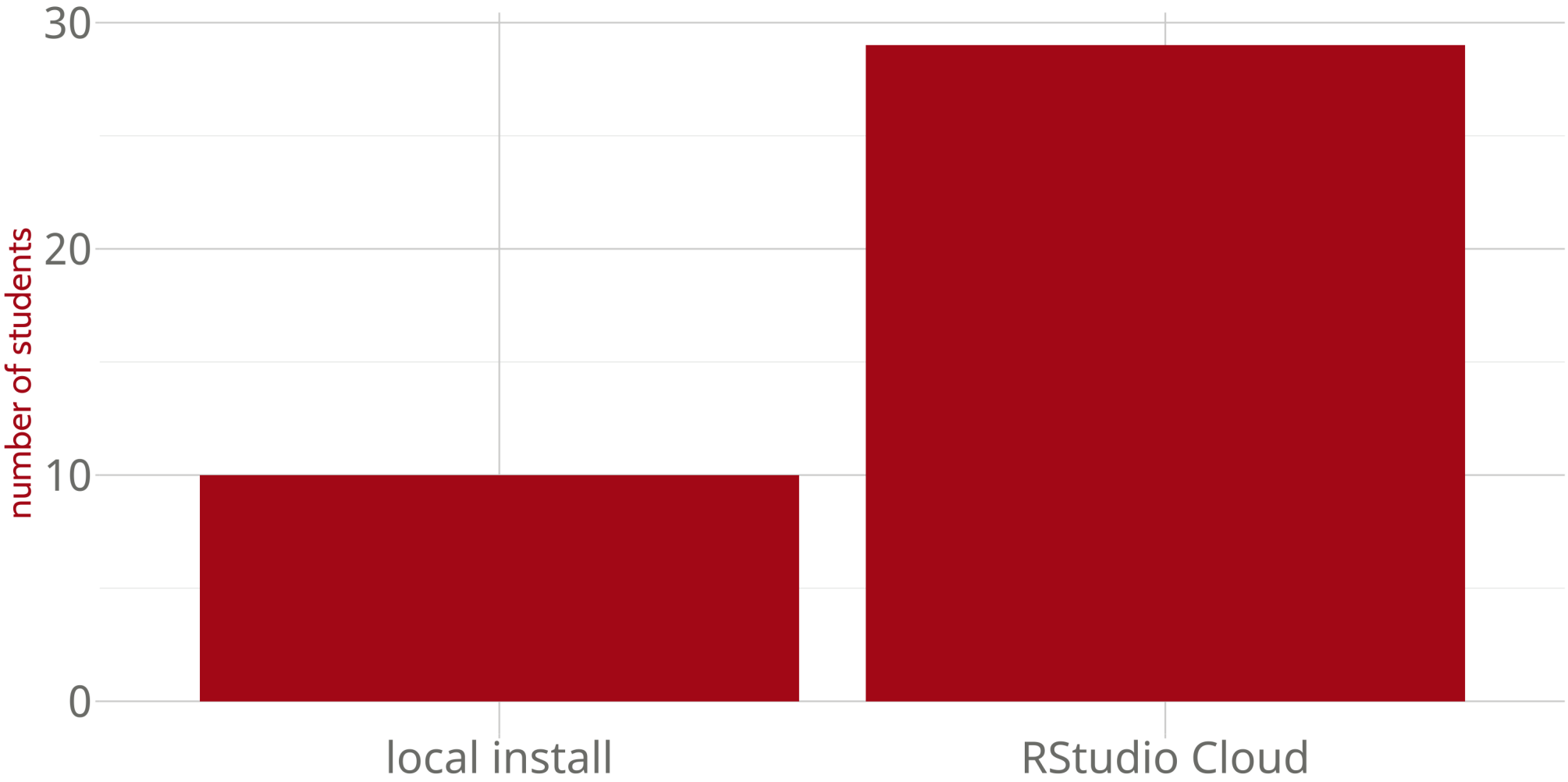
Please rate the difficulty overall:



Please rate the workload overall:



RStudio Cloud versus a local install?



Some themes:

Desire for more business applications

- We will work on this, especially on projects

Several people requested guidance on installing/working locally

- I will post some basic guidance on the course site
- Come to office hours if you want help installing R+RStudio!

Examples: make them more open-ended/challenging + work more as a group

- I will experiment with this!

Interest in modeling, statistical methods, linear regression, etc.

Mini Project 1

Mini Project 1

Use R and the tidyverse to wrangle and visualize equities data

Three parts:

1. AAPL
2. The S&P 500
3. Our Class Portfolio

Mini Project 1: data

```
sp500_companies
```

```
## # A tibble: 504 × 7
##    symbol company           identifier sedol  weight sector       local_currency
##    <chr>  <chr>             <chr>      <chr>   <dbl> <chr>        <chr>
##  1 AAPL   Apple Inc.        03783310   20462… 0.0701 Information… USD
##  2 MSFT   Microsoft Corpor… 59491810   25881… 0.0601 Information… USD
##  3 AMZN   Amazon.com Inc.   02313510   20000… 0.0349 Consumer Di… USD
##  4 GOOGL  Alphabet Inc. Cl… 02079K30   BYVY8… 0.0218 Communicati… USD
##  5 GOOG   Alphabet Inc. Cl… 02079K10   BYY88… 0.0203 Communicati… USD
##  6 TSLA   Tesla Inc         88160R10   B616C… 0.0185 Consumer Di… USD
##  7 BRK-B  Berkshire Hathaw… 08467070   20733… 0.0162 Financials   USD
##  8 NVDA   NVIDIA Corporati… 67066G10   23795… 0.0160 Information… USD
##  9 FB     Meta Platforms I… 30303M10   B7TL8… 0.0130 Communicati… USD
## 10 UNH    UnitedHealth Gro… 91324P10   29177… 0.0124 Health Care  USD
## # … with 494 more rows
```

Mini Project 1: data

```
sp500_prices
```

```
## # A tibble: 628,663 × 8
##    symbol date        open  high   low close    volume adjusted
##    <chr>  <date>     <dbl> <dbl> <dbl> <dbl>     <dbl>    <dbl>
##  1 AAPL   2017-01-03  29.0  29.1  28.7  29.0 115127600     27.3
##  2 AAPL   2017-01-04  29.0  29.1  28.9  29.0  84472400     27.3
##  3 AAPL   2017-01-05  29.0  29.2  29.0  29.2  88774400     27.4
##  4 AAPL   2017-01-06  29.2  29.5  29.1  29.5 127007600     27.7
##  5 AAPL   2017-01-09  29.5  29.9  29.5  29.7 134247600     28.0
##  6 AAPL   2017-01-10  29.7  29.8  29.6  29.8  97848400     28.0
##  7 AAPL   2017-01-11  29.7  30.0  29.6  29.9 110354400     28.1
##  8 AAPL   2017-01-12  29.7  29.8  29.6  29.8 108344800     28.0
##  9 AAPL   2017-01-13  29.8  29.9  29.7  29.8 104447600     28.0
## 10 AAPL   2017-01-17  29.6  30.1  29.6  30   137759200     28.2
## # … with 628,653 more rows
```

Mini Project 1: data

```
our_companies
```

```
## # A tibble: 22 × 2
##    name                                n
##    <chr>                           <dbl>
##  1 Allbirds Inc                        1
##  2 Alphabet Inc. Class A               1
##  3 Anheuser Busch Inbev SA             1
##  4 Apple Inc.                         11
##  5 Berkshire Hathaway Inc. Class B     1
##  6 Bumble Inc                          1
##  7 Capri Holdings Ltd                  1
##  8 Costco Wholesale Corporation        1
##  9 Electronic Arts Inc.                1
## 10 Levi Strauss & Co.                  1
## # … with 12 more rows
```

Mini Project 1: logistics

Work in groups of 3 (posted on canvas)

Write report in R Markdown that summarizes your work, presents visualizations, and discusses takeaways

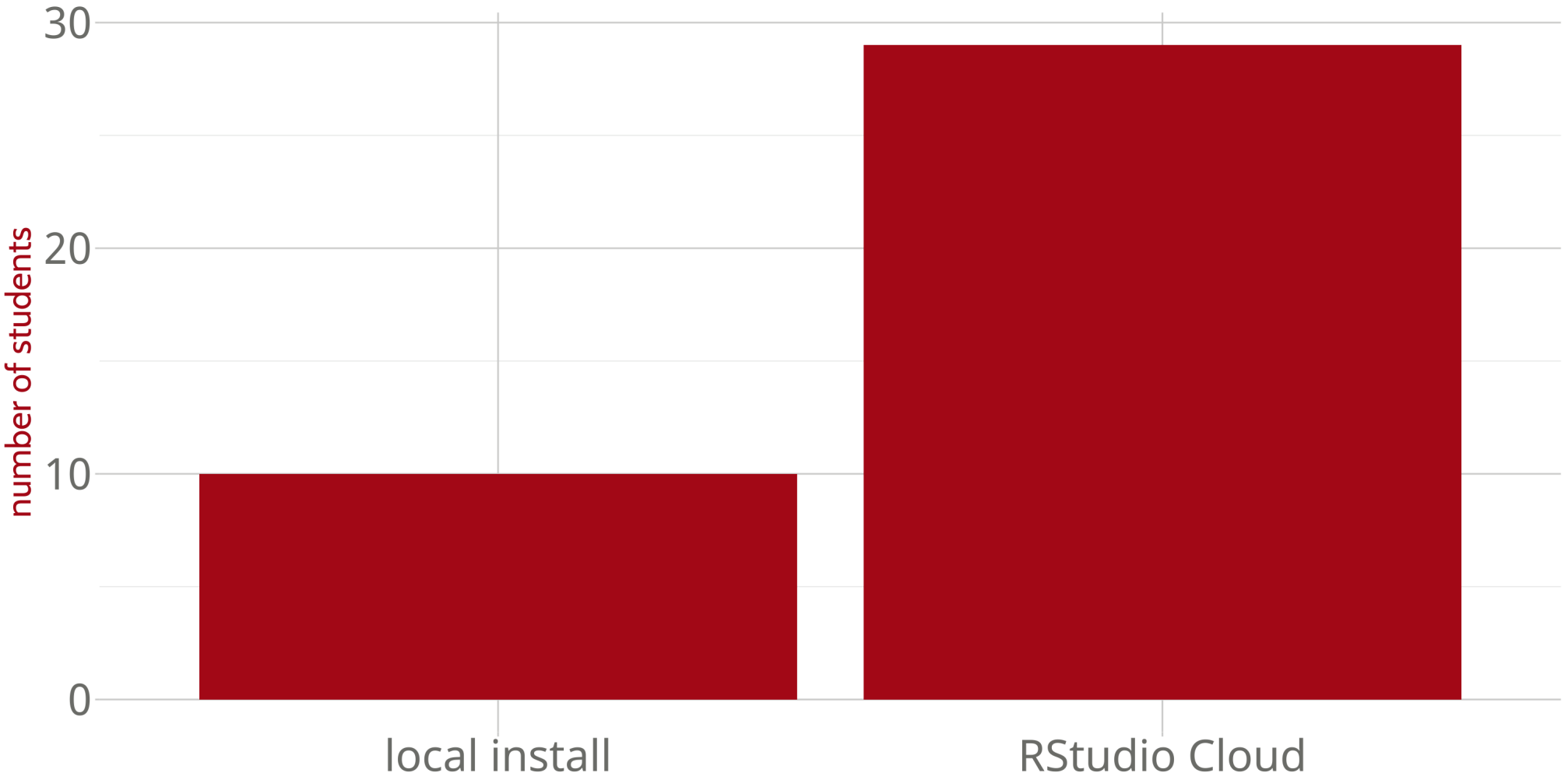
Do not use any packages outside base R and the tidyverse

No TA help for Part 3!

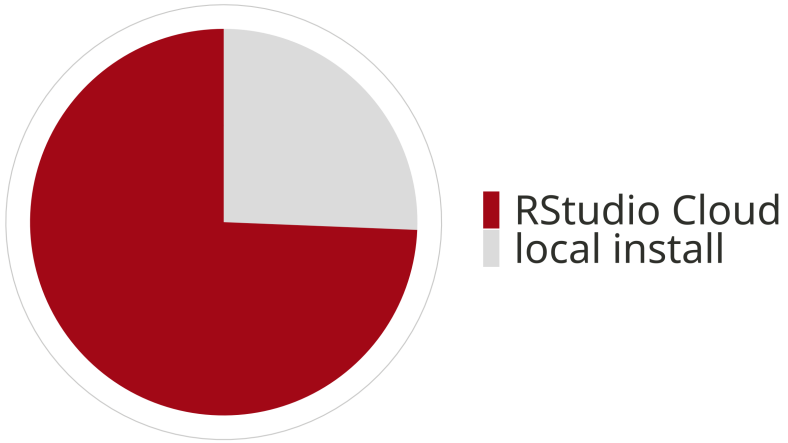
- If you try but still need help, come to my office hours

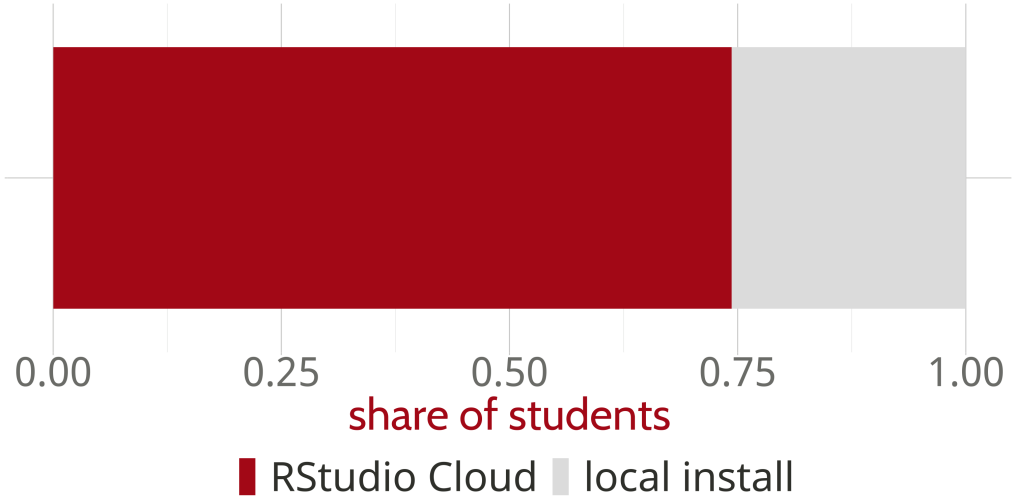
Proportions: a brief interlude

Can we improve this survey visualization?



Proportions





Pros and cons of different approaches

	Pie chart	Stacked bars	Side-by-side bars
Allows easy comparison of relative proportions	✖	✖	✔

Pros and cons of different approaches

	Pie chart	Stacked bars	Side-by-side bars
Allows easy comparison of relative proportions	✖	✖	✔
Shows data as proportions of a whole	✔	✔	✖

Pros and cons of different approaches

	Pie chart	Stacked bars	Side-by-side bars
Allows easy comparison of relative proportions	✖	✖	✔
Shows data as proportions of a whole	✔	✔	✖
Emphasizes simple fractions (1/2, 1/3, ...)	✔	✖	✖

Pros and cons of different approaches

	Pie chart	Stacked bars	Side-by-side bars
Allows easy comparison of relative proportions	✖	✖	✔
Shows data as proportions of a whole	✔	✔	✖
Emphasizes simple fractions (1/2, 1/3, ...)	✔	✖	✖
Visually appealing for small datasets	✔	✖	✔

Pros and cons of different approaches

	Pie chart	Stacked bars	Side-by-side bars
Allows easy comparison of relative proportions	✖	✖	✔
Shows data as proportions of a whole	✔	✔	✖
Emphasizes simple fractions (1/2, 1/3, ...)	✔	✖	✖
Visually appealing for small datasets	✔	✖	✔
Works well for a large number of subsets	✖	✖	✔

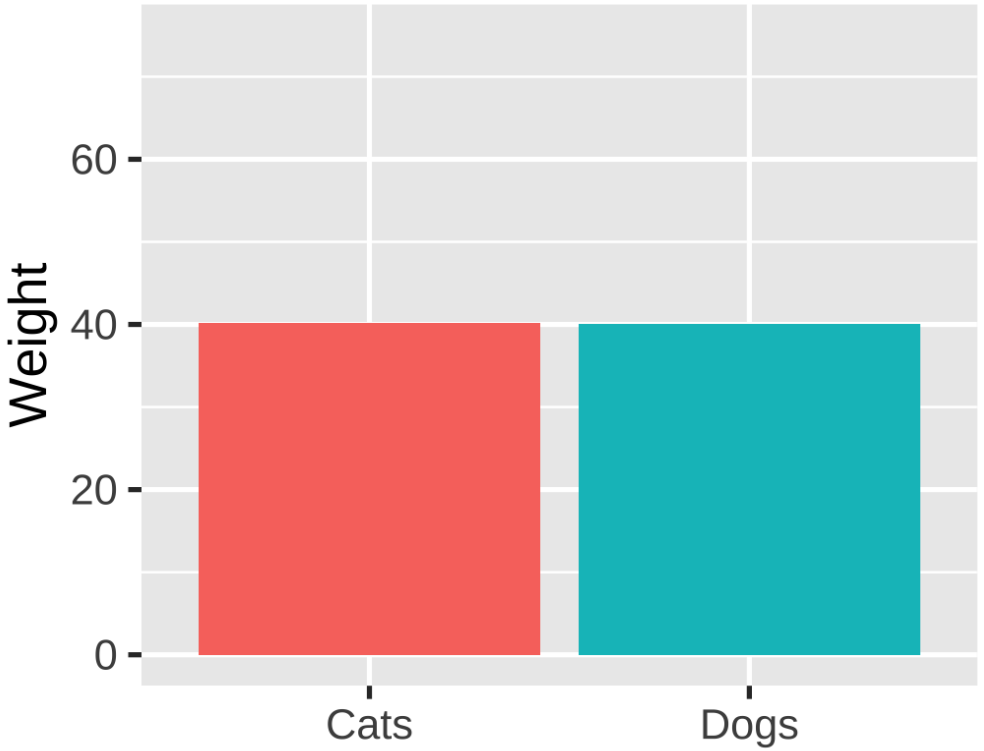
Pros and cons of different approaches

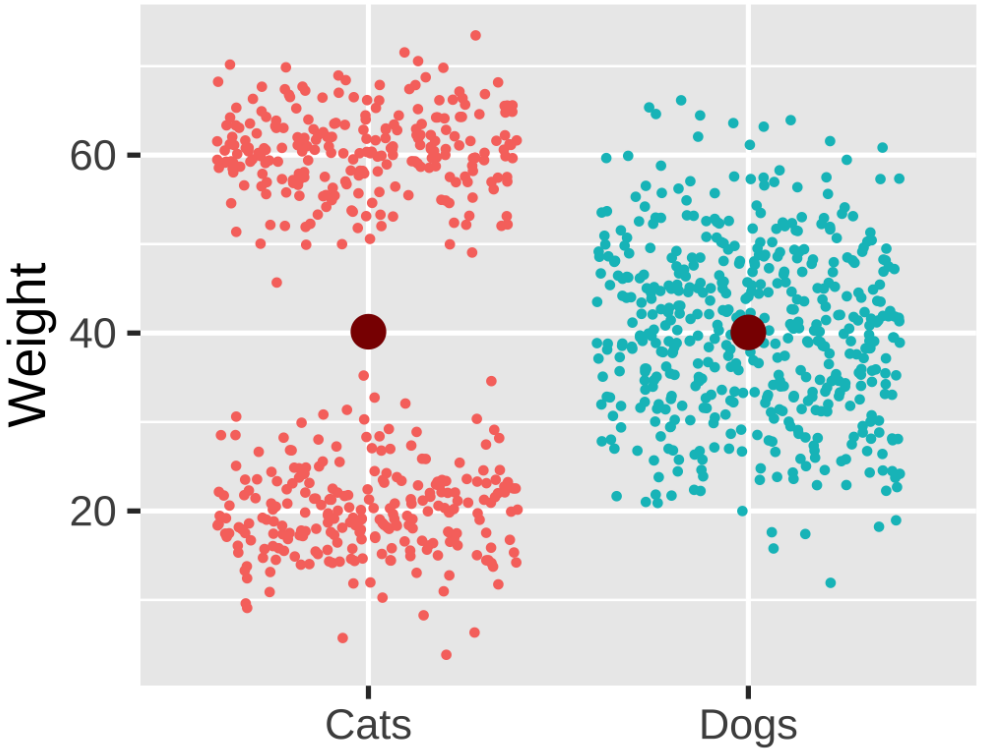
	Pie chart	Stacked bars	Side-by-side bars
Allows easy comparison of relative proportions	✖	✖	✔
Shows data as proportions of a whole	✔	✔	✖
Emphasizes simple fractions (1/2, 1/3, ...)	✔	✖	✖
Visually appealing for small datasets	✔	✖	✔
Works well for a large number of subsets	✖	✖	✔
Works well for time series and similar	✖	✔	✖

No one visualization fits all scenarios!

Distributions

Problems with single numbers





More information is (almost) always better

Avoid visualizing single numbers when you have a whole range or distribution of numbers

Uncertainty in single variables

Uncertainty across multiple variables

Uncertainty in models and simulations

Histograms

What are they?

Put data into equally spaced buckets (or "bins"), plot how many rows are in each bucket

Histograms

How would we use the grammar of graphics to make a histogram of `lifeExp`?

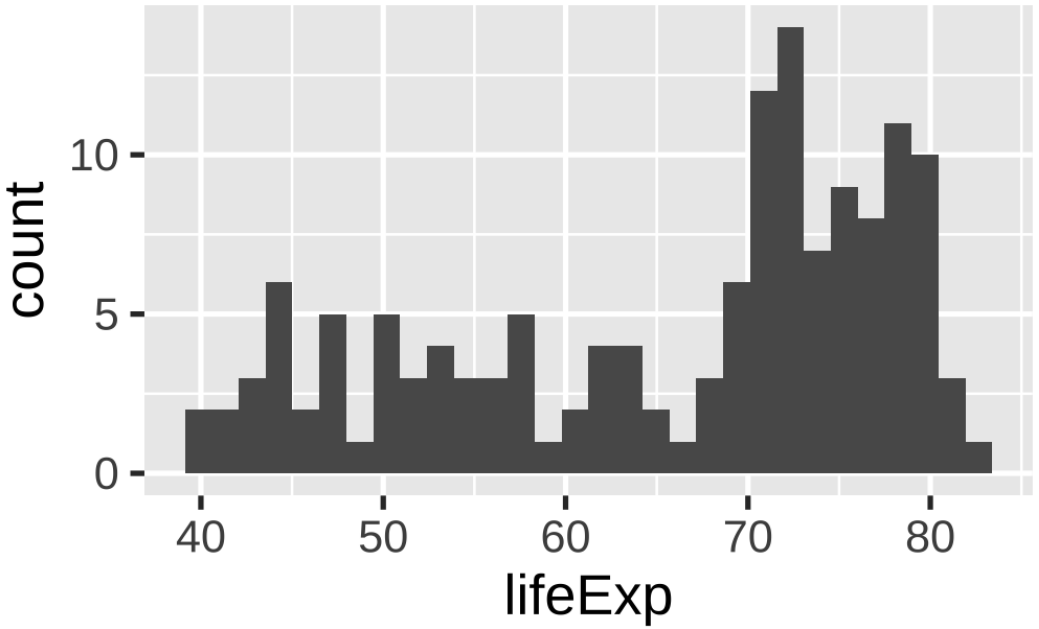
```
library(gapminder)

gapminder_2002 <- gapminder %>%
  filter(year == 2002)

head(gapminder_2002)
```

```
## # A tibble: 6 × 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       2002    42.1 25268405      727.
## 2 Albania     Europe     2002    75.7  3508512     4604.
## 3 Algeria     Africa     2002    71.0 31287142     5288.
## 4 Angola      Africa     2002    41.0 10866106     2773.
## 5 Argentina   Americas   2002    74.3 38331121     8798.
## 6 Australia   Oceania    2002    80.4 19546792    30688.
```

Histograms

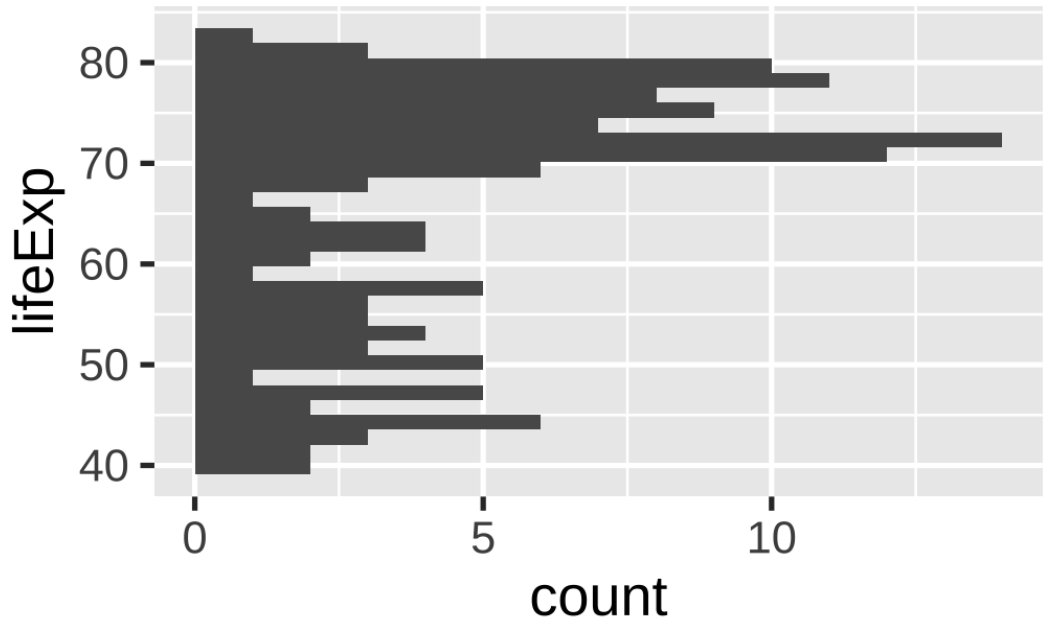
```
ggplot(gapminder_2002,
       aes(x = lifeExp)) +
  geom_histogram()
```

What if we mapped lifeExp to y?



Histograms

```
ggplot(gapminder_2002,
       aes(y = lifeExp)) +
  geom_histogram()
```

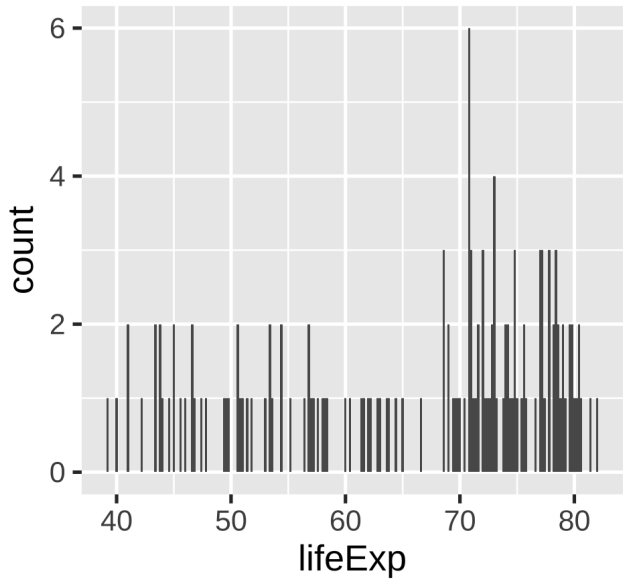


Histograms: bin width

No official rule for what makes a good bin width

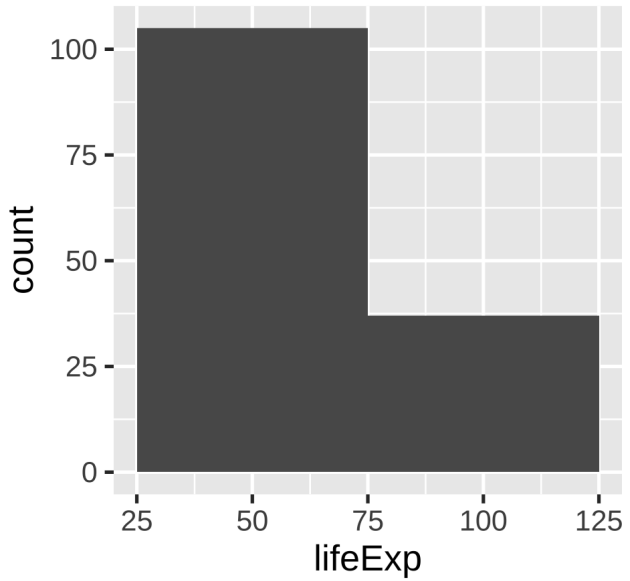
Too narrow:

```
geom_histogram(binwidth = .2)
```



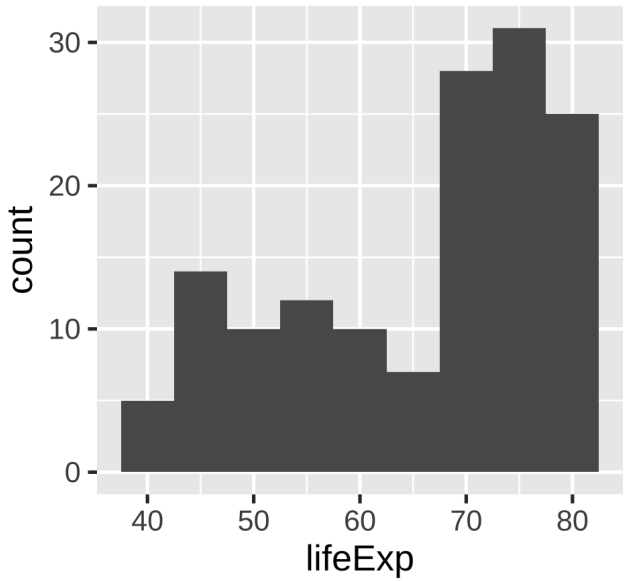
Too wide:

```
geom_histogram(binwidth = 50)
```



(One type of) just right:

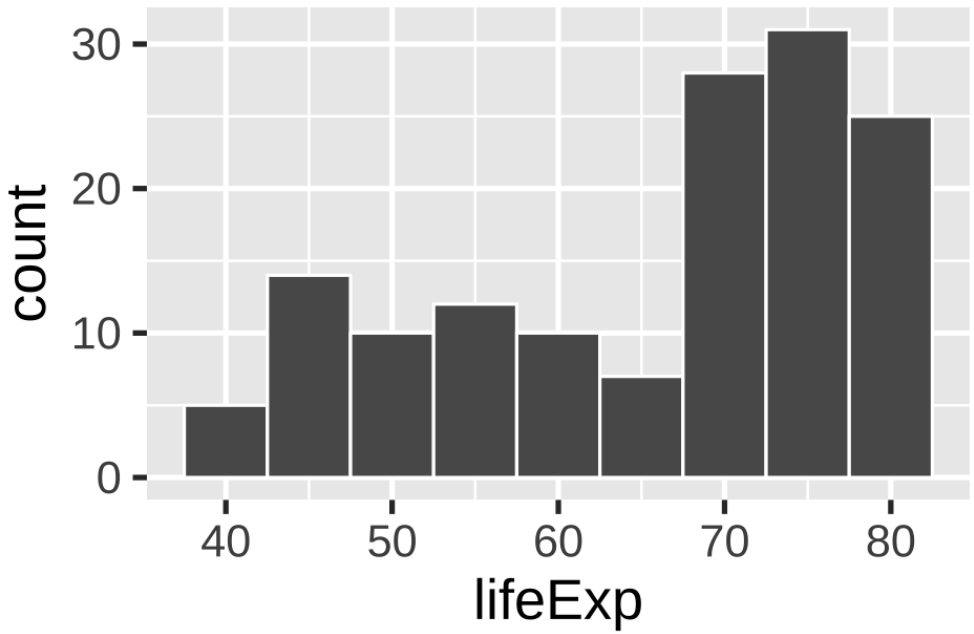
```
geom_histogram(binwidth = 5)
```



Histogram tips

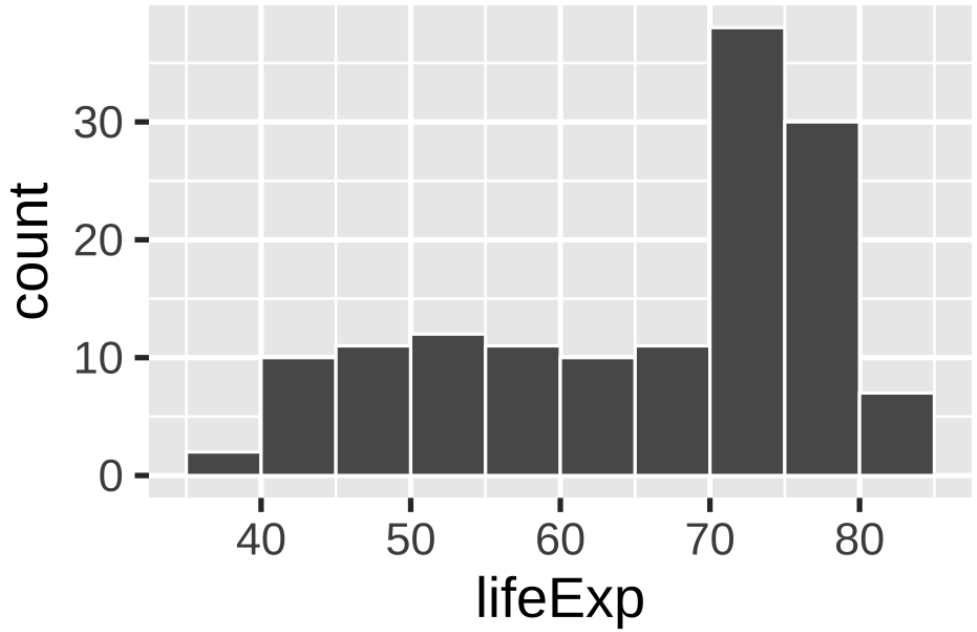
Add a border to the bars
for readability

```
geom_histogram(..., color = "white")
```



Set the boundary;
bucket now 50–55, not 47.5–52.5

```
geom_histogram(..., boundary = 50)
```



Density plots

What are they?

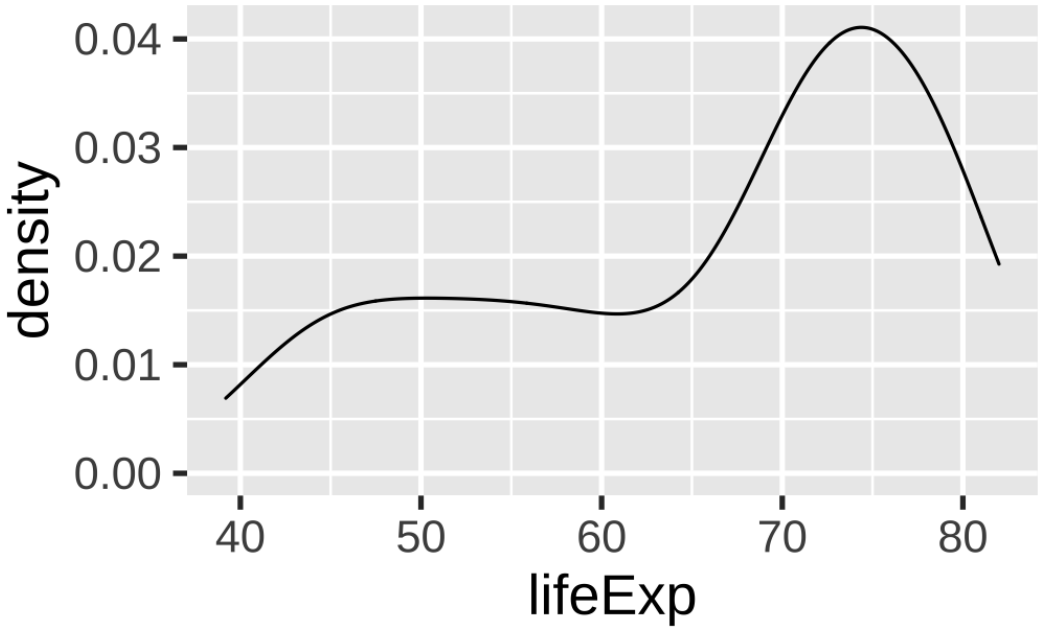
Estimates of the **probability *density* function** of a random variable

Histograms show raw counts; density plots show proportions (integrate to 1)

How would we use the grammar of graphics to make a density plot of `lifeExp`?

Density plots

```
ggplot(gapminder_2002,
       aes(x = lifeExp)) +
  geom_density()
```

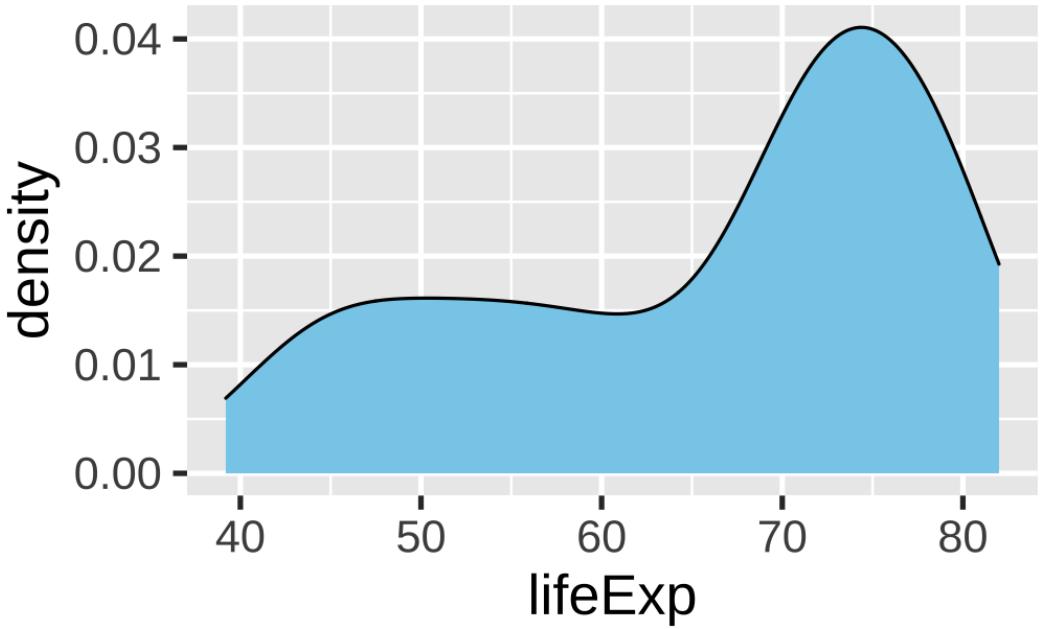


Density plots: add some color

```
ggplot(gapminder_2002,
       aes(x = lifeExp)) +
  geom_density(fill = "skyblue")
```

Reminder: we can use aesthetics as parameters inside a geom rather than inside an **aes()** statement

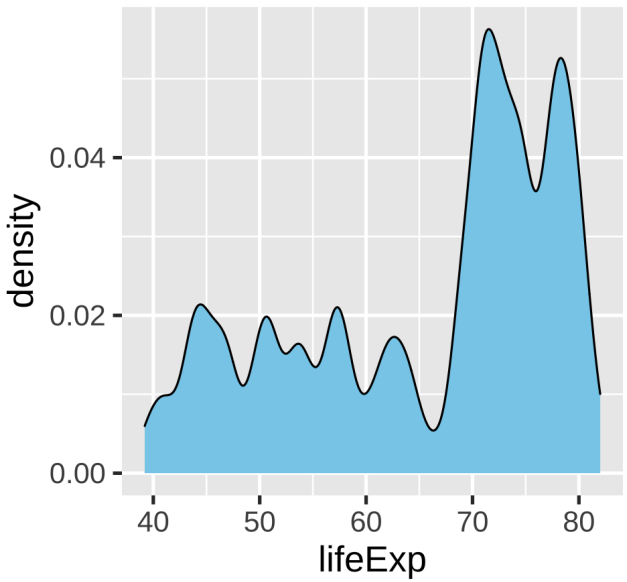
Here we used **fill = "skyblue"**



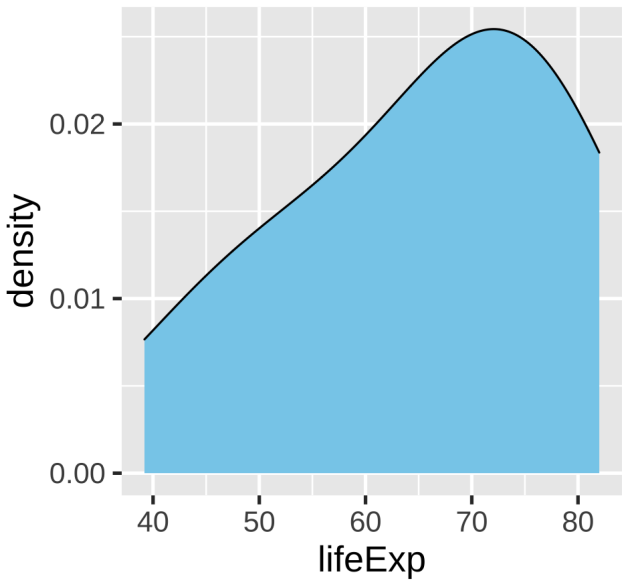
Density plots: bandwidths

Different options for calculus change the plot shape

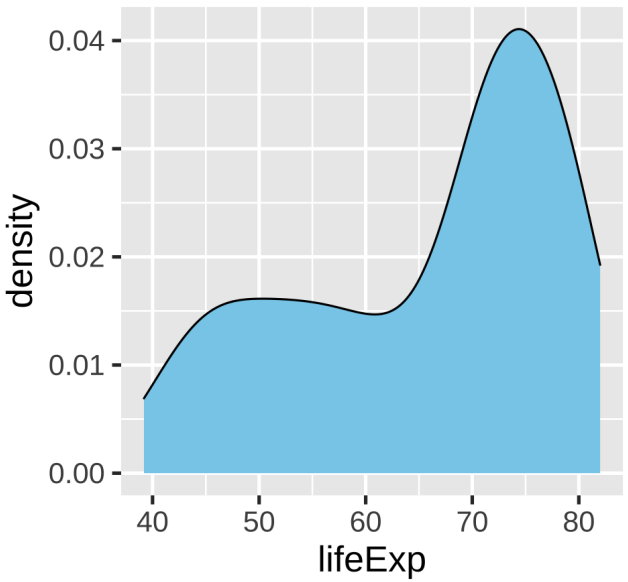
`bw = 1`



`bw = 10`



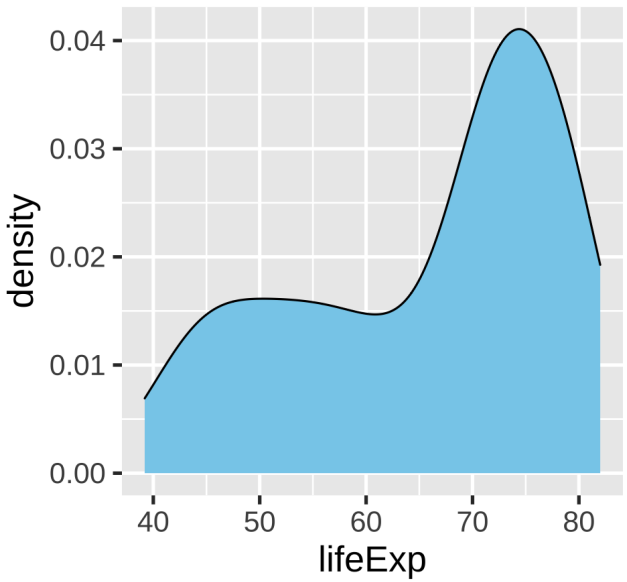
`bw = "nrd0"` (default)



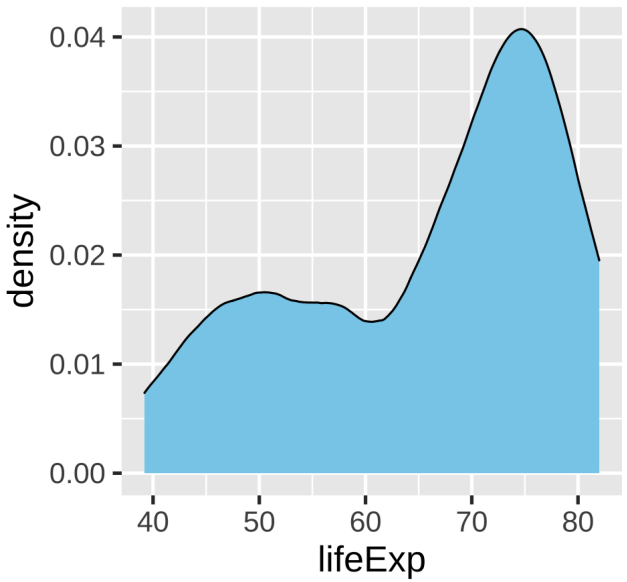
Density plots: kernels

Different options for calculus change the plot shape

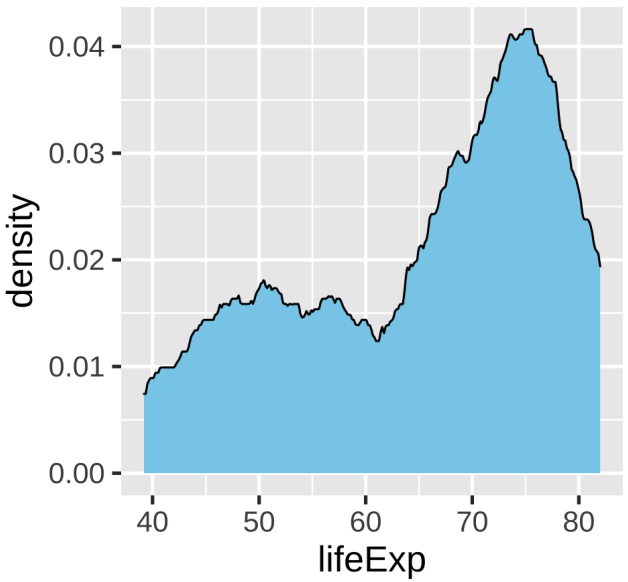
`kernel = "gaussian"`



`"epanechnikov"`



`"rectangular"`

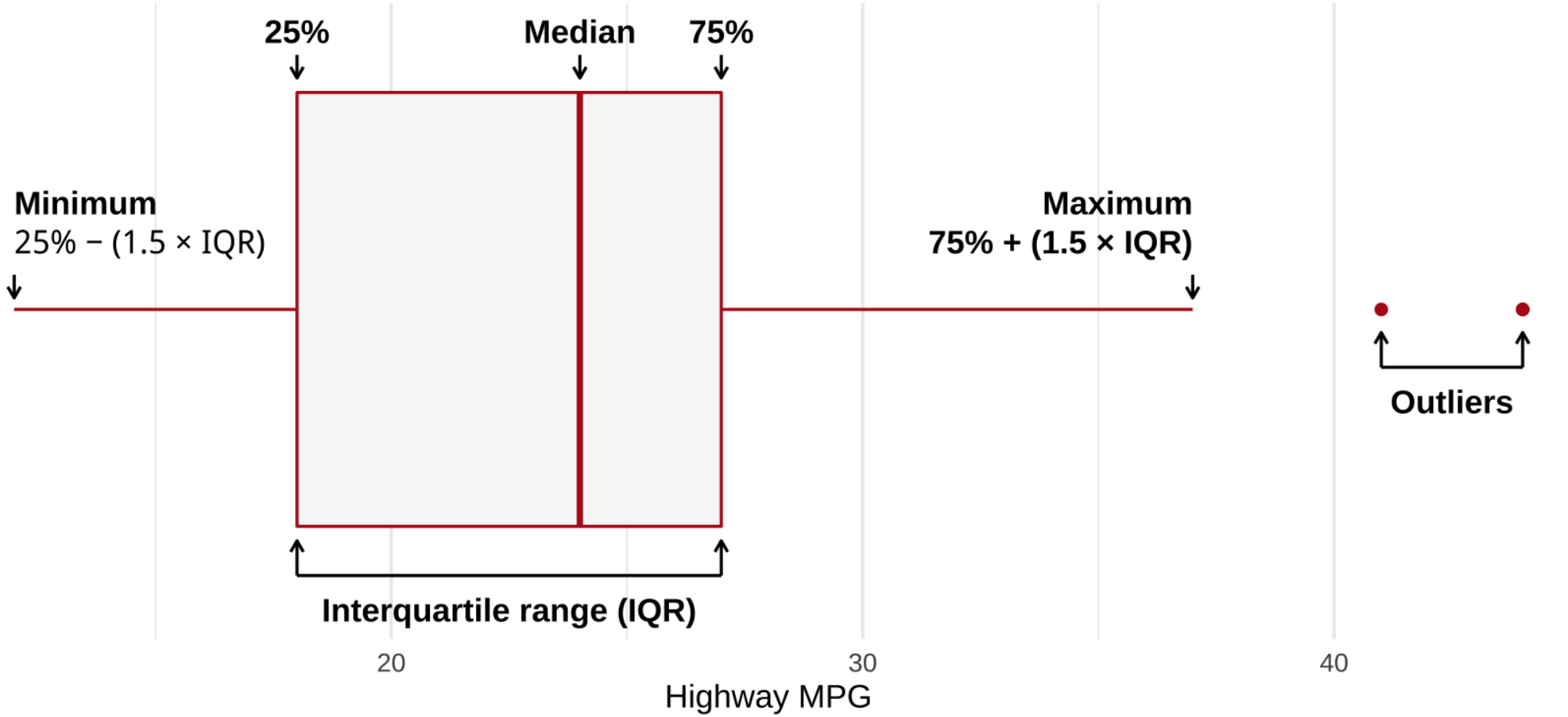


Box plots

What are they?

Graphical representations of specific points in a distribution

Box plots


25%
Median
75%
Minimum
25% − (1.5 × IQR)
Maximum
75% + (1.5 × IQR)
Outliers
Interquartile range (IQR)
20
30
40
Highway MPG

Box plots

What are they?

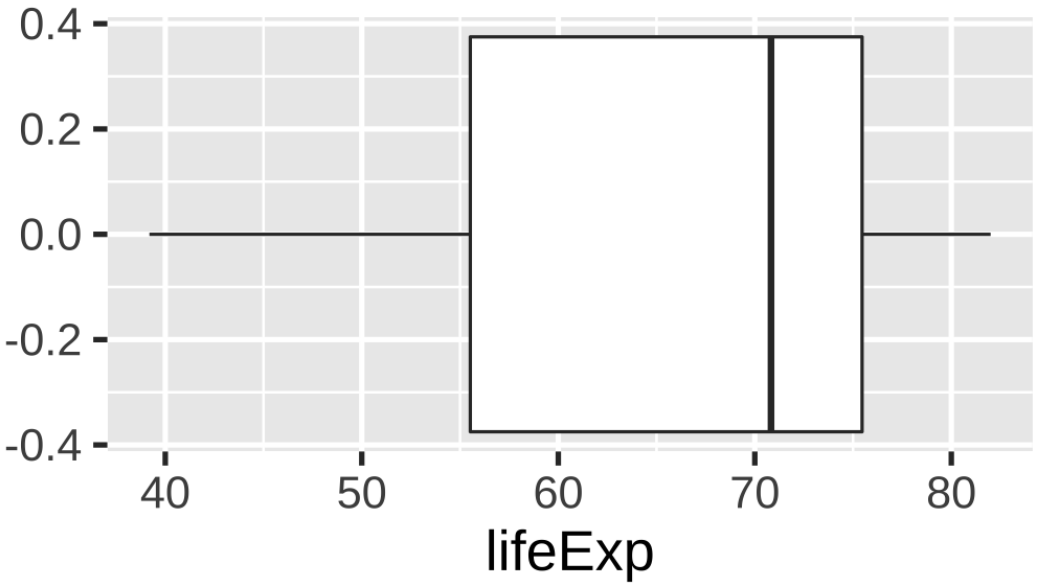
Show specific points in a distribution

How would we use the grammar of graphics to make a boxplot of `lifeExp`?

Box plots

```
ggplot(gapminder_2002,
       aes(x = lifeExp)) +
  geom_boxplot()
```

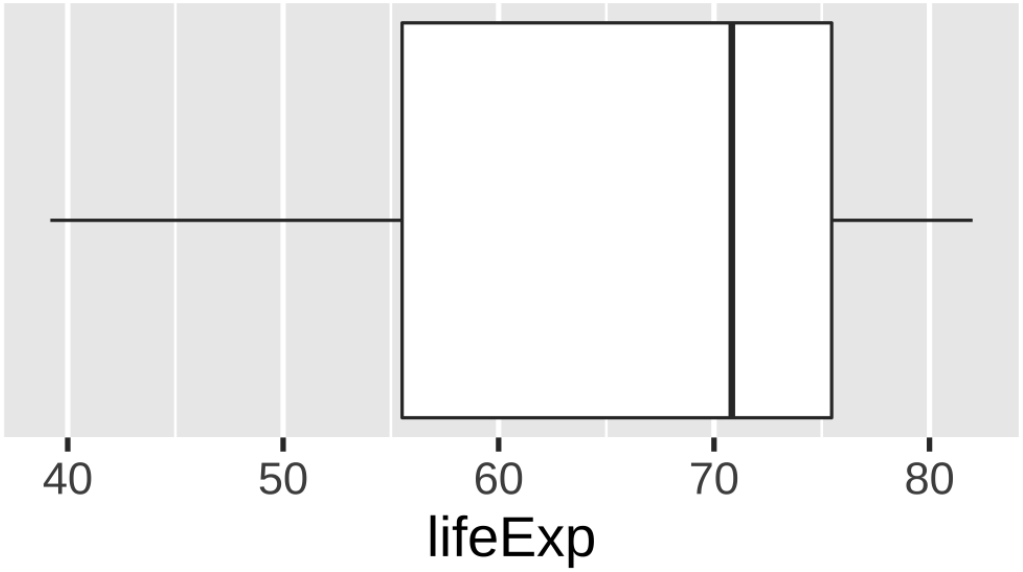
What do the y axis numbers mean?



Box plots

Use `theme()` to customize the plot for this geom

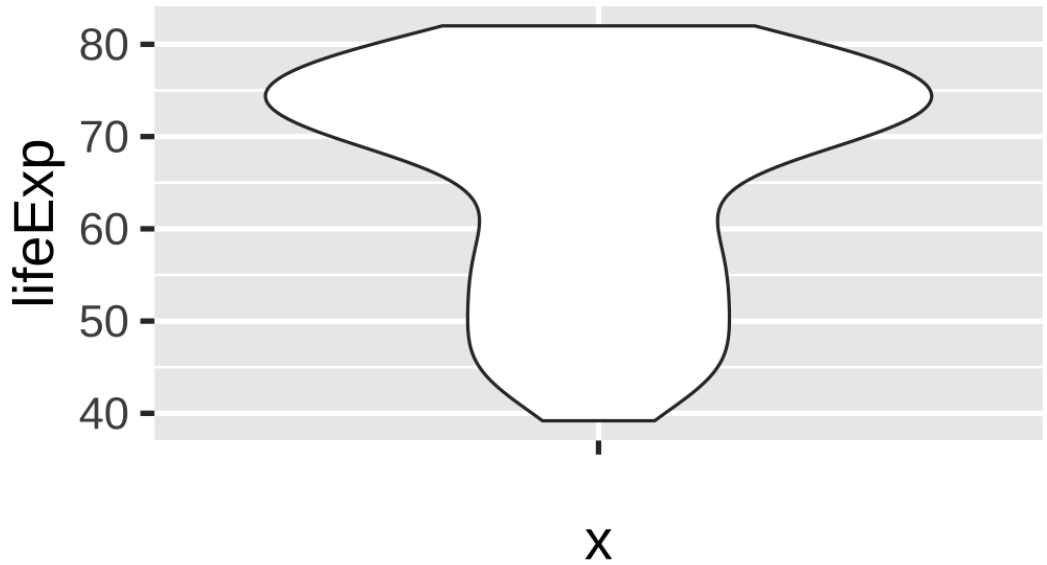
```
ggplot(gapminder_2002,
       aes(x = lifeExp)) +
  geom_boxplot() +
theme(axis.text.y = element_blank(),
      axis.ticks.y = element_blank(),
      panel.grid.major.y = element_blank(),
      panel.grid.minor.y = element_blank())
```



Violin plots

Mirror density plot and flip

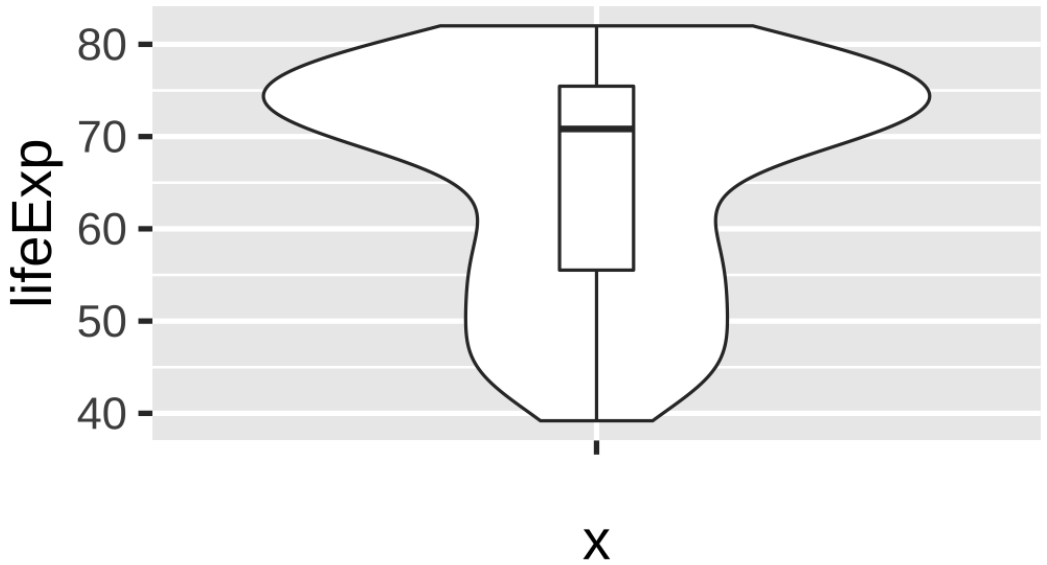
```
ggplot(gapminder_2002,
       aes(x = "",
           y = lifeExp)) +
  geom_violin()
```



Overalying geometries

We can overlay multiple geometries to provide more information

```
ggplot(gapminder_2002,
       aes(x = "",
           y = lifeExp)) +
  geom_violin() +
  geom_boxplot(width = 0.1)
```



Uncertainty across multiple variables

How could we visualize the distribution of a single variable across groups?

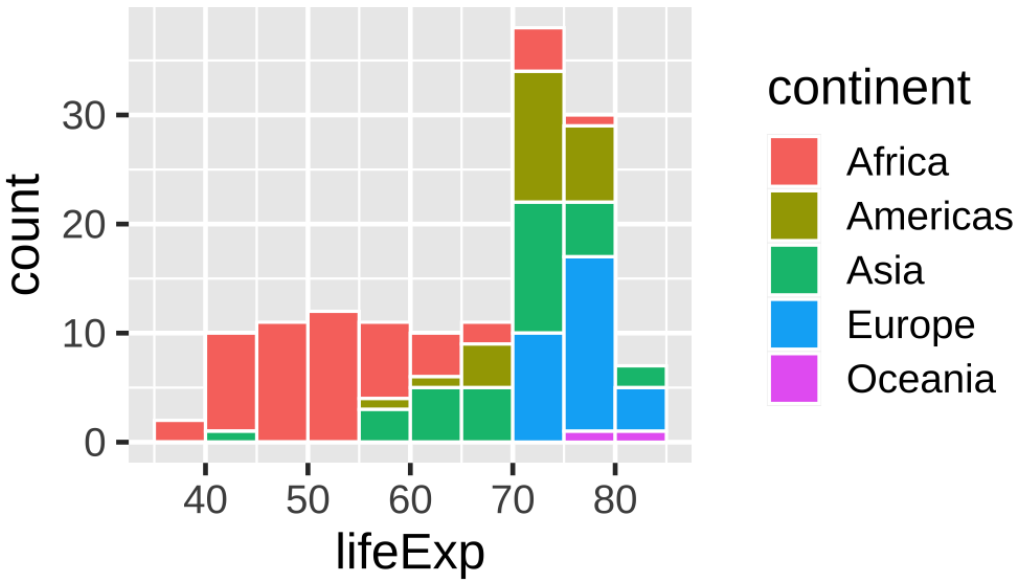
Add a `fill` aesthetic or use faceting!

Multiple histograms

Fill with a different variable

This is bad and really hard to read though

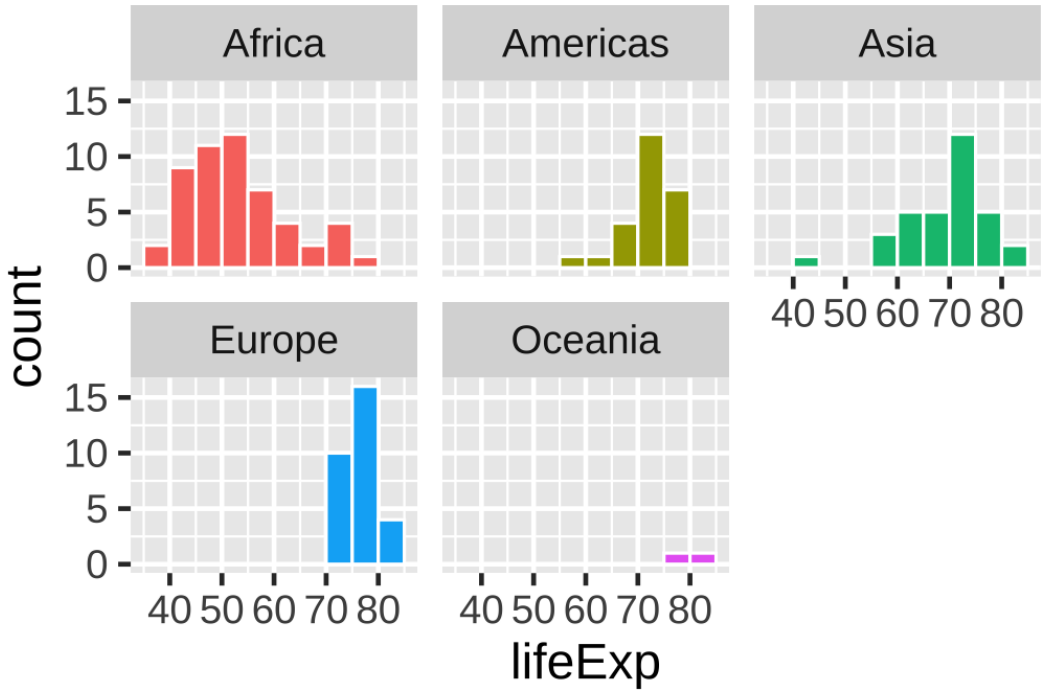
```
ggplot(gapminder_2002,
       aes(x = lifeExp,
           fill = continent)) +
  geom_histogram(binwidth = 5,
                 color = "white",
                 boundary = 50)
```



Multiple histograms

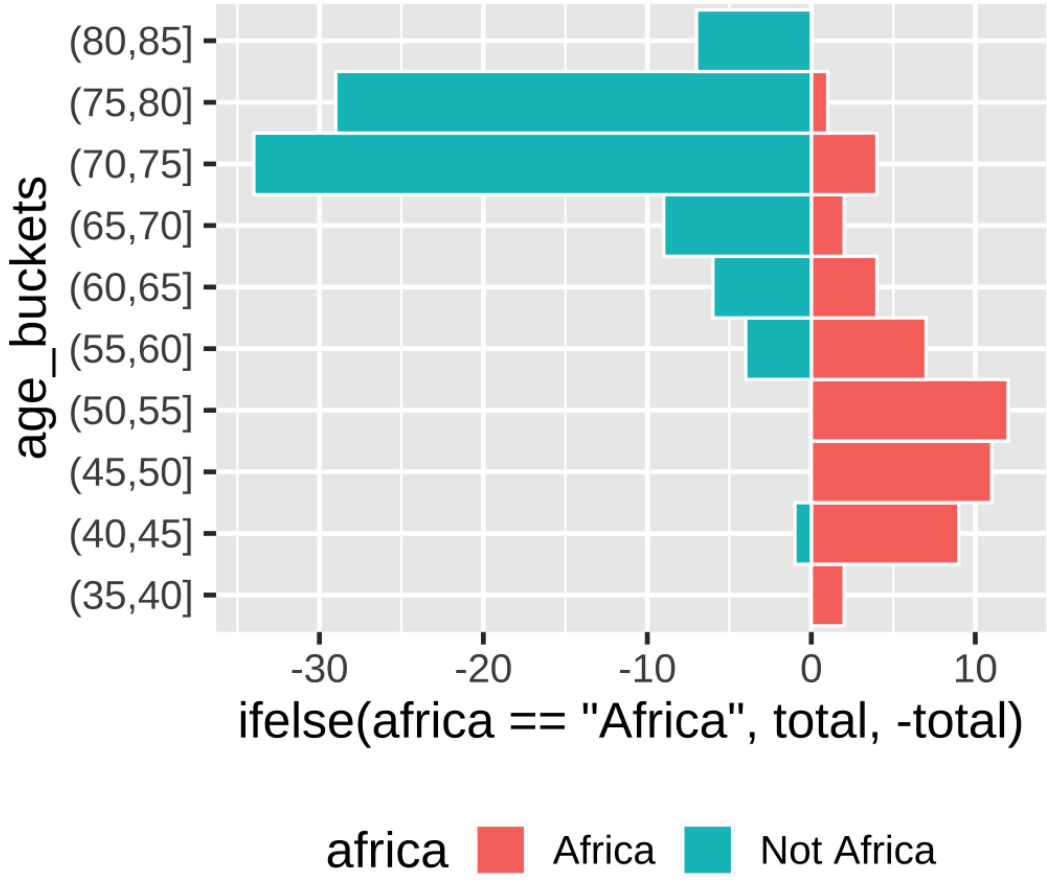
Facet with a different variable

```
ggplot(gapminder_2002,
       aes(x = lifeExp,
           fill = continent)) +
  geom_histogram(binwidth = 5,
                 color = "white",
                 boundary = 50) +
  guides(fill = "none") +
  facet_wrap(vars(continent))
```



Pyramid histograms

```
gapminder %>%
  filter(year == 2002) %>%
  mutate(africa =
           ifelse(continent == "Africa",
                  "Africa",
                  "Not Africa")) %>%
      mutate(age_buckets =
               cut(lifeExp,
                   breaks = seq(30, 90, by = 5))) %
      group_by(africa, age_buckets) %>%
      summarize(total = n()) %>%
ggplot(aes(y = age_buckets,
           x = ifelse(africa == "Africa",
                      total, -total),
           fill = africa)) +
  geom_col(width = 1, color = "white") +
  theme(legend.position = "bottom")
```

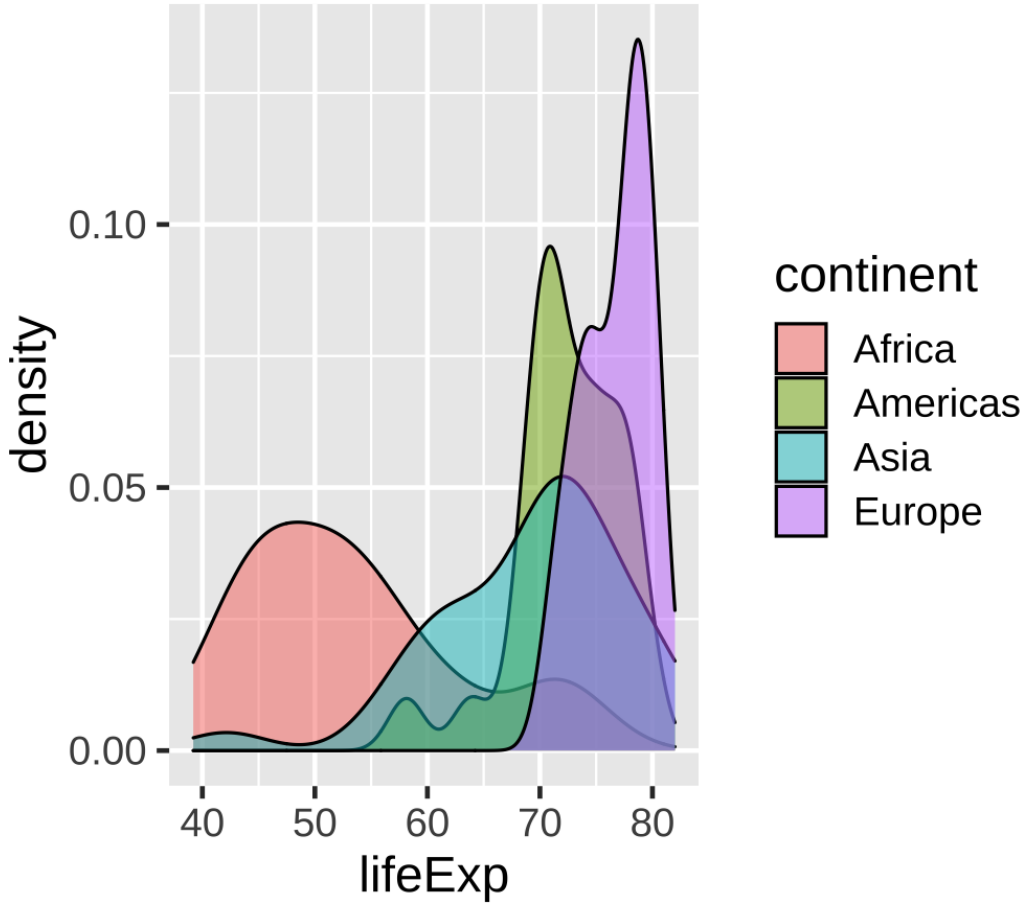


Multiple densities: Transparency

```
gapminder_2002 %>%
  filter(continent != "Oceania") %>%
  ggplot(aes(x = lifeExp,
             fill = continent)) +
  geom_density(alpha = 0.5)
```

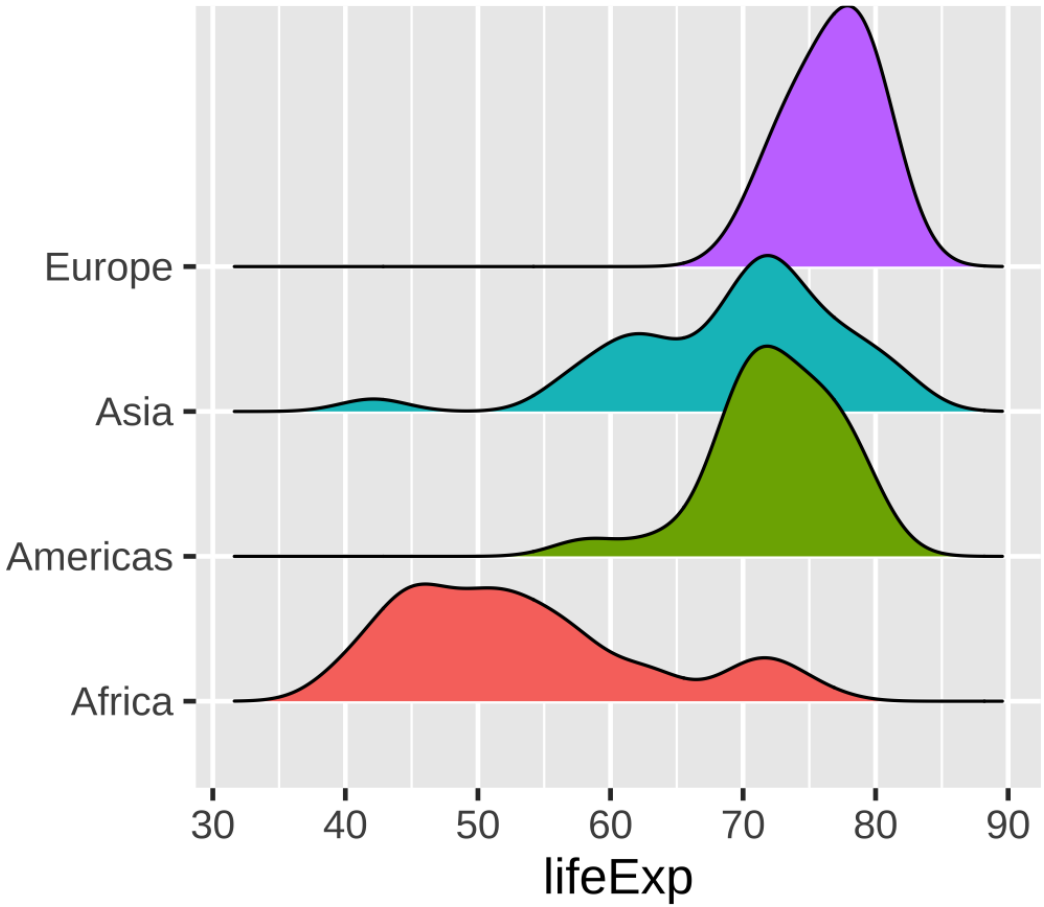
But be careful, these can get confusing quickly

With many groups, better to space them out using ridgeline plots



Multiple densities: Ridgeline plots

```
library(ggridges)

ggplot(filter(gapminder_2002,
              continent != "Oceania"),
       aes(x = lifeExp,
           fill = continent,
           y = continent)) +
  guides(fill = "none") +
  labs(y = NULL) +
  geom_density_ridges()
```

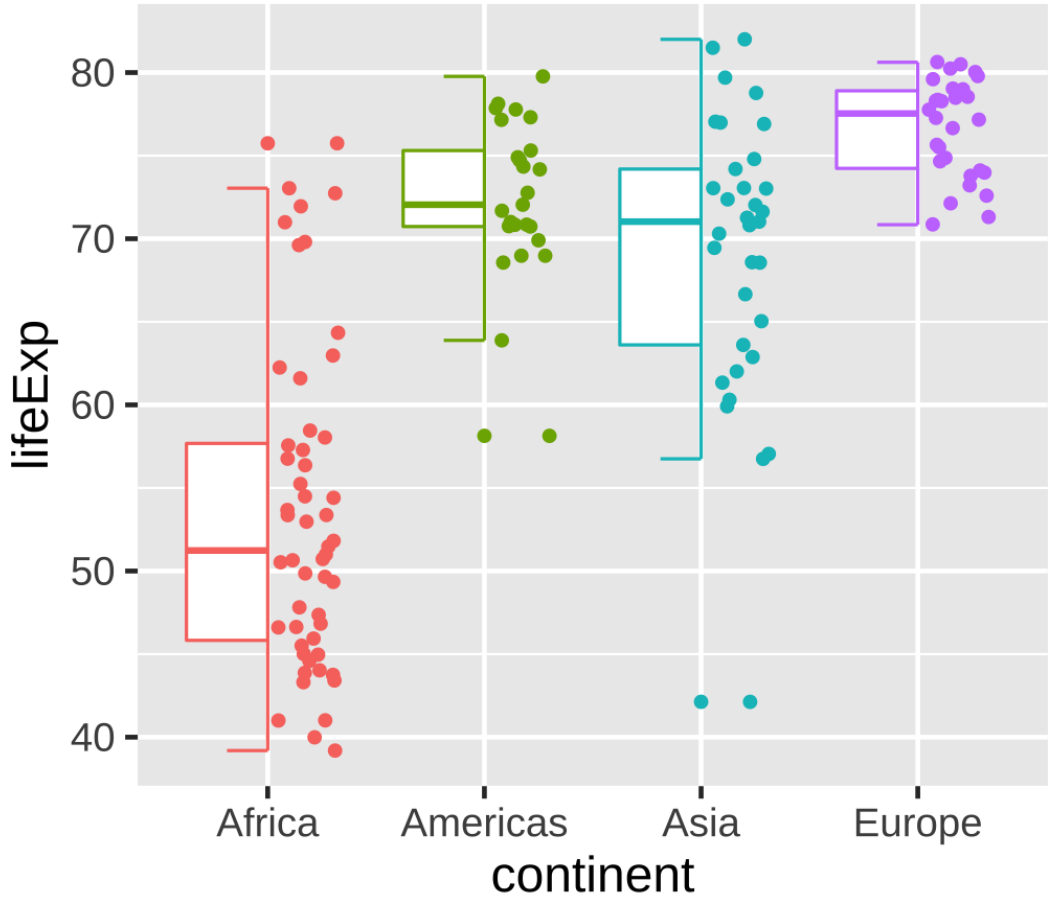


There is no explicit scale for the densities anymore (it is shared with y)

With many densities, use a single fill color to prevent distraction

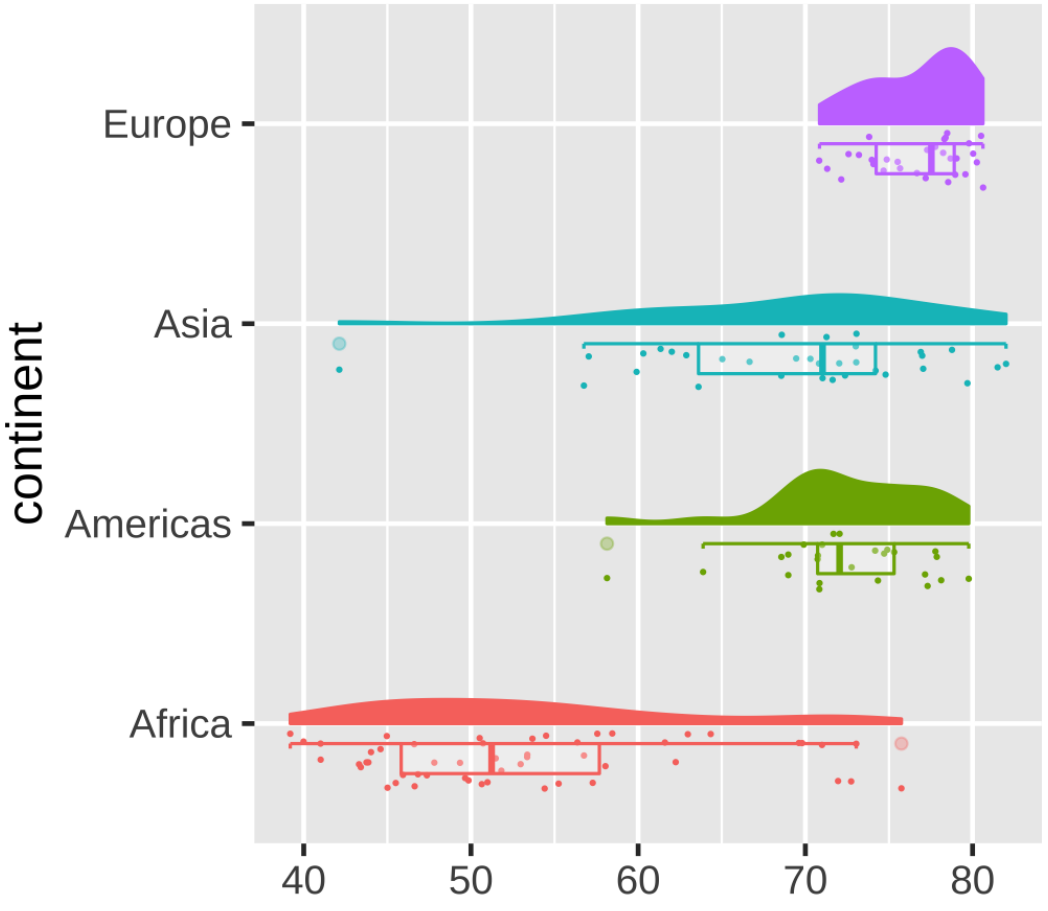
Multiple geoms: gghalves

```
library(gghalves)

ggplot(filter(gapminder_2002,
              continent != "Oceania"),
       aes(y = lifeExp,
           x = continent,
           color = continent)) +
  geom_half_boxplot(side = "l") +
  geom_half_point(side = "r") +
  guides(color = "none")
```



Multiple geoms: Raincloud plots

```
library(gghalves)

ggplot(filter(gapminder_2002,
              continent != "Oceania"),
       aes(y = lifeExp,
           x = continent,
           color = continent)) +
  geom_half_point(side = "l", size = 0.3) +
  geom_half_boxplot(side = "l", width = 0.5,
                    alpha = 0.3, nudge = 0.1) +
  geom_half_violin(aes(fill = continent),
                   side = "r") +
  guides(fill = "none", color = "none") +
  labs(y = NULL) +
  coord_flip()
```



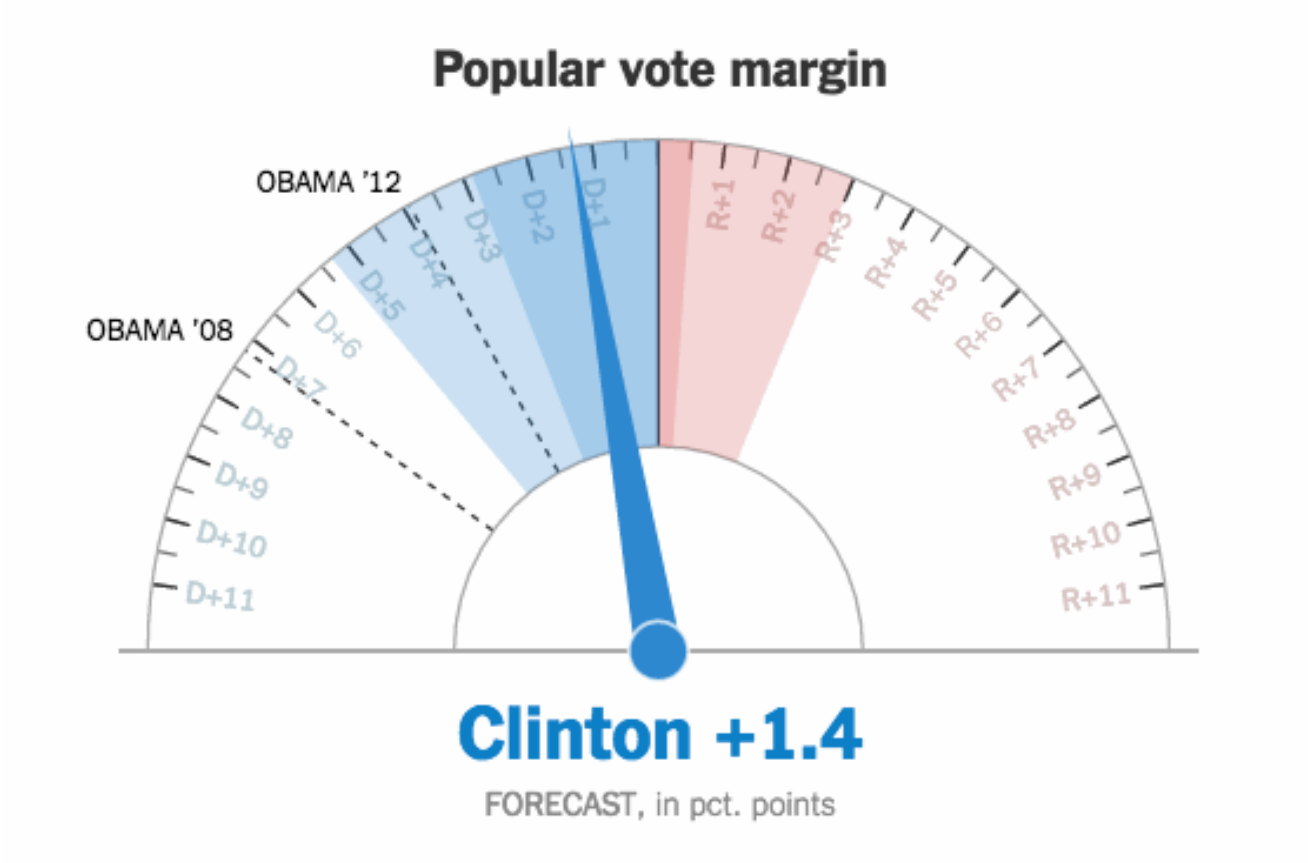
Uncertainty in models and simulations

We have already seen at least one example: `geom_smooth()`

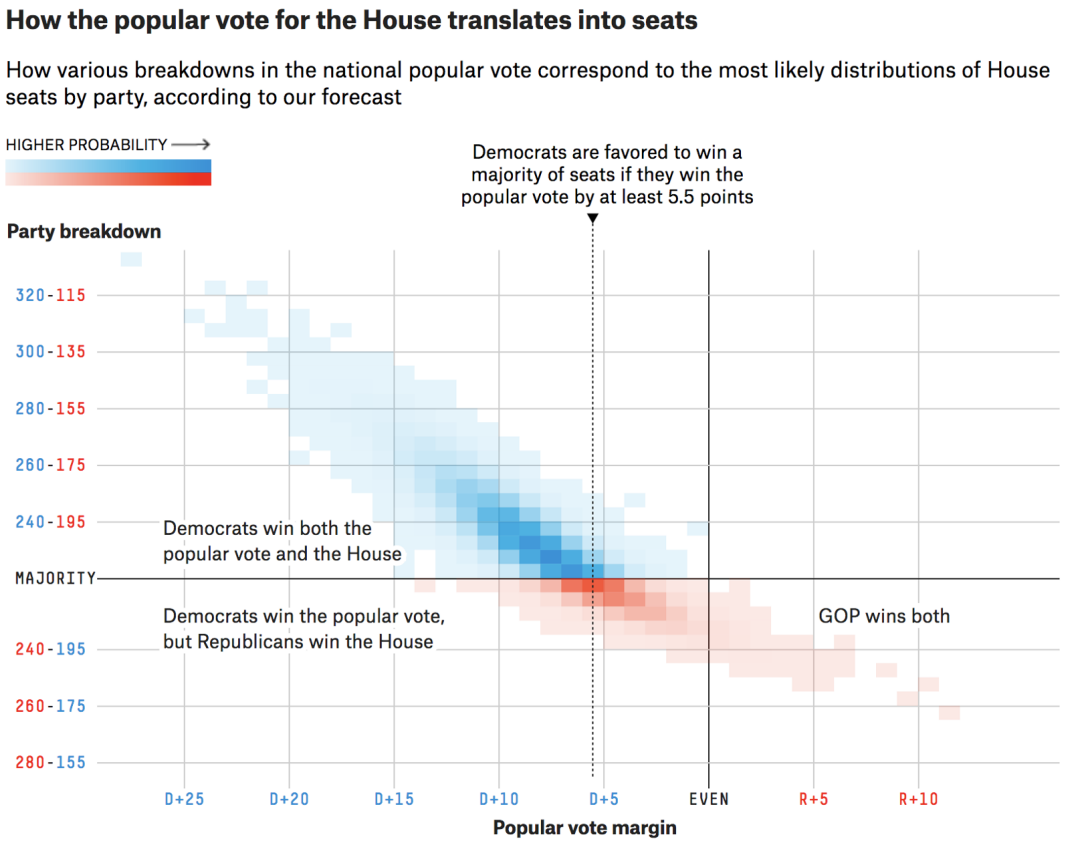
We will discuss these more next week

Until then, here are a few real-world examples

The needle


Popular vote margin
OBAMA '12
OBAMA '08
D+1
D+2
D+3
D+4
D+5
D+6
D+7
D+8
D+9
D+10
D+11
R+1
R+2
R+3
R+4
R+5
R+6
R+7
R+8
R+9
R+10
R+11
Clinton +1.4
FORECAST, in pct. points

Uncertainty in model outcomes



FiveThirtyEight's 2018 midterms model outcomes plot