Distributions

Week 8

AEM 2850 / 5850 : R for Business Analytics
Cornell Dyson
Spring 2023

Acknowledgements: Andrew Heiss, Claus Wilke

Class Participation

I got some questions about the class participation component of the course grade

Reminder from the syllabus:

> "Class participation and regular attendance are expected. Excessive absences and failure to complete weekly in-class examples will impact your final grade..."

We will use name cards to help me learn names and track attendance

I expect everyone to attempt all examples (even if absent)

As long as you attend class regularly and try all the examples, you should receive full credit for class participation (5% of final grade)

Announcements

Prelim 1 grades posted on canvas

- See canvas announcement for details
- Grading questions? Contact Hui Zhou first, then me
- Other questions? Schedule an appointment at aem2850.youcanbook.me

I will give you preliminary details on the group project today

Questions before we get started?

Plan for today

Prologue

Group project

Distributions

Proportions: cut for time

Prologue

Health and wealth revisited

Group project

Group project

Use R and the tidyverse to wrangle and visualize equities data

Multiple parts:

1. AAPL
2. The S&P 500
3. Our Class Portfolio
4. Something extra for 5850 students (TBD)

Group project: data

```r
sp500_companies
```

```
## # A tibble: 504 × 7
##    symbol company                          identifier sedol weight sector local…¹
##    <chr>  <chr>                            <chr>      <chr>  <dbl> <chr>  <chr>
##  1 AAPL   Apple Inc.                       03783310   2046… 0.0701 Infor… USD
##  2 MSFT   Microsoft Corporation            59491810   2588… 0.0601 Infor… USD
##  3 AMZN   Amazon.com Inc.                  02313510   2000… 0.0349 Consu… USD
##  4 GOOGL  Alphabet Inc. Class A            02079K30   BYVY… 0.0218 Commu… USD
##  5 GOOG   Alphabet Inc. Class C            02079K10   BYY8… 0.0203 Commu… USD
##  6 TSLA   Tesla Inc                        88160R10   B616… 0.0185 Consu… USD
##  7 BRK-B  Berkshire Hathaway Inc. Class B 08467070   2073… 0.0162 Finan… USD
##  8 NVDA   NVIDIA Corporation               67066G10   2379… 0.0160 Infor… USD
##  9 FB     Meta Platforms Inc. Class A      30303M10   B7TL… 0.0130 Commu… USD
## 10 UNH    UnitedHealth Group Incorporated 91324P10   2917… 0.0124 Healt… USD
## # … with 494 more rows, and abbreviated variable name ¹local_currency
```

Group project: 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
```

Group project: data

```
our_companies
```

```
## # A tibble: 25 × 2
##    name                                 n
##    <chr>                            <dbl>
##  1 Alphabet Inc. Class A                2
##  2 Amazon.com Inc.                      1
##  3 Apple Inc.                           8
##  4 Blackstone Inc.                      1
##  5 Boyd Gaming Corp                     1
##  6 Catalyst Pharmaceuticals, Inc.       1
##  7 China International Capital Corp     1
##  8 Chipotle Mexican Grill Inc.          1
##  9 Costco Wholesale Corporation         1
## 10 Deere & Company                      1
## # … with 15 more rows
```

Group project: overview

First: you choose groups of 3

- All group members must be in the same section (i.e., 2850 or 5850)

Write quarto report that summarizes your work, presents visualizations, and discusses takeaways

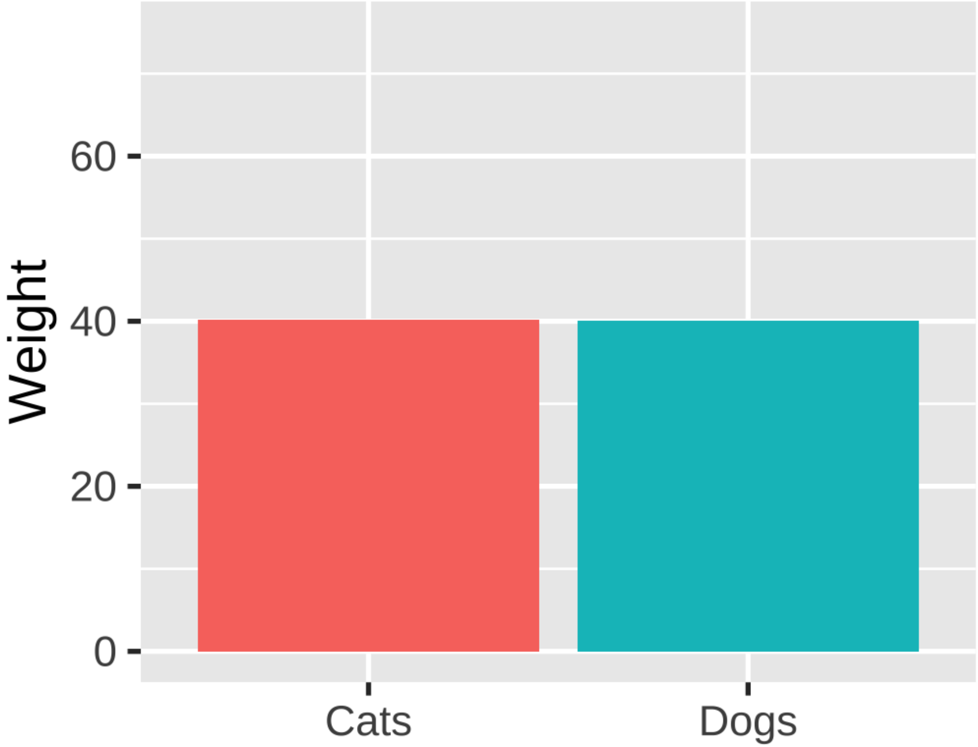
Do not use any packages outside base R and the tidyverse

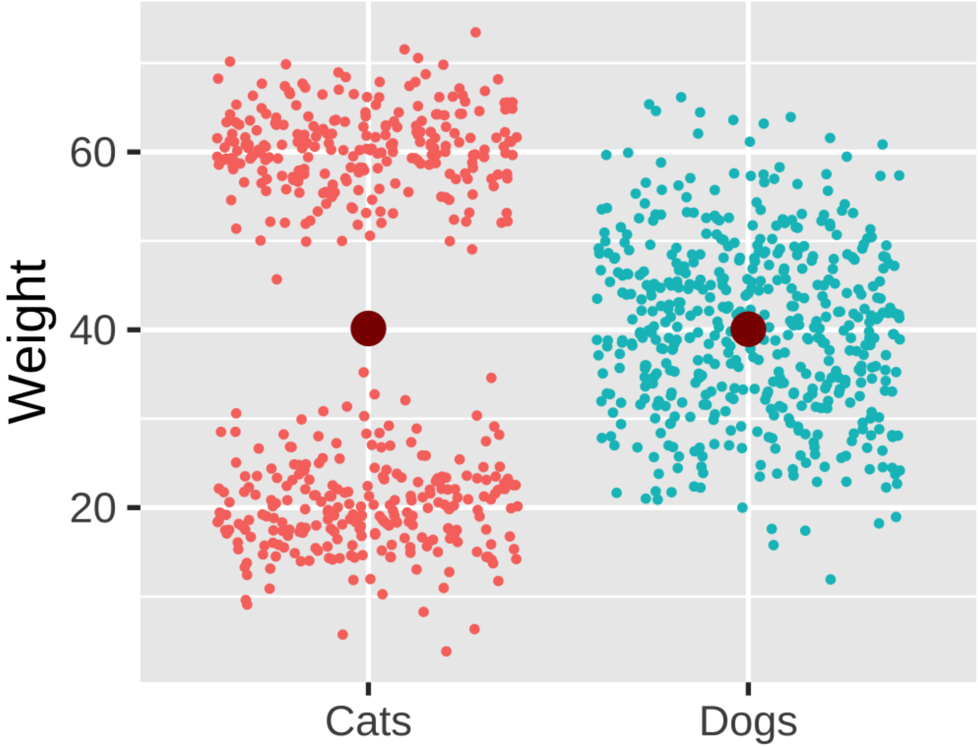
Limited TA help for Part 3!

Due: Friday, April 14 at 11:59pm (after spring break)

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

What are some common methods for visualizing distributions?

Histograms, densities, box plots, etc.

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`?

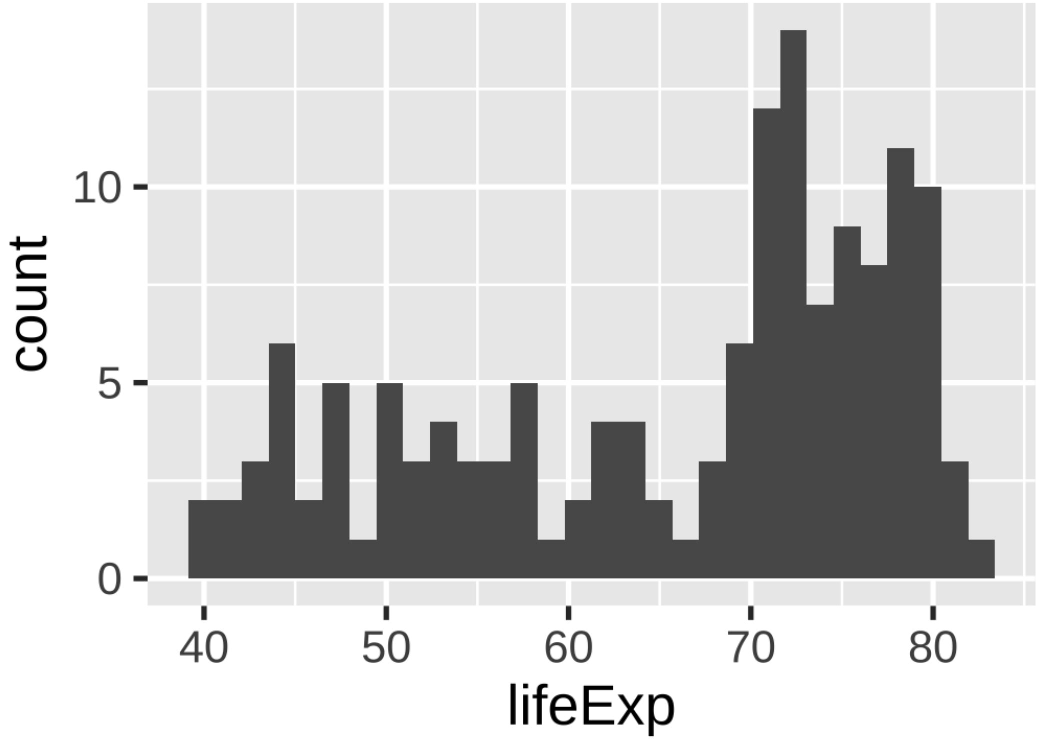
```r
library(gapminder)

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   Asia       2002    80.4 19546792    30688.
```

Histograms

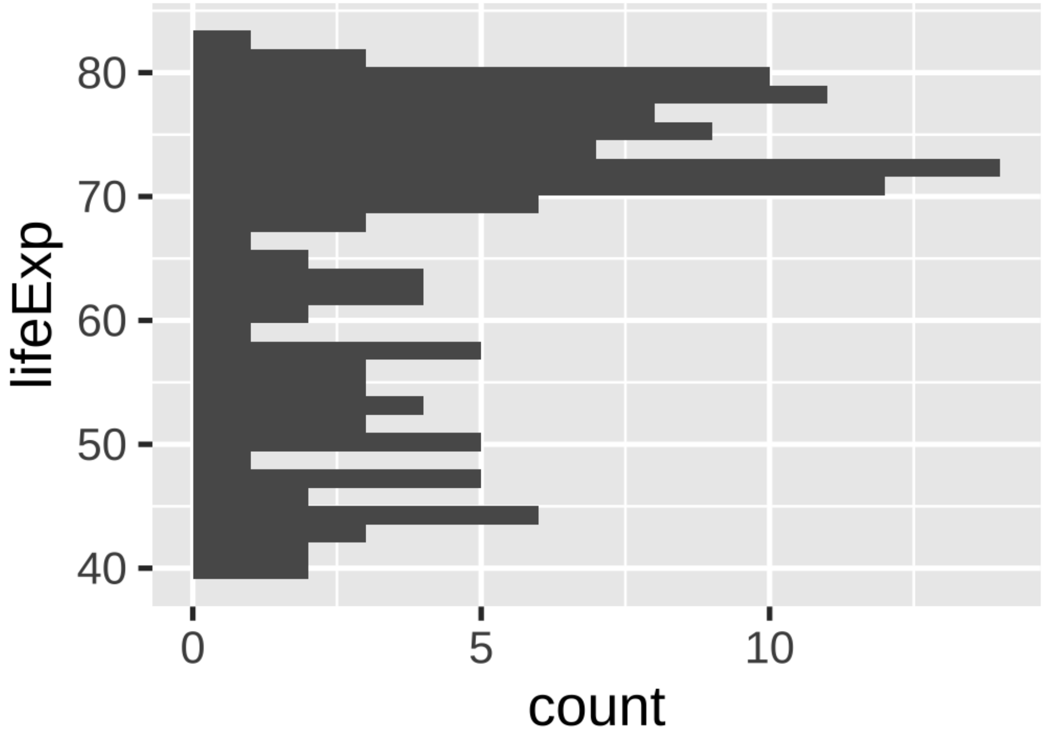
```
gapminder_2002 |>
  ggplot(aes(x = lifeExp)) +
  geom_histogram()
```

What if we mapped `lifeExp` to `y`?



Histograms

```r
gapminder_2002 |>
  ggplot(aes(y = lifeExp)) +
  geom_histogram()
```

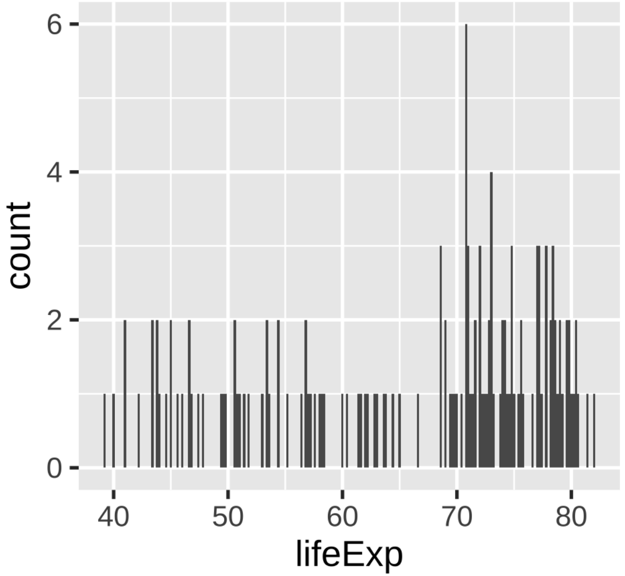


Histograms: bin width

No official rule for what makes a good bin width

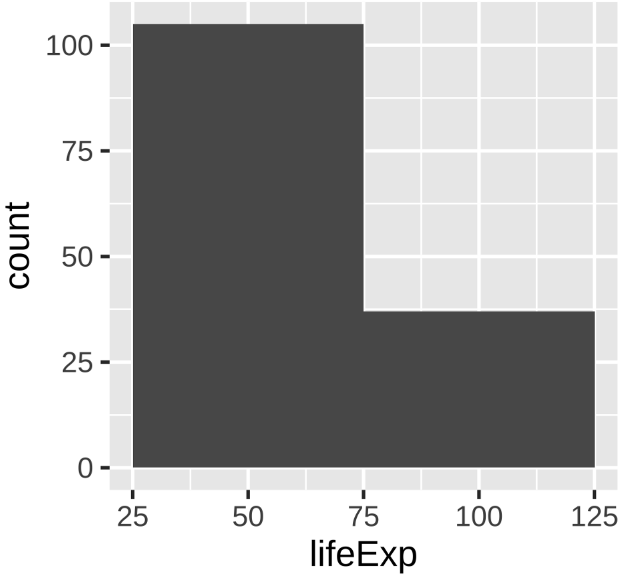
Too narrow:

`geom_histogram(binwidth = .2)`



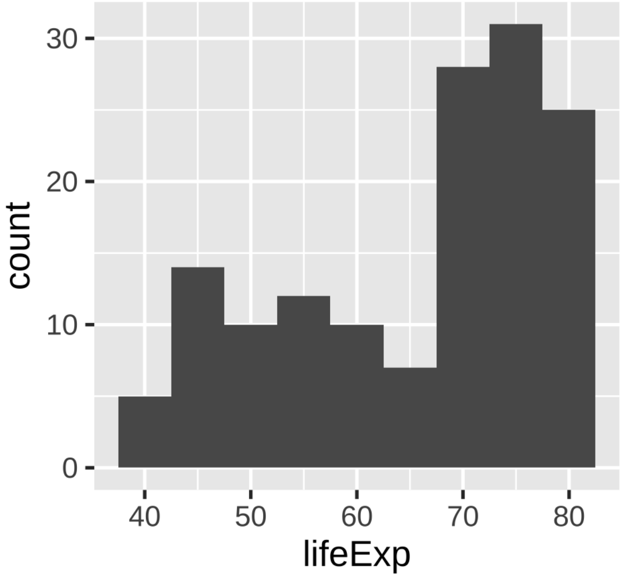
Too wide:

`geom_histogram(binwidth = 50)`



(One type of) just right:

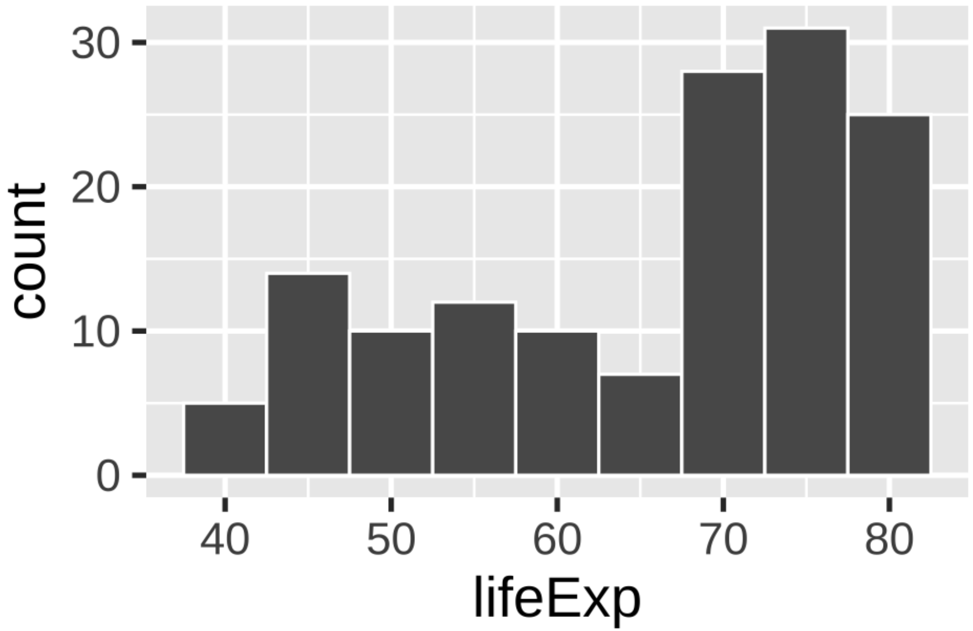
`geom_histogram(binwidth = 5)`



Histogram tips

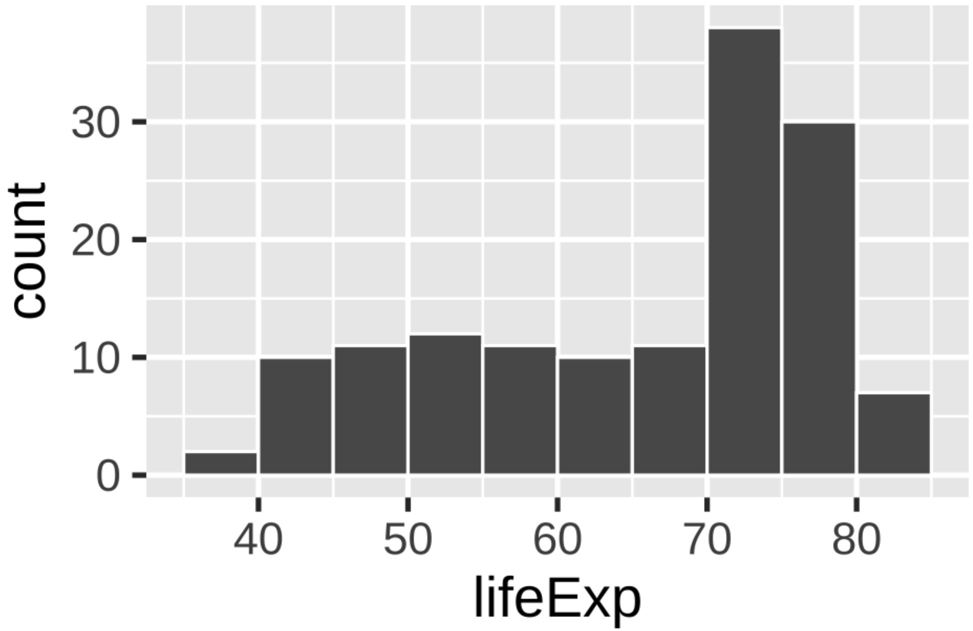
Add a border to the bars for readability

`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?

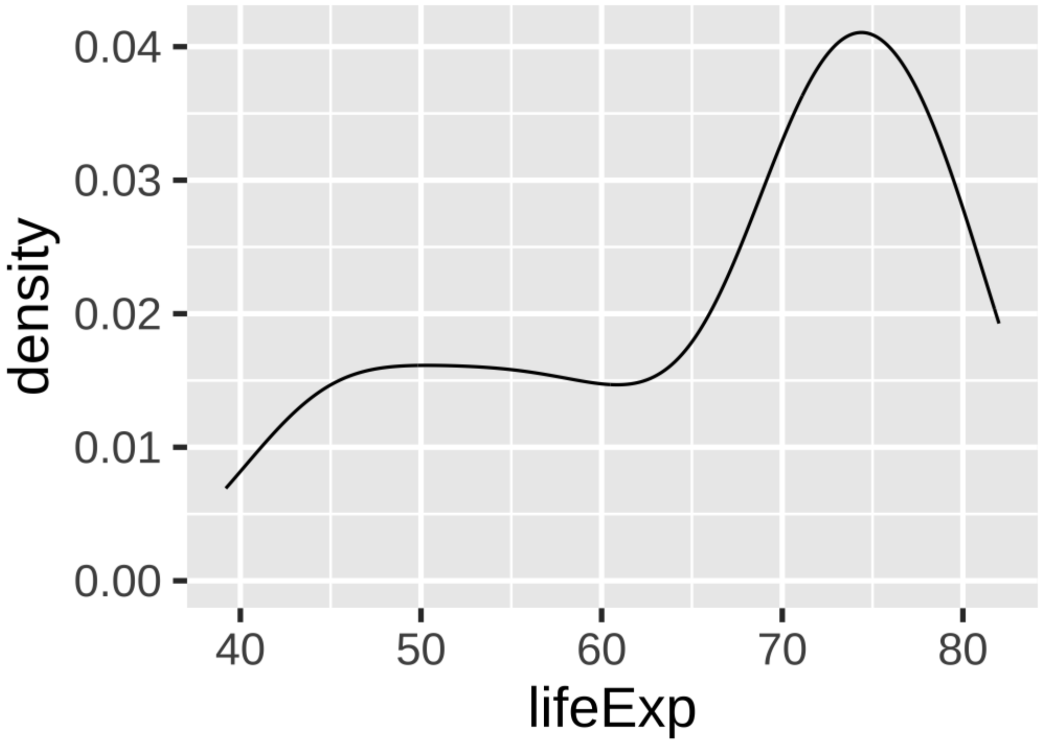
Estimates of the **probability *density* function** of a random variable

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

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp)) +
  geom_density()
```

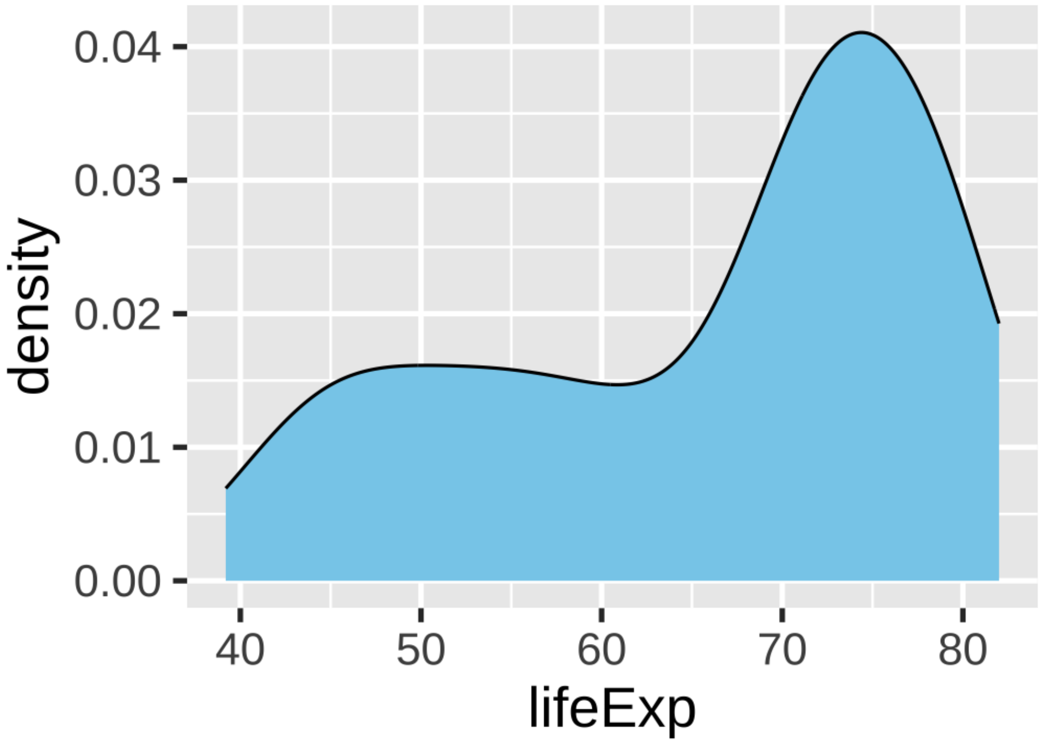


Density plots: add some color

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

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

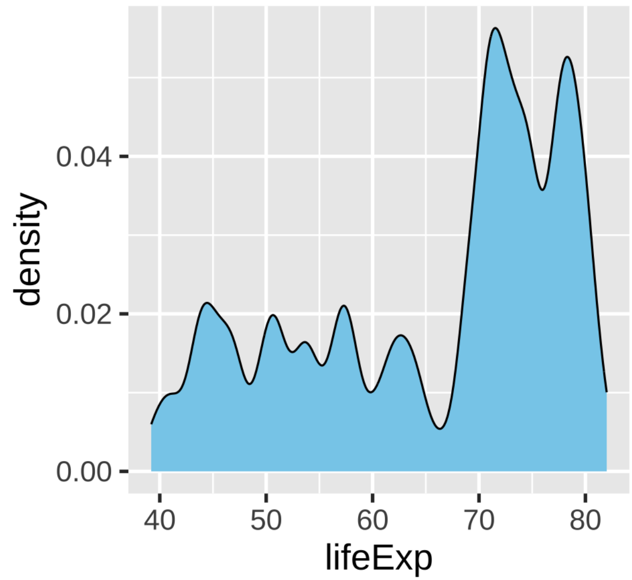
Here we used **fill = "skyblue"**



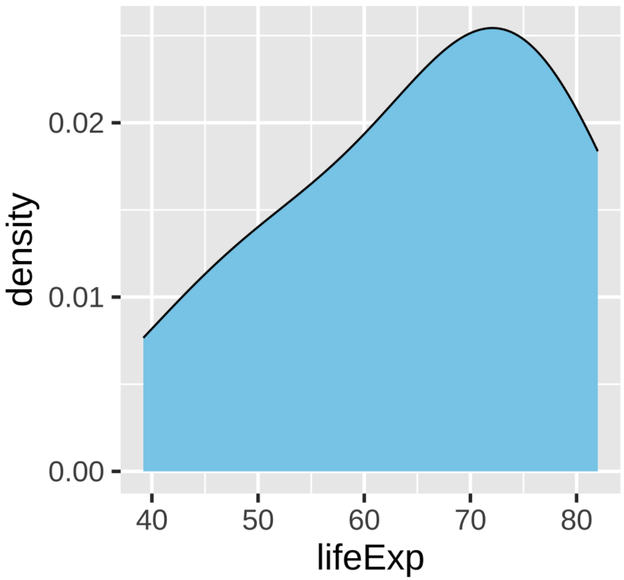
Density plots: bandwidths

Different options for calculus change the plot shape

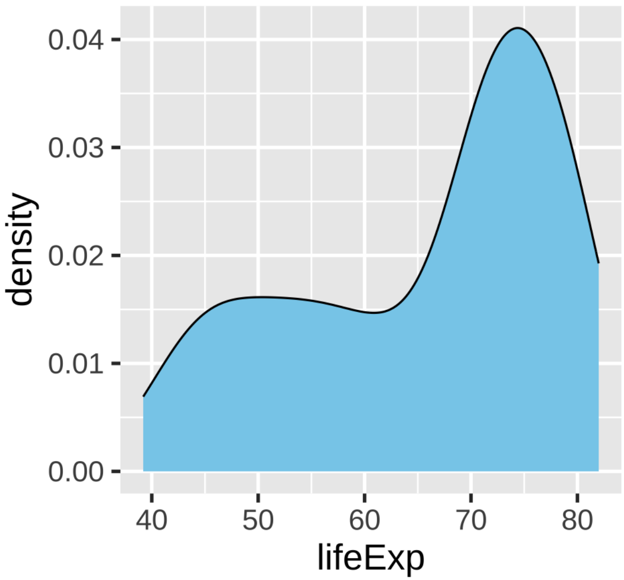
`bw = 1`



`bw = 10`



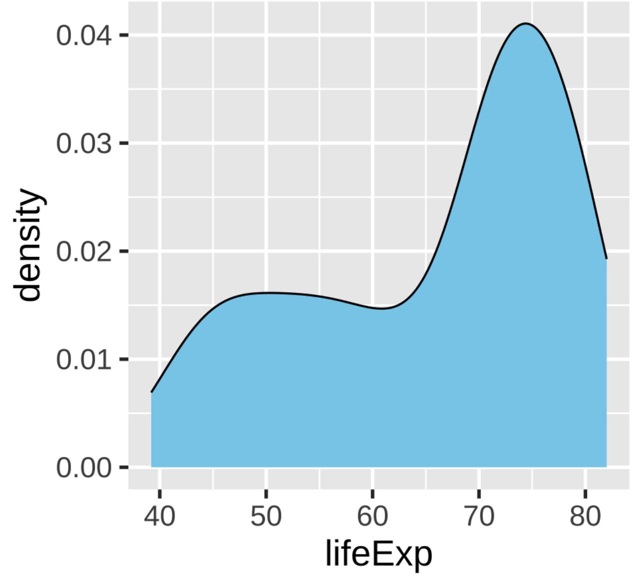
`bw = "nrd0"` (default)



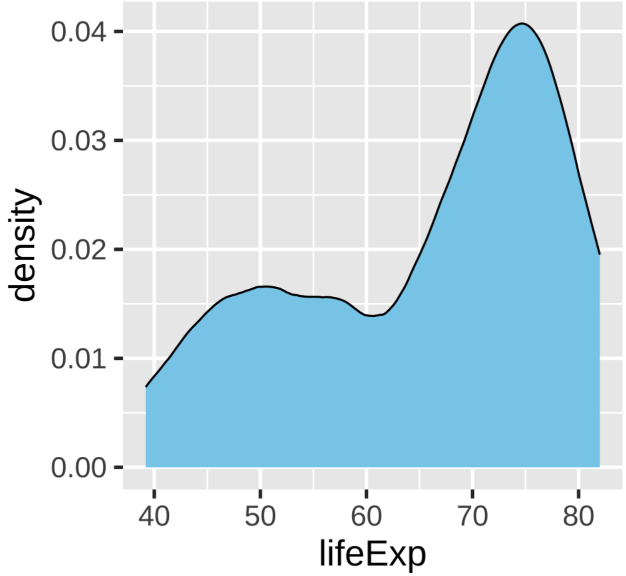
Density plots: kernels

Different options for calculus change the plot shape

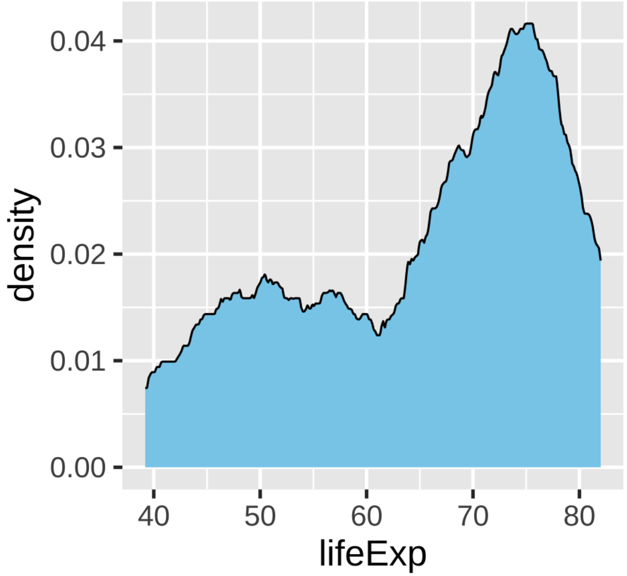
`kernel = "gaussian"`



`"epanechnikov"`



`"rectangular"`

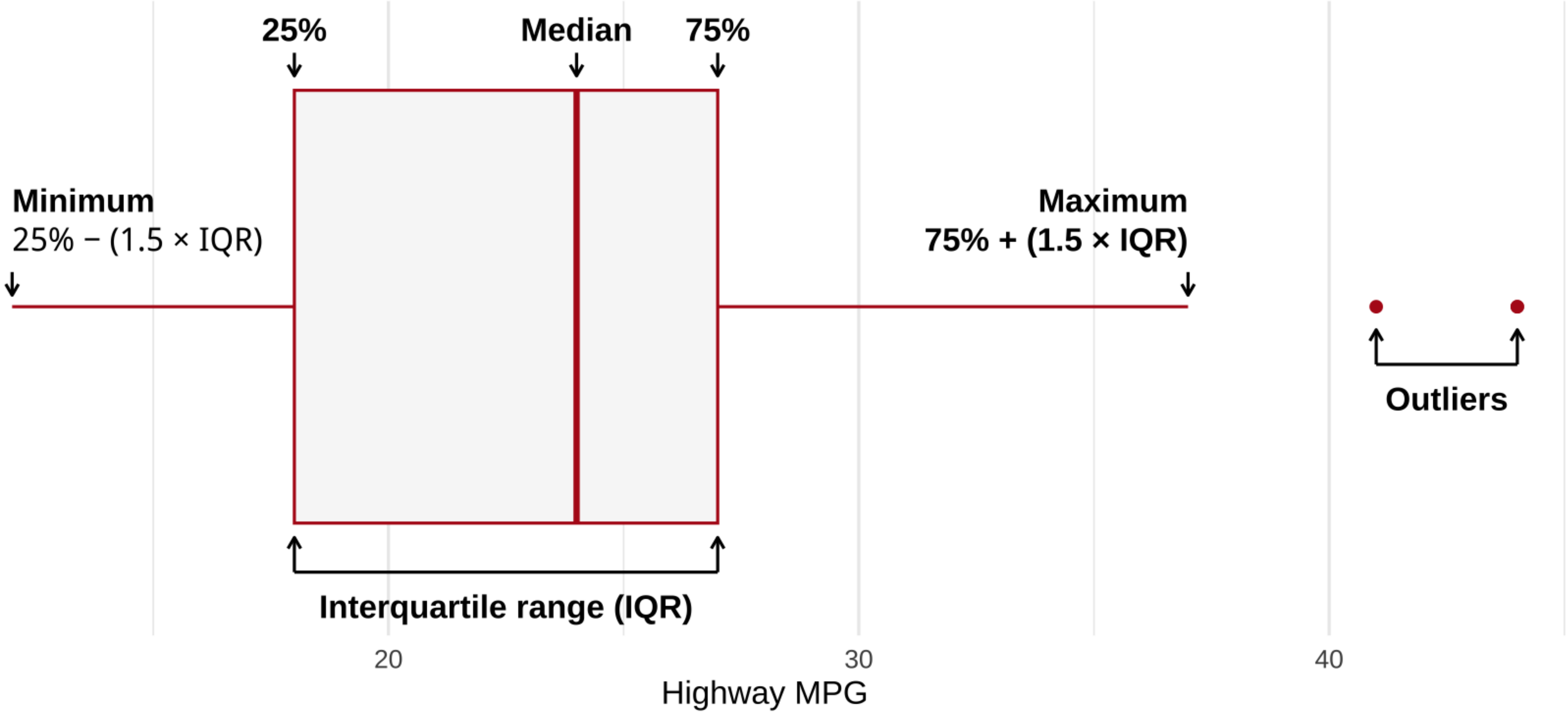


Box plots

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?

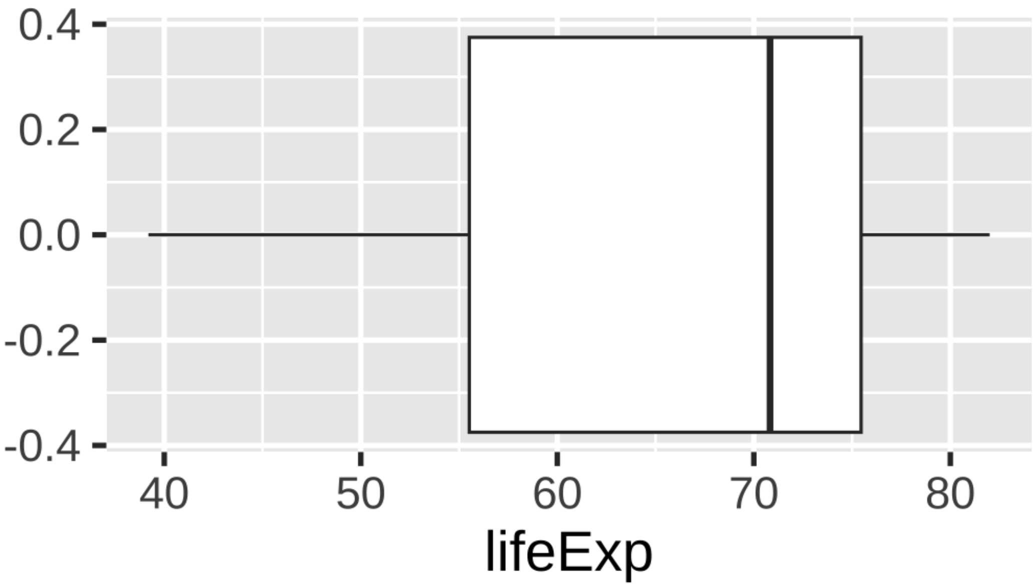
Graphical representations of specific points in a distribution

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

Box plots

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp)) +
  geom_boxplot()
```

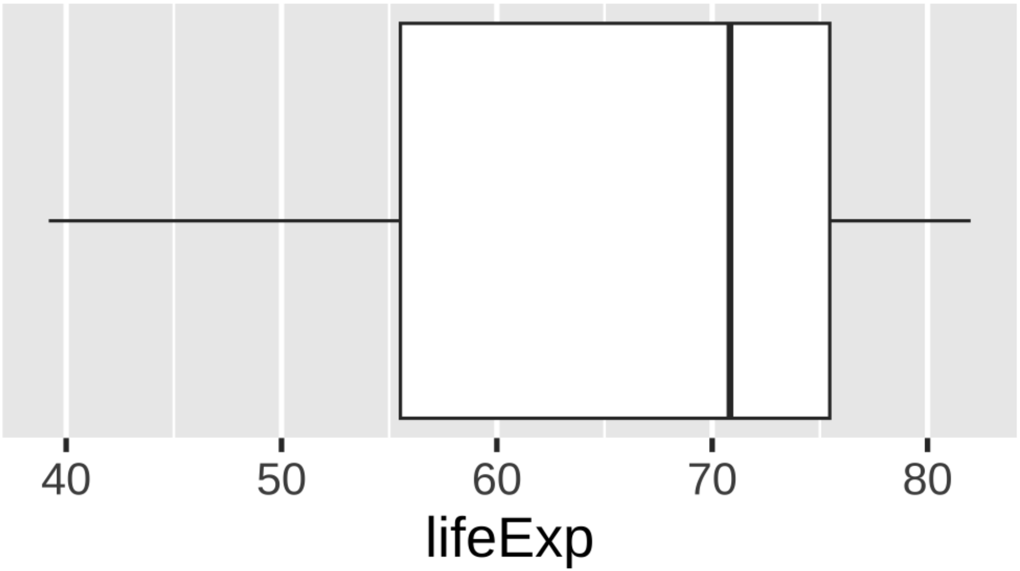
What do the y axis numbers mean?



Box plots

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

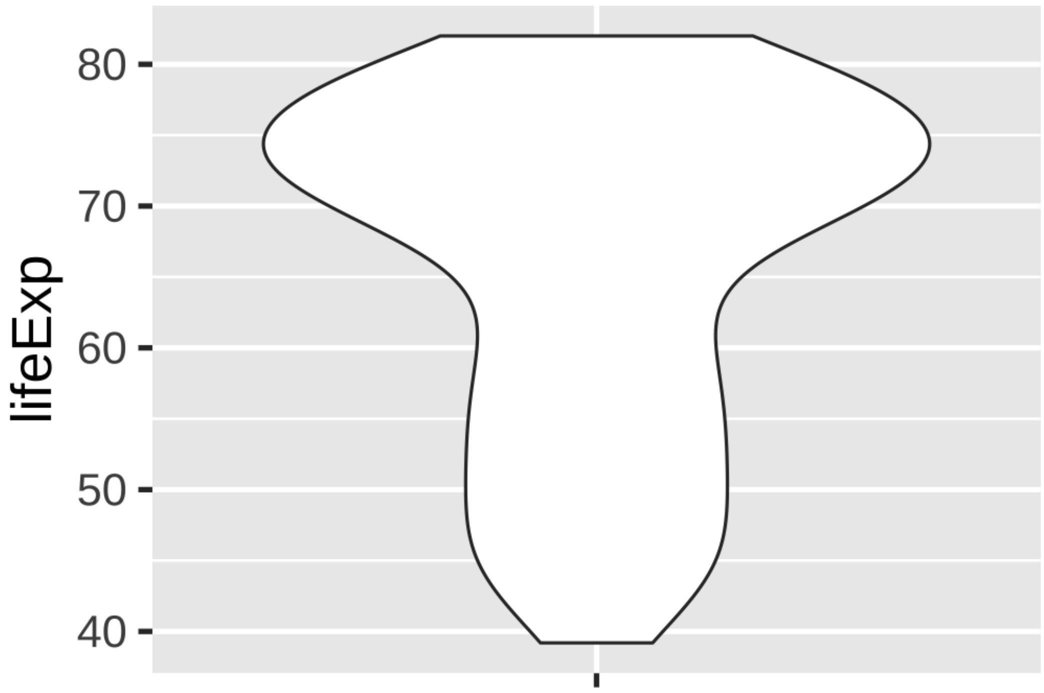
```
gapminder_2002 |>
  ggplot(aes(x = lifeExp)) +
  geom_boxplot() +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank())
```



Violin plots

Mirror density plot and flip

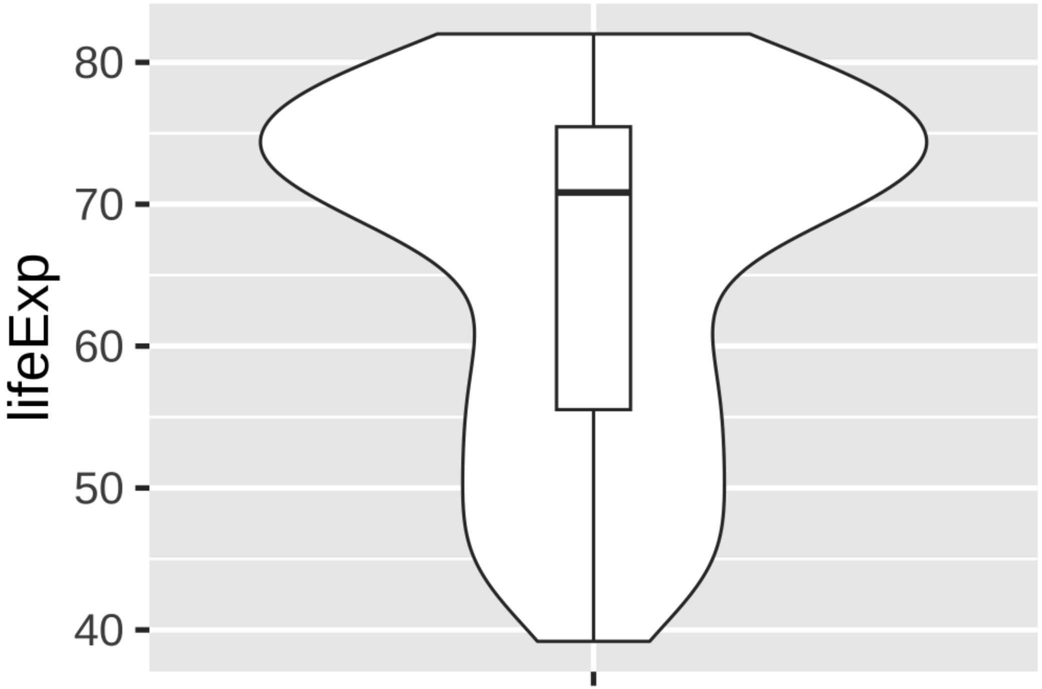
```
gapminder_2002 |>
  ggplot(aes(x = "",
             y = lifeExp)) +
  geom_violin() +
  labs(x = NULL)
```



Overalying geometries

We can overlay multiple geometries to provide more information

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



Uncertainty across multiple variables

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

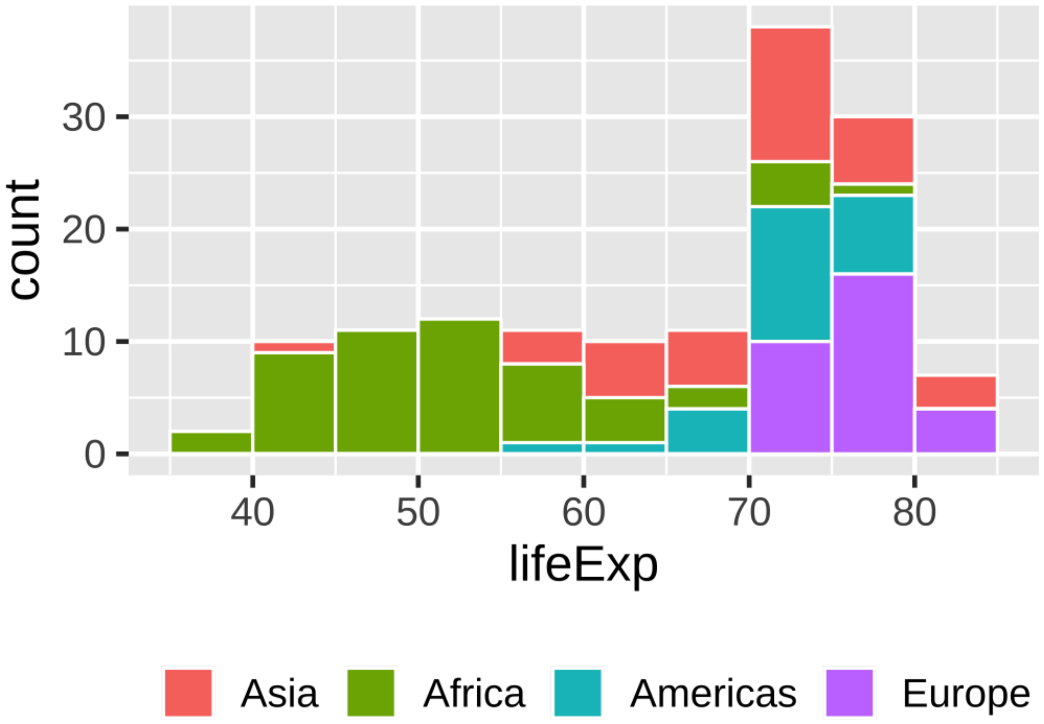
Add a `fill` aesthetic or use facets!

Multiple histograms

Fill with a different variable

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp,
             fill = continent)) +
  geom_histogram(binwidth = 5,
                 color = "white",
                 boundary = 50) +
  theme(legend.position = "bottom") +
  labs(fill = NULL)
```

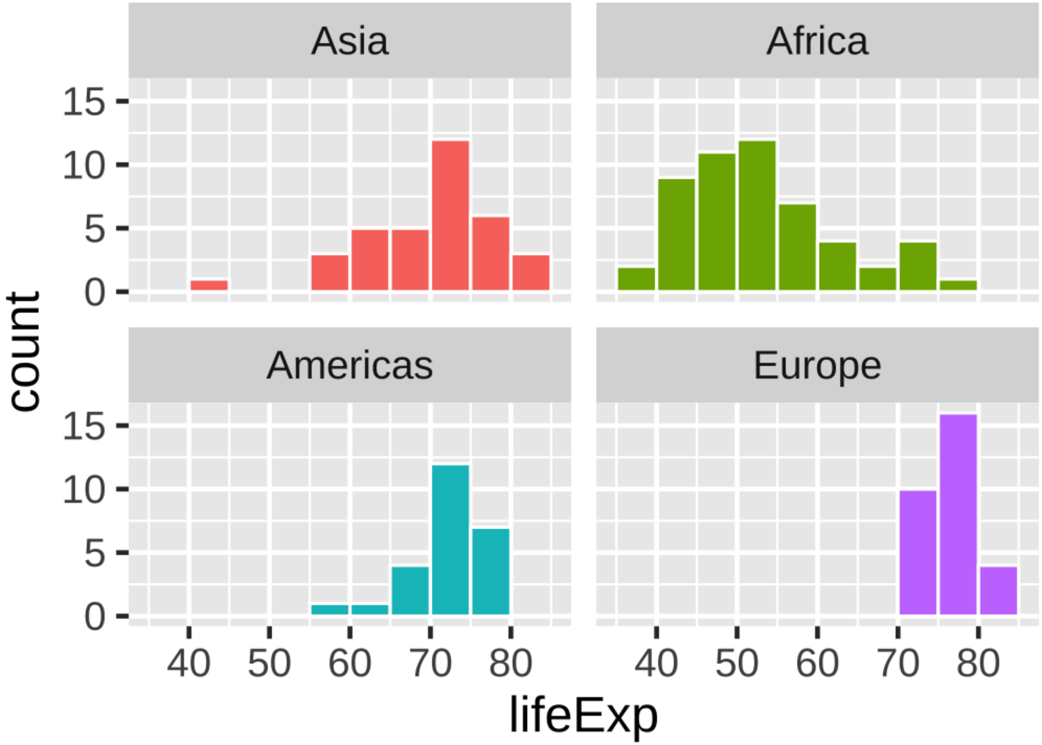
This is bad and hard to read though



Multiple histograms

Facet with a different variable

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

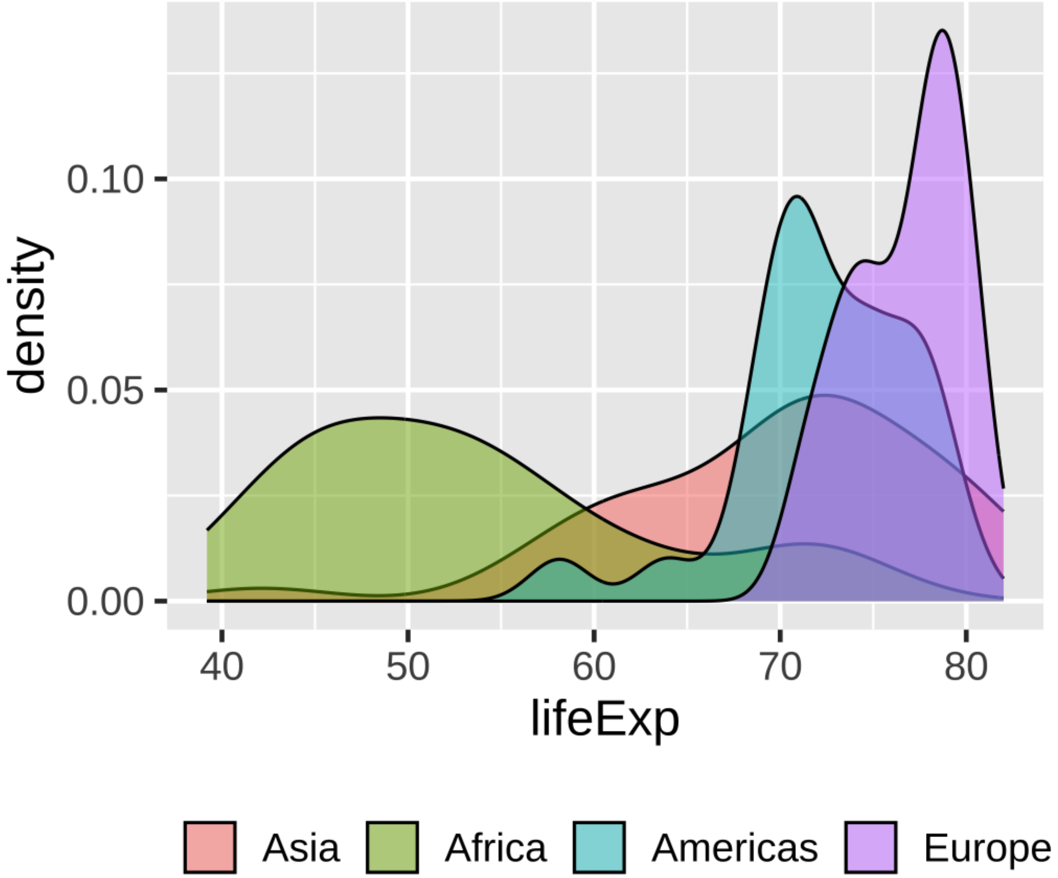


Multiple densities: Transparency

```
gapminder_2002 |>
  ggplot(aes(x = lifeExp,
             fill = continent)) +
  geom_density(alpha = 0.5) +
  theme(legend.position = "bottom") +
  labs(fill = NULL)
```

But be careful, these can get confusing quickly

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

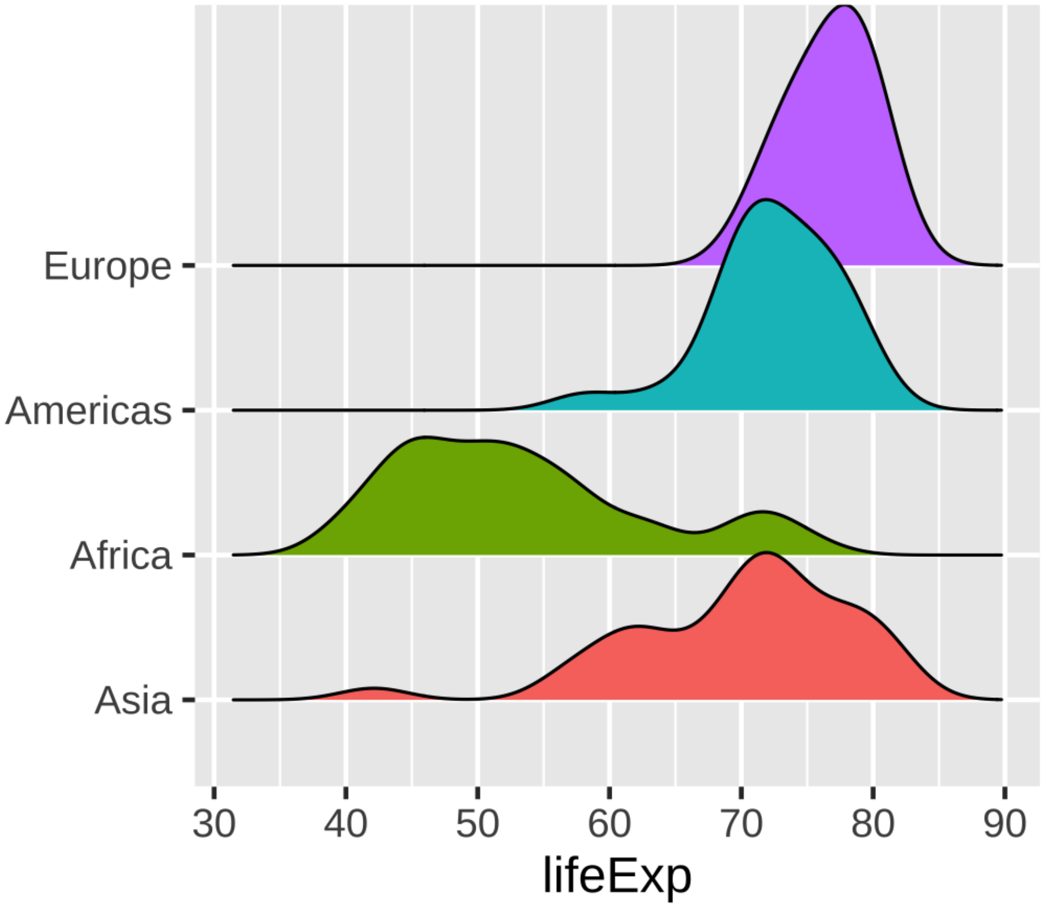


Multiple densities: Ridgeline plots

```r
library(ggridges)

gapminder_2002 |>
  ggplot(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)

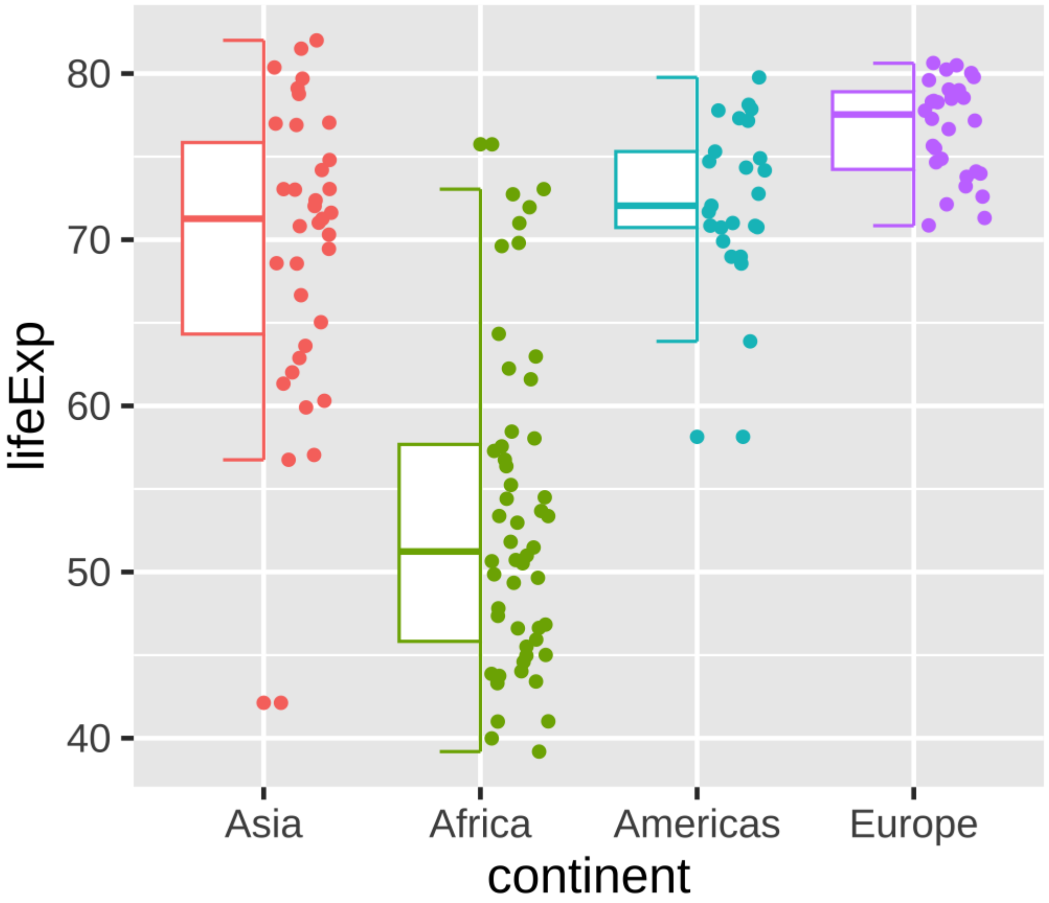
With many densities, use a single fill color to prevent distraction



example-08
remaining slides cut for time

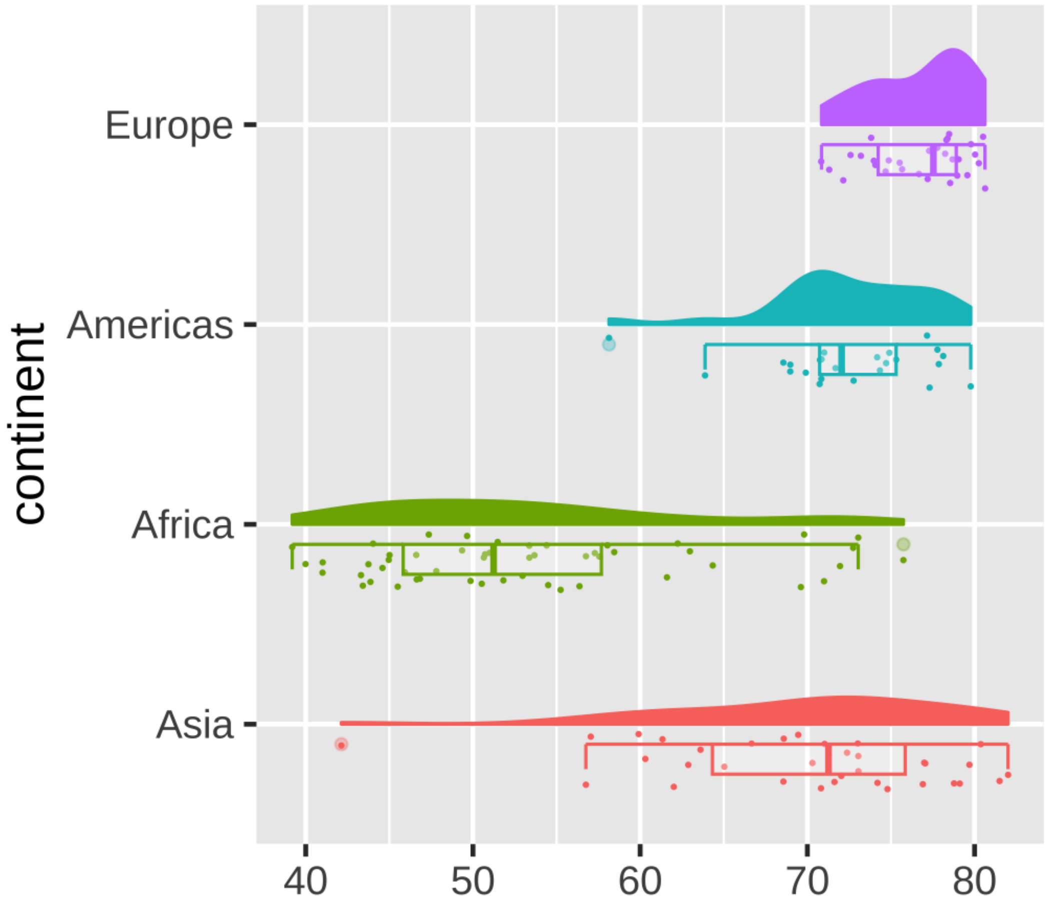
Multiple geoms: gghalves

```r
library(gghalves)

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



Multiple geoms: Raincloud plots

```r
library(gghalves)

gapminder_2002 |>
  ggplot(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()
```



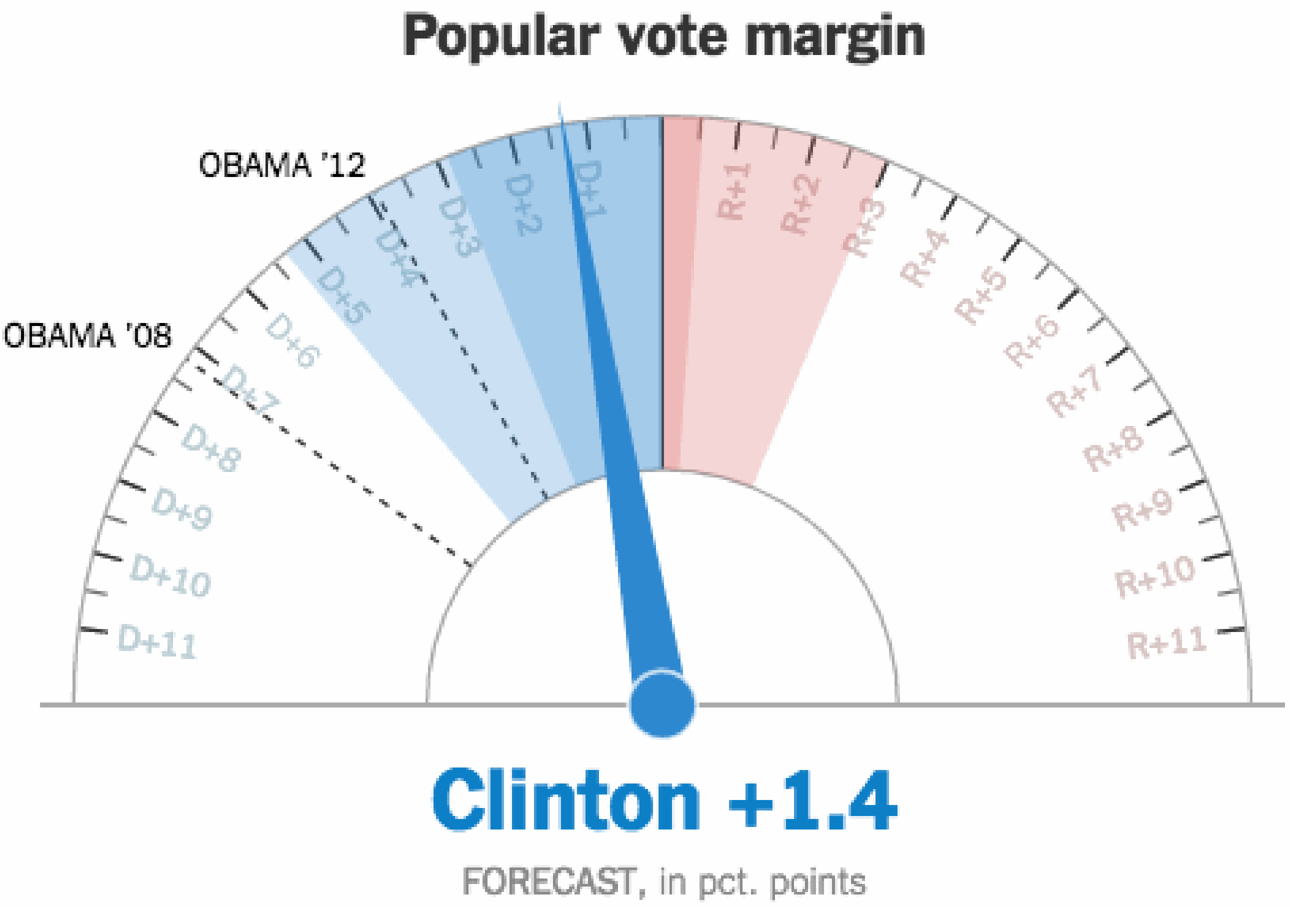
Uncertainty in models and simulations

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

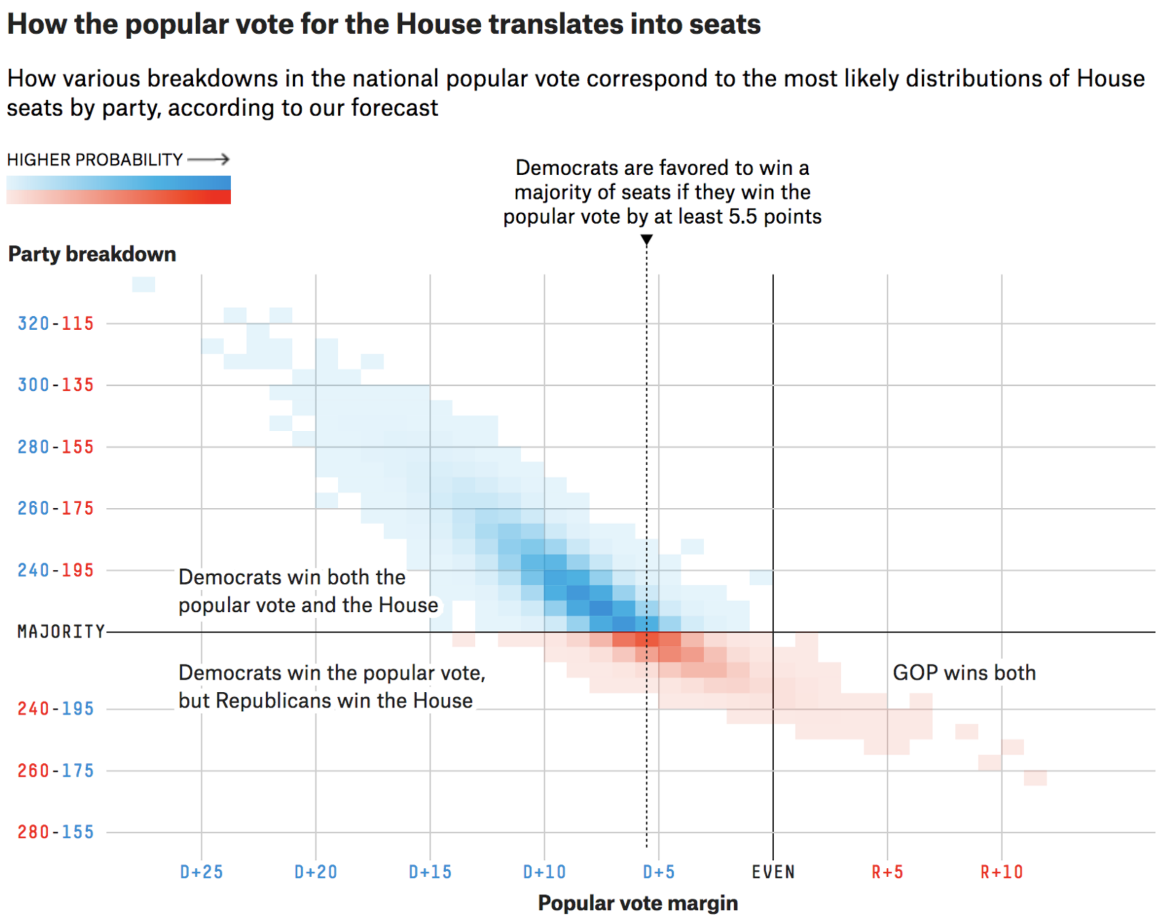
We will discuss these more in **Relationships**

Until then, here are a few real-world examples

The needle


Popular vote margin
OBAMA '12
OBAMA '08
D+11
D+10
D+9
D+8
D+7
D+6
D+5
D+4
D+3
D+2
D+1
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



How the popular vote for the House translates into seats

How various breakdowns in the national popular vote correspond to the most likely distributions of House seats by party, according to our forecast

Democrats are favored to win a majority of seats if they win the popular vote by at least 5.5 points

Democrats win both the popular vote and the House

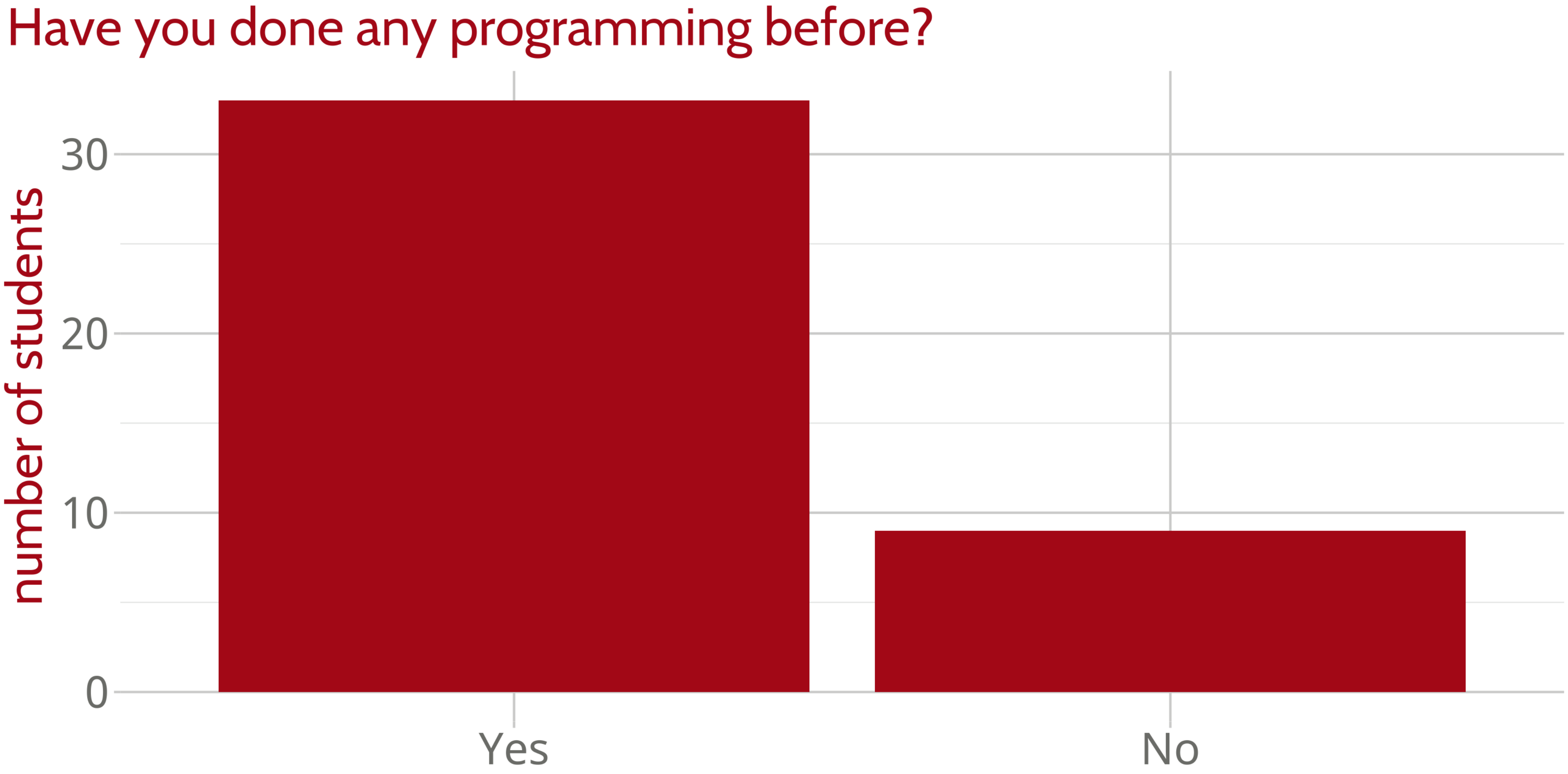
Democrats win the popular vote, but Republicans win the House

GOP wins both

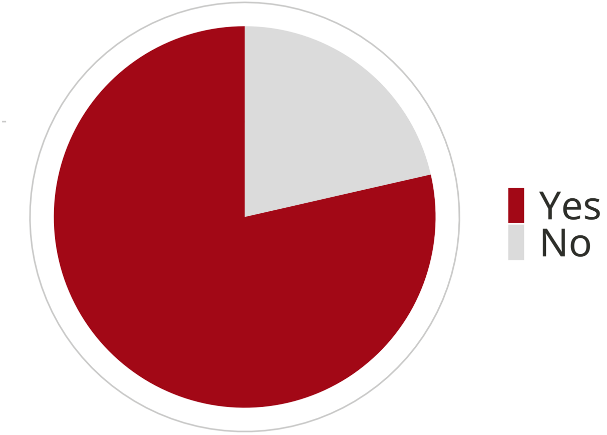
FiveThirtyEight's 2018 midterms model outcomes plot

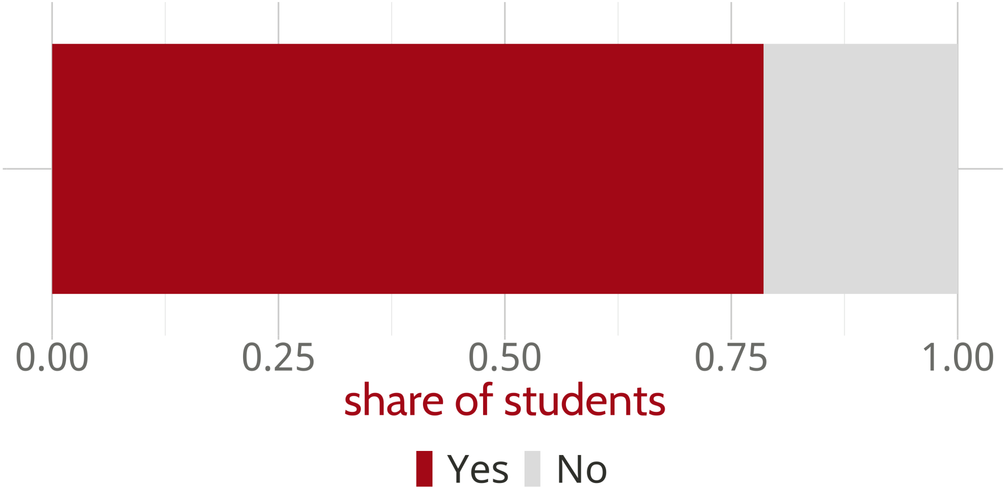
Proportions: cut for time

Can we improve this survey visualization?



Have you done any programming before?





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	✖	✖	✔

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	✔	✔	✖

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, ...)	✔	✖	✖

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	✔	✖	✔

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	✖	✖	✔

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			

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!