Introduction to Visualization

Daniel Anderson
Week 2, Class 1

Agenda

Discuss different visualizations

- Visualizing distributions
  - histograms
  - density plots
  - Empirical cumulative density plots
  - QQ plots
- Grouped data
- Visualizing amounts
  - bar plots
  - dot plots
  - heatmaps

Learning Objectives

- Understand various ways the same underlying data can be displayed
- Think through pros/cons of each
- Understand the basic structure of the code to produce the various plots

What type of data do you have?

What type of data do you have?

We'll focus primarily on standard continuous/categorical data

What type of data do you have?

We'll focus primarily on standard continuous/categorical data

What is your purpose?

What type of data do you have?

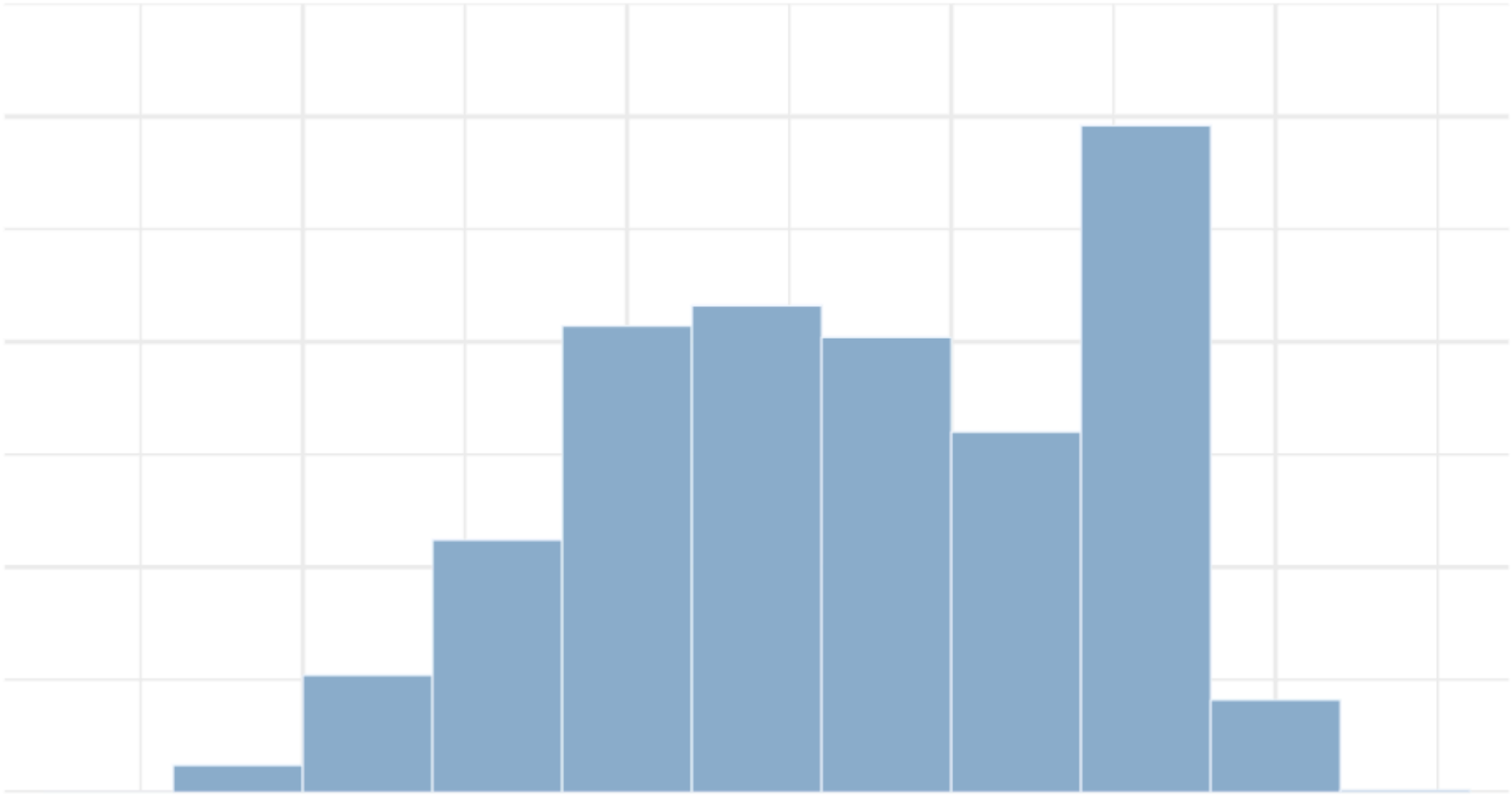
We'll focus primarily on standard continuous/categorical data

What is your purpose?

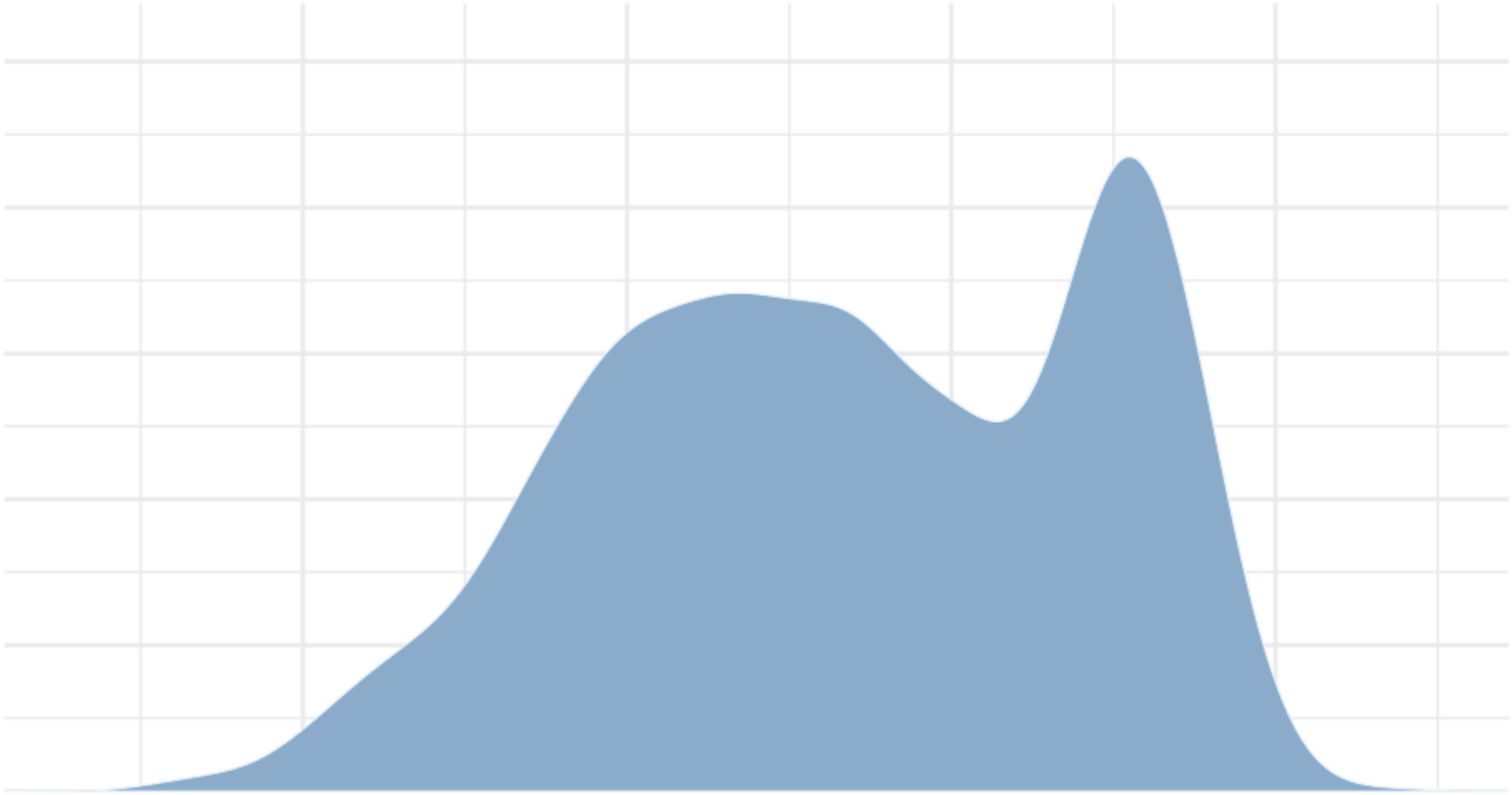
Exploratory? Communication?

One continuous variable

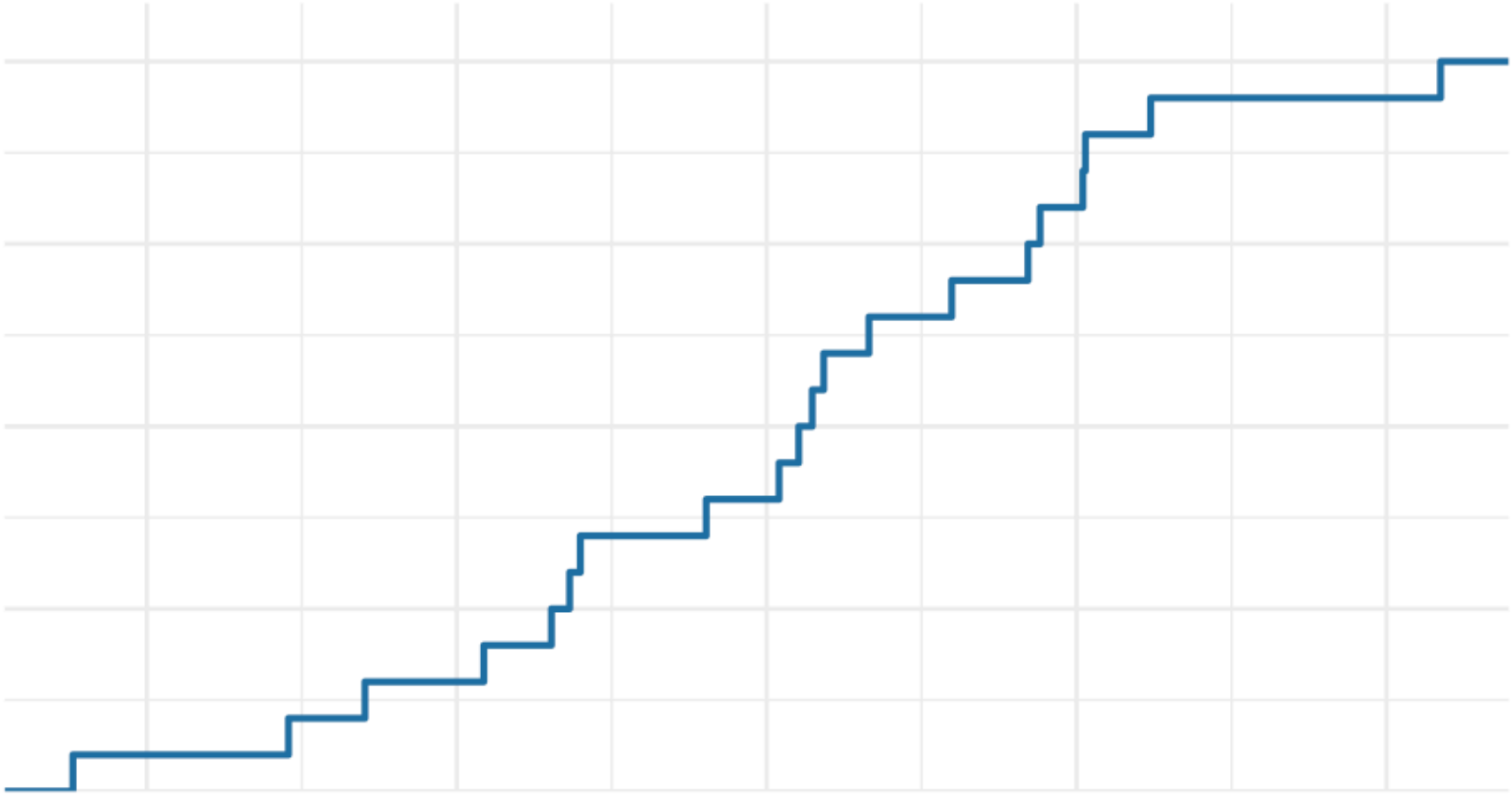
Histogram



Density plot

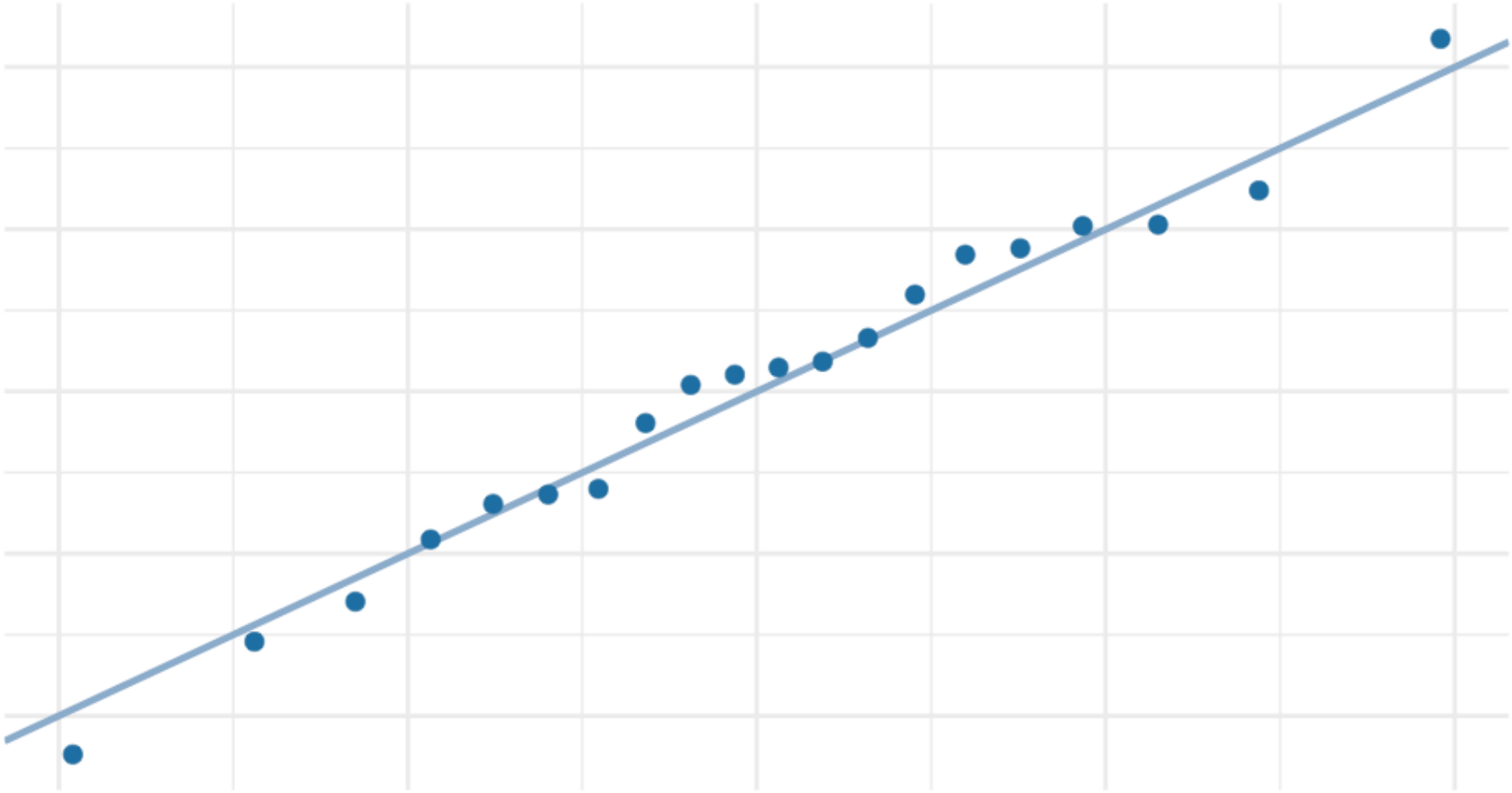


(Empirical) Cumulative Density



QQ Plot

Compare to theoretical quantiles (for normality)



Empirical examples

Titanic data

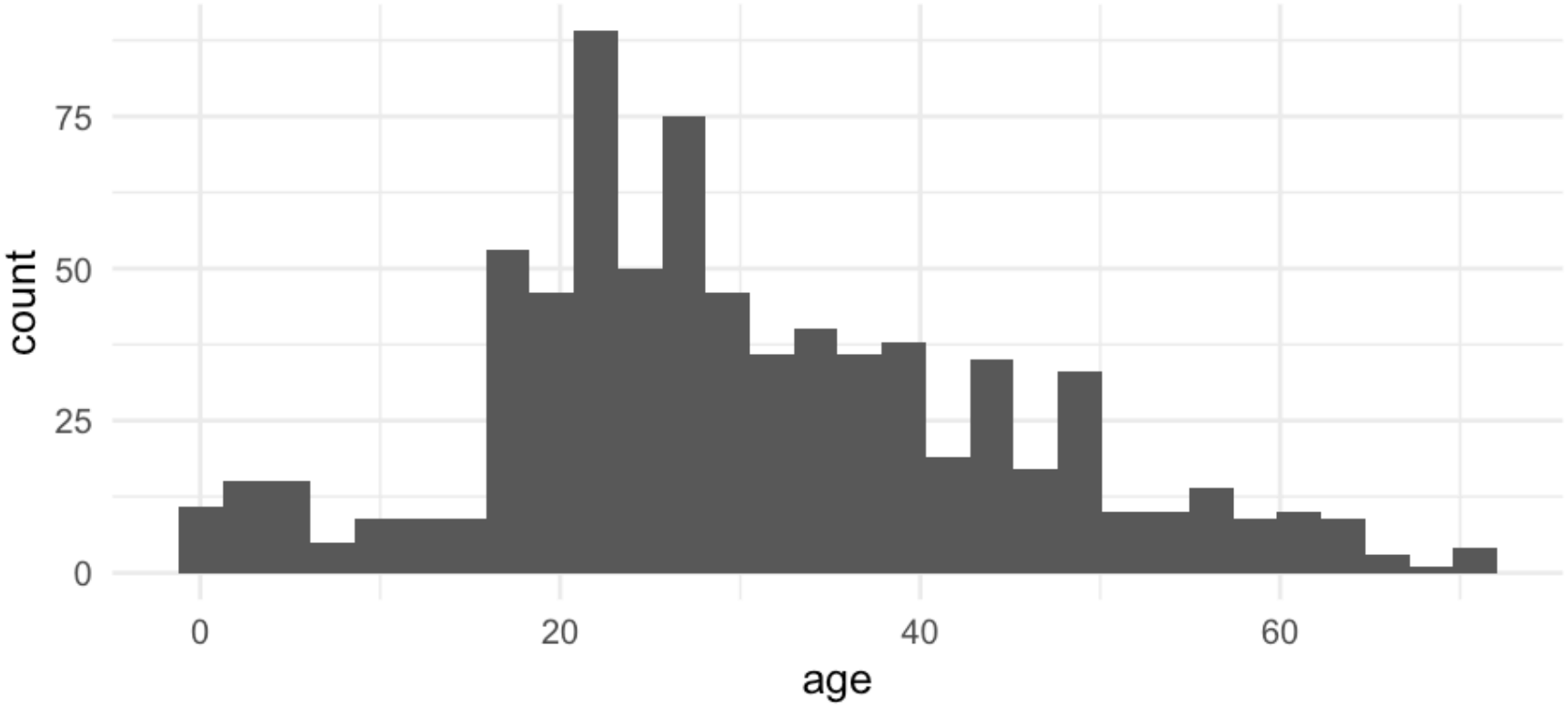
```
head(titanic)
```

```
##   class   age    sex survived
## 1   1st 29.00 female        1
## 2   1st  2.00 female        0
## 3   1st 30.00   male        0
## 4   1st 25.00 female        0
## 5   1st  0.92   male        1
## 6   1st 47.00   male        1
```

Basic histogram

```
ggplot(titanic, aes(x = age)) +
  geom_histogram()
```

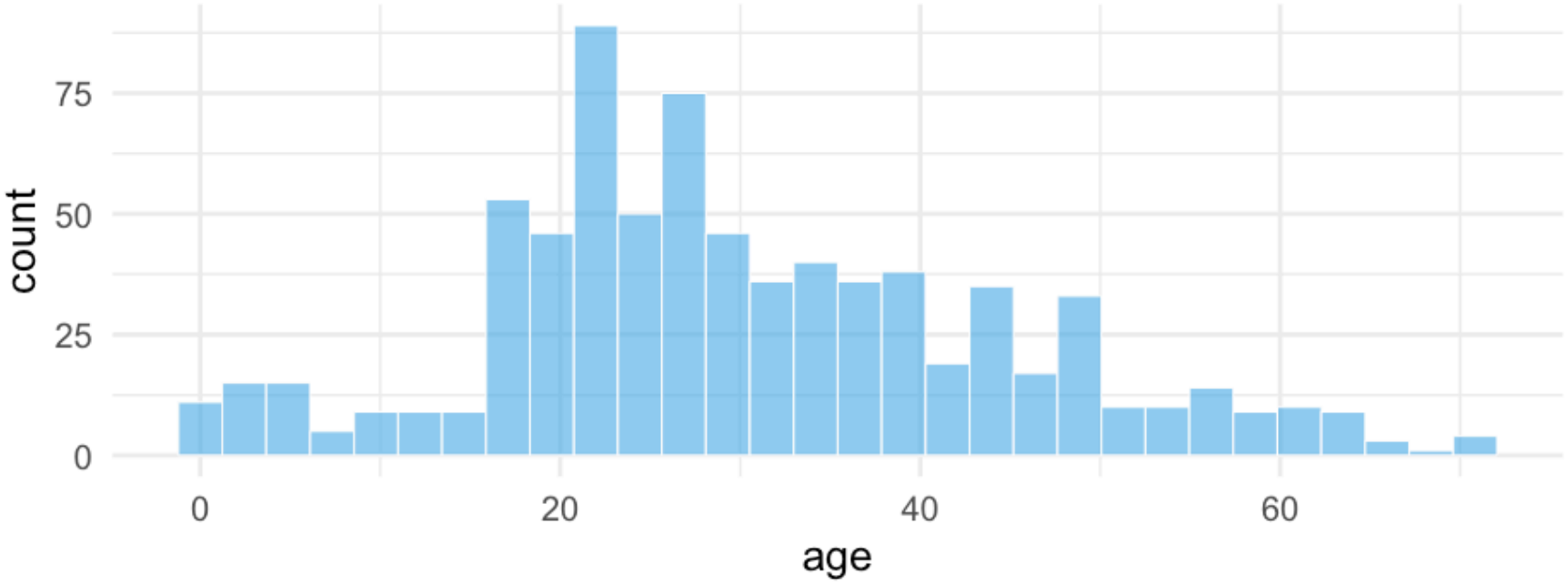
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Make it a little prettier

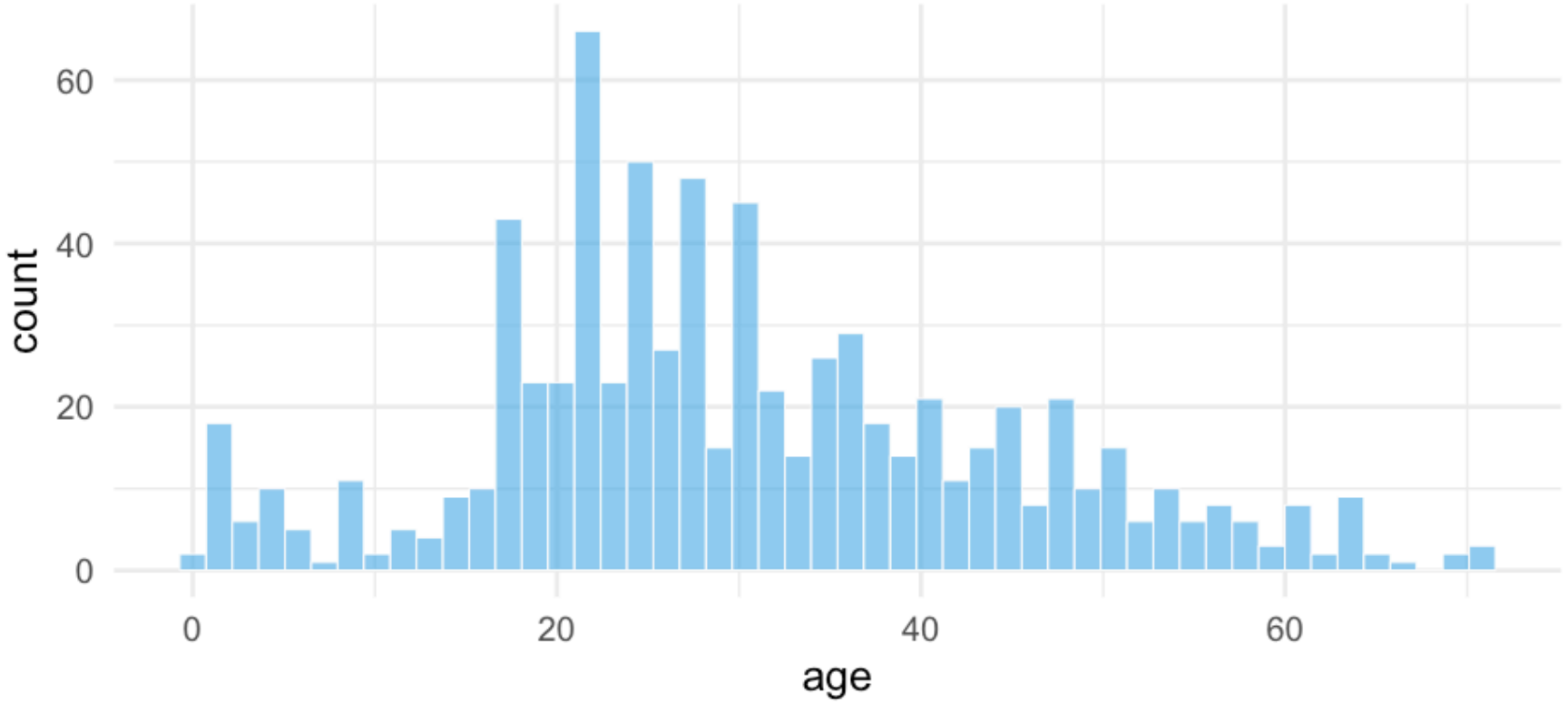
```
ggplot(titanic, aes(x = age)) +
  geom_histogram(fill = "#56B4E9",
                 color = "white",
                 alpha = 0.7)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

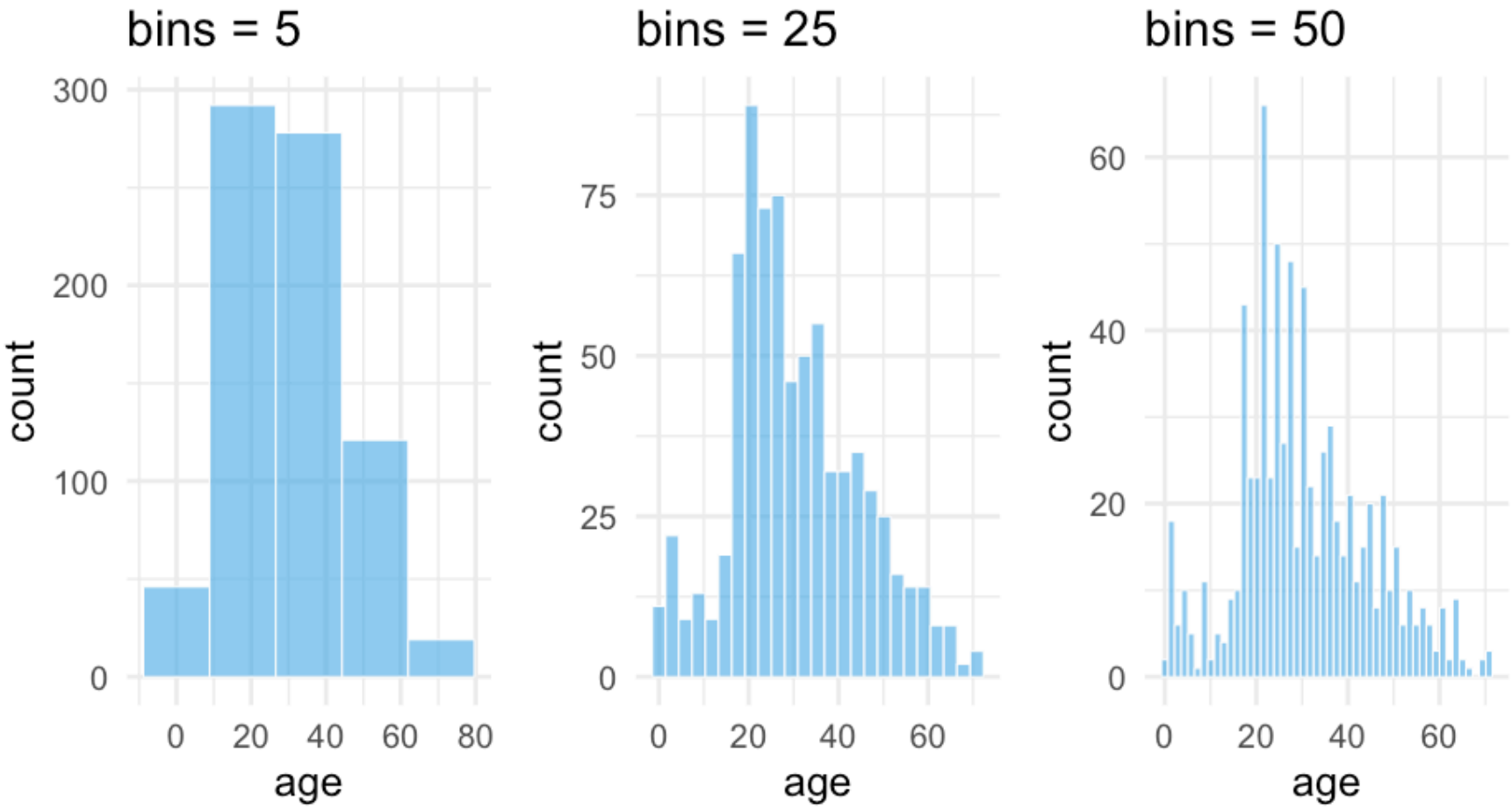


Change the number of bins

```
ggplot(titanic, aes(x = age)) +
  geom_histogram(fill = "#56B4E9",
                 color = "white",
                 alpha = 0.7,
                 bins = 50)
```



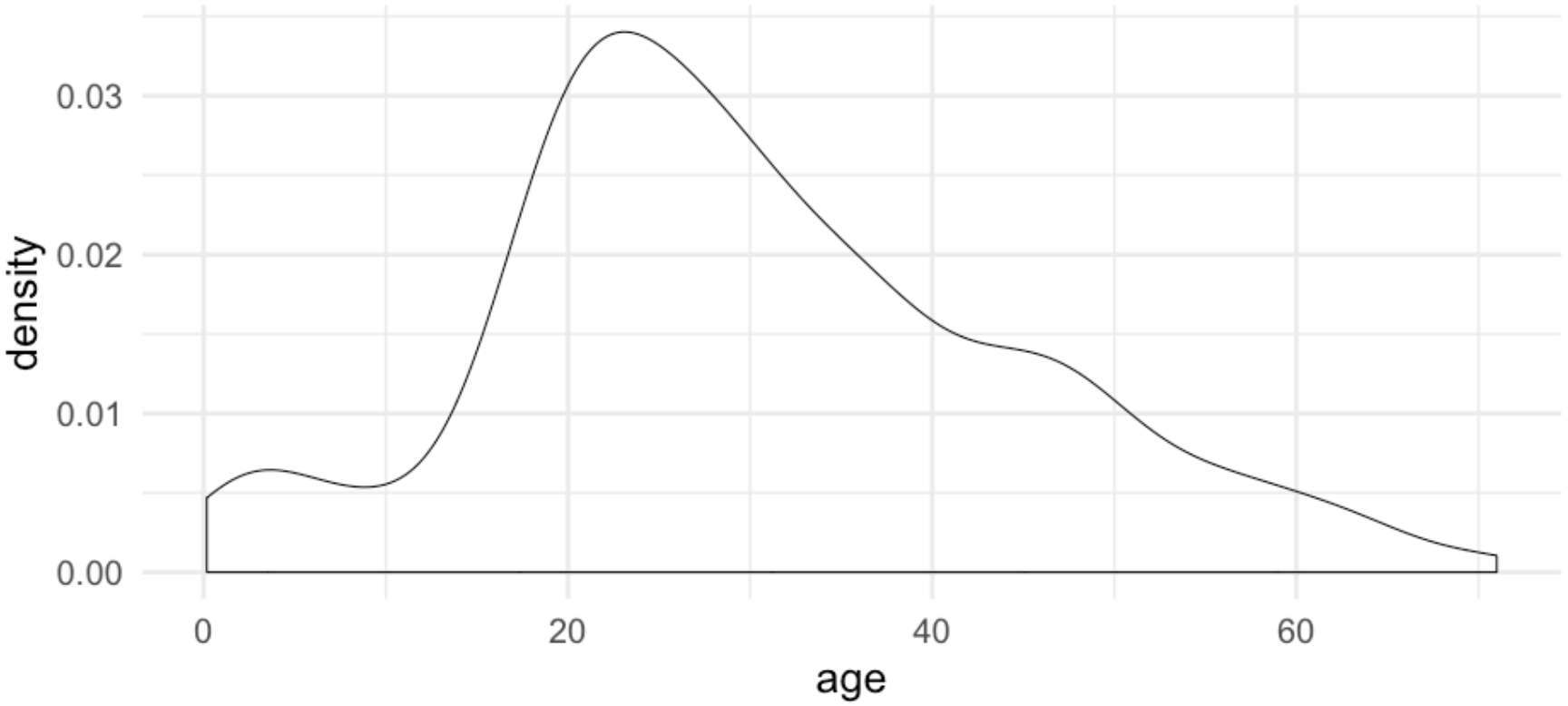
Vary the number of bins



Denisty plot

ugly 😫

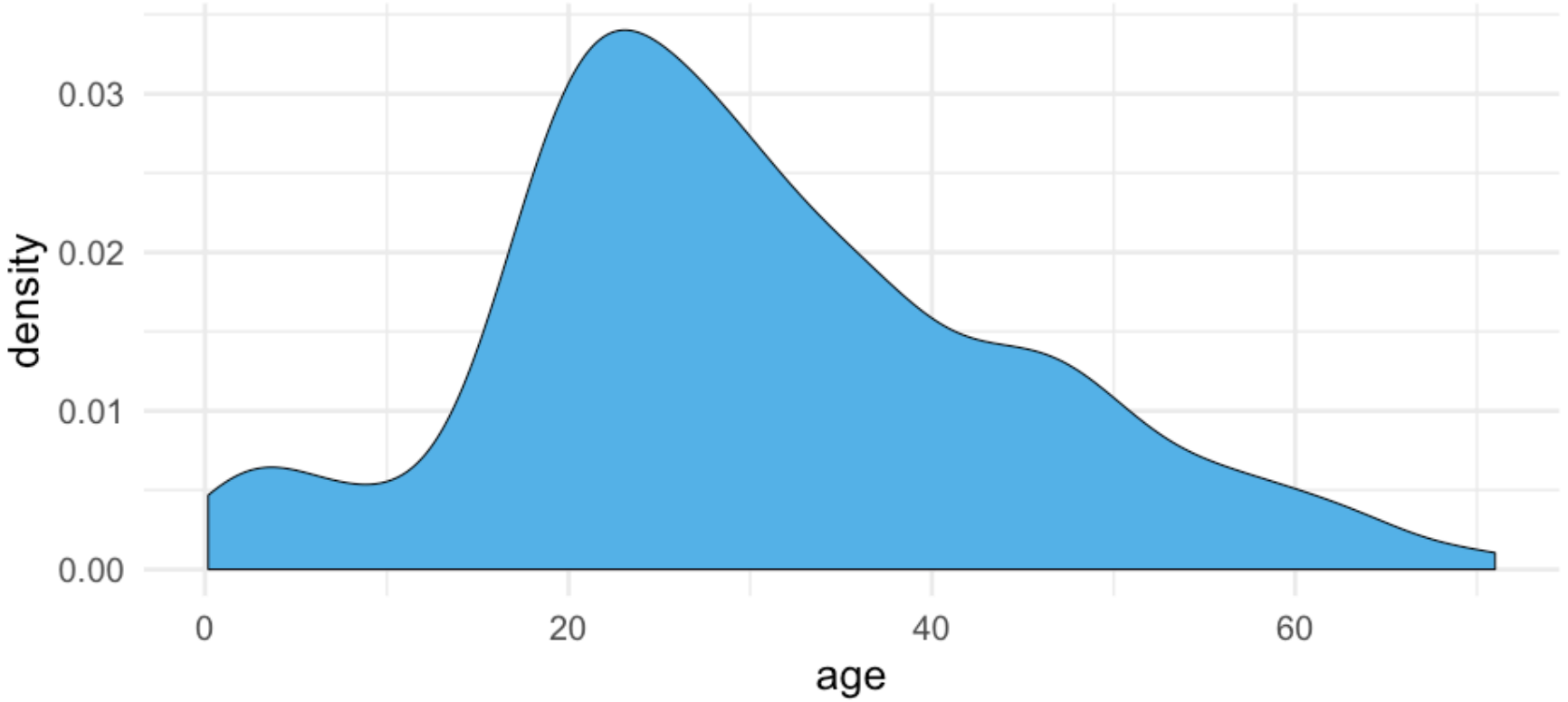
```
ggplot(titanic, aes(age)) +
  geom_density()
```



Denisty plot

change the fill

```
ggplot(titanic, aes(age)) +
  geom_density(fill = "#56B4E9")
```



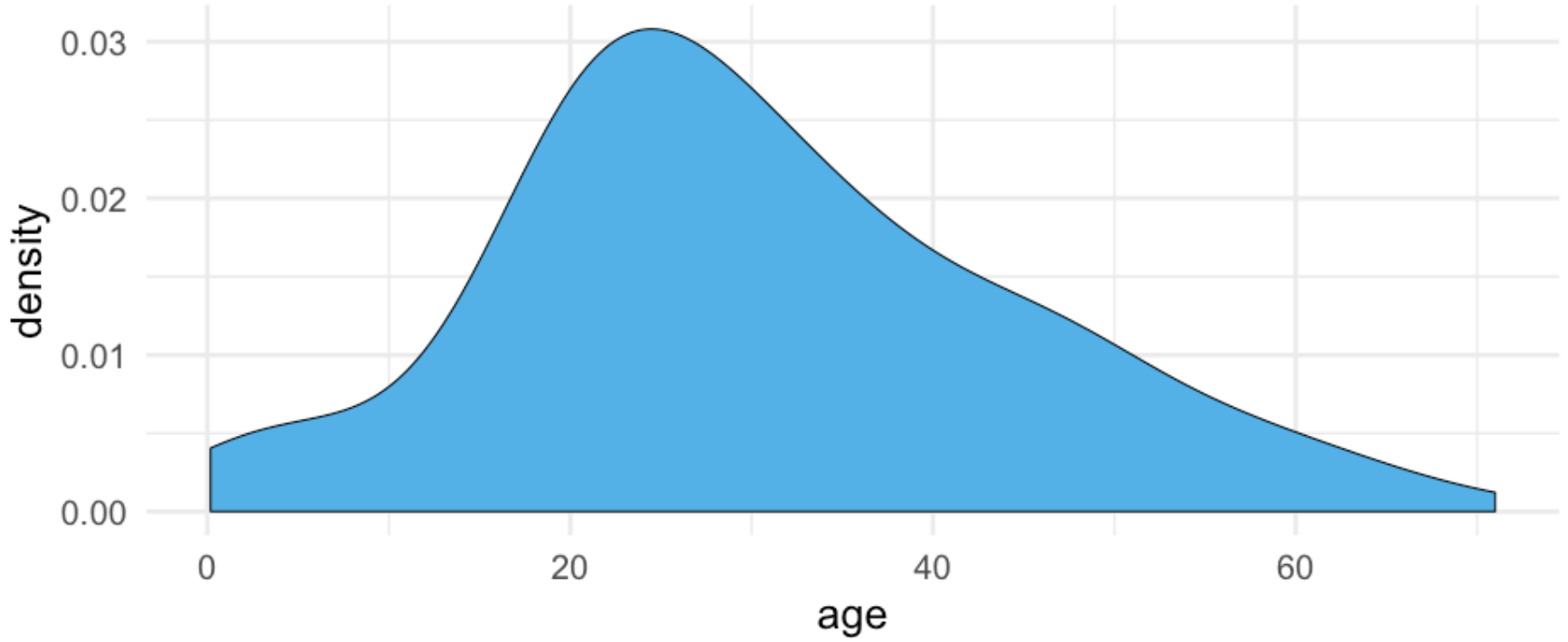
Density plot estimation

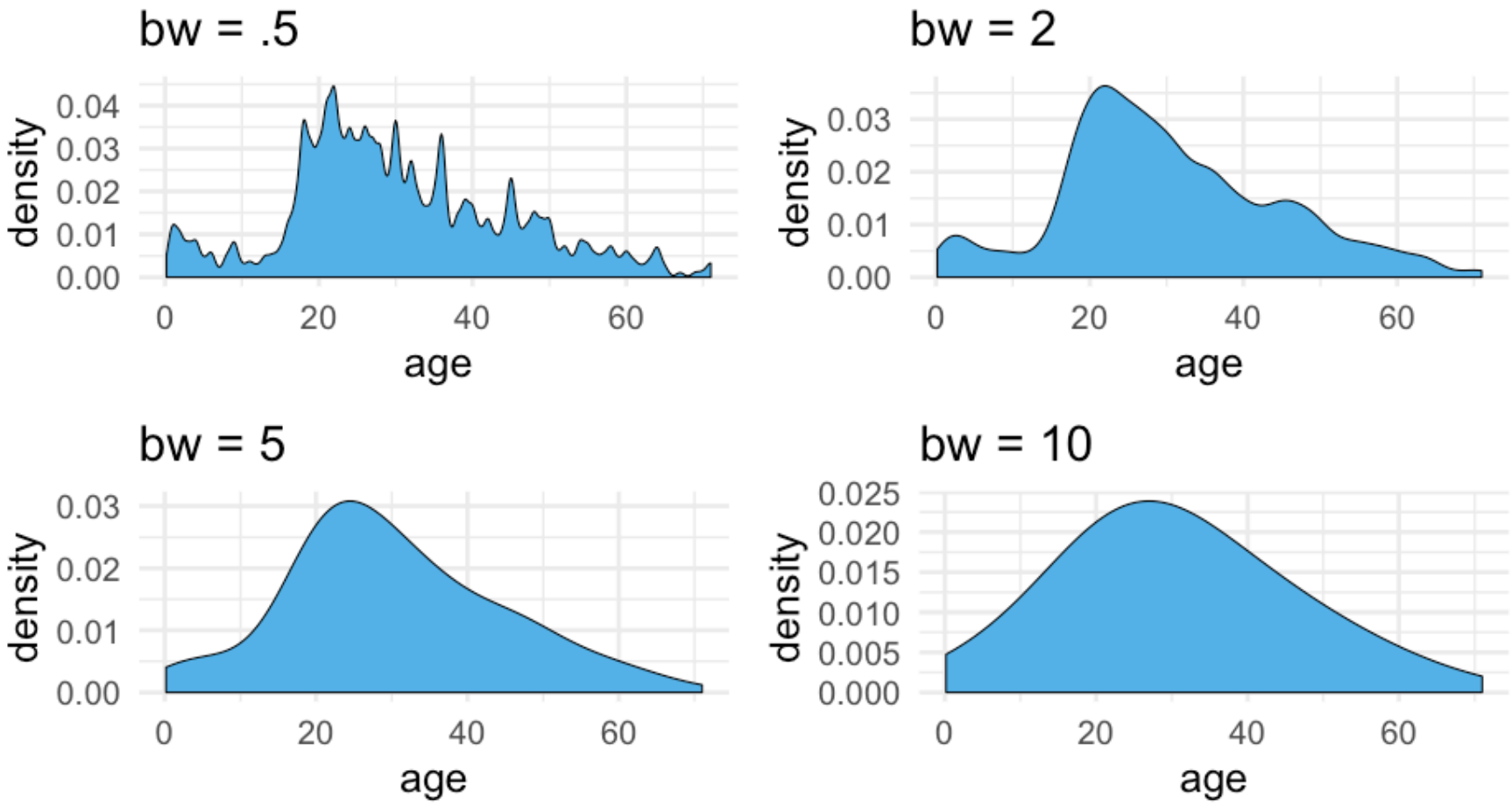
- Kernal density estimation
  - Different kernal shapes can be selected
  - Bandwidth matters most
  - Smaller bands = bend more to the data
- Approximation of the underlying continuous probability function
  - Integrates to 1.0 (y-axis is somewhat difficult to interpret)

Denisty plot

change the bandwidth

```
ggplot(titanic, aes(age)) +
  geom_density(fill = "#56B4E9",
               bw = 5)
```

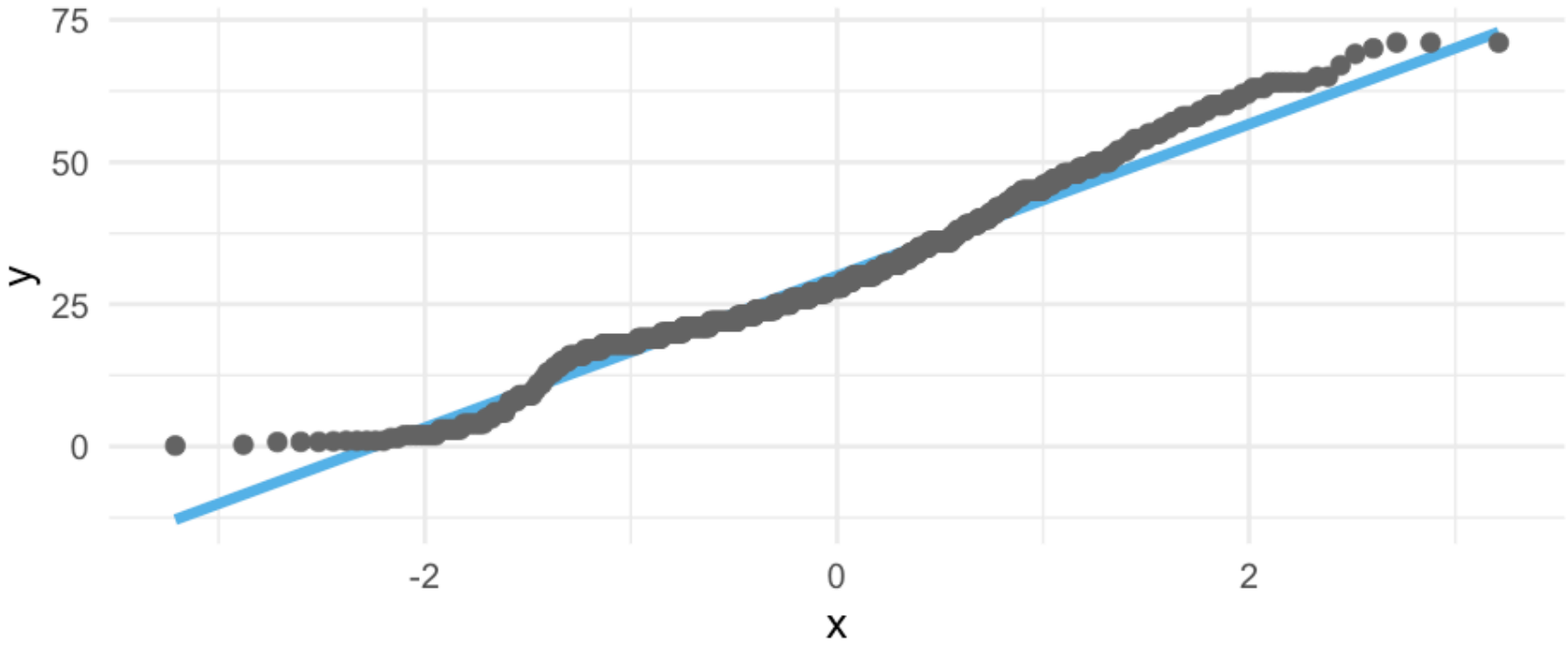



bw = .5
density
0.04
0.03
0.02
0.01
0.00
0
20
40
60
age
bw = 2
density
0.03
0.02
0.01
0.00
0
20
40
60
age
bw = 5
density
0.03
0.02
0.01
0.00
0
20
40
60
age
bw = 10
density
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
age

Quickly

How well does it approximate a normal distribution?

```
ggplot(titanic, aes(sample = age)) +
  stat_qq_line(color = "#56B4E9") +
  geom_qq(color = "gray40")
```



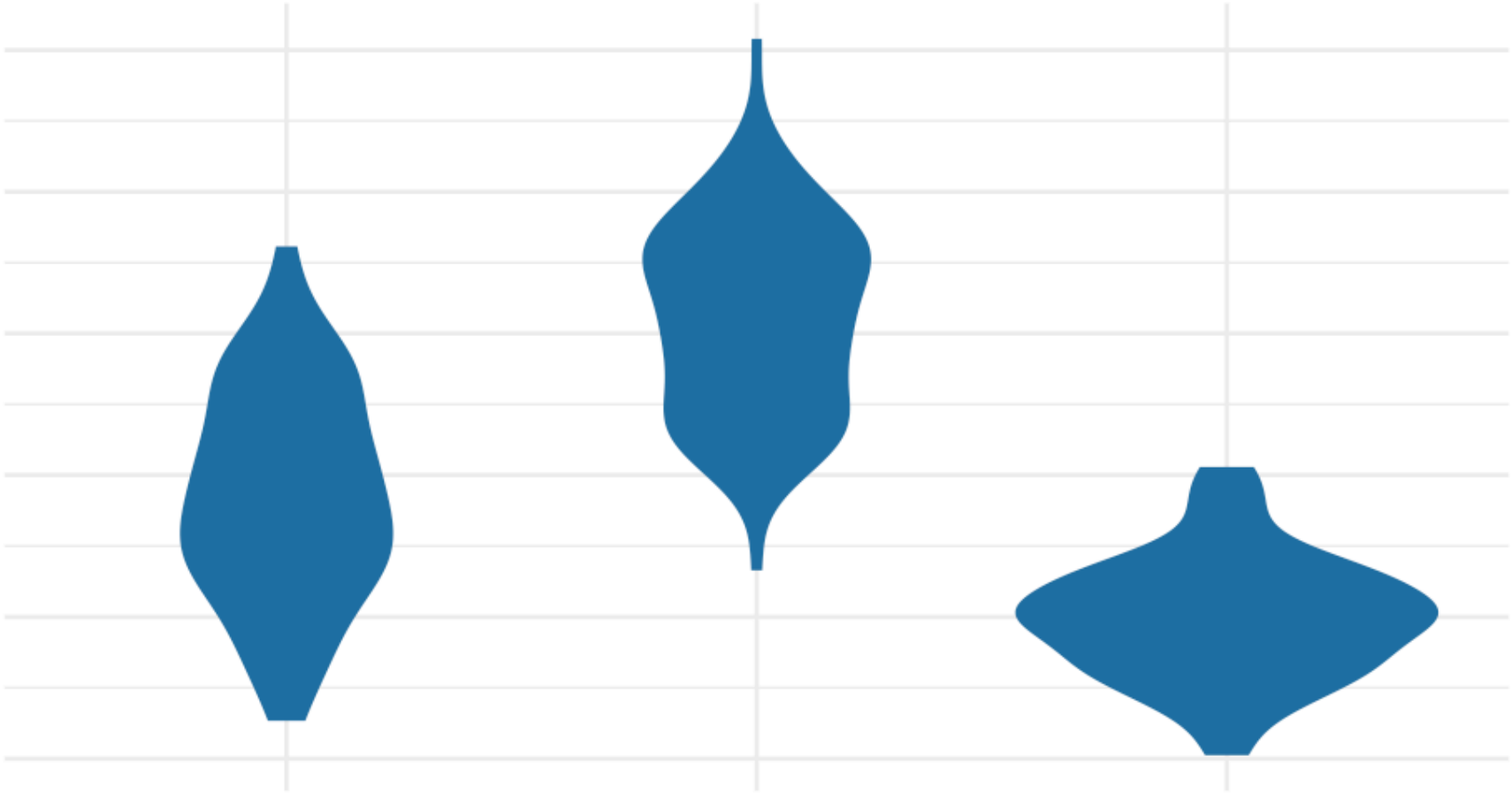
Grouped data

Distributions

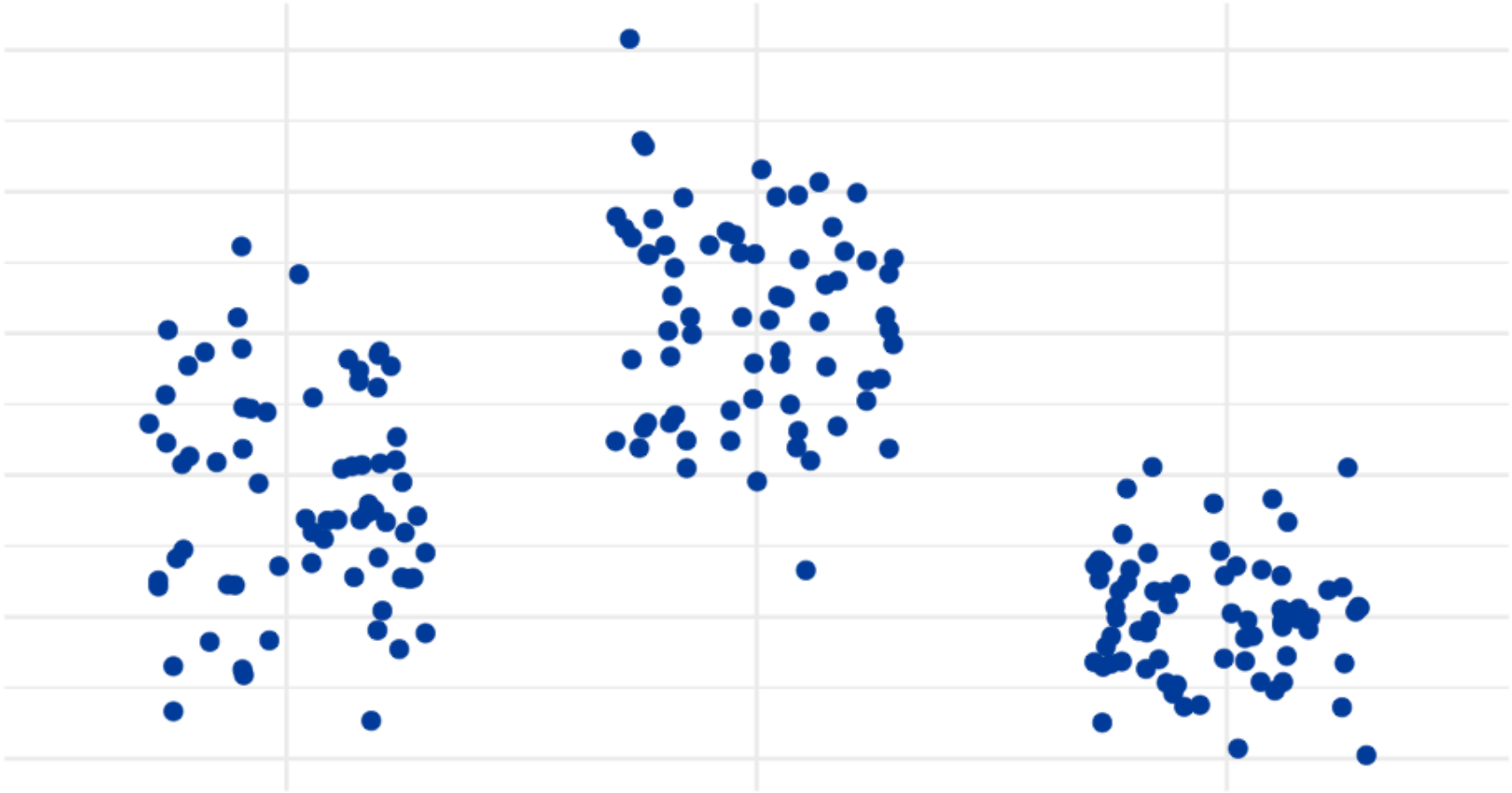
How do we display more than one distribution at a time?

Boxplots

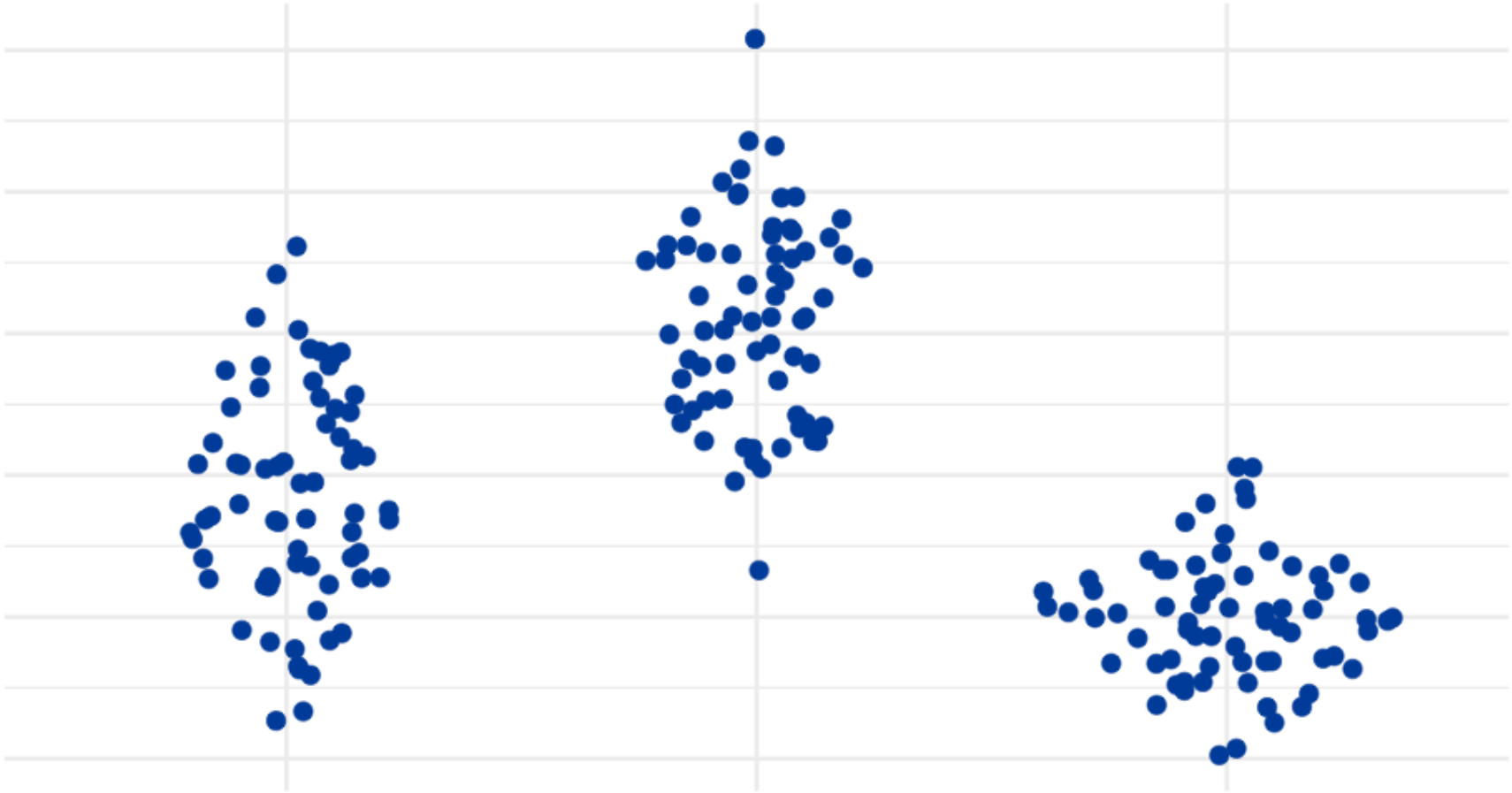
Violin plots



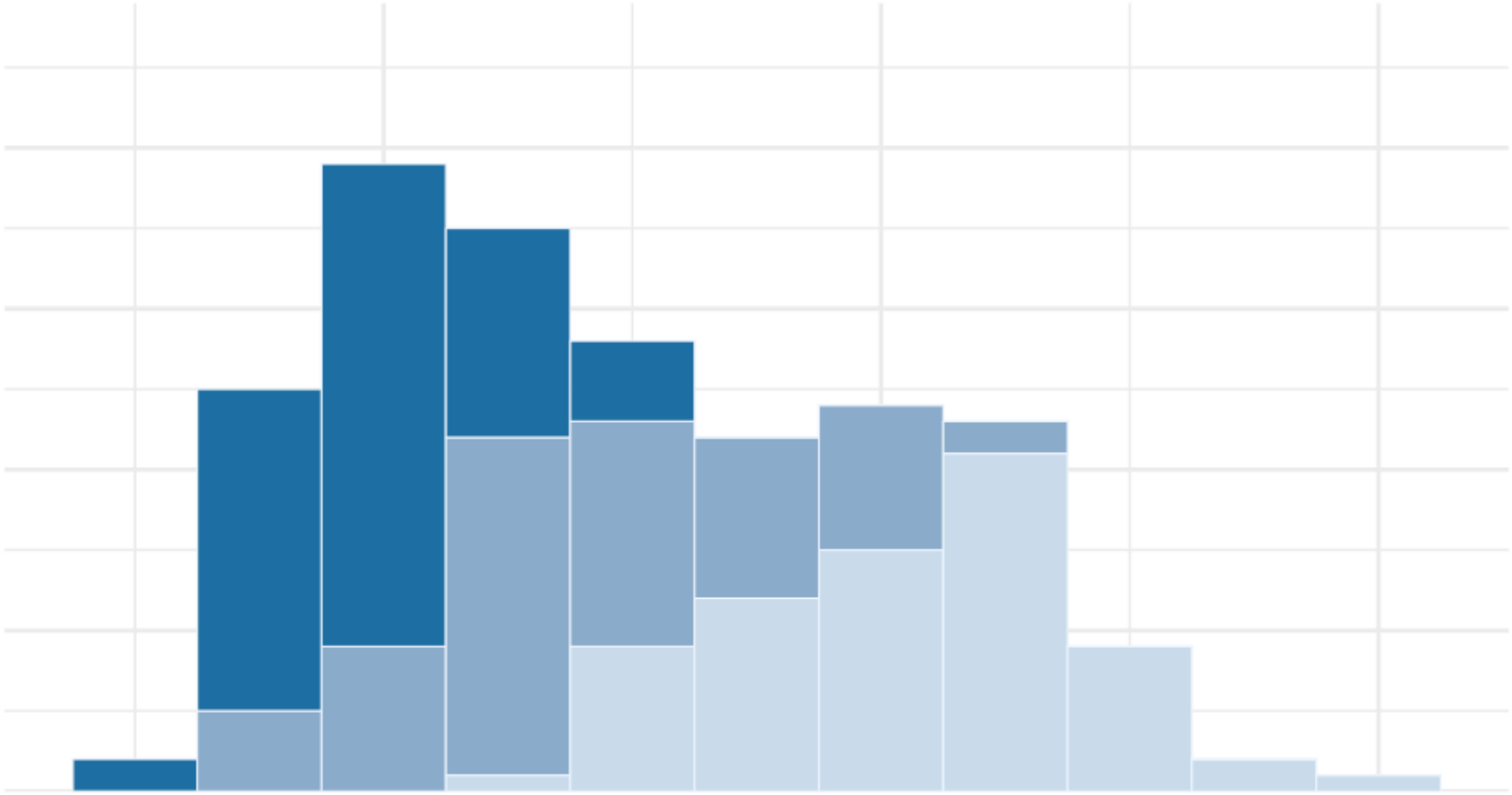
Jittered points



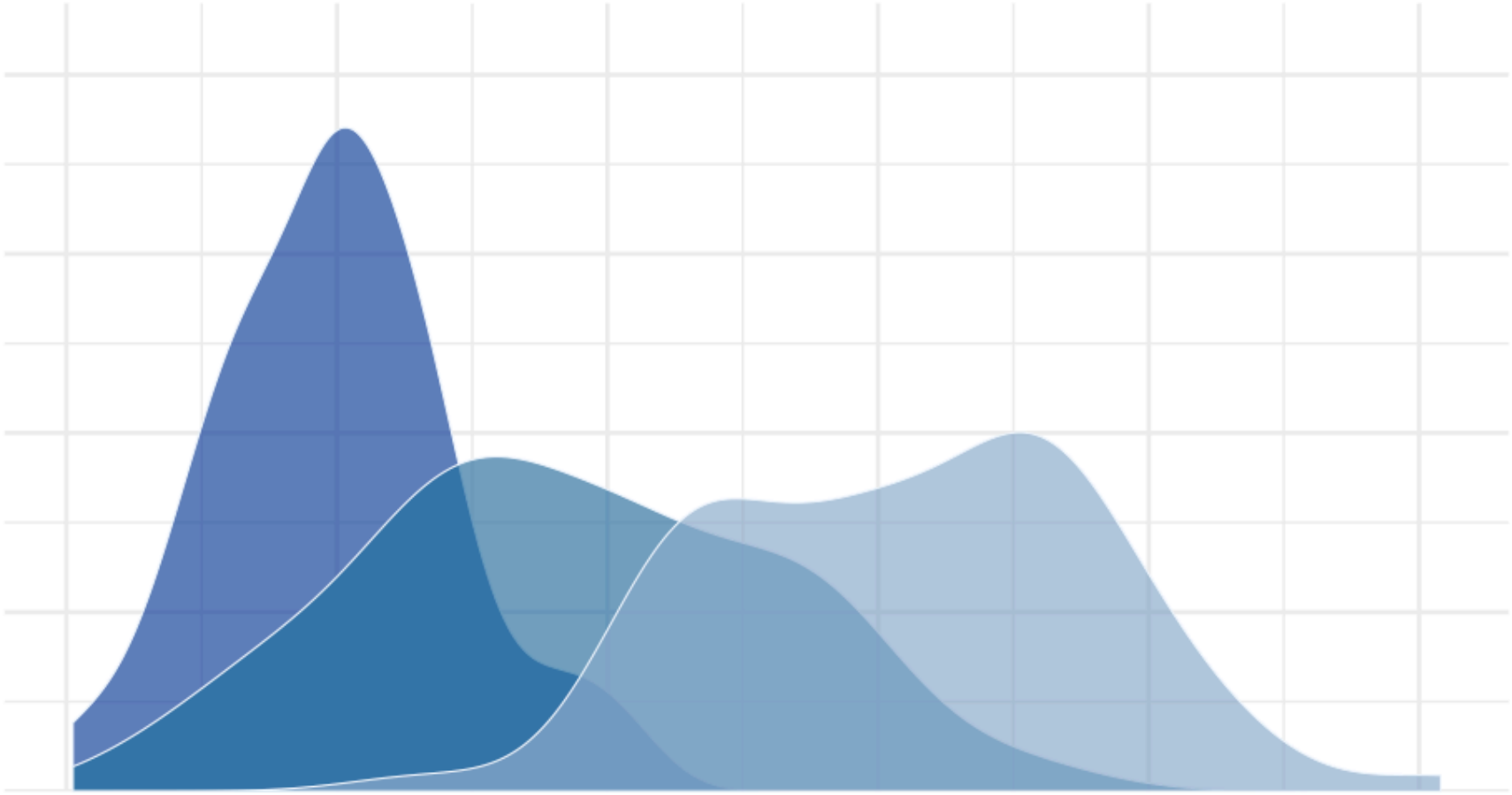
Sina plots



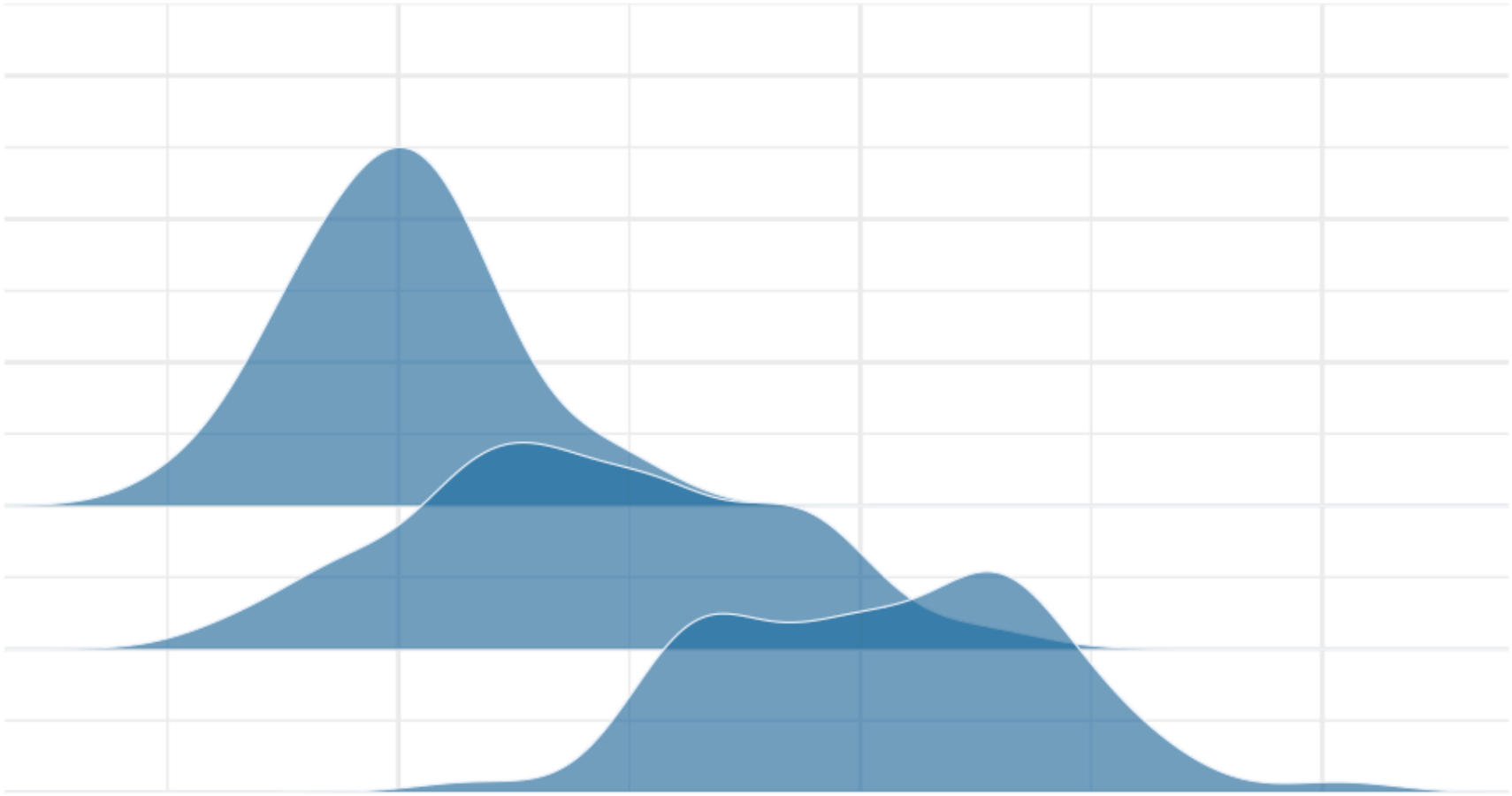
Stacked histograms



Overlapping densities



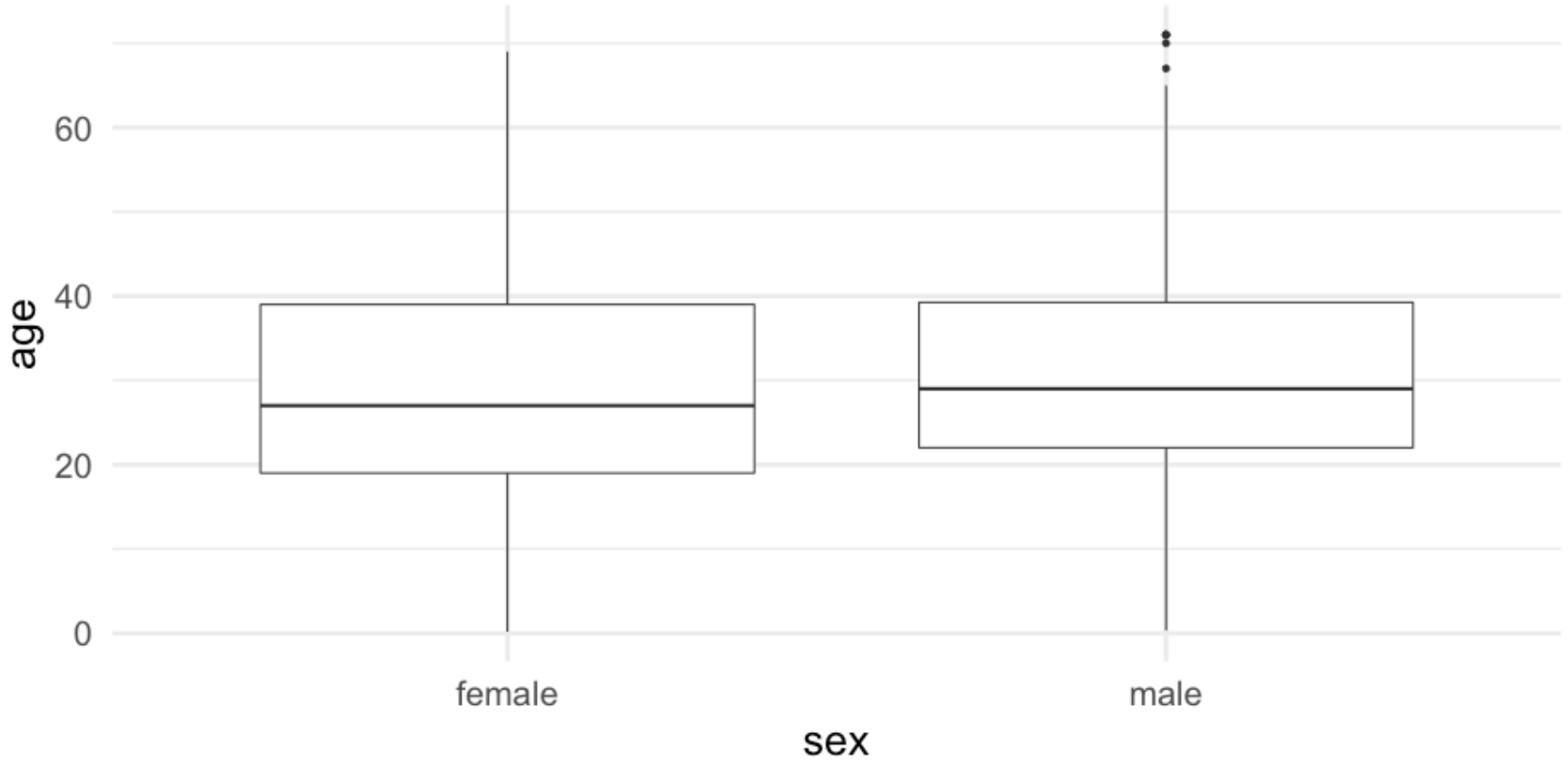
Ridgeline densities



Quick empirical examples

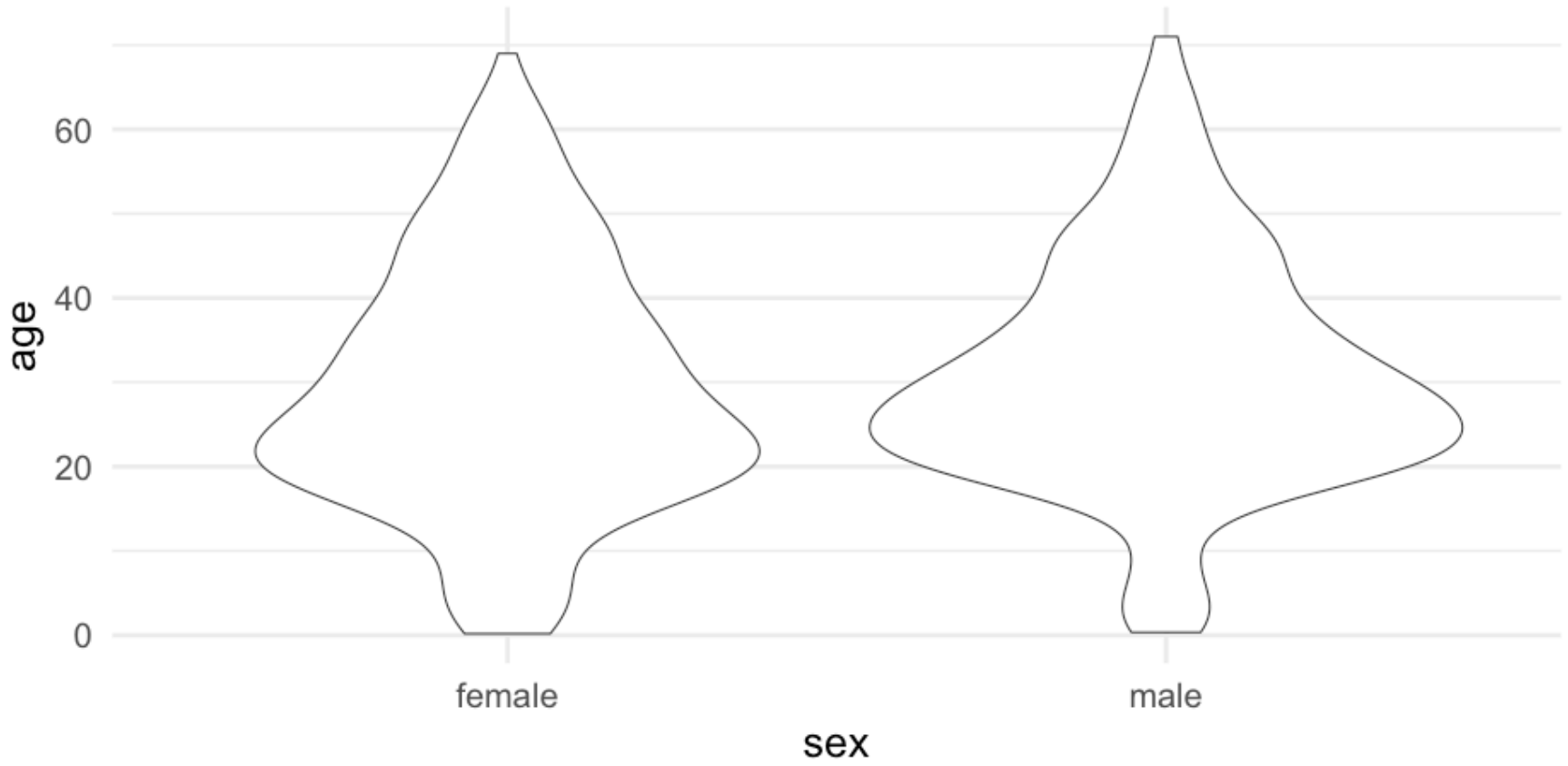
Boxplots

```
ggplot(titanic, aes(sex, age)) +
  geom_boxplot()
```



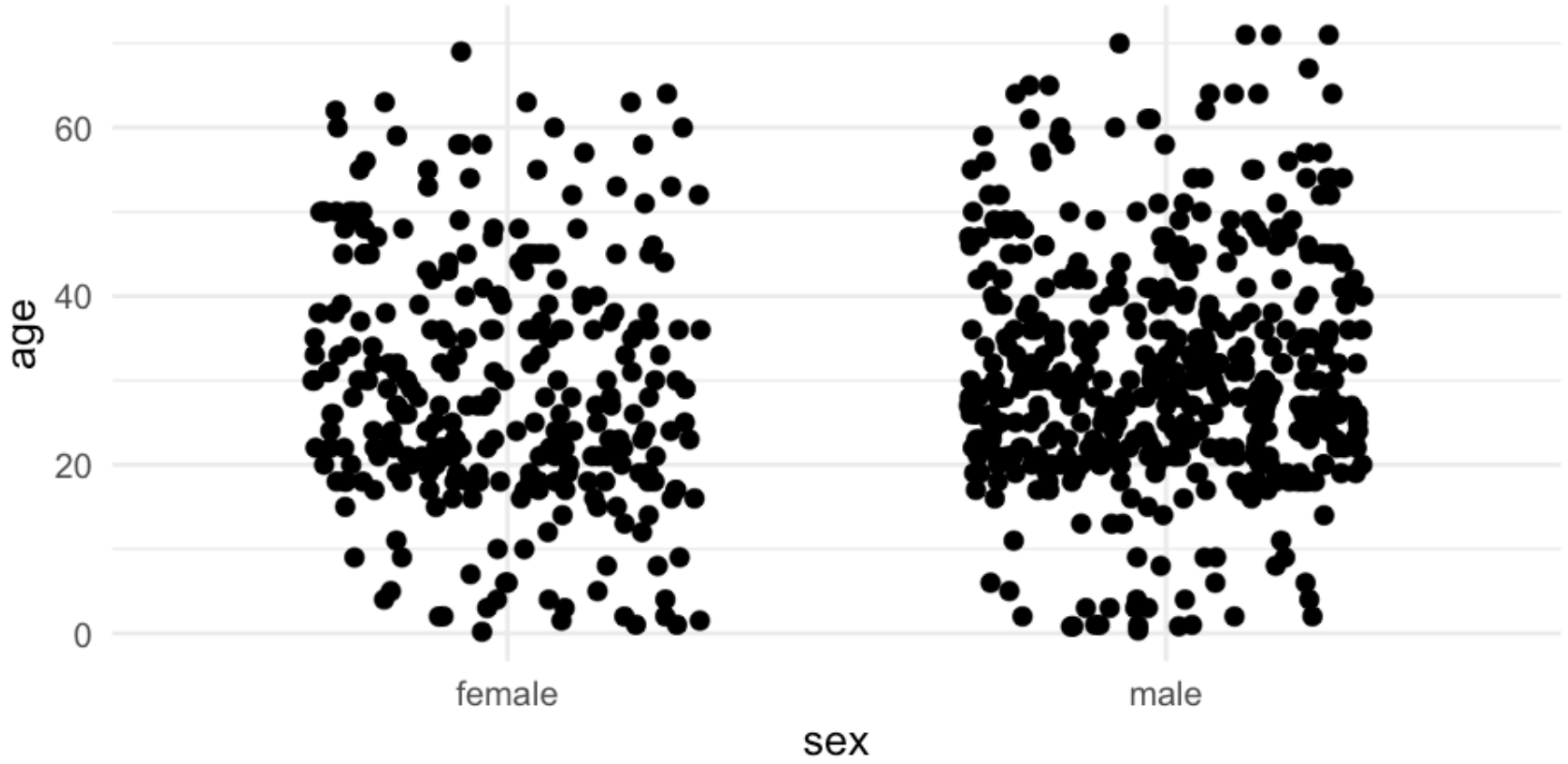
Violin plots

```
ggplot(titanic, aes(sex, age)) +
  geom_violin()
```



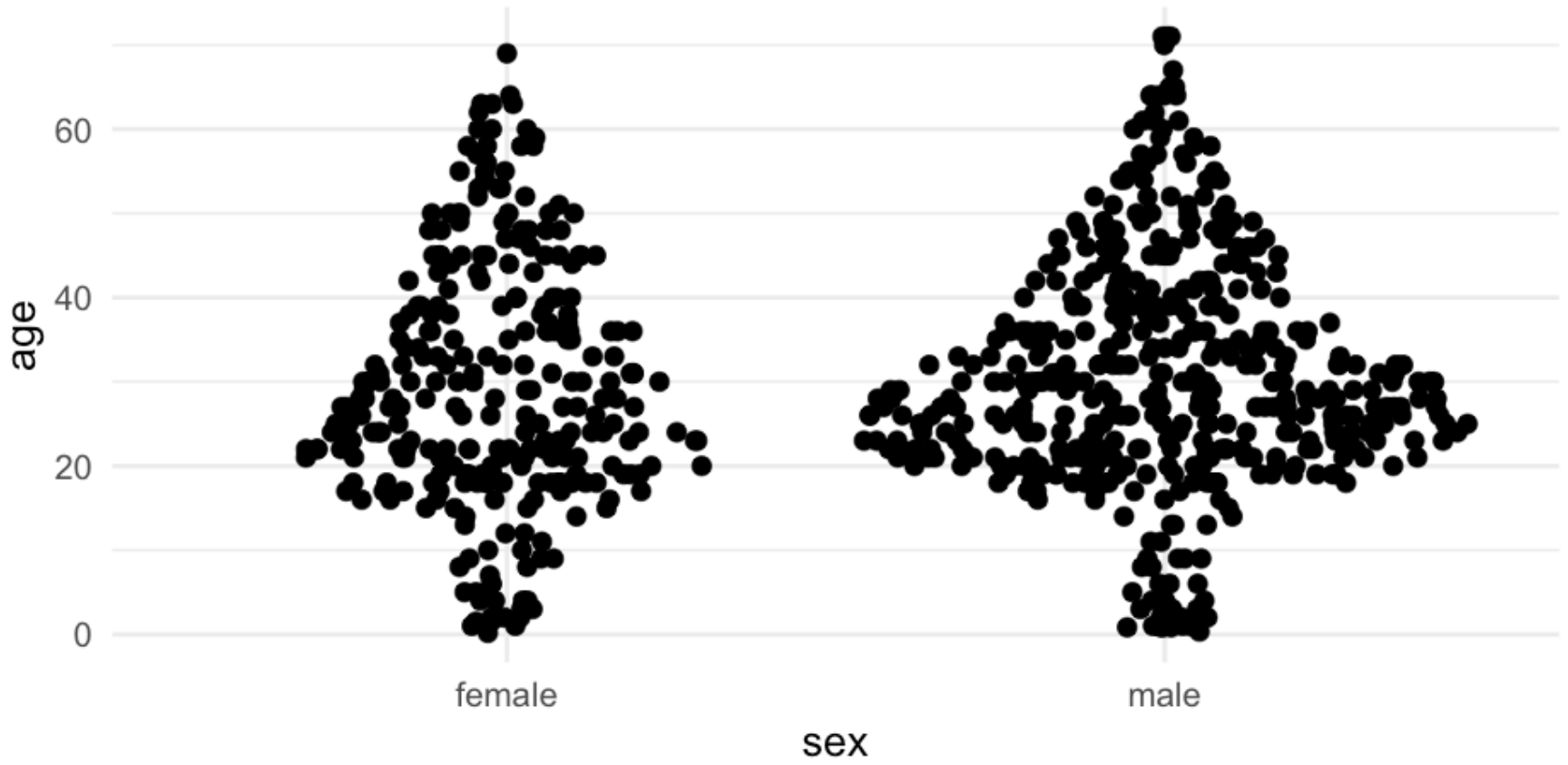
Jitterd point plots

```
ggplot(titanic, aes(sex, age)) +
  geom_jitter(width = 0.3)
```



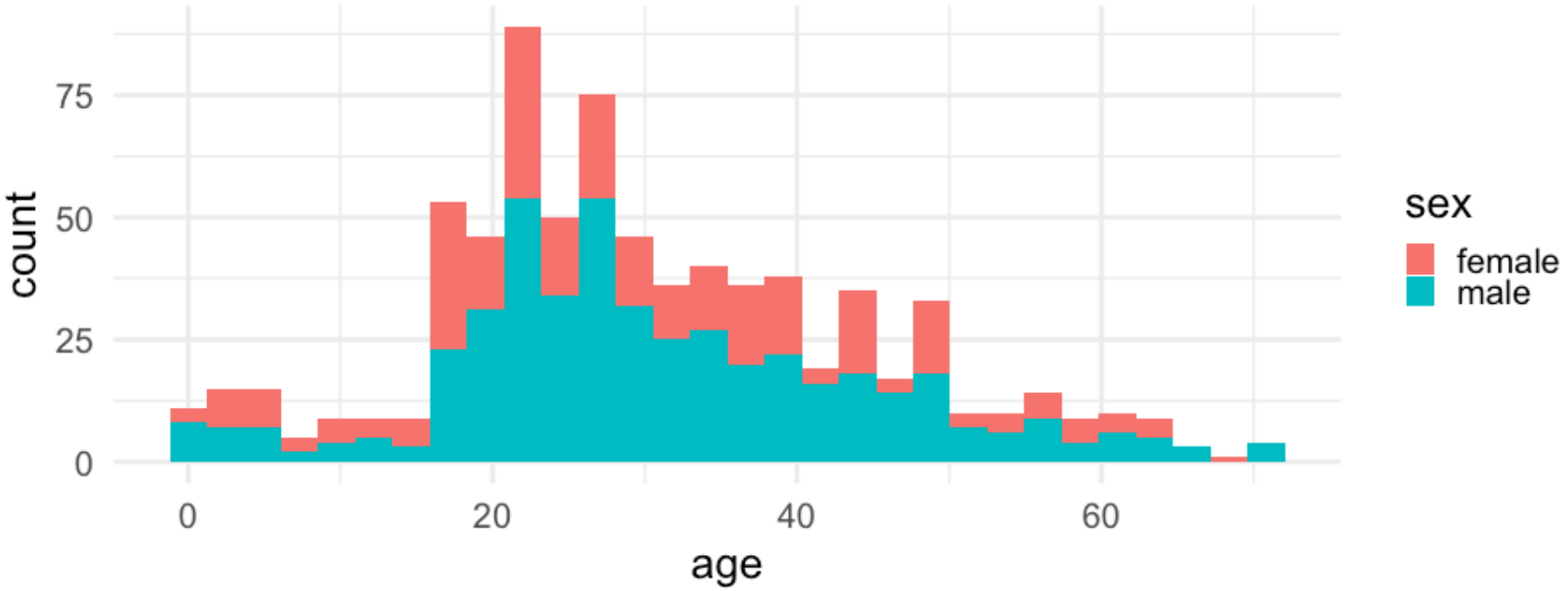
Sina plot

```
ggplot(titanic, aes(sex, age)) +
  ggforce::geom_sina()
```



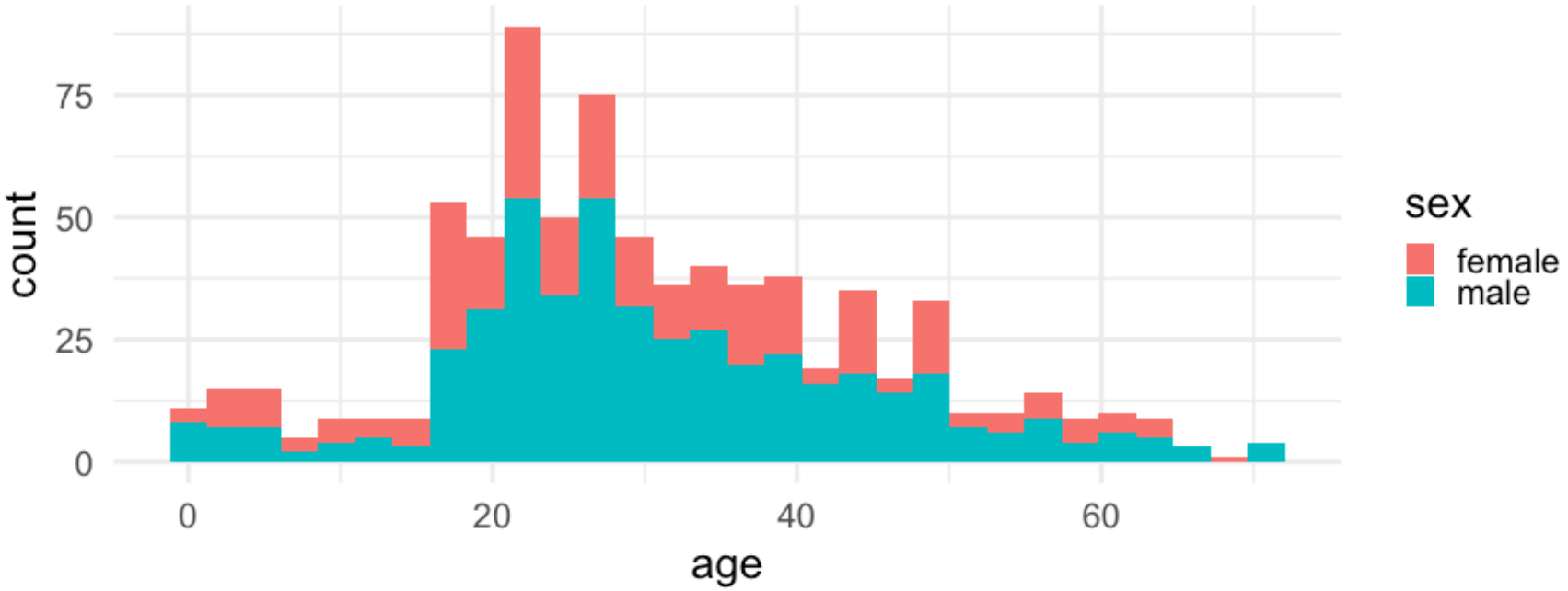
Stacked histogram

```
ggplot(titanic, aes(age)) +
  geom_histogram(aes(fill = sex))
```



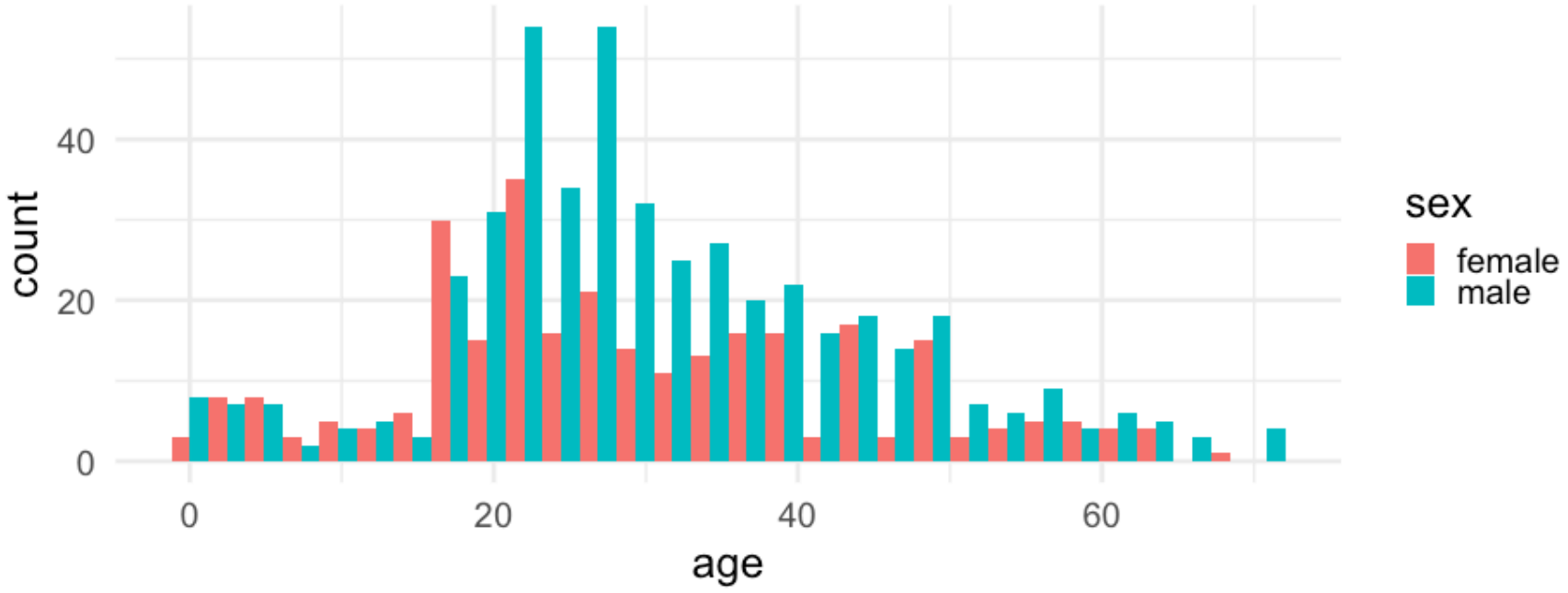
Stacked histogram

```
ggplot(titanic, aes(age)) +
  geom_histogram(aes(fill = sex))
```



Dodged

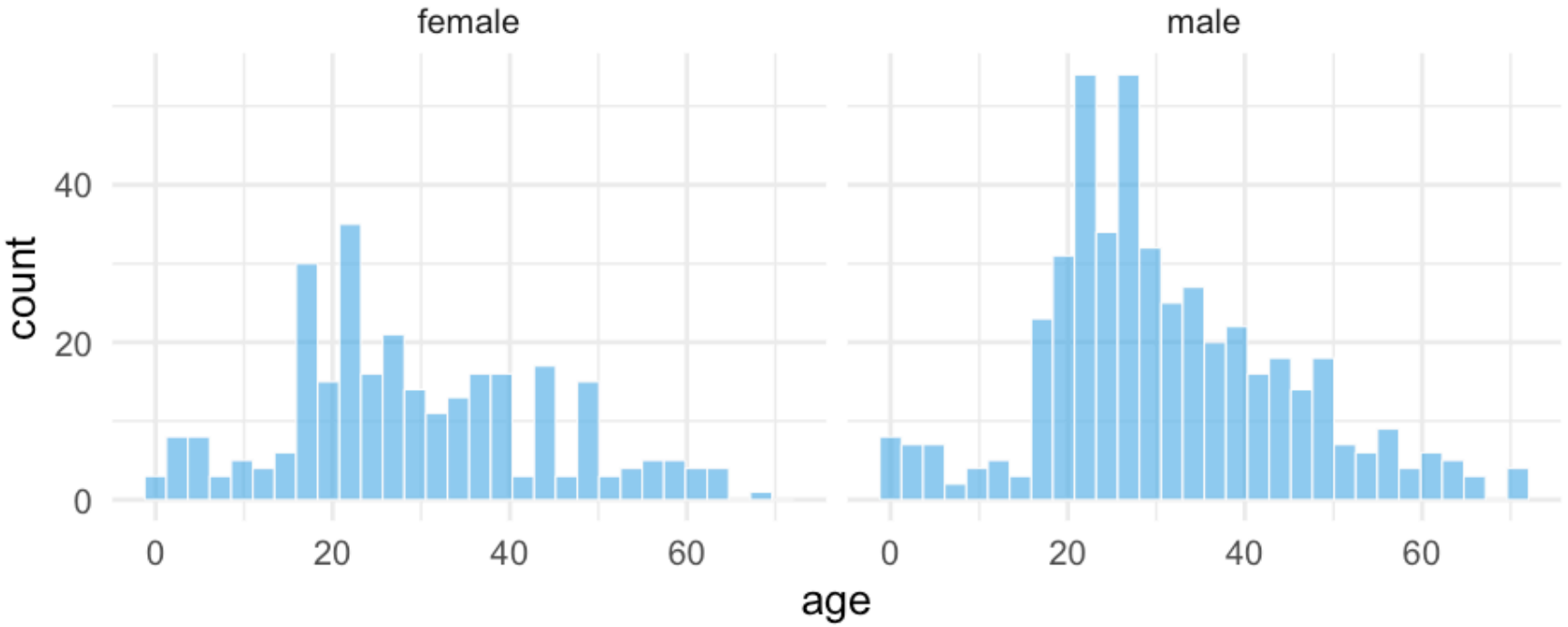
```
ggplot(titanic, aes(age)) +
  geom_histogram(aes(fill = sex),
                 position = "dodge")
```



Note `position = "dodge"` does not go into `aes` (not accessing a variable in your dataset)

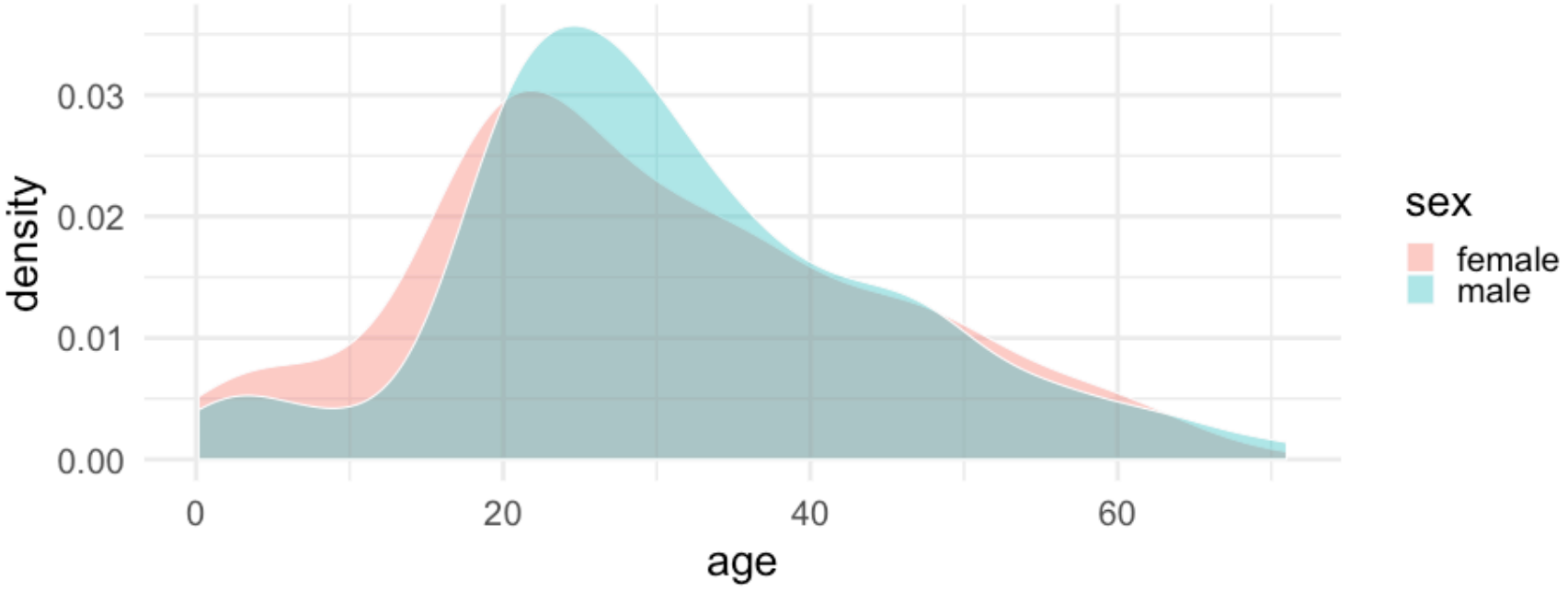
Better

```
ggplot(titanic, aes(age)) +
  geom_histogram(fill = "#56B4E9",
                 color = "white",
                 alpha = 0.7,) +
  facet_wrap(~sex)
```



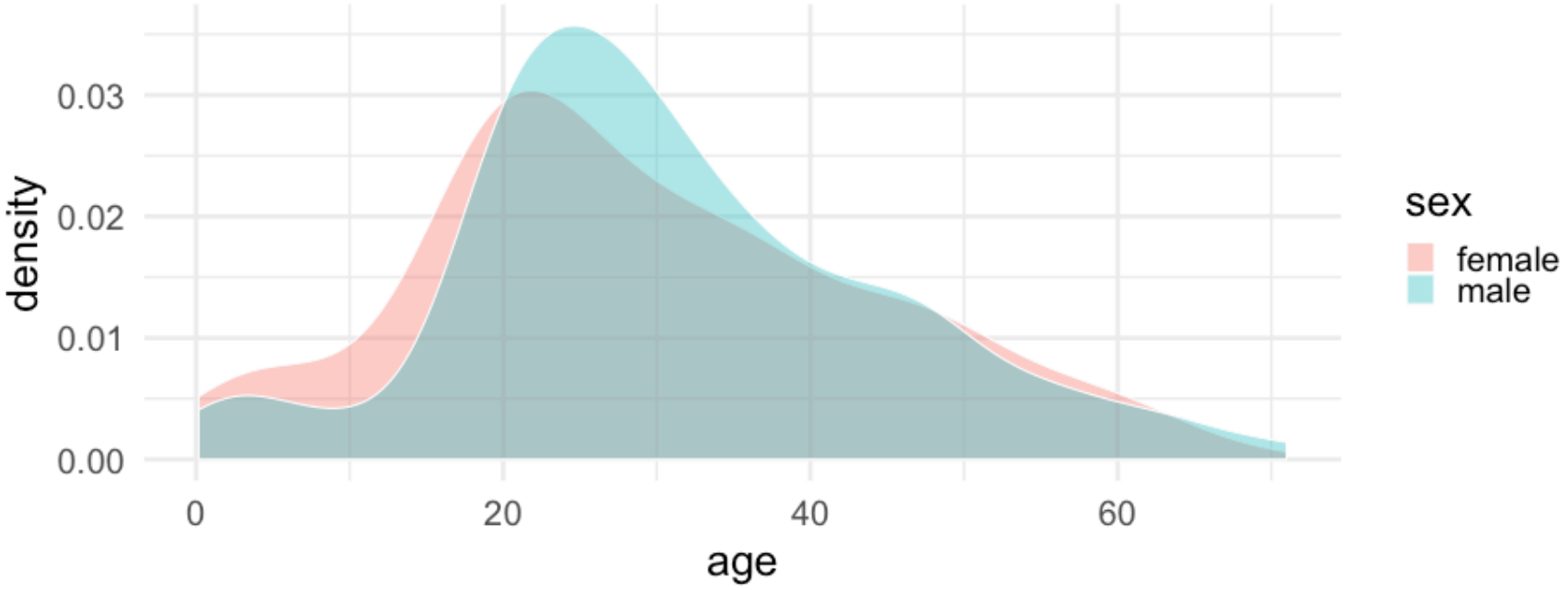
Overlapping densities

```
ggplot(titanic, aes(age)) +
  geom_density(aes(fill = sex),
               color = "white",
               alpha = 0.4)
```



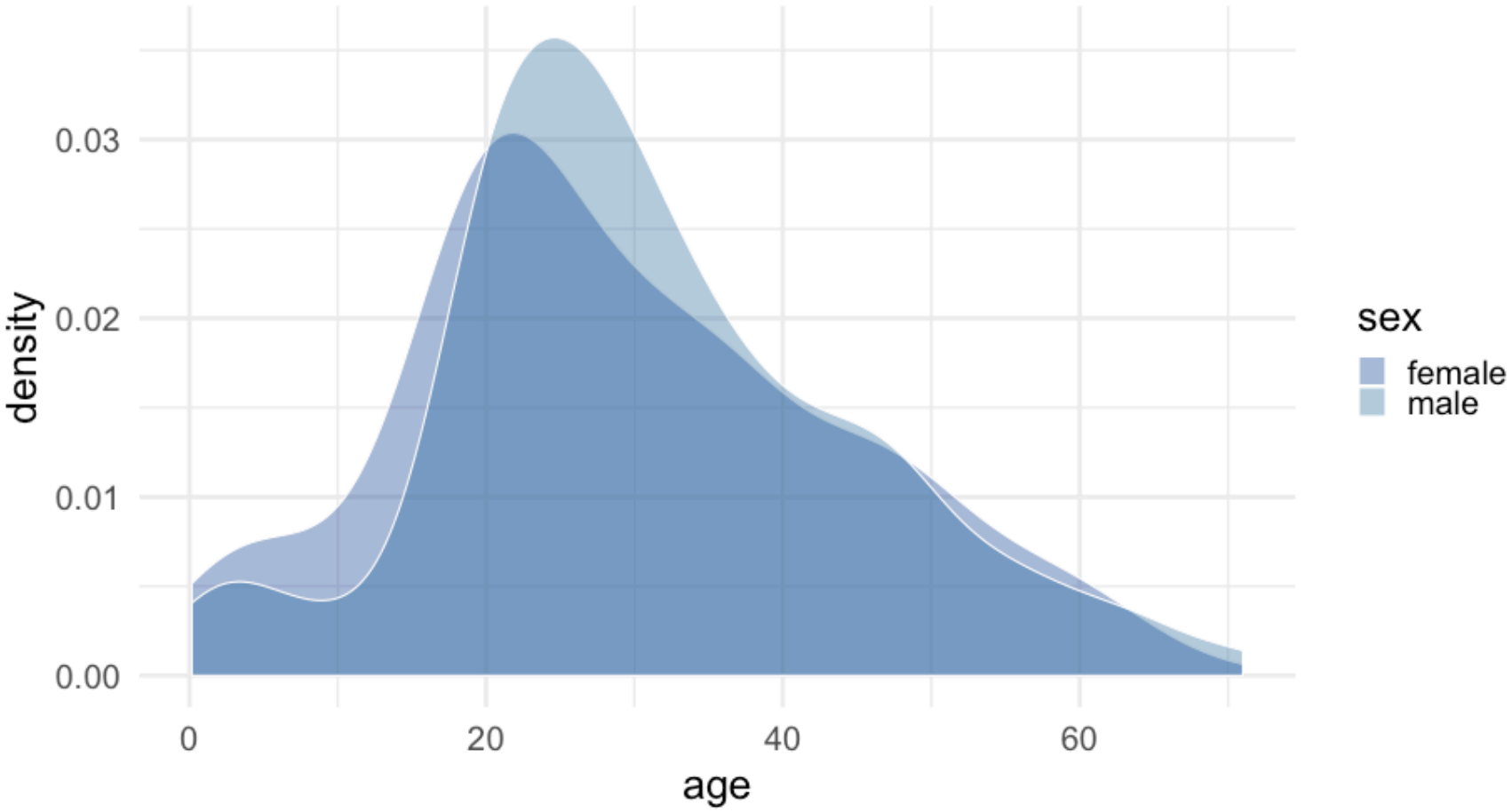
Overlapping densities

```
ggplot(titanic, aes(age)) +
  geom_density(aes(fill = sex),
               color = "white",
               alpha = 0.4)
```



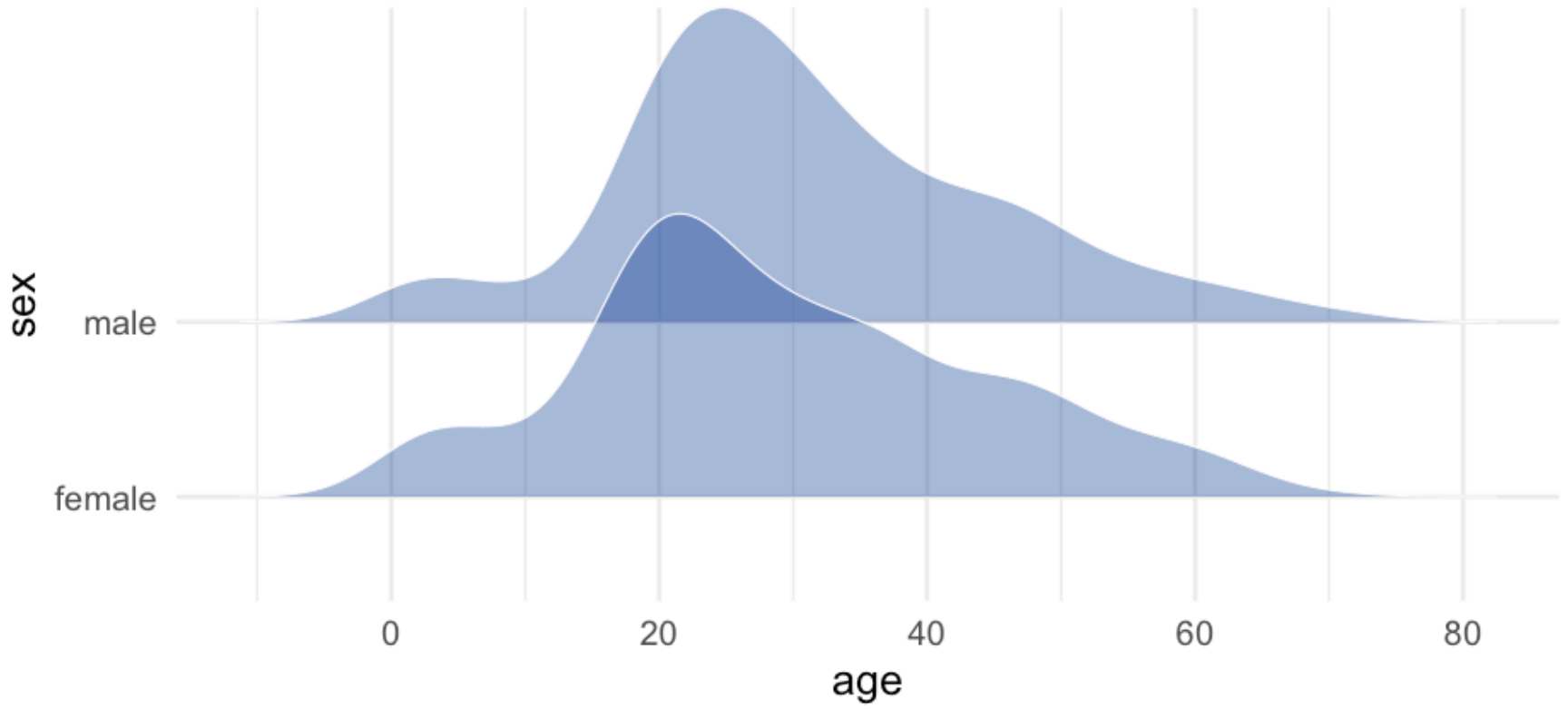
Note the default colors really don't work well in most of these

```
ggplot(titanic, aes(age)) +
  geom_density(aes(fill = sex),
               color = "white",
               alpha = 0.4) +
  scale_fill_manual(values = c("#003F9C", "#2372A3"))
```



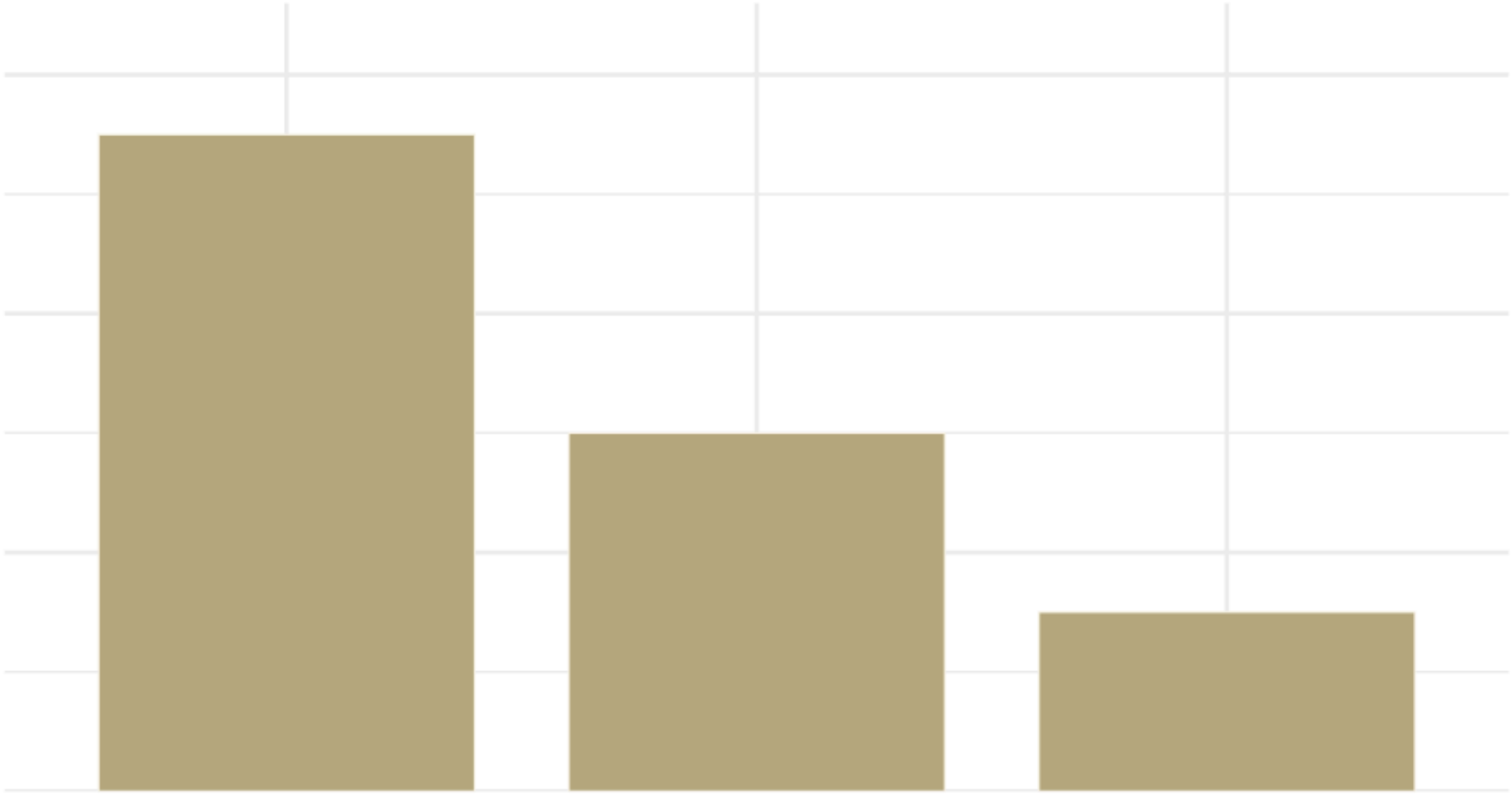
Ridgeline densities

```
ggplot(titanic, aes(age, sex)) +
  ggridges::geom_density_ridges(color = "white",
                                alpha = 0.4,
                                fill = "#003F9C")
```

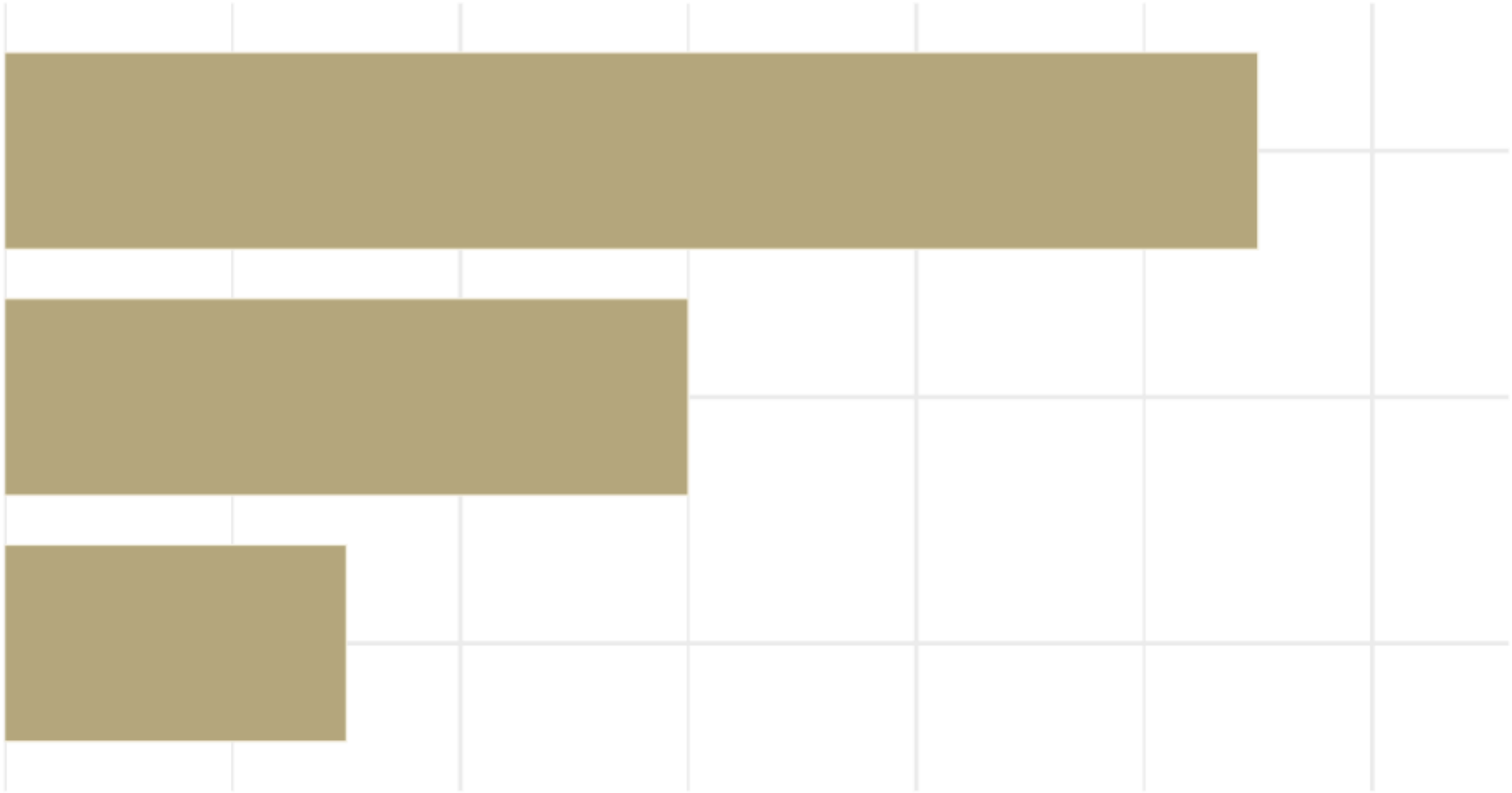


Visualizing amounts

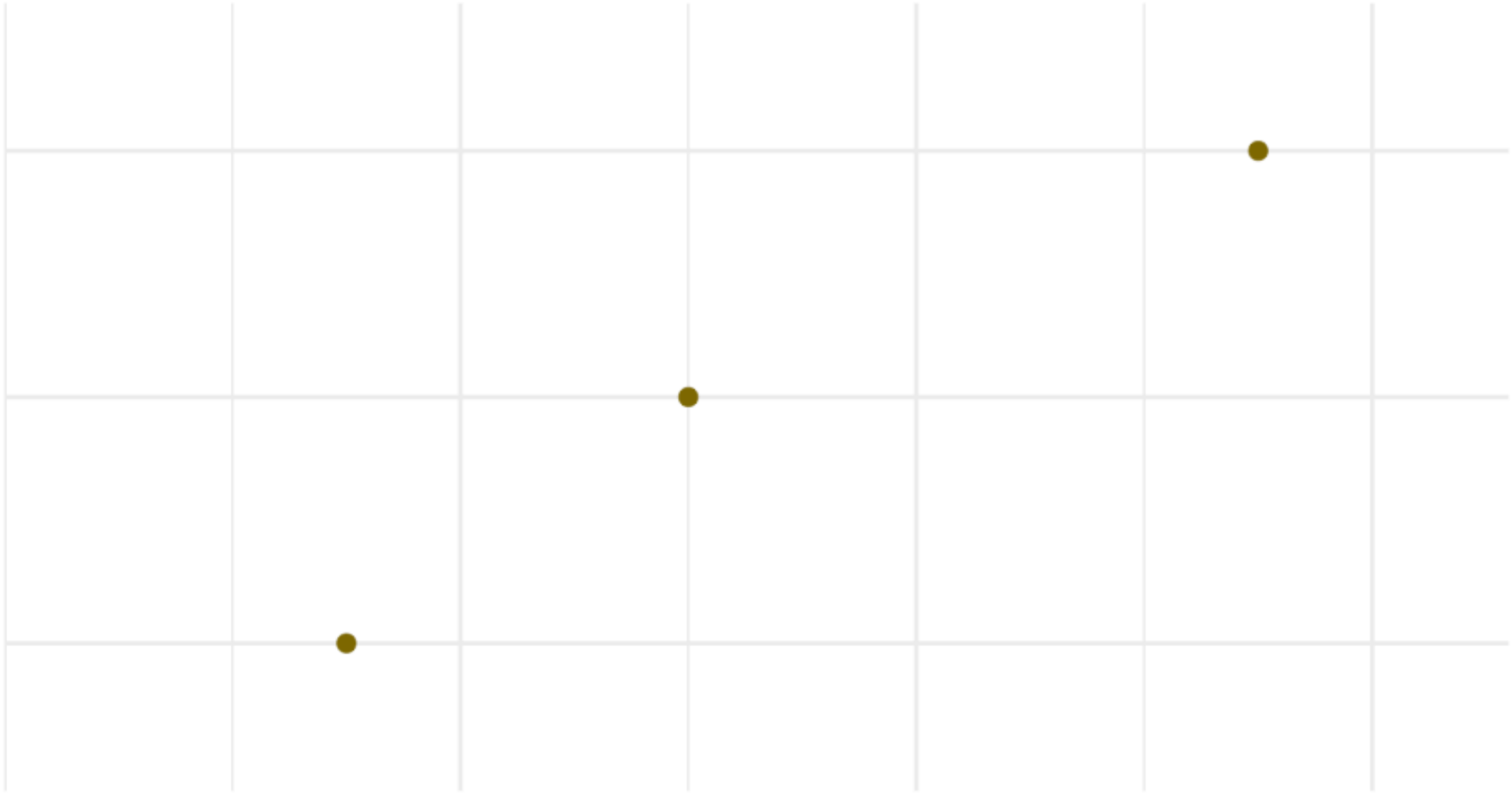
Bar plots



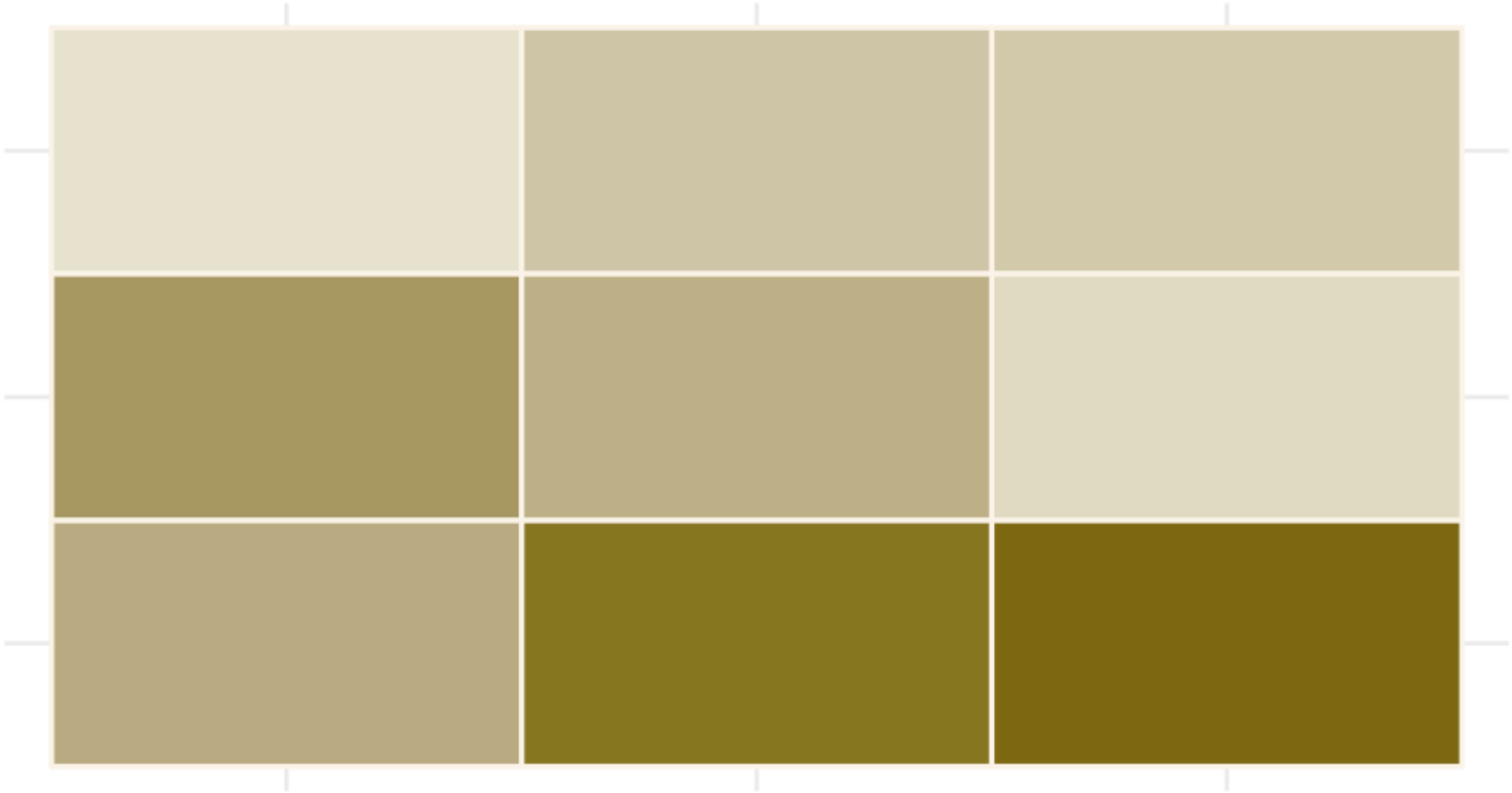
Flipped bars



Dotplot



Heatmap



Empirical examples

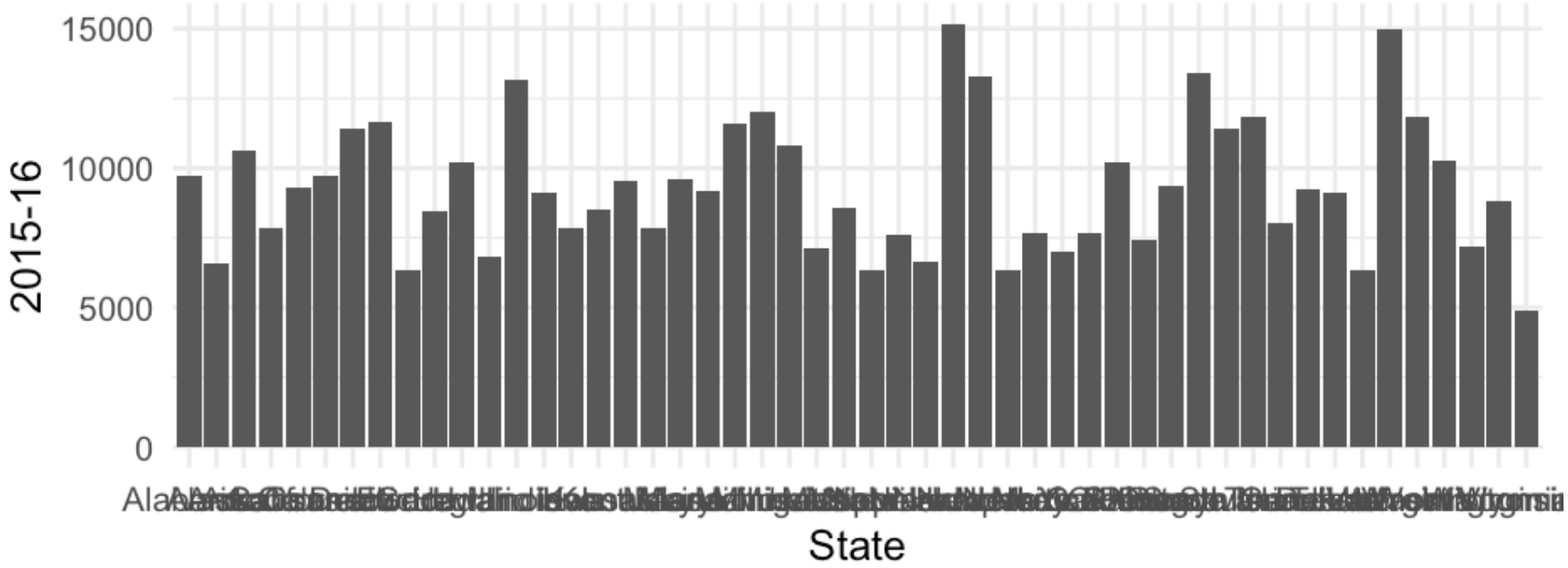
How much does college cost?

```
library(here)
library(rio)
tuition <- import(here("data", "us_avg_tuition.xlsx"),
                  setclass = "tbl_df")
head(tuition)
```

```
## # A tibble: 6 x 13
##   State `2004-05` `2005-06` `2006-07` `2007-08` `2008-09` `2009-10`
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 Alab…  5682.838  5840.550  5753.496  6008.169  6475.092  7188.954
## 2 Alas…  4328.281  4632.623  4918.501  5069.822  5075.482  5454.607
## 3 Ariz…  5138.495  5415.516  5481.419  5681.638  6058.464  7263.204
## 4 Arka…  5772.302  6082.379  6231.977  6414.900  6416.503  6627.092
## 5 Cali…  5285.921  5527.881  5334.826  5672.472  5897.888  7258.771
## 6 Colo…  4703.777  5406.967  5596.348  6227.002  6284.137  6948.473
## # … with 6 more variables: `2010-11` <dbl>, `2011-12` <dbl>,
## #   `2012-13` <dbl>, `2013-14` <dbl>, `2014-15` <dbl>, `2015-16` <dbl>
```

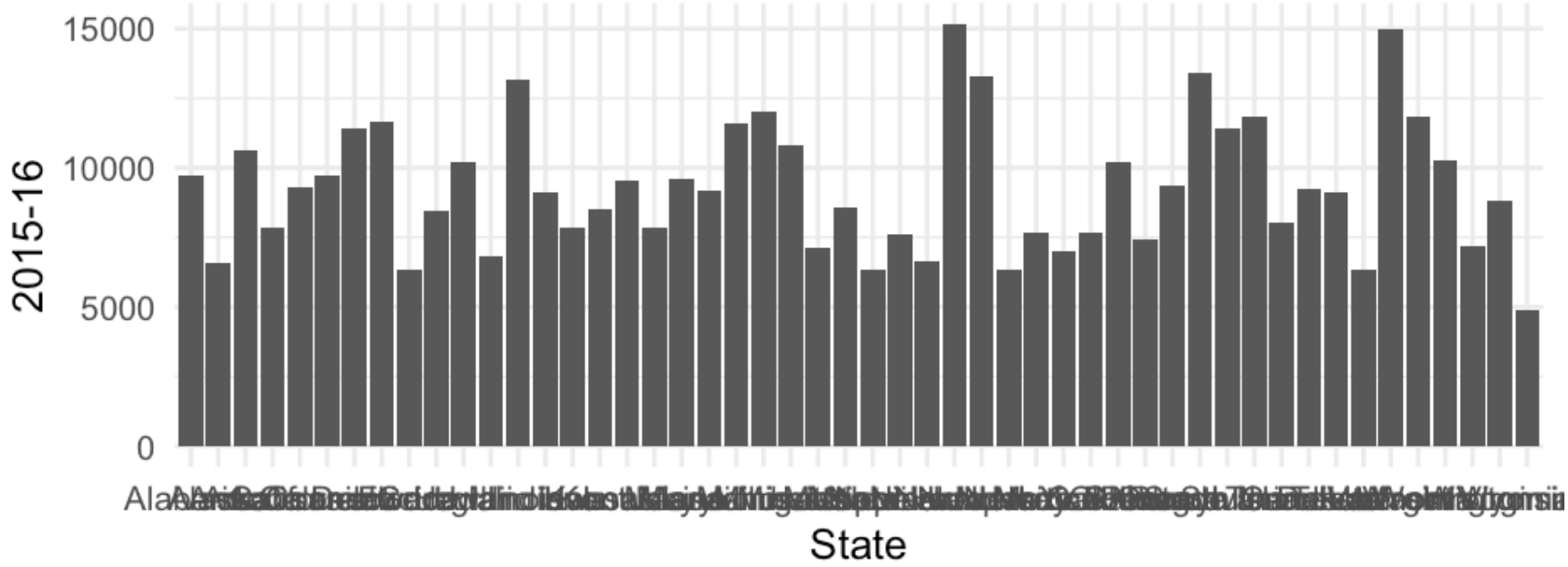
By state: 2015-16

```
ggplot(tuition, aes(State, `2015-16`)) +
  geom_col()
```



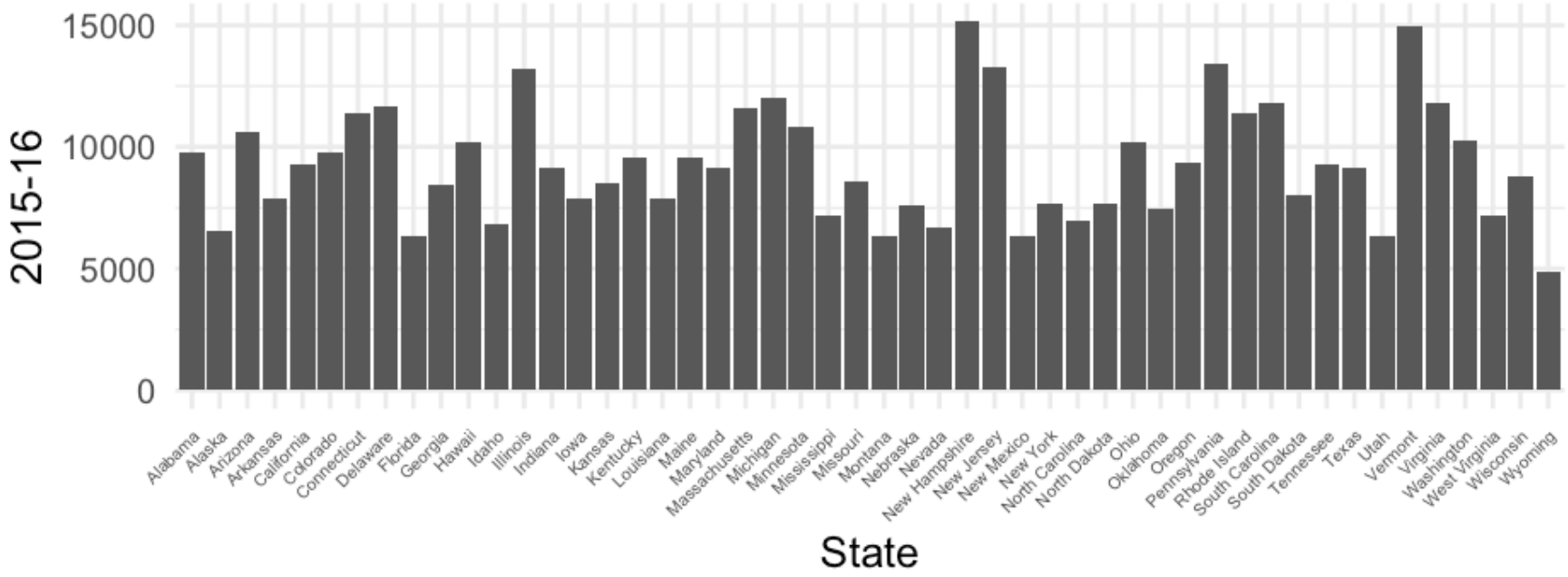
By state: 2015-16

```
ggplot(tuition, aes(State, `2015-16`)) +
  geom_col()
```



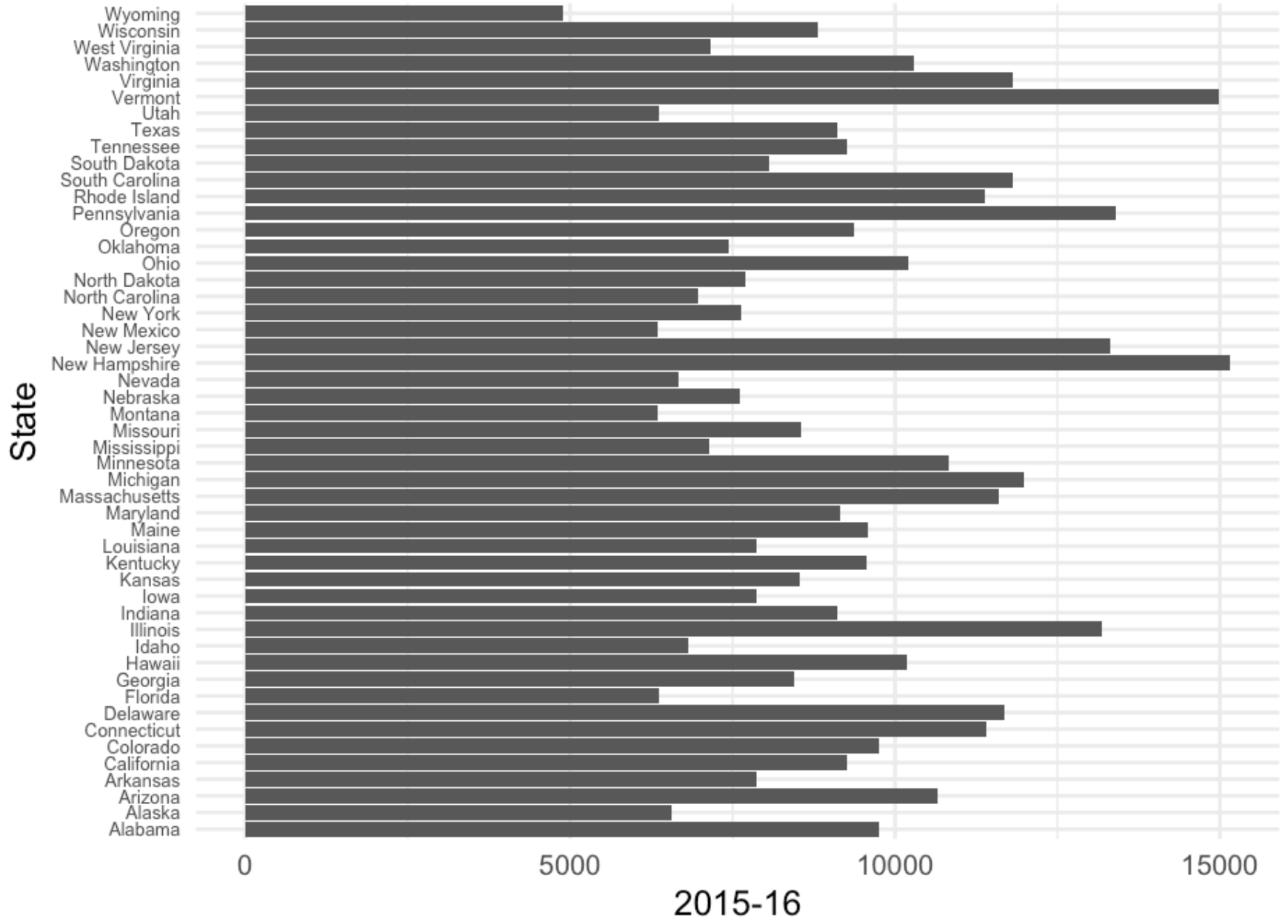
Two puke emoji version

```
ggplot(tuition, aes(State, `2015-16`)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 10))
```



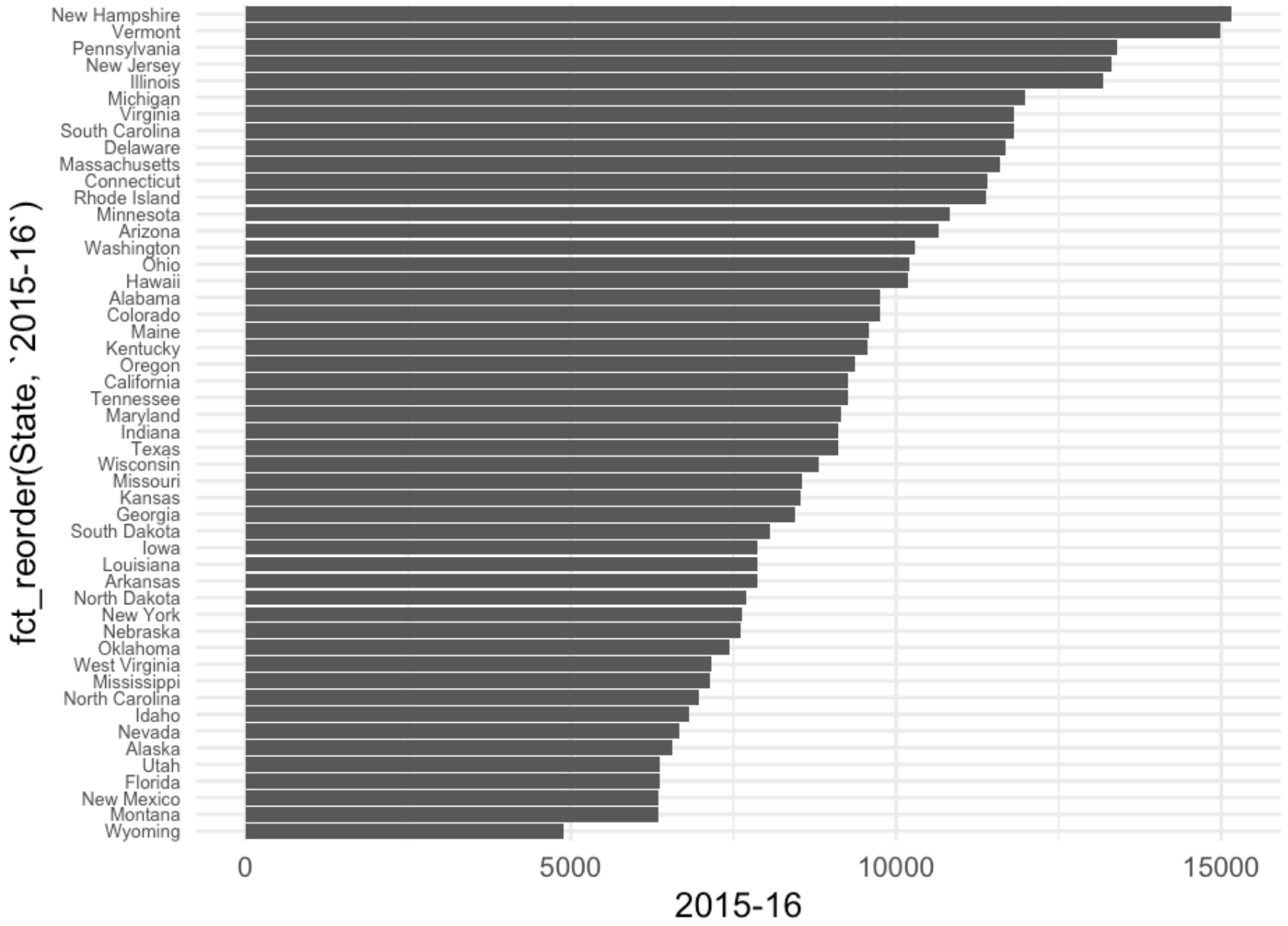
One puke emoji version

```
ggplot(tuition, aes(State, `2015-16`)) +
  geom_col() +
  coord_flip()
```


State
Wyoming
Wisconsin
West Virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Hawaii
Georgia
Florida
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alaska
Alabama
0
5000
10000
15000
2015-16

Kinda smiley version

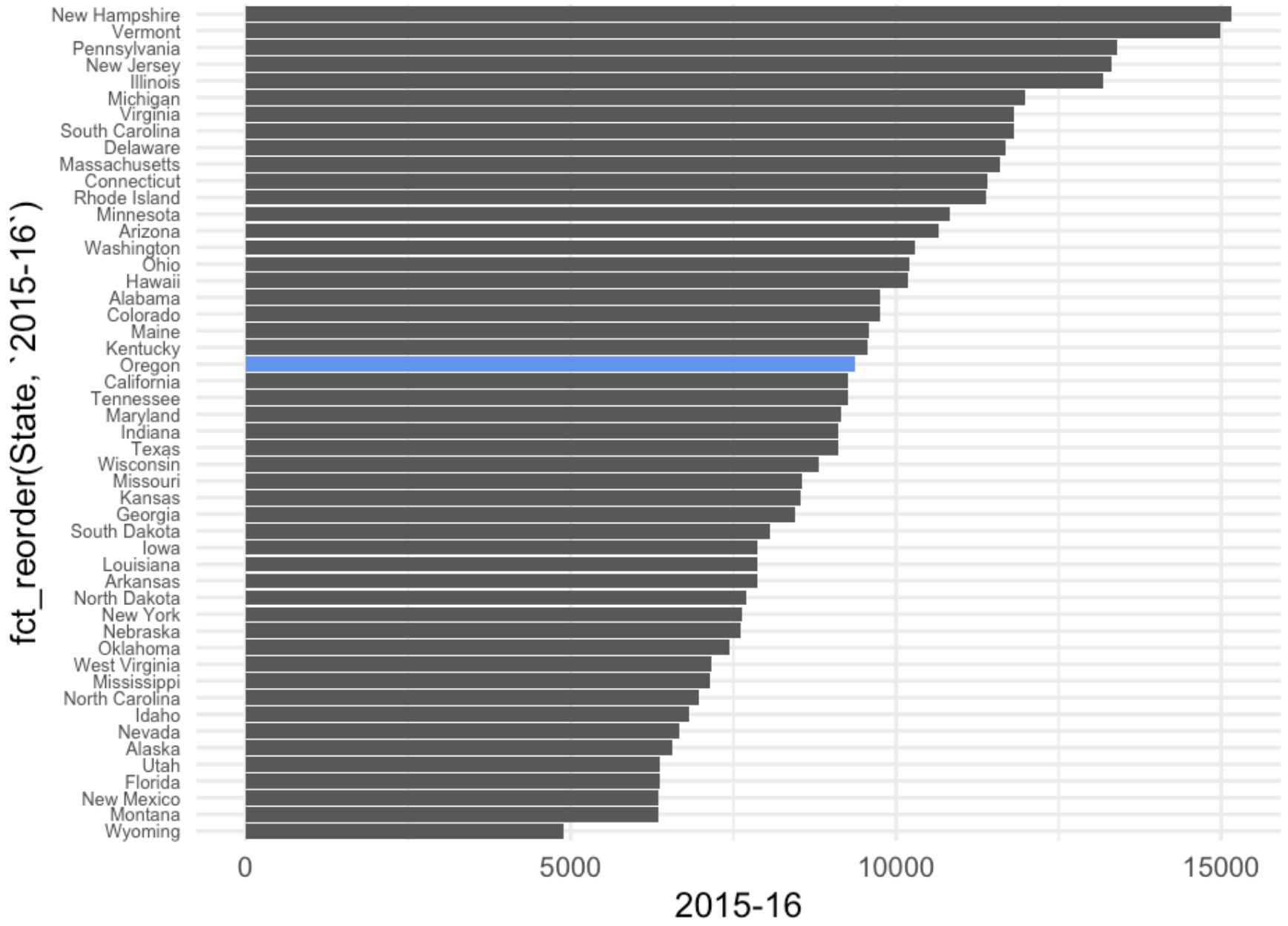
```
ggplot(tuition, aes(fct_reorder(State, `2015-16`), `2015-16`)) +
  geom_col() +
  coord_flip()
```


New Hampshire
Vermont
Pennsylvania
New Jersey
Illinois
Michigan
Virginia
South Carolina
Delaware
Massachusetts
Connecticut
Rhode Island
Minnesota
Arizona
Washington
Ohio
Hawaii
Alabama
Colorado
Maine
Kentucky
Oregon
California
Tennessee
Maryland
Indiana
Texas
Wisconsin
Missouri
Kansas
Georgia
South Dakota
Iowa
Louisiana
Arkansas
North Dakota
New York
Nebraska
Oklahoma
West Virginia
Mississippi
North Carolina
Idaho
Nevada
Alaska
Utah
Florida
New Mexico
Montana
Wyoming
fct_reorder(State, `2015-16`)
0
5000
10000
15000
2015-16

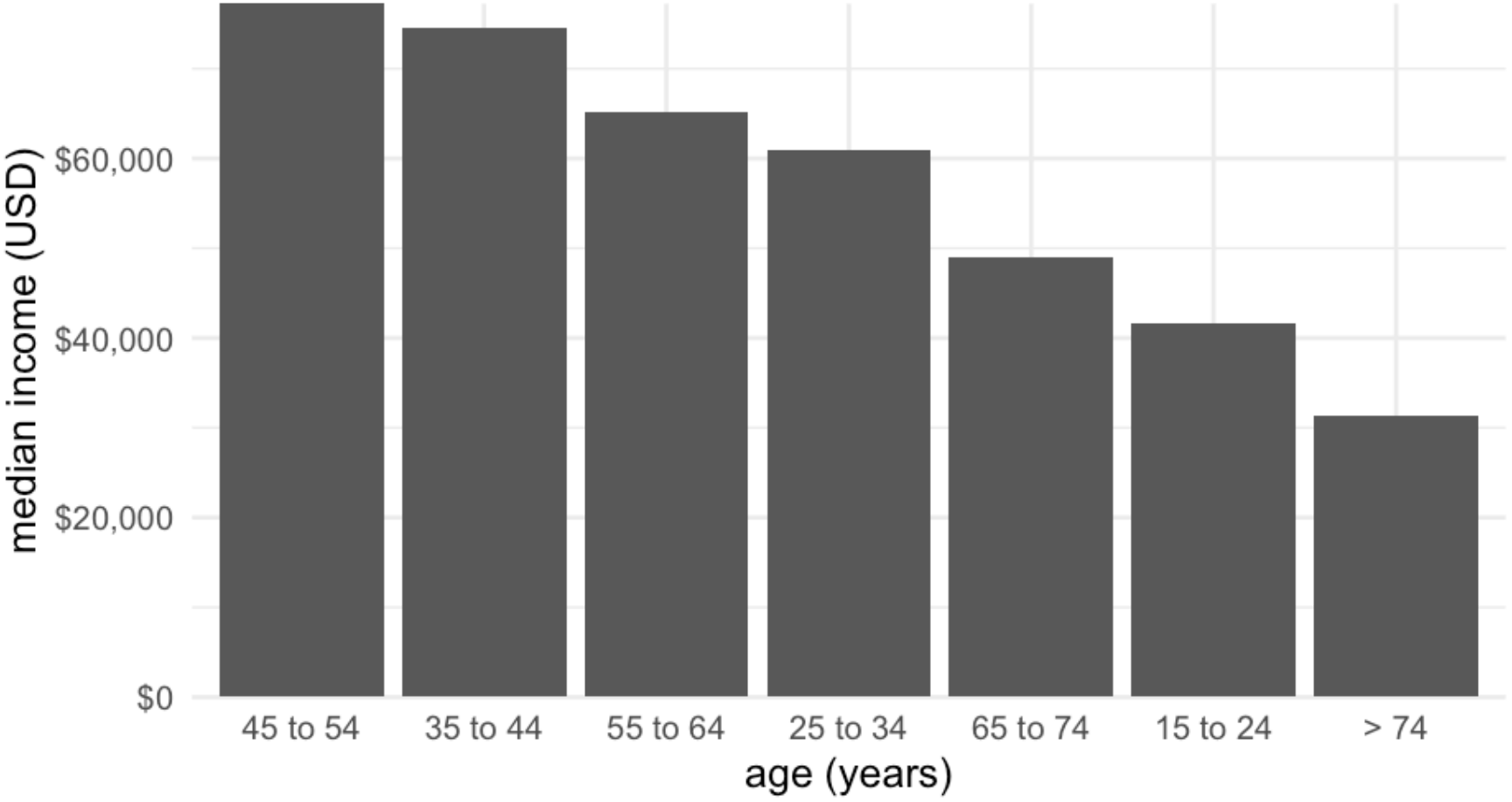
Highlight Oregon

🙂

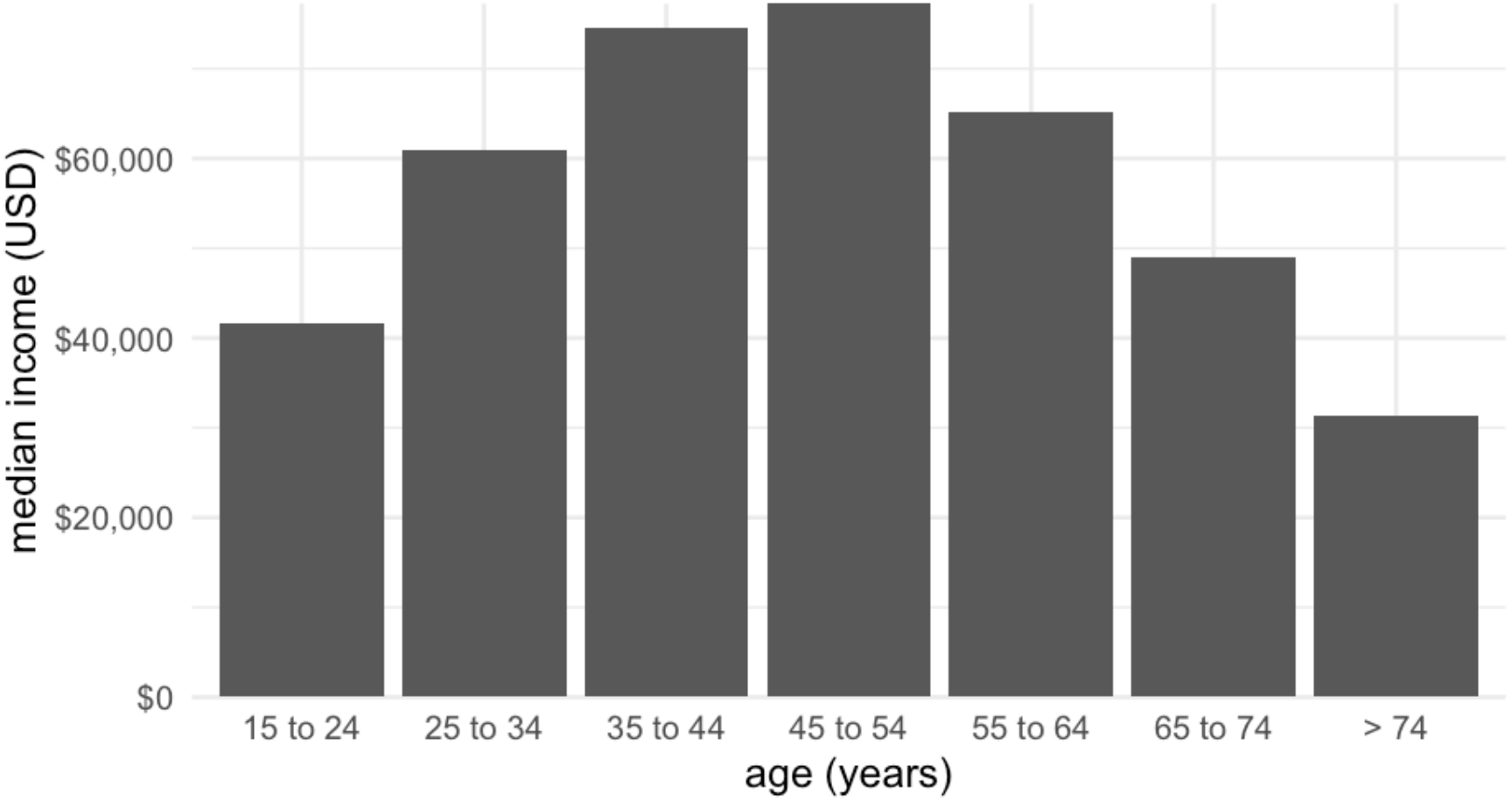
```
ggplot(tuition, aes(fct_reorder(State, `2015-16`), `2015-16`)) +
  geom_col() +
  geom_col(fill = "cornflowerblue",
           data = filter(tuition, State == "Oregon")) +
  coord_flip()
```


New Hampshire
Vermont
Pennsylvania
New Jersey
Illinois
Michigan
Virginia
South Carolina
Delaware
Massachusetts
Connecticut
Rhode Island
Minnesota
Arizona
Washington
Ohio
Hawaii
Alabama
Colorado
Maine
Kentucky
Oregon
California
Tennessee
Maryland
Indiana
Texas
Wisconsin
Missouri
Kansas
Georgia
South Dakota
Iowa
Louisiana
Arkansas
North Dakota
New York
Nebraska
Oklahoma
West Virginia
Mississippi
North Carolina
Idaho
Nevada
Alaska
Utah
Florida
New Mexico
Montana
Wyoming
fct_reorder(State, `2015-16`)
0
5000
10000
15000
2015-16

Not always good to sort



Much better


median income (USD)
$60,000
$40,000
$20,000
$0
15 to 24
25 to 34
35 to 44
45 to 54
55 to 64
65 to 74
> 74
age (years)

Averages tuition by year

How?

```
head(tuition)
```

```
## # A tibble: 6 x 13
##   State `2004-05` `2005-06` `2006-07` `2007-08` `2008-09` `2009-10`
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 Alab…  5682.838  5840.550  5753.496  6008.169  6475.092  7188.954
## 2 Alas…  4328.281  4632.623  4918.501  5069.822  5075.482  5454.607
## 3 Ariz…  5138.495  5415.516  5481.419  5681.638  6058.464  7263.204
## 4 Arka…  5772.302  6082.379  6231.977  6414.900  6416.503  6627.092
## 5 Cali…  5285.921  5527.881  5334.826  5672.472  5897.888  7258.771
## 6 Colo…  4703.777  5406.967  5596.348  6227.002  6284.137  6948.473
## # … with 6 more variables: `2010-11` <dbl>, `2011-12` <dbl>,
## #   `2012-13` <dbl>, `2013-14` <dbl>, `2014-15` <dbl>, `2015-16` <dbl>
```

Rearrange

```
tuition %>%
  gather(year, avg_tuition, `2004-05`:`2015-16`)
```

```
## # A tibble: 600 x 3
##    State       year    avg_tuition
##    <chr>       <chr>         <dbl>
##  1 Alabama     2004-05    5682.838
##  2 Alaska      2004-05    4328.281
##  3 Arizona     2004-05    5138.495
##  4 Arkansas    2004-05    5772.302
##  5 California  2004-05    5285.921
##  6 Colorado    2004-05    4703.777
##  7 Connecticut 2004-05    7983.695
##  8 Delaware    2004-05    8352.89
##  9 Florida     2004-05    3848.201
## 10 Georgia     2004-05    4298.040
## # … with 590 more rows
```

Compute summaries

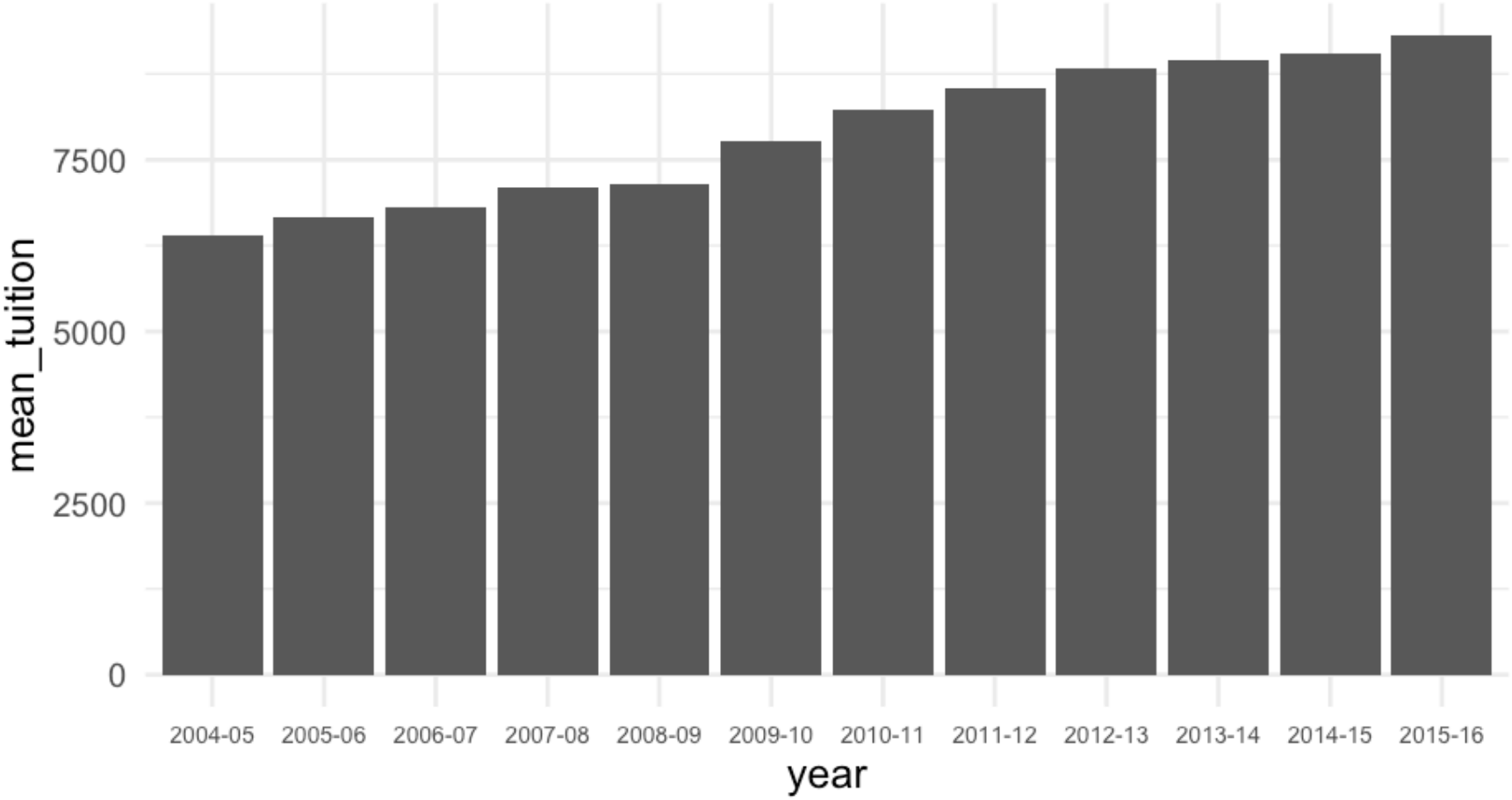
```
annual_means <- tuition %>%
  gather(year, avg_tuition, `2004-05`:`2015-16`) %>%
  group_by(year) %>%
  summarize(mean_tuition = mean(avg_tuition))

annual_means
```

```
## # A tibble: 12 x 2
##    year    mean_tuition
##    <chr>          <dbl>
##  1 2004-05     6409.564
##  2 2005-06     6654.177
##  3 2006-07     6809.914
##  4 2007-08     7085.881
##  5 2008-09     7156.560
##  6 2009-10     7761.810
##  7 2010-11     8228.834
##  8 2011-12     8539.115
##  9 2012-13     8842.357
## 10 2013-14     8947.938
## 11 2014-15     9037.357
## 12 2015-16     9317.633
```

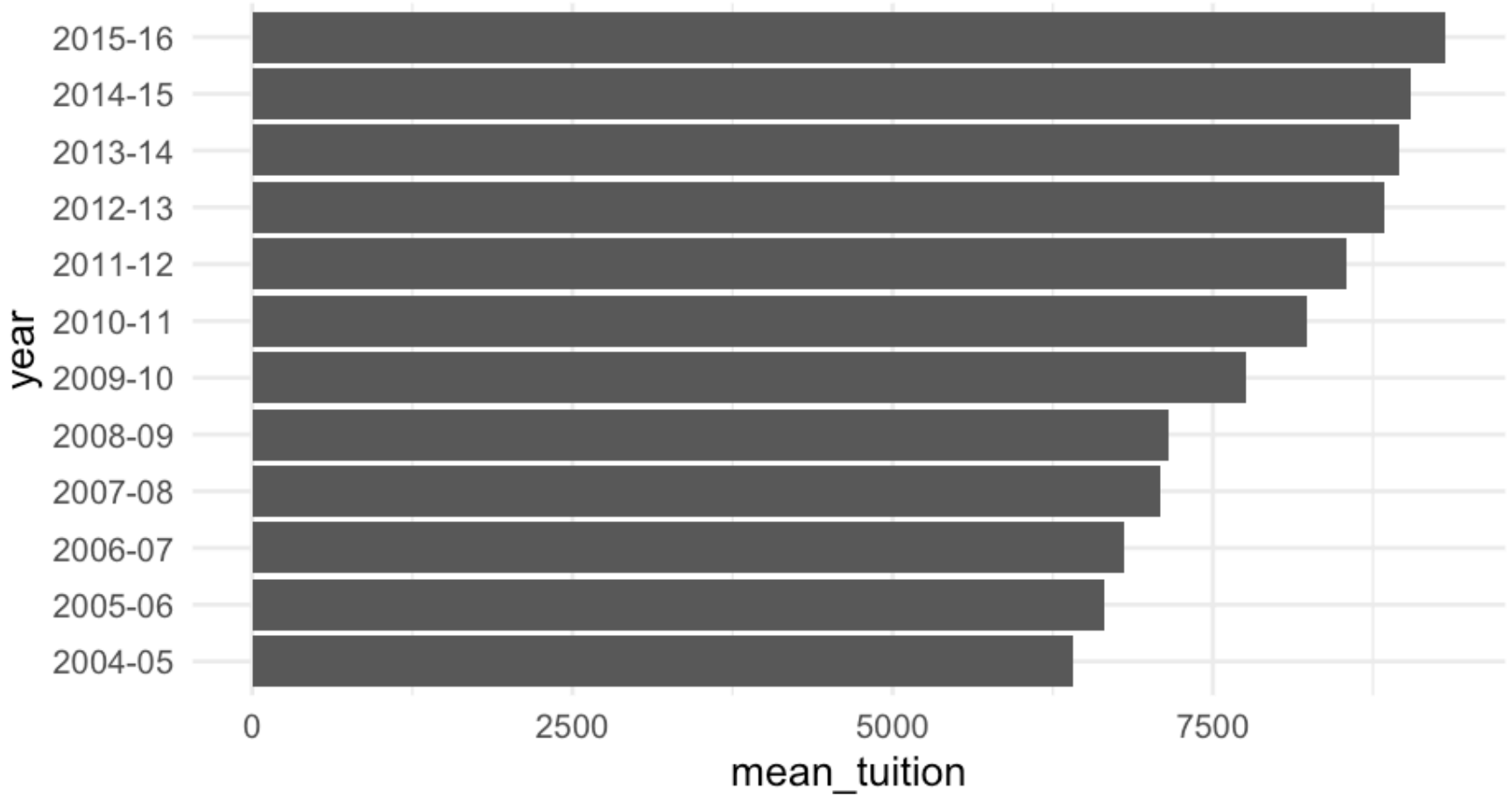
Good

```
ggplot(annual_means, aes(year, mean_tuition)) +
  geom_col()
```



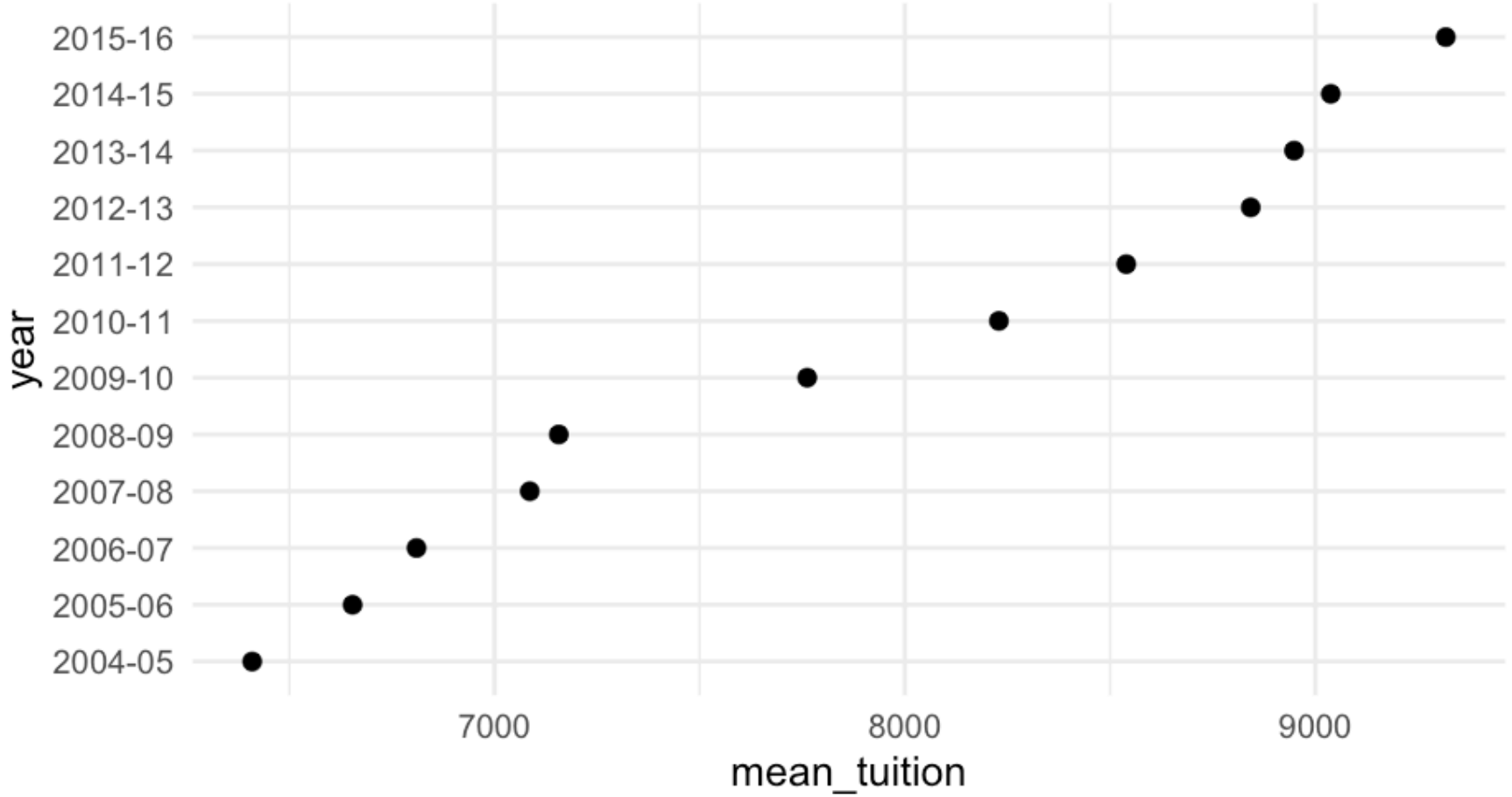
Better?

```
ggplot(annual_means, aes(year, mean_tuition)) +
  geom_col() +
  coord_flip()
```



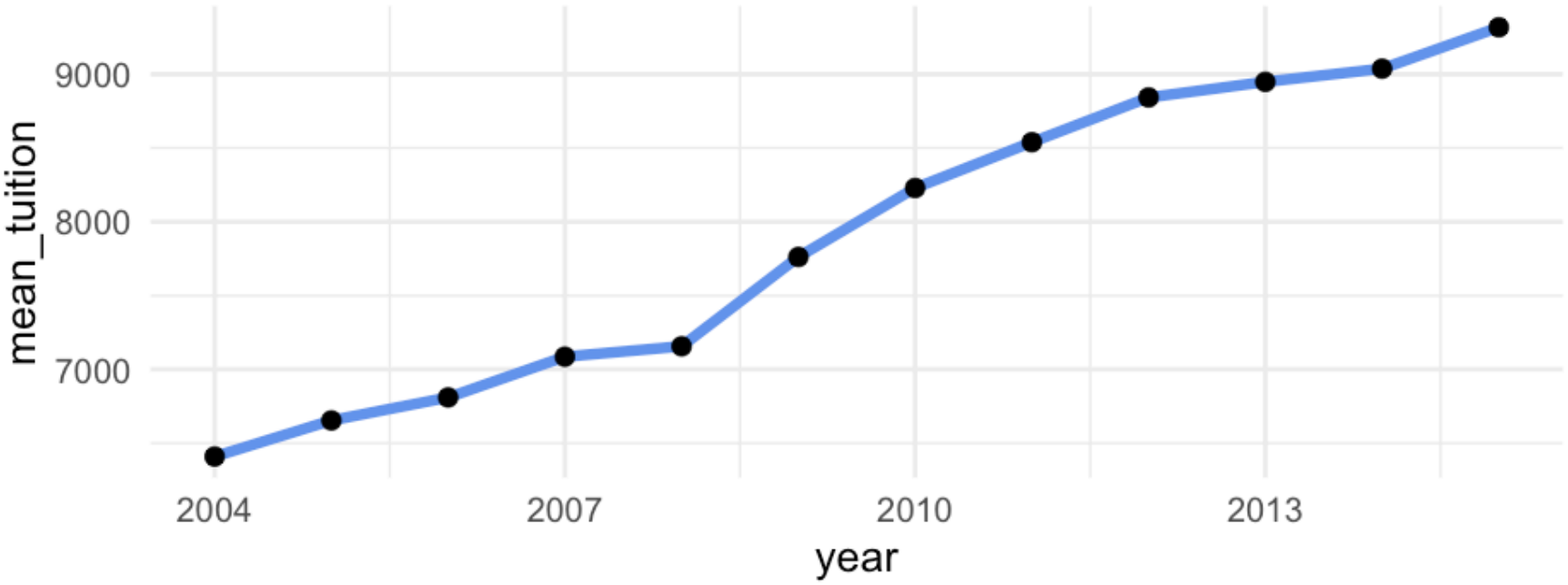
Better still?

```
ggplot(annual_means, aes(year, mean_tuition)) +
  geom_point() +
  coord_flip()
```



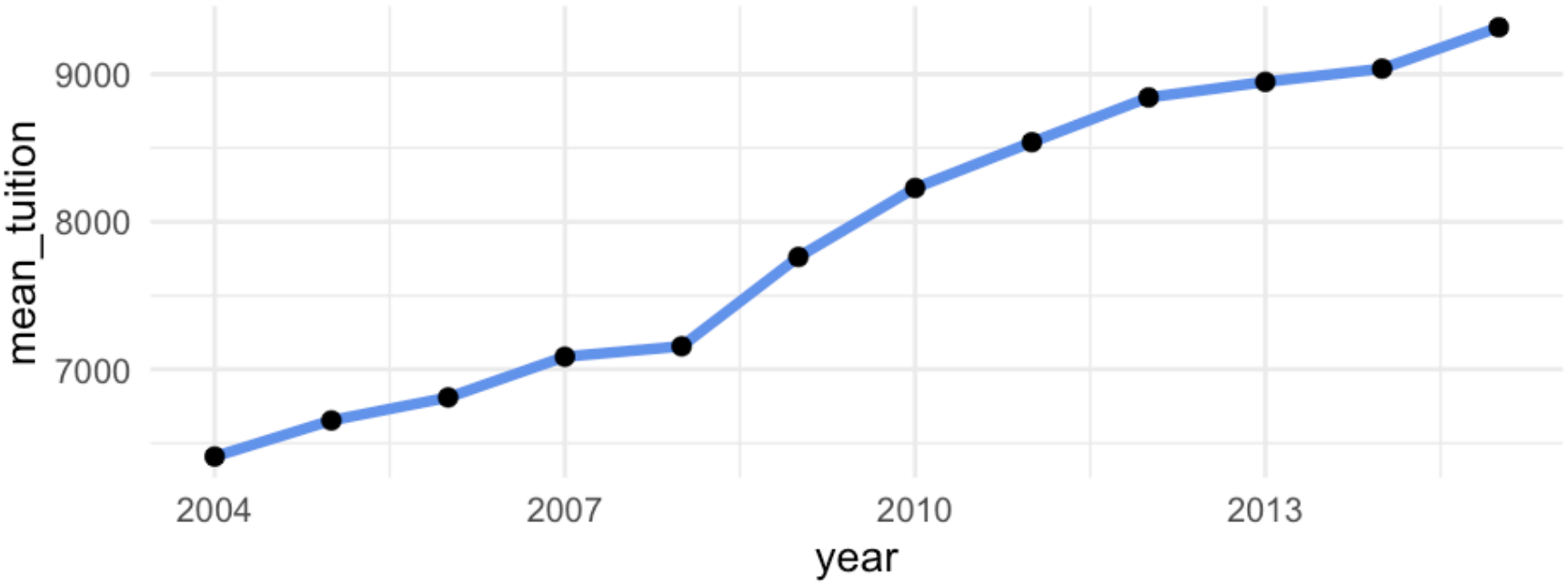
Even better

```
annual_means %>%
  mutate(year = readr::parse_number(year)) %>%
  ggplot(aes(year, mean_tuition)) +
    geom_line(color = "cornflowerblue") +
    geom_point()
```



Even better

```
annual_means %>%
  mutate(year = readr::parse_number(year)) %>%
  ggplot(aes(year, mean_tuition)) +
    geom_line(color = "cornflowerblue") +
    geom_point()
```



Treat time (year) as a continuous variable

Let's back up a bit

- Lets go back to our full data, but in a format that we can have a `year` variable.

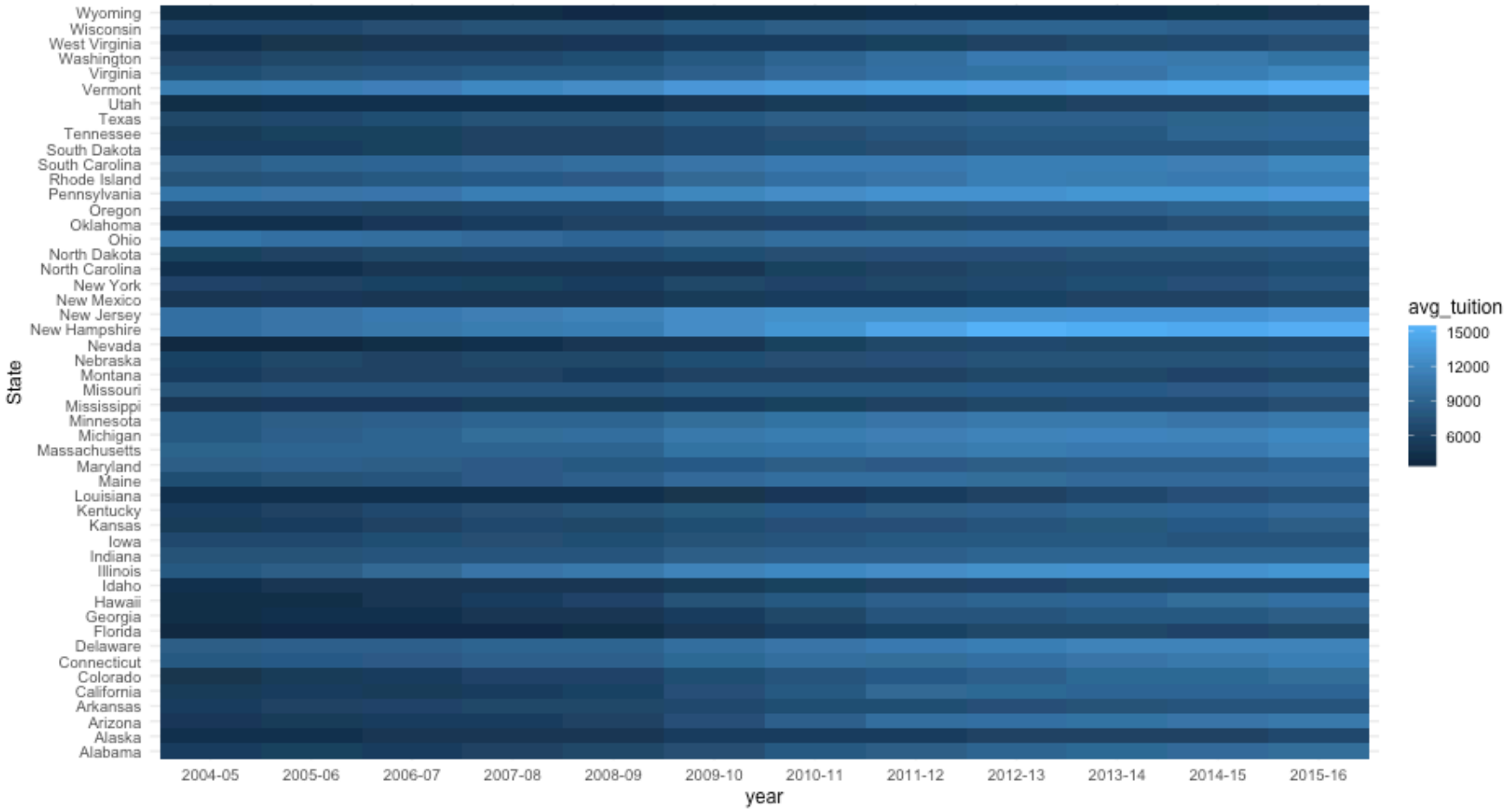
```
tuition_l <- tuition %>%
  gather(year, avg_tuition, -State)

tuition_l
```

```
## # A tibble: 600 x 3
##    State       year    avg_tuition
##    <chr>       <chr>         <dbl>
##  1 Alabama     2004-05    5682.838
##  2 Alaska      2004-05    4328.281
##  3 Arizona     2004-05    5138.495
##  4 Arkansas    2004-05    5772.302
##  5 California  2004-05    5285.921
##  6 Colorado    2004-05    4703.777
##  7 Connecticut 2004-05    7983.695
##  8 Delaware    2004-05    8352.89
##  9 Florida     2004-05    3848.201
## 10 Georgia     2004-05    4298.040
## # … with 590 more rows
```

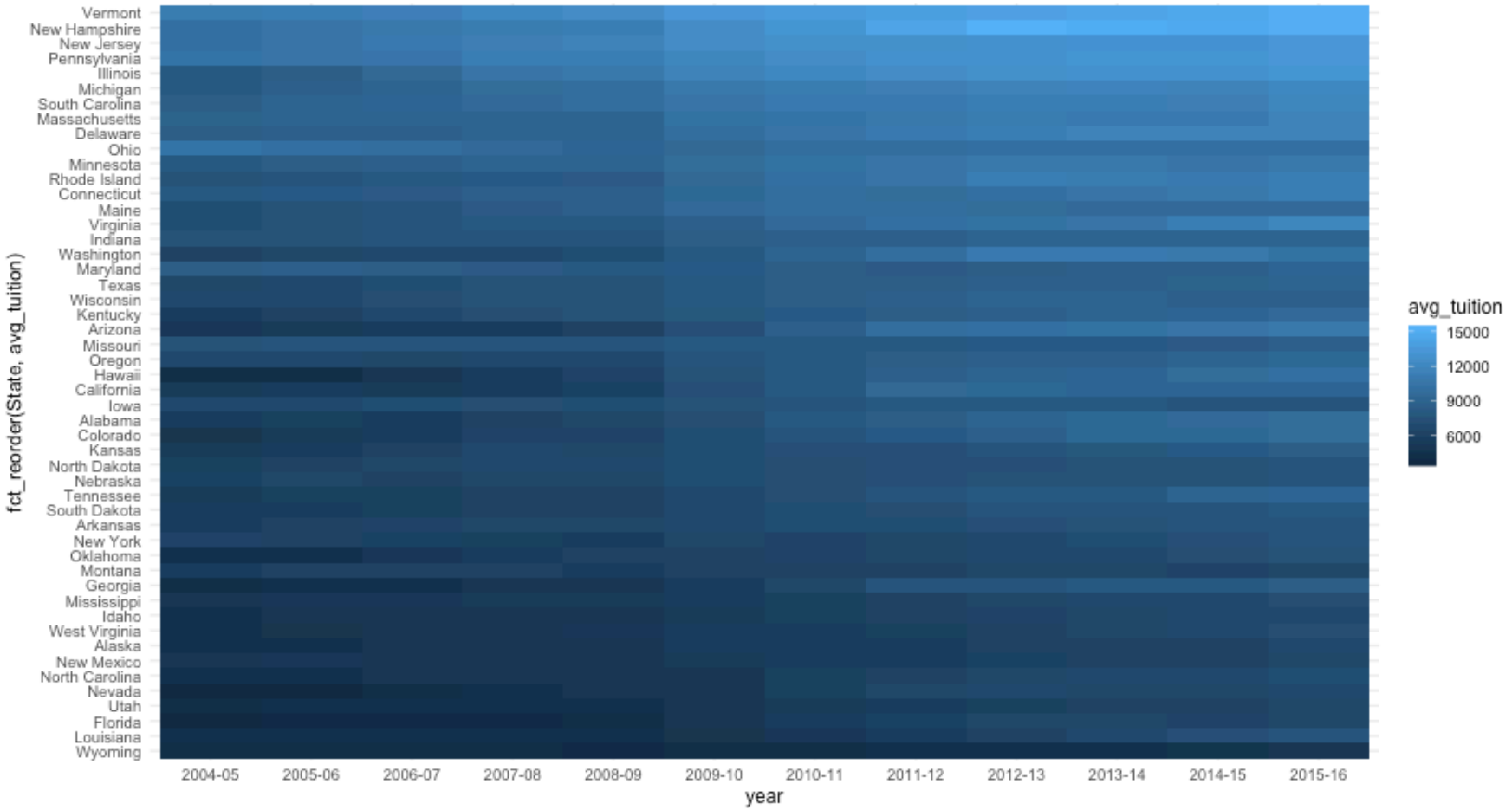
Heatmap

```
ggplot(tuition_l, aes(year, State)) +
  geom_tile(aes(fill = avg_tuition))
```



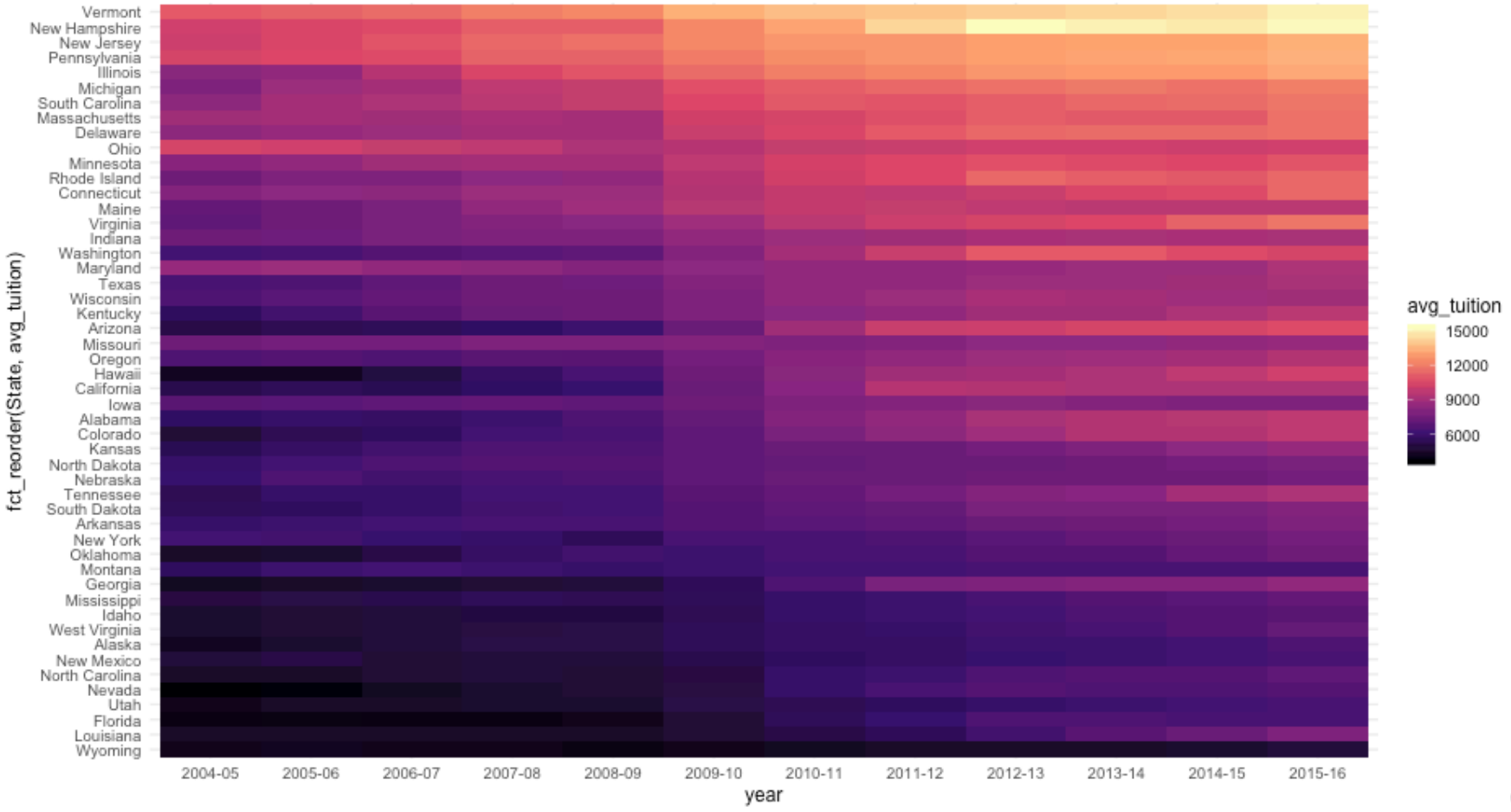
Better heatmap

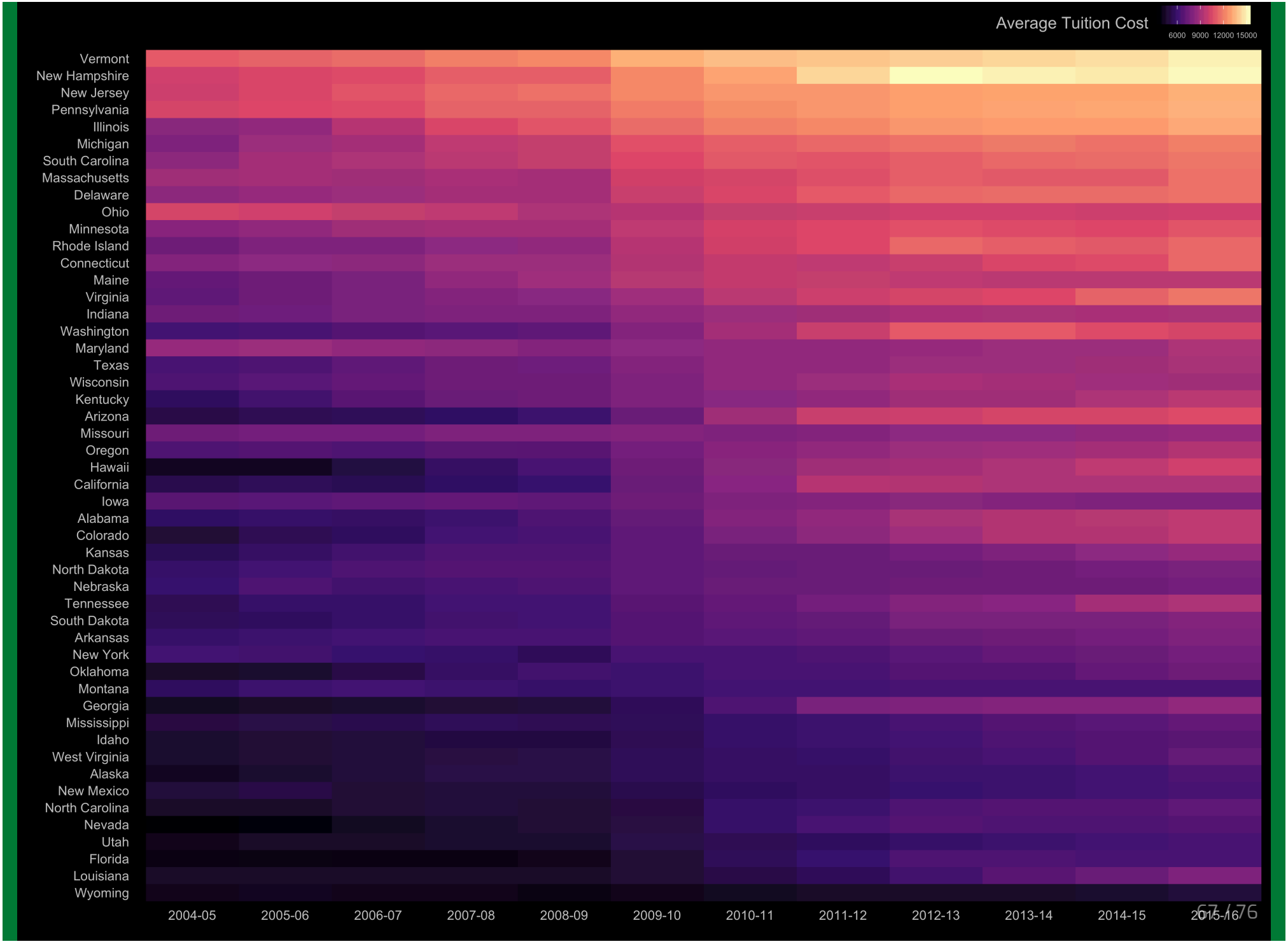
```
ggplot(tuition_l, aes(year, fct_reorder(State, avg_tuition))) +
  geom_tile(aes(fill = avg_tuition))
```



Even better heatmap

```
ggplot(tuition_l, aes(year, fct_reorder(State, avg_tuition))) +
  geom_tile(aes(fill = avg_tuition)) +
  scale_fill_viridis_c(option = "magma")
```




Average Tuition Cost
6000 9000 12000 15000
Vermont
New Hampshire
New Jersey
Pennsylvania
Illinois
Michigan
South Carolina
Massachusetts
Delaware
Ohio
Minnesota
Rhode Island
Connecticut
Maine
Virginia
Indiana
Washington
Maryland
Texas
Wisconsin
Kentucky
Arizona
Missouri
Oregon
Hawaii
California
Iowa
Alabama
Colorado
Kansas
North Dakota
Nebraska
Tennessee
South Dakota
Arkansas
New York
Oklahoma
Montana
Georgia
Mississippi
Idaho
West Virginia
Alaska
New Mexico
North Carolina
Nevada
Utah
Florida
Louisiana
Wyoming
2004-05
2005-06
2006-07
2007-08
2008-09
2009-10
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16

Quick aside

- Think about the data you have
- Given that these are state-level data, they have a geographic component

Quick aside

- Think about the data you have
- Given that these are state-level data, they have a geographic component

```
#install.packages(c("maps"))
state_data <- map_data("state") %>% # ggplot2::map_data
  rename(State = region)
head(state_data)
```

```
##        long      lat group order   State subregion
## 1 -87.46201 30.38968     1     1 alabama      <NA>
## 2 -87.48493 30.37249     1     2 alabama      <NA>
## 3 -87.52503 30.37249     1     3 alabama      <NA>
## 4 -87.53076 30.33239     1     4 alabama      <NA>
## 5 -87.57087 30.32665     1     5 alabama      <NA>
## 6 -87.58806 30.32665     1     6 alabama      <NA>
```

Join it

Obviously we'll talk more about joins later

```
tuition <- tuition %>%
  mutate(State = tolower(State))
states <- left_join(state_data, tuition)
head(states)
```

```
##        long      lat group order   State subregion  2004-05 2005-06
## 1 -87.46201 30.38968     1     1 alabama      <NA> 5682.838 5840.55
## 2 -87.48493 30.37249     1     2 alabama      <NA> 5682.838 5840.55
## 3 -87.52503 30.37249     1     3 alabama      <NA> 5682.838 5840.55
## 4 -87.53076 30.33239     1     4 alabama      <NA> 5682.838 5840.55
## 5 -87.57087 30.32665     1     5 alabama      <NA> 5682.838 5840.55
## 6 -87.58806 30.32665     1     6 alabama      <NA> 5682.838 5840.55
##    2006-07  2007-08  2008-09  2009-10  2010-11  2011-12  2012-13  2013-14
## 1 5753.496 6008.169 6475.092 7188.954 8071.134 8451.902 9098.069 9358.929
## 2 5753.496 6008.169 6475.092 7188.954 8071.134 8451.902 9098.069 9358.929
## 3 5753.496 6008.169 6475.092 7188.954 8071.134 8451.902 9098.069 9358.929
## 4 5753.496 6008.169 6475.092 7188.954 8071.134 8451.902 9098.069 9358.929
## 5 5753.496 6008.169 6475.092 7188.954 8071.134 8451.902 9098.069 9358.929
## 6 5753.496 6008.169 6475.092 7188.954 8071.134 8451.902 9098.069 9358.929
##    2014-15  2015-16
## 1 9496.084 9751.101
```

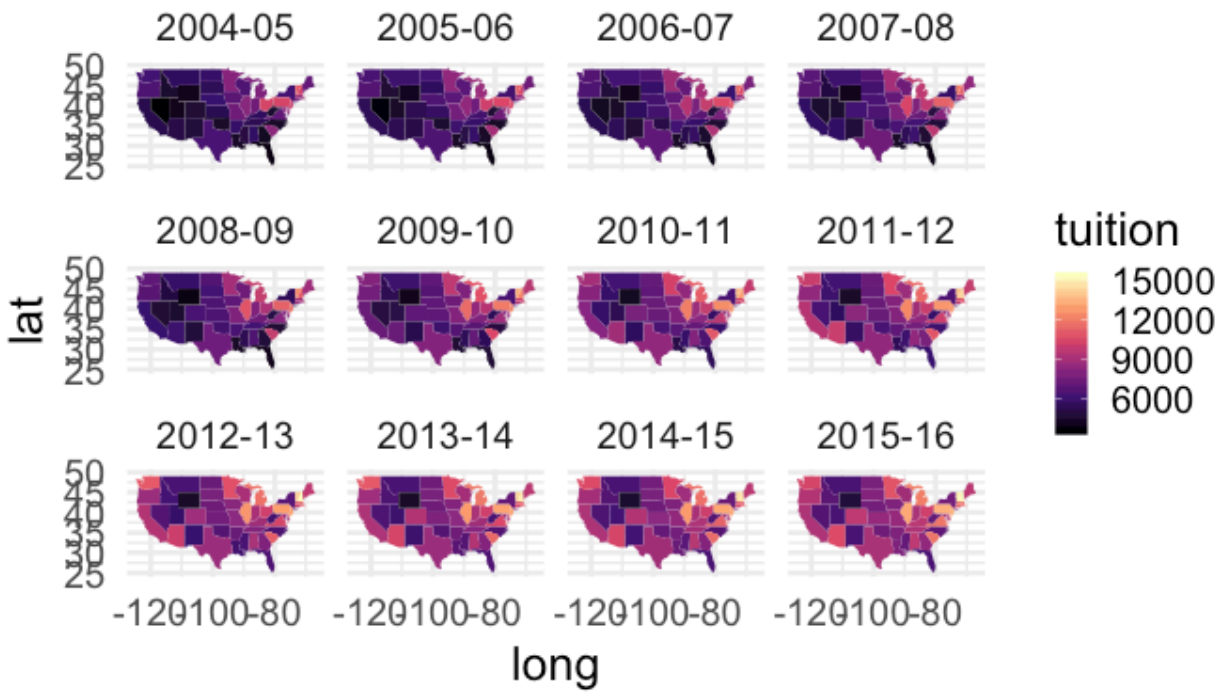
Arrange

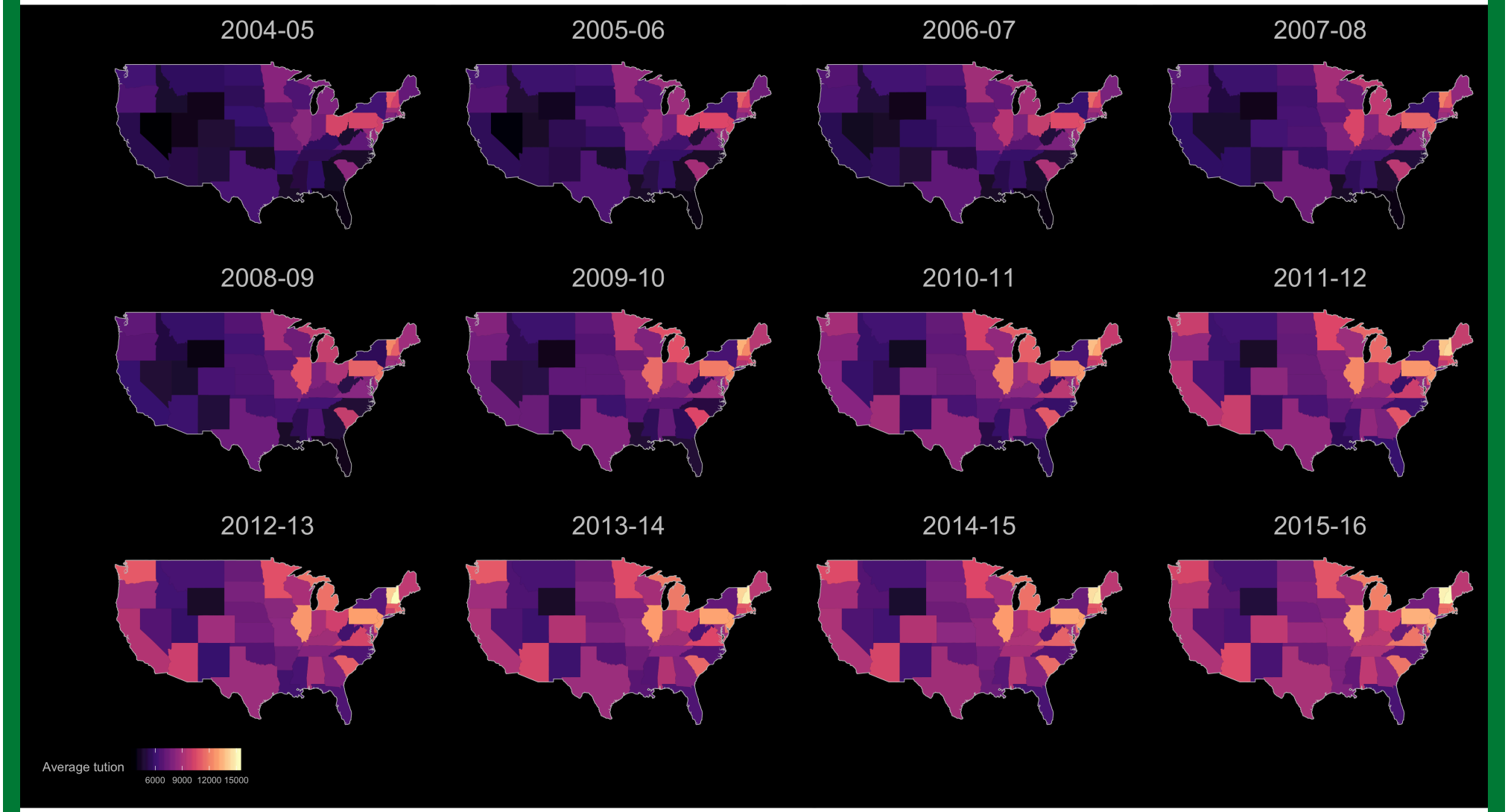
```
states <- states %>%
  gather(year, tuition, `2004-05`:`2015-16`)
head(states)
```

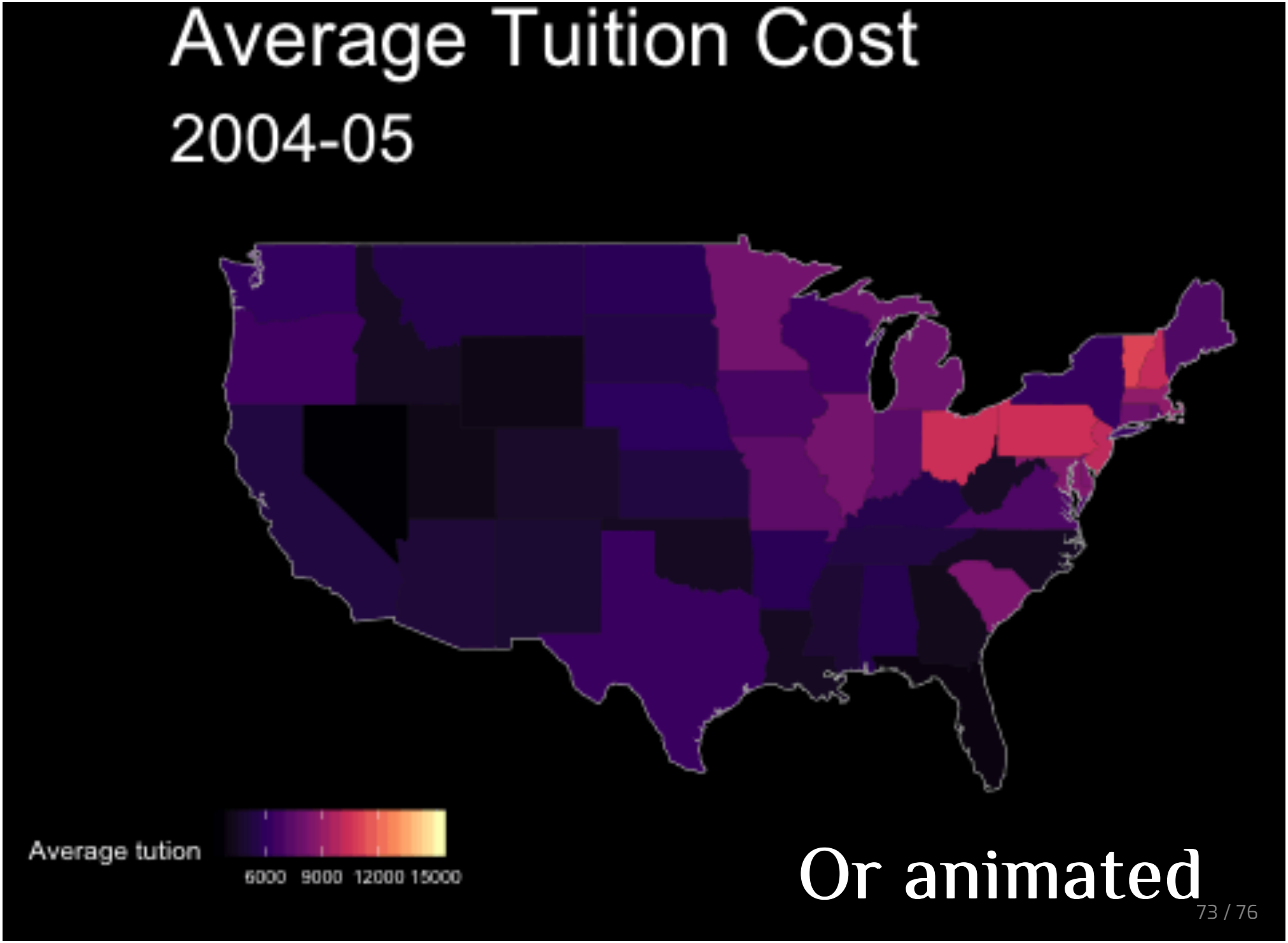
```
##        long      lat group order   State subregion    year  tuition
## 1 -87.46201 30.38968     1     1 alabama      <NA> 2004-05 5682.838
## 2 -87.48493 30.37249     1     2 alabama      <NA> 2004-05 5682.838
## 3 -87.52503 30.37249     1     3 alabama      <NA> 2004-05 5682.838
## 4 -87.53076 30.33239     1     4 alabama      <NA> 2004-05 5682.838
## 5 -87.57087 30.32665     1     5 alabama      <NA> 2004-05 5682.838
## 6 -87.58806 30.32665     1     6 alabama      <NA> 2004-05 5682.838
```

Plot

```
ggplot(states) +
  geom_polygon(aes(long, lat, group = group, fill = tuition)) +
  coord_fixed(1.3) +
  scale_fill_viridis_c(option = "magma") +
  facet_wrap(~year)
```




2004-05
2005-06
2006-07
2007-08
2008-09
2009-10
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
Average tution
6000 9000 12000 15000


Average Tuition Cost
2004-05
Average tution
6000 9000 12000 15000
Or animated

Grouped data

Grouped data

I'm sure we're running short on time

Grouped data

I'm sure we're running short on time

We just saw one a couple instances of grouped data

Grouped data

I'm sure we're running short on time

We just saw one a couple instances of grouped data

Almost all data can be displayed by group

Grouped data

I'm sure we're running short on time

We just saw one a couple instances of grouped data

Almost all data can be displayed by group

Some methods work better than others, depending on the context (e.g., stacked versus dodged versus faceted)

Grouped data

I'm sure we're running short on time

We just saw one a couple instances of grouped data

Almost all data can be displayed by group

Some methods work better than others, depending on the context (e.g., stacked versus dodged versus faceted)

We'll talk about all of this more as we go

Wrapping up

- We've got a ways to go - today was just an introduction
- We basically didn't talk about multivariate data (not even scatter plots)
- Other types of plots will be embedded within the topics later in the class

Next time

Take-home lab

Mapping data to aesthetics

- Will ask you to try out different bins and bandwidths and make a judgement
- Visualize amounts
- Will push you a little on grouping
- Feel free to work together