# Tips for effective data visualization

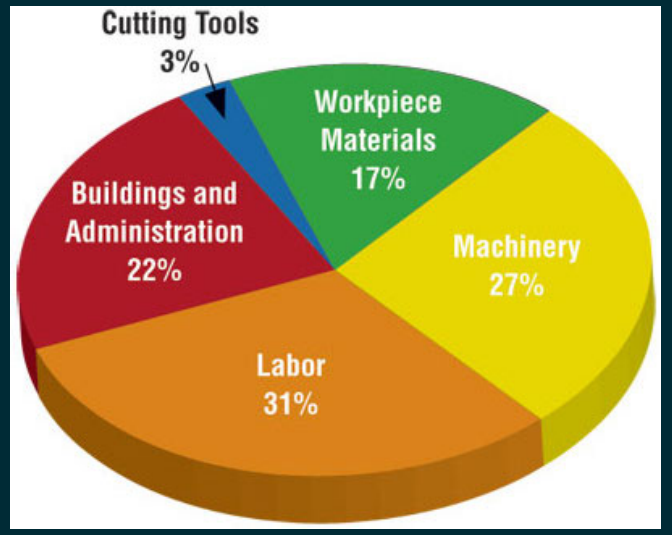
## Data Science in a Box

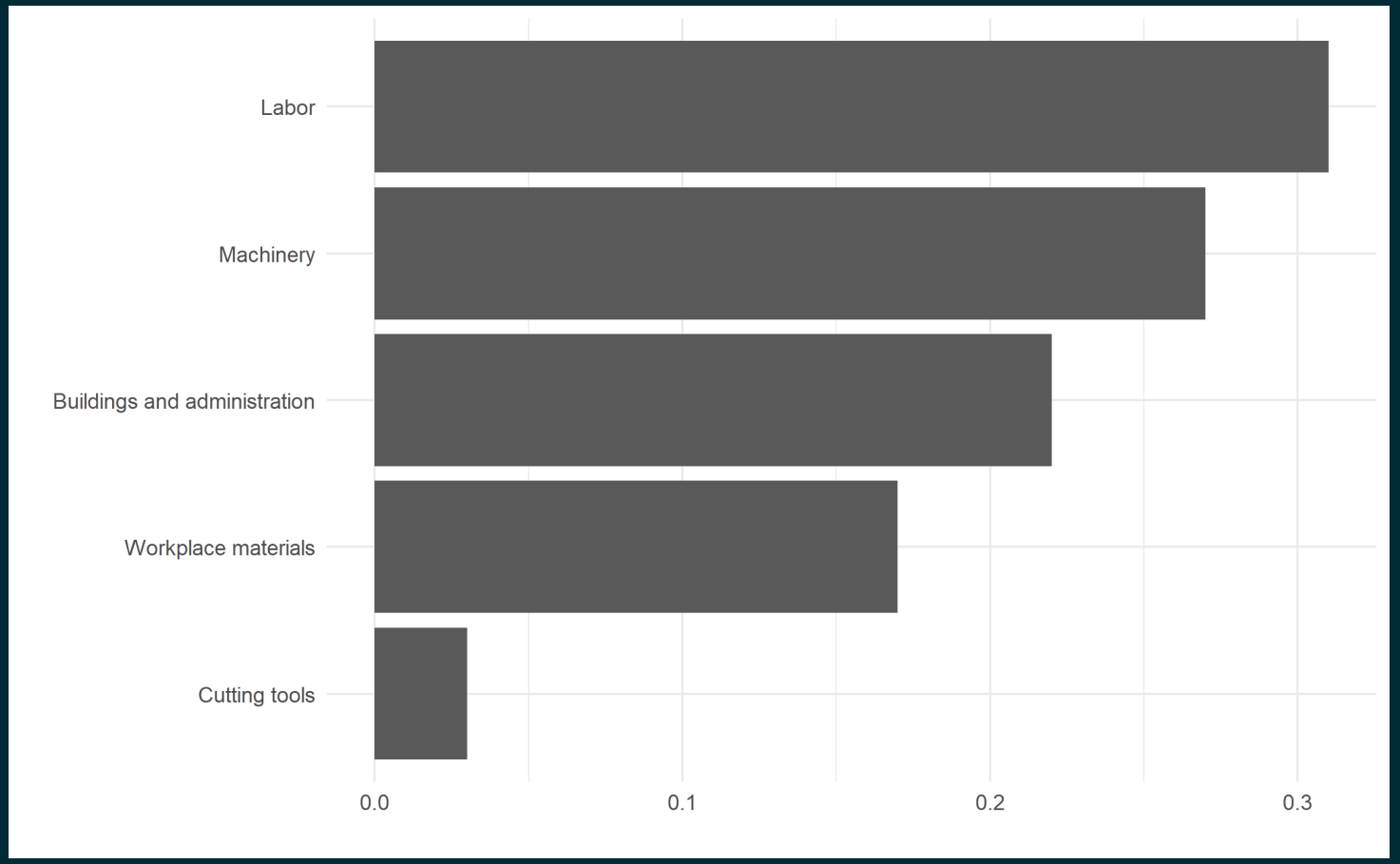
datasciencebox.org

Modified by Tyler George

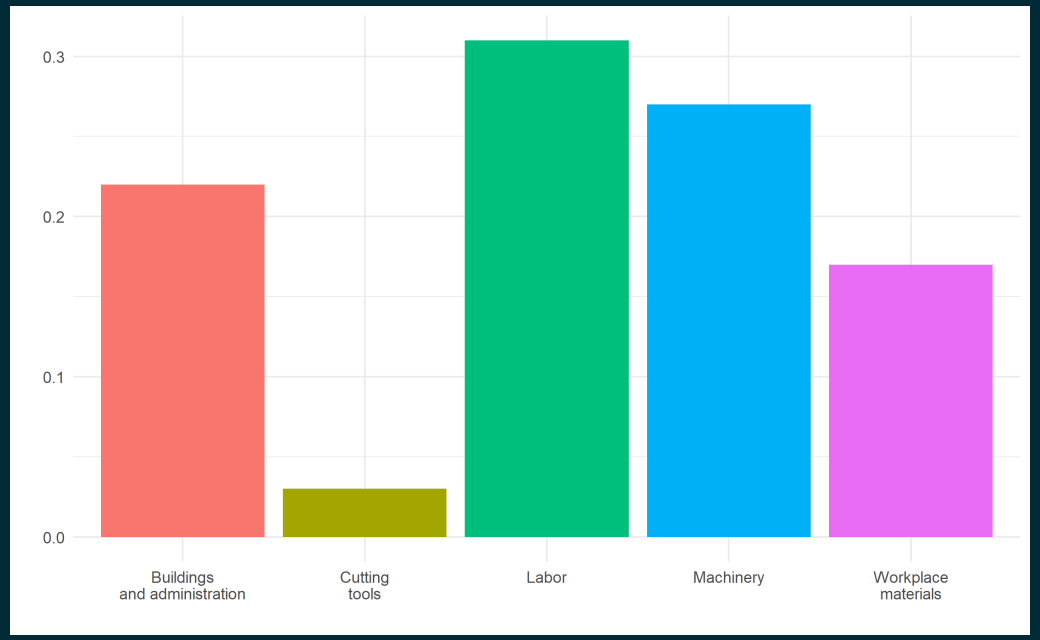
# Designing effective visualizations

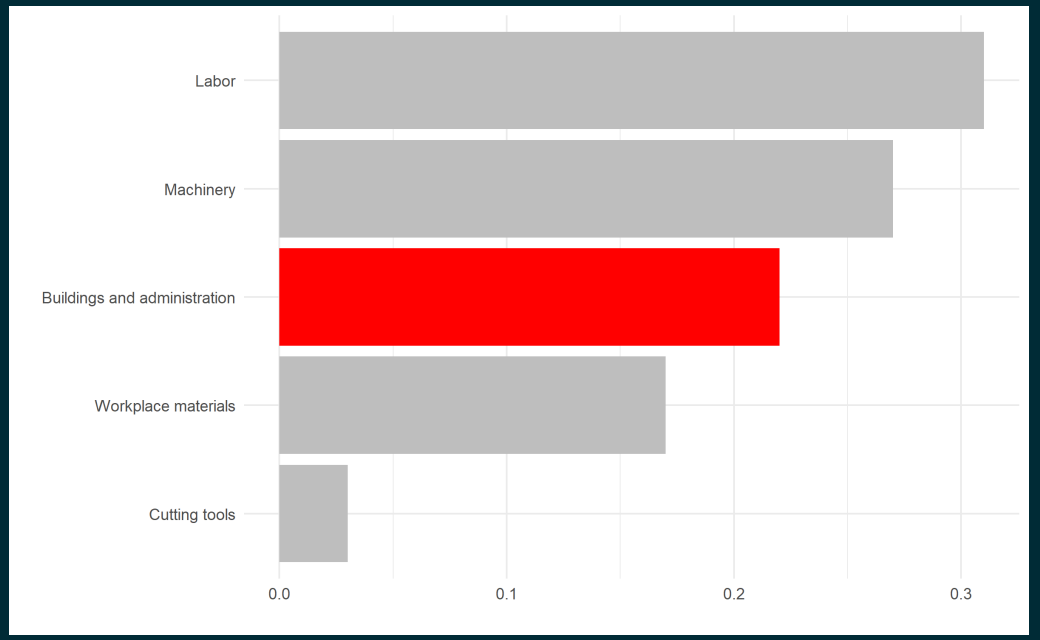
# Keep it simple



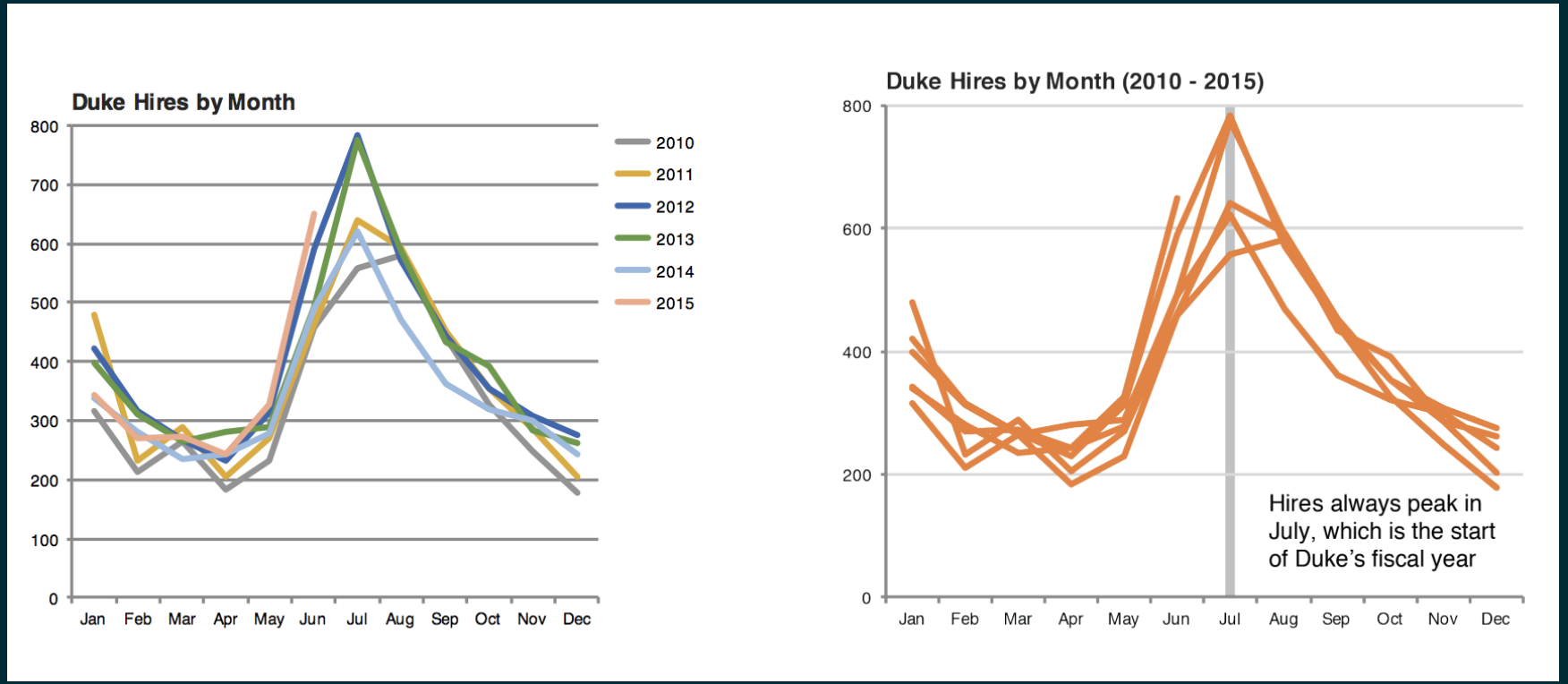



# Use color to draw attention



# Tell a story




Credit: Angela Zoss and Eric Monson, Duke DVS

# Principles for effective visualizations

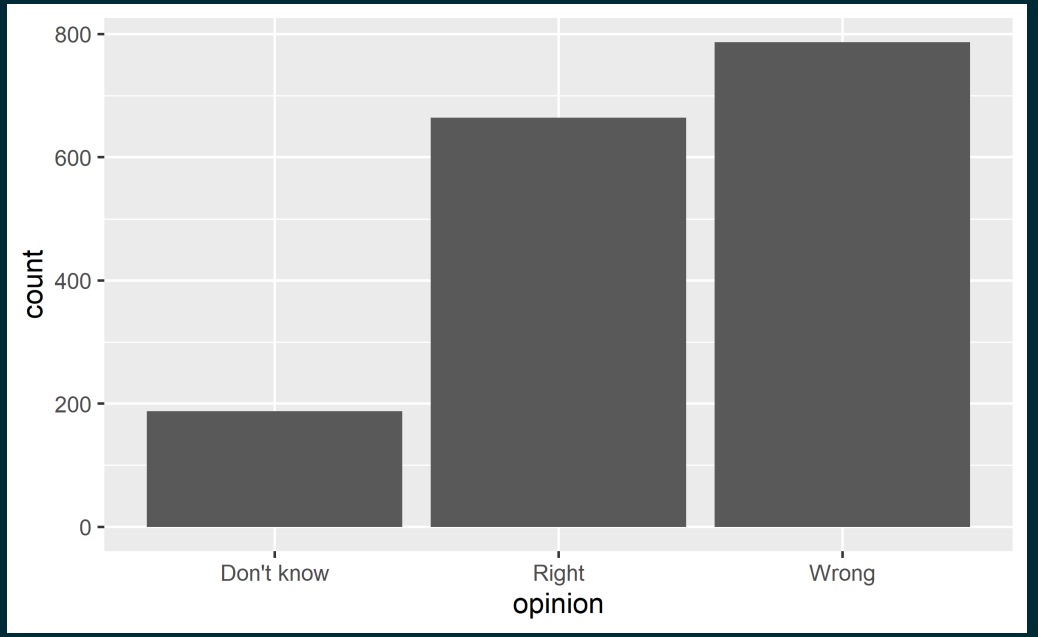
# Principles for effective visualizations

- Order matters
- Put long categories on the y-axis
- Keep scales consistent
- Select meaningful colors
- Use meaningful and nonredundant labels

# Data

In September 2019, YouGov survey asked 1,639 GB adults the following question:

> In hindsight, do you think Britain was right/wrong to vote to leave EU?
>
> - Right to leave
> - Wrong to leave
> - Don't know

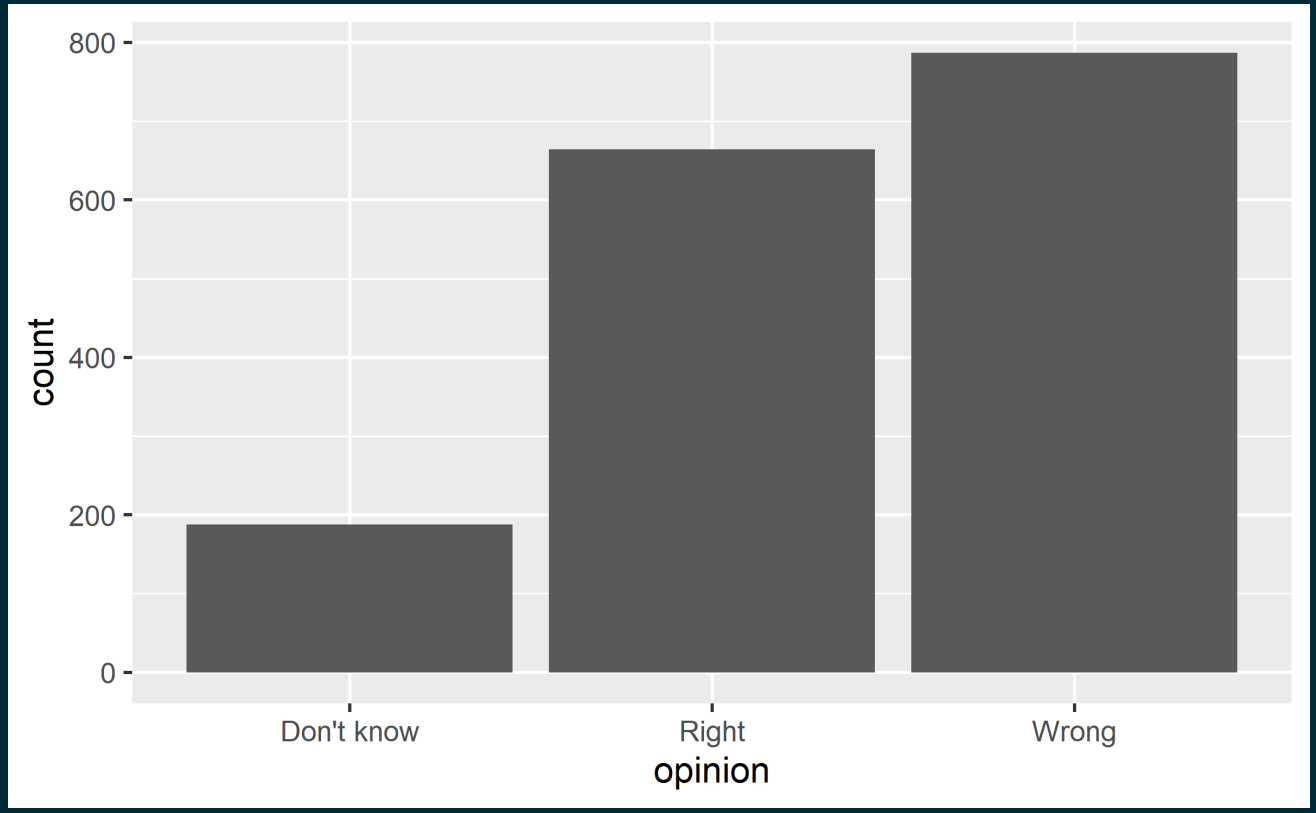



Source: YouGov Survey Results, retrieved Oct 7, 2019

# Order matters

# Alphabetical order is rarely ideal

Plot Code



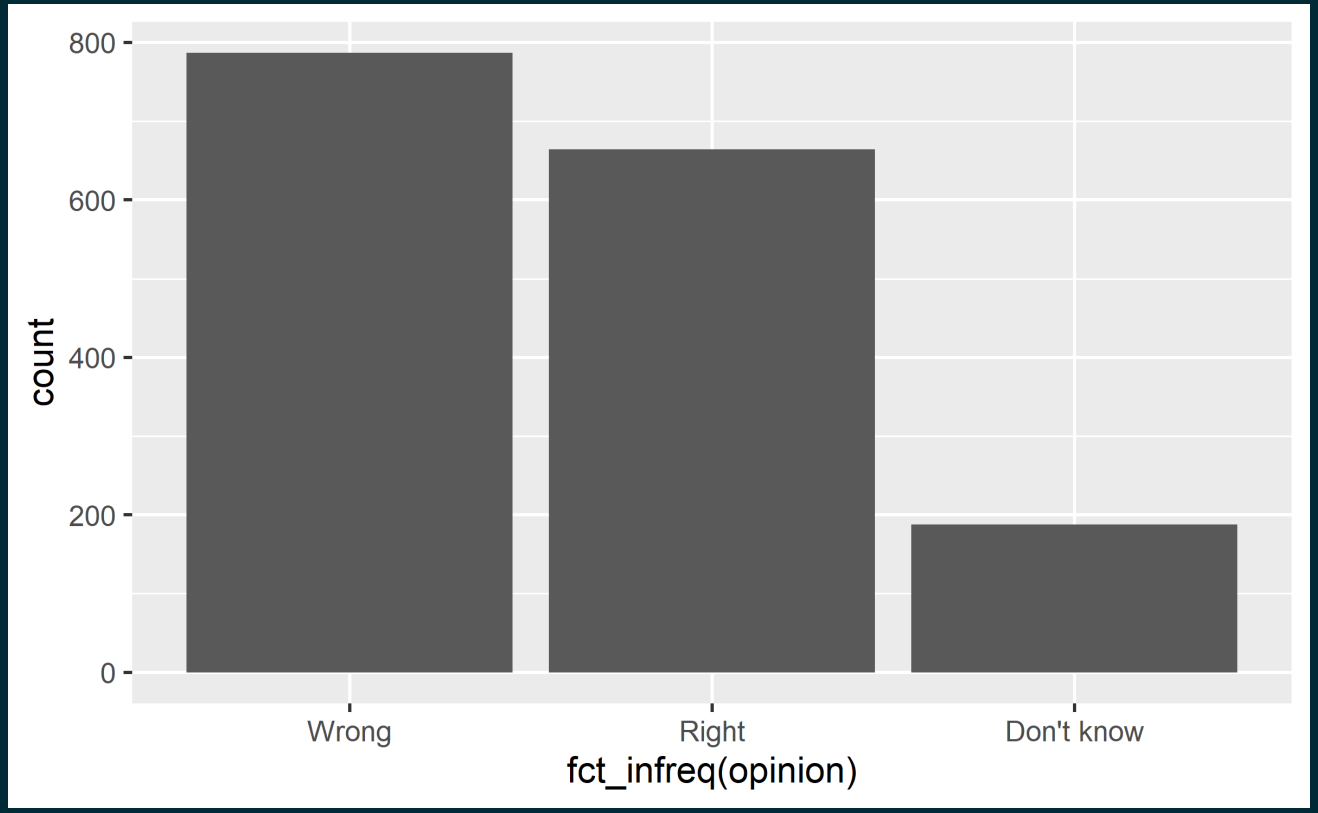
# Alphabetical order is rarely ideal

Plot | Code

```
ggplot(brexit, aes(x = opinion)) +
  geom_bar()
```

# Order by frequency

Plot Code



# Order by frequency

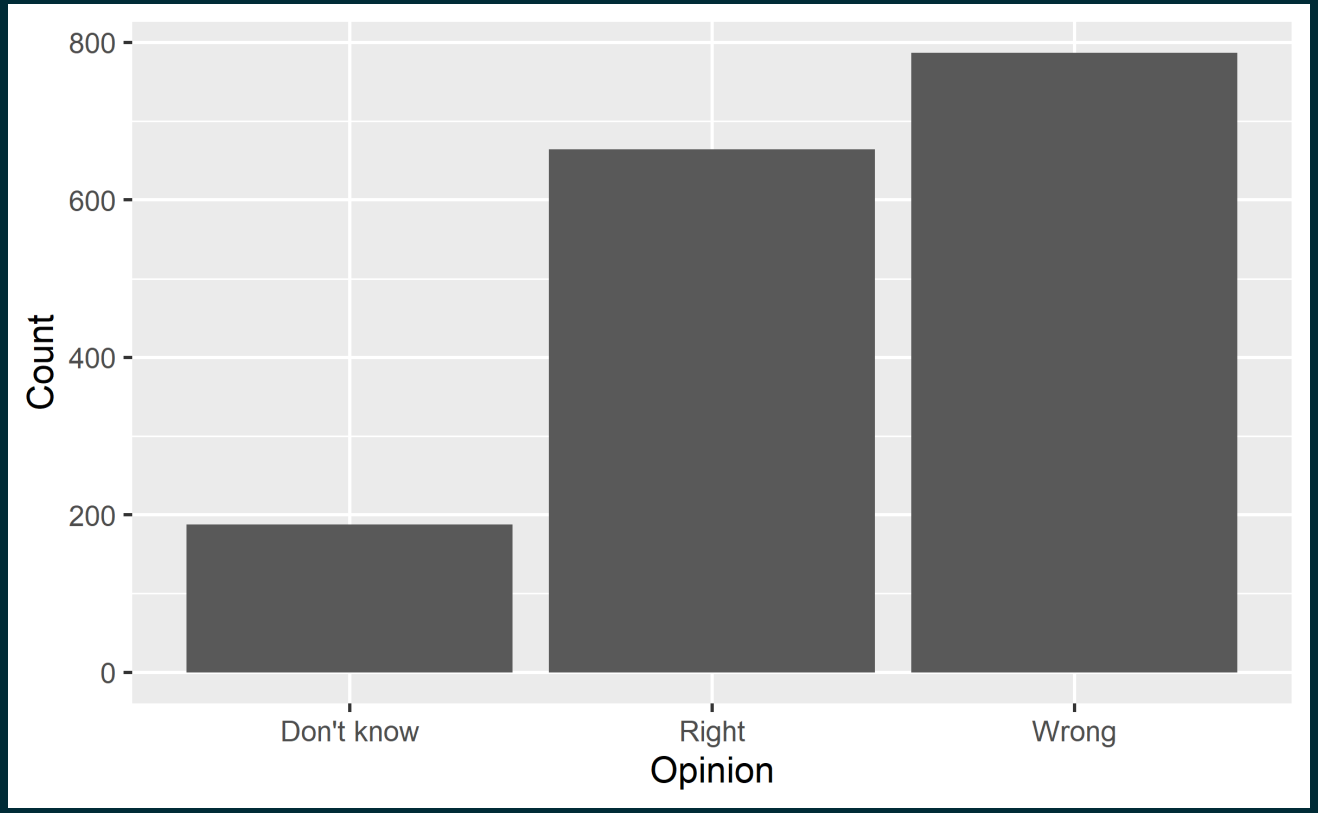
Plot | Code

`fct_infreq`: Reorder factors' levels by frequency

```
ggplot(brexit, aes(x = fct_infreq(opinion))) +
  geom_bar()
```

# Clean up labels

Plot Code



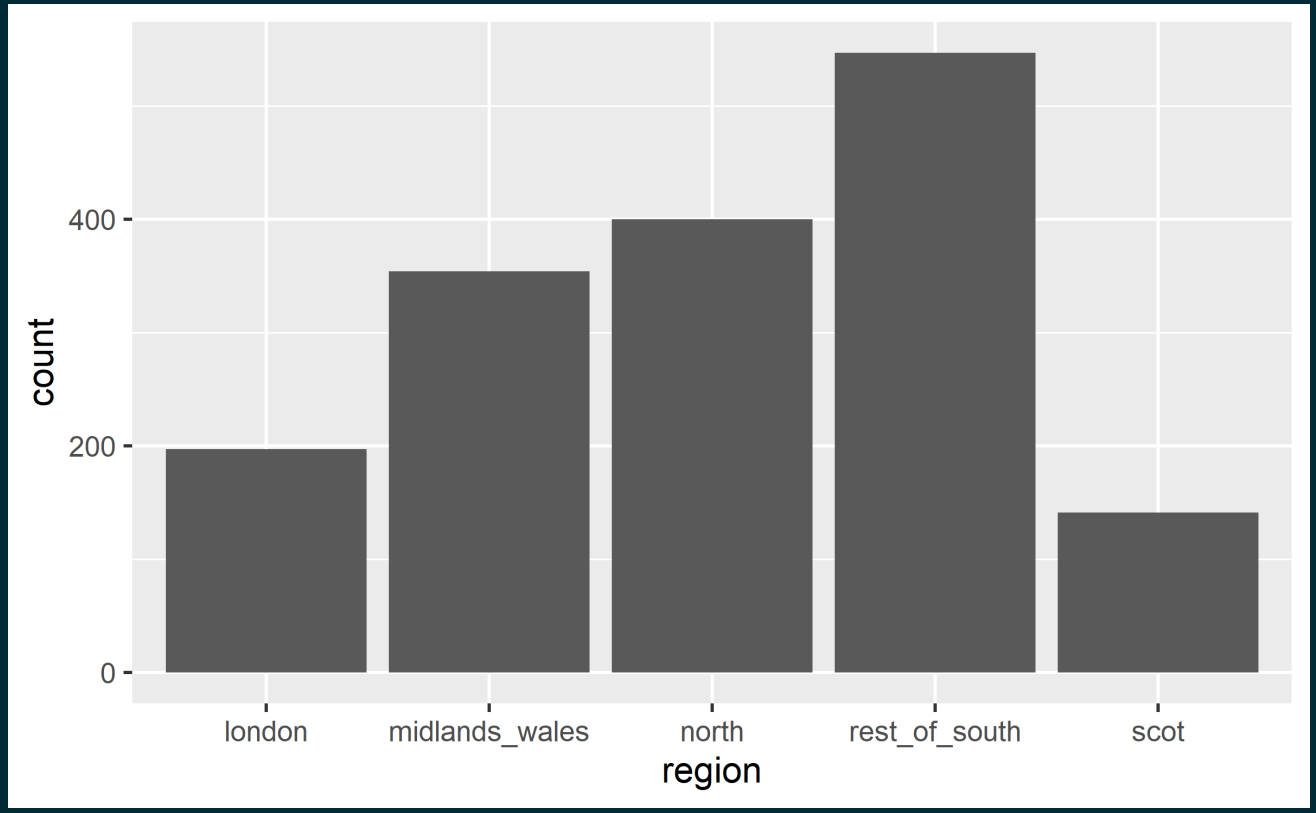
# Clean up labels

Plot | Code

```
ggplot(brexit, aes(x = opinion)) +
  geom_bar() +
  labs(
    x = "Opinion",
    y = "Count"
  )
```

# Alphabetical order is rarely ideal

Plot Code



# Alphabetical order is rarely ideal

Plot | Code

```
ggplot(brexit, aes(x = region)) +
  geom_bar()
```

# Use inherent level order

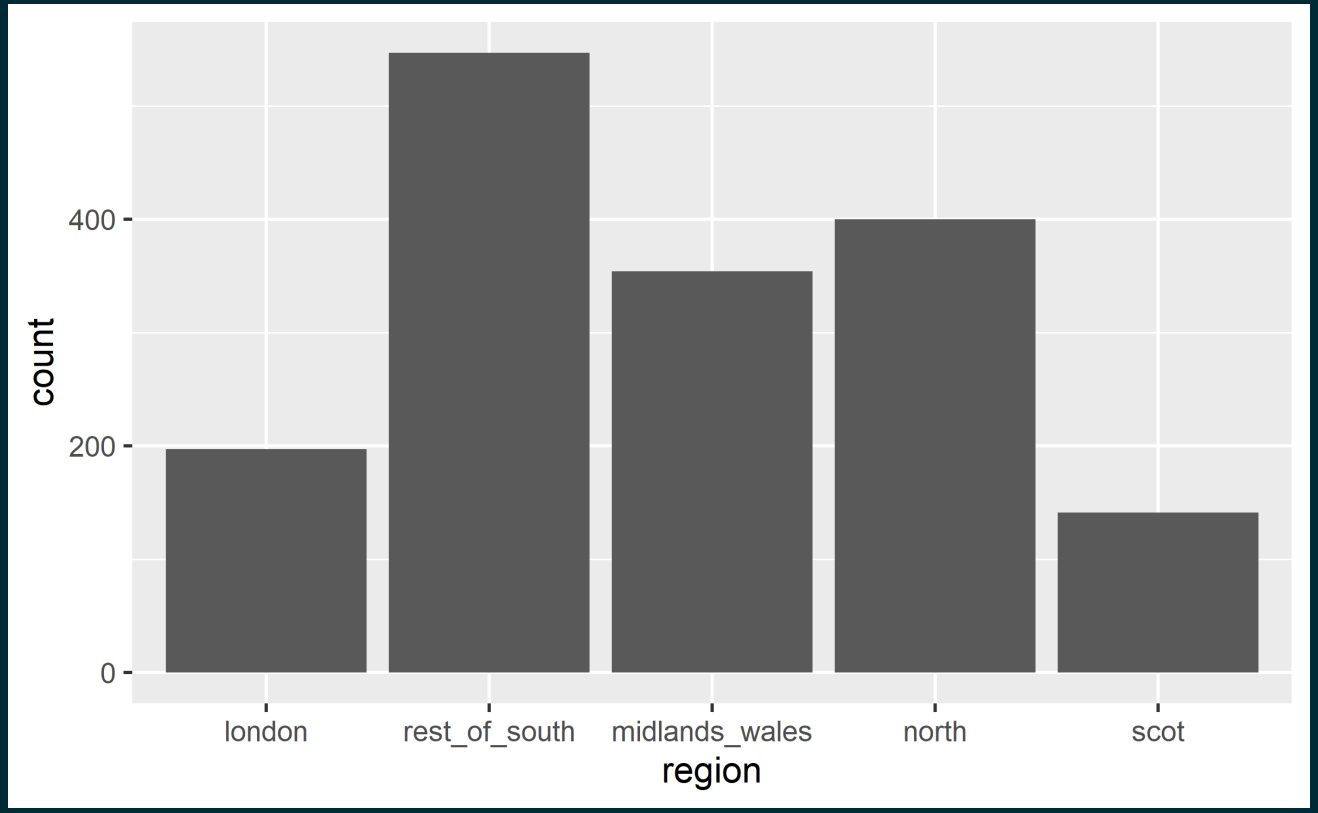
Relevel | Plot

`fct_relevel`: Reorder factor levels using a custom order

```
brexit <- brexit %>%
  mutate(
    region = fct_relevel(
      region,
      "london", "rest_of_south", "midlands_wales", "north", "scot"
    )
  )
```

# Use inherent level order

Relevel | Plot



# Clean up labels

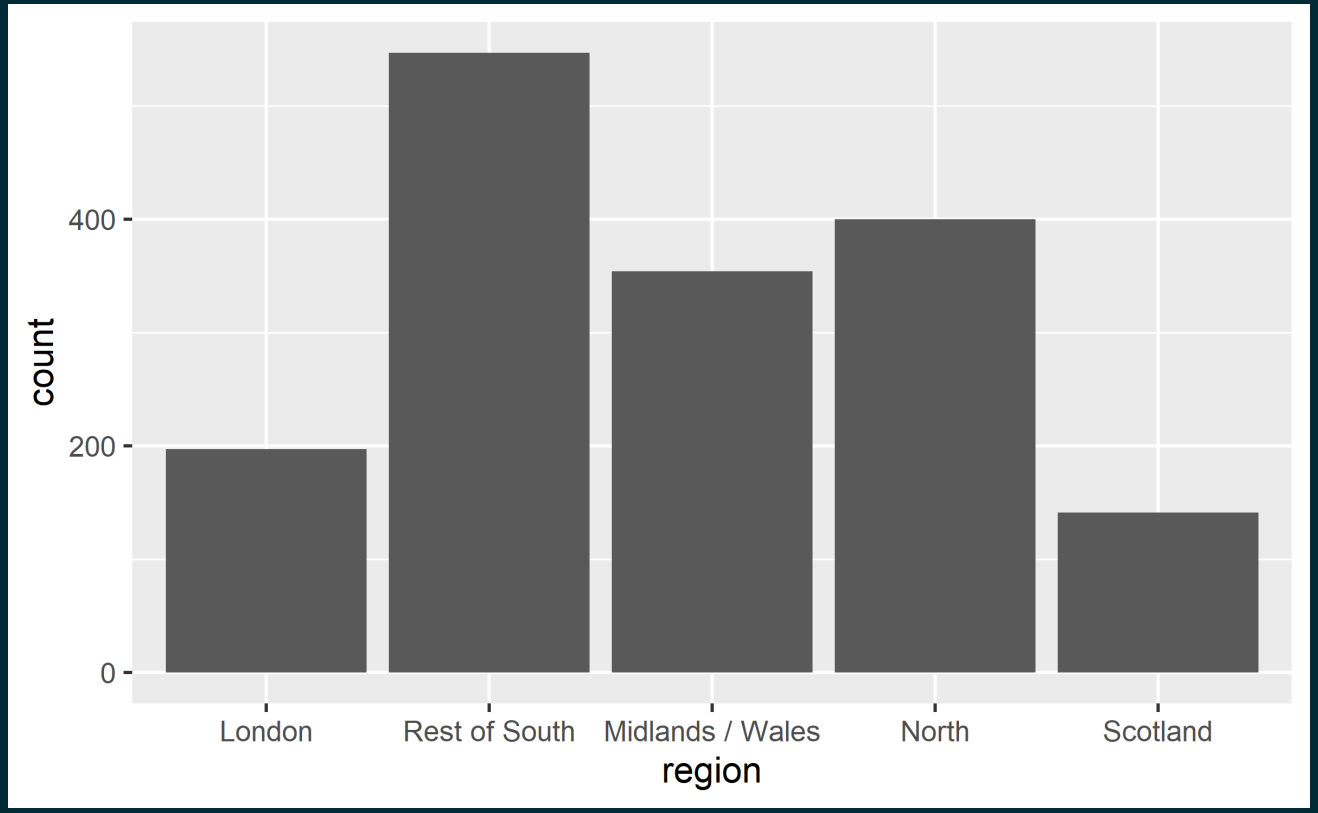
Recode | Plot

`fct_recode`: Change factor levels by hand

```
brexit <- brexit %>%
  mutate(
    region = fct_recode(
      region,
      London = "london",
      `Rest of South` = "rest_of_south",
      `Midlands / Wales` = "midlands_wales",
      North = "north",
      Scotland = "scot"
    )
  )
```

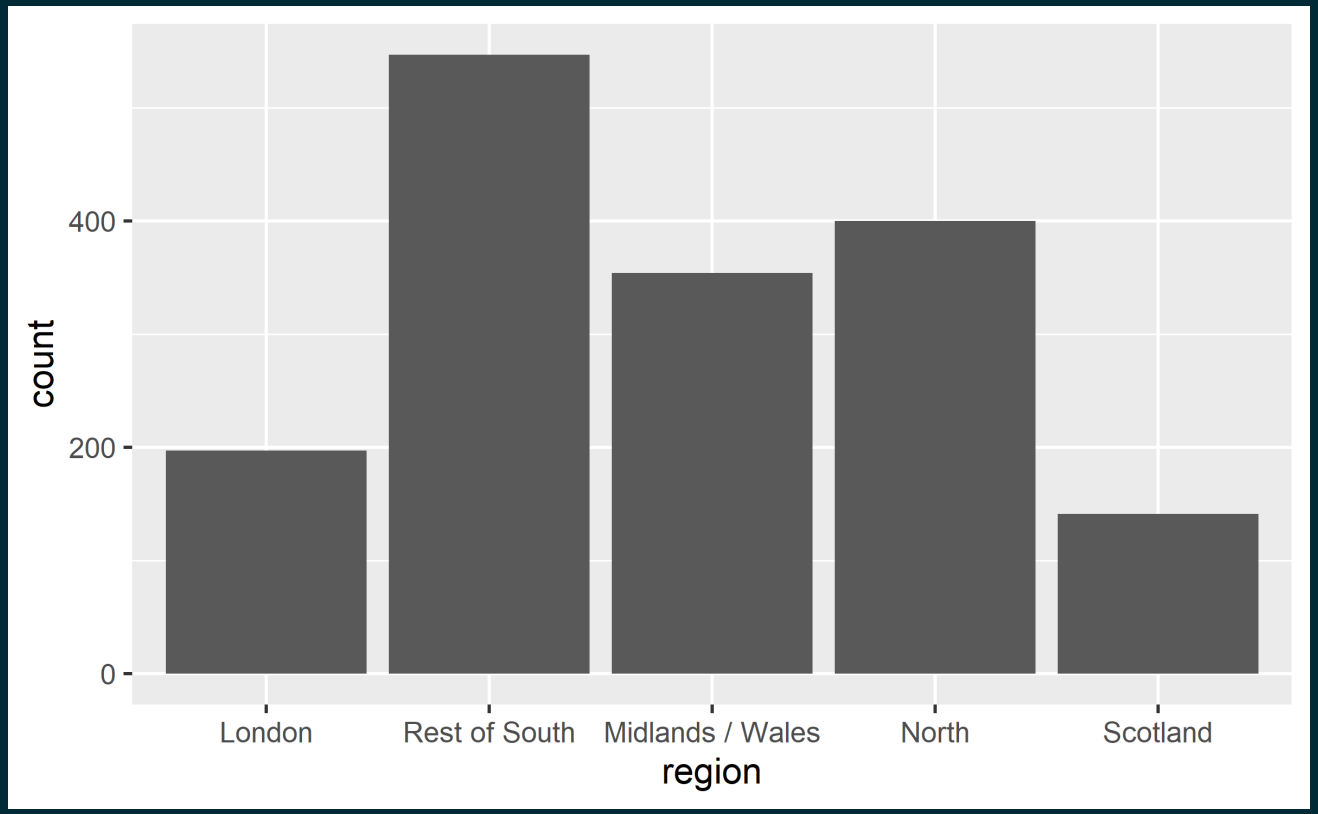
# Clean up labels

Recode | Plot



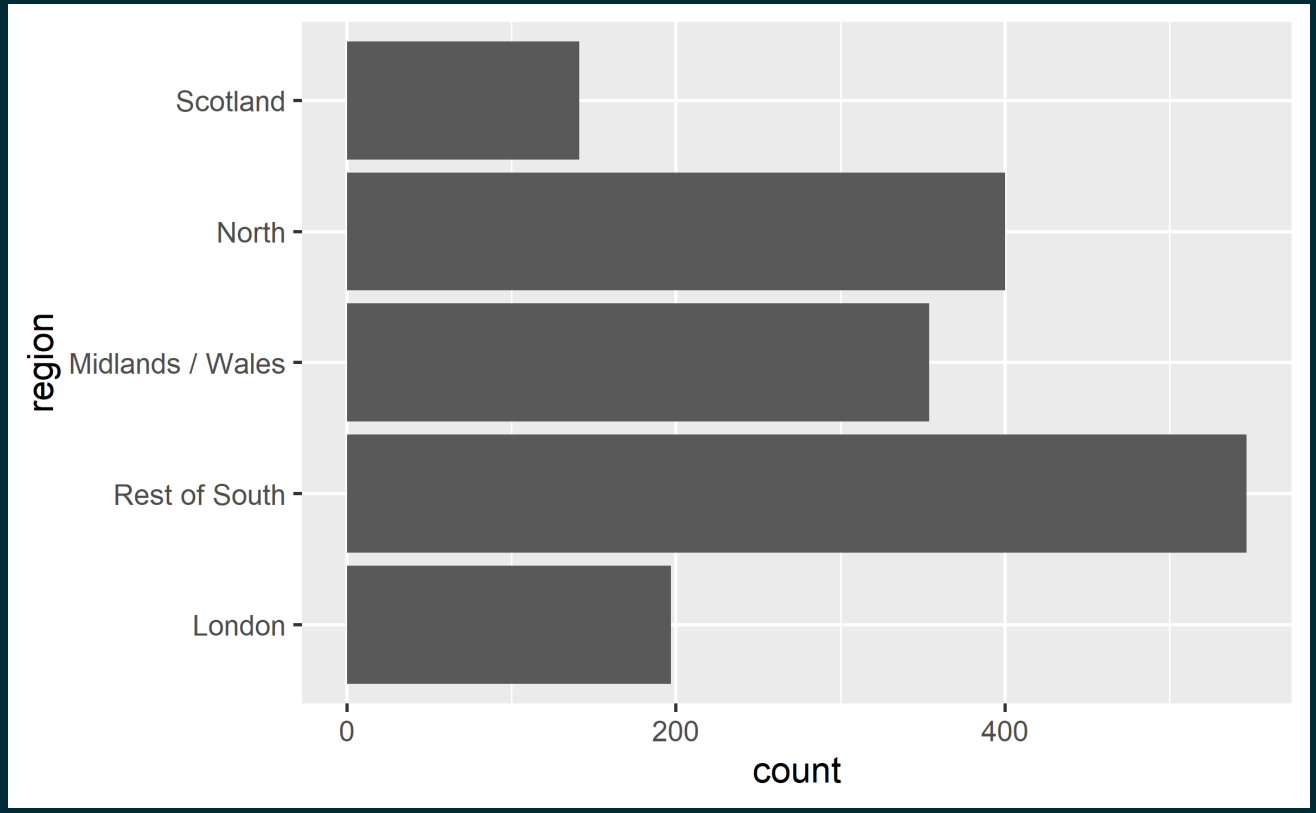
# Put long categories on the y-axis

# Long categories can be hard to read



# Move them to the y-axis

Plot Code



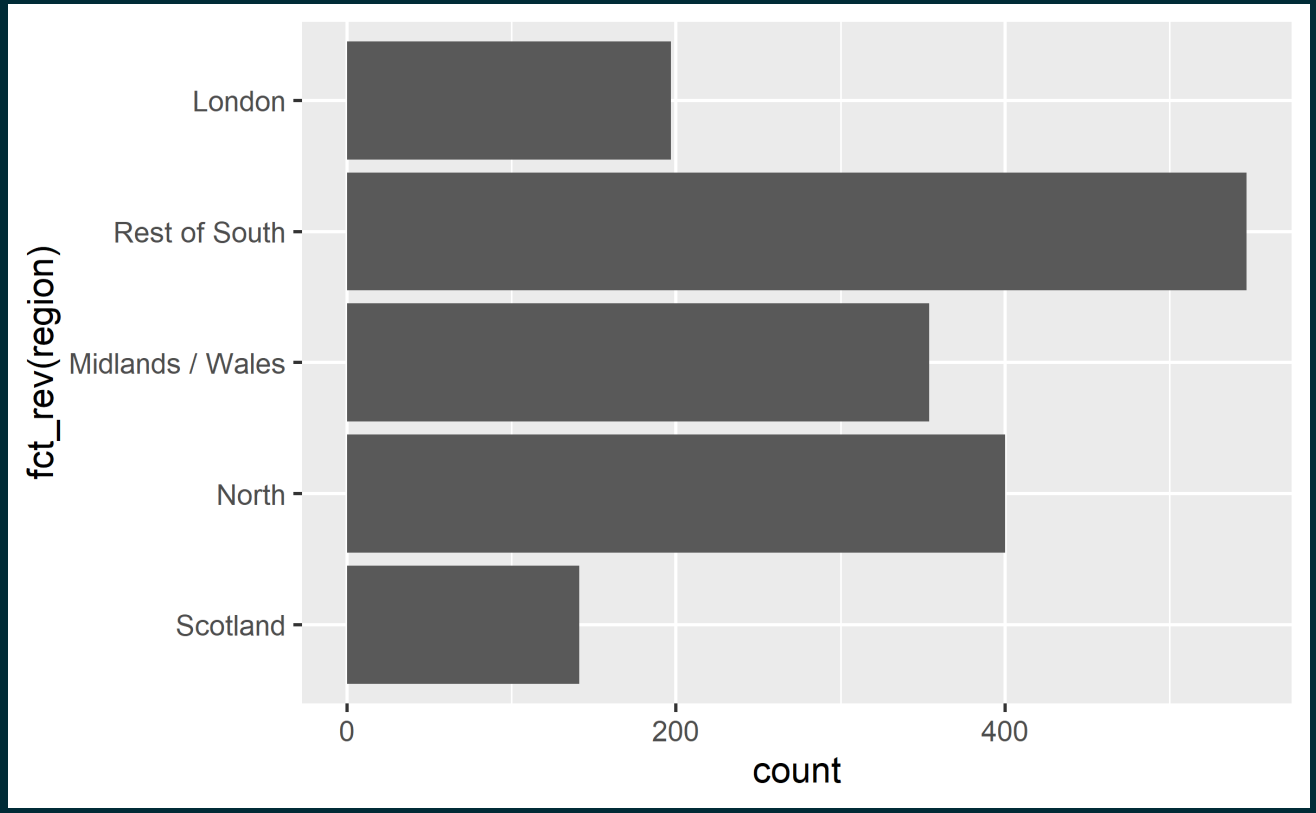
# Move them to the y-axis

Plot | Code

```
ggplot(brexit, aes(y = region)) +
  geom_bar()
```

# And reverse the order of levels

Plot Code



# And reverse the order of levels

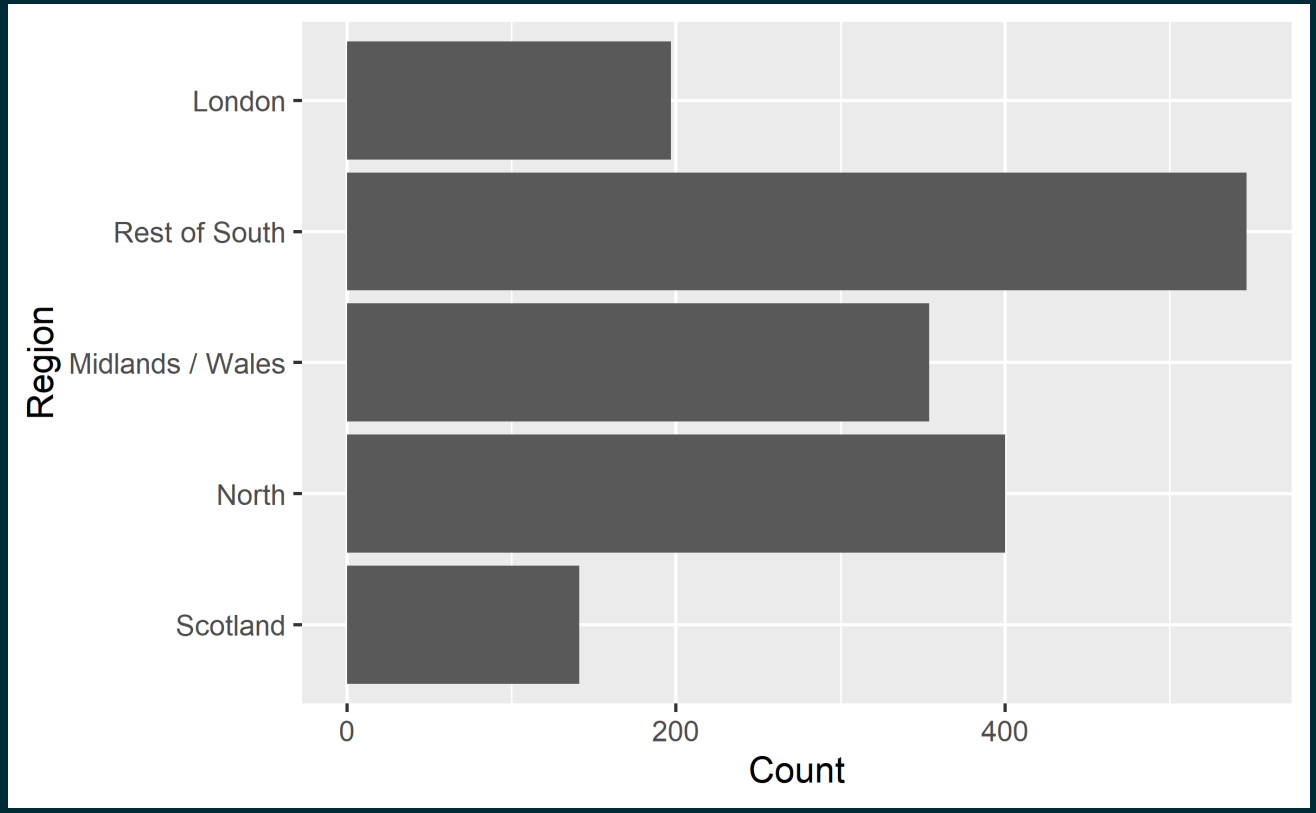
Plot | Code

`fct_rev`: Reverse order of factor levels

```
ggplot(brexit, aes(y = fct_rev(region))) +
  geom_bar()
```

# Clean up labels

Plot Code



# Clean up labels

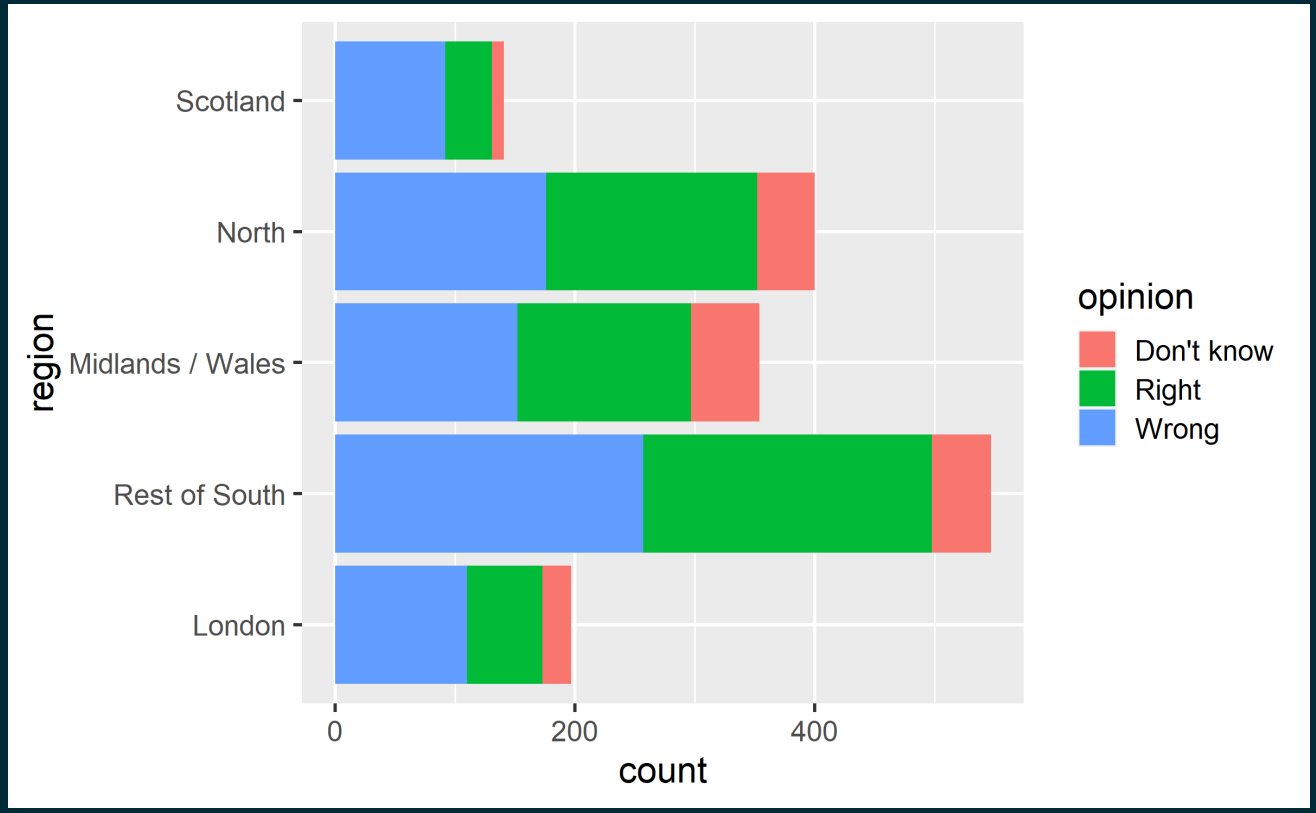
Plot | Code

```
ggplot(brexit, aes(y = fct_rev(region))) +
  geom_bar() +
  labs(
    x = "Count",
    y = "Region"
  )
```

# Pick a purpose

# Segmented bar plots can be hard to read

Plot Code



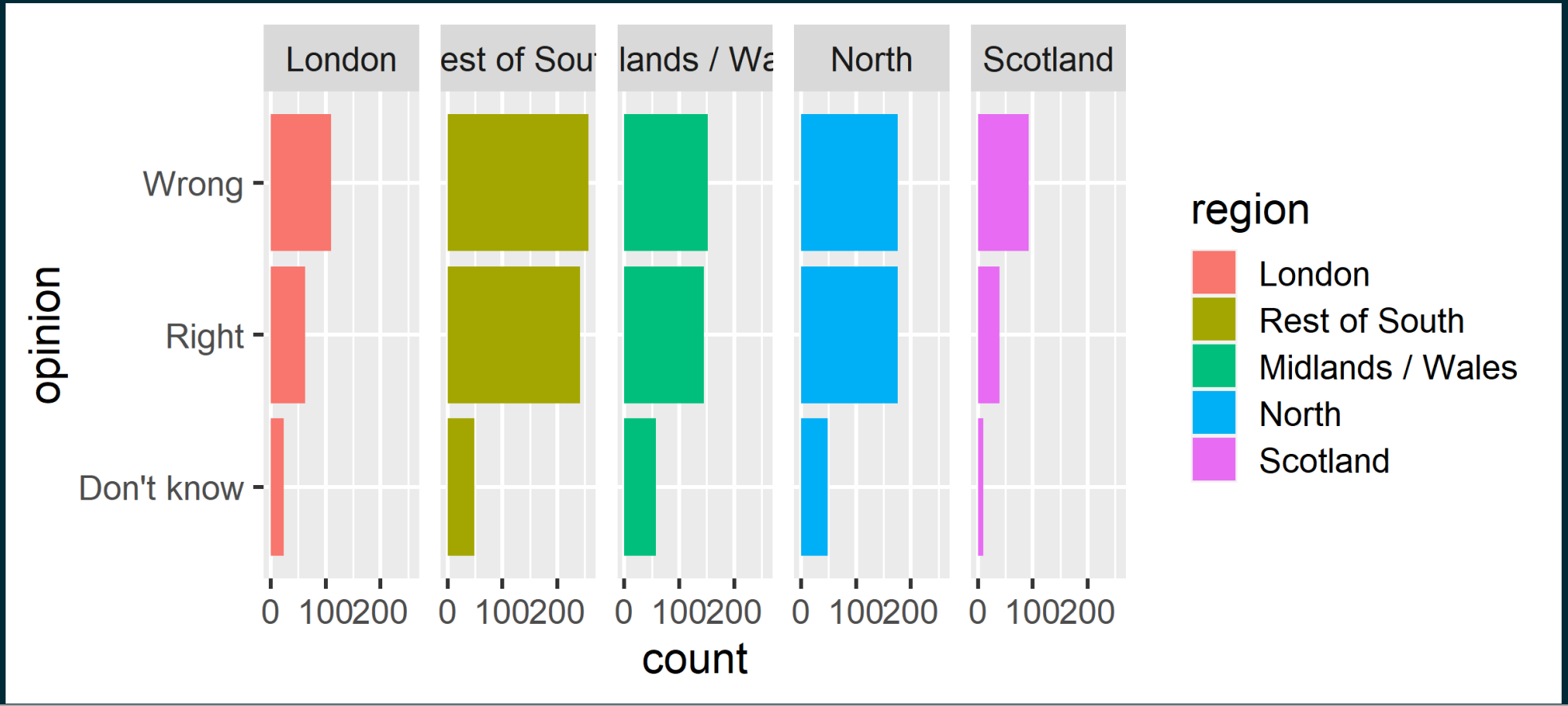
# Segmented bar plots can be hard to read

Plot | Code

```
ggplot(brexit, aes(y = region, fill = opinion)) +
  geom_bar()
```

# Use facets

Plot Code

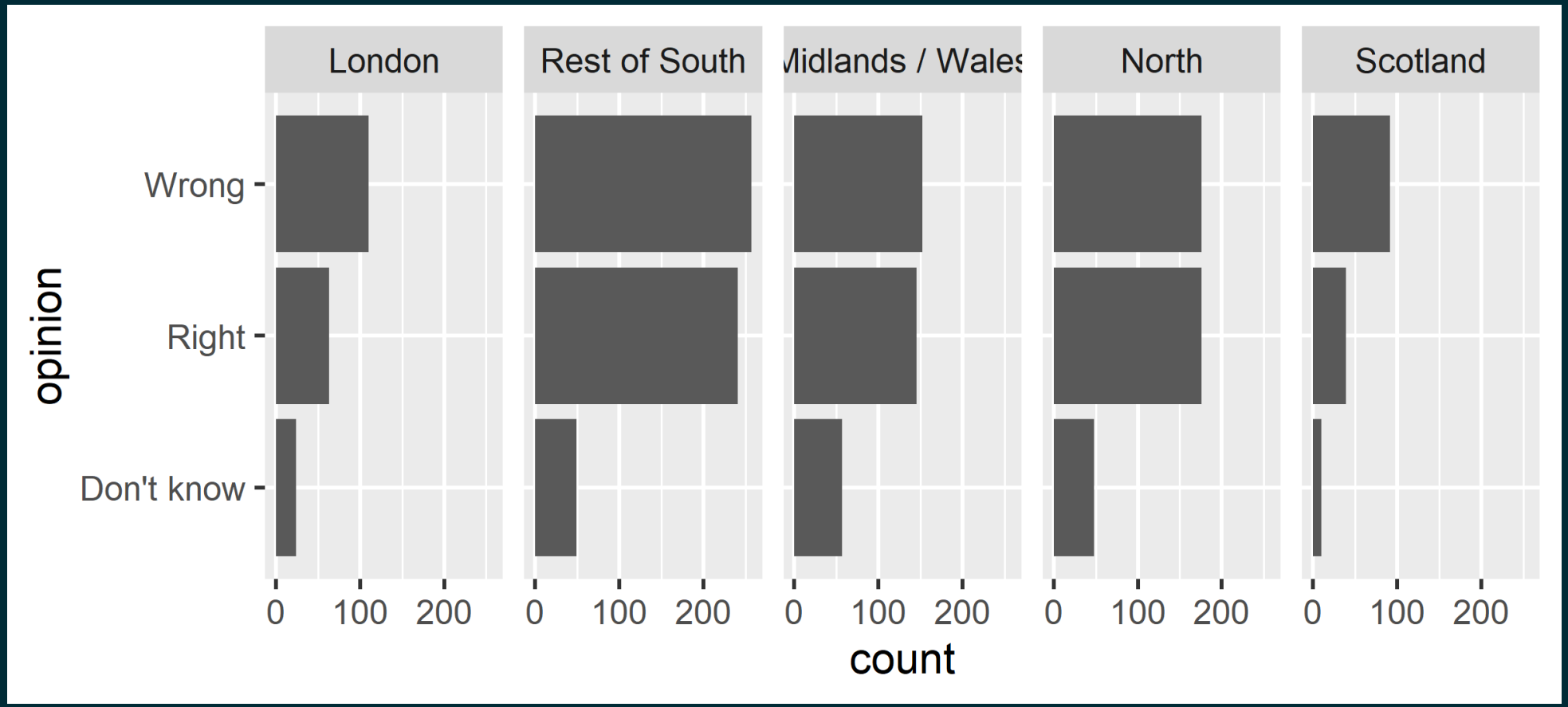


# Use facets

Plot | Code

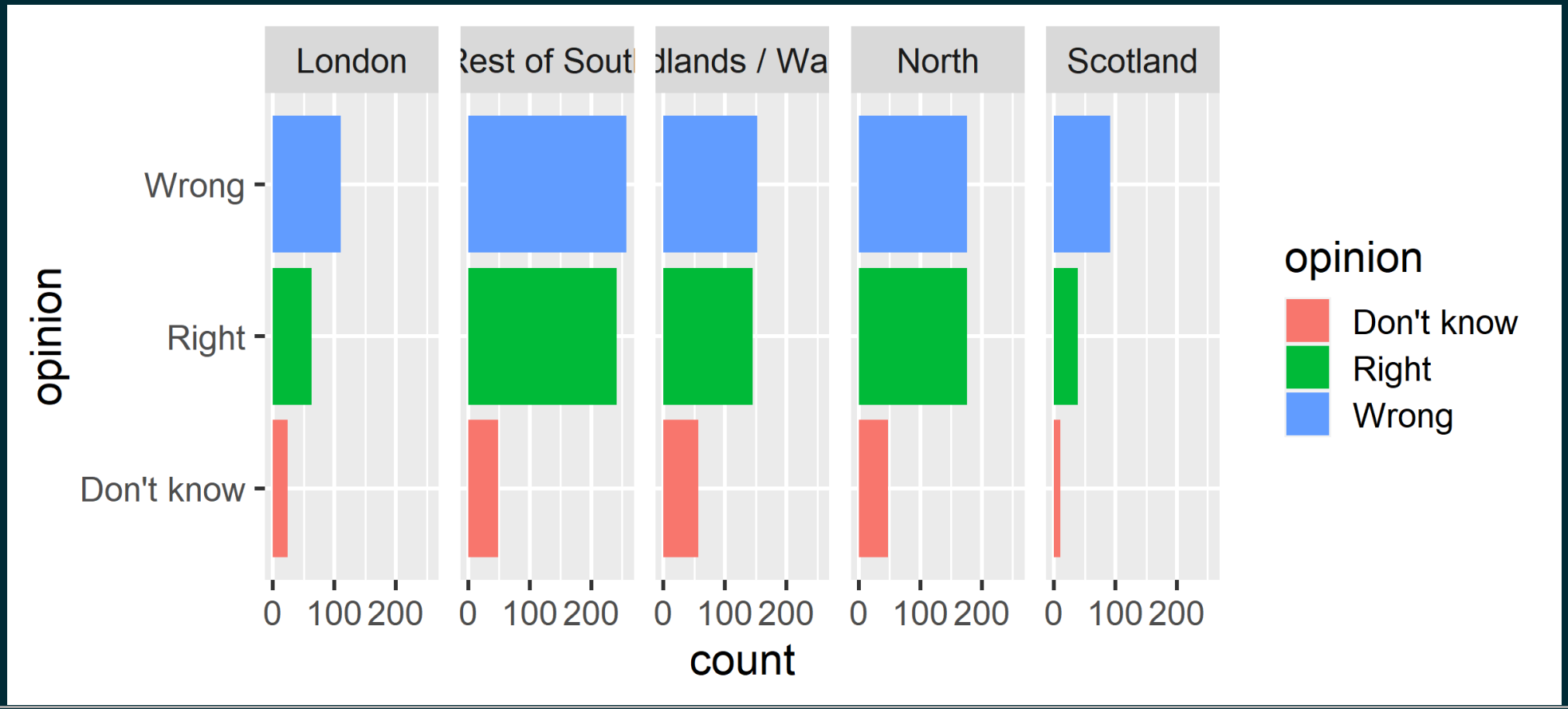
```
ggplot(brexit, aes(y = opinion, fill = region)) +
  geom_bar() +
  facet_wrap(~region, nrow = 1)
```

# Avoid redundancy?



# Redundancy can help tell a story

Plot Code



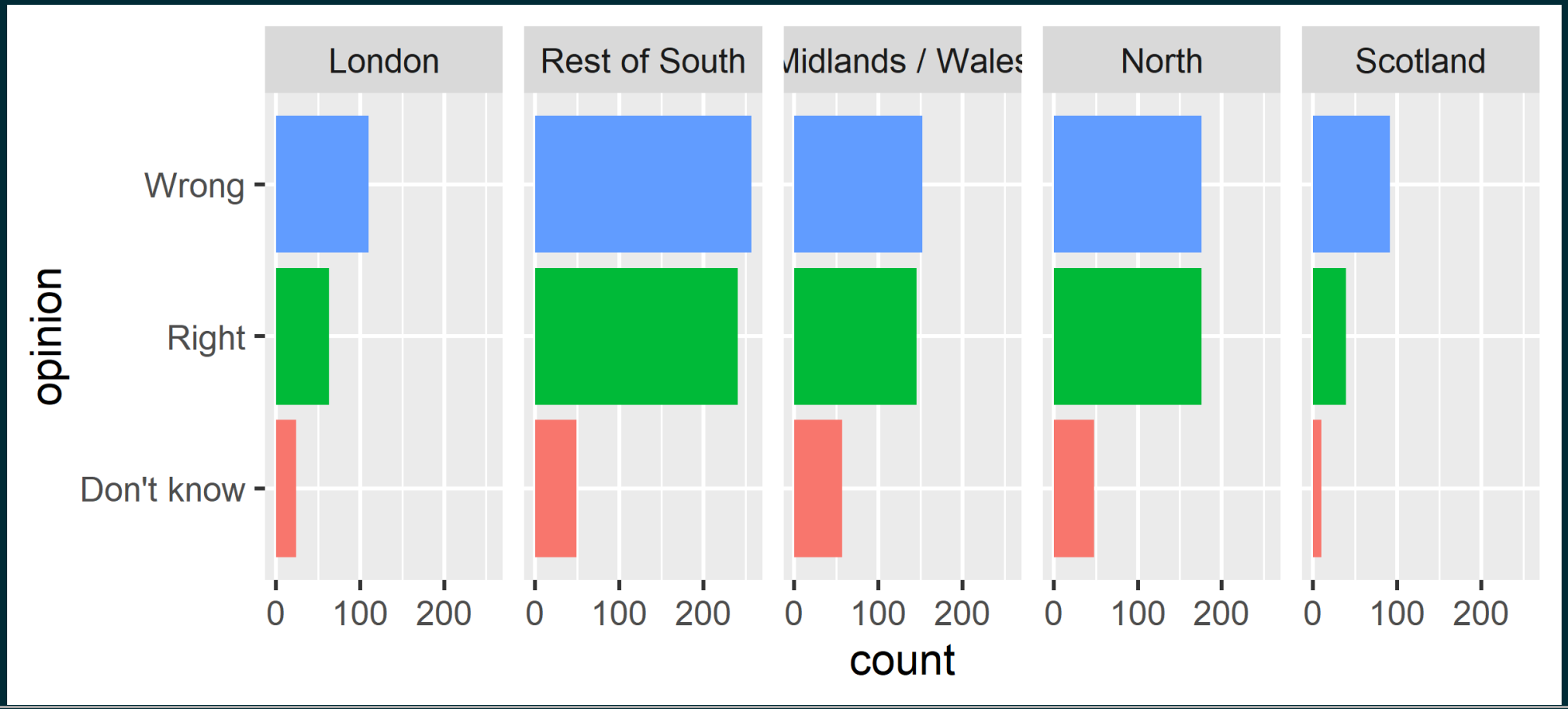
# Redundancy can help tell a story

Plot | Code

```
ggplot(brexit, aes(y = opinion, fill = opinion)) +
  geom_bar() +
  facet_wrap(~region, nrow = 1)
```

# Be selective with redundancy

Plot Code



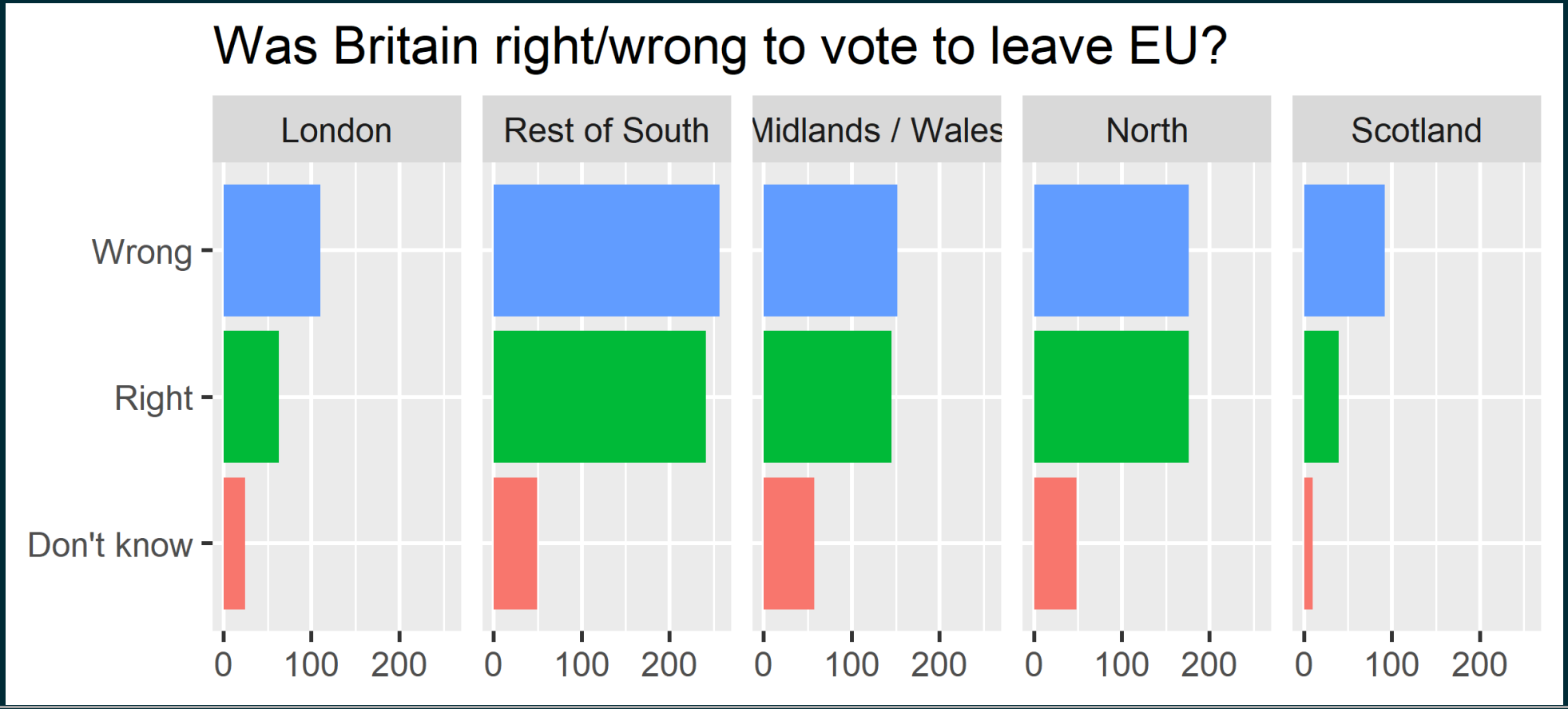
# Be selective with redundancy

Plot | Code

```
ggplot(brexit, aes(y = opinion, fill = opinion)) +
  geom_bar() +
  facet_wrap(~region, nrow = 1) +
  guides(fill = "none")
```

# Use informative labels

Plot Code



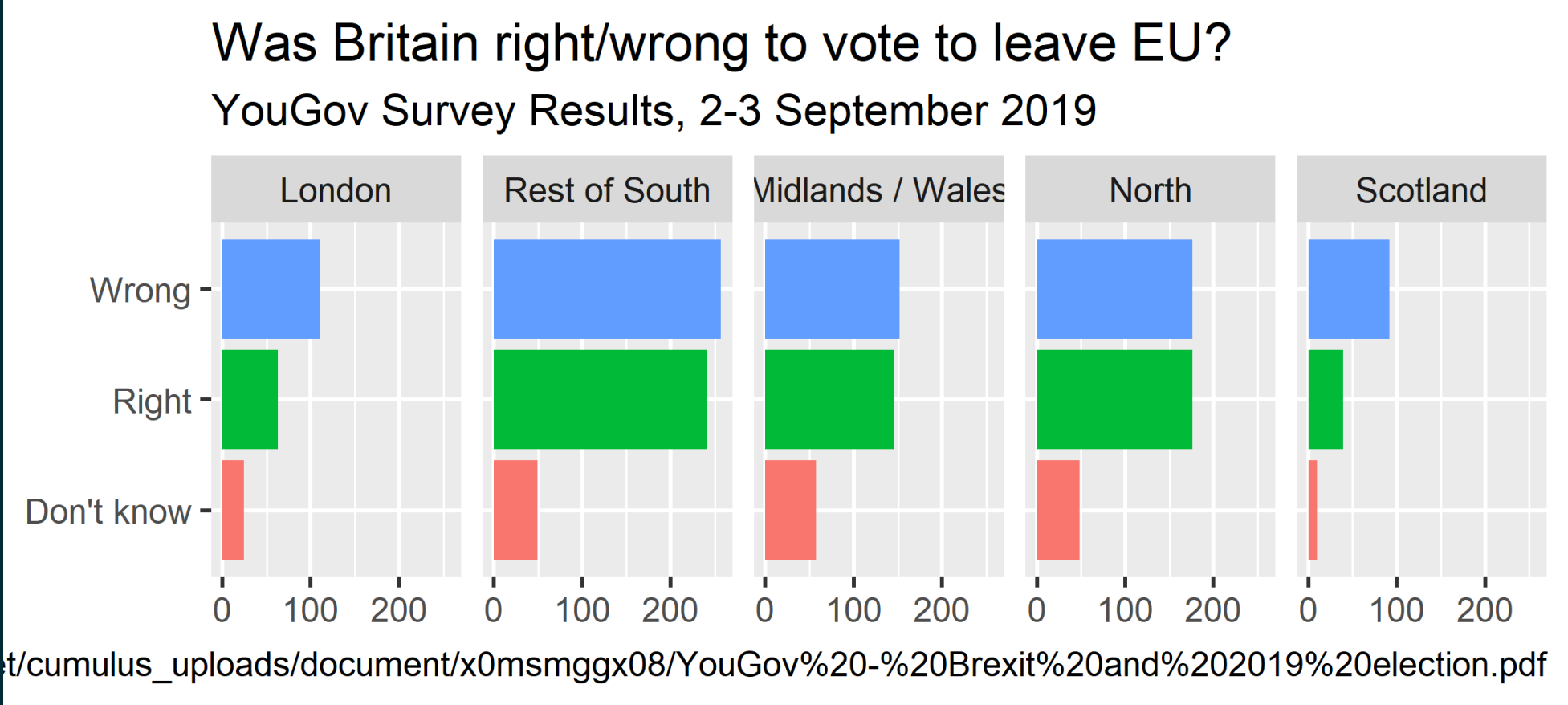
# Use informative labels

Plot | Code

```
ggplot(brexit, aes(y = opinion, fill = opinion)) +
  geom_bar() +
  facet_wrap(~region, nrow = 1) +
  guides(fill = "none") +
  labs(
    title = "Was Britain right/wrong to vote to leave EU?",
    x = NULL, y = NULL
  )
```

# A bit more info

Plot Code



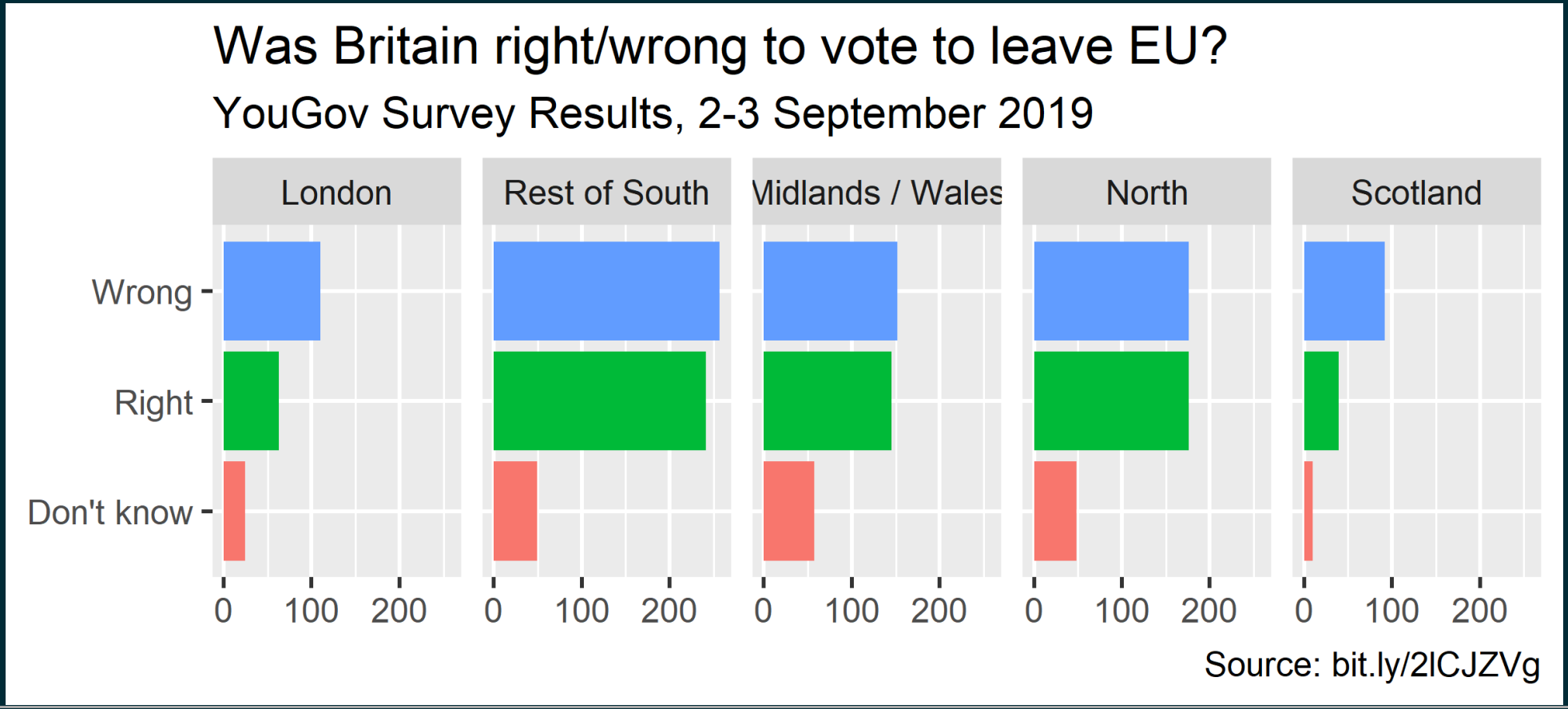
# A bit more info

Plot | Code

```
ggplot(brexit, aes(y = opinion, fill = opinion)) +
  geom_bar() +
  facet_wrap(~region, nrow = 1) +
  guides(fill = "none") +
  labs(
    title = "Was Britain right/wrong to vote to leave EU?",
    subtitle = "YouGov Survey Results, 2-3 September 2019",
    caption = "Source: https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/x0msmggx08/Y
    x = NULL, y = NULL
  )
```

# Let's do better

Plot Code



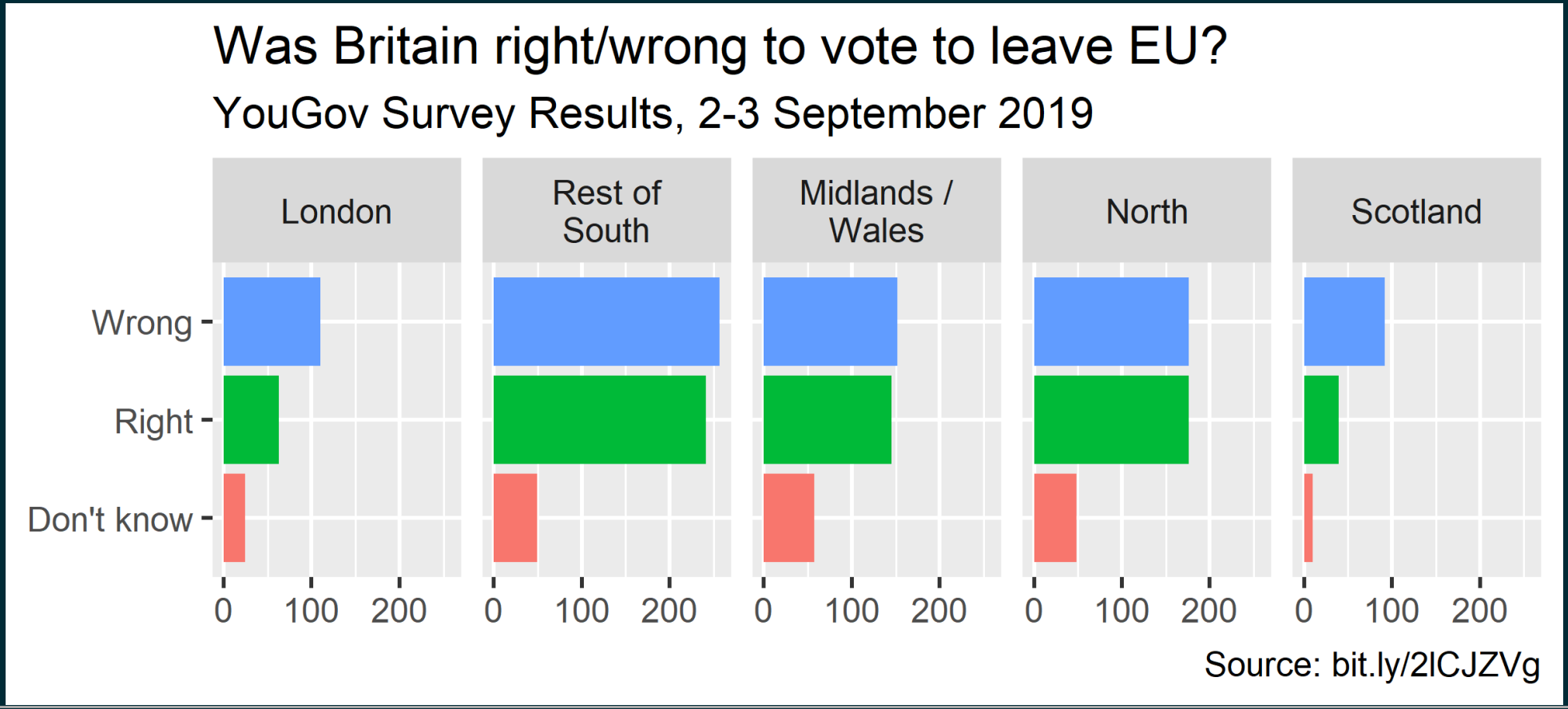
# Let's do better

Plot | Code

```
ggplot(brexit, aes(y = opinion, fill = opinion)) +
  geom_bar() +
  facet_wrap(~region, nrow = 1) +
  guides(fill = "none") +
  labs(
    title = "Was Britain right/wrong to vote to leave EU?",
    subtitle = "YouGov Survey Results, 2-3 September 2019",
    caption = "Source: bit.ly/2lCJZVg",
    x = NULL, y = NULL
  )
```

# Fix up facet labels

Plot Code



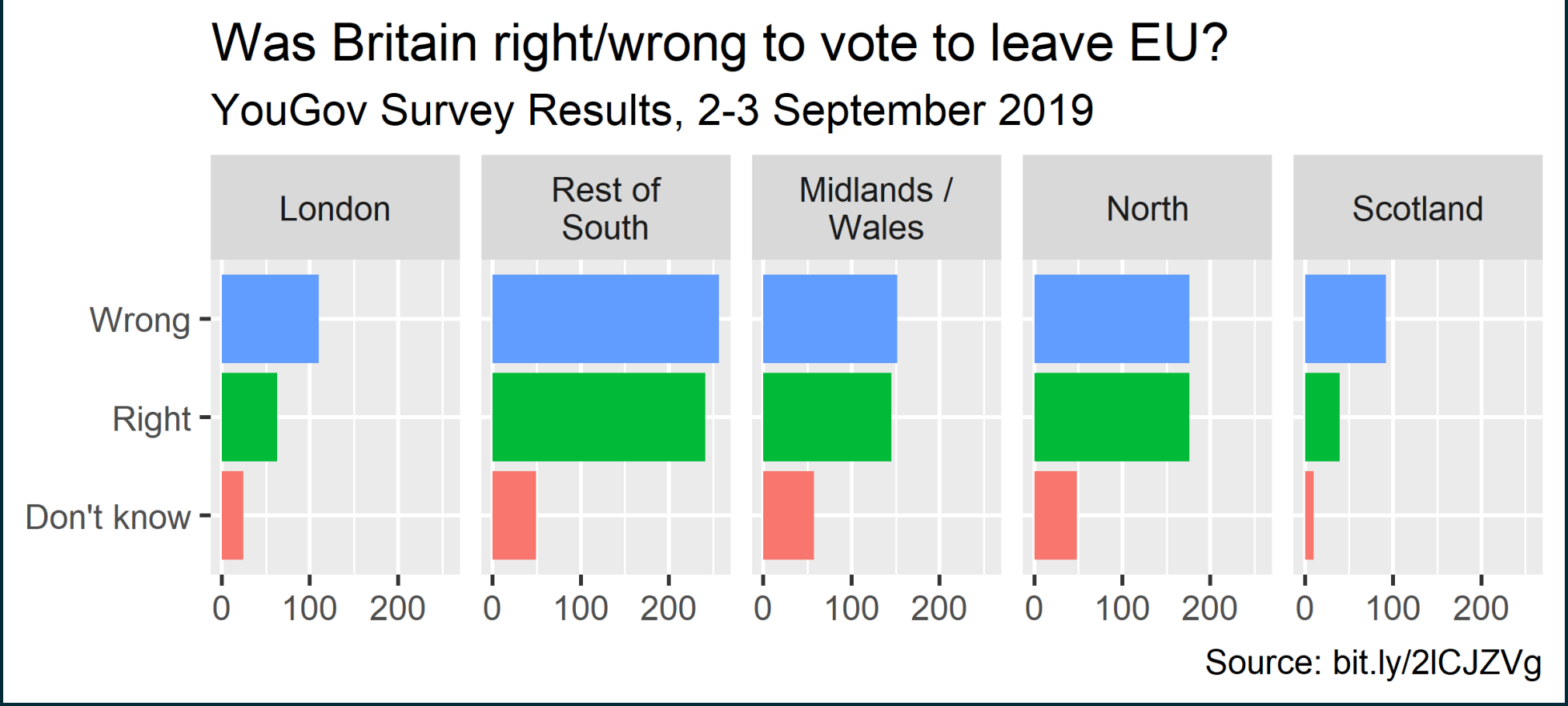
# Fix up facet labels

Plot | Code

```
ggplot(brexit, aes(y = opinion, fill = opinion)) +
  geom_bar() +
  facet_wrap(~region,
    nrow = 1,
    labeller = label_wrap_gen(width = 12)
  ) +
  guides(fill = "none") +
  labs(
    title = "Was Britain right/wrong to vote to leave EU?",
    subtitle = "YouGov Survey Results, 2-3 September 2019",
    caption = "Source: bit.ly/2lCJZVg",
    x = NULL, y = NULL
  )
```

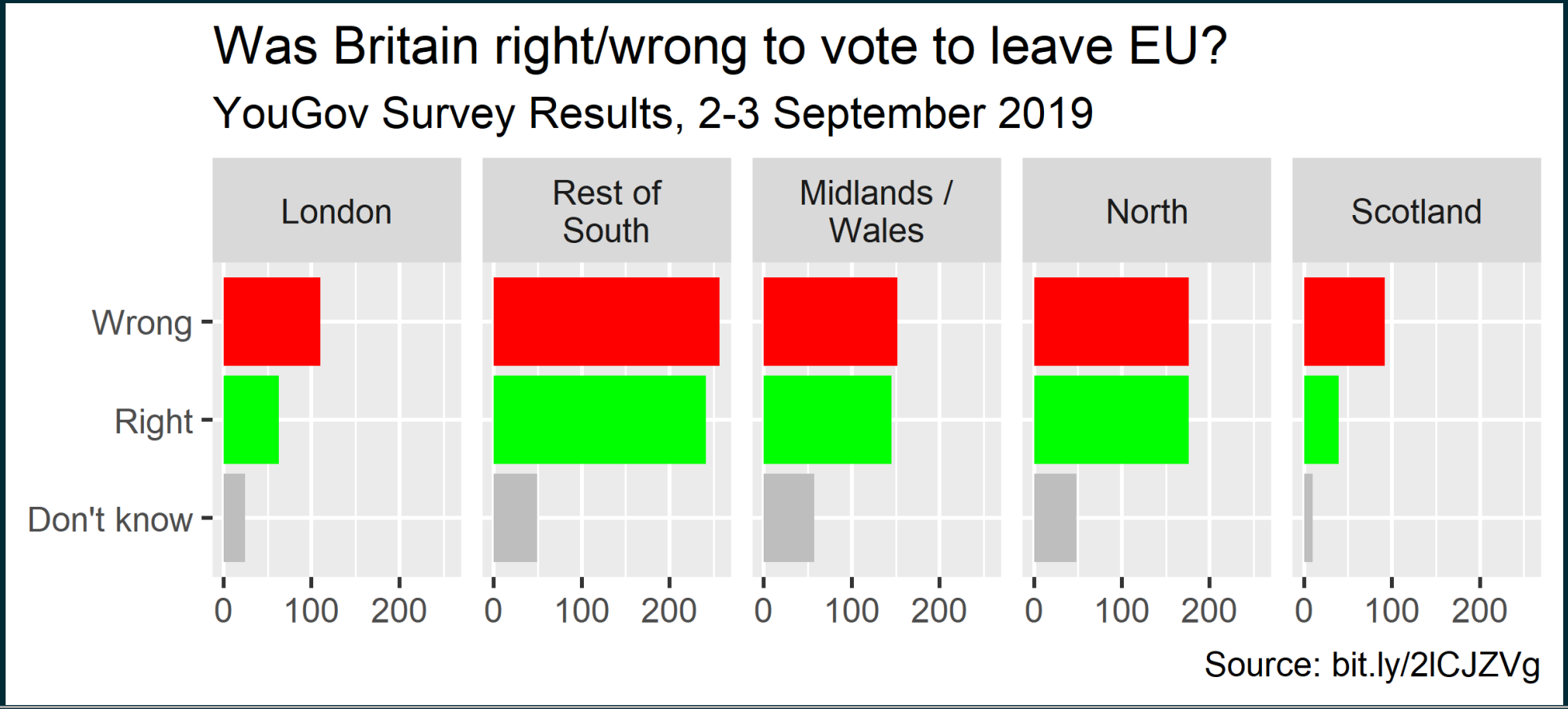
# Select meaningful colors

# Rainbow colors not always the right choice



# Manually choose colors when needed

Plot Code



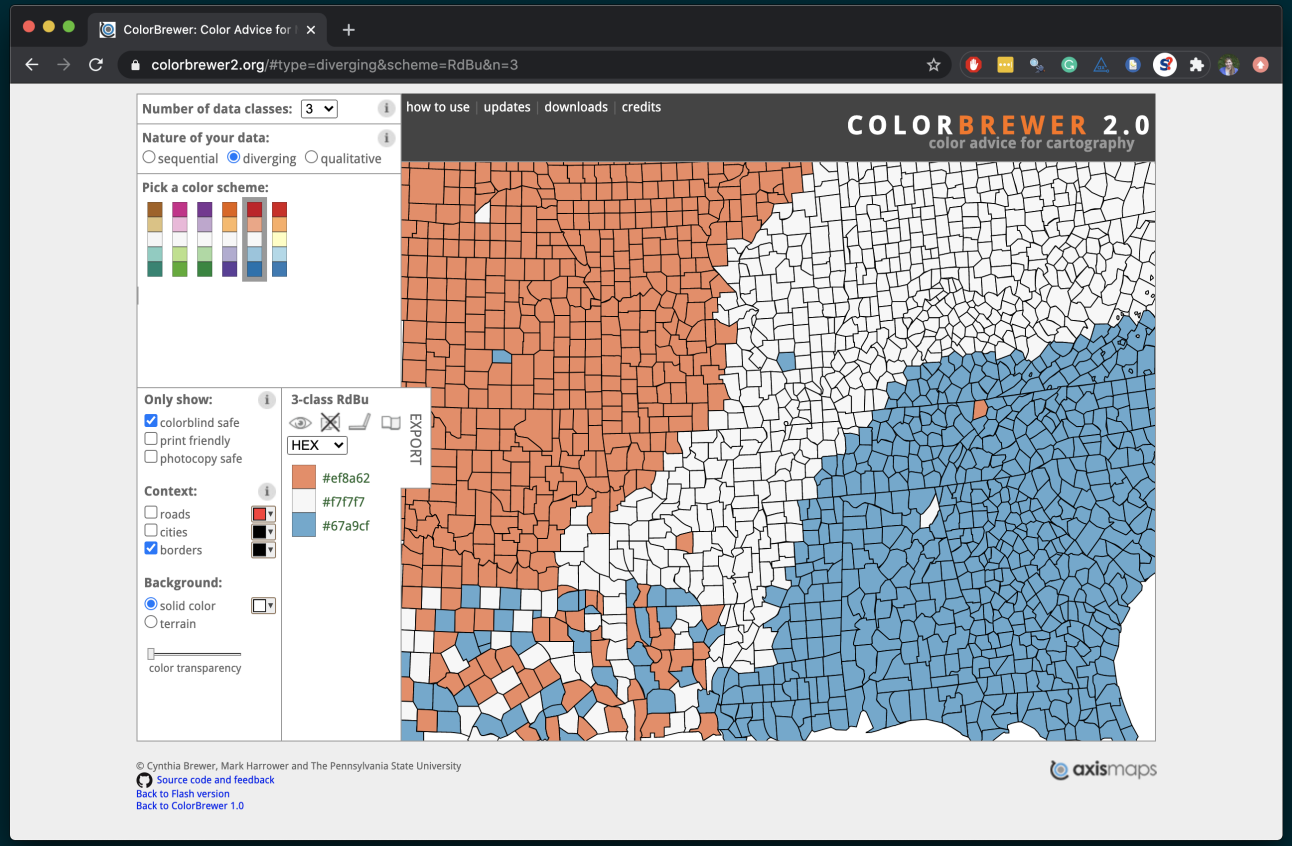
# Manually choose colors when needed

Plot | Code

```
ggplot(brexit, aes(y = opinion, fill = opinion)) +
  geom_bar() +
  facet_wrap(~region, nrow = 1, labeller = label_wrap_gen(width = 12)) +
  guides(fill = "none") +
  labs(title = "Was Britain right/wrong to vote to leave EU?",
       subtitle = "YouGov Survey Results, 2-3 September 2019",
       caption = "Source: bit.ly/2lCJZVg",
       x = NULL, y = NULL) +
  scale_fill_manual(values = c(
    "Wrong" = "red",
    "Right" = "green",
    "Don't know" = "gray"
  ))
```

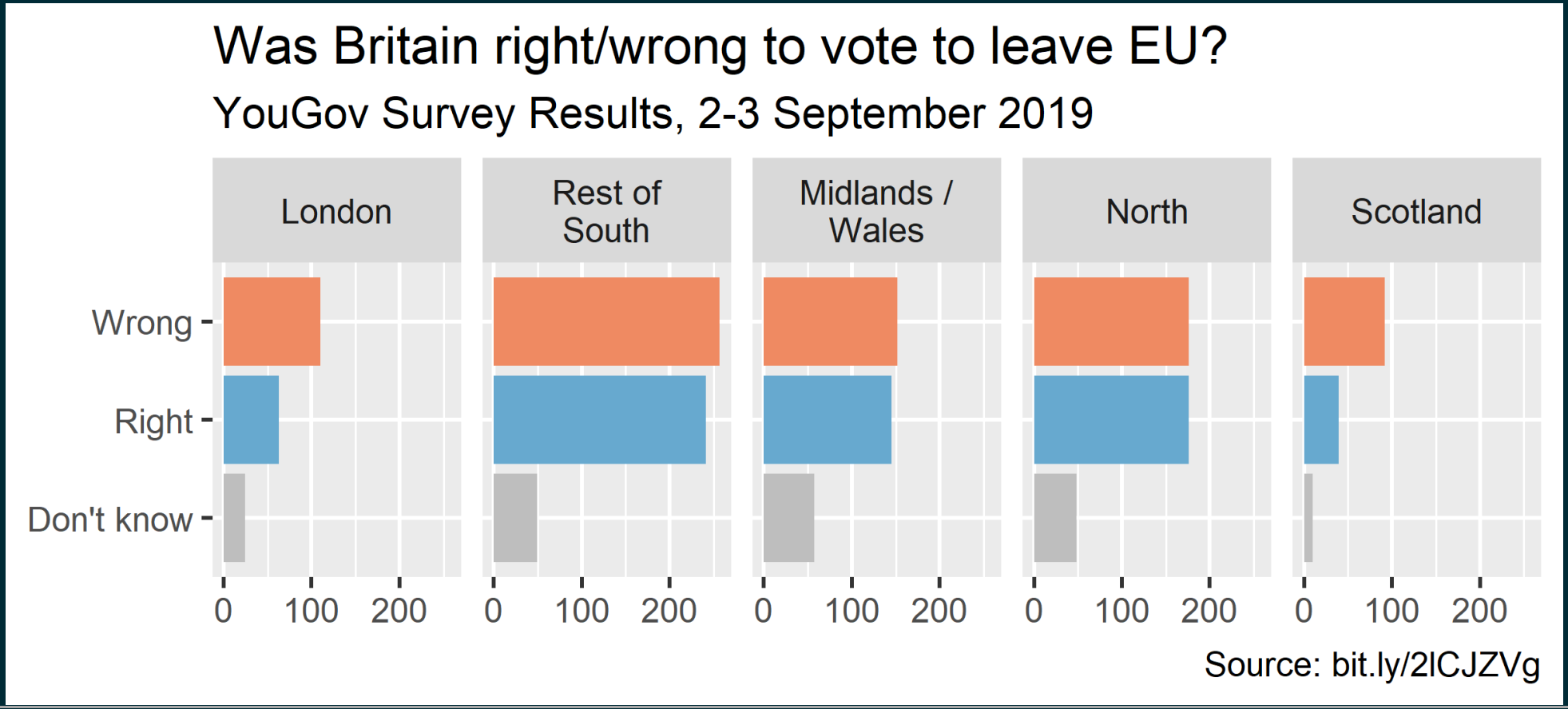
# Choosing better colors

colorbrewer2.org



# Use better colors

Plot Code



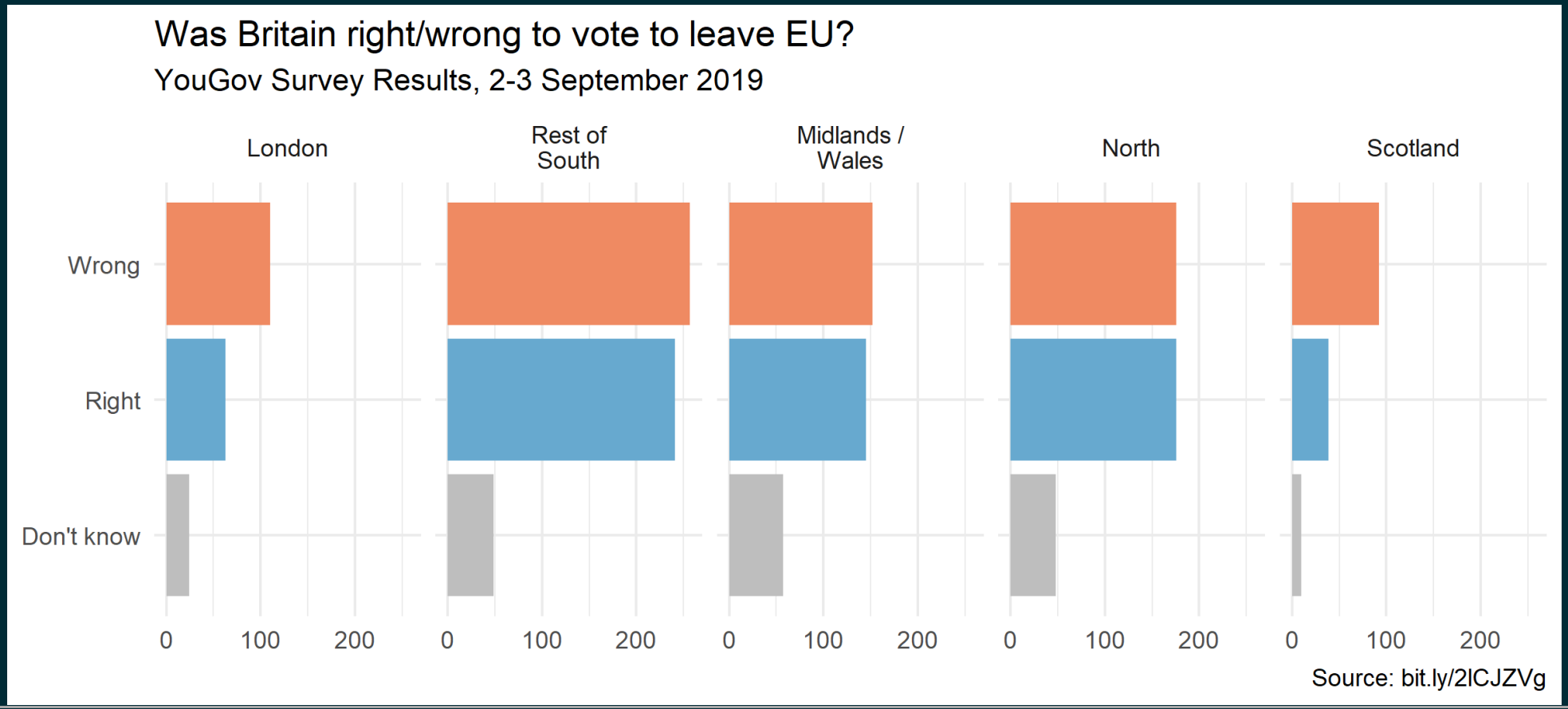
# Use better colors

Plot | Code

```
ggplot(brexit, aes(y = opinion, fill = opinion)) +
  geom_bar() +
  facet_wrap(~region, nrow = 1, labeller = label_wrap_gen(width = 12)) +
  guides(fill = "none") +
  labs(title = "Was Britain right/wrong to vote to leave EU?",
       subtitle = "YouGov Survey Results, 2-3 September 2019",
       caption = "Source: bit.ly/2lCJZVg",
       x = NULL, y = NULL) +
  scale_fill_manual(values = c(
    "Wrong" = "#ef8a62",
    "Right" = "#67a9cf",
    "Don't know" = "gray"
  ))
```

# Select theme

Plot Code



# Select theme

Plot | Code

```
ggplot(brexit, aes(y = opinion, fill = opinion)) +
  geom_bar() +
  facet_wrap(~region, nrow = 1, labeller = label_wrap_gen(width = 12)) +
  guides(fill = "none") +
  labs(title = "Was Britain right/wrong to vote to leave EU?",
       subtitle = "YouGov Survey Results, 2-3 September 2019",
       caption = "Source: bit.ly/2lCJZVg",
       x = NULL, y = NULL) +
  scale_fill_manual(values = c("Wrong" = "#ef8a62",
                               "Right" = "#67a9cf",
                               "Don't know" = "gray")) +
  theme_minimal()
```

# Your turn!

- RStudio Cloud > `AE 07 - Brexit + Telling stories with dataviz` > `brexit.Rmd`.
- Change the visualisation in three different ways to tell slightly different stories with it each time.

# Scientific studies

# Scientific studies

## Observational

- Collect data in a way that does not interfere with how the data arise ("observe")
- Establish associations

## Experimental

- Randomly assign subjects to treatments
- Establish causal connections

What type of study is the following, observational or experiment? What does that mean in terms of causal conclusions?

*Researchers studying the relationship between exercising and energy levels asked participants in their study how many times a week they exercise and whether they have high or low energy when they wake up in the morning.*

*Based on responses to the exercise question the researchers grouped people into three categories (no exercise, exercise 1-3 times a week, and exercise more than 3 times a week).*

*The researchers then compared the proportions of people who said they have high energy in the mornings across the three exercise categories.*

What type of study is the following, observational or experiment? What does that mean in terms of causal conclusions?

*Researchers studying the relationship between exercising and energy levels randomly assigned participants in their study into three groups: no exercise, exercise 1-3 times a week, and exercise more than 3 times a week.*

*After one week, participants were asked whether they have high or low energy when they wake up in the morning.*

*The researchers then compared the proportions of people who said they have high energy in the mornings across the three exercise categories.*

# Case study: Breakfast cereal keeps girls slim

> *Girls who ate breakfast of any type had a lower average body mass index (BMI), a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute with funding from the National Institutes of Health (NIH) and cereal-maker General Mills.* [...]
>
> *The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio, and Maryland who were tracked between the ages of 9 and 19.* [...]
>
> *As part of the survey, the girls were asked once a year what they had eaten during the previous three days.* [...]

Souce: Study: Cereal Keeps Girls Slim, Retrieved Sep 13, 2018.

# Explanatory and response variables

- Explanatory variable: Whether the participant ate breakfast or not
- Reponse variable: BMI of the participant

# Three possible explanations

# Three possible explanations

1. Eating breakfast causes girls to be slimmer

# Three possible explanations

1. Eating breakfast causes girls to be slimmer

2. Being slim causes girls to eat breakfast

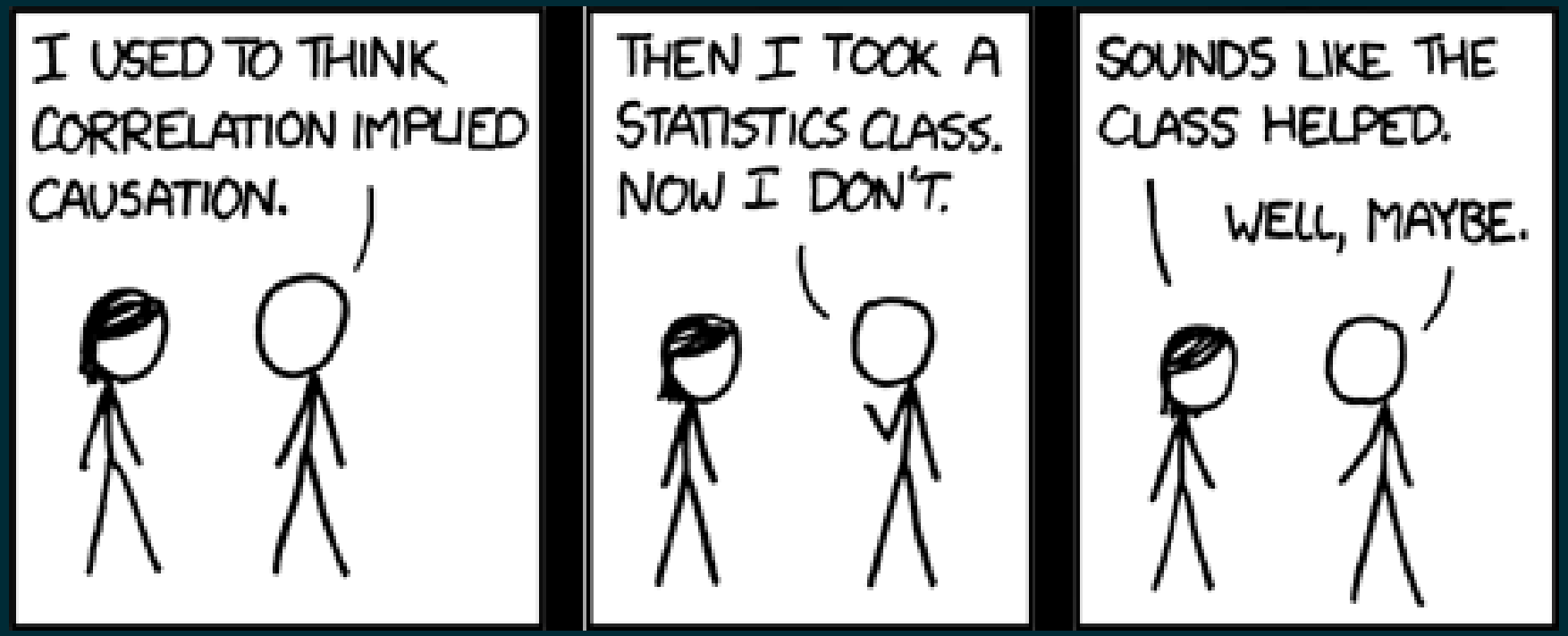
# Three possible explanations

1. Eating breakfast causes girls to be slimmer

2. Being slim causes girls to eat breakfast

3. A third variable is responsible for both -- a **confounding** variable: an extraneous variable that affects both the explanatory and the response variable, and that makes it seem like there is a relationship between them

# Correlation != causation




Randall Munroe CC BY-NC 2.5 http://xkcd.com/552/

# Studies and conclusions

| | Random assignment | No random assignment | |
|---|---|---|---|
| Random sampling | Causal and generalizable | Not causal, but generalizable | Generalizable |
| No random sampling | Causal, but not generalizable | Neither causal nor generalizable | Not generalizable |
| | Causal | Not causal | |

# Case study: Climate change survey

# Survey question

> A July 2019 YouGov survey asked 1633 GB and 1333 USA randomly selected adults which of the following statements about the global environment best describes their view:
>
> - The climate is changing and human activity is mainly responsible
> - The climate is changing and human activity is partly responsible, together with other factors
> - The climate is changing but human activity is not responsible at all
> - The climate is not changing

# Survey data

| | The climate is changing and human activity is mainly responsible | The climate is changing and human activity is partly responsible, together with other factors | The climate is changing but human activity is not responsible at all | The climate is not changing | Don't know | Sum |
|---|---|---|---|---|---|---|
| GB | 833 | 604 | 49 | 33 | 114 | 1633 |
| US | 507 | 493 | 120 | 80 | 133 | 1333 |
| Sum | 1340 | 1097 | 169 | 113 | 247 | 2966 |

Source: YouGov - International Climate Change Survey

What percent of **all respondents** think the climate is changing and human activity is mainly responsible?

| | The climate is changing and human activity is mainly responsible | The climate is changing and human activity is partly responsible, together with other factors | The climate is changing but human activity is not responsible at all | The climate is not changing | Don't know | Sum |
|---|---|---|---|---|---|---|
| GB | 833 | 604 | 49 | 33 | 114 | 1633 |
| US | 507 | 493 | 120 | 80 | 133 | 1333 |
| Sum | 1340 | 1097 | 169 | 113 | 247 | 2966 |

What percent of **all respondents** think the climate is changing and human activity is mainly responsible?

| | The climate is changing and human activity is mainly responsible | The climate is changing and human activity is partly responsible, together with other factors | The climate is changing but human activity is not responsible at all | The climate is not changing | Don't know | Sum |
|---|---|---|---|---|---|---|
| GB | 833 | 604 | 49 | 33 | 114 | 1633 |
| US | 507 | 493 | 120 | 80 | 133 | 1333 |
| Sum | 1340 | 1097 | 169 | 113 | 247 | 2966 |

```
(all <- 1340 / 2966)
```

```
## [1] 0.4517869
```

What percent of **GB respondents** think the climate is changing and human activity is mainly responsible?

| | The climate is changing and human activity is mainly responsible | The climate is changing and human activity is partly responsible, together with other factors | The climate is changing but human activity is not responsible at all | The climate is not changing | Don't know | Sum |
|---|---|---|---|---|---|---|
| GB | 833 | 604 | 49 | 33 | 114 | 1633 |
| US | 507 | 493 | 120 | 80 | 133 | 1333 |
| Sum | 1340 | 1097 | 169 | 113 | 247 | 2966 |

What percent of **GB respondents** think the climate is changing and human activity is mainly responsible?

| | The climate is changing and human activity is mainly responsible | The climate is changing and human activity is partly responsible, together with other factors | The climate is changing but human activity is not responsible at all | The climate is not changing | Don't know | Sum |
|---|---|---|---|---|---|---|
| GB | 833 | 604 | 49 | 33 | 114 | 1633 |
| US | 507 | 493 | 120 | 80 | 133 | 1333 |
| Sum | 1340 | 1097 | 169 | 113 | 247 | 2966 |

```
(gb <- 833 / 1633)
```

```
## [1] 0.5101041
```

What percent of **US respondents** think the climate is changing and human activity is mainly responsible?

| | The climate is changing and human activity is mainly responsible | The climate is changing and human activity is partly responsible, together with other factors | The climate is changing but human activity is not responsible at all | The climate is not changing | Don't know | Sum |
|---|---|---|---|---|---|---|
| GB | 833 | 604 | 49 | 33 | 114 | 1633 |
| US | 507 | 493 | 120 | 80 | 133 | 1333 |
| Sum | 1340 | 1097 | 169 | 113 | 247 | 2966 |

What percent of **US respondents** think the climate is changing and human activity is mainly responsible?

| | The climate is changing and human activity is mainly responsible | The climate is changing and human activity is partly responsible, together with other factors | The climate is changing but human activity is not responsible at all | The climate is not changing | Don't know | Sum |
|---|---|---|---|---|---|---|
| GB | 833 | 604 | 49 | 33 | 114 | 1633 |
| US | 507 | 493 | 120 | 80 | 133 | 1333 |
| Sum | 1340 | 1097 | 169 | 113 | 247 | 2966 |

```
(us <- 507 / 1333)
```

```
## [1] 0.3803451
```

> Based on the percentages we calculated, does there appear to be a relationship between country and beliefs about climate change? If yes, could there be another variable that explains this relationship?

```
all
```

```
## [1] 0.4517869
```

```
gb
```

```
## [1] 0.5101041
```

```
us
```

```
## [1] 0.3803451
```

# Conditional probability

**Notation**: $P(A|B)$: Probability of event A given event B

- What is the probability that it will be unseasonably warm tomorrow?
- What is the probability that it will be unseasonably warm tomorrow, given that it was unseasonably warm today?

# Independence

- If knowing event A happened tells you something about event B happening, or vice versa, then events A and B are not independent
- If not, they are said to be independent
- $P(A|B) = P(A)$

# Case study: Berkeley admission data

# Berkeley admission data

- Study carried out by the Graduate Division of the University of California, Berkeley in the early 70's to evaluate whether there was a gender bias in graduate admissions.
- The data come from six departments. For confidentiality we'll call them A-F.
- We have information on whether the applicant was male or female and whether they were admitted or rejected.
- First, we will evaluate whether the percentage of males admitted is indeed higher than females, overall. Next, we will calculate the same percentage for each department.

# Data

```
## # A tibble: 4,526 x 3
##    admit    gender dept
##    <fct>    <fct>  <ord>
##  1 Admitted Male   A
##  2 Admitted Male   A
##  3 Admitted Male   A
##  4 Admitted Male   A
##  5 Admitted Male   A
##  6 Admitted Male   A
##  7 Admitted Male   A
##  8 Admitted Male   A
##  9 Admitted Male   A
## 10 Admitted Male   A
## 11 Admitted Male   A
## 12 Admitted Male   A
## 13 Admitted Male   A
## 14 Admitted Male   A
## 15 Admitted Male   A
## # ... with 4,511 more rows
```

```
## # A tibble: 2 x 2
##   gender     n
##   <fct>  <int>
## 1 Female  1835
## 2 Male    2691
```

```
## # A tibble: 6 x 2
##   dept      n
##   <ord> <int>
## 1 A       933
## 2 B       585
## 3 C       918
## 4 D       792
## 5 E       584
## 6 F       714
```

```
## # A tibble: 2 x 2
##   admit        n
##   <fct>    <int>
## 1 Rejected  2771
## 2 Admitted  1755
```

What can you say about the overall gender distribution? Hint: Calculate the following probabilities: $P(Admit|Male)$ and $P(Admit|Female)$.

```
ucbadmit %>%
  count(gender, admit)
```

```
## # A tibble: 4 x 3
##   gender admit        n
##   <fct>  <fct>    <int>
## 1 Female Rejected  1278
## 2 Female Admitted   557
## 3 Male   Rejected  1493
## 4 Male   Admitted  1198
```

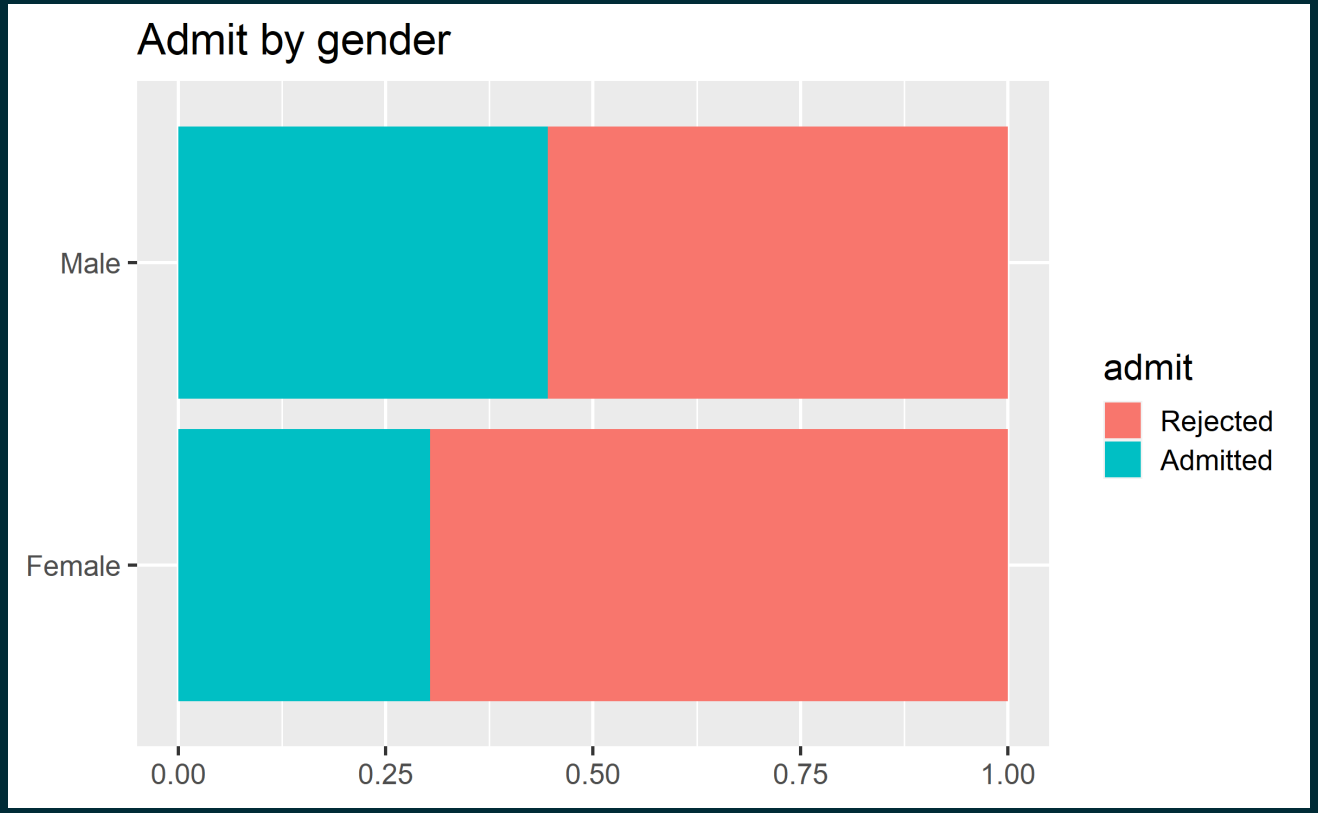
```
ucbadmit %>%
  count(gender, admit) %>%
  group_by(gender) %>%
  mutate(prop_admit = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   gender [2]
##   gender admit        n prop_admit
##   <fct>  <fct>    <int>      <dbl>
## 1 Female Rejected  1278      0.696
## 2 Female Admitted   557      0.304
## 3 Male   Rejected  1493      0.555
## 4 Male   Admitted  1198      0.445
```

- $P(Admit|Female) = 0.304$
- $P(Admit|Male) = 0.445$

# Overall gender distribution

Plot Code



# Overall gender distribution

Plot | Code

```
ggplot(ucbadmit, aes(y = gender, fill = admit)) +
  geom_bar(position = "fill") +
  labs(title = "Admit by gender",
       y = NULL, x = NULL)
```

What can you say about the gender distribution by department ?

```
ucbadmit %>%
  count(dept, gender, admit)
```

```
## # A tibble: 24 x 4
##   dept  gender admit        n
##   <ord> <fct>  <fct>    <int>
## 1 A     Female Rejected    19
## 2 A     Female Admitted    89
## 3 A     Male   Rejected   313
## 4 A     Male   Admitted   512
## 5 B     Female Rejected     8
## 6 B     Female Admitted    17
## # ... with 18 more rows
```

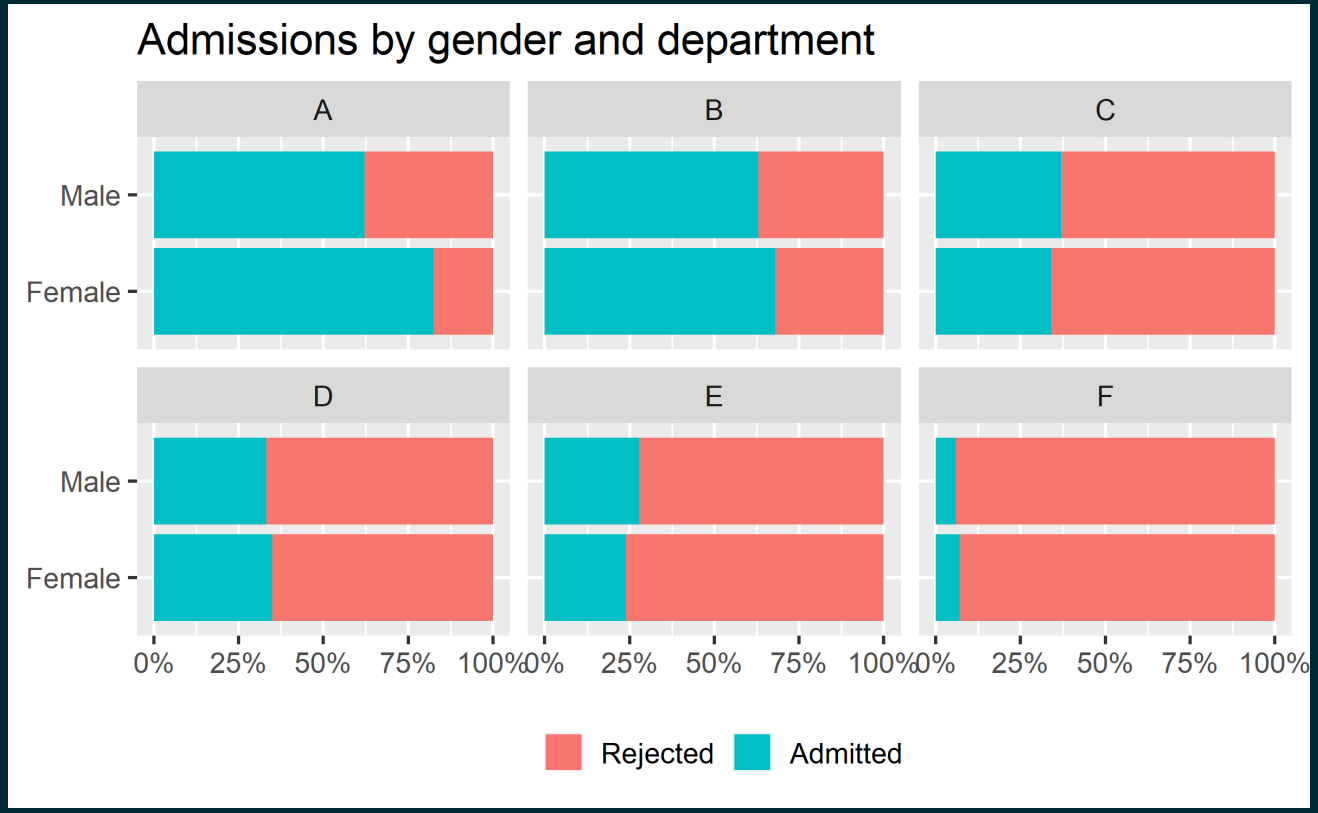
Let's try again... What can you say about the gender distribution by department?

```
ucbadmit %>%
  count(dept, gender, admit) %>%
  pivot_wider(names_from = dept, values_from = n)
```

```
## # A tibble: 4 x 8
##   gender admit        A     B     C     D     E     F
##   <fct>  <fct>    <int> <int> <int> <int> <int> <int>
## 1 Female Rejected    19     8   391   244   299   317
## 2 Female Admitted    89    17   202   131    94    24
## 3 Male   Rejected   313   207   205   279   138   351
## 4 Male   Admitted   512   353   120   138    53    22
```

# Gender distribution, by department

Plot Code

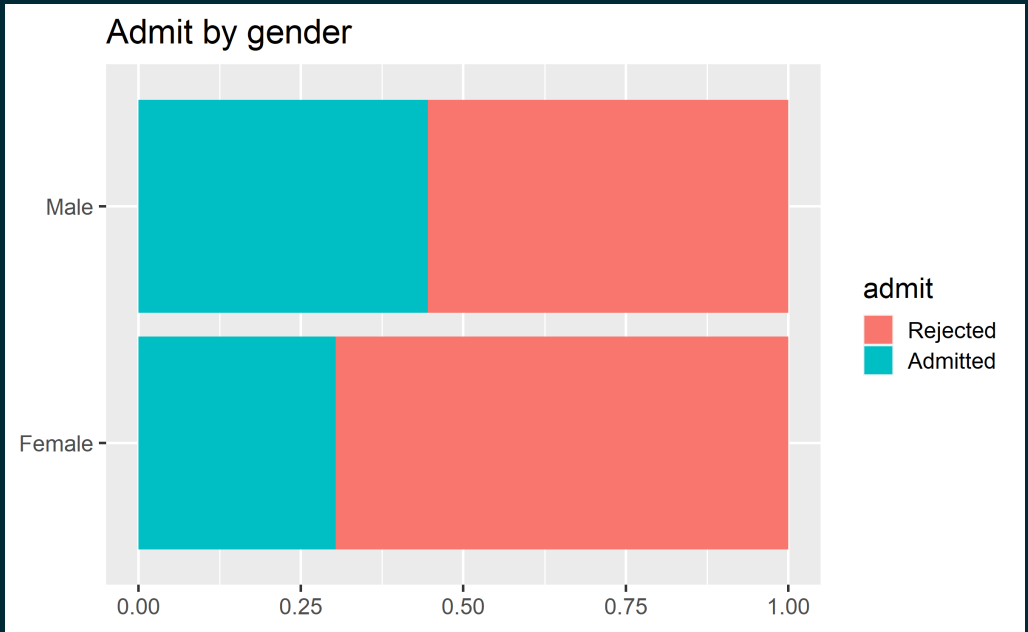


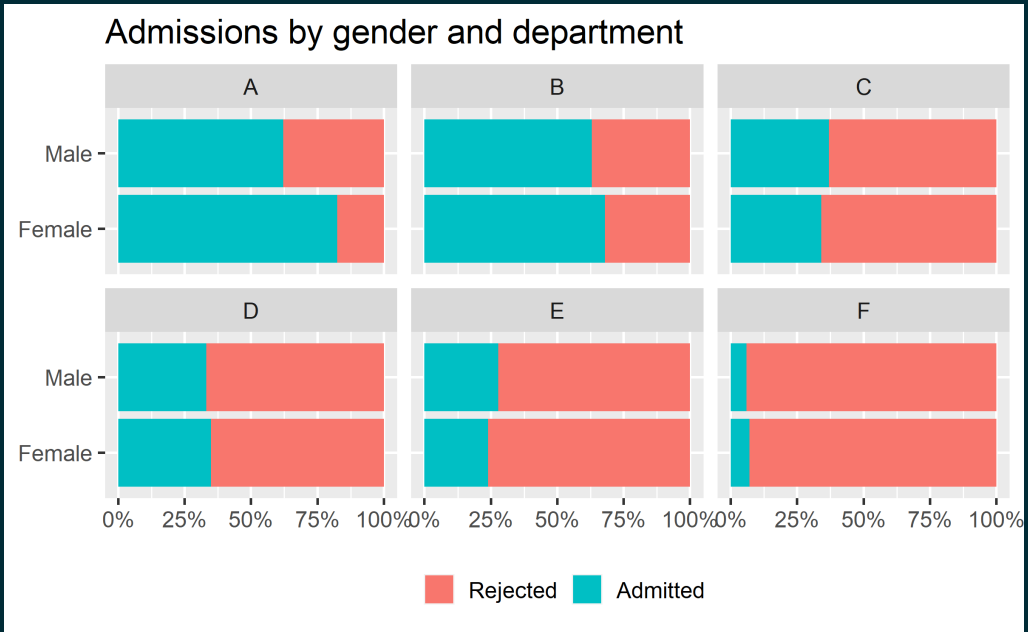
# Gender distribution, by department

Plot | Code

```
ggplot(ucbadmit, aes(y = gender, fill = admit)) +
  geom_bar(position = "fill") +
  facet_wrap(. ~ dept) +
  scale_x_continuous(labels = label_percent()) +
  labs(title = "Admissions by gender and department",
       x = NULL, y = NULL, fill = NULL) +
  theme(legend.position = "bottom")
```

# Case for gender discrimination?



# Closer look at departments

Output | Code

```
## # A tibble: 12 x 5
## # Groups:   dept, gender [12]
##    dept  gender n_admitted n_applied prop_admit
##    <ord> <fct>       <int>     <int>      <dbl>
##  1 A     Female         89       108     0.824
##  2 A     Male          512       825     0.621
##  3 B     Female         17        25     0.68
##  4 B     Male          353       560     0.630
##  5 C     Female        202       593     0.341
##  6 C     Male          120       325     0.369
##  7 D     Female        131       375     0.349
##  8 D     Male          138       417     0.331
##  9 E     Female         94       393     0.239
## 10 E     Male           53       191     0.277
## 11 F     Female         24       341     0.0704
## 12 F     Male           22       373     0.0590
```

# Closer look at departments

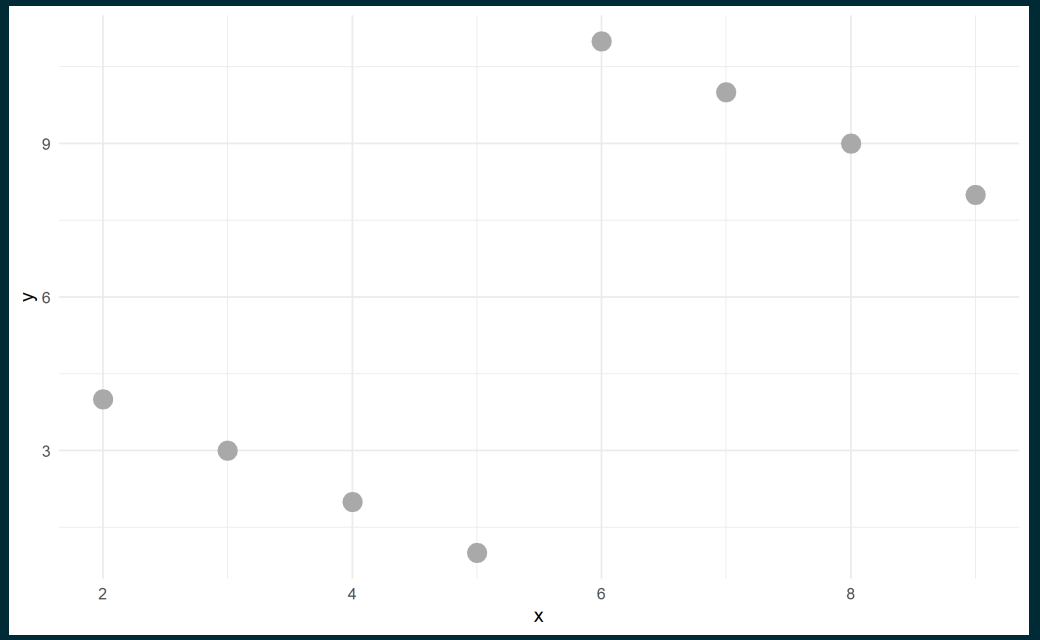
Output | Code

```
ucbadmit %>%
  count(dept, gender, admit) %>%
  group_by(dept, gender) %>%
  mutate(
    n_applied  = sum(n),
    prop_admit = n / n_applied
    ) %>%
  filter(admit == "Admitted") %>%
  rename(n_admitted = n) %>%
  select(-admit) %>%
  print(n = 12)
```

# Simpson's paradox

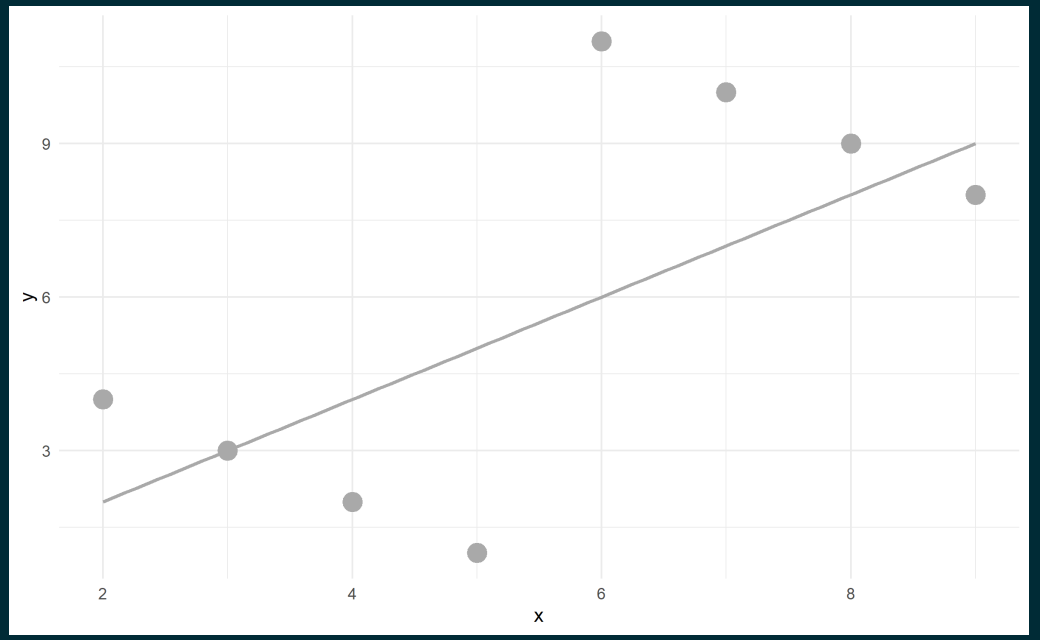
# Relationship between two variables

```
## # A tibble: 8 x 3
##       x     y z
##   <dbl> <dbl> <chr>
## 1     2     4 A
## 2     3     3 A
## 3     4     2 A
## 4     5     1 A
## 5     6    11 B
## 6     7    10 B
## # ... with 2 more rows
```



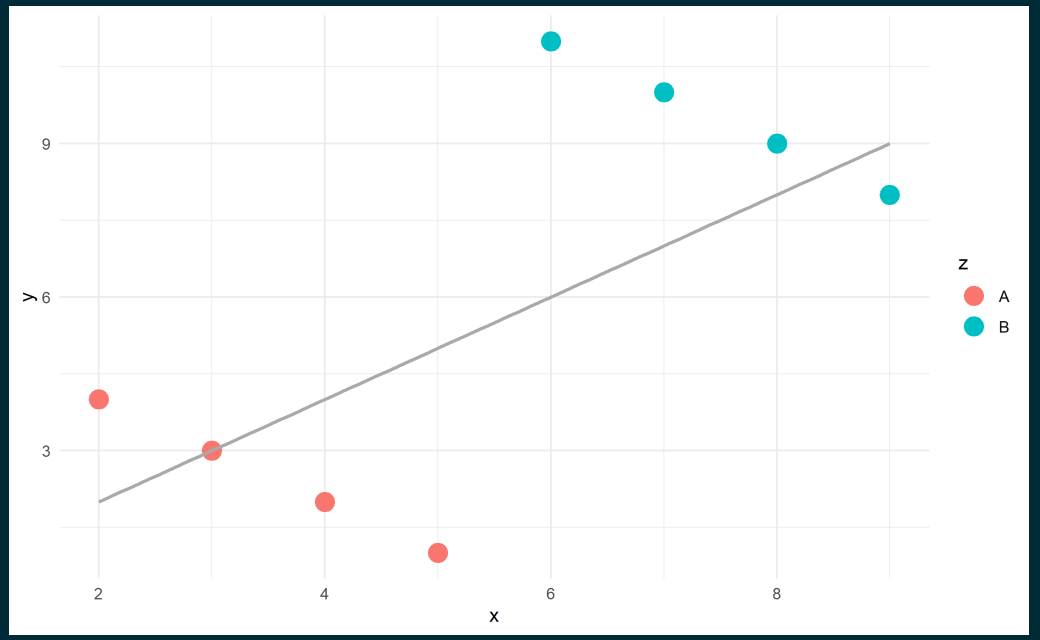
# Relationship between two variables

```
## # A tibble: 8 x 3
##       x     y z
##   <dbl> <dbl> <chr>
## 1     2     4 A
## 2     3     3 A
## 3     4     2 A
## 4     5     1 A
## 5     6    11 B
## 6     7    10 B
## # ... with 2 more rows
```



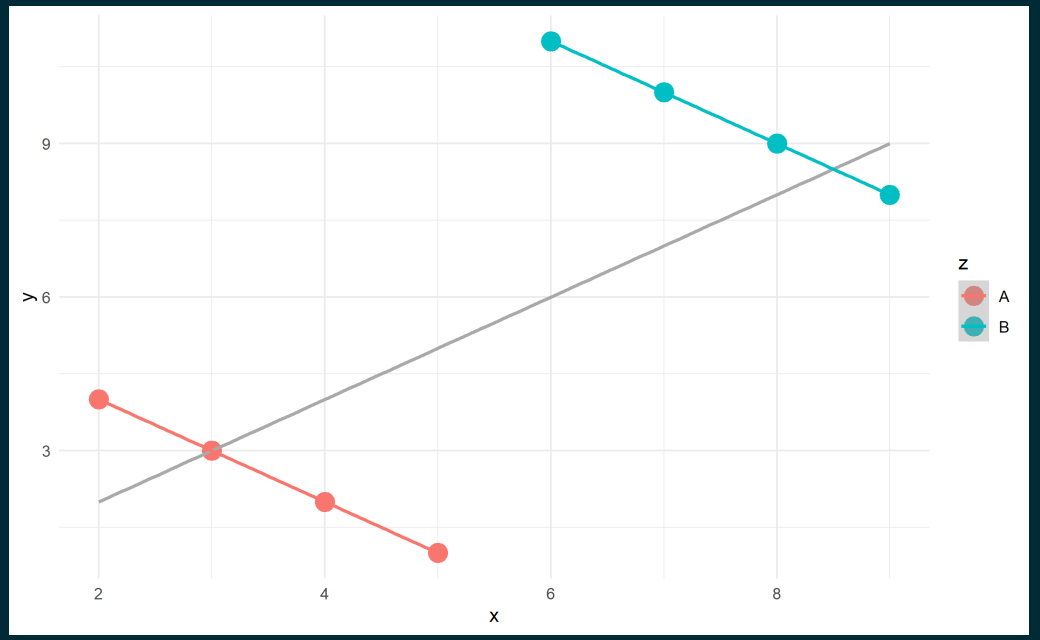
# Considering a third variable

```
## # A tibble: 8 x 3
##       x     y z
##   <dbl> <dbl> <chr>
## 1     2     4 A
## 2     3     3 A
## 3     4     2 A
## 4     5     1 A
## 5     6    11 B
## 6     7    10 B
## # ... with 2 more rows
```



# Relationship between three variables

```
## # A tibble: 8 x 3
##       x     y z
##   <dbl> <dbl> <chr>
## 1     2     4 A
## 2     3     3 A
## 3     4     2 A
## 4     5     1 A
## 5     6    11 B
## 6     7    10 B
## # ... with 2 more rows
```



# Simpson's paradox

- Not considering an important variable when studying a relationship can result in **Simpson's paradox**
- Simpson's paradox illustrates the effect that omission of an explanatory variable can have on the measure of association between another explanatory variable and a response variable
- The inclusion of a third variable in the analysis can change the apparent relationship between the other two variables

# Aside: `group_by()` and `count()`

# What does group_by() do?

`group_by()` takes an existing data frame and converts it into a grouped data frame where subsequent operations are performed "once per group"

```
ucbadmit
```

```
## # A tibble: 4,526 x 3
##   admit    gender dept
##   <fct>    <fct>  <ord>
## 1 Admitted Male   A
## 2 Admitted Male   A
## 3 Admitted Male   A
## 4 Admitted Male   A
## 5 Admitted Male   A
## 6 Admitted Male   A
## # ... with 4,520 more rows
```

```
ucbadmit %>%
  group_by(gender)
```

```
## # A tibble: 4,526 x 3
## # Groups:   gender [2]
##   admit    gender dept
##   <fct>    <fct>  <ord>
## 1 Admitted Male   A
## 2 Admitted Male   A
## 3 Admitted Male   A
## 4 Admitted Male   A
## 5 Admitted Male   A
## 6 Admitted Male   A
## # ... with 4,520 more rows
```

# What does group_by() not do?

`group_by()` does not sort the data, `arrange()` does

```
ucbadmit %>%
  group_by(gender)
```

```
## # A tibble: 4,526 x 3
## # Groups:   gender [2]
##   admit    gender dept
##   <fct>    <fct>  <ord>
## 1 Admitted Male   A
## 2 Admitted Male   A
## 3 Admitted Male   A
## 4 Admitted Male   A
## 5 Admitted Male   A
## 6 Admitted Male   A
## # ... with 4,520 more rows
```

```
ucbadmit %>%
  arrange(gender)
```

```
## # A tibble: 4,526 x 3
##   admit    gender dept
##   <fct>    <fct>  <ord>
## 1 Admitted Female A
## 2 Admitted Female A
## 3 Admitted Female A
## 4 Admitted Female A
## 5 Admitted Female A
## 6 Admitted Female A
## # ... with 4,520 more rows
```

# What does group_by() not do?

`group_by()` does not create frequency tables, `count()` does

```
ucbadmit %>%
  group_by(gender)
```

```
## # A tibble: 4,526 x 3
## # Groups:   gender [2]
##   admit    gender dept
##   <fct>    <fct>  <ord>
## 1 Admitted Male   A
## 2 Admitted Male   A
## 3 Admitted Male   A
## 4 Admitted Male   A
## 5 Admitted Male   A
## 6 Admitted Male   A
## # ... with 4,520 more rows
```

```
ucbadmit %>%
  count(gender)
```

```
## # A tibble: 2 x 2
##   gender     n
##   <fct>  <int>
## 1 Female  1835
## 2 Male    2691
```

# Undo grouping with ungroup()

```
ucbadmit %>%
  count(gender, admit) %>%
  group_by(gender) %>%
  mutate(prop_admit = n / sum(n)) %>%
  select(gender, prop_admit)
```

```
## # A tibble: 4 x 2
## # Groups:   gender [2]
##   gender prop_admit
##   <fct>       <dbl>
## 1 Female      0.696
## 2 Female      0.304
## 3 Male        0.555
## 4 Male        0.445
```

```
ucbadmit %>%
  count(gender, admit) %>%
  group_by(gender) %>%
  mutate(prop_admit = n / sum(n)) %>%
  select(gender, prop_admit) %>%
  ungroup()
```

```
## # A tibble: 4 x 2
##   gender prop_admit
##   <fct>       <dbl>
## 1 Female      0.696
## 2 Female      0.304
## 3 Male        0.555
## 4 Male        0.445
```

# count() is a short-hand

`count()` is a short-hand for `group_by()` and then `summarise()` to count the number of observations in each group

```
ucbadmit %>%
  group_by(gender) %>%
  summarise(n = n())
```

```
## # A tibble: 2 x 2
##   gender     n
##   <fct>  <int>
## 1 Female  1835
## 2 Male    2691
```

```
ucbadmit %>%
  count(gender)
```

```
## # A tibble: 2 x 2
##   gender     n
##   <fct>  <int>
## 1 Female  1835
## 2 Male    2691
```

# count can take multiple arguments

```
ucbadmit %>%
  group_by(gender, admit) %>%
  summarise(n = n())
```

```
## # A tibble: 4 x 3
## # Groups:   gender [2]
##   gender admit        n
##   <fct>  <fct>    <int>
## 1 Female Rejected  1278
## 2 Female Admitted   557
## 3 Male   Rejected  1493
## 4 Male   Admitted  1198
```

```
ucbadmit %>%
  count(gender, admit)
```

```
## # A tibble: 4 x 3
##   gender admit        n
##   <fct>  <fct>    <int>
## 1 Female Rejected  1278
## 2 Female Admitted   557
## 3 Male   Rejected  1493
## 4 Male   Admitted  1198
```

# `summarise()` after `group_by()`

- `count()` ungroups after itself
- `summarise()` peels off one layer of grouping by default, or you can specify a different behaviour

```
ucbadmit %>%
  group_by(gender, admit) %>%
  summarise(n = n())
```

```
## # A tibble: 4 x 3
## # Groups:   gender [2]
##   gender admit        n
##   <fct>  <fct>    <int>
## 1 Female Rejected  1278
## 2 Female Admitted   557
## 3 Male   Rejected  1493
## 4 Male   Admitted  1198
```

# What's in a data analysis?

# Five core activities of data analysis

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

Roger D. Peng and Elizabeth Matsui. "The Art of Data Science." A Guide for Anyone Who Works with Data. Skybrude Consulting, LLC (2015).

# Stating and refining the question

# Six types of questions

1. **Descriptive:** summarize a characteristic of a set of data
2. **Exploratory:** analyze to see if there are patterns, trends, or relationships between variables (hypothesis generating)
3. **Inferential:** analyze patterns, trends, or relationships in representative data from a population
4. **Predictive:** make predictions for individuals or groups of individuals
5. **Causal:** whether changing one factor will change another factor, on average, in a population
6. **Mechanistic:** explore "how" as opposed to whether

Jeffery T. Leek and Roger D. Peng. "What is the question?." Science 347.6228 (2015): 1314-1315.

# Ex: COVID-19 and Vitamin D

1. **Descriptive:** frequency of hospitalisations due to COVID-19 in a set of data collected from a group of individuals

# Ex: COVID-19 and Vitamin D

1. **Descriptive:** frequency of hospitalisations due to COVID-19 in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and COVID-19 hospitalisations

# Ex: COVID-19 and Vitamin D

1. **Descriptive:** frequency of hospitalisations due to COVID-19 in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and COVID-19 hospitalisations
3. **Inferential:** examine whether any relationship between taking Vitamin D supplements and COVID-19 hospitalisations found in the sample hold for the population at large

# Ex: COVID-19 and Vitamin D

1. **Descriptive:** frequency of hospitalisations due to COVID-19 in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and COVID-19 hospitalisations
3. **Inferential:** examine whether any relationship between taking Vitamin D supplements and COVID-19 hospitalisations found in the sample hold for the population at large
4. **Predictive:** what types of people will take Vitamin D supplements during the next year

# Ex: COVID-19 and Vitamin D

1. **Descriptive:** frequency of hospitalisations due to COVID-19 in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and COVID-19 hospitalisations
3. **Inferential:** examine whether any relationship between taking Vitamin D supplements and COVID-19 hospitalisations found in the sample hold for the population at large
4. **Predictive:** what types of people will take Vitamin D supplements during the next year
5. **Causal:** whether people with COVID-19 who were randomly assigned to take Vitamin D supplements or those who were not are hospitalised

# Ex: COVID-19 and Vitamin D

1. **Descriptive:** frequency of hospitalisations due to COVID-19 in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and COVID-19 hospitalisations
3. **Inferential:** examine whether any relationship between taking Vitamin D supplements and COVID-19 hospitalisations found in the sample hold for the population at large
4. **Predictive:** what types of people will take Vitamin D supplements during the next year
5. **Causal:** whether people with COVID-19 who were randomly assigned to take Vitamin D supplements or those who were not are hospitalised
6. **Mechanistic:** how increased vitamin D intake leads to a reduction in the number of viral illnesses

# Questions to data science problems

- Do you have appropriate data to answer your question?
- Do you have information on confounding variables?
- Was the data you're working with collected in a way that introduces bias?

# Questions to data science problems

- Do you have appropriate data to answer your question?
- Do you have information on confounding variables?
- Was the data you're working with collected in a way that introduces bias?

> Suppose I want to estimate the average number of children in households in Edinburgh. I conduct a survey at an elementary school in Edinburgh and ask students at this elementary school how many children, including themselves, live in their house. Then, I take the average of the responses. Is this a biased or an unbiased estimate of the number of children in households in Edinburgh? If biased, will the value be an overestimate or underestimate?

# Exploratory data analysis

# Checklist

- Formulate your question
- Read in your data
- Check the dimensions
- Look at the top and the bottom of your data
- Validate with at least one external data source
- Make a plot
- Try the easy solution first

# Formulate your question

- Consider scope:
  - Are air pollution levels higher on the east coast than on the west coast?
  - Are hourly ozone levels on average higher in New York City than they are in Los Angeles?
  - Do counties in the eastern United States have higher ozone levels than counties in the western United States?
- Most importantly: "Do I have the right data to answer this question?"

# Read in your data

- Place your data in a folder called `data`
- Read it into R with `read_csv()` or friends (`read_delim()`, `read_excel()`, etc.)

```
library(readxl)
fav_food <- read_excel("data/favourite-food.xlsx")
fav_food
```

```
## # A tibble: 5 x 6
##   `Student ID` `Full Name`  favourite.food  mealPlan  AGE   SES
##          <dbl> <chr>        <chr>           <chr>     <chr> <chr>
## 1            1 Sunil Huffm~ Strawberry yog~ Lunch on~ 4     High
## 2            2 Barclay Lynn French fries    Lunch on~ 5     Midd~
## 3            3 Jayendra Ly~ N/A             Breakfas~ 7     Low
## 4            4 Leon Rossini Anchovies       Lunch on~ 99999 Midd~
## 5            5 Chidiegwu D~ Pizza           Breakfas~ five  High
```

# clean_names()

If the variable names are malformatted, use `janitor::clean_names()`

```
library(janitor)
fav_food %>% clean_names()
```

```
## # A tibble: 5 x 6
##   student_id full_name   favourite_food   meal_plan   age   ses
##        <dbl> <chr>       <chr>            <chr>       <chr> <chr>
## 1          1 Sunil Huff~ Strawberry yogh~ Lunch only  4     High
## 2          2 Barclay Ly~ French fries     Lunch only  5     Midd~
## 3          3 Jayendra L~ N/A              Breakfast ~ 7     Low
## 4          4 Leon Rossi~ Anchovies        Lunch only  99999 Midd~
## 5          5 Chidiegwu ~ Pizza            Breakfast ~ five  High
```

# Case study: NYC Squirrels!

- The Squirrel Census is a multimedia science, design, and storytelling project focusing on the Eastern gray (*Sciurus carolinensis*). They count squirrels and present their findings to the public.
- This table contains squirrel data for each of the 3,023 sightings, including location coordinates, age, primary and secondary fur color, elevation, activities, communications, and interactions between squirrels and with humans.

```
#install_github("mine-cetinkaya-rundel/nycsquirrels18")
library(nycsquirrels18)
```

# Locate the codebook

mine-cetinkaya-rundel.github.io/nycsquirrels18/reference/squirrels.html

# Locate the codebook

mine-cetinkaya-rundel.github.io/nycsquirrels18/reference/squirrels.html

# Check the dimensions

```
dim(squirrels)
```

```
## [1] 3023   35
```

# Look at the top...

```
squirrels %>% head()
```

```
## # A tibble: 6 x 35
##    long   lat unique_squirrel_id hectare shift date
##   <dbl> <dbl> <chr>              <chr>   <chr> <date>
## 1 -74.0  40.8 13A-PM-1014-04     13A     PM    2018-10-14
## 2 -74.0  40.8 15F-PM-1010-06     15F     PM    2018-10-10
## 3 -74.0  40.8 19C-PM-1018-02     19C     PM    2018-10-18
## 4 -74.0  40.8 21B-AM-1019-04     21B     AM    2018-10-19
## 5 -74.0  40.8 23A-AM-1018-02     23A     AM    2018-10-18
## 6 -74.0  40.8 38H-PM-1012-01     38H     PM    2018-10-12
## # ... with 29 more variables: hectare_squirrel_number <dbl>,
## #   age <chr>, primary_fur_color <chr>,
## #   highlight_fur_color <chr>,
## #   combination_of_primary_and_highlight_color <chr>,
## #   color_notes <chr>, location <chr>,
## #   above_ground_sighter_measurement <chr>,
## #   specific_location <chr>, running <lgl>, chasing <lgl>, ...
```

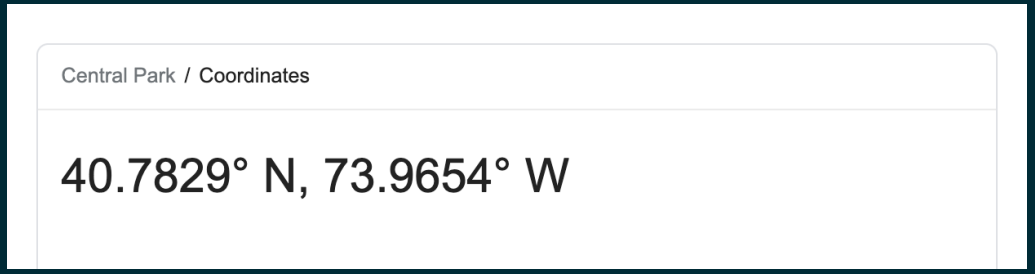
# ...and the bottom

```
squirrels %>% tail()
```

```
## # A tibble: 6 x 35
##    long   lat unique_squirrel_id hectare shift date
##   <dbl> <dbl> <chr>              <chr>   <chr> <date>
## 1 -74.0  40.8 6D-PM-1020-01      06D     PM    2018-10-20
## 2 -74.0  40.8 21H-PM-1018-01     21H     PM    2018-10-18
## 3 -74.0  40.8 31D-PM-1006-02     31D     PM    2018-10-06
## 4 -74.0  40.8 37B-AM-1018-04     37B     AM    2018-10-18
## 5 -74.0  40.8 21C-PM-1006-01     21C     PM    2018-10-06
## 6 -74.0  40.8 7G-PM-1018-04      07G     PM    2018-10-18
## # ... with 29 more variables: hectare_squirrel_number <dbl>,
## #   age <chr>, primary_fur_color <chr>,
## #   highlight_fur_color <chr>,
## #   combination_of_primary_and_highlight_color <chr>,
## #   color_notes <chr>, location <chr>,
## #   above_ground_sighter_measurement <chr>,
## #   specific_location <chr>, running <lgl>, chasing <lgl>, ...
```

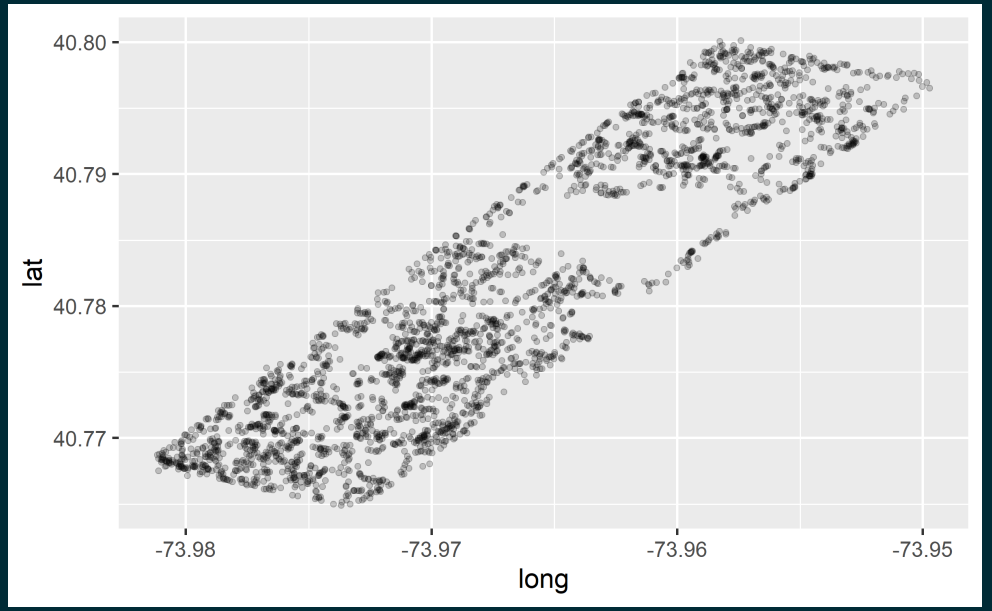
# Validate with at least one external data source

```
## # A tibble: 3,023 x 2
##     long   lat
##    <dbl> <dbl>
##  1 -74.0  40.8
##  2 -74.0  40.8
##  3 -74.0  40.8
##  4 -74.0  40.8
##  5 -74.0  40.8
##  6 -74.0  40.8
##  7 -74.0  40.8
##  8 -74.0  40.8
##  9 -74.0  40.8
## 10 -74.0  40.8
## 11 -74.0  40.8
## 12 -74.0  40.8
## 13 -74.0  40.8
## 14 -74.0  40.8
## 15 -74.0  40.8
## # ... with 3,008 more rows
```



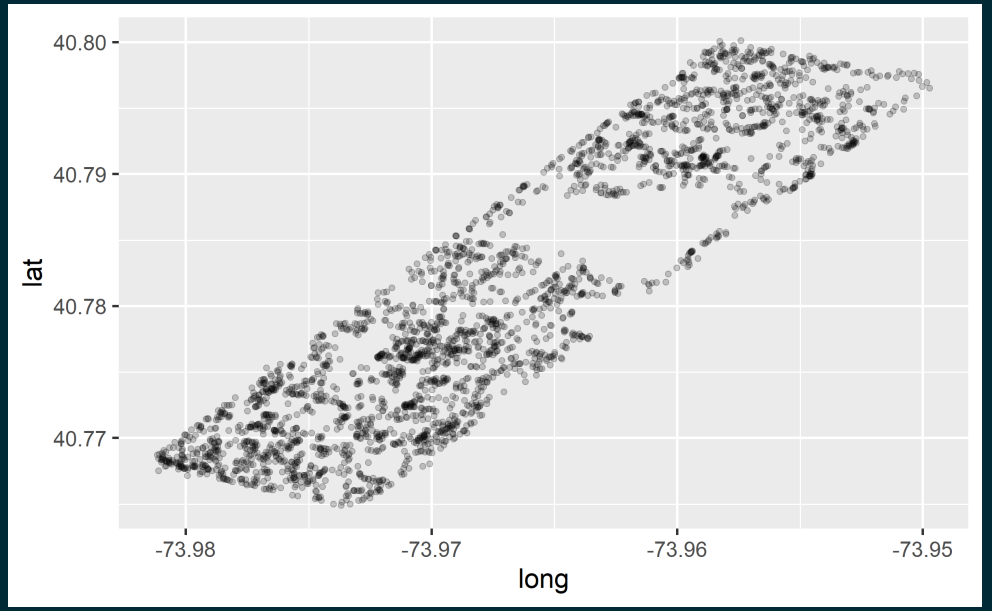
# Make a plot

```
ggplot(squirrels, aes(x = long, y = lat)) +
  geom_point(alpha = 0.2)
```



# Make a plot

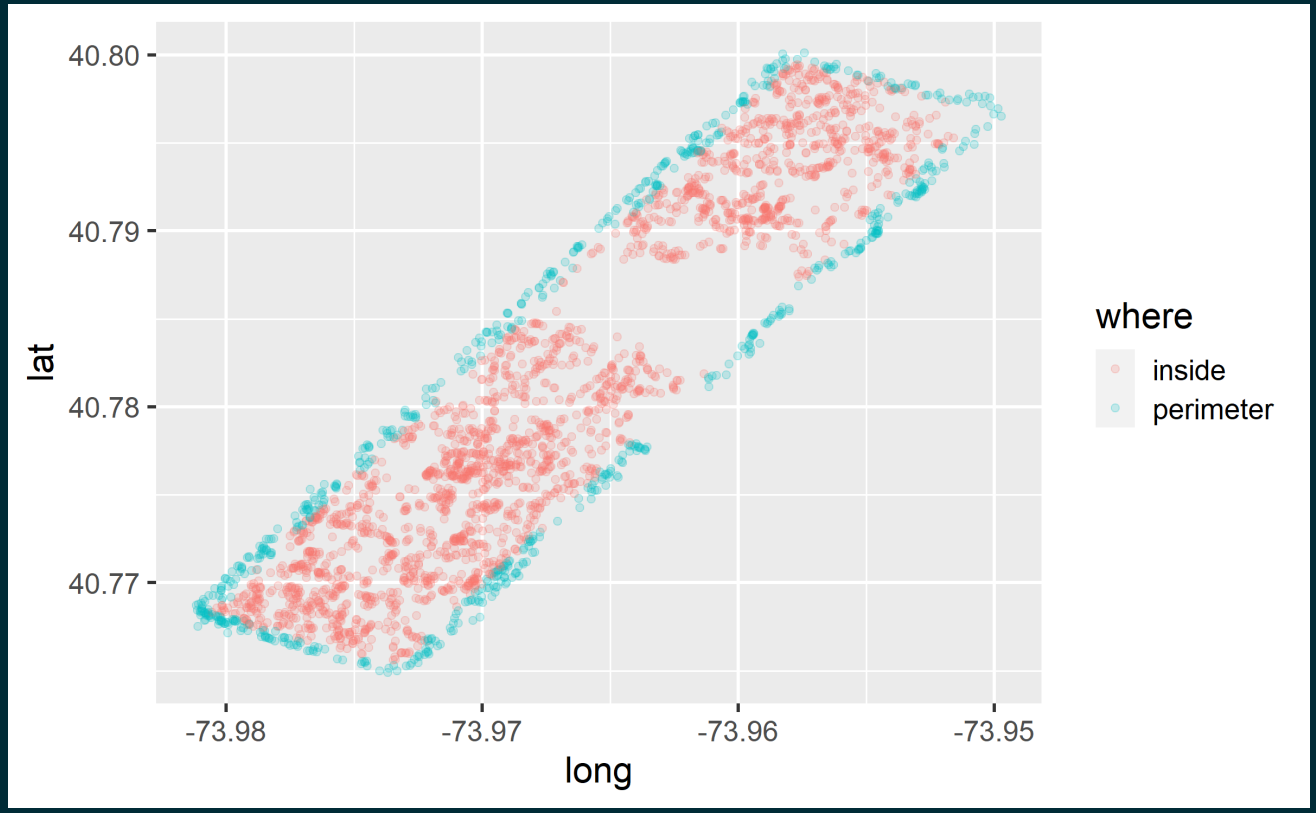
```
ggplot(squirrels, aes(x = long, y = lat)) +
  geom_point(alpha = 0.2)
```




**Hypothesis:** There will be a higher density of sightings on the perimeter than inside the park.

# Try the easy solution first

Plot Code



# Try the easy solution first

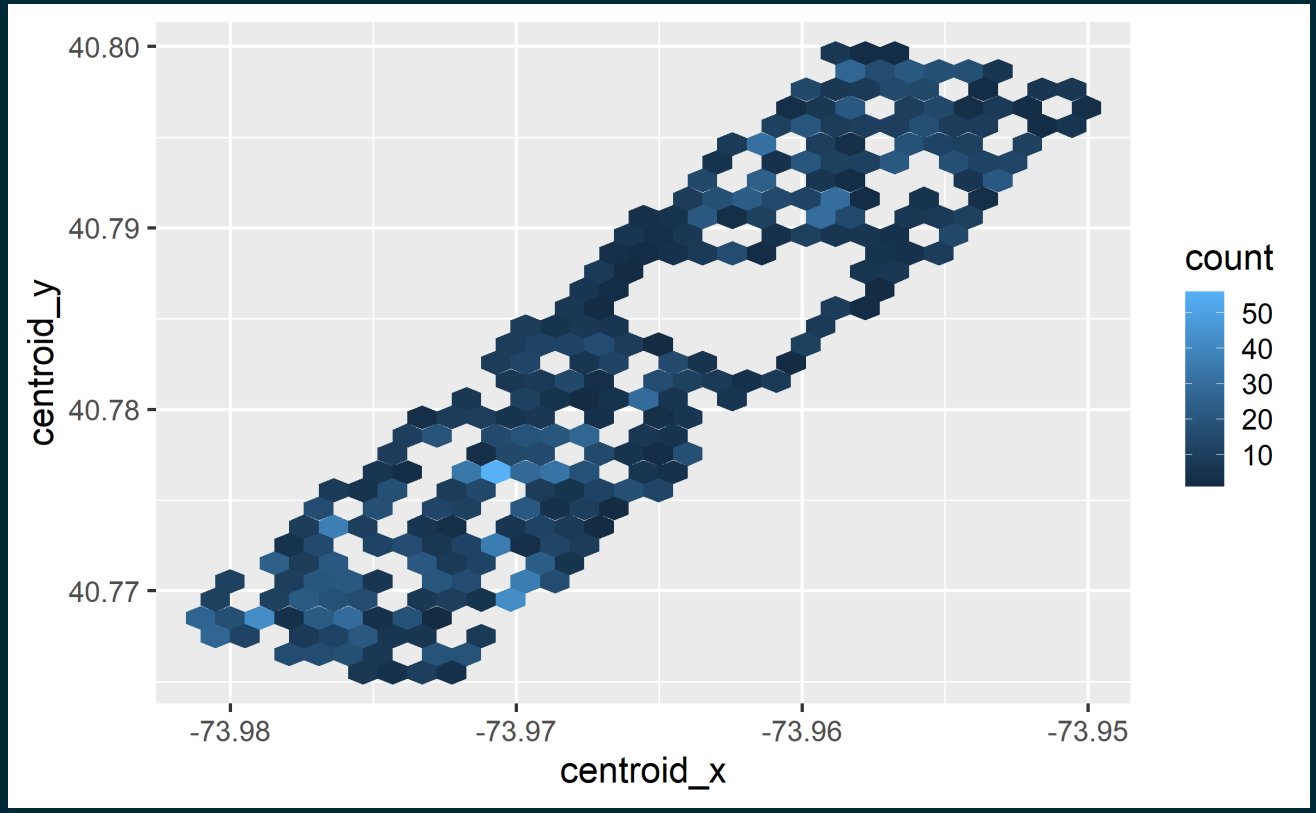
Plot | Code

```
squirrels <- squirrels %>%
  separate(hectare, into = c("NS", "EW"), sep = 2, remove = FALSE) %>%
  mutate(where = if_else(NS %in% c("01", "42") | EW %in% c("A", "I"), "perimeter", "inside"))

ggplot(squirrels, aes(x = long, y = lat, color = where)) +
  geom_point(alpha = 0.2)
```

# Then go deeper...

Plot Code



# Then go deeper...

Plot | Code

```
hectare_counts <- squirrels %>%
  group_by(hectare) %>%
  summarise(n = n())

hectare_centroids <- squirrels %>%
  group_by(hectare) %>%
  summarise(
    centroid_x = mean(long),
    centroid_y = mean(lat)
  )

squirrels %>%
  left_join(hectare_counts, by = "hectare") %>%
  left_join(hectare_centroids, by = "hectare") %>%
  ggplot(aes(x = centroid_x, y = centroid_y, color = n)) +
  geom_hex()
```

# The squirrel is staring at me!

```
squirrels %>%
  filter(str_detect(other_interactions, "star")) %>%
  select(shift, age, other_interactions)
```

```
## # A tibble: 11 x 3
##   shift age   other_interactions
##   <chr> <chr> <chr>
## 1 AM    Adult staring at us
## 2 PM    Adult he took 2 steps then turned and stared at me
## 3 PM    Adult stared
## 4 PM    Adult stared
## 5 PM    Adult stared
## 6 PM    Adult stared & then went back up tree—then ran to differ~
## # ... with 5 more rows
```

# Communicating for your audience

- Avoid: Jargon, uninterpreted results, lengthy output
- Pay attention to: Organization, presentation, flow
- Don't forget about: Code style, coding best practices, meaningful commits
- Be open to: Suggestions, feedback, taking (calculated) risks