



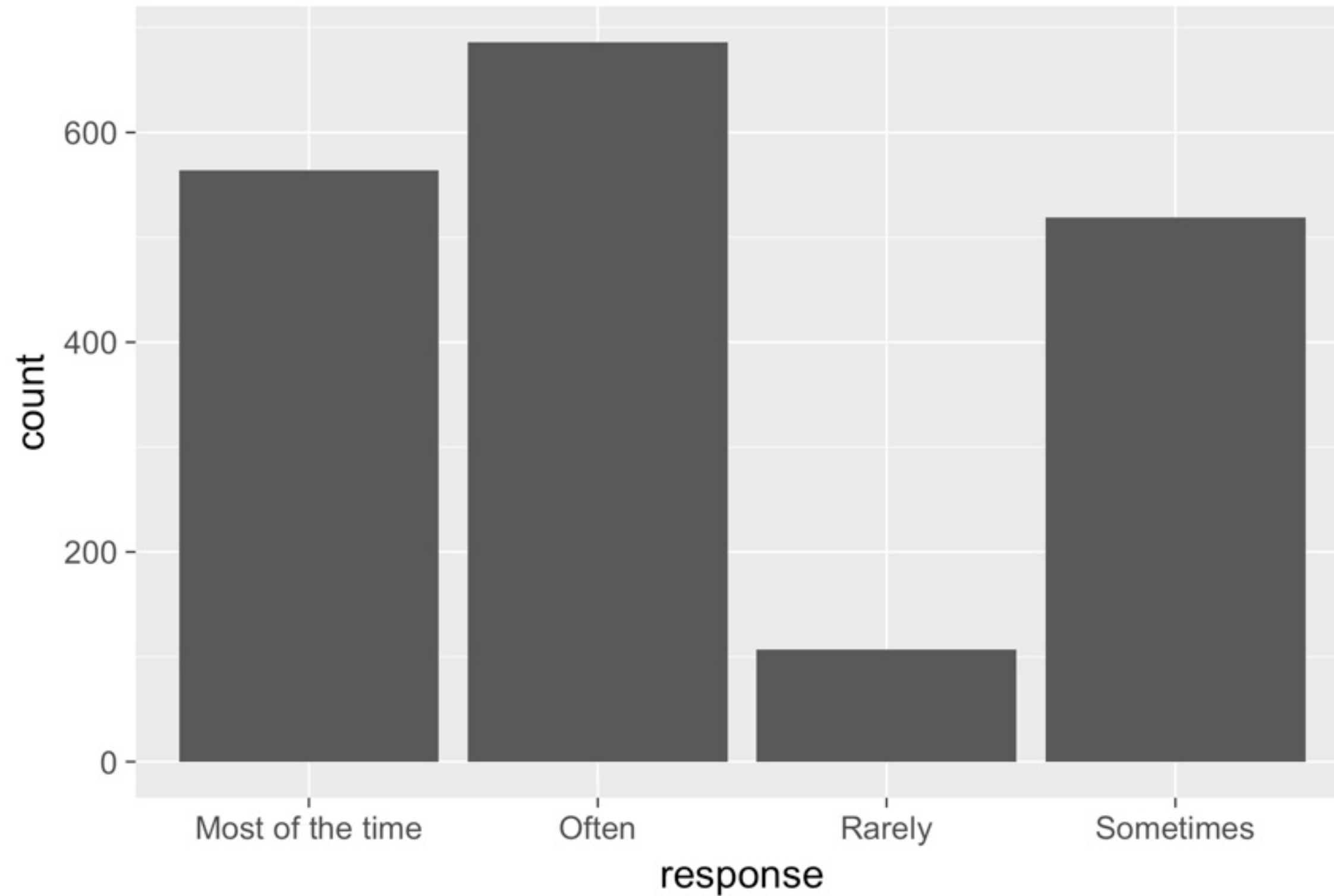
## CATEGORICAL DATA IN THE TIDYVERSE

# Reordering factors

Emily Robinson  
Data Scientist



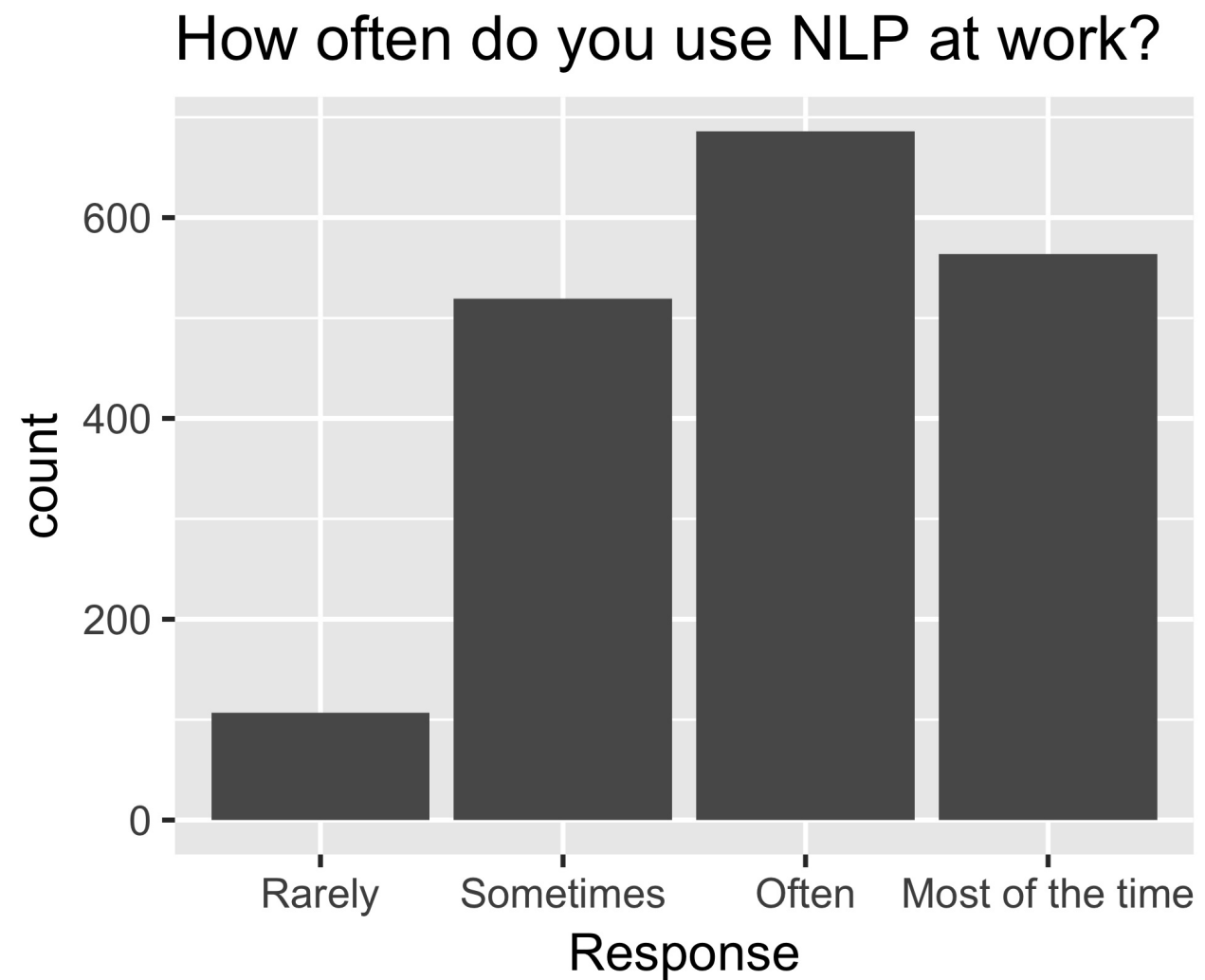
## How often do you use NLP at work?





# Corrected graph

```
ggplot(aes(nlp_frequency,  
           x = fct_relevel(response,  
                           "Rarely", "Sometimes", "Often", "Most of the time")) +  
       geom_bar())
```



# fct\_reorder()

```
nlp_frequency %>%  
  pull(response) %>%  
  levels()
```

```
[1] "Most of the time" "Often"          "Rarely"  
[4] "Sometimes"
```

```
nlp_frequency %>%  
  mutate(response = fct_relevel(response,  
                                "Often", "Most of the time")) %>%  
  pull(response) %>%  
  levels()
```

```
[1] "Often"          "Most of the time" "Rarely"  
[4] "Sometimes"
```

# Additional arguments

```
nlp_frequency %>%  
  mutate(response = fct_relevel(response,  
    "Often", "Most of the time", after = 2)) %>%  
  pull(response) %>%  
  levels()
```

```
nlp_frequency %>%  
  mutate(response = fct_relevel(response,  
    "Often", "Most of the time", after = Inf) %>%  
  pull(response) %>%  
  levels()
```

Both return:

```
[1] "Rarely"      "Sometimes"   "Often"  
[4] "Most of the time"
```



## CATEGORICAL DATA IN THE TIDYVERSE

**Let's practice!**



## CATEGORICAL DATA IN THE TIDYVERSE

# Renaming factor levels

Emily Robinson  
Data Scientist



# Introduction to FiveThirtyEight dataset

```
# A tibble: 1,040 x 27
  RespondentID travel_amount do_recline height
      <dbl> <fct>           <fct>    <fct>
1  3436139758 Once a year or le... NA        NA
2  3434278696 Once a year or le... About half the ... "6'3\"
3  3434275578 Once a year or le... Usually    "5'8\"
4  3434268208 Once a year or le... Always     "5'11\"
5  3434250245 Once a month or l... About half the ... "5'7\"
6  3434245875 Once a year or le... Usually    "5'9\"
7  3434235351 Once a month or l... Once in a while "6'2\"
8  3434218031 Once a year or le... Once in a while "6'0\"
9  3434213681 Once a year or le... Once in a while "6'0\"
10 3434172894 Once a year or le... Never      "5'6\"
# ... with 1,030 more rows, and 23 more variables:
#   children_sub_18 <fct>, middle_arm_rest_three <fct>,
#   middle_arm_rest_two <fct>, window_shade_control <fct>,
#   rude_move_seats <fct>, rude_talk <fct>,
#   times_get_up <fct>, recliner_obligation <fct>,
#   rude_recline <fct>, eliminate_recline <fct>,
#   rude_switch_seats_friend <fct>,
#   rude_switch_seats_family <fct>, rude_bathroom <fct>,
#   rude_walking <fct>, rude_baby <fct>,
#   rude_unruly_children <fct>, personal_electronics <fct>,
#   smoking <fct>, gender <fct>, age <fct>, income <fct>,
#   education <fct>, location <fct>
```





# fct\_recode()

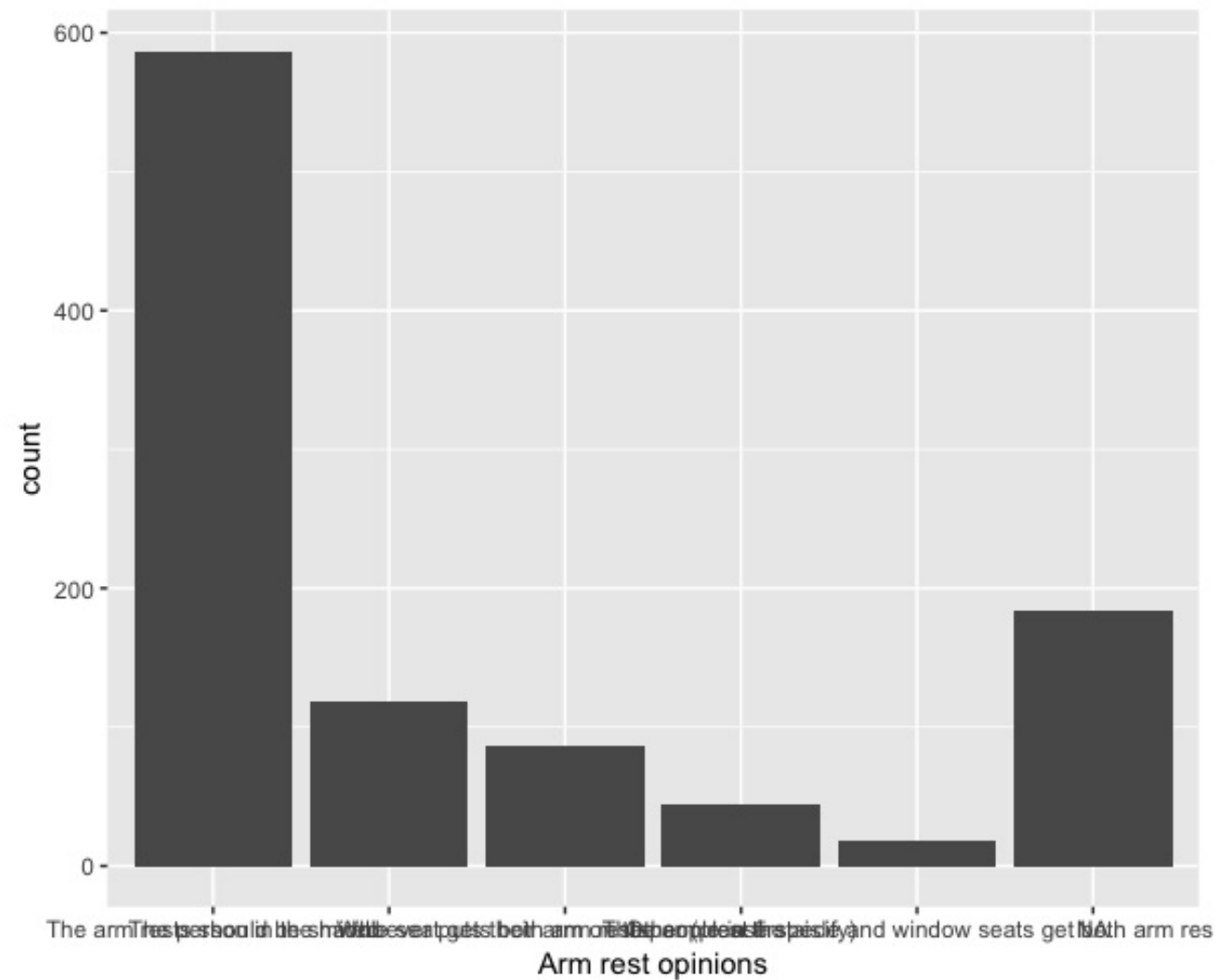
```
levels(flying_etiquette$middle_arm_rest_three)
[1] "Other (please specify)"
[2] "The arm rests should be shared"
[3] "The people in the aisle and window seats get both arm rests"
[4] "The person in the middle seat gets both arm rests"
[5] "Whoever puts their arm on the arm rest first"
```

```
ggplot(flying_etiquette, aes(x = fct_infreq(middle_arm_rest_three))) +
  geom_bar() +
  labs(x = "Arm rest opinions")
```



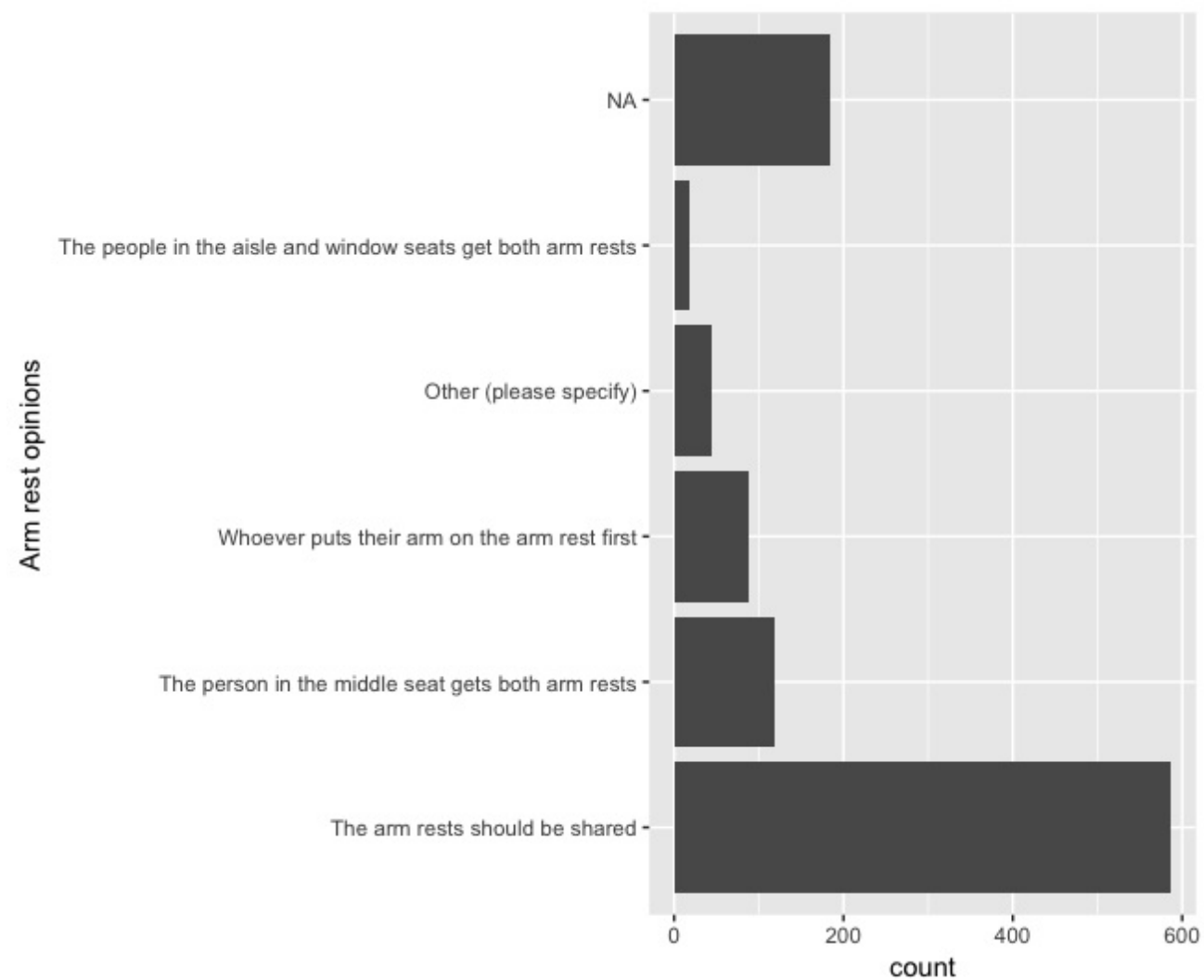
# Crowded text

```
ggplot(flying_etiquette, aes(x = fct_infreq(middle_arm_rest_three))) +  
  geom_bar() +  
  labs(x = "Arm rest opinions")
```



# Extraneous text

```
ggplot(flying_etiquette, aes(x = fct_infreq(middle_arm_rest_three))) +  
  geom_bar() +  
  coord_flip() +  
  labs(x = "Arm rest opinions")
```



# Changing with fct\_recode()

```
flying_etiquette %>%  
  mutate(middle_arm_rest_three = fct_recode(middle_arm_rest_three,  
    "Other" = "Other (please specify)",  
    "Everyone should share" = "The arm rests should be shared",  
    "Aisle and window people" =  
    "The people in the aisle and window seats get both arm rests",  
    "Middle person" = "The person in the middle seat gets both arm rests",  
    "Fastest person" = "Whoever puts their arm on the arm rest first"  
  )) %>%  
  count(middle_arm_rest_three)
```

```
# A tibble: 6 x 2  
  middle_arm_rest_three      n  
  <fct>                <int>  
1 Everyone should share    587  
2 Middle person           119  
3 Fastest person           87  
4 Other                    45  
5 Aisle and window people  18  
6 NA                      184
```



# Renaming a couple levels

```
flying_etiquette %>%  
  mutate(middle_arm_rest_three = fct_recode(middle_arm_rest_three,  
    "Everyone should share" = "The arm rests should be shared")) %>%  
  count(middle_arm_rest_three)
```

```
# A tibble: 6 x 2  
  middle_arm_rest_three      n  
  <fct>                <int>  
1 Other (please specify)      45  
2 Everyone should share     587  
3 The people in the aisle and window seats get both ...    18  
4 The person in the middle seat gets both arm rests     119  
5 Whoever puts their arm on the arm rest first         87  
6 NA                184
```

# Renaming unknown levels

```
library(flying_etiquette) %>%
  mutate(middle_arm_rest_three = fct_recode(middle_arm_rest_three,
    "Everyone should share" = "arm rests should be share")) %>%
  count(middle_arm_rest_three)
```

```
# A tibble: 6 x 2
  middle_arm_rest_three      n
  <fct>                <int>
1 Other (please specify)      45
2 The arm rests should be shared 587
3 The people in the aisle and window seats get both ...   18
4 The person in the middle seat gets both arm rests    119
5 Whoever puts their arm on the arm rest first         87
6 NA                184
```

Warning message:

Unknown levels in `f`: arm rests should be share



## CATEGORICAL DATA IN THE TIDYVERSE

**Let's practice!**

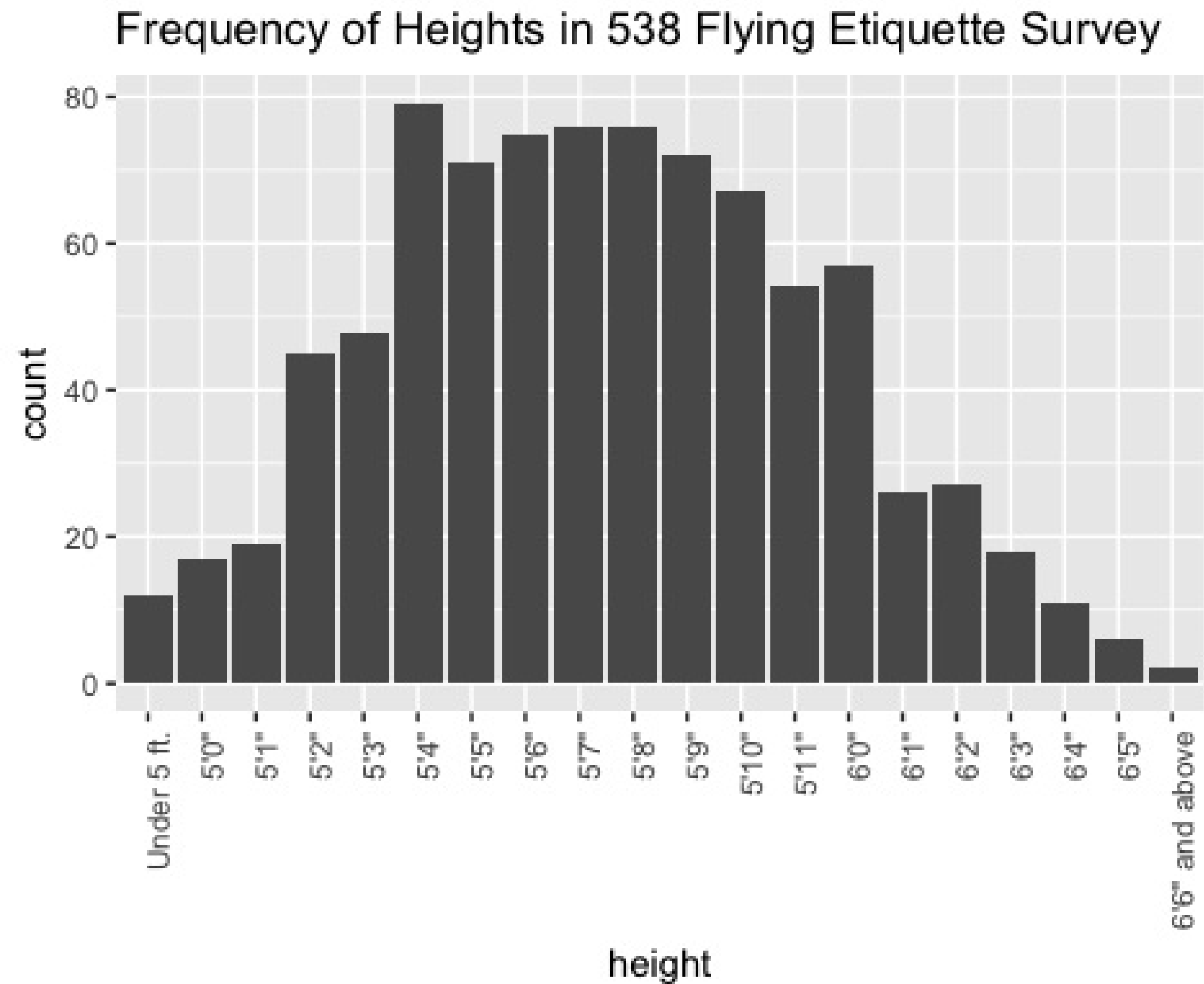


CATEGORICAL DATA IN THE TIDYVERSE

# **Collapsing factor levels**

**Emily Robinson**  
Data Scientist







# fct\_collapse()

```
flying_etiquette %>%  
  mutate(height = fct_collapse(height,  
    under_5_3 = c("Under 5 ft.", "5'0\\", "5'1\\", "5'2\\"),  
    over_6_1 = c("6'1\\", "6'2\\", "6'3\\", "6'4\\",  
      "6'5\\", "6'6\\ and above")) %>%  
  pull(height) %>%  
  levels()
```

```
[1] "under_5_3" "5'10\\" "5'11\\" "5'3\\"  
[5] "5'4\\" "5'5\\" "5'6\\" "5'7\\"  
[9] "5'8\\" "5'9\\" "6'0\\" "over_6_1"
```



# fct\_other() keep

```
flying_etiquette %>%  
  mutate(new_height = fct_other(height, keep = c("6'4\"", "5'1\""))) %>%  
  count(new_height)
```

```
# A tibble: 4 x 2  
  new_height      n  
  <fct>         <int>  
1 "5'1\""         19  
2 "6'4\""         11  
3 Other         828  
4 NA            182
```



# fct\_other() drop

```
flying_etiquette %>%  
  mutate(new_height = fct_other(height,  
    drop = c("Under 5 ft.", "5'0\\", "5'1\\", "5'2\\", "5'3\\")) %>%  
    pull(new_height) %>%  
    levels()
```

```
[1] "5'4\\" "5'5\\" "5'6\\"  
[4] "5'7\\" "5'8\\" "5'9\\"  
[7] "5'10\\" "5'11\\" "6'0\\"  
[10] "6'1\\" "6'2\\" "6'3\\"  
[13] "6'4\\" "6'5\\" "6'6\\" and above"  
[16] "Other"
```



# fct\_lump() prop

```
flying_etiquette %>%  
  mutate(new_height = fct_lump(height, prop = .08)) %>%  
  count(new_height)
```

	new_height	n
	<fct>	<int>
1	"5'4\""	79
2	"5'6\""	75
3	"5'7\""	76
4	"5'8\""	76
5	Other	552
6	NA	182



# fct\_lump() n

```
flying_etiquette %>%  
  mutate(new_height = fct_lump(height, n = 3)) %>%  
  count(new_height)
```

	new_height	n
	<fct>	<int>
1	"5'4\""	79
2	"5'7\""	76
3	"5'8\""	76
4	Other	627
5	NA	182



## CATEGORICAL DATA IN THE TIDYVERSE

**Let's practice!**