



CATEGORICAL DATA IN THE TIDYVERSE

Case study introduction

Emily Robinson
Data Scientist



Hell Is Other People In A Pressurized Metal Tube

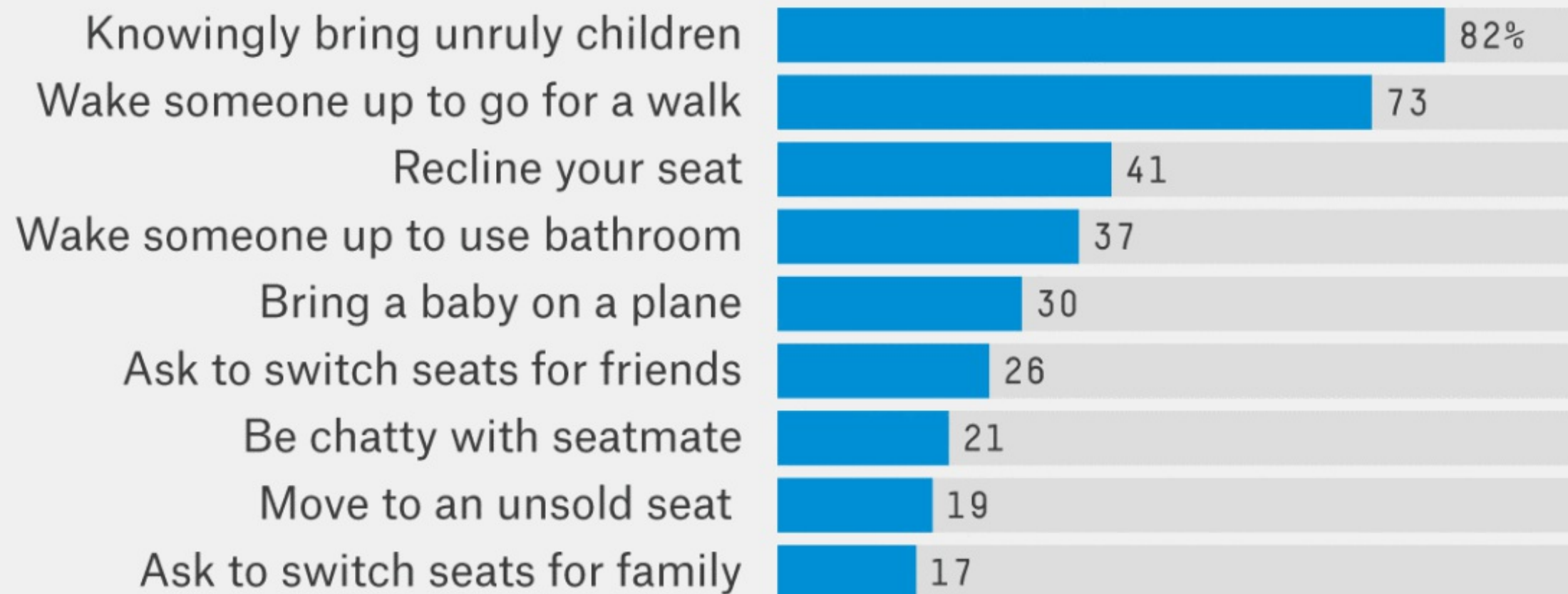
Percentage of 874 air-passenger respondents who said action is very or somewhat rude

SURVEY DATES

Aug. 29-30, 2014

NO. OF RESPONDENTS

1,040



Original dataset

```
# A tibble: 1,040 x 27
  RespondentID travel_amount do_recline
      <dbl> <chr>           <chr>
1  3436139758. Once a year or... NA
2  3434278696. Once a year or... About half t...
3  3434275578. Once a year or... Usually
4  3434268208. Once a year or... Always
# ... with 24 more variables: height <chr>,
#   children_sub_18 <chr>,
#   middle_arm_rest_three <chr>,
#   middle_arm_rest_two <chr>,
#   window_shade_control <chr>,
#   rude_move_seats <chr>, rude_talk <chr>,
#   times_get_up <chr>,
#   recliner_obligation <chr>,
#   rude_recline <chr>,
#   eliminate_recline <chr>,
#   rude_switch_seats_friend <chr>,
```

Tools recap

```
wide_data
```

```
# A tibble: 2 x 3
```

	favorite_fruit	favorite_vegetable	disliked_dessert
	<chr>	<chr>	<chr>
1	apple	carrot	cookie
2	orange	cauliflower	cake

```
wide_data %>%
```

```
  mutate_if(is.character, as.factor)
```

```
# A tibble: 2 x 3
```

	favorite_fruit	favorite_vegetable	disliked_dessert
	<fct>	<fct>	<fct>
1	apple	carrot	cookie
2	orange	cauliflower	cake



tidyr gather()

```
wide_data %>%  
  gather(column, value)  
  
# A tibble: 6 x 2  
  column      value  
  <chr>      <chr>  
1 favorite_fruit apple  
2 favorite_fruit orange  
3 favorite_vegetable carrot  
4 favorite_vegetable cauliflower  
5 disliked_dessert cookie  
6 disliked_dessert cake
```



Select helper functions

```
wide_data %>%  
  select(contains("favorite"))  
  
# A tibble: 2 x 2  
  favorite_fruit favorite_vegetable  
  <chr>         <chr>  
1 apple        carrot  
2 orange       cauliflower
```



CATEGORICAL DATA IN THE TIDYVERSE

Let's practice!



CATEGORICAL DATA IN THE TIDYVERSE

Data preparation and regex

Emily Robinson
Data Scientist

Handling long names

```
gathered_data %>%  
  distinct(response_var)  
  
# A tibble: 9 x 1  
  response_var  
  <chr>  
1 Is it rude to move to an unsold seat on a  
  plane?  
2 Generally speaking, is it rude to say  
  more than a few words to the stranger...  
3 Is it rude to recline your seat on a plane?  
4 Is it rude to ask someone to switch  
  seats with you in order to be closer to...  
5 Is it rude to ask someone to switch  
  seats with you in order to be closer to...  
6 Is it rude to wake a passenger up if  
  you are trying to go to the bathroom?  
7 Is it rude to wake a passenger up if  
  you are trying to walk around?  
8 In general, is it rude to bring a  
  baby on a plane?  
9 In general, is it rude to knowingly  
  bring unruly children on a plane?
```



Regex

```
str_detect("happy", ".")  
[1] TRUE
```

```
str_detect("happy", "h.")  
[1] TRUE
```

```
str_detect("happy", "y.")  
[1] FALSE
```



Regex

```
string <- "Statistics is the best"
```

```
str_remove(string, ".*the ")  
[1] "best"
```



CATEGORICAL DATA IN THE TIDYVERSE

Let's practice!



CATEGORICAL DATA IN THE TIDYVERSE

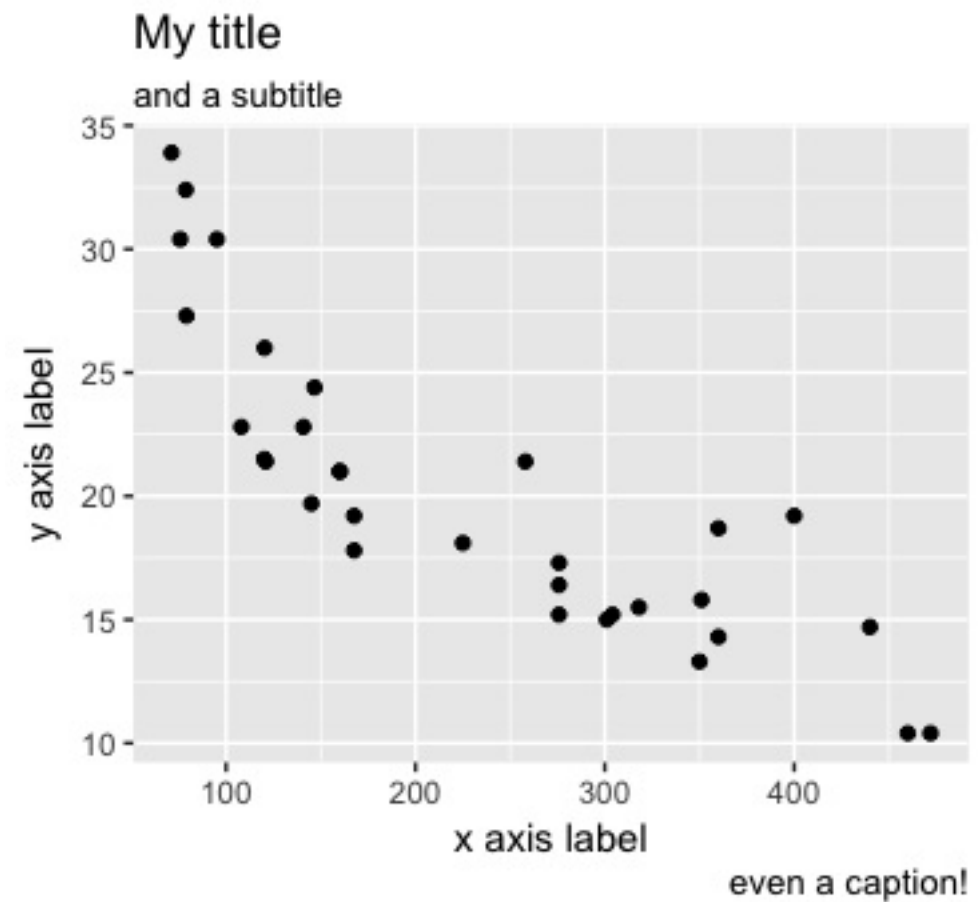
Recreating the plot

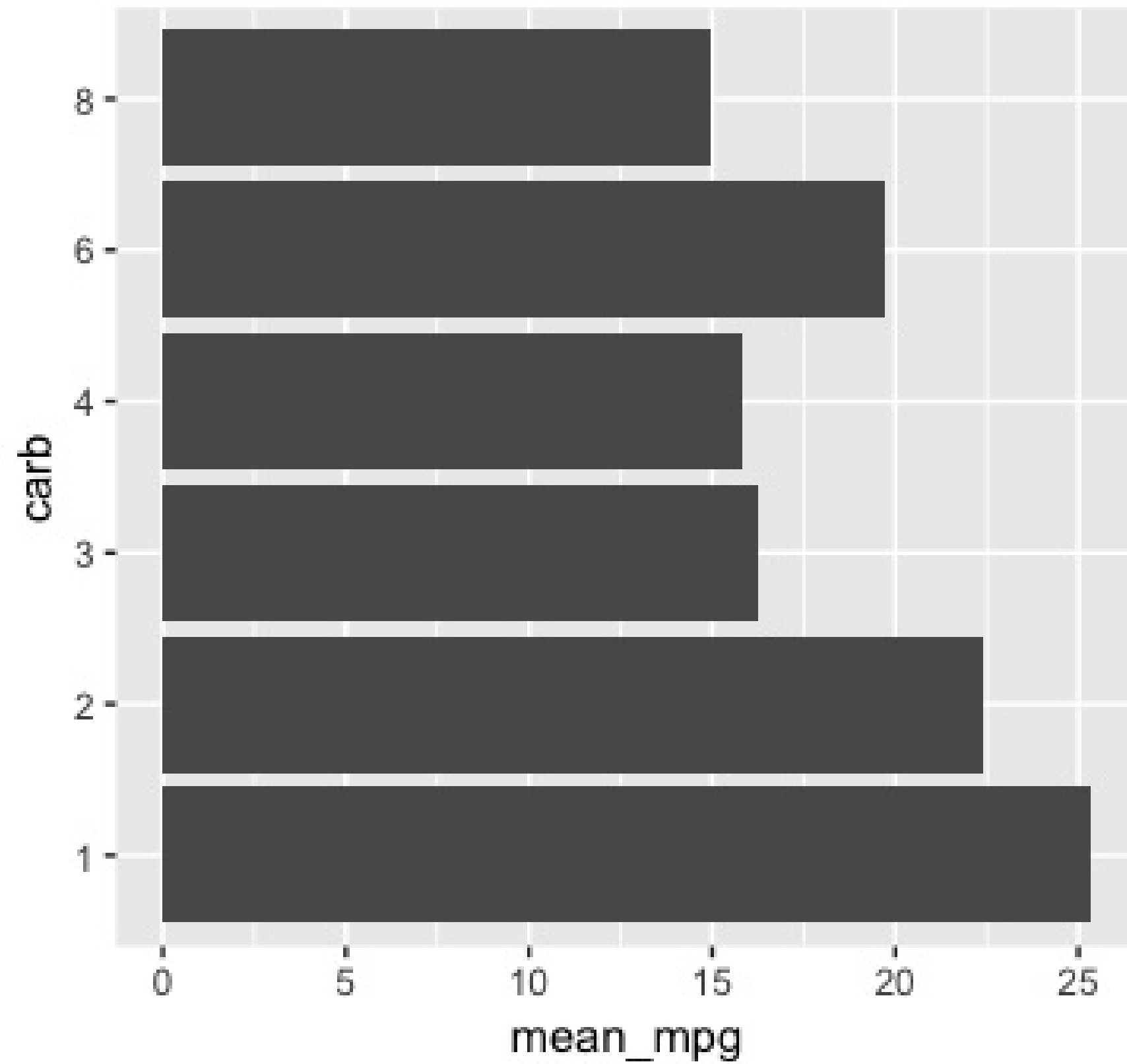
Emily Robinson
Data Scientist



Labs

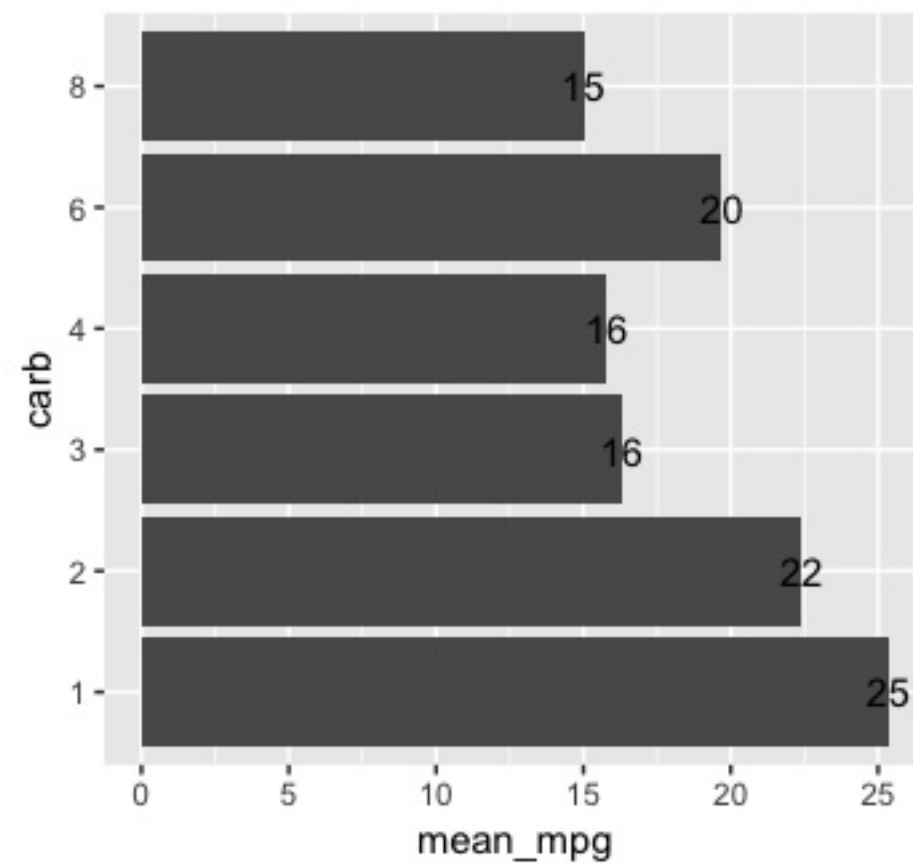
```
ggplot(mtcars, aes(displacement, mpg)) + geom_point() +  
  labs(x = "x axis label", y = "y axis label", title = "My title",  
        subtitle = "and a subtitle", caption = "even a caption!")
```





Geom_text

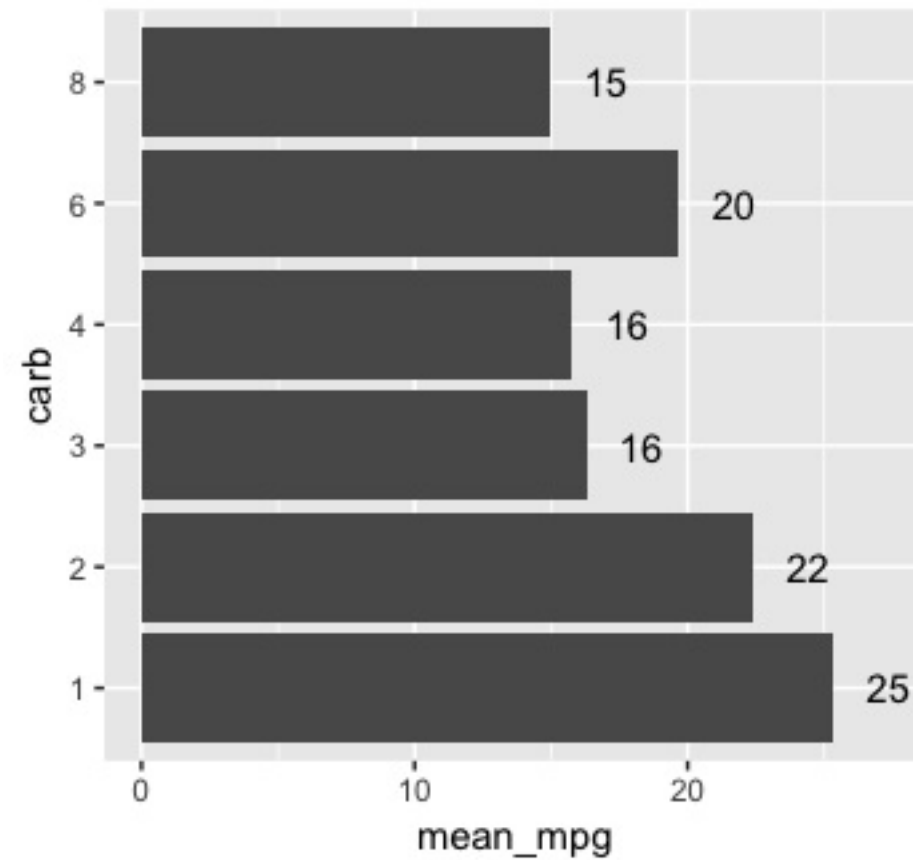
```
initial_plot +  
  geom_text(aes(label = round(mean_mpg)))
```





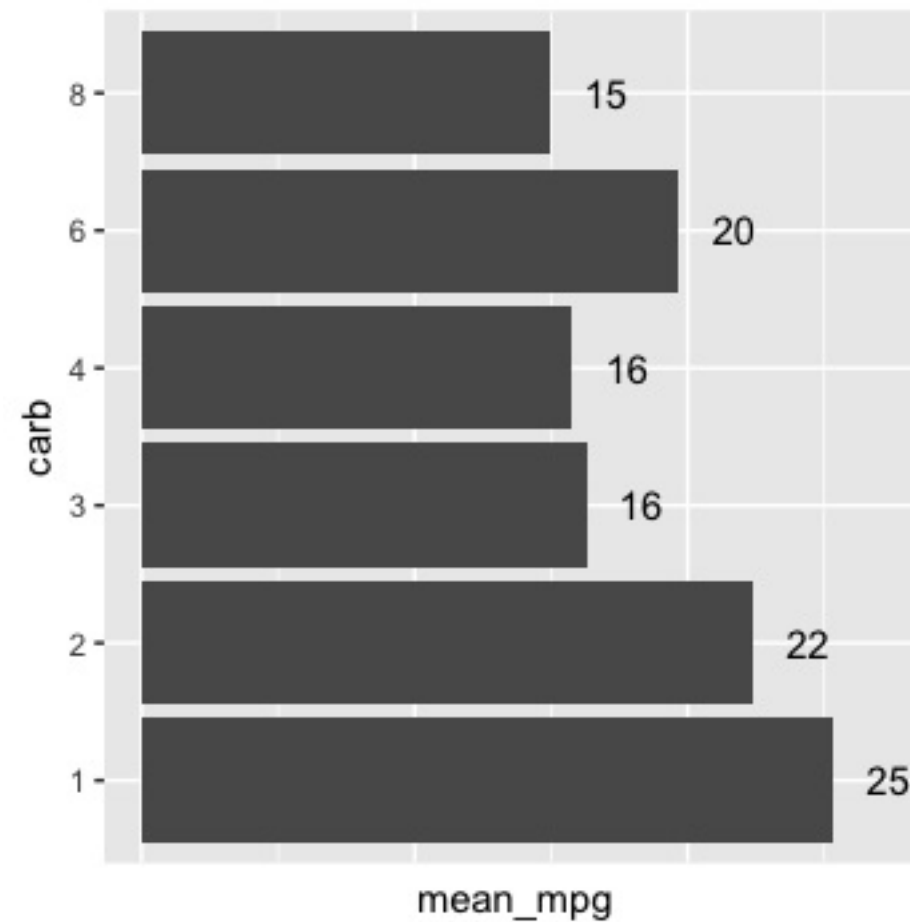
Moving text

```
initial_plot +  
  geom_text(aes(label = round(mean_mpg),  
                y = mean_mpg + 2))
```



Theme

```
initial_plot +  
  geom_text(aes(label = round(mean_mpg), y = mean_mpg + 2)) +  
  theme(axis.text.x = element_blank(),  
        axis.ticks.x = element_blank())
```





CATEGORICAL DATA IN THE TIDYVERSE

Let's practice!



CATEGORICAL DATA IN THE TIDYVERSE

Final thoughts

Emily Robinson
Data Scientist



What you've learned

- forcats functions:
 - `fct_reorder()`, `fct_collapse()`, `fct_other()`, `fct_relevel()`, `fct_rev()`, & `fct_recode()`
- tidyverse functions:
 - `case_when()`, `mutate_if()`, `gather()`, & `str_remove()`
- ggplot2 tricks:
 - `scales::percent_format()`, `labs()`, & `axis.text.x`
- Case study



CATEGORICAL DATA IN THE TIDYVERSE

Congratulations!