# Introduction to qualitative data

## CATEGORICAL DATA IN THE TIDYVERSE

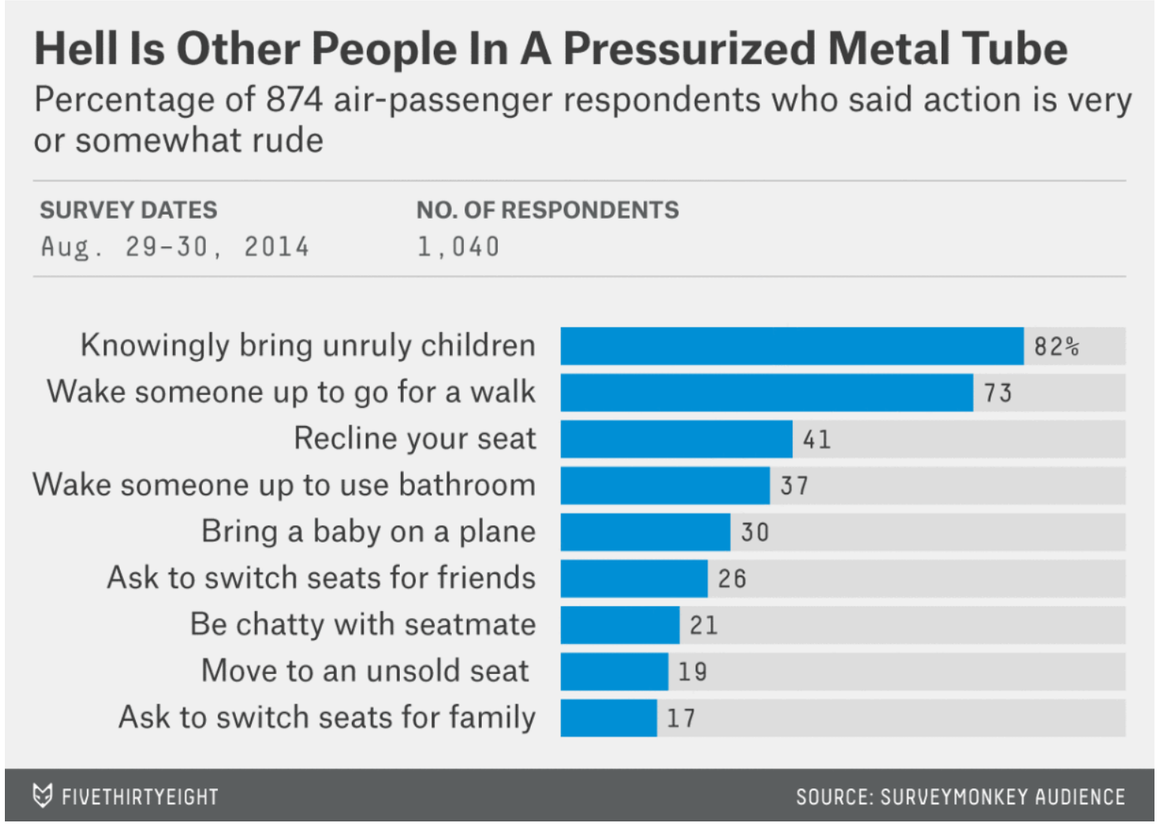
**Emily Robinson**
Data Scientist

datacamp

# Course overview

- Identifying and inspecting qualitative variables
- Working with the forcats package
- Making effective visualizations

# Final chapter




## 41% of Fliers Think You're Rude if You Recline Your Seat

# What are qualitative variables?

- Categorical vs. Ordinal data

# Categorical (nominal) data

# Ordinal data

Annual Income Options:

- "0-$50,000"
- "$50,000-150,000"
- "$150,000-500,000"
- "More than $500,000"

# Qualitative variables in R

- Names vs. question on programming languages

# Qualitative variables in R

- Look at your whole dataset

```
library(fivethirtyeight)
print(college_all_ages)
```

```
# A tibble: 173 x 11
   major_code major       major_category     total employed
        <int> <chr>       <chr>              <int>    <int>
 1       1100 General Ag… Agriculture & Na… 128148    90245
 2       1101 Agricultur… Agriculture & Na…  95326    76865
 3       1102 Agricultur… Agriculture & Na…  33955    26321
 4       1103 Animal Sci… Agriculture & Na… 103549    81177
# ... with 163 more rows, and 6 more variables:
#   employed_fulltime_yearround <int>, unemployed <int>,
#   unemployment_rate <dbl>, p25th <dbl>, median <dbl>,
#   p75th <dbl>
```

# Qualitative variables in R

- Look at your variables one at a time:

```
is.factor(college_all_ages$major_category)
```

```
FALSE
```

# Let's practice!

CATEGORICAL DATA IN THE TIDYVERSE

# Understanding your qualitative variables

## CATEGORICAL DATA IN THE TIDYVERSE

**Emily Robinson**
Data Scientist

datacamp

# Introduction to the dataset

- Dataset: Kaggle 2017 Data Science survey

```
# A tibble: 16,716 x 228
   GenderSelect        Country    Age EmploymentStatus
   <chr>               <chr>    <int> <chr>
 1 Non-binary, gender... NA          NA Employed full-time
 2 Female              United ...    30 Not employed, but lo...
 3 Male                Canada       28 Not employed, but lo…
 4 Male                United ...    56 Independent contract...
 5 Male                Taiwan       38 Employed full-time
 6 Male                Brazil       46 Employed full-time
 7 Male                United ...    35 Employed full-time
 8 Female              India        22 Employed full-time
 9 Female              Austral...    43 Employed full-time
10 Male                Russia       33 Employed full-time
# ... with 16,706 more rows, and 224 more variables:
#   StudentStatus <chr>, LearningDataScience <chr>,
```

# Converting characters to factors

```
is.character(multipleChoiceResponses$LearningDataScienceTime)
```

```
TRUE
```

```
multipleChoiceResponses %>%
    mutate_if(is.character, as.factor)
```

```
# A tibble: 16,716 x 228
   GenderSelect        Country    Age EmploymentStatus
   <fct>               <fct>    <int> <fct>
 1 Non-binary, gender  NA          NA Employed full-time
 2 Female              United ...    30 Not employed, but lo...
 3 Male                Canada      28 Not employed, but lo...
 4 Male                United ...    56 Independent contract...
# ... with 16,710 more rows, and 224 more variables:
#   StudentStatus <fct>, LearningDataScience <fct>,
#   CodeWriter <fct>, CareerSwitcher <fct>,
#   CurrentJobTitleSelect <fct>, TitleFit <fct>,
#   CurrentEmployerType <fct>, MLToolNextYearSelect <fct>,
```

# Summarising factors

- Get the number of categories (levels)

```
nlevels(multipleChoiceResponses$LearningDataScienceTime)
```

```
6
```

- Get the list of categories (levels)

```
levels(multipleChoiceResponses$LearningDataScienceTime)
```

```
[1] "< 1 year"    "1-2 years"   "10-15 years" "15+ years"
[5] "3-5 years"   "5-10 years"
```

# Summarising factors

- Get number of levels for every factor variable

```
multipleChoiceResponses %>%
  summarise_if(is.factor, nlevels)
```

```
# A tibble: 1 x 215
  GenderSelect Country EmploymentStatus StudentStatus
         <int>   <int>            <int>         <int>
1            4      52                7             2
# ... with 211 more variables: LearningDataScience <int>,
#   CodeWriter <int>, CareerSwitcher <int>,
```

# Let's practice!

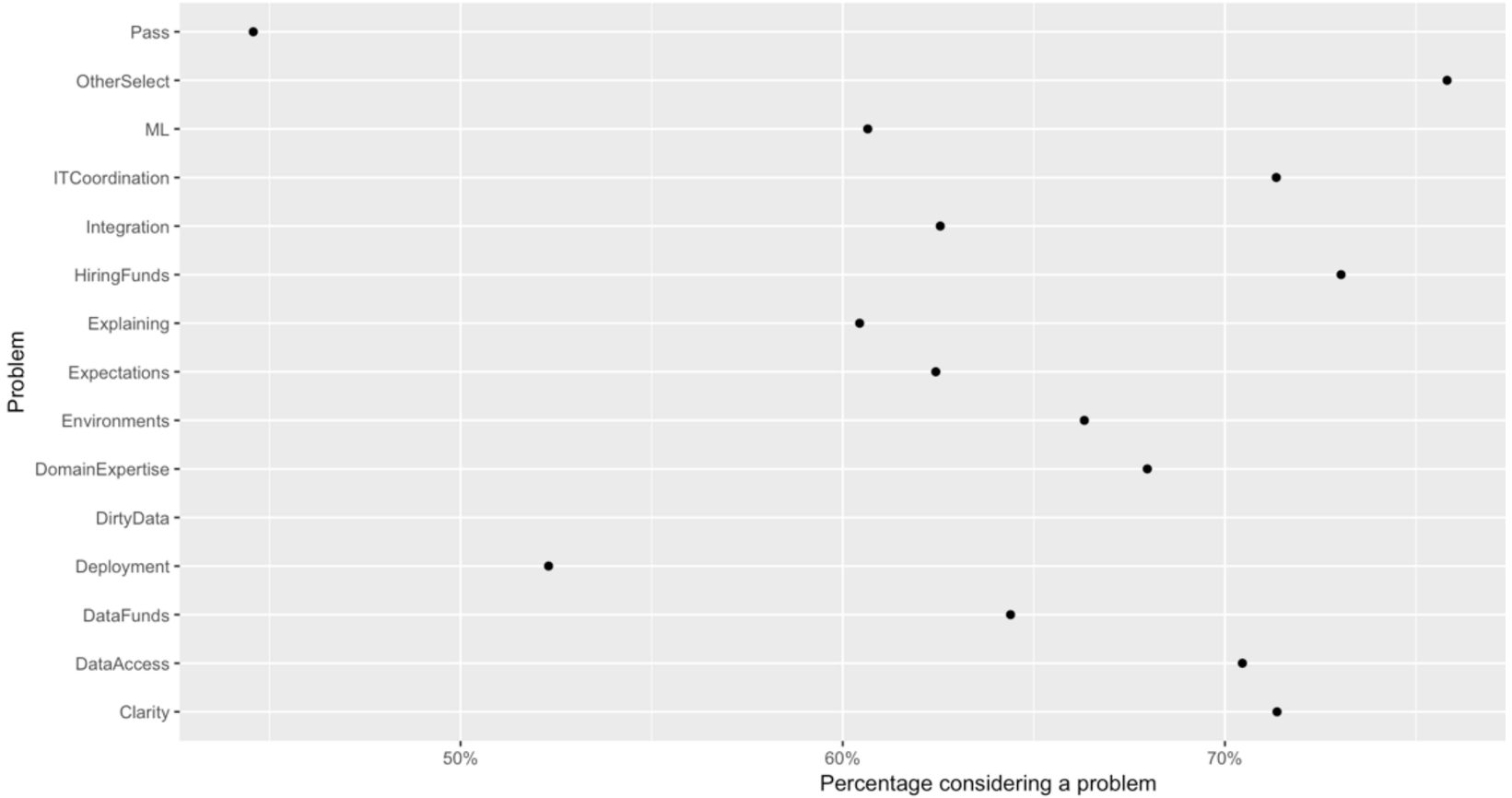
CATEGORICAL DATA IN THE TIDYVERSE

# Making better plots

## CATEGORICAL DATA IN THE TIDYVERSE

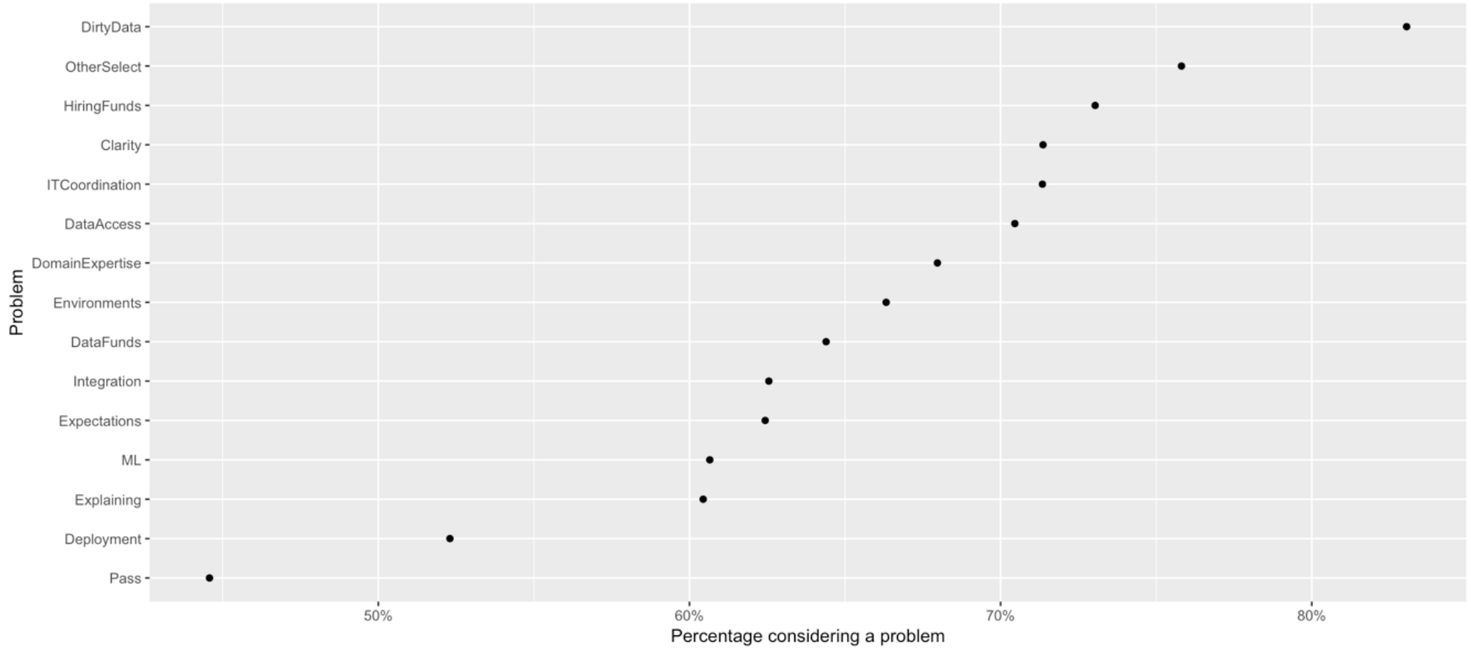
**Emily Robinson**
Data Scientist

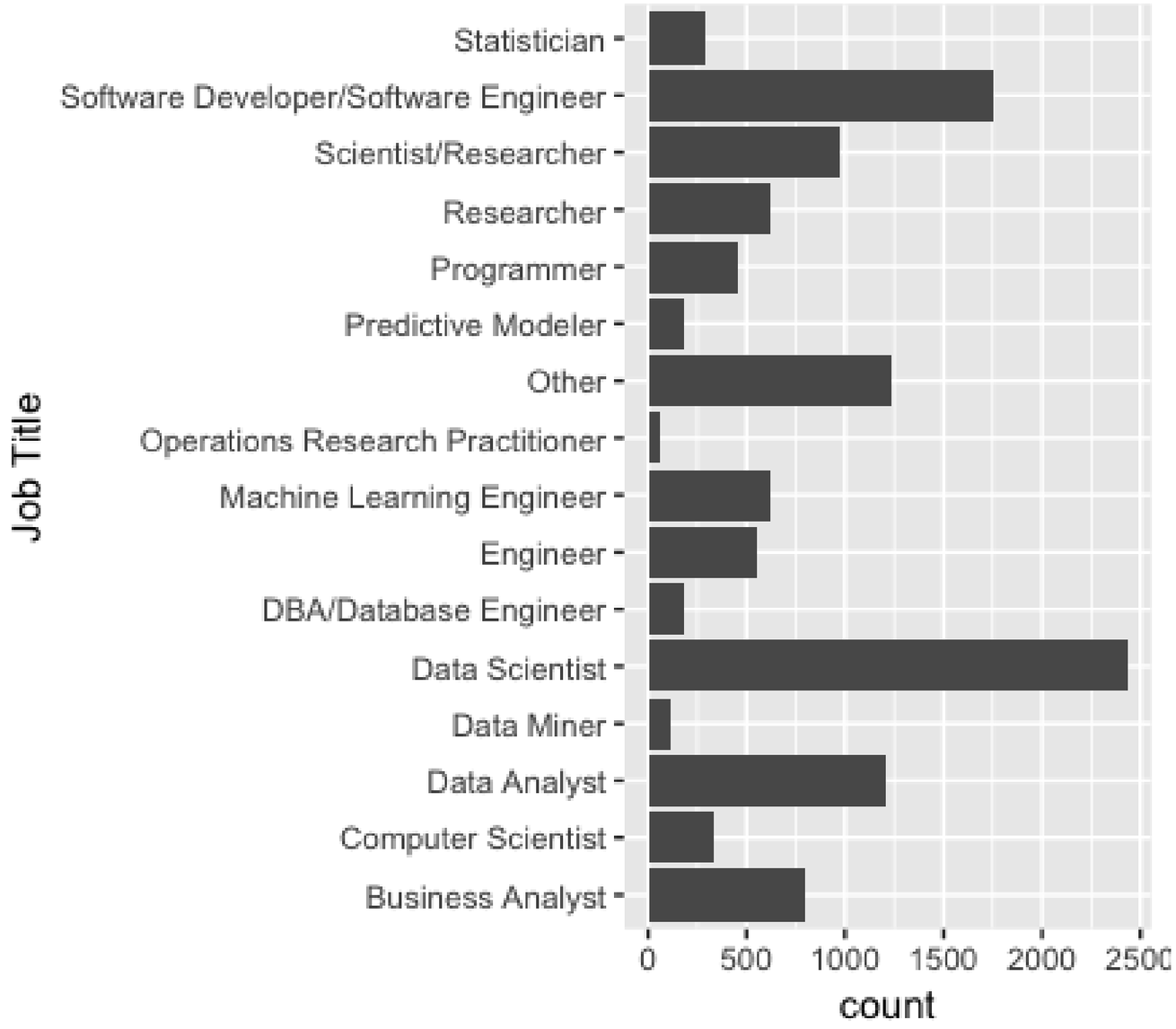
datacamp


Pass
OtherSelect
ML
ITCoordination
Integration
HiringFunds
Explaining
Expectations
Environments
DomainExpertise
DirtyData
Deployment
DataFunds
DataAccess
Clarity
Problem
50%
60%
70%
Percentage considering a problem

# Reordering factors

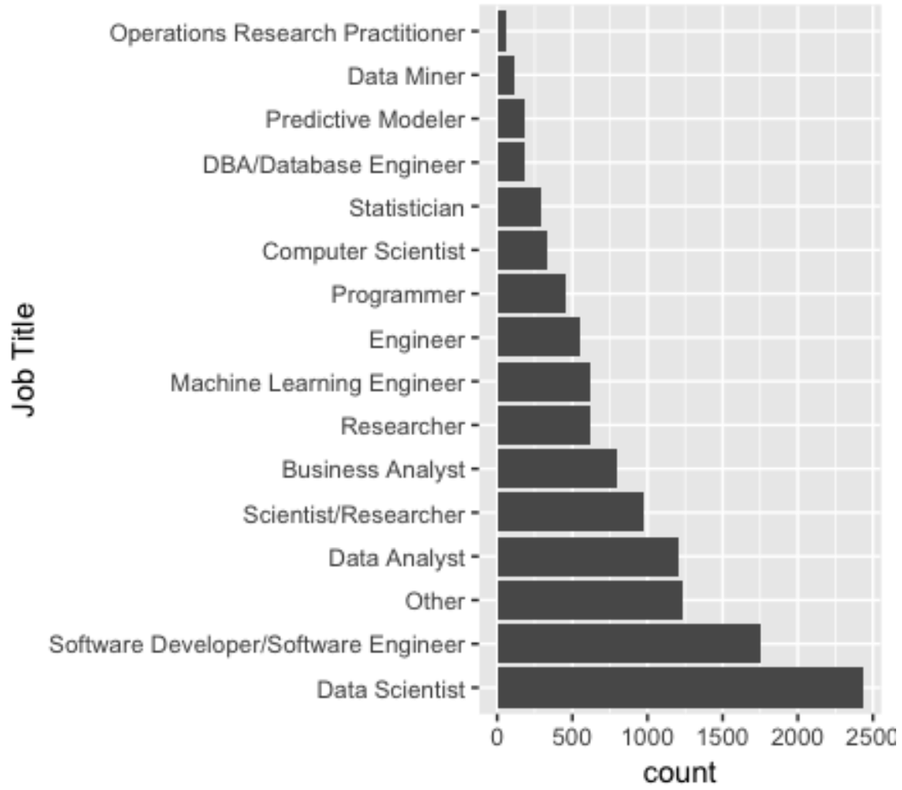
```
ggplot(WorkChallenges) +
  geom_point(aes(x = fct_reorder(question, perc_problem),
                 y = perc_problem))
```




Job Title
Statistician
Software Developer/Software Engineer
Scientist/Researcher
Researcher
Programmer
Predictive Modeler
Other
Operations Research Practitioner
Machine Learning Engineer
Engineer
DBA/Database Engineer
Data Scientist
Data Miner
Data Analyst
Computer Scientist
Business Analyst
0
500
1000
1500
2000
2500
count

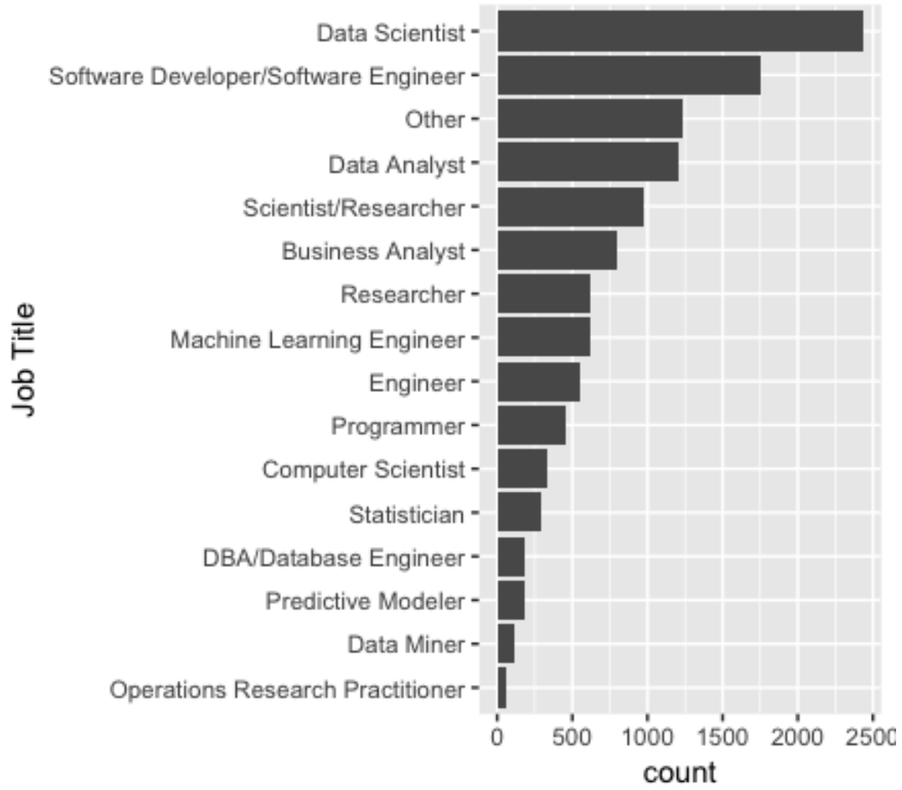
# Reordering bar chart

```
ggplot(multiple_choice_responses) +
   geom_bar(aes(x = fct_infreq(CurrentJobTitleSelect))
```



# Reversing factor levels

```
ggplot(multiple_choice_responses) +
    geom_bar(aes(x = fct_rev(fct_infreq(CurrentJobTitleSelect))))
```



# Let's practice!

CATEGORICAL DATA IN THE TIDYVERSE