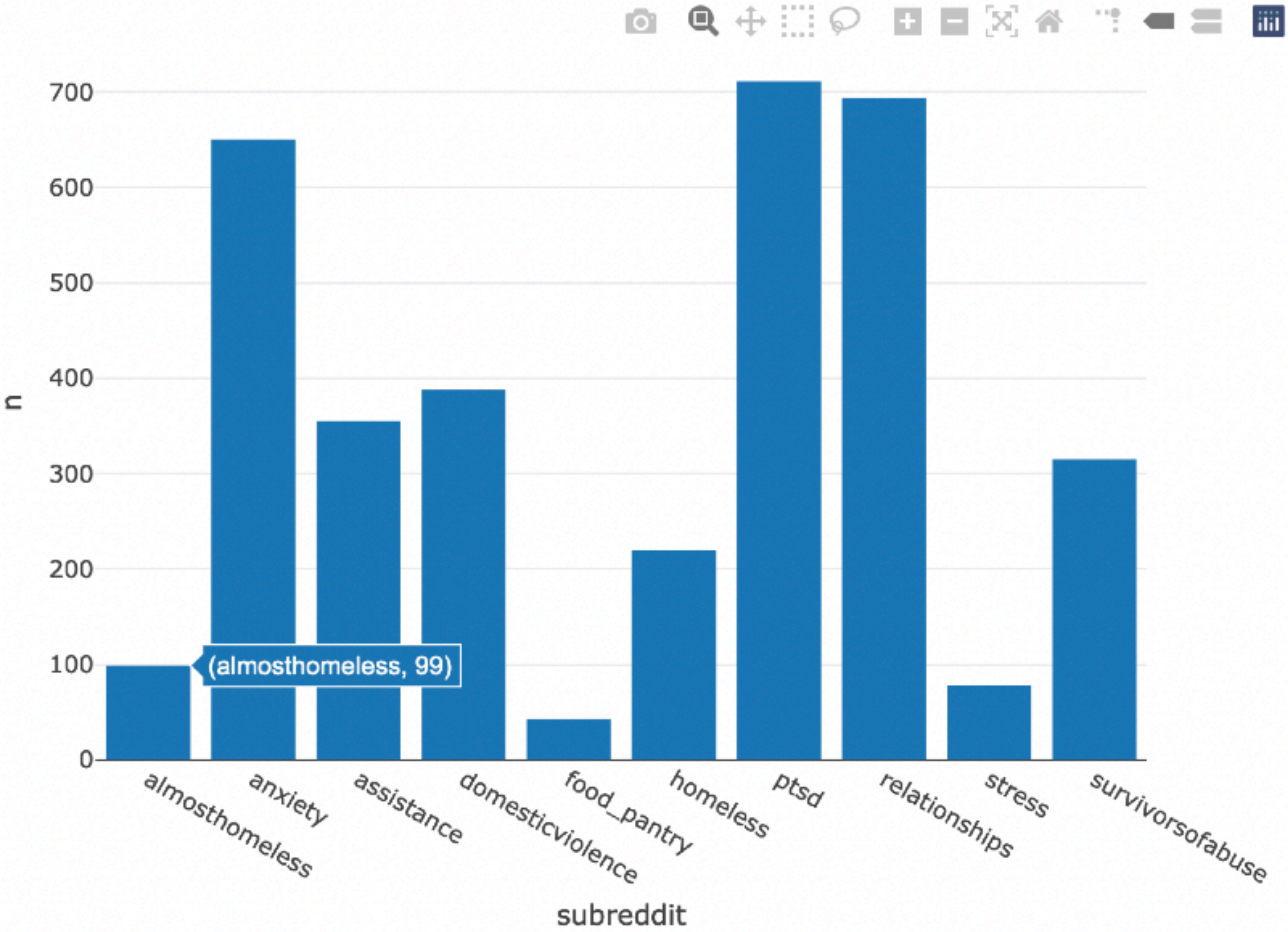
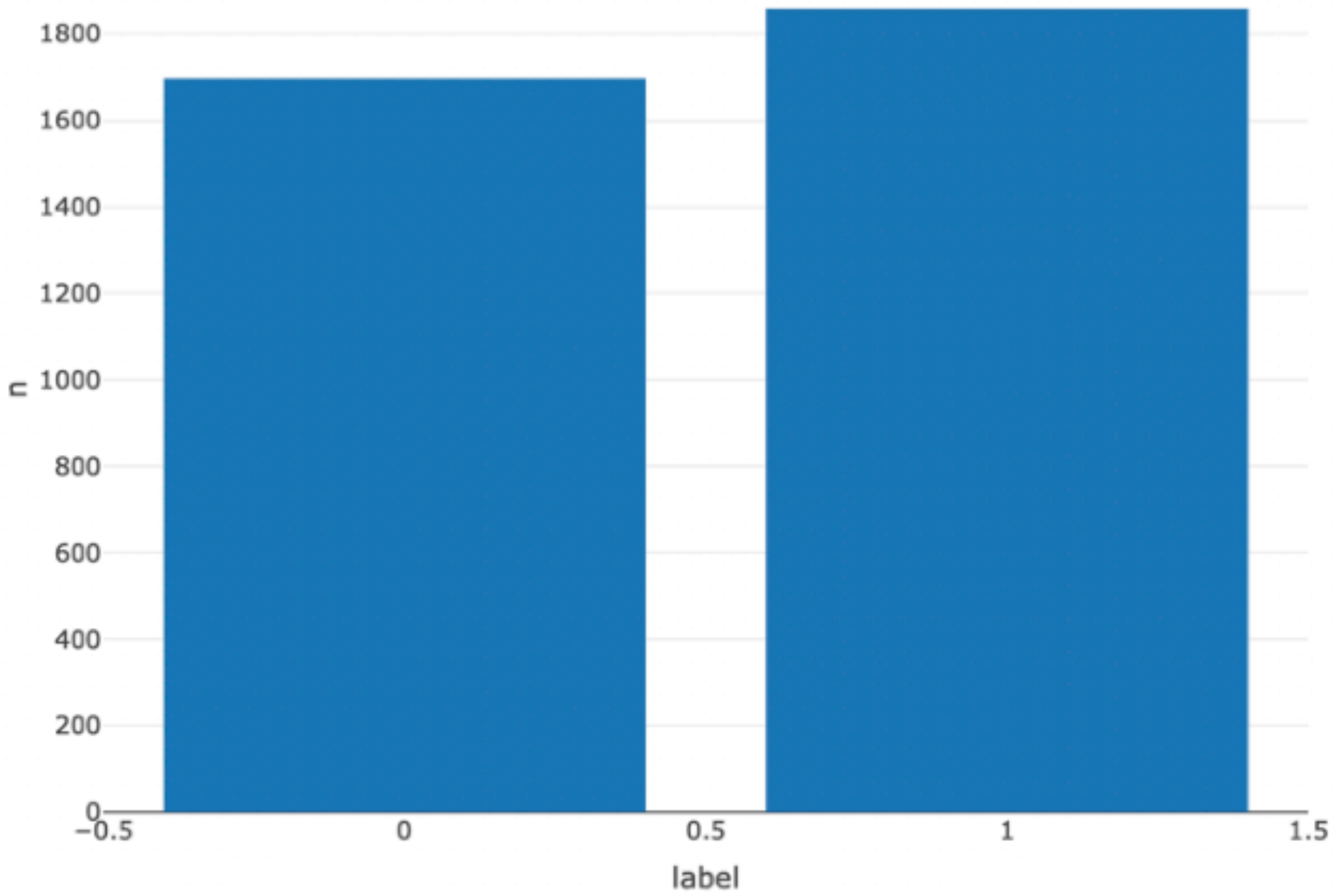
Dataset Overview

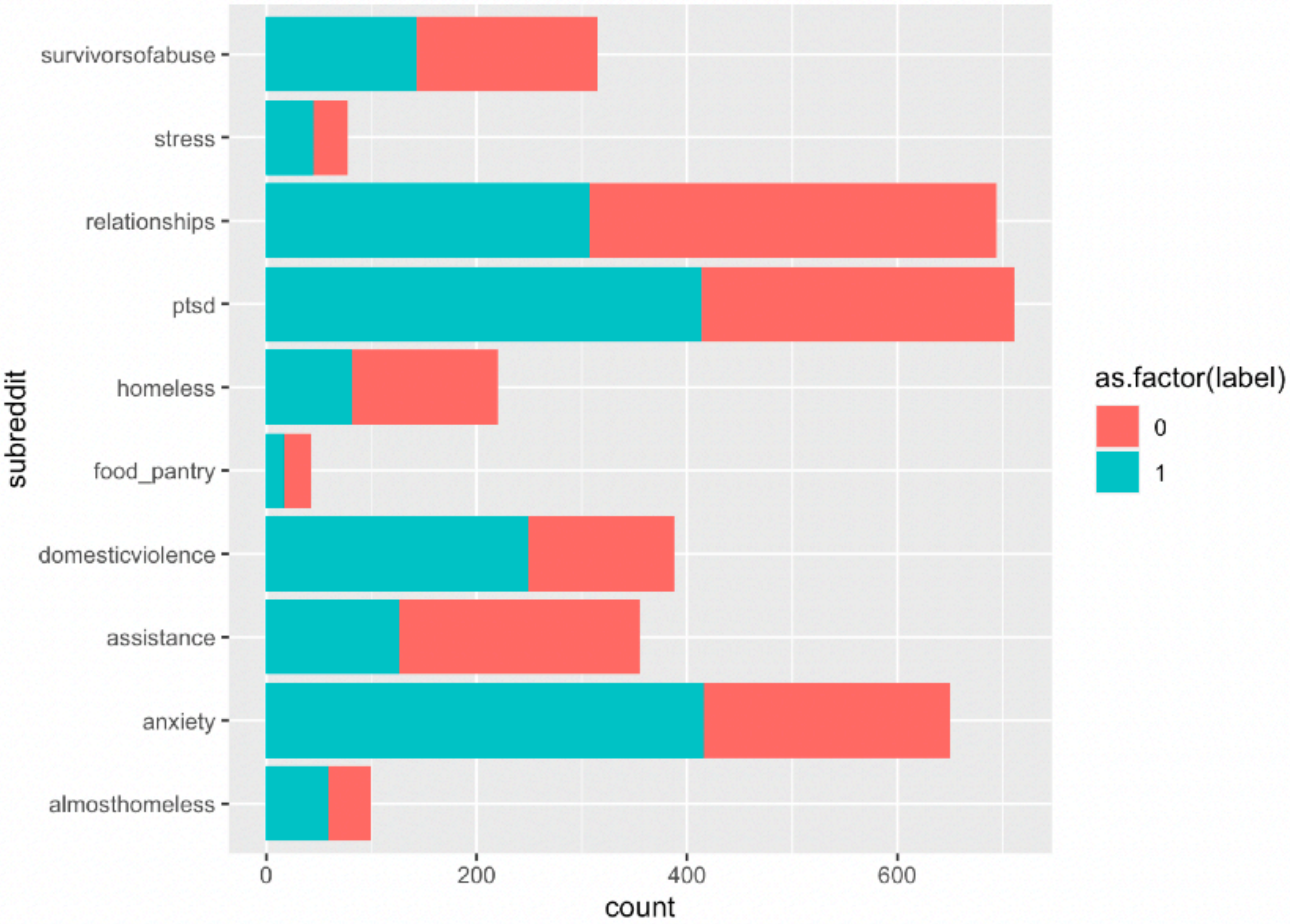
- Dataset consists of training data with 2838 posts, test data with 715 posts
- Sampled from 11 subreddits

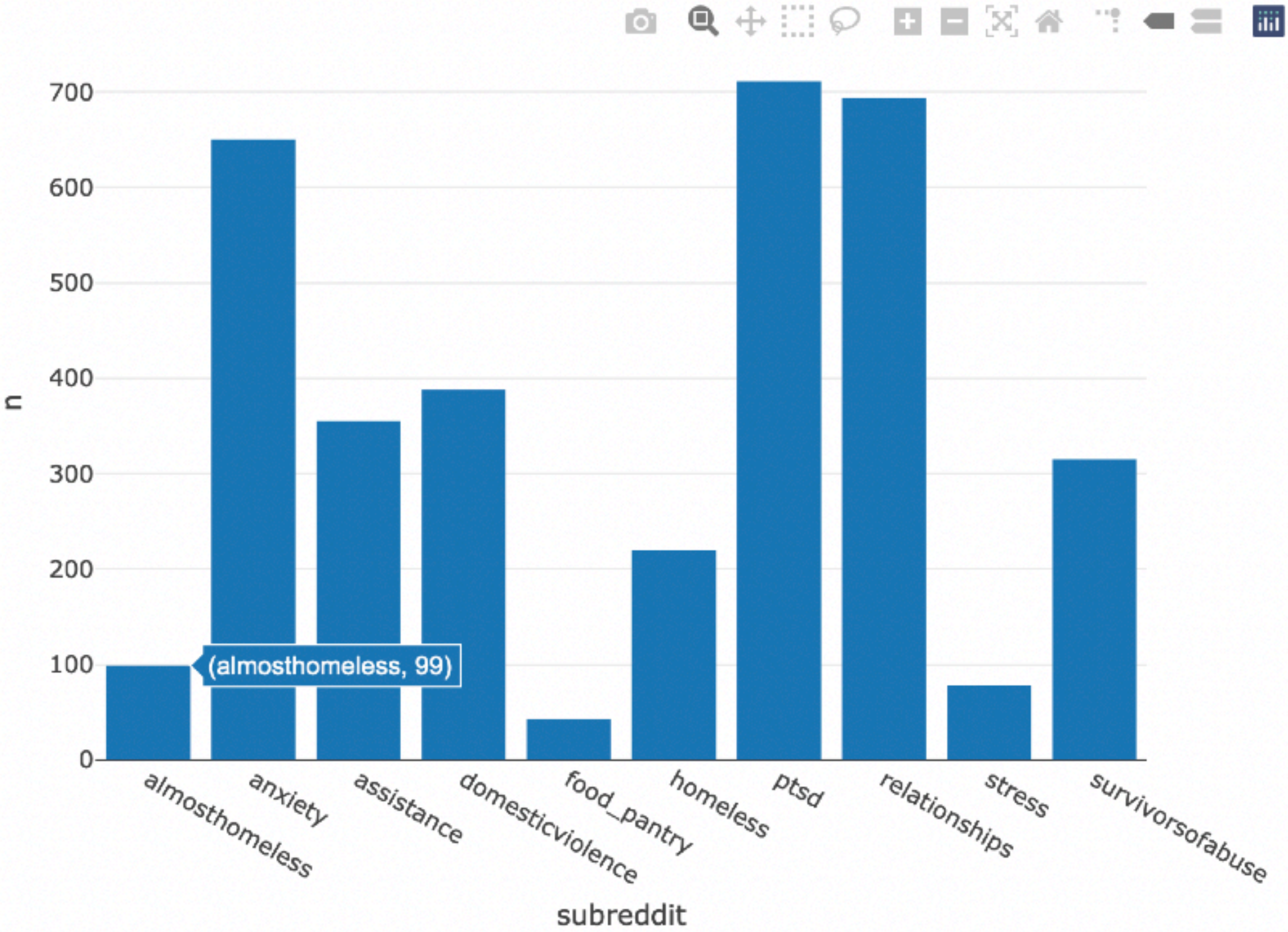

700
600
500
400
300
200
100
0
n
(almosthomeless, 99)
almosthomeless
anxiety
assistance
domesticviolence
food_pantry
homeless
ptsd
relationships
stress
survivorsofabuse
subreddit

Getting the Labels Distribution

```r
label_counts <- reddit_stress_data %>%
  group_by(label) %>%
  count()
plot_ly(label_counts, x = ~label, y = ~n, type = "bar")
```

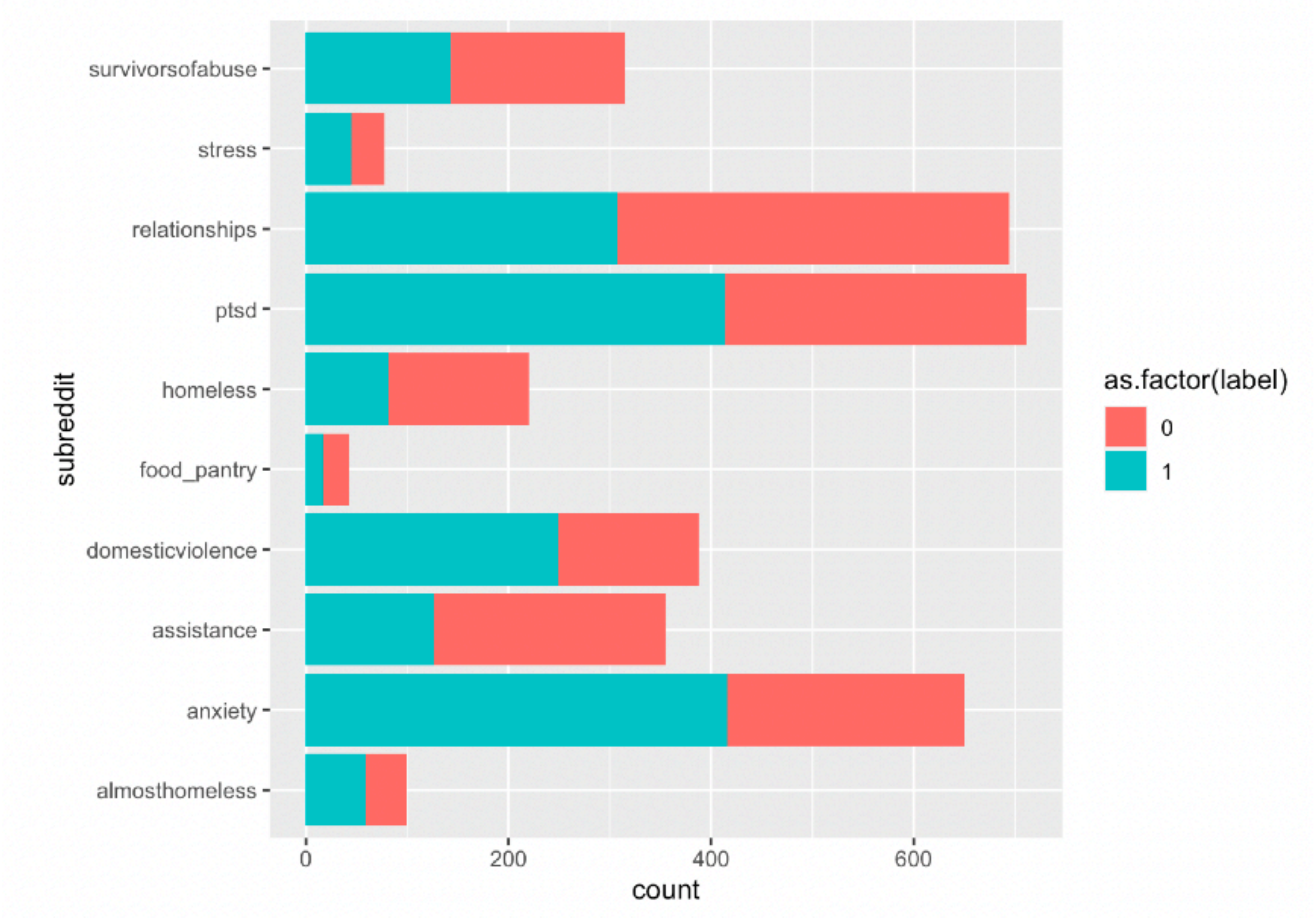



survivorsofabuse
stress
relationships
ptsd
homeless
food_pantry
domesticviolence
assistance
anxiety
almosthomeless
subreddit
0
200
400
600
count
as.factor(label)
0
1


700
600
500
400
300
200
100
0
n
(almosthomeless, 99)
almosthomeless
anxiety
assistance
domesticviolence
food_pantry
homeless
ptsd
relationships
stress
survivorsofabuse
subreddit

Dataset Overview

- Dataset consists of training data with 2838 posts, test data with 715 posts
- Sampled from 11 subreddits



Special Issues with Text-Related Data

- Text data is inherently messy.
- Firstly, computers can only process numbers, meaning that any text data needs some way to be converted to numbers.
- Some words that commonly appear in text have very little influence on the text. We call these **stop words**. (Such as “and”, “so”, “to”, etc.)
- A computer treats “cold”, “Cold”, “cold;” as separate objects, even though they are the same word.
- Additionally, there are rare words that may not appear very often but may have an influence on the text data.