Assignment 2

Group 2

2025-03-14

We will explore the data we scraped to determine next steps for analysis examining both posts and comments related to these posts.

1. Read in data from github, and tidy it to continue processing.

```r
# path to repo
#repo_path <- str_c(~/Desktop/UMICH/SURV622_Assignment/data/") # felix
repo_path <- str_c("~/repos/SURV622_Assignment/data/") # kevin


data_files <- list.files(repo_path) |>
  tibble() |>
  rename(data_file_names = 1) |>
  filter(str_detect(data_file_names, ".xlsx|.RDS")) |>
  mutate(data_names = str_remove_all(data_file_names, "\\..*"))


list2env(
  map(data_files$data_file_names, function(x){

  # read data file
  if(str_detect(x, ".xlsx")) {
    dat = read_xlsx(str_c(repo_path, x))
  } else(
    dat = readRDS(str_c(repo_path, x))
  )

  # manipualte variables
  dat = dat |>
```

```
     mutate(date_utc = ymd(date_utc),
    hour_posted = hour(as_datetime(timestamp)),
  )

}) |>
  # name the files
    set_names(data_files$data_names),

# save data to the global environment
globalenv()
)
```

```
<environment: R_GlobalEnv>
```

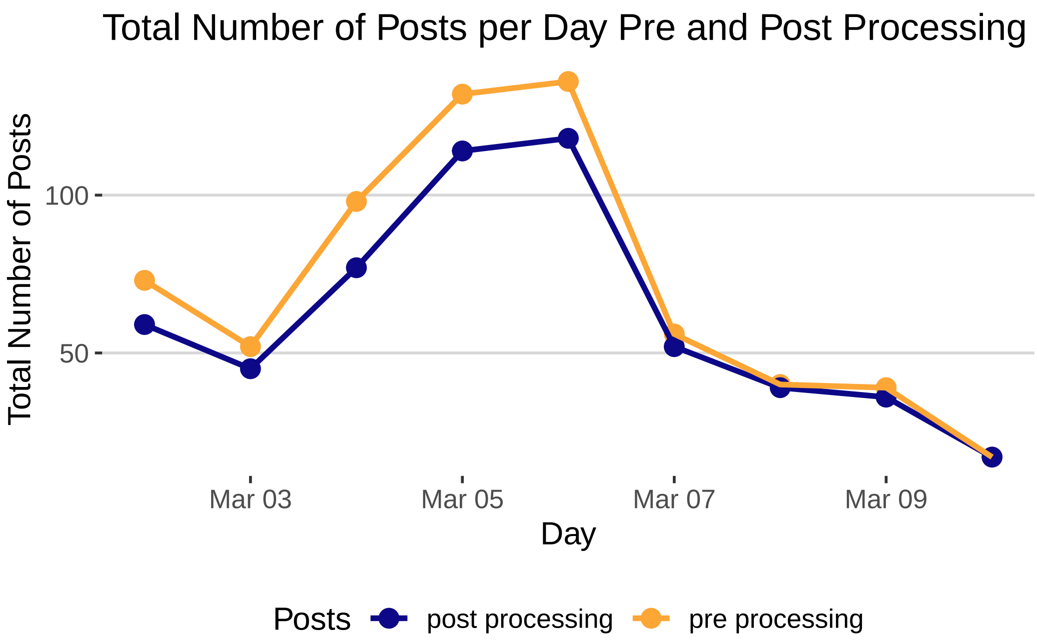
How many posts were collected in total and by day before and after processing/cleaning?

- There appears to be a noticeable peak in activity during the middle of the week. However, we don't have enough data to conclusively say it's due to the day of the week itself. It's more likely that these spikes are linked to specific events that occurred on those days, which sparked conversations.
- The two peaks in our graph could be attributed to two major news stories. On March 5, Elon Musk made a statement about wanting to "save Western civilization from empathy," and on March 6, there were reports that President Trump was limiting Musk's authority due to backlash over cuts to DOGE. These events likely had a significant impact on the volume of Reddit posts related to Elon Musk and DOGE, driving the observed spikes in activity.
- On May third, news outlets started reporting that DOGE was claiming $105 billion dollars in savings from cutting "wasteful" spending by layoffs and cutting foreign aid. This amount is controversial since since the receipts shown on their site only amount to less than $9.6 billion https://abcnews.go.com/US/doge-website-now-saved-105-billion-backtracked-earlier/story?id=119408347. This could have prompted reddit users to take to subreddits and provide their opinion or thoughts on this matter.

```
# number of total posts
total_posts <- posts_data |>
  filter(date_utc > "2025-03-01") |>
  group_by(date_utc) |>
  reframe(total = n()) |>
  mutate(Posts = "pre processing") |>
  add_row(
    posts_data_clean |>
  group_by(date_utc) |>
  reframe(total = n()) |>
  mutate(Posts = "post processing")
  )

total_posts |>
  group_by(Posts) |>
  reframe(Total = sum(total)) |>
  kable()
```

Posts	Total
post processing	557
pre processing	643

```
total_posts |>
  ggplot( aes(x = date_utc, y = total, color= Posts)) +
  geom_point(size = 3) +
  geom_line(linewidth = 1) +
  labs(
    title = "Total Number of Posts per Day Pre and Post Processing",
    x = "Day",
    y = "Total Number of Posts") +
  scale_color_viridis_d(option="C", end=.8) +
  theme_hc()
```



How many comments and by day?

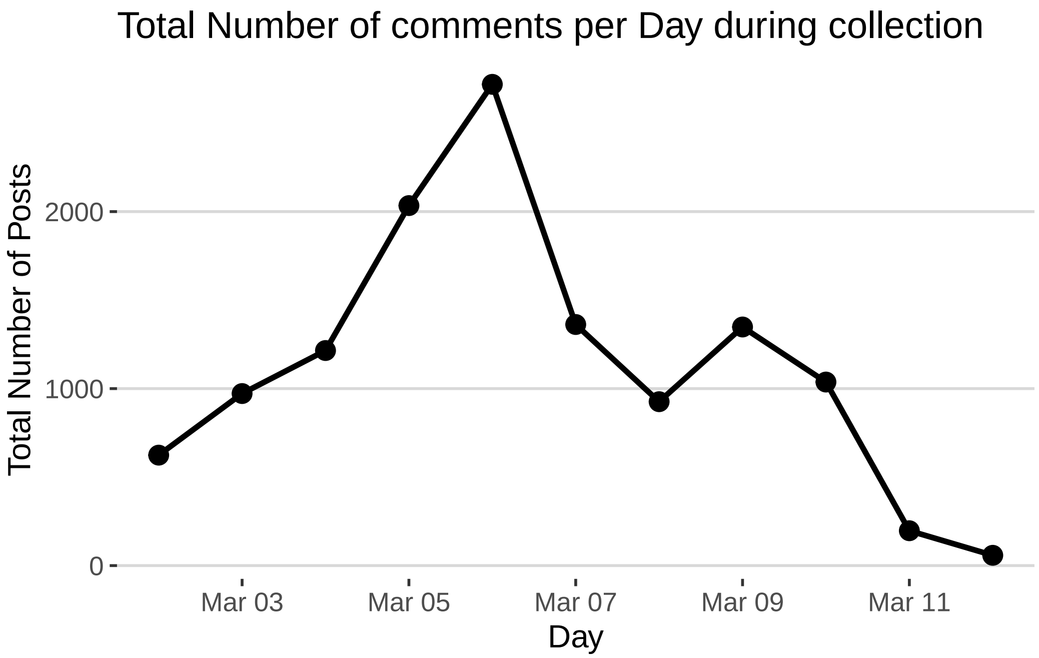
- We see a peak in the comments during March 6th, which was a Thursday, and perhaps in anticipation of DOGE policies that often are released on Friday morning and have come to be known as “Musk-acre Friday” (https://smotus.substack.com/p/friday-night-musk-acre).

```
# number of comments
comments_data_clean |> nrow() |> kable(col.names = "Total Comments")
```

Total Comments
12492

```
comments_data_clean |>
  group_by(date_utc) |>
  reframe(total = n()) |>
  ggplot( aes(x = date_utc, y = total)) +
  geom_point(size = 3) +
```

```r
  geom_line(linewidth = 1) +
  labs(
    title = "Total Number of comments per Day during collection",
    x = "Day",
    y = "Total Number of Posts") +
  scale_color_viridis_d(option="A") +
  theme_hc()
```



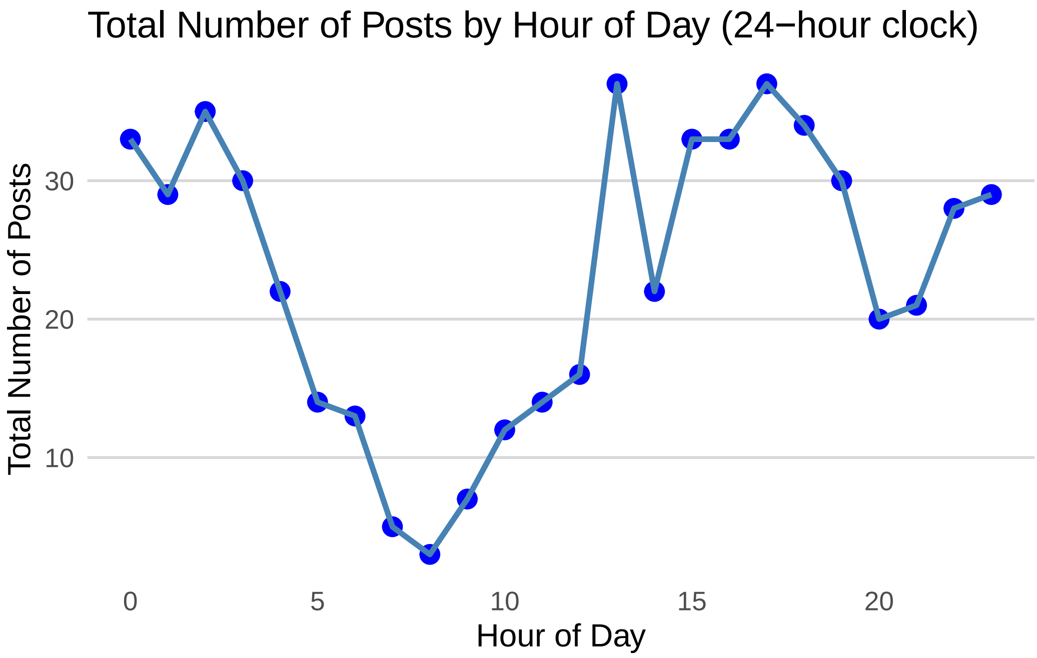
Is there a pattern with respect to the time of day or day of week when posts were created?

Is there a relationship between events and frequency of reddit posts?

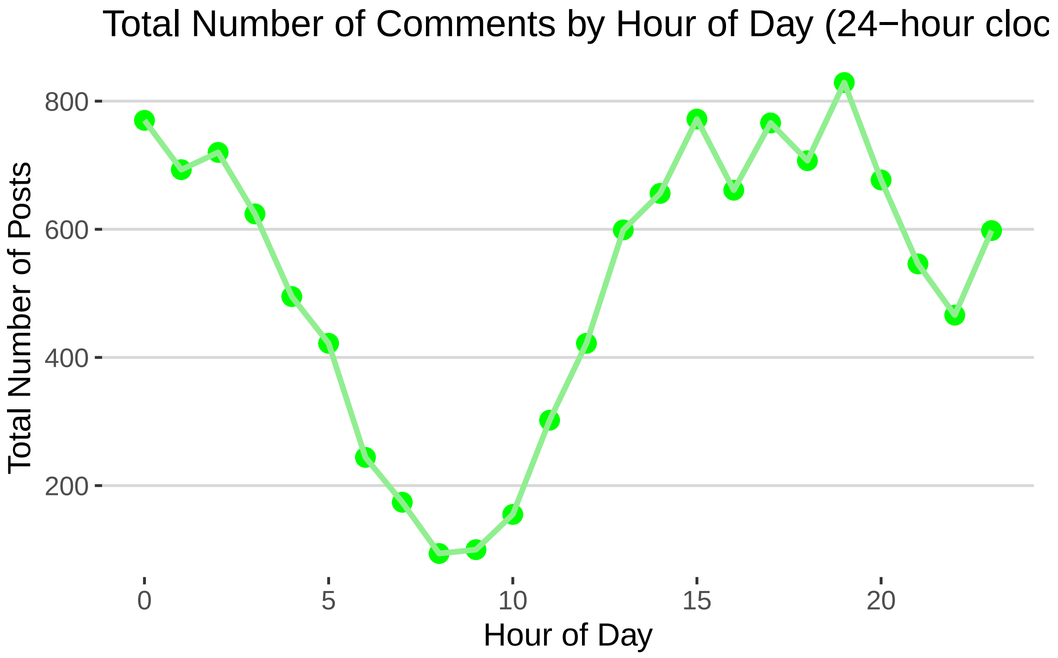
- Most posts occur during the early morning, noon, and afternoon hours. The highest peak is around lunchtime, likely due to people posting during their lunch break. There is also significant activity in the afternoon, evening, and early morning. This could be because people have more free time or stay up late, allowing them to take the time to write posts, which require more effort than quick interactions. The fewest posts are seen in the morning, which makes sense as people are typically getting ready for work, commuting, or sleeping in.

- The trend for comments mirrors that of posts. This makes sense because people are likely using Reddit at similar times for both posting and commenting. However, the volume of comments is much higher than the number of posts, as each post receives many comments from different users.

```
posts_data_clean |>
  group_by(hour_posted ) |>
  reframe(Total = n()) |>
  ggplot(aes(x=hour_posted, y=Total)) +
  # geom_bar(stat = "identity", fill = "steelblue") +
  geom_point(color = "blue", size = 3) +
  geom_line(color = "steelblue", linewidth = 1) +
  labs(
    title = "Total Number of Posts by Hour of Day (24-hour clock)",
    x = "Hour of Day",
    y = "Total Number of Posts"
  ) +
  theme_minimal() +
  theme_hc()
```



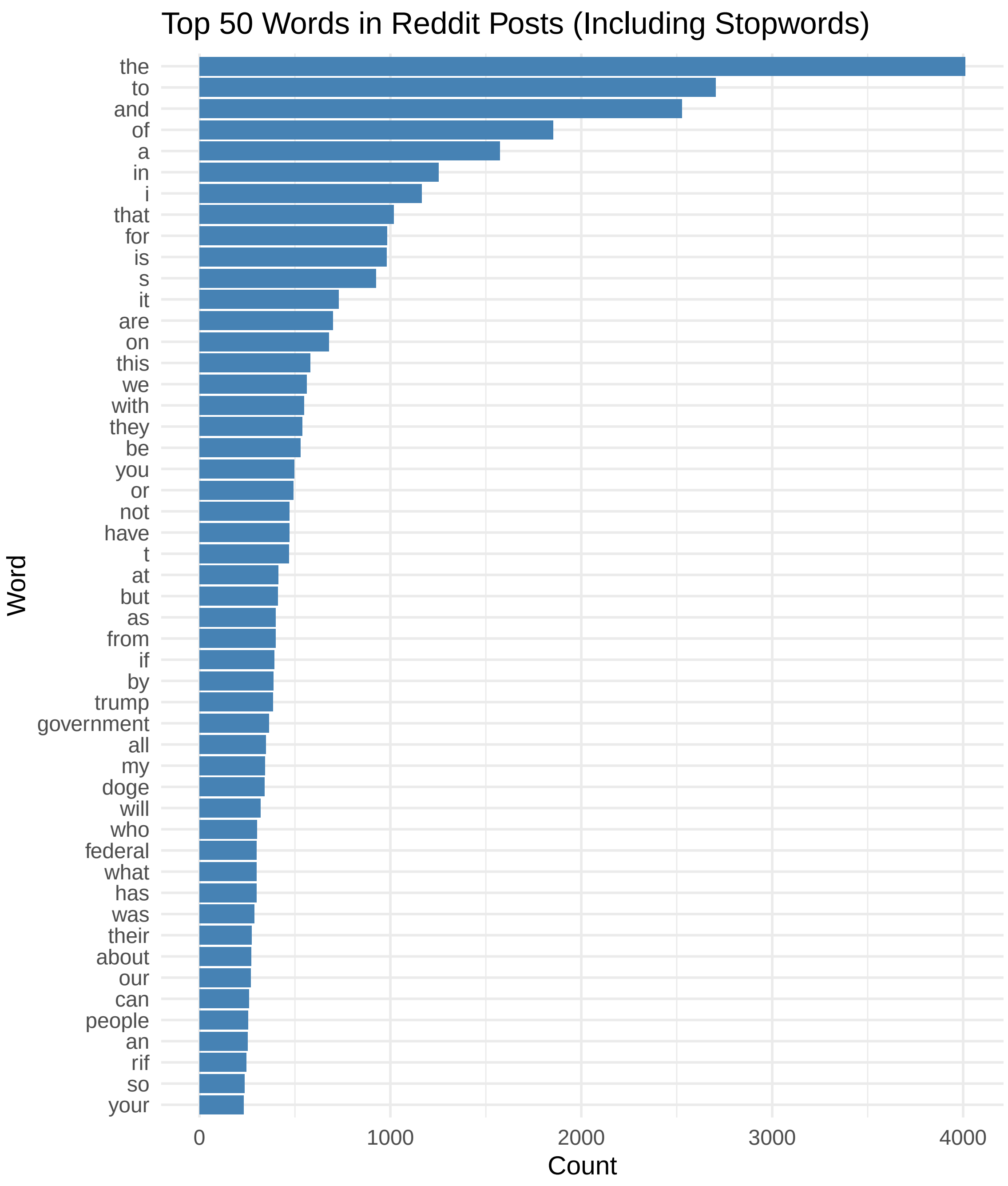
```
comments_data_clean |>
  group_by(hour_posted ) |>
  reframe(Total = n()) |>
  ggplot(aes(x=hour_posted, y=Total)) +
  # geom_bar(stat = "identity", fill = "steelblue") +
  geom_point(color = "green", size = 3) +
  geom_line(color = "lightgreen", linewidth = 1) +
  labs(
    title = "Total Number of Comments by Hour of Day (24-hour clock)",
    x = "Hour of Day",
    y = "Total Number of Posts"
  ) +
  theme_hc()
```



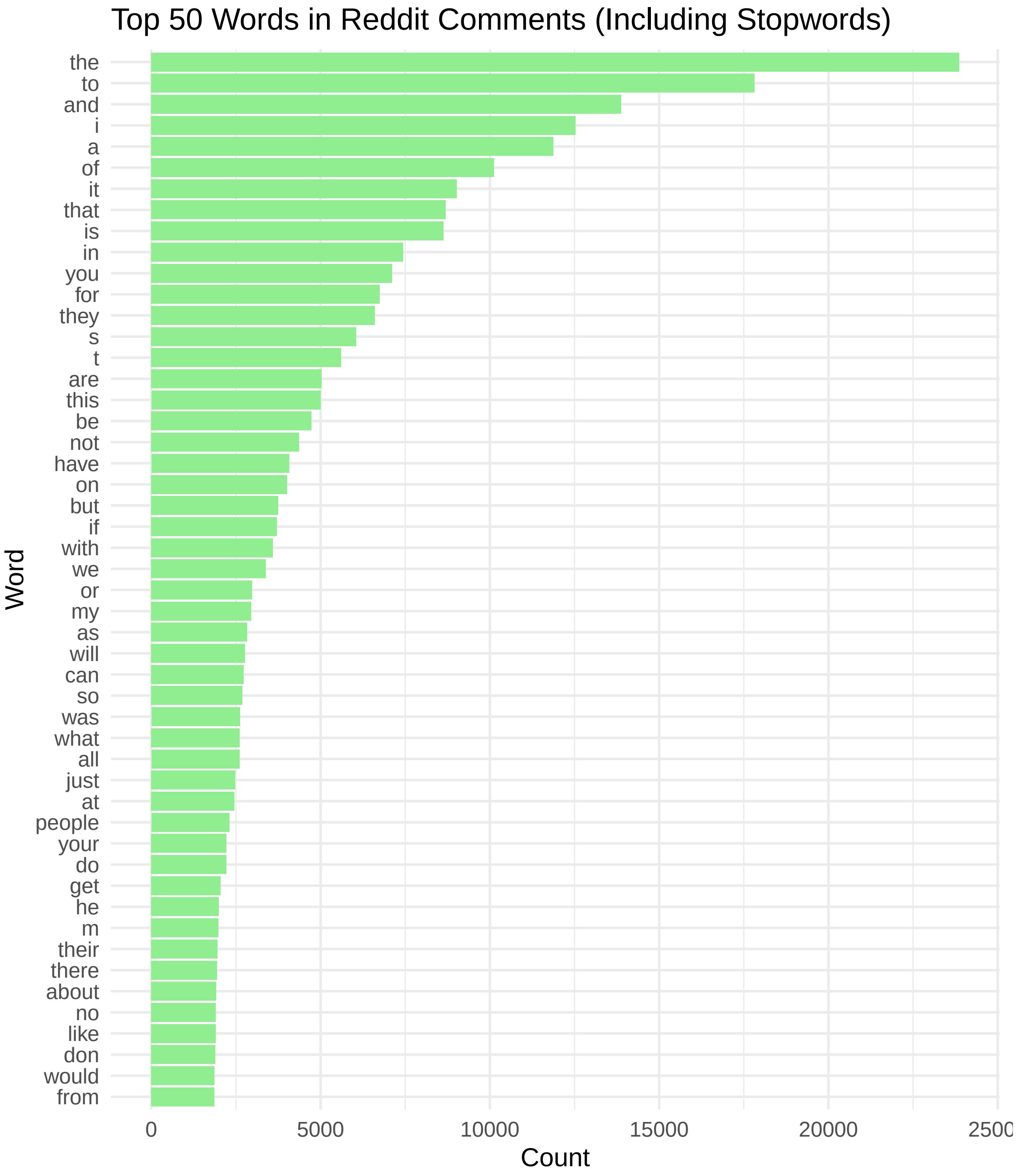
What are the words in the set of posts you have assembled that appear most frequently?

```
# tokenize words to count frequencies for reddit posts
posts_word_counts <- posts_data_clean |>
  select(title, text) |>
  pivot_longer(cols = c(title, text), values_drop_na = TRUE) |>
  unnest_tokens(word, value) |>
```

```r
  count(word, sort = TRUE) |>
  slice_max(n, n = 50)

# plot
ggplot(posts_word_counts, aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 50 Words in Reddit Posts (Including Stopwords)",
       x = "Word",
       y = "Count") +
  theme_minimal()
```



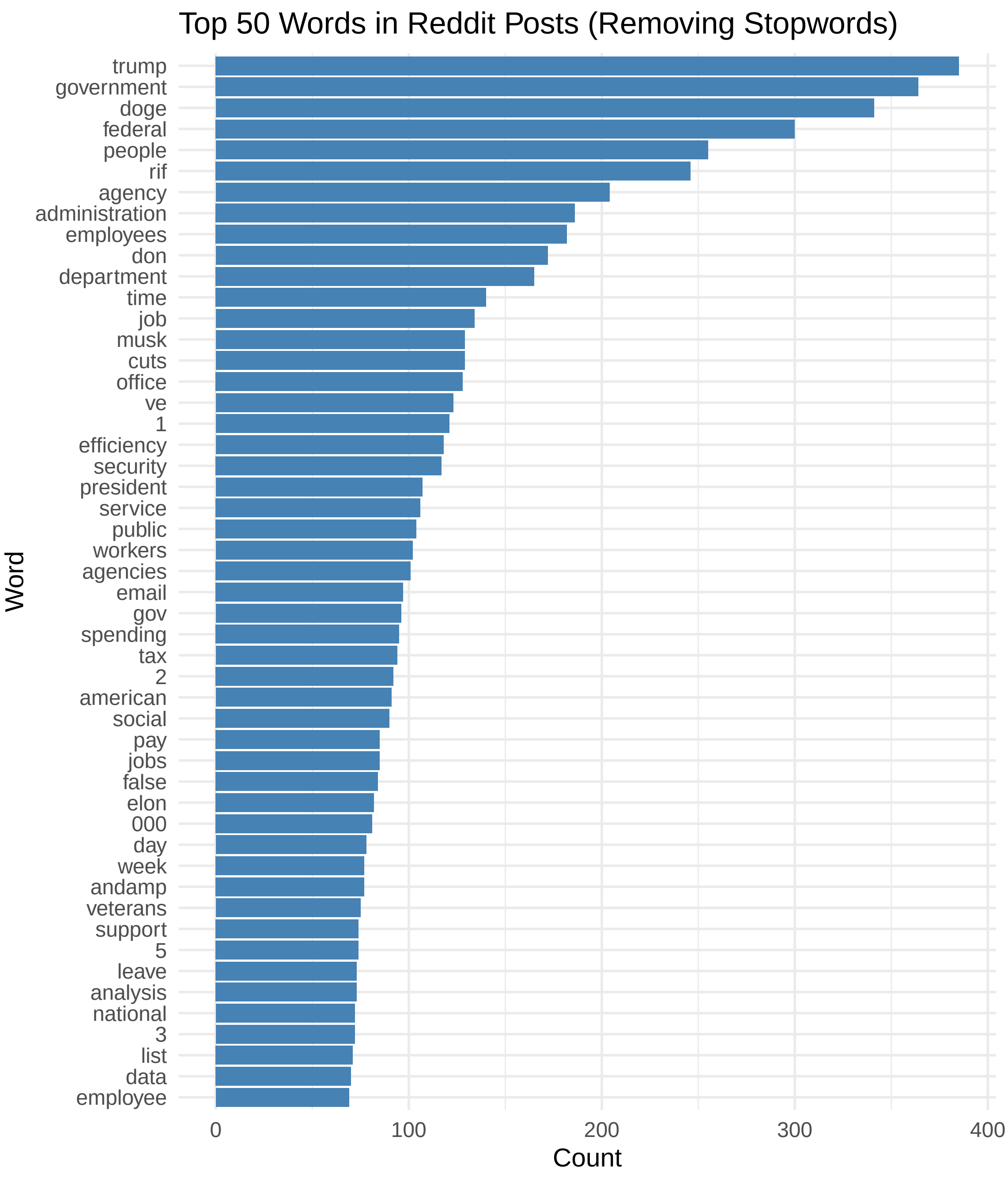
```
# tokenize words to count frequencies for reddit comments
comments_word_counts <- comments_data_clean |>
  select(comment) |>
  pivot_longer(cols = c(comment), values_drop_na = TRUE) |>
  unnest_tokens(word, value) |>
  count(word, sort = TRUE) |>
  slice_max(n, n = 50)

# plot
ggplot(comments_word_counts, aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "lightgreen") +  # Removed extra parentheses
  coord_flip() +
  labs(title = "Top 50 Words in Reddit Comments (Including Stopwords)",
       x = "Word",
       y = "Count") +
  theme_minimal()
```



The graphs above show the most frequent words in both Reddit posts and their comments. However, many of these words are stopwords, which don't provide much insight into the actual content of the discussions. To get a better understanding of what people are really talking about, we will remove these stopwords in the next step.

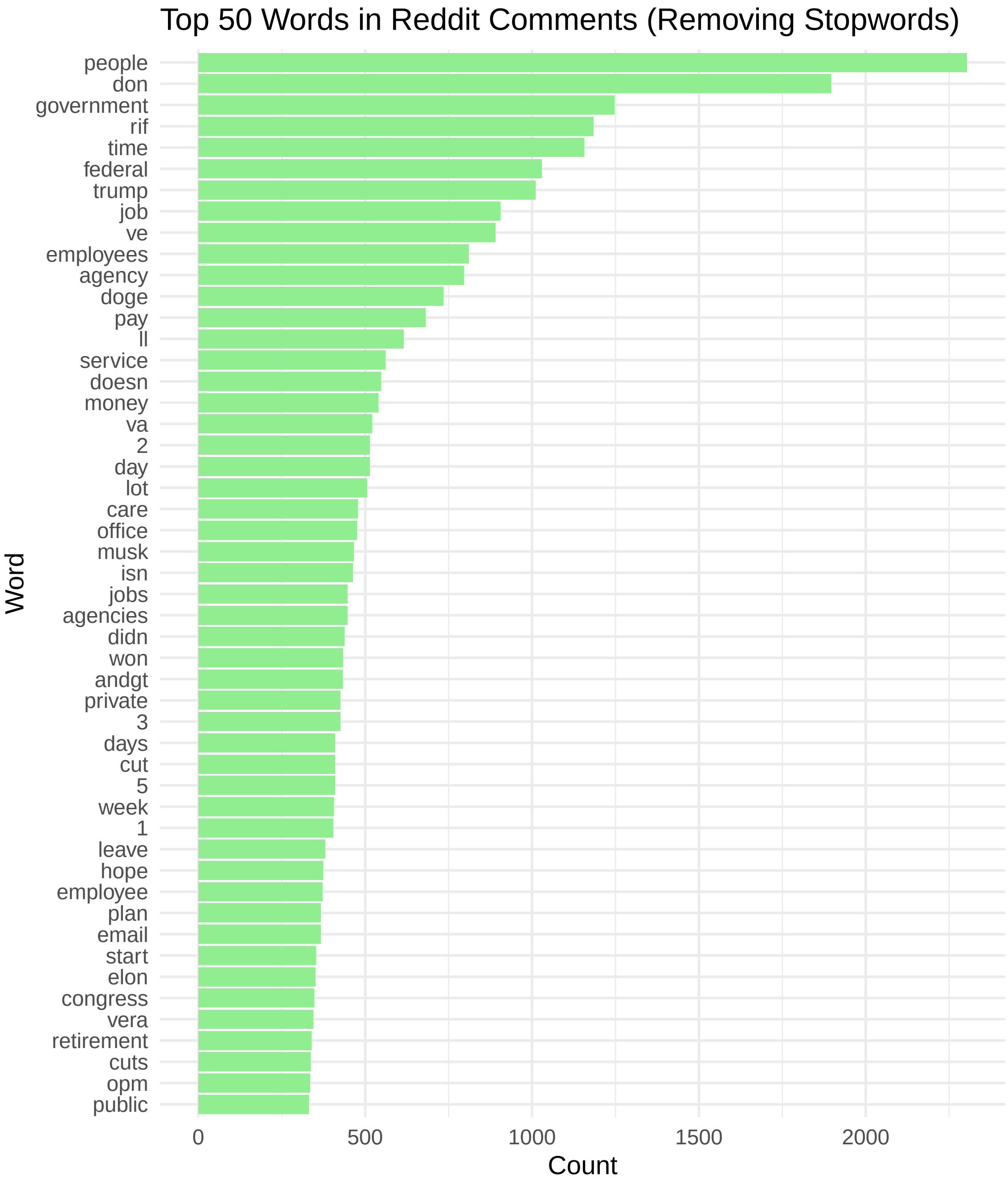
How does this change if you exclude “stop words” such as “a,” “an,” “the,” “is,” and others that are common in English sentences but are generally not informative?

```r
# tokenize words to count frequencies for reddit posts removing stop words
posts_word_counts <- posts_data_clean |>
  select(title, text) |>
  pivot_longer(cols = c(title, text), values_drop_na = TRUE) |>
  unnest_tokens(word, value) |>
  anti_join(stop_words, by = "word") |>
  count(word, sort = TRUE) |>
  slice_max(n, n = 50)

# plot
ggplot(posts_word_counts, aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 50 Words in Reddit Posts (Removing Stopwords)",
       x = "Word",
       y = "Count") +
  theme_minimal()
```



```
# tokenize words to count frequencies for Reddit comments, removing stop words
comments_word_counts <- comments_data_clean |>
  select(comment) |>
  pivot_longer(cols = c(comment), values_drop_na = TRUE) |>
```

```r
  unnest_tokens(word, value) |>
  anti_join(stop_words, by = "word") |>
  count(word, sort = TRUE) |>
  slice_max(order_by = n, n = 50)

# plot
ggplot(comments_word_counts, aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "lightgreen") +  # Removed extra parentheses
  coord_flip() +
  labs(title = "Top 50 Words in Reddit Comments (Removing Stopwords)",
       x = "Word",
       y = "Count") +
  theme_minimal()
```



The graphs above show the most frequent words used in both Reddit posts and their comment sections. We can see that many of these words are directly related to our topic, with frequent mentions of the President, DOGE, the federal government, and RIF.