Assignment 2

Group 2

2025-03-12

We will explore the data we scraped to determine next steps for analysis.

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

```
# posts data
reddit_posts <- read_xlsx(
  "~/repos/SURV622_Assignment/data/posts_data_clean.xlsx") |>
  mutate(date_utc = ymd(date_utc)) |>
  # remove variables not needed
  select(-timestamp)


# comments data
reddit_comment <- read_xlsx(
  "~/repos/SURV622_Assignment/data/comments_data_clean.xlsx") |>
    mutate(date_utc = ymd(date_utc)) |>
  # great unique post id
  arrange(date_utc) |>
  group_by(url) |>
  mutate(post_id = cur_group_id()) |>
  ungroup() |>
  # remove variables not needed
  select(-author, -timestamp, -comment_id)


# id url group IDs to the posts data
reddit_posts_2 <- reddit_posts |>
  semi_join(reddit_comment |> select(url, post_id))
```

The strings we will be processing sometimes have special characters, numbers, brackets. We need to clean these string for processing.

- We can use the fedmatch::clean_strings() function to tidy up the text. The following is an example of what will be cleaned.

```
# print example of posts
reddit_posts |>
  slice(10:15) |>
  select(text)
```

```
# A tibble: 6 x 1
  text
  <chr>
1 "Join us tonight from anywhere for 50501 DC\u0019s Virtual Event =4 LIVE at 8~
2 "From Ariella Elm [Home | Substack]( I was on the [womenforward.us]( Your Ass~
3 "Us fired feds/contractors have been going to the DC Congressional offices an~
4 "I want to preface this post by stating clearly that revolution does not mean~
5 "\u0018He needs to resign&\u0019 Obviously another Russian asset practicing D~
6 "Trump communicates publicly that Elon has free reign in DOGE to ferret out \~
```

```
# clean posts
reddit_posts |>
  slice(10:15) |>
  mutate(text = clean_strings(text)) |>
  select(text)
```

```
# A tibble: 6 x 1
  text
  <chr>
1 join us tonight from anywhere for 50501 dc s virtual event 4 live at 8 30pm e~
2 from ariella elm home substack i was on the womenforward us your assignment c~
3 us fired feds contractors have been going to the dc congressional offices and~
4 i want to preface this post by stating clearly that revolution does not mean ~
5 he needs to resignand obviously another russian asset practicing doge cancel ~
6 trump communicates publicly that elon has free reign in doge to ferret out wa~
```

- Clean posts and comment strings.

```
reddit_posts <- reddit_posts |>
  mutate(text = clean_strings(text))

reddit_comment <- reddit_comment |>
  mutate(comment = clean_strings(comment))

reddit_posts_2 <- reddit_posts_2 |>
  mutate(text=clean_strings(text))
```

We can now begin to explore the data. First we explore the posts.

```
  reddit_posts |>
  slice(364) |>
  pull(text)
```

[1] "the office of personnel management is giving an ultimatum to remote and teleworking employees who are more than 50 miles away from their official duty stations opm is directing employees in this scenario to either report to their current duty station agree to a management directed reassignment and relocate to office space in another geographic region or accept termination from their jobs"