SAL 413 HW2

Hunter Geise

2023-10-04

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
theme_set(theme_bw())
library(purrr)
library(stringr)
library(ggTimeSeries)
library(glue)
```

Question 1: a. Download the tweets.json file and load it into R as a named list, woj_tweets, using the function jsonlite::read_json(). Print Woj's most recent tweet. Do not include the metadata, just the tweet itself.

```
#a#
woj_tweets <- jsonlite::read_json("tweets.json")
print(woj_tweets[[1]]$text)
```

[1] "Philadelphia 76ers coach Doc Rivers joins The Woj Pod to discuss the Joel Embiid/James Harden partnership, shrinkin... https://t.co/ZW7udhETF1"

b. Twitter users can mention users or direct tweets to users using the @ symbol directly followed by a username following these rules. Determine the top 20 users Woj mentioned in the data set.

```
#b#
woj_tweets %>% str_extract_all("\\@\\w+") -> usernames_list
```

```
## Warning in stri_extract_all_regex(string, pattern, simplify = simplify, :
## argument is not an atomic vector; coercing
```

```
usernames_list %>% unlist() -> usernames
usernames %>% as.data.frame() -> username_df
colnames(username_df) <- c("Username")
username_count <- as.data.frame(table(username_df$Username))
ordered_username <- username_count[order(-username_count$Freq), ]
final_order <- head(ordered_username, 20)
final_order
```

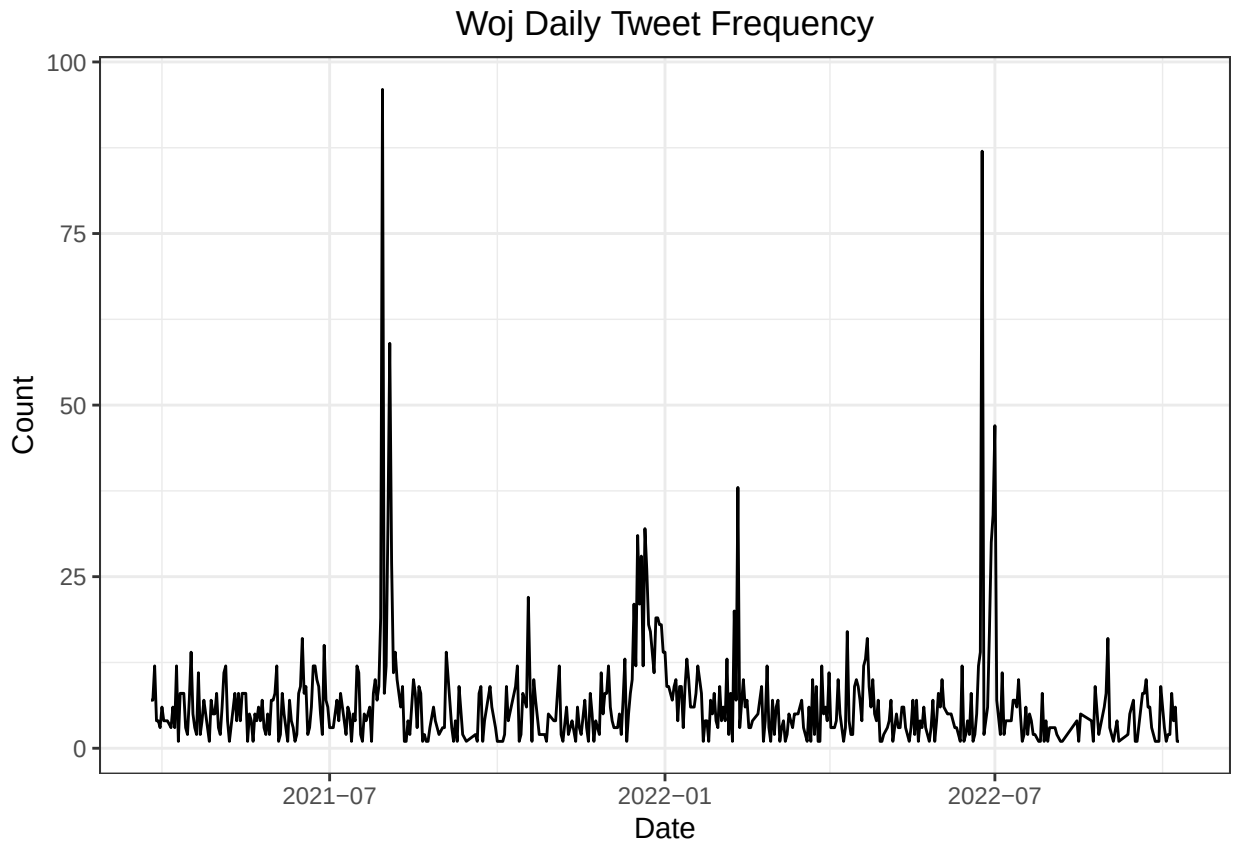
```
##                Var1 Freq
## 54     @DraftExpress   93
## 62    @espn_macmahon   82
## 220     @TimBontemps   67
## 241         @wojespn   66
## 18     @BobbyMarks42   53
## 134  @malika_andrews   51
## 142           @mcten   51
## 147    @Mike_Schmitz   47
## 14           @Baxter   30
## 182 @ramonashelburne   29
## 1     @_Andrew_Lopez   24
## 243    @ZachLowe_NBA   23
## 161    @NotoriousOHM   21
## 61             @ESPN   20
## 239   @WindhorstESPN   15
## 4      @AdamSchefter   14
## 177  @PrioritySports   14
## 60             @espn   11
## 94     @JamalCollier   11
## 101      @JeffPassan   11
```

c. Make a line chart that shows the number of tweets Woj wrote daily in the data set. Hint: Unless you want to go nuts with string processing, let lubridate do the heavy lifting. You'll need to do some, but note the following behavior of lubridate's mdy() function: Once you compute the vector of dates, put them in a single-column tibble with variable named date. Then pipe the tibble into dplyr's count(), see its documentation to learn the syntax you need. You should know where to go from there. I achieved this result in 10 lines while following the tidyverse style guide.

```
#c#
woj_tweets %>% str_extract_all("\\d{4}-\\d{2}-\\d{2}") -> woj_dates
```

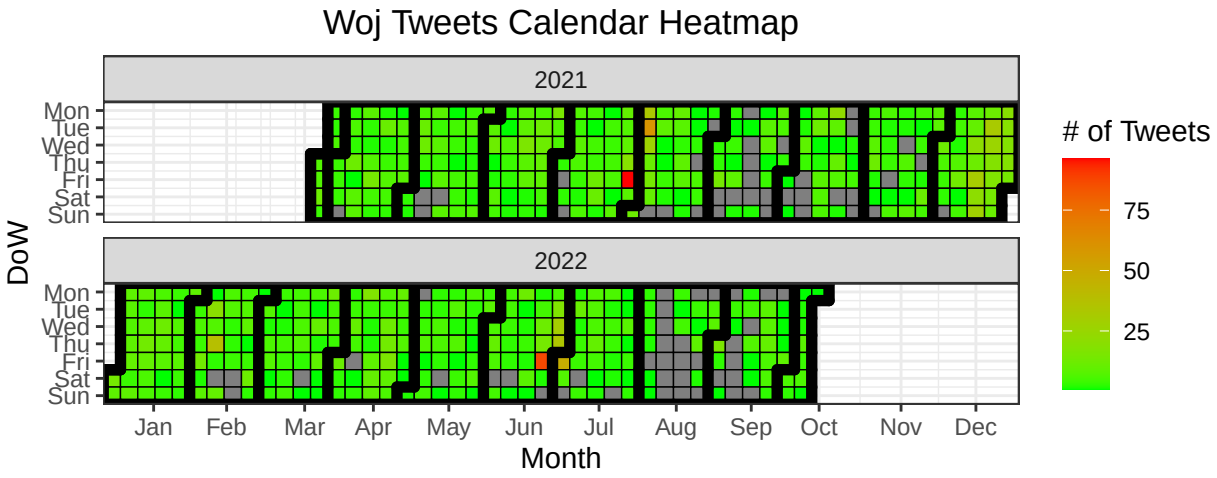
```
## Warning in stri_extract_all_regex(string, pattern, simplify = simplify, :
## argument is not an atomic vector; coercing
```

```
woj_dates %>% unlist() -> woj_dates
new_woj_dates <- lubridate::ymd(woj_dates)
new_woj_dates %>% tibble(Date = new_woj_dates) -> woj_tibble
woj_tibble %>% count(Date) -> date_count
ggplot(date_count, aes(x = Date, y = n, group = 1)) +
  geom_line() +
  labs(title = "Woj Daily Tweet Frequency") +
  ylab("Count") +
  theme(plot.title = element_text(hjust = 0.5))
```



d. Using the ggTimeSeries package, make a calendar heatmap of the tweets. (See the page to learn what that is.)

```
#d#
ggplot_calendar_heatmap(dtDateValue = date_count, "Date", "n") +
  labs(title = "Woj Tweets Calendar Heatmap") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_continuous(low = 'green', high = 'red', name = "# of Tweets") +
  facet_wrap(~Year, ncol = 1)
```



e. What were Woj's top ten most popular tweets by like (favorite) count? Print out just the tweets (no meta data) in the format "1. [tweet] ([# votes])" with new lines between each, e.g. 1. blah blah blah (1234). Use string processing to format it cleanly on the output page. You must make use of the glue package for this question.

```
#e#
woj_tweets[order(map_int(
  woj_tweets, function(i) -i$favorite_count))] -> woj_favorite_tweets

top_tweets <- head(woj_favorite_tweets, 10)

final_top_10 <- map_chr(seq_along(top_tweets), function(x) {
  glue("{x}. {top_tweets[[x]]$text} ({top_tweets[[x]]$favorite_count} Likes)")
  })

cat(final_top_10, sep = "\n")
```

1. The Brooklyn Nets are trading James Harden to the Philadelphia 76ers for Ben Simmons, Seth Curry, Andre Drummond an... https://t.co/MIsfZUGW0a (255474 Likes)
2. The Cleveland Cavaliers have acquired Donovan Mitchell in a trade, sources tell ESPN. (165296 Likes)
3. The Lakers and Wizards have agreed on the trade for Russell Westbrook, sources tell ESPN. (132729 Likes)
4. ESPN Sources: Happy New Year. (132303 Likes)
5. Free agent F Carmelo Anthony has agreed to a one-year deal with the Los Angeles Lakers, his manager Bay Frazier tells ESPN. (122931 Likes)
6. The Lakers are near a deal to acquire Washington's Russell Westbrook for Kyle Kuzma, Montrezl Harrell, Kentavious C... https://t.co/Pa0zsgiEPb (108502 Likes)
7. Memphis Grizzlies 152 Oklahoma City Thunder 79: Biggest margin of victory in NBA history. (105681 Likes)
8. Derek Chauvin verdict in Minneapolis: Guilty on all three counts in the death of George Floyd. (101413 Likes)
9. Utah is trading Rudy Gobert to Minnesota, sources tell ESPN. (96663 Likes)
10. Denver Nuggets center Nikola Jokic has been voted the NBA's Most Valuable Player for a second consecutive season, s... https://t.co/5aqvTaDFUh (79770 Likes)

f. Each tweet comes with the same fields, source, id_str, etc. Convert your woj_tweets list into a tibble with those columns. Where necessary, make appropriate changes.

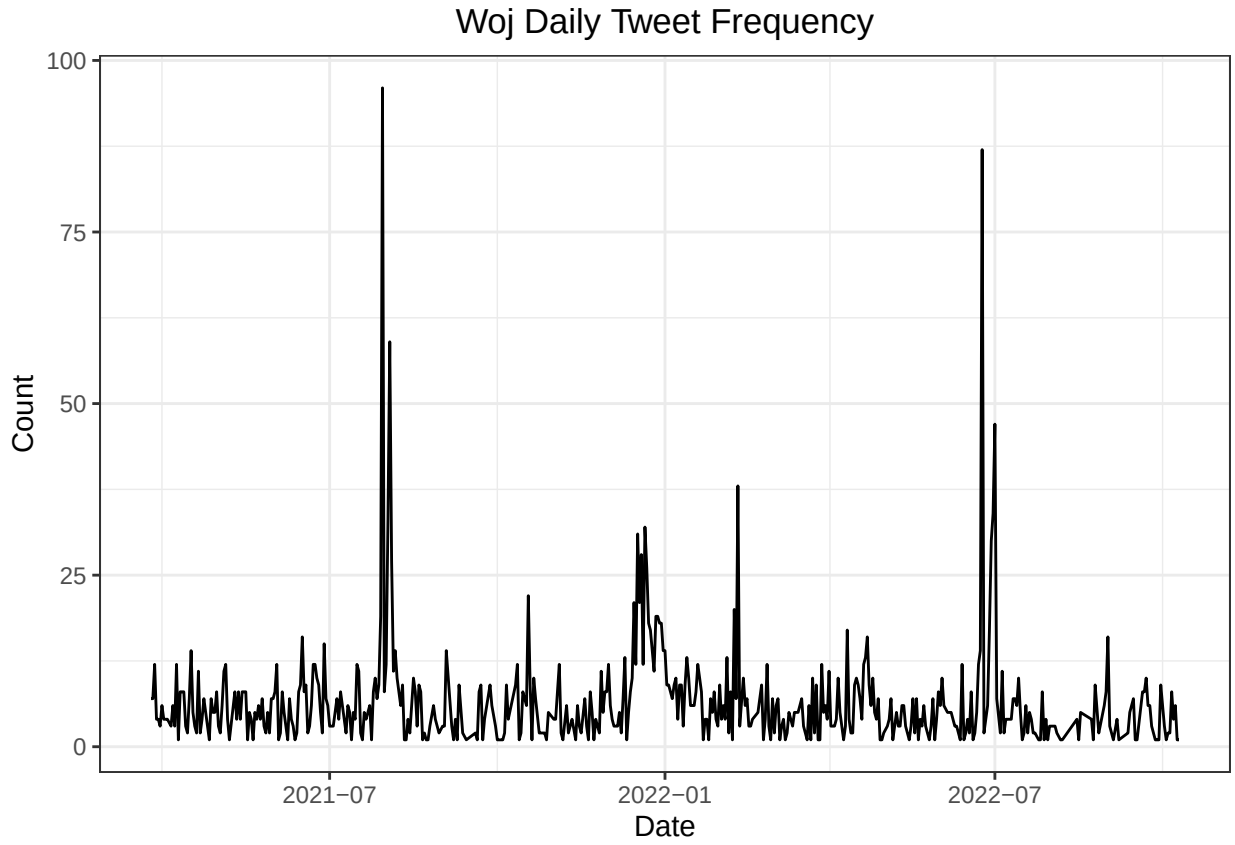
```
#f#
whole_tibble <- function(list) {
  tibble(
    Created_At = list$created_at,
    Favorite_Count = list$favorite_count,
    Followers_Count = list$follwers_count,
    ID_Str = list$id_str,
    In_Reply_To_Screen_Name = list$in_reply_to_screen_name,
    Retweet_Count = list$retweet_count,
    Screen_Name = list$screen_name,
    Text = list$text
  )
}

whole_woj_tibble <- map_dfr(woj_tweets, whole_tibble)
whole_woj_tibble
```

```
## # A tibble: 3,249 x 7
##    Created_At          Favorite_Count ID_Str     Retweet_Count Screen_Name Text
##    <chr>                        <int> <chr>              <int> <chr>       <chr>
##  1 2022-10-10 12:42:24            914 157945228~            74 wojespn     "Phi~
##  2 2022-10-09 15:24:40           7305 157913073~           640 wojespn     "The~
##  3 2022-10-08 18:12:51              0 157881067~           710 wojespn     "RT ~
##  4 2022-10-08 18:03:13          16926 157880824~          1036 wojespn     "At ~
##  5 2022-10-08 12:56:50            929 157873114~            61 wojespn     "ESP~
##  6 2022-10-08 12:34:47            906 157872559~            46 wojespn     "Sug~
##  7 2022-10-08 12:29:48           5634 157872434~           457 wojespn     "Orl~
##  8 2022-10-08 01:55:08              0 157856462~           121 wojespn     "RT ~
##  9 2022-10-07 19:33:31              0 157846858~          2816 wojespn     "RT ~
## 10 2022-10-07 13:24:37           5958 157837574~           409 wojespn     "Con~
## # i 3,239 more rows
## # i 1 more variable: In_Reply_To_Screen_Name <chr>
```

g. Bonus. Using any technique you like, Text Mining with R may help and is available to you, do something interesting with this dataset that uses your tidyverse skills.

```
#g#
ggplot(date_count, aes(x = Date, y = n, group = 1)) +
  geom_line() +
  labs(title = "Woj Daily Tweet Frequency") +
  ylab("Count") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
date_count[order(-date_count$n), ]
```

```
## # A tibble: 490 x 2
##    Date           n
##    <date>     <int>
##  1 2021-07-30    96
##  2 2022-06-24    87
##  3 2021-08-03    59
##  4 2022-07-01    47
##  5 2022-02-10    38
##  6 2022-06-30    34
##  7 2021-08-02    33
##  8 2021-12-21    32
##  9 2021-12-17    31
## 10 2022-06-29    30
## # i 480 more rows
```

Looking back at part b, it was clear that two periods where Woj's tweets were far more frequent than others. So, I ordered the final data set from part b to find what the dates were around July of 2022 and August of 2021. After doing research online, I discovered that the major influx in tweets in both periods were caused by the free agency periods and NBA drafts. It was interesting though after looking through the JSON file that I discovered the 2022 NBA Draft was on June 23rd, but all his tweets were sent out on the 24th. The same instance happened in the 2021 NBA Draft where the draft happened on July 29th, but all of Woj's tweets about the draft were on the 30th. I don't know if this was caused by a different time zone or if it was something involving the JSON file, but something is up there.