MECH476: Engineering Data Analysis in R

Chapter 6 Homework: Strings, Dates, and Tidying

Flynn Nyman

02 October, 2024

Chapter 6 Homework

For this homework assignment, you will use data from Twitter that include tweets (2011 to 2017) from Colorado senators, which can be downloaded from Canvas. Just FYI—some tweets were cut off before Twitter's character limit; just work with the data you have. The original data are from FiveThirtyEight.

When a question asks you to make a plot, remember to set a theme, title, subtitle, labels, colors, etc. It is up to you how to personalize your plots, but put in some effort and think about making the plotting approach consistent throughout the document. For example, you could use the same theme for all plots. I also like to use the subtitle as a place for the main summary for the viewer.

Question 1: Hashtags

Within a pipeline using the Colorado-only tweet data, select `text` variable and use `stringr::str_extract_all()` with a pattern of `"#(\\d|\\w)+"` to extract all of the hashtags from the tweets. This will return a list with one element. How many hashtags were used by Colorado senators?

```
## [1] 2569
```

Question 2: Fires

Colorado is on fire right now and has experienced many wildfires over the years. Let's examine senators' tweet activity related to wildfires based on hashtags. Using the character vector of hashtags you extracted in Question 1, search for the hashtags that include "fire" or "wildfire". How many hashtags included "fire"? How many included "wildfire"?

```
## [1] 16
```

```
## [1] 8
```

Question 3: Wildfires

Now, let's look at general tweets concerning wildfires. First, subset the data to a dataframe that includes tweets containing the word "wildfire" and their corresponding timestamp and user. Specifically, (a) select `text`, `date`, and `user` and (b) filter to text strings that include the word "wildfire" using `dplyr::filter()` and `stringr::str_detect()`.

Question 4: Senators

Which Colorado senator tweets more about wildfires?

```
## # A tibble: 2 x 2
##   user               n
##   <chr>          <int>
## 1 SenBennetCO       20
## 2 SenCoryGardner    13
```

Senator Bennet tweets more about wildfires. ## Question 5: Timing

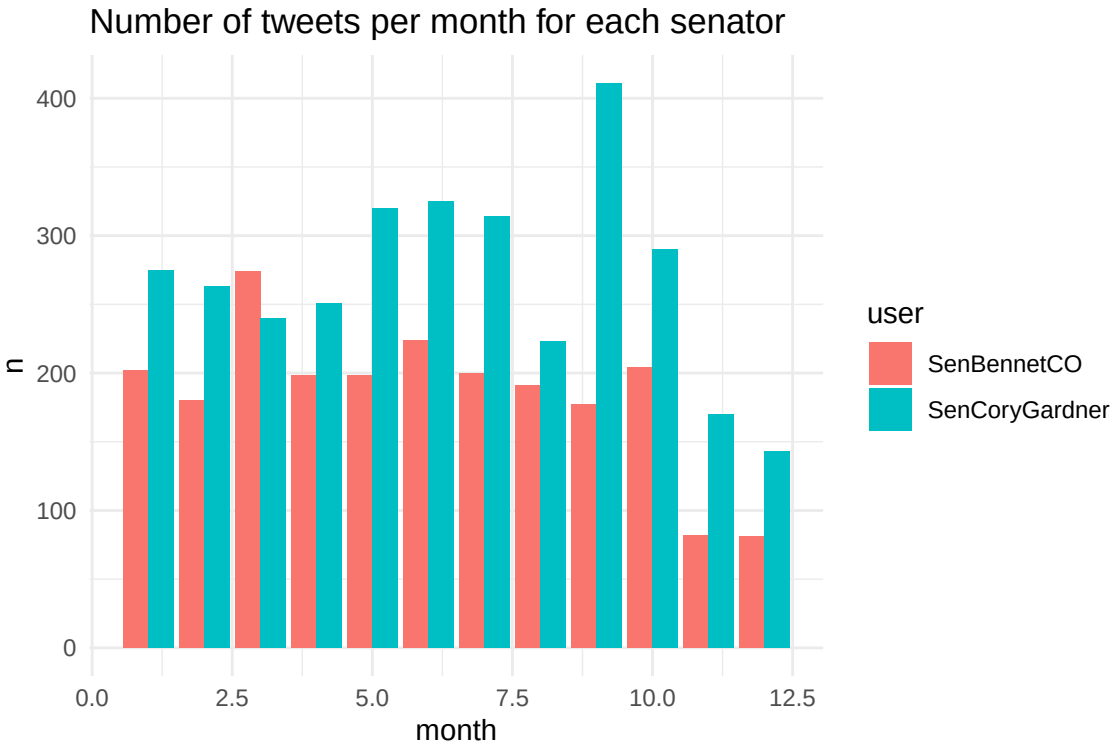
Using the same `wildfires` dataframe, create a summary table that shows the number of tweets containing the word "wildfire" by year (2011-2017). Which year has the most tweets about wildfires? Why might this be the case? (Hint: Think about what happened in the previous year.)

```
## # A tibble: 7 x 2
##    year     n
##   <dbl> <int>
## 1  2011     2
## 2  2012     3
## 3  2013    13
## 4  2014     1
## 5  2015     6
## 6  2016     5
## 7  2017     3
```

2013 has the most tweets about wildfires. After a quick search I found that there were two significant wildfires that break into the top 20 largest in Colorado. Furthermore, one of them destroyed the second most amount of homes for a wildfire in Colorado.

Question 6: Monthly tweets

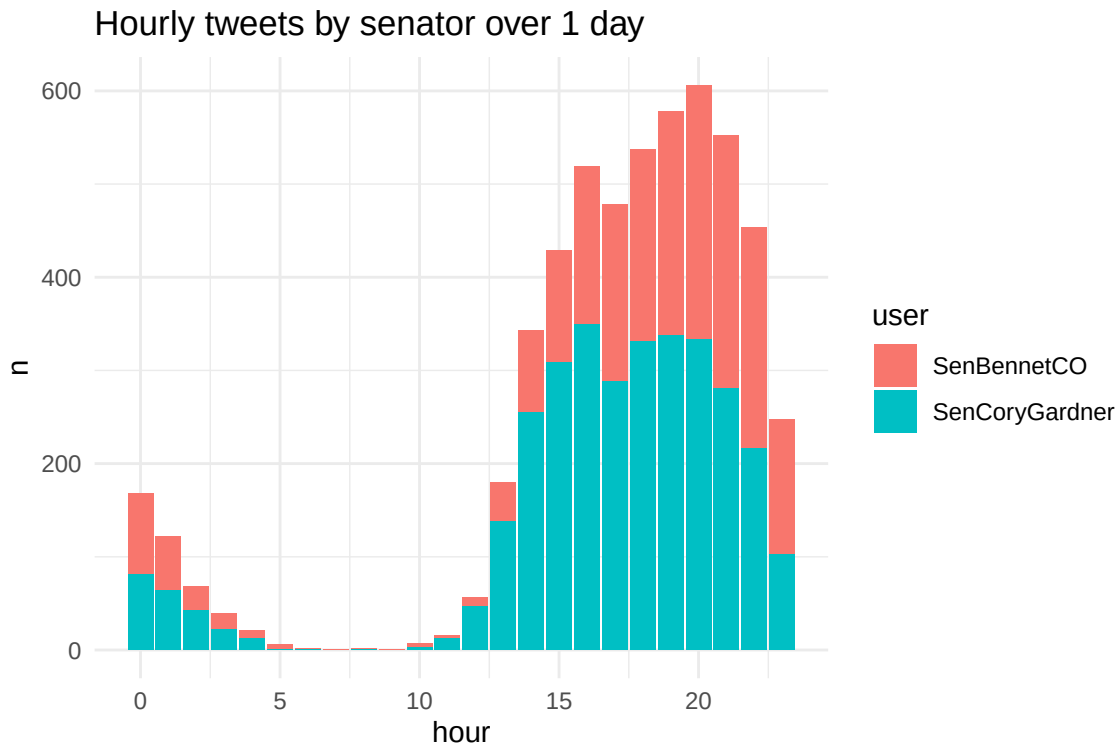
Create a bar chart that answers the question: Are Colorado senators more active at a certain time of year? Hints: Convert `month` to a factor. Fill by `user`.



Generally the senators aren't anymore active during one time of the year. Senator Cory Gardner specifically tweeted more during September through October.

Question 7: Hourly tweets

Create a histogram of tweets by hour of day to visualize when our senators are tweeting.



Appendix

```
# set global options for figures, code, warnings, and messages
knitr::opts_chunk$set(fig.width=6, fig.height=4, fig.path="../figs/",
                      echo=FALSE, warning=FALSE, message=FALSE)
library(readr)
library(stringr)
library(dplyr)
library(lubridate)
library(ggplot2)
senators_co <- read_csv("./senators_co.csv")
tweets <- read_csv("./senators_co.csv") %>%
  select(text) %>%
  as.character() %>%
  str_extract_all("#(\\d|\\w)+")

hashtags <- unlist(tweets) %>%
  length() %>%
  print()
fire <- tweets[[1]] %>%
  str_detect("fire") %>%
  sum() %>%
  print()

wildfire <- tweets[[1]] %>%
  str_detect("wildfire") %>%
  sum() %>%
  print()
# filter to tweets concerning wildfires
wildfire_tweets <- select(senators_co, text, created_at, user) %>%
  filter(str_detect(text, "wildfire"))
# number of wildfire tweets by senator
senators <- wildfire_tweets %>%
  group_by(user) %>%
  tally() %>%
  ungroup() %>%
  print()

# number of wildfire tweets by year
tweets_year <- wildfire_tweets %>%
  mutate(date = lubridate::mdy_hm(created_at),
         year = lubridate::year(date)) %>%
  group_by(year) %>%
  tally() %>%
  ungroup() %>%
  print()
# create plot of tweets by month and user
monthly_tweets <- senators_co %>%
  select(created_at, user) %>%
  mutate(date = lubridate::mdy_hm(created_at),
         month = lubridate::month(date)) %>%
  group_by(user, month) %>%
  tally() %>%
```

```
  ungroup()

ggplot(monthly_tweets, aes(x = month, y = n, fill = user)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Number of tweets per month for each senator")
# create plot of cumulative hourly tweets by senator
hourly_tweets <- senators_co %>%
  select(created_at, user) %>%
  mutate(date = lubridate::mdy_hm(created_at),
         hour = lubridate::hour(date)) %>%
  group_by(user, hour) %>%
  tally() %>%
  ungroup()

ggplot(hourly_tweets, aes(x = hour, y = n, fill = user)) +
  geom_histogram(stat = "identity") +
  theme_minimal() +
  labs(title = "Hourly tweets by senator over 1 day")
```