MECH476: Engineering Data Analysis in R

Chapter 6 Homework: Strings, Dates, and Tidying

Michael Thill

11 December, 2025

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] "Total CO senator hashtags: 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] "'Fire' Hashtags: 138"
```

```
## [1] "'Wildfire' Hashtags: 39"
```

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()`.

```
## # A tibble: 6 x 6
##   date_parsed         year_num month_name hour_of_day text                user
##   <dttm>                 <dbl> <ord>            <int> <chr>               <chr>
## 1 2017-10-06 17:26:00     2017 Oct                 17 "intro'd bill to he~ SenB~
## 2 2017-09-26 19:02:00     2017 Sep                 19 "tune in to watch @~ SenB~
## 3 2017-07-05 19:00:00     2017 Jul                 19 "as #opioid addicti~ SenB~
## 4 2016-07-12 20:05:00     2016 Jul                 20 "our thoughts and s~ SenB~
## 5 2015-08-06 22:45:00     2015 Aug                 22 "glad to see our wi~ SenB~
## 6 2015-06-11 21:25:00     2015 Jun                 21 "#tbt to speaking w~ SenB~
```

Question 4: Senators

Which Colorado senator tweets more about wildfires?

```
## [1] "CO Senator with most wildfire themed tweets: SenBennetCO"
```

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.)

Table 1: CO Senator Wildfire-Related Tweets by Year (2011–2017)

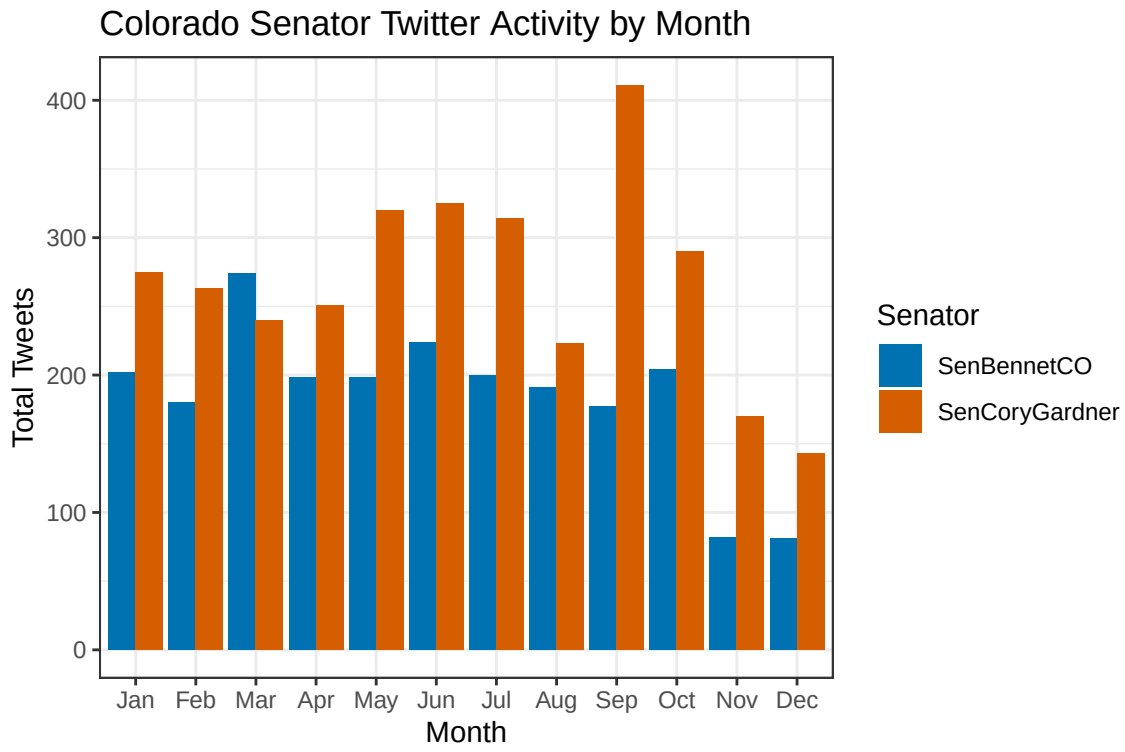
Year	Tweet Count
2013	**16**
2015	8
2016	5
2012	4
2017	3
2011	2
2014	1

```
## [1] "Year with most tweets about wildfires: 2013"
```

*** 2013 was the year with the most tweets related to wildfires, this is likely due to the droughts that had been prevalent in Colorado at the time. this would cause lots of vegetation to die and dry out winch is easy kindling for a fire to start that can spread rapidly as the whole state is very dry. ***

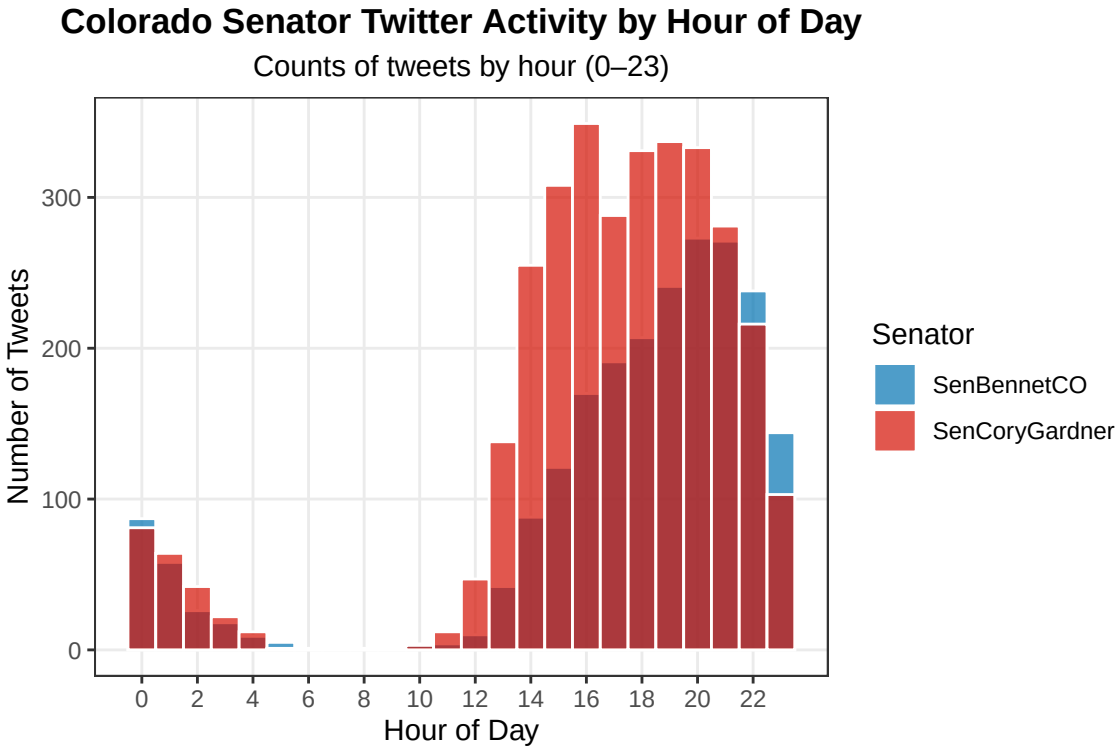
Question 6: Monthly tweets

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`.



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(tidyverse)
library(lubridate)
library(kableExtra)


twitter_raw_data <- read_csv("../data/senators.csv")

tweets <- twitter_raw_data %>%
  select(created_at, text, user, state) %>%
  drop_na() %>%
  rename(date = created_at) %>%
  mutate(text = str_to_lower(text))

co_tweets <- tweets %>%
  filter(state == "CO") %>%
  mutate(
    date_parsed = mdy_hm(date),
    year_num = year(date_parsed),
    month_name = month(date_parsed, label = TRUE, abbr = TRUE),
    hour_of_day = hour(date_parsed)
  )


# wildfire hashtag list

hashtags <- co_tweets %>%
  select(text) %>%
  pull(text) %>%
  str_extract_all(pattern="#(\\d|\\w)+") %>%
  unlist()

paste0("Total CO senator hashtags: ", length(hashtags))

hashtag_fire <- co_tweets %>%
  select(text) %>%
  pull(text) %>%
  str_extract_all(pattern="fire|wildfire") %>%
  unlist()

paste0("'Fire' Hashtags: ", length(hashtag_fire))

hashtag_wildfire <- str_subset(hashtag_fire, "wild")

paste0("'Wildfire' Hashtags: ", length(hashtag_wildfire))

# filter to tweets concerning wildfires
```

```r
wildfire_tweets <- co_tweets %>%
  select(date_parsed, year_num, month_name, hour_of_day, text, user) %>%
  filter(str_detect(text, "wildfire"))

head(wildfire_tweets)

# number of wildfire tweets by senator

Most_tweets <- wildfire_tweets %>%
  count(user, sort=TRUE) %>%
  slice_head(n=1) %>%
  pull(user)

paste0("CO Senator with most wildfire themed tweets: ", Most_tweets)

# number of wildfire tweets by year

yearly_wildfires <- wildfire_tweets %>%
  select(year_num) %>%
  filter(year_num >= 2011 & year_num <= 2017) %>%
  count(year_num, sort = TRUE) %>%
  rename(year = year_num)

yearly_wildfires_table <- yearly_wildfires %>%
  mutate(year = as.integer(year)) %>%
  rename(
    Year = year,
    `Tweet Count` = n
  ) %>%
  knitr::kable(
    caption = "CO Senator Wildfire-Related Tweets by Year (2011-2017)"
  ) %>%
  kable_styling(
    bootstrap_options = c("striped", "hover", "condensed", "responsive"),
    full_width = FALSE,
    position = "center"
  ) %>%
  row_spec(
    row = which.max(yearly_wildfires$n),
    bold = TRUE,
    color = "white",
    background = "#d9534f"
  )

yearly_wildfires_table

top_fire_year <- yearly_wildfires %>%
  slice_head(n=1) %>%
  pull(year)

paste0("Year with most tweets about wildfires: ", top_fire_year)

# create plot of tweets by month and user
```

```r
ggplot(co_tweets, aes(x = month_name, fill = user)) +
  geom_bar(position = "dodge") +
  scale_fill_manual(
    values = c(
      "SenBennetCO" = "#0072B2",
      "SenCoryGardner" = "#D55E00"
    ),
    name = "Senator"
  ) +
  labs(
    title = "Colorado Senator Twitter Activity by Month",
    x = "Month",
    y = "Total Tweets",
    fill = "Senator"
  ) +
  theme_bw()

# create plot of cumulative hourly tweets by senator
ggplot(co_tweets, aes(x = hour_of_day, fill = user)) +
  geom_histogram(
    binwidth = 1,
    position = "identity",
    alpha = 0.7,
    color = "white"
  ) +
  scale_fill_manual(
    values = c(
      "SenBennetCO" = "#0072B2",
      "SenCoryGardner" = "#D50E00"
    ),
    name = "Senator"
  ) +
  scale_x_continuous(
    breaks = seq(0, 23, by = 2),
    limits = c(-0.5, 23.5)
  ) +
  labs(
    title = "Colorado Senator Twitter Activity by Hour of Day",
    subtitle = "Counts of tweets by hour (0-23)",
    x = "Hour of Day",
    y = "Number of Tweets"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5),
    panel.grid.minor = element_blank()
  )
```