Module 12 : Mining Hurricane Harvey Tweets

Saurabh Shinde

11/9/2020

Data Importing

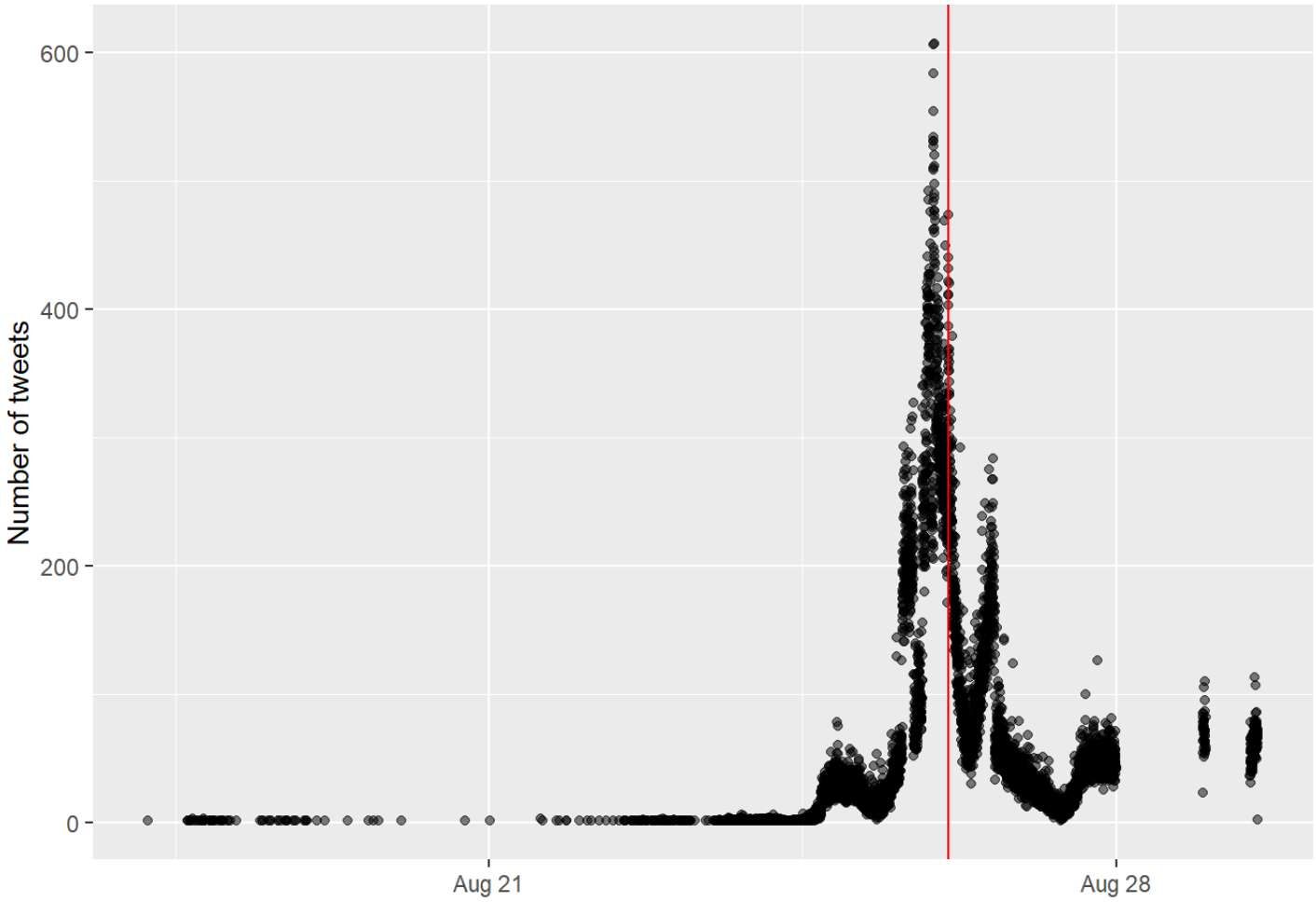
Import the data and packages necessary for analysis

```
#Clear Workspace
rm(list = ls(all = TRUE))

#Load Libraries
library(tidyverse)
library(tidytext)
library(wordcloud)
library(textdata)
library(lubridate)


#Read .csv files
harvey_tweets <- read_csv("data/hurricane_harvey_tweets.csv")
```

Part 1

When did Harvey-related tweets peak in relation to when the hurricane made landfall?

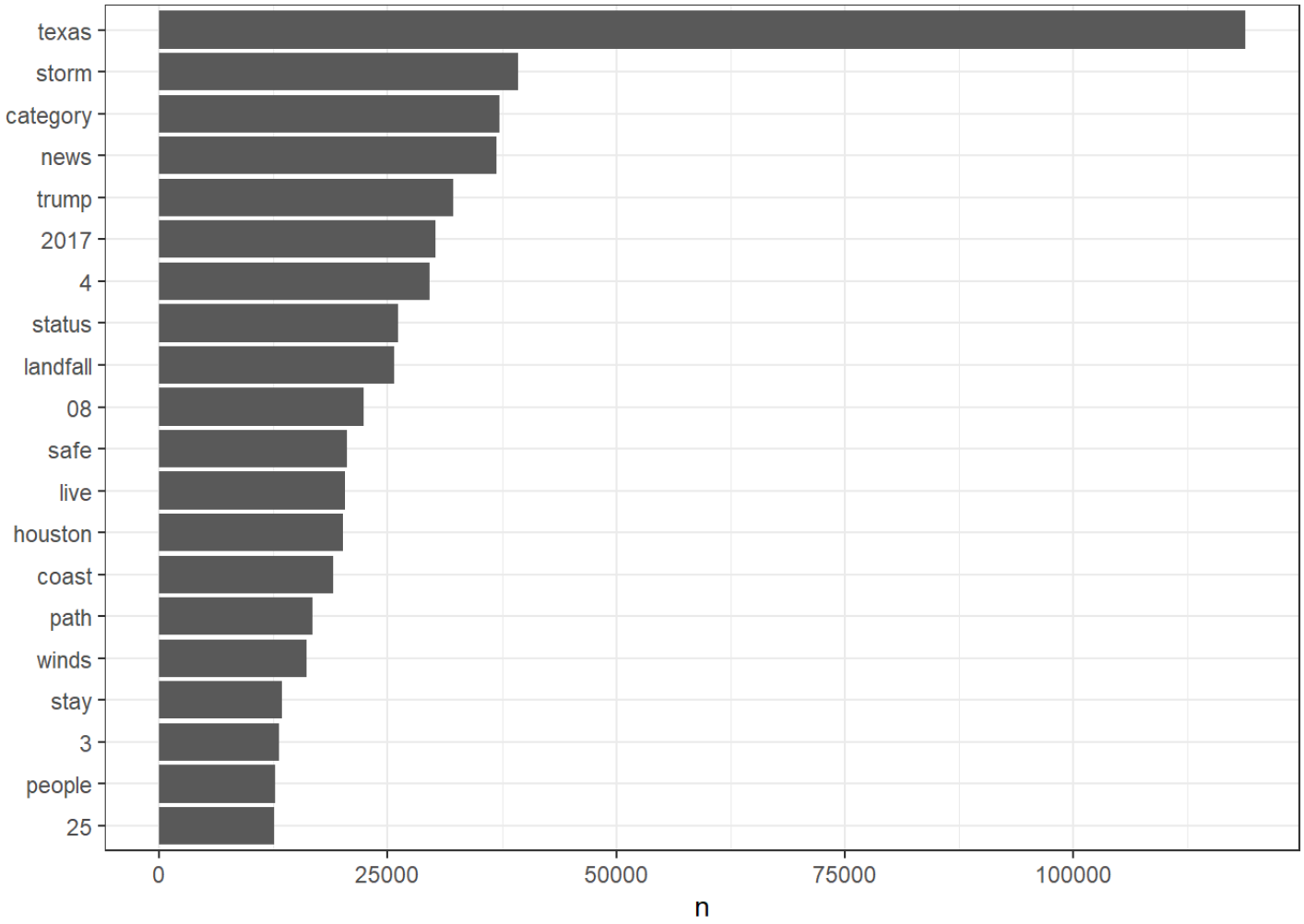
```
#Count tweets
count_tweet <- harvey_tweets %>%
  count(datetime)

#Plotting scatter plot
ggplot(count_tweet) +
  geom_point(aes(x = datetime, y = n), alpha = 0.5) +
  geom_vline(xintercept = (ymd_hms("2017-08-26 03:00:00")), color = 'red') +
  labs(y = 'Number of tweets', x = '')
```



Part 2

What are the 20 most commonly used words in the Hurricane Harvey tweets?

```
#Tokenize data
tokenize_data <- harvey_tweets %>%
  unnest_tokens(word, tweet)

#Remove stop words and custom words
data("stop_words")

remove <- bind_rows(tibble(word = c("hurricane", "harvey", "hurricaneharvey", "http",
"https", "html", "ift.tt", "pic.twitter.com", "twitter.com", "fb.me", "bit.ly", "dlvr
.it", "youtube", "youtu.be"), lexicon = c('custom')), stop_words)

tokenize_data %>%
  anti_join(remove) %>%
  count(word) %>%
  top_n(20) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word)) +
  geom_col() +
  labs(y = NULL) +
  theme_bw()
```



Part 3

What are common words used in tweets that reference refineries?

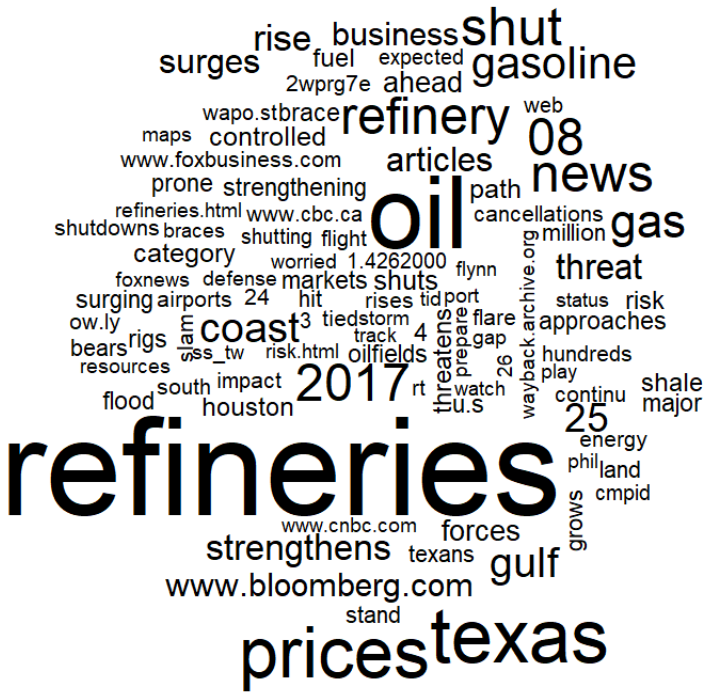
Write 2-3 sentences in your R Markdown file describing whether tweets that referenced refineries seemed to emphasize the potential economic or environmental impacts.

- By looking at the most common words in wordcloud it makes it clear that most of the tweets that referenced refineries seemed to emphasize the potential economic impact more than that of environmental impact.

```
#Filter tweets having refinery and refineries in them
ref <- c('refinery','refineries')
ref_match <- str_c(ref, collapse = '|')
ref_match
```

```
## [1] "refinery|refineries"
```

```
refinery_df <- tibble(tweet = str_subset(harvey_tweets$tweet, ref_match))

tokenize_refinery_df <- refinery_df %>%
  unnest_tokens(word, tweet)

#WordCloud
tokenize_refinery_df %>%
  anti_join(remove) %>%
  count(word, sort = TRUE) %>%
  with(wordcloud(word, n, max.words = 100))
```



Part 4

How did the average sentiment of tweets change from August 17-29, 2019?

```
afinn <- get_sentiments('afinn')

sentiment_date <- tokenize_data %>%
  anti_join(remove) %>%
  inner_join(afinn) %>%
  group_by(date) %>%
  summarise(avg_value = mean(value))

ggplot(sentiment_date) +
  geom_col(aes(x = date, y = avg_value)) +
  scale_x_date(date_breaks = "day", date_labels = "%d") +
  labs(y = 'Average sentiment', x = 'Day in August 2019') +
  theme_bw()
```