Understanding the recent pike in crime in San Francisco

Part 1

Motivation:

Off lately, the city of San Francisco has been a victim of heightened criminal activities. According to the San Francisco Chronicle, as of October 31st, the SFPD has received reports of a staggering and scarcely believable, 810 burglaries or attempted burglaries in and around the Mission District. These numbers are up by 13% from last year. And its, not just burglaries, there has been an increase in shootings, assaults, shoplifting and car break-ins to name a few. Some families have gone to extent of hiring private security to protect their property (O'Brien, C. (2021, November 7). Adding insult to injury, many businesses have decided to reduce their hours. Safeway, a major grocery and produce retailer has decided to cut down its hours from being open 24 hours to just 6 a.m. to 9 p.m. (ANN DAILEY MORENO, T. N. D. (2021, November 1).

This then possess an interesting question, what factors contribute to the rise in criminal activities in San Francisco. Could it be homelessness, could it be leniency due to Proposition 47; which was a controversial bill passed Californian voters seven years ago that made shoplifting something valued at $950 or less a misdemeanor rather than a felony; or could it have anything to do with the COVID-19 pandemic?

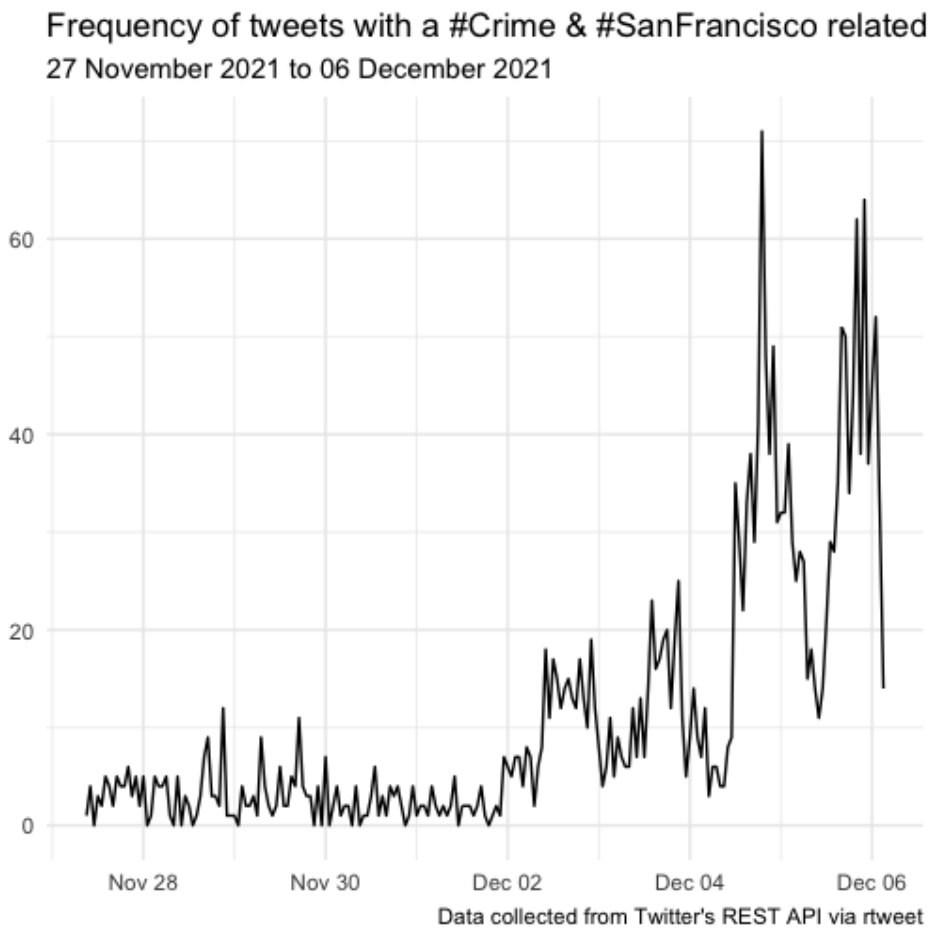
Method:

To answer the question, twitter data was gathered using Twitter's REST API and analyzed through R Programming. The data consisted of conversations people were having around these topics on Twitter that were collected using various hashtags that best represented these topics. Then using principles of text mining like correlograms and pairwise correlations, the conversational data was analyzed for over 2000 tweets. Kindly refer to Part 2 of this document for the full code.

Mechanics:

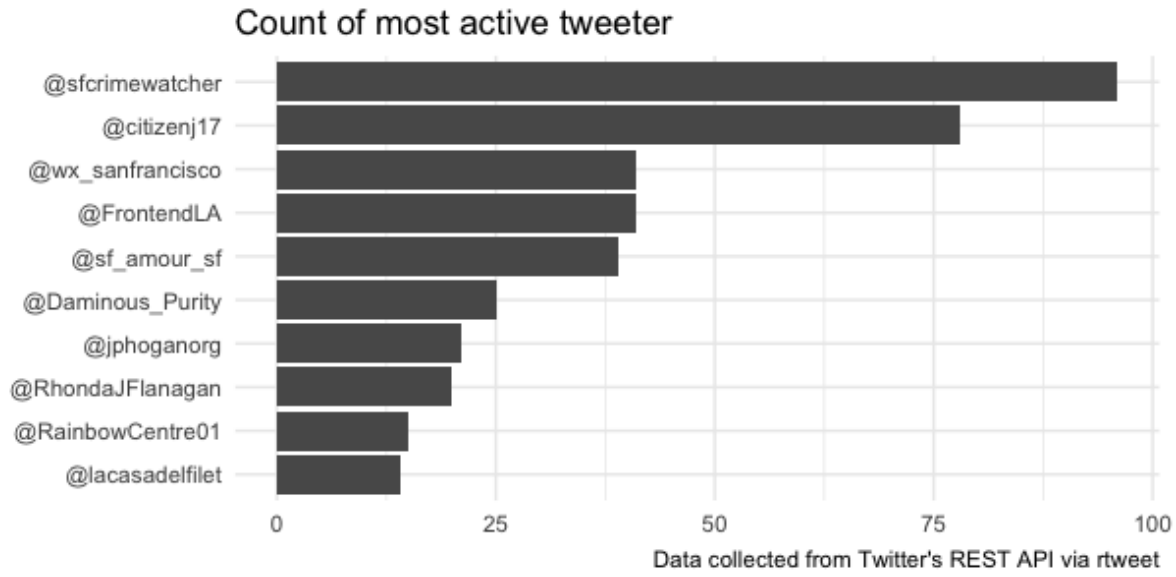
First step was to collate all the data and perform basic alterations like subseting or filtering. Since rtweet package on R can only analyse one week of data, the following analysis was conducted on data collected from 27th November to 6th December 2021.

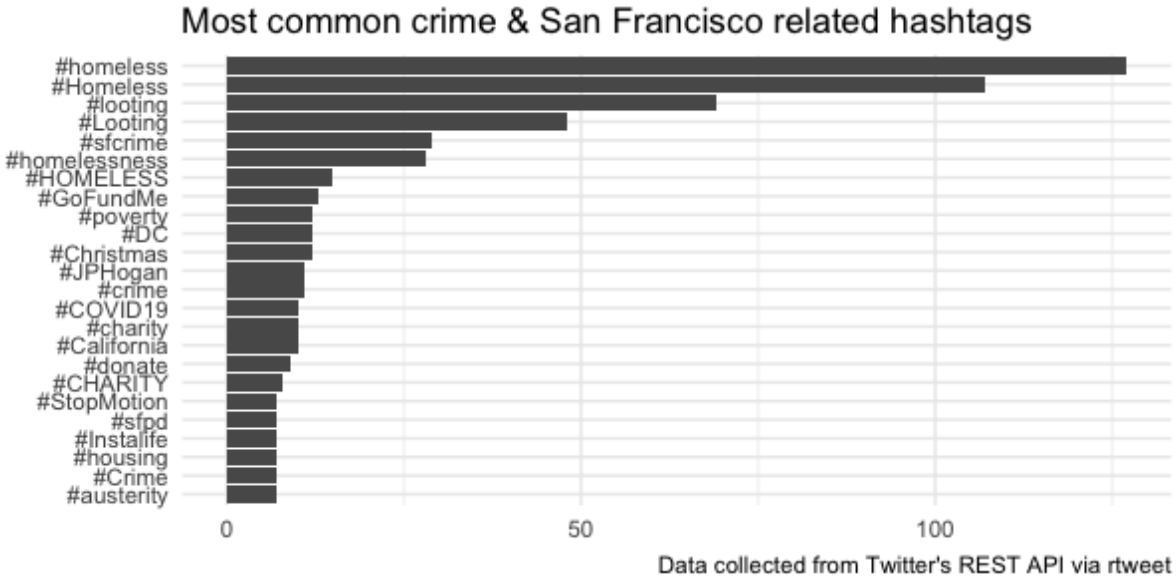
Once the appropriate columns were filtered, descriptive analysis was conducted to understand the data. Starting with the analysis on the frequency of tweets i.e., how often does tweet about crime or its related hashtags in San Francisco.



Whilst there been spikes although out November, the number of tweets increased monumentally in the first week of December. This could be indicative of the holiday season, which historically has known to be a period of high petty criminal activity.

Furthermore, steps were taken to understand the most active tweeter, most common crime hashtags related to San Francisco.



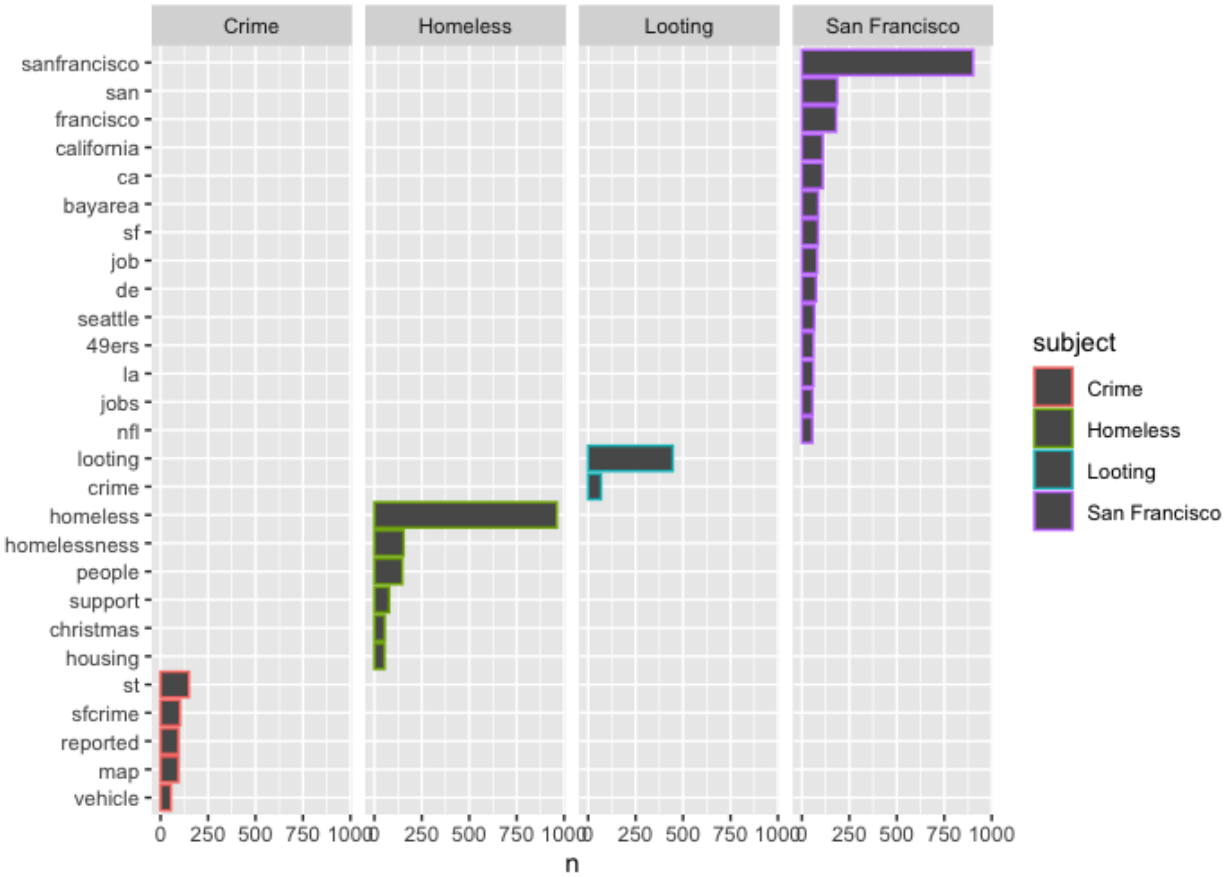


After understanding that there is in fact a relationship between the COVID-19, homelessness, and holidays (as indicated by repeated occurrence "Christmas") the following hypothesis was drafted.

$$Ho: No\ relation\ between\ crime\ and\ (homeless + COVID - 19 + Holidays)$$
$$Ha: Some\ relation\ between\ crime\ and\ (homeless + COVID - 19 + Holidays)$$

Text Mining:

The text data within the tweets were isolated by removing the links and mentions. Then text was tokenized using the tidy format, the stop words were removed, and the frequency of the tokens were plotted by subject.



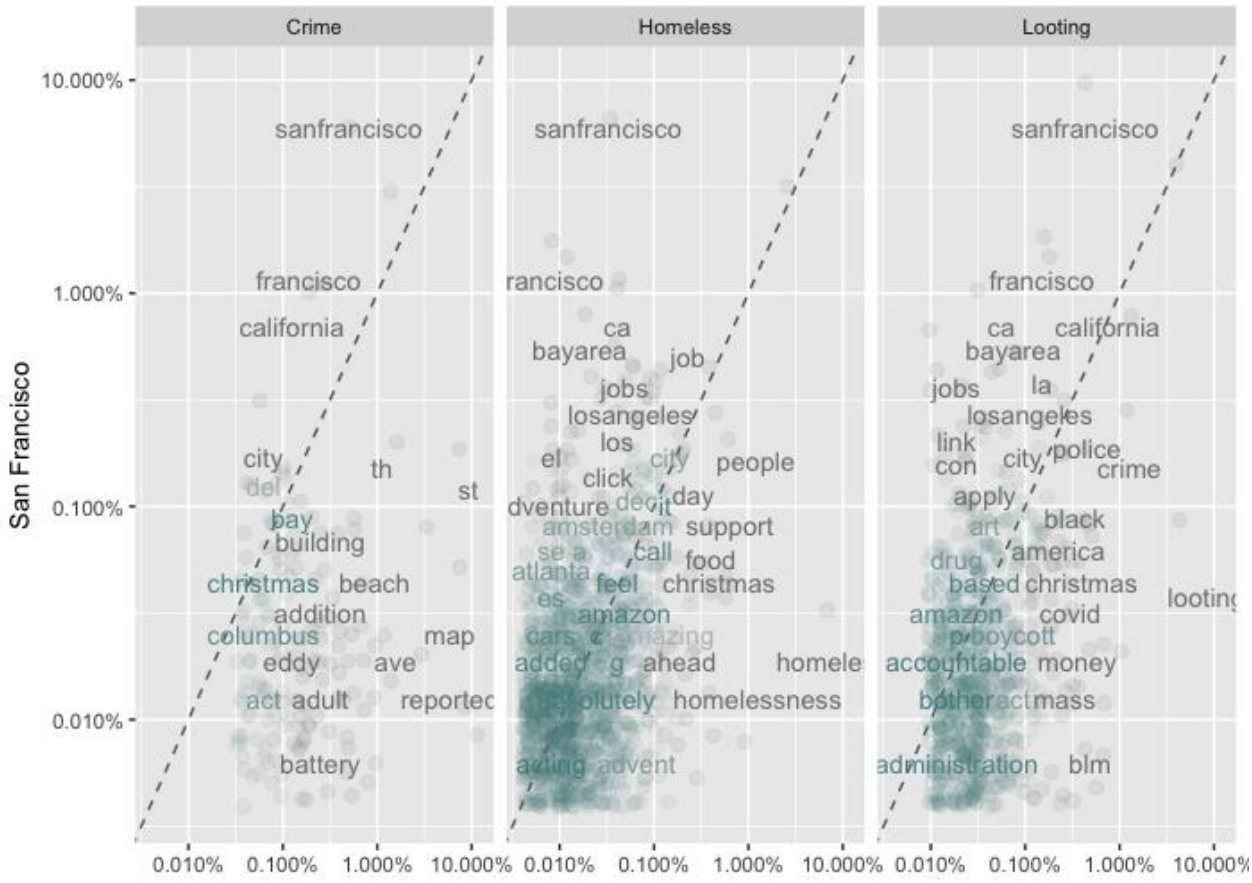
While the following gives an understanding the most contested topics within each subject groups, we need to compare the correlation between these topics to prove our hypothesis.

Using correlograms, we can compare the correlation between Crime and San Francisco, words like crime, looting, business, bay appear to be common. Similarly, in Homeless, words like crime, center and district stand out and, finally for Looting, words like bay, California, city stood out. Therefore, we can deduce that there is in fact some correlation between these topics.

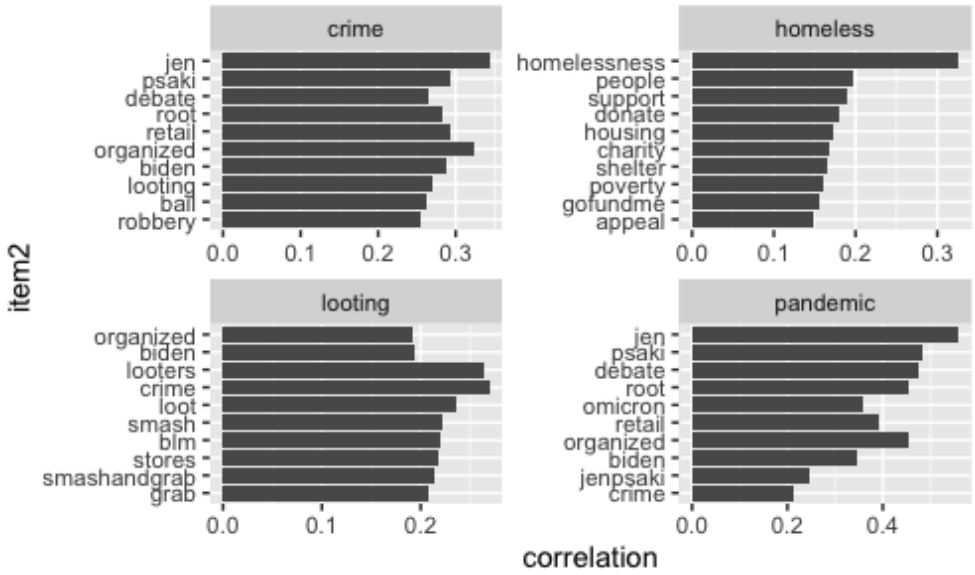
Correlation Table

Topics	Crime	Homeless	Looting
San Francisco	0.03258107	0.20936	0.1861214

But there is certain limitation to this analysis. Tokenization tends to remove context as it difficult to understand the entire range of emotion by analyzing a single word. Therefore, we need to find the correlation between a pair of words or collection of words to truly understand the true relation.



Pairwise Correlations:



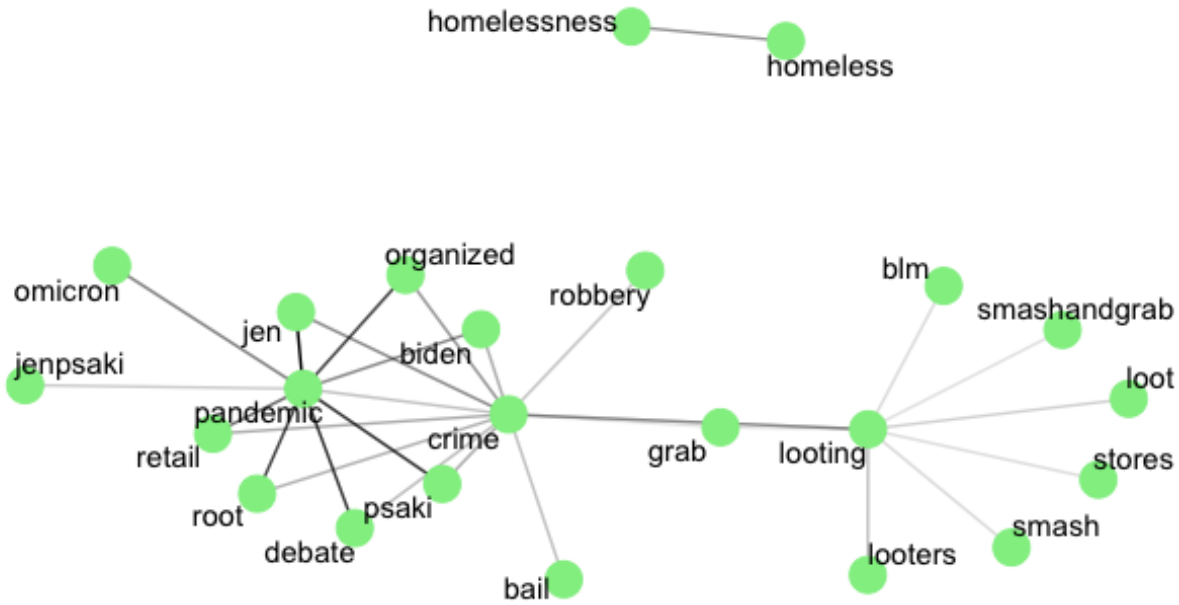
After performing a pairwise correlation we can visualize the overlaps more clearly. There is a very clear and distinct relation between the pandemic and organized crime in the city. But our initial hypothesis needs to be updated as the overlap between crime and homelessness or holidays has reduced.

Updated Hypotheses:

$$Ho: No\ relation\ between\ crime\ and\ pandemic$$

$$Ha: Some\ relation\ between\ crime\ and\ pandemic$$

To correctly visualize the correlation between these topics, we need to plot a node plot.



Message:

From the graph, we can clearly see that the scare of potential 4th wave of the new COVID-19 variant Omicron could be a possible reason the recent spike in criminal activities in San Francisco. We can also the nature of these crimes. Most crimes are organized and are of the smash and grab variety. And the most common target seems to be retail stores.

A possible explanation of the same can be due to the survival instinct of human beings. When faced with adversity, we need to prepare by hoarding on items. While this doesn't excuse their actions, it does provide an explanation.

A way to mitigate this issue would be run education camps on proper COVID protection protocols and increase assistance to these people who feel the necessity to commit a felony by giving their food and essentials.

Therefore, the updated hypothesis is proven true. There is in fact a relation between the pandemic and the recent spike crime.

Citations

1. ANN DAILEY MORENO, T. N. D. (2021, November 1). *San Francisco supermarket reduces hours due to increase in crime; an 'equity issue'*. WSET. Retrieved December 6, 2021, from https://wset.com/news/nation-world/another-san-francisco-market-reduces-hours-due-to-increase-in-crime-an-equity-issue.
2. O'Brien, C. (2021, November 7). *San Francisco Chronicle ripped for asking if residents should 'tolerate burglaries'*. Fox News. Retrieved December 6, 2021, from https://www.foxnews.com/media/san-francisco-chronicle-ripped-for-asking-if-residents-should-tolerate-burglaries.

Twitter_Analysis

Arjun Manohar

12/5/2021

Part 2

This section contains the code the output of the different sections

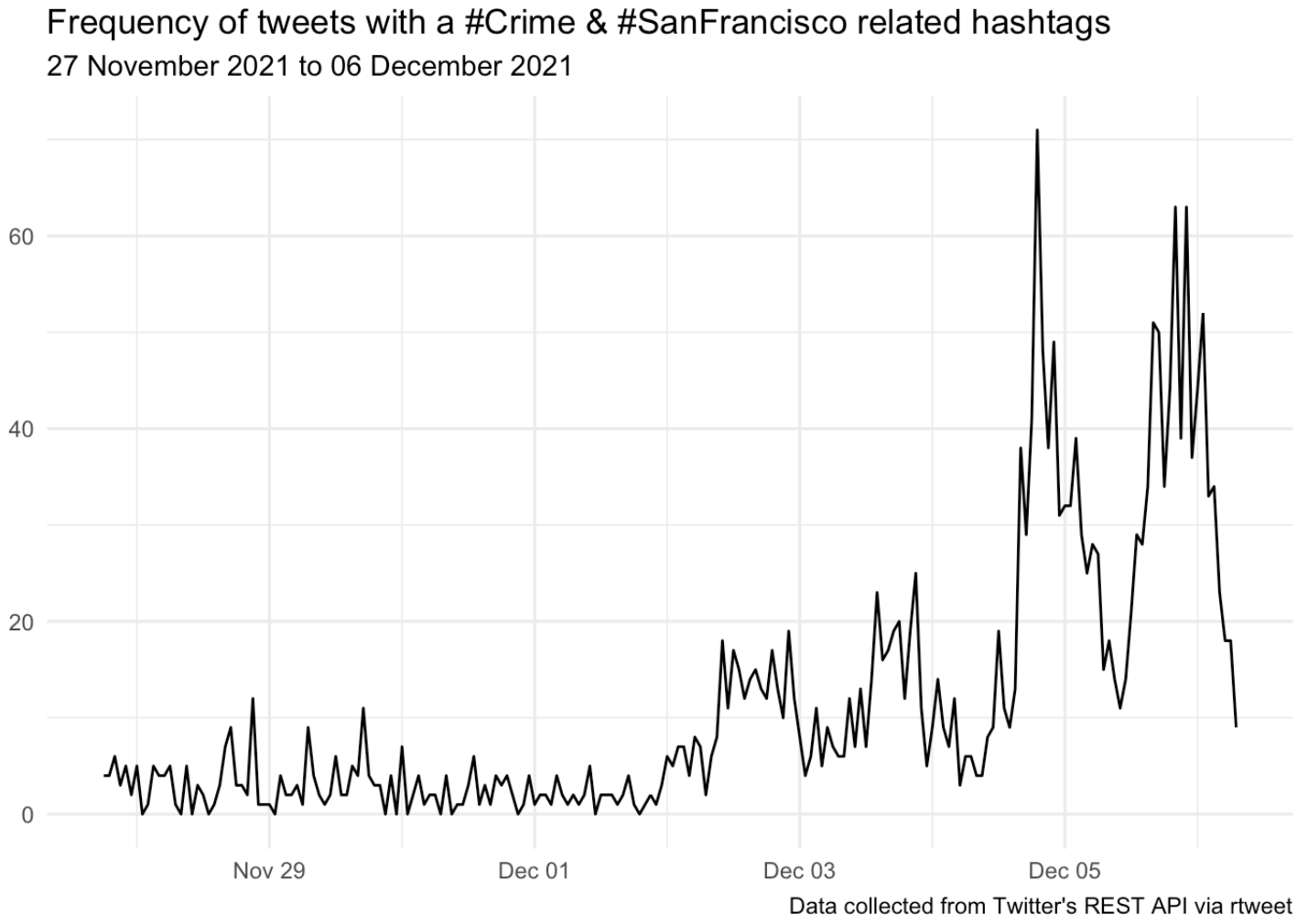
```
# install.packages("rtweet")
library(rtweet)
library(dplyr)
library(stringr)
library(ggplot2)
library(tm)
library(tidyr)

consumer_key <- "NkBpBPbhIZc9YdHnw88kTWLmx"
consumer_secret <- "VuNqU4gXeofa4MbBBX2CkZeuNwlNCTOrhIeAmr1L82b2e2EEpj"
access_token <- "1040465492-gTYOGPyjoeCNNyBnxifJEoHt93rmsf1NggMcyjA"
access_secret <- "z4zbHD9vF7mQ4wfgGom3IRqKru9GFsJSdza4sidDhTQtC"
name_of_app <- "Test_ArjunM"

twitter_token <- create_token(
  app = name_of_app,
  consumer_key = consumer_key,
  consumer_secret = consumer_secret,
  access_token = access_token,
  access_secret = access_secret)

# # hashtags <- c("#sanfrancisco", "#SanFrancisco", "#sfcrime" ,"#crime",
#               "#homeless", "#lootings", "#shoplifting", "#unsafe",
#               "police", "#SF", "#sfpd", "tenderloin", "#homelessSF")

rt <- search_tweets("#looting", n = 1000, include_rts = FALSE)
rt1 <- search_tweets("#sfcrime", n = 1000, include_rts = FALSE)
rt2 <- search_tweets("#SanFrancisco", n = 1000, include_rts = FALSE)
rt4 <- search_tweets("#homeless", n = 1000, include_rts = FALSE)


twitter_data <- bind_rows(mutate(rt, subject="Looting"),
                          mutate(rt1, subject="Crime"),
                          mutate(rt4, subject ="Homeless"),
                          mutate(rt2, subject="San Francisco"))
```

```
# let's look at the data
# names(twitter_data)
# View(twitter_data)

# subseting the data
tweet_data <- twitter_data %>%
  select("created_at", "screen_name", "text", "source",
         "favorite_count", "retweet_count", "hashtags",
         "place_full_name", "subject")

#######################
# Data Massaging
#######################
library(qdapRegex)

# clean_tweets <- data.frame()
# tweet_data$text = gsub("(f|ht)tp(s?)://(.*)[.][a-z]+", "", tweet_data$text)

tweet_data$text = gsub("@[a-z,A-z,0-9]*", "", tweet_data$text) # removing usernames f
rom the tweets
tweet_data$text <- rm_twitter_url(tweet_data$text) # removing URLs from the tweets
```

Understanding the data

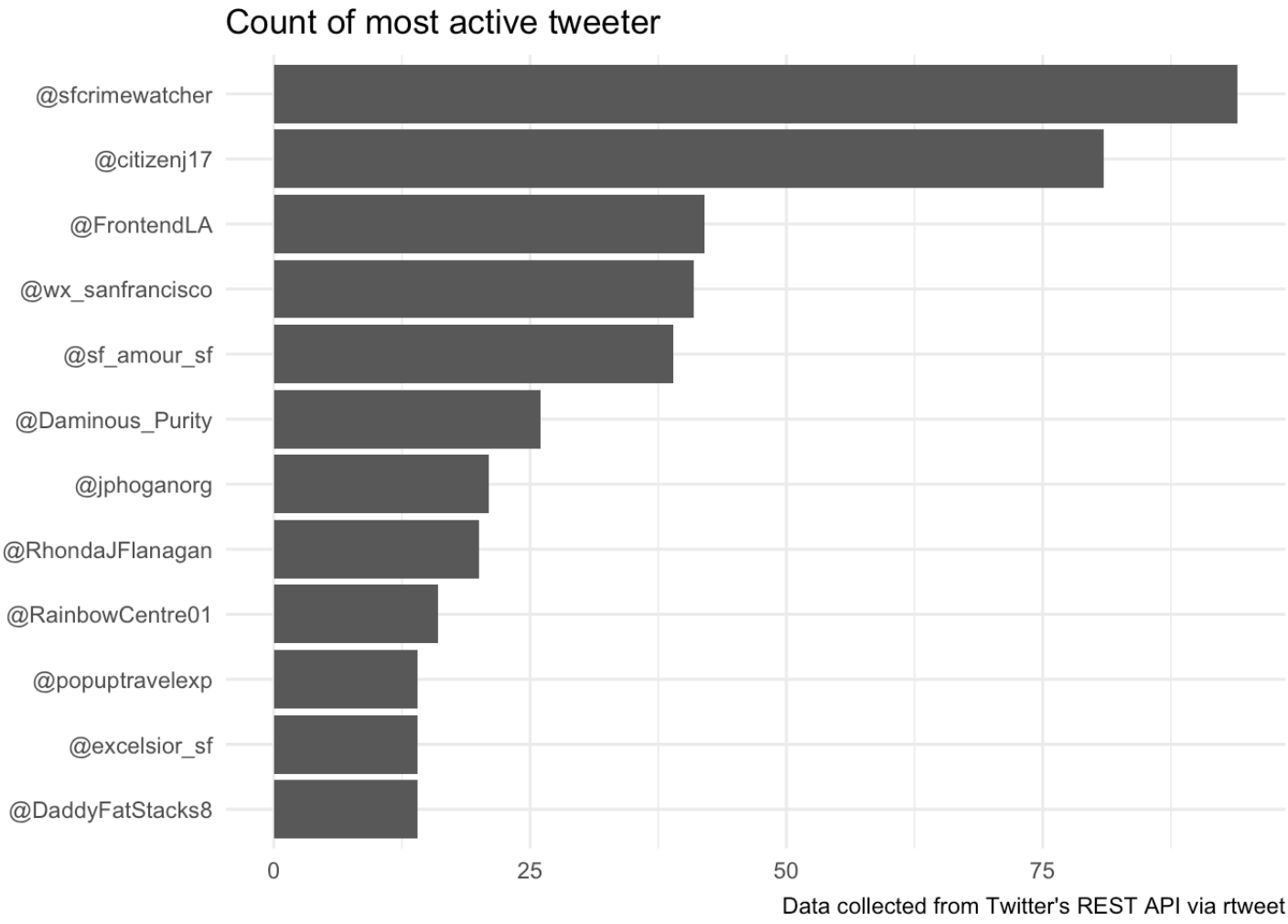
Descriptive analysis

```
# plotting the frequency of tweets
library(tidytext)
ts_plot(tweet_data, "hours") +
  labs(x = NULL, y = NULL,
       title = "Frequency of tweets with a #Crime & #SanFrancisco related hashtags",
       subtitle = paste0(format(min(tweet_data$created_at), "%d %B %Y"), " to ", form
at(max(tweet_data$created_at),"%d %B %Y")),
       caption = "Data collected from Twitter's REST API via rtweet") +
  theme_minimal()
```



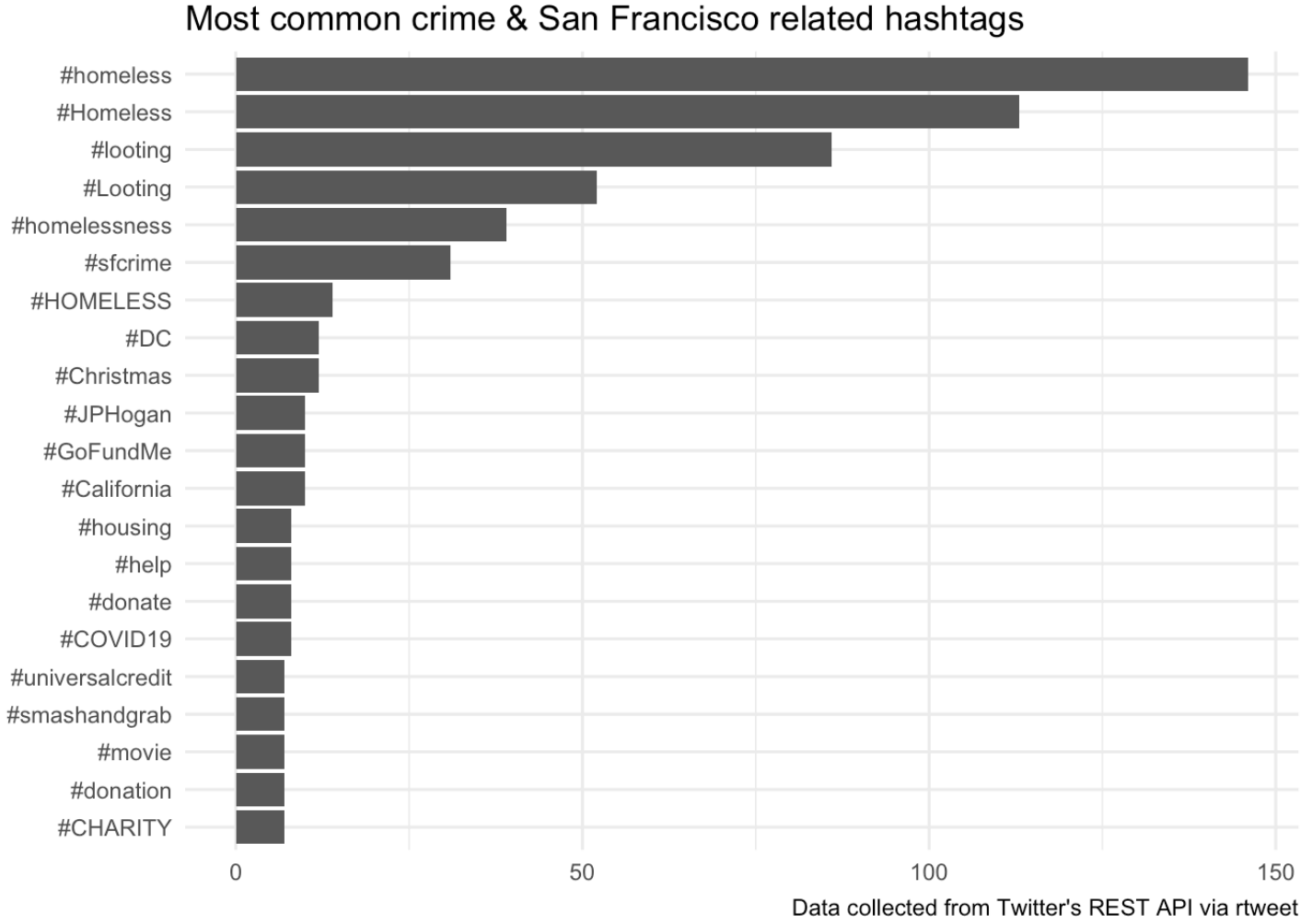

```
# plotting the most active tweeter
tweet_data %>%
  count(screen_name, sort = TRUE) %>%
  top_n(10) %>%
  mutate(screen_name = paste0("@", screen_name)) %>%
  mutate(screen_name=reorder(screen_name,n)) %>%
  ggplot(aes(screen_name,n)) + geom_col() +
  labs(x = NULL, y = NULL,
       title = "Count of most active tweeter",
       caption = "Data collected from Twitter's REST API via rtweet") +
  theme_minimal() +
  coord_flip()
```

```
## Selecting by n
```


Count of most active tweeter
@sfcrimewatcher
@citizenj17
@FrontendLA
@wx_sanfrancisco
@sf_amour_sf
@Daminous_Purity
@jphoganorg
@RhondaJFlanagan
@RainbowCentre01
@popuptravelexp
@excelsior_sf
@DaddyFatStacks8
0
25
50
75
Data collected from Twitter's REST API via rtweet

```
# # most common hashtags
# tweet_data %>%
#   unnest_tokens(hashtag, text, "tweets", to_lower = FALSE) %>%
#   filter(str_detect(hashtag, "^#"), hashtag != "#SanFrancisco", hashtag != "#sanfra
ncisco", hashtag != "#crime") %>%
#   count(hashtag, sort = TRUE) %>%
#   top_n(100)

# most common crime & San Francisco related hashtags
tweet_data %>%
  unnest_tokens(hashtag, text, "tweets", to_lower = FALSE) %>%
  filter(str_detect(hashtag, "^#"), hashtag != "#sanfrancisco",
         hashtag !="#SanFrancisco", subject == c("Crime", "Looting", "Homeless")) %>%
  count(hashtag, sort = TRUE) %>%
  top_n(20) %>%
  mutate(hashtag=reorder(hashtag,n)) %>%
  ggplot(aes(hashtag,n)) + geom_col() +
  labs(x = NULL, y = NULL,
       title = "Most common crime & San Francisco related hashtags",
       caption = "Data collected from Twitter's REST API via rtweet") +
  theme_minimal() +
  coord_flip()
```

```
## Selecting by n
```

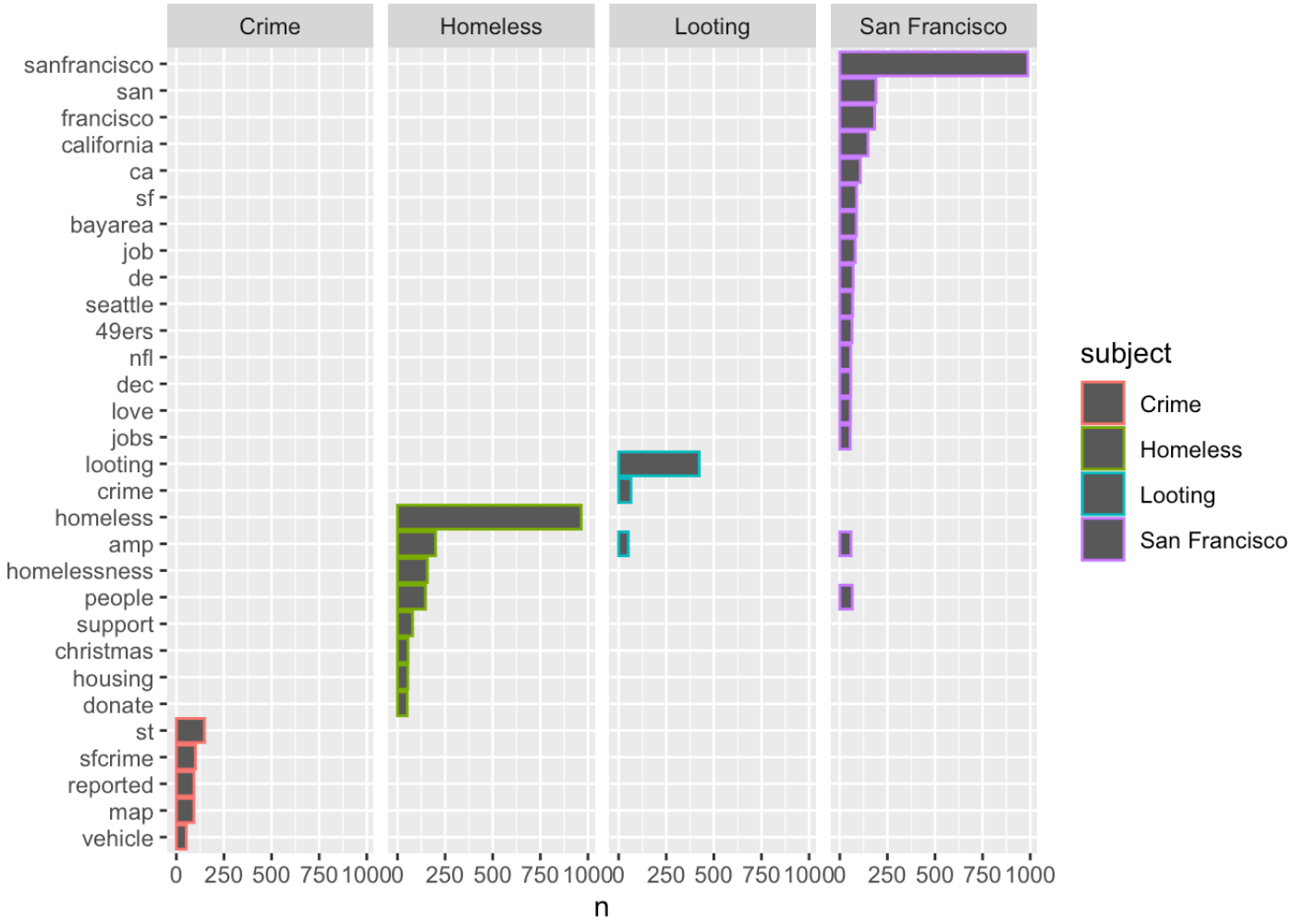


Tokenization, Tidy Format & Correlograms

```
tidy_tweet <- tweet_data %>% unnest_tokens(word, text) %>% anti_join(stop_words)
```

```
## Joining, by = "word"
```

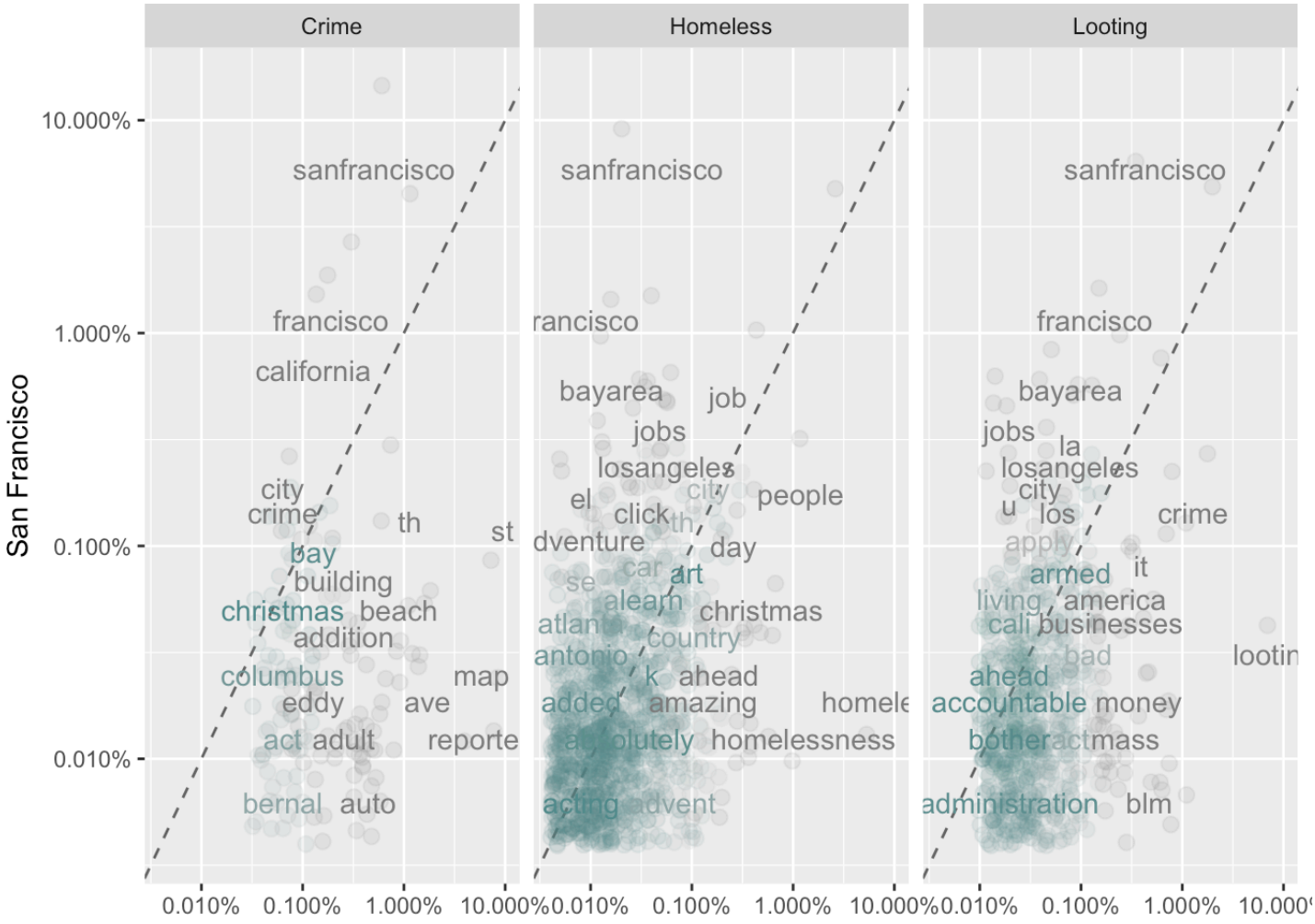
```
# plotting tokenised frequencies
tidy_tweet %>%
  group_by(subject) %>%
  count(word, sort = T) %>%
  filter(n>50) %>%
  mutate(word=reorder(word,n)) %>%
  ggplot(aes(word,n,color = subject)) + geom_col() + xlab(NULL) + coord_flip() + face
t_wrap(~subject, ncol=4)
```


Crime
Homeless
Looting
San Francisco
sanfrancisco
san
francisco
california
ca
sf
bayarea
job
de
seattle
49ers
nfl
dec
love
jobs
looting
crime
homeless
amp
homelessness
people
support
christmas
housing
donate
st
sfcrime
reported
map
vehicle
0 250 500 750 10000
250 500 750 10000
250 500 750 10000
250 500 750 1000
n
subject
Crime
Homeless
Looting
San Francisco

```
# cleaning the data - removing weather and bots
tidy_tweet <- tidy_tweet %>% filter(screen_name != "sf_amour_sf",
                                    screen_name != "wx_sanfrancisco",
                                    word != "amp", word != "fuck",
                                    word != "fucking")

# plotting correlograms
library(tidyr)
frequency <- tidy_tweet %>%
  mutate(word=str_extract(word, "[a-z']+")) %>%
  count(subject, word) %>%
  group_by(subject) %>%
  mutate(proportion = n/sum(n))%>%
  select(-n) %>%
  spread(subject, proportion) %>%
  gather(subject, proportion, `Looting`, `Crime`, `Homeless`)

# let's plot the correlograms:
library(scales)
ggplot(frequency, aes(x=proportion, y=`San Francisco`,
                      color = abs(`San Francisco`- proportion)))+
  geom_abline(color="grey40", lty=2)+
  geom_jitter(alpha=.1, size=2.5, width=0.3, height=0.3)+
  geom_text(aes(label=word), check_overlap = TRUE, vjust=1.5) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
  scale_color_gradient(limits = c(0,0.001), low = "darkslategray4", high = "gray75")+
  facet_wrap(~subject, ncol=3)+
  theme(legend.position = "none")+
  labs(y= "San Francisco", x=NULL)
```

```
## Warning: Removed 26580 rows containing missing values (geom_point).
```

```
## Warning: Removed 26583 rows containing missing values (geom_text).
```




```
cor.test(data=frequency[frequency$subject == "Crime",],
         ~proportion + `San Francisco`) # very low correlation observed
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and San Francisco
## t = 0.34593, df = 156, p-value = 0.7299
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1290122  0.1830349
## sample estimates:
##       cor
## 0.02768577
```

```
cor.test(data=frequency[frequency$subject == "Looting",],
         ~proportion + `San Francisco`) # low-moderate correlation observed
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and San Francisco
## t = 5.6398, df = 698, p-value = 2.474e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1367789 0.2785593
## sample estimates:
##       cor
## 0.2087658
```

```
cor.test(data=frequency[frequency$subject == "Homeless",],
         ~proportion + `San Francisco`) # low correlation observed
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and San Francisco
## t = 6.4356, df = 1171, p-value = 1.789e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1289534 0.2395301
## sample estimates:
##       cor
## 0.1848267
```

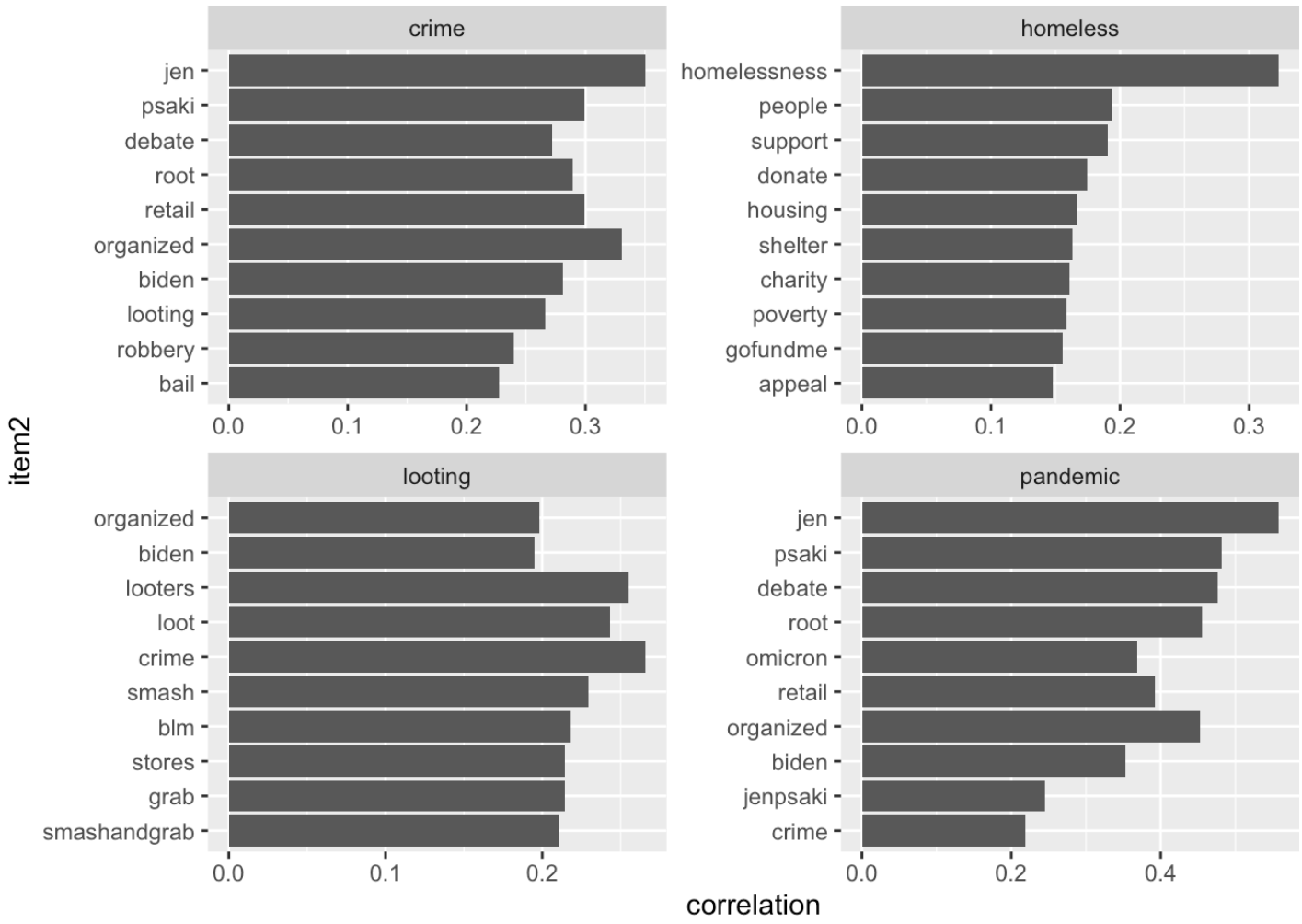
Pairwise Correlation Analysis

```
library(tidytext)
library(widyr)
new_tweet_data <- twitter_data %>% mutate(tweet_id = row_number())

# cleaning the data
new_tweet_data$text = gsub("@[a-z,A-z,0-9]*", "", tweet_data$text) # removing usernam
es from the tweets
new_tweet_data$text <- rm_twitter_url(tweet_data$text) # removing URLs from the tweet
s

# tokenising
new_tweet_tidy <- new_tweet_data %>% unnest_tokens(word, text) %>% anti_join(stop_wor
ds)
```

```
## Joining, by = "word"
```

```
new_tweet_tidy <- new_tweet_tidy %>% filter(screen_name != "sf_amour_sf",
                                     screen_name != "wx_sanfrancisco",
                                     word != "amp", word != "fuck",
                                     word != "fucking")


tweet_cors <- new_tweet_tidy %>%
  group_by(word) %>%
  filter(n() >= 5) %>%
  pairwise_cor(word, tweet_id, sort=TRUE)

tweet_cors %>%
  filter(item1 %in% c("looting", "pandemic", "crime", "homeless")) %>%
  group_by(item1) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(item2 = reorder(item2, correlation)) %>%
  ggplot(aes(item2, correlation)) +
  geom_bar(stat = "identity")+
  facet_wrap(~item1, scales = "free")+
  coord_flip()
```

```
## Selecting by correlation
```




```
# creating a correlation network
library(ggraph)
library(igraph)
```

```
##
## Attaching package: 'igraph'
```

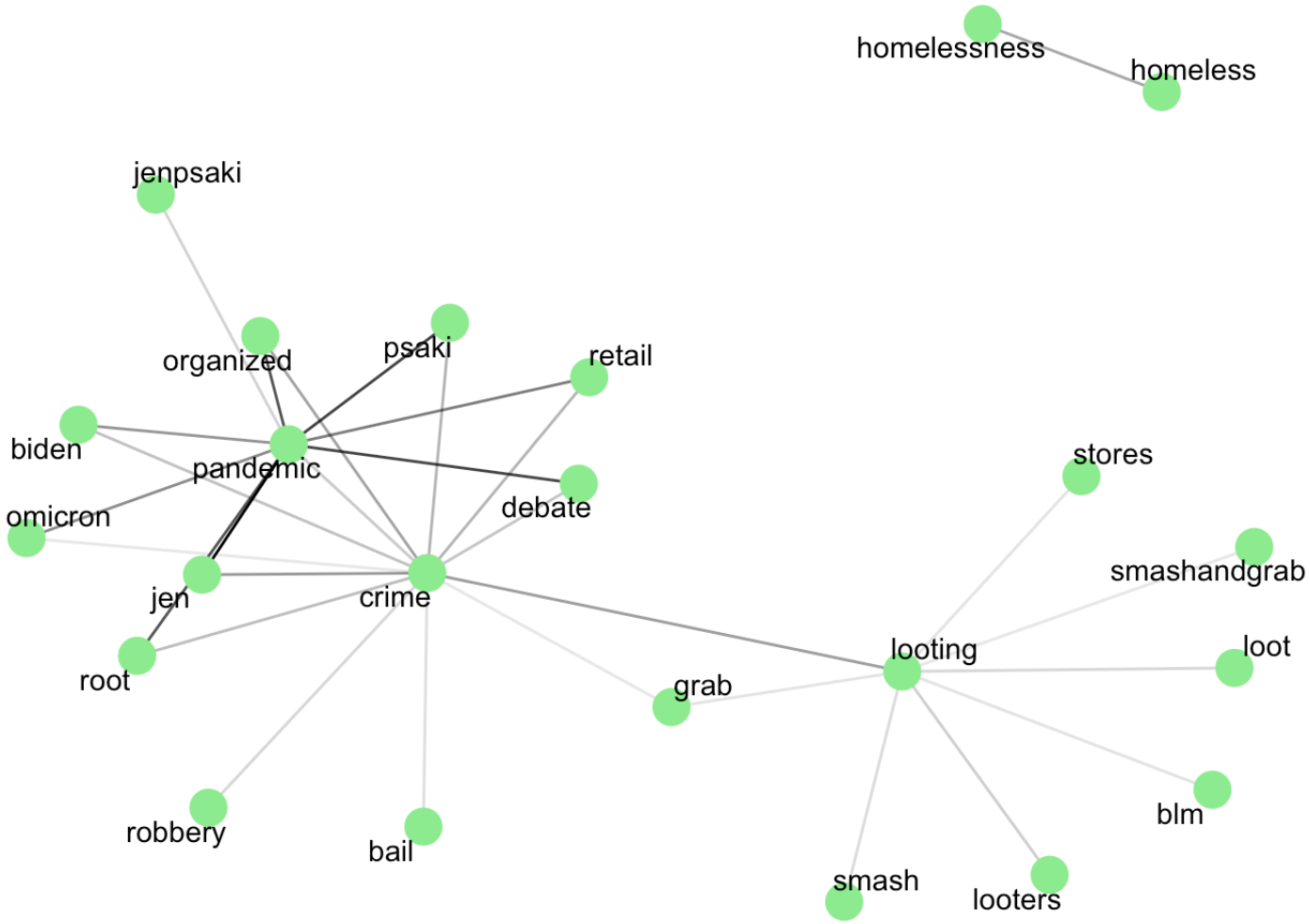
```
## The following object is masked from 'package:tidyr':
##
##     crossing
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```
tweet_cors %>%
  filter(item1 %in% c("looting", "pandemic", "crime", "homeless")) %>%
  filter(correlation >.2) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr")+
  geom_edge_link(aes(edge_alpha = correlation), show.legend=F)+
  geom_node_point(color = "lightgreen", size=6)+
  geom_node_text(aes(label=name), repel=T)+
  theme_void()
```


homelessness
homeless
jenpsaki
organized
psaki
retail
biden
pandemic
stores
omicron
debate
smashandgrab
jen
crime
looting
loot
root
grab
blm
robbery
bail
smash
looters

The End