Homework 07

Spencer Pease

5/25/2020

Questions

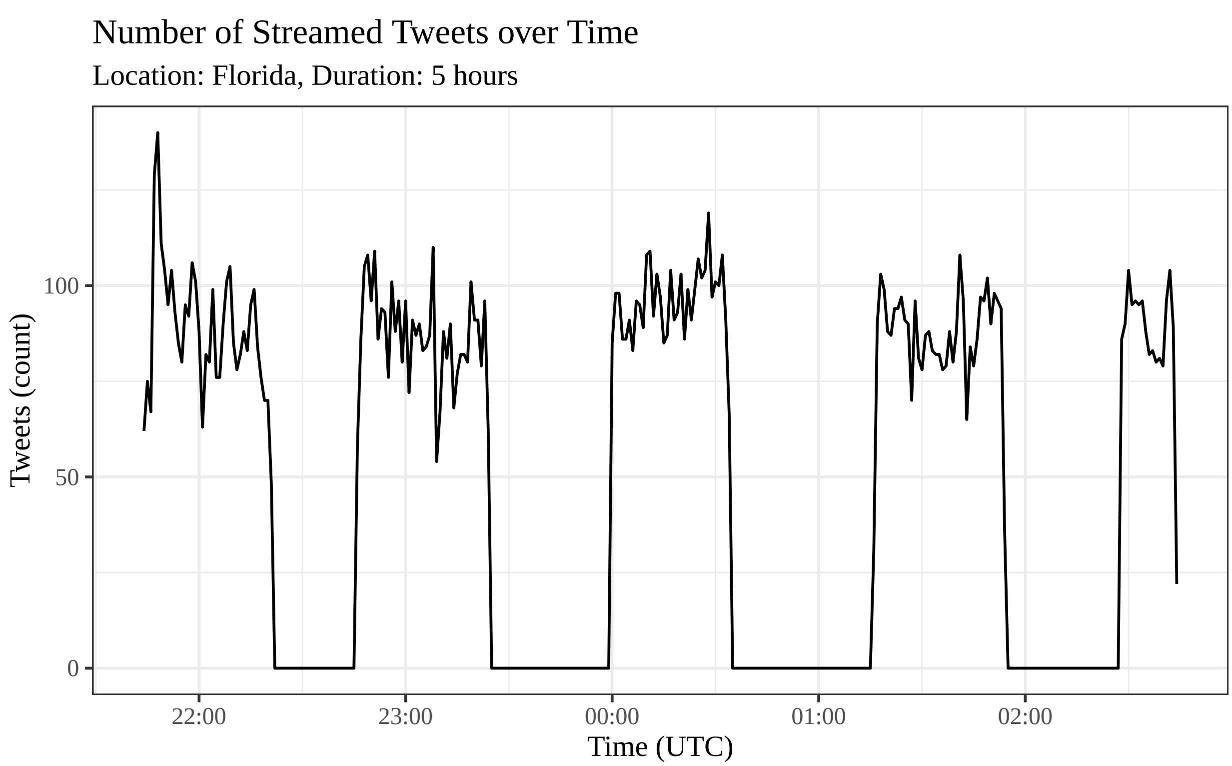
Q1

Q1.a

This report focuses on collecting Twitter data from Florida. The bounding box used to encompass Florida is $(-87.587, 24.257, -79.735, 30.983)$, created using [bboxfinder.com][http://bboxfinder.com/].

Q1.b

After streaming tweets from Florida for around 5 hours, **14888** were collected. We can inspect the distribution of these collected tweets in a time series plot:



The periods of zero tweets are likely a result of connection issues when collecting data, and not representative of the true frequency of tweets.

Q2

Q2.a

Got a census key!

Q2.b

From the *American Community Survey*, we get total population estimates for all counties within Florida:

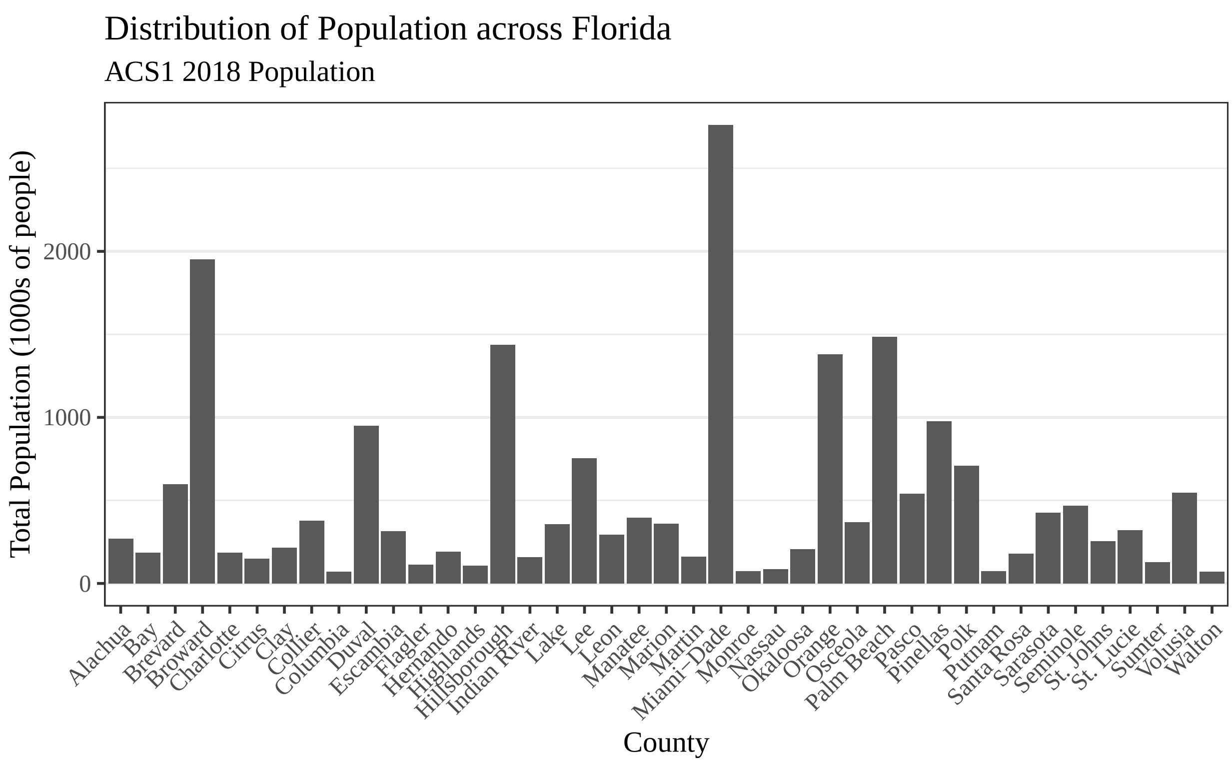
Table 1: Florida counties and total population estimated by ACS1 2018

County	State	GEOID	Pop. estimate
Alachua	Florida	12001	269956
Bay	Florida	12005	185287
Brevard	Florida	12009	596849
Broward	Florida	12011	1951260
Charlotte	Florida	12015	184998
Citrus	Florida	12017	147929
Clay	Florida	12019	216072
Collier	Florida	12021	378488
Columbia	Florida	12023	70503
Duval	Florida	12031	950181
Escambia	Florida	12033	315534
Flagler	Florida	12035	112067
Hernando	Florida	12053	190865
Highlands	Florida	12055	105424
Hillsborough	Florida	12057	1436888
Indian River	Florida	12061	157413
Lake	Florida	12069	356495
Lee	Florida	12071	754610
Leon	Florida	12073	292502
Manatee	Florida	12081	394855
Marion	Florida	12083	359977
Martin	Florida	12085	160912
Miami-Dade	Florida	12086	2761581
Monroe	Florida	12087	75027
Nassau	Florida	12089	85832
Okaloosa	Florida	12091	207269
Orange	Florida	12095	1380645
Osceola	Florida	12097	367990
Palm Beach	Florida	12099	1485941
Pasco	Florida	12101	539630
Pinellas	Florida	12103	975280
Polk	Florida	12105	708009
Putnam	Florida	12107	74163
St. Johns	Florida	12109	254261
St. Lucie	Florida	12111	321128
Santa Rosa	Florida	12113	179349
Sarasota	Florida	12115	426718
Seminole	Florida	12117	467832
Sumter	Florida	12119	128754

County	State	GEOID	Pop. estimate
Volusia	Florida	12127	547538
Walton	Florida	12131	71375

Q2.c

Estimated total populations of Florida counties can also be visualized:



Note: I tried to use a map here, but had issues adding the geometry information to the data.

Q3

Q3.a

Latitude and Longitude information can be extracted from the collected tweet data with `rtweet:lat_lng()`, which uses all geolocation information in a tweet to get a coordinate pair. This poses an issue if a twitter user has their location turned off, or manually set to a location different from where they are when using Twitter. Using all the geolocation information also means looking at the bounding box of the tweet, which may overlap with the area of interest, but not truly be in the area of interest.

We can solve the issue of tweet coming from other locations by subsetting our data, but it is difficult to avoid the issue of capturing tweets where the location is manually set to be within our region of interest.

Q3.b

After adding GEOID data to the tweets, **6522** tweets outside of Florida were dropped (43.8%).

Q3.c

With our tweets tagged with the appropriate county-level GEOID, we can investigate how many tweets are associated with each county:

Table 2: Number of collected tweets geo-coded to each Florida county

County	Pop. Estimate	Tweet Count
Alachua	269956	113
Bay	185287	75
Brevard	596849	147
Broward	1951260	1199
Charlotte	184998	56
Citrus	147929	15
Clay	216072	45
Collier	378488	54
Columbia	70503	19
Duval	950181	72
Escambia	315534	132
Flagler	112067	18
Hernando	190865	32
Highlands	105424	11
Hillsborough	1436888	727
Indian River	157413	68
Lake	356495	68
Lee	754610	163
Leon	292502	205
Manatee	394855	128
Marion	359977	60
Martin	160912	71
Miami-Dade	2761581	1489
Monroe	75027	32
Nassau	85832	28
Okaloosa	207269	66
Orange	1380645	830
Osceola	367990	90
Palm Beach	1485941	663
Pasco	539630	168
Pinellas	975280	536
Polk	708009	131
Santa Rosa	179349	42
Sarasota	426718	117
Seminole	467832	204
St. Johns	254261	76
St. Lucie	321128	122
Sumter	128754	19
Volusia	547538	255
Walton	71375	20

Q3.d

To ascertain the association between the number of tweets originating from a county and the county's population, we fit the simple linear model

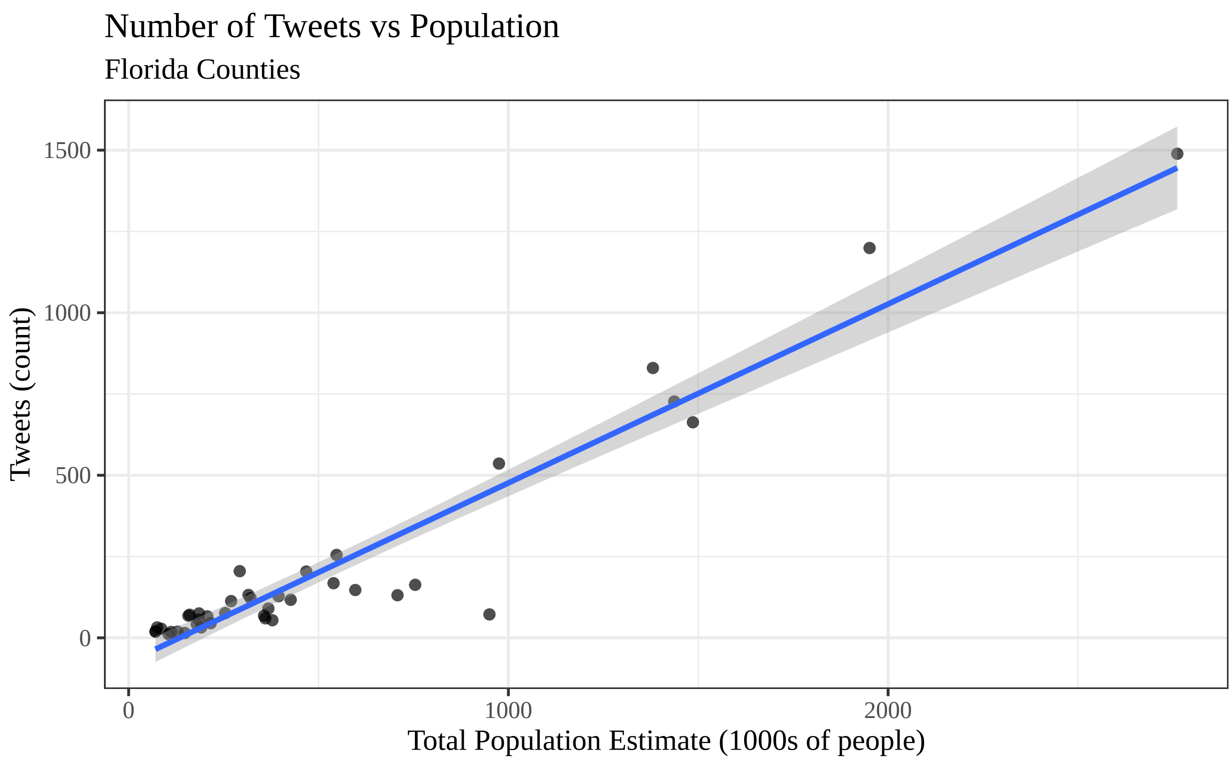
$$lm(\text{tweets} \sim \text{population})$$

which yields the parameters:

Table 3: Summary of model fitting the association between tweets and population

Term	Estimate	Std. Error	P-value	2.5% CI	97.5% CI
(Intercept)	-73.8537	20.737	0.001	-115.8337	-31.8738
population	0.0006	0.000	0.000	0.0005	0.0006

We can also visualize the model:



From the data and model, we can say that there is a statistically significant relationship between the population of a county and the number of tweets from a county. Removing the few extreme values from the data would significantly change this relationship, however, so maybe there are some underlying factors this model doesn't capture (such as tourists coming to Miami and tweeting, but not being counted as part of the population).

Appendix

```r
# Prep work ---------------------------------------------------------------

# Load libraries
library(dplyr)
library(ggplot2)
library(rtweet)
library(tidycensus)
library(tigris)


# Question 1 --------------------------------------------------------------


# Question 1a -------------------------------------------------------------

bbox_florida <- c(-87.586670, 24.256982, -79.735108, 30.983104)

# Collect Tweets
# streamed_tweets_florida <- stream_tweets(
#   q = bbox_florida,
#   timeout = (60 * 60 * 8),
#   parse = FALSE,
#   file_name = "data/rtweet_stream_florida.json",
# )
#
# tweet_tbl <- parse_stream("data/rtweet_stream_florida.json")
# saveRDS(tweet_tbl, "data/tweet_stream_florida_parsed.RDS")

tweet_tbl <- readRDS("data/tweet_stream_florida_parsed.RDS")



# Question 1b -------------------------------------------------------------

num_tweets <- nrow(tweet_tbl)

time_span_hr <- difftime(
  max(tweet_tbl$created_at), min(tweet_tbl$created_at),
  units = "hours"
  ) %>%
  as.numeric() %>%
  round(digits = 1)

ts_plot(tweet_tbl, by = "1 minutes") +
  theme_bw() +
  theme(text = element_text(family = "serif")) +
  labs(
    title = "Number of Streamed Tweets over Time",
    subtitle = paste("Location: Florida, Duration:", time_span_hr, "hours"),
    x = "Time (UTC)",
    y = "Tweets (count)"
```

```r
  )


# Question 2 ------------------------------------------------------------


# Question 2a -----------------------------------------------------------

# Load US Census API key
CENSUS_API_KEY <- readRDS("~/.uscensus_api_key.RDS")
census_api_key(CENSUS_API_KEY)


# Question 2b -----------------------------------------------------------

fla_counties <-
  get_acs(
    geography = "county",
    variables = c("Total Population" = "B01001_001"),
    year = 2018,
    state = "Florida",
    geometry = FALSE,
    survey = "acs1"
  ) %>%
  tidyr::separate(NAME, c("county", "state"), sep = ", ") %>%
  mutate(county = stringr::str_remove(county, " County")) %>%
  rename(geoid_county = GEOID, population = estimate) %>%
  select(county, state, geoid_county, population)

knitr::kable(
  fla_counties, booktabs = TRUE,
  col.names = c("County", "State", "GEOID", "Pop. estimate"),
  caption = "Florida counties and total population estimated by ACS1 2018"
)


# Question 2c -----------------------------------------------------------

ggplot(fla_counties, aes(x = county, y = population / 1000)) +
  geom_col() +
  theme_bw() +
  theme(
    text = element_text(family = "serif"),
    axis.text.x.bottom = element_text(angle = 45, hjust = 1),
    panel.grid.major.x = element_blank()
  ) +
  labs(
    title = "Distribution of Population across Florida",
    subtitle = "ACS1 2018 Population",
    x = "County",
    y = "Total Population (1000s of people)"
  )
```

```r
# Question 3 -------------------------------------------------------------


# Question 3a ------------------------------------------------------------

tweet_lat_lon_tbl <- tweet_tbl %>%
  rtweet::lat_lng() %>%
  rename(lon = lng) %>%
  select(lat, lon)


# Question 3b ------------------------------------------------------------

# Make function that won't fail when finding GEOIDs for all lat/lon pairs
safe_geolocator <- purrr::possibly(tigris::call_geolocator_latlon, NA_character_)

# Takes a while to run, so save results
# tweet_lat_lon_tbl %>%
#   purrr::pmap_chr(safe_geolocator) %>%
#   saveRDS("data/block_geoids.RDS")

block_geoids <- readRDS("data/block_geoids.RDS")

geocode_tweets <- tweet_tbl %>%
  rtweet::lat_lng() %>%
  rename(lon = lng) %>%
  mutate(
    geoid_block = block_geoids,
    geoid_county = substr(geoid_block, 1, 5)
  ) %>%
  left_join(fla_counties, by = "geoid_county") %>%
  filter(state == "Florida")

num_dropped <- num_tweets - nrow(geocode_tweets)
pct_dropped <- (num_dropped / num_tweets) * 100


# Question 3c ------------------------------------------------------------

grouped_tweets <- geocode_tweets %>%
  group_by(county, population) %>%
  summarise(tweets = n())

knitr::kable(
  grouped_tweets, booktabs = TRUE,
  col.names = c("County", "Pop. Estimate", "Tweet Count"),
  caption = "Number of collected tweets geo-coded to each Florida county"
)

# Question 3d ------------------------------------------------------------

tweet_model <- lm(tweets ~ population, data = grouped_tweets)
```

```r
model_tbl <- tweet_model %>%
  broom::tidy() %>%
  left_join(
    confint(tweet_model) %>% as_tibble(rownames = "term"),
    on = "term"
  ) %>%
  select(-statistic)

knitr::kable(
  model_tbl, booktabs = TRUE, digits = 4,
  col.names = c("Term", "Estimate", "Std. Error", "P-value", "2.5% CI", "97.5% CI"),
  caption = "Summary of model fitting the association between tweets and population"
)

ggplot(grouped_tweets, aes(x = population / 1000, y = tweets)) +
  geom_point(alpha = .7) +
  geom_smooth(method = "lm") +
  theme_bw() +
  theme(text = element_text(family = "serif")) +
  labs(
    title = "Number of Tweets vs Population",
    subtitle = "Florida Counties",
    x = "Total Population Estimate (1000s of people)",
    y = "Tweets (count)"
  )
```