# tweet_analysis

*Michael Spencer*

*10/7/2019*

## Setup

### Libraries

```
if (!require(tidyverse)) install.packages("tidyverse")
library(tidyverse)
if (!require(lubridate)) install.packages("lubridate")
library(lubridate)
if (!require(scales)) install.packages("scales")
library(scales)
```

### Parameters

```
all_tweets_pathname <- "all_tweets_clean.tsv"
filtered_tweets_pathname <- "filtered_tweets_clean.tsv"
female_names_url <- "https://5harad.com/mse125/assets/hw1/female_names.tsv.gz"
male_names_url <- "https://5harad.com/mse125/assets/hw1/male_names.tsv.gz"
```

## Load Data

```
all_tweets <-
  all_tweets_pathname %>%
  read_tsv(quote = " ")

filtered_tweets <-
  filtered_tweets_pathname %>%
  read_tsv()

female_names <-
  female_names_url %>%
  read_tsv()

male_names <-
  male_names_url %>%
  read_tsv()
```

## Functions

### find_gender function

Here we do simple string manipulation to try and clear the names of any numbers and emojis under the assumption that first names only include alpha characters. Once we have isolated estimated first names we

join our tweet data with our gender data so that we can attempt to guess the gender of users.

We make an educated guess about a user's gender by determining which gender a name occurs with more. The formula used is:

$$\frac{males\ with\ name\ x - females\ with\ name\ x}{total\ occurences\ of\ name\ x}$$

For names present in the data, this formula gives us a number from -1 to 1, with -1 indicating female, 0 indicating unkown, and 1 indicating male. If a given names generates a negative number, it is assigned female, if 0, unknown, and if a positive number, then male.

```
find_gender <- function(data) {

  data %>%

    # String manipulation to isolate first names
    mutate_if(
      is.character,
      str_replace_all,
      pattern = "[^[:alpha:]+[:space:]]", # Gets rid of non-alpha characters
      replacement = ""
    ) %>%
    mutate_if(
      is.character,
      str_extract,
      pattern = "^[:alpha:]+" # Extracts first string
    ) %>%
    mutate_if(
      is.character,
      str_to_lower # Normalizes names to lowercase for easier joining
    ) %>%

    # Gives each row a unique key so that they can later be brought back
    # together
    mutate(pair_key = as.integer(rownames(.))) %>%

    # Gathers the names to minimize the amount of processing needed when
    # inferring gender, connects each name to counts, and infers gender based
    # on the formula explained above
    gather(key = "user_type", value = "name", username, og_poster) %>%
    left_join(male_names, by = "name") %>%
    left_join(female_names, by = "name") %>%
    mutate_at(
      vars(contains("total")),
      replace_na,
      replace = 0
    ) %>%
    mutate(
      gender_p = (total_male - total_female)/(total_male + total_female),
      est_gender = case_when(
        gender_p < 0 ~ "female",
        gender_p == 0 ~ sample(c("female", "male"), replace = TRUE, size = 1),
        gender_p > 0 ~ "male",
```

```
        TRUE ~ "unknown"
      )
    ) %>%
    select(-total_male, -total_female) %>%

    # Collects newly created data and then spreads it out so that further
    # analysis can be done, keeping interacting user pairs together
    unite(col = "info", name, gender_p, est_gender) %>%
    spread(key = "user_type", value = "info") %>%
    separate(
      username,
      into = c("user_name", "user_gender_p", "user_est_gender"),
      sep = "_"
    ) %>%
    separate(
      og_poster,
      into = c("og_name", "og_gender_p", "og_est_gender"),
      sep = "_"
    ) %>%
    select(
      date, time, user_name, og_name, user_est_gender, og_est_gender,
      user_gender_p, og_gender_p
    )

}
```

## Data Prep

### Prep Gender Data

Assigning Twitter users to a given gender using only their names requires a statistical strategy. Much of this computation will be done in the following steps using the `find_gender()` function, but the first step is to identify the range of years we would like to use before aggregating the total count of each name for each gender.

Let's assume that no Twitter user is older than 90 (let's face it, not many 90 year olds are tech savvy). Let's also assume that no one younger than 5 is tweeting.

This means our Twitter users are between 5 and 90, and were thus born between 1928 and 2014. I'll filter out data to reflect that, while simultaneously aggregating the names and making them lowercase (to ensure our matching of Twitter users' names is robust to capitalization).

```
male_names <-
  male_names %>%
  filter(year >= 1928 & year <= 2014) %>%
  mutate(name = name %>% str_to_lower()) %>%
  group_by(name) %>%
  summarise(total_male = sum(count, na.rm = TRUE))

female_names <-
  female_names %>%
  filter(year >= 1928 & year <= 2014) %>%
  mutate(name = name %>% str_to_lower()) %>%
```

```
  group_by(name) %>%
  summarise(total_female = sum(count, na.rm = TRUE))
```

### Prep Tweet Data

We use the `find_gender()` function written above to prep our tweet data for homophily and time analysis.

This approach is limited in that it assumes first names appear first in multi-element strings; it's implementing a simple binary classifer (of sorts) which is not guaranteed to be perfect; and we also have no way of classifying names with non-english characters.

```
filtered_tweets_gen <- find_gender(filtered_tweets)
all_tweets_gen <- find_gender(all_tweets)
```

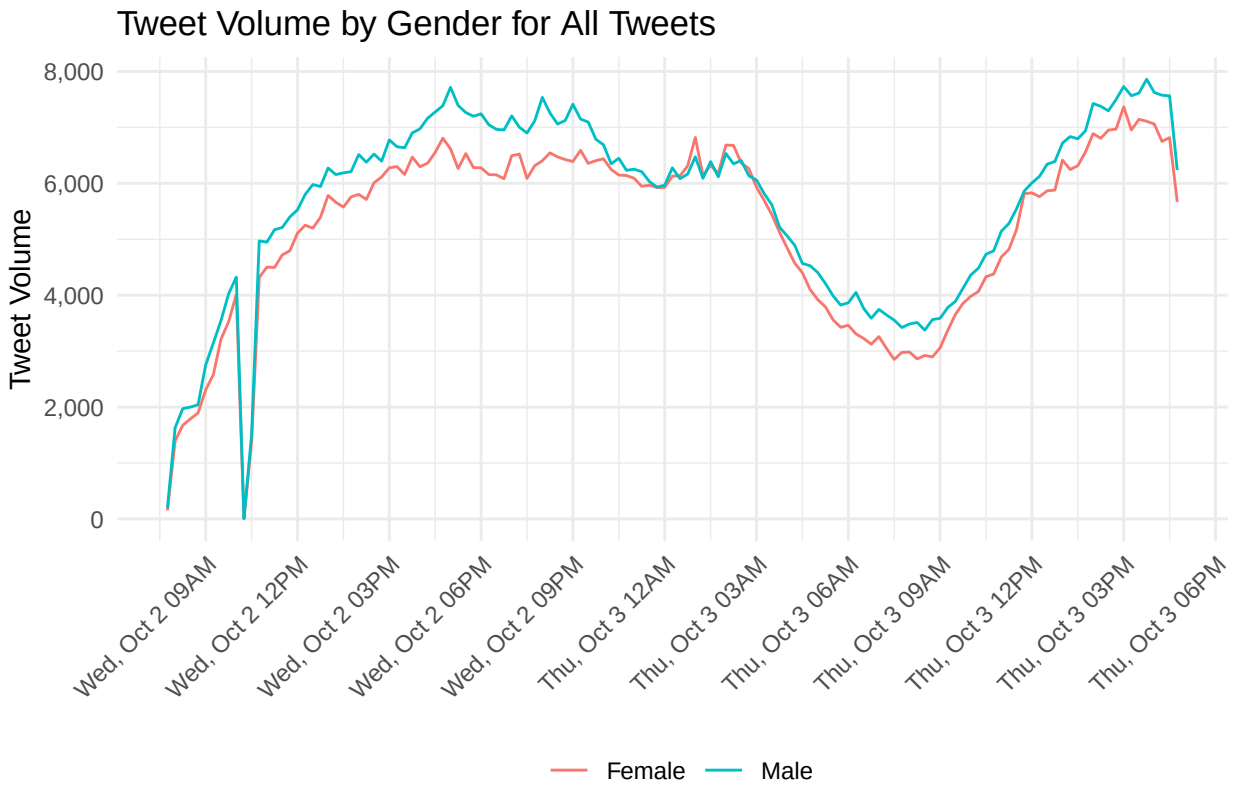
## Analysis

### Gender Trends Over Time

*Tweet Volume by Gender for All Tweets*

```
all_tweets_plot <-
  all_tweets_gen %>%
  filter(user_est_gender != "unknown") %>%
  count(date, time, user_est_gender, name = "tweets") %>%
  arrange(date, time, user_est_gender) %>%
  transmute(
    user_est_gender = user_est_gender %>% str_to_title(),
    tweets,
    datetime = ymd_hms(str_c(date, time))
  ) %>%
  ggplot(aes(datetime, tweets, color = user_est_gender)) +
  geom_line() +
  scale_x_datetime(
    breaks = "3 hours",
    date_labels = "%a, %b%e %I%p"
  ) +
  scale_y_continuous(
    labels = comma_format()
  ) +
  labs(
    title = "Tweet Volume by Gender for All Tweets",
    x = NULL,
    y = "Tweet Volume",
    caption = "Data collected from Twitter over 24 hours."
  ) +
  theme_minimal() +
  theme(
    legend.title = element_blank(),
    legend.position = "bottom",
    axis.text.x.bottom = element_text(angle = 45, hjust = 1)
  )
```

```
all_tweets_plot
```



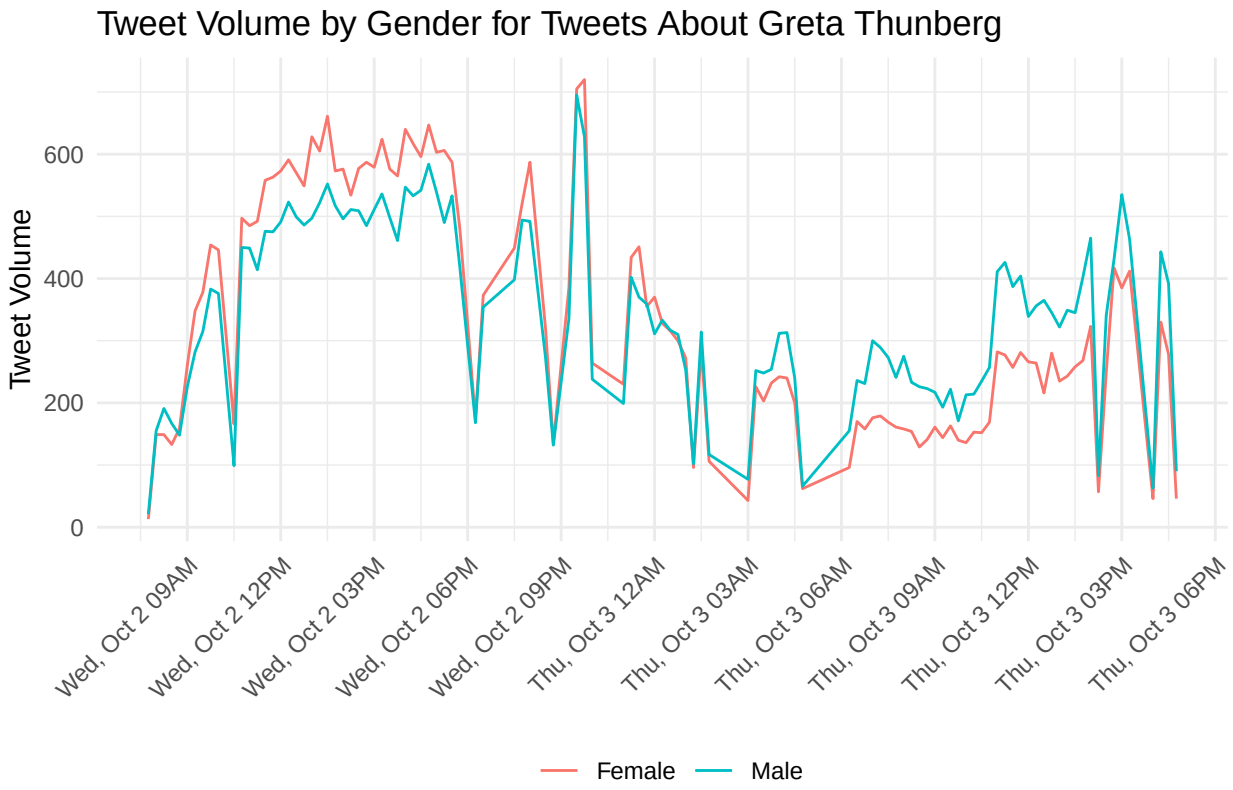

```
ggsave(plot = all_tweets_plot, file = "all_tweets_plot.pdf", width = 8, height = 5)
```

*Tweet Volume by Gender for Tweets About Greta Thunberg*

```
filtered_tweets_plot <-
  filtered_tweets_gen %>%
  filter(user_est_gender != "unknown") %>%
  count(date, time, user_est_gender, name = "tweets") %>%
  arrange(date, time, user_est_gender) %>%
  transmute(
    user_est_gender = user_est_gender %>% str_to_title(),
    tweets,
    datetime = ymd_hms(str_c(date, time))
  ) %>%
  ggplot(aes(datetime, tweets, color = user_est_gender)) +
  geom_line() +
  scale_x_datetime(
    breaks = "3 hours",
    date_labels = "%a, %b%e %I%p"
  ) +
  scale_y_continuous(
    labels = comma_format()
  ) +
```

```
  labs(
    title = "Tweet Volume by Gender for Tweets About Greta Thunberg",
    x = NULL,
    y = "Tweet Volume",
    caption = "Data collected from Twitter over 24 hours."
  ) +
  theme_minimal() +
  theme(
    legend.title = element_blank(),
    legend.position = "bottom",
    axis.text.x.bottom = element_text(angle = 45, hjust = 1)
  )

filtered_tweets_plot
```




```
ggsave(plot = filtered_tweets_plot, file = "filtered_tweets_plot.pdf", width = 8, height = 5)
```

Both trends above, for all tweets and tweets about Greta, followed a similar pattern. It appears high tweet volumes picked up around 11 AM and continued until midnight, with quieter traffic in the morning hours likely due to users' sleep schedules.

### Homophily

*Expectations*

Our goal here is to see if users of one gender retweet users of the same gender disproportionately more often than they do users of a opposite gender. To test this we will use a test similar to the one used above, in which we'll compare the expected value to the actual value. The expected value assumes that if no homophily exists, a given gender will retweet males and females proportional to their occurences in the retweet data (ie, for the set of all retweets, if males represent 20% of the original posters, then a given gender will retweet males 20% of the time). The formulas are as follows:

Expected % of retweets of gender x =

$$\frac{\#\ of\ retweets\ of\ gender\ x}{total\ \#\ of\ retweets}$$

Actual % of retweets of gender x by gender y =

$$\frac{\#\ of\ retweets\ of\ gender\ x\ by\ y}{total\ \#\ of\ retweets\ by\ y}$$

```
exp_prob <-
  all_tweets_gen %>%
  filter(user_est_gender != "unknown" & og_est_gender != "unknown") %>%
  count(og_est_gender, name = "total") %>%
  mutate(exp_prop = total / sum(total)) %>%
  arrange(og_est_gender) %>%
  pull(exp_prop)

female_exp <- exp_prob[1]
male_exp <- exp_prob[2]
```

- Expected proportion of female retweets: 38.78%
- Expected proportion of male retweets: 61.22%

*Gender Homophily in All Tweets*

To determine homophily, we calculate the actual proportions separately for each gender. We can then compare these actuals with the expected proportions to see if gender homophily exists. Namely, for a given gender, if the actual proportion of females retweeted is less then the expected proportion, that gender retweets females less often than expected. If this phenomena occured amongst male users, we would say that they demonstrate homophily because they are retweeting females less (and thus males more) than expected. The same is true of the converse situation.

In the charts below, if `user_est_gender` and `homophily_rt_more` are the same, homophily exist for that gender.

```
all_homophily_results <-
  all_tweets_gen %>%
  filter(user_est_gender != "unknown" & og_est_gender != "unknown") %>%
  count(user_est_gender, og_est_gender, name = "total") %>%
  spread(key = "og_est_gender", value = "total") %>%
  rename("female_rts" = female, "male_rts" = male) %>%
  mutate(
    female_act = female_rts / (female_rts + male_rts),
    male_act = male_rts / (female_rts + male_rts),
    fem_diff_from_exp = female_act - female_exp,
    homophily_rt_more = case_when(
```

```
      fem_diff_from_exp < 0 ~ "male",
      fem_diff_from_exp > 0 ~ "female",
      TRUE ~ "none"
    )
  )

knitr::kable(all_homophily_results, format = "markdown", digits = 4)
```

| user_est_gender | female_rts | male_rts | female_act | male_act | fem_diff_from_exp | homophily_rt_more |
|---|---|---|---|---|---|---|
| female | 73015 | 89374 | 0.4496 | 0.5504 | 0.0618 | female |
| male | 56397 | 114932 | 0.3292 | 0.6708 | -0.0586 | male |

For all tweets, females retweet females 6.18% more than expected than if homophily didn't exist. Male users retweeted male users 5.86% more than expected than if homophily didn't exist. For both genders, homophily exists to some extent.

*Gender Homophily in Tweets About Greta Thunberg*

```
filtered_homophily_results <-
  filtered_tweets_gen %>%
  filter(user_est_gender != "unknown" & og_est_gender != "unknown") %>%
  count(user_est_gender, og_est_gender, name = "total") %>%
  spread(key = "og_est_gender", value = "total") %>%
  rename("female_rts" = female, "male_rts" = male) %>%
  mutate(
    female_act = female_rts / (female_rts + male_rts),
    male_act = male_rts / (female_rts + male_rts),
    fem_diff_from_exp = female_act - female_exp,
    homophily_rt_more = case_when(
      fem_diff_from_exp < 0 ~ "male",
      fem_diff_from_exp > 0 ~ "female",
      TRUE ~ "none"
    )
  )

knitr::kable(filtered_homophily_results, format = "markdown", digits = 4)
```

| user_est_gender | female_rts | male_rts | female_act | male_act | fem_diff_from_exp | homophily_rt_more |
|---|---|---|---|---|---|---|
| female | 5807 | 3999 | 0.5922 | 0.4078 | 0.2044 | female |
| male | 5565 | 6468 | 0.4625 | 0.5375 | 0.0747 | female |

For tweets about Greta Thunberg, female users retweeted female users 20.44% more than expected than if homophily didn't exist. Male users retweeted female users 7.47% more than expected than if homophily didn't exist. In this case, homophily exists for females and is much stronger than in the standard case above. Meanwhile, for males tweeting about Greta Thunberg, homophily does not exist. These changes are likely due to the topic chosen - Greta Thunberg - and the fact that she is a female environmental activist. Many of those talking about her may be females, and those engaging with tweets about her may also be females. This, in addition to the fact that the likely majority of tweets regarding Greta are posted by female users, makes it easy to understand why we see the results we do.