W241_final_project

```
library(data.table)

library(sandwich)
library(lmtest)
library(stargazer)
library(foreign)
library(AER)
library(ggplot2)
library(patchwork)

robust_se <- function(mod, type = 'HC3') {
  sqrt(diag(vcovHC(mod, type)))
  }
```

We started with 391 subreddits where we created posts by using both gender-neutral (control) and female (treatment) usernames. Of those 391 subreddits, 205 existed for at least 24 hours and were able to be analyzed. The rest 186 subreddits have at least one control or treatment post removed, which we define as attrition. We had 47.6% attrition in total.

```
d <- fread('./final_data_r.csv')

setnames(d, "1day_upvote","upvote")
setnames(d, "1day_comment", "comment")

head(d)
```

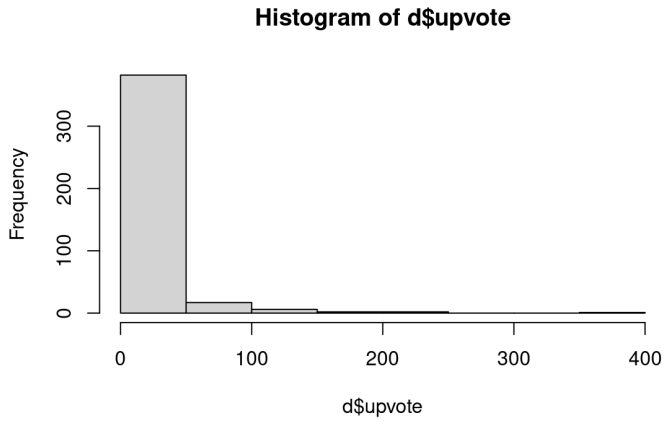
...	ID	subreddits	comments_per_day	random_assign_rank	NotUsedReason
<int>	<int>	<chr>	<int>	<int>	
1	3851	0xPolygon	110	478	
2	1027	49ers	542	745	
3	3193	ac_newhorizons	143	414	
4	654	Accounting	839	230	
5	654	Accounting - second time	839	230	
6	472	AirForce	1150	466	

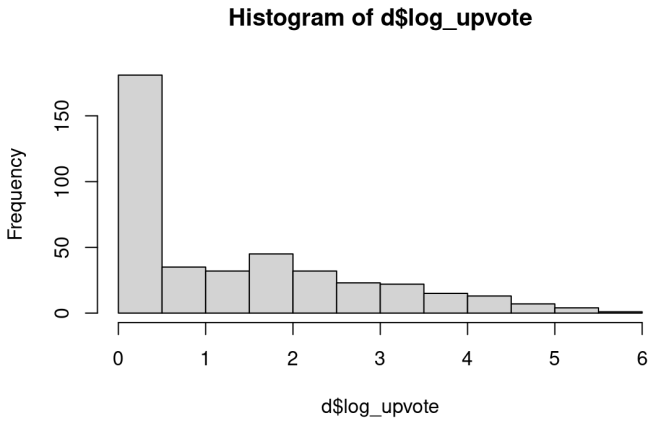
6 rows | 1-6 of 20 columns

Data exploration

The outcome that is being measured is whether the number of upvotes would vary by the account gender. First, we want to look at the number of `upvote`. As seen below, while most posts received less than 20 upvotes, we also observed that a few posts were quite popular and received hundreds of upvotes. Considering that high-

upvotes posts would attract more views and potentially more upvotes, we decided to log transform the number of upvotes. Below histograms demonstrate the original upvotes data and log transformed data.

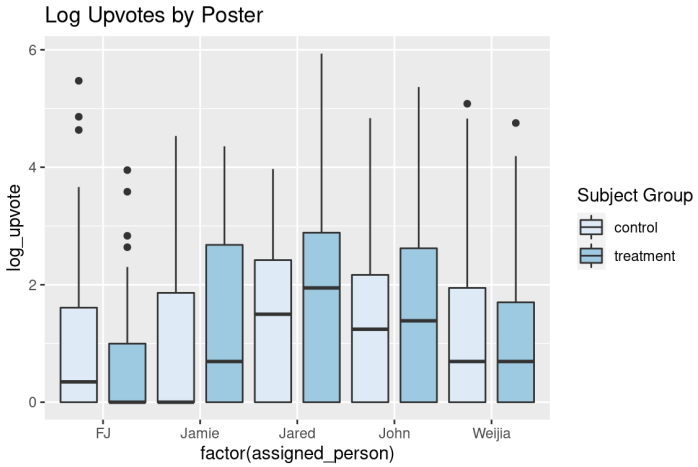




There are also a few important covariates that we want to explore before adding to the model:

`assigned_person` : we are also concerned with the potential heterogeneous treatment effects from the poster. To collect as many samples as possible, each of the team members selected the subreddits that are relatively familiar or interesting to him/her. This might inevitably result in a heterogeneous treatment effect. As seen below, when breaking down the log upvotes by poster, we can see some deviation between control and treatment groups by different posters. For the same reason as mentioned above, we would not remove outliers.

```
ggplot(data=d) +
    geom_boxplot( aes(x=factor(assigned_person), y=log_upvote, fill=factor(group)), posi
tion=position_dodge(1)) +
    theme_minimal() +
    scale_fill_brewer('Set3') +
    guides(fill=guide_legend(title="Subject Group")) +
    theme_update(plot.title = element_text(hjust = 0.5)) +
    ggtitle("Log Upvotes by Poster")
```



`treatment_order` : though we randomized the sequence of posting and allowed wash-out period to minimize the impact from time, we would still include this covariate to explore if any impact.As seen below, the effect from treatment order is not distinct – the treatment has a mixed effect on the log upvotes. Based on this finding, we will not include treatment order to the regression model.

```
ggplot(d, aes(x=log_upvote, fill=factor(treatment_order))) +
           geom_density(alpha=.3) +
            theme_update(plot.title = element_text(hjust = 0.5)) +
            ggtitle("Log Upvotes by Treatment Order (0: control first; 1: treatment firs
t)")
```

Regression

As we adopted within-subject design, we apply the fixed effect of subreddit, which has unique value for each sample. When using the linear regression model to estimate the mean difference in log upvotes between control (gender-neutral username) and treatment (female username) group, we structured models into 3 levels – the simplest model regressed log upvotes on the treatment (username gender) and the fixed effect of subreddits; the intermediate model by adding poster (`assigned_person`); and the improved model by adding interaction between poster and treatment.

```
mod_1 <- d[ , lm(log_upvote ~ group + subreddits)]
mod_2 <- d[ , lm(log_upvote ~ group + assigned_person + subreddits)]
mod_3 <- d[ , lm(log_upvote ~ group + assigned_person + group * assigned_person + subred
dits)]


stargazer(mod_1, mod_2, mod_3,
          type='text',
          se = list(
                      robust_se(mod_1),
                      robust_se(mod_2),
                      robust_se(mod_3)),
          column.labels = c('simplest model','intermediate model','improved model'),
          title = 'Username Gender Impact on Post Scores',
          omit = 'subreddits',
          add.lines = list(c("Subreddits Fixed Effects", "Yes", "Yes","Yes")),
          font.size = "tiny",
          column.sep.width = "1pt"
          )
```

```
##
## Username Gender Impact on Post Scores
## ================================================================================
==========================
##                                                                  Dependent variable:
##                                         ------------------------------------------------
--------------------------
##                                                                       log_upvote
##                                              simplest model        intermediate model
improved model
##                                                   (1)                      (2)
(3)
## --------------------------------------------------------------------------------
--------------------------
## grouptreatment                                   0.038                    0.038
-0.398
##                                                 (0.148)                  (0.148)
(0.465)
##
## assigned_personJamie                                                     4.149***
3.826***
##                                                                          (0.086)
(0.407)
##
## assigned_personJared                                                      -0.000
-0.368
##                                                                          (0.038)
(0.469)
##
## assigned_personJohn                                                      3.473***
3.245***
##                                                                          (0.903)
(0.994)
##
## assigned_personWeijia                                                     0.000
-0.137
##                                                                          (0.038)
(0.412)
##
## grouptreatment:assigned_personJamie
0.646
##
(0.571)
##
## grouptreatment:assigned_personJared
0.735
##
(0.553)
##
## grouptreatment:assigned_personJohn
0.456
##
```

```
(0.548)
##
## grouptreatment:assigned_personWeijia
0.274
##
(0.558)
##
## Constant                                    4.130***                  -0.019
0.199
##                                             (0.111)                  (0.079)
(0.371)
##
## ------------------------------------------------------------------------------------
-------------------------
## Subreddits Fixed Effects                      Yes                       Yes
Yes
## Observations                                  410                       410
410
## R2                                           0.672                     0.672
0.680
## Adjusted R2                                  0.430                     0.430
0.434
## Residual Std. Error                    1.098 (df = 235)          1.098 (df = 235)
1.094 (df = 231)
## F Statistic                         2.770*** (df = 174; 235) 2.770*** (df = 174; 23
5) 2.759*** (df = 178; 231)
## ====================================================================================
=========================
## Note:                                                                          *
p<0.1; **p<0.05; ***p<0.01
```

Breakdown by each poster

For an easy read on poster's impact on upvotes, we further broke down the improved model by each poster and the result is shown below. There is no statistically significant impact of poster on the treatment effect. And the coefficient is in line with Table 2. For example, John's posts in the treatment (female username) group received an 0.058% higher upvotes than posts in his control (gender neutral) group. This is the same effect as calculated from Table 2: -0.398 + 0.456 = 0.058.

```
mod_john <- d[assigned_person == 'John' , lm(log_upvote ~ group + subreddits )]
mod_jamie <- d[assigned_person == 'Jamie', lm(log_upvote ~ group  + subreddits)]
mod_jared <- d[assigned_person == 'Jared', lm(log_upvote ~ group  + subreddits)]
mod_weijia <- d[assigned_person == 'Weijia' , lm(log_upvote ~ group + subreddits )]
mod_fj <- d[assigned_person == 'FJ' , lm(log_upvote ~ group + subreddits)]

stargazer(mod_john, mod_jamie, mod_jared, mod_weijia, mod_fj,
          type='text',
          se = list(
                      robust_se(mod_john),
                      robust_se(mod_jamie),
                      robust_se(mod_jared),
                      robust_se(mod_weijia),
                      robust_se(mod_fj)),
          column.labels = c('John','Jamie','Jared','Weijia','FJ'),
          title = 'Username Gender Impact on Post Scores by Poster',
          omit = 'subreddits',
          add.lines = list(c("Subreddits Fixed Effects", "Yes", "Yes","Yes", "Yes","Yes"
)),
          font.size = "tiny",
          column.sep.width = "0.01pt"
          )
```

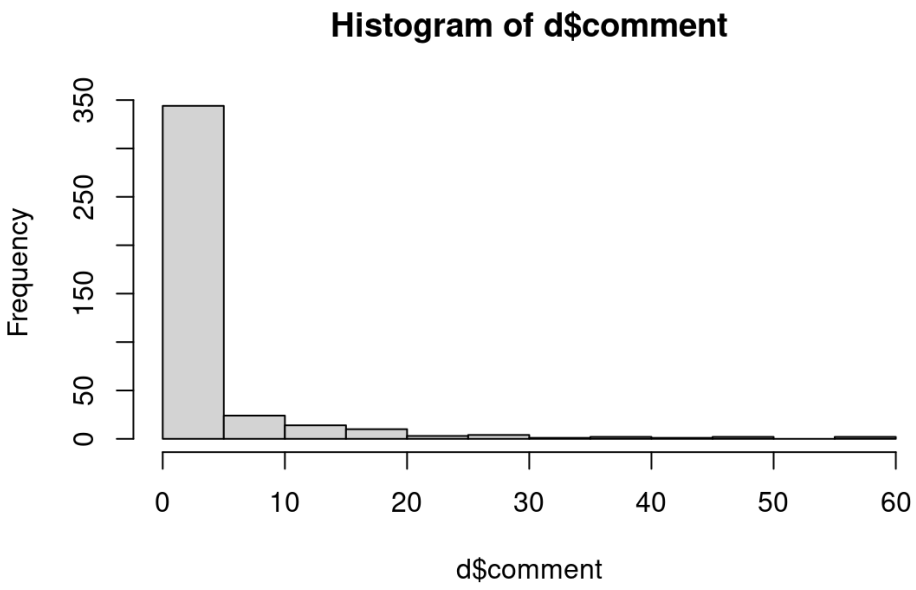
```
##
## Username Gender Impact on Post Scores by Poster
## ================================================================================================
================================================
##                                                                                 Dependent vari
able:
##                          ---------------------------------------------------------------------
-----------------------------------------------
##                                                                                        log_upvote
##                                    John                  Jamie                    Jared
Weijia                 FJ
##                                    (1)                    (2)                     (3)
(4)                   (5)
## ------------------------------------------------------------------------------------------------
-----------------------------------------------
## grouptreatment                    0.058                  0.248                   0.338
-0.124               -0.398
##                                  (0.290)                (0.331)                 (0.298)
(0.308)               (0.465)
##
## Constant                          -0.029                4.025***                4.055**
*                0.408*                0.199
##                                  (0.145)                (0.166)                 (0.160)
(0.247)               (0.269)
##
## ------------------------------------------------------------------------------------------------
-----------------------------------------------
## Subreddits Fixed Effects           Yes                    Yes                     Yes
Yes                   Yes
## Observations                       84                     68                      96
94                   68
## R2                                0.514                  0.811                   0.751
0.673                 0.562
## Adjusted R2                       0.349                  0.648                   0.537
0.366                 0.136
## Residual Std. Error         1.138 (df = 62)        0.912 (df = 36)        1.055 (df =
51)        1.025 (df = 48)       1.318 (df = 34)
## F Statistic              3.118*** (df = 21; 62) 4.986*** (df = 31; 36) 3.502*** (df =
44; 51) 2.194*** (df = 45; 48) 1.320 (df = 33; 34)
## ================================================================================================
================================================
## Note:
*p<0.1; **p<0.05; ***p<0.01
```

Regression on log_comments and comments+upvotes

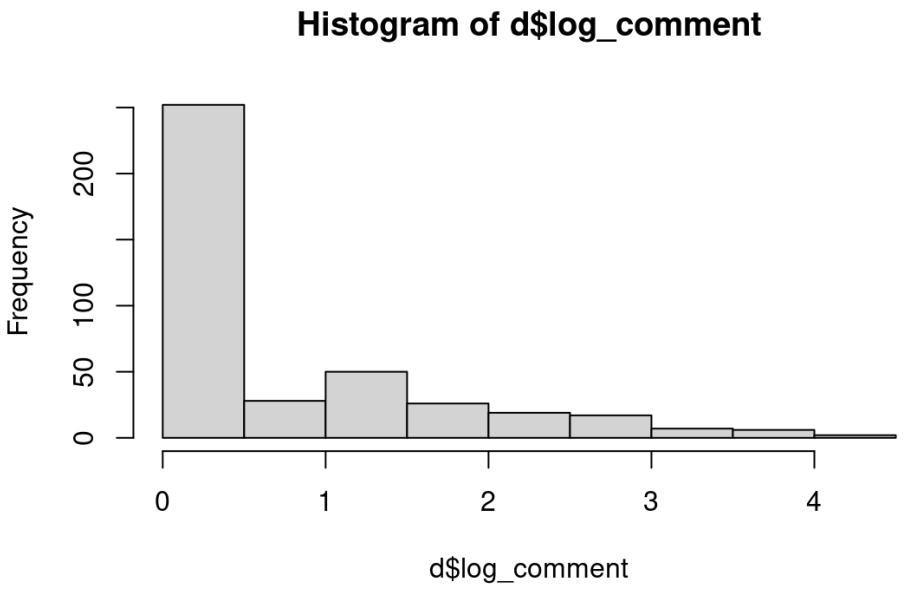
We also measured the number of comments and subset the model on each poster. The model that uses comments as the outcome variable was not informative, as there is no statistically significant impact from any of the treatment or covariate variables. For simplicity, we will keep the results with comments and the sum in the

code report.

```
hist(d$comment)
```




```
d[ , log_comment := ifelse(comment ==0, 0, log(comment))]
hist(d$log_comment)
```




```
d[ , sum_comm_upvote := upvote + comment]
d[ , log_sum_comm_upvote := ifelse(sum_comm_upvote ==0, 0, log(sum_comm_upvote))]
```

```
mod_4 <- d[ , lm(log_comment ~ group + subreddits)]
mod_5 <- d[ , lm(log_comment ~ group + assigned_person + subreddits)]
mod_6 <- d[ , lm(log_comment ~ group + assigned_person + group * assigned_person + subre
ddits)]


stargazer(mod_4, mod_5, mod_6,
          type='text',
          se = list(
                      robust_se(mod_4),
                      robust_se(mod_5),
                      robust_se(mod_6)),
          #column.labels = c('simplest model','intermediate model','improved model'),
          title = 'Username Gender Impact on Comments',
          omit = 'subreddits',
          add.lines = list(c("Subreddits Fixed Effects", "Yes", "Yes","Yes")),
          align = TRUE
          )
```

```
##
## Username Gender Impact on Comments
## ====================================================================================
==========================
##                                                                    Dependent variable:
##                                        -----------------------------------------------
--------------------------
##                                                                        log_comment
##                                                (1)                      (2)
(3)
## ------------------------------------------------------------------------------------
--------------------------
## grouptreatment                                0.006                    0.006
-0.263
##
##
## assigned_personJamie                                                   2.740
2.583
##
##
## assigned_personJared                                                   -0.000
-0.196
##
##
## assigned_personJohn                                                    2.454
2.262
##
##
## assigned_personWeijia                                                  -0.000
-0.102
##
##
## grouptreatment:assigned_personJamie
0.315
##
##
## grouptreatment:assigned_personJared
0.392
##
##
## grouptreatment:assigned_personJohn
0.386
##
##
## grouptreatment:assigned_personWeijia
0.203
##
##
## Constant                                      2.738                    -0.003
0.132
##
##
```

```
## ------------------------------------------------------------------------------------
--------------------------
## Subreddits Fixed Effects                          Yes                        Yes
Yes
## Observations                                      407                        407
407
## R2                                               0.606                      0.606
0.611
## Adjusted R2                                      0.310                      0.310
0.306
## Residual Std. Error                        0.835 (df = 232)          0.835 (df = 232)
0.837 (df = 228)
## F Statistic                              2.048*** (df = 174; 232) 2.048*** (df = 174; 23
2) 2.008*** (df = 178; 228)
## ====================================================================================
==========================
## Note:                                                                          *
p<0.1; **p<0.05; ***p<0.01
```

```
mod_7 <- d[ , lm(log_sum_comm_upvote ~ group + subreddits)]
mod_8 <- d[ , lm(log_sum_comm_upvote ~ group + assigned_person + subreddits)]
mod_9 <- d[ , lm(log_sum_comm_upvote ~ group + assigned_person + group * assigned_person
+ subreddits)]


stargazer(mod_7, mod_8, mod_9,
          type='text',
          se = list(
                      robust_se(mod_7),
                      robust_se(mod_8),
                      robust_se(mod_9)),
          #column.labels = c('simplest model','intermediate model','improved model'),
          title = 'Username Gender Impact on Comments & Upvotes',
          omit = 'subreddits',
          add.lines = list(c("Subreddits Fixed Effects", "Yes", "Yes","Yes"))
          )
```

```
##
## Username Gender Impact on Comments & Upvotes
## ================================================================================================
==========================
##                                                                 Dependent variable:
##                                     ---------------------------------------------------------
--------------------------
##                                                                 log_sum_comm_upvote
##                                              (1)                        (2)
(3)
## ------------------------------------------------------------------------------------------------
--------------------------
## grouptreatment                             0.003                      0.003
-0.555
##
##
## assigned_personJamie                                                  4.380
3.935
##
##
## assigned_personJared                                                  -0.000
-0.423
##
##
## assigned_personJohn                                                   3.784
3.483
##
##
## assigned_personWeijia                                                 0.347
0.145
##
##
## grouptreatment:assigned_personJamie
0.890
##
##
## grouptreatment:assigned_personJared
0.846
##
##
## grouptreatment:assigned_personJohn
0.602
##
##
## grouptreatment:assigned_personWeijia
0.402
##
##
## Constant                                   4.379                      -0.001
0.278
##
##
```

```
## ------------------------------------------------------------------------------------
-------------------------
## Subreddits Fixed Effects                          Yes                          Yes
Yes
## Observations                                      407                          407
407
## R2                                               0.682                        0.682
0.694
## Adjusted R2                                      0.444                        0.444
0.454
## Residual Std. Error                        1.081 (df = 232)          1.081 (df = 232)
1.071 (df = 228)
## F Statistic                              2.863*** (df = 174; 232) 2.863*** (df = 174; 23
2) 2.899*** (df = 178; 228)
## ====================================================================================
==========================
## Note:                                                                          *
p<0.1; **p<0.05; ***p<0.01
```