Alumni Survey Bounce and Open Rate

```r
library(data.table)
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

Load Data

```r
# list of people
dt_roster <- read_csv('data/alumni_scrape.csv', col_types=cols()) %>%
  select(Email = email, Name = name, Cohort = cohort, Gender = gender,
         Year_Graduation = year_graduation,
         Block = blocking_genderyear_assignments)

# attrition, bounced, round 1 post-send merged status
r1_g1_bounce <- read_csv('data/round1_0702_group1_0702_0941.csv', col_types=cols()) %>%
  select(Email, MStatus = `Merge status`) %>%
  filter(MStatus == 'BOUNCED')
r1_g2_bounce <- read_csv('data/round1_0702_group2_0702_1028.csv', col_types=cols()) %>%
  select(Email, MStatus = `Merge status`) %>%
  filter(MStatus == 'BOUNCED')
r1_g3_bounce <- read_csv('data/round1_0702_group3_0702_1055.csv', col_types=cols()) %>%
  select(Email, MStatus = `Merge status`) %>%
  filter(MStatus == 'BOUNCED')

# final status, sent, opened
r1_g1_status <- read_csv('data/round1_0702_group1_0708_1033.csv', col_types=cols()) %>%
  select(Email, MStatus = `Merge status`)
r1_g2_status <- read_csv('data/round1_0702_group2_0708_1220.csv', col_types=cols()) %>%
  select(Email, MStatus = `Merge status`)
r1_g3_status <- read_csv('data/round1_0702_group3_0708_1122.csv', col_types=cols()) %>%
  select(Email, MStatus = `Merge status`)
r2_status <- read_csv('data/round2_0709_all_0716_1117.csv', col_types=cols()) %>%
  select(Email, MStatus = `Merge status`)
r3_status <- read_csv('data/round3_0716_all_0722_1046.csv', col_types=cols()) %>%
```

```r
  select(Email, MStatus = `Merge status`)
r4_status <- read_csv('data/round4_0722_all_0723_1131.csv', col_types=cols()) %>%
  select(Email, MStatus = `Merge status`)
r5_status <- read_csv('data/round5_0723_all_0724_1219.csv', col_types=cols()) %>%
  select(Email, MStatus = `Merge status`)

# outcome, Qualtrics completion
dt_qualtrics <- read_csv('data/Qualtrics_downloaded_20210726.csv', col_types=cols()) %>%
  select(Email = RecipientEmail, Duration = `Duration (in seconds)`,
         Long = LocationLongitude, Lat = LocationLatitude) %>%
  slice(3:n()) %>%
  mutate(Duration = as.numeric(Duration),
         Long = as.numeric(Long),
         Lat = as.numeric(Lat))

# emails to exclude
ex_emails <- c('deveshkhandelwal@berkeley.edu', 'rahosbach@berkeley.edu',
               'tgao2020@berkeley.edu', 'mirza2020@berkeley.edu')
```

Combine Data

```r
dt_agg <-
  # roster does not contain us
  dt_roster %>%
  select(Email) %>%

  # bounce
  left_join(bind_rows(r1_g1_bounce, r1_g2_bounce, r1_g3_bounce) %>%
              add_column(bounce = 1) %>%
              select(-MStatus),
            by = 'Email') %>%
  replace(is.na(.), 0) %>%

  # non-compliance: email_sent status for all 5 rounds
  left_join(
    # EMAIL_SENT not balanced.
    #bind_rows(r1_g1_status %>% filter(MStatus == 'EMAIL_SENT'),
    #          r1_g2_status %>% filter(MStatus == 'EMAIL_SENT'),
    #          r1_g3_status %>% filter(MStatus == 'EMAIL_SENT')) %>%
    #  filter(!Email %in% ex_emails) %>%
    #inner_join(r2_status %>% filter(MStatus == 'EMAIL_SENT') %>% select(Email),
    #           by='Email') %>%

    # EMAIL_SENT is balanced in r3, r4 and r5
    dt_roster %>% select(Email) %>% add_column(MStatus='EMAIL_SENT') %>%

      inner_join(r3_status %>% filter(MStatus == 'EMAIL_SENT') %>% select(Email),
                 by='Email') %>%
      inner_join(r4_status %>% filter(MStatus == 'EMAIL_SENT') %>% select(Email),
                 by='Email') %>% # 4 & 5 are the same?
      inner_join(r5_status %>% filter(MStatus == 'EMAIL_SENT') %>% select(Email),
                 by='Email') %>%
      add_column(non_complier = 1) %>%
```

```r
      select(-MStatus),
    by = 'Email') %>%
  replace(is.na(.), 0) %>%

  # compliers: people who opened at least 1 round
  mutate(complier = 1 - (bounce + non_complier)) %>%

  # completion: qualtrics records
  left_join(dt_qualtrics %>% select(Email) %>% add_column(completion=1), by = 'Email') %>%
  replace(is.na(.), 0) %>%

  # override some incorrect YAMM data
  mutate(non_complier = if_else(completion == 1, 0, non_complier),
         complier = if_else(completion == 1, 1, complier)) %>%

  # meta data
  left_join(dt_roster, by = 'Email') %>%
  left_join(dt_qualtrics, by = 'Email')
```

Output Data

```r
# output
write_csv(dt_agg, 'data/agg_table.csv')
```

Quick Look

```r
dt_agg %>%
  select(Block, bounce, non_complier, complier, completion) %>%
  gather(measure, value, -Block) %>%
  group_by(Block, measure) %>%
  summarise(value = sum(value)) %>%
  spread(measure, value) %>%
  select(Block, bounce, non_complier, complier, completion) %>%
  mutate(all = bounce + non_complier + complier, .before=2)
```

```
## `summarise()` has grouped output by 'Block'. You can override using the `.groups` argument.

## # A tibble: 3 x 6
## # Groups:   Block [3]
##   Block   all bounce non_complier complier completion
##   <dbl> <dbl>  <dbl>        <dbl>    <dbl>      <dbl>
## 1     1   369     42          124      203         66
## 2     2   368     39          136      193         78
## 3     3   368     45          136      187         85
```