# w41_tweets_facts_sentiment_FN

Arriaga, Hue, Khural, Moran

July 31, 2021

# The raw data

First we will load the data.

```
# d<- data.table::fread('../data/raw/response_tp_07_31.csv')
# d<- data.table::fread('../data/raw/response_tn_07_31.csv')
# d<- data.table::fread('../data/raw/response_fp_07_31.csv')
d<- data.table::fread('../data/raw/response_fn_07_31.csv')
```

# Data cleaning

We will first review:

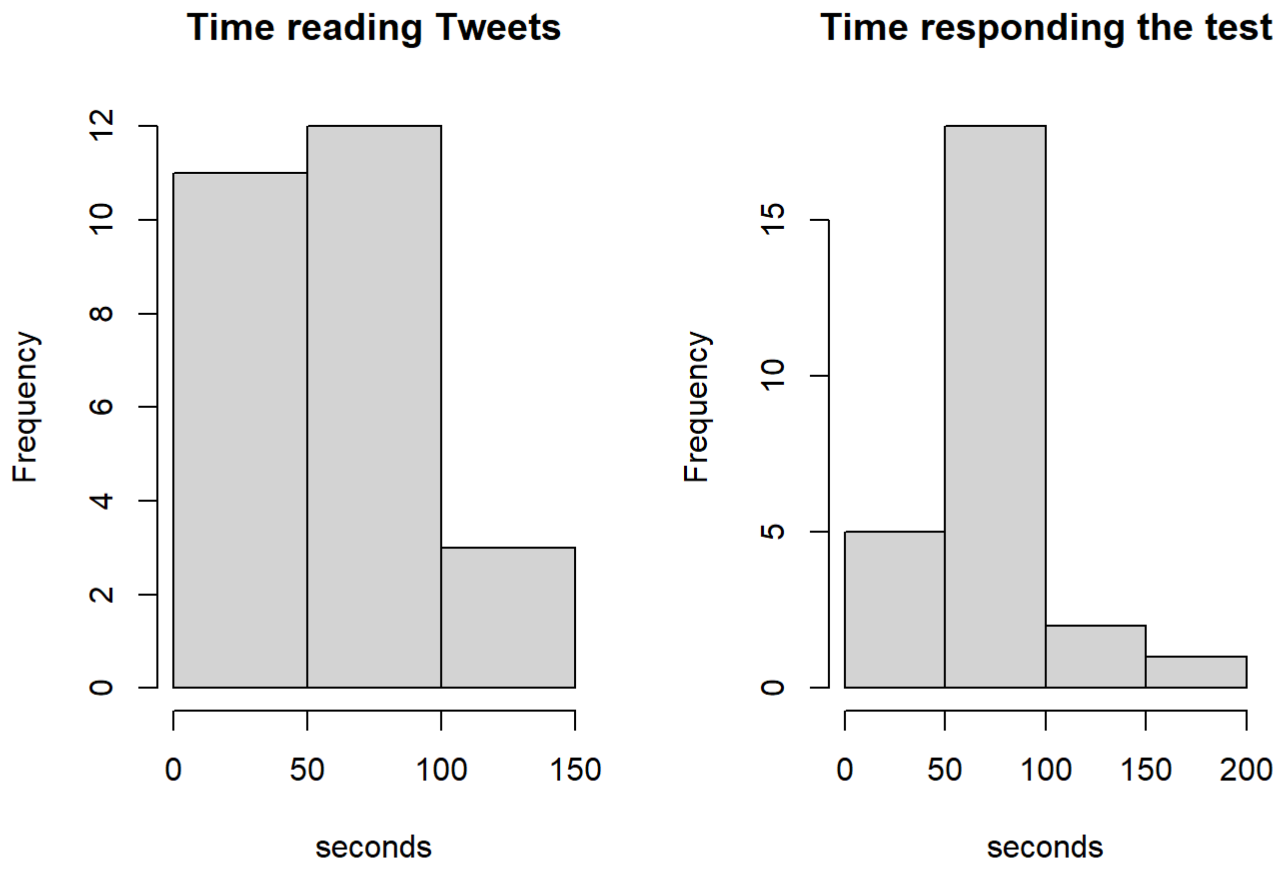
- Eliminate empty columns
- Rename columns
- Update fields that didn’t export correctly (missing data)
- Check for attrition or duplicates
- Check for fake answers: response time too quick plus attention check question wrong.
- Convert data to correct type (all are strings)

```
## [1] "We found 0  non-compliers"
```

```
## [1] "We found a  0 % attrition."
```

We can also see a histogram of time looking at the tweets and time responding the test.

```
par(mfrow=c(1,2))

h1 <- hist(d$tweet_submit_time,
     main = "Time reading Tweets",
     xlab = "seconds",
     breaks = length(d$tweet_submit_time)/5)

h2<- hist(d$test_submit_time,
     main = "Time responding the test",
     xlab = "seconds",
     breaks = length(d$test_submit_time)/5)
```



Data seems reasonably normal. Now let’s remove variable that are not of interest.

```
# Keep only columns of interest
colnames(d)
```

```
##  [1] "start_date"          "end_date"            "survey_duration"
##  [4] "finished"            "finished_date"       "id"
##  [7] "email"               "gender"              "age"
## [10] "education"           "tweet_first_click_time" "tweet_last_click_time"
## [13] "tweet_submit_time"   "tweet_click_count"   "math_q1"
## [16] "math_q2"             "test_first_click_time" "test_last_click_time"
## [19] "test_submit_time"    "test_click_count"    "stimulus"
## [22] "georgians"           "energy"              "soccer"
## [25] "fauci"               "pollution"           "election"
```

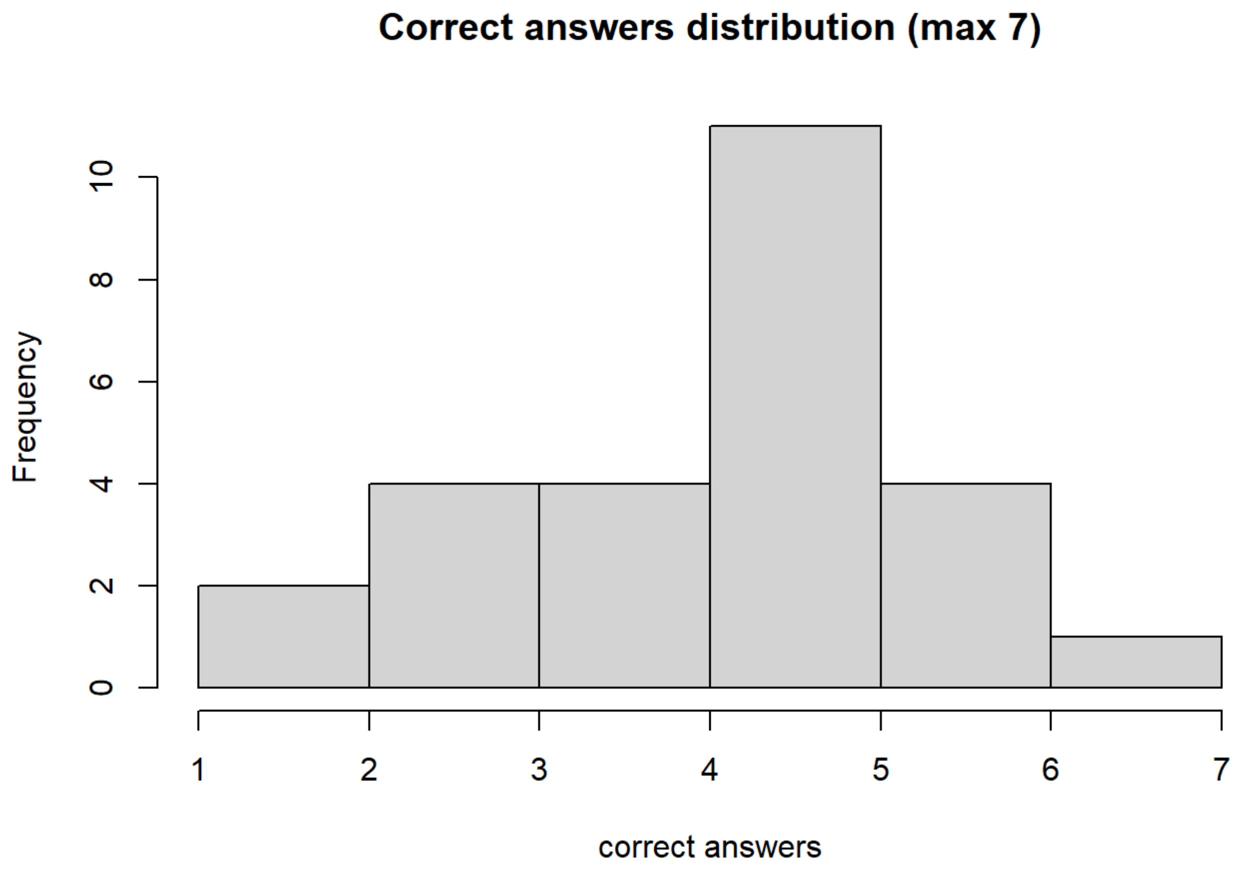
```
cols_to_remove <- c("start_date","end_date",
                    "finished","finished_date","id","email",
                    "tweet_first_click_time","tweet_last_click_time","tweet_click_count",
                    "test_first_click_time","test_last_click_time","test_click_count")

cols_to_drop_ix<-which(colnames(d) %in% cols_to_remove)

d[,cols_to_drop_ix]<- list(NULL)
```

Let’s include the treatment columns to the data.

```
## IMPORTANT Adding factors
# d$truth<- as.factor("fact")
d$truth<- as.factor("fake")
# d$sentiment<- as.factor("positive")
d$sentiment<- as.factor("negative")

# Correct answers for test: FACT
# d[ , bin_stimulus := ifelse(stimulus == "Green House Gases", yes = 1, no = 0)]
# d[ , bin_georgians := ifelse(georgians == "None of the above", yes = 1, no = 0)]
# d[ , bin_energy := ifelse(energy == "Natural Gas", yes = 1, no = 0)]
# d[ , bin_soccer := ifelse(soccer == "Unknown", yes = 1, no = 0)]
# d[ , bin_fauci := ifelse(fauci == "National Geographic Society", yes = 1, no = 0)]
# d[ , bin_pollution := ifelse(pollution == "6th largest", yes = 1, no = 0)]
# d[ , bin_election := ifelse(election == "2024", yes = 1, no = 0)]


# # Correct answers for test: FAKE
d[ , bin_stimulus := ifelse(stimulus == "Green House Gases", yes = 1, no = 0)]
d[ , bin_georgians := ifelse(georgians == "70", yes = 1, no = 0)]
d[ , bin_energy := ifelse(energy == "Wind", yes = 1, no = 0)]
d[ , bin_soccer := ifelse(soccer == "Pfizer vaccine", yes = 1, no = 0)]
d[ , bin_fauci := ifelse(fauci == "None of the above", yes = 1, no = 0)]
d[ , bin_pollution := ifelse(pollution == "1st largest", yes = 1, no = 0)]
d[ , bin_election := ifelse(election == "2024", yes = 1, no = 0)]

d$total_correct <- d%>% select(c("bin_stimulus", "bin_georgians","bin_energy","bin_soccer","bin_
fauci","bin_pollution", "bin_election")) %>% rowSums()
```

We can see the distribution of correct answers.

```
h3 <- hist(d$total_correct,
     main = "Correct answers distribution (max 7)",
     xlab = "correct answers",
     breaks = length(d$total_correct)/4)
```



Now we can save the file for future reference and stacking.

```
# Save to CSV
# write.csv(d,"../data/interim/tn_data.csv", row.names = FALSE)
# write.csv(d,"../data/interim/fp_data.csv", row.names = FALSE)
write.csv(d,"../data/interim/fn_data.csv", row.names = FALSE)
```