Generate clean US data

ddd

11/7/2020

- go back and delete dashes
- Mutual friends Tom edit
- decide what to do with unclassifiable race posts

NF

```
to.char.list <- c("participant_id", "post_tenure", "nf_group..y.n.", "event..y.n.")
```

```
#Bring in all the data and clean it

getwd()
```

```
## [1] "/Users/diagdavenport/Desktop/Synced Research/Ludwig/FB/Lockdown"
```

```
participant.log <- read.csv("Data/RA Protocol/India/COMPLETE Mastersheet_India - Participant log.csv")

# ignore the phantom row if there's no RA_id
participant.log <- participant.log %>% filter(RA_id != "")

# remove commas and safely convert factor to numeric
  participant.log$total_friends <- as.numeric(gsub(",", "", participant.log$total_friends))

# deal with gender variations
  participant.log$subject_gender_RA <- tolower(participant.log$subject_gender_RA)

# deal with race variations
    participant.log$subject_religion_RA <- tolower(participant.log$subject_religion_RA)

newsfeed.df <- read.csv("Data/RA Protocol/India/COMPLETE Mastersheet_India - NF.csv")

# ignore phantom rows
newsfeed.df <- newsfeed.df[!is.na(newsfeed.df$preference..1.7.) & !is.na(newsfeed.df$participant_id), 1

# clean up names
  colnames(newsfeed.df) <- c("participant.id",
                   "nf.order",
                   "preference",
                   "post.tenure",
                   "tenure.units",
                   "human",
                   "nf.group",
                   "event",
```

```
                    "poster.nf.caste.ra",
                    "poster.nf.religion.ra",
                    "poster.nf.gender",
                    "nf.ra.notes")

  newsfeed.df$poster.nf.caste.ra <- tolower(newsfeed.df$poster.nf.caste.ra)
  newsfeed.df$poster.nf.religion.ra <- tolower(newsfeed.df$poster.nf.religion.ra)
  newsfeed.df$poster.nf.gender <- tolower(newsfeed.df$poster.nf.gender)


newsfeed.df <- newsfeed.df %>% filter(participant.id != "")

# some ppl said and a "half"
newsfeed.df$preference <- round(newsfeed.df$preference)

# Now let's merge the nf data to the log data

m.newsfeed.df <- merge(newsfeed.df, participant.log, by.x = "participant.id", by.y = "participant_id")

num.nf.ids <- length(unique(newsfeed.df$participant.id))

m.newsfeed.df$religion.in.group <- (m.newsfeed.df$poster.nf.religion.ra == m.newsfeed.df$subject_religi
m.newsfeed.df$gender.in.group <- m.newsfeed.df$poster.nf.gender == m.newsfeed.df$subject_gender_RA

write.csv(m.newsfeed.df, file = "Temp/Clean India Data NF.csv")
```

```
pymk.df <- read.csv("Data/RA Protocol/India/COMPLETE Mastersheet_India - PYMK.csv")
pymk.df <- pymk.df[!is.na(pymk.df$familiarity..1.7.) & !is.na(pymk.df$participant_id), 1:8] # to be con

# clean up names
  colnames(pymk.df) <- c("participant.id",
                    "pymk.order",
                    "familiarity",
                    "mutual.friends",
                    "poster.pymk.caste.ra",
                    "poster.pymk.religion.ra",
                    "poster.pymk.gender",
                    "pymk.ra.notes")

  pymk.df$poster.pymk.caste.ra <- tolower(pymk.df$poster.pymk.caste.ra)
  pymk.df$poster.pymk.religion.ra <- tolower(pymk.df$poster.pymk.religion.ra)
  pymk.df$poster.pymk.gender <- tolower(pymk.df$poster.pymk.gender)

# drop more phantom rows

  pymk.df <- pymk.df %>% filter(participant.id != "")

  participant.log <- participant.log %>% filter(participant_id != "")

  # some ppl said and a "half"
  pymk.df$familiarity <- round(pymk.df$familiarity)

num.log.ids <- length(unique((participant.log$participant_id)))
num.pymk.ids <- length(unique((pymk.df$participant.id)))
```

```
# Same deal with pymk, merge to the log data

m.pymk.df <- merge(pymk.df, participant.log, by.x = "participant.id", by.y = "participant_id")

assert_that(abs(num.pymk.ids - num.log.ids) < 12)
```

```
## [1] TRUE
```

```
m.pymk.df$religion.in.group <- m.pymk.df$poster.pymk.religion.ra == m.pymk.df$subject_religion_RA
m.pymk.df$gender.in.group <- m.pymk.df$poster.pymk.gender == m.pymk.df$subject_gender_RA

write.csv(m.pymk.df, file = "Temp/Clean India Data PYMK.csv")
```

Qualtrics data

```
nf.data.raw <- m.newsfeed.df

nf.data <- nf.data.raw %>% select(participant.id,
                                  subject_gender_RA,
                                  subject_religion_RA)

nf.data <- unique(nf.data)
nf.data <- nf.data %>% dplyr::rename(uni.id = participant.id)

qualtrics1 <- read.csv("Data/Self-Assessments/India Qualtrics.csv")

combined.qualtrics <- qualtrics1

combined.qualtrics$date <- date(combined.qualtrics$StartDate)
```

```
## Warning: tz(): Don't know how to compute timezone for object of class factor;
## returning "UTC". This warning will become an error in the next major version of
## lubridate.
```

```
combined.qualtrics <- combined.qualtrics %>% filter(Status == "IP Address" &
                                                    Finished == "True" &
                                                    Q2 != "")

combined.qualtrics$primary.id <- combined.qualtrics$subject_id

merged <- merge(combined.qualtrics, nf.data, by.x = "primary.id", by.y = "uni.id", all.x = F, all.y = T

merged <- merged %>% dplyr::rename(self.race = Q3,
                                   self.race.free = Q4,
                                   gender = Q5,
                                   age = Q6,
                                   education = Q7,
                                   covid.usage = Q9,
                                   last.login = Q10,
                                   pre.covid.usage = Q11)

merged$self.race <- as.character(merged$self.race)
```

```r
merged$self.race <- ifelse(grepl(",",merged$self.race),"Two or more",merged$self.race)

dim(merged)
```

```
## [1] 199  34
```

```r
write.csv(merged, file = "Temp/Clean India qualtrics data.csv")
```