01-data_cleaning-survey1

Jiaheng Li, Anni Lin, Yuechen Shen, Yuxin Yang

28/10/2020

Preamble

Purpose: Prepare and clean the survey data downloaded from voterstudygroup.org

Author: Jiaheng Li, Anni Lin, Yuechen Shen, Yuxin Yang

Data: 22 October 2020

Contact: rohan.alexander@utoronto.ca [PROBABLY CHANGE THIS ALSO!!!!]

License: MIT

Pre-requisites:

- Need to have downloaded the data from X and save the folder that you're

interested in to inputs/data

- Don't forget to gitignore it!

Workspace setup

```
library(haven)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
setwd("~/Downloads/UofT 2018-2022/Academic/STA304/STA304_PS3")
# Read in the raw data (You might need to change this if you use a different dataset)
raw_survey_data <- read_dta("ns20200625.dta")

# Add the labels
raw_survey_data <- labelled::to_factor(raw_survey_data)
```

```
# Just keep some variables
reduced_survey_data <-
  raw_survey_data %>%
  dplyr::select(interest,
         registration,
         vote_2016,
         vote_intention,
         vote_2020,
         ideo5,
         employment,
         foreign_born,
         gender,
         census_region,
         hispanic,
         race_ethnicity,
         household_income,
         education,
         state,
         congress_district,
         age)
```

What else????

Maybe make some age-groups?

Maybe check the values?

Is vote a binary? If not, what are you going to do?

```
reduced_survey_data<-reduced_survey_data %>%
  mutate(vote_trump = ifelse(vote_2020=="Donald Trump", 1, 0)) %>%
  mutate(vote_biden = ifelse(vote_2020=="Joe Biden", 1, 0)) %>%
  #Convert state abbreviations to names
  mutate(state_name = state.name[match(state, state.abb)]) %>%
  # select variables in interest
  dplyr::select(vote_2016, gender, age, race_ethnicity, employment, state_name, vote_trump, vote_biden)
  na.omit()

reduced_survey_data$employment <- ifelse(reduced_survey_data$employment=="Full-time employed", "Employed
                                  ifelse(reduced_survey_data$employment=="Homemaker", "Not in labor forc
                                  ifelse(reduced_survey_data$employment=="Retired", "Not in labor force
                                  ifelse(reduced_survey_data$employment=="Unemployed or temporarily on l
                                  ifelse(reduced_survey_data$employment=="Part-time employed", "Employed
                                  ifelse(reduced_survey_data$employment=="Permanently disabled", "Not in
                                  ifelse(reduced_survey_data$employment=="Student", "Student",
                                  ifelse(reduced_survey_data$employment=="Self-employed", "Employed",
                                  ifelse(reduced_survey_data$employment=="Other:", "Other",
                        NA  )))))))))

reduced_survey_data$race_ethnicity <- ifelse(reduced_survey_data$race_ethnicity=="Asian (Asian Indian)"
                                      ifelse(reduced_survey_data$race_ethnicity=="Asian (Chinese)", "Eas
                                      ifelse(reduced_survey_data$race_ethnicity=="Asian (Filipino)", "O
                                      ifelse(reduced_survey_data$race_ethnicity=="Asian (Japanese)", "Ea
                                      ifelse(reduced_survey_data$race_ethnicity=="Asian (Korean)", "East
                                      ifelse(reduced_survey_data$race_ethnicity=="Asian (Vietnamese)",
                                      ifelse(reduced_survey_data$race_ethnicity=="Asian (Other)", "Other
                                      ifelse(reduced_survey_data$race_ethnicity=="Pacific Islander (Nati
                                      ifelse(reduced_survey_data$race_ethnicity=="Pacific Islander (Guam
                                      ifelse(reduced_survey_data$race_ethnicity=="Pacific Islander (Samo
                                      ifelse(reduced_survey_data$race_ethnicity=="Pacific Islander (Othe
                                      ifelse(reduced_survey_data$race_ethnicity=="Some other race", "Oth
                                      ifelse(reduced_survey_data$race_ethnicity=="White", "White",
                                      ifelse(reduced_survey_data$race_ethnicity=="Black, or African Amer
                                      ifelse(reduced_survey_data$race_ethnicity=="American Indian or Ala
                        NA  )))))))))))))))
```

```
View(reduced_survey_data)
```

Saving the survey/sample data as a csv file in my working directory

```
write_csv(reduced_survey_data, "/Users/yangyuxin/Downloads/UofT 2018-2022/Academic/STA304/STA304_PS3/su
```