ANA 515 Assignment 2

Hai Yen Nguyen

2023-06-14

Section 1: description of the data

Dataset content

The data measures various personal information, including education, race, gender, income level, and vote frequency. The dataset was collected initially through polling conducted with Ipsos' KnowledgePanel, a probability-based online panel selected to be representative of the US population. The poll was conducted from Sept. 15 to Sept. 25 among a sample of U.S. citizens that oversampled young, Black, and Hispanic respondents, with 8,327 respondents. It then was weighted according to general population benchmarks for U.S. citizens from the U.S. Census Bureau's Current Population Survey March 2019 Supplement. Following that, the voter file company Aristotle connected respondents to a voter file to better understand their voting history, utilizing the panelist's first name, surname name, zip code, and the first eight characters of their address, if appropriate, using the National Change of Address program.

With the dataset, I am hoping to figure out why many Americans don't vote and the common profiles of non-voters. By analyzing the datasets, I can acquire a complete understanding of the causes contributing to low voter. This research can be used to develop more effective campaign to urge more Americans to vote.

Dataset format

The dataset is in a delimited file format and is saved in CSV (Comma-Separated Values) file format. Delimiter data refers to the specific character or sequence of characters used to separate or delimit individual data elements within a dataset or file.

In a CSV file, each line represents a row of data, and commas separate the values within each line. This makes it a flat file with variable-width fields, as the length of each field can vary depending on the data present in that particular column.

Section 2: read the data into R & assigns it to a dataframe object

```r
#Using read.csv, which is base R function, to read data from csv file from a
URL

url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/non-
voters/nonvoters_data.csv"
fulldataset <- read.csv(url)
```

Section 3: clean the data

3A: Call dplyr library

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

3B: Cleaning and organize to include informative data

```r
#Subseting the data using "select" function
selected <- select(fulldataset, educ, race, gender, income_cat,
voter_category)

#Renaming two columns using "rename" function
selected_renamed <- rename(selected, income = income_cat, vote =
voter_category)

#Organizing the data in vote column using "arrange" function
selected_renamed_arrange <- arrange(selected_renamed, vote)

#Filter just the rarely/never in vote column from selected_renamed dataset
rarely_never <- filter(selected_renamed, vote=="rarely/never")
```

Section 4: characteristics of the data

This dataframe has 5836 rows and 5 columns. The names of the columns and a brief description of each are in the table below:

```r
#Include a table using kable from the knitr package with 2 columns:
#(1) the column name in the dataframe and
#(2) a very brief description of what each column measures



library(knitr)
columns_summary <- data.frame(
Columns = c(colnames(selected_renamed)),
Description = c(
"highest education level of respondents",
"ethinicity of respondents",
"sexual identities of respondents",
"ranged of annual income of respondents",
"voting frequency")
)

kable(columns_summary, caption = "Voters Selected Renamed Table")
```

Voters Selected Renamed Table

Columns	Description
educ	highest education level of respondents
race	ethinicity of respondents
gender	sexual identities of respondents
income	ranged of annual income of respondents
vote	voting frequency

Section 5: summary statistics

5A: Pick three columns of the dataframe using subset function.

```r
data_pick3 <- select(fulldataset, weight, Q2_1, Q2_2)
```

5B: Summary table

```r
#Use a summary function to get the following summaries of the three columns
#Include min, max, mean, and missing values
#The summary statistics are assigned to an "summarytable" object
Summarytable<-summary(data_pick3)

#Prints the output summary
print(Summarytable)
```

```
##      weight            Q2_1              Q2_2       
##  Min.   :0.2298   Min.   :-1.000   Min.   :-1.000  
##  1st Qu.:0.7932   1st Qu.: 1.000   1st Qu.: 1.000  
##  Median :0.9676   Median : 1.000   Median : 2.000  
##  Mean   :0.9910   Mean   : 1.246   Mean   : 1.705  
##  3rd Qu.:1.1696   3rd Qu.: 1.000   3rd Qu.: 2.000  
##  Max.   :3.0386   Max.   : 4.000   Max.   : 4.000  
```

5C: Number of missing values

```r
sum(is.na(data_pick3))
```

```
## [1] 0
```