Problem Set 2

Due February 7, 2023

Instructions

- Read all of these instructions closely.
- This problem set is due Tuesday, February 7, 2023 at 4pm.
- Submit files via Github:
  1. the .Rmd (R Markdown) file
  2. the knitted .pdf file
  3. anything else the particular problem set might require
- Use a copy of this file, perhaps with your name or initials appended to the file name, to write your answers to the questions. You'll see there is a designated space where your answers should begin.
- Knitting the .Rmd file to a .pdf file *as you work* will ensure your code runs without errors and is working how you expect. Knit early and often. You've already read the instruction that a knitted .pdf is required when you submit.
- Per the syllabus, I will not accept any late work. Keep in mind the two lowest problem set scores are dropped. Turn in what you have.

Overview

This problem set uses a subset of expenditures data for all campaigns and PACs available from Open Secrets for 2002 cycle. The reduced dataset is available here. (While not the point of this question, I encourage you to visit the link to see how data shared on Dropbox can be imported directly into R via its url.)

Before you being this question, you should familiarize yourself with the variables. The codebook is available here.

```
expenditures_url <- "https://www.dropbox.com/s/z6gw9lvve6jogi5/Expends2002.txt?raw=1"
df <- read.csv(expenditures_url)
```

Question 1–Working with logicals

Use R code to answer the following questions.

1a

Are any `Amount` values missing?

```
#code here
```

1b

How many observations are for refunds?

Hint: Read the codebook carefully for the `Amount` variable.

```
#code here
```

1c

What are the row indices for observations that indicate an amount spent of $1,000,000 or more?

```
#code here
```

1d

Double check that all of the `Cycle` values equal 2002.

```
#code here
```

1e

How many observations are for "Club for Growth" OR the "Madison Project" OR the "Republican National Cmte"?

```
#code here
```

Question 2–Working with dataframes

2a

Using R functions, describe the following properties of the `df` object: class, dimensions, columnnames, rownnames, and anything else you think is pertinent.

```
#code here
```

2b

For the `TransID` variable, change its column name to `Useless_Var`.

Bonus: If you want to challenge yourself, try to write code that is flexible, meaning it will work correctly if `TransID` is the 3rd variable, 20th variable, or any position in the dataframe.

```
#code here
```

2c

Remove the variables `Useless_Var` and `Source` from the dataframe.

Bonus: Make this code flexible as well.

```
#code here
```

2d

The variable `State` has many obvious errors. I've created the variable `StateWrong` with `NA` placeholders. Recode `StateWrong` to be `TRUE` if the `State` variable contains an error or a missing value, and `FALSE` otherwise.

Hint: We did a recoding exercise in the inclass activity.

Bonus: Try to use the `%in%` function. We haven't used it in class yet. It is similar to `==`. The syntax is `x %in% y`, which assesses each value of vector `x` and asks, is it equal to any of the values in vector `y`? I've included a simple example below.

```
df$StateWrong <- NA

# Example of %in%
# In words: For each letter in the alphabet,
```

```r
# check if it is it equal (TRUE) or not (FALSE)
# to A, D, or F
x <- LETTERS
y <- c("A", "D", "F")
x %in% y
```

```
##  [1]  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE
```

2e

Using the `StateWrong` variable, report how many observations in the dataset have a wrong or missing value. Then remove these observations. Confirm that you've removed the correct number of rows by checking the dimensions of the data.

```r
#code here
```

2f

Create the variable in the dataframe called `Payroll`. It should be a logical indicating whether the `Descrip` variable contains the string "payroll" *regardless* of capitalization. Report the number of `TRUE` values in this variable.

Hint: Use the `grepl` function and read the helpfile closely.

```r
#code here
```

2g

Write a function named `sum_state_exp` that takes one character argument called `state_code`. The function should return the total amount of expenditures in given state.

```r
# Write function

# After writing the function, run it for IA, IL, and CA
# sum_state_exp(state_code = "IA")
# sum_state_exp(state_code = "IL")
# sum_state_exp(state_code = "CA")
```