Problem Set 3

Data Wrangling

[YOUR NAME]

Due Date: 2024-02-02

Getting Set Up

Open `RStudio` and create a new RMarkDown file (`.Rmd`) by going to `File -> New File -> R Markdown...`. Accept defaults and save this file as `[LAST NAME]_ps3.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[LAST NAME]_ps3.Rmd` file. Then change the `author: [Your Name]` to your name.

We will be using the `MI2020_ExitPoll.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/MI2020_ExitPoll.Rds).

All of the following questions should be answered in this `.Rmd` file. There are code chunks with incomplete code that need to be filled in.

This problem set is worth 8 total points, plus two extra credit points. The point values for each question are indicated in brackets below. To receive full credit, you must have the correct code. In addition, some questions ask you to provide a written response in addition to the code.

You are free to rely on whatever resources you need to complete this problem set, including lecture notes, lecture presentations, Google, your classmates…you name it. However, the final submission must be complete by you. There are no group assignments. To submit, compiled the completed problem set and upload the PDF file to Brightspace on Friday by midnight. Also note that the TAs and professors will not respond to Campuswire posts after 5PM on Friday, so don't wait until the last minute to get started!

Good luck!

*Copy the link to ChatGPT you used here: _________________

Question 0

Require `tidyverse` and an additional package called `labelled` (remember to `install.packages("labelled")` if you don't have it yet) and load the `MI2020_ExitPoll.Rds` data to an object called `MI_raw` . (Tip: use the `read_rds()` function with the link to the raw data.)

```
# INSERT CODE HERE
```

Question 1 [2 points]

Create a new object called `MI_clean` that contains only the following variables:

- AGE10

- SEX
- PARTYID
- EDUC18
- PRMSI20
- QLT20
- LGBT
- BRNAGAIN
- LATINOS
- QRACEAI
- WEIGHT

and then list which of these variables contain missing data recorded as `NA` . How many respondents were not asked certain questions?

```
MI_clean <- MI_raw %>%
  select() # Select the requested variables
```

```
## Error in MI_raw %>% select(): could not find function "%>%"
```

```
# Identify which have missing data recorded as NA
# INSERT CODE HERE
```

> Write answer here.

Question 2 [2 points]

Are there **unit non-response** data in the `AGE10` variable? If so, how are they recorded? What about the `PARTYID` variable? How many people refused to answer both of these questions?

```
# INSERT CODE HERE
```

> Write answer here.

Question 3 [2 points]

Let's create a new variable called `preschoice` that converts `PRSMI20` to a character. To do this, install the `labelled` package if you haven't already, then use the `to_character()` function from the `labelled` package. Now `count()` the number of respondents who reported voting for each candidate. How many respondents voted for candidate Trump in 2020? How many respondents refused to tell us who they voted for?

```
# INSERT CODE HERE
```

> Write answer here.

Question 4 [1 point]

Now do the same for the `QLT20` variable, the `AGE10` variable, and the `LGBT` variable. For each variable, make the character version `Qlty` for `QLT20`, `Age` for `AGE10`, and `Lgbt_clean` for `LGBT`. Now, for each of these new variables (including `preschoice` from the previous question), replace the unit non-response label with `NA`. (**HINT**: Use `grepl()` or `str_match()` to make it easier.)

```
# QLT20
MI_clean <- MI_clean %>%
  mutate(Qlty = , # Create new variable with text
         Age = ,  # Create new variable with text
         Lgbt_clean = ) # Create new variable with text
```

```
## Error in MI_clean %>% mutate(Qlty = , Age = , Lgbt_clean = ): could not find function "%>%"
```

```
MI_clean <- MI_clean %>%
  mutate(Qlty = ifelse(grepl("SUBSTRING FOR UNIT NON-RESPONSE",Qlty),NA,Qlty), # Replace unit non-response with NA
         Lgbt_clean = ifelse(grepl("SUBSTRING FOR UNIT NON-RESPONSE",Lgbt_clean),NA,Lgbt_clean), # Replace unit non-response with NA
         Age = ifelse(grepl("SUBSTRING FOR UNIT NON-RESPONSE",Age),NA,Age), # Replace unit non-response with NA
         preschoice = ifelse(grepl("SUBSTRING FOR UNIT NON-RESPONSE",preschoice),NA,preschoice)) # Replace unit non-response with NA
```

```
## Error in MI_clean %>% mutate(Qlty = ifelse(grepl("SUBSTRING FOR UNIT NON-RESPONSE", : could not find function "%>%"
```

Question 5 [1 point]

What proportion of women supported Trump? What proportion of LGBTQ-identifying respondents supported Trump?

```
# INSERT CODE HERE
```

> Write answer here.

Extra Credit [2 points]

Plot the relationship between Trump support and gender.

```
# INSERT CODE HERE
```