Problem Set 3

Data Wrangling

[YOUR NAME]

Due Date: 2024-09-13

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_F2024/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 5 total points, plus 1 extra credit point. 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. Instructions for how to compile the output as a PDF can be found in Problem Set 0 (https://github.com/jbisbee1/DS1000_F2024/blob/main/Psets/ds1000_pset_0.pdf) and in this gif tutorial (https://github.com/jbisbee1/DS1000_F2024/blob/main/Psets/save_as_pdf.gif).

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.)

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
require(labelled)
```

```
## Loading required package: labelled
```

```
MI_raw <- read_rds('https://github.com/jbisbee1/DS1000_F2024/raw/main/data/MI2020_ExitPoll.rds')
```

Question 1 [1 point]

What is the unit of analysis in this dataset? How many variables does it have? How many observations?

> The unit of analysis is a person. There are 63 variables and 1,231 observations. [Rubric: 0 points if not attempted. 0.5 points if the unit of analysis is not an individual person responding to a survey (although fine if it is a voter, a person in Michigan, etc.). 0.75 points if the number of variables is wrong or the number of observations is wrong.]

Question 2 [1 point]

This has too much information that we don't care about. 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(AGE10,SEX,PARTYID,EDUC18,PRSMI20,QLT20,LGBT,BRNAGAIN,LATINOS,QRACEAI,WEIGHT) # Select t
he requested variables

summary(MI_clean) # Identify which have missing data recorded as NA
```

```
##      AGE10             SEX          PARTYID          EDUC18         PRSMI20
##  Min.   : 1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :0.00
##  1st Qu.: 6.000   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.00
##  Median : 8.000   Median :2.00   Median :2.000   Median :3.000   Median :1.00
##  Mean   : 8.476   Mean   :1.53   Mean   :2.236   Mean   :3.288   Mean   :1.63
##  3rd Qu.: 9.000   3rd Qu.:2.00   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:2.00
##  Max.   :99.000   Max.   :2.00   Max.   :9.000   Max.   :9.000   Max.   :9.00
##
##      QLT20            LGBT          BRNAGAIN        LATINOS
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :3.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :2.956   Mean   :2.224   Mean   :1.907   Mean   :2.175
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
##  NA's   :616     NA's   :615     NA's   :615
##     QRACEAI          WEIGHT
##  Min.   :1.000   Min.   :0.1003
##  1st Qu.:1.000   1st Qu.:0.3775
##  Median :1.000   Median :0.8020
##  Mean   :1.572   Mean   :1.0000
##  3rd Qu.:1.000   3rd Qu.:1.4498
##  Max.   :9.000   Max.   :5.0853
##
```

`QLT20`, `LGBT`, and `BRNAGAIN` have missing values stored as `NA`. 616 respondents were not asked `QLT20`, and 615 were not asked either `LGBT` or `BRNAGAIN`. [Rubric: 0 points if not attempted. 0.5 points if the student failed to use the object assignment operator (`<-`). 0.5 points if the incorrect variables were identified for missingness. 0.75 points if the wrong numbers were given for missingness.]

Question 3 [1 point]

Are there **unit non-response** *data in the `PRSMI20` variable? If so, how are they recorded? What about the `PARTYID` variable? How many people refused to answer both of these questions?*

```
MI_clean %>%
  count(PRSMI20)
```

```
## # A tibble: 6 × 2
##   PRSMI20                                     n
##   <dbl+lbl>                                <int>
## 1 0 (NA) [Will/Did not vote for president]     6
## 2 1 [Joe Biden, the Democrat]                723
## 3 2 [Donald Trump, the Republican]           459
## 4 7 [Undecided/Don’t know]                     4
## 5 8 [Refused]                                 14
## 6 9 [Another candidate]                       25
```

```
MI_clean %>%
  count(PARTYID)
```

```
## # A tibble: 5 × 2
##   PARTYID                                  n
##   <dbl+lbl>                             <int>
## 1 1 [Democrat]                            425
## 2 2 [Republican]                          280
## 3 3 [Independent]                         416
## 4 4 [Something else]                       94
## 5 9 [[DON'T READ] Don’t know/refused]      16
```

```
MI_clean %>%
  count(PRSMI20,PARTYID) %>%
  filter(PRSMI20 == 8 & PARTYID == 9)
```

```
## # A tibble: 1 × 3
##   PRSMI20     PARTYID                                  n
##   <dbl+lbl>   <dbl+lbl>                             <int>
## 1 8 [Refused] 9 [[DON'T READ] Don’t know/refused]     2
```

The unit non-response data in the `PRSMI20` variable is recorded with the number `8`. Missing data in the `PARTYID` variable is recorded with the number `9`. Only one person refused to give answers to both questions. [Rubric: 0 points for no attmept. 0.5 points for incorrect answer to how many people refused to answer both questions. 0.75 points if unit non-response code is incorrect for `PRSMI20`. (Some students might also indicate codes `0`, `7`, or `9`, but only `8` is correct.)]

Question 4 [1 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?

```
require(labelled)
MI_clean <- MI_clean %>%
  mutate(preschoice = to_character(PRSMI20))

MI_clean %>%
  count(preschoice)
```

```
## # A tibble: 6 × 2
##   preschoice                          n
##   <chr>                           <int>
## 1 Another candidate                  25
## 2 Donald Trump, the Republican      459
## 3 Joe Biden, the Democrat           723
## 4 Refused                            14
## 5 Undecided/Don't know                4
## 6 Will/Did not vote for president     6
```

459 respondents voted for candidate Trump in 2020. 14 people refused to give an answer. [Rubric: 0 points if no attempt. 0.5 points if correct written answer but code produces error. 0.5 points if correct code but no / wrong answer.]

Question 5 [1 point]

What proportion of women supported Trump?

```
# Women Trump supporters
MI_clean %>%
  drop_na(preschoice) %>%
  filter(SEX == 2) %>%
  count(preschoice) %>%
  mutate(share = n / sum(n))
```

```
## # A tibble: 6 × 3
##   preschoice                          n   share
##   <chr>                           <int>   <dbl>
## 1 Another candidate                   8 0.0123
## 2 Donald Trump, the Republican      212 0.325
## 3 Joe Biden, the Democrat           419 0.643
## 4 Refused                             7 0.0107
## 5 Undecided/Don't know                1 0.00153
## 6 Will/Did not vote for president     5 0.00767
```

```
# Alternative approach
MI_clean %>%
  drop_na(SEX,preschoice) %>%
  mutate(trumpSupp = grepl('Trump',preschoice)) %>%
  group_by(SEX) %>%
  summarise(share = mean(trumpSupp))
```

```
## # A tibble: 2 × 2
##   SEX        share
##   <dbl+lbl>  <dbl>
## 1 1 [Male]   0.427
## 2 2 [Female] 0.325
```

> 32.9% of women supported Trump. [Rubric: 0 points if no attempt. 0.5 points if correct written answer but code produces error. 0.5 points if correct code but no / wrong answer. (Note that there are multiple ways to calculate this answer.)]

Extra Credit [1 point]

Among women, which age group sees the highest support for Trump? To answer, you will need to calculate the proportion of women who supported Trump by age-group to determine which age-group had the highest Trump support among women. You will need to clean the AGE10 variable before completing this problem, just like we did with the PRSMI20 variable. Call the new variable "Age". HINT: to make your life easier (and not write a 10-level nested ifelse() function), try asking ChatGPT for help with this prompt: "I have a labelled variable in R that I want to convert to text. How can I do this?"

```
MI_clean %>%
  count(AGE10)
```

```
## # A tibble: 11 × 2
##    AGE10                          n
##    <dbl+lbl>                  <int>
##  1  1 [18 and 24,]               33
##  2  2 [25 and 29,]               28
##  3  3 [30 and 34,]               42
##  4  4 [35 and 39,]               46
##  5  5 [40 and 44,]               78
##  6  6 [45 and 49,]               83
##  7  7 [50 and 59,]              274
##  8  8 [60 and 64,]              143
##  9  9 [65 and 74,]              290
## 10 10 [75 or over?]             199
## 11 99 [[DON'T READ] Refused]    15
```

```
require(labelled)
MI_clean <- MI_clean %>%
  mutate(Age = as.character(to_factor(AGE10)))

MI_clean %>%
  count(AGE10,Age)
```

```
## # A tibble: 11 × 3
##    AGE10                     Age                     n
##    <dbl+lbl>                 <chr>               <int>
##  1  1 [18 and 24,]           18 and 24,             33
##  2  2 [25 and 29,]           25 and 29,             28
##  3  3 [30 and 34,]           30 and 34,             42
##  4  4 [35 and 39,]           35 and 39,             46
##  5  5 [40 and 44,]           40 and 44,             78
##  6  6 [45 and 49,]           45 and 49,             83
##  7  7 [50 and 59,]           50 and 59,            274
##  8  8 [60 and 64,]           60 and 64,            143
##  9  9 [65 and 74,]           65 and 74,            290
## 10 10 [75 or over?]          75 or over?           199
## 11 99 [[DON'T READ] Refused] [DON'T READ] Refused   15
```

```
MI_clean %>%
  count(Age,SEX,preschoice) %>%
  group_by(Age,SEX) %>%
  mutate(proportion = prop.table(n)) %>%
  filter(SEX == 2,
         grepl('Trump',preschoice)) %>%
  arrange(desc(proportion))
```

```
## # A tibble: 11 × 5
## # Groups:   Age, SEX [11]
##    Age                  SEX        preschoice                       n proportion
##    <chr>                <dbl+lbl>  <chr>                        <int>      <dbl>
##  1 [DON'T READ] Refused 2 [Female] Donald Trump, the Republican     4     0.571
##  2 40 and 44,           2 [Female] Donald Trump, the Republican    15     0.469
##  3 45 and 49,           2 [Female] Donald Trump, the Republican    16     0.372
##  4 75 or over?          2 [Female] Donald Trump, the Republican    41     0.369
##  5 50 and 59,           2 [Female] Donald Trump, the Republican    54     0.36
##  6 65 and 74,           2 [Female] Donald Trump, the Republican    51     0.304
##  7 35 and 39,           2 [Female] Donald Trump, the Republican     5     0.294
##  8 30 and 34,           2 [Female] Donald Trump, the Republican     4     0.267
##  9 60 and 64,           2 [Female] Donald Trump, the Republican    20     0.256
## 10 25 and 29,           2 [Female] Donald Trump, the Republican     1     0.0714
## 11 18 and 24,           2 [Female] Donald Trump, the Republican     1     0.0588
```

Among women, the age group with the greatest support for Trump is between 40 and 44 years old, followed by 45 and 49 year olds. We also see that women who refused to give their age actually had the highest support for Trump (57%). [Rubric: 0 points for no attempt. 0.5 points for preparing the cleaned version of `Age` but failed attempt to calculate proportions by age. 0.75 points for calculating proportions in the wrong way (i.e., % of Trump supporters who are women, by age). Full credit if they decided to answer that the group of women who refused their give their age had the highest support.]