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.)

```
require(tidyverse)
```

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

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.2      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
require(labelled)
```

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

```
## Warning: package 'labelled' was built under R version 4.2.3
```

```
MI_raw <- read_rds('https://github.com/jbisbee1/DS1000_S2024/raw/main/data/MI2020_ExitPoll.rds')
```

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(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
##
```

```
MI_raw %>%
  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_raw %>%
  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
```

```
MI_raw %>%
  count(SEX)
```

```
## # A tibble: 2 × 2
##   SEX            n
##   <dbl+lbl>  <int>
## 1 1 [Male]     579
## 2 2 [Female]   652
```

```
MI_raw %>%
  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
```

`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`.

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?

```
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
```

```
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(AGE10,PARTYID) %>%
  filter(AGE10 == 99 | PARTYID == 9)
```

```
## # A tibble: 13 × 3
##    AGE10                    PARTYID                                n
##    <dbl+lbl>                <dbl+lbl>                          <int>
##  1  2 [25 and 29,]          9 [[DON'T READ] Don’t know/refused]    2
##  2  3 [30 and 34,]          9 [[DON'T READ] Don’t know/refused]    1
##  3  4 [35 and 39,]          9 [[DON'T READ] Don’t know/refused]    1
##  4  6 [45 and 49,]          9 [[DON'T READ] Don’t know/refused]    1
##  5  7 [50 and 59,]          9 [[DON'T READ] Don’t know/refused]    2
##  6  8 [60 and 64,]          9 [[DON'T READ] Don’t know/refused]    2
##  7  9 [65 and 74,]          9 [[DON'T READ] Don’t know/refused]    3
##  8 10 [75 or over?]         9 [[DON'T READ] Don’t know/refused]    3
##  9 99 [[DON'T READ] Refused] 1 [Democrat]                          1
## 10 99 [[DON'T READ] Refused] 2 [Republican]                        4
## 11 99 [[DON'T READ] Refused] 3 [Independent]                       8
## 12 99 [[DON'T READ] Refused] 4 [Something else]                    1
## 13 99 [[DON'T READ] Refused] 9 [[DON'T READ] Don’t know/refused]   1
```

> The unit non-response data in the `AGE10` variable is recorded with the number `99`. Missing data in the `PARTYID` variable is recorded with the number `9`. 15 people refused to give their age and 16 people refused to give their party. (NB: Not required for full credit: Only one person refused to give answers to both questions.)

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?

```
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?

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`.

```
# Converting
MI_clean <- MI_clean %>%
  mutate(Qlty = to_character(QLT20),
         Age = to_character(AGE10),
         Lgbt_clean = to_character(LGBT))

# Looking at unit non-response codes
MI_clean %>%
  count(Qlty) # [DON'T READ] Don’t know/refused
```

```
## # A tibble: 6 × 2
##   Qlty                                n
##   <chr>                           <int>
## 1 [DON'T READ] Don’t know/refused    26
## 2 Can unite the country             125
## 3 Cares about people like me        121
## 4 Has good judgment                 205
## 5 Is a strong leader                138
## 6 <NA>                              616
```

```
MI_clean %>%
  count(Age) # [DON'T READ] Refused
```

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

```
MI_clean %>%
  count(Lgbt_clean) # [DON'T READ] Don’t know/Refused
```

```
## # A tibble: 4 × 2
##   Lgbt_clean                          n
##   <chr>                           <int>
## 1 [DON'T READ] Don’t know/Refused    23
## 2 No                                570
## 3 Yes                                23
## 4 <NA>                              615
```

```
MI_clean %>%
  count(preschoice) # Refused
```

```
## # 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
```

```
# Replacing with NAs
MI_clean <- MI_clean %>%
  mutate(Qlty = ifelse(grepl("DON'T READ",Qlty),NA,Qlty),
         Lgbt_clean = ifelse(grepl("DON'T READ",Lgbt_clean),NA,Lgbt_clean),
         Age = ifelse(grepl("DON'T READ",Age),NA,Age),
         preschoice = ifelse(grepl("Refused",preschoice),NA,preschoice))
```

Question 5 [1 point]

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

```
# LGBT Trump supporters
MI_clean %>%
  drop_na(preschoice) %>%
  filter(Lgbt_clean == 'Yes') %>%
  count(preschoice) %>%
  mutate(share = n / sum(n))
```

```
## # A tibble: 4 × 3
##   preschoice                          n  share
##   <chr>                           <int>  <dbl>
## 1 Donald Trump, the Republican        7 0.304
## 2 Joe Biden, the Democrat            14 0.609
## 3 Undecided/Don’t know                1 0.0435
## 4 Will/Did not vote for president     1 0.0435
```

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

```
## # A tibble: 5 × 3
##   preschoice                          n   share
##   <chr>                           <int>   <dbl>
## 1 Another candidate                   8 0.0124
## 2 Donald Trump, the Republican      212 0.329
## 3 Joe Biden, the Democrat           419 0.650
## 4 Undecided/Don't know                1 0.00155
## 5 Will/Did not vote for president     5 0.00775
```

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

```
## # A tibble: 2 × 2
##   Lgbt_clean share
##   <chr>      <dbl>
## 1 No         0.385
## 2 Yes        0.304
```

```
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]   0.432
## 2 2 [Female] 0.329
```

> 30.4% of LGBT-identifying voters supported Trump. 32.9% of women supported Trump.

Extra Credit [2 points]

Plot the relationship between Trump support and gender.

```
MI_clean %>%
  mutate(SEX = to_character(SEX)) %>%
  group_by(SEX) %>%
  summarise(pctTrump = mean(grepl('Trump',preschoice))) %>%
  ggplot(aes(x = SEX,y = pctTrump)) +
  geom_bar(stat = 'identity') +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Gender',
       y = '% Supporting Trump',
       title = 'Trump Support by Gender in the 2020 U.S. Presidential Election')
```