Lecture 4 Notes

2024-07-03

Wrapping up data wrangling and transitioning to univariate analysis

```
require(tidyverse)
```

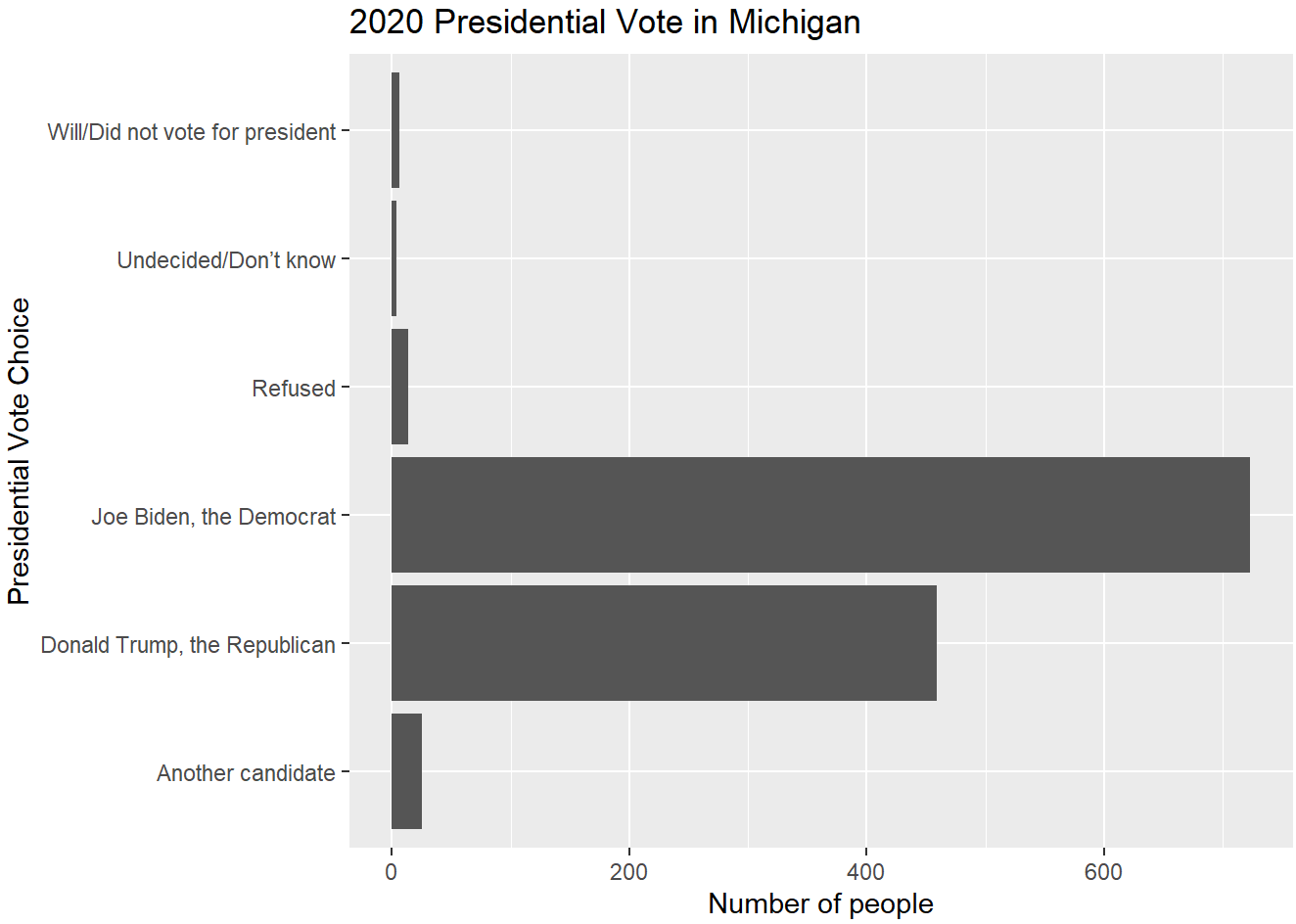
```
## Loading required package: tidyverse
```

```
## ── 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 t
o become errors
```

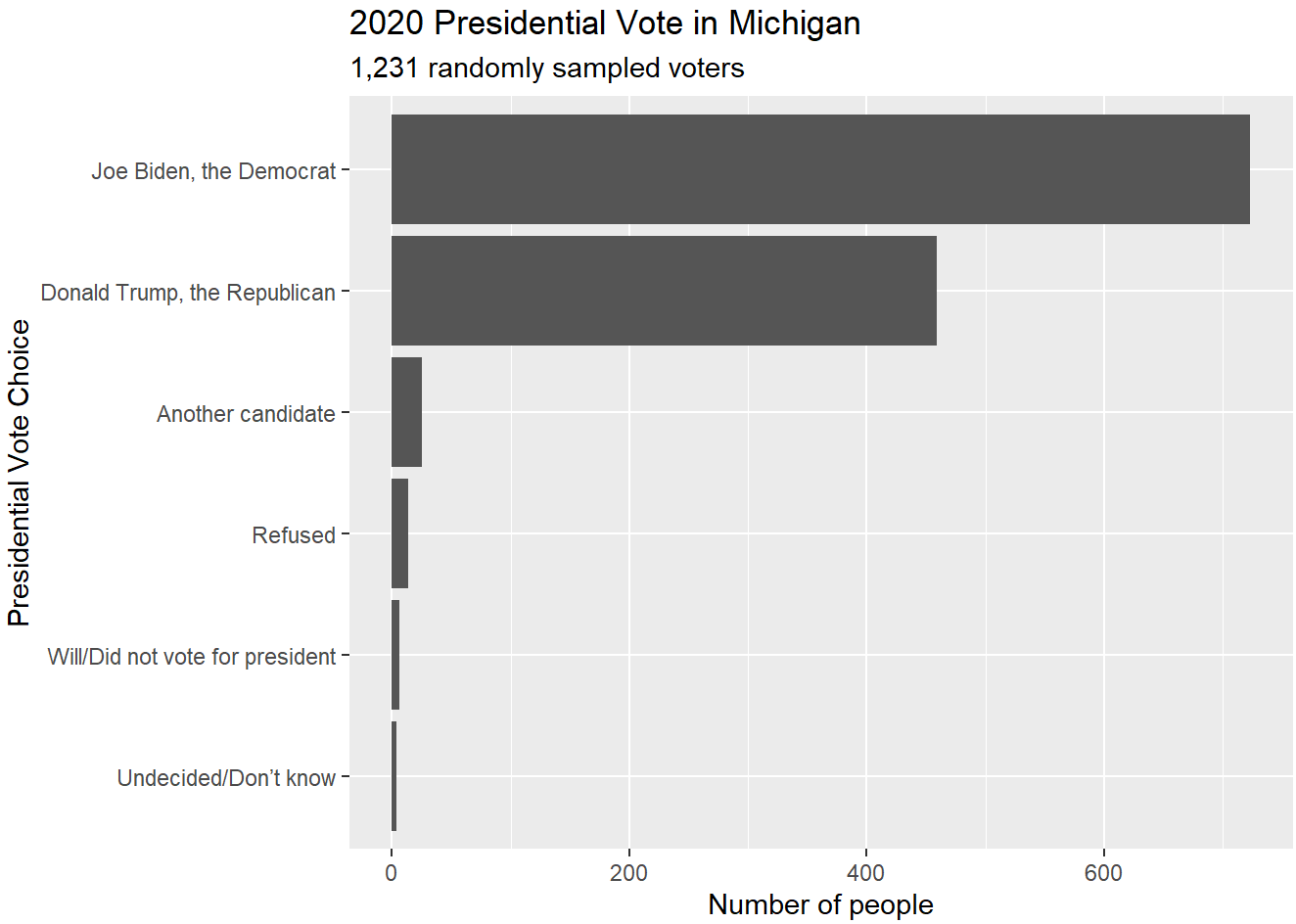
```
MI <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/MI2020_E
xitPoll_small.rds")

MI
```

```
## # A tibble: 1,231 × 14
##      SEX AGE10 PRSMI20 PARTYID WEIGHT QRACEAI EDUC18  LGBT BRNAGAIN LATINOS
##    <dbl> <dbl>   <dbl>   <dbl>  <dbl>   <dbl>  <dbl> <dbl>    <dbl>   <dbl>
##  1     2     2       1       3  0.405       1      4    NA       NA       2
##  2     2    10       1       1  1.81        2      1     2        1       2
##  3     2     7       1       1  0.860       1      5     2        2       2
##  4     1     9       1       3  0.199       1      4    NA       NA       2
##  5     2     8       1       3  0.177       1      5    NA       NA       2
##  6     2     7       1       3  0.492       1      3     2        2       2
##  7     1     9       1       1  1.37        1      3     2        1       2
##  8     1     8       1       1  1.15        1      3     2        2       2
##  9     2     6       2       2  1.50        1      4    NA       NA       2
## 10     1     8       1       1  1.30        2      4    NA       NA       2
## # ℹ 1,221 more rows
## # ℹ 4 more variables: RACISM20 <dbl>, QLT20 <fct>, preschoice <chr>,
## #   Quality <chr>
```

```
# Create barplot of preschoice
MI %>%
  ggplot(aes(y = preschoice)) +
  geom_bar() +
  labs(y = "Presidential Vote Choice",
       x = "Number of people",
       title = "2020 Presidential Vote in Michigan")
```



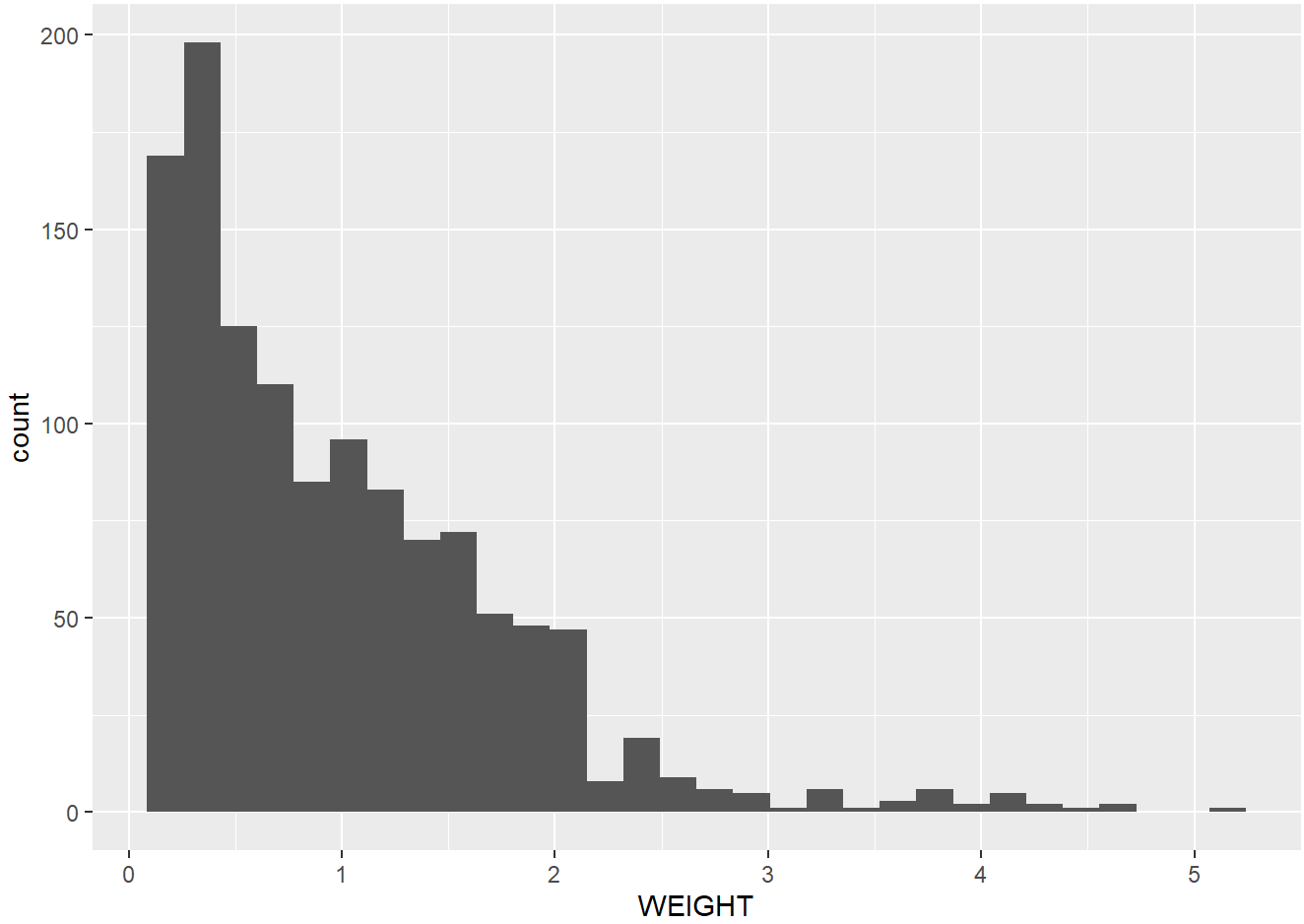

```
# Second approach...calculate count ourselves
MI %>%
  count(preschoice) %>%
  ggplot(aes(y = reorder(preschoice,n),
             x = n)) +
  geom_bar(stat = 'identity') +
    labs(y = "Presidential Vote Choice",
       x = "Number of people",
       title = "2020 Presidential Vote in Michigan",
       subtitle = "1,231 randomly sampled voters")
```



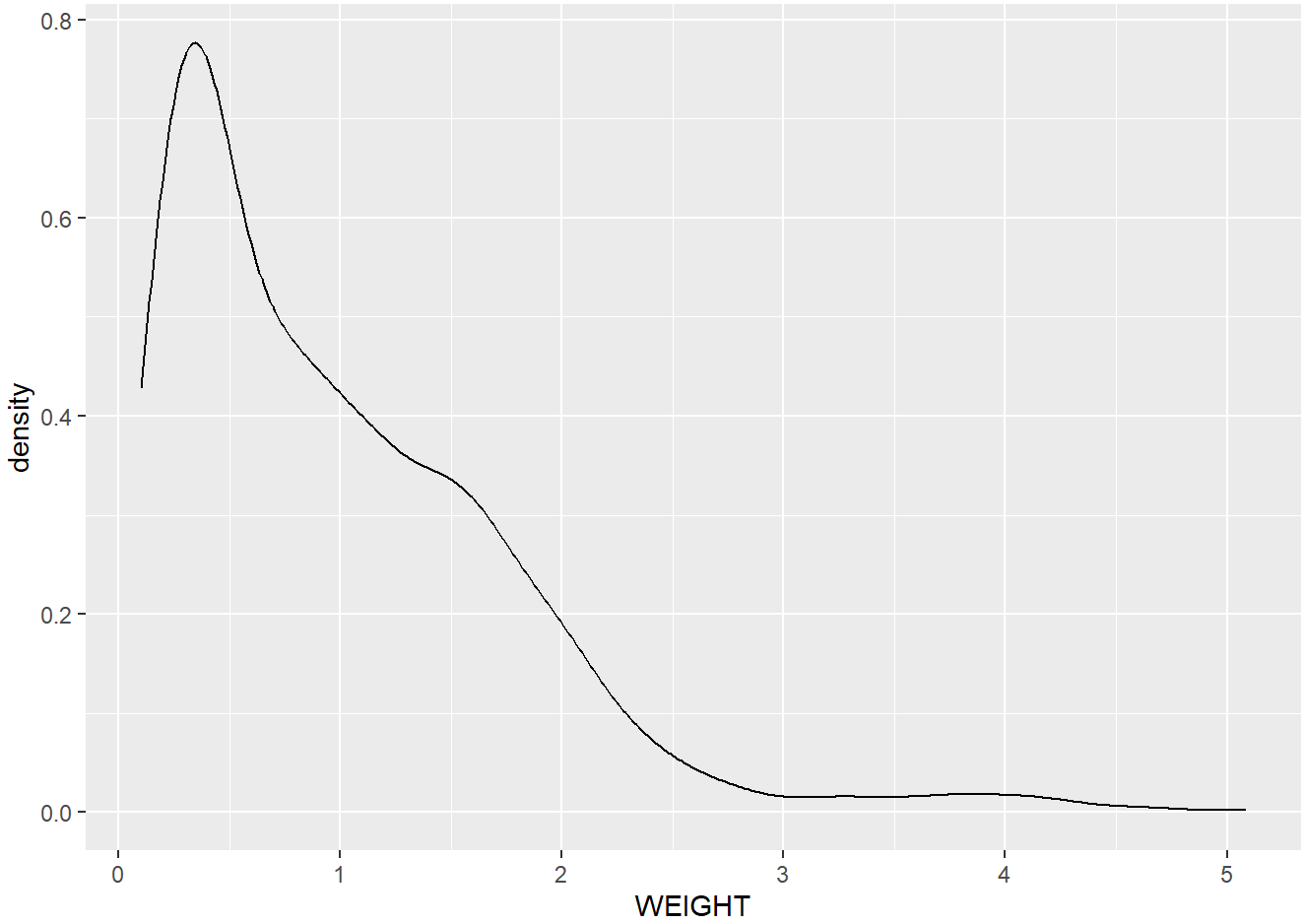
Other variables we could look at

```
View(MI)

# Univariate visualization of WEIGHT
MI %>%
  ggplot(aes(x = WEIGHT)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

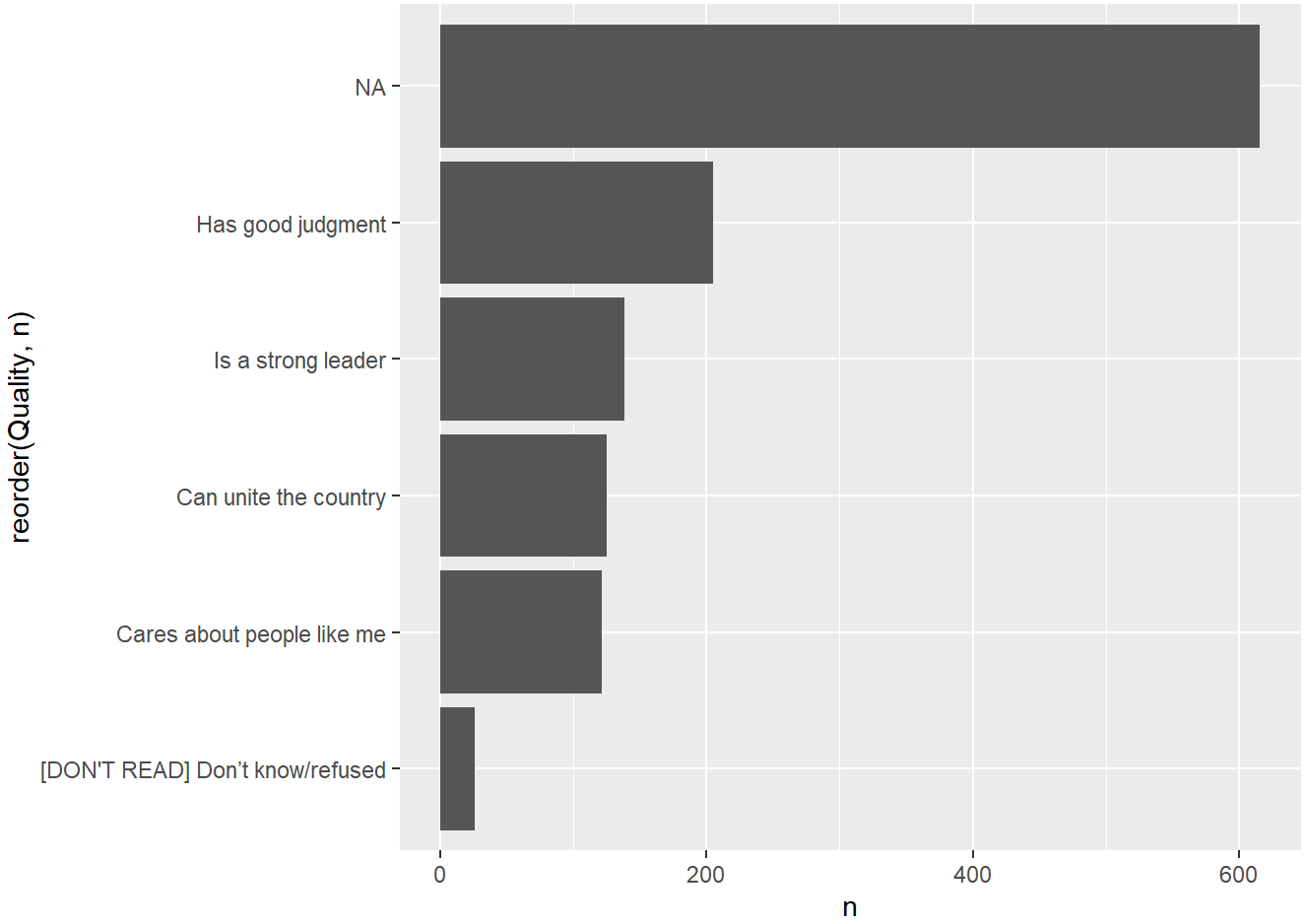



```
# Equally good (in most cases)
MI %>%
  ggplot(aes(x = WEIGHT)) +
  geom_density()
```



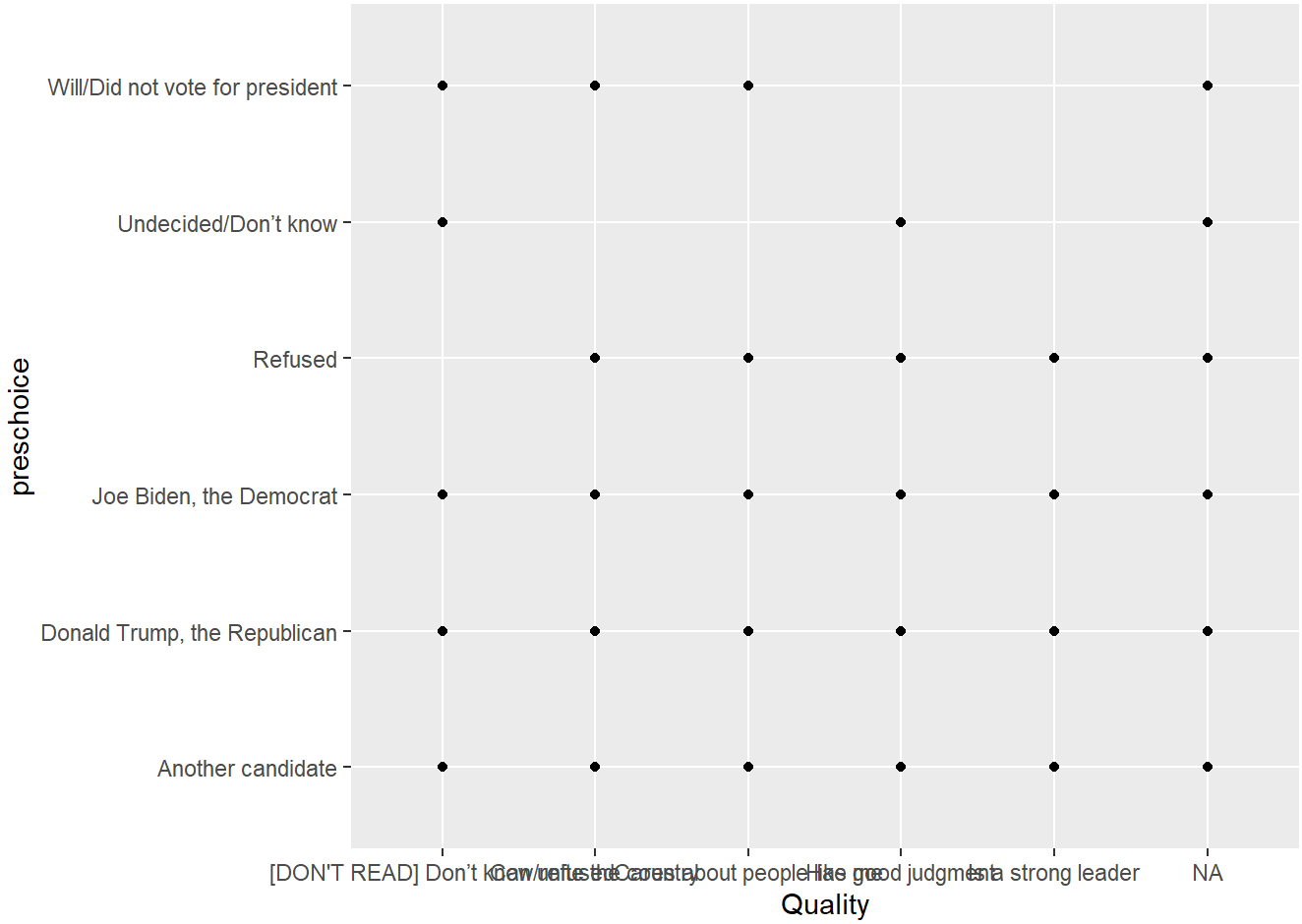
Univariate visualization of Quality

```
MI %>%
  count(Quality) %>%
  ggplot(aes(y = reorder(Quality,n),
             x = n)) +
  geom_bar(stat = 'identity')
```

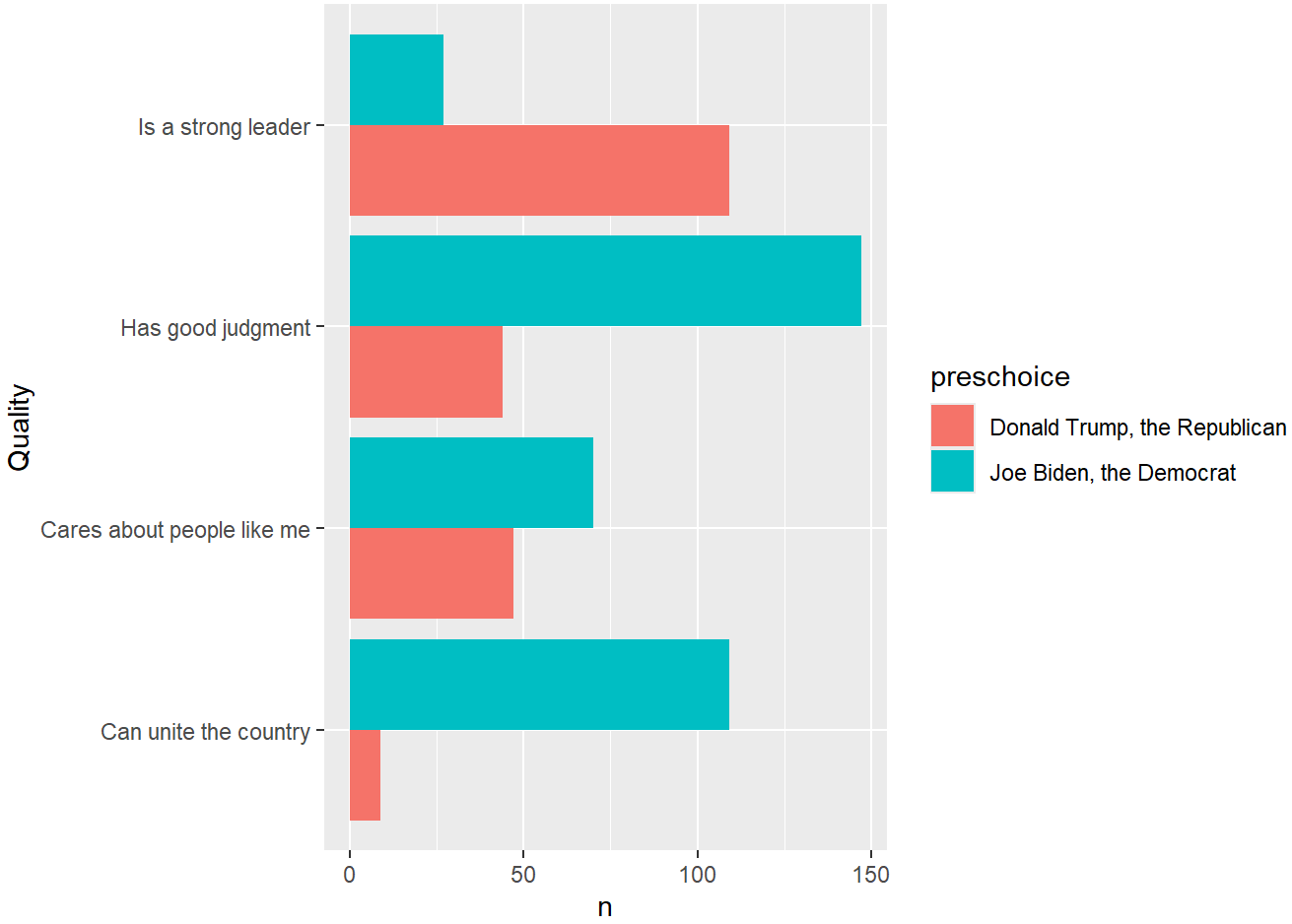


Multivariate Visualization of Quality and preschoice

```
# A bad example of using geom_point()
MI %>%
  ggplot(aes(x = Quality,
             y = preschoice)) +
  geom_point()
```




```
# Let's try geom_bar()
MI %>%
  filter(!is.na(Quality),
         !str_detect(Quality,"refuse")) %>%
  filter(str_detect(preschoice,'Trump|Biden')) %>%
  count(Quality,preschoice) %>%
  ggplot(aes(x = n,
             y = Quality,
             fill = preschoice)) +
  geom_bar(stat = 'identity',
           position = 'dodge')
```



New research question:

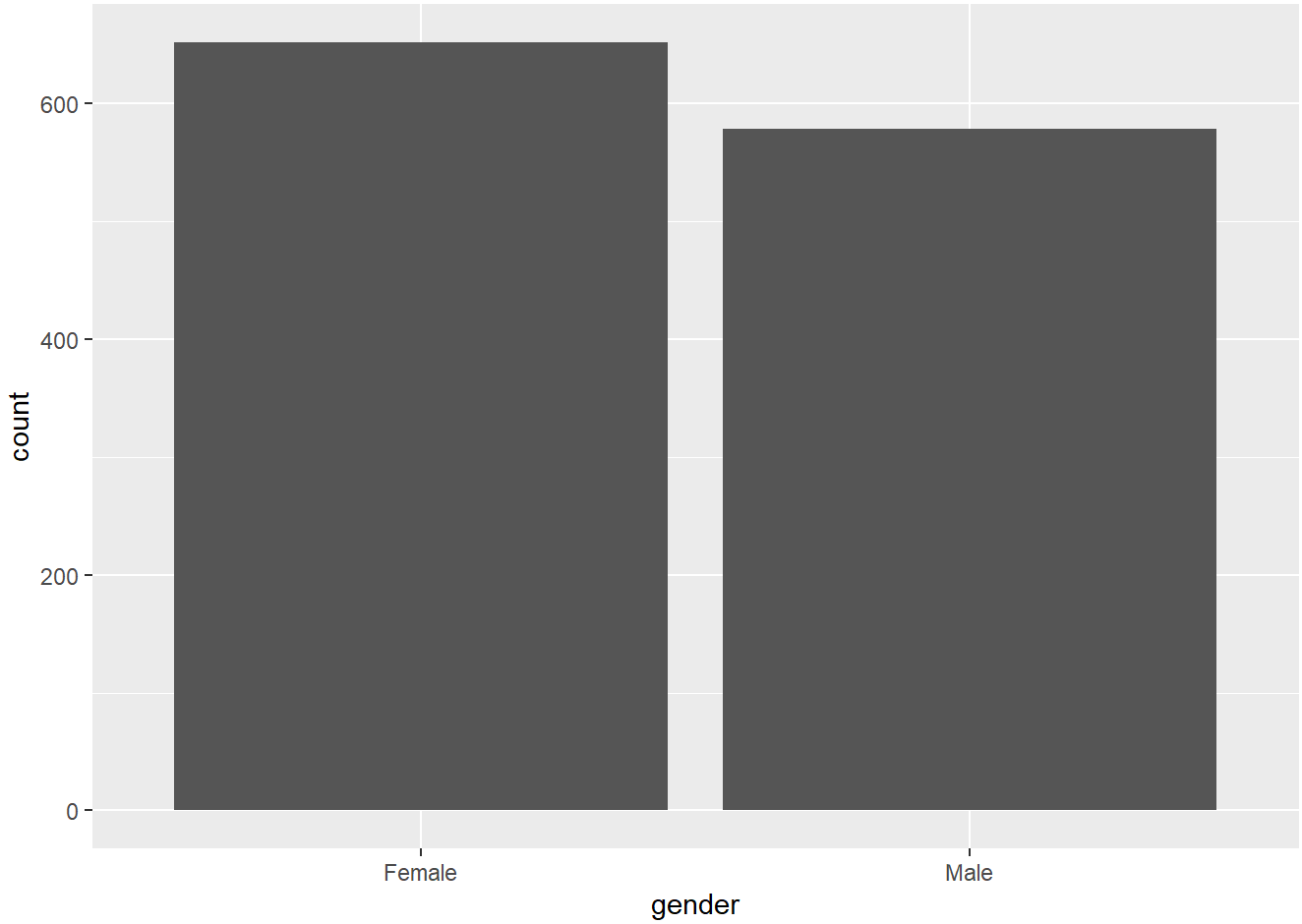
What is the relationship between sex and vote choice?

Start with univariate visualization of X variable

```
MI %>%
  count(SEX)
```

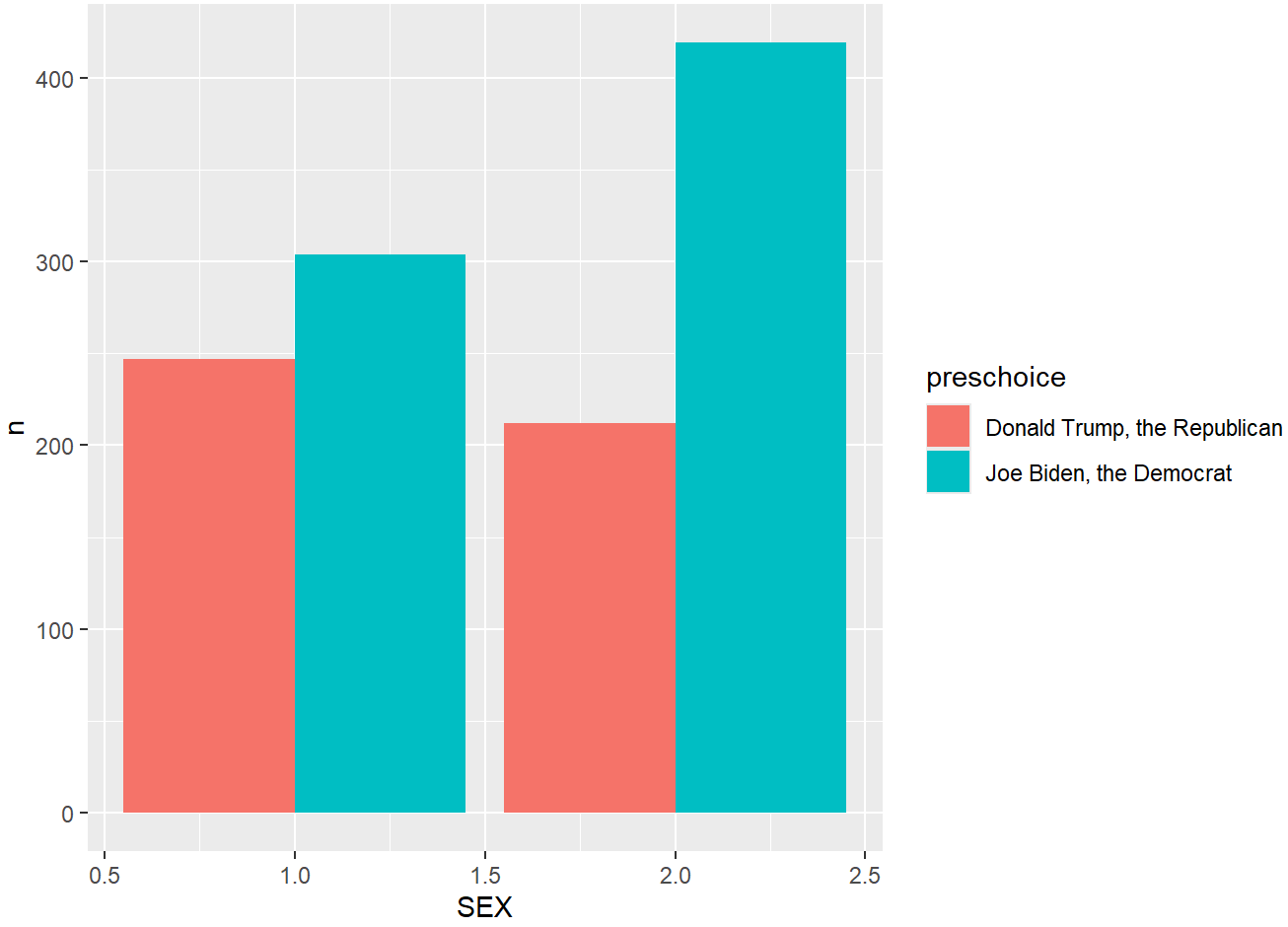
```
## # A tibble: 2 × 2
##     SEX     n
##   <dbl> <int>
## 1     1   579
## 2     2   652
```

```
MI %>%
  mutate(gender = ifelse(SEX == 1,'Male','Female')) %>%
  ggplot(aes(x = gender)) +
  geom_bar()
```

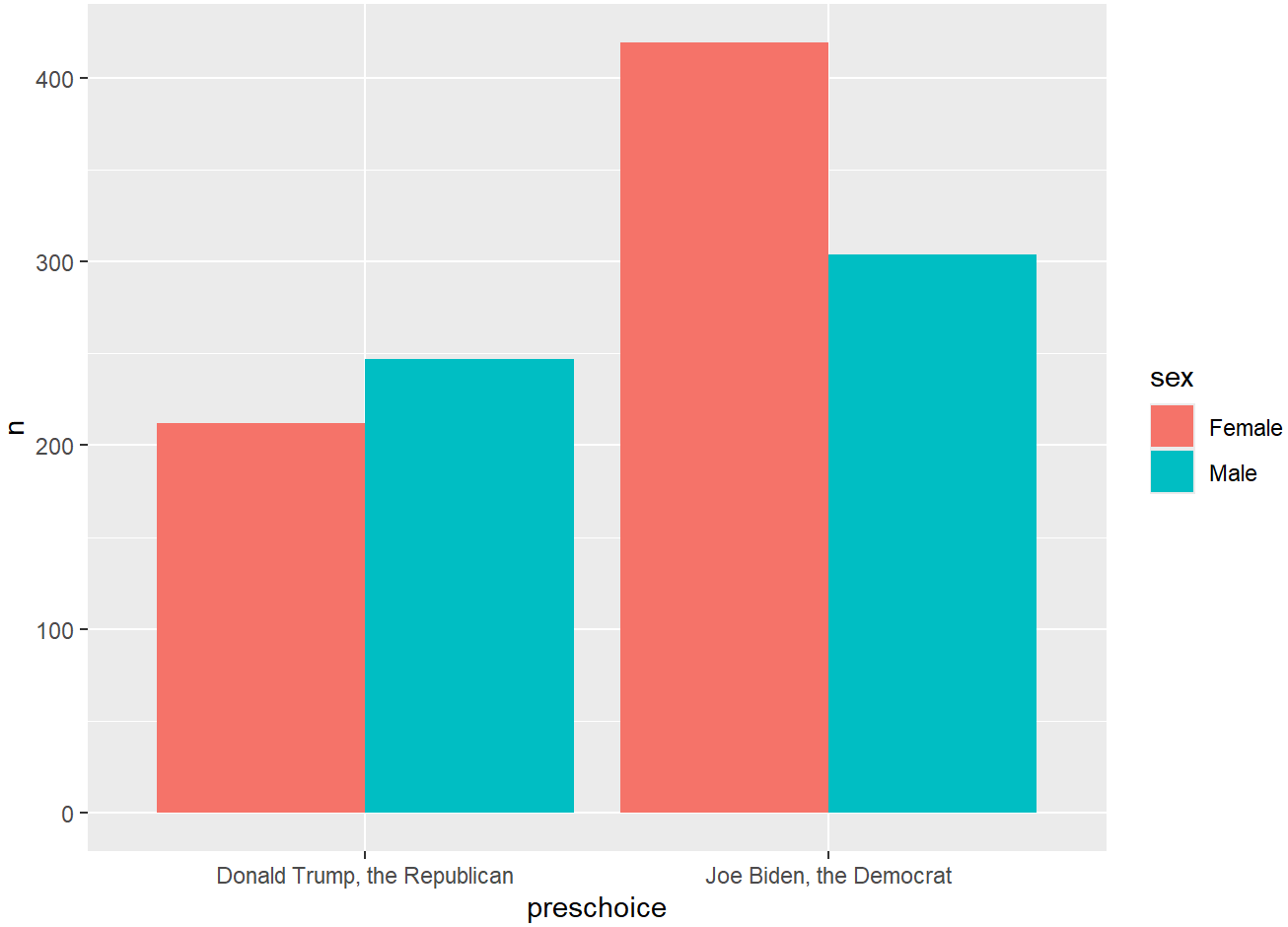


Turning to multivariate visualization

```
MI %>%
  filter(str_detect(preschoice,'Trump|Biden')) %>%
  count(SEX,preschoice) %>%
  ggplot(aes(x = SEX,
             y = n,
             fill = preschoice)) +
  geom_bar(stat = 'identity',
           position = 'dodge')
```

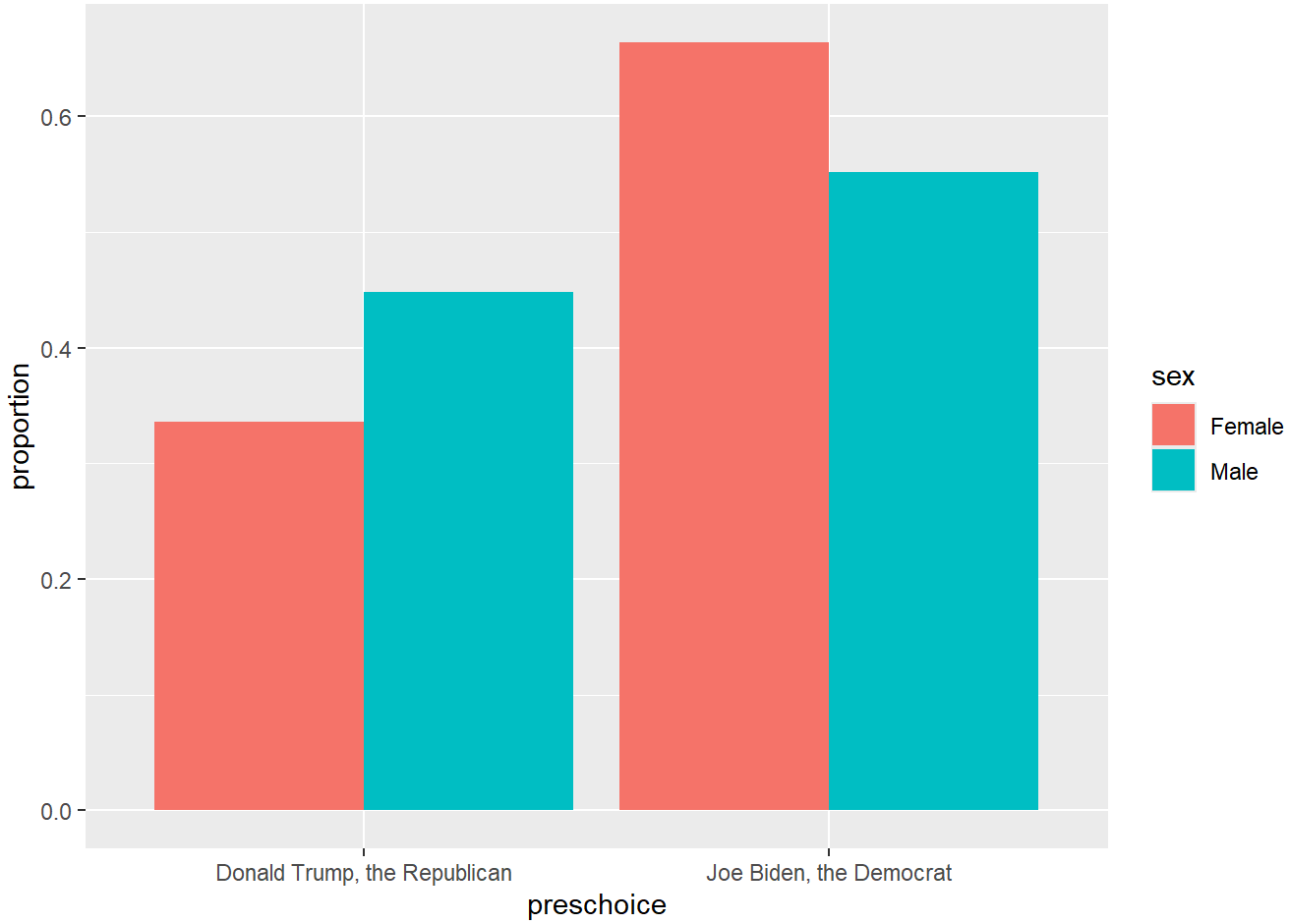



```
MI %>%
  mutate(sex = ifelse(SEX == 1,'Male','Female')) %>%
  filter(str_detect(preschoice,'Trump|Biden')) %>%
  count(sex,preschoice) %>%
  ggplot(aes(x = preschoice,
             y = n,
             fill = sex)) +
  geom_bar(stat = 'identity',
           position = 'dodge')
```



Calculating Proportions

```
# Quick wranlging the data
MI <- MI %>%
  mutate(sex = ifelse(SEX == 1,"Male",'Female'))

# Proportion calculation #1
MI %>%
  filter(str_detect(preschoice,'Biden|Trump')) %>%
  count(sex,preschoice) %>% # This counts up the number of respondents
  group_by(sex) %>% # This and the next line calculate the total by sex
  mutate(totn = sum(n)) %>%
  mutate(proportion = n / totn) %>% # This calculations the proportion
  ggplot(aes(x = preschoice,
             y = proportion,
             fill = sex)) +
  geom_bar(stat = 'identity',
           position = 'dodge')
```




```
# Proportion calculation #2
MI %>%
  filter(str_detect(preschoice,'Biden|Trump')) %>%
  count(sex,preschoice) %>%
  group_by(sex) %>%
  mutate(proportion = prop.table(n))
```

```
## # A tibble: 4 × 4
## # Groups:   sex [2]
##   sex    preschoice                          n proportion
##   <chr>  <chr>                           <int>      <dbl>
## 1 Female Donald Trump, the Republican      212      0.336
## 2 Female Joe Biden, the Democrat           419      0.664
## 3 Male   Donald Trump, the Republican      247      0.448
## 4 Male   Joe Biden, the Democrat           304      0.552
```