# DS1000: Problem Set 2

### Ilayda Koca

### 2022-09-14

# Getting Set Up

If you haven't already, create a folder for this course, and then a subfolder within for the second lecture `Topic4_DataWrangling` , and two additional subfolders within `code` and `data` .

Open `RStudio` and create a new RMarkDown file ( `.Rmd` ) by going to `File -> New File -> R Markdown...` . Change the title to `"DS1000: Problem Set 2"` and the author to your full name. Save this file as `[LAST NAME]_ps2.Rmd` to your `code` folder.

If you haven't already, download the `MI2020_ExitPoll.Rds` file from the course github page (https://github.com/jbisbee1/DS1000-F2022/blob/master/Lectures/Topic4_DataWrangling/data/MI2020_ExitPoll.rds) and save it to your `data` folder.

**NB:** Please upload a `.pdf` version of your homework to Brightspace! To do so, you can either choose the `knit` dropdown of "Knit to PDF", or you can open the standard `.html` output in your browser, then click print and choose "Print to PDF".

# Question 1

Require `tidyverse` and load the `MI2020_ExitPoll.Rds` data to `MI_raw` .

```
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   1.2.0      ✔ 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()
```

```
MI_raw <- readRDS('../data/MI2020_ExitPoll.rds')
```

# Question 2 [1 point]

How many voters were from Wayne County?

```
MI_raw %>%
  count(County) %>%
  filter(County == 'WAYNE')
```

```
## # A tibble: 1 × 2
##   County     n
##   <chr>  <int>
## 1 WAYNE    102
```

> - There were 102 voters from Wayne County elections in the 2020 presidential elections of 2020.

# Question 3 [1 points]

Who did the majority of surveyed voters support in the 2020 presidential election?

```
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
```

> - In the 2020 Michigan presidential elections, the majority of the voters supported Joe Biden of the Democrat Party.

# Question 4 [2 points]

What proportion of women supported Trump? What proportion of men supported Biden?

```
MI_raw %>%
  group_by(SEX) %>%
  summarize(biden_percentage = mean(PRSMI20 == 1), trump_percentage = mean(PRSMI20 ==
2))
```

```
## # A tibble: 2 × 3
##   SEX        biden_percentage trump_percentage
##   <dbl+lbl>             <dbl>            <dbl>
## 1 1 [Male]              0.525            0.427
## 2 2 [Female]            0.643            0.325
```

> - In the 2020 Michigan presidential elections, 32.5% of women supported Trump and 52.5% of the men supported Biden.

# Question 5 [1 point]

Create a new object called `MI_clean` that contains only the following variables: - SEX, AGE10, PARTYID, EDUC18, PRSMI20, QLT20, LGBT, BRNAGAIN, LATINOS, QRACEAI, WEIGHT

```
MI_clean <- MI_raw %>%
  select(SEX, AGE10, PARTYID, EDUC18, PRSMI20, QLT20, LGBT, BRNAGAIN, LATINOS, QRACEA
I, WEIGHT)
```

# Question 6 [1 point]

Which of these variables have missing data recorded as `NA` ?

```
MI_clean %>%
  colSums(is.na(MI_clean))
```

```
##      SEX    AGE10  PARTYID   EDUC18  PRSMI20    QLT20     LGBT BRNAGAIN
##     1883    10434     2753     4048     2006       NA       NA       NA
##  LATINOS  QRACEAI   WEIGHT
##     2678     1935     1231
```

> - QLT20, LGBT, and BRNAGAIN variables have missing data recorded as NA.

# Question 7 [1 point]

Are there **unit non-response** data in the `AGE10` variable? If so, how are they recorded?

```
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
```

> - There are 15 people who either refused to answer or didn’t read the question regarding sex at all. They are recorded so that the data is not considered NA, but hints that the question might have been confusing/embarrassing/frustrating to answer for 15 of the respondents. “[[DON’T READ] Refused]”

# Question 8 [1 point]

What about in the PARTYID variable? How is unit non-response data recorded there?

```
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
```

> - There are 16 of unit non-response data recorded for the Party id category. This data recorded in a way that reflects respondents who couldn’t answer the question either because they didn’t know who to vote for or they refused to answer because of a potential perceived thread/humiliation/intrusion. “[[DON’T READ] Don’t know/refused]”

# Question 9 [1 point]

Let’s create a new variable called `preschoice` that converts `PRSMI20` to a character. To do this, install the `sjlabelled` package and then create a new dataset called `lookup` that contains both the numeric value of the `PRSMI20` variable as well as the character label. Then merge this `lookup` dataframe to the `MI_clean tibble` with `left_join`.

```
#sjlabelled library extracts the labels as chars
sjlabelled::get_labels(MI_raw$PRSMI20)
```

```
## [1] "Will/Did not vote for president" "Joe Biden, the Democrat"
## [3] "Donald Trump, the Republican"    "Undecided/Don’t know"
## [5] "Refused"                         "Another candidate"
```

```
#create a new column that converts the value for PRSMI20 to the label
#to do this, create a lookup object containing the numeric values and labels for PRSM
I20
labels <- sjlabelled::get_labels(MI_raw$PRSMI20)
values <- sjlabelled::get_values(MI_raw$PRSMI20)
lookup <- data.frame(PRSMI20 = values, preschoice = labels)
lookup
```

```
##   PRSMI20                      preschoice
## 1       0 Will/Did not vote for president
## 2       1         Joe Biden, the Democrat
## 3       2    Donald Trump, the Republican
## 4       7            Undecided/Don’t know
## 5       8                         Refused
## 6       9               Another candidate
```

```
#Now, we can merge our data with the look-up to attach the char column to preschoice
#to merge, use the left_join() function
MI_raw <- MI_raw %>%
  left_join(lookup,by = c('PRSMI20' = 'PRSMI20'))

MI_raw %>%
  select(PRSMI20,preschoice)
```

```
## # A tibble: 1,231 × 2
##    PRSMI20                         preschoice
##    <dbl+lbl>                       <chr>
##  1 1 [Joe Biden, the Democrat]      Joe Biden, the Democrat
##  2 1 [Joe Biden, the Democrat]      Joe Biden, the Democrat
##  3 1 [Joe Biden, the Democrat]      Joe Biden, the Democrat
##  4 1 [Joe Biden, the Democrat]      Joe Biden, the Democrat
##  5 1 [Joe Biden, the Democrat]      Joe Biden, the Democrat
##  6 1 [Joe Biden, the Democrat]      Joe Biden, the Democrat
##  7 1 [Joe Biden, the Democrat]      Joe Biden, the Democrat
##  8 1 [Joe Biden, the Democrat]      Joe Biden, the Democrat
##  9 2 [Donald Trump, the Republican] Donald Trump, the Republican
## 10 1 [Joe Biden, the Democrat]      Joe Biden, the Democrat
## # … with 1,221 more rows
```

# Question 10 [1 point]

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` . EXTRA CREDIT: create a function to repeat this task easily.

```
#create a function to relabel data
relabFn <- function(data,column) {
  labels <- sjlabelled::get_labels(data[[column]])
  values <- sjlabelled::get_values(data[[column]])
  return(data.frame(orig = values,lab = labels))
}

lookupAGE10 <- relabFn(data = MI_raw,column = 'AGE10') %>%
  rename(AGE10 = orig,Age = lab)
lookupQLT20 <- relabFn(data = MI_raw,column = 'QLT20') %>%
  rename(QLT20 = orig,Qlty = lab)
lookupLGBT <- relabFn(data = MI_raw,column = 'LGBT') %>%
  rename(LGBT = orig,Lgbt_clean = lab)

lookupAGE10
```

```
##    AGE10                  Age
## 1      1           18 and 24,
## 2      2           25 and 29,
## 3      3           30 and 34,
## 4      4           35 and 39,
## 5      5           40 and 44,
## 6      6           45 and 49,
## 7      7           50 and 59,
## 8      8           60 and 64,
## 9      9           65 and 74,
## 10    10          75 or over?
## 11    99 [DON'T READ] Refused
```

```
lookupQLT20
```

```
##   QLT20                            Qlty
## 1     1           Can unite the country
## 2     2              Is a strong leader
## 3     3      Cares about people like me
## 4     4               Has good judgment
## 5     9 [DON'T READ] Don’t know/refused
```

```
lookupLGBT
```

```
##   LGBT                      Lgbt_clean
## 1    1                             Yes
## 2    2                              No
## 3    9 [DON'T READ] Don’t know/Refused
```

# Question 11 [1 point]

For each of these new variables, replace the missing data label with `NA`.

```
MI_raw %>%
  mutate(Qlty = ifelse(QLT20 == 9 ,NA, QLT20)) %>%
  count(Qlty)
```

```
## # A tibble: 5 × 2
##    Qlty     n
##   <dbl> <int>
## 1     1   125
## 2     2   138
## 3     3   121
## 4     4   205
## 5    NA   642
```

# Question 12 [2 points]

What proportion of LGBT-identifying voters supported Trump?

```
MI_raw %>%
  group_by(LGBT) %>%
  summarize( trump_percentage = mean(PRSMI20 == 2))
```

```
## # A tibble: 4 × 2
##   LGBT                                 trump_percentage
##   <dbl+lbl>                                       <dbl>
## 1  1 [Yes]                                        0.304
## 2  2 [No]                                         0.382
## 3  9 [[DON'T READ] Don’t know/Refused]            0.435
## 4 NA                                              0.364
```

> - In the 2020 Michigan presidential elections, 30.4% of the LGBT-identifying voters supported Trump.

# Question 13 [2 points]

Convert `AGE10` to a numeric variable and replace the missing data code with `NA`. What is the average age category in the data? What age bracket does this define?

```
MI_raw <- MI_raw %>%
  mutate(AGE_new = ifelse(AGE10 == 99,NA,AGE10))
MI_raw %>%
  summarise(avgAge = mean(AGE_new,na.rm=T))
```

```
## # A tibble: 1 × 1
##   avgAge
##    <dbl>
## 1   7.36
```

```
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
```

> - The average age category in the data is 7.39 which indicates the age bracket between the ages of 50-59.

# Question 14 [2 points]

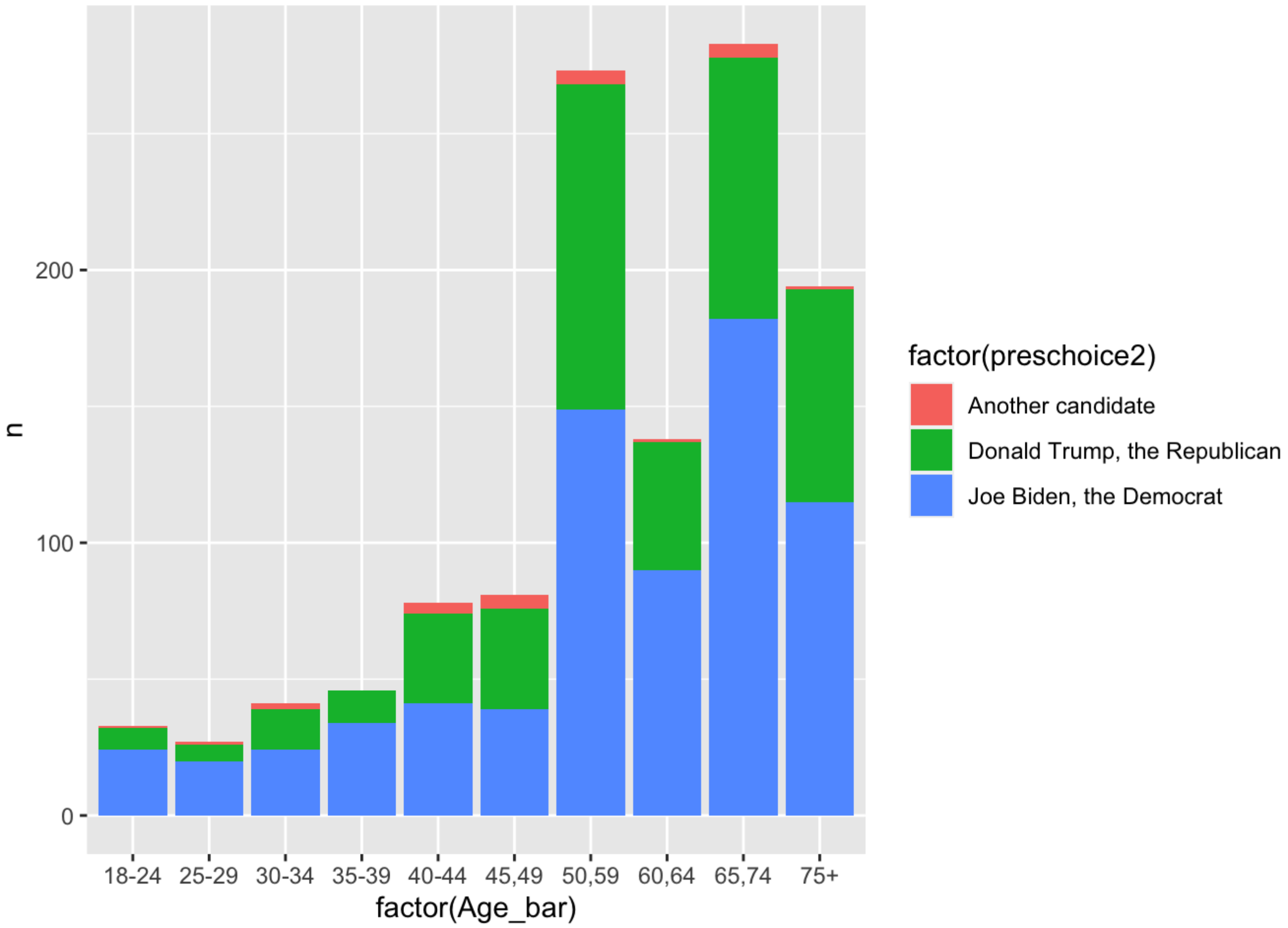
Plot the distribution of ages in the data. EXTRA CREDIT: color by the number of voters in each bracket that supported Trump, Biden, or someone else. Make sure to drop voters who didn't indicate who they voted for **AND** those who didn't indicate their age.

```
MI_raw <- MI_raw %>%
  mutate(Age_bar = ifelse(AGE10 == 1, "18-24",
                        ifelse(AGE10 ==2, "25-29",
                               ifelse(AGE10==3,"30-34",
                                      ifelse(AGE10==4,"35-39",
                                             ifelse(AGE10==5,"40-44",
                                                    ifelse(AGE10==6,"45,49",
                                                           ifelse(AGE10==7, "50,59"
,
                                                                  ifelse(AGE10==8,
"60,64",ifelse(AGE10==9,"65,74",ifelse(AGE10==10,"75+",NA)))))))))))

MI_raw <- MI_raw %>%
  mutate(preschoice2 = ifelse(preschoice == "Refused",NA,ifelse(preschoice == "Will/D
id not vote for president",NA,ifelse(preschoice == "Undecided/Don’t know",NA,preschoi
ce))))

MI_raw <- MI_raw %>%
  mutate(preschoice2 = ifelse(PRSMI20 == 7, NA, preschoice2))

MI_raw %>%
  group_by(Age_bar,preschoice2) %>%
  count(Age_bar) %>%
  drop_na(Age_bar) %>%
  drop_na(preschoice2) %>%
  ggplot(aes(x = factor(Age_bar),y = n, fill = factor(preschoice2))) +
  geom_bar(stat = "identity")
```



```
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
```

# Question 15 [3 points]

EXTRA CREDIT: In a two-way race (i.e., dropping those who voted for a candidate other than Biden or Trump), which age group most heavily favored Trump? Which most heavily favored Biden? Discuss some theories for why this might be the case. EXTRA **EXTRA** CREDIT: plot this answer.

```
relabFn <- function(data,column) {
  labels <- sjlabelled::get_labels(data[[column]])
  values <- sjlabelled::get_values(data[[column]])
  return(data.frame(orig = values,lab = labels))
}

lookupAGE10 <- relabFn(data = MI_raw,column = 'AGE10') %>%
  rename(AGE10 = orig,Age = lab)

#

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
```

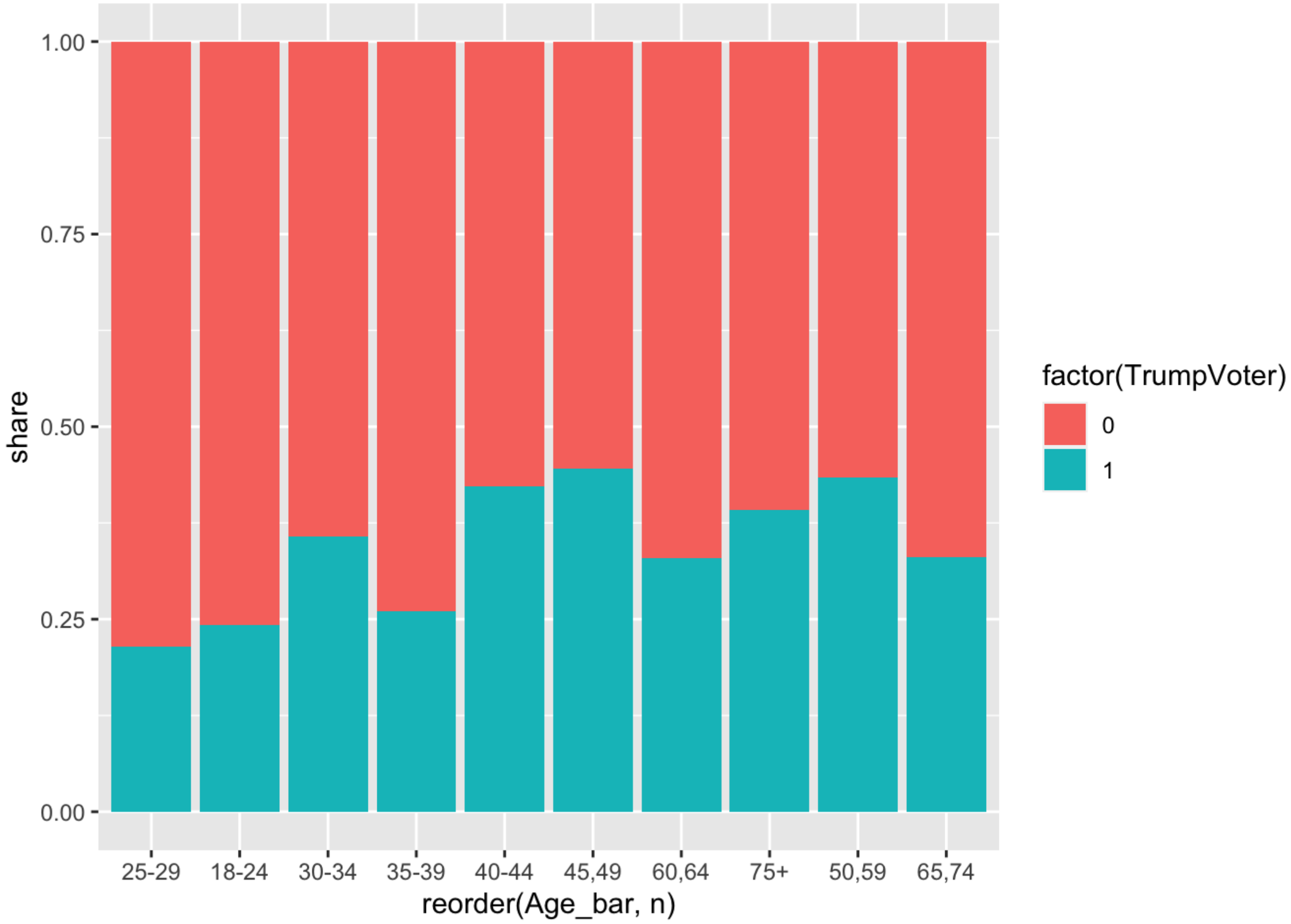
```
MI_raw %>%
  group_by(AGE10) %>%
  summarize( trump_percentage = mean(PRSMI20 == 2), biden_percentage = mean(PRSMI20 =
= 1))
```

```
## # A tibble: 11 × 3
##    AGE10                          trump_percentage biden_percentage
##    <dbl+lbl>                                 <dbl>            <dbl>
##  1  1 [18 and 24,]                           0.242            0.727
##  2  2 [25 and 29,]                           0.214            0.714
##  3  3 [30 and 34,]                           0.357            0.571
##  4  4 [35 and 39,]                           0.261            0.739
##  5  5 [40 and 44,]                           0.423            0.526
##  6  6 [45 and 49,]                           0.446            0.470
##  7  7 [50 and 59,]                           0.434            0.544
##  8  8 [60 and 64,]                           0.329            0.629
##  9  9 [65 and 74,]                           0.331            0.628
## 10 10 [75 or over?]                          0.392            0.578
## 11 99 [[DON'T READ] Refused]                 0.533            0.333
```

```
MI_raw <- MI_raw %>%
  mutate(BidenVoter = ifelse(grepl('Biden',preschoice),1,0),
         TrumpVoter = ifelse(grepl('Trump',preschoice),1,0))

MI_raw <- MI_raw %>%
  mutate(Age_bar = ifelse(AGE10 == 1, "18-24",
                          ifelse(AGE10 ==2, "25-29",
                                 ifelse(AGE10==3,"30-34",
                                        ifelse(AGE10==4,"35-39",
                                               ifelse(AGE10==5,"40-44",
                                                      ifelse(AGE10==6,"45,49",
                                                             ifelse(AGE10==7, "50,59"
,
                                                                    ifelse(AGE10==8,
"60,64",ifelse(AGE10==9,"65,74",ifelse(AGE10==10,"75+",NA))))))))))

MI_raw %>%
  group_by(Age_bar,TrumpVoter) %>%
  drop_na(Age_bar) %>%
  count() %>%
  group_by(Age_bar) %>%
  mutate(share = n / sum(n)) %>%
  ggplot(aes(x = reorder(Age_bar,n), y = share, fill = factor(TrumpVoter))) +
  geom_bar(stat = 'identity')
```



-The age group between 45-49 most heavily supported Trump. Voters between the ages of 25-29 most predominantly favored Biden in the Michigan 2020 Presidential Elections. The reason why younger people below the age 30 support Biden more heavily potentially because Millennials are more racilly diverse, more tuned in to power of networks and systems that make them more liberal as opposed to conservative. Plus they were found to favor government-run health care, student debt relief, marijuana legalization, and such issues that Democrats advocate for.