Lecture 3 Notes

2024-07-02

Which state (`stabbr`) has the lowest average admissions rate (`adm_rate`)?

```
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
```

```
df <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/sc_debt.
Rds")

df %>%
  arrange(adm_rate) %>%
  select(stabbr,adm_rate)
```

```
## # A tibble: 2,546 × 2
##    stabbr adm_rate
##    <chr>     <dbl>
##  1 NY       0
##  2 NY       0
##  3 MA       0.0197
##  4 PA       0.0393
##  5 CA       0.0434
##  6 MA       0.0464
##  7 CA       0.0511
##  8 NY       0.0545
##  9 NJ       0.0578
## 10 CT       0.0608
## # ℹ 2,536 more rows
```

```
df %>%
  group_by(stabbr) %>%
  summarise(avg_adm_rate = mean(adm_rate,na.rm=T)) %>%
  arrange(avg_adm_rate)
```

```
## # A tibble: 51 × 2
##    stabbr avg_adm_rate
##    <chr>         <dbl>
##  1 DC            0.529
##  2 MA            0.582
##  3 CT            0.589
##  4 CA            0.592
##  5 NC            0.609
##  6 RI            0.619
##  7 FL            0.620
##  8 DE            0.627
##  9 LA            0.646
## 10 GA            0.650
## # ℹ 41 more rows
```

Research Question: Do students from more selective schools (`selective`) make a higher salary (`md_earn_wne_p6`)?

```
# First, make selective using mutate() and ifelse()
df %>%
  mutate(example_new_column = adm_rate / 2)
```

```
## # A tibble: 2,546 × 17
##    unitid instnm   stabbr grad_debt_mdn control region preddeg openadmp adm_rate
##     <int> <chr>    <chr>          <int> <chr>   <chr>  <chr>      <int>    <dbl>
##  1 100654 Alabama… AL             33375 Public  South… Bachel…        2    0.918
##  2 100663 Univers… AL             22500 Public  South… Bachel…        2    0.737
##  3 100690 Amridge… AL             27334 Private South… Associ…        1   NA
##  4 100706 Univers… AL             21607 Public  South… Bachel…        2    0.826
##  5 100724 Alabama… AL             32000 Public  South… Bachel…        2    0.969
##  6 100751 The Uni… AL             23250 Public  South… Bachel…        2    0.827
##  7 100760 Central… AL             12500 Public  South… Associ…        1   NA
##  8 100812 Athens … AL             19500 Public  South… Bachel…       NA   NA
##  9 100830 Auburn … AL             24826 Public  South… Bachel…        2    0.904
## 10 100858 Auburn … AL             21281 Public  South… Bachel…        2    0.807
## # ℹ 2,536 more rows
## # ℹ 8 more variables: ccbasic <int>, sat_avg <int>, md_earn_wne_p6 <int>,
## #   ugds <int>, costt4_a <int>, selective <dbl>, research_u <dbl>,
## #   example_new_column <dbl>
```

```
# df %>%
#   summarise(mean_example = mean(example_new_column))

df <- df %>%
  mutate(selective = ifelse(adm_rate < .1,
                            'selective',
                            'non-selective'))

# Second, calculate average salary by selective
df %>%
  group_by(selective) %>%
  summarise(avg_salary = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 3 × 2
##   selective     avg_salary
##   <chr>              <dbl>
## 1 non-selective     35739.
## 2 selective         54252.
## 3 <NA>              27901.
```

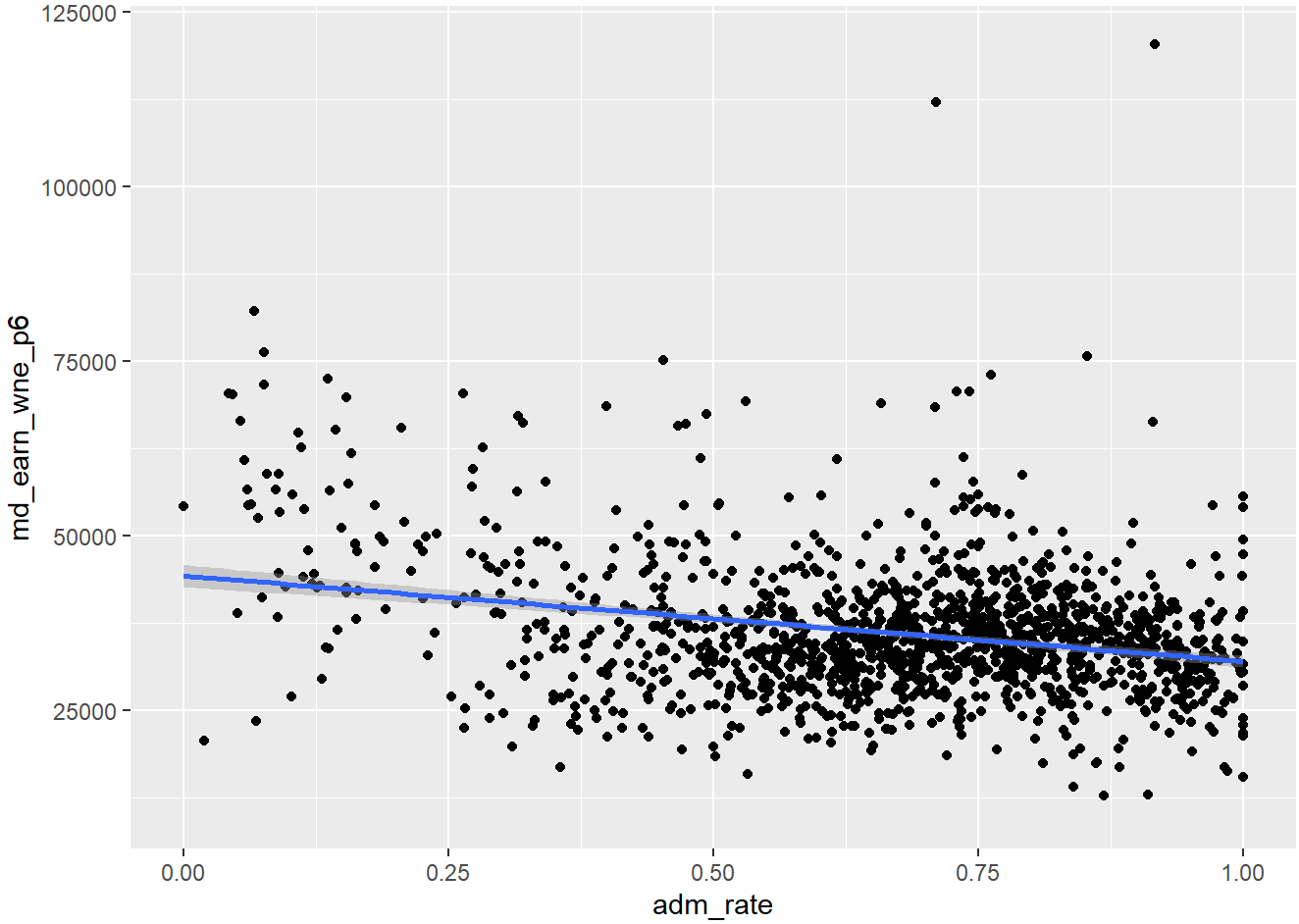
Introducing functions from ggplot2

```
df %>%
  ggplot(aes(x = adm_rate,
             y = md_earn_wne_p6)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1092 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1092 rows containing missing values or values outside the scale rang
e
## (`geom_point()`).
```



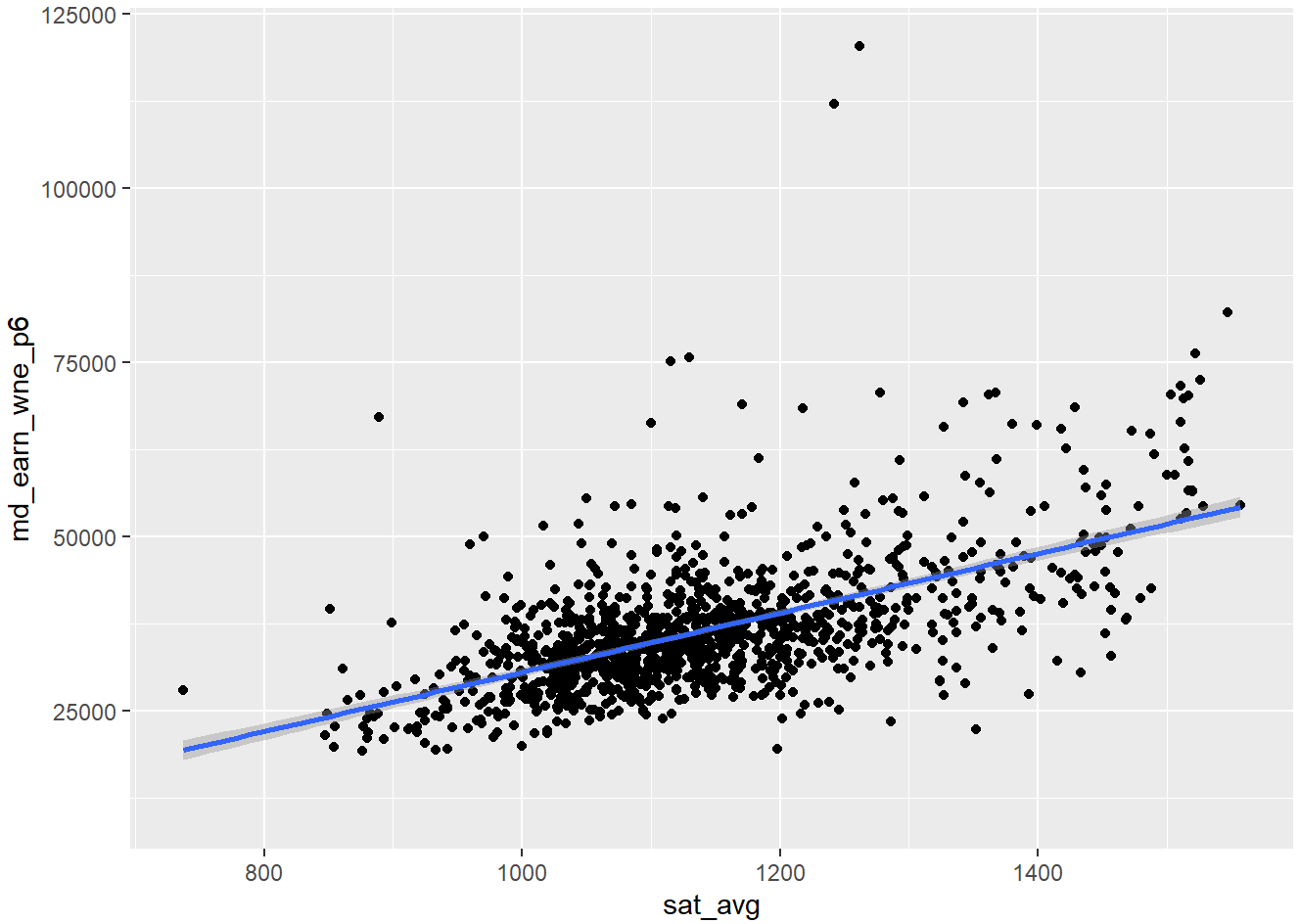
What is the relationship between SAT scores and earnings?

```
df %>%
  ggplot(aes(x = sat_avg,
             y = md_earn_wne_p6)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1348 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1348 rows containing missing values or values outside the scale rang
e
## (`geom_point()`).
```



Looking at outliers

```
df %>%
  mutate(outlier = ifelse(md_earn_wne_p6 > 100000,
                          instnm,
                          NA)) %>%
  drop_na(outlier,sat_avg) %>%
  select(instnm)
```

```
## # A tibble: 2 × 1
##   instnm
##   <chr>
## 1 University of Health Sciences and Pharmacy in St. Louis
## 2 Albany College of Pharmacy and Health Sciences
```

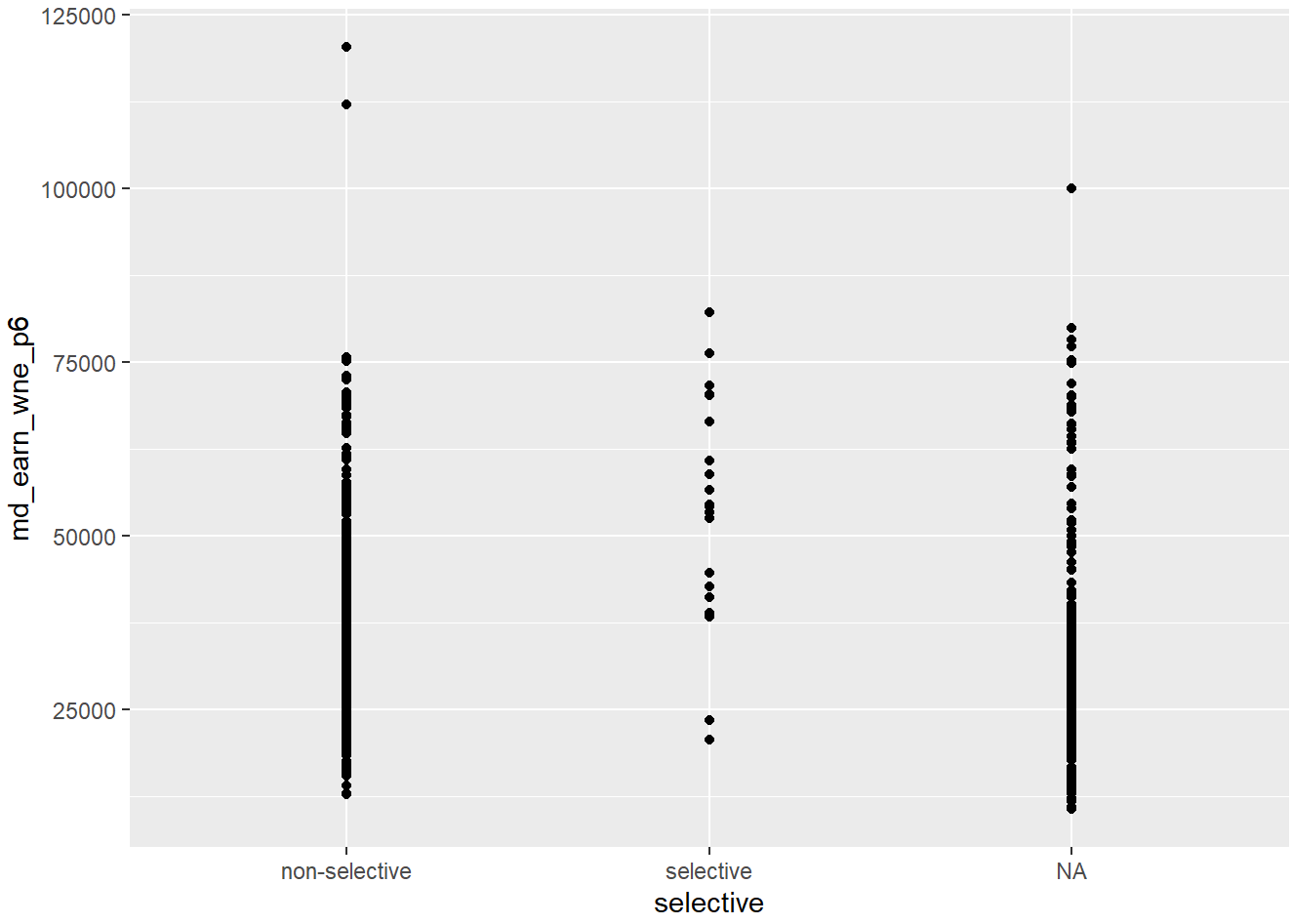
Variable “types”

```
df %>%
  select(sat_avg,selective,md_earn_wne_p6)
```

```
## # A tibble: 2,546 × 3
##    sat_avg selective     md_earn_wne_p6
##      <int> <chr>                  <int>
##  1     939 non-selective          25200
##  2    1234 non-selective          35100
##  3      NA <NA>                   30700
##  4    1319 non-selective          36200
##  5     946 non-selective          22600
##  6    1261 non-selective          37400
##  7      NA <NA>                   23100
##  8      NA <NA>                   33400
##  9    1082 non-selective          30100
## 10    1300 non-selective          39500
## # ℹ 2,536 more rows
```

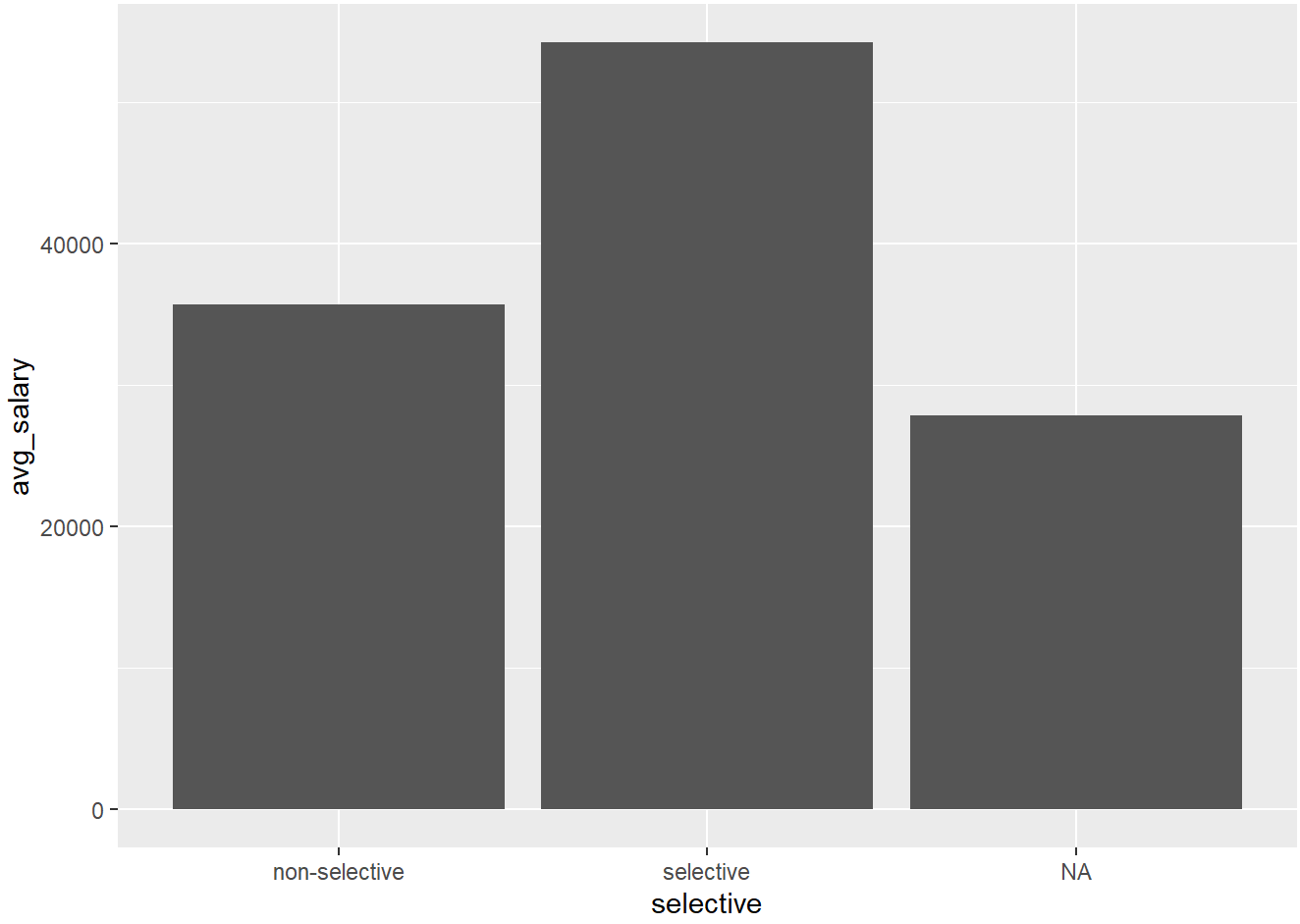
```
df %>%
  ggplot(aes(x = selective,
             y = md_earn_wne_p6)) +
  geom_point()
```

```
## Warning: Removed 240 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

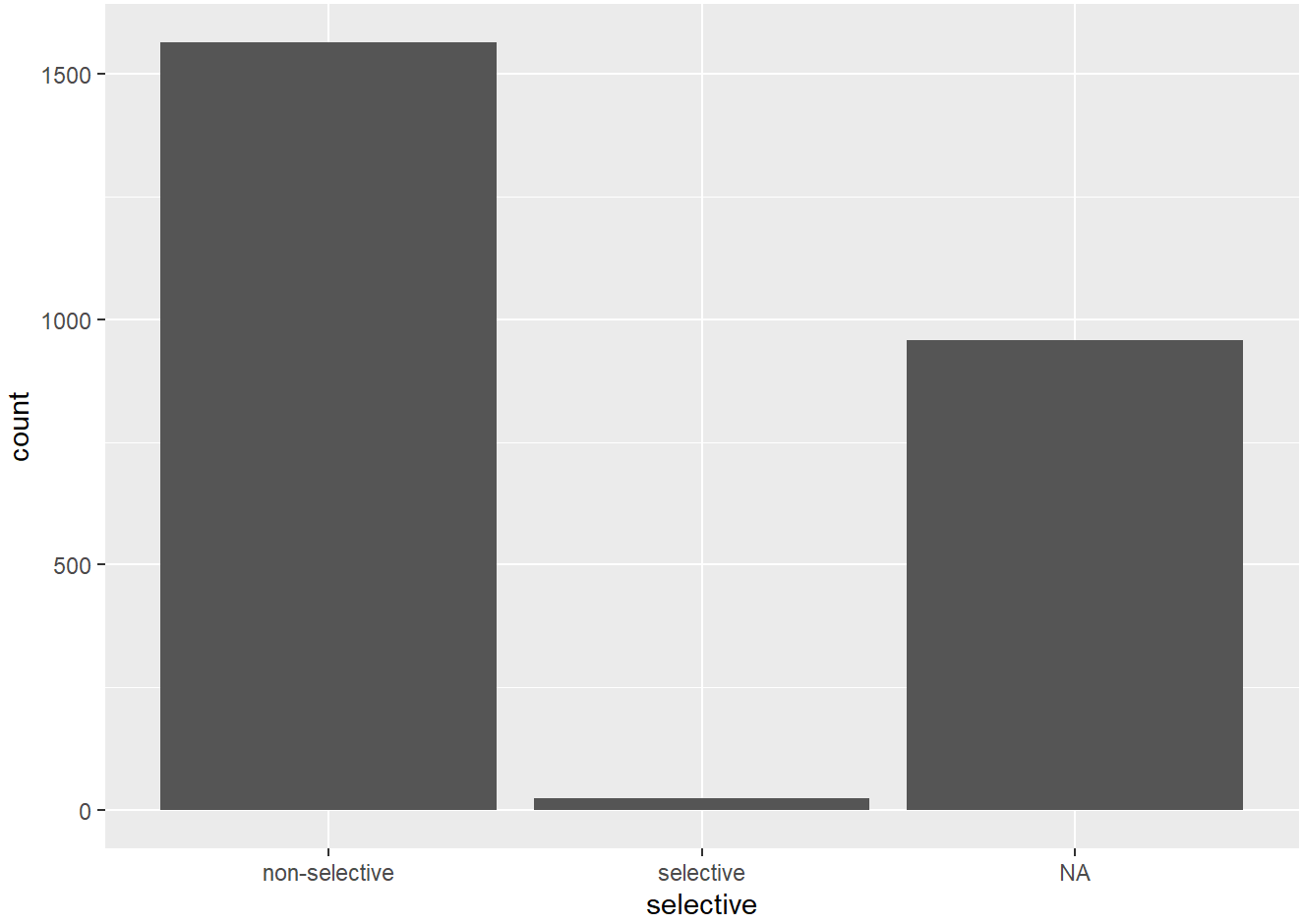


Introducing `geom_bar()`

```
df %>%
  group_by(selective) %>%
  summarise(avg_salary = mean(md_earn_wne_p6,na.rm=T)) %>%
  ggplot(aes(x = selective,
             y = avg_salary)) +
  geom_bar(stat = 'identity')
```




```
df %>%
  ggplot(aes(x = selective)) +
  geom_bar()
```



Data Wrangling:

Always start with an empty environment

```
rm(list = ls())
require(tidyverse)

MI_raw <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/MI2020_ExitPoll.rds")

MI_raw
```

```
## # A tibble: 1,231 × 63
##       ID WEIGHT LALVOTERID    GROUP   ZIP DISTRICT    Z1    S1   S2A   S2B    S3
##    <dbl>  <dbl> <chr>         <hvn> <dbl>    <dbl> <dbl> <hvn> <hvn> <hvn> <hvn>
##  1     9  0.405 LALMI6290066      3 49327        2    NA     1     2    NA     1
##  2    66  1.81  LALMI2492492      1 48234       14    NA     1     1     1    NA
##  3   225  0.860 LALMI5489814…     4 48301        9 48322     1     1     1    NA
##  4   243  0.199 LALMI5053772…     1 48130        7 48130     1     1     1    NA
##  5   286  0.177 LALMI6831689      1 49946        1    NA     1     1     2    NA
##  6   293  0.492 LALMI4019782      1 48615        4    NA     1     1     1    NA
##  7   365  1.37  LALMI4151378      1 48906        4 48813     1     1     1    NA
##  8   367  1.15  LALMI5912584      1 49442        2    NA     1     1     1    NA
##  9   388  1.50  LALMI6635050      1 48451        5    NA     1     2    NA     1
## 10   417  1.30  LALMI3567125      1 48197       12    NA     1     1     1    NA
## # ℹ 1,221 more rows
## # ℹ 52 more variables: S4 <hvn_lbl_>, VERSION <hvn_lbl_>, PRSMI20 <hvn_lbl_>,
## #   SENMI20 <hvn_lbl_>, TIME16 <hvn_lbl_>, ISSUE20 <hvn_lbl_>,
## #   QLT20 <hvn_lbl_>, TEMPBIDEN <hvn_lbl_>, TEMPTRUMP <hvn_lbl_>,
## #   CONTROLSEN <hvn_lbl_>, FINSIT <hvn_lbl_>, ECONVCORONA20 <hvn_lbl_>,
## #   FAVBIDEN2 <hvn_lbl_>, FAVTRUMP <hvn_lbl_>, FORCAND <hvn_lbl_>,
## #   NEWVOTER <hvn_lbl_>, NEC <hvn_lbl_>, HANDLEECON20 <hvn_lbl_>, …
```

Always LOOK at the data first

```
glimpse(MI_raw)
```

```
## Rows: 1,231
## Columns: 63
## $ ID              <dbl> 9, 66, 225, 243, 286, 293, 365, 367, 388, 417, 563, 572…
## $ WEIGHT          <dbl> 0.4045421, 1.8052619, 0.8601966, 0.1991648, 0.1772090, …
## $ LALVOTERID      <chr> "LALMI6290066", "LALMI2492492", "LALMI548981440", "LALM…
## $ GROUP           <hvn_lbl_> 3, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 9, 1, 4, 2, 1,…
## $ ZIP             <dbl> 49327, 48234, 48301, 48130, 49946, 48615, 48906, 49442,…
## $ DISTRICT        <dbl> 2, 14, 9, 7, 1, 4, 4, 2, 5, 12, 7, 2, 5, 9, 1, 5, 8, 2,…
## $ Z1              <dbl> NA, NA, 48322, 48130, NA, NA, 48813, NA, NA, NA, NA, NA…
## $ S1              <hvn_lbl_> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
## $ S2A             <hvn_lbl_> 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 2, 2, 1, 2,…
## $ S2B             <hvn_lbl_> NA, 1, 1, 1, 2, 1, 1, 1, NA, 1, NA, 1, 1, NA, NA, …
## $ S3              <hvn_lbl_> 1, NA, NA, NA, NA, NA, NA, NA, 1, NA, 1, NA, NA, 1…
## $ S4              <hvn_lbl_> 1, NA, NA, NA, NA, NA, NA, NA, 1, NA, 1, NA, NA, 1…
## $ VERSION         <hvn_lbl_> 1, 2, 2, 1, 1, 2, 2, 2, 1, 1, 1, 1, 1, 2, 1, 2, 2,…
## $ PRSMI20         <hvn_lbl_> 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 1, 2, 1, 1, 2,…
## $ SENMI20         <hvn_lbl_> 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 1, 2, 1, 1, 2,…
## $ TIME16          <hvn_lbl_> 5, NA, NA, 5, 5, NA, NA, NA, 5, 5, 5, 5, 5, NA, 3,…
## $ ISSUE20         <hvn_lbl_> 5, NA, NA, 9, 1, NA, NA, NA, 4, 9, 3, 3, 1, NA, 3,…
## $ QLT20           <hvn_lbl_> 4, NA, NA, 4, 3, NA, NA, NA, 3, 3, 2, 2, 3, NA, 1,…
## $ TEMPBIDEN       <hvn_lbl_> 1, NA, NA, 1, 1, NA, NA, NA, 2, 1, 2, 2, 1, NA, 1,…
## $ TEMPTRUMP       <hvn_lbl_> 2, NA, NA, 2, 2, NA, NA, NA, 1, 2, 1, 1, 2, NA, 1,…
## $ CONTROLSEN      <hvn_lbl_> 9, NA, NA, 1, 1, NA, NA, NA, 2, 1, 2, 2, 9, NA, 1,…
## $ FINSIT          <hvn_lbl_> 3, NA, NA, 3, 3, NA, NA, NA, 3, 3, 3, 3, 2, NA, 3,…
## $ ECONVCORONA20   <hvn_lbl_> 1, NA, NA, 1, 1, NA, NA, NA, 2, 1, 9, 2, 1, NA, 1,…
## $ FAVBIDEN2       <hvn_lbl_> 1, NA, NA, 1, 1, NA, NA, NA, 2, 1, 2, 2, 1, NA, 1,…
## $ FAVTRUMP        <hvn_lbl_> 2, NA, NA, 2, 2, NA, NA, NA, 1, 2, 1, 1, 2, NA, 1,…
## $ FORCAND         <hvn_lbl_> NA, 1, 1, NA, NA, 1, 2, 1, NA, NA, NA, NA, NA, 1, …
## $ NEWVOTER        <hvn_lbl_> NA, 2, 2, NA, NA, 2, 2, 2, NA, NA, NA, NA, NA, 2, …
## $ NEC             <hvn_lbl_> NA, 4, 4, NA, NA, 4, 4, 4, NA, NA, NA, NA, NA, 2, …
## $ HANDLEECON20    <hvn_lbl_> NA, 1, 1, NA, NA, 1, 1, 1, NA, NA, NA, NA, NA, 2, …
## $ HANDLECORONA20  <hvn_lbl_> NA, 1, 1, NA, NA, 1, 1, 1, NA, NA, NA, NA, NA, 2, …
## $ RACISM20        <hvn_lbl_> NA, 2, 2, NA, NA, 2, 2, 2, NA, NA, NA, NA, NA, 2, …
## $ VOTE2016        <hvn_lbl_> NA, 1, 1, NA, NA, 1, 1, 1, NA, NA, NA, NA, NA, 2, …
## $ COUNTACC        <hvn_lbl_> NA, 2, 2, NA, NA, 1, 2, 1, NA, NA, NA, NA, NA, 4, …
## $ TRUMP4          <hvn_lbl_> NA, 4, 4, NA, NA, 4, 4, 4, NA, NA, NA, NA, NA, 2, …
## $ CONTAINCOVID    <hvn_lbl_> NA, 4, 4, NA, NA, 4, 3, 4, NA, NA, NA, NA, NA, 2, …
## $ COVIDHARDSHIP   <hvn_lbl_> NA, 3, 1, NA, NA, 2, 2, 2, NA, NA, NA, NA, NA, 1, …
## $ AGE10           <hvn_lbl_> 2, 10, 7, 9, 8, 7, 9, 8, 6, 8, 9, 10, 1, 5, 9, 10,…
## $ SEX             <hvn_lbl_> 2, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 1, 1, 2, 1,…
## $ EDUC18          <hvn_lbl_> 4, 1, 5, 4, 5, 3, 3, 3, 4, 4, 5, 5, 4, 1, 1, 1, 5,…
## $ QRACEAI         <hvn_lbl_> 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 9, 1, 1, 1, 1, 1, 3,…
## $ LATINOS         <hvn_lbl_> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1,…
## $ PARTYID         <hvn_lbl_> 3, 1, 1, 3, 3, 3, 1, 1, 2, 1, 3, 2, 4, 4, 1, 1, 3,…
## $ PHIL3           <hvn_lbl_> 2, 2, 1, 9, 1, 2, 9, 2, 3, 2, 3, 3, 1, 3, 9, 2, 2,…
## $ INCOME20        <hvn_lbl_> 3, NA, NA, 9, 4, NA, NA, NA, 4, 4, 4, 2, 1, NA, 2,…
## $ BRNAGAIN        <hvn_lbl_> NA, 1, 2, NA, NA, 2, 1, 2, NA, NA, NA, NA, NA, 2, …
## $ CHILD12         <hvn_lbl_> NA, 2, 2, NA, NA, 1, 2, 1, NA, NA, NA, NA, NA, 1, …
## $ LGBT            <hvn_lbl_> NA, 2, 2, NA, NA, 2, 2, 2, NA, NA, NA, NA, NA, 2, …
## $ UNIONHH12       <hvn_lbl_> 2, NA, NA, 2, 2, NA, NA, NA, 2, 2, 2, 2, 2, NA, 2,…
## $ QN5             <hvn_lbl_> 3, 2, 3, 2, 2, 3, 2, 2, 3, 2, 2, 1, 4, 2, 2, 1, 3,…
## $ QN6A            <hvn_lbl_> NA, NA, 2, 1, NA, NA, 2, NA, NA, NA, NA, NA, 1, 2,…
```

```
## $ QN6B           <hvn_lbl_> 1, 1, NA, NA, 1, 1, NA, 1, 2, 1, 1, 1, NA, NA, 1, …
## $ QN6C           <hvn_lbl_> 4, 0, 0, 0, 2, 1, 0, 2, 1, 1, 2, 1, 4, 0, 1, 1, 1,…
## $ County         <chr> "NEWAYGO", "WAYNE", "OAKLAND", "WASHTENAW", "BARAGA", "…
## $ SMPFIPS        <dbl> 26123, 26163, 26125, 26161, 26013, 26057, 26037, 26121,…
## $ TTID           <dbl> 94989, 94079, 94040, 94468, 94564, 93952, 93930, 94021,…
## $ GEOCODE        <dbl> 5, 1, 2, 2, 5, 3, 2, 5, 2, 2, 3, 4, 3, 3, 5, 3, 2, 5, 2…
## $ a1age5         <dbl> 1, 5, 4, 5, 4, 4, 5, 4, 3, 4, 5, 5, 1, 3, 5, 5, 4, 2, 1…
## $ A1RACE         <hvn_lbll> 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 3, 1, 1, 1, 1, 1, 3,…
## $ A1RACE3        <hvn_lbll> 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 3, 1, 1, 1, 1, 1, 3,…
## $ A1SEX          <hvn_lbll> 2, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 1, 1, 2, 1,…
## $ A1EDUC         <hvn_lbll> 2, 1, 2, 2, 2, 1, 1, 1, 2, 2, 2, 2, 2, 1, 1, 1, 2,…
## $ EDUCWHITE      <hvn_lbll> 1, 4, 1, 1, 1, 2, 2, 2, 1, 3, 3, 1, 1, 2, 2, 2, 3,…
## $ VOTEMETHOD     <dbl> 1, 2, 2, 2, 3, 2, 2, 2, 1, 2, 1, 2, 2, 1, 1, 3, 3, 2, 1…
```

```
View(MI_raw)
```

Research Question: Do older people support Trump?

```
# X variable: age
colnames(MI_raw)
```

```
##  [1] "ID"             "WEIGHT"         "LALVOTERID"     "GROUP"
##  [5] "ZIP"            "DISTRICT"       "Z1"             "S1"
##  [9] "S2A"            "S2B"            "S3"             "S4"
## [13] "VERSION"        "PRSMI20"        "SENMI20"        "TIME16"
## [17] "ISSUE20"        "QLT20"          "TEMPBIDEN"      "TEMPTRUMP"
## [21] "CONTROLSEN"     "FINSIT"         "ECONVCORONA20"  "FAVBIDEN2"
## [25] "FAVTRUMP"       "FORCAND"        "NEWVOTER"       "NEC"
## [29] "HANDLEECON20"   "HANDLECORONA20" "RACISM20"       "VOTE2016"
## [33] "COUNTACC"       "TRUMP4"         "CONTAINCOVID"   "COVIDHARDSHIP"
## [37] "AGE10"          "SEX"            "EDUC18"         "QRACEAI"
## [41] "LATINOS"        "PARTYID"        "PHIL3"          "INCOME20"
## [45] "BRNAGAIN"       "CHILD12"        "LGBT"           "UNIONHH12"
## [49] "QN5"            "QN6A"           "QN6B"           "QN6C"
## [53] "County"         "SMPFIPS"        "TTID"           "GEOCODE"
## [57] "a1age5"         "A1RACE"         "A1RACE3"        "A1SEX"
## [61] "A1EDUC"         "EDUCWHITE"      "VOTEMETHOD"
```

```
MI_raw %>%
  select(AGE10)
```

```
## # A tibble: 1,231 × 1
##         AGE10
##    <hvn_lbl_>
##  1          2
##  2         10
##  3          7
##  4          9
##  5          8
##  6          7
##  7          9
##  8          8
##  9          6
## 10          8
## # ℹ 1,221 more rows
```

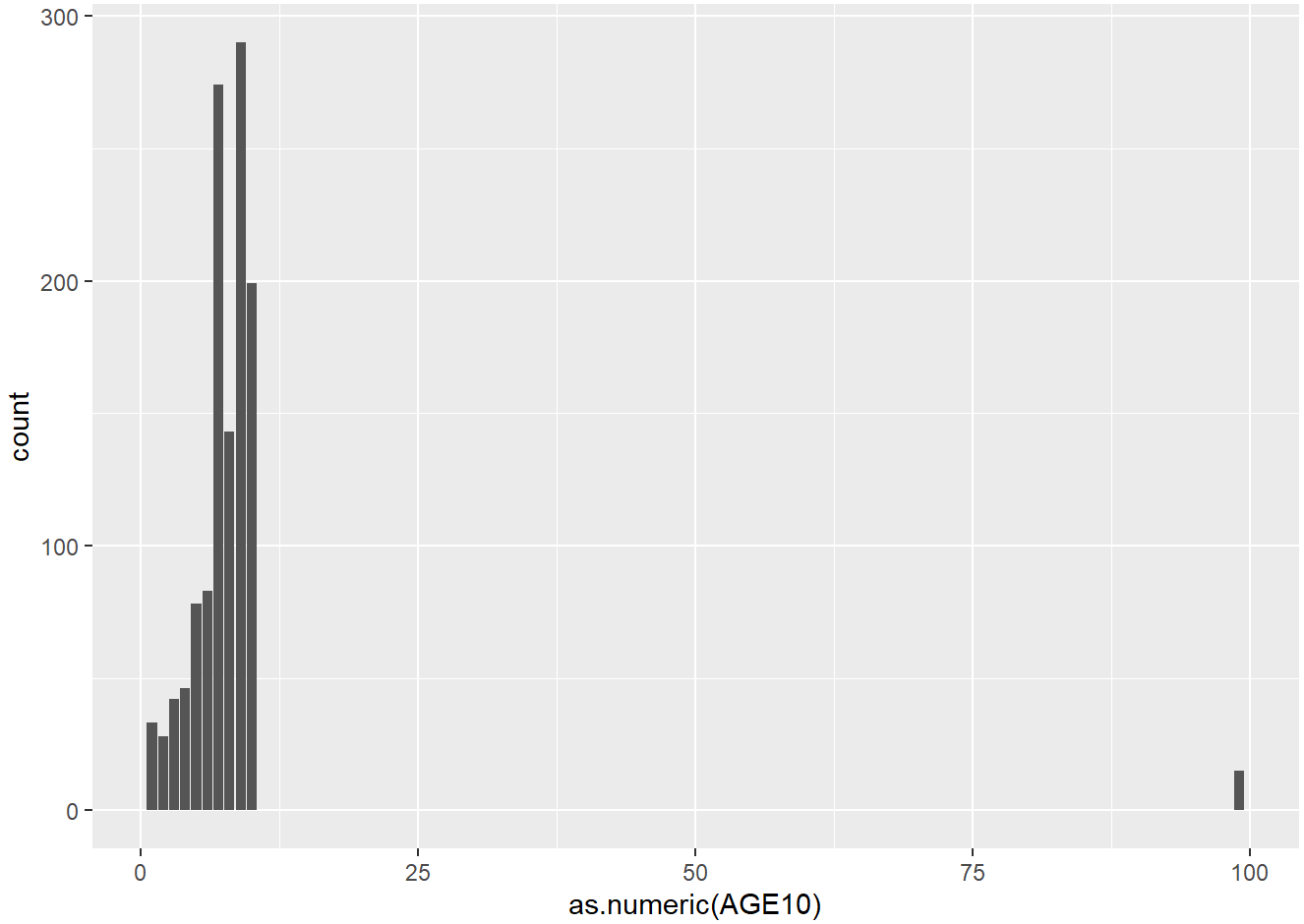
```
MI_raw %>%
  count(AGE10)
```

```
## # A tibble: 11 × 2
##         AGE10     n
##    <hvn_lbl_> <int>
##  1          1    33
##  2          2    28
##  3          3    42
##  4          4    46
##  5          5    78
##  6          6    83
##  7          7   274
##  8          8   143
##  9          9   290
## 10         10   199
## 11         99    15
```

```
# Visualize with geom_bar()
require(haven) # This is required for certain environments / package versions to convert SPSS to numeric. See TIME OUT section below and associated ChatGPT help.
```

```
## Loading required package: haven
```

```
MI_raw %>%
  ggplot(aes(x = as.numeric(AGE10))) +
  geom_bar()
```



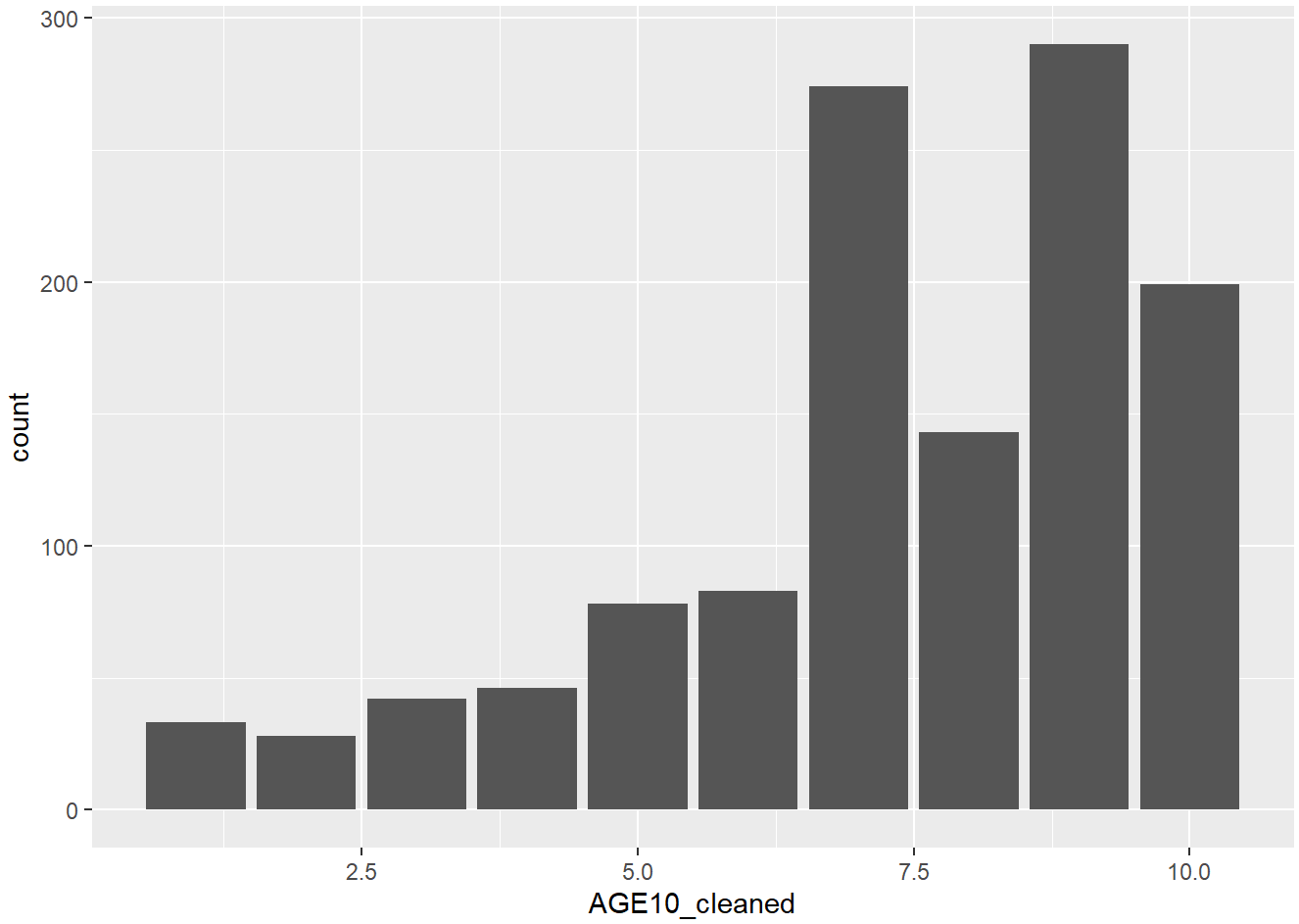

```
MI_raw %>%
  arrange(desc(AGE10)) %>%
  select(AGE10)
```

```
## # A tibble: 1,231 × 1
##    AGE10
##    <dbl+lbl>
##  1 99 [[DON'T READ] Refused]
##  2 99 [[DON'T READ] Refused]
##  3 99 [[DON'T READ] Refused]
##  4 99 [[DON'T READ] Refused]
##  5 99 [[DON'T READ] Refused]
##  6 99 [[DON'T READ] Refused]
##  7 99 [[DON'T READ] Refused]
##  8 99 [[DON'T READ] Refused]
##  9 99 [[DON'T READ] Refused]
## 10 99 [[DON'T READ] Refused]
## # ℹ 1,221 more rows
```

```
# We have "unit non-response"
MI_raw <- MI_raw %>%
  mutate(AGE10_cleaned = ifelse(AGE10 == 99,
                                NA,
                                AGE10))

MI_raw %>%
  ggplot(aes(x = AGE10_cleaned)) +
  geom_bar()
```

```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (`stat_count()`).
```




```
# Y variable: vote choice
colnames(MI_raw)
```

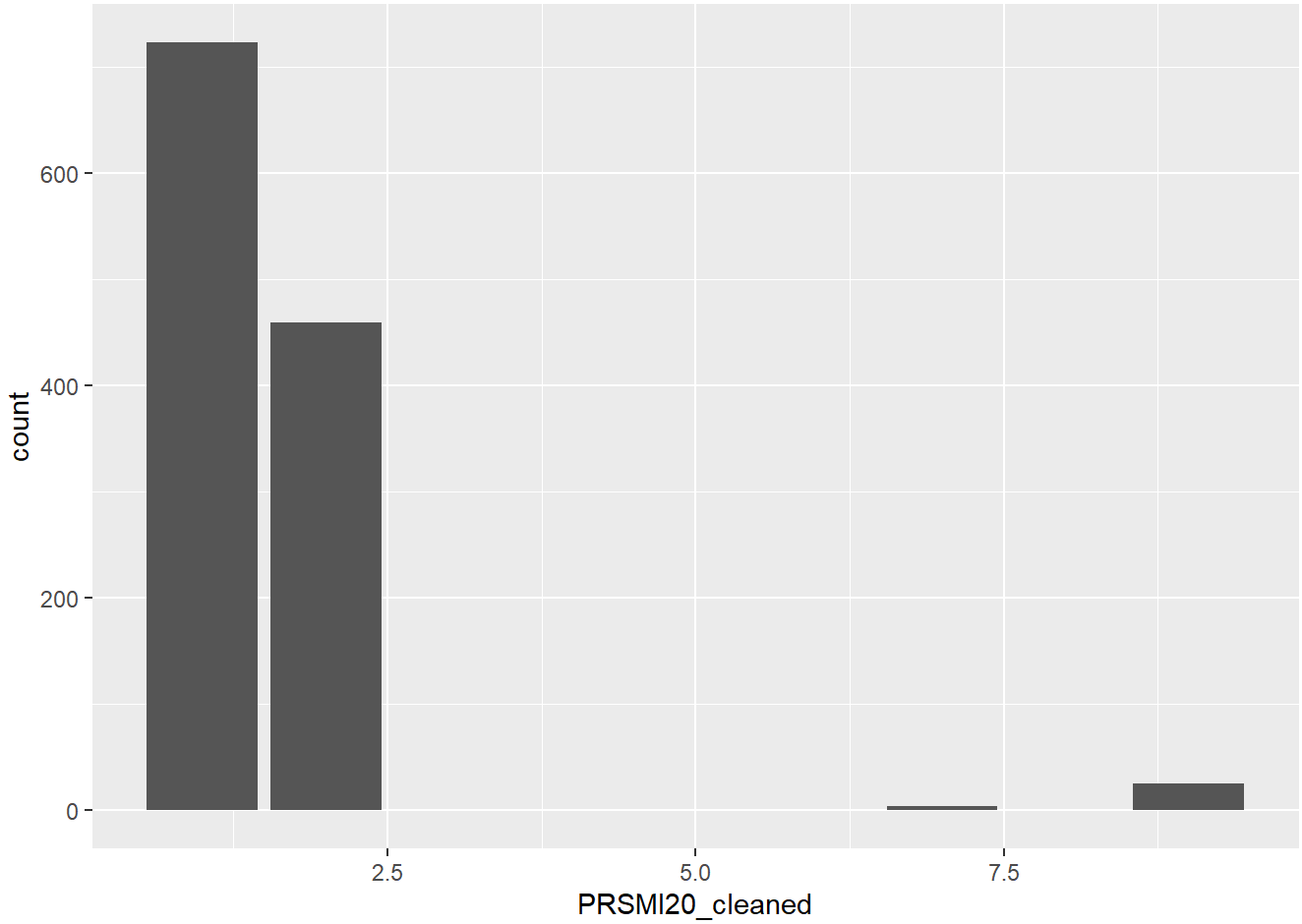
```
##  [1] "ID"             "WEIGHT"         "LALVOTERID"     "GROUP"
##  [5] "ZIP"            "DISTRICT"       "Z1"             "S1"
##  [9] "S2A"            "S2B"            "S3"             "S4"
## [13] "VERSION"        "PRSMI20"        "SENMI20"        "TIME16"
## [17] "ISSUE20"        "QLT20"          "TEMPBIDEN"      "TEMPTRUMP"
## [21] "CONTROLSEN"     "FINSIT"         "ECONVCORONA20"  "FAVBIDEN2"
## [25] "FAVTRUMP"       "FORCAND"        "NEWVOTER"       "NEC"
## [29] "HANDLEECON20"   "HANDLECORONA20" "RACISM20"       "VOTE2016"
## [33] "COUNTACC"       "TRUMP4"         "CONTAINCOVID"   "COVIDHARDSHIP"
## [37] "AGE10"          "SEX"            "EDUC18"         "QRACEAI"
## [41] "LATINOS"        "PARTYID"        "PHIL3"          "INCOME20"
## [45] "BRNAGAIN"       "CHILD12"        "LGBT"           "UNIONHH12"
## [49] "QN5"            "QN6A"           "QN6B"           "QN6C"
## [53] "County"         "SMPFIPS"        "TTID"           "GEOCODE"
## [57] "a1age5"         "A1RACE"         "A1RACE3"        "A1SEX"
## [61] "A1EDUC"         "EDUCWHITE"      "VOTEMETHOD"     "AGE10_cleaned"
```

```
MI_raw %>%
  count(PRSMI20)
```

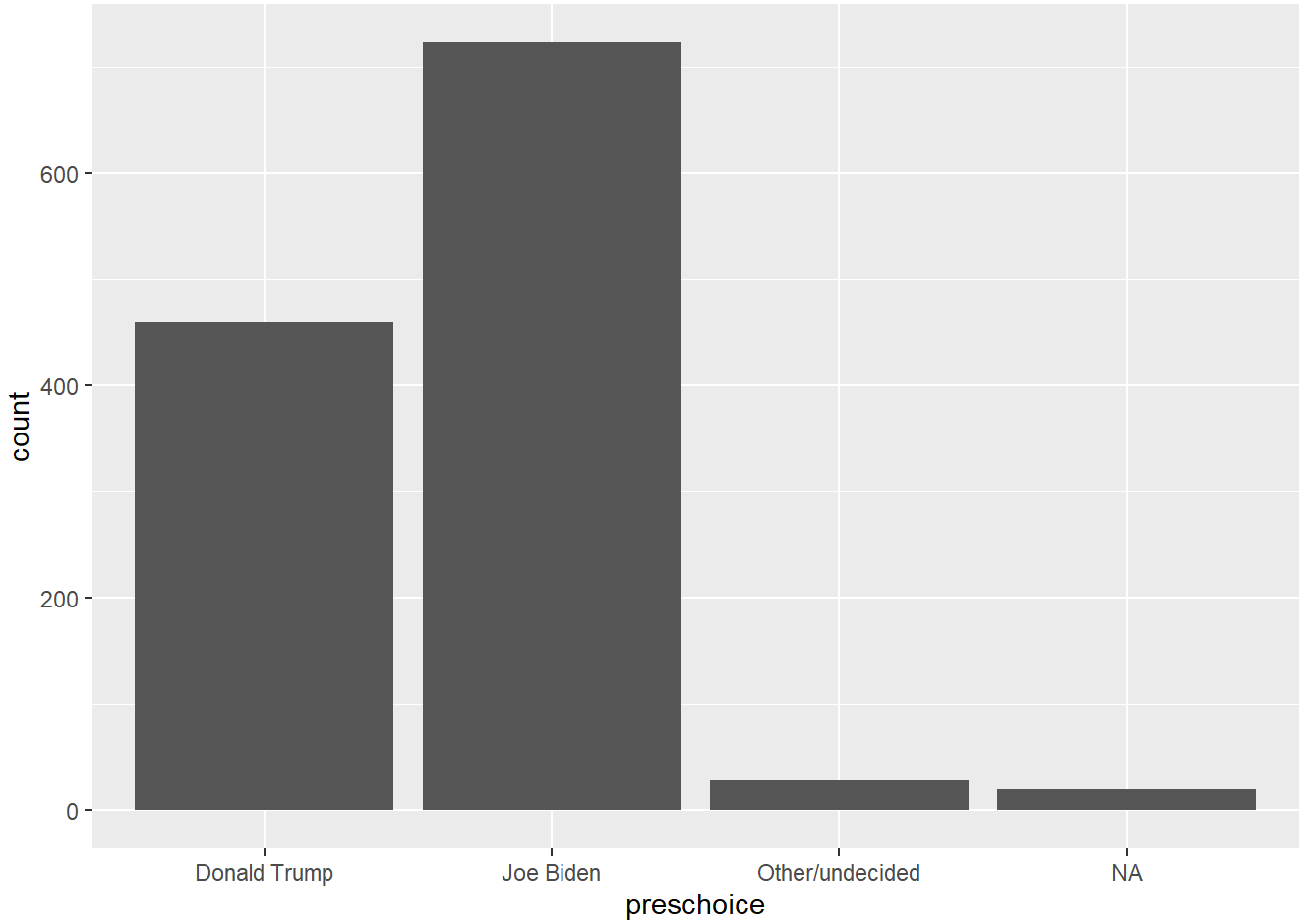
```
## # A tibble: 6 × 2
##   PRSMI20                                        n
##   <dbl+lbl>                                  <int>
## 1 0 (NA) [Will/Did not vote for president]     6
## 2 1 [Joe Biden, the Democrat]                  723
## 3 2 [Donald Trump, the Republican]             459
## 4 7 [Undecided/Don’t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

```
# Wrangling
MI_raw <- MI_raw %>%
  mutate(PRSMI20_cleaned = ifelse(PRSMI20 == 0,
                                  NA,
                                  ifelse(PRSMI20 == 8,
                                         NA,
                                         PRSMI20)))
MI_raw %>%
  ggplot(aes(x = PRSMI20_cleaned)) +
  geom_bar()
```

```
## Warning: Removed 20 rows containing non-finite outside the scale range
## (`stat_count()`).
```

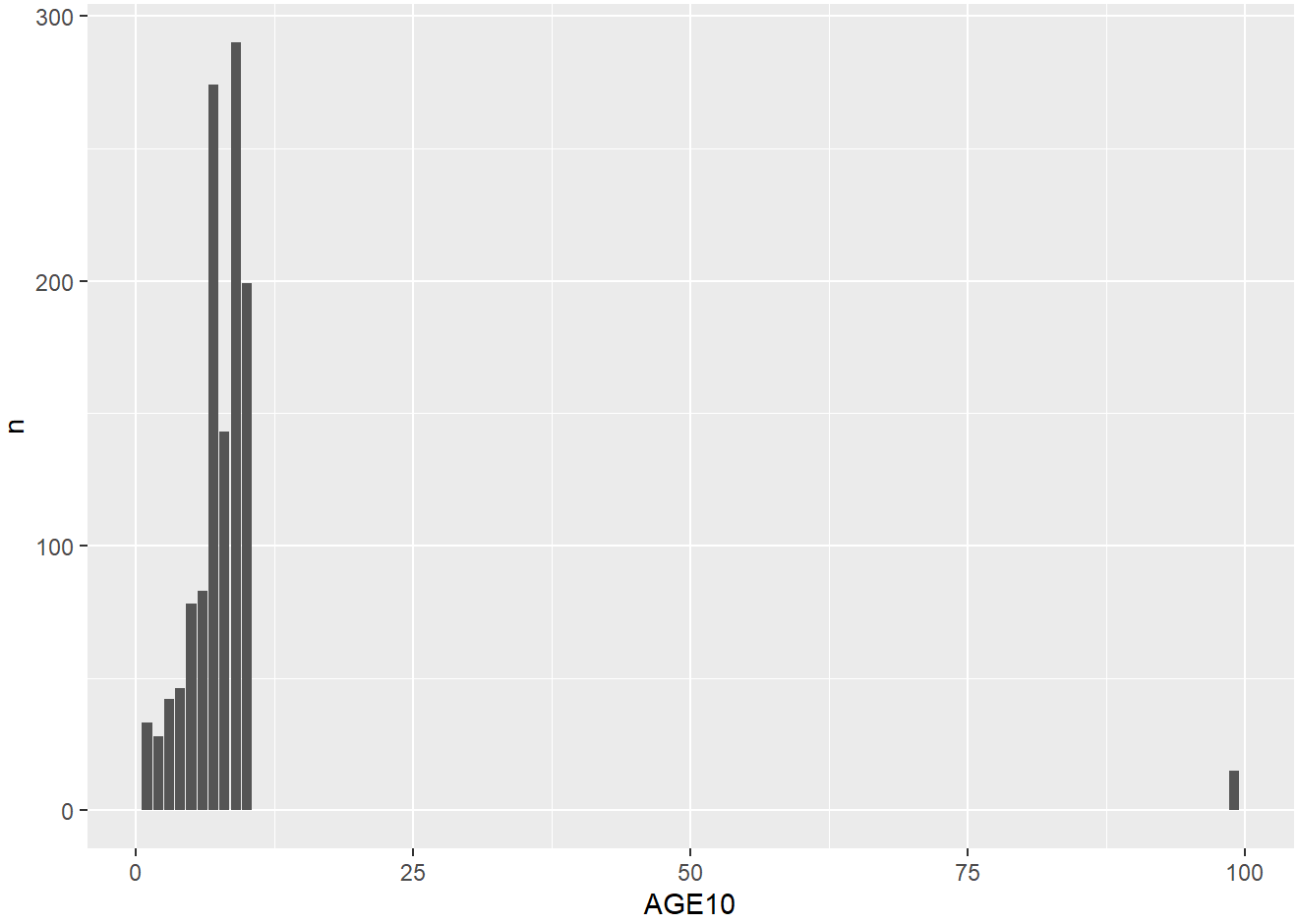



```
# Different way to categorize
MI_raw <- MI_raw %>%
  mutate(preschoice = ifelse(PRSMI20_cleaned == 1,
                             'Joe Biden',
                             ifelse(PRSMI20_cleaned == 2,
                                    'Donald Trump',
                                    'Other/undecided')))
MI_raw %>%
  ggplot(aes(x = preschoice)) +
  geom_bar()
```



TIME OUT for type errors with most of class

```
MI_raw %>%
  count(AGE10) %>%
  ggplot(aes(x = AGE10,
             y = n)) +
  geom_bar(stat = 'identity')
```



Trying chatGPT solution: https://chatgpt.com/share/a4d8867b-e436-4108-bf88-e0fa53f7a6e1 (https://chatgpt.com/share/a4d8867b-e436-4108-bf88-e0fa53f7a6e1)

```
require(haven)

MI_raw %>%
  mutate(AGE10 = as.numeric(AGE10))
```

```
## # A tibble: 1,231 × 66
##       ID WEIGHT LALVOTERID     GROUP       ZIP DISTRICT    Z1 S1
##    <dbl>  <dbl> <chr>          <dbl+lbl> <dbl>    <dbl> <dbl> <dbl+lbl>
##  1     9  0.405 LALMI6290066   3 [3]     49327        2    NA 1 [Yes]
##  2    66  1.81  LALMI2492492   1 [1]     48234       14    NA 1 [Yes]
##  3   225  0.860 LALMI548981440 4 [4]     48301        9 48322 1 [Yes]
##  4   243  0.199 LALMI505377239 1 [1]     48130        7 48130 1 [Yes]
##  5   286  0.177 LALMI6831689   1 [1]     49946        1    NA 1 [Yes]
##  6   293  0.492 LALMI4019782   1 [1]     48615        4    NA 1 [Yes]
##  7   365  1.37  LALMI4151378   1 [1]     48906        4 48813 1 [Yes]
##  8   367  1.15  LALMI5912584   1 [1]     49442        2    NA 1 [Yes]
##  9   388  1.50  LALMI6635050   1 [1]     48451        5    NA 1 [Yes]
## 10   417  1.30  LALMI3567125   1 [1]     48197       12    NA 1 [Yes]
## # ℹ 1,221 more rows
## # ℹ 58 more variables: S2A <dbl+lbl>, S2B <dbl+lbl>, S3 <dbl+lbl>,
## #   S4 <dbl+lbl>, VERSION <dbl+lbl>, PRSMI20 <dbl+lbl>, SENMI20 <dbl+lbl>,
## #   TIME16 <dbl+lbl>, ISSUE20 <dbl+lbl>, QLT20 <dbl+lbl>, TEMPBIDEN <dbl+lbl>,
## #   TEMPTRUMP <dbl+lbl>, CONTROLSEN <dbl+lbl>, FINSIT <dbl+lbl>,
## #   ECONVCORONA20 <dbl+lbl>, FAVBIDEN2 <dbl+lbl>, FAVTRUMP <dbl+lbl>,
## #   FORCAND <dbl+lbl>, NEWVOTER <dbl+lbl>, NEC <dbl+lbl>, …
```