Homework #02: Data Wrangling and Joins

due [date] 11:59 PM

Ayush Jain

02/02

Load Packages and Data

```
library(tidyverse)
library(viridis)
```

```
natunivs <- read_csv("NatUnivs.csv")
slacs <- read_csv("SLACs.csv")
presvote_pop <- read_csv("PresVote_Population.csv")
```

Exercise 1

```
full_data <- natunivs %>%
  full_join(slacs) %>%
  left_join(presvote_pop, by = c("state" = "abbrev"))
```

```
## Joining, by = c("school", "state", "rank_2022", "rank_2021", "natuniv_slac")
```

Exercise 2

```
full_data %>%
  group_by(state)%>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   state count
##   <chr> <int>
## 1 CA       18
## 2 MA       13
## 3 NY       11
## 4 PA       11
## 5 OH        5
```

Answer: The states with the most schools are California (18), Massachusetts (13), New York (11), Philadelphia (11), Ohio (5)

Exercise 3

```
presvote_pop %>%
  anti_join(full_data, by = c("abbrev" = "state")) %>%
  arrange(desc( `2020pop`)) %>%
  select(abbrev, `2020pop`)
```

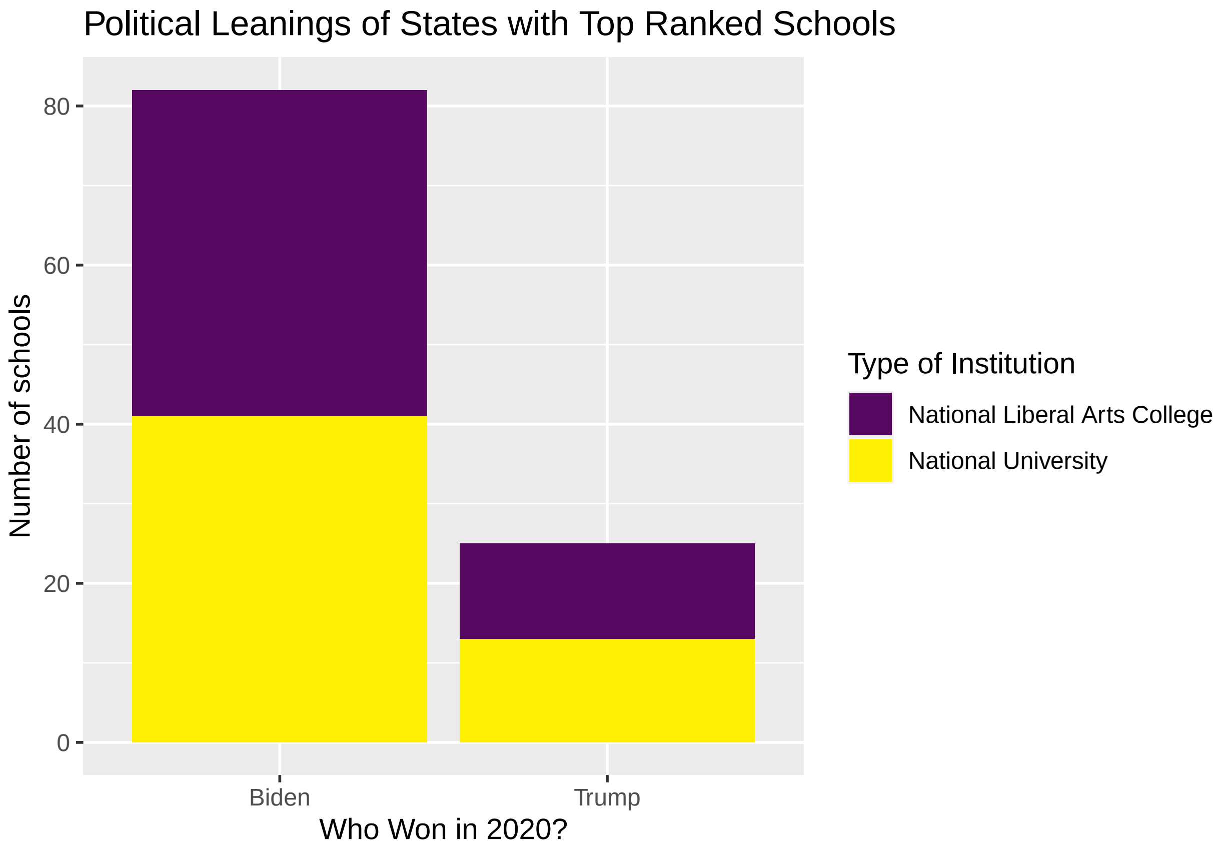
```
## # A tibble: 20 x 2
##    abbrev `2020pop`
##    <chr>      <dbl>
##  1 AZ       7151502
##  2 AL       5024279
##  3 OR       4237256
##  4 OK       3959353
##  5 UT       3271616
##  6 NV       3104614
##  7 AR       3011524
##  8 MS       2961279
##  9 KS       2937880
## 10 NM       2117522
## 11 NE       1961504
## 12 ID       1839106
## 13 WV       1793716
## 14 HI       1455271
## 15 MT       1084225
## 16 DE        989948
## 17 SD        886667
## 18 ND        779094
## 19 AK        733391
## 20 WY        576851
```

Answer: The state with the greatest population that does not have a school in the data set is Arizona (AZ) with a population of 7151502

Exercise 4

```
full_data %>%
  mutate(winner =
           ifelse(trumpvotes > bidenvotes, "Trump", "Biden")) %>%

ggplot(mapping = aes(x = winner, fill = natuniv_slac)) +
  geom_bar()+
    labs(title =
           "Political Leanings of States with Top Ranked Schools",
         x = "Who Won in 2020?",
         y = "Number of schools") +
  scale_fill_manual(name = "Type of Institution", values=c("#570861","#FFEF00"))
```

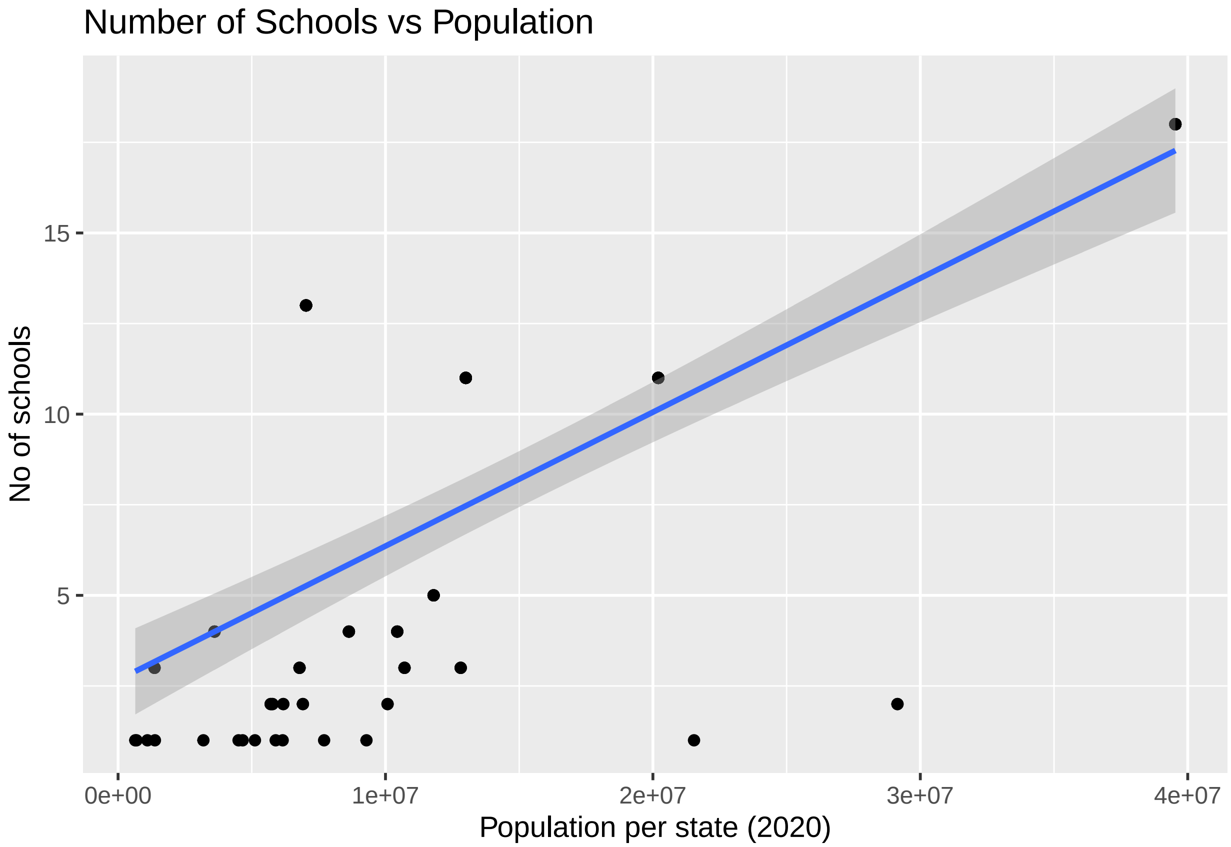


Exercise 5

```
counts <- full_data %>%
  group_by(state)%>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  right_join(full_data, by = c("state" = "state"))

  ggplot(data = counts,
         mapping = aes(x = `2020pop`, y = count)) +
  geom_point()+
    geom_smooth(method = lm) +
    labs(title = "Number of Schools vs Population",
         x = "Population per state (2020)",
         y = "No of schools")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Answer: There is a positive relation: as the population increases, the count does too. However, most of the points are not near the line which indicates that although the line of best fit does show a positive relation, it is a weak relation, with many points not fitting the trend

Exercise 6

```
full_data %>%
  group_by(state) %>%
  filter(state == "NC") %>%
  mutate(change = rank_2021 - rank_2022) %>%
  summarise(school, change)
```

```
## ‘summarise()‘ has grouped output by ’state’. You can override using the ‘.groups‘ argument.
```

```
## # A tibble: 4 x 3
## # Groups:   state [1]
##   state school                                   change
##   <chr> <chr>                                     <dbl>
## 1 NC    Duke University                               3
## 2 NC    University of North Carolina-Chapel Hill      0
## 3 NC    Wake Forest University                        0
## 4 NC    Davidson College                              2
```

Duke improved by 3 positions (Go Duke!), and Davidson College improved by 2 positions. Wake Forest and UNC did not change.

Exercise 7

```
full_data %>%
  mutate(bidenVote = ((bidenvotes)/(bidenvotes+trumpvotes)) * 100) %>%
  group_by(natuniv_slac) %>%
  summarise(meanvote = mean(bidenVote), meanpop = mean(`2020pop`))
```

```
## # A tibble: 2 x 3
##   natuniv_slac                  meanvote   meanpop
##   <chr>                            <dbl>     <dbl>
## 1 National Liberal Arts College     56.3 14018730.
## 2 National University               57.2 16101703.
```

Answer: No, the politics and population of National Liberal Arts Colleges do not differ much from National Universities