Homework #02: Data Wrangling and Joins

due [date] 11:59 PM

Dav King

1/28/22

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 <- slacs %>%
  full_join(natunivs, by = c("school", "state",
                             "rank_2022", "rank_2021", "natuniv_slac")) %>%
  inner_join(presvote_pop, by = c("state" = "abbrev"))
```

Exercise 2

```
full_data %>%
  group_by(state) %>%
  summarize(nSchools = n()) %>%
  arrange(desc(nSchools)) %>%
  slice(1:5) %>%
  select(state)
```

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

The 5 states which have the most schools in the `full_data` data set are California, Massachusetts, New York, Pennsylvania, and Ohio, respectively.

Exercise 3

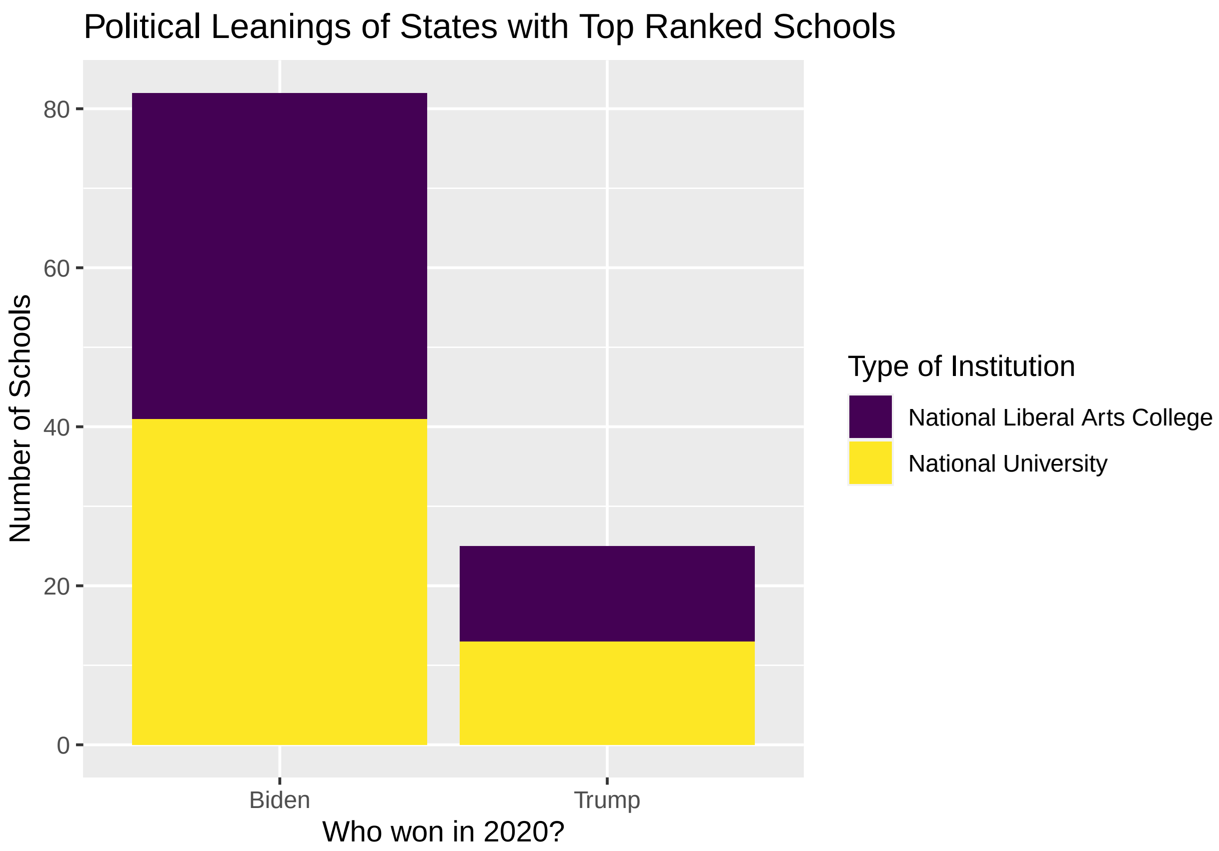
```
presvote_pop %>%
  anti_join(full_data, by = c("abbrev" = "state")) %>%
  arrange(desc(`2020pop`)) %>%
  select("abbrev", "2020pop")
```

```
## # 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
```

The state with the largest population that does not have a school in `full_data` is Arizona.

Exercise 4

```
full_data %>%
  mutate(winner = if_else(bidenvotes > trumpvotes, "Biden", "Trump")) %>%
  ggplot(., 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",
       fill = "Type of Institution") +
  scale_fill_viridis(discrete = TRUE, option = "D")
```

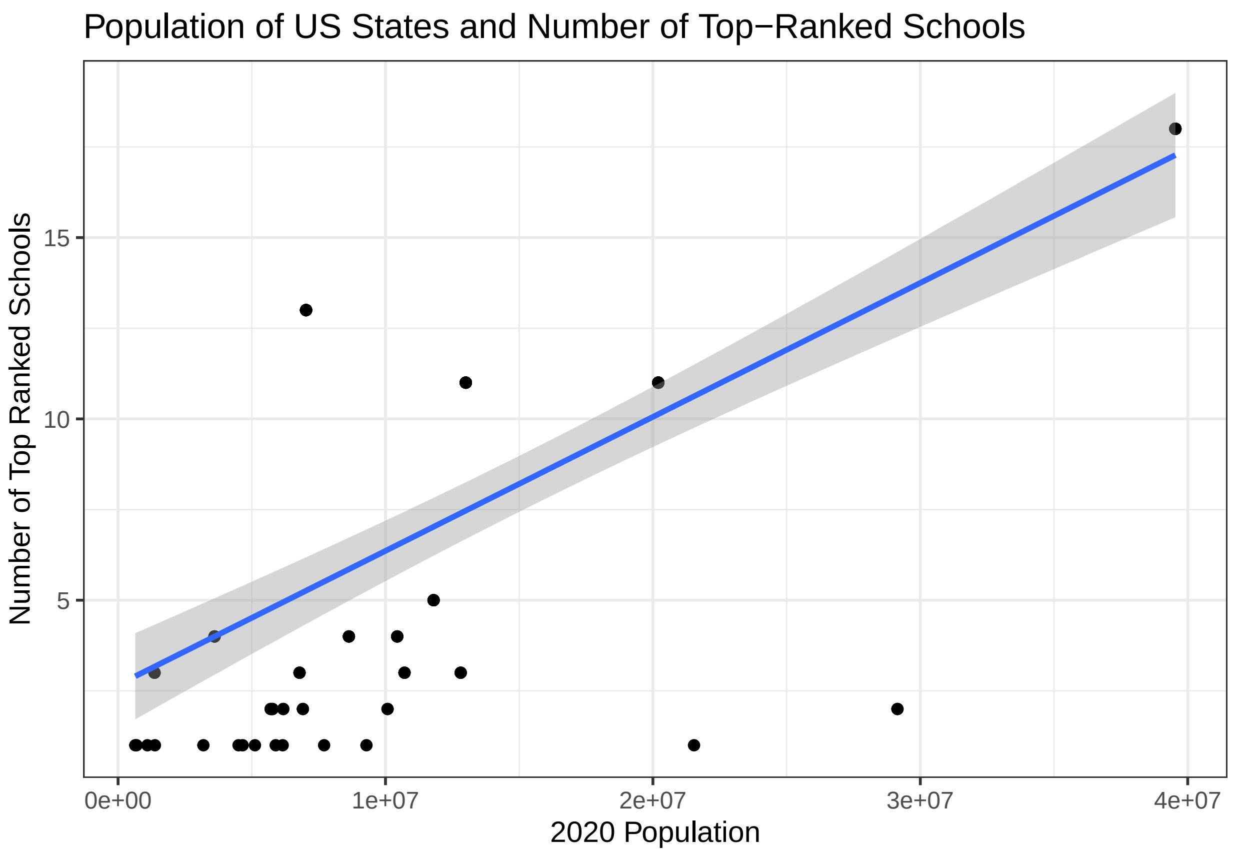


In general, the states that Biden won contained far more top ranked schools (over 80) than the states Trump won (roughly 25). In general, the proportion overall seems to be about 50/50 between National Liberal Arts Colleges and National Universities, with no notable difference in their distribution in Trump versus Biden states.

Exercise 5

```
counts <- full_data %>%
  group_by(state) %>%
  mutate(count = n())
full_data %>%
  full_join(y=counts, by = c("school")) %>%
  ggplot(., aes(x = `2020pop.x`, y = count)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(title = "Population of US States and Number of Top-Ranked Schools",
       x = "2020 Population", y = "Number of Top Ranked Schools") +
  theme_bw()
```

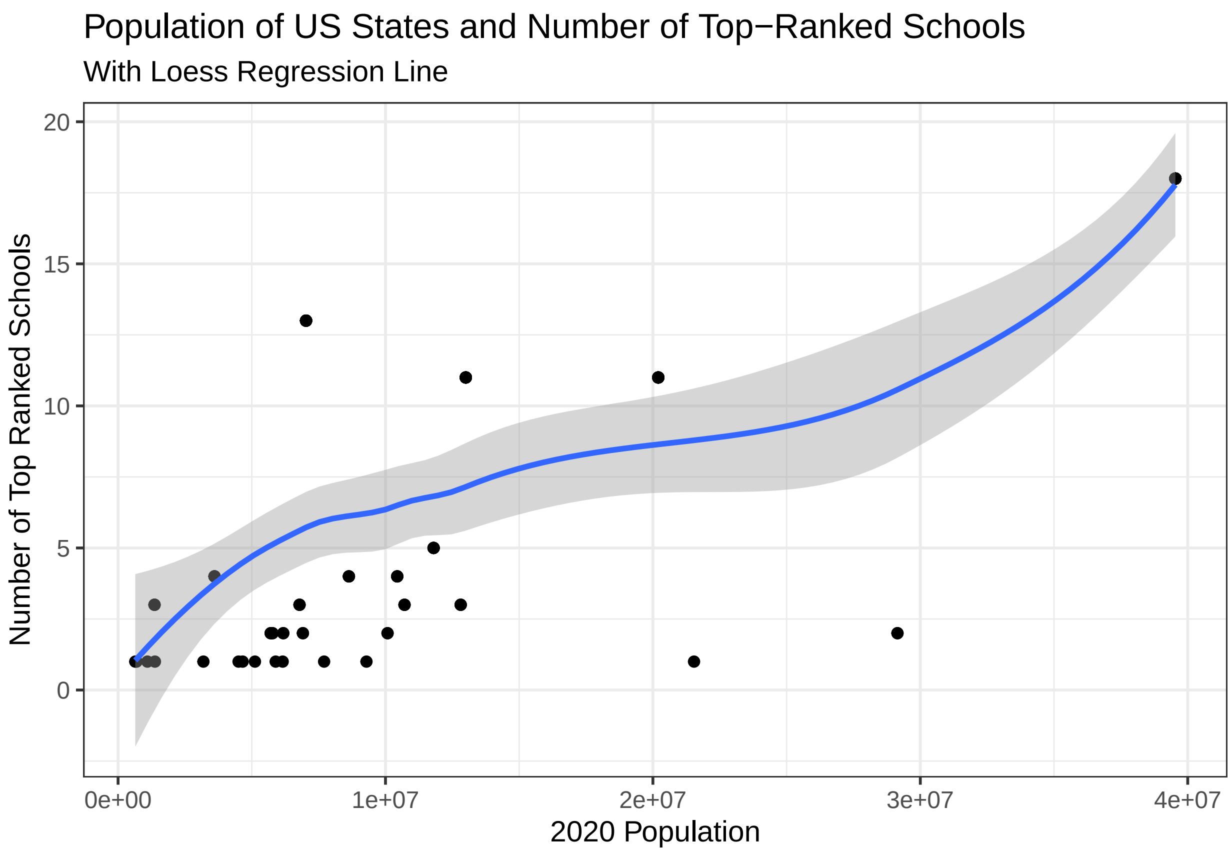
```
## 'geom_smooth()' using formula 'y ~ x'
```



This graph, using a linear regression line, shows a clear positive relationship between 2020 population and number of top ranked schools in a state. However, this appears to be something of an overplotting error. The data are not very linearly distributed, and the linear regression line owes most of its strength to a few outliers than it does a true underlying data structure (which there may well be in a relationship with a discrete variable like a number of schools). Consider, for example, the graph below - fitted with a Loess regression line instead of a linear one. It suggests that the relationship is almost cubic - certainly, not a standard linear relationship.

```
full_data %>%
  full_join(y=counts, by = c("school")) %>%
  ggplot(., aes(x = `2020pop.x`, y = count)) +
  geom_point() +
  geom_smooth(method = loess) +
  labs(title = "Population of US States and Number of Top-Ranked Schools",
       subtitle = "With Loess Regression Line",
       x = "2020 Population", y = "Number of Top Ranked Schools") +
  theme_bw()
```

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



Exercise 6

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

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

There are four schools in North Carolina in this dataset of top schools - Duke University, Davidson College, Wake Forest University, and our neighbors eight miles down the road. The “University” of North Carolina and Wake Forest did not change their rankings from 2021 to 2022, while Davidson improved 2 places and Duke increased by 3 (the Duke Difference).

Exercise 7

```
full_data %>%
  mutate(biden_share = bidenvotes / (bidenvotes + trumpvotes)) %>%
  group_by(natuniv_slac) %>%
  summarize(mean_Biden = mean(biden_share), mean_pop = mean(`2020pop`))
```

```
## # A tibble: 2 x 3
##   natuniv_slac                  mean_Biden  mean_pop
##   <chr>                              <dbl>     <dbl>
## 1 National Liberal Arts College      0.563 14018730.
## 2 National University                0.572 16101703.
```

There seem to be no real differences in Biden vote share between the two groups of colleges - 56.3% of the vote in one and 57.2% of the vote in the other. However, there is a notable difference in how populous th states containing the two types of schools are. States with National Liberal Arts Colleges tend to average around 14 million people, while states with National Universities tend to average around 16 million.