Intro to R Review Notes

Prof. Bisbee, Vanderbilt University

2023-09-13

Start with the basic

```
require(tidyverse)
```

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

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.2.1     ✓ dplyr   1.1.2
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
```

```
## Warning: package 'tibble' was built under R version 4.2.3
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
df <- read_rds("https://github.com/jbisbee1/DS1000_F2023/raw/main/Lectures/2_Intro_to_R/data/sc_debt.Rds")
```

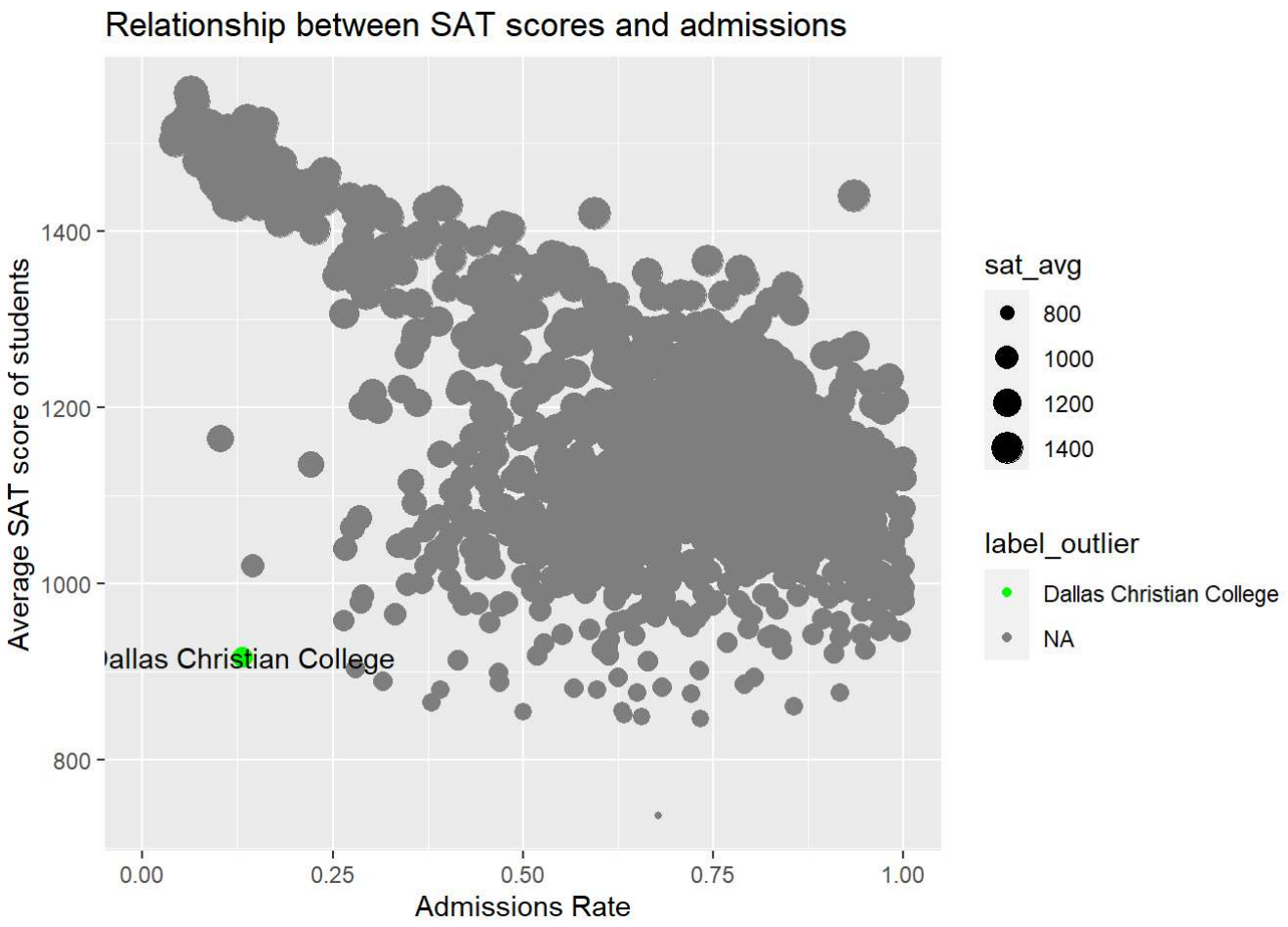
Visualize

- Show the relationship between SAT scores and admissions rates

```
df %>%
  mutate(label_outlier = ifelse(adm_rate < .2 & sat_avg < 1000,instnm,NA)) %>%
  ggplot(aes(x = adm_rate,y = sat_avg,label = label_outlier)) +
  geom_point(aes(color = label_outlier,size = sat_avg)) +
  geom_text() +
  labs(x = "Admissions Rate",
       y = "Average SAT score of students",
       title = "Relationship between SAT scores and admissions") +
  scale_color_manual(values = c('green','black'))
```

```
## Warning: Removed 1317 rows containing missing values (geom_point).
```

```
## Warning: Removed 2545 rows containing missing values (geom_text).
```



Filter work

```
df %>%
  filter((stabbr == "TX") & grepl("Community",instnm)) %>%
  select(instnm,stabbr)
```

```
## # A tibble: 9 × 2
##   instnm                                stabbr
##   <chr>                                 <chr>
## 1 Alvin Community College               TX
## 2 Austin Community College District     TX
## 3 El Paso Community College             TX
## 4 Trinity Valley Community College      TX
## 5 Houston Community College             TX
## 6 McLennan Community College            TX
## 7 Northeast Texas Community College     TX
## 8 San Jacinto Community College         TX
## 9 Collin County Community College District TX
```

```
df %>%
  filter(stabbr == "TX") %>%
  filter(str_detect(instnm,"Community"))
```

```
## # A tibble: 9 × 16
##   unitid instnm    stabbr grad_debt_mdn control region preddeg openadmp adm_rate
##    <int> <chr>     <chr>          <int> <chr>   <chr>  <chr>      <int>    <dbl>
## 1 222567 Alvin Co… TX              5750 Public  South… Associ…        1       NA
## 2 222992 Austin C… TX              9708 Public  South… Associ…        1       NA
## 3 224642 El Paso … TX              5750 Public  South… Associ…        1       NA
## 4 225308 Trinity … TX             10250 Public  South… Associ…        1       NA
## 5 225423 Houston … TX             13892 Public  South… Associ…        1       NA
## 6 226578 McLennan… TX             12000 Public  South… Associ…        1       NA
## 7 227225 Northeas… TX             12250 Public  South… Associ…        1       NA
## 8 227979 San Jaci… TX              9551 Public  South… Associ…        1       NA
## 9 247834 Collin C… TX              8478 Public  South… Associ…        1       NA
## # ℹ 7 more variables: ccbasic <int>, sat_avg <int>, md_earn_wne_p6 <int>,
## #   ugds <int>, costt4_a <int>, selective <dbl>, research_u <dbl>
```