Problem Set 1

Intro to `R`

[YOUR NAME]

Due Date: 2024-01-19

Getting Set Up

Open `RStudio` and create a new RMarkDown file (`.Rmd`) by going to `File -> New File -> R Markdown...` . Accept defaults and save this file as `[LAST NAME]_ps1.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[LAST NAME]_ps1.Rmd` file. Then change the `author: [Your Name]` to your name.

We will be using the `sc_debt.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/sc_debt.Rds).

All of the following questions should be answered in this `.Rmd` file. There are code chunks with incomplete code that need to be filled in.

This problem set is worth 8 total points, plus two extra credit points. The point values for each question are indicated in brackets below. To receive full credit, you must have the correct code. In addition, some questions ask you to provide a written response in addition to the code.

You are free to rely on whatever resources you need to complete this problem set, including lecture notes, lecture presentations, Google, your classmates…you name it. However, the final submission must be complete by you. There are no group assignments. To submit, compiled the completed problem set and upload the PDF file to Brightspace on Friday by midnight. Also note that the TAs and professors will not respond to Campuswire posts after 5PM on Friday, so don't wait until the last minute to get started!

Good luck!

*Copy the link to ChatGPT you used here: _________________

Question 0 [0 points]

Require `tidyverse` and load the `sc_debt.Rds` data by assigning it to an object named `df` .

```r
require(tidyverse) # Load 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   1.2.1      ✓ 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()
```

```
df <- read_rds("https://github.com/jbisbee1/DS1000_S2024/raw/main/data/sc_debt.Rds") # Load the
dataset
```

Question 1 [2 points]

Which school has the lowest admission rate (`adm_rate`) and which state is it in (`stabbr`)?

```
df %>%
  arrange(adm_rate) %>% # Arrange by the admission rate
  select(instnm,adm_rate,stabbr) # Select the school name, the admission rate, and the state
```

```
## # A tibble: 2,546 × 3
##    instnm                                      adm_rate stabbr
##    <chr>                                          <dbl> <chr>
##  1 Saint Elizabeth College of Nursing            0      NY
##  2 Yeshivat Hechal Shemuel                       0      NY
##  3 Hampshire College                             0.0197 MA
##  4 Curtis Institute of Music                     0.0393 PA
##  5 Stanford University                           0.0434 CA
##  6 Harvard University                            0.0464 MA
##  7 Pacific Oaks College                          0.0511 CA
##  8 Columbia University in the City of New York   0.0545 NY
##  9 Princeton University                          0.0578 NJ
## 10 Yale University                               0.0608 CT
## # … with 2,536 more rows
```

> - There are two schools with the lowest admissions rate: St. Elizabeth College of Nursing and Yeshivat Hechal Shemuel. They are both in New York, and have an admissions rate of 0%. Assuming that this is an error in the data (since how can a school not admit any students?), the lowest non-zero admissions is for Hampshire College in Massachusetts.

Question 2 [2 points]

Which are the top 10 schools by average SAT score (`sat_avg`)?

```
df %>%
  arrange(desc(sat_avg)) %>% # arrange by SAT scores in descending order
  select(instnm,sat_avg) %>% # Select the school name and SAT score
  print(n = 12) # Print the first 12 rows (EC: there is a tie)
```

```
## # A tibble: 2,546 × 2
##    instnm                                 sat_avg
##    <chr>                                    <int>
##  1 California Institute of Technology        1557
##  2 Massachusetts Institute of Technology     1547
##  3 University of Chicago                     1528
##  4 Harvey Mudd College                       1526
##  5 Duke University                           1522
##  6 Franklin W Olin College of Engineering    1522
##  7 Washington University in St Louis         1520
##  8 Rice University                           1520
##  9 Yale University                           1517
## 10 Harvard University                        1517
## 11 Princeton University                      1517
## 12 Vanderbilt University                     1515
## # … with 2,534 more rows
```

> - The top 10 schools by average SAT score are CIT, MIT, U Chicago, Harvey Mudd, Duke, Franklin Olin, WUSTL, Rice, Yale, Harvard. There is a three-way tie for the school with the 10th highest average SAT score: Princeton, Harvard, and Yale all have an average score of 1517.

Question 3 [2 points]

Create a new variable called `adm_rate_pct` *which is the admissions rate multiplied by 100 to convert from a 0-to-1 decminal to a 0-to-100 percentage point.*

```
df <- df %>%
  mutate(adm_rate_pct = adm_rate*100)
```

Question 4 [2 points]

Calculate the average SAT score and median earnings of recent graduates by state.

```
df %>%
  group_by(stabbr) %>% # Calculate state-by-state with group_by()
  summarise(sat_avg = mean(sat_avg,na.rm=T), # Summarise the average SAT
            earn_avg = mean(md_earn_wne_p6,na.rm=T)) # Summarise the average earnings
```

```
## # A tibble: 51 × 3
##    stabbr sat_avg earn_avg
##    <chr>    <dbl>    <dbl>
##  1 AK       1121    33300
##  2 AL       1123.   28082.
##  3 AR       1141.   30452.
##  4 AZ       1147.   27613.
##  5 CA       1183.   33017.
##  6 CO       1132.   33955.
##  7 CT       1194.   35994.
##  8 DC       1262    41325
##  9 DE       1043    32443.
## 10 FL       1142.   30318.
## # … with 41 more rows
```

Extra Credit [2 points]

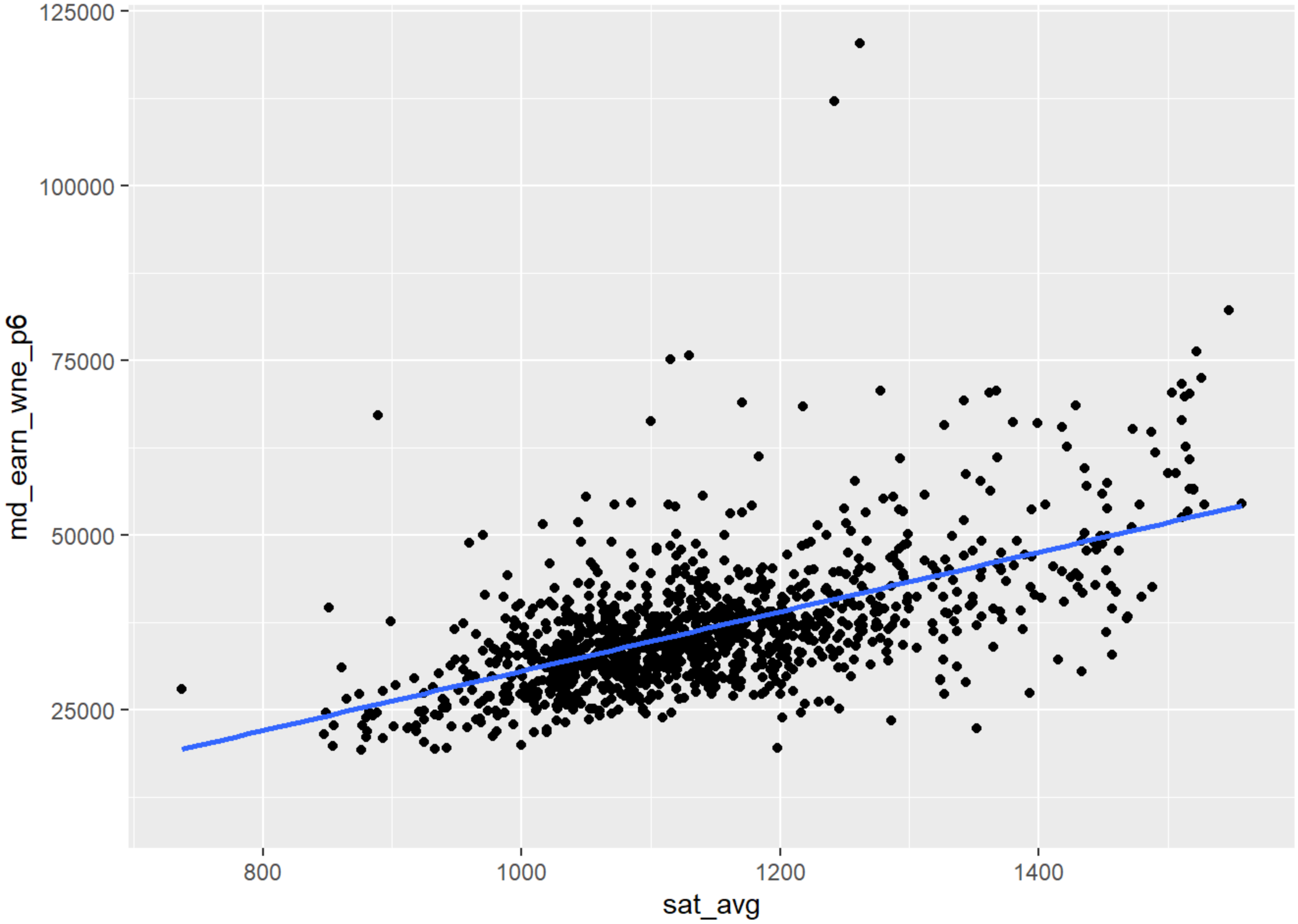
Plot the average SAT score (x-axis) against the median earnings of recent graduates (y-axis) by school, and add the line of best fit. What relationship do you observe? Why do you think this relationship exists?

```
df %>%
  ggplot(aes(x = sat_avg,y = md_earn_wne_p6)) +  # Build the plot
  geom_point() + # Add the points
  geom_smooth(method = 'lm',se = F) # Add a line of best fit
```

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

```
## Warning: Removed 1348 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1348 rows containing missing values (geom_point).
```



- I observe a positive relationship between SAT scores and earnings. I theorize that this relationship reflects the fact that SAT scores capture student abilities that are rewarded on the labor market. However, SAT scores are also correlated with many other socio-economic factors which might also improve one's earnings (i.e. social network) which are unrelated to student ability.