Problem Set 1

Intro to `R`

Alex Tomack

Due Date: 2023-01-27

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 `Problem_Set_1.Rmd` file into your `[LAST NAME]_ps1.Rmd` file. Then change the `author: [Your Name]` to your name.

If you haven't already, download the `sc_debt.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/2_Intro_to_R/data/sc_debt.Rds) and save it to your `data` folder.

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 10 total points, plus three extra credit points (one explicit and two hidden). 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 by 6PM CST on 2023/01/27 by midnight.

Good luck!

Question 1 [1 point]

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 3.4.0      ✔ purrr   1.0.0
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.5.0
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
df <- read_rds("../data/sc_debt.rds") # Load the dataset
```

Question 2 [1 point + 1 EC (hidden)]

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

```
df %>%
  arrange(adm_rate, na.rm=T) %>% # 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
```

> - Based on the table, Saint Elizabeth College of Nursing and Yeshivat Hechal Shemuel in New York have the lowest admissions rates. But, seeing as they are 0% acceptance rates, we might want to consider Hampshire College in Massachusetts instead.

Question 3 [1 point + 1 EC point (hidden)]

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
  head(df, n=10) # Print the first X rows
```

```
## # A tibble: 10 × 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
```

> - Caltech, MIT, UChicago, Harvey Mudd, Duke, Franklin Olin, WashU, Rice, Yale, Harvard are the top 10 schools with highest average SAT score.

Question 4 [1 point]

Which state is home to the school with the largest median earnings of recent graduates, and how much did they make?

```
df %>%
  arrange(desc(md_earn_wne_p6)) %>% # arrange by earnings in descending order
  select(instnm, md_earn_wne_p6, stabbr) # select the school name and earnings and the state
```

```
## # A tibble: 2,546 × 3
##    instnm                                                      md_earn…¹ stabbr
##    <chr>                                                            <int> <chr>
##  1 University of Health Sciences and Pharmacy in St. Louis        120400 MO
##  2 Albany College of Pharmacy and Health Sciences                 112100 NY
##  3 Samuel Merritt University                                      100100 CA
##  4 Massachusetts Institute of Technology                           82200 MA
##  5 Oregon Health & Science University                              80000 OR
##  6 Louisiana State University Health Sciences Center-Shreveport     78200 LA
##  7 Cochran School of Nursing                                       77300 NY
##  8 Duke University                                                 76300 NC
##  9 MCPHS University                                                75700 MA
## 10 Los Angeles County College of Nursing and Allied Health          75300 CA
## # … with 2,536 more rows, and abbreviated variable name ¹md_earn_wne_p6
```

> - The University of Health Science and Pharmacy in St Louis, Missouri, has the highest median salary of recent grads at $120,400.

Question 5 [1 point]

What is the average SAT score of the school with the highest median earnings identified in question 4?

```
df %>%
  filter(instnm == "University of Health Sciences and Pharmacy in St. Louis") %>% # Filter to th
e school identified above
  select(instnm, sat_avg) # select the school name and the SAT score
```

```
## # A tibble: 1 × 2
##   instnm                                                  sat_avg
##   <chr>                                                     <int>
## 1 University of Health Sciences and Pharmacy in St. Louis    1262
```

> - The University of Health Sciences and Pharmacy in St. Louis’ average SAT score is 1262.

Question 6 [1 point]

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(mean_sat = mean(sat_avg, na.rm=T), # Summarise the average SAT
            mean_wage=mean(md_earn_wne_p6, na.rm=T)) # Summarise the average earnings
```

```
## # A tibble: 51 × 3
##    stabbr mean_sat mean_wage
##    <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
```

Question 7 [2 points + 1 EC]

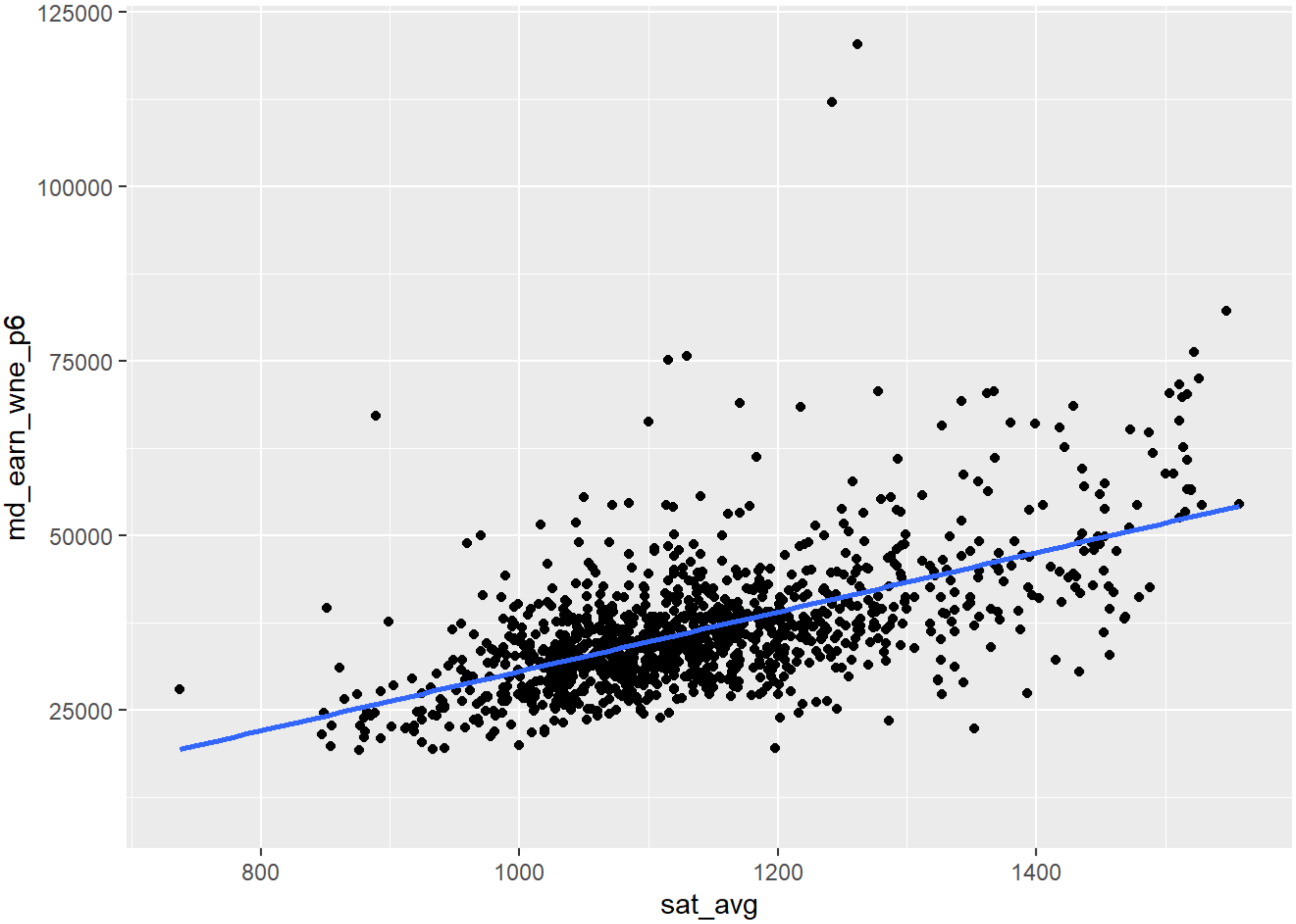
Plot the average SAT score (x-axis) against the median earnings of recent graduates (y-axis) by school. **EC: Plot the line of best fit**

```
df %>%
  ggplot(aes(x = sat_avg,y = md_earn_wne_p6)) +  # Build the plot with SAT scores on the x-axis
and earnings on the y-axis
  geom_point() +
  geom_smooth(method="lm", se=F)# Add the points
```

```
## `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()`).
```



Question 9 [2 points + 1 EC]

What relationship do you observe? Why do you think this relationship exists? EC: Is there any reason to be suspicious of this pattern?

> - There’s a moderately positive relationship between average SAT and median earnings of recent grads, indicating that graduates that scored higher on the SAT earn more. This isn’t a great conclusion, though, as it fails to take into account socioeconomic status or field of study, for instance.