Problem Set 1

Intro to R

[YOUR NAME]

Due Date: 2024-08-30

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_F2024/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 5 total points, plus 1 extra credit point. 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. Instructions for how to compile the output as a PDF can be found in Problem Set 0 (https://github.com/jbisbee1/DS1000_F2024/blob/main/Psets/ds1000_pset_0.pdf) and in this gif tutorial (https://github.com/jbisbee1/DS1000_F2024/blob/main/Psets/save_as_pdf.gif).

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` .

```
require(tidyverse) # Load tidyverse
```

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

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
df <- read_rds("https://github.com/jbisbee1/DS1000_F2024/raw/main/data/sc_debt.Rds") # Load the
dataset
```

The codebook is reproduced here.

Name	Definition
unitid	Unit ID
instnm	Institution Name
stabbr	State Abbreviation
grad_debt_mdn	Median Debt of Graduates
control	Control Public or Private
region	Census Region
preddeg	Predominant Degree Offered: Associates or Bachelors
openadmp	Open Admissions Policy: 1= Yes, 2=No,3=No 1st time students
adm_rate	Admissions Rate: proportion of applications accepted
ccbasic	Type of institution– see here (https://data.ed.gov/dataset/9dc70e6b-8426-4d71-b9d5-70ce6094a3f4/resource/658b5b83-ac9f-4e41-913e-9ba9411d7967/download/collegescorecarddatadictionary_01192021.xlsx)
selective	Institution admits fewer than 10 % of applicants, 1=Yes, 0=No
research_u	Institution is a research university 1=Yes, 0=No
sat_avg	Average Sat Scores
md_earn_wne_p6	Average Earnings of Recent Graduates
ugds	Number of undergraduates
costt4a	Average cost of attendance (tuition-grants)

Question 1 [1 point]

Which school has the highest future earnings and what is its admissions rate?

```
df %>%
  arrange(desc(md_earn_wne_p6)) %>% # Arrange by the future earnings
  select(instnm,md_earn_wne_p6,adm_rate) # Select the school name, the admission rate, and the s
tate
```

```
## # A tibble: 2,546 × 3
##    instnm                                                md_earn_wne_p6 adm_rate
##    <chr>                                                          <int>    <dbl>
##  1 University of Health Sciences and Pharmacy in St. Lo…         120400    0.917
##  2 Albany College of Pharmacy and Health Sciences                112100    0.711
##  3 Samuel Merritt University                                     100100   NA
##  4 Massachusetts Institute of Technology                          82200    0.067
##  5 Oregon Health & Science University                             80000   NA
##  6 Louisiana State University Health Sciences Center-Sh…          78200   NA
##  7 Cochran School of Nursing                                      77300   NA
##  8 Duke University                                                76300    0.076
##  9 MCPHS University                                               75700    0.853
## 10 Los Angeles County College of Nursing and Allied Hea…          75300   NA
## # ℹ 2,536 more rows
```

- The University of Health Science and Pharmacy in St. Louis has the highest future earnings and an admissions rate of almost 92%. [Rubric: 0 points if no attempt. 0.5 points if correct written answer but code produces error. 0.5 points if correct code but no / wrong answer.]

Question 2 [1 point]

How many schools are located in Massachusetts

```
df %>%
  filter(stabbr == 'MA')
```

```
## # A tibble: 91 × 16
##    unitid instnm   stabbr grad_debt_mdn control region preddeg openadmp adm_rate
##     <int> <chr>    <chr>          <int> <chr>   <chr>  <chr>      <int>    <dbl>
##  1 164368 Hult In… MA                NA Private New E… Bachel…        2    0.434
##  2 164447 America… MA             27000 Private New E… Bachel…        2    0.646
##  3 164465 Amherst… MA             13500 Private New E… Bachel…        2    0.113
##  4 164492 Anna Ma… MA             25000 Private New E… Bachel…        2    0.740
##  5 164562 Assumpt… MA             27000 Private New E… Bachel…        2    0.811
##  6 164580 Babson … MA             22985 Private New E… Bachel…        2    0.264
##  7 164614 Boston … MA             23350 Private New E… Bachel…        2    0.909
##  8 164632 Bay Pat… MA             24125 Private New E… Bachel…        2    0.718
##  9 164720 Becker … MA             26000 Private New E… Bachel…        2    0.699
## 10 164739 Bentley… MA             25000 Private New E… Bachel…        2    0.467
## # ℹ 81 more rows
## # ℹ 7 more variables: ccbasic <int>, sat_avg <int>, md_earn_wne_p6 <int>,
## #   ugds <int>, costt4_a <int>, selective <dbl>, research_u <dbl>
```

- There are 91 schools in Massachusetts. [Rubric: 0 points if no attempt. 0.5 points if correct written answer but code produces error. 0.5 points if correct code but no / wrong answer.]

Question 3 [1 point]

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. Then calculate the overall average admissions rate in the data and write this answer.

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

df %>%
  summarise(adm_rate_pct_avg = mean(adm_rate_pct,na.rm=T))
```

```
## # A tibble: 1 × 1
##   adm_rate_pct_avg
##              <dbl>
## 1             67.9
```

- The overall average admissions rate across all schools is 67.9%. [Rubric: 0 points if no attempt. 0.5 points if correct written answer but code produces error. 0.5 points if correct code but no / wrong answer. 0.75 points if code is correct but students forget `na.rm=T`, yielding an `NA`.]

Question 4 [1 point]

Calculate the average SAT score and median earnings of recent graduates by region. Which region has the highest average SAT scores and which has the highest future earnings?

```
df %>%
  group_by(region) %>% # Calculate region-by-region 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
  arrange(desc(sat_avg))
```

```
## # A tibble: 8 × 3
##   region          sat_avg earn_avg
##   <chr>             <dbl>    <dbl>
## 1 New England       1214.   37166.
## 2 Far West          1172.   32863.
## 3 Northeast         1163.   35636.
## 4 Great Lakes       1139.   33481.
## 5 Plains            1138.   33675
## 6 Rocky Mountains   1137.   30072
## 7 Southeast         1112.   29923.
## 8 Southwest         1107.   31934.
```

New England has both the highest average SAT scores and the highest future earnings. [Rubric: 0 points if no attempt. 0.5 points if correct written answer but code produces error. 0.5 points if correct code but no / wrong answer. 0.75 points if code is correct but students forget `na.rm=T`, yielding an `NA`.]

Question 5 [1 points]

Research Question: Do students who graduate from smaller schools (i.e., schools with smaller student bodies) make more money in their future careers? Make sure to give an answer and then **explain why** *you think so.*

Yes students from smaller schools will make more money. This is because smaller schools tend to have smaller classes which means that professors can work with students directly, helping them learn faster and better.

Extra Credit [1 point]

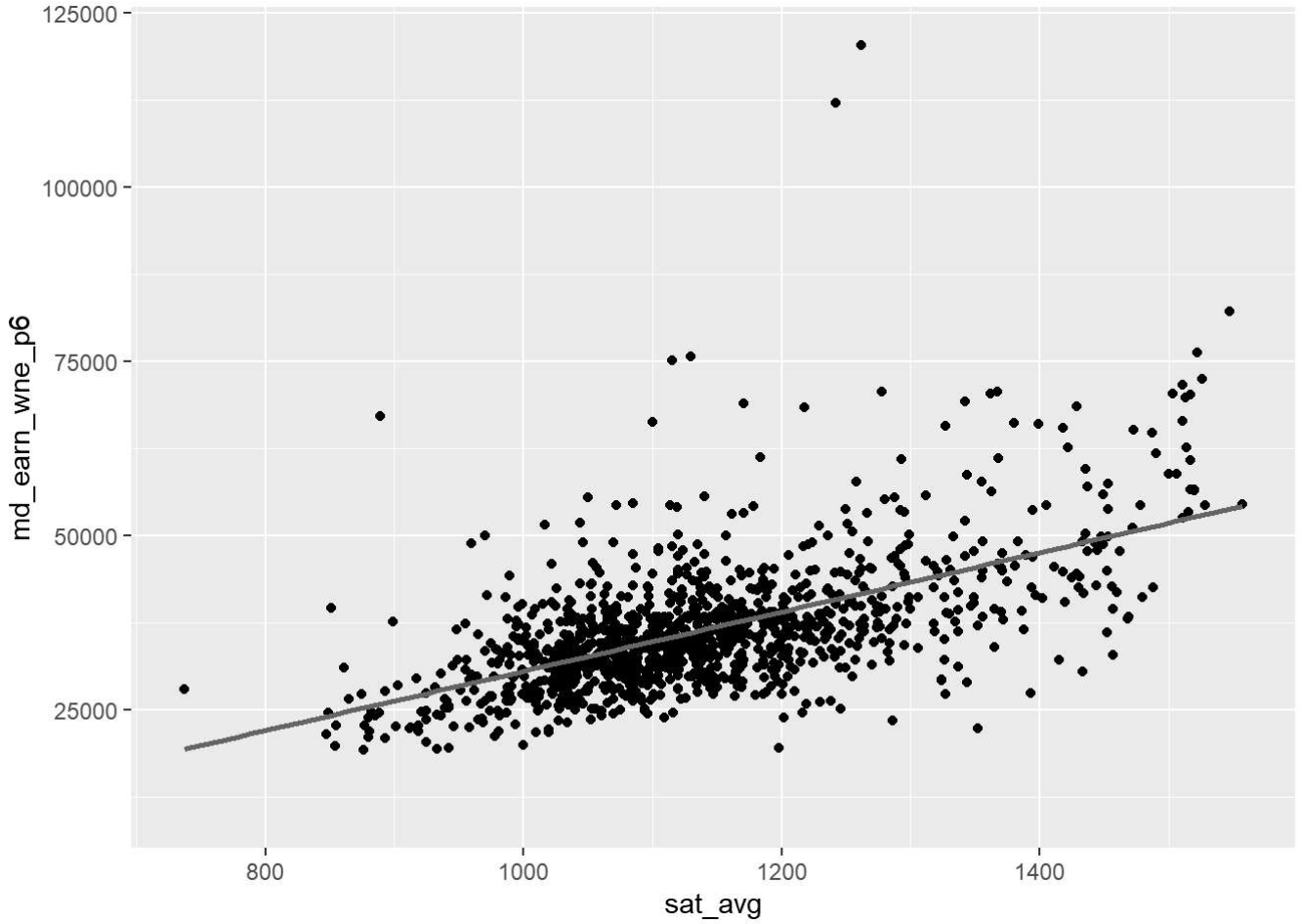
Plot the average SAT score (x-axis) against the median earnings of recent graduates (y-axis) by school, and add the line of best fit. What relationship do you observe? Why do you think this relationship exists?

```
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 outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1348 rows containing missing values or values outside the scale range
## (`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. [Rubric: 0 points if no attempt. 0.5 points if correct written answer but code produces error. 0.5 points if correct code but no / wrong answer. Written answers can vary, but should be logical.]