Problem Set 1

Intro to `R`

Jim Bisbee

Due Date: 2024-07-04

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/ISP_Data_Science_2024/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 10 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, email the knitted output to Eun Ji Kim (kej990804@snu.ac.kr (mailto:kej990804@snu.ac.kr)) **as a PDF** by the start of class on Thursday, July 4th. If you need help converting to a PDF, see this tutorial (https://github.com/jbisbee1/ISP_Data_Science_2024/blob/main/Psets/ISP_pset_0_HELPER.pdf).

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 t
o become errors
```

```
df <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/sc_debt.
Rds") # Load the dataset directly from github
```

Question 1 [1 point]

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 s
tate
```

```
## # 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
## # ℹ 2,536 more rows
```

> There are two schools: Saint Elizabeth College of Nursing and Yeshivat Hechal Shemuel. Both have admissions rates of 0, meaning they don’t accept anyone and both are located in NY.

Question 2 [1 point]

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

```
df %>%
  arrange(-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 (hint: 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
## # ℹ 2,534 more rows
```

> Here are the top 10 schools by average SAT score.

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 decimal to a 0-to-100 percentage point.*

```
df <- df %>% # Use the object assignment operator to overwrite the df object
  mutate(adm_rate_pct = adm_rate*100) # Create the new variable adm_rate_pct
```

Question 4 [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(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.
## # ℹ 41 more rows
```

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? Before looking at the data, write out what you think the answer is, and explain why you think so.

> Write a few sentences here.

Question 6 [2 points]

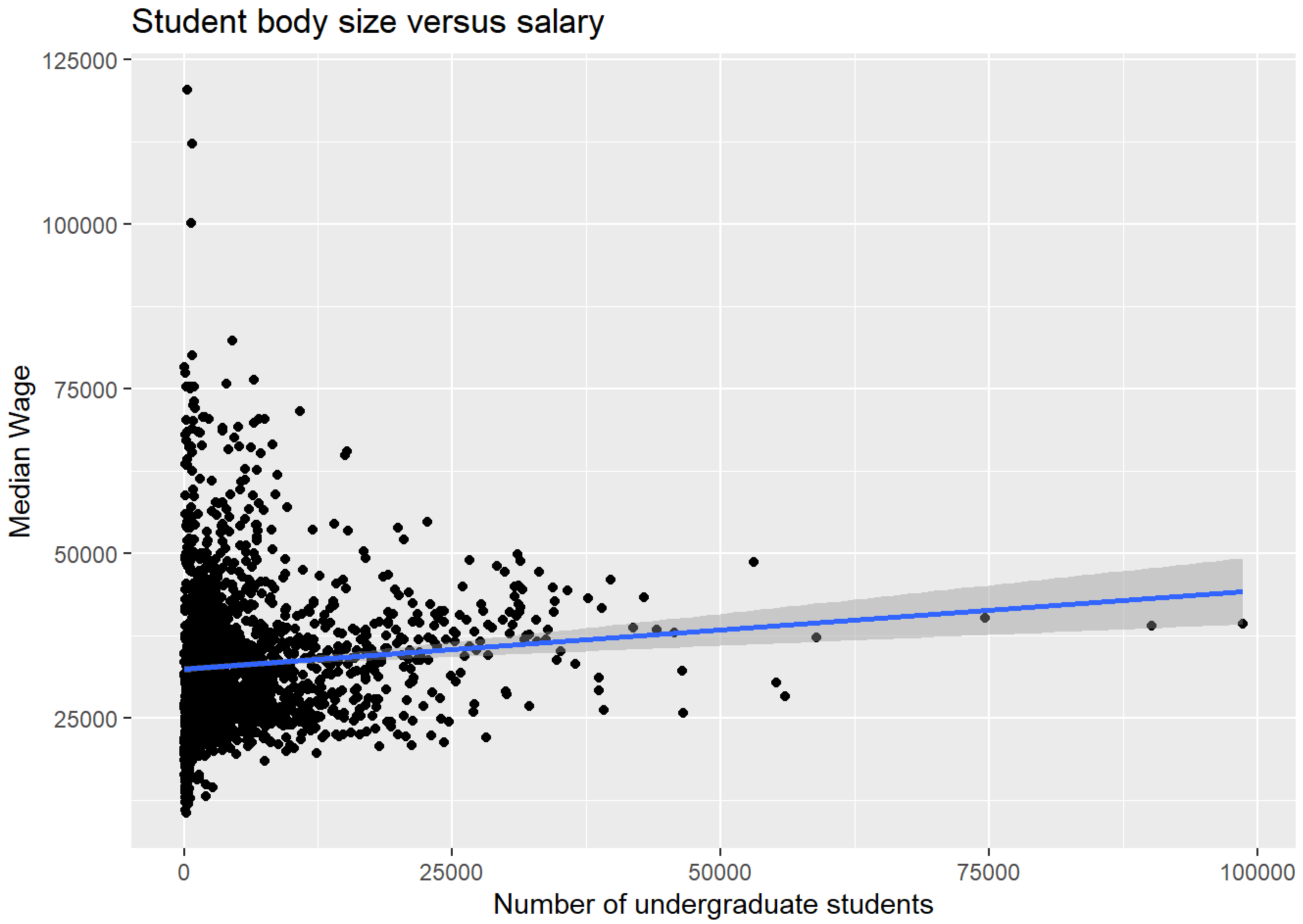
Based on this research question, what is the outcome / dependent / Y variable and what is the explanatory / independent / X variable? Create the scatterplot of the data based on this answer, along with a line of best fit. Is your answer to the research question supported?

```
View(df)

df %>%
  ggplot(aes(x = ugds, # Put the explanatory variable on the x-axis
             y = md_earn_wne_p6)) +  # Put the outcome variable on the y-axis
  geom_point() + # Create a scatterplot
  geom_smooth(method = 'lm') + # Add line of best fit
  labs(title = 'Student body size versus salary', # give the plot meaningful labels to h
elp the viewer understand it
       x = 'Number of undergraduate students',
       y = 'Median Wage')
```

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

```
## Warning: Removed 241 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Warning: Removed 241 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



> Write a few sentences here.

Question 7 [2 points]

Does this relationship change by whether the school is a research university? Using the filter() function, create two versions of the plot, one for research universities and the other for non-research universities.

```
df %>%
  filter(research_u == 0) %>% # Filter to non-research universities
  ggplot(aes(x = , # Put the explanatory variable on the x-axis
             y = )) +  # Put the outcome variable on the y-axis
  geom_point() + # Create a scatterplot
  geom_smooth() + # Add line of best fit
  labs(title = '', # give the plot meaningful labels to help the viewer understand it
       subtitle = '',
       x = '',
       y = '')
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## Error in `geom_smooth()`:
## ! Problem while computing stat.
## ℹ Error occurred in the 2nd layer.
## Caused by error in `compute_layer()`:
## ! `stat_smooth()` requires the following missing aesthetics: x and y.
```

```
df %>%
  filter(research_u == 1) %>% # Filter to research universities
  ggplot(aes(x = , # Put the explanatory variable on the x-axis
             y = )) +  # Put the outcome variable on the y-axis
  geom_point() + # Create a scatterplot
  geom_smooth() + # Add line of best fit
  labs(title = '', # give the plot meaningful labels to help the viewer understand it
       subtitle = '',
       x = '',
       y = '')
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
## Error in `geom_smooth()`:
## ! Problem while computing stat.
## ℹ Error occurred in the 2nd layer.
## Caused by error in `compute_layer()`:
## ! `stat_smooth()` requires the following missing aesthetics: x and y.
```

Question 8 [1 point]

Instead of creating two separate plots, color the points by whether the school is a research university. To do this, you first need to modify the research_u variable to be categorical (it is currently stored as numeric). To do this, use the mutate command with `ifelse()` to create a new variable called `research_u_cat` which is either “Research” if `research_u` is equal to 1, and “Non-Research” otherwise.

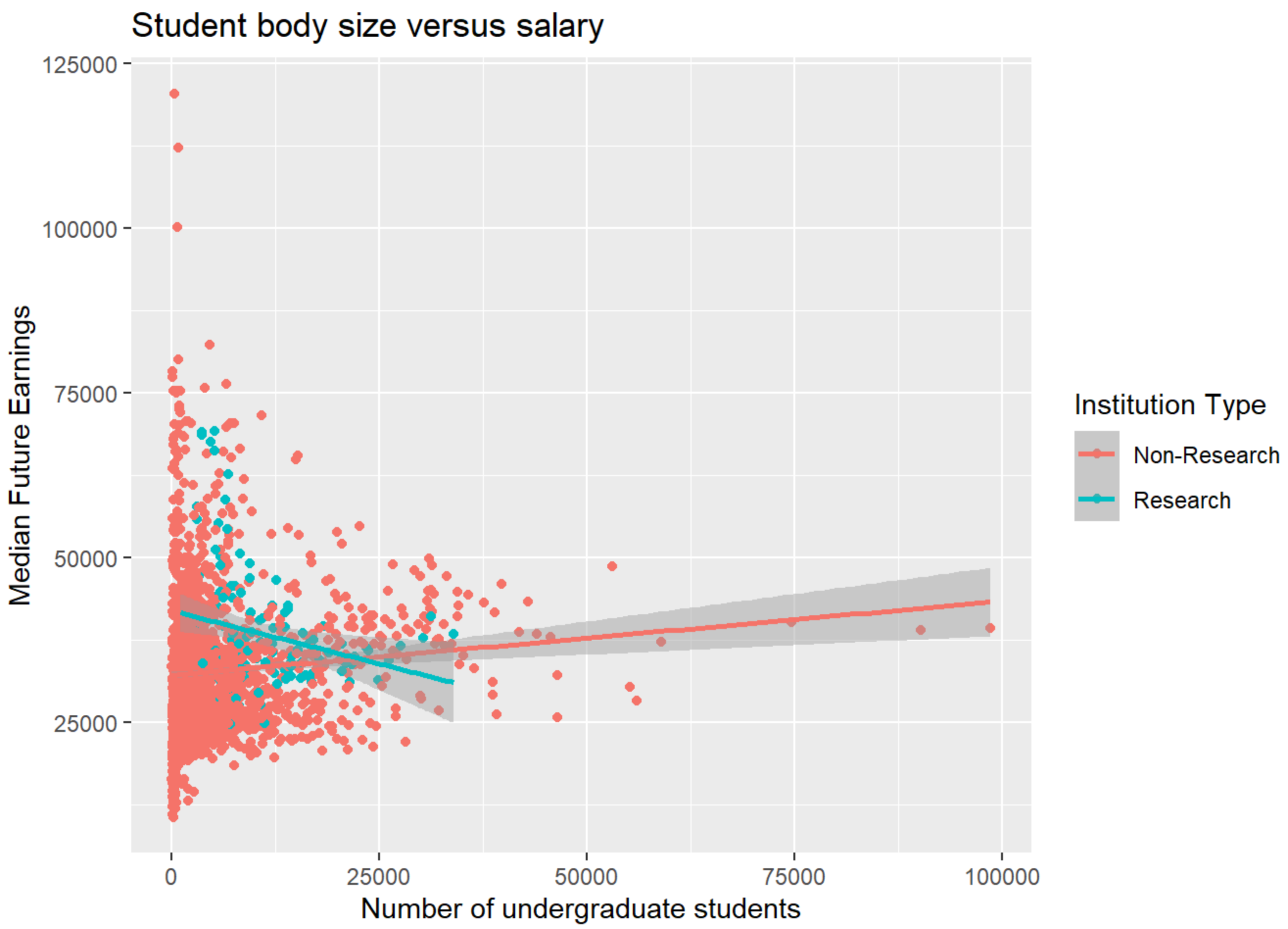
```
df <- df %>%
  mutate(research_u_cat = ifelse(research_u == 0,"Non-Research",
                                 "Research")) # Create a labeled version of the research
_u variable

df %>%
  ggplot(aes(x = ugds, # Put the explanatory variable on the x-axis
             y = md_earn_wne_p6, # Put the outcome variable on the y-axis
             color = research_u_cat)) + # Color the points by the new variable you creat
ed above
  geom_point() + # Create a scatterplot
  geom_smooth(method = 'lm') + # Add line of best fit
  labs(title = 'Student body size versus salary', # give the plot meaningful labels to h
elp the viewer understand it
       x = 'Number of undergraduate students',
       color = 'Institution Type',
       y = 'Median Future Earnings')
```

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

```
## Warning: Removed 241 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 241 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Extra Credit [2 points]

Write a short paragraph discussing your findings. What do you think is going on in these data?

> Write a few sentences here