

# Problem Set 2

## Visualization in R

[YOUR NAME]

Due Date: 2024-01-26

## 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]\_ps2.Rmd to your code folder.

Copy and paste the contents of this .Rmd file into your [LAST NAME]\_ps2.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](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, compile 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.

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

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

```
df <- read_rds("") # Load the dataset
```

```
## Error in read_rds(""): could not find function "read_rds"
```

## Question 1 [2 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 2 [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?*

```
df %>%
  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
       x = '',
       y = '')
```

```
## Error in df %>% ggplot(aes(x = , y = )): could not find function "%>%"
```

Write a few sentences here.

## Question 3 [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() %>% # 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 = '')
```

```
## Error in df %>% filter() %>% ggplot(aes(x = , y = )): could not find function "%>%"
```

```
df %>%
  filter() %>% # 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 = '')
```

```
## Error in df %>% filter() %>% ggplot(aes(x = , y = )): could not find function "%>%"
```

## Question 4 [2 points]

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()) # Create a labeled version of the research_u variable
```

```
## Error in df %>% mutate(research_u_cat = ifelse()): could not find function "%>%"
```

```
df %>%
  ggplot(aes(x = , # Put the explanatory variable on the x-axis
             y = , # Put the outcome variable on the y-axis
             color = )) + # Color the points by the new variable you created above
  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
        x = '',
        color = '',
        y = '')
```

```
## Error in df %>% ggplot(aes(x = , y = , color = )): could not find function "%>%"
```

## 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