Problem Set 2

Intro to `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).

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` .

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

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

```
## Warning: package 'tidyverse' was built under R version 4.2.3
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
## Warning: package 'tibble' was built under R version 4.2.3
```

```
## Warning: package 'tidyr' was built under R version 4.2.3
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
## Warning: package 'purrr' was built under R version 4.2.3
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
## Warning: package 'forcats' was built under R version 4.2.3
```

```
## Warning: package 'lubridate' was built under R version 4.2.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ 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 become errors
```

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

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.

> - 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.

Question 2 [2 points]

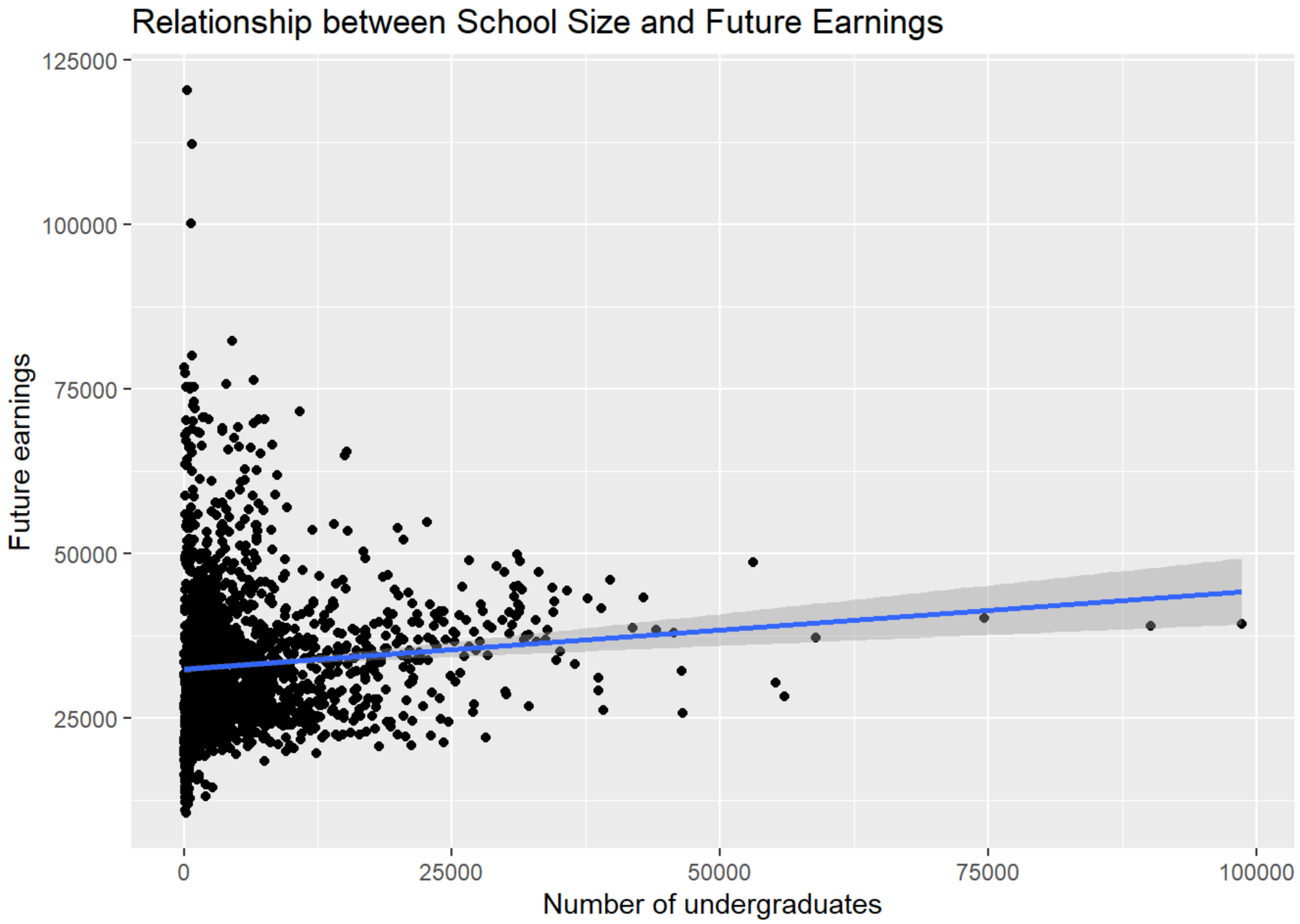
Based on this research question, what is the outcome / dependent / Y variable and what is the explanatory / independent / X variable? Create the scatterplot of the data based on this answer, along with a line of best fit. Is your answer to the research question supported?

```
df %>%
  ggplot(aes(x = 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 = 'Relationship between School Size and Future Earnings', # give the plot meaningfu
l labels to help the viewer understand it
       x = 'Number of undergraduates',
       y = 'Future earnings')
```

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

```
## Warning: Removed 241 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 241 rows containing missing values (`geom_point()`).
```



> - The outcome variable is median future earnings (`md_earn_wne_p6`) and the explanatory variable is `ugds` . There appears to be a very small positive association between school size and future earnings, which is against my hypothesis.

Question 3 [2 points]

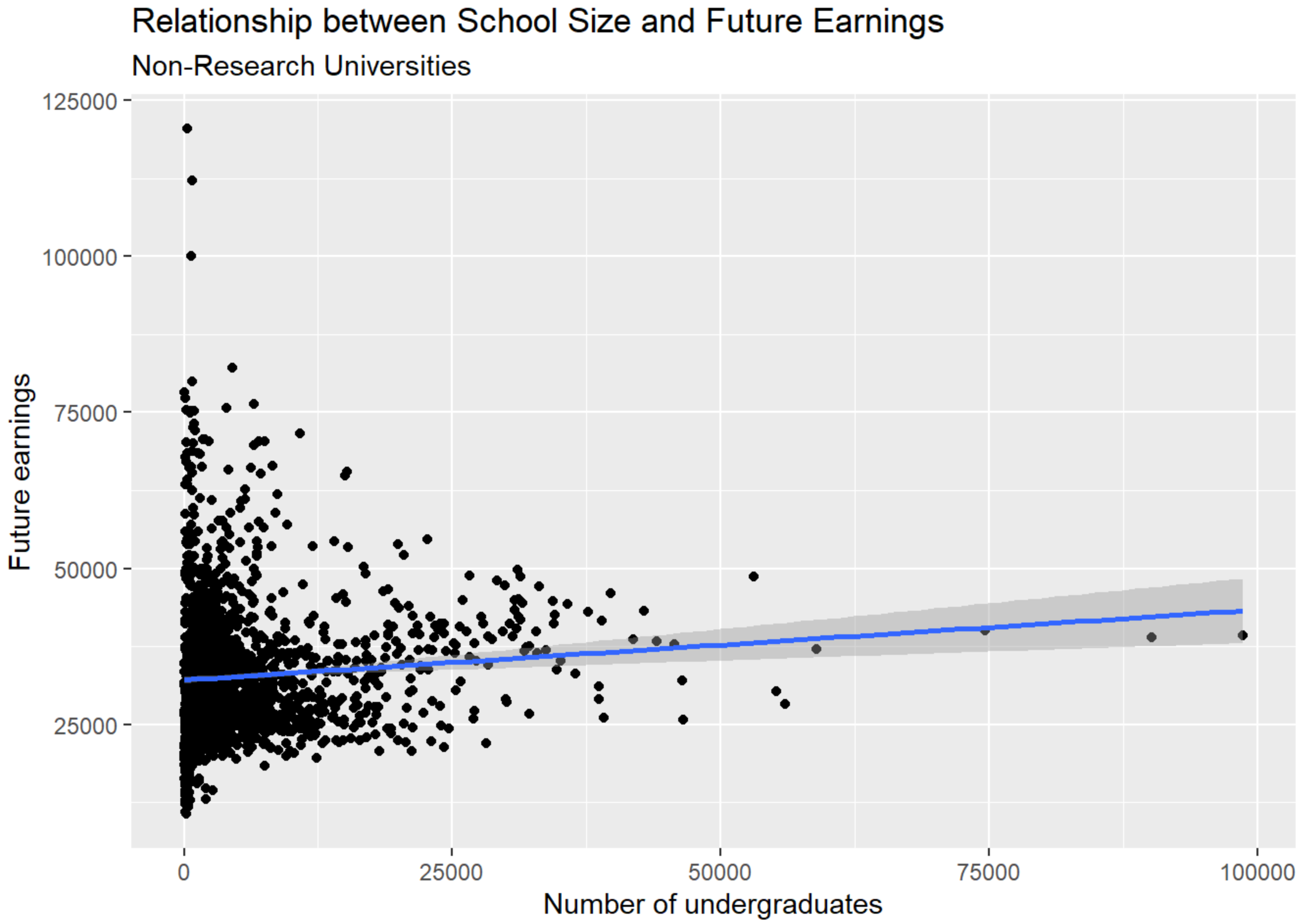
Does this relationship change by whether the school is a research university? Using the filter() function, create two versions of the plot, one for research universities and the other for non-research universities.

```
df %>%
  filter(research_u == 0) %>% # Filter to non-research universities
  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 = 'Relationship between School Size and Future Earnings', # give the plot meaningful labels to help the viewer understand it
       subtitle = 'Non-Research Universities',
       x = 'Number of undergraduates',
       y = 'Future earnings')
```

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

```
## Warning: Removed 240 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 240 rows containing missing values (`geom_point()`).
```

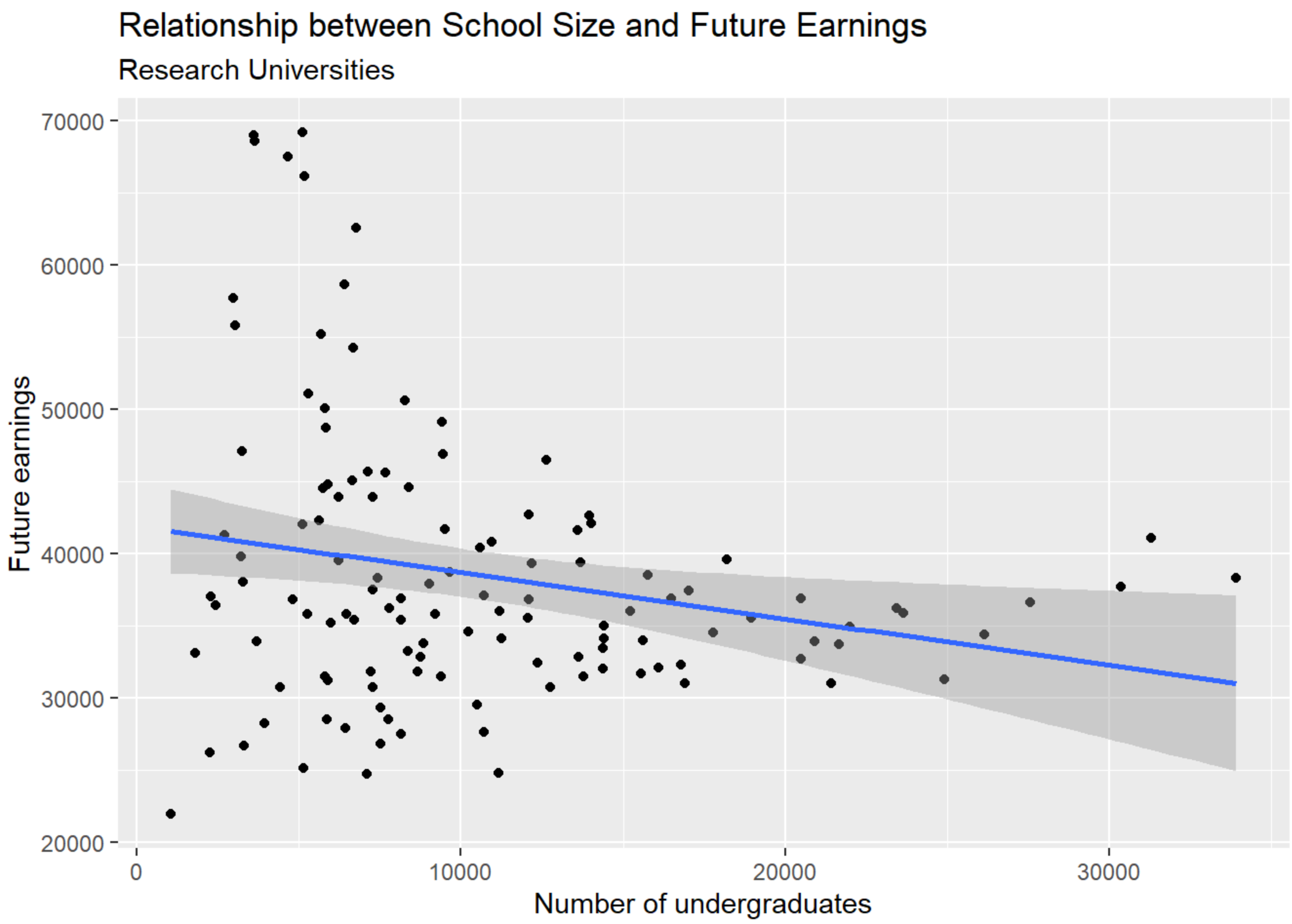


```
df %>%
  filter(research_u == 1) %>% # Filter to research universities
  ggplot(aes(x = ugds, # Put the explanatory variable on the x-axis
             y = md_earn_wne_p6)) +  # Put the outcome variable on the y-axis
  geom_point() + # Create a scatterplot
  geom_smooth(method = 'lm') + # Add line of best fit
  labs(title = 'Relationship between School Size and Future Earnings', # give the plot meaningfu
l labels to help the viewer understand it
       subtitle = 'Research Universities',
       x = 'Number of undergraduates',
       y = 'Future earnings')
```

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

```
## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Question 4 [2 points]

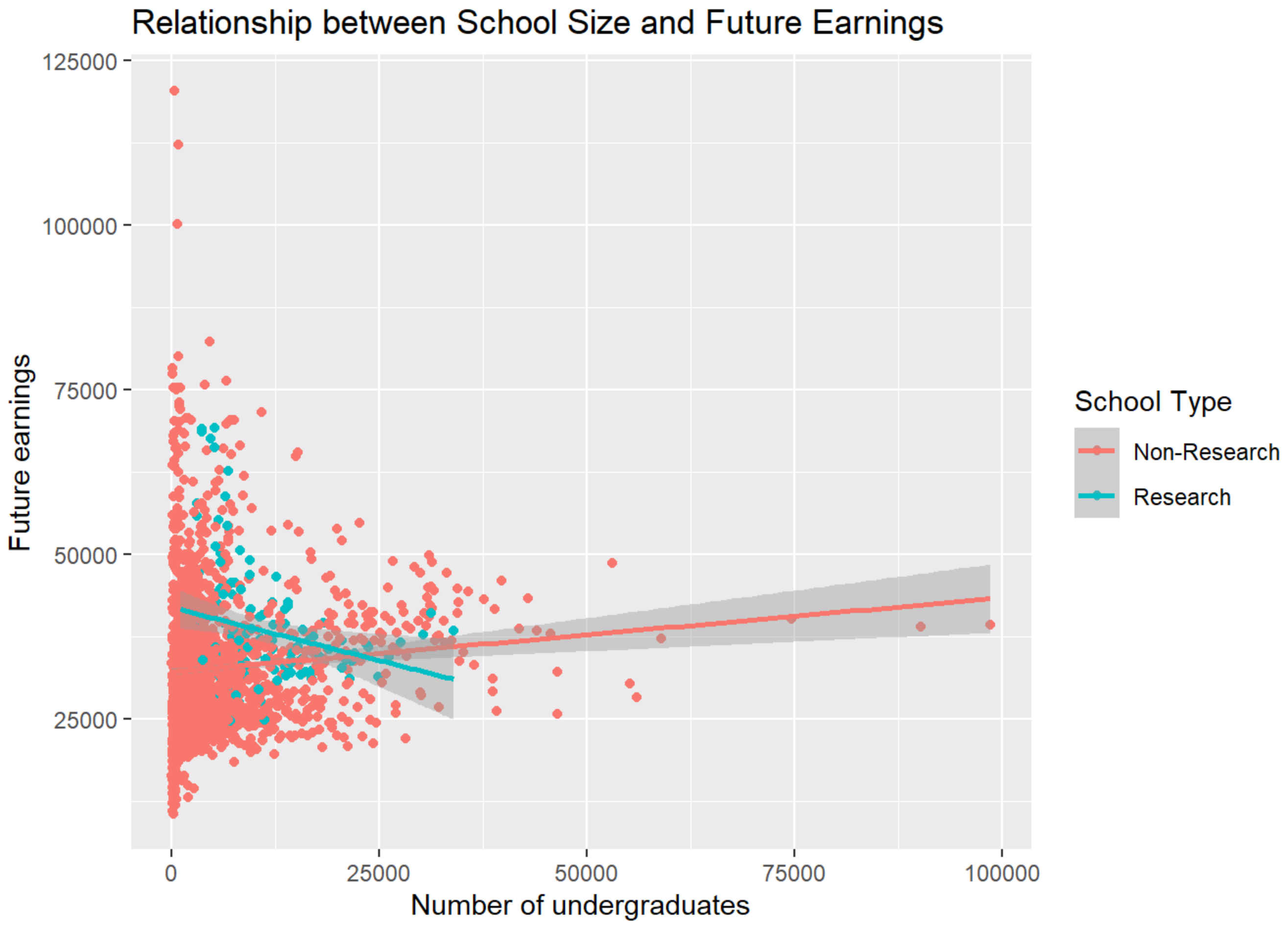
Instead of creating two separate plots, color the points by whether the school is a research university. To do this, you first need to modify the research_u variable to be categorical (it is currently stored as numeric). To do this, use the mutate command with `ifelse()` to create a new variable called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

df %>%
  ggplot(aes(x = ugds, # Put the explanatory variable on the x-axis
             y = md_earn_wne_p6,
             color = research_u_cat)) +  # Put the outcome variable on the y-axis
  geom_point() + # Create a scatterplot
  geom_smooth(method = 'lm') + # Add line of best fit
  labs(title = 'Relationship between School Size and Future Earnings', # give the plot meaningfu
l labels to help the viewer understand it
       x = 'Number of undergraduates',
       color = 'School Type',
       y = 'Future earnings')
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 241 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 241 rows containing missing values (`geom_point()`).
```



Extra Credit [2 points]

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

> - It seems that school size works in opposite directions between research and non-research universities. In research universities, graduates from smaller schools make more money, whereas graduates from larger non-research universities make more money. This might reflect the trade-off between learning valuable skills and social networks. At non-research universities, the value of education is more about building a professional network, meaning that larger schools produce graduates with larger social networks, who go on to make more money. At research universities, the value of the degree is more about the skills themselves, meaning that smaller schools provide better teaching in a more focused way, producing graduates with better skills who go on to make more money.