Problem Set 2

Mini Data Science

[YOUR NAME]

Due Date: 2024-09-06

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_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`.

```r
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_S2024/raw/main/data/sc_debt.Rds") # Load the
dataset
```

Question 1 [1 point]

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. [Rubric: 0 points if no attempt. 0.5 points if the answer is unexplained. 0.75 points if the answer is poorly explained. (Assumptions should be clear.)]

Question 2 [1 point]

Based on this research question, what is the outcome / dependent / Y variable and what is the explanatory / independent / X variable? What are their average values in the data?

```
df %>%
  summarise(avg_X = mean(ugds,na.rm=T),
            avg_earn = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 1 × 2
##   avg_X avg_earn
##   <dbl>    <dbl>
## 1 4861.   33028.
```

> - The outcome variable is median future earnings (`md_earn_wne_p6`), with an overall average value of $33,028, and the explanatory variable is `ugds` , with an overall average value of 4,861. [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]

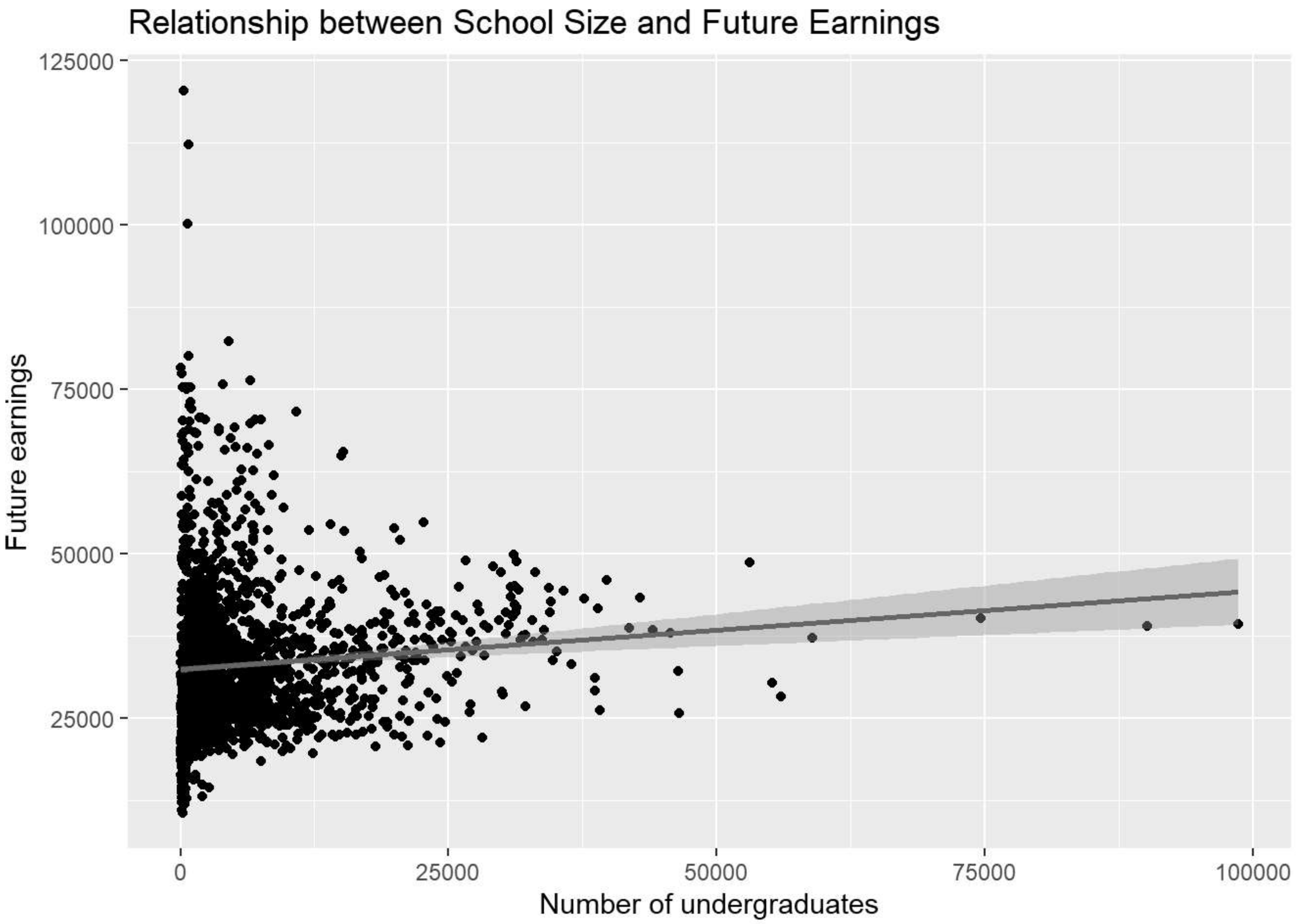
Create the scatterplot of the data that analyzes your hypothesis, along with a line of best fit. Then, describe the result. Is your answer to the research question supported?

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

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



- There appears to be a very small positive association between school size and future earnings, which is against my hypothesis. [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 and first part of written response are correct, but no reference is made to the student’s hypothesis. 0.75 points if figure isn’t labeled.]

Question 4 [1 point]

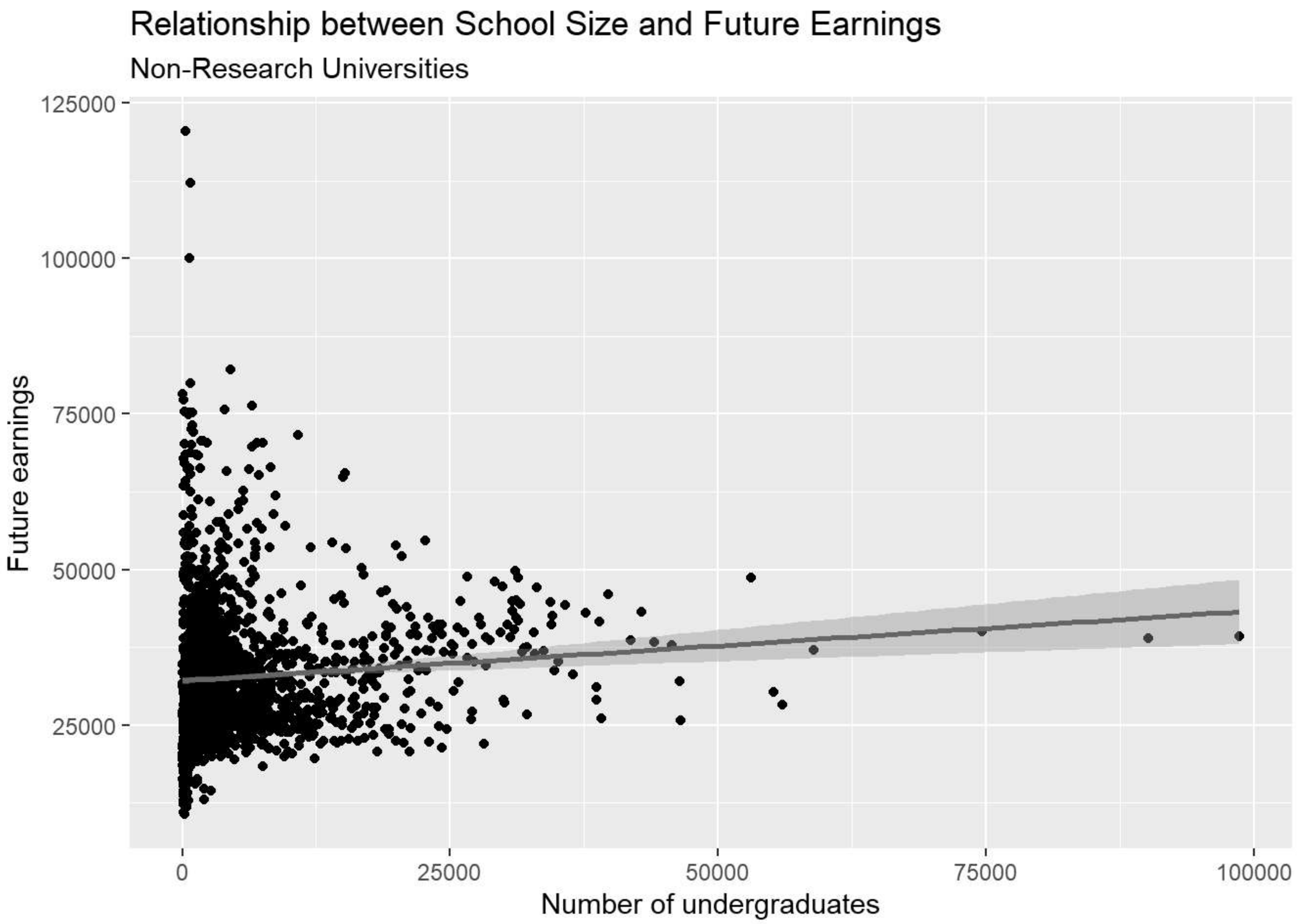
Does this relationship change by whether the school is a research university? Using the filter() function, create two versions of the plot, one for research universities and the other for non-research universities. What do you find?

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

```
## Warning: Removed 240 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

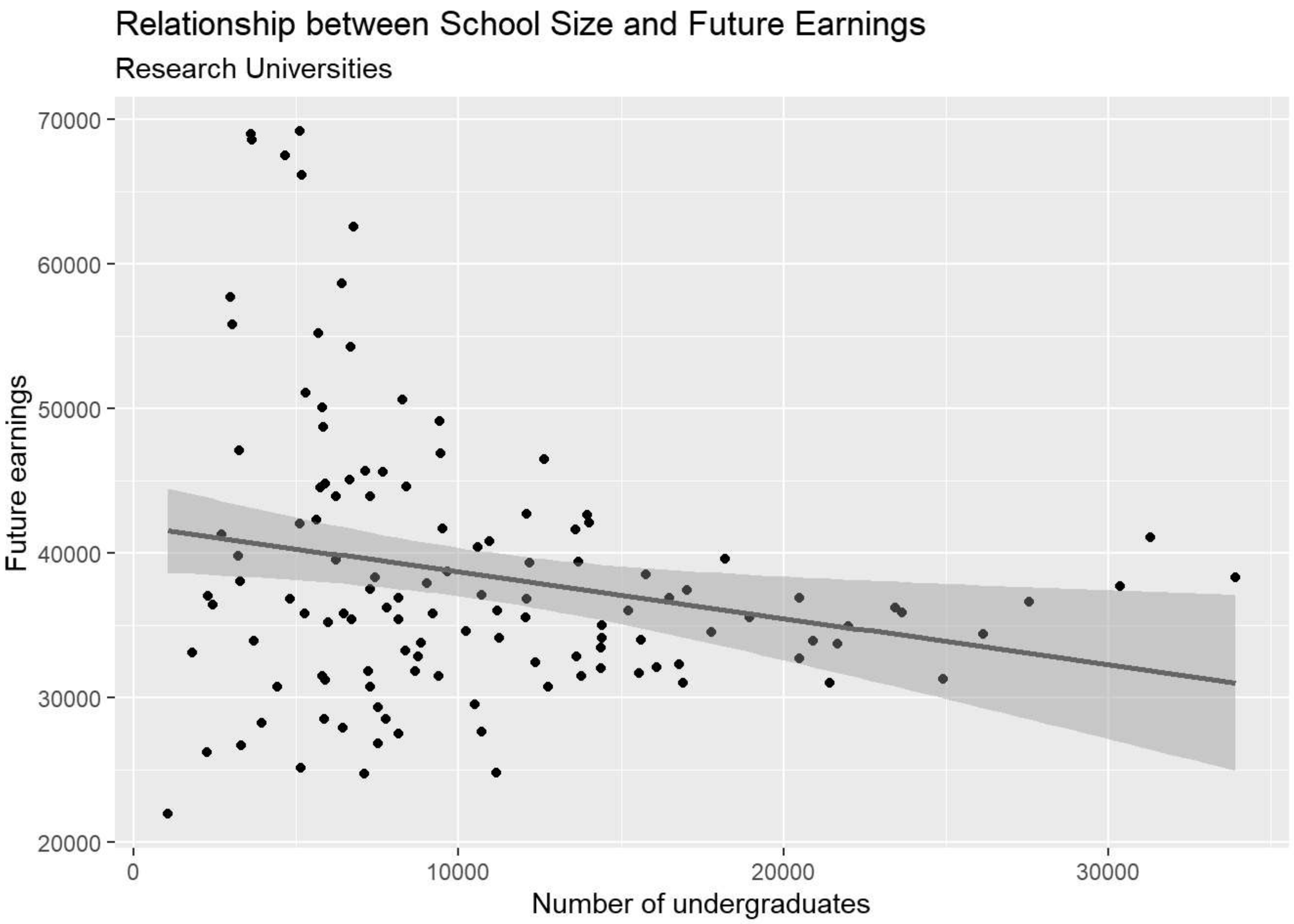


```
df %>%
  filter(research_u == 1) %>% # Filter to research universities
  ggplot(aes(x = ugds, # Put the explanatory variable on the x-axis
             y = md_earn_wne_p6)) +  # Put the outcome variable on the y-axis
  geom_point() + # Create a scatterplot
  geom_smooth(method = 'lm') + # Add line of best fit
  labs(title = 'Relationship between School Size and Future Earnings', # give the plot meaningful 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 row containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



> - The relationship does change depending on whether the school is a research university or not. In research universities, there is a negative relationship between earnings and the number of undergraduates. In non-research universities, this relationship is positive. [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 figures aren't labeled.]

Question 5 [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.

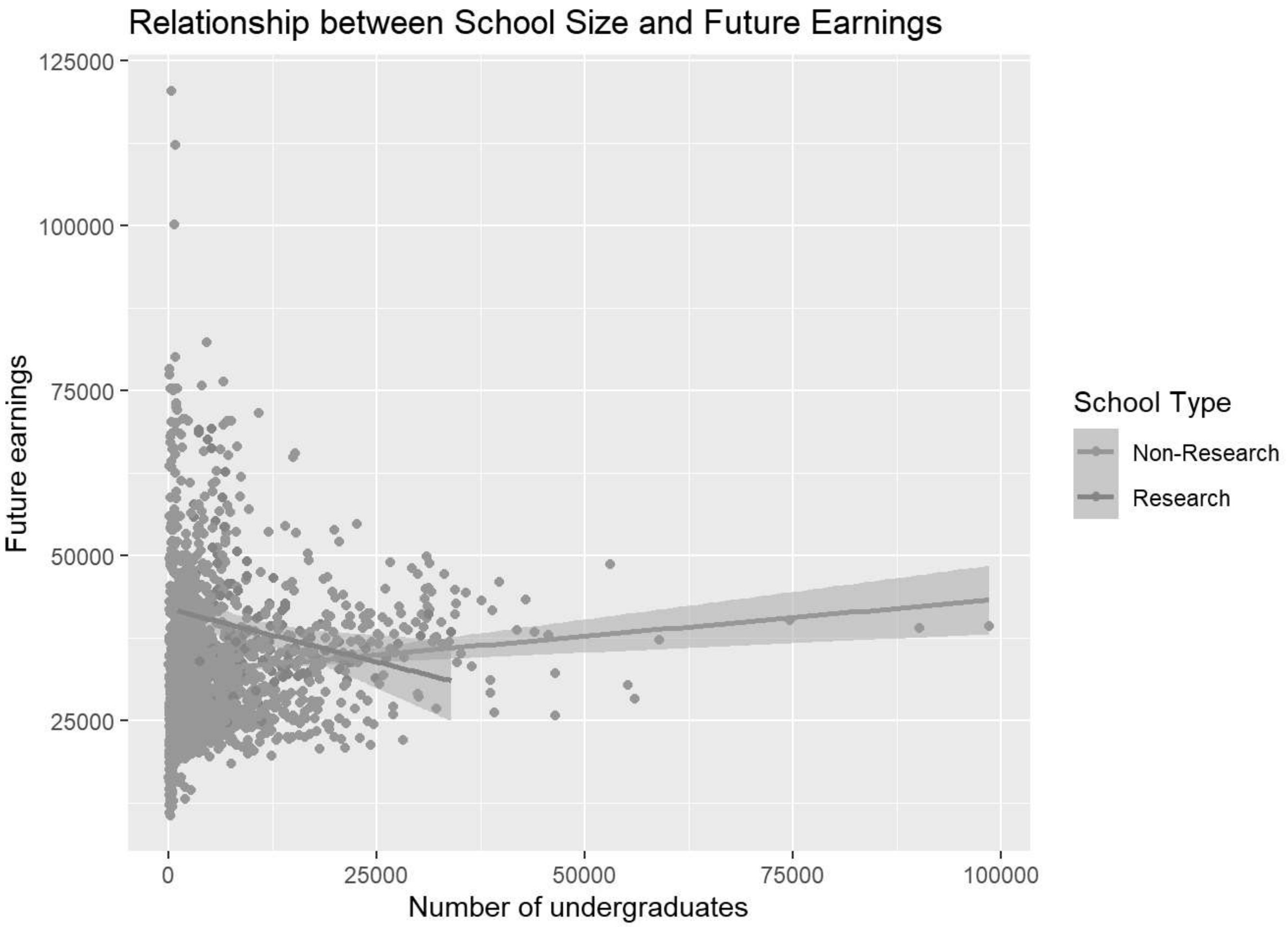
```
df <- df %>%
  mutate(research_u_cat = ifelse(research_u == 1,'Research','Non-Research'))

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

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



> [Rubric: 0 points if no attempt. 0.5 points if the `ifelse()` part doesn’t work or isn’t implemented. 0.75 points if figures aren’t labeled.]

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. [Rubric: 0 points if no attempt. 0.5 points answer isn’t clear or logical.]