Charlie Gartenberg Cleaning data

## Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see <https://quarto.org>.

## Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

# Load libraries  
library(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 3.4.4 ✔ 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 become errors

library(rio)  
# Set the root directory for not ebook chunks  
setwd = "C:/Users/gartc/Downloads/Data\_Exploration\_Rawdata (2)/Lab3\_Rawdata"

You can add options to executable code like this

# Read Google Trends data  
google\_trends\_files <- list.files(pattern = "trends\_up\_to\_.\*\\.csv")  
google\_trends <- import\_list(google\_trends\_files, rbind = TRUE, fill = TRUE)  
  
# Read Scorecard data   
scorecard <- import("Most+Recent+Cohorts+(Scorecard+Elements).csv")  
id\_name\_link <- import("id\_name\_link.csv")  
  
# Merge Scorecard data with Google Trends data  
merged\_data <- scorecard %>%  
 inner\_join(id\_name\_link, by = c("UNITID" = "unitid")) %>%  
 inner\_join(google\_trends, by = c("schname" = "schname"))

Warning in inner\_join(., google\_trends, by = c(schname = "schname")): Detected an unexpected many-to-many relationship between `x` and `y`.  
ℹ Row 1 of `x` matches multiple rows in `y`.  
ℹ Row 1241618 of `y` matches multiple rows in `x`.  
ℹ If a many-to-many relationship is expected, set `relationship =  
 "many-to-many"` to silence this warning.

# Calculate median earnings  
median\_earnings <- median(merged\_data$`md\_earn\_wne\_p10-REPORTED-EARNINGS`)

Warning in mean.default(sort(x, partial = half + 0L:1L)[half + 0L:1L]):  
argument is not numeric or logical: returning NA

# Create earnings variable and high-earning college indicator  
merged\_data <- merged\_data %>%  
 mutate(earnings = as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`),  
 high\_earning\_college = ifelse(earnings > median\_earnings, 1, 0))

Warning: There was 1 warning in `mutate()`.  
ℹ In argument: `earnings = as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`)`.  
Caused by warning:  
! NAs introduced by coercion

# Calculate SES as the average of earnings + high earning college  
merged\_data <- merged\_data %>%  
 mutate(ses = (earnings + high\_earning\_college) / 2)  
  
# Standardize Google Trends indices  
merged\_data <- merged\_data %>%  
 group\_by(keynum) %>%  
 mutate(standardized\_index = (index - mean(index)) / sd(index))  
  
# Perform DID regression  
#did\_model <- lm(standardized\_index ~ earnings \* high\_earning\_college + ses, data = merged\_data)  
  
# Summarize regression results  
#summary(did\_model)  
  
# Plotting  
#ggplot(merged\_data, aes(x = monthorweek, y = standardized\_index, color = as.factor(high\_earning\_college))) +  
 #geom\_line() +  
 #labs(x = "Month/Week", y = "Standardized Index", color = "High Earning College") +  
 #theme\_minimal()

INTRO

In this analysis, we will investigate whether the release of the College Scorecard has impacted Google search activity for low and high earning colleges. Similar to the content we have been learning in class, we will be performing a Difference-in-Differences (DID) regression analysis and have a visualization for the trends in the search activity overtime.

REGRESSION ANALYSIS

The Difference-in-Differences is vital to have because it helps with the regression analysis. In this case, we will be using the DID to estimate the effect of the College Scorecard release on Google search activity. We tackle the interaction terms between ‘monthorweek’ and post-Scorecard periods on whether or not the college is a high earning contender. We can use ‘monthorweek’ as a control variable to help with all the potential variability within the potential time trends. By having this regression model, it allows us to examine the change in search activity and whether or not high-earning colleges differentiate from low-earning colleges once the Scorecard is released

The coefficients on the Regression Analysis include:

**|intercept | post-scorecard | high earning college | interaction term|**

The intercept is a standardizerd index for low-earning colleges during the pre-scorecard period. The post-scorecard coefficient captrues the average change in the standardized index for all the colleges after the scorecards release. High earning college indicates the average difference in the standardized index between high and low earning schools pre scorecard. Finally, the interaction term (which is post scorecard period \* igh earning college) helps measure the differential change in standardized index between the high and low earning colleges after the scorecards release. Having a positive coefficient means that the high earning colleges experience greater increase in search activity compared to low earning colleges post scorecard. Vice versa for having a negative coefficient. Lastly, if the coefficient for the interaction term is NOT statistically significant, it means there is no differential change in the search activity between the high and low earning schools after the scorecard has been shown.

GRAPHICAL ANALYSIS

to visualize trends in search activity overtime for high and low earning colleges i have created a line graph. the x axis represents time (month/week) and the y axis represents standardized google trends index. I have coded different colors represent high and low earning colleges. The graph allows us to make observations on the differences that occur in search activity trends before and after the college scorecard release between high and low earning colleges.

Unfortunately my graph didn’t produce, but if there is noticeable divergence in search activity trends between the high and low earning schools after the scorecard has been released, you can assume that there are differential effects on the scorecard based on search behavior. On the other hand, if the trends remain parallel before and after the scorecard release, it means that the scorecard did not significantly impact search activity based off of the college earnings.

CONCLUSION

Overall i messed up and could not get my data to produce, but we have a good understanding on what the potential results could mean :)