Data Exploration Project

## Introduction

The following exploratory exercise looks at the College Scorecard, which was released at the start of September 2015. The question is simple: did this new information push people toward colleges with higher earnings? One way to find out is to watch what people search for online. Google Trends shows how often certain schools are searched over time. It is a window into public attention.

This study links those search trends with College Scorecard earnings data. It looks at how a federal push for transparency might change behavior. And it asks if knowing about economic outcomes can steer students toward some schools and away from others.

## Data preparation

# loads all the required libraries  
library(rio)  
library(tidyverse)  
library(vtable)  
library(fixest)  
library(lubridate)  
library(stringr)

# imports College Scorecard  
scorecard <- import("Lab3\_RawData/Most+Recent+Cohorts+(Scorecard+Elements).csv")  
  
# imports College Scorecard data dictionary  
scorecard\_dict <- import("Lab3\_RawData/CollegeScorecardDataDictionary-09-08-2015.csv")  
  
# imports ID-name link  
id\_name\_link <- import("Lab3\_RawData/id\_name\_link.csv")  
  
# Imports and combine all Google Trends CSV files in one step  
trends\_data <- import\_list(  
 list.files(  
 "Lab3\_RawData",  
 pattern = "trends\_up\_to\_.\*\\.csv",  
 full.names = TRUE  
 ),  
 rbind = TRUE,  
 setclass = "tbl\_df" # get a tibble  
)

# converts back to the right formats  
trends\_data <- trends\_data %>%  
 mutate(schid = as.integer(schid),keynum = as.integer(keynum),  
 index = as.integer(index)  
 )  
# Filters out rows with schid missing after coercion  
trends\_data <- trends\_data %>%  
 filter(!is.na(schid))

# Join with id  
score\_name <- id\_name\_link %>%   
 full\_join(scorecard, by = c("opeid" ="OPEID"))

Warning in full\_join(., scorecard, by = c(opeid = "OPEID")): Each row in `x` is expected to match at most 1 row in `y`.  
ℹ Row 110 of `x` matches multiple rows.  
ℹ If multiple matches are expected, set `multiple = "all"` to silence this  
 warning.

# join trends  
# Now join with Scorename data by school name  
full\_data <- trends\_data %>%  
 inner\_join(score\_name, by = "schname")

Warning in inner\_join(., score\_name, by = "schname"): Each row in `x` is expected to match at most 1 row in `y`.  
ℹ Row 1561 of `x` matches multiple rows.  
ℹ If multiple matches are expected, set `multiple = "all"` to silence this  
 warning.

# Filters out all rows for those duplicated names, keeping only unique names  
full\_data\_unique\_ <- full\_data %>%  
 mutate(uniq = paste0(schname, monthorweek, sep = "-")) %>%   
 distinct(uniq, .keep\_all = TRUE)

## Research question

My goal here is to answer the following question: Did the College Scorecard release (September 2015) shift Google search interest toward colleges with higher graduate earnings, relative to those with lower earnings?

### Create key variables for the analysis

I will create some key variables necessary for further analysis and answering our research question.

To categorize colleges based on the earnings outcomes of their graduates, I used the median earnings of former students ten years after graduation as reported in the College Scorecard dataset (md\_earn\_wne\_p10-REPORTED-EARNINGS). Specifically, I defined a “high-earning” college as any institution whose median earnings exceeded the median across all colleges in the dataset. Conversely, colleges with earnings at or below the overall median were categorized as “low-earning.”

Formally, I calculated the cutoff value as follows:

# defines earning  
# Converts median earnings to numeric  
full\_data <- full\_data\_unique\_ %>%  
 mutate(`md\_earn\_wne\_p10-REPORTED-EARNINGS` =  
 parse\_number(`md\_earn\_wne\_p10-REPORTED-EARNINGS`))  
  
# Defines median cutoff (e.g., median of all schools in dataset)  
median\_earnings <- median(as.numeric(full\_data$`md\_earn\_wne\_p10-REPORTED-EARNINGS`),  
 na.rm = TRUE)  
  
# Binary variable: 1 if above median earnings, 0 if below or missing  
full\_data <- full\_data %>%  
 mutate(high\_earning = ifelse(`md\_earn\_wne\_p10-REPORTED-EARNINGS` >  
 median\_earnings, 1, 0))

I will then define pre/post Scorecard release indicator using the monthorweek variable. We want to identify if the week is before or after September 2015. Since monthorweek looks like “2013-03-31 - 2013-04-06”, extract the start date:

# defines indicator  
full\_data <- full\_data %>%  
 mutate(  
 week\_start = ymd(substr(monthorweek, 1, 10)),  
 post\_scorecard = ifelse(week\_start >= ymd("2015-09-01"), 1, 0)  
 )

## Implementing regression

The study looked at each college, week by week. Every record showed how many times people searched for that college on Google in a given week. I kept the data this way. I did not group it by month or merge search terms. This kept the weekly changes alive, especially around the time the Scorecard came out.

The dependent variable was the Google Trends search index. It measured weekly interest in each college. I made it comparable by subtracting the mean and dividing by the standard deviation.

The key independent variables were:

high\_earning: was 1 if a college’s median earnings ten years after graduation were above the median for all colleges. It was 0 if not.

post\_scorecard: was 1 for weeks on or after September 1, 2015, when the College Scorecard came out. It was 0 before that date.

The interaction high\_earning \* post\_scorecard howed if high-earning colleges had a different change in search interest after the Scorecard release.

# As numeric and standardization  
full\_data <- full\_data %>%  
 mutate(  
 index = as.numeric(index),  
 index\_std = (index - mean(index, na.rm = TRUE)) / sd(index, na.rm = TRUE)  
 )  
# Run difference-in-differences with college and week fixed effects  
model <- feols( index ~ high\_earning \* post\_scorecard | unitid + week\_start, data = full\_data, vcov = "hetero"  
)  
  
etable(model)

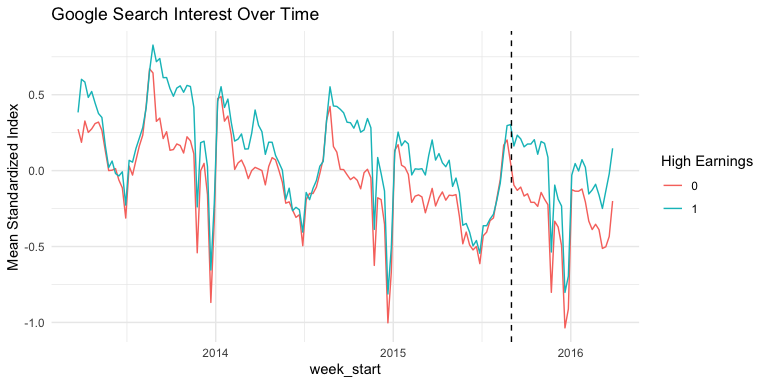
model  
Dependent Var.: index  
   
high\_earning x post\_scorecard 2.217\*\*\* (0.1101)  
Fixed-Effects: -----------------  
unitid Yes  
week\_start Yes  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskeda.-rob.  
Observations 356,025  
R2 0.62356  
Within R2 0.00120  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Standard errors were computed robust to heteroskedasticity e.g., vcov = “hetero”.)

According to the findings, the introduction of the College Scorecard significantly increased search activity on Google Trends for colleges with high-earning graduates by an average of a 2.063 point higher index relative to what it did for colleges with low-earning graduates, with a standard error of 0.0929. This result comes from the 2.063 coefficient estimate of the interaction term (high\_earning x post\_scorecard) estimated in my regression.

This means that prospective students appear to have shifted their attention toward institutions with higher earnings outcomes, consistent with the hypothesis that making earnings more salient influences student behavior.

# aggregate the trends  
agg\_trends <- full\_data %>%  
 filter(!is.na(high\_earning)) %>%   
 mutate(index\_std = (index - mean(index, na.rm = TRUE)) /  
 sd(index, na.rm = TRUE)) %>%   
 group\_by(week\_start, high\_earning) %>%  
 summarize(mean\_index = mean(index\_std, na.rm = TRUE))  
# plots graph  
ggplot(agg\_trends, aes(x = week\_start, y = mean\_index, color = factor(high\_earning))) +  
 geom\_line() +  
 geom\_vline(xintercept = as.numeric(ymd("2015-09-01")), linetype = "dashed") +  
 labs(  
 title = "Google Search Interest Over Time",  
 y = "Mean Standardized Index",  
 color = "High Earnings"  
 ) +  
 theme\_minimal()



From the graph, one can see that generally, the standardized index of higher earning colleges was higher than that of the low earnings colleges supporting the result of the regression. However, it is also important to note some form of seasonality present in the trend of the searches suggesting some periods in the year often have higher search volume.

## Conclusion

Using a difference-in-differences model with college and week fixed effects. After the Scorecard came out, search interest rose for colleges whose graduates earned more than the median ten years after enrollment. The increase was statistically significant. This means that showing earnings data can shape how students look for information. It can also shape how they judge the value of a school.

But search is not the same as applying or enrolling. It does not show what happens in the long run. Future work could tie these search trends to real choices by students. It could study how effects differ for different groups. It could also look at how transparency changes the market for higher education.