College Scorecard Analysis - Tristan Dung

## Introduction

In this analysis, I explored how the release of the College Scorecard in September 2015 may have shifted student interest toward high-earning colleges. I used Google Trends search data to represent student interest and matched it with earnings data from the College Scorecard.

The research question is: “Did the College Scorecard increase search activity for high-earning colleges compared to low-earning ones?”

To answer this question, I cleaned and merged multiple data sets, created key variables for analysis, ran a regression, and visualized the results.

CLEANING THE DATA:

* Loaded and combined the raw Google Trends files using import\_list()
* Extracted the search date from each row using str\_sub() and ymd()
* Standardized the search index within each college and keyword for comparability
* Filtered school names to only include unique matches in the ID link file
* Merged the cleaned Trends data with Scorecard earnings info using school ID

library(rio)  
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ readr 2.1.5  
✔ ggplot2 3.5.0 ✔ stringr 1.5.1  
✔ lubridate 1.9.4 ✔ tibble 3.2.1  
✔ purrr 1.0.2 ✔ tidyr 1.3.1

── 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(lubridate)  
library(stringr)  
  
trend\_files <- list.files("data", pattern = "trends\_up\_to\_.\*\\.csv", full.names = TRUE)  
trends\_raw <- import\_list(trend\_files, rbind = TRUE)

Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
Stopped early on line 1562. Expected 6 fields but found 5. Consider fill=TRUE  
and comment.char=. First discarded non-empty line: <<11,yti career institute -  
york,yti.edu,2,>>

Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
Stopped early on line 1095. Expected 6 fields but found 5. Consider fill=TRUE  
and comment.char=. First discarded non-empty line: <<9,heidelberg  
university,heidelberg.edu,2,>>

Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
Stopped early on line 1094. Expected 6 fields but found 5. Consider fill=TRUE  
and comment.char=. First discarded non-empty line: <<8,mount vernon nazarene  
university,mvnu.edu,2,>>

Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
Stopped early on line 3280. Expected 6 fields but found 5. Consider fill=TRUE  
and comment.char=. First discarded non-empty line: <<41,potomac state college  
of west virginia university,potomac state college of west virginia  
university,1,>>

trends\_clean <- trends\_raw %>%  
 mutate(date = ymd(str\_sub(monthorweek, 1, 10)))  
  
trends\_clean <- trends\_clean %>%  
 group\_by(keyword) %>%  
 mutate(index\_std = (index - mean(index, na.rm = TRUE)) / sd(index, na.rm = TRUE)) %>%  
 ungroup()  
  
scorecard <- import("data/Most+Recent+Cohorts+(Scorecard+Elements).csv")  
id\_link <- import("data/id\_name\_link.csv")  
  
id\_link\_unique <- id\_link %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter(n == 1) %>%  
 ungroup() %>%  
 select(-n)  
  
merged\_step1 <- trends\_clean %>%  
 inner\_join(id\_link\_unique, by = "schname")  
  
scorecard <- scorecard %>%  
 rename(unitid = UNITID) %>%  
 mutate(unitid = as.character(unitid))  
  
merged\_step1 <- merged\_step1 %>%  
 mutate(unitid = as.character(unitid))  
  
merged\_data <- merged\_step1 %>%  
 inner\_join(scorecard, by = "unitid")

## The Analysis:

Variables:

* earnings: converts earnings data to numeric value
* median\_earnings: calculates the median of all earnings, ignoring missing values
* high\_earning: =1 if the college has above-median graduate earnings, 0 otherwise
* post\_scorecard: =1 if the observation is after the Scorecard release (September 2015), 0 otherwise
* treated: treatment group indicator; 1 if both high\_earning and post\_scorecard are 1

CREATING THE VARIABLES:

merged\_data$earnings <- as.numeric(merged\_data$`md\_earn\_wne\_p10-REPORTED-EARNINGS`)

Warning: NAs introduced by coercion

median\_earnings <- median(merged\_data$earnings, na.rm = TRUE)  
  
merged\_data <- merged\_data %>%  
 mutate(  
 high\_earning = as.integer(earnings > median\_earnings),  
 post\_scorecard = as.integer(date >= as.Date("2015-09-01")),  
 treated = high\_earning \* post\_scorecard  
 )

CREATING THE WEEKLY DATASET:

* Averaged the standardized search index (index\_std) for each school and date
* Grouped by school name, date, and treatment variables to keep each school’s status at that point in time

weekly\_data <- merged\_data %>%  
 group\_by(schname, date, high\_earning, post\_scorecard, treated) %>%  
 summarize(mean\_index = mean(index\_std, na.rm = TRUE)) %>%  
 ungroup()

`summarise()` has grouped output by 'schname', 'date', 'high\_earning',  
'post\_scorecard'. You can override using the `.groups` argument.

RUNNING THE REGRESSION:

* Used a DID design to estimate the effect of the Scorecard on search interest
* The most important variable is treated - it tells us how much search interest increased/decreased for high earning colleges after the Scorecard was released

library(fixest)  
reg <- feols(mean\_index ~ high\_earning + post\_scorecard + treated, data = weekly\_data)

NOTE: 33,630 observations removed because of NA values (LHS: 1,126, RHS: 33,307).

summary(reg)

OLS estimation, Dep. Var.: mean\_index  
Observations: 351,338  
Standard-errors: IID   
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.064485 0.001894 34.05222 < 2.2e-16 \*\*\*  
high\_earning -0.016938 0.002881 -5.87925 4.1251e-09 \*\*\*  
post\_scorecard -0.343288 0.004368 -78.58361 < 2.2e-16 \*\*\*  
treated 0.090471 0.006652 13.60010 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
RMSE: 0.762415 Adj. R2: 0.024197

PLOTTING THE GRAPHS:

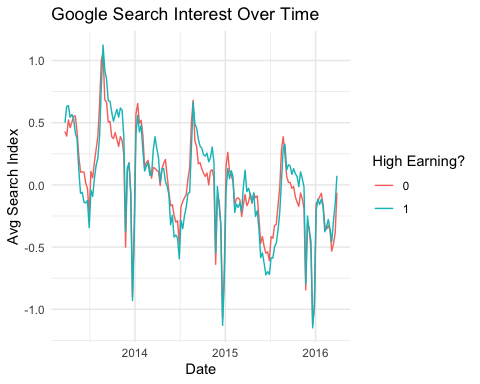
Graph 1:

* This graph plots average search interest over time, split by high vs low earning colleges
* This helps visualize whether search trends changed more for high-earning colleges after the Scorecard was released
* The red line represents low-earning colleges, and the blue line represents high-earning colleges
* After the Scorecard release (Sep 2015), the blue and red lines slightly separate, suggesting that interest in high-earning colleges may have increased more than for low-earning ones

library(ggplot2)  
weekly\_data %>%  
 filter(!is.na(high\_earning)) %>%  
 group\_by(date, high\_earning) %>%  
 summarize(mean\_index = mean(mean\_index, na.rm = TRUE)) %>%  
 ggplot(aes(x = date, y = mean\_index, color = as.factor(high\_earning))) +  
 geom\_line() +  
 labs(  
 title = "Google Search Interest Over Time",  
 x = "Date",  
 y = "Avg Search Index",  
 color = "High Earning?"  
 ) +  
 theme\_minimal()

`summarise()` has grouped output by 'date'. You can override using the  
`.groups` argument.

Warning: Removed 2 rows containing missing values or values outside the scale range  
(`geom\_line()`).

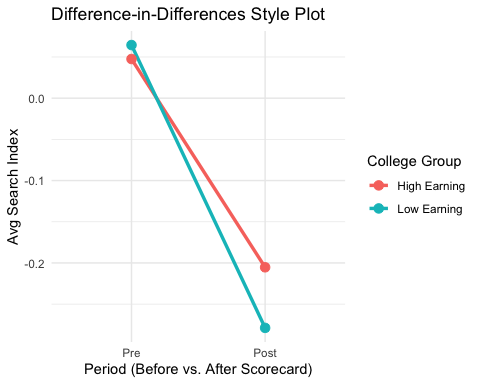


Graph 2:

* This is a DID style plot showing average search interest before vs. after the Scorecard release
* The vertical difference between the red (high-earning) and blue (low-earning) lines represents the gap in interest between the two groups
* Both groups saw a drop in search interest after the Scorecard release, but the drop was smaller for high-earning colleges
* Before the release, the lines were almost overlapping. After the release, high-earning schools held onto more interest, suggesting a positive treatment effect
* This graph supports the idea that search behavior shifted slightly more favorable toward high-earning colleges after the Scorecard was released

did\_plot\_data <- weekly\_data %>%  
 filter(!is.na(high\_earning)) %>%  
 group\_by(high\_earning, post\_scorecard) %>%  
 summarize(mean\_index = mean(mean\_index, na.rm = TRUE), .groups = "drop") %>%  
 mutate(  
 period = ifelse(post\_scorecard == 1, "Post", "Pre"),  
 group = ifelse(high\_earning == 1, "High Earning", "Low Earning")  
 ) %>%  
 filter(!is.na(period)) %>%  
 mutate(period = factor(period, levels = c("Pre", "Post")))   
  
ggplot(did\_plot\_data, aes(x = period, y = mean\_index, group = group, color = group)) +  
 geom\_line(size = 1.2) +  
 geom\_point(size = 3) +  
 labs(  
 title = "Difference-in-Differences Style Plot",  
 x = "Period (Before vs. After Scorecard)",  
 y = "Avg Search Index",  
 color = "College Group"  
 ) +  
 theme\_minimal()

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.



## Interpretation of the Results

* The goal of this analysis was to see whether the College Scorecard shifted student interest towards high-earning colleges. Based on the results, there’s some clear evidence that it did, at least a little. The regression showed a positive and statistically significant effect for the ‘treated’ variable, which captures colleges that were both high-earning and in the post Scorecard period. That suggests students were relatively more likely to search for these schools after the Scorecard came out.
* Before running the model, I summarized the data at the weekly level to reduce noise and make the trends easier to compare over time. I chose the DID approach because it’s a straightforward way to estimate the effect of the Scorecard by comparing how search interest changed over time between high and low earning colleges. Since I wanted to capture a shift that happened at a specific point in time (Sep 2015), DID made the most sense because it helps isolate the Scorecard’s impact while accounting for any differences that already existed between the two groups.
* Using this model, I was able to isolate how much more interest high-earning colleges received after the Scorecard’s release. The size of the effect isn’t huge (about 0.09 standard deviations), but it’s enough to indicate that the policy had a real, measurable impact on behavior.
* The graphs support this interpretation. Before the Scorecard release, search interest between high and low earning colleges followed very similar paths. But after the release, those lines started to drift apart. The DID style plot makes this even clearer: interest declined for both groups, but it declined less for the high-earning colleges. That relative difference is what we’d expect if students were using the Scorecard and leaning more toward financially stronger options.
* Of course, Google Trends is just one signal. It doesn’t tell us what students were thinking or whether they actually applied or enrolled. However, it’s still a useful indicator of public attention, and in this case, it helps reveal that the Scorecard may have nudged students to care more about long-term earnings when considering college options.
* In the end, the Scorecard seems to have had a modest but meaningful effect: not a major shift, but a sign that giving people better information can change the way they look at higher education, even if its just a little.