Data Exploration

Kieu Trinh

## Libraries

Load all the 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 3.5.0 ✔ 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)  
library(lubridate)  
library(fixest)

## Import Data

Import the trends files, the link file and the scorecard file to perform data cleaning and later regression.

# List all trends files  
trend\_files <- list.files(pattern = "^trends\_up\_to\_.\*\\.csv$")  
  
trends\_all <- 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,>>

id\_link <- import("id\_name\_link.csv")  
  
scorecard <- import("Most+Recent+Cohorts+(Scorecard+Elements).csv")

## Data Cleaning

Clean Google Trends data by removing missing dates, converting to monthly periods, and standardizing search indices within each school keyword pair for comparability across series.

trends\_all <- trends\_all %>%  
 filter(!is.na(monthorweek) & str\_trim(monthorweek) != "") %>%  
 mutate(date = ymd(str\_sub(monthorweek, 1, 10)))  
  
# Go by the month  
trends\_all <- trends\_all %>%  
 mutate(month = floor\_date(date, unit = "month"))  
  
#standardize index within each schname and keyword  
trends\_all <- trends\_all %>%  
 group\_by(schname, keyword) %>%  
 mutate(new\_index = (index - mean(index, na.rm = TRUE)) / sd(index, na.rm = TRUE)) %>%  
 ungroup()

Clean and convert earnings to numeric, select, rename relevant Scorecard columns, filtered to bachelor’s-degree–granting colleges with non-missing earnings.

scorecard\_clean <- scorecard %>%  
 select(  
 UNITID,  
 OPEID,  
 INSTNM,  
 PREDDEG,  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS`  
 ) %>%  
 rename(  
 schname = INSTNM,  
 degree\_type = PREDDEG,  
 earnings\_raw = `md\_earn\_wne\_p10-REPORTED-EARNINGS`  
 ) %>%  
 mutate(  
 earnings = earnings\_raw %>%  
 str\_remove\_all(",") %>%  
 str\_trim() %>%  
 na\_if("PrivacySuppressed") %>%  
 as.numeric()  
 ) %>%  
 filter(degree\_type == 3, !is.na(earnings))

Warning: There was 1 warning in `mutate()`.  
ℹ In argument: `earnings = `%>%`(...)`.  
Caused by warning in `earnings\_raw %>% str\_remove\_all(",") %>% str\_trim() %>% na\_if(  
 "PrivacySuppressed") %>% as.numeric()`:  
! NAs introduced by coercion

#lowercase key variable to match with other dataset  
scorecard\_clean <- scorecard\_clean %>%  
 rename(unitid = UNITID, opeid = OPEID)

Filter ID link data to keep only schools with unique names.

id\_link\_clean <- id\_link %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter(n == 1) %>%  
 select(-n)

Join cleaned Trends data with ID link, then merge with Scorecard data to create the final combined dataset.

trends\_id <- inner\_join(trends\_all, id\_link\_clean, by = "schname")  
trends\_scorecard\_final <- inner\_join(trends\_id, scorecard\_clean, by = c("unitid", "opeid"))

Create a post-Scorecard indicator for dates after September 2015 and a high-earning flag for colleges with earnings above the median.

trends\_scorecard\_final <- trends\_scorecard\_final %>%  
 mutate(post\_scorecard = as.integer(month >= as.Date("2015-09-01")))  
  
median\_earn <- median(trends\_scorecard\_final$earnings, na.rm = TRUE)  
  
# determine if the college is above or below medium, 1 for above, 0 for below  
trends\_scorecard\_final <- trends\_scorecard\_final %>%  
 mutate(high\_earning = as.integer(earnings >= median\_earn))

Aggregate data to the college month level, calculating average standardized search interest and keyword count for each group.

college\_monthly\_interest <- trends\_scorecard\_final %>%  
 group\_by(unitid, month, high\_earning, post\_scorecard) %>%  
 summarise(  
 avg\_index = mean(new\_index, na.rm = TRUE),  
 n\_keywords = n(),  
 .groups = "drop"  
 )

Export a single clean file

dir.create("data\_clean", showWarnings = FALSE, recursive = TRUE)  
export(college\_monthly\_interest, "data\_clean/gt\_scorecard\_panel\_monthly.csv")

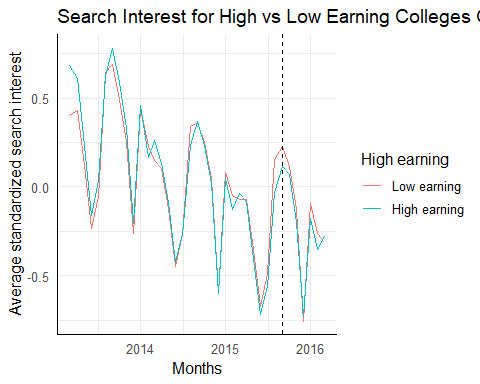
## Graph

Calculates the average standardized index for high- and low-earning colleges, and plots their trends over time with a vertical line marking the College Scorecard release in September 2015.

data <- import("data\_clean/gt\_scorecard\_panel\_monthly.csv")

Plot the high and low earning

monthly\_interest\_trends <- data %>%  
 group\_by(month, high\_earning) %>%  
 summarise(avg\_std\_index = mean(avg\_index, na.rm = TRUE), .groups = "drop") %>%  
 mutate(high\_earning = factor(high\_earning,  
 levels = c(0, 1),  
 labels = c("Low earning", "High earning")))  
  
  
ggplot(monthly\_interest\_trends, aes(month, avg\_std\_index, color=factor(high\_earning))) +  
 geom\_line() +  
 geom\_vline(xintercept = as.Date("2015-09-01"), linetype=2) +  
 labs(title = "Search Interest for High vs Low Earning Colleges Over Time",  
 color="High earning",   
 y="Average standardized search interest",   
 x="Months")+  
 theme\_minimal(base\_size = 12) +  
 theme(  
 plot.background = element\_rect(fill = "white", color = NA),  
 panel.background = element\_rect(fill = "white", color = NA),  
 axis.line = element\_line(color = "black", linewidth = 0.5)  
 )



## DID regression

scorecard\_did <- feols(  
 avg\_index ~ high\_earning \* post\_scorecard | unitid + month,  
 cluster = ~ unitid,  
 data = data  
)

The variables 'high\_earning' and 'post\_scorecard' have been removed because of collinearity (see $collin.var).

etable(scorecard\_did)

scorecard\_did  
Dependent Var.: avg\_index  
   
high\_earning x post\_scorecard -0.0674\*\* (0.0243)  
Fixed-Effects: ------------------  
unitid Yes  
month Yes  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E.: Clustered by: unitid  
Observations 51,565  
R2 0.38608  
Within R2 0.00082  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1