Data Exploration Assignment

Anh Nhan

## Libraries

library(rio)  
library(lubridate)

Attaching package: 'lubridate'

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

library(fixest)  
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 ✔ stringr 1.5.0  
✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
✔ purrr 1.0.1 ✔ tidyr 1.3.0  
✔ readr 2.1.4

── 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(haven)  
library(purrr)  
library(stringr)  
library(openxlsx)  
library(ggplot2)

# 1. Data Cleaning

## 1.1 Reading in the Google Trends data

google\_trends <- list.files(path = "Lab3\_Rawdata", pattern = "trends\_up\_to\_", full.names = TRUE)  
data\_1 <- import\_list(google\_trends, rbind = TRUE, fill = TRUE)

## 1.2 Aggregating the Google Trends data

data\_1 <- data\_1 %>%  
 mutate(date = ymd(str\_sub(monthorweek, 1, 10))) %>%   
 mutate(month = floor\_date(date, unit = 'month'))

# Standardize index  
data\_1 <- data\_1 %>%  
 group\_by(schname, keyword) %>%  
 mutate(index\_mean = mean(index, na.rm = TRUE),  
 index\_sd = sd(index, na.rm = TRUE),  
 index\_standardized = (index - index\_mean)/index\_sd)  
  
  
# Aggregate the standardized index  
data\_1\_standardized <- data\_1 %>%  
 group\_by(keyword, month) %>%  
 summarise(index\_mean\_standardized = mean(index\_standardized, na.rm = TRUE))

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

## 1.3 Reading the scorecard\_data and id\_name\_link data

scorecard\_data <- import("Lab3\_Rawdata/Most+Recent+Cohorts+(Scorecard+Elements).csv")  
  
  
# Limiting the data to colleges that predominantly grant banchelor's degrees  
scorecard\_data <- scorecard\_data %>%  
 filter(PREDDEG == 3)  
  
  
id\_name\_link <- import("Lab3\_Rawdata/id\_name\_link.csv")

## 1.4 Joining data

# Merge in the Scorecard data  
  
# Drop school name that appear more than 1  
id\_name\_link <- id\_name\_link %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter( n == 1)  
  
# Rename unitid since there is uppercase in the original scorecard\_data  
scorecard\_data <- scorecard\_data %>%  
 rename(unitid = UNITID)  
  
# Join data   
cleaned\_data <- data\_1 %>%  
 inner\_join(id\_name\_link, by = "schname") %>%  
 inner\_join(scorecard\_data, by = "unitid")

# 2. Defining “High-earnings” and “Low-earnings” colleges

# Rename a column  
cleaned\_data <- cleaned\_data %>%  
 rename(median\_earnings = `md\_earn\_wne\_p10-REPORTED-EARNINGS`)

# Converting median\_earnings to numerical  
cleaned\_data$median\_earnings <- as.numeric(as.character(cleaned\_data$median\_earnings))

Warning: NAs introduced by coercion

# Calculate the median of median earnings as the level  
median\_earnings\_level <- median(cleaned\_data$median\_earnings, na.rm = TRUE)

Since the median earning is $41,700, I assume that any earning below the median is consider “low-earnings”, and any earning above the median is consider “high-earnings”.

# Categorize colleges based on the median earnings level  
cleaned\_data <- cleaned\_data %>%  
 mutate(earnings\_category = ifelse(median\_earnings > median\_earnings\_level, "High-Earning", "Low-Earning"))

# 3. Regression model design

## 3.1 Creating a new data frame for the analysis

# Selecting variables for the analysis  
selected\_data <- cleaned\_data %>%  
 select(schname, keyword, month, index\_standardized, median\_earnings, earnings\_category)

# Remove all the NAs variable  
selected\_data <- drop\_na(selected\_data)

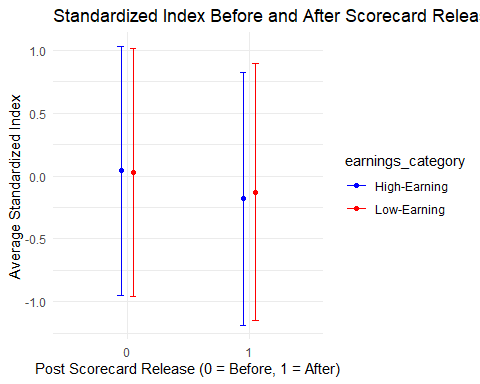
## 3.2 Regression Model

# Creating a time indicator  
selected\_data <- selected\_data %>%  
 mutate(post\_scorecard = as.integer(month >= as.Date("2015-09-01")))  
  
# Regression model  
regression\_model <- feols(index\_standardized ~ post\_scorecard \* earnings\_category, data = selected\_data)  
  
# Display the regression table  
etable(regression\_model)

regression\_model  
Dependent Var.: index\_standardized  
   
Constant 0.0420\*\*\* (0.0016)  
post\_scorecard -0.2234\*\*\* (0.0038)  
earnings\_categoryLow-Earning -0.0123\*\*\* (0.0023)  
post\_scorecard x earnings\_categoryLow-Earning 0.0658\*\*\* (0.0053)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 910,198  
R2 0.00575  
Adj. R2 0.00574  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# 4. Visualizing the graph

# Creating the plot data  
plot\_data <- selected\_data %>%  
 group\_by(post\_scorecard, earnings\_category) %>%  
 summarise(mean\_index = mean(index\_standardized, na.rm = TRUE),  
 sd\_index = sd(index\_standardized, na.rm = TRUE),  
 .groups = 'drop')  
  
  
# Visualizing the plot data  
ggplot(plot\_data, aes(x = factor(post\_scorecard), y = mean\_index, color = earnings\_category)) +  
 geom\_point(position = position\_dodge(width = 0.2)) +   
 geom\_errorbar(aes(ymin = mean\_index - sd\_index, ymax = mean\_index + sd\_index), width = 0.1, position = position\_dodge(width = 0.2)) +  
 labs(x = "Post Scorecard Release (0 = Before, 1 = After)",  
 y = "Average Standardized Index",  
 title = "Standardized Index Before and After Scorecard Release") +  
 theme\_minimal() +  
 scale\_color\_manual(values = c("High-Earning" = "blue", "Low-Earning" = "red"))



# 5. Write-up

### Introduction

This study investigates whether the College Scorecard’s release in September 2015 influenced students’ online search behaviors regarding universities. It examines if there was a noticeable shift in students searching for universities that are typically associated with higher graduate earnings compared to those with lower graduate earnings, following the Scorecard’s availability.

### Data Cleaning

The data cleaning process began with the combining and organizing of Google Trends data and the College Scorecard information to ensure accurate matching across datasets. Special attention was given to removing duplicates variables with the same name were excluded to prevent any potential confusion in the analysis.

The chosen analytical approach directly addresses the central research question by measuring the relative change in search interest for different categories of colleges due to a significant policy change. The use of Google search trends as a proxy for student interest is justified given the platform’s ubiquity and its role in college research. The decision to classify colleges into “high-earning” and “low-earning” based on median earnings is defended by the median values, which allows for a dichotomous comparison while accepting the simplification’s limitations. The data was aggregated monthly to smooth out short-term fluctuations and to capture more sustained trends. This level of aggregation was chosen over weekly data to reduce noise and over longer periods to maintain sufficient temporal resolution for observing the Scorecard’s impact.

### **Defining “High-Earning” and “Low-Earning” Colleges**

A critical aspect of the analysis involved categorizing colleges into “high-earning” and “low-earning” groups according to their graduates’ median earnings. The median value derived from the Scorecard data, $41,700, was established as the benchmark. Institutions with median graduate earnings above this figure were designated as “high-earning,” while those falling below were considered “low-earning.” This binary classification system streamlined the analysis which enabling straightforward comparisons. However, it also simplified the diverse earnings spectrum, bypassing the more intricate details of the earnings distribution.

### Regression Model

Since the regression model was designed to answer the research question, I want to go through these variables before getting in deep into the analysis.

post\_scorecard: indicates whether the search was made after the release on Scorecard

earnings\_categoryLow-Earning: the difference between high-earning and low-earning colleges based on median\_earnings

post\_scorecard x earnings\_categoryLow-Earning: this term is to identify the differential impact post\_ scorecard release on low-earning colleges.

The coefficient for post\_scorecard is -0.2234 (standard error = 0.0038), indicating a significant decrease in search interest for high-earning colleges after the Scorecard’s introduction. This suggests that the Scorecard may have led prospective students to reconsider the value proposition of such institutions. The coefficient for earnings\_categoryLow-Earning is -0.0123 (standard error = 0.0023), indicating that even before the Scorecard’s release, low-earning colleges had slightly lower search interest compared to high-earning colleges. The last term has the coefficient of 0.0658 (standard error = 0.0053) is signifies that the decrease in search interest for high-earning colleges post-Scorecard was not as pronounced for low-earning colleges. This could mean that the Scorecard may have positively influenced the perception or awareness of low-earning colleges.

### Visual Analysis

The graph illustrates the average standardized search index for high-earning (blue) and low-earning (red) colleges both before (0) and after (1) the Scorecard’s release. The lines sticking out from the dots, called error bars, show how scattered the search numbers are from the average.

Looking at the chart, there is a clear decline in search interest for high-earnings colleges post-Scorecard release, while the interest in low-earning colleges remains relatively stable. The error bars pre- and post-release, particularly for low-earning colleges, suggests that the variability in search interest is significant and could potentially affect the model’s estimates. This variation might throw off the analysis.

### Conclusion

To wrap up the project’s findings: “The introduction of the College Scorecard decreased search activity on Google Trends for colleges with high-earning graduates by 0.2234 units relative to what it did for colleges with low-earning graduates, with a standard error of 0.0038. This result comes from the post\_scorecard coefficient in my regression.” Additionally, the introduction of the College Scorecard increased search activity for low-earning colleges by 0.0658 units, as indicated by the positive interaction term coefficient with a standard error of 0.0053.

The analysis answers the research question by demonstrating that the College Scorecard has influenced Google search behavior in favor of lower-earning institutions. Although the changes is relatively small, it suggests a meaningful impact on public interest, potentially affecting college choice and, by extension, the diversity of college applications. In conclusion, the impact of the College Scorecard on student interest is significant and measurable. It appears to have achieved its intended purpose of highlighting a broader range of colleges, thus influencing student search behavior. As we can deduce from the regression model and graphical analysis, the Scorecard’s release correlates with a shift in interest towards colleges that do not top the earnings chart. This insight is crucial for policymakers and educational institutions aiming to understand and influence educational choices in the information age.