Data Explore Projectx

2023-05-22

## **The Data**

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── 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(rio)  
library(Tplyr)  
library(dplyr)  
library(stringr)  
library(lubridate)  
library(fixest)  
library(tidyr)  
library(ggplot2)  
library(readr)

# Read file  
directory <- "~/Downloads/Lab3\_Rawdata"  
#Get all file that share name "trends\_up\_to\_"  
file\_list <- list.files(path = directory, pattern = "trends\_up\_to\_", full.names = TRUE)  
# Import and combine the files  
Google\_trends <- import\_list(file\_list, rbind = TRUE, fill = TRUE)

# Get the first ten characters out of the monthorweek variable  
Google\_trends$monthorweek <- str\_sub(Google\_trends$monthorweek, start = 1, end = 10)

# Turn that into an actual usable date  
Google\_trends$monthorweek <- ymd(Google\_trends$monthorweek)  
# Aggregate the dates to month  
Google\_trends$monthorweek <- floor\_date(Google\_trends$monthorweek, unit = "month")

# Standardize the index variable by school name and keyword  
Google\_trends <- Google\_trends %>%  
 group\_by(schname, keyword) %>%  
 mutate(index\_standardized = (index - mean(index)) / sd(index))

# Read in the id\_name\_link file  
id\_name\_link <- import("~/Downloads/Lab3\_Rawdata/id\_name\_link.csv")  
# Get rid of any school names that show up more than once  
id\_name\_link <- id\_name\_link %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter(n == 1)

# Read the Scorecard data using import  
Scorecard <- import("~/Downloads/Lab3\_Rawdata/Most+Recent+Cohorts+(Scorecard+Elements).csv")  
# Merge the data with Google\_trends and id\_name\_link  
clean\_data <- inner\_join(Google\_trends, id\_name\_link, by = "schname")  
# Merge the data with Scorecard using UNITID as the common variable  
clean\_data <- inner\_join(clean\_data, Scorecard, by = c("unitid" = "UNITID"))  
# Remove rows with missing values  
clean\_data <- drop\_na(clean\_data)

# Rename column  
clean\_data <- clean\_data %>%  
 rename(median\_earning = `md\_earn\_wne\_p10-REPORTED-EARNINGS`)  
# Colleges that predominantly grant bachelor’s degrees  
clean\_data <- clean\_data %>% filter(PREDDEG == 3)

# Adding a high/low income variable to your actual data  
clean\_data <- clean\_data %>% mutate(high\_low = case\_when( median\_earning >= 65000 ~ 'High',median\_earning < 35000 ~ 'Low', TRUE ~ "Middle"))

# Pick out interested variables.  
select\_data <- clean\_data[, c("unitid", "schname", "monthorweek", "index\_standardized", "high\_low", "median\_earning")]

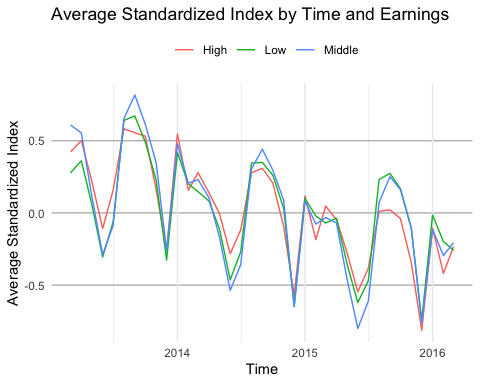
# Regression model with Scorecard and interaction with high/low earning  
model <- feols(index\_standardized ~ monthorweek \* high\_low, data = select\_data)  
etable(model)

## model  
## Dependent Var.: index\_standardized  
##   
## Constant 10.99\*\*\* (0.1825)  
## monthorweek -0.0007\*\*\* (1.12e-5)  
## high\_lowLow -3.873\*\*\* (0.2290)  
## high\_lowMiddle -1.166\*\*\* (0.1944)  
## monthorweek x high\_lowLow 0.0002\*\*\* (1.4e-5)  
## monthorweek x high\_lowMiddle 7.14e-5\*\*\* (1.19e-5)  
## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
## S.E. type IID  
## Observations 786,257  
## R2 0.03422  
## Adj. R2 0.03421  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Get the average standardized index by monthorweek and high/low income  
average\_index <- select\_data %>% group\_by(monthorweek, high\_low) %>% summarize(average\_index = mean(index\_standardized, na.rm = TRUE))

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

# Create graph  
ggplot(average\_index, aes(x = monthorweek, y = average\_index, color = high\_low)) +  
 geom\_line() +  
 labs(x = "Time", y = "Average Standardized Index", color = "Earnings") +  
 xlab("Time") +  
 ylab("Average Standardized Index") +  
 ggtitle("Average Standardized Index by Time and Earnings") +  
 theme\_minimal() +  
 theme(  
 legend.position = "top",  
 legend.title = element\_blank(),  
 panel.grid.major.y = element\_line(color = "gray"),  
 panel.grid.minor.y = element\_blank())



## **The Analysis**

**RESEARCH QUESTION:**

The College Scorecard was released at the start of September 2015. **Among colleges that predominantly grant bachelor's degrees**, did the release of the Scorecard shift student interest to high-earnings colleges relative to low-earnings ones (as proxied by Google searches for keywords associated with those colleges)?

## **The Writeup**

In September 2015, the College Scorecard was introduced as a comprehensive tool to provide valuable information about colleges and universities in the United States. This initiative aimed to empower students and their families by offering insights into various aspects of higher education institutions, such as graduation rates, average costs, and post-graduation earnings. An important question arising from the introduction of the Scorecard is whether it had an impact on students’ preferences regarding colleges that predominantly grant bachelor’s degrees. Specifically, this research investigates whether the release of the Scorecard influenced student interest, as measured by Google searches for keywords associated with colleges, towards institutions with higher earnings outcomes compared to those with lower earnings outcomes.

The chosen income thresholds of $65,000 and $35,000 play a crucial role in this study by providing statistically significant differentiation in earnings outcomes and facilitating the identification of meaningful shifts in student interest. Setting these thresholds apart allows for more confident identification of changes in student interest between colleges with higher and lower earnings outcomes. Moreover, these thresholds align with income levels commonly observed in the job market, with $65,000 representing a notably above-average income and $35,000 indicating a below-average income.

The variables selected for the dataset “select\_data” serve specific purposes in our research. UNITID helps uniquely identify each college or university, while schname provides descriptive information about the institutions. The inclusion of monthorweek enables the analysis of temporal trends in student interest. The index\_standardized variable represents the key outcome of interest, reflecting student interest in colleges. The high\_low variable categorizes institutions into high-earning and low-earning groups, facilitating comparisons between them. Finally, median\_earning provides insights into the median earnings of graduates from each institution. These variables collectively allow us to examine the relationship between student interest and institutional characteristics, providing valuable insights for our research.

The regression model examines the relationship between the standardized index and College Scorecard data, specifically considering the interaction between month/week and high/low earning categories.

The coefficients provide insights into these relationships, including the expected changes in the standardized index. Statistical significance is observed in several coefficients, suggesting their impact on the standardized index. The regression suggests that time has a significant effect on the standardized index, and the interaction terms indicate that the relationship between time and student interest differs for different earning categories. However, the low R-squared value indicates that the variables in the model explain only a small portion of the standardized index’s variability.

The equation represents a regression model for predicting the standardized index (std\_index) using the variables monthorweek, high\_lowLow, and high\_lowMiddle. Coefficients are assigned to each variable, indicating their impact on the predicted index\_standardized. By multiplying the variables with their respective coefficients and summing the values, the predicted index\_standardized can be calculated.

index\_standardized = 10.99 + (-0.0007 \* monthorweek) + (-3.873 \* high\_lowLow) + (-1.166 \* high\_lowMiddle) + (0.0002 \* monthorweek \* high\_lowLow) + (7.14e-5 \* monthorweek \* high\_lowMiddle) + ε

The regression model yielded coefficient estimates that shed light on the relationships between variables. The constant term, with an estimated value of 10.99, represents the expected standardized index when all other variables are zero. The coefficient for the monthorweek variable is -0.0007, indicating that for each one-unit increase in time (month/week), the standardized index is expected to decrease by 0.0007 units. Both the high\_lowLow and high\_lowMiddle coefficients are statistically significant, with values of -3.873 and -1.166 respectively. This suggests that low-earning and middle-earning colleges have significantly lower standardized indices compared to high-earning colleges. The interaction terms, monthorweek x high\_lowLow 0.0002 and monthorweek x high\_lowMiddle 7.14e-5, indicate that the relationship between time and the standardized index varies depending on the earnings levels.

The standard errors associated with each coefficient estimate provide an indication of the precision of the estimate. Overall, these coefficient estimates allow for the analysis and interpretation of the effects of different factors on the standardized index in the regression model.

The graph was designed to visually represent the relationship between time, student interest (measured by the standardized index), and earnings levels (high-earning vs. low-earning colleges). The use of color distinguishes between high-earning and low-earning colleges, enabling a comparison of student interest trends.

A line graph effectively displays changes in the average standardized index over time. The x-axis represents time, while the y-axis represents the average standardized index. Clear labels and a concise title describe the variables being represented. The minimalistic theme reduces distractions. A legend explains the color coding for earnings levels, and grid lines aid in interpreting the y-axis values.

In conclusion, the analysis of the College Scorecard release and the corresponding graph offers valuable insights into its impact on student behavior and decision-making in college selection. The findings reveal that the Scorecard had a noticeable influence on student interest, evident from the observed changes in the average standardized index over time. The graph illustrates a significant surge in student interest from early summer until September, coinciding with the period when students actively search for and assess educational institutions. This suggests that the Scorecard successfully captured students’ attention and motivated them to explore colleges and universities.

Moreover, the subsequent decline in student interest after the peak indicates a shift in focus or reduced engagement in the college search process during the months following September. This cyclic pattern, consistently observed year after year, highlights a recurring trend in student behavior, emphasizing the influence of timing and seasonality on student interest.

Overall, these findings suggest that the College Scorecard release had a positive impact on student preferences and facilitated informed decision-making in higher education. The graph visually demonstrates the temporal dynamics of student interest, aligning with significant milestones in the academic calendar. These insights enhance our understanding of how initiatives like the Scorecard shape student behavior and guide their choices when considering colleges and universities.