Data Exploration Assignment

Lorenzo Powell

2025-08-12

# load 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 3.5.2 ✔ tibble 3.3.0  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── 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(fixest)  
library(ggplot2)  
library(ggeffects) # for marginal effects predictions  
library(stringr)  
library(lubridate)  
library(janitor) # for clean\_names()

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(broom) # for tidy() df construction || visualization  
library(readr)  
library(rio)

To analyze the impacts of the College Scorecard on school interest we can use trends data from the most popular search engine, Google. My analysis aims to clearly indicate whether the policy implementation significantly impacts how prospective students engage with search tools for institutions considering their potential post-graduate outcomes.

The Data Exploration Assignment provided links to download the .zip file from <https://seattleu.instructure.com/courses/1622414/files/71814703/download?wrap=1>. Save to your local directory and open the file. In the following chunk, the user should define the paths prior to running this document to properly import data.

# NEW PATHS || UNCOMMENT TO RUN IN GITHUB  
# zip\_rel <- file.path("..", "OMSBA\_5300\_S25\_Data\_Exploration\_Assignment\_-LP/Data\_Exploration\_Rawdata.zip")   
# data\_dir <- file.path("..", "OMSBA\_5300\_S25\_Data\_Exploration\_Assignment\_-LP/Lab3\_Rawdata")   
#   
# if (!dir.exists(data\_dir)) {  
# unzip(zip\_rel, exdir = "..")  
# }  
#   
# dir\_path <- data\_dir  
# scorecard\_path <- file.path(dir\_path, "Most+Recent+Cohorts+(Scorecard+Elements).csv")  
# idlink\_path <- file.path(dir\_path, "id\_name\_link.csv")  
  
# paths  
dir\_path <- '/Users/lorenzopowell/Desktop/Data Exploration Assignment/Lab3\_Rawdata' # MODIFY PATH  
scorecard\_path <- file.path(dir\_path, 'Most+Recent+Cohorts+(Scorecard+Elements).csv')  
idlink\_path <- file.path(dir\_path, 'id\_name\_link.csv')  
  
# read scorecard data & id\_name\_link  
scorecard <- import(scorecard\_path) %>%  
 clean\_names()  
id\_link <- import(idlink\_path) %>%  
 clean\_names()  
  
# eliminate duplicate school entries from id\_link  
if (!'schname' %in% names(id\_link)) {  
 stop("`id\_name\_link.csv` no `schname` column present")  
}  
  
#final id\_link  
id\_link\_new<- id\_link %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 ungroup() %>%  
 filter(n == 1) %>% # filtering multi-campus schools // may want to include alternate with inclusion  
 select(-n)  
  
# load trend data  
trend\_files <- list.files(  
 path = dir\_path,  
 pattern = '^trends\_up\_to\_.\*\\.csv$',  
 full.names = TRUE)  
  
# import and bind trend files  
trends\_base <- import\_list(trend\_files, rbind = TRUE, fill = TRUE) %>%  
 clean\_names()

The Google Trends and College Scorecard data present ample information. The scorecard, released in September 2015, has highlighted and influenced student interest in institutions for favorable factors. Most notably post-graduate earning potential. Scorecard data in conjunction with associated Google search trends allows for explicit insights to be had after tidying, particularly if using a difference-in-differences approach.

As guided by the assignment guide, the trends data is prepared by parsing dates and establishing a floor, coalescing school names for simplicity, and standardizing the search index. Indices are standardized within each school × keyword (z-scores) to ensure 1 index\_std unit equals 1 SD relative to the group’s sample mean. For context, pre-standardization average school search interest stood at 47.16 with an SD of 21.50. On average, monthly search interest across institutions was 47% of the highest observed interest level.

# clean dates  
trends\_base <- trends\_base %>%  
 mutate(  
 date = ymd(str\_sub(monthorweek, 1, 10)),  
 month = floor\_date(date, unit = 'month')  
 ) %>%  
 dplyr::select(-any\_of('file'))  
  
# pull column names  
index\_col <- intersect(c('index', 'value', 'hits'), names(trends\_base)) |> first()  
keyword\_col <- intersect(c('keyword', 'search\_term', 'term'), names(trends\_base)) |> first()  
  
# coalesce columns for school names in trends to 'schname'  
trends\_base <- trends\_base %>%  
 mutate(  
 schname = coalesce(  
 !!!syms(intersect(c('schname','school\_name','college\_name','institution','name'), names(.))) # use splice operator  
 )  
 )  
  
# fail fast validation block  
if (is.null(index\_col)) stop('no index column value found')  
if (is.null(keyword\_col)) stop('no keyword column value found')  
if (!'schname' %in% names(trends\_base)) stop("could not make `schname` column in trends")  
  
trends\_ready <- trends\_base %>%   
 group\_by(schname, .data[[keyword\_col]]) %>%  
 mutate(  
 index\_std = (.data[[index\_col]] - mean(.data[[index\_col]], na.rm = TRUE)) /  
 sd(.data[[index\_col]], na.rm = TRUE)  
 ) %>%  
 ungroup()

Data is aggregrated for trialing different intervals, additionally a more robust check to evaluate keyword search trends independent of specific schools. trends\_school\_month, which allows clean joining of Scorecard & Google Trends data and difference-in-difference estimates with fixed effects for school and date, was selected for the purpose of this investigation. Keyword based aggregation cannot distinguish between pre/post policy date; day/week data porved too noisy for stable FE estimates.Keeping the monthly school-level trends data allows for easy distribution into pre & post treatment groups and additional conditional variables to be implemented later.

# schools × month  
trends\_school\_month <- trends\_ready %>%  
 group\_by(schname, month, .data[[keyword\_col]]) %>% # !!sym(keyword\_col)) %>%  
 summarise(index\_std = mean(index\_std, na.rm = TRUE), .groups = 'drop')  
  
# keywords × month ( all schools) || unutilized   
trends\_kw\_month <- trends\_ready %>%  
 group\_by(!!sym(keyword\_col), month) %>%  
 summarise(index\_std = mean(index\_std, na.rm = TRUE), .groups = 'drop')  
  
# school × week/day || unutilized   
trends\_school\_week <- trends\_ready %>%  
 group\_by(schname, date, !!sym(keyword\_col)) %>%  
 summarise(index\_std = mean(index\_std, na.rm = TRUE), .groups = 'drop')

Assembling the final data objects for exploration merges the cleaned and aggregated Google Trends data & College Scorecard data. To analyze school search trends and post graduate outcome beyond high and low earnings earnings, additional conditional variables were added for tuition cost and incurred debt. Some data exploration indicated strong colinearity (r = 0.70) between tuition cost and debt; debt was included in the final version as it is more akin structured to earnings as compared to cost in both data structure and conceptually as a defined value.

The final dataset serving as the source for future regressions is provided numerical indicators for both high-earnings. and high-debt. These conditional variables are imperative to building dataframes for useful regression models following column setup.

t\_join <- trends\_school\_month %>%  
 inner\_join(id\_link\_new, by = 'schname')  
  
# identify join keys existing in BOTH df's  
key\_cand <- c('unitid', 'opeid', 'opeid6')  
join\_key <- intersect(key\_cand, intersect(names(t\_join), names(scorecard)))  
  
if (length(join\_key) == 0) {  
 stop('No common join keys found to link to Scorecard (need one of: unitid, opeid, opeid6).')  
}  
  
# key cols must be the symmetrical || use character to avoid errors  
t\_join2 <- t\_join %>%  
 mutate(across(all\_of(join\_key), as.character))  
  
scorecard2 <- scorecard %>%  
 select(-any\_of(c('insturl','npcurl'))) %>%  
 mutate(across(all\_of(join\_key), as.character))  
  
final\_data <- t\_join2 %>%  
 inner\_join(scorecard2, by = join\_key) %>%  
 mutate(  
 cost\_num = readr::parse\_number(as.character(npt4\_pub\_average\_annual\_cost)),  
 debt\_num = readr::parse\_number(as.character(grad\_debt\_mdn\_supp)),  
 debt10yr\_num = readr::parse\_number(as.character(grad\_debt\_mdn10yr\_supp))  
 )

## Warning: There were 3 warnings in `mutate()`.  
## The first warning was:  
## ℹ In argument: `cost\_num =  
## readr::parse\_number(as.character(npt4\_pub\_average\_annual\_cost))`.  
## Caused by warning:  
## ! 193089 parsing failures.  
## row col expected actual  
## 39 -- a number NULL  
## 40 -- a number NULL  
## 41 -- a number NULL  
## 42 -- a number NULL  
## 43 -- a number NULL  
## ... ... ........ ......  
## See problems(...) for more details.  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 2 remaining warnings.

# write\_csv(final\_data, file.path(dir\_path, "trends\_scorecard\_monthly.csv"))

# define critical dates  
trend\_first\_date <- min(trends\_base$date, na.rm = TRUE)  
policy\_date <- as.Date('2015-09-12')  
  
# define earnings column  
earn\_col <- names(final\_data)[grepl('earn|wne|salary|income', names(final\_data), ignore.case = TRUE)][1] # PARE SEARCH TEARMS || only "md\_earn\_wne\_p10\_reported\_earnings" present  
stopifnot(!is.na(earn\_col))

# assign single earnings value per school || bypasses PrivacySuppressed to avoid error  
school\_earn <- final\_data %>%  
 distinct(schname, !!sym(earn\_col)) %>%  
 mutate(earn\_num = readr::parse\_number(!!sym(earn\_col))) %>% # MAY WANT TO CHANGE NAME FROM EARN NUM  
 filter(!is.na(earn\_num))

## Warning: There was 1 warning in `mutate()`.  
## ℹ In argument: `earn\_num =  
## readr::parse\_number(md\_earn\_wne\_p10\_reported\_earnings)`.  
## Caused by warning:  
## ! 268 parsing failures.  
## row col expected actual  
## 7 -- a number NULL   
## 29 -- a number PrivacySuppressed  
## 48 -- a number PrivacySuppressed  
## 49 -- a number NULL   
## 50 -- a number PrivacySuppressed  
## ... ... ........ .................  
## See problems(...) for more details.

# split into high/low earning schools || 75th percentile  
cut75 <- quantile(school\_earn$earn\_num, 0.75, na.rm = TRUE)  
  
# post-grad debt & schooling cost  
school\_cost\_debt <- final\_data %>%  
 distinct(schname, cost\_num, debt\_num, debt10yr\_num)

Performing feols() regression requires dataframe conversion; because a myriad of models were made for my analysis, I’ve retained 3 in my write-up with df and df2 being integral. The base dataframe df averages standardized search interest by school and month, then joins each school’s earnings and flags (relative to the policy\_month) and high\_earn (top-quartile earnings). df2 expands the dataset by merging the numeric cost & debt measures and creating high\_cost & high\_debt indicators using the same top-quartile rule. df\_did includes a DiD-ready set with a single did indicator for schools that are both post-policy and high-earning). These DF’s range from a simple baseline to ones containing interaction models with financial controls and still compatible with school and month fixed effects.

# add policy date and earnings vars to assist regressions  
df <- final\_data %>%  
 group\_by(schname, month) %>%  
 summarise(index\_std = mean(index\_std, na.rm = TRUE), .groups = 'drop') %>%  
 left\_join(school\_earn %>% select(schname, earn\_num), by = 'schname') %>% # join for earnings and index\_std  
 mutate(  
 post = month >= floor\_date(policy\_date, 'month'), # set policy date as floor  
 high\_earn = earn\_num >= cut75  
 )  
  
# include quantiles for cost and debt || strong colinear relationship  
df2 <- df %>%   
 left\_join(school\_cost\_debt, by = "schname") %>%  
 mutate(  
 high\_cost = cost\_num >= quantile(cost\_num, 0.75, na.rm = TRUE),  
 high\_debt = debt\_num >= quantile(debt\_num, 0.75, na.rm = TRUE)  
 )  
  
# difference-in-differences  
df\_did <- df2 %>%  
 mutate(  
 did = as.integer(post & high\_earn) # 1 if post == TRUE & high\_earn == TRUE  
 )

Key regressors are the interactions terms for post × high\_earn and post × high\_debt. Because they are time-invariant, their main effects are absorbed by the fixed effects of school; thus the identified effects are the post × ‘group’ terms. Because cost and debt are correlated (r ≈ 0.70) the joint models are interpreted with some reservation; similar mindfulness should be employed when interpreting the standard errors as they are also two-way clustered by school and month.

The initial, simple, regression model (reg1) isolates the change in standardized search interest for schools in the top 25% of graduate earnings following the College Scorecard release. The coefficient on the interaction term is positive indicating that, when controlling for school and month fixed effects, high-earning schools experienced about 0.075 SD (p = 0.06) more search interest post-policy. This suggests a policy-driven increase in student attention to higher-earning institutions.

Investigating debts influence on search interest, reg3\_dummy similarly isolates the change for the top 25% of high-debt schools. The coefficient value −0.0883 SD (p < 0.01) indicates a clearly negative and significant post-policy decline in attention for the highest-debt institutions when controlling for both fixed effects and the main effects of post-policy and debt status.

In reg\_did2 treatment is defined as ‘high-earn & post’ while controlling for affordability (tuition cost & high post-graduate debt). The previously established effect is again observable as both positive and significant 0.1620 SD (p < 0.05), nearly twice that of the baseline reg1. This is interpreted as the policy particularly increased the visibility of high-earning schools even after considering cost and debt as potential factors.

reg4 jointly examines interactions for both high-earning and high-debt classifications without the main effects for post-policy status; the model yields negative post effects for both groups—post × high\_earn −0.1055 SD & post × high\_debt −0.0684 SD (both p < 0.05). It’s implied that in the post-policy period, both high-earning and high-debt schools saw reduced search interest when these effects were estimated in the same model. The shift in sign for high-earnings compared to reg1 suggests that potential debt offsets the earnings effect when considered together.

The fully realized regression model (reg5) addresses the offset observed in reg4 by implementing terms for post, group dummies, and both interactions included. post × high\_earn is positively significant at 0.0847 SD and post × high\_debt is negatively significant at −0.0684 SD (both p < 0.05). The simplest interpretation is that the Scorecard release shifted attention toward high-earning schools and away from high-debt schools. The reg5 regression model proved optimal and anchors much of the analysis that follows.

reg1 <- feols(  
 index\_std ~ post:high\_earn | schname + month,  
 data = df,  
 vcov = ~ schname + month  
)

## NOTE: 10,844 observations removed because of NA values (LHS: 1,067, RHS: 10,844, Fixed-effects: 1,067).

## The variables 'postTRUE:high\_earnFALSE', 'postFALSE:high\_earnTRUE' and 'postTRUE:high\_earnTRUE' have been removed because of collinearity (see $collin.var).

# etable(reg1)  
  
reg3\_dummy <- feols(  
 index\_std ~ post + high\_debt + post:high\_debt | schname + month,  
 data = df2,  
 vcov = ~ schname + month  
)

## NOTE: 11,060 observations removed because of NA values (LHS: 1,067, RHS: 11,060, Fixed-effects: 1,067).

## The variables 'postTRUE' and 'high\_debtTRUE' have been removed because of collinearity (see $collin.var).

# etable(reg3\_dummy)  
  
reg\_did2 <- feols(  
 index\_std ~ did + high\_cost + high\_debt | schname + month,  
 data = df\_did,  
 vcov = ~ schname + month  
)

## NOTE: 82,437 observations removed because of NA values (LHS: 1,067, RHS: 82,256, Fixed-effects: 1,067).

## The variables 'high\_costTRUE' and 'high\_debtTRUE' have been removed because of collinearity (see $collin.var).

# etable(reg\_did2)  
  
reg4 <- feols( # THE OG OPTION  
 index\_std ~ post:high\_earn + high\_debt + post:high\_debt | schname + month,  
 data = df2,  
 vcov = ~ schname + month  
)

## NOTE: 17,404 observations removed because of NA values (LHS: 1,067, RHS: 17,404, Fixed-effects: 1,067).

## The variables 'high\_debtTRUE', 'postFALSE:high\_earnTRUE' and 'postTRUE:high\_earnTRUE' have been removed because of collinearity (see $collin.var).

# etable(reg4)  
  
reg5 <- feols( # THE GOLDEN OPTION  
 index\_std ~ post + high\_earn + post:high\_earn + high\_debt + post:high\_debt | schname + month,  
 data = df2,  
 vcov = ~ schname + month  
)

## NOTE: 17,404 observations removed because of NA values (LHS: 1,067, RHS: 17,404, Fixed-effects: 1,067).

## The variables 'postTRUE', 'high\_earnTRUE' and 'high\_debtTRUE' have been removed because of collinearity (see $collin.var).

# etable(reg5)  
  
# etable(reg1, reg3\_dummy, reg\_did2, reg4, reg5)

Being the most comprehensive regression models that adequately demonstrate the interaction effects, reg4 & reg5 were selected for plotting. reg4 isolates the post-policy × high-earn & post-policy × high-debt effects without main effects for post to highlight the independent interactions. reg5 augments its predecessor by including the main effects for post, high\_earn, and high\_debt to provide a more insights on policy impact in combination with financial considerations. Visualizations for these models could ease the assesment of how the inclusion of main effects changes the magnitude and direction of the interactions.

To properly convert the regression models for visualization, the broom library is implemented to quickly tidy and structure the model output. reg4\_tidy filters the model to retain only the key interaction terms and subsequently recode them for simpler labeling. reg5\_tidy filters for interaction terms with both post and financial classifications (high earn & debt).

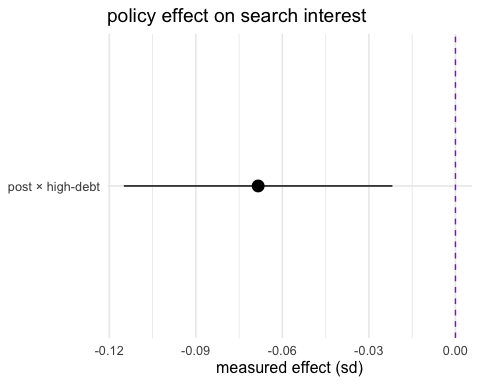
These steps towards visualization support the difference-in-differences framework by facilitating interpretation of how the Scorecard shifted search interests for high-earning and high-debt institutions, both independently and in combination with other financial factors.

# reg4  
reg4\_tidy <- tidy(reg4, conf.int = TRUE) # tidy the coeff's & SE's  
reg4\_tidy <- reg4\_tidy %>%  
 filter(term %in% c('postTRUE:high\_earnTRUE', 'high\_debtTRUE', 'postTRUE:high\_debtTRUE')) %>%  
 mutate(  
 term = recode(term, # streamline names  
 'postTRUE:high\_earnTRUE' = 'post × high-earn',  
 'high\_debtTRUE' = 'high-debt',  
 'postTRUE:high\_debtTRUE' = 'post × high-debt'  
 )  
 )  
  
# reg5  
reg5\_tidy <- tidy(reg5, conf.int = TRUE)  
reg5\_dids <- reg5\_tidy %>%  
 filter(grepl('post.\*:.\*high\_earn|high\_earn.\*:.\*post', term) |  
 grepl('post.\*:.\*high\_debt|high\_debt.\*:.\*post', term)) %>%  
 mutate(  
 label = case\_when(  
 grepl('high\_earn', term) ~ 'Post × High-earn (DiD)',  
 grepl('high\_debt', term) ~ 'Post × High-debt (DiD)',  
 TRUE ~ term  
 ),  
 label = factor(label, levels = c('Post × High-earn (DiD)',  
 'Post × High-debt (DiD)'))  
 )

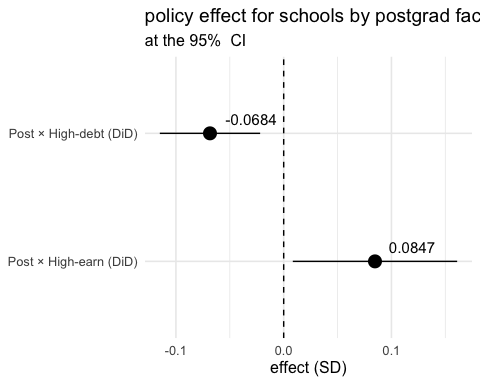
Generating visualizations continues to prove a challenge for me, however the two included provide a fair presentation of the observed effect. Plotting the reg4 model only produced the visualization for high debt institutions, despite the regression model being as previously described. The latter, reg5\_plot clearly portrays its regression model by displaying the effect of the scorecard on search interest in institutions with high earning and high debt post-graduate experiences.

Unfortunately, reg4\_plot only displays debt data, not earnings. While the plot is clear and interpretable, it lacks the utility to investigate the Scorecard’s effects on high-earning institutions. reg4\_plot ultimately is highly restricted its ability to portray how both earnings and debt shape student search interest.

# reg4  
# ONLY ASSESSES DEBT NOT EARNINGS  
reg4\_plot <- ggplot(reg4\_tidy, aes(x = term, y = estimate, ymin = conf.low, ymax = conf.high)) +  
 geom\_hline(yintercept = 0, linetype = 'dashed', color = 'purple') +  
 geom\_pointrange(size = 0.8) +  
 coord\_flip() +  
 labs(  
 title = 'policy effect on search interest',  
 x = NULL,  
 y = 'measured effect (sd)'  
 ) +  
 theme\_minimal(base\_size = 12)  
  
ggsave(  
 filename = file.path(dir\_path,'reg4\_plot.pdf'),  
 plot = reg4\_plot,  
 device = 'pdf',  
 width = 18, height = 8  
)  
  
show(reg4\_plot)



# reg5  
reg5\_plot <- ggplot(reg5\_dids, aes(x = label, y = estimate, ymin = conf.low, ymax = conf.high)) +  
 geom\_hline(yintercept = 0, linetype = 'dashed') +  
 geom\_pointrange(size = 0.9) +  
 geom\_text(aes(label = round(estimate, 4)),  
 hjust = -0.3, # horiz adj  
 vjust = -0.8, # vert adj  
 size = 4) +   
 coord\_flip() +  
 labs(  
 title = 'policy effect for schools by postgrad factors (reg5)',  
 subtitle = 'at the 95% CI',  
 x = NULL, y = 'effect (SD)'  
 ) +  
 theme\_minimal(base\_size = 12)  
  
ggsave(  
 filename = file.path(dir\_path,'reg5\_plot.pdf'),  
 plot = reg5\_plot,  
 device = 'pdf',  
 width = 18, height = 8  
)  
  
show(reg5\_plot)



# General Metrics & Summaries  
  
# summary(final\_data)  
  
# summary(school\_earn$earn\_num)  
  
# regression metric cluster  
etable(reg1, reg3\_dummy, reg\_did2, reg4, reg5)

## reg1 reg3\_dummy  
## Dependent Var.: index\_std index\_std  
##   
## postFALSE x high\_earnFALSE 0.0746. (0.0397)   
## postTRUE x high\_debtTRUE -0.0883\*\* (0.0251)  
## did   
## postTRUE x high\_earnFALSE   
## postTRUE x high\_earnTRUE   
## Fixed-Effects: ---------------- ------------------  
## schname Yes Yes  
## month Yes Yes  
## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
## S.E.: Clustered by: schn. & month by: schname & month  
## Observations 113,247 113,031  
## R2 0.34606 0.34626  
## Within R2 0.00073 0.00100  
##   
## reg\_did2 reg4  
## Dependent Var.: index\_std index\_std  
##   
## postFALSE x high\_earnFALSE -0.0207\* (0.0102)  
## postTRUE x high\_debtTRUE -0.0684\*\* (0.0230)  
## did 0.1620\* (0.0658)   
## postTRUE x high\_earnFALSE -0.1055\* (0.0454)  
## postTRUE x high\_earnTRUE   
## Fixed-Effects: ---------------- ------------------  
## schname Yes Yes  
## month Yes Yes  
## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
## S.E.: Clustered by: schn. & month by: schname & month  
## Observations 41,654 106,687  
## R2 0.44079 0.35253  
## Within R2 0.00316 0.00133  
##   
## reg5  
## Dependent Var.: index\_std  
##   
## postFALSE x high\_earnFALSE   
## postTRUE x high\_debtTRUE -0.0684\*\* (0.0230)  
## did   
## postTRUE x high\_earnFALSE   
## postTRUE x high\_earnTRUE 0.0847\* (0.0376)  
## Fixed-Effects: ------------------  
## schname Yes  
## month Yes  
## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
## S.E.: Clustered by: schname & month  
## Observations 106,687  
## R2 0.35253  
## Within R2 0.00133  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# pearson || colinearity   
# cor(final\_data$cost\_num, final\_data$debt\_num, use = "complete.obs")  
  
# mean & SD of index's [global assessments essentially useless at present]  
index\_mean <- mean(as.numeric(trends\_base$index), na.rm = TRUE)  
index\_sd <- sd(as.numeric(trends\_base$index), na.rm = TRUE)

There is a measurable and observable shift in student interest following the Scorecard’s release. High-earning institutions experienced a statistically significant increase in search interest, while high-debt institutions saw a corresponding decline. These effects persisted when controlling for notable financial factors, suggesting the policy has actively influenced prospective student behavior consistent with its intent. While the difference-in-differences approach can isolate the treatment effect, it still assumes parallel pre-policy trends and some apparent colinearity between earnings & debt classifications detracts from the findings. Despite the observed effects modesty, they are directionally consistent and statistically precise which are of greater importance at scale. However, the findings offer evidence that the Scorecard’s introduction reshaped the competitive landscape of higher education visibility, favoring institutions with stronger post-graduate financial outconmes.