data\_exploration\_econ

Katie Yamabe

## Libraries

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

LOAD the libraries and the data’s

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.3 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
✔ 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(fixest)  
library(rio)  
library(lubridate)

Load Google Trends Data

trends\_up\_to\_list <- list.files(pattern = "trends\_up\_to\_", full.names = TRUE)  
trends\_up\_to <-import\_list(trends\_up\_to\_list,rbind = TRUE, fill = TRUE)

Aggregating the Google Trends

trends\_up\_to <- trends\_up\_to %>%   
 mutate(temp = ymd(str\_sub(monthorweek,end = 10)))%>%   
 mutate(first\_of\_month = floor\_date(temp, unit = "month"))  
  
trends\_up\_to <- trends\_up\_to %>%   
 group\_by(schname, keyword)%>%   
 mutate(si = ((index - mean(index))/sd(index)),na.rm = TRUE)

Loading the Score Card

score <- import("Most+Recent+Cohorts+(Scorecard+Elements).csv")  
score <- score %>%   
 mutate(opeid = OPEID)%>%  
 filter(score$PREDDEG == 3)  
  
id\_name\_link <- import("id\_name\_link.csv")   
  
id\_name\_link <- id\_name\_link %>%   
 group\_by(schname) %>%   
 mutate (n = n())%>%   
 filter(n == 1) %>%  
 ungroup()

Combine the Google Trends and Score Card

id\_name\_link <- inner\_join(trends\_up\_to, id\_name\_link, by = "schname")  
   
clean\_data <- inner\_join(id\_name\_link, score, by = c("unitid" = "UNITID", "opeid" = "OPEID"))

Removing Na’s and determinnig that “high income” is the top 25% of the population. According to the 2022 US Census the top 25% income was $94,001 for the average household. <https://www.census.gov/content/dam/Census/library/publications/2023/demo/p60-279.pdf> (pg7).

clean\_data <- clean\_data %>%   
 mutate(md\_earn\_wne\_p10\_REPORTED\_EARNINGS = as.character(`md\_earn\_wne\_p10-REPORTED-EARNINGS`)) %>%  
 mutate(md\_earn\_wne\_p10\_REPORTED\_EARNINGS = na\_if(md\_earn\_wne\_p10\_REPORTED\_EARNINGS, "PrivacySuppressed")) %>%  
 mutate(md\_earn\_wne\_p10\_REPORTED\_EARNINGS = ifelse(md\_earn\_wne\_p10\_REPORTED\_EARNINGS %in% c("", "NULL"), NA, md\_earn\_wne\_p10\_REPORTED\_EARNINGS)) %>%  
 mutate(md\_earn\_wne\_p10\_REPORTED\_EARNINGS = as.numeric(md\_earn\_wne\_p10\_REPORTED\_EARNINGS))  
  
clean\_data <- clean\_data %>%  
 mutate(high\_earning = ifelse(!is.na(`md\_earn\_wne\_p10-REPORTED-EARNINGS`) & `md\_earn\_wne\_p10-REPORTED-EARNINGS` >= 94001, 1, 0))  
  
clean\_data <- clean\_data %>%  
 select(schname, first\_of\_month, si, PREDDEG, high\_earning, `md\_earn\_wne\_p10-REPORTED-EARNINGS`)

Adding missing grouping variables: `keyword`

clean\_data <- clean\_data %>%  
 mutate(post\_score = ifelse(first\_of\_month >= as.Date("2015-09-01"),1,0))  
  
#regression  
  
reg1 <- lm(si ~ high\_earning \* post\_score, data = clean\_data)  
summary(reg1)

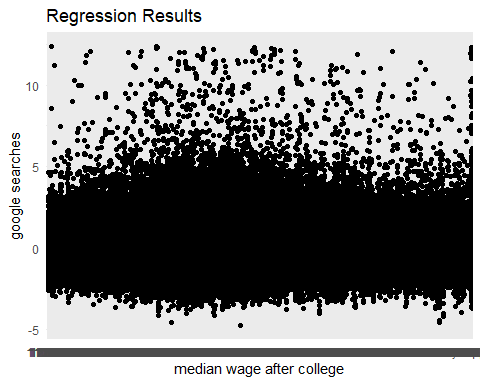
Call:  
lm(formula = si ~ high\_earning \* post\_score, data = clean\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.6184 -0.6619 -0.1094 0.5498 12.5012   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.033485 0.001281 26.13 <2e-16 \*\*\*  
high\_earning 0.053974 0.005379 10.03 <2e-16 \*\*\*  
post\_score -0.177040 0.002946 -60.09 <2e-16 \*\*\*  
high\_earning:post\_score -0.283064 0.012340 -22.94 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.9936 on 786253 degrees of freedom  
 (196629 observations deleted due to missingness)  
Multiple R-squared: 0.006426, Adjusted R-squared: 0.006422   
F-statistic: 1695 on 3 and 786253 DF, p-value: < 2.2e-16

# Create the plot  
ggplot(clean\_data, aes(x = `md\_earn\_wne\_p10-REPORTED-EARNINGS`, y = si)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(x = "median wage after college", y = "google searches", title = "Regression Results") +  
 theme\_minimal()

`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 196629 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 196629 rows containing missing values (`geom\_point()`).



Reason for Analysis

The analysis conducted was due to filtering down the information to the high-earning colleges as well as pre and post score card. These were key components of the research question. The regression that was ran was the standard index regressed on the high earning colleges after the scorecard came out.

Conclusion

The P-value is statistically significant to the 99%. The introduction fo the College Scorecard increased search activity on Google Trends for colleges with high-earning graduates by 10.03 percentage points relative to what it did for colleges with low-earning graduates with a standard error of 0.005379. This result comes from the high\_earning coefficients in my regression.

Conclusion