Data Exploration Project

## Research Question

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)?

## Libraries

library(rio)  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.1 ✔ 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(lubridate)  
library(fixest)  
library(dplyr)

## Importing Data

filelist <- list.files("Lab3\_RawData", pattern = 'trends', full.names = TRUE)  
  
my\_data <- import\_list(filelist, fill = TRUE, rbind = TRUE)  
  
head(my\_data)

schid schname keyword keynum  
1 0 young harris college young harris college 1  
2 0 young harris college young harris college 1  
3 0 young harris college young harris college 1  
4 0 young harris college young harris college 1  
5 0 young harris college young harris college 1  
6 0 young harris college young harris college 1  
 monthorweek index \_file  
1 2013-03-31 - 2013-04-06 34 Lab3\_RawData/trends\_up\_to\_finish.csv  
2 2013-04-07 - 2013-04-13 36 Lab3\_RawData/trends\_up\_to\_finish.csv  
3 2013-04-14 - 2013-04-20 45 Lab3\_RawData/trends\_up\_to\_finish.csv  
4 2013-04-21 - 2013-04-27 45 Lab3\_RawData/trends\_up\_to\_finish.csv  
5 2013-04-28 - 2013-05-04 100 Lab3\_RawData/trends\_up\_to\_finish.csv  
6 2013-05-05 - 2013-05-11 42 Lab3\_RawData/trends\_up\_to\_finish.csv

## Cleaning Data

cleaned\_data <- my\_data %>% mutate(new\_date\_variable = ymd(str\_sub(monthorweek, start = 0, end = 10)))  
  
head(cleaned\_data)

schid schname keyword keynum  
1 0 young harris college young harris college 1  
2 0 young harris college young harris college 1  
3 0 young harris college young harris college 1  
4 0 young harris college young harris college 1  
5 0 young harris college young harris college 1  
6 0 young harris college young harris college 1  
 monthorweek index \_file  
1 2013-03-31 - 2013-04-06 34 Lab3\_RawData/trends\_up\_to\_finish.csv  
2 2013-04-07 - 2013-04-13 36 Lab3\_RawData/trends\_up\_to\_finish.csv  
3 2013-04-14 - 2013-04-20 45 Lab3\_RawData/trends\_up\_to\_finish.csv  
4 2013-04-21 - 2013-04-27 45 Lab3\_RawData/trends\_up\_to\_finish.csv  
5 2013-04-28 - 2013-05-04 100 Lab3\_RawData/trends\_up\_to\_finish.csv  
6 2013-05-05 - 2013-05-11 42 Lab3\_RawData/trends\_up\_to\_finish.csv  
 new\_date\_variable  
1 2013-03-31  
2 2013-04-07  
3 2013-04-14  
4 2013-04-21  
5 2013-04-28  
6 2013-05-05

aggregated\_data <- cleaned\_data %>%  
 group\_by(schname, keyword) %>%  
 mutate(standardized\_index = (index - mean(index,na.rm = TRUE)) / sd(index, na.rm = TRUE))  
  
head(aggregated\_data)

# A tibble: 6 × 9  
# Groups: schname, keyword [1]  
 schid schname keyword keynum monthorweek index `\_file` new\_date\_variable  
 <chr> <chr> <chr> <int> <chr> <int> <chr> <date>   
1 0 young harris… young … 1 2013-03-31… 34 Lab3\_R… 2013-03-31   
2 0 young harris… young … 1 2013-04-07… 36 Lab3\_R… 2013-04-07   
3 0 young harris… young … 1 2013-04-14… 45 Lab3\_R… 2013-04-14   
4 0 young harris… young … 1 2013-04-21… 45 Lab3\_R… 2013-04-21   
5 0 young harris… young … 1 2013-04-28… 100 Lab3\_R… 2013-04-28   
6 0 young harris… young … 1 2013-05-05… 42 Lab3\_R… 2013-05-05   
# ℹ 1 more variable: standardized\_index <dbl>

### Reading in Scorecard Data

scorecard <- import('Lab3\_RawData\\Most+Recent+Cohorts+(Scorecard+Elements).csv')  
id\_name\_link <- import('Lab3\_RawData\\id\_name\_link.csv')  
  
colnames(scorecard)

[1] "UNITID"   
 [2] "OPEID"   
 [3] "opeid6"   
 [4] "INSTNM"   
 [5] "CITY"   
 [6] "STABBR"   
 [7] "INSTURL"   
 [8] "NPCURL"   
 [9] "HCM2"   
 [10] "PREDDEG"   
 [11] "CONTROL"   
 [12] "LOCALE"   
 [13] "HBCU"   
 [14] "PBI"   
 [15] "ANNHI"   
 [16] "TRIBAL"   
 [17] "AANAPII"   
 [18] "HSI"   
 [19] "NANTI"   
 [20] "MENONLY"   
 [21] "WOMENONLY"   
 [22] "RELAFFIL"   
 [23] "SATVR25"   
 [24] "SATVR75"   
 [25] "SATMT25"   
 [26] "SATMT75"   
 [27] "SATWR25"   
 [28] "SATWR75"   
 [29] "SATVRMID"   
 [30] "SATMTMID"   
 [31] "SATWRMID"   
 [32] "ACTCM25"   
 [33] "ACTCM75"   
 [34] "ACTEN25"   
 [35] "ACTEN75"   
 [36] "ACTMT25"   
 [37] "ACTMT75"   
 [38] "ACTWR25"   
 [39] "ACTWR75"   
 [40] "ACTCMMID"   
 [41] "ACTENMID"   
 [42] "ACTMTMID"   
 [43] "ACTWRMID"   
 [44] "SAT\_AVG"   
 [45] "SAT\_AVG\_ALL"   
 [46] "PCIP01"   
 [47] "PCIP03"   
 [48] "PCIP04"   
 [49] "PCIP05"   
 [50] "PCIP09"   
 [51] "PCIP10"   
 [52] "PCIP11"   
 [53] "PCIP12"   
 [54] "PCIP13"   
 [55] "PCIP14"   
 [56] "PCIP15"   
 [57] "PCIP16"   
 [58] "PCIP19"   
 [59] "PCIP22"   
 [60] "PCIP23"   
 [61] "PCIP24"   
 [62] "PCIP25"   
 [63] "PCIP26"   
 [64] "PCIP27"   
 [65] "PCIP29"   
 [66] "PCIP30"   
 [67] "PCIP31"   
 [68] "PCIP38"   
 [69] "PCIP39"   
 [70] "PCIP40"   
 [71] "PCIP41"   
 [72] "PCIP42"   
 [73] "PCIP43"   
 [74] "PCIP44"   
 [75] "PCIP45"   
 [76] "PCIP46"   
 [77] "PCIP47"   
 [78] "PCIP48"   
 [79] "PCIP49"   
 [80] "PCIP50"   
 [81] "PCIP51"   
 [82] "PCIP52"   
 [83] "PCIP54"   
 [84] "DISTANCEONLY"   
 [85] "UGDS"   
 [86] "UGDS\_WHITE"   
 [87] "UGDS\_BLACK"   
 [88] "UGDS\_HISP"   
 [89] "UGDS\_ASIAN"   
 [90] "UGDS\_AIAN"   
 [91] "UGDS\_NHPI"   
 [92] "UGDS\_2MOR"   
 [93] "UGDS\_NRA"   
 [94] "UGDS\_UNKN"   
 [95] "PPTUG\_EF"   
 [96] "CURROPER"   
 [97] "NPT4\_PUB-AVERAGE-ANNUAL-COST"   
 [98] "NPT4\_PRIV"   
 [99] "NPT41\_PUB"   
[100] "NPT42\_PUB"   
[101] "NPT43\_PUB"   
[102] "NPT44\_PUB"   
[103] "NPT45\_PUB"   
[104] "NPT41\_PRIV"   
[105] "NPT42\_PRIV"   
[106] "NPT43\_PRIV"   
[107] "NPT44\_PRIV"   
[108] "NPT45\_PRIV"   
[109] "PCTPELL"   
[110] "RET\_FT4"   
[111] "RET\_FTL4"   
[112] "RET\_PT4"   
[113] "RET\_PTL4"   
[114] "PCTFLOAN"   
[115] "UG25abv"   
[116] "GRAD\_DEBT\_MDN\_SUPP"   
[117] "GRAD\_DEBT\_MDN10YR\_SUPP"   
[118] "RPY\_3YR\_RT\_SUPP"   
[119] "C150\_4\_POOLED\_SUPP-REPORTED-GRAD-RATE"  
[120] "C200\_L4\_POOLED\_SUPP"   
[121] "md\_earn\_wne\_p10-REPORTED-EARNINGS"   
[122] "gt\_25k\_p6"

head(id\_name\_link)

unitid opeid schname  
1 180203 2517500 aaniiih nakoda college  
2 222178 353700 abilene christian university  
3 138558 154100 abraham baldwin agricultural college  
4 172866 2050300 academy college  
5 412173 3346300 academy for nursing and health occupations  
6 108232 753100 academy of art university

### Merging in Scorecard Data

id\_name\_link\_count <- id\_name\_link %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter(n == 1)  
  
combined\_data <- inner\_join(aggregated\_data, id\_name\_link\_count, by = c("schname"= "schname"))  
  
final\_data <- inner\_join(combined\_data, scorecard, by = c("unitid" = "UNITID"))  
  
head(final\_data)

# A tibble: 6 × 133  
# Groups: schname, keyword [1]  
 schid schname keyword keynum monthorweek index `\_file` new\_date\_variable  
 <chr> <chr> <chr> <int> <chr> <int> <chr> <date>   
1 0 young harris… young … 1 2013-03-31… 34 Lab3\_R… 2013-03-31   
2 0 young harris… young … 1 2013-04-07… 36 Lab3\_R… 2013-04-07   
3 0 young harris… young … 1 2013-04-14… 45 Lab3\_R… 2013-04-14   
4 0 young harris… young … 1 2013-04-21… 45 Lab3\_R… 2013-04-21   
5 0 young harris… young … 1 2013-04-28… 100 Lab3\_R… 2013-04-28   
6 0 young harris… young … 1 2013-05-05… 42 Lab3\_R… 2013-05-05   
# ℹ 125 more variables: standardized\_index <dbl>, unitid <int>, opeid <int>,  
# n <int>, OPEID <int>, opeid6 <int>, INSTNM <chr>, CITY <chr>, STABBR <chr>,  
# INSTURL <chr>, NPCURL <chr>, HCM2 <int>, PREDDEG <int>, CONTROL <int>,  
# LOCALE <chr>, HBCU <chr>, PBI <chr>, ANNHI <chr>, TRIBAL <chr>,  
# AANAPII <chr>, HSI <chr>, NANTI <chr>, MENONLY <chr>, WOMENONLY <chr>,  
# RELAFFIL <chr>, SATVR25 <chr>, SATVR75 <chr>, SATMT25 <chr>, SATMT75 <chr>,  
# SATWR25 <chr>, SATWR75 <chr>, SATVRMID <chr>, SATMTMID <chr>, …

### Limiting Data to Those Who Predominately Grant Bachelor’s Degrees

filtered\_final\_data <- final\_data %>%   
 filter(PREDDEG == 3) %>%  
 group\_by(schname, unitid, opeid, new\_date\_variable, `md\_earn\_wne\_p10-REPORTED-EARNINGS`) %>%  
 summarise(index = mean(index, na.rm = TRUE),standardized\_index = mean(standardized\_index, na.rm = TRUE))

`summarise()` has grouped output by 'schname', 'unitid', 'opeid',  
'new\_date\_variable'. You can override using the `.groups` argument.

### Mutating Reported Earnings as Numeric

filtered\_final\_data <- filtered\_final\_data %>% mutate(reported\_earnings = as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`))

Warning: There were 28842 warnings in `mutate()`.  
The first warning was:  
ℹ In argument: `reported\_earnings =  
 as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`)`.  
ℹ In group 315: `schname = "academy of couture art"`, `unitid = 475635`, `opeid  
 = 4185500`, `new\_date\_variable = 2013-03-24`.  
Caused by warning:  
! NAs introduced by coercion  
ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 28841 remaining warnings.

Here I mutated reported earnings to be a numeric value, this will be important for later when I need to differentiate between before and after the introduction of the Scorecard. It also ensures that I will not run into any possible errors in my regressions.

### Filtering Data

mean\_earnings <- mean(filtered\_final\_data$reported\_earnings, na.rm = TRUE)  
sd\_earnings <- sd(filtered\_final\_data$reported\_earnings, na.rm = TRUE)

filtered\_final\_data <- filtered\_final\_data %>% mutate(income\_level = ifelse(reported\_earnings >= mean\_earnings + sd\_earnings, 'High', 'Low'))

Here I define “high” earning colleges as any college whose reported graduate earnings were greater than or equal to one standard deviation above the mean. Therefore, “low” earning colleges would consist of all colleges with reported graduate earnings below that.

filtered\_final\_data <- filtered\_final\_data %>% filter(reported\_earnings < 40000 | reported\_earnings > 70000)

To ensure that my data contained actual high and low earning colleges as well as clean out the middle area of “average” earning colleges, I set a filter to cut off reported earnings to either below 40,000 or above 70,000. I choose this because the average salary for college graduates is around 50000-55000, however there is variation from state to state. Therefore I thought that 40,000 and 70,000 would be roughly equidistant from the mean salary.

filtered\_final\_data$month <- lubridate::month(filtered\_final\_data$new\_date\_variable)  
  
filtered\_final\_data <- filtered\_final\_data %>% mutate(after\_scorecard = ifelse(new\_date\_variable >= as.Date("2015-01-01"), 'After', 'Before'))

To be able to distinguish between before and after the College Scorecard implementation, I created a binary variable that indicates if the data point was measured before or after New Years of 2015, as the College Scorecard was introduced in 2015.

## Creating a Regression

The regression I will be employing is a Difference in Differences regression. This is because I believe that it is the best suited for this situation, as there is a clear before and after as well as two groups that could be considered treated and untreated. It will help us to answer the question of how the implementation of the Scorecard on the high earning graduate colleges.

model <- feols(index ~ after\_scorecard\*income\_level, data = filtered\_final\_data)

NOTE: 373 observations removed because of NA values (LHS: 373, RHS: 265).

etable(model)

model  
Dependent Var.: index  
   
Constant 47.47\*\*\* (0.2738)  
after\_scorecardBefore 4.636\*\*\* (0.3573)  
income\_levelLow -4.164\*\*\* (0.2816)  
after\_scorecardBefore x income\_levelLow -0.1607 (0.3675)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 137,150  
R2 0.02442  
Adj. R2 0.02440  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Firstly I will break down this regression and its coefficients. For High earning graduate colleges, 4.636 is the difference in search activity before and after the implementation of the College Scorecard. After the introduction of the College Scorecard, the difference in search activity between High and Low earning graduate colleges was -4.164. Finally, the coefficient of -0.1607 is how much larger the before and after difference between high and low earning graduate colleges. Therefore, for low earning graduate colleges, the search activity difference before and after the College Scorecard is 0.1607 less than for high earning graduate colleges.

Thus, the introduction of the College Scorecard decreased search activity on Google Trends for colleges with high-earning graduates by 0.1607 search queries relative to what it did for colleges with low-earning graduates, with a standard error of 0.3675. This result comes from the after\_scorecard x income\_level coefficient in my regression.

## Visualizations

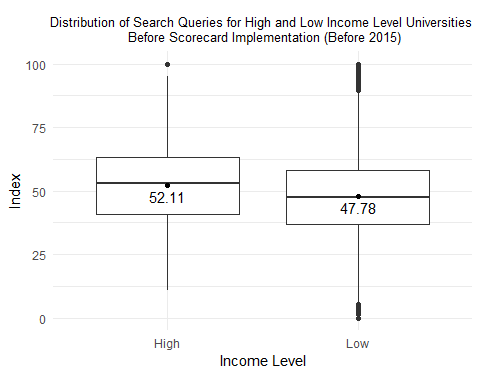
To better compare the effect of the introduction of the College Scorecard, I created two visualizations showing the distribution of search activity for High and Low earning graduates colleges for before and after the implementation.

filtered\_final\_data\_prescorecard <- filtered\_final\_data %>% filter(after\_scorecard == 'Before')  
  
filtered\_final\_data\_postscorecard <- filtered\_final\_data %>% filter(after\_scorecard == 'After')

ggplot(data = filtered\_final\_data\_prescorecard, aes(x = income\_level, y = index)) + geom\_boxplot() + theme\_minimal() + labs(x = "Income Level", y = "Index", title ="Distribution of Search Queries for High and Low Income Level Universities \n Before Scorecard Implementation (Before 2015)") + theme(aspect.ratio=1/1.5) + theme(plot.title = element\_text(size=10)) + theme(plot.title = element\_text(hjust = 0.5))+ stat\_summary(aes(y = index , label = round(..y.., 2)), fun.y = mean, geom = "text", position=position\_nudge(y=-4)) + stat\_summary(fun.y=mean, geom="point")

Warning: The `fun.y` argument of `stat\_summary()` is deprecated as of ggplot2 3.3.0.  
ℹ Please use the `fun` argument instead.

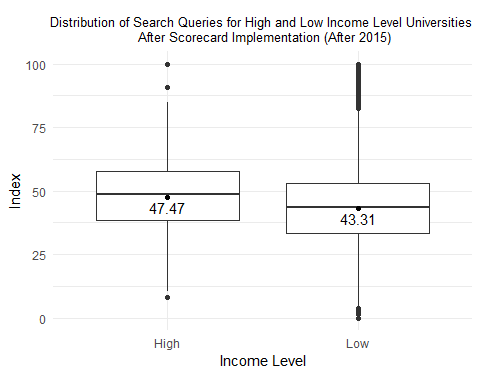
Warning: The dot-dot notation (`..y..`) was deprecated in ggplot2 3.4.0.  
ℹ Please use `after\_stat(y)` instead.



ggplot(data = filtered\_final\_data\_postscorecard, aes(x = income\_level, y = index)) + geom\_boxplot() + theme\_minimal() + labs(x = "Income Level", y = "Index", title ="Distribution of Search Queries for High and Low Income Level Universities \n After Scorecard Implementation (After 2015)") + theme(aspect.ratio=1/1.5) + theme(plot.title = element\_text(size=10)) + theme(plot.title = element\_text(hjust = 0.5)) + stat\_summary(aes(y = index , label = round(..y.., 2)), fun.y = mean, geom = "text", position=position\_nudge(y=-4)) + stat\_summary(fun.y=mean, geom="point")

Warning: Removed 108 rows containing non-finite values (`stat\_boxplot()`).

Warning: Removed 108 rows containing non-finite values (`stat\_summary()`).  
Removed 108 rows containing non-finite values (`stat\_summary()`).



What is interesting between these two visualizations is that from Before and After the implementation of the College Scorecard, search activity decreased across the board. In addition, what we can observe is that the difference in search activity between high and low income levels is lower after the introduction of the College Scorecard. Within the box and whisker plots, I have displayed the means with a point as well as a corresponding value. When taking the difference between the difference of high and low before and after the introduction of the College Scorecard, we find that the gap in search activity decreased by roughly 0.17. This corroborates my earlier findings through my regression, as the two plots demonstrate how the relative indexes decreased after the 2015 introduction of the College Scorecard, as well as how the gap in search activity between High and Low earning graduate colleges decreased after the change. It is important to note that generally, search activity seems to have been decreasing over the course of the data set.

Therefore, according to my regression and visualizations, the introduction of the College Scorecard decreased search activity for colleges with high-earning graduates and low-earning graduates alike. A possible explanation for this is that with the introduction of this resource, prospective students began to using the College Scorecard as a substitute for looking up key information on individual schools. Since the College Scorecard provides a centralized place for individuals to find information on colleges, many may not have the need to go outside of the website and continue their research.