Data Exploration Assignment

Ryan Bolen

## Libraries

library(tidyverse)

Warning: package 'tidyverse' was built under R version 4.3.2

── 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 1.9.2 ✔ 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)

Warning: package 'fixest' was built under R version 4.3.2

library(rio)

Warning: package 'rio' was built under R version 4.3.2

library(lubridate)  
library(ggplot2)  
library(scales)

Attaching package: 'scales'  
  
The following object is masked from 'package:fixest':  
  
 pvalue  
  
The following object is masked from 'package:purrr':  
  
 discard  
  
The following object is masked from 'package:readr':  
  
 col\_factor

**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)?

**Data:**

Data for this project was pulled from Google Trends which measures the number of internet searches made for a particular school and the College Scorecard where different variables related to enrollment, post graduate earnings, cost of attendance etc. are available for download and analysis.

# Getting all trends up to files in a list to pass to import/merge.  
  
trends\_up\_to\_list <- list.files(pattern = 'trends\_up\_to\_')

# Import all trends\_up\_to files and bind them into a data frame  
  
trends\_up\_to <- import\_list(trends\_up\_to\_list,rbind = TRUE, fill = TRUE)

**Standardizing search interest and joining datasets:**

Search volumes/interest for each school were transformed into standard deviations by subtracting out the mean for the ‘search index’ variable and then dividing each line by the standard deviation of the index variable. This calculation was performed for each observation and then summarized by totaling the standard deviation value for each school by name and month. Each month is reported as the first day of that given month.

# Get the first ten characters out of the month or week column which at the moment is a string.   
  
# Change the date string to an actual date and then floor it to the first day of the month.  
  
trends\_up\_to <- trends\_up\_to %>%   
 mutate(monthorweek2 = str\_sub(monthorweek,end = 10)) %>%   
 mutate(monthorweek2 = ymd(monthorweek2)) %>%   
 mutate(first\_day\_of\_month = floor\_date(monthorweek2,unit= 'month'))

# Standardize the index variable by school name  
  
trends\_up\_to <- trends\_up\_to %>% group\_by(schname,keyword) %>% mutate(std\_key = (index -mean(index,na.rm = TRUE))/sd(index,na.rm = TRUE))

# Summarize to school name and month level  
  
trends\_up\_to <- trends\_up\_to %>% group\_by(first\_day\_of\_month,schname) %>% summarize(standardized\_searches = sum(std\_key))

`summarise()` has grouped output by 'first\_day\_of\_month'. You can override  
using the `.groups` argument.

# Read in the scorecard data  
  
SC <- import('Most+Recent+Cohorts+(Scorecard+Elements).csv')  
  
# Read in the id name link data  
  
id\_name\_link <- import('id\_name\_link.csv')

# count how many times each school name pops up in id\_name\_link filter out any rows where n > 1  
  
id\_name\_link <- id\_name\_link %>% group\_by(schname) %>% mutate(n = n()) %>% filter(n == 1)

# Merge data sets - inner join trends up to (D2) on id\_name\_link that we summarised.  
  
trends\_up\_to <- trends\_up\_to %>% inner\_join(id\_name\_link, join\_by(schname))   
  
SC <- SC %>% mutate(opeid = OPEID)  
  
clean\_data <- trends\_up\_to %>% inner\_join(SC, join\_by(opeid))

Warning in inner\_join(., SC, join\_by(opeid)): Detected an unexpected many-to-many relationship between `x` and `y`.  
ℹ Row 99 of `x` matches multiple rows in `y`.  
ℹ Row 4914 of `y` matches multiple rows in `x`.  
ℹ If a many-to-many relationship is expected, set `relationship =  
 "many-to-many"` to silence this warning.

# Need to look at data before September 2015 (before card was released) and after September 2015 (after it was released)  
  
# Vars that might be good to control for:  
  
# Admission Rate 'ADM\_RATE'  
# Percentage of students who are financially independent 'DEP\_STAT\_PCT\_IND' - CANNNOT FIND  
# Dependent Variable of choice - CANNOT FIND  
# Mean earnings of students working and not enrolled 10 years after entry 'mn\_earn\_wne\_p10'  
# Graduation rate (on time is considered 6 years) - C150\_4\_POOLED\_SUPP-REPORTED-GRAD-RATE  
# Annual cost of going (average) - NPT4\_PUB-AVERAGE-ANNUAL-COST  
  
# a one-unit change in the standardized index can be understood and interpreted as a one-standard-deviation change in search interest

**High vs. low earnings bachelor’s degree granting schools:**

To investigate the research question specifically around the shift of interest to high-earnings colleges, I set a threshold for what I would consider a high earning college to be in this study. I declared that any college with earnings above the 75th percentile in earnings would be flagged as a high earnings school. Looking at the final data frame used for regression analysis, I used R to calculate the quartiles for median student earnings to determine that 48,100 is the 75th percentile value. I rounded the 75th percentile value up to 50,000, declaring that any school with median earnings 50,000 dollars or above is considered high earnings. The criteria of 50,000 was arbitrarily chosen in this case as a clean cutoff close to the 75th percentile threshold.

# Getting frame with only colleges that grant bachelor's degrees using PREDDEG variable equal to 3  
  
bach\_degree\_school\_data <- clean\_data  
  
bach\_degree\_school\_data <- clean\_data %>% filter(PREDDEG == 3)  
  
bach\_degree\_school\_data <- bach\_degree\_school\_data %>% select(first\_day\_of\_month,`md\_earn\_wne\_p10-REPORTED-EARNINGS`,schname,standardized\_searches,unitid,opeid,PREDDEG,`C150\_4\_POOLED\_SUPP-REPORTED-GRAD-RATE`,`NPT4\_PUB-AVERAGE-ANNUAL-COST`)  
  
bach\_degree\_school\_data <- bach\_degree\_school\_data %>% mutate(mnth = month(first\_day\_of\_month))  
  
# Filter to only get schools where med earnings is reported  
  
bach\_degree\_school\_data <- bach\_degree\_school\_data %>% filter(`md\_earn\_wne\_p10-REPORTED-EARNINGS` != 'PrivacySuppressed')  
  
bach\_degree\_school\_data <- bach\_degree\_school\_data %>% filter(`md\_earn\_wne\_p10-REPORTED-EARNINGS` != 'NULL')  
  
# Detirmining how the median earnings values are distributed.   
  
quantile(bach\_degree\_school\_data$earnings,type= 9)

Warning: Unknown or uninitialised column: `earnings`.

0% 25% 50% 75% 100%   
 NA NA NA NA NA

# Variable for high earnings and low earnings school - used 75K (might change to 75th percentile)  
  
median\_earnings <- 50000  
  
# Changing earnings and cost of attendance to numeric  
  
bach\_degree\_school\_data <- bach\_degree\_school\_data %>% mutate(earnings = as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`))  
  
  
bach\_degree\_school\_data <- bach\_degree\_school\_data %>% mutate(avg\_cost = as.numeric(`NPT4\_PUB-AVERAGE-ANNUAL-COST`))

Warning: There were 38 warnings in `mutate()`.  
The first warning was:  
ℹ In argument: `avg\_cost = as.numeric(`NPT4\_PUB-AVERAGE-ANNUAL-COST`)`.  
ℹ In group 1: `first\_day\_of\_month = 2013-03-01`.  
Caused by warning:  
! NAs introduced by coercion  
ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 37 remaining warnings.

# Creating a var to call out high and low earnings schools  
  
bach\_degree\_school\_data <- bach\_degree\_school\_data %>% mutate(high\_earnings\_school = earnings >= median\_earnings)

**Categorizing time periods:**

The final data set provides a variable to indicate whether the observation date is before or after the introduction of the college scorecard in 2015. If the first day of the month is greater than or equal to September 1st, 2015 then ‘after\_sc’ is True (or 1).

# Creating a var for before and after scorecard (first day of month greater or equal to september 2015)  
  
bach\_degree\_school\_data <- bach\_degree\_school\_data %>% mutate(after\_sc = ymd(first\_day\_of\_month) >= ymd('2015-09-01'))

**Interpretation based off regression:**

**Model1**

I started modeling using the ‘feols’ package in R with a simple design aimed at characterizing the general effect of the college scorecard on college interest in general. I used this formula to provide a baseline as to what effect the introduction of the college scorecard had overall and gain insight on what the search interest has historically been between high earnings and low earnings colleges:

’standardized\_searches = β0 + after\_sc β1 + high\_earnings β2

Based on this regression, I determined that the introduction of the college scorecard decreased search interest for all schools by approximately 2.6 standard deviations and that high earnings colleges have slightly less search interest relative to low earnings colleges by approximately .0041 standard deviations (independent of the scorecard introduction).

**Model2**

Capturing the change in search interest after the scorecard was introduced for high earnings schools relative to low earnings schools required extending the model to capture the interaction between the scorecard implementation and weather and whether a school is considered high earnings. I ran the following model to provide a comparison for the shift of interest between high and low schools after the college scorecard implementation date:

’standardized\_searches = after\_sc β1 + high\_earnings β2 + β3 high\_earnings \* after\_sc

The introduction of the College Scorecard decreased search activity on Google Trends for colleges with high-earning graduates by 3.213 standard deviations relative to what it did for colleges with low-earning graduates, with a standard error of 0.2015. This is summarized by the coefficient on the interaction term (β3) in the model where the number predicts a larger negative effect of the scorecard implementation on high earnings schools in the periods following the scorecard implementation. This is significant at the 95% level.

# simple model aimed at characterizing the effect of the college scorecard in general:  
model1 <- feols(standardized\_searches ~ after\_sc + high\_earnings\_school, data = bach\_degree\_school\_data)

NOTE: 754 observations removed because of NA values (LHS: 754, RHS: 573).

# model with interaction term to capture the change in interest of high earnings colleges that occurred relative to low earnings colleges after the scorecard was introduced.  
  
model2 <- feols(standardized\_searches ~ after\_sc + high\_earnings\_school + after\_sc \* high\_earnings\_school, data = bach\_degree\_school\_data)

NOTE: 754 observations removed because of NA values (LHS: 754, RHS: 573).

etable(model1)

model1  
Dependent Var.: standardized\_searches  
   
Constant 0.5097\*\*\* (0.0400)  
after\_scTRUE -2.658\*\*\* (0.0832)  
high\_earnings\_schoolTRUE -0.0041 (0.0791)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 65,652  
R2 0.01531  
Adj. R2 0.01528  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

etable(model2)

model2  
Dependent Var.: standardized\_searches  
   
Constant 0.3780\*\*\* (0.0408)  
after\_scTRUE -1.962\*\*\* (0.0938)  
high\_earnings\_schoolTRUE 0.6035\*\*\* (0.0877)  
after\_scTRUE x high\_earnings\_schoolTRUE -3.213\*\*\* (0.2015)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 65,652  
R2 0.01911  
Adj. R2 0.01906  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Conclusion and Discussion**

By plotting the average standardized search values over time and comparing high vs. low earnings schools, it’s clear to see that while search interest for all schools was dampened following the release of the college scorecard, it had a larger negative impact on high earnings institutions.

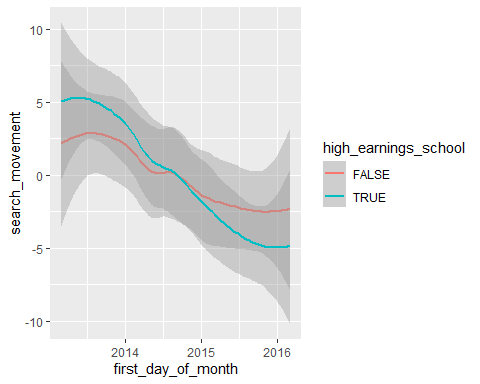
# Creating a data frame to visualize aggregated search trends on a monthly basis.   
  
monthly\_searches <- bach\_degree\_school\_data %>% group\_by(first\_day\_of\_month,high\_earnings\_school) %>% summarize(search\_movement = mean(standardized\_searches,na.rm = TRUE))

`summarise()` has grouped output by 'first\_day\_of\_month'. You can override  
using the `.groups` argument.

# Plotting search trends for high earnings vs. low earnings schools on a monthly basis.  
  
ggplot(data = monthly\_searches) +   
 geom\_smooth(mapping = aes(x = first\_day\_of\_month, y = search\_movement,color = high\_earnings\_school))

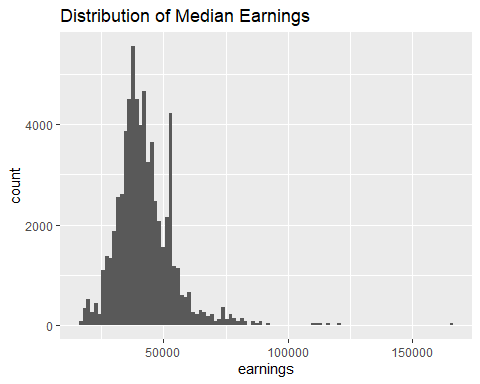
`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 2 rows containing non-finite values (`stat\_smooth()`).



Currently, most prospective students are likely to seek out lower schooling costs given the tremendously high and increasing tuition in the United States. The introduction of the college scorecard provided a way to make a greater number of comparisons to what makes a particular college appealing and allowed the population to compare similar characteristics for low cost and high-cost institutions. The improvement in transparency likely lowered interest geared towards higher cost universities (which correlate to universities with higher earnings) to universities with similar scorecard values but cheaper tuition costs. Otherwise, the majority of schools in this data set are on the lower end of median post-graduate earnings, and should contribute to a larger portion of searches in the first place.

# Distribution of earnings amongst all schools in data set.  
  
ggplot(data = bach\_degree\_school\_data, mapping = aes(x = earnings)) +   
 geom\_histogram(bins = 100)+  
 labs(title = "Distribution of Median Earnings")



It is likely that schools with higher median earnings among bachelor’s degree recipients require a higher cost of tuition because they offer degrees in more rigorous and lower admission programs. The scatter plot below demonstrates this relationship. For example, the Rosalind Franklin University of Medicine and Science is one of the institutions on the scorecard with the highest median earnings among students who received bachelor’s degrees and the average enrollment there is only approximately 2,000 students. While admission into competitive institutions seems desirable for students who want to pursue competitive degrees, most undergrads would likely opt for enrollment into a larger/cheaper school. The scorecard provided insights similar to the one below for students seeking more widely available degrees to make more informed decisions around what they would be paying for their education.

# Searches by month  
  
ggplot(data = bach\_degree\_school\_data, mapping = aes(x = earnings,y = avg\_cost)) +   
 geom\_point() +   
 labs(title = "Average Earnings vs. Average Cost") + xlim(0,120000)

Warning: Removed 47114 rows containing missing values (`geom\_point()`).

