# Data Exploration Project

Ben Carlson

5/6/2022

## Load In Libraries

```
library('dplyr')
library('readr')
library('tools')
library('purrr')
library('tidyverse')
library('lubridate')
library(ggplot2)
library(vtable)
library(fuzzyjoin)
```

## read in trends_up_to

```
trends <-  list.files(pattern = "^trends", full.names = TRUE) %>%  map(read_csv) %>%  bind_rows()
```

# Big 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)?

First, lets read in the Data,

### Now lets filter out colleges that don't predominantly grant bachelor's degrees.

```
`Most+Recent+Cohorts+(Scorecard+Elements)2` <-  `Most+Recent+Cohorts+(Scorecard+Elements)` %>% filter(PREDDEG == 3)
```

### Step two, filter colleges by high earning and low earning

I will use data from earnings after 10 years of entry for this

```
`Scorecard Data` <-  Most+Recent+Cohorts+(Scorecard+Elements)2` %>% select(UNITID, OPEID, opeid6, INSTNM, md_earn_wne_p10.REPORTED.EARNINGS)
```

Now we need to decide what is high earning and what is low earning, this is very subjective. Before anything will need to convert the string type first for the columns, since they aren't intergers.

```
ScorecardDataBad$md_earn_wne_p10.REPORTED.EARNINGS <- as.numeric(as.character(ScorecardDataBad$md_earn_wne_p10.REPORTED.EARNINGS))
```

```
## Warning: NAs introduced by coercion
```

Now lets clean up the NA values we created .

```
ScorecardData <- ScorecardData %>%  na.omit()
```

### Need to define high and low earning.

looking at a summary table, median is 40,700 for earnings after college. Since there isn't normal distribution I should cutoff per income #, not by top 100 or bottom 100 values. Lets say high earners must earn 25% more than the median, and low income must earn 25% less. 1.25 of 40,700 is 50,875, .75 is 30,525.

```
CollegesByEarnings <- ScorecardData %>% mutate(Earnings =case_when(
    md_earn_wne_p10.REPORTED.EARNINGS < 30525 ~ 'Low Earning',
    md_earn_wne_p10.REPORTED.EARNINGS > 50875 ~ 'High Earning'))
```

### Remove colleges that arent high earning or low earning

```
CollegesByEarnings <- CollegesByEarnings %>% na.omit()
```

Now We must do some relational database stuff, lets merge schname from id_name link on CollegesByEarnings.

### First we need to uncapitalize the CollegeEarnings column

```
CollegesByEarnings <- CollegesByEarnings %>% rename_with(tolower)
```

### Let's first read in the Data

```
id_name_link <- read.csv("C:/Users/Ben/Desktop/2022 Spring/ECON 4118 01/Econometrics Project/Data Exploration Project/id_name_link.csv")
```

Now merge to have the Scorecard data have the schname data in the google trends data

```
NewCollegeEarnings <-  merge(CollegesByEarnings, id_name_link, by= c('unitid','opeid'))
```

### Select the Columns want

```
NewCollegeEarnings <- NewCollegeEarnings %>% select(schname,earnings,md_earn_wne_p10.reported.earnings)
```

### Lets rename the column

```
colnames(NewCollegeEarnings)[colnames(NewCollegeEarnings) == 'md_earn_wne_p10.reported.earnings'] <- 'medianearnings10'

colnames(NewCollegeEarnings)[colnames(NewCollegeEarnings) == 'earnings'] <- 'earningsbracket'
```

### Now merge google trends with scorecard data

### First lets remove white space from both to allow the merge

```
trimws(NewCollegeEarnings$schname, which = c("both"))
```

### Now lets try to merge

```
Earningtrends <- merge(NewCollegeEarnings, trends, by= 'schname')
```

### Awesome it worked! Now lets get rid of colleges that arent high earning or low earning.

```
Earningtrendsclean <- Earningtrends %>% filter(earningsbracket == c('Low Earning','High Earning'))
```

### Now we lets add a column for data before and after the scorecard was added. Scorecard as added on September 12th 2015

```
Earningtrendsclean2 <- Earningtrendsclean %>% mutate(ScorecardRelease = case_when(
    monthorweek <= '2015-09-06 - 2015-09-12' ~ 'Before Scorecard',
    monthorweek >= '2015-09-13 - 2015-09-19' ~ 'After Scorecard'))
```

This worked even though I did not convert it into dates, I'm not complaining but I am curious why.

### Remove the Nulls

```
Earningtrendsclean2 <- Earningtrendsclean2 %>% na.omit()
```

### Now we can standardize the index for the keywords

```
Earningtrendsclean2 <- Earningtrendsclean2 %>% group_by(schname,schid) %>% mutate(std_index = (index-mean(index))/(sd(index)))
```

### Select the columns we want for the analysis

```
Earningtrendsclean3 <- Earningtrendsclean2 %>% select(schname, schid, ScorecardRelease, earningsbracket, index, std_index, monthorweek)
```

### Group by to get the average standardized index for each school, before and after the scorecard release

```
Earningtrendsclean_final <- Earningtrendsclean3 %>% group_by(schname,ScorecardRelease, earningsbracket) %>% summarise(avg_std_index = mean(std_index))
```

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

## Time for Regression!

### Because our variables are always in relation to the other variable, each of them have to interact.

```
etable(feols(avg_std_index ~ ScorecardRelease:earningsbracket, data = Earningtrendsclean_final))
```

```
## The variable 'ScorecardReleaseBefore Scorecard:earningsbracketLow Earning' has been removed because of collinearity (see $collin.var).
```

```
##                                                          feols(avg_std_ind..
## Dependent Var.:                                               avg_std_index
##
## (Intercept)                                                  0.0455* (0.0224)
## ScorecardReleaseAfterScorecard x earningsbracketHighEarning   -0.3595*** (0.0277)
## ScorecardReleaseBeforeScorecard x earningsbracketHighEarning    0.0219 (0.0277)
## ScorecardReleaseAfterScorecard x earningsbracketLowEarning     -0.2660*** (0.0277)
## ____________________________________________________________ _____________________
## S.E. type                                                                      IID
## Observations                                                                 1,108
## R2                                                                         0.24011
## Adj. R2                                                                    0.23805
## ____
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since we only have two options for x in the regression, let's try having y only have two options aswell to see if it's the same.

### First lets make a new average of our index with only two options.

```
regression <- Earningtrendsclean_final %>% group_by(ScorecardRelease, earningsbracket) %>% summarise(avg_std_index = mean(avg_std_index))
```

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

### Now lets run it.

```
etable(feols(avg_std_index ~ ScorecardRelease:earningsbracket, data = regression))
```

```
## The variable 'ScorecardReleaseBefore Scorecard:earningsbracketLow Earning' has been removed because of collinearity (see $collin.var).
```

```
##                                                          feols(avg_std_i...
## Dependent Var.:                                               avg_std_index
##
## (Intercept)                                                      0.0455 (NaN)
## ScorecardReleaseAfterScorecard x earningsbracketHighEarning     -0.3595 (NaN)
## ScorecardReleaseBeforeScorecard x earningsbracketHighEarning     0.0219 (NaN)
## ScorecardReleaseAfterScorecard x earningsbracketLowEarning      -0.2660 (NaN)
## ____________________________________________________________ _____________
## S.E. type                                                             IID
## Observations                                                            4
## R2                                                                      1
## Adj. R2                                                               NaN
## ____                                                                  Inf
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This proves that both regressions are exactly the same, with the ladder one making it much cleaner.

# Graphing time!

I want the graph to include show trends based off month so lets use the Earningtrendsclean2 dataset

### Lets clean up dataset to make the graph more readable

```
earningsgraph <- Earningtrendsclean2 %>% mutate(YearM = str_sub(monthorweek,1,7)) %>% group_by(YearM, earningsbracket) %>% summarise(avg_std_index=mean(std_index))
```

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

### Lets make the graph with ggplot

```
ggplot(earningsgraph, aes(x = YearM, y = avg_std_index, color = earningsbracket, group = earningsbracket), size = 1.5) +
  geom_line() + geom_point() + geom_vline(xintercept = 31) + geom_text(aes(x=32), label="After Scorecard", y=1, color="blue", angle=0, vjust = 0, hjust = .05, text=element_text(size=15)) + # add vertical line, with label
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1, size = 11)) + ggtitle('Monthly Search Trend s') +labs(x = "Month-Year",y = "Average Standardized Monthly Index ", color = 'Legend')
```

```
## Warning: Ignoring unknown parameters: text
```


```
# rotate labels to stop overlapping
```

## Write up

The Question I was tasked to answer was if the introduction of the College scorecard shifted interest to higher earning colleges compared to low earning ones. Specifically for colleges that predominantly grant bachelor's degrees.

First I had to decide what colleges were high or low earning. To do this I calculated the median of the median earnings after 10 years variable for each college. I used median as I wanted to reduce outliers. Based off the median I concluded that a school earning 25% more than the median was a high earning school, while 25% less was a low earning school. Deciding this was difficult as it is extremely subjective however there were three core reasons why I decided on this measure:

1. I did not want to be too restrictive as I did not want to lose statistical significance.
2. 25% is a common and clean way of division in our society, especially in regards to finance and segmenting.
3. 25% is still a high enough amount where many people would consider that in a higher bracket.

After that, I labeled each school accordingly and removed the schools not in either category. Then for each search, I categorized them either to be after to before the Scorecard release. The scorecard was released on September 12th, 2015. The date segments in the data happened to end on September 12th, so I made the following days the 'After Scorecard/Release' value. However now the issue was that the indexes cannot be compared as they are on different scales, so I standardized each keyword/school by schname to make every index on the same scale. This way I can compare how interest changes for each school.

Now to prepare my regression, I created a data set with every school, that school's Earnings bracket, and its average standardized index before the scorecard, and after the scorecard, grouped by schname. This way I can broadly see through the regression the impact that being after the scorecard had on schools depending on the earnings bracket. Running the regression I needed all coefficients to be interactions as it isn't possible in my data set for a school to not be in an earnings bracket or not be before or after a scorecard, which is why I used the ':' symbol in my regression.

Now interpreting my regression. We have a dummy variable model as both of the dependent variables are categorical data, with our omitted variable being **'ScorecardReleaseBeforeScorecard x earningsbracketLowEarning'**. Interpreting the coefficients we see that Scorecard Release actually lowered the Standardized index regardless of if the school was high or low earning. In fact, it actually lowered more for high-earning schools. If you were a high earning school, after Scorecard release index on average drops .3595 standard deviations below the mean, while it only drops .2660 for low-earning schools. Before scorecard release regardless of the earnings bracket, you were higher than the mean, High Earning was .0219 standard deviations higher, while Low Earning was .0455 standard deviations higher (it being the omitted variable it will just match the intercept).

Now this Regressions is a bit unusual as if it were to be graphed there are only possible x values you could have, either being before or after Scorecard release. this means that all y values can only be on these values, which would create a pattern of two straight lines of data points. Because of this we know that our regression is linear as in reality all our data points can be averaged down to just four observations, with two separate straight lines. To prove this I ran another regression where I got the average of the dependent variable for each independent variable combination. Running it, all the coefficients are exactly same as the first model, however with the caveat that there are no standard errors and there is a perfect R2 value.

Now the regression is telling us that there is a correlation between colleges losing interest regardless of earning, after the scorecard was released, with the values being statistically significant as the p-value is less than .001. However, is this correlation actually causation?

To figure this out I created a graph, which mapped high-earning and low earning schools' average monthly standardized index. high-earning schools were in blue, low-earning in red. The Y-axis was the Average Standardized Monthly Index, and the X-axis was the Month and Year of the observation. I also created a straight horizontal line going through 09-2015, the month that Scorecard was released.

Now observing the graph it is clear to see there is a clear negative relationship between the Month-Year and the Standardized Monthly Index (Interest). As time has progressed the interest in searching for schools had gone down, regardless of the Scorecard. Looking at the line when the Scorecard was released, there isn't any noticable change in pattern compared to the rest of the graph before relative, which makes me think that the Scorecard didn't have any impact at all in regards to searching for schools. Maybe one could infer that the release of the scorecard created an initial boost in college search interest as that month was the highest for the year. However, looking at every other year we see a similar pattern and peak for every September. What is also interesting that the graph shows us is that overall the interest in high earning schools and low earning schools is about the same every month, making both lines fairly identical.

Now to summarize, looking over my regression and chart, among colleges that predominantly grant bachelor's degrees, the release of the College Scorecard did not shift student interest to high-earning colleges relative to low-earning ones. We know this as the regression model showed reduced interest for both categories of colleges after Scorecard release. In fact, it showed a higher reduction in interest for high-earning colleges. These regression interpretations are backed-up by my chart, which shows that the introduction of the college scorecard did not change the overall pattern of the graph, and that regardless of the scorecard release, search interest in colleges was been steadily falling since 2013.