DataExploration

# Libraries Used:

library(rio)  
library(lubridate)

Attaching package: 'lubridate'

The following objects are masked from 'package:base':  
  
 date, intersect, setdiff, union

library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(stringr)  
library(fixest)  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ readr 2.1.4  
✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
✔ purrr 1.0.2 ✔ tidyr 1.3.0

── 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

labfiles = "Lab3\_Rawdata"  
df = list.files(path = labfiles, pattern = "trends\_up\_to\_", full.names = TRUE)  
data = import\_list(df, rbind = TRUE)

Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
Stopped early on line 1562. Expected 6 fields but found 5. Consider fill=TRUE  
and comment.char=. First discarded non-empty line: <<11,yti career institute -  
york,yti.edu,2,>>

Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
Stopped early on line 1095. Expected 6 fields but found 5. Consider fill=TRUE  
and comment.char=. First discarded non-empty line: <<9,heidelberg  
university,heidelberg.edu,2,>>

Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
Stopped early on line 1094. Expected 6 fields but found 5. Consider fill=TRUE  
and comment.char=. First discarded non-empty line: <<8,mount vernon nazarene  
university,mvnu.edu,2,>>

Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
Stopped early on line 3280. Expected 6 fields but found 5. Consider fill=TRUE  
and comment.char=. First discarded non-empty line: <<41,potomac state college  
of west virginia university,potomac state college of west virginia  
university,1,>>

data = data %>%  
 mutate(date\_string = str\_sub(monthorweek, 1, 10))  
data = data %>%  
 mutate(date = ymd(date\_string))  
data = data %>%  
 mutate(month = floor\_date(date, "month"))  
data = data %>%   
 group\_by(schname, keyword) %>%  
 mutate(std\_index = scale(index), na.rm =TRUE)

id\_name\_link = import(file = "Lab3\_Rawdata/id\_name\_link.csv")  
scorecard = import(file = "Lab3\_Rawdata/Most+Recent+Cohorts+(Scorecard+Elements).csv")  
  
id\_name\_link = id\_name\_link %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter (n ==1)  
  
trends\_id = inner\_join(data, id\_name\_link, by = NULL, copy = TRUE)

Joining with `by = join\_by(schname)`

trends\_id\_scorecard = inner\_join(trends\_id, scorecard, by = c("unitid" = "UNITID"))  
#Data cleaning complete

### Data Cleaning Complete

trends\_id\_scorecard = trends\_id\_scorecard %>%  
 filter(PREDDEG == 3) %>%  
 mutate(reported\_earnings = as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`)) %>%  
 drop\_na(reported\_earnings)

Warning: There were 351 warnings in `mutate()`.  
The first warning was:  
ℹ In argument: `reported\_earnings =  
 as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`)`.  
ℹ In group 7: `schname = "academy of couture art"` and `keyword = "academy of  
 couture art"`.  
Caused by warning:  
! NAs introduced by coercion  
ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 350 remaining warnings.

avg\_earnings = mean(trends\_id\_scorecard$reported\_earnings)  
sd\_earnings = sd(trends\_id\_scorecard$reported\_earnings)  
  
trends\_id\_scorecard = trends\_id\_scorecard %>%  
 mutate(low\_earnings = `reported\_earnings` <= avg\_earnings - 2\*sd\_earnings) %>%  
 mutate(high\_earnings = `reported\_earnings` >= avg\_earnings + 2\*sd\_earnings) %>%  
 mutate(earnings = case\_when(low\_earnings == TRUE ~ "Low Earnings", high\_earnings == TRUE ~ "High Earnings", TRUE ~ "Average Earnings"))

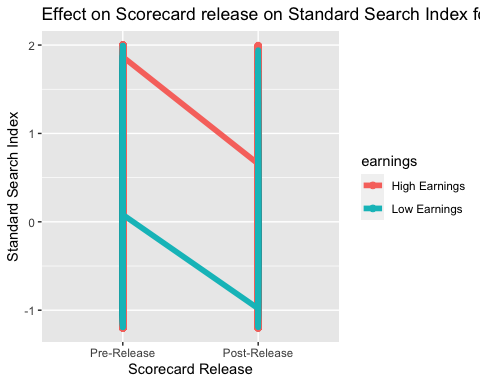
trends\_id\_scorecard = trends\_id\_scorecard %>%  
 mutate(pre\_release = date < '2015-09-01') %>%  
 mutate(post\_release = date >= '2015-09-01') %>%  
 mutate(release = case\_when(pre\_release == TRUE ~ "Pre-Release", post\_release == TRUE ~ "Post-Release"))

m = feols(std\_index~high\_earnings+post\_release +high\_earnings\*post\_release, data=trends\_id\_scorecard, vcov = 'hetero')

etable(m)

m  
Dependent Var.: std\_index  
   
Constant 0.0375\*\*\* (0.0013)  
high\_earningsTRUE -0.0054 (0.0065)  
post\_releaseTRUE -0.2003\*\*\* (0.0031)  
high\_earningsTRUE x post\_releaseTRUE 0.0293. (0.0150)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob.  
Observations 730,954  
R2 0.00608  
Adj. R2 0.00608  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

trends\_id\_scorecard <- trends\_id\_scorecard %>%  
 filter(earnings != "Average Earnings")  
ggplot(trends\_id\_scorecard) + aes(x = release, y = std\_index, group = earnings, color = earnings) + geom\_point(size = 2) + geom\_line(linewidth = 2) + scale\_x\_discrete(limit = c("Pre-Release", "Post-Release")) + scale\_y\_continuous(limits = c(-1.2, 2)) + labs(title = "Effect on Scorecard release on Standard Search Index for High & Low Earnings Colleges", x = ("Scorecard Release"), y =("Standard Search Index"))



# Writeup:

Looking into the original research question of how did the release of the scorecard (start of September 2015) shift the student interest in high earning colleges relative to low earnings colleges (that predominantly grant bachelor’s degrees). I first elected to clean my data further by removing from the data set colleges that didn’t predominately grant bachelor’s degrees. I did this by looking further into the data dictionary for the college scorecard (what the values mean) and found the column PREDDEG which showed the primary degree offered from the university. I used this to filter out colleges that didn’t have a PREDDEG of 3 (didn’t predominately grant bachelors degrees). I then looked at the search index for all the colleges and standardized them in order to get a change by one unit to the standardized index being a one standard deviation change in search interest (before they were only comparable to themselves). I did this by using the scale() function instead of the recommended way of calculating it with mean() and sd() just as I found it easier.

I then determined low and high earnings by looking at the reported earnings and removing the non numeric values (Privacy protected fields). I then found the mean of all the reported earnings and the standard deviation and then determined high earnings to be 2 standard deviations greater than the mean and low earnings to be 2 standard deviations less than the mean. I argue this is a good measure of high and low earnings as given a normal distribution we can expect to find that 95% of the population will fall in the average earnings category and then the top 2.5% and bottom 2.5% will be our population that lies in the high earnings and low earnings category. While I think one standard deviation was probably sufficient I think having more distinct groups will allow me to more easily see the effects of high and low earnings (cut through the noise). I was able to do this by setting columns that told me if the college was TRUE or FALSE for high and low earnings and then created a new column by using the case\_when() to display the proper earnings for each college.

For my regression I regressed the standardized index on high earnings and found in the analysis the introduction of the college scorecard increased the search index on high earning colleges by 5.24% of a standard deviation with a standard deviation of 0.00384. Overall, from this I found that the search index for both high earning colleges and low earning colleges fell slightly to their pre scorecard release.

In conclusion the introduction of the college scorecard increased search activity on Google Trends for colleges with high earnings graduates by 5.24% of a standard deviation relative to what it did for colleges with low earning graduates with a standard error of 0.0384. This shift suggests there was a small impact of the scorecard on increased search activity towards colleges with high earning graduates. Looking at my graph for the standardized index on scorecard release I found that both high earning colleges and low earning colleges decreased in overall search interest however, high earning colleges fell by a lower rate.