Data Exploration Project

June L

2024-02-18

## R Markdown

**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)?

**INTRODUCTION:**

The College Scorecard is an initiative by the U.S. Department of Education with a centralized location that provides insightful data and information on colleges and universities across the United States. The initiative aimed to help prospective postsecondary students make informed enrollment decisions. As a public-facing website, the College Scorecard contains essential information about colleges, including how much graduates earn. For the interest of the analysis carried out throughout this report, College Scorecard will serve as designated as a data source layered with Google Trends data to seek the answer to the research question as to whether the release of College Scorecard influenced student interest in colleges based on their earnings potential.

**DATA PREP AND STANDARDIZATION :**

For this analysis, R was utilized as a scripting language to conduct all the necessary data prep steps, including data cleaning and data merging. One important call out of the Google Trend data is that it provides the popularity of a given search term in Google over time. However, provided indices that represent the relative popularity of search terms are not directly comparable between search terms as all search terms are based on different scales. An equal amount of keyword search terms does not necessarily have the same impact from one school to another. To make the indices comparable, standardization was carried out by subtracting the mean of the index and then dividing the result by the standard deviation of the index. To ensure consistency and reduce potential variability in the regression analysis, all universities with duplicate names would be excluded from the dataset.

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(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ readr 2.1.5  
✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
✔ purrr 1.0.2 ✔ tidyr 1.3.1

── 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(fixest)  
library(vtable)

Loading required package: kableExtra  
  
Attaching package: 'kableExtra'  
  
The following object is masked from 'package:dplyr':  
  
 group\_rows

library(lmtest)

Loading required package: zoo  
  
Attaching package: 'zoo'  
  
The following objects are masked from 'package:base':  
  
 as.Date, as.Date.numeric

library(olsrr)

Attaching package: 'olsrr'  
  
The following object is masked from 'package:datasets':  
  
 rivers

knitr::opts\_knit$set(root.dir = "C:/Users/junes/Downloads")

file\_directory <- "../Downloads/Data\_Exploration\_Rawdata/Lab3\_Rawdata"  
  
# Use list.files() to get a vector of filenames  
filenames3 <- list.files(path = file\_directory, pattern = "trends\_up\_to\_", full.names = TRUE)  
  
list.files(file\_directory)

[1] "CollegeScorecardDataDictionary-09-08-2015.csv"  
 [2] "id\_name\_link.csv"   
 [3] "Most+Recent+Cohorts+(Scorecard+Elements).csv"   
 [4] "trends\_up\_to\_finish.csv"   
 [5] "trends\_up\_to\_inter\_1.csv"   
 [6] "trends\_up\_to\_inter\_2.csv"   
 [7] "trends\_up\_to\_inter\_3.csv"   
 [8] "trends\_up\_to\_inter\_4.csv"   
 [9] "trends\_up\_to\_inter\_5.csv"   
[10] "trends\_up\_to\_inter\_6.csv"   
[11] "trends\_up\_to\_UM.csv"   
[12] "trends\_up\_to\_UPhoenix.csv"   
[13] "trends\_up\_to\_UT.csv"   
[14] "trends\_up\_to\_UTMB.csv"   
[15] "trends\_up\_to\_Yorktowne.csv"

df <- import\_list(filenames3, rbind = TRUE, fill = TRUE)  
  
  
#Aggregating the google trend data with date modification  
df$first\_ten <- str\_sub(df$monthorweek, start = 1, end = 10)  
df$month\_yr <- ymd(df$first\_ten)  
df$month <- floor\_date(df$month\_yr, unit = "month")  
  
#Data modification with group\_by + mutate  
  
df <- df %>%   
 group\_by(schname, keyword) %>%   
 mutate(standarized\_index = (index - mean(index)) / sd(index))  
  
#Aggregate Keyword-month Level  
   
keyword\_month <- df %>%   
 group\_by(schname, keyword) %>%   
 summarize(mean\_stand\_ind = mean(standarized\_index))

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

#Aggregate schoolname, month level  
  
schname\_month <- df %>%   
 group\_by(schname, month) %>%   
 summarize(mean\_stand\_ind = mean(standarized\_index))

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

#Aggregate schoolname, level  
  
schname\_month <- df %>%   
 group\_by(schname, month) %>%   
 summarize(mean\_stand\_ind = mean(standarized\_index))

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

# Read in the Scorecard data  
scorecard\_data <- import(paste0(file\_directory, "/Most+Recent+Cohorts+(Scorecard+Elements).csv"))  
  
# Read in the id\_name\_link file  
id\_name\_link <- import(paste0(file\_directory, "/id\_name\_link.csv"))  
  
#Count of school name in id\_name\_link to only get unique  
  
id\_name\_link <- id\_name\_link %>%   
 group\_by(schname) %>%   
 mutate(n=n()) %>%   
 filter(n==1)  
  
#Inner Join the data  
  
scorecard\_data <- rename(scorecard\_data, opeid = OPEID)  
  
  
final\_data\_join <- inner\_join(df,id\_name\_link, by = 'schname')  
final\_data\_join <- inner\_join(final\_data\_join,scorecard\_data, by = 'opeid')

Warning in inner\_join(final\_data\_join, scorecard\_data, by = "opeid"): Detected an unexpected many-to-many relationship between `x` and `y`.  
ℹ Row 47229 of `x` matches multiple rows in `y`.  
ℹ Row 1073 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.

# Group by schname, count distinct schid values, and filter where count of distinct schid values is greater than 1  
inconsistent\_schname <- final\_data\_join %>%  
 group\_by(schname) %>%  
 summarize(distinct\_schid\_count = n\_distinct(schid)) %>%  
 filter(distinct\_schid\_count > 1)  
  
# Print rows where schname is the same while schid is different  
print(inconsistent\_schname)

# A tibble: 2,874 × 2  
 schname distinct\_schid\_count  
 <chr> <int>  
 1 abilene christian university 3  
 2 abraham baldwin agricultural college 4  
 3 academy for nursing and health occupations 2  
 4 academy of art university 3  
 5 adams state university 3  
 6 adelphi university 3  
 7 adirondack community college 3  
 8 adrian college 2  
 9 adventist university of health sciences 4  
10 agnes scott college 3  
# ℹ 2,864 more rows

final\_data\_join <- final\_data\_join %>%  
 group\_by(schname) %>%  
 filter(n\_distinct(schid) == 1) %>%  
 ungroup()  
  
vtable(final\_data\_join)

final\_data\_join

| Name | Class | Values |
| --- | --- | --- |
| schid | character |  |
| schname | character |  |
| keyword | character |  |
| keynum | integer | Num: 1 to 8 |
| monthorweek | character |  |
| index | integer | Num: 0 to 100 |
| \_file | character |  |
| first\_ten | character |  |
| month\_yr | Date | Time: 2013-03-24 to 2016-03-27 |
| month | Date | Time: 2013-03-01 to 2016-03-01 |
| standarized\_index | numeric | Num: -4.741 to 10.415 |
| unitid | integer | Num: 101286 to 482857 |
| opeid | integer | Num: 101800 to 52098870 |
| n | integer | Num: 1 to 1 |
| UNITID | integer | Num: 101286 to 48285706 |
| opeid6 | integer | Num: 1018 to 42249 |
| INSTNM | character |  |
| CITY | character |  |
| STABBR | character |  |
| INSTURL | character |  |
| NPCURL | character |  |
| HCM2 | integer | Num: 0 to 1 |
| PREDDEG | integer | Num: 0 to 3 |
| CONTROL | integer | Num: 1 to 3 |
| LOCALE | character |  |
| HBCU | character |  |
| PBI | character |  |
| ANNHI | character |  |
| TRIBAL | character |  |
| AANAPII | character |  |
| HSI | character |  |
| NANTI | character |  |
| MENONLY | character |  |
| WOMENONLY | character |  |
| RELAFFIL | character |  |
| SATVR25 | character |  |
| SATVR75 | character |  |
| SATMT25 | character |  |
| SATMT75 | character |  |
| SATWR25 | character |  |
| SATWR75 | character |  |
| SATVRMID | character |  |
| SATMTMID | character |  |
| SATWRMID | character |  |
| ACTCM25 | character |  |
| ACTCM75 | character |  |
| ACTEN25 | character |  |
| ACTEN75 | character |  |
| ACTMT25 | character |  |
| ACTMT75 | character |  |
| ACTWR25 | character |  |
| ACTWR75 | character |  |
| ACTCMMID | character |  |
| ACTENMID | character |  |
| ACTMTMID | character |  |
| ACTWRMID | character |  |
| SAT\_AVG | character |  |
| SAT\_AVG\_ALL | character |  |
| PCIP01 | character |  |
| PCIP03 | character |  |
| PCIP04 | character |  |
| PCIP05 | character |  |
| PCIP09 | character |  |
| PCIP10 | character |  |
| PCIP11 | character |  |
| PCIP12 | character |  |
| PCIP13 | character |  |
| PCIP14 | character |  |
| PCIP15 | character |  |
| PCIP16 | character |  |
| PCIP19 | character |  |
| PCIP22 | character |  |
| PCIP23 | character |  |
| PCIP24 | character |  |
| PCIP25 | character |  |
| PCIP26 | character |  |
| PCIP27 | character |  |
| PCIP29 | character |  |
| PCIP30 | character |  |
| PCIP31 | character |  |
| PCIP38 | character |  |
| PCIP39 | character |  |
| PCIP40 | character |  |
| PCIP41 | character |  |
| PCIP42 | character |  |
| PCIP43 | character |  |
| PCIP44 | character |  |
| PCIP45 | character |  |
| PCIP46 | character |  |
| PCIP47 | character |  |
| PCIP48 | character |  |
| PCIP49 | character |  |
| PCIP50 | character |  |
| PCIP51 | character |  |
| PCIP52 | character |  |
| PCIP54 | character |  |
| DISTANCEONLY | character |  |
| UGDS | character |  |
| UGDS\_WHITE | character |  |
| UGDS\_BLACK | character |  |
| UGDS\_HISP | character |  |
| UGDS\_ASIAN | character |  |
| UGDS\_AIAN | character |  |
| UGDS\_NHPI | character |  |
| UGDS\_2MOR | character |  |
| UGDS\_NRA | character |  |
| UGDS\_UNKN | character |  |
| PPTUG\_EF | character |  |
| CURROPER | integer | Num: 0 to 1 |
| NPT4\_PUB-AVERAGE-ANNUAL-COST | character |  |
| NPT4\_PRIV | character |  |
| NPT41\_PUB | character |  |
| NPT42\_PUB | character |  |
| NPT43\_PUB | character |  |
| NPT44\_PUB | character |  |
| NPT45\_PUB | character |  |
| NPT41\_PRIV | character |  |
| NPT42\_PRIV | character |  |
| NPT43\_PRIV | character |  |
| NPT44\_PRIV | character |  |
| NPT45\_PRIV | character |  |
| PCTPELL | character |  |
| RET\_FT4 | character |  |
| RET\_FTL4 | character |  |
| RET\_PT4 | character |  |
| RET\_PTL4 | character |  |
| PCTFLOAN | character |  |
| UG25abv | character |  |
| GRAD\_DEBT\_MDN\_SUPP | character |  |
| GRAD\_DEBT\_MDN10YR\_SUPP | character |  |
| RPY\_3YR\_RT\_SUPP | character |  |
| C150\_4\_POOLED\_SUPP-REPORTED-GRAD-RATE | character |  |
| C200\_L4\_POOLED\_SUPP | character |  |
| md\_earn\_wne\_p10-REPORTED-EARNINGS | character |  |
| gt\_25k\_p6 | character |  |

sumtable(final\_data\_join)

Summary Statistics

| Variable | N | Mean | Std. Dev. | Min | Pctl. 25 | Pctl. 75 | Max |
| --- | --- | --- | --- | --- | --- | --- | --- |
| keynum | 79863 | 1.9 | 1.3 | 1 | 1 | 2 | 8 |
| index | 79055 | 43 | 21 | 0 | 28 | 58 | 100 |
| standarized\_index | 53759 | 0.000000000000000012 | 1 | -4.7 | -0.69 | 0.54 | 10 |
| unitid | 79863 | 295368 | 133273 | 101286 | 177579 | 445267 | 482857 |
| opeid | 79863 | 2124835 | 4466763 | 101800 | 393805 | 2545200 | 52098870 |
| n | 79863 | 1 | 0 | 1 | 1 | 1 | 1 |
| UNITID | 79863 | 386563 | 2123192 | 101286 | 176947 | 444787 | 48285706 |
| opeid6 | 79863 | 14801 | 12524 | 1018 | 3938 | 23344 | 42249 |
| HCM2 | 79863 | 0.012 | 0.11 | 0 | 0 | 0 | 1 |
| PREDDEG | 79863 | 2.4 | 0.51 | 0 | 2 | 3 | 3 |
| CONTROL | 79863 | 2.4 | 0.69 | 1 | 2 | 3 | 3 |
| HBCU | 79863 |  |  |  |  |  |  |
| … 0 | 78614 | 98% |  |  |  |  |  |
| … 1 | 1093 | 1% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| PBI | 79863 |  |  |  |  |  |  |
| … 0 | 79707 | 100% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| ANNHI | 79863 |  |  |  |  |  |  |
| … 0 | 79551 | 100% |  |  |  |  |  |
| … 1 | 156 | 0% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| TRIBAL | 79863 |  |  |  |  |  |  |
| … 0 | 78426 | 98% |  |  |  |  |  |
| … 1 | 1281 | 2% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| AANAPII | 79863 |  |  |  |  |  |  |
| … 0 | 79229 | 99% |  |  |  |  |  |
| … 1 | 478 | 1% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| HSI | 79863 |  |  |  |  |  |  |
| … 0 | 79082 | 99% |  |  |  |  |  |
| … 1 | 625 | 1% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| NANTI | 79863 |  |  |  |  |  |  |
| … 0 | 79707 | 100% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| MENONLY | 79863 |  |  |  |  |  |  |
| … 0 | 79385 | 99% |  |  |  |  |  |
| … 1 | 322 | 0% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| WOMENONLY | 79863 |  |  |  |  |  |  |
| … 0 | 79395 | 99% |  |  |  |  |  |
| … 1 | 312 | 0% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| ACTWR25 | 79863 |  |  |  |  |  |  |
| … 10 | 158 | 0% |  |  |  |  |  |
| … 5 | 313 | 0% |  |  |  |  |  |
| … 6 | 1570 | 2% |  |  |  |  |  |
| … 7 | 313 | 0% |  |  |  |  |  |
| … 8 | 468 | 1% |  |  |  |  |  |
| … NULL | 77041 | 96% |  |  |  |  |  |
| ACTWR75 | 79863 |  |  |  |  |  |  |
| … 10 | 156 | 0% |  |  |  |  |  |
| … 11 | 157 | 0% |  |  |  |  |  |
| … 12 | 315 | 0% |  |  |  |  |  |
| … 8 | 1726 | 2% |  |  |  |  |  |
| … 9 | 468 | 1% |  |  |  |  |  |
| … NULL | 77041 | 96% |  |  |  |  |  |
| ACTWRMID | 79863 |  |  |  |  |  |  |
| … 11 | 158 | 0% |  |  |  |  |  |
| … 7 | 1569 | 2% |  |  |  |  |  |
| … 8 | 313 | 0% |  |  |  |  |  |
| … 9 | 782 | 1% |  |  |  |  |  |
| … NULL | 77041 | 96% |  |  |  |  |  |
| PCIP29 | 79863 |  |  |  |  |  |  |
| … 0 | 79394 | 99% |  |  |  |  |  |
| … 0.0039 | 157 | 0% |  |  |  |  |  |
| … 0.0204 | 156 | 0% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| DISTANCEONLY | 79863 |  |  |  |  |  |  |
| … 0 | 79228 | 99% |  |  |  |  |  |
| … 1 | 479 | 1% |  |  |  |  |  |
| … NULL | 156 | 0% |  |  |  |  |  |
| CURROPER | 79863 | 0.95 | 0.22 | 0 | 1 | 1 | 1 |

**SCORECARD VARIABLES AND DATA CLASSIFICATIONS**

As the center point of the research is to measure the causal effect of graduate earnings related to Google search trends, some barriers must be established to reduce confounding variables. Some variables that must be considered include location settings and college types that confound the causal relationships. As more jobs are centered in the nation’s metropolitan areas, schools located in urban settings have an advantage as more national headquarters are seeking new graduates.

**High-Earning and Low-Earning Classifications**

To conduct this analysis effectively, categorizing income levels is crucial due to the infinite nature of income. In the determination of income status, the median earnings of students who are employed but not enrolled for 10 years after their entry will be utilized as an indicator. College Score provides median income data for graduates at 6, 8, and 10 years post-graduation. Opting for the 10-year mark acknowledges the likelihood of individuals having settled into more stable career trajectories within their respective industries, thereby mitigating income volatility often experienced by recent graduates. By using the median earnings 10-year post-graduation value, percentiles will be used to categorize the income level as high, medium, and low.

*0-35% as Low Income* 36-80% as Medium Income \*81-100% as High Income

The below table breaks out the 10-year post-graduate average earnings into three different categories:

final\_data\_join <- final\_data\_join %>%   
 mutate(year = year(month))  
  
final\_data\_join <- final\_data\_join %>%  
 mutate(`md\_earn\_wne\_p10-REPORTED-EARNINGS` = as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`))

Warning: There was 1 warning in `mutate()`.  
ℹ In argument: `md\_earn\_wne\_p10-REPORTED-EARNINGS =  
 as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`)`.  
Caused by warning:  
! NAs introduced by coercion

income\_group <- final\_data\_join %>%   
 group\_by(year) %>%  
 select(year, `md\_earn\_wne\_p10-REPORTED-EARNINGS`) %>%  
 filter(!is.na(`md\_earn\_wne\_p10-REPORTED-EARNINGS`) &   
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` != 'PrivacySuppressed' &  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` != 'NULL') %>%   
 mutate(`md\_earn\_wne\_p10-REPORTED-EARNINGS` = as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`))  
  
final\_data\_join <- final\_data\_join %>%   
 group\_by(year) %>%  
 filter(!is.na(`md\_earn\_wne\_p10-REPORTED-EARNINGS`) &   
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` != 'PrivacySuppressed' &  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` != 'NULL') %>%   
 mutate(`md\_earn\_wne\_p10-REPORTED-EARNINGS` = as.numeric(`md\_earn\_wne\_p10-REPORTED-EARNINGS`))  
  
income\_group <- income\_group %>%  
 mutate(percentile\_group = case\_when(  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` <= quantile(`md\_earn\_wne\_p10-REPORTED-EARNINGS`, 0.35) ~ "Low",  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` > quantile(`md\_earn\_wne\_p10-REPORTED-EARNINGS`, 0.35) &   
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` <= quantile(`md\_earn\_wne\_p10-REPORTED-EARNINGS`, 0.9) ~ "Mid",  
 TRUE ~ "High"  
 ))   
  
final\_data\_join <- final\_data\_join %>%  
 mutate(percentile\_group = case\_when(  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` <= quantile(`md\_earn\_wne\_p10-REPORTED-EARNINGS`, 0.35) ~ "Low",  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` > quantile(`md\_earn\_wne\_p10-REPORTED-EARNINGS`, 0.35) &   
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` <= quantile(`md\_earn\_wne\_p10-REPORTED-EARNINGS`, 0.9) ~ "Mid",  
 TRUE ~ "High"  
 ))   
  
income\_group <- na.omit(income\_group)  
final\_data\_join <- na.omit(final\_data\_join)  
  
  
income\_group\_range1 <- income\_group %>%  
 group\_by(year,percentile\_group) %>%  
 summarize(min\_earnings = min(`md\_earn\_wne\_p10-REPORTED-EARNINGS`, na.rm = TRUE),  
 max\_earnings = max(`md\_earn\_wne\_p10-REPORTED-EARNINGS`, na.rm = TRUE))

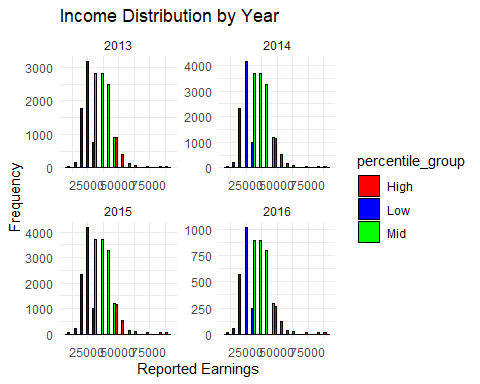
`summarise()` has grouped output by 'year'. You can override using the  
`.groups` argument.

print(income\_group\_range1)

# A tibble: 12 × 4  
# Groups: year [4]  
 year percentile\_group min\_earnings max\_earnings  
 <dbl> <chr> <dbl> <dbl>  
 1 2013 High 50500 87700  
 2 2013 Low 11600 28800  
 3 2013 Mid 28900 47500  
 4 2014 High 50500 87700  
 5 2014 Low 11600 28800  
 6 2014 Mid 28900 47500  
 7 2015 High 50500 87700  
 8 2015 Low 11600 28800  
 9 2015 Mid 28900 47500  
10 2016 High 50500 87700  
11 2016 Low 11600 28800  
12 2016 Mid 28900 47500

**DATA REPRESENTATION OF INCOME RANGE 10-yr post graduation**

ggplot(income\_group, aes(x = `md\_earn\_wne\_p10-REPORTED-EARNINGS`, fill = percentile\_group)) +  
 geom\_histogram(binwidth = 5000, position = "dodge", color = "black") +  
 facet\_wrap(~year, scales = "free") +  
 labs(title = "Income Distribution by Year",  
 x = "Reported Earnings",  
 y = "Frequency") +  
 scale\_fill\_manual(values = c("Low" = "blue", "Mid" = "green", "High" = "red")) +  
 theme\_minimal()

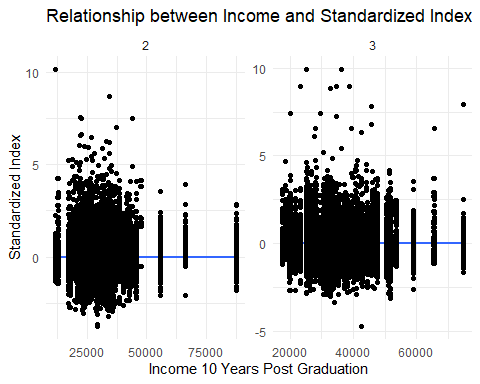


**Regression Analysis**

Before conducting the regression analysis to assess the causal relationship between reported earnings 10 years after graduation and the Google keyword search rate, the dataset will be streamlined to include only the necessary columns. The hypothesis is that the higher the income potential of a university, the more Google searches will result.As the College Scorecard was introduced in September 2015, a binary variable will be created to differentiate the post and pre-condition of the index. Additional control variables incorporated into the model encompass measures such as the distribution of degrees awarded by type, indicators of location characteristics, and income stratifications derived from percentile groupings established previously.

regression\_analysis <- final\_data\_join %>%   
 select(month, year, standarized\_index, `md\_earn\_wne\_p10-REPORTED-EARNINGS`, LOCALE, PREDDEG, percentile\_group) %>%  
 rename(income\_10 = `md\_earn\_wne\_p10-REPORTED-EARNINGS`) %>%   
 mutate(group = ifelse(month <= "2015-09-01", "pre", "post"))  
  
   
basic <- feols(standarized\_index~income\_10, data = regression\_analysis)  
  
ggplot(regression\_analysis, aes(x = income\_10, y = standarized\_index)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 facet\_wrap(~ PREDDEG, scales = "free") +  
 geom\_jitter() +   
 labs(title = "Relationship between Income and Standardized Index",  
 x = "Income 10 Years Post Graduation",  
 y = "Standardized Index") +  
 theme\_minimal()

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



pre\_post\_collegescorecard <- feols(standarized\_index~income\_10 \* group, data = regression\_analysis)  
with\_controls <-feols(standarized\_index ~ income\_10 \* group + LOCALE + PREDDEG + percentile\_group, data = regression\_analysis)  
  
  
etable(basic,pre\_post\_collegescorecard,with\_controls)

basic pre\_post\_collegesco..  
Dependent Var.: standarized\_index standarized\_index  
   
Constant 1.55e-16 (0.0159) 0.0418 (0.0388)  
income\_10 -3.58e-21 (4.46e-7) -1.07e-5\*\*\* (1.09e-6)  
grouppre -0.0488 (0.0424)  
income\_10 x grouppre 1.28e-5\*\*\* (1.19e-6)  
LOCALE12   
LOCALE13   
LOCALE21   
LOCALE22   
LOCALE23   
LOCALE31   
LOCALE32   
LOCALE33   
LOCALE41   
LOCALE42   
LOCALE43   
PREDDEG   
percentile\_groupLow   
percentile\_groupMid   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID IID  
Observations 45,790 45,790  
R2 -2.22e-16 0.02324  
Adj. R2 -2.18e-5 0.02318  
  
 with\_controls  
Dependent Var.: standarized\_index  
   
Constant 0.0363 (0.0726)  
income\_10 -1.06e-5\*\*\* (1.42e-6)  
grouppre -0.0488 (0.0424)  
income\_10 x grouppre 1.28e-5\*\*\* (1.19e-6)  
LOCALE12 0.0005 (0.0156)  
LOCALE13 0.0005 (0.0151)  
LOCALE21 0.0007 (0.0136)  
LOCALE22 0.0006 (0.0368)  
LOCALE23 0.0008 (0.0407)  
LOCALE31 0.0005 (0.0368)  
LOCALE32 0.0009 (0.0206)  
LOCALE33 0.0009 (0.0268)  
LOCALE41 0.0004 (0.0256)  
LOCALE42 0.0010 (0.0312)  
LOCALE43 0.0009 (0.0470)  
PREDDEG -0.0005 (0.0102)  
percentile\_groupLow 0.0039 (0.0352)  
percentile\_groupMid 0.0031 (0.0246)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 45,790  
R2 0.02324  
Adj. R2 0.02288  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

residuals <- residuals(with\_controls)  
bptest(residuals ~ income\_10 \* group + LOCALE + PREDDEG + percentile\_group, data = regression\_analysis)

studentized Breusch-Pagan test  
  
data: residuals ~ income\_10 \* group + LOCALE + PREDDEG + percentile\_group  
BP = 20.482, df = 17, p-value = 0.2503

**INTERPRETATION OF REGRESSION MODEL**

A total of three regression models were evaluated: \* Basic Model: Only examines the association between the median income 10 years after graduation, treated as the independent variable, and the standardized search index, designated as the dependent variable. \* pre\_post\_collegescorecard : Explores a distinguishing factor that delineates the causal relationship before and after the introduction of the College Scorecard initiative on September 1, 2015. \*with\_controls: Added additional control terms - Predominant degree awarded, Percentile Income Group, Location Type

When analyzing the outcomes of the aforementioned regression models, it becomes evident that the basic model fails to demonstrate any substantial effect on the standardized index. However, upon incorporating the interaction term distinguishing the pre and post-launch of the College Scoreboard initiative, income exhibits a notably significant negative influence on the index for the treatment group.

The negative coefficient signifies that, for the College Scorecard launch, higher income is associated with lower values of the search index while keeping other variables constant. Moreover, the smaller negative coefficient of grouppre suggests that the disparity in searching was mitigated following the release of income data.

The positive and significant coefficient of the income\_10 \* grouppre interaction term highlights a significnat relationship, indicating a reduced negative impact of median college income data on the search index. This suggests that post-graduation income may not be the primary factor influencing students’ decisions regarding college attendance.

Even with the inclusion of other control variables in the model, similar patterns persist, reinforcing the notion observed in the model that included pre and post-treatment groups. Specifically, there remains a significant negative effect of income for the treatment group, moderated by the positive and significant interaction term.

When considering heteroskadecity, this linear regression model appears to be reasonably well-specified and does not violate key assumptions. The Breusch-Pagan test for heteroskedasticity showed a high p-value, indicating the errors have constant variance and the homoskedasticity assumption is met.

**Conclusion**

The introduction of the College Scorecard increased search activity on Google Trends for colleges with high-earning graduates by 1.28 x 10^-5 standardized index units relative to what it did for colleges with low-earning graduates, with a standard error of 1.19 \* 10 ^-5. This result comes from the negative and statistically significant coefficient estimate on the income\_10 \* group variable.