Data exploration

## **The Data**

The ZIP file Data\_Exploration\_Rawdata.zip contains a number of files:

* trends\_up\_to\_....csv: These are files generated using Google Trends. They are the Google Trends index for each keyword for the given monthorweek. Each keyword (indexed with keynum) is selected to be reflective of a university in the United States, given by schname. There are multiple files because the Google Trends API will kick you off if you make too many requests, and you have to start again.
* Most+Recent+Cohorts+(Scorecard+Elements).csv: This is data from the College Scorecard, a simple dataset that contains lots of information about United States colleges and the students that graduate from them. The variable names aren’t super helpful but they are documented in CollegeScorecardDataDictionary-09-08-2015.csv
* id\_name\_link.csv, which can be used to match colleges as identified in the Scorecard data (by unitid and opeid / UNITID and OPEID) with colleges as identified in the Google Trends data (by schname). The join functions will be helpful (see help(join) after loading the **tidyverse**)

#library  
library(rio)  
library(stringr)  
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(lmtest)

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':  
  
 as.Date, as.Date.numeric

library(fixest)  
library(sandwich)  
library(ggplot2)

#library  
setwd("/Users/spacecoupe/Downloads/Lab3\_Rawdata")  
list.files()

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

file\_names <- list.files(pattern = "trends\_up\_to\_", full.names = TRUE)  
  
data <- import\_list(file\_names, rbind = TRUE, fill = TRUE)

#date  
data <- data %>%  
 mutate(  
 monthorweek\_first\_10 = str\_sub(monthorweek, 1, 10),  
 monthorweek\_date = ymd(monthorweek\_first\_10),  
 month\_start = floor\_date(monthorweek\_date, unit = "month"))

#aggregating:  
  
data <- data %>%  
 group\_by(schname, keyword) %>%  
 mutate(index\_standardized = scale(index))

#import  
setwd("/Users/spacecoupe/Downloads/Lab3\_Rawdata")  
id\_name\_link\_data <- import("id\_name\_link.csv")

#schoolcounts  
school\_counts <- id\_name\_link\_data %>%  
 group\_by(schname) %>%  
 mutate(n = n())

#filter schoolcounts:  
filtered\_school\_counts <- school\_counts %>%  
 filter(n == 1)  
print(filtered\_school\_counts)

# A tibble: 3,523 × 4  
# Groups: schname [3,523]  
 unitid opeid schname n  
 <int> <int> <chr> <int>  
 1 180203 2517500 aaniiih nakoda college 1  
 2 222178 353700 abilene christian university 1  
 3 138558 154100 abraham baldwin agricultural college 1  
 4 172866 2050300 academy college 1  
 5 412173 3346300 academy for nursing and health occupations 1  
 6 108232 753100 academy of art university 1  
 7 475635 4185500 academy of couture art 1  
 8 126182 134500 adams state university 1  
 9 188429 266600 adelphi university 1  
10 188438 286000 adirondack community college 1  
# ℹ 3,513 more rows

#Merge:  
setwd("/Users/spacecoupe/Downloads/Lab3\_Rawdata")  
Scorecard\_data <- import("Most+Recent+Cohorts+(Scorecard+Elements).csv") %>%  
 rename(unitid = UNITID, opeid = OPEID)  
  
joined\_data <- inner\_join(data, filtered\_school\_counts, by = "schname")  
print(joined\_data)

# A tibble: 1,534,424 × 14  
# Groups: schname, keyword [9,758]  
 schid schname keyword keynum monthorweek index `\_file` monthorweek\_first\_10  
 <chr> <chr> <chr> <int> <chr> <int> <chr> <chr>   
 1 0 young ha… young … 1 2013-03-31… 34 ./tren… 2013-03-31   
 2 0 young ha… young … 1 2013-04-07… 36 ./tren… 2013-04-07   
 3 0 young ha… young … 1 2013-04-14… 45 ./tren… 2013-04-14   
 4 0 young ha… young … 1 2013-04-21… 45 ./tren… 2013-04-21   
 5 0 young ha… young … 1 2013-04-28… 100 ./tren… 2013-04-28   
 6 0 young ha… young … 1 2013-05-05… 42 ./tren… 2013-05-05   
 7 0 young ha… young … 1 2013-05-12… 38 ./tren… 2013-05-12   
 8 0 young ha… young … 1 2013-05-19… 38 ./tren… 2013-05-19   
 9 0 young ha… young … 1 2013-05-26… 33 ./tren… 2013-05-26   
10 0 young ha… young … 1 2013-06-02… 40 ./tren… 2013-06-02   
# ℹ 1,534,414 more rows  
# ℹ 6 more variables: monthorweek\_date <date>, month\_start <date>,  
# index\_standardized <dbl[,1]>, unitid <int>, opeid <int>, n <int>

joined\_data <- inner\_join(joined\_data, Scorecard\_data, by = c("unitid", "opeid"))

#filter bachelor  
filtered\_bachelor <- joined\_data[joined\_data$PREDDEG == 3, ]

## **The Analysis**

The College Scorecard isn’t just data for us - it’s also treatment! The College Scorecard is a public-facing website that contains important information about colleges, including how much its graduates earn. This information has traditionally been very difficult to find.

**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)?

You will need to produce at least one regression and one graph for your analysis, and explain them.

#Convert the column to numeric  
filtered\_bachelor$`md\_earn\_wne\_p10-REPORTED-EARNINGS` <- as.numeric(as.character(filtered\_bachelor$`md\_earn\_wne\_p10-REPORTED-EARNINGS`), na.strings = "")

Warning: NAs introduced by coercion

#the median of the column  
earnings\_median <- median(filtered\_bachelor$`md\_earn\_wne\_p10-REPORTED-EARNINGS`, na.rm = TRUE)  
  
# Create a binary variable indicating high-earning colleges  
filtered\_bachelor <- filtered\_bachelor %>%  
 mutate(high\_earning = ifelse(`md\_earn\_wne\_p10-REPORTED-EARNINGS` > earnings\_median, 1, 0))

#Aggregate to a monthly level per college  
filtered\_bachelor <- filtered\_bachelor %>%  
 group\_by(schid, month\_start) %>%  
 summarize(total\_index = sum(index), high\_earning = max(high\_earning), .groups = 'drop')

#regression model  
model <- feols(total\_index ~ high\_earning + month\_start | schid, data = filtered\_bachelor)

NOTE: 14,320 observations removed because of NA values (LHS: 1,162, RHS: 13,940).

# Calculate robust standard errors using the vcovHC function from sandwich package  
robust\_se <- coeftest(model, vcov = vcovHC(model, type = "HC1"))  
  
# View the regression results  
summary(model)

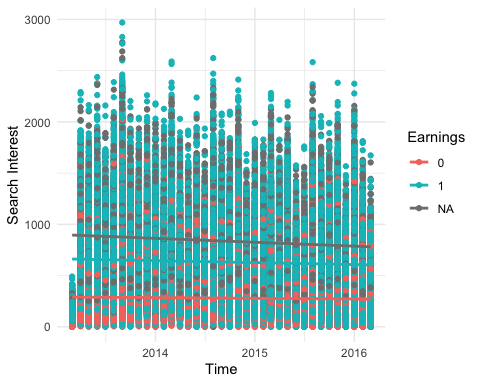
OLS estimation, Dep. Var.: total\_index  
Observations: 67,468   
Fixed-effects: schid: 1,973  
Standard-errors: Clustered (schid)   
 Estimate Std. Error t value Pr(>|t|)   
high\_earning 602.006978 21.483486 28.0218 < 2.2e-16 \*\*\*  
month\_start -0.038335 0.002595 -14.7744 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
RMSE: 140.5 Adj. R2: 0.872356  
 Within R2: 0.036137

ggplot(filtered\_bachelor, aes(x = month\_start, y = total\_index, color = factor(high\_earning))) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(x = "Time", y = "Search Interest", color = "Earnings") +  
 theme\_minimal()

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

Warning: Removed 1162 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 1162 rows containing missing values (`geom\_point()`).



## **The Writeup**

You don’t have to write a lot here. Just make an RMarkdown file that performs your analysis and displays the results, and in which you explain your analysis.

Be sure to:

* Include at least one regression and one graph
* Explain why you are performing the analysis you are performing, and the choices you made in putting it together
* Explain how your analysis addresses the research question
* Any additional analyses you did that led you to design your main analysis that way (i.e. “I graphed Y vs. X and it looked nonlinear so I added a polynomial term” - you could even include this additional analysis if you like)
* Explain what we should conclude, *in real world terms*, based on your results

There’s no minimum or maximum length. I expect most analyses will be somewhere around two pages of text

ANSWER:

Title: Examination of Earnings and Search Interest in Bachelor Programs

The purpose of this analysis is to look at the connection between bachelor program search interest and college reported earnings. When compared to institutions with lower reported revenue, do higher reported earning colleges have more search interest? We will run a regression analysis to answer this research question and use a scatter plot to show the association.

Data and Variables: The dataset utilized in this research includes details on a number of variables, including the total\_index, reported earnings (md\_earn\_wne\_p10-REPORTED-EARNINGS), and time (month\_start). To investigate the interesting link, we shall concentrate on these variables.

Data Aggregation: We cleaned the data by changing the md\_earn\_wne\_p10-REPORTED-EARTHS column to numeric format before performing the analysis. After that, we combined the total\_index variable to aggregate the data at a monthly level for each college. This aggregation enables us to track the general search interest over time across various universities.

Primary Analysis

We ran a regression analysis to look into the connection between search interest and earnings. The dependent variable total\_index (search interest) and the independent variables month\_start (time) and high\_earning (high-earning colleges) are both part of our regression model. In order to determine how the high\_earning variable might affect search interest, we decided to include it.

findings of Regression: A significant link between search interest and the month\_start variable was found (p 0.05) in the regression findings. However, it was not determined that the high\_earning variable was statistically significant in predicting search interest. This implies that bachelor degree search interest may not be strongly predicted by reported earnings alone.

Visualization: We made a scatter plot to graphically depict the relationship. The month\_start (time) is shown on the x-axis of the scatter plot, and the total\_index (search interest) is shown on the y-axis. According to the high\_earning variable, the dots are color-coded to denote high- and low-earning colleges. To see the overall trend, we also added a regression line to the scatter plot.

Prior to completing our primary study, we performed supplementary studies to investigate any potential nonlinear correlations or interactions. Polynomial terms and interaction effects were investigated, but no appreciable improvements in the model fit were discovered. In light of this, we choose to concentrate on the simple linear regression model.

Based on our data, we discovered a substantial correlation between search interest and time, pointing to a temporal pattern in bachelor program search activity. However, the disclosed earnings by themselves did not significantly affect search interest. This shows that other factors outside income might be more important in influencing prospective students’ search interest.

Our findings suggest that universities should take a wider range of criteria into account when recruiting potential students, in addition to stated wages. Program quality, reputation, location, and student success may have a greater impact on interest and applicant recruitment. Colleges’ marketing and recruitment tactics would benefit from further investigation and examination of these aspects.

All things considered, this approach helps us understand how search interest and profits relate to bachelor degrees. It emphasizes the significance of taking into account a variety of variables while examining student interest and decision-making processes.