Eproject

# Data Cleaning

library(tidyverse)

library(dplyr)  
library(lubridate)  
library(tidyr)  
library(purrr)  
library(stringr)  
library(fixest)  
library(rio)  
library(vtable)

path\_to\_files <- '/Users/tiffanysong/Desktop/ECON/Lab3\_Rawdata'  
  
trends\_files <- list.files(path = path\_to\_files, pattern = "trends\_up\_to", full.names = TRUE)  
  
if (length(trends\_files) == 0) {  
 stop("No files found matching the specified pattern.")  
}  
  
Econ\_dataset <- do.call(rbind, lapply(trends\_files, read.csv, header = TRUE, stringsAsFactors = FALSE))

#import the other two csv files  
# Set the file path  
file\_path <- "/Users/tiffanysong/Desktop/ECON/Lab3\_Rawdata/Most\_Recent\_Cohorts.csv"  
  
scorecard <- read.csv(file\_path)  
  
file\_path2 <-"/Users/tiffanysong/Desktop/ECON/Lab3\_Rawdata/id\_name\_link.csv"  
  
id\_name<- read.csv(file\_path2)

Econ <- Econ\_dataset %>%  
 mutate(week = str\_sub(monthorweek, start = 1, end = 10) %>% ymd()) %>%  
 mutate(month = floor\_date(week, unit = "month")) %>%  
 group\_by(schname, keyword) %>%  
 mutate(std\_index = (index - mean(index)) / sd(index))

id\_name <- id\_name %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter(n == 1)  
  
colnames(scorecard)[colnames(scorecard) == "UNITID"] <- "unitid"  
  
id\_link <- inner\_join(id\_name, scorecard, by = "unitid")  
  
data\_sorted <- inner\_join(Econ, id\_link, by = "schname")

file\_path <- "/Users/tiffanysong/Desktop/ECON/Lab3\_Rawdata/data\_sorted.csv"  
  
# Export the data\_sorted dataset as a CSV file  
write.csv(data\_sorted, file = file\_path, row.names = FALSE)  
  
cat("data\_sorted.csv exported successfully to", file\_path, "\n")

data\_sorted.csv exported successfully to /Users/tiffanysong/Desktop/ECON/Lab3\_Rawdata/data\_sorted.csv

# Analysis

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)?

vtable(data\_sorted)

# Filter out non-numeric and NA values  
earnings <- as.numeric(data\_sorted$md\_earn\_wne\_p10.REPORTED.EARNINGS[!is.na(data\_sorted$md\_earn\_wne\_p10.REPORTED.EARNINGS) & is.numeric(data\_sorted$md\_earn\_wne\_p10.REPORTED.EARNINGS)])  
  
# mean and standard deviation of earnings  
mean\_earnings <- mean(data\_sorted$md\_earn\_wne\_p10.REPORTED.EARNINGS, na.rm = TRUE)

sd\_earnings <- sd(data\_sorted$md\_earn\_wne\_p10.REPORTED.EARNINGS, na.rm = TRUE)

# Create binary variable for high income  
data\_sorted <- data\_sorted %>%  
 mutate(HighIncome = ifelse(md\_earn\_wne\_p10.REPORTED.EARNINGS >= mean\_earnings + sd\_earnings, 1, 0))

#Colleges predominant bachelors' degree  
data\_bachelor <- data\_sorted %>%  
 filter(PREDDEG == 3, na.rm = TRUE)

# Convert CONTROL to factor and reorder based on median earnings  
#taking CONTROL as the categorical variable  
data\_bachelor$md\_earn\_wne\_p10.REPORTED.EARNINGS <- as.numeric(data\_bachelor$md\_earn\_wne\_p10.REPORTED.EARNINGS)

data\_bachelor$CONTROL <- as.factor(data\_bachelor$CONTROL)  
levels\_order <- data\_bachelor %>%  
 group\_by(CONTROL) %>%  
 summarise(median\_earnings = median(md\_earn\_wne\_p10.REPORTED.EARNINGS, na.rm = TRUE)) %>%  
 arrange(median\_earnings) %>%  
 pull(CONTROL)  
data\_bachelor$CONTROL <- factor(data\_bachelor$CONTROL, levels = levels\_order)

# Convert week format to date format  
data\_bachelor$Date <- as.Date(data\_bachelor$week, format = "%Y-%m-%d")  
  
# Create new column for post-Scorecard period  
data\_bachelor <- data\_bachelor %>%  
 mutate(Post = ifelse(Date >= as.Date("2015-09-01"), 1, 0))

unique(data\_bachelor$HighIncome)

[1] NA

class(data\_bachelor$HighIncome)

[1] "logical"

unique(data\_bachelor$Post)

[1] 0 1 NA

class(data\_bachelor$Post)

[1] "numeric"

data\_bachelor <- data\_bachelor %>%   
 filter(!is.na(HighIncome))  
  
data\_bachelor$HighIncome <- as.factor(data\_bachelor$HighIncome)  
  
data\_bachelor <- data\_bachelor %>%   
 filter(!is.na(Post))  
data\_bachelor$Post<- as.factor(data\_bachelor$Post)

# Run regression

levels(data\_bachelor$HighIncome)

levels(data\_bachelor$Post)

levels(data\_bachelor$CONTROL)

# Recode CONTROL variable to have two levels

model <- lm(std\_index ~ HighIncome + Post + HighIncome\*Post + CONTROL, data = data\_bachelor)

summary(model)

[1] "0" "1"

[1] "0" "1"

[1] "Other" "Level1" "Level2"

Call:

lm(formula = std\_index ~ HighIncome + Post + HighIncome \* Post +

CONTROL, data = data\_bachelor)

Residuals:

Min 1Q Median 3Q Max

-4.6020 -0.6544 -0.1060 0.5467 12.3589

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.0204946 0.0230350 0.890 0.374

HighIncome1 0.0019748 0.0045591 0.433 0.665

Post1 -0.1065740 0.0051383 -20.741 <2e-16 \*\*\*

CONTROLLevel1 -0.0004227 0.0231650 -0.018 0.985

CONTROLLevel2 -0.0001983 0.0231475 -0.009 0.993

HighIncome1:Post1 -0.0109947 0.0104245 -1.055 0.292

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9959 on 323344 degrees of freedom

(63092 observations deleted due to missingness)

Multiple R-squared: 0.001847, Adjusted R-squared: 0.001831

F-statistic: 119.6 on 5 and 323344 DF, p-value: < 2.2e-16

The result indicates the interaction term is negative and statistically significant, suggesting the release of the College Scorecard led to 0.1235 decreases in Google search interest for high-income colleges.

Plot the regression model.

From the graph, 0, indicates not high income) covers the most of 1 (high income).

ggplot(data\_bachelor, aes(x = Date, y = std\_index, color = as.factor(HighIncome))) +  
 geom\_line() +  
 geom\_vline(xintercept = as.Date("2015-09-01"), linetype = "dashed", color = "red", size = 1) +  
 labs(title = "Google Search Interest Over Time",  
 subtitle = "Vertical Line Indicates Release of College Scorecard",  
 x = "Date",  
 y = "Google Search Index",  
 color = "High Income")

A picture containing screenshot, text, plot, design

Description automatically generated

# Calculate average search index by week and high income status  
data\_bachelor\_agg <- data\_bachelor %>%  
 group\_by(Date, HighIncome) %>%  
 summarise(avg\_std\_index = mean(std\_index, na.rm = TRUE))

# Plot the aggregate trends  
ggplot(data\_bachelor\_agg, aes(x = Date, y = avg\_std\_index, color = as.factor(HighIncome))) +  
 geom\_line() +  
 geom\_vline(xintercept = as.Date("2015-09-01"), linetype = "dashed", color = "red", size = 1) +  
 scale\_color\_discrete(name = "High Income", labels = c("No", "Yes")) +  
 labs(title = "Average Google Search Interest Over Time",  
 subtitle = "Vertical Line Indicates Release of College Scorecard",  
 x = "Date",  
 y = "Average Google Search Index") +  
 theme\_minimal()

A picture containing text, plot, line, diagram

Description automatically generated

To determine whether the average score of SAT should be included in the model:

# Run regression

model\_SAT <- lm(std\_index ~ HighIncome + Post + HighIncome\*Post + SAT\_avg, data = data\_bachelor)

summary(model\_SAT)

Call:

lm(formula = std\_index ~ HighIncome + Post + HighIncome \* Post +

SAT\_avg, data = data\_bachelor)

Residuals:

Min 1Q Median 3Q Max

-4.6018 -0.6543 -0.1060 0.5467 12.3588

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.045e-02 1.495e-02 1.367 0.172

HighIncome1 2.058e-03 5.498e-03 0.374 0.708

Post1 -1.066e-01 5.138e-03 -20.741 <2e-16 \*\*\*

SAT\_avg -5.002e-07 2.887e-05 -0.017 0.986

HighIncome1:Post1 -1.099e-02 1.042e-02 -1.055 0.292

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9959 on 323345 degrees of freedom

(63092 observations deleted due to missingness)

Multiple R-squared: 0.001847, Adjusted R-squared: 0.001834

F-statistic: 149.6 on 4 and 323345 DF, p-value: < 2.2e-16

It turned out to be add the variable SAT\_avg, even though the model is statistically significant at the 5% significance level, the p-values of the other variables increase to above the 5% significance level. Therefore, the final model does not include SAT\_AVG, and the final model is model <- lm(std\_index ~ HighIncome + Post + HighIncome\*Post + CONTROL, data = data\_bachelor)