Data Exploration

library(rio)  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.1 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── 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

dat <- import('cleandat.csv')

**REGRESSIONS AND GRAPH STUFF**

library(fixest)  
dat <- dat %>%  
 mutate(high\_low = case\_when(`md\_earn\_wne\_p10-REPORTED-EARNINGS` >= 60000 ~ 'High',  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` <= 50000 ~ "Low",   
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` > 50000 & `md\_earn\_wne\_p10-REPORTED-EARNINGS` < 60000 ~ "Medium"))   
  
dat <- dat %>%  
 mutate(after = date >= ym('2015-09'))  
  
m1 <- feols(index\_std ~ monthorweek + high\_low + after, data = dat)

NOTE: 8,475 observations removed because of NA values (LHS: 8,475, RHS: 7,583).

summary(m1)

OLS estimation, Dep. Var.: index\_std  
Observations: 1,525,949   
Standard-errors: IID   
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 12.688719 0.05516395 230.018317 < 2.2e-16 \*\*\*  
monthorweek -0.000779 0.00000340 -229.251037 < 2.2e-16 \*\*\*  
high\_lowLow 0.000163 0.00265327 0.061321 0.95110   
high\_lowMedium -0.000757 0.00350107 -0.216091 0.82892   
afterTRUE 0.130731 0.00273878 47.733160 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
RMSE: 0.973475 Adj. R2: 0.046245

dat\_flat <- dat %>%  
 group\_by(after, date, high\_low) %>%  
 summarize(index\_std = mean(index\_std))

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

ggplot(dat\_flat, aes(x = date, y = index\_std, color = high\_low)) + geom\_line() +   
 geom\_vline(xintercept = ym('2015-09'))

Warning: Removed 3 rows containing missing values (`geom\_line()`).



m2 <- feols(index\_std ~ date + high\_low + after, data = dat)

NOTE: 8,475 observations removed because of NA values (LHS: 8,475, RHS: 7,583).

etable(m2)

m2  
Dependent Var.: index\_std  
   
Constant 12.81\*\*\* (0.0552)  
date -0.0008\*\*\* (3.39e-6)  
high\_lowLow 0.0002 (0.0027)  
high\_lowMedium -0.0008 (0.0035)  
afterTRUE 0.1348\*\*\* (0.0027)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 1,525,949  
R2 0.04687  
Adj. R2 0.04687  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

m3 <- feols(index\_std ~ after + high\_low + after\*high\_low, data = dat)

NOTE: 8,475 observations removed because of NA values (LHS: 8,475, RHS: 7,583).

etable(m3)

m3  
Dependent Var.: index\_std  
   
Constant 0.0550\*\*\* (0.0028)  
afterTRUE -0.2912\*\*\* (0.0065)  
high\_lowLow -0.0019 (0.0030)  
high\_lowMedium 0.0214\*\*\* (0.0040)  
afterTRUE x high\_lowLow 0.0104 (0.0069)  
afterTRUE x high\_lowMedium -0.1157\*\*\* (0.0091)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 1,525,949  
R2 0.01362  
Adj. R2 0.01362  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**WRITE UP**

**Explanation:** When defining variables, I established high earnings as earnings greater than or equal to $60,000, low earnings as earnings less than or equal to $50,000, and medium earnings as any earnings between $50,000 and $60,000. These cutoff points were based on average post-college earnings in the United States, which is approximately $55,000. I denoted these three variables using the variable name “high\_low.”

The initial graph illustrates the trends in the index before and after the introduction of the scorecard in September 2015. While the graph does not specifically isolate the effect of the scorecard on colleges with high or low earnings, it demonstrates the relationship between the index/outcome and high, medium, and low earning colleges over time, both before and after the introduction of the scorecard. The graph indicates a downward trend in the index before the scorecard was introduced, which continued after its introduction in September 2015. This conclusion is drawn by considering the inclusion of a line in the graph indicating the timing of the scorecard introduction relative to the index trends of post-college earnings.

The next step involved isolating the effect of the scorecard on student interest (index) in high-earning or low-earning colleges using dummy variables for the scorecard and high earnings. This was accomplished through a model derived from the equation: d(index)/d(scorecard) = b1 + b3\*high\_low. Based on this graph, I conducted two regressions, m2 and m3. m2 examines the general effect of the scorecard on student interest across high-earning and low-earning colleges, while m3 delves further into the changes in student interest before and after the scorecard for both high-earning and low-earning colleges.

**Interpretation:** A one unit change in the variable afterTRUE, holding other variables constant, is associated with a decrease of -0.2912 units in the index. In other words, for each college included in the scorecard after September 2015, there is a decrease of -0.2912 units in the likelihood that the college is searched on Google.

A one unit change in the variable high\_lowMedium, controlling for other variables, results in an increase of 0.0214 units in the index. This implies that for each medium-earning college, there is an increase of 0.0214 units in the likelihood that the college is searched on Google.

A one unit change in the interaction variable afterTRUE x high\_lowMedium is associated with a decrease of -0.1157 units in the index. This means that for each medium-earning college included in the scorecard, there is a decrease of -0.1157 units in the likelihood that the college is searched on Google, holding all other variables constant.

**Real-world interpretation:**

The introduction of the Scorecard had a significant negative impact on student interest in colleges that were included in the Scorecard. For each college included in the Scorecard after September 2015, there was a decrease in the likelihood that the college was searched on Google. This suggests that the release of the Scorecard may have influenced students to explore colleges less frequently or shifted their attention away from the included colleges.

Colleges with medium earnings, falling between the range of $50,000 and $60,000, experienced a slight increase in student interest. Each medium-earning college had a modest increase in the likelihood of being searched on Google. This implies that the Scorecard may have had a limited positive effect on generating interest in medium-earning colleges among students.

However, when considering the interaction between the Scorecard and medium-earning colleges, there was a decrease in student interest. For each medium-earning college included in the Scorecard, there was a decrease in the likelihood of being searched on Google. This suggests that the Scorecard’s influence on student interest in medium-earning colleges may have had a negative impact, counteracting the general increase observed for medium-earning colleges.