Data Exploration Analysis

## Data Exploration Analysis

For this analysis we have been asked to investigate, “Did the release of the College Scorecard shift student interest towards high earnings colleges vs. low earnings colleges?”

After cleaning up the necessary data, we need to be able to differentiate between “High Earning Colleges” and “Low Earning Colleges”. I will be using the median earnings after 10 years to set a threshold to differentiate between high and low earning colleges. Additionally, I’ve decided to create a dummy variable that has 1 = low earning colleges and 0 = high earning colleges to help me describe the differences in interest between the categories.

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(readr)  
library(fixest)  
library(ggplot2)

#filtered\_data$`med\_earning` <- as.numeric(filtered\_data$`med\_earning`)  
  
#threshold <- quantile(filtered\_data$med\_earning, 0.50, na.rm = TRUE)  
  
#filtered\_data <- filtered\_data %>%  
 #mutate(income\_cat = ifelse(med\_earning > threshold, 0,1))

Now that our thresholds and dummy variables have been created, I need to filter my month down to show post Scorecard data and the pre Scorecard data. The reason being because we can see the effect the Scorecard data had on interest of high earning and low earning colleges using the Standardized Index we created.

#data filtered down to show dates post intro of Scorecard data (Sept. 9, 2015)  
  
  
#execute regression  
#m1 <- feols(standardize\_index ~ month, data = filtered\_data)  
  
#summary(m1)

A screenshot of a computer

Description automatically generated

We can see that after running the regression above, that the interest level of Low Earning colleges after each month post introduction of Scorecard data decreases slightly by 0.001517 as noted by the coefficient month.

#extracting the coefficients  
  
  
#ggplot(filtered\_data, aes(x = month, y = standardize\_index)) +  
# geom\_smooth(method = 'lm', se = FALSE, color = 'blue') +  
# labs(title = 'Scatter Plot with Regression Line')  
# x = "Month"  
# y = "Standardized Index" +  
# theme\_minimal()

A graph with a line and a blue line

Description automatically generated

The visual of the regression model solidifies what we stated. For each month that is after September 2015, the interest to Low Earning colleges decreased.

I decided to keep my regression and visual simple, because I believe I set up my month variable to be able to differentiate between pre and post introduction of Scorecard data. I was able to filter my data set down to show data only after the introduction of the Scorecard data (Sept. 2015). This allowed me to conclude that the interest between high and low earning colleges really didn’t change much after the introduction of the Scorecard data.

So, we should can conclude that *the introduction of the College Scorecard slightly decreased interest in Low Earning colleges by 0.001517 units relative to what it did for High Earning colleges, with a s.e. of* 0.000051. These *results comes from the month coefficient in my regression.*