OfficalQAnalysis

## Analysis of Data Exploration Project

Loading in library’s

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(tidyverse)

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --  
v forcats 1.0.0 v readr 2.1.4  
v ggplot2 3.4.1 v stringr 1.5.0  
v lubridate 1.9.2 v tibble 3.2.1  
v purrr 1.0.1 v tidyr 1.3.0

-- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
x dplyr::filter() masks stats::filter()  
x dplyr::lag() masks stats::lag()  
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(lubridate)  
library(rio)  
library(stringr)  
library(ggplot2)  
library(fixest)

setwd("C:/Users/rking/OneDrive/Documents/Econometrics/Data exploration project/Data\_Exploration\_Rawdata/")  
CData <- import('final\_data.xlsx')

In order to explore the earnings I need to remove the privacy secured ones and NULL’s

CData2 <- rename(CData, earnings = 'md\_earn\_wne\_p10-reported-earnings')  
  
  
CData2 <- CData2 %>%  
 filter(!str\_detect(earnings, "PrivacySuppressed|NULL"))

Trying to identify what high earning and low earning will be defined at

# Looking for mean and SD of income  
  
CData4 <- as.integer(CData2$earnings)  
mean(CData4, na.rm = TRUE)

[1] 42804.79

sd(CData4, na.rm = TRUE)

[1] 11833.26

# Making categories for earning level  
CData2 <- CData2 %>%  
 mutate(earning\_level = case\_when(  
 earnings <= 33200 ~ 'low',  
 earnings >= 56800 ~ 'high',  
 TRUE ~ 'Medium'  
 ))

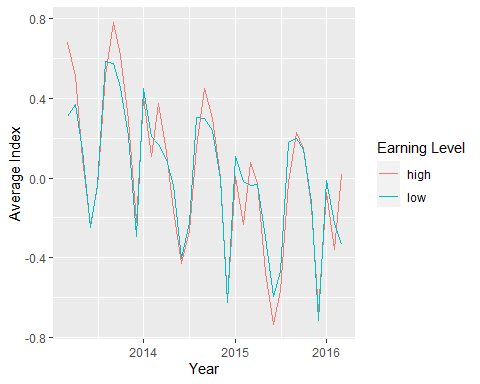
Making date groups, between before and after the scorecard was introduced

#was introduced on 09/01/2015  
CData2 <- CData2 %>%  
 mutate(SCEffect = case\_when(  
 monthorweek <= '2015-09-01' ~ 'before',  
 earnings > '2015-09-01' ~ 'after'  
 ))  
  
#Filtering how meduim earning schools and NA's  
  
highvlow <- filter(CData2, earning\_level != 'Medium')  
highvlow <- na.omit(highvlow)

highvlow2 <- highvlow %>% group\_by(SCEffect, earning\_level, monthorweek) %>%  
 summarize(mean\_index = mean(new\_index))

`summarise()` has grouped output by 'SCEffect', 'earning\_level'. You can  
override using the `.groups` argument.

ggplot(data = highvlow2, aes(x=monthorweek, y = mean\_index, color= earning\_level)) + geom\_line() + labs(x = "Year", y = "Average Index", color = "Earning Level")



Creating binary variables for my regression

regData <- highvlow %>%   
 mutate(high\_earners = ifelse(earning\_level == "high", 1, 0),  
 low\_earners = ifelse(earning\_level == "low", 1, 0))  
  
regData <- regData %>%   
 mutate(screlease = ifelse(monthorweek < '2015-09-01', 0, 1))

Running the regression

reggg <- feols(new\_index ~ high\_earners : screlease + low\_earners : screlease + monthorweek + locale , data = regData)  
etable(reggg)

reggg  
Dependent Var.: new\_index  
   
Constant 9.608\*\*\* (0.3152)  
monthorweek -6.82e-9\*\*\* (2.25e-10)  
locale12 0.0002 (0.0157)  
locale13 -0.0059 (0.0174)  
locale21 -0.0033 (0.0142)  
locale22 0.0018 (0.0393)  
locale23 -0.0041 (0.0481)  
locale31 -0.0024 (0.0321)  
locale32 -0.0027 (0.0198)  
locale33 -0.0021 (0.0204)  
locale41 -0.0023 (0.0230)  
locale42 0.0055 (0.0307)  
locale43 0.0011 (0.0470)  
high\_earners x screlease 0.1003\*\*\* (0.0246)  
screlease x low\_earners 0.0695\*\*\* (0.0191)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 13,638  
R2 0.08547  
Adj. R2 0.08453  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**The write-up:**

With such a large amount of data I decided to group the aggregated index by month so that I had a month-to-month scale for change in index because I though having week by week would be harder to see on a chart. In order to create my groups for earning levels I took the mean income of the entire df and it was $42,804, then the mean income between 2013 and 2016 according to the “Social Security Administration” was $47,028. Looking at these two numbers I chose to use $45,000 as the true average of incomes. I took the standard deviation and that was $11,800 and I used this to determine my high and low earning groups. The low earning group will be one SD below the mean ($33,200) and the high earning group will be one SD above ($56,800).

When starting this analysis I wanted to get a look at how the data was so that when I ran my regression I could have a sense of if I was on the right track or not. I decided to group by earning level and the monthorweek variable so that I could get averages for the index scores based on the month and earning level. Putting this on a line graph was very helpful since I could see the changes over time and the potential effects that the score card had. The main find that I had from this visual was that after the score card was introduced both index’s did their usual drop but high earning colleges made a sharper rebound than the lower earning colleges.

In order to make the regression I decided to change my data frame a little bit and add binary variables because they are very effective in comparing means. My regression model had both “high\_earners : screlease” and low\_earners : screlease, because I really only wanted to get the interaction between the two of them, because this would show me the true difference in high and low earners before and after the scorecard. When adding control variables I added the date because every year economic swings happen so I didn’t want that alter my results, then I added locale as well, because this is a measure of population of the city that the school is in and that could cause a positive bias for schools that are in bigger cities. I tried controlling for school size as well, but when I did it dropped the adjusted R Squared significantly so I decided to leave it out of my final model.

We can say that there is a statistically significant relationship between high and low earning groups and the release of the college scorecard because both interactions are significant at the 0.001 significance level. When comparing the two interactions, we can see that with a one unit increase of scorecard there is a 0.1003 unit increase to high earning colleges being searched and with a one unit increase of scorecard there is a 0.0695 unit increase on low earning colleges being searched, so long as all variables listed in the controls are contolled. This means that the introduction of the college scorecard did positively shift interest to high earning colleges.