Data Exploration Assignment Writeup

Jeremy Sternlicht

## Introduction

This analysis investigates the impact of the College Scorecard’s release in September 2015 on student interest in U.S. colleges that predominantly grant bachelor’s degrees. Specifically, this project aims to answer the research question: did the release of the Scorecard shift student interest to high-earnings colleges relative to low-earnings ones? Utilizing Google Trends data to proxy for student interest and College Scorecard data for graduate earnings, regression analysis is conducted to understand this dynamic.

## Methodology

I utilize linear regression models to analyze the standardized search index from Google Trends data. Three models are used in my analysis to address the research question by comparing the results between them. These models are a combined model considering all colleges while adjusting for locale differences and two separate models for colleges categorized as “high earnings” and “low earnings” based on their graduates’ median earnings, each model also adjusting for locale variations.

For graphical representation, scatter plots were generated to visualize the trends over time in search interest, with regression lines indicating the trend over time by locale category for each of the three regression models.

## Rationale of Analysis Design

This analysis is designed to answer the research question by directly comparing changes in search interest for “high earnings” and “low earnings” colleges before and after the College Scorecard’s release. The decision to segregate the data by earnings category is directly informed by the research question, which seeks to understand whether the Scorecard’s release shifted interest towards colleges with higher graduate earnings as opposed to those with lower graduate earnings. For the Scorecard data on graduate earnings, I decided to define “high earnings” colleges as anything where their graduates’ median earnings data is above the mean graduate earnings data for all colleges in the dataset. “Low earning” colleges are defined as any colleges whose median earnings data is below the mean earnings data for all colleges in the dataset. The inclusion of locale as a control variable in the regression models and the examination of its interaction with time periods (before and after release of Scorecard data) addresses potential regional variations that could confound the analysis. Additionally, this provides further insight into regional differences in the impact of the release of Scorecard data for the different earnings cat

The decision to produce scatter plots, accompanied by regression lines per locale, provides a visual complement to the regression analysis. This allows for an intuitive understanding of trends over time and across different locales.

## Code

# Data Exploration Assignment Part 2  
# Regression Models and Graphs  
  
# Load necessary libraries  
library(tidyverse)

Warning: package 'lubridate' was built under R version 4.3.2

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.3 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
✔ purrr 1.0.2   
── 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

library(rio)

Warning: package 'rio' was built under R version 4.3.2

library(ggplot2)  
  
# Import the cleaned data from the "Processed Data" directory  
final\_filtered\_data <- rio::import(file.path("../Processed Data", "final\_filtered\_data.csv"))  
  
# Regression model with interaction between LOCALE and time\_period  
model\_combined <- lm(standardized\_index ~ time\_period \* factor(LOCALE), data = final\_filtered\_data)  
summary(model\_combined)

Call:  
lm(formula = standardized\_index ~ time\_period \* factor(LOCALE),   
 data = final\_filtered\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.6749 -0.4961 -0.0586 0.4238 10.4012   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.073016 0.003402 21.463 < 2e-16 \*\*\*  
time\_period -0.389935 0.007862 -49.600 < 2e-16 \*\*\*  
factor(LOCALE)12 -0.021983 0.005719 -3.844 0.000121 \*\*\*  
factor(LOCALE)13 -0.034517 0.005704 -6.051 1.44e-09 \*\*\*  
factor(LOCALE)21 -0.004010 0.005119 -0.783 0.433389   
factor(LOCALE)22 -0.062502 0.010899 -5.735 9.78e-09 \*\*\*  
factor(LOCALE)23 -0.053877 0.013358 -4.033 5.50e-05 \*\*\*  
factor(LOCALE)31 -0.051296 0.009825 -5.221 1.78e-07 \*\*\*  
factor(LOCALE)32 -0.057689 0.006920 -8.336 < 2e-16 \*\*\*  
factor(LOCALE)33 -0.086420 0.007498 -11.526 < 2e-16 \*\*\*  
factor(LOCALE)41 -0.039978 0.010571 -3.782 0.000156 \*\*\*  
factor(LOCALE)42 -0.048750 0.015061 -3.237 0.001209 \*\*   
factor(LOCALE)43 -0.088958 0.018916 -4.703 2.57e-06 \*\*\*  
time\_period:factor(LOCALE)12 0.116894 0.013224 8.840 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)13 0.184420 0.013180 13.993 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)21 0.022130 0.011823 1.872 0.061245 .   
time\_period:factor(LOCALE)22 0.333852 0.025173 13.262 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)23 0.288164 0.030806 9.354 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)31 0.273633 0.022732 12.037 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)32 0.308173 0.015985 19.279 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)33 0.461096 0.017287 26.673 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)41 0.213764 0.024412 8.756 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)42 0.261846 0.034614 7.565 3.90e-14 \*\*\*  
time\_period:factor(LOCALE)43 0.475735 0.043877 10.842 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.7563 on 242348 degrees of freedom  
 (726 observations deleted due to missingness)  
Multiple R-squared: 0.02184, Adjusted R-squared: 0.02175   
F-statistic: 235.3 on 23 and 242348 DF, p-value: < 2.2e-16

# Filter for Low Earnings colleges  
low\_earnings\_data <- final\_filtered\_data %>%  
 filter(earnings\_category == "Low Earnings")  
  
# Regression model with interaction between LOCALE and time\_period for Low Earnings colleges  
model\_low\_earnings\_locale\_interaction <- lm(standardized\_index ~ time\_period \* factor(LOCALE), data = low\_earnings\_data)  
summary(model\_low\_earnings\_locale\_interaction)

Call:  
lm(formula = standardized\_index ~ time\_period \* factor(LOCALE),   
 data = low\_earnings\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.3950 -0.4963 -0.0605 0.4253 10.3403   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.076789 0.004637 16.559 < 2e-16 \*\*\*  
time\_period -0.410574 0.010723 -38.290 < 2e-16 \*\*\*  
factor(LOCALE)12 -0.023919 0.007937 -3.014 0.002581 \*\*   
factor(LOCALE)13 -0.039154 0.007743 -5.056 4.28e-07 \*\*\*  
factor(LOCALE)21 -0.002286 0.007247 -0.315 0.752452   
factor(LOCALE)22 -0.086156 0.014669 -5.874 4.27e-09 \*\*\*  
factor(LOCALE)23 -0.061630 0.021072 -2.925 0.003448 \*\*   
factor(LOCALE)31 -0.049300 0.011989 -4.112 3.92e-05 \*\*\*  
factor(LOCALE)32 -0.057453 0.008423 -6.821 9.09e-12 \*\*\*  
factor(LOCALE)33 -0.091678 0.008887 -10.315 < 2e-16 \*\*\*  
factor(LOCALE)41 -0.048650 0.012746 -3.817 0.000135 \*\*\*  
factor(LOCALE)42 -0.046761 0.016338 -2.862 0.004210 \*\*   
factor(LOCALE)43 -0.092731 0.019463 -4.764 1.90e-06 \*\*\*  
time\_period:factor(LOCALE)12 0.127803 0.018354 6.963 3.34e-12 \*\*\*  
time\_period:factor(LOCALE)13 0.210207 0.017880 11.756 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)21 0.014014 0.016735 0.837 0.402367   
time\_period:factor(LOCALE)22 0.460499 0.033869 13.596 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)23 0.329854 0.048631 6.783 1.18e-11 \*\*\*  
time\_period:factor(LOCALE)31 0.263065 0.027766 9.475 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)32 0.307512 0.019456 15.806 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)33 0.489488 0.020485 23.895 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)41 0.260567 0.029434 8.853 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)42 0.251816 0.037585 6.700 2.09e-11 \*\*\*  
time\_period:factor(LOCALE)43 0.496374 0.045145 10.995 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.7683 on 142486 degrees of freedom  
 (432 observations deleted due to missingness)  
Multiple R-squared: 0.02148, Adjusted R-squared: 0.02132   
F-statistic: 136 on 23 and 142486 DF, p-value: < 2.2e-16

# Filter for High Earnings colleges  
high\_earnings\_data <- final\_filtered\_data %>%  
 filter(earnings\_category == "High Earnings")  
  
# Regression model with interaction between LOCALE and time\_period for High Earnings colleges  
model\_high\_earnings\_locale\_interaction <- lm(standardized\_index ~ time\_period \* factor(LOCALE), data = high\_earnings\_data)  
summary(model\_high\_earnings\_locale\_interaction)

Call:  
lm(formula = standardized\_index ~ time\_period \* factor(LOCALE),   
 data = high\_earnings\_data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.7293 -0.4960 -0.0562 0.4217 10.4032   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.068304 0.004983 13.708 < 2e-16 \*\*\*  
time\_period -0.364223 0.011506 -31.654 < 2e-16 \*\*\*  
factor(LOCALE)12 -0.019305 0.008205 -2.353 0.018634 \*   
factor(LOCALE)13 -0.028693 0.008398 -3.416 0.000635 \*\*\*  
factor(LOCALE)21 -0.004512 0.007213 -0.626 0.531571   
factor(LOCALE)22 -0.031291 0.016229 -1.928 0.053848 .   
factor(LOCALE)23 -0.046426 0.017131 -2.710 0.006727 \*\*   
factor(LOCALE)31 -0.061220 0.017650 -3.469 0.000524 \*\*\*  
factor(LOCALE)32 -0.065526 0.012958 -5.057 4.27e-07 \*\*\*  
factor(LOCALE)33 -0.075707 0.015474 -4.892 9.97e-07 \*\*\*  
factor(LOCALE)41 -0.021784 0.019578 -1.113 0.265852   
factor(LOCALE)42 -0.099030 0.046796 -2.116 0.034329 \*   
time\_period:factor(LOCALE)12 0.101972 0.018969 5.376 7.65e-08 \*\*\*  
time\_period:factor(LOCALE)13 0.152045 0.019422 7.828 4.99e-15 \*\*\*  
time\_period:factor(LOCALE)21 0.023768 0.016659 1.427 0.153657   
time\_period:factor(LOCALE)22 0.166589 0.037499 4.442 8.90e-06 \*\*\*  
time\_period:factor(LOCALE)23 0.248008 0.039489 6.280 3.39e-10 \*\*\*  
time\_period:factor(LOCALE)31 0.326502 0.040731 8.016 1.10e-15 \*\*\*  
time\_period:factor(LOCALE)32 0.349368 0.029960 11.661 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)33 0.403799 0.035772 11.288 < 2e-16 \*\*\*  
time\_period:factor(LOCALE)41 0.115984 0.045224 2.565 0.010329 \*   
time\_period:factor(LOCALE)42 0.523999 0.106726 4.910 9.13e-07 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.7386 on 99840 degrees of freedom  
 (294 observations deleted due to missingness)  
Multiple R-squared: 0.02313, Adjusted R-squared: 0.02293   
F-statistic: 112.6 on 21 and 99840 DF, p-value: < 2.2e-16

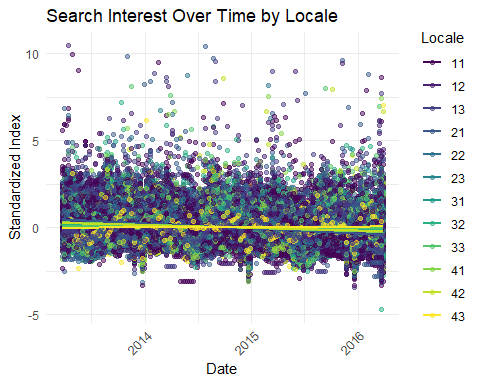
Then scatter plots were generated for each regression model using the following code:

# Graphing  
# Scatter plot with regression lines for combined dataset  
ggplot(final\_filtered\_data, aes(x = date, y = standardized\_index, color = factor(LOCALE))) +  
 geom\_point(alpha = 0.5) + # Adjust alpha for point transparency if needed  
 geom\_smooth(method = "lm", se = FALSE) + # Adds a linear regression line without confidence interval  
 scale\_color\_viridis\_d() + # Optional: Use a different color scale for clarity  
 labs(x = "Date", y = "Standardized Index", color = "Locale",  
 title = "Search Interest Over Time by Locale") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

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

Warning: Removed 726 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 726 rows containing missing values (`geom\_point()`).

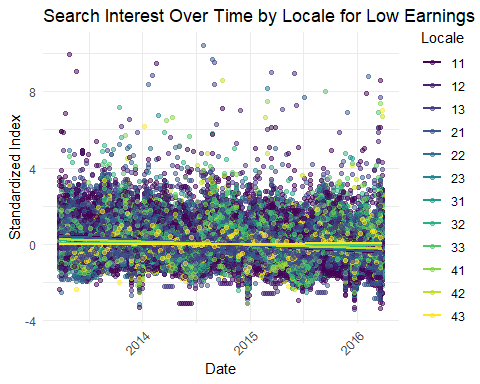


# Filter for Low Earnings colleges  
low\_earnings\_data <- final\_filtered\_data %>%  
 filter(earnings\_category == "Low Earnings")  
  
# Scatter plot with regression lines for Low Earnings colleges  
ggplot(low\_earnings\_data, aes(x = date, y = standardized\_index, color = factor(LOCALE))) +  
 geom\_point(alpha = 0.5) + # Adjust transparency with alpha  
 geom\_smooth(method = "lm", se = FALSE, aes(group = LOCALE)) + # Add regression line for each LOCALE  
 scale\_color\_viridis\_d() + # Use a discrete color scale  
 labs(x = "Date", y = "Standardized Index", color = "Locale",  
 title = "Search Interest Over Time by Locale for Low Earnings Colleges") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1), legend.position = "right")

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

Warning: Removed 432 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 432 rows containing missing values (`geom\_point()`).

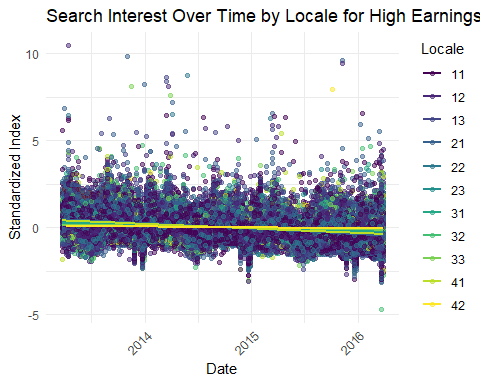


# Filter for High Earnings colleges  
high\_earnings\_data <- final\_filtered\_data %>%  
 filter(earnings\_category == "High Earnings")  
  
# Scatter plot with regression lines for High Earnings colleges  
ggplot(high\_earnings\_data, aes(x = date, y = standardized\_index, color = factor(LOCALE))) +  
 geom\_point(alpha = 0.5) + # Adjust transparency with alpha  
 geom\_smooth(method = "lm", se = FALSE, aes(group = LOCALE)) + # Add regression line for each LOCALE  
 scale\_color\_viridis\_d() + # Use a discrete color scale  
 labs(x = "Date", y = "Standardized Index", color = "Locale",  
 title = "Search Interest Over Time by Locale for High Earnings Colleges") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1), legend.position = "right")

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

Warning: Removed 294 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 294 rows containing missing values (`geom\_point()`).



## Results

The analysis conducted across three distinct regression models offers nuanced insights into how the introduction of the College Scorecard influenced Google search interest in colleges, with particular attention to differences between high-earning and low-earning colleges and the impact of locale variations.

The combined model indicates a general decrease in search interest post-Scorecard release, with a significant coefficient for ‘time\_period’ (-0.389935, SE = 0.007862, p < 2e-16), suggesting a broad decline across all colleges.

The model for “low earnings” colleges further shows the impact of the release of the Scorecard data, showing a slightly more significant decrease in search interest (‘time\_period’ coefficient = -0.410574, SE = 0.010723, p < 2e-16) compared to the combined model. This suggests that low earnings colleges experienced a slightly steeper decline in Google search interest post-Scorecard release.

Conversely, high earnings colleges exhibited a smaller decrease in search interest (‘time\_period’ coefficient = -0.364223, SE = 0.011506, p < 2e-16) compared to the combined model, suggesting that the Scorecard’s release may have shifted public interest slightly towards these institutions.

## Conclusion

Since the release of the Scorecard data in September 2015, the overall level of student interest in colleges that predominantly grant bachelor’s degrees (as measured by Google Trends data) decreased by 0.389935 units of the standardized index. Using this as a benchmark, student interest in “low earnings” colleges decreased by a larger amount, 0.410574 units, since the release of the Scorecard data. In comparison, student interest in “high earnings” colleges only decreased by 0.364223 units since the release of the Scorecard data, which is a lesser amount than the overall decrease of all colleges in this dataset and a significantly lesser amount than the decrease in student interest for “low earnings” colleges. These results come from the ‘time\_period’ coefficient in each of the three regression models. Comparing this coefficient across models reveals if the Scorecard possibly had a more substantial impact on one earnings category of colleges over the other, and the results suggest that the release of the Scorecard data may have shifted student interest more towards high-earnings colleges compared to low-earnings ones.

In real world terms, these findings highlight the Scorecard’s potential role in shaping perceptions of value in higher education, suggesting that students may increasingly seek out institutions that offer better earnings outcomes and a better return on their educational investment. This analysis underscores the need for further research into the multifaceted factors that influence student interest in specific colleges after the release of public data like the College Scorecard.

Additionally, the statistically significant locale-specific trends observed across all models highlight the importance of geographic context in understanding the impact of public information like the College Scorecard in shaping public interest and behavior concerning college education choices.