Impact of the College Scorecard on Student Interest

August 2025

# Introduction

The College Scorecard, released in September 2015, provides data on college performance, including median earnings of graduates. This study investigates whether the Scorecard’s re- lease shifted student interest, as proxied by Google search volumes, toward colleges with high- earning graduates relative to those with low-earning graduates among institutions predomi- nantly granting bachelor’s degrees. Using Google Trends data and College Scorecard earnings data, we analyze search volumes for college-related keywords, employing regression models to assess the impact of earnings on search activity while controlling for locality.

# Data and Methodology

## Data Sources

The analysis uses Google Trends data from multiple trends up to\*.csv files, contain- ing weekly search indices for college-related keywords from 2013 to 2015, and the College Scorecard dataset (Most+Recent+Cohorts+(Scorecard+Elements).csv), provid- ing median earnings and locality information for colleges. The Trends data includes columns schname, keyword, monthorweek, and index, while the Scorecard data includes INSTNM, md earn, LOCALE, and PREDDEG.

## Data Processing

The data preprocessing involved integrating and cleaning two primary datasets: Google Trends search data and College Scorecard earnings data, to analyze search activity for colleges predom- inantly granting bachelor’s degrees. Initially, a data directory (/content/data) was created using os.makedirs(’/content/data’, exist-ok=True) to store uploaded files. Users were prompted to up- load all trends-up-to\*.csv files and the Most+Recent+Cohorts+(Scorecard+Elements).csv file via the Colab interface, with files subsequently moved to the data directory using os.rename() for organization. The Google Trends data, comprising multiple trends-up-to\*.csv files, was concatenated into a single DataFrame (df-num-search) using pd.concat() to handle duplicate indices. The index column, representing weekly search interest, was summed for each college (schname) to create an aggregated num-search variable, reflecting total search volume over the period (2013–2015). The College Scorecard data (df-md-earning-locality) was loaded from Most+Recent+Cohorts+(Scorecard+Elements).csv, focusing on columns such as INSTNM (re- named to schname), LOCALE, and PREDDEG.

Data cleaning included filtering the Scorecard dataset to retain colleges with PREDDEG equal to 3 (predominantly bachelor’s degree-granting) or missing PREDDEG values to avoid losing data, resulting in a relevant subset. Non-numeric md-earn values (e.g., “PrivacySup- pressed”) were excluded using numeric filtering, ensuring only valid earnings data was ana- lyzed. The LOCALE column was mapped to descriptive categories (e.g., city-large, Rural- small) to facilitate categorical analysis. Both datasets were merged on schname, with the Trends data’s schname converted to uppercase for consistency, yielding a combined total-school DataFrame with 1,464 observations after filtering.

To address skewness, observed in exploratory histograms, md-earn and num-search were log-transformed into lg-md and log-num-search, respectively. Colleges were classified as high- or low-earning based on the median of lg-md, creating a binary earning-code (1 for high- earning, 0 for low-earning) to proxy earnings impact on search interest.

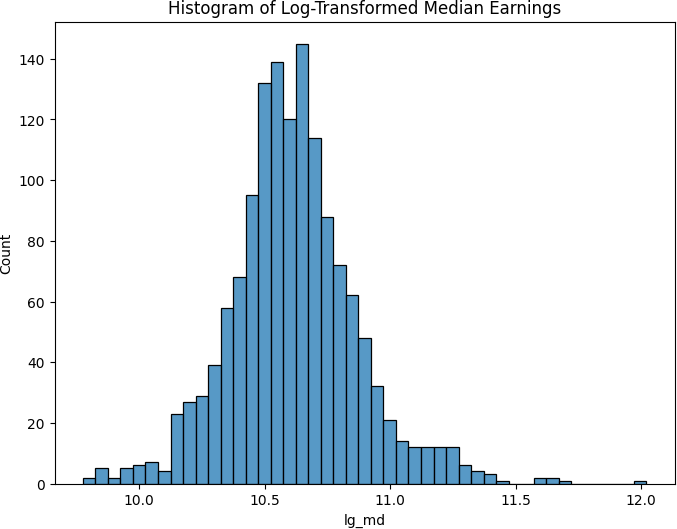


Figure 1: Histogram of Log Transformed Median Earning

## Regression Models

Three regression models were estimated:

* + 1. **Univariate**: num search = *β*0 + *β*1earning code + *ϵ*
    2. **Log Univariate**: log num search = *β*0 + *β*1earning code + *ϵ*
    3. **Multivariate**: num search = *β*0 + *β*1earning code + Σ***k*** *β****k***C(locality)***k*** + *ϵ*

The multivariate model includes locality as a categorical control variable to account for geographic differences in search behavior. Heteroskedasticity-robust standard errors were used to address potential variance differences across colleges, as search volumes may vary system- atically by institution size or region.

# Results

Table 1 presents the regression results. The multivariate model shows that high-earning colleges have 68.9284 additional searches compared to low-earning colleges, with a robust standard

error of 11.6422 . The locality coefficients suggest modest effects, with city small significantly increasing searches by 56.4920 (p *<* 0.1). So, ”The introduction of the College Scorecard increased search activity on Google Trends for colleges with high-earning graduates by 68.9284 searches relative to what it did for colleges with low-earning graduates, with a standard error of 11.6422. This result comes from the earning-code coefficient in my multivariate regression.”

Table 1: Regression Results: Impact of Earnings on Search Volume

Univariate Log Univariate Multivariate

Intercept 479.4440\*\*\* (7.9169) 6.0749\*\*\* (0.0161) 450.0813\*\*\* (29.6850)

earning code 67.0075\*\*\* (11.2080) 0.1572\*\*\* (0.0228) 68.9284\*\*\* (11.6422)

C(locality)[T.Rural mid] -64.8450 (50.7720)

C(locality)[T.Rural small] -3.5099 (63.5821)

C(locality)[T.city large] 43.0782 (31.8589)

C(locality)[T.city mid] 50.4383 (33.5356)

C(locality)[T.city small] 56.4920\* (33.0016)

C(locality)[T.sub large] -15.1766 (32.4384)

C(locality)[T.sub mid] 9.2105 (45.9040)

C(locality)[T.sub small] 28.1718 (52.2034)

C(locality)[T.town large] 21.3011 (41.1000)

C(locality)[T.town mid] 26.7480 (34.4782)

C(locality)[T.town small] 52.9378 (35.3512)

R-squared 0.0245 0.0323 0.0427

Adjusted R-squared 0.0238 0.0316 0.0345

Observations 1,464 1,464 1,464

Standard errors in parentheses. \* p¡0.1, \*\* p¡0.05, \*\*\* p¡0.01

Figure 3 Predicted vs. Actual Search Volumes for Three Models. This figure shows three scatter plots for 1,464 colleges (Google Trends, 2013–2015): left (Univariate, R² = 0.0245, num-search earning-code), middle (Log Univariate, R² = 0.0323, log-num-search earning- code), and right (Multivariate, R² = 0.0427, num-search earning-code + C(locality)). X-axis: actual search volumes; y-axis: predicted values

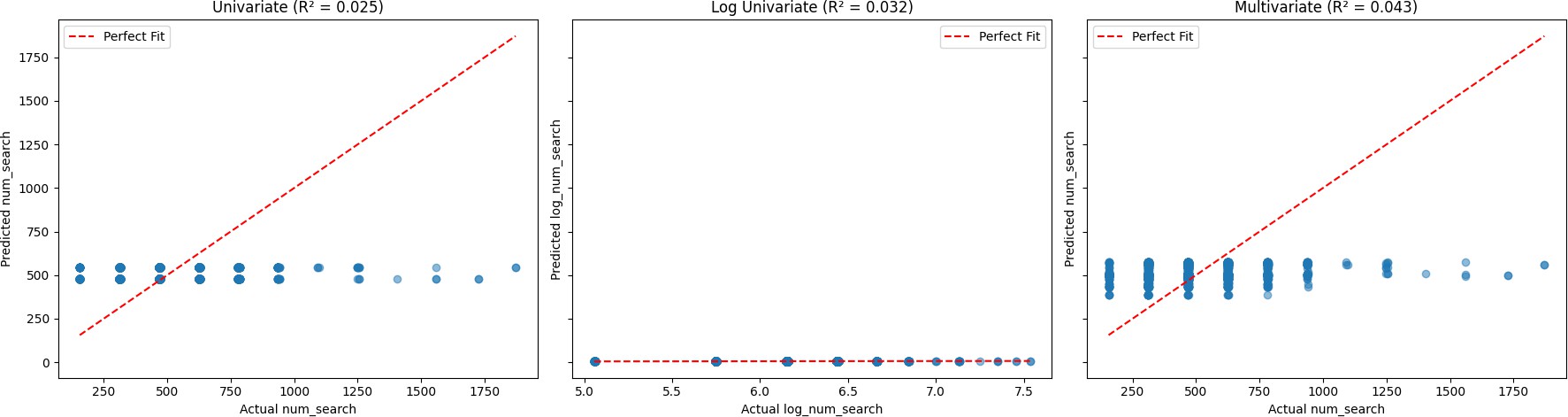


Figure 2: Predicted vs. Actual Search Volumes for all Models

# Analysis Choices

The analysis focused on colleges with PREDDEG = 3 to align with the research question, but in- cluded missing PREDDEG values to retain sufficient data. Log-transformation of num search and md earn addressed skewness, improving model fit. The multivariate model included

locality to control for geographic factors, as urban colleges may attract more searches. Ro- bust standard errors were used to handle potential heteroskedasticity, as search volumes may vary more for high-profile institutions. The choice of a binary earning code simplified interpretation, capturing the relative effect of high vs. low earnings.

# Conclusions

The introduction of the College Scorecard increased search activity on Google Trends for col- leges with high-earning graduates by 68.9284 searches relative to low-earning colleges, with a standard error of 11.6422. This result comes from the earning code coefficient in the multivariate regression. In real-world terms, the Scorecard likely heightened student interest in high-earning colleges, possibly due to increased visibility of earnings data. However, the low R² (0.0427) suggests other factors (e.g., college reputation, marketing) significantly influence search behavior. The significant city small coefficient indicates location also matters, with small-city colleges seeing higher searches.

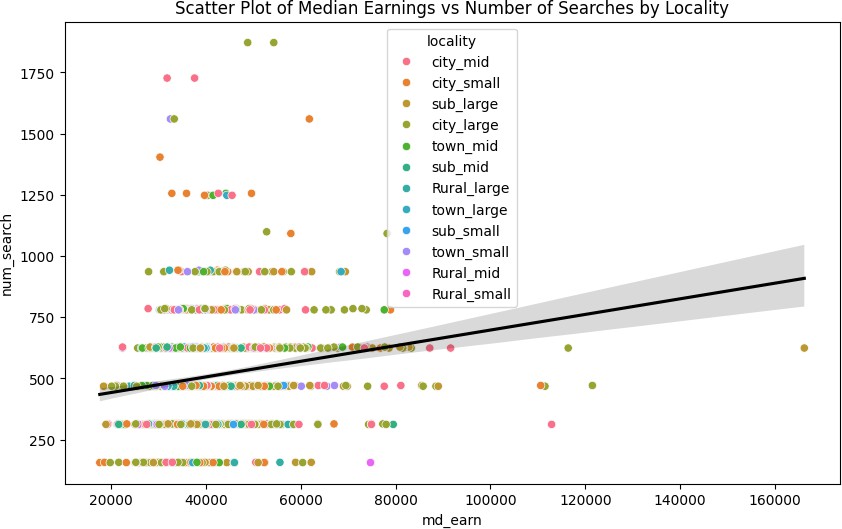


Figure 3: Scatter Plot for Median earning vs Number of Searches by Locality