Impact of the College Scorecard on Student Interest

August 2025

1 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 release shifted student interest, as proxied by Google search volumes, toward colleges with high-earning graduates relative to those with low-earning graduates among institutions predominantly 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.

2 Data and Methodology

2.1 Data Sources

The analysis uses Google Trends data from multiple `trends_up_to*.csv` files, containing weekly search indices for college-related keywords from 2013 to 2015, and the College Scorecard dataset (`Most+Recent+Cohorts+(Scorecard+Elements).csv`), providing 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`.

2.2 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 predominantly 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 upload 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 (renamed 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., "PrivacySuppressed") were excluded using numeric filtering, ensuring only valid earnings data was analyzed. 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.

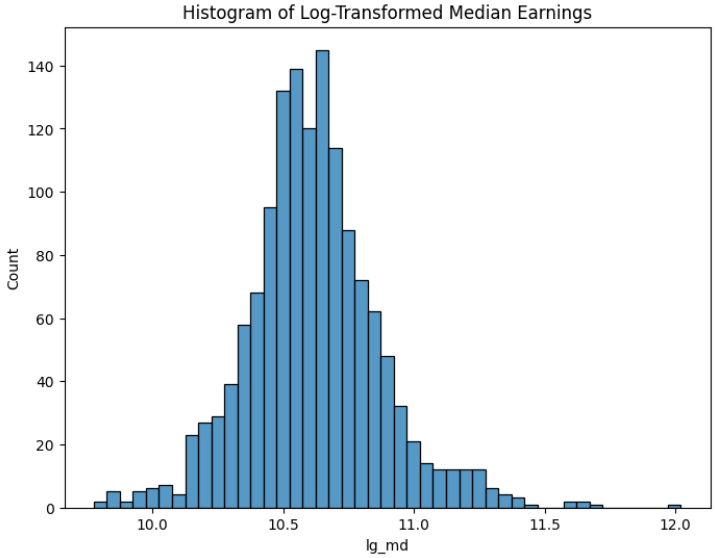

Figure 1: Histogram of Log Transformed Median Earning

2.3 Regression Models

Three regression models were estimated:

1. **Univariate**: $\text{num\_search} = \beta_0 + \beta_1 \text{earning\_code} + \epsilon$
2. **Log Univariate**: $\text{log\_num\_search} = \beta_0 + \beta_1 \text{earning\_code} + \epsilon$
3. **Multivariate**: $\text{num\_search} = \beta_0 + \beta_1 \text{earning\_code} + \sum_k \beta_k \text{C(locality)}_k + \epsilon$

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 systematically by institution size or region.

3 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, $R^2 = 0.0245$, num-search earning-code), middle (Log Univariate, $R^2 = 0.0323$, log-num-search earning-code), and right (Multivariate, $R^2 = 0.0427$, num-search earning-code + C(locality)). X-axis: actual search volumes; y-axis: predicted values

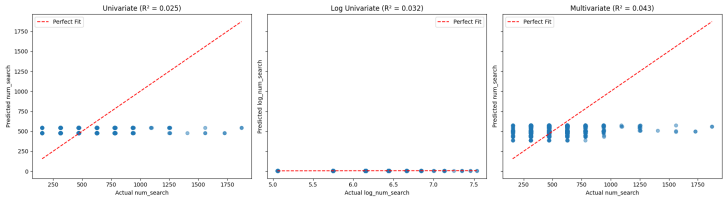

Figure 2: Predicted vs. Actual Search Volumes for all Models

4 Analysis Choices

The analysis focused on colleges with `PREDDEG` = 3 to align with the research question, but included 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. Robust 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.

5 Conclusions

The introduction of the College Scorecard increased search activity on Google Trends for colleges 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 R^2 (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.

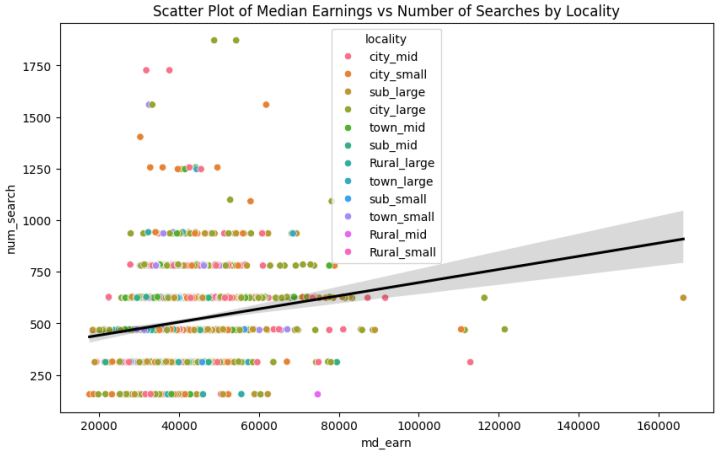

Figure 3: Scatter Plot for Median earning vs Number of Searches by Locality