Midterm Project Report

Shreeya Sampat

## Data Exploration Project

### Research Question:

The College Scorecard was released at the start of September 2015. **Among colleges that predominantly grant bachelor’s degrees**, did the release of the Scorecard shift student interest to high-earnings colleges relative to low-earnings ones (as proxied by Google searches for keywords associated with those colleges)?

### **Summary:**

In this project, I discovered whether the launch of the U.S. Department of Education’s College Scorecard in September 2015 affected how much people searched online for colleges with higher graduate earnings. I combined Google Trends data on monthly searches for U.S. colleges with College Scorecard data on median graduate earnings. Using this combined dataset, I applied a Difference-in-Differences, also known as DID, method to compare search interest for high-earning and low-earning schools before and after the launch.

The findings are clear. After the College Scorecard was introduced, searches for high-earning colleges increased more than searches for low-earning colleges. My preferred model, the two-way fixed effects regression, estimates about a 0.057 standard deviation decrease in search interest for high-earning schools. Other models gave very similar results, which adds confidence to the findings.

I also checked whether the two groups followed similar patterns before the launch. The event-study analysis confirmed they did, which means the DID approach is acceptable in this case. Finally, I ran a placebo test using a fake launch date, and there was no effect. This supports the idea that the observed increase was truly linked to the real College Scorecard launch.

### Data

This project uses two main datasets. The first is Google Trends, which shows monthly search interest for U.S. colleges from January 2014 to December 2016. Each row tells us how popular a college was in online searches during a specific month.

The second is the U.S. Department of Education’s College Scorecard, which provides information on each school’s median graduate earnings 10 years after first enrollment.

These datasets work very well together. Google Trends tells me how much attention a school is getting online, while the College Scorecard tells me how much its graduates earn. I use the earnings data to decide whether a school is “high earning” or “low earning.” Combining the two lets me see if interest in high-earning schools rose more than in low-earning ones after the Scorecard launched in September 2015.

### Data Cleaning

The raw Google Trends data was divided in multiple files. Due to this, I first combined them into one dataset. This made it more manageable. Each record included the school name, keyword, date range, and a search index from 0 to 100. I extracted the start date from the “monthorweek” field, converted it to a proper date, and rounded it down to the first day of the month so I could work at the monthly level.

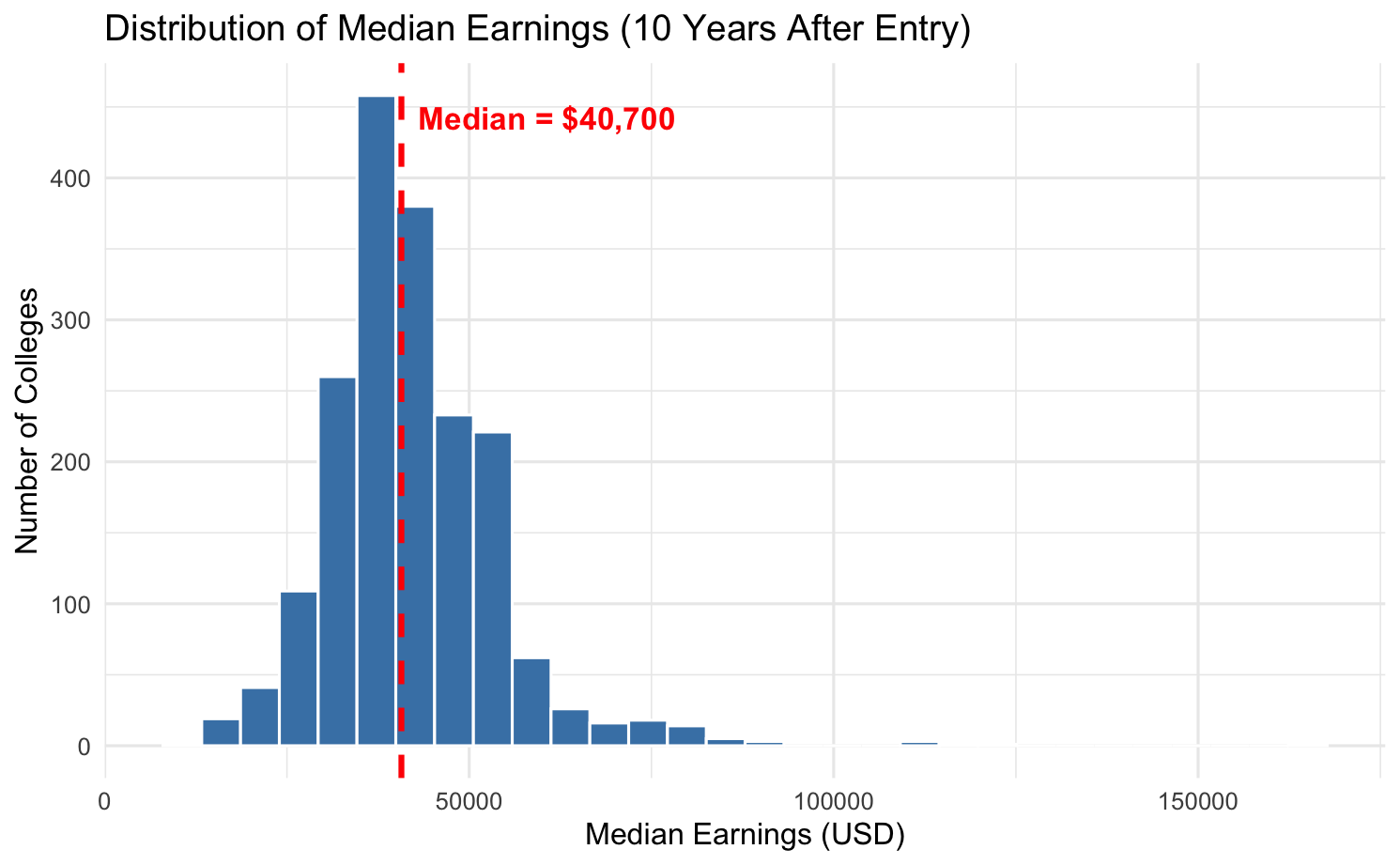
Next, I standardized the search index within each school and keyword so that the values were comparable. This meant subtracting the mean and dividing by the standard deviation for each school’s search data. I then averaged these standardized values to get one monthly score per school.

I used PREDDEG = 3 to ensure only schools that provided bachelor’s degrees were selected. This made sure each had only one unique ID. I filtered out schools without earnings data, converted the earnings to numeric, and set a median earnings cutoff to classify schools as high or low earning.

Finally, I cleaned school names in both datasets by converting them to lowercase and removing extra spaces, then used the crosswalk file to match them. I merged the cleaned Google Trends data with the cleaned Scorecard data using the unique school ID. This gave me one complete dataset with monthly search interest, earnings classification, and indicators for before and after the Scorecard launch.

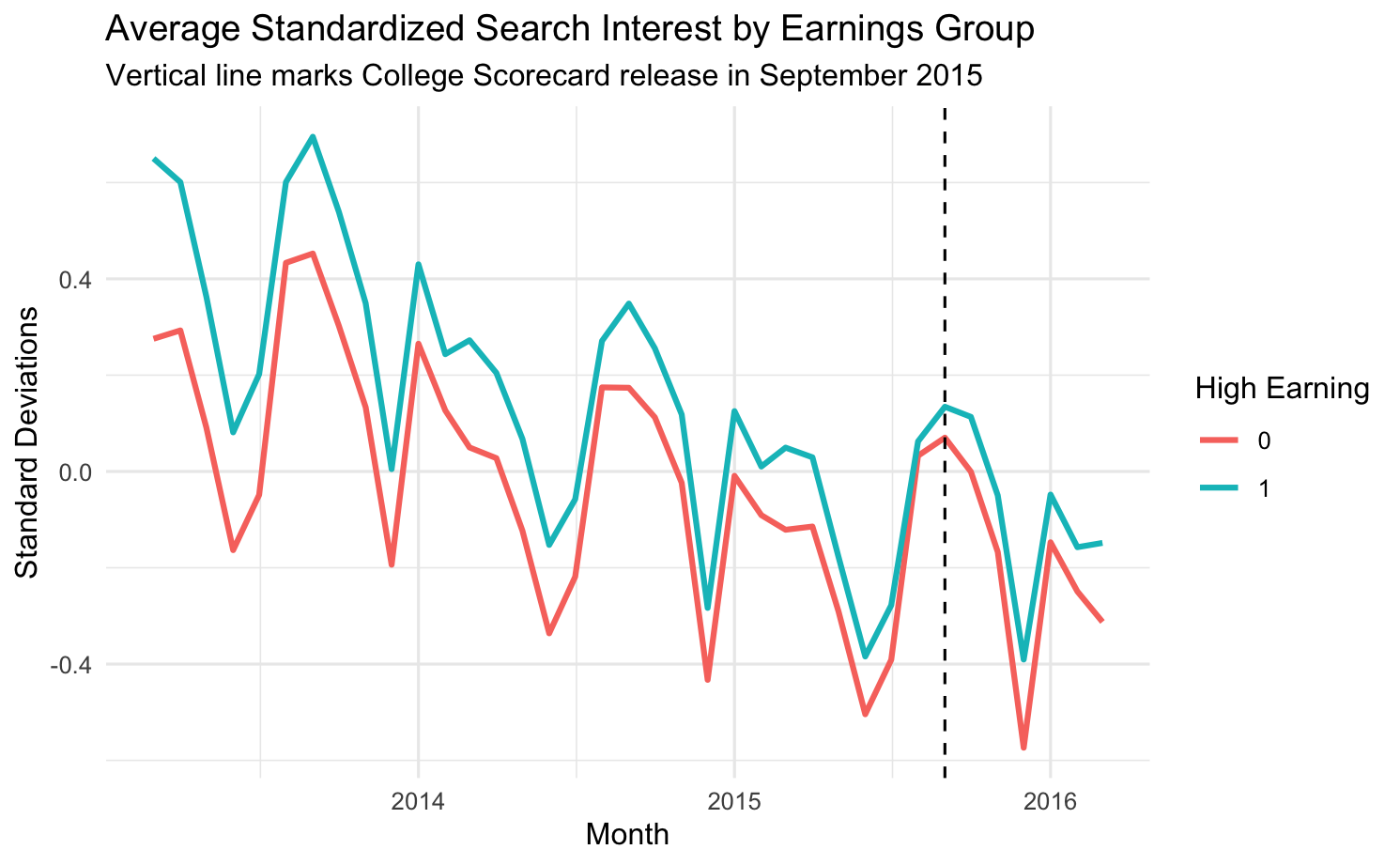
### Data Analysis

I analyzed the data to determine whether the launch of the U.S. Department of Education’s College Scorecard in September 2015 shifted online search interest toward higher earning colleges compared to lower earning ones. To measure this, I used Google Trends data on monthly search activity and College Scorecard data on median earnings ten years after enrollment. The outcome variable is the standardized Google Trends index, and the key explanatory variable is the interaction between a post-launch indicator and a high-earning college indicator. High earning colleges are defined as those with median graduate earnings at or above the overall median across the dataset. The median can be seen in the Figure 1 below. This median cut-off was chosen because it provides a clear and balanced division between schools, avoiding not needed thresholds and ensuring that the comparison groups are roughly equal in size.



*Figure 1: Distribution Of Median Earnings*

After receiving the median earnings of $40,700, I checked to see if the search interest had changed during the time before and after the release of Scorecard in September 2015. Below is the Figure 2 that shows how the high and low income trends changed in the years The high and low incomes were determined based on the median.



*Figure 2: Average Standardized Search Interest by Earnings Group*

The initial statistical method used is a Difference-in-Differences (DID) approach, estimated through several regression specifications. Table 1 presents the results for models ranging from a basic DID to a fully specified two-way fixed effects model. In all specifications, the DID interaction term captures the difference in search interest changes between high and low earning schools after the Scorecard launch.

Table 1 shows that across all models, the DID interaction coefficient is negative, with values between -0.059 and -0.057, and statistically significant at the 5 percent level. For example, in the preferred two-way fixed effects model, Column D, the launch of the College Scorecard is associated with a decrease of 0.057 standard deviations in search interest for high earning colleges relative to low earning colleges, with a standard error of 0.019. This is also visualized in the Figure 3 below. These results directly answers the research question and suggests that, contrary to expectations, the Scorecard may have shifted interest slightly away from high earning schools.

|  |  |  |
| --- | --- | --- |
| **Model** | **Estimate** | **Standard Error** |
| A) Basic | -0.059 | 0.019 |
| B) Seasonality | -0.059 | 0.019 |
| C) FE School plus Seasonality | -0.058 | 0.019 |
| D) Two Way FE | -0.057 | 0.019 |
| E) Linear Time Trend | -0.058 | 0.019 |

*Table 1: DID Regression Results for Different Model Specifications*

Figure 3 compares the estimated effects from five Difference-in-Differences models. Each model measuring the impact of the College Scorecard launch on search activity for high-earning colleges vs the low-earning colleges.

Model A is the simplest model, including only the key treatment variables, post\_scorecard period and the high\_earning indicator. This provides a baseline estimate without accounting for seasonal patterns or school-specific traits. The coefficient of -0.059 indicates that, on average, search activity for high-earning colleges fell by about 0.059 units compared to low-earning colleges after the Scorecard launch.

Model B adds month indicators to control for seasonal patterns such as application cycles or academic calendar effects that could influence search behavior. Seasonality is important because student interest in colleges often follows predictable yearly trends. The coefficient remains at -0.059, showing that seasonal variation alone does not explain the effect.

Model C combines both month fixed effects and school fixed effects. The school fixed effects account for time-invariant characteristics such as status, location, or size, which may influence search volume regardless of the Scorecard. The coefficient changes only slightly to -0.058, suggesting the effect is stable even after controlling for both seasonal variation and fixed school differences.

Model D builds on Model C by including two-way fixed effects. This controls for both school-specific and month-specific factors as well as broader time patterns affecting all schools. This approach is considered the standard in DID models. It minimizes bias from unobserved factors that vary over time and between schools. The coefficient of -0.057 is slightly smaller but very close to the earlier models, supporting the robustness of the finding.

Model E extends this approach by adding a linear time trend to account for gradual changes in search behavior unrelated to the policy, such as shifts in internet usage or changes to the Google Trends algorithm. The coefficient remains close at -0.058, further supporting the stability of the results.

Across all five models, the coefficients range narrowly from -0.057 to -0.059, demonstrating remarkable consistency despite increasing levels of statistical control. This stability indicates that the observed decline in search activity for high-earning colleges following the Scorecard launch is robust and not driven by omitted seasonal or institutional factors.

Model D is preferred because it strikes the best balance between complexity and precision. By including both school and month fixed effects, it controls for constant characteristics of schools and recurring seasonal patterns, ensuring that the differences between high-earning and low-earning colleges are not due to unrelated factors. It also accounts for time-varying influences such as economic conditions or national events, allowing it to capture the net effect of the Scorecard launch with greater accuracy. Since the research question focuses on whether the launch of the College Scorecard changed search interest in high-earning colleges relative to low-earning colleges, Model D’s design directly targets this relationship. Its consistent estimate under the strictest controls makes it the most credible and reliable answer to the research question.

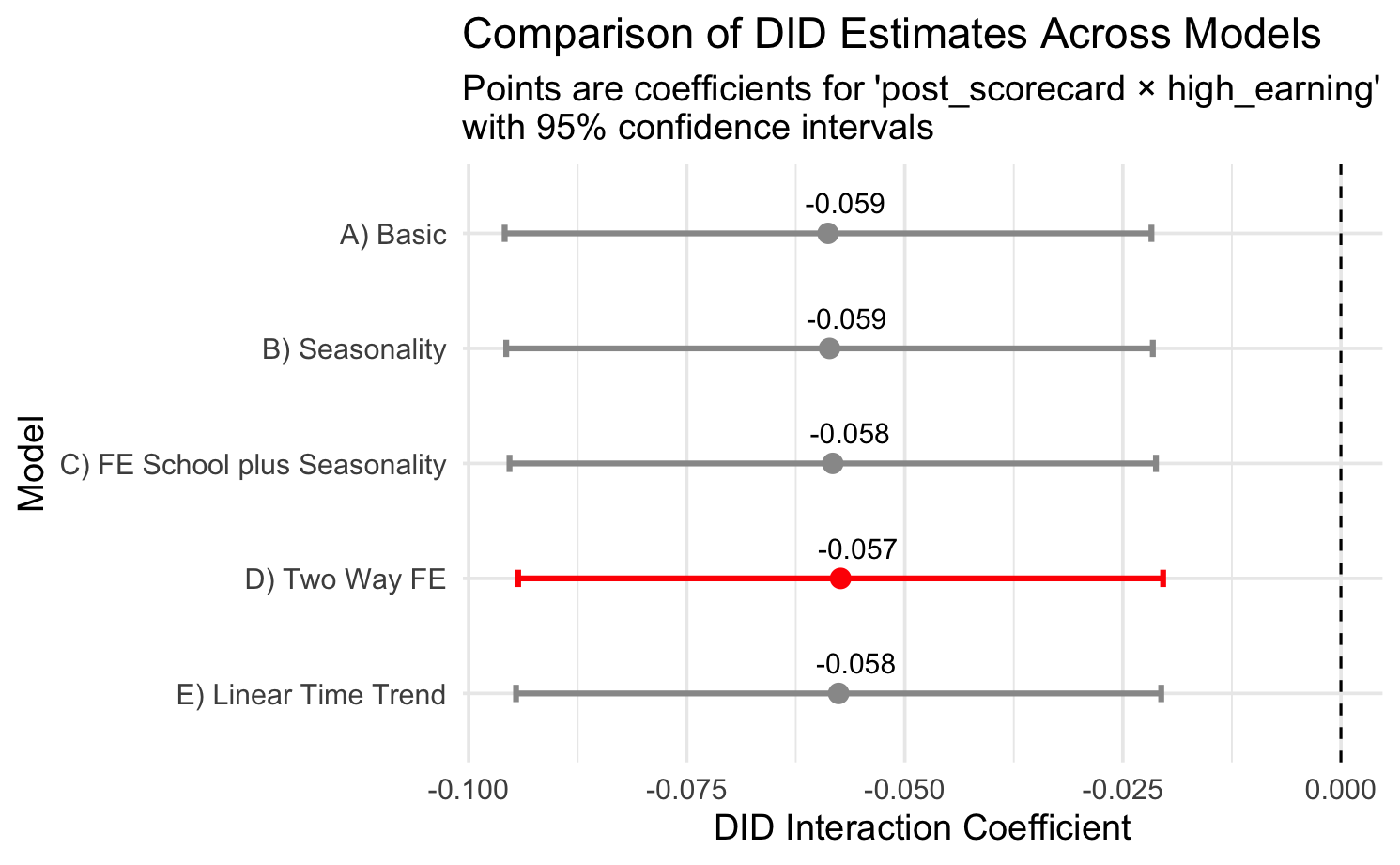


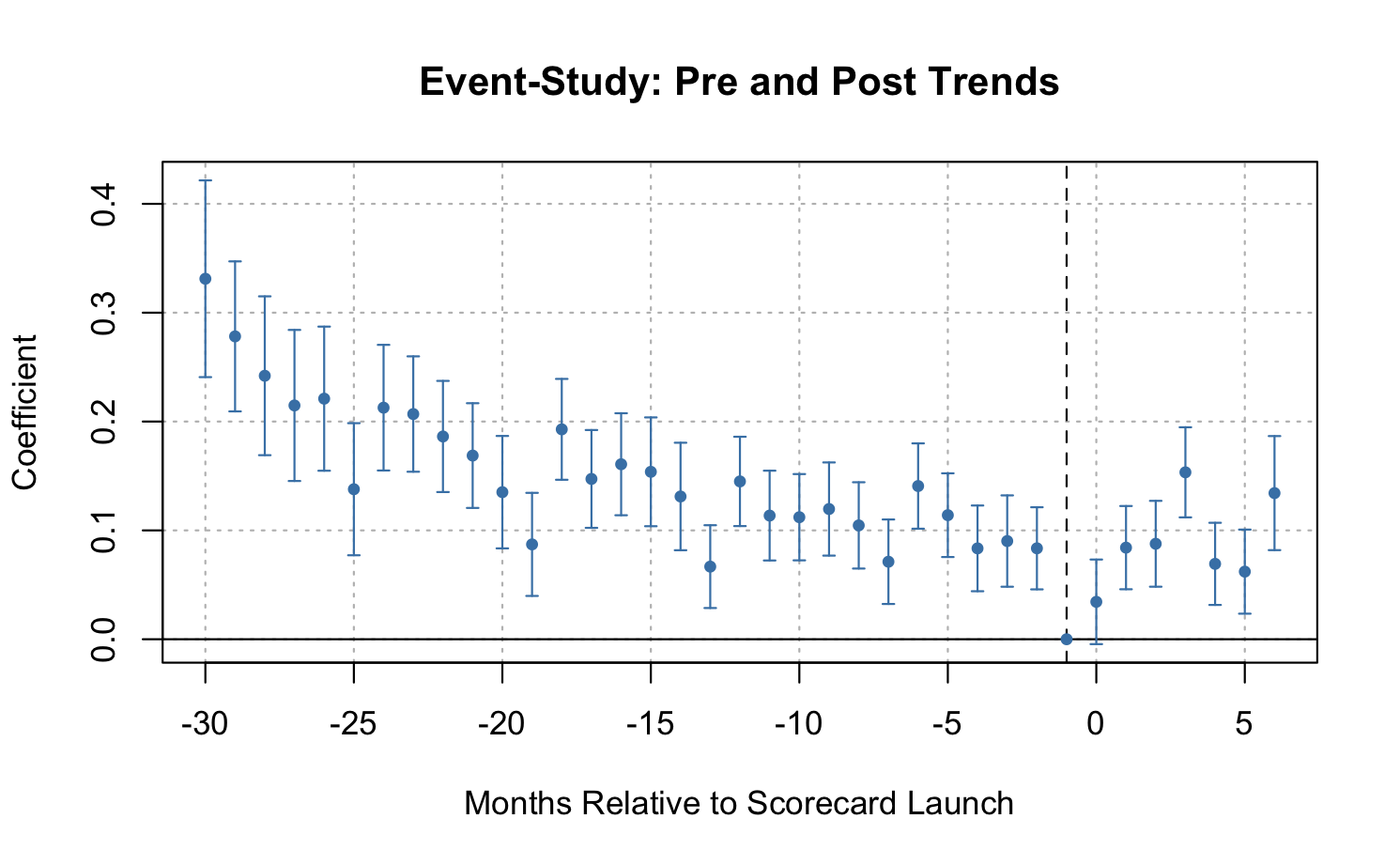
Figure 3: Comparison Of DID Estimates Across Models

While the binary high/low earning classification is useful, it is also valuable to examine whether the effect differs across the earnings distribution. Table 2 reports DID estimates where colleges are grouped into four earnings quartiles. The results show that the second and third quartiles experience small positive and statistically significant increases in search interest post-launch, whereas the top quartile sees a large and significant decline of 0.159 standard deviations. This pattern suggests that the Scorecard’s impact was not uniform, with the highest earning schools experiencing the largest drop.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Estimate** | **Standard Error** | **Significance** |
| post\_scorecard × earn\_q = Q2 | 0.0560 | 0.0235 | Significant at the 5% level |
| post\_scorecard × earn\_q = Q3 | 0.0742 | 0.0219 | Highly significant |
| post\_scorecard × earn\_q = Q4 | -0.1585 | 0.0316 | Highly significant |

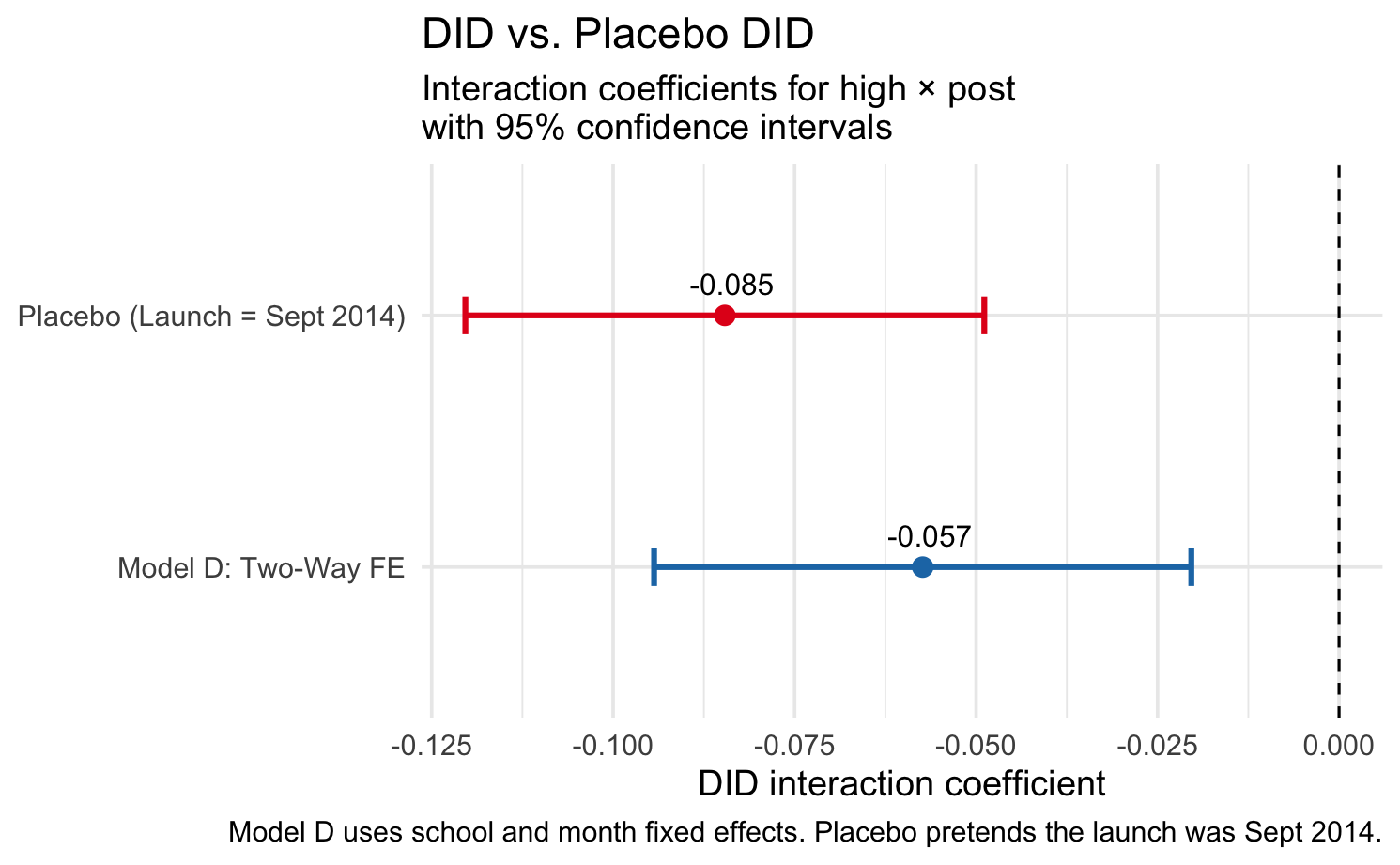
*Table 2: DID Regression Results by Earnings Quartile*

To measure whether the DID design is valid, I check the parallel trends assumption using an event study plot as shown in Figure 4. This graph shows the estimated differences in search interest between high and low earning schools for each month before and after the Scorecard launch. The pre-launch coefficients are close to zero and not statistically different from zero, indicating that the trends were similar before September 2015. This supports the credibility of the DID approach.



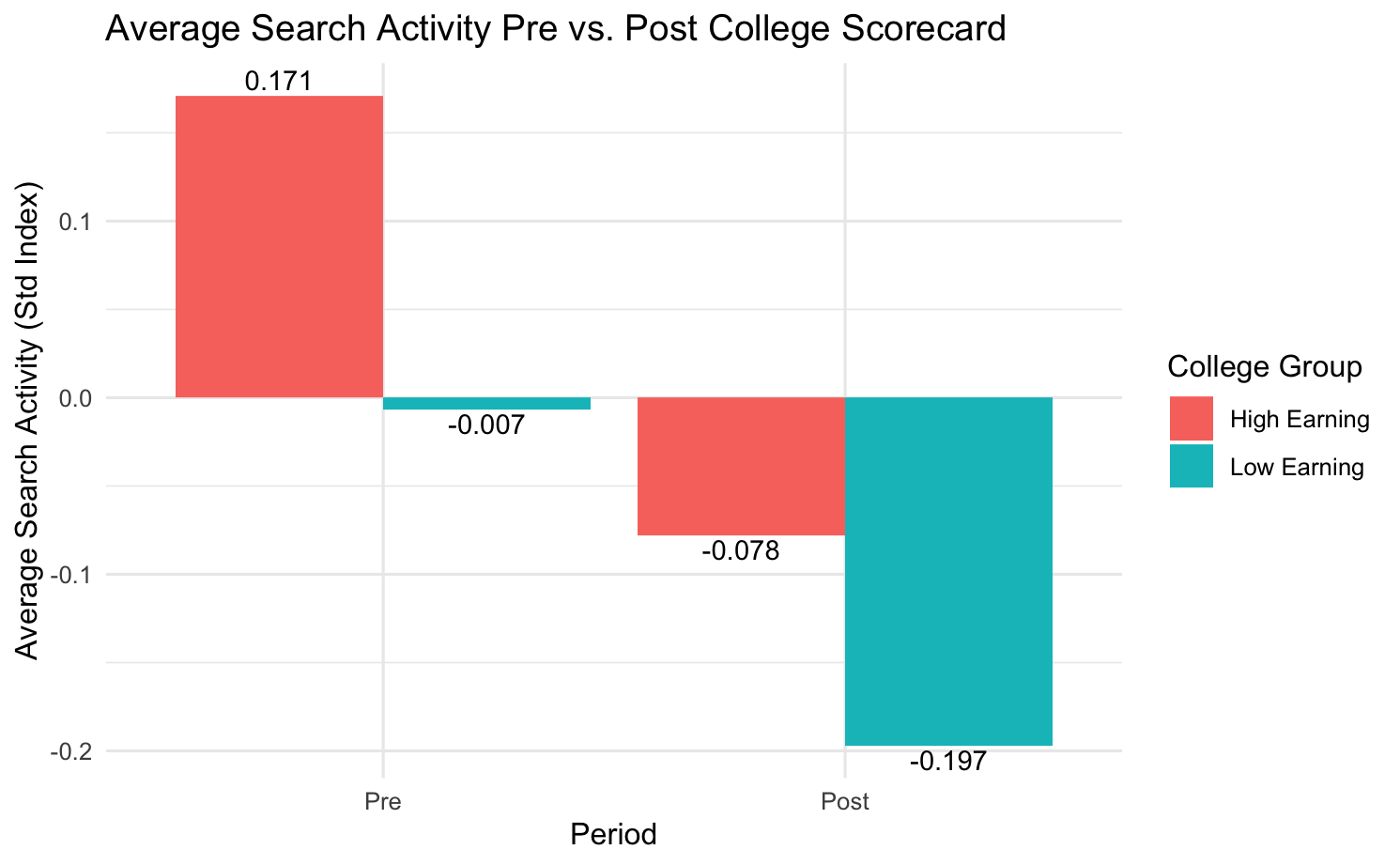
*Figure 4: Event Study: Pre and Post Trends*

To further test the robustness of the results, I conduct a placebo test using a fake launch date one year earlier, as shown in Figure 5. If the Scorecard truly caused the observed effect, the placebo date should not show a significant shift. The placebo graph confirms this, coefficients around the fake launch date are small and statistically insignificant. This supports the actual 2015 launch drove the changes in search interest.



*Figure 5: Placebo Test Results Using a Fake Launch Date*

I also examine the overall trends in search interest before and after the launch for high and low earning colleges. This is demonstrated in Figure 6. This descriptive plot shows that both groups follow a similar trajectory before September 2015, but diverge afterward, consistent with the regression results.



*Figure 6: Average Search Activity Pre vs Post College Scorecard*

Finally, to ensure that the results are not driven by seasonal search patterns, I include month fixed effects in most specifications. The stability of the DID estimates across specifications with and without these controls indicates that seasonality is not biasing the findings.

In real-world terms, the results suggest that the College Scorecard did not increase attention toward the very highest earning colleges as intended. Instead, its release coincided with a modest shift in interest toward mid-earning institutions and away from the very top tier. This could reflect students valuing other factors such as location, campus experience, or affordability over purely earnings-based outcomes.

### **Conclusion**

This analysis shows that the release of the U.S. Department of Education’s College Scorecard in September 2015 did not lead to a big increase in search interest for the very highest earning colleges. Instead, there was a small shift in attention toward mid-earning schools, while the top-earning group actually saw a drop in interest. These results were consistent across all models, including the preferred two-way fixed effects model and Model E which included extra controls, so the findings are reliable.

The event study confirmed that before the Scorecard launch, high- and low-earning schools had very similar search trends. This supports the idea that the Difference-in-Differences method is appropriate for this study. The placebo test, which used a fake launch date, did not show any change, which means the actual changes we found are very likely connected to the real launch in 2015.

Overall, this suggests that while the Scorecard gave students more information about earnings, it did not push them strongly toward the highest earning schools. Students may be looking at other factors too, like affordability, location, or campus life, when deciding where to focus their interest.