Tuition and Salary Analysis

About: The data project aims to analyze the relationship between college tuition costs and post-college earnings, using datasets that explore factors such as geographic location, college degree type, and potential salary impact on tuition costs and earning potential. The project seeks to shed light on the complex relationship between education and income in the US and could have significant implications for policymakers, educators, and individuals seeking to make informed decisions about their education and career paths.

Data Creation Range:
January 2014 - 2020
Created By:
Sammy Bharadwaj, Aryan Palave, Zhikai Li
Published: 5/7/23
Source:
https://www.kaggle.com/datasets/wsj/college-salaries,
https://www.kaggle.com/datasets/jessemostipak/college-tuition-diversity-and-pay

Use Cases

1. How does the cost of college tuition vary by geographic location, degree type, and income bracket?
2. What is the relationship between college tuition costs and post-college earnings?
3. How do factors such as college degree type, geographic location, and potential salary impact the cost of tuition and the earning potential of graduates?
4. What are the potential societal implications of rising college tuition costs?
5. What insights can be gained from the data on post-college salaries, including variations by college type, major, and region?
6. How can policymakers, educators, and individuals use the findings of this dataset to make informed decisions about higher education and career paths?

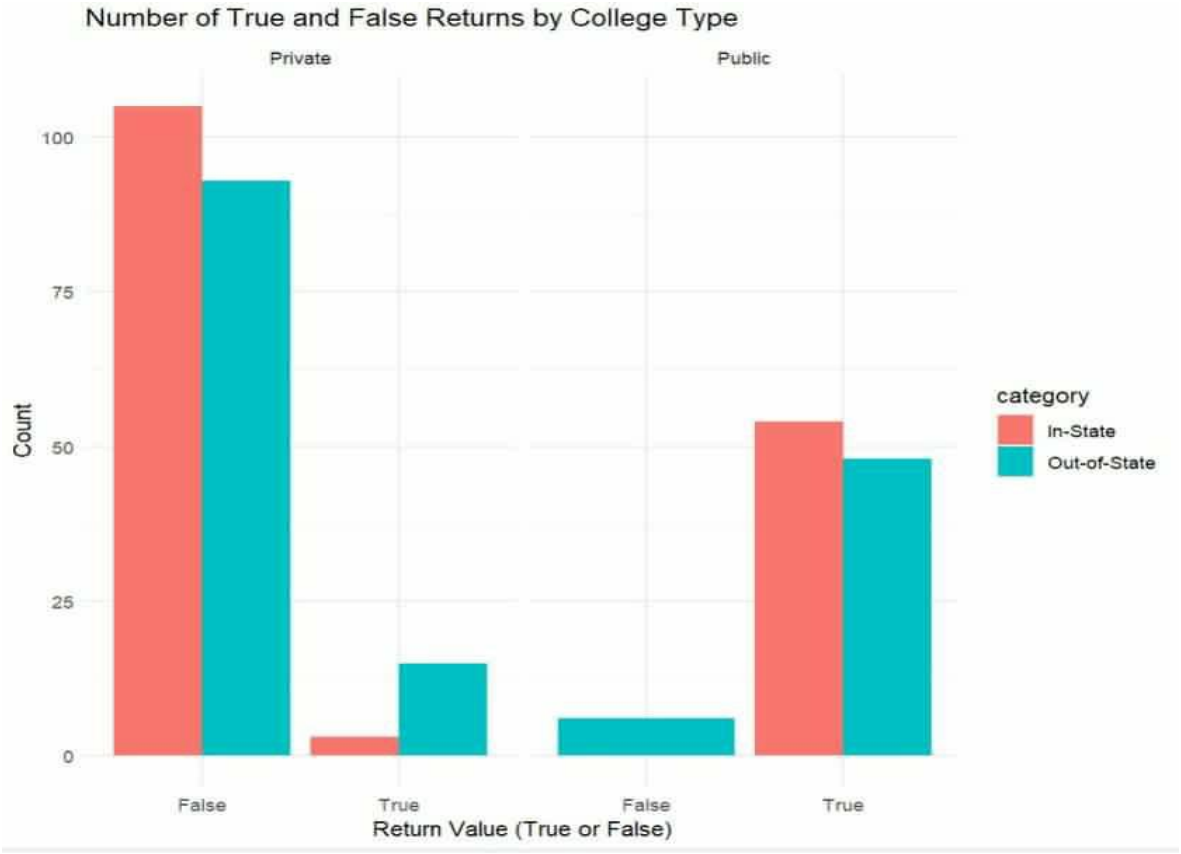
Tuition and Salary Analysis

Dataset Summarized: This code reads in two datasets on college tuition costs and salaries by region, and cleans the data to remove dollar signs and convert columns to numeric format. It then joins the two datasets on the college name variable and creates new columns calculating the ratio of mid-career median salary to total tuition cost for both in-state and out-of-state students. It also creates categorical return variables based on whether the salary-to-tuition ratio is above or below the mean. The dataset is then grouped and summarized based on in-state vs. out-of-state tuition, with summary statistics calculated for median mid-career salary, average salary-to-tuition ratio, and counts of whether the salary-to-tuition ratio is above or below the mean. The resulting summary data frame is exported to a csv file.

Data Collection process: The data is sourced from PayScale's College Salary Report.
PayScale collects data from online compensation surveys completed by individuals who have attended colleges and universities in the US, and uses it to generate salary reports for different institutions. The surveys ask for information about education, job title, employer, experience, and other factors that may influence compensation.

Tuition and Salary Analysis

Subplot distribution:



Tuition and Salary Analysis

Are there domains or industries in which this dataset should not be used?

For example, this dataset should not be used to make individual decisions about which college to attend or which major to choose, as it only provides aggregate data and may not reflect the specific experiences of each student.

Which communities, groups, or identities are represented in this dataset?

The dataset includes information about colleges and universities across the United States. It does not explicitly represent any particular community, group, or identity. However, the dataset could potentially be used to analyze patterns and trends in higher education across different regions, types of institutions, and demographic groups, such as students' race/ethnicity, gender, or socioeconomic status.

What concerns might you have about extrapolating trends or making generalized inferences from this dataset at a population level?

One potential concern with extrapolating trends or making generalized inferences from this dataset at a population level is that it may not be fully representative of the entire population. The dataset includes data from a limited number of universities and colleges in the United States, and these institutions may not be representative of the full range of educational institutions in the country.

Describe any known mitigation strategies, technical or otherwise, to address the concerns stated above.

One technical mitigation strategy to address concerns about generalizing inferences from this dataset could be to ensure that the dataset is representative of the population being studied. This could involve using a representative sampling method and collecting data from a diverse range of sources to ensure that the dataset includes a variety of perspectives and experiences.

Tuition and Salary Analysis

Are there concerns around using this data to make decisions or predictions at the individual level?

Yes, there are concerns around using this data to make decisions or predictions at the individual level. For example, while this dataset provides information about median salaries and tuition costs for different colleges and universities, it does not take into account the unique circumstances and characteristics of individual students or graduates. Therefore, making decisions or predictions about an individual's future earning potential or student debt based solely on this dataset could lead to inaccuracies or unfair outcomes. It is important to use this data in conjunction with other relevant information and to consider individual circumstances before making any decisions or predictions at the individual level.

Describe any known mitigation strategies, technical or otherwise, to address the concerns stated above

Anonymization: Personal information such as names, addresses, and other identifying information can be removed from the dataset to prevent the dataset from being linked to any specific individual.