Data Open Championship 2022

Team 1

Murtaza Nomani[1], Nange Li[2], Pengyun Li[3], Enming Zhang[4]

[1]Georgia Institute of Technology
[2]Brown University
[3]Columbia University
[4]Cornell University

December 4th, 2022

More Than Fairness: Budget Guideline for U.S. Post-secondary Public Institutions

Part I: Non-Technical Executive Summary

Introduction

Educating the public is an imperative objective of post-secondary educational institutions in the United States. Statistics have shown a continuous increase in postsecondary education enrollment within the US over the past 12 years. However, many critics question the high cost of education and exclusion of underrepresented minority groups at these post-secondary education institutions.

Motivation

Accessibility of education is crucial to achieving educational equity. In the United States, the tuition for post-secondary education is a heavy burden for students from lower-income families or under-represented groups. To support those talented prospective students, public institutions in the U.S. provide financial aid and scholarships. However, less out-of-pocket payment from students leads to less tuition fees for institutions as tuition is one of the major funding sources of American educational institutes [1]. Since academic institutions are responsible for their own budgets, they must allocate resources to ensure fiscal health while providing educational opportunities to students from different ethnic, income, and gender backgrounds.

As we review socio-economical literature on educational equity and resource allocation in the post-secondary education system [2, 3, 4], we take the initiative to extract the financial aid policies in post-secondary institutions, measure their efforts to promote equity, build a quantitative method, and navigate institutions through the budget planning tradeoff.

Main Objective

In this project, our main objective is to guide public post-secondary institutions in the US to optimize the trade off between educational fairness and profitability. We focus on public institutions only because of the data availability. However, the methodology and model we developed are applicable to any institution if given enough datasets. We hope that our findings can help institutions to develop a win-win strategy, and bring more impact on promoting a well educated workforce in the United States in the long run.

Method Overview

With a comprehensive data exploration on the given dataset, we construct a fairness measurement and a profitability measure for individual public schools. The fairness measurement included gender, ethnicity and income level. The profitability measurement included relevant and recurring income and expenses.

We cluster all public institutions into five clusters based on their similarities. We expect comparable utility values for schools in the same cluster. Our team uses a utility function and indifference curve as a measurement to identify institutions that achieved an overall higher level of fairness and profitability. We create a relatively clear boundary between the optimized group and the non-optimized group. We then analyze the income and expense structure of optimized and non-optimized public institutions to drive an optimal income and spending breakdown for public universities.

Key Findings

There is sufficient statistical evidence to support the significance of profitability on the impact on education fairness. The overall evidence implies that public colleges should allocate more resources to students and earn more from tuition to achieve a higher overall level of fairness and profitability. Factors like research, though an important component for institutional development, may not contribute from the perspective of achieving education fairness. Our findings could guide all public colleges to optimize their educational fairness and profitability. Statistical tests endorse the significance and robustness of our findings.

Besides our core findings, we also discover the racial diversity breakdown stabilizes slowly over the years and financial support provided to students generally increases from 2015 to 2020. Education fairness, as defined by the OECD criteria [6], has been increasing from 2015 to 2020.

Part II: Technical Exposition

1 Data Descriptions and Preprocessing

Our team has explored all the available datasets and filtered the relevant ones to preprocess. Given the massive missing values for American private institutions, we decide to focus on datasets related to public institutions.

The final datasets we have created in detail are shown below:

- **financial_f1_profit.csv**

  Variables are extracted and re-processed from *F_F1A_1415-1920_data.csv* (Public institutions) with the resulting dataset containing 13 columns including 'unitid', 'Federal operating grants and contracts', 'State operating grants and contracts', 'Local operating grants and contracts', 'Federal appropriations', 'State appropriations', 'Local appropriations, education district taxes, and similar support', 'Federal nonoperating grants', 'State nonoperating grants', 'Local nonoperating grants', 'year', 'edu_fairness_score', and 'profit'. Detailed construction of the ‘profit’ variable will be discussed under the methodology section.

- **match_features_key_z.csv**

  Selected variables are extracted across various datasets to indicate key features associated with an institution regardless of the institutional type.

  This dataset contains 22 columns, including 'unitid', 'year', 'Enrolled total', 'Applicants total', 'Admissions total', 'Secondary school GPA', 'Secondary school rank', 'Total employees', 'Sector of institution', 'Institutional category', 'Level of institution', 'Control of institution', 'SAT Evidence-Based Reading and Writing 25th percentile score', 'SAT Evidence-Based Reading and Writing 75th percentile score', 'SAT Math 25th percentile score', 'SAT Math 75th percentile score', 'ACT Composite 25th percentile score', 'ACT Composite 75th percentile score', 'ACT English 25th percentile score', 'ACT English 75th percentile score', 'ACT Math 25th percentile score', and 'ACT Math 75th percentile score'.

- **final_education_fairness.csv**

  We extract columns from *SFA_1415-2021_data.csv* (Student financial aid and net price data)*,* *C_B_2015-2021_data.csv* (Number of students receiving awards/degrees broken down by race/ethnicity and gender), and *EFFY_2015-2021_data.csv* (12-month headcount broken down by race/ethnicity, gender and level of student), and external data from Statista [5] to get the nationwide population profile data. Gini indexes of each component of educational fairness are computed and an education fairness score is calculated from the weighted summation of the indexes. The detailed algorithm for deriving gini indexes will be discussed in the methodology section.

This final dataset contains 7 columns, including ‘year', 'unitid', 'Gini_gender', 'Gini_ethnics', 'Gini_income', 'edu_fairness_score'.

- **budget_eval.csv**

  This dataset is considered the final dataset for drawing results. Columns include 'Tuition and fees, after deducting discounts and allowances', 'Private operating grants and contracts', 'Sales and services of educational activities', 'Gifts, including contributions from affiliated organizations', 'Student services - Current year total', 'Instruction - Current year total', 'Scholarships and fellowships expenses -- Current year total', 'Research - Current year total', 'Academic support - Current year total', 'unitid', 'year', 'Total Profit', 'gen_profit', 'label', 'Gini_gender', 'Gini_ethnics', 'Gini_income', 'edu_fairness_score', 'institution_type', and 'is_good'.

To summarize, the *financial_f1_profit.csv* includes all useful information as well as confounding factors to evaluate the profitability of public institutions. The *final_education_fairness.csv* provides the education fairness score as well as the gini indexes for gender, ethnicity, and income level respectively. The *match_features_key_z.csv* is mainly used for non-parametric score matching to include the key features of institutions. The *budget_eval.csv* is the final dataset containing all metrics to define profitability and education fairness of public institutions.

2 Exploratory Data Analysis

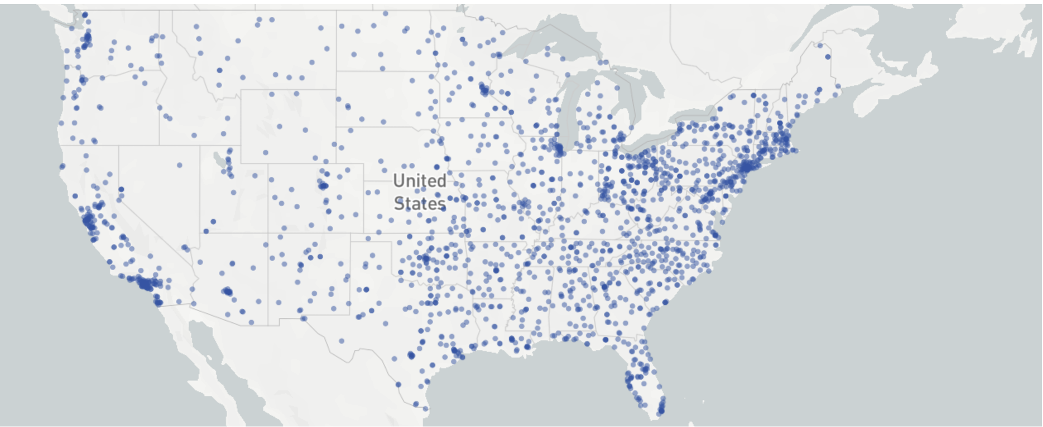

Figure 1. Geographical Distribution of U.S. Public Institutions

To start, we visualize the distribution of post-secondary public institutions in the U.S. and find that the majority of schools are located on the east coast region of the country, with the middle part having the least number of institutions.

According to the Review of Equity in Education conducted by OECD in 2012 [6], fairness in education assumes that personal and social status (i.e. gender, socio-economic position or ethnic origin) should not stand in the way of achieving educational potential. With this definition, our team starts the exploratory data analysis on the public institution students' gender, ethnicity, and income level.

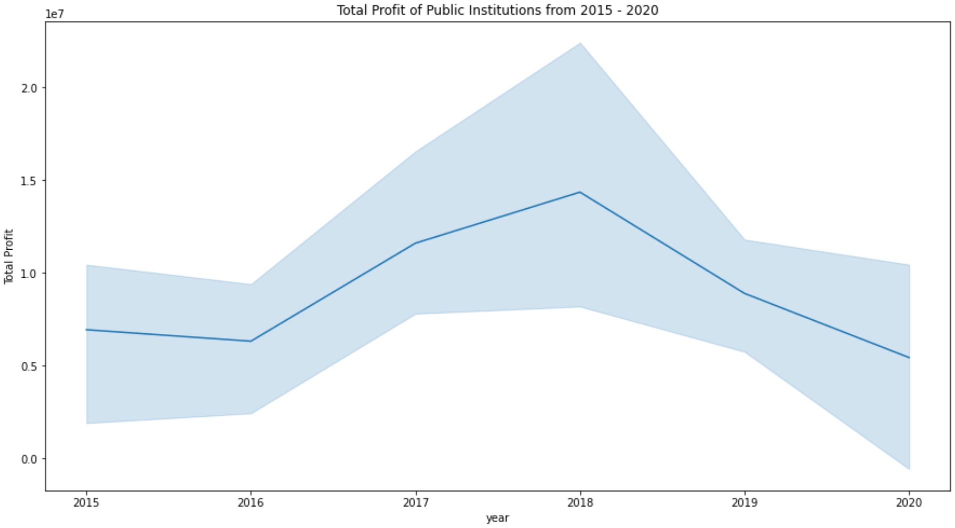

Figure 2. Line plot of total profit of public institutions from 2015-2020

By visualizing the total profit statistics from 2015 to 2020, there is an increasing trend starting from 2016 that reaches the peak in 2018. The total profit decreases after 2018. However, this trend may be largely affected by how profit is defined. Our own profit models for further analysis will be discussed later.

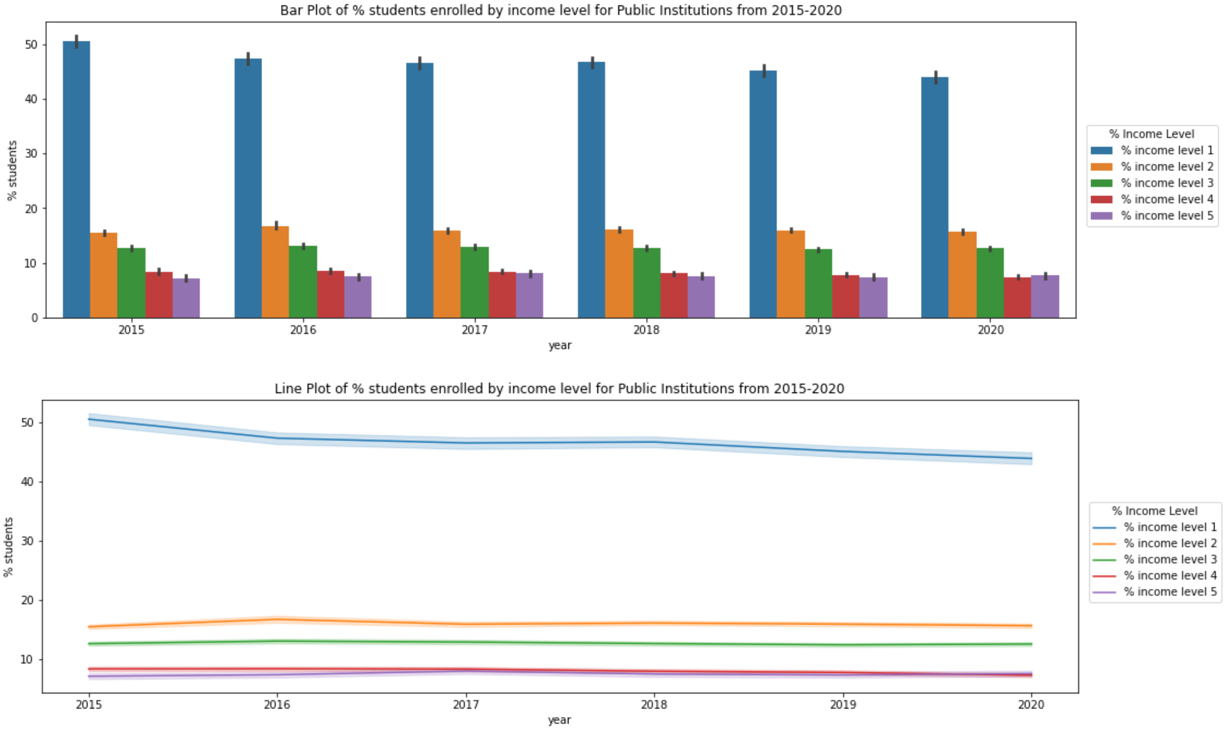

Figure 3. Barplot (top) and Lineplot (bottom) of % students enrolled by income level for public institutions from 2015-2020

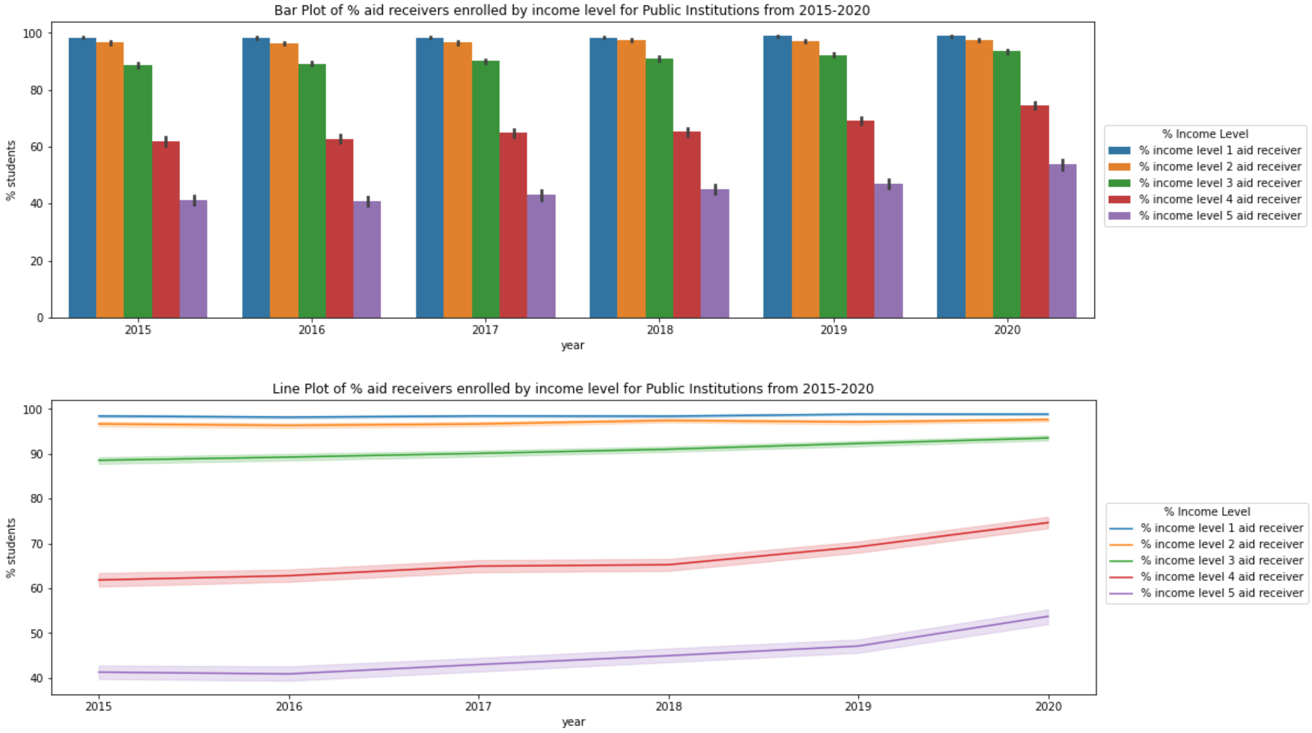

Figure 4. Barplot (top) and Lineplot (bottom) of % students aid receivers by income level for public institutions from 2015-2020

As seen from Figure 3 above, income level 1 (0- 30,000 USD) has the most students. From 2015 to 2020, this percentage has been gradually decreasing. For other income levels, the enrollment rates remain comparatively the same. This phenomenon may be attributed to the fact that the overall socio-economic status of the U.S. population has been increasing over the years, leading to a decrease in the proportion of very low-income people in the country, which in turn affects the corresponding enrollment rates.

To investigate the student support provided by institutions for each income level, Figure 4 has shown that the percentage of grants and aid receivers coming from income level 1 (0-30,000 USD), level 2 (30,001 - 48,000 USD), and level 3 (48,001 - 75,000 USD) remain high over the years. For students from level 4 (75,001-110,000 USD) and level 5 (110,001 USD and more), a greater percentage of students are receiving funds over the years, which suggests stronger financial support provided by public institutions.

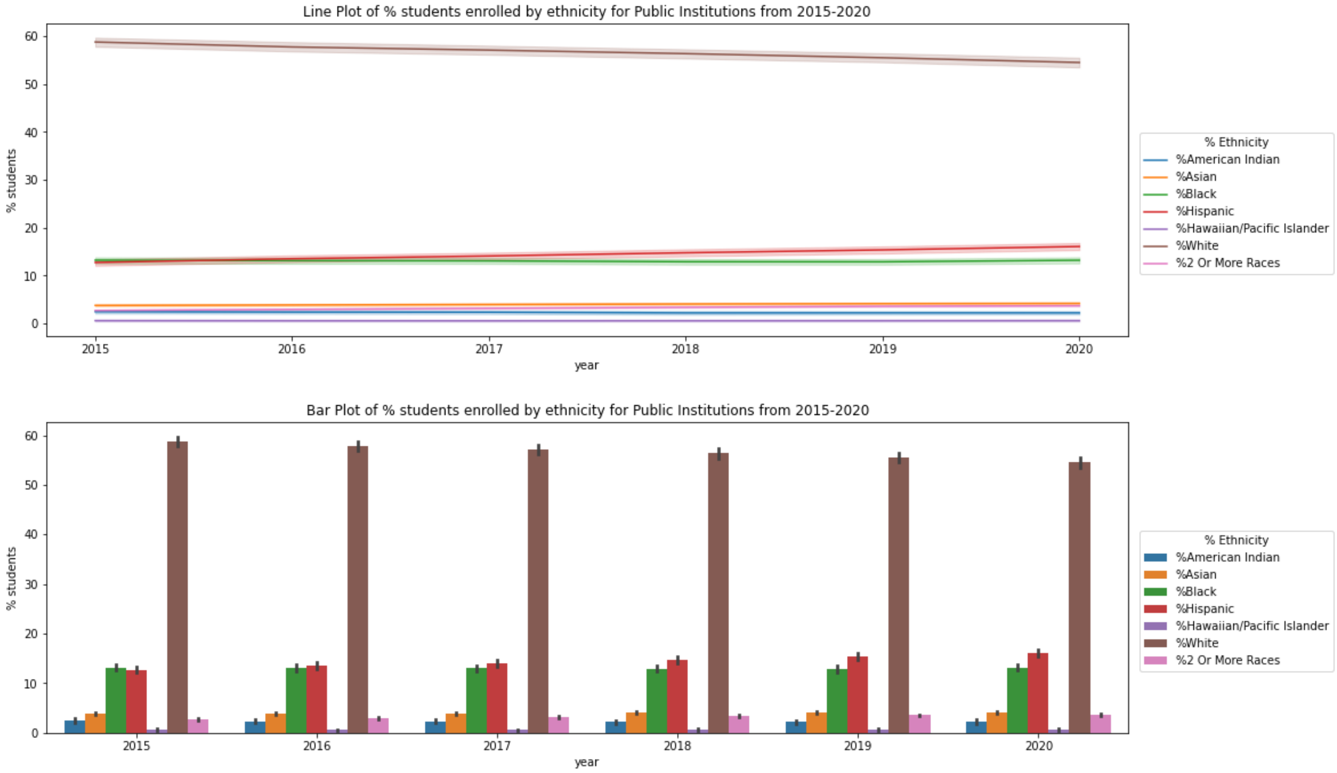

Figure 5. Barplot (top) and Lineplot (bottom) of % students by ethnicity for public institutions from 2015-2020

Figure 5 above shows that the percentage of students coming from different racial groups are steady over the years.

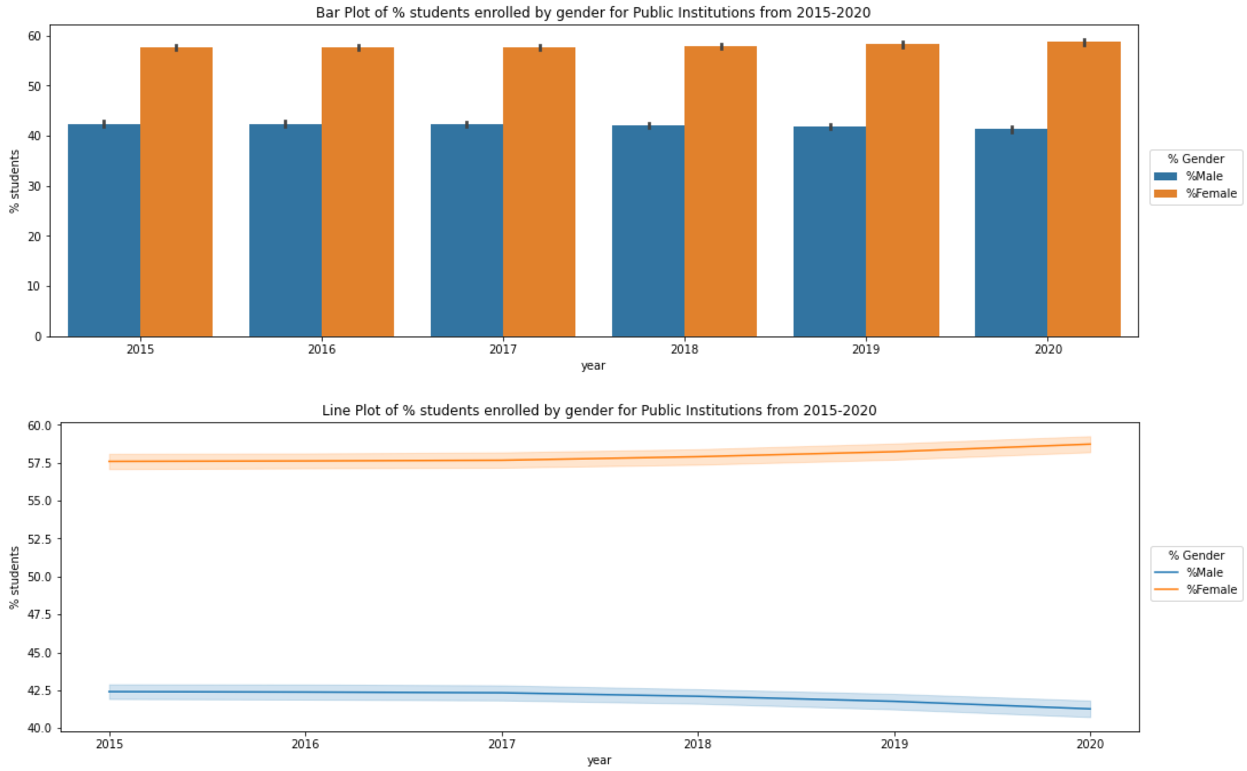

Figure 6. Barplot (top) and Lineplot (bottom) of % students by gender for public institutions from 2015-2020

As for gender, Figure 6 shows that the percentage of female students is always higher than males with the gap keeps widening over the years. This diverging trend shows a potential sign of greater gender inequality in education. Thus, we included the national gender ratio into our methodology for evaluating fairness.

3 Methodology

3.1 Metrics Definition

3.1.1 Profitability

The next step is to construct a new measurement for profitability. We choose income and expense features that are recurring and more related to fairness. Federal and State level income are more related to schools' research work and in-state student enrollments rather than the fairness that we focus on in this paper.

The new profitability measure is the sum of all the incomes subtracting the sum of all the expenses from Table 1.

Inflow	Tuition and Fees	Private Operating Grants and Contracts	Sales and Services of Educational Activities	Gifts	-
Outflow	Student Service	Instruction	Scholarship and Fellowship	Research	Academic Support

Table 1. Inflow and outflow components for institution profitability

3.1.2 Education Fairness

- **Gini Impurity Index**

We use the Gini impurity index as the metric to evaluate the educational fairness of an institution. To start with, we calculate the percentage of student enrollment by their gender, ethnicity, and income level respectively. The values are then scaled by the percentage of the US population from each income level, gender, ethnicity, and income level respectively. The scaling is essential to be made on the country level as for instance, having a 50% White and 50% Black students at an institution in a country with 90% White and 10% Black population should not be considered a fair scenario. The scaled ratios are then normalized to ensure a range from 0 to 1, and the resulting values are then used as the probability values in the Gini impurity formula to derive a final Gini index value.

$$Gini_{gender} = 1 - [(\alpha \frac{p_{institution}(male)}{p_{USA}(male)})^2 + (\alpha \frac{p_{institution}(female)}{p_{USA}(female)})^2],$$

$$Gini_{ethnic} = 1 - \sum (\beta \frac{p_{institution}(ethnic_k)}{p_{USA}(ethnic_k)})^2 ,$$

$$Gini_{income} = 1 - \sum (\gamma \frac{p_{institution}(income\_level_i)}{p_{USA}(income\_level_i)})^2$$

where α, β, γ refers to normalization parameters; k refers to different racial groups; $i \in \{1, 2, 3, 4, 5\}$.

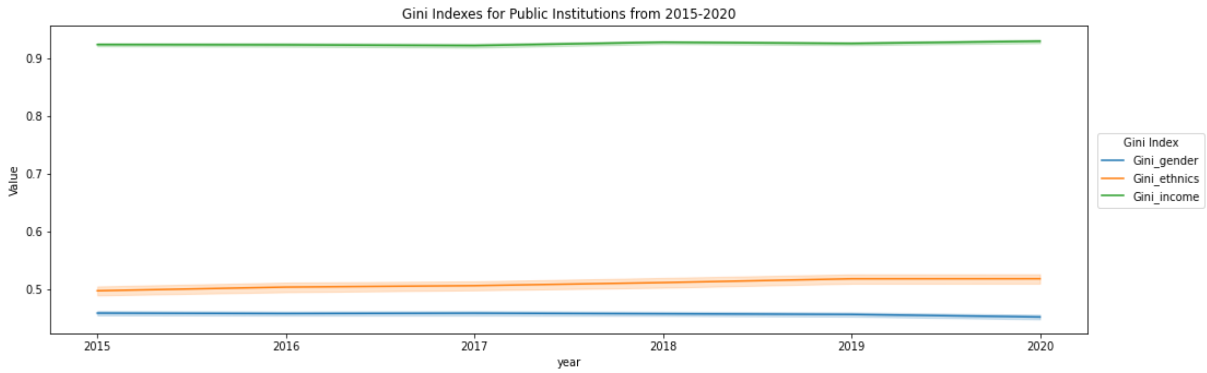

Figure 7. Line Plot of Gini Indexes for Public Institutions from 2015-2020

The line plots of Gini indexes in Figure 7 have shown that the indices for ethnicity and income have been slightly increasing from 2015-2020 while the index for gender has been slightly decreasing over the years.

- **Education Fairness Score**

We assume that the three components to define education fairness are of equal importance. The education fairness score is thus defined as:

$$Fairness = \frac{Gini_{gender} + Gini_{income} + Gini_{ethnic}}{3}$$

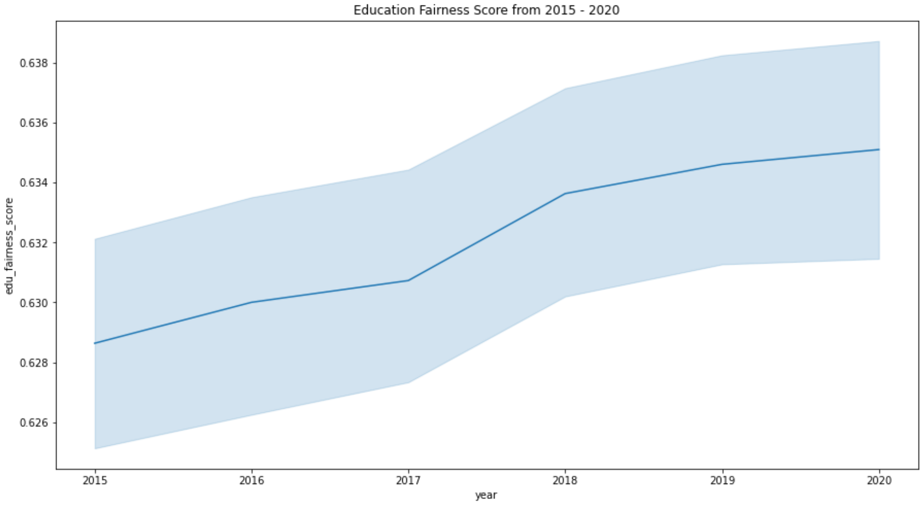

Figure 8. Line plot of Education Fairness Score from 2015 - 2020

A higher score indicates better equality achieved in education. The plot in Figure 8 indicates that the level of education fairness has been increasing from 2015 to 2020.

3.1.3 Adjusted correlation

We account for confounding variables such as Federal-level and State-level funding by running a linear model of fairness and profitability. The linear regression results show that profitability and fairness have a negative relationship indicated by the negative coefficient = -0.0048. The p-value is < 0.0001, proving that such a negative relationship is robust.

3.2 Score Matching with t-SNE

To further analyze the engaged institutions in groups, the first step of our modeling part is a non-parametric score matching. We try to avoid biases and select the features that apply to all institutions. The key features are selected based on three categories: school size, type, and admission selectivity of the universities.

For the school size attributes, we pick the number of applicants, number of employees, and number of total students. For the institution type, we select university level and control categorical data that indicate if they are private, public, 2 years, or 4 years. For the selectivity of universities, we choose SAT, ACT score percentiles, and admission acceptance rate. With those related features ready, we applied Z-score standardization for all the numerical features and one hot encoding for all the categorical features. That is,

$$Z_{feature} = \frac{X - \bar{X}}{Var(X)}$$

While it is ideal to preprocess features and apply unsupervised clustering on all the data points, we need to address the existing missing values on our selected key features. To avoid potential bias and decreasing variance issues, we decide not to impute the missing institution features, switching to a semi-supervised approach instead. Those data points without missing values are treated as training data (27.2%), which will be learned to obtain the clusters. Then the remaining 82.8% of the data points will be compared based on the k-nearest neighbors in the training set, using whatever non-missing features this data point has to match a cluster.

T-SNE, short for t-distributed stochastic neighbor embedding, is an efficient algorithm for learning the structural information of the data points while keeping the pairwise similarity. Before mapping, we have a perplexity measurement for each pair of high-dimensional data points (x_i, x_j) with Gaussian density:

$$P_{i,j}(\sigma) = \frac{exp\{(x_i - x_j)^2/2\sigma^2\}}{\sum_{k\neq j} exp\{(x_k - x_j)^2/2\sigma^2\}}$$

Where σ is a positive value extracted from solving equations. Then another measurement is defined on the mapped data with lower dimension. In our case, each institution's feature vector x will be mapped to a point y in the 2-D space, and the quantitative measurement after mapping with Cauchy density is:

$$Q_{i,j} = \frac{(1+|y_i - y_j|^2)^{-1}}{\sum_{k\neq j}(1+|y_k - y_j|^2)^{-1}}$$

To find the best Q, we can use the KL divergence as the loss function to apply any stochastic optimization methods:

$$Loss = KL(P, Q) = cross\ entropy(P, Q) - entropy(P) = \sum_{(i,j),\, i\neq j} P_{i,j} \log \frac{P_{i,j}}{Q_{i,j}}$$

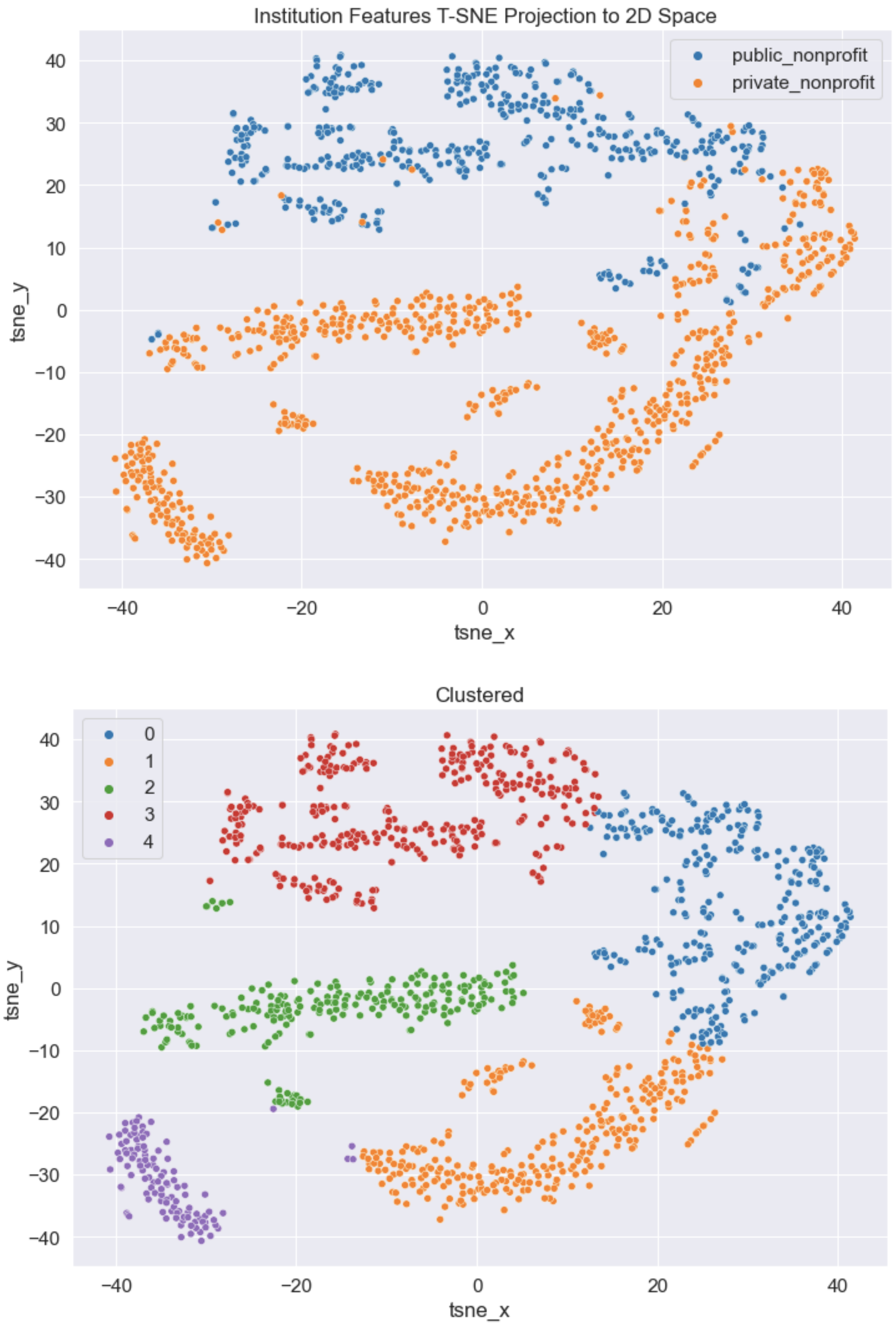

Figure 9. The 2-D mapped institution feature vectors and label assigned in clusters

Visualizing institution features above, we can infer that there are five basic components after t-SNE dimension reduction in the training data. Given this division, we assign labels and generate five clusters, with which we can map the remaining data points to the nearest training cluster using KNN. This mapping is implemented on a projection of a feature subset, where feature imputation is not necessarily applied.

3.3 Indifference Curve

After the score matching section, we want to identify schools with the most efficient resource allocation, i.e. how to adjust the tradeoff between fairness and profitability. In the Economics field, a utility function accompanied by an indifference curve is a common model for this optimization scenario. Given an institution's fairness index and profitability value, we have the following equation to represent a group of indifference curves:

$$\mathit{fairness}^{\alpha} \cdot \mathit{profitability}^{1-\alpha} = C$$

Where α is a parameter ranging between 0 and 1, indicating an institution's emphasis on each of the two indices. By adjusting α we are changing the slope of a curve, while by adjusting the right-hand side C we are moving a curve towards top-right or bottom-left. Here we'd like to put same weights on both fairness and profitability, setting $\alpha = 0.5$, and the above equation then is transformed to:

$$\mathit{fairness}^{1/2} \cdot \mathit{profitability}^{1/2} = C$$

Or equivalently,

$$\sqrt{\mathit{fairness} \cdot \mathit{profitability}} = C$$

Note that the two indices under the square root have been normalized to (0, 1) so that there will not be any invalid square root concern.

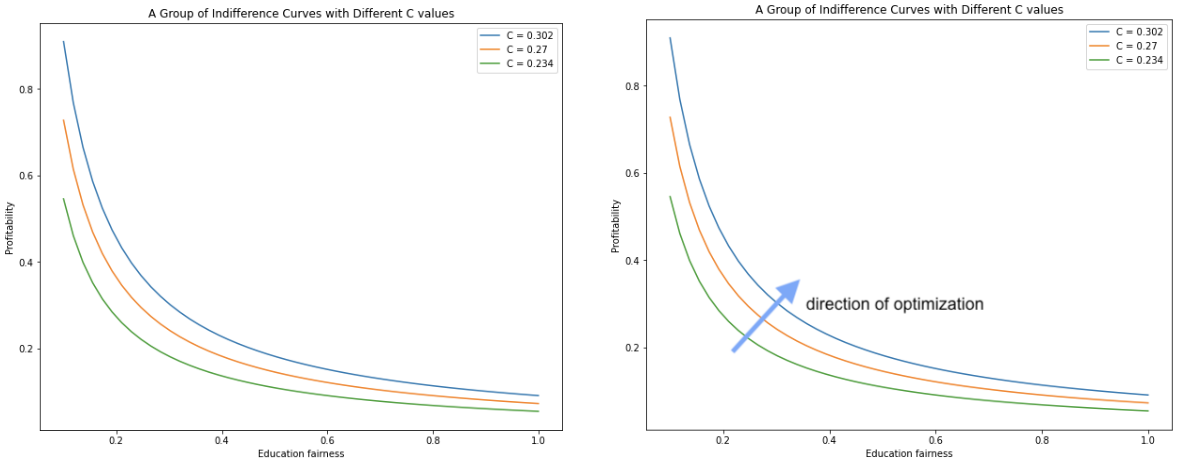

Figure 10. Given indifference curve group, the moving trends while changing C

In Figure 10, we can assert that *education fairness, profitability* pair points on the blue curve (C = 0.302) have better performance than points on the orange or green curves. If we fix an education fairness level, the profitability is increasing from the green line to the blue line. Thus, this economic model gives us a guideline on how to cross-evaluate the utility of the involved post-secondary institutions.

Within each cluster that we generated from t-SNE and matching, we already have the assumption that those institutions are structurally closer to each other in the feature space, therefore we would expect the utility values of them to be comparable. We set the median utility value of a cluster as a baseline, and classify the utilities in the same cluster to “optimized budget” if the value is above median, otherwise “not optimized budget”.

4 Results

The OLS regression results after standardizing all variables indicate a negative adjusted coefficient of profitability -0.0048 with a p-value smaller than 0.0001, suggesting that there is sufficient statistical evidence to support the significance of profit to impact education fairness. The overall evidence from breakdowns of income and expense budget structure shows that a school that allocates more resources to students and earns more from tuition tends to achieve a higher overall level of fairness and profitability. The Kolmogorov-Smirnov test is applied at the end which showed that the income and expense percentage differences are robust (p-value < 0.0001).

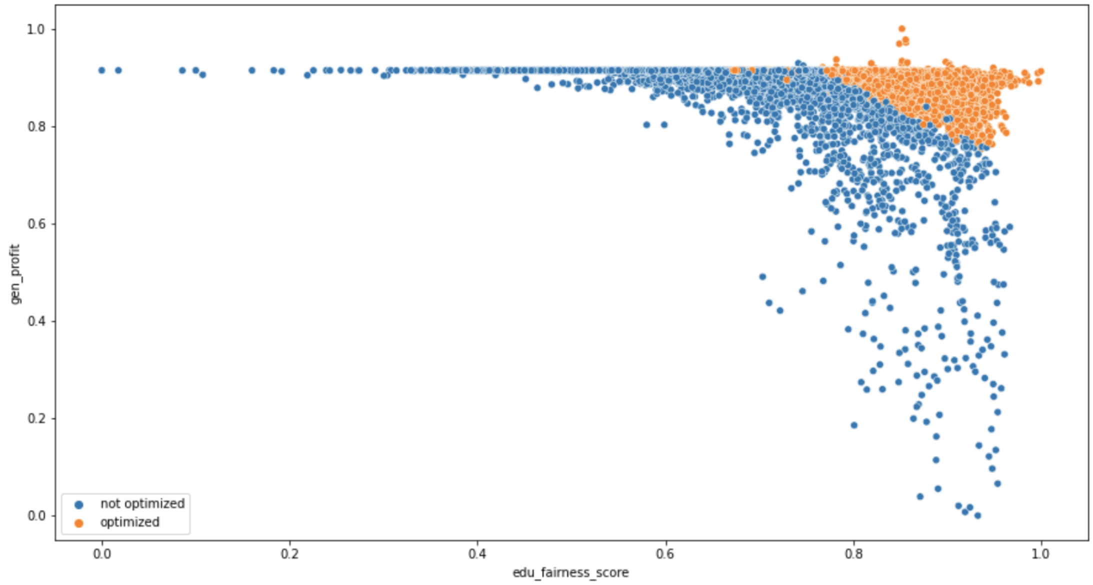

Figure 11. Scatter plot of of the tradeoff between normalized profitability and education fairness score, mounted on optimized and non-optimized institutions respectively

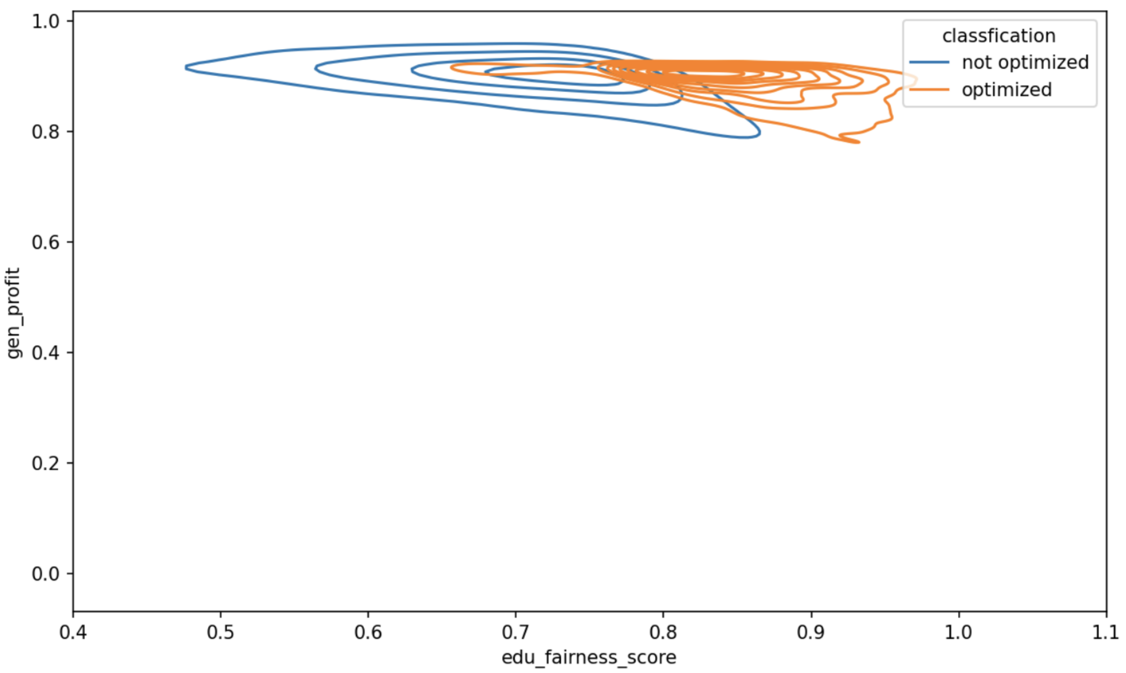

Figure 12. KDE plot of the tradeoff between profitability and education fairness score, mounted on optimized and non-optimized institutions respectively

The scatter plot and kernel density estimation (KDE) plot both contain each individual schools' education fairness score and profitability. After applying the indifference curve as a boundary, we are able to identify the schools that achieved an overall higher level of fairness and profitability, with a relatively clear boundary between two groups of data points (Kolmogorov-Smirnov test p-value < 0.0001 for both educational fairness score and profitability between optimized and non-optimized). To interpret our results, we analyze the income and expense structure of optimized and non-optimized schools.

Income (Mean)	**Optimized Institutions**	**Not Optimized Institutions**
Tuition	0.904	0.876
Private Grants and Contracts	0.032	0.039
Sale and Services of Educational Activities	0.026	0.037
Gift	0.038	0.047

Table 2. Income Structure breakdown

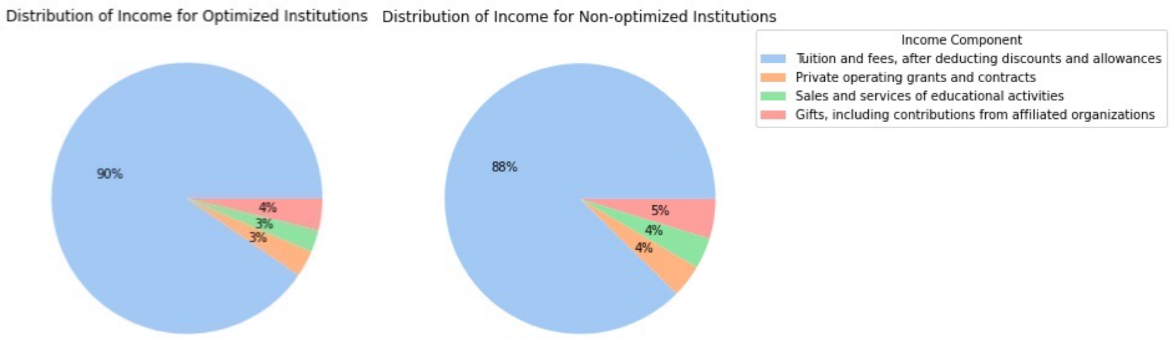

Figure 13. Pie Charts of distribution of income for optimized (left) & non-optimized (right) Institutions

From the income structure breakdown table as well as the pie charts in Figure 13, the optimized schools focus on generating revenue from Tuition, while the not optimized schools secure funding from non-tuition sources. This implies that institutions should focus on students and tuition to optimize fairness and profitability.

Expense (Mean)	**Optimized Institutions**	**Not Optimized Institutions**
Student Service	0.153	0.130
Instruction	0.588	0.607
Scholarships and fellowships	0.121	0.098
Research	0.012	0.044
Academic Support	0.125	0.121

Table 3: Expense Structure breakdown

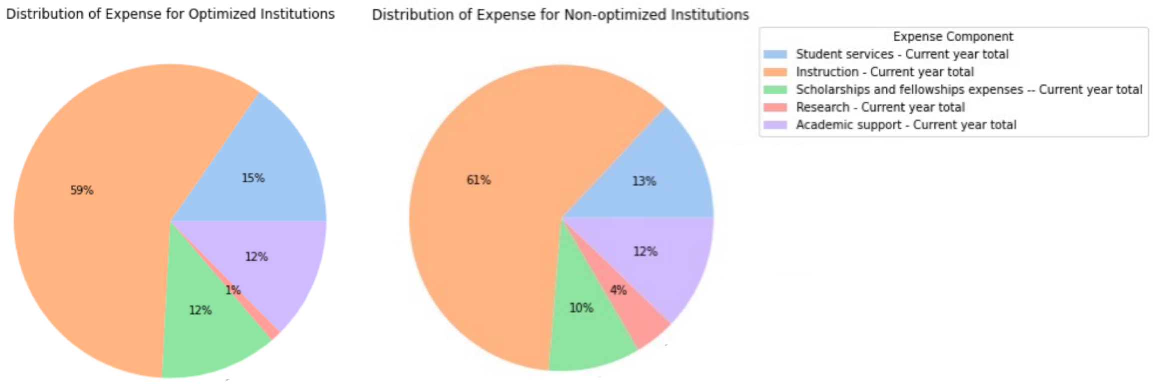

Figure 14. Pie Charts of distribution of expenses for optimized (left) & non-optimized (right) Institutions

Non-optimized institutions are spending more on research and instruction compared to optimized ones. On the other hand, optimized institutions spend more on student services, scholarships and fellowships. We apply the Kolmogorov-Smirnov test and the income and expense percentage differences are robust as all the p-values are smaller than 0.0001.

The overall evidence from income and expense structure shows that a school allocating more resources to students and earning more from tuition tends to achieve a higher overall level of fairness and profitability. Factors like research, though usually considered as an important component for institutional development, may not contribute from the perspective of achieving better education fairness.

Part III: Discussion

Strengths

The main strength of our analysis is the social impact it can bring to American post-secondary institutions. Although our analysis in this report is only for public institutions, it also applies to private institutions since the method is generalizable. Many studies analyze the equity of institutions without taking into account the cost of making the institution more fair. By taking into account the institutional budget, our analysis gives post-secondary education institutions practical policy advice on where to allocate their funds in a way that promotes equity.

Weaknesses

Given the limits posed by the extremely low availability of income related statistics of private institutions in the United States, we have to narrow down our target institutions to public ones only. Though we could choose to simply remove the income component of the gini calculation or adopt a separate fairness rating system for private institutions, it would sacrifice the accuracy, consistency, and applicability of our project.

Besides, the modeling results would be improved with more fine grained datasets. For instance, many datasets have either by ethnicity or by income-level statistics, but none of them provide both information. The datasets on income-level also divide income levels into four levels rather than the traditional seven income levels that are featured in the national household incomes data set.

Future Work

Given the increasing availability of public datasets in education in recent years, we could explore and integrate more relevant factors in our model to assess the fairness of education. Also, our current work is for within-US institutions. One future direction is to extend the scope of the research to other countries and apply more time series analysis in the future.

Additionally, to more accurately calculate the fairness of an institution, we could incorporate data from the Common Data Set and Tax Return Data to analyze equity in financial aid rewards of an institution. Adding a separate factor to the gini coefficient that reflects how many students in financial need are able to receive aid would make our analysis of equitability in educational institutions more accurate.

Our analysis uses the IPEDS dataset for student enrollment, which according to [7], is inaccurate in its financial aid rewards numbers for certain institutions. Taking advantage of an expanded dataset that includes financial aid rewards and tax records such as the *Opportunity Insights* dataset featured in [7], we could further refine our estimate score of the “fairness” of an institution.

Part IV: References

[1] Liefner, I. Funding, resource allocation, and performance in higher education systems. *Higher Education* **46**, 469–489 (2003). https://doi.org/10.1023/A:1027381906977.

[2] Buckner, E., Zhang, Y. The quantity-quality tradeoff: a cross-national, longitudinal analysis of national student-faculty ratios in higher education. *High Education* **82**, 39–60 (2021). https://doi.org/10.1007/s10734-020-00621-3.

[3] Kenneth J. Arrow (1993) Excellence and Equity in Higher Education, Education Economics, 1:1, 5-12, DOI: 10.108/09645299300000020.

[4] Boaz Shulruf, Rolf Turner, John Hattie, A Dual admission model for equity in higher education: a multi-cohort longitudinal study, Procedia - Social and Behavioral Sciences, Volume 1, Issue 1, 2009, Pages 2416-2420.

[5] https://www.statista.com/. Statista dataset for national-wide gender, ethnics, and income level distribution.

[6] OECD, Equity and Quality in Education: Supporting Disadvantaged Students and Schools, (2012)

[7] Weintraut, B., Hill, C. B., Kurzweil, M., & Pisacreta, E. D. (2020, November 16). Comparing Public Institution-Level Data on Students’ Family Income and Financial Aid. https://doi.org/10.18665/sr.314398