Unveiling Diversity Dynamics: Exploring Racial and Gender Diversity, Fields of Study, and Alumni Outcomes in U.S. Higher Education

SI 618 Team Project FA23

Authors:
Yi-Chun Wang (ritaycw); section 001
Je-Ching Liao (jeching); section 101
Yi Hsien Wu (yihsien); section 101

Motivation

Our project aims to dissect the relationships between racial and gender diversity, fields of study, and alumni outcomes in US higher education, providing actionable insights for institutions committed to Diversity, Equity, and Inclusivity (DEI). As a group of international students, we seek to reveal patterns in racial and gender diversity across institutions, examining correlations with academic disciplines and degree levels. For example, we would like to explore stereotypes like higher Asian student percentages in engineering programs. Additionally, we intend to investigate whether institutions fostering diversity have alumni with better salary potentials. Through statistical analysis, our project results can be used to inform recruitment strategies and contribute to the broader conversation on cultivating inclusive academic environments.

Data Sources

There are four datasets involved in this project, one primary and three secondary. The primary dataset is called ‘diversity_school.csv’ (referred to in the code as dfDiversity), containing information about various institutions, including their institution names, total student enrollment, and the state in which they are located. It also includes data related to student enrollment categorized by groups, race, and gender. This dataset allows for the geographical categorization of schools and offers insights into the composition and diversity of the student body.

The first two secondary datasets are ‘tuition_cost.csv’ and ‘salary_potential.csv’, referred to in the code as dfTuition and dfSalary. The dataset ‘tuition_cost.csv’ provides information about various educational institutions, including school name, location details (state name and abbreviation), the type of institution (public, private, for-profit), the duration of degree programs (4-year or 2-year), and financial details such as room and board costs, in-state and out-of-state tuition fees, as well as the total costs for in-state and out-of-state residents, combining tuition and room and board expenses. This data compares the costs of attending different schools and understands the financial aspects of higher education in various states. The dataset ‘salary_potential.csv’ provides information on potential salary ranks, school names, and state names for various educational institutions. It includes estimated early career and mid-career pay in USD, as well as data on the percentage of alumni who believe they are contributing to making the world a better place and the percentage of the student body engaged in STEM (Science, Technology, Engineering, and Mathematics) fields. It is a dataset for assessing the earning potential and social impact of different schools within the United States. The above three datasets are all retrieved, in .csv format, from a Kaggle dataset page (College tuition, diversity, and pay (kaggle.com)).

The final dataset bears the name ‘FieldOfStudyData1718_1819_PP.csv’, referred to in the code as dfFieldOfStudy. This dataset is abbreviated as the ‘field_of_study.csv’ in our proposal. The valid columns for us in the dataset are Column no.3 to Column no.9. These columns contain information about various educational institutions, including school name, their type (public, private, for-profit), the number of main degree programs they offer, department details (code and name), and information about the degree programs they offer, such as the degree level and degree name. This dataset is retrieved as .csv format from the College Scorecard of the U.S. Department of Education (Data Home | College Scorecard (ed.gov)).

Data Manipulation Methods

Missing values:

Our initial data manipulation step involved addressing missing values. Given that the large dfFieldOfStudy dataset contained numerous privacy-suppressed data related to debts that were not available, we selectively excluded those unavailable columns by only retaining columns with available data regarding school information, fields of study, and levels of degree, i.e., columns named “UNITID”, “OPEID6”, “INSTNM”, “CONTROL”, “CIPCODE”, “CIPDESC”, “CREDLEV”, and “CREDDESC”.

Additionally, the dfSalary dataset presented 33 missing values in the "make_world_better_percent" column. Since our analysis did not focus on this column, we did not address these missing values in the beginning. However, for machine learning purposes, where predictions were required for "make_world_better_percent" scores of unavailable institutions, we separated the rows with missing values as the testing dataset and retained the rest as the training dataset. Upon merging the dfSalary dataset, which included the "make_world_better_percent" column, with dfInstitutionDiversity and dfTuition for the preparation of training and testing datasets, the inner join operation resulted in the removal of some missing values in the "make_world_better_percent" column, ultimately reducing the testing datasets to only 17 rows of data.

Diversity Index (DI):

We applied Simpson’s Diversity Index (SDI) infinite version, a formula commonly utilized in AP Biology classes for calculating diversity in situations involving large sample sizes. This formula generates values ranging from 0 to 1, representing no to maximum diversity. In our analysis, we implemented this formula to compute diversity and created a new column called "diversity_index" within the dfDiversity dataset.

$$D_s = 1 - \sum \left(\frac{n}{N}\right)^2$$

D_s = Diversity Index
n = Number of individuals for each species
N = Total number of all individuals

Data Merging:

Given that all datasets share the common "name" column representing school names, we performed inner joins on the datasets as required for analyzing specific questions. For example, we merged dfTuition and dfDiversity to address inquiries related to the correlation between institutional tuition fees and diversity. Similarly, the merging of dfSalary and dfInstitutionDiversity was conducted to explore associations between alumni salary outcomes and institutional diversity indices. In addition, we merged dfStatesAbbr into dfDiversity to include state abbreviations, enhancing the visual clarity of the plots.

Analysis

Topic 1: Simply analyzing enrollment by gender and racial groups of educational diversity in the United States (Q1-Q3)

From Figure 1, the barplot for the top 10 states with the highest percentage of women enrollment. Each of the top 10 states, namely Virginia, Nebraska, Mississippi, Kentucky, Louisiana, Ohio, Florida, Idaho, Wisconsin, and Georgia, boasts a female enrollment rate surpassing 50%. Interestingly, except Idaho, most of these states are geographically situated in the eastern part of the US.

From Figure 2, the barplot for the top 10 states with the highest percentage of student enrollment for different groups, we observe interesting trends in enrollment rates for different groups of races. For the enrollment rate for American native students, the highest state is Montana (~30%), with North Dakota and Alaska following, and then New Mexico and South Dakota. This trend coincides with the proximity of the location of the states to the native American population. For the enrollment rate for Asian students, we see an extremely high rate for the state of Hawaii(~25%), reasonably due to its proximity to Asia and its history with Asian settlers. Following Hawaii, west-coast states such as California, Nevada, and Washington all have high Asian student enrollment rates. For the enrollment rate for African American native students, the top 10 states are all part of the historical Southern United States, scoring between 25% to 50%. For the enrollment rate for Hispanic students, those states with connections with the previous Spanish colonization and Mexico are in the top 10 states list, notably in the top 5. For the enrollment rate for Pacific Islanders, it is not surprising that Hawaii scores way ahead. For the enrollment rate for White students, the top 10 states are mostly in the northern part of the country, with similar rates between 70% to 80%. For the enrollment rate for students with two or more races, Hawaii again scores way ahead, owing to the diversity of the people living there. Finally, we examine the total trend for Minorities, with Hawaii coming in first; the following states are mostly states with a large population of Hispanic or African American students.

In order to measure the diversity of each educational institution, we introduce the Diversity Index. According to Figure 3, the histogram of the distribution of the Diversity Index for all institutions is left-skewed, attaining its highest point at 0.6. This implies that a notable proportion of institutions possess diversity scores below this peak. The overall range of Diversity Index values spans from 0.0 to 0.85.

Topic 2: Examining tuition disparities, institutional types, and their impact on diversity in U.S. higher education (Q4-Q6)

From Figure 4, the barplot of differences between in-state and out-of-state fees, among all the US states, institutions in Colorado have the largest difference between their in-state and out-of-state tuition fees, which is more than $8000 on average. This substantial disparity in tuition fees highlights the financial challenge for non-residential students to pursue higher education in Colorado. In contrast, the states of Iowa, South Dakota, and Minnesota have the smallest gaps between in-state and out-of-state tuition fees, averaging less than $1000. This suggests a more accessible and equitable educational environment for students from outside of these states, potentially fostering a more diverse student population.

Figures 5 and 6, the regression plots of relationship between tuition and diversity index, indicate positive correlations between both in-state and out-of-state tuition fees and the institution's diversity index, meaning that institutions with higher tuitions are more likely to have higher diversity indices. Further regression analysis we conducted shows that both positive relationships are statistically significant, with both p-values lower than 0.05. We speculate that

the positive correlation between tuition fees and diversity indices derives from higher-cost institutions offering better educational resources, which attracts a more diverse student population seeking enhanced educational opportunities.

According to Figure 7, the boxplots of enrollment rates of three types of institutions categorized by gender, it becomes evident that the median enrollment rates for women in all three types of institutions are nearly identical. However, for-profit institutions reveal a more extensive interquartile range (IQR). This expanded IQR signifies a heightened variability and spread of data points within for-profit institutions. Conversely, when considering the smaller IQR for Public Institutions, it suggests a more concentrated distribution of data. Furthermore, both Public and Private institutions display outliers in their enrollment rate data.

According to Figure 8, the boxplots of enrollment rates in three institution types by gender, it is revealed that, on average, for-profit institutions have higher median rates compared to the other two. Public institutions display a broader interquartile range (IQR), indicating increased variability. Private institutions, conversely, exhibit more outliers, signifying significant deviations in enrollment rates. This expanded IQR within for-profit institutions signifies a broader range of enrollment rates, highlighting the diverse distribution of data in this sector. Moreover, when examining private institutions, a notable observation is the presence of a relatively higher number of outliers, indicating instances where enrollment rates deviate significantly from the overall trend.

Topic 3: Analyzing the impact of racial diversity and STEM programs on alumni career outcomes (Q7-Q10)

Figures 9 and 10, the regression plots of relationship between diversity index and career pay, show positive relationships between both salary outcomes in early and mid-career pay and the institution's diversity index, meaning that institutions with higher diversity index are more likely to have alumni with better salary outcomes. Further regression analysis we conducted shows that both positive correlations are statistically significant, as evidenced by both p-values below 0.05. We speculate that the positive correlation between career pay and diversity indices may stem from the concept that more diverse institutions foster an environment where students learn to adapt effectively to diverse workplaces. This adaptability is valuable in today's interconnected professional landscape, potentially contributing to higher career pay for individuals who have experienced a diverse educational setting. The correlation suggests a potential link between diversity exposure during education and the ability to navigate and excel in varied work environments.

Figure 13, the regression plot of relationship between diversity index and "make_world_better_place" percentage, indicates a negative relationship between the institution's diversity index and the percentage of their alumni feeling that they make the world a better place. Further regression analysis we conducted shows that the negative relationship is statistically significant, as evidenced by the p-value below 0.05. We speculate that as more students from different backgrounds, mostly less privileged backgrounds, study at institutions with higher diversity, the alumni from the institutions are reasonably less confident in making the world a better place.

Figure 11, the regression plot of relationship between STEM percentage and career pay, indicates a positive relationship between the percentage of STEM programs in the institution and its alumni's early career pay, meaning that institutions with a higher percentage of STEM programs foster students with better salaries in their early careers. Further regression analysis shows that the positive correlation is statistically significant due to the p-value of below 0.05. We speculate that the positive relationship may arise from the high demand for STEM-related skills in the job market.

Figure 12, the regression plot of relationship between STEM percentage and "make_world_better_place" percentage, indicates a negative relationship between the percentage of STEM programs in the institution and the percentage of their alumni feeling that they make the world a better place. Further regression analysis we conducted

shows that the negative correlation is statistically significant, as evidenced by the p-value below 0.05. We speculate that the negative relationship may arise from the specialized nature of STEM education, which often focuses on technical skills rather than emphasizing societal impact. Therefore, graduates from institutions with a higher percentage of STEM programs may perceive their roles as task-oriented instead of actually bringing a positive impact to the world themselves.

Topic 4: Exploring connections between racial diversity, areas of study, and degree levels in higher education (Q11-Q12)

From Figure 14, the regression plots of relationship between percentage of different majors and diversity index, we can see the distribution of the percentage of each kind of major all concentrated on the spectrum's lower end. While we might conclude that the diversity index does not correlate with the percentage since a low percentage of a specific field of study happens in institutions with all levels of diversity index score. However, as we look at institutions with higher percentages of Literature, Arts, and Bio-related majors, they tend to also have higher diversity indices, with an obvious outlier at the bottom right of the plot of literature majors. The same does not apply to the percentage of engineering, science, and management majors. We can also see many institutions with 100% science majors or 100% arts majors.

Figure 15, the regression plots of relationship between percentage of different degrees and diversity index, shows the relationships between percentages of different institution degrees and diversity indices. Notably, the results for Master's Degrees, Doctoral Degrees, and Graduate Certificates show slightly positive relationships with diversity indices. Conversely, results in other groups show either no correlation or slight negative associations with diversity indices. To validate the statistical significance of these relationships, regression analyses were conducted. The results indicate that only the positive relationship between the percentage of institution Doctoral Degrees and its Diversity Index was statistically significant, with a p-value below 0.05 (p-value = 0.037), while other groups did not show statistical significance.

We initially speculated that higher percentages of higher education levels might be positively correlated with higher diversity indices due to the expectation that advanced academic programs, such as Master's Degrees, Doctoral Degrees, and Graduate Certificates, often attract a more diverse cohort (e.g., international students). The results from both visualizations and statistical analyses aligned with our initial speculation, confirming the expected positive relationship. However, it's worth noting that statistical significance was only confirmed within the Doctoral Degrees group.

Topic 5: Machine Learning Exercise: predict "make_world_better_place_percent" with the tuition, salary, diversity index, degree length, and school type (Q13)

There are 17 schools with missing values on the percentage of students who think they are making the world a better place. We can predict the values for these schools by comparing the R-squared score and root mean squared error from the results of the linear regression model, logistic regression model, and gradient boosting regressor. From the result, we can see that gradient boosting regressor prevails in both metrics. From Table 1, we could see the prediction for the 17 schools.

Topic 6: Dimension Reduction Exercise: institution clustering based on reducing columns including salary, tuition, STEM percentage, diversity index, etc. (Q14)

The data is reduced into 3 dimensions with principal component analysis, upon which agglomerative clustering of four clusters is conducted. The colored result is plotted as Figure 16, the scatterplot of dimension-reduced data with agglomerative clustering results; we can see from the plot that Cluster 0, 1, and 2, are quite concentrated, while Cluster 3

is scattered. The clusters do not greatly overlap each other. Using numerical analysis (mean value for each column), from Table 2, Cluster 2 has the lowest potential salary, STEM percentage, diversity index, and tuition, while Cluster 1 has the highest of the same data. Clusters 0 and 3 have similar potential pay and tuition, with Cluster 0 having a higher "make the world a better place" percentage, but a lower stem percentage and diversity index.

Furthermore, k-means clustering is also conducted on the dimension-reduced dataset. Since the 2-cluster k-means clustering has the highest silhouette score (about 0.411), we select it as the optimal number of clusters. From Figure 17, the scatterplot of dimension-reduced data with k-means clustering results, we can see the 2-cluster clustering yields two well-defined clusters with their members not encroaching on the area of the other cluster. Using numerical analysis (mean value for each column), from Table 3, we can see Cluster 0 has a lower salary, tuition, STEM percentage, and diversity index, but a higher "make the world a better place" percentage.

Visualization

Figure 1

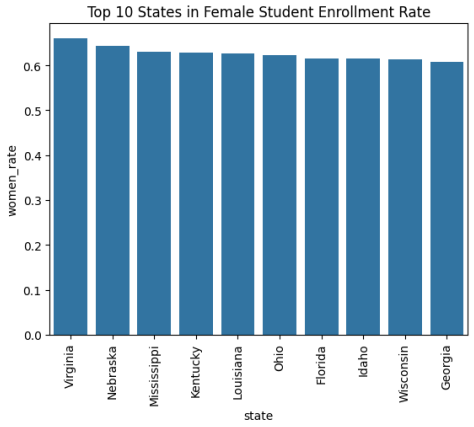

Figure 2

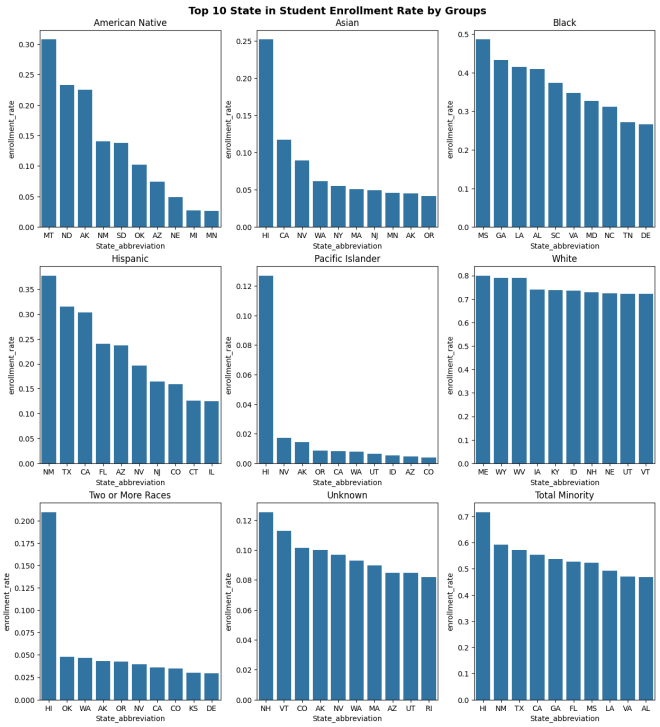

Figure 3

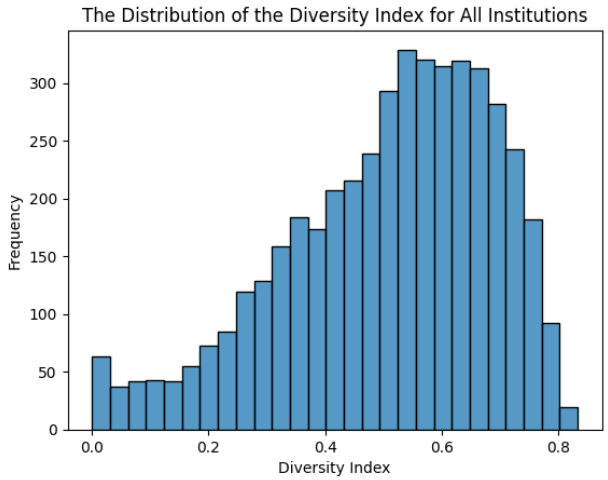

Figure 4

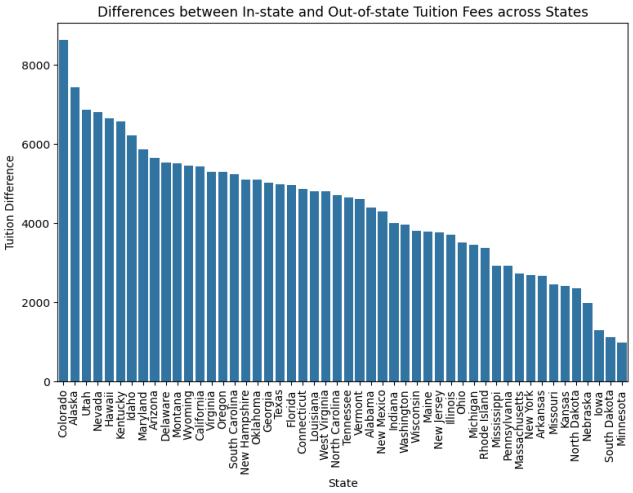

Figure 5

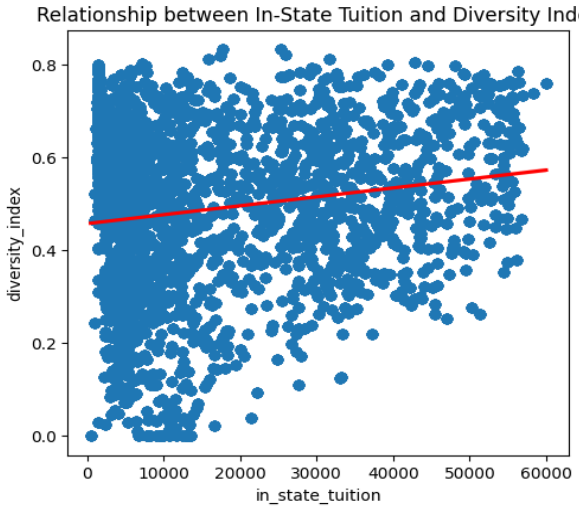

Figure 6

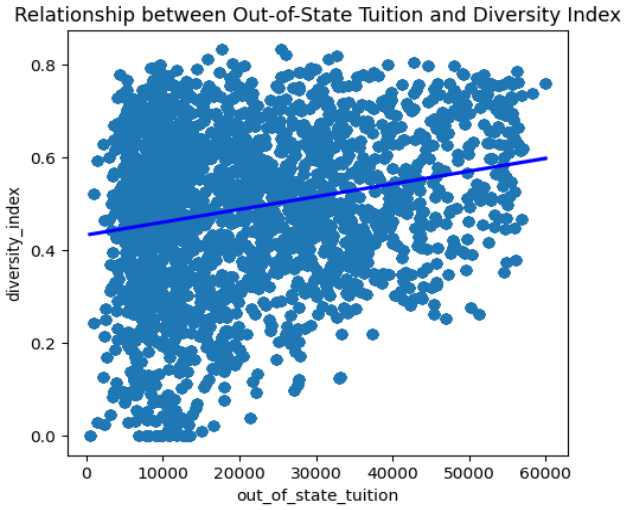

Figure 7

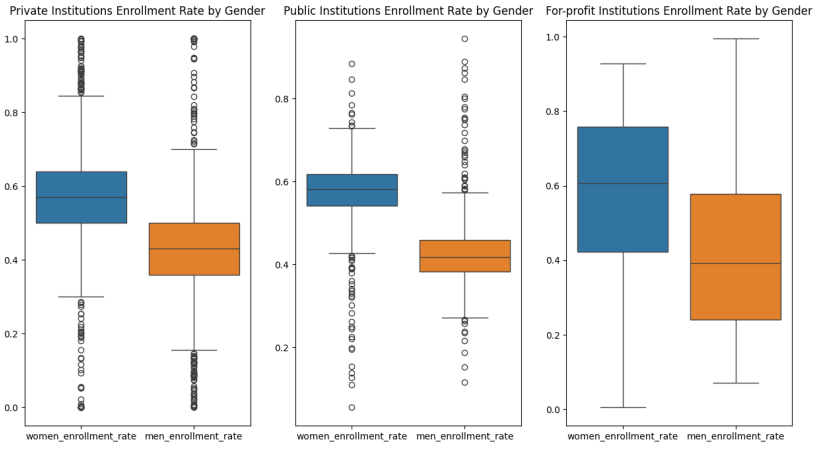

Figure 8

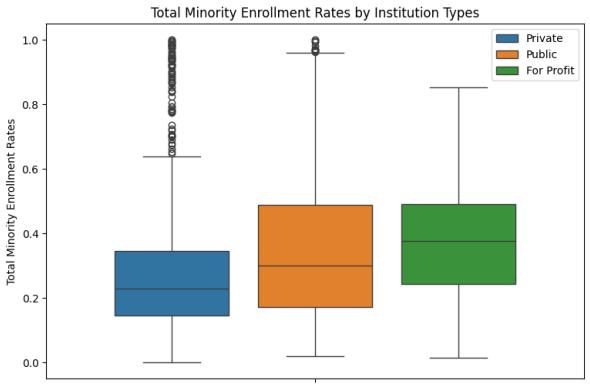

Figure 9

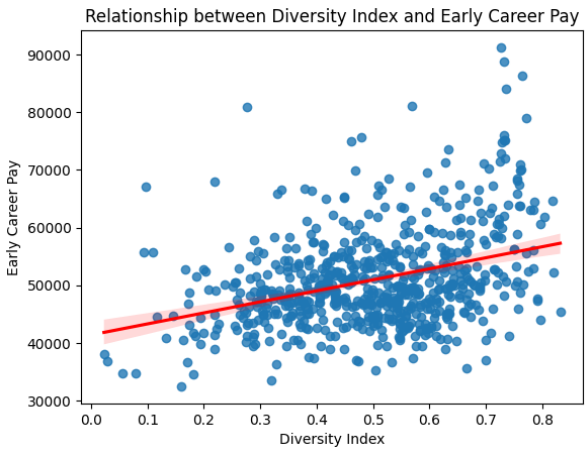

Figure 10

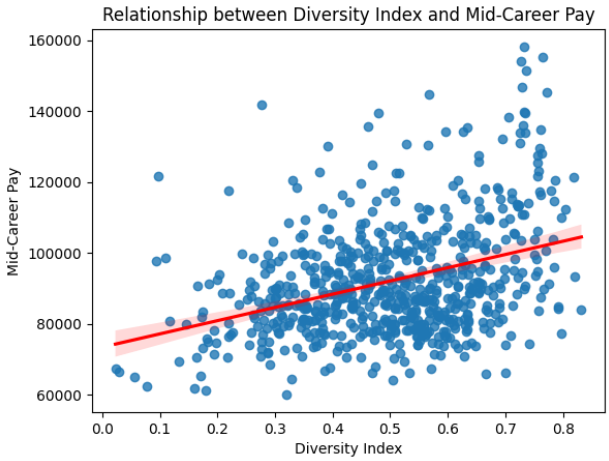

Figure 11

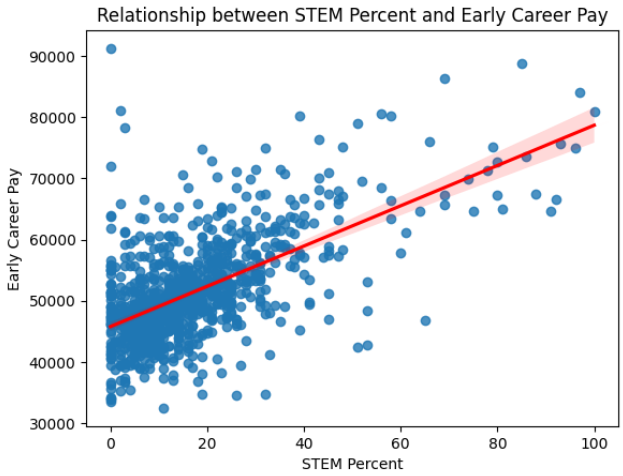

Figure 12

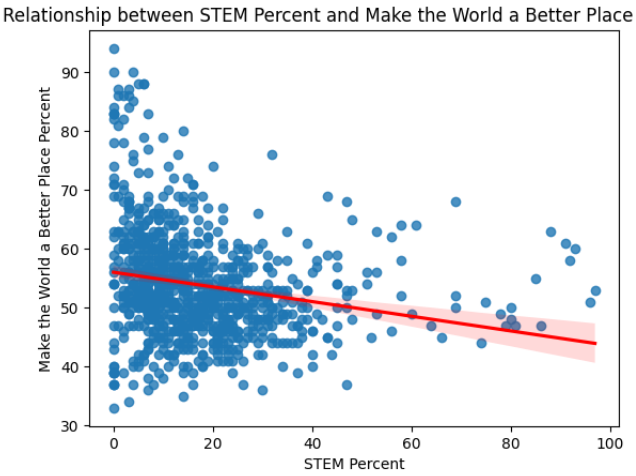

Figure 13

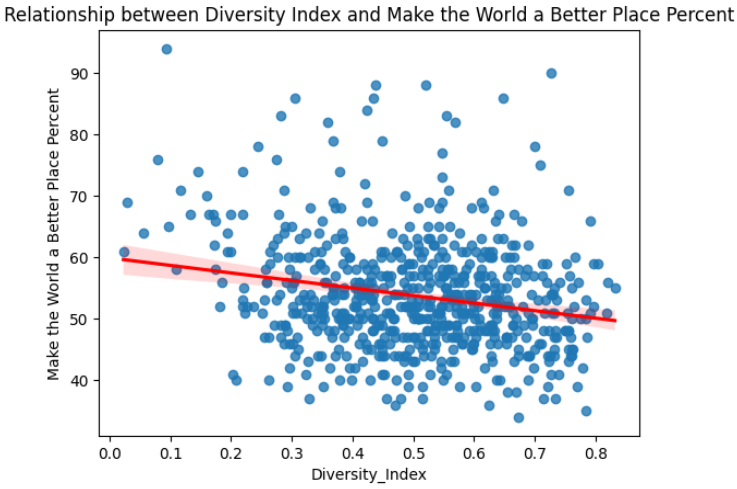

Figure 14

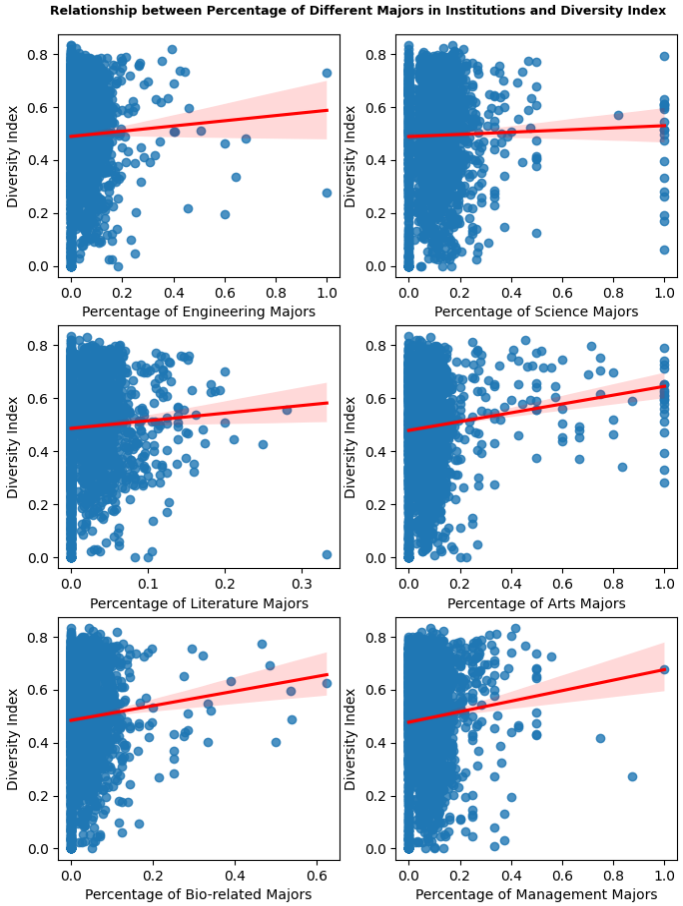

Figure 15

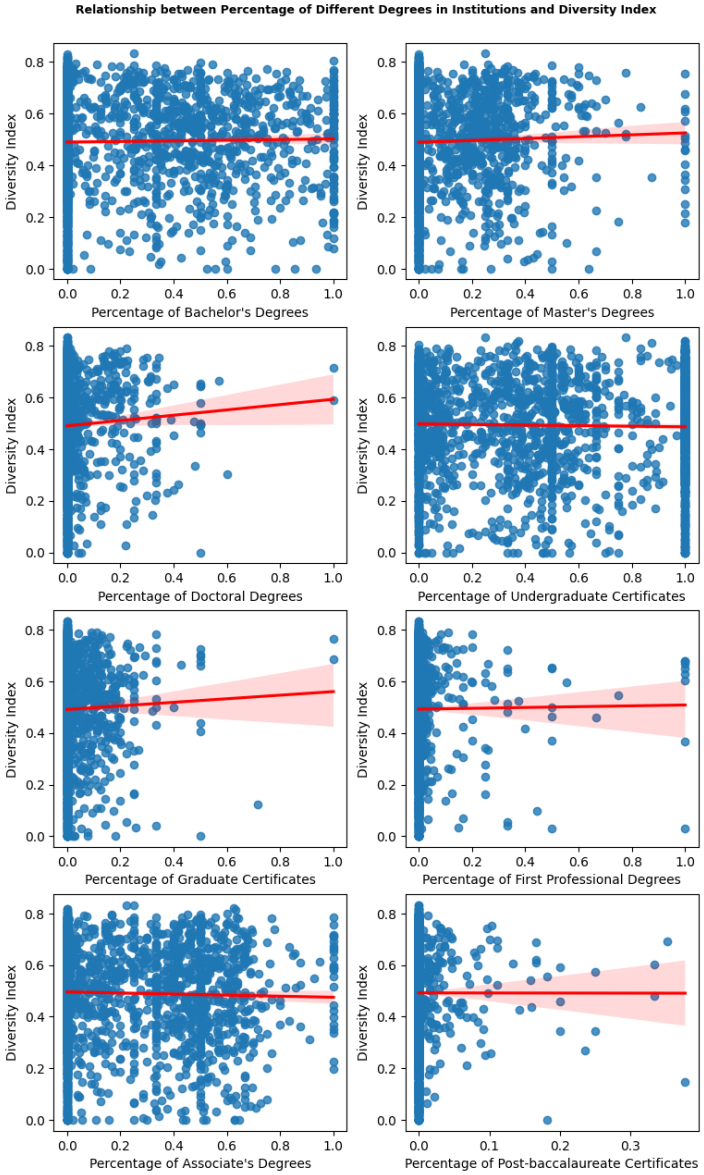

Figure 16

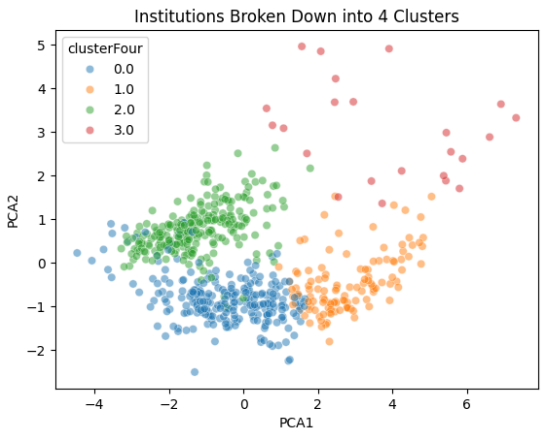

Figure 17

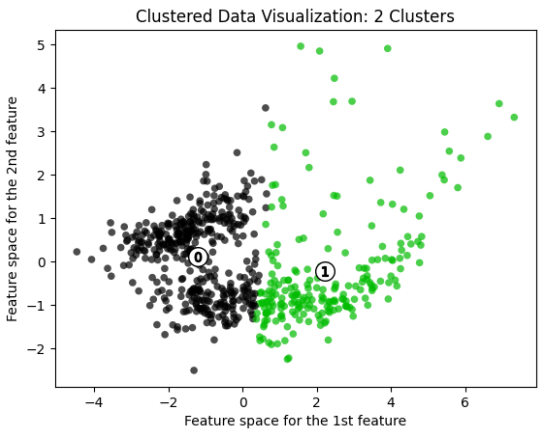

Table 1

Predicted “make_world_better” Percentage of 17 Institutions with Initial Missing Values

name	Make_world_better_percent (%)
Amridge University	60.72
Arizona Christian University	58.31
Central Baptist College	60.70
Philander Smith College	58.33
Nazarene Bible College	65.38
Holy Cross College	57.44
Allen College	67.74
Maharishi University of Management	53.71
Manhattan Christian College	68.47
Central Christian College of Kansas	57.30
Thomas More College of Liberal Arts	64.55
Webb Institute	52.57
Kettering College	78.29
Pacific Northwest College of Art	48.87
Goddard College	68.29
Bastyr University	60.35
Clark College	58.56

Table 2

Mean Values of Features of Four Clusters of Institutions by Agglomerative Clustering

cluster	early_career_pay ($)	mid_career_pay ($)	make_world_better_percent (%)	stem_percent (%)	diversity_index	in_state_tuition ($)	out_of_state_tuition ($)
0	50272.0	91017.0	53.58	17.13	0.49	26799.0	31152.0
1	55386.0	100308.0	53.95	18.61	0.57	31895.0	35707.0
2	49812.0	90000.0	53.56	16.01	0.48	24765.0	29476.0
3	51850.0	93764.0	49.50	17.36	0.58	26206.0	31811.0

Table 3

Mean Values of Features of Four Clusters of Institutions by K-means Clustering

cluster	early_career_pay ($)	mid_career_pay ($)	make_world_better_percent (%)	stem_percent (%)	diversity_index	in_state_tuition ($)	out_of_state_tuition ($)
0	49828.0	90212.0	53.73	16.66	0.48	25497.0	29980.0
1	53468.0	96603.0	53.05	17.56	0.55	29581.0	33943.0

Statement of Work

Our group worked well together, demonstrating unwavering commitment, effective communication, and adherence to project deadlines. Regular synchronous collaboration on Zoom facilitated efficient discussions and enabled prompt resolution of any arising questions. Each team member contributed to the project by leveraging their diverse skills, focusing on areas where they excelled. This collective effort and interaction were important in achieving our project goals successfully. The statement of work for each team member is listed below:

Yi-Chun Wang (ritaycw); section 001:

- Data cleaning and manipulation: code 40%, report 70%
- Visualization: code 30%, report 30%
- Statistical analysis: code 30%, report 30%
- Formatting and proofreading: code 10%, report 30%

Je-Ching Liao (jeching); section 101:

- Data cleaning and manipulation: code 40%, report 15%
- Visualization: code 50%, report 25%
- Statistical analysis: code 50%, report 20%
- Formatting and proofreading: code 80%, report 35%

Yi Hsien Wu (yihsien); section 101:

- Data cleaning and manipulation: code 20%, report 15%
- Visualization: code 20%, report 45%
- Statistical analysis: code 20%, report 50%
- Formatting and proofreading: code 10%, report 30%

References

A python list of all US state abbreviations. (n.d.). Gist. https://gist.github.com/JeffPaine/3083347

College tuition, diversity, and pay. (2020, March 9). Kaggle. https://www.kaggle.com/datasets/jessemostipak/college-tuition-diversity-and-pay?fbclid=IwAR12jFjcMK1UAquA7da7H2oLxy52TlcdGqURHh-dIxgAjH8lcBGp3QMWZ4c&select=diversity_school.csv

Data Home | College Scorecard. (n.d.). https://collegescorecard.ed.gov/data/?fbclid=IwAR2Z_LfBwuh2tXLfN0pn9zTT58ZR-Ihj8nILSrFqIyFe1Au8jXsEAhrmf8g

Pappas, J. (2020, November 20). How to calculate Simpson's Diversity Index (AP Biology). *Biology Simulations*. https://www.biologysimulations.com/post/how-to-calculate-simpson-s-diversity-index-ap-biology