Milestone 3 Write-up

Group 65 - Minghao Chen, Owen Lin, Jacob Yu, Xinjie Yi

1. Summary of the Data

Source: College Scorecard provided by U.S. Department of Education (Link)

Release Date: October 10, 2023

The College Scorecard dataset is a comprehensive collection of higher education institutions and field-of-study information intended to assist prospective students, families, and educational policymakers. It promotes transparency by highlighting college performance metrics and outcomes.

Data Structure:

- Institution-Level Data: This includes a broad range of data elements that provide detailed information about individual higher education institutions.
- Field-of-Study-Level Data: This offers in-depth insights into specific fields of study, detailing course outcomes and other relevant statistics at various institutions.

For our project, we are focusing on the Institution-Level Data from the latest release. Comprising over 3,232 variables, the dataset encompasses a diverse range of data types, including both numerical and categorical variables. We select the 10 most representative variables that are interpretable and interesting to analyze and draw predictions before conducting exploratory data analysis to filter out some variables to form a more concise subset. The dataset thus far contains 6,543 rows and 10 columns.

Here are the explanation and descriptive statistics for the 10 variables:

1) INSTNM (Institution Name): String (as an ID for each data point)
2) INEXPFTE (Instructional Expenditure per Full-Time Equivalent Student): Float
3) ADM_RATE_ALL (Overall Admission Rate): Float
4) MD_EARN_WNE_4YR (Median Earnings of Students Working and Not Enrolled 4 Years After Graduation): Float
5) MEDIAN_HH_INC (Median Household Income): Float
6) FAMINC (Average Family Income): Float
7) MD_FAMINC (Median Family Income): Float
8) STUFACR (Student-Faculty Ratio): Float
9) UGDS_MEN (Undergraduate Men Percentage): Float
10) UGDS_WOMEN (Undergraduate Women Percentage): Float

(ISTNM is excluded in the description since it will not be involved in our data analysis and predictions)

	INEXPFTE	ADM_RATE_ALL	MD_EARN_WNE_4YR	MEDIAN_HH_INC	FAMINC	MD_FAMINC	STUFACR	UGDS_MEN	UGDS_WOMEN
count	6024.000000	2224.000000	5637.000000	5145.000000	5911.000000	5911.000000	5758.000000	5769.000000	5769.000000
mean	8759.862716	0.728743	40373.035835	58177.592089	40125.036241	29465.280156	14.967871	0.349934	0.647640
std	13306.209710	0.226722	16731.101079	12827.502697	23777.164631	19886.012235	6.998611	0.246280	0.247743
min	0.000000	0.000000	8915.000000	15790.530000	321.385321	0.000000	1.000000	0.000000	0.000000
25%	3854.500000	0.612580	27864.000000	50299.150000	23560.779029	16869.000000	10.000000	0.134700	0.530800
50%	6377.500000	0.780102	39282.000000	57880.380000	33136.902181	23028.000000	14.000000	0.361100	0.638000
75%	9894.500000	0.900904	49321.000000	66472.610000	50432.813711	35952.000000	18.000000	0.467500	0.863000
max	598759.000000	1.000000	139418.000000	100870.750000	174263.250000	179864.000000	112.000000	1.000000	1.000000

We examined the number of missing entries per variable. “Admission Rates” has a very high missing rate of over 50%. Our group thinks that with such missingness, no imputation method would truly solve the problem. So, even though it might be a significant variable, our group decided to not include it due to this reason.

2. Exploratory Data Analysis

On the list of 8 variables (since we have excluded “Institution Name” and “Admission Rates”), we conducted EDA to gain insights into their relationships.

1) Histograms

   To analyze the distribution of the variables and decide on the imputation methods, we plot (and try to fit with a curve) some histograms of some selected variables (See next page).

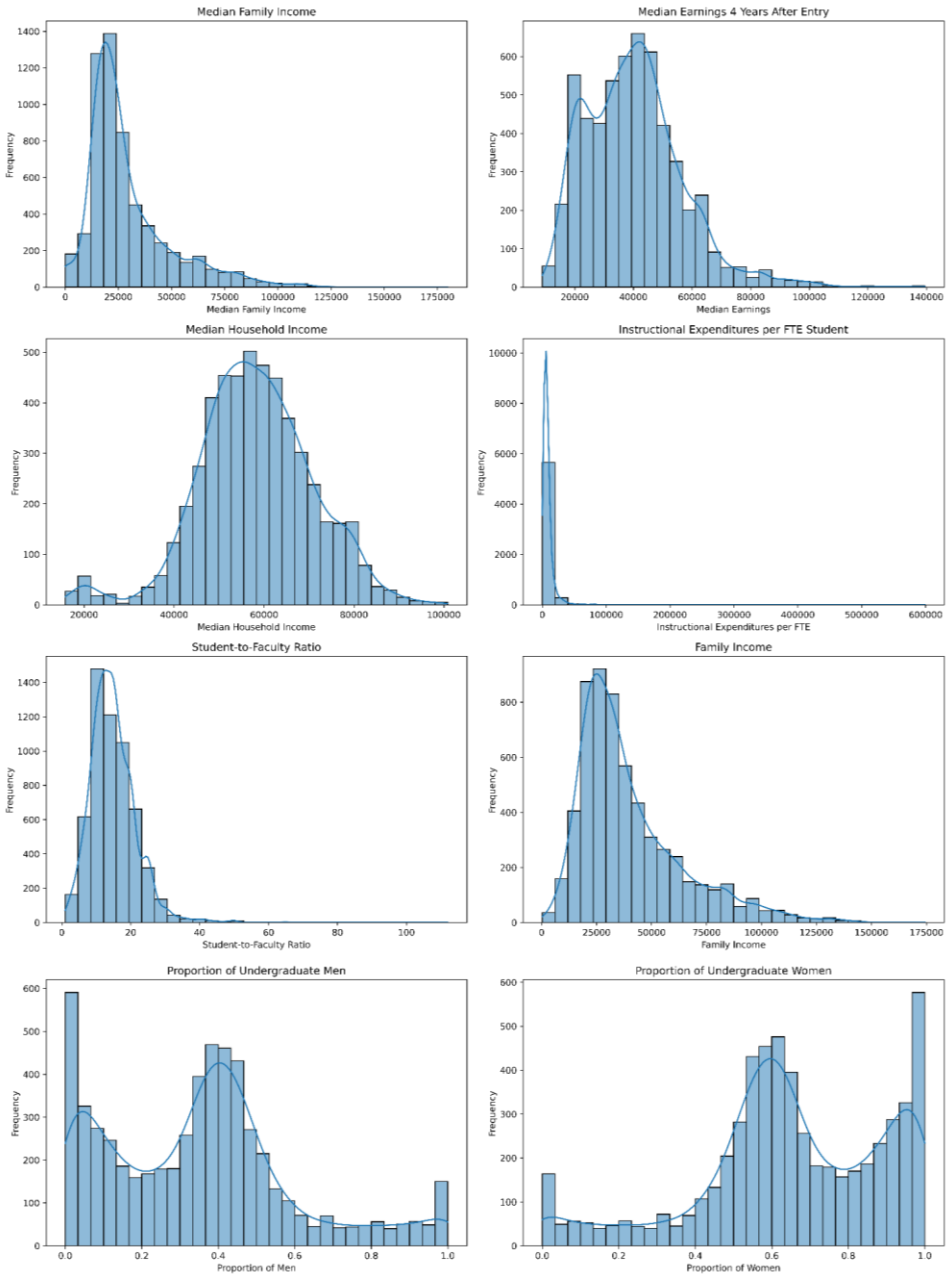
   We further selected variables by analyzing the distributions shown by the histograms. "Median Household Income" was chosen over "Median Family Income" and "Average Family Income" because it not only serves as a good proxy for both but also exhibits the most normal distribution, which could be pivotal in our statistical modeling.

   When considering gender proportions, we recognized that the variables "Proportion of Undergraduate Men" and "Proportion of Undergraduate Women" are complementary. We opted to include only "Proportion of Undergraduate Men" for simplicity, acknowledging that one could infer the other. Additionally, our group thought about the inclusion of non-binary genders into our model (by subtracting the proportion of men and women from 1). However, due to the minority proportion of the group, we were afraid that the measurement error would exceed the actual proportion.

   Consequently, we narrowed our variables to five: "Median Household Income," "Median Earnings 4 Years After Entry," "Instructional Expenditures per FTE," "Student-to-Faculty Ratio," and "Proportion of Undergraduate Men."

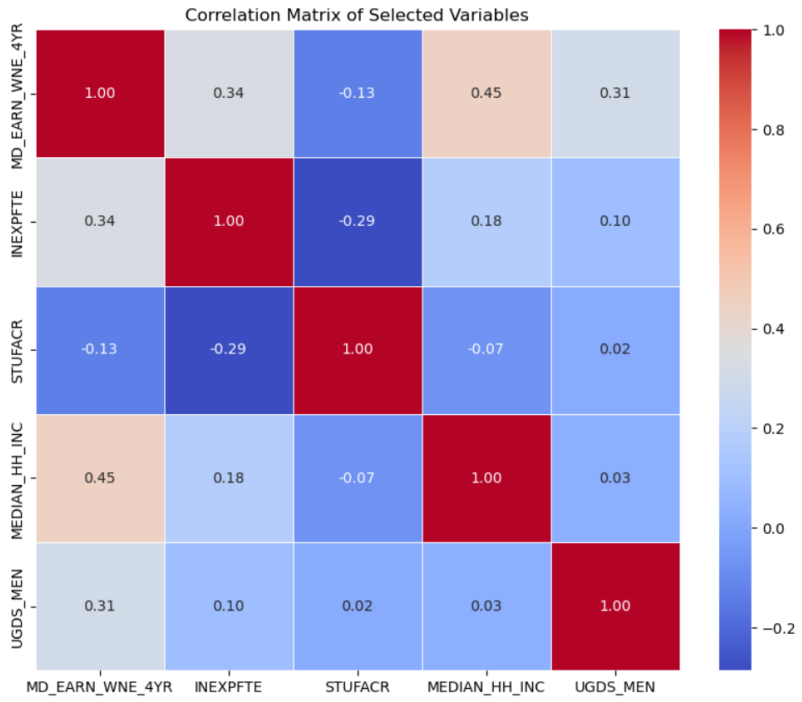
   For missing data, we plan to perform imputation according to the different natures of each variable:

- We have observed that for 'Instructional Expenditures per FTE' and ‘Median Earnings 4 Years After Entry’, the distributions show some degree of skewness, suggesting that median imputation would be a better choice.
- As for the ‘Median Household Income’ and ‘Student-to-faculty Ratio’, their distributions seem fairly normal, so both mean and median imputation would be suitable.
- As for the ‘Proportion of Undergraduate Men’, it shows a bi-modal pattern with two peaks, so we will choose the mode imputation or the KNN imputation.



2) Correlation Analysis

Next, we computed a heatmap to analyze the correlation between our 5 selected variables:



Note that only "Median Earnings 4 Years After Entry" has relatively high correlations with other variables. When "Median Earnings 4 Years After Entry" is not involved, the highest pairwise correlation is between "Student Faculty Ratio" and "Faculty Expenditure per Student" (this makes intuitive sense because as we have fewer students, there is more expenditure per student). We might want to do variance inflation factor (VIF) analysis later on, but our group thinks correlation under 0.3 is not severe enough to yield collinearity problems.

Based on these observations, we select "Median Earnings 4 Years After Entry" as our response variable. As a sanity check, other variables do have some kind of relationship with it intuitively. We can thereby formulate our research question as follows: "How are the median earnings 4 years after entry of a US college affected by its faculty expenditure per student, student-to-faculty ratio, median household income, and the proportion of male students?"

3. Baseline Model and Implementation Plan

Given that our response variable is continuous, we are addressing a regression problem. Our initial approach will involve training a linear regression model as our baseline. This simple and

interpretable model will help us establish a fundamental understanding of the relationship between predictors and the response variable.

For the implementation, we will begin by preparing the dataset, which includes splitting the data into training and testing sets, selecting key variables based on domain knowledge and preliminary data analysis, handling missing values through appropriate imputation methods, and normalizing or standardizing the variables to ensure consistent scale. The model will be trained using the training set and evaluated using the testing set. We will employ cross-validation to validate the model's performance and ensure it is not overfitting. Our model will be trained on the Mean Squared Error (MSE) objective, and its performance will be evaluated using R-squared. This iterative process of testing, evaluation, and refinement will help improve the model's performance, and the findings, along with insights derived from the model, will be thoroughly documented for review and further decision-making. After initial trials, we might as well try regression models using the non-parametric K-nearest neighbors, decision trees, and ensemble methods.

4. 209 Component

We are contemplating enhancing our project with additional analyses and features. Firstly, we aim to incorporate models not discussed in our class. Specifically, for regression analysis, we are considering Support Vector Regression (SVR). SVR, derived from Support Vector Machines (SVMs), is tailored for max-margin regression tasks. Additionally, we plan to introduce an interactive element to the project. This could be particularly beneficial for college applicants looking to find suitable schools. They would have the ability to input a school name, and our model would then output a list of schools that best align with the input school. This can be achieved by computing the distance in some latent space, or running some (hierarchical) clustering algorithm to obtain groups of similar schools from coarse to fine.