Project Proposal

Proposal Statement

We are Kaleb Cervantes, Vidya Pingali, and Felicia Kuan, three juniors at SCU interested in analyzing how colleges in America give **financial aid** to students.

Financial aid is funding exclusively for college students in US. Financial aid is available from federal, state, educational institutions, and private agencies, and can be awarded in the forms of grants, education loans, work-study and scholarships. In order to apply for federal financial aid, students must first complete the Free Application for Federal Student Aid (FAFSA). Colleges then use a student's FAFSA along with the college's resources to determine the aid given to a student.

We think this is interesting as college admissions is very relevant, especially in the wake of all the college scandals going on. Additionally, student loans and student debt is a very hot topic and we want to get a deeper insight into how colleges assign grants to students. Many students end up going to a particular college solely based on the aid and scholarships it gives them, regardless of whether or not they wanted to go there. Many students across the nation depend on financial aid to go to college.

Santa Clara University is a very expensive school, costing approximately $70,000 a year. Since it is a private university, we know many students receive some sort of financial aid to attend here. It would be interesting to us, as SCU students, to know how our school gives financial aid to its students and see how it compares to other schools. It would be interesting to compare SCU's financial aid process to other schools, such as UCs, other private schools similar in size, or other Jesuit schools.

We want to predict how much percentage of their accepted freshmen students a college will give financial aid to based on a number of factors such as:

- Location (state, zip code, county, etc.)
- Private or public school
- Tuition fees
- Historically black college
- Religious affiliation
- Racial breakdown of students (this can be correlated to family income)
- Types of degrees the college offers
    - 2-year vs 4-year
    - Bachelor's vs graduate degree
- Number of total students

- Average ACT/SAT scores of freshmen
    - Scores shouldn't technically matter for a student's financial aid since it's based on family income. However, scores definitely determine how elite a school is, which can affect their endowment.
- Carnegie classification (classifies accredited colleges into certain categories)
- Degree of urbanization (city, town, rural, suburb)
- Acceptance rate, etc.

Since financial aid usually carries over from year to year, we are just looking at the percentage of freshmen that receive any financial aid.

We plan to use the "American University Data" IPEDS dataset. We originally found it on Kaggle (https://www.kaggle.com/sumithbhongale/american-university-data-ipeds-dataset/home), which lists information about it, but the actual dataset is found on Tableau Public (https://public.tableau.com/en-us/s/resources). This dataset is from 2013.

Data Collection Proposal

We are obtaining our data from Tableau Public and it is called "American University Data" IPEDS dataset. It is a public dataset based on admissions from 2013.

The dataset is 1534 rows and 145 columns large. The rows list the colleges so it has data from over 1500 colleges, all in America. The columns list the features, such as number of accepted applicants, tuition, types of degrees offered, etc. When we start analyzing the data, we will delete unnecessary columns.

The raw data is stored as an Excel spreadsheet but for this project's purposes, we are going to convert it into a Python dataframe. Yes, we will be processing the original data to get it into an easier, more compressed format. Luckily, the Excel spreadsheet was pretty organized and easy to read so it was simple and quick to convert it into a Python dataframe. When we actually start the project, we will delete quite a few columns, of course, that are not significant to our project's goals.

There are numerous columns about grants such as:

- **Percent of freshmen receiving any financial aid**
- Percent of freshmen receiving federal, state, local or institutional grant aid
- Percent of freshmen receiving federal grant aid
- Percent of freshmen receiving Pell grants
- Percent of freshmen receiving other federal grant aid
- Percent of freshmen receiving state/local grant aid

- Percent of freshmen receiving institutional grant aid
- Percent of freshmen receiving student loan aid
- Percent of freshmen receiving federal student loans
- Percent of freshmen receiving other loan aid

We could have chosen to do any of these aids but we chose financial aid because that is the general category and the ones most people, including us, are familiar with in terms of student aid. Therefore, we can delete the rest of the columns because we will not be analyzing those.

For example, there are columns for the state the college is in, as well as the zip code, county, state abbreviation, FIPS state code, as well as latitude and longitude coordinates. They all give us the same information so we can combine some of them to help us. We think the state is useful as well as the zip code. For example, SCU and Loyola Marymount University are often compared as being very similar. They are both private, relatively small in size, Jesuit affiliated, and in California, but in different zip codes and want to see how that might affect the financial aid given by the universities.

There are numerous other columns we can delete as well about graduate enrollment because we are focusing on freshmen and do not see a huge correlation between them, even though there might be one. We can delete the year column because the whole dataset is from 2013. We can also delete "Sector of Institution" and "Level of Institution" because they are all 4 years or more colleges and just repeats the same information as "Control of Institution." We can also delete "Tribal college" because none of the colleges in our dataset are.

Then, we'll have to clean some of the data points, particularly the ones that give us ambiguous information. For example, for a certain university that offers a doctor's degree, there are several types of the degree-- research, professional practice, and other. These all have different attributes and some of the data simply says "implied no" instead of a clear "yes" or "no." For these types of ambiguous data, we will likely have to write code to parse and change these responses to either a hard no, or actually go through and verify the information ourselves.

For now, we plan to simulate our data by splitting our data, about 80% training and 20% testing. It would be hard for us to make up fake colleges with all these factors to test our data so we think it's a better idea to just split our dataset.

Evaluation Report

We chose this problem because we wanted to do something related to college as it is pertinent in our lives and everyone in the class as well. From the dataset we chose, the features that stood out

to us the most were the ones regarding aid as it is applicable in today's economic and political spheres.

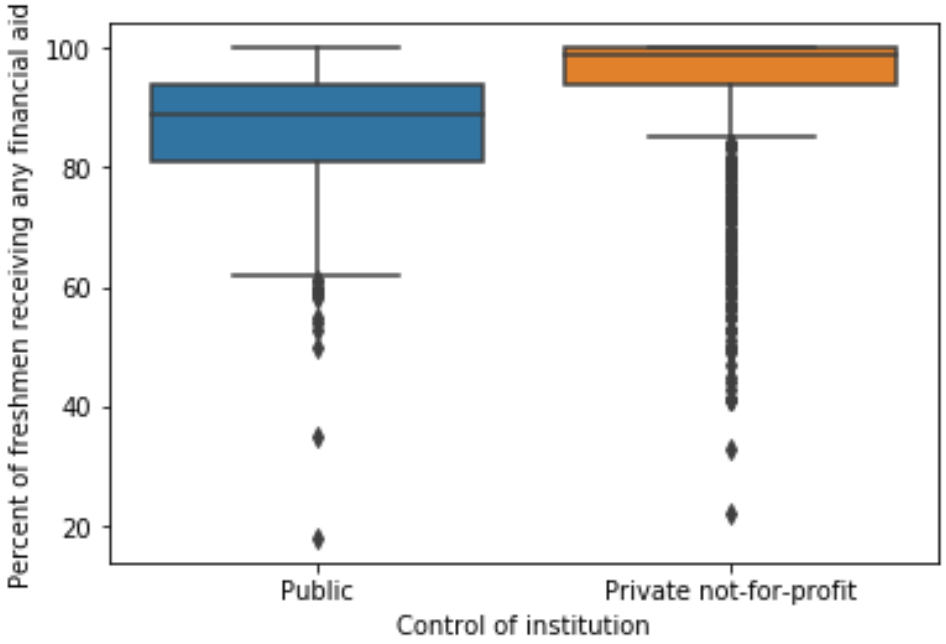
Before we even start our analysis, we can hypothesize that private schools give more students financial aid due to large endowment from donors and alumni. Although it's not represented by the data, we predict that the amount of famous and successful alumni for a university will also influence the percentage of financial aid the student body will receive, particularly because these alumni will have more to donate. However, since this is not represented by the data, we can only make assumptions about this after we determine which universities give the highest percentage of financial aid.

Religiously affiliated schools such as SCU also give many students financial aid because of donations from religious institutions, so we predict that the religious affiliation of the school will influence percentage of financial aid.

We also hypothesize that degree of urbanization will be interesting to note. There is a big correlation to control of institution and size of school. Schools in cities are often big public schools and might not give a lot of financial aid, whereas smaller, usually private, schools in rural or suburban areas generally give more students aid. For example, Ivy Leagues or the liberal arts colleges in South California such as Claremont McKenna or Scripps College are all in smaller towns and will probably give more students aid.

First, we need to begin by parsing the data. Even though the dataset is an Excel spreadsheet, it's not very convenient to keep reading it and is fairly large. We are going to convert it into a Python dataframe and work on it in Colaboratory so we can all collaborate on the same project. Additionally, we're going to use the attribute "state" to help us categorize each of the universities into their respective states. This will help us make predictions in the next step.
Our main metric will be the percentage of students who receive any type of financial aid, given to us for each school by the attribute "Percent of freshmen receiving any financial aid." If we want the actual number of students who receive financial aid at each university, we can use that percentage and do some simple arithmetic using the total amount of enrolled freshmen to obtain that value.

In our attempt to identify which attributes have the most influence on the main metric, we plotted several attributes together on plots. First, we plotted the amount of students from public and private universities and the percentage of freshmen that receive any financial aid. This graph is displayed below.



Just like we predicted, private schools give more students financial aid than public schools do. This dots represent the outliers and while it might seem like there are a lot, the majority of the data points lie in the boxes. The horizontal line in the box is the average. The average for public universities is around 89.56% and the average for private universities is around 98.83%, which is a considerable disparity.

Then, the Machine Learning method we could try using Multivariate Regression because we have many variables and factors we are using to predict our outcome. For example, we can use a college's state, zip code, control of institution, and degree of urbanization to determine the percentage of students that will receive financial aid.

However, we have tentatively decided to use was K-Nearest Neighbors Linear Regression. We observed that our data is continuous-- not categorical-- so decision trees would not be useful to make predictions. Therefore, since the data is continuous, it would be most helpful to predict based on a regression line. However, we have decided not to use a simple regression line because we observed that due to how different types of universities are run, there are various factors that more important to some that are likely not as important to others.

For example, the amount of funding a state university receives is dependent on the state (location), which will undoubtedly affect how much financial aid is available to allocate to students. The location of the school is not as important to religious universities, as they do not rely on state or city funding, so the location attribute will carry less weight. We guess that

depending on the type of university, the attributes to are observing to influence the aid given to students will carry different types of weight, but we don't exactly know how much weight each of these attributes will amount to for the final decision. In an attempt to reconcile this complication, we are going to use KNN Linear Regression to group universities with others of a similar type-- for example all the private, religiously affiliated and rural universities closer together.

In conclusion, we are very excited to find the results of our project. We think it is applicable and hope the class will find it interesting too.