Student Loan Project Report

Kemal Fidan, Logan Courtney, and Chandler Lampe

***Abstract*— This document breaks down our project which analyzes student loan data. We explain our objective, what gave us the motivation to do this project, where we obtained our data and what we plan to do with it, our findings, each of our group member's responsibilities, the timeline we expect to follow in order to complete this project, and the expected outcome we hope to achieve.**

I. OBJECTIVE

Using several data sets, we will be analyzing different aspects of the data to find correlations and trends in student loans. We expect to find general trends which will be interesting information for people who have student loans or are about to have student loans.

The main objective in our project is to find if any trends exist in our data sets, and if so, how those affect student loans. We also have a stretch goal set for the project. As a long-term objective (if we have time for it), we want to see if we're able to predict student loans given some background about a borrower.

Specifically, this project is interested in the following trends:

1) Average Loan Per State
2) Average Loan For Different Types of Institutions (Public, Private, etc)
3) Average Loan Vs Year – look at change over year
4) Average Loan Vs Cost of School
5) Average Loan Vs Size of School
6) Loan Differences for Rank of Institution
7) Average Loan Vs ACT/SAT
8) Average Loan Vs School's Gender %
9) Average Loan Vs School's Race %
10) Average Loan Vs Size of City
11) Average Loan Vs Region

II. MOTIVATION

A large amount of people have student loans and new people receive student loans every year. As students, this is a topic that we can relate to very closely and have a personal touch with. For this reason, we wanted to look at the data and find useful information that could educate the general population on the statistics and analysis of student loans. Futhermore, Figure 1 below shows how student loan is still prevalent today.

Also, based on our preliminary research, there are not that many analyses on student loan data that we could find. So, we hope to provide new material and results that are backed by a couple data sets. Our final motivation with this project is to see if there's any attention we can bring to inequalities of student loan. This way, federal or private loans can help those that are disadvantaged.

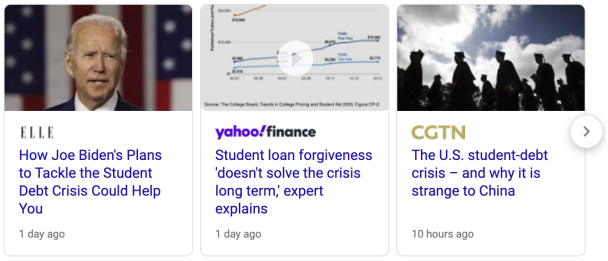

Fig. 1. Google search results for student loans. This shows how the topic concerns students and the economy today.

III. METHODS

A. Data

We got our main data set from https://studentaid.gov/data-center/student/title-iv. This student aid data set comes from the Federal Student Aid Office, which is a branch of the United States Department of Education. Since the Department of Education is a part of the United States government, the data set that is provided by them is extremely reliable and is expected to be 100% truthful.

In detail, this student aid resource provides a large amount of loan data dating from 1999 to 2020. Each row of data contains many features, like school name, school type, number of loan recipients, and the amount of the total loan for that school.

We also used a smaller data set containing demographics of universities around the United States. This data set from Kaggle https://www.kaggle.com/sumithbhongale/american-university-data-ipeds-dataset/home? contains more information related to a school. Things such as ACT percentiles, enrollment numbers, and student ethnicity will be found in this data set.

Though there are 2 separate datasets, each row in both datasets contains a school name. This school name is used as a foriegn-key like entity to combine the datasets together during runtime. To ensure that the maximum number of schools were matched, some preprocessing was doen to each of the school's name. First, the string was converted to lower case. Next, stop words were removed from every school's name. This was done so that the names of the schools would be most similar. Some stop words include: 'the', '(', ')', 'a', 'of', 'in', 'at', ',', "s'.

From these data sets we will be diving into some of the questions listed above. It's important to note that when we say "the average amount of loans", we mean the average

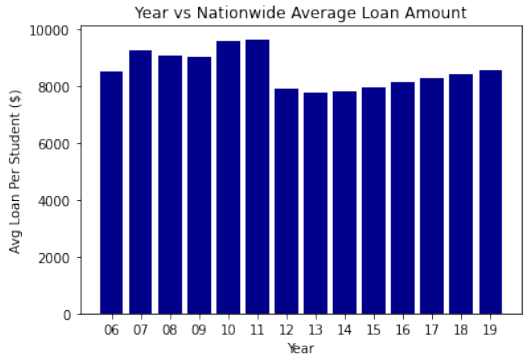

Fig. 2. Average loan per year nationwide

loan per person. The range of our comparisons is limited to two data sets as of now, but hopefully we will be able to find more data that can correlate to our data. Our goal is to combine these data sets into a single data set that contains demographics and student loan data.

B. Models

The primary tool that was used was xlrd. Xlrd is a Python package that makes it very easy to read Excel sheets. With xlrd, we were able to open dataset files in code, collect the data, and then aggregate the data however we needed to. Pandas dataframes were also used in several places to represent the data. One improvement that could have been done was to read the Excel file directly into a dataframe. The Excel files contained a strange header format that was difficult for the pandas library to read reliably.

IV. RESULTS

Our general findings look at how average student loans have changed over the years. Figure 2 shows this information in a bar graph. It's possible to see a steady increase in student loans over time. In 2012, there is a sharp decrease in the average loan per student nationwide. We concluded that that was an election year and that the Student Loan Forgiveness Act got passed, which played a big factor in the student loan amount's decrease in 2012. Overall, we were surprised to only see a steady increase in student loans.

We expected there to be sharp increase in the loans taken out since the student loan debt crisis is always a hot topic in the news. However, it was reasoned that loans are tied directly to the cost of tuition, the area, etc, and the numbers portrayed in the news is typically the outstanding loan debt. In other words, there is another concern on whether students are paying back their loans. It made more sense there was only a steady increase.

Next, we analyzed state-level loan metrics. Generally, cities like California and New York are known to be higher cost of living areas. Whereas states like Tennessee are relatively cheaper. For this reason, we assumed that students would have to borrow more loans for the higher cost of living. However, as seen in Figure 3, Tennessee's and New York's state averages for loans are almost equal from 2006 to

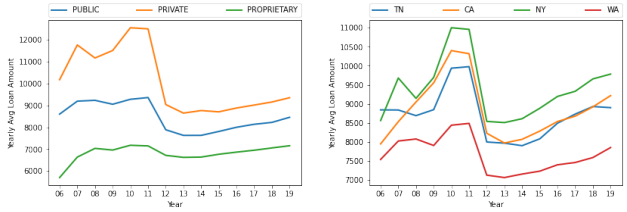

Fig. 3. School type and state vs average loan amount

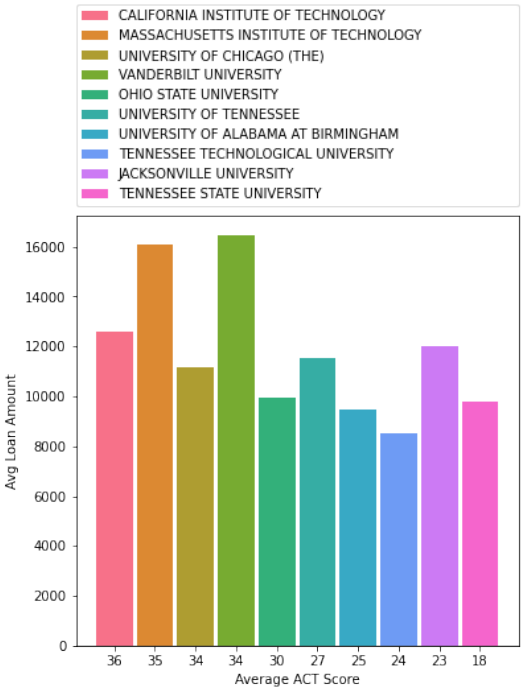

Fig. 4. Average loan for select schools

2019. This was a surprising finding. Our team concluded that schools in New York are generally higher class, and therefore might offer more scholarships to offset tuition costs.

The type of schools are also shown in Figure 3. This graph generally follows what we expected. Private schools are known to have higher tuition since they're for-profit. Though it was expected that private schools might offer more monetary aid to students, the average loan taken out by private school students is greater than public and proprietary school students.

Next, the effect of a school's average ACT score was analyzed. This was tough to hypothesize since smarter students probably go to better and more expensive school, but also might receive more aid in the process. Though there is a trend where lower ACT scores correspond to lower loan rates, there are some outliers like MIT and Vanderbilt. University of Chicago is known to be one of the most expensive colleges in the US, however, compared to other school's in its ACT score range, the students of University of Chicago borrow relatively less money.

Race and gender metrics were also analyzed on a school-level granularity. Figure 5 shows how the makeup of the

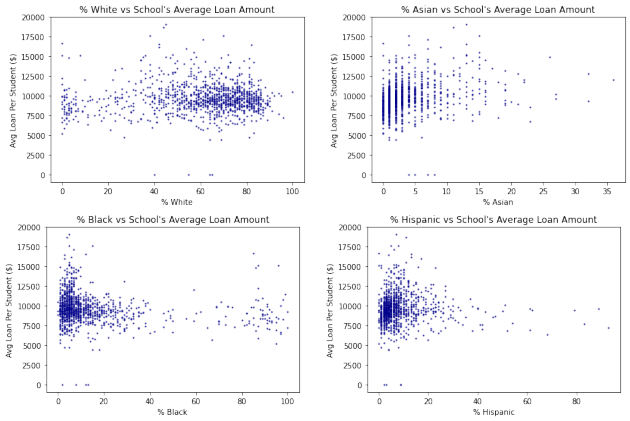

Fig. 5. School's % race vs average loan distributions

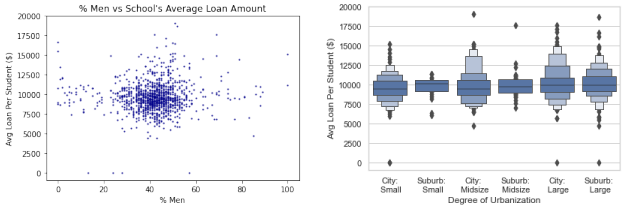

Fig. 6. School's % male and school's size graphs

race in a school corresponds to the average loan amount. In these graphs, it's difficult to see much of a trend in the data. At first glance, schools that have low Black and Hispanic races tend to have a higher average loan amount. However, this could be attributed to the fact that there is more data in those regions, so there is more of a change to get noisy data.

As for the effect of a school's gender makeup, Figure 6 shows that there isn't much of an effect of gender on a school's average loan amount. Figure 6 also shows the effect that the city's size might have on the loan taken out by students. This was gathered and split into 6 different categories. City and Suburb each have 3 different sizes:

1) Small: 0-100,000 population
2) Medium: 100,000-250,000 population
3) Large: 250,000+ population

School data based on these sizes is shown in figure 6. Schools that are located in a suburban area tend to have a less spread amount of loans per year. However, all school types have an a median at around 10k loans per year per student on average.

Figure 7 shows how each region varies in the amount of loans taken out. While the regions tend to trend together, the pattern shows which have the most amount of loans and which have the least. From the data, it can be seen that the northeast has the highest amount of loans. As discussed earlier, this trend might have something to do with the size of the cities. The northeast is more highly populated than the other regions. The southeast, on average, had the fewer amount of loans. Since the southeast is less populated than other regions, this makes since.

Figure 8 displays the top 10 largest colleges in the U.S

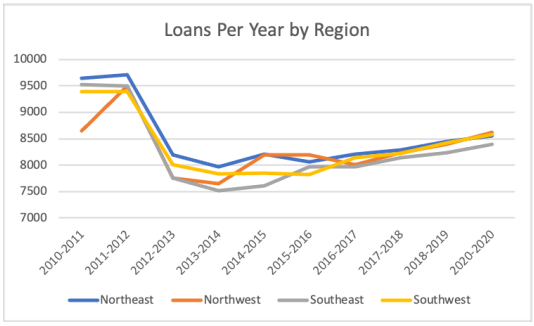

Fig. 7. Average loan per Region

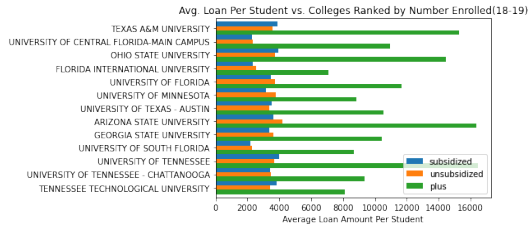

Fig. 8. Average loan vs. Largest Colleges

based on the number of students enrolled. The bottom three colleges were added to provide a familiar reference. The colleges are ordered largest to smallest, with Texas A&M being the largest and University of South Florida being the smallest of the ten. It can be seen that the subsidized and unsubsidized loans were relatively similar in all of the colleges. However, the plus loans were very spread out and didn't correlate to the largest colleges. A larger list of colleges ranked in size could have allowed for an obvious trend to show.

We believed college ranking could have an effect on the average amount of student loans a student takes out. In Figure 9, the graph list the top 9 ranked colleges in the U.S. Again, the bottom three colleges are for a familiar reference. The data seems to show that the average amount of loan is decreasing as the rank decreases, but I would say it is again too small of a list to fully conclude that. Compared to the average loan amount of UT students, we can see that you can expect to add $5,000 to that if you were attending one of the top 9 ranked colleges.

With a more extensive list of colleges in rank of tuition

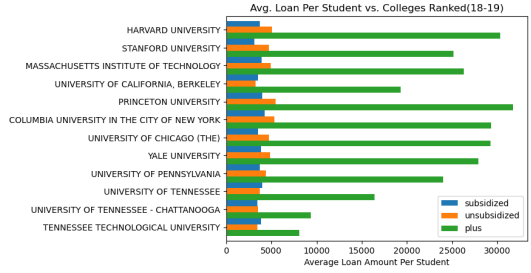

Fig. 9. Average loan vs. Ranked Colleges

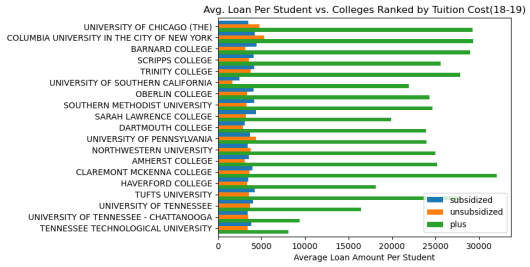

Fig. 10. Average loan vs. Tuition cost of Colleges

cost, we were able to compile the data into Figure 10. Again the data shows that most of the subsidized and unsubsidized loans amounts are similar. The data also shows that the amount of loans increases as the cost of the tuition increases, with the exception of private institutions. This was the expected outcome, but we thought that scholarships could be a factor as tuition costs increased.

V. ISSUES

Our primary issue that we encountered was that it was difficult to merge datasets together. The dataset from the federal government only contained loan data, school name, and school location. However, to merge other data such as ACT scores, it was challenging to combine completely separate datasets. We decided that we didn't need the entire dataset of ACT scores, so we handpicked a range of ACT scores and used those instead of merging datasets.

Finding datasets were also challenging. Since we were answering a wide range of questions, it was sometimes challenging to find prepared dataset with the information needed to answer the questions. We could have harvested our own data, but it wouldn't be worth going through all that trouble just to answer one of the many questions we had. For questions like the ACT dataset, metrics were collected manually. We wished that we could find data around ethnicity, race, and other demographics for each person so we could see how each of these effect loans on a more fine granularity.

VI. LIMITATIONS

Since there is multiple ways of getting a loan, like local, federal, private aid, it would be great to get the entirety of student loans. However, the data in our data sets only considers federal loans. Also, since our data is from the United States Department of Education, we only consider student loans within the United States.

VII. FUTURE WORK

A stretch goal that our team didn't have time to implement was a machine learning algorithm to predict student loan data given some information about the borrower. We found it challenging to find information about student demographics. However, with that information, the school the student is interested in attending, and other financial factors, it would definitely be possible to predict (on a student-to-student basis) how much loans the student would need to allocate. A better way to do this task would be to college data on a student-level granularity rather than school-level.

Other ideas we had in our post-project analysis was to keep data in a database. Our aggregation of the student loan data was done manually in Python. Also, the files were very large so sometimes it was hard to view contents in Excel. Finally, the information we gathered would be beneficial to the public. A site or article to hand info to the public would be a nice method of passing on what we've learned in this project.

VIII. ORG CHART

This section contains a rough timeline and responsibilities for each member.

A. Responsibilities

Our responsibilities were evenly distributed throughout all of the group members. Since we were investigating a lot of different questions, each member was be assigned a different set of problems to analyze. This way, the work for the project can be split evenly through the members.

In addition to that, we had officers to facilitate timely delivery of the project. The roles and assigned questions are shown below:

- Logan Courtney
  - Team Leader
  - Average Loan Vs Cost of School
  - Average Loan Vs Size of School
  - Loan Differences for Rank of Institution
- Kemal Fidan
  - Project Manager
  - Average Loan Per State
  - Average Loan For Different Types of Institutions (Public, Private, etc)
  - Average Loan Vs Year – look at change over year
  - Average Loan Vs ACT/SAT
  - Average Loan Vs School's Gender %
  - Average Loan Vs School's Race %
  - Average Loan Vs Size of City
- Chandler Lampe
  - Deadline Manager
  - Average Loan Vs Region

With these responsibilities, we made sure that members are working on their tasks, that we meet deadlines, and that our project was finished in time.

B. Timeline

Information about the timeline was gathered from GitHub commits, which provide an exact date that we worked on a certain task.

Oct 16	Started adding datasets
Oct 22	Assigned questions
Oct 27	Upload student gov data
Oct 31	Started development for questions
Nov 9	Finished first round of questions
Nov 10	Focus on analysis and graphs
Nov 20	Assigned second round of questions
Nov 22	Finished second round of questions
Nov 26	Focus on analysis and graphs
Nov 26	Start presentation slides and report

1) First, we had to find appropriate datasets to answer our questions. We allocated a solid week to find a good first iteration of the datasets. We thought that it would be good to go ahead and start coding, and if we needed more data, we could find more datasets for our questions.
2) Second, we began coding. During this time we made the functions that analyzed the data, printed the results, and made graphs. The functions were created to be generic, that way the same function could be used to analyze different features of the dataset. Each team member was be assigned certain analytics to work on, and was be expected to finish those functions in a timely fashion.
3) Third, we began to chart the data and make our results presentable. We used matplotlib and seaborn for our charts.