Student Loan Project Proposal

Kemal Fidan, Logan Courtney, and Chandler Lampe

***Abstract*— This document breaks down our project to analyze 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, 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.

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.

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.

III. DATA

We will be getting 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 will also be using 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 religion, ACT percentiles, enrollment numbers, and student ethnicity are included in this data set.

From these data sets we will be comparing some of the following:

- Average loan per recipient vs. state of college
- Average amount of loans vs. type of colleges (public or private)
- Average loan vs. ethnicity
- Average loan vs. ethnicity vs. state of college
- Cost of school vs the amount of loans a school has
- Demographics vs the amount of loans a school has
- Number enrolled vs the amount of loans a school has
- State the school is in vs the amount of loans a school has. Loan per person here?
- Type of institution (public, private, other) vs amount of loans
- ACT score vs the amount of loans a school has. Expect the trend to be that higher ACT schools have more funding thus students owe less dept.
- Age of loan recipients over the past few years to see if there's a trend like less older people
- Average amount of loans per recipient

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. A

IV. RESPONSIBILITIES

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

In addition to that, we'll have officers to facilitate timely delivery of the project. We have set some roles below:

- Team Leader: Logan Courtney
- Project Manager: Kemal Fidan
- Deadline Manager: Chandler Lampe

With these responsibilities, we can make sure that members are working on their tasks, that we meet deadlines, and finish our project in time.

V. TIMELINE

1) First, we will have to process the data. This includes reading and parsing multiple files and removing erroneous or nonessential data. This part of the project might take a larger percentage of the time due to the amount of files and rows in our data set. This step will also include the combination of different data sets into a single master data set.
2) Second, we will begin the analysis. During this time we will be making the functions that will analyze the data and print the results. The functions we create will be generic, that way the same function can be used to analyze different features of the data set. Each team member will be assigned certain analytics to work on, and will be expected to finish those functions in a timely fashion. This part shouldn't take long because we will be using math libraries like NumPy to quickly do the math, assuming data pre-processing goes smoothly.
3) Third, we will begin to chart the data and make our results presentable. We will probably be using matplotlib for our charts. We expect this be easy but we know matplotlib can be difficult at times.
4) After we do the finishing touches, we would also like to implement a machine learning algorithm to predict student loan data given some information about the borrower. However, this is a stretch goal and isn't well defined exactly. This machine learning application is meant to be applied in addition to our findings. Because of this, we might formulate the question after our findings.

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. OUTCOME

At the end of this project, we would like to have neat data analysis. We plan on either using a Python notebook or moving our graphs and charts to PowerPoint or something similar. This way, the data can be seen in a user friendly way, and can be easy to follow. Additionally, we would like to address anything interesting that came up while analyzing the data. For example, if some states, schools, or ethnicities have a strange student loan debt amount, we would like to present why we think that is.