BUDT 758X Project Proposal: Exploring data analytics in the Student debt crisis in the US.

Group 29 (Tachy Terps 🦕):
Prachi Singh, Madhura Dighe, Chinmay Gupta

Introduction:

The US Student loan debts have reached a whopping $1.6 Trillion and continue growing exponentially as more and more students apply for Federal aids and grants. The student loans have crossed the total amounts of auto loans and credit card loans! The student debt crisis has caught the attention of a lot of top bureaucrats including President Trump, Elizabeth Warren, and Bernie Sanders. This resulted in Elizabeth Warren introduced a bill titled the "Student Loan Debt Relief Act of 2019." The bill stated a plan that would cancel student loan debt for more than 95% of borrowers, and would entirely cancel student loan debt for more than 75% of Americans with student loan debt.

The population of interest is undergraduate and graduate students of US educational institutions who avail education loans. As the sampling is random and has enough observations the results of this study could be generalized for the population of interest.

Hasan Minaj (a popular standup comedian) recently spoke about the crisis on his show The Patriot Act. You should definitely watch it to grasp the depth of the problem: Click Here!

Questions of Interest:

- What is the student debt amount per year in the United States? Can we see a considerable growth?
- Which educational institution/city/state has the highest average student debt? Do students from for-profit institutions have a similar loan repayment rate across different states?
- Is there a relationship between the SAT score and the loan repayment rate for students?
- Can we identify an evident relationship between the loan repayment rate across students from different race (i.e. Caucasians, Hispanics, African-Americans, and Asians)? Does this relationship provide any inference about the economic disparity that exists in the society between different races?
- Does the institution a student graduate from effect the loan repayment rate because of the income earned after graduating?
- What are the changes observed on the loan repayment rate after earning for 10 years?

- Does a pay gap exist amongst genders graduating from the same institution? Are similar trends observed across different states/institutions?
- Do students at private for-profit institutions have a lower rate of federal loan repayment compared to students at non-profit institutions?
- What will the loan repayment rate look like in the coming years? Are new reforms in the education system going to affect the loan repayment rate in the future? Can the loan repayment rates for-profit and non-profit students match in the next 10 years time?

Data Processing and Analysis

Dataset Description:

A large portion of the data is collected from The College Scorecard website while other crucial information is gathered from the federal financial aid and earnings website (US Department of Education) and data.gov. This data provides insights into the performance of schools eligible to receive federal financial aid. There are 7804 records available in total but as the data has missing, invalid or null values, it requires to be cleansed and processed to be streamlined to necessary information. The files include data from the year 1996 through 2017 for all degree-granting institutions of higher education but we will be using data from 2007 to 2017 to form inferences from 10 years of data.

The important columns for the analysis that we will primarily be working with are listed below: INSTNM, CITY, STABBR, ZIP, LATITUDE, LONGITUDE, MENONLY, WOMENONLY, ADM_RATE, SAT_AVG, UGDS_WHITE, UGDS_BLACK, UGDS_HISP, UGDS_ASIAN, TUITIONFEE_IN, TUITIONFEE_OUT, TUITIONFEE_PROG, TUITFTE, INEXPFTE

Source: https://collegescorecard.ed.gov/data/

Data Processing Tasks:

- Merging datasets
- Data Cleaning
- Managing and changing missing data with a dummy and average values respectively
- Sorting, filtering data records
- Data visualization
- Data interpretation
- Regression analysis

Data Analysis:

The questions we will answer are- What kinds of visualizations and modeling do we need to do to answer our questions of interest? Maps? Charts? Will your visuals be static or interactive?
The answers are that we will be creating:

1. A static heat map of the United States to showcase the average student debt per state and draw related inferences from it.
2. A time series graph to display the rise in student debts over the years and gather relevant information from the same to answer the questions stated earlier.
3. Finding correlations between factors.
4. Histograms for comparisons.
5. A linear model to understand regression and possibly predict if the loan repayment rates for-profit and nonprofit students match in the next 10 years time due to new reforms in the education industry.

Expected Findings

We are expecting to understand why the student debt crisis has become a hot topic right now, especially with the upcoming elections. We are expecting to draw information from the correlation between different factors like repayment rates, loan amounts based on state or city or institution etc. We expect to infer if any racial disparities, gender bias, or any other factors exist that can provide a granular look into the loan repayment problem that exists in the country.

Project Timeline

Task	**Task Lead**	**Due Date**
Data Cleaning, Merging and Extraction	Prachi	10/31
Handling NULL values and missing values	Chinmay	11/05
Data Formatting	Madhura	11/12
Data Visualization	Prachi	12/01
Data Interpretation	Madhura	12/03