**UNIVERSITY OF NORTH TEXAS**

**ADTA 5130**

|  |  |
| --- | --- |
| B.E. Computer Science Engineering with Specialisation in Data Science &amp;  Analytics  DATA ANALYTICS I  -GROUP PROJECT - Hypothesis test of average student debt in the United States | Group 21  Team members  -Christopher Villagran  -Luis G Cardenas  -Oscar Vazquez  **-Sanjay R Manda** |

Objectives

In recent years, American student debt has increased dramatically, reaching record highs of more than 1.7 trillion dollars, although this practice can represent an opportunity to obtain university studies and in turn a higher paying job, it can also consider a way to get into debt, a debt that can be taken to pay off the years of greater productivity of the citizen; The problem lies not in the value of having participated in the higher education system, but the problem lies in its debt-laden, burdensome cost.

Due to the fact that many times minorities or low-income people do not have experience with the higher education system, they can become victims of predatory practices of institutions that make individuals take loans for tens of thousands of dollars to cover tuition.

More recently, an article published in the Wall Street Journal, November 9, 2021, by Bannon and Fuller1, details how the University of Southern California used its reputation as an elite school to push a $115,000 online master’s degree in social work, again for unsuspecting individuals.

To inform the public, the following analysis examines the U. S. Department of Education’s College Scorecard dataset for the 2018 and 2019 fiscal years. Insights from an analysis of this dataset may be used to mitigate long-term and burdensome student loan debt

Which brings us to the next questions:

Student debt differs between first-generation and non-first-generation students?

Does the income of students differ between the first and second years of work after graduation?

What type of school issues more debt? (Private, non-profit, for-profit, or public)

What type of major and degree has the strongest relationship to debt/earnings?

Is the average debt directly related to the zip code?

Dataset information

The U. S. Department of Education’s College Scorecard dataset for the 2018 and 2019 fiscal years. The dataset includes records for all U.S. universities, majors and level of degree offered, the mean and median student loan debt upon graduation, and the mean and median earnings for the 1st and 2nd years of work post-graduation. This consists of 6296 different institutions, out of which 3649 reported the number of students that had received Stafford and Grad PLUS student loans, the median student loan debt, and generational information (first or not first generation).

The institutions reporting include. Additional information for these schools includes geographical information, and accreditation information.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Generation | Opeid6 | Debt N | Med\_debt | Sch Deg | Accredagency |
| Denotes is a student is first generation or not | The 6-digit OPE ID for institution | Number of students reporting debt | Median student debt reported by the given group. Median Stafford and Grad PLUS loan debt disbursed at this institution | School degree level:  0-non-degree-granting  1 certificate degree  2-Associate degree  3-Bachelor's degree  4-Graduate degrees | Accreditation for the postsecondary education participant system institution |

Data cleansing

Table

Description automatically generated with low confidence

The initial dataset was downloaded from the U. S. Department of Education’s database. At the point of ingestion, it contained several duplicate rows for each institution. There was a duplicate row for each breach. For example, the University of Phoenix reported 37 branches.

For each unique institution we kept only one row of data. This helped us avoid potential issues with our averages since we did not want to double count any institutions.

Institutions that did not report number of students in debt, the median student dent and generational information were removed from this dataset.

Hypothesis testing

We believe that the amount of debt accrued as well as the number of students going into debt will vary by reporting generation.

This can be broken down into two problems, and the corresponding hypothesis are as follow.

HA = First Generation median student debt ≠ Not-first generation median student debt.

H0= First generation median student debt = Not-first generation median student debt.

Then let’s take a look at number of students reporting debt

HA = First Generation number of students in debt ≠ Not-first generation number of students in debt.

H0= First Generation number of students in debt = Not-first generation number of students in debt.

Methodology

Since our dataset is comprehensive for the institutions reporting in the US. We saw it fit to take a random sample of these data and analyze it. This was to prevent the risk of replacement where one student can report debt in two institutions, therefore distorting the results. This resulted in a random sample consisting of 1249 different institutions.

Once this random sample was gathered, we decided to complete some summary statistics to get more familiar with the data. This will help us decide which method will best suit our efforts to answer our questions.

First we analyzed the median student debt by institution.

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

The distribution of the student debt by generation and the corresponding boxplots shows a close to normal distribution. To test our hypothesis, we can use a two-sample t-test to see if there is a statistical difference in these sample.

Table

Description automatically generated Chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

The t-test results show very similar variances in the student median debt distributions, and near normal Q-Q plots. However, they also show a disappointing result. The p-values for both Pooled and Satterthwaite are higher than our preselect value of 0.05 this means that there is high probability that another random sample would yield similar results. We cannot reject the null hypothesis. **We do not have enough evidence to prove that the median student debt reported by first generation students is different than that reported by not-first generation students.**

While disappointing our results are interesting none the less. It does make sense that students who take loans for college will take loans to cover their studies which should be the same regardless of what generation you are while attending school.

Is there a difference by generation in the number of students undertaking debt? We will explore that question below.

Chart, histogram

Description automatically generated Box and whisker chart

Description automatically generated

The histogram and box plots for the student distribution by generation shows a right skewed distribution with several outliers. To verify, we generated a Q-Q plot as well. Here we can see a separation for the normal line at the upper end of the scale.

Chart, line chart

Description automatically generated

To eliminate those outliers and attempt to normalize the data we took natural log of number of students in debt. Then we ran a t-test based on that new transformed column.

Chart, histogram

Description automatically generated Table

Description automatically generated

Chart, line chart

Description automatically generated

Taking the natural log of the number of students in debt helped normalize the data to make it viable to apply a t-test to compare the two groups. First, we look at the distribution of the log\_of\_debtN, it appears normal. The variance also appears similar. Finally, in this case the p-value for this comparison is less than our threshold of 0.05 which gives us enough evidence to reject the null hypothesis. The means of the number of students in debt is statistically different.

Table

Description automatically generatedDistribution analysis between average debt and income upon one year of graduation.

Chart, histogram

Description automatically generated

p-Value <0.005

Confidence limits @ 95% 13335- 13771

Table

Description automatically generatedDistribution analysis: UNIVARIATE PROCEDURE

Graphical user interface

Description automatically generated

Graphical user interface

Description automatically generatedp-Value <0.005

Distribution analysis of median debt and reported earnings upon one year of graduation

Table

Description automatically generatedp-Value Shapiro-Wilk test= <0.0001

Chart, histogram

Description automatically generated

Distribution analysis of median debt and reported earnings upon one year of graduation

Graphical user interface

Description automatically generatedTable

Description automatically generatedP-Value Cramer test= <0.005

Graphical user interface

Description automatically generated

Boxplots median debt and reported earnings upon one year of graduation

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Table

Description automatically generatedCorrelation analysis

A picture containing graphical user interface

Description automatically generated

Correlation Analysis

Median Debt/ Credlev

Graphical user interface, chart

Description automatically generated

Median debt/ Zip code

Scatter chart

Description automatically generated

Correlation Analysis

Earnings upon 1st year of grad/ Credlev

Chart

Description automatically generated

Earnings upon 1st year of grad/ Zip code

Chart

Description automatically generated with medium confidence

Earnings upon 1st year of grad/ Median debt

Graphical user interface, diagram, engineering drawing, scatter chart

Description automatically generated

Conclusions

Student debt differs between first-generation and non-first-generation students?

We fail to reject the null hypotheses: **median student debt reported by first generation students is different than that reported by not-first generation students.**

Does the income of students differ between the first and second years of work after graduation?

What type of school issues more debt? (Private, non-profit, for-profit, or public)

What type of major and degree has the strongest relationship to debt/earnings?

Is the average debt directly related to the zip code?