Does the pay post graduation outweigh the amount of loans generated?

Felipe Rodriguez

2023-03-03

Introduction

In the fall of 2020, there were 15.85 million undergrad students registered for the fall semester (Hanson (2023)). To many, a bachelor's degree is considered to be an initial milestone to begin a long term career and create financial stability. A typical four year degree might not seem like the right path to some, but a big factor influencing this decision is the cost of attendance and the necessity to take out student loans. In this study, the pay of an undergraduate student, post graduation, will be analyzed and compared to the amount of average debt acquired by student to understand the correlations between the two.

A few research question that are worth investigating:

Which institutions have the highest paid undergraduates?

Out of those institutions, what is the average student debt?

Which institutions have the highest debt?

Does highest debt and highest pay have a correlation?

Is there a correlation between student debt and diversity in the institution?

Approach

The approach taken will be to identify data that contains salaries after graduation from a four year university. There will also be a need to have information regarding student loans, the pay each major is most likely to have, and demographic information. Using this data, we can join income and debt per institution to analyze the amounts. With the demographic information, an analysis can be done to see if there is a correlation in amount of minorities and total debt.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows).

The main sources for my data will be as follows:

- Brown, M. (2022, January 19). Student Loan Debt by School by State Report. LendEDU. Retrieved February 9, 2023, from https://lendedu.com/blog/student-loan-debt-by-school-by-state/

- Devastator, T. (2022, November 23). The schools that create the most student debt. Kaggle. Retrieved February 9, 2023, from https://www.kaggle.com/datasets/thedevastator/the-schools-that-create-the-most-student-debt
- Mostipak, J. (2020, March 9). College tuition, diversity, and pay. Kaggle. Retrieved February 10, 2023, from https://www.kaggle.com/datasets/jessemostipak/college-tuition-diversity-and-pay
- Payscale. (2021). Best universities and Colleges. The Best Universities For a Bachelor's Degree. Retrieved February 9, 2023, from https://www.payscale.com/college-salary-report/bachelors?search=nebraska

The source LendEDU contains loan information on student loans throughout the United States. This data set contains student loan information for states, private schools and public schools. The main columns that make this data a notable source will be Institution, State, and Average Student Loan debt per borrower. The data set provided by Payscale contains average salaries post graduation. The columns in this data set that will be used are School Name and Early Career pay. The last two data sets are in Kaggle. One contains data on schools that create the most debt and the other contains demographic information of school. These can be used to compare the percentage of minorities in relation to the amount of debt the school has to understand if there is a correlation of schools where there are more minorities and debt.

Required Packages

The anticipated packages that will be needed will be as follows:

- readxl
- dplyr
- purrr
- lm.beta
- ggplot2

The packages listed above will aid in plotting, visualizing, and creating calculations. These packages are anticipated are subject to change depending on the need of the study and the results uncovered. Any changes will be documented.

Plots and Table Needs

The plots that are anticipated are a bar graph and a scatter plot. Using a bar graph, the total amount of races can be visualized to see demographic information. With a scatter plot, an analysis can be done to view the amount of average debt per student and average early career salary to understand if there is a relationship between the two.

A table will need to be created containing all the data joined on University Name or Institution. There will need to be the following columns in the joined data set:

- School
- Average Student Debt
- Early Career Pay
- Total Students
- Minority Students

The stated plots and columns are anticipated in the study, but more can be created or added as necessary to the analysis being conducted.

Future steps

The next step will be to merge the data into one data set. One item that will need to be studied is renaming columns to have one consistent column to join on. The data sets have the University name as the following columns: Institution, Name, and School Name. These will need to be changed to University Name and then joined on the new name created. Another step that will be needed is to extract total students and total minority from the demographic data set. Total minority is a row field as a opposed to a column, this will need to be transformed.

How to import and clean my data

The data imported will come from four sources that have been stated above. All of the sources have been inspected and the column representing the name of the institution or university has been renamed to 'University' for all data sources. Once read in, the data can be analyzed to see if any changes need to be made.

```
## # A tibble: 6 x 6
##   University                          School~1 Early~2 Mid-C~3 Perce~4 Perce~5
##   <chr>                               <chr>      <dbl>   <dbl> <chr>     <dbl>
## 1 Massachusetts Institute of Technology Enginee~   93700  167200 0.51       0.68
## 2 Harvey Mudd College                 Enginee~   97700  166600 0.5600~    0.75
## 3 Princeton University                Ivy Lea~   81800  161500 0.48       0.49
## 4 United States Naval Academy         Enginee~   83700  160100 0.61       0.57
## 5 Stanford University                 Enginee~   87100  156500 0.5500~    0.5
## 6 Harvard University                  Ivy Lea~   80900  156200 0.49       0.2
## # ... with abbreviated variable names 1: 'School Type', 2: 'Early Career Pay',
## #   3: 'Mid-Career Pay', 4: 'Percent High Meaning', 5: 'Percent STEM Degrees'
```

At a glance, no changes are needed to the initial data set. Joining can begin with the next data sets on the column "University." The next data sets can be read in and merge to the initial data set.

```
setwd("/Users/feliperodriguez/Library/CloudStorage/OneDrive-BellevueUniversity/DSC 520 Statistics/Week
devastator_data <- read.csv("Devastator_Student Loan Debt by School 2020-2021.csv")
payscale_devastator <- merge(x=payscale_data, y=devastator_data, by="University", all.x=TRUE)
lendedu_data <- read.csv("lednedu_school_loan_per_borrower.csv")
payscale_devastator_lendedu <- merge(x=payscale_devastator, y=lendedu_data, by="University", all.x=TRUE
diversity <- read.csv("archive/diversity_school.csv")
data <- merge(x=payscale_devastator_lendedu, y=diversity, by="University", all.x=TRUE)
```

What does the final data set look like?

The final data set set contains all of the data merged into one location. There were some fields that did not exist for certain Universities. Omitting any NA or null values helped clean up the data. The data contains multiple records due to there being type of schools, ethnicity, etc. The final data set gives an overview of the information provided by all the sources. Below are the columns of the final data set.

```
data2 <- na.omit(data)
final_data <- data2[!duplicated(data2), ]
data.frame(`ColumnNames` = colnames(final_data))
```

```
##                                ColumnNames
## 1                               University
## 2                              School Type
## 3                         Early Career Pay
## 4                           Mid-Career Pay
## 5                       Percent High Meaning
## 6                       Percent STEM Degrees
## 7                                    index
## 8                                   OPE.ID
## 9                                     City
## 10                                   State
## 11                                Zip.Code
## 12                               School.Type
## 13                                Loan.Type
## 14                               Recipients
## 15             Number.of.Loans.Originated
## 16             Amount.of.Loans.Originated
## 17                Number.of.Disbursements
## 18                Amount.of.Disbursements
## 19 Average.Student.Loan.Debt.Per.Borrower
## 20                         total_enrollment
## 21                                 category
## 22                               enrollment
```

What information is not self-evident?

In the final data set, there are many variables that give various data of the universities that may be used, but for the purpose of this study, will not be needed. The information that is still needed to be analyzed is Early Career Pay versus Average Student Debt per Borrower. A calculation done to understand the difference in pay versus loan can be added. Another item that is not self-evident, is the correlation of minorities to the amount of student loans per borrower.

What are different ways you could look at this data?

The data can be looked at in different subsets with information related to the study. The initial join was to get all the data in one consolidated output. The data can be viewed by University or sorted by highest early career income or debt. With that, different subsets can be made. Two subsets that will be created are:

- University, Early Career Pay, and Average Student Debt per Borrower
- University, Average Student Debt per borrower, and Total Minority

These two subsets will answer some of the initial questions posed at the beginning of the study.

How do you plan to slice and dice the data?

As mentioned prior, the data can be divided into two subsets. The first subset that will be created will include three columns University, Early Career Pay, and Average Student Debt per Borrower, this table will be called income_debt. This subset of data will need to be removed of any duplicate values to keep a clean output and only see one record per university that have recorded data.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
income_debt <- select(final_data, University, `Early Career Pay`, Average.Student.Loan.Debt.Per.Borrowe
income_debt <- unique(income_debt)
```

Once adjusted, a small sample of data can be viewed.

```
head(income_debt, n=3L)
```

```
##              University Early Career Pay Average.Student.Loan.Debt.Per.Borrower
## 717   Auburn University            59100                                  29331
## 967      Babson College            77800                                  37866
## 2067 Bennington College            50200                                  29443
```

The next subset that will be made will only contain the following columns University, Average Student Debt per borrower, Total Minority and enrollment. The method used for this subset will include a filter when the column Category has the value "Total Minority." This will give the total minority for that school. Once those records are selected, a subset can be created to only include the columns we need.

```
minority_income_debt <- final_data[final_data$category== 'Total Minority',]
minority_income_debt <-select(minority_income_debt, University, `Early Career Pay`, Average.Student.Loa
minority_income_debt <- unique(minority_income_debt)
head(minority_income_debt, n=3L)
```

```
##              University Early Career Pay Average.Student.Loan.Debt.Per.Borrower
## 717   Auburn University            59100                                  29331
## 977      Babson College            77800                                  37866
## 2068 Bennington College            50200                                  29443
##            category enrollment
## 717  Total Minority       3269
## 977  Total Minority        701
## 2068 Total Minority        110
```

How could you summarize your data to answer key questions?

With the subsets now created, they can be used to answer the following questions:

- Which institutions have the highest paid undergraduates?
- Out of those institutions, what is the average student debt?
- Which institutions have the highest debt?

Using the income_debt subset of data, sorting the data by Early Career Pay descending can give the universities with highest paid salaries and show the student debt in those schools. The top five universities with the highest Early Career Pay are showed below.

```
ordered_income_debt <- income_debt[order(-income_debt$`Early Career Pay`),]
head(ordered_income_debt, n=5L)
```

```
##                University Early Career Pay
## 27285 Stanford University            87100
## 11239  Harvard University            80900
## 967        Babson College            77800
## 2177   Bentley University            72500
## 22052      Pomona College            70200
##       Average.Student.Loan.Debt.Per.Borrower
## 27285                                  22897
## 11239                                   6170
## 967                                    37866
## 2177                                   35187
## 22052                                  18829
```

The same practice can be done on the column Average.Student.Loan.Debt.Per.Borrower to display the schools with the highest debt per borrower. The top five universities with the highest average debt per borrower are showed below.

```
ordered_income_debt2 <- income_debt[order(-income_debt$Average.Student.Loan.Debt.Per.Borrower
),]
head(ordered_income_debt2[ , c(1, 3)], n=5L)
```

```
##                       University Average.Student.Loan.Debt.Per.Borrower
## 16074 Loyola University Maryland                                  41443
## 14938           Le Moyne College                                  40522
## 28308          Temple University                                  38634
## 967               Babson College                                  37866
## 19470      North Central College                                  37396
```

There are two remaining questions:

- Does highest debt and highest pay have a correlation?
- Is there a correlation between student debt and diversity in the institution?

Initially, to begin to dig into the question about race and debt, the schools with highest minorities will need to be analyzed. By ordering the minority_income_debt by the amount of total minorities, an initial analysis can be done. The top five schools with the most minority are displayed below.

```
ordered_minority_debt_income <- minority_income_debt[order(-minority_income_debt$enrollment),]
head(ordered_minority_debt_income[, c(1, 3:5)], n=5L)
```

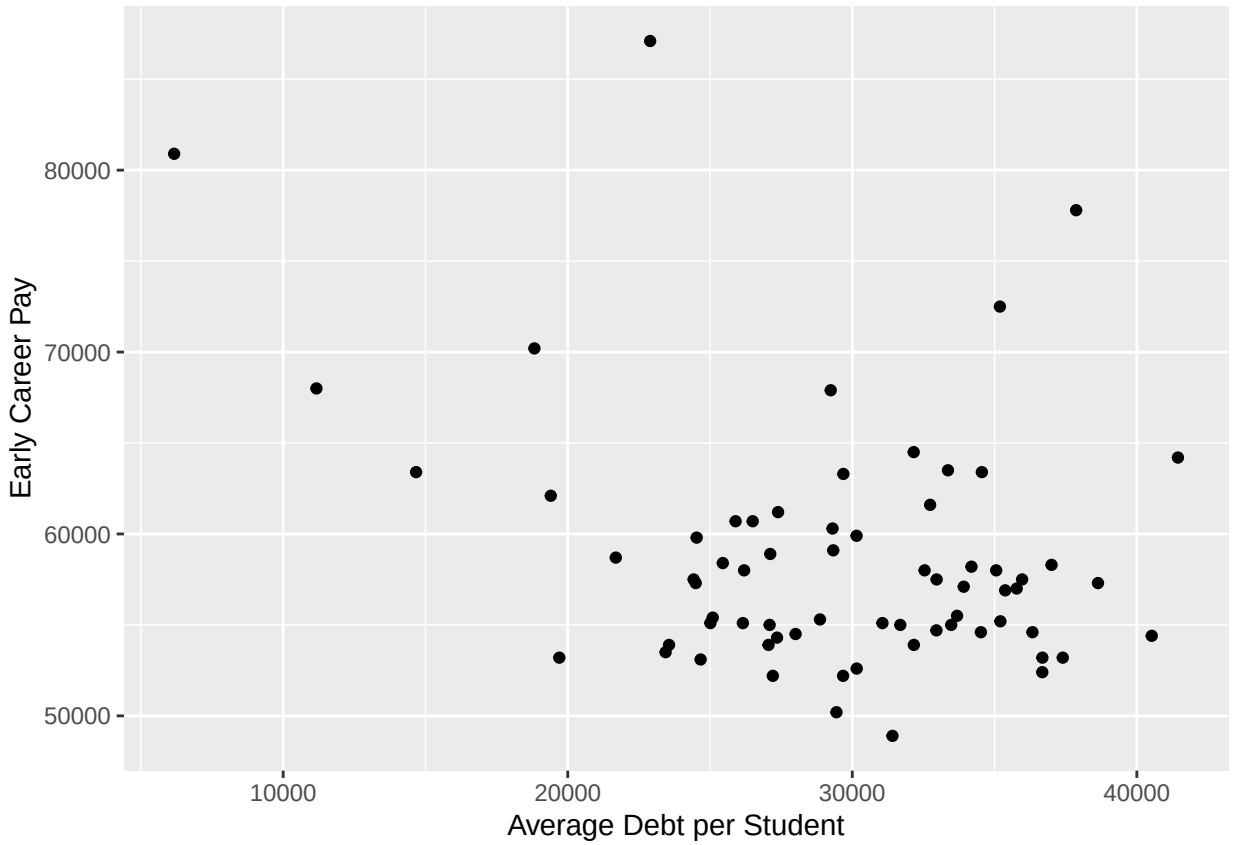
```
##                             University Average.Student.Loan.Debt.Per.Borrower
## 8954  Florida International University                                   19705
## 9886          Georgia State University                                   28864
## 19140              New York University                                   29242
## 8834       Florida Atlantic University                                   23439
## 9541           George Mason University                                   33362
##             category enrollment
## 8954  Total Minority      39763
## 9886  Total Minority      19336
## 19140 Total Minority      15342
## 8834  Total Minority      15268
## 9541  Total Minority      12977
```

To further understand the correlation between race and debt, additional analysis is required and will be explored in the sections to follow.

What types of plots and tables will help you to illustrate the findings to your questions?

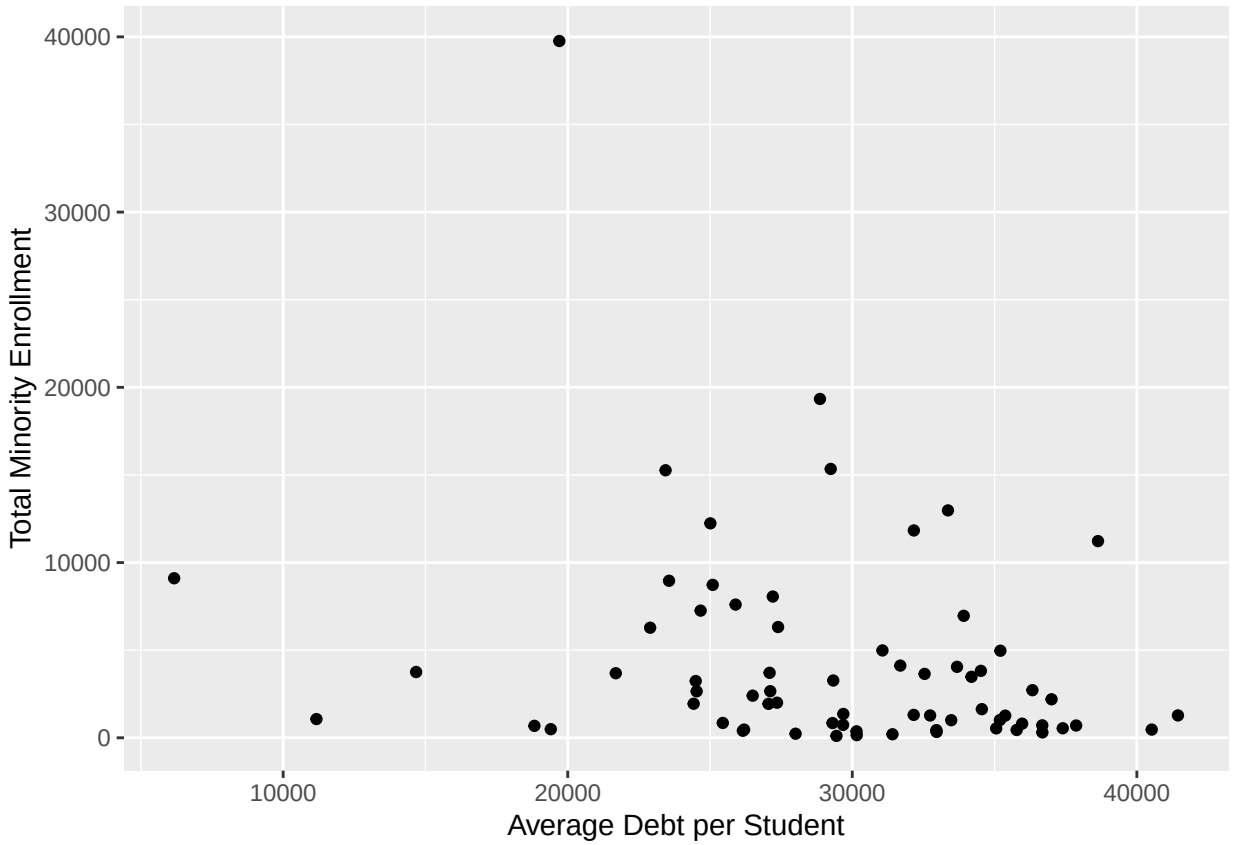
The plots that can give a good visualization of the data will be scatter plots. These plots will give a quick view of the data at question. The first plot created will show Early Career Pay versus Average Student debt income.

```
library(ggplot2)
ggplot(income_debt, aes(income_debt$Average.Student.Loan.Debt.Per.Borrower, income_debt$`Early Career Pa
```



The next plot displayed will show Minorities versus Average Student Debt Income.

```
ggplot(income_debt, aes(income_debt$Average.Student.Loan.Debt.Per.Borrower, minority_income_debt$enrolln
```



Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

The machine learning technique that can be applied to this study to answer the research questions posed is a linear model. Using a linear model or linear regression, a predictive analysis can be done understand the roll of tuition and diversity in student debt. These analysis can be conducted together, to identify which plays a bigger role, and plotted separately, to display the results individually.

Questions for future steps.

The next steps should involve building a regression model to display correlations of the two different variables. With that, research will need to be done to understand how to create the linear regression model accurately to determine the predictor and outcome. More importantly, a method will need to be established to plot the predictions versus the original data.

Analysis.

With the data, we see some trends at a glance on how the student loans and salary interact. An interesting insight is the top 5 universities with the highest paid graduates are not the same as the top 5 with the highest debt, which can be indication that there is not a correlation between the two. When looking at diversity within students and early career pay, there also does not seem to be a correlation between the two. These initial findings can give a "sneak peak" to the data.

Implications.

Some of the implications with this study revolved around the data gathered. When looking at the amount of universities captured, the total amount of universities included is a small quantity which could provide bias in the data. With a wider range of universities, there could be more trends that cannot be seen. The source data is also taken from a wide range of years, 2014-2018. More recent data was not available. With newer data and more consistent data, the quality of this analysis could also be improved.

Limitations.

As mentioned, this study was limited to the data available. This project can be improved by adding or finding additional sources that provide more data for universities, or even selecting the universities with the highest student count to lead to a wider range of data. Additionally, a model should be built in order to understand further understand the correlations between the fields. Using the model, predictions can be plotted to understand the trends of income, debt and diversity.

Concluding Remarks

This study conducted has began to uncover the trends between student loans in universities and salaries post graduation. Initially, the findings show that the the amount of loans generated by a university and the amount of pay post graduation are not dependent on one another or correlated. However, as mentioned prior, a regression model to predict these two variables would give further insight. Afterwards, the model can be tested for accuracy. Also, a wider range of data will can help with trend analysis. With these two recommendations, the analysis of student loans and pay post graduation can continue and be further explored.

References

Hanson, &amp, M. 2023. *College Enrollment Statistics [2023]: Total + by Demographic*. Education Data Initiative. https://educationdata.org/college-enrollment-statistics .