Spring 2022

An Analysis of Long-Term Financial Feasibility of Higher Education

Jacqueline Boyd Lerman
Bard College

An Analysis of Long-Term Financial Feasibility of Higher Education

A Senior Project submitted to
The Division of Science, Mathematics, and Computing
Of
Bard College

By
Jacqueline Boyd Lerman

Annandale-On-Hudson, NY
May, 2022

Acknowledgements

I would like to thank professor Kerri-Ann Norton for all of her support and guidance throughout this project. I would also like to thank all of the computer science faculty for sparking my passion in this field and the never-ending support and encouragement they have provided. This project, and all I have achieved and learned here at Bard, would not have been possible without them.

Table of Contents

Abstract

This project explores the long-term financial feasibility of higher education. With rising costs of higher education and so many choices surrounding a degree such as degree type, sector of institution one attends, student loans one takes out, and field of study, it can be hard to discover which path will be most profitable long-term. This project analyzes data from the National Center of Education Statistics to see if there are existing relationships between these variables that contribute to different experiences in higher education and financial outcomes, specifically relating to future income and student loan payments. To do this I use various statistical tools and models such as multiple linear regression, tests for correlation, Kmeans clustering, and ANOVA testing. While most of these tests showed little or no relationship or significance, through clustering I found that those who get both an associate's and a bachelor's in the same field, make on average significantly more than those who get an associate's and bachelor's degrees in different fields. I also start the creation of a summary statistics interface with the intent to display data in a way that those with minimal scientific background could understand in the hopes that this project will continue to spark conversations around the inaccessibility of data surrounding higher education and the realistic outcomes that different paths through higher education will provide.

1
Introduction

1.1 Overview

Higher Education is often seen as a means of advancing one's career into a more advanced and hopefully more profitable path. With so many options available to those searching for a higher education, it can be difficult to determine what opportunity will be the most financially viable in the long term. There are many factors that contribute to what type of institution or program a person may choose including age, socio-economic standing, family history, and gender, among others. Full-time work and additional home life responsibilities of those who may be returning to education later in life or have no choice but to work through school only adds to the difficulty of searching out and completing a form of higher education in hopes of advancing into a more profitable career path. The obstacles faced when making the choice of what form of higher education to pursue when it comes to type of degree, field of study, and sector of institution (i.e. public, private non-profit, private-for-profit, etc.) will only continue to grow far beyond these initial road blocks. I will endeavor to analyze the options available to people in varying socio-economic situations and challenges and see what forms of higher education are the most financially feasible and worthwhile in the initial years following a

degree in addition to highlighting the flaws in our current system of higher education in America. To begin I would like to look at many of the initial obstacles faced when starting a pursuit for higher education.

1.2 The Education Market

One obstacle is the intimidating expanse of different forms of higher education that one could pursue, along with the social stigmas that they may have attached. Between bachelor's degrees, associate degrees, community college, private universities, for-profit institutions, evening classes, part-time programs, specific certifications courses, and more, it's not as simple as it once appeared as young teenagers pushed into the American classic that is the four-year bachelor's degree at a reputable institution. In an ideal world, one could dedicate a fair amount of time to research the specific career path you want to pursue, and find local programs and support available to you. If you were to ask anyone to start the research process I can assure you they would all start at the same place: Google.

While I am not here to extensively argue the deeper flaws of everyone's favorite search engine, I would like to point out some upsetting realities that are present to those seeking higher education looking for an affordable way to advance career paths. Much like everything in America, higher education is a highly profitable market, and as such there is intense marketing done by for-profit institutions fighting to be at the top of that google search page. There are an

estimated 7,550 for profit institutions in the united states, with a strong emphasis on the word estimated as until the late 1990's, little effort had been made to fully track down for profit institutions in the United States (Cellini, 2012). These for-profit institutions are loud and can appear at a glance much more appealing than other local programs or certifications courses, but often offer less financial aid and higher costs of tuition to students. While the average cost of a two-year associate's degree at a for profit college is $35,000, the same associate's degree at a comparable community college will only cost an average of $8,300. This is also reflected in the median debt that a student at a for-profit college graduates with which sits at $32,700 while the average debt of the students graduating from a private non-profit college is $24,600. Loans drop even lower for those graduating from a public college or university with a median debt of $20,000 (Suevon, 2012).Those who don't have the support, knowledge, or time to understand the reality of these for-profit schools are much more likely to fall victim to their lure of low initial payments, night classes and part-time offerings. We see this reflected in the statistics of the types of students that attend these schools as Stephanie Riegg Cellini discusses in "For-Profit Higher Education: An Assessment of Costs and Benefits" (2012),

> For profit students have less parental involvement in their education, higher levels of high school absenteeism, and are more likely to be young parents than students in other sectors. Deming, Goldin, and Katz(forthcoming) corroborate the patterns found previously, exploiting new data on first-time college freshmen from the 2004/09 Beginning Postsecondary Students Longitudinal Study. For profit students in this sample are more likely to be female, black, and/or Hispanic relative to students in other sectors. Compared with those in community colleges, for -profit students are disproportionately single parents, have much lower family incomes, and they are almost twice as likely to have a general education diploma (GED), rather than a high school diploma

We start to see an exploitation of vulnerable populations as the above quote shows us a correlation of those who don't have family support, higher incomes, or a high school diploma falling into the financial trap of the for-profit institution. With 96 percent of those enrolled in for-profit schools taking out loans, compared to the 57 percent at four-year private non-profit colleges, most of these students will leave their higher education experience with an overwhelming amount of debt (The For-Profit Higher Education Industry, By the Numbers — ProPublica, n.d.). Of students who took on loans in 2009, 26 percent had to default on at least one payment over the five years with especially high default rates for students from for-profit institutions as well as those from lower income households (Cebula & Koch, 2021).

1.3 The Value of a Degree

From an early age there is an intense amount of pressure placed on student's exams such as the SAT and ACT in addition to maintaining a high GPA, community service hours, and extra-circulars, all with the hope of getting into your dream school and maybe, if you're lucky, with enough financial aid that you can actually attend. What is not discussed as often are the other options available. While community colleges offer degrees at a much lower cost to the student, there is a stigma attached along with them. There is a deeply flawed illusion in this country that the more you pay for something, the better quality it must be. We apply this logic to material goods on a daily basis, but people often forget that sometimes you are paying for a brand, an image, a marketing department, not higher quality goods. While there is notably a difference in the community college experience (often due to community colleges at times unable to pay professors as competitive wages as private institutions), how much of a difference

is there in the degree that you hold as you graduate your program? This greatly depends on the field that you want to go into as more and more success can be found in technical career paths without requiring a degree from a private institution or even a full 4-year bachelor's degree, however, there are certain paths, often including research and academia that will require more extensive higher education to reach a financially successful point in your career. So how do you know what degree is going to make you the most successful in your field while still being financially feasible? That is the question that this senior project is trying to explore through data analysis in a way that is approachable to those outside the world of STEM.

1.4 Inaccessibility of Research

The unfortunate answer to this question requires research. Since the ideal institution and degree varies so much from field to field, there is no one simple guide out there to help make this difficult choice. This brings us to the obstacles of time, knowledge, and resources. If someone is already working full time or taking care of kids or in a complicated living situation, it might not be possible to have hours of time at a computer with WIFI to do this research. A google search or relying on someone else's knowledge may be all they have. This lack of knowledge can lead to people going into higher education unaware of the debt they may accumulate or how much that degree or certification will actually correlate to a direct increase in salary to pay off this new debt. This is a major problem in our country and it is important to start having these conversations surrounding the transparency of cost-benefit reward in higher education as well as the accessibility of the knowledge and statistics that we do have access to but are not easily found or interpreted by the general population.

1.5 Pre-existing Debt

Another obstacle to consider is that of the disadvantages of pre-existing debt. When these for-profit institutions have made their money and a student walks away with their degree, they may also be walking away with a mountain of student loans. They have a degree which theoretically might be able to get them a higher paying job, but there is also an additional monthly expense that is about to follow them for a very long time. In fact, student loans are one of the very few debts that one can't claim bankruptcy on, meaning that they will follow a person around for life. The question is if the extra money they're making is really worth it. Even if they do manage to advance into a more profitable career path, how many years is it going to take before one sees the financial benefits of your degree? What does one do if they have entered into a field that isn't hiring? Suddenly they could be stuck in the same position they were in before all of this started but now with even more debt accumulating. They sought out education in hopes of alleviating debt and have fallen victim to an expensive and sometimes misleading schooling market. But just how expensive is it? How much debt does higher education lead to? As I discuss in the next paragraph, there is a drastically increasing trend in the cost of higher education, and it has not been for the better.

1.6 Rising Cost and Inaccessibility of Higher Education

The cost of college tuition and fees increased over three times faster than the overall Consumer Price Index and 60 percent faster than medical care costs between 2000 and 2019

(Cebula & Koch, 2021). In a world where essentials are becoming unreachable to certain socio-economic classes, it is unfortunate that many forms of higher education continue to also be more and more unrealistic for those in lower income households. This is why it is important to have conversations around the reality of higher education and how we can spread awareness of options available to those who can't afford the traditional path of education pushed on Americans. I would like to mention an unpleasant reality to the work I have set out to achieve. It is possible that this project seems to imply that there is in fact some form of higher education available to those in any socio-economic situations that will lead to long term financial success. This is unfortunately not always the case, and, in this work, I do not intend to imply that there are any easy solutions to those trying to seek out an affordable education in the realities of financial or social struggles. Regardless of results that I am able to generate, a large part of what's available to a student is dependent on the specific locality of that individual. Those in more vulnerable populations rarely have the luxury of moving to be closer to the ideal program for them which only continues to limit their options for education. Through my project, I hope to spread awareness of this struggle and discuss some of the financial dangers and challenges to seeking out higher education in America.

1.7 Where CS Meets Education Research: Educational Data Mining

The field of Educational Data Mining (EMD) is one that has been continuing to grow over the past few decades. The basic concept of EDM is the extraction and analysis of massive datasets to discover patterns in education. It is the intersection between Education research and Knowledge Discovery in Databases (KDD). The methods that are used in EDM are essentially the same as any form of Data Mining, this includes prediction, classification, regression,

sequential pattern mining, clustering, and more. Anita Chaware, professor at SNDT Womens University in India discusses various works in the field of Computer Science that have centered their study around EDM. These examples include using decision tree algorithms to predict the failure or success of school children, using EDM to analyze learning management systems such as Moodle with outlier and social network analysis to detect outliers among students who may need different forms of learning, as well as attempting to predict the most appropriate major (meaning students who will succeed best academically in a given major) for students entering higher education (Chaware, n.d.). It is a field that is rich in any imaginable form of data mining being applied to education related datasets to find patterns and predications within the world of education. As a result of the applications of EDM being so widespread, they can be used for both the benefit of students as well as the benefit of institutions, which may or may not always be a good thing.

A study in 2018 explored the possibility of predicting 4-year college graduation from student applications alone using machine learning models (Hutt et al., 2018). They broke down the categories contributing to college graduation to Person and Family, Academics + Standardized Tests, Extracurriculars + Work Experience, Honors, Teacher Report + Secondary School Report, and Institutional Graduation Rates. They then used a combination of machine learning techniques including random forests, naive Bayes, logistic regression and gradient boosted decision trees, to name a few. They noted that logistic regression was actually one of their worst performing techniques which was interesting since that is considered one of the more traditional analytic approaches used in this area. While there is no further discussion in the matter, it seems to imply that there may be a need to reanalyze approaches in EDM as advances in machine learning are made. In the end, they were able to accurately predict over 2/3rds of

student's graduation status from their application data alone. The applications of this research are where there seem to be some interesting conversations. They specify that their model should not be used by college admissions to make decisions regarding the acceptance of students, however, it seems like somewhat of an inevitability in this field that once these models exist, there will be institutions that take advantage of this data.

1.8 Program Overview

The program that I have designed for my project has two main components. The first being a display of summary statistics relating to average monthly income and loans based on what type of undergraduate degree and major one might have. The intent behind this side of the program is a user-friendly way to quickly view statistics surrounding the financial outcomes of varying types of degrees. The second part of my program is a prediction model that predicts whether or not one will be employed given the type of degree, gender, field of study, and type of institution (referring to 4-year public nonprofit institutes, 2-year for profits institutes…etc.) that is chosen. For this I have used a regression-based model that I will go into further detail about in the following section. The choice to create a regression-based model was made due to the overall size of my dataset not allowing for some more sophisticated machine-learning based techniques as well as the purpose of this project being based in further understanding relationships between my chosen variables. Machine learning models are made to create the most accurate predictions possible, while statistical models are designed to infer about relationships between variables. The

intent behind this side of the program is to go one step past what the average user can see by simply looking at data or summary statistics. For my data I used two different datasets both collected from the National Center for Education Statistics. While all of the data that I used is publicly available, it is not given in a format that is clear and easy for one who might not have experience in looking at data to understand and interpret. My hope is that my work will provide an easier way for people to understand and learn more about the challenges and reality of what it truly means to commit considerable time and money into a degree.

2
Methods

2.1 Data Collection and Pre-Processing

There are several datasets that were used to complete this project. The first dataset that I used for my regression models was taken from the National Center for Education Statistics' Educational Longitudinal Study of 2002(ELS:2002). This study followed students in tenth grade in 2002 throughout their secondary and postsecondary years as well as immediate years following postsecondary education. From this dataset I collected the following data:

- Gender
- Income
- Unemployment Status
- Monthly Loan Payments
- Sector of Educational Institution
- Degree Type
- Major of Degree

In this study, income refers to only the total year's income of the respondent in 2011, not the complete household income. The unemployment metric refers to a period of three months or more unemployed while actively seeking employment. Sector of Educational Institution refers to

the institution attended for postsecondary education being either public, private for-profit, or private not-for-profit as well as the length of the programs offered as either two or four years. Degree type includes the categories Certificate or diploma, Associate's degree, Bachelor's degree or Post-bachelor's certificate, Master's degree or Post-master's certificate, as well as Doctoral degree, however, my focus is primarily on undergraduate degrees.

To pre-process this data, I started by isolating the variables needed to build my model and eliminated all errors in the dataset, which included survey legitimate skips, N/As, nonrespondents, as well as item legitimate skips. Unfortunately, this brought the initial dataset size of around 16,000 to about 6,000. In this process I lost all data for the following fields of study: mathematics and statistics, parks/recreation/leisure/fitness studies, philosophy and religious studies, and physical sciences. I also chose to eliminate any respondents that were not labeled as unemployed but reported a yearly income of zero dollars. I did this because the survey's definition of unemployment included the active search for employment meaning that anyone who was choosing to no longer work or who was not the one providing household income would be marked as making zero dollars but would not be marked as unemployed.

The other datasets that I used to generate my summary statistic interface are also from the Nation Center for Education Statistics (NCES), more specifically, the 2008/18 Baccalaureate and Beyond Longitudinal Study (B&B:08/18) as well as the 2012/14 Beginning Postsecondary Students Longitudinal Study (BPS:12/14). Both of these studies follow graduates from U.S colleges and universities throughout and after postsecondary education. Due to privacy laws, the access that I had to this dataset was limited to summary statistics, however, the reason that I chose these datasets was that the size was over ten times larger than the ELS:2002 study and while I couldn't benefit from the larger study in building my regression models, I wanted to at

least use the larger dataset to make my summary statistic interface to provide as accurate information as possible.

2.2 Tools & Software Used

The entirety of this project was created using Python 3.8 in Spyder. I used the package Pandas to store my datasets as dataframes and the library Matplotlib for graphing. I also used the scikit-learn library to build my regression models as well as implement my clustering techniques. All datasets are stored in excel and a large portion of my data pre-processing was completed in excel.

2.3 Building a Summary Statistics Interface

To begin my summary statistics interface, I created two methods as documented in the chart below, one that calculates an average monthly loan payment given an age, field, gender, degree type, and sector (referring to public, private for-profit, private not-for-profit, etc.) and another that calculates an average monthly income given a degree type and a field/major of study. To make these initial calculations I queried my dataframes and then calculated combined weighted averages to produce my results except for a few cases where additional steps were required.

For my dataset on associate's and certificate degree loan information I was limited to the total principle amount of the loan rather than monthly payments. To find the monthly payment, I used the national average student loan interest rate from 2018 to match when the rest of my data was collected, and calculated the amount of monthly interest that would be paid in addition to the monthly amount towards the principle loan. It is also worth noting that due to the constraints of the associate's and certificate degree data from the BPS:12/14 study that the loan averages for those categories are based solely off of the sector of the institution where postsecondary education was received. Due to this, my summary statistics for loan payment for associate's and certificate degrees solely reflect changes in loans based on sector of institution.

To find the average monthly income given a degree type and major, I chose to calculate the post-tax monthly income instead of pre-tax. I wanted the amount that is displayed as monthly income to show net-pay so that it could be compared directly to monthly expenses and loan payments without having to calculate how much one would pay in taxes. To do this I used the national average tax rates for different income brackets and subtracted the corresponding amount from the initial income.

For the interface itself, I have initially kept it incredibly simple with the idea that in the future it could be further developed, possibly even into a web application. I gave the user an option to pull up summary statistics from either Associate's, Bachelor's, or Certificate degrees and following that choice there are two graphs that will be displayed. The first is a bar graph showing the average monthly income for all available fields of study for that given degree type. There is also text besides the graph listing the fields that don't have enough data to produce these averages. This was a limitation due to my access to only the public versions of these datasets summary statistics which meant that I was at the mercy of the NCES to decide when they didn't

have enough data to confidently produce an average. The second graph for bachelor's degree summary statistics is a bar graph that shows the average monthly loan payment for all available fields of study for that given degree type as well. For the other two types of degrees it shows sector of institution compared to monthly loan payments.

Method Chart for Summary Statistic Calculations

METHOD NAME	DESCRIPTON
CALCMONTHLYLOANPAYMENT(AGE, FIELD, GENDER, DEGREETYPE ,INSTITUTE)	Input of one integer and four strings used to calculate mean monthly loan payment. Output is finalMonthlyLoanTotal of type float. Given string inputs must match to available inputs for each category as shown below.
FINDAVERAGEINCOME(DEGREETYPE, FIELD)	Input of two strings to calculate mean monthly income given DegreeType and Field. Output is averageMonthlyIncome of type float. The given String inputs must match to available inputs for each category as shown below. For this method, Field inputs available differ for DegreeType "Bachelor" due to data availability.

Figure 2.1 Documentation for methods created to calculate monthly loan payments and average income. Table displaying the name of the method created as well as general description.

Available Inputs for Summary Statistics Methods

CALCMONTHLYLOANPAYMENT():

INPUT NAME	AVAILABLE INPUTS
FIELD	"STEM major", "Computer and information sciences","Engineering and engineering technology ", "Biology and science and mathematics", "Non-STEM major", "General studies and other", "Social sciences ", "Humanities ", "Health care fields ", "Business ", "Education "
GENDER	"Male","Female","Gender minority/other"
DEGREETYPE	"Bachelor","Associate","Certificate"
INSTITUTE	"Public","Private nonprofit", "Private for-profit"

FINDAVERAGEINCOME():

INPUT NAME	AVAILABLE INPUTS
DEGREETYPE	"Bachelor","Associate","Certificate"
FIELD("BACHELOR")	"STEM","Mathematics","Natural science","Engineering/engineering technology","Computer/information sciences","Non-STEM","Social/behavioral sciences","Humanities","Health care","Business","Education","Other"

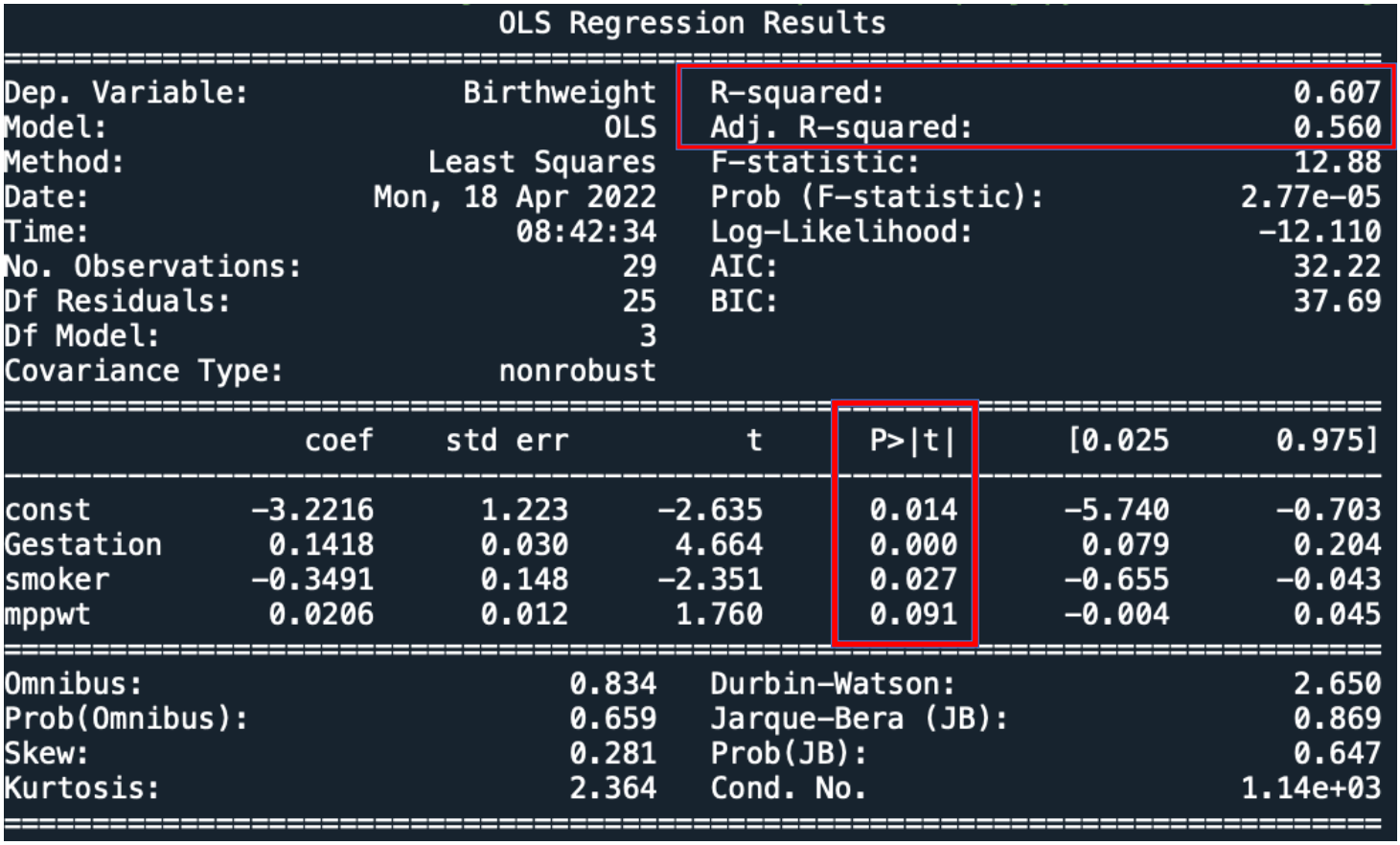
FIELD(“ASSOCIATE”) OR FIELD(“CERTIFICATE”)	"Health care","Personal and consumer services","Manufacturing,construction,repair,and transportation","Other applied fields","Engineering and engineering technology","Business","Undecided","Military technology and protective services","Social sciences and humanities","Computer and information sciences","General studies and other fields","Biology and physical science, science technology, math, agriculture"

Figure 2.2 Documentation of available inputs for each method used in calculating summary statistics. Available input lists are specific to dataset category names.

2.4 Regression Models

My hypothesis going into this work was that there would be a significant relationship between degree types, majors, sectors, and genders to both income and student loan payments. To test this hypothesis as well as gather more information regarding the degree of these relationships, I built several multiple linear regression models. To build my regression models I used the library scikit-learn’s LinearRegression class which computes an ordinary least squares linear regression, as well as their methods to split data into randomly selected testing and training subsets. To evaluate what the output should look like as well as confirm that my methods were

accurate, I first built a model on a sample dataset that contained both binary and categorical variables from the University of Sheffield's dataset collection for teaching that was intended to show an example of a working regression model that I could compare to their results. The dataset that I chose provided information on the birth weight and length of a baby as well as the length of gestation, smoking status of the mother, and the mother and father's height. For this sample model I made the dependent variable the weight of the baby at birth (in lbs.) and made smoker (smoking status of mother), mppwt (weight of mother before pregnancy), and gestation(Gestational age in weeks) into my independent variables. I then used the existing methods to split my data. For this model I chose a standard 80/20 split for training and test data. Once I had split my data, I simply used the LinearRegression() class described above and printed off the summary of the training set regression results of the model. The output is as shown below, with the sections most relevant highlighted.



OLS Regression Results

Dep. Variable:	Birthweight	R-squared:	0.607
Model:	OLS	Adj. R-squared:	0.560
Method:	Least Squares	F-statistic:	12.88
Date:	Mon, 18 Apr 2022	Prob (F-statistic):	2.77e-05
Time:	08:42:34	Log-Likelihood:	-12.110
No. Observations:	29	AIC:	32.22
Df Residuals:	25	BIC:	37.69
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.2216	1.223	-2.635	0.014	-5.740	-0.703
Gestation	0.1418	0.030	4.664	0.000	0.079	0.204
smoker	-0.3491	0.148	-2.351	0.027	-0.655	-0.043
mppwt	0.0206	0.012	1.760	0.091	-0.004	0.045

Omnibus:	0.834	Durbin-Watson:	2.650
Prob(Omnibus):	0.659	Jarque-Bera (JB):	0.869
Skew:	0.281	Prob(JB):	0.647
Kurtosis:	2.364	Cond. No.	1.14e+03

Figure 2.3 Regression results of training data where birthweight is the dependent variable and gestation, smoking status of the mother, and mother pre-pregnancy weight are independent variables.

I compared my results to The University of Sheffield's results in their regression tutorial with the same data to find that my R squared values were nearly the same with their R-squared value of 0.6104 and adjusted R-squared value at 0.5796. The slight difference between our results can be accounted for in the pre-processing of their data as they rounded their data and chose a 70/30 split for testing and training sets whereas I completed an 80/20 split and did not round the data.

There are several values in the summary above that inform us of the significance of relationships between our dependent and independent variables. One of these values is the p-

value which gives us levels of statistical significance, shown above in the highlighted column P > |t|. Using a standard alpha level of 0.05, by checking if our p-values are less than our alpha, we can say that the weight of the mother pre-pregnancy is not statistically significant in regards to its linear relation to the weight of a baby at birth. This informs us that if we were to continue forward analyzing this dataset, it might be worthwhile to drop this variable from our model. Other important values to analyze from our summary above are the R squared and adjusted R squared values. R squared shows how much variance can be accounted for or explained by the model. For this value, the higher R squared value the better the fit the model is. In our sample model above, we have a R squared value of 0.607, which means that 60% of variance can be explained by our independent variables. The adjusted R squared value is the same thing except that it only includes the statistically significant variables in the calculation. If we look at the adjusted R squared we see our value drops to 0.560, or 56%, which is still enough to show a moderate level of correlation. The last step in linear regression was to plot our predictions using our reserved testing data. For this I used the Seaborn library's regression plotting to plot the regression line.

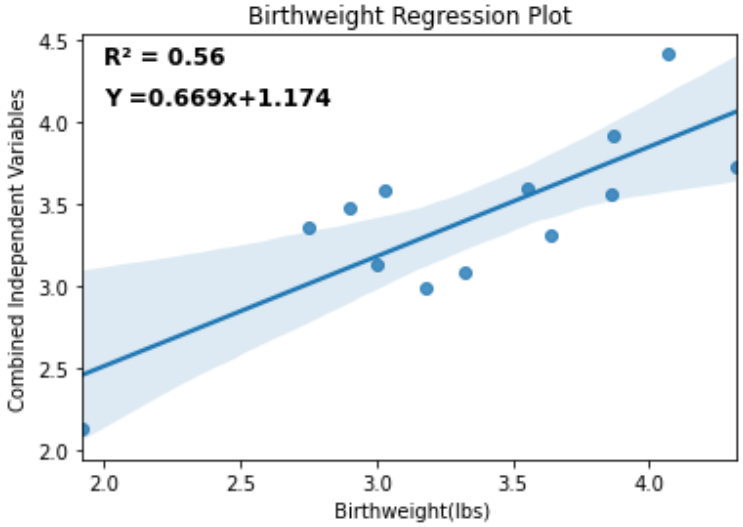


Figure 2.4 Regression plot for birthweight related to gestation, smoking status of the mother, and mother pre-pregnancy weight. Regression line is shown as well as observed values, R-squared value, and equation of the regression line.

Since this dataset is rather small, it is difficult to see much from the above however this type of visualization will be more helpful in analyzing the results of my actual, much larger dataset later on. To summarize what the above graph is showing we can say that the x-axis represents our single dependent variable while the y-axis is a numerical representation of the combinations of our independent variables. The line showed is our line of best fit from multiple linear regression analysis and the blue band surrounding it shows a 95% confident interval.

2.5 Statistical Testing for Correlation & Clusters

After building my regression models with my actual dataset, I found that there was not a significant correlation between variables and thus was unable to produce accurate prediction. I

will discuss the details of these results in the following sections, however, due to the regression models not accurately being able to predict income or student loans, I wanted to see if there was correlation or any significant relationships between any of my variables. To do this I ran further tests for correlation and relationships between all variables used in the regression models. I ran correlation tests on all subsets of my data to determine if there was a relationship between specific variables when isolated from each other. For comparisons including at least one set of ordinal data I used Spearman's Rank correlation and for degree of relation between my dichotomous and continuous variables I used point-biserial correlation. After tests for correlation I wanted to see if there were subsets of my dataset that could be analyzed to find correlation, as well as look for relationships between two sets of categorical data. To do this I used Kmeans clustering. For the significant clusters, I ran a one-way ANOVA test to determine whether or not I could find any significant differences in the means of these subsets in relation to my independent variables: income and monthly student loan payments. I then performed a Tukey's Post Hoc (HSD) test to find which specific subsets had a mean significantly higher or lower than the other subsets of data.

3
Results

3.1 Summary Statistic Results

In order to observe the relationship between a degree type and the corresponding income or loan payments in a manner that requires little scientific background knowledge, my summary statistics are displayed in bar graphs for each specific degree type. It is important to note that the data for mean loan payments was from the 2008/18 Baccalaureate and Beyond Longitudinal Study (B&B:08/18) while data used for the mean income was from the 2012/14 Beginning Postsecondary Students Longitudinal Study (BPS:12/14). As a result of this, the categories of fields of study are not the same in each dataset and therefore not the same in the output figures. This will remain the case in all output figures for my summary statistics. The following are the figures that my summary statistics outputs when given the input “Bachelors”:

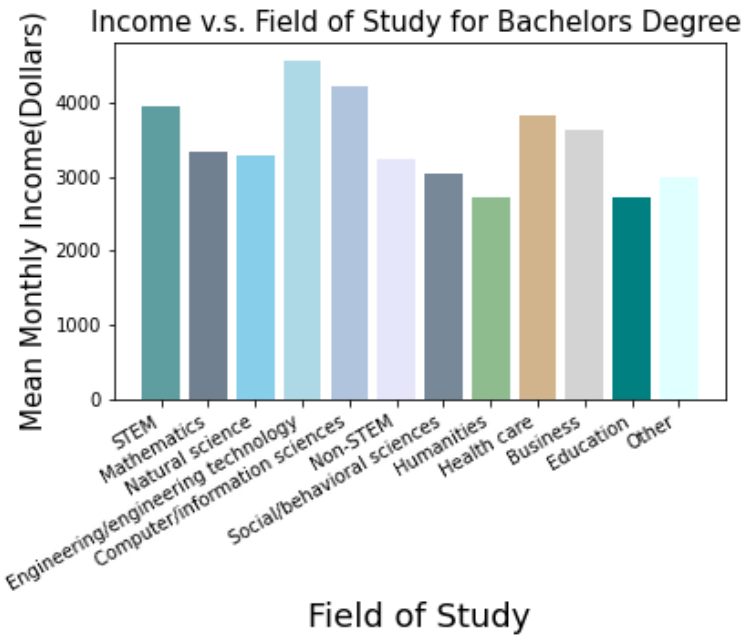


Figure 3.1 Mean monthly income after completion of a bachelor's degree in various fields of study

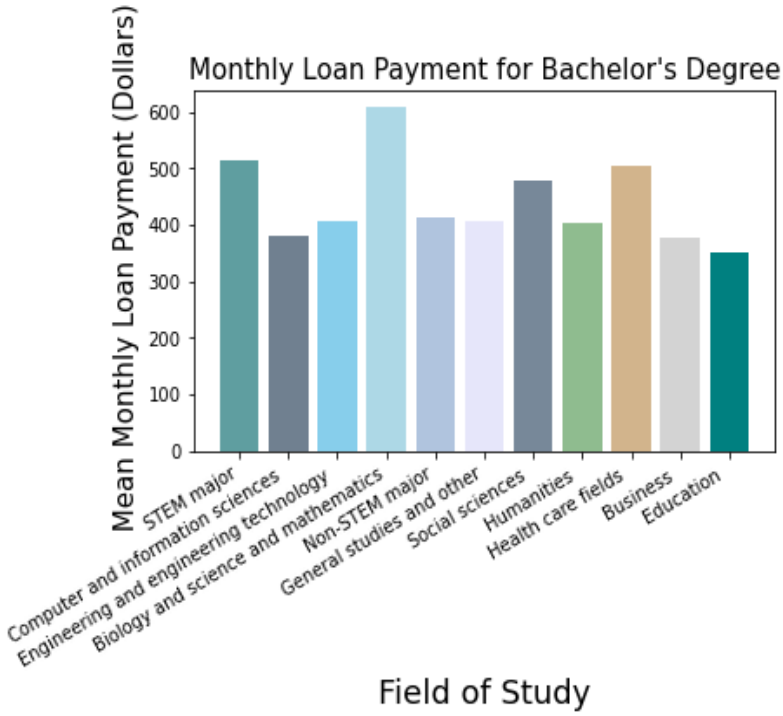


Figure 3.2 Mean monthly student loan payment after completion of a bachelor's degree in various fields of study

Our first figure giving summary statistics for Bachelor's degree data shows us that engineering/engineering technology as well as Computer/information sciences lead to a higher monthly income. The lowest monthly average comes from Humanities studies. When we look at the second figure produced relating to loan payment, we see that biology and science and mathematics have the highest monthly loan payment

We then observe the relationship between associate's degrees and income or student loan payments. The one difference is that the second figure will now show the mean loan payments for associate's degrees given the sector of the institution attended. The following figures are what my summary statistics outputs when given the input "Associates":

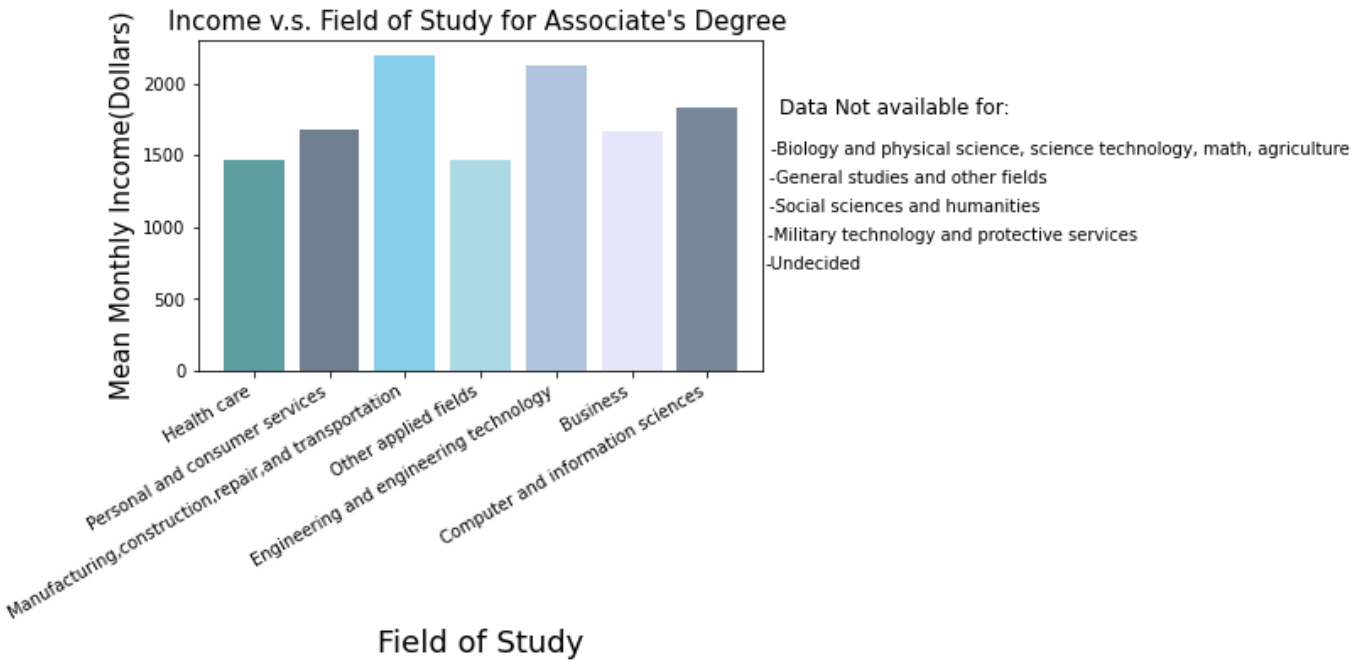


Figure 3.3 Mean monthly income after completion of an associate's degree in various fields of study

Figure 3.4 Mean monthly student loan payment after completion of an associate's degree at institutions in various sectors

The first figure providing summary statistics for Associate's degree income also has Engineering and engineering technology as well as Computer and information sciences in the top three categories for highest monthly income, however, the category: Manufacturing, construction, repair, and transportation takes the number one spot for highest income. It is also worth noting that this was simply not a given category for those completing bachelor's degrees. This reflects most trade schools or programs being a two-year associate's degree. The second figure providing a summary of monthly loan payments shows that private for-profit institutions result in the highest monthly loan payments. This reflects what was discussed around the cons of private for-profit institutions, notably being that they provide less financial aid and are more expensive then public institutions. We see the mean monthly loan payment at private for-profit institutions as almost $150 more than Associate's degrees at public institutions. It is interesting that when it comes to private nonprofit institutions, we only see a slight decrease in the mean monthly loan payment when compared to private for-profit institutions.

We then observe the relationship between certificates less than an associate's degree and income or student loan payments. The following figures are what my summary statistics outputs when given the input "Certificate":

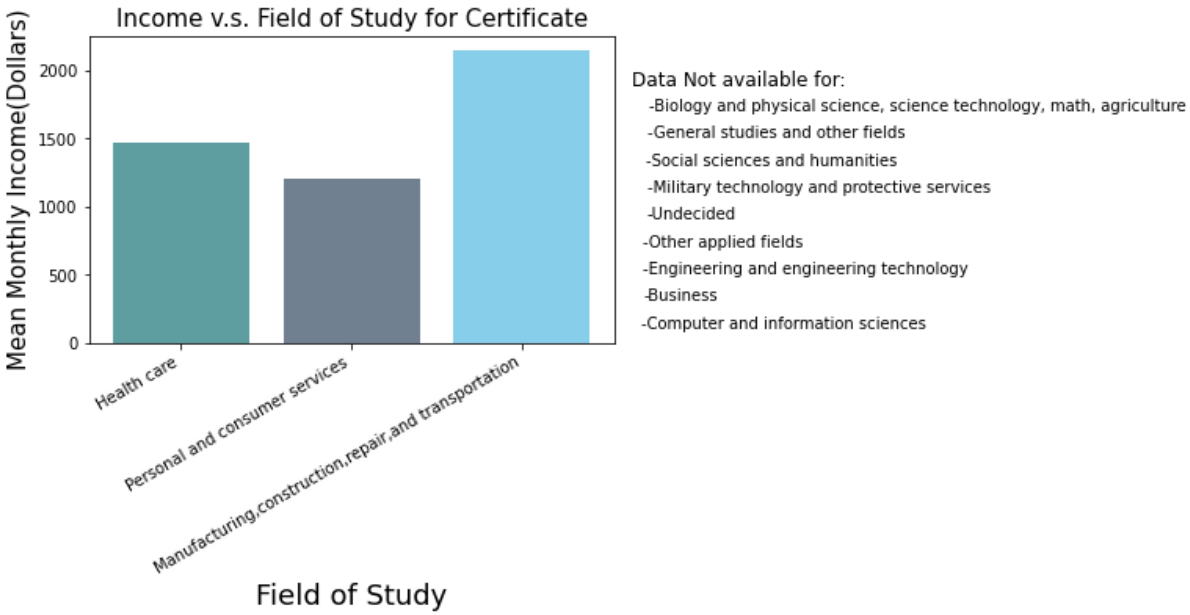


Figure 3.5 Mean monthly income after completion of a certificate in various fields of study

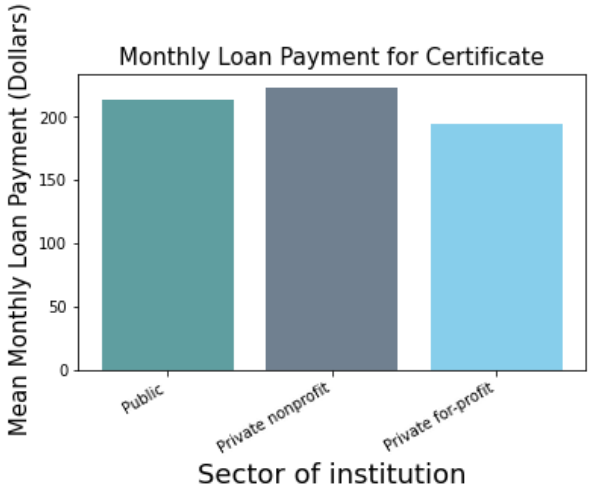


Figure 3.6 Mean monthly student loan payment after completion of a certificate at institutions in various sectors

Our results also show that the mean monthly income for those in health care with a Certificate sits close to $1500 a month, the same field of study in an associate's degree also sits at almost $1500 a month. The same field of study in a bachelor's degree raises the mean monthly income by almost $3000. The monthly loan payments for certificates does slightly contradict what we found in loan payments for associate's degrees in the sense that there doesn't seem to be that large of a difference between any of the sectors. This could be an accurate reflection of the population; however, this could also be a result of this being the category with the least amount of data available. Something interesting to note in the results of monthly mean income for Certificate degrees/certifications is that there is significantly less data available. This could be due to certificate programs not always being affiliated to larger universities which can result in a lack of reporting this type of data. It is also possible that our results can tell us what the more common certificate certifications are, in this case: health care, personal and consumer services, and the field of manufacturing, construction, repair, and transportation.

3.2 Multiple Linear Regression Model Results

Multiple linear regression was carried out in order to predict income and loan payments from our independent variables as well as further understand the degree of the relationship between an individual's student loans and their gender, income, type of degree, sector of institution attended, and major of degree attained. The results indicated that the model was not a significant predictor of student loans, $F(6,1884) = 28.59$, $p = 7.02e\text{-}33$, with an R-squared of

0.081. With only 8.1% of variance in the data explained by the independent variables, the model produced a statistically weak association. When comparing the p-values to our alpha value of 0.05, there was a significant relationship between Loans and Income ($t = 2.646$, $p = 0.008$). There was also a significant relationship between Loans and Degree Type ($t = 11.903$, $p = 0.000$). There was no significant relationship found between the remaining variables which all contained $p > 0.05$ and $-2 < t < 2$.

OLS Regression Results

Dep. Variable:	Loans	R-squared:	0.083
Model:	OLS	Adj. R-squared:	0.081
Method:	Least Squares	F-statistic:	28.59
Date:	Sat, 09 Apr 2022	Prob (F-statistic):	7.02e-33
Time:	16:57:26	Log-Likelihood:	-13419.
No. Observations:	1891	AIC:	2.685e+04
Df Residuals:	1884	BIC:	2.689e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.7179	38.675	2.165	0.031	7.868	159.568
Income	0.0008	0.000	2.646	0.008	0.000	0.001
F3TZHIGHDEG(DegreeType)	82.4174	6.924	11.903	0.000	68.838	95.997
F2PS1SEC(Sector)	-1.6302	6.025	-0.271	0.787	-13.446	10.185
BYSEX	-25.2557	13.783	-1.832	0.067	-52.287	1.775
F3TZASC1CIP2(Associates Dgree Major)	-0.0787	0.483	-0.163	0.871	-1.027	0.869
F3TZBCH1CIP2(Bachelors Major)	-0.1327	0.377	-0.352	0.725	-0.872	0.606

Omnibus:	1012.404	Durbin-Watson:	1.990
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8135.007
Skew:	2.412	Prob(JB):	0.00
Kurtosis:	11.943	Cond. No.	2.32e+05

Figure 3.7 Multiple linear regression results of training data where monthly student loan payment is the dependent variable. Monthly income, degree type, sector of institution attended, associate's degree major, and bachelor's degree major are the independent variables

A regression plot was generated including a line of best fit graphed with our remaining test data predictions to produce the following:

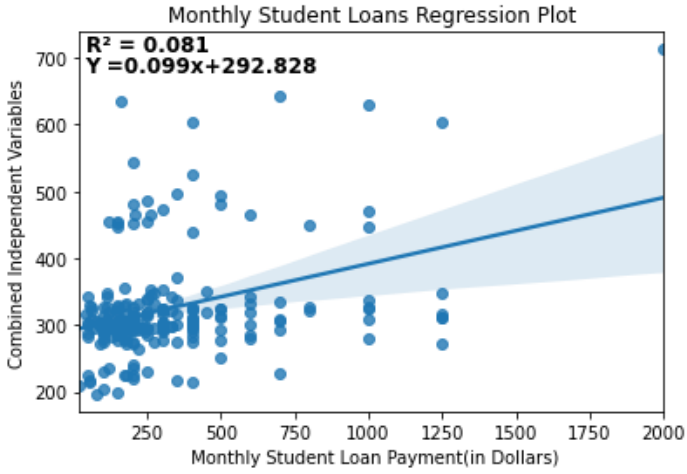
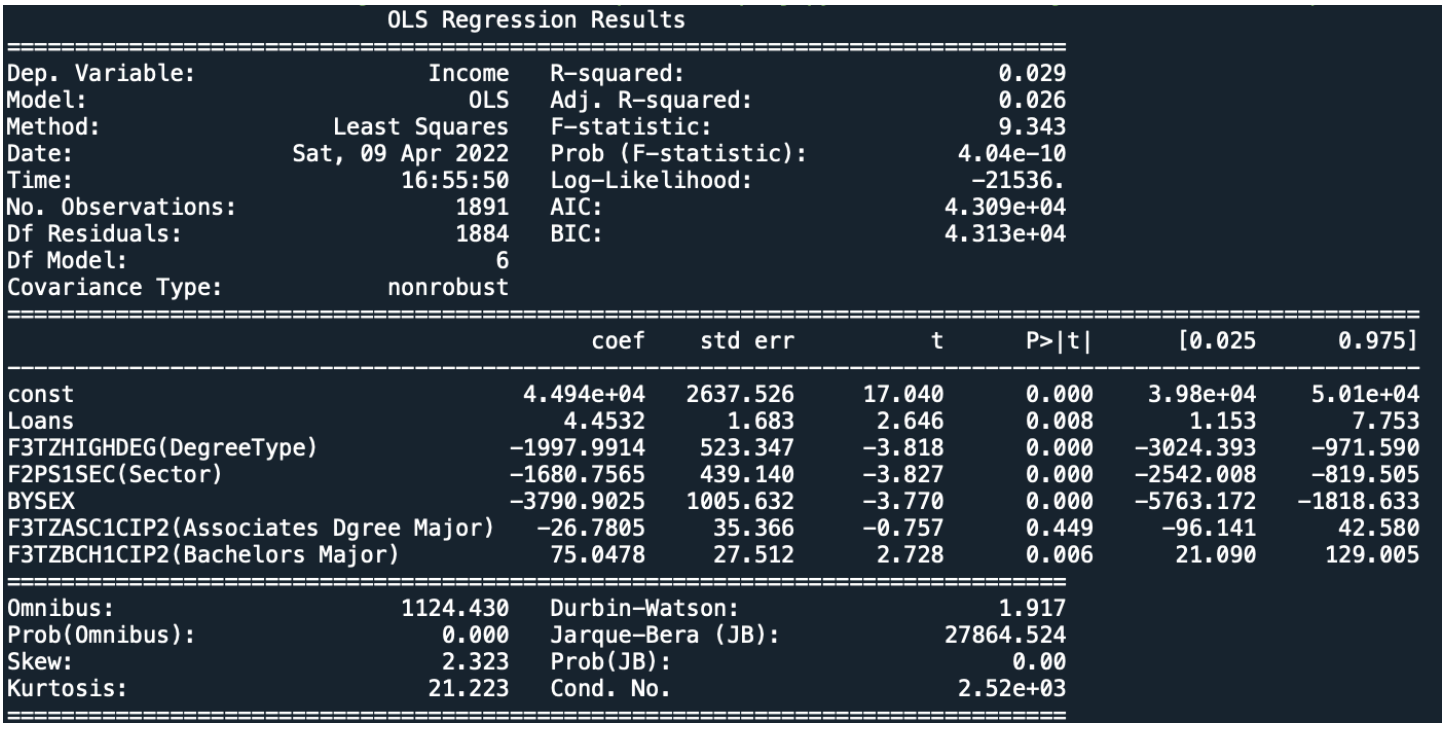


Figure 3.8 Regression plot for monthly student loan payments related monthly income, degree type, sector of institution attended, associate's degree major, and bachelor's degree major. Regression line is shown as well as observed values, R-squared value, and equation of the regression line.

There are several notable characteristics to this graph that help us to interpret our regression results. The first is that there is a clear outlier in the top right corner of the plot. This could be a result of several different factors, and in future testing this outlier should be further studied and potentially removed. The outlier in this case is causing our line of best fit to have a much steeper slope than the rest of the predictions suggest. If we look solely at the trend of the predictions, we see an almost horizontal line of spread. A nearly horizontal line of best fit implies that there is little to no relation between our variables, rather, that there is more of a random scattering of points on a grid then a traceable trend. Another interesting observation is the large cluster of points on the bottom left of the plot. It is hard to tell exactly what could be the cause of this cluster, however it is possible that our dataset simply has a disproportionate amount of data showing lower monthly student loan payments. Given our low R-squared value and visual observations from the plot above, our model does not fit our data.

Multiple linear regression was carried out in order to explore the relationship between an individual's income and their gender, student loans, type of degree, sector of institution attended, and major of degree attained. The results indicated that the model was not a significant predictor of student loans, F (6,1884) = 9.343, p = 4.04e-10, with an R-squared of 0.026. With only 2.6% of variance in the data explained by the independent variables, the model produced a statistically very weak association. There was a significant relationship between Income and Loans (t = 2.646, p = 0.008), Income and Degree Type (t = -3.818, p = 0.000), Income and Sector (t = -3.827, p = 0.000), Income and Sex (t = -3.770, p = 0.000), and Income and Bachelor's Major (t = 2.728, p = 0.006). There was There was no significant relationship found between the remaining variables which contained $p > 0.05$ and $-2 < t < 2$.



OLS Regression Results

Dep. Variable:	Income	R-squared:	0.029
Model:	OLS	Adj. R-squared:	0.026
Method:	Least Squares	F-statistic:	9.343
Date:	Sat, 09 Apr 2022	Prob (F-statistic):	4.04e-10
Time:	16:55:50	Log-Likelihood:	-21536.
No. Observations:	1891	AIC:	4.309e+04
Df Residuals:	1884	BIC:	4.313e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.494e+04	2637.526	17.040	0.000	3.98e+04	5.01e+04
Loans	4.4532	1.683	2.646	0.008	1.153	7.753
F3TZHIGHDEG(DegreeType)	-1997.9914	523.347	-3.818	0.000	-3024.393	-971.590
F2PS1SEC(Sector)	-1680.7565	439.140	-3.827	0.000	-2542.008	-819.505
BYSEX	-3790.9025	1005.632	-3.770	0.000	-5763.172	-1818.633
F3TZASC1CIP2(Associates Dgree Major)	-26.7805	35.366	-0.757	0.449	-96.141	42.580
F3TZBCH1CIP2(Bachelors Major)	75.0478	27.512	2.728	0.006	21.090	129.005

Omnibus:	1124.430	Durbin-Watson:	1.917
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27864.524
Skew:	2.323	Prob(JB):	0.00
Kurtosis:	21.223	Cond. No.	2.52e+03

Figure 3.9 Multiple linear regression results of training data where monthly income is the dependent variable. Monthly student loan payments, degree type, sector of institution attended, associate's degree major, and bachelor's degree major are the independent variables

A regression plot was generated including a line of best fit graphed with our remaining test data predictions to produce the following:

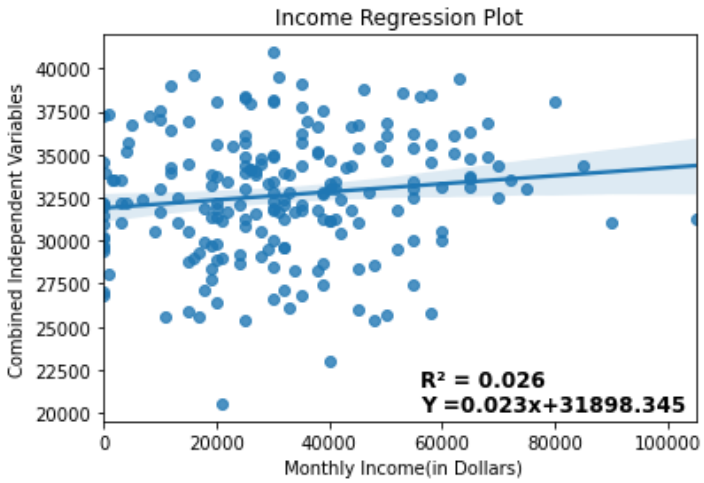


Figure 3.10 Regression plot for monthly income related monthly student loan payments, degree type, sector of institution attended, associate's degree major, and bachelor's degree major. Regression line is shown as well as observed values, R-squared value, and equation of the regression line.

In the above plot we see similar trends to our previous regression plot. The slop is very small which further indicates the lack of relationship between variables. In addition, in this plot we can see that there is an overall lack of any trend with our points appearing to be somewhat randomly scattered across our grid. Given our low R-squared value and visual observations from the plot above, our model does not fit our data.

3.3 Correlation & Clustering Results

The following is a correlation heatmap of all variables from the dataset used for building the regression models. The entire heatmap is showing Spearman's coefficient of correlation for each possible pair of variables. Since some of these variables are dichotomous or categorical, neither of which are applicable to a Spearman's test, there is only a small portion of the heatmap that we can accurately read. This section is outlined in red. We can note that there are no coefficients of correlation above 0.23 which is considered insignificant. The full results for each of my Spearman's correlation tests are as follows:

Spearman's rank correlation was computed to assess the relationship between Income and Sector, Income and Gender, Income and Degree Type, Income and Associate's Major, Income and Bachelor's Major, Loans and Sector, Loans and Gender, Loans and Degree Type, Loans and Associate's Major, and Loans and Bachelor's Major. Between Income and Sector there was a negative correlation between the two variables, $r = -0.063$, $p = 0.004$. Between Income and Gender there was a negative correlation, $r = -0.074$, $p = 0.001$. Between Income and Degree Type there was a negative correlation, $r = -0.001$, $p = 0.947$. Between Income and Associate's Major there was a negative correlation, $r = -0.09$, $p = 3.586$. Between Income and Bachelor's Major there was a positive correlation, $r = 0.11$, $p = 4.828$. Between Loans and Sector there was a negative correlation, $r = -0.036$, $p = 0.113$. Between Loans and Gender there was a negative correlation, $r = -0.0551$, $p = 0.012$. Between Loans and Degree Type there was a positive correlation, $r = 0.227$, $p = 5.468$. Between Loans and Associate's Major there was a negative correlation, $r = -0.136$, $p = 3.374$. Between Loans and Bachelor's Major there was a

positive correlation, r = 0.09, p = 3.263. None of the Spearman's test for correlation showed significant correlation.

Point-biserial correlation was computed to assess the relationship between Unemployment Status and Income, and Unemployment Status and Loans. Between Unemployment Status and Income there was a negative correlation, r = -0.221, p = 9.971. Between Unemployment Status and Loans there was a negative correlation, r = -0.010, p = 0.652. None of the point-biserial correlation tests showed significant correlation.

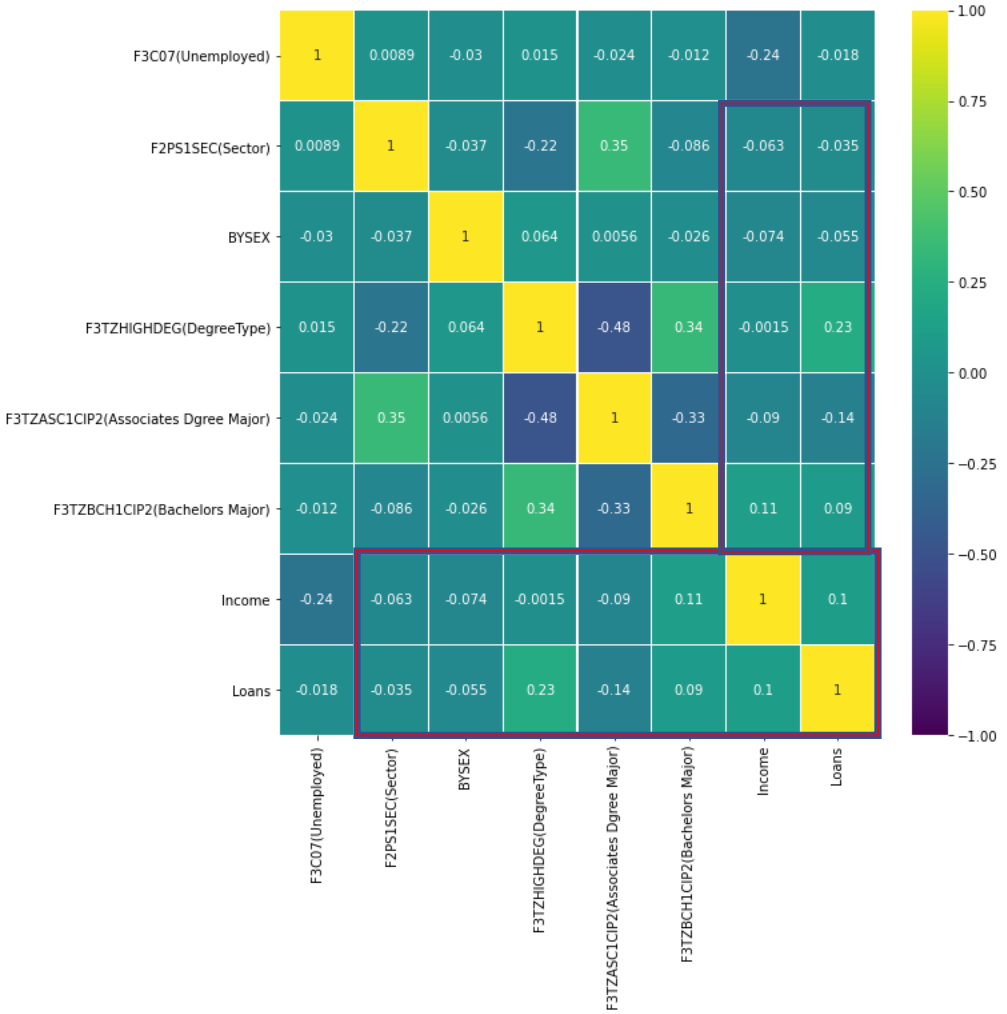


Figure 3.11 Correlation heatmap showing Spearman's coefficient of correlation. Sections not highlighted in red are not accurate due to inaccurate data type for Spearman's rank correlation.

In order to examine relationships between two categorical, non-ordinal variables, I performed a series of Kmeans clustering to analyze any potential relationships. All of the clusters looked like the example below for Loans V.S Sector with the exception of Associate's Major compared to Bachelor's Major.

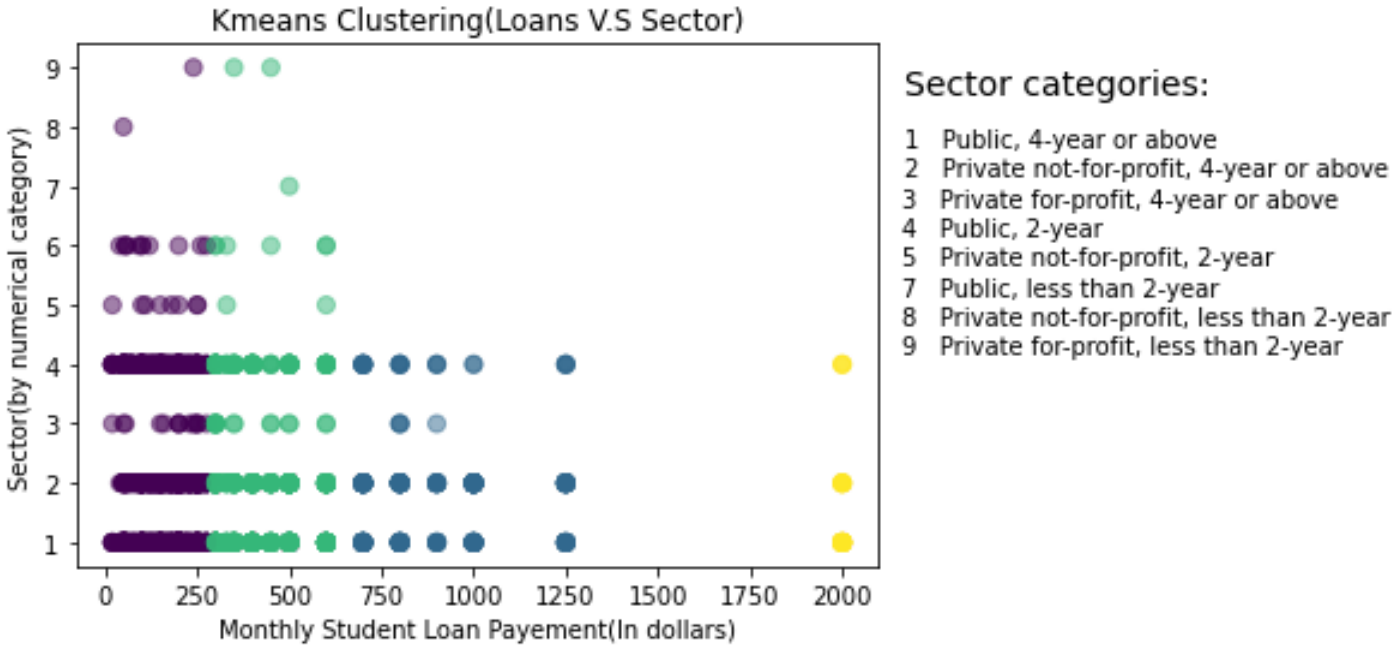


Figure 3.12 Kmeans clustering of sector of institution attended as well as monthly student loan payments

Clusters in the above graph are simply highlighting the different ranges of payment in monthly student loans. There is no interesting or relevant clustering present. Clusters in the graph below, however, show a clear division into 3 categories: those who received an Associate's and Bachelor's in education and stayed in education (blue cluster, bottom-left), those who received an Associate's and Bachelor's in Business/management/marketing(yellow cluster, top-right), and then those who majored in different fields for their Associate's and Bachelor's degrees (Green and purple clusters, top-left and bottom-right). For the purpose of further analysis

of these clusters I combined the top-left and bottom right clusters into one category as switching majors.

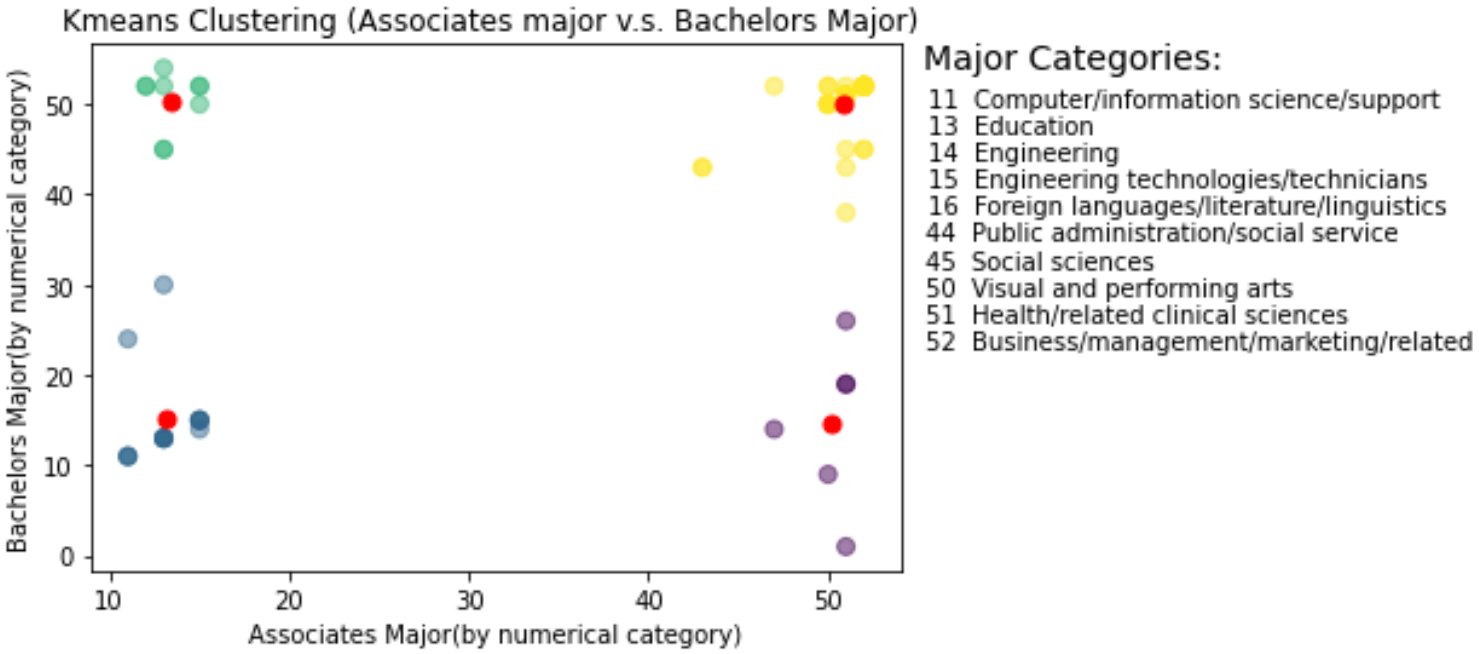


Figure 3.13 Kmeans clustering of associate's major as well as bachelor's major

To test if there was a significance in the data in these clusters to my independent variable of Income, I completed a one-way ANOVA test on the three clusters of data with the following results:

F-Value: 11.707331416560043 P-Value 9.100344771190413e-05

Since our p-value is less than our chosen significance level of 0.05, we can say that the difference between some of the means are significant. I then performed a Tukey HSD comparison to further explore the difference in means. I got the following results where group zero is the clusters that switched majors, group one is the cluster that stayed in education, and group two is the cluster that stayed in business/management/marketing:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-1793.3333	0.9	-15612.3001	12025.6334	False
0	2	11800.0	0.0383	516.8609	23083.1391	True
1	2	13593.3333	0.0142	2310.1942	24876.4725	True

From our results we can say that there is a significant difference in the means between groups zero and two showing that those who switched majors made a mean of $11800 less than those who stayed in business/management/marketing. We can also say that there is a significant difference in the means between groups one and two showing that those who stayed in business/management/marketing for both an Associate's and Bachelor's degree made more than those who stayed in education for an Associate's and Bachelor's degree.

4
Discussion

4.1 Dataset Dilemmas

This project and the methods it contains have been dependent on one crucial component: Datasets, and the variables I selected to use for my testing. In my results, we have now seen that there was no significant correlation found between any given variables. We will now explore the possibility of what could have led to this conclusion in regards to potential errors found in my selection of variables or datasets, what variables could have been missing from my regression, the overall challenges of sourcing public-use datasets, as well what it would mean if my results are in fact an accurate representation of the population.

The selection of these specific datasets was made for several reasons. The first being an issue of public access, time, and money. I was in search of a dataset that was publicly available at little to no cost. The 2008/18 Baccalaureate and Beyond Longitudinal Study (B&B:08/18) as well as the 2012/14 Beginning Postsecondary Students Longitudinal Study (BPS:12/14) were in fact much larger datasets that had more potential variables which may have produced a more accurate

model or shown more correlations and relationship present. Unfortunately, since I could only access the summary statistics version of this dataset, I had to keep looking for a more feasible option to continue with my modeling. The other factor that I was considering was that I wanted to find a national study that would represent a subset of the population of the United States as a whole, rather than a subset of a state or county's population where there could be other geographical factors at play. These requirements led me to The Educational Longitudinal Study of 2002(ELS:2002). While this dataset was significantly smaller, it contained what I believed to be enough variables to potentially create the useable model that I had initially hoped to achieve. One factor that I didn't consider when choosing this dataset was looking at the number of errors as well as hidden data that was in the dataset. As a result of this being a public-use dataset there were certain variables that I was unable to use or that had been coarsened for disclosure avoidance. This had two effects on the usability of this data. The first was that once I had removed all the errors from my desired variables, I was left with about a quarter of the original size of the dataset. This meant that I had to be careful when adding variables because each new variable meant a couple hundred less usable data points that would have to be thrown out due to errors. The second was that the choice of variables was more limited then I would have liked. Unfortunately, several variables that I had hoped to use were looking at an individual's finances which are usually one of the more protected areas of data in public-use files for disclosure avoidance.

The selection of variables in a multiple linear regression model combine to make both a complete and accurate model. When I chose my variables from the datasets, I was looking to create a full picture of everything that contributes towards student loans and post-graduation income. To do this, I had combined what I had read in literature surrounding this topic to see

what has already been shown to have some sort of connection as well as the key components that distinguish one postsecondary education from another to provide a complete picture of what parts of higher education lead to changes in income and student loans.

It is a strong possibility that there are many more variables that I wasn't able to account for that have a significant impact on income and student loans that would have provided me the ability to make a strong predication model. Other factors that we discussed at the beginning of this paper included the effect of the pre-existing socioeconomic standing of a student's family relating often to the highest level of education and income achieved by parents or other household members. When starting on a job search, those in your network are a key resource to find connections, begin networking, and maybe even find potential referrals. If students come from a family that doesn't have these types of connections then these students might have lower initial salaries post-graduation, regardless of the specifics of the degree itself. The connection between networking and income has been explored greatly in the field of economics, Dale Mortensen and Tara Vishwanath (1995) modeled this very network of connections and found that the equilibrium wage distribution is higher if the probability of the offer coming from a contact is higher. This study was done in 1995 and with the boom of technology it would be interesting to see how much of this still reigns true today, however, it is possible that pre-existing networking ability and connections is a necessary, albeit difficult to calculate, variable in the equation of the financial outcomes of higher education.

We have now addressed several factors that could have contributed to my dataset and regression model not being an accurate representation of the population with the given data. While it may be impossible to prove, I would also like to discuss the possibility of my data being an accurate representation of the population and what that would mean in regards to my results.

When it comes to the regression models, there was far too much variance that could not be accounted for by the predictor variables, showing a weakness in the model's ability to accurately make predictions. We saw in the regression plots how this might be the case as, and especially in regards to our Income regression, the plotted predictions seemed to be a collection of mostly random points scattered across a grid. It is possible that a portion of this noise could be accounted for by a problem commonly faced in the world of real data being that humans are simply at times unpredictable. It is possible that there was not enough data to lessen the effect of human noise. This is more likely to be the case due to the fact that we were able to find a large number of independent variables that did have a significant relationship to our dependent variable.

Taking a look at the results from correlation testing paints an interesting picture if we assume that the dataset is in fact an accurate representation of the population. There was no correlation found. This would imply that it doesn't really matter what type of degree you get or what field you study in terms of finances, it means that the variables that do affect these outcomes are still out there, but for us, it means we can say that it's not any of the variables we studied. This also means that there isn't direct correlation between student loans and the sector of institution you attend, which defies a fair amount of the studies that have already been done in fields of sociology and economy.

4.2 Problematic Data Holes

One downside to the data that I had to work with was its lack of representation among certain areas of higher education, notably for specific humanities majors, musicians, dancers, and artists. While there is currently work being done to improve this gap, it is an area that continues to need attention to be able to continue work in the field of educational data mining.

4.3 Future Adjustments and Improvements

There are still several aspects of these datasets that have yet to be fully explored, notably the possibility of finding trends within smaller subsets of data similar to how I found a statistically significant relationship after isolating subsets through clustering. It would be wise to continue in this work by clustering based off of other non-categorical variables and combinations that I ran out of time to attempt. In future improvements to this project I would also try out several other adaptations to my regression models in an attempt for more accurate predictions. One of these would be analyzing, isolating, and eliminating several outliers found in analysis of the regression models and plots. An additional improvement would be eliminating the independent variables that were not found to be statistically significant and rebuilding the models without them.

4.4 Summary Statistics and Lack of Interface

The initial intent behind creating a summary statistics interface was to make it easier for the average user to learn more about expected income and loans associated with their desired degree and area of study. Due to a lack of data for certain fields of study, my inability to produce prediction models with any level of accuracy, as well as the time constraints of this project, there is still considerable work to be done to create an interface that would achieve the initial goal of this work. In an ideal world, this interface would be developed into a web application which would be far more accessible to those outside of computer science than a python file. Further data would also need to be collected and applied to my methods to decrease the number of fields missing as well as add to the statistical strength of this work in regards to producing an accurate representation of the population. While I had initially hoped that this project would result in the creation of this interface, I was unable to achieve this goal. Instead, I have produced a thorough analysis of higher education datasets that can be a starting point to the continued research and exploration of this field as well as highlight where this future research is best targeted, notably, in fields of study and degree types that are less common in existing higher education datasets.

References

Cebula, R. J., & Koch, J. v. (2021). The Crisis in Public Higher Education: A New Perspective. *American Journal of Economics and Sociology*, *80*(1), 113–131. https://doi.org/10.1111/ajes.12373

Cellini, S. R. (2012). FOR-PROFIT HIGHER EDUCATION: AN ASSESSMENT OF COSTS AND BENEFITS. In *National Tax Journal* (Vol. 65, Issue 1).

Hutt, S., Gardener, M., Kamentz, D., Duckworth, A. L., & D'Mello, S. K. (2018). Prospectively predicting 4-Year college graduation from student applications. *ACM International Conference Proceeding Series*, 280–289. https://doi.org/10.1145/3170358.3170395

Lee, S. (n.d.). *The For-Profit Higher Education Industry, By the Numbers — ProPublica*. Retrieved December 7, 2021, from https://www.propublica.org/article/the-for-profit-higher-education-industry-by-the-numbers

Mortensen, D. T., & Vishwanath, T. (1995). Personal Contact and Earnings: It Is Who You Know! *Labour Economics* , *1*(1), 187–201.

Mrs, A., & Chaware, G. (n.d.). Educational Data Mining: An Emerging Trends in Education. *International Journal of Advanced Research in Computer Science*, *2*(6). www.ijarcs.info

The For-Profit Higher Education Industry, By the Numbers — ProPublica. (n.d.). Retrieved December 7, 2021, from https://www.propublica.org/article/the-for-profit-higher-education-industry-by-the-numbers