Group Project Report - Kristine Arens, Cody Burger, Michelle Cutlip, Ben Huntington and Matthew Pelham

The Situation

Our team analyzed data from post-secondary institutions to answer the question: Which factors contribute to higher or lower student loan default rates? We sought to answer three sub-questions:

1. What effect do tuition cost and program length have on an institution's default rate?
2. Is there a correlation between school type and default rate?
3. Does a school's geographic location correlate with its default rate?

Data from the Department of Education Federal Student Aid and the National Center for Education Statistics were joined into a single data frame and analyzed using R to answer these questions.

The Data

Post-secondary loan default data, located on the Department of Education's Federal Student Aid site, included number of students in default, number in repayment, the school's default rate, rate type, ethnic affiliation of the school, and school location (further referred to as "Peps" data). The data contains official cohort default rates published for schools participating in the Title IV student financial assistance programs for fiscal years 2012, 2013 and 2014. The dataset was strengthened with data from the National Center for Education Statistics (further referred to as "IPEDS" data). From this site, key items were pulled including: geographic coordinates, attendance, cost, average grant money received per student, average loan amount received per student, and percentage of students receiving loans. The Peps data was indexed by an OPEID which is the Department of Education's identifier. The IPEDS data was indexed by UNIT_ID. A mapping dataset which contained both the UNIT_ID and OPEID was located so that the two datasets could be merged. The U.S. Department of Education Plan and Policy Development Guidance for Public Access allows and encourages researchers to maximize the data made available to the public.

Importing and combining the data using R

The openlsx and dplyr libraries were loaded. The mapping dataset was imported to a dataframe titled "codes". The loan default data was read into the "peps300" dataframe, and the OPEID number was changed to an integer which removed the leading 0's and the paste function was used to add 2 0's to the end so that its format matched the "codes" dataset. The codes and default data were then merged into the codes_peps300 dataset. IPEDS data was pulled from 7 datasets on their site, and each set was loaded onto its own Excel sheet. Each sheet was read in as a dataframe using "read.xlsx" and then merged into the full dataframe called "full_df" using UNIT_ID as the linking source. Columns with 2014 data were removed, and eight total joins were completed to combine the Peps data.

Preparing the data for analysis

The data required minor cleaning before analyzing. Column names were changed to more meaningful terms, the data set was reduced to complete cases, and the school type was changed to a factor. After purging incomplete cases, the number of school type levels was reduced from six to three: Public, Private (non-profit), and Proprietary institutions. The group decided to analyze the average of 2012 and 2013 data for the amount of tuition and the default rates and created a new variable, "Tuition_Binned". In the Peps dataset, the program length was a numeric value ranging from 3 to 12 which provided no insight, so a new variable was generated for "Program Meaning". A subset of 7 variables out of the original 31 was selected from the full dataset to use in the analysis, and analysis was limited to 2012 and 2013.

Description of R Scripts

The project consists of 4 script files: project.R, dataload.R, defaultanalysis.R, geographicanalysis.R and schooltypeanalysis.R. To produce the analysis, one can copy all of the files to the same directory and perform a Source on the project.R file. The openxlsx package must be installed before running the scripts, and to do so, the first line in the project.R script can be uncommented and this command run. Below is a summary of the contents of the script files.

project.R – This is the main script file and should be the starting point for running functions in the others. It sources the other files to make all functions available in the global environment. Running this file as-is will produce sample analysis plots and tables, but the contents of this file can be modified to produce different outputs by calling functions in different ways.

dataload.R – This script file contains two functions – load.data and unflatten.data. The load.data function loads the data from the sources into data frames. The data from the original sources contain repeating groups of attributes for years 2012 and 2013, and the unflatten.data function turns the repeating groups into two data sets – one for 2012 data and one for 2013, then recombines the data into one single data set. All functions used by the group work on the data frame were produced by the unflatten.data function. The load.data and unflatten.data functions should be called in that order to generate the data used for analysis.

defaultanalysis.R – This file contains the functions used in our analysis of the question – do the program length and tuition cost impact an institution's default rate?

geographicanalysis.R – This file contains the functions used in our analysis of the question - are there any relationships between the geographic location of institutions and the default rate?

schooltypeanalysis.R - This file contains the functions used in our analysis of the question - did the type of university play a part in the default rates at institutions across America?

Global challenges

The group faced several challenges during the project, the first of which was locating data. Our initial plans to use our group members' post-secondary employers' data didn't pan out, so we took to the internet. The default loan data was located, and the group was very interested in studying what variables impacted a school's default percentage. The data didn't appear to need much cleaning, but it also didn't include many variables. It did include the OPEID, and further research ensued to see if additional information regarding

the institutions could be located. Data regarding schools was located on the IPEDS site but was indexed by UNIT_ID. A table was found containing both the OPEID and UNIT_ID so that the two data sources could be merged. Our next challenge came when we discovered that the OPEID was not set up the same way on the datasets. Once those were identified and transformed in R, the data could be merged. Communication and file sharing proved to be another challenge. We initially set up a group on ICON with folders for documents, scripts, and datasets. We quickly found that group members weren't getting updated when comments were added to a discussion item, and it was difficult to manage script versions. We changed our communication to group email and began using GitHub to share files which solved this obstacle.

The Analysis

Analysis 1: What effect do tuition cost and program length have on an institution's default rate?

A summary table was generated to compare the cost of tuition against the length of program using the mean default rate as the measurement (Table 1). It illustrates that shorter programs tend to have higher default rates. A rank-order function was applied to determine which program lengths had the highest default rates (Table 2). Non-Degree (1 yr.), and Associate's Degree's had the highest default rates, where Master's Degrees and Non-Degree (3+yr.) were among the lowest default rates. The table can be interpreted in ascending order – the lower the number, the lower the default rate rank.

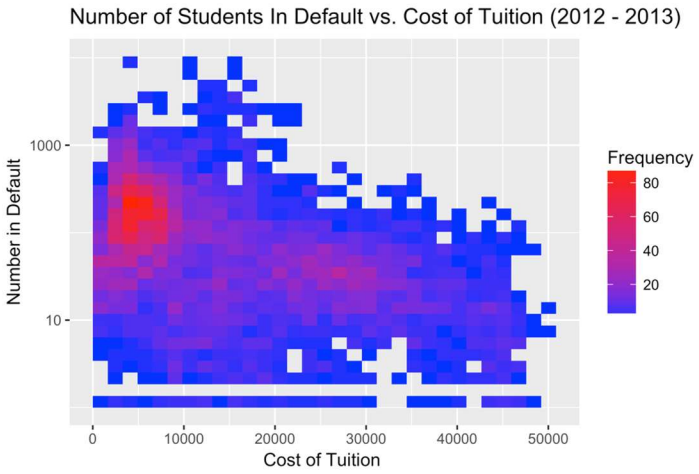
Table 1:

	Prog.Meaning	$1-$10,000	$10,001-$20,000	$20,001-$30,000	$30,001-$40,000	$40,001-$50,000
1	Non-Degree(1 yr)	14.600000	14.034483	0.000000	NA	NA
2	Non-Degree(2 yr)	11.185714	9.634375	8.300000	NA	NA
3	Associate's Degree	17.637187	13.377880	11.005000	7.360000	NA
4	Bachelor's Degree	15.986184	13.020530	8.621875	5.045361	1.807042
5	Master's Degree	8.543638	7.979474	6.419624	3.997658	1.954737
6	Non-Degree(3 yr +)	4.866667	7.626087	NA	NA	NA

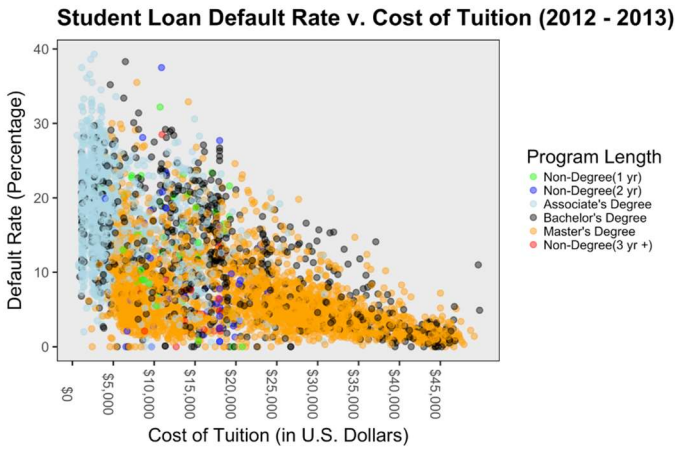
Table 2:

	Prog.Meaning	$1-$10,000	$10,001-$20,000	$20,001-$30,000	$30,001-$40,000	$40,001-$50,000
1	Non-Degree (<1 yr)	3	NA	NA	NA	NA
2	Non-Degree(1 yr)	6	7	1	NA	NA
3	Non-Degree(2 yr)	5	4	4	NA	NA
4	Associate's Degree	8	6	6	3	NA
5	Bachelor's Degree	7	5	5	2	2
6	First Professional Degree	9	1	2	NA	1
7	Master's Degree	4	3	3	1	3
8	Non-Degree(3 yr +)	1	2	NA	NA	NA
9	Two-Year Transfer	2	NA	NA	NA	NA

A heat map was generated to determine the strength of the relationship between the cost of tuition and the amount of the average default. The results were staggering showing a strong inverse correlation—as tuition increases, the number in default decreases.

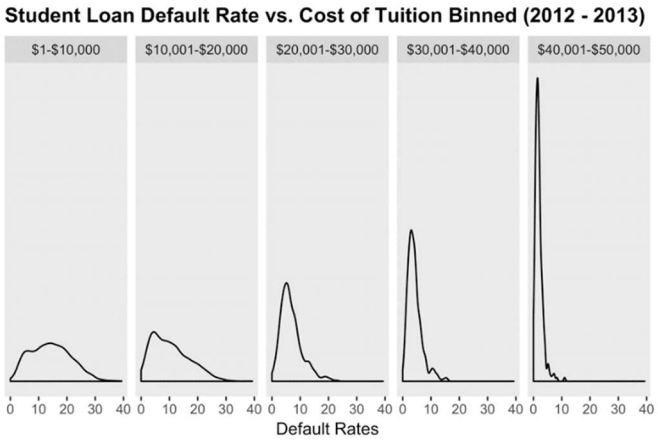


A scatter-plot was created measuring the mean default percentage on the cost of tuition. The findings were consistent with the previous analysis - the lower the tuition, the higher the default rate. The program with the lowest tuition, Associate's Degrees, had the highest default rates. Conversely, the degree with much higher tuition, Master's Degrees, showed lower default rates.



The analysis showed that lower tuition amounts were indicative of higher default rates. To measure the strength of this relationship, a density plot was utilized to measure the default rates across the binned tuition prices. The lower tuition bins were heavily skewed across the spectrum of the default rates and had

more variation in default rates. The higher tuition bins tend to have less variation among their default rates, and the overall default rates were much lower. The plot below shows that as the tuition bins increase, the density skews towards a lower default rate.



Analysis 2: Does the type of university play a part in the default rates at institutions?

To capture a simple snapshot of default rates among the three remaining school types, a table was created to summarize the mean default rate based on school type (Table 3, below). Private institutions had the lowest percentage of students in default in years 2012 and 2013, while Public institutions had the highest percentage of students in default for both years at 14.08% and 13.76% respectively.

Table 3:

	School.Type	2012	2013
1	Public	14.077096	13.758402
2	Private	6.799744	6.733731
3	Proprietary	12.790429	12.008251

Table 4:

	School.Type	2012	2013
1	Public	0.5658683	0.5593727
2	Private	0.9148211	0.9207836
3	Proprietary	0.6468647	0.6765677

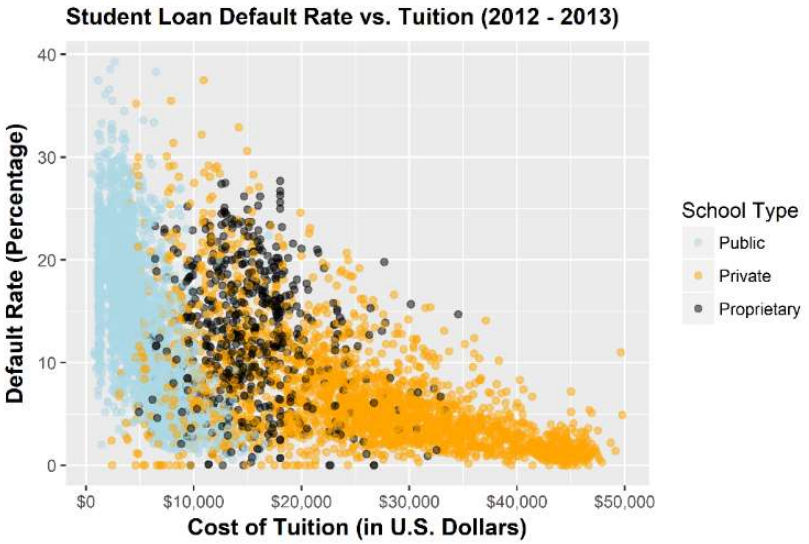
A deeper analysis was conducted to assess how these institutions were performing as it relates to Department of Education (DOE) performance standards. Institutions that maintain a cohort default rate of 15% or below for three consecutive years have greater autonomy with when and how they disburse financial aid awards. To identify which school types were most likely to meet this DOE benchmark, a table was developed grouping together two variables—school type and year, to find the percentage of each school type with default rates 15 % or below (Table 4, above). Private institutions led the way in achieving

this metric for years 2012 and 2013. Public institutions performed worst in this area posting sub-15 % default rates in 2012 and 2013.

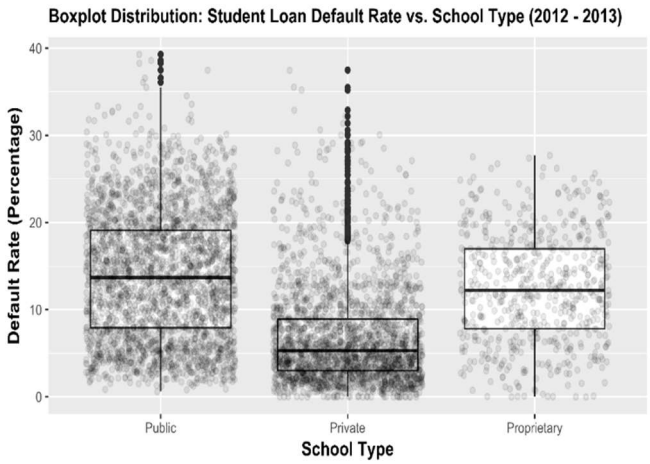
The DOE also monitors institutions with default rates at or above 30 %, and these institutions could be subject to the loss of eligibility to participate in federal financial aid programs. A table was generated grouped by school type and year, and included percentage of each school type with default rates over 30% to identify which school types could be considered most 'at risk' of losing eligibility (Table 5). While all school types boast low percentages in this area, proprietary institutions performed best at 0%. Public institutions trailed behind their counterparts as approximately 1.5 % of schools had default rates of at least 30 % in 2012 and .8% in 2013. Table 5:

	School.Type	2012	2013
1	Public	0.014970060	0.008215086
2	Private	0.005962521	0.001703578
3	Proprietary	0.000000000	0.000000000

In addition to summarizing data from the School Type variable, plots were also generated to visualize how variables were correlated with one another. A scatter plot measures the default rate over cost of tuition for years 2012 and 2013. Within the scatter plot, school type is identified by color. The plot shows clear separation between the three school types. Public Institutions boast a lower cost of tuition, but higher default rates, while private institutions show a higher cost of tuition, but lower default rates. A cluster of proprietary institutions occupies the space in between.



A boxplot with jitter was generated to compare default rate against school type. This plot identifies how observations were distributed among the three school types. It shows that the default rates of public institutions are evenly distributed between all default rates. Public institutions also had the highest median default rate. Private institutions had the lowest median default rate, with a majority observing default rates at or below 5 %. This plot also shows population density of each school type. Public institutions were the most represented school type while proprietary institutions were the least represented.

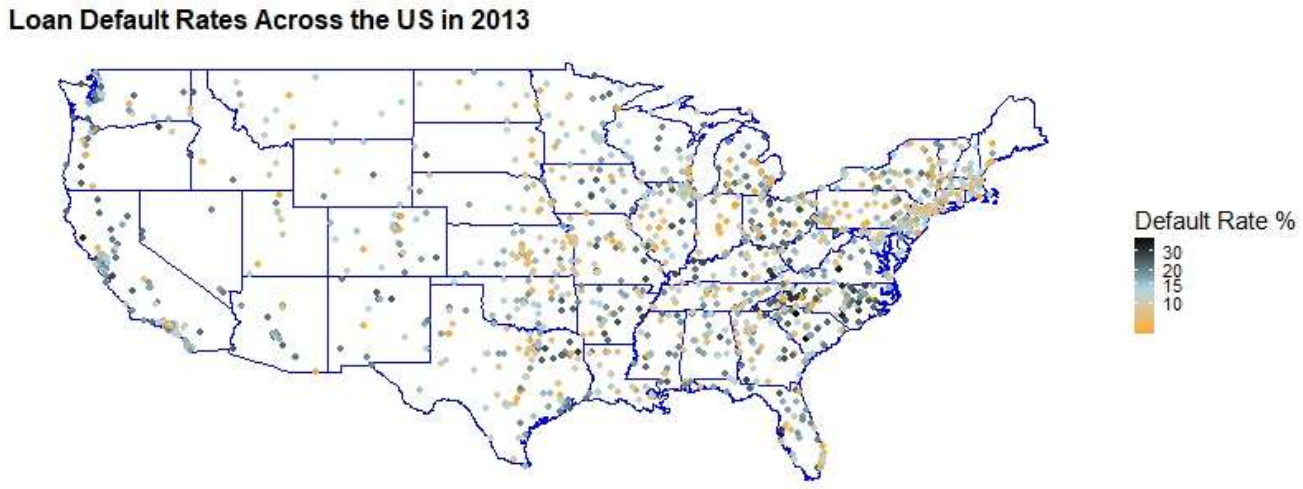


Through this analysis, school type and default rate appear to be correlated. Private institutions outperform both Public institutions and Proprietary institutions when considering some of the most important metrics the DOE assesses. In addition, public institutions are *least* likely to meet significant DOE benchmarks. Unfortunately, the analysis doesn't capture a complete picture of the influence of school type may have on default rates. The levels that are observed in the school type variable are fairly broad. In reality, there are sub-levels within identified DOE levels that would provide greater insight on the significance of this variable (for instance community colleges and public 4-year universities fall within the *same* level of this dataset, but are different in several facets). Through exploratory analysis, our data offers an interesting perspective on the correlation between school type and default rate. Ultimately, however, added levels to the school type definitions would provide the most valuable insight on how the two variables are related.

Analysis 3: What geographic areas have the highest and lowest default rates?

It was decided the best way to answer the question around relationships of geographic location to student loan default rates was to plot the data on a U.S. map with state outlines, enabling us to determine whether specific locations showed higher density of default rates.

Data needed to be adjusted to provide the best viewable maps. One of those adjustments included the exclusion of default rates that were less than 5%. This resulted in a reduction of 647 schools. Removing these schools allowed for a less "cluttered" graphic. Below is an example of the final graphic for student default rates across the US for 2013. The R script allows for user input of the state and year, or the default of the entire U.S. in 2013.



There were several obstacles faced when creating this graphic, namely understanding how to create a ggplot with the U.S. map and state outlines. Another issue was the data contained schools located in all 50 states as well as U.S. Territories. At first these longitude and latitude plots seems like outliers, but they were in fact legitimate locations. For the purposes of this graphic, the U.S. Territories as well as Alaska and Hawaii were removed. Another challenge was getting the scale of data adjusted in a way the graphic was visually impactful. The overall CohortDefaultRate, which represents the student loan default rate for

the given school, ranged from 0% - 39.3% with a large grouping of schools in the 10-15% range. This caused the graphic to look heavy on one color range. To overcome this problem, a log of the default rate was performed within the graphic to better distribute the color ranges. The graphic shows the South-East region of the U.S. appear to have the highest density of default loan rates. The North-East appears to have the lowest default rates.

Summary of analysis

The analysis completed for each of our initial questions provided insight into what factors influenced an institution's default rate. Below is a summary:

Question 1- Influence of tuition cost and program length on default rates: Both tuition cost and program length are significantly negatively correlated with loan default rates. The lower the tuition cost and the shorter the program length, the higher the institution's default rate.

Question 2 - Correlations between school type and default rate: Public institutions have the highest default rates while Private institutions have the lowest default rates.

Question 3 - What geographic areas have the highest and lowest default rates: States in the North-East region of the U.S. have lower default rates, while states in the South-East region have higher default rates. This may be due to a higher cost of living and higher salaries in the U.S.'s North-East region.