**Data Programming in R -** 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? In addition to exploring the overall relationship between the cost of tuition on default rates, we sought to answer three sub-questions:

1. Did the program length influence the overall cost of attendance and what effect did the program length have on the overall default rates?
2. Did the type of university (public, private, proprietary) play a part in the default rates at institutions across America?
3. Are there any relationships between the geographic location of institutions and the default rate?

Data from the Department of Education and National Center for Education Statistics and Integrated Post-Secondary Educational Data System were joined into a single data frame and analyzed using R. Prior to our research, the general hypothesis was that higher tuition rates would lead to higher default rates. Quite surprisingly, our research provided findings of the opposite.

The Data

Post-secondary loan default data was found on the Department of Education’s Federal Student Aid site which included number of students in default, number in repayment, the school’s default rate, rate type, ethnic affiliation of the school, and school location (this will further be referred to as “PEPs” data). The specific dataset used was set 300 which coveredofficial 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 Integrated Postsecondary Education Data Statistics division of the National Center for Education Statistics (this will further be referred to as “IPEDS” data). This site hosted many datasets, and it was decided to only pull key items per school for years 2012 and 2013 including: geographic coordinates, attendance, cost, average grant money received per student, average loan amount received per student, percentage of students receiving loans. The data dictionaries for each dataset were also downloaded.

Use of the Data

The IES Policies are aligned with the U.S. Department of Education's Plan and Policy Development Guidance for Public Access, which was approved October 21, 2016. ED encourages researchers to maximize the data made available to the public.

Aligning the Data

The PEPs data was indexed by an OPEID which is the Department of Education’s identifier. The IPEDS data was indexed by UNIT\_IT. An intermediary dataset which contained both the UNIT\_ID and OPEID was located so that the two datasets could be merged.

Importing and combining the Data using R

The openlsx and dplyr libraries were loaded. The intermediary 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, and then merged into the full dataframe called “full\_df” using UNIT\_ID as the linking source.

Preparing the data

The data required some cleaning before analysis could begin. The column names were changed to more meaningful terms. The data set was reduced to include only complete cases, and the school type was changed to a factor for analysis. We chose 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”, which breaks the tuition into bins of $10,000 segments ranging from $0 - $50,000. In the PEPs dataset, the program length was simply a numeric value ranging from 3 to 12 which provided no insight to the user about the type of program. A new variable was generated for “Program Meaning” and classified each program as: Non-Degree(1 yr.), Non-Degree(2 yr.), Associate's Degree, Bachelor's Degree, Master's Degree, and Non-Degree(3 yr. +).

# Questions to Answer

# Did the type of university (public, private, proprietary) play a part in the default rates at institutions across America?

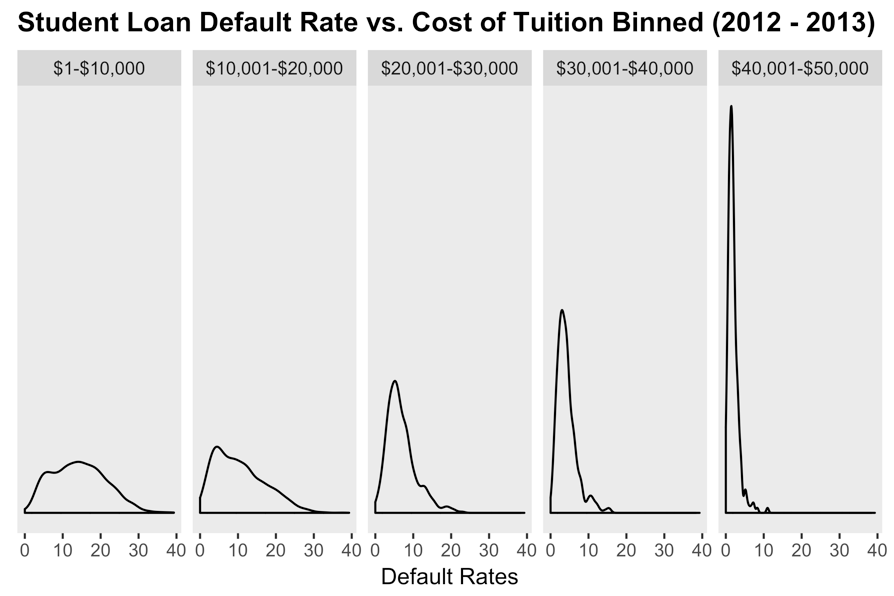
A summary table was generated to compare the cost of tuition against the length of program using the mean default rate as the measurement tool between the two variables. The summary table (*Appendix 1*) illustrates that shorter programs tend to have higher default rates. A rank-order function was applied to determine which program lengths consistently had the highest default rates. Non-Degree (1 yr), and Associate’s Degree’s had the highest default rates. On the contrary, Master’s Degrees and Non-Degree (3+ yr.) were consistently among the programs among the lowest default rates. The table can be interpreted in ascending order—the lower the number, the lower the default rate in rank order.

A heat map (*Appendix 2*) 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 as the relationship between the two variables shows a strong inverse correlation—as tuition increases, the default rates decrease.

To capture a visual representation of the percentage of students in default among the programs, a scatter-plot was produced (*Appendix 3*). The scatterplot measures the mean default percentage on the cost of tuition on a continuous scale. The findings of the scatter plot were consistent with the previous finding that the lower the tuition, the higher the default rates. The program with the lowest tuition, had the highest default rates—Associate’s Degrees. Conversely, Master’s Degrees which had much higher tuition showed lower default rates.

Through our analysis, it was evident that lower tuition amounts were indicative of higher default rates. However, it was unclear how strong that relationship was. To measure this, 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 (zero percent, up to forty percent) while 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. A higher frequency of default in any one area would appear as a wider and smoother curve on the density-plot (*Plot 1*).

*Plot 1: Frequency of default rates vs. Cost of Tuition (binned)*



## Summary of analysis

Although no statistical analysis was performed, we believe that there is a significant correlation between the cost of tuition and the chance of a student defaulting on those loans. From our analysis, we know that those higher default rates are associated with two particular programs: Non-degree (1 yr.) and Associate’s Degree’s. The curriculum offered by these degrees are often associated with lower paying jobs after graduation. There is most likely a correlation between the degree obtained, job qualifications post-graduation, salary, and ability to repay loans. Conversely, we know through our analysis that Master’s Degree’s and Non-Degree (3+ yr.) have a lower default rate. One could conclude that this is related to higher paying jobs post-graduation and a better ability to repay those loans.

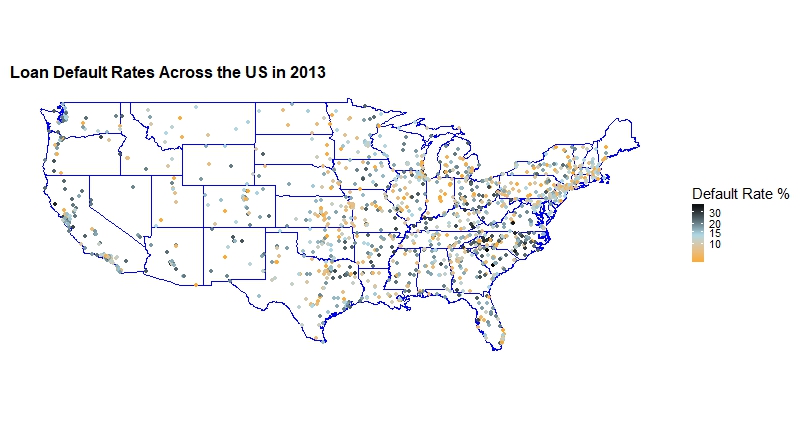
# Are there any relationships between the geographic location of institutions and the default rate?

## Summary of Analysis

To prepare the data for evaluation, a subset of data which included 7 variables out of the original 31 was selected from the full dataset. To gain a better understanding of the data, visual observations were made on the layout and types of data available for use.

It was decided the best way to answer the question around relationships of geographic location to student loan default rates, it was determined the best way to visualize the answer was to plot the data on a US map with state outlines, so we could see if 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:



## Obstacles Faced

There were several obstacles faced when creating this graphic. The most time-consuming struggle was understanding how to actually create a ggplot with the US map and state outlines. This was a new type of coding we hadn’t specifically explored, so with the help of Google, code was developed and a map was created. Another issue was the data contained schools located in all 50 states as well as US Territories such as Guam, US Virgin Islands and Puerto Rico. At first these longitude and latitude plots seems like outliers, but they were in fact legitimate locations. For the purposes of this graphic, the US Territories as well as Alaska and Hawaii were removed. Finally, a significant amount of time was spent 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, we performed a log of the default rate within the graphic to better distribute the color ranges.

## Conclusion of Analysis – Answering the Question

After visually looking at the US graphic, it shows the central states located on the east coast of the US seems to show the highest density of default loan rates. These states include North Carolina, South Carolina, Virginia, and Kentucky.

## Global Obstacles Faced

The group faced many challenges during the project, the first of which was locating data. We agreed on using school data for our project, but when our initial plans to use our group members’ employers’ data didn’t pan out we took to the internet. The default loan data was located, and it was well-organized and didn’t appear to need much cleaning. It didn’t include much demographic or school data, but it did include the OPEID. Further research ensued, and data was located on the IPEDS site regarding the schools which was indexed by UNIT\_ID. Fortunately, a table was found that contained both the OPEID and UNIT\_ID so that the two data sources could be merged. Our next challenge came when we discovered that the OPE\_ID was not set up the same way on the datasets. In one there were leading 0’s, and in the other there were trailing 0’s. Once those were identified and transformed in R, the data could be merged.

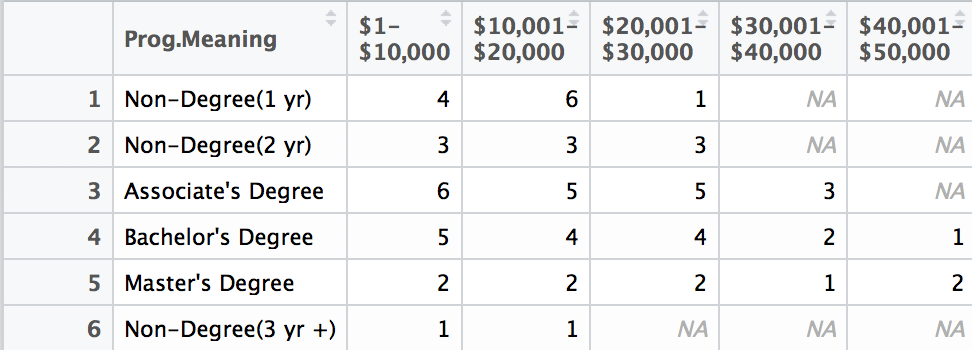
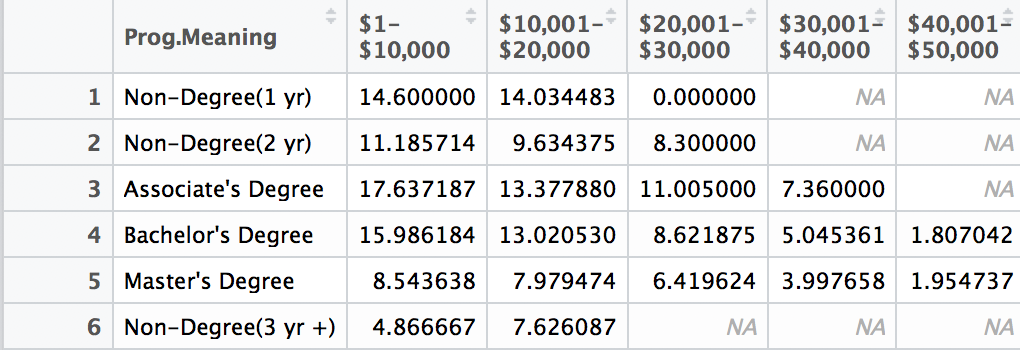
When analyzation of the data began, it became apparent that there was not much variation in the default rates. We realized that the specific dataset we had imported was a subset of the full dataset and only included schools with a default rate of <15%. We then downloaded the full dataset and changed our import commands in R to pull in the full default dataset.

Communication proved to be another challenge. We initially set up a group on ICON with folders for documents, scripts, and datasets. We quickly found that we weren’t getting updated when additional 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.

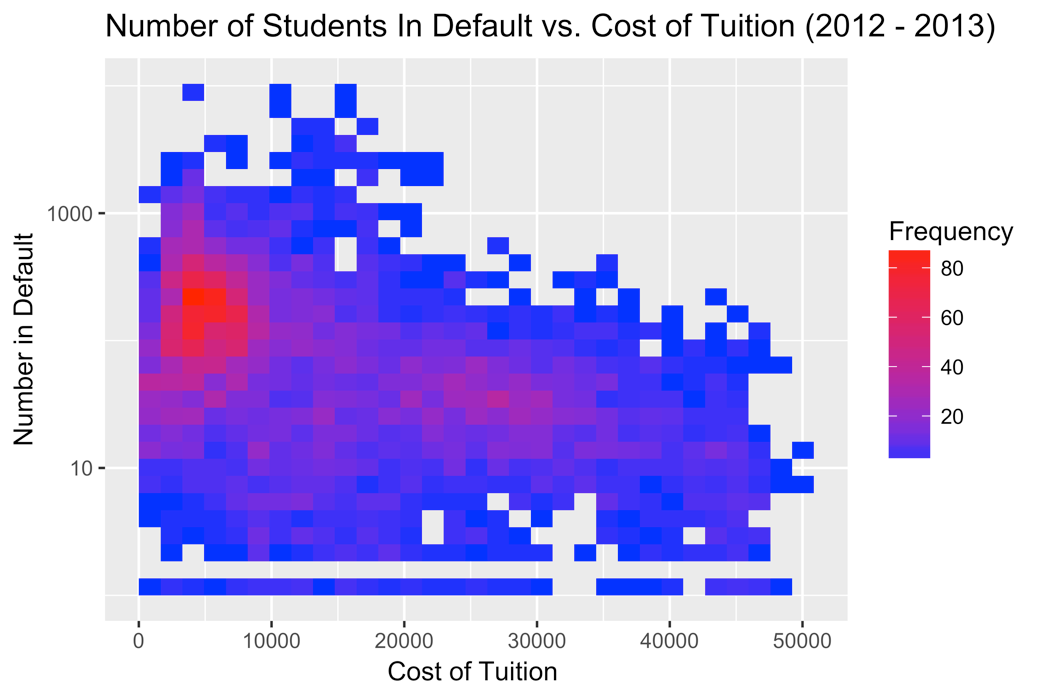
## Conclusion of the Analysis

Appendix

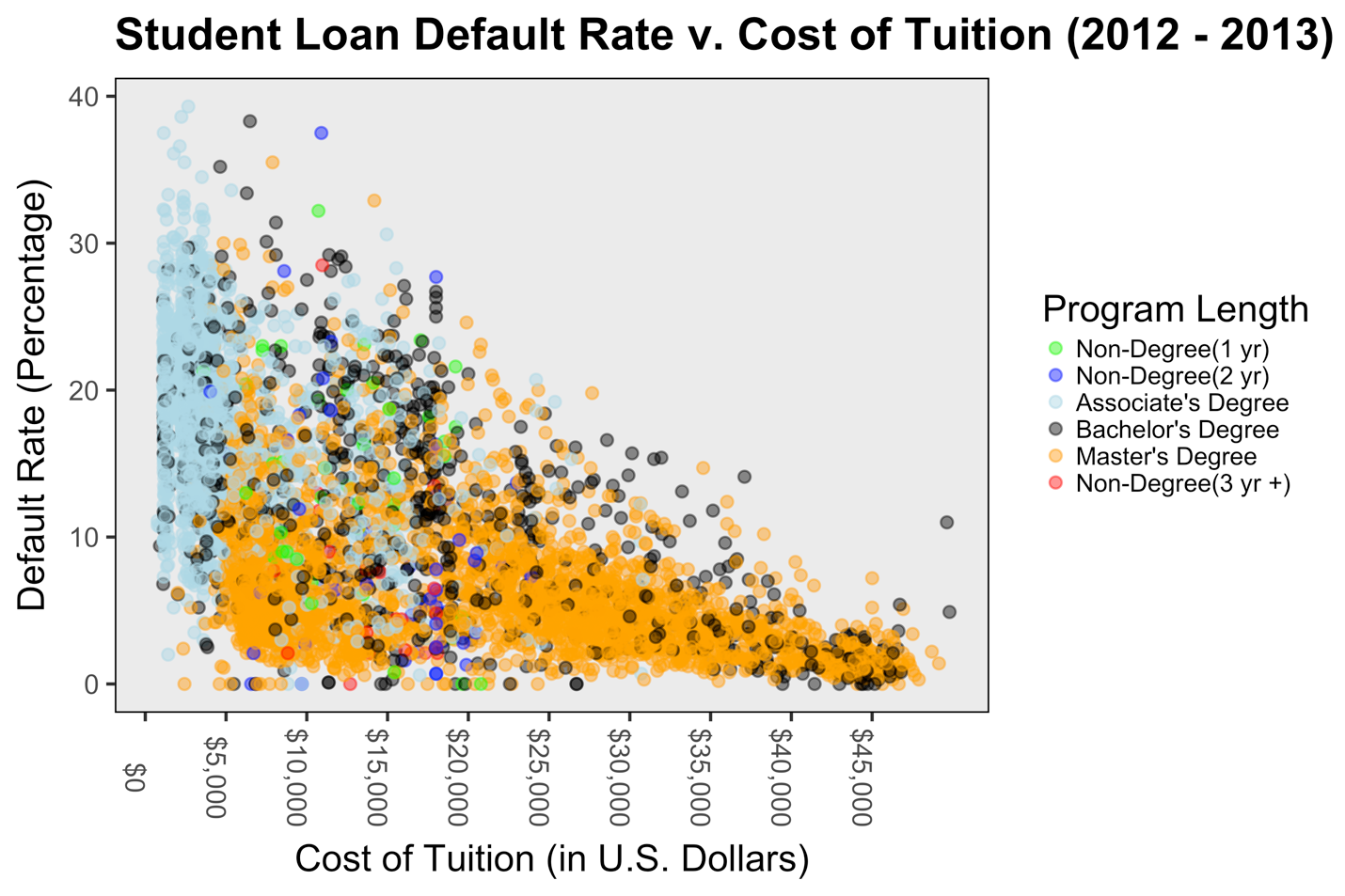
1. Summary Table: Program Meaning vs. Binned Tuition Price



2) Heat Map: Number of Students in Default Measured Against Cost of Tuition



3) Scatterplot: Mean Default Rate on Cost of Tuition by Program Length



First Professional Degree🡪 6 observations Collapsed into the Master’s Degree

Two-Year Transfer🡪 2 observations Collapsed into Bachelor’s

Non-degree <1 yr🡪 2 observations Collapsed into Non degree

Reduced summary tables into 6 objects from 9