**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? We hoped to gain insight into how default rates for institutions change over time, which geographic areas have the highest default rates, potential correlations between school type and default rate, and the influence of cost of attendance on default rates. Data from the Department of Education, National Center for Education Statistics and Integrated Post-Secondary Educational Data System were joined into a single data frame and analyzed using R. A statistical analysis was completed on variables including state, school type, program length, attendance, and cost of attendance, among others to answer these questions.

The Data

The team agreed that we wanted to study school student loan default data, and data was located on the Department of Education’s Federal Student Aid site which contained number of students in default, number in repayment, the school’s default rate, rate type, ethnic affiliation of the school, and school location. We strengthened our dataset with data from the Integrated Postsecondary Education Data Statistics division of the National Center for Education Statistics. We located additional data on the schools including geographic coordinates, attendance, cost, average grant money received per student, average loan amount received per student, and the percentage of students receiving loans.

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 loan default 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 was 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 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.

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 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.

Summary of findings