**Introduction**

One of the biggest ways people measure the quality of a university is how well the students do after attending the institution. This wellbeing is often equated to money in the forms of earnings and indebtedness. In other words, the question at the front of everyone’s mind is: Was it all worth it? Is it a good investment to go to this school? Will I drown in debt after leaving school? To answer this question, it makes sense to look at student’s earnings and indebtedness.

**Methodology**

The publicly available datasets that could help answer these questions were the College Scorecard’s API (from the U.S. Department of Education) and the Post-Secondary Employment Outcomes (PSEO) database (from the U.S. Census Bureau). The College Scorecard’s data was segmented by male and female as well as the year of repayment (1,3,5, or 7) and was from the year 2020-21. The PSEO data was segmented by cohorts (e.g. 2016-2018), percentile of earnings (25th, 50th, or 75th), major (officially the cip-code), year after graduate, and degree level. The PSEO dataset is limited to public institutions and does not have institutions from significant states like California, Illinois, and Florida. For this study, the year 5 of indebtedness was used along with the 2016-2018 cohort so that the year of debt repayment would fall within that cohort.

The technique employed to analyze the dataset was [k-means clustering](https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1). The goal of the technique is to segment the data into groups whose observations are the most related (based off their characteristics) so that they can be analyzed further. This was accomplished by using R’s native **kmeans()** function and integrating into Tableau using the its analytics extension. In order to ensure that the answers were replicable, the command **set.seed()** was used so that the results would stay consistent. The attributes used for the clustering were the debt repayment and earnings. In order to do clusters for the majors Computer Science and Engineering, the earnings percentiles, and the different genders for debt, 12 clusters were created.

Because of the limitations of joining the datasets, a one-to-one comparison was not possible, i.e., the debt rates do not apply to the majors or the percentiles, and likewise the earnings do not apply to men and women. It only makes sense to compare the institutions to one another, not the groups within the institution. A useful analogy is a study comparing people’s hair color and eye color. A clustering analysis may show that there are groups that have similar variations of these characteristics, but it does not have anything to say about the relationship between those characteristics. Dataset is limited in that it does not show the genetic relationship between these traits, it just shows where they are appearing together.

**Results**

**Definitions of the variables:**

The number of clusters refers to the amount of times the institution ended up in the same cluster as Georgia Tech. There were 6 different clusters for each major, and 12 when the majors were looked at together.

The Earn Diff Avg refers to the average difference in earnings that the institution had from the cluster mean, which can be thought of as how below/above average it was.

The Rate Diff Avg likewise refers to the average difference in debt repayment from the cluster mean and can be interpreted similarly to Earn Diff Avg.

**Tables**

A screenshot of a graph

Description automatically generated with low confidence

**With us (all):**

This shows the clustering data for both the engineering and CS clusters, which shows which schools overall were the most similar to Georgia Tech. Most of the school’s rates were relatively close to (or just below) the cluster averages, with the exception of the Colorado School of Mines and UVA. This also applied to earnings, where most schools were just below the average with UVA being a significant exception.

**With us (eng):**

A screenshot of a graph

Description automatically generated with low confidence

This shows the clustering data for engineering clusters, which shows the schools that were most similar to Georgia Tech. Most of the schools were within a 5.5% difference from the cluster average, with the main exception being the Colorado School of Mines. UVA was again above average with an average difference of +5.36%. Most of the were within roughly $2k of the cluster average, with the exceptions being UVA with its earnings being $4.32 above the average, and Missouri University of Science and Technology at $2.98k below the average.

**With us (CS):**

A screenshot of a graph

Description automatically generated with low confidence

This table shows the clustering data for the CS clusters, which shows the schools that were most similar to Georgia Tech. A little over half of the schools are within 4% of the cluster average, with the notable schools that are outside of this being UMass Amherst (falling short by 6.03%), Colorado School of Mines (exceeding by 6.33%), and Binghamton (falling short by 5.55%) (Oregon State does not fit this mainly because it is only in cluster). With earnings, it is notable that UVA is on average $16.88k above the cluster average, which may help in explaining why most of schools are at least $5k below the cluster average (when considering that UVA was in every cluster).

A screenshot of a graph

Description automatically generated with low confidence

**Cluster Tables**

A picture containing screenshot, text, colorfulness

Description automatically generated

This table shows specific statistics about each cluster that was used to analyze the data. It can be seen that earnings are linked to the percentile they are in and that there is some variation in the rates, but this is not super significant to the results overall. The purpose of this is to make the method of this much more clear and give context to all of this talk about clusters.

**Elbow Chart:**

This chart shows the within-cluster-sum-of-squares for each cluster which shows in essence shows close the cluster overall is to its center. As the number of clusters increases, the WCSS will decrease. This technique involves finding the point at which the WCSS stops significantly decreasing and choosing that point as the number of clusters. There is some flexibility in this approach in that in this chart because there is not much difference in the measure between using 7 or 8 clusters, therefore either will suffice.

For this study, the goal was to have Georgia Tech’s numbers as close to the sample average as possible (which was difficult for CS), therefore cluster numbers were chosen in accordance with that goal. By doing this, it is possible to get the most representative and similar schools to   
Georgia Tech (allowing in schools that are below and above it but still near it).

**Discussion**

**The “Public Ivy’s”**

Many of the schools that ended in Georgia Tech’s clusters were highly regarded public institutions who have particularly strong STEM programs and are often nationally regarded for them. This makes sense given the limitations of the data set, but there is still a lot of interesting things to discover by delving deeper into these institutions.

Two of the institutions most commonly in Georgia Tech’s clusters (UVA and William & Mary) were in Richard Moll’s original 1985 list of “public Ivy’s” (from his book *Public Ivies: A Guide to America's Best Public Undergraduate Colleges and Universities*). The list is by no means an official measure, but it is worth pointing out that it does identify a group of institutions that have many characteristics in common. The most important similarities are that they are among the United States’ well regarded and academically rigorous public institutions, and in many cases their programs are able to compete (sometimes outcompete) institutions from the Ivy League, which are often thought to be the best academic institutions in the US. Given how well Georgia Tech’s programs perform in rankings, it seems reasonable to start comparing it to schools whose programs are regarded in a similar sense.

When looking at the rankings from Niche, the average ranking for the schools in our Top 10 was #34 among best public institutions, with 4 in the top 20 and 1 in the top 5 (Georgia Tech is ranked #3). According to the site, Georgia Tech’s Engineering and Computer Science programs are ranked the 6th best in the nation each, which outranks all of the schools in the Ivy League except Princeton and outranks all of the “Public Ivies” in the 1985 and 2001 lists. Other sites that rank universities are likely to differ to some degree, but the value of using these lists is that it shows what other schools are in the same “neighborhood” as Georgia Tech (not unlike the clustering method).

**Applications**

The dashboard itself has some practical students for students interested in many different majors that are interested in researching the earnings of schools as well as the debt repayment rate. In other words, one can investigate financial of large groups of students for many public institutions. Here is an example use case:

There is a student interested in pursuing a degree in CS that is also a woman. Based off of the dashboard, she can see that students from a certain cohort in a certain major (it also can include all majors) earned x amount at the 25th, 50th, or 75th percentile, and that x percent of all the women that graduated roughly within that cohort were able to pay (or continue paying) their debt back. It can also be interpreted as the probability that the user will pay their debt back based off that cohort.

At an institutional level, it is also useful in identifying institutions that can be considered as peers. Georgia Tech, which is a public university world renowned for its Computer Science and many engineering programs (among others), seems to have many attributes in common with institutions like the Colorado School of Mines and UVA, which have the characteristics described above.

**Pros & Cons**

**Restrictions**

The main limitations of the data set are that one can’t do one-to-one comparisons of debt and earnings, the dataset is significantly limited in terms of which schools it has, and there are data missing for certain years.

Since the debt and earnings measures cannot be exactly matched (e.g., there is no repayment rate for statistic for the 25th percentile of computer science graduates) it limits which analytical techniques one can use. Had there been this one-to-one match, prediction using methods like linear regression, decision trees, or LASSO would have been possible. Additionally, since the earnings cohorts encompass three years, the debt rates are at best an approximate match (i.e., they fall within the cohort).

This dataset is limited because it does not have data for public institutions from several US states, including but not limited to California, Illinois, and Florida. It is common knowledge that many of the most prestigious public institutions in the United States reside there. Adding these schools would give a much more holistic set of clusters and could point to more institutions that Georgia Tech is related to.

Another limitation is that data is not available for certain years from certain schools. Some of the most pertinent examples of this (for Georgia Tech) are there not being data for U Wisconsin, U Michigan, and UT Austin for the 2016-2018 cohort. All these schools are in Georgia Tech’s peer group, and when the cohort year changed, all of these schools ended up in Georgia Tech’s cluster x times (still need do this study).

**Benefits**

The main benefits of this dashboard and study are that they allow an institution or a person to find schools with similar trait. Using the clustering method allows the user to circumvent needing a one to one comparison between repayment rates and earnings.

**Improvements/Future Plans**

**One to one**

Ways to improve this study in the future are to acquire data that allow one-to-one comparisons of earnings and debt repayment rates. This would allow one to get a deeper insight into the financial health of individual groups of students and to help predict the financial health of them.

**Employment**

It would also be interesting to look at the schools in the clusters whose earnings and debt were similar and see what kinds of jobs/companies they end up working in. This could provide more insight into how related these groups (from this study) are what significance these clusters have. The underlying question behind a study like this would be: how similar are the ways that these groups of students are earning money and maintaining their financial health?

**Broader Cluster**

Another approach that utilizes the same data is to do a cluster of the variables that were present in it but instead combine them. For example, rather than do three different cluster test for each percentile, one could combine them into one data frame and then cluster based off of that. This would allow a broader view of the institutions and could reveal relationships that were not found in this study. The main downside of this approach is that it may more computationally expensive and will take longer to obtain results.

**Federal Data**

Another approach is to use publicly available data to supplement the current data set. These data include but are not limited to unemployment data and local economic growth data. The purpose of including these data would be to see the effect that the local environment (i.e., local economic conditions) have on the outcomes of students at universities in the area. Another approach is to link the institutions to their data in IPEDS to use other variables to create clusters of institutions.

**Implications**

What this study implies is that Georgia Tech’s students are having relatively similar financial outcomes to the students from these institutions. It is possible that Georgia Tech’s graduates are competing for jobs with students from these similar institutions, therefore it would serve the institution to understand these programs.