Analysis of post graduate outcomes in terms of monetary success 10 years post graduation

By Alexander Milberger
SI 370

I. Motivation:

As a student who wishes to achieve success in life, I wanted to analyze the monetary outcomes students from various institutions. I defined success by using the weighted average of earnings 10 years after graduation. Many students wish to find success in life, whether that be through family, their professional life, or monetary earnings. While it seems petty to define success in terms of weighted average of earnings 10 years after graduation, from my personal experience, especially in a capitalist society, people who make/have more money are considered more successful. To find schools that foster the capability to give students high earning potential, I needed to isolate only the relevant variables that are the most correlated with the weighted average of earnings after 10 years.

Research Questions:

1. What factors play the largest role in determining post graduate success?
2. Does the type of institution play a role in in post graduate outcomes?
3. Are there any evident clusters in the data? And how do they relate to post graduate success?
4. Can outcomes be predicted using correlation of principle components or MDS dimensions?

II. Data Source:

For this project, I chose the dataset published by College Score Card. The data set contains over 7000 rows and 1700 columns. The most recent college scorecard data was downloaded as a csv file. The only changes that were made prior to analysis was the removal and replacement of Null and Missing values with 0s. The data that was replaced with 0s was data that that had the value Privacy Suppressed, or a NaN . Also, the addition of several weighted variables was added to my dataset. The weighted variables that were added describe the weighted average earnings of students 10 years after graduation, the weighted income of dependent family members, and the weighted income of independents. Because of the way I defined success in my research, the main variable of interest was relating to the weighted average earnings after 10 years of graduation. In my analysis, I only used institutions, which were predominantly bachelor awarding, main campuses, and not for profit.

III. Research Questions and Methods:

Initial data cleaning was done in an iPython Notebook. The Pandas and Matplotlib libraries were heavily used for this process. Nothing was interesting after the initial data cleansing. After cleansing, correlation plots relating the newly added weighted variables against each other. From this I determined the variable related to an independents income was not very correlated to the variables of interest, and concluded that it would be mostly useless in my analysis.

Q1: What factors play the largest role in defining post graduate success?

To get an initial Idea of what variables I could use, that would relate the most to post graduate success, I ran multiple correlation matrix. To get an idea of what type of correlation to use, I plotted the distribution of my variable of interest, and It wasn't normal. Because of this, I decided to use spearman's method of correlation to perform my correlation analysis on my dataset. Through running multiple correlation tests I could reduce the number of variables in my dataset from over 1700 columns, then to 25 of the most correlated with the weighted post graduate outcomes. After this I identified 4 of the variables that were the most related for further analysis. After determining the variables, I was most interested in, I plotted the distribution off all relevant variables, which were all skewed left, and not

normally distributed. To get the most accurate data, I ended up using the smaller data frame with the 4 most correlated variables to the weighted average earnings after 10 years. The 4 variables I isolated were already normalized/ adjusted to use in analysis and produced similar results to a normalized version of the dataset.

Q2: Does the type of institution play a role in in post graduate outcomes?

To address this, I plotted the distribution of weighted earnings after 10 years of graduation for public institutions and private institutions. Although the analysis was not very comprehensive, the results proved to be interesting nonetheless.

Q3: Are there any evident clusters in the data? And how do they relate to post graduate success?

To answer this question, I employed the use of k-means clustering. Through using the silhouette method, and incorporating graphical visualization, I was able determine the best number of clusters to be 3. Initially I employed the clustering and k means analysis on my 25 most important variables, and then on data frame with only the four most relevant variables. The results of these proved to be very similar, reaffirming that variables "D100_4, D150_4, and dependents family income", were the most related to post graduate earnings after 10 years.

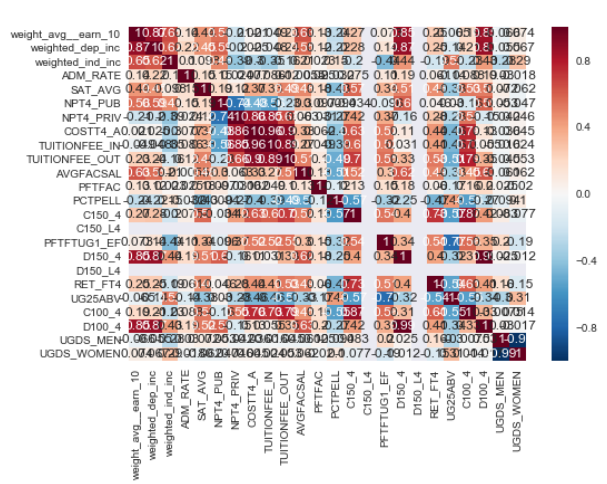
Q4: Can outcomes be predicted using correlation of principle components or MDS dimensions?

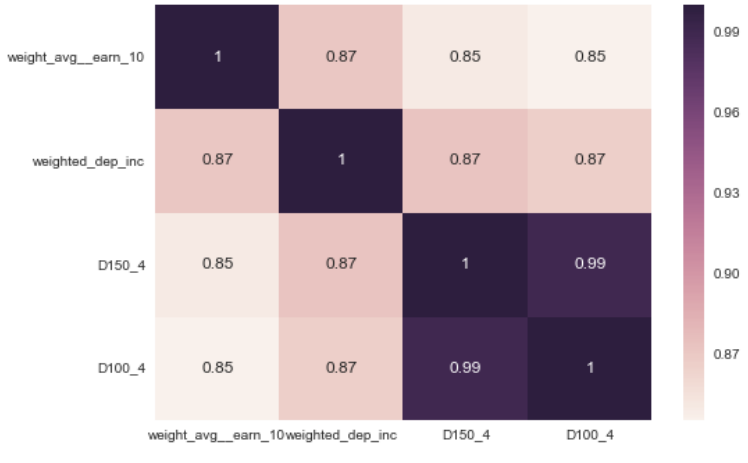
To determine this I ran MDS with 2 dimensions on my larger set of relevant variables, and on my small set of the most relevant variables. Because the results were similar, I choose the data collected from the data frame with only the four most relevant variables. The results between for my MDS analysis on my normalized data frame of the four most relevant variables and on the non-normalized data frame were very similar. This reaffirms the notion that the four most relevant variables already took normalization into account. I also ran PCA analysis with 2 principle components on the same dataset to see/reaffirm which schools were similar.

Beyond this, I also incorporated K means clustering with MDS analysis to visualize the relation of the different clusters.

IV Analysis

Q1: Out of the 1700 plus rows I could find 25 variables that had moderate to strong correlations to my variables of interest. From this I could identify the 4 most important factors that relate the most to postgraduate outcomes.



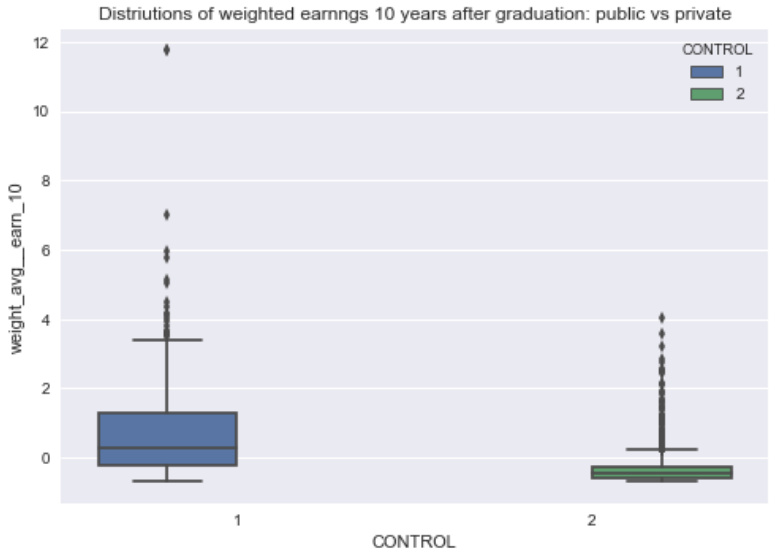


The first correlation matrix shows the 25 most related variables to post graduate outcomes. These variables detail values related to: Admission rate, Sat average score, Price of the institution, Tuition in and out of state, percentage of faculty to students, percentage of students with Pell grants, Completion rates of 100% and 150%, along with their adjusted values, among a few others.

The second correlation matrix shows the variables that were the most correlated to average earnings 10 years after graduation. The variables in this matrix all have a correlation of 0.85 or greater, which means they are thoroughly related as to how I defined post graduate success. These variables detail the weighted average earnings 10 years after graduation, the weighted income of the families of dependent students, and the adjusted completion of 100% and 150% within 4 years.

This shows that these are the best indicators of a student's post graduate outcomes, and were used to address my remaining research questions.

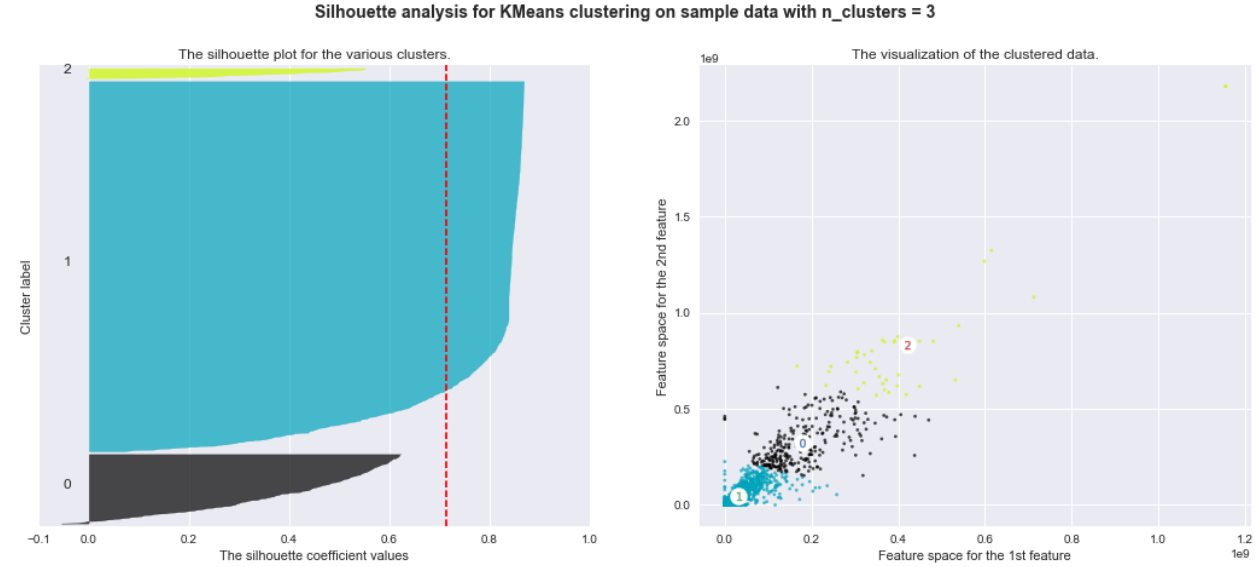
Q2: I found that the type of institution is related to post graduate outcomes, it appears as if public institutions make more money on average, while private institutions do not. I think this can be attributed to the fact that there are a lot of liberal arts colleges that are private compared to the likes of private ivy league institutions.



Shows that students from public institutions have a larger distribution of earnings, but larger potential to earn more

Q3: Through the silhouette method of calculating k means, I identified that there were 3 main clusters in my data. The image below is the analysis using my smaller data set of the four most relevant variables.

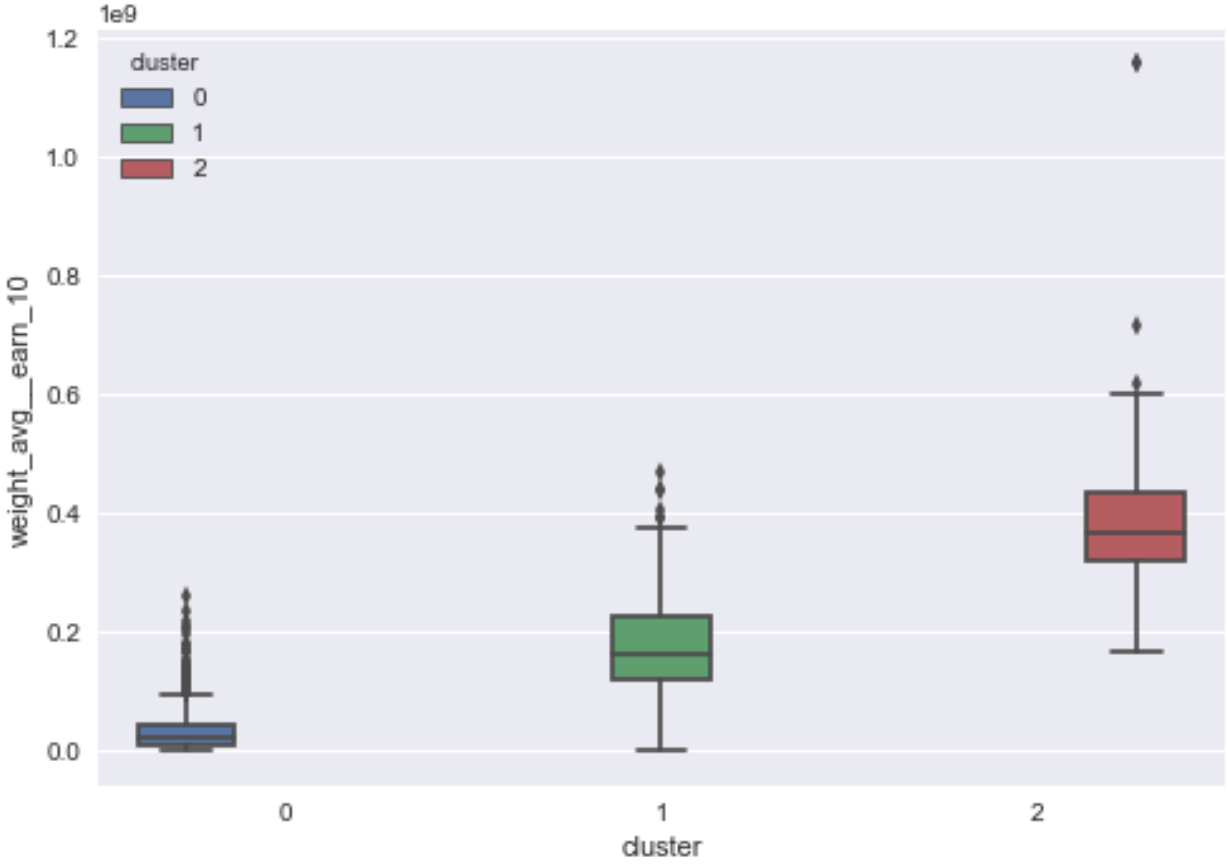
Through this analysis we can see three clear clusters within the data. The visualization of clustered data shows that the clusters are indeed related to each other, based my list of the most relevant variables. Because clustering with the larger dataset and the smaller data set of relevant variables produced very similar clustering results, I decided to focus on the four most relevant variables that were contained in the smaller data. This next image shows the silhouette analysis with the larger set of 25 variables deemed relevant though my analysis. The average silhouette score is for the smaller dataset containing only the most relevant variables is 0.713498703114., The average silhouette score for the larger dataset of 25 relevant variables is 0.701237844108.



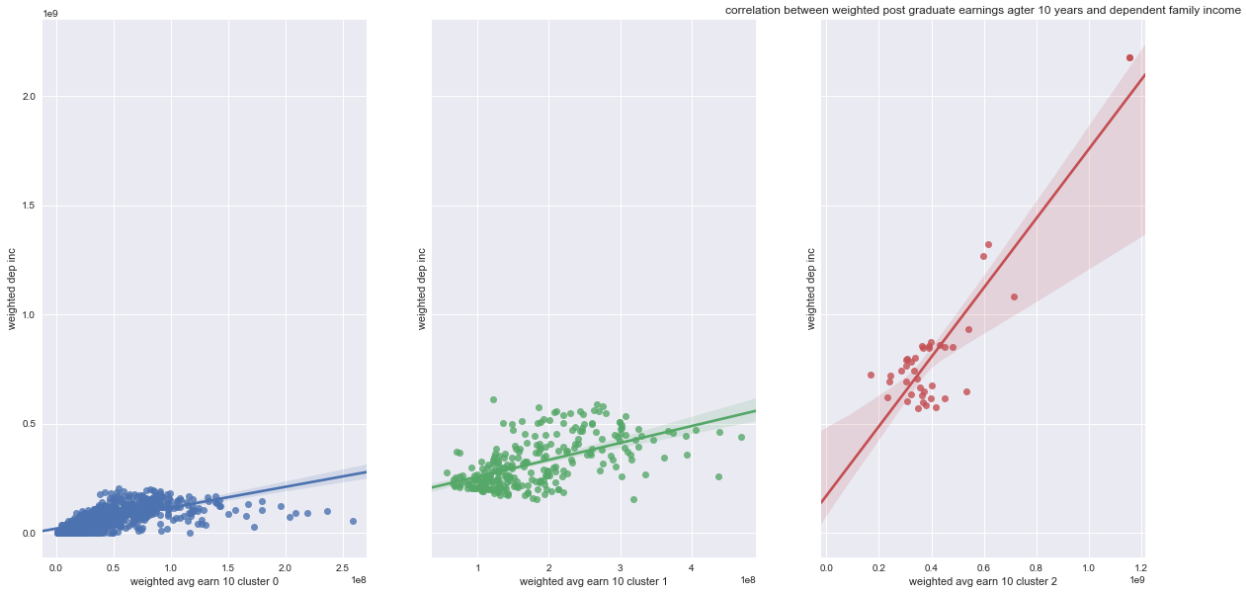
From these images, we can see that both datasets of relevant variables, big, and small, produce similar clustering results.

A distribution analysis of the different clusters shows the average earnings 10 years after graduation. From this we can see institutions in cluster 0 have the least amount of earnings, which is followed by

cluster 1 with a slightly higher potential for earnings, and cluster 2, which shows to have the highest earning potential.



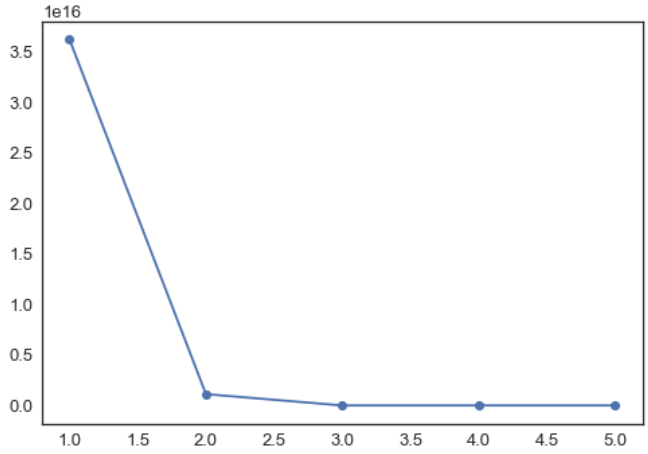
To further determine how correlated each cluster is to my benchmark of post graduate, I ran correlation analysis between weighted family dependent income and weighted earnings 10 years after graduation.



R squared for cluster 0: 0.601483061217 , R squared for cluster 1: 0.334058460503 , R squared for cluster 2: 0.182778510036

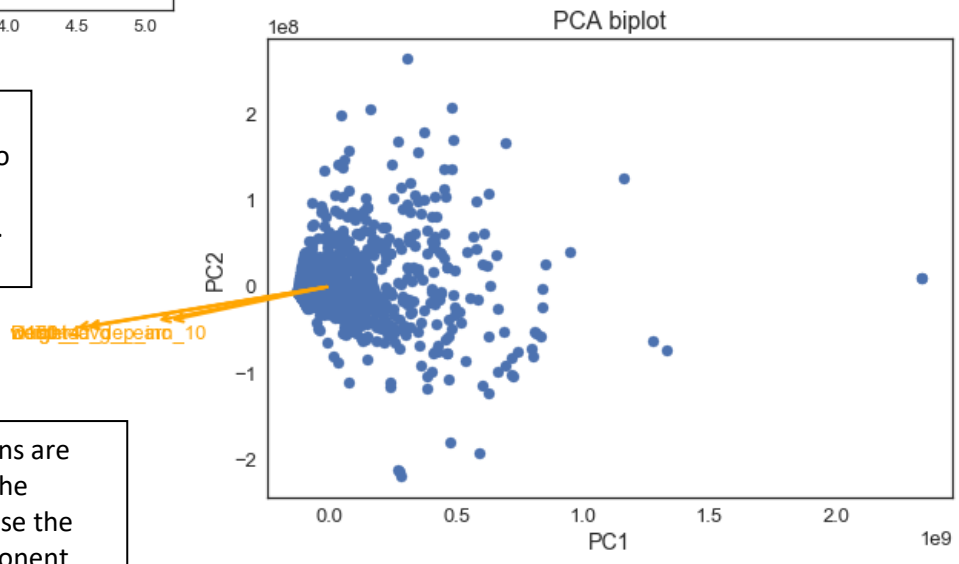
These results tell us that dependents' family income is more associated with earning potential for institutions with lesser earnings, as shown in cluster 0. As the schools earning potential rises, the less of an influence a dependent's family income has on a student's ability earn becomes. From this I can deduce, that the other most relevant variables, such as adjusted completion rate, play large roles in an institution's capability to produce high earning professionals.

Q4: I found that with both my larger and small set of relevant variables can be explained with only two dimensions. This means that potential earnings can be predicted within each cluster, based on which variables have the highest correlations to earnings. First through principle component analysis I show that all my related variables correlate highly with each other through two principle components. This also helps to highlight similar schools, with similar earning potentials.



Through the elbow plot to the **left**, I can determine that variation between institutions can be explained though two principle components

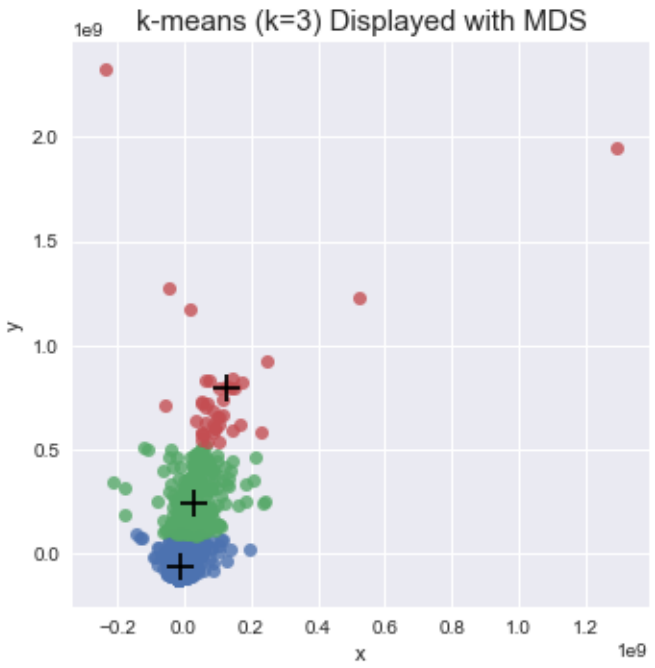
The PCA plot to the **right** reaffirms the relation of similar institutions to the most relevant variables that were preciously determined by me.

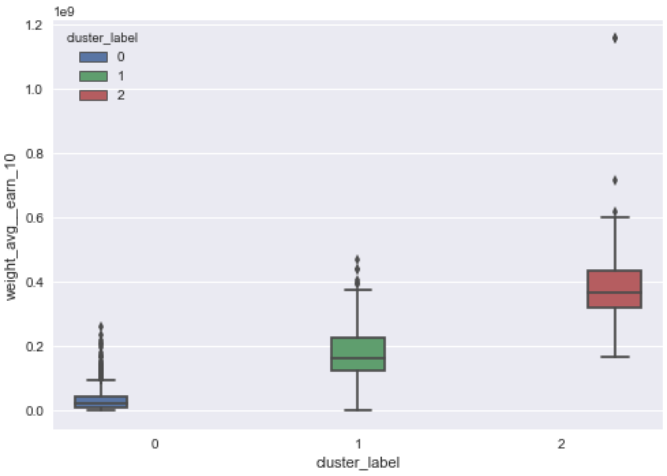


Although we can see what institutions are similar, it's hard to determine how the institutions are similar. This is because the variables used in the principle component analysis are closely related.

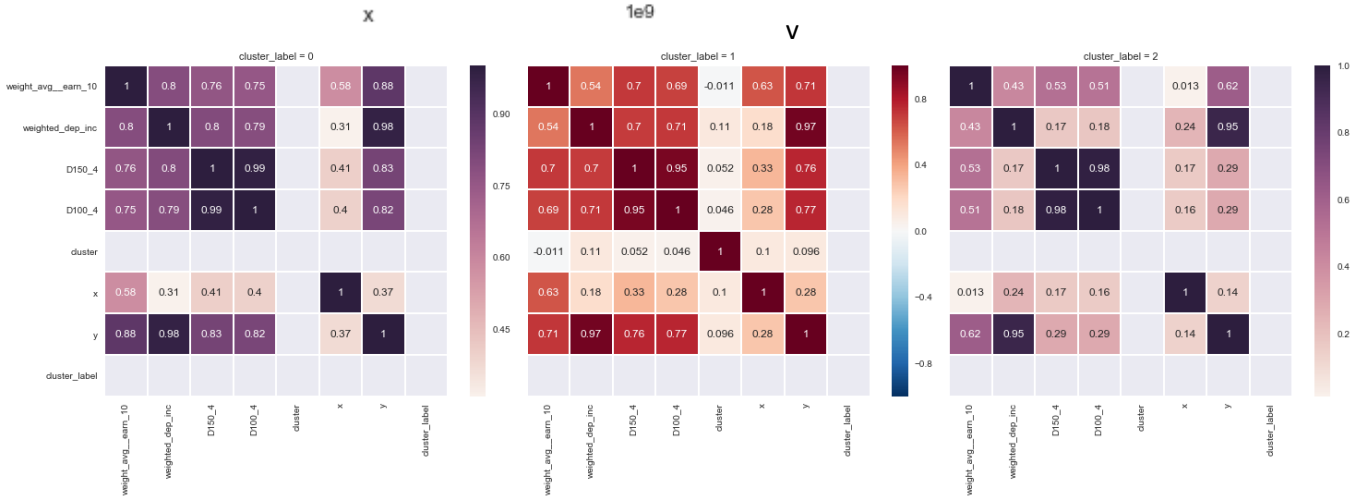
To get a better Idea as to how each institution relates, I employed MDS with 2 dimensions and clustering to visualize my clusters and distributions. Through these methods I correlated clusters and earning potentials to the most influential variables.

Figure M: We see similar results from my previous clustering analysis, as well as similar distributions of potential earnings in each of the clusters





Using the different clusters after running correlation analysis, I depict how the variables in each cluster relate the most to potential earnings, as well as the dimensions X and Y, that were created through MDS



From the correlations of each of the clusters, we can see earnings in cluster 0 are moderately correlated with the Y dimension (in Figure M) , meaning the Hight of a cluster 0 institution on the Y axis is a good indicator of earning potential for cluster 0 schools. As the earning potential of students rises between the clusters, there is less correlation to the x and y dimensions. This can mean that potential earnings in cluster 1 and 2 schools are more reliant on completion rates than on dependents family income, as is with the cluster 0. I repeated the correlation, clustering, and MDS analysis on the data that was normalized, only to get results that were very similar. This shows that the variables used in analysis accounted for skewedness through weighting, and creation of adjusted variables like *D100_4*.

V. Future Directives / next steps:

Considering the almost limitless actions you can take with a dataset of the magnitude that is college scorecard, my next step would be to train a classifier to identify schools who allow for higher earnings. In addition to this, the evaluation of nominal values with other relevant values could be useful to categorizing what nominal factors relates to an institutions capability to give students large or little earning potential.