College Graduate Success Factors

Jayme Green

1. BACKGROUND

The US Department of Education's College Scorecard emerged in 2013 after former-President Obama commissioned it to hold colleges accountable as well as allow the US population to see where they will "get the most bang [money, success] for your educational buck".[5] The College Scorecard holds a lot of information about colleges and the students that go to those colleges. This data includes student demographics, student economic backgrounds, past students current median incomes, and numerous other data. This vast amount of data, collected by a reputable source, allows for numerous analysis to occur.

With the data public, announced and prepared by the federal government, and about the hot-topic of college pricing, numerous different mediums are using this data and doing analysis. These include the three main people: news companies, educational universities, and individuals. All have done analysis on this data, but with so much data provided, many do not fully copy each other in their analysis.

The news companies focused mainly on what a person can look up as well as focusing on the price. For example, the National Public Radio (NPR) published an article, "The New College Scorecard: NPR Does Some Math", reports on the best colleges according to cost of attendance, 4-year graduation rate, and 10-year median income of graduates.[2] It seems that this information was mostly generated through querying and using a formula. The three main attributes were looked up for every college and a score was calculated for each college based on a formula using these statistics. The data was sorted on this formed score. While crude and easy, it provides basic analysis on some of the better colleges in the United States. Many news sources did similar, easy analysis using the College Scorecard.

Universities also looked at this data. While they were motivated by providing educational analysis, they also have a conflicted agenda; they want their institution to look the best in their analysis. With this known bias, the universities' studies have to be questioned on integrity. For example, a vocational school analyzed the data to determine that numerous graduates are getting low incomes that correlate with jobs that only require a high school diploma. It uses this analysis to emphasize that there will always need car mechanics and that they (a car mechanics school) provide skills more important that most colleges.[1] While this may be one of the most egregious bias, it cannot be expected that every other college does not have the same motivation. Therefore, college research (which is usually one of the most trusted) cannot be trusted fully for non-bias results.

Finally, individuals - like myself - have also done analysis based on the College Scorecard. A community at Kaggle have been exploring this data for years and have discovered interesting information. While none are exactly my topic, many are close that I can base my research on. For example, Michael L. Thompson explored the potential earnings of an individual depending on their background and future choices (assuming they continue with their decisions) with different college's median income after graduation.[7] He looked at typed of college (private/public, for-profit/not-for-profit, etc) as well as average SAT scores which I might look into for my research. Additionally, he used a naive-Bayes probability to determine how good an attribute can help to determine this value proposition. Similarly, I can use this thought and use naive-Bayes to determine how much each of my attributes influence median income of graduates. I will reference Thompson's papers, as well as other individuals, to get further ideas on how to further my progress.

Numerous people have done analysis on the US Department of Education's College Scorecard which I can reference to see their progression and results. Yet, my analysis seems to be more unique which will bring more insight in this data.

2. OVERVIEW

For the project, I decided to switch from 2017 US Primary data to the US Department of Education's College Scorecard. [3] The 2017 US Primary data was interesting, but not enough to motivate me. Some many people have done analysis on this data - and published it - that I have seem hundreds of these studies. As a result, I did not think I would be discovering anything and just regurgitating the same study. After seeing a graph on r/dataisbeautiful on Reddit using the College Scorecard, it intrigued me and made me look up this data.[6] If I was looking up the data in my ever-shrinking free time at the end of the semester, it would be much more interesting to analyze than my previous choice. As a result, I switched my dataset to the US Department of Education's College Scorecard.

Using the US Department of Education's College Scorecard, I have decide to look into the main factors that lead to higher income of individuals 10 years after entry of college. Some of the factors I will look into will include demographics including race and gender statistics of the college. Additionally, economic background will also be looked at like the poverty rate of the college as well as the median household income of the college goers. Associated with this data emerges in the median income at 10 years after beginning college. This statistic will be the main target variable as I want to discover the biggest factors that influence it.

Additionally, I may look into the differences between male and female income after college. This statistic was also measured by the College Scorecard for every college. Although it would to expected for the factors to be the same, differences may occur and bring about interesting information about the different factors that lead to success in the different genders.

In the beginning of the exploration, I focus on discovering basic information about the data. This mainly involves clustering techniques such as agglomerative and K-Means. As this basic information reveals, the study will continue and will use more data mining techniques to discover additional information. At the end of the study, I will be able to determine the most contributing factors to success of college graduates.

3. DATA CLEANING AND PRE-PROCESSING

The data that I received was huge. The US Department of Education's College Scorecard has data from all colleges throughout the years including college statistics and student statistics.[3] While this data allows for a lot of analysis, it is too much for this project. As a result, I pre-determined what data I was interested in.

The US Department of Education provided data documentation that describes the data.[4] In this document, it provides the column name, what the column is describing, and what file that column resides. Using this, I decided on a handful of useful attributes for this study. Therefore, I reduced the number of attributes from 100+ to 16 columns. These columns had demographic information on students, wealth of the students before college, and the median income of the students after 10 years after entry of college. With these rows, I would be able to start understanding what the biggest factors leading to high income would be with those results influencing what attributes I should focus on as well as if I want to incorporate different factors.

These 16 columns contained over 5,700 amount of values with each row representing a different college. Unfortunately, some colleges chose not to respond (a NULL) or chose to not disclose the information (Privacy Suppressed). Wanting to cluster this data using agglomerative or K-Means, I needed all data to be present for the row in order to accurately track the center of mass. With this need, I removed all rows that did not hold all of the attributes. As this was the first stage analysis and understanding, I do not have enough domain and data knowledge to make accurate predictions of these missing values. After this removal of rows, I was left with approximately 3,500 rows. While a large amount of colleges were removed, a sample size of 3,500 should provide accurate information. Although, this removal could cause an unknown bias that I will continue to question throughout my study.

This basic cleaning provided suitable data for my beginning research. As I continue, my data cleaning and pre-processing will most likely continue.

4. RESULTS

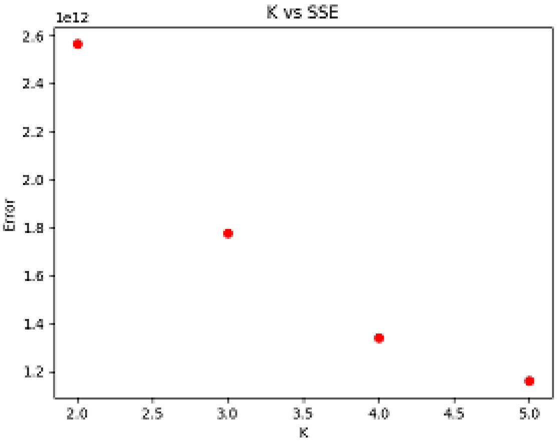
I first wanted to get some more knowledge on the data itself. This led me to clustering as the clusters would reveal some basic knowledge on the most important factors that distinguishes wealth after college and the factors correlating with it. This would help determine the most important attributes as well as the least important attributes which can be thrown away and/or replaced.

4.1 Agglomeration Clustering

At first, I tried to use agglomeration clustering algorithm I used on my previous homeworks. My algorithm was tailored to the homework data which was significantly smaller than mine (3500+ rows with 16 columns). The algorithm made a distance matrix for every point to every other point. This made the matrix 3500x3500 causing the complexity to become too great for my computer. After 2 hours and no distance matrix not complete for the first iteration of 3,000+, I decided to stop the algorithm and try another approach. In the future, I may modify the algorithm to perform better in the future.

4.2 K-Means

Secondly, I used K=Means to determine the clusters. This ran much more easily than agglomeration despite doing over 200 iterations to make sure the smallest error was picked. The algorithm discovered that there are 4 natural clusters within the colleges as shown below.



Using 4 clusters, K-Means was run to find these clusters. With this amount of data, the algorithm took about 2 hours to run to produce the results below. All of the starred fields (*) represent the percent of population from the student's zip code where the statistic is true. Additionally, the abbreviation ME10 means that the field measures the median income of the cohort after 10 years of entering college.

Cluster	Statistic	Amount
Cluster 1 (1432)	Number of Students	1961
	Percent of Students Age 24+	32
	Percent White*	78
	Percent Black*	11
	Percent Asian*	3
	Percent Hispanic*	10
	Percent with BA*	16
	Percent Graduate*	9
	Percent Born in US*	89
	Household Income of Parents	63,960
	Poverty Rate*	8
	Percent Unemployment Rate*	3
	Students ME10	38,086
	75 Percentile ME10	55,444
	Females ME10	37,612
	Males ME10	48,026
Cluster 2 (405)	Number of Students	9649
	Percent of Students Age 24+	23
	Percent White*	78
	Percent Black*	10
	Percent Asian*	5
	Percent Hispanic*	10
	Percent with BA*	19
	Percent Graduate*	11
	Percent Born in US*	87
	Household Income of Parents	72,804
	Poverty Rate*	7
	Percent Unemployment Rate*	3
	Students ME10	53,941
	75 Percentile ME10	78,551
	Females ME10	54,984
	Males ME10	69,836
Cluster 3 (38)	Number of Students	678
	Percent of Students Age 24+	52
	Percent White*	75
	Percent Black*	11
	Percent Asian*	6
	Percent Hispanic*	12
	Percent with BA*	20
	Percent Graduate*	13
	Percent Born in US*	84
	Household Income of Parents	71,513
	Poverty Rate*	8
	Percent Unemployment Rate*	3
	Students ME10	98,908
	75 Percentile ME10	155,852
	Females ME10	102,768
	Males ME10	155,505
Cluster 4 (1635)	Number of Students	1905
	Percent of Students Age 24+	46
	Percent White*	72
	Percent Black*	16
	Percent Asian*	2
	Percent Hispanic*	16
	Percent with BA*	12
	Percent Graduate*	6
	Percent Born in US*	89
	Household Income of Parents	50,259
	Poverty Rate*	14
	Percent Unemployment Rate*	4
	Students ME10	26,596
	75 Percentile ME10	40,217
	Females ME10	26,809
	Males ME10	34,881

This clustering reveals a lot of information about the colleges and students. The four cohorts have drastically different median income of graduates after 10 years at 38,086, 53,941, 98,907, and 26,596. This seems to reveal that their are four different types of colleges having four drastically different medians. Looking at the differences between the attributes can reveal some of the more and least important attributes.

Some of the most important attributes (the ones that have the largest differences between the four clusters) are the amount of graduates from the zip code, median wealth of parents/guardians, percent of students over 24 years old at start, poverty rate, and median earnings by student including all 4 attributes that measure it.The amount of graduates varied between the clusters being between 6-13 percent with the higher percentage correlating with the higher wealth of the students. The median wealth of the parents and guardians varied between 50,259-72,804 dollars which correlates with the higher wealth of the students. Interestingly, this correlation does not reflect perfectly as the highest wealth of the parents did not reflect the highest wealth of the student for the higher wealth cases. This will have to explored further on why this exists. Additionally, the percent of students over 24 years old at start also greatly influences the clustering as it varied from 23-52 percent. Unlike the rest of the statistics mentioned here, the correlation does not necessarily show whether the student would have higher salaries after college. With this unclear impact, I will continue to have this attribute in my analysis, but note it for future evaluations. Poverty rate also seemed to have a big impact. Varying from 7-14 percent, the poverty rate acted similarly to the average income of parents and correlated with the data. Lastly, all of the income measurements of the students after 10 years correlated together; the cluster with the higher median income after 10 years also had to highest median income of males, females, and the 75th percentile. While this data correlated together, it does not differentiate as I wanted and will probably be aggregated into a male:female difference.

The least important attributes seem to be percent Black, percent Asian, percent Hispanic, and unemployment rate. These factors vary slightly between cohorts. The unemployment rate between the clusters only varies by one percent. This attribute, not impacting much, will be taken out of the analysis. Also, the percent of different races does not significantly differ unlike the percent white (which varies 7 percent). To make these statistics more useful, and fully measure their impact, I will aggregate them together in a new category called non-white students. This may provide more impact as well as free up a few columns to introduce new attributes.

The start of my analysis has brought a lot of interesting statistics, but provides more clarification and areas to explore as well. This will be explored further and explained in the final draft.

4.3 K-Means Round 2

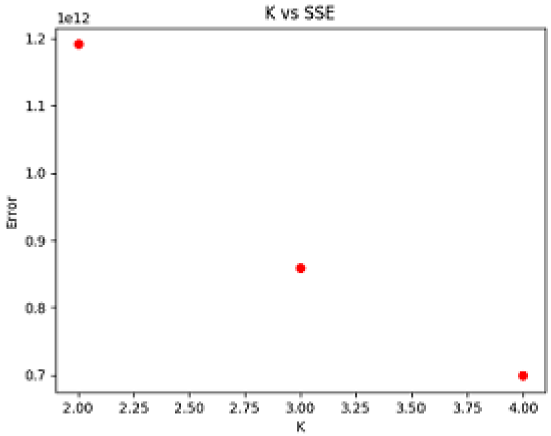
As mentioned before, the previous K-Means emphasized what some of the splits in the data are. Additionally, these splits revealed some of the most and least important attributes. As a result, I cleaned the data as a response.

One of the major changes was to aggregate the demographics together. It seemed like percent white had an impact on the clusters while the others did not. I suspect that the percent white had a larger impact because it varied so much compared to the rest of the other percentages. In order to test this hypothesis, as well as not remove the information fully, I decided to aggregate all of these attributes into a single attribute called non-white. Because someone could fall into multiply categories (white and Hispanic), these attributes do no add to 100.

Another attribute I added from aggregation was Male-Female income difference. While both were interesting and seemingly important, the main use of this attribute was to measure the difference in income. Therefore, I decided to combine these attributes showing the difference between them. This will be interesting to see if this difference attribute has more impact in this form. Also, making 2 attributes into 1 allows me to add more attributes and avoid the curse of dimensionality.

With more room for different attributes, I looked to put in a few more attributes that I hypothesize what would impact the success of the graduates' salaries. Unfortunately, some of these statistics I looked for were unavailable or had a significant amount of Null's. These included SAT/ACT scores of applicants and admission rates to the universities. While I think these would have significant impact on success, I have no choice as these are not provided. Yet, there were other statistics that are important that also were included.

Some of the additional attributes that I included are the median debt accumulated during college, percent of STEM majors given, percent of non-STEM major given, and the type of university. The median debt will be interesting to see as some high-paying jobs require a lot of schooling usually pushing students to a lot of debt. It could also reveal much different choices (or background) of students from different clusters. The STEM and non-STEM percentage was aggregated from over 50 different attributes in the dataset. Some of these majors, like agricultural sciences, were unconventional leading to these fields not just including the stereotypical STEM, non-STEM majors. Lastly, I included a what type of university separating the universities into three categories: public, private, and private for-profit. With these new attributes, I did the same analysis as before to see the impact.



Interestingly, this round of K-Means had 3 natural clusters unlike the 4 natural clusters previously. The results of the K-Means is below in the table:

Cluster	Statistic	Amount
Cluster 1	Percent of Students Age 24+	39
	Percent White*	74
	Percent Non-White*	31
	Percent Graduate*	8
	Percent Born in US*	89
	Household Income of Parents	57,246
	Poverty Rate*	11
	Students ME10	33,777
	75 Percentile ME10	49,943
	MF Difference ME10	10,020
	Type	1.87
	Debt Median	11,813
	Percent STEM	65
	Percent Non-Stem	61
Cluster 2	Percent of Students Age 24+	34
	Percent White*	74
	Percent Non-White*	31
	Percent Graduate*	9
	Percent Born in US*	86
	Household Income of Parents	62,611
	Poverty Rate*	10
	Students ME10	37,977
	75 Percentile ME10	56.027
	MF Difference ME10	13,404
	Type	1.84
	Debt Median	11,813
	Percent STEM	62
	Percent Non-Stem	44
Cluster 3	Percent of Students Age 24+	34
	Percent White*	77
	Percent Non-White*	26
	Percent Graduate*	8
	Percent Born in US*	89
	Household Income of Parents	37,397
	Poverty Rate*	9
	Students ME10	37,397
	75 Percentile ME10	55,044
	MF Difference ME10	10,628
	Type	1.74
	Debt Median	11,813
	Percent STEM	63
	Percent Non-Stem	43

Upon seeing these results, I realized some of my mistakes in creating this dataset. Unlike the first round of K-Means, the clusters were not truly defined well. The first K-Means had each cluster with drastic differences in median wealth of students after 10 years and the other categories. While there are some differences here, they are not too drastic.

The biggest flaw in my creation was the type. This field was provided as a single field with integers representing the difference. With 1 representing public, 2 as private, and 3 as private for-profit, these values' non-discrete center of masses does not create correct implication. As seen above, the types are all continuous values around like 1.74. What does this mean? Having it as a single attribute makes this ambiguous. As a result, I plan on splitting the attribute into three separate attributes as boolean values; if a college matches the type it will be a 1 and if not it will represented as a 0. This should disambiguate the information.

4.4 Cross-Correlations

Previously I was using K-Means to understand the data. As I did not know anything about it, K-Means brought interesting results and enhanced my knowledge the first time. The second time indicated the flaws of relying on K-Means fully for understanding the impact of each attribute. With this as well as review of cross-correlation in class, I decided to do cross-correlation analysis for the attributes I used previous with the type changes I mention above. The results are followed between each attribute and median income of graduate after 10 years:

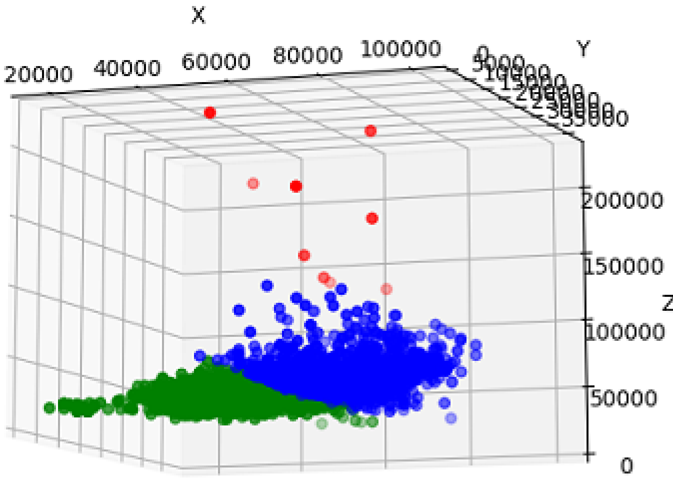
Attribute	CC Coefficient
Percent of Students Age 24+	-0.38
Percent White*	0.18
Percent Non-White*	-0.28
Percent Graduate*	0.56
Percent Born in US*	-0.009
Household Income of Parents	0.57
Poverty Rate*	-0.41
75 Percentile ME10	0.54
MF Difference ME10	0.50
Public	-0.07
Private	0.33
For-Profit	-0.28
Debt Median	0.50
Percent STEM	0.08
Percent Non-Stem	0.06

These results reveal great detail about the impact of each attribute on the median income of graduates after 10 years. The least impactful attributes are Born US, public school, STEM, as well as Non-STEM. These results are not too surprising. While immigrants are usually not as well off as native students, they are often encouraged to do well in school; this would have positive impact on their salaries as well as their lower wealth decreasing their success. This balance of both would provide non-correlation to wealth. Also, public schools vary greatly in quality. Some are the greatest schools are public as well as some of the worst. This variety would cause great variety in median wealth of students. Lastly, and most surprising, is the little impact of major choice in STEM or non-STEM. While there are some non-STEM majors that can make a significant amount of money (law, business, etc) it is heavily believed that STEM majors generally make more money. This difference could be caused by including the non-stereotyped majors. While computer science and engineering majors do make more money and are STEM, there could be other STEM majors that do not make that much money. With these attributes being insignificant, I will remove them in future analysis.

The most important attributes seem to be household income of parents, percent graduates in zip code, 75 percentile of wealth, male-female difference, and debt median. The percentage of graduate degrees and household incomes seemed to be linked with each other. People with graduate degrees usually make more money which would result in the zip code of the parents more expensive, therefore, implying that the parents' household income to be high. The male-female difference, while unfortunate, does make sense to scale as the wealth increases; with higher wealth there is more room for differences to exist. Also, these higher paid positions are, most likely, management and high-paying jobs (like engineering and computer science) where men usually reside more often. The median debt also makes sense as the more advanced degrees require more years of schooling which will require the student to acquire more debt. All of these attributes make sense to have impact and will be kept and looked at as I continue the analysis.

4.5 K-Means of Most Important Attributes

With the most important attributes discovered, I wanted to be able to visualize and see some of the differences in a graphic way. While I was able to see the clusters before in a chart, with numerous attributes, I could graph the results. Using the most important attributes of household income and graduate percentage with median income, I did K-Means on the data.



The clusters are very different than I expected. Firstly, the clusters are very different in terms of data points. The green points are the colleges where the students have the lowest parental household income (x) and the lowest median debt (y). This cluster is the biggest with 2,577 universities. The blue cluster has a higher parental household income, higher median debt, as well as slightly higher graduate income after 10 years with 1,624 points. While these clusters are separate, they are relatively similar with many points on the edge of the cluster definitions. The red cluster remains the most unique as it only has 9 points with significantly higher graduate income after 10 years. These outliers require their own cluster with universities like Stanford and Harvard. These schools are very unique compared to the majority and incredibly in the minority. In the graph, I had to select every 5th point in the blue and green clusters in order for them to be readable as well not overwhelm the graph itself.

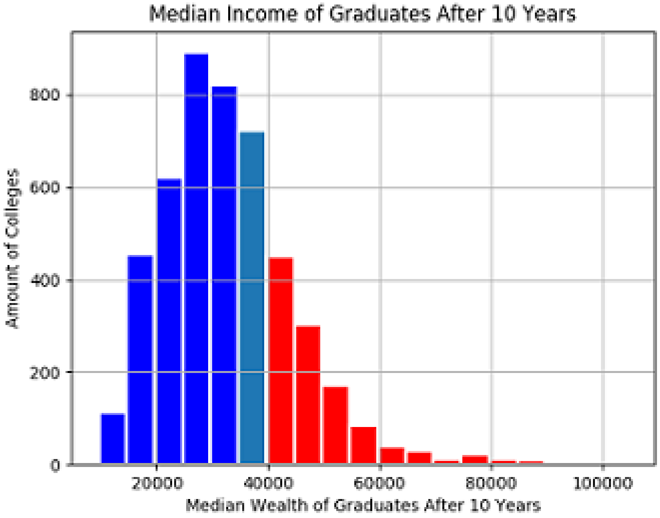
With so few very successful schools, I wondered why are they the minority. These schools do not have greatly varied household income and median debt compared to the other two clusters. What are the differences?

4.6 Ostu's Method

As the median wealth seems to be quite varied, I was curious on why these splits resulted; what differences are there between the wealthier and poorer graduates? To answer this question, I decided to do Ostu's Method on the median wealth of graduates after 10 years and find the two different clusters.

For Ostu's Method, I decided to have the bins as sets of 2's. While this created a large amount of bins requiring a lot of runtime, I was able to leave the program running overnight and get a more accurate answer. For the graph below, I decided to make the bins as 5,000 to make it more readable. Additionally, with the few, higher income outliers, the graph gets expended greatly. But with only one school in those bins, the histogram's scale cannot show this. As a result, I decided to maximize the income after 10 years (the x-axis) to 105,000 where the amount of school in that bracket drops drastically.

Below is the graph of all of the schools' results of the median graduate after 10 years of graduation.



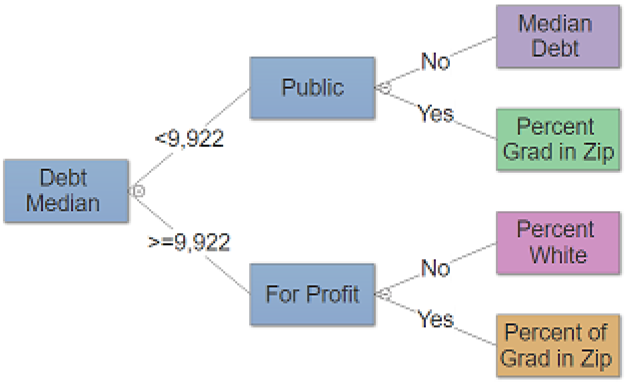
As you can see from the graph, the large amount of median wealth after 10 years visually averages around 30,000 dollars. The blue portions of the graph represents the data points less than Ostu's Methods results of the threshold. The red represents the data points higher than that. The threshold of 38,000 dollars is in between the bin of 35,000 dollars and 40,000 dollars. With more points in the blue region, the turquoise-like bar represents the bar where the threshold is.

Best Threshold	38,000
Best, Mixed Variance	2,632,052

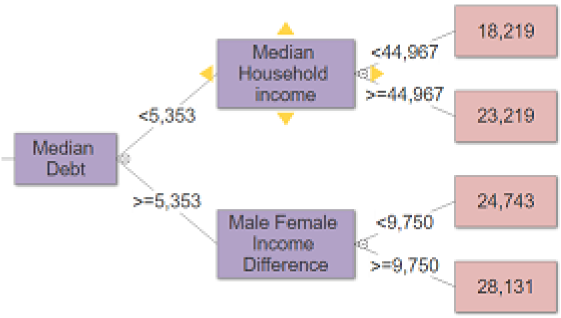
Interesting about the data, is the variance between the two clusters. While the blue cluster has the points centered around its center, the red cluster has a much larger variance. As mentioned above, there are even more points not shown in graph that have even larger median income of graduates after 10 years. With the vast differences in the red clusters, I suspect their classifier will be less certain than the blue cluster. To test this hypothesis and to explore the data further, I split the data into two separate .csv files based on the clusters reflected here.

4.7 Classifier on Cluster 1 of Ostu's Method

Upon splitting the data into two groups based on Ostu's Method, I decided to use Weka to configure a decision tree. Originally, I let the algorithm run without setting a depth or changing any of the original parameters. This led to an extremely large decision tree to be generated with over 10 levels on depth. This was obviously too large and unreadable with the tree being too specific to analyze the most impact. Therefore, I decided to limit the depth to 5. This made it much more readable, but still unable to translated well into this report as well as still being very large. As a result, I went to maximum depth of 4 where it emphasizes the important factors while still classifying successfully a majority of the time. To make it readable, the decision tree was broken up into 5 separate sections which are color coded to represent where they originally came from.

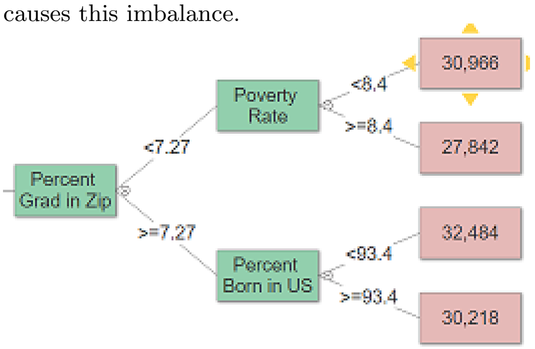


The first split comes from debt median of 9,922 dollars. After this, those with less than 9,922 dollars ask whether they were public or not. This makes sense as public schools are generally cheaper than private and for-profit. From there it defines and goes into the other graphs that are below. If they are public, they go into the green graph where it explores percent of graduates in zip code. If they did not, then they go into the purple graph where it looks into further defining median debt. On the other side, if the median debt was greater than or equal to 9,922 it splits into whether for-profit or not. This also makes sense as for-profit students usually acquire a lot of debt. If they went to for-profit schools, then it will be looked at in the orange graph where the percentage of graduate degrees in zip code of the students. Lastly, if they did not go to a for-profit school, then it will explored in the pink graph where they check percent white of the zip code the students were in. These paths will be fully explored below leading to the income of the students after 10 years after graduating.



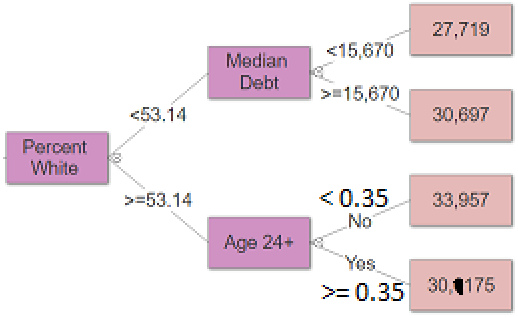
The purple subsection branches from the main graph. In order to get here, the colleges need to have had a median debt less than 9,922 as well as being a non-public university. In this subsection, it further defines the median wealth with having another split at 5,353 dollars. If they has less than 5,353 dollars, the last deciding factor was median income of parents. If their parents made less than 44,967 dollars, the students averaged 18,219 dollars after 10 years while if their parents made greater the students averaged 23,219 dollars. While both are low salaries, the higher amount of parents salaries coincides to higher children salaries as discovered in the cross-correlation coefficient above.

The other side of the subsection results from the median wealth of the students being greater than or equal to 9,922 dollars. The next split results from the male, female income difference at 9,750 dollars. The colleges with less difference have an average income after 10 years of graduates at 24,743 dollars and greater and equal to with 28,131 dollars. This surprises me greatly with the male, female difference. Assuming the school has equal male and female enrollment, the females could only be making 22,000 dollars while the males make 32,000 dollars (with 10,000 difference). These differences are drastic especially at this low income making me curious if these are specialized schools or something that causes this imbalance.



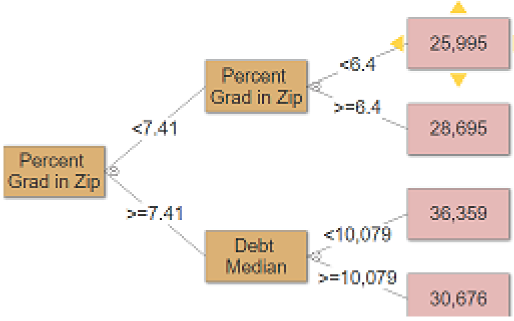
In the green subsection, containing the highest median salary after 10 years graduated, these colleges had median debt less than 9,922 dollars and from a public school. They then switch depending on percent of graduate students in their zip code at 7.27 percent. For those with less, the poverty rate than comes into account with those less than 8.4 percent earning 30,965 dollars. Those with the higher poverty earn 27,842 dollars. These numbers do make sense as surrounded by graduates do lead less poverty and vice versa the opposite. Interestingly, the colleges with average of 27,842 dollars is around the middle of this section of the data. Yet, the students were not surrounded by a lot of graduates and had high poverty rates around them. I suspect that these schools would be associated with trade jobs or other decent paying blue-collar jobs.

Meanwhile, if the students from a college had more percent of graduates from their home zip code greater than 7.27 percent then the most important factor became percent of the people in their home zip code that was born in the United States. If less than 93.4 percent then the student earned 32,484 dollars on average while the greater earned 30,218 dollars. Interestingly, the group with higher immigrants made more money than the non-immigrants. This helps justify the belief that immigrant families seek out opportunity more than non-immigrants. Yet, this result does not conclude anything and that aspect will have to explored in another study to prove or disprove.



In the pink subsection, these colleges had a median debt of graduates greater than or equal to 9,922 dollars and from not from for-profit schools. If the school had students from zip codes that had less than 53.14 percent white ethnicity and had a median debt less than 15,670 dollars, then the students averaged 27,719 dollars. Otherwise, if the students had more debt, then they would have an average income of 30,697 dollars. This is an unfortunate path as minority students have a lot of debt with very little income. Interestingly these students did not attend a for-profit school and must have accumulated this debt going to a public or private school. With these results, I suspect that the students did not pick a profitable major and had little to no money to spend for college.

On the other side of the pink subsection, the students were from zip codes of more than 53.14 percent white. The schools where the students were less than 35 percent over age 24 when starting earned 33,957 dollars while those with less, older students had an average earning of 30,175 dollars. This surprised me as I thought the higher aged students would earn more money as they are already established compared to their younger counterparts. Yet, I can see where this statistic can be true as it is possible that the older students want a "do-over" and have gone back to school to do so.



In the orange subsection, these colleges had median debt of graduates greater than or equal to 9,922 dollars and from for-profit schools. The upper portion of the graph contained colleges where students had less than 7.41 percent of graduates degrees in their zip code. These students were also further defined by if they had less than 6.4 percent graduates in their zip code making 25,995 dollars while those with greater made 28,696 dollars. As seen before, there emerges a trend of more graduates in your zip code, the more success you, on average, achieve. This trend emphasizes how important environment of the students is to future success.

In the bottom of the orange subsection, the students have greater than or equal to 7.41 percent of graduates in their zip codes. They are further broken down into how much debt they accurred during college. If they have accurred less than 10,079 dollars of debt they averaged 36,359 dollars while their counterparts with more debt averaged 30,676 dollars of income after 10 years. Interestingly, the cohort with less debt made the most money. This could be because of the for-profit format or due to the students being more calculated with money as they went through the college system. Yet, it is unfortunate for the students with more than 10,079 dollars in debt as that makes up over 30 percent of their income (after 10 years!) and will be a great burden on them to pay it off.

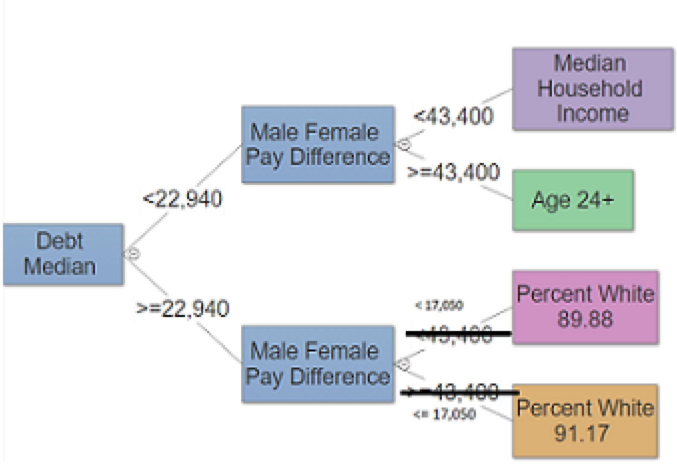
While these incomes are all of the lowest of the data, they still vary greatly. The highest incomes emerges with 36,359 dollars and the lowest one at 18,219 dollars. The highest one had the path of having a median debt between 9,922 and 10,079 dollars, from at for-profit school, and had 7.41+ percent of graduate degrees from their zip code. Meanwhile, the lowest cohort have debt less than 5,353 dollars, did not go to a public school, and have a parental household income of less than 44,967 dollars. These paths vary widely and have large differences in money. While the larger income accumulated about double the debt than the lower income (5,000 dollars more), the higher income cohort makes double the amount than the lower cohort at 18,000 dollars more. While this implies that more debt relates to more income, there was another subsection that had more than 10,079 dollars in debt that made less than this highest region.

The biggest difference between the two incomes emerge from household income and graduates in the region. While both did not measure each category, both of these categories heavily influence each other. People with graduate degrees make significantly more money than the average. With the lower money path having very low income, this implies that they have low amount of graduates in their zip codes. Meanwhile, the highest income bracket has more than 7.41 percent of people in their region with graduate degrees implying that their zip code is much wealthier than the other. These differences emphasize how important the student's environment impacts their success. This detail becomes reinforced in observations as wealthy people often stay wealthy and poor usually remain poor with some, rarer cases of class mobility. Unfortunately, stagnant mobility emerges probable in the data analysis done.

4.8 Classifier on Cluster 2 of Ostu's Method

The above section focused on all the data points (colleges) where the average student's income was less than 38,000 dollars which was discovered by Ostu's Method to be the best threshold to split the data based on mixed variance. This section will focus on the data points that have greater than 38,000 thousand dollars for average income of students. As mentioned previously, this group has much more variety as the ranges exist from 38,000 thousand dollars to over 100,000 dollars.

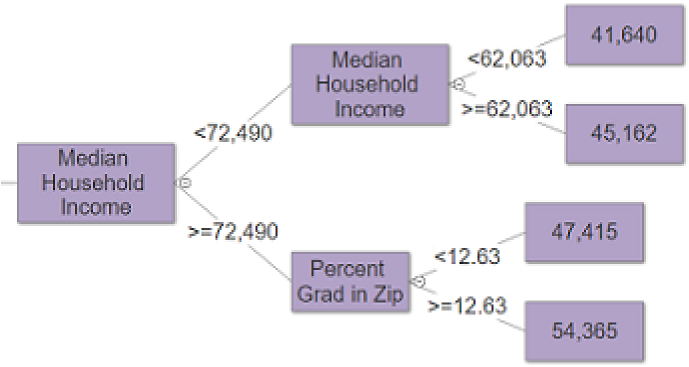
The tree was derived using Weka's decision tree functionality. As mentioned above, I decided to restrict the depth to a maximum of 4. Having tried no limit as well as a limit of a depth of 5, the trees became much too large as well as too specific. As I am not trying to predict and just trying to find the generalities, I want the trees to be general and reveal trends that differentiate the lowest and highest paid individuals.



Above shows the first part of the tree for this subsection. As structured above, I split the tree into four additional subsections to make it more readable that correspond to the color of the leaf. The first main split comes from median debt of the students. With less than 22,940 dollars in debt, the students were broke into groups depending on male-female difference in income. If the difference was less than 43,400 dollars, then the group went into the purple subgroup where they were fully split by median parental household income. If the male-female difference was greater than 43,400 dollars, then they were deemed to be in the green subgroup breaking them upon percent of them who are older than age 24 at entry. The most interesting break here was the large break in male-female difference. At 43,400 dollars, this divide is greater than any of the incomes in the lower cohorts. Additionally, these schools must be incredibly niche, as this difference is very significant.

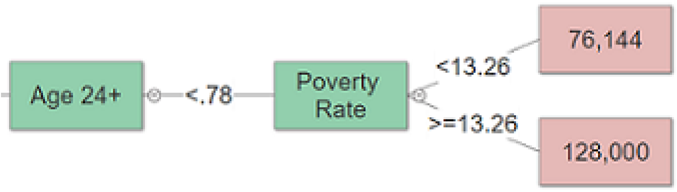
In the lower portion of the tree, the schools have an average debt of greater than or equal to 22,940 dollars. The next factor was also male-female difference in income after 10 years. This split is less drastic at 17,050 dollars. Those colleges with less than 17,050 dollar difference went into the pink subgroup to be further broken up by the percent white their home zip code was. The colleges with greater than 17,050 dollars of male-female difference in income went into the orange subgroup where they also split based on the percent white of their zip code. Interestingly, those with more debt had less male-female difference in income on average. It also is intriguing about how this more-debt group also both converged on percent white being a strong indicator of income.

Below will explore each subsection of the decision tree with the paragraphs about the section below the colored picture of it.



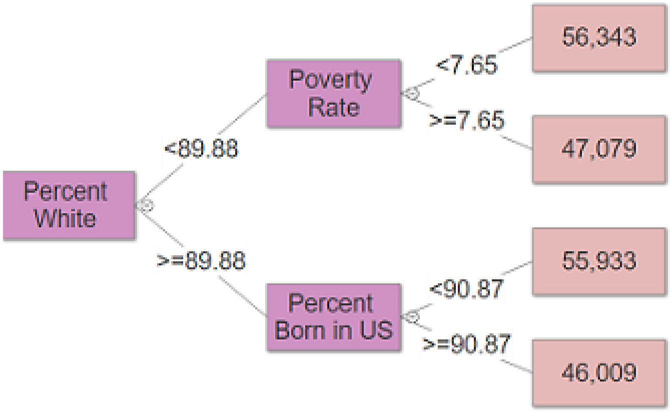
This subsection resulted from the colleges' average debt of students being less than 22,543 dollars as well as the median male-female income difference being less than 43,400 dollars. The next divide in the data came from median household of income and whether the college's average was greater than (or equal to) or less than 72,940 dollars. Those with less than that amount were further broken down into a further definition of household income of 62,063 dollars. With those families' earning less than that, the students averaged earning 41,640 dollars while the student's whose families had higher income earned about 45,162 dollars. Again, this difference comes from parents income. The higher the student's parents income is, the more money the student usually makes.

On the other side of the graph, the students have higher parental household income at greater than or equal to 72,940 dollars. The next deciding factor comes in the percent of graduate students in the zip code that the student resided in. With less than 12.63 percent, the students had a median salary after 10 years of 47,415 dollars while those with greater than or equal to having an average income of 54,365 dollars. Interestingly, as you look down the purple subgraph, the incomes of students go up. Additionally, as you look down the purple subgraph, the affluence of the student's neighborhood growing up goes up. The environment of the student's childhood greatly determines the wealth of the students as the become independent adults.



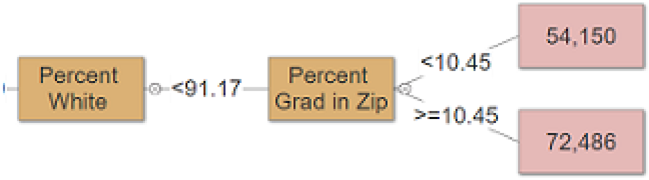
The green subsection originated from the average college debt of students being less than 22,543 dollars as well as the median male-female income difference being greater than 43,400 dollars. The unique green cluster has less splits than the previous sub-trees. Unfortunately, the software I used to generate the trees was limited to not be able to show this difference and was, therefore, excluded from the sub-tree. At the first split of percent of students older than 23 years old, if the school had greater than or equal to, then the income of the students after 10 years after graduation was 167,400 dollars. This wage, by far, rushes ahead of the others as the highest salary. This group has much more older students possibly being a result from older professionals getting more degrees because they can or getting a masters or other specialization to boost their income at their job. This would explain the extremely high income as well as the older population.

The other side of the tree contains the students where they have less than 78 percent of students greater than age 23. From this break, the students were split further into poverty rate. The students from zip codes of poverty rates less than 13.26 percent earned about 76,144 dollars compared with the students from zip code that had poverty rates greater than 13.26 percent earning about 128,000 dollars. These results are incredibly interesting as they contradict the previous trends; the students who came more poverty areas earned more money than the ones who did not. While they both made a great deal of income compared to their peers, the students of less affluent areas earned about 50,000 dollars more. This outlier proves that while difficult and against the odds, their is a possibility of social mobility in the United States college system.



This pink subsection came about by having the data points which contained the debt median being greater than or equal to 22,940 dollars as well as the male-female income difference being less than 17,050 dollars. The data is further split based on the percent white of the student's zip code where they grew up. If their hometown was less than 89.88 percent white, then the next factor was poverty rate of their hometown. If their zip code that they grew up had a poverty rate of less than 7.65 percent, then the students made a median of 56,343 dollars while those with the poverty rate of higher than 7.65 percent had a median income of 47,079 dollars after 10 years. Again (and again), the more affluent the child's situation growing up, the more successful they were as adults in terms of salary.

On the other side of the graph, the students had lived in zip code where the people were greater than or equal to 89.88 percent white. This was further split based on how many people from their zip code growing up were born in the United States. With less than 90.87 percent, the students usually averaged 55,933 dollars at the end of 10 years after college. Those with greater than 90.87 percent, meanwhile, had an average salary of 46,009 percent. Interestingly, the higher amount of immigrants resulted in higher incomes. Yet, this split was the successor of the split guaranteeing that 89.88 percent of the people in the zip code were white. Together, the majority of the immigrants in the subsection are white. This distinguishing could be explored more to determine what are the true factors that led to this split with much greater income.



Finally, the orange subsection resulted from the student's average debt being greater than or equal to 22,940 dollars and the male-female income difference after 10 years being greater than or equal to 17,050 dollars. Similar to the green subsection, this unique tree has a split that determines the wealth of the students at an earlier depth than the rest of the decisions. With the limits of the software I used, I could not portray this information clearly and, therefore, left it out of the graph above. This missing split emerges from the first diverge at the percent white of the student's zip code growing up. If they had greater than 91.17 percent white in their zip code then they would earn, on average, 41,800 dollars 10 years after graduation. This salary becomes one of the lowest of this group and with the prerequisite of having 22,940 dollars in debt as well as the male-female income difference being 17,050 dollars. This means that a female in this cohort could have an income of 28,000 dollars with 22,940 dollars in debt. These schools do not seem to be best for students especially females.

In the shown portion of the orange subsection, the schools have less than 91.17 percent white where the student's zip code. The determinate split emerges from the percent of graduates in the zip code of the students. If the students have less than 10.45 percent of people with graduate degrees in their zip code where they grew up, then they average having 54,150 dollars after 10 years. Meanwhile, the students with greater than or equal to 10.45 percent of graduates in their hometown zip code average a salary of 72,486 dollars. The students who were surrounded more with people with graduate degrees often surpass their peers who were less likely to be near graduates.

The differences in income after 10 years are drastic in this cohort. The lowest average salary brings in 41,639 dollars while the highest earns 167,400 dollars. The lowest earner was differentiated mostly by lower household income and less debt. The lowest earners were twice split into lower groups where household income of parents were accounted for. Additionally, these universities had students with less than 22,940 dollars in debt. While they do not make that much, the lower debt could suggest that they did not get advanced degrees as they would have accumulated more debt due to their parent's low income. Although, this cannot be determined without further investigation.

The high earners of this cohort are vast outliers compared to their peers. The two highest at 167,000 and 128,000 dollars are, respectfully, the highest incomes by far for 10 years after graduation. Yet, these leafs are outliers as they only are one school respectfully. These schools were so unique that the decision tree had to give them a separate path. Due to this, I will disregard them for this analysis as they are not showing trends of a group of universities and students.

The next highest salary cohort emerges at 76,144 dollars. They are apart of the universities which have a debt median of less than 22,939 dollars, the male-female income difference is greater than or equal to 43,400 dollars, the percent of students older than 23 the start is lower than 0.78 percent, and the poverty rate is less than 13.26 percent. None of these are great indicators to point out what this cohort represents due to the high variability. This selection was near the other two outliers which swayed the decisions. If this cohort had an average of 5 percent poverty while the one, outlier school had 13.27 poverty levels. This could cause the split to be at 13.26 percent (where it is currently) or 5.01 percent and still have the same splits. The most telling feature emerges from the male-female difference. Being so high, we can adjust the median income after 10 years to be 97,000 dollars for males and 55,000 dollars for females. This, most likely, emerges from schools who have a high amount of majors which make a lot of money, but are male dominated. For example, computer science majors make much more money than other majors, yet if you go inside the classroom, the vast majority is male. This would drive up the male salaries while the females could be skewed by the same, but reversed, occurrence (high amount of females in a major that does not make a large amount of money). This difference could be further explored for additional research.

5. FURTHER RESEARCH

With a limited amount of time to do this project as well as limiting myself to the College Scorecard, there are many areas of further research that can be explored. The major areas that I would have personally looked into without a time constraint would be adding more attributes about the student, looking at geographical impact on salaries, as well as exploring certain cohorts much more in depth.

5.1 More Attributes of Students

One of the areas of further exploration will be of more attributes/factors. While the College Scorecard has a lot of information and attributes, it has a lot of the attributes as 'NULL' or 'Privacy Suppressed'.[3] In this report, I disregarded all non-values. Unfortunately, this cleaning removed close to 2,000 schools for certain analysis. This amount of removal could have skewed the analysis and/or not revealed a trend that exists the schools not included. These missing fields also had influence in what attributes I could pick in this study.

During the second phase of K-Means, I had the ability to fill in more attributes. Unfortunately, many of the attributes I was interested in (mentioned below in each subsection) were missing over half of the time or were not measured in this dataset. This required me to look at secondary statistics that were interesting, but I hypothesize not as important as some of the others I wanted to measure.

5.1.1 SAT/ACT Scored

SAT and ACT scores are often represented as the student's academic amplitude. While argued in its accuracy, it remains as one of the only organized, national test that almost every student takes. While the scores usually trend with wealth of the parents due to paid tutors, it will still indicate the average academic intelligence of the students. This would have been useful as I do not have a measurement in this type of achievement. Also, the tests are often broken down into different sections with both measuring math and English skills (as well as science, writing, and other skills depending on which one you take), which would be interesting to explore which section correlates greater with salary. Additionally, it would be interesting to explore how much impact a test done in the late teens impact someones salary in their late 20s, early 30s. If I had more time as well as have the data accessible, I would definitely put this attribute in the study.

5.1.2 High School Rank

One of the flaws of the SAT/ACT scores would emerge if certain students are not good test takers as well as being heavily influence by parental wealth. Looking at high school rank as well would compliment the scores as well explore the impact of high schools on the salaries of the students after 10 years.

While a lot of high schools are segregated based wealth, they can often have very diverse populations. As public high schools do not follow town lines, a high school in a wealthy area can cut into a much poorer neighbor town. Additionally, private and charter schools often offer scholarships and financial aid to underprivileged students. With this allowance of wealth differences, it will help measure the impact on the high schools themselves.

Unfortunately, high school rank emerges vague. How do you measure effectiveness of a high school. While this ranking uncovers many questions on best, there are many places that do the rankings based on different categories. If I was trying to include this rank, I would average all of these rankings together. While one may favor AP scores another may favor college acceptance which the other factors in the hardships of the high schools and how well they perform in consideration to these hardships. Doing an average would remove some of the extreme and/or biased rankings and provide a relatively reliable ranking which will help in the analysis of success after college.

5.1.3 College Acceptance Rate

College acceptance rate changes drastically between colleges. Some elite, private colleges accept less than 10 percent of all applicants while some community colleges accept over 90 percent of all applicants. This high rejection number allows these exclusive colleges to pick the most successful students (from their definition) which would definitely impact the salary after graduation. Some of these hand-picked students would have probably made a lot of money regardless of what college they went to if they went to college at all. This would inflate the colleges' success numbers as well as place the student's earned success onto the college wrongfully. With college acceptance, it would further explore this disparity as well as give more context on the different cohorts.

5.1.4 College GPA

An aspect that is not accounted for in my study emerges in how well the students do in college. It is believed that the students' success relies on their GPA, but this depends on the how hard the college grades. Some colleges inflate their grades so that the average student has a 3.5+ GPA while others grade much tougher. While this would be hard to measure, it would reveal whether the inflation has an impact on the salaries of the students. If it does not effect salary, then grades are probably not reflective of the student. Otherwise, we can see which schools prepare their students the greatest.

If I were to include college GPA, I would also include other metrics which would compliment. For example, I would look at number of internships taken during college or another attribute that would measure college success. This would help explore how much the college has impact on the students' salary 10 years after graduation.

5.2 Geography

While not conventional, geography could reveal latent variables that are unable to be measured directly. Most students take jobs near their college. Two people getting the same job but one in Kansas and another in San Francisco would have drastically different salaries simply based on price of living difference. The measurements of this study simply record salary after 10 years, but do not consider geography. I very wealthy person in the mid-west United States could only make 60,000 dollars as they would be significantly wealthier than everyone around them. Similarly, that 60,000 dollars would have that same person struggle if they lived in expensive places like San Francisco or New York City. These factors will have to be explored to fully understand the data.

As the data measures colleges and not individuals, it would be impossible to show where the students work after college and their weighted salary based on this. Instead, knowing that the majority of students get jobs close to their college, I would take into consideration the state and/or region the college is. Unfortunately, that would create a categorical attribute that cannot be used in numerous analysis. It would also not be too conclusive as a many states have very varied standards of living within the state. For example, Rochester and New York City, while in the same state, have drastically different rent prices. In Rochester, I spend (divided by my roommates) 1400 dollars a month for a 3 bedroom, 1 bath in one of the nicest parts of Rochester. My brother spends 3,500 dollars a month for a 1 bedroom, 1 bath in an up-and-coming part of Brooklyn. These prices alone require a great amount of salary difference to afford and this does not include food or any additional expenses.

Looking at geography emerges incredibly important to put into context these incomes, but this evaluation becomes complex and needs to be further explored to understand the best way to do so.

5.3 In-Depth Exploration on Certain Schools

With the time restraint on this project, I discovered interesting trends in certain clusters, but could not fully explore them. For example, in the higher-salary classifying section, there was a cohort which was an outlier as they had higher poverty rates, yet had high salaries. This contradicted every other cluster. I would explore why this happened.

Additionally, I would explore a sample of the three best, salaries regions and the the worst, lowest salaried regions. Minimizing my data to just the extremes will allow me to see what makes these regions extremes. Attributes within the two clusters might be shared and some might be opposites. Doing a full in-depth exploration could reveal a lot of differences revealing the biggest influences on monetary success after college.

6. CONCLUSION

Upon finishing the analysis in the scope of this project, numerous trends have emerged that link to the student's monetary success after college. Although more research should be done before finalizing these observations, numerous trends emerged including the impact of pre-college parental affluence, the impact on age at different wealth ranges, and the other, important factors in college success.

6.1 Affluence of Student Before College

The affluence of the parents greatly impact the success of the students. Throughout the study, the highest salaries always had the largest amount of parental household income and graduate degrees in their zip code.

In the first K-Means analysis, the two highest incomes also had the two highest household incomes and graduates percent. The cluster with the largest income for the students after 10 years after graduation, had a parental household income of 71,513 dollars and percent of graduates in the zip code at 13 percent. This drastically differs from the lowest income cluster where the household income was 50,259 dollars and a percent of graduate degrees in their zip code at 6 percent.

Additionally, throughout both of the decisions tree, the splits of higher percent of graduates in the hometown zip code as well as higher parental household income led to higher student income. These correlations are confirmed as their cross-correlation coefficients are extremely high with household income of parents having a 0.54 coefficient and graduates per zip code had a coefficient of 0.56.

Interestingly, it seems that parental household income just has to be over a certain amount while graduate increasingly impacts the income of the students 10 years after graduation. For the first K-Means analysis, the best student income of 98,908 dollars actually has slightly less parental income at 71,513 dollars than the second best student income. This second group had a much lower student income at 53,941 dollars while their parents made more money at 72,804 dollars and with the percent of graduate degrees the zip code of 11 percent. This difference implies that while the individual's affluence was important, it was more important of the surrounding neighborhood to be affluent.

The students who grew up wealthy often stayed wealthy. Additionally, being surrounded by graduates, the students aspired and acquired more wealth than their peers who were not.

6.2 Age 24

Another interesting attribute was whether the student entering was older than 24 years old. Despite the cross-correlation coefficient being negative at -0.38 percent, it sometimes has a positive impact on some of the cohorts in the higher incomes. For example, the largest income in the decision tree comes from having a high amount of students above age 24 enrolled with students making 167,400 dollars after 10 years of graduating. The possible outlier might be because of the majority of age 24 year old going back to college are trying to re-define their career resulting in low salaries. But this wealthy cohort most likely are going back to school during the job and, therefore, get a vertical or horizontal promotion due to going back. This explanation does make sense for these results, but cannot be fully concluded without further exploration.

6.3 Most Important Attributes

Using all of the analysis done I was able to determine some of the most important attributes contributing to monetary success of graduates after 10 years. In the decision trees, the only attributes used for the splits were median debt, male-female difference in income, household income, percent of graduates from zip code, the number of students age 24 or more at entry, as well as the poverty rate and percent white where the students came from. Every other attribute was not used in any of the splits making it not very big in differentiating.

While age 24, household income, and percent graduates of the area they grew up are explained above, the rest of the attributes do have importance, but not as much as the ones above. The next important factors emerge in median debt and male-female difference. The median debt correlates greatly with more advanced degrees, as the longer the student is in school the more debt they can accrue. Similarly the male-female difference becomes emphasized at the larger incomes. A large disparity requires high wealth. So while these attributes are important, they represent latent variables that correlate and accumulating wealth does not equate to more income.

Poverty rates seemed to impact the lowest incomes the most and this seems to be similar to graduate rates except having the opposite effects. The percent white seems to correlate with poverty rate slightly as minorities often have higher poverty rates. All of these attributes have impact, but like median debt and male-female difference probably measure other latent variables that can be explored and remedied to fix this difference.

7. FINAL REMARKS

College can catapult a person through income brackets. But, as concluded in this study, this rarely happens and often the wealth becomes heavily influenced by a person's environment before they even enter college. While this study is crude compared to an official study, these results are concerning. Social and class mobility is key for society and the lack of so develops into a caste system. Hopefully with these results, other researches and more powerful people can recognize this trends and research more about what exactly causes this phenomenon and how we can decrease it more often.

8. REFERENCES

[1] U. T. Institute. Are old assumptions about college, careers and the path to success outdated?

[2] A. Kamenetz. The new college scorecard: Npr does some math. September 2015.

[3] U. D. of Education. ”college scorecard data”.

[4] U. D. of Education. ”data documentation”.

[5] P. Office. Education department releases college scorecard to help students choose best college for them. February 2013.

[6] Reddit. ”data is beautiful”.

[7] M. L. Thompson. College scorecard: Earnings premium and value proposition. Janurary 2017.