An Analysis of Student Success in Higher Education

Abstract:

Higher education can have profoundly positive effects on the lives of individuals who successfully graduate. As a result, it is vital to understand what factors contribute to being successful in higher education. This study aims to discover and analyze these factors in two different ways. Using a dataset containing information about university students, this analysis will first use k-prototypes clustering to find patterns that relate certain characteristics to academic success. This study will then deploy a multiple linear regression model to more specifically find which characteristics most significantly affect academic success. The clustering and regression analyses ultimately yielded different conclusions. The clustering model mainly found patterns between having a significant other and obtaining high grades and between having many siblings and obtaining high grades. The regression analysis concluded that the only factor that had a significant effect on academic success was reading non-academic readings often. Overall, limitations from the dataset and our methods led to inconclusive results. As a result, this study concluded that relating characteristics to academic success is far more complex than originally anticipated.

Introduction:

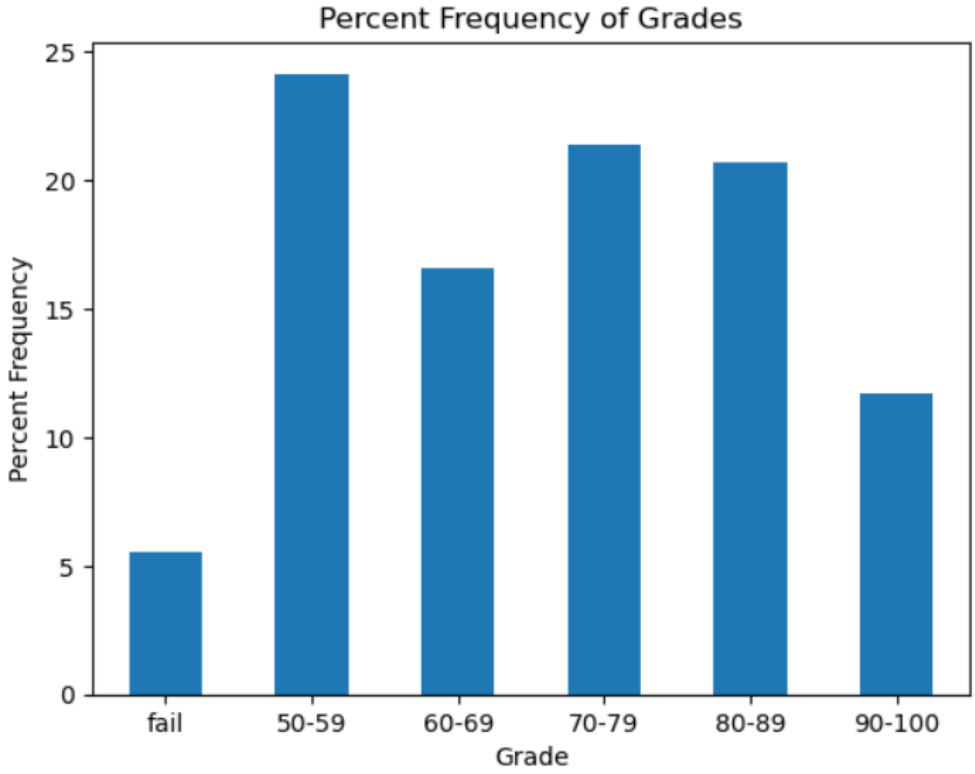
Higher education has been proven to have numerous benefits. Of course, higher education tends to leave students more well-rounded due to increased learning on a wide variety of subjects. In addition to this, the emphasis and detail that major programs apply to their respective fields allow students to seek work in the future with a set of specialized skills and experiences that are different from those of the general population. Not only does higher education positively affect the learning and knowledge of individuals, but it also is an opportunity for profound personal growth. While in higher education, students improve their organization skills while learning how to manage their time, their communication skills in having to speak with professors, classmates, and advisors, and their personal skills in having to deal with new roommates. Additionally, students in higher education tend to be exposed to a high amount of cultural diversity that prepares them for joining the workforce in the future. Perhaps the most significant benefit of higher education is the difference in salary received compared to those without higher education. According to a Georgetown University report that cited a 2002 U.S. Census Bureau study, on average, the lifetime earnings of individuals with a Bachelor's degree is 75% higher than the lifetime earnings of individuals with just high school diplomas. This disparity grew to 84% in 2011 when it was found that a Bachelor's degree on average earns 2.8 million dollars over the course of a career [1].
https://cew.georgetown.edu/wp-content/uploads/collegepayoff-completed.pdf (page 1)

Similar trends are seemingly present in Europe as a 2018 study by the U.K. The Department of Education found that the median salary of individuals with postgraduate degrees is about 10,000 pounds higher than that of non-graduates. https://assets.publishing.service.gov.uk/media/5cc0672040f0b640357127a5/GLMS_2018_publication_main_text.pdf (page 1)
Overall, it seems that higher education tends to result in significantly increased earnings over the course of one's career. Clearly, higher education can be incredibly beneficial. As a result, it is vitally important that the factors that contribute to success in higher education are fully understood, as these factors facilitate the attainment of university degrees and thus their long-term benefits. The purpose of this work is to find and discuss the factors that contribute to success in higher education. In order to find the factors, a dataset containing the personal, family, and educational information of 145 students from Cyprus' Near East University will be used in two ways. The first of which is a clustering analysis aiming to answer the question of whether or not grouping the students in the dataset based on their characteristics reveals any patterns related to academic success. In other words, clustering will be performed on the dataset in order to group together similar students based on their characteristics. The clusters containing the most students with higher grades will then be analyzed to find potential characteristics that contribute to higher grades. We hypothesize that there will be clear patterns relating to which characteristics contribute to academic success based on the clusters. The second way the data will be used is through a regression analysis with the purpose of more clearly answering the question of which characteristics most contribute to academic success. We hypothesize that the regression analysis will conclude that features like the number of hours studied weekly, the frequency at which academic reading is done, attendance to seminars and classes, the way exams are prepared for, and the frequency of taking notes in class have the most significant impact on students' success. The results of the regression analysis could ultimately support or oppose the results of the clustering analysis. After the completion of the clustering analysis, it is clear that our hypothesis is not fully supported by the results as the clusters formed did not result in many clear patterns. This being said, some patterns could still be derived from the clusters. The results of the multiple regression model posed interesting results which do not support our hypothesis. These results also opposed the results of the clustering analysis.

Materials and Methods:

As mentioned in the introduction, a dataset containing information about 145 students from Cyprus' Near East University was used to answer our research questions. This dataset can be found on the UCI Machine Learning Archive and is titled *Higher Education Students Performance Evaluation*. While conducting exploratory data analysis we found the grade

distribution of the dataset which is displayed in Figure X.



Clearly, grades between 50 and 59 are the most common while the fail grade is the least common. Before beginning the clustering analysis, some feature engineering was performed on the dataset to make it more readable. Prior to editing, every single feature in the dataset was represented by a number rather than text. Additionally, since all of the features are categorical, each value of each feature was also represented by a number. These two factors made the dataset impossible to understand when beginning to work with the data. As a result, every feature was replaced with text that described itself, and each value was also changed into being represented as text. Following this, the salary and flip-classroom features were removed from the dataset as the frequency with which salary was being paid was unspecified and the flip-classroom feature did not seem relevant to the analysis. After this, binning was performed on the final grade feature. Prior to binning, the final grade feature contained letter grades using the grading scale in Cyprus. In this system, each letter grade does not correspond to grade ranges of equal width. As a result, the final grade feature was altered to contain five bins of width ten with the first bin beginning at fifty and the final ending at 100. The only value that did not fit into a bin was the ‘fail’ value which indicated that a student failed their class. This value was left as is. Binning was ultimately done to make finding patterns from the clustering analysis easier to see and to make results easier to understand for the reader.

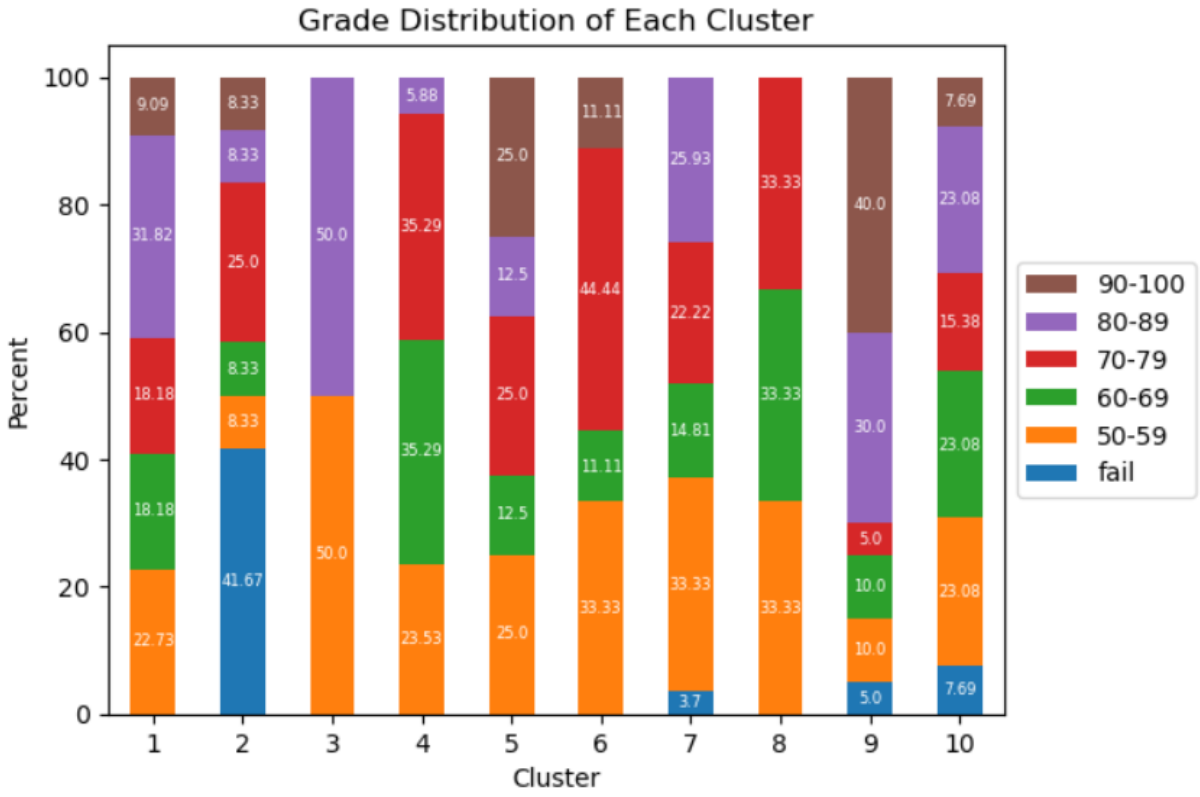
In order to simplify the clustering algorithm and to further make finding patterns easier, clustering was done three separate times on three different subsets of the data. The three subsets separated the features into three categories: personal, family, and education. The personal subset contained the features containing personal information such as age and sex. The family subset contained features pertaining to family matters such as number of siblings and parental marital status. Of course, the education subset contained features regarding

education such as weekly study hours and frequency of attendance to class. The final step completed before clustering was to numerically encode the ordinal categorical variables so that their relationship relative to each other could be better captured by the clustering algorithm. For example, when clustering the personal subset, the age variable was numerically encoded in the following way: the original feature contained '18-21', '22-25', and 'above 26' as its three values, and these values were changed to be 0, 1 and 2, respectively. These numbers were specifically chosen in order to avoid issues with scaling. All features that were numerically encoded were done in the same manner to avoid scaling issues. Of course, numerically encoding some of the features meant that now there were both categorical and numerical variables in the dataset. As a result, k-prototypes clustering was used to cluster the data. K-prototype clustering is a hybrid clustering technique commonly used when working with categorical and numerical data. This method uses Hamming distance for categorical variables and Euclidean distance for numerical variables. This method requires the use of a predetermined number of clusters. This number was in part chosen by using the elbow curve method which is displayed in Figure A1. Because the elbow curve method did not show one clear optimal number of clusters, we used domain knowledge in conjunction with the curve in order to choose ten as the number of clusters. Overall, ten seemed to be an optimal number of clusters given the amount of students in the dataset. In order to find patterns in the clusters, the percent frequency of each grade was plotted using a stacked bar plot. This made it obvious which clusters contained higher than average grades and thus the feature values in these clusters were analyzed for trends.

Before creating the multiple regression model, there was also a need for some preprocessing of the data. Notably, our dependent variable, final grade, had to be changed to be numerical so that we could use a linear regression model. We specifically chose to do a regression model here as we thought the results would be more intuitive when using a simpler model. Of course, in changing the final grade feature to be numeric, the predictive power of the model decreases due to us using estimated final grade values. For the purpose of our research question, we accepted this consequence as the main goal of the model was to find the most significant feature that relates to final grade, not to precisely predict final grade. For the independent variables, we decided to use all of the variables in the dataset except for our dependent variable, final grade, and the GPA feature as this feature is likely correlated to final grade. Because all of the independent variables were categorical, we also had to produce dummy variables in order to be able to use the model. In order to refine the model, we removed the least significant features (by p-value) one by one until only significant features remained. We used .05 as the level of significance.

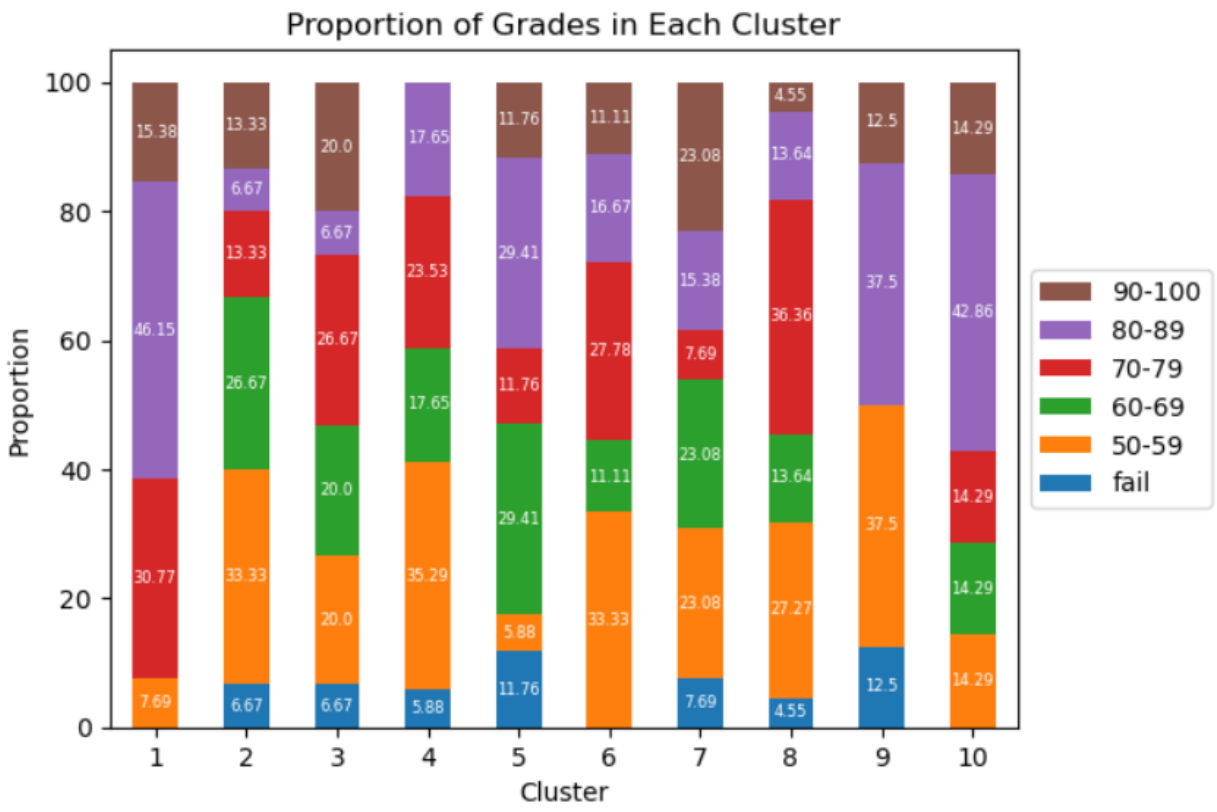
Results:

The personal subset of the data contained the following features: age, sex, type of high school, scholarship type, additional work (yes or no), additional activity (yes or no), partner status, transportation type, and accommodation type. Clustering the personal subset of the data yielded clusters with the grade distributions shown in Figure X.



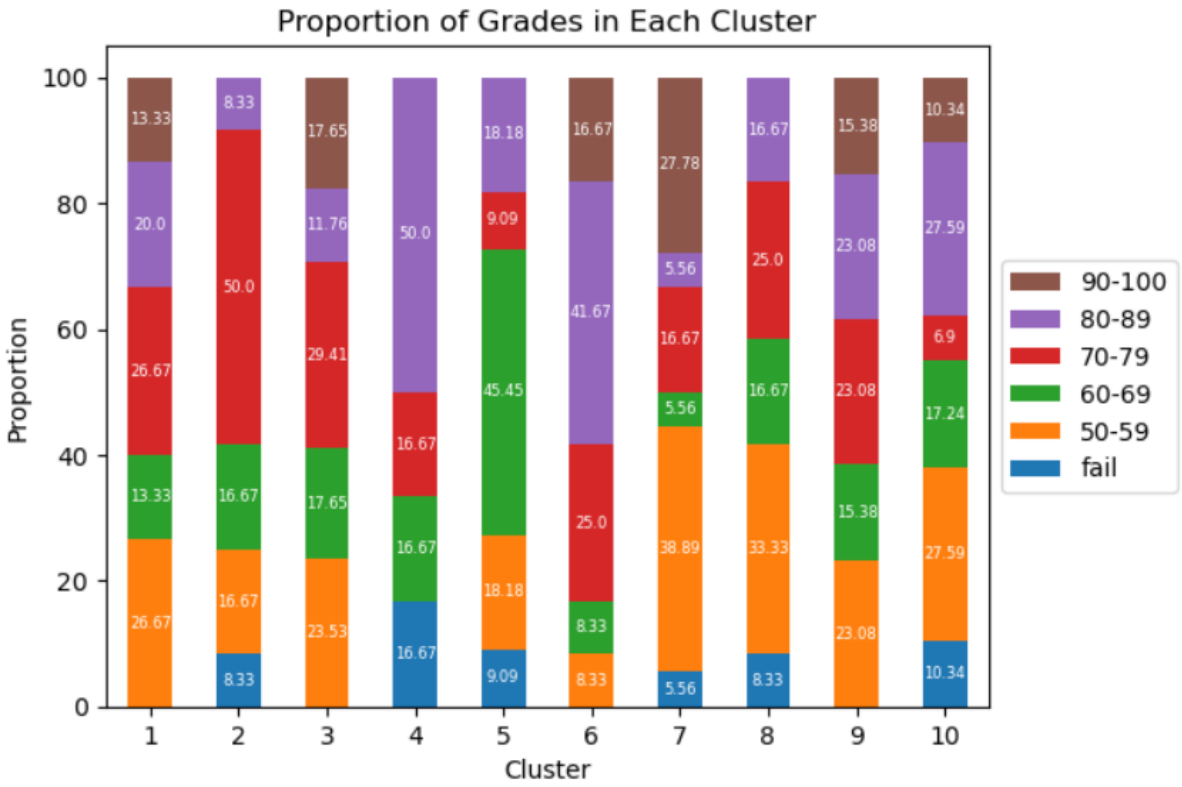
The above bar plot shows that clusters 1, 3, and 9 have the highest amount of high grades (we are considering any grade 80-100 a high grade). Analyzing cluster 1, we find that most of the students attended a state high school, had a 50% scholarship, did not partake in any additional work or activity, had a partner, took the bus to university, and lived in a rental accommodation. In cluster 3, most students had a 100% scholarship, did partake in additional work and activity, all had a partner, and lived in a college dorm. In cluster 9, most students attended a state high school, all had a 75% scholarship, did not partake in additional work or activities, did not have a partner, and took the bus to university. Overall, these clusters show the pattern that most of the students with high grades attended a state high school and had a scholarship of at least 50%. While this might suggest that these factors contribute to success in university, it is generally unclear if this is the case given that the majority of students in the entire dataset have these characteristics. One interesting pattern is that most of the students with high grades had a partner despite the fact that less than half of the students in the entire dataset share this characteristic. This could reflect the fact that having a partner often gives individuals more support, especially when away from home at a university. This increased support could then lead to increased motivation to obtain high grades.

The family subset contained the following features: highest level of education achieved by mother, highest level of education achieved by father, amount of siblings, parental marital status, mother's occupation, and father's occupation. Clustering the family subset of the data yielded clusters with the grade distributions shown in Figure X.



The above bar plot shows that clusters 1 and 10 have the highest percentage of high grades. The majority of students in cluster 1 had mothers who only attended primary school, fathers who attended high school, at least four siblings, married parents, and mothers who were housewives. Meanwhile, students in cluster 10 had more educated mothers as all attended high school or higher education. The fathers, however, were less educated as all only attended primary or secondary school. The majority of students in this cluster also had married parents, more than four siblings, and mothers who were housewives. Overall, the majority of students in the entire dataset had married parents and mothers who were housewives, so it is unclear how impactful these characteristics are in relation to achieving success in higher education. Additionally, the conflicting amounts of education for mothers and fathers in each cluster suggest that this also may not have a significant effect on academic performance. The most interesting pattern in these clusters is that the majority of the students that had high grades had at least four siblings as only about one-third of the entire dataset share this characteristic. This suggests a similar pattern to that of having a partner as having multiple siblings likely also increases the amount of support students have when at university. This ultimately could explain the increase in grades.

The education subset contained the following features: amount of weekly study hours, reading frequency (non-academic readings), academic reading frequency, seminar attendance, impact of projects, class attendance, how studying is done (alone, with a friend, etc.), when studying is done relative to exam date, note taking frequency, listening in class frequency, and impact of discussion. Clustering the education subset of the data yielded clusters with the grade distributions shown in Figure X.



The above bar plot shows that clusters 4 and 6 have the highest percentage of high grades. Cluster 4 contains students who read non-academic readings occasionally, often read academic readings, viewed projects as having a positive impact, studied close to the exam date, and sometimes listened in class. The students in cluster 6 read both academic and non-academic readings occasionally, attended seminars, almost always attended class, mostly studied alone and close to the exam date, almost always listened in class, and viewed projects and discussions positively. Overall, these clusters yielded conflicting results and thus do not contain many patterns. The patterns that they do show are present throughout the entire dataset and thus not indicative of obtaining higher grades.

const	101.4318	6.014	16.865	0.000	89.535
weekly_study_hours_>20 hours	-23.3825	7.129	-3.280	0.001	-37.484
reading_frequency_often	7.7147	3.391	2.275	0.025	1.007
high_school_type_private	-8.6266	3.029	-2.848	0.005	-14.619
scholarship_type_50%	-9.3574	2.456	-3.810	0.000	-14.216
scholarship_type_full	-10.2557	3.334	-3.076	0.003	-16.851
additional_work_yes	-8.5199	2.371	-3.593	0.000	-13.211
transportation_method_other	-11.8200	3.360	-3.518	0.001	-18.467
housing_rental	-4.8897	2.337	-2.092	0.038	-9.513
mother_education_masters	-17.0900	9.974	-1.713	0.089	-36.820
parental_status_divorced	-15.5404	6.511	-2.387	0.018	-28.419
parental_status_married	-12.7252	5.353	-2.377	0.019	-23.314
mother_occupation_self-employment	-18.2680	9.162	-1.994	0.048	-36.392

Figure X shows the results of the multiple linear regression model. According to the model, the only feature that had a significant positive relationship with final grade is reading

non-academic readings often. Meanwhile, multiple features were found to have a significant negative impact on final grade. These features are studying over twenty hours a week, going to a private high school, receiving a fifty percent scholarship or full scholarship, having additional work, living in a rental house, having a mother with a master's degree, having either divorced or married parents (as opposed to having one or more deceased parents), and having a self-employed mother. Ultimately, these results were very unexpected. This can likely be explained by multiple factors. Firsty, when checking the model assumptions, we found that the model was not fully accurate. Specifically, the model failed the assumptions of homoscedasticity and independence of residuals. This is likely due to the complicated nature of having many categorical variables in the dataset as predicting a continuous variable from multiple categorical variables tends to be difficult. Additionally, many of the variables had a high class imbalance which likely further caused the model to be inaccurate. One other potential factor is that the data may have been somewhat inaccurate in that it was collected from student responses.

Discussion:

In summary, the clustering analysis did reveal some patterns in regard to achieving high grades in higher education. For example, when clustering the personal subset of the dataset, we found a potential relationship between having a partner and being more successful academically. Also, clustering the family subset of the dataset revealed that having many siblings could also be related to obtaining higher grades. Ultimately, these two patterns point to a positive relationship between having support and being successful in higher education. Other than this, there were not many other particularly clear or interesting trends that the clustering analysis revealed. Because we only found a few seemingly meaningful patterns, we can conclude that our findings only somewhat support our hypothesis which stated that we would find patterns related to which characteristics contribute to academic success based on the clusters. Overall, the clustering showed that students with similar characteristics do not necessarily get similar grades.

Our regression model provided us with unexpected and likely inaccurate results. As a result, they were not the results we hypothesized. Our hypothesis was that features such as the number of hours studied weekly, the frequency at which academic reading is done, attendance to seminars and classes, the way exams are prepared for, and the frequency of taking notes in class would have the most significant impact on students' success. The results concluded that none of these features were significant. Additionally, the regression model also opposed the results of the clustering analysis as neither having a partner nor having many siblings were deemed to be significant.

Conclusion:

In summary, this analysis aimed to find which characteristics led to being successful in higher education. In a broader context, these findings could then be used to help students succeed academically and thus enjoy the benefits of graduating from higher education. Unfortunately, this analysis did not lead to any meaningful or conclusive results. As discussed, this may be due to limitations of the regression model and the data. One thing that we can conclude from our results is that human behavior is very distinct by nature. Human behavior is

difficult to predict because every person is unique. While there may be trends and patterns in certain data, truly being able to understand human behavior is a difficult task. To attempt to understand this topic better in the future, further research can be conducted. Firstly, the use of a more complex model than multiple linear regression would likely be more useful in answering our research questions given the complex nature of this task. Perhaps a feature importance evaluation of a random forest model would bear more meaningful results in regard to which characteristics most significantly affect student success in higher education. Additionally, it could be interesting to answer our research questions when using graduation to measure academic success rather than the final grade from one class. This would, of course, require the use of a different dataset.

Appendix: