Undergraduate Student Retention - 2013 Cohort

Yushan Juang, Amina Mahrouch

Embry-Riddle Aeronautical University

Table of Content

List of Figures

List of Tables

Abstract

This report applies data science principles and information visualization techniques to inform Embry-Riddle Aeronautical University on characteristics that may influence undergraduate student retention rates. Comparable characteristics in student retention by motivational, environmental, and financial factors may lead to an innovative solution to student retention.

Introduction

Student retention rates have been an important research topic in education, most prominently within universities. Internationally, universities have been developing new strategies to study possible ways to increase the rate of retention. This proposal intends to demonstrate retention rates (and therefore attrition) and identify success factors using information visualization techniques. The atomic analysis will investigate the persistent dynamic and common attributes contributing to retaining the 2013 first-time undergraduate degree-seeking freshmen cohort at Embry-Riddle Aeronautical University (ERAU). Identifying value-added qualities will influence new Enrollment Management Group (EMG) strategies that may profit future student success. ERAU's president, Dr. Butler, started a strategic initiative program focused on enrollment management, enhancing the student experience, and aiding in student success. With a motto that states, "Students First," it is beneficial for the university's academic reputation to understand what circumstances influence student retention. According to Institution Research (IR), retention data show an 11.3 percent increase in first-year retention from 2009 to 2018; however, many factors could contribute to these trends (i.e., environmental, psychological, cognitive).

To investigate the retention rates of undergraduate students, datasets were provided by the ERAU Raw-Data Warehouse (EDW) and Eagle Card services. The datasets contained student senses, financial data, campus solutions, and Academic Advancement Center (AAC) student utilization data. An overview of domain cleansing and exploration is required to present the correlations, variables, and distribution of clusters.

Varying factors can present challenges in detecting parallel influences and privacy restrictions, data overload, and biases. For example, standardized tests are another critical variable; nevertheless, it is apparent that some students take either the SAT, ACT or both; this can stand as a concern when creating an ethical correlation from the data frame. To overcome some of these challenges, the nature of the scope will focus on the cognitive, economic, and motivational aspects of the atomic retention analysis. Furthermore, student names, identification numbers, and other personal information was removed, and each student had one record observation per student ID. The process in which the data was cleansed will be discussed in more detail in the data preparation section.

The objective of this report will be to cover the process of investigating the possible factors that contribute to the graduation rate (retention rate) of students starting with the data preparation, which includes data reduction, transformation, dimensioning, and other essential attributes in preparing the

dataset. The Data Analysis Approach (EDA) will be referenced to demonstrate relationships between value-added parameters. Furthermore, a visualization conclusion will be presented and analyzed concerning cognitive, economic, and motivational factors containing variables that affect the retention trend and show the distribution and variance (uncertainty) based on the three C Principles: Correctness, Clarity, and Conciseness. Observations will be recorded and shown through different data visualization techniques.

Data Preparation

Data preparation is the most critical part of any information visualization study. To have a cohesive and reliable dataset, the data must be pre-processed, cleansed, tidy-ed, and combined. Initially, four datasets were provided by the Eagle Card office and EDW that include: Senses, Financial Aid, Academic Advancement, and Campus Solution.

The card swipe data from the Eagle Card office is used for maintaining the student Eagle Card service. The EDW dataset is used to provide overall functionality about student records. The Senses and Campus Solution data were combined. All personal information was deleted upon receiving the data resulting in three datasets: The senses and Campus Solutions (SCS) dataset, Financial Aid dataset, and the Academic Advancement Center (AAC) dataset.

Data Transformation and Cleansing

Financial Aid data show 927 records of students from the 2013 cohort, with the expected family contribution from 2013 to 2019.

AAC card swipe data from 2013 to 2019 provides 350,102 records with 11 variables. These variables include the date, campus ID, name, place where the card is swiped, account information, transaction, and the reader. Transformation for the card swipe data using the following procedure:

1. Extract only the date and ID of each record and only keep the records where the card is swiped in labs on campus—this part is done on Excel.
2. Import this dataset in R, then format the dates extracted from the raw data to convert the function since the original ones cannot be recognized by R.

The goal is to provide organized data so that each ID has a row where the number of one card swiped by the corresponding student each year is recorded; thus, the summarized function in R is used by the priority of ID and year of the date.

SCS dataset contains 6435 records of 937 students with 31 variables. These (student) variables include demographics such as the encrypted ID, gender, ethnicity, age, international status, veteran

status, college/- major, campus residency status, admit type, and term. In addition, the dataset provided measurable variable such as high school converted GPA and rank, the scores of standardized tests, academic year and a term indicating where the dynamic information belongs, cumulative GPA of undergraduate, graduate, and doctoral (if applicable), but most importantly, graduation status.

Within the SCS data, untidiness is detected in the dataset. The number of records for each ID is varied, contingent on how many semesters the students stayed at ERAU. Additionally, the static information of some students indicates to be inconsistent between records. Take the ethnicity data as an example; it is shown as 'Two or more races' in some records while 'Race/Ethnicity Unknown' in others belonging to the same student. Moreover, some unexpected blanks can be filled according to other records from the same student. The edge cases are detected as well; namely, the status of a student can vary between semesters like on/off-campus, admit type, and veteran status. The sparsity of master and doctoral data also increases the complexity of the cleansing process. The simplest way to manage master and doctoral data sparsity are by selecting the cases with admitting dates that span the Summer 2013 and Fall 2013 semesters.

At first glance, the anomalies of the dataset were not evident. Methods were proposed and implemented by R to detect such incongruities and manipulate the original dataset. Initially, the records were separated by each semester by the "filter()" function in R to render one record per student. Subsequently, the "full join()" function was performed on static variables. Due to the untidiness mentioned above, there will be more than one record in the standard dataset. Thus, the preferred way of tackling this untidiness is to go back to the original dataset and eliminate inconsistencies manually. It is an iterative task, running an R script and manipulating the dataset repeatedly until there is no overlapped record.

In the case of the inconsistency removed from the data, numerous edge cases would still be present, impeding the ongoing cleansing job. The variables that exhibit edge cases are major and on/off-campus. The notion is to select the first change, namely, only keep the original status and the status changed simultaneously with the change flag, the change date. For admit type, international student flag, and veteran type, only six students encountered such edge; thus, it is appropriate to delete without affecting the overall dataset.

Regarding the cumulative GPA at the decision point, the strategy is to keep the first and last-term GPA, regardless of graduation status. Lastly, the sparsity of all GPAs is avoided, keeping all with the justification that student time at ERAU various.

After such considerations, it is very straightforward to manipulate the data in R. First; a new data frame is created so that each student ID has one row. Then, a for loop is used to store all variables. For static variables, now it is safe to use the data in the first record of students since inconsistencies have been eliminated.

Some sparsity was found in important variables such as the standardized test scores and high school converted GPA when examining the dataset. They are essential for retention analysis and cannot be fabricated by interpolation techniques; therefore, cases without high school GPA and cases without at least one standardized test score will be eliminated. After such manipulation, there are still 857 observations. Lastly, the variables about standardized test scores are considered. If a student only takes the ACT, it can be converted into an SAT equivalency score by individual SATV and SATM equivalence. If one takes both tests, the conversion will also be made, and only the highest score will be kept. In order to join the datasets into a coherent database, the SCS data must be cleansed and combined with financial data and card swipe data by matching the ID.

Data Exploration

Exploration of data and finding trends in smaller groups is essential before data mining and visualization can be performed to select the correct tools in the analysis or reprocessing. Exploration Data Analysis (EDA) philosophy in the graphical analysis was applied to extract significant variables, evaluate underlying assumptions, and detect anomalies. This approach has many advantages in the formulation of valid engineering and scientific conclusions. A combination of all stated datasets has been flattened, resulting in the following parameters: Ethnicity, Academic Program, College, Veteran Status, High school GPA, Standardized Test Scores, Estimated Family Contribution (EFC), First-Semester GPA, Undergraduate Degree GPA, Retention by year, Graduation Status, AAC, International Status, and Age. A decomposition of the cohort in sub-cohorts of similar demographics and parameters has been demonstrated to show the yield observations. Subsequently, consideration within cognitive, environmental, and financial factors are examined and displayed using R, Tableau, and Excel for visualization and techniques such as correlation, distribution, and variance.

Overview

The overarching goal of this analysis is to report on trends that will assist the management and operations departments at ERAU in identifying any underlying trends of student retention.

Figure 1. Overview of graduation rates

Figure 1 displays the total student body analyzed in this report. Five hundred fifty-three of which graduated are shown in green, and the remaining 304 who did not, are shown in red. Moreover, the graduation rate broken down by major can be observed in Figure 2. The data shows Aeronautics majors (COA) and Space Physics majors (COAS) retention rates are the lowest, with only 20% and 25% of students graduating, respectively.

Figure 2. Graduation rates by major

It is to be advised that the data in which these were taken used the initial declared major. It can be seen in Figure 2. that approximately 65 percent of students graduate across the board.

Demographic

Before interpreting the factors contributing to the overall objective, an overview of the data is provided to explore possible leads and clusters of similar trends that can then point to a more refined observation. These observation parameters include various groupings of age, sex, ethnicity, international status, major, and graduation rates.

Error! Reference source not found. displays the undergraduate count per college grouped by sex a nd national status generated using Excel, where International Resident is denoted by "I," and National Resident is denoted by "N."

Table 1. Percent of total enrollment demographic by sex and national status

College	Female (%)	Male (%)	Total (%)
Arts and Science	2.92	3.50	6.42
I	0.12	0.12	0.23
N	2.8	3.38	6.18
Aviation	6.77	33.37	40.14
I	0.58	1.63	2.22
N	6.18	3.174	37.95
Business	1.52	1.75	3.27
I	0.58	0.70	1.28
N	0.93	1.05	1.98
Engineering	8.75	41.42	50.18
I	0.70	2.57	3.27
N	8.05	38.86	46.91
Total	**19.95**	**80.05**	**100.00**

Furthermore, the span of student demographics throughout the four colleges: College of Aviation (COA), College of Engineering (COE), College of Arts and Sciences (COAS), and College of Business (COB) is shown in Figure 3.

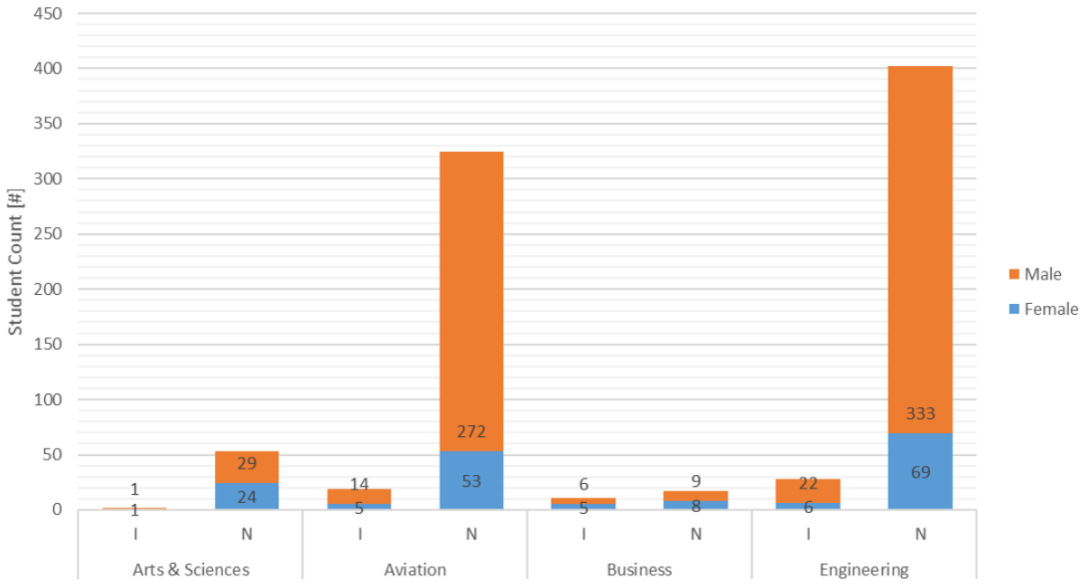

Figure 3. Student Enrollment by sex, college, and national status

It is apparent from Figure 3 that there is a higher number of male students compared to female students and a large variation in national and international student counts, more prominently in COA and COE. The majority of student enrollment is in engineering (430) and aviation (344), making up more than half of the total enrollment. To get a breakdown of the overall distribution in enrollment, refer to Table 1 above. COE and COA enrollment makes up approximately 90 percent of the total enrollment.

Now that the spread of clusters has been defined, a general idea of enrollments within the colleges produced using a similar approach as Figure 4; the graduation rates broken down by sex and ethnicity have been illustrated using Tableau. It is to be noted that the ethnic group "American Indian/Alaskan Native" and "Native Hawaiian/Pacific Islander" have been omitted from Figure 4.

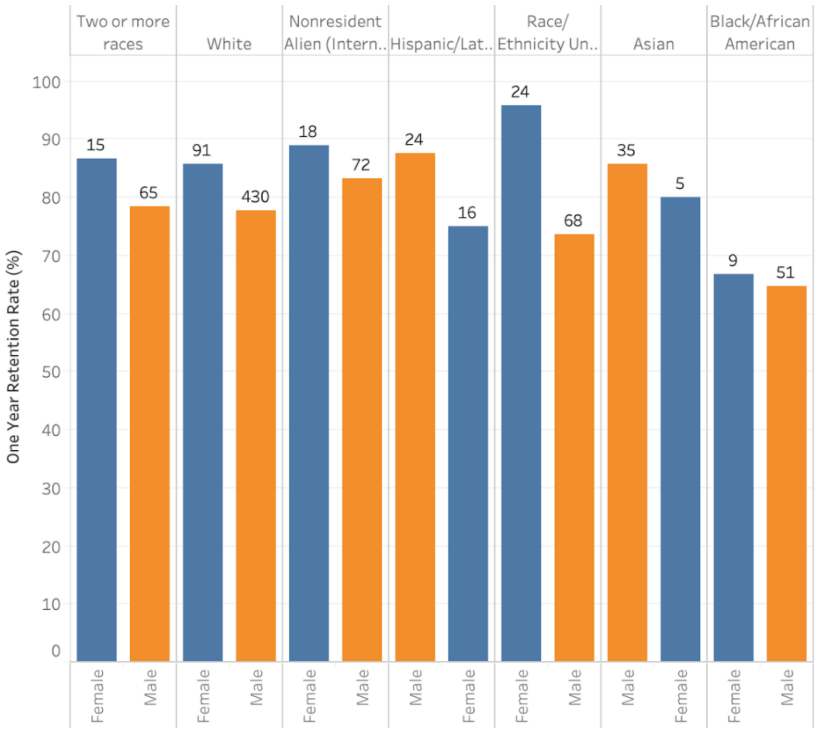

Figure 4. Graduation rate by sex and ethnicity

The observed trends seen in Figure 4 provide an interesting insight into the spread of graduation rates within the different ethnic groups, along with the female graduation percentage showing higher than male for five out of the seven groups. However, in the Nonresidential Alien and Hispanic/Latino ethnic group, the male students had a higher graduation percentage. There is still not enough information to determine any possible root cause(s).

Cognitive Factors

Cognitive factors in education have been proven to impact student academic performance, especially during college. Such factors can be present in the students' expectations for success based on previous success in high school while maintaining the same work ethic. The cognitive variables in this study include First-term GPA, Highschool (HS) GPA, and Standardize Test Scores. However, a more in-depth observation starts with a correlation between numerical data, as shown in Figure 5.

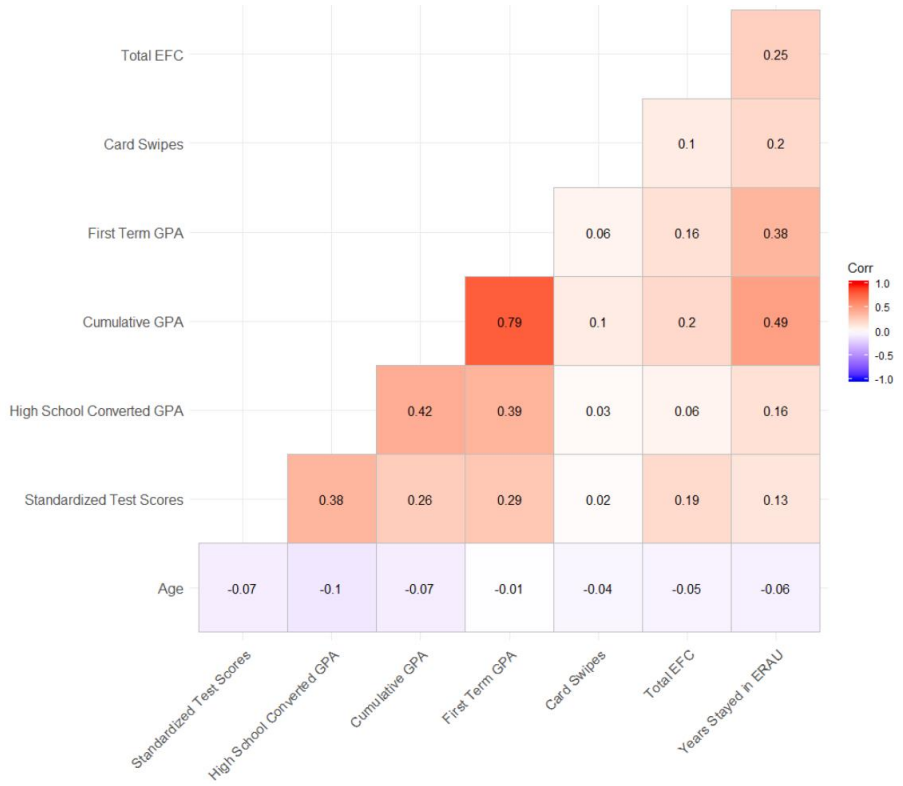

Figure 5. Numerical correlation of cognitive variables

A correlation between the HS GPA and first-term GPA can be seen immediately. Moreover, the correlation in HS GPA versus the first-term GPA by graduation status can be seen in Figure 6.

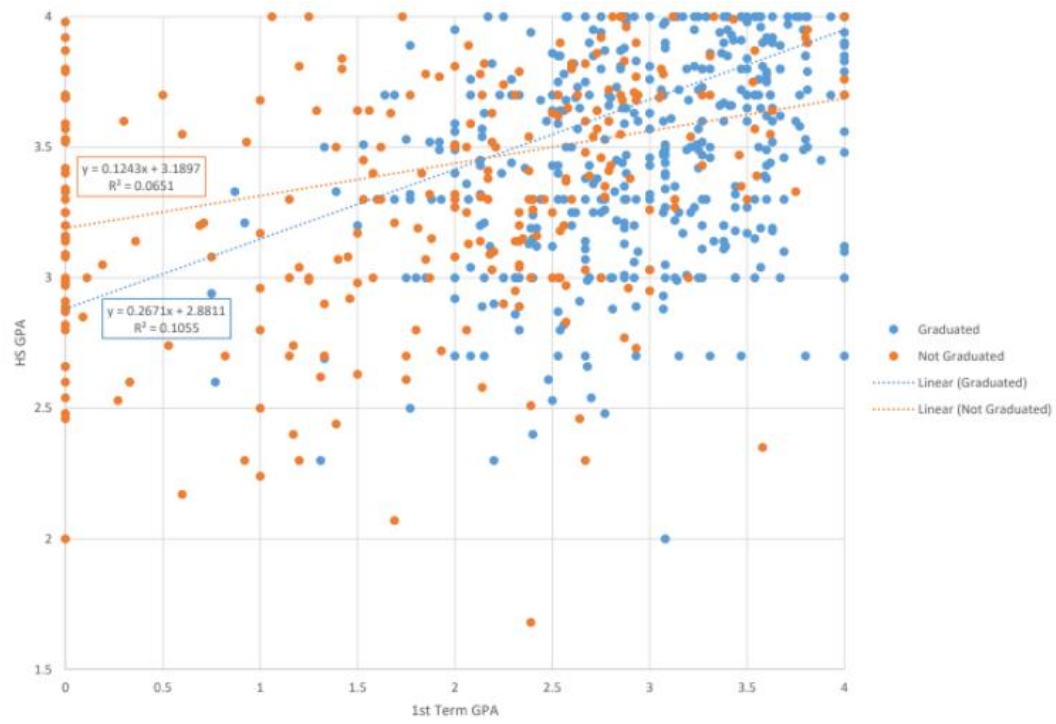

Figure 6. Correlation between high school GPA and first-term GPA by graduation status

It can be observed that the students who had an HS GPA much greater than the first term were more likely to not graduate, with some anomalies, of course. Furthermore, students who had both their HS and first-term GPAs greater than 3.0 are shown to have a higher graduation rate.

Motivational Factors

Motivational factors can predict a students' success rate, i.e., the student who shows initiative in on-campus resources like SGA, tutoring, or holding an on-campus job, is predicted to be more likely retained. This trend can be seen in the AAC data by the overall graduation rate. Figure 7 displays that 84% of students who went to the AAC at least once in their academic career graduated.

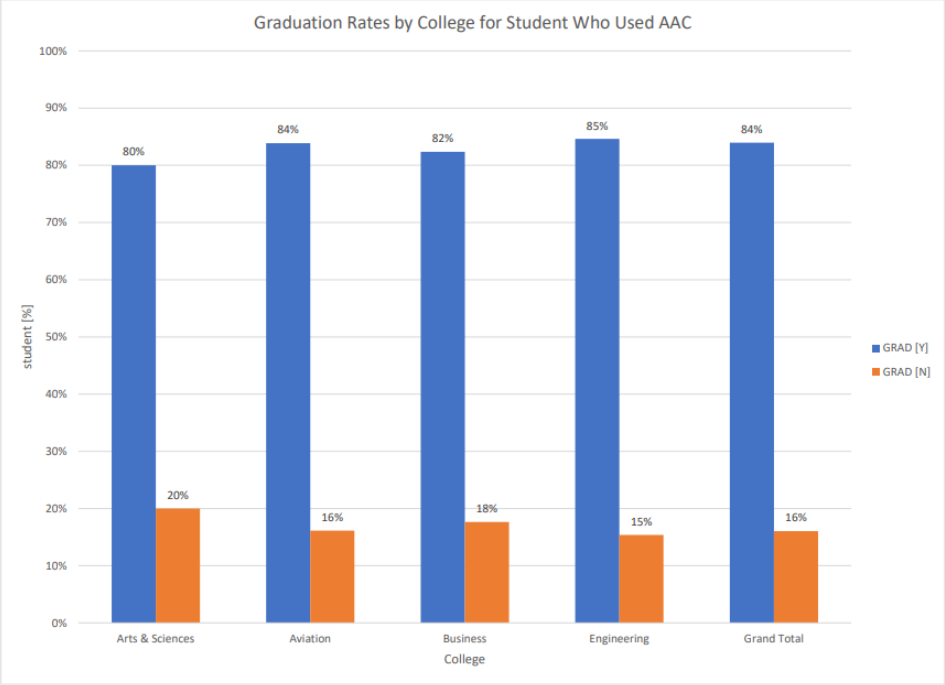

Figure 7. Graduation rates by AAC [Y] and college

This academic correlation can also be demonstrated from the AAC [Y] data subtracted from 100 percent, as seen in Figure 8, which displays the graduation rate of a student who did not attend the AAC during their time at ERAU.

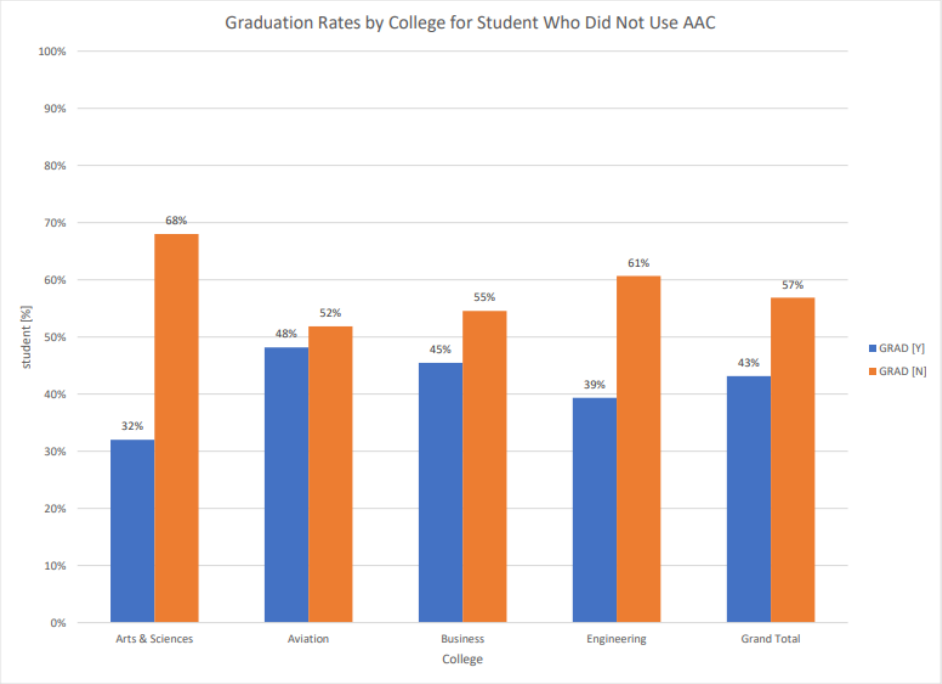

Figure 8. Graduation rates by AAC [N] and college

Referring to Figure 7, it is apparent that students who did attend the AAC had a total average of 84% graduation rates with minor deviations and anomalies. In Figure 8, the students show a non-motivational demeanor concerning tutoring, had a 43 percent graduation rate. This motivation factor in students doubled the graduation percentage. An illustration of the entire graduating students who used the AAC services (in green) and those who did not (in red) can be observed in Figure 9.

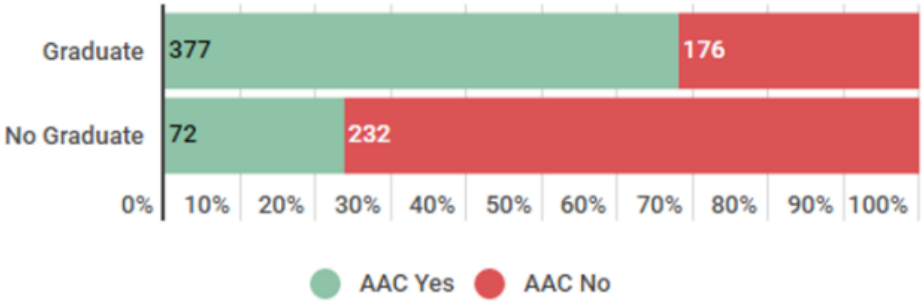

Figure 9. Graduation rates by AAC overall

A key takeaway from this is that these services can indeed help with retention, as more than 75 percent of students who graduated did use the services AAC provides. In contrast, the same trend seems to follow for the number of students who did not graduate and did not use the AAC.

In addition, Figure 10 shows the distribution of cumulative GPA and graduation status. The data speaks volumes as below a certain threshold, around the 2.0 mark, there is a greater propensity for students not to graduate. In contrast, most of the distribution when it came to a' Yes' in graduation status came from individuals with an average cumulative GPA of around and above 3.0.

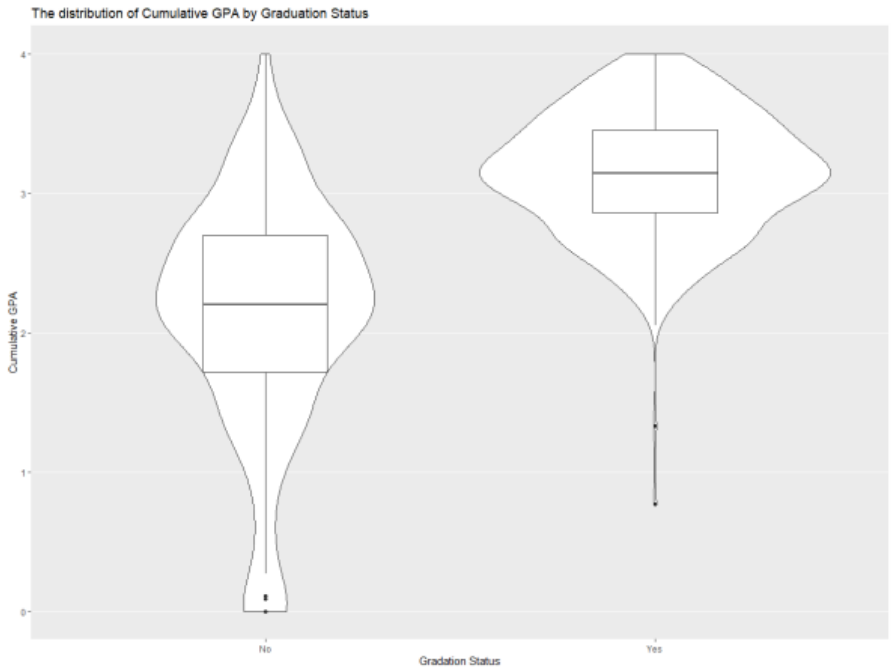

Figure 10. Cumulative GPA by graduation status

Lastly, as shown in Table 2 below, there seems to be a significant correlation between students who changed majors and graduating regardless of whether the change occurs from STEM to non-STEM or vice-versa.

Table 2. Graduation rate by major change and STEM/non-STEM

	Graduated	**Not Graduated**	**Total**
Major Changed	231	56	287
STEM - STEM	123	31	154
STEM - NON-STEM	102	24	126
NON-STEM - STEM	6	1	7
Did Not Change Major	302	248	570
Total	**533**	**304**	**857**

In many ways, this information may be due to the motivation factor stated above. It is assumed that if a student changes their major, then that shows that the student is motivated to graduate, find the right program for them, and has enough drive to explore another possibilism before quitting.

Financial Factors

Financial stress can play an enormous role in academic behavior and retention rates. For instance, if a student has little to no family contribution, then the likelihood of graduation is far lower. However, like many other factors, this is not the one-size-fits-all, rather a general statement.

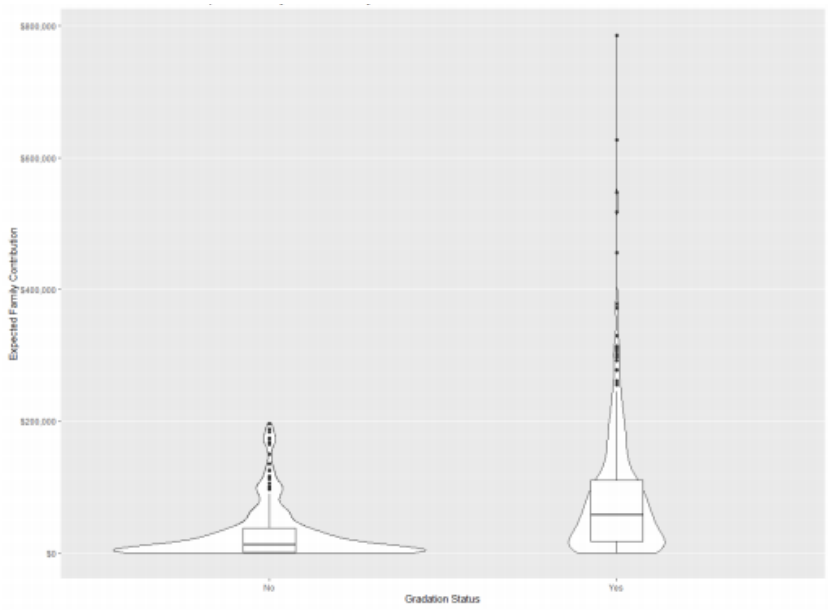

Figure 11. Total estimated family contribution by graduation status

Figure 12 illustrates the EFC for the 2013 cohort graduation status. It can be observed that, on average, the students who graduated had a larger amount of EFC than the student who did not.

Conclusion

In conclusion, many factors can play a role in the retention rate in education. Some trends seem more probable than others, such as the Academic Advancement Center utilization and the cognitive factors associated with the first-term and high school GPA. Nevertheless, there are still many more factors and backgrounds that were not taken into account. For example, with more data, the generations constantly change with environmental, social, and psychological factors; it is hard to pinpoint an exact answer or just one factor. It is useful to use some of the visuals shown to make informed decisions on what to continue and implement throughout the upcoming semesters and track changes based on environmental and motivational factors.

References

Steven Lehr, Hong Liu (2016) *Journal*, Use Educational Data Mining to Predict Undergraduate Retention.