ANALYSIS OF FACTORS CONTRIBUTING TO STUDENT'S ACADEMIC PERFORMANCE IN COLLEGE'S STEM PROGRAM

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

COLLEGE OF ARTS AND SCIENCES

BY

ABIMBOLA AYENI, B.S.

DENTON, TEXAS

MAY 2022

DEDICATION

I am dedicating this thesis project first to God Almighty for giving me the grace and wisdom to choose the right counsel in guiding me during the duration of this thesis, knowing I have come to the end of the master's degree program gives me pure joy.

To my mother, Dr, Mrs. Eniola Ayeni, I love you. Thank you for being the shoulder to cry on, the ears I rant to when all I needed was a listening ear to unburden the stress of graduate school.

To my sister, Oluwatoyin Ayeni, we made it; bagging this master's degree means we bagged it both. I love you.

ACKNOWLEDGEMENTS

I have immense gratitude for every committee member. Dr Sides has been the best teacher anyone could ask for. With every question I asked about confounding concepts, he always had an applicable life scenario to relate the explanation of the concept to, and that made understanding it easier. Dr. Falley, I am in awe of you. You are shrewd in your field, and you easily relate the complex concepts for students to understand. Dr. Wheeler, thank you for guiding me throughout the complex and confusing regulations that the university mandates. Having you in the committee has helped me ensure the correct processes are followed every step of the way. Finally, Dr. Hamner, I am grateful for the guidance you gave during the beginning of this process. You laid the foundation upon which this thesis was founded.

ABSTRACT

ABIMBOLA A. AYENI

ANALYSIS OF FACTORS CONTRIBUTING TO STUDENT'S ACADEMIC PERFORMANCE IN COLLEGE'S STEM PROGRAM

MAY 2022

The purpose of this study is to analyze different variables in understanding how they affect the academic performance of students taking STEM courses at a mid-sized college in the southern United States. These factors are classified into pre-college variables and college variables. The pre-college variables are ACT/SAT scores, gender, ethnicity, and first-generation student or not. The college variables are college GPA, Pell grant recipient, the student's continuance with the degree major in sophomore year. Implementing regression analysis and using backward regression to remove variables that are not correlated to the response variable (CGPA), the researcher was able to arrive at the best model using the three variables: ACT mathematics, ACT English, and gender. The resulting variables selected show how the value of R^2 has helped in making sense of the model selected by 14.2%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The most essential asset for any educational institution is its students. The students' academic performance helps in making sure that quality graduates are produced who then, in turn, become leaders and help contribute to the social and economic growth of any country (Abaidoo, 2018). Several researchers have shown enormous interest in the analysis of a student's academic performance, factors affecting them, and different approaches that can be implemented to help improve the overall academic performance of these students.

Significance of the Study

According to Merhi et al. (2018), the first 2 years of higher education (with focus on the first year) have been identified as the most important and defining years for students' success, resilience, and retention with regards to academic performance and life after school. During this time, the students will learn or unlearn skills and develop attitudes, whether positive or negative, towards their academic courses, which eventually shapes their future engagement in said fields. This in turn leads to the development of perceptions about their personalities, which contributes to their character definition as students.

Unfortunately, the dropout rates are higher during these times, creating an increasing amount of research that focuses more on the dynamics of student's engagement and the methods to use in implementing new tactics that would improve academic performance during these formative years (Ribeiro et al., 2019). Although the general dropout rates are high, there is an unbalanced ratio with those in the science, technology, engineering, and mathematics (STEM) courses. The United States has a lower ratio of STEM to non-STEM undergraduate degree completion (fewer than 1 in 6), and STEM jobs are speculated to increase by 17% in the next

decade in comparison to non-STEM jobs (Beede et al., 2011). Also, the percentage of students that declare STEM-related majors in college is on a continuous decrease; approximately 30% of students declare a STEM major in college (Chen, 2009; Chen & Ho, 2012), and this percentage reduced even further because many students start off with a STEM major and decide to switch to a non-STEM related major either in their freshman or sophomore year.

This study aims to analyze the academic performance of the population focus of Texas Woman's University freshmen and sophomores in the United States, and it uses factors such as pre-college and college variables to have an informed conclusion about how students strive to achieve good academic performances while enrolled in STEM courses.

Purpose of the Study

The purpose of this study is to analyze different variables in understanding how they affect the academic performance (where success is defined in this research as students who continued their major from freshman year to the sophomore year and maintained or improved their GPA) of students (where the focus would be on freshman and sophomores) taking STEM courses (with focus on chemistry, microbiology, biochemistry, mathematics, and computer science courses) at a mid-sized college in the southern Unites States. These factors are classified into pre-college variables and college variables listed below.

The pre-college variables include:

- ACT/SAT scores
- Gender
- Ethnicity

- First generation student or not (First generation is defined as the parents having no college degree and Not first generation is defined as the parents having some level of college degree)

The college variables include:

- College GPA

Research Questions

While implementing this study, the following research questions were addressed:

- Are students' grades in STEM courses significantly different in their first-year vs second year?
- Are students' academic performance significantly different by the student's family's educational level (that is first generation or not)?
- Are students' grades in STEM courses significantly different when comparing demographic data?
- What factors lead to better academic performance in STEM courses?

CHAPTER II

LITERATURE REVIEW

History of Academic Performance

Academic performance assessment dates as far back as 2000 BC when government officials were selected by examinations in China. The academic performance assessment process surfaced again in the fifth century; this time, teachers in Athens included questions that were made for the purpose of evaluation in their pedagogical toolkit. Now, fast forward to modern academic assessment, which originated in the United States and dates to the 1830s. Educational advocates Horace Mann and Samuel G. Howe conducted a standardized test that they used to evaluate student's performance and progress in Boston (a standardized test that later became known as the Boston Survey; Tian & Sun, 2018).

Between 1887 and 1898, Joseph Rice assessed a large school district of 33,000 students' spelling abilities and argued that these students did not make substantial progress in learning which eventually led to the invention of the unit of measurement for academic performance in 1900 by Edward Thorndike. For this unit, Edward used a scale design that implemented the statistical principles "equidistance and psychological measurement" that eventually birthed the academic performance assessment scale, which was based on quantifiability, objectivity, and standardization. In 1983, the National Commission on Excellence in Education recommended students take a core academic performance assessment scale, which was based on quantifiability, objectivity, and standardization. In 1983, the National Commission on Excellence in Education recommended students take a core academic curriculum called the New Basics, which led to an increase in enrollment from 42.5% to 68.8%. Thus, standardized academic performance assessments such as the Scholastic Assessment Test (SAT) in 1934, the American College Test (ACT) in 1959 (Tian & Sun, 2018), and many others emerged.

With the above, one can see that academic performance plays an important role in promoting educational equity, and without an assessment for academic performance it would be challenging to know if basic education has been achieved by the students. In the context of modern education, academic assessment has specific features of its own. This specificity lies in its close connection with people's efforts to promote equity and quality in education, to adjust education policy, as well as to make classroom interventions (Tian & Sun, 2018).

Academic performance can thus be defined as measuring a student's ability to learn and implement the knowledge into achievement across various academic subjects. This measurement is typically done by using factors such as classroom performance, graduation rates of the students, and results from standardized tests. Noble (1991) discovered that using a singular factor to determine academic success was not as efficient as combining two or more factors.

History of STEM in the United States

STEM is an acronym for the science, technology, engineering, and mathematics fields of education. It can be traced back to the Morrill Act of 1862 but was introduced in 2001 by the scientific administrators at National Science Foundation (NSF) with the previous acronym of SMET. In no time, reports were published showing that United States schools were not having students enroll and graduate from STEM programs as much as other countries, thus creating more focus on how to solve the disparity. In no time, programs like Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA) were implemented. In 2006, PISA's result showed that the United States ranked 21st in a 30-country panel on assessment of scientific knowledge and competency. With this, a list of 10 action points was developed with its top three recommendations being the following:

1. Increase America's talent pool by improving the quality of K-12 science and mathematics education;
2. Strengthen the teacher's skills with additional training in sciences, mathematics, and technological courses; and
3. Increase the number of students prepared to enter college and graduate with STEM degrees.

CHAPTER III

METHODOLOGY

Data Collection

The data collected for this research was de-identified data from the Institutional Research and Improvement Office of the college under study. The data consists of three First-Time-In-College (FTIC) cohorts; that is, three sets of undergraduate students entering college for the first time and starting in the fall semester (Fall 2016, Fall 2017, and Fall 2018). The grades provided for these students span over two full years (fall, spring, and summer equate to one full academic year). The variables in the dataset are briefly explained below:

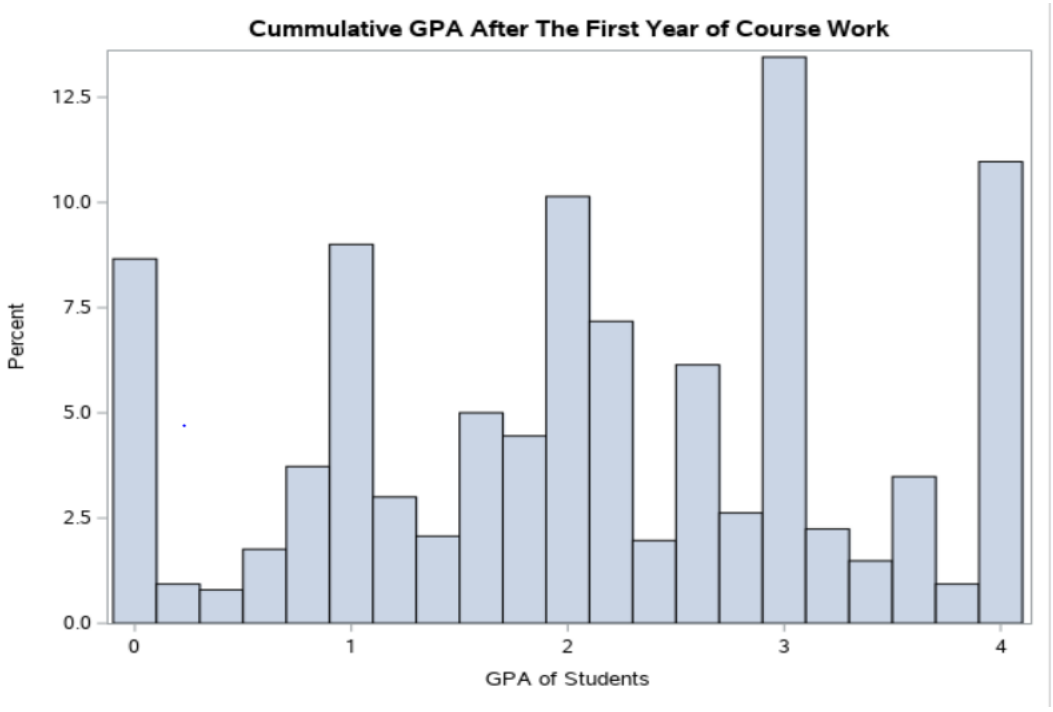
- SAS_STU_ID: this is a quantitative variable that represents each student's identification number. The ID number has been de-identified.
- YEAR: as mentioned earlier, the dataset consists of three different start years (2016, 2017, and 2018). Each dataset consisted of two academic years (including the summer semester). To easily differentiate between each start year, the dataset was classified according to the start year. For example, 2016 would represent students whose first semester started in fall of 2016.
- SAT and ACT Scores: this is a quantitative variable. It represents SAT Verbal and Mathematics scores as well as ACT English and Mathematics scores
- FULL_TIME_COHORT: this variable represents students who are enrolled full time and those that are enrolled part-time. This is a binary variable.
- PERSIST: this variable represents the students who chose to continue with their degree program in the sophomore year with the university. They are represented by 1, while the students who dropped out after their first year are represented by 0.

- ETHNICITY: This is a qualitative variable. For this dataset, ethnicity is broken into five groups (Asian, Black, Hispanic, White, and Others) with the highest population being Hispanic students with 41.24% of the population.
- GENDER COHORT: This is a qualitative variable with the values represented as either male (M) or female (F). The dataset has a larger population of female students (91.5%) and a smaller population of male students (8.5%).
- PELL_COHORT: This represents students who have not received a bachelor's degree and have a high documented financial need. In addition to the aforementioned conditions, for a student to continue receiving these funds, the said student will be limited to a maximum of 12 full time semester hours. For this dataset, students in this category are represented with a 1, where 0 represents students that have not been awarded these funds.
- STEM_MAJOR: STEM courses in this college are offered in various departments (courses include but are not limited to biological science, physical science, mathematical sciences, computer and information science, geosciences, engineering, and technology areas associated with the preceding fields). One of the college goals is to increase the number of enrollments into the STEM field, and to achieve this several incentives are attached to students who enroll for these courses (affirming the need to analyze students' performance in this field of study). The STEM_MAJOR variable is a binary variable with a yes/no indicator for students who are (or are not) in a STEM major.
- GPA: This is the cumulative grade point average of each student. The starting semester is represented as CUM_GPA_1; this is the GPA before the commencement

of course work that semester (for example, if starting semester is Fall 2016, then CUM_GPA_1 is the GPA the student has before the end of course work for their first semester). Most students in this fall semester have a blank starting GPA (indicating no GPA). Students whose GPAs are not 0 have college credit going into their first semester. SP_CUM_GPA_1 represents the cumulative GPA entering a student's second semester of course work, and the remaining variables follow a similar pattern. It should be noted that most of the students who started their first semester with dual credit or AP credits have GPAs between 3 and 4 (representing a left-skewed histogram). After the academic course work of the first year is completed and with the inclusion of students entering the university without college credit, more spread can be seen in figure 1 below.

Figure 1

Histogram of Students Cumulative GPA after the First Year



Statistical Analysis

Data analysis was conducted using statistical software. Regression analysis was used in modelling the relationship between variables and to calculate the correlation coefficients of variables being considered.

CHAPTER IV

RESULTS

Description of Results

Descriptive statistics were used to describe the basic features of the data. First, a quantitative variable that allows for easy statistical computation was selected: SAT and ACT scores. Using the statistical analysis system (SAS software), the mean, median, and standard deviation of SAT scores (by type) and ACT scores (by type) by cohort, STEM major status, and gender were calculated.

Using the variable Year to classify the results, the mean, median, and standard deviation of SAT scores were analyzed. Table 1 shows, we see that student SAT scores for verbal and mathematics improved over the years with the increase in the mean and median values across the three years. Further, the standard deviations are smaller as the years increase, indicating the data points move closer together over time. Box plots are shown in Figure 2.

Table 1

Scholastic Assessment Test (SAT) Verbal and Mathematics Scores for 2016, 2017, and 2018

Analysis Variable	Cohort	Mean	Median	Standard Deviation
SAT Verbal	2016	462.02	450	81.58
	2017	497.41	500	75.29
	2018	536.21	540	73.39
SAT Math	2016	469.07	460	82.12
	2017	484.72	490	76.35
	2018	522.92	520	76.28

Figure 2

Boxplots of SAT Verbal and Mathematics Scores for 2016, 2017, and 2018

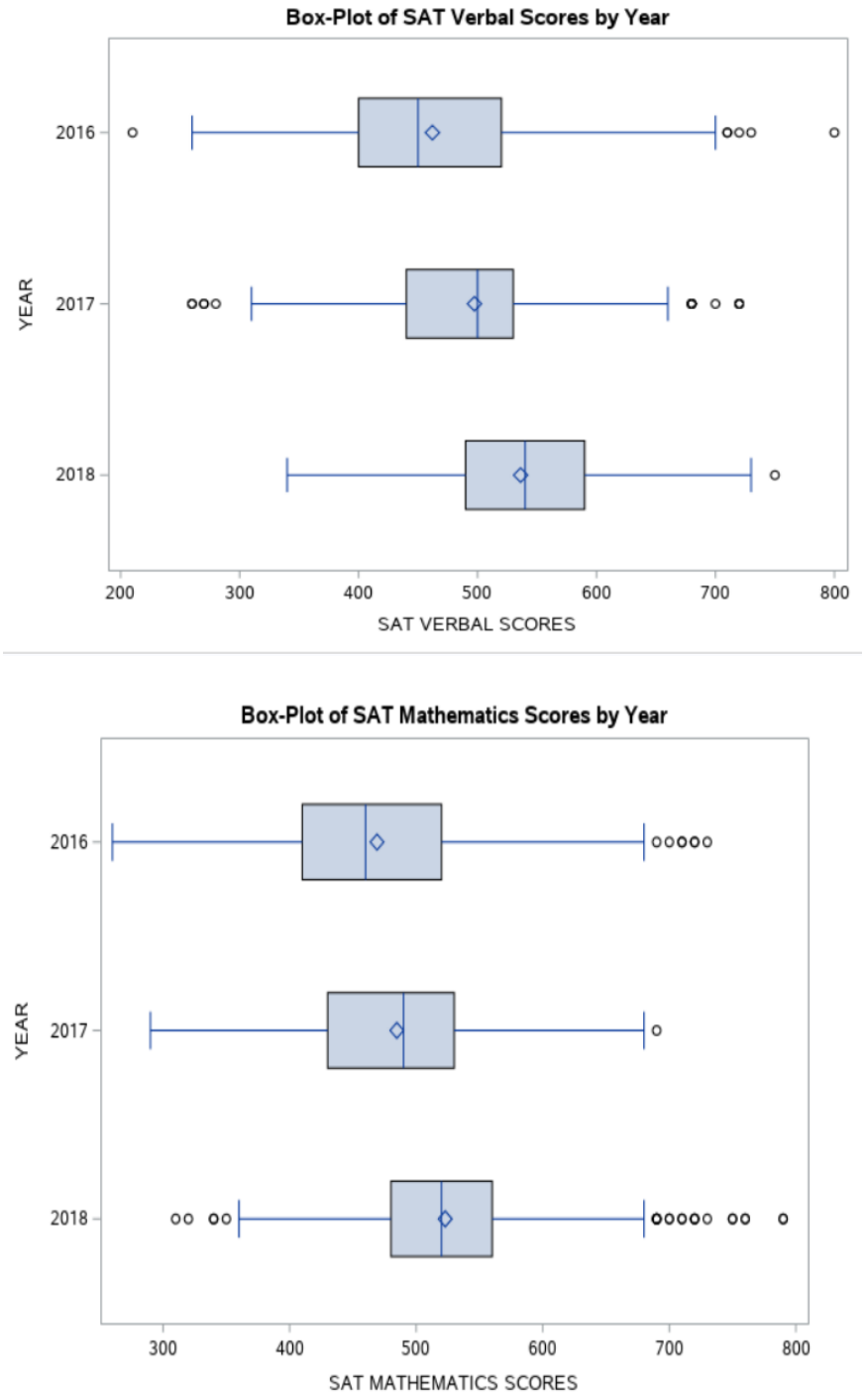


Figure 3 shows summary statistics (that is, minimum score, mean, standard deviation, and maximum value) for SAT and ACT scores during the 3 years of data. From this table, the lowest

and highest SAT verbal score admitted occurred in 2016 while for SAT mathematics, the lowest was in 2016 and the highest recorded score happened in 2018. The ACT English scores registered its lowest score in 2018 and the highest in 2016 and 2018. ACT mathematics recorded the lowest and highest score in 2017.

The same analysis of SAT and ACT scores was done against the STEM major variable (see Table 2). From these results, STEM majors on average have a higher mean score for all forms of the SAT and ACT.

Figure 3

Summary Statistics of ACT and SAT Scores by Years

YEAR	N Obs	Variable	Label	Mean	Std Dev	Minimum	Maximum	N
2016	1053	SAT_V	SAT_V	462.0240964	81.5841312	210.0000000	800.0000000	830
		SAT_M	SAT_M	469.0722892	82.1234005	260.0000000	730.0000000	830
		ACT_M	ACT_M	19.8595041	3.9923442	13.0000000	31.0000000	484
		ACT_E	ACT_E	19.2685950	5.0396118	7.0000000	36.0000000	484
2017	967	SAT_V	SAT_V	497.4132492	75.2852157	260.0000000	720.0000000	634
		SAT_M	SAT_M	484.7244094	76.3545066	290.0000000	690.0000000	635
		ACT_M	ACT_M	19.8706897	4.0584928	9.0000000	34.0000000	464
		ACT_E	ACT_E	19.1875000	5.1291501	7.0000000	35.0000000	464
2018	920	SAT_V	SAT_V	536.1983471	73.3861533	340.0000000	750.0000000	726
		SAT_M	SAT_M	522.9201102	76.2754479	310.0000000	790.0000000	726
		ACT_M	ACT_M	19.9881235	4.0127596	12.0000000	33.0000000	421
		ACT_E	ACT_E	19.6104513	5.2199512	6.0000000	36.0000000	421

Table 2

Average Scholastic Assessment Test (SAT) and American College Test (ACT) Scores Analyzed by Science Technology Engineering and Mathematics (STEM) Major

	Not STEM Major	STEM Major
SAT Verbal	496.51	498.96
SAT Mathematics	490.24	498.80
	Not STEM Major	STEM Major
ACT English	19.33	19.83
ACT Mathematics	19.41	20.41

Next, using gender, Table 3 shows that male students have higher mean scores in both types of standardized testing than the female students on both examinations.

Table 3

Average Scholastic Assessment Test (SAT) and American College Test (ACT) Scores Analyzed by Gender

	Females	Males
SAT Verbal	495.37	523.10
SAT Mathematics	488.15	527.45
	Females	Males
ACT English	19.27	19.96
ACT Mathematics	19.75	21.78

Another variable of interest used in descriptive analysis is GPA. The initial data set used for this research provided the researcher with the student's overall GPAs (stated earlier), which was calculated based on all courses that a student has completed. Since this research focuses on STEM courses (chemistry, biology, biochemistry, mathematics, and computer science) that the students completed during their first and second year in college, the GPA for STEM only was calculated. To get the students' STEM GPA for the first two academic years, the researcher first had to arrange all courses according to first fall, first spring, and first summer semester, and then align the second-year courses as was had done with the first years. Next, the course grades were extracted from the dataset. On the SAS platform, using conditional if-else statements, the extracted grades were converted for all STEM courses in the dataset to its equivalent number representation. That is, A is equivalent to a 4, B is equivalent to a 3, C is equivalent to a 2, D is equivalent to a 1, F is equivalent to a 0, and every other grade is regarded as null.

Next, the number of class hours was included for each STEM course enrolled in by the students. Using Excel, the last digit in the course codes (which represents the class hours for the course) were extracted; for example, MATH XXX3 has 3 class hours (3 being the last numeric value in the course code). A column was then created to store the number of hours for each course. Using SAS, the next step was to multiply the number of hours by the letter grade (which by now has been converted to its numeric equivalent) and then divide that by the summation of all course hours for each student. Mathematically, this calculation is represented by

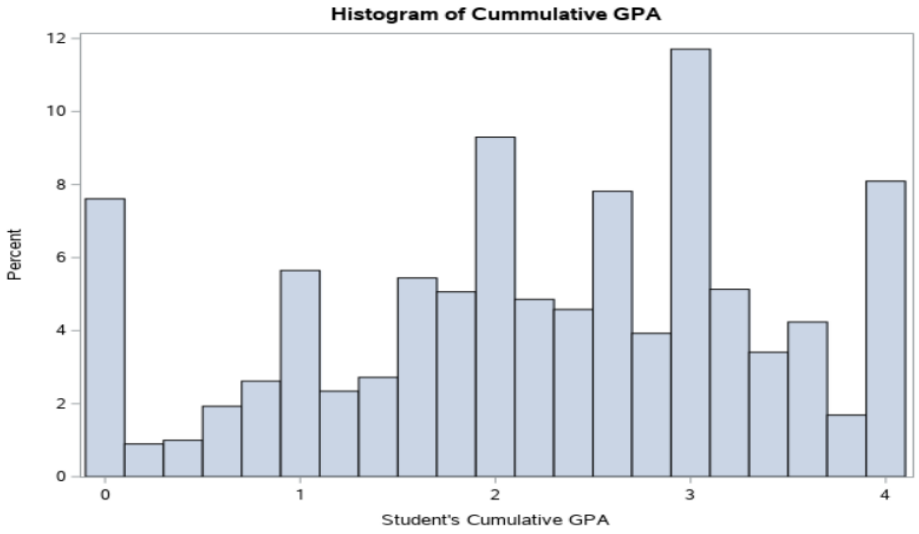
$$STEM\ GPA = \sum_{i=1}^{n} \frac{h_i\ g_i}{h_i}$$

$$= \frac{h_1 g_1 + h_2 g_2 + h_3 g_3 + \cdots h_n g_n}{h_1 + h_2 + h_3 + \cdots + h_n} \qquad (3.1),$$

where n represents the number of STEM courses a student takes in a given time period. After Equation 3.1 was implemented, the cumulative GPA for students with regards to the STEM courses was obtained; this is displayed in the histogram in Figure 4.

Figure 4

Cumulative STEM GPA of Students



Testing STEM Majors against Non-STEM Majors

The idea of a two-sample test is to compare two population averages by juxtaposing the independent samples. The samples in this case were STEM major students (mean value represented by μ_1) and non-STEM major students (mean value represented by μ_0), and the goal was to test if the STEM GPA of both categories of STEM major students are different. That is, the researcher tested $H_0 : \mu_0 = \mu_1$ vs $H_1 : \mu_0 \neq \mu_1$ (the mean of the two populations are the same vs there is a difference between the first and second mean).

Next, the assumptions of this test were checked. For a two-sample *t*-test, the first assumption stated that there must be one continuous dependent variable (which in this case was the STEM GPA) and one categorical independent dichotomous variable (STEM major).

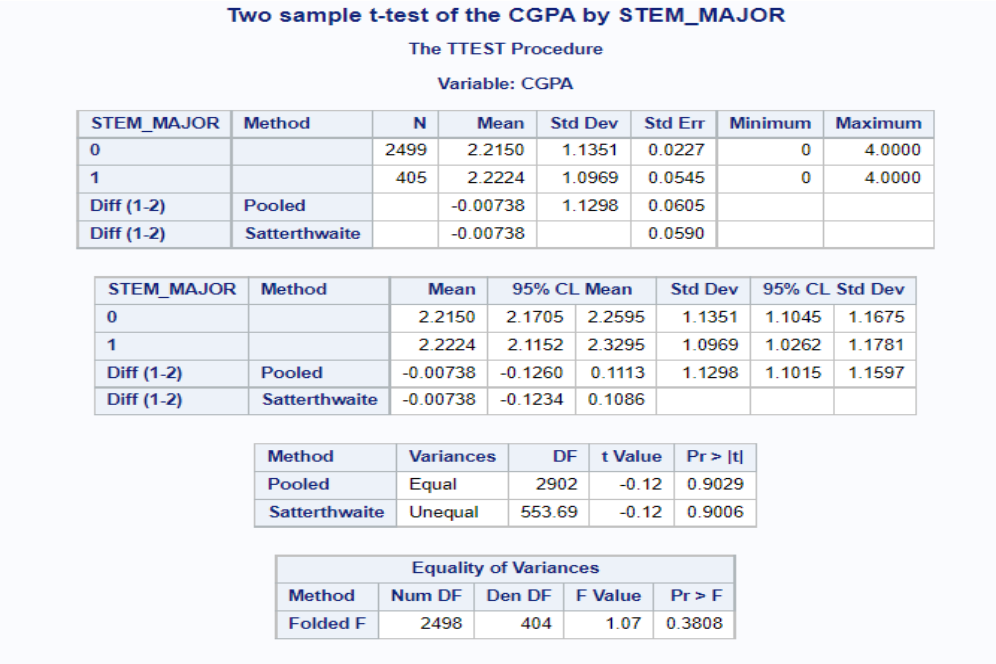
Secondly, the samples of interest must be independent (which they were). Lastly, the researcher confirmed if both sample means followed a normal distribution. While the normality assumption could be checked by plotting histograms of the data to see the shape of the initial populations, another method to confirm normalcy of the sample means was by invoking the Central Limit Theorem. In this case, n_0 (the sample size of non-STEM major students) equaled 2531 and n_1 (the sample size of STEM major students) equaled 409; thus, the researcher confirmed normality of both sample means.

Next, the researcher determined if they want to use the pooled or Satterthwaite method, which were appropriate depending on whether or not there were equal variances. The pooled method was preferred because it had a higher power and should be used when it cannot be concluded that the variances are unequal. Thus, a separate test is run to determine this, where the researcher can use the pooled method if the *p*-value under the equality of variance section is greater than the value of alpha. The Satterthwaite method is used in the case of unequal variances (the *p*-value is less than the value of alpha). The results as seen in Figure 5 showed the *p*-value to be 0.3808 (greater than the value of alpha = 0.05, indicating that the researcher could not conclude unequal variances), so the researcher was allowed to run a pooled variance *t*-test.

With this, the mean STEM GPA of non-STEM majors $(\bar{x}_0)$ was 2.22 and the mean of students who are STEM majors $(\bar{x}_1)$ was 2.22. After running the *t*-test using a pooled variance, the test statistic was given as -0.12 and *p*-value of 0.9029. Because the *p*-value of 0.9029 was greater than any reasonable choice for alpha, the decision was to fail to reject the null hypothesis. The conclusion would then be that the researcher could not conclude that the average STEM GPA of students who are STEM majors is different from those who are not STEM majors.

Figure 5

Two Sample Testing Result



Two sample t-test of the CGPA by STEM_MAJOR

The TTEST Procedure

Variable: CGPA

STEM_MAJOR	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		2499	2.2150	1.1351	0.0227	0	4.0000
1		405	2.2224	1.0969	0.0545	0	4.0000
Diff (1-2)	Pooled		-0.00738	1.1298	0.0605		
Diff (1-2)	Satterthwaite		-0.00738		0.0590		

STEM_MAJOR	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		2.2150	2.1705	2.2595	1.1351	1.1045	1.1675
1		2.2224	2.1152	2.3295	1.0969	1.0262	1.1781
Diff (1-2)	Pooled	-0.00738	-0.1260	0.1113	1.1298	1.1015	1.1597
Diff (1-2)	Satterthwaite	-0.00738	-0.1234	0.1086			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2902	-0.12	0.9029
Satterthwaite	Unequal	553.69	-0.12	0.9006

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2498	404	1.07	0.3808

Testing Ethnicity

ANOVA (Analysis of Variance) is a test done as a measure of knowing whether there is any statistical difference between the means of three or more independent groups. This test was best suited for the mean STEM GPA values by ethnicity since the data showed five ethnic groups (where Table 4 shows their means). The following hypotheses were tested:

H_0 : *All population means are the same*

H_1 : *At least one population mean is different*

Table 4

Resulting Mean Cumulative Grade Point Average (CGPA) Values from Data Mining Process for Ethnicity

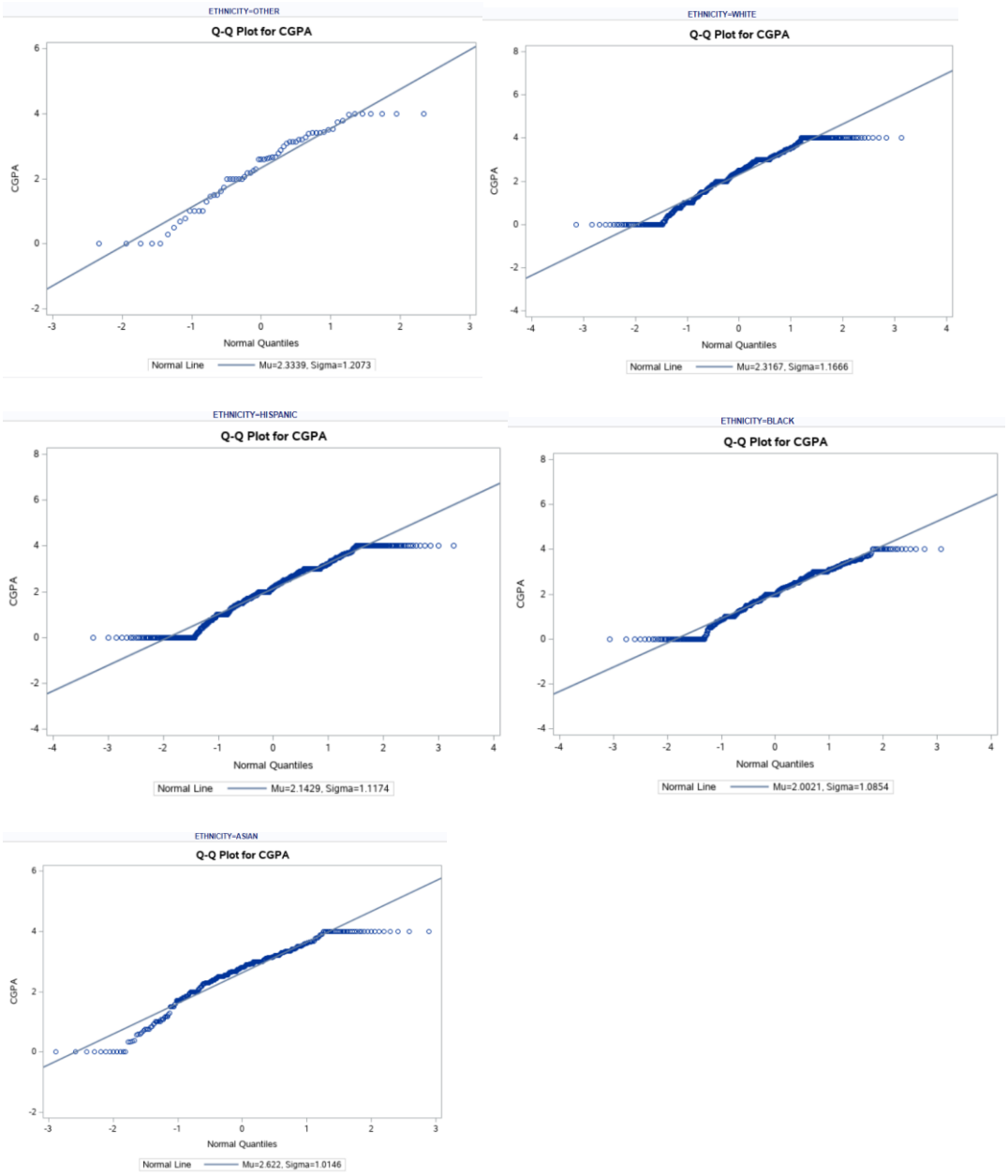
Variable	Variable Values	Mean Variable
Ethnicity	Asian	2.62
	Black	2.00
	Hispanic	2.14
	White	2.32
	Other	2.33

The first assumption for the ANOVA testing stated that a dependent continuous variable (GPA) and an independent categorical variable (ethnicity) were needed. The next assumption required the confirmation of the normality for each of the groups; that is, if the samples were normally distributed or not (here the researcher must note that the ANOVA test was relatively robust even when data was not normally distributed). The researcher implemented Q-Q plots to explain better the nature of normality for the categories (see Figure 6).

From the resulting Q-Q plots above, there was some bunching around the GPA points 0 and 4, and this created non-normality in the groups, likely due to the number of hours students in the dataset had. A normal dataset of GPA would have at least 120 hours, but this dataset focused on the first 2 years of STEM academic work (with many students having 3, 4, or 6 hours). So, the researcher concluded from the Q-Q plot that the STEM GPAs by ethnicity were not exactly normal, but they also do not deviate too much from normality. Because this was a loose assumption, the researcher was okay to proceed with this assumption.

Figure 6

Q-Q Plot of CGPA by Ethnicity



The next assumption required the confirmation of the equality of variances across groups. The researcher carried out a test of homogeneity of variances by implementing the hovtest command in SAS, with its output represented as Levene's test of homogeneity in Table 5. The researcher assumed homogeneity of variance (that is failing to reject the null hypothesis that

states that all variances are equal) if the *p*-value was greater than alpha and failed to assume homogeneity of variance (by concluding the alternative hypothesis of unequal variances) if the *p*-value was less than alpha. Because the *p*-value is 0.0059, the researcher cannot conclude that there were equal variances; since this was one of the ANOVA assumptions, this lack of equal variances might have affected the accuracy of the results.

Table 5

Levene's Test of Homogeneity of Cumulative Grade Point Average (CGPA) Variances

Source	DF	Sum of Square	Mean Square	F- Value	Pr > F
Ethnicity	4	30.19	7.55	3.63	0.0059
Error	2893	6015.9	2.08		

The results of the ANOVA test are shown in Table 6. The ANOVA testing gave an F-statistic of 19.24 and a *p*-value smaller than 0.0001. With the *p*-value less than alpha, the researcher rejected the null hypothesis and concluded the alternative hypothesis that at least one of the groups has a different mean than the other groups.

Table 6

Analysis of Variance (ANOVA) Testing of Cumulative Grade Point Average (CGPA) by Ethnicity

Source	DF	Sum of Square	Mean Square	F- Value	Pr > F
Ethnicity	4	95.62	23.91	19.24	<.0001
Error	2893	3593.55	1.24		
Corrected Total	2897	3689.17			

Next, the researcher performed the post hoc test known as pairwise comparisons; a pairwise comparison was made among all pairs of ethnic groups to determine where there was statistical significance. The result of the pairwise comparisons (see Figure 7) show that of all pairwise comparisons that could be made among the groups, there was statistical significance between the Asian and White, Asian and Hispanic, Asian and Black, White and Hispanic, and White and Black ethnic groups.

Figure 7

Post Hoc Test on CGPA by Ethnicity

Comparisons significant at the 0.05 level are indicated by ***.

ETHNICITY Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
ASIAN - OTHER	0.28811	-0.13014	0.70637	
ASIAN - WHITE	0.30530	0.10338	0.50723	***
ASIAN - HISPANIC	0.47913	0.28999	0.66827	***
ASIAN - BLACK	0.61988	0.41019	0.82958	***
OTHER - ASIAN	-0.28811	-0.70637	0.13014	
OTHER - WHITE	0.01719	-0.38234	0.41672	
OTHER - HISPANIC	0.19102	-0.20220	0.58424	
OTHER - BLACK	0.33177	-0.07174	0.73528	
WHITE - ASIAN	-0.30530	-0.50723	-0.10338	***
WHITE - OTHER	-0.01719	-0.41672	0.38234	
WHITE - HISPANIC	0.17383	0.03079	0.31687	***
WHITE - BLACK	0.31458	0.14529	0.48387	***
HISPANIC - ASIAN	-0.47913	-0.66827	-0.28999	***
HISPANIC - OTHER	-0.19102	-0.58424	0.20220	
HISPANIC - WHITE	-0.17383	-0.31687	-0.03079	***
HISPANIC - BLACK	0.14075	-0.01306	0.29457	
BLACK - ASIAN	-0.61988	-0.82958	-0.41019	***
BLACK - OTHER	-0.33177	-0.73528	0.07174	
BLACK - WHITE	-0.31458	-0.48387	-0.14529	***
BLACK - HISPANIC	-0.14075	-0.29457	0.01306	

Regression Analysis

Regression analysis models the relationships between a response variable and one or more independent predictor variables. It provides an equation that allows for predictions about future data, and, most importantly to the purpose of this study, potentially explains why students succeed or fail in their first two years in STEM courses. Since the researcher considered more than one variable, the multiple regression model was implemented. The first step was to represent the variables in a scatter matrix, which helped to determine if there was a linear correlation between the variables that were supposed to be independent. The scatter matrix in Figure 7 shows the response variable CGPA (cumulative STEM GPA) and with some continuous independent variables: ACT English, ACT mathematics, SAT verbal, and SAT mathematics.

With the scatter matrix, the researcher had the strongest positive correlation between SAT and ACT mathematics and the SAT and ACT English (which could indicate multicollinearity). This discovery makes sense as one would expect that a student that does well in SAT mathematics also does well in ACT mathematics and vice versa. The effect of this relationship is considered later when assessing the model. For the regression model, the researcher included the binary variables that were left out of the scatter matrix (a scatter matrix including binary variables shows nothing interesting) and the variables mentioned earlier.

With any regression model, there are issues that might occur that need to be addressed; one of these concerns is multicollinearity. Multicollinearity is defined as a statistical phenomenon where there is the existence of a strong relationship between predictor variables; simply put, when predictors are highly correlated, there is multicollinearity. This is considered a problem if present in the model whose primary purpose is explanation as opposed to prediction

because while the predictions will remain largely unchanged, the model does not know which of the correlated variables to associate changes in the response variable to.

Figure 8

Scatter Matrix of Continuous Variables

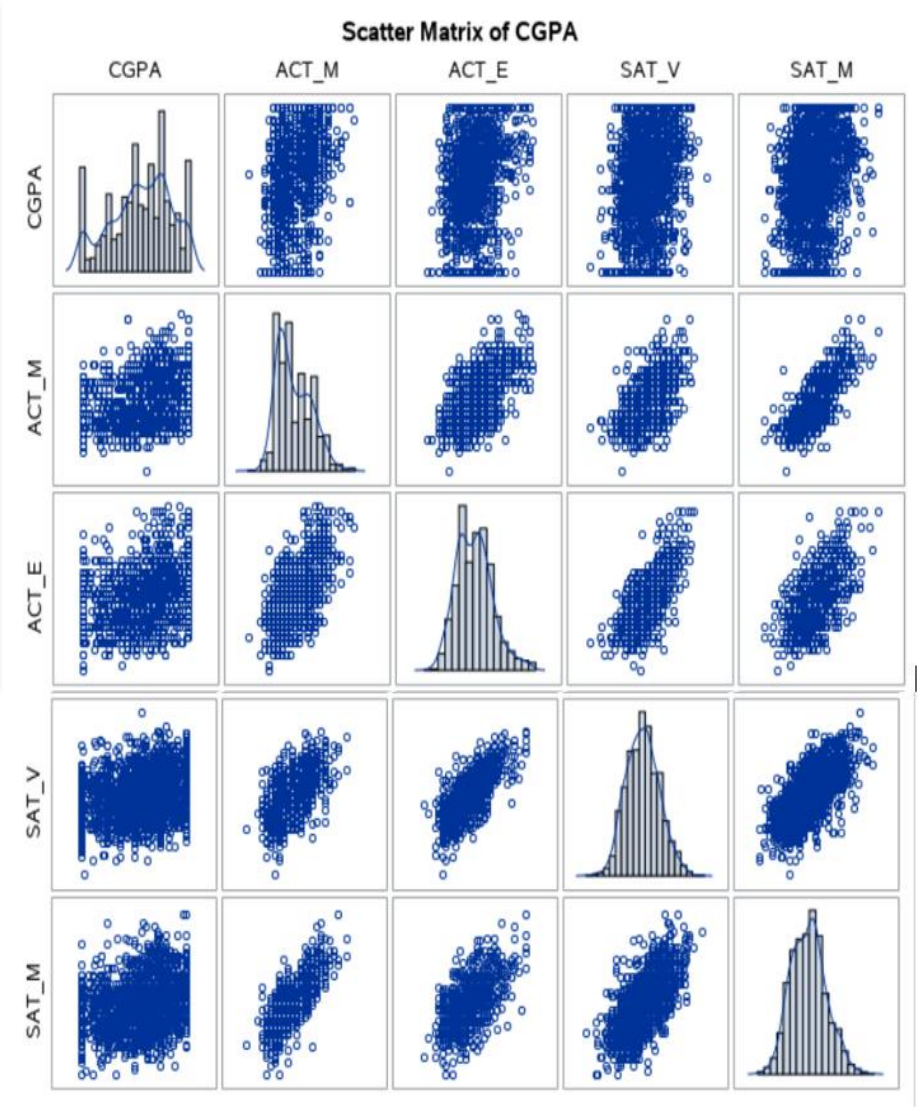


Figure 9

Output of the First Regression Model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	196.53433	24.56679	28.47	<.0001
Error	748	645.46621	0.86292		
Corrected Total	756	842.00053			

Root MSE	0.92894	R-Square	0.2334
Dependent Mean	2.36808	Adj R-Sq	0.2252
Coeff Var	39.22737		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.15442	0.29235	-0.53	0.5975
ACT_M	ACT_M	1	0.04205	0.01420	2.96	0.0032
ACT_E	ACT_E	1	0.03494	0.01071	3.26	0.0012
SAT_V	SAT_V	1	-0.00059054	0.00063483	-0.93	0.3526
SAT_M	SAT_M	1	0.00086651	0.00072955	1.19	0.2353
GENDER_COHORT	GENDER_COHORT	1	0.24537	0.12853	1.91	0.0566
FIRST_GEN_NODEG_0	FIRST_GEN_NODEG_0	1	0.05722	0.07891	0.73	0.4685
Pell_cohort	Pell_cohort	1	-0.07252	0.07857	-0.92	0.3563
PERSIST_1	PERSIST_1	1	0.84019	0.08632	9.73	<.0001

Because the model for this research was conducted to help explain the variables that contribute to a student's success in STEM courses, the researcher did not want there to be multicollinearity between the variables because the presence of multicollinearity would result in instability of the regression model estimates for the coefficients.

The researcher first explored whether the model had multicollinearity by implementing a correlation matrix. With the results of the correlation matrix in Figure 10, the researcher wanted to watch out for variables with large correlation values. Noting that all the standardized test scores had reasonably large correlations was an indicator that at least one of them should be removed from the model.

The researcher then proceeded further into examining multicollinearity by implementing the Variance Inflation Factor (VIF) command in SAS. When checking for multicollinearity, the researcher wanted to make sure that no VIF was above 10, but preferably none were above 3. As seen in Figure 11, the largest VIF in this model was SAT mathematics (with VIF of 3.13), which confirmed that the researcher needed to explore eliminating at least one of the math and verbal based standardized scores.

To find a better model, the researcher implemented backwards regression on the first model that contained all the variables by eliminating variables that were not significant and had a high *p*-values, which helped the researcher arrive at a model with the best adjusted R^2 value. Doing this, the researcher removed the SAT variables, which improved the adjusted R^2 from 22.5% to 22.6% (and solved the multicollinearity concern of having too many standardized test variables). Continued implementation of the backward regression system by removing first-generation status, Pell grant, and the persist variables gave the best model with an adjusted R^2 of 14.2% (see Figure 12).

The best model is thus:

$$\widehat{CGPA} = -0.28 + (0.08 * ACT_M) + (0.02 * ACT_E) + (0.58 * G) \qquad (4.1),$$

where T_M , ACT_E, G represents ACT mathematics, ACT English, and gender variables, respectively. The resulting variables selected show how R^2 has helped make sense of the model selected. The value of R^2 tells the percent of the variability for CGPA, which was explained by the model; the caveat was that about 86% of the variability was unaccounted for from the model. Therefore, the variables that help explain a student's STEM course performance are ACT mathematics, ACT English, and gender.

Figure 10

Output of the Correlation Matrix

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	CGPA	ACT_M	ACT_E	SAT_V	SAT_M	GENDER_COHORT	FIRST_GEN_NODEG_0	Pell_cohort	PERSIST_1
CGPA	1.00000 2904	0.34314 <.0001 1354	0.28839 <.0001 1354	0.23479 <.0001 2164	0.31014 <.0001 2165	0.08506 <.0001 2904	-0.12380 <.0001 2823	-0.08559 <.0001 2904	0.36310 <.0001 2904
ACT_M ACT_M	0.34314 <.0001 1354	1.00000 1369	0.64115 <.0001 1369	0.57212 <.0001 777	0.76865 <.0001 777	-0.13289 <.0001 1369	-0.23133 <.0001 1345	-0.21999 <.0001 1369	0.14800 <.0001 1369
ACT_E ACT_E	0.28839 <.0001 1354	0.64115 <.0001 1369	1.00000 1369	0.70949 <.0001 777	0.60945 <.0001 777	-0.03403 0.2082 1369	-0.32252 <.0001 1345	-0.30291 <.0001 1369	0.11209 <.0001 1369
SAT_V SAT_V	0.23479 <.0001 2164	0.57212 <.0001 777	0.70949 <.0001 777	1.00000 2190	0.67154 <.0001 2190	-0.05915 0.0056 2190	-0.24209 <.0001 2128	-0.21392 <.0001 2190	0.09999 <.0001 2190
SAT_M SAT_M	0.31014 <.0001 2165	0.76865 <.0001 777	0.60945 <.0001 777	0.67154 <.0001 2190	1.00000 2191	-0.13320 <.0001 2191	-0.21402 <.0001 2129	-0.18627 <.0001 2191	0.12300 <.0001 2191
GENDER_COHORT GENDER_COHORT	0.08506 <.0001 2904	-0.13289 <.0001 1369	-0.03403 0.2082 1369	-0.05915 0.0056 2190	-0.13320 <.0001 2191	1.00000 2940	0.01093 0.5592 2858	0.04384 0.0175 2940	0.04155 0.0243 2940
FIRST_GEN_NODEG_0 FIRST_GEN_NODEG_0	-0.12380 <.0001 2823	-0.23133 <.0001 1345	-0.32252 <.0001 1345	-0.24209 <.0001 2128	-0.21402 <.0001 2129	0.01093 0.5592 2858	1.00000 2858	0.36832 <.0001 2858	-0.02140 0.2528 2858
Pell_cohort Pell_cohort	-0.08559 <.0001 2904	-0.21999 <.0001 1369	-0.30291 <.0001 1369	-0.21392 <.0001 2190	-0.18627 <.0001 2191	0.04384 0.0175 2940	0.36832 <.0001 2858	1.00000 2940	0.01838 0.3192 2940
PERSIST_1 PERSIST_1	0.36310 <.0001 2904	0.14800 <.0001 1369	0.11209 <.0001 1369	0.09999 <.0001 2190	0.12300 <.0001 2191	0.04155 0.0243 2940	-0.02140 0.2528 2858	0.01838 0.3192 2940	1.00000 2940

Figure 11

Output of VIF to Analyze Multicollinearity

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-0.15442	0.29235	-0.53	0.5975	.	0
ACT_M	ACT_M	1	0.04205	0.01420	2.96	0.0032	0.35181	2.84243
ACT_E	ACT_E	1	0.03494	0.01071	3.26	0.0012	0.38365	2.60656
SAT_V	SAT_V	1	-0.00059054	0.00063483	-0.93	0.3526	0.39198	2.55116
SAT_M	SAT_M	1	0.00086651	0.00072955	1.19	0.2353	0.31983	3.12667
GENDER_COHORT	GENDER_COHORT	1	0.24537	0.12853	1.91	0.0566	0.96020	1.04145
FIRST_GEN_NODEG_0	FIRST_GEN_NODEG_0	1	0.05722	0.07891	0.73	0.4685	0.74509	1.34212
Pell_cohort	Pell_cohort	1	-0.07252	0.07857	-0.92	0.3563	0.74946	1.33429
PERSIST_1	PERSIST_1	1	0.84019	0.08632	9.73	<.0001	0.96772	1.03336

Figure 12

Output of Regression Analysis for the Best Model

Best Regression Model

The REG Procedure
Model: MODEL1
Dependent Variable: CGPA

Number of Observations Read	3035
Number of Observations Used	1354
Number of Observations with Missing Values	1681

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	239.01623	79.67208	75.76	<.0001
Error	1350	1419.73335	1.05165		
Corrected Total	1353	1658.74958			

Root MSE	1.02550	R-Square	0.1441
Dependent Mean	2.30497	Adj R-Sq	0.1422
Coeff Var	44.49082		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.28044	0.18404	-1.52	0.1278
ACT_M	ACT_M	1	0.08112	0.00912	8.89	<.0001
ACT_E	ACT_E	1	0.02250	0.00709	3.17	0.0015
GENDER_COHORT	GENDER_COHORT	1	0.57760	0.10729	5.38	<.0001

CHAPTER V

CONCLUSION AND RECOMMENDATIONS

This analysis shows that students' average STEM GPA in their first year was lower than the average scores in the second year; this indicates an improvement as their academic journey progresses. Further, the resulting mean STEM GPA of students who were not from a first-generation family had a higher mean value than those from students who were first-generation university students. It was shown that there was a significant difference in students' grade in STEM courses by demographic data. The result of the post hoc test shows that there was statistical significance between the Asian and White, Asian and Hispanic, Asian and Black, White and Hispanic, and White and Black ethnic groups

As was shown in the regression analysis, the variables were struggling to predict how well the students do in STEM courses. Factors like the amount of time spent studying for a course, the amount of practice work completed, performance in high school STEM classes, course type, and time all likely play a large role in predicting student STEM success. Further, it is known that standardized tests are not the best measuring sticks; wealthier students (correlated to race and sometimes gender) do better on these tests because they are better prepared for them (Hess, 2019; Leilan & Sara, 2014). However, while a better approach would include more variables, the model created here gives a glimpse into what factors help predict a student's success in STEM courses.

It should be noted that for this research, the students GPA were calculated by making the students have equal weights, meaning that a student that takes four STEM courses would have the same weight when calculating overall GPA as a student who took one STEM course. An alternative method to calculating this would be to have the weighted average GPA calculated by

courses. This way the student taking more STEM courses would be weighed higher than the student taking one STEM course (because the weight would be equal by course hours instead of by student). However, due to the data the researcher had at their disposal and rigorous nature of the later route, it was deemed that the results that would have been obtained using the later route would be time consuming to implement and be of minimal difference to the result within the present analysis. Thus, the researcher decided to continue the implementation of this research with the first route mentioned. For future analysis, more time could be spent on analyzing the second route and results could be compared to identify significant differences in obtained results.

REFERENCES

Abaidoo, A. (2018). *Factors contributing to academic performance of students in a Junior High School* (Publication No. 9783668841062) [Master's thesis, University of Education] Grin Verlag. https://www.grin.com/document/450284

Beede, D., Doms, M., Khan, B., Langdon, D., & McKittrick, G. (2011). *STEM: Good jobs now and for the future* (Issue Brief 03-11). U.S. Department of Commerce, Economics and Statistics Administration. https://files.eric.ed.gov/fulltext/ED522129.pdf

Chen, X. (2009). *Students who study science, technology, engineering and Mathematics (STEM) in postsecondary education* (NCES 2009-161). National Centre for Education Statistics. U.S. Department of Education

Chen, X., & Ho, P. (2012). *STEM in postsecondary education; Entrance, attrition, and course taking among 2003-04 beginning postsecondary students* (NCES 2013 – 152). National Center for Education Statistics. U.S. Department of Education

Hess, A. J. (2019, October 3). Rich students get better SAT scores-here's why. *CNBC*. https://www.cnbc.com/2019/10/03/rich-students-get-better-sat-scores-heres-why.html

Leilan S. & Sara D. (2015). *Impact of household income on standardized test scores*. [Undergraduate honors thesis, Georgia Institute of Technology]. SMARTech Repository. https://www.smartech.gatech.edu/bitstream/handle/1853/54227/impact_of_income_on_test_scores_dadasheu.pdf?sequence=1&isAllowed=y

Merhi, R., Sánchez-El Vira-Paniagua, A., & Palací, F. J. (2018). The role of psychological strengths, coping strategies and well-being in the prediction of academic engagement and burnout in first-year university students. *Revista Acción Psicologica*, *15(2)*, 51–68. https://doi.org/: 10.5944/ap.15.2.21831

Noble, J. P. (1991). *Predicting college grades from ACT assessment scores and high school course work and grade information* (Report No. 91-3 [50291930]). American College Testing

Ribeiro L., Rosário P., Núñez J. C., Gaeta M., & Fuentes S. (2019). First-year students background and academic achievement: The mediating role of student engagement. *Frontiers in psychology, 10(6),* 5-13. https://doi.org/10.3389/fpsyg.2019.02669

Tian H., & Sun Z. (2018). *Historical development of academic achievement assessment (1st ed.).* Academic achievement assessment. https://doi.org/10.1007/978-3-662-56198-0_1

APPENDIX

SAS CODES

```
/*codes to install the three datasets */

/*Fall16*/
%web_drop_table(WORK.fall16);


FILENAME REFFILE '/home/u54850522/FALLL16.xlsx';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLSX
	OUT=WORK.fall16;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.fall16; RUN;


%web_open_table(WORK.fall16);


/*FALL17*/

%web_drop_table(WORK.fall17);


FILENAME REFFILE '/home/u54850522/FALLL17.xlsx';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLSX
	OUT=WORK.fall17;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.fall17; RUN;


%web_open_table(WORK.fall17);


/*FALL18*/
```

```
%web_drop_table(WORK.FALL18);


FILENAME REFFILE '/home/u54850522/FALLL18.xlsx';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLSX
	OUT=WORK.fall18;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.FALL18; RUN;


%web_open_table(WORK.FALL18);



/*Code to combine all three datasets*/
/*STEP 1: To combine all three table I first have to sort each dataset*/

proc sort data = work.fall16;
	by sas_stu_id;
	run;

proc sort data = work.fall17;
	by sas_stu_id;
	run;

proc sort data = work.fall18;
	by sas_stu_id;
	run;


/*STEP 2: combine all three using a new data name*/
data theData;
merge work.fall16
		work.fall17
		work.fall18;
	by sas_stu_id;
	run;

/*Print out first 10*/
proc print data =thedata (obs= 300);
run;
/*Now I want to make some basic statitical manipulation of my data*/
```

```
/*- I first want to fine the mean, median and STD of the dataset
by ethnicity*/

/*I choose three varibale for the group by function*/


proc means data=WORK.THEDATA chartype mean std min max n vardef=df;
        var SAT_V SAT_M ACT_M ACT_E;
        class YEAR;
run;


/*GROUP A*/
proc sort data = work.thedata;
        by year;
        run;
proc means mean median std;
        by year;
        var sat_v;
        run;

proc means mean median std;
        by year;
        var sat_m;
        run;

/*ACT   */
 proc means mean median std;
        by year;
        var act_e;
        run;
 proc means mean median std;
        by year;
        var act_m;
        run;




/*Group B*/
proc sort data = work.thedata;
        by gender_cohort;
        run;
proc means mean median std;
        by gender_cohort;
```

```
            var sat_v;
            run;

proc means mean median std;
            by gender_cohort;
            var sat_m;
            run;

/*ACT   */
 proc means mean median std;
            by gender_cohort;
            var act_e;
            run;
 proc means mean median std;
            by gender_cohort;
            var act_m;
            run;

/**/
/*Group C*/
proc sort data = work.thedata;
            by full_time_cohort;
            run;
proc means mean median std;
            by full_time_cohort;
            var sat_v;
            run;
 proc means mean median std;
            by full_time_cohort;
            var sat_m;
            run;
/*ACT   */
 proc means mean median std;
            by full_time_cohort;
            var act_e;
            run;
 proc means mean median std;
            by full_time_cohort;
            var act_m;
            run;

/*Group D*/
proc sort data = work.thedata;
            by stem_major;
            run;
 proc means mean median std;
```

```
            by stem_major;
            var sat_v;
            run;
 proc means mean median std;
            by stem_major;
            var sat_m;
            run;

/*ACT  */
 proc means mean median std;
            by stem_major;
            var act_e;
            run;
 proc means mean median std;
            by stem_major;
            var act_m;
            run;

/*GROUP D*/
proc sort data = work.thedata;
            by first_gen_nodeg_0;
            run;

proc means mean median std;
            by first_gen_nodeg_0;
            var sat_v;
            run;
 proc means mean median std;
            by first_gen_nodeg_0;
            var sat_m;
            run;

/*ACT   */

proc means mean median std;
            by first_gen_nodeg_0;
            var act_e;
            run;
 proc means mean median std;
            by first_gen_nodeg_0;
            var act_m;
            run;

/*mean statistics of ACT and SAT scores*/
proc means mean median std;
            var act_e;
```

```
	run;
proc means mean median std;
	var act_m;
	run;
proc means mean median std;
	var sat_v;
	run;
proc means mean median std;
	var sat_m;
	run;

 proc means mean std;
	var cum_gpa_1;
	run;


/*Data Visualization of variables*/


proc sgplot;
	histogram cum_gpa_1;
	label cum_gpa_1 = 'First Fall Cummulative GPA';
	title 'Histogram of Fall Cumulative GPA';
	run;



proc sgplot;
	histogram sp_cum_gpa_1;
	label cum_gpa_1 = 'First Spring Cummulative GPA';
	title 'Histogram of Spring Cummulative GPA';
	run;

proc sgplot;
	histogram cum_gpa_2;
	label cum_gpa_1 = 'Second Year Fall Cummulative GPA';
	title 'Histogram of 2nd Year Fall Cummulative GPA';
	run;

proc sgplot;
	histogram sp_cum_gpa_2;
	label cum_gpa_1 = 'Second Year Spring Cummulative GPA';
	title 'Histogram of 2nd Year Spring Cummulative GPA';
	run;

proc sgplot;
```

```
	histogram cum_gpa_3;
	label cum_gpa_1 = 'Third Year Fall Cummulative GPA';
	title 'Histogram of 3rd Year Fall Cummulative GPA';
	run;

proc sgplot;
	histogram sp_cum_gpa_3;
	label cum_gpa_1 = 'Third Year Spring Cummulative GPA';
	title 'Histogram of 3rd Year Spring Cummulative GPA';
	run;




/*Box-plot*/

proc sgplot;
	hbox sat_v / category = ethnicity;
	xaxis label = 'SAT VERBAL SCORES';
	title 'Boxplot of SAT Verbal by Etnicity';
	run;

proc sgplot;
	hbox sat_m / category = ethnicity;
	xaxis label = 'SAT MATHEMATICS SCORES';
	title 'Boxplot of SAT Mathematics by Etnicity';
	run;

proc sgplot;
	hbox act_e / category = ethnicity;
	xaxis label = 'ACT ENGLISH SCORES';
	title 'Boxplot of ACT English by Etnicity';
	run;

proc sgplot;
	hbox act_m / category = ethnicity;
	xaxis label = 'ACT MATHEMATICS SCORES';
	title 'Boxplot of ACT Mathematics by Etnicity';
	run;

/*Yearly boxplot of SAT and ACT scores*/

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.THEDATA;
```

```
        hbox SAT_V / category=YEAR;
        xaxis label = 'SAT VERBAL SCORES';
        title 'Box-Plot of SAT Verbal Scores by Year';
        run;

proc sgplot data=WORK.THEDATA;
        hbox SAT_M / category=YEAR;
        xaxis label = 'SAT MATHEMATICS SCORES';
        title 'Box-Plot of SAT Mathematics Scores by Year';
        run;

proc sgplot data=WORK.THEDATA;
        hbox ACT_E / category=YEAR;
        xaxis label = 'ACT ENGLISH SCORES';
        title 'Box-Plot of ACT English Scores by Year';
        run;

proc sgplot data=WORK.THEDATA;
        hbox ACT_M / category=YEAR;
        xaxis label = 'ACT MATHEMATICS SCORES';
        title 'Box-Plot of ACT Mathematics Scores by Year';
        run;

ods graphics / reset;



/*by stem_major*/
proc sgplot;
        hbox sat_v / category = stem_major;
        title 'boxplot of sat verbal by stem major';
        run;

proc sgplot;
        hbox sat_m / category = stem_major;
        title 'boxplot of sat mathematics by stem major';
        run;

proc sgplot;
        hbox act_e / category = stem_major;
        title 'boxplot of act english by stem major';
        run;

proc sgplot;
        hbox act_m / category =stem_major;
        title 'boxplot of act mathematics by stem major';
```

```
	run;

proc sgplot;
	hbox sp_cum_gpa_1 / category = ethnicity;
	title 'Boxplot of Sat by Etnicity';
	run;

/*by year; I DONT THINK THIS IS RELEVANT IN WRITE UPS*/
proc sgplot;
	hbox sat_v / category = year;
	title 'Boxplot of SAT Verbal by YEAR';
	run;

proc sgplot;
	hbox sat_m / category = year;
	title 'Boxplot of SAT Mathematics by Year';
	run;

proc sgplot;
	hbox act_e / category = year;
	title 'Boxplot of ACT English by Year';
	run;

proc sgplot;
	hbox act_m / category =year;
	title 'Boxplot of ACT Mathematics by Year';
	run;




/*pie chart*/

proc template;
	define statgraph SASStudio.Pie;
		begingraph;
		layout region;
		piechart category=ETHNICITY /  stat=pct;
		title 'Pie Chart of the Dataset by Ethnicity';
		endlayout;
		endgraph;
	end;
run;

ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgrender template=SASStudio.Pie data=WORK.THEDATA;
```

```
run;

ods graphics / reset;


proc template;
	define statgraph SASStudio.Pie;
		begingraph;
		layout region;
		piechart category=stem_major / stat=pct;
		title 'pie chart of the dataset by stem_major';
		endlayout;
		endgraph;
	end;
run;

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgrender template=SASStudio.Pie data=WORK.THEDATA;
run;

ods graphics / reset;

/*pie chart of gender*/

/* Define Pie template */
proc template;
	define statgraph SASStudio.Pie;
		begingraph;
		layout region;
		piechart category=GENDER_COHORT / stat=pct;
		title 'Pie Chart of the Dataset by Gender';
		endlayout;
		endgraph;
	end;
run;

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgrender template=SASStudio.Pie data=WORK.THEDATA;
run;

ods graphics / reset;


data work.thedata;
```

```
set work.thedata;
length Grades $10;
if CRS_FA_GRADE1 = 'DA' then Grades = 4;
else if CRS_FA_GRADE1 = 'A' then Grades = 4;
else if CRS_FA_GRADE1 = 'DB' then Grades = 3;
else if CRS_FA_GRADE1 = 'B' then Grades = 3;
else if CRS_FA_GRADE1 = 'DC' then Grades = 2;
else if CRS_FA_GRADE1 = 'C' then Grades = 2;
else if CRS_FA_GRADE1 = 'DD' then Grades = 1;
else if CRS_FA_GRADE1 = 'D' then Grades = 1;
else if CRS_FA_GRADE1 = 'DF' then Grades = 0;
else if CRS_FA_GRADE1 = 'F' then Grades = 0;
else if CRS_FA_GRADE1 = 'W' then Grades = '.';
else Grades = '.';


length Grades1 $10;
if                    CRS_FA_GRADE2 = 'DA' then       Grades1 = 4;
else if CRS_FA_GRADE2 = 'A' then  Grades1 = 4;
else if CRS_FA_GRADE2 = 'DB' then          Grades1 = 3;
else if CRS_FA_GRADE2 = 'B' then  Grades1 = 3;
else if CRS_FA_GRADE2 = 'DC' then          Grades1 = 2;
else if CRS_FA_GRADE2 = 'C' then  Grades1 = 2;
else if CRS_FA_GRADE2 = 'DD' then          Grades1 = 1;
else if CRS_FA_GRADE2 = 'D' then  Grades1 = 1;
else if CRS_FA_GRADE2 = 'DF' then          Grades1 = 0;
else if CRS_FA_GRADE2 = 'F' then  Grades1 = 0;
else if CRS_FA_GRADE2 = 'W' then          Grades1 = '.';
else Grades1 = '.';


length Grades2 $10;
if                    CRS_FA_GRADE3 = 'DA' then       Grades2 = 4;
else if CRS_FA_GRADE3 = 'A' then  Grades2 = 4;
else if CRS_FA_GRADE3 = 'DB' then          Grades2 = 3;
else if CRS_FA_GRADE3 = 'B' then  Grades2 = 3;
else if CRS_FA_GRADE3 = 'DC' then          Grades2 = 2;
else if CRS_FA_GRADE3 = 'C' then  Grades2 = 2;
else if CRS_FA_GRADE3 = 'DD' then          Grades2 = 1;
else if CRS_FA_GRADE3 = 'D' then  Grades2 = 1;
else if CRS_FA_GRADE3 = 'DF' then          Grades2 = 0;
else if CRS_FA_GRADE3 = 'F' then  Grades2 = 0;
else if CRS_FA_GRADE3 = 'W' then          Grades2 = '.';
else Grades2 = '.';
```

```
length Grades3 $10;
if                  CRS_FA_GRADE4 = 'DA' then       Grades3 = 4;
else if CRS_FA_GRADE4 = 'A' then  Grades3 = 4;
else if CRS_FA_GRADE4 = 'DB' then          Grades3 = 3;
else if CRS_FA_GRADE4 = 'B' then  Grades3 = 3;
else if CRS_FA_GRADE4 = 'DC' then          Grades3 = 2;
else if CRS_FA_GRADE4 = 'C' then  Grades3 = 2;
else if CRS_FA_GRADE4 = 'DD' then          Grades3 = 1;
else if CRS_FA_GRADE4 = 'D' then  Grades3 = 1;
else if CRS_FA_GRADE4 = 'DF' then          Grades3 = 0;
else if CRS_FA_GRADE4 = 'F' then  Grades3 = 0;
else if CRS_FA_GRADE4 = 'W' then           Grades3 = '.';
else Grades3 = '.';


length Grades4 $10;
if                  CRS_FA_GRADE4 = 'DA' then       Grades4 = 4;
else if CRS_FA_GRADE4 = 'A' then  Grades4 = 4;
else if CRS_FA_GRADE4 = 'DB' then          Grades4 = 3;
else if CRS_FA_GRADE4 = 'B' then  Grades4 = 3;
else if CRS_FA_GRADE4 = 'DC' then          Grades4 = 2;
else if CRS_FA_GRADE4 = 'C' then  Grades4 = 2;
else if CRS_FA_GRADE4 = 'DD' then          Grades4 = 1;
else if CRS_FA_GRADE4 = 'D' then  Grades4 = 1;
else if CRS_FA_GRADE4 = 'DF' then          Grades4 = 0;
else if CRS_FA_GRADE4 = 'F' then  Grades4 = 0;
else if CRS_FA_GRADE4 = 'W' then           Grades4 = '.';
else Grades4 = '.';


length Grades5 $10;
if                  CRS_FA_GRADE5 = 'DA' then       Grades5 = 4;
else if CRS_FA_GRADE5 = 'A' then  Grades5 = 4;
else if CRS_FA_GRADE5 = 'DB' then          Grades5 = 3;
else if CRS_FA_GRADE5 = 'B' then  Grades5 = 3;
else if CRS_FA_GRADE5 = 'DC' then          Grades5 = 2;
else if CRS_FA_GRADE5 = 'C' then  Grades5 = 2;
else if CRS_FA_GRADE5 = 'DD' then          Grades5 = 1;
else if CRS_FA_GRADE5 = 'D' then  Grades5 = 1;
else if CRS_FA_GRADE5 = 'DF' then          Grades5 = 0;
else if CRS_FA_GRADE5 = 'F' then  Grades5 = 0;
else if CRS_FA_GRADE5 = 'W' then           Grades5 = '.';
else Grades5 = '.';


length Grades6 $10;
```

```
if                  CRS_FA_GRADE6 = 'DA' then       Grades6 = 4;
else if CRS_FA_GRADE6 = 'A' then  Grades6 = 4;
else if CRS_FA_GRADE6 = 'DB' then          Grades6 = 3;
else if CRS_FA_GRADE6 = 'B' then  Grades6 = 3;
else if CRS_FA_GRADE6 = 'DC' then          Grades6 = 2;
else if CRS_FA_GRADE6 = 'C' then  Grades6 = 2;
else if CRS_FA_GRADE6 = 'DD' then          Grades6 = 1;
else if CRS_FA_GRADE6 = 'D' then  Grades6 = 1;
else if CRS_FA_GRADE6 = 'DF' then          Grades6 = 0;
else if CRS_FA_GRADE6 = 'F' then  Grades6 = 0;
else if CRS_FA_GRADE6 = 'W' then          Grades6 = '.';
else Grades6 = '.';


length Grades7 $10;
if                  CRS_SP_GRADE1 = 'DA' then       Grades7 = 4;
else if CRS_SP_GRADE1 = 'A' then  Grades7 = 4;
else if CRS_SP_GRADE1 = 'DB' then          Grades7 = 3;
else if CRS_SP_GRADE1 = 'B' then  Grades7 = 3;
else if CRS_SP_GRADE1 = 'DC' then          Grades7 = 2;
else if CRS_SP_GRADE1 = 'C' then  Grades7 = 2;
else if CRS_SP_GRADE1 = 'DD' then          Grades7 = 1;
else if CRS_SP_GRADE1 = 'D' then  Grades7 = 1;
else if CRS_SP_GRADE1 = 'DF' then          Grades7 = 0;
else if CRS_SP_GRADE1 = 'F' then  Grades7 = 0;
else if CRS_SP_GRADE1 = 'W' then Grades7 = '.';
else Grades7 = '.';


length Grades8 $10;
if                  CRS_SP_GRADE2 = 'DA' then       Grades8 = 4;
else if CRS_SP_GRADE2 = 'A' then  Grades8 = 4;
else if CRS_SP_GRADE2 = 'DB' then          Grades8 = 3;
else if CRS_SP_GRADE2 = 'B' then  Grades8 = 3;
else if CRS_SP_GRADE2 = 'DC' then          Grades8 = 2;
else if CRS_SP_GRADE2 = 'C' then  Grades8 = 2;
else if CRS_SP_GRADE2 = 'DD' then          Grades8 = 1;
else if CRS_SP_GRADE2 = 'D' then  Grades8 = 1;
else if CRS_SP_GRADE2 = 'DF' then          Grades8 = 0;
else if CRS_SP_GRADE2 = 'F' then  Grades8 = 0;
else if CRS_SP_GRADE2 = 'W' then Grades8 = '.';
else Grades8 = '.';


length Grades9 $10;
if                  CRS_SP_GRADE3 = 'DA' then       Grades9 = 4;
```

```
else if CRS_SP_GRADE3 = 'A' then  Grades9 = 4;
else if CRS_SP_GRADE3 = 'DB' then          Grades9 = 3;
else if CRS_SP_GRADE3 = 'B' then  Grades9 = 3;
else if CRS_SP_GRADE3 = 'DC' then          Grades9 = 2;
else if CRS_SP_GRADE3 = 'C' then  Grades9 = 2;
else if CRS_SP_GRADE3 = 'DD' then          Grades9 = 1;
else if CRS_SP_GRADE3 = 'D' then  Grades9 = 1;
else if CRS_SP_GRADE3 = 'DF' then          Grades9 = 0;
else if CRS_SP_GRADE3 = 'F' then  Grades9 = 0;
else if CRS_SP_GRADE3 = 'W' then Grades9 = '.';
else Grades9 = '.';


length Grades10 $10;
if                  CRS_SP_GRADE4 = 'DA' then       Grades10 = 4;
else if CRS_SP_GRADE4 = 'A' then  Grades10 = 4;
else if CRS_SP_GRADE4 = 'DB' then          Grades10 = 3;
else if CRS_SP_GRADE4 = 'B' then  Grades10 = 3;
else if CRS_SP_GRADE4 = 'DC' then          Grades10 = 2;
else if CRS_SP_GRADE4 = 'C' then  Grades10 = 2;
else if CRS_SP_GRADE4 = 'DD' then          Grades10 = 1;
else if CRS_SP_GRADE4 = 'D' then  Grades10 = 1;
else if CRS_SP_GRADE4 = 'DF' then          Grades10 = 0;
else if CRS_SP_GRADE4 = 'F' then  Grades10 = 0;
else if CRS_SP_GRADE4 = 'W' then Grades10 = '.';
else Grades10 = '.';


length Grades11 $10;
if                  CRS_SP_GRADE5 = 'DA' then       Grades11 = 4;
else if CRS_SP_GRADE5 = 'A' then  Grades11 = 4;
else if CRS_SP_GRADE5 = 'DB' then          Grades11 = 3;
else if CRS_SP_GRADE5 = 'B' then  Grades11 = 3;
else if CRS_SP_GRADE5 = 'DC' then          Grades11 = 2;
else if CRS_SP_GRADE5 = 'C' then  Grades11 = 2;
else if CRS_SP_GRADE5 = 'DD' then          Grades11 = 1;
else if CRS_SP_GRADE5 = 'D' then  Grades11 = 1;
else if CRS_SP_GRADE5 = 'DF' then          Grades11 = 0;
else if CRS_SP_GRADE5 = 'F' then  Grades11 = 0;
else if CRS_SP_GRADE5 = 'W' then Grades11 = '.';
else Grades11 = '.';


length Grades12 $10;
if                  CRS_SP_GRADE6 = 'DA' then       Grades12 = 4;
```

```
else if CRS_SP_GRADE6 = 'A' then  Grades12 = 4;
else if CRS_SP_GRADE6 = 'DB' then          Grades12 = 3;
else if CRS_SP_GRADE6 = 'B' then  Grades12 = 3;
else if CRS_SP_GRADE6 = 'DC' then          Grades12 = 2;
else if CRS_SP_GRADE6 = 'C' then  Grades12 = 2;
else if CRS_SP_GRADE6 = 'DD' then          Grades12 = 1;
else if CRS_SP_GRADE6 = 'D' then  Grades12 = 1;
else if CRS_SP_GRADE6 = 'DF' then          Grades12 = 0;
else if CRS_SP_GRADE6 = 'F' then   Grades12 = 0;
else if CRS_SP_GRADE6 = 'W' then Grades12 = '.';
else Grades12 = '.';



length Grades13 $10;
if                  CRS_SP_GRADE7 = 'DA' then       Grades13 = 4;
else if CRS_SP_GRADE7 = 'A' then  Grades13 = 4;
else if CRS_SP_GRADE7 = 'DB' then          Grades13 = 3;
else if CRS_SP_GRADE7 = 'B' then  Grades13 = 3;
else if CRS_SP_GRADE7 = 'DC' then          Grades13 = 2;
else if CRS_SP_GRADE7 = 'C' then  Grades13 = 2;
else if CRS_SP_GRADE7 = 'DD' then          Grades13 = 1;
else if CRS_SP_GRADE7 = 'D' then  Grades13 = 1;
else if CRS_SP_GRADE7 = 'DF' then          Grades13 = 0;
else if CRS_SP_GRADE7 = 'F' then   Grades13 = 0;
else if CRS_SP_GRADE7 = 'W' then Grades13 = '.';
else Grades13 = '.';


length Grades14 $10;
if                  CRS_SU_GRADE1 = 'DA' then       Grades14 = 4;
else if CRS_SU_GRADE1 = 'A' then  Grades14 = 4;
else if CRS_SU_GRADE1 = 'DB' then          Grades14 = 3;
else if CRS_SU_GRADE1 = 'B' then  Grades14 = 3;
else if CRS_SU_GRADE1 = 'DC' then          Grades14 = 2;
else if CRS_SU_GRADE1 = 'C' then  Grades14 = 2;
else if CRS_SU_GRADE1 = 'DD' then          Grades14 = 1;
else if CRS_SU_GRADE1 = 'D' then  Grades14 = 1;
else if CRS_SU_GRADE1 = 'DF' then          Grades14 = 0;
else if CRS_SU_GRADE1 = 'F' then  Grades14 = 0;
else if CRS_SU_GRADE1 = 'W' then          Grades14 = '.';
else Grades14 = '.';


length Grades15 $10;
if                  CRS_SU_GRADE2 = 'DA' then       Grades15 = 4;
```

```
else if CRS_SU_GRADE2 = 'A' then  Grades15 = 4;
else if CRS_SU_GRADE2 = 'DB' then          Grades15 = 3;
else if CRS_SU_GRADE2 = 'B' then  Grades15 = 3;
else if CRS_SU_GRADE2 = 'DC' then          Grades15 = 2;
else if CRS_SU_GRADE2 = 'C' then  Grades15 = 2;
else if CRS_SU_GRADE2 = 'DD' then          Grades15 = 1;
else if CRS_SU_GRADE2 = 'D' then  Grades15 = 1;
else if CRS_SU_GRADE2 = 'DF' then          Grades15 = 0;
else if CRS_SU_GRADE2 = 'F' then  Grades15 = 0;
else if CRS_SU_GRADE2 = 'W' then          Grades15 = '.';
else Grades15 = '.';



length Grades16 $10;
if                  CRS_SU_GRADE3 = 'DA' then       Grades16 = 4;
else if CRS_SU_GRADE3 = 'A' then  Grades16 = 4;
else if CRS_SU_GRADE3 = 'DB' then          Grades16 = 3;
else if CRS_SU_GRADE3 = 'B' then  Grades16 = 3;
else if CRS_SU_GRADE3 = 'DC' then          Grades16 = 2;
else if CRS_SU_GRADE3 = 'C' then  Grades16 = 2;
else if CRS_SU_GRADE3 = 'DD' then          Grades16 = 1;
else if CRS_SU_GRADE3 = 'D' then  Grades16 = 1;
else if CRS_SU_GRADE3 = 'DF' then          Grades16 = 0;
else if CRS_SU_GRADE3 = 'F' then  Grades16 = 0;
else if CRS_SU_GRADE3 = 'W' then          Grades16 = '.';
else Grades16 = '.';


length Grades17 $10;
if                  CRS_SU_GRADE4 = 'DA' then       Grades17 = 4;
else if CRS_SU_GRADE4 = 'A' then  Grades17 = 4;
else if CRS_SU_GRADE4 = 'DB' then          Grades17 = 3;
else if CRS_SU_GRADE4 = 'B' then  Grades17 = 3;
else if CRS_SU_GRADE4 = 'DC' then          Grades17 = 2;
else if CRS_SU_GRADE4 = 'C' then  Grades17 = 2;
else if CRS_SU_GRADE4 = 'DD' then          Grades17 = 1;
else if CRS_SU_GRADE4 = 'D' then  Grades17 = 1;
else if CRS_SU_GRADE4 = 'DF' then          Grades17 = 0;
else if CRS_SU_GRADE4 = 'F' then  Grades17 = 0;
else if CRS_SU_GRADE4 = 'W' then          Grades17 = '.';
else Grades17 = '.';


length Grades18 $10;
if                  CRS_SU_GRADE5 = 'DA' then       Grades18 = 4;
```

```
else if CRS_SU_GRADE5 = 'A' then  Grades18 = 4;
else if CRS_SU_GRADE5 = 'DB' then          Grades18 = 3;
else if CRS_SU_GRADE5 = 'B' then  Grades18 = 3;
else if CRS_SU_GRADE5 = 'DC' then          Grades18 = 2;
else if CRS_SU_GRADE5 = 'C' then  Grades18 = 2;
else if CRS_SU_GRADE5 = 'DD' then          Grades18 = 1;
else if CRS_SU_GRADE5 = 'D' then  Grades18 = 1;
else if CRS_SU_GRADE5 = 'DF' then          Grades18 = 0;
else if CRS_SU_GRADE5 = 'F' then  Grades18 = 0;
else if CRS_SU_GRADE5 = 'W' then           Grades18 = '.';
else Grades18 = '.';


length Grades19 $10;
if                  CRS2_FA_GRADE1 = 'DA' then      Grades19 = 4;
else if CRS2_FA_GRADE1 = 'A' then          Grades19 = 4;
else if CRS2_FA_GRADE1 = 'DB' then         Grades19 = 3;
else if CRS2_FA_GRADE1 = 'B' then          Grades19 = 3;
else if CRS2_FA_GRADE1 = 'DC' then         Grades19 = 2;
else if CRS2_FA_GRADE1 = 'C' then          Grades19 = 2;
else if CRS2_FA_GRADE1 = 'DD' then         Grades19 = 1;
else if CRS2_FA_GRADE1 = 'D' then          Grades19 = 1;
else if CRS2_FA_GRADE1 = 'DF' then         Grades19 = 0;
else if CRS2_FA_GRADE1 = 'F' then          Grades19 = 0;
else if CRS2_FA_GRADE1 = 'W' then          Grades19 = '.';
else Grades19 = '.';



length Grades20 $10;
if                  CRS2_FA_GRADE2 = 'DA' then      Grades20 = 4;
else if CRS2_FA_GRADE2 = 'A' then          Grades20 = 4;
else if CRS2_FA_GRADE2 = 'DB' then         Grades20 = 3;
else if CRS2_FA_GRADE2 = 'B' then          Grades20 = 3;
else if CRS2_FA_GRADE2 = 'DC' then         Grades20 = 2;
else if CRS2_FA_GRADE2 = 'C' then          Grades20 = 2;
else if CRS2_FA_GRADE2 = 'DD' then         Grades20 = 1;
else if CRS2_FA_GRADE2 = 'D' then          Grades20 = 1;
else if CRS2_FA_GRADE2 = 'DF' then         Grades20 = 0;
else if CRS2_FA_GRADE2 = 'F' then          Grades20 = 0;
else if CRS2_FA_GRADE2 = 'W' then          Grades20 = '.';
else Grades20 = '.';


length Grades21 $10;
if                  CRS2_FA_GRADE3 = 'DA' then      Grades21 = 4;
```

```
else if CRS2_FA_GRADE3 = 'A' then          Grades21 = 4;
else if CRS2_FA_GRADE3 = 'DB' then         Grades21 = 3;
else if CRS2_FA_GRADE3 = 'B' then          Grades21 = 3;
else if CRS2_FA_GRADE3 = 'DC' then         Grades21 = 2;
else if CRS2_FA_GRADE3 = 'C' then          Grades21 = 2;
else if CRS2_FA_GRADE3 = 'DD' then         Grades21 = 1;
else if CRS2_FA_GRADE3 = 'D' then          Grades21 = 1;
else if CRS2_FA_GRADE3 = 'DF' then         Grades21 = 0;
else if CRS2_FA_GRADE3 = 'F' then          Grades21 = 0;
else if CRS2_FA_GRADE3 = 'W' then          Grades21 = '.';
else Grades21 = '.';


length Grades22 $10;
if                  CRS2_FA_GRADE4 = 'DA' then     Grades22 = 4;
else if CRS2_FA_GRADE4 = 'A' then          Grades22 = 4;
else if CRS2_FA_GRADE4 = 'DB' then         Grades22 = 3;
else if CRS2_FA_GRADE4 = 'B' then          Grades22 = 3;
else if CRS2_FA_GRADE4 = 'DC' then         Grades22 = 2;
else if CRS2_FA_GRADE4 = 'C' then          Grades22 = 2;
else if CRS2_FA_GRADE4 = 'DD' then         Grades22 = 1;
else if CRS2_FA_GRADE4 = 'D' then          Grades22 = 1;
else if CRS2_FA_GRADE4 = 'DF' then         Grades22 = 0;
else if CRS2_FA_GRADE4 = 'F' then          Grades22 = 0;
else if CRS2_FA_GRADE4 = 'W' then          Grades22 = '.';
else Grades22 = '.';


length Grades23 $10;
if                  CRS2_FA_GRADE5 = 'DA' then     Grades23 = 4;
else if CRS2_FA_GRADE5 = 'A' then          Grades23 = 4;
else if CRS2_FA_GRADE5 = 'DB' then         Grades23 = 3;
else if CRS2_FA_GRADE5 = 'B' then          Grades23 = 3;
else if CRS2_FA_GRADE5 = 'DC' then         Grades23 = 2;
else if CRS2_FA_GRADE5 = 'C' then          Grades23 = 2;
else if CRS2_FA_GRADE5 = 'DD' then         Grades23 = 1;
else if CRS2_FA_GRADE5 = 'D' then          Grades23 = 1;
else if CRS2_FA_GRADE5 = 'DF' then         Grades23 = 0;
else if CRS2_FA_GRADE5 = 'F' then          Grades23 = 0;
else if CRS2_FA_GRADE5 = 'W' then          Grades23 = '.';
else Grades23 = '.';


length Grades24 $10;
if                  CRS2_FA_GRADE6 = 'DA' then     Grades24 = 4;
else if CRS2_FA_GRADE6 = 'A' then          Grades24 = 4;
```

```
else if CRS2_FA_GRADE6 = 'DB' then          Grades24 = 3;
else if CRS2_FA_GRADE6 = 'B' then           Grades24 = 3;
else if CRS2_FA_GRADE6 = 'DC' then          Grades24 = 2;
else if CRS2_FA_GRADE6 = 'C' then           Grades24 = 2;
else if CRS2_FA_GRADE6 = 'DD' then          Grades24 = 1;
else if CRS2_FA_GRADE6 = 'D' then           Grades24 = 1;
else if CRS2_FA_GRADE6 = 'DF' then          Grades24 = 0;
else if CRS2_FA_GRADE6 = 'F' then           Grades24 = 0;
else if CRS2_FA_GRADE6 = 'W' then           Grades24 = '.';
else Grades24 = '.';


length Grades25 $10;
if                  CRS2_FA_GRADE7 = 'DA' then      Grades25 = 4;
else if CRS2_FA_GRADE7 = 'A' then           Grades25 = 4;
else if CRS2_FA_GRADE7 = 'DB' then          Grades25 = 3;
else if CRS2_FA_GRADE7 = 'B' then           Grades25 = 3;
else if CRS2_FA_GRADE7 = 'DC' then          Grades25 = 2;
else if CRS2_FA_GRADE7 = 'C' then           Grades25 = 2;
else if CRS2_FA_GRADE7 = 'DD' then          Grades25 = 1;
else if CRS2_FA_GRADE7 = 'D' then           Grades25 = 1;
else if CRS2_FA_GRADE7 = 'DF' then          Grades25 = 0;
else if CRS2_FA_GRADE7 = 'F' then           Grades25 = 0;
else if CRS2_FA_GRADE7 = 'W' then           Grades25 = '.';
else Grades25 = '.';


length Grades26 $10;
if                  CRS2_FA_GRADE8 = 'DA' then      Grades26 = 4;
else if CRS2_FA_GRADE8 = 'A' then           Grades26 = 4;
else if CRS2_FA_GRADE8 = 'DB' then          Grades26 = 3;
else if CRS2_FA_GRADE8 = 'B' then           Grades26 = 3;
else if CRS2_FA_GRADE8 = 'DC' then          Grades26 = 2;
else if CRS2_FA_GRADE8 = 'C' then           Grades26 = 2;
else if CRS2_FA_GRADE8 = 'DD' then          Grades26 = 1;
else if CRS2_FA_GRADE8 = 'D' then           Grades26 = 1;
else if CRS2_FA_GRADE8 = 'DF' then          Grades26 = 0;
else if CRS2_FA_GRADE8 = 'F' then           Grades26 = 0;
else if CRS2_FA_GRADE8 = 'W' then           Grades26 = '.';
else Grades26 = '.';


length Grades27 $10;
if                  CRS2_SP_GRADE1 = 'DA' then      Grades27 = 4;
else if CRS2_SP_GRADE1 = 'A' then           Grades27 = 4;
```

```
else if CRS2_SP_GRADE1 = 'DB' then          Grades27 = 3;
else if CRS2_SP_GRADE1 = 'B' then           Grades27 = 3;
else if CRS2_SP_GRADE1 = 'DC' then          Grades27 = 2;
else if CRS2_SP_GRADE1 = 'C' then           Grades27 = 2;
else if CRS2_SP_GRADE1 = 'DD' then          Grades27 = 1;
else if CRS2_SP_GRADE1 = 'D' then           Grades27 = 1;
else if CRS2_SP_GRADE1 = 'DF' then          Grades27 = 0;
else if CRS2_SP_GRADE1 = 'F' then           Grades27 = 0;
else if CRS2_SP_GRADE1 = 'W' then           Grades27 = '.';
else Grades27 = '.';


length Grades28 $10;
if                  CRS2_SP_GRADE2 = 'DA' then      Grades28 = 4;
else if CRS2_SP_GRADE2 = 'A' then           Grades28 = 4;
else if CRS2_SP_GRADE2 = 'DB' then          Grades28 = 3;
else if CRS2_SP_GRADE2 = 'B' then           Grades28 = 3;
else if CRS2_SP_GRADE2 = 'DC' then          Grades28 = 2;
else if CRS2_SP_GRADE2 = 'C' then           Grades28 = 2;
else if CRS2_SP_GRADE2 = 'DD' then          Grades28 = 1;
else if CRS2_SP_GRADE2 = 'D' then           Grades28 = 1;
else if CRS2_SP_GRADE2 = 'DF' then          Grades28 = 0;
else if CRS2_SP_GRADE2 = 'F' then           Grades28 = 0;
else if CRS2_SP_GRADE2 = 'W' then           Grades28 = '.';
else Grades28 = '.';


length Grades29 $10;
if                  CRS2_SP_GRADE3 = 'DA' then      Grades29 = 4;
else if CRS2_SP_GRADE3 = 'A' then           Grades29 = 4;
else if CRS2_SP_GRADE3 = 'DB' then          Grades29 = 3;
else if CRS2_SP_GRADE3 = 'B' then           Grades29 = 3;
else if CRS2_SP_GRADE3 = 'DC' then          Grades29 = 2;
else if CRS2_SP_GRADE3 = 'C' then           Grades29 = 2;
else if CRS2_SP_GRADE3 = 'DD' then          Grades29 = 1;
else if CRS2_SP_GRADE3 = 'D' then           Grades29 = 1;
else if CRS2_SP_GRADE3 = 'DF' then          Grades29 = 0;
else if CRS2_SP_GRADE3 = 'F' then           Grades29 = 0;
else if CRS2_SP_GRADE3 = 'W' then           Grades29 = '.';
else Grades29 = '.';


length Grades30 $10;
if                  CRS2_SP_GRADE4 = 'DA' then      Grades30 = 4;
else if CRS2_SP_GRADE4 = 'A' then           Grades30 = 4;
else if CRS2_SP_GRADE4 = 'DB' then          Grades30 = 3;
```

```
else if CRS2_SP_GRADE4 = 'B' then          Grades30 = 3;
else if CRS2_SP_GRADE4 = 'DC' then         Grades30 = 2;
else if CRS2_SP_GRADE4 = 'C' then          Grades30 = 2;
else if CRS2_SP_GRADE4 = 'DD' then         Grades30 = 1;
else if CRS2_SP_GRADE4 = 'D' then          Grades30 = 1;
else if CRS2_SP_GRADE4 = 'DF' then         Grades30 = 0;
else if CRS2_SP_GRADE4 = 'F' then          Grades30 = 0;
else if CRS2_SP_GRADE4 = 'W' then          Grades30 = '.';
else Grades30 = '.';


length Grades31 $10;
if                  CRS2_SP_GRADE5 = 'DA' then     Grades31 = 4;
else if CRS2_SP_GRADE5 = 'A' then          Grades31 = 4;
else if CRS2_SP_GRADE5 = 'DB' then         Grades31 = 3;
else if CRS2_SP_GRADE5 = 'B' then          Grades31 = 3;
else if CRS2_SP_GRADE5 = 'DC' then         Grades31 = 2;
else if CRS2_SP_GRADE5 = 'C' then          Grades31 = 2;
else if CRS2_SP_GRADE5 = 'DD' then         Grades31 = 1;
else if CRS2_SP_GRADE5 = 'D' then          Grades31 = 1;
else if CRS2_SP_GRADE5 = 'DF' then         Grades31 = 0;
else if CRS2_SP_GRADE5 = 'F' then          Grades31 = 0;
else if CRS2_SP_GRADE5 = 'W' then          Grades31 = '.';
else Grades31 = '.';


length Grades32 $10;
if                  CRS2_SP_GRADE6 = 'DA' then     Grades32 = 4;
else if CRS2_SP_GRADE6 = 'A' then          Grades32 = 4;
else if CRS2_SP_GRADE6 = 'DB' then         Grades32 = 3;
else if CRS2_SP_GRADE6 = 'B' then          Grades32 = 3;
else if CRS2_SP_GRADE6 = 'DC' then         Grades32 = 2;
else if CRS2_SP_GRADE6 = 'C' then          Grades32 = 2;
else if CRS2_SP_GRADE6 = 'DD' then         Grades32 = 1;
else if CRS2_SP_GRADE6 = 'D' then          Grades32 = 1;
else if CRS2_SP_GRADE6 = 'DF' then         Grades32 = 0;
else if CRS2_SP_GRADE6 = 'F' then          Grades32 = 0;
else if CRS2_SP_GRADE6 = 'W' then          Grades32 = '.';
else Grades32 = '.';


length Grades33 $10;
if                  CRS2_SP_GRADE7 = 'DA' then     Grades33 = 4;
else if CRS2_SP_GRADE7 = 'A' then          Grades33 = 4;
else if CRS2_SP_GRADE7 = 'DB' then         Grades33 = 3;
else if CRS2_SP_GRADE7 = 'B' then          Grades33 = 3;
```

```
else if CRS2_SP_GRADE7 = 'DC' then          Grades33 = 2;
else if CRS2_SP_GRADE7 = 'C' then           Grades33 = 2;
else if CRS2_SP_GRADE7 = 'DD' then          Grades33 = 1;
else if CRS2_SP_GRADE7 = 'D' then           Grades33 = 1;
else if CRS2_SP_GRADE7 = 'DF' then          Grades33 = 0;
else if CRS2_SP_GRADE7 = 'F' then           Grades33 = 0;
else if CRS2_SP_GRADE7 = 'W' then           Grades33 = '.';
else Grades33 = '.';


length Grades34 $10;
if                  CRS2_SU_GRADE1 = 'DA' then      Grades34 = 4;
else if CRS2_SU_GRADE1 = 'A' then           Grades34 = 4;
else if CRS2_SU_GRADE1 = 'DB' then          Grades34 = 3;
else if CRS2_SU_GRADE1 = 'B' then           Grades34 = 3;
else if CRS2_SU_GRADE1 = 'DC' then          Grades34 = 2;
else if CRS2_SU_GRADE1 = 'C' then           Grades34 = 2;
else if CRS2_SU_GRADE1 = 'DD' then          Grades34 = 1;
else if CRS2_SU_GRADE1 = 'D' then           Grades34 = 1;
else if CRS2_SU_GRADE1 = 'DF' then          Grades34 = 0;
else if CRS2_SU_GRADE1 = 'F' then           Grades34 = 0;
else if CRS2_SU_GRADE1 = 'W' then           Grades34 = '.';
else Grades34 = '.';


length Grades35 $10;
if                  CRS2_SU_GRADE2 = 'DA' then      Grades35 = 4;
else if CRS2_SU_GRADE2 = 'A' then           Grades35 = 4;
else if CRS2_SU_GRADE2 = 'DB' then          Grades35 = 3;
else if CRS2_SU_GRADE2 = 'B' then           Grades35 = 3;
else if CRS2_SU_GRADE2 = 'DC' then          Grades35 = 2;
else if CRS2_SU_GRADE2 = 'C' then           Grades35 = 2;
else if CRS2_SU_GRADE2 = 'DD' then          Grades35 = 1;
else if CRS2_SU_GRADE2 = 'D' then           Grades35 = 1;
else if CRS2_SU_GRADE2 = 'DF' then          Grades35 = 0;
else if CRS2_SU_GRADE2 = 'F' then           Grades35 = 0;
else if CRS2_SU_GRADE2 = 'W' then           Grades35 = '.';
else Grades35 = '.';


length Grades36 $10;
if                  CRS2_SU_GRADE3 = 'DA' then      Grades36 = 4;
else if CRS2_SU_GRADE3 = 'A' then           Grades36 = 4;
else if CRS2_SU_GRADE3 = 'DB' then          Grades36 = 3;
else if CRS2_SU_GRADE3 = 'B' then           Grades36 = 3;
```

```
else if CRS2_SU_GRADE3 = 'DC' then          Grades36 = 2;
else if CRS2_SU_GRADE3 = 'C' then           Grades36 = 2;
else if CRS2_SU_GRADE3 = 'DD' then          Grades36 = 1;
else if CRS2_SU_GRADE3 = 'D' then           Grades36 = 1;
else if CRS2_SU_GRADE3 = 'DF' then          Grades36 = 0;
else if CRS2_SU_GRADE3 = 'F' then           Grades36 = 0;
else if CRS2_SU_GRADE3 = 'W' then           Grades36 = '.';
else Grades36 = '.';


length Grades37 $10;
if                  CRS2_SU_GRADE4 = 'DA' then      Grades37 = 4;
else if CRS2_SU_GRADE4 = 'A' then           Grades37 = 4;
else if CRS2_SU_GRADE4 = 'DB' then          Grades37 = 3;
else if CRS2_SU_GRADE4 = 'B' then           Grades37 = 3;
else if CRS2_SU_GRADE4 = 'DC' then          Grades37 = 2;
else if CRS2_SU_GRADE4 = 'C' then           Grades37 = 2;
else if CRS2_SU_GRADE4 = 'DD' then          Grades37 = 1;
else if CRS2_SU_GRADE4 = 'D' then           Grades37 = 1;
else if CRS2_SU_GRADE4 = 'DF' then          Grades37 = 0;
else if CRS2_SU_GRADE4 = 'F' then           Grades37 = 0;
else if CRS2_SU_GRADE4 = 'W' then           Grades37 = '.';
else Grades37 = '.';



length Grades38 $10;
if                  CRS2_SU_GRADE5 = 'DA' then      Grades38 = 4;
else if CRS2_SU_GRADE5 = 'A' then           Grades38 = 4;
else if CRS2_SU_GRADE5 = 'DB' then          Grades38 = 3;
else if CRS2_SU_GRADE5 = 'B' then           Grades38 = 3;
else if CRS2_SU_GRADE5 = 'DC' then          Grades38 = 2;
else if CRS2_SU_GRADE5 = 'C' then           Grades38 = 2;
else if CRS2_SU_GRADE5 = 'DD' then          Grades38 = 1;
else if CRS2_SU_GRADE5 = 'D' then           Grades38 = 1;
else if CRS2_SU_GRADE5 = 'DF' then          Grades38 = 0;
else if CRS2_SU_GRADE5 = 'F' then           Grades38 = 0;
else if CRS2_SU_GRADE5 = 'W' then           Grades38 = '.';
else Grades38 = '.';


RUN;


/*now we multiply the grades by the number of hours
```

I want to set the results of the multiplication in a new
data folder
New folder name: new
Old folder name: theData*/

```
data new;
        set theData;
        by sas_stu_id;
        G1 = Grades * CRS_FA1_HR;
        G2 = Grades2 * CRS_FA2_HR;
        G3 = Grades3 * CRS_FA3_HR;
        G4 = Grades4 * CRS_FA4_HR;
        G5 = Grades5 * CRS_FA5_HR;
        G6 = Grades6 * CRS_FA6_HR;

        G7  = Grades7 * CRS_SP1_HR;
        G8  = Grades8 * CRS_SP2_HR;
        G9  = Grades9 * CRS_SP3_HR;
        G10 = Grades10 * CRS_SP4_HR;
        G11 = Grades11 * CRS_SP5_HR;
        G12 = Grades12 * CRS_SP6_HR;

        G13 = Grades14 * CRS_SU1_HR;
        G14 = Grades15 * CRS_SU2_HR;
        G15 = Grades16 * CRS_SU3_HR;
        G16 = Grades17 * CRS_SU4_HR;
        G17 = Grades18 * CRS_SU5_HR;

        G18 = Grades19 * CRS2_FA1_HR;
        G19 = Grades20 * CRS2_FA2_HR;
        G20 = Grades21 * CRS2_FA3_HR;
        G21 = Grades22 * CRS2_FA4_HR;
        G22 = Grades23 * CRS2_FA5_HR;
        G23 = Grades24 * CRS2_FA6_HR;

        G24 = Grades27 * CRS2_SP1_HR;
        G25 = Grades28 * CRS2_SP2_HR;
        G26 = Grades29 * CRS2_SP3_HR;
        G27 = Grades30 * CRS2_SP4_HR;
        G28 = Grades31 * CRS2_SP5_HR;
        G29 = Grades32 * CRS2_SP6_HR;
        G30 = Grades33 * CRS2_SP7_HR;

        G31 = Grades34 * CRS2_SU1_HR;
        G32 = Grades35 * CRS2_SU2_HR;
        G33 = Grades36 * CRS2_SU3_HR;
```

```
        G34 = Grades37 * CRS2_SU4_HR;
run;


/*Here we sum up all the course hours each student has registered for thus far
HR = total number of hours*/

/*make sure to run it all at once*/

data new2;
        set new;
        by sas_stu_id;

HR= sum(CRS_FA1_HR, CRS_FA2_HR, CRS_FA3_HR, CRS_FA4_HR, CRS_FA5_HR,
CRS_FA6_HR,
        CRS_SP1_HR, CRS_SP2_HR, CRS_SP3_HR, CRS_SP4_HR, CRS_SP5_HR,
CRS_SP6_HR,
        CRS_SU1_HR, CRS_SU2_HR, CRS_SU3_HR, CRS_SU4_HR, CRS_SU5_HR,
        CRS2_FA1_HR, CRS2_FA2_HR, CRS2_FA3_HR, CRS2_FA4_HR, CRS2_FA5_HR,
CRS2_FA6_HR,
        CRS2_SP1_HR, CRS2_SP2_HR, CRS2_SP3_HR, CRS2_SP4_HR, CRS2_SP5_HR,
CRS2_SP6_HR, CRS2_SP6_HR,
        CRS2_SU1_HR, CRS2_SU2_HR, CRS2_SU3_HR, CRS2_SU4_HR);


/*Here we sum up the 'grade*no of hours of the course h1g1+ h2g2....+HzGz
GHs = sum of grade * course_hours*/


GHs = sum(G1, G2, G3, G4, G5, G6,
                        G7, G8, G9, G10, G11, G12,
                        G13, G14, G15, G16, G17,
                        G18, G19, G20, G21, G22, G23,
                        G24, G25, G26, G27, G28, G29, G30,
                        G31, G32, G33, G34);


/*Now to calculate each students CGPA from the no of hours
and grade for courses calculated above in first and second year*/



CGPA = GHs / HR ;
run;

/*----------------------------------------*/
```

```
/*Calculating mean median and std,
of CGPA by variables
*/


proc sort data = new2;
        by stem_major;
        run;

proc means mean median std;
        by stem_major;
        var cgpa;
        run;



proc means mean std;
        var CGPA;
        run;


proc sort data = new2;
        by gender_cohort;
        run;
proc means mean median std;
        by gender_cohort;
        var CGPA;
        run;


proc sort data = new2;
        by ethnicity;
        run;

proc means mean median std;
        by ethnicity;
        var CGPA;
        run;


proc sort data = new2;
        by year;
        run;

proc means mean median std;
        by year;
```

```
        var CGPA;
        run;


proc sort data = new2;
        by first_gen_nodeg_0;
        run;

proc means mean median std;
        by first_gen_nodeg_0;
        var CGPA;
        run;


proc sort data = new2;
        by pell_cohort;
        run;

proc means mean median std;
        by pell_cohort;
        var CGPA;
        run;

proc sort data = new2;
        by persist_1;
        run;

proc means mean median std;
        by persist_1;
        var CGPA;
        run;

/*--------------------------------------*/

/*code to confirm normality of the variable
using qq-plot*/

proc univariate data=new2 normal;
                qqplot CGPA /Normal(mu=est sigma=est color=red l=1);
                        run;

/*---------------*/

/*------------------------*/

/*Performing a two sample test on CGPA by stem_major
```

```
default alpha is 0.05*/

proc freq data= theData;
   table stem_major;

  run;


proc ttest h0=0 side=2 ;
            title "Two sample t-test of the CGPA by STEM_MAJOR" ;
            class stem_major;
            var CGPA;
            run;


/*THE ANOVA*/
/*because we ran code without 'means' (line 1291), and found p-value
less than predetermined alpha (0.05),
we then added the line : 'means Ethnicity /tukey';

hoctest helps navigate wether we have equal variances or not.
tukey test does the pairwise comparison when variances arent equal
*/
proc anova;
            class ethnicity;
            model CGPA = Ethnicity;
            title "ANOVA Testing of CGPA by Ethnicity" ;
            means Ethnicity /hoctest welch;
            means Ethnicity /tukey;
run;


/*To carry out test of normality of variance assumptions (for the ANOVA)
I have to change the categorical variable 'ethinicity'
to numeric*/

proc sort data = new2;
            by ethnicity;
            run;

proc univariate data= new2 normal;
                        qqplot CGPA /Normal(mu=est sigma=est color=red l=1);
                        by ethnicity;
                        run;
/*
```

```
****making a scatter matrix for the dataset****
*/
/*first is correct, second plot looks funny******/

proc sgscatter data= new2;
	title "Scatter Matrix of CGPA" ;
	matrix CGPA act_m act_e sat_v sat_m / diagonal=(histogram kernel);

run;


proc sgscatter data=new2;
  title "Scatterplot Matrix for CGPA Data";
  matrix CGPA gender_cohort first_gen_nodeg_0 pell_cohort persist_1
  / diagonal=(histogram kernel);

run;


/*----------------------------------------*/




/*THIS IS OTHER PRACTICE; not NEEDED*/
proc sgscatter data= new2;
	title "Scatter Matrix of CGPA" ;
	matrix CGPA act_m act_e sat_v sat_m / group= CGPA diagonal=(histogram kernel);

run;


proc sgscatter data= new2;
	title "Scatter Matrix of CGPA" ;
	matrix CGPA act_m act_e sat_v sat_m /group= stem_major diagonal=(histogram
kernel) ;
run;

proc sgscatter data= new2;
	title "Scatter Matrix of CGPA" ;
	matrix CGPA act_m act_e sat_v sat_m /group= ethnicity diagonal=(histogram kernel) ;

run;

proc sgscatter data= new2;
	title "Scatter Matrix of CGPA" ;
```

```
	matrix CGPA act_m act_e sat_v sat_m /group= pell_cohort diagonal=(histogram
kernel) ;
run;

proc sgscatter data= new2;
	title "Scatter Matrix of CGPA" ;
	matrix CGPA act_m act_e sat_v sat_m /group= gender_cohort diagonal=(histogram
kernel) ;
run;

proc sgscatter data= new2;
	title "Scatter Matrix of CGPA" ;
	matrix CGPA act_m act_e sat_v sat_m /group= first_gen_nodeg_0
diagonal=(histogram kernel) ;
run;

/*----------------------------------------------*/




/*Now we calculate a regression model*/
/*-------------------------*/
/*first model*/
proc reg;
	model CGPA = act_m act_e sat_v sat_m gender_cohort
	first_gen_nodeg_0 pell_cohort persist_1;
	run;

proc reg;
	model CGPA = act_m act_e sat_v sat_m;
	run;

/*-----------------*/
proc reg;
	model CGPA = sat_v sat_m gender_cohort first_gen_nodeg_0 pell_cohort persist_1;;
	run;


proc reg;
	model CGPA = sat_v sat_m gender_cohort ;
	run;

proc reg;
	model CGPA = act_m act_e gender_cohort first_gen_nodeg_0 pell_cohort persist_1;;
	run;
```

```
proc reg;
        model CGPA = act_m act_e first_gen_nodeg_0 pell_cohort persist_1;;
        run;
/*-----------------------------*/



/*second model* *THIS IS THE BEST MODEL*/
proc reg;
        model CGPA = act_m act_e gender_cohort persist_1;
        title 'Best Regression Model';
        run;

proc reg;
        model CGPA = act_m act_e gender_cohort ;
        title 'Best Regression Model';
        run;


/*third model*/
proc reg;
        model CGPA = act_m act_e persist_1;
        title 'Best Regression Model';
        run;

proc reg;
        model CGPA = act_m gender_cohort persist_1;
        run;

/*fourth model*/
proc reg;
        model CGPA = act_m act_e gender_cohort;
        run;

/*-----------------------------*/
/*correlation matric to see if correlation os higher than 0.8 FIRST*/
proc corr;
        var CGPA act_m act_e sat_v sat_m gender_cohort
        first_gen_nodeg_0 pell_cohort persist_1;
        Title 'Correlation Matrix of the Regression Model Variables';
        run;



/*colinearity test VIF & TOL FOR THE REGRESSION MODEL*/
```

```
proc reg;
        model CGPA = act_m act_e sat_v sat_m gender_cohort
        first_gen_nodeg_0 pell_cohort persist_1
        /  tol vif collin;
        run;
```