Topic: A study of pattern between the academic behaviour of students

Introduction

As many of us are curious about if there is any pattern between the academic behaviour of students, we are going to study this topic by looking into 3 different criteria.

Here are some technical terms that we need to know. Correlation coefficient is the measure of the degree of linear dependence between two quantitative variables. If correlation coefficient is close to 0, it means that there is no or weak relationship between two variables. If magnitude of correlation coefficient is close to 1, it means that there is strong relationship between two variables. Density graph is a visualisation of distribution of data over a continuous interval or time period. Two-sample t-test is proposed to discern the difference between two sample means to make conclusions about differences in the parameters of the target populations. Boxplot is a graphical methodology to display the median, quartiles, and extremes of a data set on a number line to show the distribution of the data. Scatter plot is a mathematical diagram using to show the relationship between a set of data.

This set of data is an observation study. It is downloaded from GitHub. The data is collected by a professor at a midwestern university for the sake of creating a model to estimate students' GPA for their first year of the university.

The primary objective is to study if students who have high GPA in high school get high GPA in the first-year college. The secondary objective is to study if there are correlation between family's education level and students' academic performance. The tertiary objective is to study if gender affecting our choices in selecting subjects.

Method

<u>Primary objective:</u>

The FirstYearGPA data includes details from a study of 219 first year students at a midwestern college. In this criteria, we will look into two set of data which are first-year college GPA and high

school GPA. Those two sets of data reflects academic performance of students in the first year of their college and high school respectively and we will examine them accordingly. As grade A describes excellent performance and 3.7 is the minimum grade point that can get from A range, those GPA who score above or equal to 3.7 have excellent performance on average. Thus ,we will consider high GPA to be the GPA that is higher or equal to 3.7 and those who score below this boundary will be considered as low GPA.

Since we are going to investigate if there is a relationship between the academic performance of students in high school and college, the most appropriate analysis to answer this question is Pearson's correlation coefficient. Pearson's correlation coefficient is a measure of the degree of linear dependence between two quantitative variables. We can see how strong variables are correlated by checking the value of Pearson's correlation coefficient. Moreover, bar char and scatter plot are also used to analyze the data and draw conclusion from them. Both of them are mathematical diagram to examine the data and see the implication behind the data by evaluating them.

Secondary objective:

To evaluate this question, 4 criteria is used to make a conclusion. familyedu stands for the family education levels. GPA stands for First-year college GPA on a 0.0 to 4.0 scale .HSGPA stands for High school GPA on a 0.0 to 4.0 scale. Sumsat stands for Sum of Verbal/critical reading SAT score and Math SAT score. The values of GPA, HSGPA and sumsat show each student's academic performances. The Higher the value ,the better the academic performance for all three criteria. If familyedu has a value of 1, it means that the student is the first in his/her family to attend college, so we assume their family's education level is lower. If familyedu has a value of 0, it means that the student is not the first in his/her family to attend college, so we assume their family's education level is higher .

In order to examine this question, the most appropriate analysis to answer this question is correlation coefficient. Correlation coefficient is a measure of the strength of linear relationship between two variables. If magnitude of the correlation coefficient is close to 0, it means that there is no or weak

relationship between two variables. If magnitude of the correlation coefficient is close to 1, it means that there is strong relationship between two variables. Furthermore, boxplot is also used to study the data. It is a kind of graphical methodology to display the median, quartiles, and extremes of a data set on a number line to show the distribution of the data.

Tertiary objective:

As our target is to explore the relationship between gender and the credit hours earned in different areas, we select the data of "number of credit hours earned" and separating them by gender to make comparison. First, we choose to plot the histogram to have an overview on the differences on the same subject with different gender and the differences on the same gender with different subject. We also plot the mean of the data in each graph, we found that the mean of male students having credit hour on social science (~7.44) is slightly greater than female students (~7.09), the mean of male students having credit hour on humanity (~12.96) is slightly less than female students (~13.24). In other view, we see that both male and female students having more credit hours on humanity rather than social science. After that, we want to use one-sided two-sample t-test to test our hypothesis. We want to perform two test, the first test is 'male students have more credit hour in social science than female students' with null hypothesis is 'both population mean are equal' and alternative hypothesis is 'the population mean difference is greater than 0'. The second test is 'male students have fewer credit hour in humanity than female students' with null hypothesis is 'both population mean are equal' and alternative hypothesis is 'the population mean difference is less than 0'. Before doing so, Shapiro–Wilk test and Bartlett's test should be done in order to check the normality assumptions as well as homogeneity of variance respectively. We will test the distribution at significance level 0.05 for convention purpose.

Result

Primary objective:

From the analysis that we have done below, it is clear to see that the Pearson's correlation coefficient is equal to 0.4468873 which is less than 0.5. This indicates that the high school GPA has a weak

positive linear relationship with the first-year college GPA. Students who score high in high school doesn't imply that they will get high GPA in their first year of college. Based on the data summarised in table 1 and figure 2, we can see that ,in the sample of 219 students, 73 of them score higher or equal to 3.7 in their high school but only 13 of them can get high GPA(>=3.7) in their first year of college. More than 80% of student who score greater than or equal to 3.7 in high school get a GPA below 3.7 in their first year of college. However, those who score low in high school is likely to get low GPA(<3.7) in their first year of college. According to the data in the table 1 and figure 2, 146 of them score below 3.7 in their high school. Only 4 students who score below 3.7 in high school can score above or equal to 3.7 in their first year of college. More than 95% of them score below 3.7 in the first year of college.

Secondary objective:

Based on the data shown in table 1 and figures 7, 8 and 9, the correlation coefficient of sum of SAT score and family education level is a -0.250557 which is negatively weak correlation. This implies that students whose family has low educational level tend to have lower sum of SAT score. The correlation coefficient of high school GPA and family education level is 0.0641858 which is positively weak correlation. This indicates that students whose family has low educational level tend to have higher high school GPA. Correlation coefficient of first year college GPA and family education level is -0.156577 which is negatively weak correlation. This show that students whose family has low educational level tend to have lower high school GPA.

Tertiary objective:

According to figure 3 and 4, we can see that in the humanity course, females are slightly getting more credit hours than males, while the males overall getting more credit hours than females over the social sciences area. As we want to perform a one-sided two-sample T test, we must check the normality of all data set and the variance. We use the shapiro.test() to test the normality. The p-value of the test result of male students' credit hours on social science is 0.00053, which is not significant enough to say that it follows the normal distribution. The p-value of the test result of male students' credit hours

on humanity is 0.00024, p-value of the test result of female students' credit hours on social science is 0.00028, p-value of the test result of female students' credit hours on humanity is 0.0094. All four test results are saying that they are not following the normal distribution at significance level of 0.05. However, the p-value of Bartlett's test on total credit hours on social science and humanity of male and female students in Bartlett's test in which the p-value of each of them are greater than 0.05 as table 1 shown. Unfortunately, we cannot perform the one-sided two-sample T test to get more evidence to support our observations and hypothesis.

Conclusion

Primary objective:

The relationship between students' academic performance in first year college and high school is weak. Student who perform well in high school can't be guarantee that their GPA in the first year of college will score above or equal to 3.7 again but those who score below 3.7 in high school is likely to score below 3.7 in their first year of college.

Secondary objective:

All three relationships are very weak with no consistency. Therefore, there is not enough evidence to draw a conclusion that there is relationship between education level of students' family and student's performance in academics.

Tertiary objective:

After we have constructed the graph and making comparison, we can conclude that we observed there is diversity on males and females credit hours on social science and humanity. The differences between gender is very minor while the differences between subjects is much larger. We failed to use the one-sided two-sample T test to verify our hypothesis and observations. Hope that we can have more knowledge to test the data which is not following the normal distribution than we can discover more on this objective.

Graph:

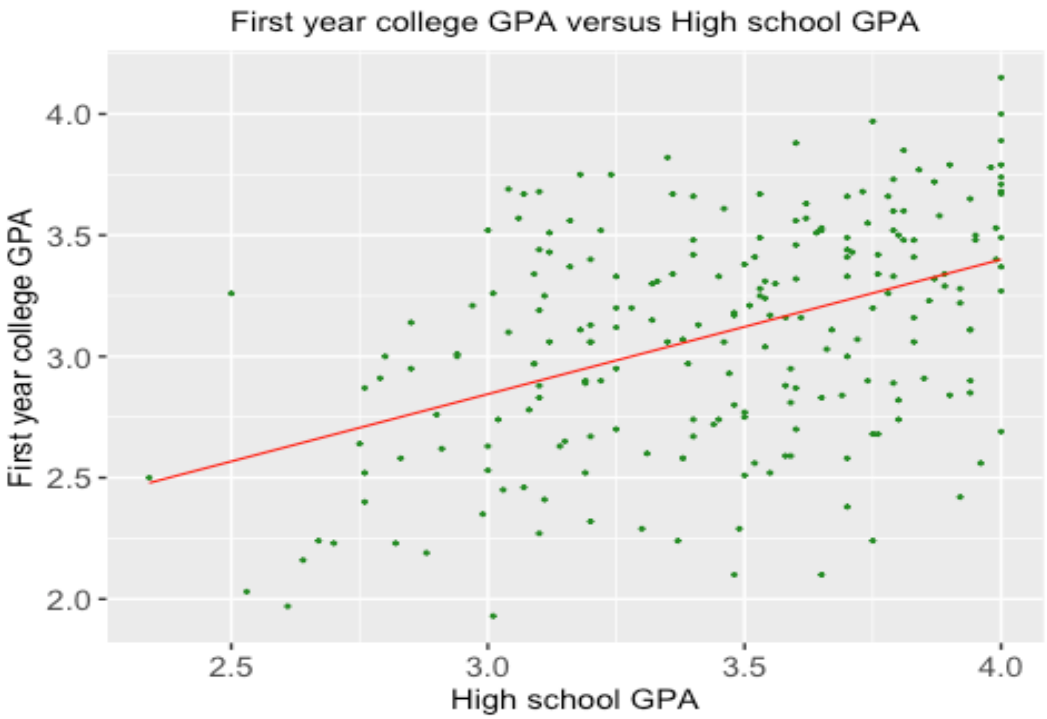

Figure 1

Formula Pearson's Correlation Coefficient r for sample X and sample Y

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

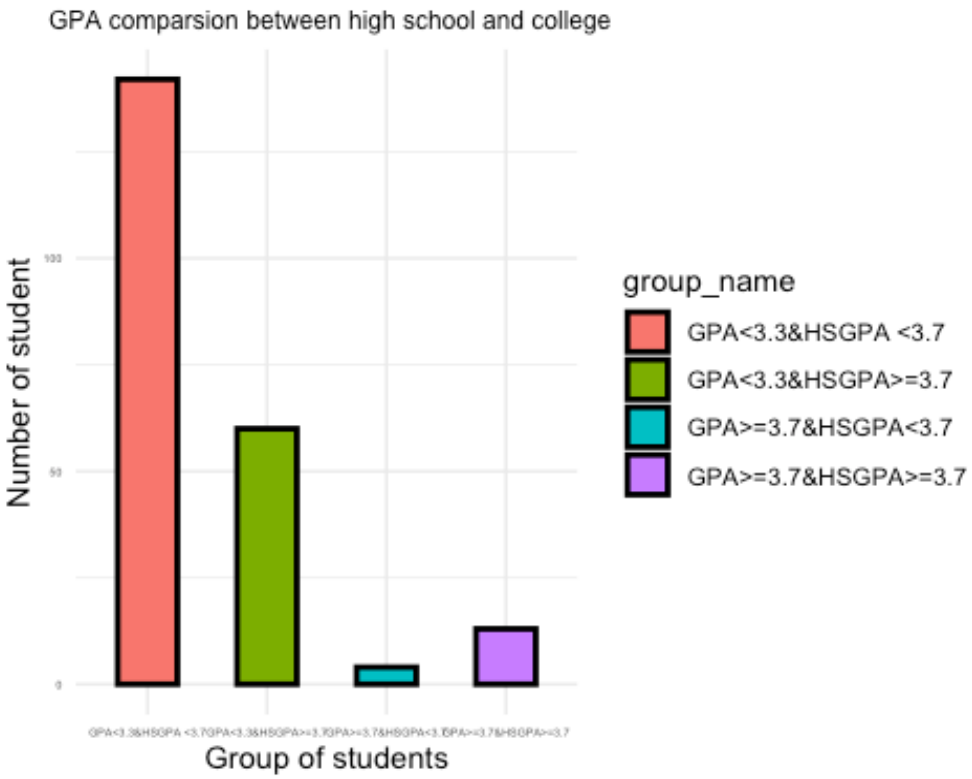

Figure 2

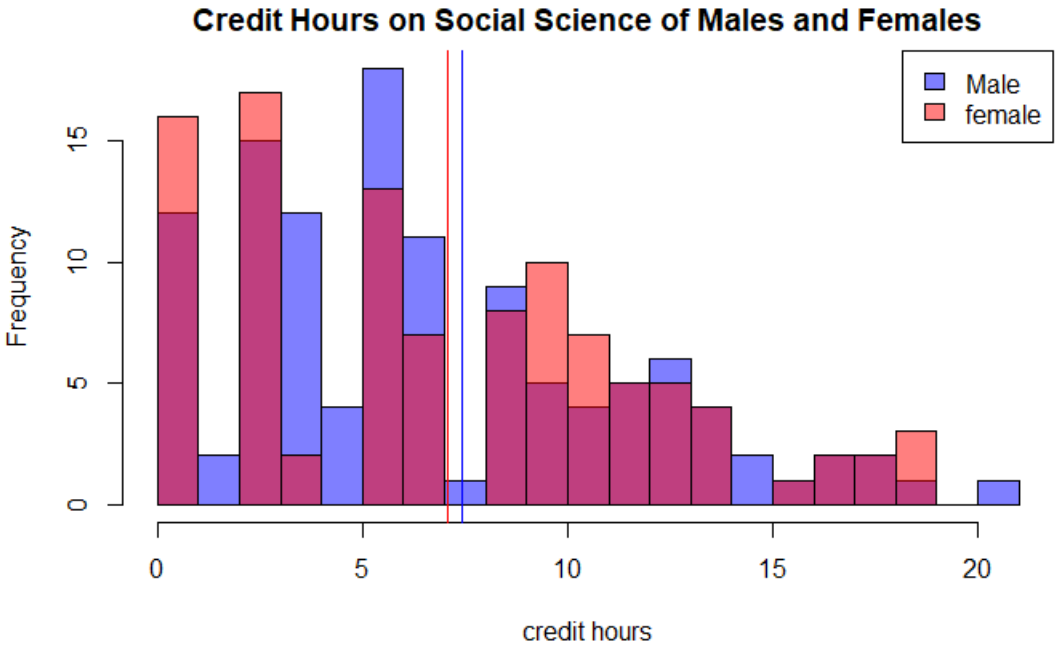

Figure 3

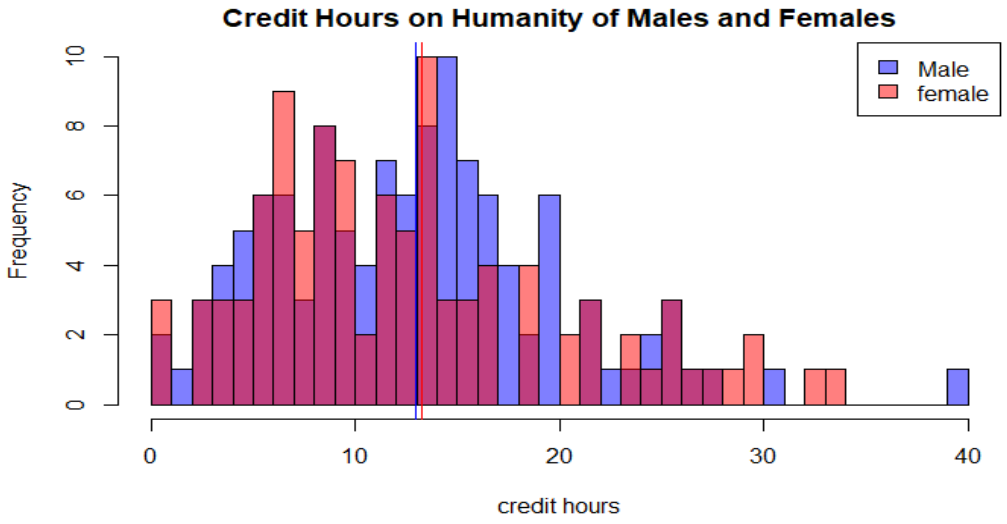

Figure 4

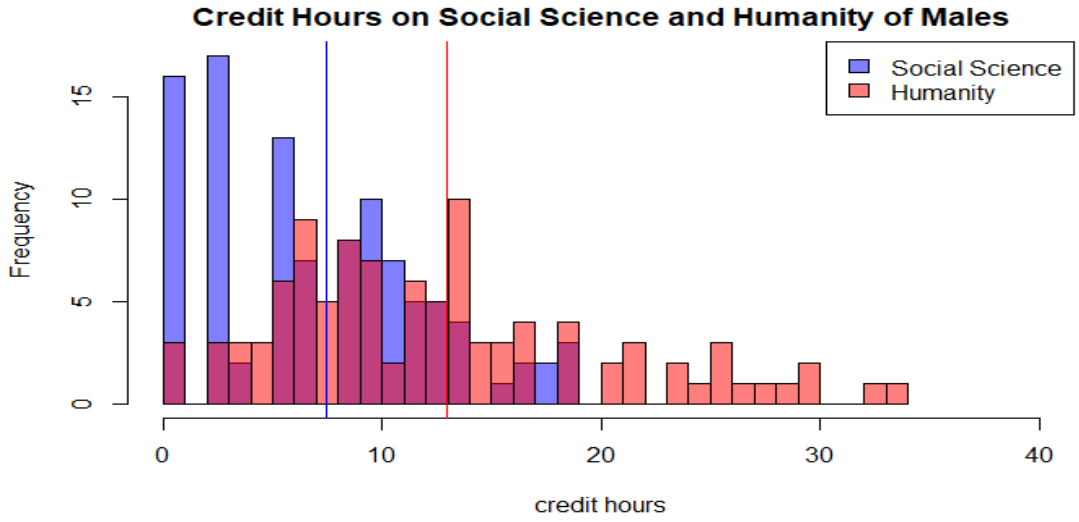

Figure 5

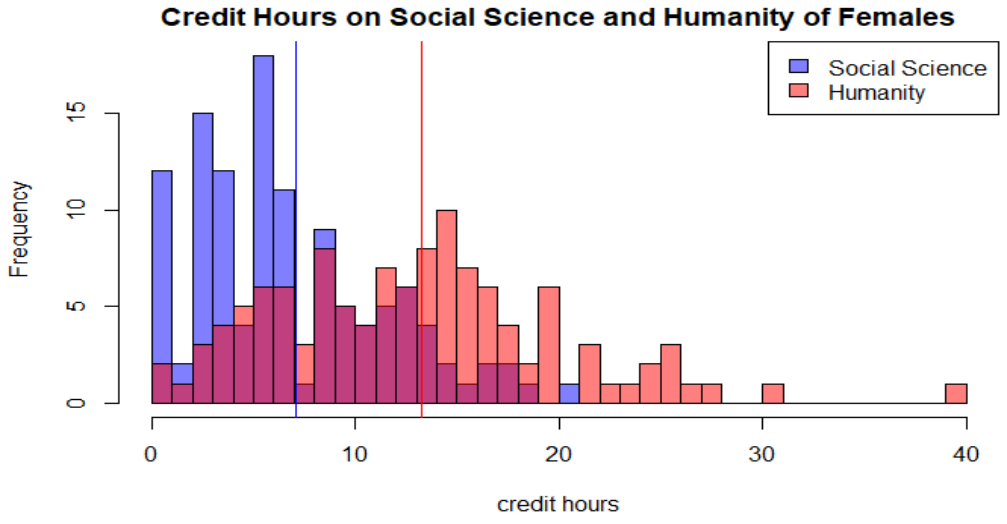

Figure 6

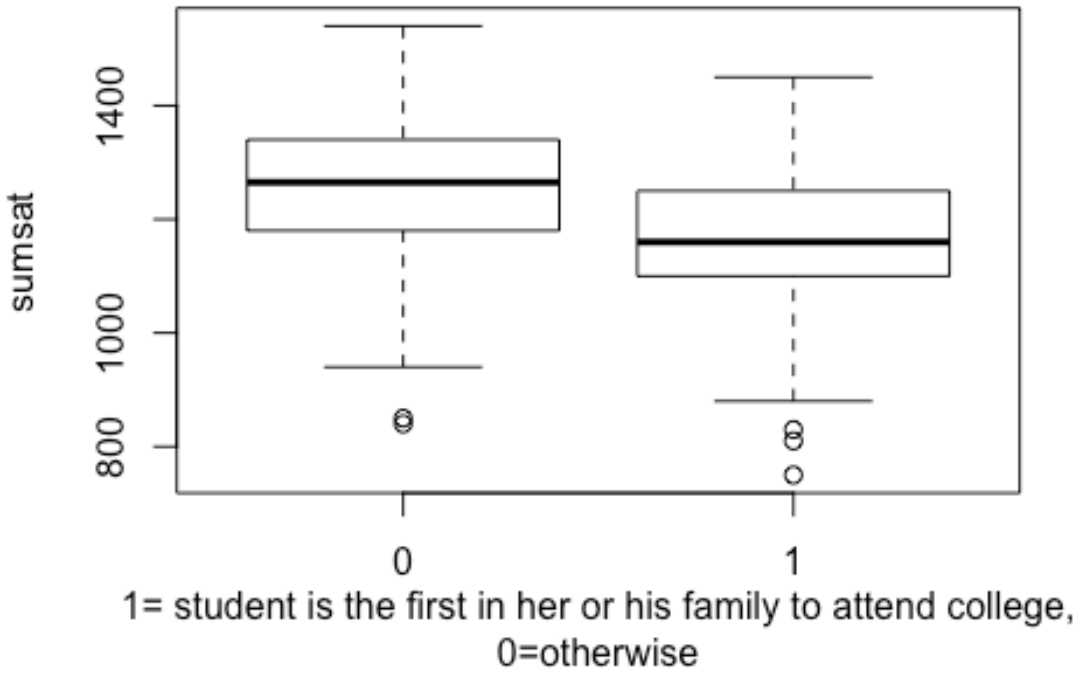

Figure 7

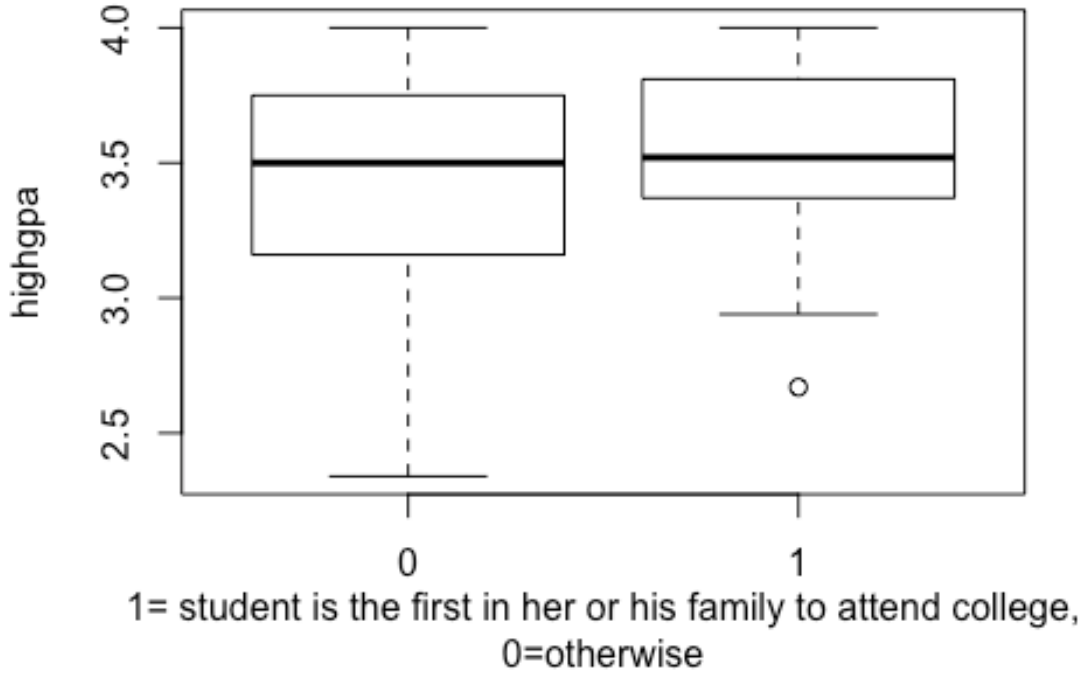

Figure 8

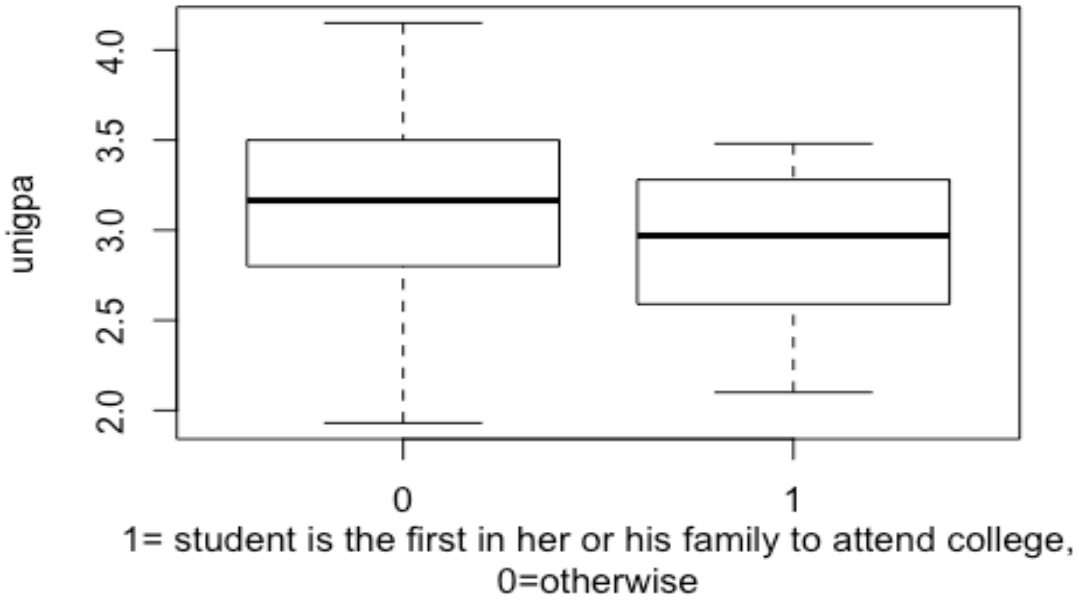

Figure 9

Appendix

Item	value
Pearson's correlation coefficient for high school GPA and first year college GPA	0.4468873
Pearson's correlation coefficient for sum of SAT score and family education level	-0.250557
Pearson's correlation coefficient for family education level and high school GPA	0.0641858
Pearson's correlation coefficient for family education level and first year college GPA	-0.156577
The p-value of Bartlett's test on credit hours on social science of male and female students	0.3965
The p-value of Bartlett's test on credit hours on humanity of male and female student	0.3413
Mean of credit hours on social science of male students	7.436275
Mean of credit hours on humanity of male students	12.96275
Mean of credit hours on social science of female students	7.08547
Mean of credit hours on humanity of female students	13.23504
The p-value of shapiro test on credit hours on social science of male students	0.0005277
The p-value of shapiro test on credit hours on humanity of male students	0.000241
The p-value of shapiro test on credit hours on social science of female students	0.0002822
The p-value of shapiro test on credit hours on humanity of female students	0.009397

Table 1