Homework 1C - DATA-312

Jeffrey Williams

26, March 2022

Abstract

This writeup explores the dataset, FirstYearGPA, from the R library, Stat2Data. Categorical variables in this dataset utilize boolean values for identification. These variables are as follows: Male, FirstGen, White, CollegeBound. Numerical variables are as follows: GPA, HSGPA, SATV, SATM, HU, SS.

Accessible as a library in R, this dataset comprises information from "a sample of 2019 first year students at a midwestern college", with the original intention of constructing an informed prediction of their first year GPA using various categorical and numerical variables.

NUMERICAL VARIABLES:

GPA: Represents the grade point average of any given student in their first year of college.

HSGPA: Represents the grade point average of any given student in high school.

SATV: Represents the SAT score of any given student with respect to the Verbal/Literacy portion.

SATM: Represents the SAT score of any given student with respect to the Mathematics portion.

HU: Representative of how many credit hours any given student has earned in high school humanities courses.

SS: Representative of how many credit hours any given student has earned in high school social science courses.

CATEGORICAL VARIABLES:

FirstGen: Boolean variable (0/1) representative of whether student is first in their family to attend college.

White: Boolean variable (0/1) representative of whether student identifies as White.

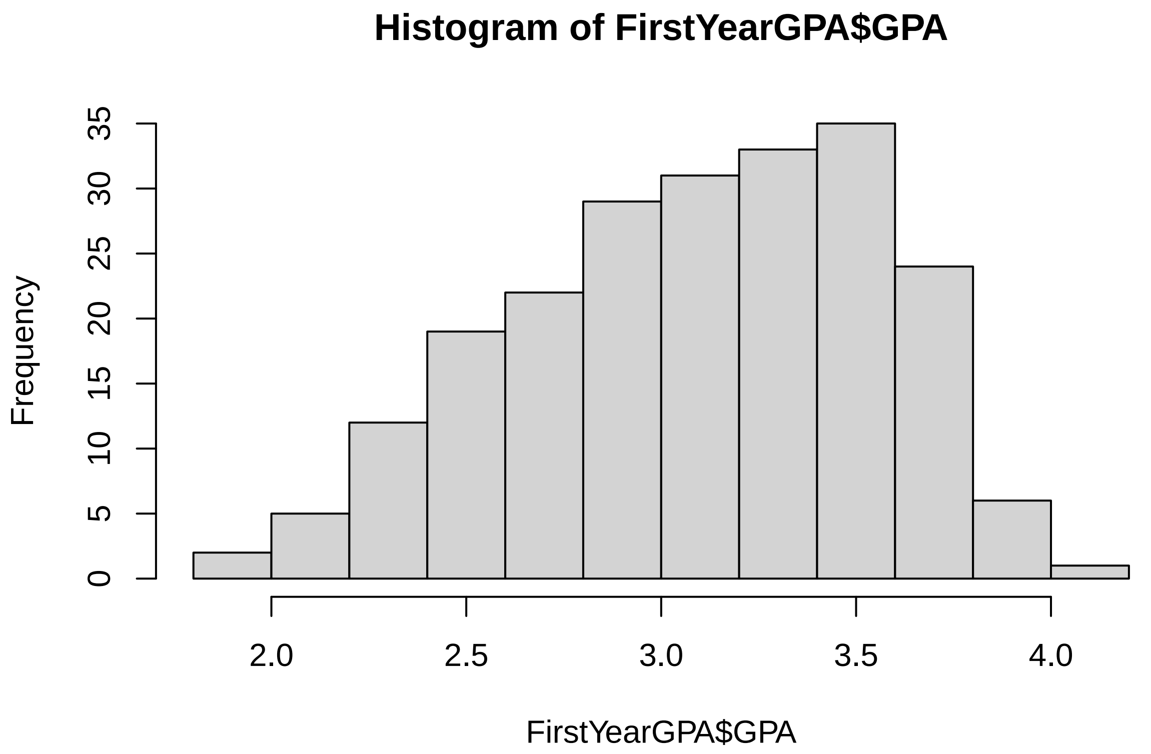
CollegeBound: Boolean variable (0/1) representative of whether student attended a high school where 50% of the student body (or greater) indicates intention to attend college.

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 422014 22.6     883898 47.3   643711 34.4
## Vcells 771643  5.9    8388608 64.0  1650190 12.6
```

```
##    GPA HSGPA SATV SATM Male   HU   SS FirstGen White CollegeBound
## 1 3.06  3.83  680  770    1  3.0  9.0        1     1            1
## 2 4.15  4.00  740  720    0  9.0  3.0        0     1            1
## 3 3.41  3.70  640  570    0 16.0 13.0        0     0            1
## 4 3.21  3.51  740  700    0 22.0  0.0        0     1            1
## 5 3.48  3.83  610  610    0 30.5  1.5        0     1            1
## 6 2.95  3.25  600  570    0 18.0  3.0        0     1            1
```

```
##       GPA            HSGPA            SATV            SATM
##  Min.   :1.930   Min.   :2.340   Min.   :260.0   Min.   :430.0
##  1st Qu.:2.745   1st Qu.:3.170   1st Qu.:565.0   1st Qu.:580.0
##  Median :3.150   Median :3.500   Median :610.0   Median :640.0
##  Mean   :3.096   Mean   :3.453   Mean   :605.1   Mean   :634.3
##  3rd Qu.:3.480   3rd Qu.:3.760   3rd Qu.:670.0   3rd Qu.:690.0
##  Max.   :4.150   Max.   :4.000   Max.   :740.0   Max.   :800.0
##       Male              HU              SS            FirstGen
##  Min.   :0.0000   Min.   : 0.00   Min.   : 0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 8.00   1st Qu.: 3.000   1st Qu.:0.0000
##  Median :0.0000   Median :13.00   Median : 6.000   Median :0.0000
##  Mean   :0.4658   Mean   :13.11   Mean   : 7.249   Mean   :0.1142
##  3rd Qu.:1.0000   3rd Qu.:17.00   3rd Qu.:11.000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :40.00   Max.   :21.000   Max.   :1.0000
##      White       CollegeBound
##  Min.   :0.00   Min.   :0.0000
##  1st Qu.:1.00   1st Qu.:1.0000
##  Median :1.00   Median :1.0000
##  Mean   :0.79   Mean   :0.9224
##  3rd Qu.:1.00   3rd Qu.:1.0000
##  Max.   :1.00   Max.   :1.0000
```

```
##          GPA        HSGPA         SATV         SATM         Male           HU
##    "numeric"    "numeric"    "integer"    "integer"    "integer"    "numeric"
##           SS     FirstGen        White CollegeBound
##    "numeric"    "integer"    "integer"    "integer"
```



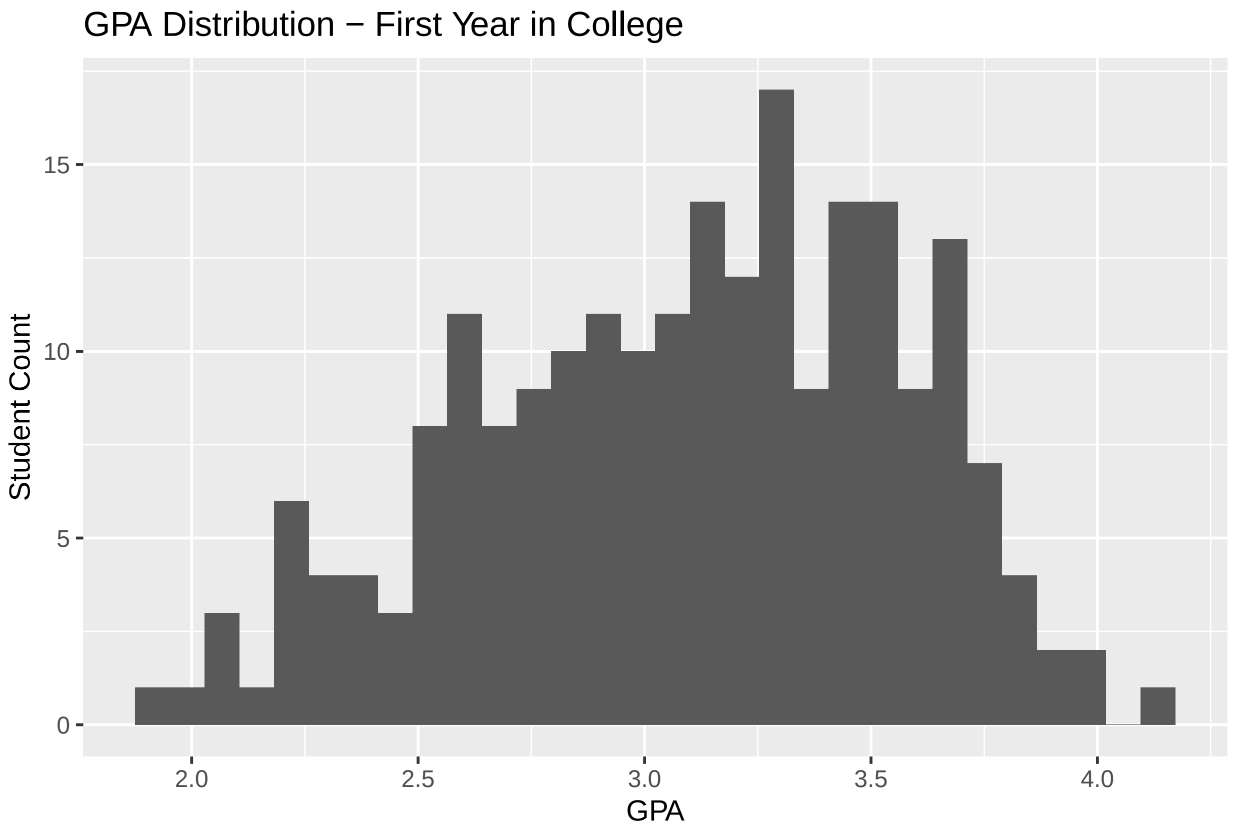
```
## [1] 3.096164
```

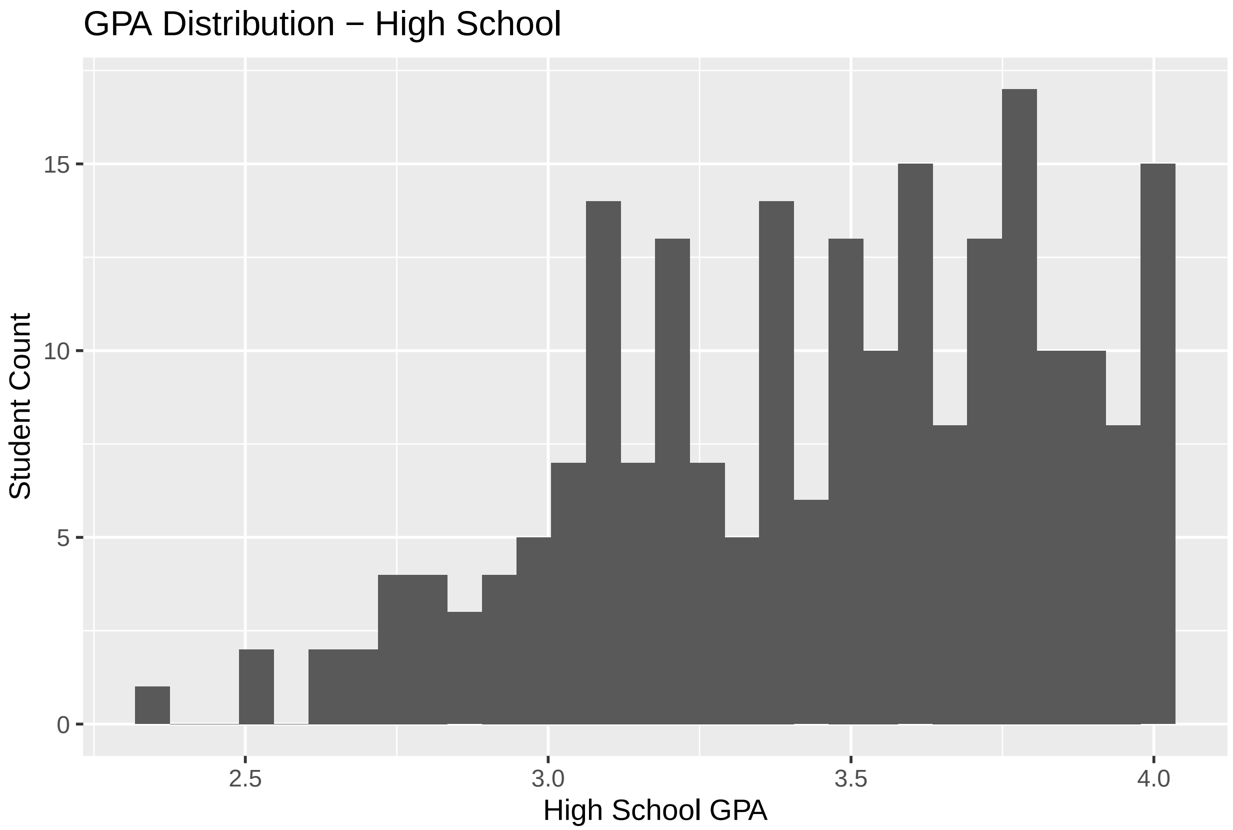
```
## [1] 0.2166678
```

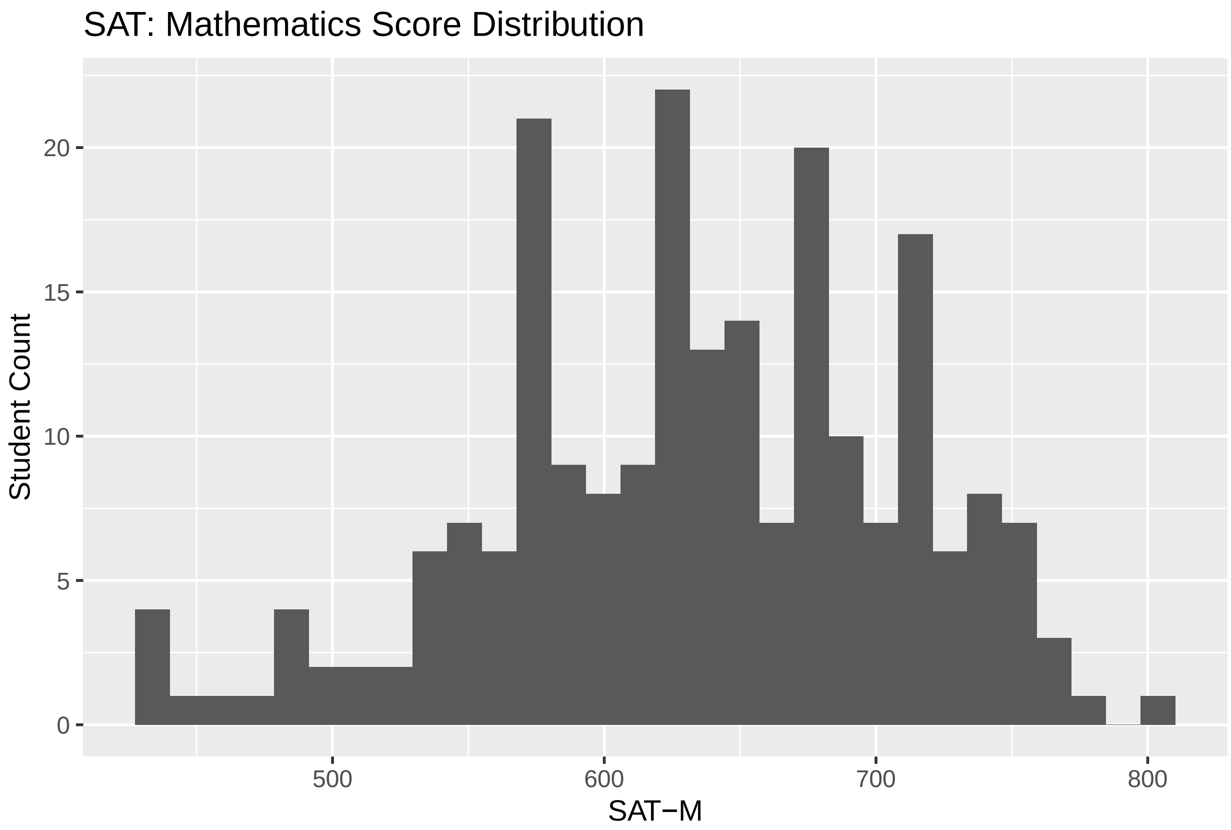
```
## [1] 0.4654759
```

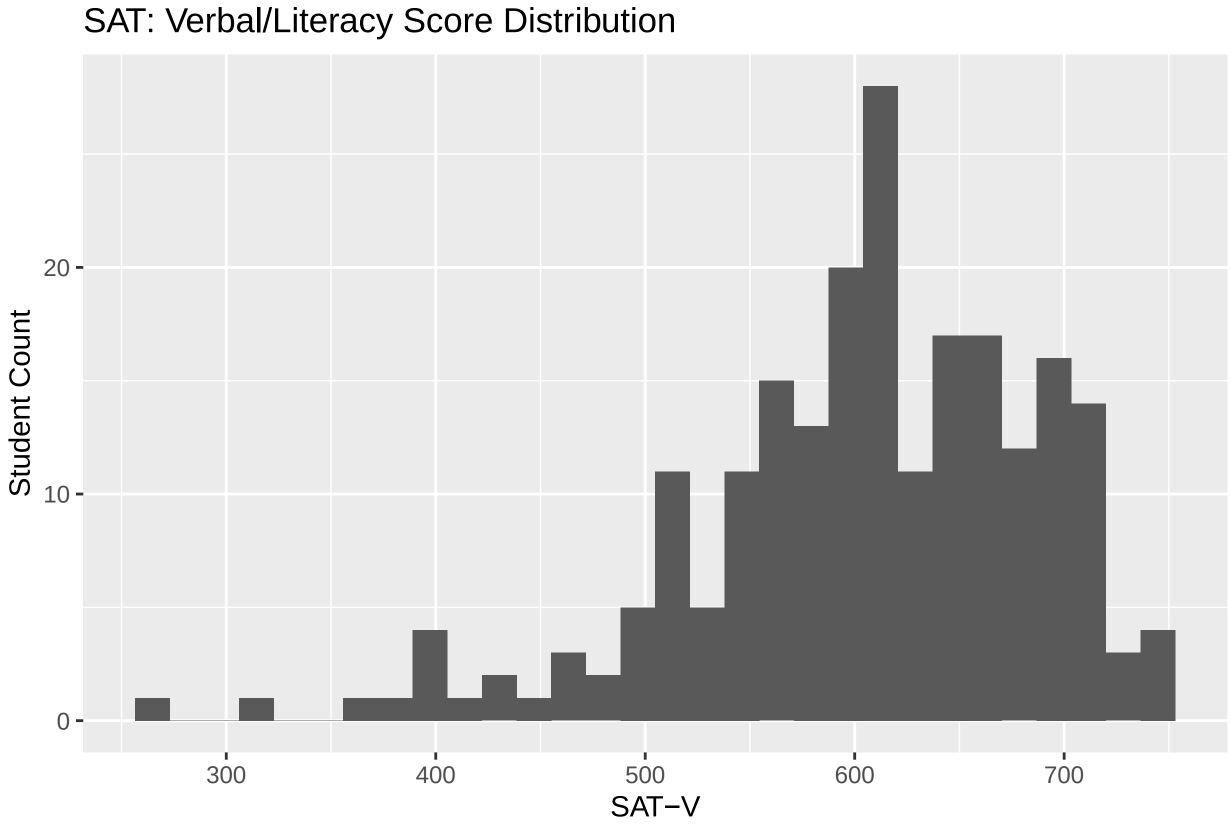
```
## [1] 0.4468873
```

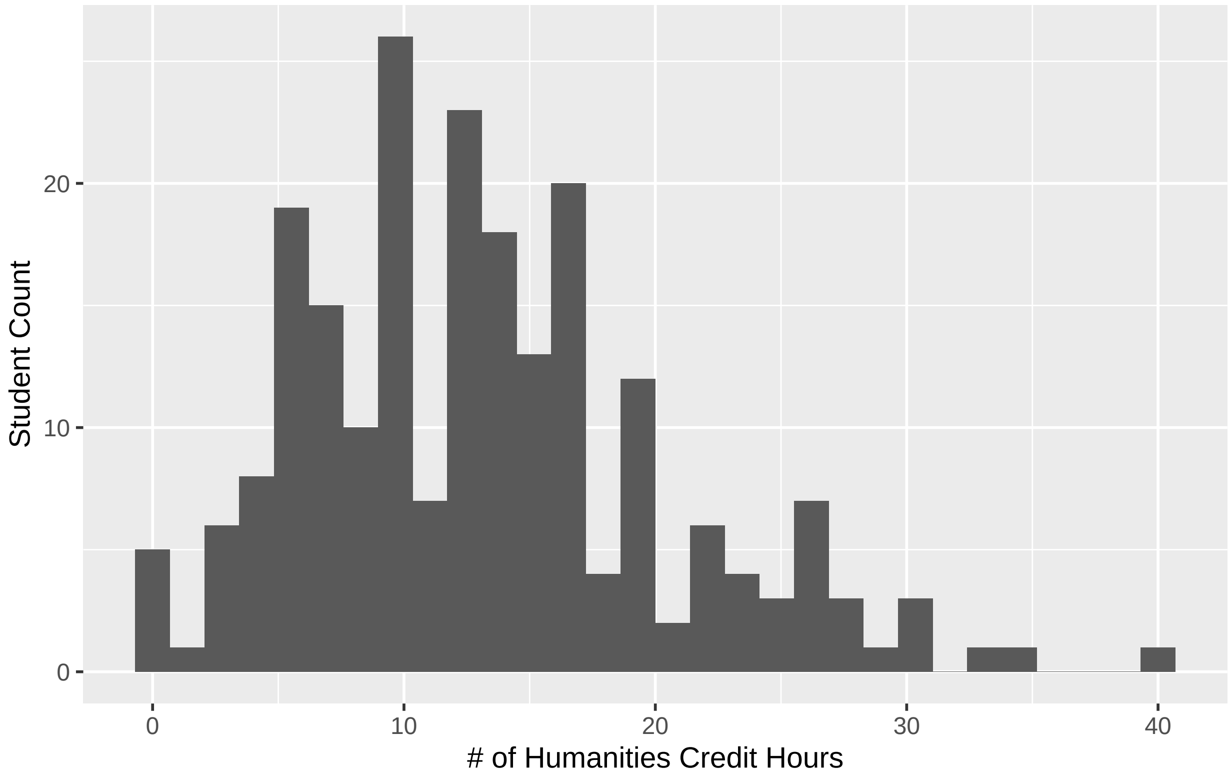
The following provides visual representation of the distributions of the variables in the dataset that are classified as numerical. These distributions permit further conceptualization of the standings of students at the closure of their high school careers with respect to their GPA, SAT scores, and the numbers of credit hours they have earned for various types of classes.


GPA Distribution − First Year in College
Student Count
15
10
5
0
2.0
2.5
3.0
3.5
4.0
GPA

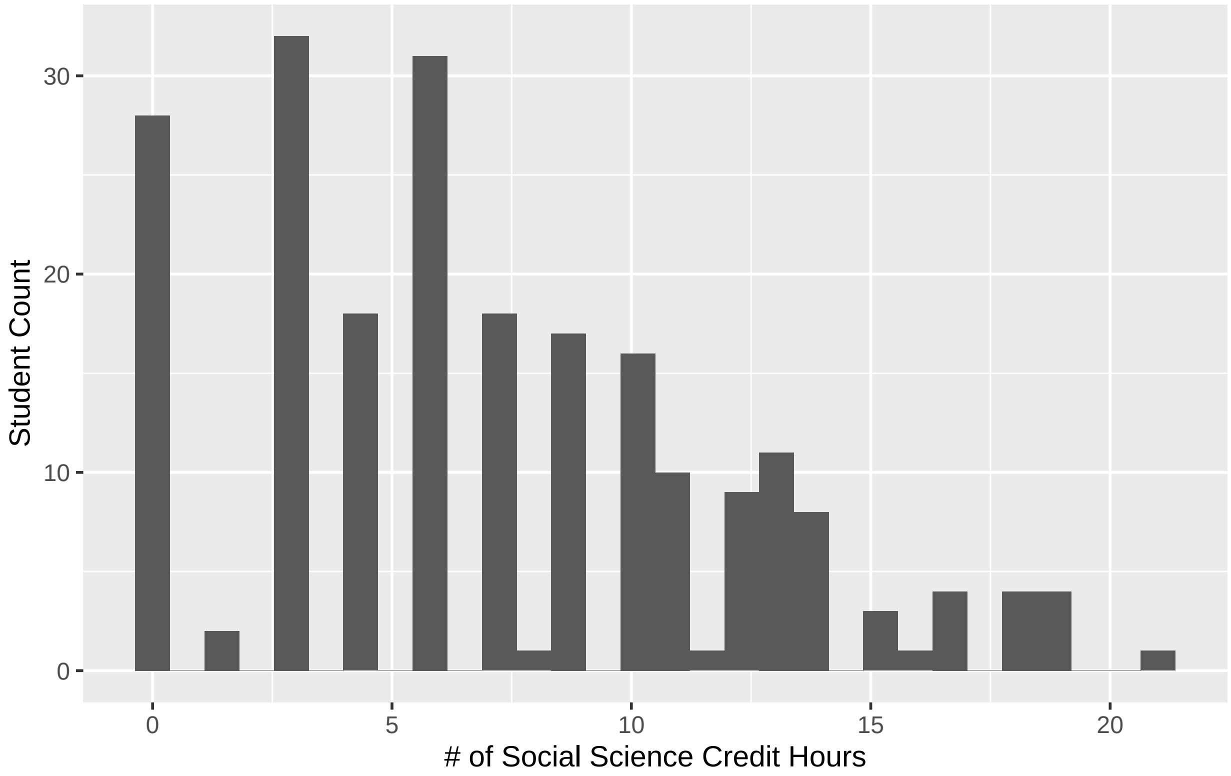

GPA Distribution − High School
Student Count
15
10
5
0
2.5
3.0
3.5
4.0
High School GPA


SAT: Mathematics Score Distribution


SAT: Verbal/Literacy Score Distribution

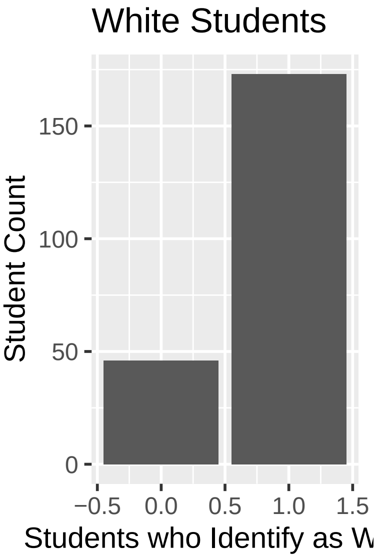


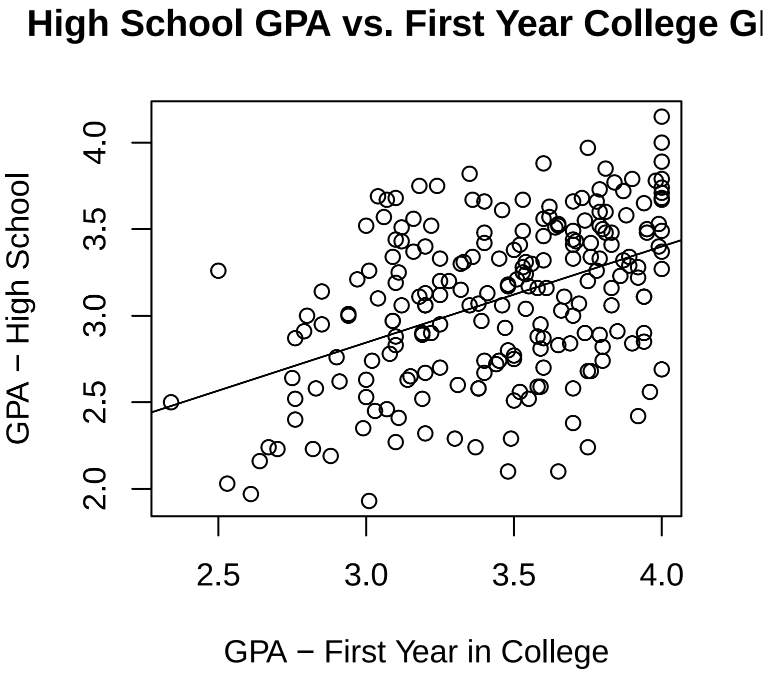
Credit Hour Distribution for High School Humanities Courses



Credit Hour Distribution for High School Social Science Courses

The following provides appropriate visual representation of the distributions of categorical variables identified in the dataset. As previously stated, these categorical variables consist strictly of boolean values, with 0 indicating falsity of a certain identity/status of a student, and 1 indicating that a student identifies with a given condition.

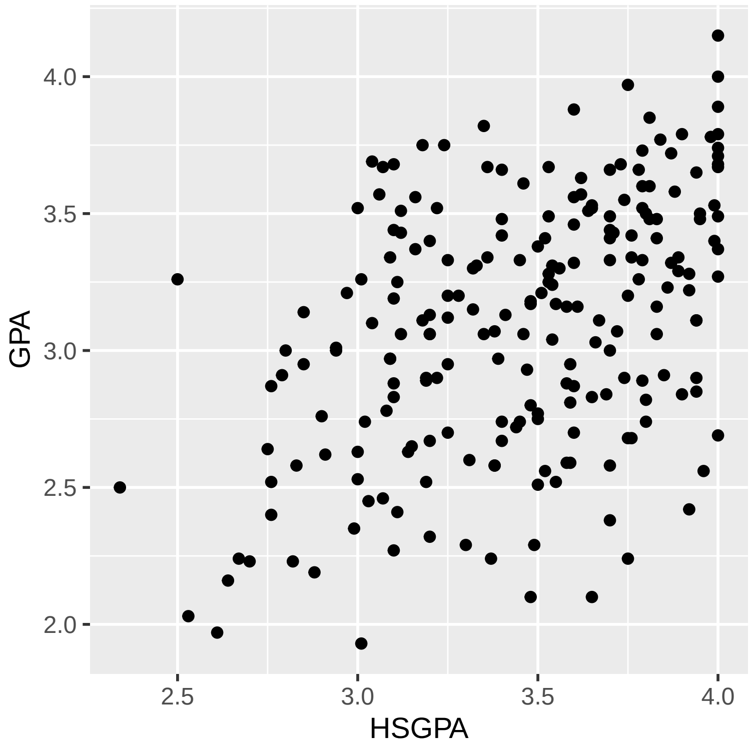


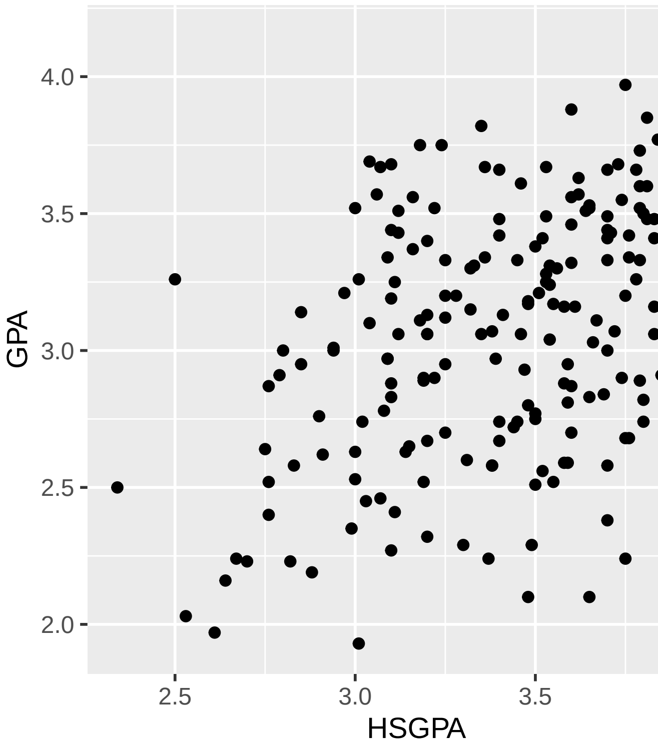


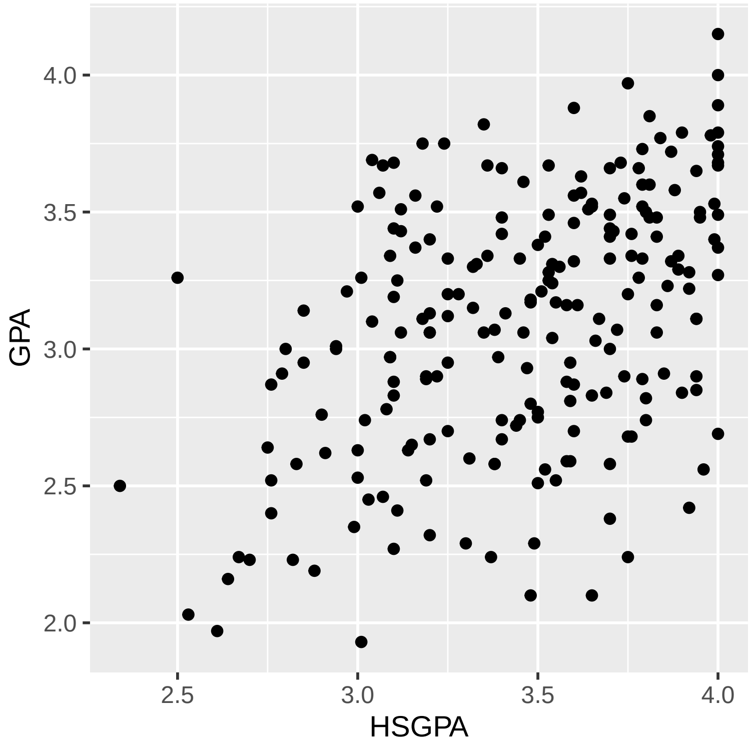
```
##
## Call:
## lm(formula = FirstYearGPA$GPA ~ FirstYearGPA$HSGPA)
##
## Residuals:
```

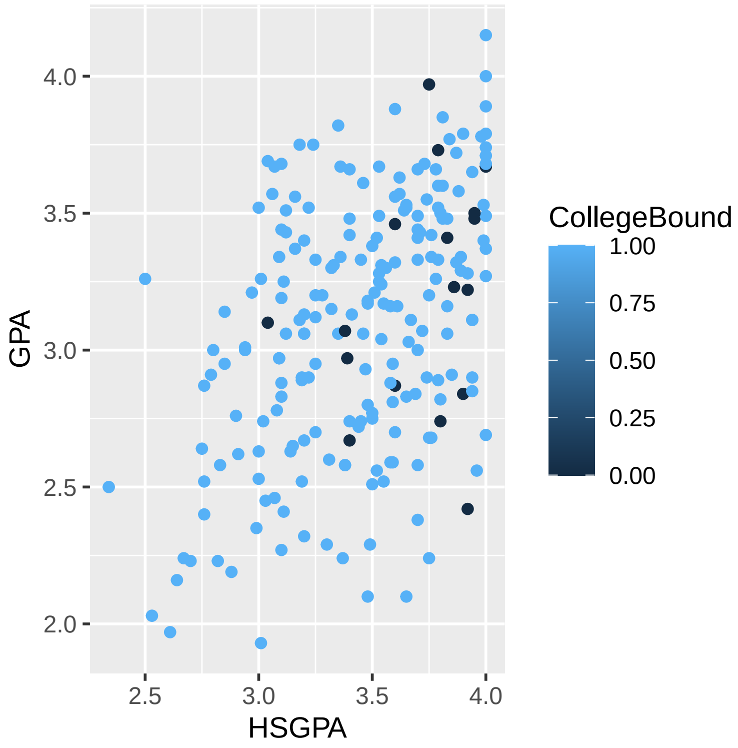
```
##      Min       1Q   Median       3Q      Max
## -1.10565 -0.31329  0.05871  0.29485  0.82291
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.17985    0.26194   4.504 1.09e-05 ***
## FirstYearGPA$HSGPA  0.55501    0.07542   7.359 3.78e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4174 on 217 degrees of freedom
## Multiple R-squared:  0.1997, Adjusted R-squared:  0.196
## F-statistic: 54.15 on 1 and 217 DF,  p-value: 3.783e-12
```

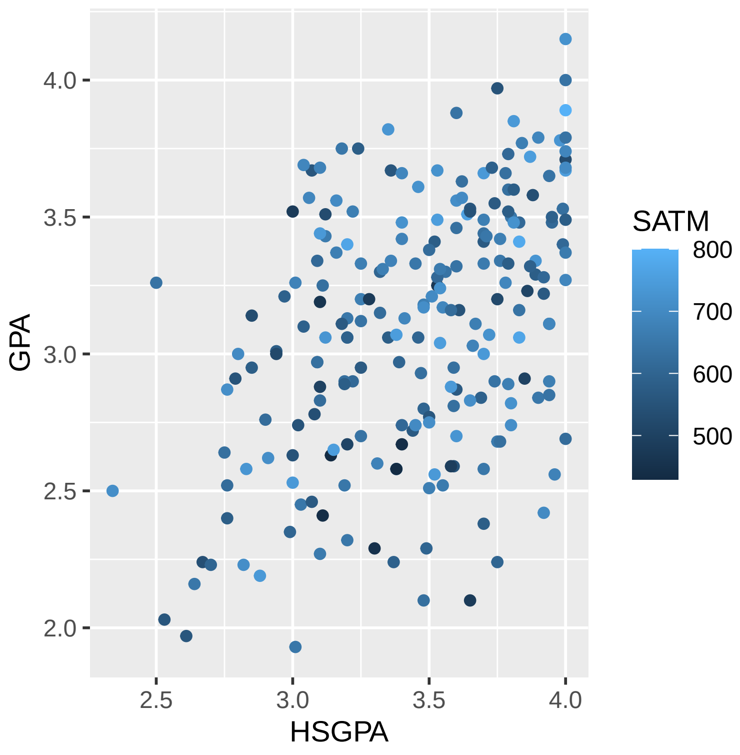
```
##        (Intercept) FirstYearGPA$HSGPA
##          1.1798507          0.5550125
```

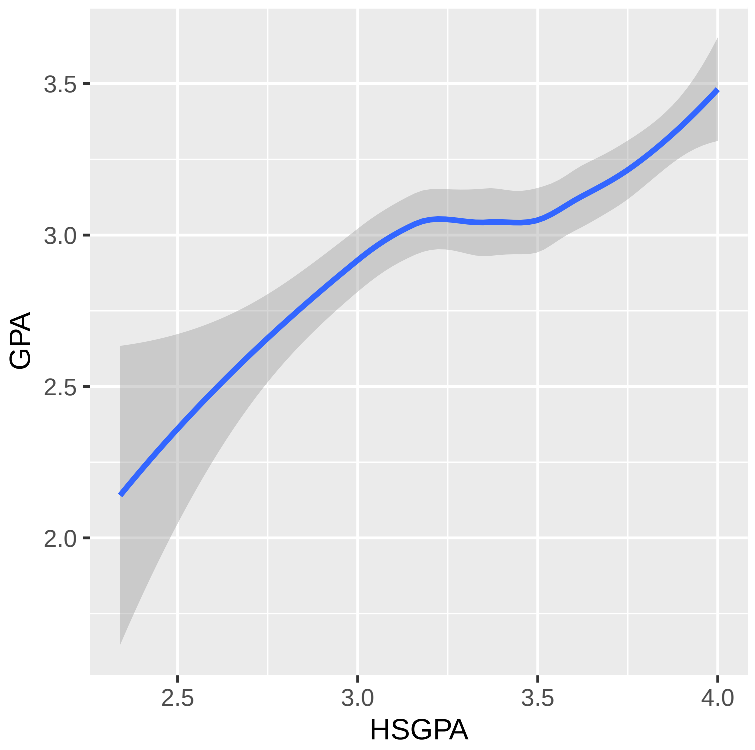


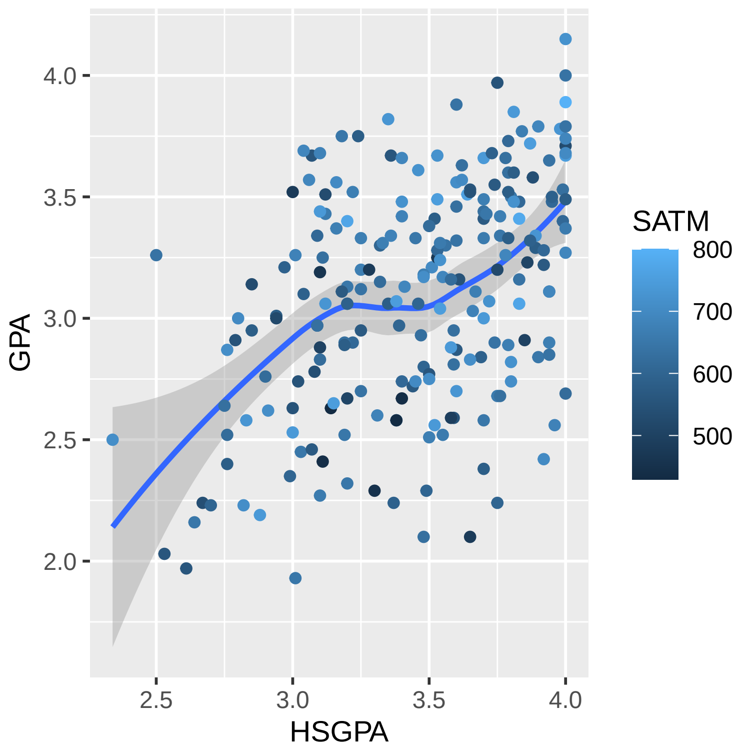


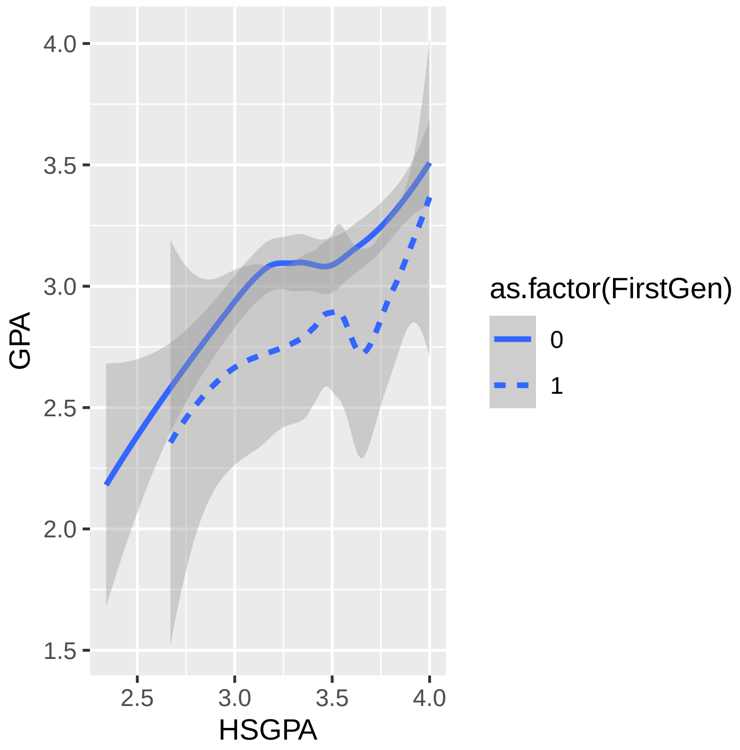

GPA
4.0
3.5
3.0
2.5
2.0
2.5
3.0
3.5
4.0
HSGPA

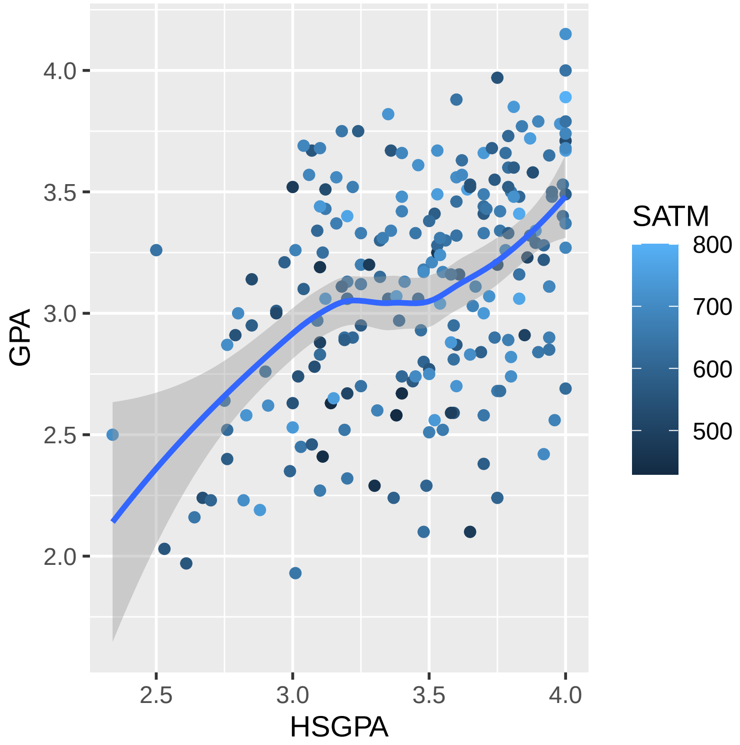

GPA
4.0
3.5
3.0
2.5
2.0
2.5
3.0
3.5
4.0
HSGPA
CollegeBound
1.00
0.75
0.50
0.25
0.00


4.0
3.5
3.0
2.5
2.0
GPA
2.5
3.0
3.5
4.0
HSGPA
SATM
800
700
600
500

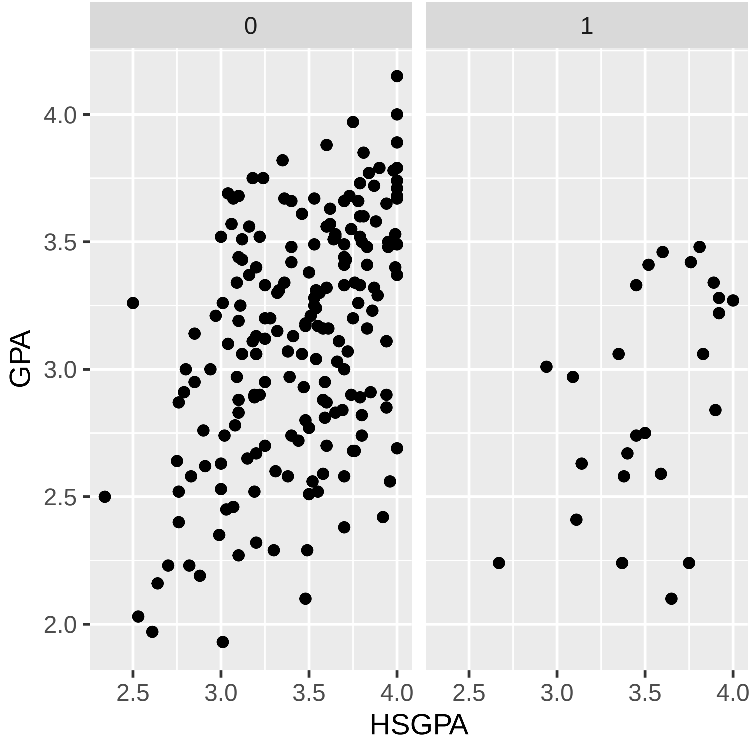

3.5
3.0
2.5
2.0
GPA
2.5
3.0
3.5
4.0
HSGPA


GPA
4.0
3.5
3.0
2.5
2.0
2.5
3.0
3.5
4.0
HSGPA
SATM
800
700
600
500

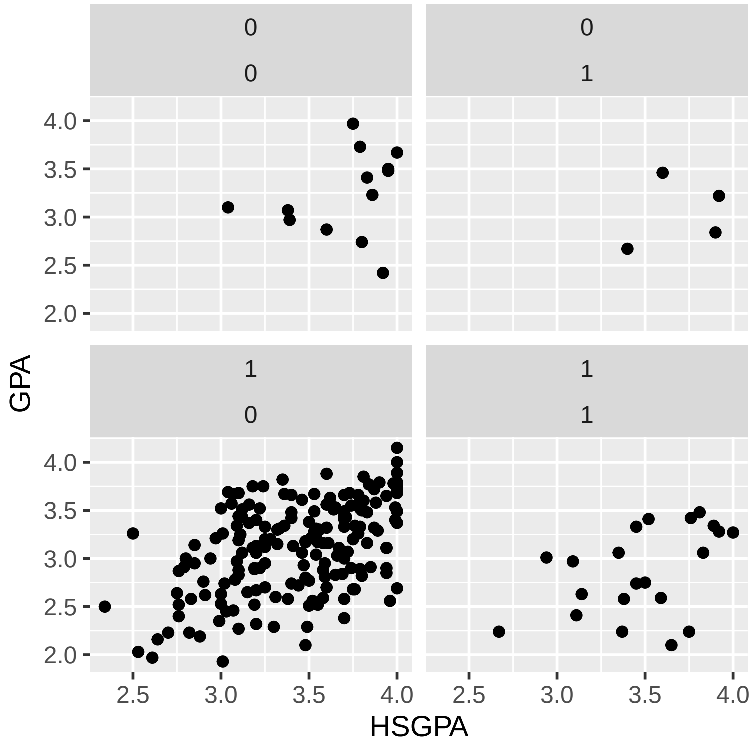

GPA
4.0
3.5
3.0
2.5
2.0
1.5
2.5
3.0
3.5
4.0
HSGPA
as.factor(FirstGen)
0
1

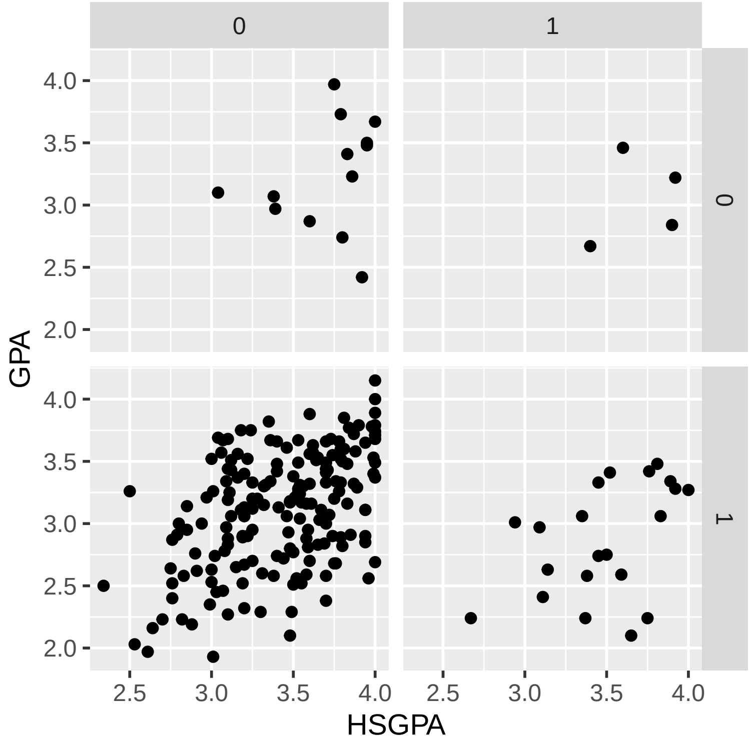


4.0
3.5
3.0
2.5
2.0
GPA
2.5
3.0
3.5
4.0
HSGPA
SATM
800
700
600
500

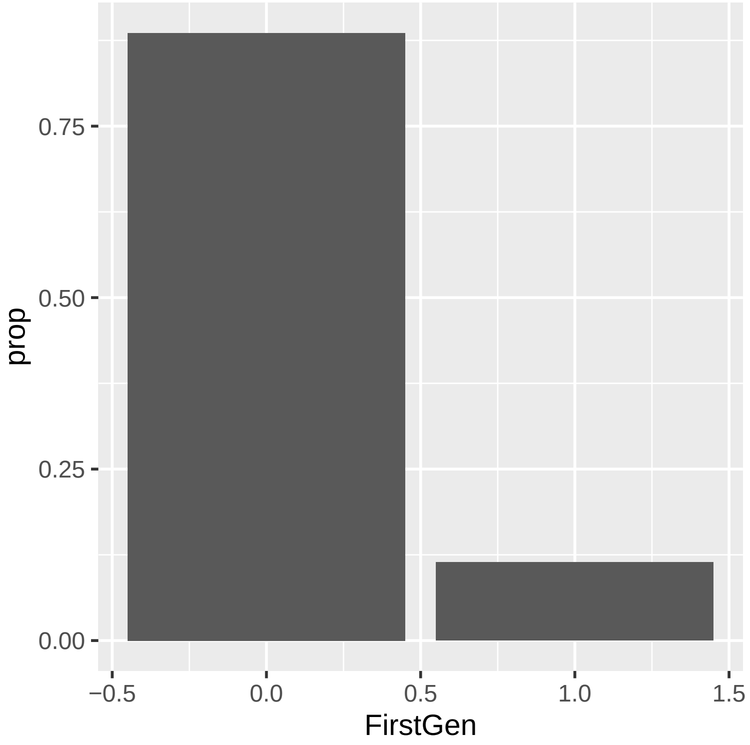


0
1
4.0
3.5
3.0
2.5
2.0
GPA
2.5
3.0
3.5
4.0
2.5
3.0
3.5
4.0
HSGPA





```
##
##  Welch Two Sample t-test
##
## data:  nonwhite$GPA and iswhite$GPA
## t = -3.8836, df = 62.441, p-value = 0.0002511
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4865808 -0.1559198
## sample estimates:
## mean of x mean of y
##  2.842391  3.163642
```



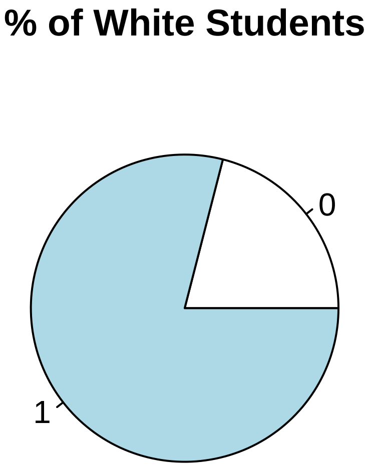
```
##
##       0   1
##   0  34 160
##   1  12  13
```

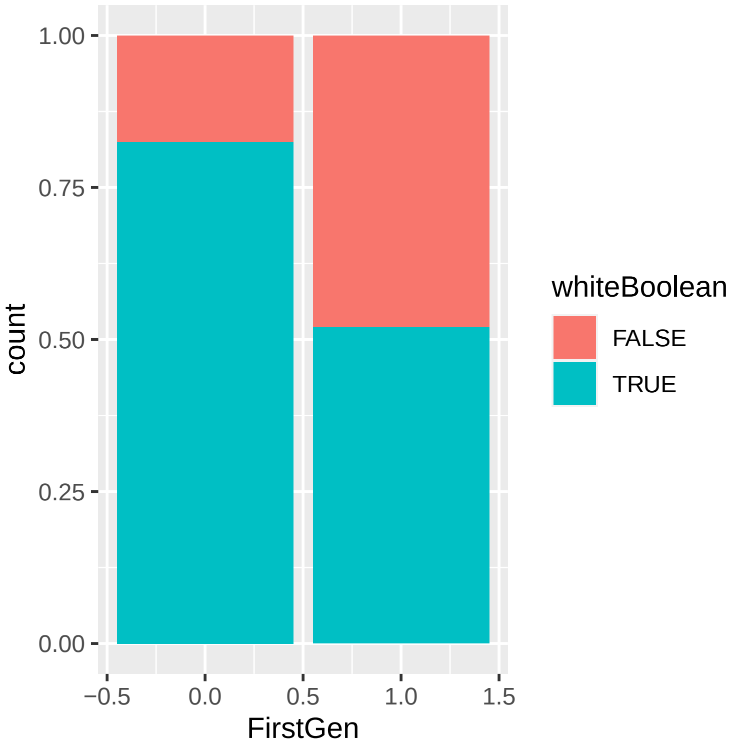
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  newtable
## X-squared = 10.626, df = 1, p-value = 0.001115
```

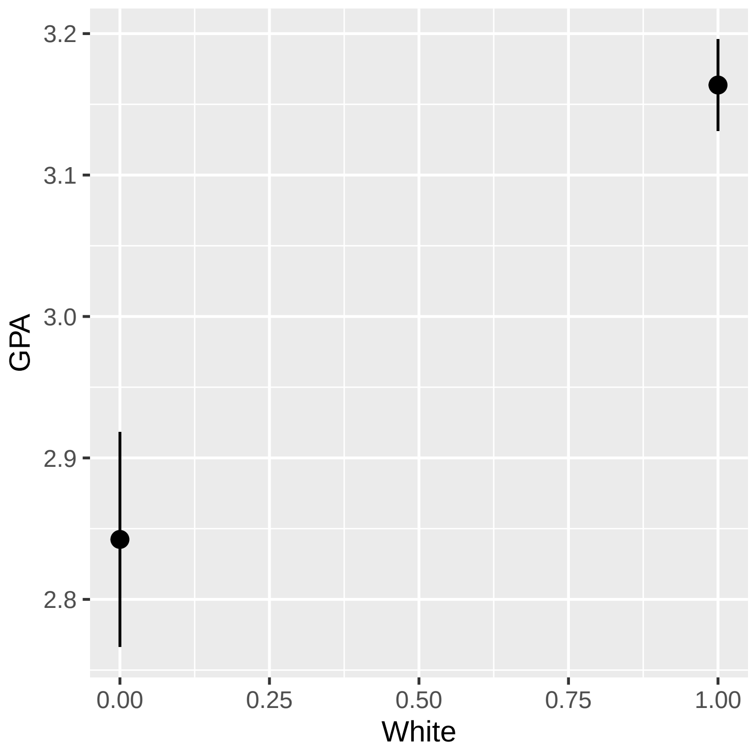
```
##
##       0   1
##   0  34 160
##   1  12  13
```

```
##
##             0         1
##   0 40.748858 153.25114
##   1  5.251142  19.74886
```

```
##
##          0      1
##   0 40.75 153.25
##   1  5.25  19.75
```






3.2
3.1
3.0
2.9
2.8
GPA
0.00
0.25
0.50
0.75
1.00
White