# REPORT FINDINGS

1. **SAT and GPA scores with respect to student success (Graduation Rates -6 Year Rate): Exploratory Analysis:**
   1. Upon first glance we see that public high school students have higher GPAs but lower SAT scores, and graduate at lower rates in comparison to their private school counterparts ([SAT Graph](#SAT))
   2. Also, female students seem to score higher on the SAT overall, and have higher entering GPAs but seem to graduate at about the same rate as their male counterparts ([SAT graph](#SAT), [GPA graph](#GPA))
   3. Lastly, female private school students seem to have the best chance for success according to our box plots, but most notably, those with high SAT scores seem to graduate at higher rates ([SAT graph](#SAT), [GPA graph](#GPA))
2. **Feature Selection in R: Finding the most important variables**

Before we began our analysis, we made sure to clean the data in various ways;

* NAs/blanks were handled one of two ways:
  + For the SAT scores the NA/blank values were interpolated using the average of the column provided, which allowed us to keep the same dimension
  + For categorical items the NAs were largely omitted when passing it into the model
* IDs were dropped due to their non-causal relationship to student success
* Year admitted was dropped due to the fact that these were all the same and had no important impact on our model
* All other variables of categorical nature were recoded with a numerical placeholder to allow this information to be passed into the model
  1. Feature selection using Random Forest in R
     1. An effective and widely used feature selection technique, we used the Random Forest (RF) algorithm to use random decision trees to pull out the most important features of the dataset
        + 2 separate R libraries were chosen to ensure a robust RF algorithm was used
        + A visual and tabular output are provided ([variable ranks](#feature_selection))
     2. From this feature selection method we found that the three most important variables by order of importance are:

Math was taken through all 4 years of high school

The number of transfer units earned by the student from another institution

High school GPA of student

* 1. Feature selection using XGBoost in R
     1. Another widely used feature selection algorithm, we used XGBoost for its computational speed as well as its gradient boosting qualities, and as a check for our findings from our RF algorithm
        + in this case we used the XGboost tool from the caret library in R which gave a highly readable ranking of the most important features of our model ([XGBoost output](#xgb_feature))
     2. From this feature selection method, we found that the three most important variables by order of importance are:

Math was taken through all 4 years of high school

The number of transfer units earned by the student from another institution

High school GPA of student

From these features we can use a logistic regression to give insight to our research question: What factors should be observed when attempting to predict student success (graduating within 6 years)?

* Three models were investigated in R (using a binary logistic regression) with graduated/non-grad rates as the outcome variable. All of the models showed some degree of statistical significance, but upon review of the residual plots, we found that the data may not be completely random where it needs to be so the following results should be treated with caution:
  + Model 1 used whether Math was taken through all 4 years of High School
    - Finding: There was high significance within this model, .
  + Model 2 used Transferred Units and whether a student took math all 4 years
    - Finding: There was high significance, and a high Chi-squared value, so we may reject the null hypothesis

.

* + Model 3 used Transferred Units, whether the student took Math all 4 years of high school, and their high school GPA into account
    - Finding: This model showed that these variables were statistically significant with the exception of GPA, and with a large chi-squared value, we can reject the null for all but high school GPA, which was not a significant feature selected out of our RF algorithm

,

1. **Challenges to Analysis**
   1. Compute Time
      1. With larger datasets we begin to run into runtime issues, forcing us to use more powerful computation methods such as deploying cloud technology to outsource our own hardware limitations
   2. Multicollinearity
      1. When data is increased by several orders of magnitude we may find some variables are actually closely related, some may even allow for perfect or near-perfect correlations, which could skew data and disallow a reasonable model to be produced; if we wish to just make predictions then this *may* be overlooked but for completeness we would use a variance inflation factors function to measure VIF levels, if they’re above 5 we could standardize the variables by subtracting the mean to better center the variables and reduce multicollinearity.
      2. LASSO (or LASSO regularization) and ridge regression could also make for alternatives as well
   3. Normality of Data
      1. Larger data also runs the risk of being non-normal, with skewed distributions, exponentials, log, long-tail, etc. and with more distributions we must find the correct tools to either normalize the data itself, or find an applicable model for the distribution in question.

Figure 1: SAT score and graduation

Figure 2: HS GPA and graduation

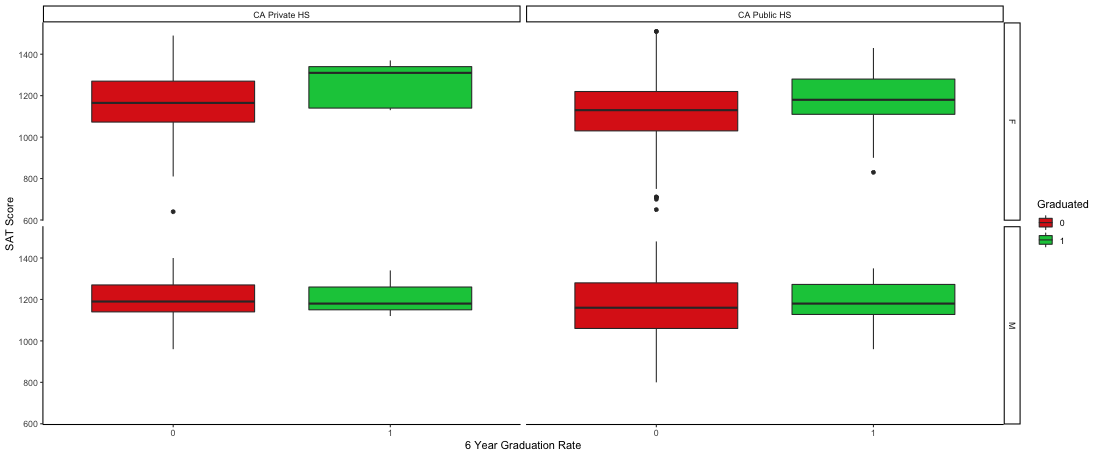
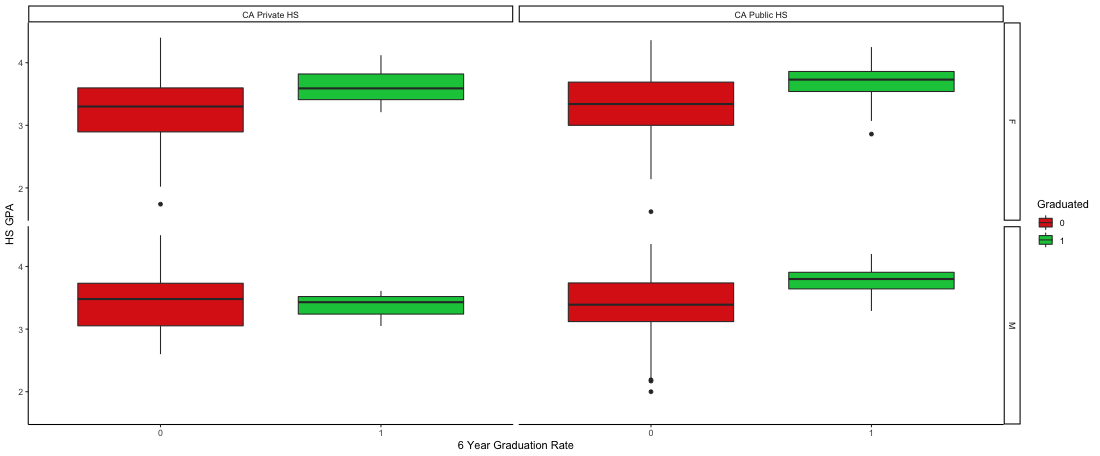
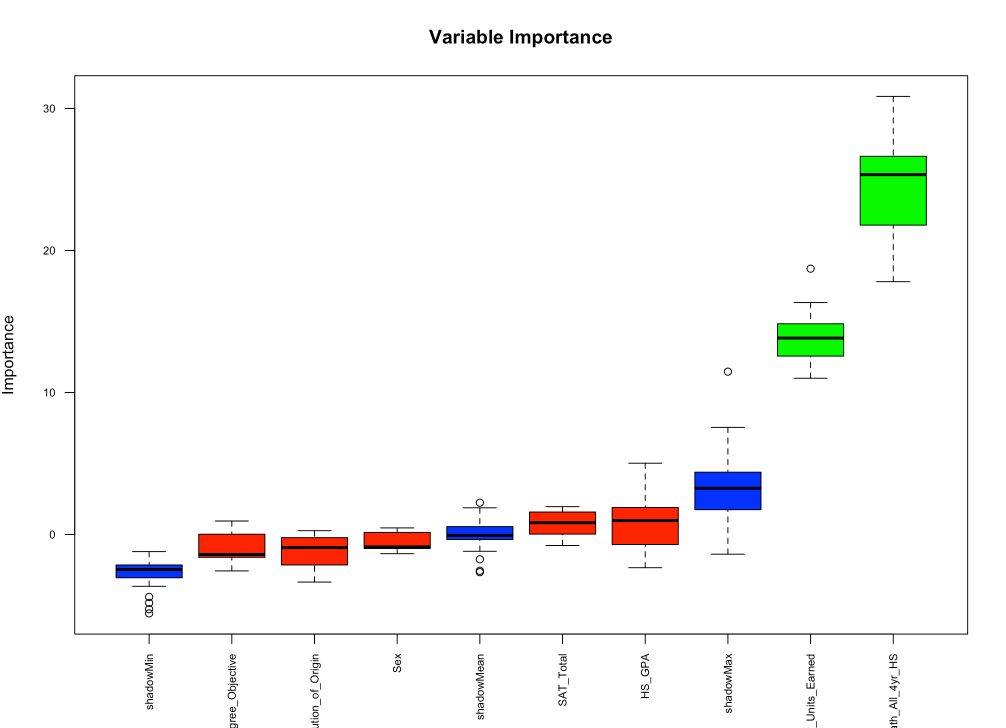


Figure 3: Graph of Random Forest Feature Rankings



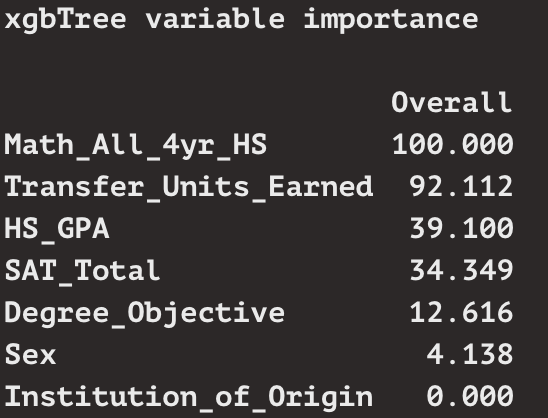


Figure 4: R output of XGBoost Feature Rankings