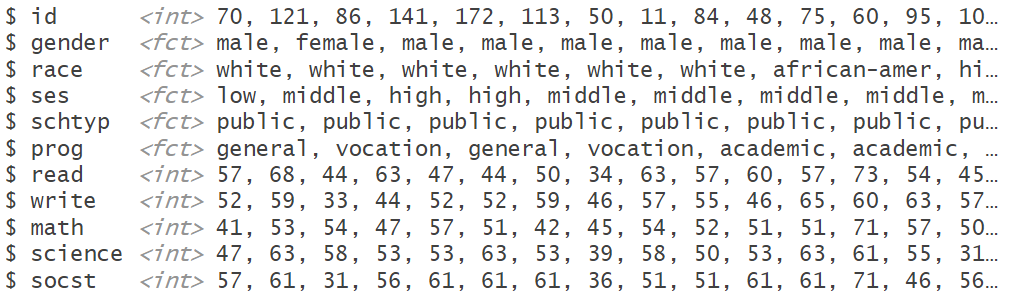
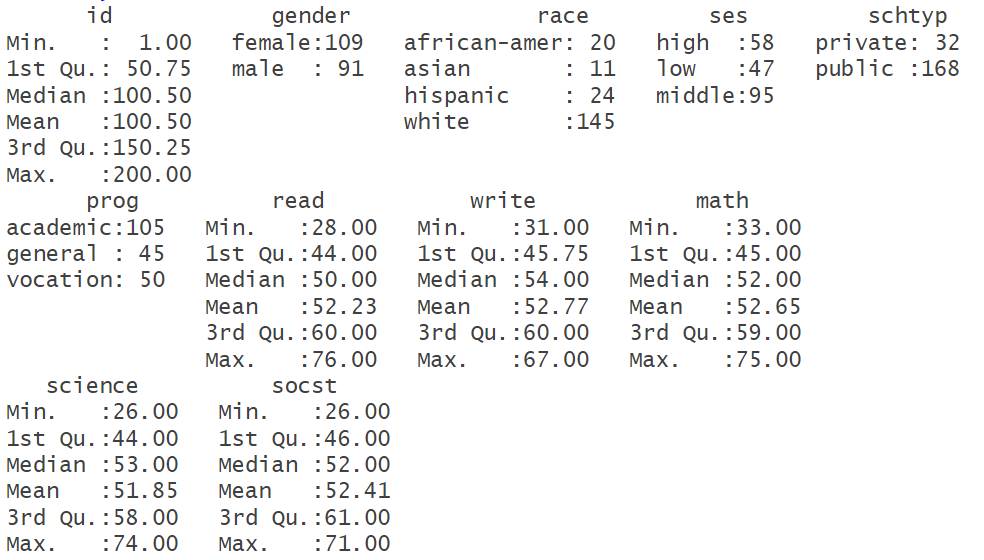
**Predicting Student Program Choices Using Random Forest Classification**

**OBJECTIVE**  
This project aims to predict a student's chosen academic program (general, academic, or vocational) using variables selected through a feature selection process. Using the hsb dataset from the faraway package, we explore patterns in the data through visualizations and summary statistics. A Random Forest classification model is then trained using 10-fold cross-validation to assess model accuracy and optimize the mtry hyperparameter. The final model focuses on five subject score variables, selected based on their importance during the feature selection.

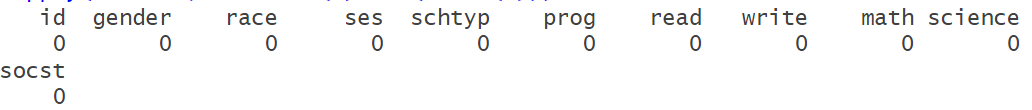
**OVERVIEW OF THE DATA**



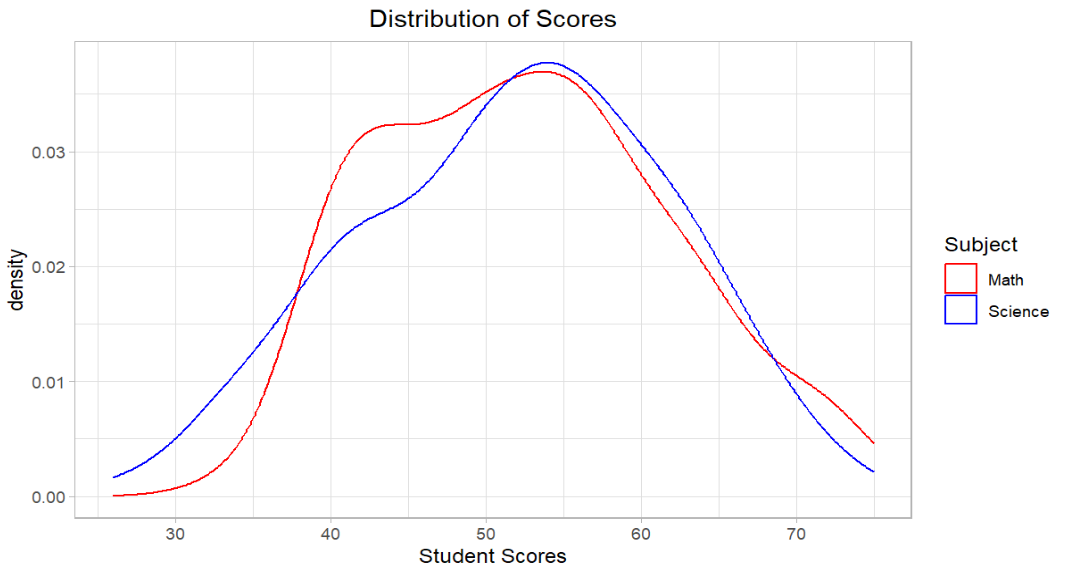
**SUMMARY STATISTICS**

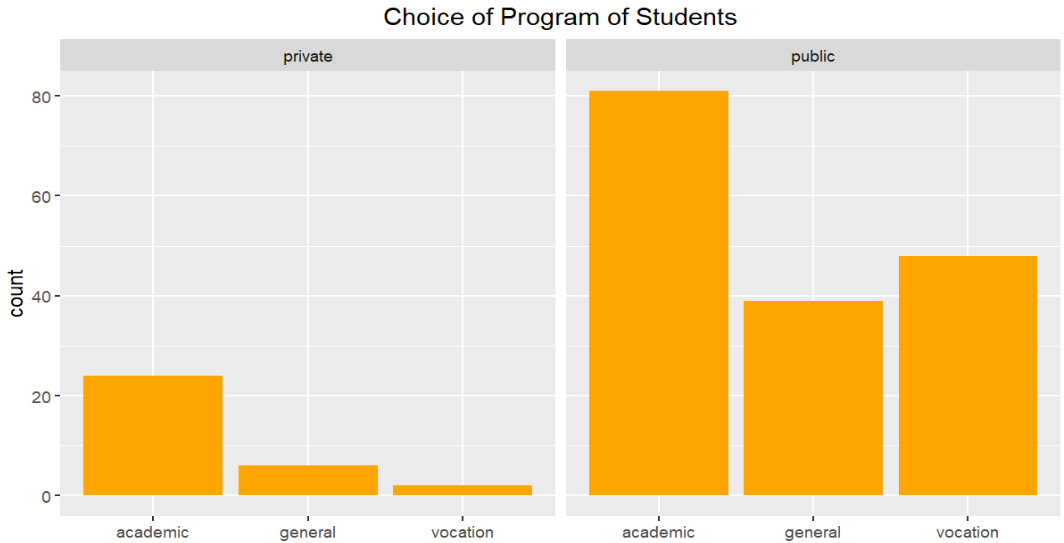


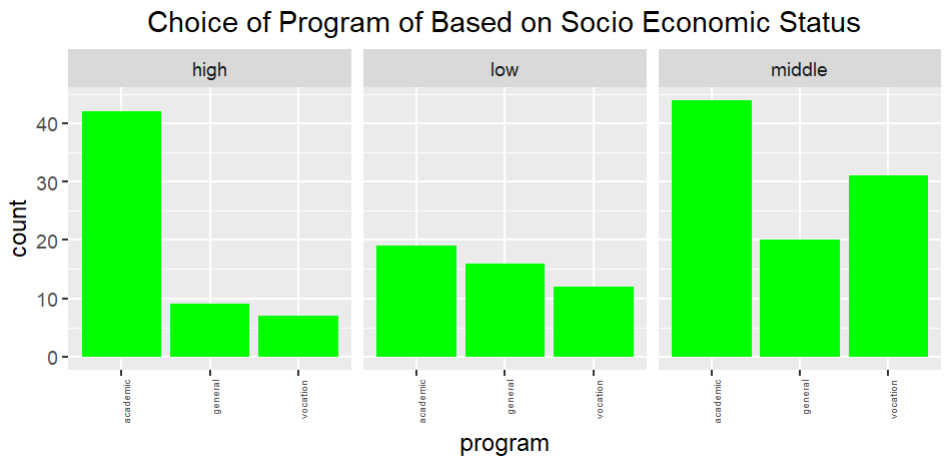
**CHECKING OF NAs**

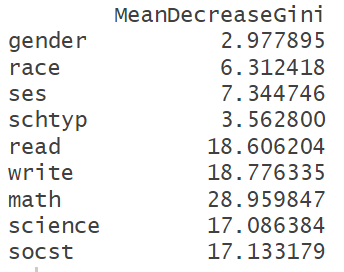


**DATA VISUALIZATION**

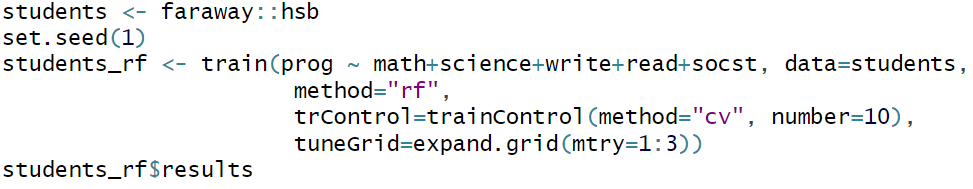
****

****

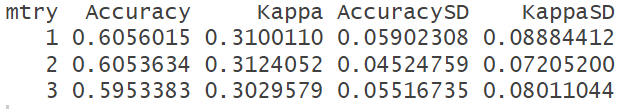
****

**FEATURE SELECTION BASED ON VARIABLE IMPORTANCE  
**

**PREDICTIVE MODEL USING RANDOM FOREST WITH CROSS-VALIDATION**

****

**MODEL PERFORMANCE**

****

**SUMMARY OF RESULT**

We performed exploratory data analysis (EDA) on the *hsb* dataset from *faraway package* to understand the distribution of student program choices across socioeconomic status (low, middle, high) and school types (public or private). Visualizations highlighted key program trends and the distributions of subject scores for math and science using ggplot.

A Random Forest model was built using five subject-related features (math, science, write, read, and socst) based on the feature selection process. Using 10-fold cross-validation, the optimal model was selected with the best-performing mtry value (mtry = 1). The final model demonstrates a moderate ability to predict a student's program type (60% accuracy), reflecting the relationship between academic performance and educational program choices.