**Introduction to statistical learning: Project Proposal**

**Student Performance Data Set**

**Team Members:**

**Sandeep Reddy Salkuti (16296868)**

**Sumanth Medavarapu(16295321)**

**PardhaSaradhi Ramineni(16300893)**

**SaiChand Patchala(16292087)**

**Poojasree Reddem(16296515)**

**Project Description:**

We are willing to find student final grades based on considering different attributes. This data approaches student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). The two datasets were modeled under binary/five-level classification and regression tasks. The target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful.

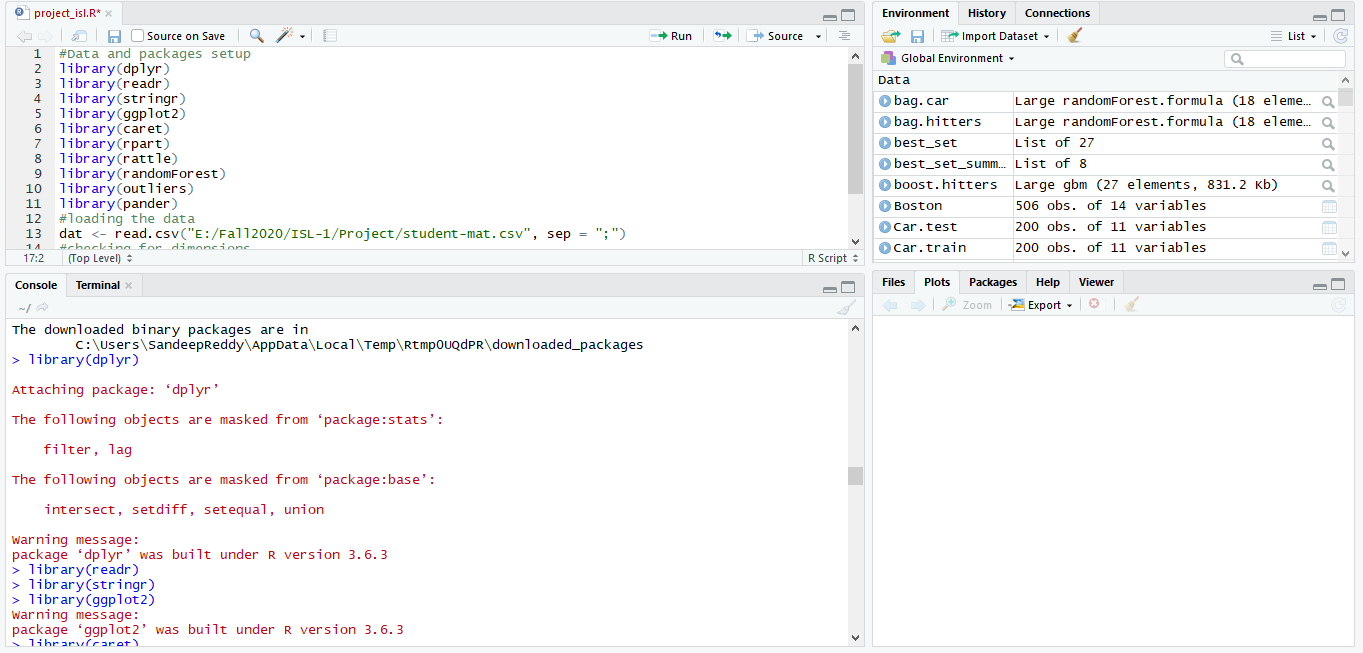
**Goal:**

The classification goal is to predict student performance in secondary education (high school)

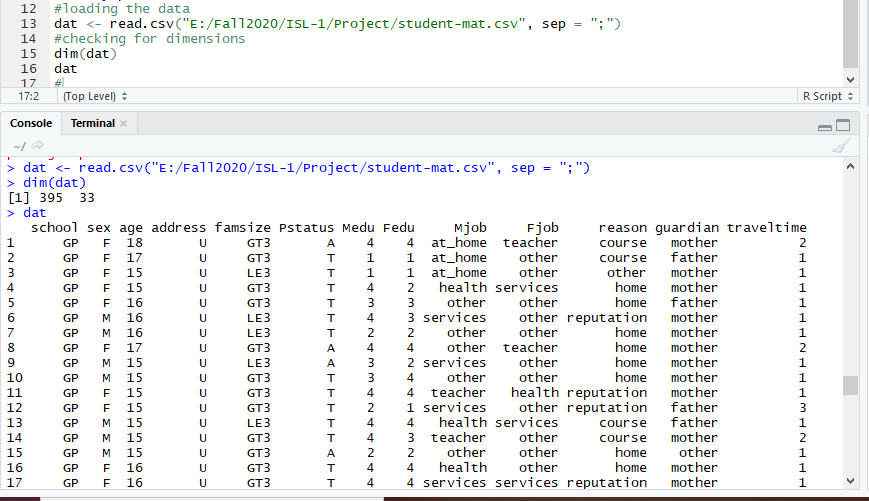
**Dataset:**

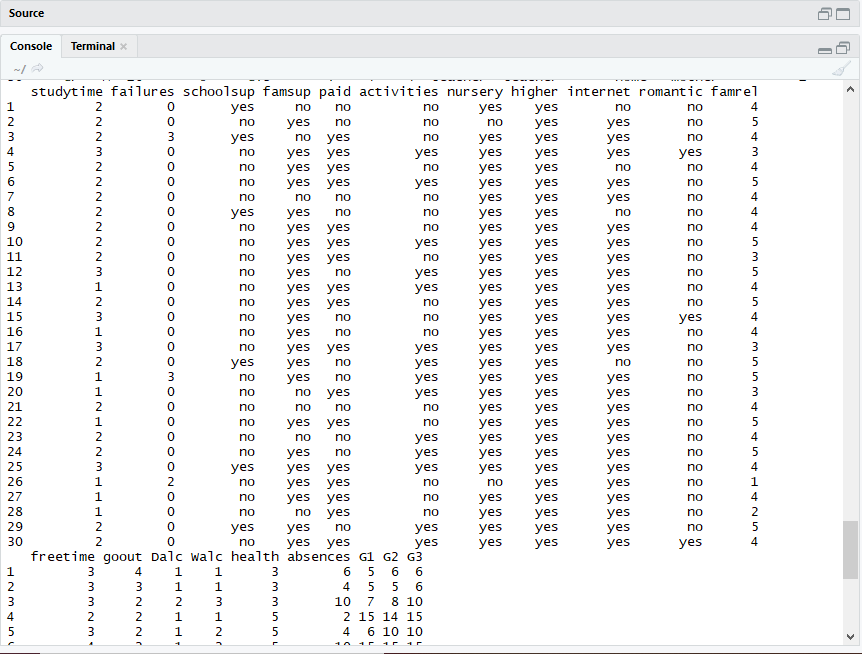
It is the Student Performance dataset uploaded originally in the UCI Machine Learning Repository. The dataset gives information about a student grades, demographic, social and school related features.

**Loading libraries:**

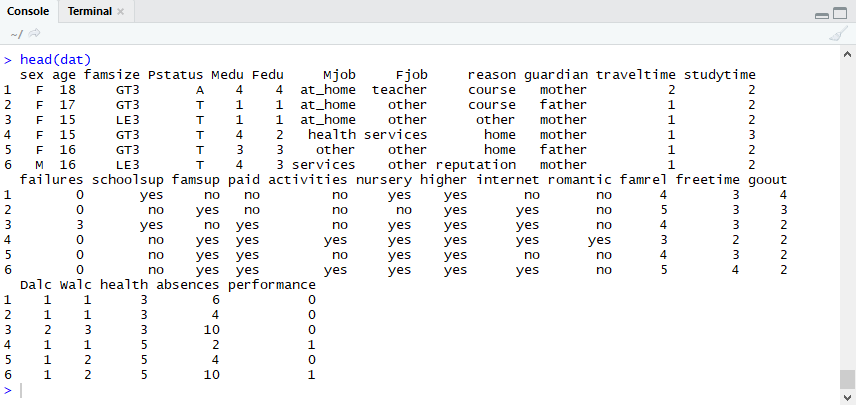


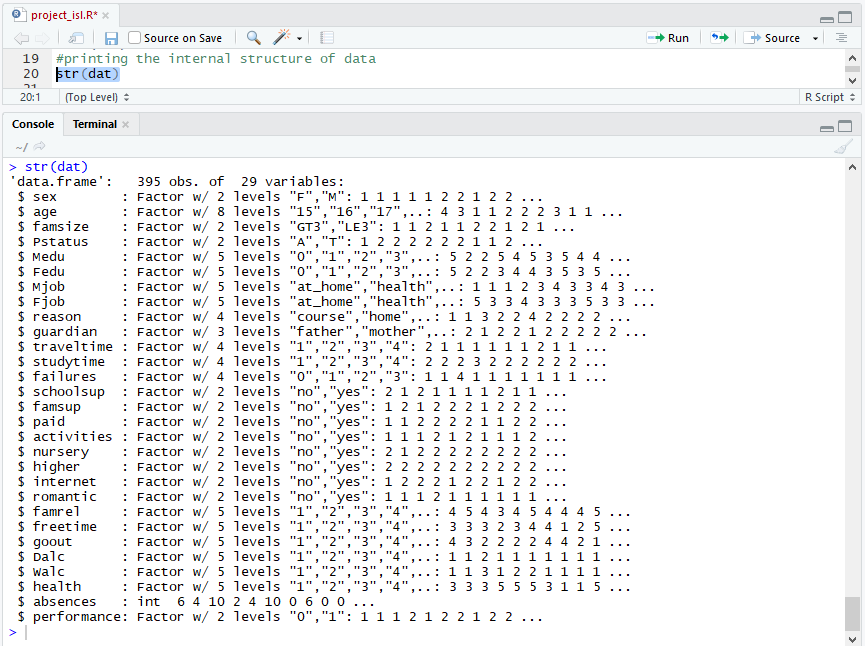
**Loading dataset and viewing the dimensions :**

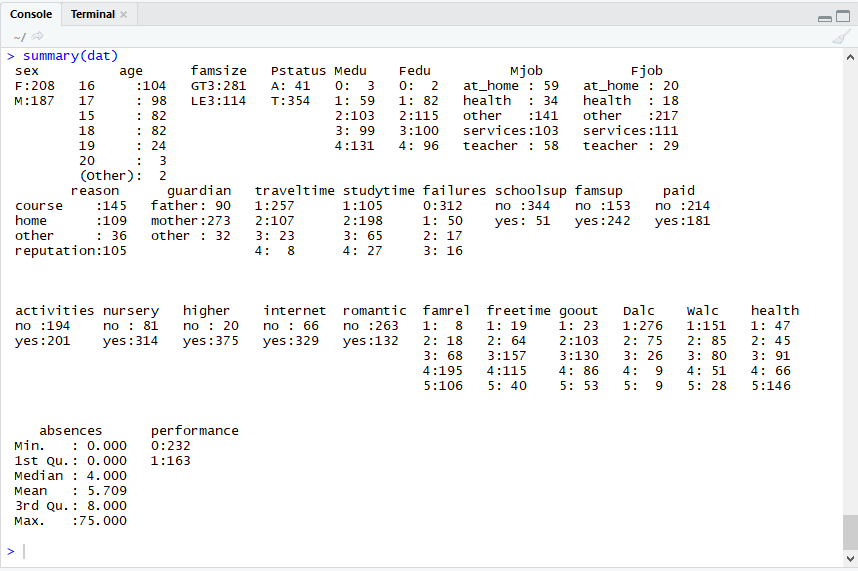




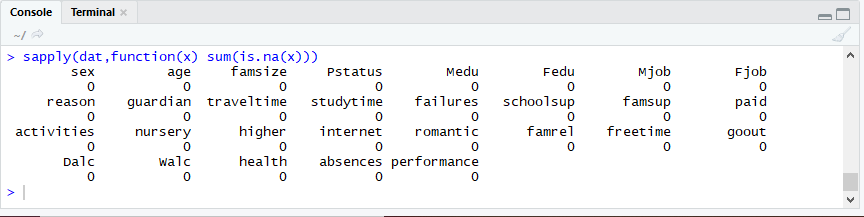
Printing data, internal structure and summary



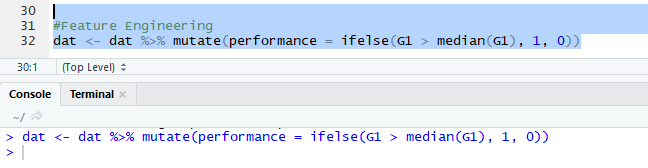




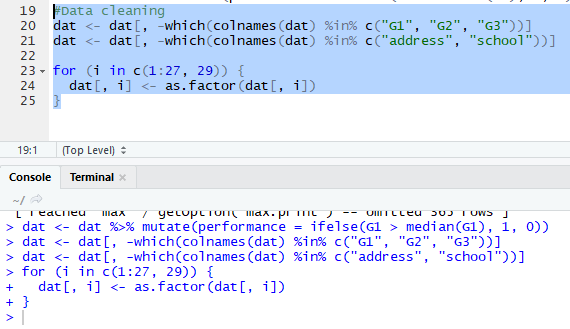
Checking for any missing values



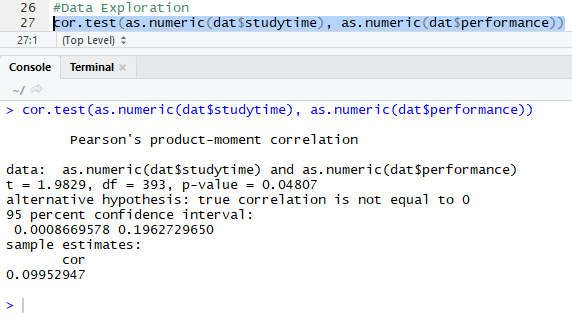
**Feature engineering:**



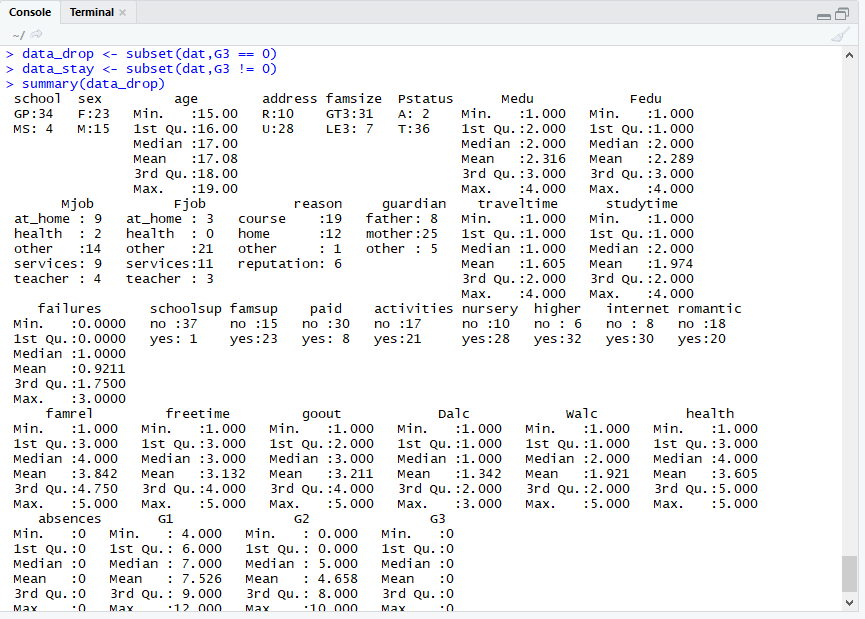
**Data Cleaning:**

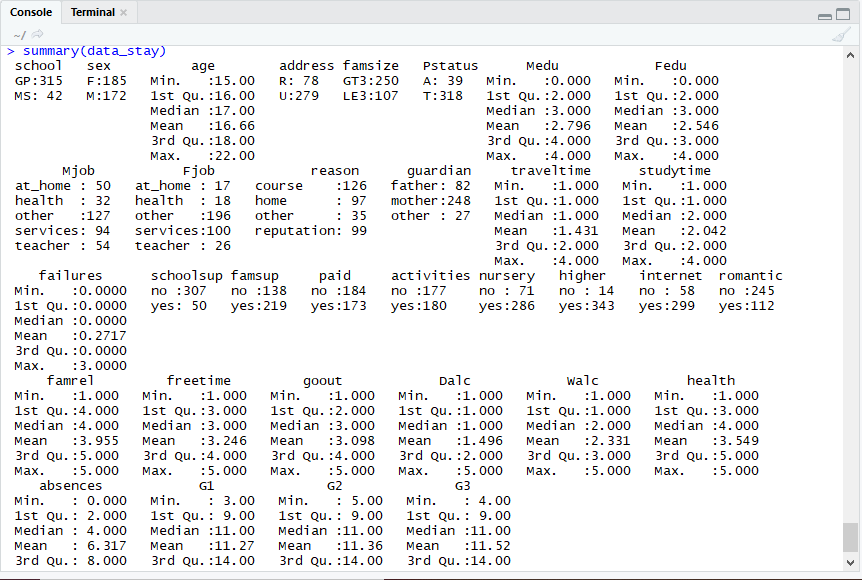


**Data Exploration:**

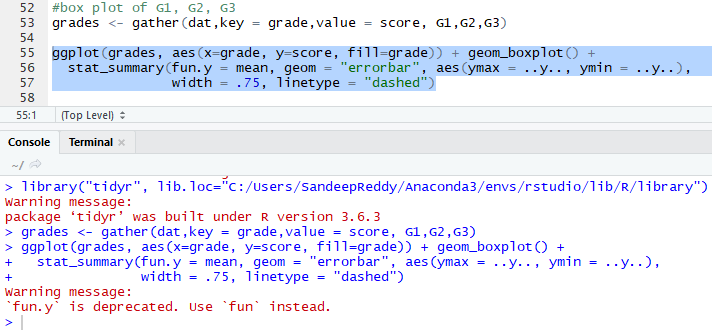


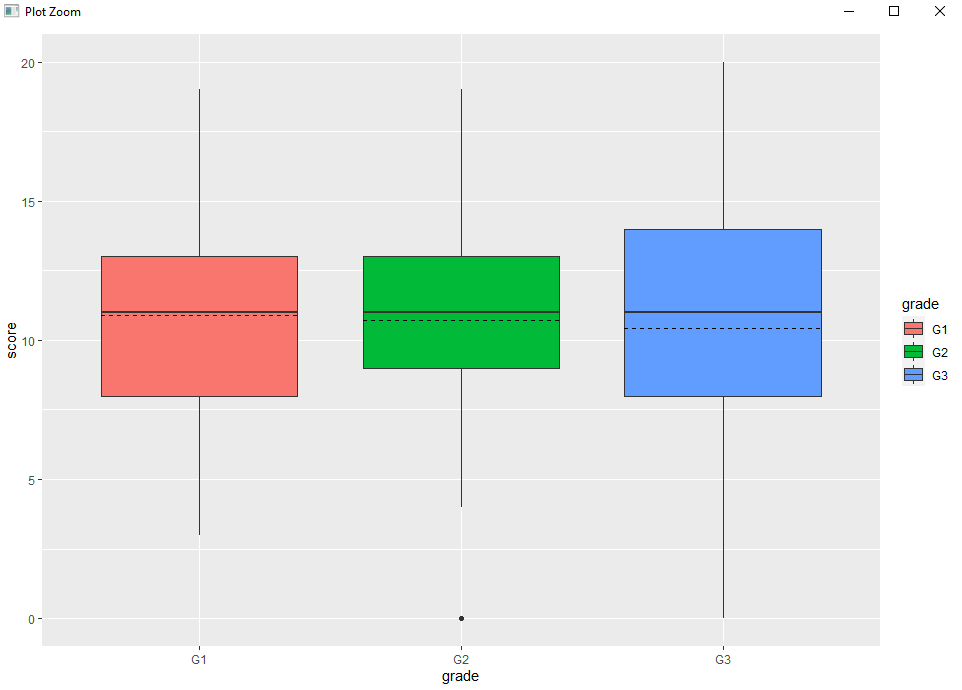
**Data Visualization:**





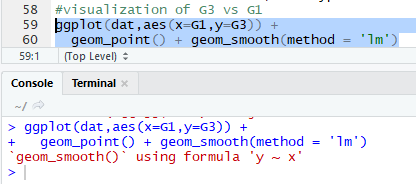
Box Plot of G1, G2, G3

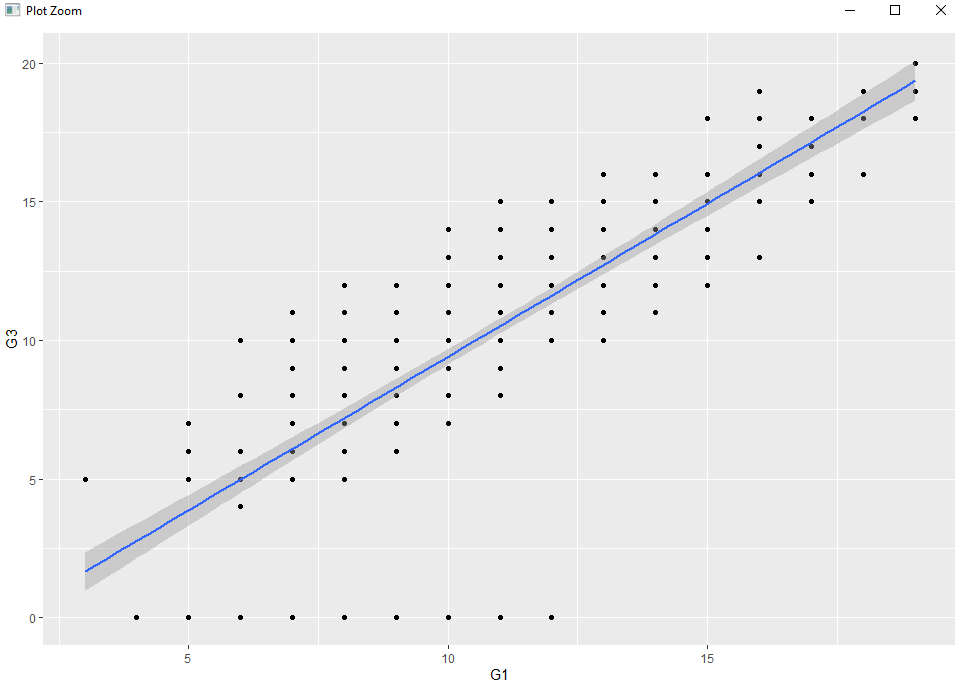




Visualization of relationship between G3 and other predictor variables

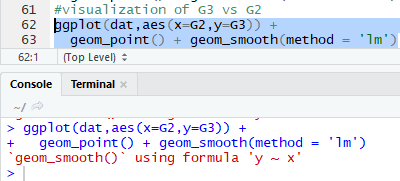
G3 versus G1

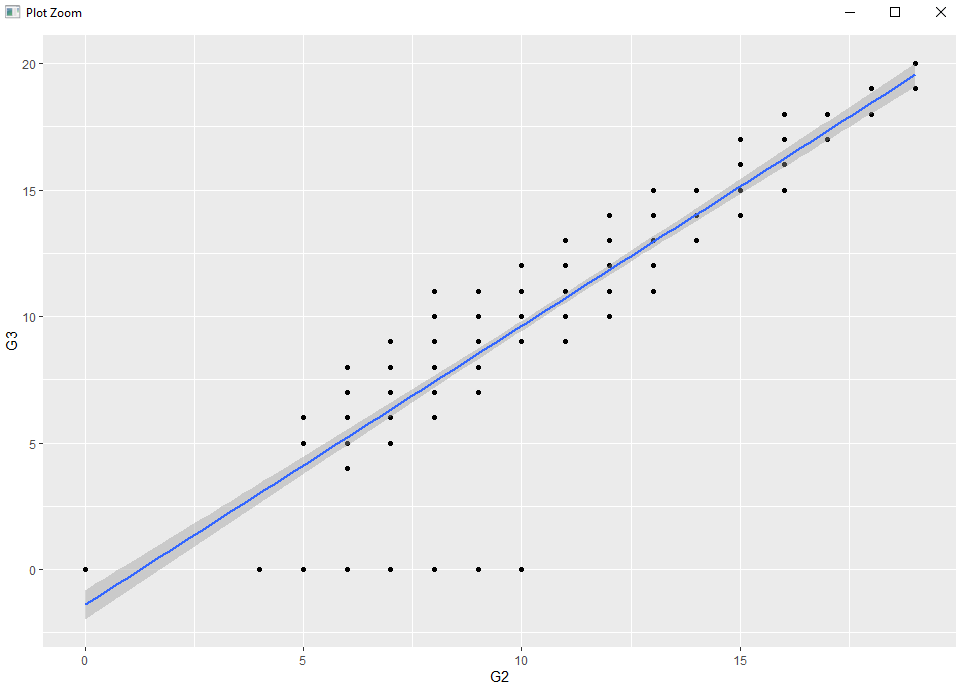




The graph above shows strong linear relationship.

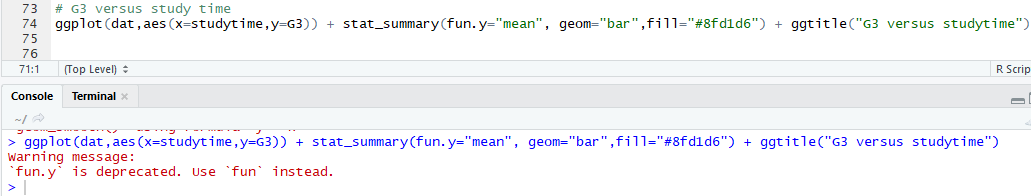
Visualization of G3 versus G2

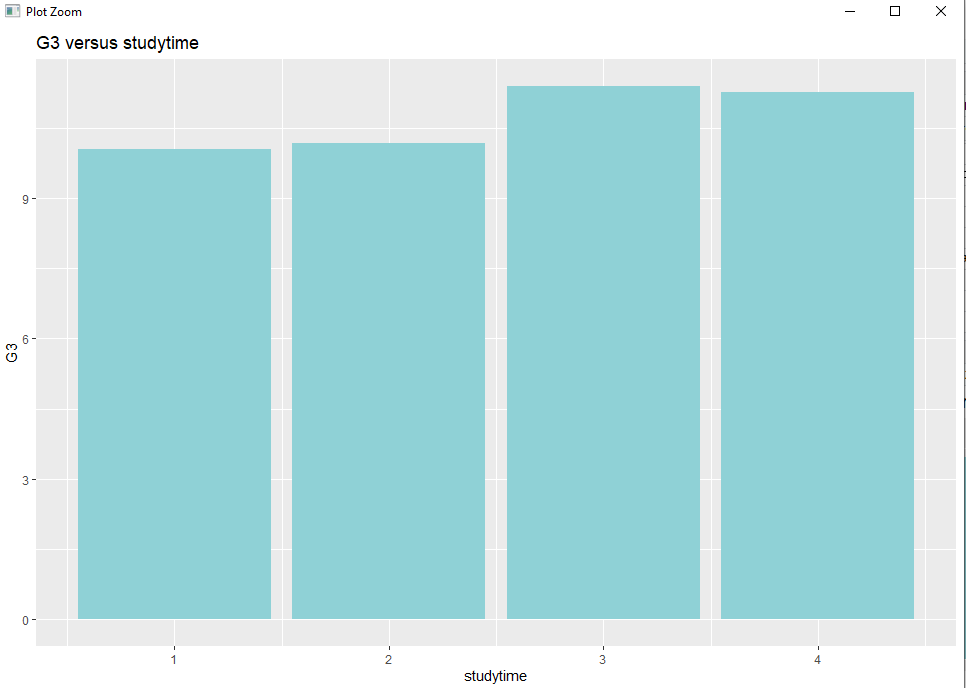




The above graph shows an even stronger relationship between G3 and G2

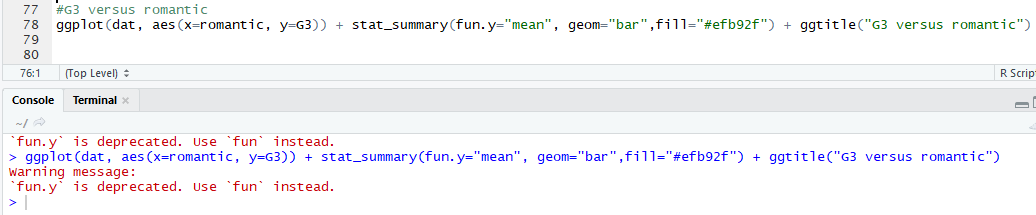
G3 versus study time

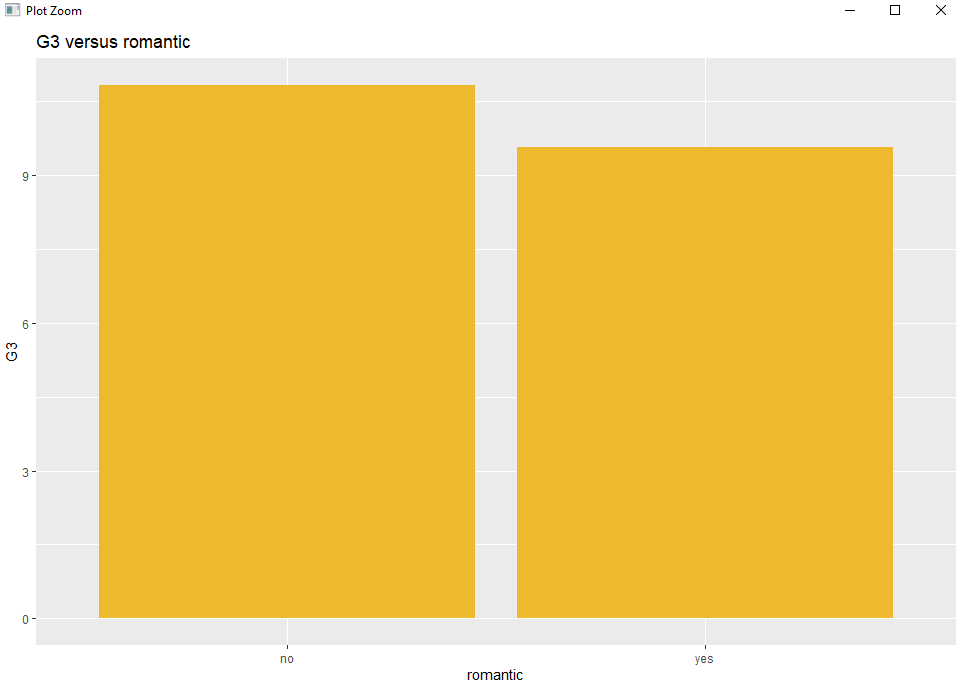




From above graph students who study for longer time get better final grades than students who study for shorter time.

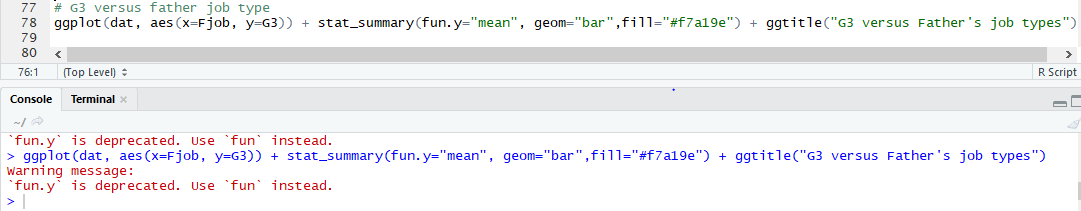
G3 versus romantic

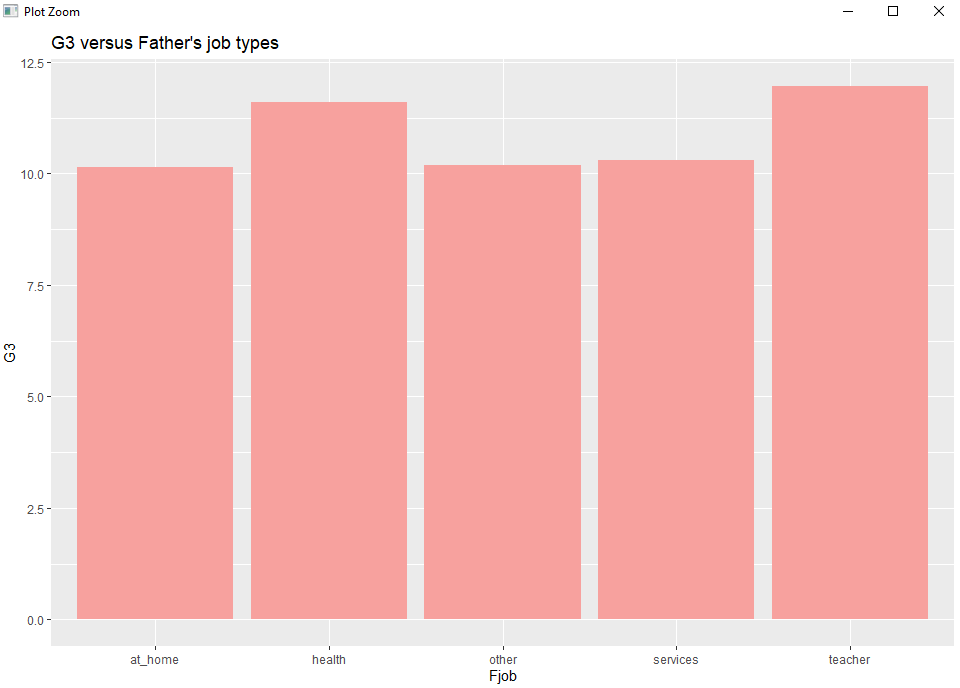




Single students have better scores

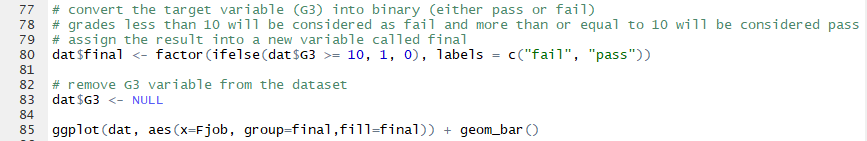
G3 versus father job types

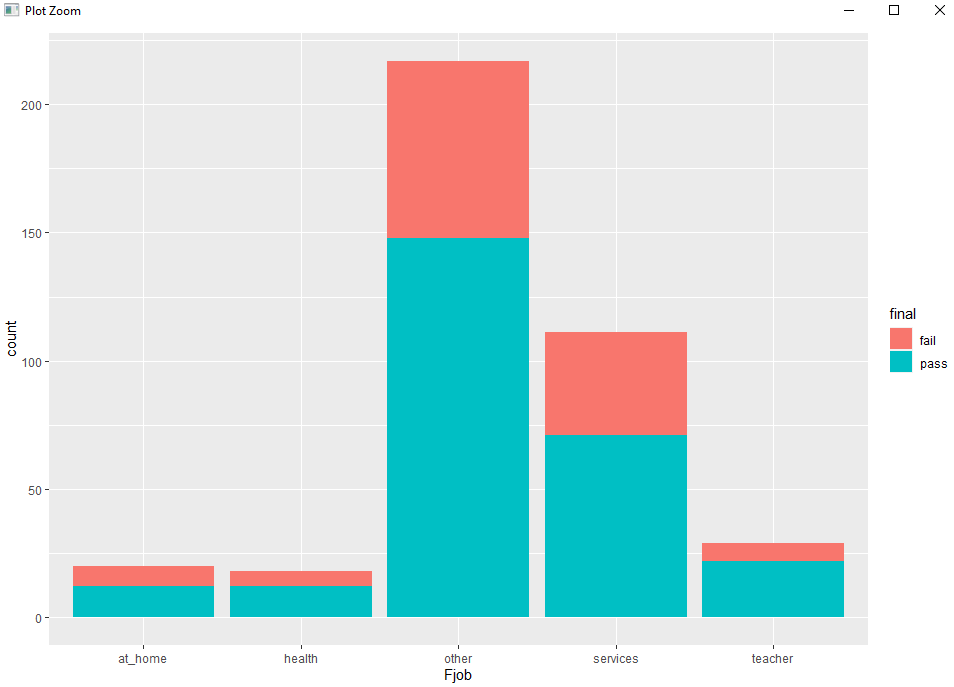




Students whose father works in education or healthcare field have slightly higher final grade

Bar chart showing count of pass and fail for fathers job types

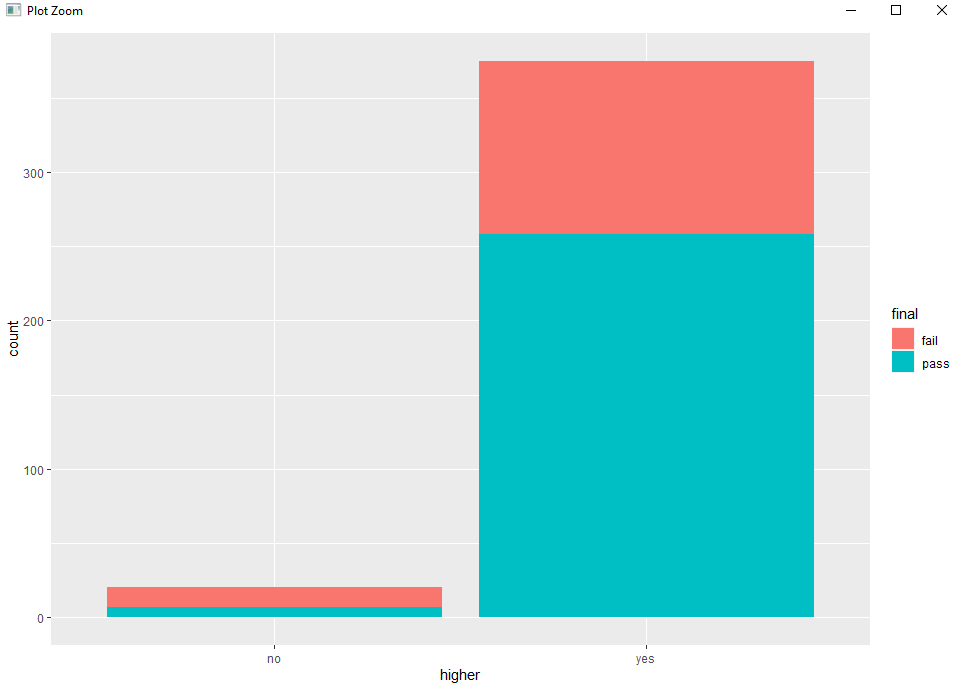




The above graph makes it easier to see the relationship between father’s occupation and the student’s performance

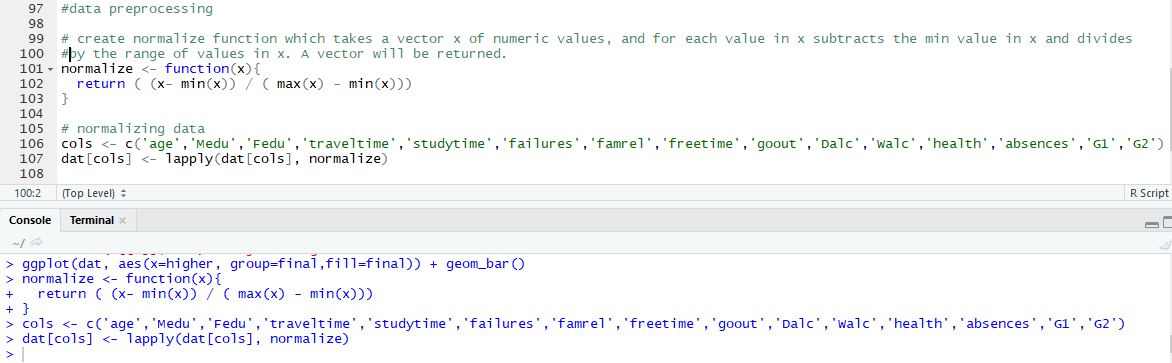
Student’s who want to pursue higher education



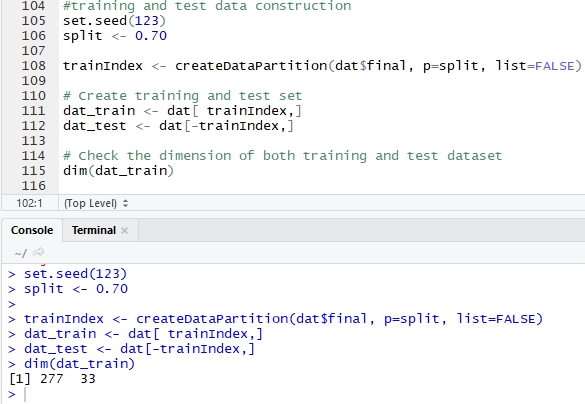


Students who do not plan to pursue higher education are more likely to fail and vice versa. Yes = 258 pass & 117 fail No = 7 pass & 13 fail

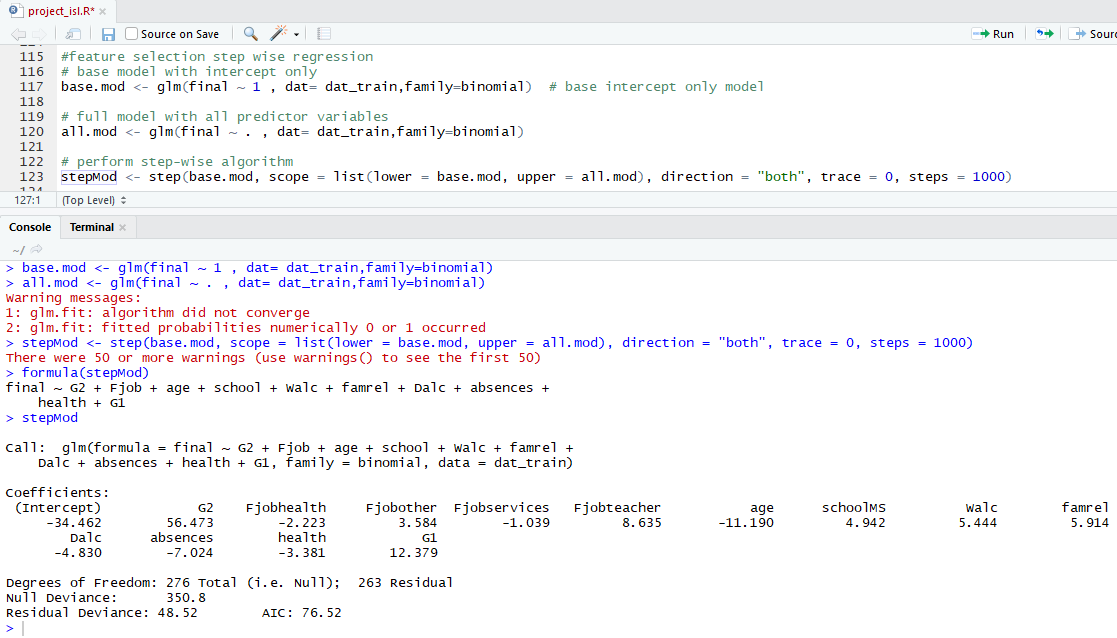
**Data preprocessing:**

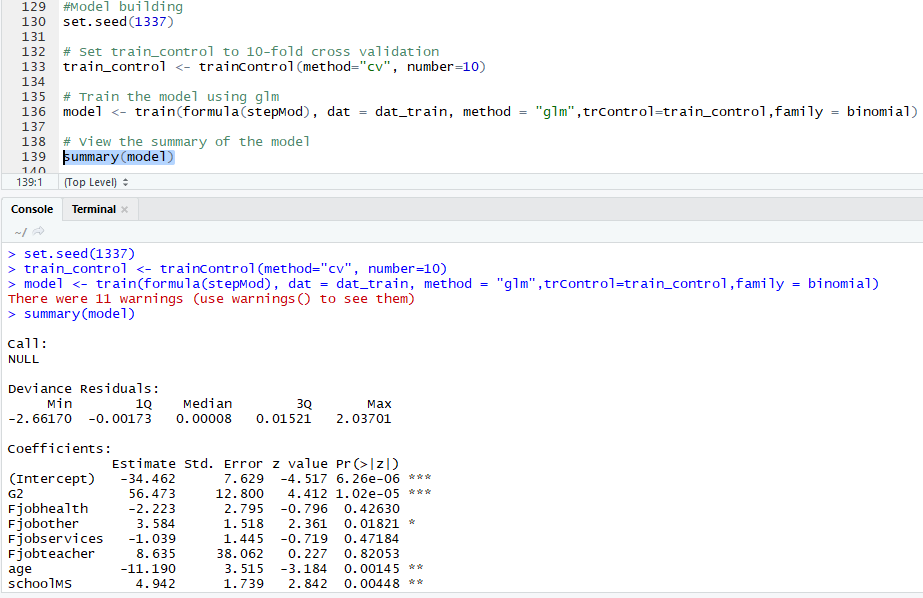


**Training and test data construction:**

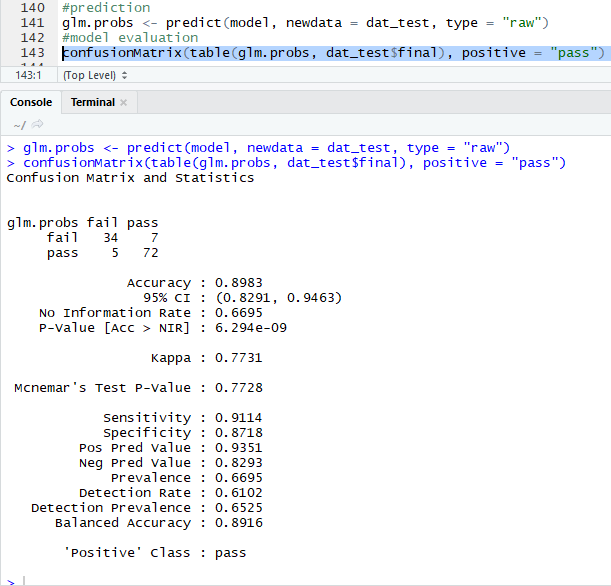






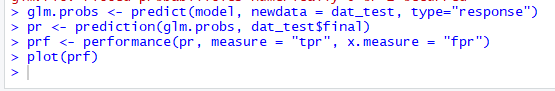


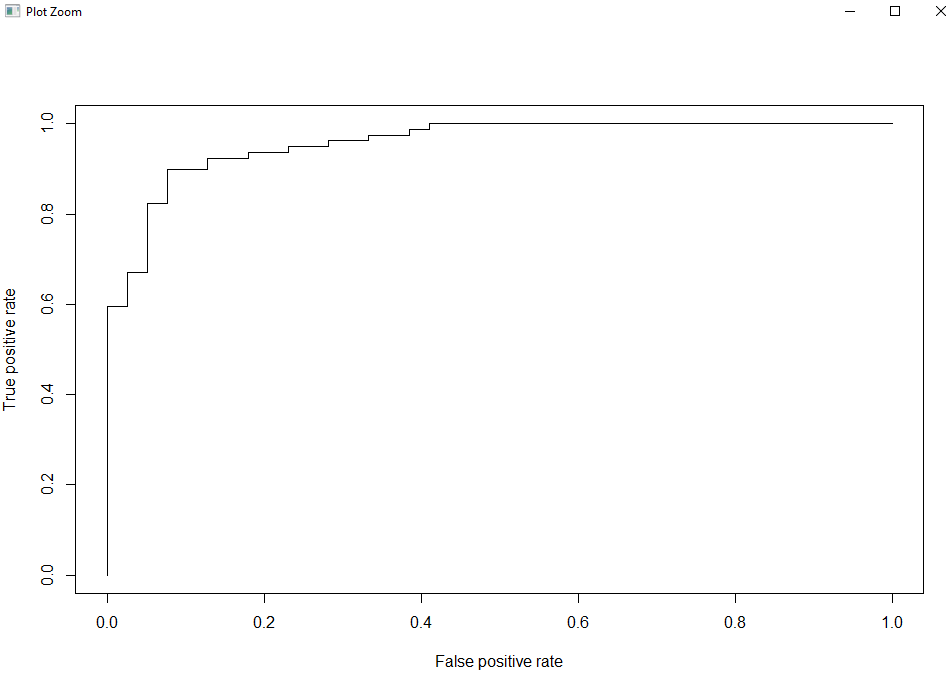
**Model evaluation:**

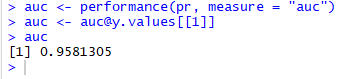


Accuracy of model is 89%

ROC curve:

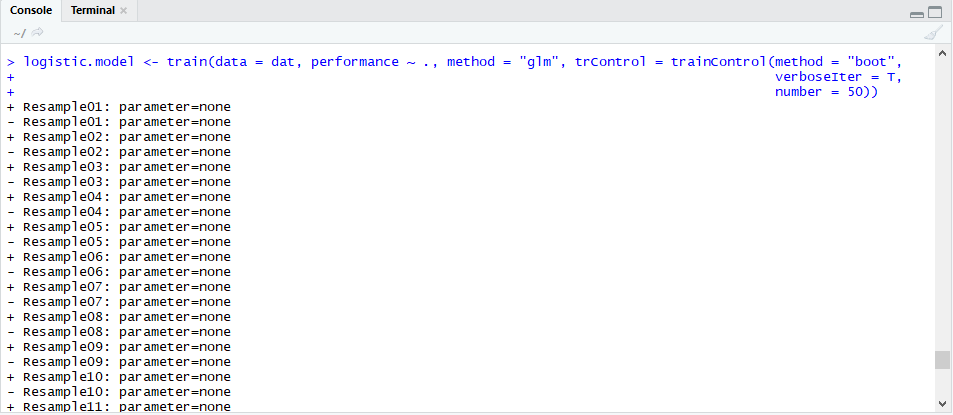






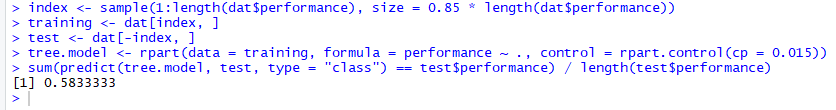
The auc value indicates that the model have 95% chance of accurately discriminate between positive and negative classes.

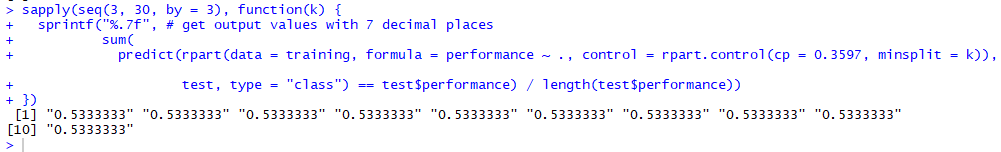
**Logistic Regression:**



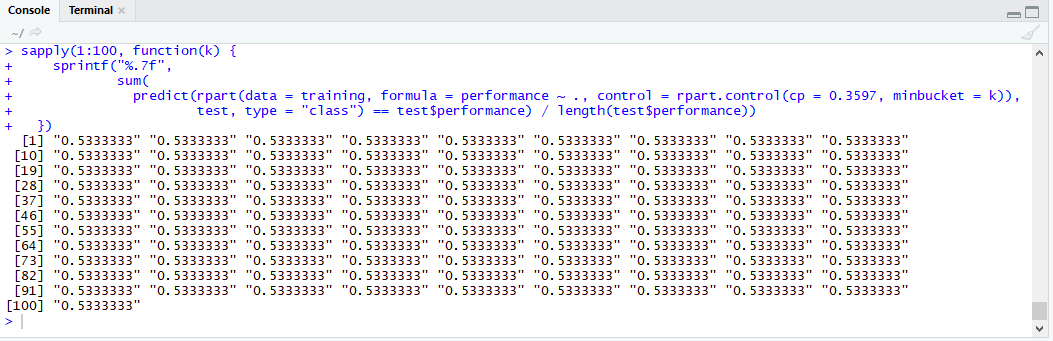
We get 61.7% accuracy rate

**Decision Tree:**





We somehow get upto 60% accuracy rate

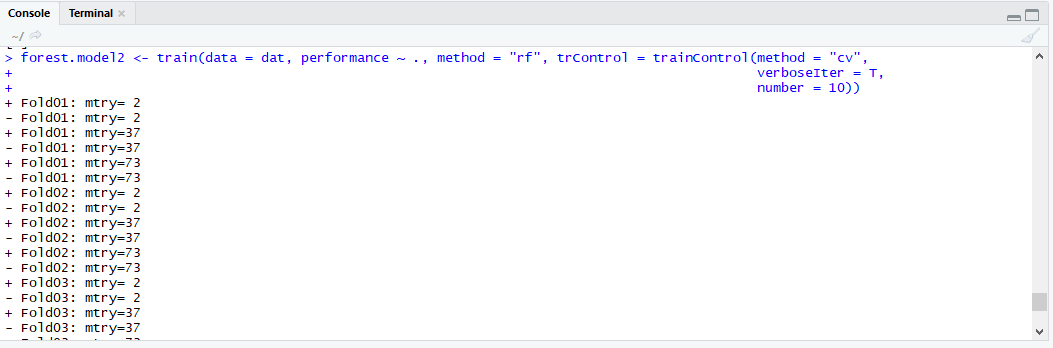


From above none of the values produce a deviation from constant accuracy rate of 60%

**Random Forest:**



We get 68% accuracy rate which is higher so far



After doing cross validation accuracy is 64.3% which is smaller value than 68% we got initially.

**SVM:**



It gets 67.3% which is remarkable among all accuracies.

**Conclusion:**

Out of all models after doing data cleaning, data preprocessing and checking accuracies of all classification algorithms SVM yields better for the data set.