

**ANALYSIS ON FACTORS INFLUENCING STUDENT GRADES**

A Project Report

Presented to

The Department of Data Analytics Engineering

George Mason University

In Partial Fulfillment

Of the Requirements for the

Masters Degree in

**Data Analytics Engineering**

By

Bhargav Ram Nara -G01223760

Kuldip Gadapa -G01239022

Sai Tejaswi Bellapukonda -G01181390

George Mason University

4400 University Dr, Fairfax, VA 22030

**Abstract:**

This is a survey held between courses Portuguese and math language in secondary school. There are interesting things like gender, social, study information along with this alcohol consumed by the students are also included in the data set. The project intends to find the grades of students based on the consumption of alcohol by the students. First, we preprocessed the data then we saw the attributes which are used for predicting the final grades student gets. Then we combined both the datasets, and we correlate various characteristics in the datasets to find the main features which affect the grades of the students. Then we also check whether the consumption of alcohol by the students on weekdays and weekends is a strong predictor or not. Finally, we use different types of classifiers, such as logistic regression, Random forest, Gradient Boosting, and Decision trees, to find the best model.

**Dataset Description:**

Source: The Dataset was taken from UCI Machine Learning Repository, but was initially published in proceedings of the 5th future business technology conference (FUBUTEC2008) by P. Cortez and A.Silva. The dataset contains two CSV files, one is for the Portuguese course, and the other is for a math course. The dataset contains 33 columns in both the math and Portuguese datasets. There are 395 students in a math course, and 649 are registered in Portuguese courses out of this 1044 students, 382 students are enrolled for both the courses.

There are different features present in the dataset, such as age, weekend, weekday alcohol consumption, Father and Mother’s education, Health status, Family size of the Students, Internet access at home, time allocated by the student for both traveling and studying. There are some features that depend on other features, and there are some independent features. The main feature of the dataset is alcohol consumption, which is ranging values between 1 to 10, where 1 represents the low level of alcohol consumption, and 10 represents the maximum amount of alcohol consumption.

**Preprocessing:**

There are no null values in our dataset, but we have removed some features such as an address, school, Mother and Fathers job, Family size, and Guardian. We have extracted these features because there is no importance for these features in the analysis part. We have introduced GG and AG variables in the dataset, and also we split the data into a training set with 80% samples and testing set with 20% samples.

**Exploratory Analysis:**

Analysis is very crucial for exploring the dataset. To get an in-depth analysis, we have performed correlation tests. We have used this correlation test to measure the linear dependence among the two variables. It is also known as a parametric correlation test, as it is dependent on the distributed data. It can only be used when both x and y are in normal distribution form. Then we have performed some visualizations and generated summary statistics for some columns. We have used the box plot for each period of grades for each understanding. Later we have also created bar graphs for alcohol consumption based on the gender and final grades.

**Principle Component Analysis**

PCA is the best analysis to identify the variable of categorical type and datatype that represents most of the information in a level or feature format. Specific to our data set, it is observed that a column gives lot of records that shows varied trends of relation with the overall data columns. We can consider age as a reference where the students have a range of 8 years starting from 15 to having a maximum of 22 i.e., it shows a normal distribution on overall count. On contrast, variables like paid and nursery have a less significance on the outcome but they could also be responsible for a cluster of cases. A feature like paid; extra pay for a class, could be considered as 94% of the observances show that students do not opt for an extra class.

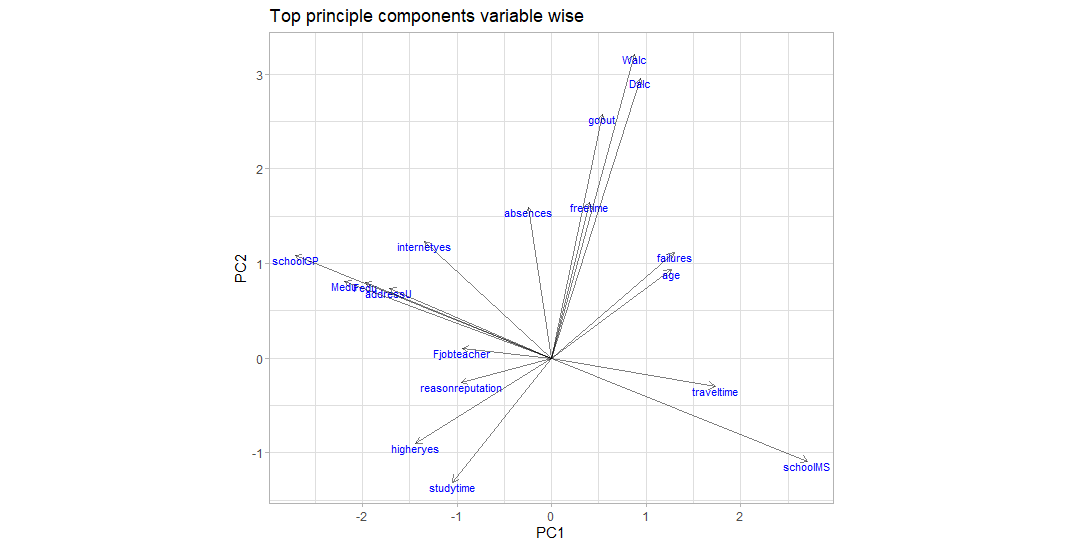
PCA ignores all these real world representations and it only watches out the principle components which are optimised for the highest variation irrespective of the variable kind. A case like the name of the school they study would be same for all the students as the data is collected by a single school or adding any field with all similar values does not add any meaning or description of relations amoung the students.

After the data transormation of Gaverage and GG we find that they can really make insights on our analysis.This transformation helps us in having a good understanding about the corelation importance for categorical variables. I have used spearman corelation to commute the corelation factor instead of any hypothesis or chi-square test.As this is a common survey or research data collection it is not specific that the data has good corelation but nearly one third (10) variables shows a value above median. Now we set a dataframe for factors having abs-cor value greater than 0.08i.e; above mean range. There is a low correlation factor between the variables seen and it would not be easy to predict the grades if we set the grade factor fue to multicolinearity. Here we observe that schoolIMS and SchoolIGS are highly correlated and the pairs like father and mother eduction, Walc and Dalc also have a good correlation factor.

On fitting a LM we observe the factor significance of each variable. This plot explains us about how the varience factor in PCA is seen for variables above the mean level and the 3rd quartile shows a good increase in varience. We see that the first 10 variables show us a good explanation about the data more than half have been represented by the top 5 variables.

Using PCA model is really valuable for this kind of data because it is we can omit the least correlated factors and apart from it we can also reduce the multi dimensionality factor. It can also draw crutial insights like which variables tend to signify the most importance to the data goal and

displaying the Top principle components. PCA can be treated as a best practice for having an overall idea and important factors in our obervation. Apart from that we can consider pca to interpret even more dificut tasks and cateogorical data. The prediction factor could be increased. We also get the positive and negative factors that affects the students overall performance. Students who mostly drink during the week and have failures in the past tend to have more less grades and it is predictable. It is considered as a valid outcome.



**Multiple Linear Model**

The MLR is an extension to the simple linear model. It is used to fot the prediction factor on a variable based on the multiple distinct predictor variable. In our case we have taken the average of grade to be our outcome or response variable and used several other predictor variables to measure the association in between the predictor and outcomes.

We have initially fit the models in between the students grades to obser the fit in between them and made other analysis on the factors that most reflected or influenced the grade.

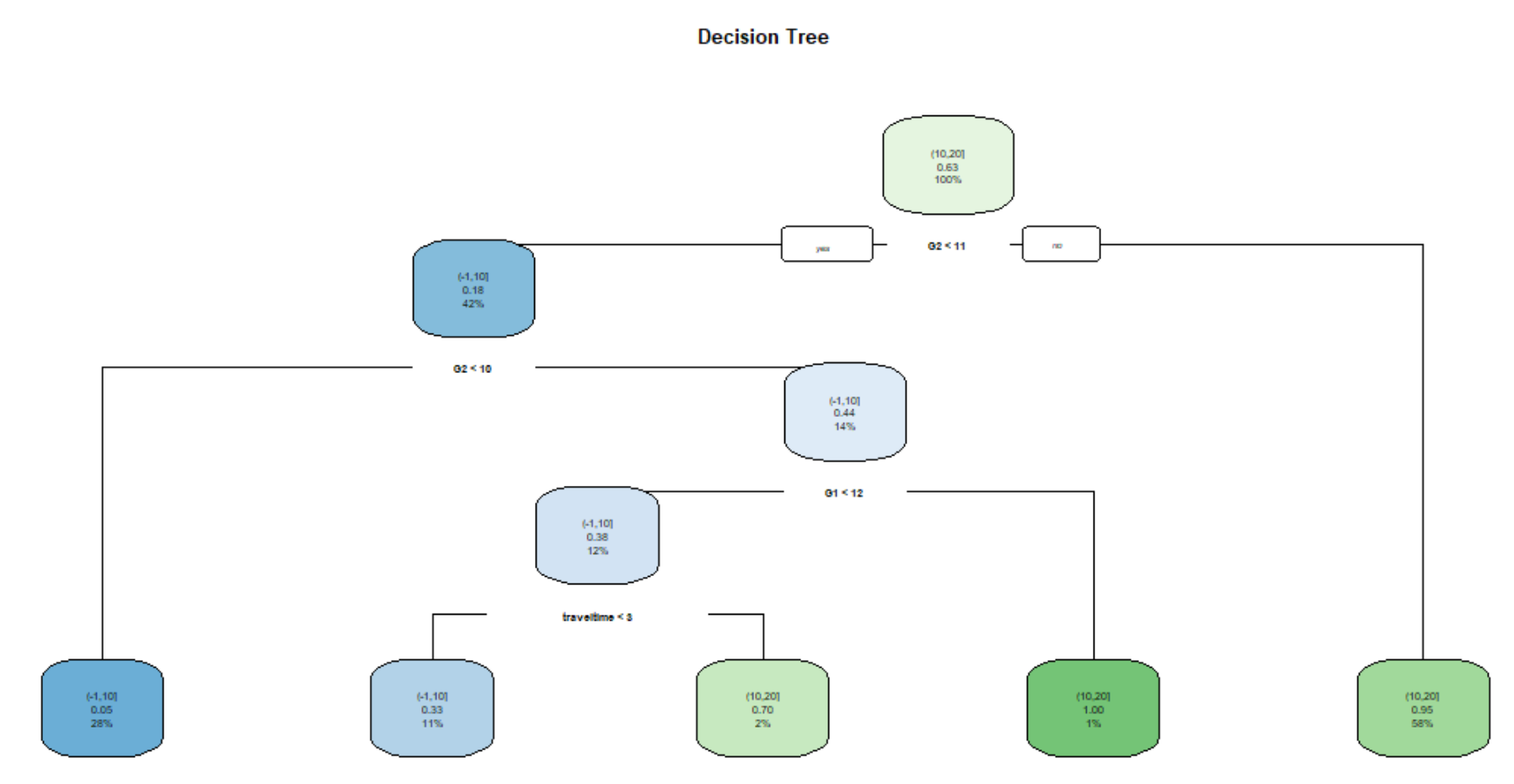
**Logistic Regression:**

We have used logistic regression, which allows us to estimate the probability of the categorical results based on one or more predictor variables. One can say that when there is a presence of the predictor variable, there is a probability increase or decrease in the outcome by some specific percentage. Initially, we consider predicting the period grades. Though we got an accuracy of 93%, the AIC value we got is 293. For reducing the AIC value, we have used the backward method by reducing the independent variables, which resulted in AIC value as the 276 and accuracy of 94%. We consider only the first model, as there is only one percent change.

**Decision Tree**

Decision Trees are the versatile Machine Learning algorithms that can perform both classification and regression tasks. They are very powerful and are capable of fitting complex datasets. They are also fundamental components of random forests, the most important machine learning algorithms available today.

The benefits of handling and modelling decision trees are, The predictions on future decisions can be made and the uncertainity on confidence impacted is solved. The measurement of risks and varying components is analysed. The proper understanding of sequencing and interrelations for tasks and events is understood. The accuracy we found using decision tree is 93.5%



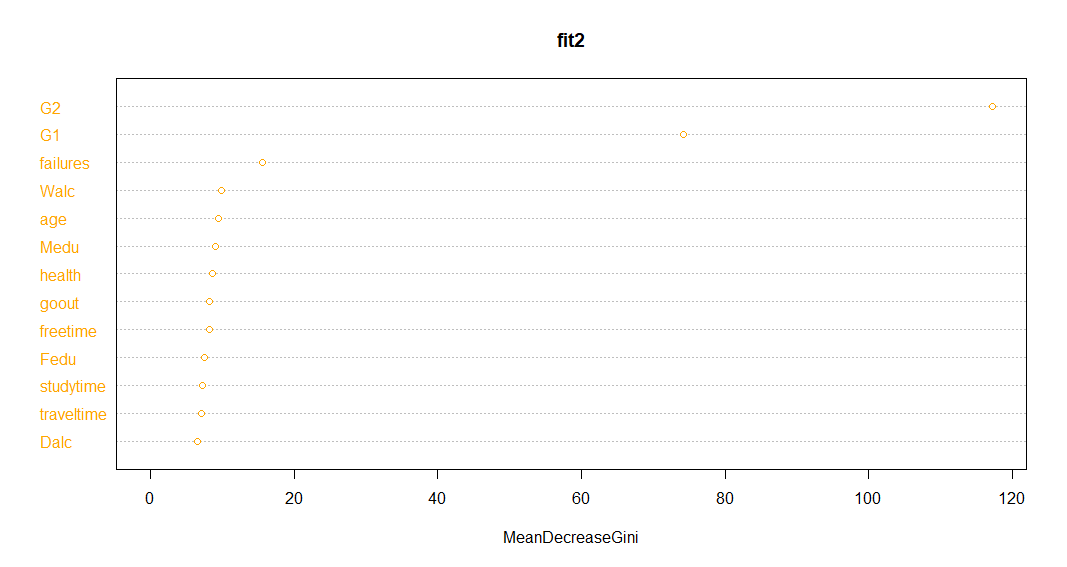
**Random Forest**

A random forest is an ensemble of decision trees for regression. Each of the regression tree models is learned on a different set of rows or records and/ a different set of columns describing the attributes, whereby the latter can also be a bit/byte/double vector descriptor. The output model describes an ensemble of regression tree models and is applied in the corresponding predictor node using a simple mean of the individual predictions.

The Random Forest implements Breiman’s random forest algorithm for the classification and regression. It can also be used in unsupervised mode for assessing proximities with in the data points. Random forest is a average multiple deep decision tree that is trained using a training set and with the goal of overcomming the over-fitting problem of the classification or simple decision tree.

It prevents the generalized arrangements of tree formations, decesion tree overfitting and gives the accurate fit but it may not be true for the validation stage. Random forest can be used for both continuous and categorical variables.

In the process we implemented random forest modelfit using 80% of the data as train data. The sample is fit to make the selection on the specified variables and we have performed the bagging to decrease the OOB error rate and the votes of that class. The accuracy we find using this model is 93.77%.



**RESULTS**

The main important criteria that effects the grades of the students is considered to be drinking on the weekend. The boys are mostly failing compared to gilrs because the girls had a low alcohol consumption during the weekends.Approximately 75% of the male students perform no better than the median female students except some outliers. The male students performance is not upto the mark when alcohol consumption is taken as a factor.

On observations made using principle component analysis, it is clear that the main reason behind low grades for students is due to the alcohol consumption. Factors like the school they study and the educational qualifications of their parebts also affects their performance in the examinations. This is because, if they are not intrested in their school or subject they pay least effort in understanding the concepts and would not perform well in exams. Similarly, if the students parents are not having minimum qualifications they may not be able to help their children while preperation or would assign their work for the students by cauisng a deviation. The predicted accuracies in all models are varied with slight difference. From all the above implemented models, Random Forest is best.

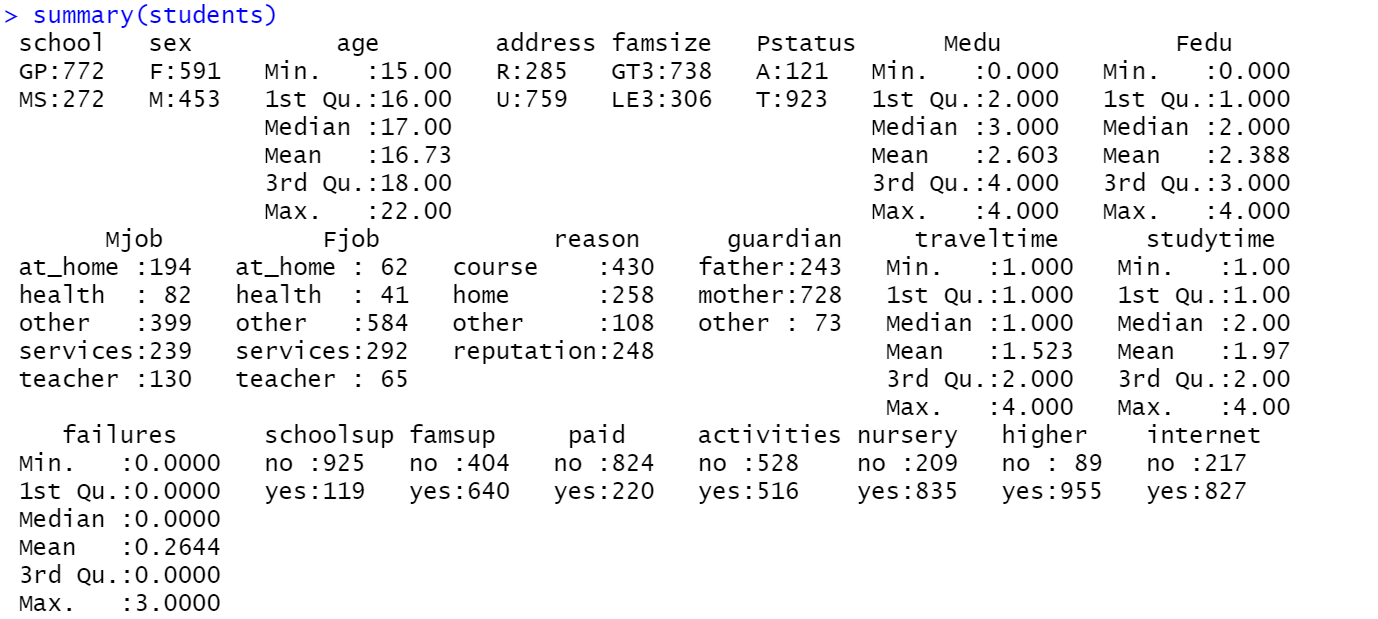
# References

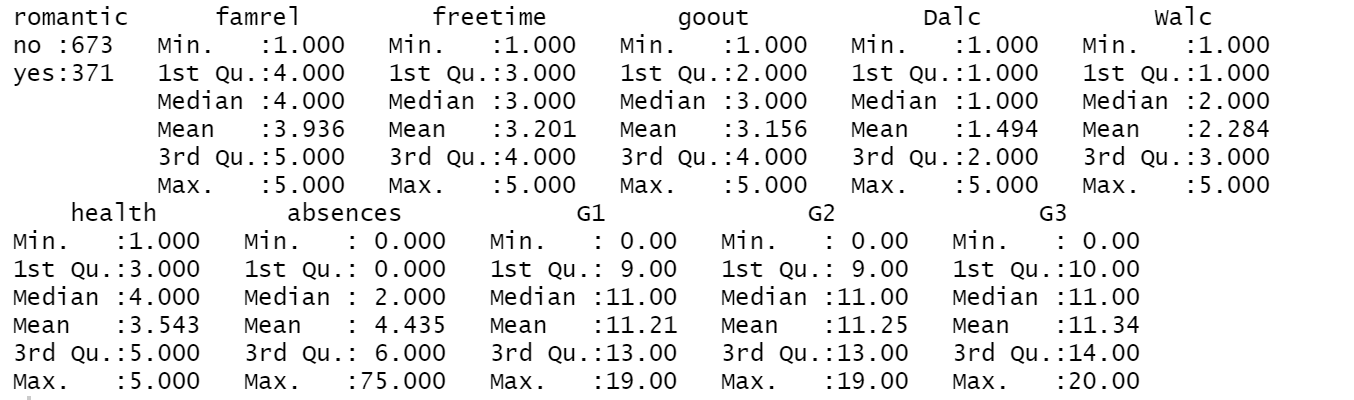
Julian. (2018). *Student Alcohol Consumption*. Retrieved from RPubs by RStudio: https://www.rpubs.com/heke0495/419722

Lyons, K. (2016). *UCI-Student-Alcohol-Consumption*. Retrieved from GitHub, Inc.: https://github.com/KMLDS/UCI-Student-Alcohol-Consumption

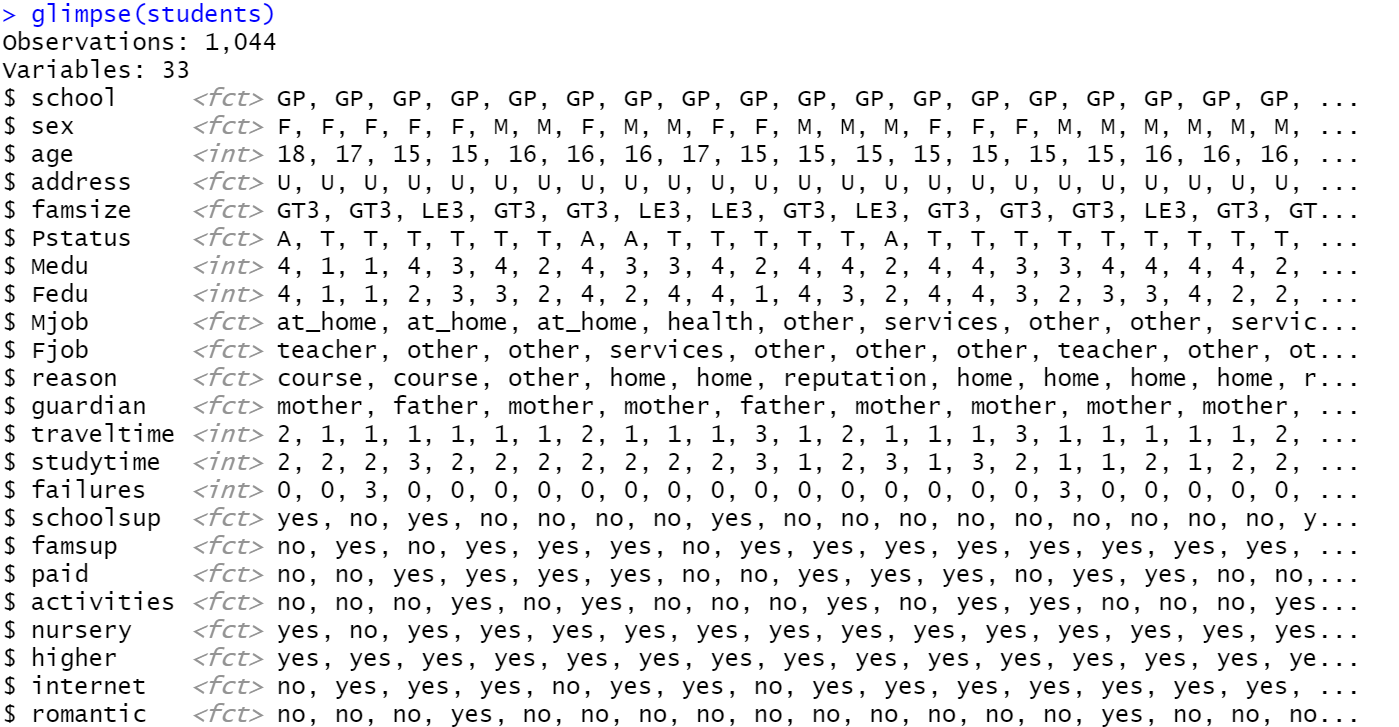
*Student Alcohol Consumption*. (n.d.). Retrieved from kaggle: https://www.kaggle.com/uciml/student-alcohol-consumption

**Appendix**

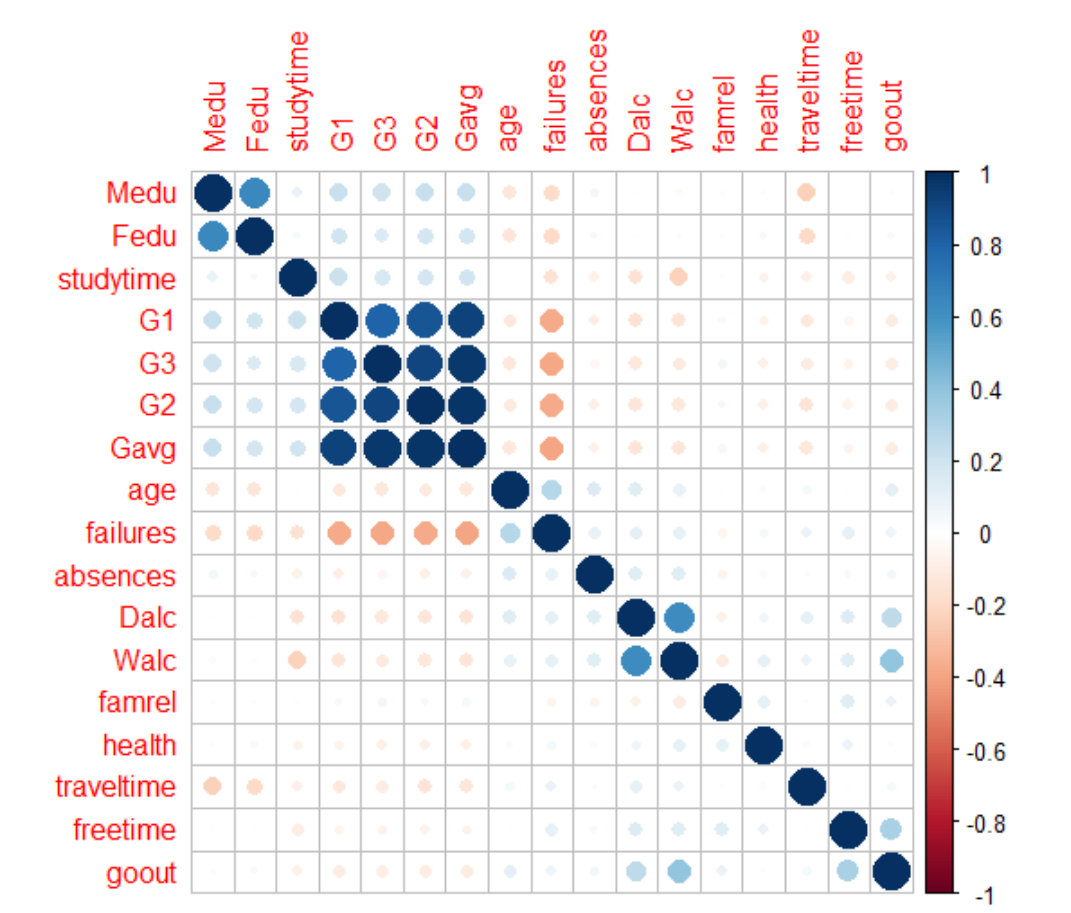




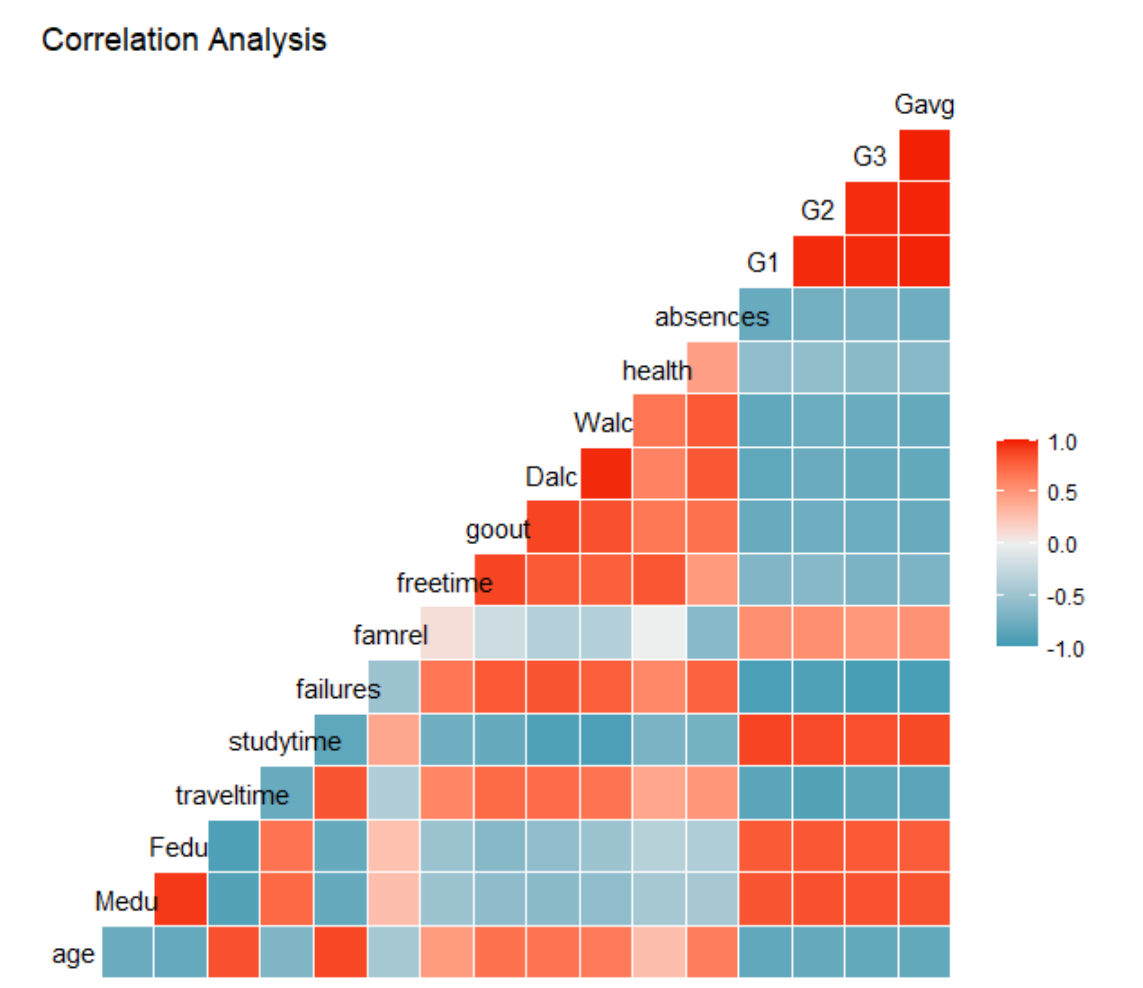
Summary statistics of Original data

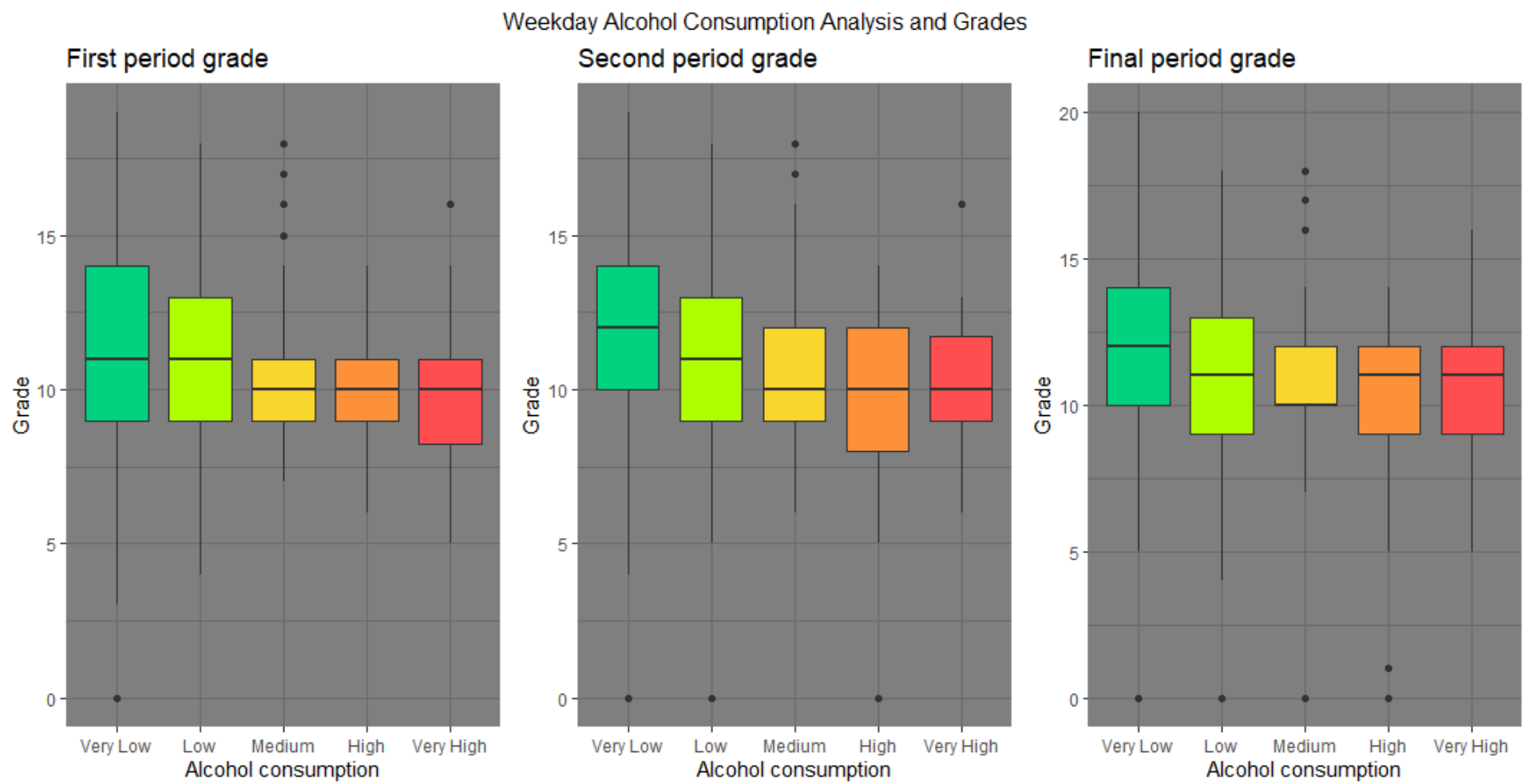


Glimpse of Students data

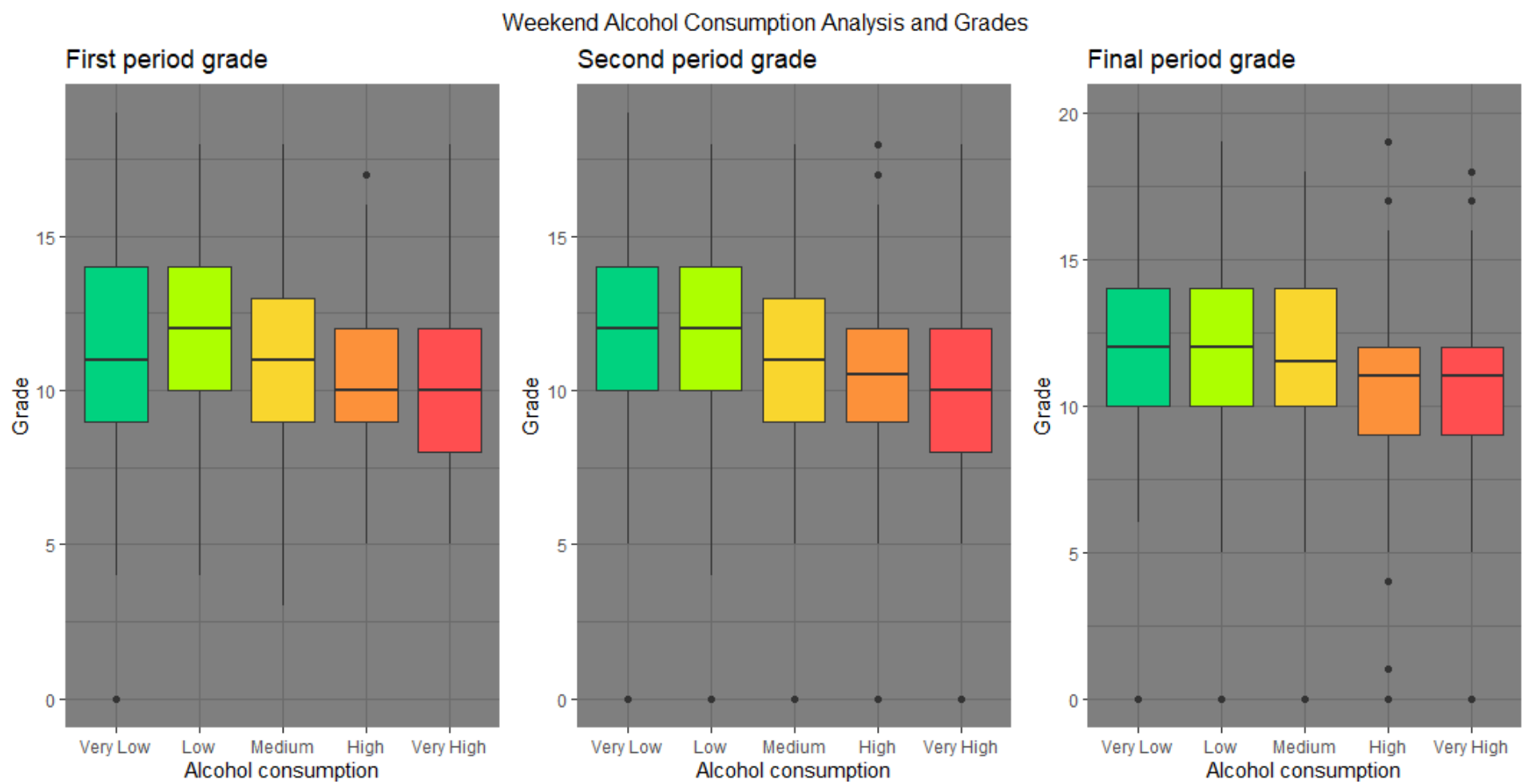


Correlation plot

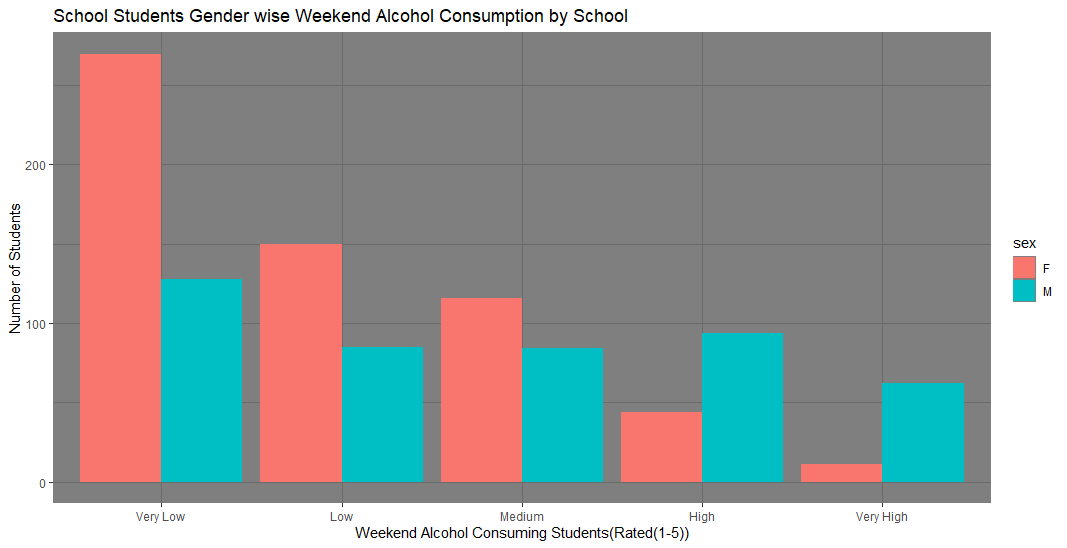




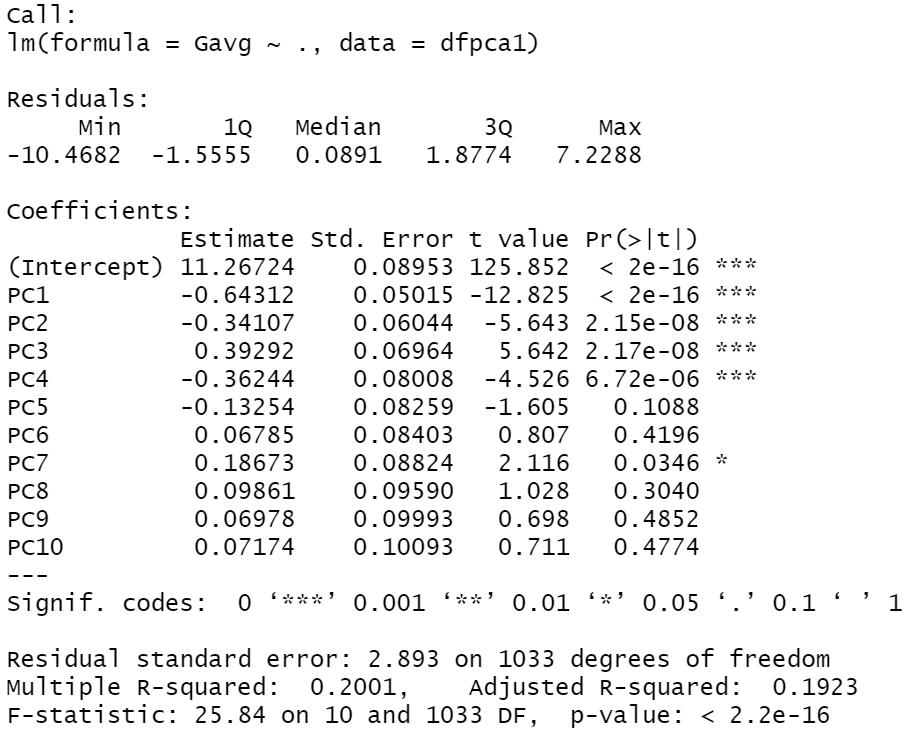
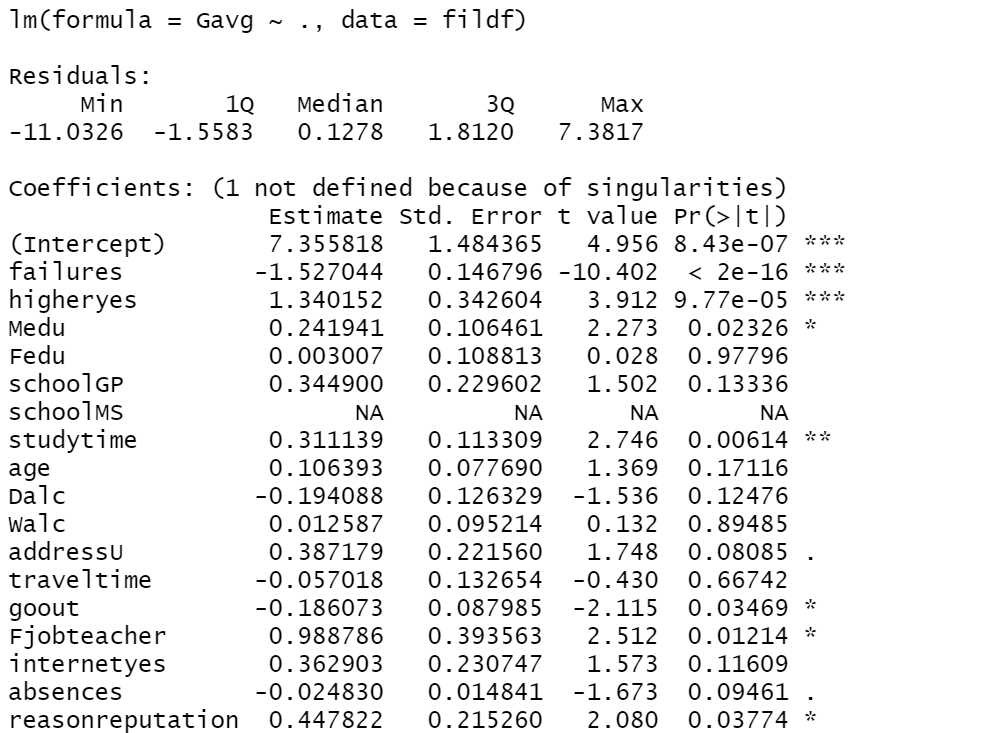
The box plot represents weekday alcohol consumption analysis and grades

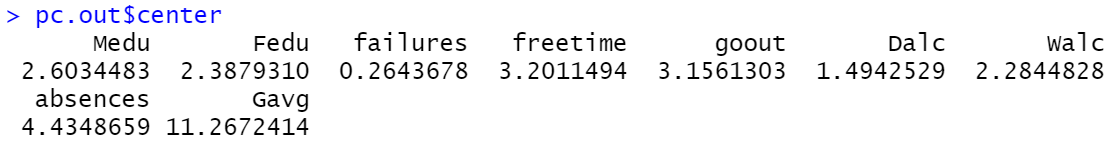


The box plot represents weekend alcohol consumption analysis and grades

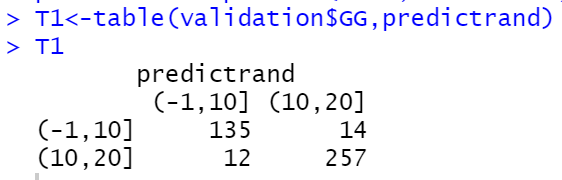


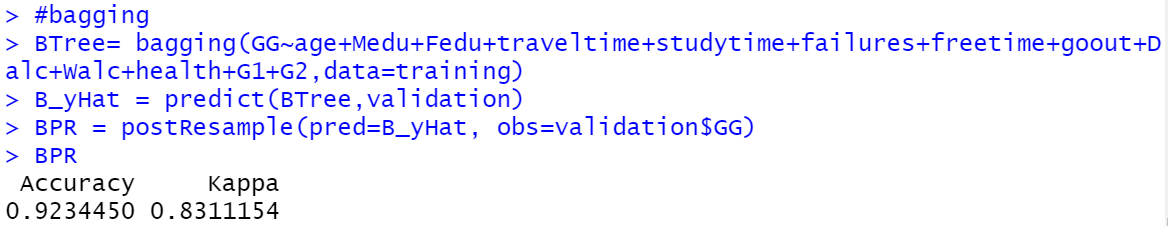
The bar plot represents students genderwise weekend alcohol consumption

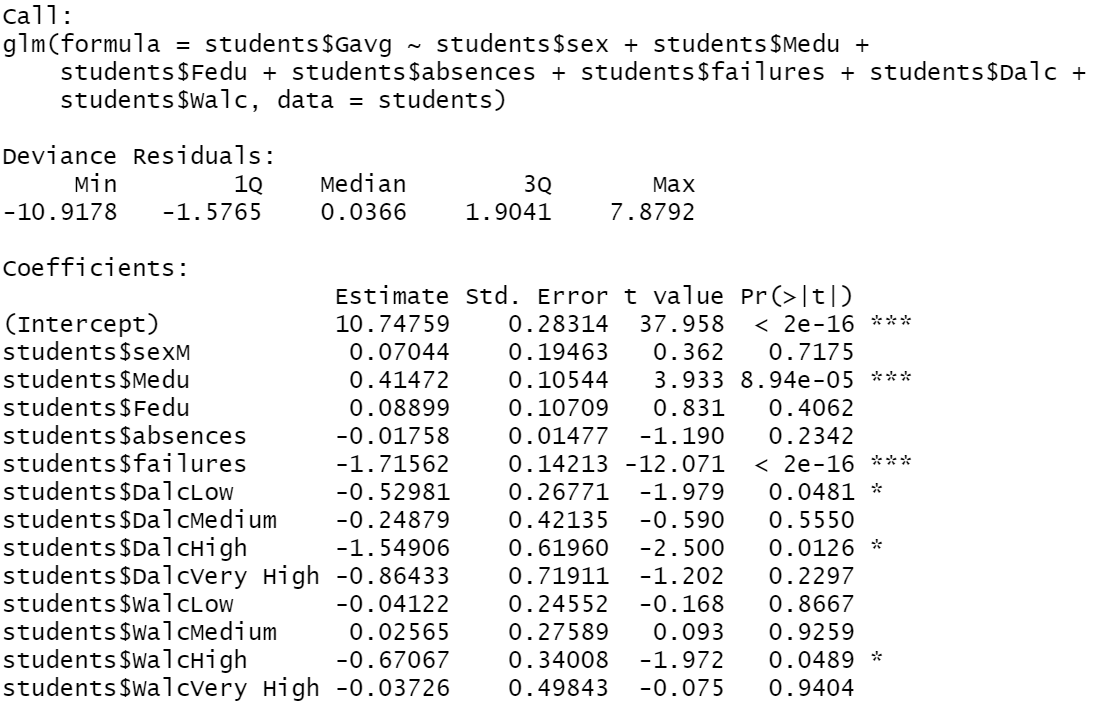


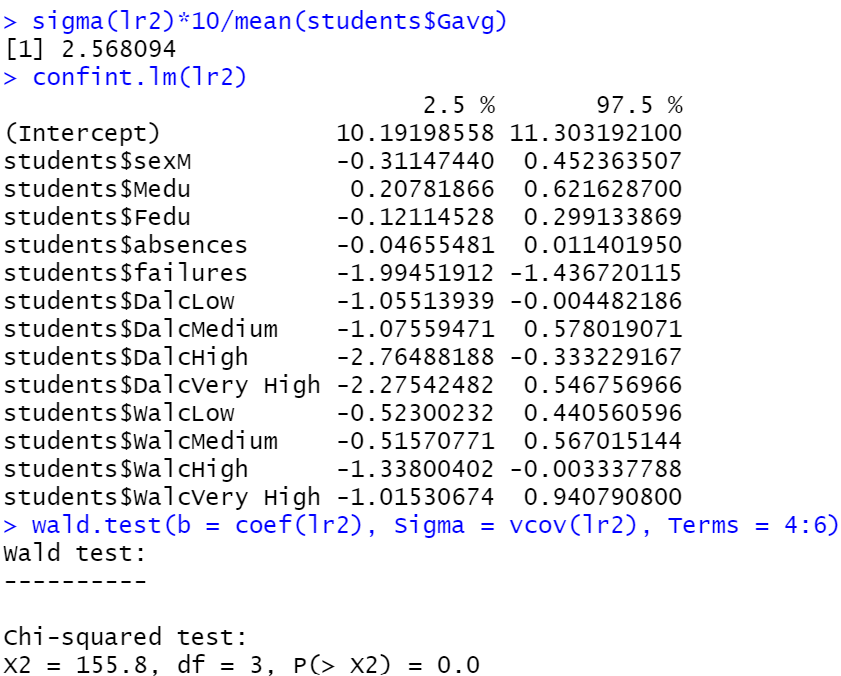


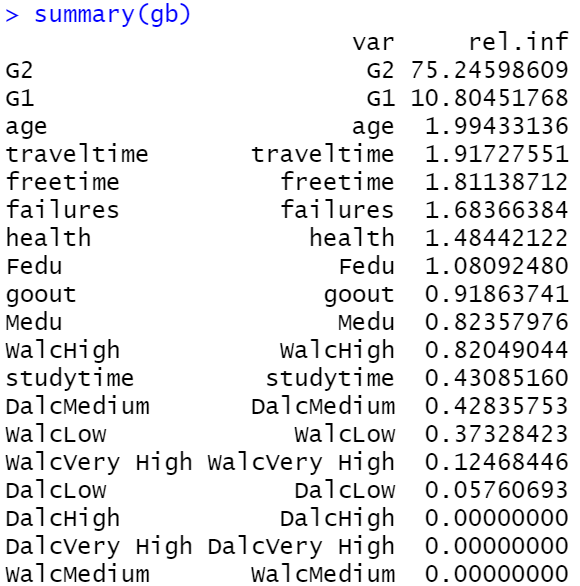
Summary statistics for priciple component analysis



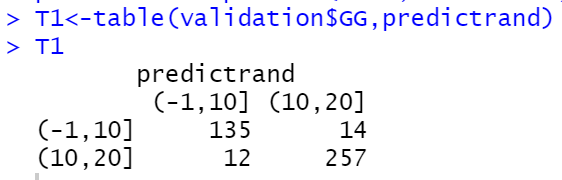


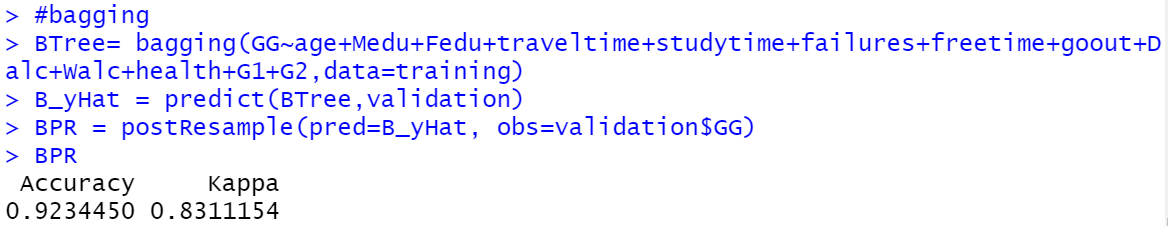




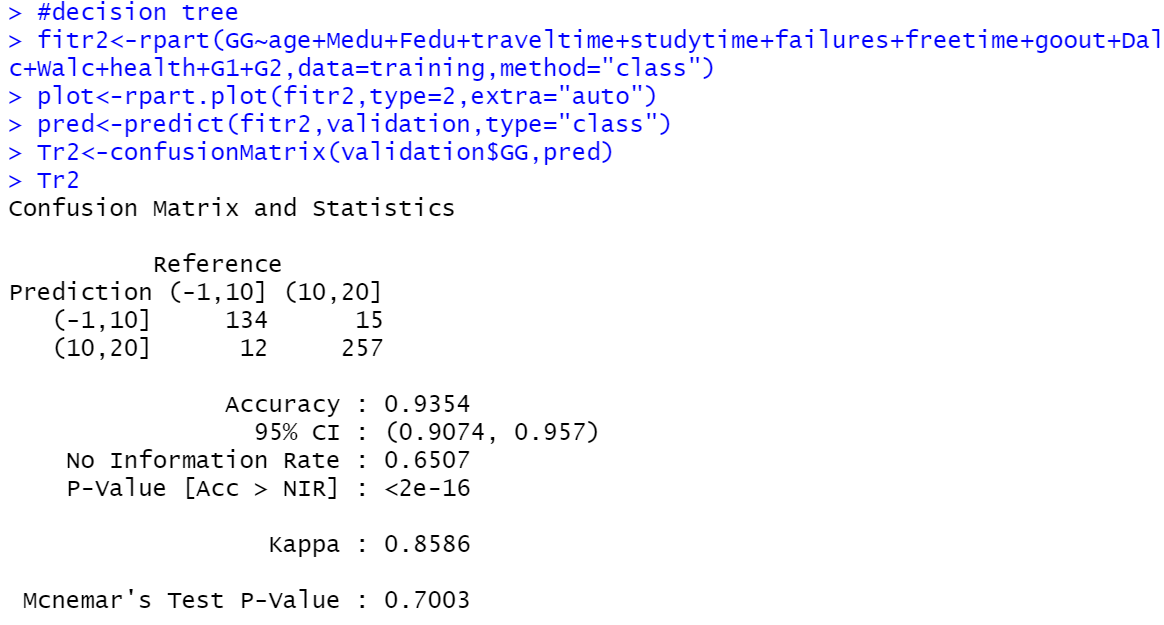


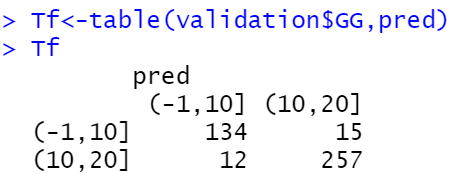
Summary statistics for logistic regression





Summary statistics for random forest model





Summary statistics for decision tree