DATA MINING & VISUALIZATION
ANALYSIS OF STUDENT PERFORMANCE USING RANDOM FOREST MODEL

By:

Devin Augustin	**2440094352**
Audrey Tabitha Ariani	**2440082812**
Jocelyn Wievin	**2440063025**
Samuel Sebastian	**2440085316**
Rico Frenaldi Tokanto	**2440114373**

COMPUTER SCIENCE & STATISTICS
BINUS UNIVERSITY

TABLE OF CONTENTS

Abstract 2

Background: 2

Objectives: 3

Methodology: 3

Result: 4

- Head and Summary of Dataset using R: 4
- Data cleaning: 5
- Correlation Heatmap: 6
- Categorical Data: 8
- Numerical data: 16
- Categorize the Final Grade (G3) 25
- Split the data into two for training and testing data 25
- Random Forest Classifier 26

Discussion: 33

Conclusion: 33

Resources: 34

1. Abstract

The Random Forest Model is a classification algorithm that consists of decision trees which uses bagging and features randomness for each tree to create an uncorrelated forest of trees. This prediction will be more accurate than any of the individual trees. The Random Forest Model has a lot more parameters than Linear Models. Therefore, the Random Forest Model is more suitable to analyze the data set of student performances in Portugal, collected from two different secondary schools. The goal of this study is to predict the final score of the students, which will be done by data exploration with R, data cleaning, checking the correlation between variables, data testing, and then summarizing the result. The model and method used in this study has proved to be effective and accurate.

2. Background:

Education is a powerful tool that can be used to drive the world's economic growth, prosperity, and equality across different groups of people. One of the most fundamental stages of education is completing a secondary degree before pursuing a higher degree, where those who show a higher performance or aptitude are, more often than not, predicted to achieve a higher level of accomplishment or success in their area of studies/career(s). A wide range of institutions recognize that academic performance is a product of many different factors and in order to determine the most impactful ones, they must be measured relative to their final grades.

This study uses two datasets within the "Student Performance Data Set" collected from two different secondary schools in Portugal, where one of them stores the data of 395 students pursuing the Mathematics course and 649 students pursuing the Portuguese course. Out of 1044 students, 382 of them pursued both, and those students are the sample in this case study. Each of the two dataset measures 30 different variables. The variables and their description are as follows:

- school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
- sex - student's sex (binary: "F" - female or "M" - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: "U" - urban or "R" - rural)
- famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
- Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
- guardian - student's guardian (nominal: "mother", "father" or "other")
- traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)

- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)
- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)
- G3 - final grade (numeric: from 0 to 20, output target)

Additionally, the independent variables are the grades of each student in each of the Mathematics and Portuguese courses.

The purpose of this case study is to study the relationship of each variable and how much they affect the final score of the students. In order to do so, we chose the random forest learning classification model. Random forest is a supervised learning algorithm often used in regression (numerical target variable) and classification (categorical target variable) problems, and it will suffice the purpose of this study which is to predict the performance/final score of students based on several impactful factors.

3. Objectives:

a. Predict students performance in secondary school
b. Determine the variables with strong correlation to students performance

4. Methodology:

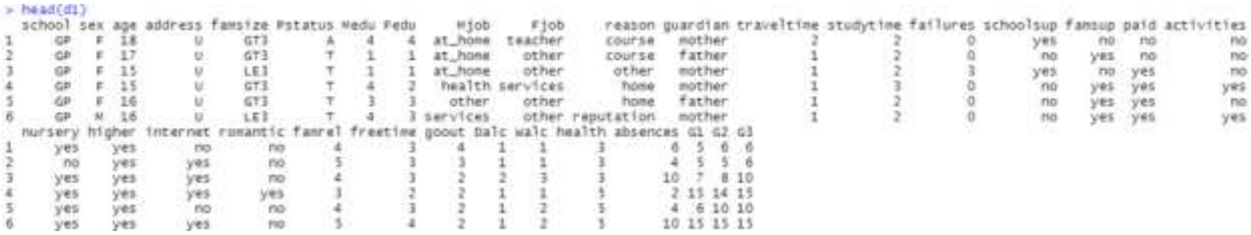
In order to predict students' performance, it is necessary to do exploratory data analysis beforehand in order to acknowledge the basic characteristics of the data. Afterwards, we would generate basic visualizations such as the correlation heatmaps for the numerical categories and individual boxplot and violin plots for each categorical variable based on their groups. Next, we would categorize the target variable into seven groups to minimize the errors and split the sample dataset into an 80:20 ratio for training and testing on a random tree classification model. Random tree classification was chosen because it was the most suitable for a dataset where the variables do not have linear relationship with another. Finally, we would receive the summary of the classification model.

5. Result:

a. Head and Summary of Dataset using R:

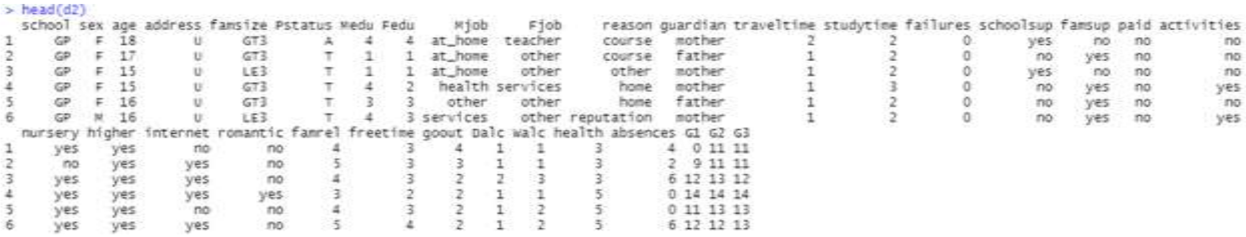
Head:

- D1 (Students Studying Mathematics)



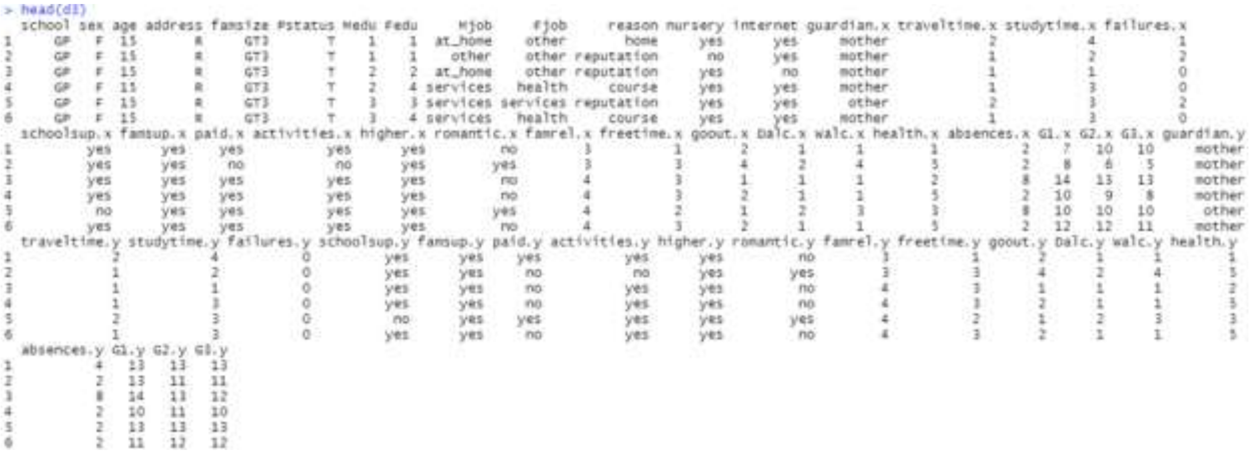
```
> head(d1)
  school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason guardian traveltime studytime failures schoolsup famsup paid activities
1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course   mother          2         2        0       yes     no   no         no
2     GP   F  17       U     GT3       T    1    1  at_home    other     course   father          1         2        0        no    yes   no         no
3     GP   F  15       U     LE3       T    1    1  at_home    other      other   mother          1         2        3       yes     no  yes         no
4     GP   F  15       U     GT3       T    4    2   health services       home   mother          1         3        0        no    yes  yes        yes
5     GP   F  16       U     GT3       T    3    3    other    other       home   father          1         2        0        no    yes  yes         no
6     GP   M  16       U     LE3       T    4    3 services    other reputation   mother          1         2        0        no    yes  yes        yes
  nursery higher internet romantic famrel freetime goout Dalc Walc health absences G1 G2 G3
1     yes    yes       no       no      4        3     4    1    1      3        6  5  6  6
2      no    yes      yes       no      5        3     3    1    1      3        4  5  5  6
3     yes    yes      yes       no      4        3     2    2    3      3       10  7  8 10
4     yes    yes      yes      yes      3        2     2    1    1      5        2 15 14 15
5     yes    yes       no       no      4        3     2    1    2      5        4  6 10 10
6     yes    yes      yes       no      5        4     2    1    2      5       10 15 15 15
```

- D2 (Students Studying Portuguese)



```
> head(d2)
  school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason guardian traveltime studytime failures schoolsup famsup paid activities
1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course   mother          2         2        0       yes     no   no         no
2     GP   F  17       U     GT3       T    1    1  at_home    other     course   father          1         2        0        no    yes   no         no
3     GP   F  15       U     LE3       T    1    1  at_home    other      other   mother          1         2        0       yes     no   no         no
4     GP   F  15       U     GT3       T    4    2   health services       home   mother          1         3        0        no    yes   no        yes
5     GP   F  16       U     GT3       T    3    3    other    other       home   father          1         2        0        no    yes   no         no
6     GP   M  16       U     LE3       T    4    3 services    other reputation   mother          1         2        0        no    yes   no        yes
  nursery higher internet romantic famrel freetime goout Dalc Walc health absences G1 G2 G3
1     yes    yes       no       no      4        3     4    1    1      3        4  0 11 11
2      no    yes      yes       no      5        3     3    1    1      3        2  9 11 11
3     yes    yes      yes       no      4        3     2    2    3      3        6 12 13 12
4     yes    yes      yes      yes      3        2     2    1    1      5        0 14 14 14
5     yes    yes       no       no      4        3     2    1    2      5        0 11 13 13
6     yes    yes      yes       no      5        4     2    1    2      5        6 12 12 13
```

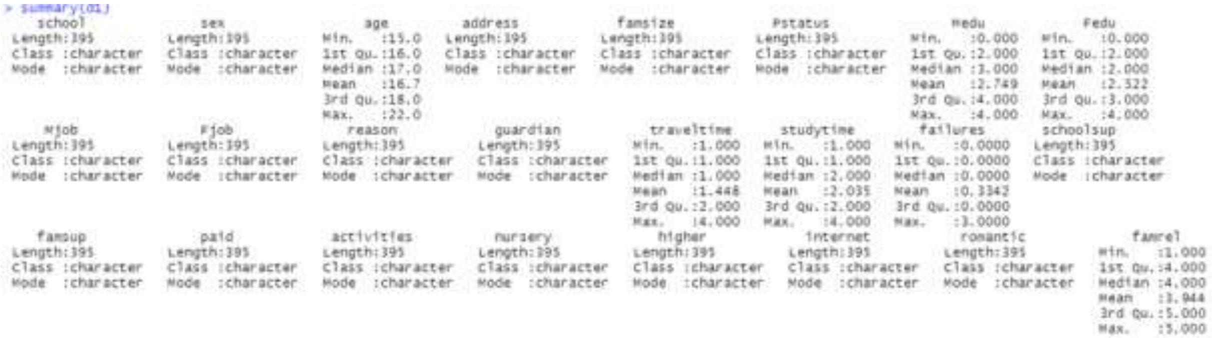
- D3 (Students Studying Both Mathematics & Portuguese)



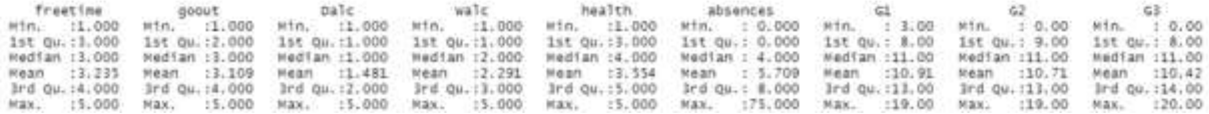
```
> head(d3)
  school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason nursery internet guardian.x traveltime.x studytime.x failures.x
1     GP   F  15       R     GT3       T    1    1  at_home    other       home     yes      yes     mother            2           4          1
2     GP   F  15       R     GT3       T    1    1    other    other reputation      no      yes     mother            1           2          2
3     GP   F  15       R     GT3       T    2    2  at_home    other reputation     yes       no     mother            1           1          0
4     GP   F  15       R     GT3       T    2    4 services   health     course     yes      yes     mother            1           3          0
5     GP   F  15       R     GT3       T    3    3 services services reputation     yes      yes      other            2           3          2
6     GP   F  15       R     GT3       T    3    4 services   health     course     yes      yes     mother            1           3          0
  schoolsup.x famsup.x paid.x activities.x higher.x romantic.x famrel.x freetime.x goout.x Dalc.x Walc.x health.x absences.x G1.x G2.x G3.x guardian.y
1         yes      yes    yes          yes      yes         no        3          1       2      1      1        1          2    7   10   10     mother
2         yes      yes     no           no      yes        yes        3          3       4      2      4        5          2    8    6    5     mother
3         yes      yes    yes          yes      yes         no        4          3       1      1      1        2          8   14   13   13     mother
4         yes      yes    yes          yes      yes         no        4          3       2      1      1        5          2   10    9    8     mother
5          no      yes    yes          yes      yes        yes        4          2       1      2      3        3          8   10   10   10      other
6         yes      yes    yes          yes      yes         no        4          3       2      1      1        5          2   12   12   11     mother
  traveltime.y studytime.y failures.y schoolsup.y famsup.y paid.y activities.y higher.y romantic.y famrel.y freetime.y goout.y Dalc.y Walc.y health.y
1            2           4          0         yes      yes    yes          yes      yes         no        3          1       2      1      1        1
2            1           2          0         yes      yes     no           no      yes        yes        3          3       4      2      4        5
3            1           1          0         yes      yes     no          yes      yes         no        4          3       1      1      1        2
4            1           3          0         yes      yes     no          yes      yes         no        4          3       2      1      1        5
5            2           3          0          no      yes    yes          yes      yes        yes        4          2       1      2      3        3
6            1           3          0         yes      yes     no          yes      yes         no        4          3       2      1      1        5
  absences.y G1.y G2.y G3.y
1          4   13   13   13
2          2   13   11   11
3          8   14   13   12
4          2   10   11   10
5          2   13   13   13
6          2   11   12   12
```

Summary:

- D1 (Students Studying Mathematics)

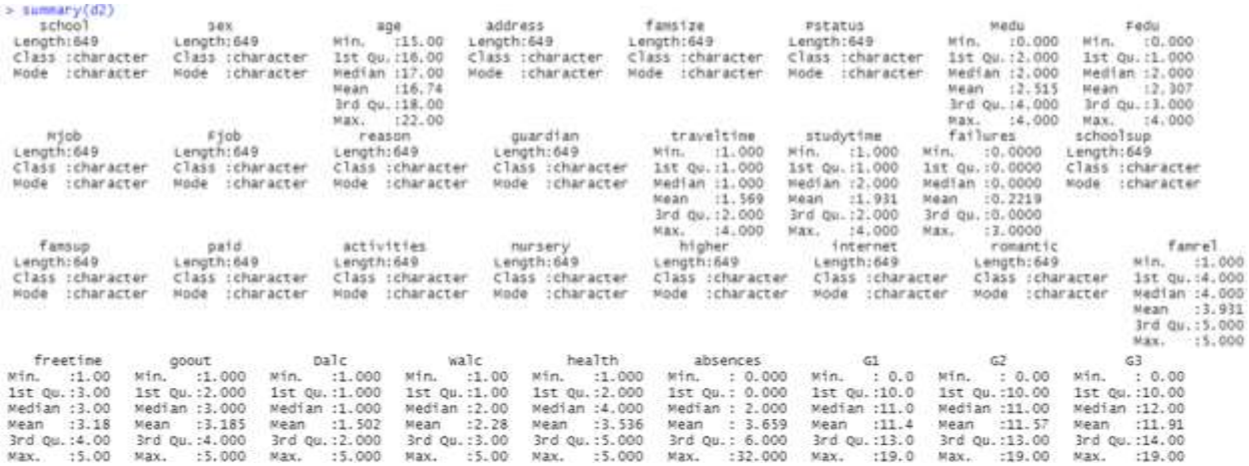


```
> summary(d1)
    school              sex                 age          address            famsize            Pstatus               Medu            Fedu      
 Length:395         Length:395         Min.   :15.0   Length:395         Length:395         Length:395         Min.   :0.000   Min.   :0.000  
 Class :character   Class :character   1st Qu.:16.0   Class :character   Class :character   Class :character   1st Qu.:2.000   1st Qu.:2.000  
 Mode  :character   Mode  :character   Median :17.0   Mode  :character   Mode  :character   Mode  :character   Median :3.000   Median :2.000  
                                       Mean   :16.7                                                            Mean   :2.749   Mean   :2.322  
                                       3rd Qu.:18.0                                                            3rd Qu.:4.000   3rd Qu.:3.000  
                                       Max.   :22.0                                                            Max.   :4.000   Max.   :4.000  
     Mjob               Fjob              reason            guardian           traveltime      studytime        failures       schoolsup        
 Length:395         Length:395         Length:395         Length:395         Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:395        
 Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.000   Median :2.000   Median :0.0000   Mode  :character  
                                                                             Mean   :1.448   Mean   :2.035   Mean   :0.3342                     
                                                                             3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000                     
                                                                             Max.   :4.000   Max.   :4.000   Max.   :3.0000                     
    famsup              paid            activities          nursery             higher            internet           romantic             famrel     
 Length:395         Length:395         Length:395         Length:395         Length:395         Length:395         Length:395         Min.   :1.000  
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:4.000  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :4.000  
                                                                                                                                      Mean   :3.944  
                                                                                                                                      3rd Qu.:5.000  
                                                                                                                                      Max.   :5.000  
```



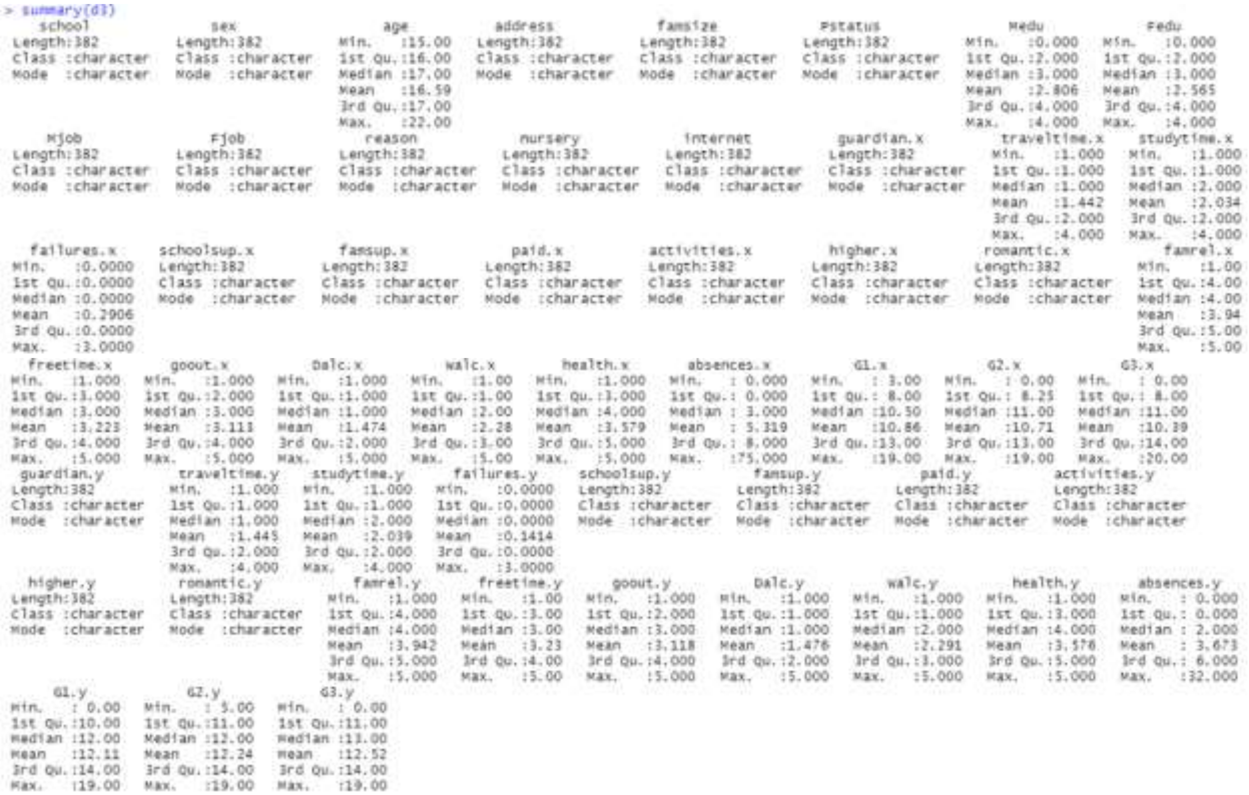
```
   freetime         goout            Dalc            Walc           health          absences            G1              G2              G3       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00  
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00  
 Median :3.000   Median :3.000   Median :1.000   Median :2.000   Median :4.000   Median : 4.000   Median :11.00   Median :11.00   Median :11.00  
 Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291   Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71   Mean   :10.42  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00  
```

- D2 (Students Studying Portuguese)



```
> summary(d2)
    school              sex                 age          address            famsize            Pstatus               Medu            Fedu      
 Length:649         Length:649         Min.   :15.00   Length:649         Length:649         Length:649         Min.   :0.000   Min.   :0.000  
 Class :character   Class :character   1st Qu.:16.00   Class :character   Class :character   Class :character   1st Qu.:2.000   1st Qu.:1.000  
 Mode  :character   Mode  :character   Median :17.00   Mode  :character   Mode  :character   Mode  :character   Median :2.000   Median :2.000  
                                       Mean   :16.74                                                            Mean   :2.515   Mean   :2.307  
                                       3rd Qu.:18.00                                                            3rd Qu.:4.000   3rd Qu.:3.000  
                                       Max.   :22.00                                                            Max.   :4.000   Max.   :4.000  
     Mjob               Fjob              reason            guardian          traveltime      studytime        failures        schoolsup        
 Length:649         Length:649         Length:649         Length:649         Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:649        
 Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.000   Median :2.000   Median :0.0000   Mode  :character  
                                                                             Mean   :1.569   Mean   :1.931   Mean   :0.2219                     
                                                                             3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000                     
                                                                             Max.   :4.000   Max.   :4.000   Max.   :3.0000                     
    famsup              paid            activities          nursery             higher            internet           romantic             famrel     
 Length:649         Length:649         Length:649         Length:649         Length:649         Length:649         Length:649         Min.   :1.000  
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:4.000  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :4.000  
                                                                                                                                      Mean   :3.931  
                                                                                                                                      3rd Qu.:5.000  
                                                                                                                                      Max.   :5.000  
    freetime         goout            Dalc             Walc           health          absences            G1             G2              G3       
 Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00  
 1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000   1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00  
 Median :3.00   Median :3.000   Median :1.000   Median :2.00   Median :4.000   Median : 2.000   Median :11.0   Median :11.00   Median :12.00  
 Mean   :3.18   Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean   :3.536   Mean   : 3.659   Mean   :11.4   Mean   :11.57   Mean   :11.91  
 3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000   3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00  
 Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :32.000   Max.   :19.0   Max.   :19.00   Max.   :19.00  
```

- D3 (Students Studying Both Mathematics & Portuguese)

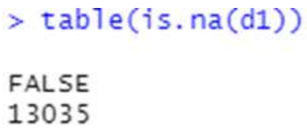


```
> summary(d3)
    school              sex                 age          address            famsize            Pstatus               Medu            Fedu      
 Length:382         Length:382         Min.   :15.00   Length:382         Length:382         Length:382         Min.   :0.000   Min.   :0.000  
 Class :character   Class :character   1st Qu.:16.00   Class :character   Class :character   Class :character   1st Qu.:2.000   1st Qu.:2.000  
 Mode  :character   Mode  :character   Median :17.00   Mode  :character   Mode  :character   Mode  :character   Median :3.000   Median :3.000  
                                       Mean   :16.59                                                            Mean   :2.806   Mean   :2.565  
                                       3rd Qu.:17.00                                                            3rd Qu.:4.000   3rd Qu.:4.000  
                                       Max.   :22.00                                                            Max.   :4.000   Max.   :4.000  
     Mjob               Fjob              reason            nursery            internet          guardian.x         traveltime.x    studytime.x   
 Length:382         Length:382         Length:382         Length:382         Length:382         Length:382         Min.   :1.000   Min.   :1.000  
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.000   Median :2.000  
                                                                                                                   Mean   :1.442   Mean   :2.034  
                                                                                                                   3rd Qu.:2.000   3rd Qu.:2.000  
                                                                                                                   Max.   :4.000   Max.   :4.000  
   failures.x      schoolsup.x          famsup.x            paid.x           activities.x        higher.x          romantic.x           famrel.x   
 Min.   :0.0000   Length:382         Length:382         Length:382         Length:382         Length:382         Length:382         Min.   :1.00  
 1st Qu.:0.0000   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:4.00  
 Median :0.0000   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :4.00  
 Mean   :0.2906                                                                                                                     Mean   :3.94  
 3rd Qu.:0.0000                                                                                                                     3rd Qu.:5.00  
 Max.   :3.0000                                                                                                                     Max.   :5.00  
   freetime.x       goout.x          Dalc.x          Walc.x         health.x       absences.x          G1.x            G2.x            G3.x      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00  
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 8.25   1st Qu.: 8.00  
 Median :3.000   Median :3.000   Median :1.000   Median :2.00   Median :4.000   Median : 3.000   Median :10.50   Median :11.00   Median :11.00  
 Mean   :3.223   Mean   :3.113   Mean   :1.474   Mean   :2.28   Mean   :3.579   Mean   : 5.319   Mean   :10.86   Mean   :10.71   Mean   :10.39  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00  
  guardian.y         traveltime.y    studytime.y      failures.y      schoolsup.y          famsup.y            paid.y           activities.y      
 Length:382         Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:382         Length:382         Length:382         Length:382        
 Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Median :1.000   Median :2.000   Median :0.0000   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                    Mean   :1.445   Mean   :2.039   Mean   :0.1414                                                                              
                    3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000                                                                              
                    Max.   :4.000   Max.   :4.000   Max.   :3.0000                                                                              
   higher.y          romantic.y           famrel.y       freetime.y       goout.y          Dalc.y          Walc.y         health.y       absences.y    
 Length:382         Length:382         Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000  
 Class :character   Class :character   1st Qu.:4.000   1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000  
 Mode  :character   Mode  :character   Median :4.000   Median :3.00   Median :3.000   Median :1.000   Median :2.000   Median :4.000   Median : 2.000  
                                       Mean   :3.942   Mean   :3.23   Mean   :3.118   Mean   :1.476   Mean   :2.291   Mean   :3.576   Mean   : 3.673  
                                       3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 6.000  
                                       Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :32.000  
      G1.y            G2.y            G3.y      
 Min.   : 0.00   Min.   : 5.00   Min.   : 0.00  
 1st Qu.:10.00   1st Qu.:11.00   1st Qu.:11.00  
 Median :12.00   Median :12.00   Median :13.00  
 Mean   :12.11   Mean   :12.24   Mean   :12.52  
 3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00  
 Max.   :19.00   Max.   :19.00   Max.   :19.00  
```

b. Data cleaning:

<u>Missing Data:</u>

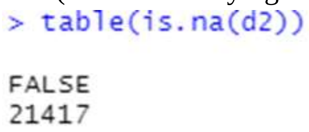
-D1 (Students Studying Mathematics)



```
> table(is.na(d1))

FALSE 
13035 
```

There is no missing data in dataset d1.

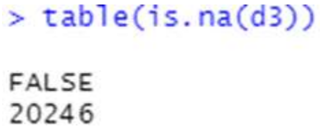
-D2 (Students Studying Portuguese)



```
> table(is.na(d2))

FALSE 
21417 
```

There is no missing data in dataset d2.

-D3 (Students Studying Both Mathematics & Portuguese)

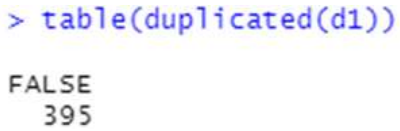


```
> table(is.na(d3))

FALSE 
20246 
```

There is no missing data in dataset d3.

Duplicated Data:

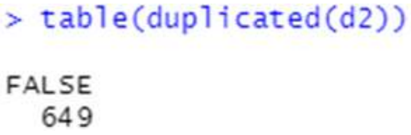
-D1 (Students Studying Mathematics)



```
> table(duplicated(d1))

FALSE 
  395 
```

There is no duplicated data in dataset d1.

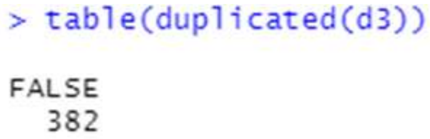
-D2 (Students Studying Portuguese)



```
> table(duplicated(d2))

FALSE 
  649 
```

There is no duplicated data in dataset d2.

-D3 (Students Studying Both Mathematics & Portuguese)



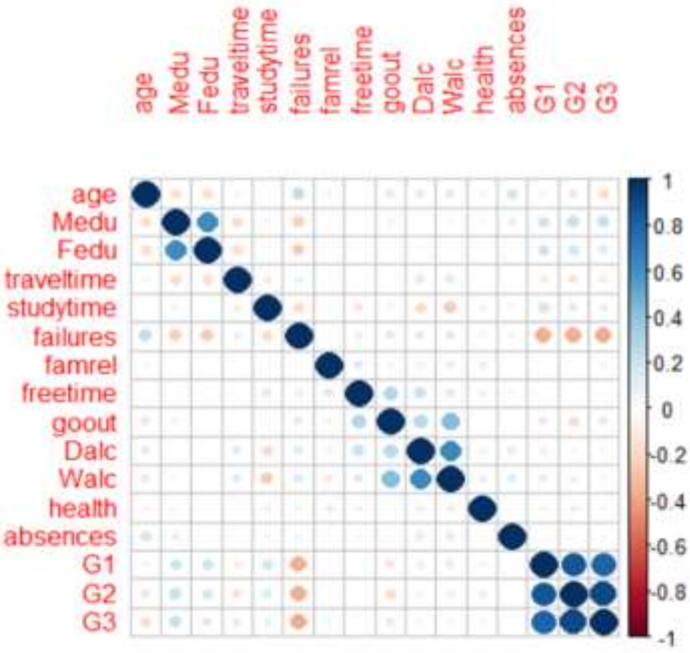
```
> table(duplicated(d3))

FALSE 
  382 
```

There is no duplicated data in dataset d3.

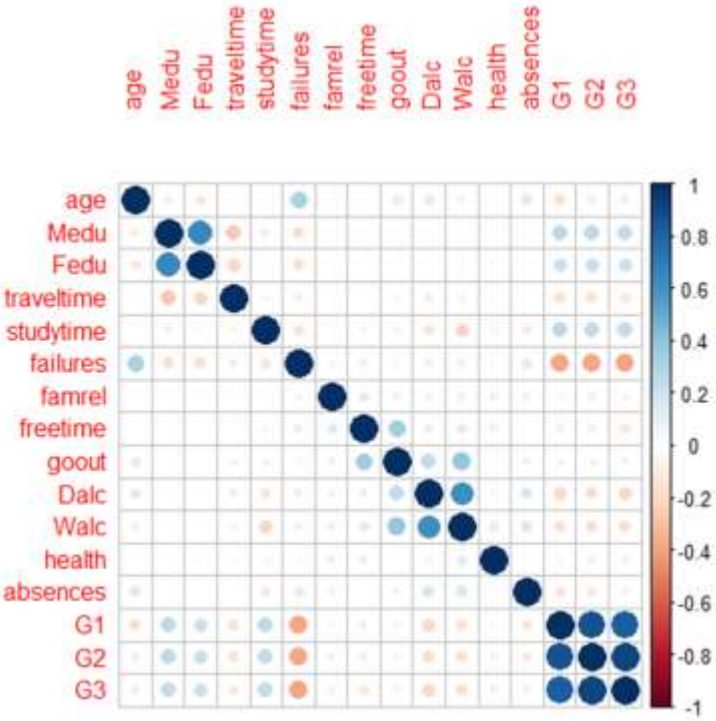
c. Correlation Heatmap:

- D1

d1.numeric<-d1[,sapply(d1, is.numeric)]
corrplot::corrplot(cor(d1.numeric))

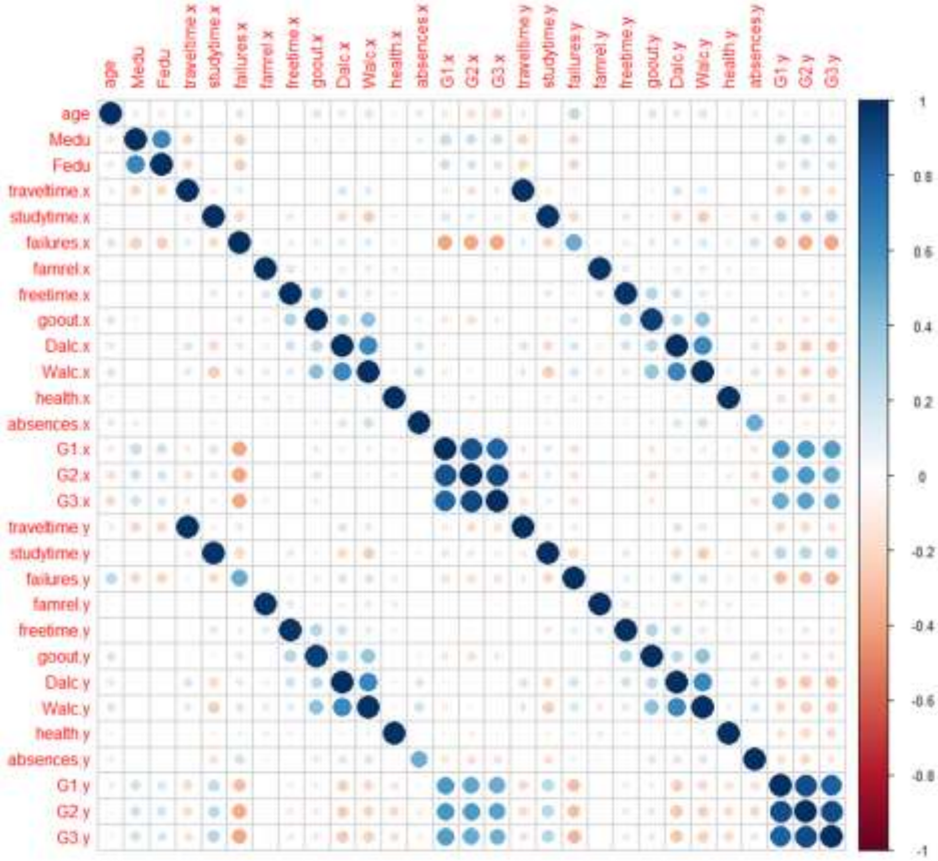


- D2
d2.numeric<-d2[,sapply(d2, is.numeric)]
corrplot::corrplot(cor(d2.numeric))



-D3
d3.numeric<-d3[,sapply(d3, is.numeric)]

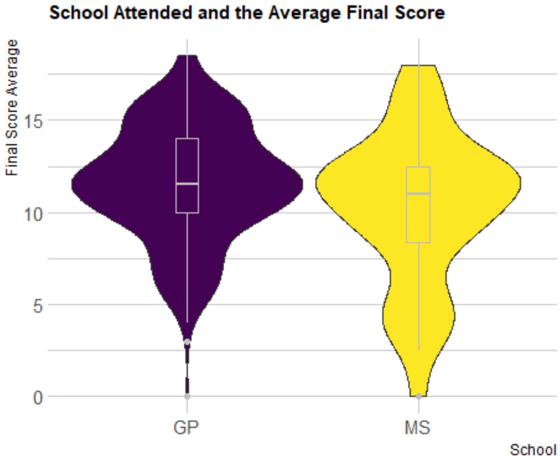
corrplot::corrplot(cor(d3.numeric))



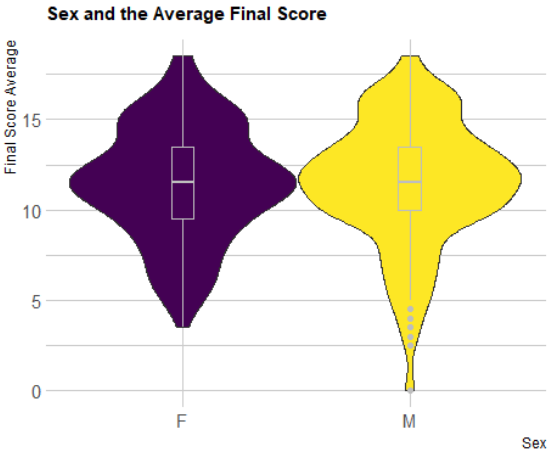
d. Categorical Data:
Using ggplot2 and dplyr packages, the groups within each variable could be visualized with boxplot wrapped by violin plot by using the following syntax:

```
sample_size = d3 %>% group_by(variable_name) %>% summarise(num=n())

d3 %>%
  left_join(sample_size) %>%
  mutate(myaxis = paste0(variable_name, "\n", "n=", num)) %>%
  ggplot( aes(x=variable_name, y=target_variable_name, fill=variable_name)) +
  geom_violin(width=1.0) +
  geom_boxplot(width=0.1, color="grey", alpha=1) +
  scale_fill_viridis(discrete = TRUE) +
  theme_ipsum() +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("...") +
  xlab(“variable_name”) + ylab(target_variable_name)
```

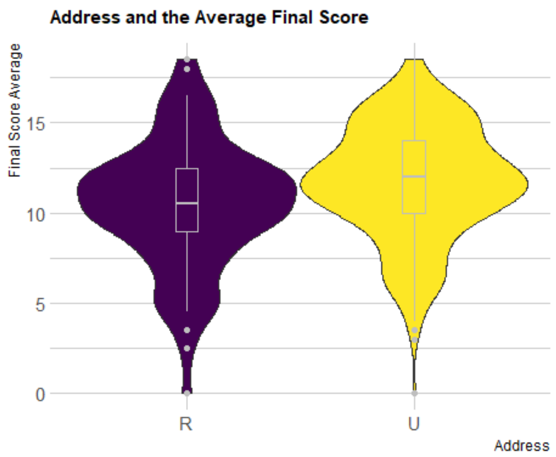
i. School (GP or MS):



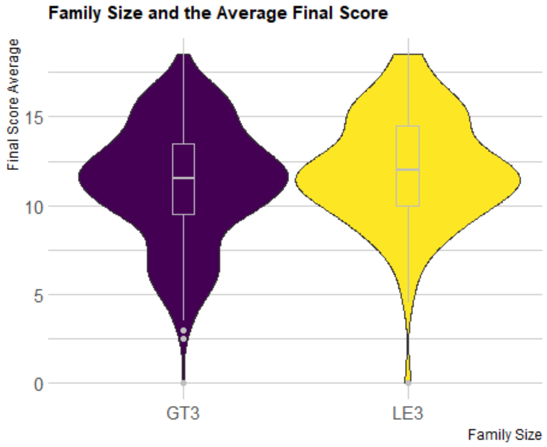
ii. Sex (Male or Female):



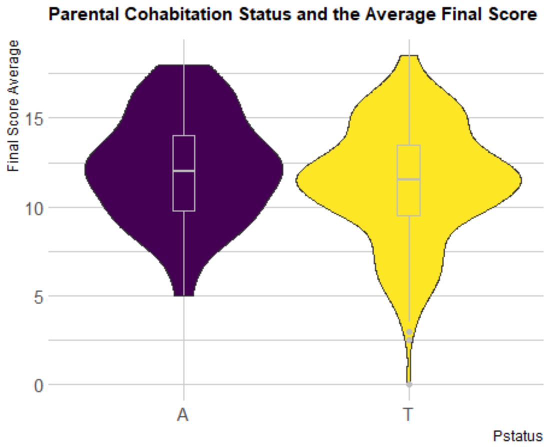
iii. Address (Urban or Rural):



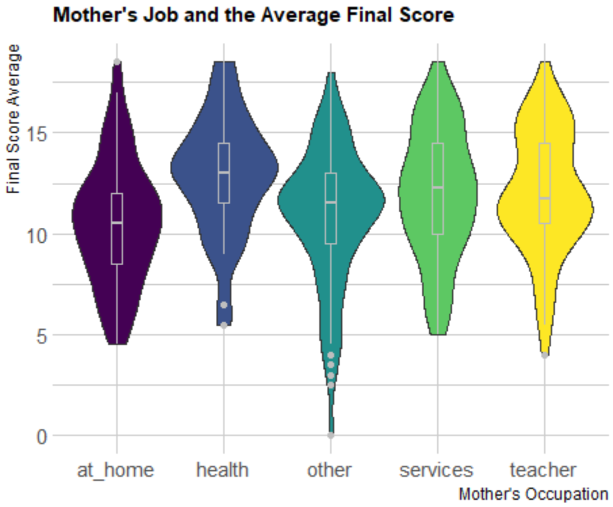
iv. Family Size (Less than 3, 3 or greater):



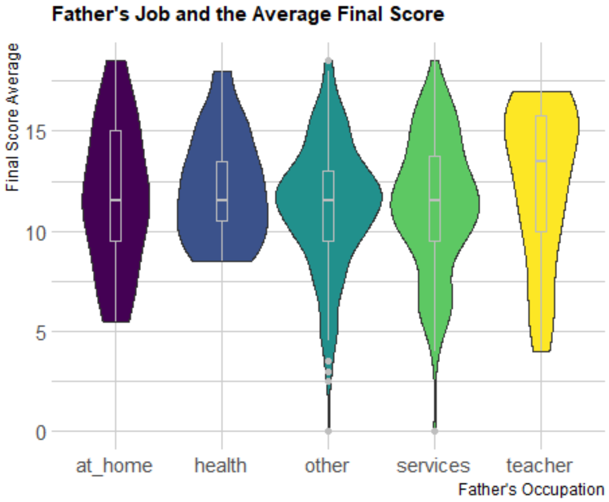
v. Parents Cohabitation Status (Apart or Together):



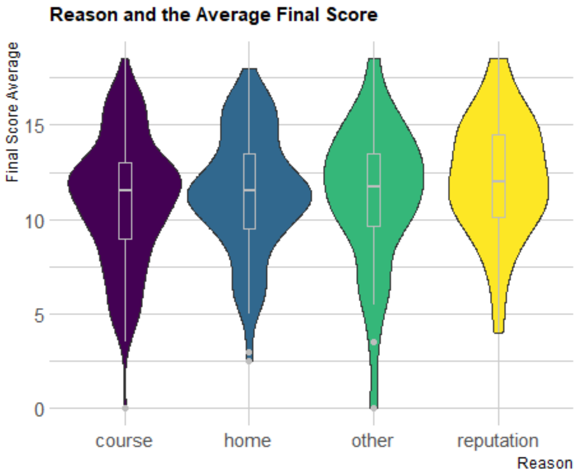
vi. Mother's Job (stay at home, healthcare, other, services, teacher):



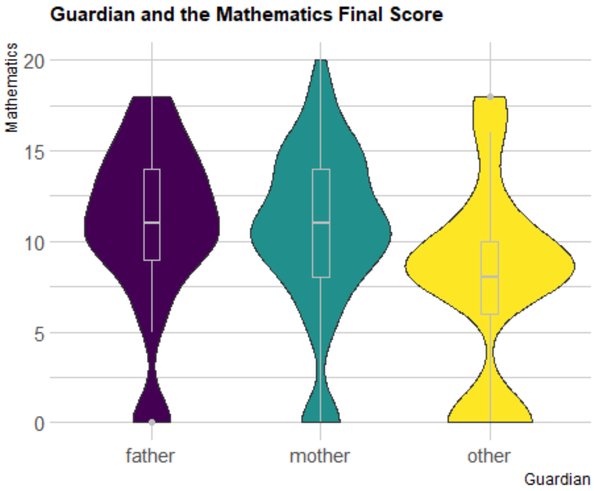
vii. Father's Job (stay at home, healthcare, other, services, teacher):

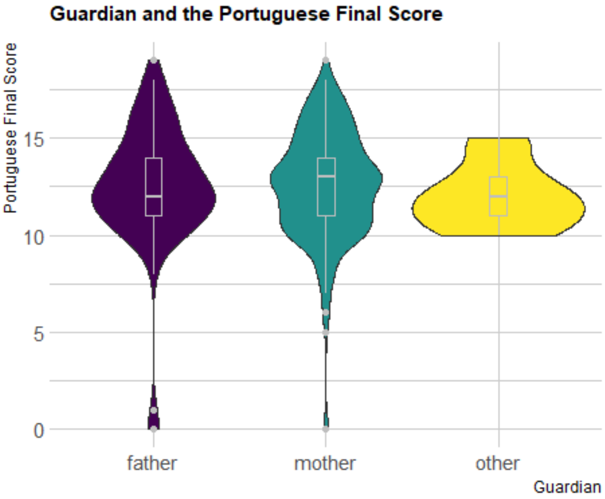


viii. Reason (course, distance from home, other, reputation):

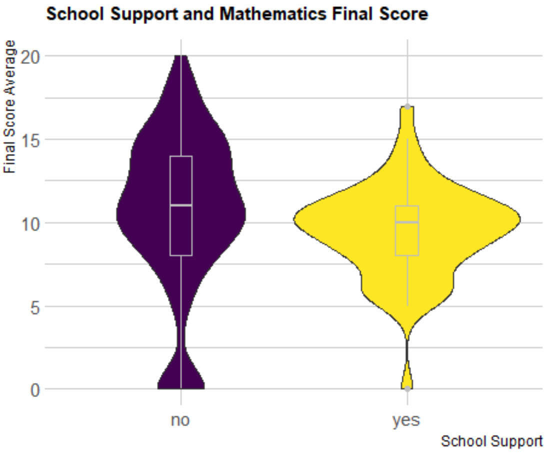


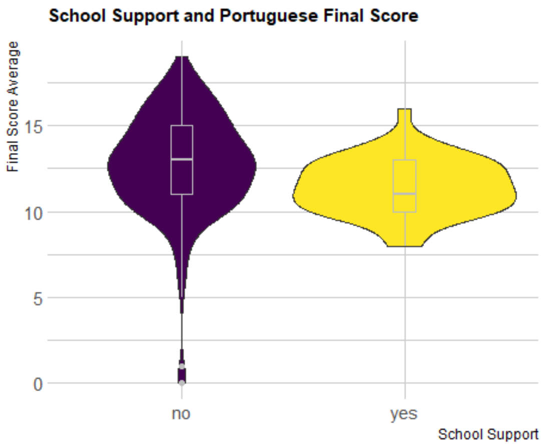
ix. Guardian (father, mother, other):



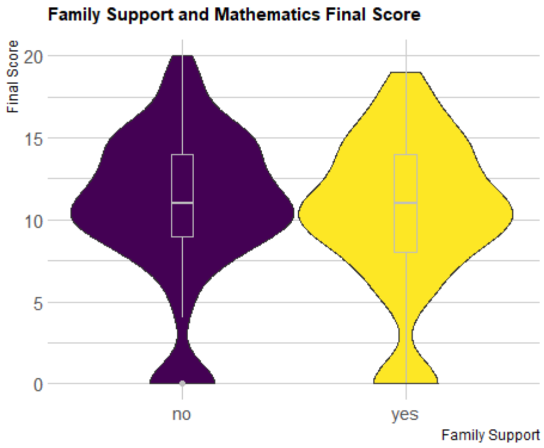


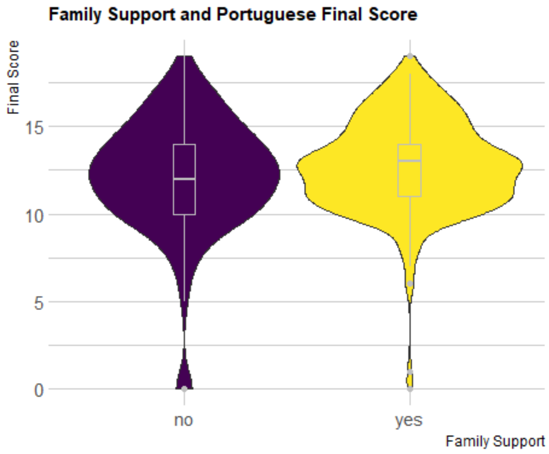
x. School Support (yes or no):



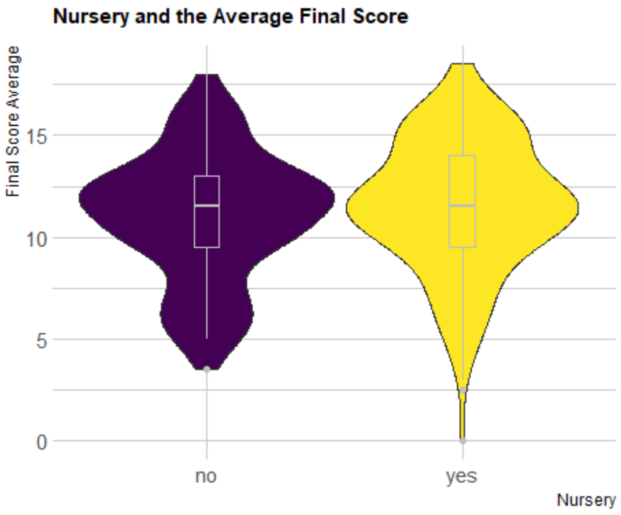


xi. Family Support (yes or no):

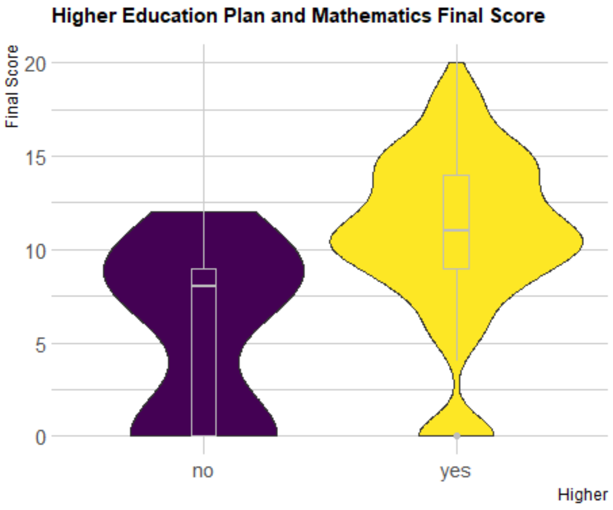


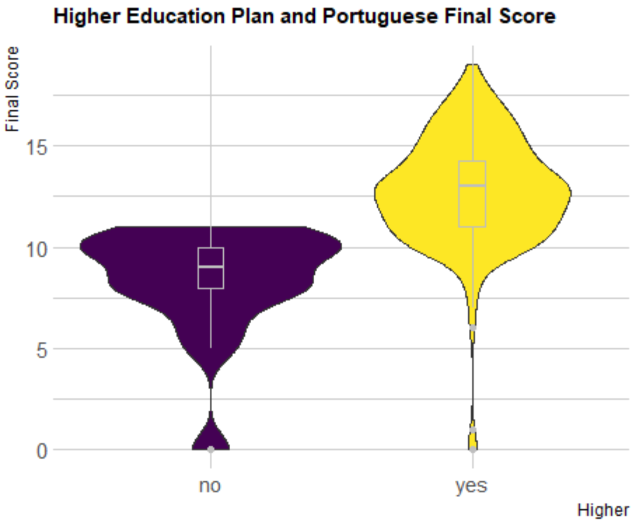


xii. Nursery (yes or no):

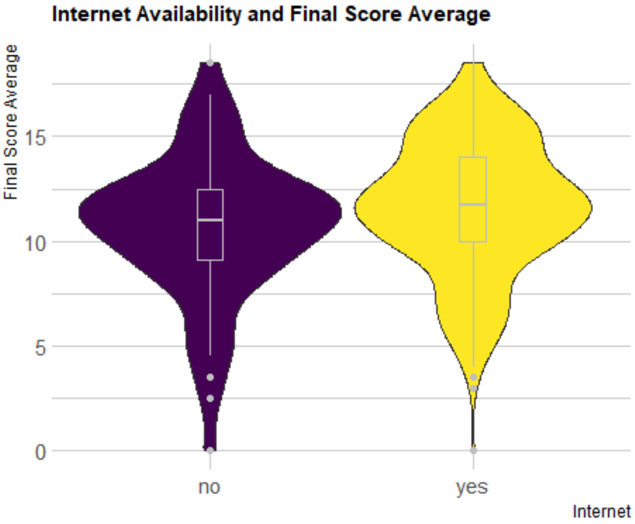


xiii. Higher Education (yes or no):

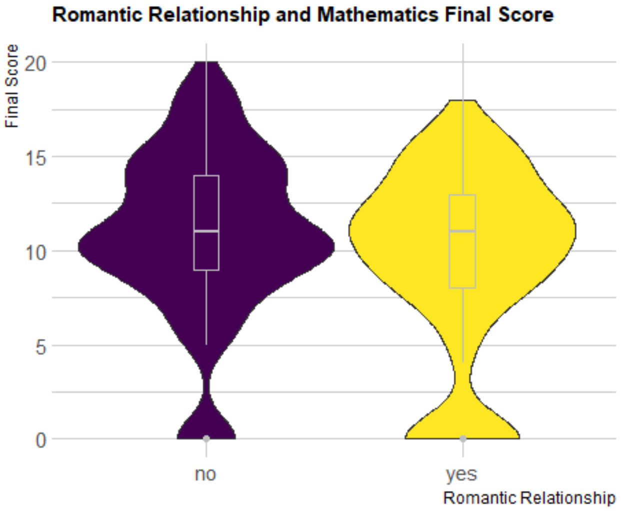


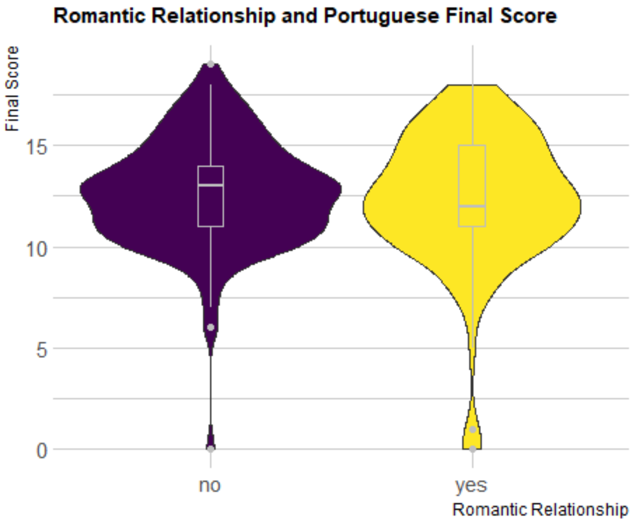


xiv. Internet Connection (yes or no):



xv. Romantic (yes or no):

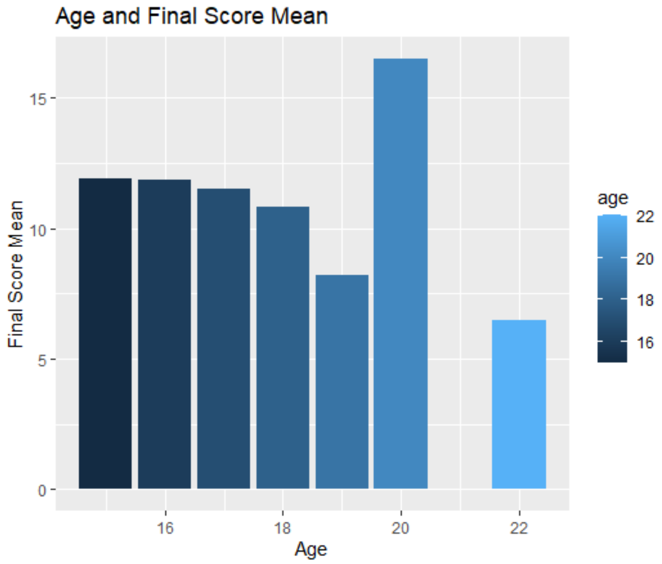




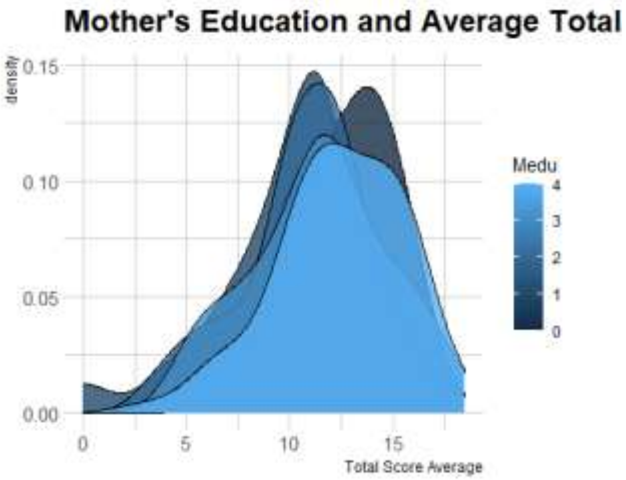
e. Numerical data:
Some variables are better visualized by using boxplots, distribution plots, or scatter plots. The code for bar plot and distribution plot respectively are:

```
d3 %>%
  group_by(variable_name) %>%
  summarise(finalScoreMean = mean(target_variable_name)) %>%
  ggplot(data = ., aes(x=variable_name, y=finalScoreMean, fill = variable_name)) +
  geom_bar(stat = 'identity') +
  ggtitle("...") +
  xlab(“variable_name”) + ylab("Final Score Mean")

d3 %>%
  ggplot(data=., aes(x=target_variable_name, group=variable_name, fill=variable_name))
+
  geom_density(adjust=1.2, alpha=0.8) +
  xlab("Total Score Average") +
  ggtitle("...") +
  theme_ipsum()
```

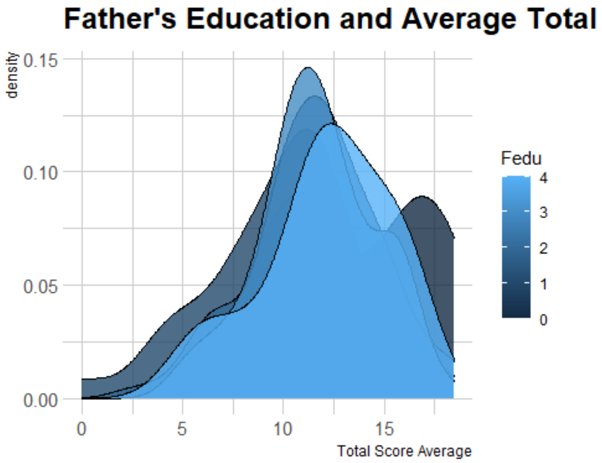
i. Age:



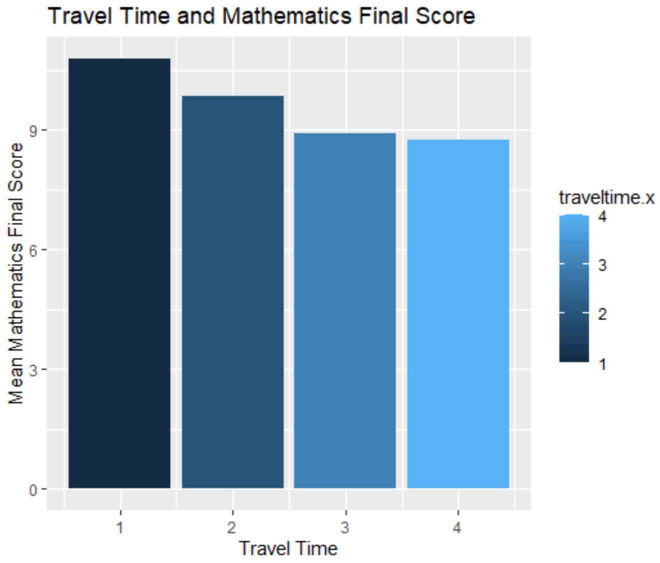
ii. Mother’s Education:

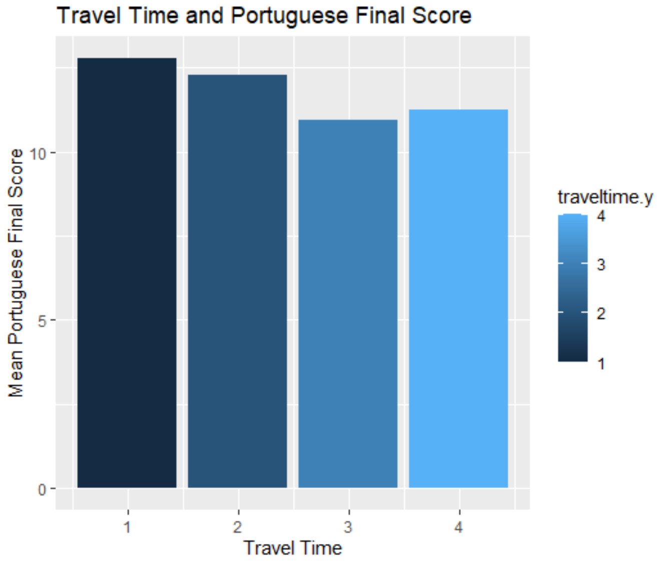


iii. Father’s Education:

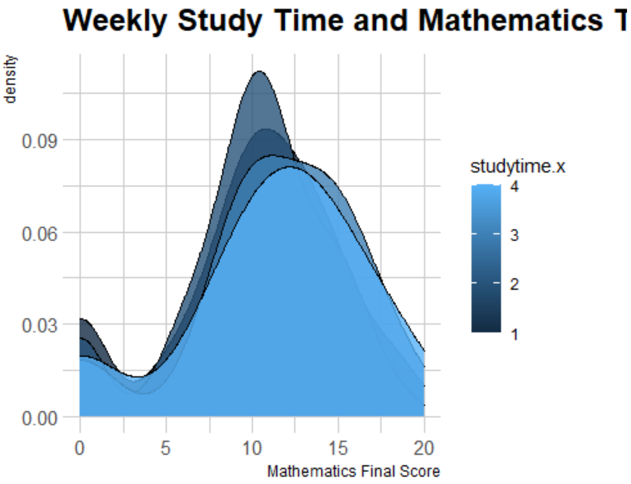


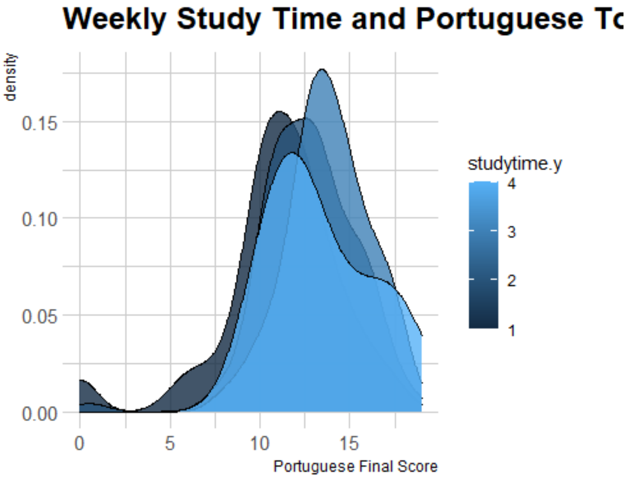
iv. Travel time:



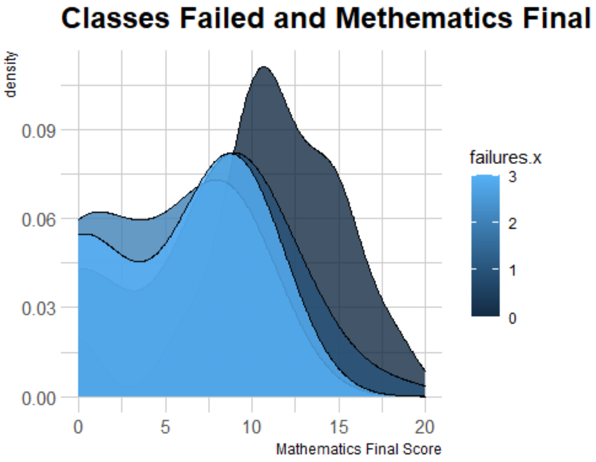


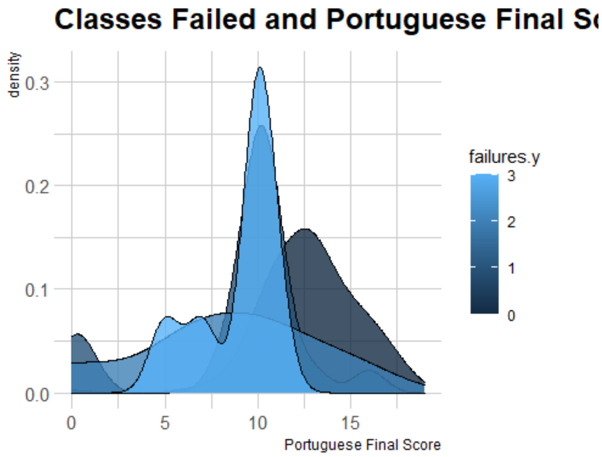
v. Study time:



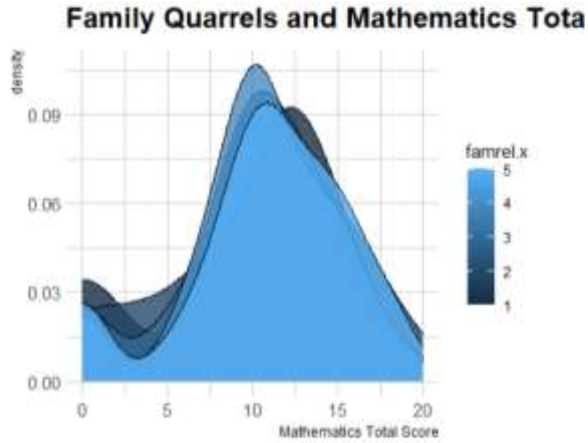


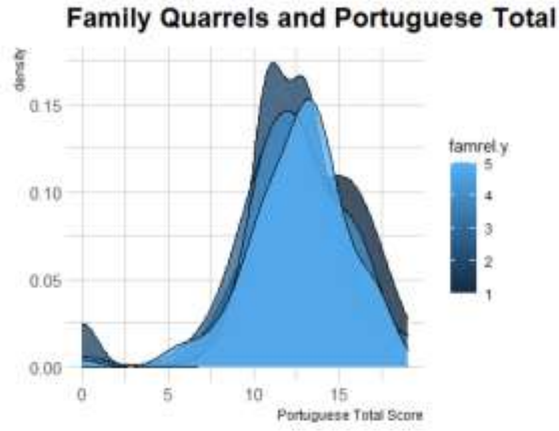
vi. Classes Failed:



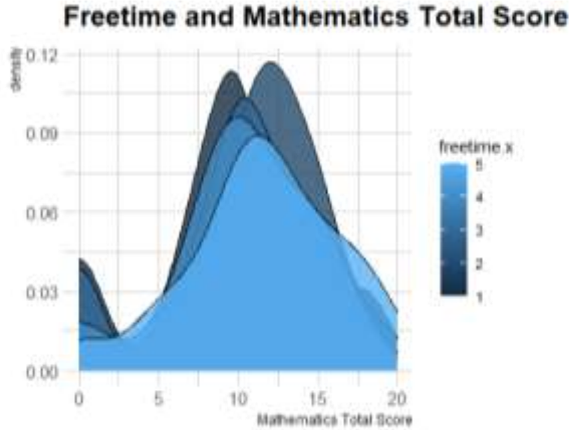


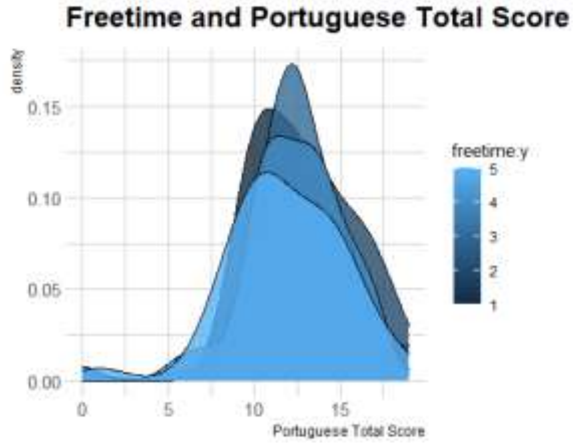
vii. Family Quarrels:



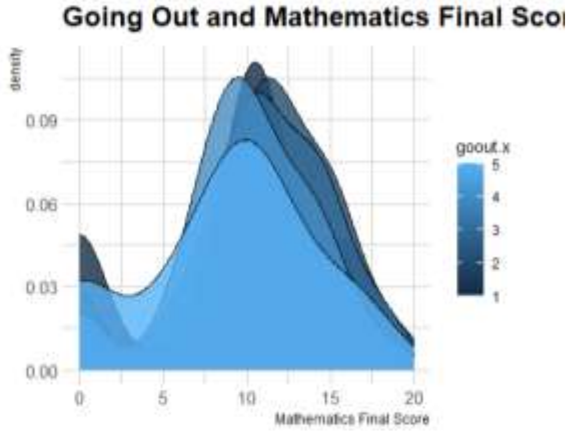


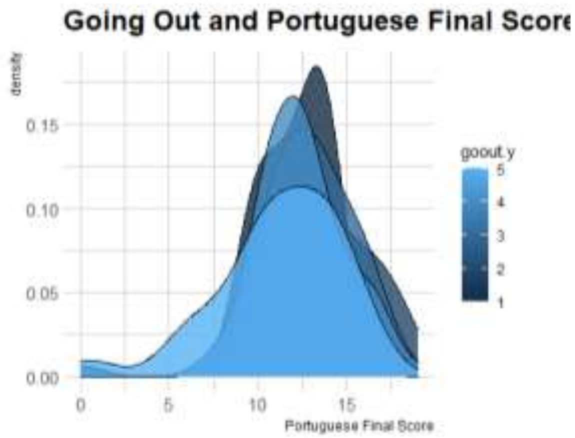
viii. Free Time:


Freetime and Mathematics Total Score

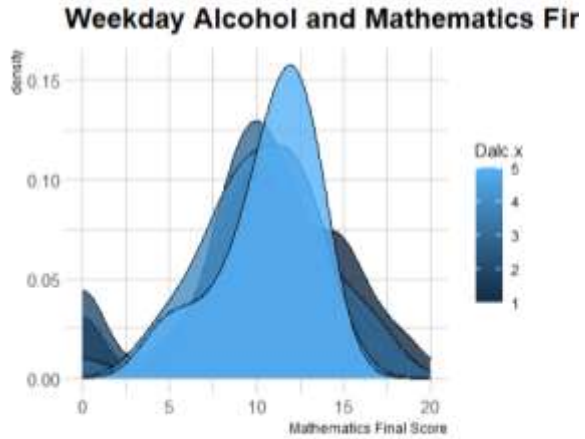

Freetime and Portuguese Total Score

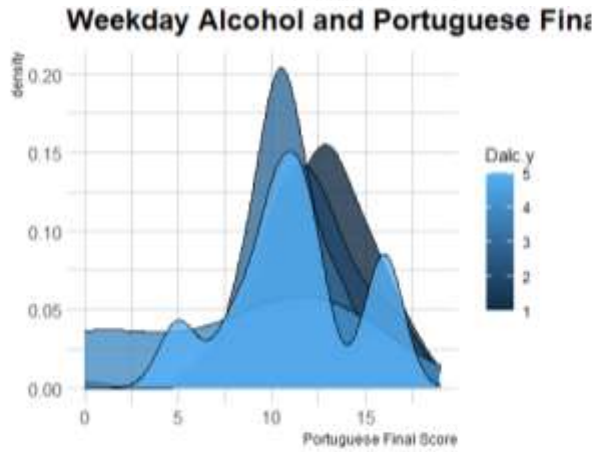
ix. Go Out:


Going Out and Mathematics Final Scoı

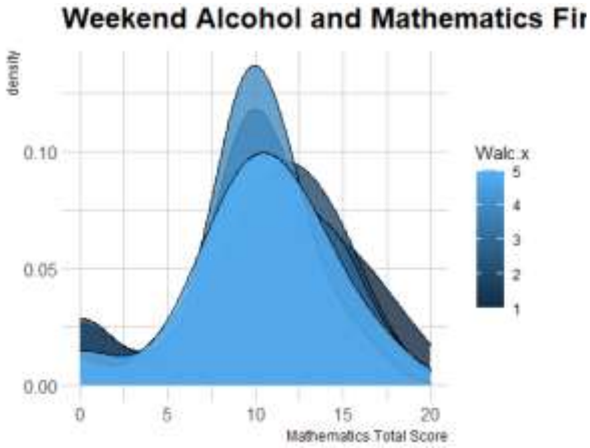


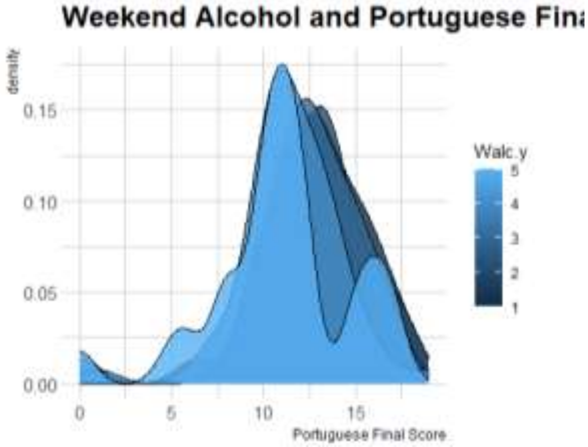
x. Weekday Alcohol Consumption:



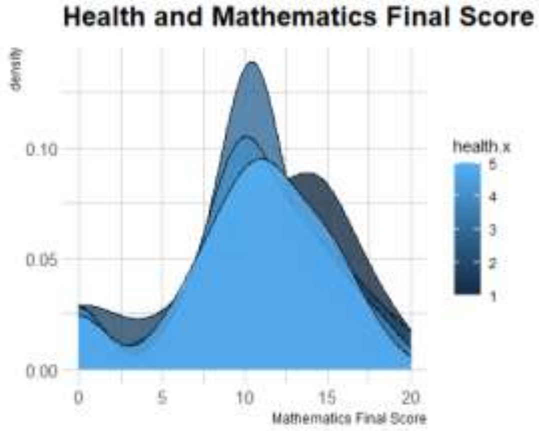


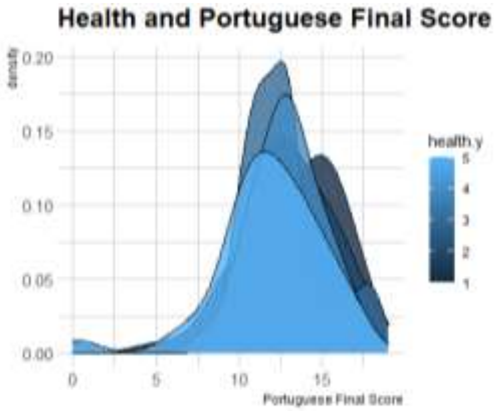
xi. Weekend Alcohol Consumption:


Weekend Alcohol and Mathematics Fir
density
0.10
0.05
0.00
0
5
10
15
20
Mathematics Total Score
Walc.x
5
4
3
2
1


Weekend Alcohol and Portuguese Fin
density
0.15
0.10
0.05
0.00
0
5
10
15
Portuguese Final Score
Walc.y
5
4
3
2
1

xii. Health:

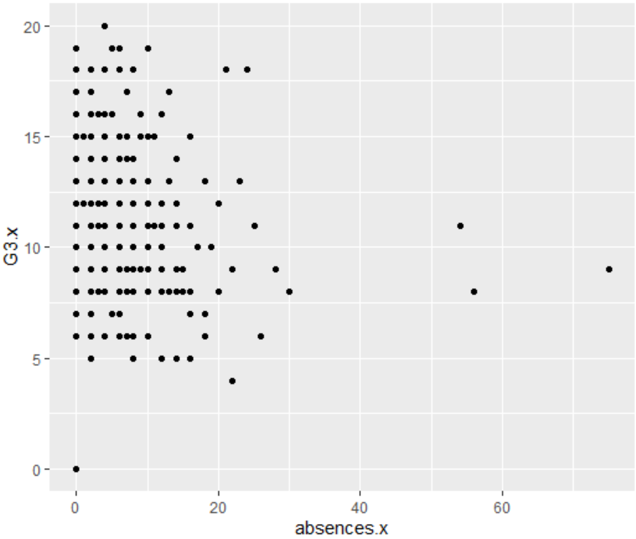

Health and Mathematics Final Score
density
0.10
0.05
0.00
0
5
10
15
20
Mathematics Final Score
health.x
5
4
3
2
1



xiii. Absences:

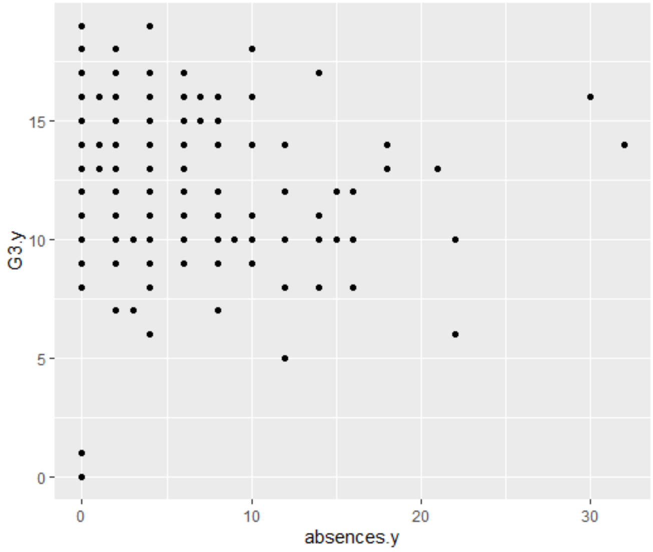
- Mathematics Score:

```
ggplot(d3_new, aes(x=absences.x, y=G3.x)) +
  geom_point()
```



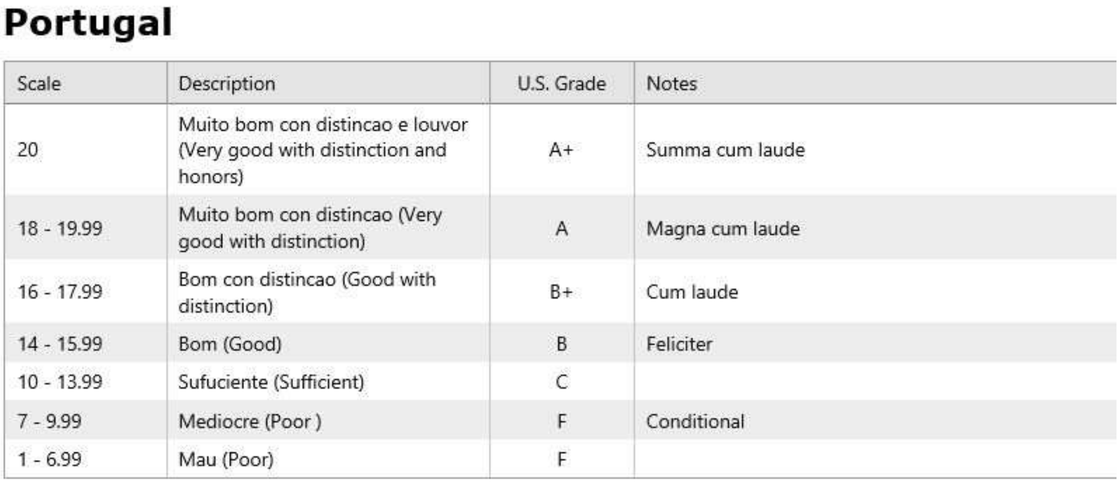
- Portuguese Score:

```
ggplot(d3_new, aes(x=absences.y, y=G3.y)) +
  geom_point()
```



f. Categorize the Final Grade (G3)

   i. The final grade from the data is split into 7 categories based on the following table

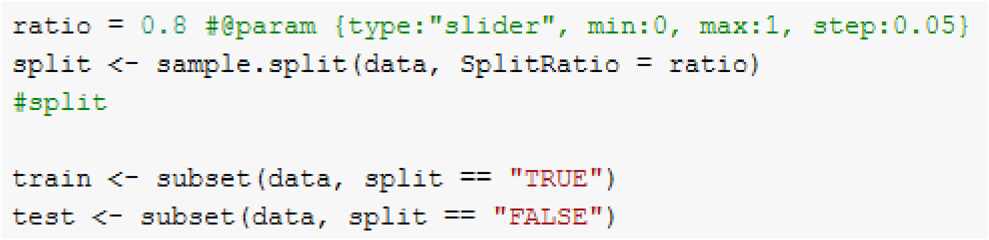


Portugal

Scale	Description	U.S. Grade	Notes
20	Muito bom con distincao e louvor (Very good with distinction and honors)	A+	Summa cum laude
18 - 19.99	Muito bom con distincao (Very good with distinction)	A	Magna cum laude
16 - 17.99	Bom con distincao (Good with distinction)	B+	Cum laude
14 - 15.99	Bom (Good)	B	Feliciter
10 - 13.99	Sufuciente (Sufficient)	C	
7 - 9.99	Mediocre (Poor)	F	Conditional
1 - 6.99	Mau (Poor)	F	

   ii. Categorizing the final grades from (1-20) to (F - A+) will help decrease the error.

g. Split the data into two for training and testing data

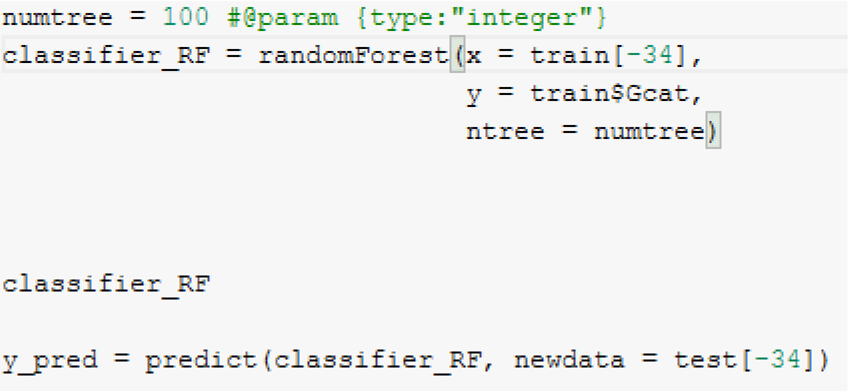


```
ratio = 0.8 #@param {type:"slider", min:0, max:1, step:0.05}
split <- sample.split(data, SplitRatio = ratio)
#split

train <- subset(data, split == "TRUE")
test <- subset(data, split == "FALSE")
```

   i. Samples from each data are split into training and testing data with 80:20 ratio

ii. The split is done using *caTools* package since it offers an easier way to split a data using a predefined ratio

h. Random Forest Classifier

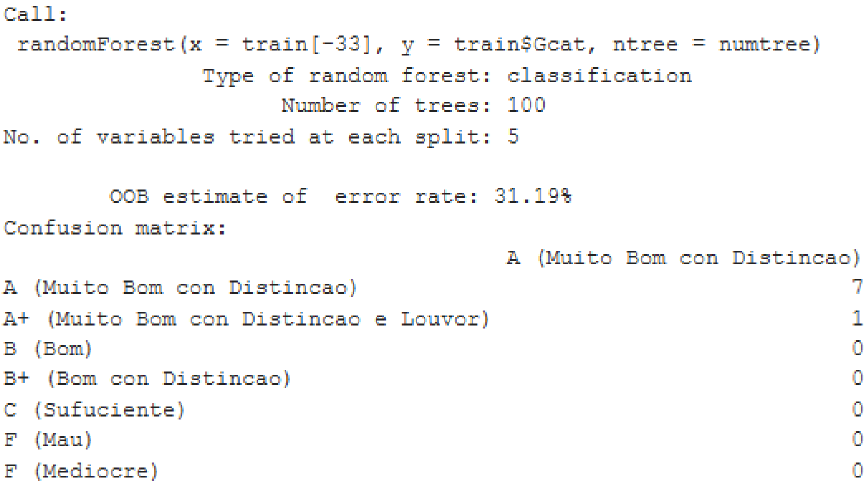


```
numtree = 100 #@param {type:"integer"}
classifier_RF = randomForest(x = train[-34],
                             y = train$Gcat,
                             ntree = numtree)



classifier_RF

y_pred = predict(classifier_RF, newdata = test[-34])
```

i. A Random Forest Classifier with 100 trees is trained using the training data and tested on the testing data

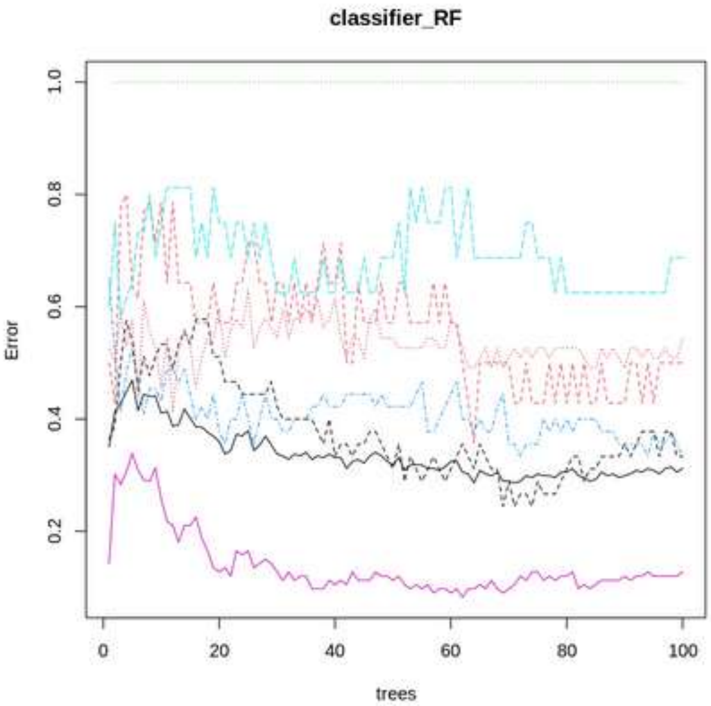
ii. The Random Forest Classifier is done using *randomForest* package

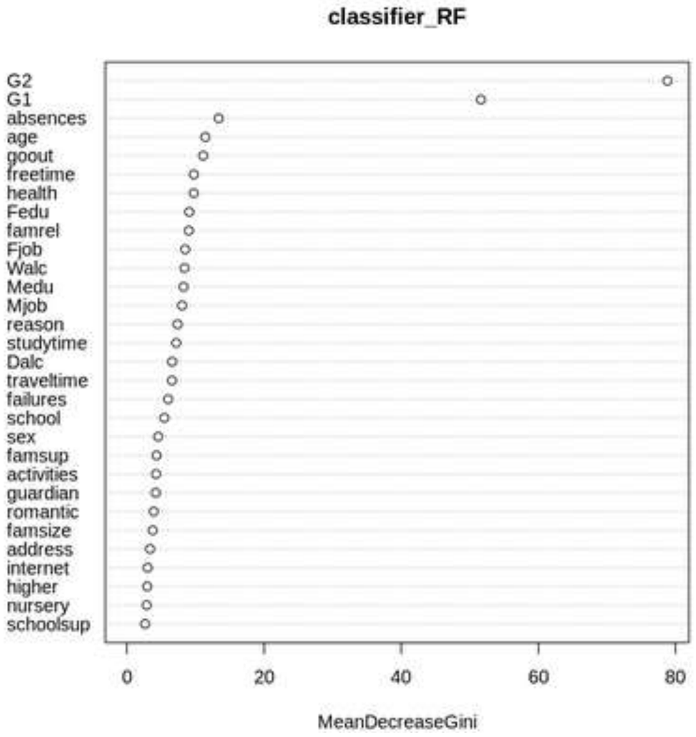
1. D1 (Math)



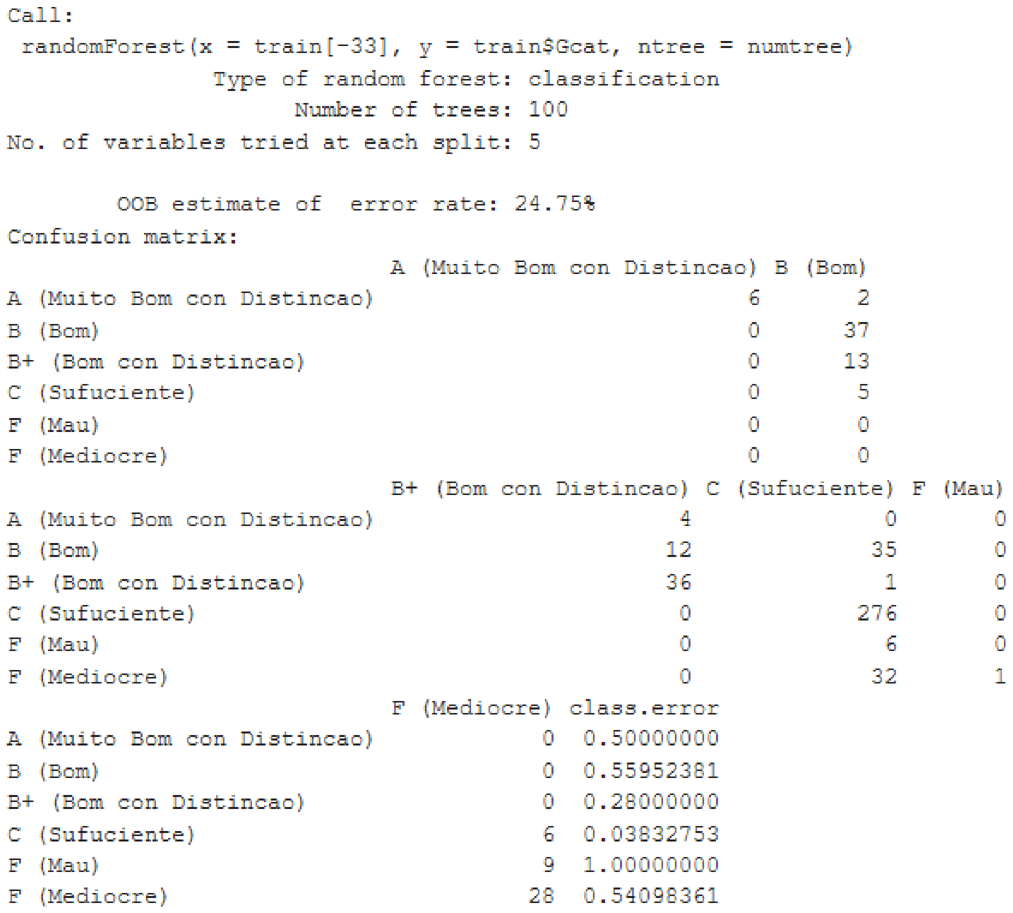
```
Call:
 randomForest(x = train[-33], y = train$Gcat, ntree = numtree) 
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 5

        OOB estimate of  error rate: 31.19%
Confusion matrix:
                                      A (Muito Bom con Distincao)
A (Muito Bom con Distincao)                                     7
A+ (Muito Bom con Distincao e Louvor)                           1
B (Bom)                                                         0
B+ (Bom con Distincao)                                          0
C (Sufuciente)                                                  0
F (Mau)                                                         0
F (Mediocre)                                                    0
```

```
A (Muito Bom con Distincao)                                               0
A+ (Muito Bom con Distincao e Louvor)                                     0
B (Bom)                                                                   0
B+ (Bom con Distincao)                                                    0
C (Sufuciente)                                                            0
F (Mau)                                                                   0
F (Mediocre)                                                              0
                                      B (Bom) B+ (Bom con Distincao)
A (Muito Bom con Distincao)                 7                      0
A+ (Muito Bom con Distincao e Louvor)       0                      0
B (Bom)                                    30                      2
B+ (Bom con Distincao)                     11                      5
C (Sufuciente)                              5                      0
F (Mau)                                     0                      0
F (Mediocre)                                0                      0
                                      C (Sufuciente) F (Mau) F (Mediocre)
A (Muito Bom con Distincao)                        0       0            0
A+ (Muito Bom con Distincao e Louvor)              0       0            0
B (Bom)                                           13       0            0
B+ (Bom con Distincao)                             0       0            0
C (Sufuciente)                                   116       0           12
F (Mau)                                            3      30           12
F (Mediocre)                                      18      13           26
                                      class.error
A (Muito Bom con Distincao)             0.5000000
A+ (Muito Bom con Distincao e Louvor)   1.0000000
B (Bom)                                 0.3333333
B+ (Bom con Distincao)                  0.6875000
C (Sufuciente)                          0.1278195
F (Mau)                                 0.3333333
F (Mediocre)                            0.5438596
```

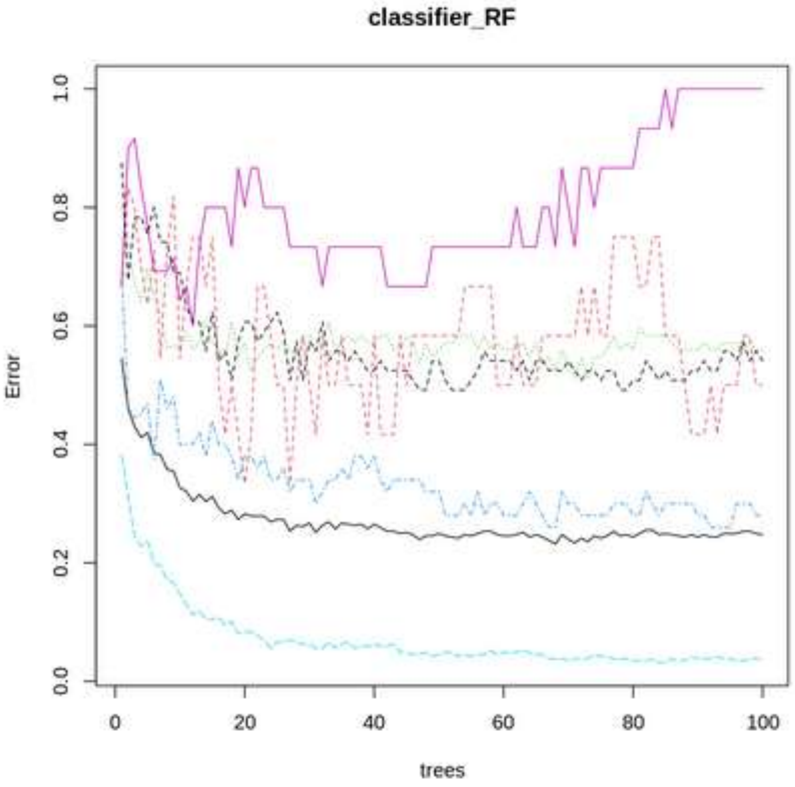


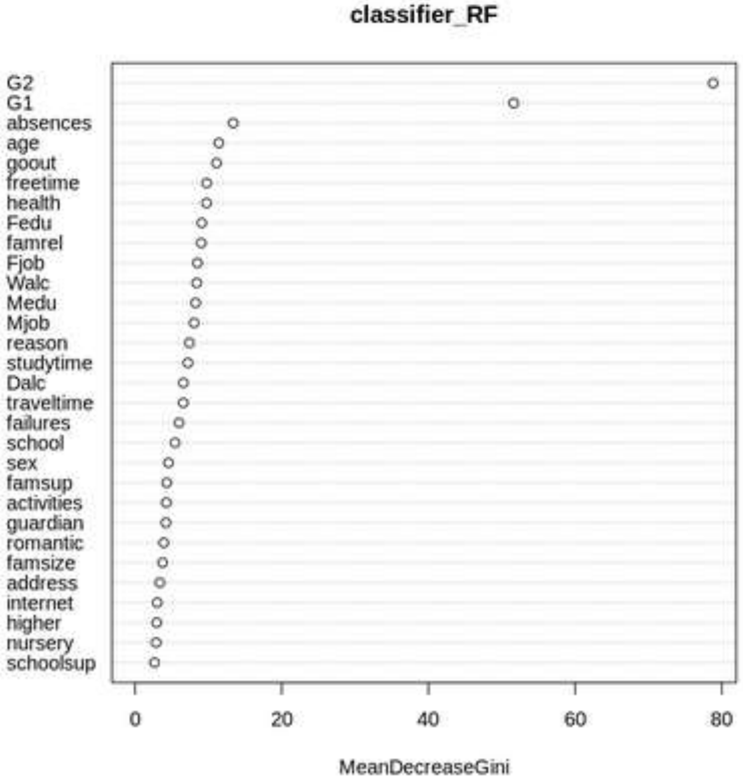


2. D2 (Portuguese)

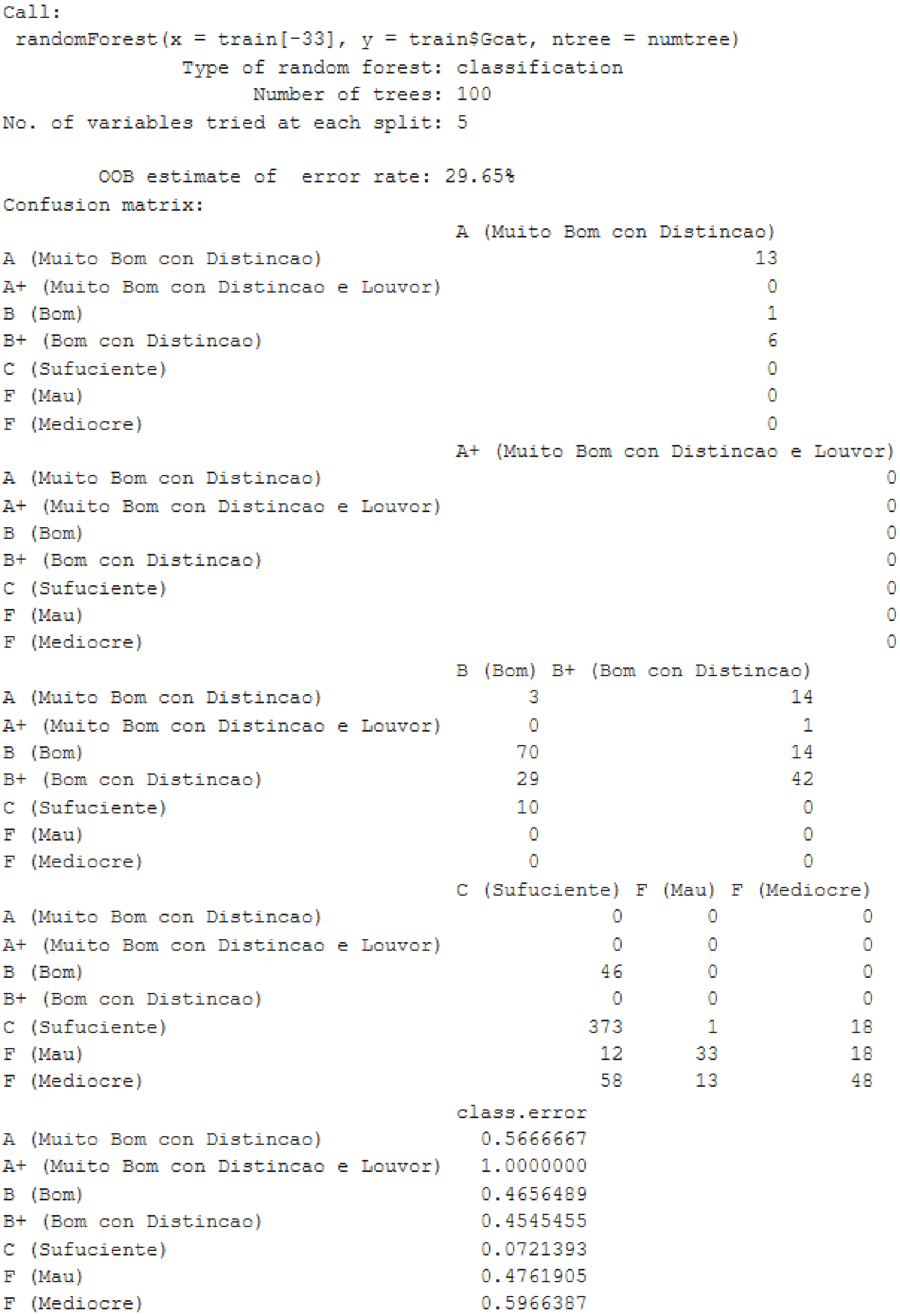


```
Call:
 randomForest(x = train[-33], y = train$Gcat, ntree = numtree) 
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 5

        OOB estimate of  error rate: 24.75%
Confusion matrix:
                            A (Muito Bom con Distincao) B (Bom)
A (Muito Bom con Distincao)                           6       2
B (Bom)                                               0      37
B+ (Bom con Distincao)                                0      13
C (Sufuciente)                                        0       5
F (Mau)                                               0       0
F (Mediocre)                                          0       0
                            B+ (Bom con Distincao) C (Sufuciente) F (Mau)
A (Muito Bom con Distincao)                      4              0       0
B (Bom)                                         12             35       0
B+ (Bom con Distincao)                          36              1       0
C (Sufuciente)                                   0            276       0
F (Mau)                                          0              6       0
F (Mediocre)                                     0             32       1
                            F (Mediocre) class.error
A (Muito Bom con Distincao)            0  0.50000000
B (Bom)                                0  0.55952381
B+ (Bom con Distincao)                 0  0.28000000
C (Sufuciente)                         6  0.03832753
F (Mau)                                9  1.00000000
F (Mediocre)                          28  0.54098361
```

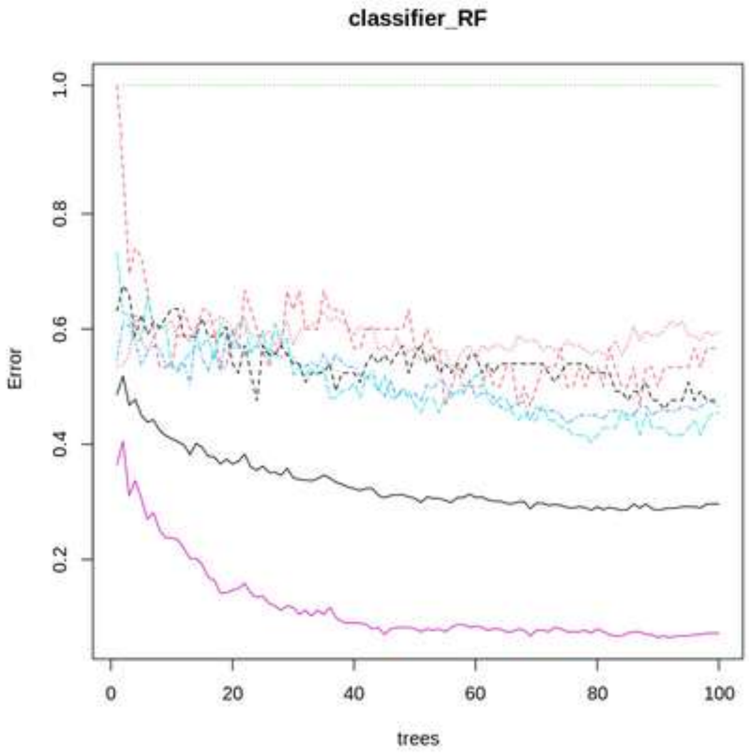

classifier_RF
Error
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
trees

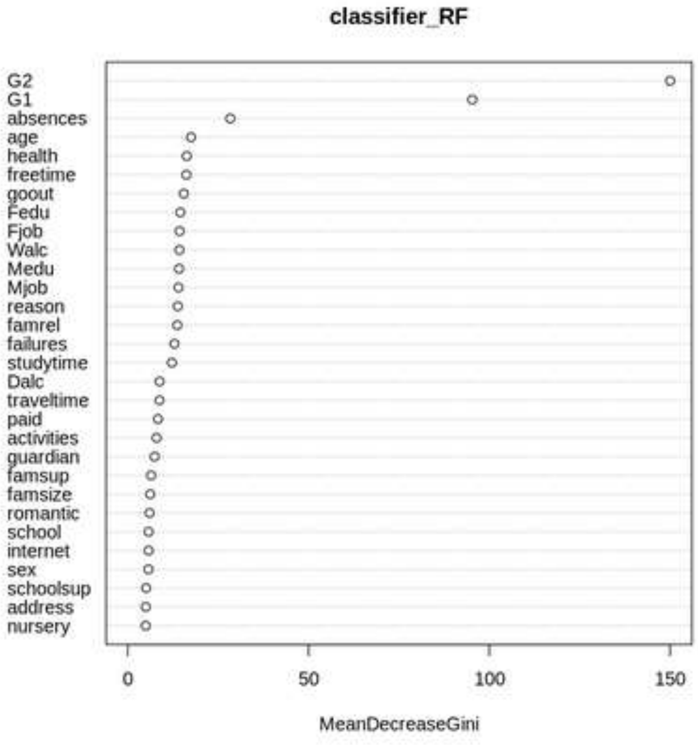

classifier_RF
G2
G1
absences
age
goout
freetime
health
Fedu
famrel
Fjob
Walc
Medu
Mjob
reason
studytime
Dalc
traveltime
failures
school
sex
famsup
activities
guardian
romantic
famsize
address
internet
higher
nursery
schoolsup
0
20
40
60
80
MeanDecreaseGini

3. D3 (Both)



```
Call:
 randomForest(x = train[-33], y = train$Gcat, ntree = numtree)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 5

        OOB estimate of  error rate: 29.65%
Confusion matrix:
                                      A (Muito Bom con Distincao)
A (Muito Bom con Distincao)                                    13
A+ (Muito Bom con Distincao e Louvor)                           0
B (Bom)                                                         1
B+ (Bom con Distincao)                                          6
C (Sufuciente)                                                  0
F (Mau)                                                         0
F (Mediocre)                                                    0
                                      A+ (Muito Bom con Distincao e Louvor)
A (Muito Bom con Distincao)                                               0
A+ (Muito Bom con Distincao e Louvor)                                     0
B (Bom)                                                                   0
B+ (Bom con Distincao)                                                    0
C (Sufuciente)                                                            0
F (Mau)                                                                   0
F (Mediocre)                                                              0
                                      B (Bom) B+ (Bom con Distincao)
A (Muito Bom con Distincao)                 3                     14
A+ (Muito Bom con Distincao e Louvor)       0                      1
B (Bom)                                    70                     14
B+ (Bom con Distincao)                     29                     42
C (Sufuciente)                             10                      0
F (Mau)                                     0                      0
F (Mediocre)                                0                      0
                                      C (Sufuciente) F (Mau) F (Mediocre)
A (Muito Bom con Distincao)                        0       0            0
A+ (Muito Bom con Distincao e Louvor)              0       0            0
B (Bom)                                           46       0            0
B+ (Bom con Distincao)                             0       0            0
C (Sufuciente)                                   373       1           18
F (Mau)                                           12      33           18
F (Mediocre)                                      58      13           48
                                      class.error
A (Muito Bom con Distincao)             0.5666667
A+ (Muito Bom con Distincao e Louvor)   1.0000000
B (Bom)                                 0.4656489
B+ (Bom con Distincao)                  0.4545455
C (Sufuciente)                          0.0721393
F (Mau)                                 0.4761905
F (Mediocre)                            0.5966387
```


classifier_RF
Error
1.0
0.8
0.6
0.4
0.2
0
20
40
60
80
100
trees


classifier_RF
G2
G1
absences
age
health
freetime
goout
Fedu
Fjob
Walc
Medu
Mjob
reason
famrel
failures
studytime
Dalc
traveltime
paid
activities
guardian
famsup
famsize
romantic
school
internet
sex
schoolsup
address
nursery
0
50
100
150
MeanDecreaseGini

6. Discussion:

The dataset contains a large number of columns (30 variables) in each of d1 and d2, before being merged into d3. At a glance with the correlation heatmap, the majority of the numerical variables (not including G1, G2, and other variables with Mathematics and Portuguese counterparts) do not have a strong relationship with G3, the final score. Upon visualizing the categorical data and their distribution of final scores, it would be tempting to draw rough conclusions about their relationship to G3 or their respective subject's final scores. For example, it would have been easy to conclude that students that are pursuing a higher degree would significantly outperform those who do not. However, it is also necessary to consider their sizes: students who are not pursuing a higher degree are drastically outnumbered by those who are pursuing one. Therefore, it would have been fatal to directly hypothesize that "higher" is a strong determining factor of their academic performance.

Additionally, it is also against the presumption that "failures" would have higher importance to the learning model than "absences". Looking at the correlation heatmap above, it is visible that "failures" have a more significant relationship to the final score than "absences". Contrastingly, the mean decrease gini reports that "absences" hold higher importance to the random forest classification model than "failures" since "absences" have a higher mean decrease of gini.

An improvement that can be done to further reduce the errors is to get a bigger sample than 382 students. With the random forest classification model, it was possible to develop a model with a low error rate of 3.38% (a high number of errors for the A+ group was detected since no student received a perfect score). The confusion matrix also shows the learning model was sufficient in correctly labeling an element. It is not always necessary to have a large sample in order to build a classification model with high accuracy and low rate of error. However in this case, the majority of the variables show weak relationship with the target variable (final score) thus, a larger sample size would be beneficial to further study the relationship between the variables. In addition, it would also be helpful to do feature engineering and not utilize variables with weak correlation with the target variable and choose the ones with strong correlations instead.

7. Conclusion:

It is obvious that the final results of the students are directly correlated to their first class score (G1) and their second class score (G2) since they make up the final score (G3). Aside from those variables "absences", "age", "health", "freetime", and "goout" can form a great explanation in order to explain students' final scores. It suggests that "absences" shows students' learning commitment, and those who spend more time studying in class would generally perform better than those with a higher number of class absences. It closely supports "freetime" and "goout", which suggests that students with more free time tend to go out more and not focus on their studies as much. The variable "age" may suggest that older students place high priority to pass their secondary education in order to attend higher education and start their career compared to their younger peers.

Living conditions at home also may affect students' academic performance such that their parents' education (Fedu and Medu) and occupations (Fjob and Mjob) may support student's learning i.e. parents with higher level of education will have higher expectations of their children with higher knowledge resources as well, and students with teachers as parents also have higher knowledge resources. However, if their parents have a lot of quarrels (famrel), the students would endure more stress and focus less on their studies. Furthermore, "health" is also an important feature to consider since generally speaking, students with illnesses are preoccupied with medical related activities and healthier students are more well suited to attend classes and focus on their education.

There really is not one singular factor that mainly affects the performance of secondary students in Portugal. Using the random forest classification, it suggests that students perform better when they regularly attend their classes and put more effort into their studies. This emphasizes the points that students who put enough time to study would be more likely to succeed. Other factors that may support this are their age, health, amount of free time, how often they go out, parents' educational background and jobs. This will lead them to perform well in their first class test (G1) and second class test (G2), which will give them a high overall test result (G3).

8. Resources:

The dataset used: https://archive.ics.uci.edu/ml/datasets/Student+Performance