Big data: - Final project

Submission by: מוסא תחאוך

Id: 311590707

Introduction

The subject of this project is investigation of a possible correlation between student performance in Portuguese language classes and the student's socioeconomic status.

Dataset:
http://archive.ics.uci.edu/ml/datasets/Student+Performance

The dataset is comprised of 649(columns) different instances, each representing a student studying Portuguese language classes.
Each student has 33 attributes (rows) which are the included in a separate file (student.txt).

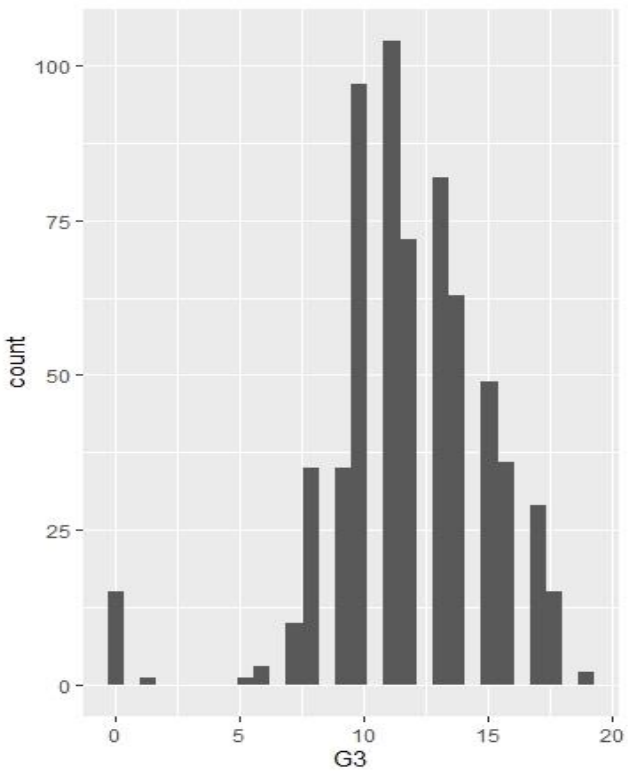
General description of procedural plan:
Data processing, data exploration, data distribution, find any correlations and use found correlations for
linear regression test to see if there is any sample of the data that shows effects on student's performance, and if there is, how strong the relations are between the two.
This can be extended to a multivariate linear regression and multiple linear regression
Possibility of using statistical inference to deduct any form of correlations
Furthermore, try to plot and draw out and find any correlations between columns or find any trending in data (groups of columns that correlate to the performance of students), find the most influencing variables and see if, with the given information, we can deduct the performance of any student to see if our conclusions from the data we analysed can be an indication to student's performances or not.

*Do note that most of the data processing and exploration will be mainly in the R script and the report will mainly include basic plots and main data processing results from the R script itself. (meaning not all plotting and results will be included in the report, only some as means of demonstration)

Data processing/exploration:

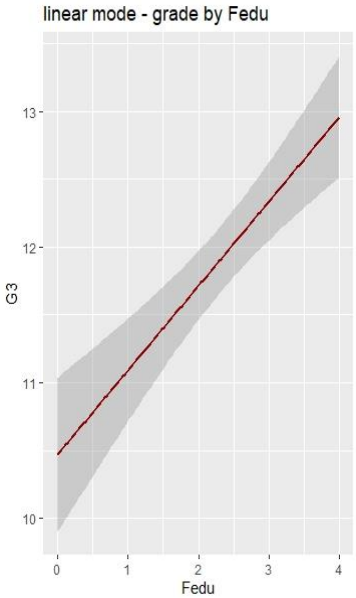
We'll start by simply filtering out columns that we'd consider irrelevant or redundant, which will be "failures", "reason" and both G1 and G2.
G3 is the consequence result of both G1 and G2 so we don't want them.
Failures are redundant as they don't represent each class and are too strongly dependent on G3, meaning a G3 of failure will be accounted for in Failures.
"reason" is also redundant as we don't care about the options present in there.

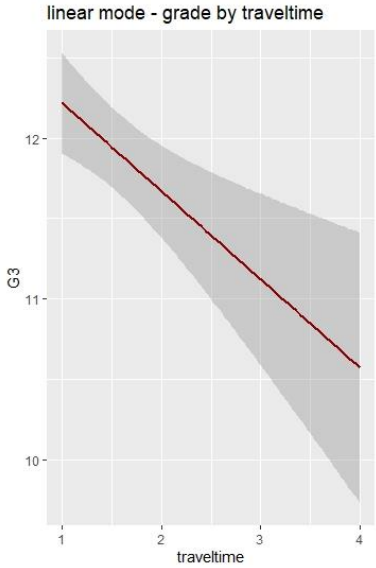
A quick exploration shows us the following.

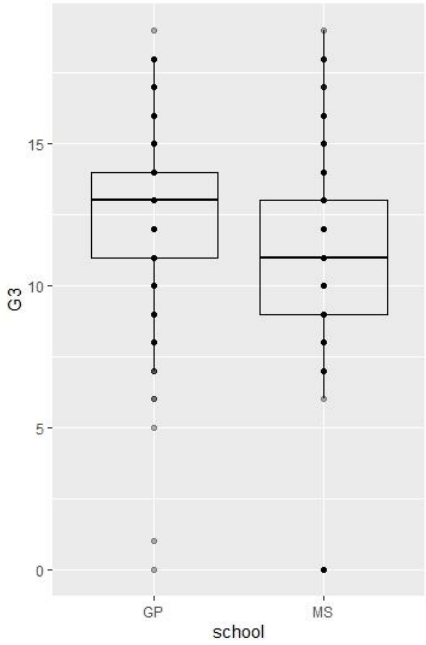


Distribution of grades, most students seem to hover a bit above 10.

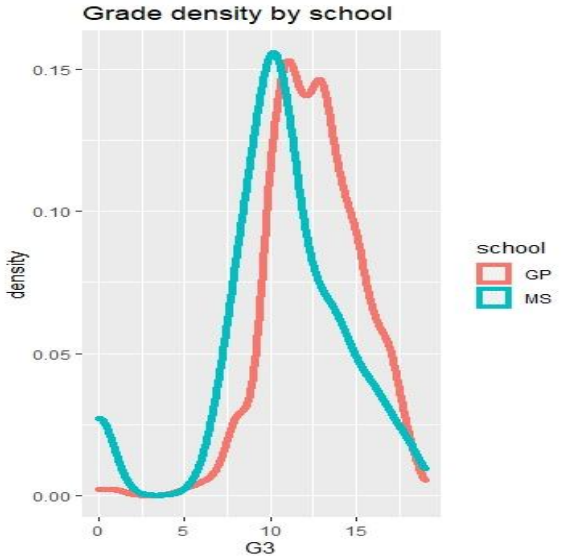
Checking the linear models of grades by Medu,Fedu,famrel,traveltime and studytime
We can see the graph following what would be logical to us.

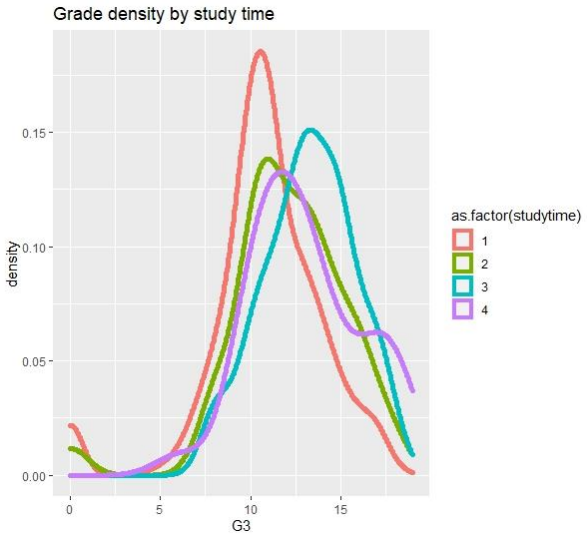






We can see in the boxplot that GP school seems to have abit of a higher scoring than MS school.





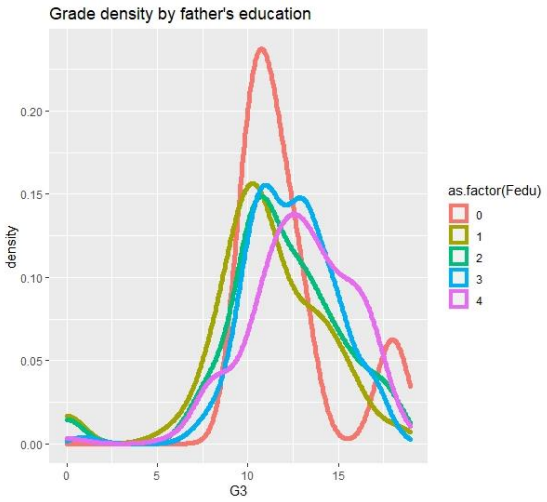
Grade density by school:

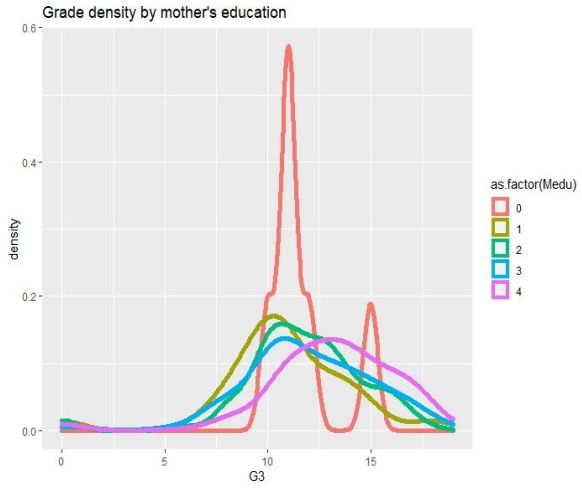
As the boxplot above, we can see in this density graphing of grades by schools that GP school seems to have more students scoring around 12 than MS school.

Grade density by study time:

Most students study around 1 hour and seem to average a score around 10

students who study 2 and 4 hours seems to score the same, except of a minor part of those who study 4 hours and those who study 3 hours seem consistent in their scoring and seem to be aligned between those who study 1 hours (who are below the 3 hour studiers) and the minority of 4 hour studiers who score the highest.

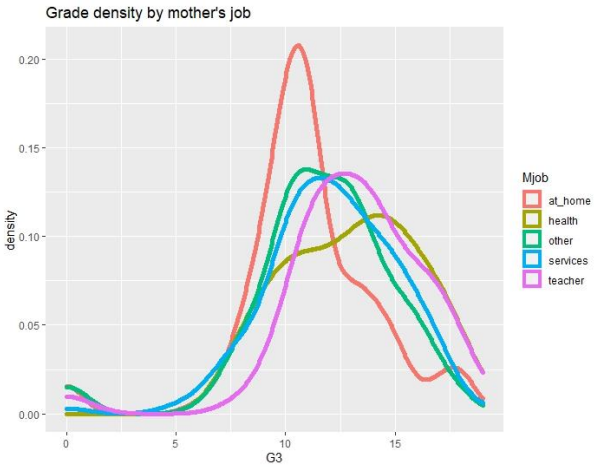


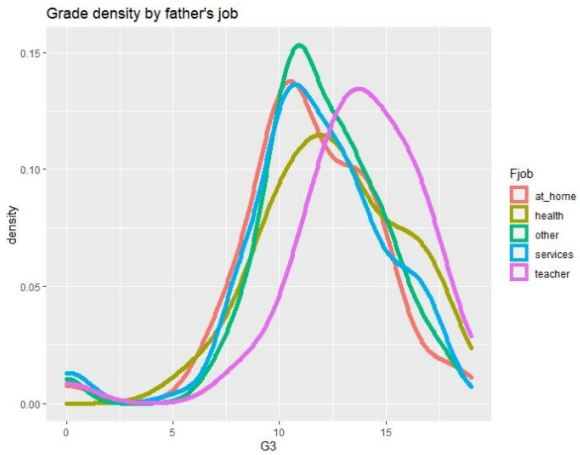


Parents education:

Note the bump at score 15 in grades by mother's education, it seems a high number of students with mothers with the lowest education, score abnormally high, maybe it is correlated to them being at home?

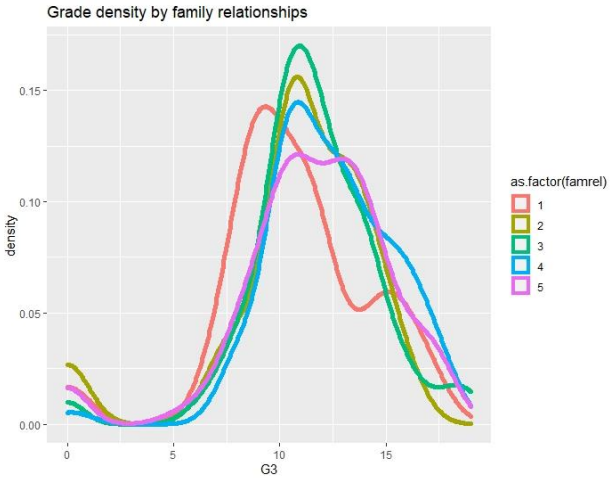
Same thing with Fedu graph.

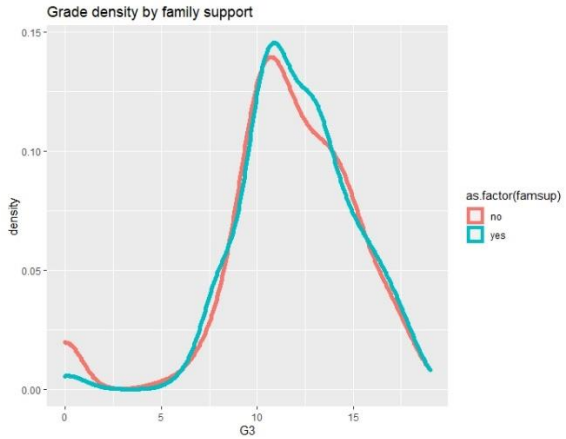




Parents job:

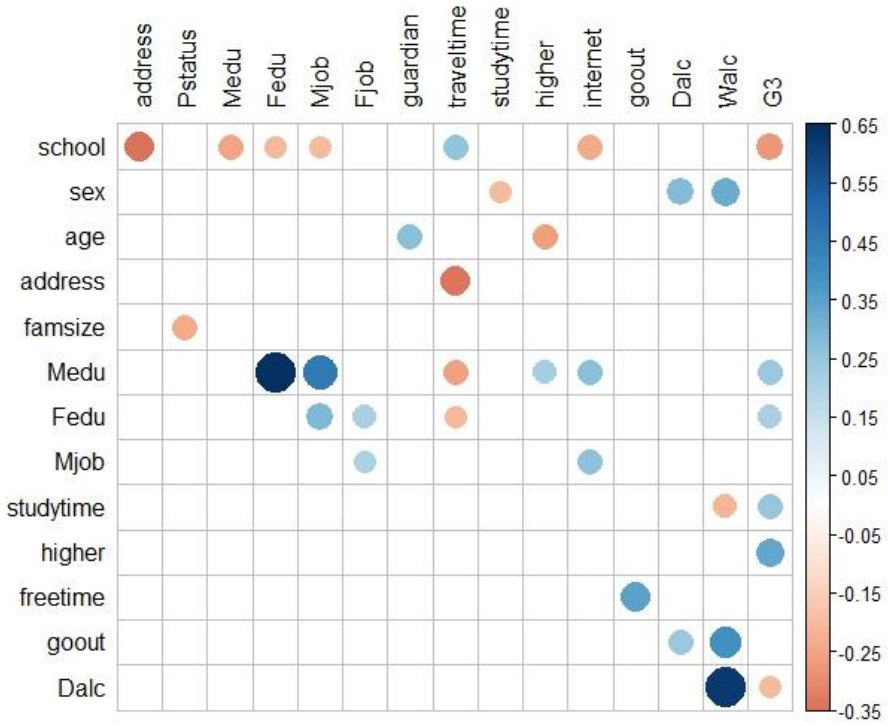
It seems the parent education graph and parents job correlate in a nice way, a portion of parents who are at home have kids who score as high as any other highest scoring student's.
Maybe those are students who have good family support\relationship with their parents





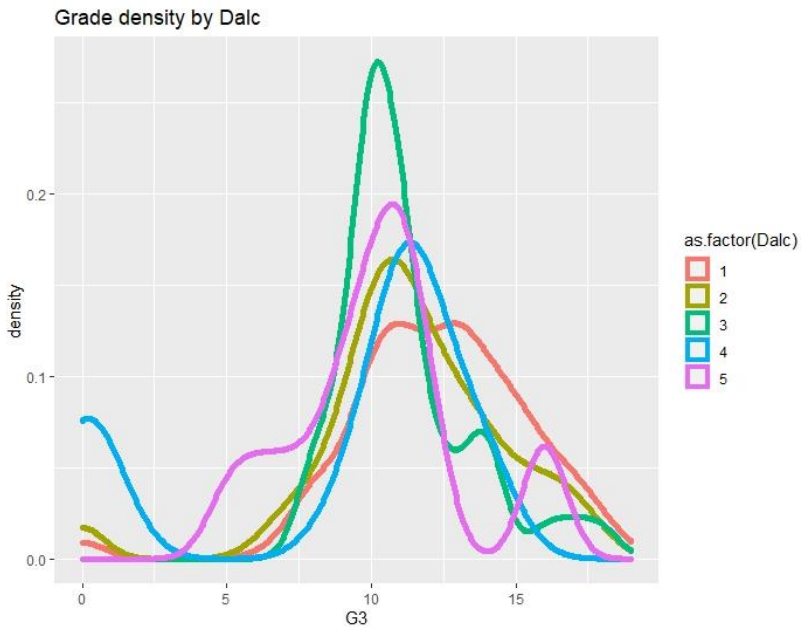
Family relations and support:

Family support doesn't give us any indications, while family relationships do show the same bump from the previous graphs, which could mean that a healthy and positive family structure can be more or as indicative of high grades as things like parent education, which we'd logically wouldn't conclude.



<u>Correlations between all data variables:</u>

An overall correlations processing between all variables lets us see the big picture but mostly what's important to us, the correlation of every variables to G3.
We can see negative correlation between school and Dalc with G3 and positive correlation with Medu, Fedu, study time and higher with G3, it seems the family relationship correlation we saw might be too "weak" for us to consider.



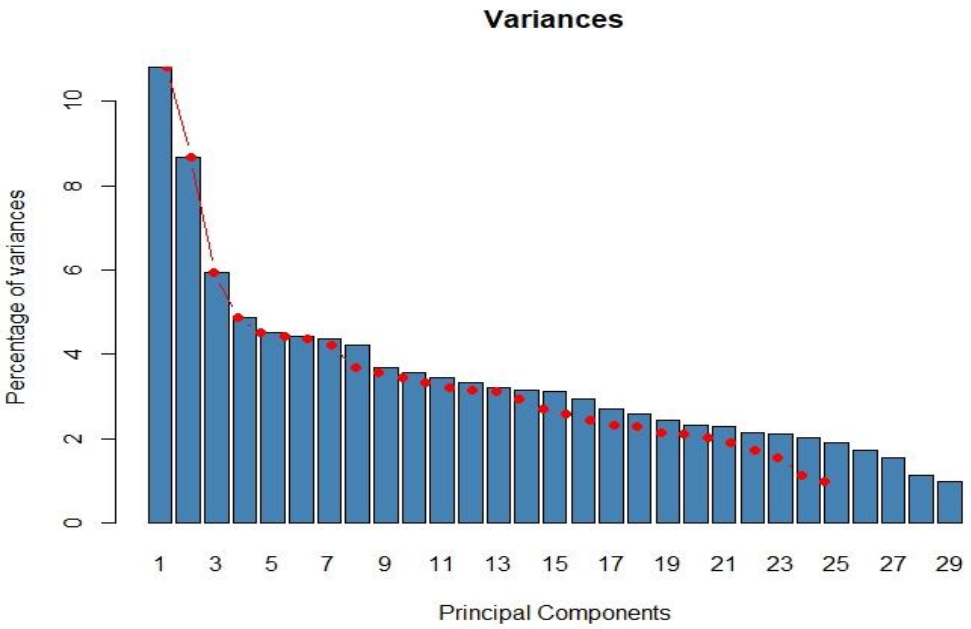
multiple variables linear model:

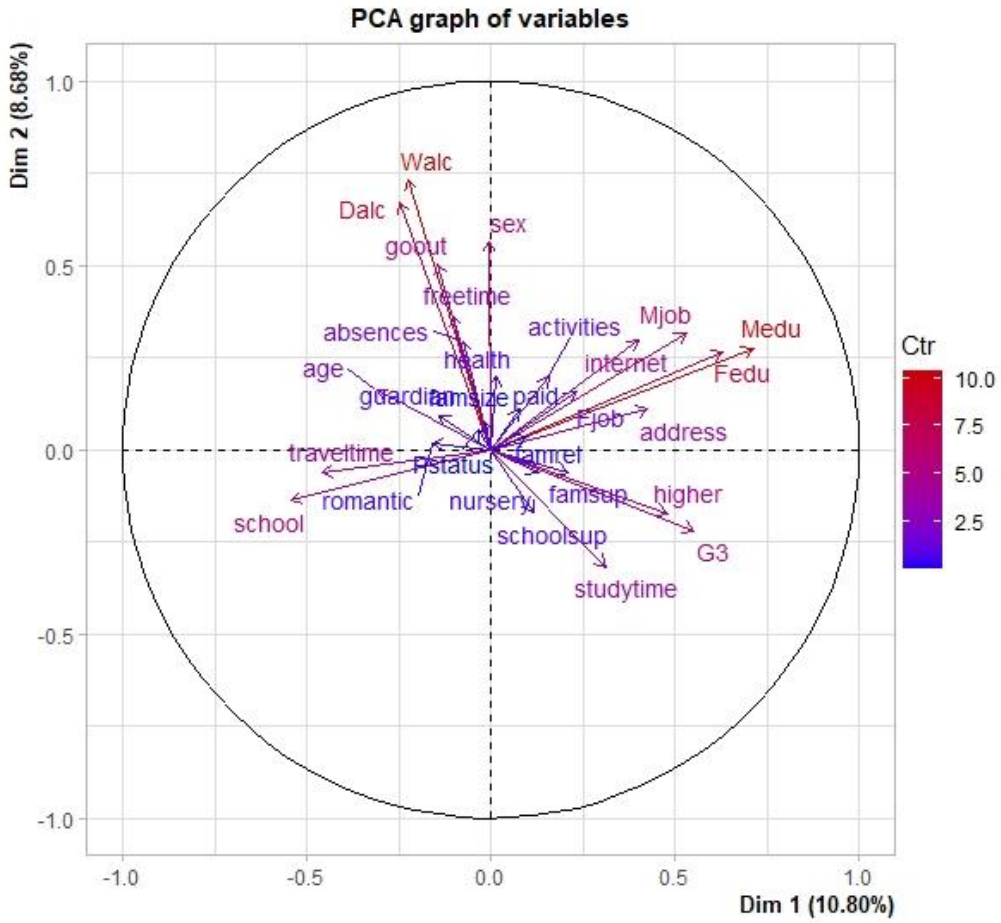
```
lm(formula = G3 ~ ., data = dt)
Residuals:
    Min      1Q   Median      3Q     Max 
-12.4551  -1.3938  -0.0095   1.6261   8.1129 
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    10.63795    2.03475   5.228 2.35e-07 ***
schoolMS       -1.37958    0.27233  -5.066 5.39e-07 ***
sexM           -0.76719    0.25794  -2.974 0.003052 ** 
age            -0.01088    0.10260  -0.106 0.915622    
addressU        0.27482    0.26922   1.021 0.307763    
famsizeLE3      0.45375    0.25306   1.793 0.073461 .  
PstatusT        0.19626    0.35932   0.546 0.585122    
Medu            0.06953    0.15679   0.443 0.657591    
Fedu            0.21869    0.14272   1.532 0.125952    
Mjobhealth      0.72240    0.55478   1.302 0.193358    
Mjobother       0.13488    0.31292   0.431 0.666604    
Mjobservices    0.35930    0.38618   0.930 0.352533    
Mjobteacher     0.59840    0.52026   1.150 0.250513    
Fjobhealth     -0.73903    0.77537  -0.953 0.340900    
Fjobother      -0.27904    0.47181  -0.591 0.554450    
Fjobservices   -0.77046    0.49586  -1.554 0.120751    
Fjobteacher     0.53720    0.69587   0.772 0.440420    
guardianmother -0.38047    0.27489  -1.384 0.166837    
guardianother  -0.22395    0.54878  -0.408 0.683356    
traveltime      0.01539    0.16413   0.094 0.925337    
studytime       0.51221    0.14344   3.571 0.000384 ***
schoolsupyes   -1.53469    0.37537  -4.088 4.92e-05 ***
famsupyes      -0.08364    0.23564  -0.355 0.722764    
paidyes        -0.67418    0.47451  -1.421 0.155882    
activitiesyes   0.20506    0.22848   0.898 0.369800    
nurseryyes     -0.15270    0.28139  -0.543 0.587567    
higheryes       2.22545    0.39067   5.696 1.90e-08 ***
internetyes     0.39709    0.28567   1.390 0.165022    
romanticyes    -0.45035    0.23710  -1.899 0.057977 .  
famrel          0.20332    0.12023   1.691 0.091337 .  
freetime       -0.21047    0.11583  -1.817 0.069704 .  
goout          -0.03650    0.11135  -0.328 0.743170    
Dalc           -0.22032    0.15727  -1.401 0.161759    
Walc           -0.07818    0.12259  -0.638 0.523895    
health         -0.20603    0.07910  -2.605 0.009422 ** 
absences       -0.04689    0.02566  -1.827 0.068178 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.766 on 613 degrees of freedom
Multiple R-squared:  0.3064,	Adjusted R-squared:  0.2668 
F-statistic: 7.736 on 35 and 613 DF,  p-value: < 2.2e-16

lm_summary_dt$adj.r.squared
[1] 0.2667744
```

Firstly, we’ll note that according to our results, adj.r.squared = 0.2667744,
Secondly, the p.value we got is minuscule, being below the required value for it to be statistically significant, meaning we can determine that there is a variable in our dataset that is correlated to G3, which is good news for us!.

Principal component analysis

We'll first see that, estim_ncp returns 2 meaning the first two dimensions are the most significant, which are hardly 19% of the variables, with can see that in the screeplot as well





Secondly, we're gonna look at the clusters
Group1: higher, G3, studytime, schoolsup, famsup,nursey
Group2: Mjob, Medu, Fedu, address, internt, Fjob, activities
Group3: Walc, Dalc, goout, freetime, absences
Group4: traveltime, school, romantic

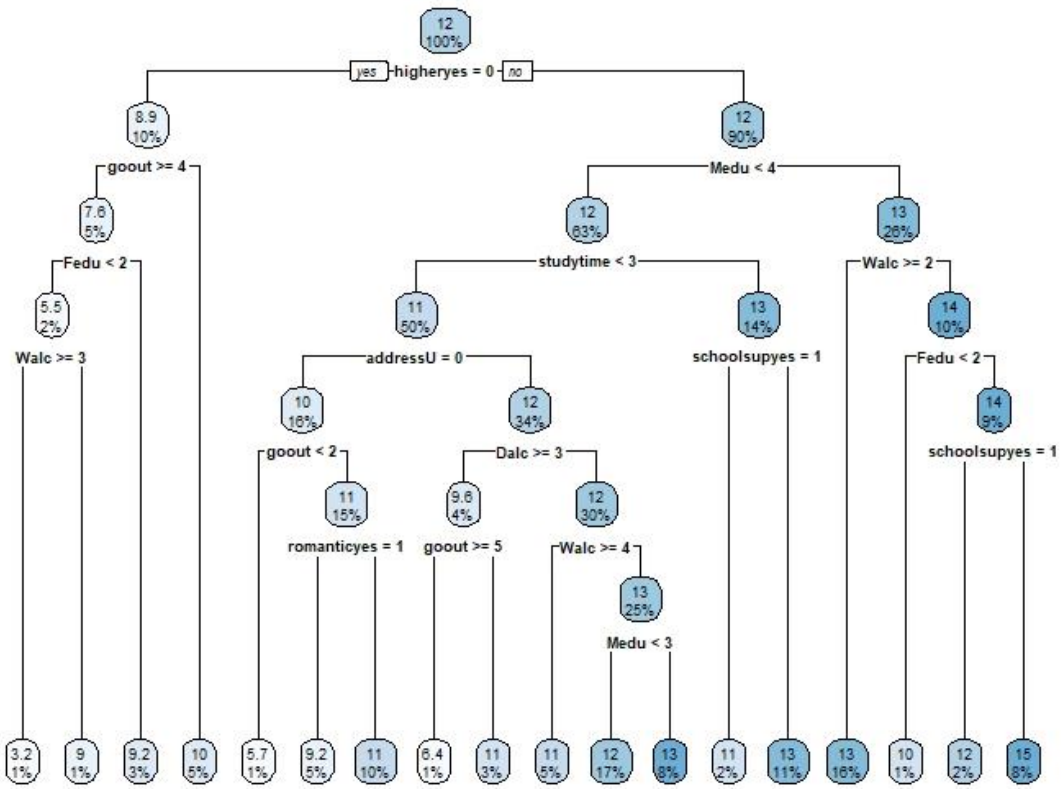
Group1 is anti correlated to group4
Group1 is semi correlated to group2 which is anti correlated to group4
(semi meaning they correlated in dim1 but anti-correlated in dim2)

Decision/bagging/random forest tree model predictions:

Decision tree:

With the analysis of PCA, we picked the highest contributors to the first two dimensions And filtered out school, age and sex, on the assumption that these aren't variables strongly correlated to the subject of socioeconomic, or at least shouldn't be.

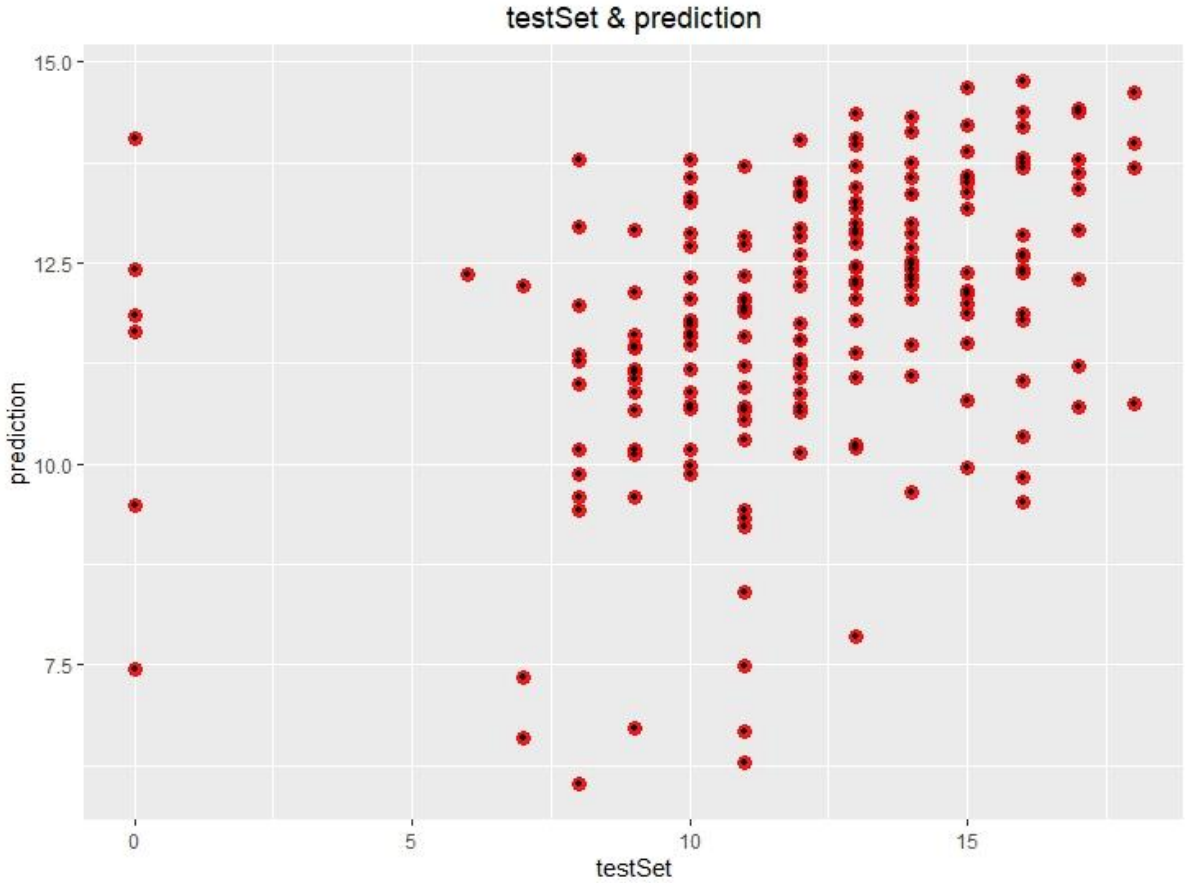
In each model we'll see the "Root-mean-square deviation" of each model, and we'll see what the model thinks are the strongest contributors to determining G3



RMSD = 3.73382, meaning the mean of all the deviations between the predicated value and the actual value is 3.73382, so on average if the actual score of a student was 15, the model would've (on average) predicated either 18.73 or 11.67, which seems pretty bad considering the values are between 0-20, most valuable variables are, from first to last:

Absences, activitiesyes, address, Dalc famrel famsupyes Fedu Fjobservices freetime goout guardianmother higheryes internetyes Medu nurseryyes paidyes romanticyes schoolsupyes studytime traveltime Walc

Tree bag:

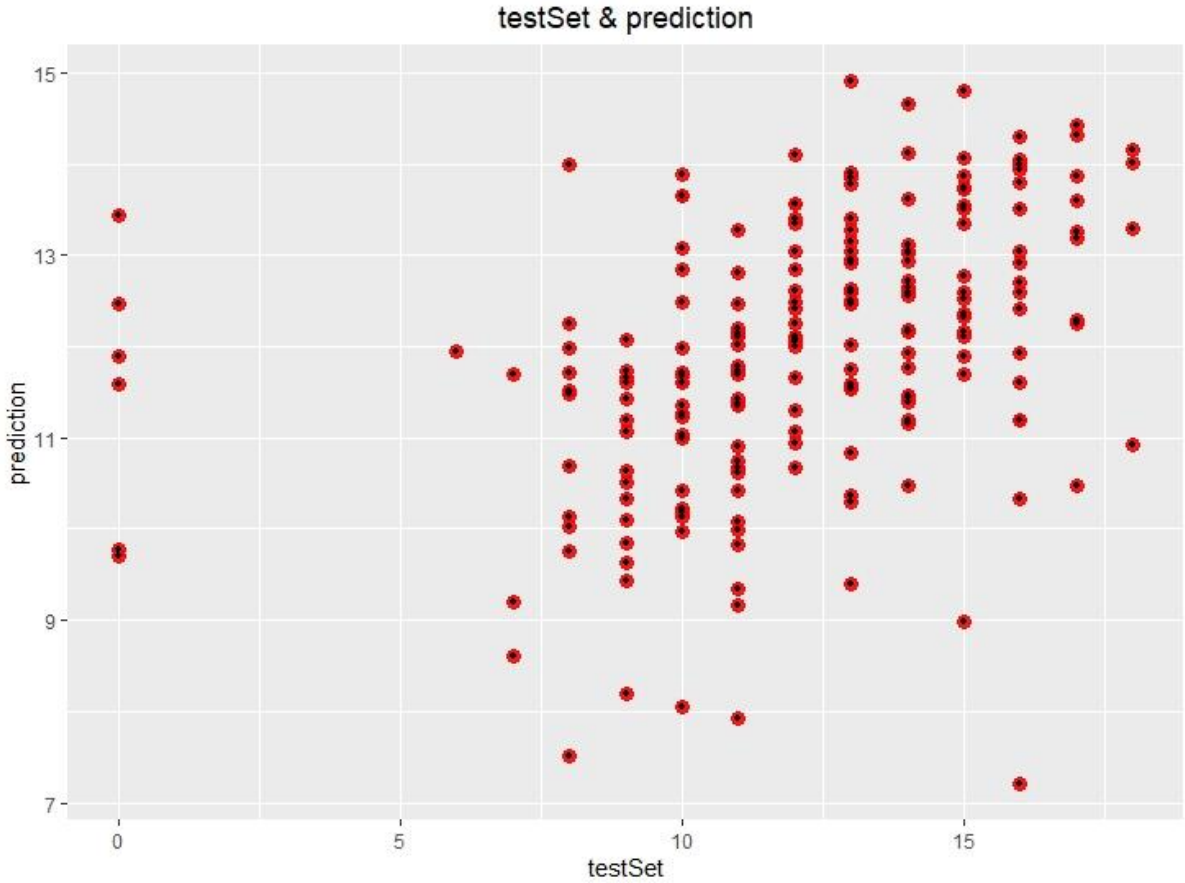


RMSD = 3.205063

It seems treebagging is an improvement over decision tree model

```
  only 20 most important variables shown (out of 29)

               Overall
Walc            100.00
goout            83.79
Medu             77.16
traveltime       75.48
Fedu             72.54
absences         62.07
studytime        58.22
freetime         51.63
addressU         45.09
Fjobservices     43.04
Dalc             42.52
activitiesyes    38.56
famrel           38.17
Fjobother        36.02
guardianmother   33.60
famsupyes        31.06
Mjobother        28.08
schoolsupyes     22.75
Mjobteacher      22.26
romanticyes      20.73
```

Random Forest:



RMSD = 3.172892 even more of an improvement over both bagtree and decision tree with the most important variables being:-

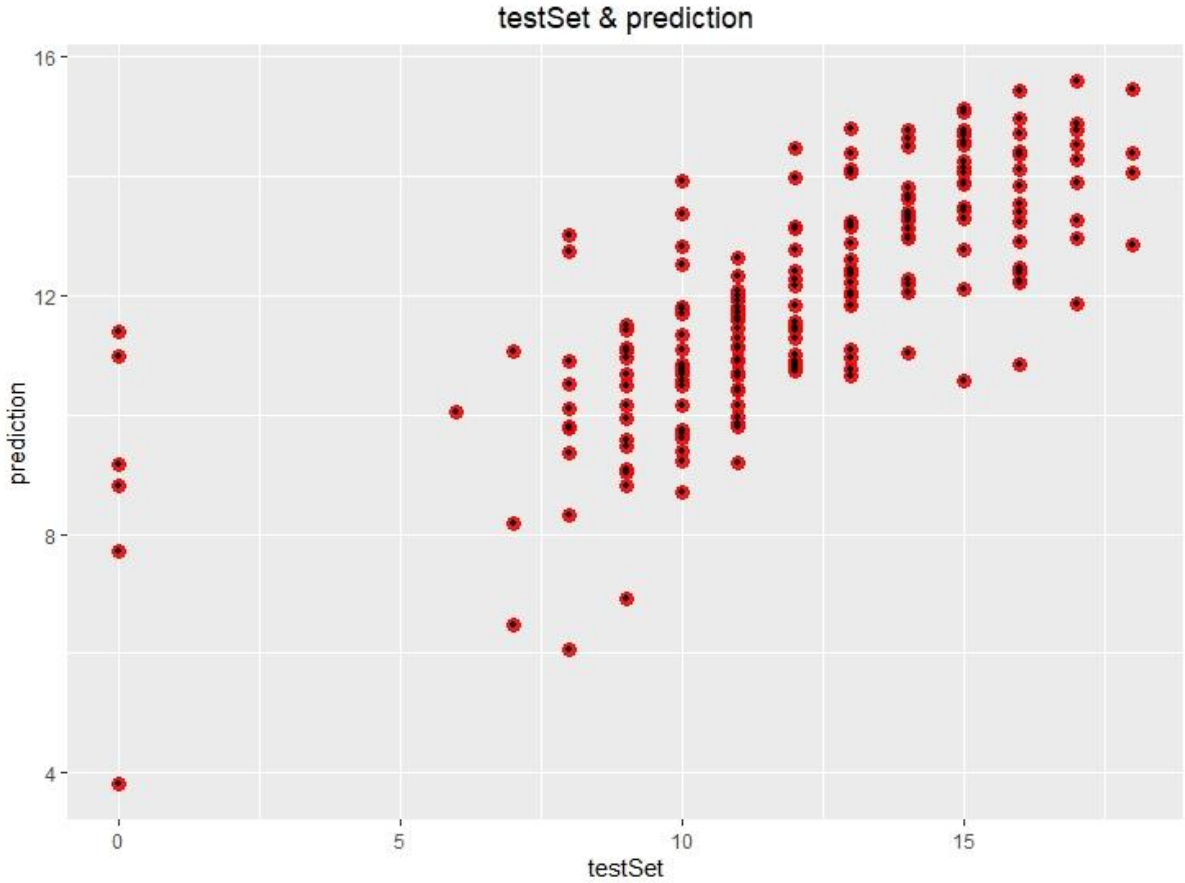
```
  only 20 most important variables showr

               Overall
Medu            100.00
Walc             97.61
goout            91.30
higheryes        90.86
absences         88.76
freetime         85.32
Fedu             84.23
studytime        76.38
Dalc             63.88
famrel           62.31
traveltime       48.30
addressU         43.89
Fjobservices     31.19
romanticyes      30.47
famsupyes        28.79
activitiesyes    27.35
internetyes      26.47
schoolsupyes     24.58
Fjobother        23.14
nurseryyes       21.01
```

Random Forest with PCA:

Lastly we'll see if we can improve our predication accuracy with the help of PCA with random forest tree model.

RMSD = 2.420946

Significant improvement compared to the other three models!




```
  only 20 most important variables shown

              Overall
Medu           100.00
Walc            97.61
goout           91.30
higheryes       90.86
absences        88.76
freetime        85.32
Fedu            84.23
studytime       76.38
Dalc            63.88
famrel          62.31
traveltime      48.30
addressU        43.89
Fjobservices    31.19
romanticyes     30.47
famsupyes       28.79
activitiesyes   27.35
internetyes     26.47
schoolsupyes    24.58
Fjobother       23.14
nurseryyes      21.01
```

In conclusion

Our main subject was whether or not there is correlation between student performance in Portuguese language classes and the student's socioeconomic status.

With the initial exploration it seemed there are apparent and logical correlations with G3 (famsup,Medu,Fedu,traveltime etc) that later were revealed to be insignificant, and things that were insignificant were determined to be significant by our models, though few variables stayed relevant, such as Medu traveltime and studytime, though the more we explored the more the variables changes, variables like Walc, goout and absences looked more strongly correlated.
Though we would've liked to see RMSD lower than 2.0 we nearly got there.
And my final conclusion would be we answered the question as a semi-yes, meaning we can see some correlation between student performances in Portuguese language classes and the student's socioeconomic status, but we cannot conclude with certainty what variables indicate the correlation between them.