DATA621 Final Project

Ivan Tikhonov, Seung Min Song, Alice Frieddman

Abstract

This study investigates the external factors affecting students' academic performance in secondary schools' math and Portuguese language courses. Using a quantitative data set obtained through a survey, the study analyzes various external variables, such as students' health, social and demographic factors, study habits, and other relevant factors to understand their impact on academic performance.

The study uses Multiple Linear Regression, Random Forrest, and a Decision Treet to examine the relationship between a dependent variable (final grade) and multiple independent variables. Ultimately the non-parametric results have more predictive value and success in modelling this complex phenomenon. The results of this study indicate that future researchers should consider nonparametric approaches to gain greater predicitive value.

Introduction

This study may be relevant to you as it explores the external factors that have an impact on students' grades. By analyzing the various external variables that can influence academic performance, this study provides valuable insights that can help parents, educators, and policymakers to better support students' success in school.

The data used in this study was collected from a survey of secondary school students and includes information on various external variables and their impact on students' grades. The data covers a range of subjects, including math, language, and other core academic areas, providing a comprehensive understanding of the factors that influence student performance across different subjects.

Literature Review

Academic success is a widely studied phenomenon. Previous studies have used Pearson's partial correlations, moderated linear regression, or analysis of variance between groups (ANOVA) to determine the relationship between factors such as demographics (e.g. sex), lifestyle habits (e.g. screen time), and physical and cognitive abilities (e.g. cardiovascular fitness measured by VO2 Max) to predict academic achievement. All of the studies presented below attempt to understand factors associated with a *student* rather than a school or teacher that my predict academic success, and so (as is the case with our data set) look at cross-section of students at a single school, so that the school itself is not considered a factor. This is different from our data set which includes two high schools.

A large number of factors have been identified in previous studies to predict academic achievement including sex, screen time, sleep, cell-phone use, maternal education, IQ; however, the effect typically found is low (explaining between approximately 6% and 50% of variation). This indicates that other factors not included in the study (which may include student-specific habits like study habits or school-related factors like school quality) may impact variation, and that the true model is likely complex.

Most studies in the literature reviewed take a linear approach; however, as acamic achievement is a complex phenomenon, new nonparametric approaches not previously available to researchers may prove more fruitful.

Additional details of the studies reviewed appear in the appendix.

Data Source

The data source is from a Kaggle and represents observations from 662 students at two high schools. The data is initially split into two data sets based on which subject is being observed (Portuguese language or math).

We will use the data set to predict academic achievement represented by the final grades, `final_grades`. A data dictionary is included in Appendix 1.

Data Exploration

Data Preparation

The first task is to merge the data sets, identifying which students are identical and which grades correspond to which class. In order to predict achievement in school independent of subject, with df_both, using a target `final_grade`, which is the average of both classes. Because students have answered with different number of absences in the different classes, we have also averaged the absences in this data set. There are 370 records with identical features in both datasets outside of the those specific to the subject. The field `paid` refers to subject-specific paid tutoring, and is changed to `paid_tutor` to indicate a paid tutor in either subject for the merged analysis.

```
## Joining with ‘by = join_by(school, sex, age, address, famsize, Pstatus, Medu,
## Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, schoolsup, famsup,
## activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc,
## Walc, health)‘
```

Include:

- EDA
  - The dataframe df_mat contains 395 rows with 33 columns, representing values related to a math class
  - The data set df_por contains 649 rows with 33 columns, representing values related to a Portuguese class
  - The combined data set df_both contains 370 rows with 44 columns.
  - The data types of the columns include integers (int64) and objects (object).
  - The columns represent various features of the student data set such as personal and family characteristics, academic performance, and social activities.
  - Some of the columns include: **school**, **sex**, **age**, **address**, **family size**, **parents’ education**, **mother and father’s occupation**, **travel time**, **study time**, **number of failures**, **support received from school**, **family**, **extra-curricular**, **activities**, **health**, and **academic grades**.

Overall Statistics

To gain comprehensive insights, we can merge *df_mat* and *df_por* data sets and analyze their collective statistics by looking at studends in both classes.

When examining the distinctive variables in the **df_both**, a few noteworthy ones stand out:

- failures:
  - The average is shown as 0.28, but considering that the maximum value is 3, it seems that some students have failed in several subjects in the previous semester.

    - The skew value is 2.75, which is a very large positive value, indicating that the data is skewed to the right.

- travel time:
    - The average is 1.45, and it seems that most students take a relatively short time to school.
    - The skewness value is 1.63, which is positive and indicates that the data is skewed to the right.
- absences:
    - The average is 5.38, indicating that on average students missed about 4 days.
    - Since the maximum value is 75, which is a very large value, it is possible that some students have many absences.
    - The skewness value is 4.03, which is a very large positive value, indicating that the data is very skewed to the right.
- Dalc (workday drinking):
    - The average is 1.48, and most students seem to be drinking relatively little on workdays.
    - The skewness value is 2.19, which is positive and indicates that the data is skewed to the right.
- Average Drinking: During the week, students average about 2.29 units of drinking.
    - The standard deviation of the drinking amount is 1.29, showing the degree of scattering of the data. Most students appear to have weekday alcohol consumption spread around the average.
    - The skewness value has a positive value of 0.61, and the distribution type is skewed to the right. That said, some students are likely to consume relatively more alcohol during the week.
- health:
    - The average is 3.56, and the average health status of students seems to be in the middle.
    - A negative skewness value of -0.51 indicates that the data is slightly skewed to the left.

‘A full descriptive summary of variables in included in Appendix 2.

Basic Plots

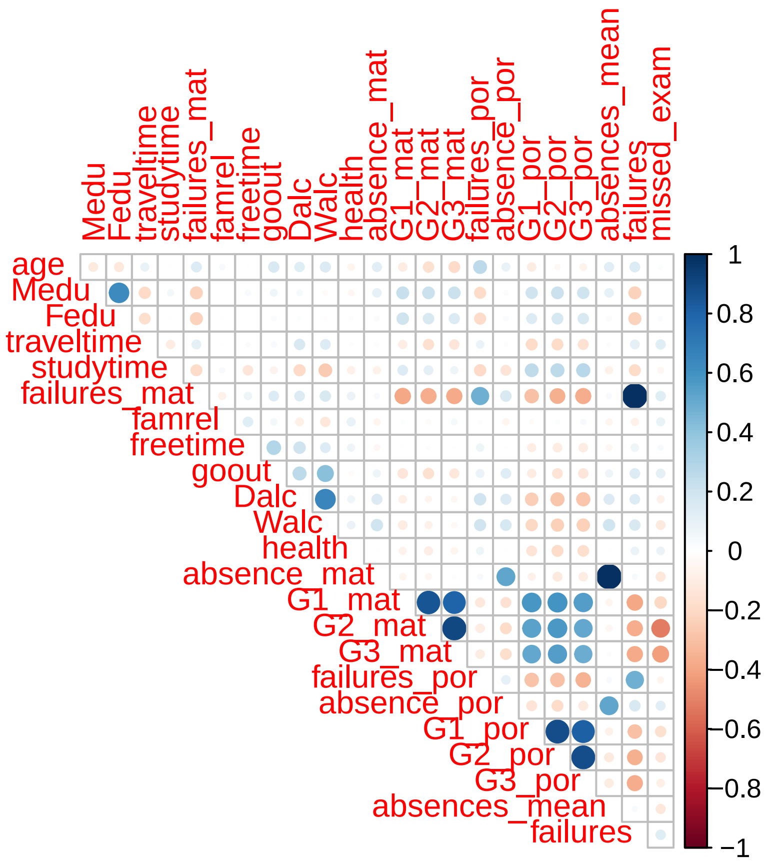
Basic plots of numeric variables help us understand if they have a normal distribution – necessary to undestand if a particular model type will be appropriate. In general, we can see most features are not normally distributed, meaning that linear regression is not likely to have good predictive value. Basic plots are included in Appendix 3.

Correlation

When considering the distribution of *traveltime*, *studytime*, *failures*, *famrel*, *freetime*, *goout*, *Dalc*, *Walc*, *health*, and *absences* across different categories, it becomes evident that these variables show a relatively even distribution within each G3 group. Grades and absences between the classse are highly correlated, so it makes sense to combine these for futher analysis.

We establish the presences or absence of multicollinearity in order to again determine the suitability of different model approaches, or determine if we should consider dropping some variables.

We can see that mother’s and father’s education (`Medu` and `Fadu`) as well as daily and weekly alcholol consumption (`Dalc` and `Walc`) are highly correlate.



Missing values

The data set is complete with no missing values; however in merging the two datasets we had to make a decision for how to deal with variables that are subject specific. this treatment is described below.

Renaming columns

We change `famsup` to `famEdsup`, change `Dacl` to `dayAlc`, change `Walc` to `weekAlc`, and change `goout` to `friendtime` so that it is easier to understand. We also change the name of binary values to reflect the 1 value, e.g

Adding new columns

We used the dplyr::mutate() method to create a new columns in the `df_both` data set: `final_grade`: an average of the final grade for both classes `absences_mean`: average absences reported across both classes `parentEdu`: sum of `Medu` + `Fedu` to remove a collinearity and combine parental education into a single variable `paid_tutor`: this column is marked as "yes" if the student receives paid tutoring in either subject `missed_exam`: this column indicates whether a student missed an exam in any class

Replacing binary values

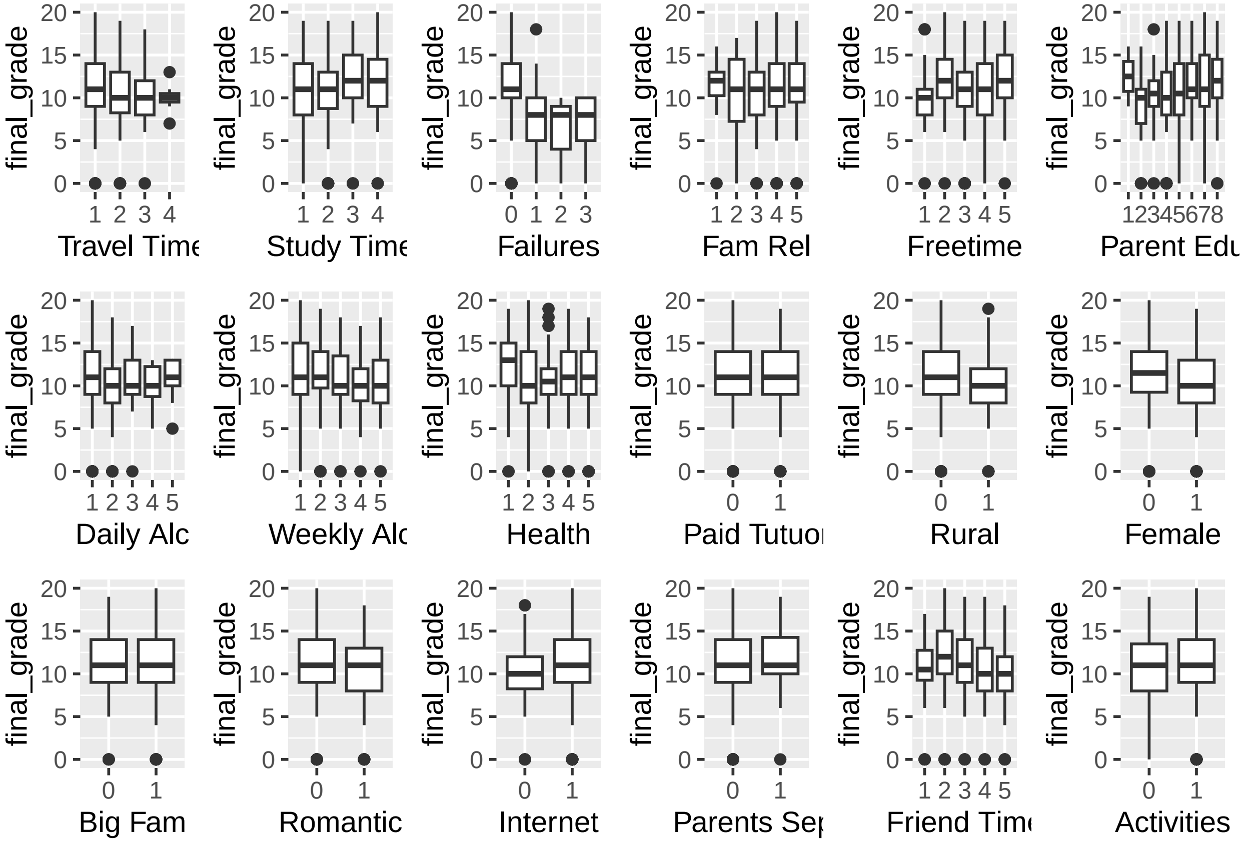
For easier computation and interpretation of results, we replace all binary data ("yes", "no"; "rural" or "urban"; "m" or "f") with 1 or 0.

Remove rows

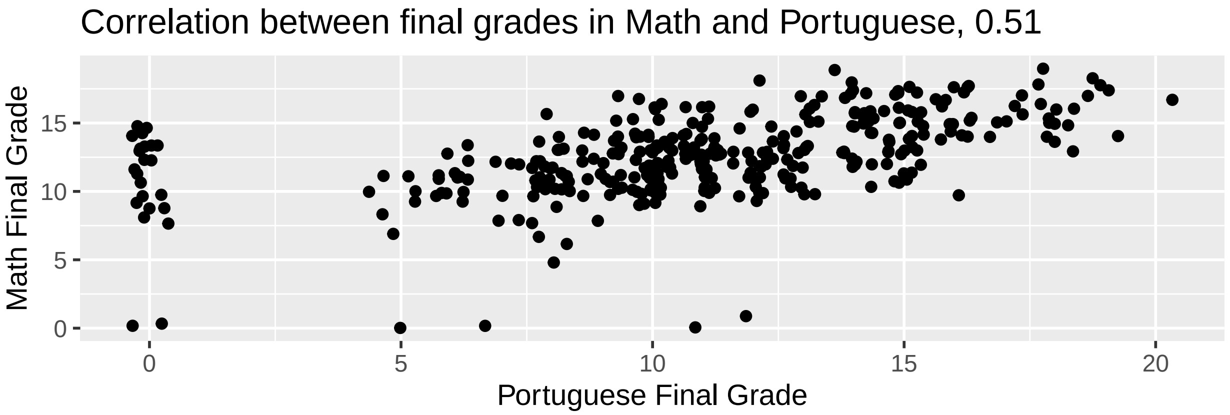
If the student received a 0 on the final exam, we remove them as they received their final grade due to a missing exam as opposed to other factors.

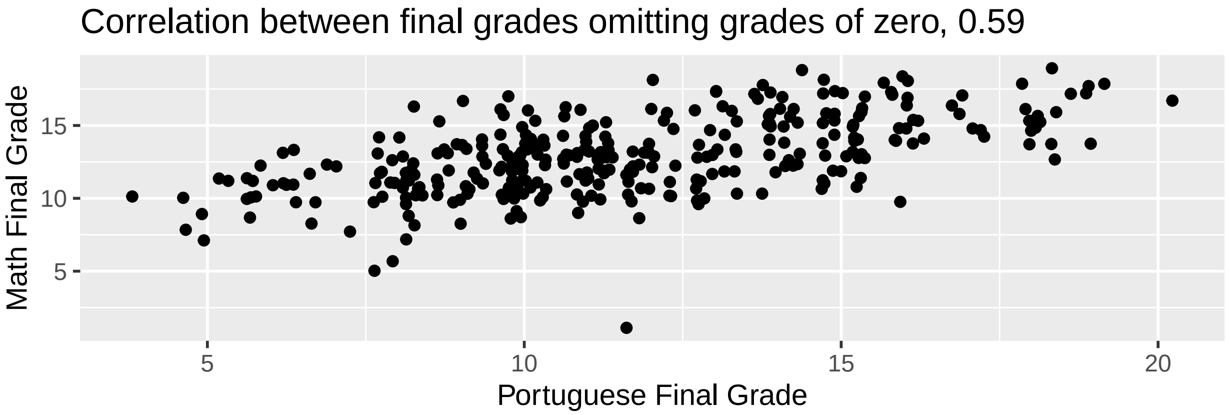
Data Visualization

We can use box plots to understand the impact of factor variables on the target final grade. We can see that `failures` jumps out as having significant predictive value, whle the presence of absence of a paid tutor makes little difference



In order to determine how best to measure overall student achievement across the classes, we evaluate the correlation between `G3_mat` and `G3_por`. In general, it appears that these correlation between these two subjects is quite high, especially if we omit final grades of “0” from consideration; this is helpful because it gives us a sense of the upper bound of how much variation in achievement can be explained by other features – this correlation, after all, is for final grades for different classes but for the same students.


Correlation between final grades in Math and Portuguese, 0.51


Correlation between final grades omitting grades of zero, 0.59

Additional data visualizaton can be found in Appendix 5

Models

We will attempt to predict the target, `final_grade` using three approaches: multiple linear regression, random forrest, and a decision tree.

Model 1

Multiple Linear Regrssion We will first attempt the simplest case multiple linear regression using the full subset method to choose the best linear model. Using the `leaps` package and 3-fold cross validation from the `caret` package, we can select the best performing model among possile linear regression strategies.

What is very interesting about this result is that the same best subset is not created for each fold – this indicates that the relationships are not especially strong, except for those few that appear each time. The only consistent result with this methodology is to include `failures`: Students who have previously failed a class are more likely to fail again.

A visualization and ANOVA confirm this results – students who have previously failed are significantly likely to have a lower final grade than students who have never failed.

By evaluating the BIC curves, we can determine that the minimum BIC is between 3 and 6 features. Using the rule of thumb of the fewest feautres within 1 SD, we will choose 3 features.

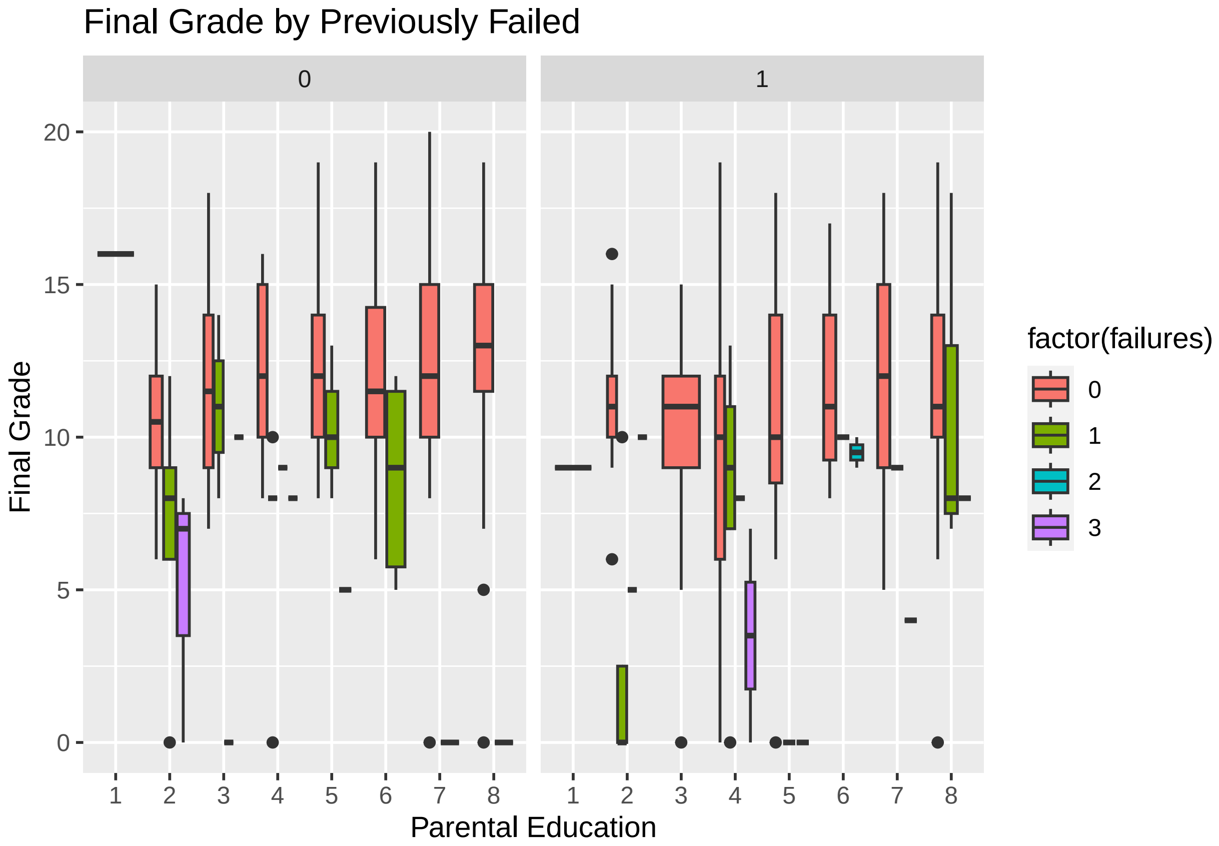
The coefficients in the best subset with 3 features are `failure`, `sex_f` and `parentEDU` – not dissimilar from results from previous studies.

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 2 linear dependencies found

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 2 linear dependencies found

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 2 linear dependencies found
```

Using a 3-fold validation, the mean correlation between predicted and actual final grades is 0.39 and the R squared is 17%.



We can also attempt to use stepwise regression to see if there is a different result. In fact the result is very similar with a .39 correlation between test and predicted results and multiple R-squared of 0.20. RMSE for both are almost identical at 4.12 and 4.10 respectively.

```
## Start:  AIC=630.26
## final_grade ~ sex_F + failures + parentEdu
##
##                       Df Sum of Sq    RSS    AIC
## <none>                             3251.3 630.26
## + sex_F:failures       1     15.10 3236.2 631.15
## + failures:parentEdu   1      6.49 3244.8 631.78
## + sex_F:parentEdu      1      5.12 3246.2 631.88
## - sex_F                1     55.82 3307.1 632.31
## - parentEdu            1     86.47 3337.7 634.51
## - failures             1    324.99 3576.3 650.93
```

```
##
## Call:
## lm(formula = final_grade ~ sex_F + failures + parentEdu, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7219  -1.9125   0.2303   2.4291   8.5662
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.0958     0.8542  11.819  < 2e-16 ***
## sex_F        -0.9751     0.4865  -2.004   0.0462 *
## failures     -1.7744     0.3669  -4.836  2.4e-06 ***
## parentEdu     0.3283     0.1316   2.495   0.0133 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.728 on 234 degrees of freedom
## Multiple R-squared:  0.1612, Adjusted R-squared:  0.1505
## F-statistic:    15 on 3 and 234 DF,  p-value: 5.835e-09
```

```
## [1] 0.4461409
```

```
## [1] 4.204069
```

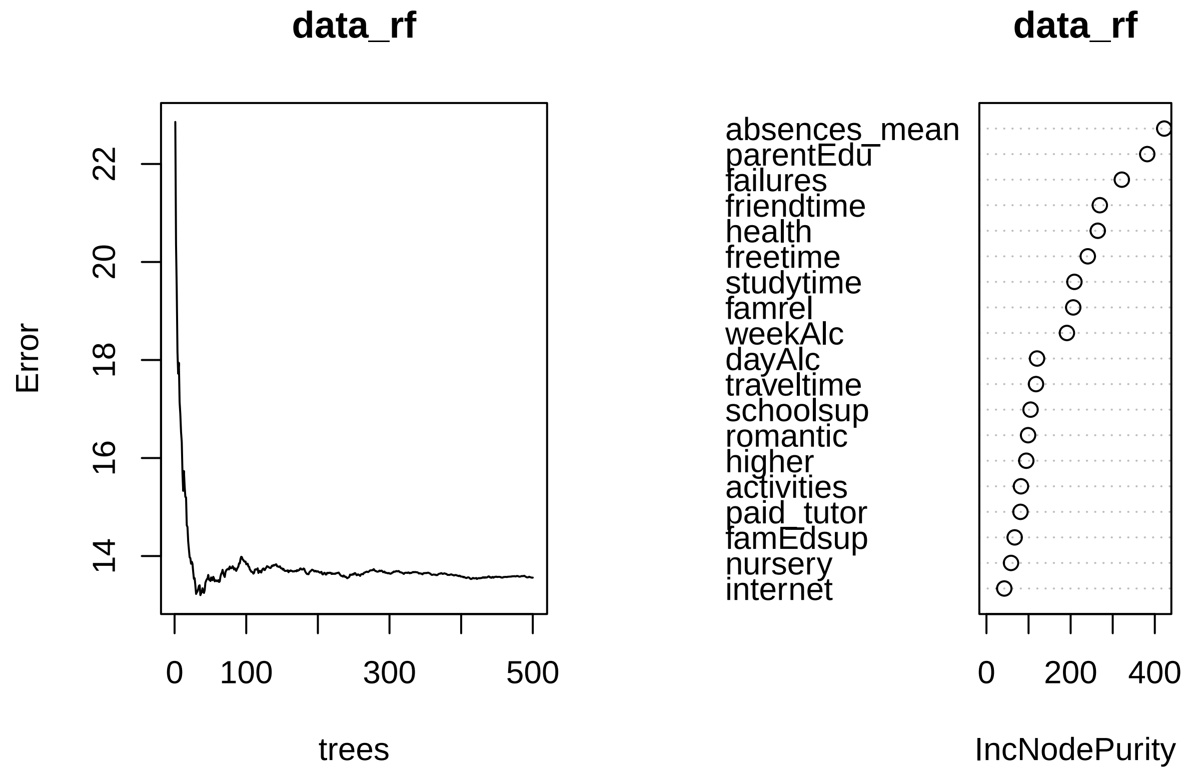
```
## [1] 4.204069
```

Model 2

Random Forest Model

Random Forrest has superior test results, with a correlation between predicted and actual results of 0.57 and a lower RMSE of 3.79.

```
##
## Call:
##  randomForest(formula = final_grade ~ traveltime + studytime +      failures + schoolsup + famEdsup
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##           Mean of squared residuals: 13.56023
##                     % Var explained: 16.74
```

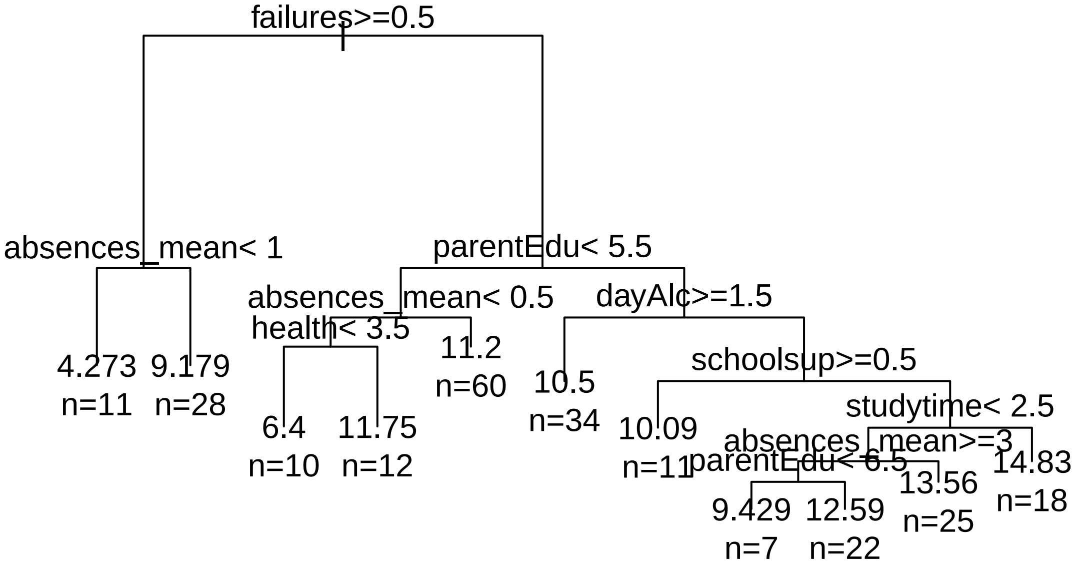


```
## null device
##           1
```

Model 3

Decision tree model

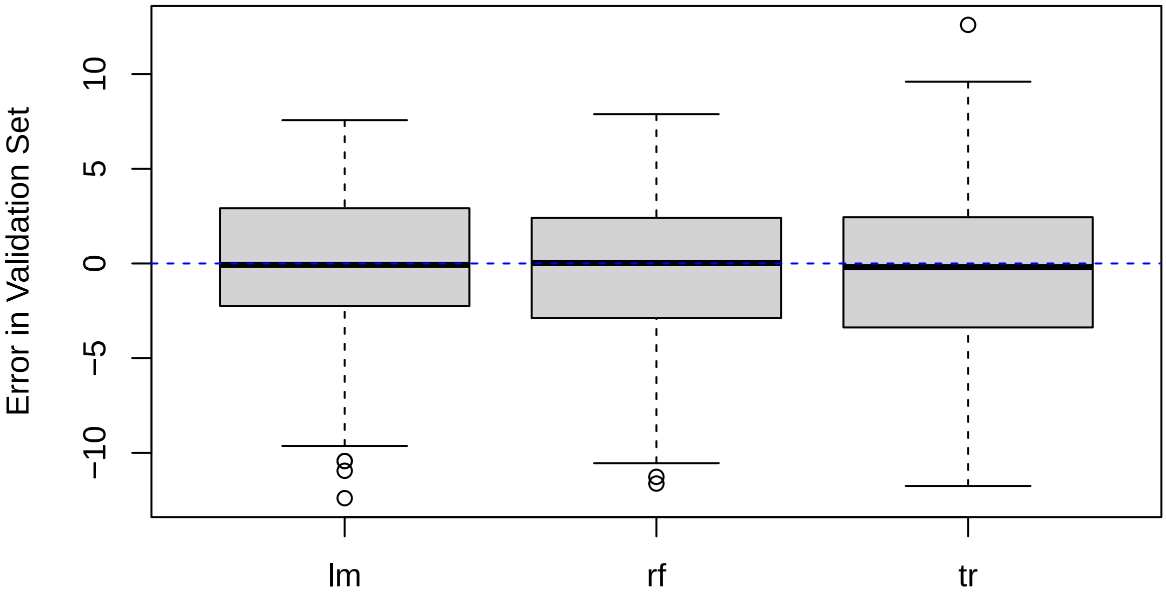
```
## n= 238
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##   1) root 238 3876.31900 10.915970
##     2) failures>=0.5 39  646.35900  7.794872
##       4) absences_mean< 1 11  256.18180  4.272727 *
##       5) absences_mean>=1 28  200.10710  9.178571 *
##     3) failures< 0.5 199 2775.59800 11.527640
##       6) parentEdu< 5.5 82 1209.37800 10.695120
##        12) absences_mean< 0.5 22  732.77270  9.318182
##          24) health< 3.5 10  300.40000  6.400000 *
##          25) health>=3.5 12  276.25000 11.750000 *
##        13) absences_mean>=0.5 60  419.60000 11.200000 *
##       7) parentEdu>=5.5 117 1469.55600 12.111110
##        14) dayAlc>=1.5 34  528.50000 10.500000 *
##        15) dayAlc< 1.5 83  816.65060 12.771080
##           30) schoolsup>=0.5 11   56.90909 10.090910 *
##           31) schoolsup< 0.5 72  668.65280 13.180560  
##             62) studytime< 2.5 54  442.59260 12.629630  
##              124) absences_mean>=3 29  226.13790 11.827590  
##                248) parentEdu< 6.5 7   17.71429  9.428571 *
##                249) parentEdu>=6.5 22  155.31820 12.590910 *
##              125) absences_mean< 3 25  176.16000 13.560000 *
##             63) studytime>=2.5 18  160.50000 14.833330 *
```

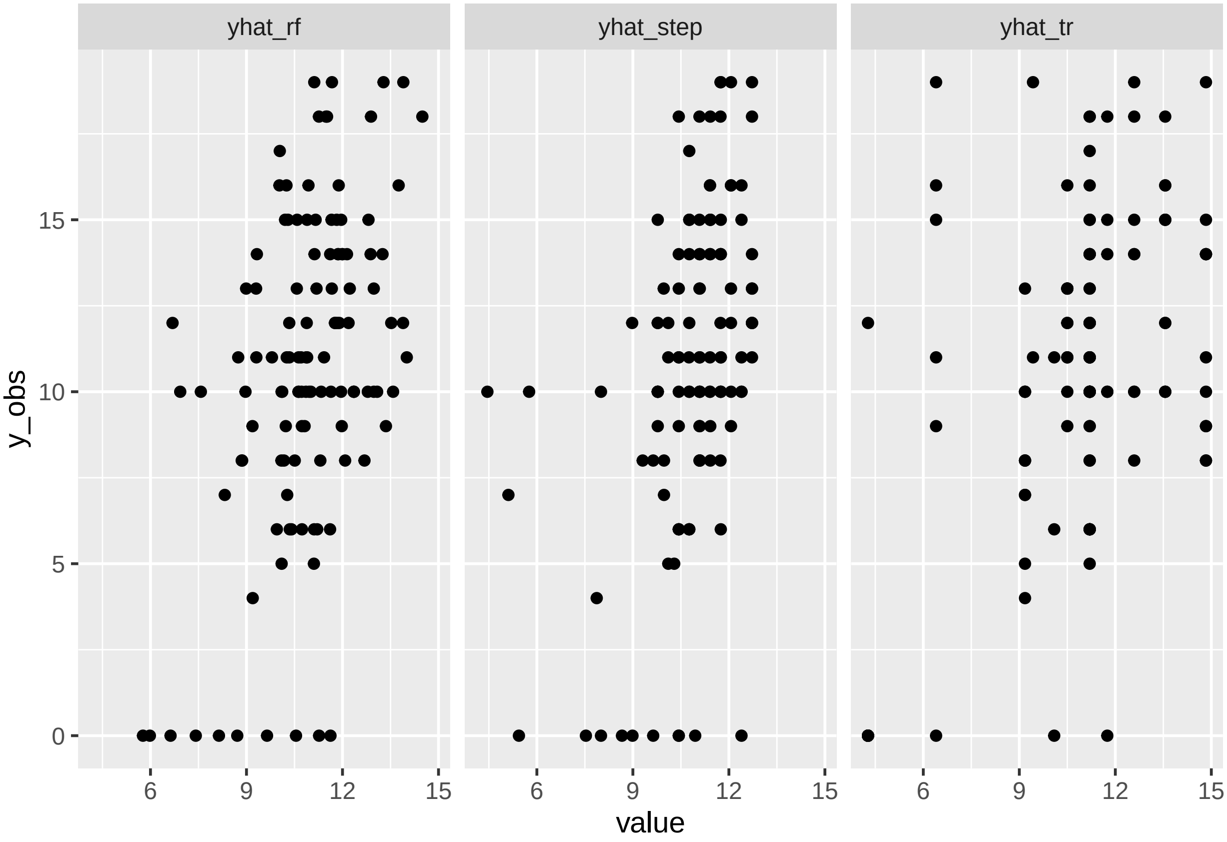


In a complex model, a decision tree can be effective. This model results in correlation between predicted and actual final grades of `cor(y_obs, yhat_tr)` as well as an RMSE of `rmse(y_obs, yhat_tr)`.

Model Selection

The best model is not one with the highest R-squared, but the one that perfoms the best on test data. In fact, all of these models performed similarly, but the non-linear models had a slightly better correlation.


Error in Validation Set
10
5
0
−5
−10
lm
rf
tr



Conclusion

The most significant predictor for the students final grade turns out to be having previousy failed a class in a linear model. This is an interesting and actionable results, as students who have previously failed are a known population to whom extra school support can be given. School support, thankfully, also rises to the top of the predictors. The best predictive model is the Decision Tree, but the differences are in fact extemely subtle. Decision Tree modeling is harder to interpret or explain.

Ultimately, the choice of the best model depends upon the goal of the study. For a policy maker, the simpler linear model may be preferred.

Appendix 1: Data Dictionary

Adapted from Kaggle: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

- *absences* - number of school absences (numeric: from 0 to 93)
- *absences_mean* - mean reported absences from both classes for students in Portuguese and Math
- *activities* - extra-curricular activities (binary: yes or no)
- *address* - student's home address type (binary: 'U' - urban or 'R' - rural)
- *age* - student's age (numeric: from 15 to 22)
- *Dalc / dayAlc* - workday alcohol consumption (numeric: from 1 - very low to 5 - very high) *renamed
- *failures* - number of past class failures (numeric: n if 1<=n<3, else 4)
- *famsize* - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- *famsup / famEdsup*- family educational support (binary: yes or no) *renamed

- *Fedu* - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- *Fjob* - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- *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) *renamed
- *goout* / *friendtime* - going out with friends (numeric: from 1 - very low to 5 - very high)
- *guardian* - student's guardian (nominal: 'mother', 'father' or 'other')
- *health* - current health status (numeric: from 1 - very bad to 5 - very good)
- *higher* - wants to take higher education (binary: yes or no)
- *internet* - Internet access at home (binary: yes or no)
- *Medu* - mother'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')
- *nursery* - attended nursery school (binary: yes or no)
- *parentEdu* - sum of mother's + fathers education level
- *paid* - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- *Pstatus* - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- *reason* - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- *romantic* - with a romantic relationship (binary: yes or no)
- *school* - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- *schoolsup* - extra educational support (binary: yes or no)
- *sex* - student's sex (binary: 'F' - female or 'M' - male)
- *studytime* - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- *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)
- *Walc* / *weekAlc* - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) *renamed

These grades are related with the course subject, Math or Portuguese and are dropped for the joint analysis:

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)

This is the target variable: *final_grade* - final grade averaged from both classes (numeric: from 0 to 20, output target)

Appendix 2: Summary Statistics

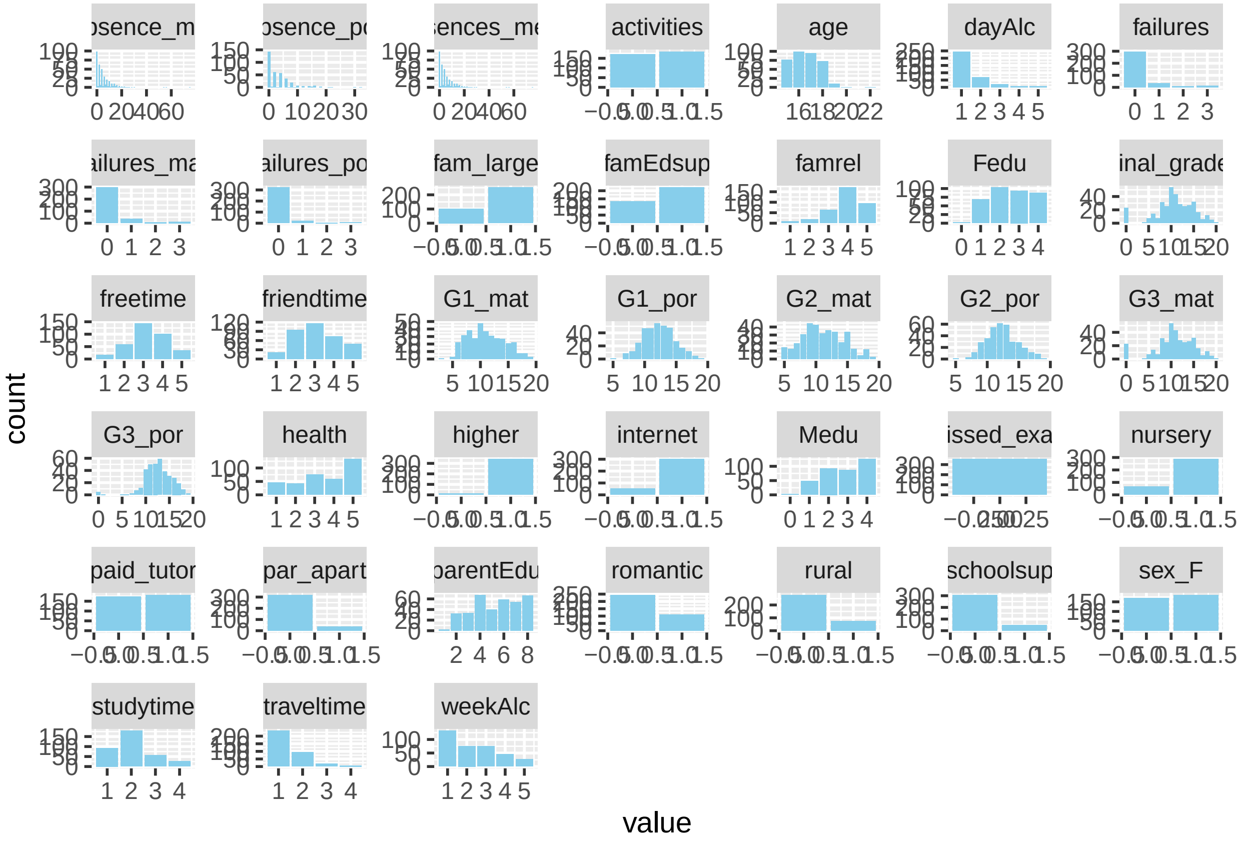
```
df_both
```

```
##                vars   n  mean   sd median trimmed  mad min max range  skew
## school*           1 370  1.11 0.31      1    1.01 0.00   1   2     1  2.56
## sex*              2 370  1.47 0.50      1    1.47 0.00   1   2     1  0.11
## age               3 370 16.58 1.18     17   16.54 1.48  15  22     7  0.41
## address*          4 370  1.78 0.41      2    1.85 0.00   1   2     1 -1.35
## famsize*          5 370  1.28 0.45      1    1.23 0.00   1   2     1  0.97
## Pstatus*          6 370  1.90 0.30      2    2.00 0.00   1   2     1 -2.61
## Medu              7 370  2.80 1.08      3    2.89 1.48   0   4     4 -0.38
## Fedu              8 370  2.56 1.09      3    2.58 1.48   0   4     4 -0.08
## Mjob*             9 370  3.18 1.22      3    3.23 1.48   1   5     4 -0.32
## Fjob*            10 370  3.30 0.85      3    3.32 0.00   1   5     4 -0.27
## reason*         11 370  2.26 1.21      2    2.20 1.48   1   4     3  0.41
## guardian*       12 370  1.81 0.49      2    1.83 0.00   1   3     2 -0.38
## traveltime      13 370  1.45 0.70      1    1.31 0.00   1   4     3  1.63
## studytime       14 370  2.04 0.85      2    1.96 0.74   1   4     3  0.64
## failures_mat    15 370  0.28 0.71      0    0.08 0.00   0   3     3  2.75
## schoolsup*      16 370  1.13 0.34      1    1.04 0.00   1   2     1  2.16
## famsup*         17 370  1.62 0.48      2    1.66 0.00   1   2     1 -0.51
## paid_mat*       18 370  1.47 0.50      1    1.46 0.00   1   2     1  0.12
## activities*     19 370  1.52 0.50      2    1.52 0.00   1   2     1 -0.06
## nursery*        20 370  1.81 0.40      2    1.88 0.00   1   2     1 -1.54
## higher*         21 370  1.96 0.20      2    2.00 0.00   1   2     1 -4.47
## internet*       22 370  1.85 0.36      2    1.93 0.00   1   2     1 -1.91
## romantic*       23 370  1.32 0.47      1    1.28 0.00   1   2     1  0.76
## famrel          24 370  3.94 0.91      4    4.03 1.48   1   5     4 -0.94
## freetime        25 370  3.22 0.99      3    3.22 1.48   1   5     4 -0.14
## goout           26 370  3.12 1.13      3    3.09 1.48   1   5     4  0.12
## Dalc            27 370  1.48 0.90      1    1.27 0.00   1   5     4  2.19
## Walc            28 370  2.29 1.29      2    2.15 1.48   1   5     4  0.61
## health          29 370  3.56 1.41      4    3.70 1.48   1   5     4 -0.51
## absence_mat     30 370  5.38 7.67      4    4.00 5.93   0  75    75  4.03
## G1_mat          31 370 10.89 3.35     11   10.77 4.45   3  19    16  0.27
## G2_mat          32 370 10.75 3.80     11   10.87 2.97   0  19    19 -0.40
## G3_mat          33 370 10.46 4.61     11   10.89 4.45   0  20    20 -0.70
## failures_por    34 370  0.13 0.49      0    0.00 0.00   0   3     3  4.30
## paid_por*       35 370  1.07 0.25      1    1.00 0.00   1   2     1  3.43
## absence_por     36 370  3.63 4.83      2    2.65 2.97   0  32    32  2.18
## G1_por          37 370 12.14 2.55     12   12.11 2.97   0  19    19 -0.16
## G2_por          38 370 12.27 2.47     12   12.18 2.97   5  19    14  0.26
## G3_por          39 370 12.55 2.94     13   12.65 2.97   0  19    19 -1.00
## absences_mean   40 370  5.38 7.67      4    4.00 5.93   0  75    75  4.03
## final_grade     41 370 10.46 4.61     11   10.89 4.45   0  20    20 -0.70
## failures        42 370  0.28 0.71      0    0.08 0.00   0   3     3  2.75
## paid_tutor*     43 370  1.50 0.50      1    1.50 0.00   1   2     1  0.01
## missed_exam     44 370  0.04 0.18      0    0.00 0.00   0   1     1  5.03
##               kurtosis   se
## school*           4.56 0.02
## sex*             -1.99 0.03
## age               0.08 0.06
## address*         -0.17 0.02
## famsize*         -1.06 0.02
## Pstatus*          4.81 0.02
## Medu             -1.03 0.06
## Fedu             -1.19 0.06
## Mjob*            -0.68 0.06
## Fjob*             0.98 0.04
## reason*          -1.40 0.06
## guardian*         0.18 0.03
## traveltime        2.39 0.04
## studytime        -0.06 0.04
## failures_mat      6.87 0.04
## schoolsup*        2.67 0.02
## famsup*          -1.74 0.03
## paid_mat*        -1.99 0.03
## activities*      -2.00 0.03
## nursery*          0.36 0.02
## higher*          18.06 0.01
## internet*         1.65 0.02
## romantic*        -1.43 0.02
## famrel            1.01 0.05
## freetime         -0.29 0.05
## goout            -0.82 0.06
## Dalc              4.72 0.05
## Walc             -0.80 0.07
## health           -1.04 0.07
## absence_mat      26.42 0.40
## G1_mat           -0.70 0.17
## G2_mat            0.48 0.20
## G3_mat            0.32 0.24
## failures_por     19.48 0.03
## paid_por*         9.80 0.01
## absence_por       6.50 0.25
## G1_por            0.77 0.13
## G2_por           -0.21 0.13
## G3_por            3.55 0.15
## absences_mean    26.42 0.40
## final_grade       0.32 0.24
## failures          6.87 0.04
## paid_tutor*      -2.01 0.03
## missed_exam      23.35 0.01
```

Appendix 3: Basic Plots

Students in Both Classes

```
## Warning in geom_histogram(stat = "count", fill = color): Ignoring unknown
## parameters: ‘binwidth‘, ‘bins‘, and ‘pad‘
```



Students in Math `basicp(df_mat, "pink")`

Students in Portuguese `basicp(df_por, "darkblue")`

Appendix 3: References

Data source accessed: Kaggle competition https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

Data source original citation: P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. Accessed via Kaggle

Title: Application of Multiple Linear Regression Identifying Contributing Factors in Students'Academic Achievement Authors: Dg Siti Nurisya Sahirah Binti Ag Isha and Siti Rahayu Binti Mohd Hashim Proceedings of the International Conference on Mathematical Sciences and Statistics 2022 (ICMSS 2022)At: Selangor, Malaysia Mathematics With Economics Programme, Faculty of Science and Natural Resources, December 2022 https://www.researchgate.net/publication/366929441_Application_of_Multiple_Linear_Regression_in_Identifying_Contributing_Factors_in_Students%27_Academic_Achievement

This study aimed to identify significant factors that contribute to students' academic success by analyzing internal and external factors. The study involved 327 final-year undergraduate students and found that self-esteem, intelligence, and maternal education were significant factors affecting students' achievement.

In this research, they made three different models to see which factors were most important for academic achievement. They found that self-esteem, IQ, and maternal education were the most important factors.

A Study on Academic Achievement and Personality of Secondary School Students Authors: Dr. Suvarna V. D.and Dr H. S. Ganesha Bhata1 Research in Pedagogy, v6 n1 p99-108, 2016 https://files.eric.ed.gov/fulltext/EJ1149330.pdf

This study uses ANOVA and the Pearson's product-moment coefficient to test hypotheses related to differences in academic achievement between different groups of students (including both demographics and responses to a personality test) using a data set of approximately 300 secondary school students.The results find that the students' age and gender affect their achievement levels, but that other demographic categories and (interestingly) personality characteristics do not.

Predictors of Academic Performance in High School Students: The Longitudinal ASAP Study

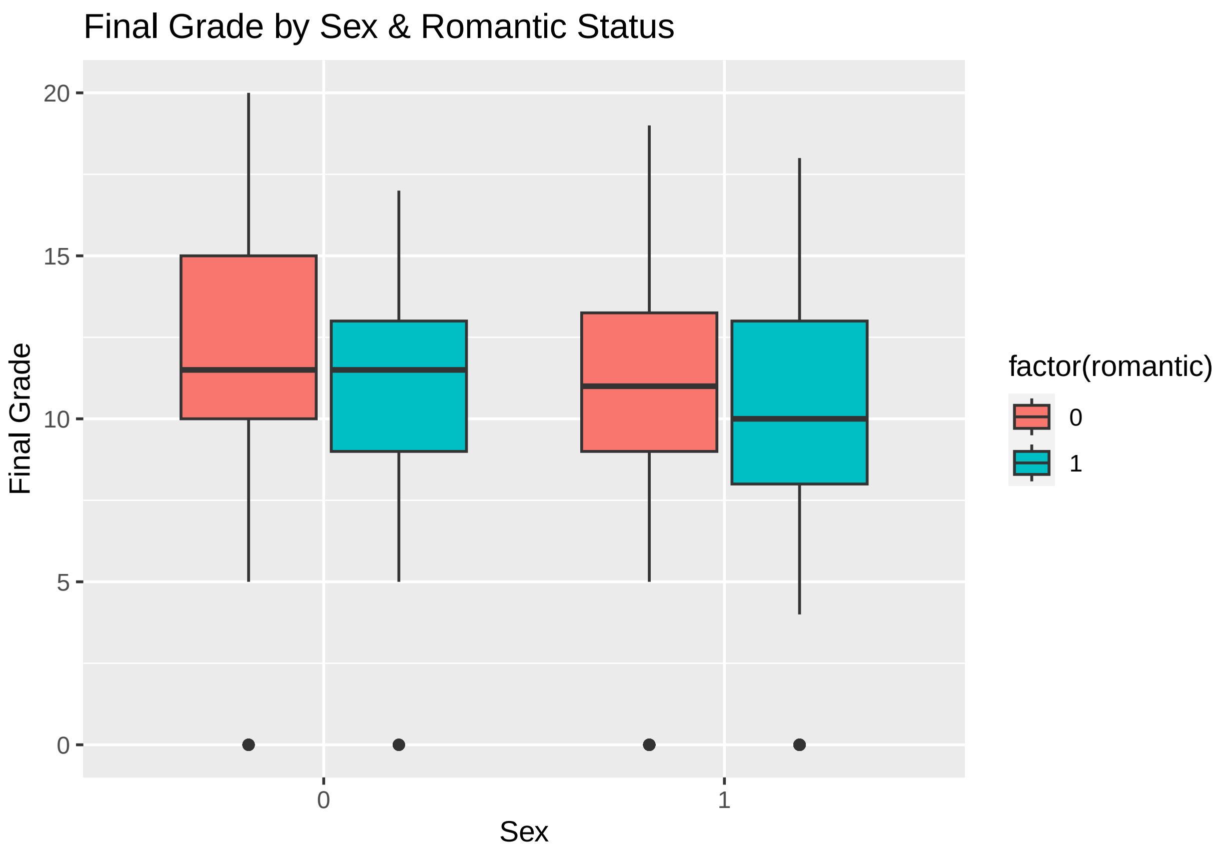
Authors: Marie-Maude Dubuc, Mylene Aubertin-Leheudr, and Antony D. Karelis Published online 2022 May 1, International Journal of Exercise https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9365103/#:~:text=Finally%2C%20psychological%20factors%20such%20as,42%2C%2043%2C%2048)

This study used moderated multivariate linear regression, separately comparing male and female students, to evaluate the impact of social, physical, and cognitive factors to explain variation in academic achievement among a cohort of 185 high-school students evaluated at a single high school over three years. The researchers controlled for demographic factors such as race, income, and ethnicity and use both a cross-sectional and longitudinal approach. In their results, they found that sex, cardiovascular fitness (measured by VO2 Max), and working memory were important predictors; however, they found that results differed when controlling for sex, school subject, and study design indicating that academic performance is in fact a highly complex phenomenon.

Appendix 4: Data Visualization

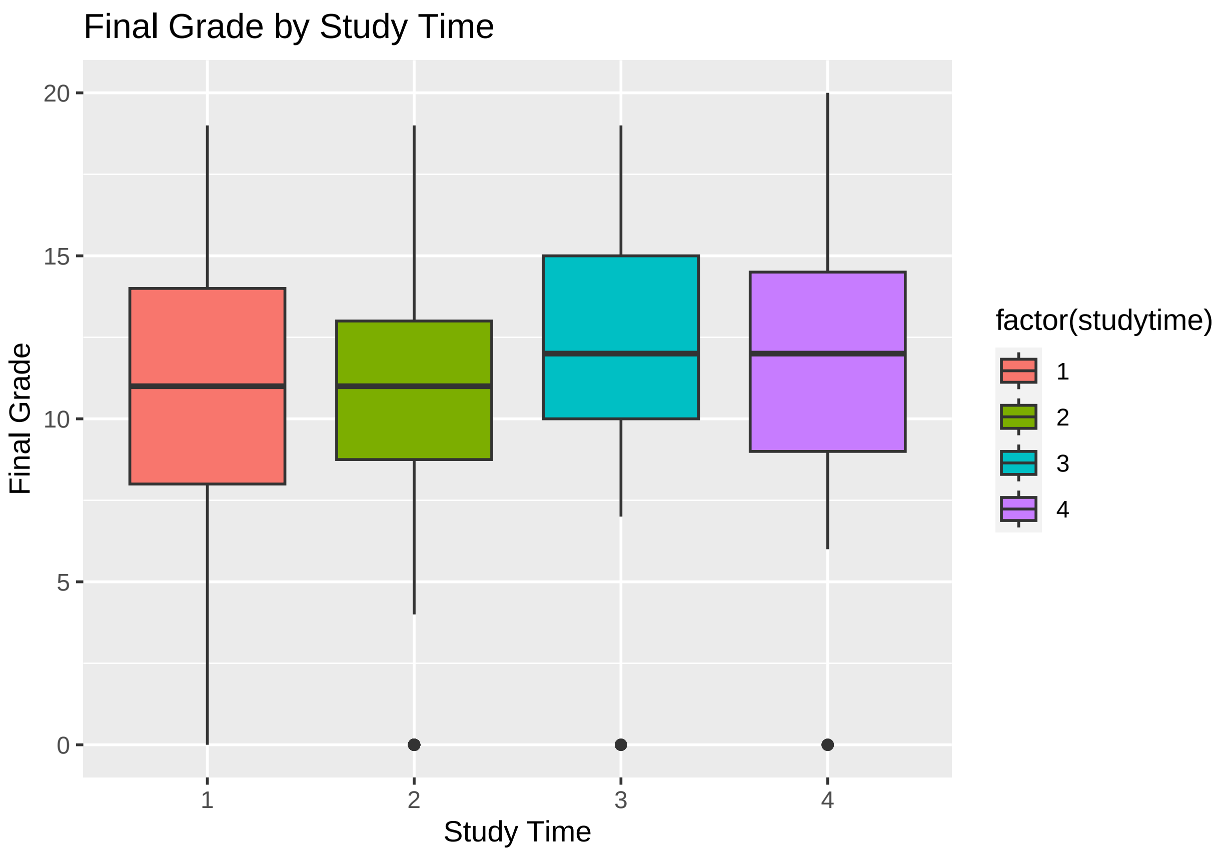
Romantic status

In general, students who are in a romantic relationship may have a more challenging time focusing on their studies compared to those who are not. A plot indicates that this may be the case – especially male students who are not romantically involved appear to have a slightly higher final grade; ANOVA confirms that romantic status and sex both have significant predictive value.



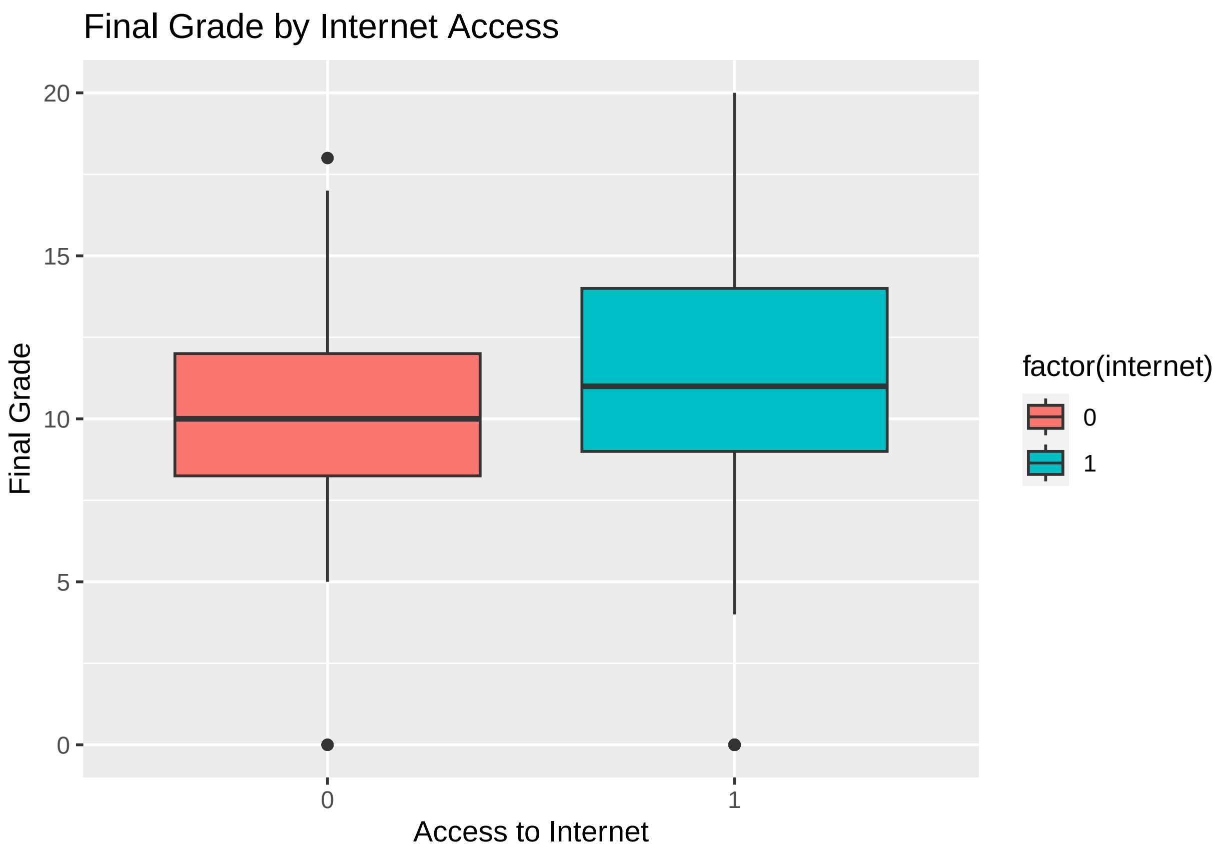
Study Time

Although at first glance the findings suggest that students who study for a longer duration have a higher average score compared to those who study for a shorter duration, study time is not predictive according to ANOVA and we cannot rule out chance.



Internet

The following graph indicates that the difference in academic performance between students with and without Internet access is negligible; again, however, ANOVA results indicate that we cannot rule out the null hypothesis that chance may explain the difference in means between the two groups.


Final Grade by Internet Access
20
15
10
5
0
Final Grade
0
1
Access to Internet
factor(internet)
0
1