DATA621 Final Project: Predicting Academic Achievement

Group 5

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 Tree 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 predictive 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 two schools and includes grades for two core subjects, math and language, providing a cross-sectional understanding of the factors that influence student performance.

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 academic 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 competition and represents observations from 662 students at two high schools in Portugal. 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.

Adding new columns

Fields which are specific to one class or another are aggregated or merged in the final data set: `final_grade`: an average of the final grade for both classes `absences_mean`: average absences reported across both classes `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

We additionally added the following fields to deal with multicollinearity found in the initial data analysis: `parentEdu`: sum of `Medu` + `Fedu` to remove a co-linearity and combine parental education into a single variable `studentAlc`: combination of weekly and daily student alcohol consumption

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. This is used to filter out students from the final analysis as this grade is presumed to indicate absence rather than student potential for achievement

Missing values

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

Renaming columns

We change `famsup` to `famEdsup`, 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 `sex` becomes `sex_F`.

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.

```r
df_mat <- read.csv("https://raw.githubusercontent.com/seung-m1nsong/602/main/Final/student-mat.csv")

df_por <- read.csv("https://raw.githubusercontent.com/seung-m1nsong/602/main/Final/student-por.csv")

df_both <- dplyr::inner_join(
  df_mat %>%
    rename(
      G1_mat = G1,
      G2_mat = G2,
      G3_mat = G3,
      absence_mat = absences,
      failures_mat = failures,
      paid_mat = paid
    ),
  df_por%>%
    rename(
      G1_por = G1,
      G2_por = G2,
      G3_por = G3,
      absence_por = absences,
      failures_por = failures,
      paid_por = paid
    )
  ) %>% rowwise() %>%
      mutate(
        absences_mean = mean(absence_mat, absence_por),
        final_grade = mean(G3_mat, G3_por),
        failures = mean(failures_mat, failures_por),
        paid_tutor = if_else(paid_mat=="yes" | paid_por=="yes", "yes", "no"),
        missed_exam = as.numeric(if_else(G3_por==0 | G3_mat==0, 1, 0))
      )
```

```r
df_both <- df_both %>%
  mutate(
  parentEdu = (Medu + Fedu),
  alc = (Dalc + Walc)) %>%
    select(-Dalc, -Walc, -Medu, -Fedu)
```

```r
binary <- function(df, col, level1="yes", level0="no"){
  df[[col]][df[[col]]==level1] <- 1
  df[[col]][df[[col]]==level0] <- 0

  return(df[[col]] %>% as.numeric())
}

df_both$schoolsup <- binary(df_both, "schoolsup")

df_both$famsup <- binary(df_both, "famsup")

df_both$paid_tutor <- binary(df_both, "paid_tutor")

df_both$activities <- binary(df_both, "activities")
```

```r
df_both$nursery <- binary(df_both, "nursery")

df_both$higher <- binary(df_both, "higher")

df_both$internet <- binary(df_both, "internet")

df_both$romantic <- binary(df_both, "romantic")

df_both$sex <- binary(df_both, "sex", "F", "M")

df_both$Pstatus <- binary(df_both, "Pstatus", "A", "T")

df_both$address <- binary(df_both, "address", "R", "U")

df_both$famsize <- binary(df_both, "famsize", "GT3", "LE3")

df_both$school <- binary(df_both, "school", "GP", "MS")
```

```r
df_both <- df_both %>% rename(
  famEdsup = famsup,
  sex_F = sex,
  friedtime = goout,
  rural = address,
  par_apart = Pstatus,
  friendtime = goout,
  fam_large = famsize,
  school_GP = school
)
```

```r
df_both <- df_both %>% filter(missed_exam == 0) %>% select(-missed_exam)
```

```r
descrip <- describe(df_both)
```

Prepared Data Set EDA

The resulting data set includes: *332 rows with 41 columns.* Various features of the student data set such as personal and family characteristics, academic performance, and social activities: **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

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

- failures:
  - The average is shown as 0.21, but considering that the maximum value is 3, it seems that some students have failed in several subjects in the previous semester. This excludes students who missed their fianl exam, as descibed above.
  - The skew value is 3.27, which is a very large positive value, indicating that the data is skewed to the right.

- travel time:
  - The average is 1.42, and it seems that most students take a relatively short time to school.
  - The skewness value is 1.73, which is positive and indicates that the data is skewed to the right.
- absences:
  - The average is 5.91, 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 3.98, which is a very large positive value, indicating that the data is very skewed to the right.
- alc (drinking):
  - The average is 3.83, and most students seem to be drinking relatively little on workdays.
  - The skewness value is 1.16, which is slightly positive and indicates that the data is skewed to the right.
- health:
  - The average is 3.55, and the average health status of students seems to be in the middle.
  - A negative skewness value of -0.52 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.

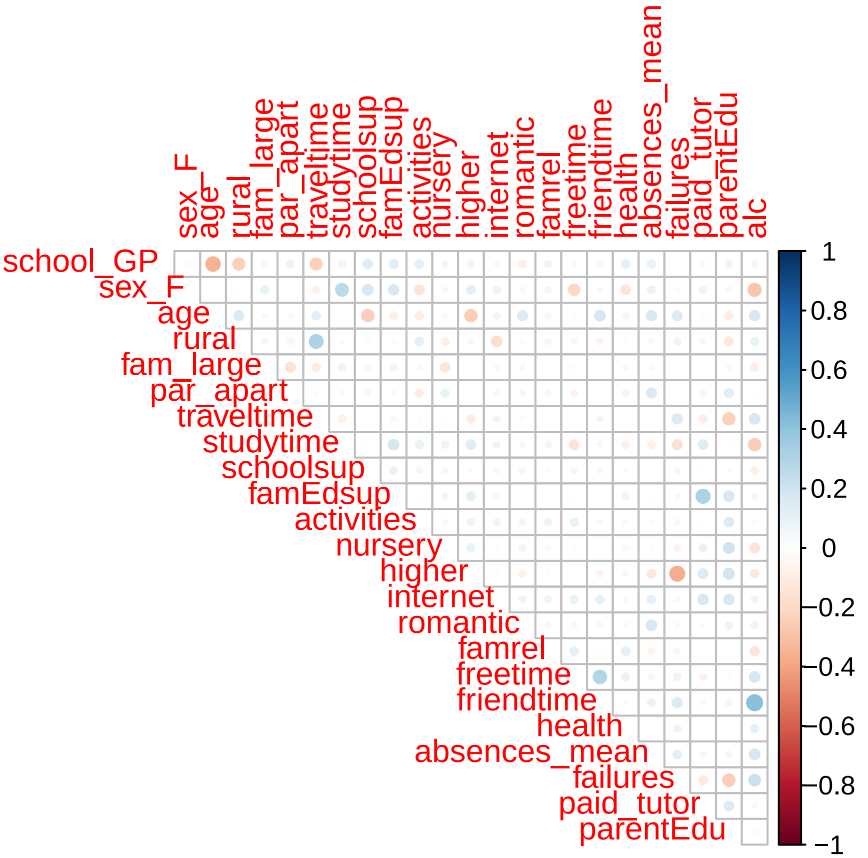
Correlation

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.

```
df_both %>%
  select(-final_grade, -ends_with("_por"), -ends_with("_mat")) %>% select_if(is.numeric) %>%
  cor() %>%
  corrplot(
  type="upper", diag = FALSE
  )
```

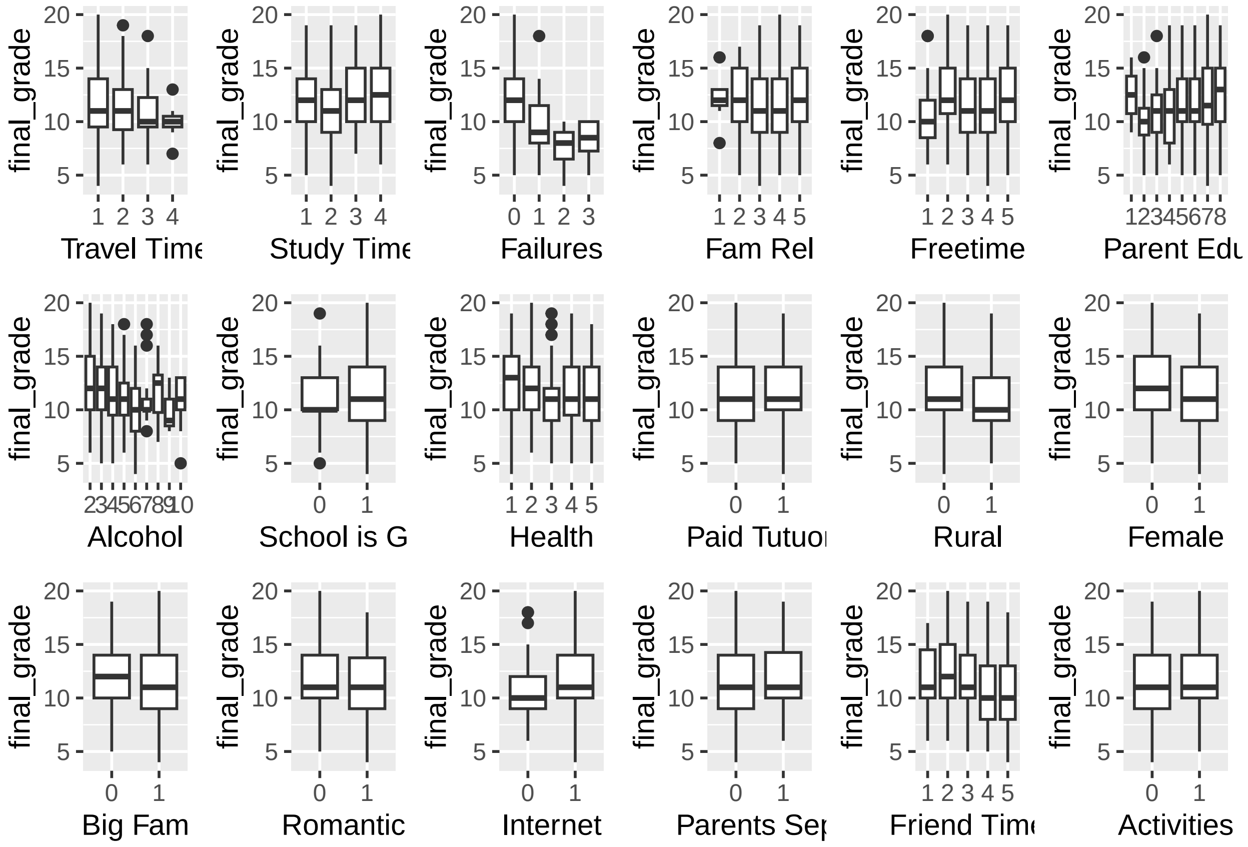


Data Visualization

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

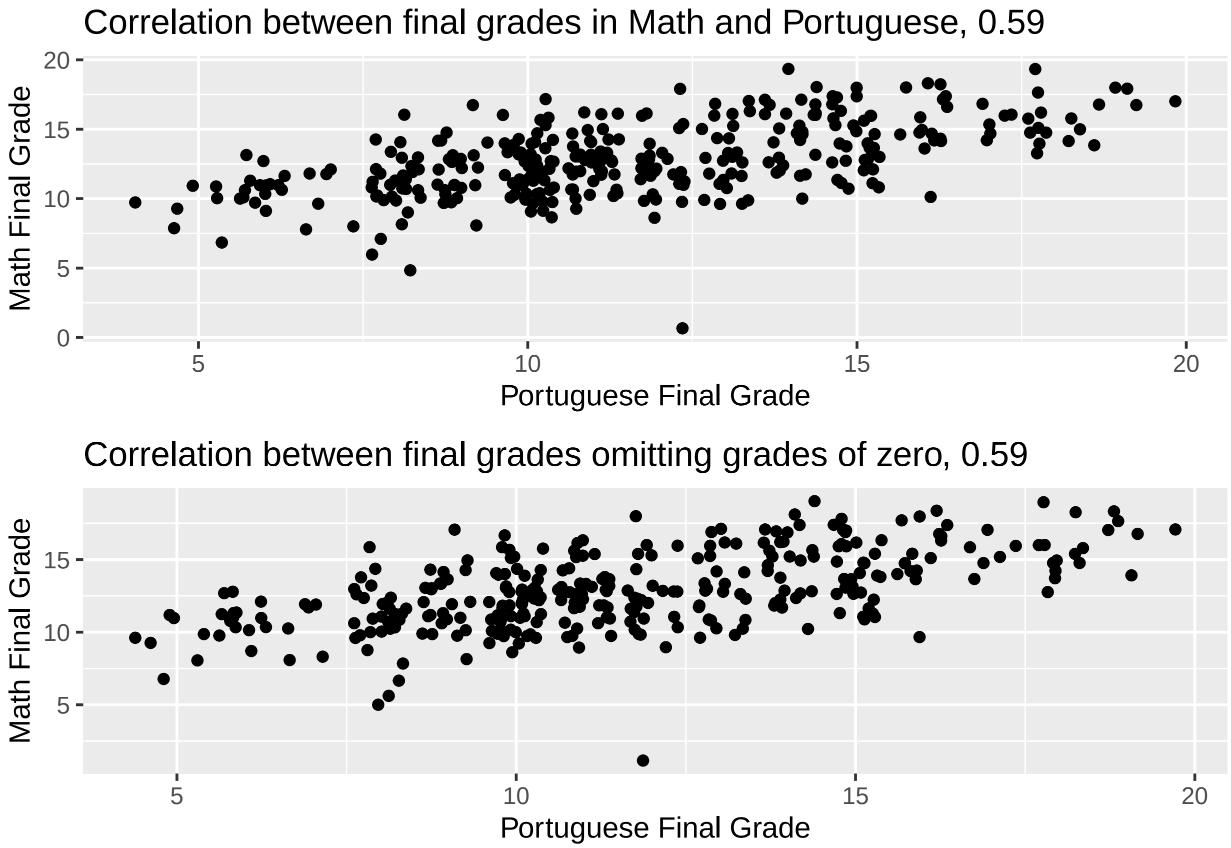
```
p1 <- df_both %>% ggplot(aes(factor(traveltime), final_grade)) + geom_boxplot() + xlab("Travel Time")
p2 <- df_both %>% ggplot(aes(factor(studytime), final_grade)) + geom_boxplot() + xlab("Study Time")
p3 <- df_both %>% ggplot(aes( factor(failures), final_grade)) + geom_boxplot() + xlab("Failures")
p4 <- df_both %>% ggplot( aes(factor(famrel), final_grade)) + geom_boxplot() + xlab("Fam Rel")
p5 <- df_both %>% ggplot( aes(factor(freetime), final_grade)) + geom_boxplot()+ xlab("Freetime")
p6 <- df_both %>% ggplot( aes(factor(parentEdu), final_grade)) + geom_boxplot() + xlab("Parent Edu")
p7 <- df_both %>% ggplot( aes(factor(alc), final_grade)) + geom_boxplot() + xlab("Alcohol")
p8 <- df_both %>% ggplot( aes(factor(school_GP), final_grade)) + geom_boxplot() + xlab("School is GP")
p9 <- df_both %>% ggplot( aes(factor(health), final_grade)) + geom_boxplot() + xlab("Health")
p10<- df_both %>% ggplot( aes(factor(paid_tutor), final_grade)) + geom_boxplot() + xlab("Paid Tutuor")
p11<- df_both %>% ggplot( aes(factor(rural), final_grade)) + geom_boxplot() + xlab("Rural")
p12<- df_both %>% ggplot( aes(factor(sex_F), final_grade)) + geom_boxplot() + xlab("Female")
p13<- df_both %>% ggplot( aes(factor(fam_large), final_grade)) + geom_boxplot() + xlab("Big Fam")
p14<- df_both %>% ggplot( aes(factor(romantic), final_grade)) + geom_boxplot() + xlab("Romantic")
p15<- df_both %>% ggplot( aes(factor(internet), final_grade)) + geom_boxplot() + xlab("Internet")
p16<- df_both %>% ggplot( aes(factor(par_apart), final_grade)) + geom_boxplot() + xlab("Parents Sep.")
p17<- df_both %>% ggplot( aes(factor(friendtime), final_grade)) + geom_boxplot() + xlab("Friend Time")
p18<- df_both %>% ggplot( aes(factor(activities), final_grade)) + geom_boxplot() + xlab("Activities")
```

```r
grid.arrange(p1, p2, p3, p4, p5, p6, p7,p8,p9,p10, p11, p12,p13, p14, p15, p16, p17, p18, ncol=6)
```



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.

```r
x <- "Portuguese Final Grade"
y <- "Math Final Grade"
final_grades_coef <- cor(df_both$G3_mat, df_both$G3_por, use="pairwise.complete.obs")
pg1 <- df_both %>% na.omit() %>% ggplot(aes(G3_mat, G3_por)) +
  geom_point(position="jitter") +
  ggtitle(
    paste(
    "Correlation between final grades in Math and Portuguese,",
    round(final_grades_coef, 2)
    )) +
  xlab(x) +
  ylab(y)


nonzeroes <- df_both %>% na.omit() %>% subset(G3_por > 0 & G3_mat >0)
final_grades_coef <- cor(nonzeroes$G3_mat, nonzeroes$G3_por, use="pairwise.complete.obs")
pg2 <- nonzeroes  %>%
  ggplot(aes(G3_mat, G3_por)) +
  geom_point(position="jitter") +
  ggtitle(
    paste(
    "Correlation between final grades omitting grades of zero,",
    round(final_grades_coef, 2)
    )) +
  xlab(x) +
  ylab(y)
grid.arrange(pg1, pg2)
```



Additional data visualizaton can be found in the Appendix.

Model Builging and Evaluation

We will attempt to predict the target, `final_grade` using three approaches: multiple linear regression, random forest, and a decision tree, and evaluate results based on validation with a test set withheld from model building.

```
set.seed(123)
# Create a list of indices for the 3 folds that can be used for all models
folds <- createFolds(df_both$final_grade, k = 3, list = TRUE, returnTrain = TRUE)
train_data <- df_both[folds[[1]], ] %>% select(-ends_with("_por"), -ends_with("_mat")) #drop unwanted f
test_data <- df_both[-folds[[1]], ]
```

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 possible 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 features within 1 SD, we will choose 3 features.

The coefficients in the best subset with 3 features are `failures`, `schoolsup` and `absences` – not dissimilar from results from previous studies. Depending on which exact split is used, other features may result in this answer, indicating that the results may not be very robust.

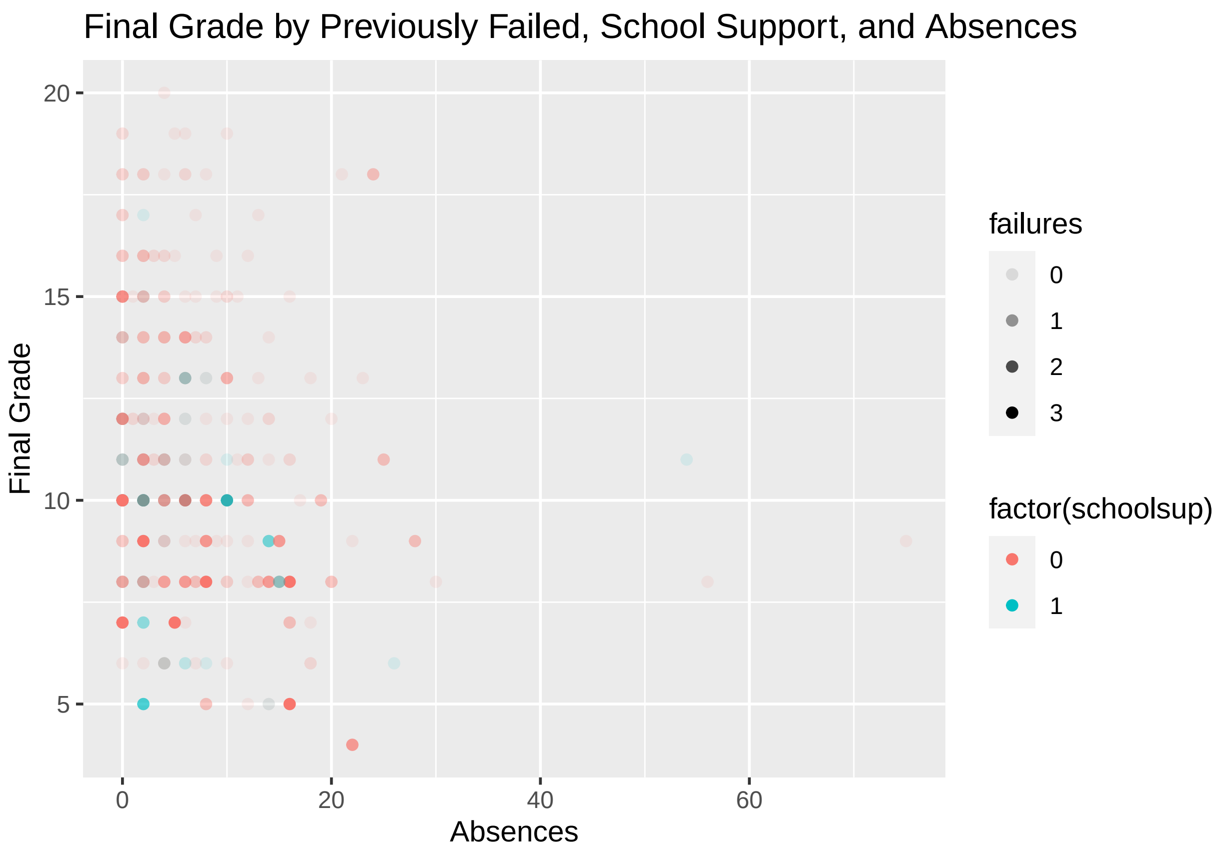
```
library(leaps)

# Specify the predictor variables you want to consider
predictors <- names(df_both %>% select(-final_grade, -ends_with("_por"), -ends_with("_mat")))  # Exclud

predictions <- c()
# Write a function to repeat with each fold
bestsubset <- function(data, pred, i) {
    # Subset data into training and testing sets based on the folds
      train_data <- data[folds[[i]], ]
      test_data <- data[-folds[[i]], ]

    # Convert the target variable to a matrix or vector ***using the fold**
    target <- as.matrix(train_data$final_grade)

    # Run the best subset selection from leaps
    subset_model <- leaps::regsubsets(target ~ ., data = train_data[, pred])

    return(subset_model)

}
bestsub <- bestsubset(df_both, predictors, 2) #using custom function from above
coef(bestsub, 3)
```

```
# Run a loop to evaluate how large the model will be using 3-fold cross validation

for(i in 1:3){
  bestsub <- bestsubset(df_both, predictors, i)
  m <- summary(bestsub)$bic
  plot(1:8, m) +
    abline(h= mean(m)-sd(m))
}
```

```r
lm_bestsub <- lm(final_grade ~ absences_mean + failures + schoolsup, data = train_data)
yhat_bestsub <- predict(lm_bestsub, test_data)
y_obs <- test_data$final_grade

runXval <- function(i){
  train_data <- df_both[folds[[i]], ]
  test_data <- df_both[-folds[[i]], ]
  lm <- lm(final_grade ~ sex_F + failures + parentEdu, data = train_data)

  predict <- predict(lm, test_data)
  r.squared <- summary(lm)$r.squared
  cor <- cor(predict, test_data$final_grade)
  metrics <- c(cor, r.squared)
  return(metrics)
}

cors <- c()
r.sq <- c()

for (i in 1:3){
  cors_bestsub <- c(cors, runXval(i)[[1]])
  r.sq_bestsub <- c(r.sq, runXval(i)[[2]])
}

mean(cors_bestsub)
mean(r.sq_bestsub)
cor_bestsub <- cor(yhat_bestsub, y_obs)

#plot(y_obs, yhat_bestsub)
```

Using a 3-fold validation, the mean correlation between predicted and actual final grades is 0.34 and the R squared is 0%.

```r
df_both %>%
  ggplot(aes(y=final_grade, x=absences_mean, alpha=failures, color=factor(schoolsup))) +
  geom_point() +
  xlab("Absences")+
  ylab("Final Grade") +
  ggtitle("Final Grade by Previously Failed, School Support, and Absences")
```



```
stats::aov(final_grade ~ failures + absences_mean + schoolsup, df_both) %>%
summary()
```

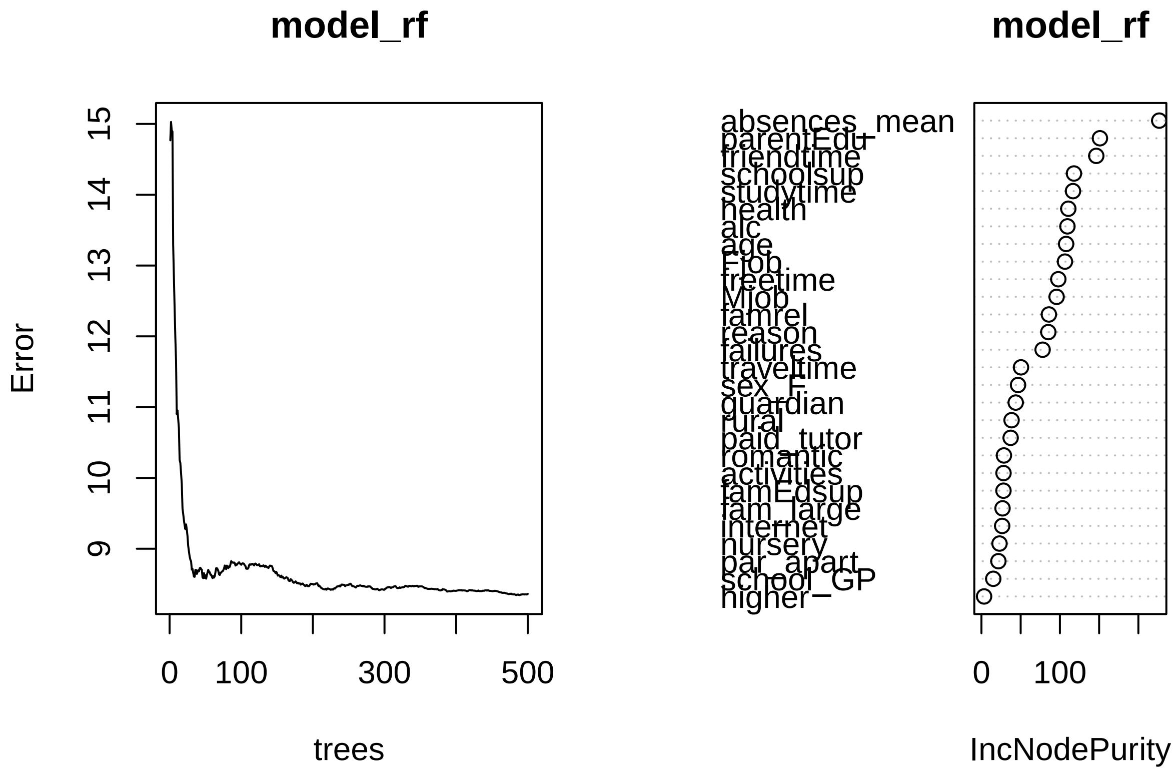
We can also attempt to use stepwise regression to see if there is a different result. Interestingly, this results in a different and much larger model – perhaps because it is judged on AIC rather than BIC, which penalizes the inclusion of additional features more.

In fact the result is very similar with a correlation between test and predicted results of 0.3715614 and multiple R-squared of 0.31. RMSE is close but higher at 3.312686compared to 3.107402

Model 2

Random Forest Model

```
##
## Call:
##  randomForest(formula = final_grade ~ ., data = train_data)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 9
##
##           Mean of squared residuals: 8.361846
##                     % Var explained: 17.59
```

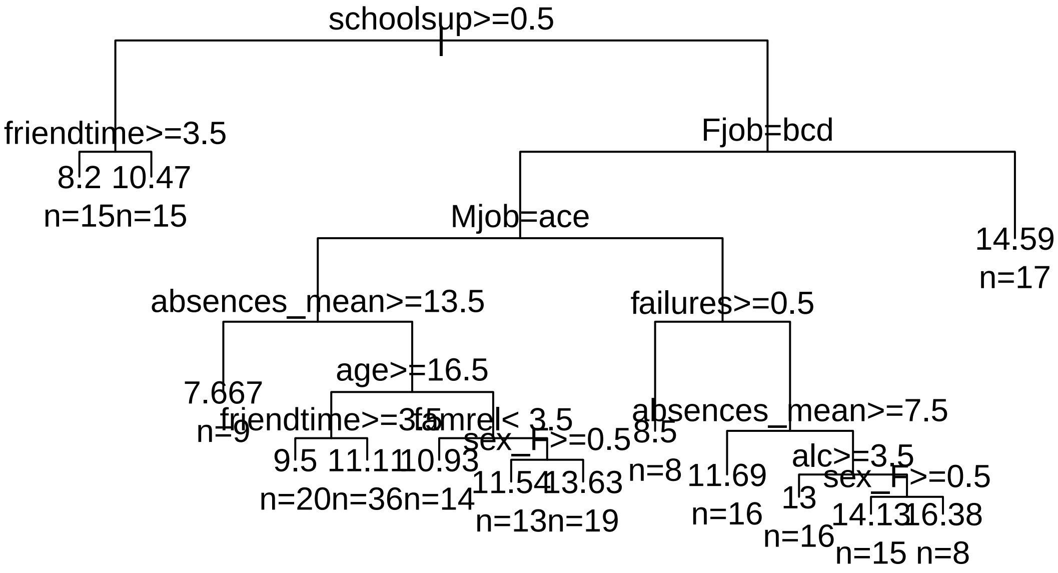


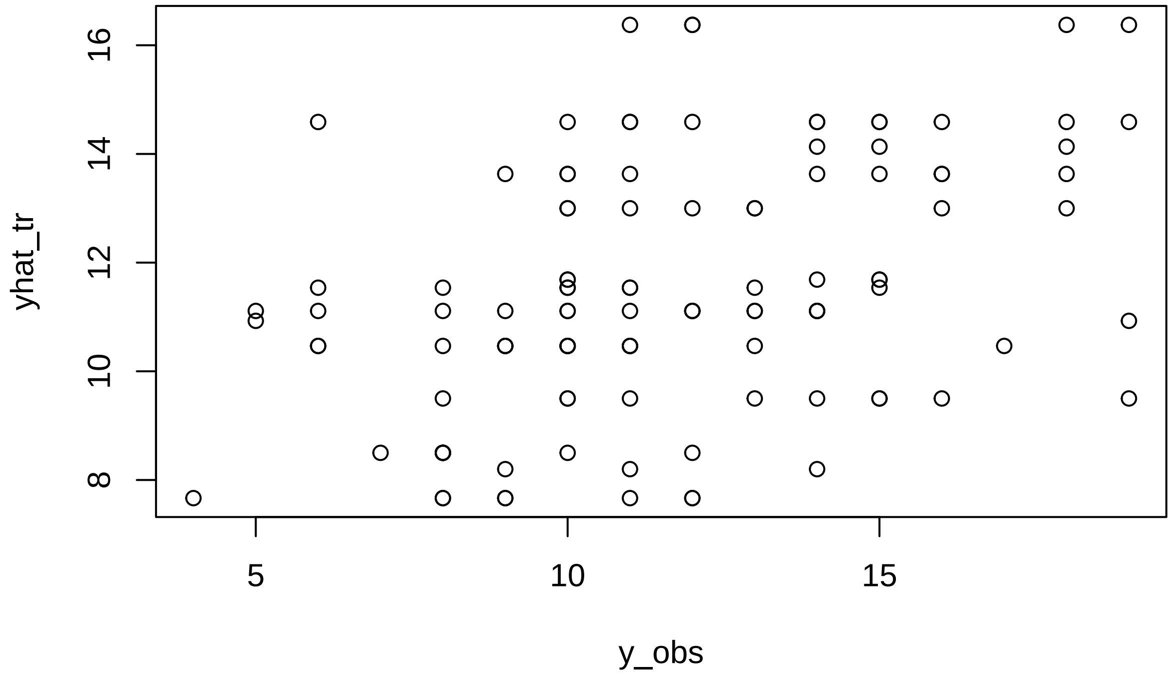
```
## null device
##           1
```

Model 3

Decision tree model

```
## n= 221
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##   1) root 221 2242.42500 11.561090
##     2) schoolsup>=0.5 30  186.66670  9.333333
##       4) friendtime>=3.5 15   80.40000  8.200000 *
##       5) friendtime< 3.5 15   67.73333 10.466670 *
##     3) schoolsup< 0.5 191 1883.48700 11.910990
##       6) Fjob=health,other,services 174 1623.61500 11.649430
##        12) Mjob=at_home,other,teacher 111  842.00000 11.000000
##          24) absences_mean>=13.5 9   28.00000  7.666667 *
##          25) absences_mean< 13.5 102  705.17650 11.294120
##            50) age>=16.5 56  327.92860 10.535710
##             100) friendtime>=3.5 20  131.00000  9.500000 *
##             101) friendtime< 3.5 36  163.55560 11.111110 *
##            51) age< 16.5 46  305.82610 12.217390
##             102) famrel< 3.5 14   94.92857 10.928570 *
##             103) famrel>=3.5 32  177.46880 12.781250
##               206) sex_F>=0.5 13   69.23077 11.538460 *
##               207) sex_F< 0.5 19   74.42105 13.631580 *
##        13) Mjob=health,services 63  652.31750 12.793650
##          26) failures>=0.5 8   20.00000  8.500000 *
##          27) failures< 0.5 55  463.38180 13.418180
##            54) absences_mean>=7.5 16  123.43750 11.687500 *
##            55) absences_mean< 7.5 39  272.35900 14.128210
##             110) alc>=3.5 16   90.00000 13.000000 *
##             111) alc< 3.5 23  147.82610 14.913040
##               222) sex_F>=0.5 15   87.73333 14.133330 *
##               223) sex_F< 0.5 8   33.87500 16.375000 *
##       7) Fjob=at_home,teacher 17  126.11760 14.588240 *
```



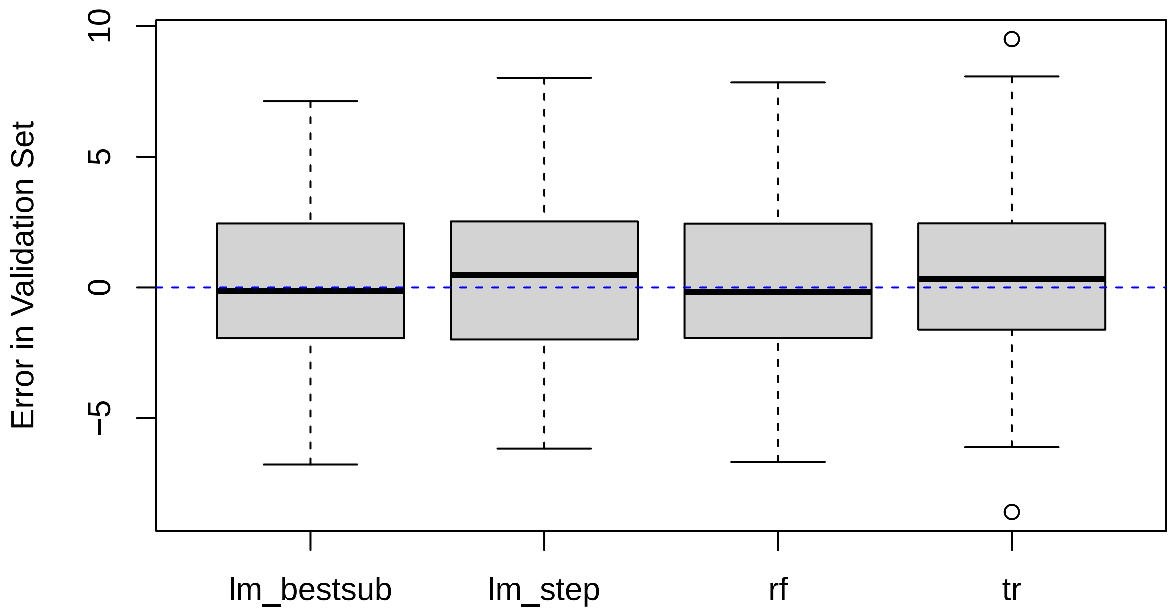


In a complex model, a decision tree can be effective. This model results in correlation between predicted and actual final grades of 0.4144219 as well as an RMSE of 3.244109.

Model Selection

The best model is not one with the highest R-squared, but the one that performs the best on test data. In fact, all of these models performed very similarly.

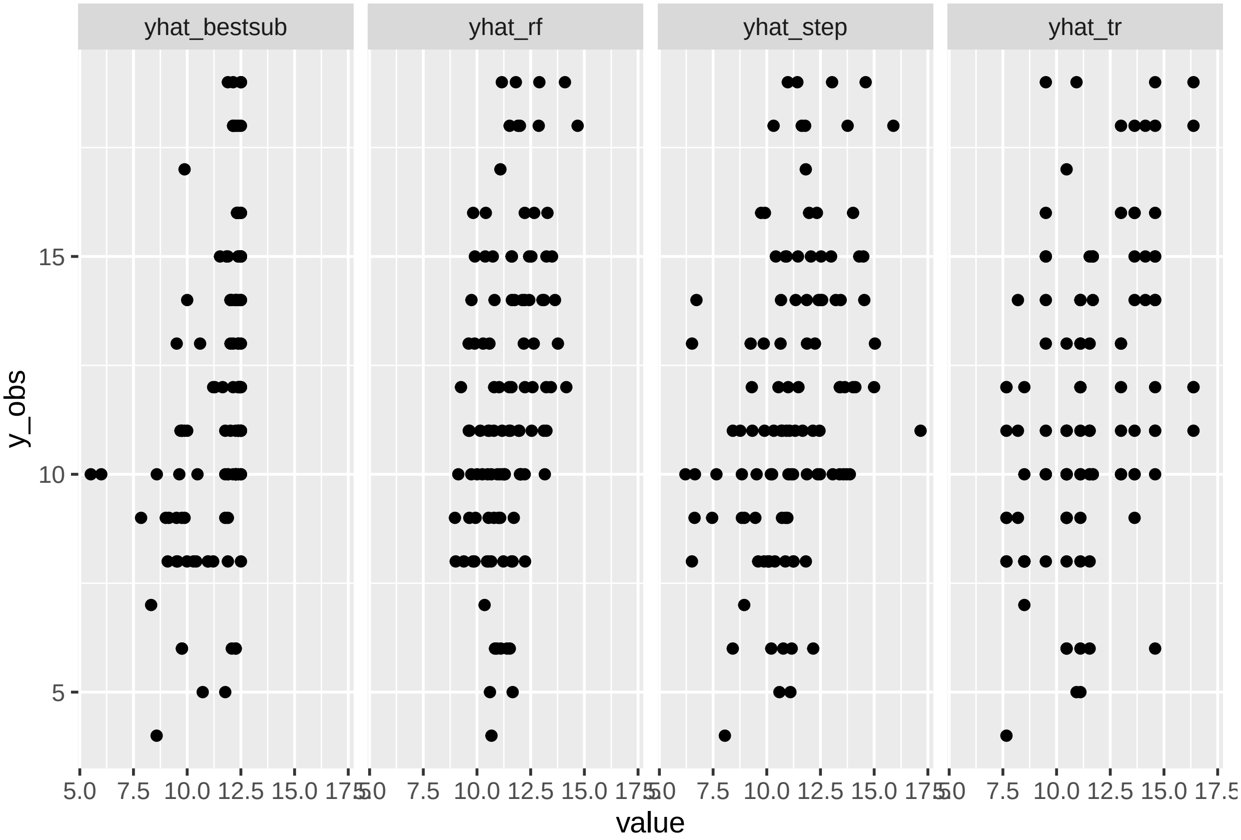
```
boxplot(list(lm_bestsub = y_obs - yhat_bestsub,
             lm_step = y_obs-yhat_step,
             rf = y_obs-yhat_rf,
             tr = y_obs - yhat_tr),
        ylab="Error in Validation Set",
        title = "Error by Model")

abline(h=0, lty=2, col='blue')
```



```
results <- tibble(y_obs, yhat_bestsub, yhat_step, yhat_tr, yhat_rf)

Models <- c("Best Subset", "Stepwise Selection", "Random Forrest", "Decision Tree")
Correlation <- c(cor_bestsub, cor_step, cor_rf, cor_tr)
RMSE <- c(rmse_bestsub, rmse_step, rmse_rf, rmse_tr)

compare <- tibble(Models, Correlation, RMSE)

compare
```

```
## # A tibble: 4 x 3
##   Models             Correlation  RMSE
##   <chr>                    <dbl> <dbl>
## 1 Best Subset              0.424  3.11
## 2 Stepwise Selection       0.372  3.31
## 3 Random Forrest           0.391  3.16
## 4 Decision Tree            0.414  3.24
```

```
results %>% gather(key, value, -y_obs) %>% ggplot(aes(x=value, y=y_obs)) + geom_point() + facet_grid(~ke
```



Conclusion

The most significant predictors in the Best Subset model for the students final grade turns out to be having previously failed a class, presence or absence of school support, and absences. This is an interesting and actionable set of 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 by a hair in terms of correlation between predicted and actual results, is the Random Forest, but the differences are in fact extremely subtle and unlikely to be replicated exactly. Decision Tree and Random Forest modeling are harder to interpret or explain.

Ultimately, the choice of the best model depends upon the goal of the study. For a policy maker, the simplest linear model may be preferred as it results in the clearest and most defensible policy prescriptions: * Intervene with students who have previously failed * Offer school support * Intervene to prevent or mitigate absences

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)
- *alc* - combined `Dalc` + `Walc`

- *age* - student’s age (numeric: from 15 to 22)
- *Dalc* - workday alcohol consumption (numeric: from 1 - very low to 5 - very high) *removed
- *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) *removed
- *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) *removed
- *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* - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) *removed

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`

```
descrip
```

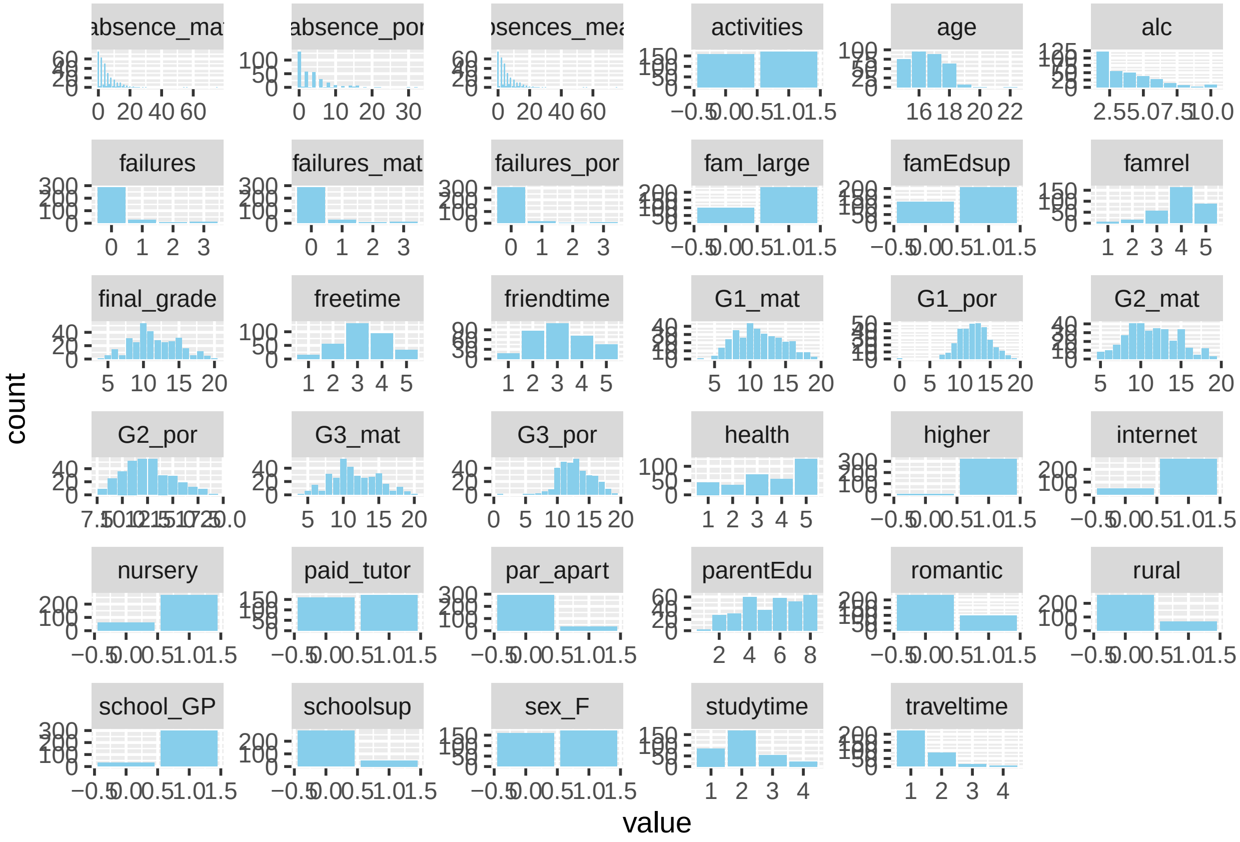
```
##                vars   n  mean   sd median trimmed  mad min max range  skew
## school_GP         1 332  0.90 0.30      1    1.00 0.00   0   1     1 -2.72
## sex_F            2 332  0.52 0.50      1    0.52 0.00   0   1     1 -0.07
## age              3 332 16.52 1.16     16   16.47 1.48  15  22     7  0.50
## rural            4 332  0.21 0.41      0    0.14 0.00   0   1     1  1.43
## fam_large        5 332  0.70 0.46      1    0.75 0.00   0   1     1 -0.88
## par_apart        6 332  0.11 0.31      0    0.01 0.00   0   1     1  2.51
## Mjob*            7 332  3.22 1.22      3    3.27 1.48   1   5     4 -0.33
## Fjob*            8 332  3.30 0.84      3    3.32 0.00   1   5     4 -0.24
## reason*          9 332  2.30 1.22      2    2.25 1.48   1   4     3  0.35
## guardian*       10 332  1.80 0.48      2    1.83 0.00   1   3     2 -0.47
## traveltime      11 332  1.42 0.68      1    1.29 0.00   1   4     3  1.73
## studytime       12 332  2.05 0.84      2    1.97 0.00   1   4     3  0.63
## failures_mat    13 332  0.21 0.63      0    0.04 0.00   0   3     3  3.27
## schoolsup       14 332  0.14 0.35      0    0.06 0.00   0   1     1  2.01
## famEdsup        15 332  0.63 0.48      1    0.66 0.00   0   1     1 -0.52
## paid_mat*       16 332  1.49 0.50      1    1.49 0.00   1   2     1  0.02
## activities      17 332  0.52 0.50      1    0.52 0.00   0   1     1 -0.07
## nursery         18 332  0.82 0.39      1    0.89 0.00   0   1     1 -1.63
## higher          19 332  0.97 0.17      1    1.00 0.00   0   1     1 -5.47
## internet        20 332  0.85 0.36      1    0.94 0.00   0   1     1 -1.94
## romantic        21 332  0.30 0.46      0    0.24 0.00   0   1     1  0.89
## famrel          22 332  3.94 0.90      4    4.04 1.48   1   5     4 -0.96
## freetime        23 332  3.23 1.00      3    3.22 1.48   1   5     4 -0.13
## friendtime      24 332  3.11 1.11      3    3.09 1.48   1   5     4  0.13
## health          25 332  3.55 1.42      4    3.69 1.48   1   5     4 -0.52
## absence_mat     26 332  5.91 7.87      4    4.55 5.93   0  75    75  3.98
## G1_mat          27 332 11.27 3.26     11   11.18 3.71   3  19    16  0.21
## G2_mat          28 332 11.42 3.19     11   11.35 2.97   5  19    14  0.19
## G3_mat          29 332 11.59 3.27     11   11.53 2.97   4  20    16  0.20
## failures_por    30 332  0.12 0.49      0    0.00 0.00   0   3     3  4.60
## paid_por*       31 332  1.07 0.25      1    1.00 0.00   1   2     1  3.38
## absence_por     32 332  3.48 4.70      2    2.54 2.97   0  32    32  2.33
## G1_por          33 332 12.34 2.51     12   12.32 2.97   0  19    19 -0.20
## G2_por          34 332 12.49 2.39     12   12.40 2.97   8  19    11  0.36
## G3_por          35 332 12.85 2.57     13   12.80 2.97   1  19    18 -0.15
## absences_mean   36 332  5.91 7.87      4    4.55 5.93   0  75    75  3.98
## final_grade     37 332 11.59 3.27     11   11.53 2.97   4  20    16  0.20
## failures        38 332  0.21 0.63      0    0.04 0.00   0   3     3  3.27
## paid_tutor      39 332  0.52 0.50      1    0.53 0.00   0   1     1 -0.08
## parentEdu       40 332  5.42 1.94      6    5.52 2.97   1   8     7 -0.24
## alc             41 332  3.83 2.02      3    3.50 1.48   2  10     8  1.16
##               kurtosis   se
## school_GP         5.43 0.02
## sex_F            -2.00 0.03
## age               0.42 0.06
## rural             0.06 0.02
## fam_large        -1.23 0.03
## par_apart         4.30 0.02
## Mjob*            -0.65 0.07
## Fjob*             0.96 0.05
## reason*          -1.46 0.07
## guardian*         0.20 0.03
## traveltime        2.89 0.04
## studytime        -0.03 0.05
## failures_mat     10.41 0.03
## schoolsup          2.06 0.02
## famEdsup          -1.73 0.03
## paid_mat*         -2.01 0.03
## activities        -2.00 0.03
## nursery            0.65 0.02
## higher            28.04 0.01
## internet           1.79 0.02
## romantic          -1.20 0.03
## famrel             1.11 0.05
## freetime          -0.36 0.05
## friendtime        -0.78 0.06
## health            -1.05 0.08
## absence_mat       25.44 0.43
## G1_mat            -0.68 0.18
## G2_mat            -0.58 0.18
## G3_mat            -0.49 0.18
## failures_por      21.78 0.03
## paid_por*          9.43 0.01
## absence_por        7.77 0.26
## G1_por             1.01 0.14
## G2_por            -0.41 0.13
## G3_por             0.74 0.14
## absences_mean     25.44 0.43
## final_grade       -0.49 0.18
## failures          10.41 0.03
## paid_tutor        -2.00 0.03
## parentEdu         -1.09 0.11
## alc                0.84 0.11
```

Appendix 3: Basic Plots

```
basicp <- function(df, color){
df_long <- df %>% select_if(is.numeric)
p <- df_long %>% gather() %>% ggplot(aes(x= value)) + geom_histogram(stat = "count", fill = color) + fac
print(p)
}

basicp (df_both, "skyblue")
```

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

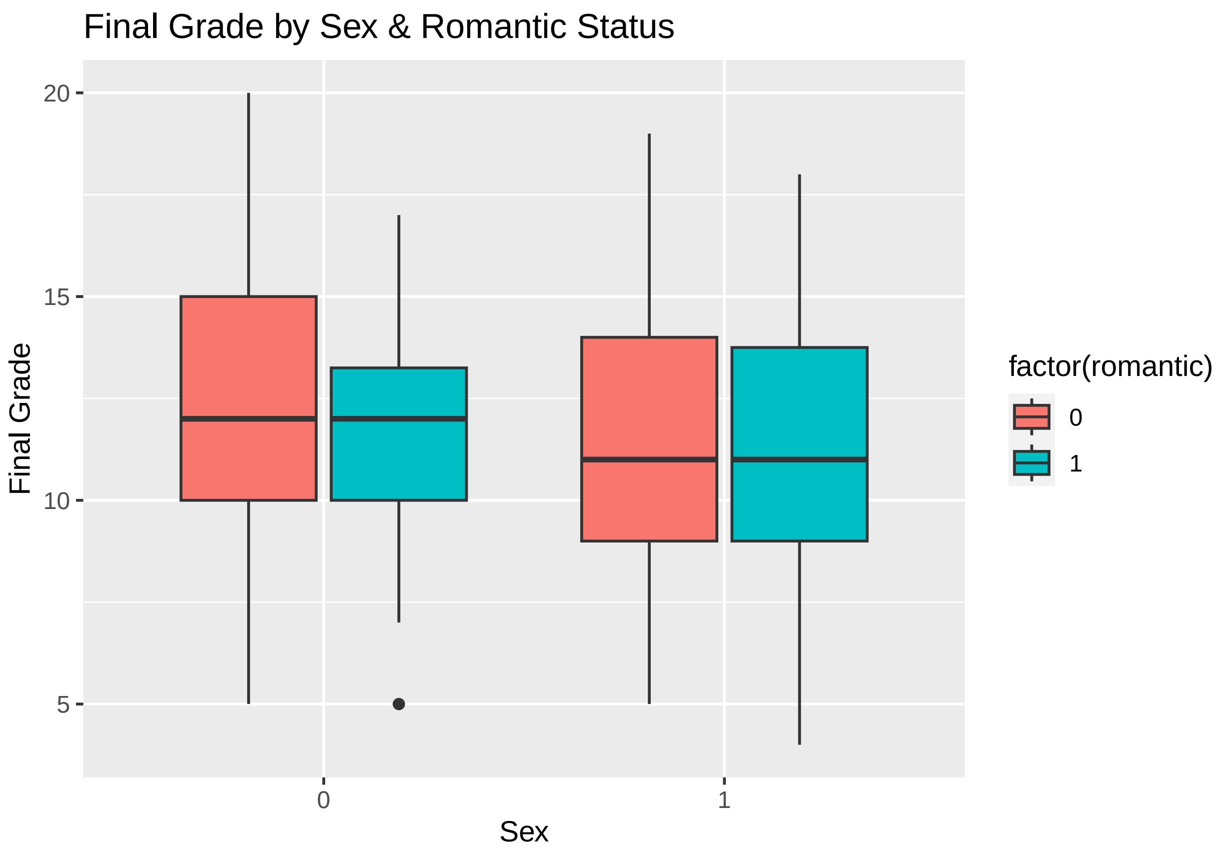
Authors: Marie-Maude Dubuc, Mylene Aubertin-Leheudr, and Antony D. Karelis Published online 2022 May 1, International Journal of Exercise https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9365103/#:~:text=Finally%2C%20psychological%20factors%20such%20as,42%2C%2043%2C%2048)

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

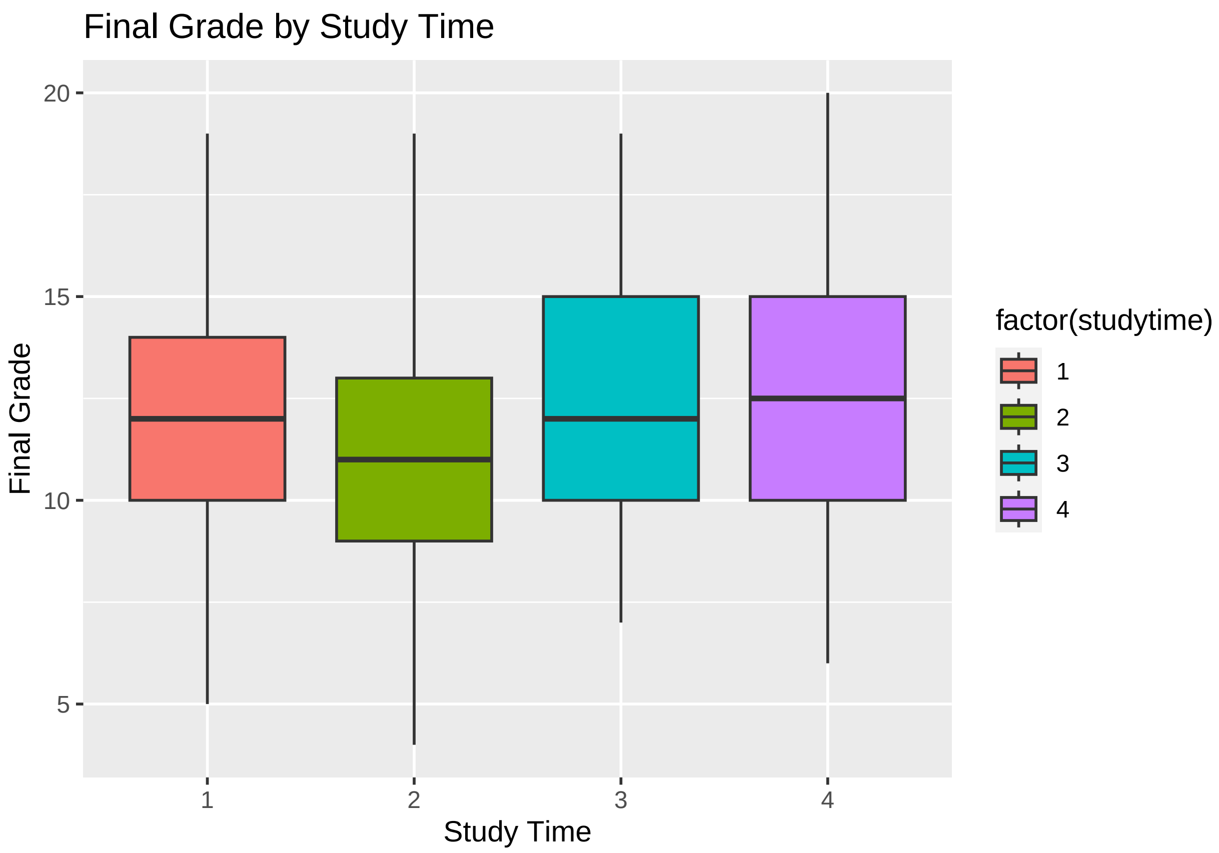
In general, students who are in a romantic relationship may have a more challenging time focusing on their studies compared to those who are not. A plot indicates that this may be the case – especially male students who are not romantically involved appear to have a slightly higher final grade; ANOVA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



```
stats::aov(final_grade ~ romantic + sex_F, df_both) %>%
summary()
```

Study Time

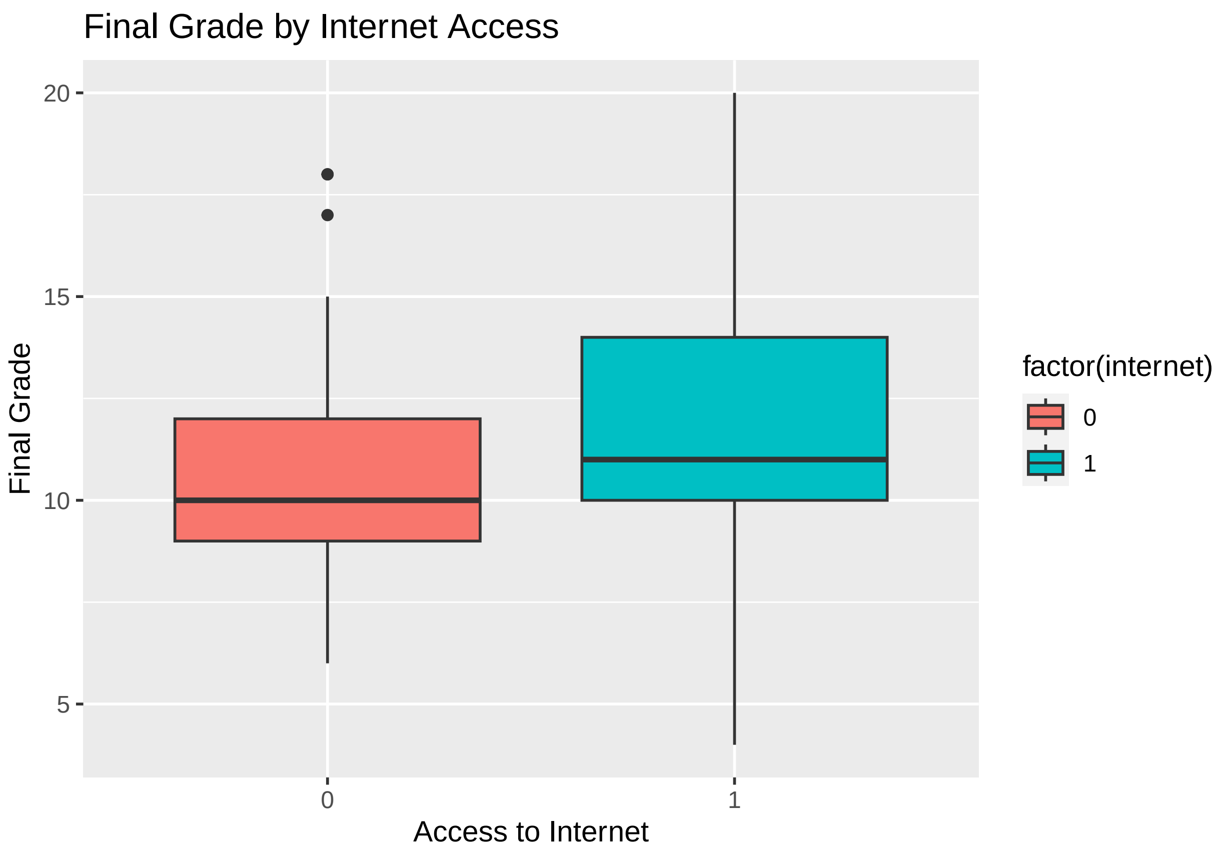
Although at first glance the findings suggest that students who study for a longer duration have a higher average score compared to those who study for a shorter duration, study time is not predictive according to ANOVA and we cannot rule out chance.



```
stats::aov(final_grade ~ studytime, df_both) %>%
summary()
```

Internet

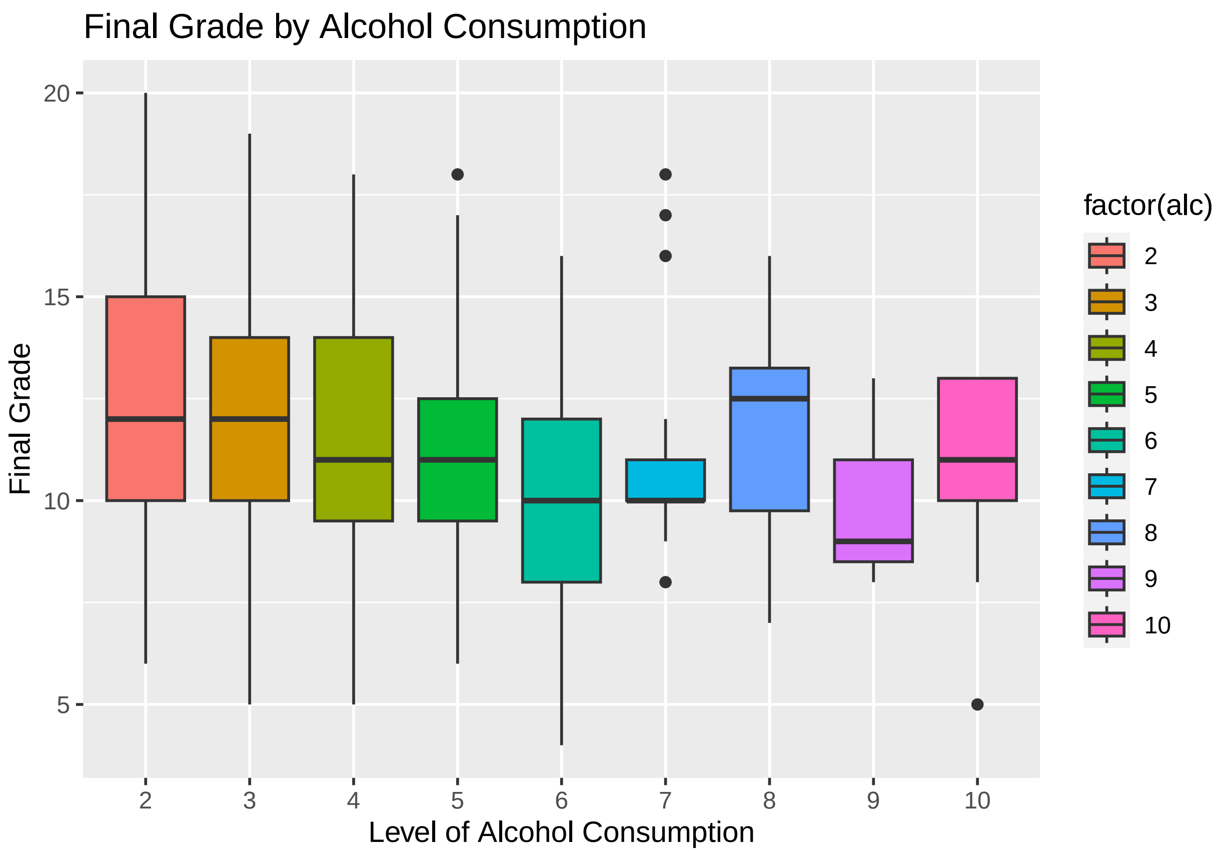
The following graph indicates that the difference in academic performance between students with and without Internet access is negligible; again, however, ANOVA results indicate that we cannot rule out the null hypothesis that chance may explain the difference in means between the two groups.



```
stats::aov(final_grade ~ internet, df_both) %>%
summary()
```

Alcohol

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.



```
stats::aov(final_grade ~ alc, df_both) %>%
summary()
```