BST 260 Final Report

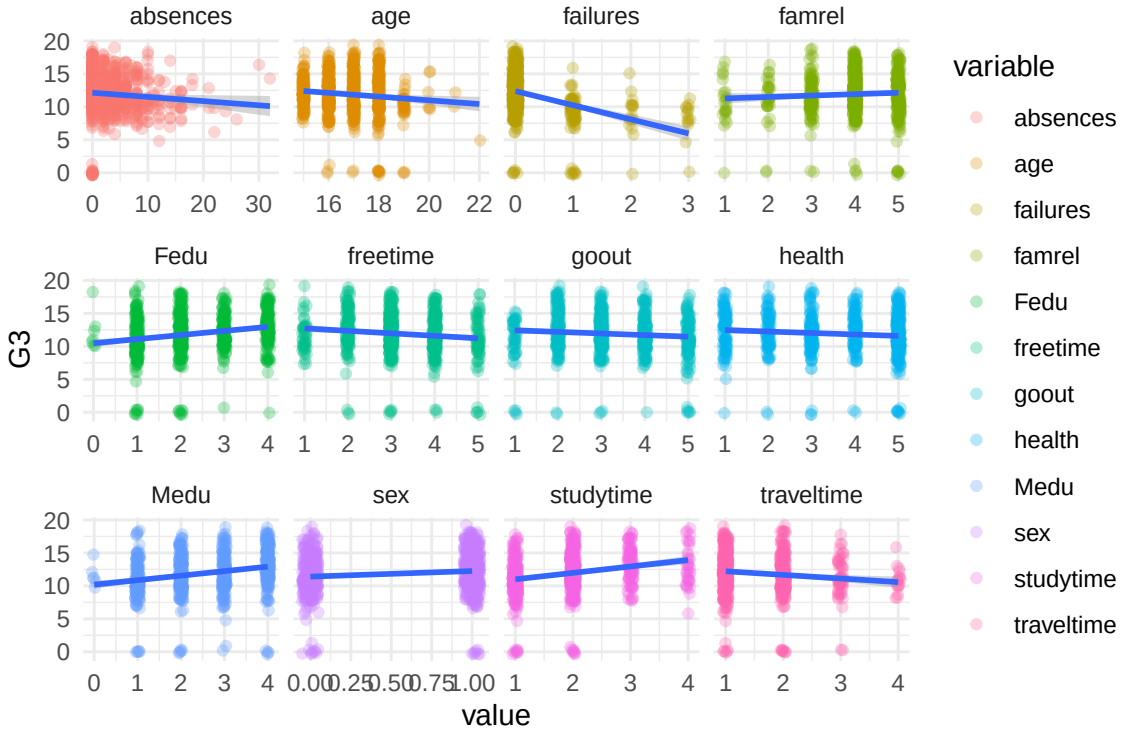
Rebecca Hurwitz

2022-12-14

Introduction

Various factors affect a student's ability to succeed in school, ranging from study time to family situations to community influences to health, and so on. For this BST 260 Final Project, I aimed to use data science to predict Portuguese high school students' final grades based on various factors in their lives. The grading scale in Portugal differs from that of the US – using a scale of 0 – 20, rather than letter grades, we can observe a more precise measurement of performance (*"The Portuguese Grading System", 2022*).

The dataset I used for this project is from UC Irvine's Machine Learning Repository (*Yilmaz and Sekeroglu*). It consists of 649 high school students in Portugal with 33 attributes each, acquired via school reports and questionnaires. Some particular variables of interest, which I used to explore students' final grade in Portuguese class, were weekly hours spent studying, age, sex, mother's education, father's education, quality of family relationships, hours of free time per week, previous classes failed, and current health status. The first facet-wrap plot (Figure 1, larger in Appendix) outlines some exploratory data analysis I completed using linear regression to get a better understanding of the data. The y-axis (G3: Final Grade) ranges from 0 (fail) to 20 (excellent) and is the same for each covariate, while the x-axis varies by covariate. Also included is a "table1" summary of all of the extracted data (Figure 2).



While in a perfect world, every child would experience a level playing field and be born with equal opportunities to succeed in school, we can see here that factors outside of a student's control can contribute to their performance in an academic setting. This EDA led me to solidify my research question: Is there a significant association between time spent studying per week and final grades, and what other external factors can affect this relationship?

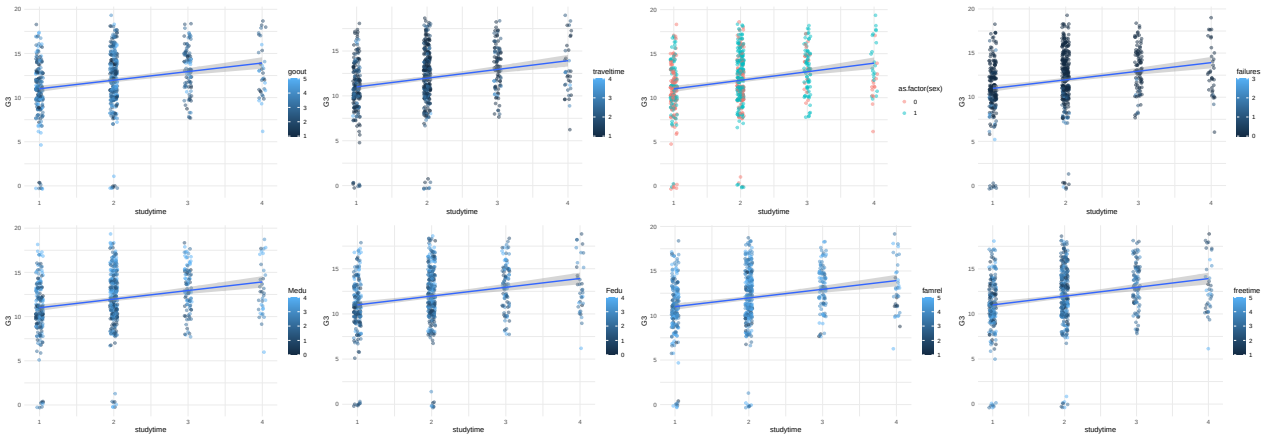
This plot piqued my interest and prompted me to further explore the data by conducting two major analyses: fitting linear models and conducting random forest machine learning. I chose linear modeling because of the vast number of attributes (33) – I believe that linearly modeling these would be informative because we can see the effects of each attribute. Next, I chose random forest machine learning because I have never learned about it before BST 260, and really wanted to challenge myself by applying this technique. Random forest also is a very sound way to parse through a lot of data due to our many covariates, and perhaps improve prediction performance by reducing instability in models (*Irizarry, "Chapter 31 Examples of Algorithms | Introduction to Data Science"*).

Results

Linear Modeling:

To begin the process of linear modeling, I conducted analyses to ensure that my data met the LINE assumptions (linearity, independence, normality, and equal variance (homoscedasticity))(*Lake, Topic 2 | Applied Linear Regression"*). Via a QQ-plot and a histogram (Figures 3 and 4), I verified that these assumptions were approximately met, confirming that the normal approximation was useful in this case (*Irizarry, "Chapter 18 Linear Models | Introduction to Data Science"*). Starting with a very big model (as seen in Figure 5), I then used forward model selection to identify important and relevant covariates.

Looking at the significance of these covariates, I moved to simplify my model by eliminating variables either not of interest or not deemed significant (p-value > 0.05), including age, family size, address, guardian, parental marital status, travel time to school, and days spent going out. Next, because we know that association is not causation, it was critical to analyze potential confounding variables. I used both the classical definition (via DAGs) and the statistical definition (via the 10% rule) to determine that none of my variables of interest were confounding, and thus, we do not see any effect modification either. In fitting these models, I created various plots to visually stratify each variable to further assess for confounding, of which I found none (Figures 7 - 15, tiled here but larger in Appendix).



For this analysis, I started with a basic model looking at the base question: the potential for an association between study time per week and final course grade. From there, I created 6 linear models to assess the other covariates for confounding. Next, I then used forward selection to create 4 more models, adding in each covariate, to determine the best model based on AIC and Adjusted R^2 value. I dug deeper by trying a quadratic model next, followed by a logistic model, neither of which had a lower AIC or higher Adjusted

R^2 value than the current champion, Model 8. Thus, of these thirteen models, Model 8, as seen below, had the strongest fit (Figure 6). Because of the many covariates, visualizing these models simultaneously in two dimensions proved difficult. Figure 16 plots all thirteen models on a study time versus G3 (final grade) plot, demonstrating the indecipherability and drawbacks of selecting a model visually, instead of using the above diagnostics I opted to utilize.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	9.704	0.737	13.17	3.252e-35	* * *
studytime	0.6325	0.1392	4.543	6.623e-06	* * *
sex	0.5213	0.2382	2.189	0.02898	*
Medu	0.359	0.1305	2.751	0.006102	* *
Fedu	0.2315	0.1334	1.735	0.08314	
famrel	0.212	0.1186	1.787	0.07437	
failures	-1.716	0.1946	-8.814	1.13e-17	* * *
freetime	-0.2015	0.1087	-1.855	0.06406	
health	-0.1627	0.07836	-2.076	0.03825	*

Further answering the research question, we can see that all of the above covariates have an effect (positive for increased study time, sex (if female), greater years of parents' education, higher quality family relationships; negative for having more free time, having previously failed more classes, and interestingly, reporting better health) on our outcome, final grade.

As mentioned, I found the best model to be Model 8 (Figure 6, below), which included the covariates *studytime*, *sex*, *Medu*, *Fedu*, *famrel*, *freetime*, *failures*, and *health* to have both the highest Adjusted R^2 at 0.2314 and the lowest AIC at 3204.031. One thing to note that this is a low R^2, which is not ideal, but is the best one I got with this data. This answers another part of our question, as clearly, hours per week spent studying is a statistically significant predictor of final course grade according to these data. We can interpret this relationship in the following manner:

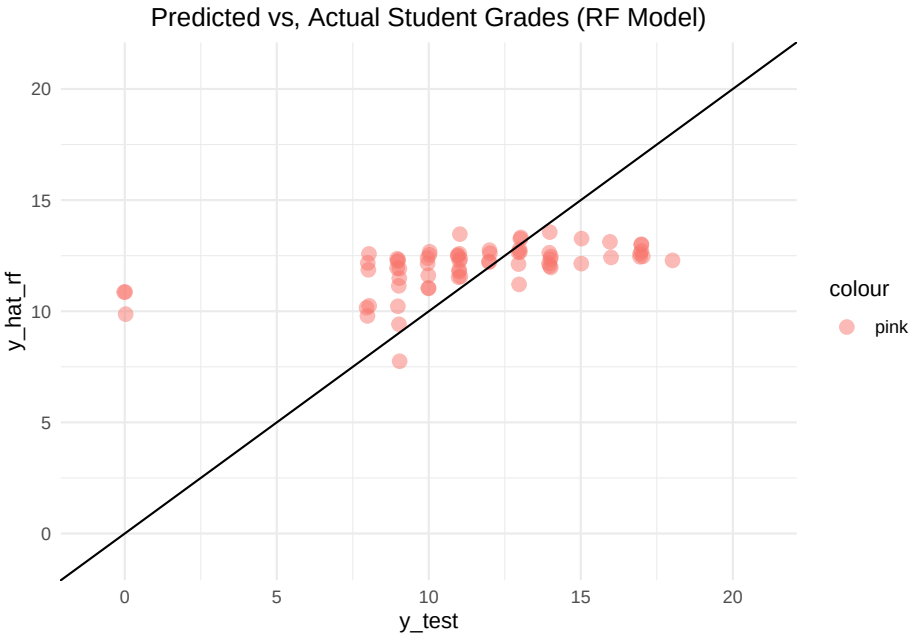
For every 1-hour increase in hours spent studying per week, final course grade increases by 0.6325 points (of 20) on average, according to these data, holding sex, mother's education, father's education, quality of family relationships, previous class failures, free time, and health constant.

Random Forest:

More advanced than decision trees, random forests utilize bootstrapping as well as the selection of random features to be built into in each tree (*Irizarry, "Chapter 31 Examples of Algorithms | Introduction to Data Science"*). This method fits very well with my project because my dataset has a large number of features, and I want to be sure not to overfit my model.

Using a 90/10 split create train and test sets, I built a random forest model with 10-fold cross validation to predict G3 final grades, using only the covariates in my model of best fit, Model 8. I minimized my node size and I chose 75 trees as the error rate seemed to stabilize between 60 and 80 trees (see Error Rate vs. Trees Plot, Figure 17). I chose 1000 as my sampling number as my sample size was only 649. Figure 18, below, provides a graph of my predicted values against the students' actual test grades. I arrived at an RMSE of 3.36698, meaning that my final predictions were on average, 3.37 grade points away from the real final grades. While this model isn't 100% perfect, the fact that it is less than 3.5 points off from a total of 20 possible grade points still presents valuable insights.

In addition, we can see via the alpha shading that there appear to be slightly more points above the y = x line than below it, indicating that my model is slightly overestimating final grades. Interestingly, to the far left, it is clear that the test dataset included three 0 grades, but the predicted dataset did not, further confirming the overestimating nature of the model.



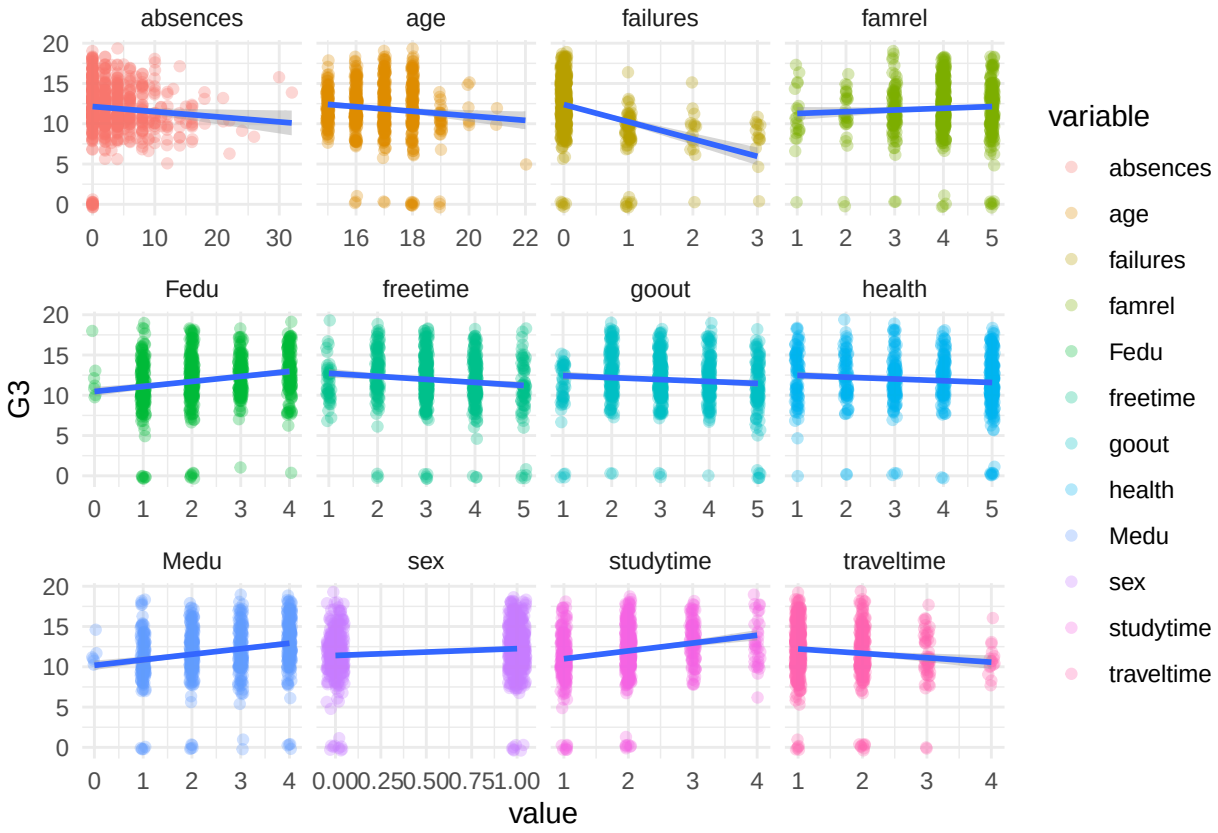
Conclusion

The purpose of this project was to see if it was possible to predict final grades of high school students in Portugal using various aspects of their lives. In order to explore this question, I delved deep into data analysis as I learned to do this semester in BST 260. I performed exploratory data analysis, fit 13 linear models, and created a random forest model. While the adjusted R^2 of the linear model was low and the RMSE of the random forest was a bit high, it is important to note that there is still statistical power in these results. The models are somewhat successful in answering the research question, although it is clear that they can be improved upon with more time and resources.

My final models were limited in their power to predict final grades, but they do provide some insight into the great importance of prioritizing study time and maintaining positive family relationships, among other study habits. Future analyses should include more students over a longer period of time, collecting additional information and variables, and applying more advanced machine learning techniques to make the next models more accurate. With more information and improved models, it may possible to further elucidate what factors affect students' learning and advocate for local, national, and global changes to improve the academic performance of students around the world.

Figures and Tables

Figure 1: Facet-Wrapped Linear Regression Plots of Twelve Covariates vs. G3 Final Grades



	Top Quintile (17-20)	Fourth Quintile (13-16)	Third Quintile (9-12)	Second Quintile (5-8)	Bottom Quintile (1-4)	Overall
	(N=46)	(N=230)	(N=308)	(N=49)	(N=16)	(N=649)
factor(sex)						
0	13 (28.3%)	80 (34.8%)	142 (46.1%)	22 (44.9%)	9 (56.3%)	266 (41.0%)
1	33 (71.7%)	150 (65.2%)	166 (53.9%)	27 (55.1%)	7 (43.8%)	383 (59.0%)
age						
Mean (SD)	17.1 (0.849)	16.6 (1.10)	16.7 (1.32)	17.0 (1.22)	17.7 (1.01)	16.7 (1.22)
Median [Min, Max]	17.0 [15.0, 18.0]	17.0 [15.0, 20.0]	17.0 [15.0, 21.0]	17.0 [15.0, 22.0]	18.0 [16.0, 19.0]	17.0 [15.0, 22.0]
famsize						
GT3	31 (67.4%)	168 (73.0%)	207 (67.2%)	37 (75.5%)	14 (87.5%)	457 (70.4%)
LE3	15 (32.6%)	62 (27.0%)	101 (32.8%)	12 (24.5%)	2 (12.5%)	192 (29.6%)
Pstatus						
Mean (SD)	0.891 (0.315)	0.865 (0.342)	0.886 (0.318)	0.857 (0.354)	0.875 (0.342)	0.877 (0.329)
Median [Min, Max]	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]
Medu						
Mean (SD)	3.17 (1.04)	2.77 (1.10)	2.27 (1.10)	2.29 (1.06)	2.31 (1.20)	2.51 (1.13)
Median [Min, Max]	4.00 [1.00, 4.00]	3.00 [0, 4.00]	2.00 [0, 4.00]	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]	2.00 [0, 4.00]
Fedu						
Mean (SD)	2.63 (1.10)	2.54 (1.10)	2.13 (1.06)	2.16 (1.16)	1.75 (0.856)	2.31 (1.10)
Median [Min, Max]	2.00 [0, 4.00]	2.50 [0, 4.00]	2.00 [0, 4.00]	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]	2.00 [0, 4.00]
Mjob						
at_home	8 (17.4%)	28 (12.2%)	84 (27.3%)	11 (22.4%)	4 (25.0%)	135 (20.8%)
health	7 (15.2%)	21 (9.1%)	16 (5.2%)	4 (8.2%)	0 (0%)	48 (7.4%)
other	11 (23.9%)	97 (42.2%)	121 (39.3%)	20 (40.8%)	9 (56.3%)	258 (39.8%)
services	9 (19.6%)	52 (22.6%)	61 (19.8%)	13 (26.5%)	1 (6.3%)	136 (21.0%)
teacher	11 (23.9%)	32 (13.9%)	26 (8.4%)	1 (2.0%)	2 (12.5%)	72 (11.1%)
Fjob						
at_home	2 (4.3%)	14 (6.1%)	20 (6.5%)	5 (10.2%)	1 (6.3%)	42 (6.5%)
health	3 (6.5%)	8 (3.5%)	10 (3.2%)	2 (4.1%)	0 (0%)	23 (3.5%)
other	21 (45.7%)	131 (57.0%)	185 (60.1%)	22 (44.9%)	8 (50.0%)	367 (56.5%)
services	14 (30.4%)	57 (24.8%)	85 (27.6%)	19 (38.8%)	6 (37.5%)	181 (27.9%)
teacher	6 (13.0%)	20 (8.7%)	8 (2.6%)	1 (2.0%)	1 (6.3%)	36 (5.5%)
guardian						
father	11 (23.9%)	57 (24.8%)	72 (23.4%)	9 (18.4%)	4 (25.0%)	153 (23.6%)
mother	35 (76.1%)	164 (71.3%)	208 (67.5%)	37 (75.5%)	11 (68.8%)	455 (70.1%)
other	0 (0%)	9 (3.9%)	28 (9.1%)	3 (6.1%)	1 (6.3%)	41 (6.3%)
traveltime						
Mean (SD)	1.33 (0.519)	1.48 (0.698)	1.67 (0.792)	1.53 (0.819)	1.75 (0.683)	1.57 (0.749)
Median [Min, Max]	1.00 [1.00, 3.00]	1.00 [1.00, 4.00]	1.50 [1.00, 4.00]	1.00 [1.00, 4.00]	2.00 [1.00, 3.00]	1.00 [1.00, 4.00]
studytime						
Mean (SD)	2.39 (0.930)	2.10 (0.789)	1.80 (0.819)	1.69 (0.742)	1.50 (0.516)	1.93 (0.830)
Median [Min, Max]	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]	1.50 [1.00, 2.00]	2.00 [1.00, 4.00]
failures						
Mean (SD)	0 (0)	0.0261 (0.185)	0.260 (0.618)	0.918 (1.04)	0.813 (0.834)	0.222 (0.593)
Median [Min, Max]	0 [0, 0]	0 [0, 2.00]	0 [0, 3.00]	1.00 [0, 3.00]	1.00 [0, 3.00]	0 [0, 3.00]
famrel						
Mean (SD)	4.13 (0.778)	4.01 (0.872)	3.85 (0.983)	3.88 (1.15)	3.88 (1.31)	3.93 (0.956)
Median [Min, Max]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]
freetime						
Mean (SD)	3.02 (0.977)	3.07 (1.03)	3.22 (1.06)	3.45 (1.04)	3.63 (1.15)	3.18 (1.05)
Median [Min, Max]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	3.50 [2.00, 5.00]	3.00 [1.00, 5.00]
goout						
Mean (SD)	2.96 (0.988)	3.08 (1.16)	3.24 (1.18)	3.51 (1.16)	3.19 (1.68)	3.18 (1.18)
Median [Min, Max]	3.00 [2.00, 5.00]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]
Dalc						
Mean (SD)	1.24 (0.524)	1.28 (0.668)	1.66 (1.03)	1.65 (1.16)	2.00 (1.26)	1.50 (0.925)
Median [Min, Max]	1.00 [1.00, 3.00]	1.00 [1.00, 5.00]	1.00 [1.00, 5.00]	1.00 [1.00, 5.00]	1.50 [1.00, 4.00]	1.00 [1.00, 5.00]
Walc						
Mean (SD)	1.89 (1.06)	2.03 (1.14)	2.44 (1.34)	2.65 (1.44)	2.75 (1.34)	2.28 (1.28)
Median [Min, Max]	2.00 [1.00, 5.00]	2.00 [1.00, 5.00]	2.00 [1.00, 5.00]	2.00 [1.00, 5.00]	3.00 [1.00, 5.00]	2.00 [1.00, 5.00]
health						
Mean (SD)	3.33 (1.49)	3.32 (1.49)	3.71 (1.37)	3.53 (1.60)	3.88 (1.36)	3.54 (1.45)
Median [Min, Max]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.50 [1.00, 5.00]	4.00 [1.00, 5.00]
absences						
Mean (SD)	2.17 (3.35)	2.97 (4.39)	4.15 (4.59)	6.41 (5.96)	0 (0)	3.66 (4.64)
Median [Min, Max]	0 [0, 14.0]	2.00 [0, 32.0]	2.00 [0, 24.0]	4.00 [0, 26.0]	0 [0, 0]	2.00 [0, 32.0]
G1						
Mean (SD)	16.2 (1.44)	13.2 (1.51)	10.2 (1.65)	7.53 (1.32)	7.25 (1.88)	11.4 (2.75)
Median [Min, Max]	16.0 [13.0, 19.0]	13.0 [9.00, 17.0]	10.0 [0, 14.0]	7.00 [4.00, 10.0]	7.00 [4.00, 11.0]	11.0 [0, 19.0]
G2						
Mean (SD)	16.9 (1.13)	13.5 (1.32)	10.3 (1.22)	7.57 (1.06)	4.19 (3.97)	11.6 (2.91)
Median [Min, Max]	17.0 [14.0, 19.0]	13.0 [10.0, 17.0]	10.0 [7.00, 13.0]	8.00 [5.00, 9.00]	5.50 [0, 10.0]	11.0 [0, 19.0]

(ABOVE) Figure 2: Table1 Stratified Demographics/Covariates Table

Figure 3: Histogram of G3 Final Grades

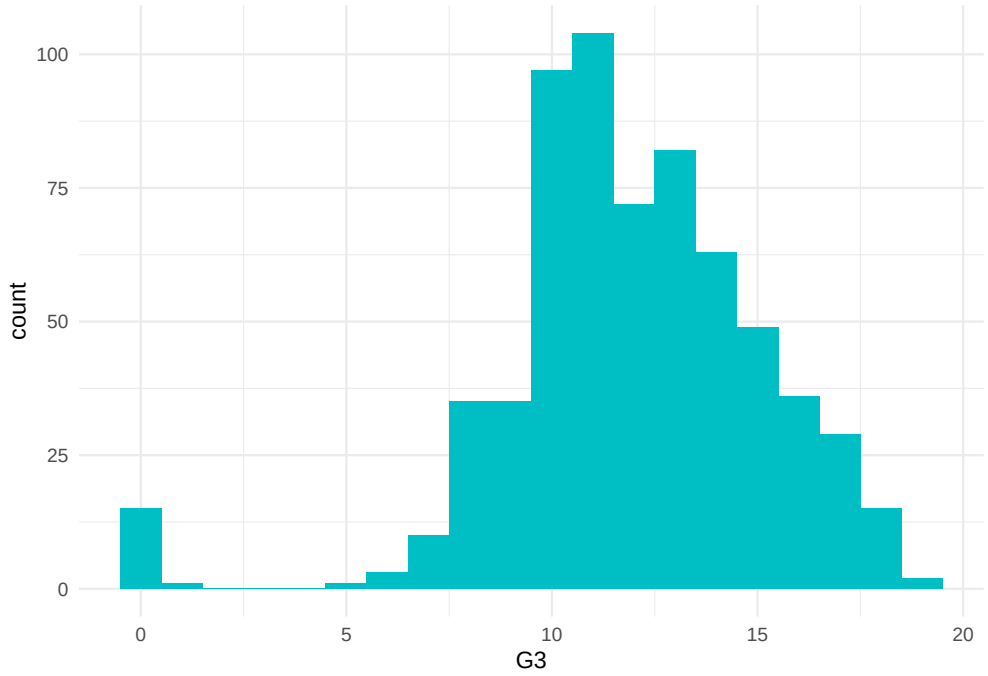


Figure 4: QQ-Plot of G3 Final Grades

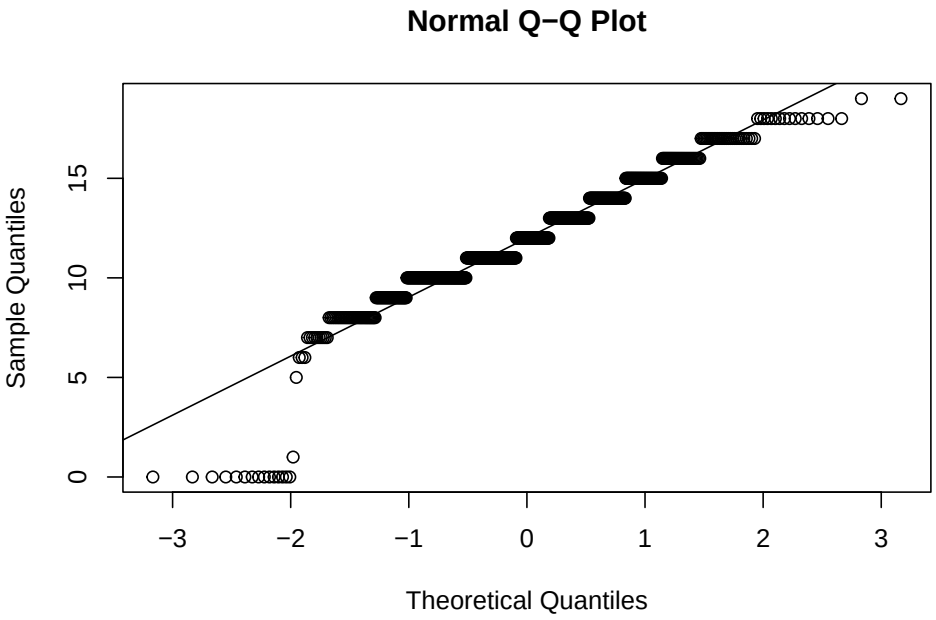


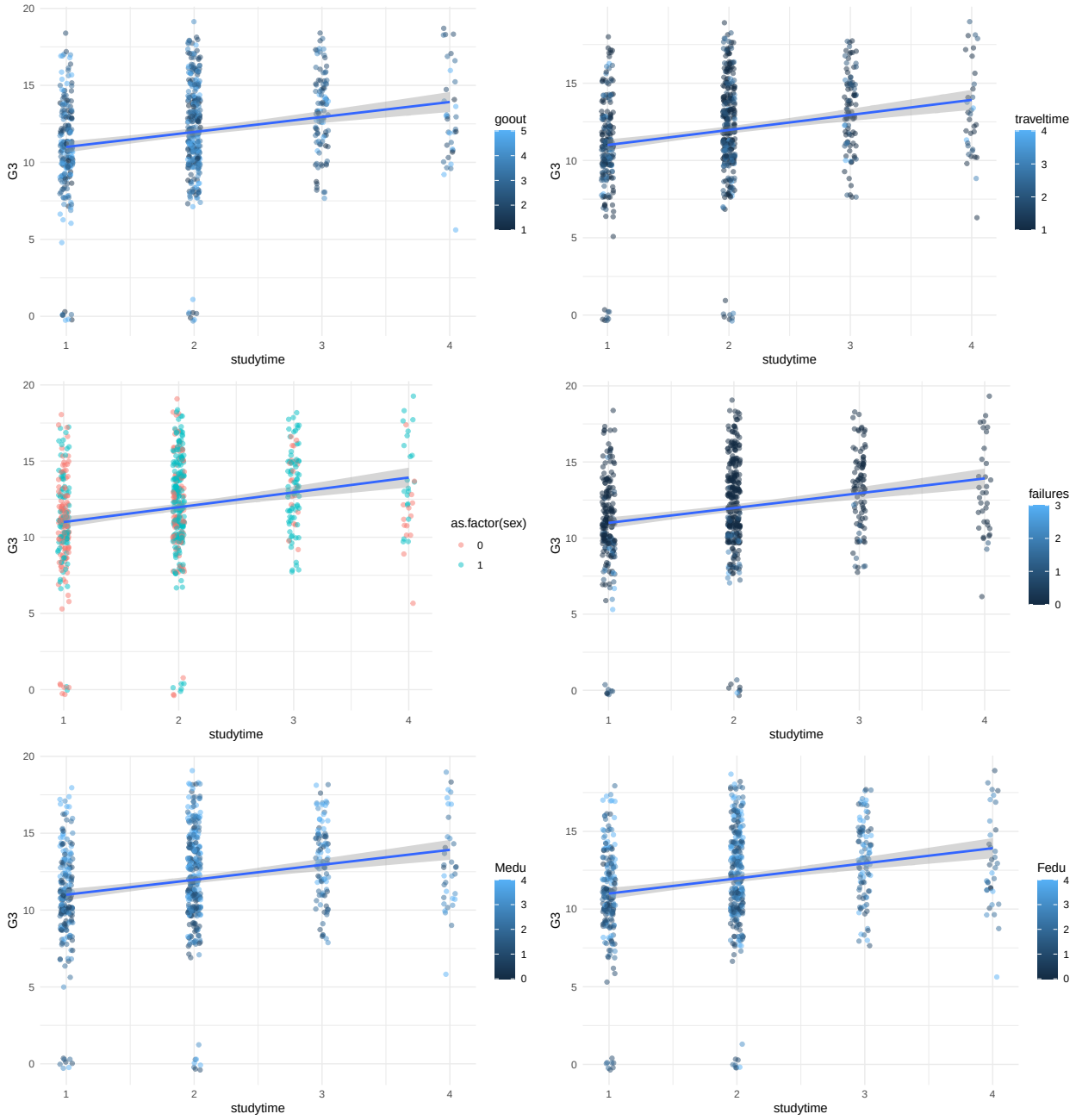
Figure 5: Linear Regression Summary Table of all Analyzed Covariates (Model 0)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.354	1.906	4.383	1.373e-05	* * *
studytime	0.5913	0.1391	4.25	2.465e-05	* * *
age	0.05333	0.1018	0.524	0.6005	
sex	0.5837	0.241	2.422	0.01571	*
addressU	0.7782	0.2601	2.992	0.00288	* *
famsizeLE3	0.2923	0.2534	1.154	0.2491	
failures	-1.726	0.2066	-8.356	4.113e-16	* * *
guardianmother	-0.1547	0.2713	-0.5702	0.5688	
guardianother	0.125	0.5425	0.2304	0.8179	
Pstatus	0.3256	0.3568	0.9127	0.3618	
Medu	0.3313	0.1339	2.475	0.01359	*
Fedu	0.2235	0.1347	1.659	0.09761	
traveltime	-0.01137	0.1635	-0.06958	0.9446	
famrel	0.2338	0.1185	1.973	0.04898	*
freetime	-0.1284	0.1154	-1.112	0.2664	
goout	-0.1561	0.1022	-1.526	0.1274	
health	-0.1703	0.07824	-2.177	0.02984	*

Figure 6: Linear Regression Summary Table of Final Model (Model 8)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	9.704	0.737	13.17	3.252e-35	* * *
studytime	0.6325	0.1392	4.543	6.623e-06	* * *
sex	0.5213	0.2382	2.189	0.02898	*
Medu	0.359	0.1305	2.751	0.006102	* *
Fedu	0.2315	0.1334	1.735	0.08314	
famrel	0.212	0.1186	1.787	0.07437	
failures	-1.716	0.1946	-8.814	1.13e-17	* * *
freetime	-0.2015	0.1087	-1.855	0.06406	
health	-0.1627	0.07836	-2.076	0.03825	*

Figures 7 - 15: Visual Stratification of Covariates to Check Further for Confounding



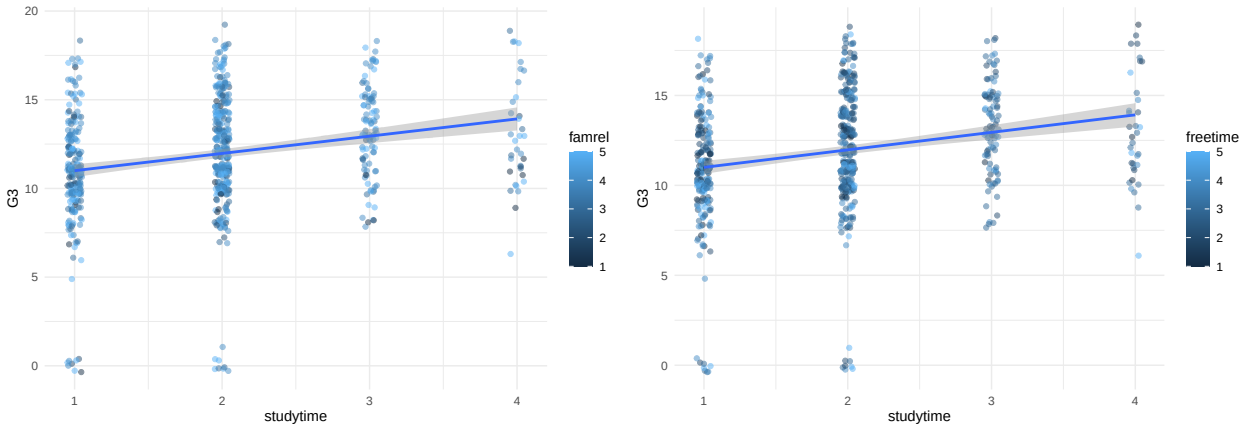


Figure 16: Thirteen Overlayed Linear Regression Models, Showing G3 Final Grades vs. Study Time

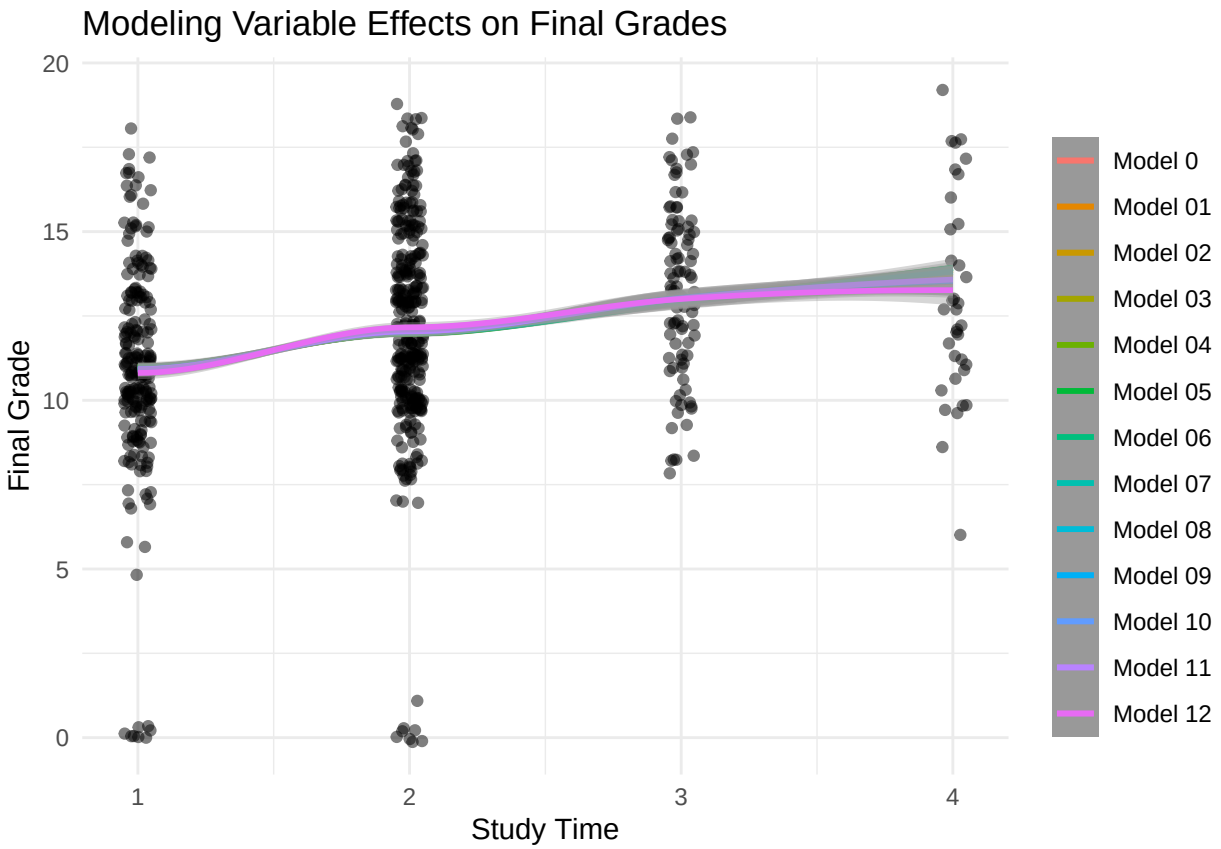


Figure 17: Random Forest Error Rate vs. Number of Trees

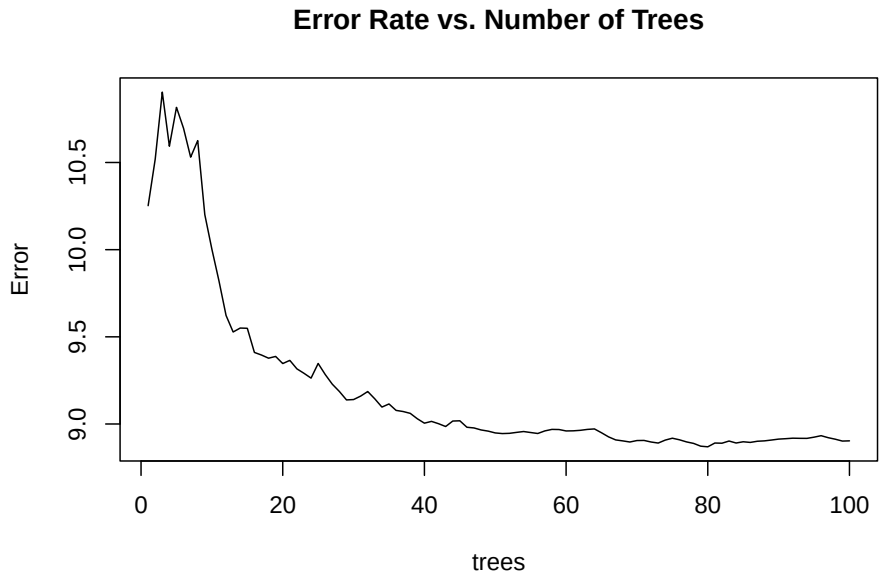
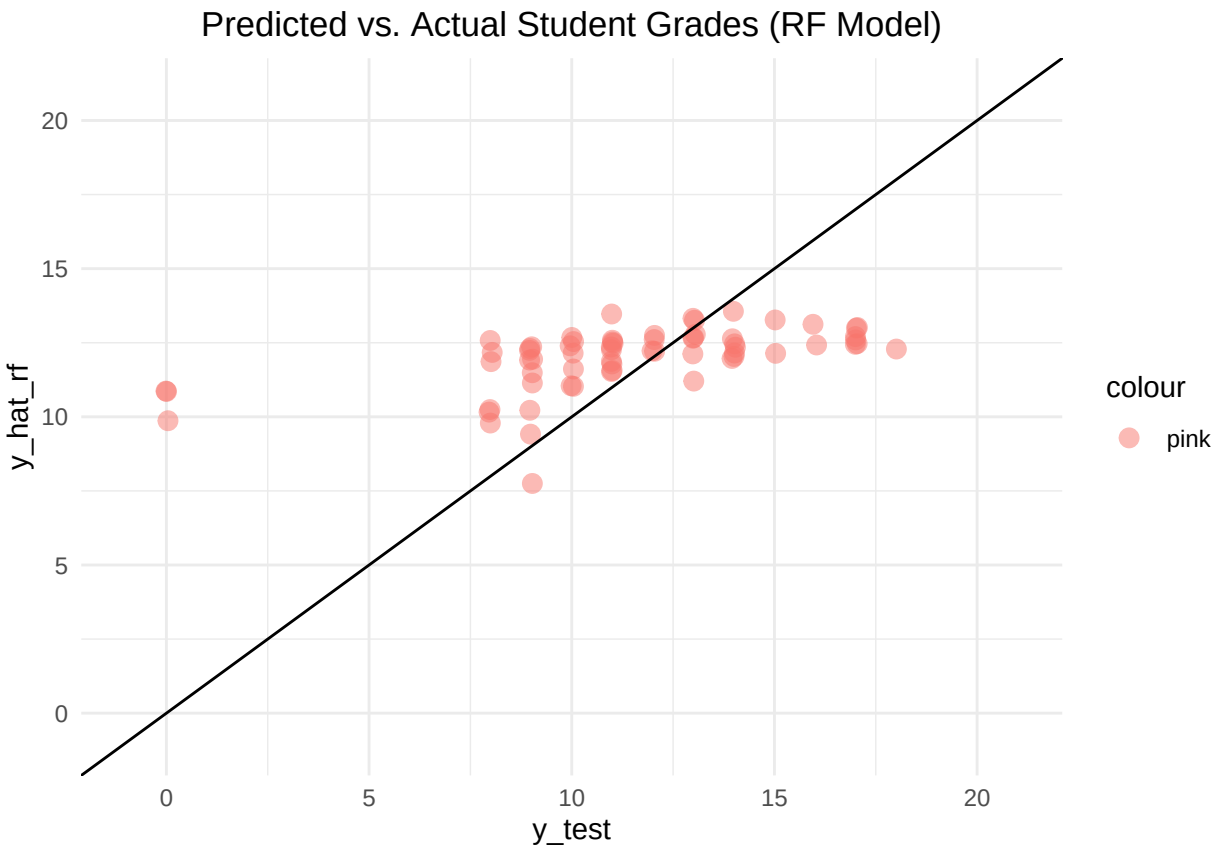


Figure 18: Predicted (y_hat_rf) vs. Actual (y_test) Final Student Grades



References

Irizarry, Rafael A. “Chapter 18 Linear Models | Introduction to Data Science.” Rafalab.dfci.harvard.edu, 2022, rafalab.dfci.harvard.edu/dsbook/linear-models.html.

Irizarry, Rafael A.“Chapter 31 Examples of Algorithms | Introduction to Data Science.” Rafalab.dfci.harvard.edu, 2022, rafalab.dfci.harvard.edu/dsbook/examples-of-algorithms.html.

Lake, Erin. “Topic 2 | Applied Linear Regression.” 2022.

“The Portuguese Grading System.” Www.studyineurope.eu, 2020, www.studyineurope.eu/ study-in-portugal/grades.

Yilmaz, Nevriye, and Boran Sekeroglu. “UCI Machine Learning Repository: Higher Education Students Performance Evaluation Dataset Data Set.”Archive.ics.uci.edu, archive.ics.uci.edu/ /ml/datasets/Higher+Education+Students+P Accessed 1 Dec. 2022.

Appendix Code

```
library(tidyverse)
library(lubridate)
library(rvest)
library(ggplot2)
library(dplyr)
library(corrplot)
library(broom)
library(caret)
library(matrixStats)
library(randomForest)
library(codyn)
library(dslabs)
library(sjPlot)
library(sjPlot)
library(gt)
library(gtExtras)
library(jtools)
library(texreg)
library(car)
library(GGally)
library(table1)
#install.packages("stargazer")
library(stargazer)
#install.packages("flextable")
library(flextable)
library(kableExtra)
library(pander)
#install.packages("float")
library(float)

# Data wrangling and EDA

# Set working directory and load data
setwd(file.path("/Users/rebeccahurwitz/Desktop/Harvard/Courses/"
                ,"Semester I Fall 2022 Courses/BST 260 - Intro to Data Science/BST260FinalProject"))

# Reading in the .csv file
studentdata <- read.csv("student-por.csv", sep = ";")
head(studentdata)
studentdatalong <- studentdata %>% select(G3, sex, age, Medu, Fedu, traveltime, studytime, failures,
                                          freetime, goout, famrel, health, absences)
studentdatalong <- gather(studentdatalong, key = "variable", value = "value", -G3)

head(studentdatalong)

newstudentdatalong <- studentdata %>% select(G3, age, traveltime, studytime)
newstudentdatalong <- gather(newstudentdatalong, key = "variable", value = "value", -G3)

# Recoding variables
studentdata <- studentdata %>%
      mutate(Pstatus = ifelse(Pstatus == "A",0,1))
```

```r
studentdata <- studentdata %>%
      mutate(sex = ifelse(sex == "M",0,1))
studentdata$studytime_squared <- studentdata$studytime^2

# Linear

studentdata |>
  ggplot(aes(studytime, G3, col = freetime)) +
  geom_point() +
  facet_grid(traveltime~failures)

studentdata$gooutfac <- as.factor(studentdata$goout)

studentdata$famrelfac <- as.factor(studentdata$famrel)

studentdata$healthfac <- as.factor(studentdata$health)

studentdata$freetimefac <- as.factor(studentdata$freetime)

studentdata$studytimefac <- as.factor(studentdata$studytime)



parentmodel <- lm(G3 ~ Pstatus + famrel, data = studentdata)
summary(parentmodel)


# Visualizing possible confounders

# Study time with go out color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = goout), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with travel time color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = traveltime), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with sex color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = as.factor(sex)), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with failures color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = failures), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()
```

```r
# Study time with Medu color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = Medu), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with Fedu color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = Fedu), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with famrel color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = famrel), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with freetime color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = freetime), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Travel time w study time color
studentdata %>% ggplot(aes(x = traveltime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = studytimefac), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Failures
studentdata %>% ggplot(aes(x = failures, y = G3)) +
  geom_point(position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Famrel
studentdata %>% ggplot(aes(x = famrel, y = G3)) +
  geom_point(position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Freetime w sex color
studentdata %>% ggplot(aes(x = freetime, y = G3)) +
  geom_point(aes(col = sex), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# G1+G2 avg w famrel color
studentdata %>% ggplot(aes(x = (G2+G1)/2, y = G3)) +
  geom_point(aes(col = famrelfac), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()
```

```r
studentdata %>% ggplot(aes(x = health, y = G3)) +
  geom_point(alpha = 0.5, aes(col = famrelfac), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = famrelfac), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()


studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_boxplot()

studentdatalong <- studentdata %>% select(G3, sex, age, Medu, Fedu,
                                          traveltime, studytime, failures,
                                          freetime, goout, famrel, health, absences)
studentdatalong <- gather(studentdatalong, key = "variable", value = "value", -G3)

head(studentdatalong)

newstudentdatalong <- studentdata %>% select(G3, age, traveltime, studytime)
newstudentdatalong <- gather(newstudentdatalong, key = "variable", value = "value", -G3)

YES <- ggplot(studentdatalong, aes(x = value, y = G3)) +
  geom_point(alpha = 0.5,aes(col = variable), position = position_jitter(w = 0.05)) +
  geom_smooth(method = "lm") +
  facet_wrap(~ variable, scales = "free_x") +
  theme_minimal()

# Checking Normality
qqnorm(studentdata$G3)
qqline(studentdata$G3)

ggplot(data = studentdata, aes(G3)) +
  geom_histogram(bins = 20, fill = "#00BFC4") +
  theme_minimal()



# Adjusting covariates
studentdata$studytime_squared <- studentdata$studytime^2
studentdata$age_squared <- studentdata$age^2

ggplot(data = studentdata, aes(G3)) +
  stat_qq(aes(sample=G3)) +
  facet_wrap(~G3)

model1A <- glm(G3 ~ studytime + age + studentdata$sex + address + famsize +
                 guardian + studentdata$Pstatus + Medu + Fedu + traveltime +
                 famrel + freetime + goout + health, data = studentdata)
summary(model1A)
```

```
glance(model1A)

model0 <- lm(G3 ~ studytime + age + sex + address + famsize + failures +
               guardian + Pstatus + Medu + Fedu + traveltime + famrel +
               freetime + goout + health, data = studentdata)
summary(model0)
glance(model0)
model0 %>%
    tab_model()


texreg(model0)

model1 <- lm(G3 ~ studytime + age + sex + Pstatus + Medu + Fedu + traveltime +
               famrel + freetime + goout + health, data = studentdata)
summary(model1)
glance(model1)

basic_model <- lm(G3 ~ studytime, data = studentdata)

model01 <- lm(G3 ~ studytime + sex, data = studentdata) #sex not
summary(model01)
AIC(model01)
abs(100 * ((coef(basic_model)[2] - coef(model01)[2]) /
             coef(model01)[2])) > 10

model02 <- lm(G3 ~ studytime + Medu, data = studentdata) #medu not
summary(model02)
AIC(model02)
abs(100 * ((coef(basic_model)[2] - coef(model02)[2]) /
             coef(model02)[2])) > 10

model03 <- lm(G3 ~ studytime + Fedu, data = studentdata) #fedu not
summary(model03)
AIC(model03)
abs(100 * ((coef(basic_model)[2] - coef(model03)[2]) /
             coef(model03)[2])) > 10

model04 <- lm(G3 ~ studytime + famrel, data = studentdata) #famrel not
summary(model04)
AIC(model04)
abs(100 * ((coef(basic_model)[2] - coef(model04)[2]) /
             coef(model04)[2])) > 10

model05 <- lm(G3 ~ studytime + freetime, data = studentdata) #freetime not
summary(model05)
AIC(model05)
abs(100 * ((coef(basic_model)[2] - coef(model05)[2]) /
             coef(model05)[2])) > 10

model06 <- lm(G3 ~ studytime + health, data = studentdata) #health not
summary(model06)
AIC(model06)
```

```
abs(100 * ((coef(basic_model)[2] - coef(model06)[2]) /
             coef(model06)[2])) > 10

model07 <- lm(G3 ~ studytime + failures, data = studentdata) #health not
summary(model07)
AIC(model07)
abs(100 * ((coef(basic_model)[2] - coef(model07)[2]) /
             coef(model07)[2])) > 10

model08 <- lm(G3 ~ studytime + sex + Medu + Fedu + famrel + failures +
                freetime + health, data = studentdata)
summary(model08) #0.2314
AIC(model08) #3204.031
lm8 <- model08 %>%
    tab_model()


model09 <- lm(G3 ~ studytime + sex + Medu + Fedu + famrel + failures +
                freetime, data = studentdata)
summary(model09) #0.1275
AIC(model09) #3206.388

model10 <- lm(G3 ~ studytime + sex + Medu + Fedu + famrel + failures,
              data = studentdata)
summary(model10) #0.2241
AIC(model10) #3208.24

model11 <- lm(G3 ~ studytime + sex + Medu + Fedu + failures,
              data = studentdata)
summary(model11) #0.2231
AIC(model11) #3208.09

model12 <- lm(G3 ~ studytime + studytime_squared + sex + Medu + Fedu +
                famrel + failures + freetime + health, data = studentdata)
summary(model12) #0.2313
AIC(model12) #3205.098

model13 <- glm(G3 ~ studytime + sex + Medu + Fedu + famrel + freetime +
                 health + failures, data = studentdata)
summary(model13) #
AIC(model13) #3204.031
glance(model13)

#MODEL 8 WINS


model2 <- lm(G3 ~ studytime + age + sex + Medu + Fedu + famrel + freetime +
               health, data = studentdata)
summary(model2)
glance(model2)

model3 <- lm(G3 ~ studytime + studytime_squared + age + age_squared + sex + Medu
             + Fedu + famrel + freetime + health, data = studentdata)
```

```
summary(model3)
glance(model3)

model4 <- lm(G3 ~ studytime + age + sex + Medu + Fedu + famrel +
               freetime, data = studentdata)
summary(model4)
glance(model4)


model5 <- lm(G3 ~ studytime + studytime_squared + age + age_squared,
             data = studentdata)
summary(model5)
glance(model5)

ggplot(data = studentdata, aes(studytime, G3)) +
  geom_point(alpha = 0.5, position = position_jitter(w = 0.05)) +
  geom_smooth(aes(y = predict(model0), color="Model 0")) +
  geom_smooth(aes(y = predict(model01), color="Model 01")) +
  geom_smooth(aes(y = predict(model02), color="Model 02")) +
  geom_smooth(aes(y = predict(model03), color="Model 03")) +
  geom_smooth(aes(y = predict(model04), color="Model 04")) +
  geom_smooth(aes(y = predict(model05), color="Model 05")) +
  geom_smooth(aes(y = predict(model06), color="Model 06")) +
  geom_smooth(aes(y = predict(model07), color="Model 07")) +
  geom_smooth(aes(y = predict(model08), color="Model 08")) +
  geom_smooth(aes(y = predict(model09), color="Model 09")) +
  geom_smooth(aes(y = predict(model10), color="Model 10")) +
  geom_smooth(aes(y = predict(model11), color="Model 11")) +
  geom_smooth(aes(y = predict(model12), color="Model 12")) +
  xlab("Study Time") +
  ylab("Final Grade") +
  scale_colour_manual("",
                      breaks = c("Model 0", "Model 01", "Model 02", "Model 03",
                                 "Model 04", "Model 05", "Model 06", "Model 07",
                                 "Model 08", "Model 09", "Model 10", "Model 11",
                                 "Model 12"),
                      values = c("#F8766D", "#E58700", "#C99800", "#A3A500",
                                 "#6BB100", "#00BA38", "#00BF7D", "#00C0AF",
                                 "#00BCD8", "#00B0F6", "#619CFF", "#B983FF",
                                 "#E76BF3")) +
  labs(title="Modeling Variable Effects on Final Grades") +
  theme_minimal()



model2 <- lm(G3 ~ studytime + studytime_squared, data = studentdata)
summary(model2)

model3 <- lm(G3 ~ Pstatus, data = studentdata)
summary(model3)


ggplot(data = studentdata, aes(studytime, G3)) +
```

```
  geom_point(alpha = 0.5, aes(col = age), position = position_jitter(w = 0.05)) +
  theme_minimal() +
  geom_smooth(aes(y=predict(model1))) +
  geom_smooth(aes(y=predict(model3))) +
  xlab("studytime") +
  ylab("G3")
  labs(title="G3 vs. Study time")

# "Hand"-calculating linear model
summary_stats <- studentdata %>% summarize(avg_studytime = mean(studytime),
            s_studytime = sd(studytime),
            avg_grade = mean(G3),
            s_grade = sd(G3),
            r = cor(studytime, G3))
summary_stats

reg_line <- summary_stats %>%  summarize(slope = r*s_grade/s_studytime,
                           intercept = avg_grade - slope*avg_studytime)

p <- studentdata |>
  ggplot(aes(studytime, G3)) +
  geom_point(alpha = 0.5)
p

p + geom_abline(intercept = reg_line$intercept, slope = reg_line$slope)
p + geom_smooth(method = "lm")


avPlots(model5)
ggpairs(studentdata, columns = 1:3)



ggplot(newstudentdatalong, aes(x = value, y = G3) ) +
  geom_point(aes(col = variable)) +
  stat_smooth() +
  facet_wrap(~ variable, scales = "free_x") +
  theme_minimal()


ggplot(studentdatalong, aes(x = value, y = G3) ) +
  geom_point(aes(col = variable)) +
  stat_smooth() +
  facet_grid(~ variable, scales = "free_x") +
  theme_minimal()

# Machine Learning!!!!!!!

set.seed(1996)
indices <- sample(1:649,floor(.9*649), replace = FALSE)
my_train <- studentdata[indices, ]
my_test <- studentdata[-indices, ]
```

```r
set.seed(1996)
x <- my_train %>%
  select(studytime, sex, Medu, Fedu, famrel, failures, freetime, health)
y <- my_train %>%
  select(G3) %>% unlist() %>% as.numeric()

#as.numeric(unlist(studentdata[split1 == 0, ]


x_test <- my_test %>%
  select(studytime, sex, Medu, Fedu, famrel, failures, freetime, health)
y_test <- my_test %>%
  select(G3) %>% unlist() %>% as.numeric()

#%>% unlist() %>% as.numeric()
#as.numeric(unlist(studentdata[split1 == 1, ]
#)) %>% as.factor()

colnames(x) <- 1:ncol(x)
colnames(x_test) <- colnames(x)

control <- trainControl(method = "cv", number = 10)
grid <- data.frame(mtry = c(1:10))
train_rf <-  train(x, y,
                   method = "rf",
                   ntree = 75,
                   trControl = control,
                   tuneGrid = grid,
                   nSamp = 1000)

fit_rf <- randomForest::randomForest(x, y,
                        mtry = train_rf$bestTune$mtry,
                        ntree = 75)
plot(fit_rf)

y_test <- y_test
y_hat_rf <- predict(fit_rf, x_test)

#importance(fit_rf)
#varImpPlot(fit_rf)
#fit_rf$importance


sqrt(mean((y_hat_rf-y_test)^2))

#sqrt(mean((y_hat_rf2-y_test2)^2))

combined <- as.data.frame(cbind(y_hat_rf, y_test))
#combined$y_hat_rf <- as.factor(combined$y_hat_rf)
#combined$y_test <- as.factor(combined$y_test)

treegraph1 <- combined %>% ggplot(aes(y_test, y_hat_rf)) +
  geom_point(alpha = 0.5, aes(col = "pink"), position = position_jitter(w = 0.05)) +
```

```
  geom_abline(slope = 1, aes(col = "gray")) +
  xlim(-1, 21) +
  ylim(-1, 21) +
  labs(title = "Actual vs. Predicted Student Grades (RF Model)") +
  theme_minimal()




####### covariate change
set.seed(1996)
x2 <- my_train %>%
  select(studytime, sex, Medu, Fedu, famrel, freetime, health)
y2 <- my_train %>%
  select(G3) %>% unlist() %>% as.numeric()
 #as.numeric(unlist(studentdata[split1 == 0, ]


x_test2 <- my_test %>%
  select(studytime, sex, Medu, Fedu, famrel, freetime, health)
y_test2 <- my_test %>%
  select(G3) %>% unlist() %>% as.numeric()
#%>% unlist() %>% as.numeric()
#as.numeric(unlist(studentdata[split1 == 1, ]
#)) %>% as.factor()

colnames(x2) <- 1:ncol(x2)
colnames(x_test2) <- colnames(x2)

control2 <- trainControl(method = "cv", number = 10)
grid2 <- data.frame(mtry = c(1:10))
train_rf2 <-  train(x, y,
                    method = "rf",
                    ntree = 100,
                    trControl = control2,
                    tuneGrid = grid2,
                    nSamp = 1000)

fit_rf2 <- randomForest::randomForest(x2, y2,
                         mtry = train_rf2$bestTune$mtry,
                         ntree = 100)
plot(fit_rf2)

y_test2 <- y_test2
y_hat_rf2 <- predict(fit_rf, x_test)

sqrt(mean((y_hat_rf2-y_test2)^2))

#levels(y_hat_rf) <- levels(y_test)

#str(as.numeric(y_test))
#str(as.numeric(y_hat_rf))
```

```r
#relevel(y_hat_rf, ref = "M")
y_hat_rf2 <- factor(y_hat_rf, levels=c(1:65))



#cm <- confusionMatrix(y_hat_rf, y_test)
combined2 <- as.data.frame(cbind(y_hat_rf2, y_test2))
#combined$y_hat_rf <- as.factor(combined$y_hat_rf)
#combined$y_test <- as.factor(combined$y_test)

combined2 %>% ggplot(aes(y_test2, y_hat_rf2)) +
  geom_point(alpha = 0.5, aes(color = "#FF62BC"), position = position_jitter(w = 0.05)) +
  xlim(0, 20) +
  ylim(0, 20) +
  labs(title = "Actual vs. Predicted Student Grades (RF Model 2)") +
  theme_minimal()

#levels(combined$y_hat_rf)
#levels(combined$y_test)









#max(train_rf[["results"]][["Accuracy"]])

#cm <- confusionMatrix(as.factor(y_hat_rf), as.factor(y_test))

#levels(as.factor(y_hat_rf))
#levels(as.factor(y_test))

class(y_hat_rf)
class(y)
class(y_test)

levels(y_hat_rf) <- levels(y_test)

nzv <- nearZeroVar(x)
col_index <- setdiff(1:ncol(x), nzv)
x_final <- x[ ,col_index]

#confusionMatrix(y_hat_rf,as.factor(y_test))

#table(y_hat_rf)
#table(y_test)

#confusionMatrix(
```

```
 # factor(y_hat_rf, levels = 1:18),
  #factor(y_test, levels = 1:18))

#cm$overall["Accuracy"]



class(y_test)
class(y_hat_rf)

#confusionMatrix(pred,as.factor(testing$Final))

#varImp(y_hat_rf)
#imp <- varImp(fit_rf)
#impDF <- data.frame(imp[1])
#rownames(impDF)[order(impDF$Overall, decreasing=TRUE)]

#col_index <- varImp(fit_rf)$importance %>%
 # mutate(names=row.names(.)) %>%
  #arrange(-Overall)


#varImp(y_hat_rf, scale = TRUE)$importance
plot(y_hat_rf, top = 20)

#imp <- importance(fit_rf)
#mat <- rep(0, ncol(x))
#mat[x] <- imp
#image(matrix(mat, 28, 28))


train_rf <- randomForest(y ~ ., data=mnist_27$train)

#confusionMatrix(predict(train_rf, mnist_27$test),


#set.seed(1996)
#rf <-randomForest(G3~studytime + age + Medu + Fedu + famrel +
#failures + freetime + health, data=studentdata, mtry = train_rf$bestTune$mtry, \
#importance=TRUE,ntree=75)
#print(rf)
#Evaluate variable importance
#importance(rf)
#varImpPlot(rf)

#imp=as.data.frame(importance(fit_rf))
#imp=cbind(vars=rownames(imp),imp)

YES <- ggplot(studentdatalong, aes(x = value, y = G3)) +
  geom_point(alpha = 0.3,aes(col = variable), position = position_jitter(w = 0.05)) +
  geom_smooth(method = "lm") +
  facet_wrap(~ variable, scales = "free_x") +
  theme_minimal()
```

```
YES
# Study time with go out color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = goout), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with travel time color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = traveltime), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with sex color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = as.factor(sex)), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with failures color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = failures), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with Medu color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = Medu), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with Fedu color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = Fedu), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with famrel color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = famrel), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with freetime color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = freetime), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

plswork <- model08 %>% pander(add.significance.stars = T, caption = "")
plswork
treegraph1 <- combined %>% ggplot(aes(y_test, y_hat_rf)) +
  geom_point(alpha = 0.5, aes(col = "pink"), position = position_jitter(w = 0.05), size=3) +
```

```
  geom_abline(slope = 1, aes(col = "gray")) +
  xlim(-1, 21) +
  ylim(-1, 21) +
  labs(title = "Predicted vs, Actual Student Grades (RF Model)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
treegraph1

YES <- ggplot(studentdatalong, aes(x = value, y = G3)) +
  geom_point(alpha = 0.3,aes(col = variable), position = position_jitter(w = 0.05)) +
  geom_smooth(method = "lm") +
  facet_wrap(~ variable, scales = "free_x") +
  theme_minimal()

YES
studentdata$catG3 <- as.factor(ifelse(studentdata$G3 > 16, 'Top Quintile (17-20)',
                      ifelse(studentdata$G3 > 12, 'Fourth Quintile (13-16)',
                      ifelse(studentdata$G3 > 8, 'Third Quintile (9-12)',
                      ifelse(studentdata$G3 > 4, 'Second Quintile (5-8)', 'Bottom Quintile (1-4)')))))

tableyay <- table1(~ factor(sex) + age + famsize + Pstatus + Medu + Fedu + Mjob + Fjob + guardian + trav

#tableyay <- t1flex(tableyay)
t1kable(tableyay) %>% kable_styling(latex_options="scale_down")
ggplot(data = studentdata, aes(G3)) +
  geom_histogram(bins = 20, fill = "#00BFC4") +
  theme_minimal()
qqnorm(studentdata$G3)
qqline(studentdata$G3)

model0 %>% pander(add.significance.stars = T, caption = "")
plswork <- model08 %>% pander(add.significance.stars = T, caption = "")
plswork
# Study time with go out color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = goout), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with travel time color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = traveltime), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with sex color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5, aes(col = as.factor(sex)), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with failures color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
```

```
  geom_point(alpha = 0.5, aes(col = failures), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with Medu color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = Medu), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with Fedu color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = Fedu), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with famrel color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = famrel), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

# Study time with freetime color
studentdata %>% ggplot(aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.5,aes(col = freetime), position = position_jitter(w = 0.05)) +
  geom_smooth(method = lm) +
  theme_minimal()

ggplot(data = studentdata, aes(studytime, G3)) +
  geom_point(alpha = 0.5, position = position_jitter(w = 0.05)) +
  geom_smooth(aes(y = predict(model0), color="Model 0")) +
  geom_smooth(aes(y = predict(model01), color="Model 01")) +
  geom_smooth(aes(y = predict(model02), color="Model 02")) +
  geom_smooth(aes(y = predict(model03), color="Model 03")) +
  geom_smooth(aes(y = predict(model04), color="Model 04")) +
  geom_smooth(aes(y = predict(model05), color="Model 05")) +
  geom_smooth(aes(y = predict(model06), color="Model 06")) +
  geom_smooth(aes(y = predict(model07), color="Model 07")) +
  geom_smooth(aes(y = predict(model08), color="Model 08")) +
  geom_smooth(aes(y = predict(model09), color="Model 09")) +
  geom_smooth(aes(y = predict(model10), color="Model 10")) +
  geom_smooth(aes(y = predict(model11), color="Model 11")) +
  geom_smooth(aes(y = predict(model12), color="Model 12")) +
  xlab("Study Time") +
  ylab("Final Grade") +
  scale_colour_manual("",
                      breaks = c("Model 0", "Model 01", "Model 02", "Model 03", "Model 04", "Model 05",
                      values = c("#F8766D", "#E58700", "#C99800", "#A3A500", "#6BB100", "#00BA38", "#00
  labs(title="Modeling Variable Effects on Final Grades") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_minimal()
plot(fit_rf2, main = "Error Rate vs. Number of Trees")
treegraph1 <- combined %>% ggplot(aes(y_test, y_hat_rf)) +
```

```
  geom_point(alpha = 0.5, aes(col = "pink"), position = position_jitter(w = 0.05), size=3) +
  geom_abline(slope = 1, aes(col = "gray")) +
  xlim(-1, 21) +
  ylim(-1, 21) +
  labs(title = "Predicted vs. Actual Student Grades (RF Model)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
treegraph1
```