Investigating High School Student Performance

Sasha Sabater

05/09/23

Introduction

What Do Students Need? Investigating High School Performance

In this project, I aim to gain insight on what aspects affect a student's school performance. Specifically, this project analyzes a dataset compiled from two Portuguese high schools, and contains 33 variables that range from school performance and information to demographic and social factors. More details on this study, and how its data was compiled, can be found in this public paper. Documentation on the variables and their values can be found here.

```
#first going to install and load in all necessary packages
#install.packages("tidyverse")
#install.packages("tidymodels")
library(tidyverse)
library(tidymodels)
```

Exploring The Data

The dataset is split in two - one for math scores and one for Portuguese scores.They contain all of the same variables, of which there are 33. I choose to focus on just the Portuguese dataset for this project.

```
port_df <- read_delim("data/student-por.csv")
```

Most of the columns' data are stored as either a double or character data type, when they should in fact be categorized as factors. Many of the columns are questionnaire questions, where the student was asked to rank say, their quality of familial relationship on a scale of 1 to 5. Additionally, there are columns, like "activities" and "internet" that are answered either character strings Yes or No, meaning I should convert almost all of the columns into categorical variables before conducting any kind of classification model.

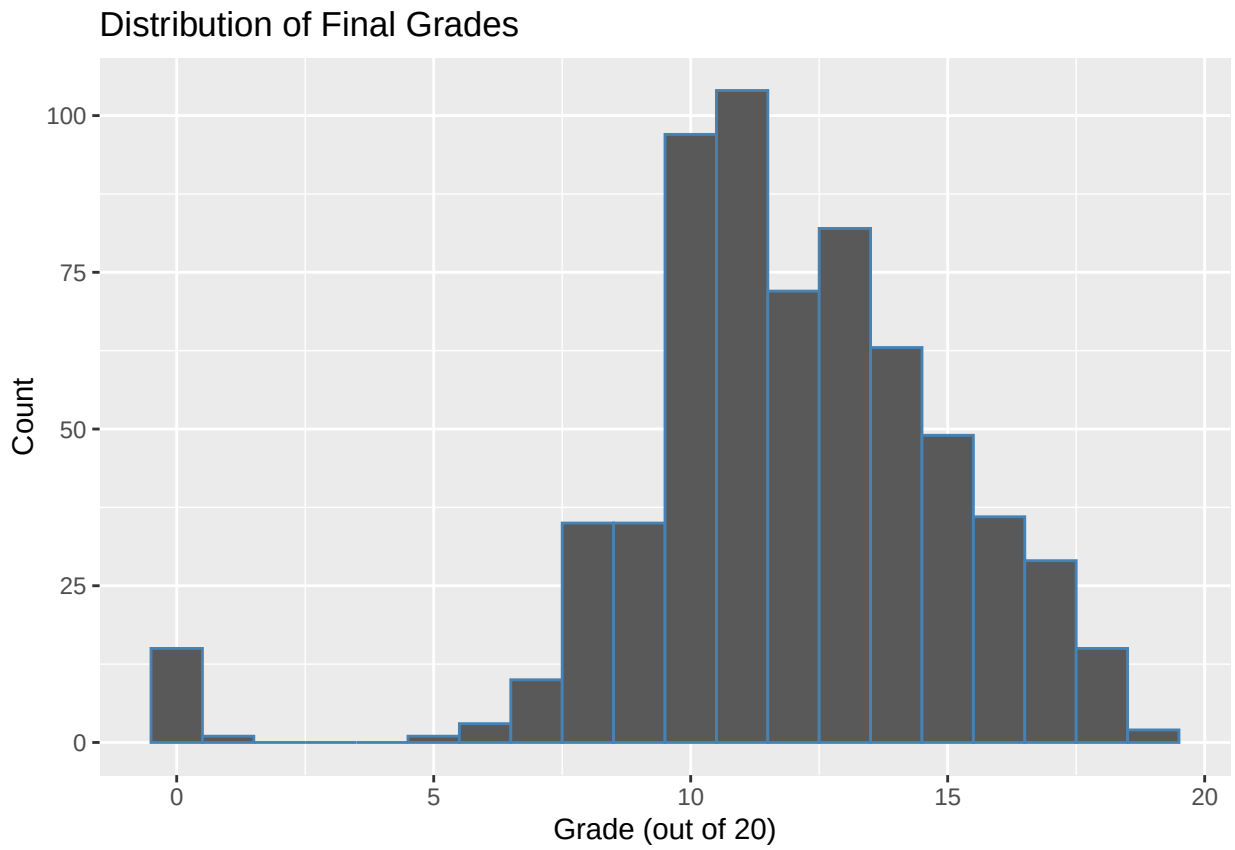
However, before modeling I explore the set to gain some insights on its variables.

```
##  [1] "school"     "sex"        "age"        "address"    "famsize"
##  [6] "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"
## [11] "reason"     "guardian"   "traveltime" "studytime"  "failures"
## [16] "schoolsup"  "famsup"     "paid"       "activities" "nursery"
## [21] "higher"     "internet"   "romantic"   "famrel"     "freetime"
## [26] "goout"      "Dalc"       "Walc"       "health"     "absences"
## [31] "G1"         "G2"         "G3"
```

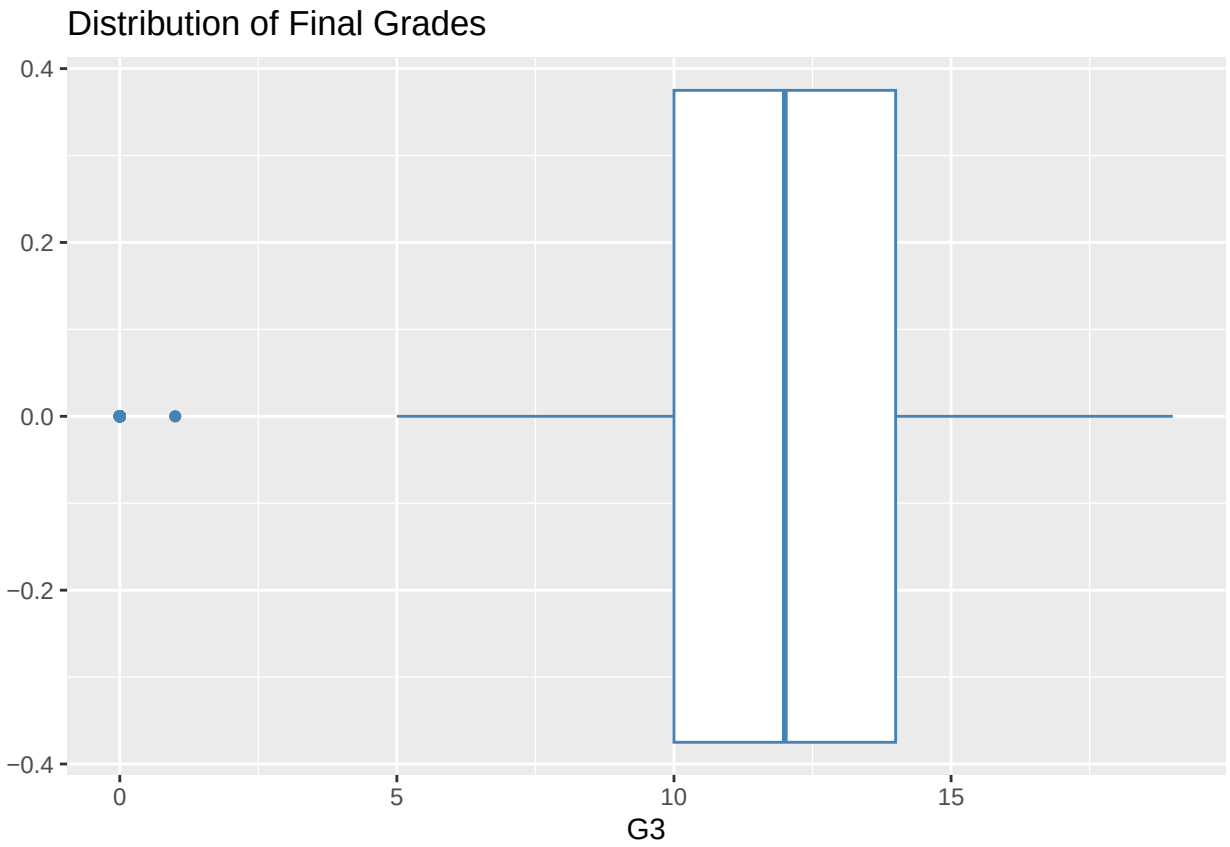
I am going to exclude the following columns from analysis: school, sex, age, address, Pstatus, Mjob, Fjob, reason, guardian, G1, and G2 (please note that the variables "G1", "G2", and "G3" refer to term grades, with G1 and G2 being the students grade in the first two terms of the year, and G3 being the final grade in the course, and is scored out of 20). I am mostly interested in the categorical variables about the students' social and familial life - for example I do not wish to include whether their parents are separated or whether they have parents(which is why I am omitting the columns "Pstatus" and "guardian"), but I do want to include variables like "goout" and "famrel" which categorizes whether they go out with friends or the quality of their familial relationships respectively.

```
port_df <- port_df %>%
  select(-school, -sex, -age, -address, -Pstatus, -Mjob, -Fjob, -reason, -guardian, -G1, -G2)
```

```
#the following plots to see the distribution of the variable in question
port_df %>%
  ggplot(aes(x = G3)) +
  geom_histogram(bins = 20, color = "steelblue") +
  labs(title = "Distribution of Final Grades",
       x = "Grade (out of 20)",
       y = "Count")
```




```
port_df %>%
  ggplot(aes(x = G3))+
  geom_boxplot(color = "steelblue")+
  labs(title = "Distribution of Final Grades")
```



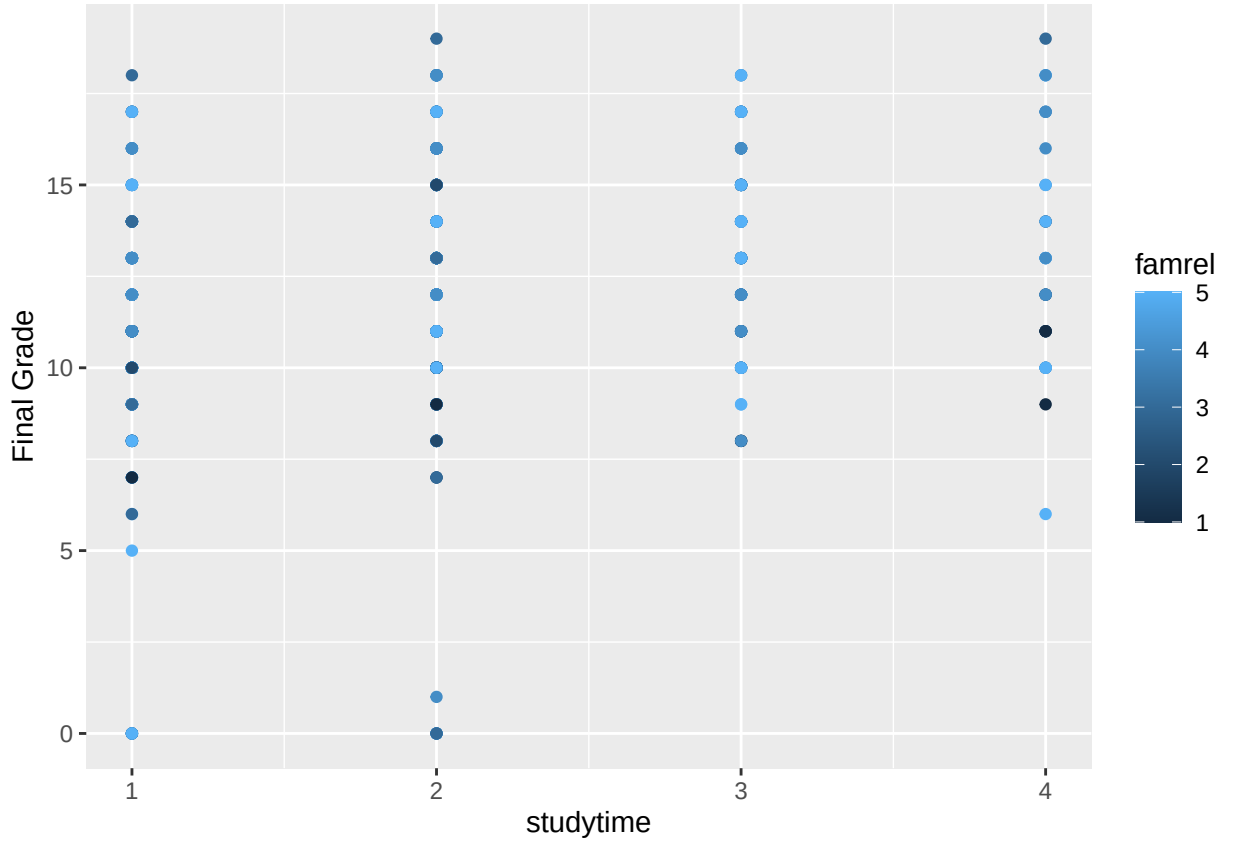

```
#creating a custom summarizing function
stats_summary <- function(df, var){
  df %>%
    summarise(mean = mean({{var}}),
              median = median({{var}}),
              sd = sd({{var}}),
              min = min({{var}}),
              max = max({{var}}))
}
#creating a function to group_by an explanatory variable and summarise a response variable
group_summary <- function(df, exp_var, resp_var){
  df %>%
    group_by({{exp_var}}) %>%
    summarise(mean = mean({{resp_var}}),
              median = median({{resp_var}}),
              sd = sd({{resp_var}}),
              min = min({{resp_var}}),
              max = max({{resp_var}}))
}

stats_summary(port_df, G3)
```

```
## # A tibble: 1 x 5
##    mean median    sd   min   max
##   <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1  11.9     12  3.23     0    19
```

The histogram illustrates that this variable is normally distributed, albeit a bit left skewed, with a singular peak happening around the mean/median (since they are so close in valuable they can both be taken as a measure of center). We can also see that there are a few outliers occurring in the lower end of the distribution.

```
port_df %>%
  ggplot(aes(x = studytime, y = G3, color = famrel)) +
  geom_point() +
  labs(y = "Final Grade")
```



Presumption might lead us to think that those who spend more time studying would get higher grades, but we can see there seems to be (direct or obvious) correlation/trend between an increase in study time and an increase in grade performance.

```
group_summary(port_df, failures, G3)
```

```
## # A tibble: 4 x 6
##   failures  mean median    sd   min   max
##      <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1        0 12.5    12    2.83     0    19
## 2        1  8.64   10    3.44     0    16
## 3        2  8.81    9.5  3.21     0    15
## 4        3  8.07    8.5  2.79     0    11
```

Unsurprisingly, it seems that there is a trend that students who have failed classes before tend to score lower overall. It seems that there is a steep drop off in average grade just after the first level of the failure variable

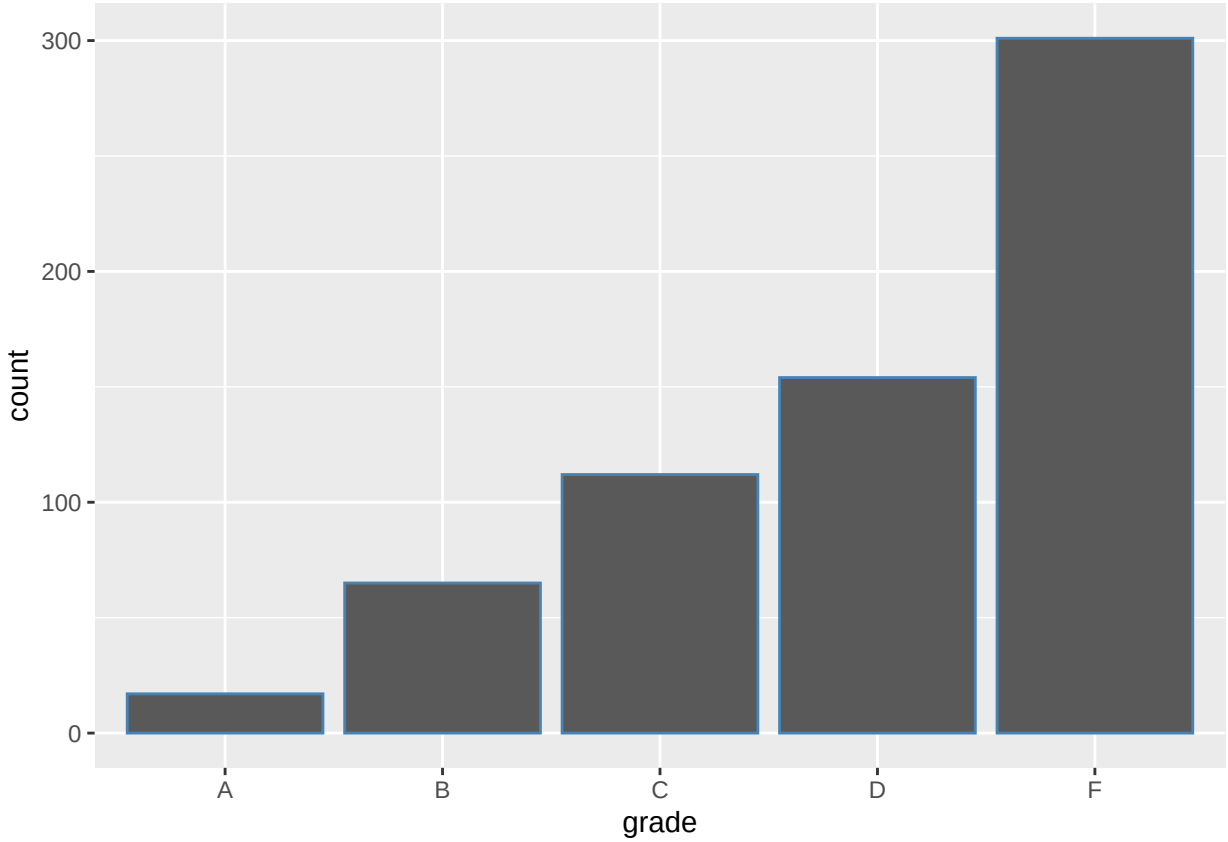
- those who have failed at least one class have significantly lower summary statistics than those who have never failed a class before.

All of the variables in question are categorical, and I am going to convert G3 from numeric to letter grades to better capture its categorical nature, and I am going to change its name to "grade".

```
#converting G3
port_df <- port_df %>%
  mutate(grade = case_when(
    G3 >= 18 ~ "A",
    G3 >= 16 ~ "B",
    G3 >= 14 ~ "C",
    G3 >= 12 ~ "D",
    TRUE ~ "F"
  )) %>%
  select(-G3)
```

```
port_df %>%
  ggplot(aes(x = grade)) +
  geom_histogram(color = "steelblue", stat = "count")
```

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



The vast majority of these students have failed

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.