Analysis of Subject Proficiency Differentials in High School Students

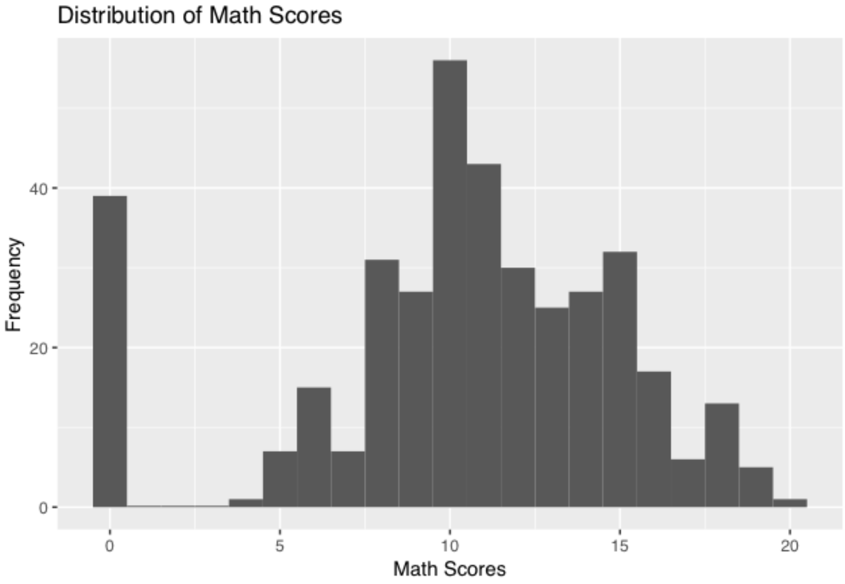
Jeffrey Cheng

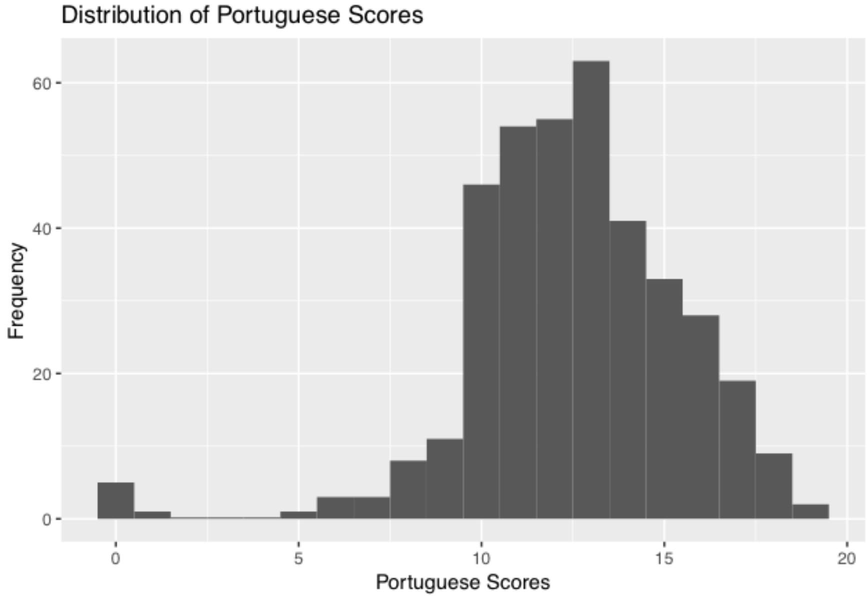
05-30-19

1 Introduction

This project analyzes a data set of 382 high school students who speak Portuguese. Students took both a Portuguese and a Mathematics course over the year and their quarterly and final grades are displayed, alongside various lifestyle statistics that the students were polled on.

2 Examining the Data





Math scores are generally lower than Portuguese scores, with more failures as well.

3 Math $\iff$ Portuguese Correlation

The first step was creating a simple linear regression model between a student's math score and his or her Portuguese Score. The results of the model and a 95 percent confidence interval for the slope of the model are displayed below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8203     0.9205   0.891    0.373
G3.y          0.7644     0.0716  10.676   <2e-16 ***
---

Residual standard error: 4.116 on 380 degrees of freedom
Multiple R-squared:  0.2307,  Adjusted R-squared:  0.2287
F-statistic:   114 on 1 and 380 DF,  p-value: < 2.2e-16

                 2.5 \%   97.5 \%
(Intercept) -0.9896571 2.630338
G3.y         0.6236251 0.905189
```

We can see here that math scores and Portuguese scores are clearly positively correlated. The confidence interval shows us that we are extremely confident that the correlation between math and Portuguese scores is actually positive. However, the R^2 value is fairly low. We are going to investigate this further by looking at students with a good score for one course and a low score for another.

4 Unusual Cases

There are only 2 students with a high math score and a low Portuguese score. Both of these students failed to take the final, leaving them with failing scores for the class. However, this is only the case for 1 out of 382 students and their result cannot be extrapolated. What is more interesting to look at is the students with mid/high Portuguese scores and low math scores, of which there are 26.

````
‘‘‘{r}
lowmath <- data %>%
  filter(G3.x < 6, G3.y > 10)
dim(lowmath)
````

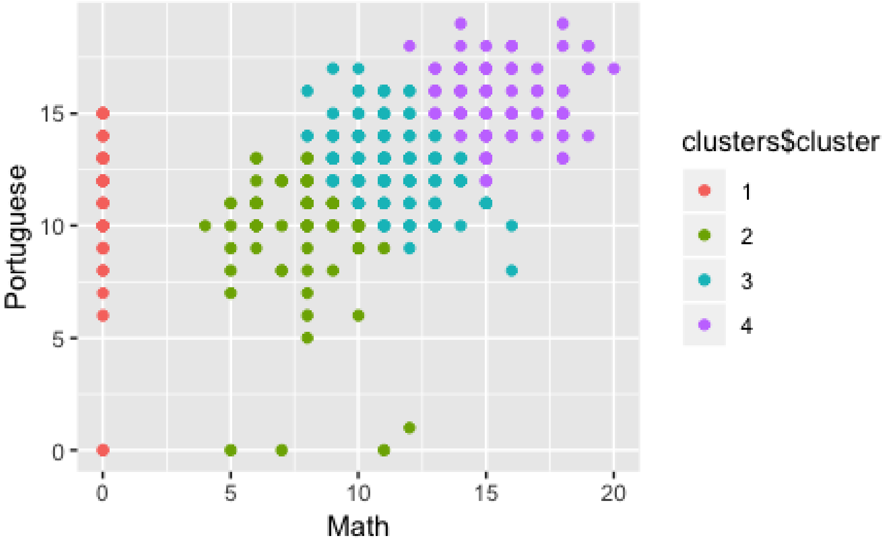
These 26 students were defined by having a math score lower than 6 and a Portuguese score higher than 10. Next, two linear regression models were set up: number of past classes failed vs. math score for the entire population, and number of past classes failed vs. math score for the subset of 26 students with low math scores but high Portuguese ones. The results of these models are also displayed below.

```
Call:
lm(formula = G3.x ~ failures.x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0983  -2.0983  -0.0983   2.9017   9.3481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.0983     0.2391  46.425  < 2e-16 ***
failures.x   -2.4464     0.3048  -8.027 1.26e-14 ***
---

Residual standard error: 4.34 on 380 degrees of freedom
Multiple R-squared:  0.145, Adjusted R-squared:  0.1427
F-statistic: 64.43 on 1 and 380 DF,  p-value: 1.264e-14

Call:
lm(formula = G3.x ~ failures.x, data = lowmath)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6759 -0.4276 -0.3034 -0.3034  4.6966

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3034     0.3201   0.948    0.353
failures.x    0.1241     0.2614   0.475    0.639

Residual standard error: 1.38 on 24 degrees of freedom
Multiple R-squared:  0.00931, Adjusted R-squared:  -0.03197
F-statistic: 0.2255 on 1 and 24 DF,  p-value: 0.6391
```

What's interesting to note here is that for the entire population, the number of previous classes failed is negatively correlated to both math and Portuguese scores. However, the unusual 26 students show no correlation between number of past courses failed and math/Portuguese scores. This implies that these students' struggle is not with the Portuguese content or study habits in general, but with this mathematics content in particular. So we have a good 10% of this population who have generally good study habits and high scores in Portuguese, but are struggling with this mathematics content.

5 Clustering Analysis

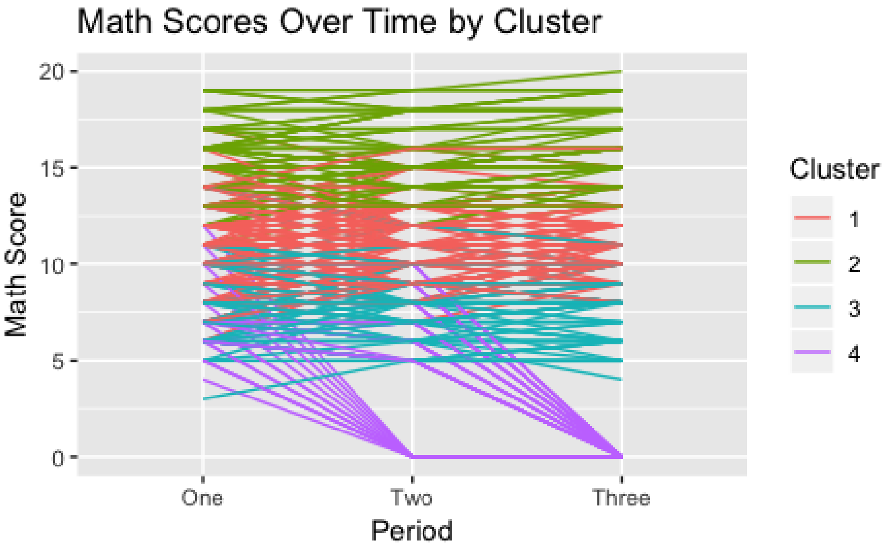
The next step was to use the kmeans clustering algorithm to cluster the students into four groups.

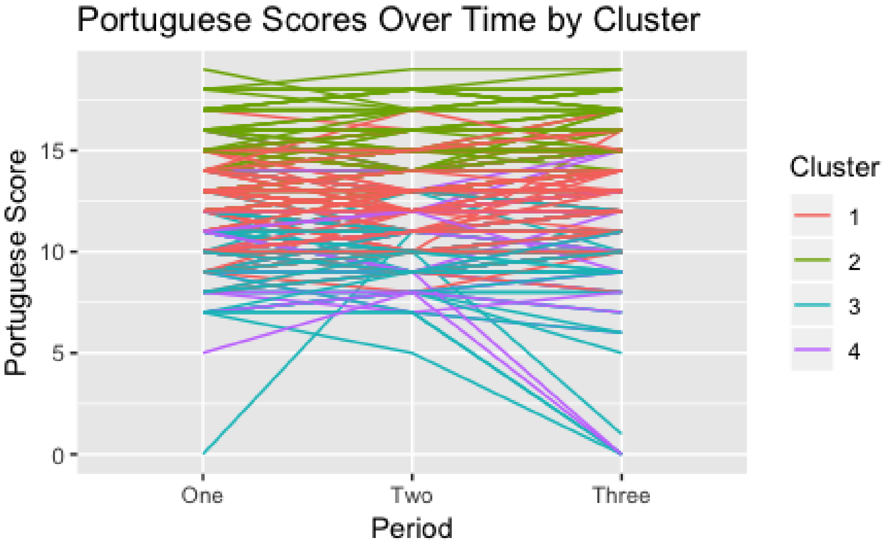


The students are clustered into roughly four groups.

1. A group with high proficiency in both subjects
2. A group with medium proficiency in both subjects
3. A group with low proficiency in both subjects
4. A group with varying Portuguese scores but no math proficiency

Because each student's scores over three periods are recorded, individual students progress can be tracked over time. The below graphs shows each student's math progression and their respective cluster.





The first thing to note is that there is much more upward mobility in Portuguese scores. In math scores, students have either slight upward and downward fluctuations while remaining stable or negative development. In Portuguese scores, that pattern remains the same but there are a decent amount of students who start with lower scores and are able to increase them by the end.

6 Conclusion

Among this high school class, there is clearly a difference in proficiency in math vs proficiency in Portuguese. The students as a whole are generally worse in math, whatever the reason may be for that. Furthermore, there is less of an ability to improve at math as opposed to Portuguese. Math scores over the course of the class remained relatively stable or dropped off, while some students were able to improve their Portuguese scores. This could be due to a number of factors: the quality of the mathematics class/faculty vs. that of the Portuguese class/faculty comes to mind. However, due to the linear regression models, it seems most likely that there is just a largely number of students who are unable to understand the mathematics content than the Portuguese content. This could explain the relatively higher decline in scores over the course of the classes for the math class as well: students who were initially able to do well in the class may have trouble with harder concepts later in the class and drop off.