The Relationship Between Romantic Relationships and School Absences in Math

Ghayasudin Ghayas

National University (NU)

Prof. Matthew Vanderbilt

Oct 26,2024

Purpose and Scope

The primary purpose of this analysis is to investigate the factors contributing to student absences in math classes, with a specific focus on whether students involved in a romantic relationship tend to have more absences. The analysis aims to provide insights that help the client understand the dynamics that lead to absenteeism, thus enabling more targeted interventions. The dataset used for this analysis is myData retrieval from kggale.come[1]

Data Generating Process (DGP) Story:

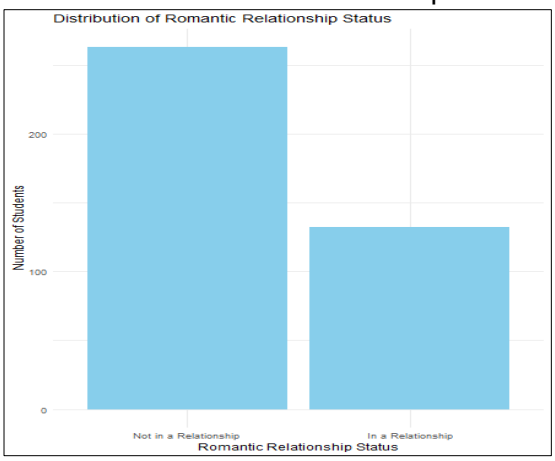
Absenteeism in school can result from a variety of factors including academic performance, social interactions, family circumstances, and extracurricular activities. Romantic relationships, which often emerge during adolescence, might impact student attendance. Students in romantic relationships could prioritize spending time with their partner over attending class, potentially leading to more absences. Furthermore, emotional stress or distraction from these relationships could contribute to reduced focus on academics, increasing the likelihood of absenteeism.

Research Question

The research question driving this study is: *Does being in a romantic relationship significantly affect the number of absences students have in math classes?* The hypothesis is that students in romantic relationships will have more absences compared to those who are not involved in such relationships.

Rationale:

Romantic relationships during school years can lead to shifts in priorities, affecting time management and attention to academic responsibilities. As these relationships often require emotional and time investments, students may skip classes to spend time with their partners, or the emotional challenges may distract them from academic focus, leading to more absences.



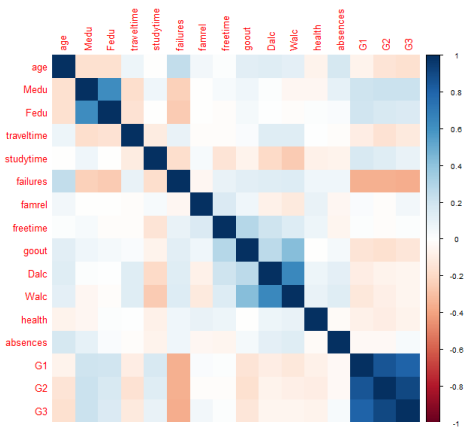
But in current date the majority of students (66.6%) reported not being in a romantic relationship, while a smaller portion (33.4%)

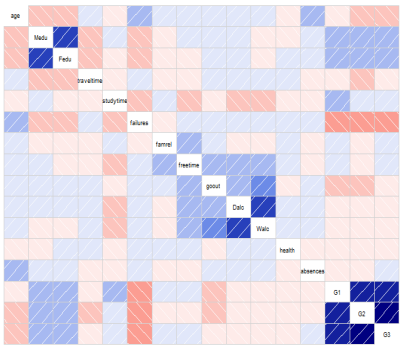
[1] 1 UCI Machine Learning Repository. (n.d.). *Student alcohol consumption*. Retrieved October 26, 2024, from https://www.kaggle.com/uciml/student-alcohol-consumption ml/student-alcohol-consumption

indicated that they are in one. This insight could be useful for understanding the social dynamics among the students in your dataset.

Corplot and Corgram:

The Corplot easily shows the correlation of numerical data. We can observe that the color coding effectively illustrates these correlations. The dark blue diagonal indicates a high correlation between the G1, G2, and G3 grades, which intuitively makes sense; if a student has good grades in the first and second grades, they are likely to achieve good grades in the final grade as well. Conversely, poor grades in G1 and G2 will likely result in a lower grade in G3. Additionally, the feature related to the number of past classes shows a strong correlation with the parental education levels, suggesting that parental education significantly influences the students' academic performance.





The Corgram also displays a wealth of information. The main distinction between Corplot and Corgram is that Corplot works exclusively with numerical data, whereas Corgram can work with data frames, providing a more comprehensive view of the relationships within the data.

Five-Number Summaries of Quantitative Demographic Variables

The age distribution of the dataset reveals that students range from 15 to 22 years old, with a median age of 17 and a mean of 16.7, indicating that the majority are in their late teens. In terms of absences, while most students have low absence rates, with the first quartile and median at 0 and 4 absences respectively, there is a notable skew in the data: the mean is 5.7, but some students have up to 75 absences, indicating a wide range and some high absenteeism that could impact overall trends in the dataset.

Summary- Number of the Quantitative Demographic Variables

1- Age

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15.0	16.0	17.0	16.7	18.0	22.0

2- Absence

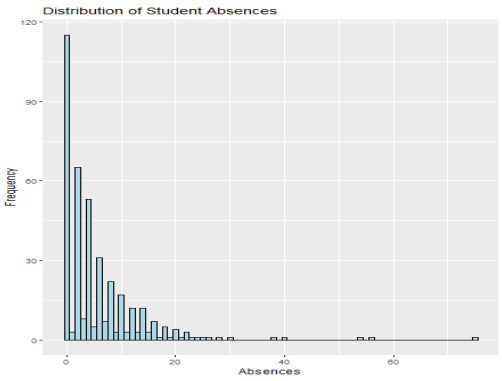
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	4.000	5.709	8.000	75.000

Histogram of Absences

The histogram for student absences is right-skewed, indicating that most students have a low number of absences, while a smaller group has significantly higher absence rates. This skew suggests that excessive absences are relatively uncommon in the dataset.

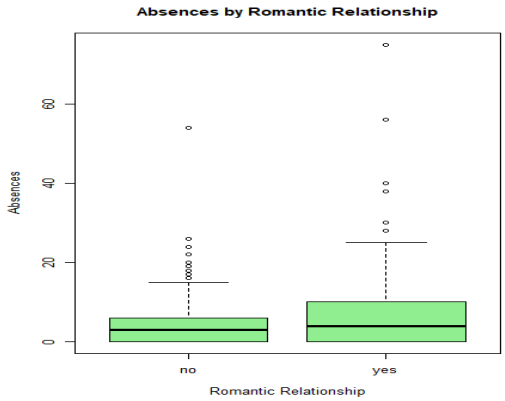
Visualization of Research Question

The box plot shows that students with and without a romantic relationship have similar distributions of absences, with both groups having a median around the same level. However, both groups show a wide range in absence counts, with several outliers, especially for students in relationships, where absences extend to over 60. This suggests that while having a romantic relationship does not significantly change the median absences, there may be more variability in absence rates among students in relationships.



Research Question: Absences Explained by Romantic Relationship

Empty Model



The output from the empty model, which predicts student absences based solely on the mean, reveals that the average number of absences is approximately 5.71 days. The model's intercept is statistically significant ($p < 0.001$), indicating that this average is a meaningful baseline. However, the residuals show a large range, with a standard error of 8.003, suggesting considerable variation in absences among students that isn't explained by the model. This model serves as a simple starting point for comparison with more complex models that might include additional predictors.

Explanatory Model

```
> myData <- read.csv(file = "student-mat-1.csv", header = TRUE)
>  empty_model <- lm(absences ~ 1, data = myData)
> summary(empty_model)

Call:
lm(formula = absences ~ 1, data = myData)

Residuals:
   Min     1Q Median     3Q    Max
-5.709 -5.709 -1.709  2.291 69.291

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.7089     0.4027   14.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.003 on 394 degrees of freedom
```

The calculated F-ratio of 9.47 compares the explanatory power of the "explanatory model" (a model with predictors) against the "empty model" (a model without predictors, only the mean). An F-ratio of 9.47 suggests that the explanatory model provides a significantly better fit to the data than the empty model, as it explains more of the variance in absences than the mean alone. If this F-ratio is significantly higher than what would be expected by chance, it indicates that the added predictors in the explanatory model improve the model's performance in explaining the variation in absences.

Model Statement:

The Welch Two Sample t-test compares absences between students not in a romantic relationship (group "no") and those in a relationship (group "yes"). The test yields a t-value of -2.576 and a p-value of 0.01083, indicating a statistically significant difference in mean absences between the two groups ($p < 0.05$). Specifically, students in a relationship have a higher average of 7.44 absences, while those not in a relationship average 4.84 absences. The Proportion Reduction in Error (PRE) value of **0.0235268** indicates that the model explains approximately **2.35%** of the variability in student absences beyond what would be predicted by the mean. This relatively low PRE suggests that the model does not account for a significant portion of the variance in absences, indicating that other factors not included in the analysis may play a more substantial role. Given that PRE values range from **0** (no explanatory power) to **1** (perfect explanatory power), a value of **0.0235** reflects limited explanatory capability. Therefore, it may be beneficial to explore additional variables,

interactions, or different modeling techniques to better capture the complexity of factors influencing student absences. Additionally, ensuring that categorical variables, such as romantic relationships, are appropriately coded and tested for interactions with other relevant variables may enhance model performance.

Conclusion

From the models, we find that romantic relationships are a significant predictor of student absences, with students in relationships missing more school days on average. However, the effect size is moderate, and there may be other unmeasured factors affecting absences.

Limitations:
Potential limitations include missing data, self-reported relationship status, and the fact that the analysis only captures correlations, not causation.