Does Student Romantic Involvement Explain Absenteeism?

Randall C. Crawford

National University

ANA600 Fundamentals of Analytics

Abstract

Per client request, the objective of this analysis is to understand a potential cause of student absences in math classes. The client hypothesizes that students who are in a romantic relationship will tend to have more absences. Data was provided for the analysis, and was obtained in a survey of student math courses in a secondary school. It contains a lot of interesting social, gender, and study information about 395 students. For the purpose of this analysis, only two variables will be used in the evaluation:

23. absences – number of school absences (numeric: 0 to 75)
30. romantic – with a romantic relationship (binary: yes or no)

Does Student Romantic Involvement Explain Absenteeism?

When considering the potential causes of student absenteeism, it seems logical that many factors could contribute to it, and not just romantic involvement. The data provided has many other variables that could be included in the analysis:

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. 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)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)

I just wanted to provide some ideas as to what is available from the survey that was taken, and note that much of this data could possibly be used to explain absenteeism. Regarding the specific objective, romantic involvement could potentially be considered time-consuming, possibly putting some demands on a student's schedule. On the other hand, having an intimate partner can be a blessing, when it comes to sharing the responsibilities of everyday tasks, possibly removing some demands from a student's schedule. It should be interesting to see what the data reveals.

The Research Question

In trying to better understand the relationship of the outcome variable for absenteeism, *absences*, and the explanatory variable for romantic involvement, *romantic*, a word equation should be put together to help generalize what we are after with creating a linear model.

$$\text{absences} = B_0 + B_1(\text{romantic}) + \text{error}$$

The first step will be to statistically evaluate the outcome and explanatory variable data, and create some visualizations concerning them. The second step will be to create an empty model to better understand the outcome variable, *absences*, and then create an explanatory model including the outcome variable, *absences*, and the explanatory variable, *romantic,* similar to the word equation depicted above. The third and final step will involve model comparisons, and the drawing of conclusions.

Table of Contents

- Page 1 – Document Cover Page
- Page 2 – Abstract
- Page 3 & 4 – Does Student Romantic Involvement Explain Absenteeism?
- Page 4 – The Research Question
- Page 5 – Table of Contents
- Page 6 & 7 – Data Description and Visualization
- Page 8 – Constructing The Empty and Explanatory Models
- Page 9 – Explanatory Model Results and Conclusions
- Page 10 – References
- Page 11 & 12 – Tables
- Page 13, 14, 15, & 16 – Figures

Data Description and Visualization

Generate Frequency of the Categorical Variable for Romantic Involvement.

Out of 395 students, 263 students were not romantically involved, and 132 students were romantically involved (see **Table 1**).

Five Number Summary of the Quantitative Variable for Absences.

0 absences seemed to be the most frequent result, *mode,* and represented the *minimum* value. Although, 4 absences fell in the center of the data set, *median*. The average value was 5.709, *mean*. 75 absences represented the *maximum* value, but the majority of results fell with 0-8 absences, *IQR* (see **Table 2)**.

Evaluate Histogram of the Quantitative Variable for Absences.

The shape appears to be mostly unimodal with a notable apex at the *mode* (0), and there are some outliers above 20 absences. The *mode*, illustrated by the dark green vertical line, is less than the *median* (4), illustrated by the center line in the purple box, that is less than the *mean* (5.709), illustrated by the dark red vertical line, indicating a Right skew to the distribution. The purple box represents the Inner Quartile Range (*IQR*) from 0-8, and accounts for the majority of the results. The spread makes sense based upon the general desire for students to succeed at their studies and the benefit garnered from attending lectures (see **Figure 1)**.

Generate Descriptive Statistics of Absences Respecting Romantic Involvement.

Both choices for romantic involvement (yes/no) had 0 absences as the most frequent result, *mode,* and represented the *minimum* value. *Median* values were reasonably close with 4 absences for yes and 3 absences for no. *Mean* values had a notable difference with 7.439 for yes and 4.840 for no, and *Standard Deviation* values differed with 10.788 for yes and 5.989 for no, for which both could bode well for the client's hypothesis (see **Table 3).**

Evaluate Absences Histograms Faceted by Romantic Involvement.

The histogram for those responding with yes to romantic involvement shows more exaggerated Right skew and a larger Inner Quartile Range (IQR), purple box, than the histogram for those responding with no to romantic involvement (see **Figure 2**).

Evaluate Absences Boxplots Faceted by Romantic Involvement.

The boxplots with jitterplot overlays give a more pronounce, but similar story as the histograms with boxplot overlays, when looking at the respective Inner Quartile Ranges (IQRs), purple boxes, and the dispersal of values (see **Figure 3**).

Evaluate Absences and Romantic Involvement Scatterplot.

The scatterplot gives a more linear view of the dispersal of values, but visually shows the difference in Absences *means* between no and yes romantic involvement (see **Figure 4**).

Constructing the Empty and Explanatory Models

Fit the Empty Model for the Outcome Variable for Absences.

$$Y = b_0 + e$$

Y = estimated absences, b_0 = mean of 5.709 absences, e = error

The empty model to estimate absences was created in R for comparative reference to the explanatory model. It had a sum of squares equal to 25,236 absences squared. This will be our baseline to see what adding the explanatory variable for romantic involvement has accomplished (see **Table 4**).

Recode the Explanatory Variable for Romantic Involvement.

A numeric variable, *romantic_num*, had to be created from the string variable, *romantic*, with "no" = 0 and "yes" = 1, such that the target Explanatory Model would function correctly.

Fit the Explanatory Model for the Outcome Variable for Absences and the Explanatory Variable for Romantic Involvement.

$$Y = b_0 + b_1(X_1) + e$$

Y = estimated absences, b_0 = mean of 4.840 absences with a "no" romantic value, b_1 = 2.599 the difference between the "yes" and "no" romantic value means, X_1 = the binary value of 0 or 1 for "no" and "yes" romantic values, e = error

The explanatory model to estimate absences with romantic involvement was created in R. The p-value of 0.0022 associated with *romantic_num* variable in the model had a less than *alpha* = 0.003, which is better than 99.7% confidence level demonstrating significance (see **Table 5**).

Explanatory Model Results and Conclusions

Evaluating Explanatory Model Results.

Though the model F statistic of 9.469 and the p-value of 0.0022 indicate significance, the sum of square difference compared to the empty model is notable:

PRE = 0.0235 = 593.711 (SS-Model Error Reduction) / 25235.519 (SS-Empty Model)

The PRE of 0.0235 can be translated to 2.35% of error reduction in the explanatory model. This leaves 97.65% of error to be explained. The Cohen's D value of 0.328 fell within the small and medium size effect standards (see **Table 6**).

Conclusions

The difference in *mean* absences noted in **Table 3** and illustrated in **Figure 4** was a telltale sign that romantic involvement would at least explain some of what was being seen in absences. The significance of romantic involvement was confirmed to a 99% confidence level, and the Cohen's D value indicated a small to medium size effect. Although, with the small PRE value, romantic involvement only explains 2.35% of the model. As previously mentioned in the introduction, there is plenty of room for other variables to be added to help explain absences, but looks like romantic involvement should be one of them.

References

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.

University of Camerino

Fabio Pagnotta, fabio.pagnotta@studenti.unicam.it

Hossain Mohammad Amran, mohammadamran.hossain@studenti.unicam.it

Tables

Table 1: Frequency of *romantic*

romantic	
no	yes
263	132

Table 2: Five Number Summary for *absences*

absences - Five Number Summary					
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0	0	4	5.709	8	75

Note: Inner Quartile Range (*IQR*) = 3rd Quartile – 1st Quartile

Table 3: Descriptive Statistics for *absences* respecting *romantic*

absences by romantic - Descriptive Statistics							Standard Deviation	Sample Count
romantic	Minimum	1st Quartile	Median	3rd Quartile	Maximum	Mean		
no	0	0	3	6	54	4.840	5.989	263
yes	0	0	4	10	75	7.439	10.788	132

Table 4: ANOVA for the Empty Model for *absences*

Analysis of Variance Table

Response: absences

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	394	25236	64.05	N/A	N/A

Table 5: ANOVA for the Explanatory Model for *absences* with *romantic*

Analysis of Variance Table

Response: absences

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Significance
romantic_num	1	593.7	593.71	9.4688	0.002236	**
Residuals	393	24641.8	62.7	N/A	N/A	N/A

Significance Codes:	'***'< 0.001	'**'< 0.01	'*'<0.05	'.'< 0.1

Table 6: SUPERNOVA with Cohen's D for the Explanatory Model for *absences* with *romantic*

Analysis of Variance Table (Type III SS)
Model: absences ~ romantic_num

	SS	df	MS	F	PRE	p	Cohen's D
Model (error reduced)	593.711	1	593.711	9.469	0.0235	0.0022	0.328
Error (from model)	24641.808	393	62.702	N/A	N/A	N/A	N/A
Total (empty model)	25235.519	394	64.05	N/A	N/A	N/A	N/A

Relevant Confidence Levels	99.9%	99.7%	99.0%	95.0%	90.0%	
Corresponding Alpha	0.001	0.003	0.01	0.05	0.1	p < Alpha for significance

Effect Size	Small	Medium	Large
Cohen's D	0.2	0.5	0.8

Figure 1: Histogram with Boxplot Overlay for *absences*

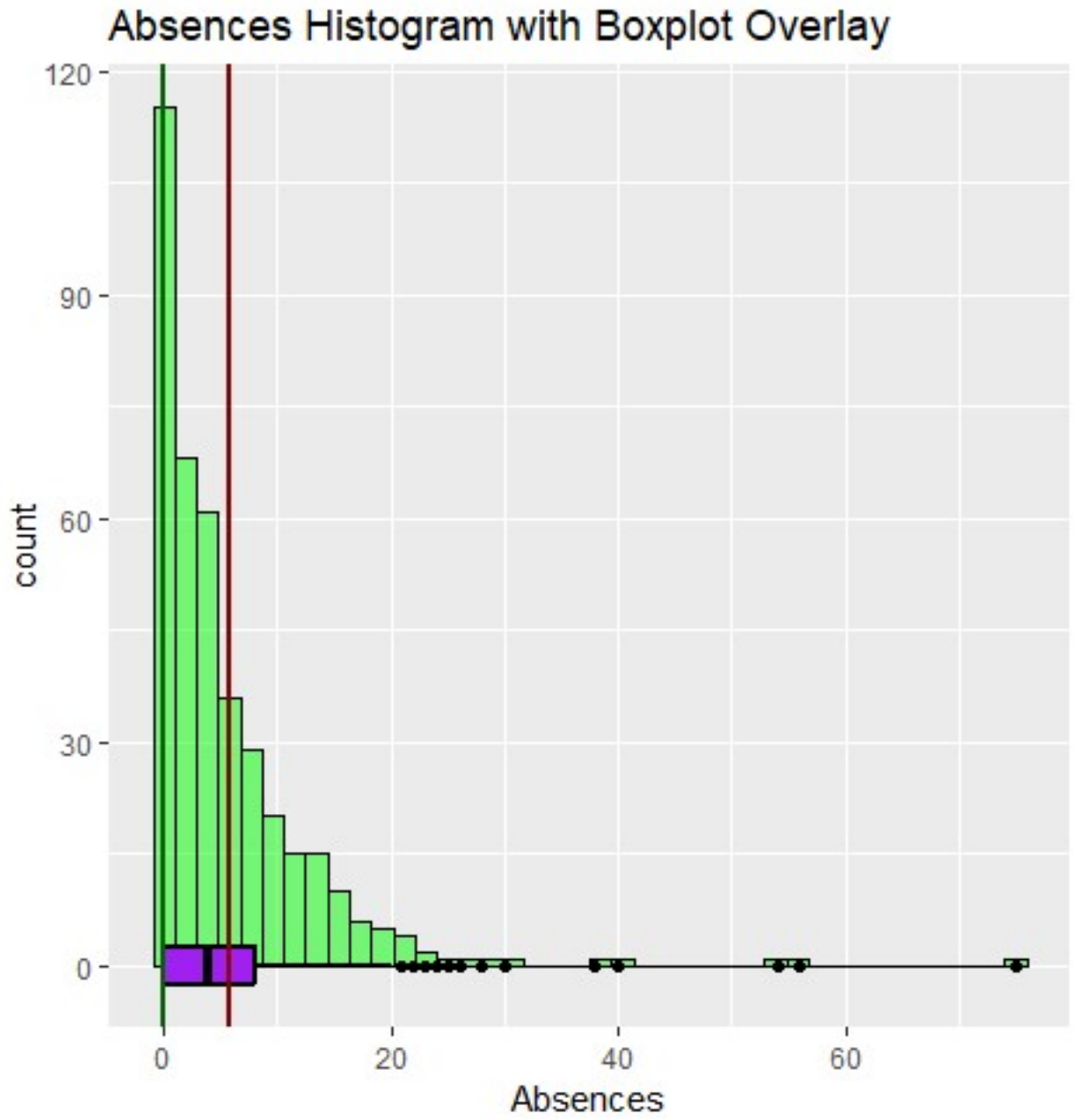

Figure 2: *absences* Histograms with Boxplot Overlays faceted by *romantic*

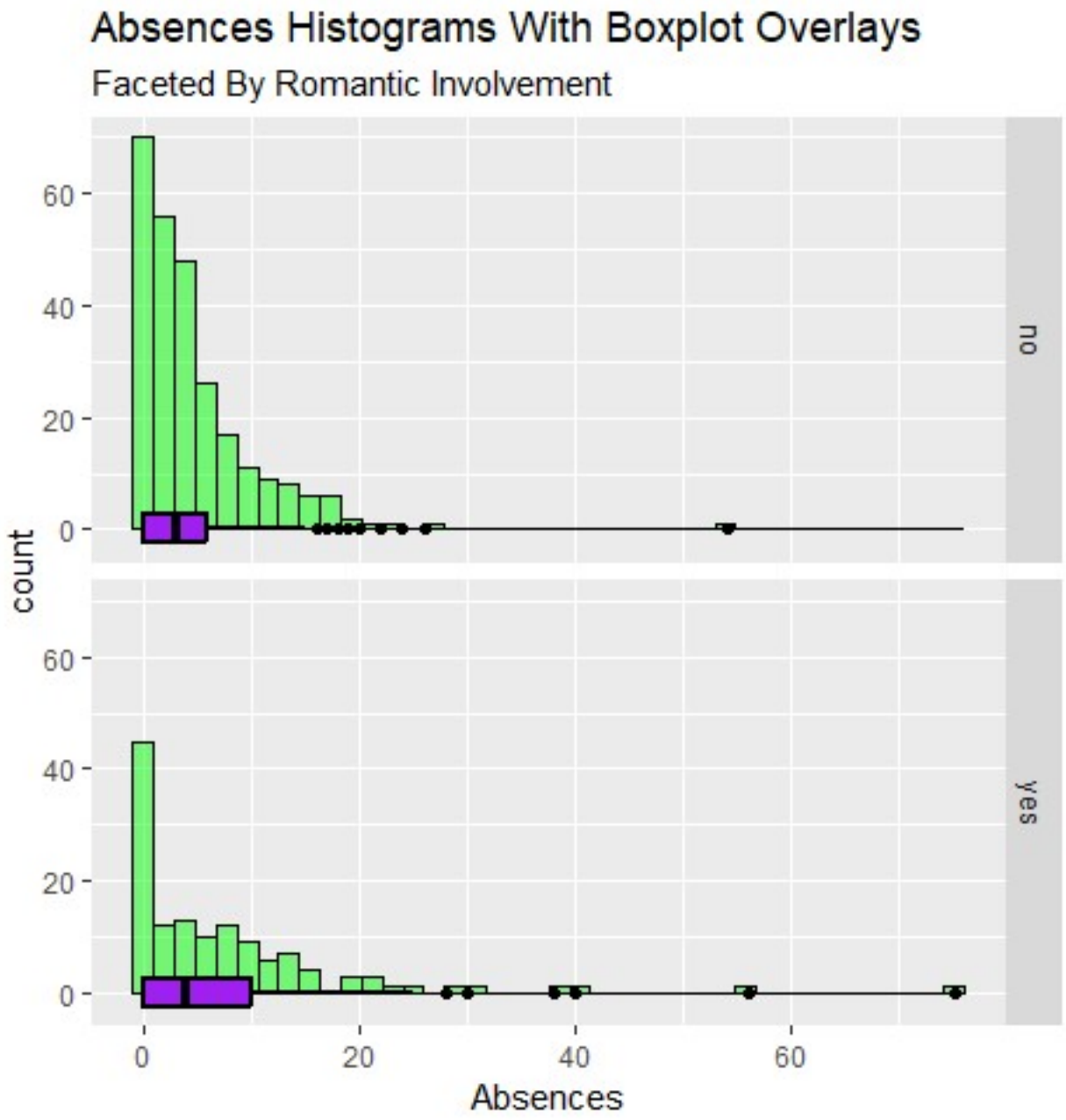

Figure 3: *absences* Boxplots with Jitterplot Overlays faceted by *romantic*

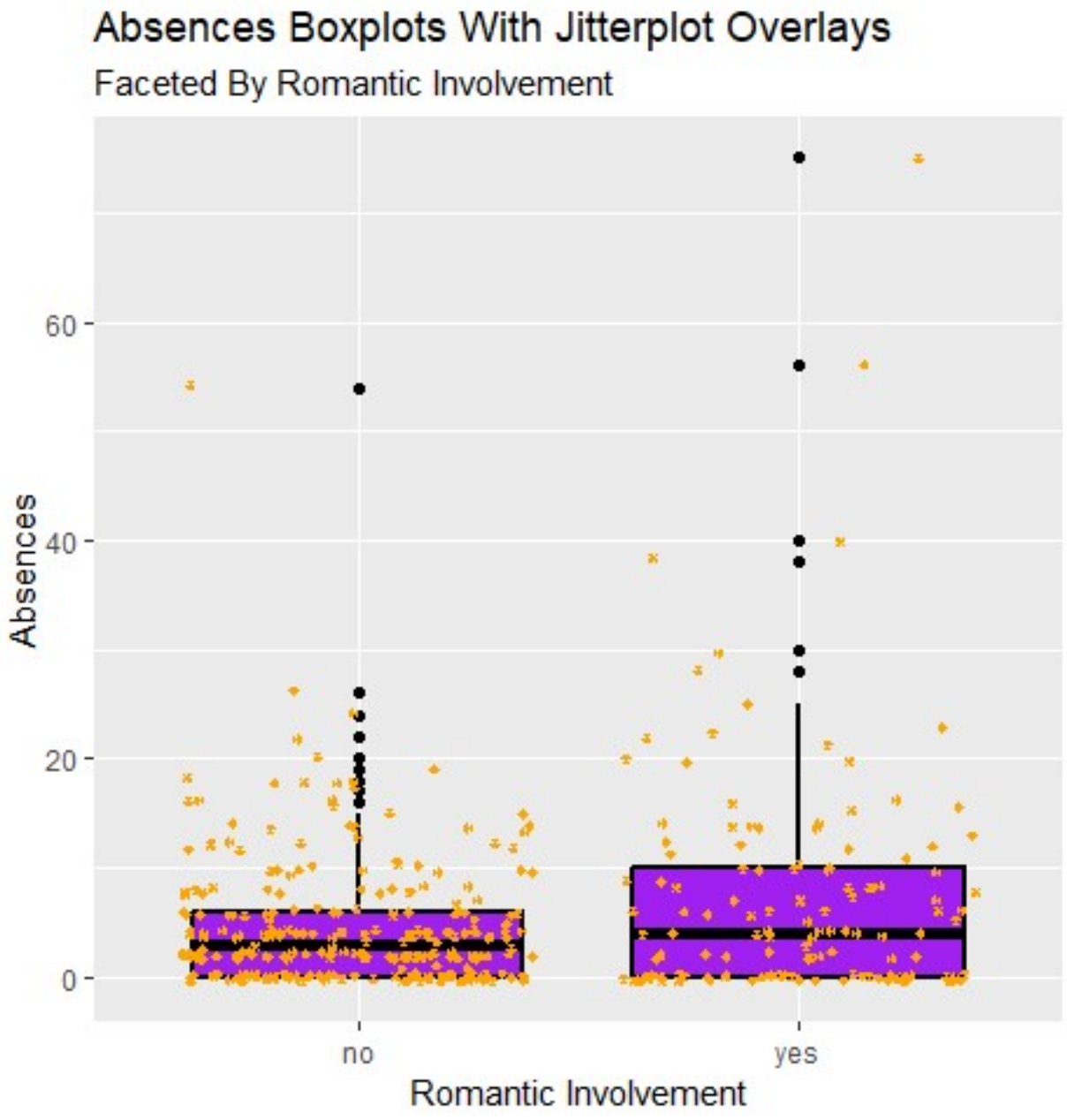

Figure 4: *absences* and *romantic* Scatterplot