Engagement EDA

William Morgan

13 July, 2018

Purpose:

Run preliminary correlational analyses to uncover possible relationships between student outcomes and various measures of faculty engagement. This will inform our approach to modeling later.

Forms of engagement to inspect:

- Presence of a Hallway Conversation board
  - Blackboard discussion forums sometimes have boards to emulate hallway conversations. These boards are meant to imitate natural discussion that would happen between students and instructors between classes
- Faculty Post Consistency
- Post Quantity

Outcomes of interest:

- Student Perception of Faculty Engagement (`instr_score`)
  - These scores come from course evaluations; there are three questions pertaining to faculty presence and engagement which we average into a single score
- Student Posts (`posts_per_student`)
  - The average number of posts per student throughout the duration of the course
- Student Posts Consistency (`stu_post_consistency`)
- Student Grades (`avg_grade`)
  - Average grade received in a course-section

General Outline

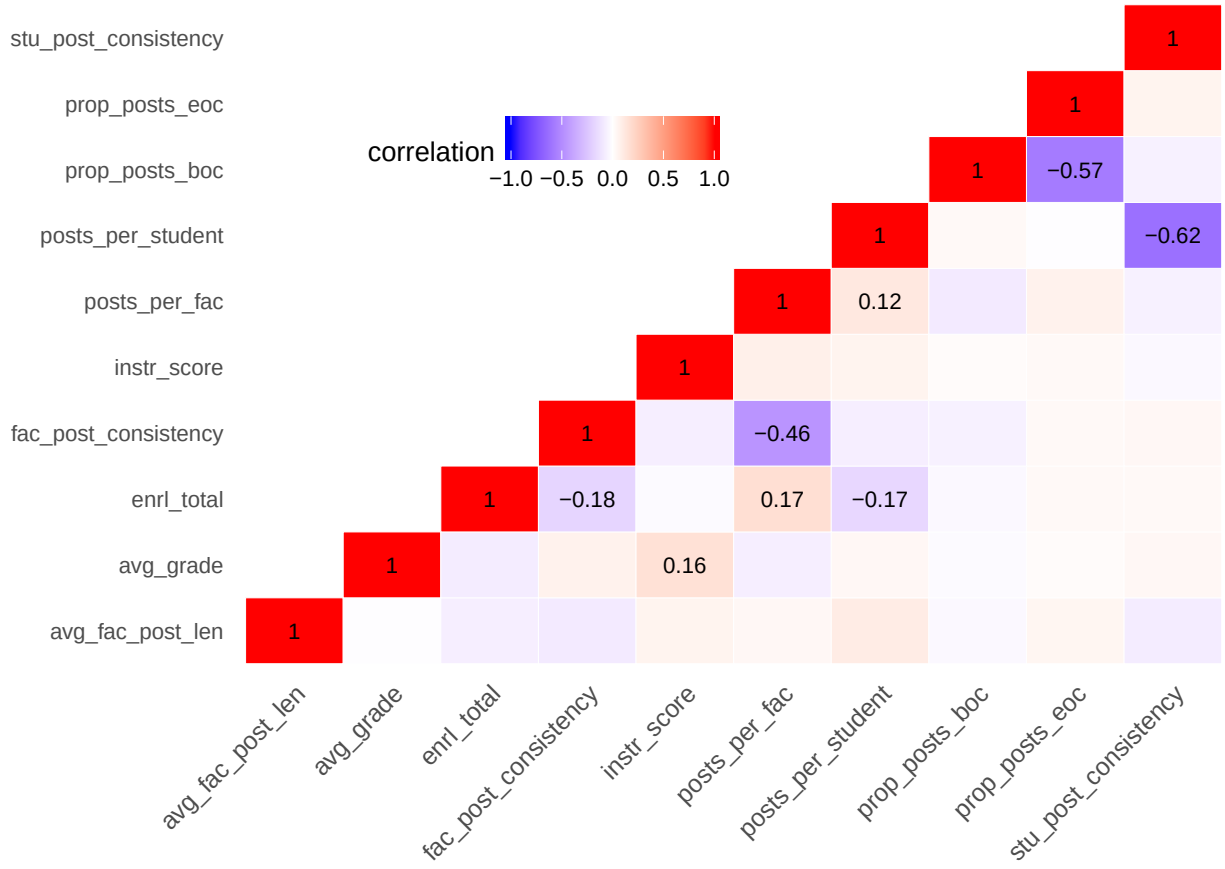
The goal is to understand which features correlate most with the several outcomes. As a first pass we'll create a correlation heatmap relating the continuous features to each of the outcomes. This will give us some basic insight on any apparent linear relationships. Next we'll move on to plotting the features individually to get a sense of any nonlinear relationships. In particular, we are looking for evidence favoring the inclusion of polynomial terms in a regression. We'll conclude with similar plots for the categorical variables (i.e. plotting a feature against a particular outcome).

Before moving on, let's preview the data to remind ourselves what we'll be working with.

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    3019 obs. of  14 variables:
##  $ avg_fac_post_len    : num  336 840 157 568 182 ...
##  $ avg_grade           : num  3.29 3.09 3 3.36 3.07 ...
##  $ course_id           : chr  "2015SummerB-X-AST111-42994-42993" "2015SummerA-X-OGL300-43650-44146" "
##  $ enrl_total          : num  28 19 7 119 89 30 72 36 50 137 ...
##  $ fac_post_consistency: num  0.364 0.363 1.029 0.348 0.182 ...
##  $ has_hallway         : Factor w/ 2 levels "0","1": 2 2 1 1 2 2 2 2 2 2 ...
##  $ instr_score         : num  4.03 4.45 4.96 4.74 4.43 ...
```

```
##  $ posts_per_fac       : num  14 13 4 34 9.33 ...
##  $ posts_per_student   : num  2.29 22.58 1.43 7.84 19.75 ...
##  $ prop_posts_boc      : num  0.393 0.769 1 0.353 0.821 ...
##  $ prop_posts_eoc      : num  0.25 0 0 0.176 0 ...
##  $ session_a           : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 2 1 1 1 ...
##  $ stu_post_consistency: num  4.09 0.51 2.352 1.773 0.862 ...
##  $ upper_division      : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 1 2 2 1 ...
```

Correlation Heatmap

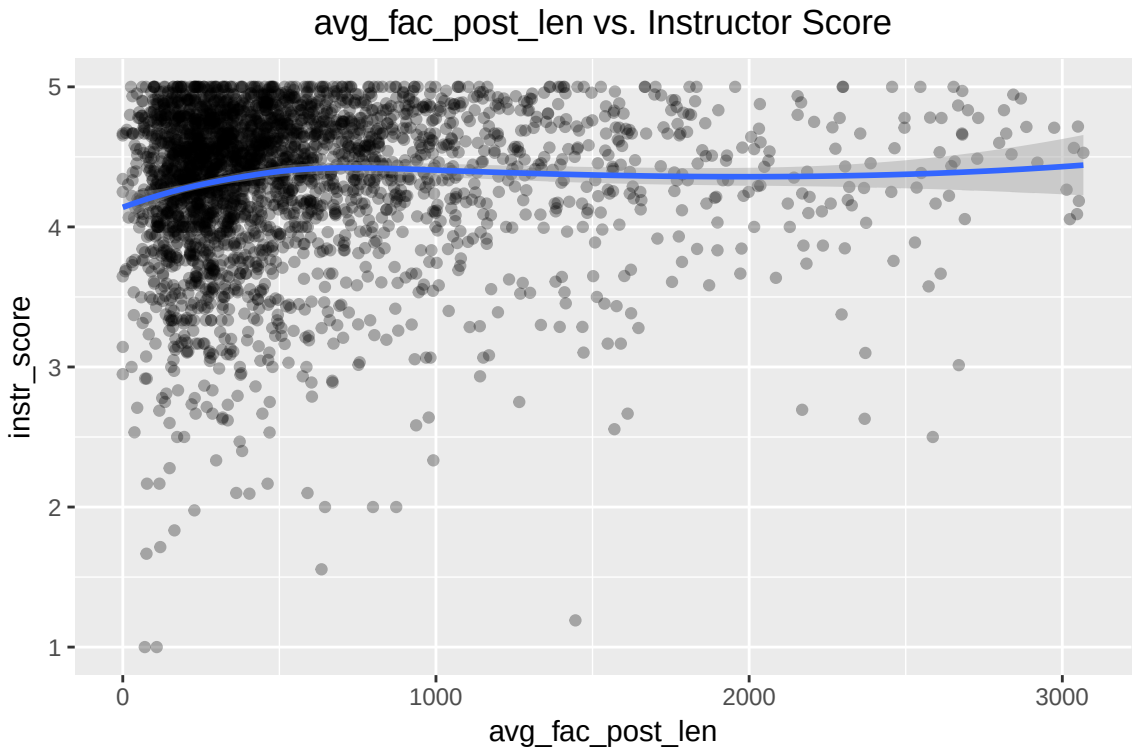


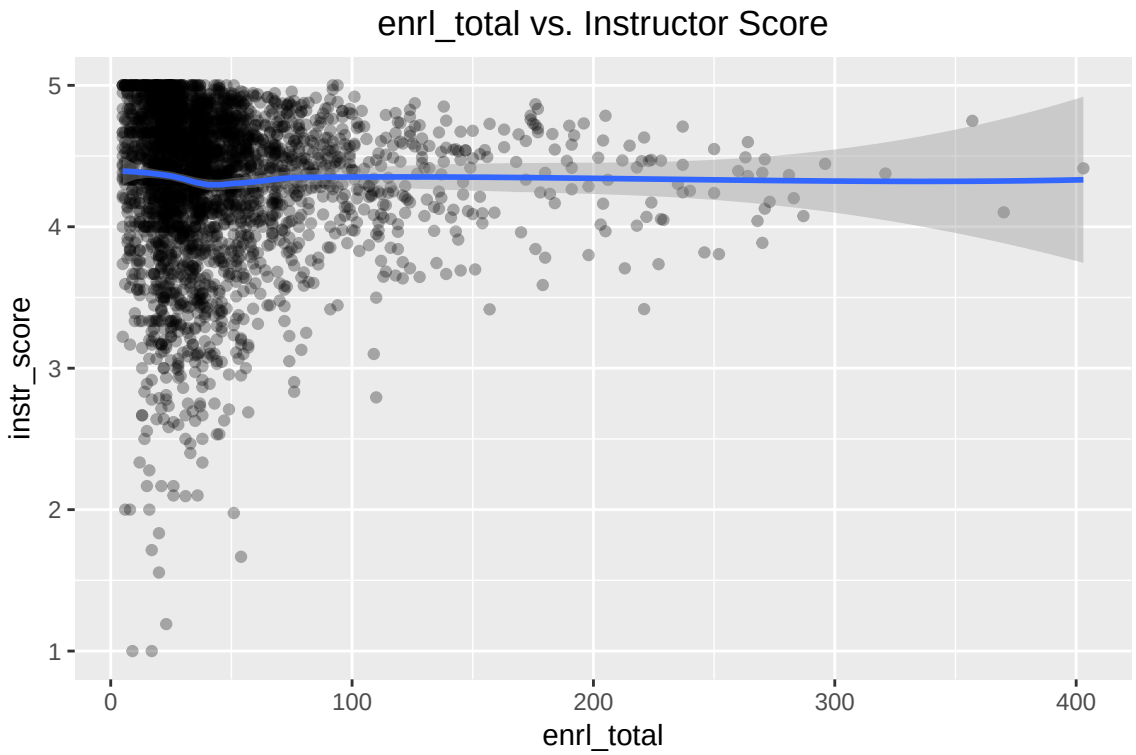
Discussion

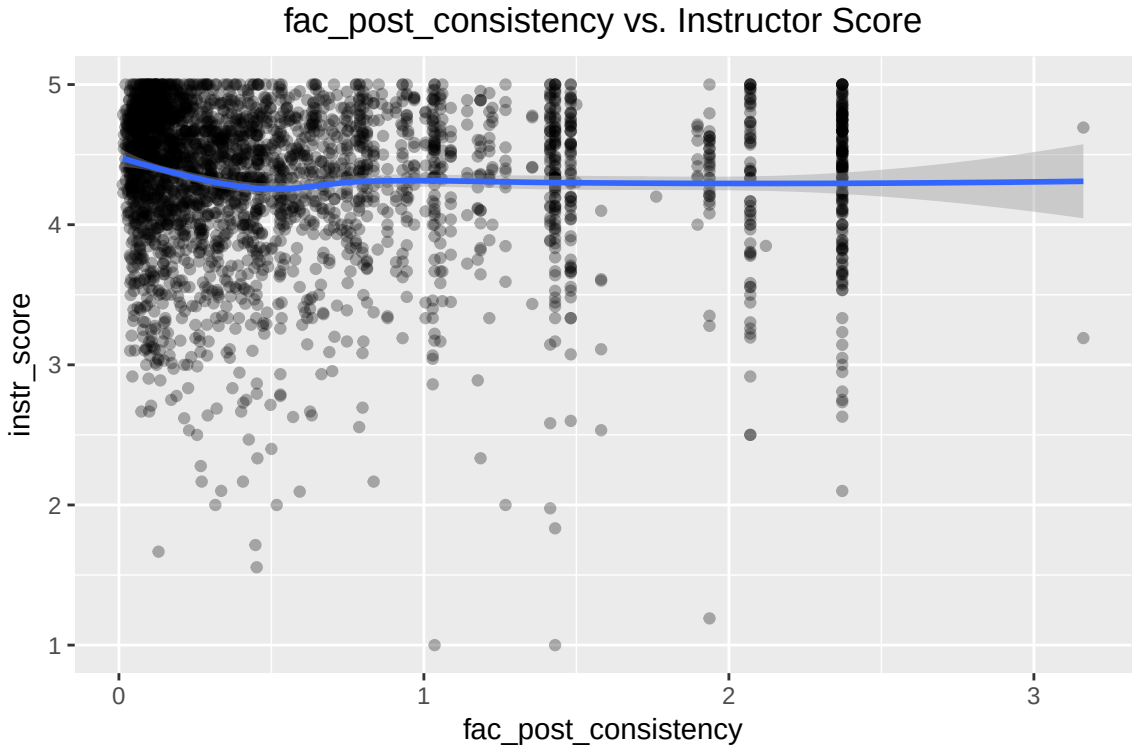
`instr_score` appears to be unrelated with every variable we were considering to include in a regression model. This doesn't mean we should throw it out necessarily, but it does indicate that is has a very very weak linear association with all other variables.`posts_per_student` has a somewhat negative relationship with the number of students in the course and a slightly negative correlation with the proportion of faculty posts in the first two weeks (`prop_posts_boc`) but for the most part the situation is the same. `avg_grade` unfortunately appears to be just as uncorrelated as the other outcomes. We'll have to move on to doing more general plots to uncover something. Just like the others, `stu_post_consistency` does not seem to have any strong correlations with the explanatory variables.

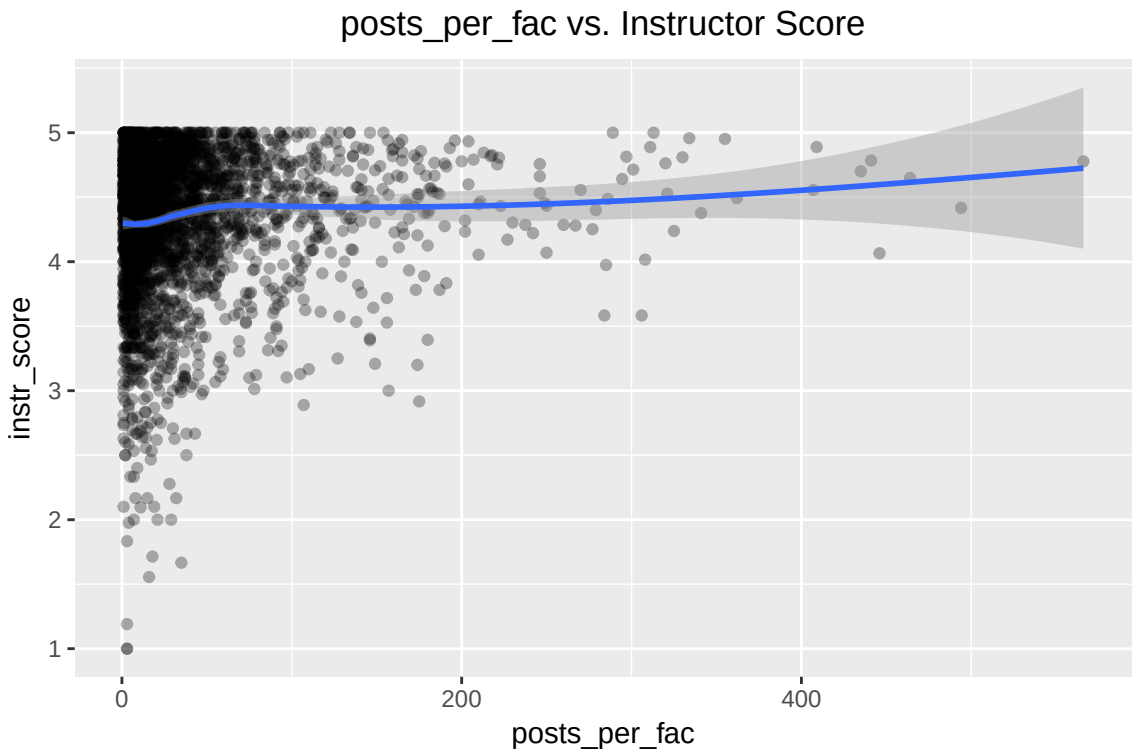
Two-way plots

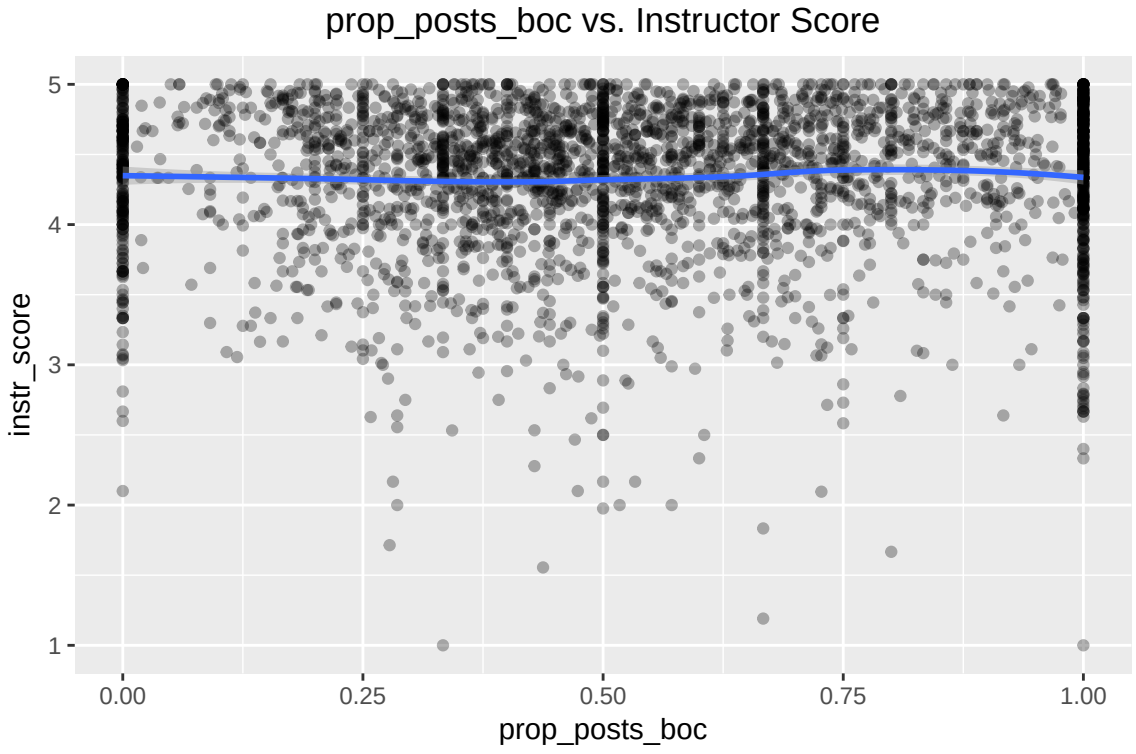
Instructor Score plots

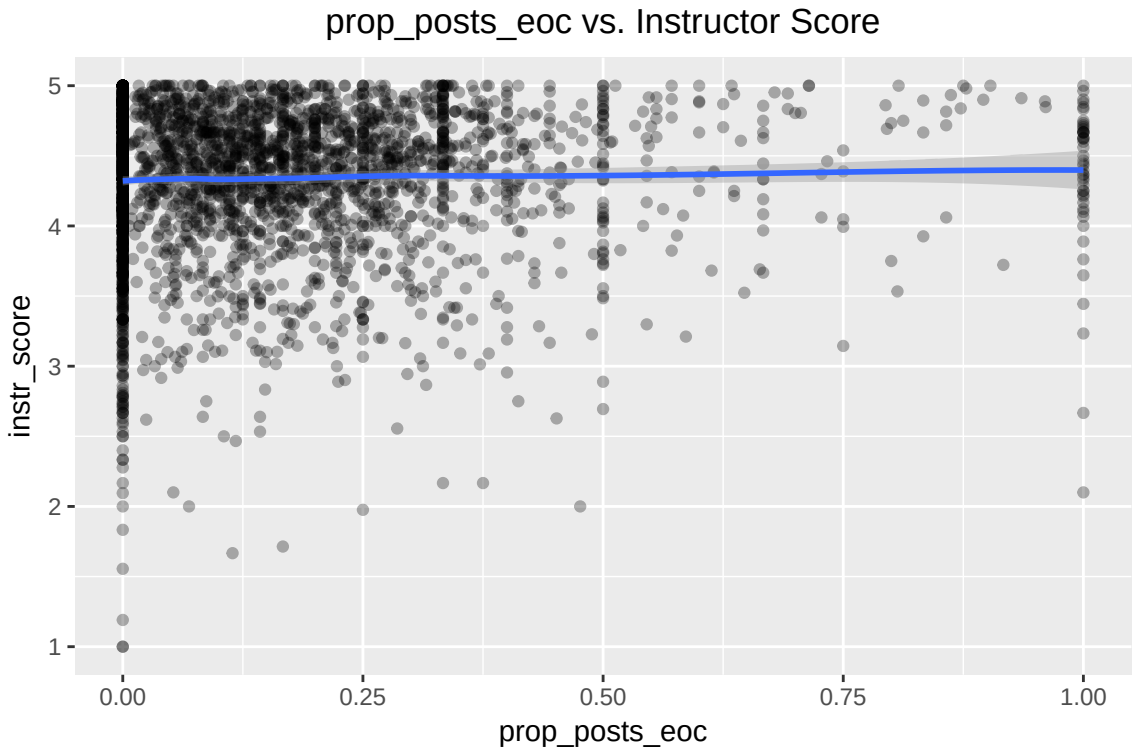



enrl_total vs. Instructor Score
instr_score
enrl_total
5
4
3
2
1
0
100
200
300
400


fac_post_consistency vs. Instructor Score
instr_score
fac_post_consistency
5
4
3
2
1
0
1
2
3


posts_per_fac vs. Instructor Score
instr_score
5
4
3
2
1
0
200
400
posts_per_fac

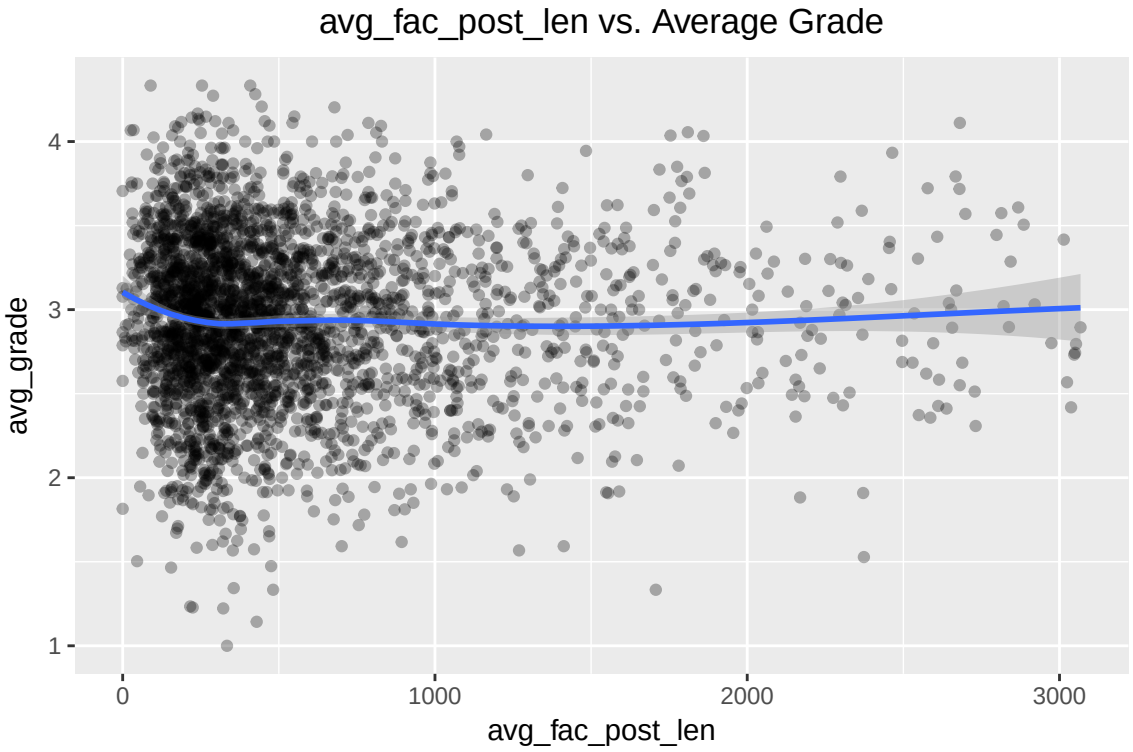

prop_posts_boc vs. Instructor Score
instr_score
5
4
3
2
1
0.00
0.25
0.50
0.75
1.00
prop_posts_boc

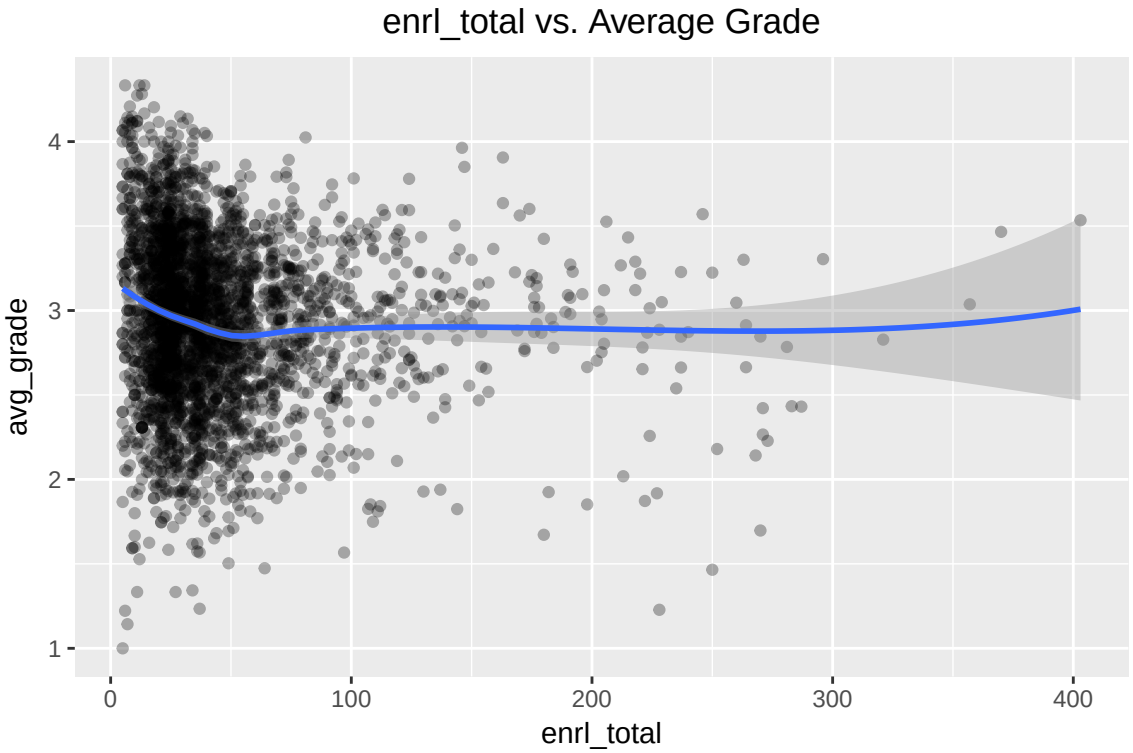


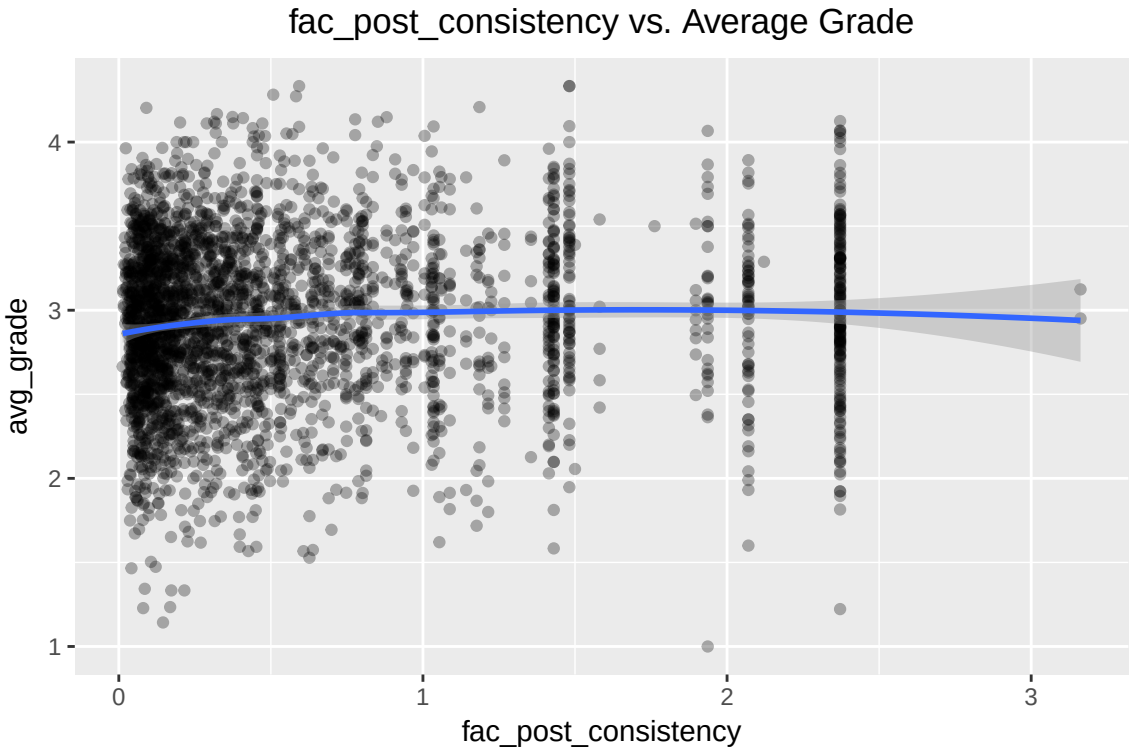
Discussion

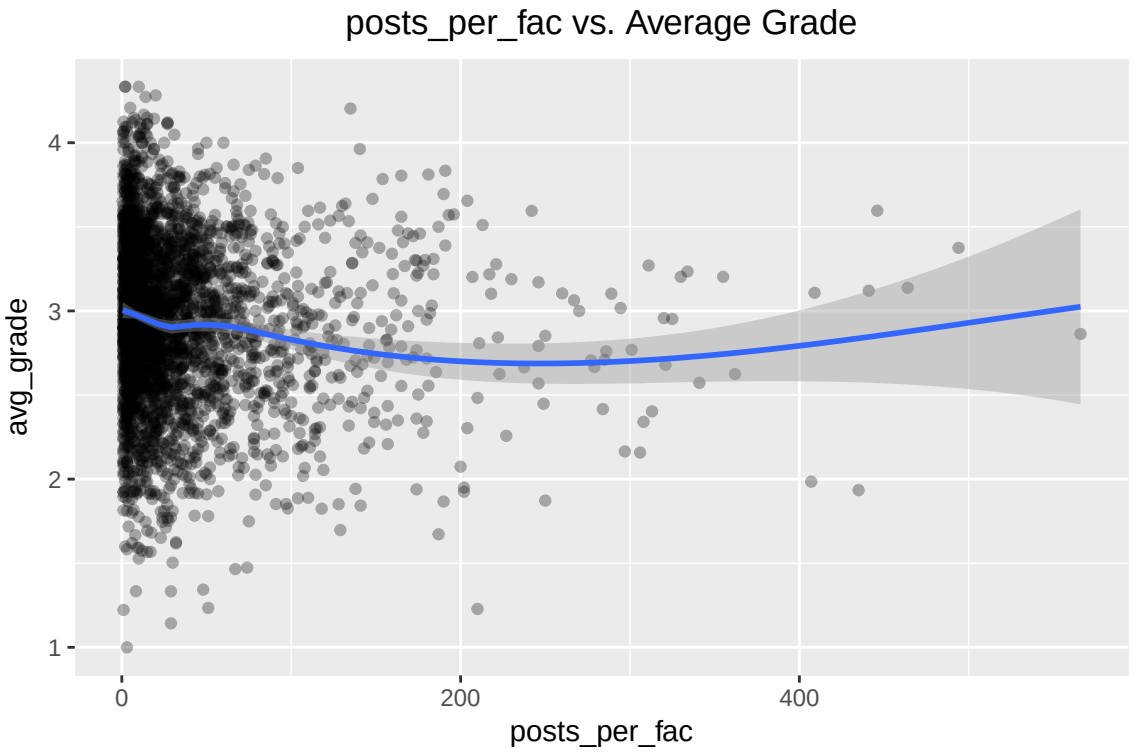
- Surprisingly, each of the six variables appears to have practically no effect on the instructor evaluation scores

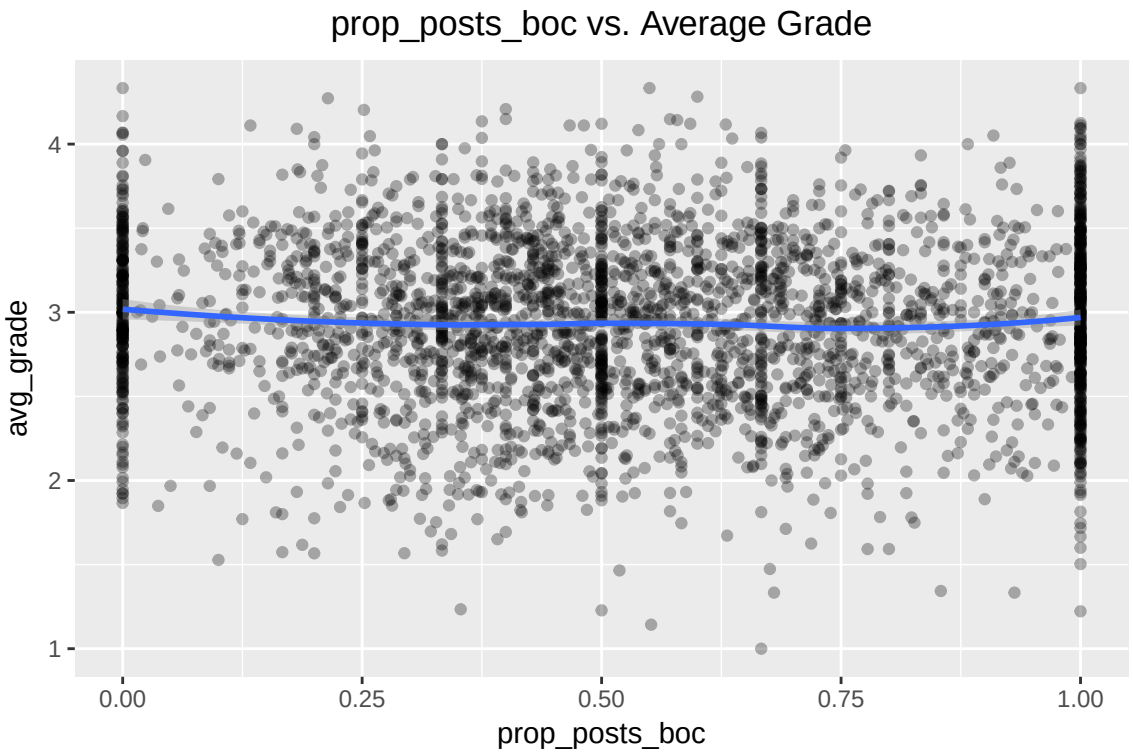
Average Grade

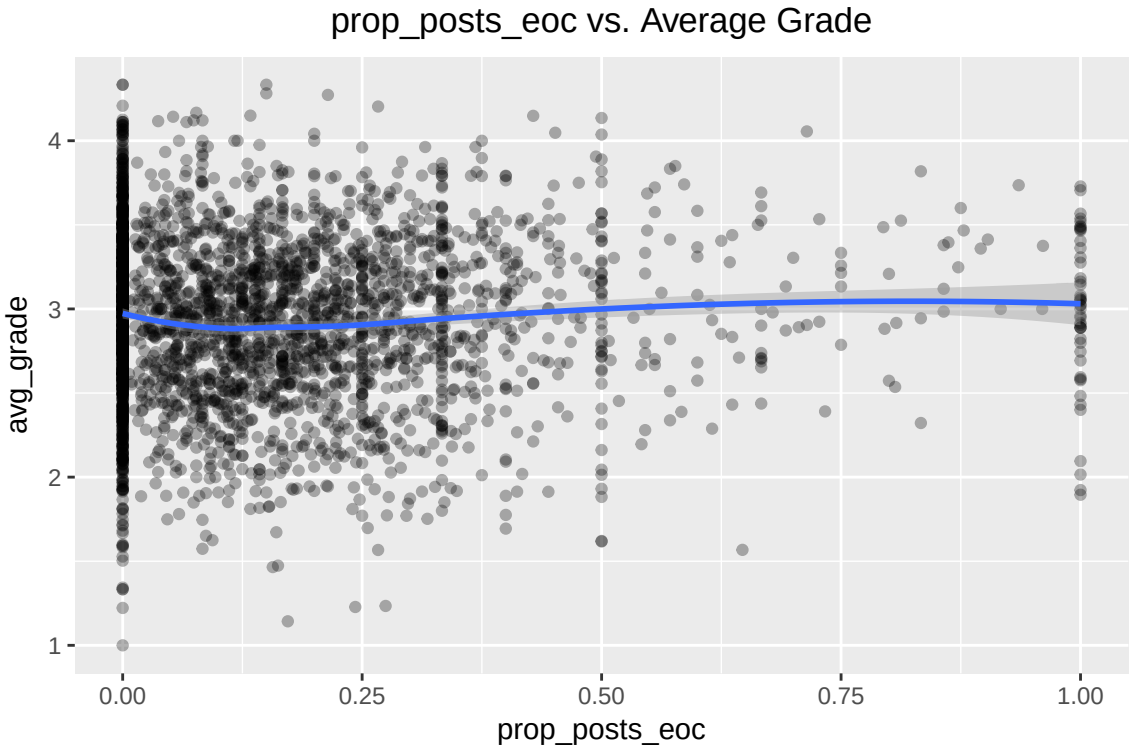





fac_post_consistency vs. Average Grade
avg_grade
4
3
2
1
0
1
2
3
fac_post_consistency


posts_per_fac vs. Average Grade
avg_grade
4
3
2
1
0
200
400
posts_per_fac

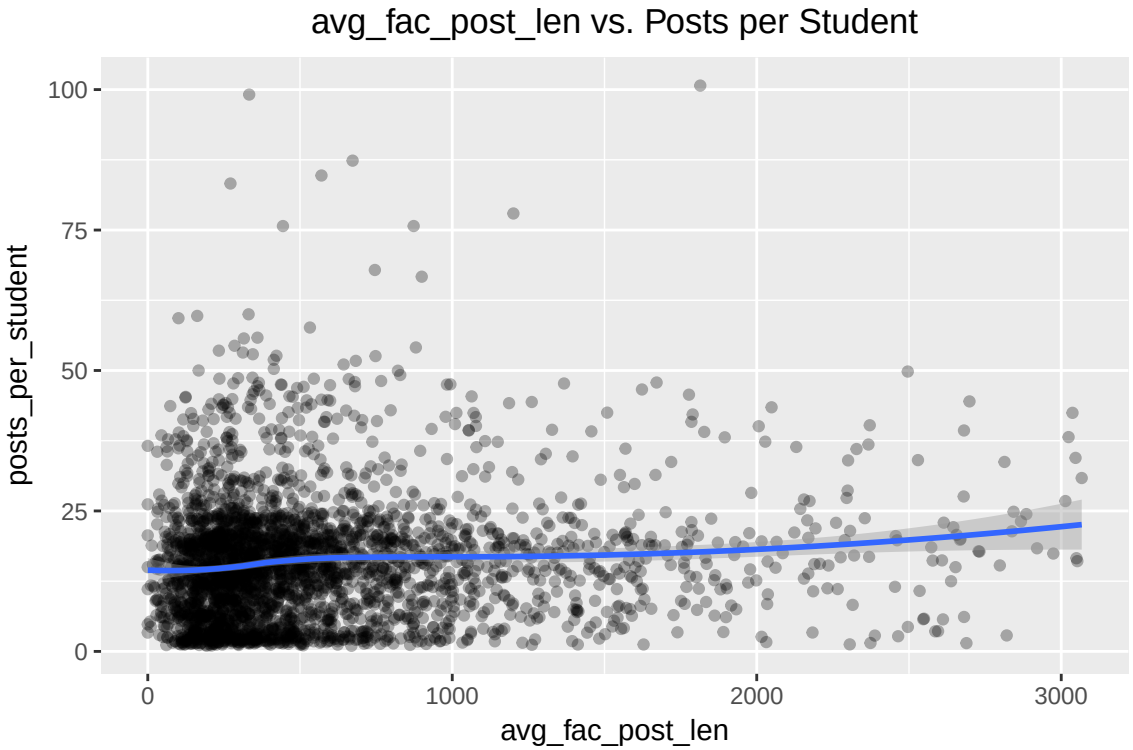


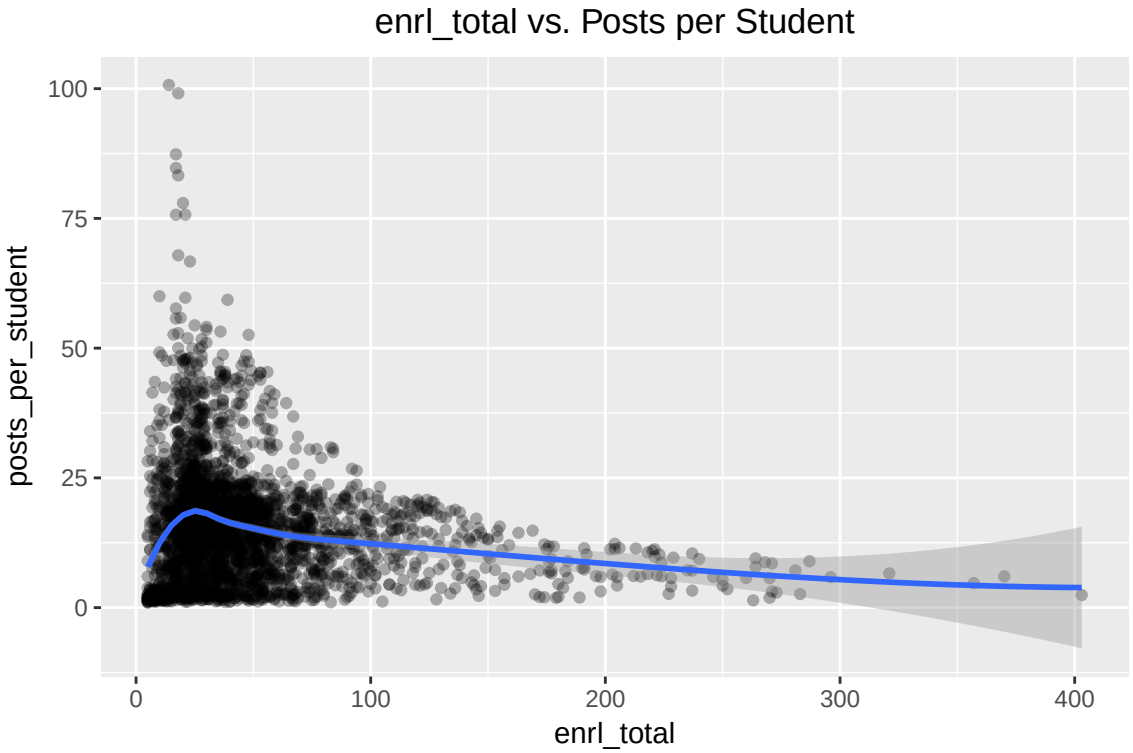


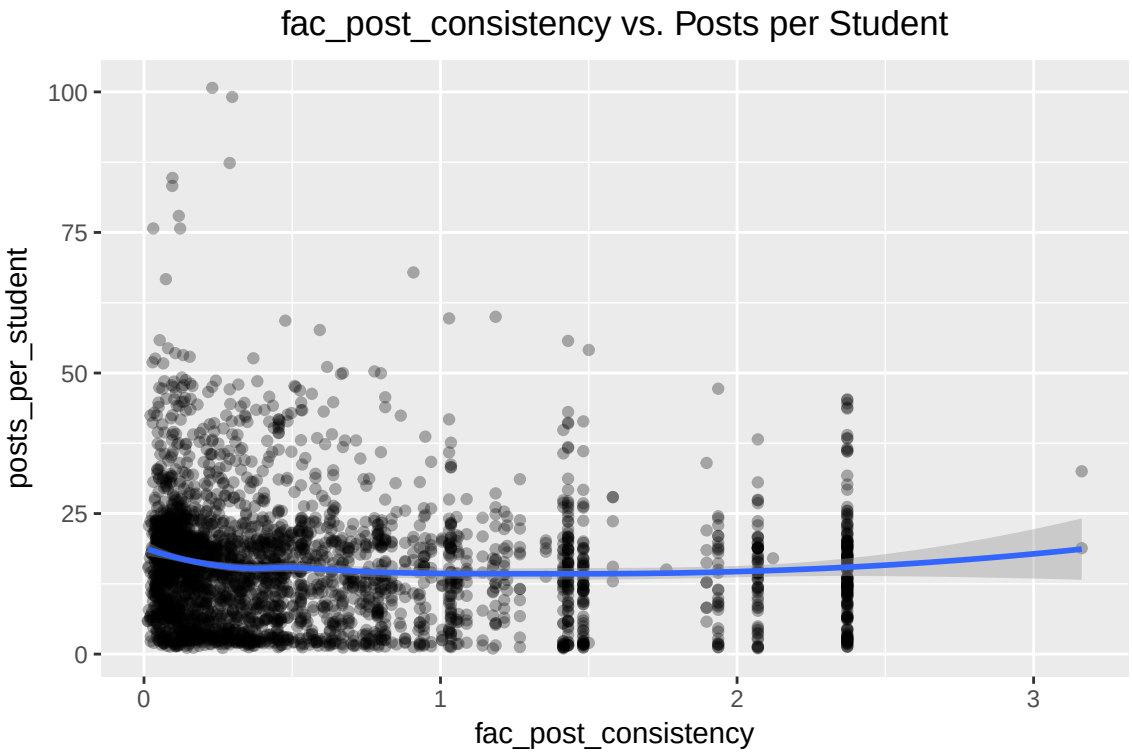
Discussion

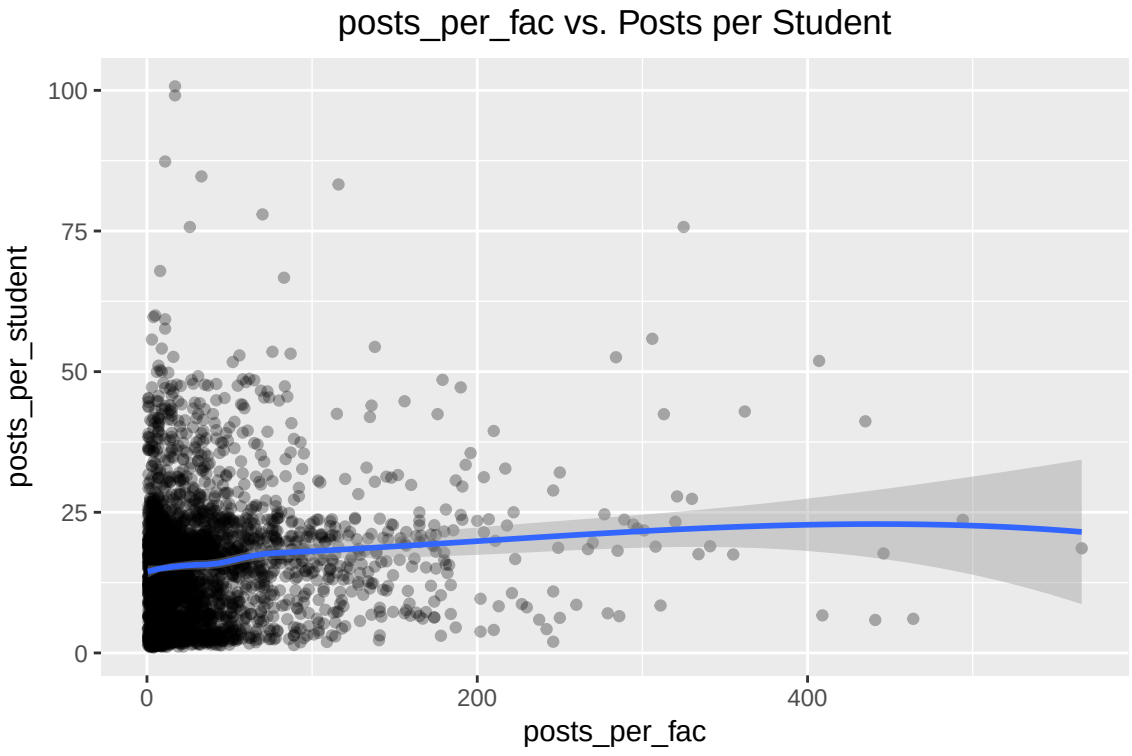
- Like the previous outcome, none of these plots suggest that the variables we thought to be important will be relevant in predicting grade.

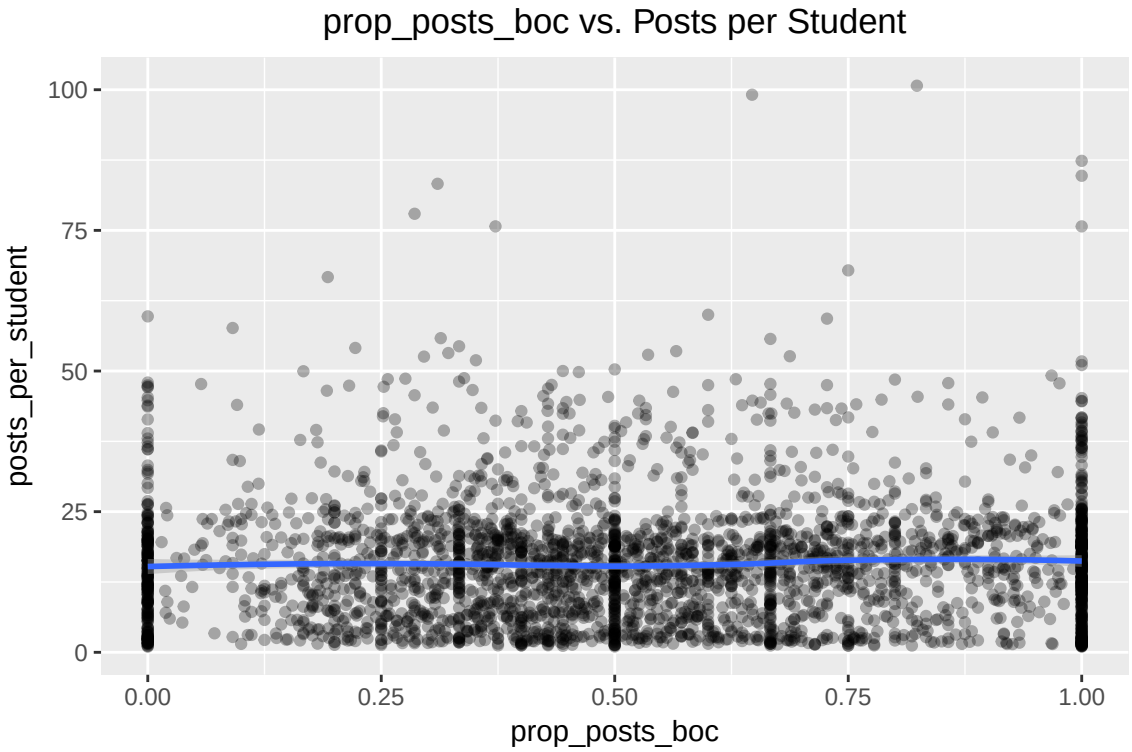
Posts per Student

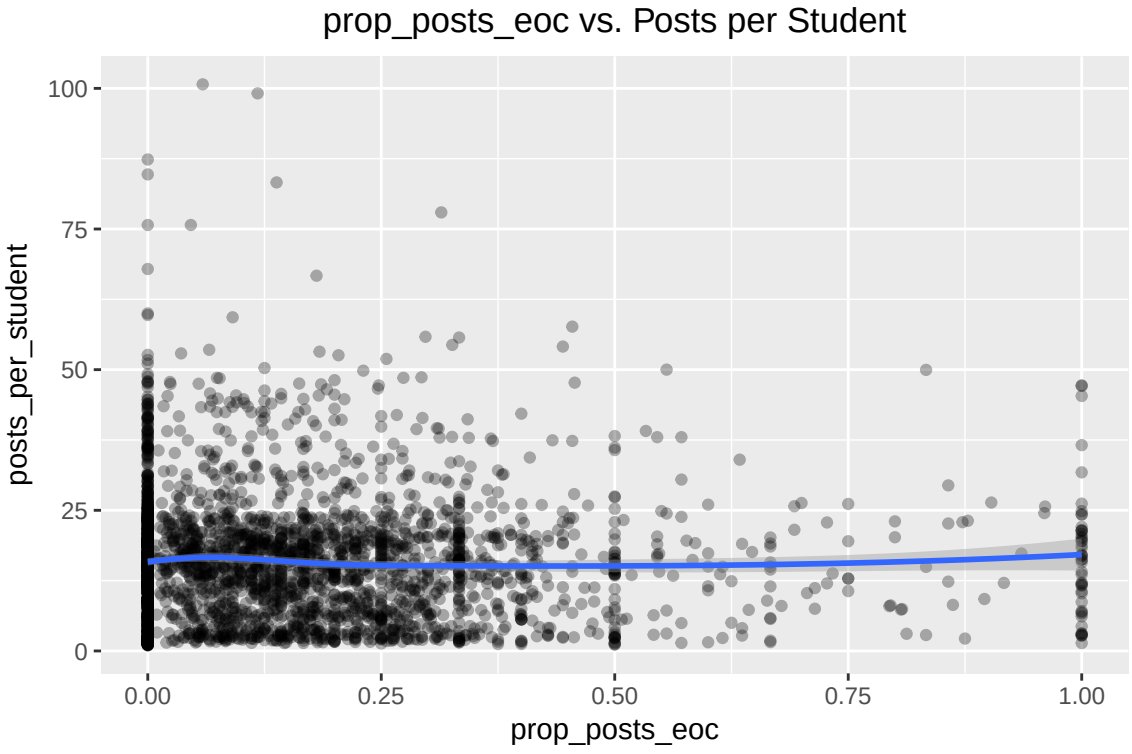





fac_post_consistency vs. Posts per Student
posts_per_student
100
75
50
25
0
0
1
2
3
fac_post_consistency


posts_per_fac vs. Posts per Student
posts_per_student
100
75
50
25
0
0
200
400
posts_per_fac



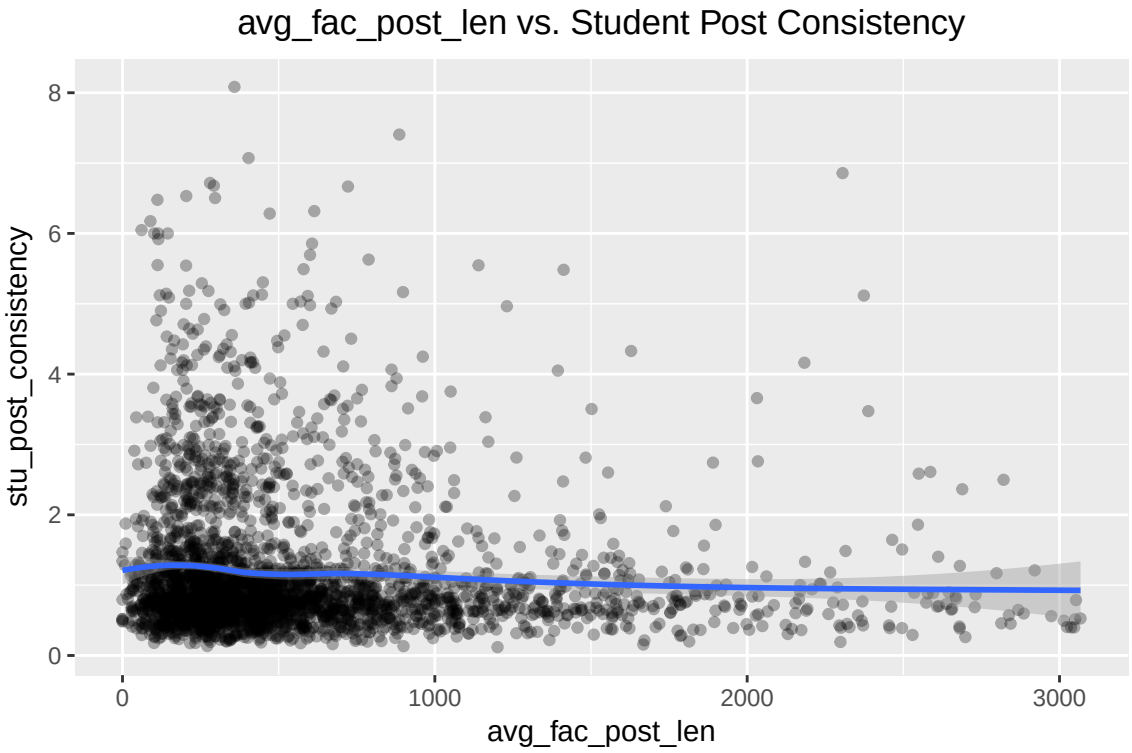


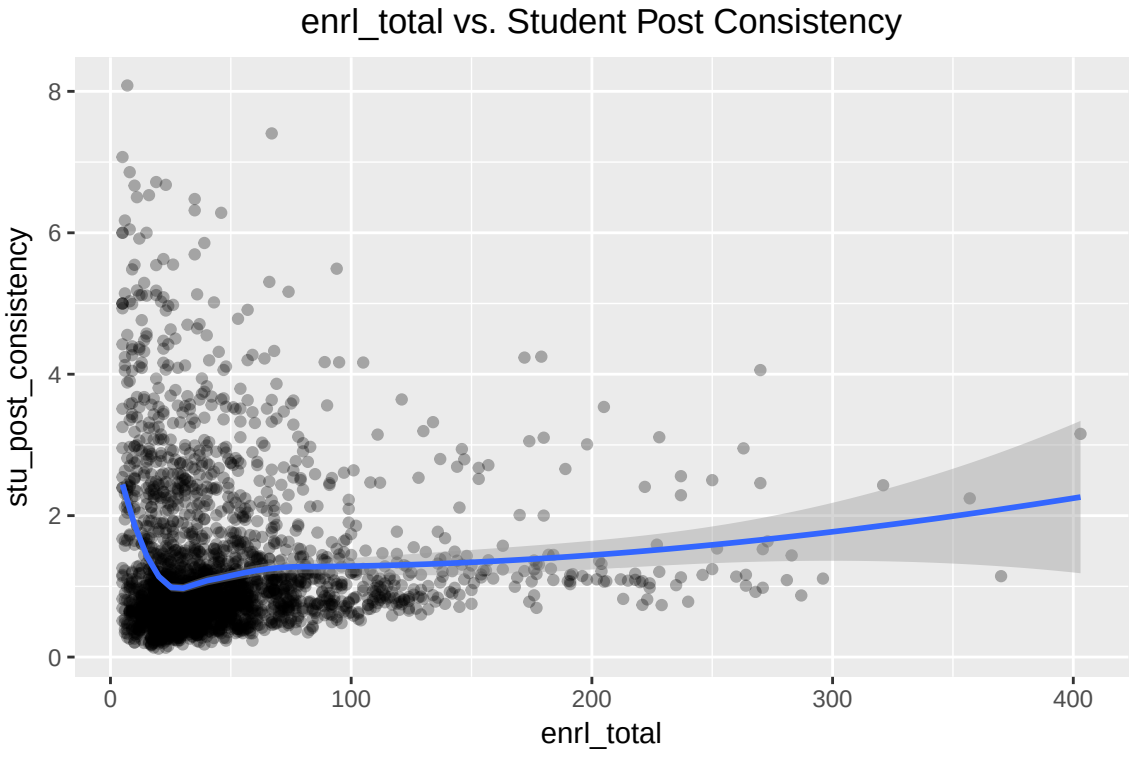
Discussion

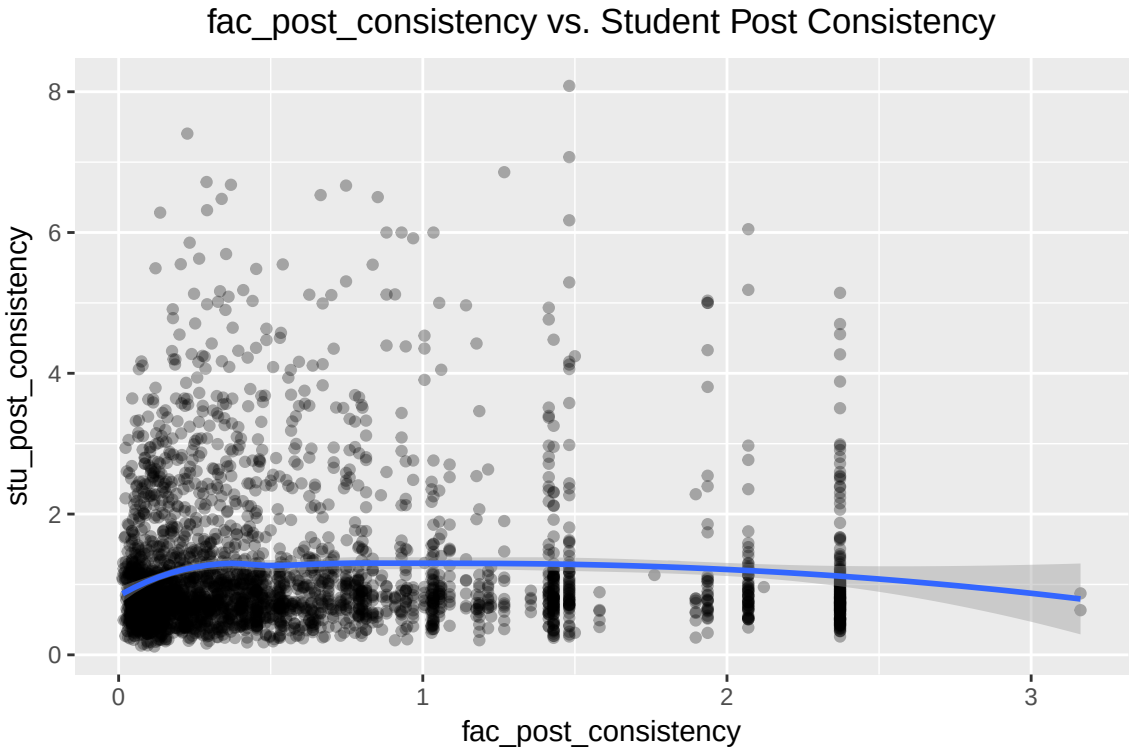
- Disappointingly, none of the variables here seem to be strongly related to the number of posts per student. This seems to be completely opposite to our expectations, but at the very least we now have

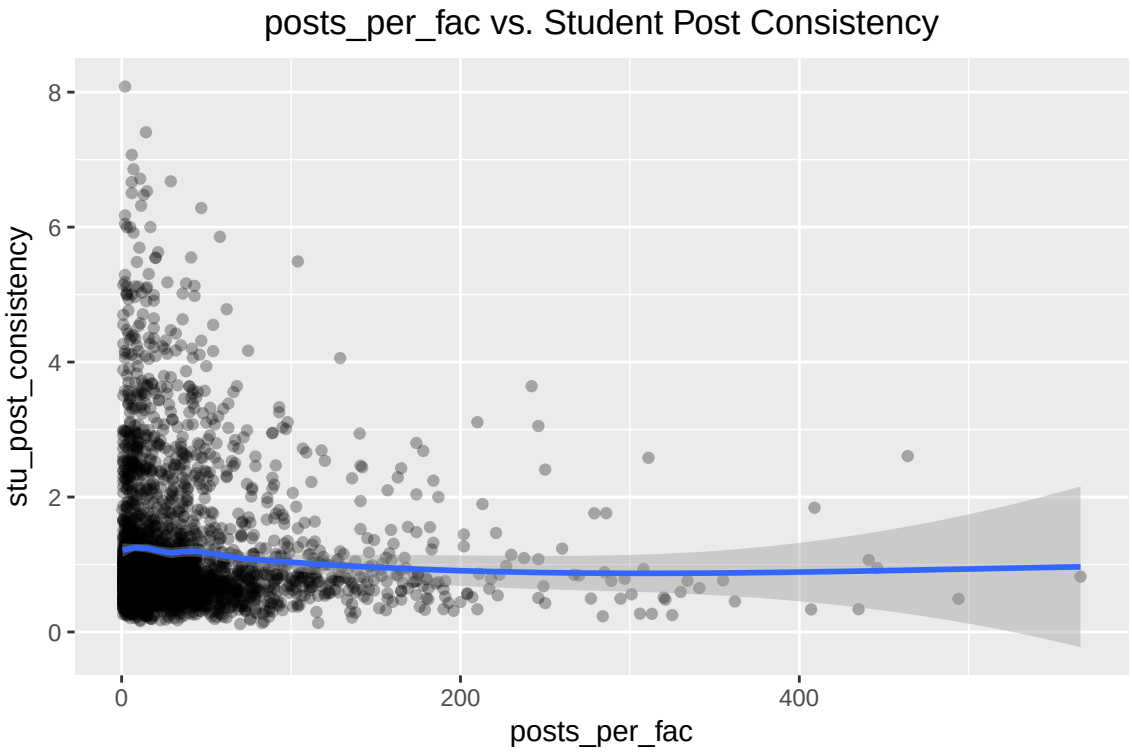
some reason to start looking for other possible explanations

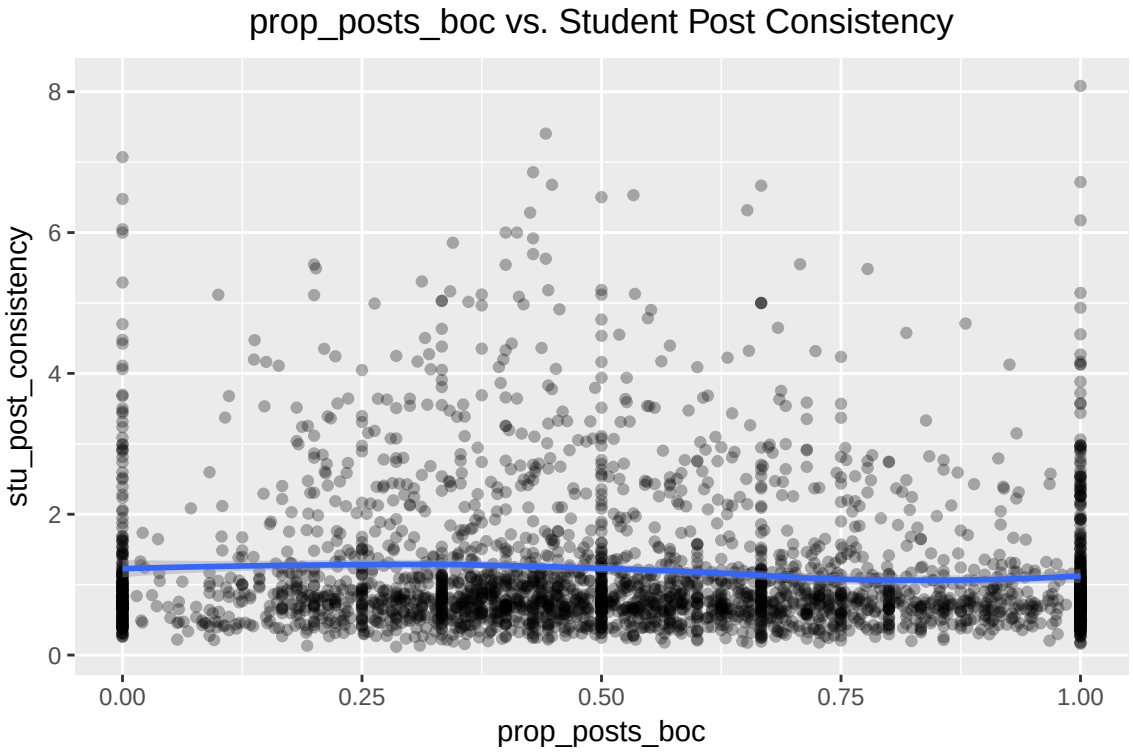
Student Post Consistency

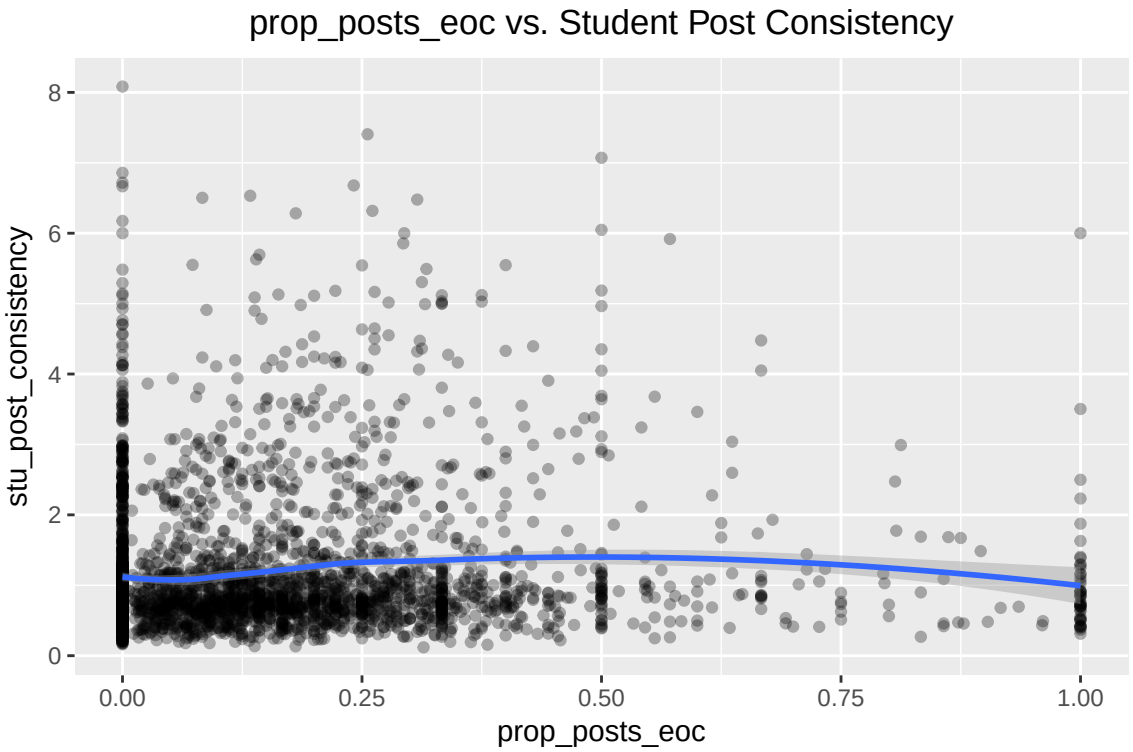





fac_post_consistency vs. Student Post Consistency
stu_post_consistency
fac_post_consistency
0
2
4
6
8
0
1
2
3


posts_per_fac vs. Student Post Consistency
stu_post_consistency
posts_per_fac
0
2
4
6
8
0
200
400





Discussion

- It appears that for very low enrollment courses student post consistency is higher than average. Other than this relatively small finding, the rest of the variables do not appear to show any significant relation

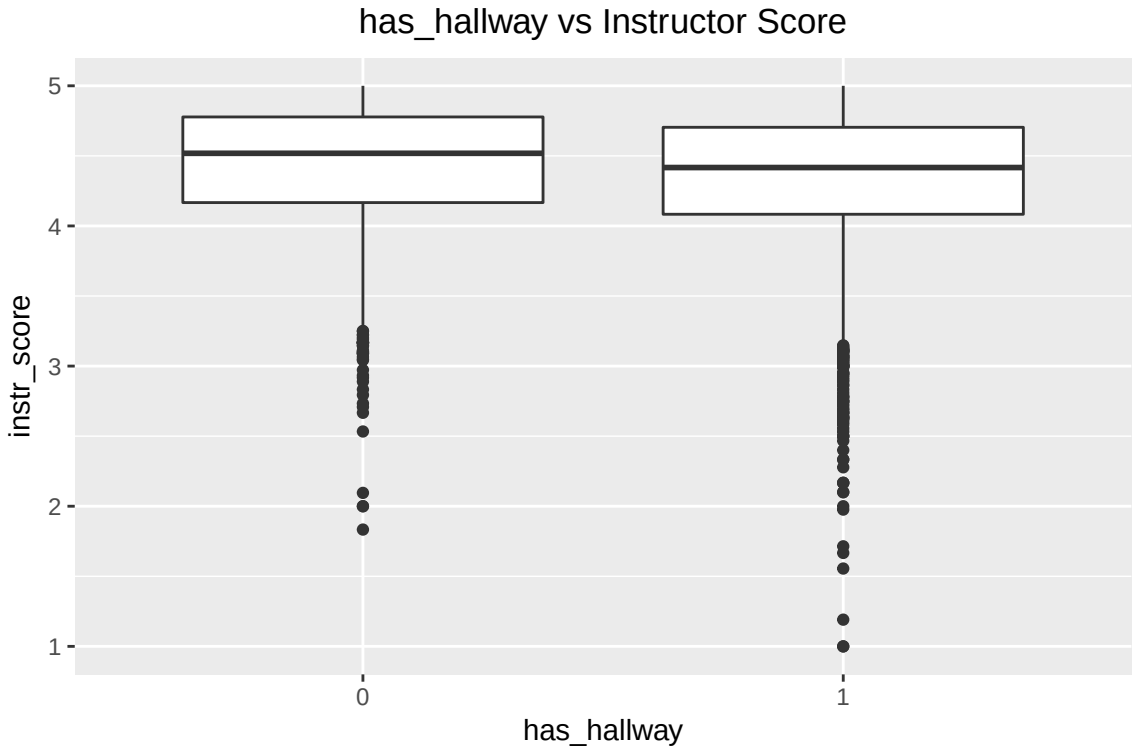
with student post consistency.

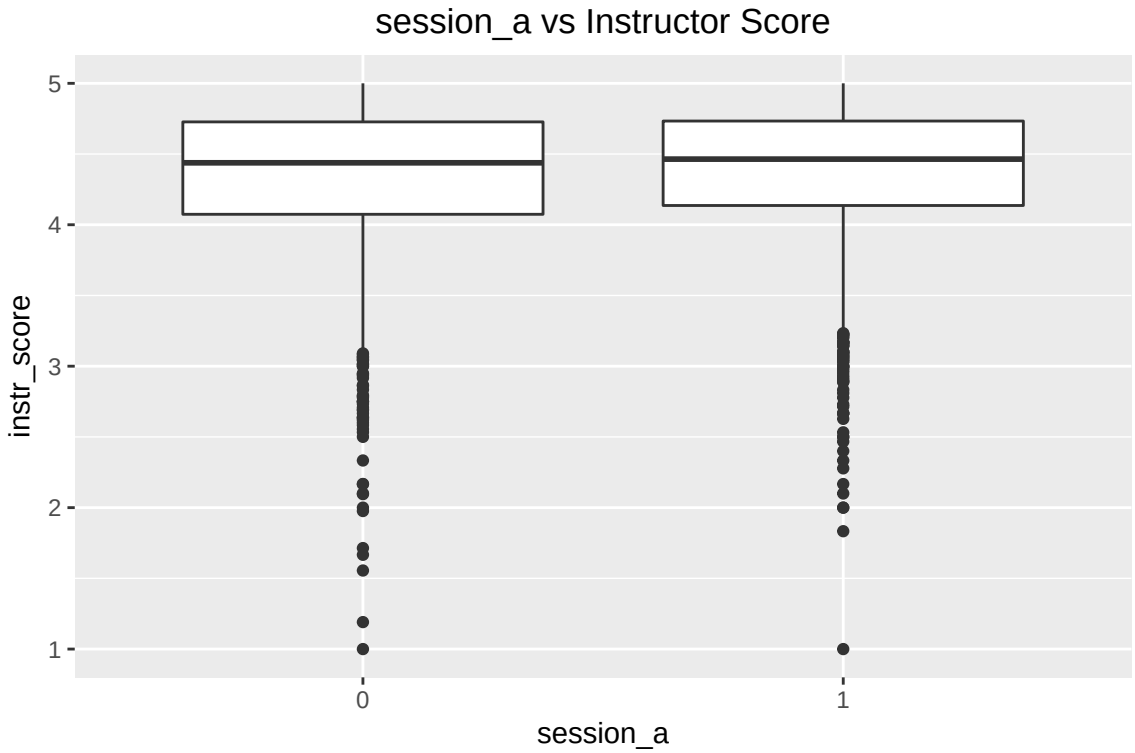
Binary Variable Exploration

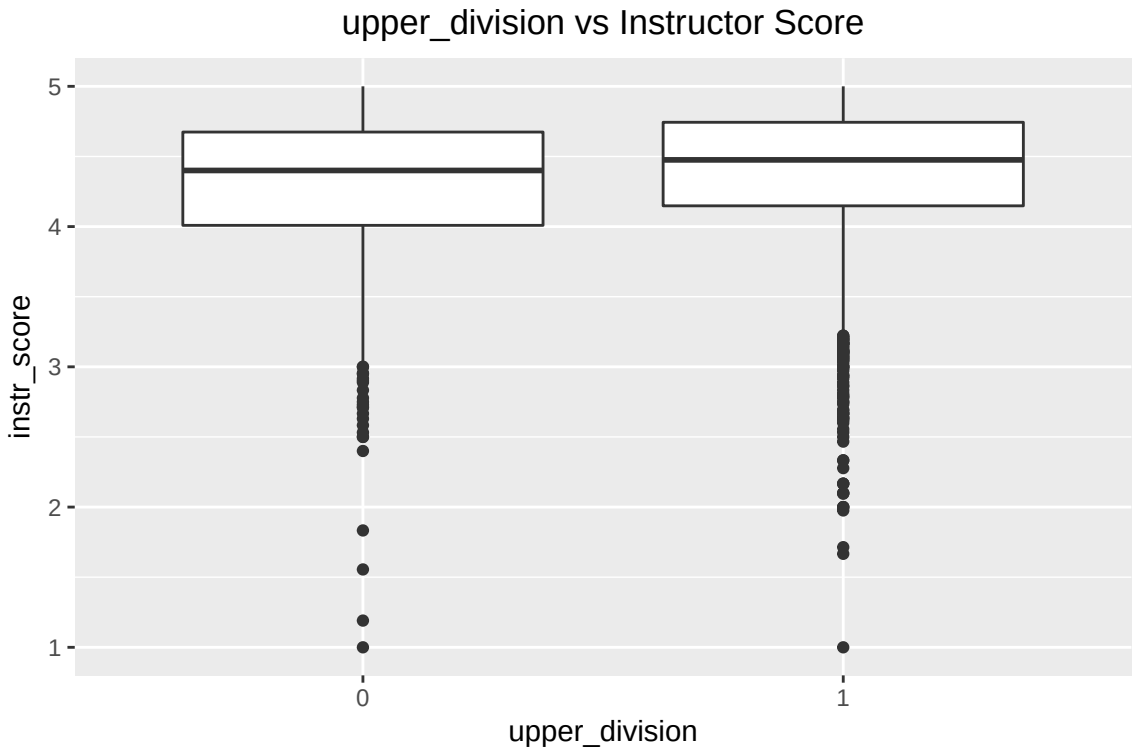
Up until this point we've ignored the three binary variables in the data - `has_hallway`, `session_a`, and `upper_division`. We defined the first earlier, and the last should be self-explanatory, but the second indicator tells us if the course-section occurred during the first 8-week session of the semester or the second.

Now that we have practically abandoned all hope with the continuous variables, we need to see if the same pattern is going on with the binary variables.

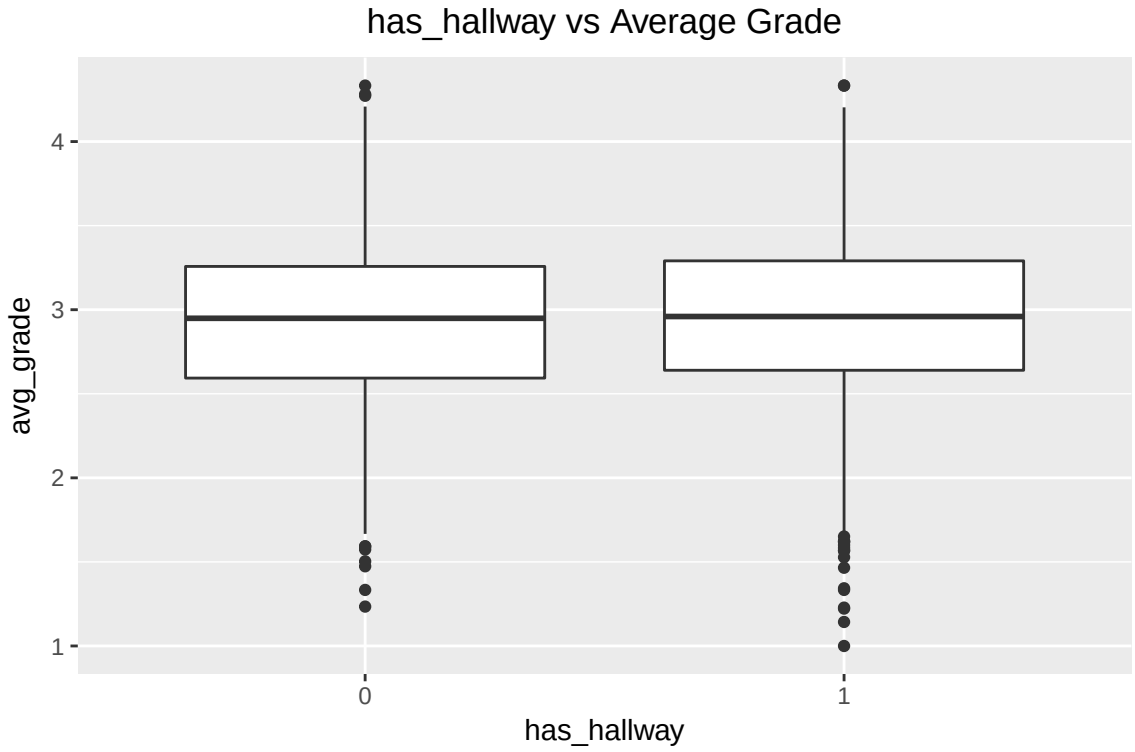
Instructor Score

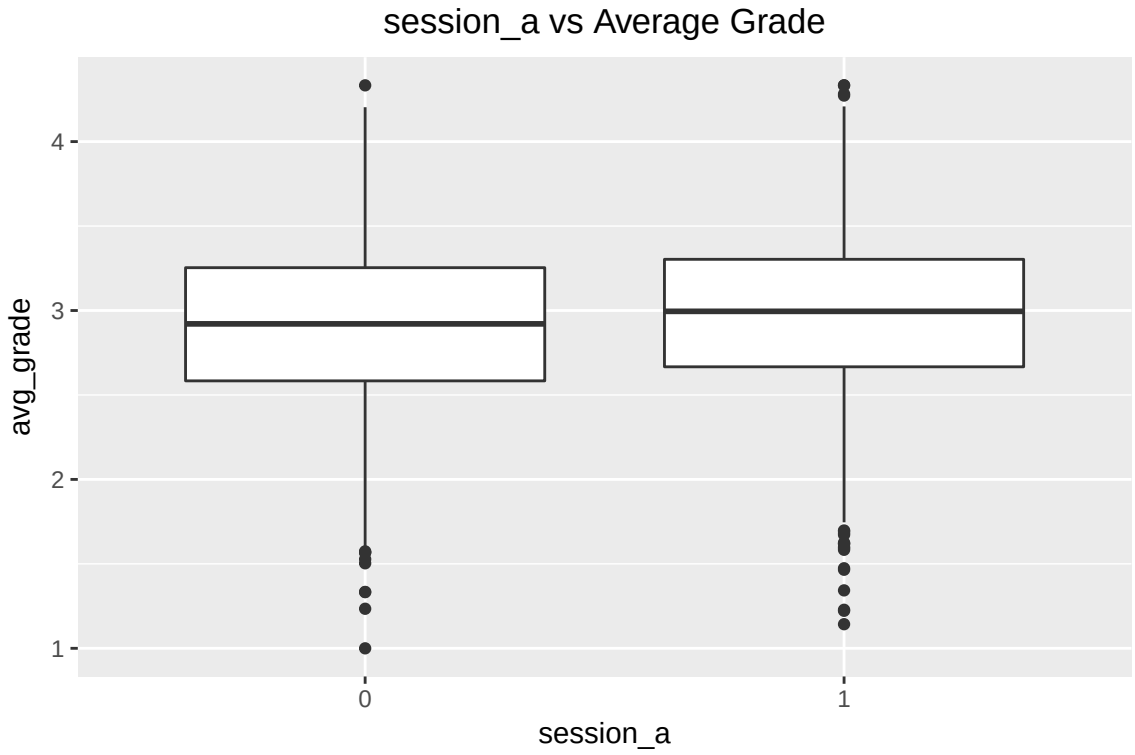


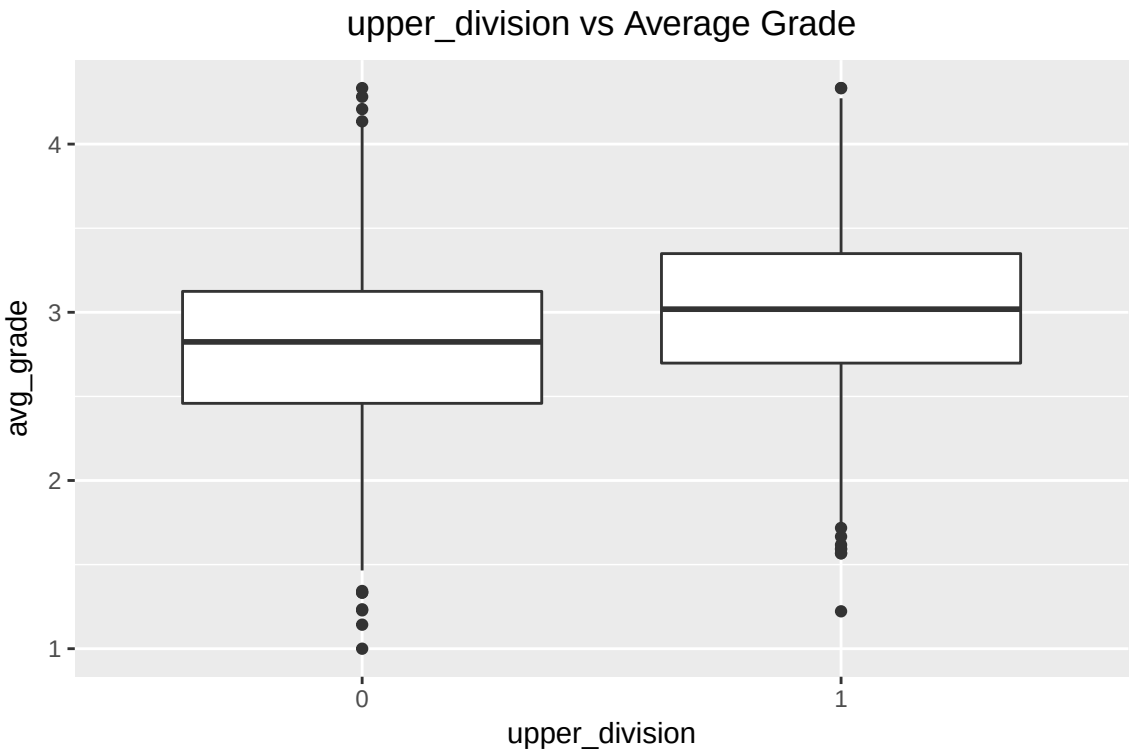

session_a vs Instructor Score
instr_score
5
4
3
2
1
0
1
session_a


upper_division vs Instructor Score
instr_score
5
4
3
2
1
0
1
upper_division

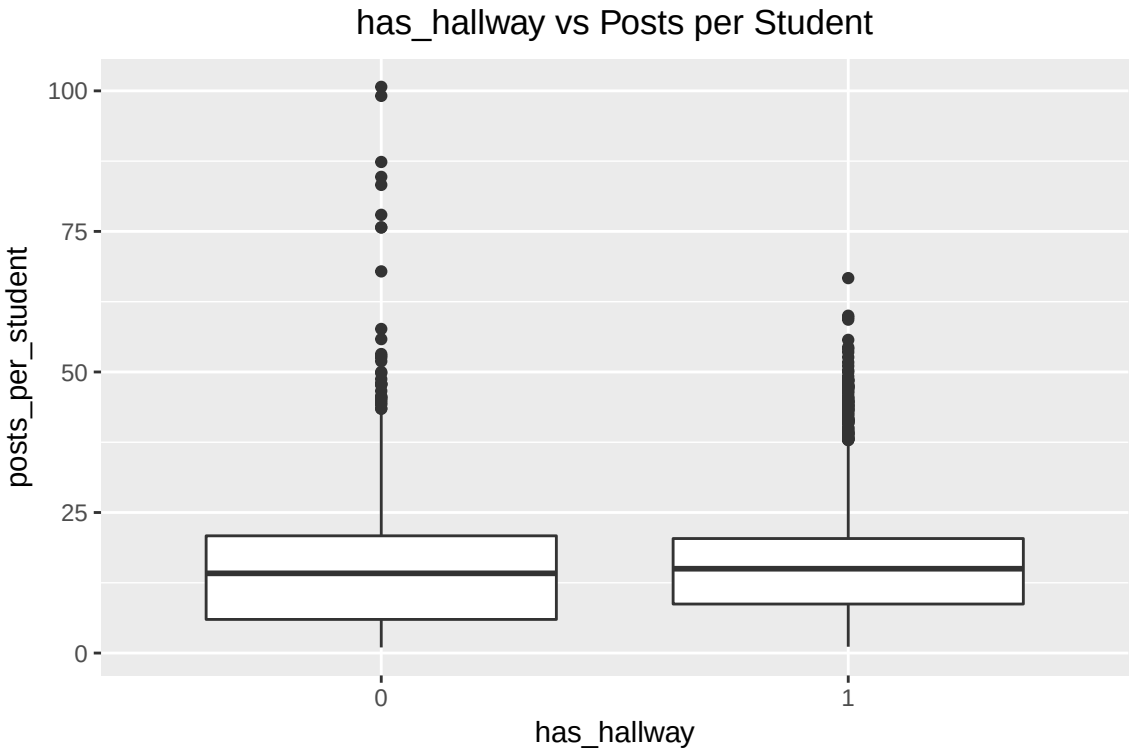
Average Grade

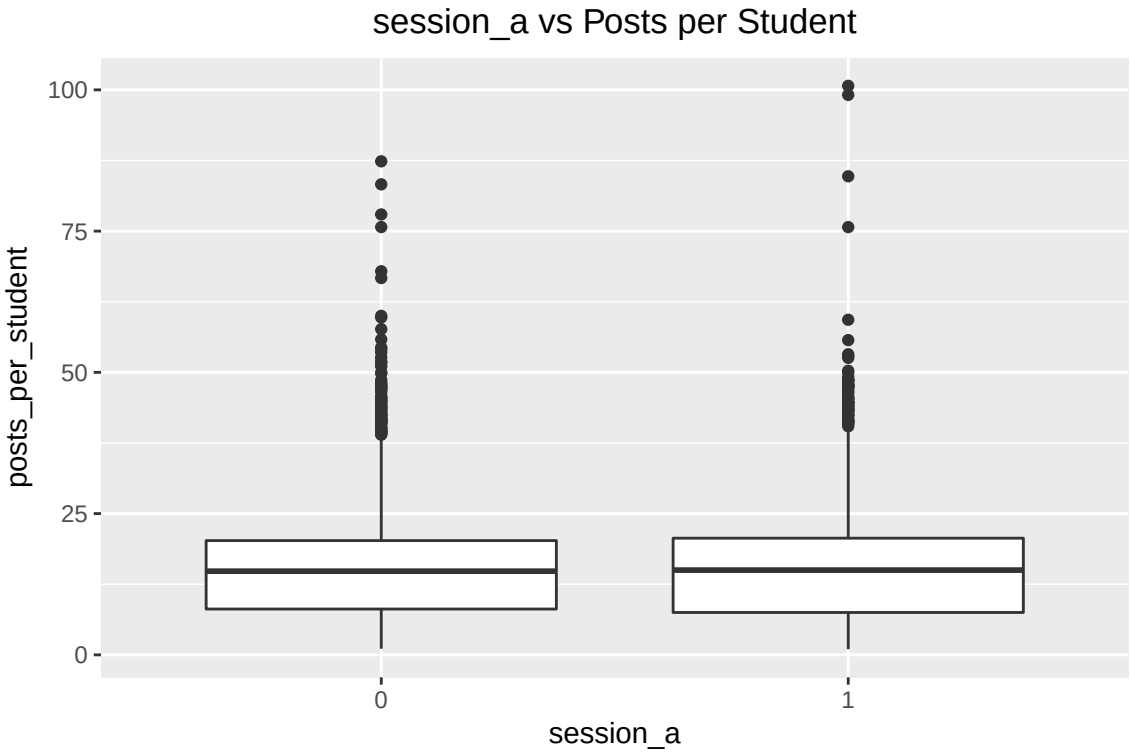


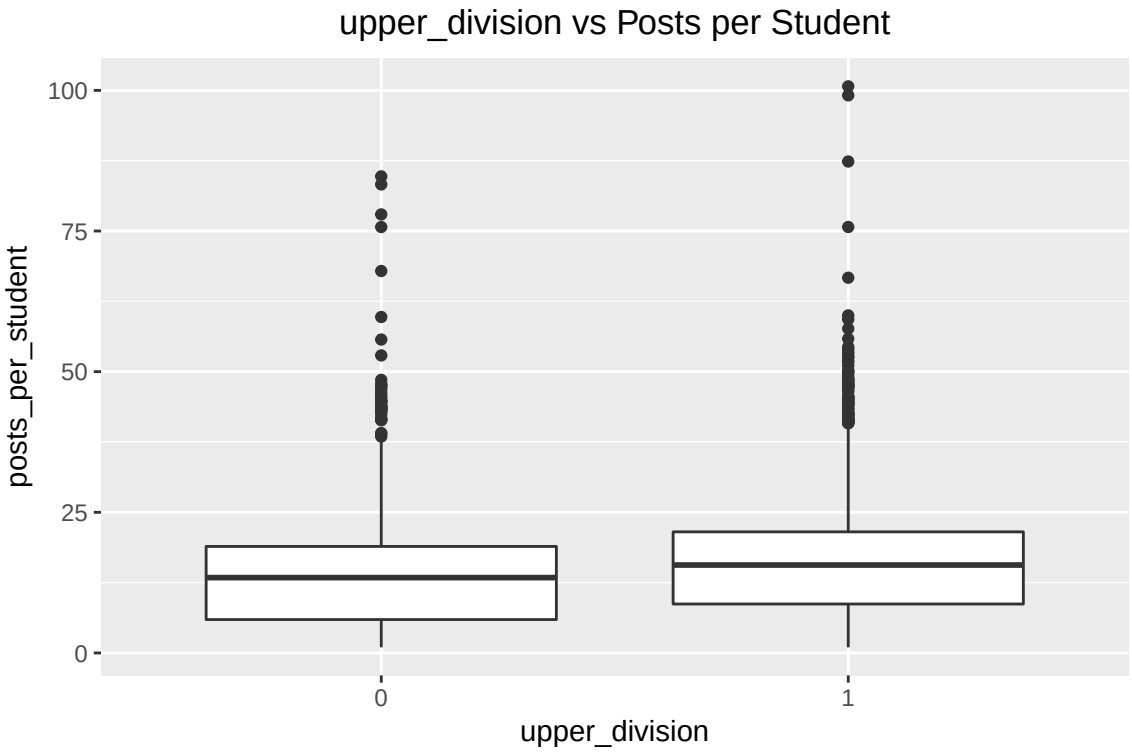




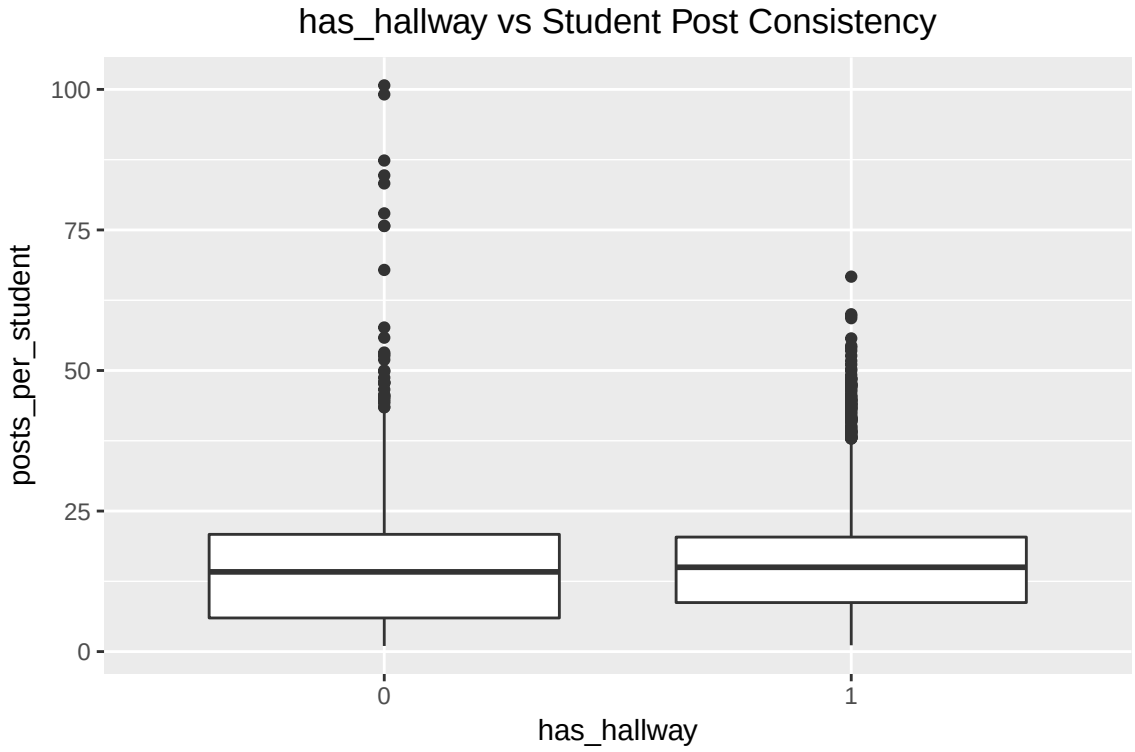
Posts per Student

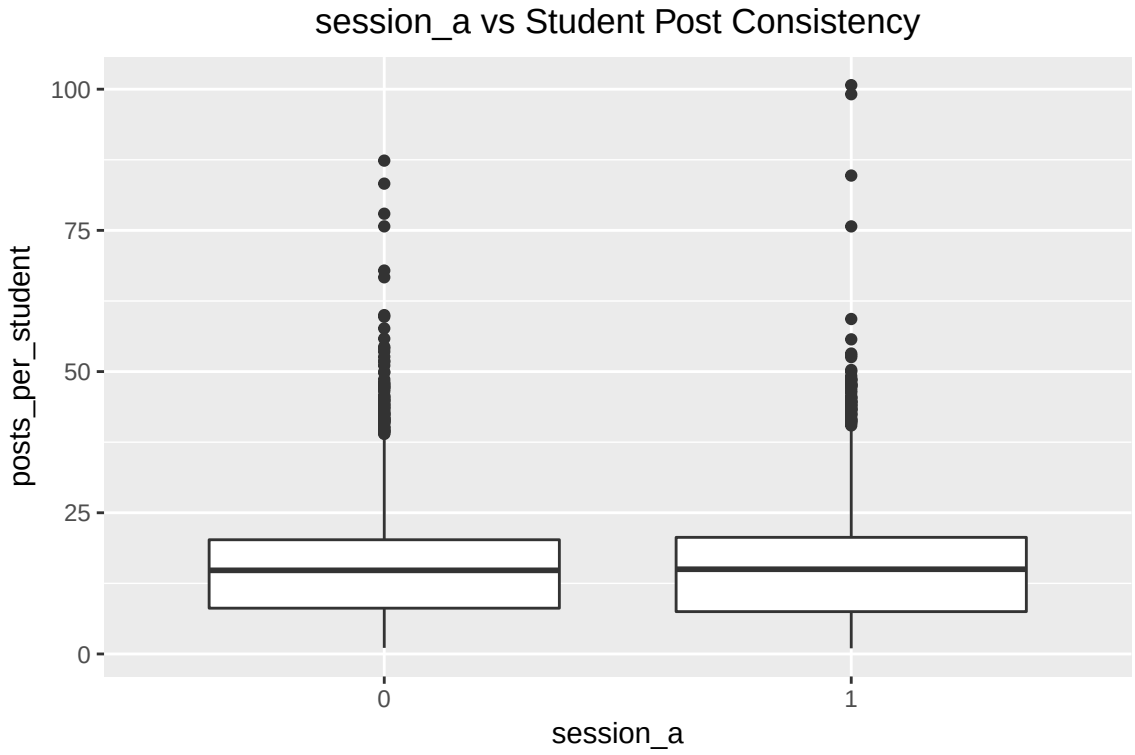


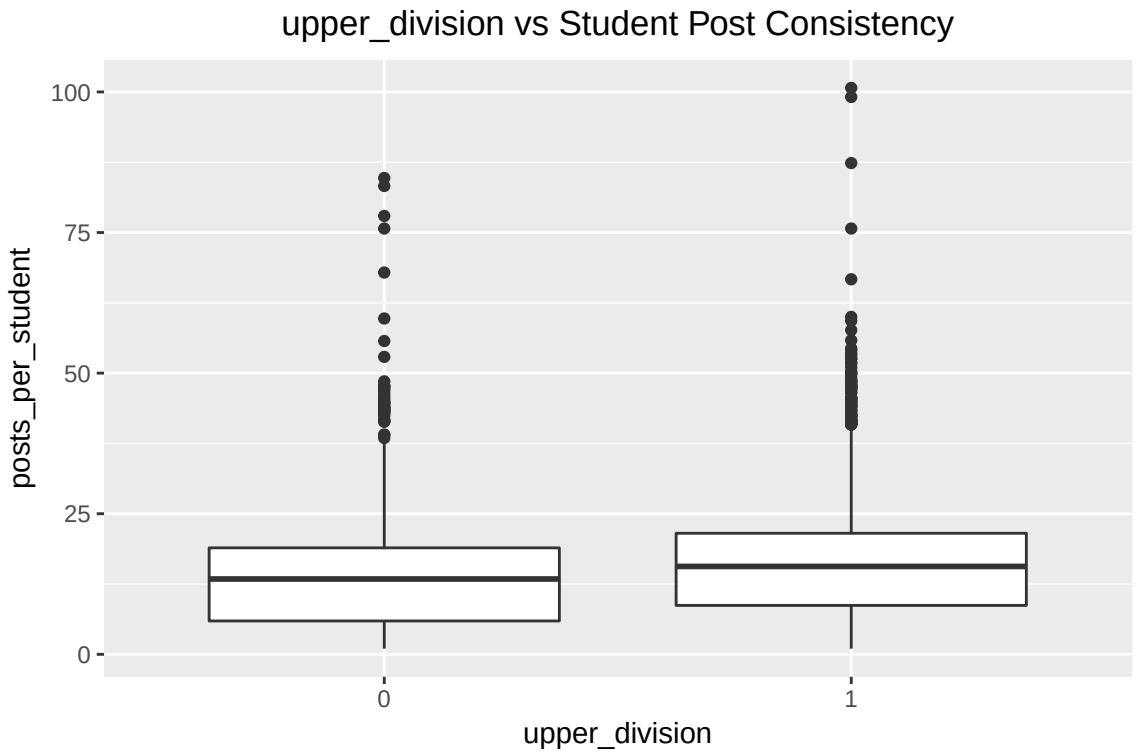

session_a vs Posts per Student
posts_per_student
100
75
50
25
0
0
1
session_a


upper_division vs Posts per Student
posts_per_student
100
75
50
25
0
0
1
upper_division

Student Post Consistency







Discussion

Based on these plots, it is fair to say that we have evidence against using any of these variables to explain variation in the outcomes. None of the box plots revealed noticeable differences in outcomes between the levels of the binary indicator, so it is unlikely that they will be useful in a regression.

Conclusion

These are rather disappointing results, but this is not necessarily the end of the line. We can still move forward with some modeling to see if any sort of signal can be extracted from these variables and if all else fails we can revisit the data extraction and the feature engineering.