Engagement EDA

William Morgan

24 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:

- Faculty Post Consistency (`fac_consistency`)
- Post Quantity (`ppf`)
- Post Timing (`prop_posts_boc`)
  - The proportion of posts in the course that happen in the first two weeks
- Post Length (`avg_fac_len`)

Outcomes of interest:

- Student Engagement:
  - Posts per student (`pps`)
  - Student post consistency (`stu_consistency`)
- Grade Outcomes:
  - Average grade received (`avg_grade`)
  - Withdrawal rates (`wdrw_rate`)
- Instructor evaluation scores (`instr_score`)
  - there are three questions on course evaluation surveys pertaining to faculty presence and engagement which we average into a single score

General Outline

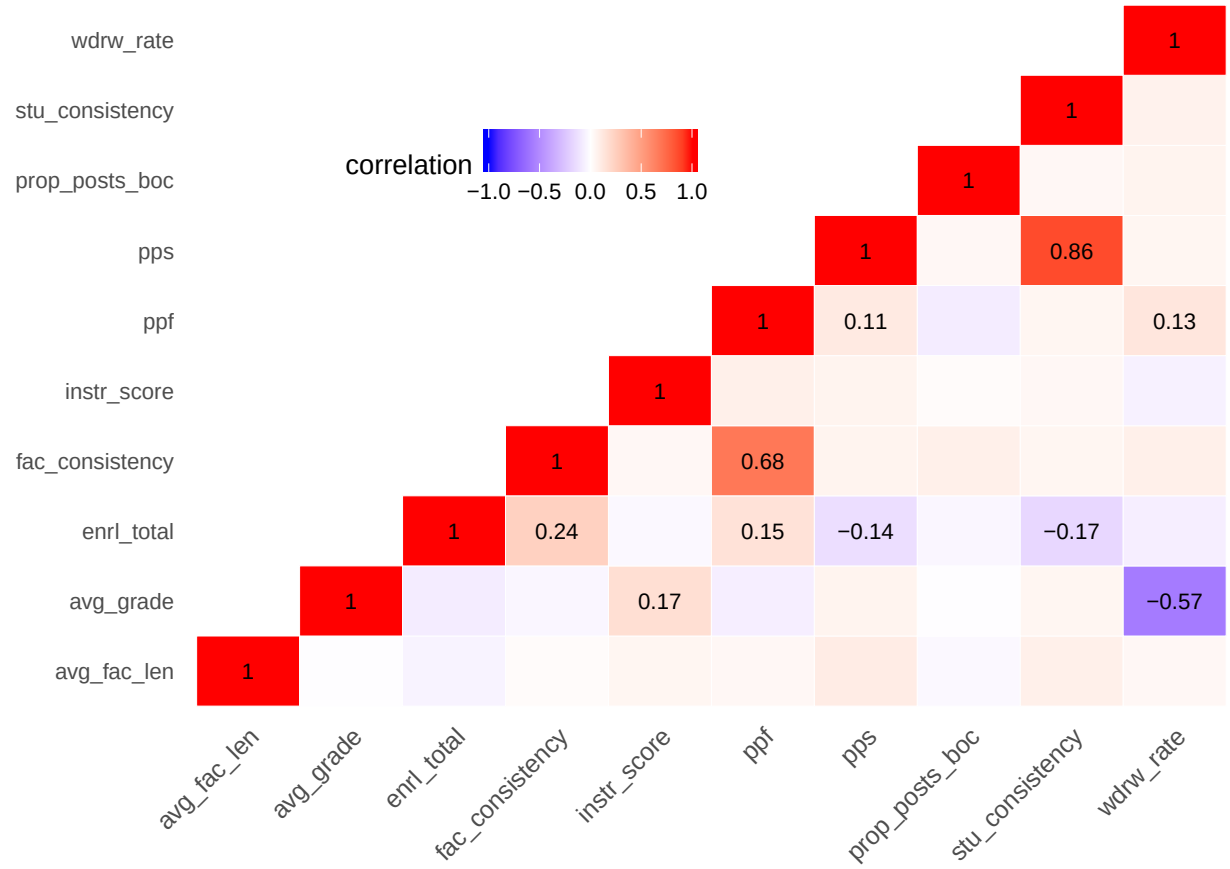
The goal is to understand which features correlate most with the several outcomes. As a first pass we'll create a correlation heatmap relating the continuous features to each of the outcomes. This will give us some basic insight on any apparent linear relationships. Next we'll move on to plotting the 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':    2852 obs. of  13 variables:
##  $ avg_fac_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" "2015
##  $ enrl_total     : num  28 19 7 119 89 72 36 50 137 21 ...
##  $ fac_consistency: num  7.556 7.6 0.944 8.278 30.322 ...
##  $ 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 ...
##  $ ppf            : num  14 13 4 34 9.33 ...
##  $ pps            : num  2.29 22.58 1.43 7.84 19.75 ...
##  $ prop_posts_boc : num  0.393 0.769 1 0.353 0.821 ...
##  $ stu_consistency: num  0.244 1.962 0.425 0.564 1.16 ...
##  $ upper_division : Factor w/ 2 levels "0","1": 2 2 1 1 2 2 1 1 2 2 ...
##  $ wdrw_rate      : num  0 0.0526 0 0.0084 0.0562 ...
```

Correlation Heatmap

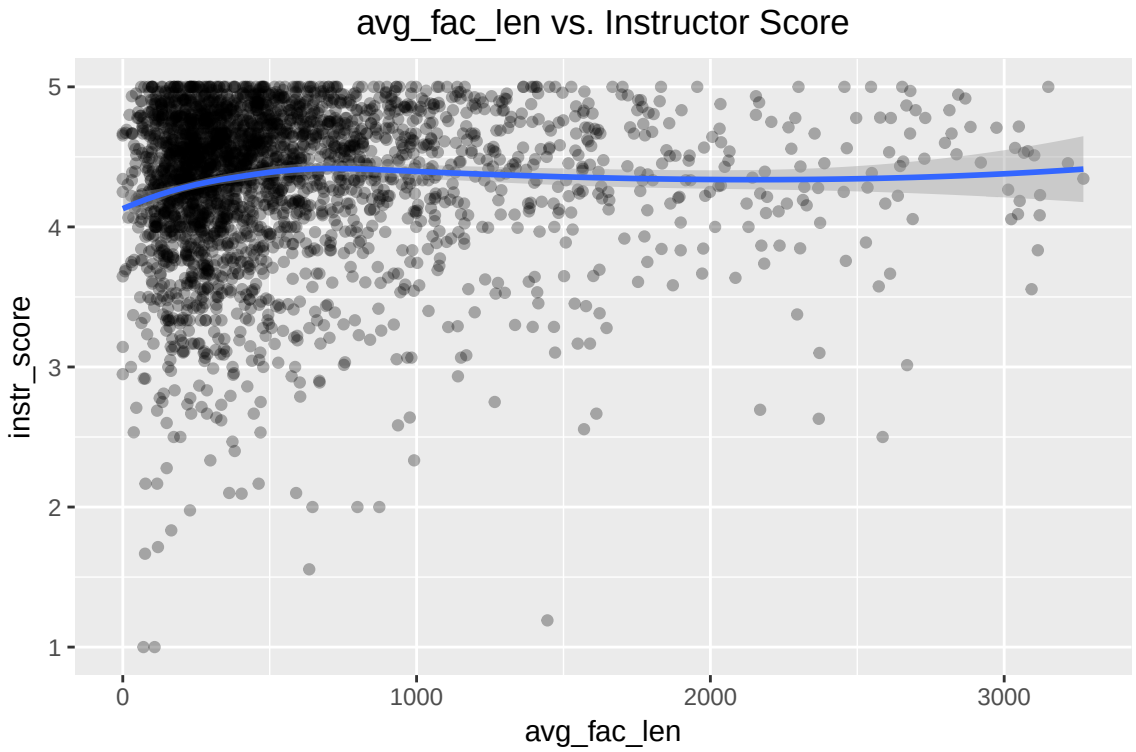


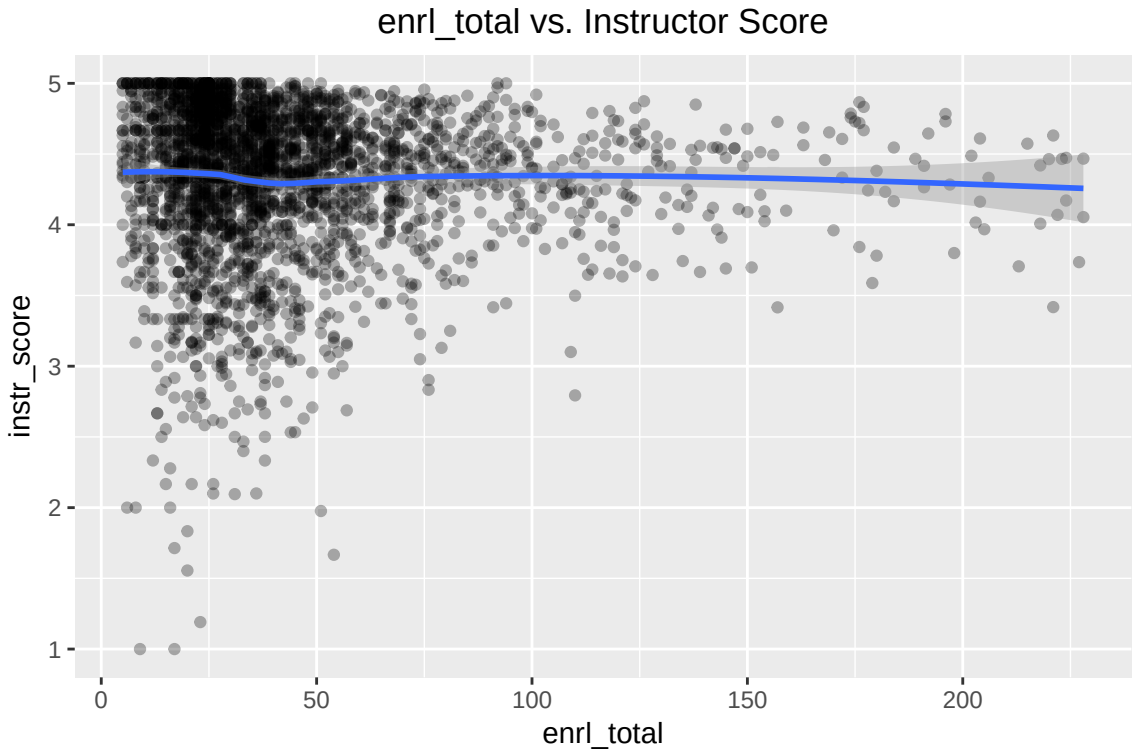
Discussion

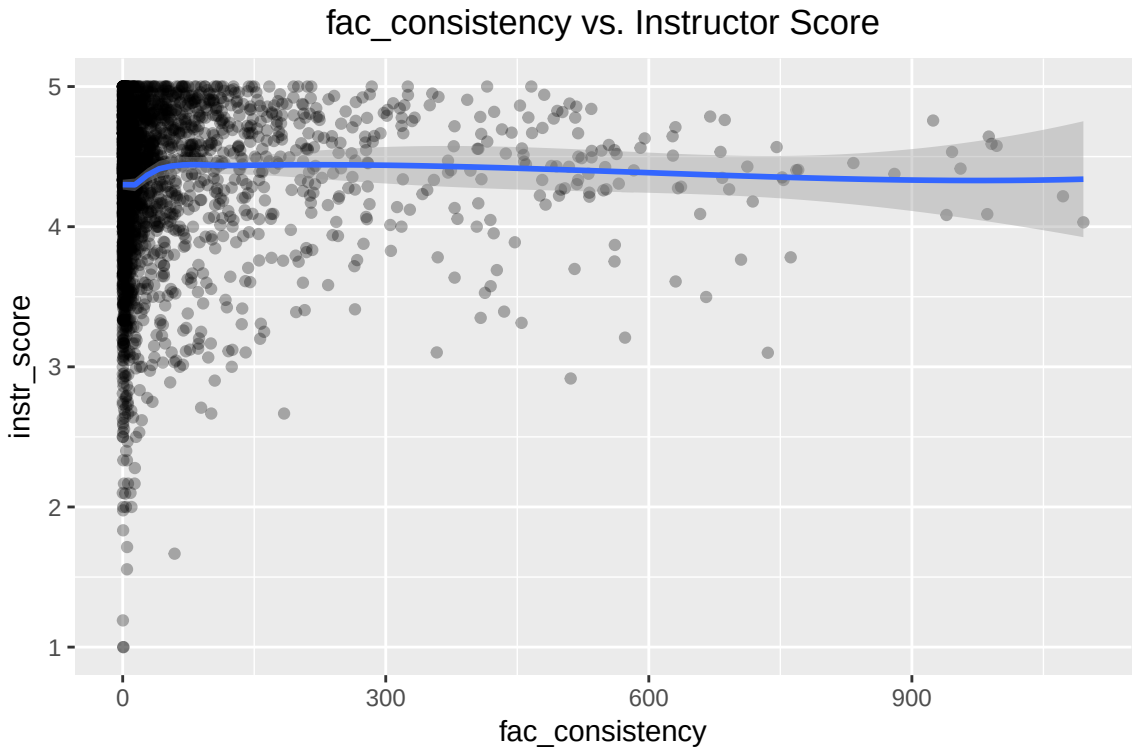
`instr_score` appears to be unrelated with every variable we were considering to include in a regression model. This doesn't mean we should throw it out as an outcome, but it does indicate that is has a very weak linear association with all other variables. `pps` has a slightly negative relationship with the number of students in the course and a positive one with `ppf`. `avg_grade` unfortunately appears to be uncorrelated with all forms of faculty engagement. We'll have to move on to doing more general plots to uncover something. Just like the others, `stu_consistency` does not seem to have any strong correlations with the explanatory variables besides the class size. Lastly, the withdrawal rate and posts per faculty variables have a slight positive relationship.

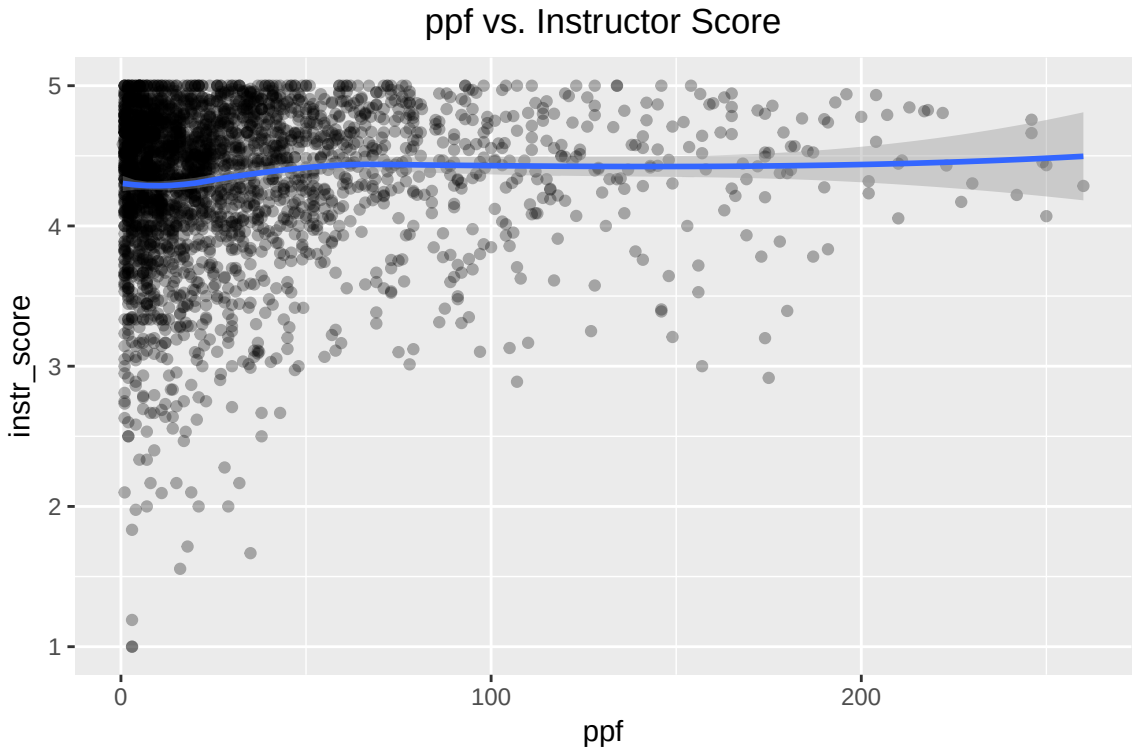
Two-way plots

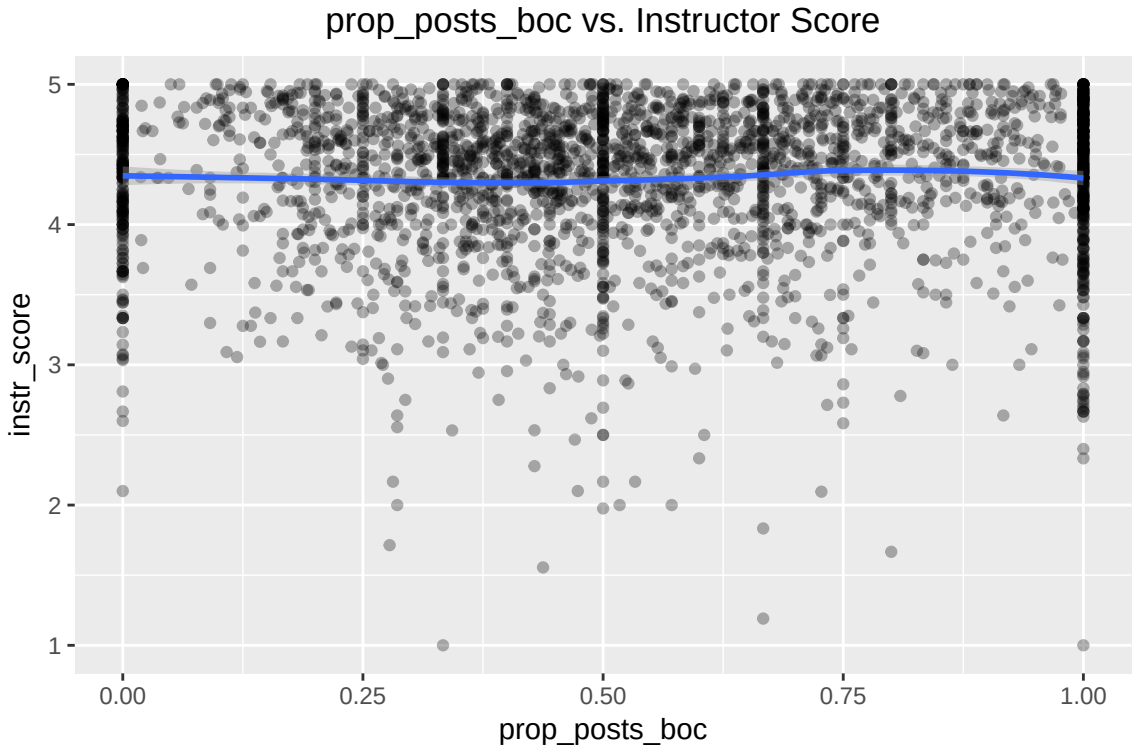
Instructor Score plots




enrl_total vs. Instructor Score
instr_score
enrl_total
5
4
3
2
1
0
50
100
150
200


fac_consistency vs. Instructor Score
instr_score
fac_consistency
5
4
3
2
1
0
300
600
900

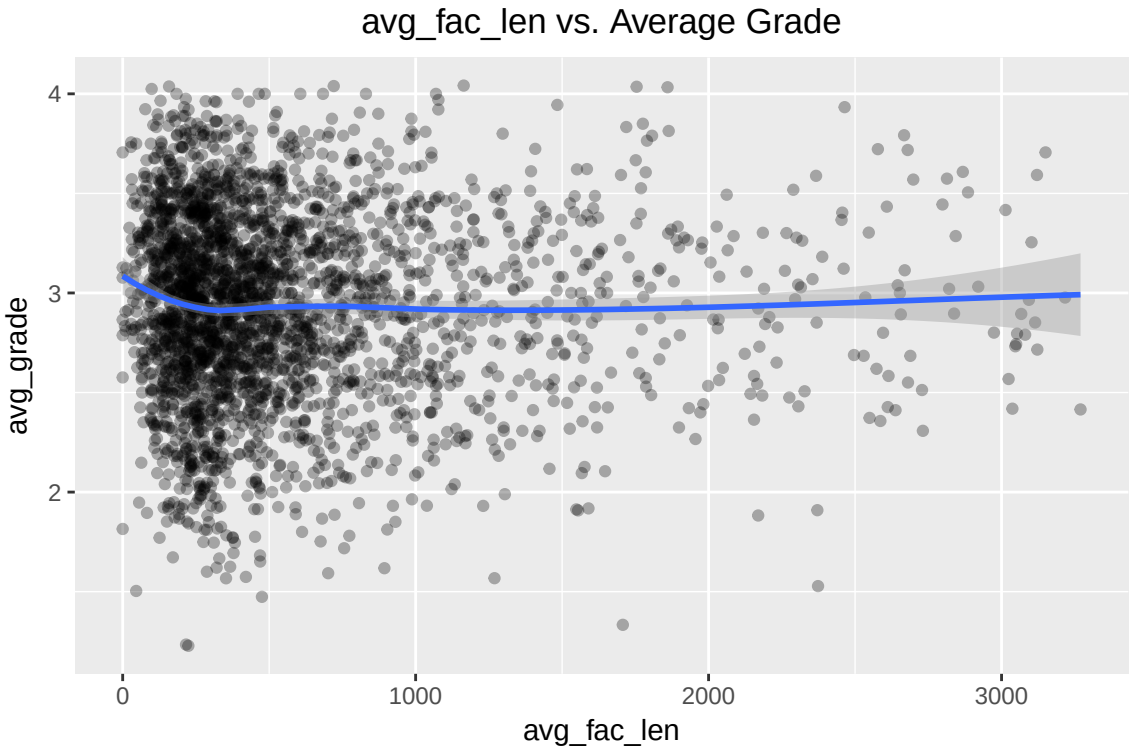


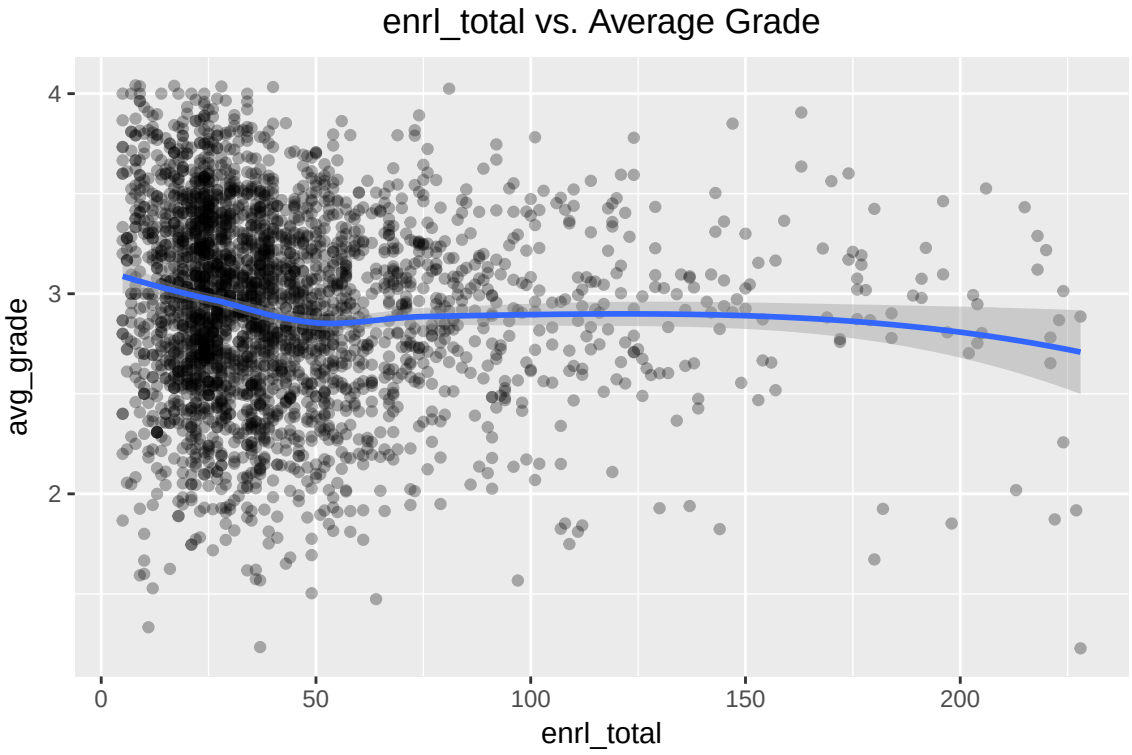


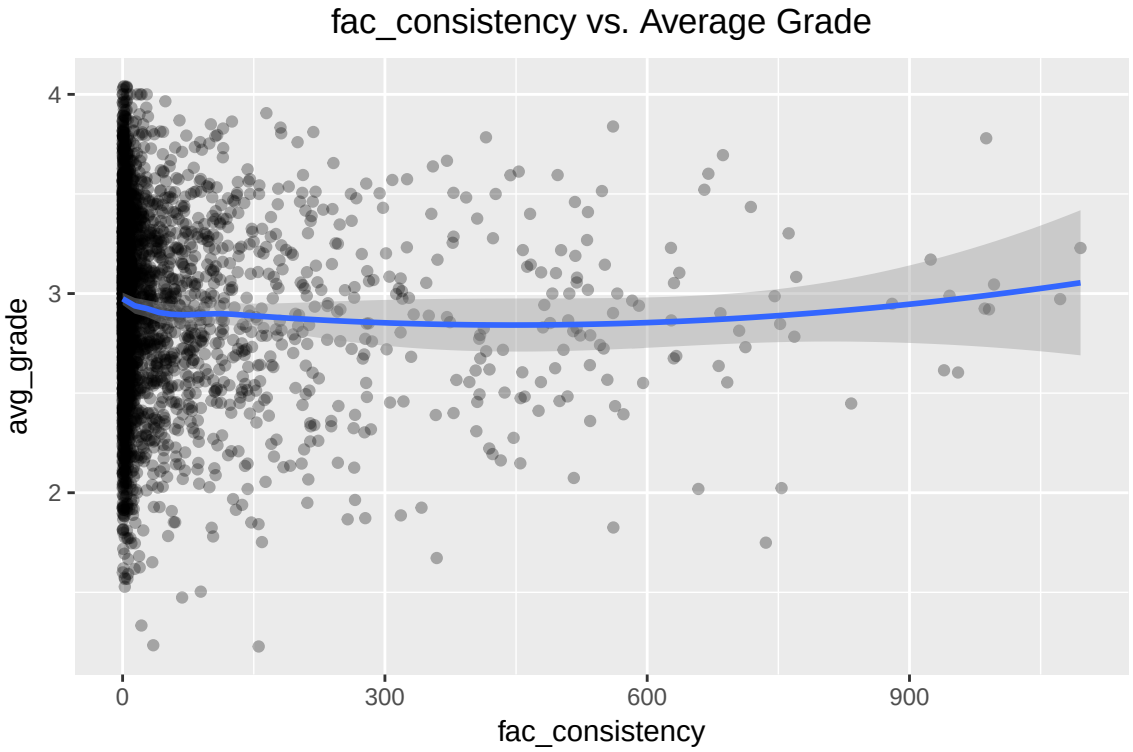
Discussion

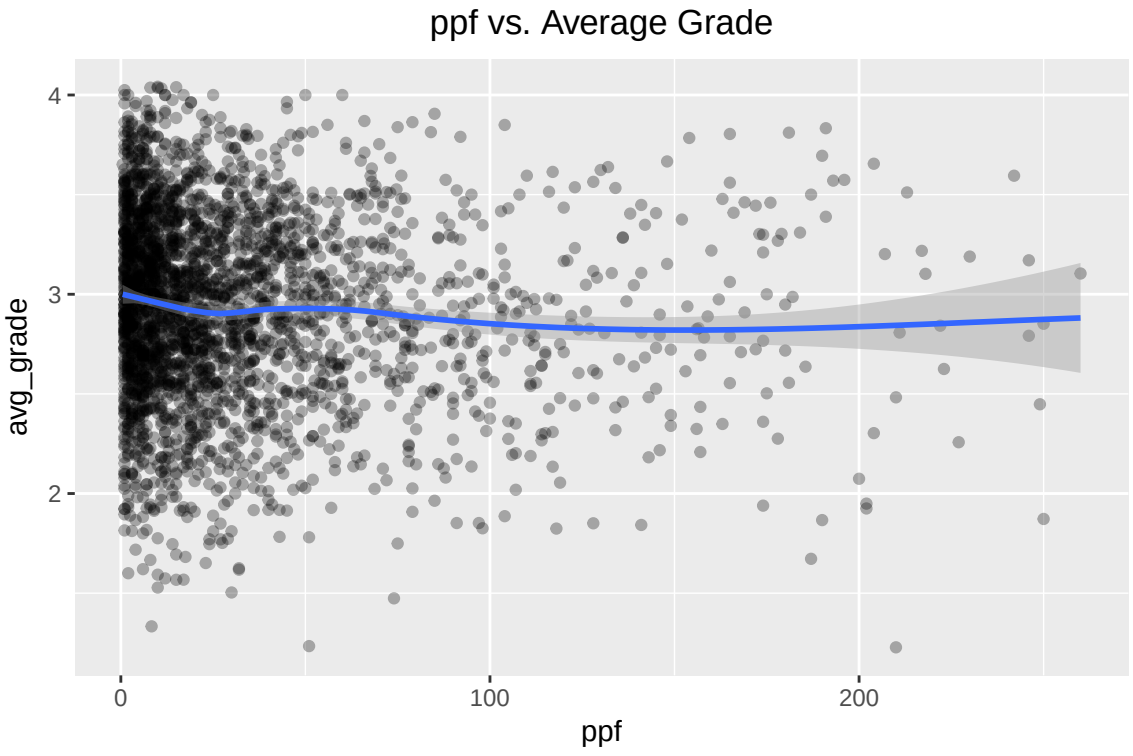
- Surprisingly, each of the six variables appears to have practically no effect on the instructor evaluation scores

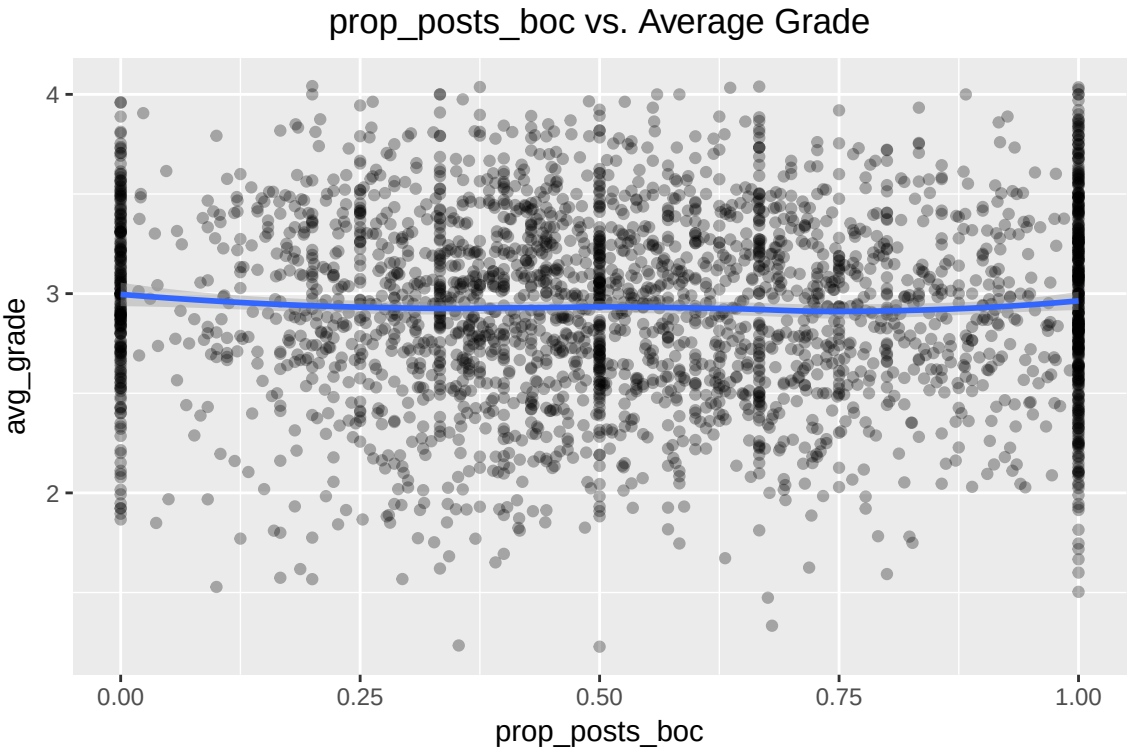
Average Grade






fac_consistency vs. Average Grade
avg_grade
4
3
2
0
300
600
900
fac_consistency

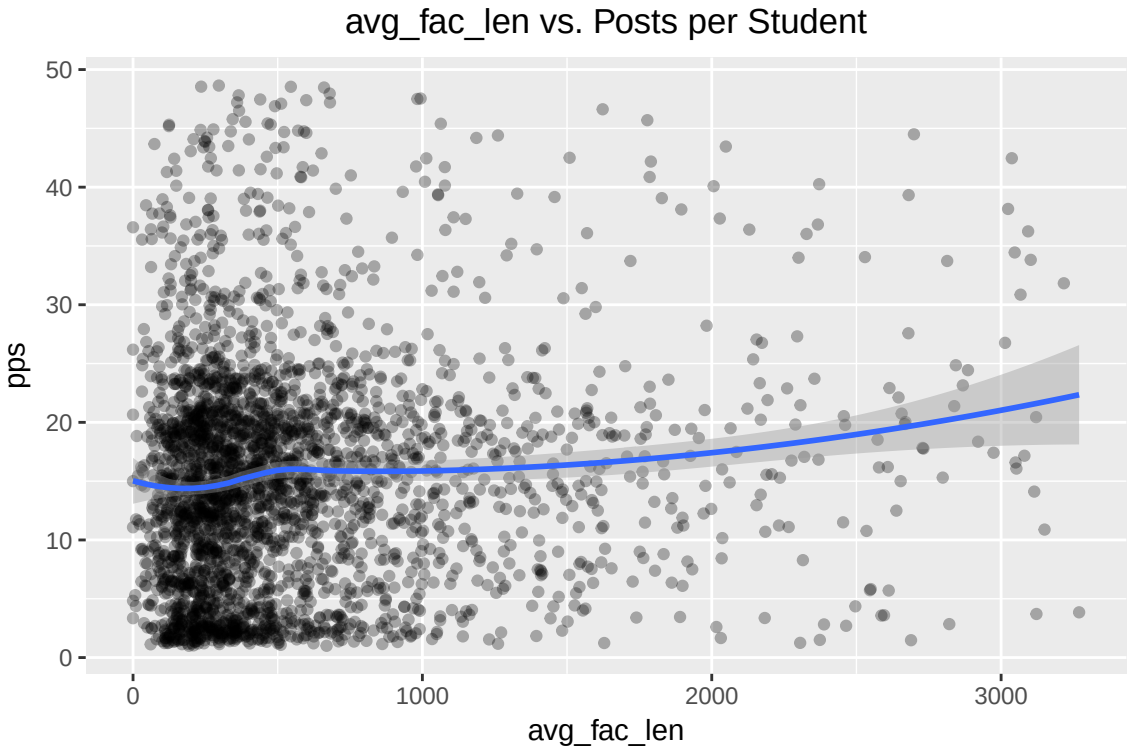

ppf vs. Average Grade
avg_grade
4
3
2
0
100
200
ppf

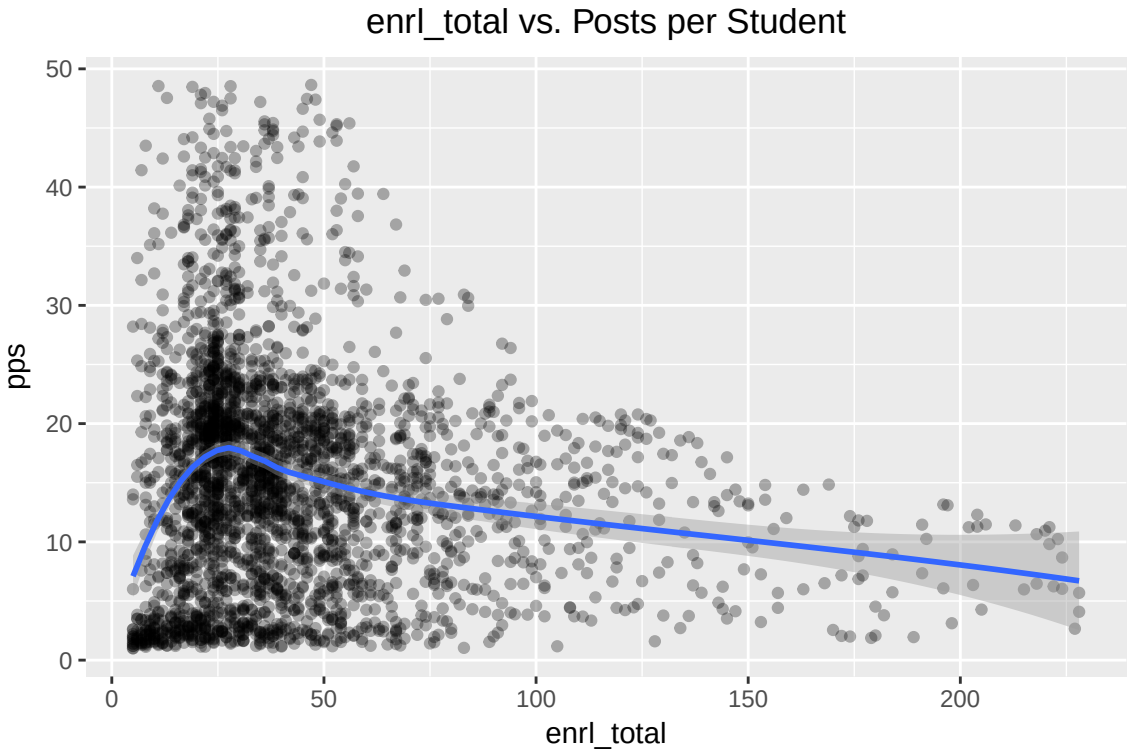


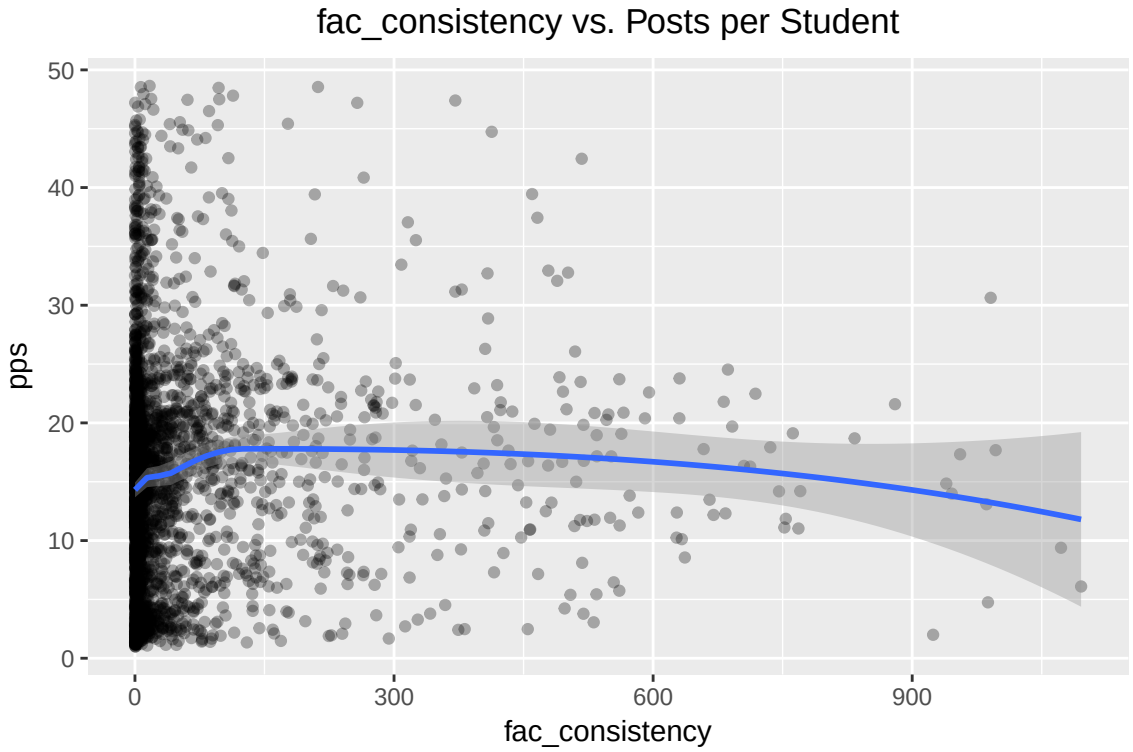
Discussion

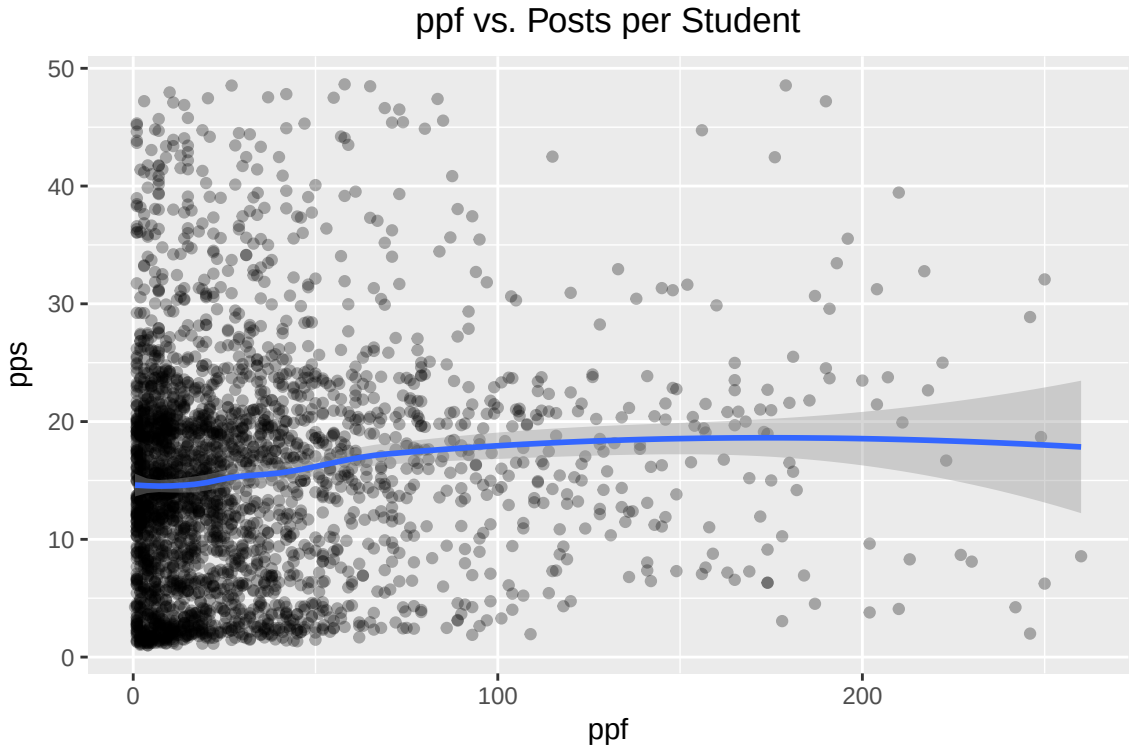
- Like the previous outcome, none of these plots suggest that the variables we thought to be important will be relevant in predicting grade.

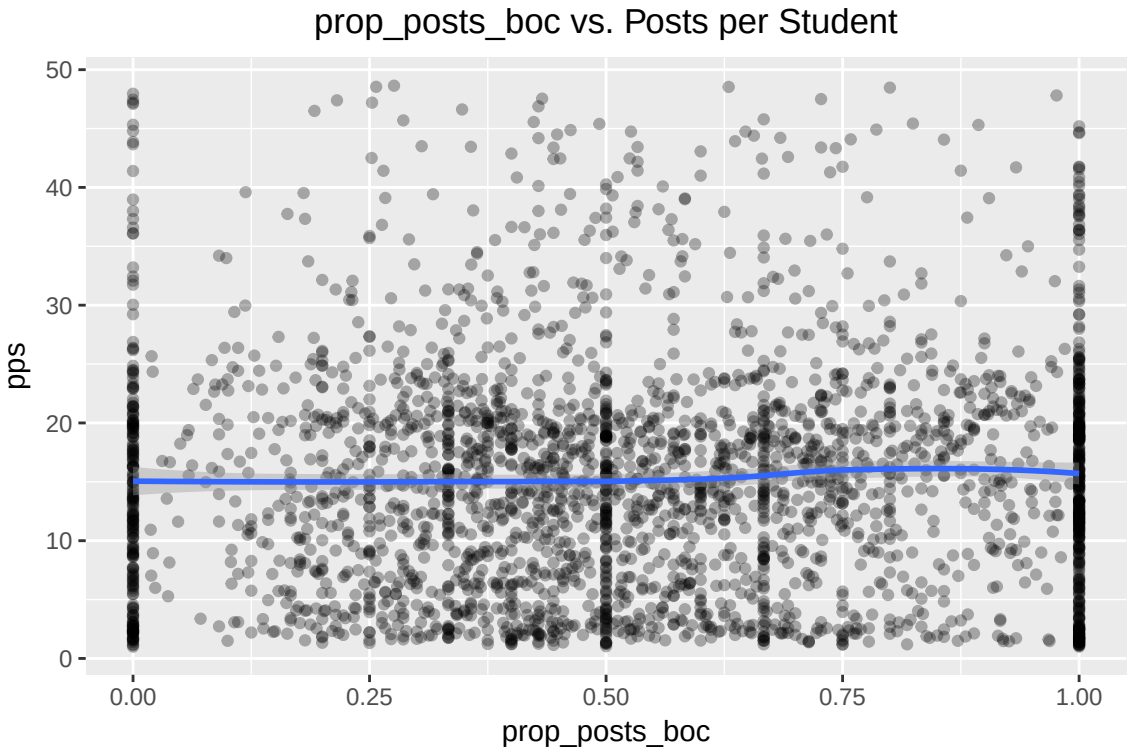
Posts per Student






fac_consistency vs. Posts per Student
pps
fac_consistency
0
10
20
30
40
50
0
300
600
900

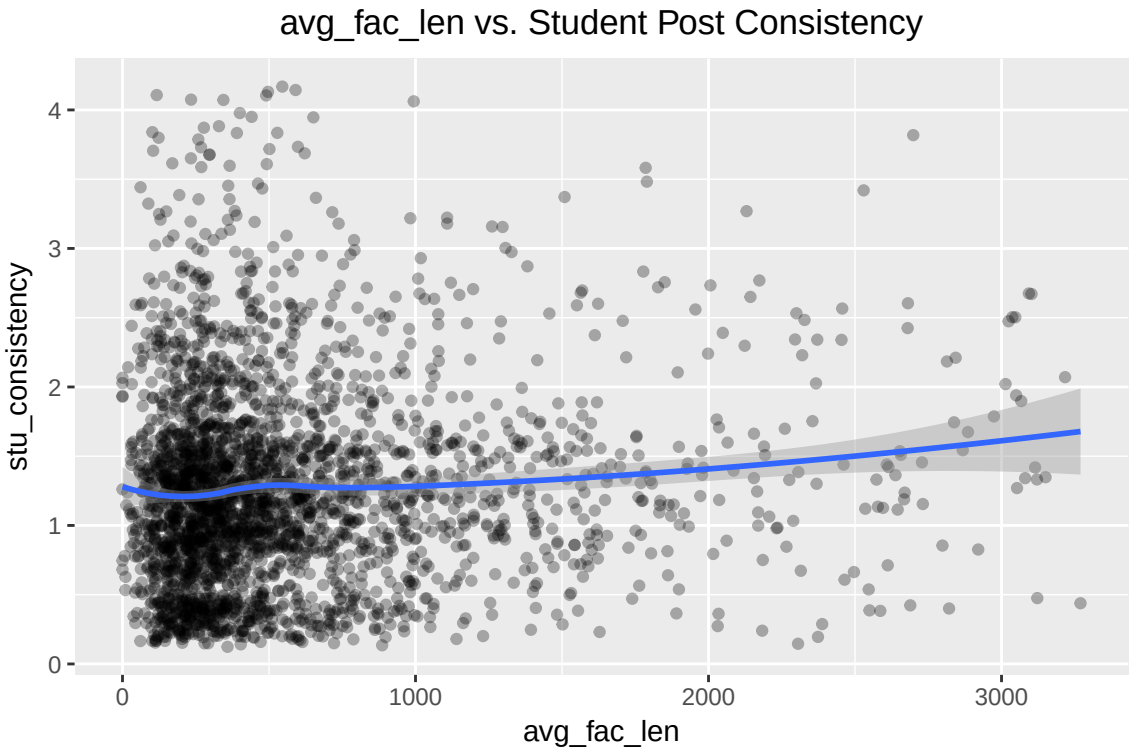

ppf vs. Posts per Student
pps
ppf
0
10
20
30
40
50
0
100
200

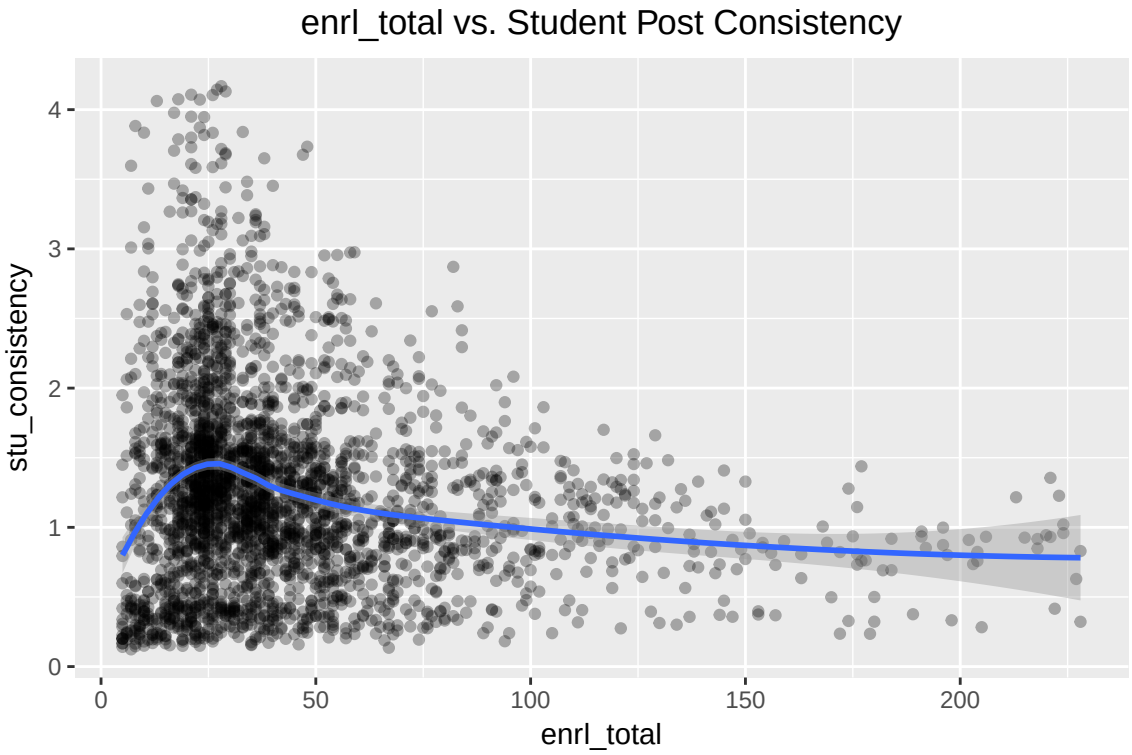


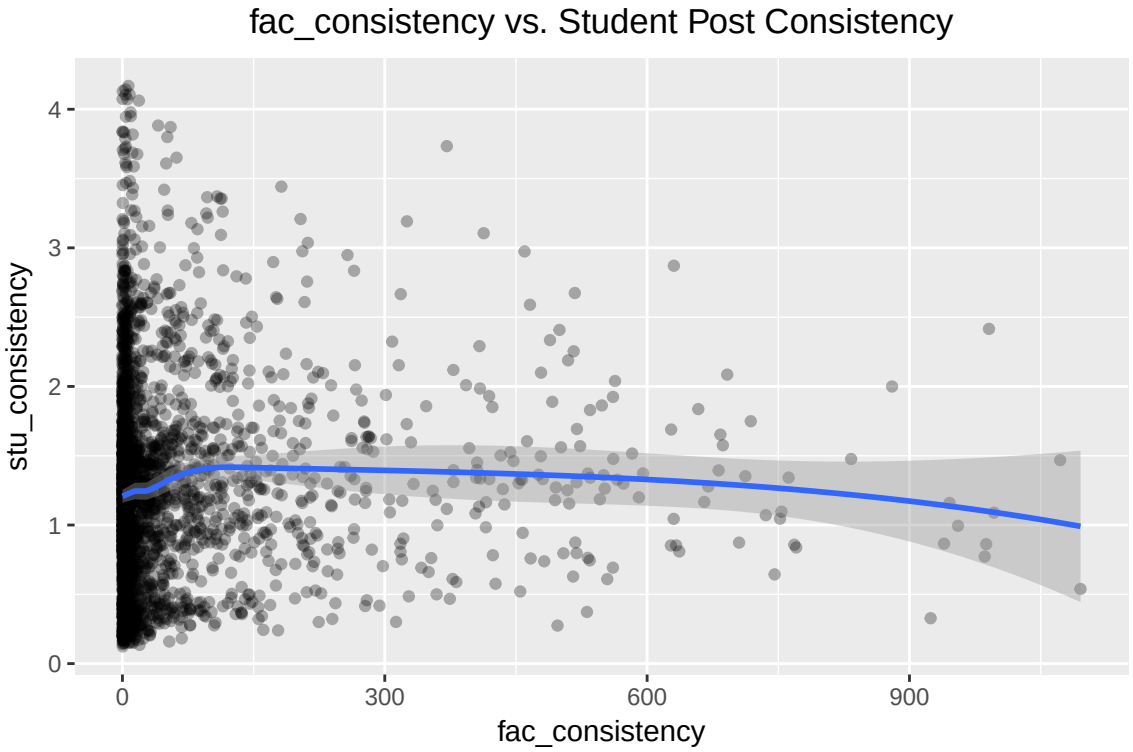
Discussion

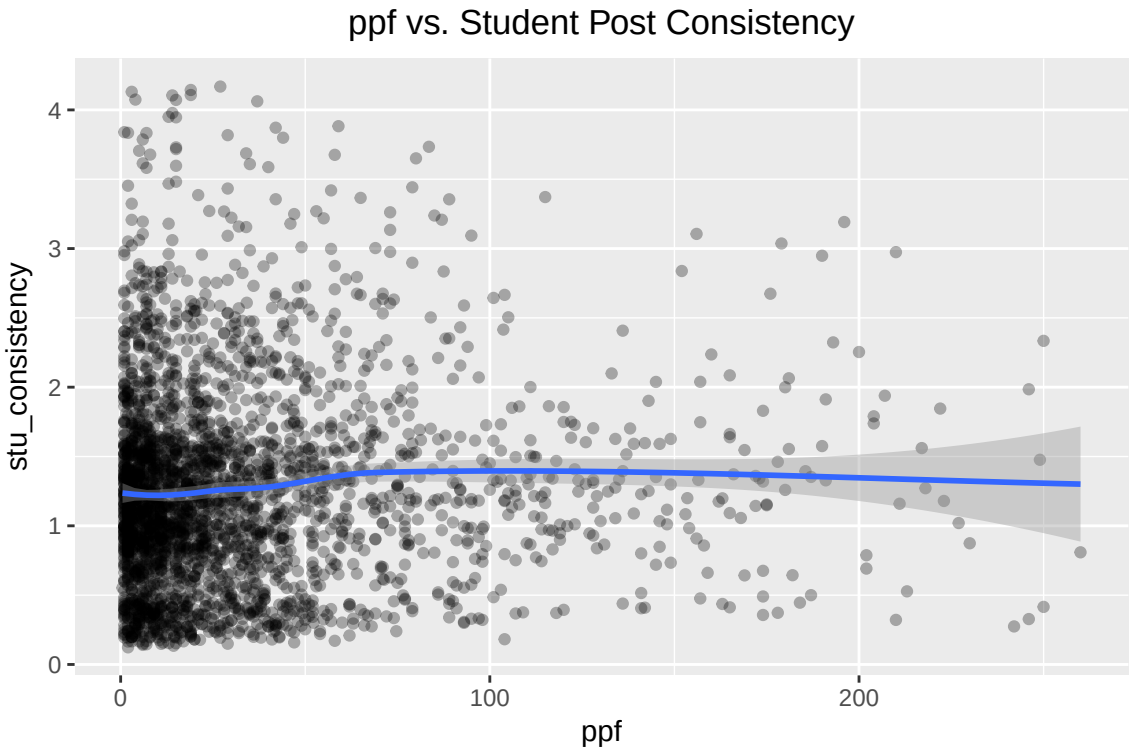
- Disappointingly, none of the variables here seem to be strongly related to the number of posts per student. The average faculty post length seems to have a small positive relationship while `enrl_total` and `fac_consistency` have a share a negative one.

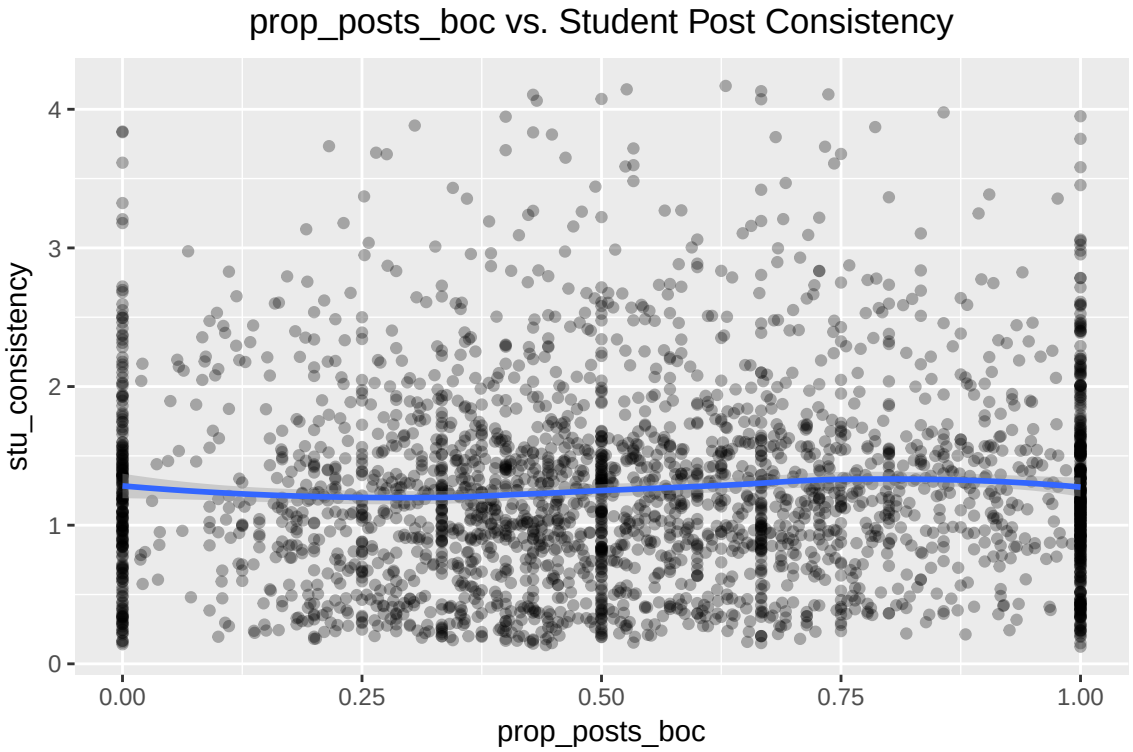
Student Post Consistency






fac_consistency vs. Student Post Consistency
stu_consistency
fac_consistency
0
1
2
3
4
0
300
600
900


ppf vs. Student Post Consistency
stu_consistency
ppf
0
1
2
3
4
0
100
200



Discussion

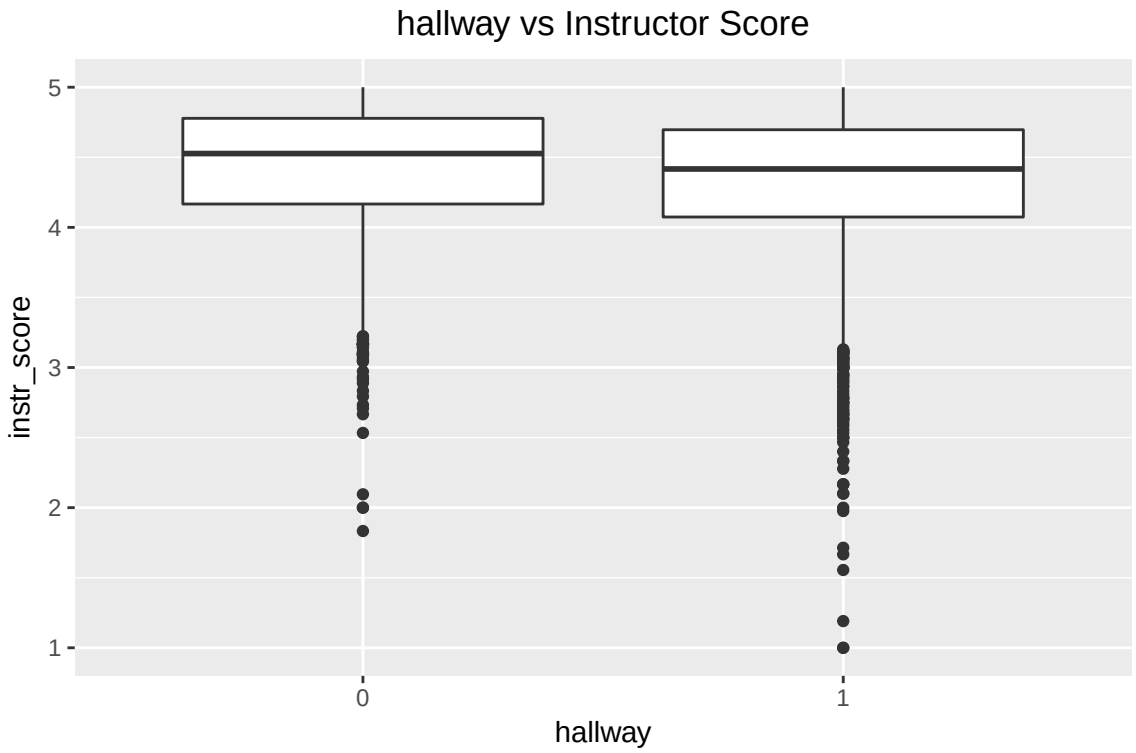
- It appears that for very low enrollment courses student post consistency is higher than average. Other than this relatively small finding, the rest of the variables do not appear to show any significant relation with student post consistency.

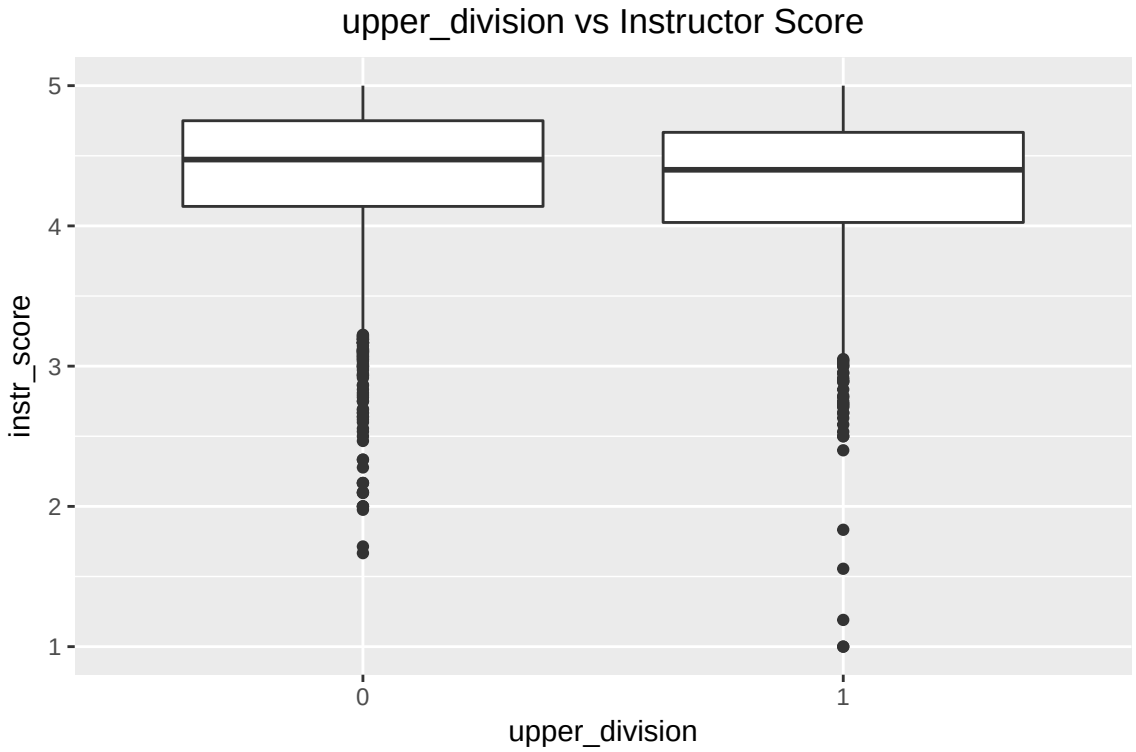
Binary Variable Exploration

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.

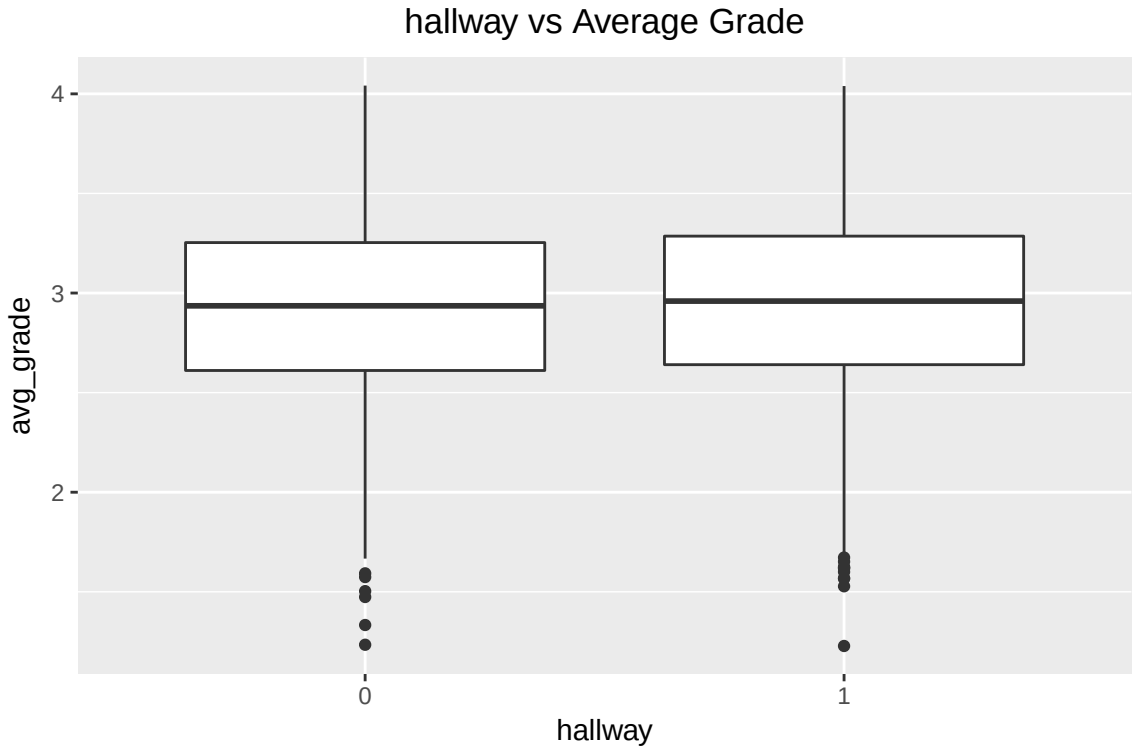
Now that we have practically abandoned all hope with the continuous variables, we need to see if the same pattern is going on with the binary variables.

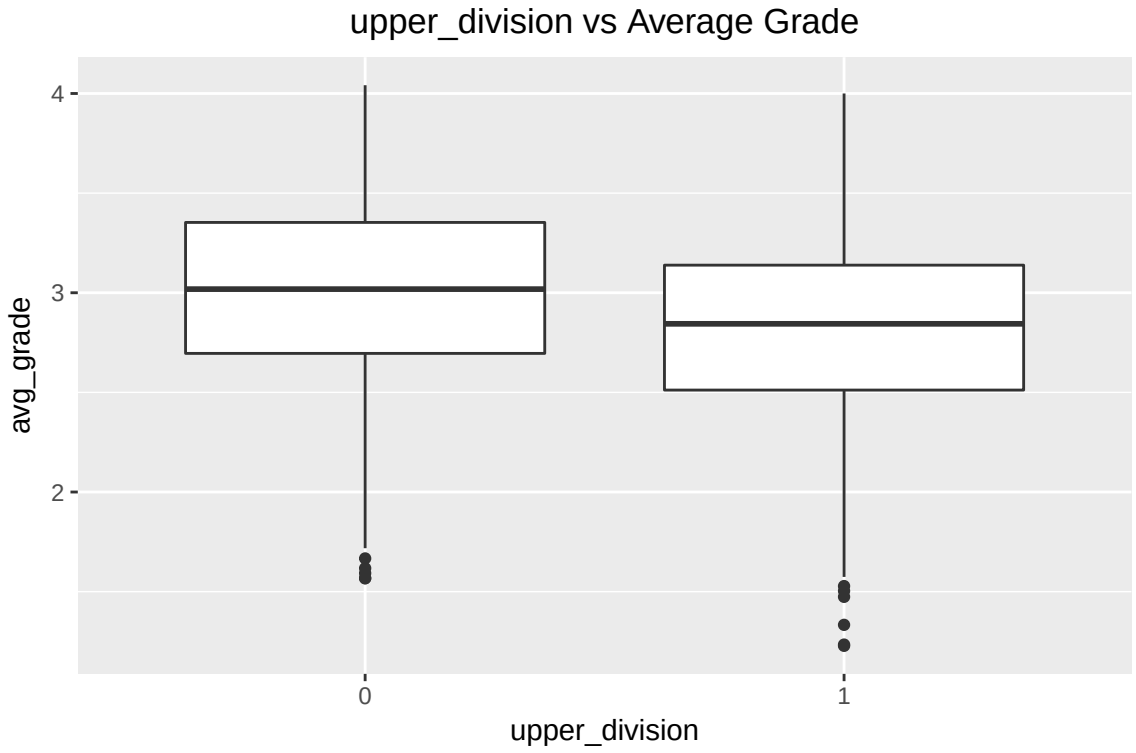
Instructor Score



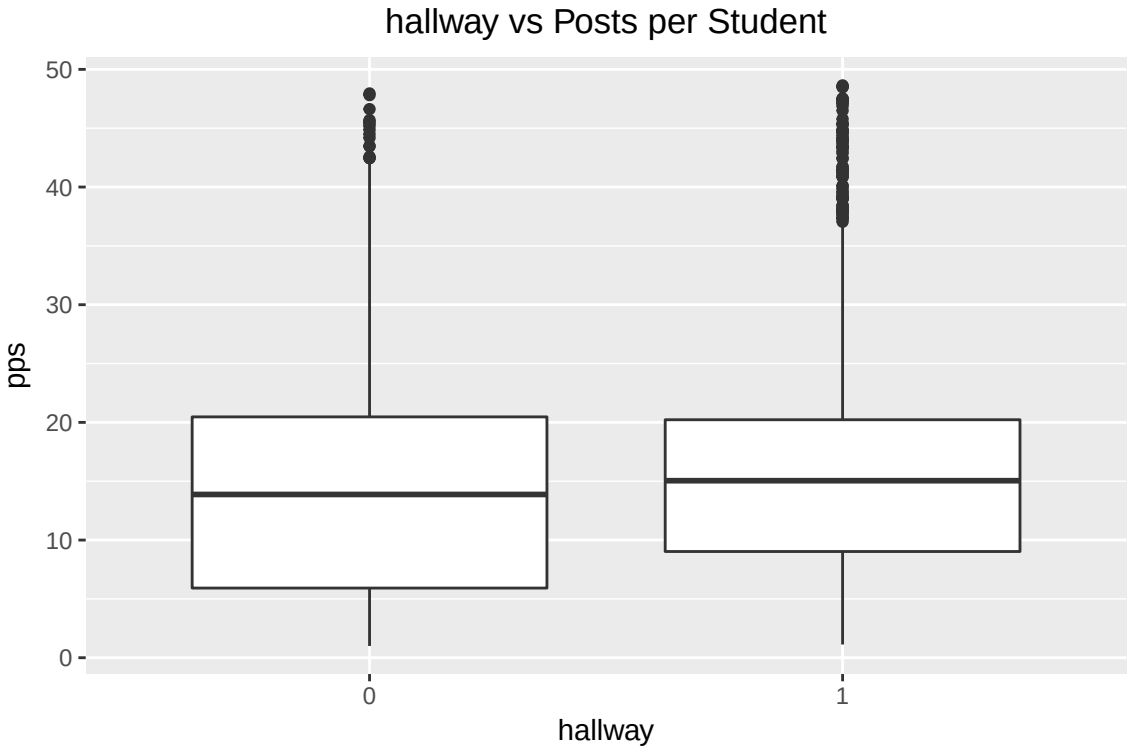


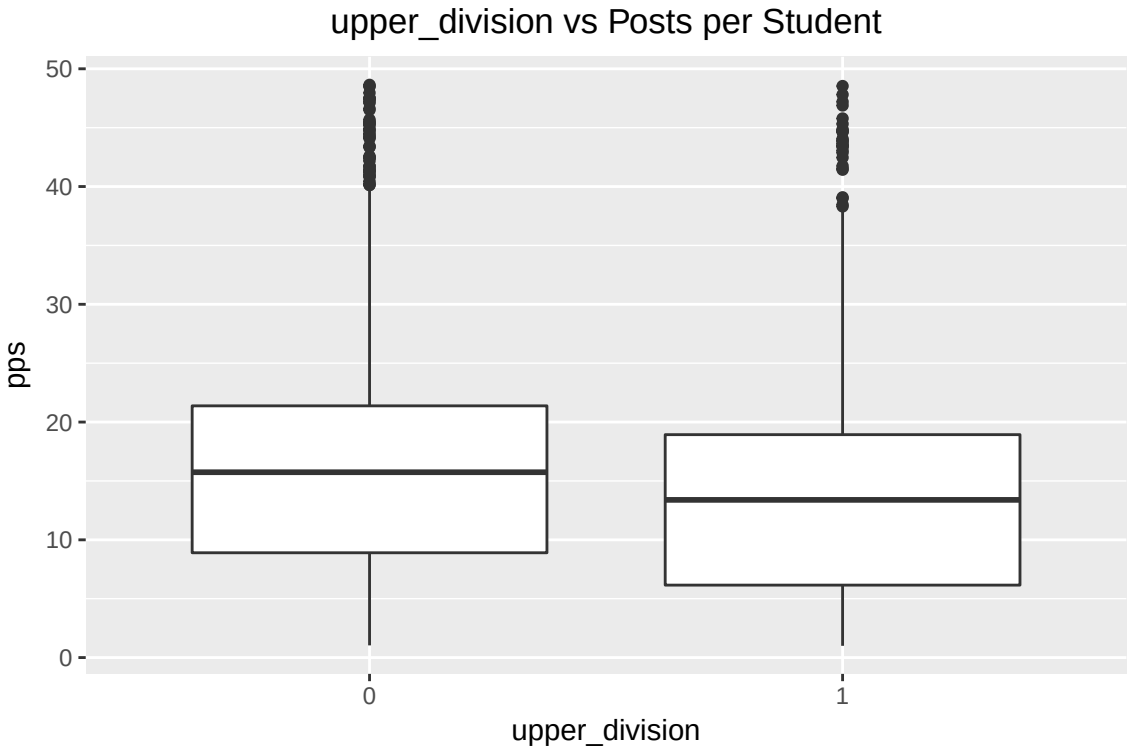
Average Grade



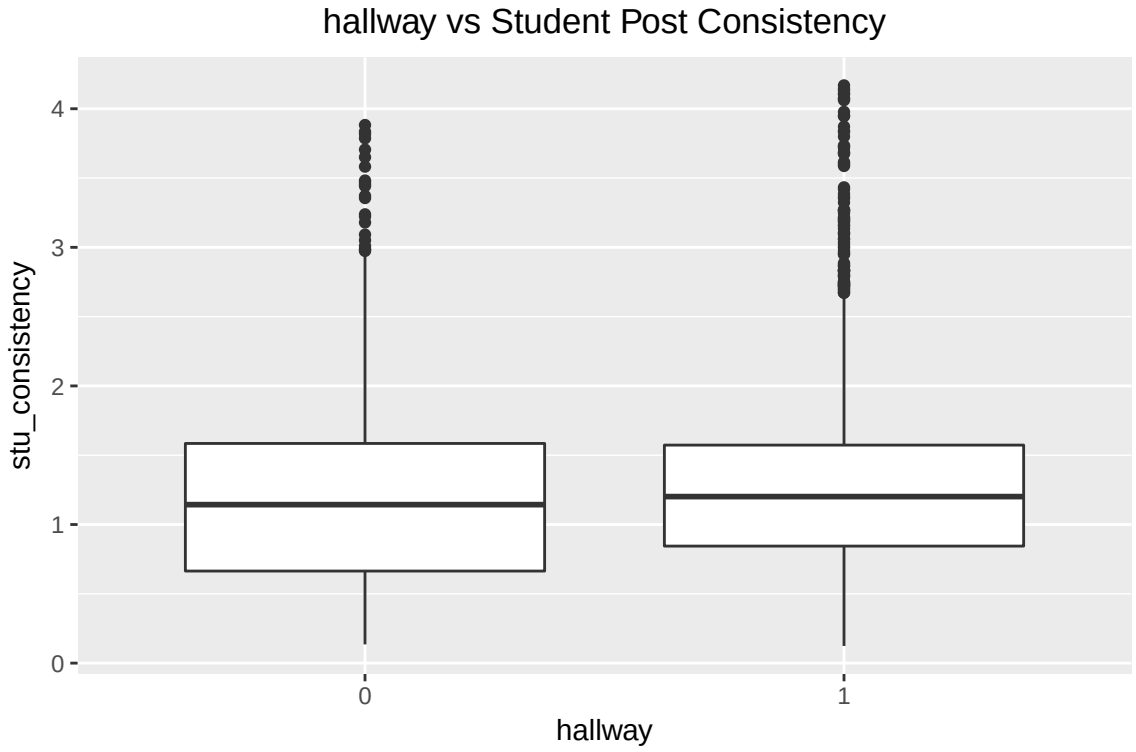


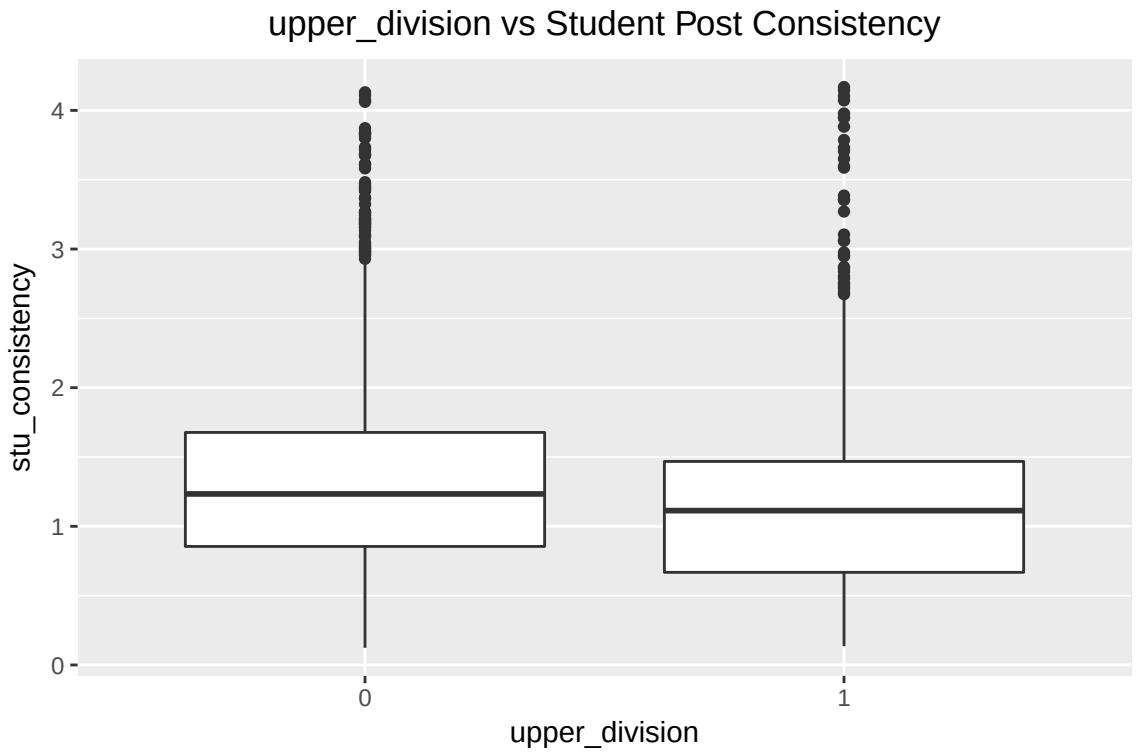
Posts per Student





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.