Abstract

Kaggle is a data modeling competition service, where participants compete to build a model with lower predictive error than other participants. Several years ago they released a reduced service that enables instructors to run competitions in a classroom setting. This paper describes the results of an experiment to determine if the participating in a predictive modeling competition enhances learning. The evidence suggests it does. In addition, students were surveyed to examine if the competition improved engagement and interest in the class.

Keywords: technology, statistical modeling, data science, statistics education

1 Introduction

Kaggle is well-known for the richly funded data competitinns, where participants compete to score the lowest error in their model fitting. Recently, they have made it possible to run in-class competitions, private and limited to members of a university course. This work explores how student participation in these challenges improves performance, engagement, interest

Campbell & Austin (2002) Schubert et al. (2013; Chi et al. 1981)

2 Data collection

2.1 Melbourne price data set

This data set was built in two stages. First the original sales in the Melbourne metropole ware downloaded then a few artificial features were add.

The original data was taken from the weekly report published by Domain and compiled by Home Price Guide (R) *COMENT: Found the symbol.* The reports are published as the pdf document. To read them into R we used the XXX package *COMENT: reference to our new package for reading those reports.* The weekly report contains the information about the properties full address, number of bedrooms, property type (house, unit/ duplex, townhouse, development site or other residential), sold price, type of sale (sold, sold prior, passed in, no bid, vendor bid, withdrawn prior to auction, sold after auction and N/A - price or highest bid not available) and agent name. We used the weekly reports from 2nd of February 2013 to 17th of December 2016. We also removed all the records without the price. Over all we had 75,367 sold properties for the data competition.

To enrich this data and make the model building process more interesting we add a few 'artificial' features, namely – number of visitors, the average rating gave to the property by the visitors, number of car spaces, number of baths, land size and house size (builded area).

2.2 Spam data set

3 Participants

- To examine different type of students (undergraduate vs post graduate) and different learning environments (different university) the study run across three different subjects and two universities, Monash University (ETC3250 and ETC5242) and University of Melbourne (MAST90083 subject).
- Time period: semester two 2016
- MAST90083: 63 enrolled students; postgraduate level; background: math & stats, IT and actuarial science;
- ETC5242
- ETC3250

4 Methodology

5 Results

5.1 Test scores

5.1.1 MAST90083

We have examined two normalizations. Once, we normalized the score for the question (or group of questions) by the maximum possible score for the question (or the group of questions), denoted as PTQ. We also normalized by the total exam score, denoted as PTE. The PTQ and the PTE scores were calculated for each student for the following four clusters of questions

- Questions related to classification methods
- Questions related to regression methods

- Questions related to the classification and regression methods
- Questions related to other topics that covered during the subject but unrelated to the data competition

In addition, the PTQ and the PTE scores were calculated to each of the questions in the first and the second clasters.

Figure 1 shows the boxplots for the PTQ scores and the PTE scores for each group of students[1].

The plots in the left column summarize the PTQ scores and the plots in the right column summarize the PTE scores. The four questions clusters are corresponding to the four rows in the figure.

Examining the figure, we can see clearly the positive correlation between student's scores and the type of the data competition. Namely, the median score (PTQ and PTE) of the students that took part in the regression related competition is higher for regression questions. Similarly, for the students that took part in the classification related competitions. There is no significant different in the medians of the two groups of students for the question related to the general question covered both the classification and the regression methods. As well as for the all other question related to other topics covered during the subject.

The statistical significance of this results was examined via the "permutation tests" ... [TO COMPLETE].

5.1.2 Examining individual questions

In the final exam two questions were related to the classification methods (Q1 and Q10) and five questions were related to regression methods (Q5, Q6, Q8, Q15 and Q16).

Figure 2 shows the scores distribution for the classification methods questions: Q1 and Q10.

[1]The left boxplot is related to the students that took part in the data competition related to the regression methods, the Melbourne Price competition. The right boxplot is related to the students that took part in the data competition related to the calssification methods, the Spam classification competition.

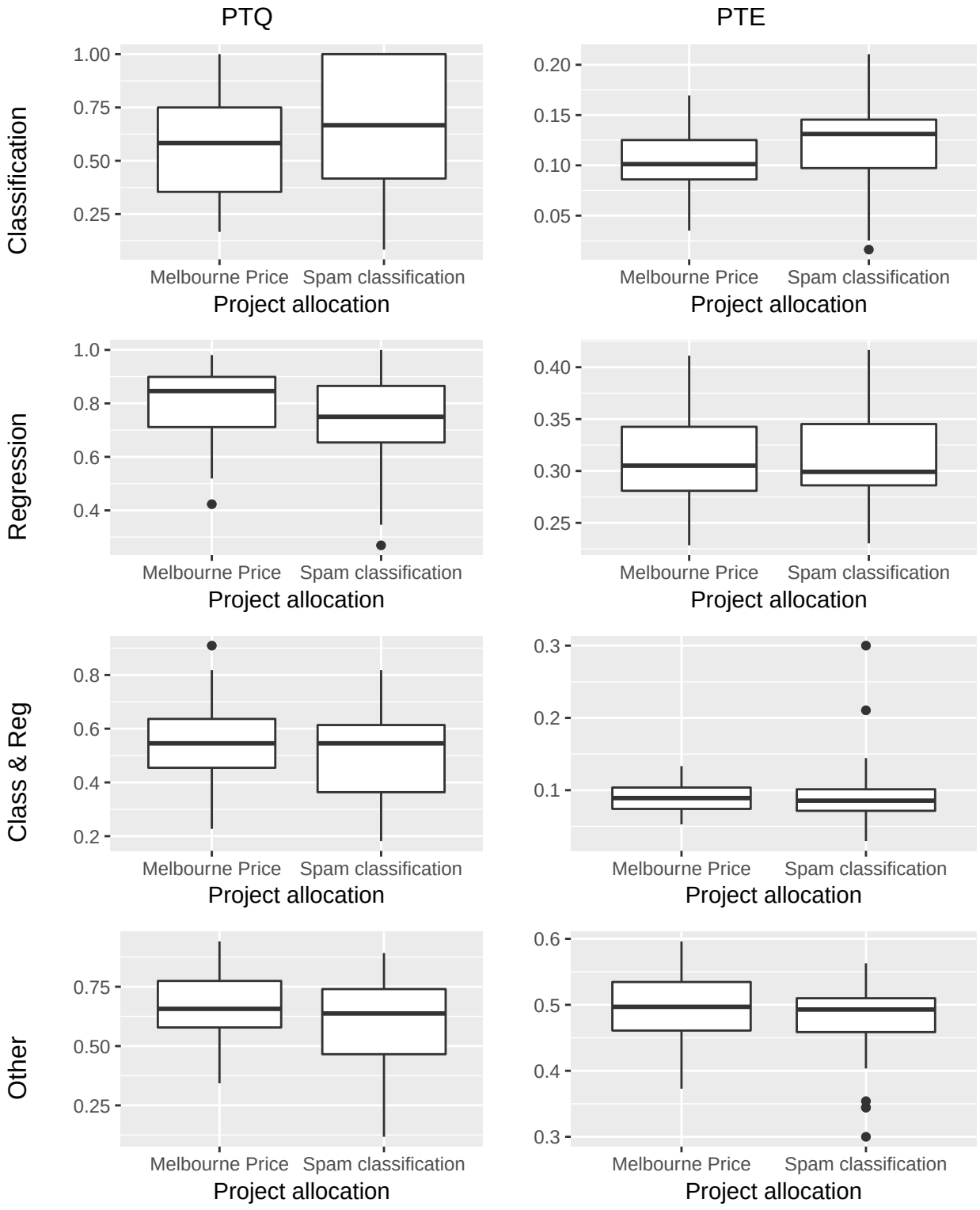

Figure 1: PTQ and PTE scores distributions over different question groups

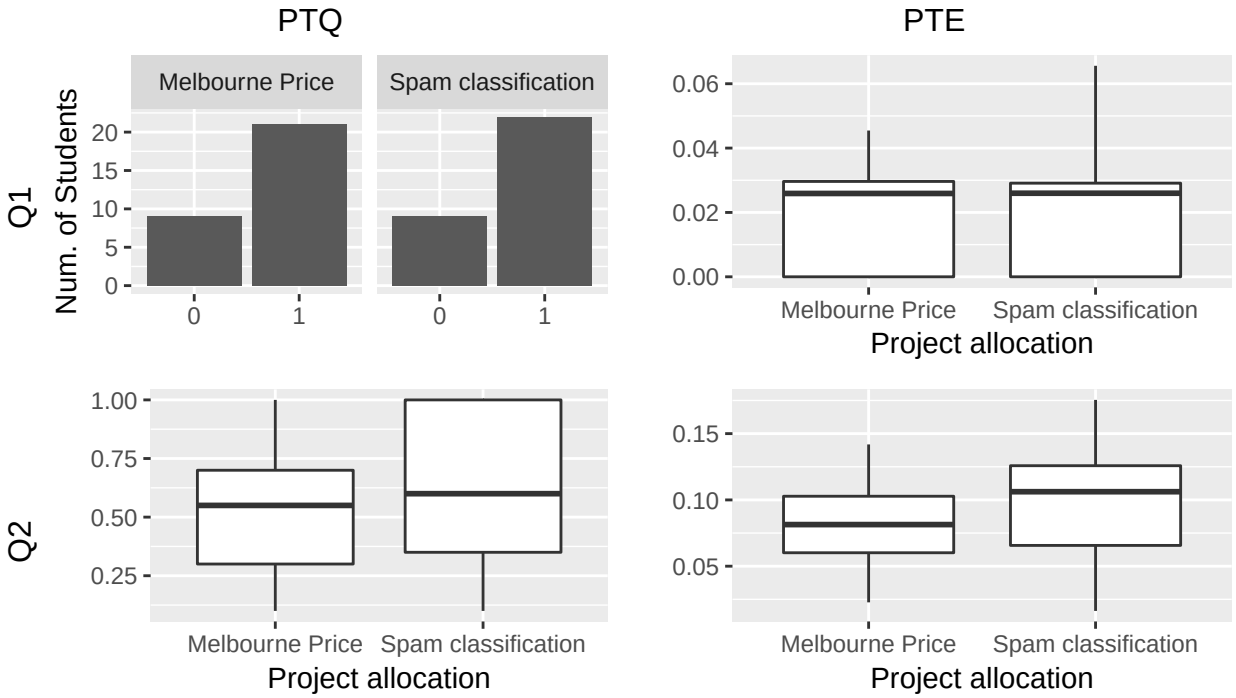

Figure 2: PTQ and PTE scores distributions for classification questions

Q1 was a multi-chose question, worth only 2 points (out of 100). Both plots, of the PTQ and the PTE scores show no difference between the two groups of students.

Q10, was a relatedly large question, worth 10 points. Examining the medians of the PTQ scores indicate only mild advantage for the students from the classification competition. However, looking on the PTE scores suggests that the classification questions was much easier for the students from the classification competition comparing to all the other questions in the exam.

Figure 3 shows the scores distribution for the regression methods questions: Q5, Q6, Q8, Q15 and Q16. There was no difference in performances of the two groups in the Q5, a small 2 point, question. Interestingly, for question Q6, 6 point question that asked to explain different line of code, the median PTQ scores were similar in the two groups. However, the lower 50 percent of student's scores was spread over the much lower scores. In fact, the second quartile of the scores of students participated in the classification competition covered the range of the scores of the 50 percent of student's scores participated the regression competition. This suggest that participation in the competition help the student to remember the R command better. Similar behaviour can be observed in the

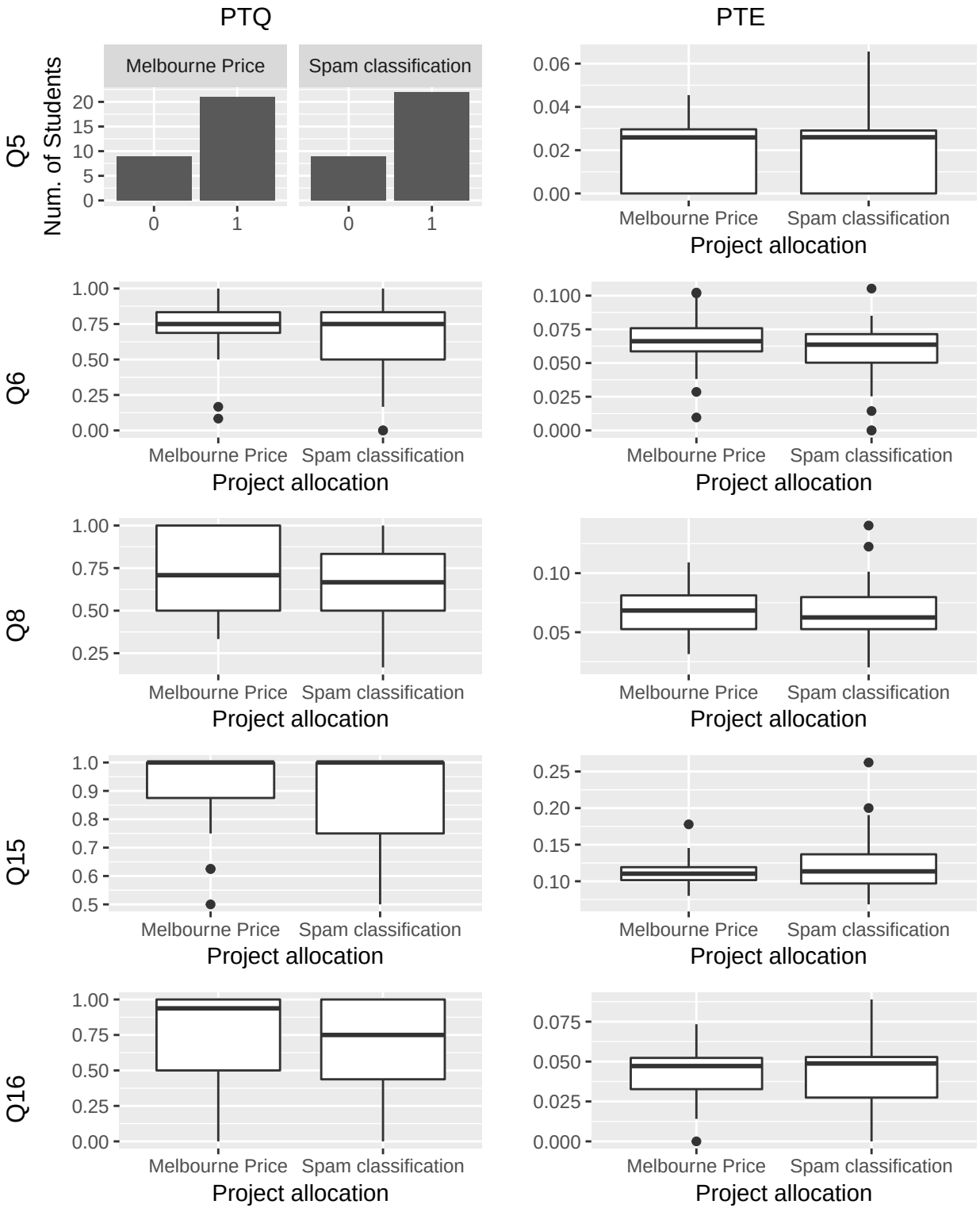

Figure 3: PTQ and PTE scores distributions for regression questions

performances in Q15, and easy 8 point question, dealing with the differences between Lasso and Ridge regressions.

In the performances in Q8, a 8 point fairly easy question, that dealing with the regression trees, a 'mirror' behaviour can be observed. The median PTQ scores were similar in the two groups, however the third quartile of the scores of students participated the regression competition spread in the range of the 3^{rd} and the 4^{rd} quartile of the students' scores participated the classification competition.

Finally, looking on the PTQ scores, in the Q16, 4 point question, that required a deep understanding of the GAM method, one of the regression technic, the students participated the regression competition had a clear advantage over their pears participated the classification competition.

Examining all the remind exam questions weren't arise any different between the two groups of students.

5.2 Engagement

To examine the correlation between student's engagement levels and the performances in the exam we plot the number of submission during the competition versus the performances in the exam, Figure 4. Once again we exam the performances based on two normalizations. Once normalizing by the total possible marks for the relevant cluster of questions[2] (PTQ) and once normalizing by total exam marks (PTE).

In Figure 4 we can clearly see a weak positive correlation between the number of submissions during the data competition and the scores normalized to the total possible marks for the relevant cluster of questions. This suggest that as more engaged the student was with the competition, the question about the methods relevant to her competition were easier to her. There is no correlation between the number of submission and the marks for the relevant cluster of questions normalized by the total exam marks (PTE). *BECAUSE ... the questions that unrelated to the data competitions (51 points)? students put less*

[2]For the students participated in the Melbourne Price competition is the cluster of regression questions. For the students participated in the Spam Classification competition is the cluster of questions about classification methods.

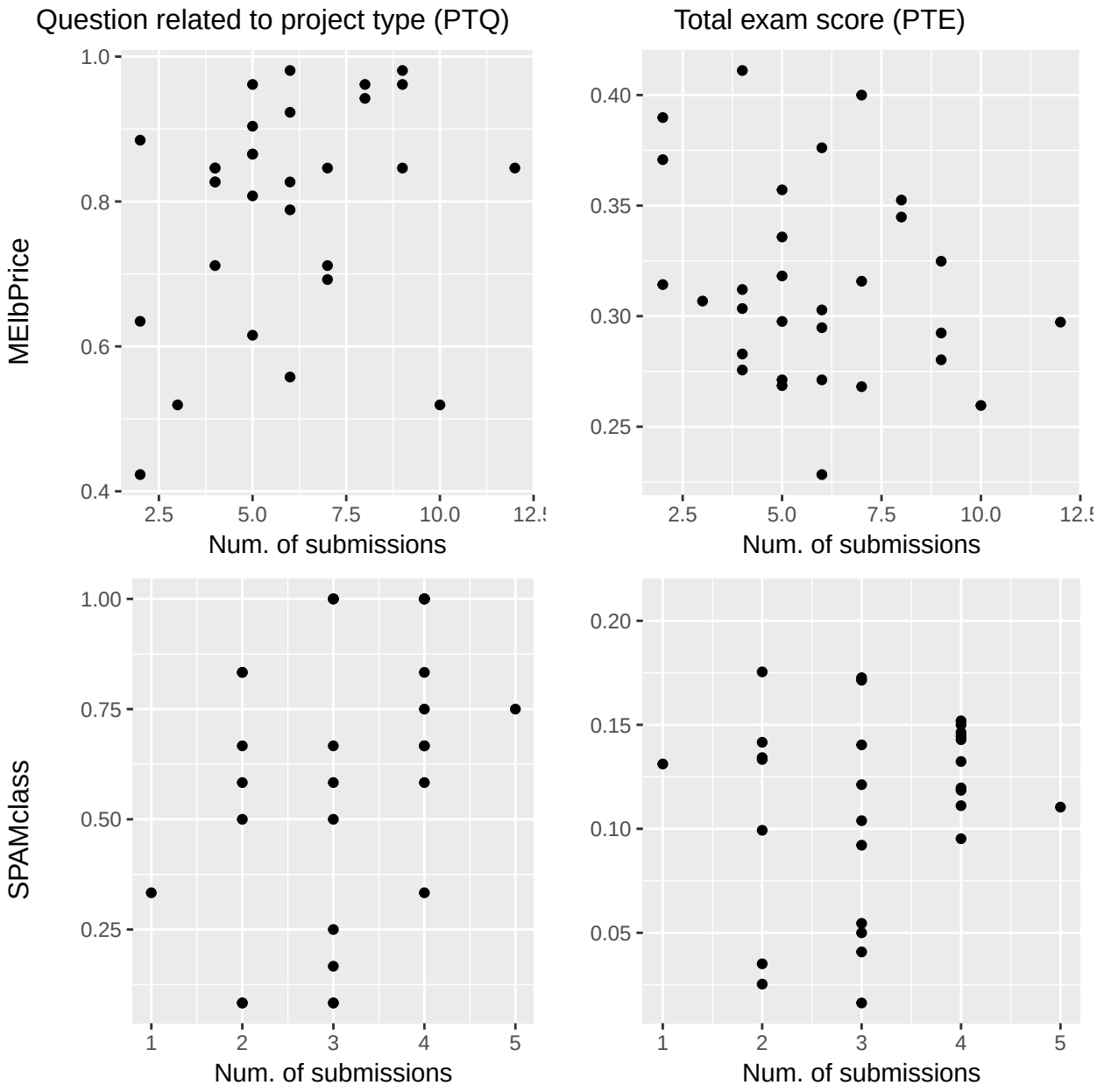

Figure 4: Number of submission vs exam grades

effort to learn other material? Harder questions?

We didn't found any evidence for correlation between the performances in the competition (final score) and the performances in the exam. This suggest that the single fact of participation improve the students marks in the exam. Not necessarily better students in the competition have grater chances to success in the exam.

5.3 Interest

6 Discussion

References

Campbell, J. I. & Austin, S. (2002), 'Effects of response time deadlines on adults' strategy choices for simple addition', *Memory & Cognition* **30**(6), 988–994.

Chi, M. T., Feltovich, P. J. & Glaser, R. (1981), 'Categorization and representation of physics problems by experts and novices', *Cognitive science* **5**(2), 121–152.

Schubert, C. C., Denmark, T. K., Crandall, B., Grome, A. & Pappas, J. (2013), 'Characterizing novice-expert differences in macrocognition: an exploratory study of cognitive work in the emergency department', *Annals of emergency medicine* **61**(1), 96–109.