

# Kaggle-in-class Data Challenges can Boost Student Learning

Julia Polak

Department of Statistics, University of Melbourne  
and

Dianne Cook

Department of Econometrics and Business Statistics, Monash University

March 22, 2018

## 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:* instructional technology, statistical modeling, data science, statistics education, data mining

# 1 Introduction

Kaggle is well-known for the richly funded data competitions, 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 Experimental setup

### 2.1 Data collection

Explain experimental design: two competitions, two locations, two courses, one with students randomised into two groups, the other with control group of students and regression challenge, and exam questions. (Not sure how to incorporate ETC3250 students at the moment.)

### 2.2 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 2017
- MAST90083:

63 enrolled students; postgraduate level; background: math & stats, IT and actuarial science;

Students were randomly assigned to one of two groups, regression (R) or classification (C). Each group will have a data competition designed for that topic, made available through kaggle in class (<https://inclass.kaggle.com>). Students worked individually to

build predictive models, and submit predictions for the first 16 days. On the 16<sup>th</sup> day of the competition students received an email that they allowed to establish groups and continue the competition as a group. The competition was continued for additional 7 days. In case they chose to form a group, the final score of the group in total was accounted as their individual mark towards the final grade for the subject and analysed in this project. Only 3 groups were formed, two groups with 3 students and one group with two students. All other students chose to continue the competition individually.

Note that the data competition was compulsory to all the students. The volunteering part was the questionnaire at the end of the semester asking for feedback on aspects of the class such as their competition experience, their favourite topic of the semester, and their enjoyment of the material in the subject.

- ETC5242
- ETC3250

## 2.3 Competitions

## 2.4 Used Platform

Kaggle in class (<https://inclass.kaggle.com>)...

# 3 Methodology

# 4 Results

## 4.1 Performance

### 4.1.1 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 clusters.

Figure 1 shows the boxplots for the *PTQ* scores and the *PTE* scores for each group of students. 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 classification methods, the Spam classification competition.

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].

##	V1	mean_reg	mean_class	dif	pv_perm	pv_ttest
----	----	----------	------------	-----	---------	----------

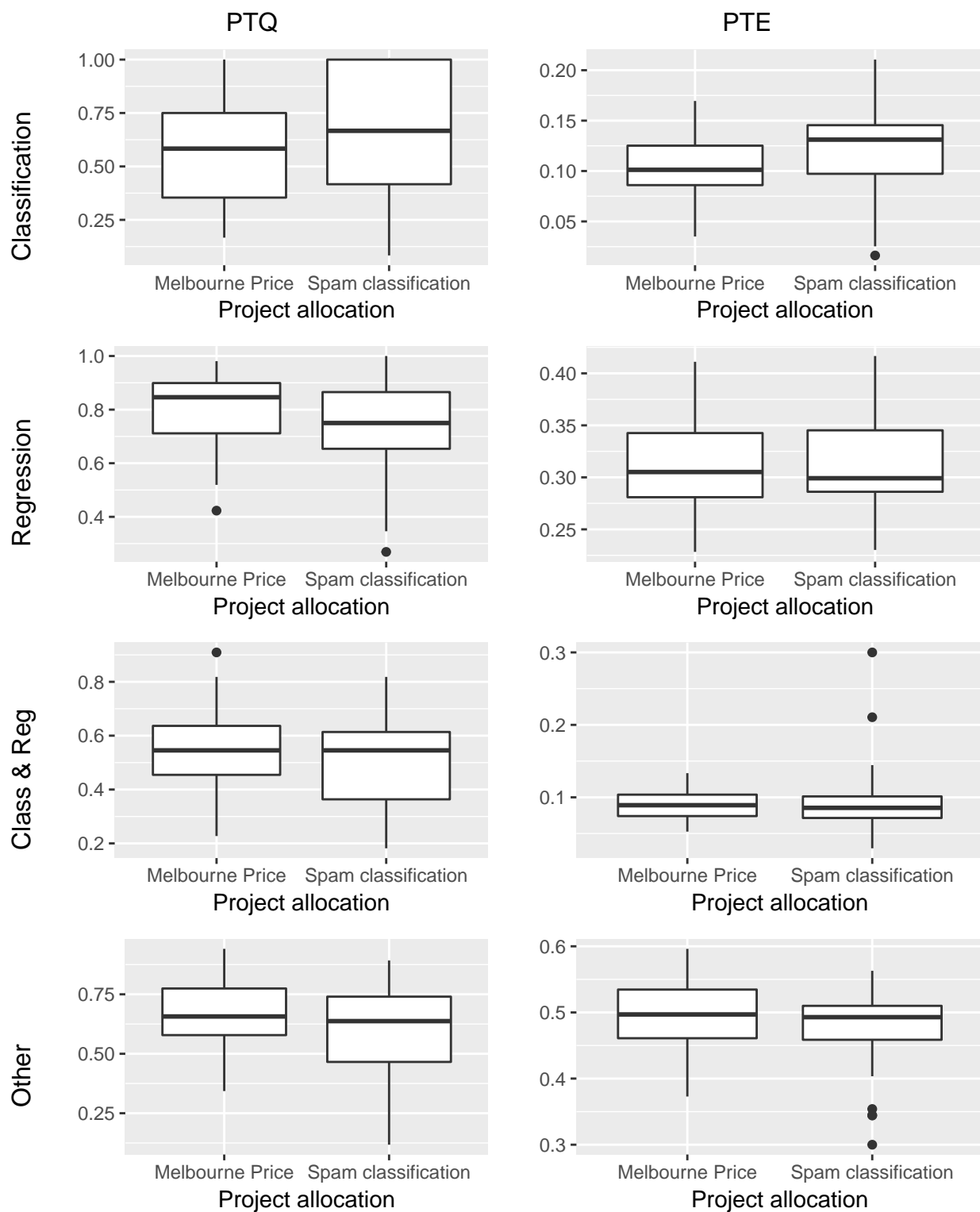


Figure 1: PTQ and PTE scores distributions over different question groups

## 1		Mean MelbP	Mean Spam	Diff P.v.	Perm P.v.	t-test
## 2		PTQ	PTQ	PTQ	PTQ	PTQ
## 3	Classification	0.5833333	0.6397849	-0.05645161	0.739	0.8813086
## 4	Regression	0.7974359	0.7282878	0.06914806	0.058	0.02953775
## 5	Class & Reg	0.5515152	0.4985337	0.05298143	0.11	0.179733
## 6	Other	0.6486928	0.5882353	0.06045752	0.083	0.9954901
##	V7	mean_reg.1	mean_class.1	dif.1	pv_perm.1	pv_ttest.1
## 1		Mean MelbP	Mean Spam	Diff P.v.	Perm P.v.	t-test
## 2		PTE	PTE	PTE	PTE	PTE
## 3		0.1019125	0.1175769	-0.01566437	0.935	0.9672623
## 4		0.3141449	0.3127369	0.001407996	0.438	0.2909268
## 5		0.0911207	0.0968283	-0.005707603	0.684	0.208796
## 6		0.4928219	0.4728579	0.01996398	0.1	0.4674309

Next we examined the statistical significance of the mean differences between the performances of the two group of students.

### 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.

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.

##	V1	mean_reg	mean_class	dif	pv_perm	pv_ttest	V7
## 1		Mean MelbP	Mean Spam	Diff P.v.	Perm P.v.	t-test	
## 2		PTQ	PTQ	PTQ	PTQ	PTQ	

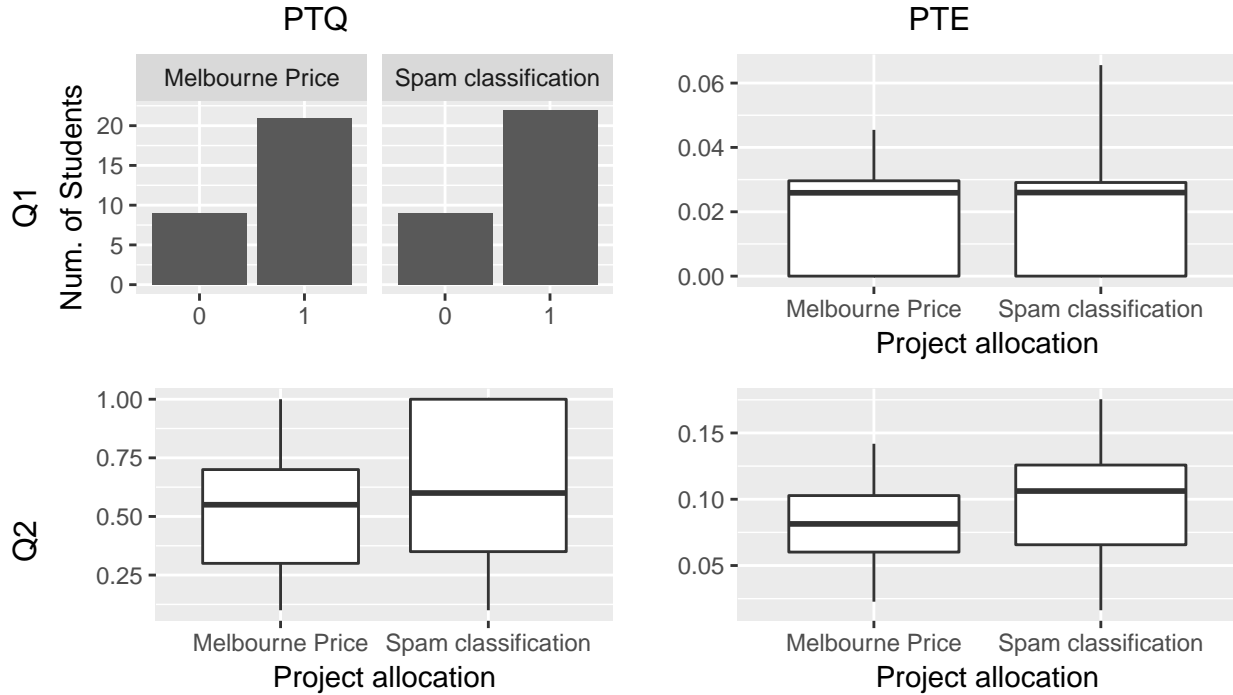


Figure 2: PTQ and PTE scores distributions for classification questions

```
## 3 Q10      0.56  0.6258065 -0.06580645    0.781  0.8378724
##   mean_reg.1 mean_class.1      dif.1 pv_perm.1  pv_ttest.1
## 1 Mean MelbP   Mean Spam      Diff P.v. Perm P.v. t-test
## 2           PTE           PTE           PTE           PTE           PTE
## 3 0.08117103   0.0955668 -0.01439576    0.924  0.6867794
```

Next we examined the statistical significance of the mean differences between the performances of the two group of students for Q10.

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

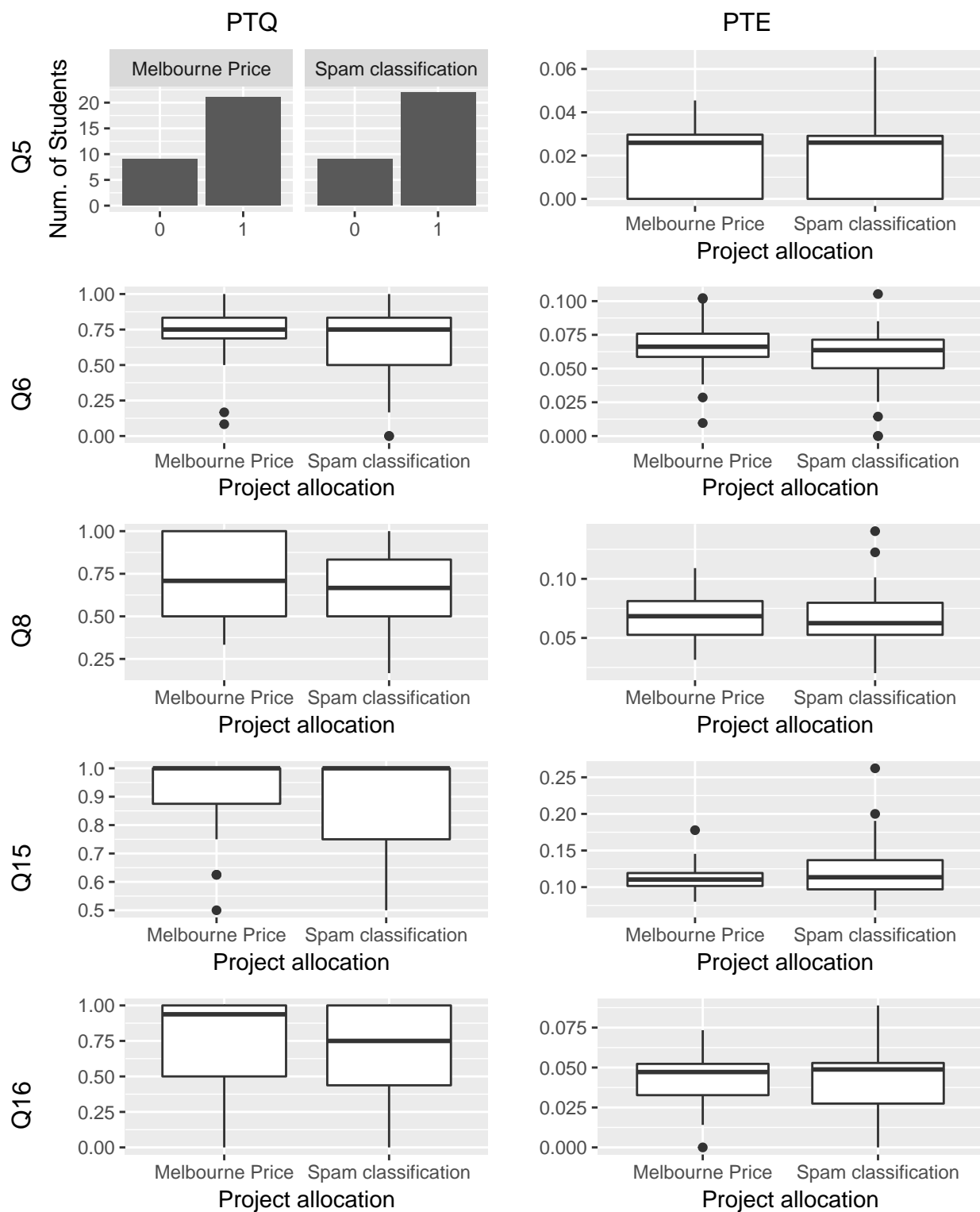


Figure 3: PTQ and PTE scores distributions for regression questions



student to remember the R command better. Similar behaviour can be observed in the 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<sup>rd</sup> and the 4<sup>rd</sup> 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 peers participated the classification competition.

##	V1	mean_reg	mean_class	dif	pv_perm	pv_ttest	V7	mean_reg.1			
## 1	Mean	MelbP	Mean	Spam	Diff	P.v.	Perm	P.v.	t-test	Mean	MelbP
## 2		PTQ		PTQ		PTQ		PTQ			PTE
## 3	Q6	0.7277778	0.6344086	0.09336918		0.084		0.6860259		0.06610254	
## 4	Q8	0.7361111	0.655914	0.08019713		0.109		0.03625119		0.06648805	
## 5	Q15	0.9125	0.875	0.0375		0.205		0.06208776		0.1118238	
## 6	Q16	0.7541667	0.6532258	0.1009409		0.128		0.9455191		0.04427506	
##	mean_class.1		dif.1	pv_perm.1		pv_ttest.1					
## 1	Mean	Spam		Diff	P.v.	Perm	P.v.	t-test			
## 2		PTE		PTE		PTE		PTE			
## 3	0.05770137		0.008401174		0.053		0.9908918				
## 4	0.06706914		-0.0005810942		0.53		0.1343112				
## 5	0.1209734		-0.009149623		0.856		0.290243				
## 6	0.03987538		0.004399678		0.215		0.2063405				

Next we examined the statistical significance of the mean differences between the performances of the two group of students for Q6, Q8, Q15 and Q16.

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

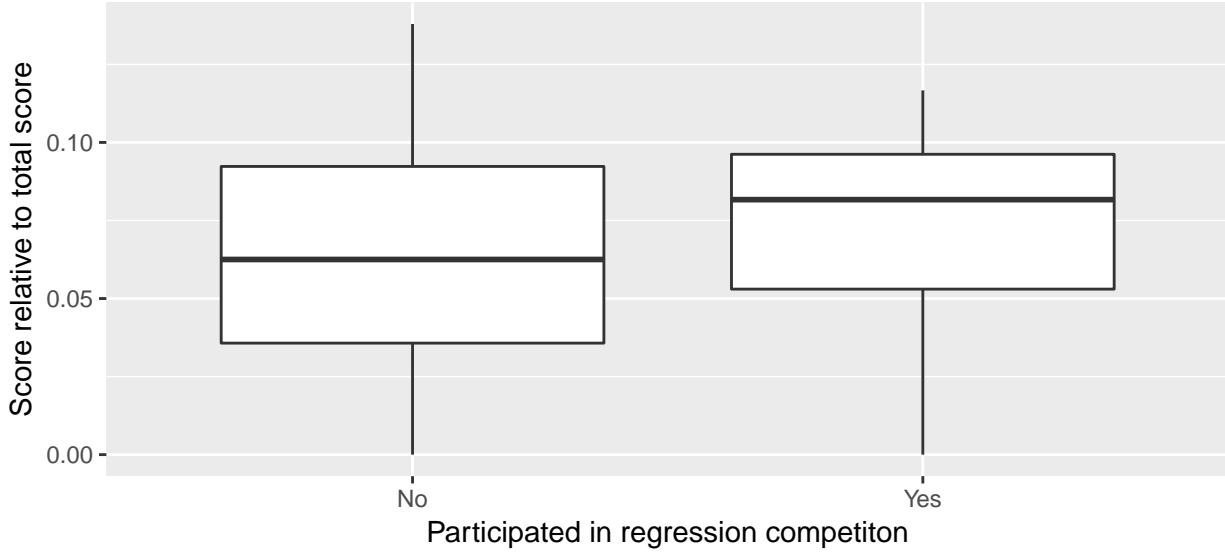


Figure 4: Performance for regression question relative to total exam score for students who did and didn't do the regression data competition in the Statistical Thinking course at Monash.

#### 4.1.2 Monash students

Figure 4 shows the results for students in the Statistical Thinking course at Monash University. Only the post-graduate students participated in the regression competition, as their additional assessment requirement. Scores for the question on regression in the final exam were compared with the total exam score (PTE defined earlier). The boxplots suggest that the students who participated in the challenge performed relatively better on the regression question than expected given their total exam performance: the median is higher and there is less variability.

On average the students who participated in the kaggle challenge scored 0.0090629 higher than those that didn't, 0.0720037 in comparison to 0.0629408. Using a permutation  $t$ -test, this corresponds to a significant difference in means, with  $p$ -value of 0.078. A conventional  $t$ -test, yields similar results with a  $p$ -value of 0.0797066.

## 4.2 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 5. Once again we exam the performances based on two normalizations. Once normalizing by the total possible marks for the relevant cluster of questions ( $PTQ$ ) and once normalizing by total exam marks ( $PTE$ ). 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.

In Figure 5 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 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.

## 4.3 Interest

# 5 Discussion

# 6 Acknowledgments

This project (title: Effect of Data Competition on Learning Experience) has been approved by the Faculty of Science Human Ethics Advisory Group University of Melbourne (Ethics ID: 1749858.1 on September the 4th, 2017) and by XXX Ethics Advisory Group Monash

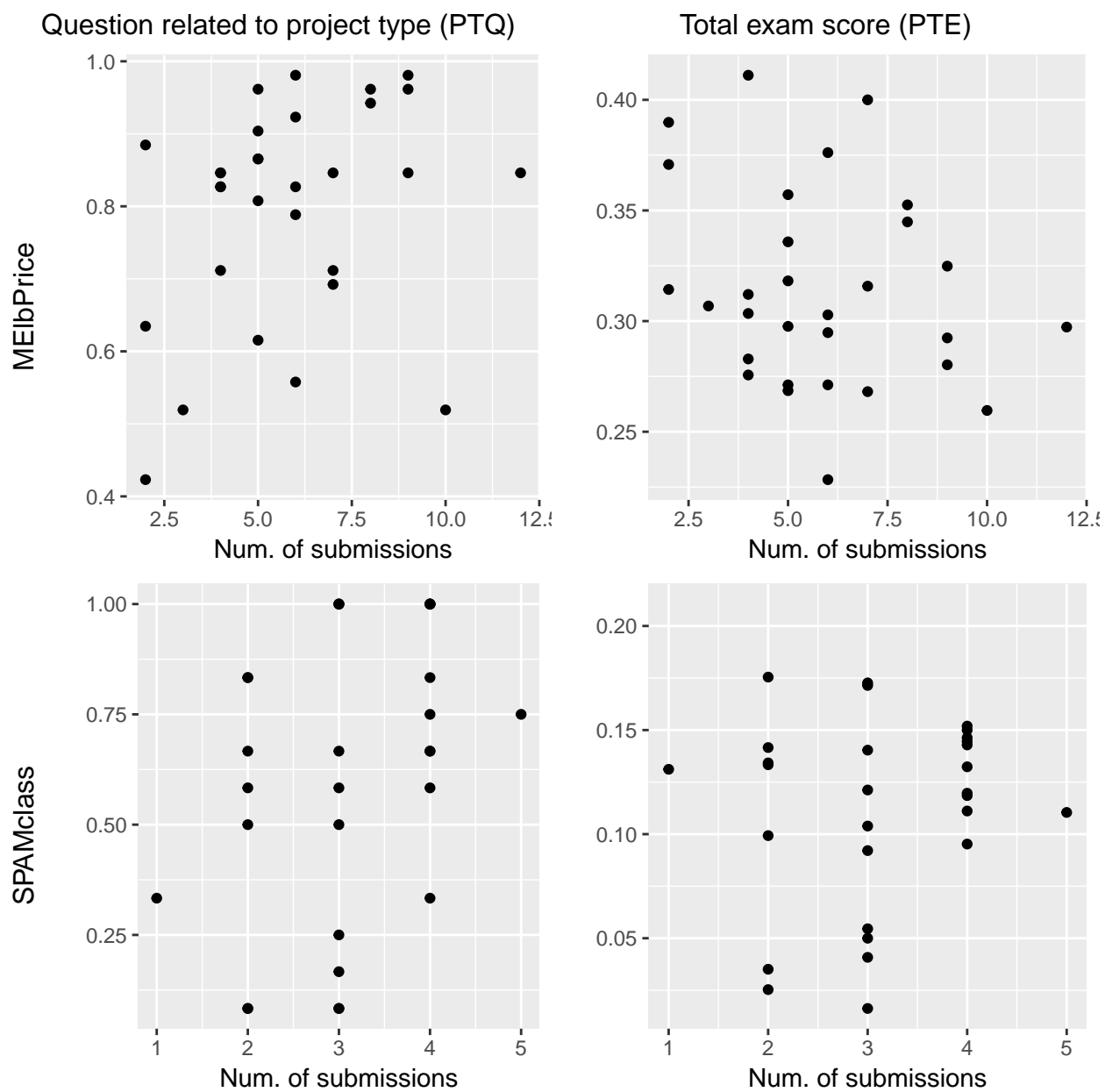


Figure 5: Number of submission vs exam grades

University (Ethics ID: XXX on XXX, 2017).

## 7 References

## 8 Stuff to hang onto for now

## 9 Verifications

This section will be just long enough to illustrate what a full page of text looks like, for margins and spacing.

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

## 10 Appendix

There were two data sets used for competitions, one was a regression problem and the other a classification problem.

### 10.1 Melbourne price data set

This data contained records of auction prices for residential properties in Melbourne. The data was compiled from auction reports collected between Feb 2, 2013 and Dec 17, 2016.

Auction reports were published weekly by Domain and compiled by Home Price Guide (R) *COMMENT: Found the symbol in pdf format.* To read them into R we used the XXX package *COMMENT: reference to our new package for reading those reports.* The reports contained the following information about the properties sold during the week: 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 also removed all the records without the price. Over all we had 75,367 sold properties for the data competition.

To enrich this data, to make it more realistic we added several additional features, number of visitors, the average rating given to the property by the visitors, number of car spaces, number of baths, land size and house size. It is possible that some of this information might have been extracted from other web sites, given that we had the full address of the property, but the time required to research that was too daunting to tackle for the upcoming teaching period.

## 10.2 Spam data set

## 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.