Report on MOOC dataset

Before attempting to answer either of the questions it is necessary to assess what types of variables we are dealing with within the dataset:

Nominal	**Ordinal**	**Interval**	**Ratio**
StudentID	Age group, degree classification	Registration, attendance, exams 1-4, coursework 1-8	VLE clicks, online engagement, raw final mark, moderated mark

StudentID, age groups, degree classification, registration, exams 1-4, coursework 1-8, VLE clicks, attendance and moderated marks are considered discrete data as they only take on specific values and most are finite, whereas raw final marks are continuous data as they can take on any value within a range.

Hypotheses

1. Registering early shows that the student is keen and performs well.
2. There is a difference in degree classifications across age groups.

Let us examine the first hypothesis by determining which values are needed and how they can be broken down. As registration is, according to hypothesis 1, the main factor that influences the keenness and performance of the students, then this will be our independent or control variable.

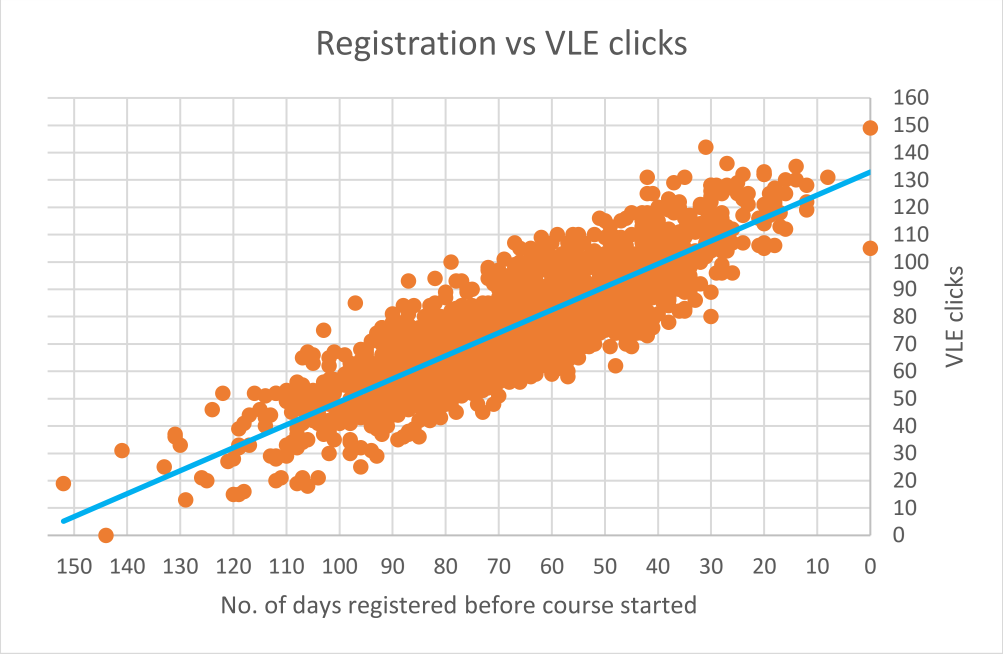
Keenness and performance are the variables which we are measuring and so these are our dependent variables. Regarding performance, I decided to choose raw final marks as the dependent variable and there are several reasons for this. Firstly, the exam and coursework marks are individual measures and it isn’t clear if these are percentages or aggregates, nor is it clear whether one exam is worth more than another. Therefore, I thought it best to exclude these values from our study as we already have a totalled average of all these marks in the raw final mark and the moderated mark. Degree classification is also excluded as, being an ordinal variable, it lacks even more precision than either the raw final mark or the moderated mark.

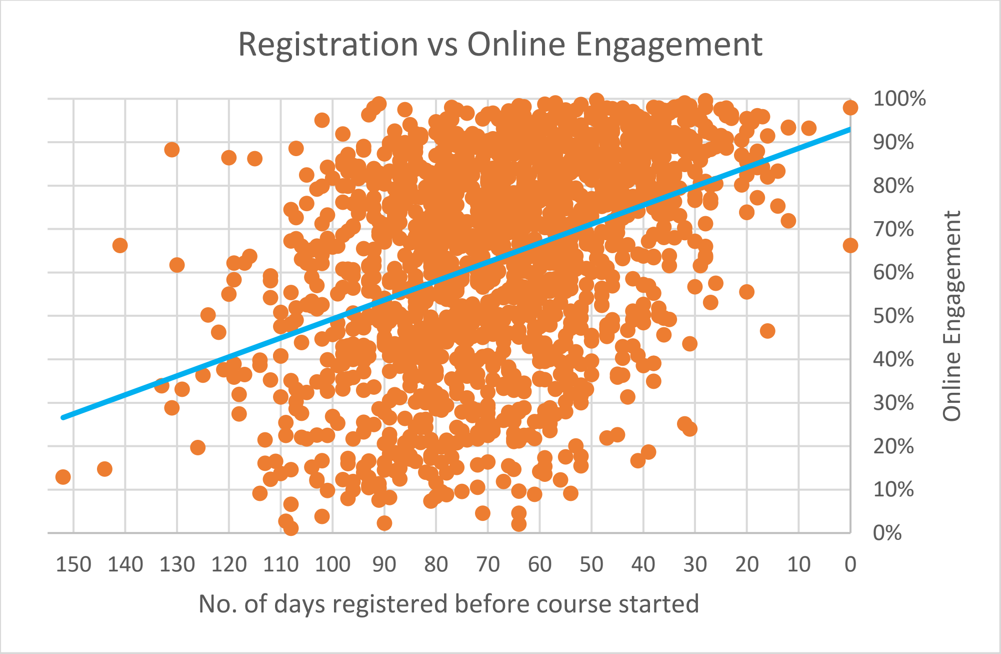
The moderated marks however present other challenges: these are approximates used for calculating the degree classification through grade boundaries and result in some students being given the same moderated mark even though their raw final mark is different. See example below:

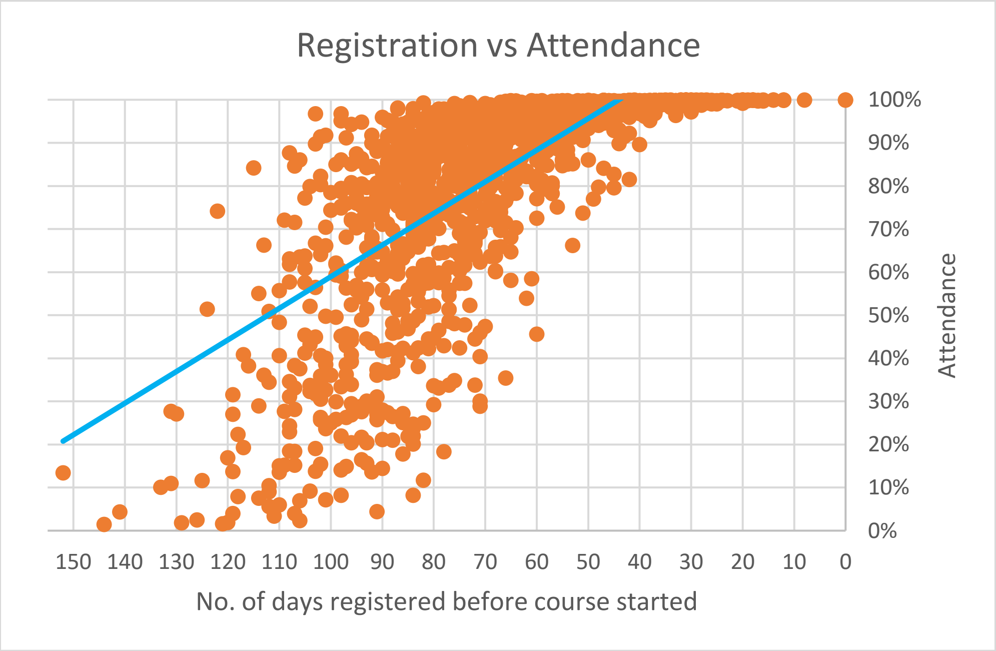
StudentID	Raw Final Mark	Moderated Mark
1630090	73.575	72
9130032	72.575	72
1810151	72.375	72

We can see that the 3 students above would be given the same moderated mark which would affect the results of our study. The raw final mark is therefore the best measure of overall performance of a student for this type of study.

Next, let us look at keenness. This is more difficult to define, as we are presented with 3 possible variables we can use: VLE clicks, Online Engagement and Attendance. I decided to see if there was any clear pattern between when students registered for the course and these 3 criteria:


Registration vs VLE clicks
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
VLE clicks
150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0
No. of days registered before course started


Registration vs Online Engagement
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Online Engagement
150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0
No. of days registered before course started


Registration vs Attendance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Attendance
150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0
No. of days registered before course started

From observation, it appears that there is little correlation between registration and online engagement, possibly some correlation between registration and attendance and a strong, negative correlation between registration and VLE clicks. We also see by analysing the z-values that registration is normally distributed, with 67.6% of the values falling within one standard deviation of the mean, which is very close to the theoretical value of 66.8%. The same applies to the raw final mark (68.2%) and VLE clicks (68.3%) whereas online engagement and attendance, at 63.7% and 21.7% respectively, are not. When examining the skew and kurtosis of these criteria, we also see that the skew is close to zero and kurtosis is low (around 3) for registration (0.104, 3.01), VLE clicks (0.109, 3.04) and raw final mark (-0.329, 2.97) respectively. For online engagement and attendance, we see however that there is a shift away from normal distribution and so I conducted a 1-sample Kolmogorov-Smirnov and Shapiro-Wilk test on all listed variables above to test their normality, the results of which can be seen below:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
VLEClicks	.031	1624	<.001	.998	1624	.013
Raw Final Mark	.042	1624	<.001	.992	1624	<.001
Registration	.018	1624	.200*	.998	1624	.146
Online Engagement	.083	1624	<.001	.952	1624	<.001
Attendance	.229	1624	<.001	.736	1624	<.001

*. This is a lower bound of the true significance.

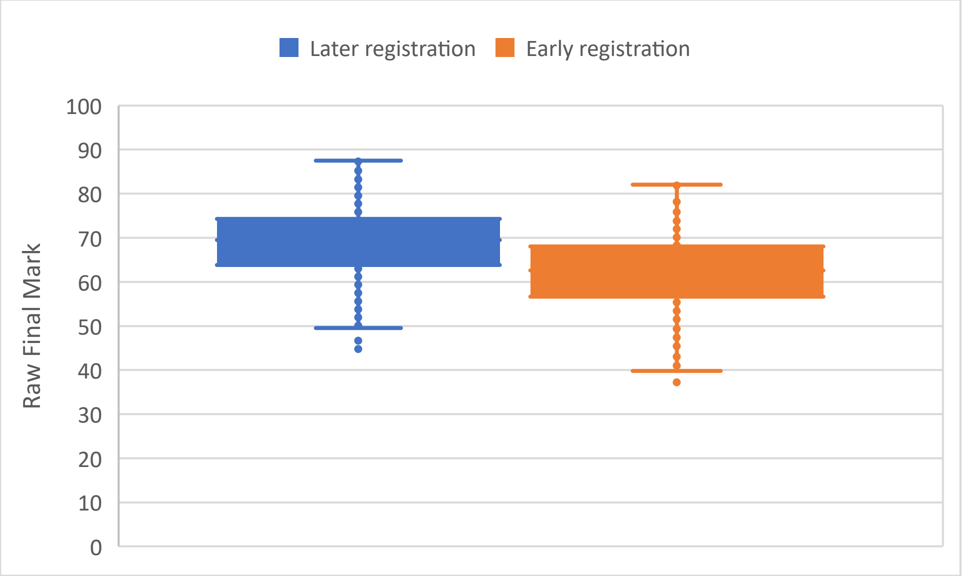
a. Lilliefors Significance Correction

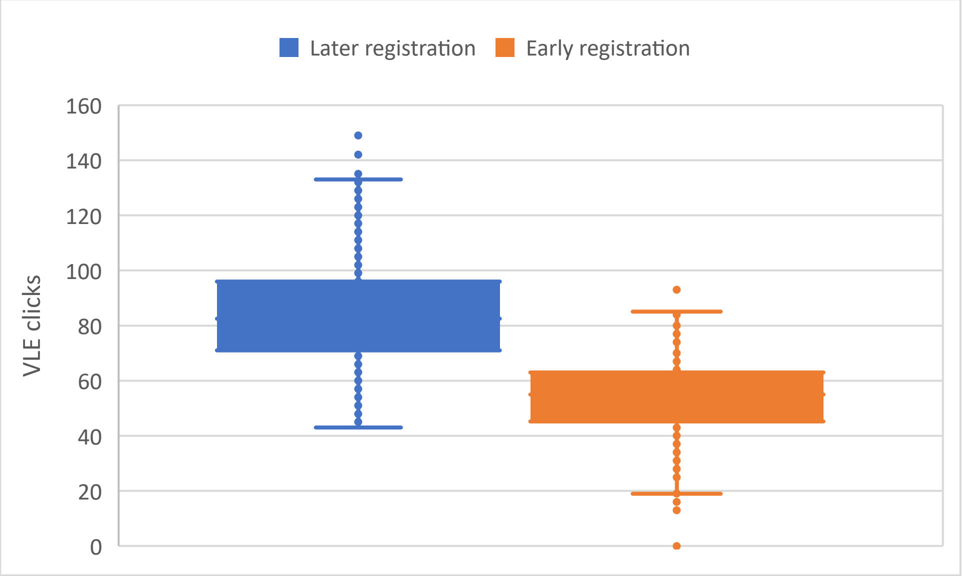
As expected, with p-values of more than 0.05, the more sensitive Shapiro-Wilk test shows that both VLE clicks and registration follow a normal pattern of distribution, whereas online engagement and attendance both appear to be non-parametric data. More surprisingly though, was that the raw final mark does not pass the test for normality. Nonetheless, after running both a Pearson correlation and a Spearman correlation in SPSS I discovered some interesting results among the association between the different variables. The strongest correlations were between registration and VLE clicks at -0.853 using the Pearson Correlation, and registration and attendance at -0.827 using the Spearman Correlation. Both were strong, negative correlations with p-values below 0.01 which recommend that we should reject the null hypothesis and we can perhaps interpret that the later students register, the better their attendance and the keener they are at engaging with the course in terms of their VLE clicks.

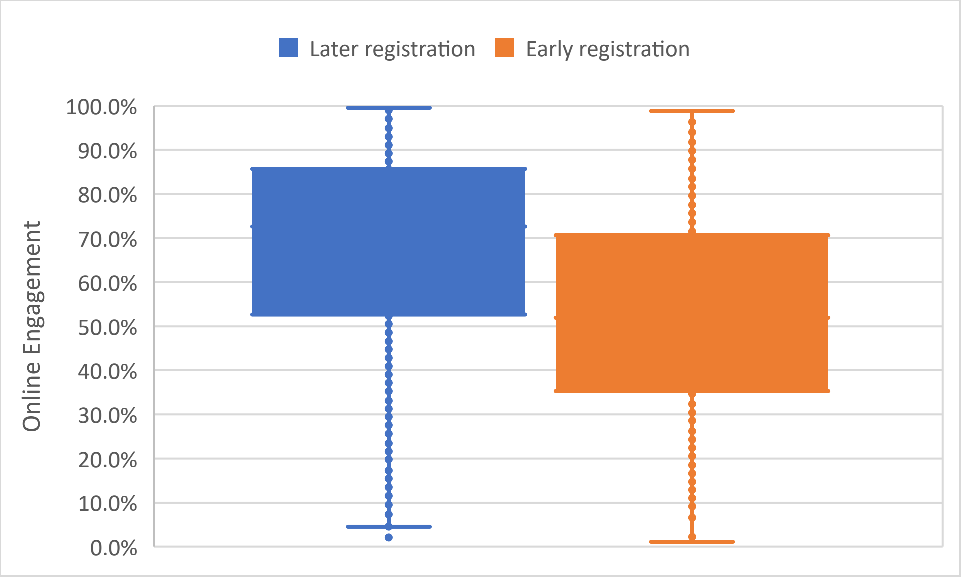
Correlation coefficients for the other variables have weaker values of between 0.358 to 0.463 and -0.408 to -0.455. Again, all of which had p-values of less than 0.01 which assumes that there is an association between each of the variables. A stepwise, multivariate regression analysis also revealed that each of the variables has some effect on the other, particularly for the raw final mark, which is the dependent variable I chose to measure performance. Registration, when measured alone against the raw final mark, had a coefficient of -0.455 which suggests that it has a direct, negative effect on the performance of the students. However, when VLE clicks, online engagement and attendance were inputted into the model, they each gave a coefficient average of 0.167 while the registration dropped to -0.135. All significance testing resulted in p-values of 0.02 or lower. This shows that the measures of keenness are positively affecting the raw final mark, and so it is possible that how early the student registered for the course is affecting their keenness which then, in turn affects their performance.

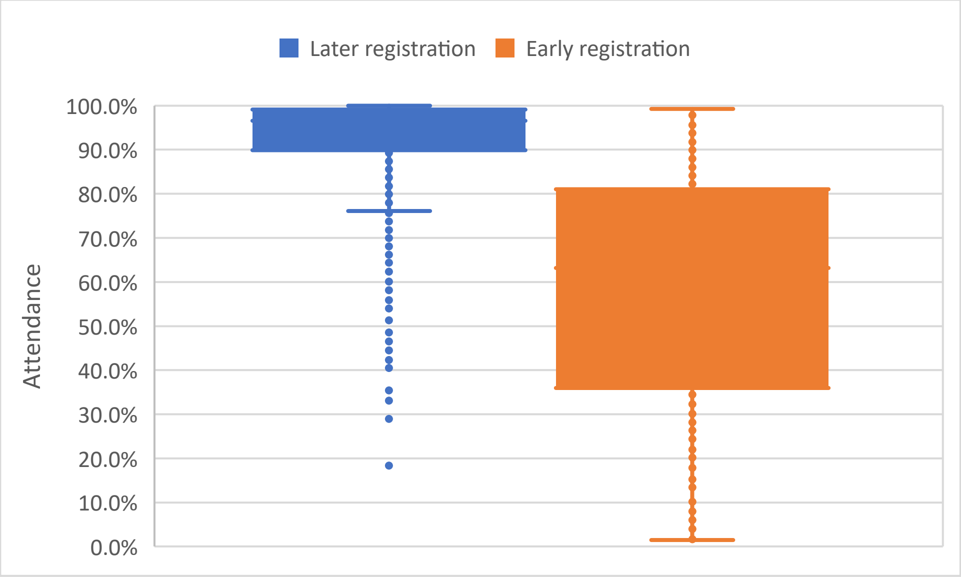
Having shown that there are some strong, negative correlations between registration & keenness and registration & performance, and that keenness itself also positively correlates with performance, we can reject the null hypothesis that there is no correlation between when a student registers for the course, how keen they are and how well they perform at the end of the course. However, the alternative hypothesis states that: ‘registering early shows that the student is keen and performs well’. It is therefore important to define what ‘early’ is in numerical terms. After calculating the interquartile range for registration, I found that the upper half, or Q3 gives a value of 82. Any values that are 82 or higher I will thus define as early registration. I settled on this measure of scale as I found the others to be unsuited for the task. The mean and median for example, give values of 66.9 and 67 respectively, which represent an average of the registration time. However, the hypothesis requires a definition of early and later registration, and I believe that choosing a location which is in the middle of the dataset will ignore what might be considered the average registration time for students. Quartiles on the other hand better represent the students who tend towards the outliers that registered particularly early. Finally, the standard deviation and the range weren’t used as they are badly affected by outliers, especially in this case where there are a small number of values at 0.

Using ≥82 as the definition of early registration, we can see how this compares to the students that registered later regarding raw final marks, VLE clicks, online engagement and attendance:



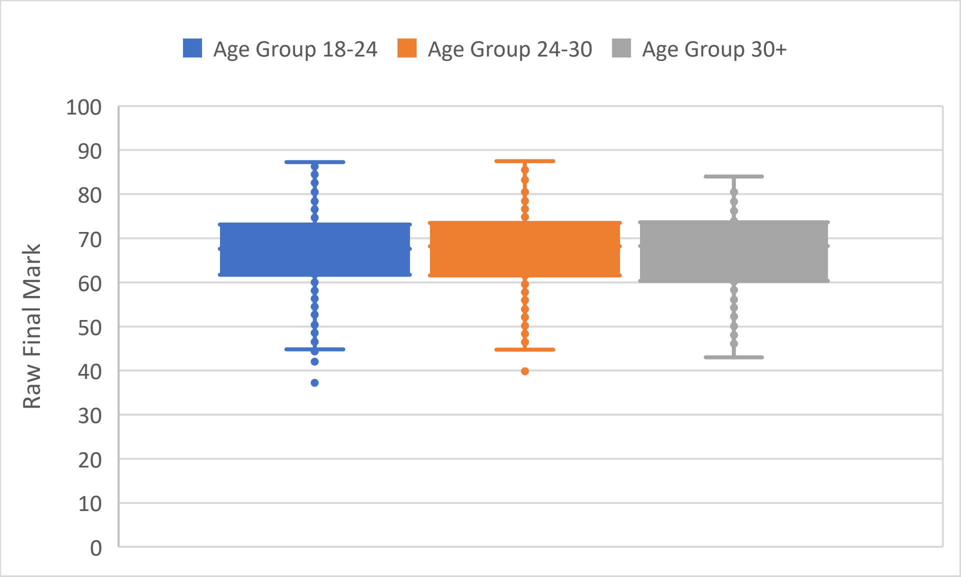






The box and whisker plots show that there is a large amount of dispersion when comparing registration to online engagement and attendance. As commented on earlier when I highlighted the higher skew and rather high kurtosis of both variables, in addition to them being non-parametric data, there are a large number of outliers which suggests that they may not be the most reliable means of rejecting the null hypothesis, even if they appear to do so. Furthermore, there are many factors which contribute to students' attendance rate and this is perhaps not the most accurate way of showing keenness. Instead, after performing a t-test of both equal and unequal variances on VLE clicks and raw final mark, which returned p-values less than 0.00000…, we can reject the null hypothesis and say that there is a difference in both keenness (VLE clicks) and performance (raw final mark) between those who registered early versus those who registered later. In addition, the correlation analysis suggested that the alternative hypothesis could also be incorrect and that the reverse may be occurring: registering later actually shows that students are keen and perform well, while registering early shows the opposite.

Finally, it could also be possible that there is another independent variable that is affecting student performance and keenness – age. However, when I compared the raw final mark across age groups there does not appear to be any significant variation:



This is also due to age being sorted into groups instead of individual values, which results in 3 ordinal variables that don't offer much insight. Rather, it is better to establish if age is a factor by analysing our second hypothesis. This states that 'there is a difference in degree classifications across age groups.'

After counting the frequency of students by their age group and what degree classification they received, we can see the results below:

	Degree Classification			
Age Group	**1st**	**2:1**	**2:2**	**Total**
18-24	366	421	176	963
24-30	197	197	101	495
30+	69	59	38	166
Total	632	677	315	**1624**

By using a χ² test, we can assess whether or not there is an association between being in a certain age group and being awarded a certain degree classification. I created this χ² distribution table below to calculate if this was the case:

	Expected values			
Age Group	**1st**	**2:1**	**2:2**	**Total**
18-24	374.76	401.45	186.79	0.59
24-30	192.64	206.35	96.01	0.30
30+	64.60	69.20	32.20	0.10
Total	0.39	0.42	0.19	1

After performing a χ² test on the actual values in the yellow table and the expected values in the pink table, the test returned a p-value of 0.247689775. As this is higher than 0.05, this suggests that we cannot reject the null hypothesis and we can conclude that there is likely no association between age group and degree classification, as was seen in the box and whisker plot above.

Further study

It would be useful to have a better breakdown of the age groups as this would allow more detailed analysis of the association between it and other variables such as degree classification. Furthermore, regarding hypothesis 1, a survey to see why students register when they do, could help explain why there is a negative correlation between the date of registration and keenness/performance.