# Group 10 CSC8633: OULAD Investigation

Annie Sames, Junzhe Zhao, Yimiao Wang, Mayank Baraksar, Rosemary Finnegan, Avi Gupta

22/03/2022

## Abstract

*Context:*

The Learning and Teaching Development Service at Newcastle University wants to be able to help students gain greater autonomy to manage their own learning experience to achieve their maximum potential. Learning analytics provide insights to institutions about the behaviour of their students, and the interactions that they have with course material. Institutions such as universities are then able to alter courses to better the students' learning experience, and ultimately improve their grades.

*Objective:*

This project will use the OULAD dataset which contains information about 22 courses on the Open University run in 2013 and 2014, along with the interactions that the students enrolled on these courses had with the VLE. Multiple problems will be investigated, including the effect of the module lengths on the performance and engagement of students, the withdrawal rates and the behaviour of the students who withdraw, the effect of the time of submission on result and the overarching relationship between interaction and score.

*Method:*

A variety of techniques will be used to explore these problems, but predominantly, exploratory data analysis will be used. Correlation tests, such as Pearson and t-tests will be used to test the significance of the differences and relationships among these variables. A predictive model will then be created that allows the user to predict the student's final results based on their interaction,

*Results:*

This research found that there is an overarching relationship between score and interactions, measured using the clicks. It has also found that when the activities mainly consist of 'resources', students are more likely to pass or get a distinction, compared to when there is more 'outcontent' as this is when students are most likely to fail. It has also been found that students are most likely to withdraw nearer the beginning of the courses, and the interactions that the students have with the VLE seems to have a significant effect on whether or not a student will withdraw. This research also found that module FFF had the most interactions. The interaction with the VLE fluctuates throughout each of the courses significantly. The module with the largest number of clicks was the module FFF2014J, and the module with the smallest number of clicks was the module GGG2014J. In terms of the dates when assessments are submitted, the later submissions seemed to obtain lower results in their assessments.

*Novelty:*

This project is a more data analysis focused project which with comparison to previous research, has not been done many times. Being able to give concrete findings to institutions, such as the LTDS, will benefit them as they will be able to use this information to make the relevant changes to their course.

# 1. Business Understanding

The Open University is one of the largest online learning universities in the world, with around 170,000 students currently registered on various programmes (Kuzilek, Hlosta & Zdrahal, 2017). Material is taught to students via the Virtual Learning Environment (VLE), and they are assessed in many ways, for example online quizzes. Analysing the data from these programmes is essential, as it allows institutions to improve their courses, increase student engagement, and help students to get the most out of their modules. This process is known as learning analytics.

## 1.1 Learning Analytics

The aim of learning analytics is to collect and analyse learners' data, to improve overall learning experience. This allows institutions to better teach students and optimise the learning materials that they offer (Clow, 2013). Learning analytics usually starts with the collection of data on student engagement, and once this is done, it can be broken down and analysed to highlight any patterns in the data. This helps institutions, such as online universities, to know which courses have the highest levels of student engagement, and the possible reasons for this, ultimately assisting them in achieving their goal of improving overall student experience. In some cases, this data can also be used to carry out predictive analytics by making models to predict future trends in the behaviour of the students or their results.

## 1.2 Previous Literature

This dataset has been the subject of extensive research into learning analytics, and one example of this is the research carried out by Kuzilek et al in 2018 who used Markov chain modelling to identify behavioural patterns leading to drop-out or withdrawal. They found many interesting patterns in the behaviour of the students with the VLE, particularly with the submission of the first assessments for each of the modules. The proportion of students withdrawing from their courses when they had had no interactions with the VLE, was twice the amount of the number of students with at least some interactions with the VLE. In terms of the submission of the first assessment, this research found that the submission of the first assessment for a module was a good predictor of the student's final grade. Kuzilek et al found that those who submitted the first assessments also tended not to withdraw from the courses if they had some interaction with the VLE. This is compared to the students who did not submit the first assessment, but still interacted with the VLE, as these students have a probability of withdrawing of 0.25. The findings from this research motivated the research undertaken in this project focusing on the students who withdraw and their scores and interaction with the VLE. These findings are consistent with those of Hlosta, Zdrahal and Zendulka (2017) who found that students who do not submit assignments have a 90% chance of dropping out of a course.

In addition to the research by Kuzilek et al (2018), research by Hong, Wei and Yang (2017) found that the number of students who withdraw from courses heavily outweighs the number of students who submit assignments. They used three two-layer machine classifiers- support vector machines, random forest, and logistic regression to create predictive models for the drop-out rates of students enrolled on the courses and found that their predictive models had a precision score of 97%.

To help guide the aim of this research, the methods of some other previous learning analytics research were studied, included that of Haiyang et al (2018). This research used time series classification based on the students' behaviour and activities with the VLE. They also used the time series forest classification algorithm to propose a predictive model for the students dropping out. Time series analysis was also used in a study by Chen and Wu (2020) who created a predictive model for student dropout by combining neural networks and long short-term memory. Also, it has been found that the length of time of study affects the performance of students, as those who have enrol on longer courses tend to perform better (Ukpong & George, 2013).

### 1.3 Problem Statement

The aim of this investigation was to provide the Learning and Teaching Development Service (LTDS) at Newcastle University with useful insights into how they can help students to gain greater autonomy to manage their own learning experience. This will allow students studying at Newcastle University to seek the maximum benefits from the university services and it will also enhance student engagement and achievement. Being able to pinpoint what courses and at what people engage more, drop out and perform better is useful information for institutions such as universities. Having access to this information will allow Newcastle University to optimise the engagement of their students across all their courses, as they will be able to make appropriate improvements, and increase their performance.

### 1.4 Project Aims

The aims of this research are to provide an insight into the behaviour of students enrolled on these courses, and the analysis has been broken down into several areas. The investigation will first analyse the number of students enrolled on each of the courses, as well as the performance in each of these. The effect of the module length on the scores and the interaction with the VLE will also be investigated. Next, the behaviour of the students withdrawing from the course will be analysed, along with the relationship between the clicks and score. The project will then go onto to look at the effect of submission dates on the student's performance. This analysis will contribute to the creation of a predictive model that can be used to predict the student withdrawals and the student performance. A conclusion will then be made as to what seems to have the most significant effect on student performance and engagement.

### 1.5 Project Plan

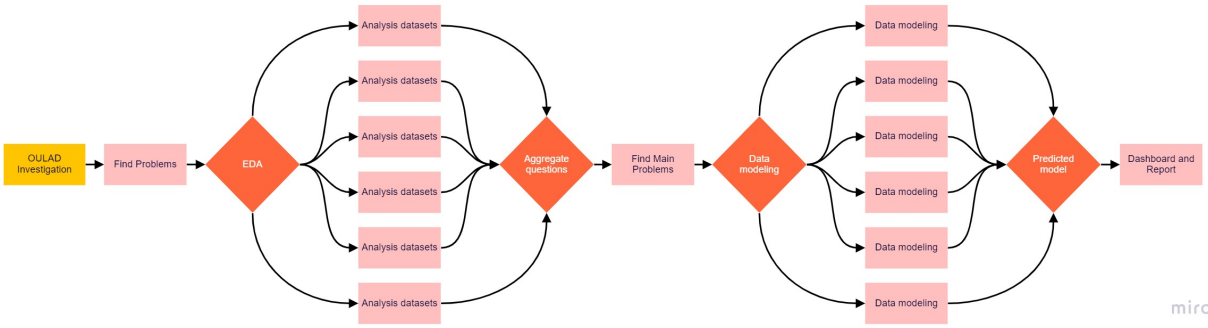


Figure 1: Flow chart of project plan

The CRISP-DM framework was implemented in this project (see Appendix A) in order to design a reproducible data analysis pipeline for the dataset. Figure 1 shows the project plan. Preliminary exploratory data analysis was carried out, and then the core issues were identified. The analysis was then broken up amongst the project team and the mathematical and statistical analysis was carried out. The project management on Github also adopts this approach, starting with a question in the main branch, then dividing the issues into smaller six branches. The data analysis and modelling from each of the branches was merged to compile a final project report.

## 2. Data Understanding

The dataset that will be used in this research is the Open University Learning Analytics Dataset (OULAD), which contains data on 22 courses and the 32,593 students who were enrolled on these (Kuzilek, Hlosta &

Zdrahal, 2017). As well as this, the dataset contains information on the 10,655,280 interactions that were had between the students and the Virtual Learning Environment (VLE) across the courses that were run in February and October of 2013 and 2014. The typical length of presentations is around nine months. Students are able to register for modules from a few months before the module starts up to two weeks before, and once registered on the module, students have access to the relevant resources. Students are also able to withdraw from modules throughout the duration of the course. The content of the course is taught using the VLE with a variety of activities, including 'dataplus', 'oucollaborate' and 'URL'. The student's performance is measured via several assessments, and at the end of the module there is usually a final exam. Some of the data that has been provided includes the student's scores for each assessment, the number of interactions the student has with the material in each day and the number of students who withdraw from each of the courses. In the case of this research, the demographic information has been disregarded to avoid ethical issues. An in-depth description of the dataset can be found in appendix B.

# 3. Data Preparation

According to the business plan, four data frames were prepared to answer questions posed (the other two are shown in the appendix C):

1. Exploring the relationship between final result, assessment score and sum clicks:

(1) In the 'studentVle' file, the columns of `code_module`, `code_presentation`, `id_student`, and `sum_click` were selected, and the first three columns are grouped, based on the count of total number of clicks for each student participating in the course.
(2) In the 'studentAssessment' file, students with `is_banked` of 1 are removed to ensure no grade migration. This was merged with the assessments and the weighted scores were calculated. The rows containing 'NA' values were replaced with '0'. The total assessment scores were calculated for each of the students, based on the student ID numbers.
(3) The `code_module`, `code_presentation`, `id_student`, and `final_result` columns from the 'studentInfo' were selected and and the assessment score and clicks were calculated for each student.

Table 1 shows the first eight rows of the data frame. There are 32,593 records and 6 variables.

| code_module | code_presentation | id_student | final_result | final_assessment | sum_click |
|---|---|---|---|---|---|
| AAA | 2013J | 11391 | Pass | 82.4 | 934 |
| AAA | 2013J | 28400 | Pass | 65.4 | 1435 |
| AAA | 2013J | 30268 | Withdrawn | 0 | 281 |
| AAA | 2013J | 31604 | Pass | 76.3 | 2158 |
| AAA | 2013J | 38053 | Pass | 55 | 1034 |
| AAA | 2013J | 38053 | Pass | 66.9 | 2445 |
| AAA | 2013J | 38053 | Pass | 67.8 | 1492 |
| AAA | 2013J | 38053 | Pass | 72.5 | 1428 |

Table 1: Student final result with assessments and clicks

2. Exploring the relationship between submission date, assessment score and final result: One main dataframe was created and modified as needed during analysis to analyse the relationship between submission date and assessment score.

(1) In the assessment table, the columns `id assessment`, `id student`, `date submitted`, and `score` were chosen to search for correlations.
(2) The assessment dataset was then merged with courses to determine the deadline for each assessment based on the `id assessment`.

(3) To obtain the final result, the calculated dataframe was combined with `studentInfo` dataset. However, no demographic variables used.
(4) A new variable `date_submitted_perc` was calculated to bring different deadlines of assessment on same scale by calculating the percentage of `date_submitted` which was submission date of student by the deadline of that particular assessment `date` from courses dataset.
(5) Another new variable `submissionTypeLabel` was created as a category variable with two possible values of `late` and `not late`.
(6) Also, `assessment_final_result` was also a categorical variable that was calculated based on assessment score where the possible values were `pass` and `fail`

Table 2 shows the first three rows and key columns.There are 10655280 records and 10 variables. This table was generated after merging assessment, courses and studentInfo assessment.

| id_assessment | id_student | date_submitted | score | code_modules | date | submissionType | AssessmentResult |
|---|---|---|---|---|---|---|---|
| 37443 | 643435 | 250 | 78 | AAA | 251 | Late | Pass |
| 1752 | 11391 | 18 | 39 | BBB | 19 | Not Late | Fail |
| 14993 | 420604 | 25 | 90 | FFF | 23 | Note Late | Pass |

Table 2: Dataset required for time series analysis

# 4. Data Analysis

## 4.1 Preliminary data analysis

### 4.1.1 Student Enrolment

To explore the number of students enrolled onto each module in each presentation, a data frame containing this information was first made. This information was then plotted on the graph shown in figure 2. From figure 2, it is evident that the module with the highest number of students enrolled onto it was CCC in 2014B with 2498 students and the course with the least number of students enrolled onto it is AAA starting in October 2014. This provides vital information to the Open University as it is clear which courses are more popular than others and may provide them some indication as to what courses they might want to advertise more. It is also interesting to note from figure 2 that the courses in 2014 have a lot more students enrolled on them compared to the courses in 2013. The 2014 courses had an average of 1431 students enrolled, compared to 2013 which had an average of 758 students enrolled. A t-test was conducted to compare these groups, which revealed a significant difference between the number of students enrolled in 2013 against 2014, with a p-value of 0.02519.

From figure 2, it is also clear that the courses starting in January are less popular than the courses starting in February. Again, this is seen with the February courses having an average of 1059 students enrolled compared to the October courses having an average of 1130.714 students enrolled. However, a t-test indicated that this difference was not significant as it generated a p-value of 0.8372. This indicates that these differences are likely due to chance, and there is not an actual difference between the enrolments at each period of the year. This will later be explored in context to the withdrawals.

### 4.1.2 Student Performance

To understand the spread of results of the students in each module and presentation, the graph in figure 3 was plotted. The total number of students achieving a distinction, pass, fail and those withdrawing have been shown in table 3. This table demonstrates that 9.28% of students achieve a distinction, 37.93% of students pass, 21.64% of students fail and 31.16% of students withdraw. It is evident that the course with the highest percentage of students passing or achieving a distinction was AAA starting in October 2013,

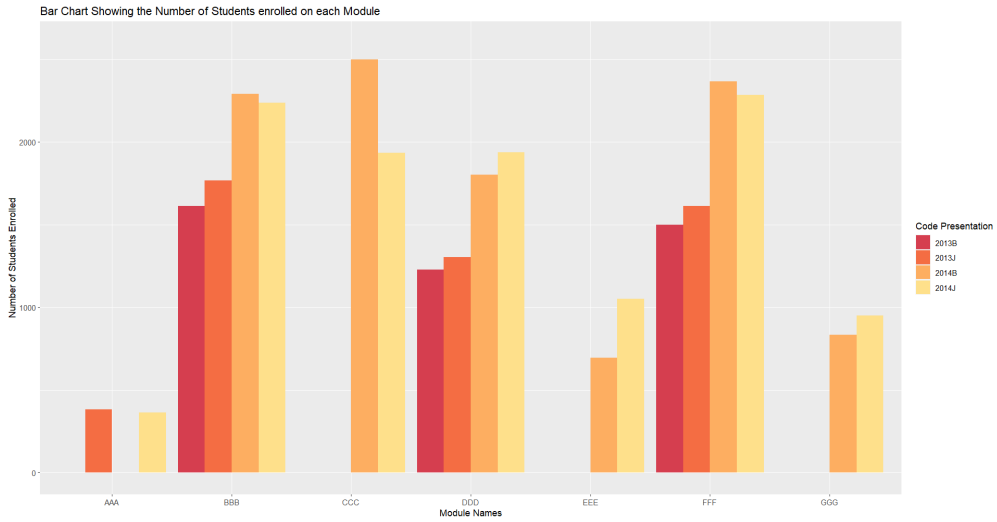


Figure 2: Graph showing the number of students enrolled on each module

with nearly 75% of the enrolled students passing or getting a distinction. This can be compared to the results from module CCC, starting in February 2014, where only around 30% of people achieved a pass or distinction. Looking at distinctions specifically, the course with the highest percentage of students achieving a distinction is GGG, and specifically the code presentation '2014J'. Figure 3 also illustrates that the course with the highest percentage of people dropping out is CCC, starting in February 2014, and the course with the lowest rates of drop out is GGG across all three presentation codes, but in particular '2013J'. There may be several reasons for these differences, including course difficulty, assessment type, or even the people enrolled on the course, however, something which can be analysed is the activity types for each of the courses.

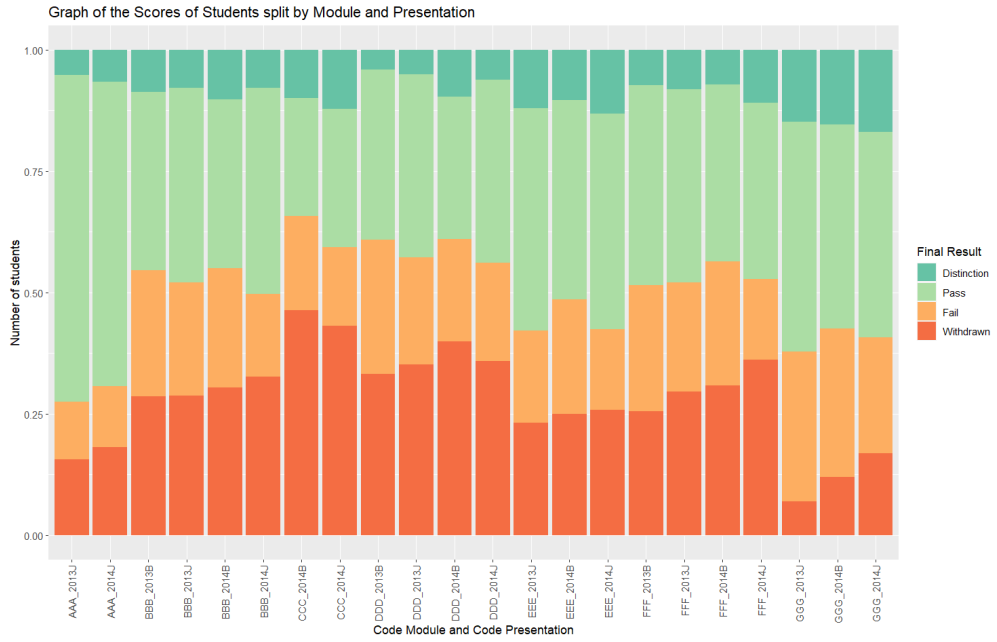


Figure 3: Graph showing the percentage of students' final result on each module

As shown in figure 3, the module with the highest percentage of people getting a pass and distinction (AAA in 2013J), is primarily made up of the 'resource' activity type. This may suggest that when the course

| Grade | Number of Students |
|---|---|
| Distinction | 3024 |
| Fail | 7052 |
| Pass | 12361 |
| Withdrawn | 10156 |

Table 3: Student Performance

uses the 'resource' activities more, students tend to perform better. This in contrast to the module CCC starting in 2014B which despite also being made up of a lot of 'resource' activities, also had a large quantity of 'oucontent'. This may show that this type of learning is not as effective for students, and they do not engage with it as well because when this is the case, the course had a lot more withdrawals. Looking at the courses with higher percentages of students achieving a distinction, it is clear that these modules are predominantly made up of 'resource' activities, again reinforcing the idea that the 'resource' activities may be most beneficial and engaging for students.

### 4.1.3 Students Withdrawals

A further analysis concerned the data from students withdrawing from each of the modules. It may be important for the Open University to have this data as it may help them to reduce the numbers of students withdrawing from courses. This is crucial as it will improve the overall outlook of the Open University if they reduce their withdrawal rates, as well as saving valuable resources and time. From the withdrawals data, the OU will also be able to gain insights into what potential interventions can be put into place to reduce students withdrawing from courses, by analysing when students are most likely to withdraw and the behavioural patterns of those that do.

Firstly, it was paramount to identify if there was a certain time of year that students enrolled onto these courses are most likely to withdraw from the courses. To do this, the data frame for the enrolments was subsetted by month, start date, and module code. A table for the number of people withdrawing from courses throughout the year was created, and these showed that there is a lot of variation in the withdrawals throughout the duration of the year, and indicated that some days had a lot more people withdrawing on them than others.

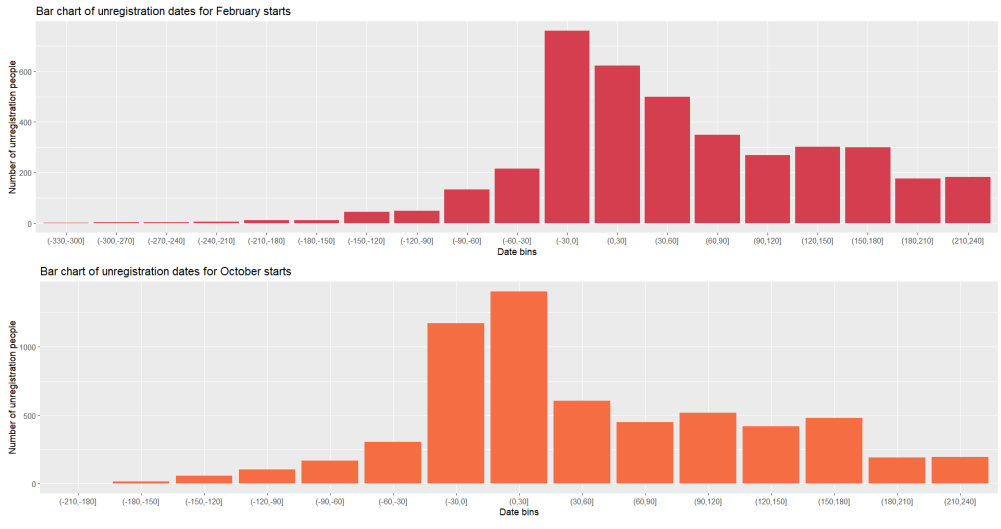


Figure 4: Bar chart of unregistration dates

Charts were plotted of the data divided into 'bins' of 30 days that approximate to one-month periods. Figure 4 for February and October starts showed a similar pattern, with many people dropping out just before, or

close to, the start date of the course and some continuing to drop out for many months after the course had started. Withdrawals for February starts tended to be slightly later than those for October starts. Figure 12(see Appendix D.1) was plotted to illustrate the number of withdrawals by day, for each start month. These showed small fluctuations from day to day, with clear 'spikes' close to day zero for both start months and a smaller increase from about day 200 for February starts.

| code module | Top five dates |
|---|---|
| AAA | 58, 128, 166, 172, 175 |
| BBB | 12, 0, 27, -7, -1 |
| CCC | 12, 0, 31, 32, -1 |
| DDD | 12, 0, -7, 27, 5 |
| EEE | 12, 0, -8, -5, -1 |
| FFF | 12, 0, 27, -1, -5 |
| GGG | 12, 25, 45, -8, -3 |

Table 4: Top five withdrawn dates in each module

The five most frequent withdrawal dates were obtained, and these have been shown in table 4. Day 0 is the module start date. In six of the seven modules (all excluding AAA), the most popular date to withdraw overall was 12. In five of the seven (all except AAA and GGG) the second most popular withdrawal date was 0. One reason for this could be that students who withdraw fewer than 14 days from the module start receive a full refund; see the OU's fee rules that were valid at the time the data was collected. Another reason could be that students started studying for a module but then realised the module was not what they expected and so they quite quickly withdrew. However, not everyone who withdrew from a module did so early; some students withdrew after hundreds of days. Module AAA seems to be an outlier with a very different pattern of student withdrawal dates as compared to the other modules. It would be interesting to examine the reasons for this if further data were collected. For modules that start in February, the five most popular withdrawal days were day 0, 12 and 27 of the course, but also seven days and the day before the course started. Similar to the courses starting in February, two of the most popular withdrawal dates for the courses starting in October were day 0 and day 12, but also day five of the course. For October courses, it was also common to see students withdraw five days and one day before the course starts.

#### 4.1.4 Student Daily Activity (Sum of Clicks)

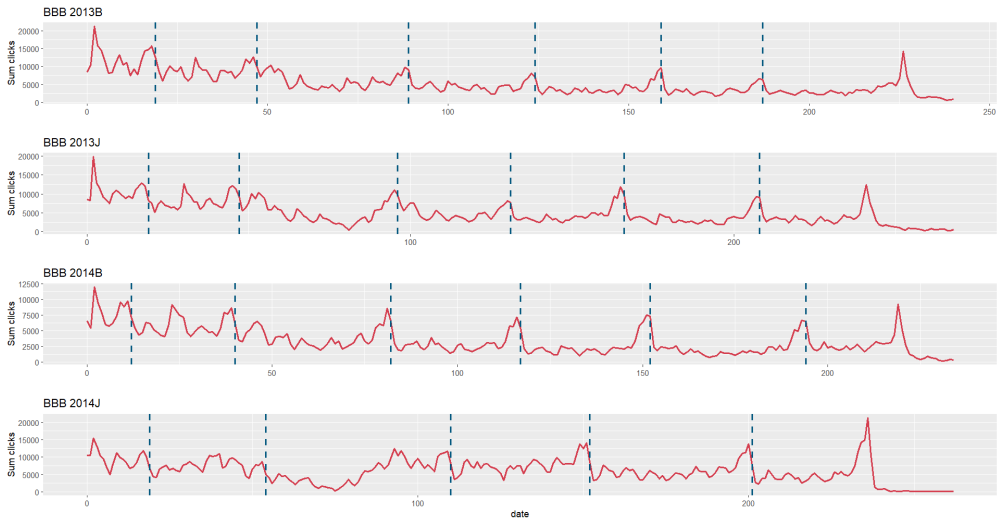


Figure 5: Time Series of Sum Clicks in Module B

The total number of clicks for each day were calculated as well as the exam dates for the TMA assessments. Figure 5 shows the change in the total number of clicks over time, and this is an illustration of how active the students are throughout the duration of each of the courses. The blue line on each of these graphs represent the assessment days in each of the modules. The daily activity level decreases as the course goes on, but appears to increase around the time of the assessments. After the assessment dates, the activity level decreases again. Figure 5 shows the assessment days for module BBB, and the daily interactions have beeen shown around the time of each assessment. In the BBB module, the code for the demo module with the maximum number of clicks is BBB2014J and the maximum number of clicks in all four modules is around the day after the module starts. In BBB2014J, the total average number of clicks is concentrated within from 5000 to 10,000, the maximum number of clicks is 21336 (date=236), and the minimum number of clicks is 58 (date=259). The module has four segments with clicks above 12,500, indicating that the module has four segments of assessment, all of which have clicks at the time of non-assessment at 100 clicks or so. For module BBB starting in October 2014, the second and third assessments have a noticable decrease in activity around the dates they are due. This is in comparison to the module BBB starting in October 2013 and February 2013 where the time between the second and third assessments is shorter, the decrease in activity around this time is not as prominant. In conclusion, the scheduling of the exam dates and the difficulty of the exams all impact student activity, and the relevance of the final student performance will be explicitly analysed through the exam results subsequently.The details of each module can be seen in the appendix D.2

#### 4.1.5 Students' preference for activity type

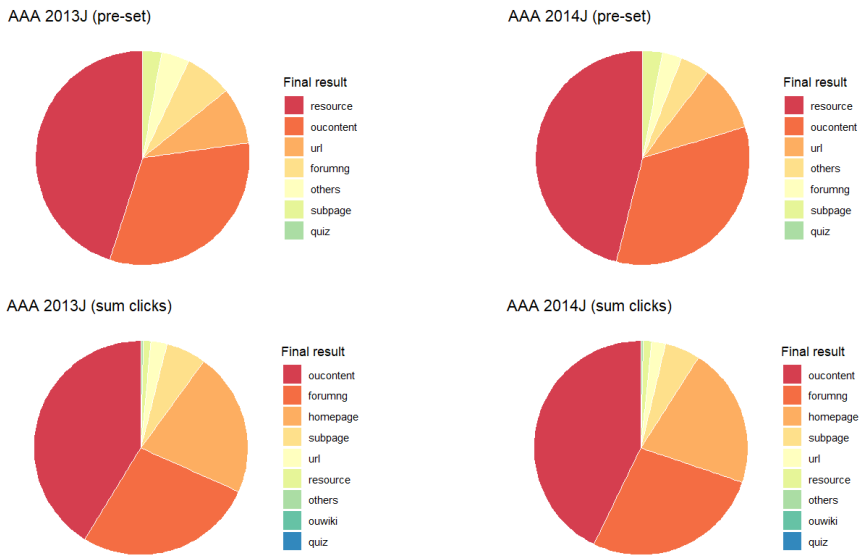


Figure 6: Pie chart of pre-set and students' preference for activity type in module A

Each course has various activity types of interactive content, and there are pre-set percentages within the course. By counting the number of times students click on the types within the course, we can calculate which activity types students prefer to interact with. Figure 6 shows the pre-set percentage of activity types and the student preference percentage for Module A. In the pre-settings, AAA 2014J has changed little compared to AAA 2013J, with a decrease in the proportion of forumng and an increase in the proportion of other types. In student preferences, students have increased the number of clicks on oucontent. In contrast to the pre-settings, students prefer oucontent, forumng, and homepage, with resource accounting for more of the pre-settings. Table 19 and 20 (Appendix D.3) shows all the profiles for all the courses. It can be used to make subsequent adjustments to the percentage of activity type to increase the motivation of student interaction.

#### 4.1.6 Students performance depending on submission date

The major goal of this investigation is to determine if students have adequate time to complete exams on this platform. After reviewing the dataset, it was discovered that the final outcome was positively connected with the assessment scores. When students performed well in assessments, they received high grades on the final test. Following the visualisation of the assessments dataset, it was discovered that 29% of students were submitting late and missing their assessment deadlines, while 71% of students were able to complete assessments on time as it can be seen in Figure 7. It's reasonable to presume that students who missed their deadline did so because of the course's difficulty or because the time allotted for submitting evaluations on the platform was insufficient.

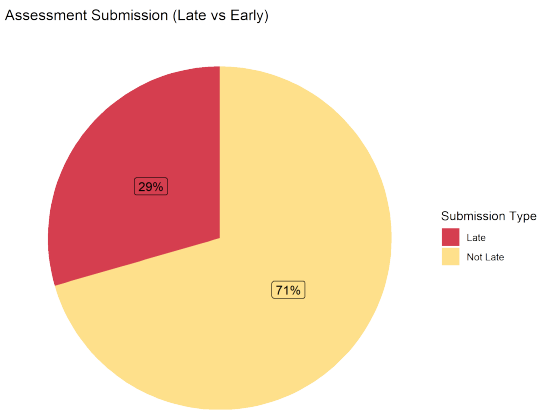


Figure 7: Pie chart of students' late and on-time assessments submission

Figure 13(Appendix D.4) shows that there was one module, "CCC," in which the number of students who submitted late was more than the number of students who submitted on time. The ratio of late vs. on time submission was fairly close in modules like "BBB" and "DDD." Based on this depiction, it's safe to deduce that only a few modules on the platform took less time to submit assessments. To better understand the implications of this close ratio, consider the scatterplot which is presented in Figure 14(Appendix D.4), which shows the average score of students with respect to the date of submission. In the cases of modules "FFF" and "GGG," the scores were evenly distributed, implying that the time of submission had no effect on assessment scores where the difference in ratio of submitting on time and late was close.Also, there were no significant drop in scores. In case of "BBB". there were lot of students which submitted very near to deadline but no significant drops in scores. It is possible that the time allotted for evaluations was insufficient, but the module was most likely simple to grade. However, in the instance of the "CCC" and "DD" module, where students frequently submitted late, the results decreased as the deadline approached. In general, the assessment results can be viewed as motivation to continue the course on this platform, which aided students in achieving high marks on the final test. To increase platform engagement, it's also critical to maintain exams competitive. However, the timeframe for assessments should be reasonable for students, as this will help them get higher assessment scores and desire to complete the course.

#### 4.1.7 Student Disability

The problem lies in the course module. As interpreted from the analysis disabled students that are withdrawing from the course has sufficient amount of interaction which suggest that they are facing problem with a specific module. Specific analysis is available in Appendix G.1.

#### 4.1.8 Number of Attempts

The problem is similar to the issues faced in 4.1.7. Even after interacting with material some of the students are not able to pass their assessment which suggest that they are finding the module material difficult, or they are not able to perform well in a particular assessment which can be causing this. Specific analysis is available in Appendix G.2.

### 4.2 Data Correlation Analysis

#### 4.2.1 Investigating the Effect of Module Length

One of the questions posed was how the length of the course affects the performance of the students and their interaction with the VLE. Each of the modules ran for a different number of days, ranging from 234 days for BBB starting in February 2014 to 269 for AAA, CCC, EEE, FFF and GGG starting in October 2014.

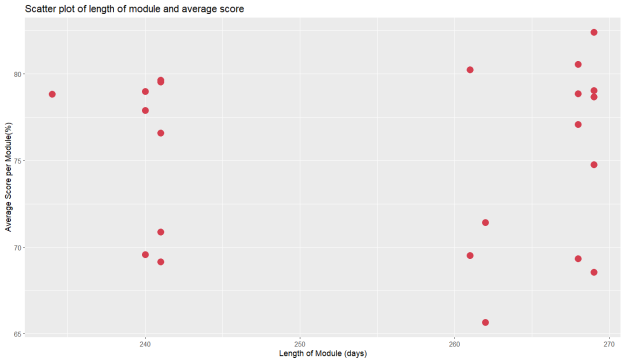


Figure 8: Graph showing the relationship between student performance and the length of the module

**4.2.1.1 Module length and Performance** The effect of the module length on performance was then investigated, and the graph shown in figure 8 was plotted. From this, it is evident that there is no obvious relationship between the length of the module and the final results of the students. To support this, a Pearson's correlation test was carried out which indicated a correlation of -0.01078 between the average score and the module length, and this test had a significance value of 0.962. This non-significant result highlights that there is no true relationship between the module length and the scores of the students. Despite this non- significant result, it is interesting to observe the patterns shown in figure 8. The courses are roughly split into two groups based on their lengths - those around 240 days, and those between 260 and 270 days. There is no difference in the performance of the students in both groups. The mean score for the modules above 250 days is 75.09% whereas the mean score for the modules shorter than 250 days is 75.68% and to test whether the difference between these groups is significant, a t-test was carried out. This test revealed a p-value of 0.7848, indicating that there is no real difference between the scores of the people on the courses shorter than 250 days and longer than 250 days. Even though there is no significant difference, these observations may be beneficial to the Open University as they demonstrate that there is no real benefit to increasing the length of the module as it does not improve the students' final results. Institutions may want to consider shortening their courses to around 240 days as this may save money on resources and may in turn increase the students' engagement.

**4.2.1.2 Engagement with the VLE** Further to this, the effect of the length of modules on the student's interaction with the VLE was investigated. Initially, the total clicks per module were plotted against the module length, however it was quickly realised that this would not be an appropriate measure due to the

modules all varying in length. To effectively standardise the data, the clicks per day were used instead, and this is what has been plotted in figure 15 (Appendix E.4). A Pearson's correlation test was used to assess the significance of the relationship between the clicks per day and the average score revealed a significant correlation of 0.4339079, with a p-value of 0.04363. This means that as the clicks per day increase, the average score also increases, so despite courses being longer having no effect on score, if more interactions are made within that time frame, the overall score is likely to be higher.

### 4.2.2 Relationship between final result and sum clicks

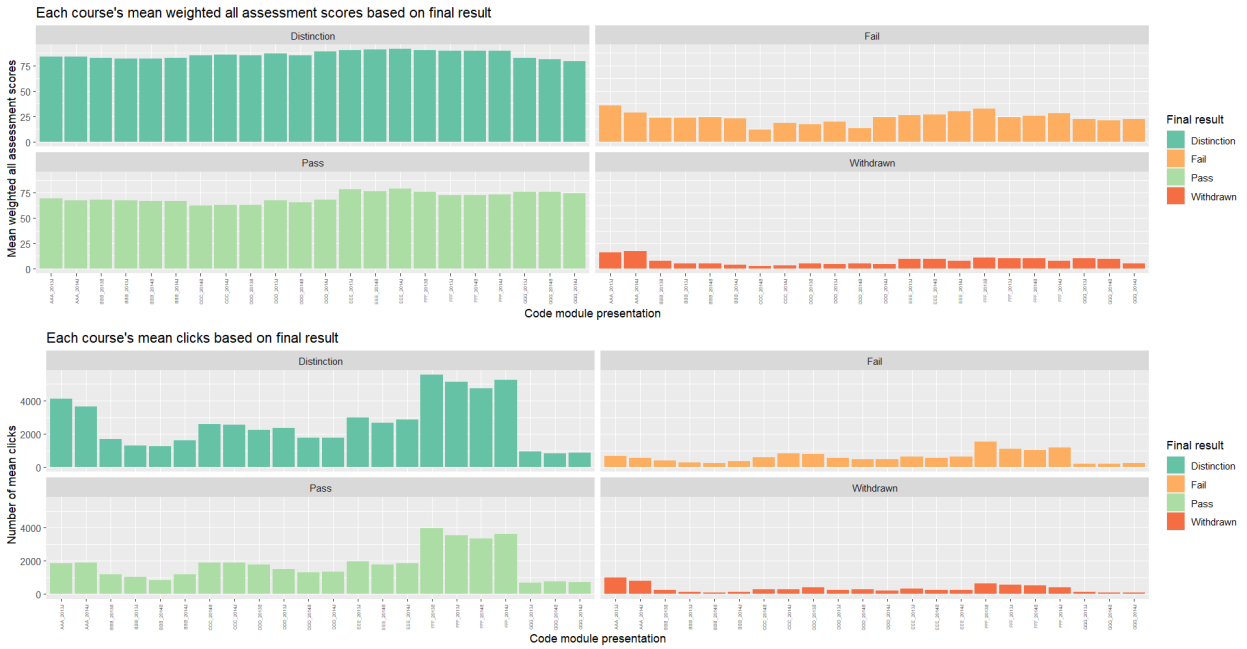


Figure 9: Each course's mean clicks and weighted assessments' score with final result

Figure 9 has been plotted and this shows the scores for each of the modules split by module code and presentation. Although the categories 'Fail', 'Pass' and 'Distinction' are shown separately here, the focus will be on the 'Withdrawn' category.

The differences in the behaviour of the students who did and did not withdraw were examined to identify whether there are significant differences between them. The data was firstly tested for normality using a histogram plot, a Q-Q plot and a Shapiro test, then the differences in variance between the two samples were examined. It became apparent that the data needed to be split by module code. Once the data was split up, it was tested for normality and this test illustrated that it followed a normal distribution. The difference between the variances was not significant. Subsequently, table 21 (Appendix E.1) shows the results of performing a t-test on the final result split by withdrawal status. For all of the module presentations, the p-value was below 0.05 and in almost all cases, the p-value was close to zero. Therefore the conclusion can be made that a student's sum of clicks has a very significant effect on whether a student does or does not withdraw from a module.

### 4.2.3 Relationship between final result and all weighted assessments' score

The relationship between the final result and the weighted summed assessment score was explored using the data processing used in 4.2.2. The top of figure 9 clearly demonstrates that there are significant differences between the weighted summed assessment score for different final results in each module. Table 22 (Appendix

E.2) shows a t-test was performed on the final result by withdrawn status. For all of the module presentations, the p-value was below 0.05 and in almost all cases, the p-value was close to zero. Therefore, a student's weighted summed assessment score has a very significant effect on whether a student does or does not withdraw from a module.

#### 4.2.4 Relationship between final result and submitted date

To further understand the effects of the platform's limited assessment time, a few tests were run to determine the relationship between the date of submission and the assessment score. The procedure began with an assessment of the data's normalcy. The Anderson Darling test was used since there were so many records. Qqplot and histograms were also created to support the normality results. Following a review of the findings, it was determined that the dataset did not have a normal distribution and normality was not presents. After that, the correlation was chosen based on the dataset's random distribution. Kendall rank correlation and Spearman correlation tests were used, as well as a P value test. Theoretically, there were two possibilities: There is no association between the date of submission and the assessment score, according to the null hypothesis. Alternative hypothesis was that there was a link between the date of submission and the date of submission.

| Method | Correlation Value | p-value |
|---|---|---|
| Spearman | -0.2617756 | <2.2e-16 |
| Kendall rank | -0.1830759 | <2.2e-16 |

Table 5: Results of correlation test

The spearman and kendall correlation values were -0.26 and -0.18, respectively, as shown in the results table 5. Both numbers indicate a negative relationship between the date of submission and the final result. The negative figure indicates that as the day of submission advances, the final score decreases. Also, the p-value resulted as significant and it can be assumed that null hypothesis is not true. As a result, students who procrastinated, completed late submissions due to insufficient time allotted for specific assessments or due to the complexity of the course, assessment marks and final exam scores tend to drop, decreasing students' enthusiasm to continue the course.

## 5. Model

According to the fourth part of the correlation study, three explanatory variables and one response variable were selected. The three explanatory variables were each student's total number of clicks, the summation of the weighted assessment score, the total number of late submissions, and the response variable was the student's final result. Table 6 was created to show the first five rows of the datasets for model fitting. The three explanatory variables were normalized due to the wide range of values of the three explanatory variables. The final results were set to 1 for pass and distinction, and 0 for fail and withdrawn.

| Student ID | Late submissions | Weighted-Score | Sum clicks | Final result |
|---|---|---|---|---|
| 6516 | 0 | 63.5 | 2791 | 1 |
| 8462 | 1 | 17.4 | 646 | 0 |
| 11391 | 0 | 82.4 | 934 | 1 |
| 23629 | 3 | 16.7 | 161 | 0 |
| 23698 | 4 | 75.0 | 910 | 1 |

Table 6: Datasets for model fitting(first five rows)

## 5.1 Logistic Regression

The binary logistic regression model has been used to predict the binary response based on one or more predictor variables. In classification problems, the logistic regression essentially uses a logistic function that sets explanatory variables in a sigmoid function, and the logistic regression's range is bounded between 0 and 1. The logistic function is:

$$Pr(Y = 1|\underline{X} = \underline{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}$$

The logistic function uses a loss function called maximum likelihood estimation(MLE), a conditional probability. In this report, if the probability is greater than 0.5, the response variables will be classified as 0, and otherwise will be assigned 1.

The dataset was split into a test and training set of data, with 18,666 records in the training set and 8,000 records in the test set. The model was fitted to the training data, and hypothesis tests were carried out; all of which outputted a p-value of less than 0.05, indicating significance. These four predictors are very valuable in terms of predicting a student's final result, i.e. whether the student will pass (pass or distinction) or fail (fail or withdrawn). The coefficients of all predictors is:

$$Pr(Y = 1|\underline{X} = \underline{x}) = \frac{e^{0.34+0.22*LateSubmission+3.79*Score+0.18*Click}}{1 + e^{0.34+0.22*LateSubmission+3.79*Score+0.18*Click}}$$

The probability of the final result being 'pass', increases as the values of the explanatory variables increases, with *Score* having the most significant effect. The table 7 shows the confusion matrix. The test error is 8.225%.

| | failing(0) | passing(1) |
|---|---|---|
| failing(0) | 2890 | 457 |
| passing(1) | 201 | 4452 |

Table 7: Confusion matrix of testing set in logistic regression

## 5.2 10-fold Cross-validation

Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. The basic form of cross-validation is k-fold cross-validation, and 10-fold cross-validation($K = 10$)is the most common. There are different models to fit, including logistic regression, logistic regression with L1 and L2 regularization, LDA(Linear Discriminant Analysis), QDA(Quadratic Discriminant Analysis) and SVM(Support Vector Machines). The ten test errors of each $k-th$ fold can be obtained and computed the weighted average errors by the folds' sizes to compare the performance fairly. Table 8 shows the results of 10-fold cross-validation. The best model is logistic regression with ridge penalty, which is 7.875% test error and 96.733% AUC.

| Model name | Test error(%) | AUC |
|---|---|---|
| Logistic regression | 8.018% | 96.732% |
| LDA | 7.950% | 91.142% |
| QDA | 8.831% | 90.659% |
| SVM | 8.111% | 90.896% |
| Logistic regression with LASSO penalty | 7.931% | 96.734% |
| Logistic regression with ridge penalty | 7.875% | 96.733% |

Table 8: Test errors and AUC of 10-fold cross-validation in different models

Figure 16(Appendix F) shows the ROC-AUC curve based on the training and testing set. The logistic regression has the highest AUC, which is 96.60%.

# 6. Deployment

All the analysis and models are deployed into the dashboard in figure 10. This dashboard is split into three parts. The first part shows the basic information for the module presentation BBB 2014B and compares this to the previous presentation of the same module BBB 2013B. Red, downwards pointing arrows indicate a decrease as compared to the previous module presentation, while green, upwards pointing arrows indicate an increase.

The second part shows the final result: Distinction, Pass, Fail or Withdrawn. The pie chart on the right shows the type of activity that the student prefers to click on.

The third part shows the student's sum of clicks by date, their weighted assessment scores and the prediction model that indicates whether a student is likely to pass or fail the module. A predicted final result of 0 indicates that the student is likely not to pass the module (a final result of Fail or Withdrawn) while 1 indicates that they are likely to pass (a final result of Distinction or Pass). In this model, there are three factors: sum of clicks, late submission and assessment score; the predicted final result can only be either 0 or 1. If the predicted result for a student is 0, the system could send an email encouraging the student to interact more with the Virtual Learning Environment (VLE) or other targeted interventions could be taken with that student.

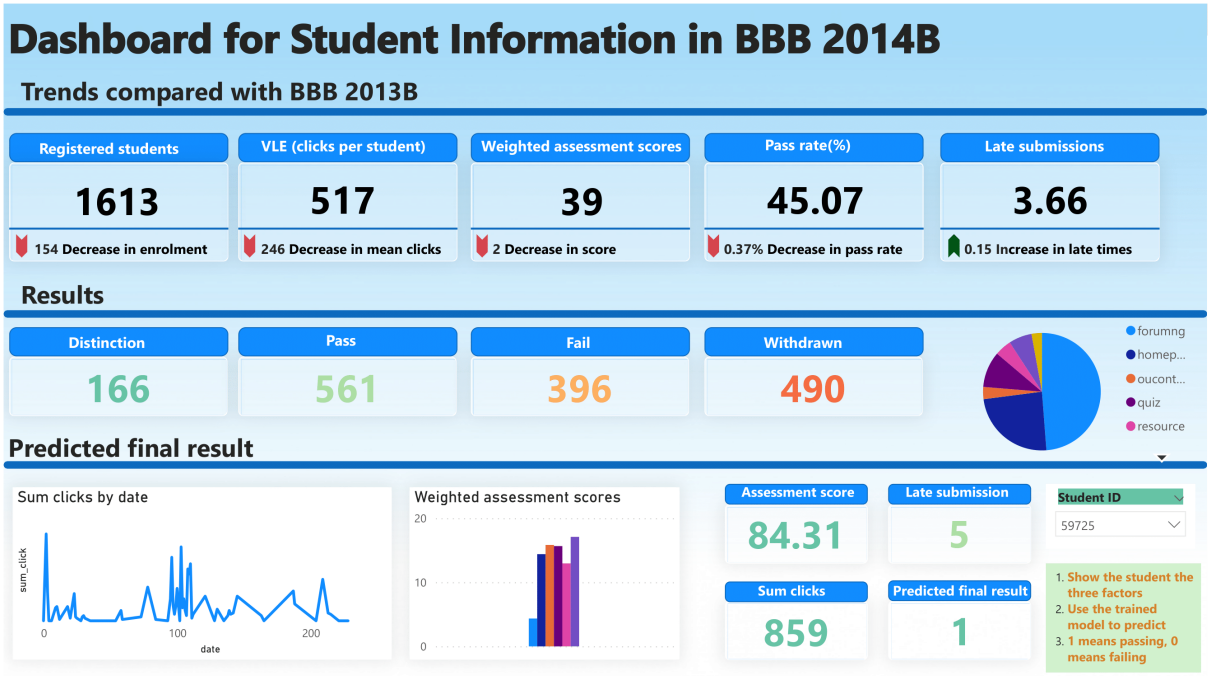


Figure 10: Dashboard

# 7. Conclusions

To conclude, this project has identified several interesting patterns in the results. It has been observed that the most popular course (course with the most students enrolled) was CCC, starting in February 2014, and the least popular was AAA in both 2013 and 2014 courses. There was a significant difference found between the number of students enrolled onto the courses in 2013 compared to 2014, however no difference between the number of students enrolling in February compared to October. This project has also found that the

courses where there was a lot of 'resource' activities had the most students passing and getting distinctions, however the courses where the 'oucontent' was high had a large percentage of students withdrawing. Going forward, this is useful for the institutions as they will be able to adjust the contents of their course to ensure people are achieving the best marks they can. In addition, there was no significant relationship found between the module length and performance of the students, however there was a significant relationship found between the clicks per day and the performance of the students.

After reviewing the visual data plots and statistical analyses, it was possible to establish that the submission time was related to the assessment score. The average score of students who submitted late decreased. There were a few modules where it was discovered that the number of students who submitted late was more than the number of students who submitted on time. It was also discovered that there was a significant decline in score in those modules. It's safe to conclude that the pupils didn't have enough time to complete their assignments. As a result, students who submitted their papers early had a better chance of getting good grades. As a result, because there is a relationship between final exam score and assessment score, students who procrastinated had a lower chance of getting good scores in the final exam.

The analysis of the withdrawal dates found that most, but not all, withdrawals took place near the module start date, which is labelled as day 0. For modules that start in February, the five most popular withdrawal dates were 0 and 27 days after the course had started and one, seven and 12 days before the course starts. For modules with an October start date, the most popular withdrawal dates were 0 and 12 days after the course had started, as well as one and five days before it started. A student's interaction with the VLE also was shown to have a significant effect on whether a student is likely to withdraw from a module or not.

In addition, it was found that between all the modules, 'FFF' had the most student interaction. The number of clicks on the material by students fluctuates significantly throughout the course. The module with the largest number of clicks was FFF2014J, and the module with the smallest number of clicks was GGG2014J. When the number of clicks reached the maximum, the number of days after the start of the module demonstration The highest concentration of clicks was observed between days 20 and 25, and then also between 230 and 260. This may suggest to the university that should try and encourage more students to engage with the material at the start of the course, as well as keeping up with it in between assessments.

*Assumptions:*

During this research, several assumptions have been made about the data, and in turn, this may have caused some problems. One assumption that was made was that the exams took place on the last day of the module. This was because the exam dates were not readily available in the OULAD dataset, meaning that when plottting the time series of clicks according to assessments, it had to be assumed that the final fluctuation in interaction observed at the end of the module was due to upcoming exams. In addition, the assumption was also made that these courses can all be directly compared, meaning that it was assumed that none of the courses were compulsory or significantly easier or harder than others.

*Future Research:*

Future research may need to consider other factors that are likely to influence student interaction, including how many other modules students are studying at the time or the amount of times they have resat the modules. Including these variables in a final predictive model may also allow for an even more accurate prediction of a student's final grade.

# 8. References

2022. [online] Available at: https://help.open.ac.uk/documents/policies/fee-rules/files/69/fee-rules-2012.pdf [Accessed 22 March 2022].

Clow, D., 2013. An overview of learning analytics. Teaching in Higher Education, 18(6), pp.683-695.

Hong, B., Wei, Z. and Yang, Y., 2017, August. Discovering learning behavior patterns to predict dropout in MOOC. In 2017 12th International Conference on Computer Science and Education (ICCSE) (pp. 700-704). IEEE.

Kuzilek, J., Hlosta, M. and Zdrahal, Z., 2017. Open university learning analytics dataset. Scientific data, 4(1), pp.1-8.

Kuzilek, J., Vaclavek, J., Fuglik, V. and Zdrahal, Z., 2018, September. Student drop-out modelling using virtual learning environment behaviour data. In European conference on technology enhanced learning (pp. 166-171). Springer, Cham.

Ukponh, D.E. and George, I.N. 2013. Length of Study- Time Behaviour and Academic Achievement of Social Studies Education Students in the University of Uyo. International Education Studies, 6(3), pp.172-178.

# Appendix

## Appendix A. Business understanding

Figure 11 shows the framework of CRISP-DM.

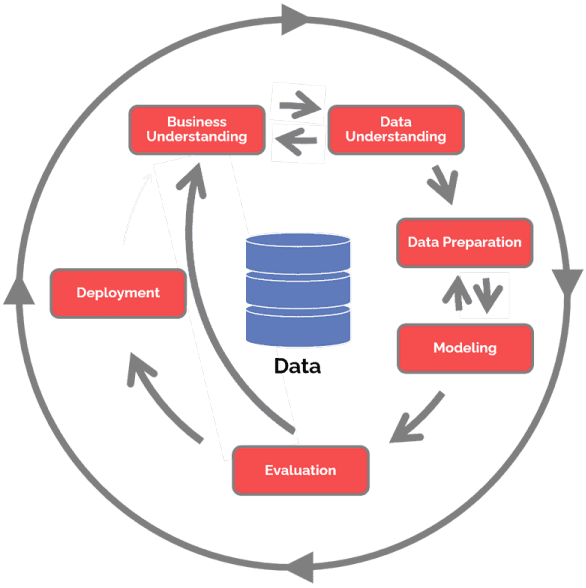


Figure 11: CRISP-DMn

## Appendix B. Data understanding

(Collecting initial data, Describing data, Exploring data, Verifying data quality)

Table 9 shows assessments' information, and the date, weight, and id_assessment are integer and the rest are character. There are 206 records and 6 variables, and eleven of these exams have a submission date of NA.

| code_module | code_presentation | id_assessment | assessment_type | date | weight |
|---|---|---|---|---|---|
| AAA | 2013J | 1752 | TMA | 19 | 10 |
| AAA | 2013J | 1753 | TMA | 54 | 20 |
| AAA | 2013J | 1754 | TMA | 117 | 20 |
| AAA | 2013J | 1755 | TMA | 166 | 20 |
| AAA | 2013J | 1756 | TMA | 215 | 30 |

Table 9: Assessments (first five rows)

- code_module: identification code of the module, to which the assessment belongs.
- code_presentation: identification code of the presentation, to which the assessment belongs.
- id_assessment: identification number of the assessment.
- assessment_type: type of assessment. Three types of assessments exist: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
- date: information about the final submission date of the assessment calculated as the number of days since the start of the module-presentation. The starting date of the presentation has number 0 (zero).
- weight: weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%.

Table 10 shows student assessments' scores, and all data types are integer. There are 173,912 records and 5 variables, and 173 of these students scored NA.

| id_assessment | id_student | date_submitted | is_banked | score |
|---|---|---|---|---|
| 1752 | 11391 | 18 | 0 | 78 |
| 1752 | 28400 | 22 | 0 | 70 |
| 1752 | 31604 | 17 | 0 | 72 |
| 1752 | 38053 | 26 | 0 | 69 |
| 1752 | 45462 | 19 | 0 | 79 |

Table 10: Student assessment (first five rows)

- id_assessment: the identification number of the assessment.
- id_student: a unique identification number for the student.
- date_submitted: the date of student submission, measured as the number of days since the start of the module presentation.
- is_banked: a status flag indicating that the assessment result has been transferred from a previous presentation.
- score: the student's score in this assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail. The marks are in the range from 0 to 100.

Table 11 shows information about the available materials, id_site, week_from, and week_to are integer and the rest are character. There are 6,364 records and 6 variables, and 10,486 of date of material use week have a submission date of NA.

- id_site: an identification number of the material.

| id_site | code_module | code_presentation | activity_type | week_from | week_to |
|---|---|---|---|---|---|
| 546943 | AAA | 2013J | resource | NA | NA |
| 546712 | AAA | 2013J | oucontent | NA | NA |
| 546998 | AAA | 2013J | resource | NA | NA |
| 546888 | AAA | 2013J | url | NA | NA |
| 546614 | AAA | 2013J | resource | NA | NA |

Table 11: Vle (first five rows)

- code_module: an identification code for module.
- code_presentation: the identification code of presentation.
- activity_type: the role associated with the module material.
- week_from: the week from which the material is planned to be used.
- week_to: week until which the material is planned to be used.

Table 12 shows information about each student's interactions with the materials, and all data types are integer. There are 1,065,5280 records and 6 variables, and no NA value is available.

| code_module | code_presentation | id_student | id_site | date | sum_click |
|---|---|---|---|---|---|
| AAA | 2013J | 28400 | 546652 | -10 | 4 |
| AAA | 2013J | 28400 | 546652 | -10 | 1 |
| AAA | 2013J | 28400 | 546652 | -10 | 1 |
| AAA | 2013J | 28400 | 546614 | -10 | 11 |
| AAA | 2013J | 28400 | 546714 | -10 | 1 |

Table 12: Student vle (first five rows)

- code_module: an identification code for a module.
- code_presentation: the identification code of the module presentation.
- id_student: a unique identification number for the student.
- id_site: an identification number for the VLE material.
- date: the date of student's interaction with the material measured as the number of days since the start of the module-presentation.
- sum_click: the number of times a student interacts with the material in that day.

Table 13 contains the list of all available modules and their presentations, and courses' Length are integer and the rest are character. There are 22 records and 3 variables, and no NA value is available.

| code_module | code_presentation | module_presentation_length |
|---|---|---|
| AAA | 2013J | 268 |
| AAA | 2014J | 269 |
| BBB | 2013J | 268 |
| BBB | 2014J | 262 |
| BBB | 2013B | 240 |

Table 13: courses (first five rows)

- code_module: code name of the module, which serves as the identifier.
- code_presentation: code name of the presentation. It consists of the year and "B" for the presentation starting in February and "J" for the presentation starting in October.
- length: length of the module-presentation in days.

Table 14 contains information about the time when the student registered for the module presentation, and registration and unregistration dates are integer, and others are character. There are 32,593 records and 5 variables, and 22566 of registration and unregistration dates are NA.

| code_module | code_presentation | id_student | date_registration | date_unregistration |
|---|---|---|---|---|
| AAA | 2013J | 11391 | -159 | NA |
| AAA | 2013J | 28400 | -53 | NA |
| AAA | 2013J | 30268 | -92 | 12 |
| AAA | 2013J | 31604 | -52 | NA |
| AAA | 2013J | 38053 | -176 | NA |

Table 14: Student registration (first five rows)

- code_module: an identification code for a module.
- code_presentation: the identification code of the presentation.
- id_student: a unique identification number for the student.
- date_registration: the date of student's registration on the module presentation, this is the number of days measured relative to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).
- date_unregistration: date of student unregistration from the module presentation, this is the number of days measured relative to the start of the module-presentation. Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final_result column in the studentInfo.csv file.

Table 15 contains demographic information about the students together with their results, and id_student, studied credits, number of previous attempts are integer, and others are character. There are 32,593 records and 7 variables, and no NA value is available.

| code_module | code_presentation | id_student | attempts | studied_credits | disability | final_result |
|---|---|---|---|---|---|---|
| AAA | 2013J | 11391 | 0 | 240 | N | Pass |
| AAA | 2013J | 28400 | 0 | 60 | N | Pass |
| AAA | 2013J | 30268 | 0 | 60 | Y | Withdrawn |
| AAA | 2013J | 31604 | 0 | 60 | N | Pass |
| AAA | 2013J | 38053 | 0 | 60 | N | Pass |

Table 15: Student information (first five rows)

- code_module: an identification code for a module on which the student is registered.
- code_presentation: the identification code of the presentation during which the student is registered on the module.
- id_student: a unique identification number for the student.
- num_of_prev_attempts: the number times the student has attempted this module.
- studied_credits: the total number of credits for the modules the student is currently studying.
- disability: indicates whether the student has declared a disability.
- final_result: student's final result in the module-presentation.

## Appendix C. Data Preparation

1. Exploring the most popular activity types: In order to analyse the percentage of different activity types pre-set for the course and student interaction, two tables were created as follows:

(1) In the VLE table, the columns of `code_module`, `code_presentation`, and `activity_type` were selected to calculate the proportion of each activity type.
(2) Secondly, in the VLE table, the `activity_type` and `id_site` columns were extreacted and combined with the 'studentVLE' data frame. In this data frame, the `code_module`, `code_presentation`, and `activity_type` columns were grouped to calculate the percentage of each activity types' clicks.

Table 16 shows the first three rows. There are 22 records and 8 variables. Based on the percentage of activity types in the 22 courses, the six types that usually account for the highest percentage are selected, and all other types are grouped into others.

| code name | forumng | oucontent | resource | subpage | url | quiz | others |
|---|---|---|---|---|---|---|---|
| AAA_2013J | 7.109 | 32.227 | 45.023 | 2.843 | 8.530 | 0.000 | 4.265 |
| AAA_2014J | 2.970 | 33.663 | 46.039 | 2.970 | 9.900 | 0.000 | 4.455 |
| BBB_2013B | 5.396 | 0.317 | 74.920 | 11.746 | 4.761 | 1.587 | 1.269 |

Table 16: Percentage of pre-course setting activity

Table 17 shows the first three rows.There are 22 records and 10 variables. Based on the percentage of activity types in the 22 courses, the eight types that usually account for the highest percentage are selected, and all other types are grouped into others.

| code name | forumng | homepage | oucontent | resource | subpage | url | quiz | ouwiki | others |
|---|---|---|---|---|---|---|---|---|---|
| AAA_2013J | 27.064 | 21.641 | 41.246 | 1.135 | 6.110 | 2.429 | 0.000 | 0.000 | 0.371 |
| AAA_2014J | 27.014 | 21.127 | 42.783 | 1.263 | 5.342 | 2.110 | 0.000 | 0.000 | 0.357 |
| BBB_2013B | 60.175 | 21.660 | 0.102 | 3.453 | 5.925 | 1.472 | 6.959 | 0.000 | 0.252 |

Table 17: Percentage of student's clicking activity type

## Appendix D. Preliminary data analysis

### Appendix D.1 Students Withdraw Data

Figure 12 shows time Series of unregistrtion dates.

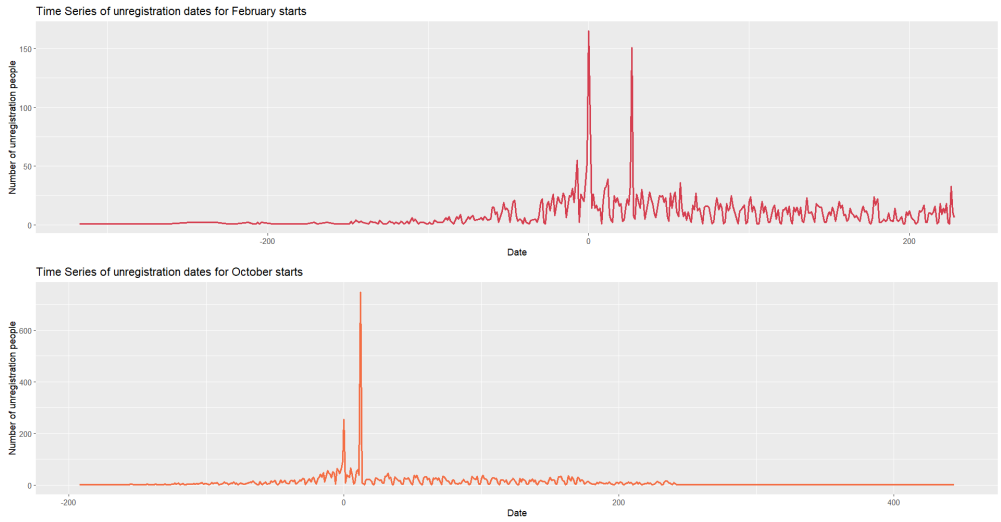


Figure 12: Time Series of unregistrtion dates

### Appendix D.2 Student Daily Activity (Sum of Clicks)

Table 18 shows the summary of student daily activity. Since each student has different daily clicks on the presentation materials, make a time series diagram for the total clicks of each module, and check that the most clicks during that period. First analyze the AAA2014J module. It can be seen from the sequence diagram that the total average clicks are concentrated within [2000, 4000]. When date=18, the maximum clicks are sum_click=7974, date=259, and the minimum clicks are sum_click =44. And at date=241, the click volume has increased sharply, sum click=5853, it can be inferred that the students are highly motivated to learn about the module on the 18th day after the module demonstration, and on the 18th day after the module starts At 241 days, the number of clicks increased significantly due to exam review and other reasons. In the BBB module, the code for the demo module with the maximum number of hits is BBB2014J and, of the four modules, the maximum number of hits is around the day after the start of the module. BBB2014J The total average number of clicks is concentrated within [5000,10000], the maximum number of clicks is sum_click=21336,date=236, the minimum number of clicks is sum_click=58,date=259. The module has four segments with clicks above 12500, indicating that the module has four segments of assessment, all with clicks around 100 clicks or so. In CCC2014J, the difference in the number of clicks is evident from the chronological chart, which shows that the number of clicks before and after each assessment reaches 50,000, indicating that students are not usually motivated to learn the module, with the maximum number of clicks being sum_click=50543,date=18,and the minimum number of clicks being sum_click=367,date=266. The number of clicks in DDD2013B is decreasing, with a maximum of sum_click=18871, date=24, and a minimum of sum_click=192, date=238..In EEE2014J, the number of clicks increases gradually from date=90 days, with a maximum of sum_click=26164, date=166 and a minimum of sum_click=113, date=266. The average number of clicks in the FFF module was above 20,000, indicating that the number of people choosing the module was high and that students were more interested in the module, with the maximum number of clicks concentrated around 200 days after the start of the module. fff2014b The maximum number of clicks was sum_click=41598, date=226,and the minimum number of clicks was sum_click=251, date=238. FFF2014J Maximum clicks are sum_click=73321, date=240, minimum clicks are sum_click=333, date=266.

| module | Max_click | Date | Min_click | Date |
|---|---|---|---|---|
| AAA2014J | 7974 | 18 | 44 | 259 |
| AAA2013J | 7881 | 18 | 36 | 259 |
| BBB2014J | 21336 | 236 | 58 | 259 |
| BBB2014B | 12010 | 2 | 187 | 231 |
| BBB2013B | 21256 | 2 | 640 | 238 |
| BBB2013J | 19872 | 2 | 257 | 259 |
| CCC2014B | 42454 | 18 | 523 | 238 |
| CCC2014J | 50543 | 18 | 367 | 266 |
| DDD2013B | 18871 | 24 | 192 | 238 |
| DDD2013J | 21474 | 233 | 158 | 259 |
| DDD2014B | 12591 | 24 | 186 | 238 |
| DDD2014J | 17775 | 2 | 143 | 259 |
| EEE2013J | 25061 | 19 | 51 | 259 |
| EEE2014B | 15547 | 19 | 238 | 35 |
| EEE2014J | 26164 | 166 | 113 | 266 |
| FFF2013B | 56837 | 221 | 299 | 231 |
| FFF2013J | 58175 | 18 | 301 | 267 |
| FFF2014B | 41598 | 226 | 251 | 238 |
| FFF2014J | 73321 | 240 | 333 | 266 |
| GGG2013J | 6925 | 228 | 21 | 246 |
| GGG2914B | 7405 | 221 | 28 | 238 |
| GGG2014J | 6419 | 228 | 14 | 252 |

Table 18: Date of sum clicks in eac module

### Appendix D.3 Students' preference for activity type

Table 19 shows percentage of sum clicks based on each activity type in each module, and table 20 shows percentage of course pre-set activity type in each module.

### Appendix D.4 Students performance depending on submission date

| Module Name | forumng | homepage | oucontent | resource | subpage | url | quiz | ouwiki | others |
|---|---|---|---|---|---|---|---|---|---|
| AAA 2013J | 27.07 | 21.64 | 41.25 | 1.14 | 6.11 | 2.43 | 0 | 0 | 0.37 |
| AAA 2014J | 27.02 | 21.13 | 42.78 | 1.26 | 5.34 | 2.11 | 0 | 0 | 0.36 |
| BBB 2013B | 60.17 | 21.66 | 0.1 | 3.45 | 5.92 | 1.47 | 6.96 | 0 | 0.25 |
| BBB 2013J | 50.94 | 25.02 | 2.53 | 3.96 | 5.91 | 2.8 | 8.61 | 0 | 0.23 |
| BBB 2014B | 48.7 | 24.13 | 3.36 | 4.61 | 6.32 | 2.84 | 9.88 | 0 | 0.16 |
| BBB 2014J | 11.36 | 17.84 | 51.2 | 4.42 | 3.56 | 0.2 | 10.05 | 0 | 1.37 |
| CCC 2014B | 13.54 | 16.58 | 10.96 | 4.61 | 8.17 | 0.58 | 45.32 | 0 | 0.24 |
| CCC 2014J | 13.86 | 15.86 | 11.08 | 4.68 | 6.74 | 0.61 | 46.83 | 0 | 0.34 |
| DDD 2013B | 22.58 | 27.04 | 11.78 | 5.32 | 18.46 | 4.88 | 0 | 6.36 | 3.58 |
| DDD 2013J | 25.48 | 25.94 | 12.51 | 4.78 | 19.01 | 2.36 | 0 | 6.72 | 3.2 |
| DDD 2014B | 23.32 | 24.85 | 13.15 | 5.88 | 20.2 | 2.44 | 0 | 7.31 | 2.86 |
| DDD 2014J | 35.58 | 25.5 | 12.41 | 6.13 | 13.8 | 2.17 | 0 | 0.02 | 4.4 |
| EEE 2013J | 21.62 | 15.93 | 36.79 | 1.19 | 2.06 | 2.38 | 8.47 | 11.17 | 0.39 |
| EEE 2014B | 14.7 | 16.38 | 41.21 | 1.7 | 2.4 | 2.92 | 8.74 | 11.72 | 0.24 |
| EEE 2014J | 20.93 | 16.93 | 37.54 | 1.24 | 1.97 | 2.01 | 7.42 | 11.69 | 0.28 |
| FFF 2013B | 13.47 | 13.9 | 37.82 | 1.3 | 9.37 | 1.08 | 20.38 | 0.92 | 1.75 |
| FFF 2014J | 13.06 | 13.57 | 38.2 | 1.25 | 8.93 | 0.92 | 21.96 | 0.88 | 1.22 |
| FFF 2014B | 11.86 | 13.31 | 37.95 | 1.46 | 9.23 | 0.96 | 23.07 | 0.97 | 1.18 |
| FFF 2014J | 13.06 | 13.57 | 38.2 | 1.25 | 8.93 | 0.92 | 21.96 | 0.88 | 1.22 |
| GGG 2013J | 18.77 | 21.61 | 35.13 | 4.6 | 3.63 | 0 | 15.62 | 0 | 0.64 |
| GGG 2014B | 20.53 | 21.22 | 33.51 | 5.42 | 3.96 | 0 | 14.93 | 0 | 0.43 |
| GGG 2014J | 14.12 | 20.83 | 39.05 | 6.54 | 4.38 | 0 | 14.35 | 0 | 0.74 |

Table 19: Percentage of sum clicks based on each activity type in each module

| Module Name | forumng | oucontent | resource | subpage | url | quiz | others |
|---|---|---|---|---|---|---|---|
| AAA 2013J | 7.11 | 32.23 | 45.02 | 2.84 | 8.53 | 0 | 4.27 |
| AAA 2014J | 2.97 | 33.66 | 46.04 | 2.97 | 9.9 | 0 | 4.46 |
| BBB 2013B | 5.4 | 0.32 | 74.92 | 11.75 | 4.76 | 1.59 | 1.27 |
| BBB 2013J | 5.92 | 0.93 | 73.52 | 11.84 | 4.67 | 1.56 | 1.56 |
| BBB 2014B | 5.47 | 0.96 | 74.28 | 11.9 | 4.5 | 1.61 | 1.29 |
| BBB 2014J | 1.45 | 33.82 | 50.24 | 4.83 | 2.9 | 1.93 | 4.83 |
| CCC 2014B | 4.59 | 23.98 | 39.8 | 14.29 | 6.63 | 8.16 | 2.55 |
| CCC 2014J | 4.04 | 26.01 | 38.12 | 13.9 | 8.52 | 7.17 | 2.24 |
| DDD 2013B | 2.57 | 1.4 | 42.52 | 26.64 | 21.5 | 0 | 5.37 |
| DDD 2013J | 3.46 | 2.81 | 38.53 | 41.99 | 9.52 | 0 | 3.68 |
| DDD 2014B | 2.87 | 2.87 | 39.07 | 42.6 | 9.05 | 0 | 3.53 |
| DDD 2014J | 3.56 | 3.56 | 46.3 | 30.14 | 12.6 | 0 | 3.84 |
| EEE 2013J | 4.46 | 42.86 | 30.36 | 6.25 | 6.25 | 1.79 | 8.04 |
| EEE 2014B | 4.59 | 42.2 | 29.36 | 6.42 | 6.42 | 1.83 | 9.17 |
| EEE 2014J | 4.35 | 39.13 | 32.17 | 6.09 | 6.09 | 1.74 | 10.43 |
| FFF 2013B | 0.98 | 21.26 | 25.39 | 10.83 | 24.61 | 2.17 | 14.76 |
| FFF 2014J | 1.55 | 22.79 | 17.7 | 11.5 | 29.2 | 2.43 | 14.82 |
| FFF 2014B | 1.26 | 21.13 | 22.8 | 11.09 | 27.82 | 2.3 | 13.6 |
| FFF 2014J | 1.55 | 22.79 | 17.7 | 11.5 | 29.2 | 2.43 | 14.82 |
| GGG 2013J | 1.46 | 17.52 | 69.34 | 3.65 | 0 | 6.57 | 1.46 |
| GGG 2014B | 1.61 | 20.16 | 65.32 | 4.03 | 0 | 7.26 | 1.61 |
| GGG 2014J | 1.89 | 24.53 | 59.43 | 4.72 | 0 | 7.55 | 1.89 |

Table 20: Percentage of course pre-set activity type in each module

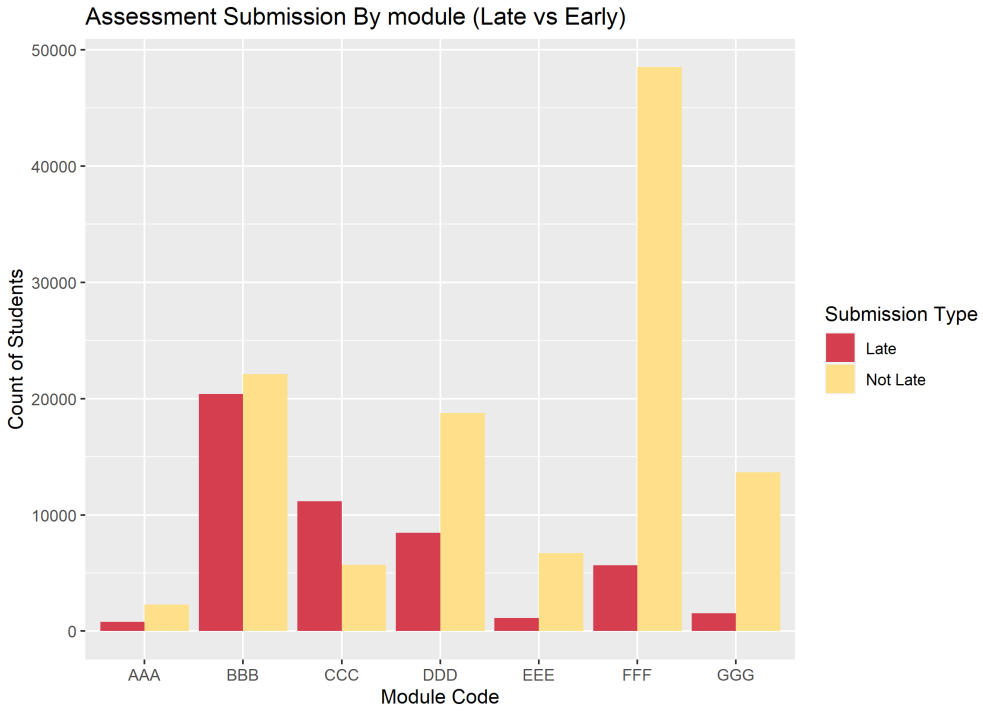


Figure 13: Grouped Bar chart of students' late and on-time assessments submission module wise

## Appendix E. Data Analysis

### Appendix E.1 Relationship between final result and sum clicks

### Appendix E.2 Relationship between final result and all weighted assessments' score

### Appendix E.3 Data about module length with different presentation Code along with Assessments count

### Appendix E.4 Engagement with the VLE

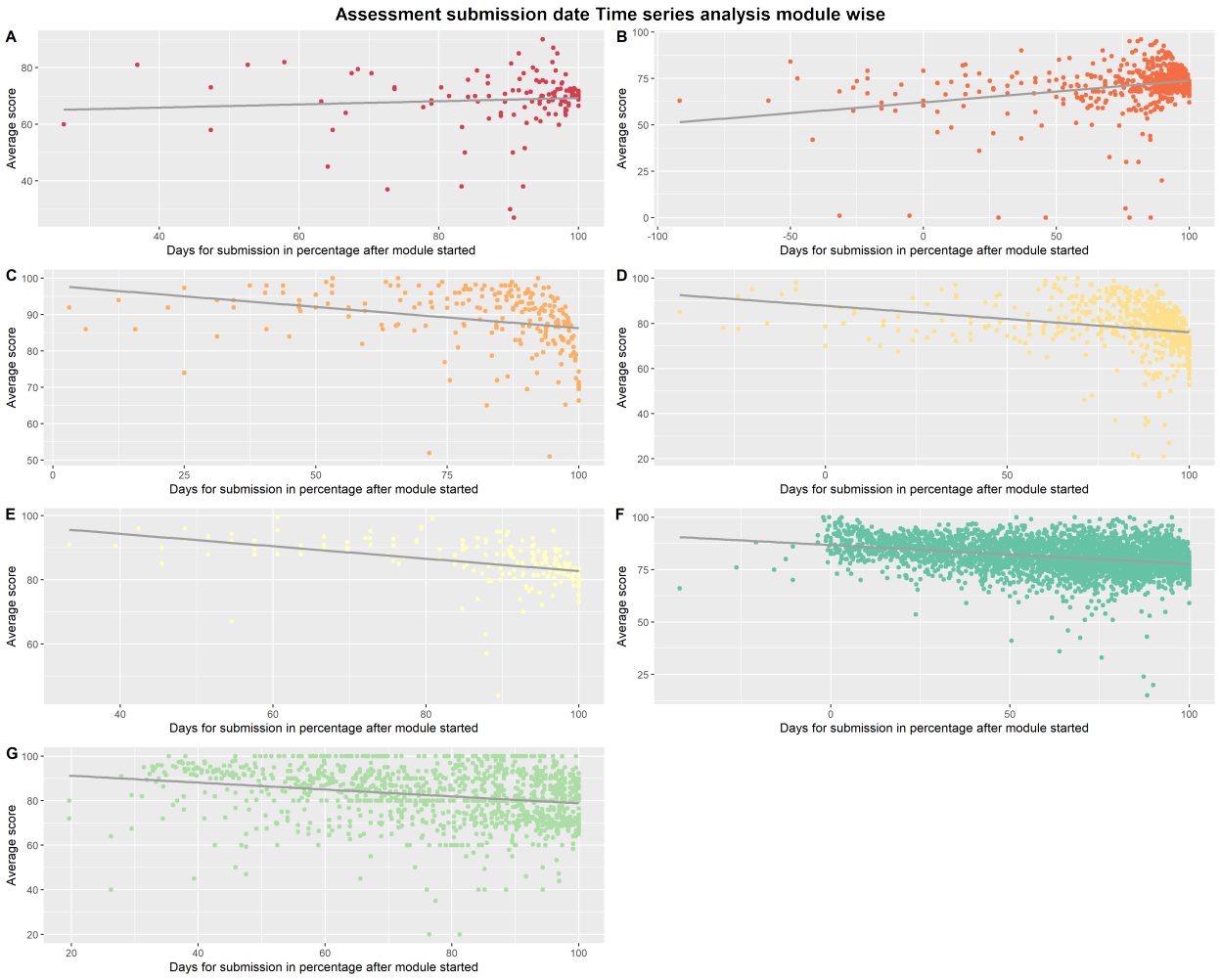


Figure 14: Scatterplot of students' late and on-time assessments submission module wise

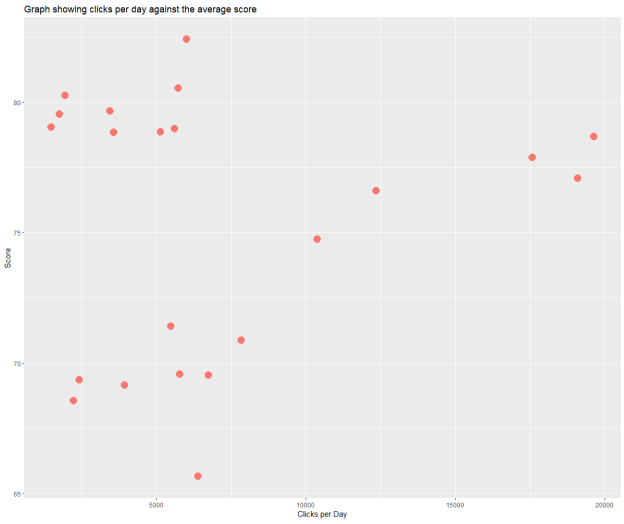


Figure 15: Graph showing the relationship between interaction and the length of the module

| Module presentation | Mean (withdrawn) | Mean (not withdrawn) | t-value | p-value |
|---|---|---|---|---|
| AAA2013J | 987.32 | 1824.32 | 3.50 | 0.0005177 |
| AAA2014J | 812.68 | 1821.14 | 4.03 | 6.88e-05 |
| BBB2013B | 248.24 | 968.74 | 10.63 | <2.2e-16 |
| BBB2013J | 129.44 | 813.12 | 12.51 | <2.2e-16 |
| BBB2014B | 96.49 | 700.43 | 12.08 | <2.2e-16 |
| BBB2014J | 125.68 | 1023.73 | 25.51 | <2.2e-16 |
| CCC2014B | 303.09 | 1557.81 | 20.75 | <2.2e-16 |
| CCC2014J | 305.44 | 1733.99 | 24.41 | <2.2e-16 |
| DDD2013B | 395.97 | 1396.56 | 15.78 | <2.2e-16 |
| DDD2013J | 272.21 | 1250.47 | 20.26 | <2.2e-16 |
| DDD2014B | 303.68 | 1082.46 | 16.64 | <2.2e-16 |
| DDD2014J | 225.43 | 1117.56 | 16.44 | <2.2e-16 |
| EEE2013J | 314.47 | 1804.13 | 14.93 | <2.2e-16 |
| EEE2014B | 246.51 | 1515.15 | 13.05 | <2.2e-16 |
| EEE2014J | 237.43 | 1750.09 | 19.48 | <2.2e-16 |
| FFF2013B | 654.93 | 3284.19 | 21.29 | <2.2e-16 |
| FFF2013J | 570.03 | 2942.58 | 24.90 | <2.2e-16 |
| FFF2014B | 544.48 | 2624.35 | 20.33 | <2.2e-16 |
| FFF2014J | 406.40 | 3267.77 | 33.02 | <2.2e-16 |
| GGG2013J | 126.15 | 565.20 | 6.08 | 1.724e-09 |
| GGG2014B | 97.84 | 566.69 | 8.14 | 1.465e-15 |
| GGG2014J | 82.90 | 624.69 | 10.37 | <2.2e-16 |

Table 21: The t-test value of sum clicks in two groups

| Module | Mean (withdrawn) | Mean (not withdrawn) | t-value | p-value |
|---|---|---|---|---|
| AAA2013J | 65.7 | 16.19 | 20.386 | <2.2e-16 |
| AAA2014J | 62.76 | 16.99 | 16.089 | <2.2e-16 |
| BBB2013B | 53.69 | 7.62 | 46.334 | <2.2e-16 |
| BBB2013J | 54.88 | 5.09 | 44.118 | <2.2e-16 |
| BBB2014B | 53.88 | 4.96 | 38.055 | <2.2e-16 |
| BBB2014J | 57.64 | 4.1 | 55.194 | <2.2e-16 |
| CCC2014B | 48.4 | 2.34 | 44.89 | <2.2e-16 |
| CCC2014J | 55.2 | 2.99 | 56.684 | <2.2e-16 |
| DDD2013B | 45.36 | 4.88 | 29.464 | <2.2e-16 |
| DDD2013J | 52.58 | 4.65 | 43.178 | <2.2e-16 |
| DDD2014B | 50.15 | 5 | 31.717 | <2.2e-16 |
| DDD2014J | 56.2 | 4.29 | 61.936 | <2.2e-16 |
| EEE2013J | 67.2 | 9.84 | 29.223 | <2.2e-16 |
| EEE2014B | 62.53 | 9.41 | 27.908 | <2.2e-16 |
| EEE2014J | 70.07 | 7.79 | 37.31 | <2.2e-16 |
| FFF2013B | 62.01 | 11.07 | 35.448 | <2.2e-16 |
| FFF2013J | 58.9 | 10.31 | 41.148 | <2.2e-16 |
| FFF2014B | 56.92 | 10.14 | 40.506 | <2.2e-16 |
| FFF2014J | 64.11 | 7.68 | 57.686 | <2.2e-16 |
| GGG2013J | 59.1 | 10.55 | 12.357 | <2.2e-16 |
| GGG2014B | 57.72 | 9.63 | 14.818 | <2.2e-16 |
| GGG2014J | 60.41 | 4.93 | 20.616 | <2.2e-16 |

Table 22: The t-test value of weighted assessments' score in two groups

| Module Code | Presentation Code | Module Length | Assessments Count |
|---|---|---|---|
| AAA | 2013J | 268 | 5 |
| | 2014J | 269 | 5 |
| BBB | 2013B | 240 | 11 |
| | 2013J | 268 | 11 |
| | 2014B | 234 | 11 |
| | 2014J | 262 | 5 |
| CCC | 2014B | 241 | 8 |
| | 2014J | 269 | 8 |
| DDD | 2013B | 240 | 14 |
| | 2013J | 261 | 7 |
| | 2014B | 241 | 7 |
| | 2014J | 262 | 6 |
| EEE | 2013J | 268 | 5 |
| | 2014B | 241 | 5 |
| | 2014J | 269 | 5 |
| FFF | 2013B | 240 | 13 |
| | 2013J | 268 | 13 |
| | 2014B | 241 | 13 |
| | 2014J | 269 | 13 |
| GGG | 2013J | 261 | 10 |
| | 2014B | 241 | 10 |
| | 2014J | 269 | 10 |

Table 23: Module Presentation Length with assessments count

## Appendix F. Model

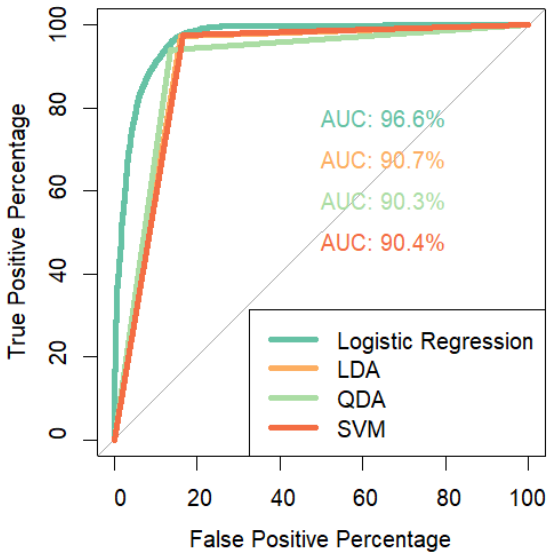


Figure 16: Plot of ROC-AUC curve based on fitted training dataset

## Appendix G. Eda extension study

### Appendix G.1 Student Disability

The next stage of the analysis will be focusing on those students who are disabled. As per figure you can see the distribution of disabled and non-disabled students per course. The reason for this analysis is to take the variation into account. This variation is related to the difference in the problem that a disabled student suffers from. This problem is unique and cannot be always treated by implementing a solution which was interpreted by considering the common issues that a non-disabled student encounter. As per the above table you can see that there is a significant difference between these two categories. Although the population of disabled student is not high as compared to non-disabled student but still their involvement in the course needs to consider as there is a huge difference between the problems they encounter and the problem a non-disabled student can face. The main goal of this analysis was to determine whether the interaction of these students with the course material is affecting their result or not. As per the problem statement that has been discussed above the idea here is that if the students are able to secure good grades, then they will be motivated to engage more with the materials. But before focusing separately on those students who are disabled a common graph was plotted in which it showed the engagement of every student in each course.

The idea behind this analysis was to see the interactions of disabled & non-disabled students in each course. As per the graph we can see that the interaction in modules “GGG2014B” is lowest as compared to other course modules whereas module “FFF2014” have the highest interactions whereas for disabled students the statistics is different. The highest interaction for disabled students is “FFF2013J” and the low number of interactions is in the module “GGG2014J”. Based on this result further analysis was directed towards the results of these students.

As you can see in this graph for both disabled and non-disabled students the number of interactions is high for the “pass” category of result and is low for “withdrawn” category of result. In the case of disabled students, the lowest number of clicks (or interaction) belongs to the result category “distinction” which suggest that the interaction of those disabled students who are withdrawing from the course is high as compared to non-disabled students. To justify this, result a pie chart was also plotted which displays the interaction percentage of all students and their result.

According to the chart it can be deduced that the significant difference between the interaction percentage

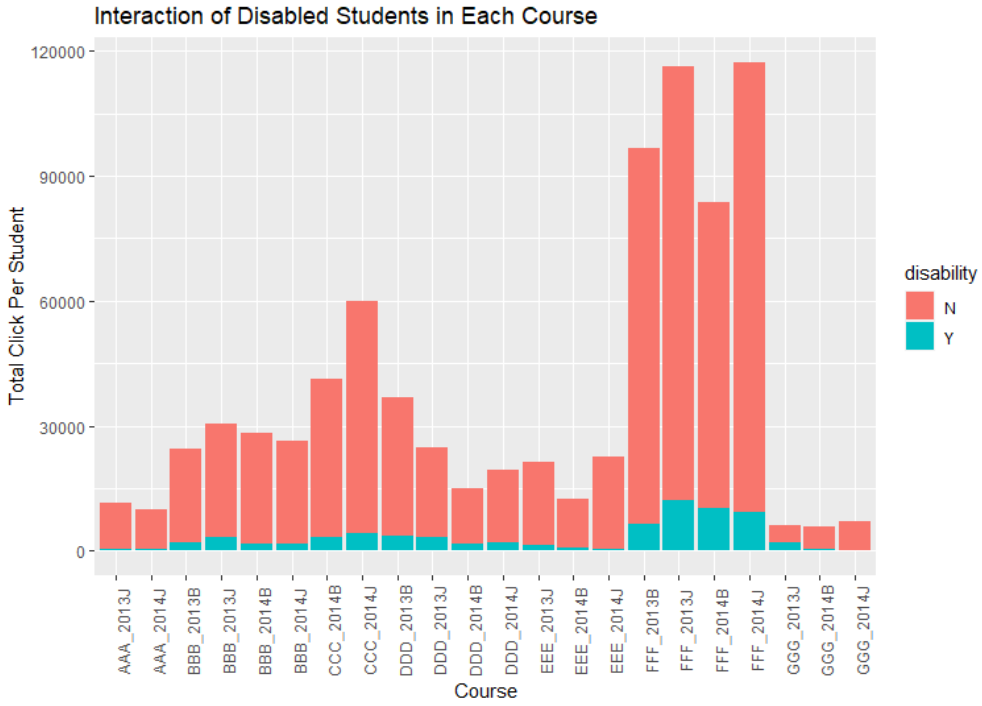


Figure 17: Graph showing the interaction of disabled and non-disabled students in each course

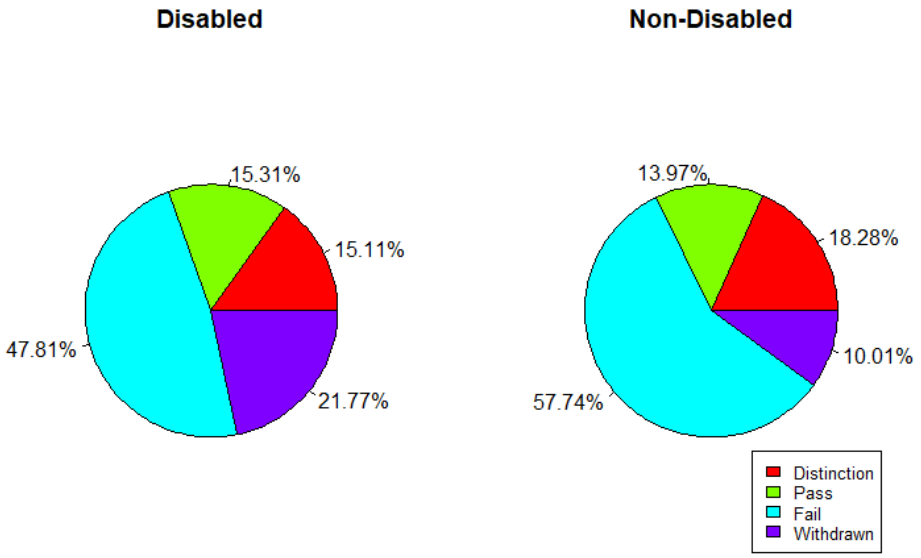


Figure 18: Chart showing the interaction percentage of all students and their results

| Result | P-Value |
|---|---|
| Distinction | 0.06024 |
| Pass | 0.0597 |
| Fail | 0.02911 |
| Withdrawn | 0.2256 |

Table 24: P-Values

can be seen in the withdrawn category of the result. To summarize those students who are disabled are interacting more with the material but still withdrawing form the course as compared to the students who are not disabled. To support this analysis a hypothesis test was conducted in which the interaction percentage was compared for both students using p-value was calculated applying the t-test method. Here the null hypothesis was considered that there is a significant difference between the interaction percentage of disabled and non-disabled students for all result categories. This hypothesis will be true if the p-value is $> 0.05$. As per the table it can be seen that the null hypothesis is true for distinction, pass, and withdrawn category of result. Although for distinction & pass category the value is near to 0.05 whereas for the withdrawn category the difference is huge.

The problem lies in the course module. As interpreted from the analysis disabled students that are withdrawing from the course has sufficient amount of interaction which suggest that they are facing problem with a specific module. As per the graph you can see that their highest interaction is with the module "FFF_2013J" which suggest that this module can be the reason for their withdrawal. Although it cannot be said with a certainty that this module is causing problem so the best way to overcome this issue is to conduct research whose focus will be on all those things that can make a module difficult and after analysing the data that will be the generated result of this research a proper pipeline can be constructed which solve their problems.

### Appendix G.2 Number of Attempts

For this next analysis the question is focused on the number of attempts which was taken by every student to pass their assessment. The above graph indicates the number of students who passed on their first try which is much higher than those students who took 1 or more than 1 attempt to clear their assessment whereas some of the students were not able to pass their assessment after taking 1 or more than one attempt. As per the graph less than 3000 students took another attempt. It can also be seen that the number of students is declining with each attempt. For the further analysis a pie chart was plotted. According to the above plotted chart it can be seen that there is a huge difference between the interaction percentage of those students who failed even after taking various attempts. This suggest that even though their interaction is more than the last time but still not able to pass their assessment. To support this analysis another hypothesis test was conducted in which the null hypothesis was same as the one considered in 4.1.8 which states that there is a significant difference between the interaction percentage of those students who didn't took any attempt and those who took 1 or more than 1 attempt.

As per the table, the null hypothesis has been rejected for the pass category of the result which confirms the above analysis that even though students are interacting more with the material they are still not able to clear their assessment.

Here the problem is similar to the issues faced in 4.1.7. Even after interacting with material some of the students are not able to pass their assessment which suggest that they are finding the module material difficult, or they are not able to perform well in a particular assessment which can be causing this. Although this hypothesis is based on an assumption so the best way to construct a solution is to again conduct a research and after analysing that data we can land on a proper solution.

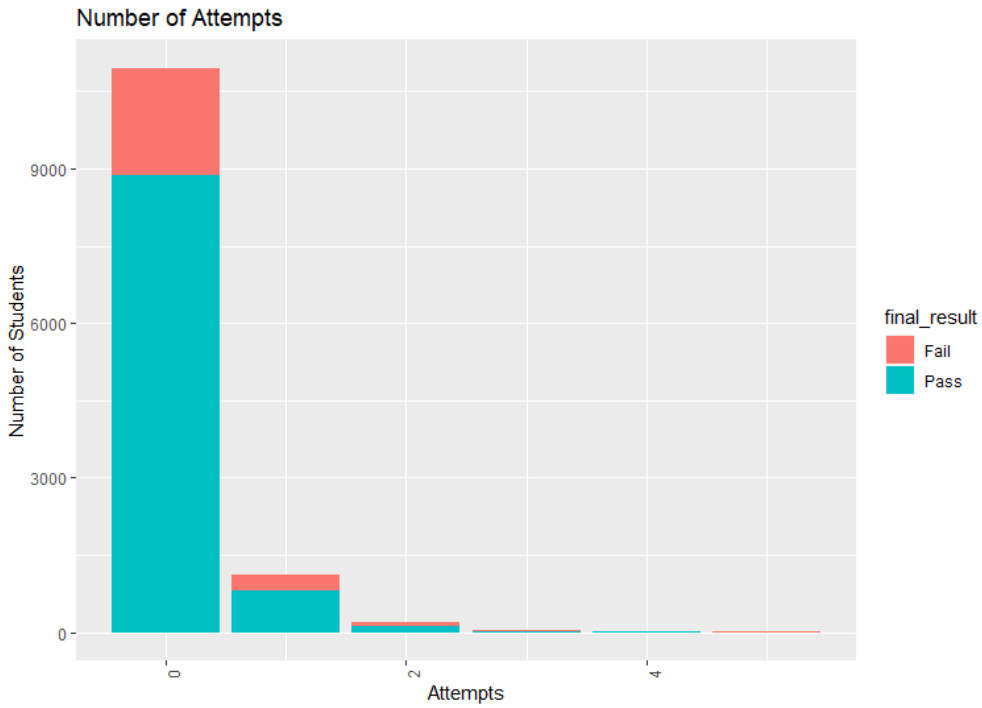


Figure 19: Graph showing the number of attempts and their results

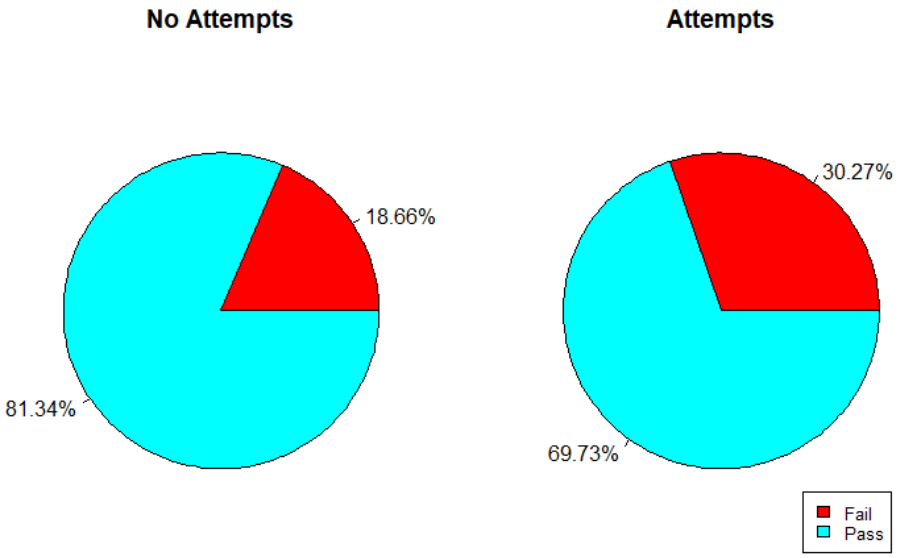


Figure 20: Chart showing the interaction for the students who took various attempts and who didn't