Analysis on Cyber Security Learning course

Mayank Baraskar

Abstract

This study uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology and R to analyze a Massive Open Online Course (MOOC) on cyber security for the year 2018. Massive open online courses (MOOCs) have been widely employed in the field of education and have lately been promoted as a result of the COVID-19 epidemic. The primary goal of this research is to determine why enrollment in the course has decreased by examining their interactions as well as the completion rate of the course work. This study intends to better understand why students drop out of class and, as a result, improve the course process for the benefit of both students and professors. It is well established that different learners use learning materials in different ways. Some students, for example, tend to complete their course work, whereas others attempt to learn crucial concepts from the entire course work and abandon the rest. This study, on the other hand, can be utilized to figure out the most prevalent student learning patterns and reasons for dropping out. The study was conducted using data from students who enrolled in the Cyber Security online course in 2018. The course was offered in three distinct intakes during the year, in February, June, and September. The goal is to improve the course work for future intakes with the help of visualization of this analysis so that students can get better quality material and presentation of total course work with a lower leaving rate.

Introduction

The use of online learning and techniques such as Massive Open Online Courses (MOOCs) has increased significantly as a result of the COVID-19 epidemic. FutureLearn is a learning platform that allows students to enroll in any field course and complete their coursework online. Students can learn from home or anywhere else because the course materials are easily available online. They are not required to attend traditional classrooms.

There are many benefits to using this platform, the most important of which is flexibility, since students may replay content and look over course work anytime they choose. MOOC courses were created to improve learner performance and learning outcomes. When a student enrolls in a course, he or she is required to do course work, which includes notes, videos, and tests. However, course analytics have revealed that, despite these benefits, many students abandon the course halfway through and, even if they finish, do not perform well. Each intake, the completion rate decreases in tandem with the number of people enrolled. Because the number of students enrolled is bigger than in traditional classrooms, and they come from all over the world, it will be difficult for professors to grasp each student's condition. Because not every student will be comfortable with the course work, material, or any other reason, students make a choice either to drop out or refuse to enroll in future intakes and courses. (Yu, Wu, and Liu, 2019).

The goal of this study is to learn about students' behavior patterns during the course, using analysis to determine when most students are likely to drop out and what the most prevalent reasons are for dropping out. The investigation will be based on data from three different intakes of FutureLearn's cyber security course, which is one of the MOOC learning platforms. The year 2018 was chosen from the entire data set since it had the leaving responses for various intakes in 2018, as opposed to 2016 and 2017, which will improve the accuracy of interpreting the learner's comments.

Research method

We will mine data using the CRISP-DM approach in this investigation. It stands for cross-industry data mining methodology. It is a very adaptable, reliable, and well-proven methodology. It's a six-step cycle strategy made up of events with various goals. We can adjust the sequence of events based on the business requirements because it is highly adaptable. As shown in Figure 1, there are primarily six steps, each of which has a distinct purpose.

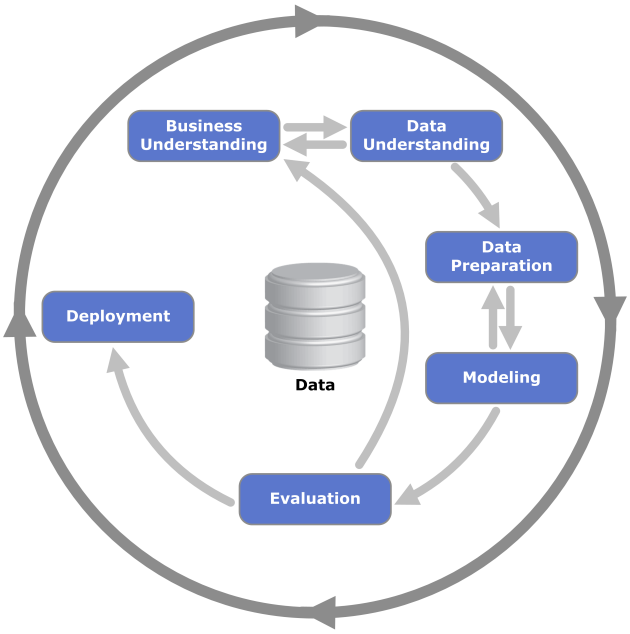

Figure 1: The CRISP-DM Model (Yaacob, Nasir, Yaacob and Sobri, 2019).

- **Business Understading**: This step focuses mostly on the business perspective and its requirements. After reviewing the situation, such as data availability, there are primarily two purposes for analysis in this study. The initial goal is to determine when the majority of students will be leaving the course. The second goal is to figure out why you dropped out of school. The overarching purpose is to do an exploratory data analysis on students and generate a report that will help institute enhance course content and increase general engagement.

- **Data understanding**: Understanding the data is the second phase in the CRISP DM process. The major goal of this stage is to identify, collect, and evaluate the data set that FutureLearn has provided to us. R was used to load the entire data set into Project Template. After the data was loaded, it was evaluated to determine the data format, number of records, and noise level in order to clean the data further. In this step, visuals for learning metrics were created, which FutureLearn can use to optimise the course process.

- **Data Preparation**: The preparation of data is the third process. Data munging is another name for it. The final data set for the visualisation challenge was prepared in this stage. Unnecessary data was removed from the whole data set. The chosen data was cleaned and assembled according to the specifications. Because the data was split into numerous files, it was combined, and joins were employed in some cases. The data was also formatted, with string numbers being converted to numeric values for mathematical calculations.
- **Modelling and Evaluation**: Machine learning and statistics are mostly involved in this step. The first phase in the modelling process is to choose methodologies, which might include algorithms such as regression, classification, and others. The model is fitted to the data in order to produce precise

predictions. The accuracy is calculated at the evaluation step, which divides the entire data set into training and test sets. The model will be installed into a training set and put through its paces on a test set. The decision can be made if these models met the business criteria via evaluation. However, because the focus of this study is on data analysis, neither of these stages will be included.

- **Deployement**: The report, which comprises the data analysis and the final presentation for the outcome, is prepared during the deployment step.(CRISP-DM - Data Science Process Alliance, 2021)

Data Description

Learners who enrolled in FutureLearn's cyber security course had their data utilised. Because there was insufficient data for analysis for the previous year, the data for the year 2018 was used, with three different intakes in the months of February, June, and September. Multiple variables were used for data analysis, and they are listed in the table below with descriptions. This definition can be used as a reference throughout the report because these attributes were employed throughout the research. This table can also show which data attributes were accessible and how they were used for analysis. The data properties included in the table are a comprehensive list; nonetheless, they are used in various data frames.

Sr No.	Attribute Name	Description
1	learner_id	Unique ID which is alloted to learner at time of enrollment
2	enrolled_at	Timestamp of when the student is enrolled
3	fully_participated_at	Timestamp for when the learner completed the course
4	left_at	Timestamp for when the learner left the course
5	leaving_reason	Reason for learner leaving the course
6	last_completed_step_at	Timestamp for when the learner completed the last step
7	fully_participated_at	Timestamp for when the learner completed the course
8	last_completed_step	Last completed step number
9	last_completed_step_week_no	The week number when learner completed the last step
10	last_completed_step_no	The step number when learner completed the last step
11	correct	Boolean attribute whether the quiz response was correct or not

Analysis

As two business objectives have been specified in this article, there are two CRISP-DM cycles. The steps of CRISP-DM have been followed and discussed separately for each business aim.

Business Objective 1 - An analysis of when the majority of learners drop out of the course.

There are numerous topics that will be explored in this business purpose. The main goal is to figure out when the majority of students abandon the course halfway through. Despite the fact that MOOCs are flexible, allowing students to play content or complete course work whenever they choose, unlike traditional classrooms, students can opt to leave or unenroll from the course.
The procedure was advanced to the second level of Data understanding because our business aim for this cycle has been sorted. The data came from FutureLearn's records, which covered seven runs with various input and years. The procedure was advanced to the second level of Data understanding because our business aim for this cycle has been sorted. The data came from records provided by FutureLearn, which covered seven separate runs with different input and years. Students' enrollment, archetype survey responses, quiz responses, step activity, weekly attitude survey responses, leaving survey responses, video statistics, and team members were all part of the data for each intake. Because the goal of this business

is to figure out when students choose to drop out, the focus is on enrollment data and survey responses. Unfortunately, there was no data from previous intakes for the departing survey, but there were responses for the year 2018, thus this analysis will be done on three separate intakes from that year to gain a better understanding and learn learners behavioural patterns. The raw data of enrollments and survey responses was in the following format after collecting the data for 2018:

Enrollment Data

```
head(feb_2018_enrollments, 4)
```

```
##                             learner_id             enrolled_at
## 1 480d5ab0-b755-4ea7-8374-100c25b920c4 2018-03-26 12:41:18 UTC
## 2 46a4c71e-8819-4c5a-8164-2f786186b9fb 2018-01-02 20:57:49 UTC
## 3 afaf7031-a708-49f6-a823-37eb8e4bc721 2018-03-31 01:03:01 UTC
## 4 41ad5273-6853-43d7-b615-06f8e808cea1 2018-09-10 12:08:49 UTC
##             unenrolled_at    role fully_participated_at purchased_statement_at
## 1 2018-10-19 09:57:54 UTC learner
## 2 2018-10-19 10:31:15 UTC learner
## 3 2018-10-14 04:09:49 UTC learner
## 4                         learner
##    gender country age_range highest_education_level employment_status
## 1    male      GB     46-55               tertiary working_full_time
## 2  female      GR     26-35     university_masters full_time_student
## 3 Unknown Unknown   Unknown                Unknown           Unknown
## 4 Unknown Unknown   Unknown                Unknown           Unknown
##                   employment_area detected_country
## 1     it_and_information_services               GB
## 2 accountancy_banking_and_finance               GR
## 3                         Unknown               VE
## 4                         Unknown               --
```

There are 3544 observations of 13 variables in the enrollments data, as can be seen. For this data, the variables were cleaned and prepared as follows.

- **learner_id**: Because there are no null values in this variable, no cleaning or inclusion/exclusion was performed.
- **enrolled_at**: Because the values are timestamps of when the student enrolled and there are no NA or null values, no cleaning or inclusion/exclusion was necessary.
- **unenrolled_at**: Because the data will be left joined with the survey response which has left_at attribute, and unenrolled at variables appears to be inconsistent, this column was removed.
- **role**: Because the focus of this variable is on learner role rather than organisation admin, the data was filtered for role admin.
- **fully_participated_at**: This variable is needed to figure out what the completion rate was, so it will be included. However, there are NA values that can be assumed to mean that the learner did not complete the course, and if the value has a timestamp, the learner completed the course. This variable is transformed into a categorical variable with the categories "Completed" and "Not Completed."
  There are some variables that aren't needed because they're not part of our analysis for the business goal, so they've been left out. Due to unknown values and NA values, the data was also very inconsistent. They are *purchased_statement_at*, *gender*, *country*, *age_range*, *highest_education_level*, *employment_status*, *employment_area*, *detected_country*

After transforming and cleaning enrollment data, the subset looks like:

```
head(feb_2018_enrollments_final, 4)
```

```
##                             learner_id             enrolled_at    role
## 1 480d5ab0-b755-4ea7-8374-100c25b920c4 2018-03-26 12:41:18 UTC learner
## 2 46a4c71e-8819-4c5a-8164-2f786186b9fb 2018-01-02 20:57:49 UTC learner
## 3 afaf7031-a708-49f6-a823-37eb8e4bc721 2018-03-31 01:03:01 UTC learner
## 4 41ad5273-6853-43d7-b615-06f8e808cea1 2018-09-10 12:08:49 UTC learner
##   fully_participated_at
## 1         Not Completed
## 2         Not Completed
## 3         Not Completed
## 4         Not Completed
```

Leaving Survey Response Data

```
head(feb_2018_leaving_survey, 4)
```

```
##      id                             learner_id                 left_at
## 1 34003 8853543d-b930-43e0-a3cb-99de6b9ea4f3 2018-01-30 20:04:59 UTC
## 2 38604 b170480c-7ed6-434f-8827-73cfa130ece1 2018-02-05 03:01:39 UTC
## 3 39016 92b8485e-b2a6-4345-8721-3e93142b469d 2018-02-05 10:30:03 UTC
## 4 39241 f4cae359-0966-4a6d-830f-4218861f2fd9 2018-02-05 13:27:14 UTC
##                                leaving_reason last_completed_step_at
## 1                          I prefer not to say
## 2 The course required more time than I realised
## 3                                       Other
## 4                                       Other
##   last_completed_step last_completed_week_number last_completed_step_number
## 1                  NA                         NA                         NA
## 2                  NA                         NA                         NA
## 3                  NA                         NA                         NA
## 4                  NA                         NA                         NA
```

There are 173 observations of 8 variables in the enrollments data, as can be seen. For this data, the variables were cleaned and prepared as follows.

- **id**: This variable is for storing unique ID while recording leaving survey, however this will be excluded from the filtered data
- **learner_id**: Because there are no null values in this variable, no cleaning or inclusion/exclusion was performed. However, this variable was used for joining enrollment and leaving survey response data
- **left at**: Because the values are timestamps of when the student left the course and there are no NA or null values, no cleaning or inclusion/exclusion was necessary.
- **leaving reason**: This variable contains the reasons for the learner's withdrawal from the course. For the purposes of accurate analysis, we will omit the complete record for some NA values. Furthermore, when the reason was recorded, the string was not converted, resulting in responses like "The course wasn't what I expected," which were transformed to "The course wasn't what I expected" for better visualisation.

For better visualisation and accuracy, we omitted complete rows for some variables that contained NA values. They are *last_completed_step_at*, *last_completed_step*, *last_completed_week_number*, *last_completed_step_number*

After transforming and cleaning enrollment data, the subset looks like:

```
head(feb_2018_leaving_survey_final, 4)
```

```
##                               learner_id                left_at
## 6  a04b5f29-c283-419f-b262-5797c9220aa9 2018-02-05 18:12:27 UTC
## 11 3ae76b85-153a-4fe6-8a24-46d795e2608c 2018-02-06 20:00:27 UTC
## 14 f6ec6af3-d021-4422-a7c7-82012145f590 2018-02-07 12:49:19 UTC
## 19 5086b814-3121-43e7-999b-926113ae98d2 2018-02-09 18:31:44 UTC
##                          leaving_reason  last_completed_step_at
## 6                                 Other 2018-02-05 18:12:15 UTC
## 11 The course wasnâ\200\231t what I expected 2018-02-06 19:04:22 UTC
## 14             The course was too easy 2018-02-06 11:17:02 UTC
## 19                 I prefer not to say 2018-02-09 18:24:29 UTC
##    last_completed_step last_completed_week_number last_completed_step_number
## 6                  1.2                          1                          2
## 11                 1.1                          1                          1
## 14                 1.2                          1                          2
## 19                 1.3                          1                          3
```

The data preparation and data understanding are done simultaneously because the study is based on the CRISP-DM model for exploratory data analysis and the main objective is to visualise and explore the data rather than modelling it. Now that the final enrollment and survey response data is clean and filtered according to the business objective, the left join was used to merge the data from both tables. Because a left join was used after merging the data and not all of the students responded to the learning survey, the NA values were again omitted throughout the data. After merging the data, this were the columns which were created:

```
colnames(feb_2018_merged_data)
```

```
##  [1] "learner_id"                "enrolled_at"
##  [3] "role"                      "fully_participated_at"
##  [5] "left_at"                   "leaving_reason"
##  [7] "last_completed_step_at"    "last_completed_step"
##  [9] "last_completed_week_number" "last_completed_step_number"
## [11] "intake"
```

Because the data set for all of the intakes was quite similar, the entire data preparation and cleaning process was repeated separately for June and September intakes. Because the data set was similar for all intakes, the entire data preparation and cleaning process was repeated separately for June and September intakes. In addition, the datasets from February, June, and September were concatenated, yielding 152 observations across 11 variables. As a result, we were finished with the data understanding and data preparation and were ready to move on to the exploratory data analysis. (Ford, 2021).

Here are some of the questions that arose while looking at the data while going through cleaned and prepared data that can help us achieve our business goal of determining when the majority of students tend to drop out of the course.

- **What was the course completion rate, in terms of understanding whether learners actually left the course?**

To understand this question, the visualization was made in form of bar chart where x axis was kept as intake and y axis as percentage of completion rate.

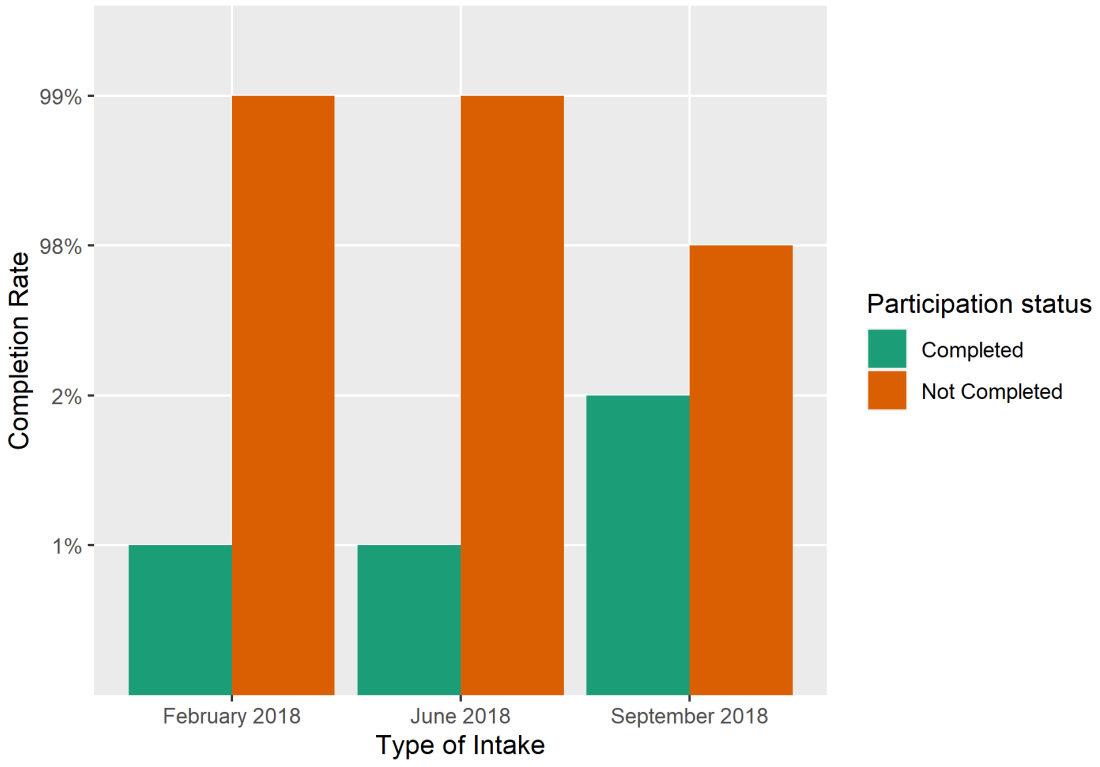

Figure 2: Comparison of completion rate month wise

As can be seen in Figure 2, the completion rate for each intake in 2018 was extremely low. In the months of February and June, the completion rate was only about 1%, and in September, it was only about 2%. It can be concluded that the majority of students chose to drop out of the course for various reasons. This analysis leads us to the next question: at what point did learners decide to drop out of the course?

- **Which week did the most and least people drop out of the course?**

Since we arrived at this conclusion based on a month-by-month analysis of completion rates, it is clear that the majority of students dropped out of the course. The only question now is when. In this step, we'll analyse the aggregate data from various intakes to gain a better understanding of the situation and see if any patterns emerge. For this step, a graph was created with the x axis representing the week number and the y axis representing the sum of learners who dropped out during that week.

As shown in Figure 3, the drop rate is highest in Week 1, at 53%, and lowest in Week 2. In Week 3, the drop increased to 31 percent. This analysis indicates that the majority of students choose to drop out of the course during Week 1, when the entire course work is typically introduced. It is reasonable to assume that the majority of people do not want to continue with the course because they dislike the course content or the format in which it is presented. As the investigation progresses, an analysis of the reason will be carried out. However, because the percentage of people who drop out of the course in week 2 has dropped significantly, it can be assumed that those who choose to stay in the course intend to finish it. However, several students dropped out of the course by week 3, indicating that they were either no longer interested or considered the course to be too long.

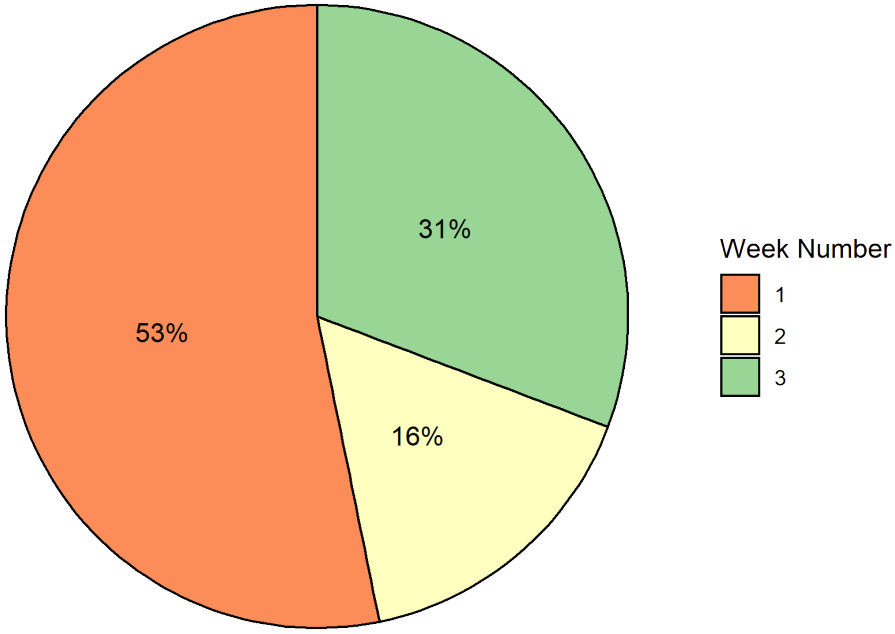

Figure 3: Student leaving percentage week-wise

- **At what step number did the majority and least number of persons abandon the course?**

Since then, it has been noted that the bulk of students dropped out of the course in week one, with the fewest dropping out in week two. As a result, going ahead, the goal is to see if there are any patterns in stepwise data. As a result, the data from the leaving survey is utilised to determine the number of students that dropped out of each phase. In figure 4, the total number of students left in each step was calculated and presented as a bar chart, with the y axis representing the number of students and the x axis representing the step numbers. To improve the accuracy of discovering the pattern among the majority of students, the step in which only one student quit the course was ignored.

After looking at the graph, it's clear that the majority of students are leaving at step 3.20, which is the last item that deals with the course's conclusion. As a result, it's possible that a significant percentage of students completed the course but departed early to submit the survey rather than waiting until the end of the course, implying that they actually digested the course, which appears to be a positive indicator. However, the second biggest number of students abandoning the course is found in the penultimate step of week 1, which is part of the rounding up of the course given in week 1. It appears that the student tried the course for the first week but decided to drop out since they didn't like it or didn't find the curriculum relevant. There are also a large number of students who dropped out right after the first step, which is to say, "Welcome to the course session." They abandoned the course because they were either dissatisfied with the course or wasn't what the learners expected when they went through this session. There could be a variety of reasons for leaving, but the most typical interpretation appears to be that they completed the course for week 1, which served as a trial, but were unable to continue owing to some cause.

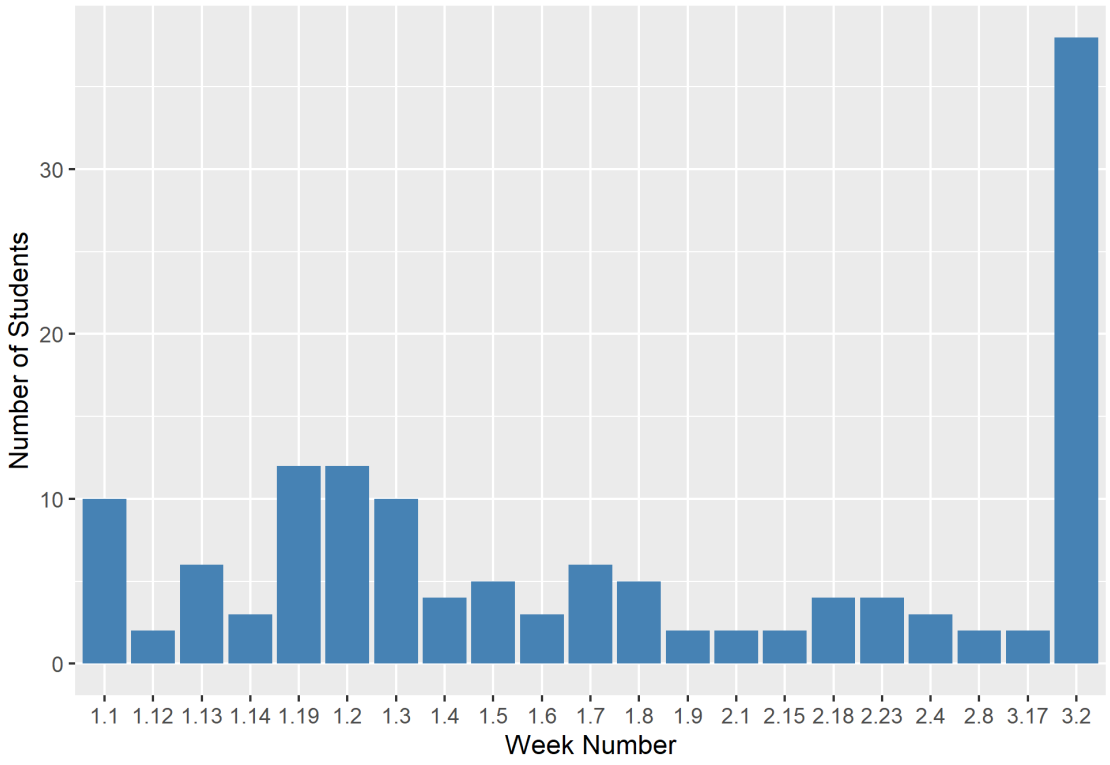

Figure 4: Comparison of number of students leaving in different steps

Business Objective 2 - An analysis of why the majority of students left the course.

When the majority of students abandon the course halfway through the preceding business objective and cycle of CRISP-DM, the exploratory data analysis was finished and the conclusion was formed. In addition, data analysis based on learner patterns revealed the most likely reasons. The goal of this objective is to figure out what the main reasons are for students dropping out of classes. It can assist professors and course administrators in improving course work and processes in order to promote student engagement and enrollment. A similar CRISP-DM process was used for this business aim as it was for the first. However, the data set that will be used for this data analysis is the same as that which was cleaned and prepared for the previous business purpose as below.

```
colnames(feb_2018_merged_data)
```

```
##  [1] "learner_id"                "enrolled_at"
##  [3] "role"                      "fully_participated_at"
##  [5] "left_at"                   "leaving_reason"
##  [7] "last_completed_step_at"    "last_completed_step"
##  [9] "last_completed_week_number" "last_completed_step_number"
## [11] "intake"
```

There is one additional data set, however, that requires comprehension and preparation in order to help the data analysis. It's a collection of quiz responses. The purpose of this data set was to better understand the learners' performance and see whether there was any pattern or correlation with the feedback.

Question Response Data

In the same way that the data from three intakes for the year 2018 is used, the question response data is used in the same way. For February intake, there are 18752 observations of 10 variables. The data looks something like below:

```
head(feb_2018_question_response, 4)
```

```
##                              learner_id correct
## 1 06b2174a-9624-407c-83ba-d26e999b5fc9    true
## 2 42b5a416-a42a-4929-aac3-6c551cacaf15    true
## 3 80186114-623e-4708-a61f-00dc2d99d70d    true
## 4 f91b7ed8-6f24-4640-93c9-432ecaaf7a74    true
```

Since our main focus in this business objective is to find the reason behind leaving the course and this data might not seem to be relevant, however, this data is going to be as supporting factor. So, the variables which are required are just 2 which are as follows.

- **learner_id**: Because there are no null values in this variable, no cleaning or inclusion/exclusion was performed. However, this variable was used for joining enrollment and leaving survey response data
- **correct**: This variable is for storing whether the quiz response given by learner is correct or not. It is a boolean variable and no NA or null values are present hence no omitting required for this data set.

Rest of the variables will be removed from the data set. This data set will be merged in the same way that the business objective was merged. The information from three intakes will be combined and the percentage was calculated for comparison. The clean and merged data looks like.

```
head(all_month_res_perf, 2)
```

```
## # A tibble: 2 x 5
##   correct              n  perc Percentage_labels intake
##   <chr>            <int> <dbl> <chr>             <chr>
## 1 Incorrect Answer  8000 0.427 43%               February 2018
## 2 Correct Answer   10752 0.573 57%               February 2018
```

Here are some of the questions that arose while looking at the data while going through cleaned and prepared data that can help us achieve our business goal of determining why the majority of students tend to drop out of the course.

- **Can course difficulty be a factor in students dropping out?**

Based on the trend of learners abandoning the course in week 1, it's possible that either the learners were dissatisfied with the course or the difficulty level was either too easy or too challenging, causing the learner to exit the course. Because we have data on quiz response for all of the intakes, the goal of this study is to evaluate the data to see if difficulty was a factor. The average is calculated for all intakes where students either answered properly or incorrectly to the quiz question that is part of the course work in this question.

To visualise this, a dogde style bar plot was used, with the x axis representing different intakes and the y axis representing the percentage of students that gave accurate and incorrect responses.

As shown in Figure 5, the performance of students has been very consistent with each intake. When comparing the results to those from traditional classrooms, we can conclude that the students who enrolled in this course and attempted the quiz questions did not do particularly well. On average, only 57 percent of students are able to provide correct answers, but 43 percent provide incorrect responses. Given that the

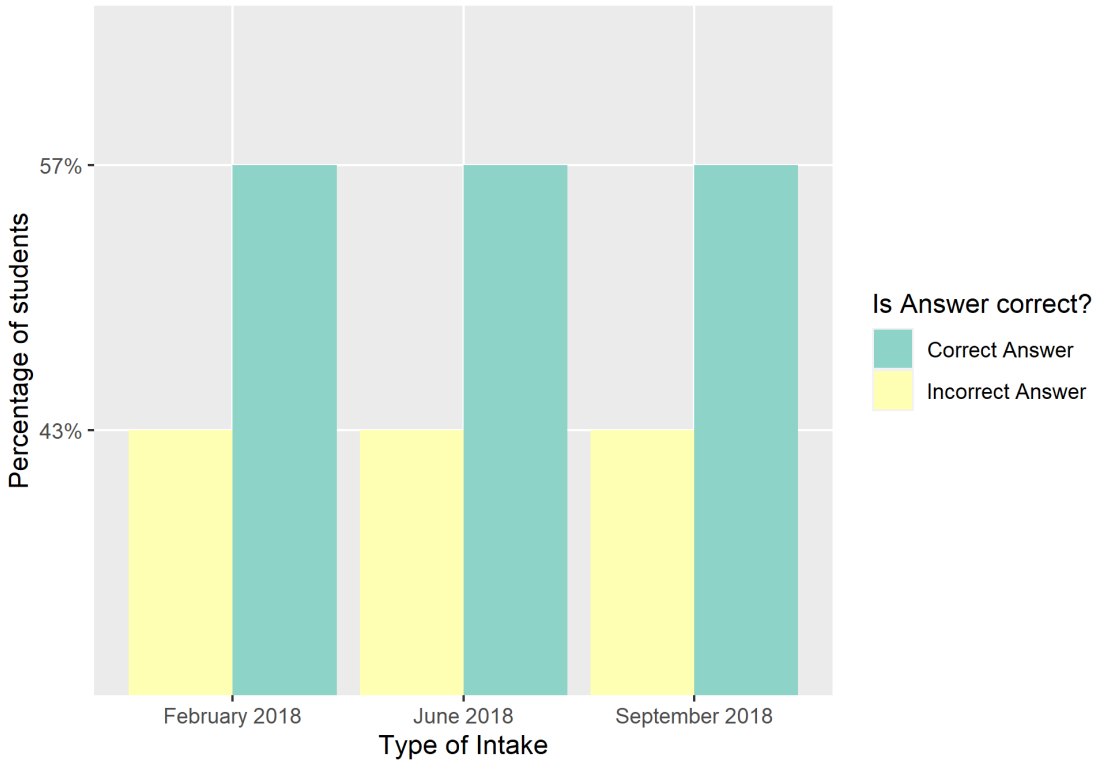

Figure 5: Comparison of quiz reason survey data

number of students enrolled in online courses is higher than in traditional classes, this ratio of incorrect answers suggests that either students do not understand the course material or it is becoming difficult to cover it.

As a result of this question, there's a good chance that the course's difficulty level is one of the reasons why students drop out. The reasons for leaving the survey are analysed in the next question, which can help us better understand this relationship.

- **What was the reason for dropping out of the class?**

The data set included the reasons for abandoning the course, as well as the step and week number at which the student left. When a student chooses to leave, the student is requested to complete a survey in which he or she can select one of seven possible responses.The categorical responses were as follows.

- I don't have enough time
- I prefer not to say
- The course required more time than I realised
- The course was too hard
- The course wasn't what I expected
- The course won't help me reach my goals
- Other

To better comprehend this, the data set was presented as a pie chart, with the percentage of students giving categorical values as reasons as illustrated in Figure 6.

As can be seen in Figure 6, the two justifications "Other" and "I prefer not to say" were eliminated because they didn't specify the particular reason. However, it can be deduced from the graph that the length of the

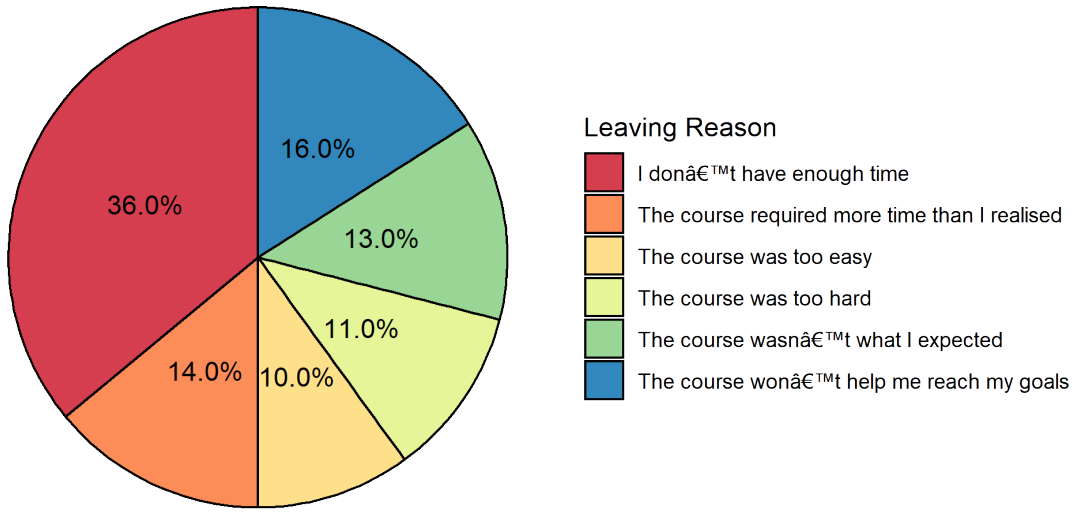

Figure 6: Comparison of leave response of students

course was the primary cause for students dropping out. The percentages on the pie chart are calculated by taking the average of all three intakes. Given that most students drop out of the course in Week 1, as determined by the first business objective, it is reasonable to suppose that students review the future course work in the introduction lectures and conclude that the course is too long for them. As a result, they decide to drop out of the class. Adding to this, 14% of students said the course needed more work than they expected, which is quite similar and points to the length of course work as a problem.

As previously observed in previous analyses of course difficulty, and based on the leaving reason of the course being difficult, it can be concluded that there is a positive correlation between difficulty and the ratio of students leaving the course. This course appears to be overly difficult, according to the second most prevalent explanation. Unfortunately, another majority appears to find the course to be either extremely difficult or impossible to complete, as seen by their abandonment of the course halfway through. Furthermore, some students felt that the course did not live up to their expectations. The general conclusion of this analysis is that the coursework needs to be enhanced, with the syllabus being kept brief, crisp, and relevant and easy to understand so that learners can engage more in the course and the participation rate rises.

Conclusion:

MOOCs, which are currently popular and have a significant role in COVID 19, have demonstrated that online courses, in addition to traditional classrooms, are an effective way of learning. Traditional classrooms and online learning, on the other hand, each have their own set of benefits and drawbacks. The goal of this research was to decipher the data provided by Futurelearn, a MOOC platform, in order to gain insight into students' engagement and performance in online learning.The study was carried out utilizing CRISP-DM methodology with the help of enrollment, leaving survey response and quiz response data.

Initially, the goal is to locate the problem in the data collection in order to establish the appropriate business aim. There was data in a set of seven where the runs were done on various intakes from various years. The enrollments were the most compelling reason to begin the analysis. Each intake and year, there was a considerable drop in enrollments. Furthermore, it was discovered that the majority of those who enrolled in the course chose to drop out, resulting in a relatively low completion rate. Before boosting enrollments, it's important to understand why students choose to drop out midway through a course despite the fact that they could study something from home. As a result, the first questions that came to mind were when and why? The first inquiry was when did the majority of the students drop out of the course after enrolling. The reason for leaving the course was the second question.

This study attempted to understand the completion rate, which turned out to be quite low, averaging approximately 1.5 percent for three intakes, while going through the analysis of the first question, which is when the majority of students abandoned the course. We studied the number of students who left the course in each week because we have the data for learners leaving week by week. The highest number of dropouts was seen in week one. It was deduced that the student used week 1 as a trial week and was likely dissatisfied with the course content for a variety of reasons. Those who choose to continue after week 1 had a considerable decrease in dropout in week 2 and a slight increase in week 3. However, because week 3 was the last week of the course, it was presumed that the majority of students had completed it. To look into it further, we looked at the student dropouts one by one, and the assumption week 3 appears to be correct. The majority of dropouts occurred in the last step, which shows that the learner completed all coursework but elected to quit the course with a survey rather than fully finishing it. Another finding was that the majority of dropouts in week one occurred in the final step of that week. Given a result, the presumption that the learner is dissatisfied with the course appears to be correct, as the final step in week 1 was to wrap up the introduction.

The next big concern was why so many students dropped out of the course. Is difficulty a role in dropping out of a course before moving on to the next question? Since the data for quiz response was available, it was determined that more than 40% of students were providing wrong answers, which is not a good performance. There's a potential the students dropped out because the course was too challenging for them, but that's just a guess. Moving forward, what were your reasons for dropping out of the class? Because there was data from the departing survey responses, an analysis was conducted, and it was discovered that the majority of students found the course to be too long, although the difficulty of the course was the second most common reason for leaving. Because we made a guess, it appears to be correct. Following this exploratory data analysis, it can be concluded that in order to enhance the number of enrollments as well as the completion rate, the courses should be maintained brief, crisp, and easy to grasp.

References

- Preidys, S.; Sakalauskas, L. 2010. Analysis of students' study activities in virtual learning environments using data mining methods, Technological and Economic Development of Economy 16(1): 94–108.
- Yu, C., Wu, J. and Liu, A., 2019. Predicting Learning Outcomes with MOOC Clickstreams. Education Sciences, [online] 9(2), p.104. Available at: https://www.mdpi.com/2227-7102/9/2/104/htm.
- Yaacob, W., Nasir, S., Yaacob, W. and Sobri, N., 2019. Supervised data mining approach for predicting student performance. Indonesian Journal of Electrical Engineering and Computer Science, 16(3), p.1584.
- Data Science Process Alliance. 2021. CRISP-DM - Data Science Process Alliance. [online] Available at: https://www.datascience-pm.com/crisp-dm-2/ [Accessed 2 December 2021].
- Ford, C., 2021. Data Wrangling in R: Combining, Merging and Reshaping Data. [online] Clayford.github.io. Available at: https://clayford.github.io/dwir/dwr_05_combine_merge_rehsape_data.html [Accessed 3 December 2021].
- Brillinger, D., Preisler, H., Ager, A. and Kie, J., 2004. An exploratory data analysis (EDA) of the paths of moving animals. Journal of Statistical Planning and Inference, [online] 122(1-2), pp.43-63. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0378375803002404 [Accessed 3 December 2021].
- Reproducible data science techniques in actuarial work. 2021. Exercise 1. [online] Available at: https://philipdarke.com/reproducible-actuarial-work/exercise1 [Accessed 3 December 2021].
- Taylor & Francis. 2021. An overview of learning analytics. [online] Available at: https://doi.org/10.1080/13562517.2013.827653 [Accessed 3 December 2021].
- Chatti, M., Dyckhoff, A., Schroeder, U. and Thüs, H., 2012. A reference model for learning analytics. International Journal of Technology Enhanced Learning, 4(5/6), p.318.