FutureLearn Data Mining and Analysis

170339779

Project Report

Business Understanding

FutureLearn, a for profit online education provider, have provided us with a large dataset to analyse and draw conclusions from. To draw relevant and meaningful conclusions for FurtureLearn we must first outline FutureLearn's business objectives. As FutureLearn is a for profit education provider it is clear that their long term objectives include the financial stability and profitability of the company. To this end we might consider that the metric most important to FutureLearn, when considering this free course, would be one that measures the number of students purchasing a completion certificate. Indeed this is one measure which we may keep track of over the course. We later learn that only 54 purchases by non-repeat students occur out of 5313 enrolled students over the 2 last runs. It is possible that costs are low enough per course that this low number of sales for a individual courses is sufficient for the company to reach a profit. However we also must consider that perhaps the course serves the purpose of enticing students with a free course and familiarising them with the FutureLearn learning platform so that in the future they might return for a paid course. In brief the course may effectively serve as a form of advertisement for FutureLearn.

Regardless, in keeping with both of the possible objectives for the course is the need for the course to engage students. We assume an engaged student is more likely to return to FutureLearn in the future and purchase another of their courses and, importantly for the data we have, is more likely to purchase a certificate at the end of the course. By measuring engagement in a step dependent way we would be able to determine if any steps are failing to engage students. Additionally, we should identify subgroups so that differences between them might be viewed. To ensure that the conclusions are as relevant to the company going forward as possible we will only consider the last 2 runs of the course (run 6 and 7) as the course noticeably changes over the 7 runs provided. Therefore our plan is two use run 6 and 7 data to find useful estimates of engagement.

Data Understanding and Data Preparation (Business Understanding relavant)

```
n_for_sentiment#prints number of experience ratings
```

```
## [1] "180 responses to the weekly sentiment survery"
```

There are various ways which we considered measuring engagement using the datasets available. One useful measure may have been weekly sentiment scores, however the number of respondents for each week was determined to be insufficient (n=180) across the two datasets being used. Additionally it experience ratings tended to be more positive than negative and so the number of responses which could have been used for improvement purposes (i.e scores of 1) were even smaller than the number of responses. In the future FutureLearn could consider giving a wider scale from 1 to 10 so that more variation in responses could be recorded. With this in mind we decided to define engagement by its opposite, disengagement. Completing no more of the course was considered to be a lack of engagement, so was unenrolling and a lack of step

completion (second cycle). Using these measures in combination with retention time (n= 2406, calculated by working out the time difference between enrollment and completing their last step) provided enough usable data to perform our analysis. Retention time was considered a measure of engagement with greater values being an indicator of further engagement. We consider 21 days of retention to be and important boundary as this is a 3 week course. We also initially attempt to use metrics from the question response dataset to provide an estimator of engagement.

To achieve this we prepared our primary dataset called uni_ids (See Summary 1) which is composed of information from the following datasets: enrollments, step activity and question response. Dates were converted to numerical form for handling them. The numerical form is the distance in seconds from 1/1/1970. The pass column is True if the student has a pass date and False if the student has an unenroll date. the pass column is NA if the student has neither a pass nor a unenroll date. The mean refers to the number of correct answers answered in the quiz questions divided by the total number of answered questions. The Q_count is the number of Quiz questions answered. The question score is the number of quesetions answered correctly. The last step complete refers to the last step completed (maximum date step completed date for each student that completed a step). The letter code is that last step completed converted into letters so that it can be sorted with 1.1 becoming "aa" and 2.1 becoming "ba" (each number indexes the alphabet). Retention time in days is the retention time in days (retention time in seconds/86400). Country refers to the detected country of the student. Week_number refers to the week in which they completed their last step and question score is simply the number of right answers a student achieved. The purchase, na_pass and retention_TF variable was added into the preprocessing during the second cycle and so are ignored throughout the first cycle but are boolean and booleanesque variables. NAs are retained in uni_ids but are not considered when graphs are plotted unless stated (NAs removed). Later in the second cycle na_pass was created which identifys NAs as FALSE from the pass column as they did not fully participate. Figure 1 shows that the retention time in days is heavily skewed to the right in the mold of something like a poisson distribution which is what we might expect for this data. The data for retention time presents no immediate quality issues, although it is somewhat unexpected that some students completed steps past 80 days. Repeat students who are in run 6 and 7 are removed from uni_ids and analysed separately.

```
hist(uni_ids$retention_time_days, xlab="Retention Time in Days"
     , main="Histogram of Retention Time in Days")#Histogram plotted
```
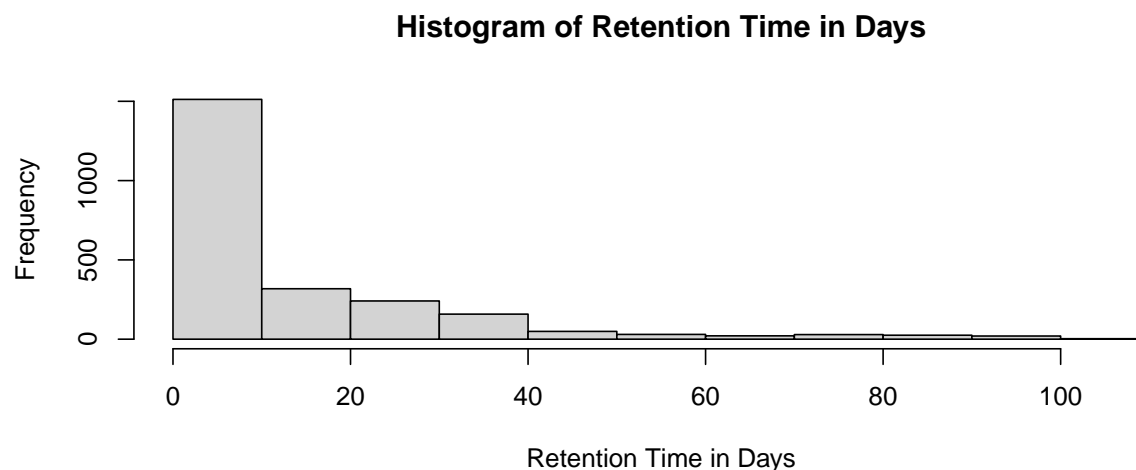

Figure 1: Histogram showing retention time in days for n=2046 students.

```
summary(uni_ids)#summary of uni_ids
```

```
##    learner_id          pass_date              purchase
##  Length:5313      Min.    :1.529e+09   Min.    :2018-04-11 10:00:57
##  Class :character 1st Qu.:1.535e+09   1st Qu.:2018-07-13 16:20:17
##  Mode  :character Median :1.539e+09   Median :2018-08-15 18:32:38
##                   Mean    :1.537e+09   Mean    :2018-08-11 21:45:17
##                   3rd Qu.:1.540e+09   3rd Qu.:2018-09-20 16:57:37
##                   Max.    :1.541e+09   Max.    :2018-10-29 11:57:40
##                   NA's    :5245       NA's    :5259
##  unenroll_date        pass           mean           Q_count
##  Min.    :1.524e+09  Mode :logical  Min.    :0.000  Min.    : 1.00
##  1st Qu.:1.532e+09  FALSE:267      1st Qu.:0.500  1st Qu.: 8.00
##  Median :1.535e+09  TRUE :68       Median :0.611  Median :13.00
##  Mean    :1.535e+09  NA's :4978     Mean    :0.621  Mean    :15.53
##  3rd Qu.:1.538e+09                 3rd Qu.:0.750  3rd Qu.:21.00
##  Max.    :1.541e+09                 Max.    :1.000  Max.    :63.00
##  NA's    :5044                      NA's    :4046  NA's    :4046
##  last_step_completed letter_code      week_number    date_of_last
##  3.20   : 420        Length:5313      Min.    :1.0   Min.    :1.529e+09
##  1.1    : 256        Class :character 1st Qu.:1.0   1st Qu.:1.531e+09
##  1.2    : 225        Mode  :character Median :1.0   Median :1.534e+09
##  1.3    : 217                         Mean    :1.6   Mean    :1.535e+09
##  1.19   : 133                         3rd Qu.:2.0   3rd Qu.:1.538e+09
##  (Other):1155                         Max.    :3.0   Max.    :1.541e+09
##  NA's   :2907                         NA's    :2907  NA's    :2907
##  question_score   retention_time_days   country         na_pass
##  Min.    : 0.000  Min.    :  0.0001  Length:5313       Mode :logical
##  1st Qu.: 6.000  1st Qu.:  0.0102  Class :character   FALSE:5245
##  Median : 7.000  Median :  2.3311  Mode  :character   TRUE :68
##  Mean    : 8.896  Mean    : 12.4554
##  3rd Qu.:13.000  3rd Qu.: 19.1052
##  Max.    :22.000  Max.    :109.2686
##  NA's    :4046   NA's    :2907
##   retention_21    purchase_TF
##  Min.    :0.0000  Mode :logical
##  1st Qu.:0.0000  FALSE:5259
##  Median :0.0000  TRUE :54
##  Mean    :0.2265
##  3rd Qu.:0.0000
##  Max.    :1.0000
##  NA's    :2907
```

Summary 1. A summary of all of the data collected into the primary dataset used throughout. The data frame uni_ids is named after its 1st column filled with learner ids.

Modelling/Deployment ## Step Analysis
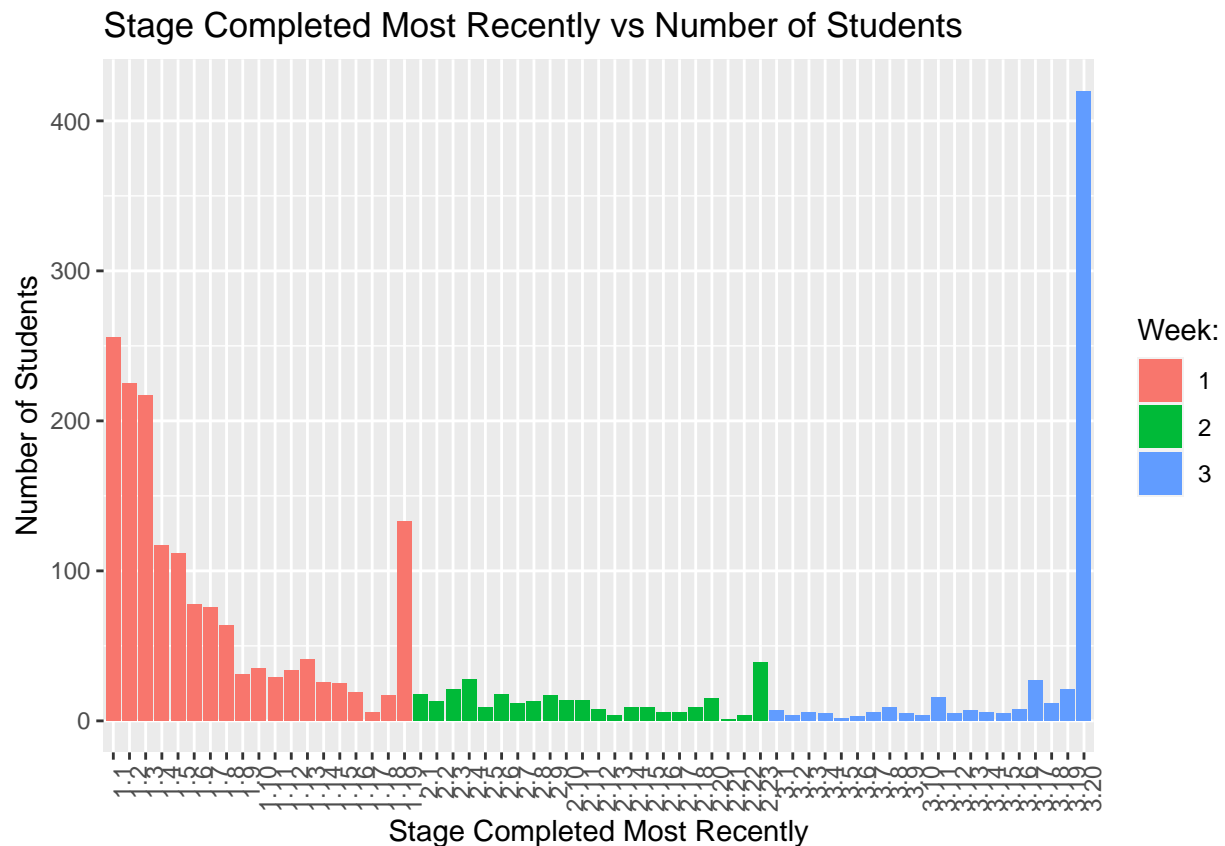
```
stage_complete #plot graph
```


Figure 2: Barchart Graph showing the stage completed most recently with the height of the bars the number of students that completed that step last

FutureLearn have an interest in students finishing the course so that they might opt into buying the certificate at the end of the process. In order to complete the course a student must complete the pre-requesite steps over the 3 weeks. To see how far students get in the course we plotted a bar chart of the course steps against the number of students who's most recently completed stage was that step (See Figure 2). We assume that the last step completed is the furtherest step along that the student completed. It is worth noting that week 1 appears to have the most people continue no further. Additionally there is a spike at the very end of the 1st two weeks indicating that after that week the student completed no more steps. FutureLearn could focus their efforts on smoothing over the periods between the weeks as much as possible so that students are ready to rejoin the following week. This could be achieved by reminding the students of the upcoming week and next load of steps. However, an alternative method might be to experiment with where the hardest parts of the workload fall for each week. In other words FurtureLearn could investigate as to whether putting the hardest part of the coursework in the middle part of the steps for a given week prevents the spike at the end of the week.

Alternative Measures of Engagement

```
knitr::kable(correlation_matrix, caption=corr_caption)#table creation
```

Table 1: A correlation matrix of the whether or somone fully participated (pass), the number of right questions they answered (question_score), the number of questions they answered (Q_count) and the number of questions they got right out of the number of questions they answered (mean).

	pass	question_score	Q_count	mean
pass	1.0000000	0.7927972	0.7789884	-0.3480387
question_score	0.7927972	1.0000000	0.8521685	-0.2857244
Q_count	0.7789884	0.8521685	1.0000000	-0.6260739
mean	-0.3480387	-0.2857244	-0.6260739	1.0000000

Having considered how the individual steps correlate to student progression, we decided to investigate how other variables correlate to finishing all the steps (pass). Although only dealing with n=68 Table 1 shows that Q_count may positively correlate to finishing the course. This would be expected as people who finish the course answer more questions. That this is not as strong a correlation as it could be indicates that some people who pass the course may not have answered more questions than those who stopped. The small negative correlation between mean and pass may be due to those who finished only a on question and got it right (thus mean=1) and then did not finish any more of the course. Therefore it is clear that mean (quiz percentage of correct answers) would not be a good metric for identifying engagement as it cannot distinguish those who answer many questions and score well from those who score well and answer many questions. We the attempted to visualise part of the results in Figure 3. Figure 3 is not very useful as it can be expected that those who completed the course answered more questions but it does confirm this. It might be useful to consider the number of questions answered in a certain amount of time for a model predicting participction but this is not an objective of ours. Ultimately, we consider these variables uninformative as they do not provide useful insight for FutureLearn or explain more than might have been expected upon further consideration. This may be due to the manner in which we decided to approach and consider these variables.

```
mean_vs_Q_count#plot graph
```

Video Step Analysis

```
plot(rep(c(5,10,25,50,75,95,100),13), t(cyber.security.7_video.stats[,9:15]),
    type="p", col=1:13, ylab= "Percentage still Viewing- %",
    xlab = "Percentage of Video Watched- %",
    main= "7th Dataset Video Stats Watchtime vs Students Still Watching")
```

```
#plot graph and setup
```

```
video_boxplot#plot graph
```

Some of the steps monitored in Figure 1 are videos that the students can watch. Video steps perform well in Figure 1 (no radical peaks) except for step 1.1 (a Welcome video) which is likely not the fault of the video
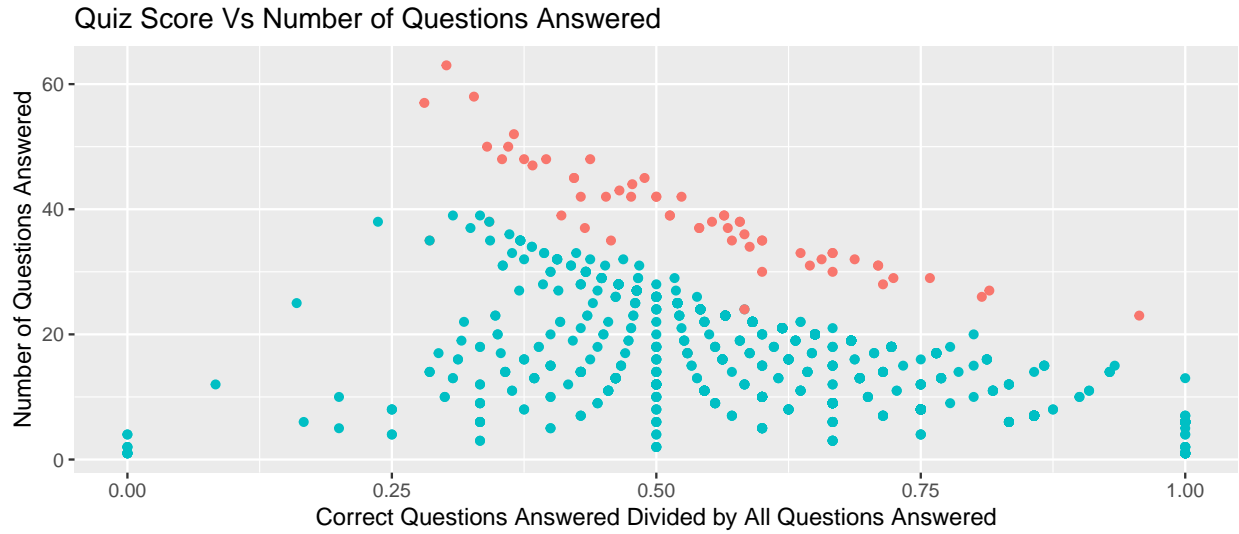
Figure 3: Plot of percentage right answers out of all answers vs the number of questions answered. Coloured in Red are people who fully particpated whilst those in blue did not. Students who did not answer any questions are not considered but those who had neither a full participaation date nor a unenrollment date are considered to have not fully participated.
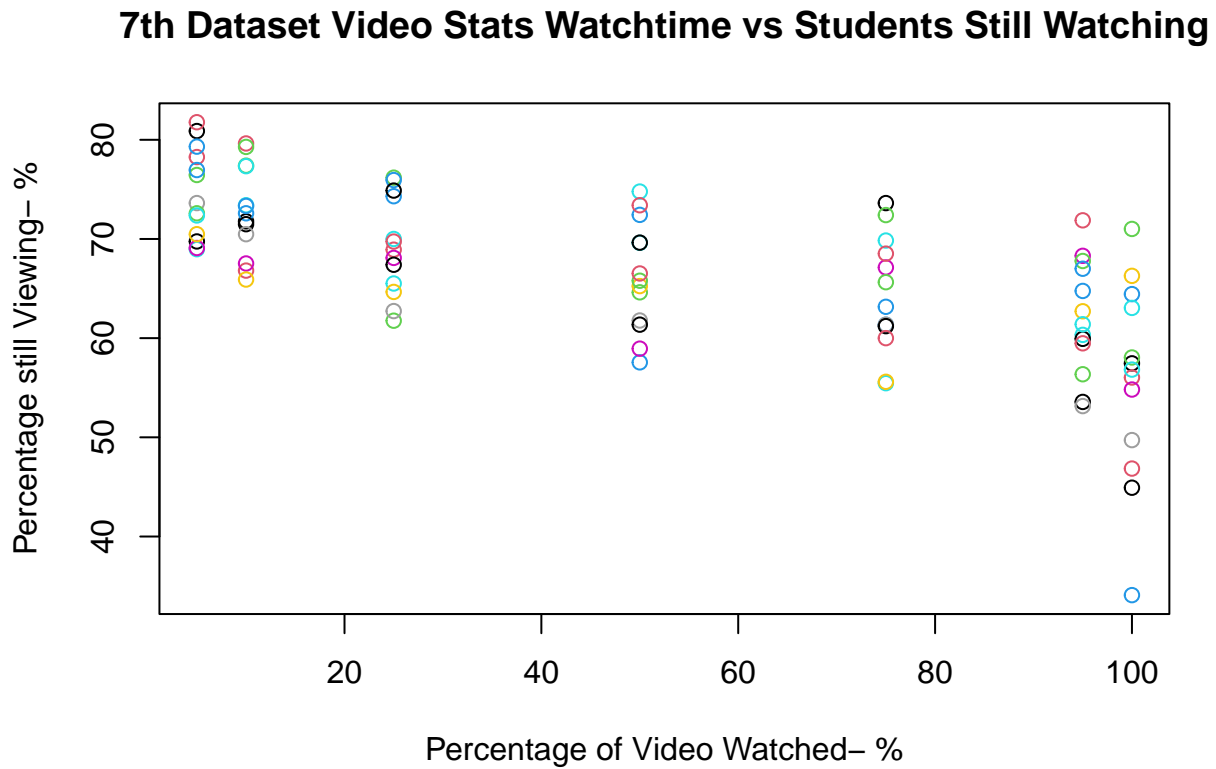

Figure 4: Plot of Percentage of Video shown vs Percentage of students still Watching, colours vary for differeent videos

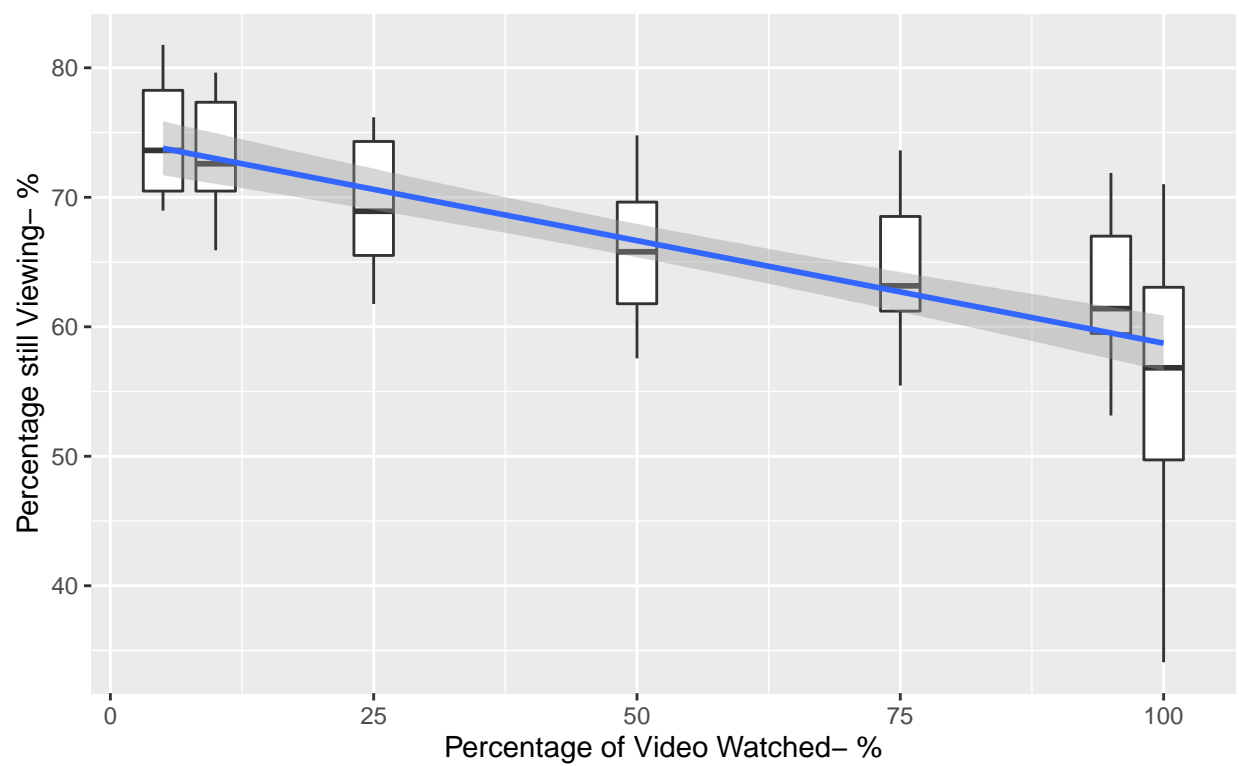7th Dataset Video Stats Watchtime vs Students Still Watching Boxplot and Linear Regression

Figure 5: Boxplots of viewership at specific proportions of the video. A line is fitted by linear regression to show the general trend.

itself but part of a rapid downward trend in the first week. However, it is worth examining the videos to ensure that no particular video radically underperforms. Figures 4 and 5 show that as the videos continue the range for people still watching the videos increases. There is a noticeable downward trend across the graph especially after the first 10% and last 95% of the video. The first 5% also has a drop from 100% which is not made obvious on the graph. A drop before the last 5% could be the that students feel that the relevant material has already passed. A drop after the first 5% could mean that students were not engaged enough at the start or felt that it is irrelevant. In total the no video seems to be performing extremely poorly when compared to he other videos for the first 95% of the video, although the last 5% shows greater range of range of viewer dropout. It might be worth considering which videos perform poorly in their last 5% and either shorten them or consider if the last 5% is contributing much.

Repeat Students

We note in the data preparation and understanding segment that the same students can be found in runs 6 and 7 (repeat students). We assume that these repeat students have taken the course at least twice. They have not been included in our analysis so far and so now we seek to understand what happened to these students and whether they represent a lack of engagement or an extreme engagement. 5 repeat students fully participated in run 6 one of which fully participated again in run 7. It seems that the repeat students finished the course the first time at a rate greater than the non-repeat students which might be expected if they liked the course enough to complete it the first time and then enroll again. A greater proportion of repeat students unenrolled in their first attempt than the single attempt sample. Perhaps due to not enough time in the first attempt and then trying again the second time once they thought they had time. However, nobody who had not already finished the course in their first attempt finished the course in their second attempt suggesting that whatever the reason their first attempt ended prematurely may have also ended their second attempt. That 5 finished the course out of 102 (4.9%) is impressive when compared to the non repeat students which might (a statistical test could verify significance) indicate that these are more motivated individuals. The reason for leaving the first and second time may be found in the leaving survey. We analyse these reasons in Table 2 and 3.

```
cat(statmentm2, "\n", statementm1, "\n", statment0, "\n", statement1,
    "\n", statment2, "\n", statment3, "\n", statment4, "\n", statement5,
    "\n", statment6, "\n", statment7, "\n", statment8, "\n", purchases, "\n")
```

```
## 102  Repeated their studies (total)
##  1  Completed the course twice
##  13 Unenrolled in their first and second attempt
##  5 fully participated the first time
##  1 fully participated the second time
##  28 unenrolled the first time
##  15 unenrolled the second time
##  74 Assumed did not finish the course the first time
##  87 Assumed did not finish the course the second time
##  1.28% of the non repeat students finished the course
##  5.06% of the non repeat students unenrolled from the course
##  62 Purchases of a certificate including 2 repeat students who bought twice
```

```
#stetements to be printed
```

Summary 2: shows a summary of the repeat student data with information from the non repeat students for comparison. Also includes Purchase from all students. We assume students did not finish if they lack a fully participated date (empty values). ## Leaving Reasons

```
leaving_reason<-left_join(repeat_students,
                          cyber.security.6_leaving.survey.responses,
                          by=c("learner_id"))
knitr::kable(leaving_reason[!is.na(leaving_reason$leaving_reason),
                            "leaving_reason"],
caption="Table of leaving reasons for repeat students in the 6th datset (their first monitored attempt)")
```

Table 2: Table of leaving reasons for repeat students in the 6th datset (their first monitored attempt)

leaving_reason
I prefer not to say
I prefer not to say
I prefer not to say
The course was too easy
I don't have enough time
I don't have enough time
I don't have enough time
I don't have enough time
I don't have enough time
I prefer not to say

```
#creates table of leaving reasons
```

```
leaving_reason2<-left_join(repeat_students, cyber.security.7_leaving.survey.responses, by=c("learner_id
knitr::kable(leaving_reason2[!is.na(leaving_reason2$leaving_reason), "leaving_reason"], caption = "Table
```

Table 3: Table of leaving reasons for repeat students in the 7th datset (their second monitored attempt)

leaving_reason
I don't have enough time
Other
I don't have enough time
Other
I don't have enough time
I don't have enough time
I don't have enough time
I don't have enough time
I don't have enough time
I prefer not to say

```
#creates table of leaving reasons
```

Half of all the replies from repeat students in Table 2 say that they did not have enough time for the course as we suggested previously. Only one says that the course was too easy and the others preferred not to say. For the second attempt it seems that once again the take away from the repeat students is that they left because they did not have enough time. Repeat students may require a less intensive course that could be better managed with a shorter amount of free time. Therefore, we can conclude that a large portion of the

repeat students are highly engaged but lack the time to finish the course, it may be worth extending the course for these students and spreading out the workload to aid completion rates. We visualise the leaving reasons for single attempt students in figure 6 and Table 4.
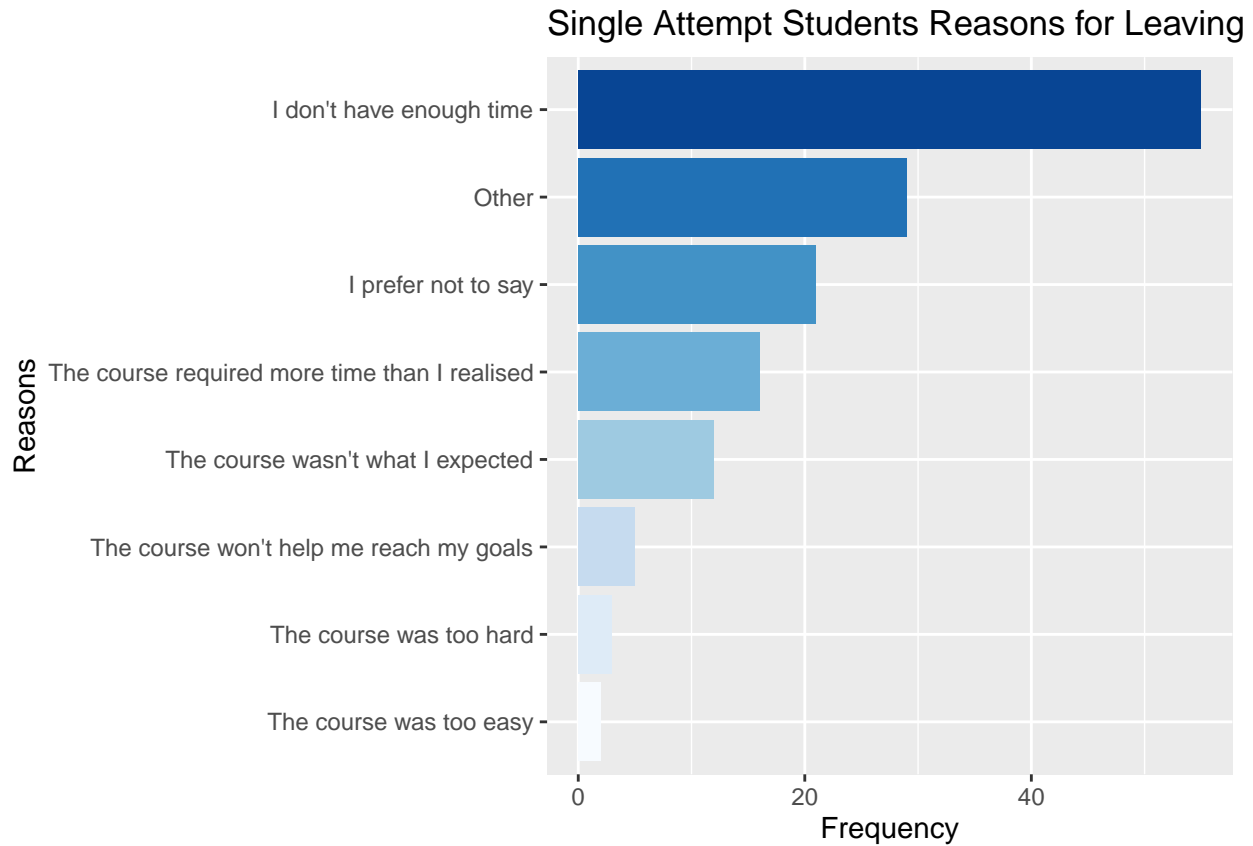
```
single_reason #plot graph
```

Single Attempt Students Reasons for Leaving

Figure 6: Barchart showing the reasons students who attempted the course only once gave for leaving the course.

```
knitr::kable(general_leaving_reason, caption="The reason and frequency that reason was given for leaving
```

Table 4: The reason and frequency that reason was given for leaving the course by single attempt students (NAs have been removed).

reasons	frequency
NA	0
The course was too easy	2
The course was too hard	3
The course won't help me reach my goals	5
The course wasn't what I expected	12
The course required more time than I realised	16
I prefer not to say	21
Other	29

10

reasons	frequency
I don't have enough time	55
Total	143

```
#creates table
```

The non-repeat students frequently state that their reason for leaving was that they did not have enough time much like the repeat students. As there are many more non repeat students this provides stronger evidence for another longer and more spread out course. Alternatively FutureLearn could aim to set expectations with students early so that they can expect to invest more time and prepare accordingly. The "Other" group likely consists of multiple reasons for leaving more specific to the student. The course requiring more time than was realized was the reason given by 16 of the 143 respondents (11%). This may further indicates that expectations for the course time-wise need to be managed early which may even encourage those who would drop out otherwise to be retained for longer. It may also be worth trimming videos that fail to retain viewers after a certain amount of time so that the student does not feel that their time is being wasted. Reassessing which content is vital or could be streamlined could also help the students feel like the time they spend is worthwhile whilst simultaneously decreasing the time required to work on the course.

Retention Time

```
pass_graph#plots graph
```

Now that it is clear why students are leaving (not engaging) we analyse our last measurements of engagement retention time in Figures 7 and 8. It is clear that retention time will correlate with whether or not an individual has a date for fully participating in the course (n=67) or an individual does not have a date for fully participating(n=2339) as shown in Figure 7 (students had to have a retention time to be included). Figure 7 shows that comparison but it also shows that there are individuals who did not achieve a fully participated date and were retained for longer than those that did fully participate. Perhaps this is due to a lack of time leading them to return again and again but at larger intervals than the students who had enough time to fully participate. The median for retention time in days for the fully participated category of students is 22.88 days (2d.p, Interquartile range: 26.49 2d.p), which is close to the 3 full weeks so is not unexpected. Those that did not have a date for full participation had a median of 2.06 days (2d.p, Interquartile range: 18.35 2d.p). Median is selected here as the distribution appears skewed. Overall this line of investigation only goes to confirm what one could reasonably expect from the data. Although Identifying why highly retained individuals still did not achieve a fully participated date may be a reasonable line of further investigation. Additionally at least one individual scored incredibly low in retention time (minimum 0.00782 days 3s.f) for the fully participated group, which may indicate an issue in the quality of our data. We should designate this individual as an outline in our dataset. Additionally our sample size for those who fully participated is very small and it may be better to only consider retention time.

Country Group Analysis

```
student_country_plot#plots graph
```

Figure 8 shows a basic comparison between those detected to be International students and who were detected to be in Great Britain. To clarify this measure assumes that the detected country will be the same as the country of origin for that student which is most likely not the case for all students. However, using detected country of origin instead of reported country of origin can give us a more complete view as
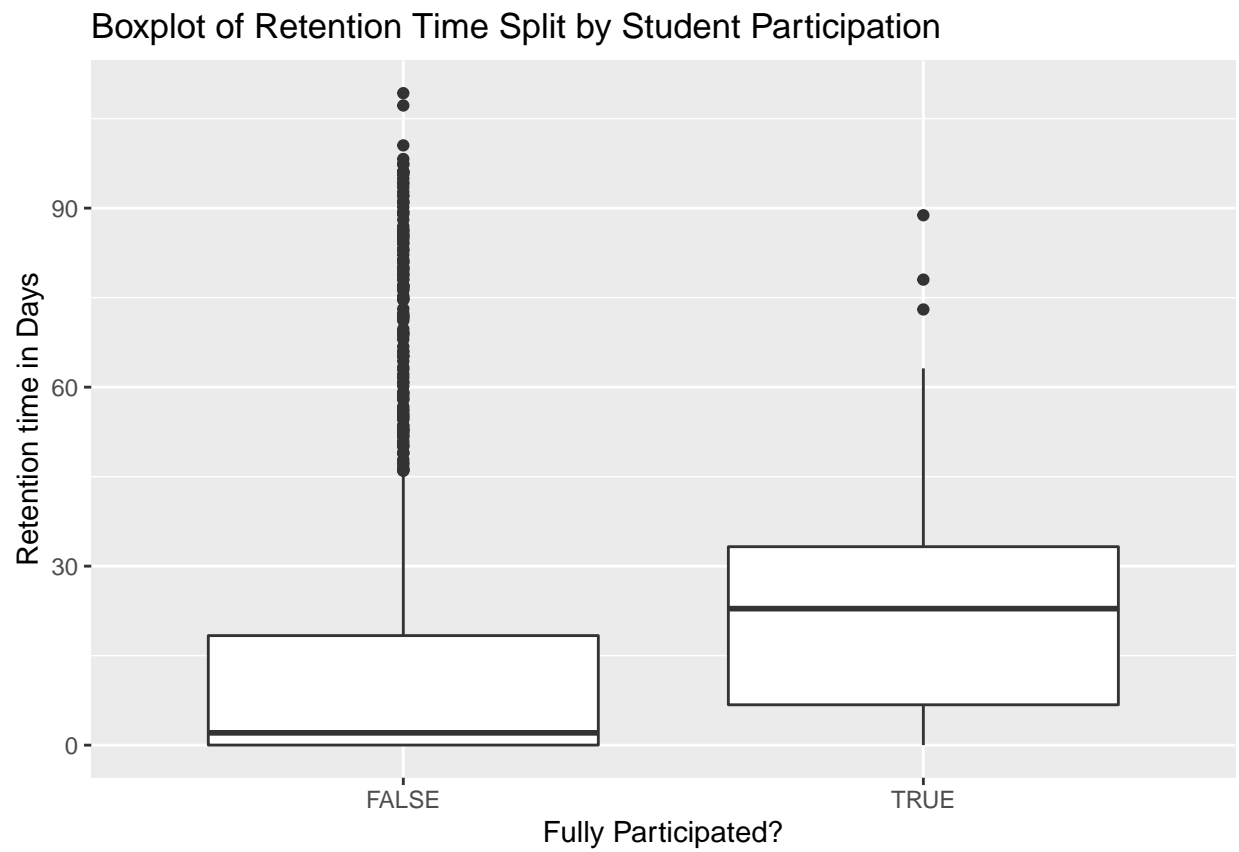
Figure 7: Boxplots showing retention time of students who fully participated in the course against those who did not
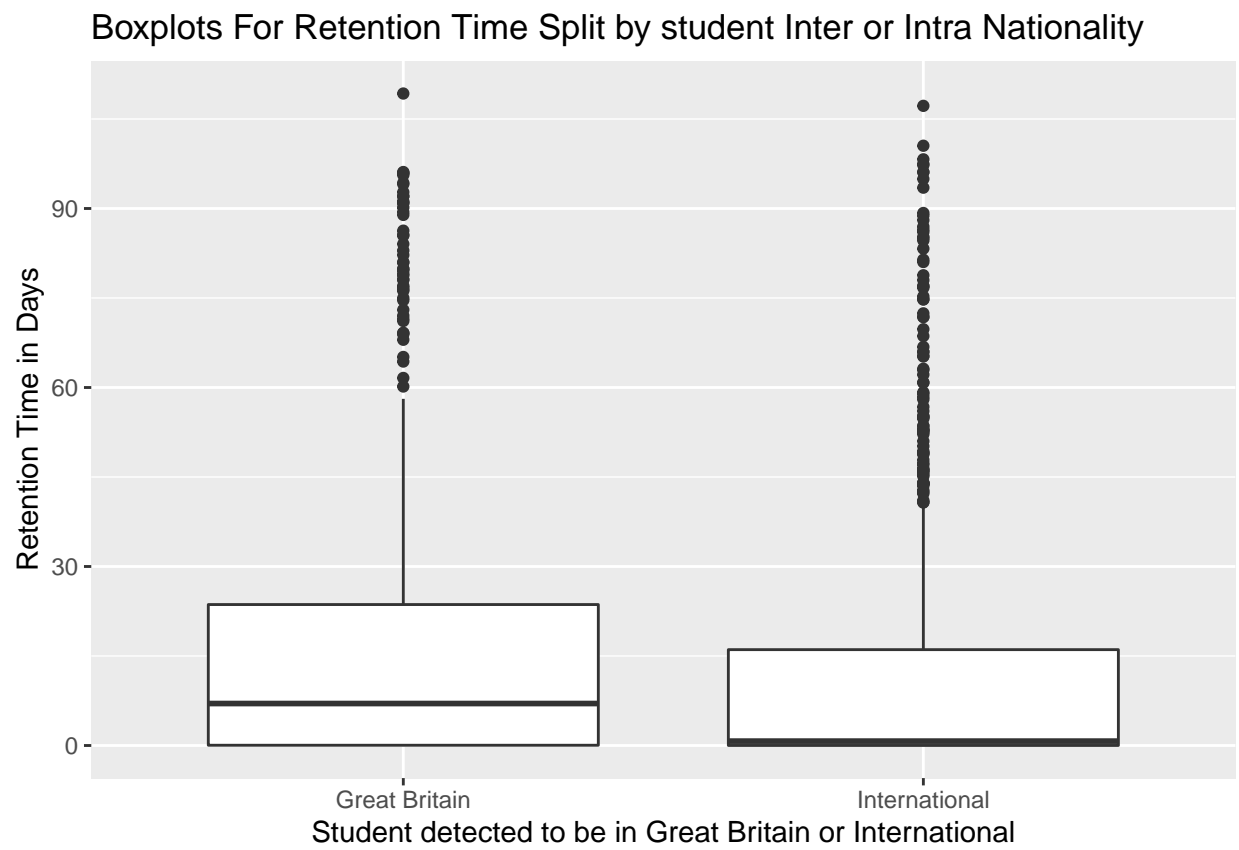
Figure 8: Boxplots of retention time for Great Britain and International students

the many people did not provide their country. All further analysis will assume that, whilst virtual private networks and simply being in a different country could change the detected country, the average student's country of origin will be that of the detected country. Therefore conclusions should be viewed with this assumption in mind. There was a lower median retention time for detected International students (n=1600) than those detected to be in Great Britain (n= 784). Therefore, engagement is greater for students within Great Britain. The aim of this analysis is to find ways that FutureLearn can better cater to all students and how they might better support those students. Therefore, further analysis into why the median is much lower for international students might be beneficial. Comparing majoritively English speaking countries to other countries may allow for FutureLearn to determine if a language barrier plays a role in the lower median for International students.

Second Cycle

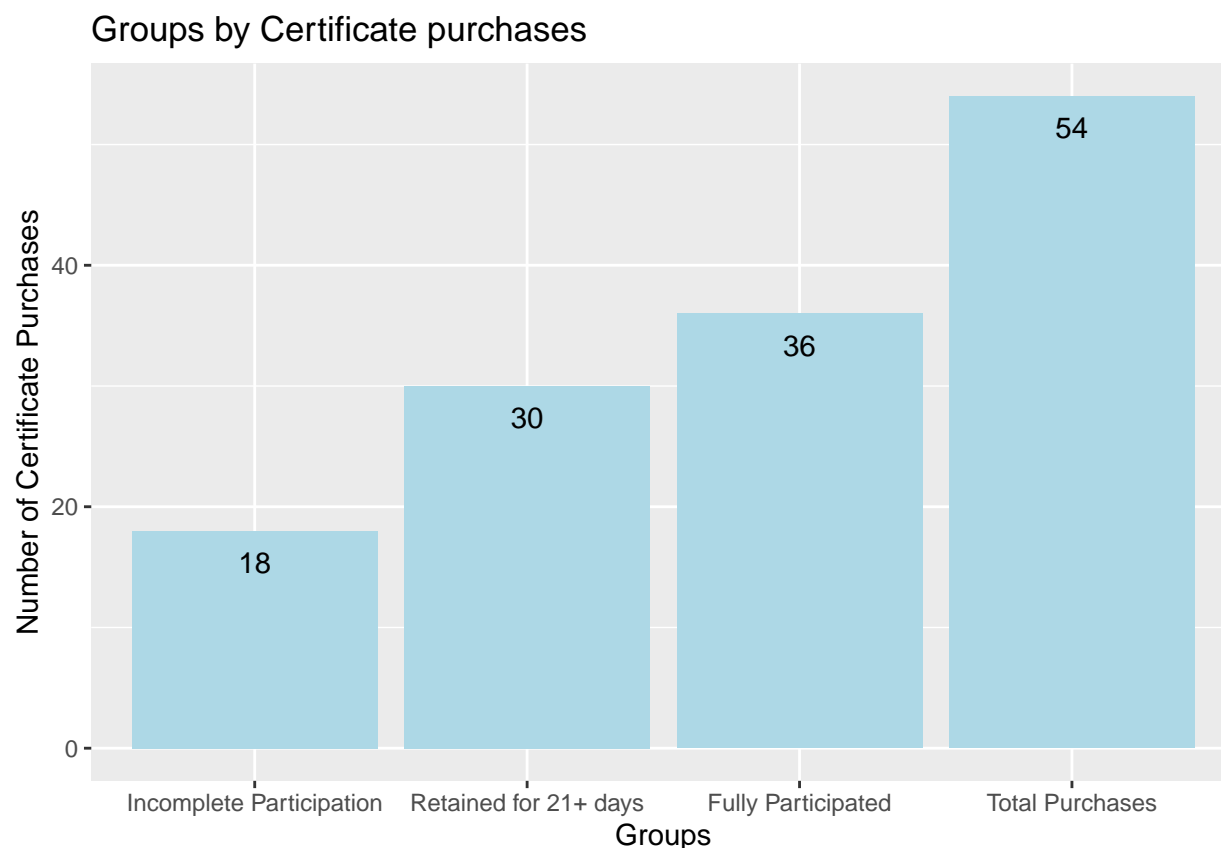### Purchases

```
purchase_bar#plots graph
```


Figure 9: Barchart showing certificate purchases by Group for Non-Repeat Students

For all the analysis so far we have assumed that purchasing the course certificate was more likely if the individual finished the course. Before continuing we should confirm that assumption. We prepared participatin data on the non repeat studetnt to analyse this. With the limited dataset currently being used we show

that twice as many people who fully participated in the course purchased the certificate (Figure 9). We also showed that those with a retention time of 21 days or more made up over half of the people who bought the course. As we only utilise a dataset with only 54 people purchasing the certificate the proportions found here could be non representative of which people will purchase certificates in the future. However, these categories are very likely to represent an increase in the likelihood that a person buys the certificate as determined by logistic regression (see Table 5).

```
knitr::kable(probs_data, caption= "Output from 2 separate logistic regressions given that a student full
```

Table 5: Output from 2 separate logistic regressions given that a student fully participated in the course or was retained for 21 days or more, against wheather or not they purchased the certificate (1 being a purchase, 0 being no purchase)

Condition	Probability When Not	Probability Given Conditon
Fully Participated	0.0034318	0.5294118
Retained >= 21 days	0.0102096	0.0550459

Table 5 shows the result of 2 separate logistic regressions over the two conditions. The probabilities given were found with the logit link function (lib/helpers.R). Both logistic regressions were shown to be significant at an alpha of 0.001. The probability of purchasing the certificate given that the student fully participated was over 0.5 and would be considered high enough in classification models to classify that person as a purchaser. while that was not the aim of this analysis it is clear that both of these conditions have an impact on whether or not a student would buy the course. The much lower increase in the probability of students purchasing the certificate after 21 or more days of retention is likely due to the fact that many students were retained for over 21 days that did not then purchase the certificate.

Alternative View on Step Analysis

```
steps_completed_chart#plot graph
```

Figure 2 showed the last step students managed to complete, but this does not show the entire story. Individuals could complete any combination of steps and therefore some information could be lost by only reporting the last step completed. We prpared data from the steps dataset where we counted the numbe of times that step was completed and the number of times it was started. Here we see that the dropout after each week is real but are reminded that compared to the large number of students to begin with this is a rather small drop off each time. The most consistent week appears to be week 3 where all bar one of the steps has a relatively stable number of completions and is confirmed by our first graph. Step 3.18 is severely underperforming perhaps due to it being the only clearly marked test on the course. Futurelearn might consider rebranding step 3.18 so that it is called a Quiz as other quizzes such as step 2.8 do not see such a dramatic decrease in participation (although it does see a relative dip in completions compared to how many people started it). The same slight decrease in completions occurs on step 3.11 another quiz. If step 3.18 is a particularly long step they might consider breaking it up across the third week to get more people to finish the whole thing. It is not unexpected that many people do not finish the first step as their is a high dropout rate in week 1 as shown in the first graph.

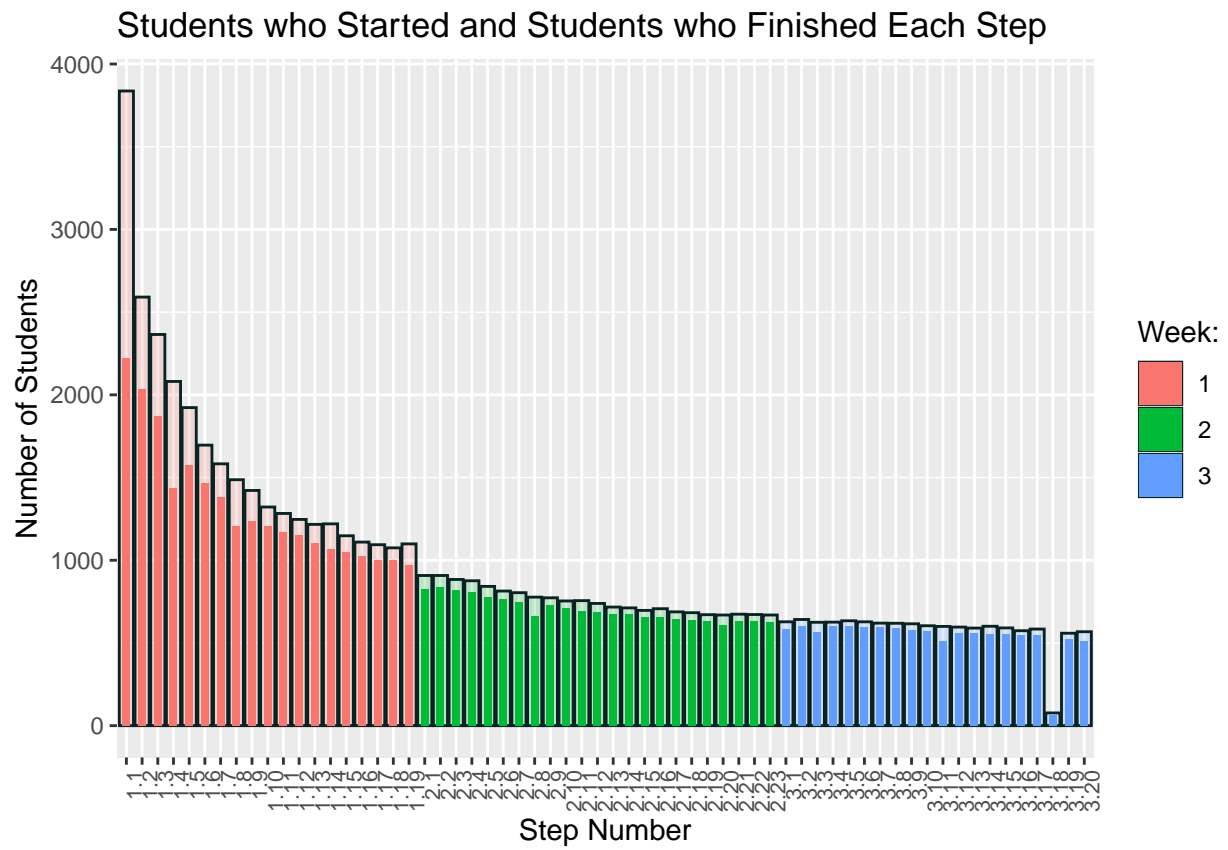## Drilling into Country Group Analysis

Figure 10: This graph shows the number of students that completed a task (dark colours) out of the number of students that started the task (light coloured and outlined in black) n = 3892.

```
student_language_plot#plots graph
```

Retention Time by Whether the Student's Detected Country Primarily Speaks English
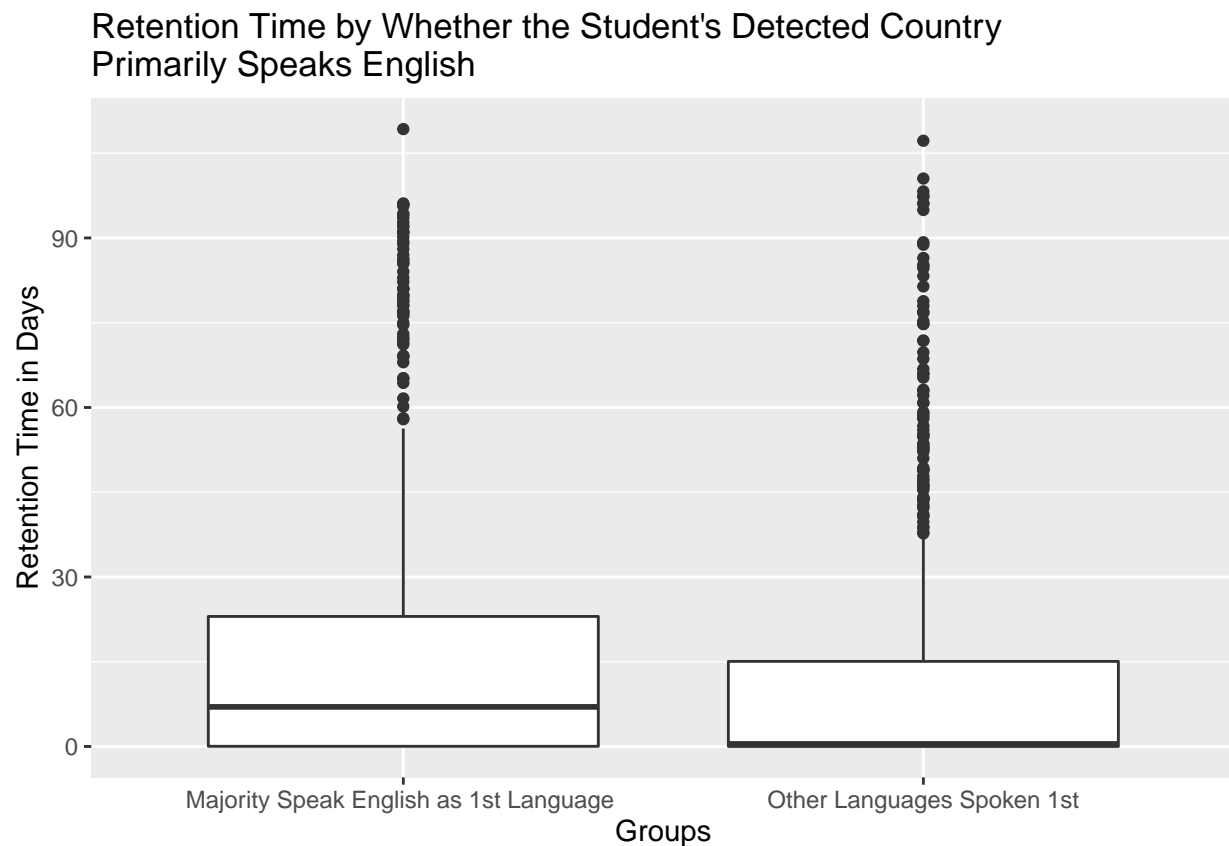

Figure 11: Boxplot showing retention time for those who were detected to be in countries that majoritively speak English as a first language and those who were detected to be from countries that did not majoritively speak English as a first language

The plan for this report includes a search for sub groups and Figure 8 shows a difference between international students and British students (detected). Therfore, we have investigated this phenomenon further dividing between countries that majoritively speak English as a first language (MSEFL). MSEFL countries were thought to be the following: Great Britian, Canada, the USA, Ireland, New Zealand and Australia. It seems that overall students from a MSEFL country were retained for a longer median time than those from countries that did not speak English as a first language. The difference between the two groups is not so large as the median for each group appears to be in the interquartile range of the other. However information may be lost when taking such a top down view and further analysis might consider building linear regression models to determine which factors have the biggest predictive impact on retention time. Although for the scope of this report it is simply enough to be aware of this possible difference. It may also be beneficial for FutureLearn to be aware of the countries making up the majority of their course and so below we show countries with greater than 25 students on this course (Figure 12). To maximize the benefit to the students FutureLearn could consider which non MSEFL countries had the most students (Figure 13).
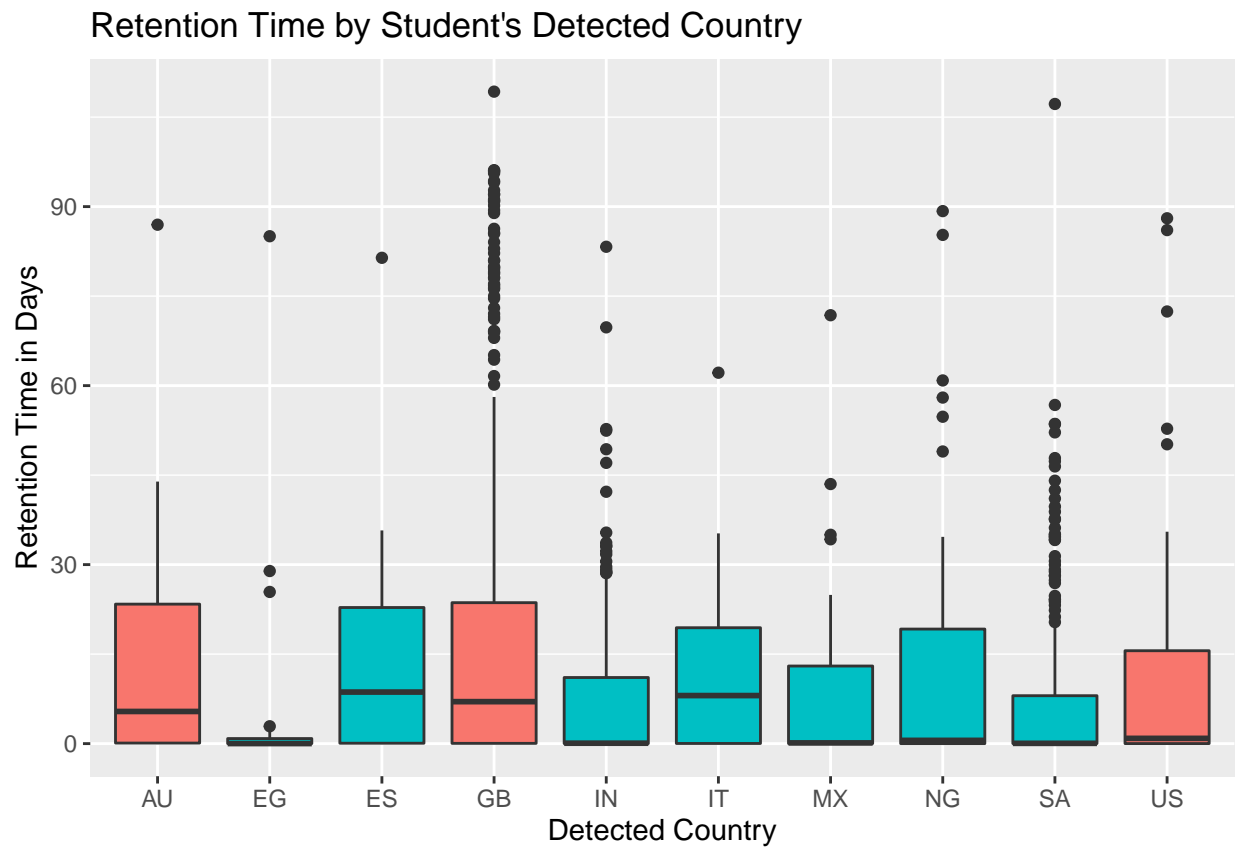
```
student_language_plot_25#plots graph
```

Figure 12: Boxplots of students retention time by the country they were detected to be in. Filled in red if thought to come from a English first language speaking majority country and blue/turquoise if not. Only countries with 25 or more students that had retention times are included in this plot

```
plot_countries#plots graph
```

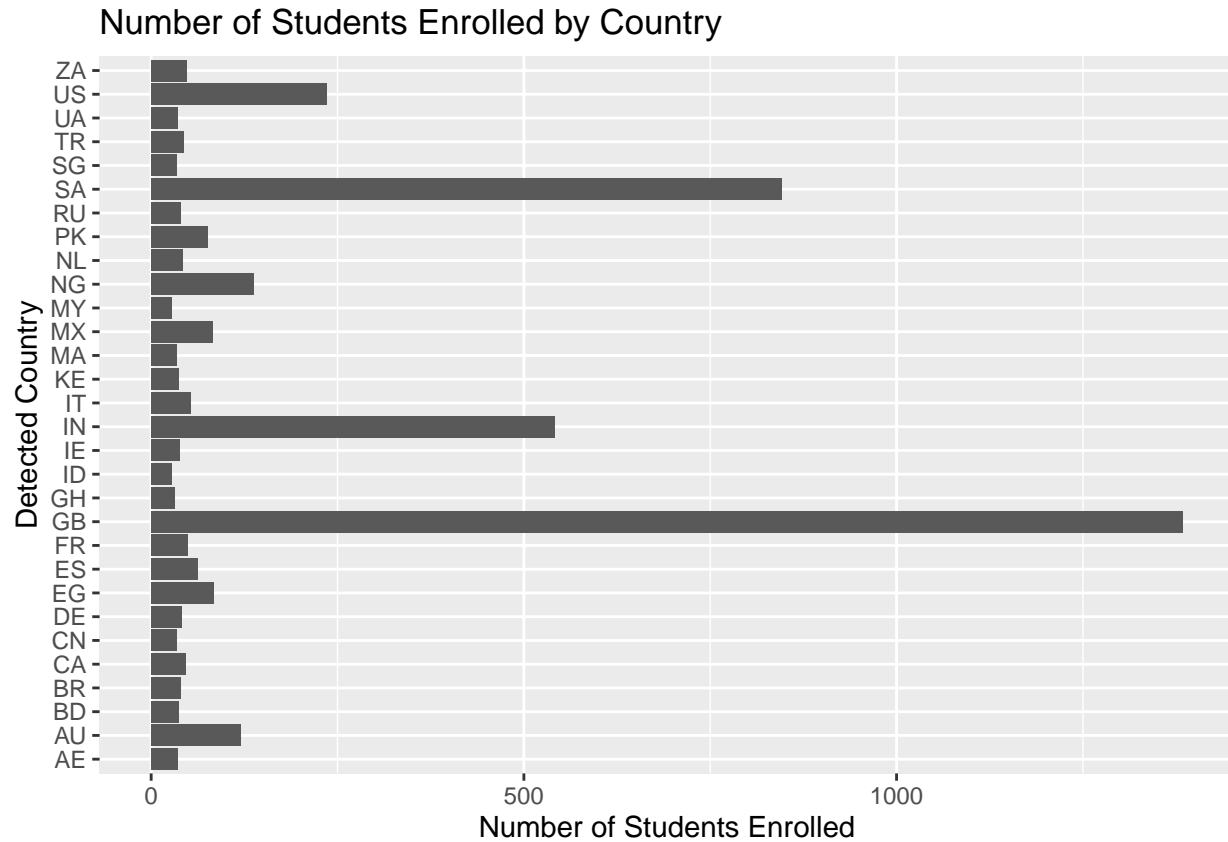## Number of Students Enrolled by Country

Figure 13: Barchart showing the number of students that enrolled in the course if that country had 25 or more students enrolled

We prepared data from countries that had 25 or more students with retention times to create Figure 12. The cutoff of 25 was determined arbitrarily, but is required to observe some spread on the boxplot. We also prepared data from countries with 25 or more students enrolled and plotted their count in Figure 13. Figure 12 shows that there appears to be a difference in retention times for MSEFL countries compared to other countries as 2 out of the 3 appear to have higher medians than most other countries countries. Notably SA (Saudi Arabia) students and IN (India) students make up the next greatest proportion of students after GB (Great Britain, see Figure 13). FutureLearn should consider this important as these two countries have poor median retention times as seen in the previous graphs. FutureLearn might consider these international students language with captions (for a lower cost solution) if they do not already or with a full dub (potentially higher cost). Additionally, FutureLearn might consider if cultural differences effect the retention times for US (United States) Students as they have a poor median retention time compared to their English speaking counterparts. The US is the 4th biggest student base. Italy and Spain have high retention median times but do not have a high number of students enrolled on the course suggesting that they are well accommodated but that FutureLearn are not reaching enough people within these countries. Perhaps some marketing towards students in Italy and Spain would provide FutureLearn with more highly retained students. Alternatively, the few students that do enroll are the most motivated from these countries. Regardless, further investigation into Italian and Spanish students should be considered by FutureLearn.

Conclusions

In this Report we have analysed the data provided by FutureLearn and suggested actions throughout that they could take to improve engagements or areas that require further investigation. We have identified a subgroups in the sample by country, by MSEFL and by repeat students that provide further insight into the barriers that face the students when completing the course. We have also identified key stages at which students may leave or not complete the course. This being the spike in final step completed after each week and the drop in participation for the test step. We have suggested that FutureLearn trim and streamline their course when possible especially when concerning their videos. We have also suggested that FurtureLearn consider rebranding their test step as a Quiz to increase the number of students participating. This rebranding should likely be deeper than just changing the label but also be a change in structure of the test so that it better resembles a Quiz. Finally, We note that due to the limited resources that FutureLearn has the most efficient ways to improve engagement might be to increase the language support for students from countries that have the most students enrolled such as Saudi Arabia.