

Project

170339779

Project Report

Business Understanding

Data provided by FutureLearn, a for profit online education provider, have provided us with a large dataset to analyse and draw conclusions from. #Clarify Business objectives here

```
summary(uni_ids)
```

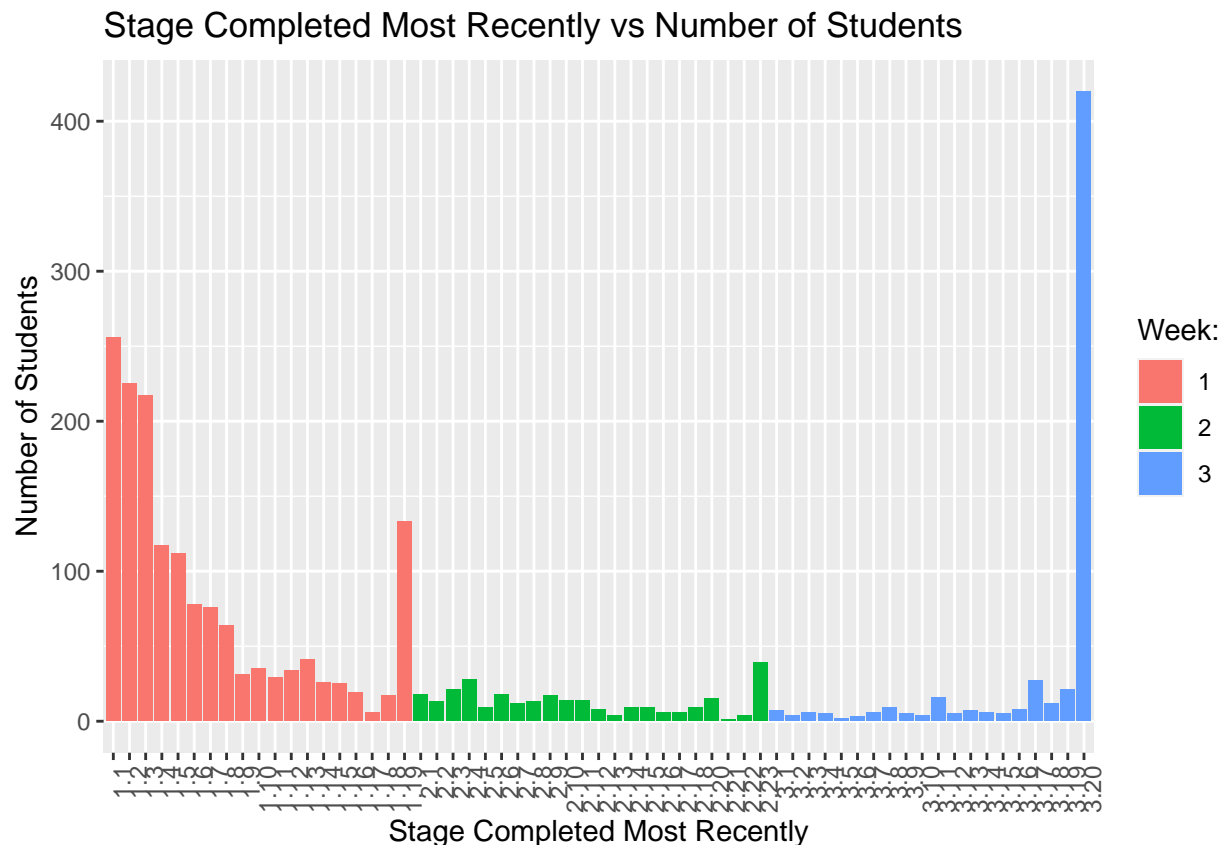
```
## learner_id      pass_date      purchase
## Length:5313     Min.      :1.529e+09   Min.      :2018-04-11 10:00:57
## Class :character 1st Qu.:1.535e+09   1st Qu.:2018-07-13 16:20:17
## Mode  :character Median :1.539e+09   Median :2018-08-15 18:32:38
##                      Mean  :1.537e+09   Mean  :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      purchase_TF
## Min.      : 0.000 Min.      : 0.0001 Length:5313 Mode :logical
## 1st Qu.: 6.000 1st Qu.: 0.0102 Class :character FALSE:5259
## Median : 7.000 Median : 2.3311 Mode  :character TRUE :54
## 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
## na_pass      retention_21
```

```
## Mode :logical   Min.    :0.0000
## FALSE:5245      1st Qu.:0.0000
## TRUE :68        Median :0.0000
##                Mean    :0.2265
##                3rd Qu.:0.0000
##                Max.    :1.0000
##                NA's    :2907
```

A summary of the data collected so far. Pass date and unenroll date have been converted to numerical to aid in the processing of data. 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 last step complete refers to the last step completed. 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). 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 variable was added into the preprocessing during the second cycle and so is ignored throughout the first cycle.

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-requisite 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 below). We used datasets 6 and 7 to achieve the figure immediately below.

```
#scale_fill_manual uni_ids$last_step_completed
stage_complete
```



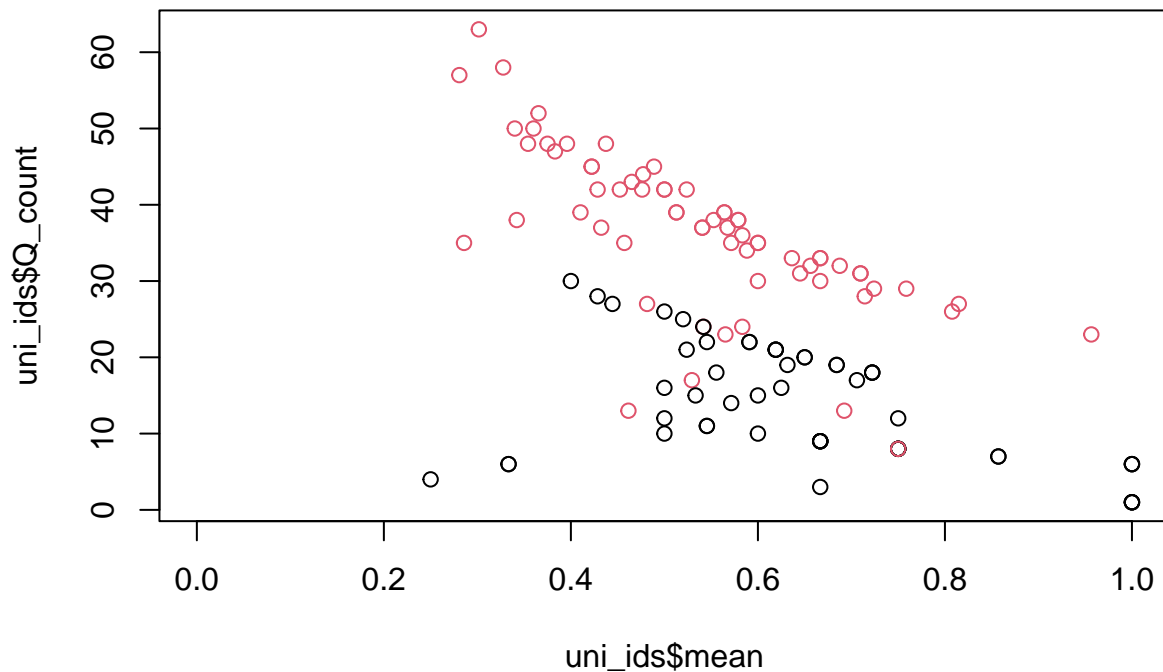
It is worth noting of the graph above that week 1 in general appears to have the most people continue no further. Additionally there is a spike at the very end of the 1st two weeks indicating that after that week the student completed no more steps. FutureLearn could focus their efforts on smoothing 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 FutureLearn 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.

```
correlation_matrix
```

```
##               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
```

Although only dealing with n=68 the above 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.

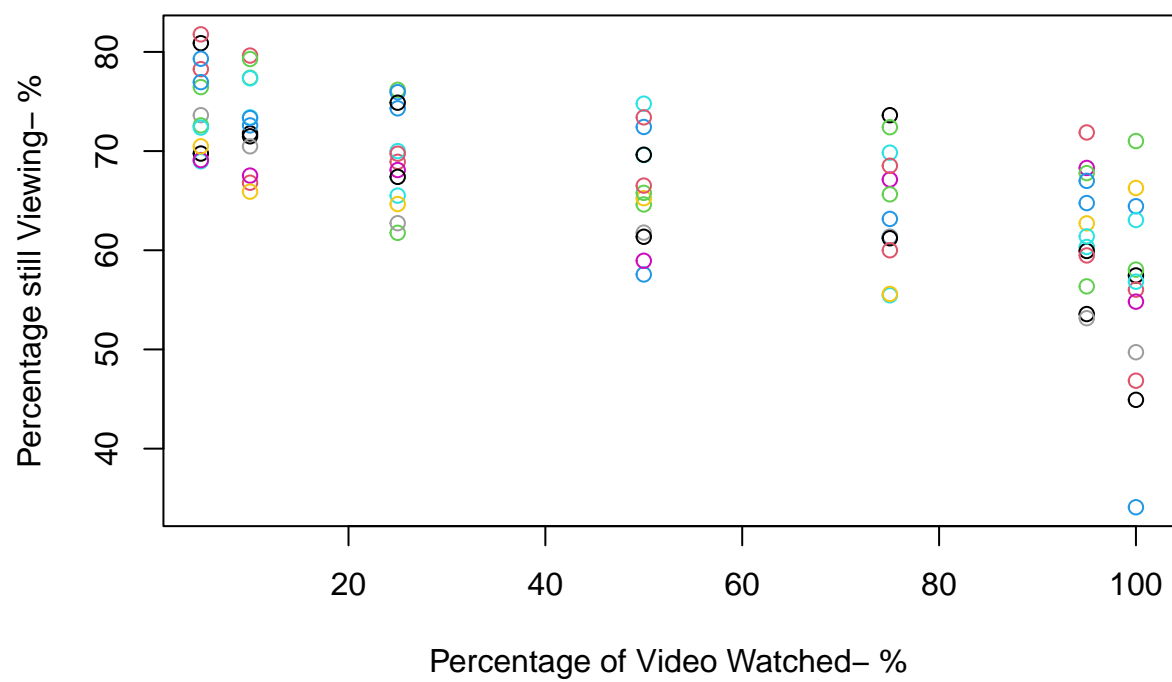
```
plot(uni_ids$mean, uni_ids$Q_count, col=uni_ids$pass+1)
```



Plotting percentage of correct answers divided by 100 against the number of questions answered. The at points are coloured red if we know that they fully participated. This plot is not very useful as it can be expected that those who completed the course answered more questions completed the course but it does again confirm this.

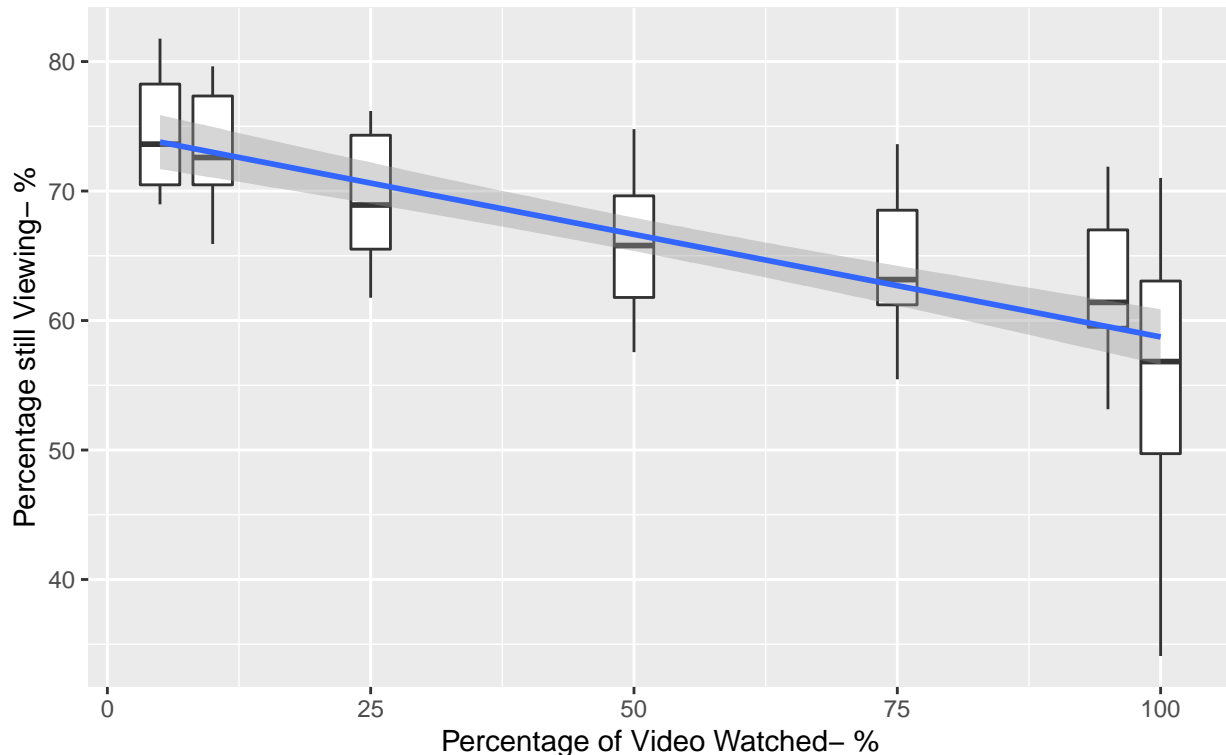
```
plot(rep(c(5,10,25,50,75,95,100),13), t(cyber.security.7_video.stats[,9:15]),
     type="p", col=1:13, ylab= "Percentage still Viewing- %",
     xlab = "Percentage of Video Watched- %",
     main= "7th Dataset Video Stats Watchtime vs Students Still Watching")
```

7th Dataset Video Stats Watchtime vs Students Still Watching



video_boxplot

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



This graph shows that as the videos continue the range of the percentage of people watching the videos increases. There is a noticeable downward trend across the graph especially after the first 10% and last 95% of the video. 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 the 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. This graph serves as the first rough look into the video stats dataset. Viewing this same data with lines connecting the dots is misleading as R does not connect the points together well.

```
cat(statement2, "\n", statement1, "\n", statement0, "\n", statement1,
    "\n", statement2, "\n", statement3, "\n", statement4, "\n", statement5,
    "\n", statement6, "\n", statement7, "\n", statement8, "\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
```

Students who were found to be on the course in datasets 6 and 7 are assumed to 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. 5 of the students who have a recorded fully participated date were in both datasets and one of them went on to have another fully participated date. 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. The reason for leaving the first and second time may be found in the leaving survey.

```
leaving_reason<-left_join(repeat_students,
                          cyber.security.6_leaving.survey.responses,
                          by=c("learner_id"))
leaving_reason[!is.na(leaving_reason$leaving_reason), "leaving_reason"]
```

```
## # A tibble: 10 x 1
##   leaving_reason
##   <chr>
## 1 I prefer not to say
## 2 I prefer not to say
## 3 I prefer not to say
## 4 The course was too easy
## 5 I donâ&#39;t have enough time
## 6 I donâ&#39;t have enough time
## 7 I donâ&#39;t have enough time
## 8 I donâ&#39;t have enough time
## 9 I donâ&#39;t have enough time
## 10 I prefer not to say
```

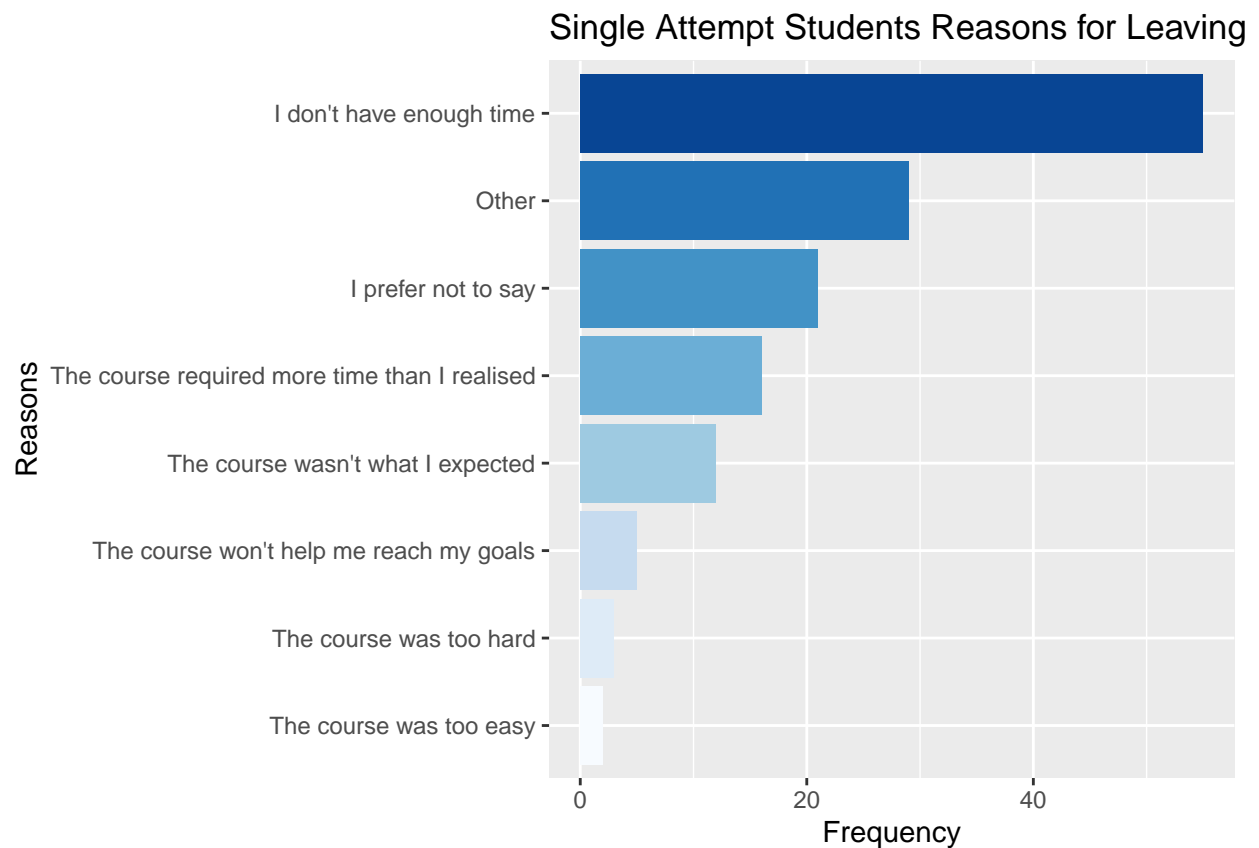
We can see the reasons given by the students that repeated their studies above (if they left a reason). Half say that they did not have enough time for the course as we suggested previously. Only one says that the course was too easy and the others preferred not to say.

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

```
## # A tibble: 10 x 1
##   leaving_reason
##   <chr>
## 1 I donâ&#39;t have enough time
## 2 Other
## 3 I donâ&#39;t have enough time
## 4 Other
## 5 I donâ&#39;t have enough time
## 6 I donâ&#39;t have enough time
## 7 I donâ&#39;t have enough time
## 8 I donâ&#39;t have enough time
## 9 I donâ&#39;t have enough time
## 10 I prefer 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.

single_reason



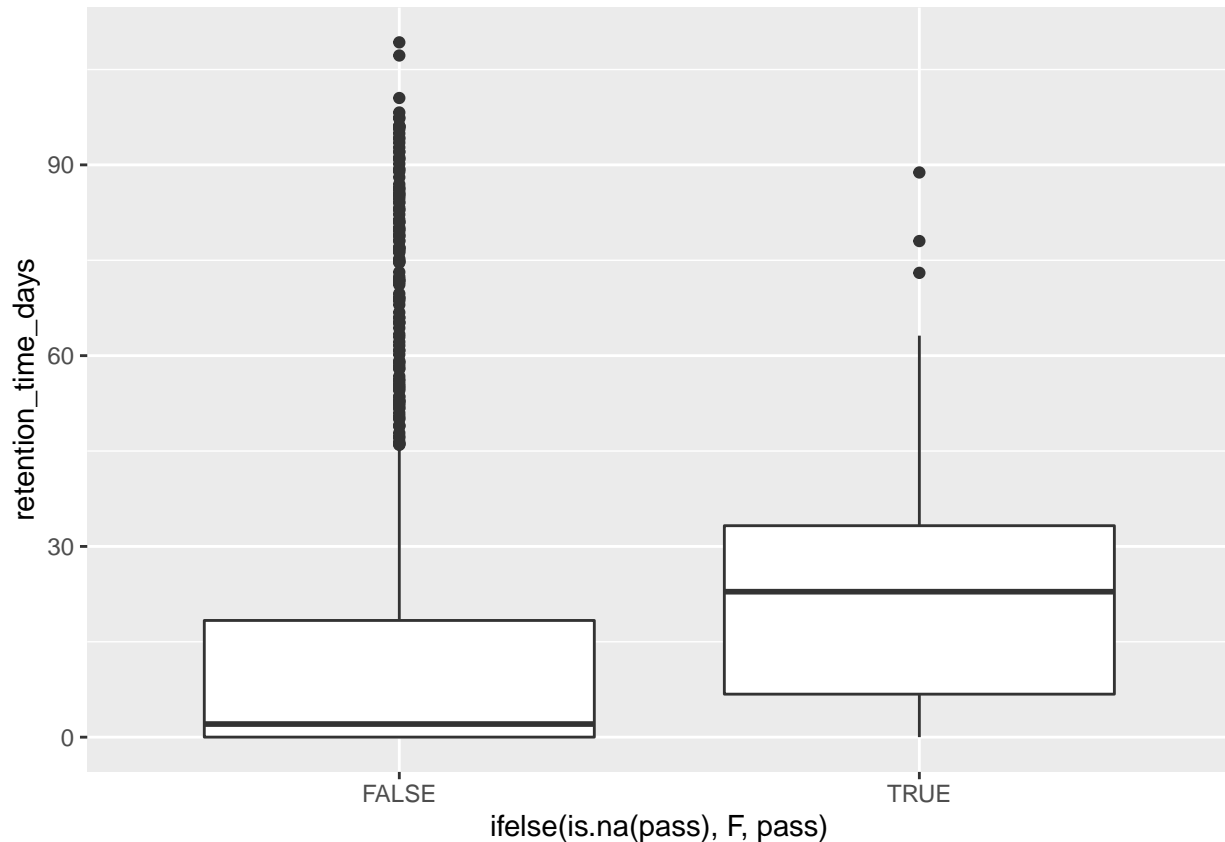
general_leaving_reason

##	reasons	frequency
## 1	<NA>	0
## 2	The course was too easy	2
## 3	The course was too hard	3
## 4	The course won't help me reach my goals	5
## 5	The course wasn't what I expected	12
## 6	The course required more time than I realised	16
## 7	I prefer not to say	21
## 8	Other	29
## 9	I don't have enough time	55
## 10	Total	143

The non-repeat students frequently state that their reason for leaving was that they did not have enough time much like the repeat students. "Other" 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 indicate that expectations for the course timewise need to be managed early which may even encourage those who would drop out otherwise to be retained for longer. It may also be

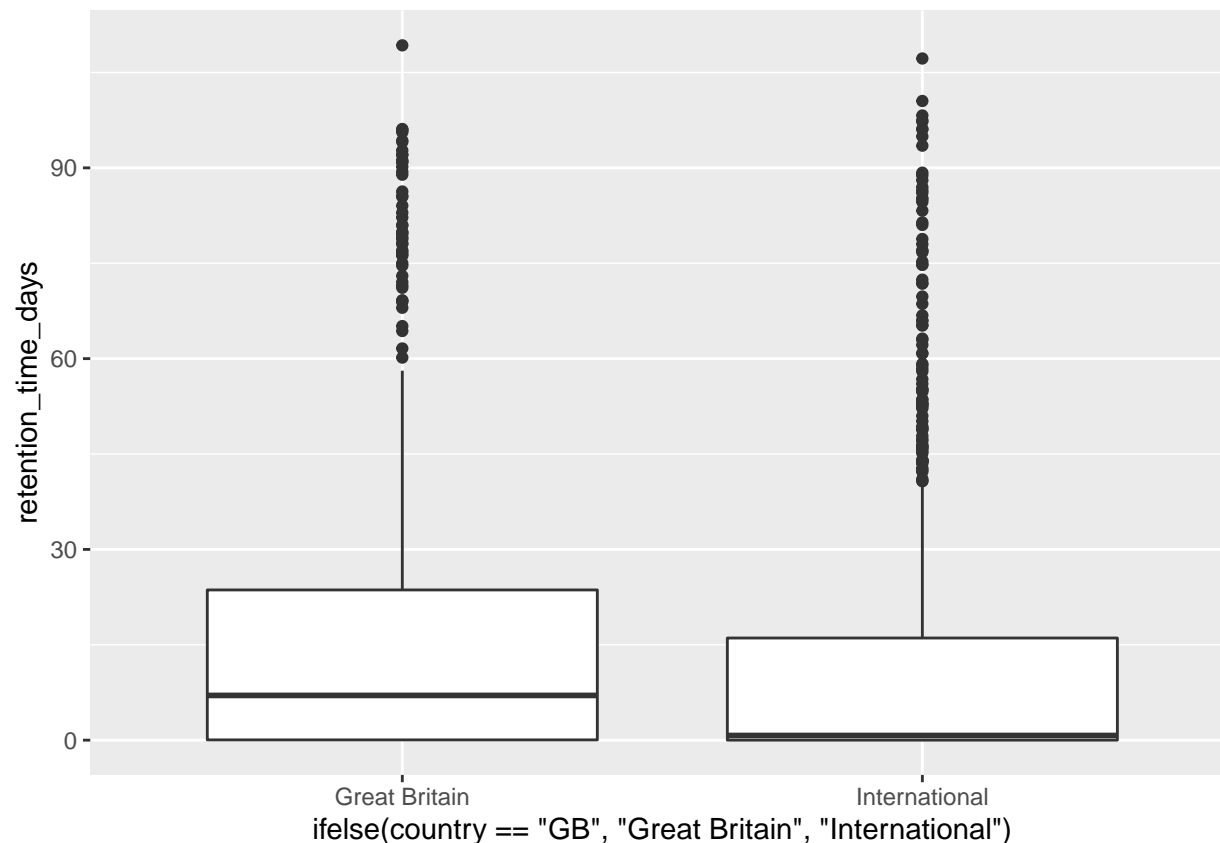
worth trimming videos that fail to retain viewers after a certain amount of time so that the student does not feel that their time is being wasted. Reassessing which content is vital or could be streamlined could also help the students feel like the time they spend is worthwhile whilst simultaneously decreasing the time required to work on the course.

pass_graph



It is clear that retention time (retention time = the difference between the last task completed date and enrollment date) 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 ($n=2339$). This graph 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 archive 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 outlier in our dataset. Additionally our sample size for those who fully participated is very small and it may be better to only consider

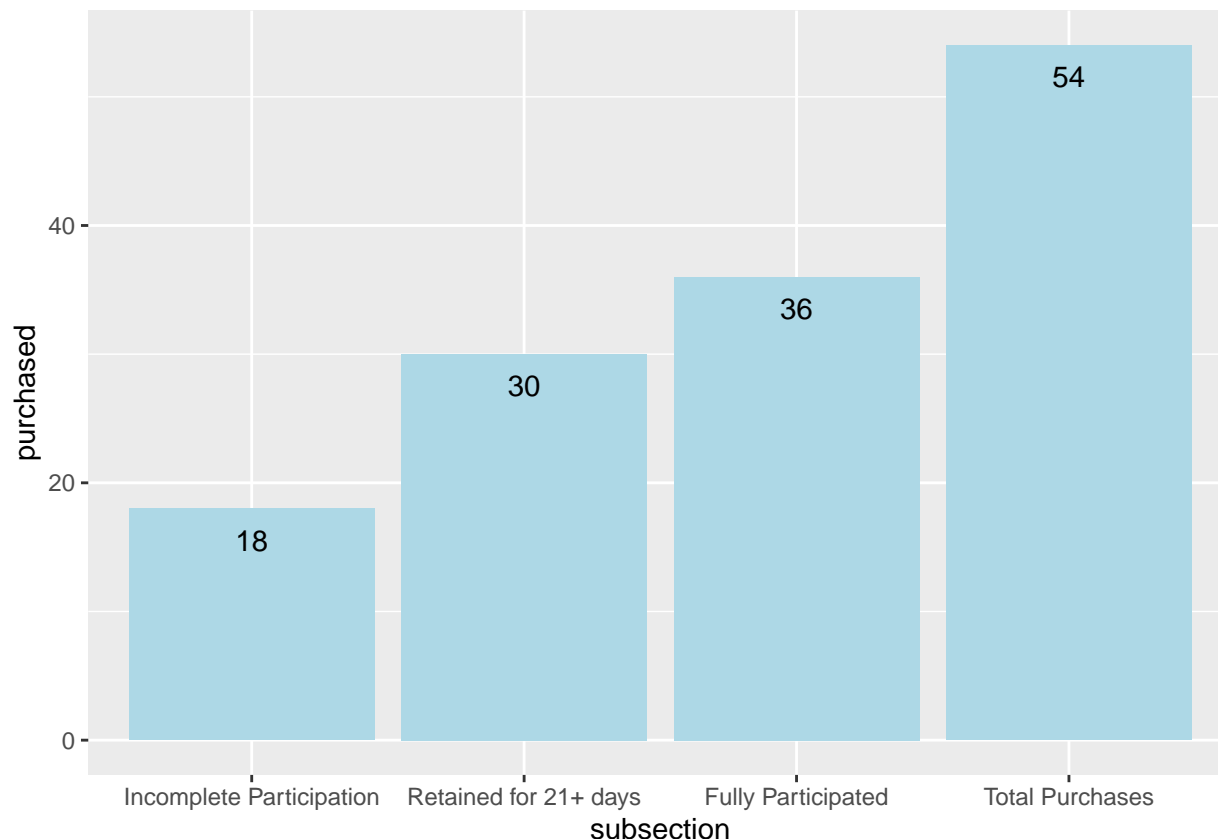
student_country_plot



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 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=3926) than those detected to be in Great Britain (n= 1384). 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

```
purchase_bar
```



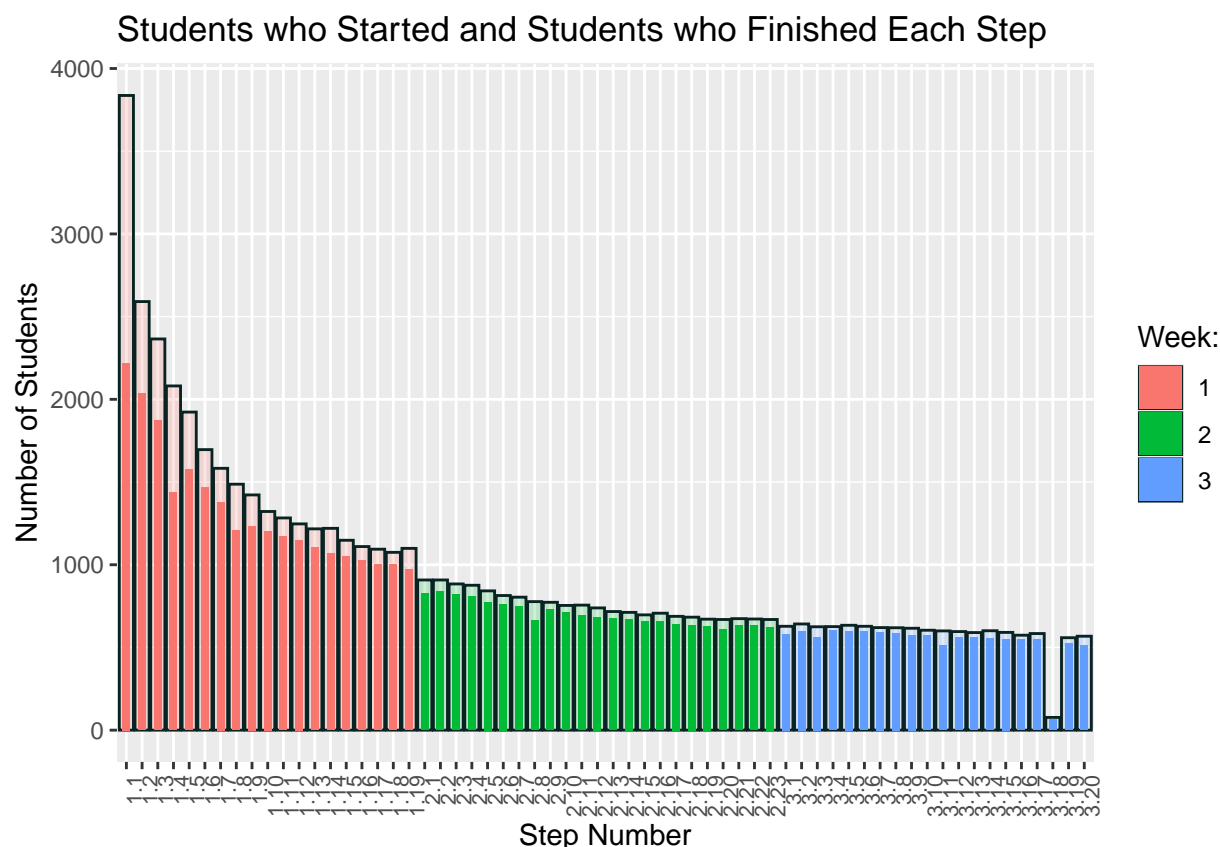
For all the analysis so far we have assumed that purchasing the course certificate was more likely if the individual finished the course. Before continuing we should confirm that assumption. With the limited dataset currently being used we show that twice as many people who fully participated in the course purchased the certificate. We also showed that those with a retention time of 21 days or 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 below)

```
probs_data
```

```
##          Condition Probability.When.Not Probability.Given
## 1 Fully Participated      0.00343184      0.52941176
## 2 Retained >= 21 days    0.01020956      0.05504587
```

The data frame above 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 certificat 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.

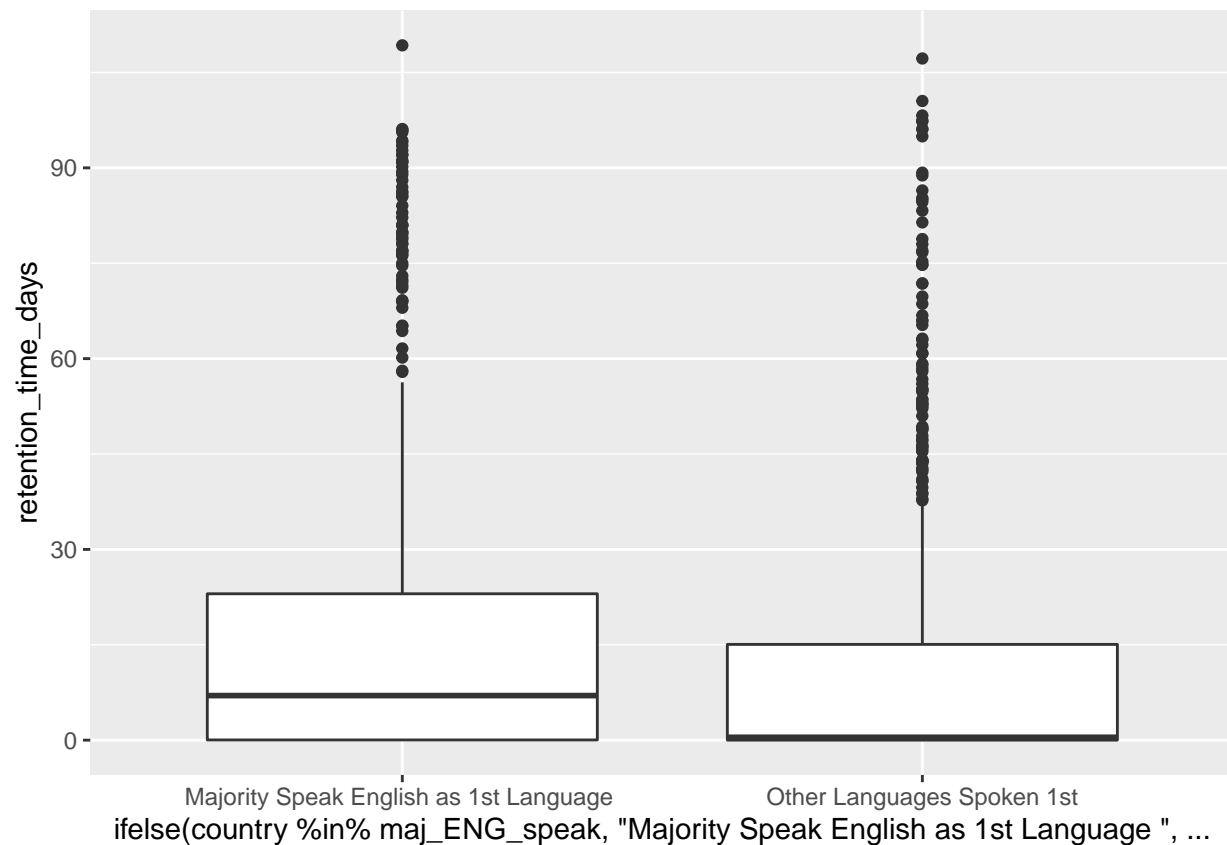
```
steps_completed_chart
```



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

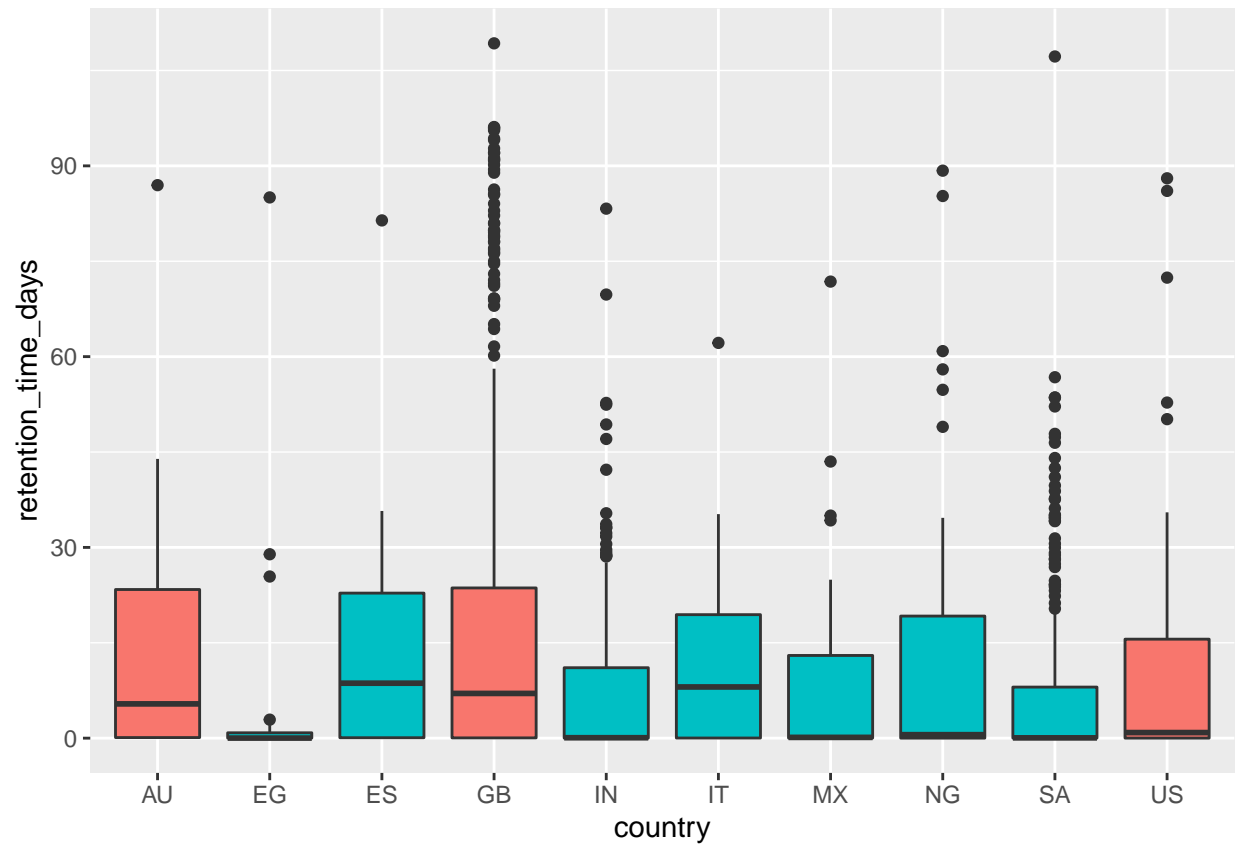
Previously we showed last step students managed to complete, but this does not show the entire story. Individuals could complete any combination of steps and therefore some information could be lost by only reporting the last step completed. Here we see that the dropout after each week is real but are 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 but 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 there is a high dropout rate in week 1 as shown in the first graph.

```
student_language_plot
```



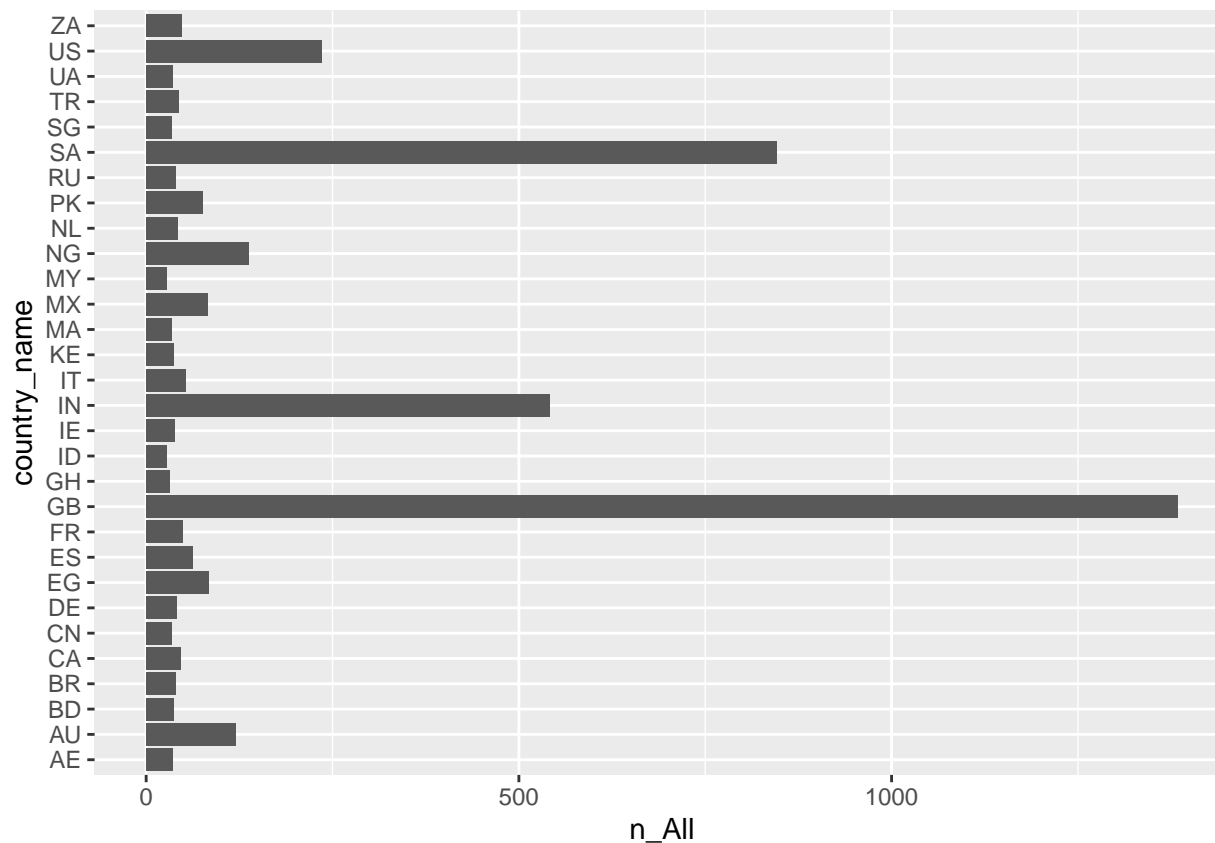
It seems that overall students from a country that majoritively speaks English as a first language (MSEFL) were retained for a longer median time than those from countries that did not speak English as a first language. However Information may be lost when taking such a top down view and further analysis 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. To maximize the benefit to the students FutureLearn could consider which non MSEFL countries had the most students

student_language_plot_25



plot of countries vs retention time in days. The

```
plot_countries
```



The plot above shows each country with 25 or more students and then the number of students from that country. Notably SA (Saudi Arabia) students and IN (India) students make up the next greatest proportion of students after GB (Great Britain). FutureLearn should consider this important as these two countries have poor median retention times as seen in the previous graphs. 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.