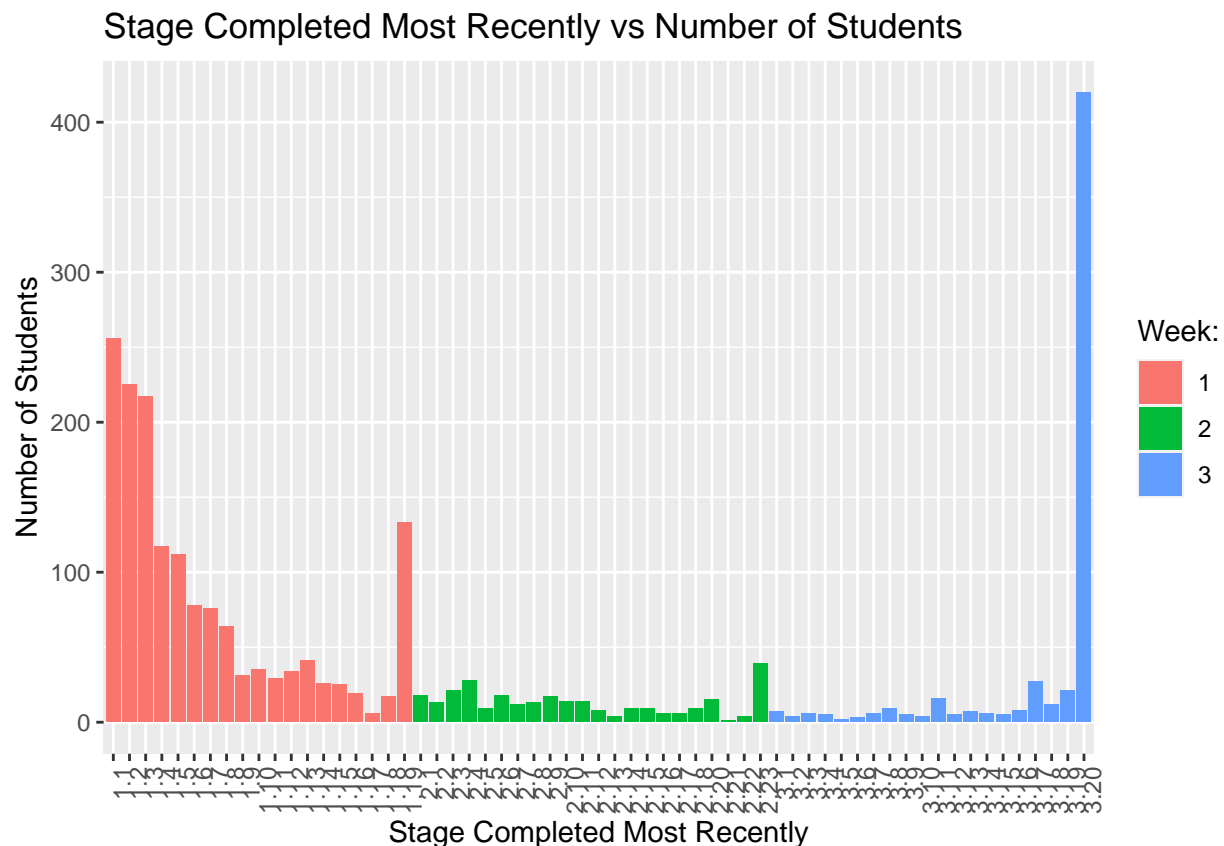


Project

Project Report

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.

```
bar_step<-ggplot(data=uni_ids[!is.na(uni_ids$last_step_completed),], aes(last_step_completed, fill=as.c
bar_step+ggtitle("Stage Completed Most Recently vs Number of Students") + labs(x="Stage Completed Most R
```



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.

```
summary(uni_ids)
```

```
## learner_id      pass_date      unenroll_date      pass
## Length:5313      Min.       :1.529e+09      Min.       :1.524e+09      Mode :logical
## Class :character  1st Qu.:1.535e+09      1st Qu.:1.532e+09      FALSE:267
## Mode  :character  Median :1.539e+09      Median :1.535e+09      TRUE :68
##                               Mean  :1.537e+09      Mean  :1.535e+09      NA's :4978
##                               3rd Qu.:1.540e+09      3rd Qu.:1.538e+09
##                               Max.   :1.541e+09      Max.   :1.541e+09
##                               NA's   :5245          NA's   :5044
##      mean      Q_count      last_step_completed      letter_code
## Min.       :0.000      Min.       : 1.00      3.20      : 420      Length:5313
## 1st Qu.:0.500      1st Qu.: 8.00      1.1       : 256      Class :character
## Median :0.611      Median :13.00      1.2       : 225      Mode  :character
## Mean    :0.621      Mean    :15.53      1.3       : 217
## 3rd Qu.:0.750      3rd Qu.:21.00      1.19      : 133
## Max.    :1.000      Max.    :63.00      (Other):1155
## NA's    :4046      NA's    :4046      NA's    :2907
## week_number      date_of_last      question_score
## Min.       :1.0      Min.       :1.529e+09      Min.       : 0.000
## 1st Qu.:1.0      1st Qu.:1.531e+09      1st Qu.: 6.000
## Median :1.0      Median :1.534e+09      Median : 7.000
## Mean    :1.6      Mean    :1.535e+09      Mean    : 8.896
## 3rd Qu.:2.0      3rd Qu.:1.538e+09      3rd Qu.:13.000
## Max.    :3.0      Max.    :1.541e+09      Max.    :22.000
## NA's    :2907      NA's    :2907      NA's    :4046
```

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.

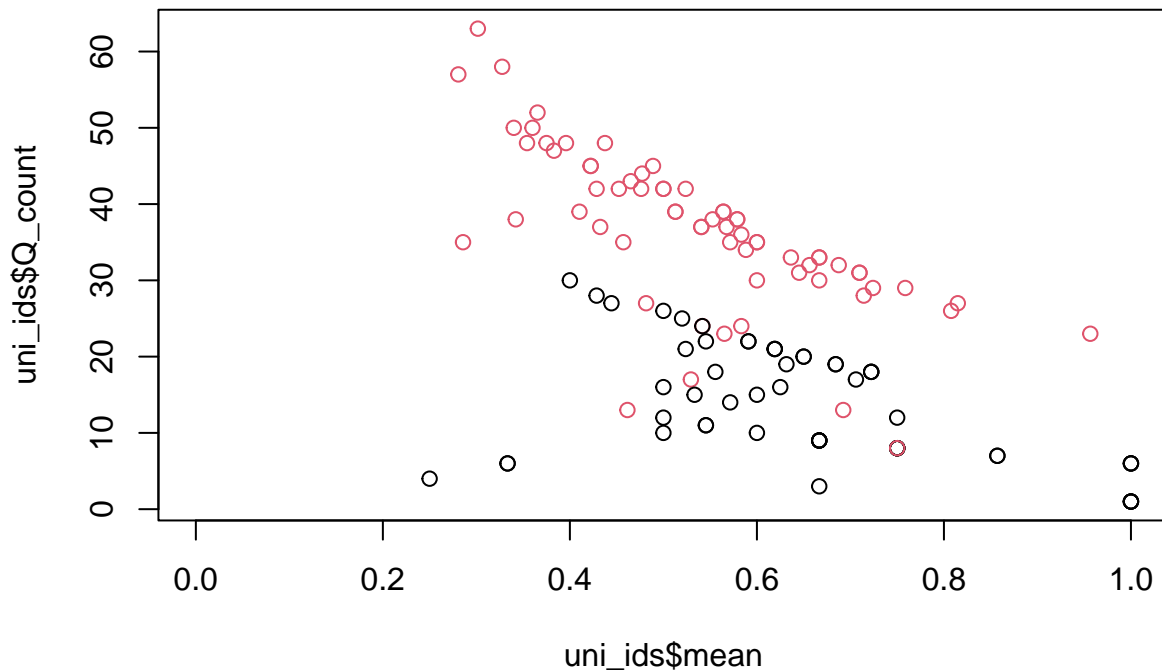
```
uni_ids[,c("pass", "question_score", "Q_count", "mean")]%>%
  as.matrix(nrow=nrow(uni_ids), ncol=3)%>%
  cor(use = "complete.obs")
```

```
##      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.

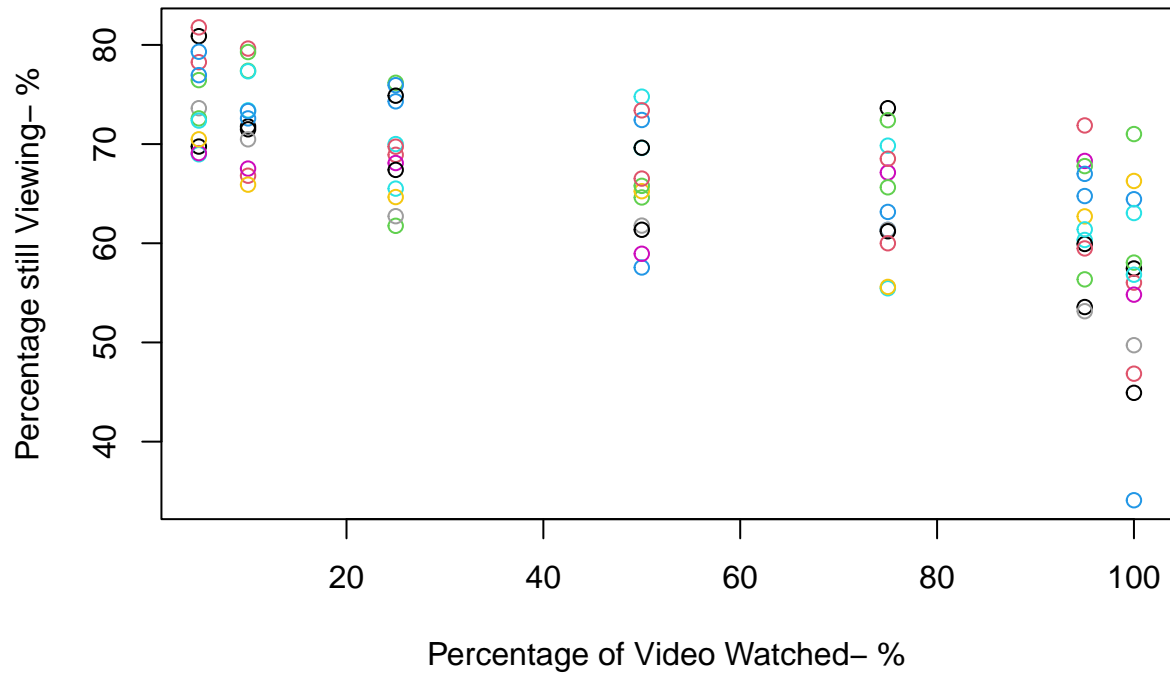
```
plot(uni_ids$mean, uni_ids$Q_count, col=uni_ids$pass+1)
```



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.

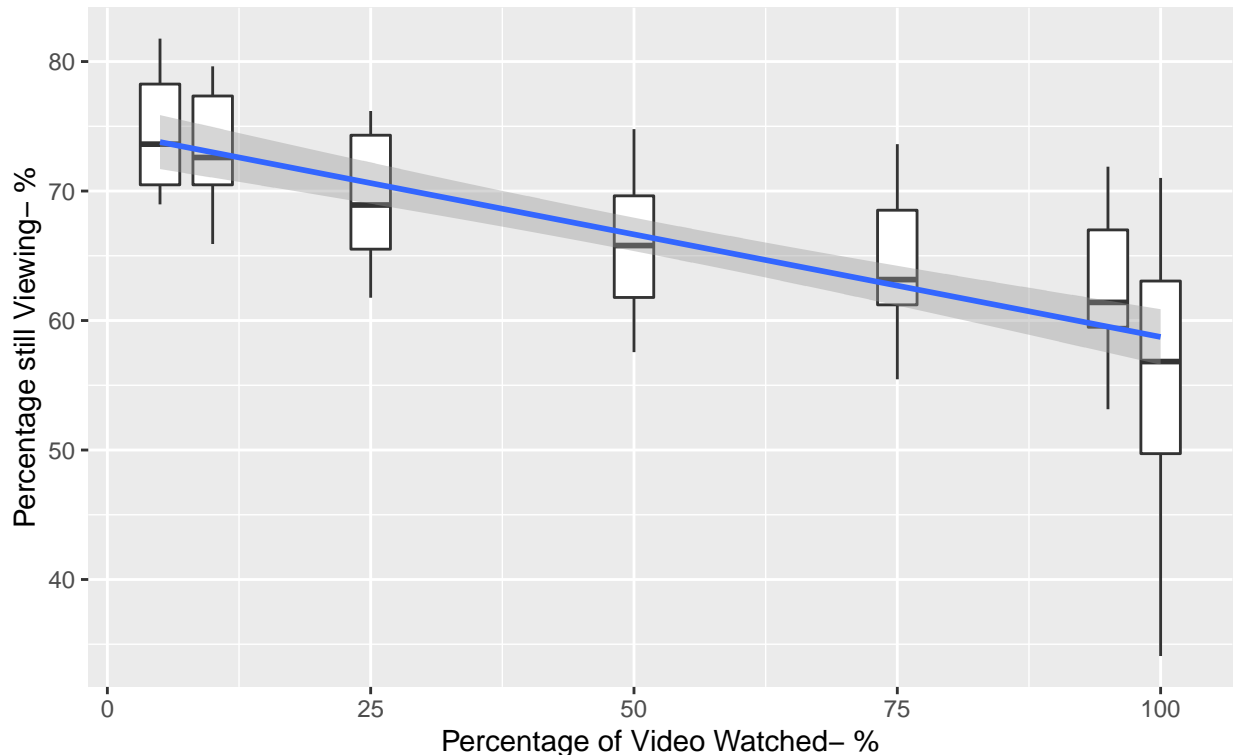
```
plot(rep(c(5,10,25,50,75,95,100),13), t(cyber.security.7_video.stats[,9:15]), type="p", col=1:13, ylab=
```

7th Dataset Video Stats Watchtime vs Students Still Watching



```
t_video<-data.frame(t(cyber.security.7_video.stats[,9:15]))
t_video$percent<-c(5,10,25,50,75,95,100)
colnames(t_video)<-paste(c(rep("X",13), "percent"), c(1:13, ""), sep="")
t_video<-pivot_longer(t_video, colnames(t_video)[-14])
boxplot_graph<- ggplot(data=t_video, aes(x=percent, y=value))
boxplot_graph+geom_boxplot(aes(group=percent))+ labs(title= "7th Dataset Video Stats Watchtime vs Student",
x="Percentage of Video Watched- %",
y="Percentage still Viewing- %")+geom_smooth(method="lm")
```

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.

```
first_attempt<-left_join(repeat_students,cyber.security.6_enrolments, by=c("learner_id"))
second_attempt<-left_join(repeat_students,cyber.security.7_enrolments, by=c("learner_id"))

statement2<-paste(nrow(first_attempt), " Repeated their studies (total)")

statementm1<-second_attempt$learner_id[second_attempt$fully_participated_at!=""] %in% first_attempt$lea
sum() %>%
paste(" Completed the course twice")

statement0<-second_attempt$learner_id[second_attempt$unenrolled_at!=""] %in% first_attempt$learner_id[fi
sum() %>%
paste(" Unenrolled in their first and second attempt", sep="")
```

```

statement1<-paste(sum(first_attempt$fully_participated_at!=""), "fully participated the first time", sep=" ")
statement2<-sum(second_attempt$fully_participated_at!="") %>%
  paste("fully participated the second time", sep=" ")
statement3<-sum(first_attempt$unenrolled_at!="") %>%
  paste("unenrolled the first time", sep=" ")
statement4<-sum(second_attempt$unenrolled_at!="") %>%
  paste("unenrolled the second time", sep=" ")
statement5<-sum(first_attempt$unenrolled_at=="") %>%
  paste("Assumed did not finish the course the first time", sep=" ")
statement6<-sum(second_attempt$unenrolled_at=="") %>%
  paste("Assumed did not finish the course the second time", sep=" ")
store_score<-mean(ifelse(is.na(uni_ids$pass)| uni_ids$pass==F, 0,1 )) %>%
  round(4)
statement7<-paste(store_score*100,"% of the non repeat students finished the course", sep=" ")
unenroll_score<-mean(ifelse(uni_ids$unenroll_date=="|is.na(uni_ids$unenroll_date), 0,1 )) %>%
  round(4)
statement8<-paste(unenroll_score*100, "% of the non repeat students unenrolled from the course", sep=" ")

print(cat(statementm2, "\n", statementm1, "\n", statment0, "\n", statement1, "\n", statment2, "\n", statment3, "\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
## NULL

```

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", "dataset_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€t have enough time
## 6 I donâ€t have enough time
## 7 I donâ€t have enough time
## 8 I donâ€t have enough time
## 9 I donâ€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.

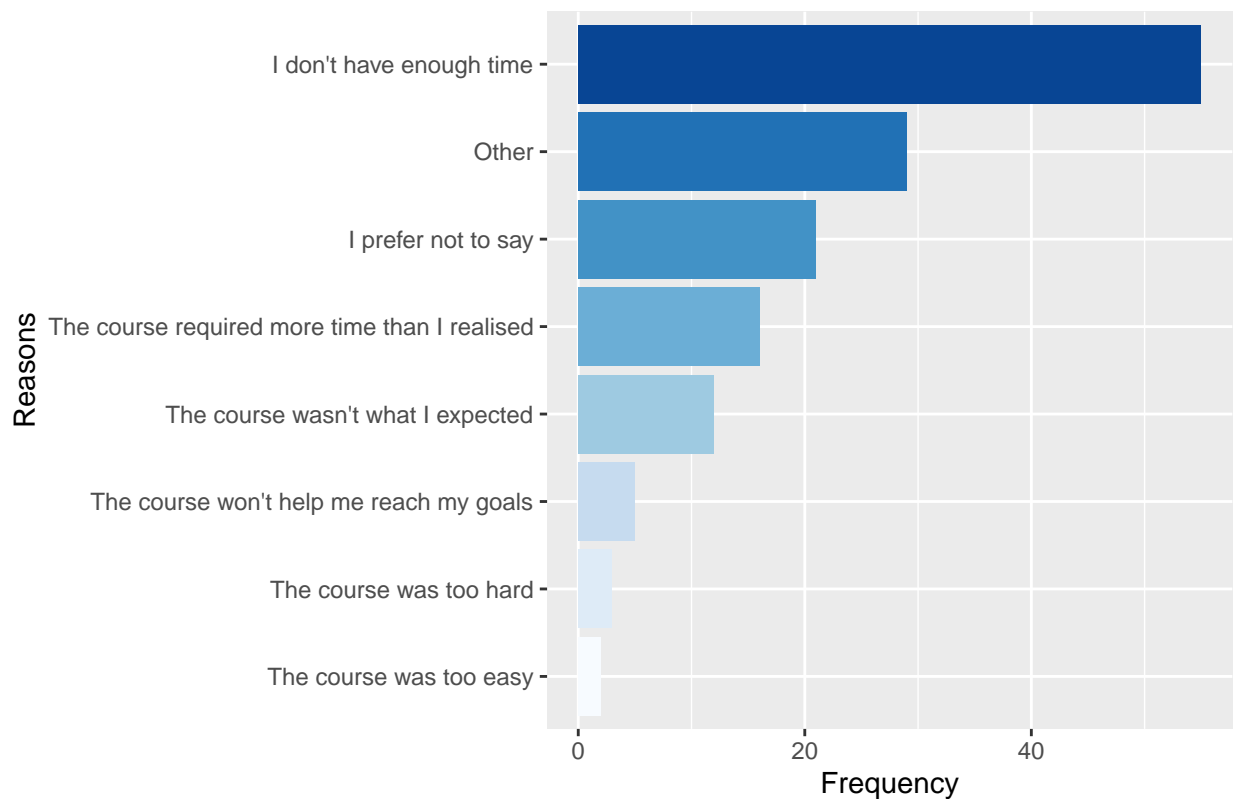
```
leaving_reason2<-left_join(repeat_students, cyber.security.7_leaving.survey.responses, by=c("learner_id", "leaving_reason"))
leaving_reason2[!is.na(leaving_reason2$leaving_reason), "leaving_reason"]
```

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

```
#We remove total from the bar chart as it is not relevant, NA has no occurrences due to the way we made this dataset so it can be ignored as well
total<-which(general_leaving_reason[,1]=="Total"| is.na(general_leaving_reason[,1]))
bar_step_reason<-ggplot(data=general_leaving_reason[-total,], aes(reasons, frequency, fill=reasons))
bar_step_reason+scale_fill_brewer(palette = "Blues")+ theme(legend.position = "none",axis.text.x = element_text(angle = 45))
```

Single Attempt Students Reasons for Leaving



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

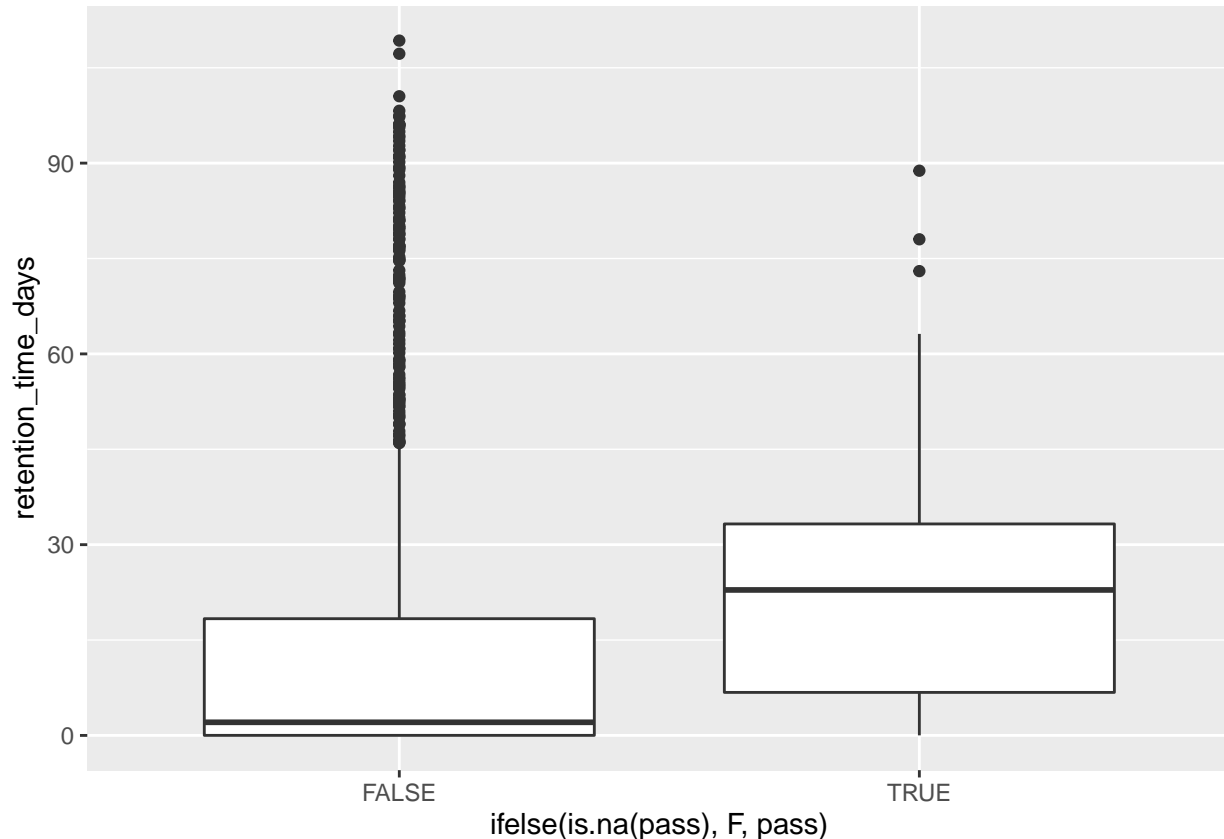
The non-repeat students frequently state that their reason for leaving was that they did not have enough time much like the repeat students. “Other” likely consists of multiple reasons for leaving more specific to the student. The course requiring more time than was realized was the reason given by 16 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.


```

another<-uni_ids%>%
  left_join(union(cyber.security.6_enrolments, cyber.security.7_enrolments), by=c("learner_id"))
if(mean(uni_ids$learner_id==another$learner_id)==1){
  uni_ids$retention_time_days<-(uni_ids$date_of_last/86400)-(as.numeric(as.POSIXct(another$enrolled_at ,f
new_boxplot_graph<- ggplot(data=uni_ids, aes(x=ifelse(is.na(pass), F, pass), y=retention_time_days))
new_boxplot_graph+geom_boxplot(aes(group=ifelse(is.na(pass), F, pass)))

```

Warning: Removed 2907 rows containing non-finite values (stat_boxplot).



```

#length(which(ifelse(is.na(uni_ids$pass), F, uni_ids$pass==T)==F & !is.na(uni_ids$retention_time_days)))
#length(which(ifelse(is.na(uni_ids$pass), F, uni_ids$pass==T)==T & !is.na(uni_ids$retention_time_days)))
#commented out code to find sample sizes

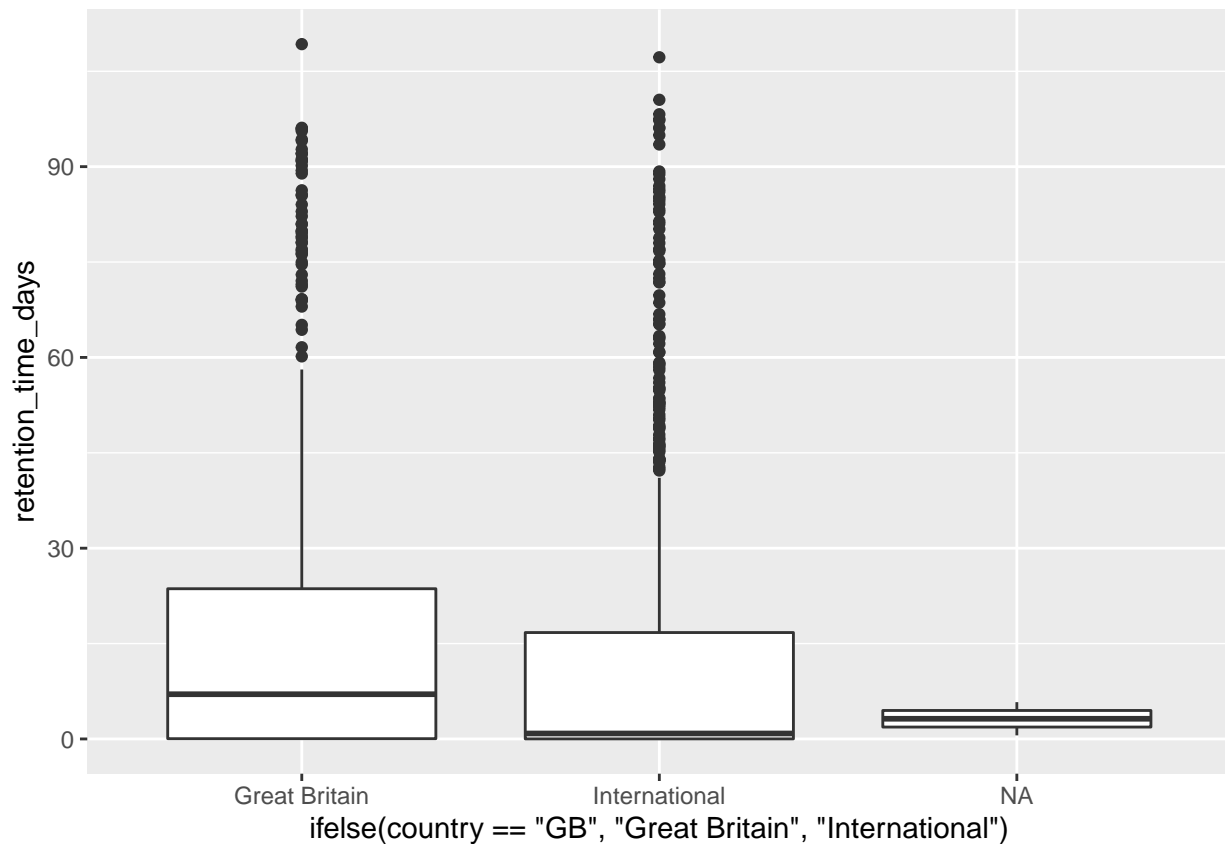
```

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 outline in our dataset. Additionally our sample size for those who fully participated is very small and it may be better to only consider

```
uni_ids$country<-another$detected_country
new_boxplot_graph<- ggplot(data=uni_ids, aes(x=ifelse(country=="GB", "Great Britain", "International"),
new_boxplot_graph+geom_boxplot(aes(group= ifelse(country=="GB", "Great Britain", "International")))
```

```
## Warning: Removed 2907 rows containing non-finite values (stat_boxplot).
```



```
#length(which(ifelse(uni_ids$country=="GB", "Great Britain", "International")=="Great Britain" & !is.na
#length(which(ifelse(uni_ids$country=="GB", "Great Britain",
#"International")=="International" & !is.na(uni_ids$country)))
#Commented out code to find the sample sizes
```

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 English speaking countries to non-English speaking countries may allow for FutureLearn to determine if a language barrier plays a role in the lower median for International students. # Second Cycle