CSC8631

Elliott Trott 180259511

09/05/2022

Abstract -

Over a seven year period data was collected on participants of FutureLearn MOOCs. Deciphering the data and revealing any variables that are closely linked with engagement could lead to a more efficient implementation of change. Studying student behaviors is vital for businesses and online courses, therefore, this report primarily focuses on the patterns of individuals who have attempted to complete the course.

Key Aims -

- Describing what the data is telling us
- Understanding any indicators for lack of participation
- Identifying areas that are putting students off
- Suggesting improvements based on statistical analysis
- Analysis of temporal effects on submission and results

Contents

Introduction -

Studying collected data is one important aspect of improving a business. Considering MOOC data provided by FutureLearn, the focus of this report will be studying and attempting to understand student behaviors. Research will continue in this area, enabling courses, and, in the same way, universities, to make more informed decisions on how to help their students.

The investigation was carried out over the data using a CRISP-DM frame work. One of the first steps in the process is understanding the data from the businesses perspective, this includes considering the data with certain questions in mind. For example, it is important to know what FutureLearn's aims are and what information they would like to reveal about the course and their students.

The Data -

Data over the 7 year period was provided by FutureLearn in the form of 53 csv files. There were 6 different types of files that were present throughout all 7 years, and an additional 2 different types were provided for later years. The data collected included information about students, for example: their gender, learner id, whether they completed the course, whether they purchased the statement and at what times they submitted question responses. Paired with statistics taken from students watching the videos and leaving responses from students who did not complete the course, the data set was fairly extensive.

To better represent the composition of the data, a segment of one of the csvs is provided below.

```
## # A tibble: 6 x 13
##   learner_id   enrolled_at unenrolled_at role  fully_participa~ purchased_state~
##   <chr>        <chr>       <chr>         <chr> <chr>            <chr>
## 1 f0ebc6f6-0f~ 2018-10-30~ <NA>          lear~ <NA>             <NA>
## 2 0fa1c614-8a~ 2018-10-25~ <NA>          lear~ <NA>             <NA>
## 3 a0ac585a-eb~ 2018-10-23~ <NA>          lear~ <NA>             <NA>
## 4 93562e76-6e~ 2018-11-01~ <NA>          lear~ <NA>             <NA>
## 5 d1ff90bd-b9~ 2018-10-11~ <NA>          lear~ <NA>             <NA>
## 6 cd362ad5-60~ 2018-10-11~ <NA>          lear~ <NA>             <NA>
## # ... with 7 more variables: gender <chr>, country <chr>, age_range <chr>,
## #   highest_education_level <chr>, employment_status <chr>,
## #   employment_area <chr>, detected_country <chr>
```

Above is a snapshot of a typical csv file that was used to complete the analysis over this report. This particular snapshot displays the enrollment data for (the final year we have data for) year '7' of the course. The dimensions for each csv file vary, therefore, merging and splicing of data sets was often necessary to enable the most accurate measurements. As an example of how varied the sets were, below are the outputs from requesting the dimensions of two different csv files

```
dim(cyber_security_6_enrolments)
```

```
## [1] 3175   13
```

```
dim(cyber_security_5_video_stats)
```

```
## [1] 13 28
```

These outputs describe just how varied each csv file is. The amount of data that was processed was vast. So vast, in fact, that many columns have not been considered for this report. The data that was used was chosen carefully and purposefully to provide desired end products.

Basic analysis -

Over the FutureLearn data set, it is imperative to understand a few basic principles and figures. It is also vital to have the CRISP-DM framework in mind, for the business, increasing participation numbers will be of high priority. The first analysis that took place was to compute participation numbers and numbers of individuals who purchased statements annually.

```
# table showing the number of participants by year:
TableForParticipantNumbers
```

```
##   Participant.Numbers Year
## 1               14394    1
## 2                6488    2
## 3                3361    3
## 4                3992    4
## 5                3544    5
## 6                3175    6
## 7                2342    7
```

Figures Above showing the participant numbers by year. The first year clearly has the greatest number of participants, it very rapidly declined and is now plateauing around 2500-3500. This is the first major concern for the business that we can immediately see. FutureLearn wishes for a greater number of participants, instead we have witnessed the total drop from 14394 to 3361 (a decrease of around 78%) in just 2 years.

A calculation for the mean average over all 7 years was also computed.

```
## [1] 5328
```

The mean output here shows only 2 of the 7 years total above it. It is imperative for the business to increase participation numbers to at least this mean average again.

Increasing participation is one main aim from the businesses perspective, another key criteria that denotes successful business would be the number of individuals willing to purchase the statement at the end of the course. Below is a table that describes totals of purchases by year.

```
# table for showing the number of purchases of the statement by year
TableForPurchasedNumbers
```

```
##   Number.Purchased Year
## 1               84    1
## 2               38    2
## 3               30    3
## 4               40    4
## 5               35    5
## 6               39    6
## 7               23    7
```

The year with the most purchases of the statement was the first year. Given that the number of participants for this particular year was essentially 220% of the next highest total participants for a one year period, this was to be expected.

```
## [1] 41.28571
```

The mean average was also computed for numbers of purchases of the year. The number over each year is something we need to rectify, and is one of the focal points for the remainder of the analysis: “What can we do increase this?”.

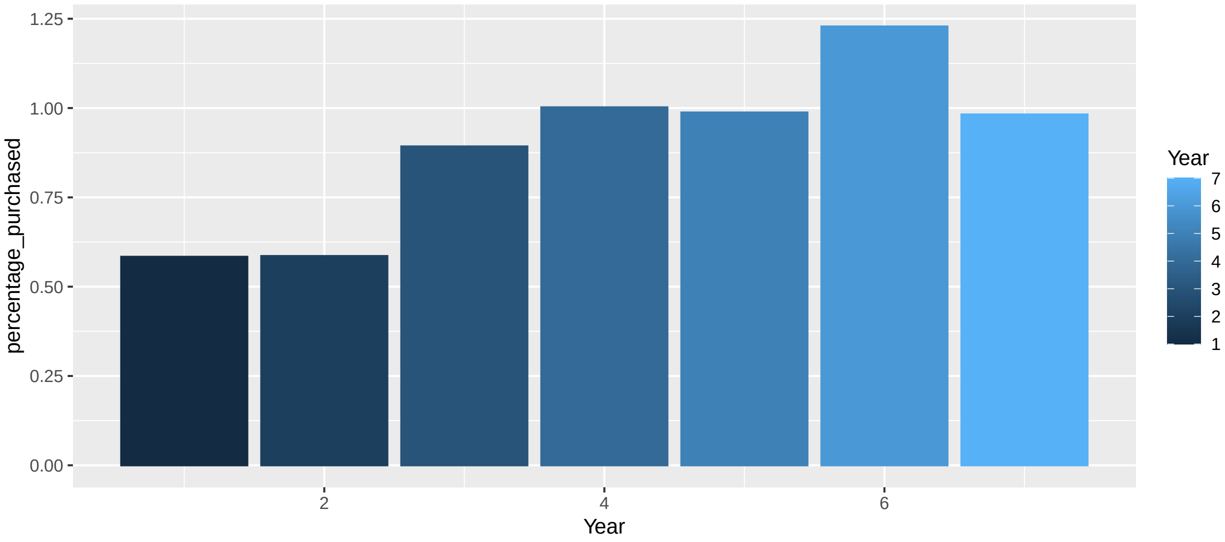


Figure 1: percentages for purchases by year

So we can visualize this more appropriately, a graphical plot (Figure 1) has been afforded displaying percentages of individuals who purchase the statement for each year.

From figure 1, the year that appeared to have the highest percentages of purchases per student is year 6. Although the figure is still under 1.25%. From the businesses perspective it does look as if percentages are increasing for this statistic, however, because numbers are so low for participation in recent years the total for purchases is minimal. FutureLearn should aim to continue increasing the percentage of purchases while simultaneously increasing participation numbers.

Years that displayed especially bad conversion percentages are year 1 and year 2.

A good business model would be to narrow down unsuccessful implementations of change by studying percentages between years and comparing the alterations made. Further investigation is required to more precisely state where the unsuccessful changes were made. However, by studying demographic data, step activity data and leaving responses we aim to do just that.

Demographic Analysis -

It is important for FutureLearn to answer a few questions which are quite common in the demographic analysis of ones audience. What demographics are we appealing to (who is our target audience)? Which groups do we need to work harder for?

Understanding these questions can allow for better growth and can also have a beneficial impact on reputation and brand image. Increasing the scope of the audience can be the basis of entire business models. Conversely, understanding which audiences you are appealing to and targeting those niches, for example, if you are primarily attracting university graduates (which is the case in this study) it may be another approach to tailor the courses to suit the needs of degree participants.

The first study that was performed was over data relating to gender.

Figure 2 displays the number of individuals who completed the course (Blue) and the number of those who started the course and did not finish (Pink) by gender. Although the plot does not represent a large skew in one direction or another, it does suggest FutureLearn should do its utmost to be more inclusive of females who are taking the course. The results for studying the demographic data are positive, if there was a large correlation between individuals who are failing to complete the course and gender, it would represent a much

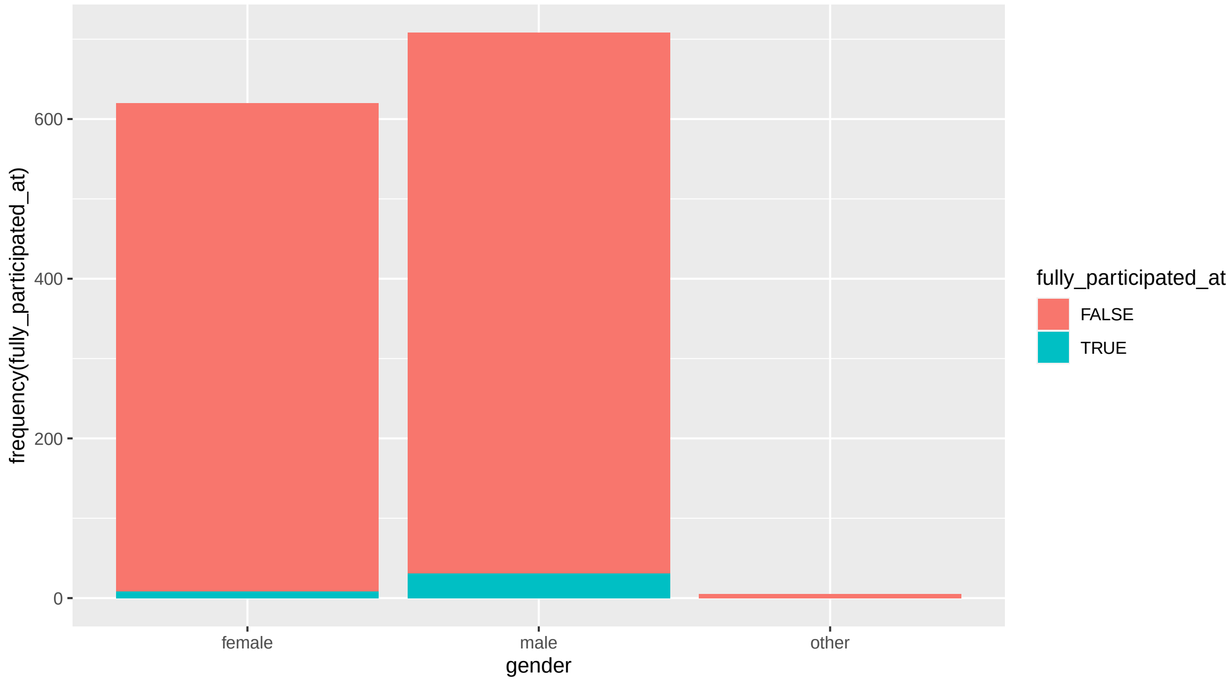


Figure 2: Frequency of full participation against gender

larger problem within the business. Therefore, to study and analyse more conclusive reasons as to why students are not completing the course, more demographic data must be studied, alongside investigation into leaving responses from the individuals who have taken part in the course.

Figure 3 describes the relationship between education level and numbers for participation alongside numbers for full completions of the course. It is evident from this graphical output that the course is dominated by those whose highest education level is university degrees. Unsurprisingly, the highest number of completions from a single group also relates to university graduates. It is important for a business to be aware of the individuals their product is attracting. There is clear separation between these groups and depending on FutureLearn's goals they should attempt to appeal to their wider audiences. There was not an apprentice who started the course who achieved a full completion, this is something FutureLearn should aim to correct within the next iteration of the course.

Leaving responses -

Why are individuals not completing the course? Data was merged from years 4 through 7 to allow for more rounded analysis, this was followed by individual examination for each year. Unfortunately, there is no complete data set from the first year of leaving responses. This years data would have been preferable as, after this year, the number of participants decreased massively.However, even without this there is extensive data and valuable information was provided by the analysis of such sets.

Figure 4: From this analysis we can understand a few of the reasons why participants are leaving part way through the course. FutureLearn wish to retain as many students as possible, therefore taking heed of the most common responses such as “I don't have enough time” is critical. The course should consider methods to alleviate the numbers who feel this way. To narrow down the areas in which students are coming to these decisions further investigitave methods were used.

Some alternative analysis could be performed in a similar area. Namely, investigating which parts of the course are responsible for individuals not completing the course. With this information strategies could be put in place focused around weeks that display a large number of dropouts.

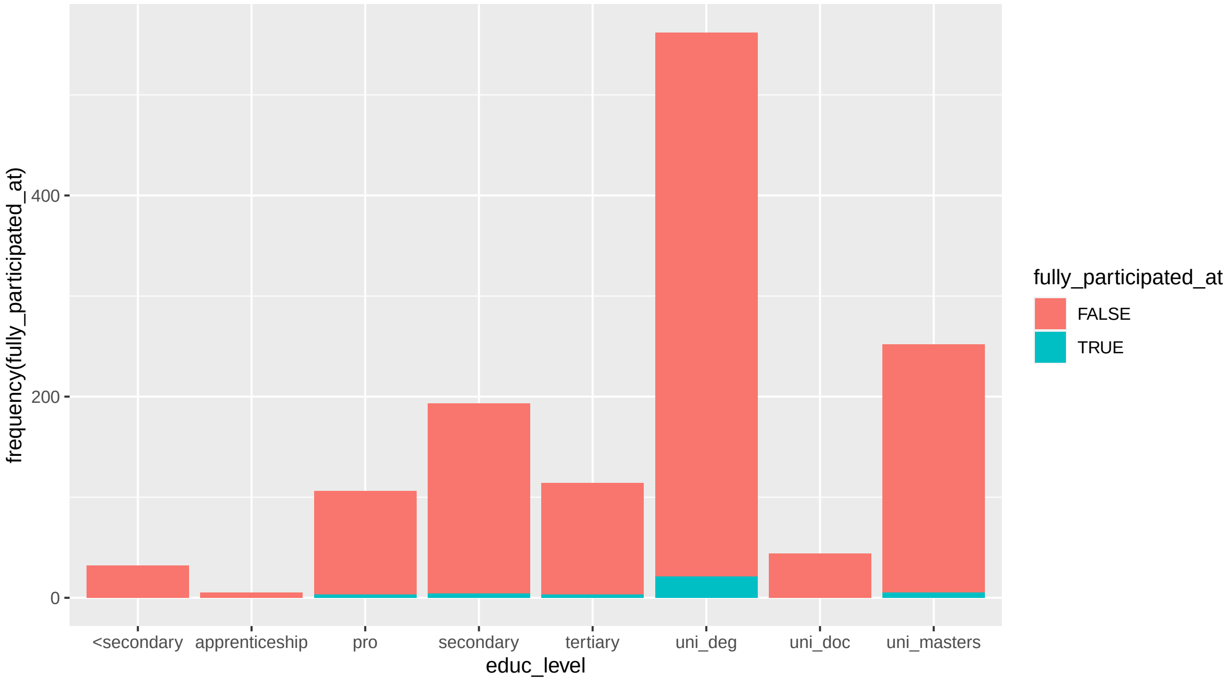


Figure 3: Frequency of full participation against highest education level

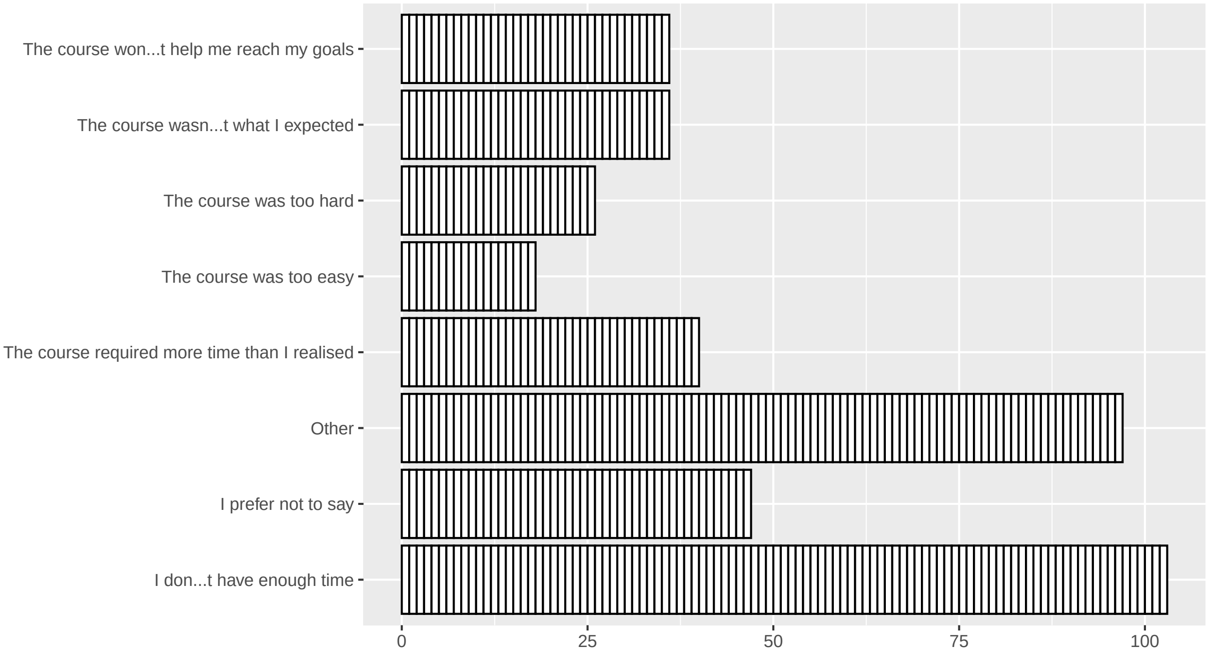


Figure 4: Frequency of Leaving reasons across combined years

Using the same data sets of leaving responses the following evidence was afforded. (Figure 5)

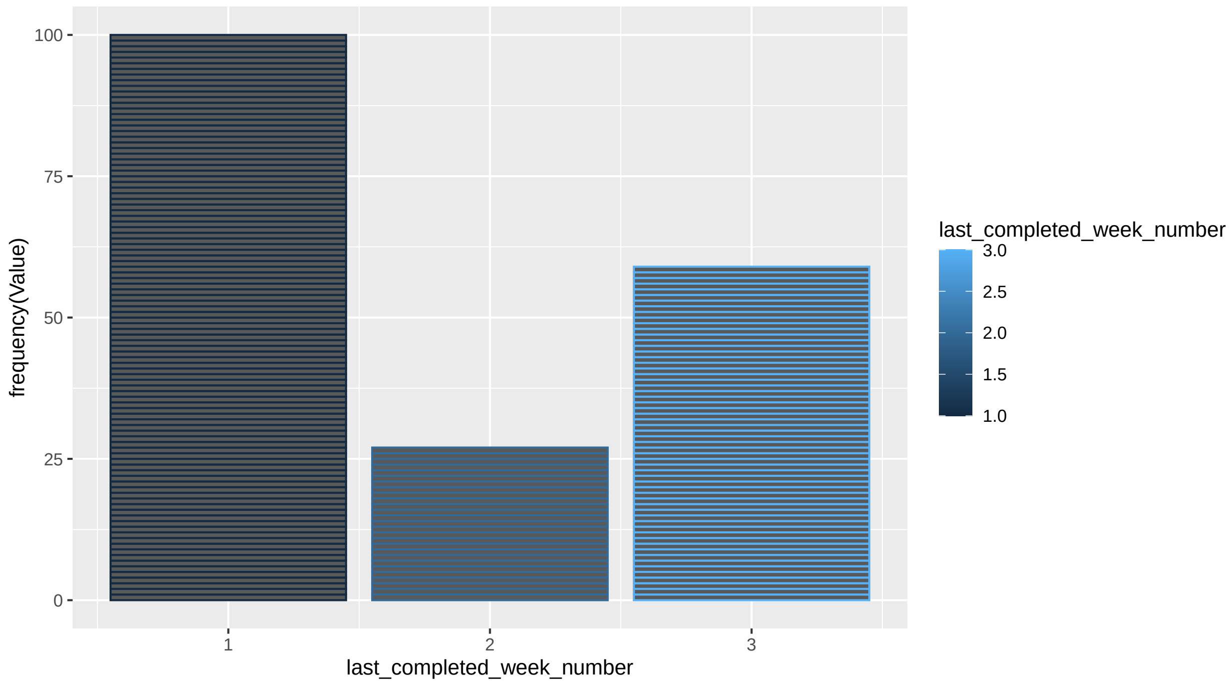


Figure 5: Frequency of individuals leaving course against week number,

Figure 5 describes the number of individuals who gave leaving responses between the years 4-7, and the week number in which they left. It is evident that the most frequent week for leaving responses is week 1. It could follow that week 1 is the most responsible for students leaving the course, which is why more specific analysis was required for week 1 alone.

The leaving responses for week 1 only (between the years 4-7) were investigated further, to afford the breakdown of reasons students were leaving the course. (Figure 6)

The trends from the first plot continue here in figure 6. The most common reason given for students leaving the course is “I don’t have enough time”. Much like the first plot, more students said the course was too hard than too easy, therefore, to ensure more students can complete the course and do not leave early, FutureLearn should adopt a strategy to make the content from the first week easier for the pupils to understand. Sadly, it may not be optimal to provide extra content as a result of the large number of participants who gave the most popular reason of not having enough time. To consider this first week in more detail, a more in depth analysis of step activity returned specific results about which step numbers may be a cause for concern.

Step activity -

A key aim from FutureLearn’s perspective is increasing student participation, ultimately leading to an increased number of individuals completing the course. To investigate the numbers of participation it was vital to analyse the step activity data. This data was collected electronically, providing information about when students were first visiting step numbers and when students were completing each step (if at all). The data collected in this area could demonstrate if there were any steps that were particularly hard to complete. In an instance where large numbers of students are unable to finish certain steps, these steps should be modified appropriately to increase users engagement.

##

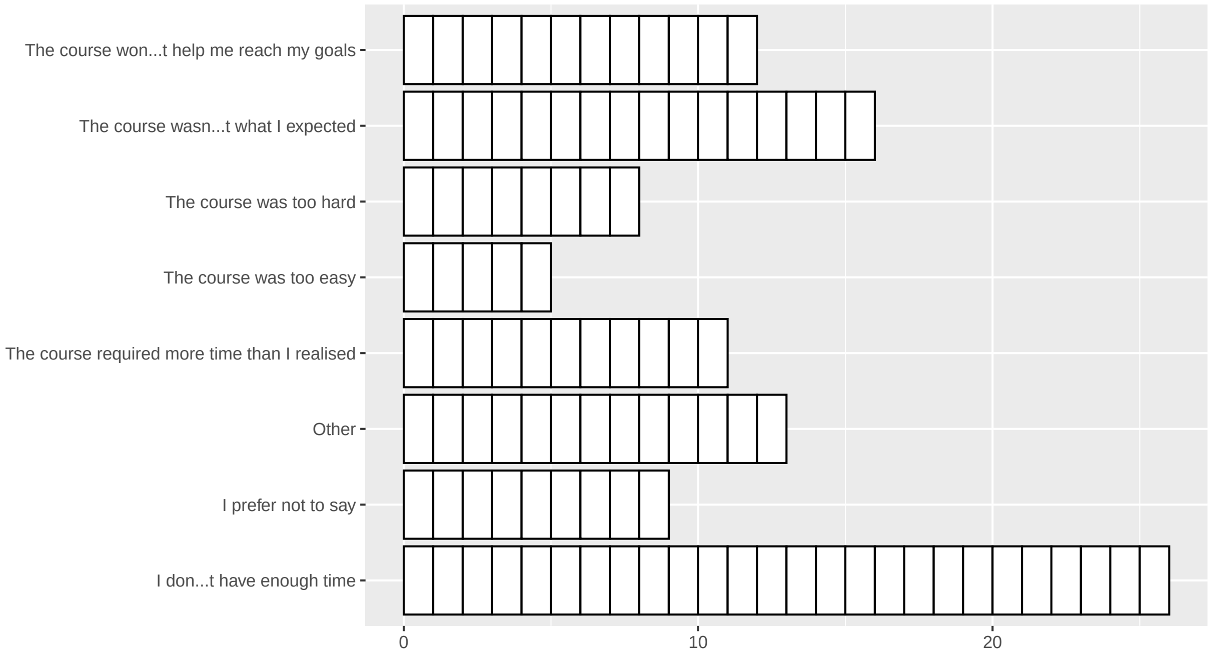


Figure 6: Frequency of Leaving reasons across week 1

```
##   1.1 1.11 1.12 1.13 1.14 1.15 1.16 1.17 1.18 1.19  1.2  1.3  1.4  1.5  1.6  1.7
## 2928  666  644  631  639  592  573  562  550  562 1444 1297 1132 1050  914  851
##   1.8  1.9  2.1 2.11 2.12 2.13 2.14 2.15 2.16 2.17 2.18 2.19  2.2 2.21 2.22 2.23
##   790  746  867  388  383  370  367  357  364  355  353  349  816  349  347  341
##   2.3  2.4  2.5  2.6  2.7  2.8  2.9  3.1 3.11 3.12 3.13 3.14 3.15 3.16 3.17 3.18
##   457  459  438  419  415  401  400  634  307  304  302  304  299  292  292   28
## 3.19  3.2  3.3  3.4  3.5  3.6  3.7  3.8  3.9
##   282  617  321  325  326  322  317  316  318
```

The table above displays the number of students who participated in we each step for year 6, i.e., how many students are starting each step. It would be worrying for the business if after certain steps values dropped significantly. This would represent a problem specific to this step, and it would be vital for FutureLearn to counteract this. The output does not show us clearly whether or not this is occurring within the dataset, therefore more analysis is required.

To provide more insight binary encoding was completed over the time stamp column of “Last_completed_at”. This meant the cells in which NA points were residing have all been set to F (False/0) and any cell that did not contain “NA” was replaced with a value of T (True/1), allowing us to record percentages of completion for each individual step.

The data from years 6 and 7 is the most recent available to us. Therefore, in some respects, it is the most appropriate data to consider for investigation. The step activity data from both years was considered and comparisons were made between them. In order to visualize them more appropriately, the week numbers have been paired together, therefore week 1 for both years 6 and 7 will be the first two plots seen, and so on.

WEEK1

Figure 7 describes numbers of completions for step numbers in week 1 for both year 6 and year 7 of the course. As we have seen from the LEAVING RESPONSE data, week 1 poses the largest problem for FutureLearn, as it was the week in which most of the students were leaving. The number of students unable to complete the first step has decreased from year 6-7 which is a massive positive for FutureLearn. However, the number

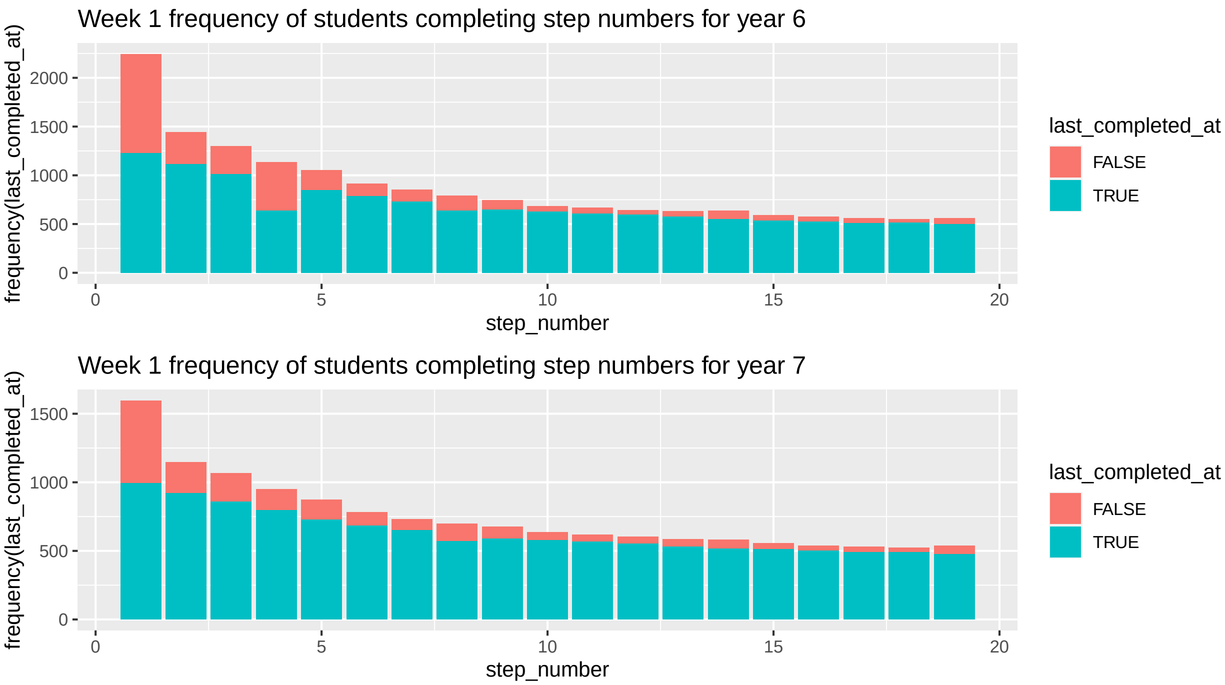


Figure 7: Week 1

is still far too great. This trend of large quantities of individuals not completing steps continues throughout the first 10 steps. Although step 4 does show some improvement the students who could be losing interest in the course are too numerous during this first week. The analysis has revealed that the first week should be the focal point for most of the changes to the course in the future.

WEEK2

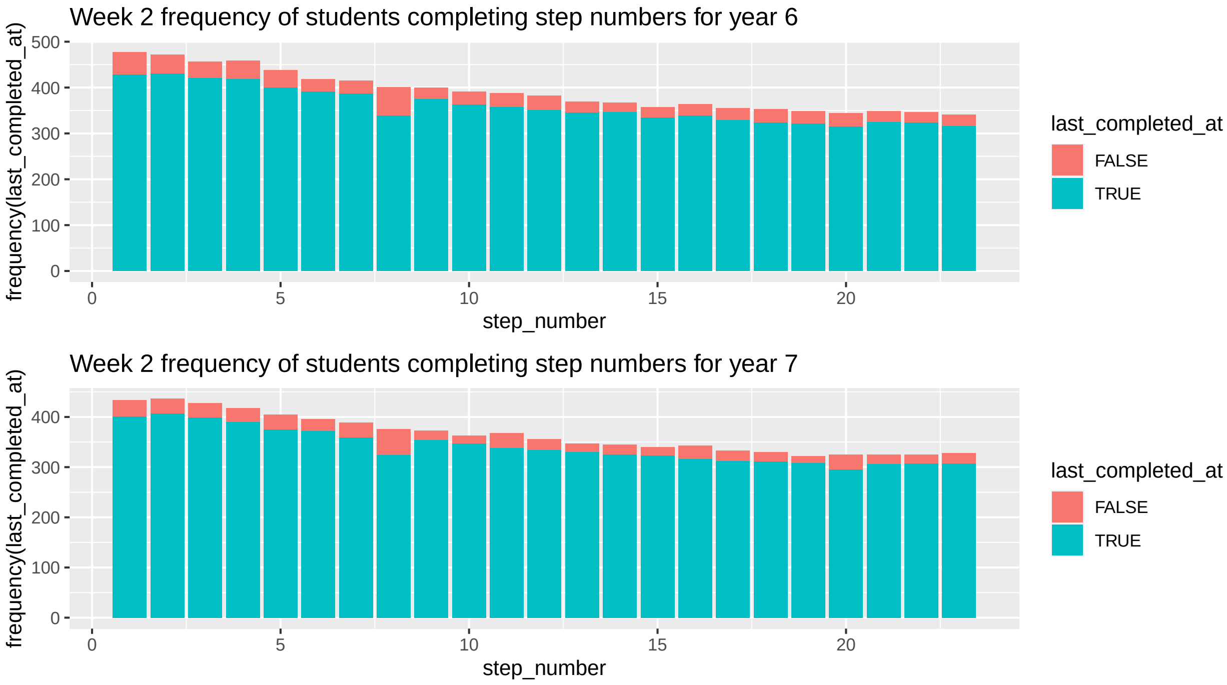
Figure 8 describes numbers of completions for step numbers in week 2 for both year 6 and year 7 of the course. It is expected of the second week of this course to show fewer individuals leaving. Compared to week 1 it is certainly better, however, a downward skew (slope) is still evident across this set, which equates to students steadily leaving the course still.

It is evident that steps 1, 2 and 8 in particular are failing. Participation numbers for step 3 are much lower than in 2, therefore, even though step 2 looks as if it is allowing a good percentage of completions between this step and the next there is a loss of interest. This is the case for both years. The business should be looking to minimize the number of students who are leaving the course. To address this, students should be asked about the highlighted steps.

WEEK3

Figure 9 describes numbers of completions for step numbers in week 3 for both year 6 and year 7 of the course. Step 11 should be the focal point for any changes within this week. This week shows a plateau of student losses as there is no downward slope present. Of the 3 weeks this is definitely the least in need of adjustment. It can be assumed that step 18 is anomalous, as from step 19 the numbers are increased again.

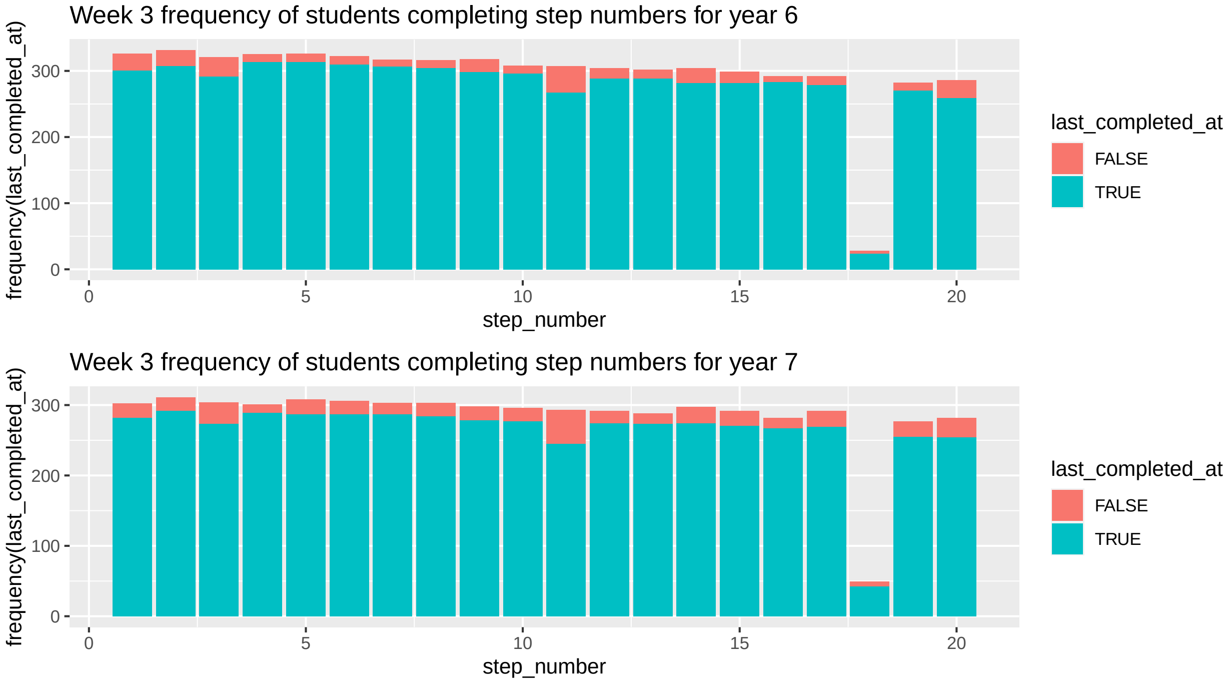
As we have discussed, it is important for FutureLearn to be aware of any specific steps that show a large number of students starting and not completing them. Steps that were most concerning throughout all 3 weeks were the 1-5 and 8 of week 1; steps 1,4 and 8 from week 2; 1,3 and 11 from the third week. FutureLearn should consider these steps specifically.



Week 2 frequency of students completing step numbers for year 6

Week 2 frequency of students completing step numbers for year 7

Figure 8: Week 2



Week 3 frequency of students completing step numbers for year 6

Week 3 frequency of students completing step numbers for year 7

Figure 9: Week 3

Video statistics -

What can the video statistics tell us? Does video length impact student's concentration and willingness to finish them?

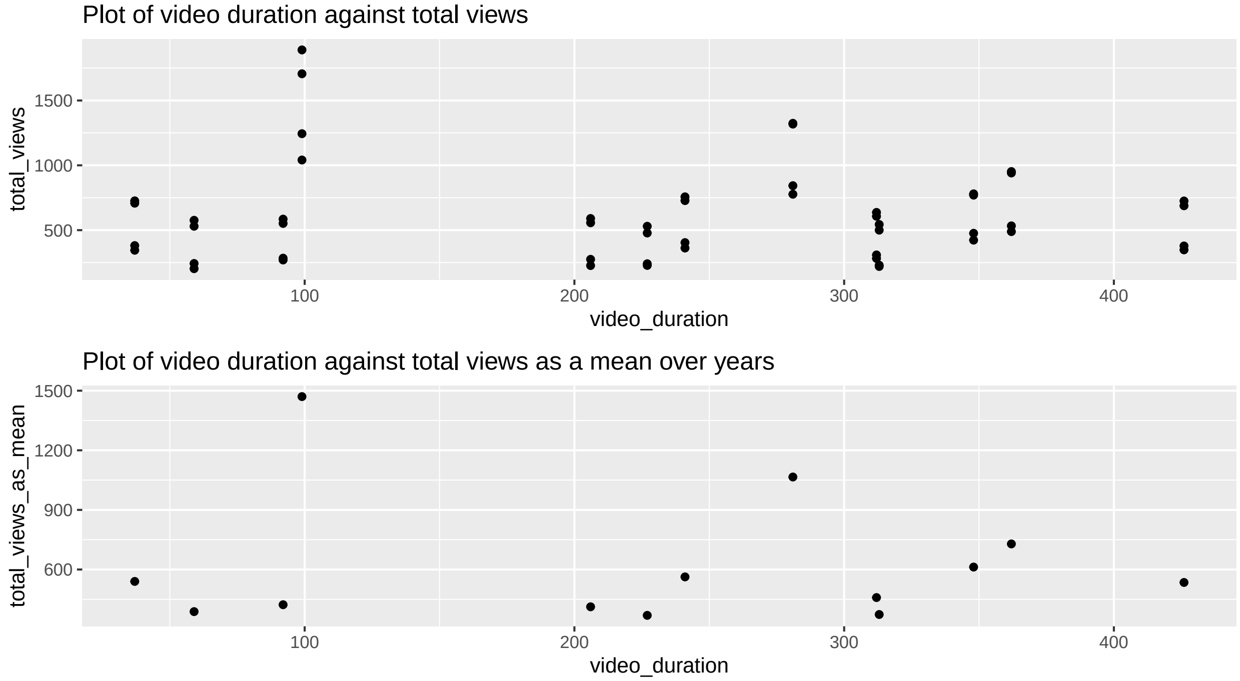


Figure 10: video duration plots

For figure 10, The above figure (top) is a plot of video duration against total views. This output suggests video duration does not effect total views as there is no correlation between the two variables graphically. To describe this more effectively, means were collected for each video, the plot showing this is represented by the bottom figure.

Total views does not relate to individuals who watched the whole video or even the majority. Therefore, a more appropriate investigation would be to see if video length effects the percentage of students who watch the video through to the end.

```
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```

Figure 11 is a graphical representation of the relationship between video duration and mean values for percentages of individuals who have watched 100% of the video. The trend shows an interesting dip around 300 seconds. It is unlikely, however, that this trend is a causal relationship, this is because this data studied only shows the frequency, not the percentage.

Student behaviors can often not be explained and sometimes show no pattern at all. However, this can still prove valuable information. In the case of video analysis, it would be very worrying if there was a very strong correlation between video length and student engagement. Therefore, FutureLearn can see these findings as a positive representation of their business

Implementing the CRISP-DM framework, it is important to consider other aspects of student behavior in relation to the video statistics. For example, what devices are being used most frequently each year? This question was explored and the results afforded are featured in figure 12.

Figure 12 shows the comparative frequencies for the devices of 'Desktop', 'Mobile' and 'Tablet' over each of the 5 years we have video statistics for. The black trend line shows the skew of individuals who are

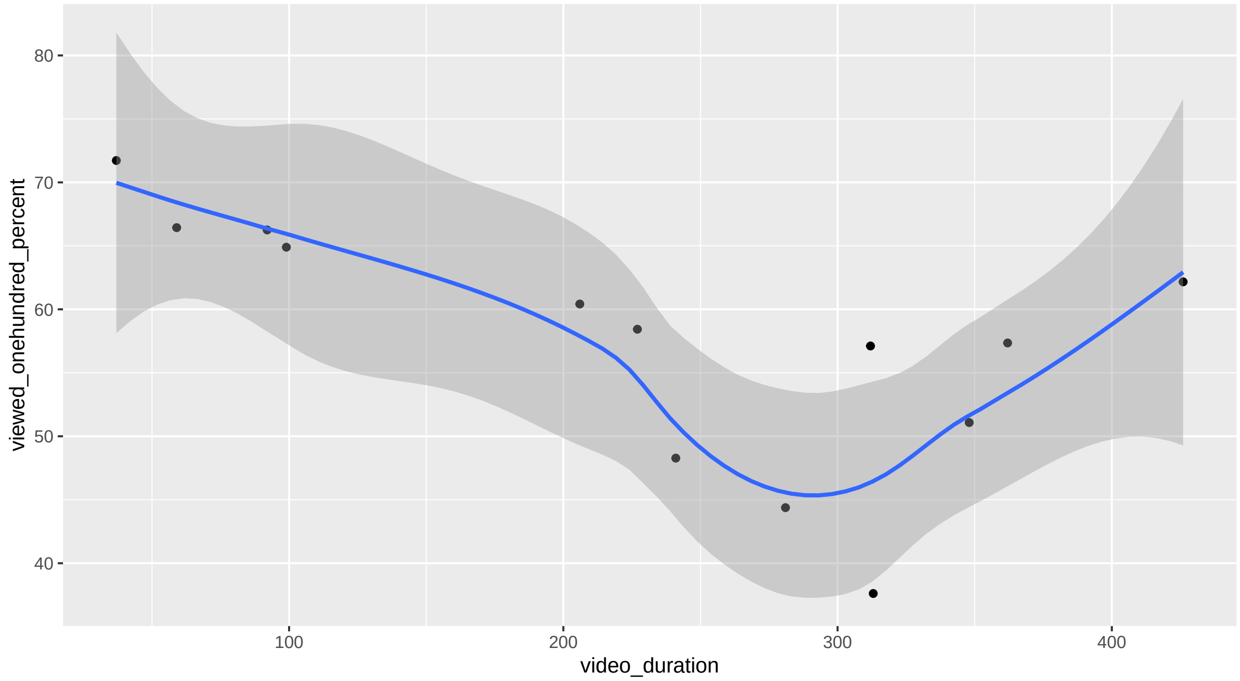


Figure 11: Total full interactions against video duration

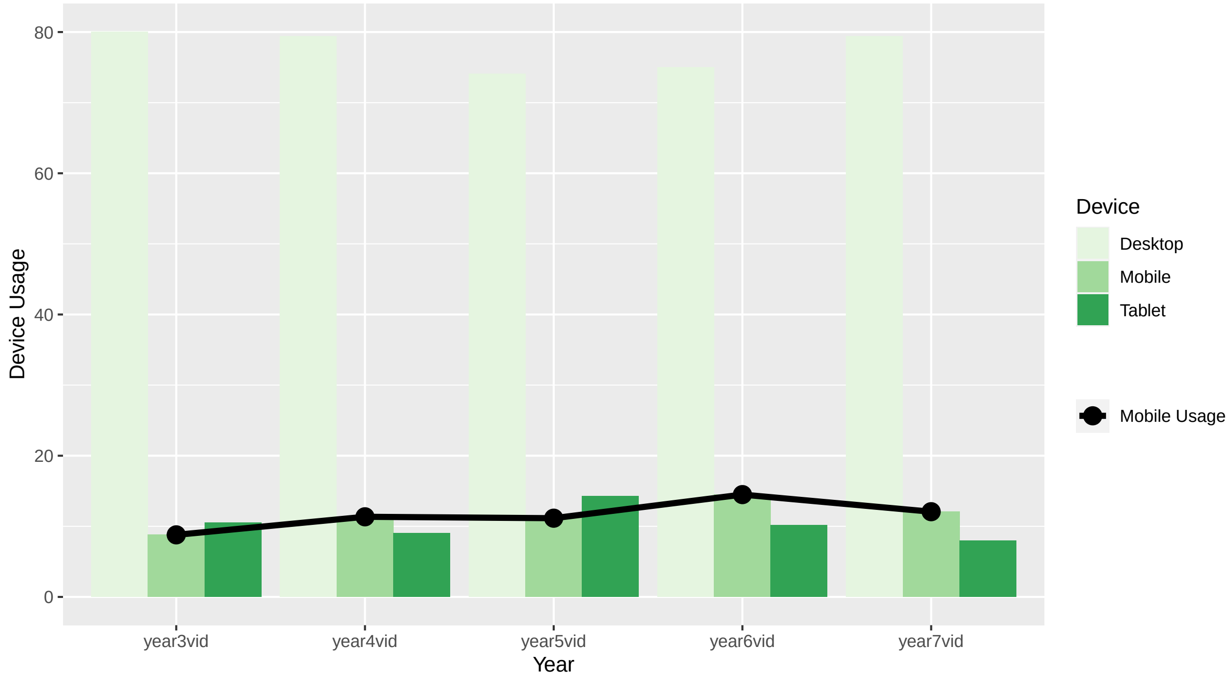


Figure 12: Device usage against year

choosing to view the videos on their mobile devices. Although the majority are choosing to utilize their desktop devices for such activities there is an obvious trend that an increasing number of students are using their mobiles. In light of this it may be a wise decision for FutureLearn to properly optimize parts of their course to be suited to mobile usage as well as regular desktop usage. They should strive to advertise this to the students, ensuring they are aware the course can be completed in parts on mobile devices. This may in turn impact those who have stated in leaving responses "I don't have enough time", if they realise they can complete the course on-the-go without caveats.

Does time of submission effect outcomes? -

Investigation was also carried out on student behaviors centered around the times individuals were submitting question responses. To understand indicators for successful learning outcomes is of utmost importance for FutureLearn. This analysis is useful for both students and the business, for example if the results were to show that early submission had a large correlative effect with getting incorrect answers: FutureLearn could reiterate to the students that taking their time with questions is vital. Moreover, if the students were privy to some of the outputs it would decrease the likelihood of them failing due to early submission.

To better understand student behaviors, some basic analysis was completed to compute and identify patterns in how the individuals were submitting their work. The first studies undertaken in this area were to see if students were more likely to submit answers in the morning or afternoon, and if there was any correlation between this and their responses.

```
## [1] 8105
```

```
## [1] 13011
```

There is a far greater number of submissions occurring in the afternoon/evening. Statistically speaking the afternoon accounts for nearly 62% of all submissions. However, for this one data set there are only 21000 pieces of data, data from the years following this were also joined together in order to verify if this statistic was accurate over a greater time period.

The pie chart (Figure 13) represents the percentages of submission numbers based on the time of day. It is clear there are a much larger proportion of students submitting work in the pm. This would suggest students tend to be more active in the afternoon and the evening, if there are any live parts to the course, it would therefore be most effective to hold them in the afternoon and the evening.

Following from this, plots were afforded to visualize how the time of day at submission effected results. (Figure 14)

(NOTE VARYING Y AXIS for figure 14)

```
# table for percentages for both am and pm
percentage_table_am_pm
```

```
##                am       pm
## %_TRUE  55.49257 57.04922
## %_FALSE 44.50743 42.95078
```

The output from the above code chunk displays a percentage table for correct and incorrect question responses between, and including, years 4-7 from the question response csv files provided by FutureLearn.There is minimal difference between the time frames. Although it does appear that students who submit answers in the pm show better percentages, the difference is only marginal (less than 2%). This output does not warrant any further investigation as to whether the time of day effects submission results.

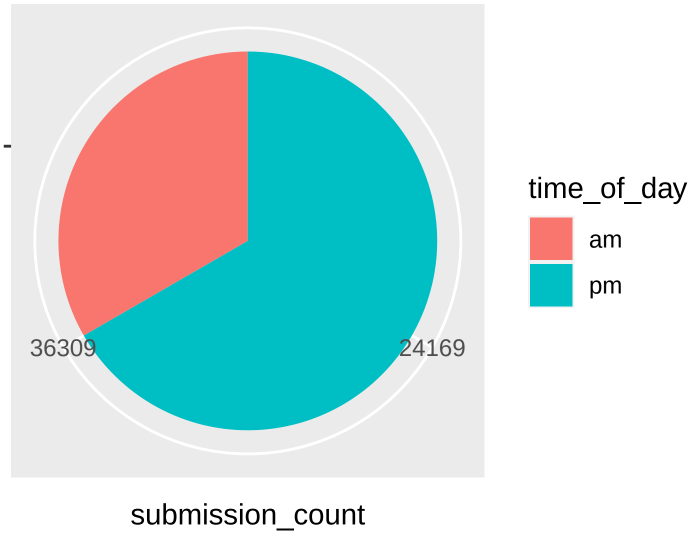


Figure 13: Chart of submissions by time of day

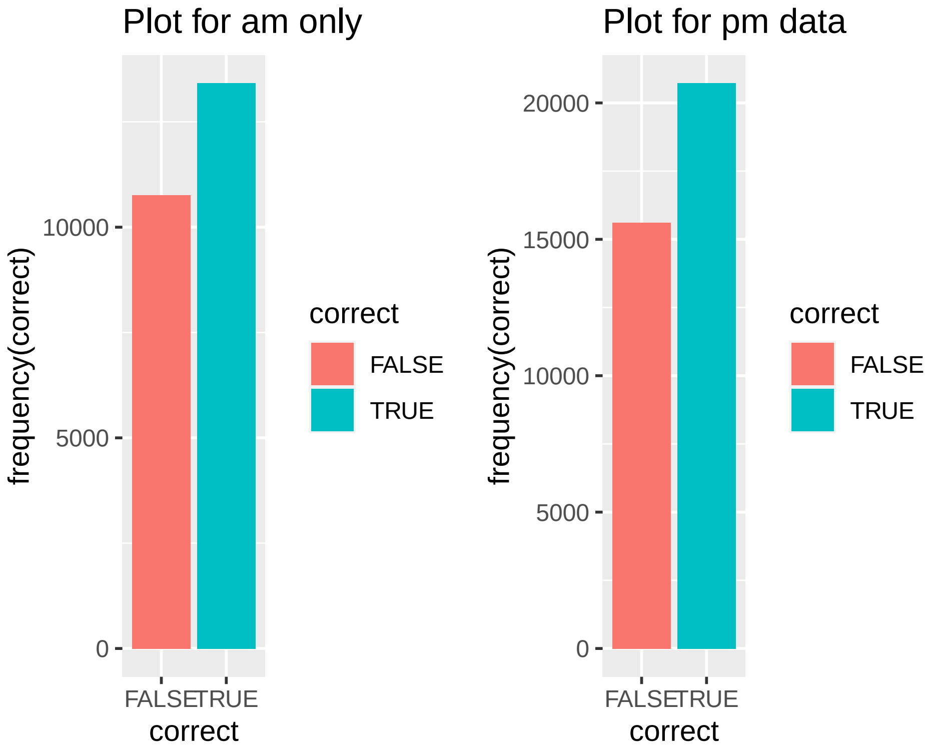


Figure 14: AM/PM timeframe plots of T/F frequencies

Despite the expected results over the study of time of day, the question response csvs were not useless, more analysis was completed as they had a substantial amount of data to offer. Another investigative technique performed over the question response csvs was to plot the frequency of correct and wrong answers over time. (Figure 15)

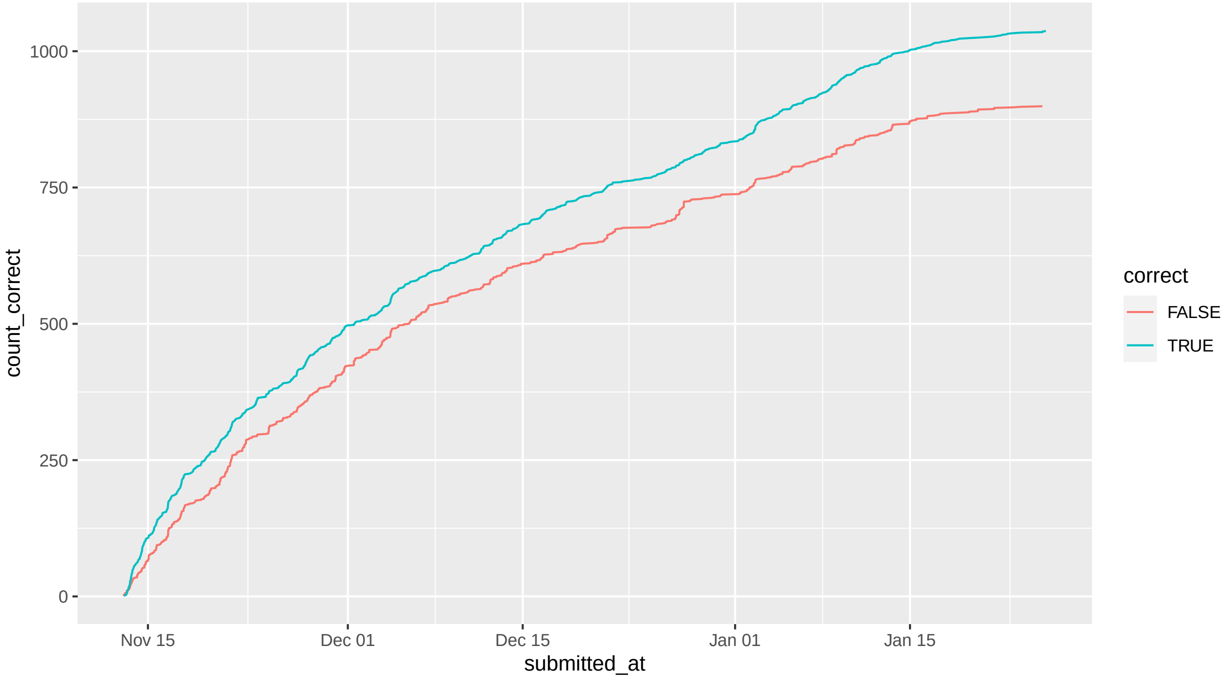


Figure 15: Plot of T/F answers against submission date for question 1

Figure 15. For the first question from the data collected during year 4, there appears to be a very small correlation between time and correct submissions. The cumulative total of false responses starts to slow closer to the deadline date, thus the gap between TRUE and FALSE answers increases further after December 15th.

To see if this is the case across the whole data set more investigation was required. As a solution, a collection of plots was produced all considering the same variables. (Figure 16). (Included in appendix).

Figure 16 (Included in appendix) shows graphical plots of questions 2,3,4 and 6 for the 4th year of the course. Both question 2 and 4 follow a similar, although more exaggerated, trend to that of the question 1 responses. The closer to the submission deadline students answered the question, the more likely it was for them to achieve a correct answer. As we have discussed, one of the goals for FutureLearn is to ensure their students are successful. With this in mind, it is of the upmost importance to encourage students to not rush through questions. The plot for question 5 from this data set has not been used, this question was anomalous for the purpose of our investigation (all individuals who participated achieved a correct answer).

Conclusions -

Whilst some areas of the analysis yielded unsurprising results, the information afforded over this report could provide some useful guidance to FutureLearn. Investigative techniques and methods that reveal patterns around student behaviors are necessary to improve the course.

From the analysis of both leaving responses and step activity it is obvious that the first week of the course is responsible for the loss of the majority of its students. For example, during the step activity analysis week 1 showed a clear negative skew and the large amount of students who started step 1 (upwards of 2000 for year 6) had dwindled (to just higher than 500) by the end of the week. The business should make a more clear attempt to ask those in week 1 what changes could be made to keep their learners engaged and wanting to participate.

Temporal analysis of submission results showed some correlation between early submission and greater chances to produce incorrect answers. FutureLearn should reiterate the importance of taking time with questions to ensure students are achieving the highest marks available. This may also have a positive effect on the leaving response data. It was clear that a greater number of individuals left the course because it 'was too hard' (26 between years 4-7) than the number who left because it 'was too easy' (18 over the same period).

Key takeaways:

- Important to continue research in this area

- The first week of the course proved to be the most problematic, therefore improvements should be centered around week 1

- Step activity analysis returned the steps that were most concerning are: 1.1-5 Primarily, and also: 1.8, 2.1, 2.4, 2.8, 3.1, 3.3 and 3.11

- Early submissions can effect student grades, the students should be made aware of this fact

- The course should take into account the number of students who leave the course due to not having enough time, and they should produce countermeasures to alleviate this

- University degree holding students are the courses main users, FutureLearn should take this into account when altering the course

- There is a clear rise in frequency for individuals using mobile devices to complete parts of the course, therefore the course should be appropriately optimized.

Appendix

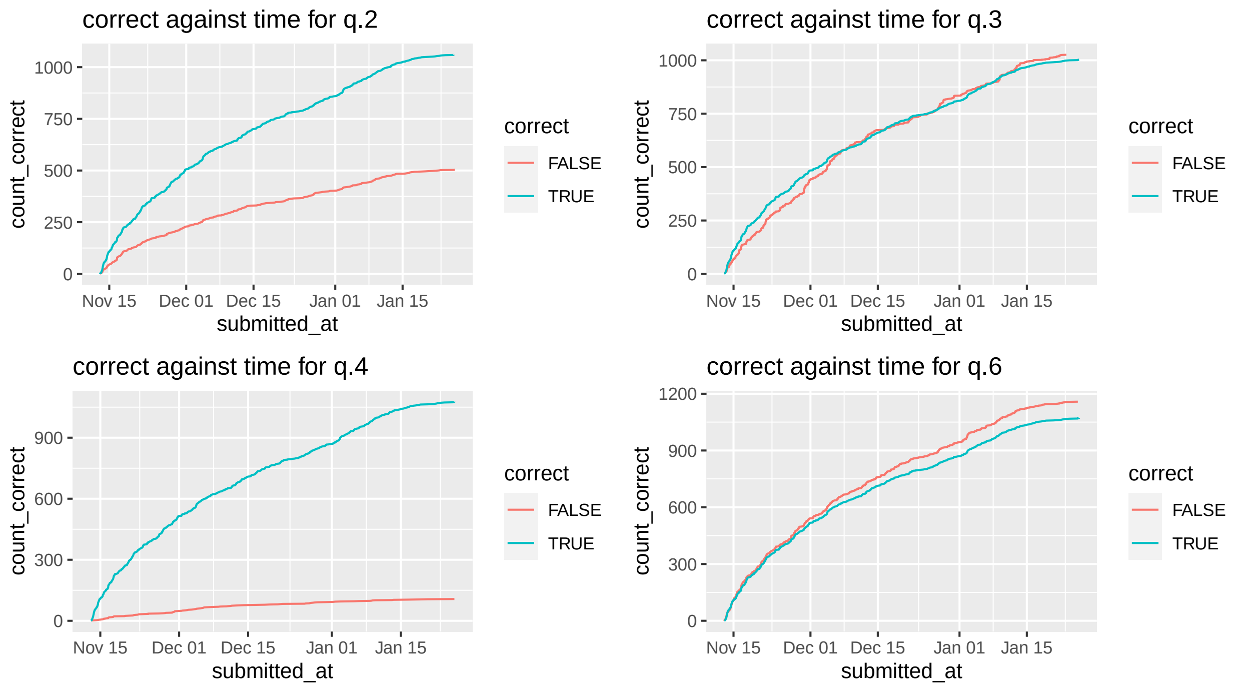


Figure 16: Plot of T/F answers against submission date