

Learning Analytics

Exploratory Data Analysis on the cyber security course from FutureLearn

Thamizhiniyan Pugazhenthhi - 200941620

02/12/2021

INTRODUCTION

We generate data on practically every aspect of our lives, which can be harnessed and analysed to reveal powerful insights about our behavior, preferences, and future actions, and the amount, velocity, and diversity of data we generate is fast increasing. Students in higher education leave a digital footprint during their studies that tells us about their learning and experiences at university. Universities can use this information to better understand how students learn and improve the student experience. This is referred to as “Learning Analytics”¹. Learning analytics data could act as a measure of the learner’s commitment to learning, training and employment of staff and teaching intensity.²

DATASET DESCRIPTION

In Future Learn MOOC data set³, we have different sets of data files from seven different runs of cyber security course on different time frames from future learn website. Each set contains enrollments, question response, step activity, team members, video stats, weekly sentiment survey, leaving survey and archetypes.

Enrollment data file contains the data about each person who enrolled in each run of the course. The column of the data file reads the designed id to each enrolled, the time stamp which they enrolled and un enrolled, role of each enrolled, fully participated at, purchased statement at, gender, country, age range, highest education level, employment status, employment area and country from which they are accessing from. Question response data file contains the data about the questions answered by each learner in the quiz they had attended. The columns are learner id, number of the quiz question, type of the question, week number, step number, response they had given, submitted at and whether it is correct or not.

Step number data file explains the data about the step each learner had visited. The data are learner id, step number, week number, first time they had visited the step and the time they had completed the step at. Weekly survey sentiment data file contains the data about the rating and review after each week. It reads the columns id, responded at, week number, experience rating and reason behind the rating.

Leaving survey data file contains the data behind leaving. It explains the data about the learner id, time they left at, leaving reason, last completed at, last completed step number and week number. Video stats contains the data which deals with the video associated with the course. It has everything about the videos like the question it associated with, title, duration, total views, downloads, caption views, transcript views, HD view, percentage of the video viewed – five percent, ten, twenty-five, fifty, seventy-five, ninety-five, hundred, console in which it had been viewed – device, desktop, mobile, TV, tablet, unknown data and the continent from which it had been viewed – Europe, Oceania, Asia, North America, South America, Africa and South America.

Learning Archetypes represent user behavior and general characteristics, and an archetype can be a general guideline about end user’s behavior and are general guidelines on how a user behaves⁴. The types of archetypes are Fixers, Vitalizers, Advancers, Preparers, Explorers, Flourishers, Hobbyists⁵. Learning

Archetypes data file segregates learners with different archetypes. It contains the data which has columns – learner id, responded at and the type of archetype the learner is.

From the data set, we will draw different analysis in terms of similarities, differences, positives, negatives, and the improvements they can make along with the recommendations we can tell the future learn team to enhance the user experience.

EXPLORATORY DATA ANALYSIS

For this analysis, we will analyse about the enrolments, question responses and leaving responses. With this, we can find the number of correct and wrong responses given by learners, when they left the course, which point in the course they left from, the reason why they left at first place.

First, we will analyse the number of people enrolled in each run of the course. By looking at ‘Figure 1’ plot, we can say that the enrolment numbers decreasing with each run. The number of enrolled in final run is $1/7^{\text{th}}$ of the first one. This shows that the people lost interest in the course over the period. This is very alarming as the team at Future learn should look at it carefully and increase the number of courses relevant to the current market.

Moreover, they need to elevate the standard and difficulty of the course. They should be flexible as people from different background can pick any course with the difficulty level they wanted for each course. It will change the appeal of the course among the learners. The course should be engaging and will be an interactive one to be in connect with the audience from the start to end.

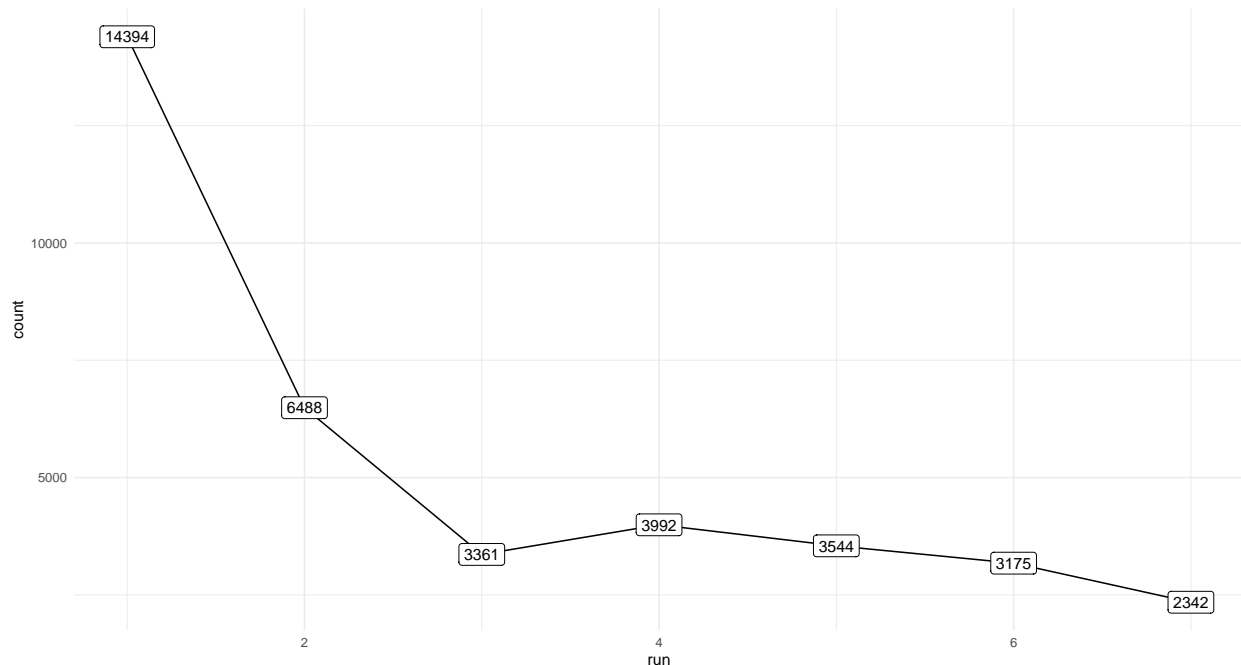


Figure 1: Comparing number of people enrolled in each run of the course

Now, we will consider enrolled people from all seven runs of cyber security course as a single data frame for some plots and as a separate run for some plots. We will classify the enrollments into four categories – age, gender, highest education level and employment status and plot the data to get clear view about the trend of enrollments. When we see all four plots, one factor that is very evident is the rate of “Unknown” entries across all four columns of enrollment data. The team at Future Learn need to get all details of an enrolled learner without any compromise. They are able to understand the trend of learners and their background if they get those details. Moreover, it will improve the team to concentrate on areas in which they are lacking and improve even more on areas in which they are good at.

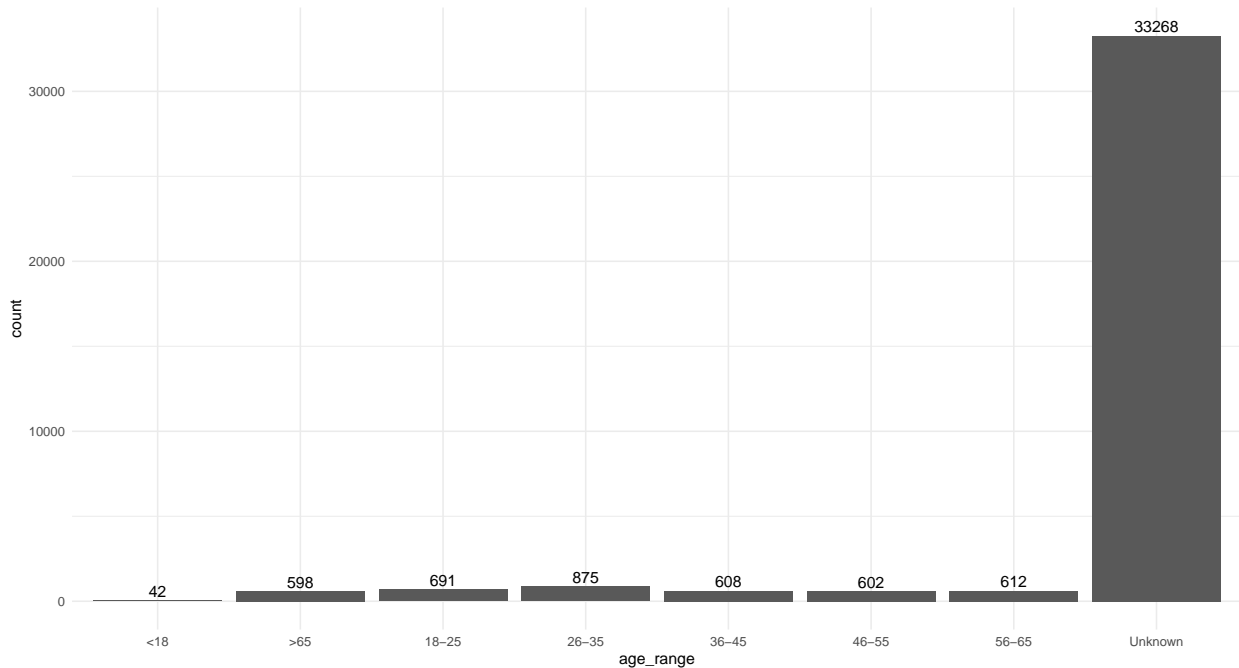


Figure 2: Comparing learners from different age group

From the ‘Figure 2’ plot regarding the age range, we cannot see any discrete differences in the block of age range across ages. Leave “Unknown” group aside. If they set the personal details entry as mandatory, we can fetch even more clear picture about the age range. They should create even more courses or tweak cyber security course according to the learners who represent majority of the age range.

From the ‘Figure 3’ plot regarding the gender of the learners, we cannot see any obvious differences between the gender - male and female. It is good since the team don’t need any promotional strategy to attract any set of people. On the other hand, non-binary and other gender people had enrolled in lesser number compared to other genders. The team need to work on promotional strategy to cover non-binary people since we as an organizer cannot neglect people of any population. But again ‘Unknown’ played major role in disturbed our analysis to some extent.

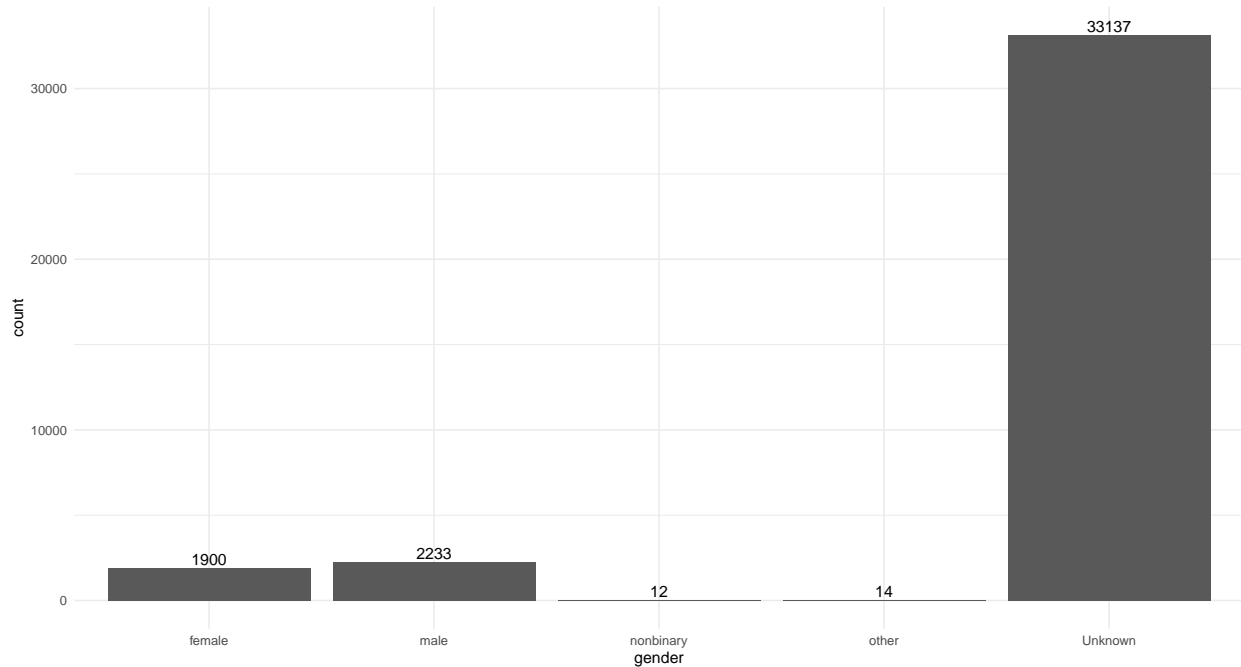


Figure 3: Comparing learners from different gender

From the ‘Figure 4’ plot about the learners from their highest education level, we can see that most of the learners are from the university - bachelors, masters, and doctorate. But there is a smaller number of learners enrolled from the school level, the team should investigate this to shift their focus towards the school students. The Future Learn team should work on content which attracts school students which is the biggest market to capture. Courses like GRE, GMAT, SAT, TOEFL are famous among school students and the students from all parts of the world will enrol the course since it has overall appeal across the globe.

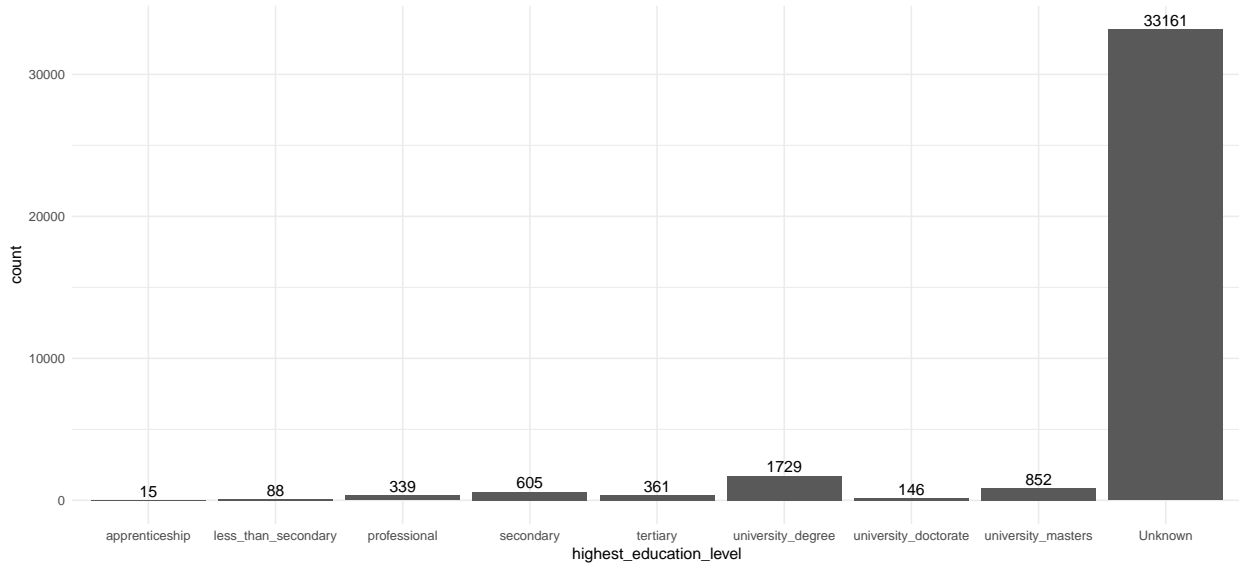


Figure 4: Comparing learners with their highest education level

From the 'Figure 5' plot about the learners with their employment status, we can see that it is obvious that the highest numbers of learners who are in their full-time job. But it is very strange that the second highest number of learners are retired people. As usual, the students had enrolled in the course in high number. But they need to concentrate on people who are not working or looking for job since the number is very low. We can explore that market considering the people are unemployed usually needs something to study or learn new to enhance their careers. The number of self-employed is comparatively less and we can pull people who are self-employed by including the course dedicated to them.

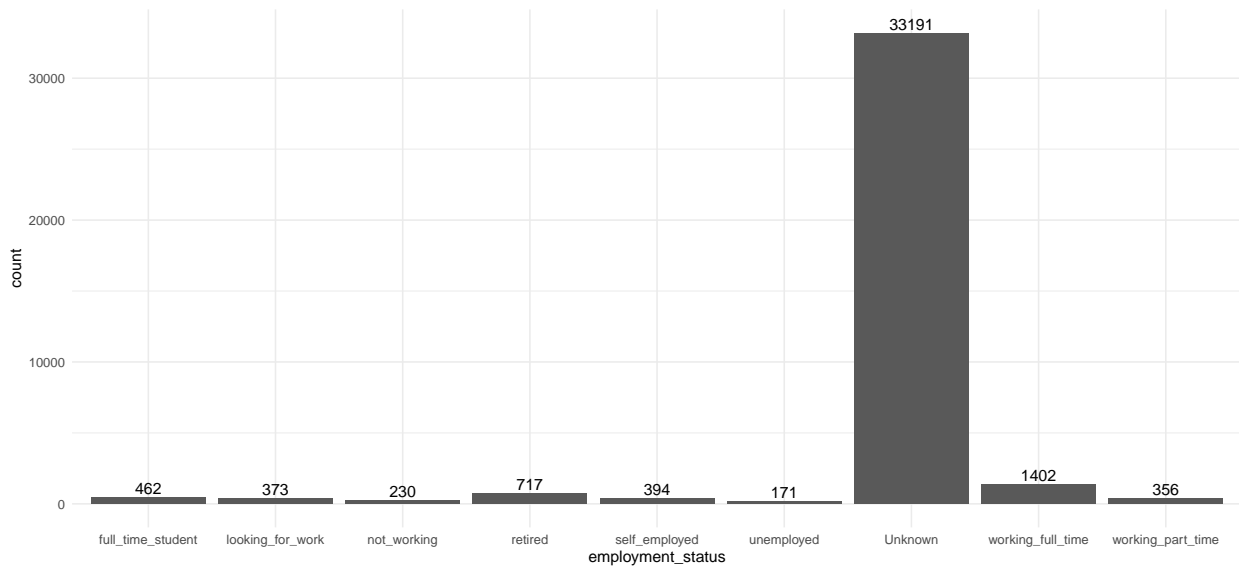


Figure 5: Comparing learners from their employment status

Now, we are going little deep into the analysis by connecting enrolment data set with question response. Here, we can calculate different aspects of the connection between the learners and how they respond to the question from the course. As we saw earlier, we can find how the learners approaches the course and how far they understand the concepts by the response they had given for each quiz associated with the step.

In this step of analysis, we will compare the number of correct responses with the number of wrong responses. From ‘Figure 6’, it is positive to some extent as the number of correct responses is higher than the number of wrong responses. But the negative point here is the difference between them. since the difference between them is non negligible, the team must take this very seriously. The number of wrong responses shows that many learners do not understand the course well enough to give the right answer. Since the quizzes were multiple choice questions, we cannot even that as a credit. There is a chance that the right answer could be a luck. It is understandable that the correct and wrong responses will not affect the business at Future learn in day-to-day basis. But if you look at the larger picture, it is not good in terms of knowledge sharing.

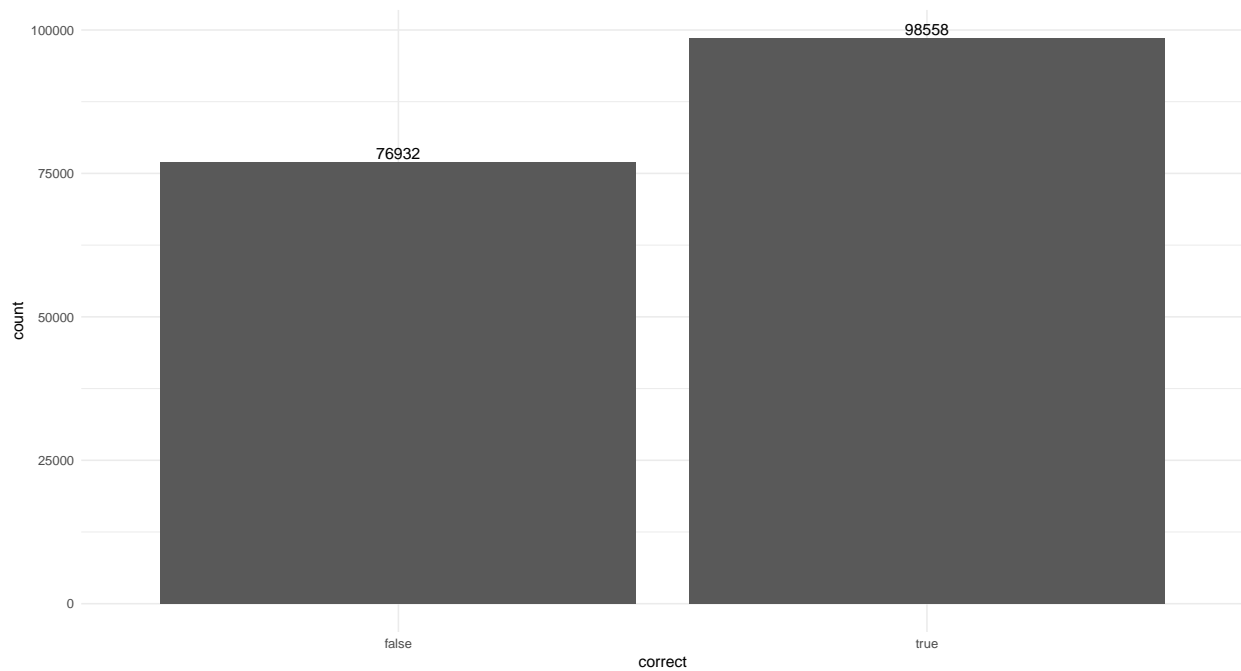


Figure 6: Comparing number of quiz responses with the response

From ‘Figure 7’ plot, we will compare the same correct and wrong responses but with each question. We can see that number of wrong responses is higher in the later part of the course. It is very evident that the learners are losing interest in the course as they travel with the course. The team should improve the course structure in the “Security in the future home” part of the module.

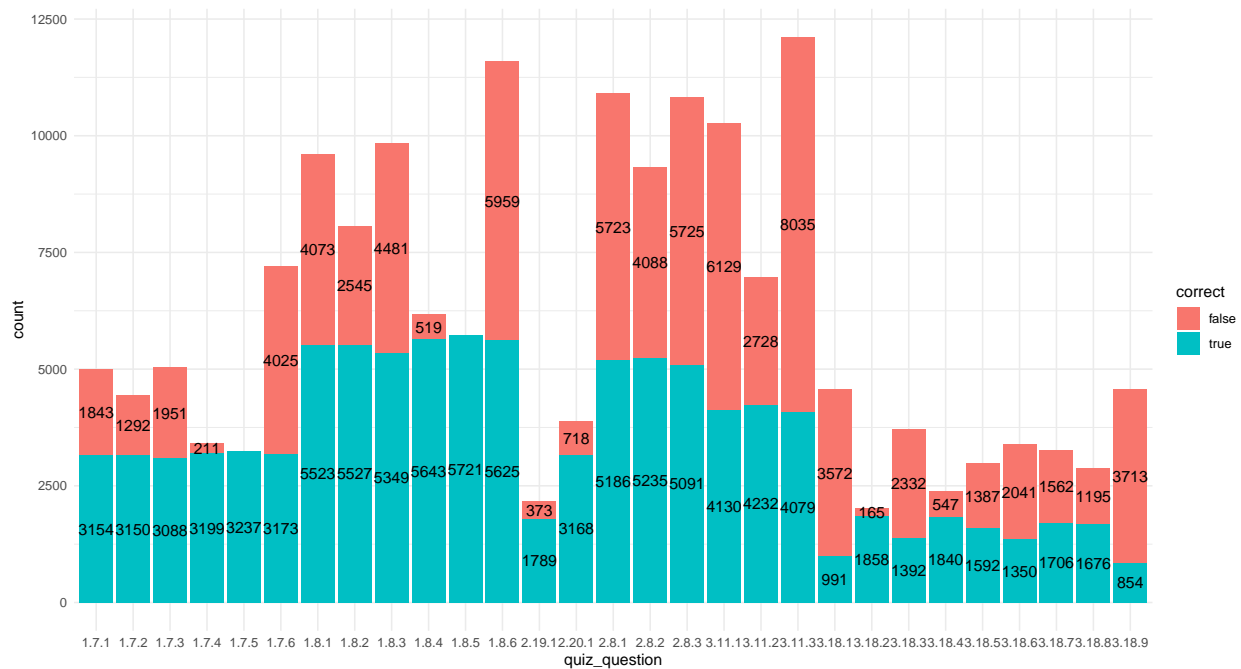


Figure 7: Comparing number of quiz responses for each question

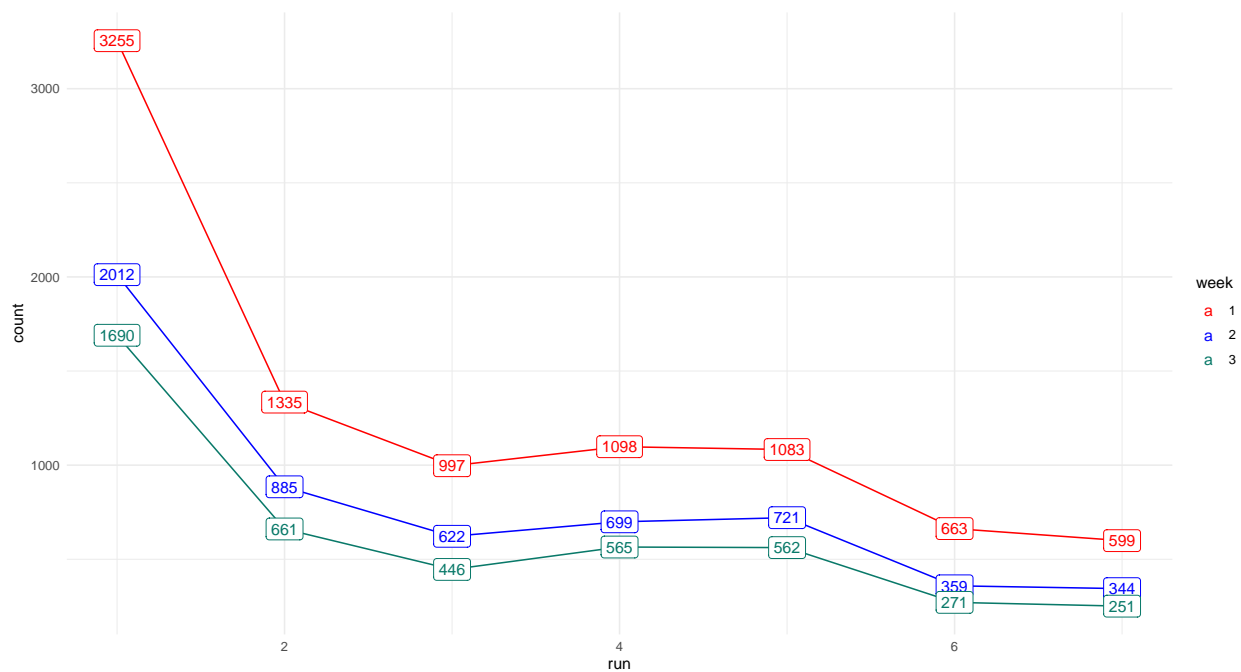


Figure 8: Comparing number of correct responses in each run of the course

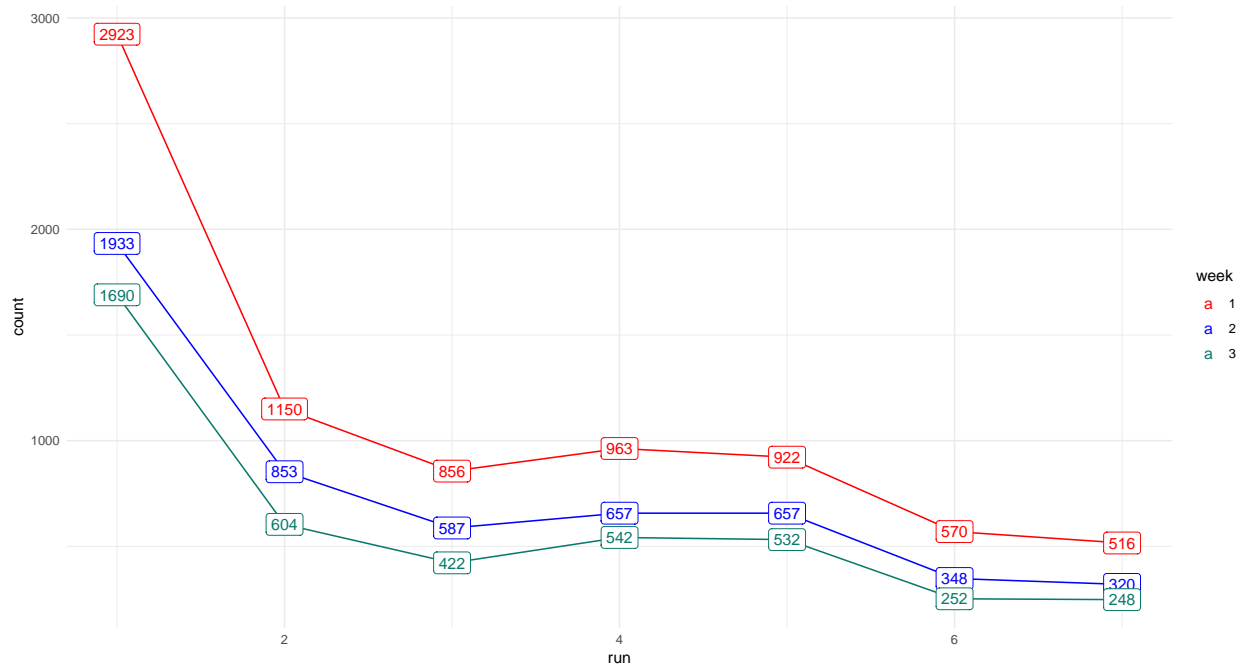


Figure 9: Comparing number of wrong responses in each run of the course

From the ‘Figure 8’ and ‘Figure 9’ plots, we can see the number of correct and wrong responses for each run respectively.

Now, we are going to connect enrolments and question responses with leaving responses. It will broaden our analysis to certain extent. As we saw earlier, we can find the reason why they left from the course, when they left, and from which point of the course they left.

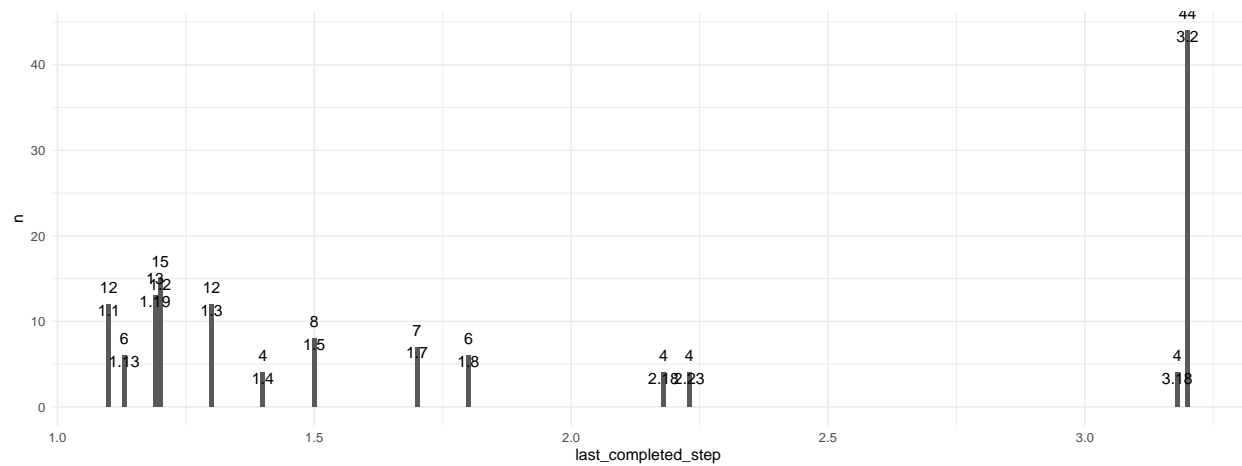


Figure 10: Comparing number of people left the course in each step of the course - top 10

From the ‘Figure 10’, we can say that most of the learners who left from the course were left at the step 3.2. Around 44 people left the course at that step. It must not be a coincidence. When we look at the step in depth, it has a video explains “Devices in the future home”. There is a chance that the video might be boring, or the course might hit a low point in terms of interaction and engagement at that point. The team at Future learn should take this seriously and work on this step a little more compared to another step.

Also, we can notice there are people who left the course in initial stages in considerable number which is also a bad factor. There might be several reasons which we will see later but this is not good. They got bored in the initial stages and left without experiencing the complete essence of the course.

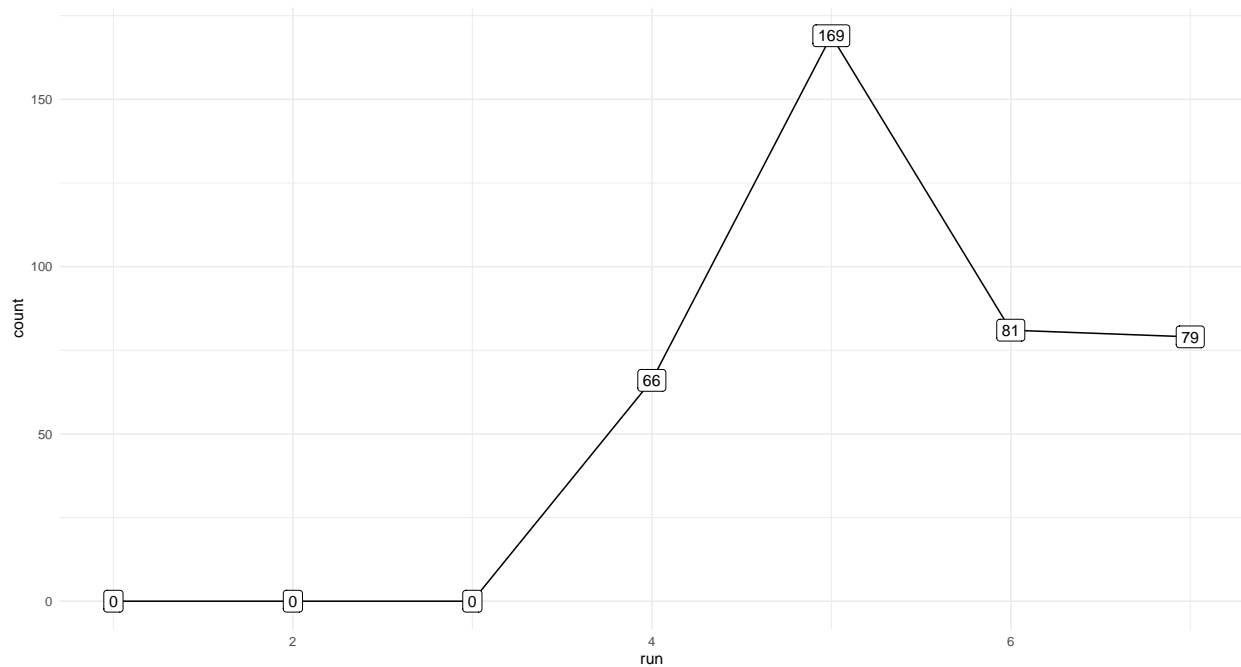


Figure 11: Comparing number of people left the course in each run

The plot ‘Figure 11’ explains about the number of learners left the course in each run. Since we do not have any data about the people left from first run to third run. We will consider the data only from fourth run to seventh run. The number of people left in fifth run of the course is higher than other runs even though the enrolled are comparatively lesser than the fourth run. The difference between people left in fourth and fifth run is more than 100. We can confirm that the fifth run had not performed well and produced good results compared to other runs.

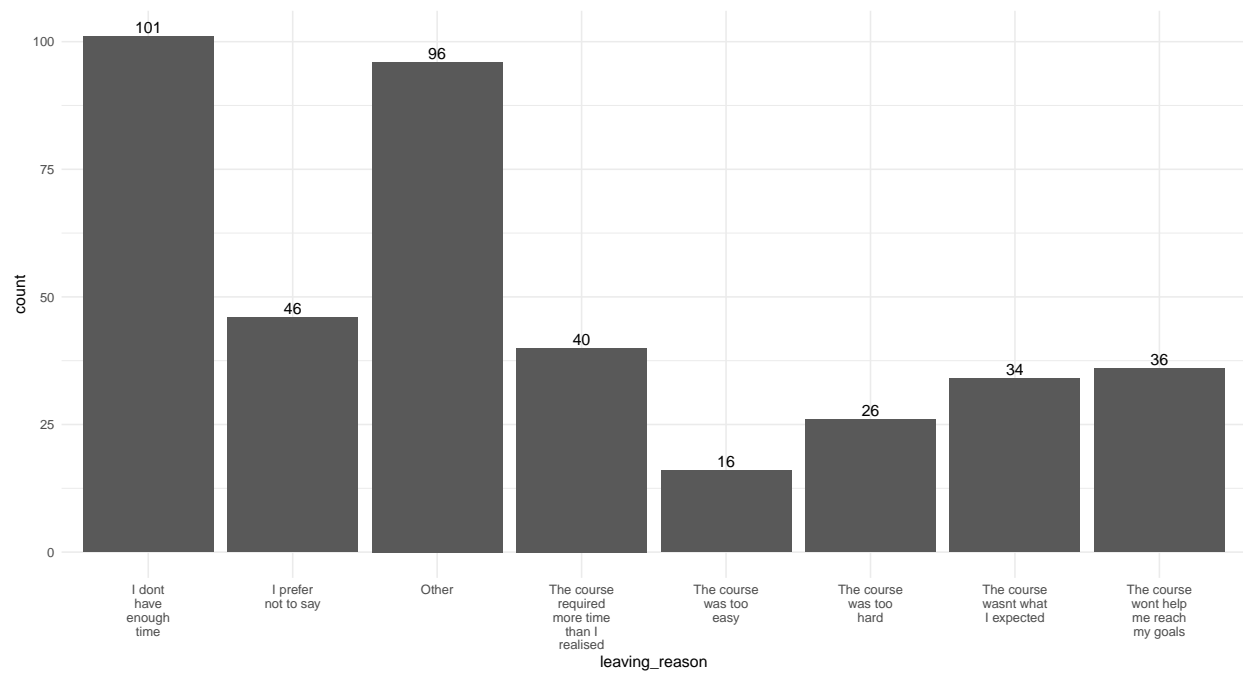


Figure 12: Comparing number of people left the course with their reason for leaving

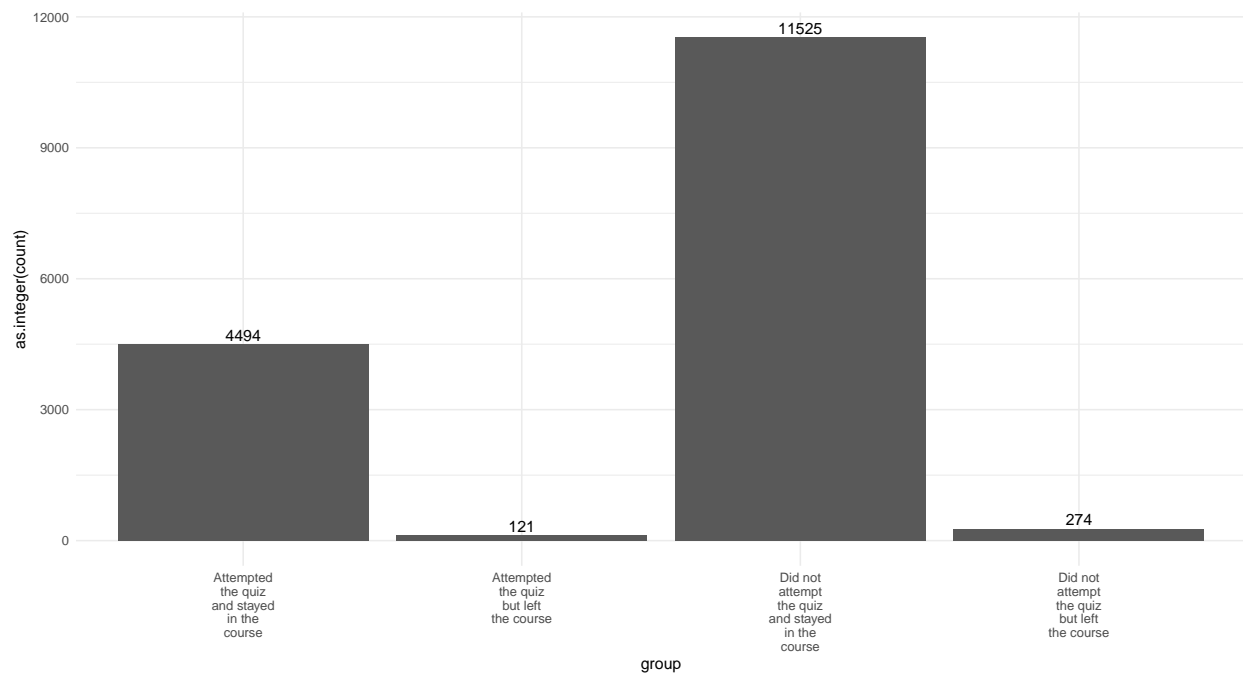


Figure 13: Understanding learner's behavior

From ‘Figure 12’ plot, we can identify the number of learners left the course with the reason of leaving they mentioned. As we can see, the highest number of learners left the course because they do not have enough time. The team need to make the course flexible according to the time the learner has. The fast-track batches are one such recommendation they can consider. It is closely followed by the reason ‘Other’. Each learner would have different answers to choose the ‘Other’. We cannot neglect that, but we do not have enough data to analyse it. The reasons ‘I prefer not to say’ and ‘The course required more time than I realised’ was selected by 46 and 40 learners respectively. This denotes that many people consider the course as too long to complete.

If we bring the reasons ‘The course was too easy’, ‘The course was not what I expected’ and ‘The course will not help me reach my goals’ under same roof, we can say all these reasons explains more or less the same point and combined total of 86 which is an alarming number. The team should consider these reasons and think about increasing the difficulty of the course by including new and advanced concepts or design special courses for such people with advanced concepts.

Moreover, we grouped the learners into four groups - learners who attempted the quiz and stayed in the course, learners who attempted the quiz but left the course, learners who did not attempt the quiz but left the course and learners who did not attend the quiz but left the course. The reason behind this grouping is to identify the learner’s behavior and the real reason why they left the course.

We can see this plot in ‘Figure 13’. Since the highest number of people who did not attempt the quiz and stayed in the course, we can say that there are many learners who are neither actively attempting the quiz nor they left the course. When we compare the people who left the course, we can say that the numbers are high in people who did not attempt the quiz but left the course. This is obvious that they do not understand the course or do not have interest in the course which probably be the reason behind the highest number in this group.

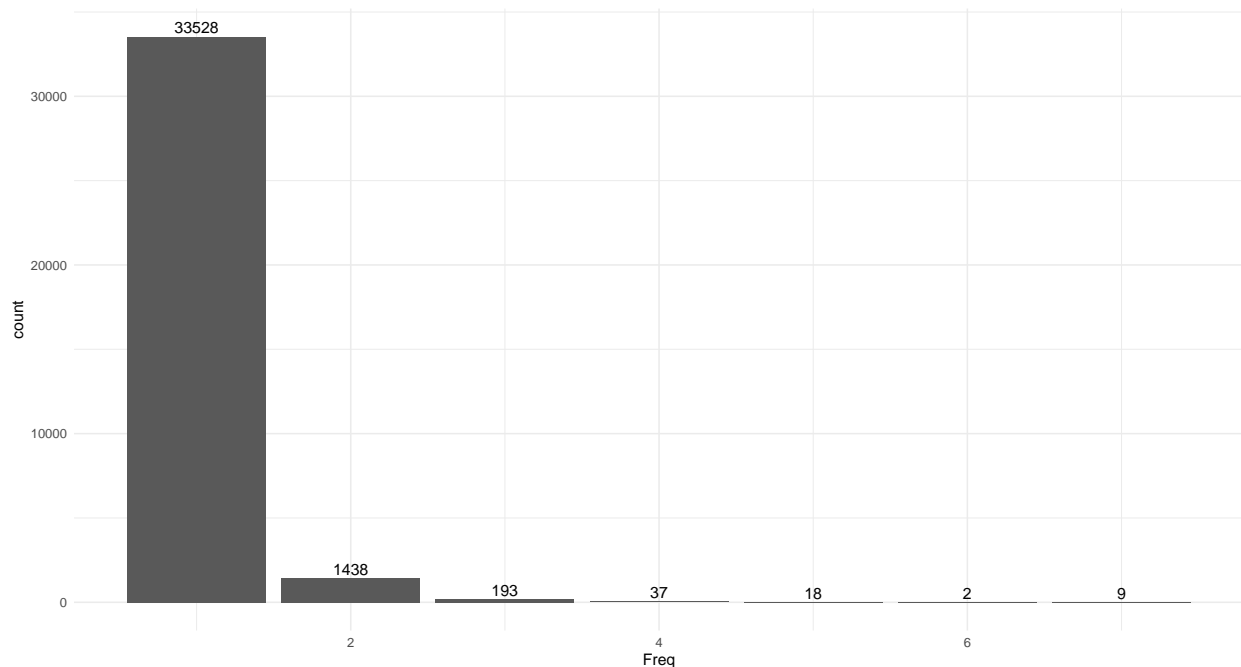


Figure 14: Comparing the number of people enrolled in all runs of the course

From ‘Figure 14’ plot, we can compare the number of people enrolled in all runs of the courses. As we can the number of people enrolled only in one run of the course is higher compared to other numbers. This is the expected one as this is the good sign since the number of learners who enrolled in one run of the course is more than 90% percent of total enrollment.

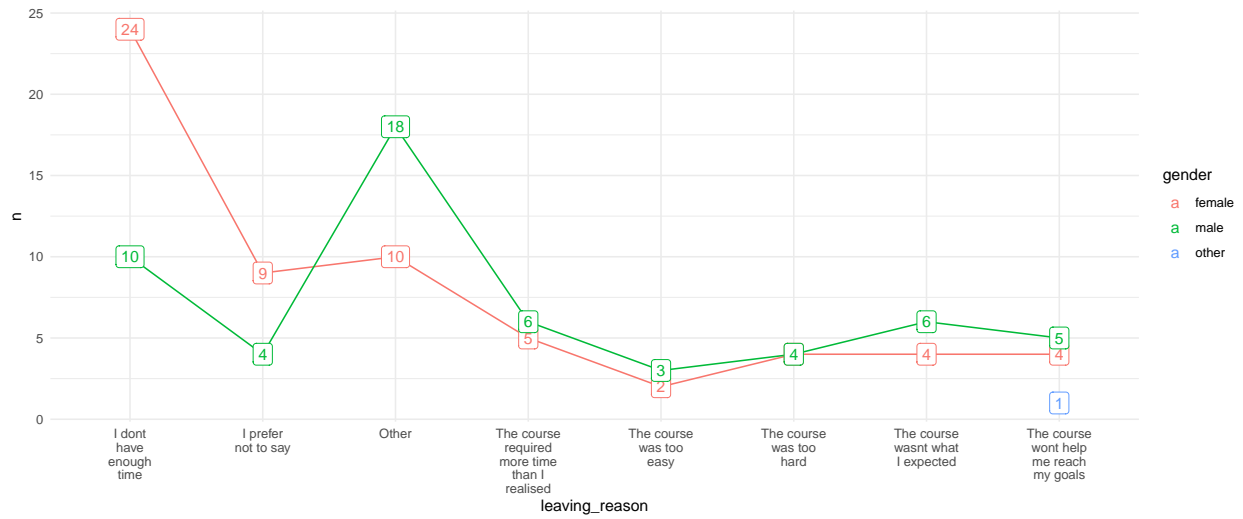


Figure 15: Correlating learner's reason for leaving with their gender

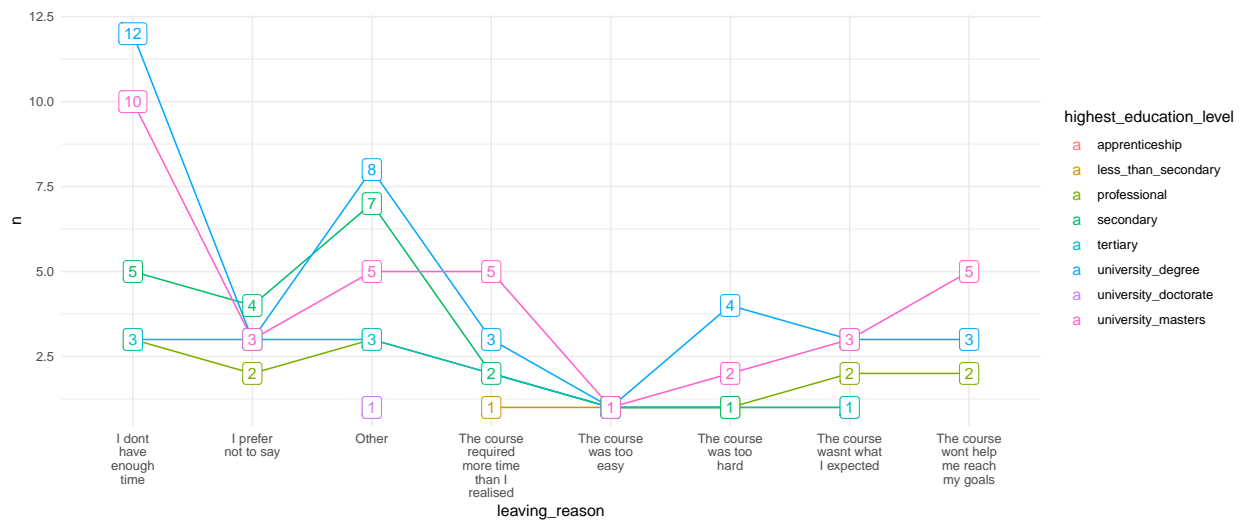


Figure 16: Correlating learner's reason for leaving with their highest education level

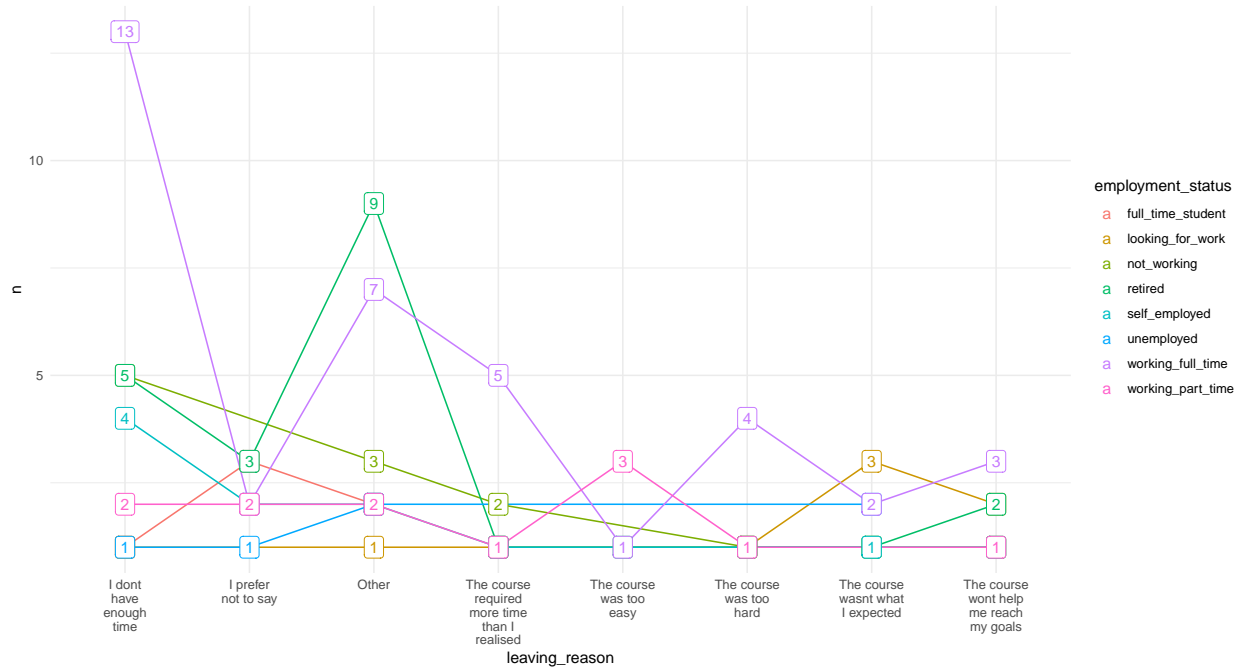


Figure 17: Correlating learner's reason for leaving with their employment status

Next, we will correlate leaving reason with gender, highest education level and employment status. The 'Figure 15', 'Figure 16' and 'Figure 17' plots deals with comparison of gender, highest education level and employment status respectively.

Let's consider one plot at a time. Gender comparison shows that the female left for the reason 'I do not have enough time' more than any other reason. But the reason 'Other' is more for male learners. They had different sub reasons for choosing 'Other'. The team should investigate this and can give a crisp version of full course according to their timetable. This will engage many female learners to actively enrol in their course with their preferred span.

The comparison between learners with their highest education level shows that the students at their university degrees did not have enough time to complete the course. It is acceptable since they already have their own curriculum to complete, and they are not considering this course as an extra knowledge gain. Like for female learners, it would be good if the team design a crisper version of course to fit within the timetable of learners who do not have enough time to complete the course. In addition to that, we can see around five people left the course with the masters education level with the reason that the course will not help them to reach their goals. It means that the course does not have a mass appeal to all possible learners. As mentioned earlier, the team should increase the difficulty of the course to avoid these kinds of comments from learners.

Moreover, the comparison between learners from different employment status shows that the people who are working full time do not have enough time to complete the course. For all these problems, the short and crisper version of course is the only solution. The team should really focus on the length of the course since it is the major concern among different sets of learners. The team should know that most of the learners will enrol in this course if they need any additional skill, if they want any promotion in their current workplace, if they want to spend their leisure time effectively by gaining some knowledge, if they need to crack any interview and so on. This will not be the primary focus in their daily routine. So, it would be ideal if the team concentrates on the duration of the course and include some advanced topics to cover users who already have basic knowledge about the topics.

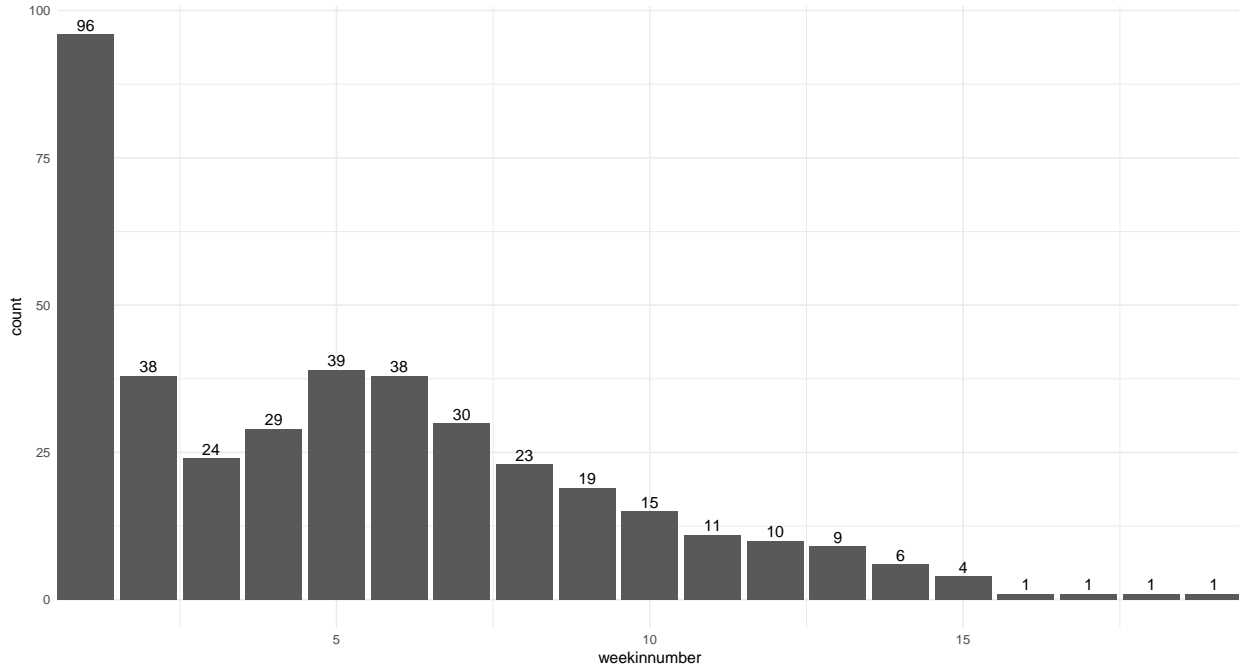


Figure 18: Comparing number of learners against number of weeks they stayed in the course

In this final part of analysis, we will compare the learners who left the course with the number of weeks they stayed in the course. We can achieve the result with the difference of the time stamp they left at, and the time stamp they enrolled at. With the 'Figure 18' plot, we can see that the number of people left at the first week is higher than the rest of the plot. It suggests that the learners who enrolled the course and left the course within a week when they see the length of the course and thinks that they could not complete the course or fit in the course within their timetable. This might be the reason of hike in the first week of the course. The course is designed to complete within three weeks but if we look at the plot, we can see that majority of the learners who left the course tried to finish the course by looking at the number who left the course after three weeks. They stayed in the course with a hope to complete the course. Since it was not possible for them, they left the course after a try.

RECOMMENDATIONS

From the exploratory data analysis, we have performed on the cyber security data set from future learn, we can say that the course performs moderate. It does not perform up to the expectations. We gave several recommendations throughout the analysis. Let's summarise all the recommendations here for team at future learn to understand effectively.

The recommendations are

1. Since number of 'Unknown' is high during enrolment process, it is very difficult to analyse the data. The team should set the personal details field as mandatory which will in turn easy to understand the user background and tweak the course and set the advertisements according to the requirements. This will cut the cost of the organisation on promotions and publicity.
2. The number of wrong responses in quiz question is little lower than the number of correct responses which is a good sign but not very good sign. To resolve this, the team should focus on the course content and design the course in such a way that the interaction with learners should be high. This will reduce the number of wrong responses in general.

3. Majority of the learners enrolled in the course without any active participation in the quizzes. The knowledge would improve only if they attempted the quiz in each step of the course. The quiz should become mandatory to test their knowledge in that topic and should set a pass mark to complete the course. Without completion of that course, they should not allow to navigate the next step of the course.
4. Majority of the learners left the course because they do not have enough time to complete the course. The team should design either the crisper version of the course or the course should be fast tracked one to fit in the learner's timetable. We should give options to the learners from which they can choose the duration of the course. It will make way for the people who have lesser time in their daily routine to enroll in the course.
5. The next highest reason for learners to leave the course is the difficulty. Many learners who studied till university degree consider this course as basic and easy. Their expectations are very high, but the course did not fulfill that. The team should either consider increasing the difficulty of the course in general or consider designing a course for the people who expect an advanced level of topics for the same course. It should be designed for the learner to choose from. Ideally, the course should be very flexible as the learner can choose the duration of the course they need and the difficulty of the course they want. This will fulfill all sets of learners from all over the world.

CONCLUSION

From the above analysis, we drew different analysis in terms of similarities, differences, positives, negatives, and the improvements they can make along with the recommendations we can tell the future learn team to enhance the user experience. With the recommendations, the team can improve the future analysis with the user background, improve the course content with necessary tweaks in terms of interaction, make the quizzes mandatory and can design the course as flexible as possible. This will bring them more new learners from all over the world with correct promotion strategies and most importantly, the people who enrolled in the course will be very happy and satisfied with the way the course has been designed. The word of mouth will be good with more positive reviews which will in turn increase the mass appeal of the course in general.

REFERENCES

1. <https://www.policyconnect.org.uk/research/report-bricks-clicks-potential-data-and-analytics-higher-education>
2. BIS Green Paper (6 November 2015), 'Higher education: teaching excellence, social mobility and student choice', pg.27-28.
3. <https://www.futurelearn.com/>
4. <https://disprz.com/understanding-learner-personas/>
5. <https://ugc.futurelearn.com/uploads/files/2f/ce/2fce94dd-54e6-4614-972d-ce9c3a6d7d91/Learner-Archetypes.pdf>