CSC8631 Report

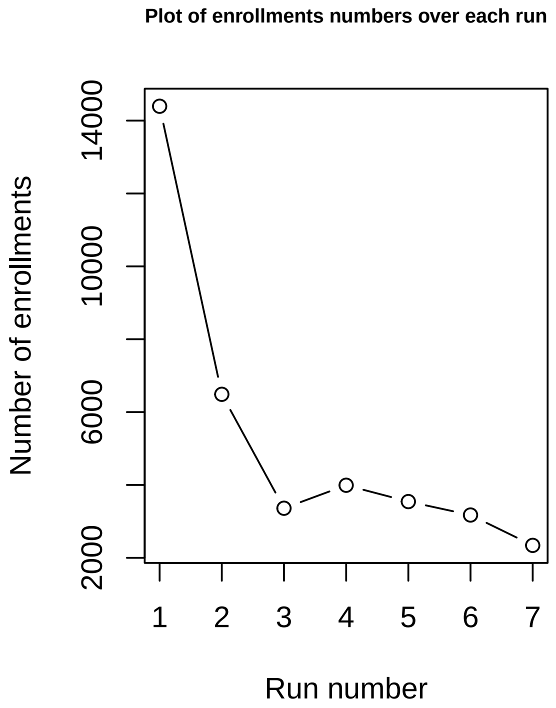
Introduction and business understanding

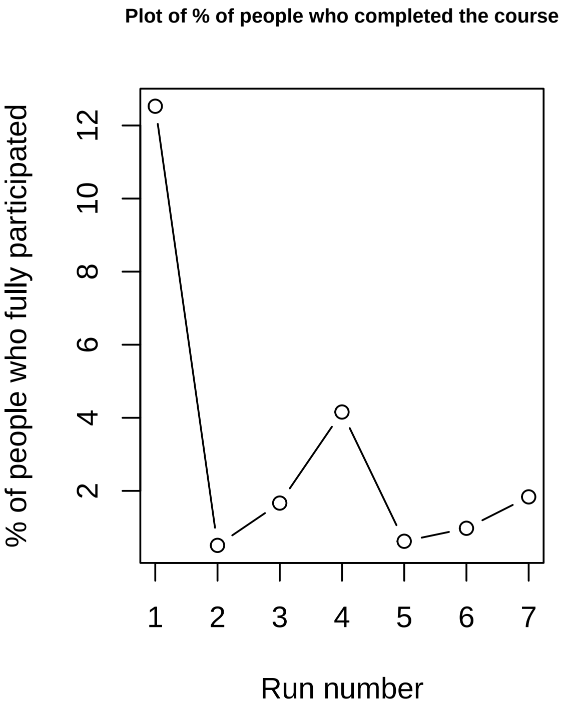
This project includes data from an online course provided by Newcastle University on cyber security. These data were collected over seven different runs of the course, with the first run starting on 5th of September 2016, and the seventh run on 10th of September 2018. The course is divided in three sections, with each section having specific steps that the students can engage in. These steps include videos, articles, quizzes, exercises and discussions were students can participate. Then at the end of the third section there is a test, which students can complete in order to test their understanding. In general all runs have the same steps, with runs 1 and 2 having some additional ones. Runs 3-7 have the exact same sections. The data collected from the seven runs, include enrollment numbers, demographics of enrolled people, video statistics, responses on quizzes and test as well as other information. The aim of this project is to analyse the data and extracting interesting insights that will be helpful for the course developers to understand better what works and what does not. This report will focus mainly on participants engagement and will give insights into what affects engagement. This analysis will aim into providing the course developers, with suggestions on which age group, or gender might be more interested in the course, which sections are more popular and what sections might need to be removed or adjusted, due to low engagement.

Data Understanding and Preparation

CRISP DM cycle 1

As mentioned before there are different data sets for each run. In general for all the runs most data sets are common. For examples for all runs there data sets that include information on the enrollments, data sets about step activity, question responses as well as survey responses. For runs 3-7 there is also an additional data set that includes statistics about the videos. For this analysis the main data sets used are the enrollments, step activity, and video statistics data set. Here it should be noted that some of these data sets contain unknown row entries. These unknown entries will not be removed from the data set, since by removing them, some conclusions reached may not be representative of the data. However, for some analysis these unknown or empty rows will be removed. So after having mentioned the above information, one of the first things to look at is the number of people involved in each run, the number of people that completed the course. In order to generate the plots below all the seven enrollments data sets were used:

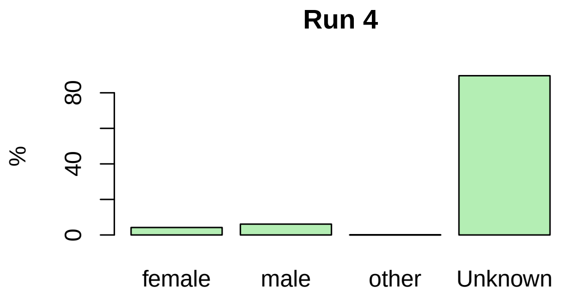

Plot of enrollments numbers over each run

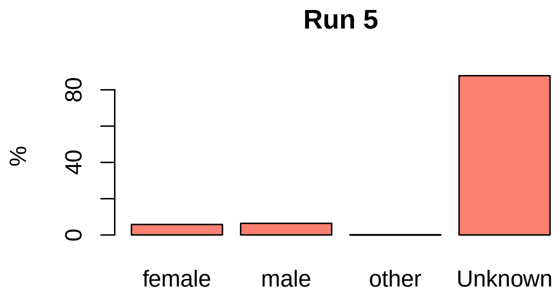

Plot of % of people who completed the course

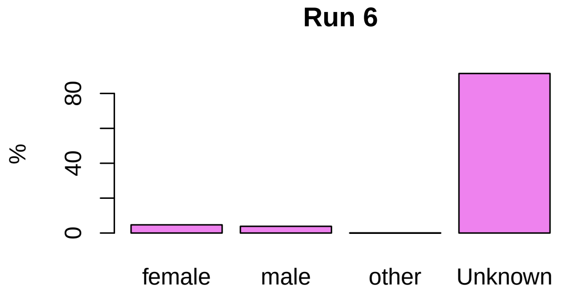
From the right plot it is clear that run 1 and 2 were the ones with the higher number of participants. Runs 3-7 seem to have a similar number of participants, with run 7 having the lowest number. In general after run 4, there is a steady decline in the number of participants. On the other hand, looking at the right plot, a different trend is observed. Run 1 has the higher completion rate, but run 2 seems to have the lowest even though it had a relatively high number of participants. Moreover for runs 5-7 even though there was a declining trend in enrollments there is an increasing trend for completion. So, event though run 7 is the one with the lowest enrollments, it has higher completion rate from runs 2,3,5 and 6. In addition run 4 seems to have a higher completion rate than expected, since it is much higher than that of run 5, even though the number of participants was similar. Since runs 4-7 have quite similar numbers of enrollments and are also the most recent ones, when compared to runs 1-3 it would be interesting to investigate what could drive this difference in completion. Two interesting areas to look at are gender and age. So the questions below arise:

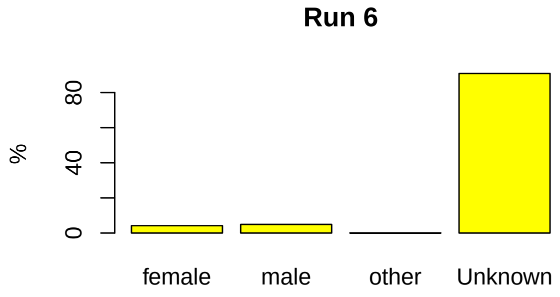
- Does gender affects initial engagement and completion of the course?
- Does age affects initial engagement and completion of the course?

To get an idea of how the gender differs across the hour runs the following bar plots are generated. These bar plots include data for all people that participated and not just the ones that completed the course:

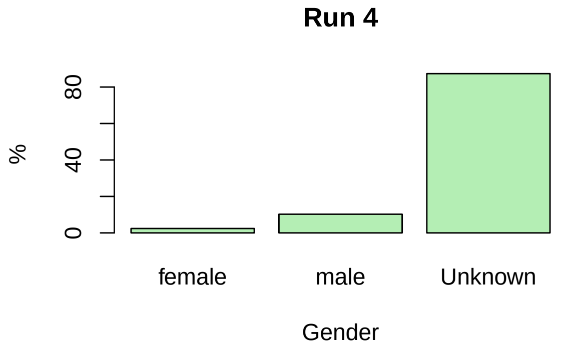


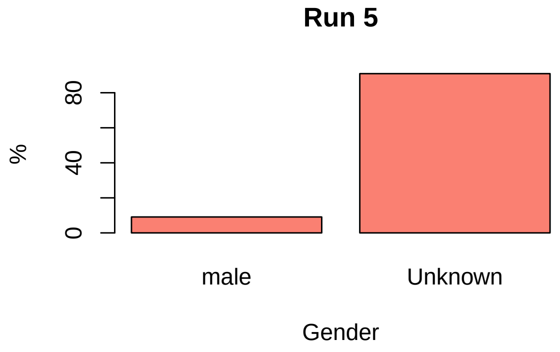


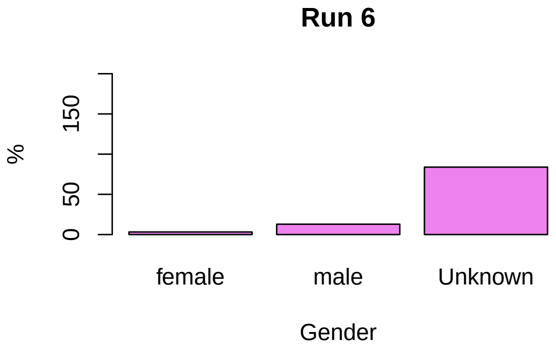


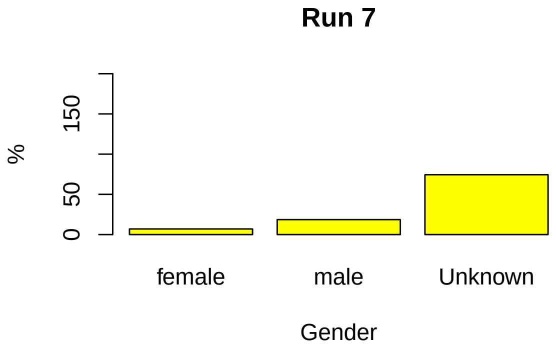


In general for all four runs, most people's gender is unknown. This is not ideal as the conclusions made are just from a sample of the population and not the whole population. So the conclusions may not be representative of the population. However, for run 4 which had the highest completion rate there seems to be more male participants compared to female. For the other runs there is approximately the same percentage numbers of males and females, with run 6 being the only run among these four runs, with more females.So there does not seem to be any significant difference in gender for the people who took part, in the different runs. However it would be interesting to also investigate whether there were differences in gender among the people that completed the course. The results are shown below:



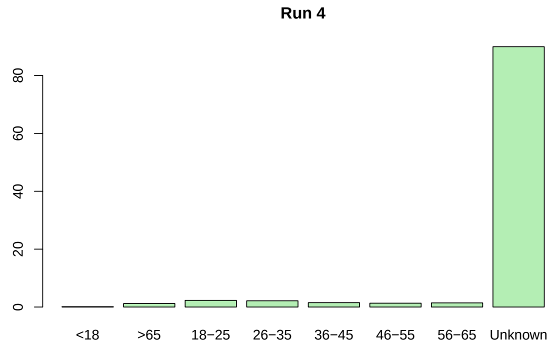


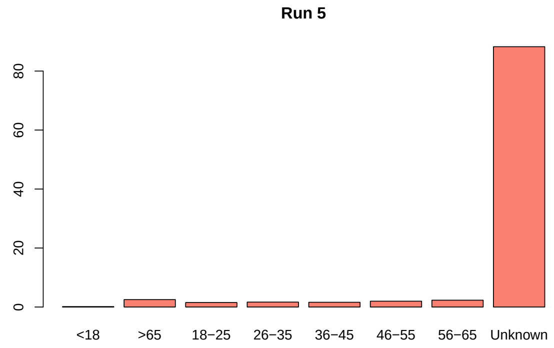


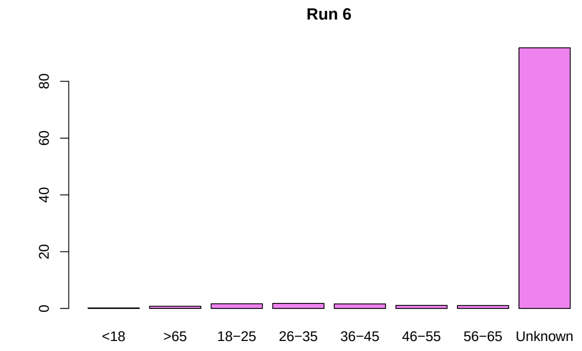


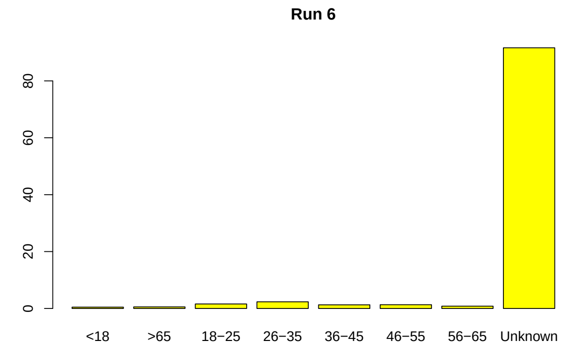
In run 4, 6 and 7 , of those who registered their gender there were more males compared to females that completed the course. However the data is not sufficient since more than 80% of the people that completed the course had not registered their gender. Especially for run 4 and 7 which have the highest completion rate among the 4, this pattern is more visible. In run 5 again this pattern is also present. However in this run only 22 people had completed the course, only from those people only 2 had registered their gender. In general there is some evidence that in the runs with higher completion there were more males who completed the course even though there was not such an obvious difference in the number of males and females that enrolled. This means that for some reason female participants tend to stop engaging with the course at some point, and maybe the target goup of the course developers if they are trying to imporove completion would be males. However since there are many unknown entries in all runs the fact gender may be affecting completion is just an assumption.

Other than gender, the age range would also be an interesting factor to look at. So one more question of whether there is a difference in age ranges between more and less successful runs arises. The plots below show the age ranges for the people who enrolled in each course:

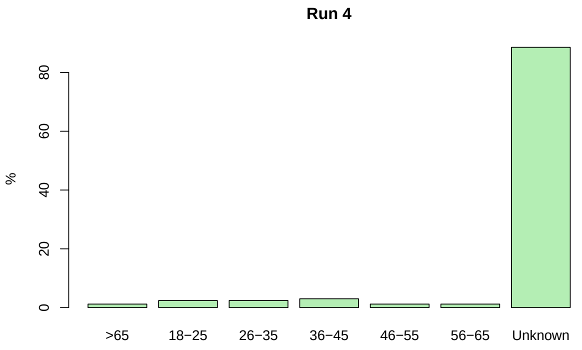


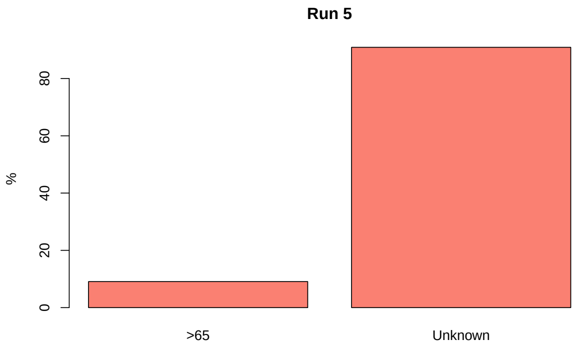


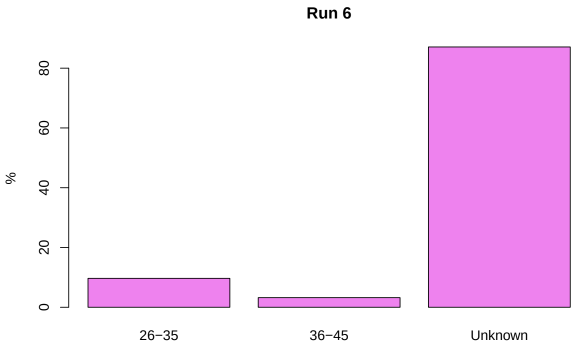


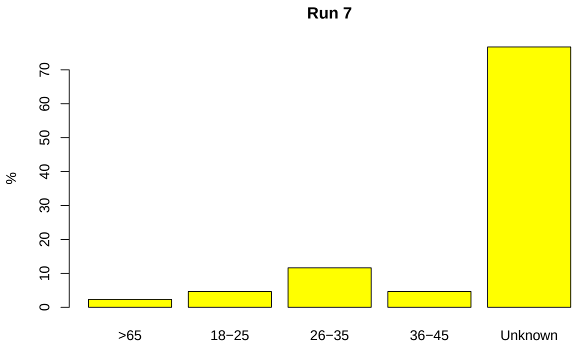


Again most age ranges of the people that enrolled are unknown so just like before the conclusions reached are just an assumption and may not be representative of the truth. Looking at the plots however, there is an interesting finding. It seems like run in 5, which is the run with the lowest completion rate, the majority of registered age ranges, lies in the 46-55, 56-65 and >65 age ranges, while in all the other runs there seem to be more people in the 18-25 and 26-35 age ranges. So there seems to be some evidence that younger people tend to engage more with the course, compared to older people. Moreover it would be interesting to look at the age ranges of people who completed the course in these runs. The results are summarised in the plots below:







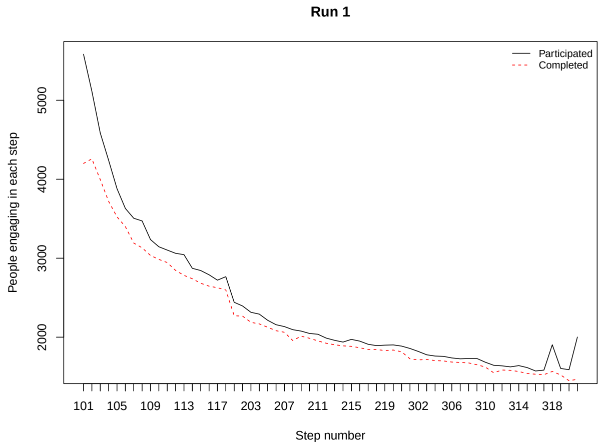


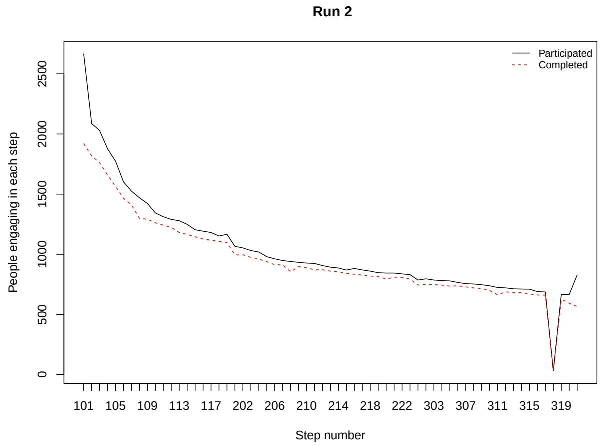
For both run 6 and 7 it seems that completion rate was the highest among people that are 26-35 years old. Moreover in all runs, except run 5, it seems that completion rate is also higher among people who are 36-45. People who are 18-25, don't seem to have very high completion rates. Moreover, what is very interesting is that in runs 6 and 7, among the people who completed the course and registered their age, all of them are below 45 years old. So this could be again an evidence that in general younger people between 26 and 45 years old seem to be the ones most interested in the course. As stated above, both in gender and age range data, there are many missing values that may make the analysis inaccurate. This may be the end of the first CRISP DM cycle since there are not any more data that could help investigate further whether age and gender affect engagement.

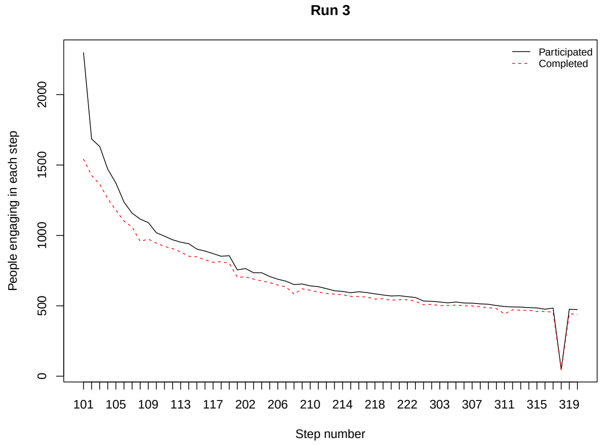
Since there are not any more insights to be extracted about age and gender, regarding engagement, the next thing to look at is step activity. The data sets for step activity include the step number (each section of the course), and there is an entry for each participant who engaged in this step. Looking at the step activity could help answer some questions that involve engagement. For example:

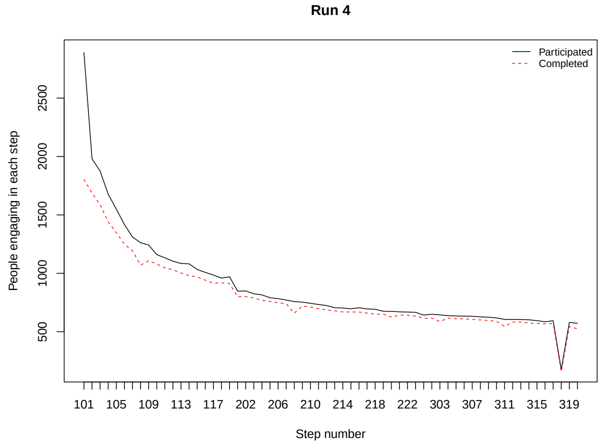
- How does engagement varies over the weeks?
- Are there any particular steps that participants find more/less interesting?
- If yes, is there any difference between these steps and the others?

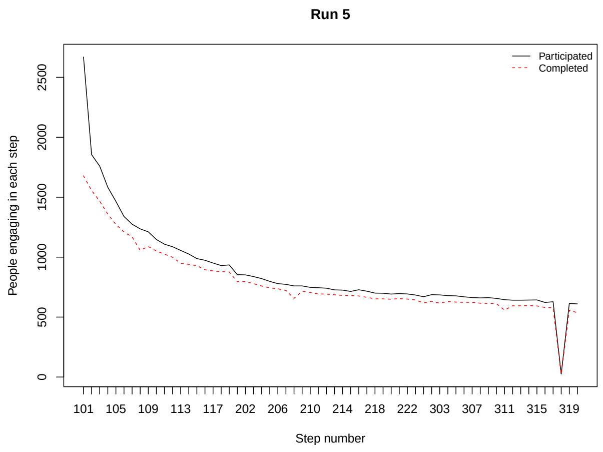
The plots below, show how engagement of all participants, and participants who completed the steps, varies among the 7 runs:

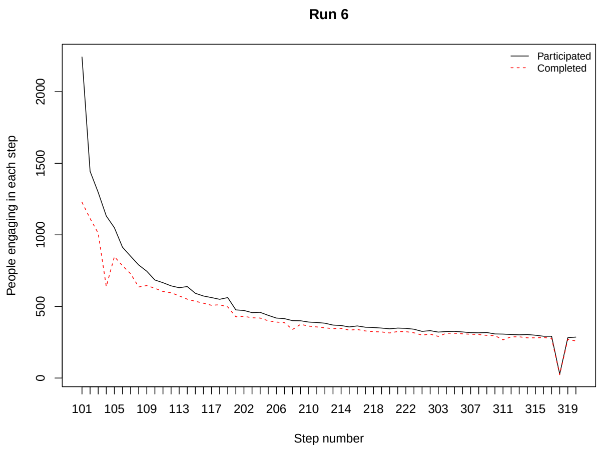


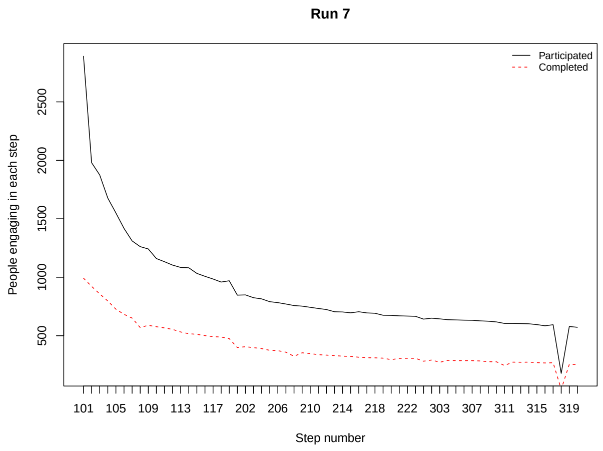







Run 5


Run 6


Run 7

In general we observe a declining trend in engagement, both for people who started the steps and the ones that completed them. For all runs except run 1, there seems to be a stabilization in engagement from the begging of week 2. In run 1, there is a more steep decline through out all of weeks. Moreover, for all runs there is quite a difference in engagement and completion of the first step, so it seems that there is a proportion of participants, that stop engaging after the first step. After the first step, all runs except run 7 there is not much difference between people who started the steps and people who finished them. So this means that in general participants who started the steps seemed to find them interesting throughout. However, there are some peaks and also some points were engagements between starting and finishing the step differs. It is important to identify these points as this could give an insight to the steps that participants engage more/less. The steps identified are:

- Step 1.8 in all runs (Quiz)
- Step 2.8 in all runs (Quiz)
- Step 3.11 in all runs (Quiz)
- Step 3.18 in all runs (Test)
- Step 3.21 in runs 1 & 2 (Glossary and references)

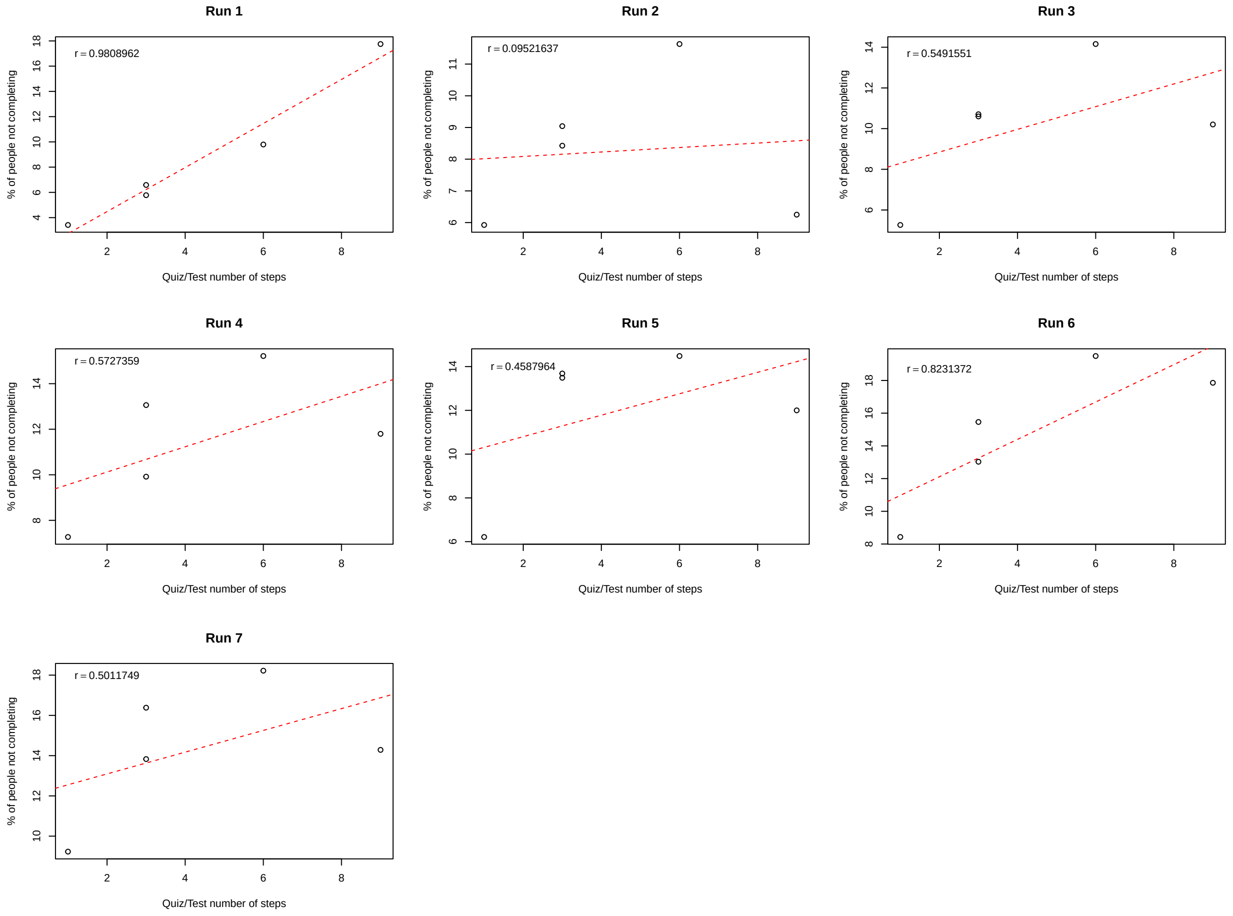
The results are very interesting. It seems like parts of the course where participants have to complete actions, rather than just watching a video, or reading an articles, have lower completion, and it seems like people are starting those section, but are not completing them. However looking closer at the course structure, it seems that step 2.20 is also a quiz. However engagement for this quiz (step) does not seem to deviate much from completion of the step. So why is this happening? Looking more closely at the quizzes and the test, the following are found:

Quiz_and_test	step_number
1.80	6
2.80	3
2.20	1
3.11	3
3.18	9

So it is observed that quiz number 2.20 only has one question, while the test has 9 questions and the other quizzes have 3 and 6 question. So the following question arises:

- Is it posible that engagement on the quizzes and the test is affected by the number of questions?

In order to investigate that the following plot is produced:

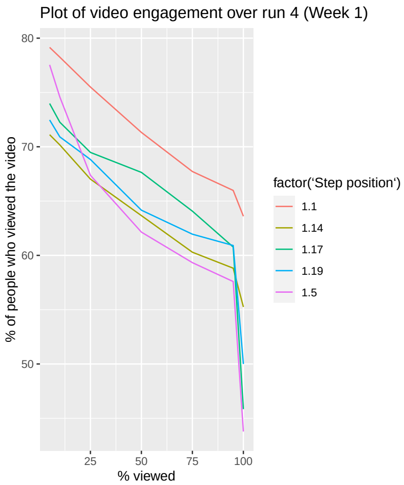


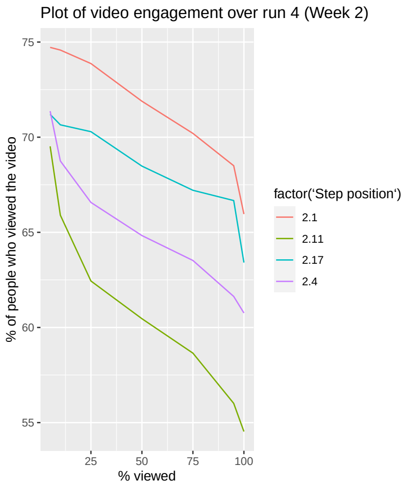
Looking at the plots and the correlation coefficients for each run, it is clear that for all runs except run 2 there is a moderate to strong positive correlation between the percentage of people not completing the quizzes/test and the number of quiz/test question. In this case it seems that run 2 is an outlier. So this means that in general the more questions a quiz or a test contains the more probable it is that the participant will not complete it. So, the course developers should consider creating shorter quizzes in order to boost engagement. Looking step activity for investigating engagement generated some very interesting results that can help course developers understand better what affects people's engagement. However a very interesting part of the data sets are the video statistics that are provided for runs 3-7.

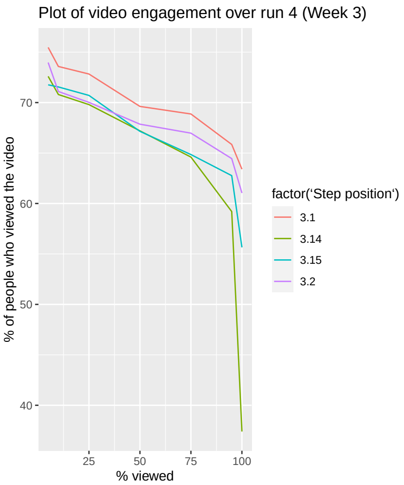
The video statistics data sets can help answer the questions below:

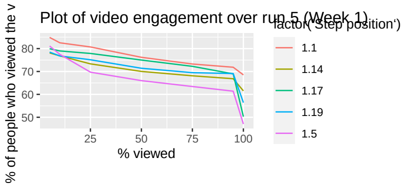
- Are there any particular videos that participants seem to engage more/less when compared to others?
- Is this pattern the same in every run?
- If there are videos with higher/lower engagement, what factor(s) drive this higher/lower popularity?

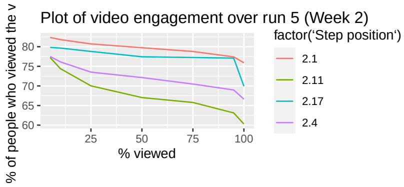
For the analysis below data from runs 4-7 are used. Run 3 is excluded because it exhibits simialar patters to the other 4. Moreover, runs 4-7 are most recent when compared to run 3. The plots below show the percentage of people who watched up to a specific percentage of the video duration. The videos are separated by week, in order to make the plots more readable:

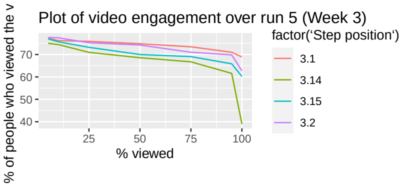




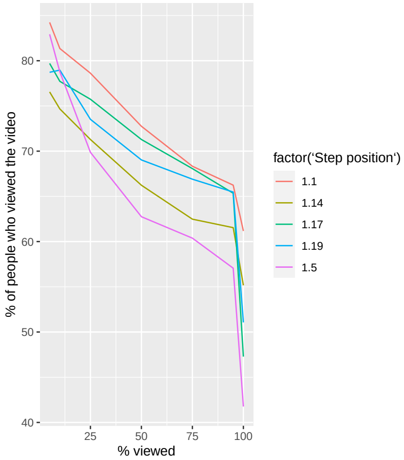




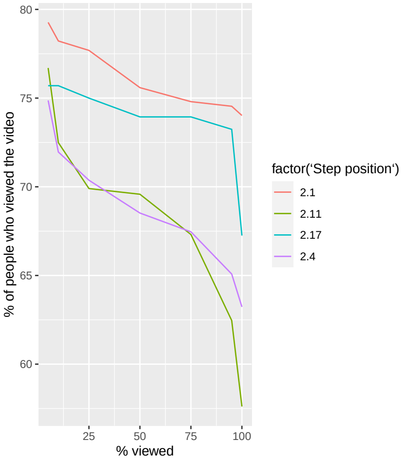




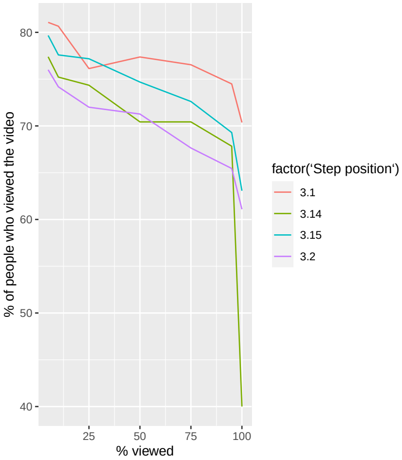
Plot of video engagement over run 6 (Week 1)



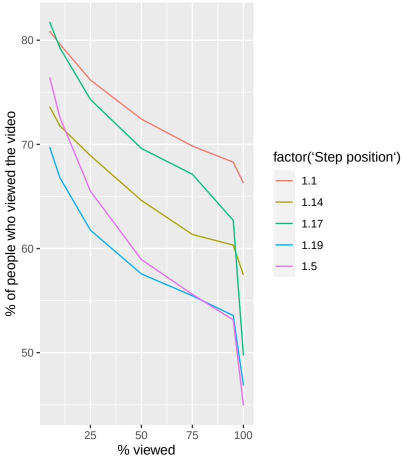
Plot of video engagement over run 6 (Week 2)



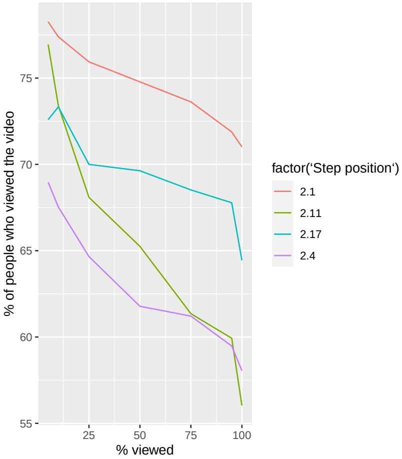
Plot of video engagement over run 6 (Week 3)



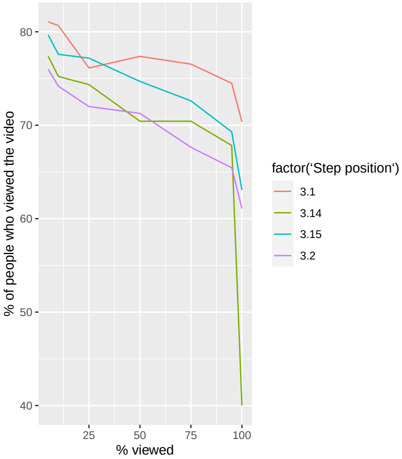
Plot of video engagement over run 7 (Week 1)



Plot of video engagement over run 7 (Week 2)



Plot of video engagement over run 6 (Week 3)


% of people who viewed the video
% viewed
factor(‘Step position‘)