report eda

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:

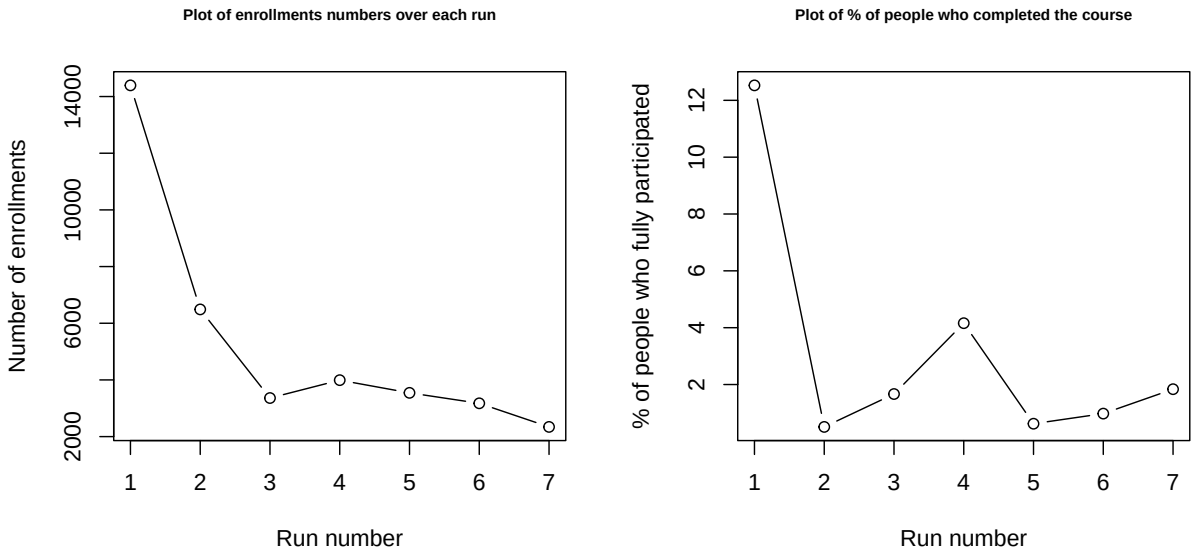

Figure 1: Plots of number of enrollments (left) and percentage of participants completing the course (right) for all 7 runs

From the figure 1, 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, from figure 2, 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.

From figure 3, 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

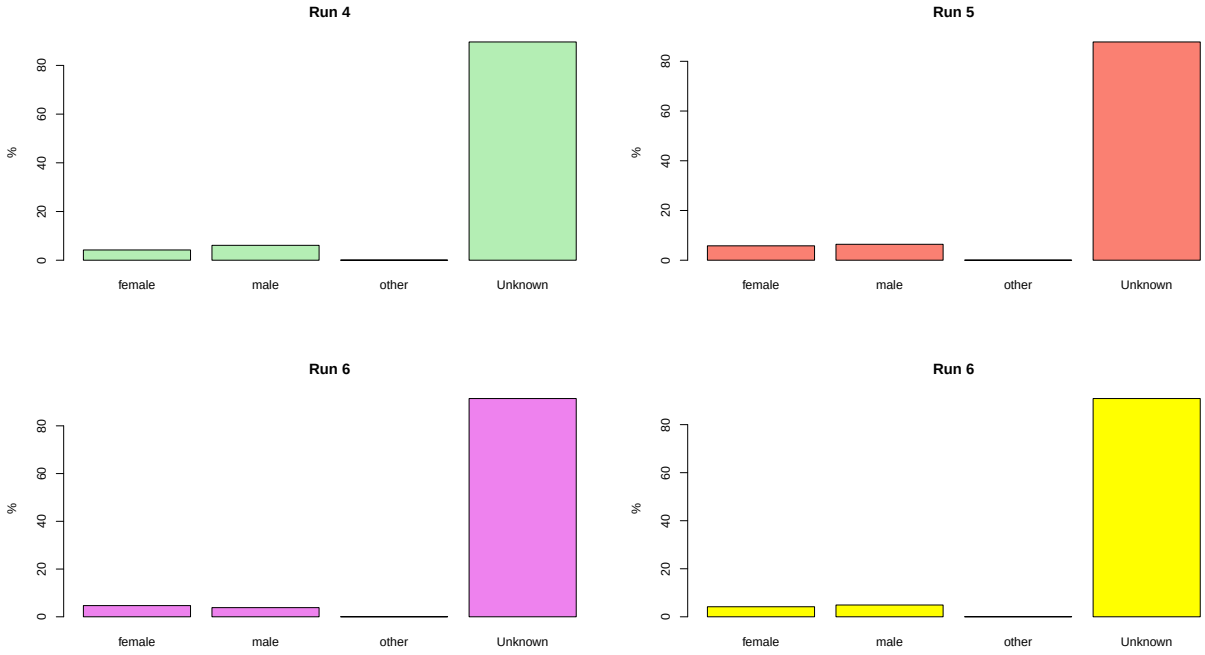

Figure 2: Gender statistics for participants enrolled in runs 4-7

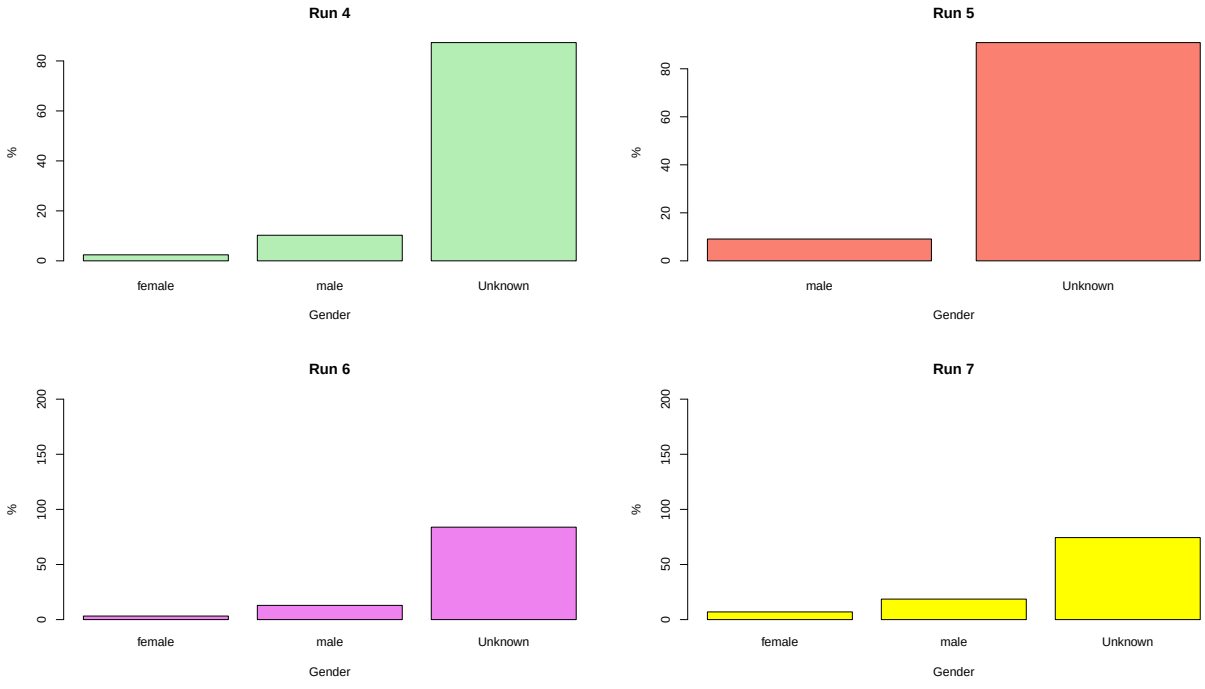

Figure 3: Gender statistics for participants that completed in runs 4-7

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:

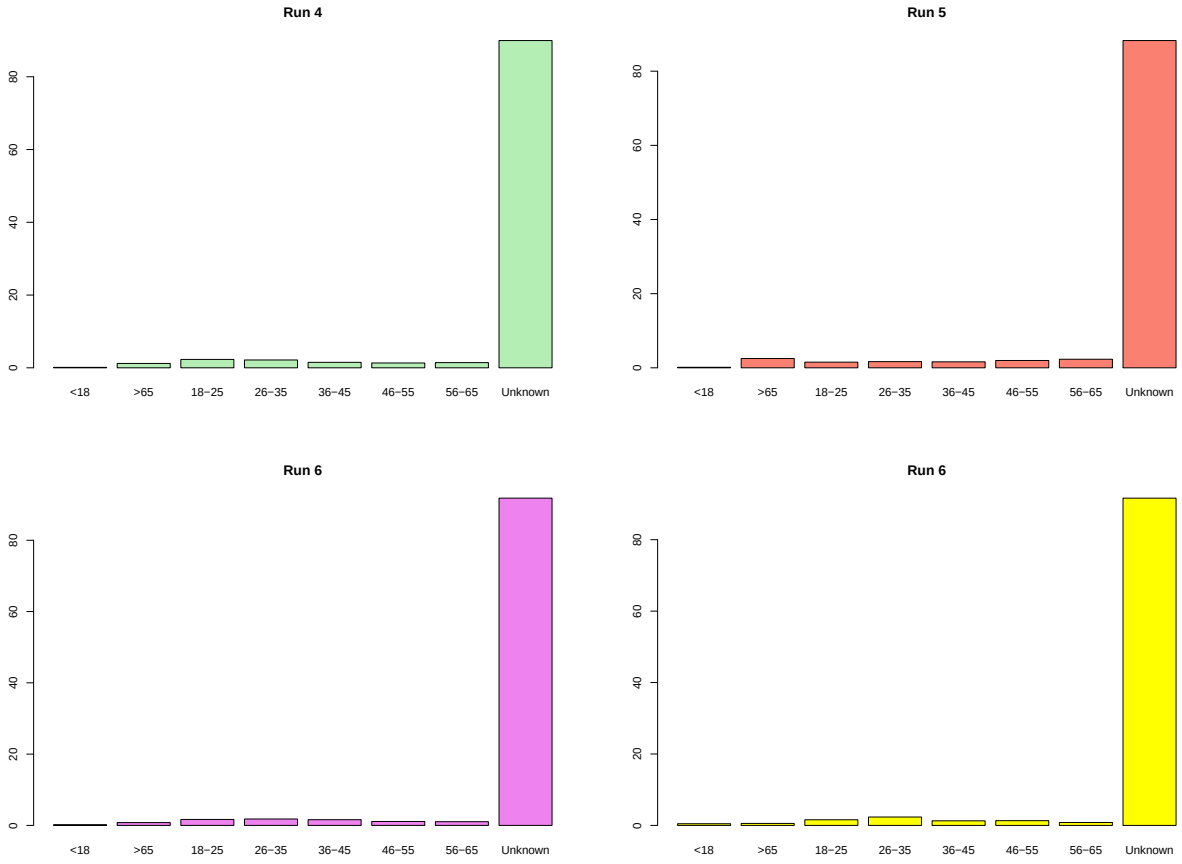

Figure 4: Age statistics for participants enrolled in runs 4-7

From figure 4, 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.

From figure 5, 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

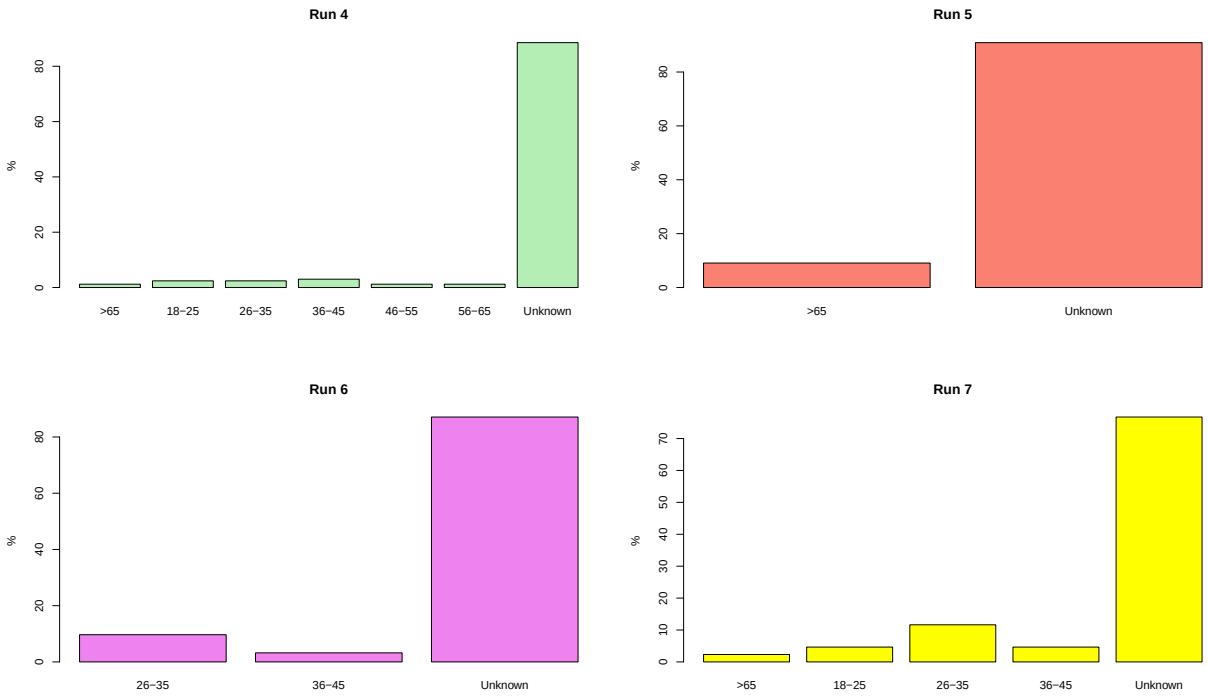

Figure 5: Age statistics for participants enrolled in runs 4-7

who are 36-45. People who are 18-25, don't seem to have very high completion rates, so it could be the case that the course developers may want to target more people on the 26 - 45 age range in order to boost completion. 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.

CRISP DM cycle 2

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?

Figure 6, shows how engagement of all participants, and participants who completed the steps, varies among the 7 runs.

In general, looking at figure 6, a declining trend in engagement is observed, 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.

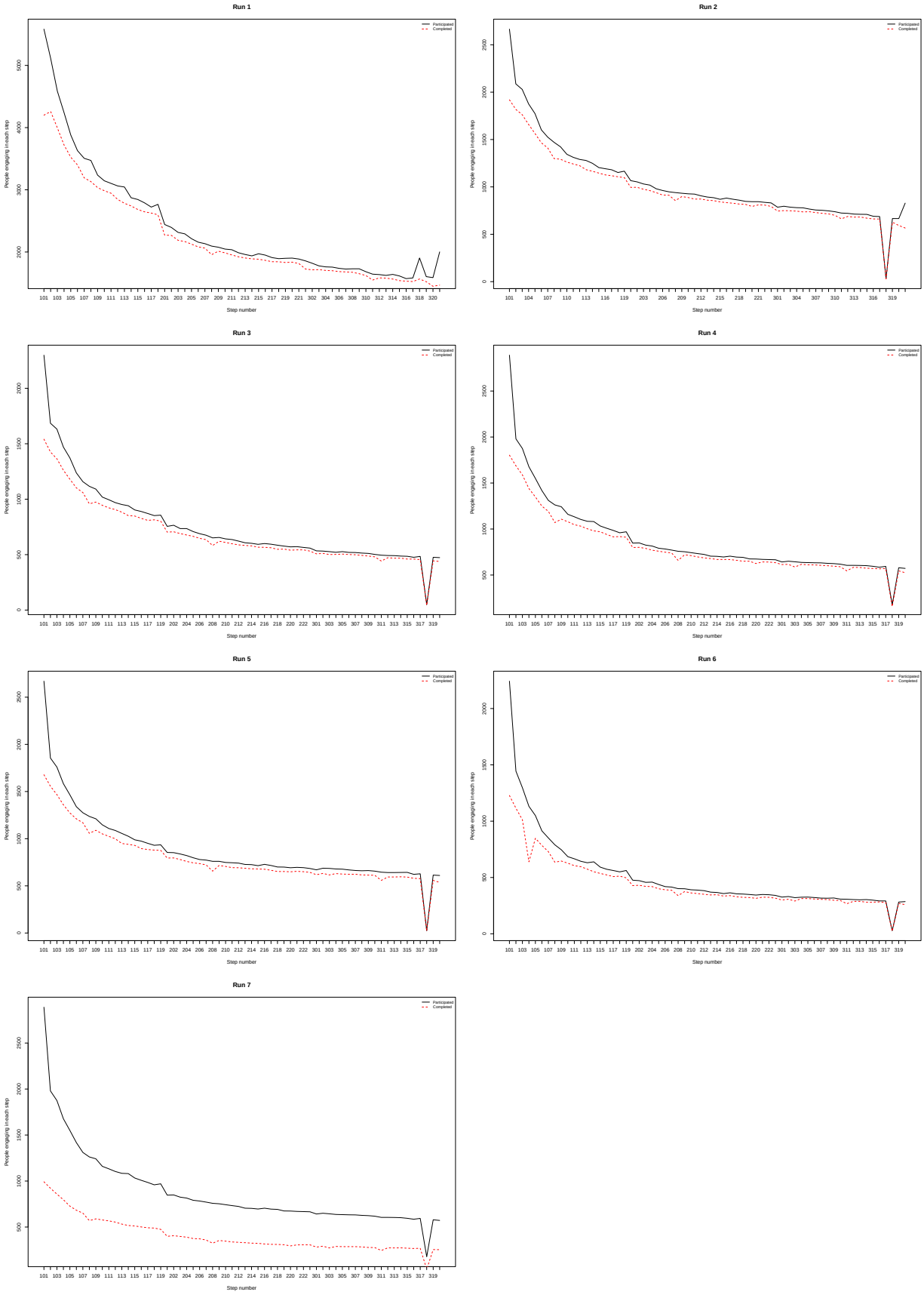

Figure 6: Plots of step number against number of people who engaged in each step (solid line) and people who engaged and finished the step (dashed line) for all 7 runs

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?

Table 1: Table showing the quiz number and the number of question it contains

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

From the table 1 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 possible that engagement on the quizzes and the test is affected by the number of questions? In order to investigate that the figure 7 is produced:

Looking at figure 7 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.

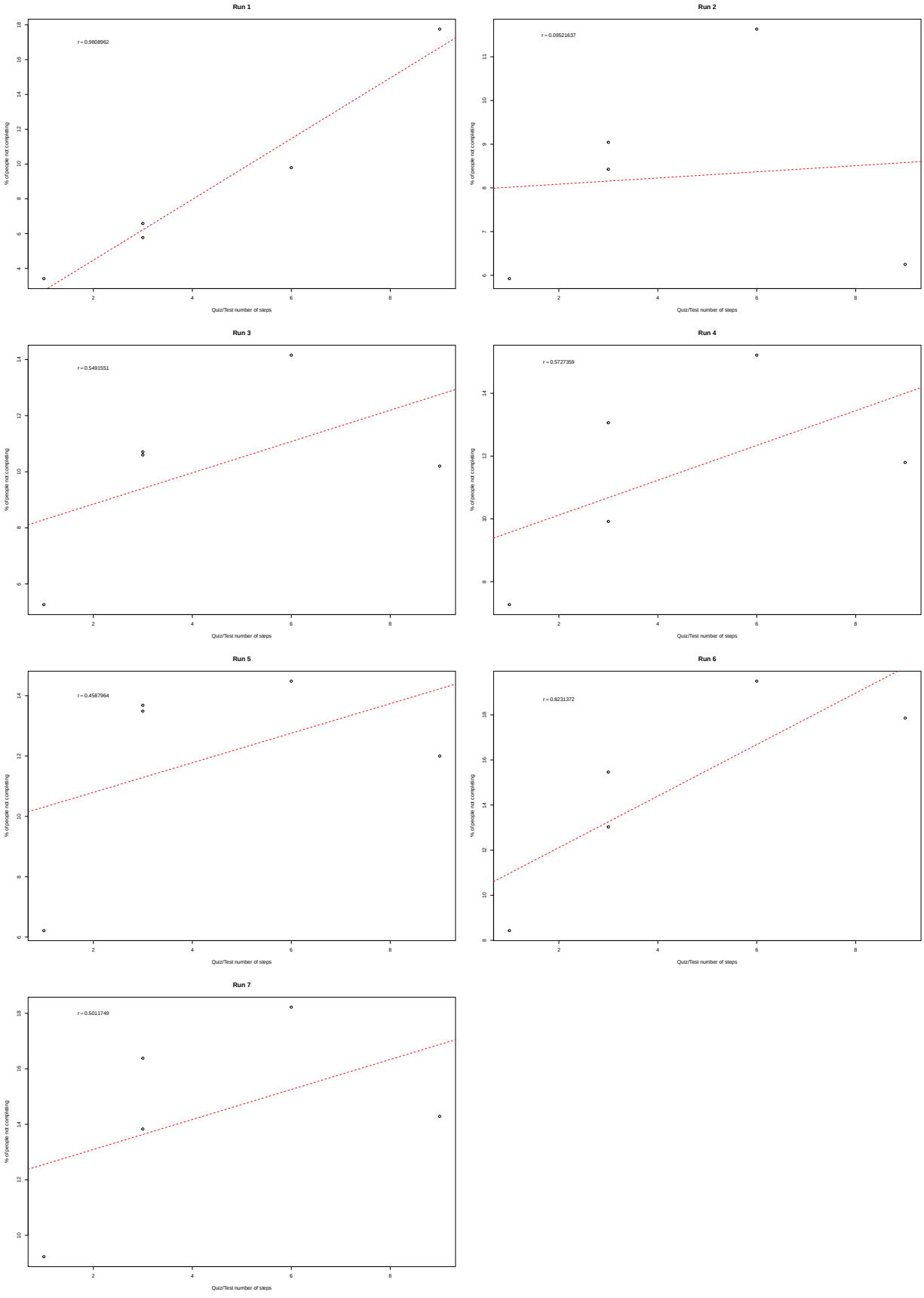

Figure 7: Plots showing the relationship of quiz/test length against engagement over all 7 runs

CRISP DM cycle 3

A third 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:

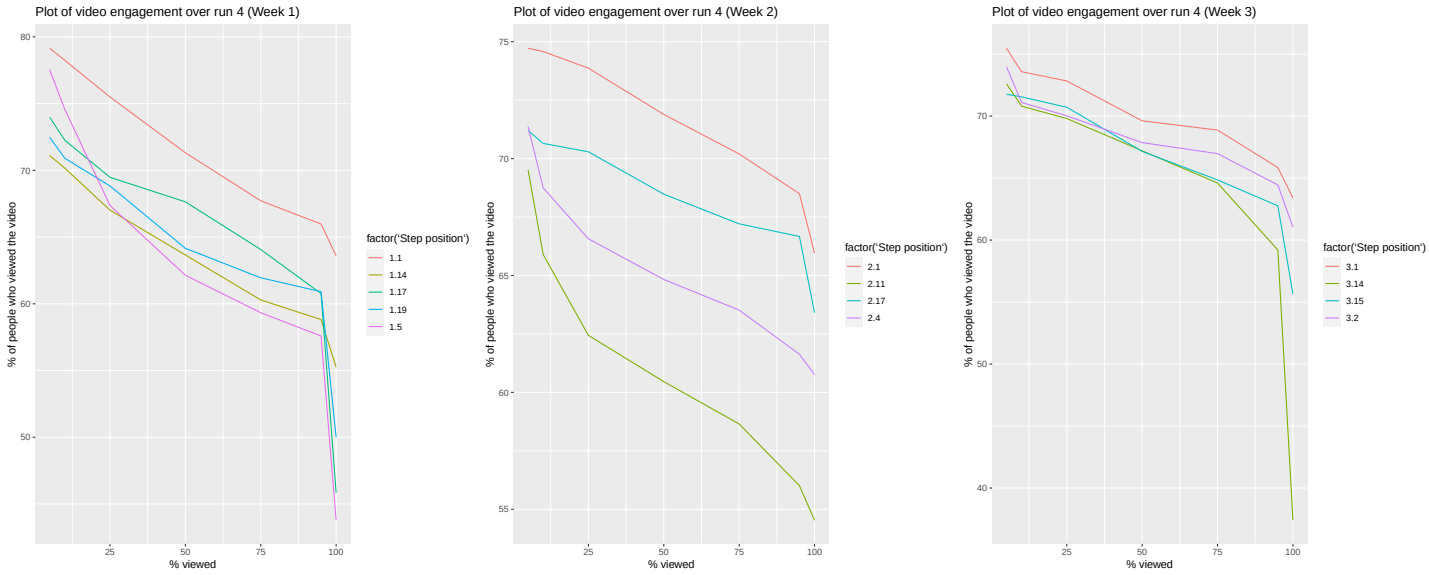

Figure 8: Percentage of people engaging with each video in run 4

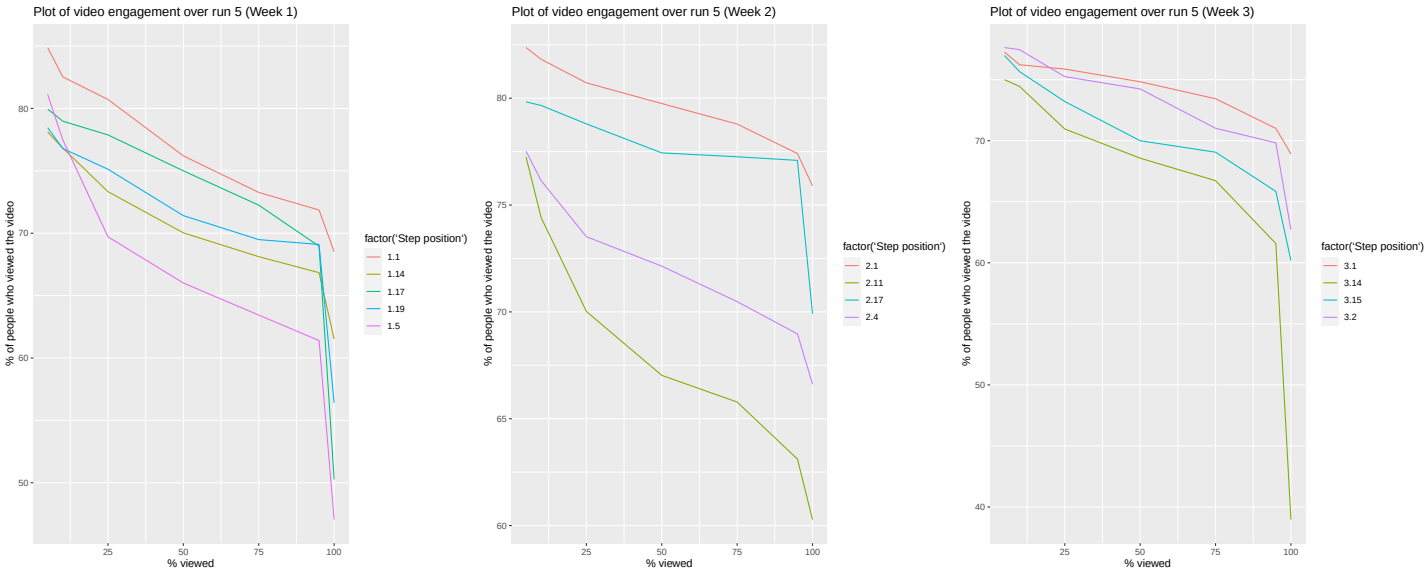

Figure 9: Percentage of people engaging with each video in run 5

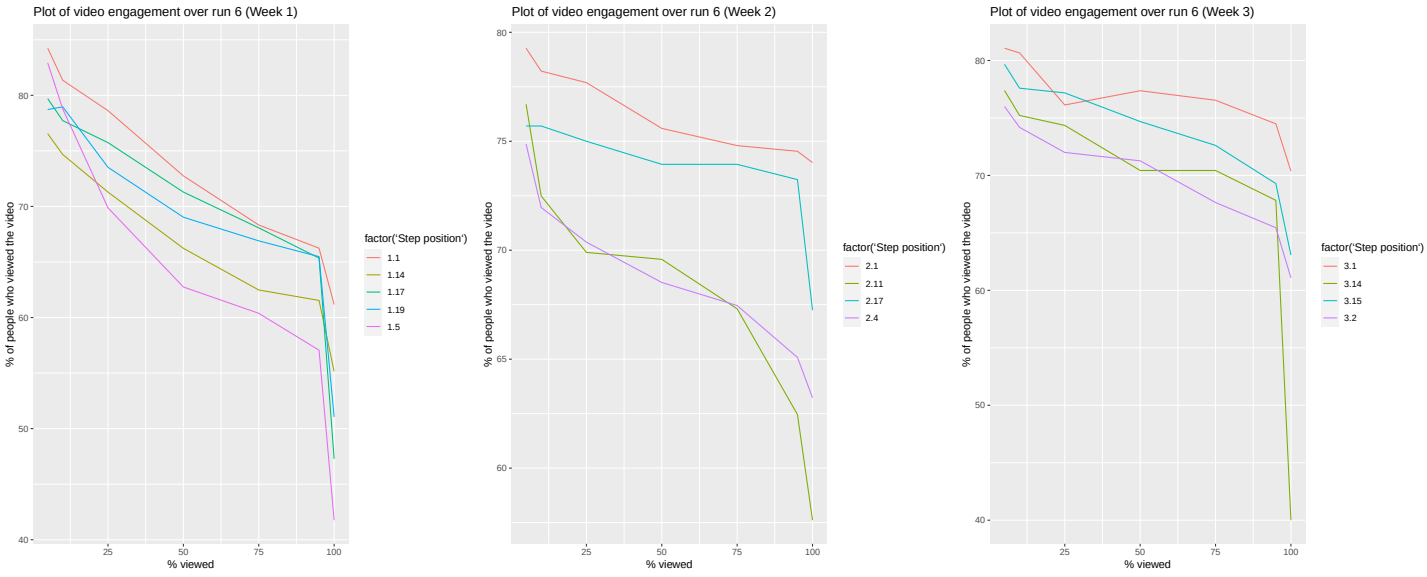

Figure 10: Percentage of people engaging with each video in run 6

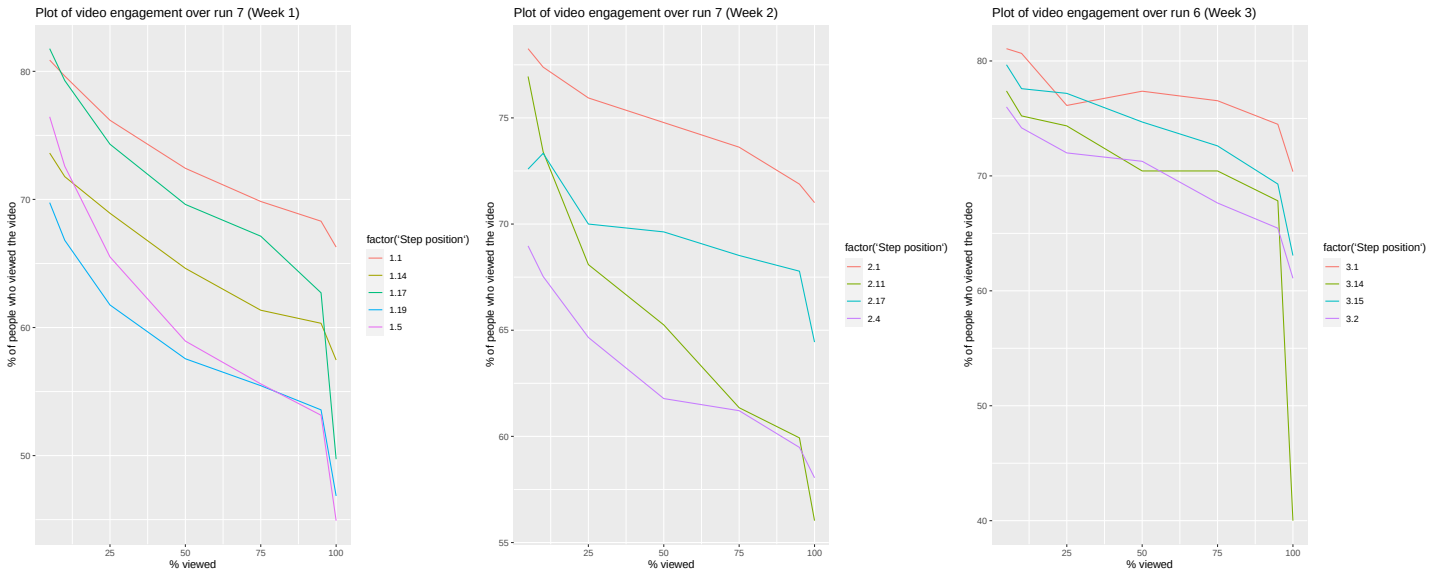

Figure 11: Percentage of people engaging with each video in run 7

Looking at figures 8-11, the first thing to notice, is that for all runs and for all weeks, there is a very steep decline in video engagement, on the last 5% of the the videos. So in general most people tend to watch up to 95% of the videos, and then leave the step. If these last seconds of the videos, contain just references, titles and trademarks etc, it is normal for people to leave, especially if this is the same in all videos. However if in this last 5% important information is being shared, then it would be better if the important details of the videos, were mentioned towards the beginning or the middle of the videos. So for more accurate conclusions it would be better to look at the engagement up to 95% of the videos, for explaining the graphs as well as for further analysis.

In general, looking at all runs, for week 1 there is the most decline in engagement for video 1.5 For week 2 there is a steep decline in engagement for videos 2.4 and 2.11 and for week 3 in general there is quite a constant engagement in the videos, with video 3.14 and 3.2 being the videos where more people left, compared to the other two videos of the week. However from the previous plots it was found that the engagement in steps during week 3 was lower and more constant, so probably the people that have stayed through week 3 are interested more in the course, tend to watch more parts of the videos. Even though there similar pattern along the 4 runs, there are also some differences. So the question arises:

- Why do some videos have higher and more constant engagement?
- Could duration play a role in that?

Table 2 below shows the duration of the videos.

Table 2: Table showing the video number and it's corresponding duration

step_position	duration_in_seconds
1.10	99
1.14	362
1.17	241
1.19	348
1.50	281
2.10	37
2.11	312
2.17	92
2.40	426
3.10	59
3.14	313
3.15	227
3.20	206

From table 2, it seems that some videos have longer duration. In order to see if this affects engagement, the correlation between duration and engagement up to 95% of the videos was computed and figure 12 was produced:

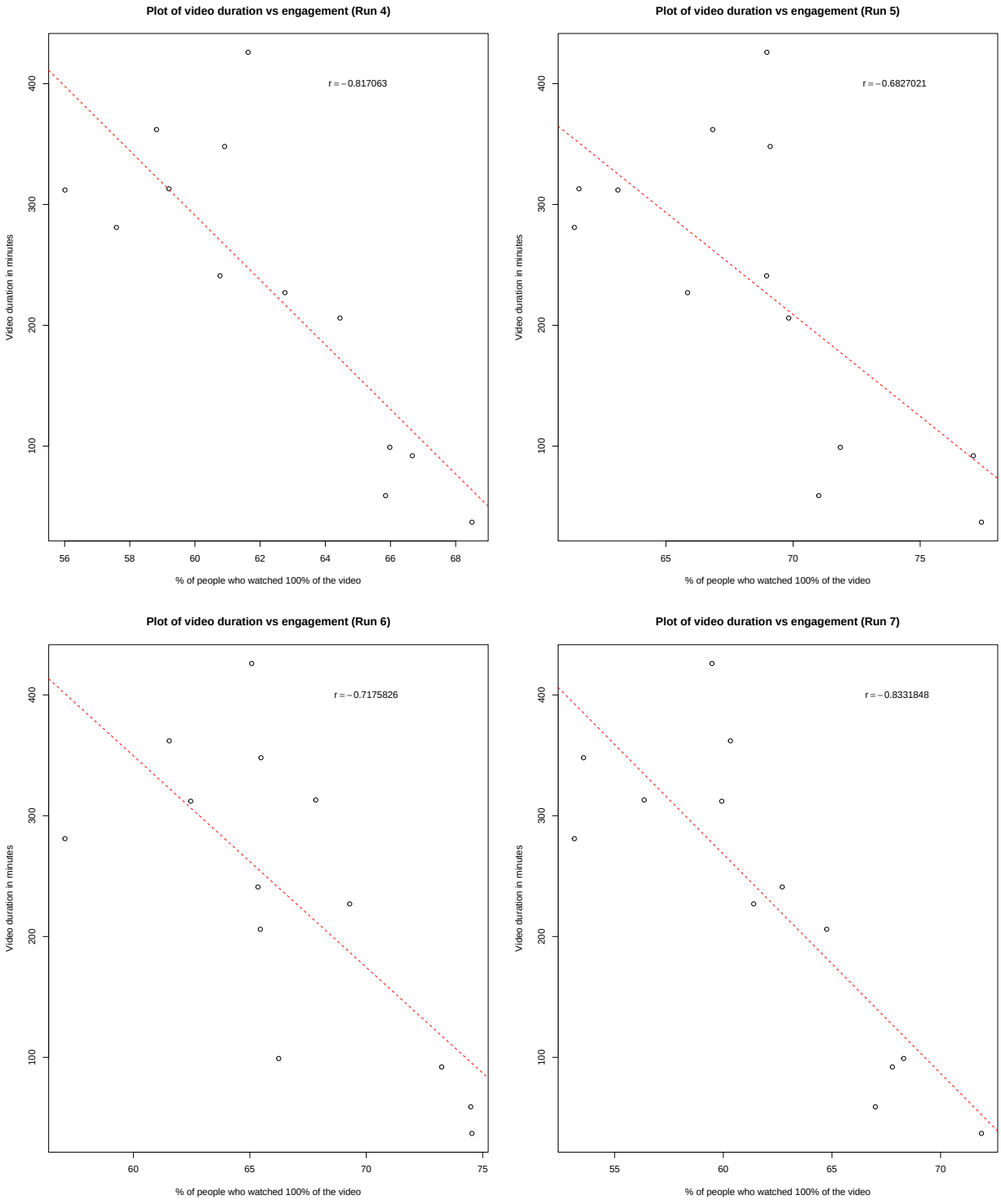

Figure 12: Plots showing the relationship between video duration and engagement over all 7 runs

Looking at figure 12 and the coefficients of correlations, it is clear that runs 4,6 and 7 have a strong negative correlation between duration and engagement, while run 5 has a moderate negative correlation. This means that in general, for all runs, the longer the duration of the video, the lower the engagement. Before it was found that in run 5, there were more older people that signed up compared to the other runs, so this could mean that older people do not mind longer videos, while young ones prefer shorter ones. However this is just an assumption, since the data for the age ranges were just a sample of the population.

In order to get a better idea of which videos are more popular the figure 13 was generated. This figure shows the percentage of people that started watching a video, but left before watching up to 95% of it.

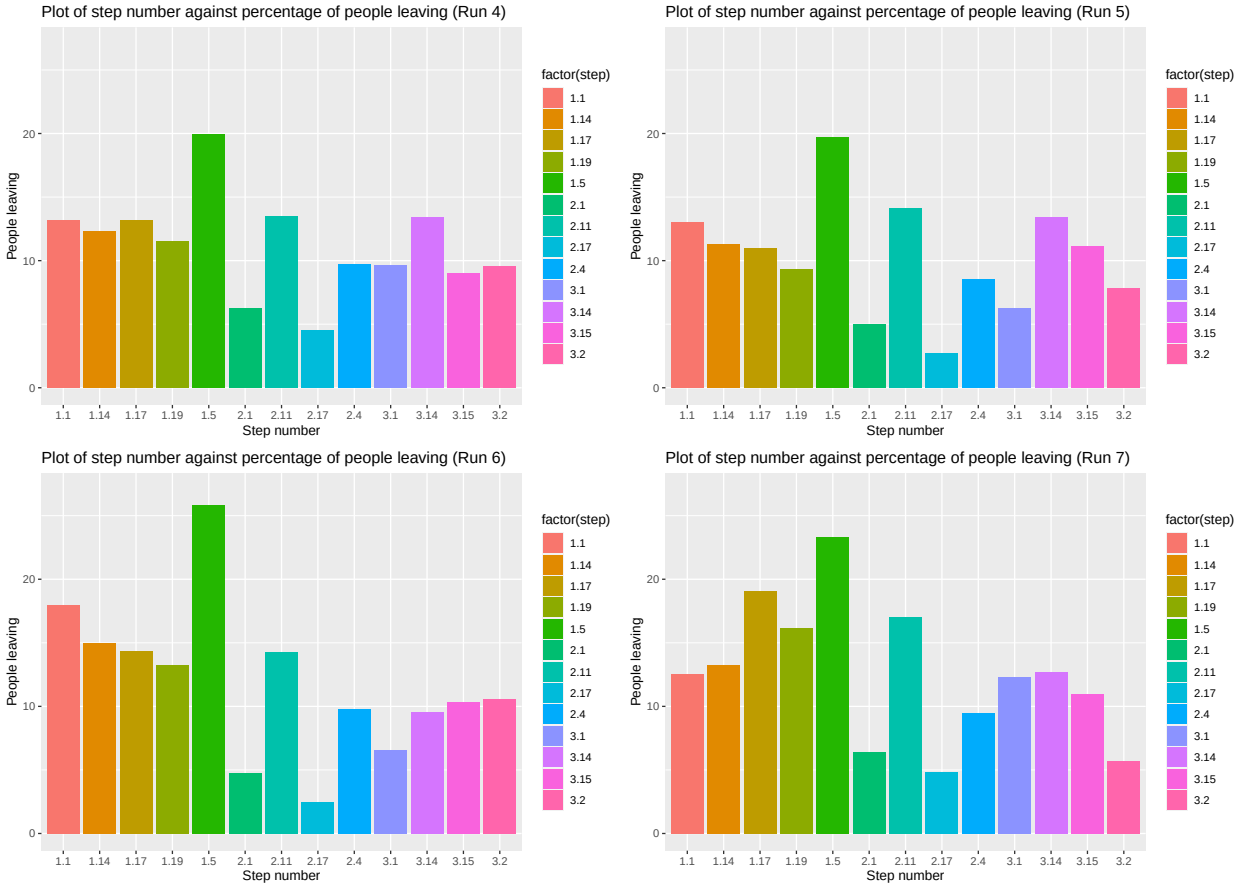

Figure 13: Barplot of percentage of people that droped out of each video for runs 4-7

The first observation, from figure 14, that is common for all runs is that video 1.5, which has duration of 281 seconds, has the largest dropout rate among all videos in week 1 and all videos in general. Then for week 2 video 2.11, which has duration of 312 seconds, has the largest dropout and for week 3 except for run 6, video 3.14, which has duration of 313 seconds, is the one with the highest dropout. As expected these are videos are ones of longer duration. Now the videos with the lowest dropouts among all weeks for most runs are videos 2.1, which lasts 37 seconds, and 2.17, which lasts 92 seconds.

Also what is quite interesting is that video 2.4 has the longest duration, but it is not the one with the highest dropout, in any of the runs, and compared to other videos it has quite a better dropout rate. So the theme of the video seems to be something that people are interested in and the course developers may want to introduce more related material to the course. Moreover video number 1.10, which lasts 99 seconds seems to have higher dropout rate in most runs except run 7, when compared to video number 1.14 which last 362 seconds. So again the content of the video, may be an area that people who participate in this course are more interested in. Finally, if the videos have different people delivering them, this could also be

a factor affecting engagement. So for example one of the lectures may be preferred by the participants and that is why some videos may be longer but have higher engagement than expected. However this is just an assumption, that may not apply in the course.

Conclusions

For the analysis above three cycles of the CRISP DM methodology were used. The first cycle investigates whether differences observed between participation and completion are connected with the gender or age range of the participants. This first cycle analyses primarily runs 4-7, because they have quite similar number of participants but different completion rates. What was found was that even though the number of males and females who participate in the course are quite similar, there is a difference in the gender of the participants that actually complete the course. It seems like more males tend to complete the course when compared to females. Moreover looking at the age range, it was found that one of the runs with the lowest participation had more participants that were in the age range of 46-55, 56-65 and >65, when compared to most successful runs, in which more younger people in the 18-25 and 26-35 age ranges had participated. So this is an indication that female participants stop engaging at some point of the course.However when looking at the people who completed the course, it seems that in general the higher completion percentage was in people that belong in the 26-35 and 36-45 age ranges. From the two findings above what can be suggested to the course developers, if their aim is to achieve higher completion rates, would be to target more males and also people in the 26-45 age range. However when making decisions based on these conclusions, care should be taken since in general most participants had not registered their age or gender, so the analysis was made only on a sample of the participants and may not be very representative of the participants population. So what the course developers could do in order to get some more accurate conclusions, would be to make the age range and gender field compulsory for the participants. Moreover there are other fields like employment status and higher education limit, that the participants can fill, but again is not compulsory. Making these field compulsory on registration would give an even better idea of the target group of the course.

The second cycle, investigated participants engagement with the steps of the course. What was found was that in general there is a steeper decline in engagement during the first week, and then for the second and third week there is a stabilization. Moreover, in general people who engage with the steps seem to be finishing them. However it was found that there were some specific steps were people engage with them but do not finish them. These steps were identified to be the quizzes and the test. So it seems that participants do not enjoy that much these kind of activities. In addition, a very interesting founding was that, participants seem to engage more with shorter quizzes (that have less number of questions), rather than longer quizzes. So course developers could look to make these activities shorter, or replace them with other activities that participants seem to enjoy more.

The third cycle, looked at the video statistics for runs 4-7 and tried to investigate which videos are more popular and why. The main result of this analysis was that in general participants prefer shorter videos. Even though duration seemed to directly affect engagement, it was found that some videos which were quite long, had a better engagement than shorter videos. This was consistent in the four runs, so participants may be interested more in the content of these videos, and course developers could look to include more related material, and maybe replace some videos, that maybe short but do not have a big engagement. Moreover,if the videos have different people delivering them, that could be a reason why some longer videos may be more popular. Finally, combing the analysis of cycle 1, with that of cycle 3, there was an indication that younger people prefer shorter videos. However this was just an assumption, since the results from the first cycle were based on a sample of the population. It was also found that most participants do not engage with the last 5% of the videos, which could lead to participants missing important information, if these are mentioned towards the end of the video. So the course developers should try to include interesting and important information towards the begging or the middle of the videos.

To sum up, all three cycles showed some interesting results that can explain which factors affect engagement. Moreover, some suggestions for the course developers were proposed, in order to boost engagement in future runs, or just understand why some things did not work as expected.