CRISP-DM MOOC Investigation: An Analysis of the Video Statistics in the Cyber Security FutureLearn course

Annie Sames

03/12/2021

FutureLearn is an online education platform that offers a variety of courses. These Massive Open Online Courses (MOOCs) can be accessed from any device, and the course offers a variety of resources, including online lectures, journal articles, and quizzes, that allow students from around the world to begin studying a range of subjects. In order to optimise learning and better the knowledge that is obtained from these courses, constant analysis of the data is necassary. One area which may be of particular interest to researchers and the company are the video statistics, and this may include information on how many peoplem enrolled on the course watch the videos, how long they watch them for and the devices that are most commonly used to watch these. Such analysis helps companies such as FutureLearn are able to improve their courses. In the case of video statistics, this enables them to increase the number of viewers and make the videos more efficient.

CRISP-DM methodology (Chapman et al, 2000) will be used for this investigation, and multiple rounds of this will be conducted. This is broken down into six stages:

- Business understanding: to understand the business objectives and rules of the company
- Data understanding: to collect and understand the data in question
- Data preparation: to select the relevant data and prepare it for analysis
- Modelling: build the relevant models for the data
- Evaluation: review the process carried out and evaluate whether the results have achieved the intended goal
- Deployment: implementing the findings with the business objectives

For the purposes of this report, this investigation will focus on the business understanding, data understanding, data preparation and evaluation stages.

Business Understanding

The first phase of CRISP-DM is business understanding. In the context of this research, this is aimed at defining the objectives of FutureLearn courses and understanding what the aims of FutureLearn as a company are. MOOCs are designed to be accessible to everyone and FutureLearn say that they should be a social experience that people enjoy. The primary aim of these courses is to deliver the material in a user-friendly way that people enjoy, as well as in a way that maintains a high quality of learning.

One way in which FutureLearn courses teach material is through the use of videos and making these videos as high quality and efficient as possible is of utmost importance to FutureLearn. This includes maximising the number of students watching the videos and the length of time that they engage with them for. Data from FutureLearn courses has shown that a large proportion of people do not watch the videos, so it would be beneficial for FutureLearn to change this as this is likely to impact the quality of the learning that the students receive. The aim for this project will be to advise FutureLearn on a length of time that their videos should be in order to increase the percentage of people enrolled on the course watching them.

The data available for this investigation has been collected from the FutureLearn course titled 'Cyber Security: Safety at Home, Online, in Life', set up by Newcastle University. This is a three-week online course, which has been designed to require three hours of study per week. It is aimed at people who have some knowledge of cyber security and IT knowledge and an interest in the application cyber security to the state. On this specific course, the highest percentage of people watching 5% of the duration of a video is only around 80%, meaning that at least 20% of people do not watch even the first 5% of all the videos in a course. The highest percentage of people on the course at any given stage watching the full videos is also only around 70%, showing that at no point in the course does everyone watch the videos for the full duration. This project will investigate the videos from this course, with the intention of improving the percentage of people watching the videos.

Data Mining Goals:

The goal of this investigation is to identify on average, how much of each video is watched, and from this, investigate whether there is an amount of time where people stop watching videos. From this, a suggestion for a length of time that videos should be can be given to FutureLearn, in order to maximise the amount of people watching them. Alternatively, if videos are not able to be cut to this length, it can be used as a guide as to where the most important information should be shown in the video to maximise the amount of people hearing that information. This should be beneficial to the students enrolled on the course as they are more likely to learn the important information, and it will also be better for FutureLearn as students' overall outlook on the course should be better and this should in turn encourage more people to enroll on the course in the future.

It is important to investigate whether there is a point at which most people stop watching videos because the data from cyber security courses shows that only a small number of people watch each video 100% of the way through. Also, a brief look at the data suggested that as durations increased, viewing percentages significantly decreased. Looking at this from a psychological perspective, it might be hypothesised that this may be because the attention span of humans for online videos can sometimes be as little as eight seconds, meaning that most of the videos in FutureLearn courses are significantly longer than this. Ultimately, reducing the length of the videos to a length more close to the average attention span of humans will be the goal; however, this must be done within reason, to allow for a sufficient amount of material to be taught in each video.

Project Plan:

To complete this project and achieve the goals set, several steps will be carried out. Firstly, the relevant data files will be selected from the large set of datafiles from the Cyber Security FutureLearn course supplied, and from this the relevant columns of data will be extracted. Next, the relationship between video duration and the percentage of people watching the videos will be identified, and from this, the point at which people stop watching videos should be determined in order to make a suggestion for the length of time that videos should be. This should achieve the overall business aim of increasing the amount of people watching the videos as the videos can be made shorter to match the optimal length of time, which will increase the number of people watching them. Multiple iterations of the CRISP-DM methodology may need to be carried out in order to achieve the goal. The outcome of each stage will be evaluated and compared against the business goals to assess whether or not a further cycle of CRISP-DM is required.

Data Understanding

The data used in this project was collected from seven runs of the Cyber Security FutureLearn course running from September 2016 through to September 2018. Eight data files were available for each of the seven runs and these included information on student demographics, survey responses, video statistics, question responses, step activity and weekly sentiment video statistics. The course overviews were also available for each of the runs, and these were used as a starting point for this project in order to become familiar with the course structures and gain an understanding of how a FutureLearn course works. Additionally, this helped identify the differences between runs, which may have altered the focus of the investigation.

The video statistics files, which contained information on the video step numbers and durations (in seconds) for each of the videos, were the focus of this investigation. For each of the thirteen videos, the percentage of people watching the video were given at seven different percentages of the total duration (5%, 10%, 25%, 50%, 75%, 95% and 100% of the total time). Also included in the video statistics files was information about how many people used transcripts and captions, downloaded the videos, viewed the videos in HD, and the devices they used. They included the percentages of people who watched videos across all seven continents and the total views for each video.

As an initial data exploration exercise, the correlations between the video durations and the percentage of people watching full videos, across runs three, four, five and seven were calculated (see table 1).

Table 1: Table of Correlations between duration and percentage of people who watched the full video

	Correlation
Run 3	-0.6221957
Run 4	-0.5666553
Run 5	-0.5522303
Run 7	-0.6383265

From table 1, it is evident that there is a negative correlation between video duration and the percentage of people who watch 100% of the videos, across runs three, four, five and seven. The strongest correlation was observed in run seven, with a correlation of -0.638. As expected, as the duration of the video increases, the percentage of people watching 100% of the video decreases, providing further motivation for this research in order to reduce the strength of this negative correlation and increase the percentage of people watching the full videos. The relationship between these has been shown in figure 1.

As part of the initial data understanding phase, it was also important to look at the durations of the videos, and from this, it was found that the duration of the videos in run seven of the FutureLearn course ranged from 99 to 426 seconds. The mean duration of videos in run seven was 231 seconds. The range in durations is consistent with the large range in differences in the percentages of people who are watching the full videos which range from 34.09% to 71.01%. A table of the video durations and views can be seen in table 2. Using the course overviews, it was seen that the video with the lowest percentage of people watching the video was the video in step 3.14 of the course and the video with the highest percentage of people watching the full video was in step 2.10 of the course. From this, it is also clear that the video with the longest duration was not the video with the lowest percentage of people watching 100% of the video, so the possible explanation for this will be investigated later in the project.

Table 2: Table of video durations and views

Video Duration	Total Views	% of People Watching Full Video
99	1041	66.28
362	489	57.46
241	362	49.72
348	476	46.85
281	777	44.92
37	345	71.01
312	282	56.03
92	270	64.44

Video Duration	Total Views	% of People Watching Full Video
426	348	58.05
59	203	63.05
313	220	34.09
227	228	54.82
206	227	56.83

The percentage of people watching 100% of the videos in run seven ranged from 71.01% to as little as 34.09% whereas the percentage of people watching the videos for 5% of the duration ranged from 81.77%, indicating that there at least 20% of people on the course at each stage are not even watching 5% of the video. This provides further support for this research as it will increase to a perentage much closer to 100%. Identifying the extent to which the relationship between video duration and views is true and for what length videos the duration has the most significant effect, will be primary focus for this research project.

Data Quality:

The quality of this data should be high as it was sourced directly from FutureLearn, who are a reputable company. For the runs with files for video statistics (three, four, five and seven), the files are complete and contain no missing observations. The video statistics files for runs one, two and six are not complete meaning that these could not be used in analysis. From the basic level of initial analysis that has been carried out, no unusual data or patterns were returned, suggesting that the quality of this data is high and that it is a reasonable dataset to use for investigation.

Data Preparation

To facilitate analysis, several data pre-processing steps were undertaken to streamline and transform the data set into an appropriate format, meaning that it only included the data that was relevant to this project. At this stage, the only datafile that was used was the run seven video statistics file, as the project solely focused on the run seven videos, meaning that no files needed to be for the completion of this research. All the information needed for analysis was contained in this one file. The course overview for run seven was also used as a reference point for information on the steps of the videos and the structures of each week. The choice to only use run seven data was reasonable as this was the most recent run of the course, and the run in which video duration was most highly correlated with the percentages of people watching the full videos.

As the original data set contained many columns which were not needed, the primary stage of the data pre-processing involved removing the columns that were not needed for analysis, including the continent views, the devices people used to watch the videos and data on the use of transcript and captions. A new dataset was created (‘seven_info’) containing only the information necessary; step number, video duration, total views, and view percentages for 5%, 10%, 25%, 50%, 75%, 95% and 100% of the duration. At this stage, the step numbers were also all converted to numbers (one to thirteen), as these corresponded to the order in which the videos were shown in the course. The course overview confirmed this.

The total views information was quickly disregarded as it was discovered that the total views were not an accurate representation of the proportion of people watching the videos as there were varying numnbers of people enrolled on the course at each step. The total views ranged from 1041 for the first video in week 1 to 203 to the first video in week 3, and this will mainly be due to the fact that people dropped out throughout the course. Because of this, it was a more sensible decision to use the percentages of people on the course at each step for the video views. These had already been given in the original data set and this is what was used for the remainder of analysis. As well as this, by using percentages for the views, it allowed the videos to be compared, despite the fact that the number of people on the course at each step varied.

A new list was made containing information for each of the duration percentages was created at this stage. These lists contained the video duration and the percentages of people watching the video each of these. This completed the data processing stage of the CRISP-DM methodology for this cycle, as the data was in a format that allowed a suitable graph to be plotted. A graph of the video duration and the viewing percentages split up by the duration percentages has been plotted (see figure 2).

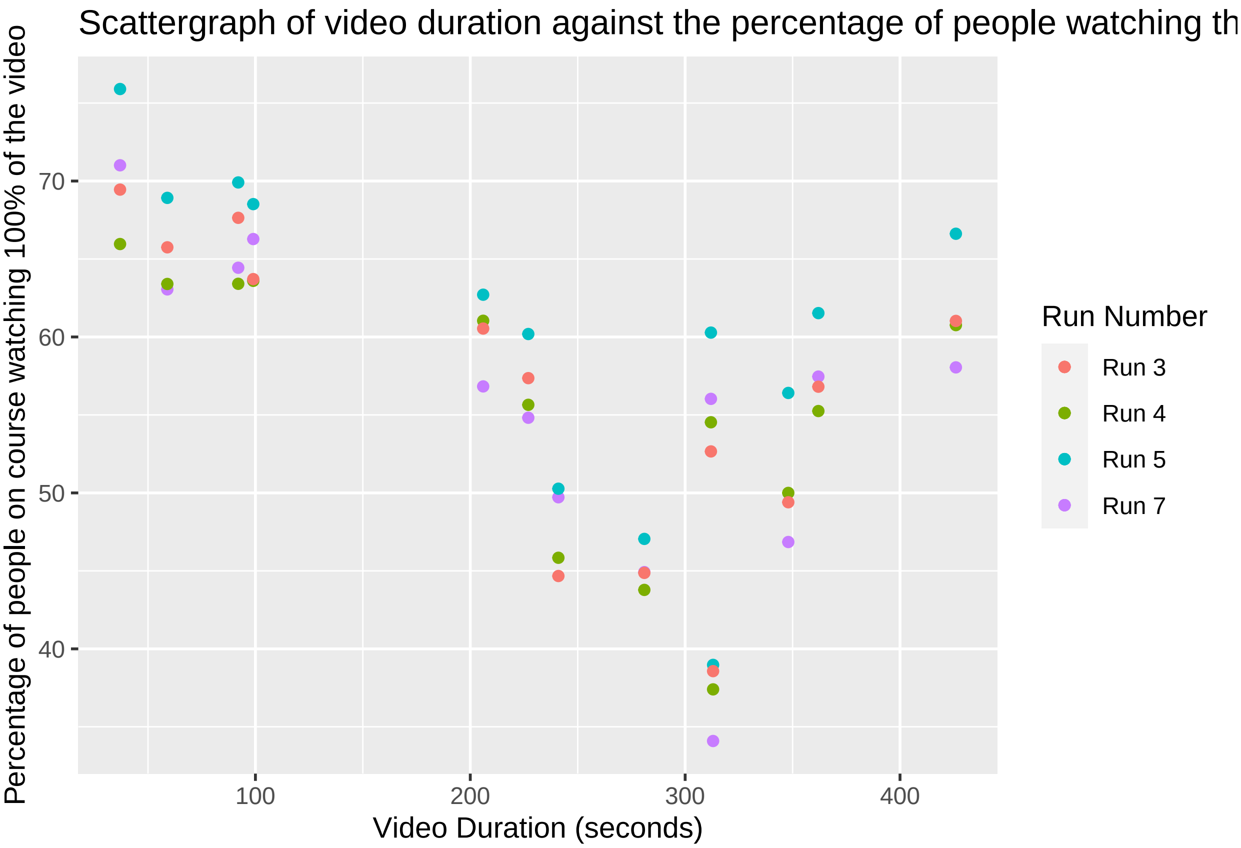

Figure 1: Scatterplot of video duration and percentage of people watching full videos

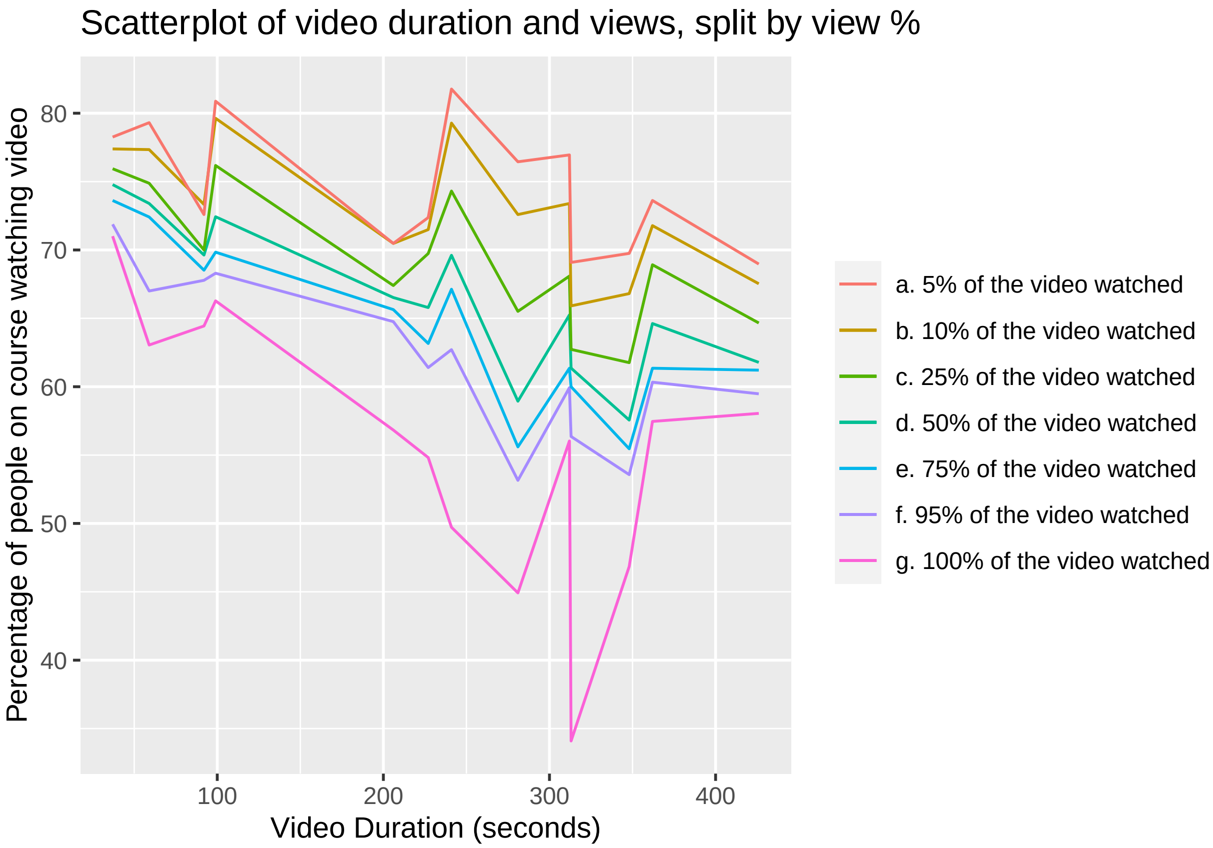

Figure 2: Graph of video duration against percentage views, split by percentage of video watched

Evaluation of Results:

Figure 2 demonstrates the effect that video duration has on the percentage of people watching 100% of the videos, and this is that as video duration increases, people are less likely to watch the video in full. This reinforces the negative correlation that was found earlier in the project. As expected, the line for 5% watched is higher than the 10% watched line, which is higher than the 25% watched line and so on. This has helped to achieve one of the original goals, as the relationship between video duration and views has been successfully determined.

Interestingly, the line on the graph for people watching 100% of the video also seems to fluctuate the most, suggesting that this is the duration percentage in which video duration has the most profound effect. The lines for the 95% of the duration watched and 100% watched are the lines that are the most different, and this difference may be explained by several factors, one of which being that to be counted as a 100% view, the person must watch till the very end of the video. It is likely that most people did not do this, meaning that they would be classed as a 95% view. This may provide an explanation as to why the difference between 95% and 100% is so large as the majority of people in the 95% group may have actually watched the full video but may not have stayed on it up until the very end.

There appears to be a sudden decrease in views at around 100 seconds, suggesting that this may be the point in which people lose concentration, however it is also seen that there seems to be an increase after videos with a 300 second duration. Therefore, there may be some other factors affecting the view percentages, as it is clear that videos with longer durations do not always have lower amounts of views.

The methods used in this cycle are reliable and allow for these results to be reproduced. However, it could be argued that this cycle of CRISP-DM was not as important as first thought as even though this cycle has demonstrated that video duration does have an effect on view percentage, it has not highlighted a particular length of video that people are more likely to watch in full and, as such, has not achieved the original aim of the project. As a result of this, a second cycle of CRISP-DM was carried out to investigate this further, and look into one of the possible factors causing the fluctutations observed in figure 2.

CYCLE 2:

Despite the results of the previous cycle of CRISP-DM highlighting the effect of duration on the percentage of people watching the videos, it did not achieve the original goal of finding the times at which people are most likely to watch the full video. Also, due to the duration clearly not having a consistent effect on the view percentages, other factors needed to be investigated, which in this case is step position. This second cycle of CRISP-DM was carried out, with the same business criteria and aims, but this altered the goal slightly to look deeper into the effects of the step number on views.

Data Preparation

In order to carry out this second cycle of CRISPM-DM, a new list containing the video numbers step and views at each of the duration percentages was created. The video numbers were also converted to numeric variables in R. The mean duration of videos for each of week were calculated at this stage. As well as this, the percentage changes for views at the start of each week compared to the end of the week were calculated and added to the data frame. This was done by calculating the difference in view percentage at the start and end of the week, dividing this by the percentage at the start of the week and multiplying it by 100 (see table 3)

An analysis of the scatter graph shown in figure 3 was performed, to determine the relationship between the step number and the percentage of people viewing the videos, with the lines indicating the start of each new week.

Evaluation of Results:

Figure 3 clearly shows that the percentage of people watching the full video at the start of the week is a lot higher than the percentage of people watching the full video at the end of the week. This is true for all video duration percentages, in all three weeks, but especially week 1. The views in week 1 decreased by 32% over

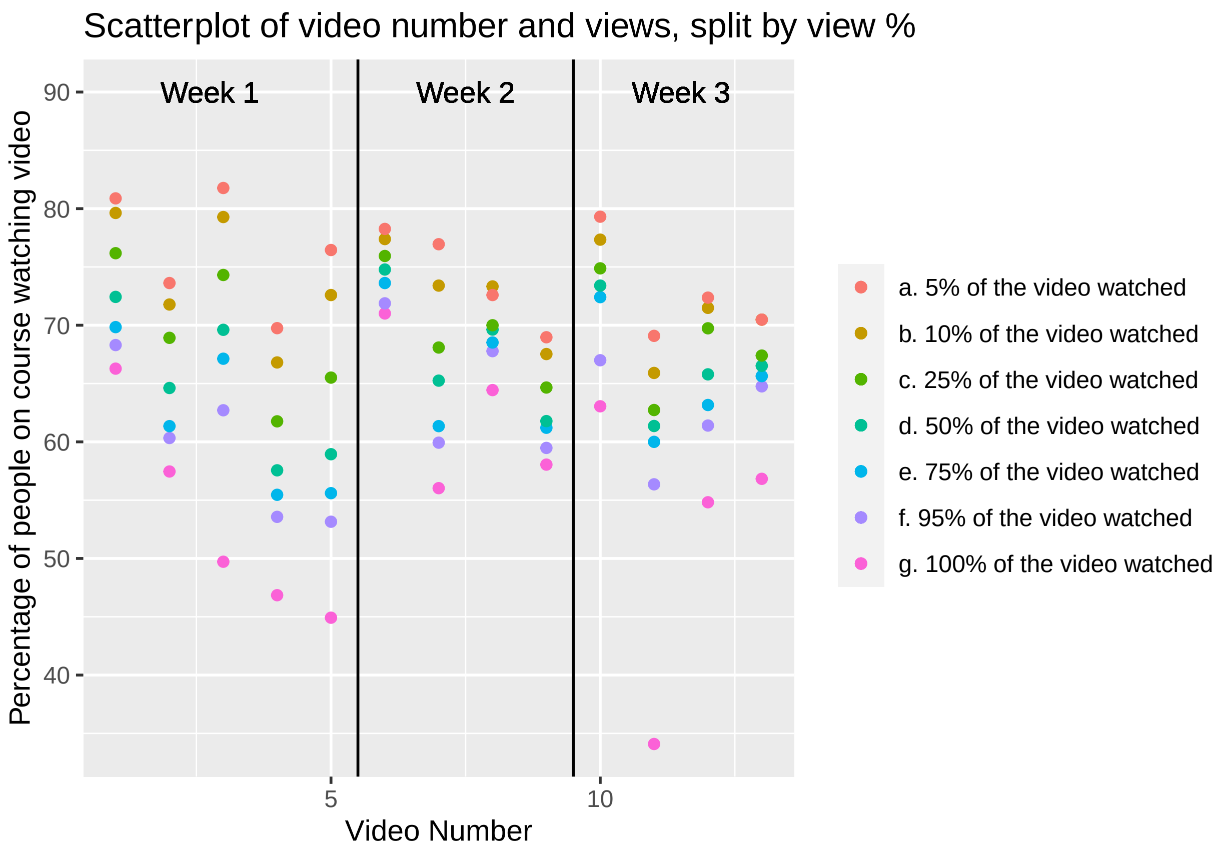

Figure 3: Scatterplot of the video number against percentage of people, split by percentage of video watched

the course of the week, whereas the views for week 2 only decreased 18% and the views for week 3 only fell by 10%. To further investigate the significant decrease in the views for week 1, the number of videos were also calculated, and have been shown in table 4.

Table 3: Weekly Video Statistics

	Number of Videos per week	Mean duration (seconds)	% Change in people watching full video at start and end of week
Week 1	5	266.20	-32.226916
Week 2	4	216.75	-18.250951
Week 3	4	201.25	-9.865186

From table 3, it could be suggested that the combined effect of mean video duration and number of videos per week is the reason for the significant decrease in views in week 1, as this is the week of the course with the highest mean duration of videos (266.2 seconds) and the most videos (5 videos). From this, it may be suggested to FutureLearn that they should aim to not put more than four videos in any given week as going over this will significantly reduce the views of these videos.

Figure 3 also demonstrates the effect that the start of a new week has on viewing percentages. The viewing percentages at the start of a new week are much higher than the viewing percentage at the end of the week before. For example, the percentage of people watching the video at the start week 2 is a lot higher than the percentage of people watching the full video at the end of week 1, as there is a difference of 26.1% in the percentage of people watching 100% of the video at the end of week 1 compared to the start of week 2. This clearly illustrates that people are more likely to watch the full videos at the start of the week compared to the end.

The research thus far has shown that the video duration and number of videos per week do both have an effect on viewing percentage, and further that this effect may be combined. This cycle of CRISP-DM has highlighted the fact that video duration definitely does have an effect on viewing percentages to a certain degree. Despite this, the project goal has still not been achieved, so a third cycle of CRISP-DM will be required to pinpoint a duration for which videos should be. These findings may help contribute to research concerning some of the business objectives, as it is in FutureLearn's best interests to increase the number of people watching the videos in each week, so by not putting more than four videos in a week, it should help to increase the percentage of people watching the full videos.

CYCLE 3:

A third and final cycle of CRISP-DM was carried out, but this time focused on using actual time periods rather than percentages of the total duration for each video. This goals were altered slightly, allowing the focus of this cycle to be more on the exact durations with intention of meeting the business aims.

Data Preparation:

As this cycle required slightly more information than what was already available at this stage, some more data processing steps were carried out. The exact durations at each of the duration percentages were calculated, by multiplying each of the duration percentages by the total duration of each of the thirteen videos. This information was then put into a set of lists, along with the view percentages at each of these durations. The plot of this information can be seen in figure 4.

Evaluation:

An initial analysis of figure 4 suggests that that there is a point at which the view percentages for each of the videos drops quite significantly. The percentage of people watching the videos after 100 seconds is quite

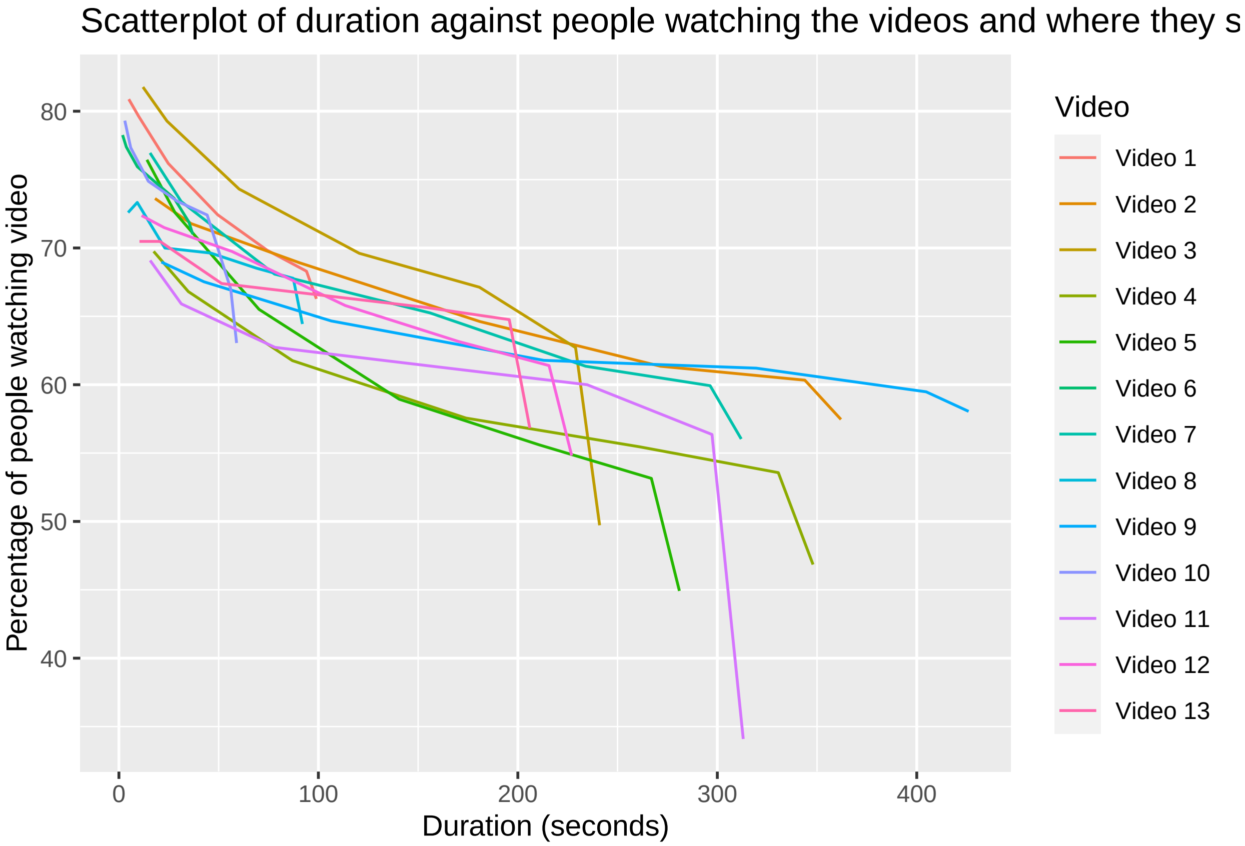

Figure 4: Scatterplot of exact duration and view percentages

low, suggesting that this might be a good time to set as a maximum length of videos. For the majority of the videos, at least 65% of people were watching them at 100 seconds. This is an attainable target for FutureLearn and this would allow them to get at least 65% of people watching the full videos across the course. Alternatively, if the videos are required to be longer than this, this duration could be used as a guide for where to place the most important information in the video. This ensures that the majority of people are still getting taught the key material even if they are not watching the videos in full. This third cycle has met the business criteria and achieved the original project goals.

Final Evaluation:

Through the completion of these three CRISP-DM cycles, the aim of this project has been achieved, as the optimal time for videos has been identified. A suggestion for the length of time that videos should be has been given. FutureLearn should aim to keep their videos to a maximum of 100 seconds, and if they are longer, the most information in the first 100 seconds of a video in order for as many people to watch it as possible. Even though the first two cycles of CRISP-DM did not directly answer the original question, they did provide some useful insight into the effects of duration and number of videos per week on views. It seems that duration does affect the views of the video, however after 300 seconds, duration does not seem to have as significant of an effect on views as first thought. This is what led to the investigation into the effect of step number, which demonstrated that the number of videos per week affects views.

The data-mining techniques that were used in this study were efficient and allowed the goal to be achieved. This process involved extracting the data for durations, step number and views from the original data file and plotting several graphs to represent the relationships between these. Looking back at this, it could be argued that not all of the cycles were needed, in particular cycle two, as this did not provide any information about the durations of videos where people are most likely to stop watching videos. However, it did provide some useful information about the effects of the number of videos per week, which may be used as the starting point for a future data mining project.

Even though a decision about the length of time that videos should be has been reached, there are a few things that may be affecting the reliability of this conclusion that need to be considered. One of which is that for a view to be counted as 100% of the video, the person has to have watched it right up until the very end of the video, meaning that even stopping the video one second before the end would class this person as having viewed only 95% of the video, despite the fact that all the teaching in the video may have stopped. This means that in the future, the 95% duration views may need to be looked at instead. Another factor that is likely to have affected results is the contents of the videos, as for example, people are more likely to watch more of a video if they find it interesting. Thirdly, it is likely that many people would have preferred to just use the transcripts of the video rather than watching the video itself, meaning that even though they did not watch the video, they still learnt the material being taught in it. When designing the videos, FutureLearn should consider all aspects that have been highlighted in this research- number of videos per week, length of the videos and where to put the key information in the video.

To conclude, this investigation has reached a decision about the length of time that videos should be. FutureLearn should aim to keep videos to a maximum of 100 seconds, and if they are longer, then the key information should be put in the first 100 seconds of the video. By implementing these suggestions, the overall viewing percentages of the videos for this FutureLearn course should increase, and this is then likely to improve overall student performance. This project has provided a starting point for a plethora of future research into the effects of duration and number of videos on views and future research may include looking at the combined effects of video duration and number of videos per week and trying to find the optimum combination of these to maximise views.

References:

Chapman, P., Clinton. J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Talking Tree Creative. (2020). Finding the Optimal Video Length | TalkingTree Creative. [online] Available at: https://www.talkingtreecreative.com/blog/video-marketing-2/the-impact-of-video-length-on-engagement/ [Accessed 1 December 2021].