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

Annie Sames

18/11/2021

FutureLearn is an online education platform that runs a variety of different online courses for people to enrol in. These Massive Open Online Courses (MOOCs) can be accessed on any electronic device and give people across the world the opportunity to study many different subjects, through a range of resources including online lectures, articles, and quizzes. In order to optimise the learning and improve the outcome of these courses, analysis of the data for these courses must take place. One area of the courses which can be explored is the videos, including how many people watch them and how long people tend to watch them for. Doing this, will enable companies such as FutureLearn to increase the number of people watching these courses and allow them to make them as efficient at delivering the information as possible.

This investigation will implement CRISP-DM methodology and will cycle through multiple rounds of this. The stages for this methodology are as follows:

- 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

However, for the purposes of this report, this investigation will predominantly 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 FutureLearn as a company are. FutureLearn courses are designed to be accessible to everyone and should be a social experience that people enjoy, therefore the primary aims of these courses are to deliver the material in a user-friendly way that people enjoy, as well to maintain the quality of learning.

Throughout FutureLearn courses, several videos are shown to the students, as part of how new material is taught to the students. As these videos play a key role in delivering material, making these videos as high quality and efficient as possible is crucial, and this includes maximising the number of people who are watching these videos, with the hope that watching more of the videos will better a students' understanding of the material and will therefore improve their overall outlook on the course.

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.

Project Goals:

The goal of this investigation is to identify the amount of time that people tend to watch the videos for, and from this, investigate whether there is a certain amount of time after which most people tend to stop watching the video. From this, a suggestion for a length of time that videos in FutureLearn courses should be to maximise the amount of people watching them can be give. Alternatively, it should give some indication to FutureLearn about the time in which the most important information should be given, ideally the times in which most people watch the videos. This will be beneficial to FutureLearn as it should increase the amount of people watching the full videos at each step of the course which will improve the outcome of the learning taking place in the course.

One of the motives for investigating whether there is a point at which a lot of people stop watching videos is that, from the cyber security course data, the percentage of people who watch 100% of each video is quite low. Also, from the data, it is observed that as the durations increased, the viewing percentages tended to decrease. 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 humans tend to engage in videos for. Reducing the length of the videos closer to the attention span of humans would ultimately be the goal, however within reason as to allow an appropriate amount of information to be delivered in each video.

Project Plan:

To complete this project, several steps will be carried out to achieve the most accurate results and conclusions. Firstly, the relationship between video duration and the percentage of people watching the videos needs to be identified, and from this an analysis can be carried out into what times people normally stop watching the videos. This will enable a conclusion to be reached as this data will highlight the length that FutureLearn should make their videos.

Data Understanding:

The data that was used in this project was collected from seven runs of the Cyber Security FutureLearn course, starting in September 2016 and running until September 2018. For each of the seven runs, eight data files were available which included information on student demographics, survey responses, video statistics, question responses, step activity and weekly sentiment video statistics. The course overviews for each of the runs were also available, and these were used a starting point for this project to become familiar with the structure of the course and gain some insight into a FutureLearn course works. This also helped identify changes between the runs as they could have potentially altered the focus of the investigation.

This investigation focused on the video statistics files, and these contained information on the step numbers of the each of the videos, as well as the durations (in seconds) for these. For each of the thirteen videos, the views were given as percentages of people on the course at the time and had been given for several percentages of the total duration (5%, 10%, 25%, 50%, 75%, 95% and 100% of the total time). The video statistics files also contained information on the number of people who used transcripts and captions, downloaded the video, viewed it n HD and the devices that were used to watch them. Data on the percentages of people who watched the videos across all seven continents was also given in this file.

As an initial data exploration exercise, the correlations between the video durations and the percentage of people watching 100% of the 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 between duration and % people watching full video
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 the video duration and the percentage of people who watch the full videos, and this is seen across all the runs. As expected, the duration increases, the amount of people who watch the full video decreases, further motivating the need for this research. As a correlation across all runs was evident, this was then plotted in a scattergraph (see figure 1).

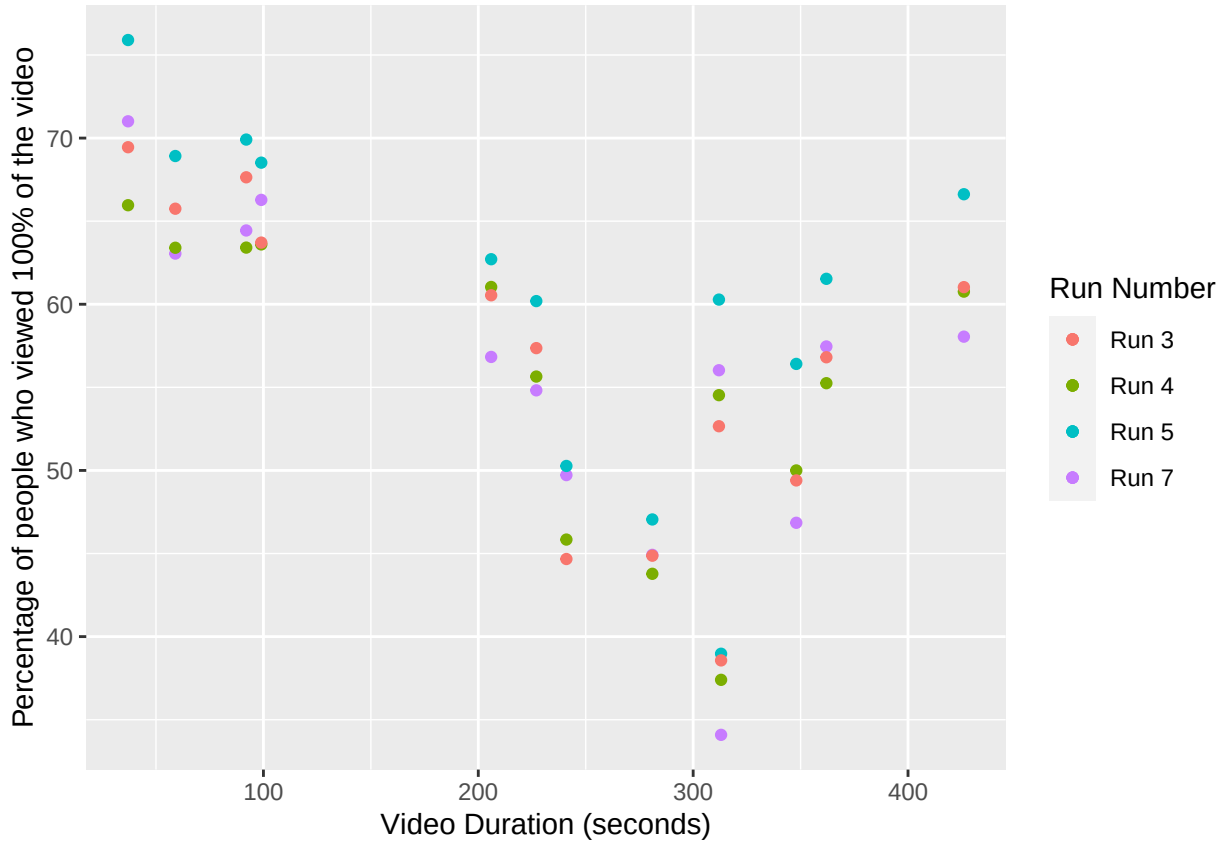

Figure 1: Scatterplot of video duration and the percentage of people who watch the full video across runs 3, 4, 5 and 7

Another part of the initial data understanding phase involved looking at the durations of the videos, and from this it was seen that the durations have a large range, ranging from 99 seconds to 426 seconds. These also initially seem to be consistent with the large range in differences of the percentages of people who watch the full videos. A table of the video durations and the percentage of people watching the full videos is shown in table 2.

Table 2: Table of video durations and views

video_duration	total_views	viewed_onehundred_percent
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
426	348	58.05
59	203	63.05
313	220	34.09
227	228	54.82
206	227	56.83

Identifying whether this is true and for what lengths videos the effect of the duration is most significant will be primary focus for this research project.

The quality of this data should be high as it has been sourced directly from FutureLearn, who are a reputable company. Also, the video statistics files for runs three, four, five and seven have no missing observations. However, one aspect of the data that was missing were the video statistics files for runs one, two and six, meaning that the video statistics for these runs could not be used for the project. At the basic level of initial analysis, no unusual or unexpected data or patterns were highlighted from the graphs or from the raw numeric data, suggesting that it is of high validity.

Data Preparation:

For analysis to take place, several data pre-processing steps were undertaken to streamline and transform the dataset into the correct format and to only include the information that was needed for this project. At this stage, the only datafile that included in this project was the run seven video statistics data, as this project has solely focused on the videos in run seven and all the information needed to carry out this project was already contained in the file. The course overview for run seven was also used throughout the investigation as a reference point for the steps of the videos and the structures of each week. The video durations and view percentages were already found to have a negative correlation at this stage (see figure 1), providing further reasoning for the use of just the run seven data, as well as the fact that it is the most recent and up to date run of the course. This means that no data files were merged at this stage of the project.

As the original data contained many columns which were not needed in this case, the primary stage of the data pre-processing involved removing the columns that were not needed for analysis, including the continent views, which devices people use and the transcript and caption data. From this, a new dataset was created ('seven_info') which contained 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 through the course. The course overview confirmed this.

In addition, to this the total views column was quickly disregarded as well as the total views were extremely different for each of the videos and does not actually give an indication of the proportion of the people on the course who were watching the videos (see table 3). Throughout the remainder of the project, the percentage of views at each duration were used instead of number of views as this allowed each of the videos to be compared, even though they all have different numbers of people watching the videos. These percentages had already been given in the original datafile, so there was no need for any calculations here.

Table 3: Table of total views for each video

Video Number	Total Views
1	1041
2	489
3	362
4	476
5	777
6	345
7	282
8	270
9	348
10	203
11	220
12	228
13	227

A new list of duration percentages was also created at this stage, and these contained columns for video duration and the each of the percentages for people watching the video at each of the duration percentages. This was the final step before the investigation could start as this allowed a graph of the video durations and the viewing percentages to be plotted (see figure 2).

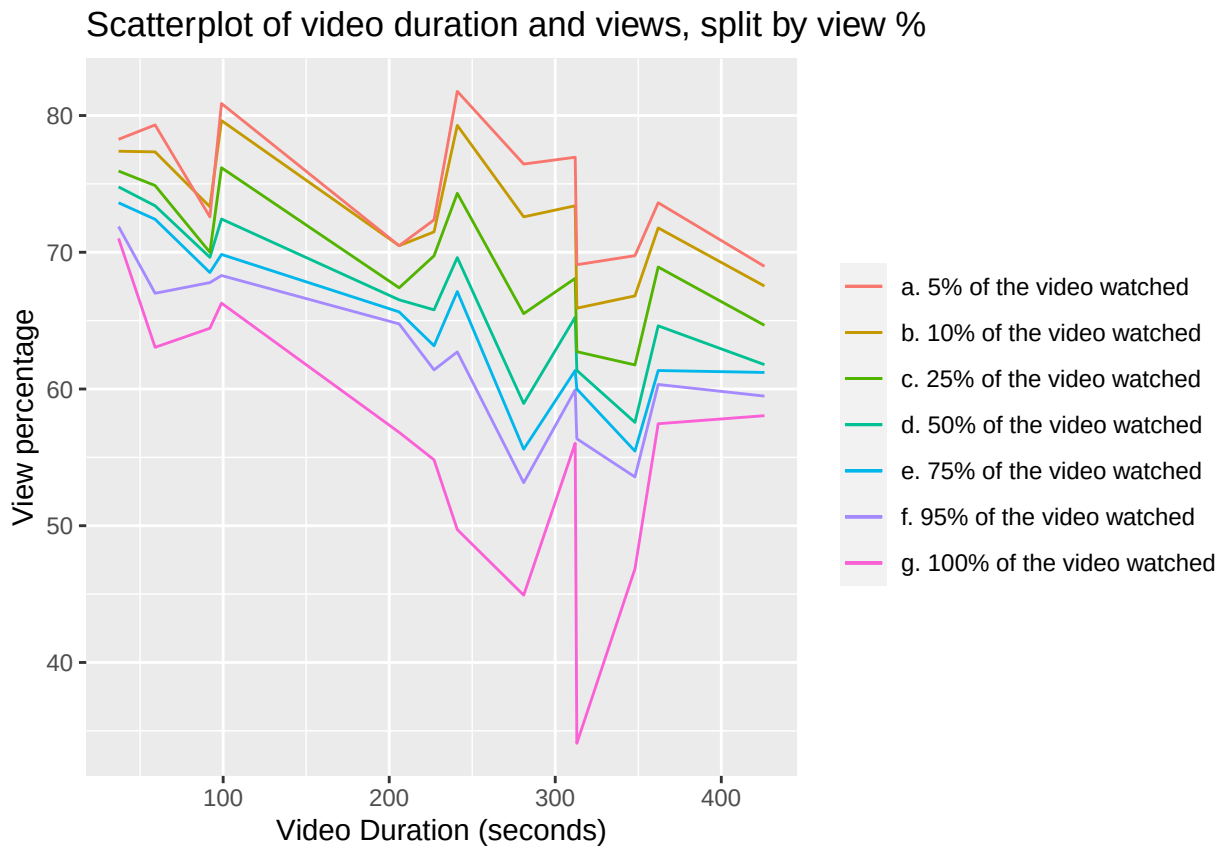

Figure 2: Graph showing the video duration against percentage of people watching, split by percentage of video watched

An analysis of this graph clearly shows that as video duration increases, the percentage of people watching the videos at each of the video durations decreases, reinforcing this negative correlation. 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.

Interestingly, the line for the people watching 100% of the video also seems to be the one with the most fluctuations and the one that video duration appears to have the biggest impact on. The biggest difference between duration percentages is between 95% and 100% of the videos, and this may be due to several reasons such as people clicking off the video a couple of seconds before the end meaning that it would not be counted as 100% duration, however this is something which will be discussed later. This graph also shows a sudden decrease in people watching 100% of the videos after around 310 seconds, and then a gradual increase after this, suggesting that there may be some other factors that affect views.

Even though this analysis has shown that there is a relationship between video duration and view percentages, it has not helped to highlight a particular length of videos that people watch and therefore 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.

CYCLE 2:

Despite the results of the previous cycle of CRISP-DM highlighting the effect that the duration of the videos has on the percentage of people watching them, it has not achieved the original goal of finding the times at which people are most likely to watch the full video. Also, as the duration did not always seem to have a consistent effect on the view percentages, a further variable needed to be investigated, which in this case is step position. As a result of this, a second cycle of CRISP-DM was carried out, but this time focused on identifying the effect that the step number of the videos has on the view percentages to see if a pattern between step number and view percentage could be identified.

Data Preparation To do this, a new list containing the step positions and the duration percentages was created. The step positions had already all been assigned video numbers, however, to plot these in order on the graph, these also were converted to numeric variables in R.

To identify the relationship between step number and amount of people viewing the video, a scatter graph was plotted, with the addition of lines showing the start of each new week (see figure 3).

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, but particularly week 1 with a percentage decrease in people watching the full video at the start of the week compared to the end of 32%. These percentage changes are shown in table 4. To further investigate the significant decrease in week 1, the number of videos in each week have also been shown in table 4, confirming the fact that the number of videos per week also has an effect as week 1 has the most videos and the biggest percentage decrease in people watching the full video.

Table 4: Table with number of videos per week and percentage change in views at beginning and end of week

	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

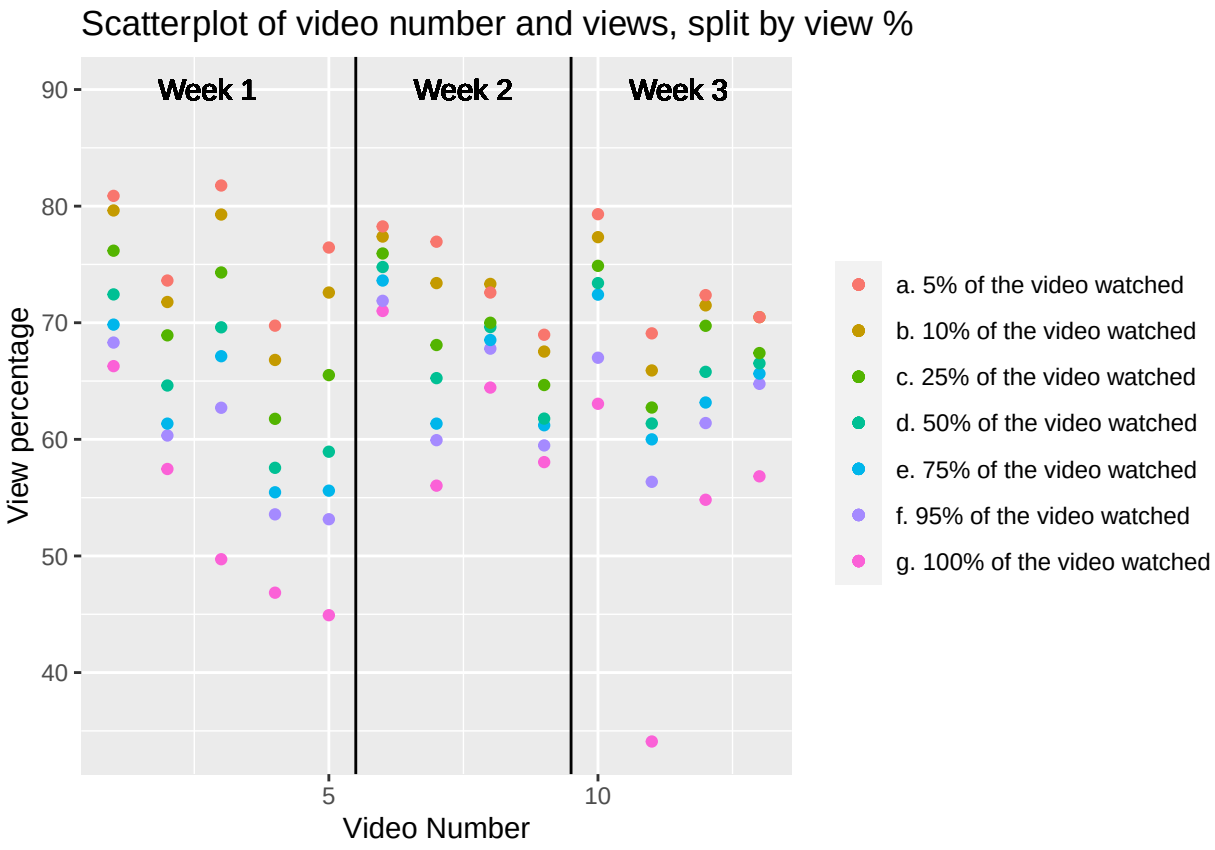

Figure 3: Scatterplot of the video number against percentage of people, split by percentage of video watched

Further, combining this with the previous research and looking into the effects of the duration of the videos (see table 4), it is seen that again the week with the biggest percentage decrease (week 1), is also the week with the highest mean duration of videos. The mean duration of the video and the number of videos have both therefore been found to affect viewing percentages, which again is to be expected, due to fatigue over the week or boredom.

Figure 3 also demonstrates the effect that new weeks have on viewing percentages as 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.

The research thus far, has shown that the video duration and number of videos per week do have an effect on viewing percentage. However, even though this has produced some interesting findings, it has still not contributed to determining a length of video that would encourage as many people to watch the full thing as possible.

CYCLE 3: