CSC8631 Report

Selina So

02/12/2021

1. Business Understanding

1.1 Business Objectives

1.1.1 Background

Learning Analytics is a study of the *"measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the information system in which it occurs"* (Shi and Cristea 2018). The benefits of Learning Analytics is that it will provide insights to the factors which influences learners' retention. Learners' retention is one of the key drivers for institutes to implement Learning Analytics, as retaining students and their associated fees has a significant economical impact on the institutions' income (Shacklock 2016). The insights from the Learning Analytics will enable course designers from educational institutes and MOOC (Massive open online course) providers to make informed decisions on the design and improvements of their courses, thus improving the learning environment for learners and drive more influx of learners enrolling.

FutureLearn is an MOOC provider, which collaborates with universities globally to offer online courses. Since their launch in 2013, they have attracted over seven million learners across the world (www.futurelearn.com). With a global reach of this extent, it is therefore crucial for FutureLearn to understand their performance in engaging with learners and providing an enhanced learning experience, which will retain and improve the learners' retention rate. The insights derived from Learning Analytics will therefore enable FutureLearn to understand areas of design or improvements which could create a positive impact for FutureLearn, their collaborators and their learners, in addition, to understand the key factors that could influence the retention rate of students.

1.1.2 Business Objectives

This study will investigate the *Cyber Security Safety at Home, Online, in Life* online course, which is a course delivered by Newcastle University on the FutureLearn platform (www.futurelearn.com/courses/cyber-security). There are many factors which could influence the learners' retention rate. Data from activities, such as videos, could act as engagement indicators of the learners and potentially allow early detection of learners' disengagement (Bote-Lorenzo and Gómez-Sánchez 2017). In this study, we will examine data from the *Cyber Security Safety at Home, Online, in Life* course to understand the factors which influence the learners' retention rate. We will look into if and how the continent the learners come from has an impact on the engagement of learners with the course. This insight will help to ensure that informed decisions can be taken to ensure the course appeals to learners from across the world.

1.2 Assess situation

The *Cyber Security Safety at Home, Online, in Life* course data is provided by FutureLearn (via Newcastle University). There are 53 .csv data files which provide information on the learners and the course. The remaining 7 .pdf files provide an overview of the course structure.

The course is divided into three weekly blocks of study. For each weekly block, there are a number of steps to complete. The first week block contains 18 steps, and the second and third week blocks contains 21 steps. (Shi and Cristea 2018) (www.futurelearn.com/courses/cyber-security). Each blocks consist of a combination of videos, articles, exercises, discussions, quizzes and tests for the learners to complete throughout the course.

The 53 .csv data files are split into 7 runs. Each run represents different time-frames of when the data was collected. All courses ran between mid-2016 and mid-2018. All runs consists of the following data files: *'Archetype-survey-responses,' 'Enrolments,' 'Leaving-survey-responses,' 'Question-response,' 'Step-activity,' 'Weekly-sentiment-survey-response, 'Team-members,' 'Video-stats'*, with the exception of run 1, which does not contain the *'Team-members'* and *'Video-stats'* datafiles, and run 2, which does not contain the *'Video-stats'* data file.

1.2.1 Inventory of Resources

The following sections will list the resources available to the project.

1.2.1.1 Software Sources The following software packages have been used:

- R
- RStudio
- Git

The use of these software sources are described below under *'1.2.2.1 Requirements'*.

1.2.1.2 Sources of Data and Knowledge

- The CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) will be applied to structure the project life cycle ([*TO DO] link the CRISP-DM guide*)(Chapman et al. 2000).
- 53 .csv data files and 7 .pdf files on the *Cyber Security Safety at Home, Online, in Life* course, provided by FutureLearn (via Newcastle University).

1.2.1.3 Personnel Sources The following personnel will be used for expert domain knowledge and technical support as well as to gain stakeholder guidance on the business perspective.

- Newcastle University lecturers (Dr Matthew Forshaw and Joe Matthews)
- External teaching experts

1.2.2 Requirements, assumptions, and constraints

1.2.2.1 Requirements Applying best programming practice is crucial for this project to enable reproducibility. Therefore the following software and packages will be implemented to apply best practice:

- R: Used for all data analysis
- RStudio: Integrated development environment to develop report

- ProjectTemplate: R package to automate project file structure
- RMarkdown: R package to produce the report
- ggplot2: R package to produce data visualizations
- Git: Used for version control

There are legal obligations and privacy policies, such as *GDPR (General Data Protection Regulation)* and the *Data Protection Act (2018)*, to consider before using the data ([*TODO: check link to where this is discussed]* see Section 1.2.2.2 Assumptions).

1.2.2.2 Assumptions The following assumptions have been made on the data:

- Assumed that full consent to use the data for this study has been provided by FutureLearn. To comply with the legal and ethical standards, we will ensure any potentially identifying data will be anonymised to reduce the likelihood that an individual could be identified.
- Assumed that although *'Video_stats'* data are not provided for run 1 and 2, this does not indicate that video learning material were not used for these runs.
- As the data is provided by FutureLearn (via Newcastle University), it is assumed that the data provides an accurate and reliable reflection of the learners taking the course.
- There were no descriptions for the data, therefore assumptions will be made as to what the data means. The assumptions are clearly stated in the appropriate sections.

1.2.2.3 Constraints The project is to be completed by 3rd December 2021.

Due to the time constraints, the key phases from the CRISP-DM methodology which require focus are *Business Understanding*, *Data Understanding*, *Data Preparation* and *Evaluation*.

1.3 Data Mining Goals

For this study, we will investigate the course data and initially decide on a set of data to analyse in more detail to understand students' engagement with the online course. This set of data will be chosen according to (a) the richness of information contained in the data, and (b) the completeness of the data that is available. Based on this, the most promising lines of investigation will be decided.

The goal is to derive insight from the data on engagement and retention during the course which will enable Newcastle University and FutureLearn to potentially modify and improve the course content to achieve optimised learner engagement and retention.

1.4 Project plan

We will utilize the CRISP-DM process to understand the data and ensuring that the insights meet the business objectives. We will perform initial investigation of the data (using a combination of simple descriptive statistical and visualization techniques) to identify potential trends and form hypotheses. Depending on the outcome of the previous phase, this shall initiate further in-depth analysis with the vision of providing better understanding and valuable insights for Newcastle University and FutureLearn.

Throughout the course of the study and depending on the outcome of the results, certain phases of the CRISP-DM methodology will be re-iterated multiply times, to further support the previous findings.

1.4.1 Initial assessment of tools and techniques

The use of the CRISP-DM methodology is useful to provide structure to the lifecycle of this study and ensuring the analysis remains relevant to the business objective.

R was very useful in enabling the analysis of the data. In addition, the packages provided by R allow the project to be reproducible with minimal effort. The key package is ProjectTemplate, which can automatically build the directory to structure the project and the files, and can automatically load data and libraries.

Git is a useful software to ensure that all creations and changes are tracked. Therefore one can revert back to a specific version of the project or change if required.

Simple statistical description and visualisation of the data provides insight into the nature and quality of available data, and will enable fast decisions on the next steps in the analysis.

Discussions with domain experts provided technical support for the delivery of the data analysis, and critical feedback to ensure the project objectives are aligned to the requirements of key stakeholders.

2. Data Understanding (Initial Observation)

2.1 Collect initial data

The 53 .csv data files of the *Cyber Security Safety at Home, Online, in Life* course data is provided by FutureLearn (via Newcastle University) and are loaded and cached in the *'data'* folder, as designed by the ProjectTemplate package.

There are 7 runs of data. Each run of data was measured several months apart from each other, between mid-2016 and mid-2018. Below is a brief description and qualitative assessment of each set of data files.

- *'Archetype-survey-responses'*: The data files consist of two sets of Id-related columns, a datetime column and a categorical column *'archetype'*. As no descriptions were provided, it is difficult to deduce the exact meaning of the archetype data without further information. In addition, the data files for runs 1-3 are either empty or incomplete.
- *'Enrolments'*: The data files contain information of learners' ID, enrollment and unenrollment date and time stamps. In addition, the files contain categorical data on gender, country, age, education and employment status and detected country. The columns *'fully_participated_at'* and *'purchased_statement_at'* contain date and time stamps, however the significance of these columns is unclear without further information. There are many *'Unknown'* in the data files.
- *'Leaving-survey-responses'*: The data files contain information about learner ID, leaving date, and learners' feedback, which was given as what is assumed to be pre-set selection of feedback options. The assumption was made as the comments all fall within a small set of answers (e.g. 'I prefer not to say,' 'The course required more time than I realised,' 'The course was too hard,' 'Other'). It contains information on the last step completed when response was provided, however the data is incomplete as many rows contain missing data. The data files for runs 1 to 3 are empty.
- *'Question-response'*: The data files contain information of the learners' ID and the quiz questions which the learners have attempted, plus the submission date. There is a *'correct'* column containing boolean datatype of 'True/ False' which represents whether the learners have answered correctly. There is a *'response'* column which contains a selection of numerical values and a *'cloze_response'* which appears to be empty, therefore the meaning of this column is unclear. It is unclear whether the learners have to answer all questions correct, as there is more than 1 number in the *'response'* column, to obtain a 'True/ False' under the *'correct'* column. Therefore further information is required in order to analyse these data sets.

- *'Step-activity'*: This data file contains information on the learner_id and the date and time stamps of when they had first visit and last completed a particular step. There is missing data in the *'last_completed_at'* column.
- *'Weekly-sentiment-survey-response'*: the data files contain an ID column, with the date and time stamps of the individuals' responses. It is unclear what the IDs refer to as they do not specify that the column is for learners' ID, therefore more information is required. The *'reason'* column contains free text, which is interesting. However data files from run 1-4 are empty, run 5 contains only 1 incomplete data row and the remaining runs (6 and 7) contain a mixture of incomplete and unstructured text data for the free-text *'reason'* column.
- *'Team-members'* (not available for run 1): the data files contain IDs, the team role and user role of the individuals. The *'first_name'* and *'last_name'* columns have been anonymised by the provider of the data file by removing the names of the individual learners and replacing them with the words 'First' and 'Last' for the columns *'first_name'* and *'last_name'*, respectively.
- *'Video-stats'* (not included in run 1 and 2): the data files contain information on the video topics, and numerical data on the number of viewers per video, devices and features used to view the videos, how long individuals watched the videos and the percentage of viewers from each continent. The contents of this data file will be examined in more detail in Section 2.2.

2.1.1 Initial data file selection

The initial data file of interest was the *'Weekly-sentiment-survey-response'* data. These data would have been useful to combine with the data from the *'Leaving-survey-responses'* data files to potentially obtain a mixture of feedback which will provide direct insight into the learners' experience to the course provider. The *'reason'* column was interesting as it contains free text, therefore provided a direct source of feedback that could be analysed using Natural Language Processing (NLP). However, after an initial assessment, it appears that the data contained positively biased views of the individuals' experience with the course. The unstructured free text were also of poor quality as they contained incomplete sentences, single word feedback with lack of context and text that contains random symbols. In addition, there were many missing feedback data and due to only runs 6 and 7 containing a small selection of feedback, there is therefore insufficient data to draw substantial conclusions with the data. The *'Leaving-survey-responses'* data files also contain pre-selected responses, which indicates that the learners were only allowed to select from a very narrow range of opinions. We concluded that the data file does not provide meaningful insight into the sentiments of the learners. Lastly, the two sets of data files do not contain the same type of ID columns. The *'Leaving-survey-responses'* data files contained *'Learner_ID'*, which leads to the assumption that the responses were provided by the learners. The *'Weekly-sentiment-survey-response'* data files contained *'ID'*, which we cannot assume that the responses were from the learners only, as other data files (such as, the *'Team-members'* data files) have already shown that 'educators,' 'mentors,' etc all have an ID number assigned to each one of them. The responses could potentially come from individuals with conflicted interest with the course, therefore the responses will not be a reliable source of information to draw conclusions on the course.

2.1.2 Final data file selection

As data from videos could act as engagement indicators, Newcastle University and FutureLearn could therefore use the video data to understand students' engagement with the course, and consequently the retention rate of the learners throughout the course. This will enable the course provider to make informed decisions on the design and improvements of the course. To achieve this, the focus will be on the '*video.stats*' data files provided by FutureLearn. These data sets are only available for 5 (out of 7) runs. Therefore runs 1 and 2 will not be considered in this study as no data is available. There appears to be no missing values from the data, therefore the data files for '*video.stats*' are complete.

2.2 Describe data

Displaying below is the list of column names of the '*video.stats*.'

```
##  [1] "step_position"                  "title"
##  [3] "video_duration"                 "total_views"
##  [5] "total_downloads"                "total_caption_views"
##  [7] "total_transcript_views"         "viewed_hd"
##  [9] "viewed_five_percent"            "viewed_ten_percent"
## [11] "viewed_twentyfive_percent"      "viewed_fifty_percent"
## [13] "viewed_seventyfive_percent"     "viewed_ninetyfive_percent"
## [15] "viewed_onehundred_percent"      "console_device_percentage"
## [17] "desktop_device_percentage"      "mobile_device_percentage"
## [19] "tv_device_percentage"           "tablet_device_percentage"
## [21] "unknown_device_percentage"      "europe_views_percentage"
## [23] "oceania_views_percentage"       "asia_views_percentage"
## [25] "north_america_views_percentage" "south_america_views_percentage"
## [27] "africa_views_percentage"        "antarctica_views_percentage"
```

There are 13 rows of data in each '*video.stats*' data file, one row corresponding to each video content throughout the course.

There are 28 columns. The first column of each dataset describes the weekly block and the step position of where a particular video is located within the course. The second column contains strings of data which refers to the video title, which we assume describes the video content. The remaining columns (columns 3-28) contain numerical data, some of which are percentage values. Care will have to be taken when merging the runs together for the columns containing percentage values. The percentage values are averaged across runs using the weighted average to account for the different number of learners in each run.

We will make two assumptions for this project: (1) The number of views is equal to the number of learners, i.e. they either do not include the views of tutors or hosts or these are neglible, and (2) each viewer only viewed each video once. The data files contain information on the video topics and the absolute number of viewers. This will inform which topics appears to attract the most number of learners. One can also use the range of columns on the 'viewed_percentage,' to deduct how the number of learners' engagement changes throughout the duration of the videos. Finally, from the percentage of viewers across the continents, we can observe how engaged the learners from different continents have been throughout the course. Therefore the data file provide a rich source of information to gain insight into the learners behaviour to inform the business objectives of the project. Due to time constraints during this project, the data on the type of devices, downloads, as well as the number of HD users will not be selected for this study. These columns may be interesting for further analysis on how to optimise the content of videos to ensure they are in the right format and have the correct features.

- Basic statistical analysis shows that there is some skewness in the *'total_views'* data, ranging from a minimum of 446 total views to 1659 total views. By analysing the number of learners for each topic, it potentially show whether certain topics attract more viewers.
- Basic statistical analysis shows that there is some skewness of the data in columns *'asia_views_percentage'* and *'africa_views_percentage'*, therefore further investigation can potentially show if there are higher drop out rates from some continents compared to others.
- The unit for the values of the video duration is included in the course overview .pdf files. The shortest and longest videos ranges from 37 seconds to 426 seconds (~7 min). The average time of videos is 231 seconds (~4 min). The duration of the video could possibly impact the learners' engagement to the course.

- From the 'viewed_percent' columns, it is clear from the summary that there is a downward trend in the number of viewers throughout the duration of the videos. Further analysis may help to identify if a maximal duration of the video exist in order to keep learners engaged.

2.3 Explore data

The following initial visualisations will be used to explore the data and help determine the areas to consider for further investigation. As the first set of video data begins from run 3, the data file 'cyber-security-3_video-stats.csv' will be used for the initial exploration.

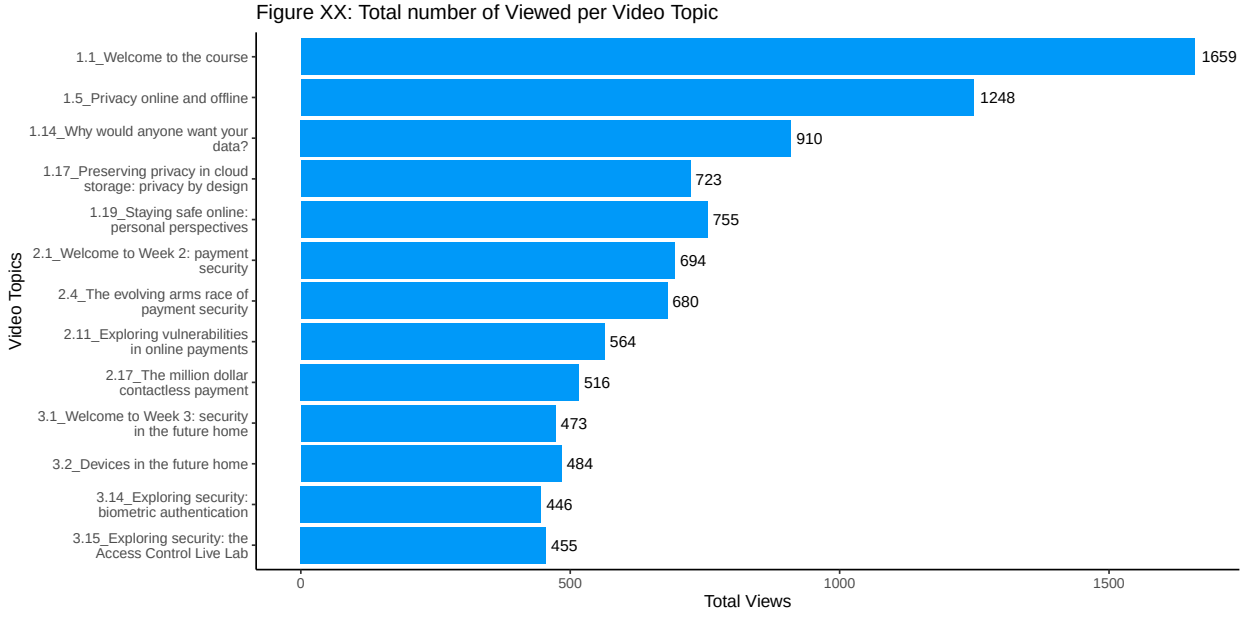
The two first columns *'step_position'* and *'title'* will be combined to allow quick reference to the order of which the videos appears throughout the course. This new column will be called *'step_title'*

2.3.1 Number of viewers for each video topic

Part of the data mining goal is to understand whether there is a trend on learners' engagement, therefore we will look at how the number of viewers have changed throughout the course.

Figure XX shows the number of viewers for each video based on the topic of the videos and the absolute number of views for each other these videos (run 3 data).

Figure XX: Total number of Viewed per Video Topic

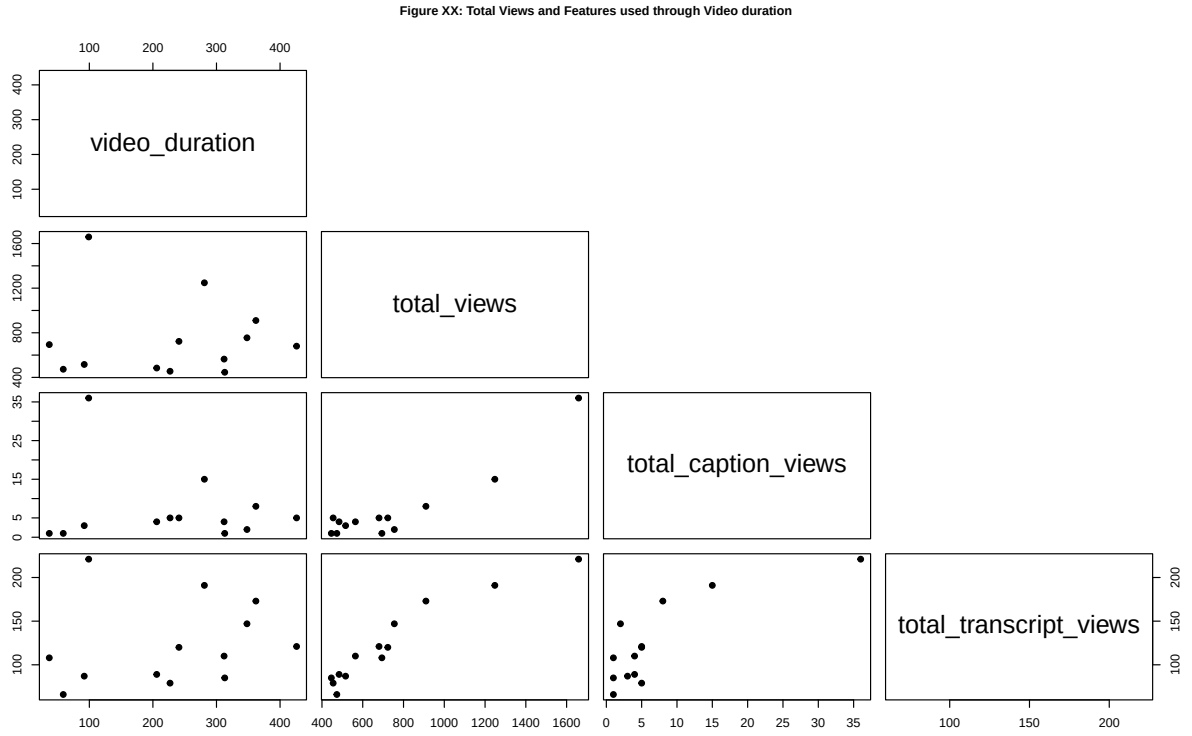


The bar chart clearly indicated that as the learners progresses throughout the course, there is a drop of viewers watching the videos. As the learners are watching from different continents, we will investigate if this trend continues in a similar manner for the learners from different continents and whether certain topics appear to attract learners more.

2.3.2 Total number of viewing and features used through the duration of videos (TO DELETE??)

A scatterplot matrix will be used to visualize any pairs of relationships. For the plots below the headings, the headings will be the x-axis, and the corresponding rows will be the y-axis.

Figure XX attempts to demonstrate the relationships between the length of the videos and the number of views, captions and transcripts for each video. It is assumed that the column '*total_transcript_views*' refer to the number of learners reading the transcript version of the videos rather than watching the videos.



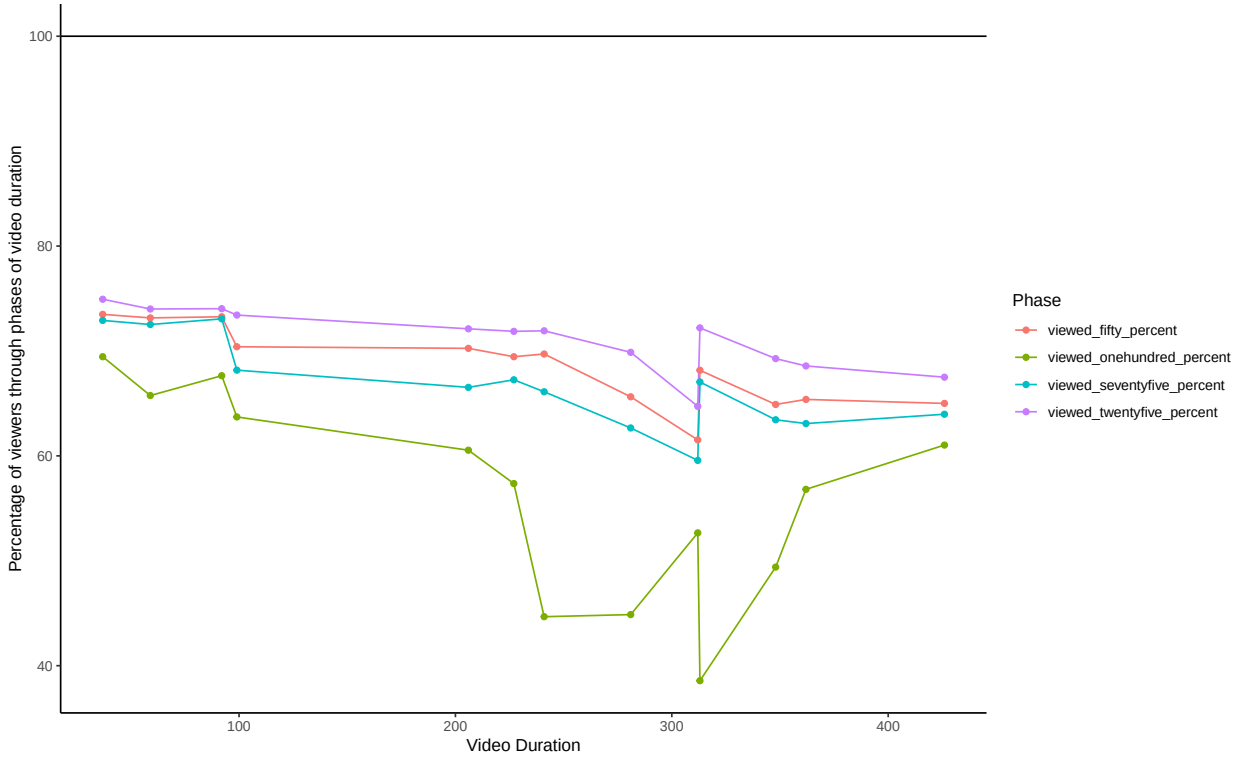
There are no obvious relationships observed between the length of the videos and the amount of views and features used for the videos. As the total number of views increases, so does the number of captions used and transcripts used, which is to be expected. As there are no unexpected patterns that would indicate a dependence of e.g. the number of transcript reads on video duration, we will not continue with investigating these columns in our study.

2.3.3 Number of viewers for video duration

The key plots of interest are on the left column. The y-axis for the percentage of viewers is decreasing throughout the duration of the videos. Further investigation will look at how large the decrease is (??) and at what point throughout the duration of the videos do we see the largest drop of viewers. This can show how engaged can learners stay for and if the duration of the videos have an impact of how long the learners stay engaged for.–>

Figure XX is a line chart to demonstrate the relationship between the video duration and the percentage of viewers who watched 25%, 50%, 75% and 100% of the video.

Figure XX: Number of viewers watched 100% of videos



Each line represents how many learners have watched a certain percentage of the videos. This is plotted against the duration of the videos on the x-axis.

The graph shows that there is a steady drop off of learners watching the entire video. The majority of the drop off of learners happens between 75% and 100% of the video duration. There is potentially a tendency of learners to drop off more if the videos are longer, however there is too much noise in the data to be certain.

Further investigation will look at this behaviour with data from all runs. This will help understanding two potential points: (1) Is a point throughout the duration of the videos (e.g. between 75% and 100%), independent of total duration of the video, where the drop off rate is the largest? (2) Is there a maximum duration of the video after which the learners are decide not to watch the video? This will help answer the question if (1) the content after a certain point in the videos is not of interest for the learners, and (2) how long learners stay engaged for.

2.3.4 Hypothesis

The initial exploration in the previous Section has shown a number of interesting areas which we will further analyse, particularly around the viewing rates from learners of different continents and whether the video topics and length of the videos are a contributing factor to the learners' retention rate. We will investigate the video data from the course to answer the following questions:

1. Does the *duration* of the videos have an impact on the viewing rates across different continents?
2. Does the *content* of the videos have an impact on the viewing rates across different continents?
3. Is there a correlation between duration of videos and drop out rate of the learners?

2.4 Data Quality

The data sets from each run have been compared and it was found that they all contain the same number of rows and columns, and the labels for rows and columns are consistent across all 5 runs. The data sets are

mostly complete with no visible missing data. The format, variables and completeness of the data files are all consistent. Therefore the quality of the data is good and the merging of the data files will be straightforward.

The column *'antarctica_views_percentage'* only contains values of 0. Based on this, it can be assumed that there are no learners from Antarctica, which is reasonable as according to the *World Population Review*, there are only 1000-4000 seasonal residents in Antarctica (www.worldpopulationreview.com/continents/antarctica-population).

3. Data preparation and modelling

In the following Sections, we will discuss the steps that have been taken to analyse the data to answer the three questions in Section 2.3.4. The analysis has been done in 2 cycles for each question. In the first cycle, we analysed the data from run 3 only. Based on the results of the first cycle, we decided on the most promising route of analysis and performed the analysis taking runs 3-7 into account for the second cycle. For each of these questions, we will discuss the data preparation and modelling steps separately (Sections xx-yy [CHECK SECTION NUMBERS]), before evaluating the results in Section xy [CHECK SECTION NUMBERS].

3.1 Data Preparation and Modelling for Q1: Does the duration of the videos have an impact on the viewing rates across different continents?

3.1.1 Select data

The first hypothesis will investigate on whether the duration of the videos have an impact on the viewing rates across different continents. Therefore it could indicate whether the length of the videos will have a negative or positive impact on the learners' engagement globally. The key columns for this investigation will be *'video_duration,' 'total_views,' 'europe_views_percentage,' 'oceania_views_percentage,' 'asia_views_percentage,' 'north_america_views_percentage,' 'south_america_views_percentage,' 'africa_views_percentage' and 'antarctica_views_percentage'* from the *'video_stats'* data file, which has been assigned to the *'run3unite'* variable for the first cycle and *'run3unite,' 'run4unite,' 'run5unite,' 'run6unite,' 'run7unite'* for the second cycle. Table xy [CHECK TABLE NUMBER] shows the raw data for run 3. The same selection of columns has been used for all remaining runs to include in the second cycle of analysis.

```
## # A tibble: 3 x 9
##   video_duration total_views europe_views_per~ oceania_views_p~ asia_views_perc~
##            <int>       <int>             <dbl>            <dbl>            <dbl>
## 1             99        1659              55.2             2.29             16.1
## 2            362         910              65.4             2.86             10.2
## 3            241         723              66.2             3.18              9.82
## # ... with 4 more variables: north_america_views_percentage <dbl>,
## #   south_america_views_percentage <dbl>, africa_views_percentage <dbl>,
## #   antarctica_views_percentage <dbl>
```

The continent columns currently shows the split percentage of viewers from each continent for each video, i.e. the data only shows which proportion of viewers for a particular video come from a particular continent. The data does not show how the absolute number of students and relative viewing rates change within a continents.

To see how the viewing rates changes within each continent throughout the course, we will need to transform the values from these columns. We will calculate the absolute number of viewings from each continent for each video.

3.1.2 Construct data

In order to calculate the absolute number of viewings from each continent for each video (i.e. each row), a function (**col_3_summary**) was developed to take the absolute number of learners from the '*Total Viewed*' column, divided by 100, then multiply by the current percentage viewed value from each continent. An example is the calculation of absolute number of views of the first video from African viewers, which is calculated as (*Total Viewed*[row 2]/100) x *africa_views_percentage*[row 2]. The resulting data frame is stored in the *'cache'* folder. Table xy [ADD TABLE REF] shows the absolute number of viewers from each continent for run 3. The data in all remaining runs has been treated in the same way to be included in the second cycle of analysis.

```
## # A tibble: 3 x 7
##   europe_views oceania_views asia_views north_america_views south_america_views
##          <dbl>         <dbl>      <dbl>               <dbl>               <dbl>
## 1         915.          38.0      267.                193.                 50.9
## 2         595.          26.0       93.0               103.                 23.0
## 3         479.          23.0       71.0                77.0                16.0
## # ... with 2 more variables: africa_views <dbl>, antarctica_views <dbl>
```

The resulting data frame then undergoes another transformation to calculate the relative number of views (percentage) for each abolute continent value. For example, for the Europe views, we will take the absolute value from the Europe column and divide it by the highest absolute value from within Europe, then multiply by 100 to calculate the percentage of viewers from within the continent for each video. The resulting percentage data frame is stored in the *'cache'*

3.1.3 Clean data

These values were then combined with the original '*video_duration*' column and re-organised so that the columns match the original data set. [CHECK THE FOLLWOING STATEMENT FOR ACCURACY:] Values are given with no decimal places to increase the readability of the results. Since the modelling and evaluation in this project is mainly based on visualisations and qualitative assessments, the accuracy of the data analysis is not affected by this decision. The calculation of the percentages has resulted in the values from the *'antarctica_views_percentage'* being '*NaN*.' Since we know that there were no viewers from Antarctica, all '*NaN*' values have been replaced by 0. Table xy [ADD TABLE REF] shows the relative number of viewers in percentage from each continent for run 3. The data in all remaining runs has been treated in the same way to be included in the second cycle of analysis.

```
## # A tibble: 3 x 8
##   video_duration europe_views_percentage oceania_views_perce~ asia_views_percen~
##            <dbl>                   <dbl>                <dbl>              <dbl>
## 1             99                     100                  100                100
## 2            362                      65                   69                 35
## 3            241                      52                   61                 27
## # ... with 4 more variables: north_america_views_percentage <dbl>,
## #   south_america_views_percentage <dbl>, africa_views_percentage <dbl>,
## #   antarctica_views_percentage <dbl>
```

3.1.4 Integrate data

Prior to calculating the percentage of viewers within each continent, to sum of the absolute actual viewers for all runs for it to be added together, thus ensuring that the average is weighted according to the absolute number of learners within each run.

3.1.5 Format data

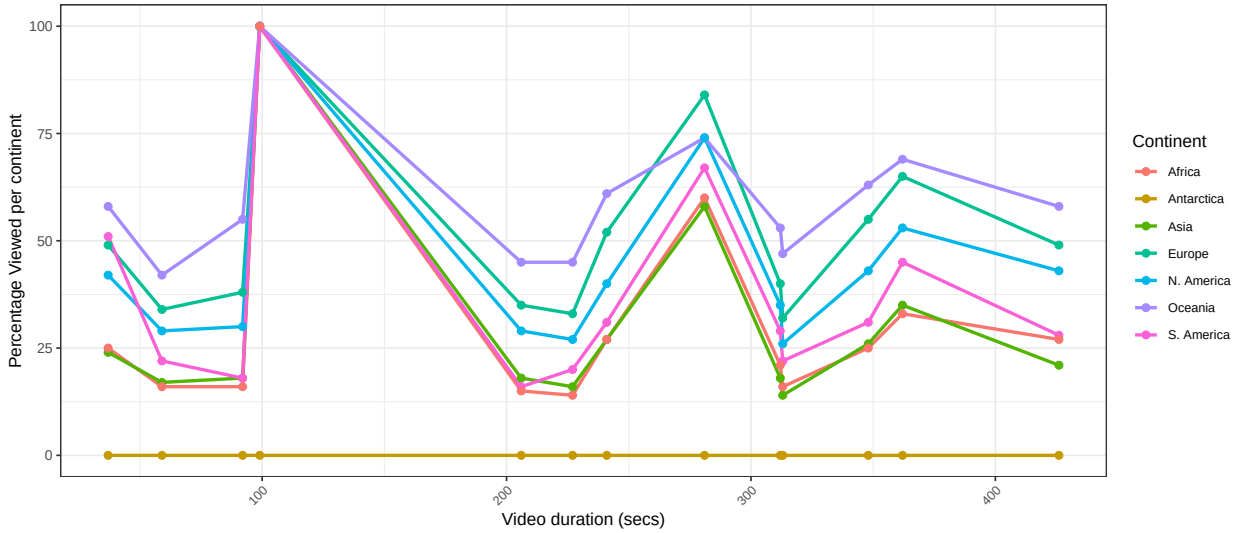
Prior to plotting the data for each cycle, the resulting dataframes are transformed to a 'long' format, i.e., a column for every variable, and a row for every observation (www.datacarpentry.org), to enable the use of ggplot2 to plot the data (see Section 3.1.6 *'Modelling'*).

3.1.6 Modelling

[CLARIFY THIS SENTENCE: DO YOU MEAN THE NUMBER OF VIEWS WITHIN EACH CONTINENT, OR THE NUMBER OF VIEWS THAT ALLOWS YOU TO COMPARE WITHIN A VIDEO?]

Figure xy [CHECK REF] shows a plot of the relative number of views as calculated in *Section 3.1.2 Construct data*, against the video duration, i.e. the data that allows to draw comparisons of the behaviour within a continent. This plot shows whether the viewers of a particular continent show an increased or decreased drop-out rate compared to the viewers of other continents. As described in *Section 3.1.2 Construct data*, all data is normalised to the most viewed video (Welcome to the course, 99 seconds duration).

Figure XX: Percentage of viewers for Video duration



The percentage of viewers from all continents fluctuate and appears to be independent of the duration of the videos. However it appears as if viewing rates from Africa, Asia and South America show a steeper decline from the *Welcome* video to all other videos in compared to the other continents. Viewing rates from Oceania show the smallest decline. Due to potential noise on the data, the same plot is plotted using the data for all runs (Figure xy [CHECk REF])

Figure XX: Percentage of viewers for Video duration

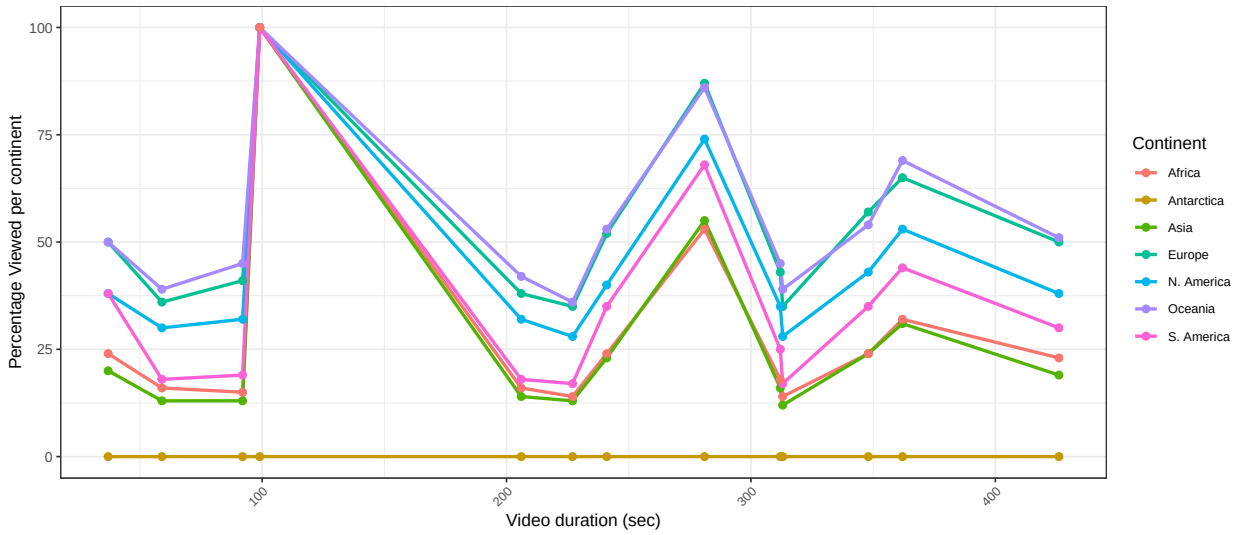


Figure XX shows the same pattern as the data of run 3 (Figure xy [CHECK REF]), in which there is no correlation between the duration of the videos and the number of viewers from each continents. However, it appears as if the steepest decline in viewing numbers can be observed for learners from Africa and Asia. South American viewing rates seemed to show a steeper decline for videos below approx. 220 seconds, whereas they increase compared to the other continents for longer video durations.

In summary, there is no clear correlation between the relative number of views within each continent and the video duration.

Again the graph does show that there are a higher percentage of viewers from Europe and Oceania for most videos than from Asia and Africa.??? Therefore the duration of the videos appears to not have a direct impact on keeping learners' engaged or keeping the learners to stay throughout the video (??)

We will investigate on whether there are correlations between the topics of the video and the number of viewers from each continent in Section 3.2.

3.2 Data Preparation and Modelling for Q2: Does the content of the videos have an impact on the viewing rates across different continents?

To answer the second question, we will investigate whether the topics of the videos have an impact on the viewing rates across different continents. This could indicate whether the learners from certain continents engage more with certain topics from the course or whether the learners find certain topics less interesting.

3.2.1 Select Data

For the same reasons as explained in Section 3.1.1, the continent columns within the original *'video_stats'* data sets show the percentage of viewers from different continents for each video. To answer question 2, we want to firstly plot a bar chart using the absolute number of viewers from each continent. As explained in Section 3.1.1, we calculated the absolute number of learners by dividing the '*Total Viewed*' column by 100 and multiply by the percentage viewed value for each continent. For the first cycle of analysis, we will use the data from run 3 only. For the second cycle of analysis, we will combine data from runs 3-7.

3.2.2 Construct Data

The columns of the new data set will be renamed to the corresponding continents. This will then be combined with the *'step_title'* column, which as discussed in *'2.3 Explore data'*, consist of combining the

two first columns *'step_position'* and *'title'* from the *'video_stats'* data set,to allow quick reference to the order of which the videos appears throughout the course. The data set will be re-organised so that the columns match the original data set. Table xy [ADD TABLE REF] shows the absolute number of viewers from each continent for run 3. The data in all remaining runs has been treated in the same way to be included in the second cycle of analysis.

```
## # A tibble: 3 x 8
##   step_title               europe_views oceania_views asia_views n_america_views
##   <chr>                           <dbl>         <dbl>      <dbl>           <dbl>
## 1 1.1_Welcome to the cour~          915            38        267             193
## 2 1.14_Why would anyone w~          595            26         93             103
## 3 1.17_Preserving privacy~          479            23         71              77
## # ... with 3 more variables: s_america_views <dbl>, africa_views <dbl>,
## #   antarctica_views <dbl>
```

Secondly, we will plot a line graph to show how the relative number of views changes throughout the course. As mentioned in *Section '3.1.2 Construct Data'*, we will take the data set with the absolute number of viewers, then calculate the relative number of view within each continent relative to the highest number of views across all videos within that continent. The resulting percentage data frame is stored in the *'cache'*.

3.2.3 Clean data

As described in Section *'3.1.3 Clean Data'*, the new dataframe of percentage values will be combined with the '*step_title*' column and re-organised so that the columns match the original data set. Values are given with no decimal places to increase the readability of the results as explained in Section *'3.1.3 Clean Data'*.

The calculation of the percentages has resulted in the values from the *'antarctica_views_percentage'* being '*NaN*.' Since we know that there were no viewers from Antarctica, all '*NaN*' values have been replaced by 0. The resulting percentage data frame is stored in the *'cache'*

```
as_tibble(continent_3_pct_ro[1:3,c(1,3:9)])
```

```
## # A tibble: 3 x 8
##   step_title europe_views_pe~ oceania_views_p~ asia_views_perc~ north_america_v~
##   <chr>                 <dbl>            <dbl>            <dbl>            <dbl>
## 1 1.1_Welco~              100              100              100              100
## 2 1.14_Why ~               65               69               35               53
## 3 1.17_Pres~               52               61               27               40
## # ... with 3 more variables: south_america_views_percentage <dbl>,
## #   africa_views_percentage <dbl>, antarctica_views_percentage <dbl>
```

3.2.4 Integrate data

Prior to calculating the percentage of viewers within each continent, to sum of the absolute actual viewers for all runs for it to be added together, thus ensuring that the average is weighted according to the absolute number of learners within each run.

3.2.5 Format data

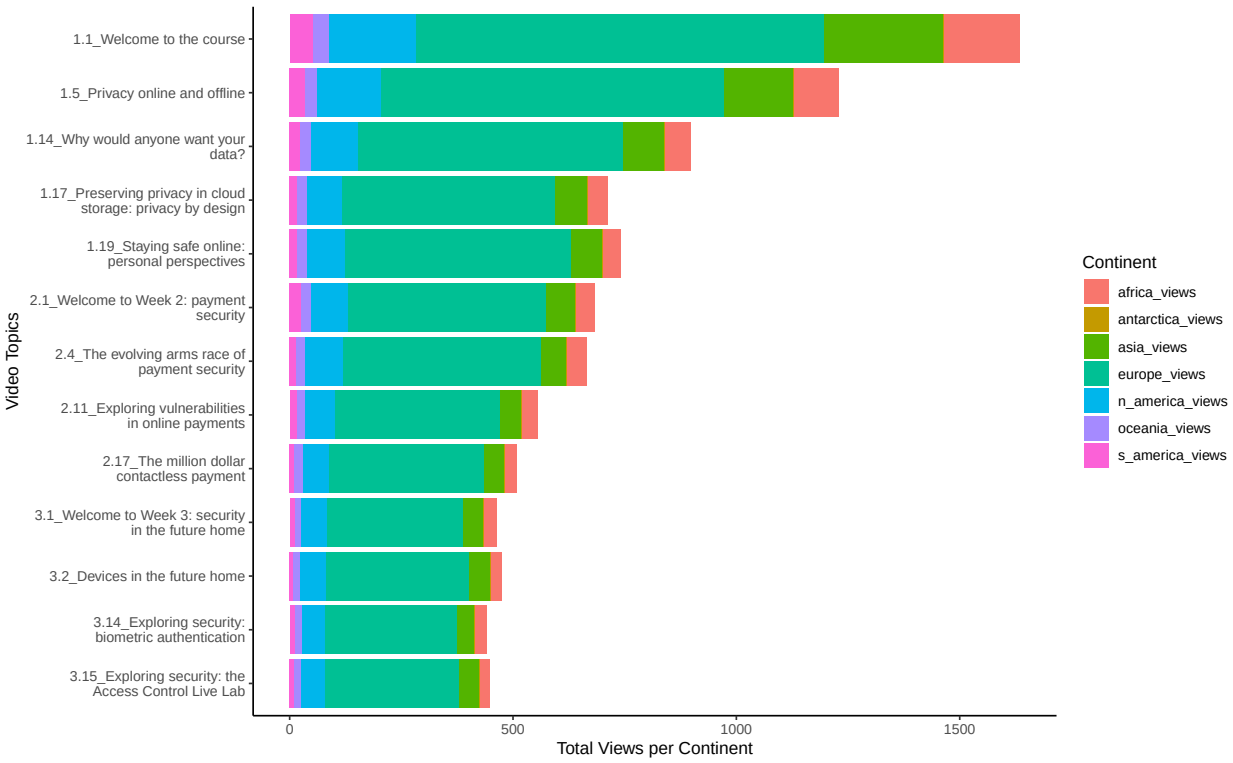
Prior to plotting the data for each cycle, the resulting dataframes are transformed to a 'long' format, i.e., a column for every variable, and a row for every observation (www.datacarpentry.org), to enable the use of ggplot2 to plot the data (see Section 3.1.6 *'Modelling'*).

3.2.6 Modelling

The video topics will be reordered to the same order as how they run throughout the course for the plots. This will show how the number of viewers have changed throughout the course.

Figure XX is a bar chart to show the relationship between the video topics and the number of viewers from across different continents for run 3. The colours in each bar are split according to the continent of the viewers, as shown in the key.

Figure XX: Total continental viewers on video topics

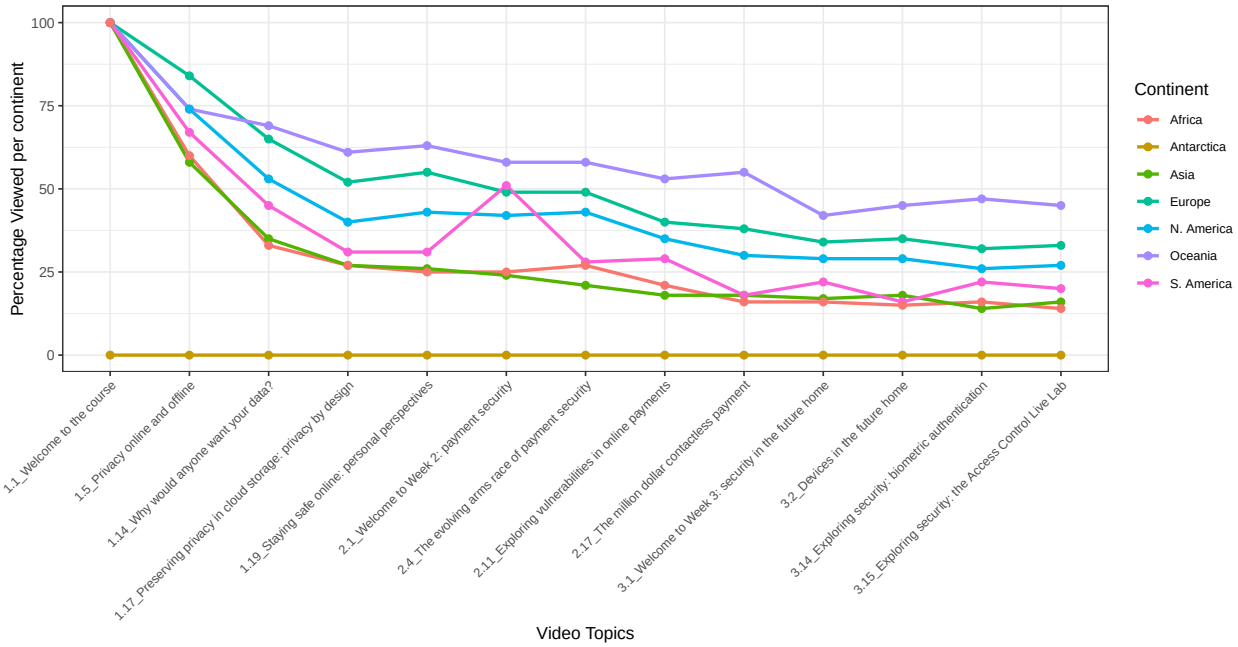


There are clear observations that the number of viewers decrease throughout the course. The largest drop in viewers occur throughout week one of the course, of which the number of viewers fell from 1659 to just over 700 viewers by the end of the first block, which is roughly 54.5% drop in viewers within a week. The drop from week two to week three is not as severe, from 694 to just under 450 viewers by the last video of the course, which is roughly 34.4% drop throughout week two to week three.

There are three videos, in which the number of viewers slightly increased compared to the previous video, for example for video *'1.19_Staying safe online: personal perspectives,' '3.2_Devices in the future house'* and *'3.15_Exploring security: the Access Control Live Lab'*.

The same data, but with the relative number of viewers from each continent, were plotted on a line chart in Figure xy [CHECK REF]. Each line represents a different continent.

Figure XX: Percentage of viewers for Video Topics

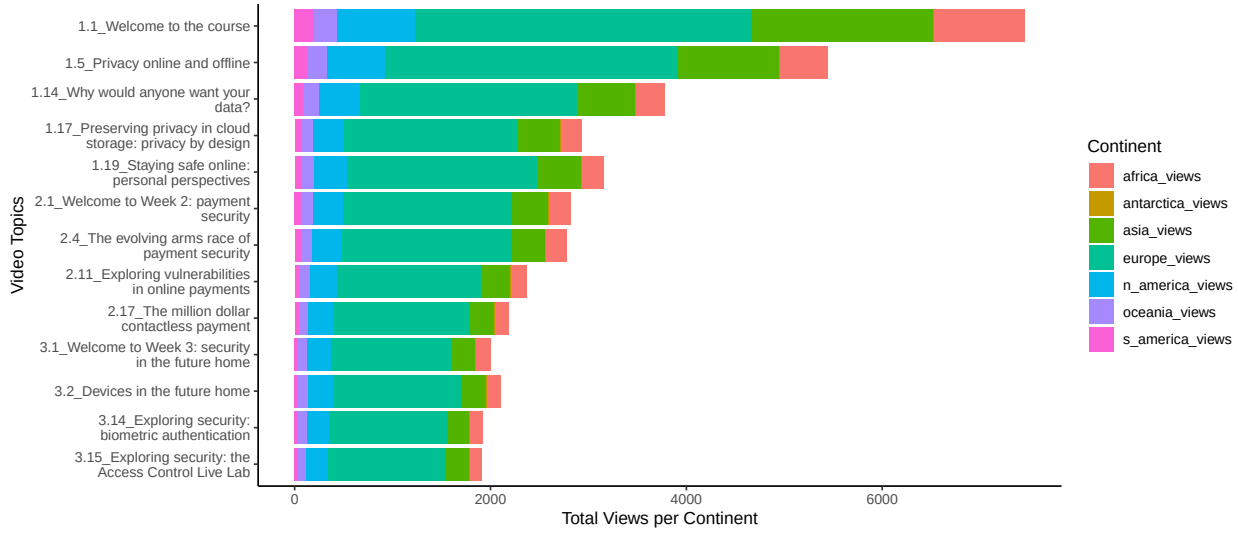


The line chart shows a steady drop of viewers throughout the course. More learners from Oceania have continued in the course and watched the videos compared to Europe, which was already seen in Section 3.1.6, Figure xy [FIRST PLOT IN THAT SECTION]. Asia and Africa show the lowest relative number of viewers.

There is an outlier from South American viewers on video *'2.1_Welcome to Week 2: payment security'*. This data point is showing a sudden increase of viewers from South America for this particular topic in the second block of the course.

There is likely to be noise on the data which obscures trends. For the second cycle of analysis, we will plot these graphs again using the data from all of the runs combined. The result is shown in Figure xy [ADD REF].

Figure XX: Total continental viewers on video topics



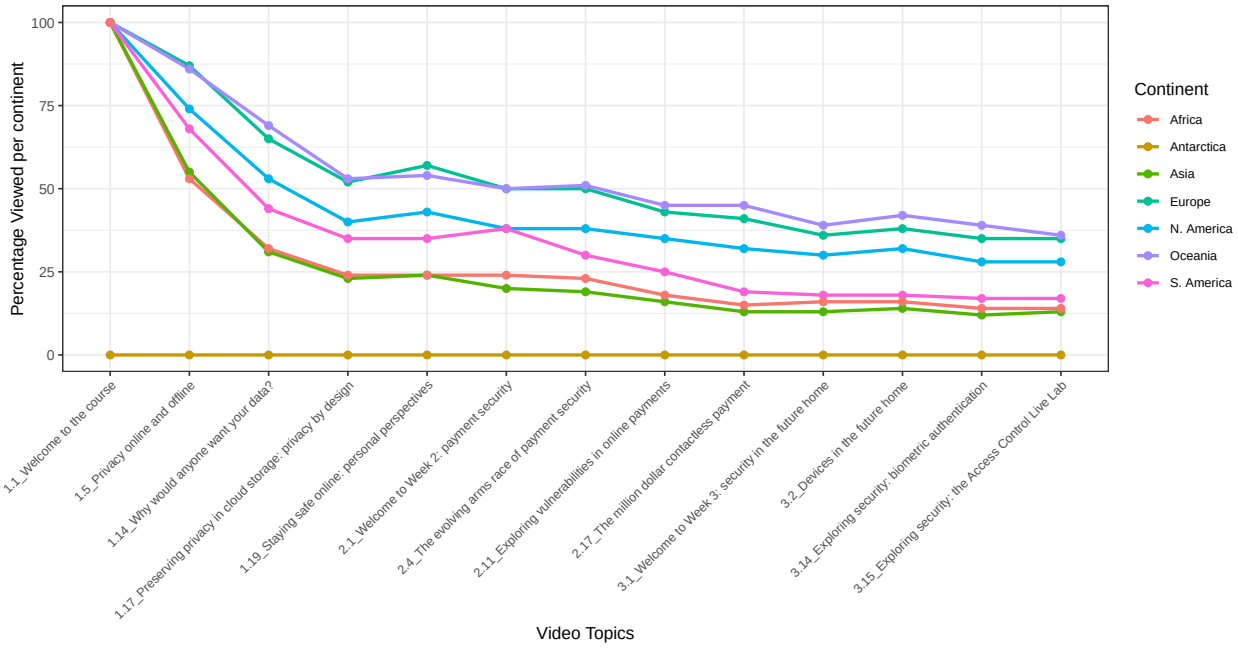
From the initial observation of the graph, the trend of the viewers from all runs combined, follow the same pattern as for run 3 viewers, with the largest drop in viewers occuring throughout week one of the course. The data shows that the drop in viewers throughout week one is very similar to run 3, from 7458 (at the

start of week one) to 3154 (end of week one), which equates to 57.7% in week one alone. Then a smaller gradual decrease of viewers between week two to week three, from 2816 to 1910, which is 32.2%, can be observed. Again, this is a very similar drop as was seen for run 3.

The combination of data for all runs shows that there are two videos in which the number of viewers slightly increased compared to the previous video, these are *'1.19_Staying safe online: personal perspectives'* and *'3.2_Devices in the future house'*.

Just as for run 3, the percentage viewers from each continent of all the runs were plotted on a line chart in Figure xy [ADD REF]. Each line represents a different continent as shown in the key.

Figure XX: Total percentage of viewers for Video Topics



The line chart with the data from all runs still shows a steady drop of viewers throughout the course. However, the gap between the percentage of learners from Oceania and Europe are a lot closer compared to what was seen in run 3. The relative numbers of views from Oceania and Europe decrease at a similar rate. The trend that the steepest decline in the relative number of viewers is seen for viewers from Asia and Africa is more pronounced when all data sets are combined.

The outlier on video *'2.1_Welcome to Week 2: payment security'* from South America that has been observed for run 3 is visible in Figure xy [ADD REF] too. The outlier is showing a smaller increase of viewers from South America for this particular topic in the second block of the course, with all the runs combined. It is unclear whether this is statistically relevant or due to noise.

3.3 Data Preparation and Modelling for Q3: Is there a correlation between duration of videos and drop out rate of the learners?

The last question is around whether there are any correlations between the duration of the videos and drop out rate of the learners while they are watching the video, as we have seen in Section 2.3.3, Figure xy [ADD REF] using the run 3 data. We want to understand whether the length of the videos will have an impact on how much of the entire video the learners will watch, e.g., how many learners have viewed 25%, 50%, 75% or 100% of the videos. This could potentially indicate whether the course provider should consider optimising the length of the videos to keep learners engaged.

3.3.1 Select data

The key columns for this investigation will be *'video_duration,' 'total_views,' 'viewed_twentyfive_percentage,' 'viewed_fifty_percentage,' 'viewed_seventyfive_percentage,' and 'viewed_onehundred_percentage'* from the *'video_stats'* data sets.

The range of 'viewed_percentage' columns shows the percentage of viewers who have viewed 25%, 50%, 75% and 100% of the videos.

The 5%, 10% and 95% range of 'viewed_percentage' columns have been excluded from this investigation because the selected columns provided an even split of data collection points. A prelimary plot of all data showed that there is no additional information in these columns, and their inclusions obscures the trends that can be observed.

3.3.2 Construct data

In order to combine the relative values within the 'viewed_percentage' columns from all of the runs, we will need to create a new data set that contains the absolute number of viewers. This will allow us to calculate the weighted average as explained in Section 3.1.4. The absolute number of views will be calculated using a similar function to question 1 and question 2, we will take the total number of learners from the '*Total Viewed*' column, divided by 100, then multiply by the relative number of views value from each cell of the 'viewed_percentage' columns. This will calculate the absolute number of viewers who have watched 25%, 50%, 75% and 100% of the videos.

To combine the data sets, the new data set containing the absolute number of views for each 'viewed_percentage' column in all runs will be added together. The resulting data frame is stored in the *'cache'*. The new dataset is then transformed to re-calculate the percentage for each of the 'viewed_percentage' column and then be combined with the *'video_duration'* column. The columns of this data set will be re-organised and renamed to the corresponding relative views (in percent) so that they match the original data set.

Values are given with no decimal places to increase the readability of the results as explained in Section *'3.1.3 Clean Data'*. The resulting percentage data frame is stored in the *'cache'*. We will use this data set to plot a line graph to show how the percentage of viewers have changed against the duration of the videos.

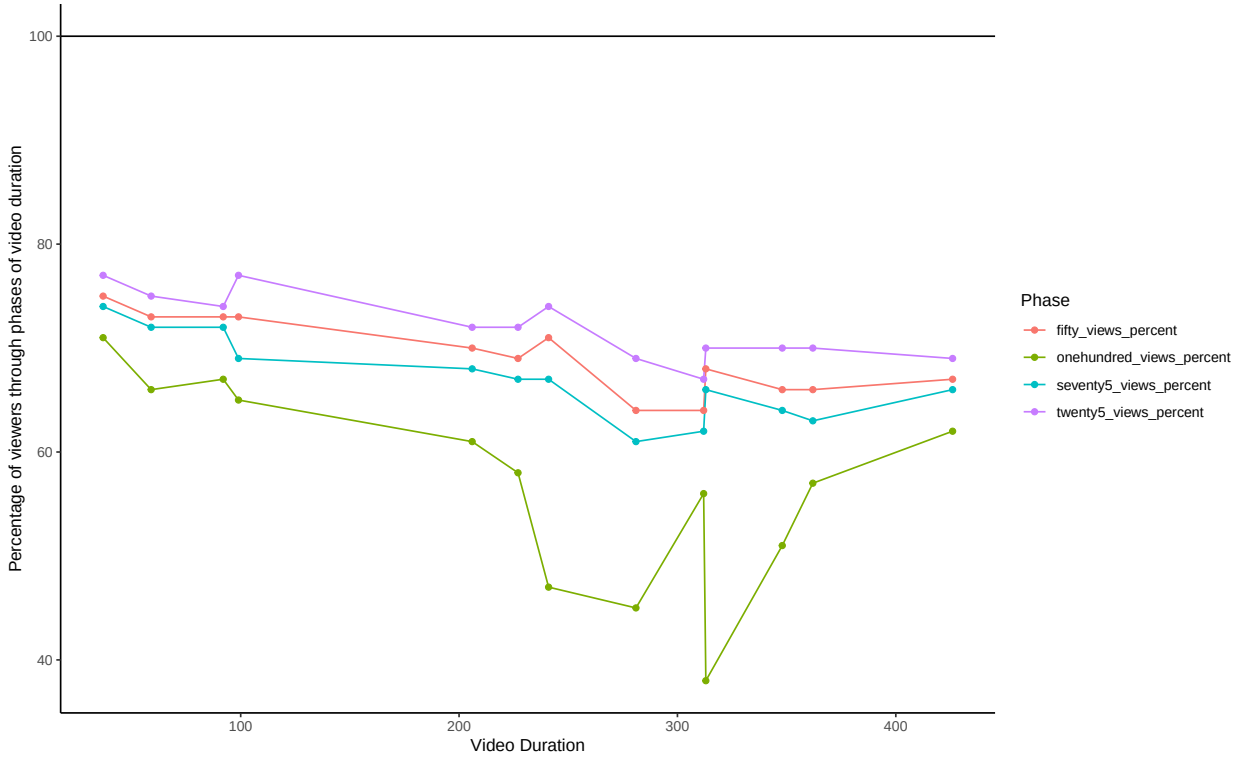
3.3.3 Format Data

The resulting dataframe is then transformed to a 'long' format to enable the use of ggplot2 to plot the data.

3.3.4 Modelling

Figure XX [ADD REF] demonstrates the relationship between the video duration and the number of views throughout the duration of the videos, at 25%, 50%, 75% and 100% of each video, for the data of all runs combined.

Figure XX: Number of viewers watched 100% of videos



This graph shows the drop off rate for all runs and it appears to show very similar behaviour of the learners drop off rate as was seen in run 3 (Section xy, Figure xy [ADD REFS]. It appears as if three videos have particularly high dropout rates between 75% and 100% of the video duration. However there does not seem to be any correlation to the video duration. More data would be required to establish whether the trend is real or due to noise. If it is real, it would be interesting to analyse the video content to find potential triggers for the drop-outs

9. Evaluation

9.1 Evaluate results - Assessment of data mining results with respect to business success criteria

The study was to understand the potential factors that could affect the learners' engagement to a course through the video material provided.

Question 1: The results shown appear to indicate that there are no correlations between the length of the videos and how engaged the learners from each continent were. The length of the video does not seem to be a deciding factor for learner's engagement and retention.

Question 2: The results clearly show that most learners will watch the first video of the course, and the number of learners viewing the videos throughout the course drop significantly within the first week. However between week two and three the drop off of learners is not as high. It is unclear whether the data provided by FutureLearn also collects the number of learners watching preview of videos for the course, or whether FutureLearn gives free access to their learning material for a short period of time, which could reflect on the sudden drop of learners watching the videos within the first week. WHAT WERE THE TOPICS OF THESE BLOCK WEEKS?

The line chart in Figure xy [ADD REF], also indicates that there is a clear interest for video 2.1 from learners

based in South America. The topic was *Welcome to Week 2: payment security.* Given more time, we could do further analysis to gain insights about the learners' from continents to understand whether there are other factors which could affect the learners' engagement for certain topics, there factors could include political factors of the continent, types of jobs and interest in cyber security within those continents.

Question 3: The results potentially show that the long a video is the higher risk of the learners' dropping off the videos earlier during the videos. As most learners drop off after 75% into a video, this could help course designers and providers to decide whether to ensure that important course content are placed earlier on the videos. Alternatively, it would be interesting to analyse the video contents in the last 25% of the video for the outliers as discussed in Section 3.3.4. This could help understand if there are any triggers for the drop-outs to inform the future design of the videos.

It would be interesting to see if the following would have an impact on the data: If the decision is to be made to make important content to appear towards the end of the videos, then a contents page could be provided in the video, to inform the learners what to expect throughout a video and this could potentially keep learners more engaged throughout the video. Depending on the interest of the content, it could affect the number of learners watching the entire length of the videos.

Summary: Based on the data and the analysis in this report, it appears that learners engagement with the course decreases throughout the course (and throughout the duration of the videos). Most learners appear to stop viewing the videos after the first week of the course. For the learners who are seriously interested in the course, they are likely to stay on for week two and week three, hence we see a smaller drop off rate to the course for those two weeks.

9.2 Review process

9.3 Determine next steps

Given more time, the data could be combined with the e.g. the _enrolments.csv data files to understand more about the types of learners from each continents, and if the data allows possibly split into separate countries. This could provide further insight if certain topics are of more interest for learners of certain continents and from different countries, which could inform the design of any future courses.

Did we find correlations to answer the business question??

Whats the future work?? May have to do some more analysis to compare with other MOOC courses

Week 2 block is mainly on cybersecurity of payment infrastructure. Could be people are more interested on how to protect digital payments or there might be more people looking to work or already working in the cyber security / financial sector and are keen to learn about these topics. More investigation will need to be made.

10. References

Bote-Lorenzo, Miguel L., and Eduardo Gómez-Sánchez. 2017. "Predicting the Decrease of Engagement Indicators in a MOOC." Journal Article, 143–47. https://doi.org/10.1145/3027385.3027387.

Chapman, Peter, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer, and Richard Wirth. 2000. "CRISP-DM 1.0: Step-by-Step Data Mining Guide." Report.

Shacklock, Xanthe. 2016. "From Bricks to Clicks - the Potential of Data and Analytics in Higher Education." Report. Higher Education Commission. https://www.policyconnect.org.uk/research/report-bricks-clicks-potential-data-and-analytics-higher-education#0.

Shi, Lei, and Alexandra I Cristea. 2018. "Demographic Indicators Influencing Learning Activities in MOOCs: Learning Analytics of FutureLearn Courses." Journal Article.