CSC8631 Report

Selina So

23/11/2021

Business Objectives

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, 2018)". Learning analytics 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 (Xanthe 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. Consequently 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 the insights driven from Learning Analytics, it will help FutureLearn identify areas which will improve the retention rate and learners' engagement.

Business objective

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, Gomez-Sanchez, 2017). In this study, the focus will be on the video lectures provider by FutureLearn, which are generally used to form part of a course.

This study will examine the Cyber Security online course, which is divided into three weekly blocks of study. The course consist of a combination of videos, articles, exercises, discussions, quizzes and tests.

There are a number of steps to complete for each weekly block. The first week block contains 18 steps, and the second and third week blocks contains 21 steps. (Shi, 2018)

Inventory of Resources

The CRISP-DM methodology (Cross-Industrie Standard Process for Data Mining) will be applied to achieve the objective of this study (link the CRISP-DM guide). The key phases of focus from the process are Business Understanding, Data Understanding, Data Preparation and Evaluation.

Data Mining Goals

For this study, we will investigate the videos 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?

to insert reference use this notation [@RN22] - this is not working, try later.

Data Understanding (Initial Obsevration)

The raw data was provided by FutureLearn on their Cyber Security course. There are seven runs of data, each run of data were measured several months apart from each other. There were no descriptions for the data, therefore assumptions will be made as to what the data means.

As the study is based on the use of video material, therefore the datafiles with the title containing '*video.stats*' would be used. There are only 5 (out of 7) runs, which contains the '*video.stats*' datafiles. Therefore runs 1 and 2 will be eliminated from this study as no data are available.

Below is the list of column names in the data.

```
names(run3unite)
```

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

There are 13 rows for each datafile, one row corresponding to each video content throughout the course.

There are 28 columns, and a combination of columns will be selected for particular analysis.

The dataset is mostly complete with no visible missing data.

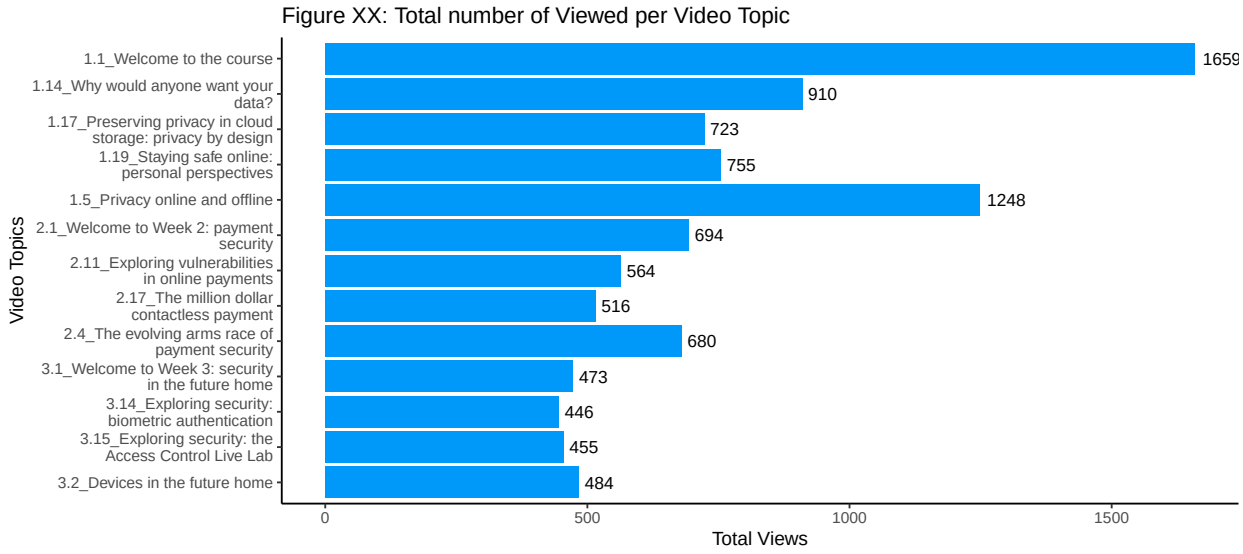
The data are mostly continuous data, other than for the first two columns, which shows the step of where the video content is located at and the title of the video content.These two columns will be combined to allow quick reference to the order of which the videos appears throughout the course.

As the study is interested in the number of views across the continent and the drop out, therefore the columns relating to viewing in HD and different devices will be removed. Other remaining columns will remain as they may contain relevant information for the study.

The following initial visualizations will help determine the areas to consider for further investigation.

Number of viewers for each video topic

Figure XX shows the number of viewers for each video based on the topic of the videos.



Finding pairs of relationships

We considered just the number of learners watching each video for the whole duration of the video, using the earliest video dataset (based on other datasets of run 3, this appears to roughly cover the time period between Jul 2017- Nov 2017).

The scatterplot matrices will be used to visualize any pairs of relationships of all of the different variables within the data.

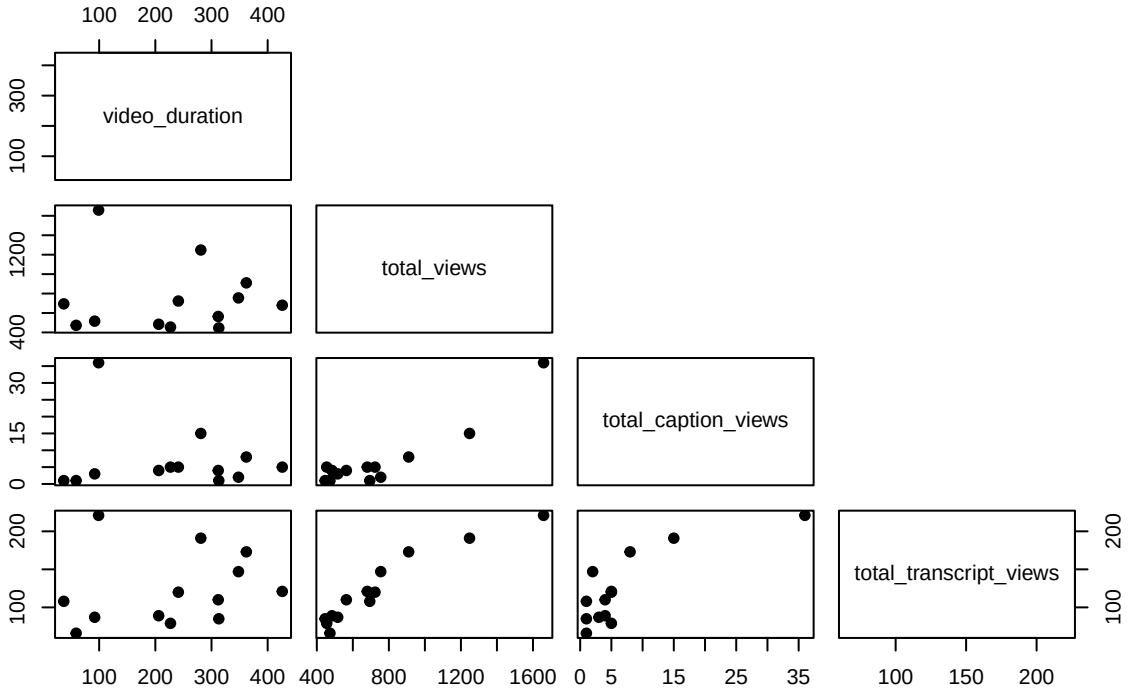
The plots below the headings, the headings will be the x-axis, and the corresponding rows will be the y-axis.

Question 3 Is there a correlation between duration of videos and drop out rate of the learners?

Total number of viewing and features used through the duration of videos

Figure XX scatterplot matrix attempts to demonstrate the relationships between the length of the videos and the amount of views and features (for example, downloads/ captions/ transcripts) used 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.

Figure XX: Total Views and Features used through Video duration



There are no obvious relationship observed between the length of the videos and the amount of views and features used for the videos. However as the total number of views increases, so does the number of downloads, captions used and transcripts used, which is to be expected.

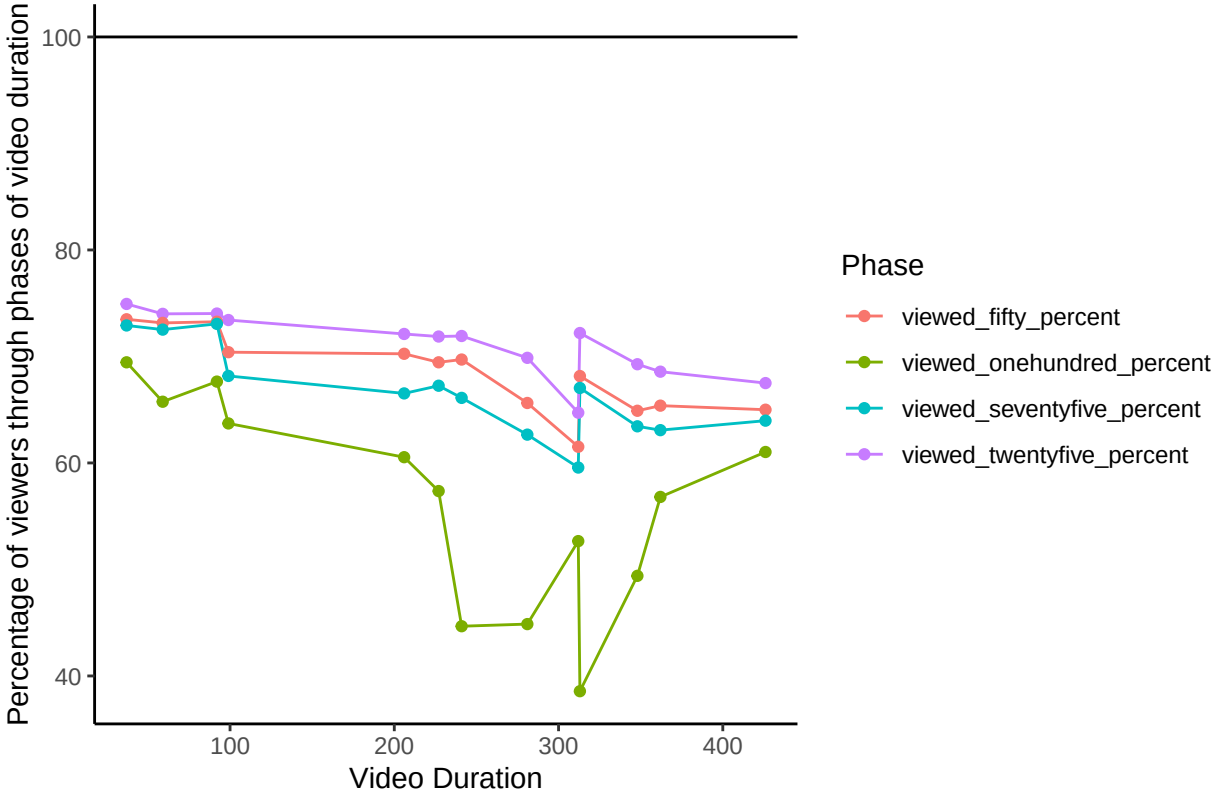
Percentage viewed whole duration of videos/ Number of viewers at 100 percent duration of the videos

Figure XX attempts to demonstrates the relationship between the video duration and the number of views throughout the duration of the videos, at 5/10/25/50/75/95/100 percent of each videos.

Through observation of the matrix, there are potentially interesting patterns on the far left column of plots. However the plots are very noisy and will require more data to make further statements, therefore further investigation will be required. The other columns do not show unexpected behaviour.

Figure XX shows the number of viewers who have stayed to view the videos to the end.

Figure XX: Number of viewers watched 100% of videos



There is potentially a pattern observed on the relationship between the video duration and the number of viewers who have stayed for the whole duration of the videos. However all the runs will need to be included to enable any statements to be made.

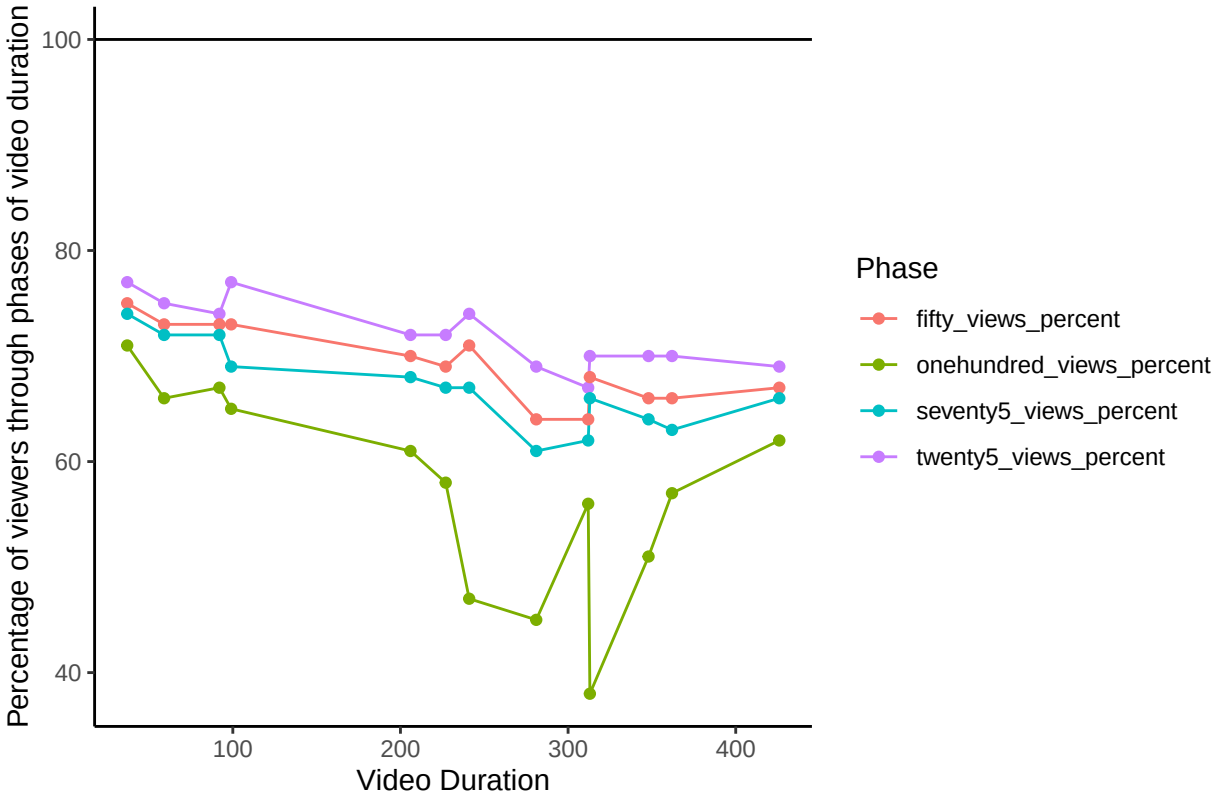
Data Preparation

the 'step_position' and the 'title' columns will be combined to allow quick reference to the order of the videos, which it appears throughout the course. This will also result in the 'step_position' data type being changed from numerical to character.

The data pre-processing codes are located in the 'munge' folder.

Percentage viewed whole duration of videos/ Number of viewers at 100 percent duration of the videos for all runs

Figure XX: Number of viewers watched 100% of videos



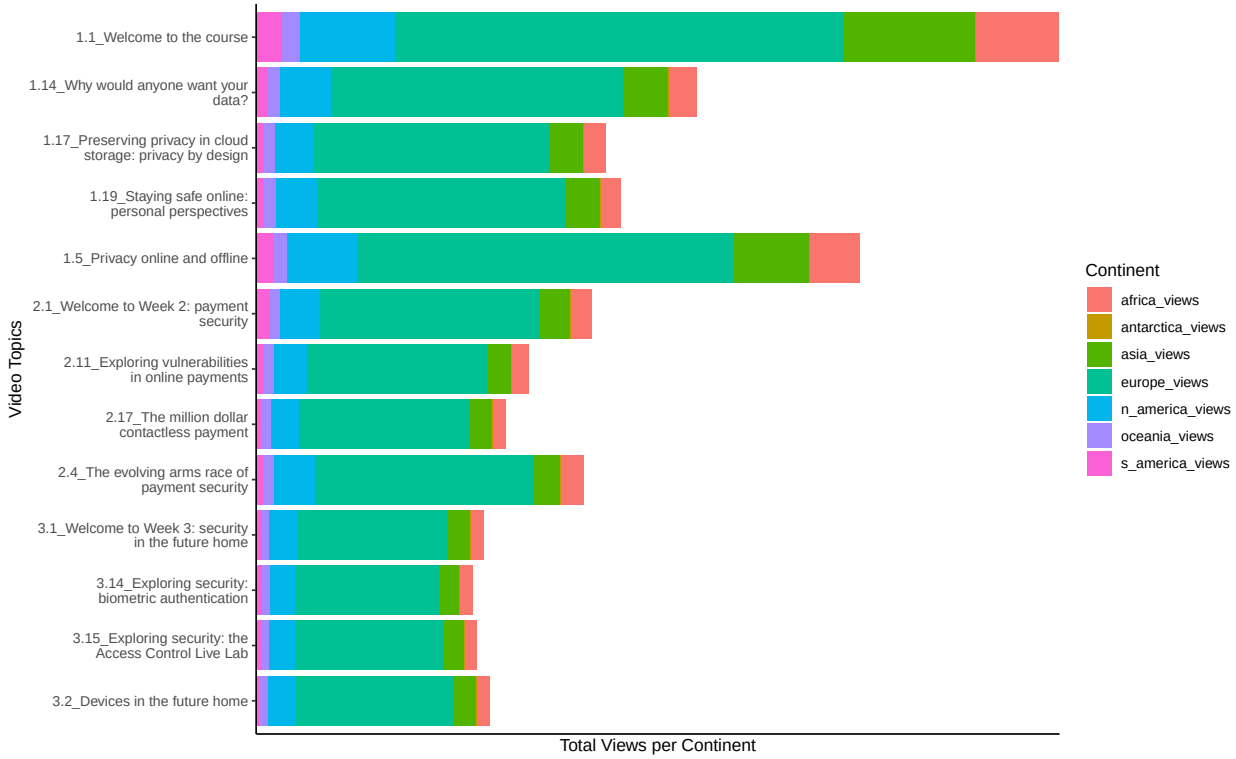
Question 2. Does the content of the videos have an impact on the viewing rates across different continents? using absolute values

Data understanding

Worldwide views of videos

Figure XX attempts to demonstrates the relationship between the video duration and the number of views from across different continents.

Figure XX: Total continental viewers on video topics



There are no obvious behaviours to be seen, therefore the data will be transformed to compare the views within each continent throughout the course rather than the viewings from each continent within each video.Therefore this could show the drop out rate from each continent throughout the course.

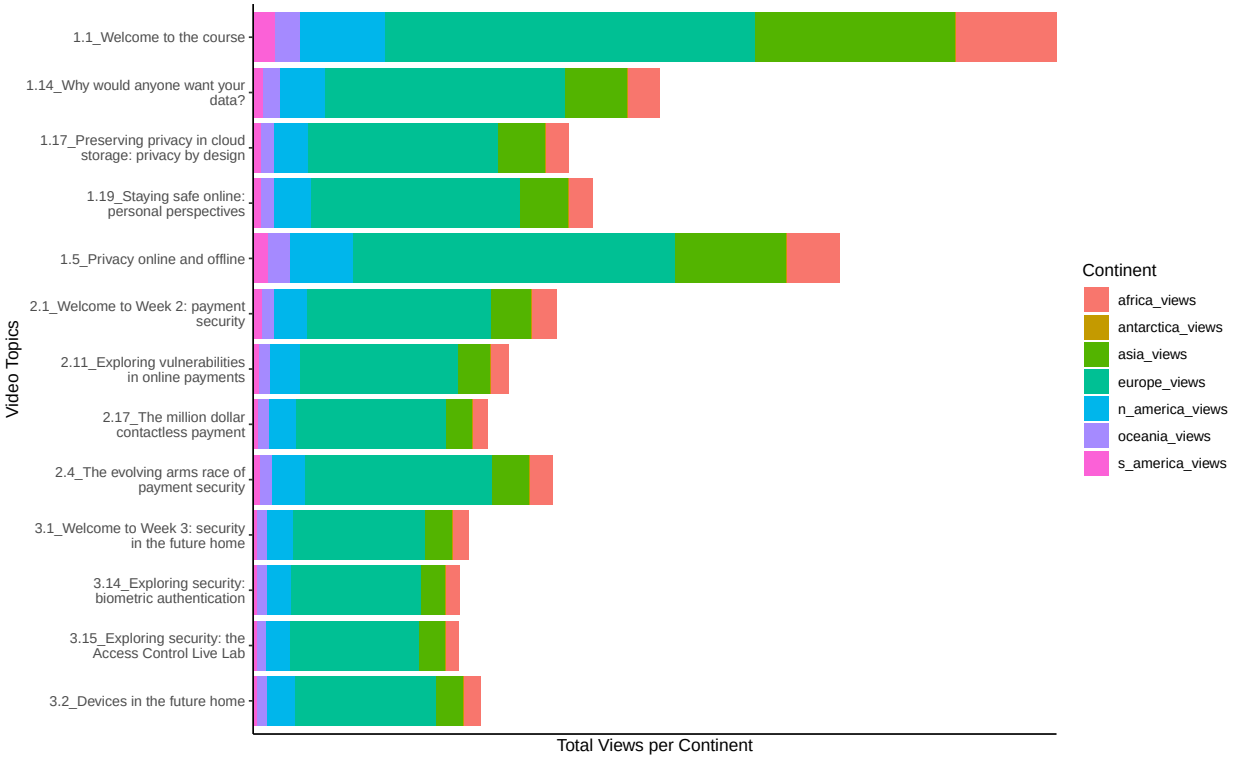
In addition, the relative views from each continent appears to be stable. There appears to be some outliers from the far left column of plots. It is unlikely related to the duration of the videos because the outliers appear random, therefore further investigation will be made on whether the video topics could be related to these outliers.

Final investigation - include all runs

Data Prepare

Worldwide views of videos using all runs as bars and absolute values

Figure XX: Total continental viewers on video topics



Question 2. Does the content of the videos have an impact on the viewing rates across different continents? Using percentage values

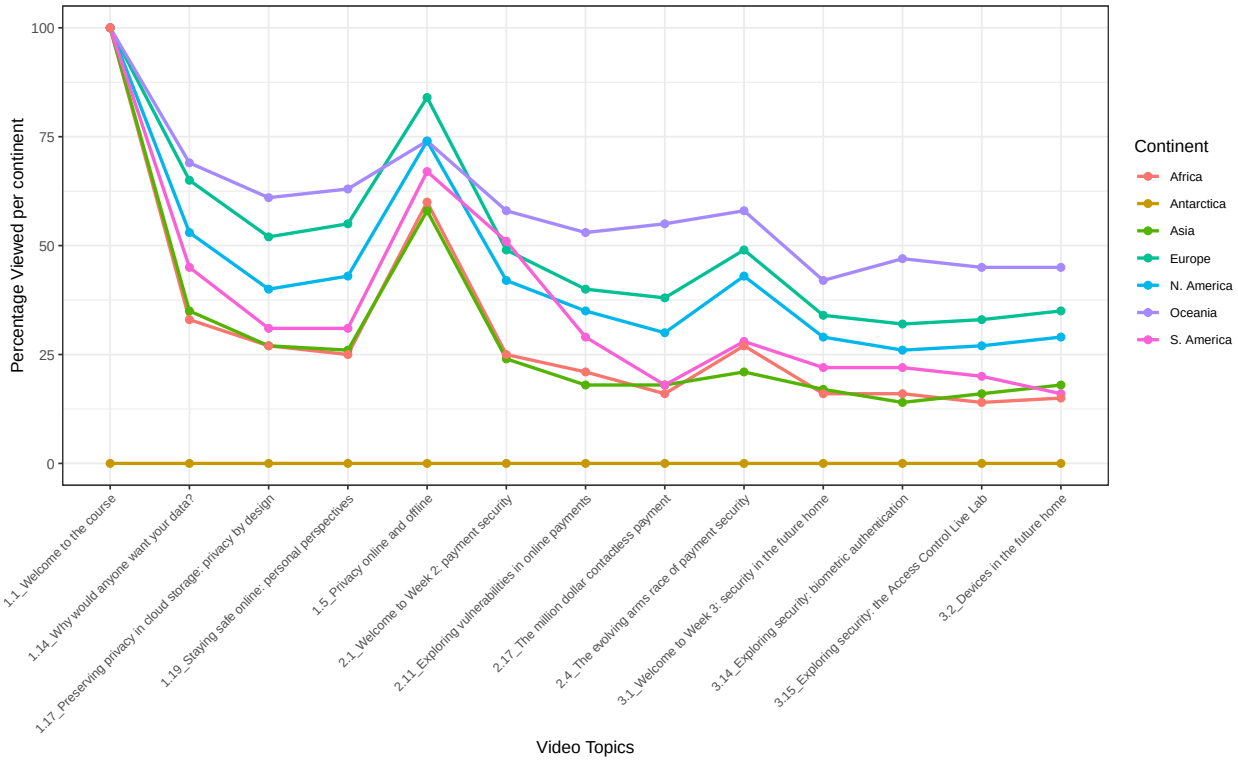
Actual views from continent

The actual viewings from each continent for each video is calculated. This is by taking the assumed total learners (by taking the highest figure from '*Total Viewed*' column), divided by 100, then multiply by the current percentage viewed value from each continent. This is the reverse percentage calculation for each continent (?)

The percentage viewed from each continent, throughout the course is calculated.Then combined with the original '*step_title*' and '*video_duration*' columns. Values are round to full numbers and '*NaN*' are replace with 0.

Line graph to show how the percentage viewers have changed throughout the duration of the course, based on how many viewers watched the videos throughout the course.

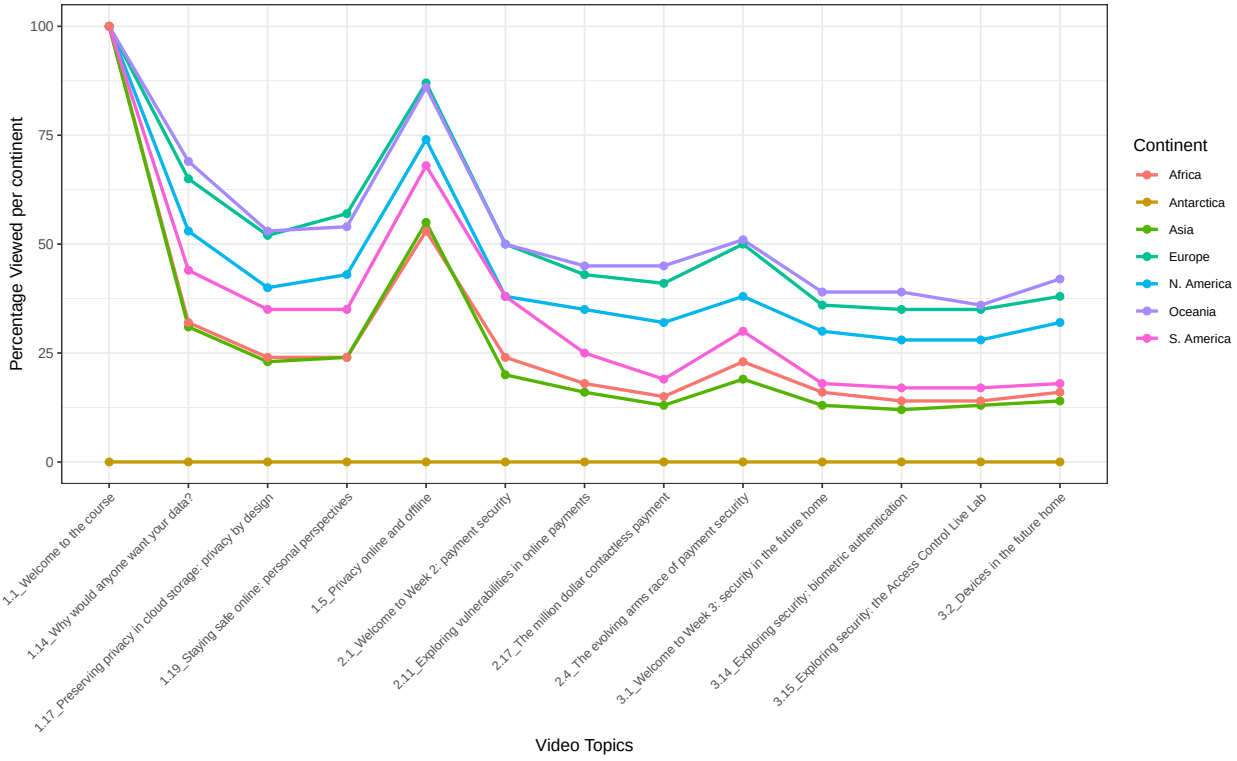
Figure XX: Percentage of viewers for Video Topics



The graph shows that for certain continents (Europe, North America and Oceania), there appears to be more views, therefore more engagements, on the topics within week 2 block of the course. On the other hand, there were more engagements from the learners of Africa and Asia continent, then a steady drop of viewers throughout the course. South America showed a dramatic uptake of viewers for the 1st topic of week 2, then a reduction of engagement throughout most of the course.

Actual views from continent using all runs and percentage

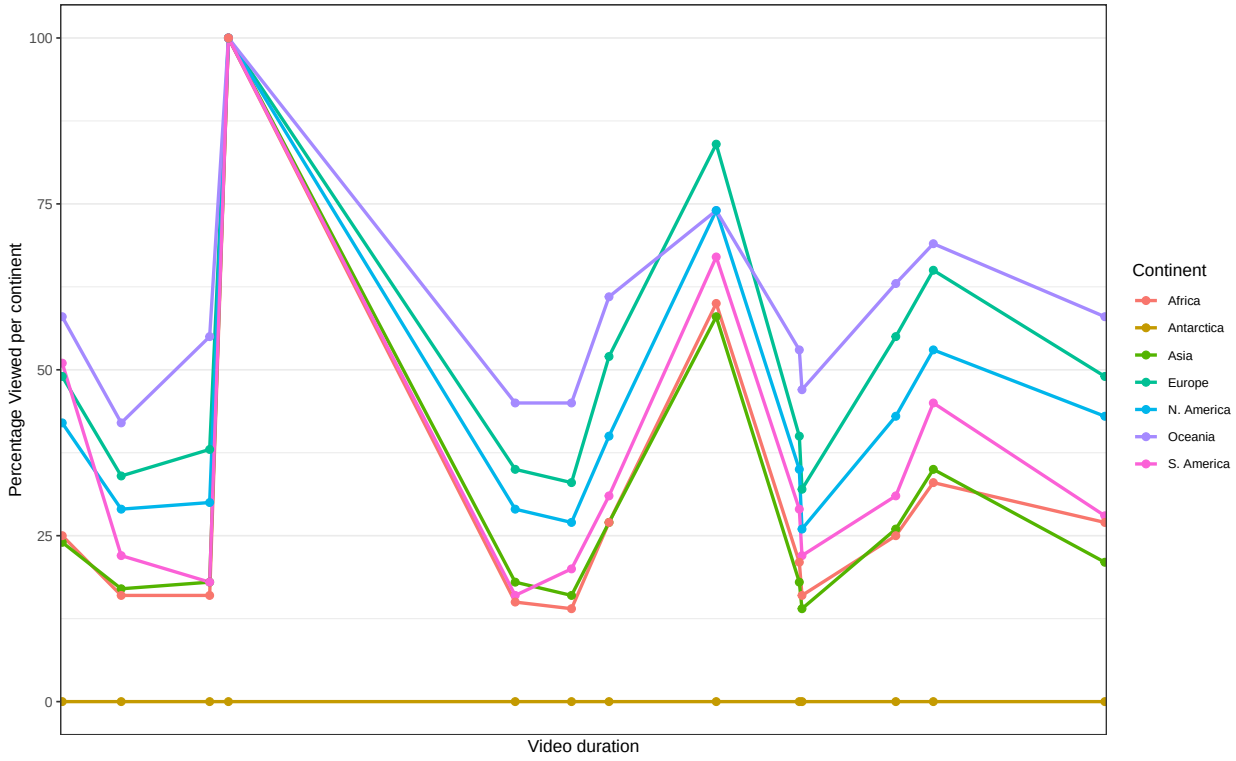
Figure XX: Total percentage of viewers for Video Topics



Question 1. Does the duration of the videos have an impact on the viewing rates across different continents?

Data understanding

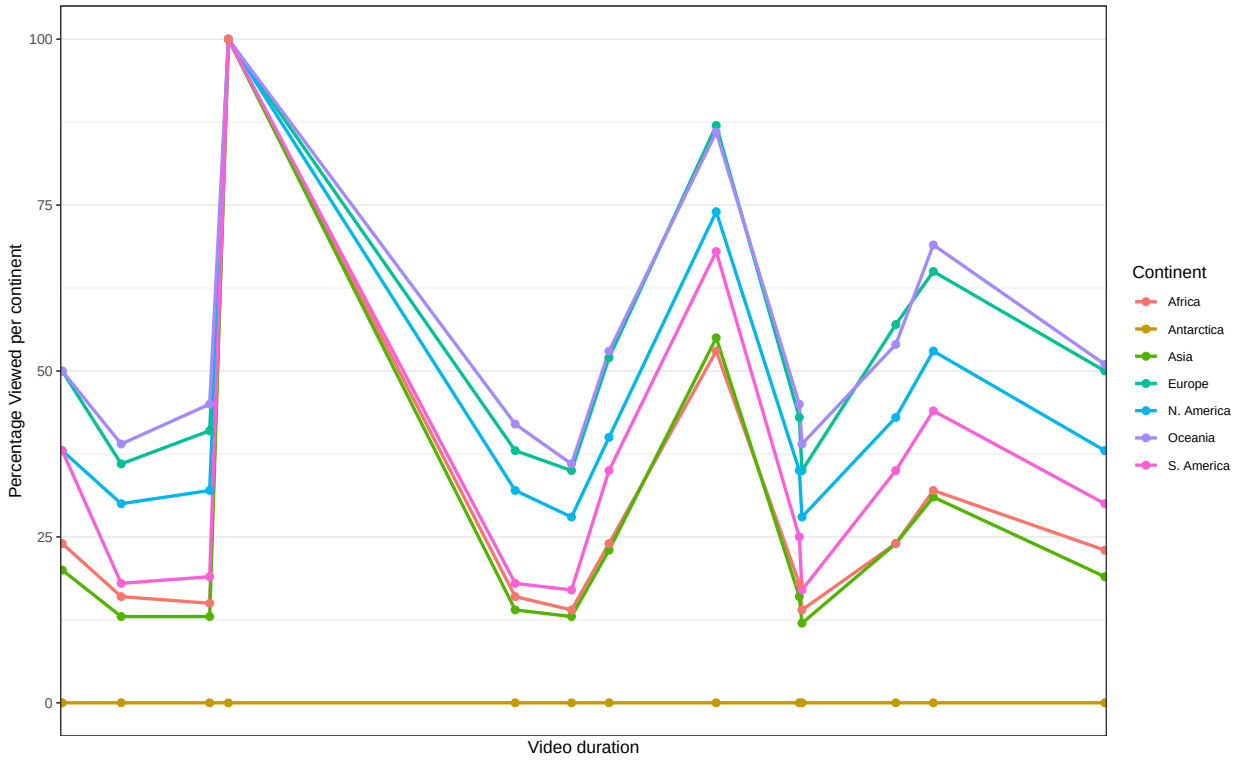
Figure XX: Percentage of viewers for Video duration



no correlation

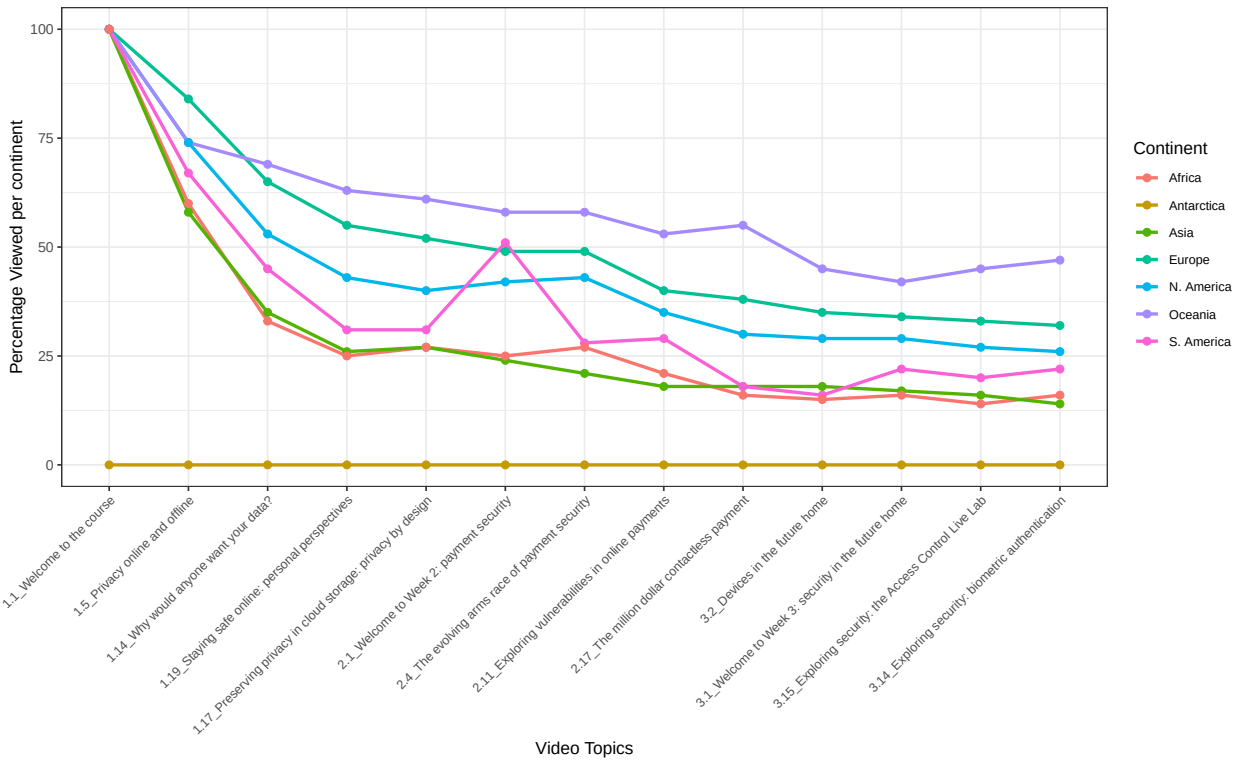
prove with all runs

Figure XX: Percentage of viewers for Video duration



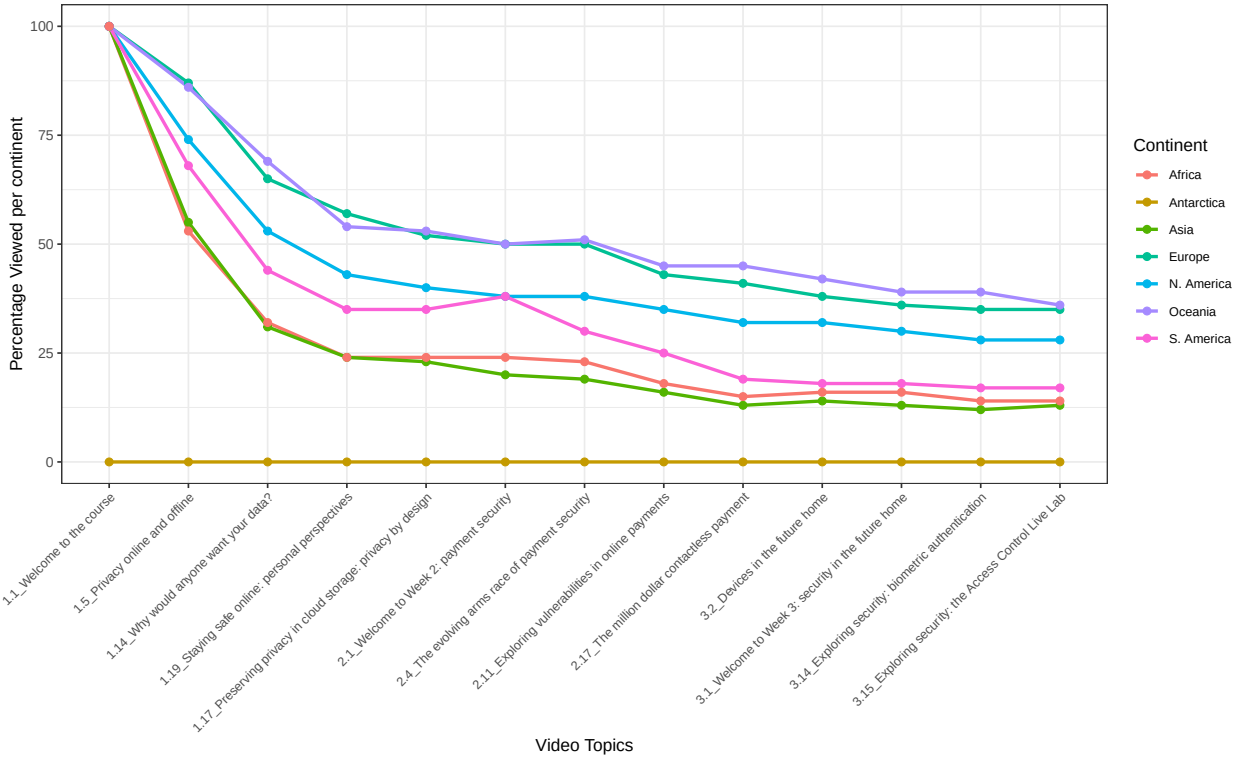
line drop out of continent vs topic

Figure XX: Total percentage of viewers from high to lowest in run 3



line drop out of continent vs topic for all runs

Figure XX: Total percentage of viewers from high to lowest in all runs



Evaluation

Recommendation

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.

References