CSC8631 Report

Selina So

23/11/2021

Business Objectives

Determine business objectives (background (record knowns about business situation), business obj (what does the customer want to accomplish, uncover factors that can influence outcome of project), business success criteria)

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)". 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 (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.

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).

This study will examine the data from the *Cyber Security Safety at Home, Online, in Life* online course, which is a course delivered by Newcastle University on the FutureLearn platform (https://www.futurelearn.com/courses/cyber-security). The course consist of a combination of videos, articles, exercises, discussions, quizzes and tests. As data from videos could act as engagement indicators, Newcastle University and FutureLearn could therefore utilize this data to aid the quality improvement of the course. To achieve this, the focus will be on the video lecture data provided by FutureLearn, which are generally used to form part of a course.

The *Cyber Security Safety at Home, Online, in Life* 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, 2018, futurelearn site)

Business success criteria (describe criteria for successful outcome to project from business view, This might be quite specific and able to be measured objectively, for example, reduction of customer churn to a certain level, or it might be general and subjective, such as "give useful insights into the relationships." In the latter case, it should be indicated who makes the subjective judgment. include???)

Insights into the relationship of engagements across different continents. Is the course able to reach a broad range of learners from across the world, from the material provided in the course.

assess situation (e resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and project plan.??)

Inventory of Resources (List the resources available to the project)

- 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 (maybe add Modelling)
- The key personnel that will be utilize for expert knowledge and as stakeholders for the study will be the University educators and teaching assistants.
- The *Cyber Security Safety at Home, Online, in Life* course data is provided by FutureLearn. There are 53 csv data files and 7 pdf files which provides an overview of the course structure.
- hardware platforms (computing resources)
- The softwares used will be R, RStudio and ggplot2.

Requirements, assumptions, and constraints (legal issues. As part of this output, make sure that you are allowed to use the data. List the assumptions made by the project. These may be assumptions about the data that can be verified during data mining, but may also include non-verifiable assumptions about the business related to the project. It is particularly important to list the latter if it will affect the validity of the results. ???? List the constraints on the project. dataset constraints)

Data Mining Goals (A data mining goal might be "Predict how many widgets a customer will buy, given their purchases over thepast three years, demographic information (age, salary, city, etc.), and the price of the item." Describe the intended outputs of the project that enable the achievement of the business objectives. Define the criteria for a successful outcome to the project in technical terms—for example, a certain level of predictive accuracy or a propensity-to-purchase profile with a given degree of "lift."???)

For this study, we will investigate the data and initially decide on a set of data to analyse in more detail to understand students's 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/amount of data that is available. Based on this, the most promising lines of investigation will be decided.

Project plan (Describe the intended plan for achieving the data mining goals and thereby achieving the business goals. The plan should specify the steps to be performed during the rest of the project, including the initial selection of tools and techniques.????)

Initial assessment of tools and techniques (At the end of the first phase, an initial assessment of tools and techniques should be performed. assess tools and techniques early in the process since the selection of tools and techniques may influence the entire project.???)

Data Understanding (Initial Obsevration)

collect initial data (n includes data loading, if necessary for data understanding. For example, if you use a specific tool for data understanding, it makes perfect sense to load your data into this tool. This effort possibly leads to initial data preparation steps.Note: if you acquire multiple data sources, integration is an additional issue, either here or in the later data preparation phase.?? Initial data collection report (List the dataset(s) acquired, together with their locations, the methods used to acquire them, and any problems encountered. Record problems encountered and any resolutions achieved. This will aid with future replication of this project or with the execution of similar future projects.????Describe all the various data used for the project, and include any selection requirements for more detailed data. The data collection report should also define whether some attributes are relatively more important than others. Remember that any assessment of data quality should be made not just of the individual data sources but also of any data that results from merging data sources. Because of inconsistencies between the sources, merged data may present problems that do not exist in the individual data sources.)

Describe data (Examine the "gross" or "surface" properties of the acquired data and report on the results. including the format of the data, the quantity of data (for example, the number of records and fields in each table), the identities of the fields, and any other surfacefeatures which have been discovered. Evaluate whether the data acquired satisfies the relevant requirements. Volumetric analysis of data, Attribute types and values, Check accessibility and availability of attributes: Check attribute types (numeric, symbolic, taxonomy, etc. Check attribute value ranges Analyze attribute correlations Understand the meaning of each attribute and attribute value in business terms For each attribute, compute basic statistics (e.g., compute distribution, average, max, min, standard deviation, variance, mode, skewness, etc.) Analyze basic statistics and relate the results to their meaning in business terms Decide if the attribute is relevant for the specific data mining goal)

Explore data (This task addresses data mining questions using querying, visualization, and reporting techniques. These include distribution of key attributes (for example, the target attribute of a prediction task) relationshipsbetween pairs or small numbers of attributes, results of simple aggregations, properties of significantsub-populations, and simple statistical analyses. These analyses may directly address the data mining goals; they may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed for further analysis.Describe results of this task, including first findings or initial hypothesis and their impact on the remainder of the project. If appropriate, include graphs and plots to indicate data characteristics that suggest furtherexamination of interesting data subsets.Analyze

data, therefore assumptions will be made as to what the data means.

We looked at the 'feedback words' as they provide potentially a direct source of feedback that could be analysied using Natural Language Processing (NLP) but decided against using it. There isn't enough diversity within the data to draw conclusions about the student's engagement with the course.

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.

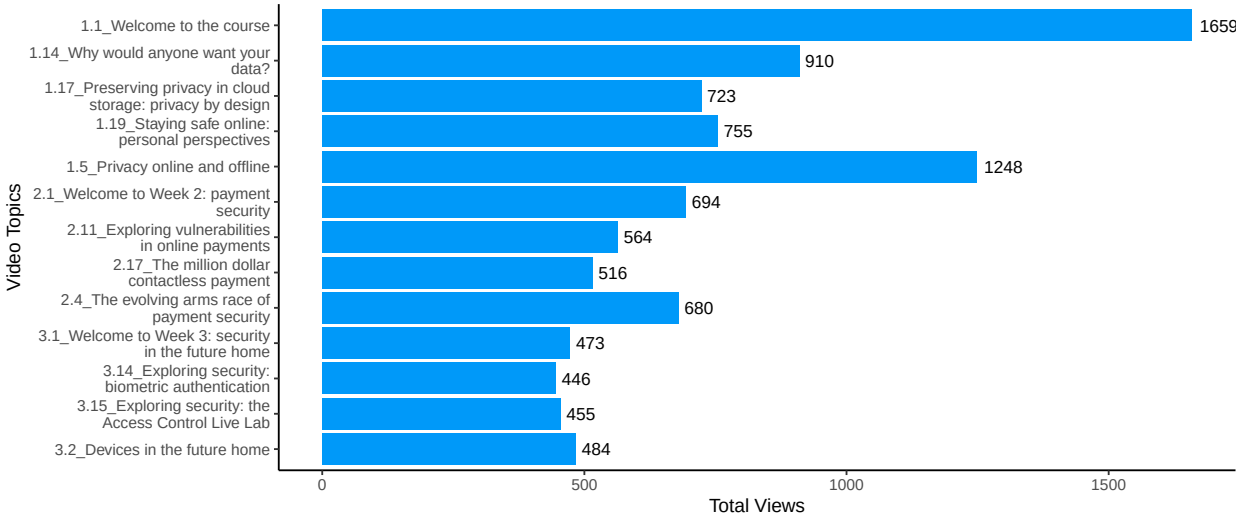
As the study is interested in the number of views across the continent and the drop out, therefore the columns relating to viewing in HD and different devices will be removed. Other remaining columns will remain as they may contain relevant information for the study.

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.

Figure XX: Total number of Viewed per Video Topic



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.

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?

Data Preparation

Dataset (These are the dataset(s) produced by the data preparation phase, which will be used for modeling or the major analysis work of the project.Describe the dataset(s) that will be used for the modeling and the major analysis work of the project)

select data (Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types. Note that data selection covers selectionof attributes (columns) as well as selection of records (rows) in a table. Rationale for inclusion/exclusion: List the data to be included/excluded and the reasons for these decisions.)

Clean data (Raise the data quality to the level required by the selected analysis techniques. This may involve selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling.Describe what decisions and actions were taken to address the data quality problems reported during the Verify Data Quality task of the Data Understanding phase. Transformations of the data for cleaning purposes and the possible impact on the analysis results should be considered.)

construct data (such as the production of derived attributes or entire new records, or transformed values for existing attributes. Derived attributes are new attributes that are constructed from one or more existing attributes in the same record. Example: area = length * width.Describe the creation of completely new records. Example: Create records for customers who made no purchase during the past year. There was no reason to have such records in the raw data, but for modeling purposes it might make sense to explicitly represent the fact that certain customers made zero purchases.)

integrate data (methods whereby information is combined from multiple tables or records to create new records or values.Merged data also covers aggregations. Aggregation refers to operations in which new values are computed by summarizing information from multiple records and/or tables)

format data (Formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool.Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict. It might be important to change the order of the records in the dataset. Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute)

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.

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.

Run 3 data prep Info

All runs data prep info

Modelling

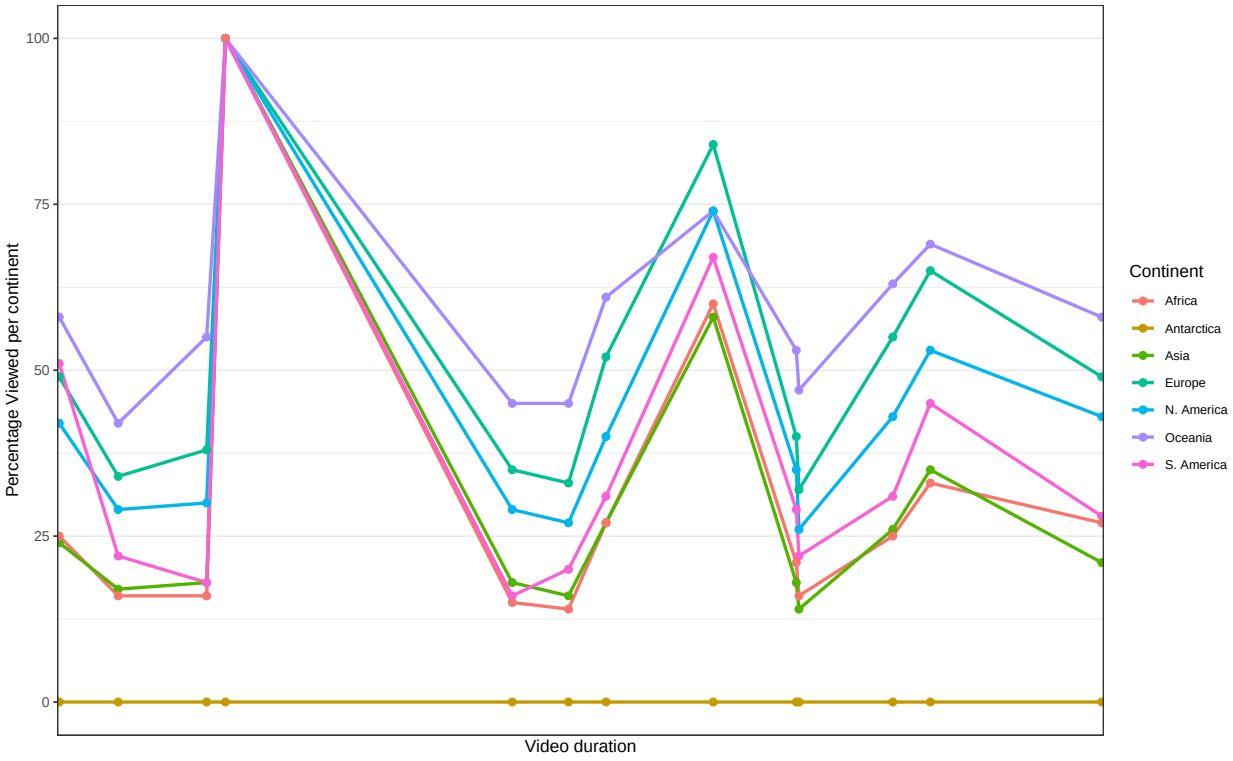
assess model (r interprets the models according to his domain knowledge, the data mining success criteria, and the desired test design. The data mining engineer judges the success of the application of modeling and discovery techniques technically; he contacts business analysts and domain experts later in order to discuss the data mining results in the business context???????)

State conclusions regarding patterns in the data (if any); sometimes the model reveals important facts about the data without a separate assessment process (e.g., that the output or conclusion is duplicated in one of the inputs)

Run 3

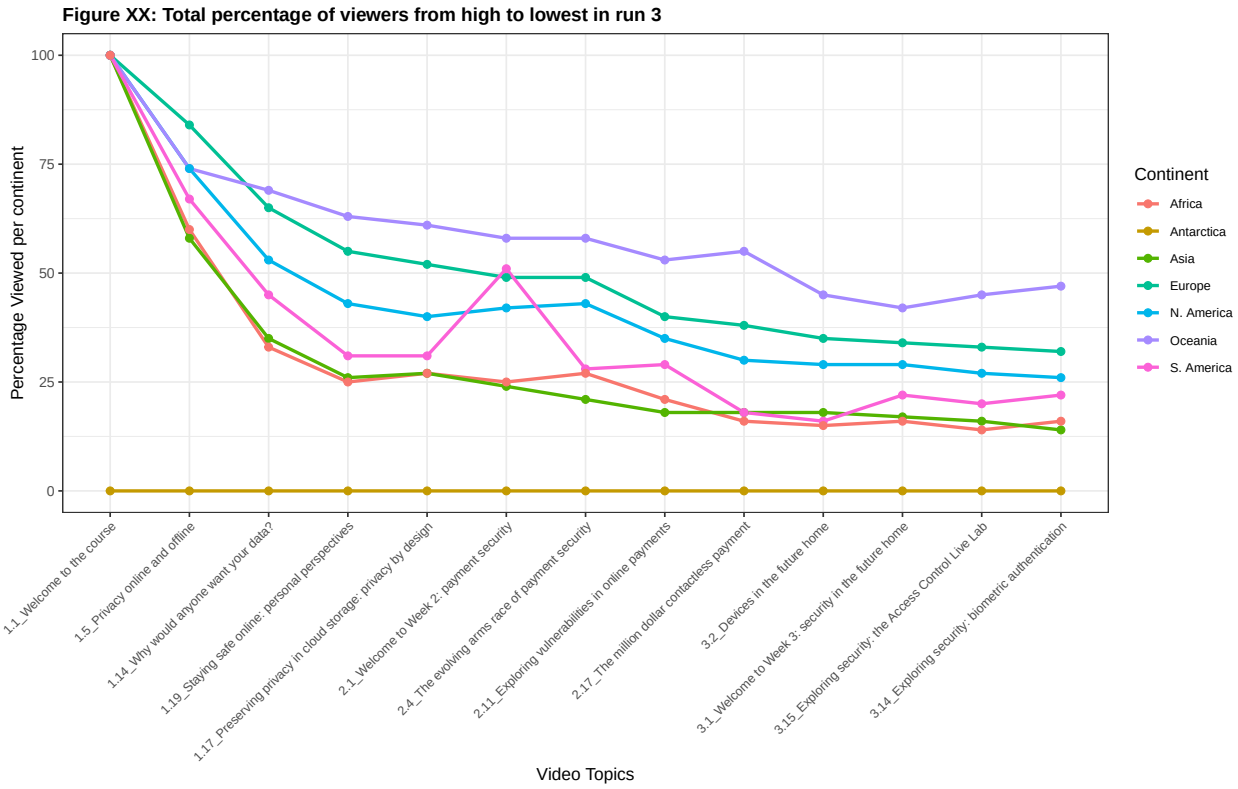
Question 1. Does the duration of the videos have an impact on the viewing rates across different continents?

Figure XX: Percentage of viewers for Video duration



no correlation

line drop out of continent vs topic

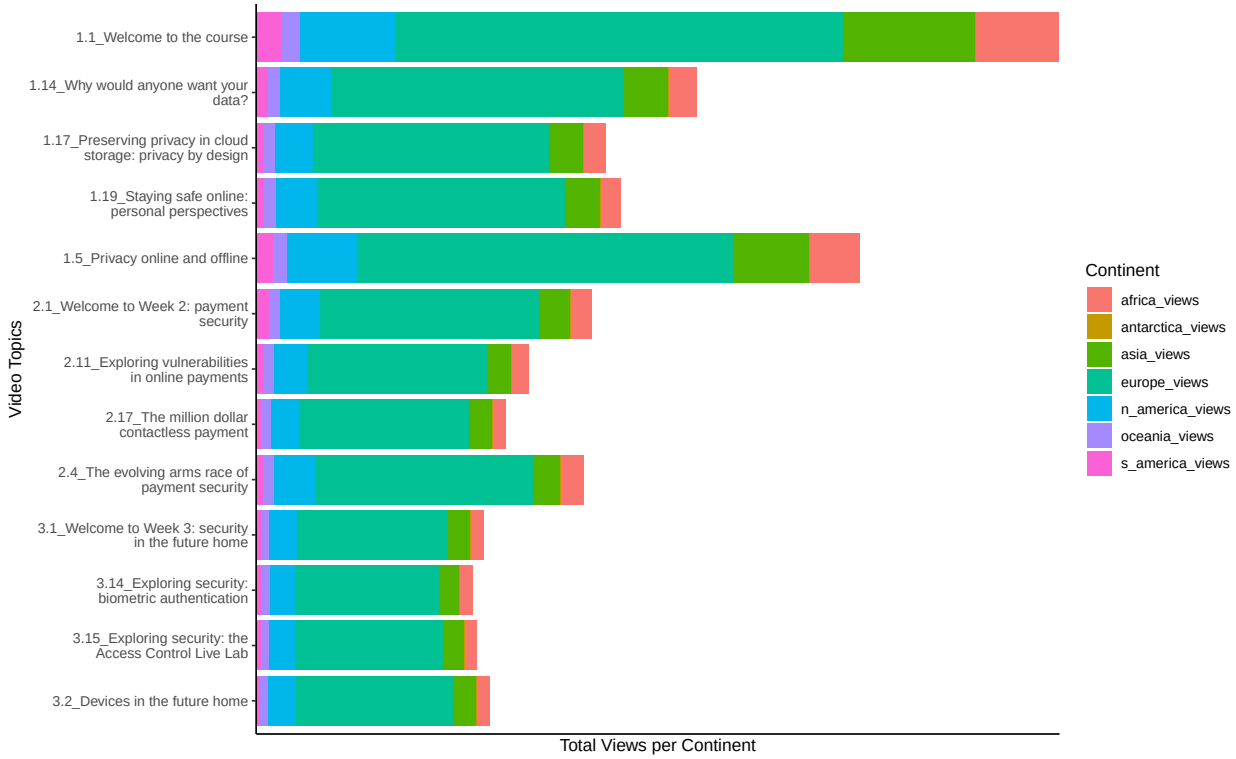


Question 2. Does the content of the videos have an impact on the viewing rates across different continents? using absolute values

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



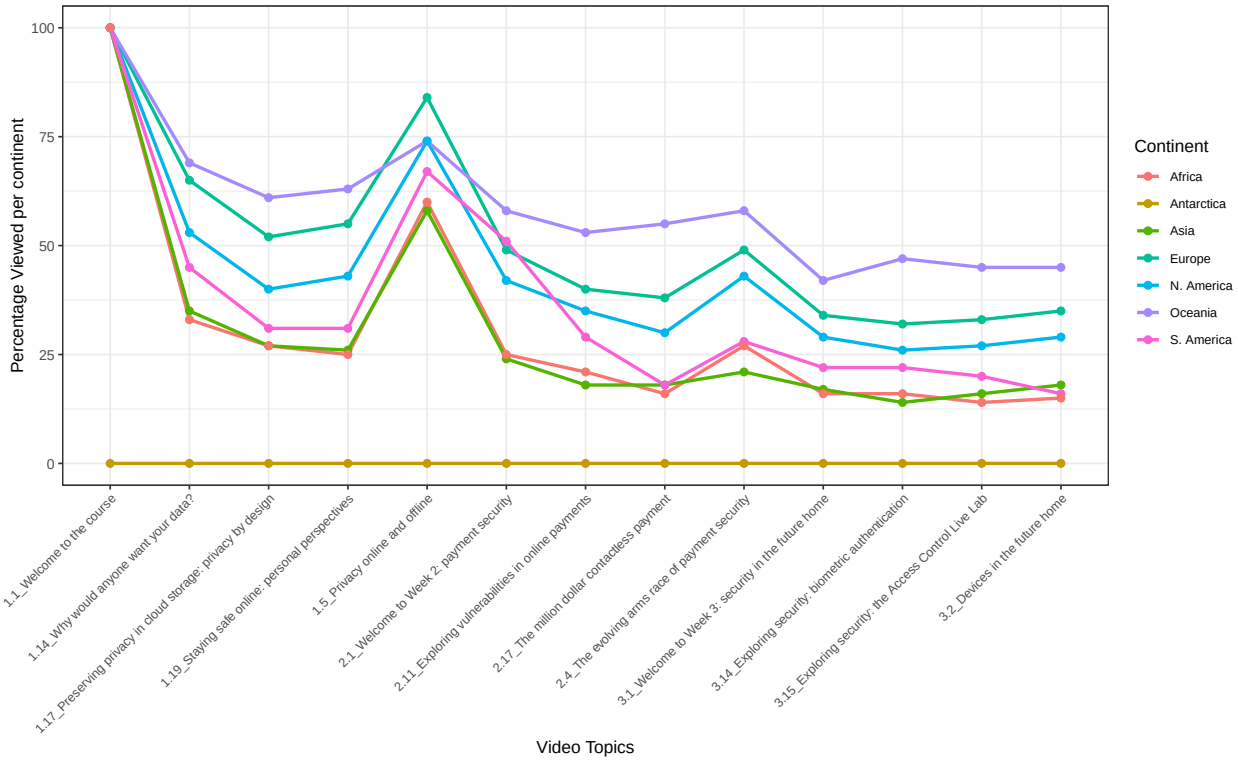
There are no obvious behaviours to be seen, therefore the data will be transformed to compare the views within each continent throughout the course rather than the viewings from each continent within each video.Therefore this could show the drop out rate from each continent throughout the course.

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.

Actual views from continent

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



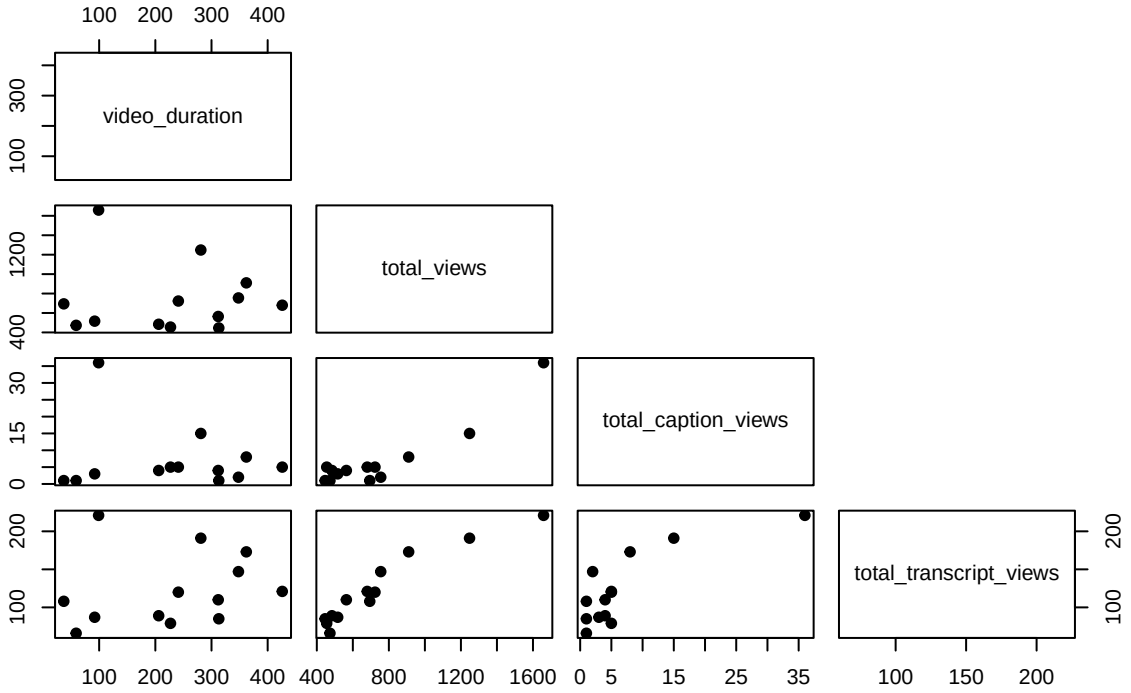
The graph shows that for certain continents (Europe, North America and Oceania), there appears to be more views, therefore more engagements, on the topics within week 2 block of the course. On the other hand, there were more engagements from the learners of Africa and Asia continent, then a steady drop 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.

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.

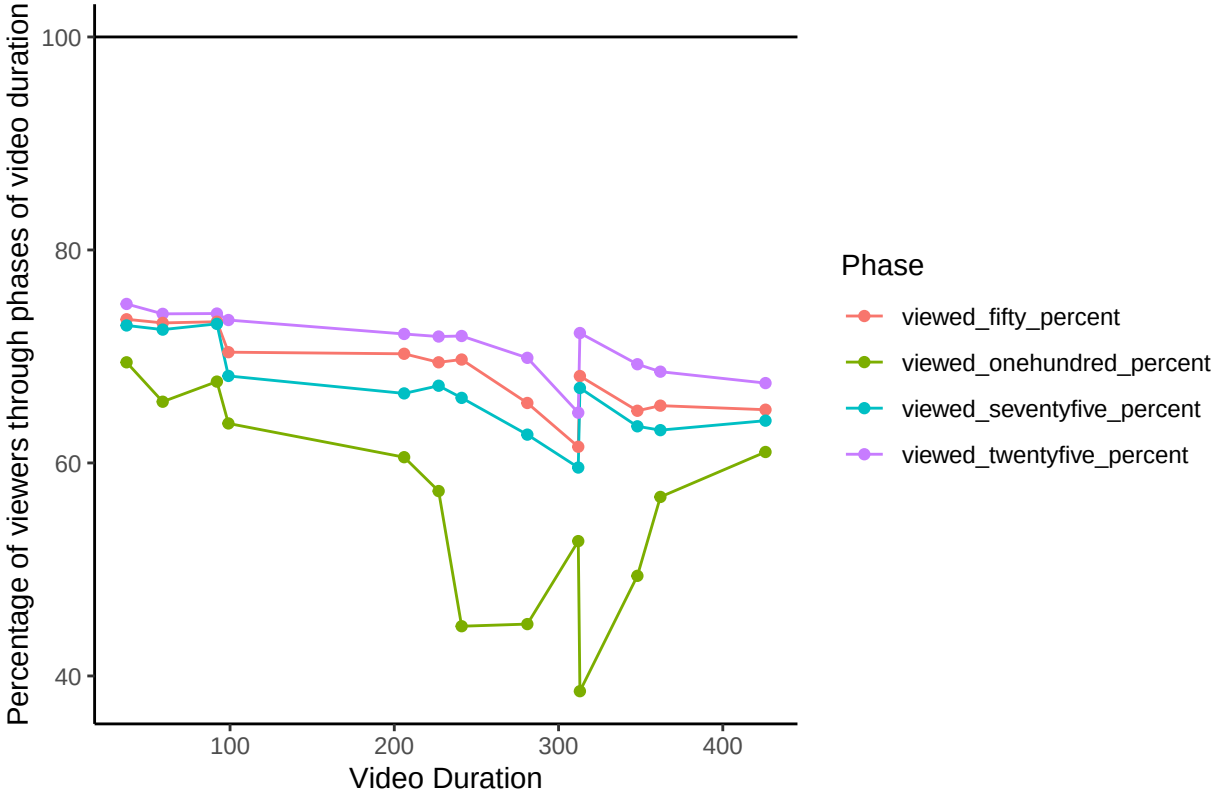
Percentage viewed whole duration of videos/ Number of viewers at 100 percent duration of the videos

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



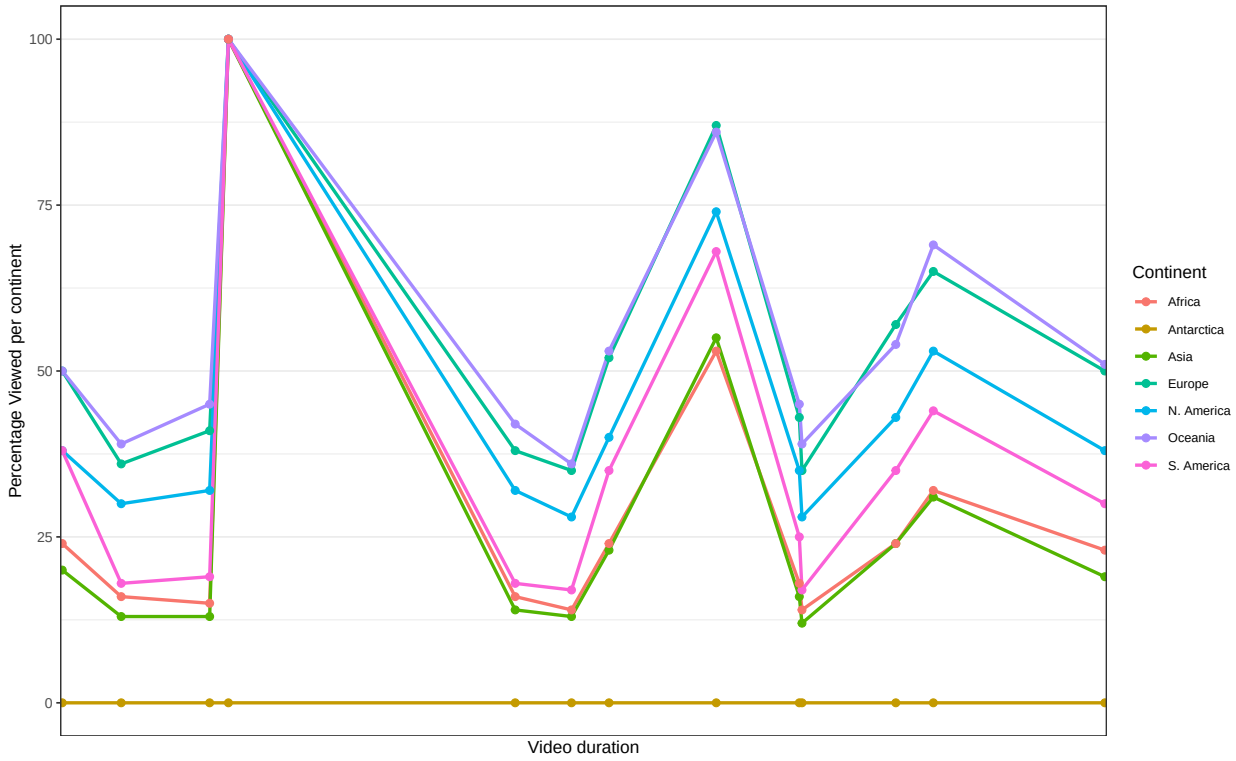
There is potentially a pattern observed on the relationship between the video duration and the number of viewers who have stayed for the whole duration of the videos. However all the runs will need to be included to enable any statements to be made.

All runs

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

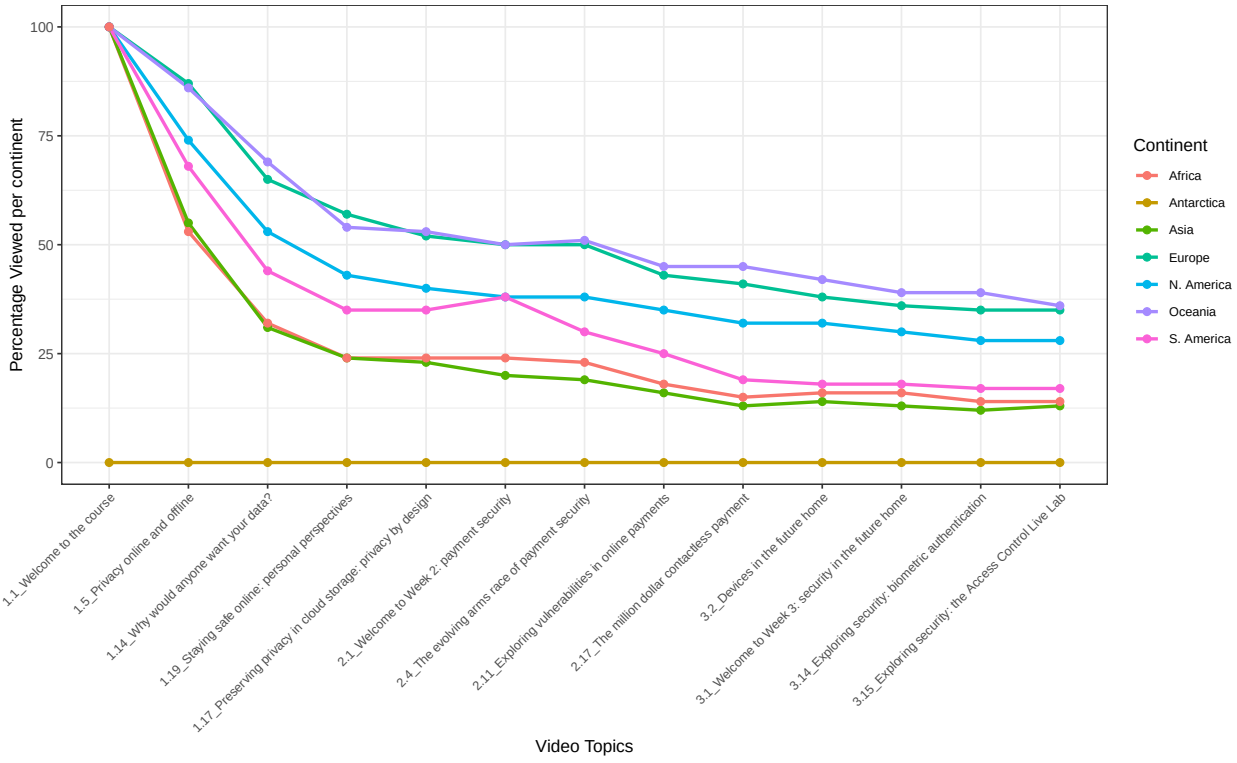
prove with all runs

Figure XX: Percentage of viewers for Video duration



line drop out of continent vs topic for all runs

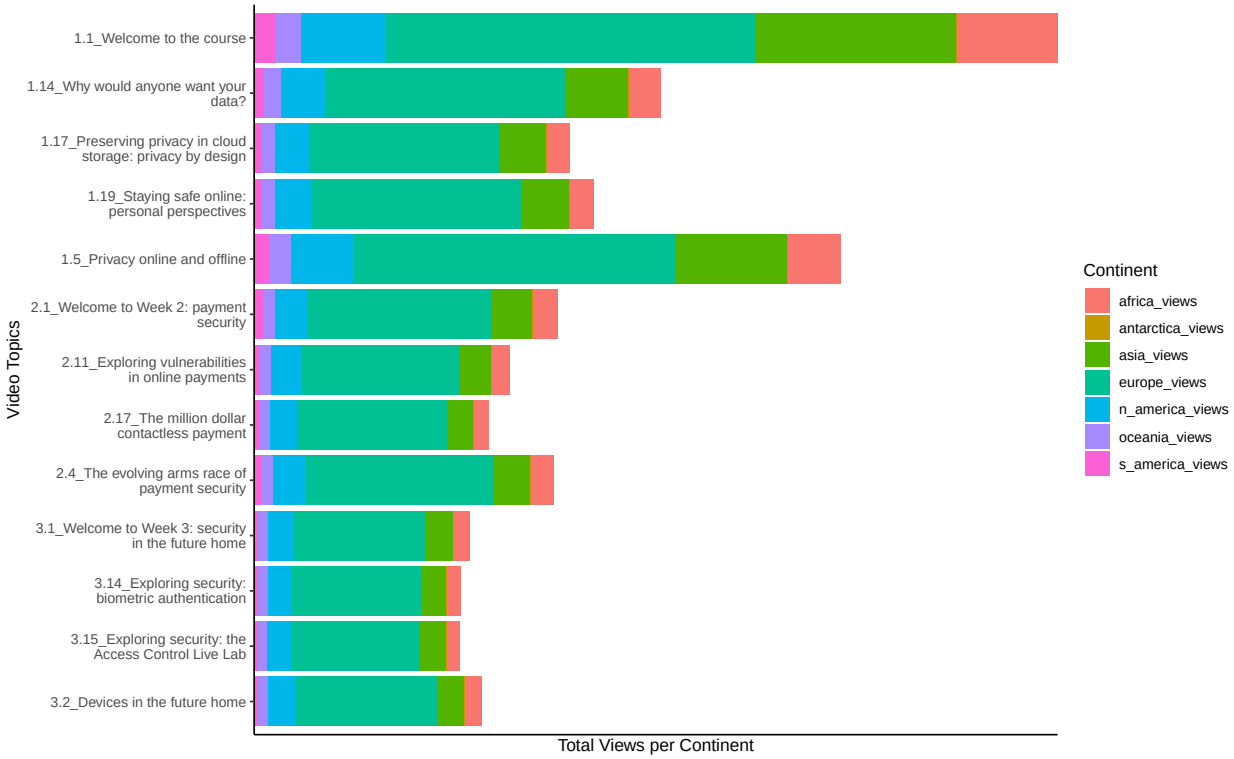
Figure XX: Total percentage of viewers from high to lowest in all runs



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

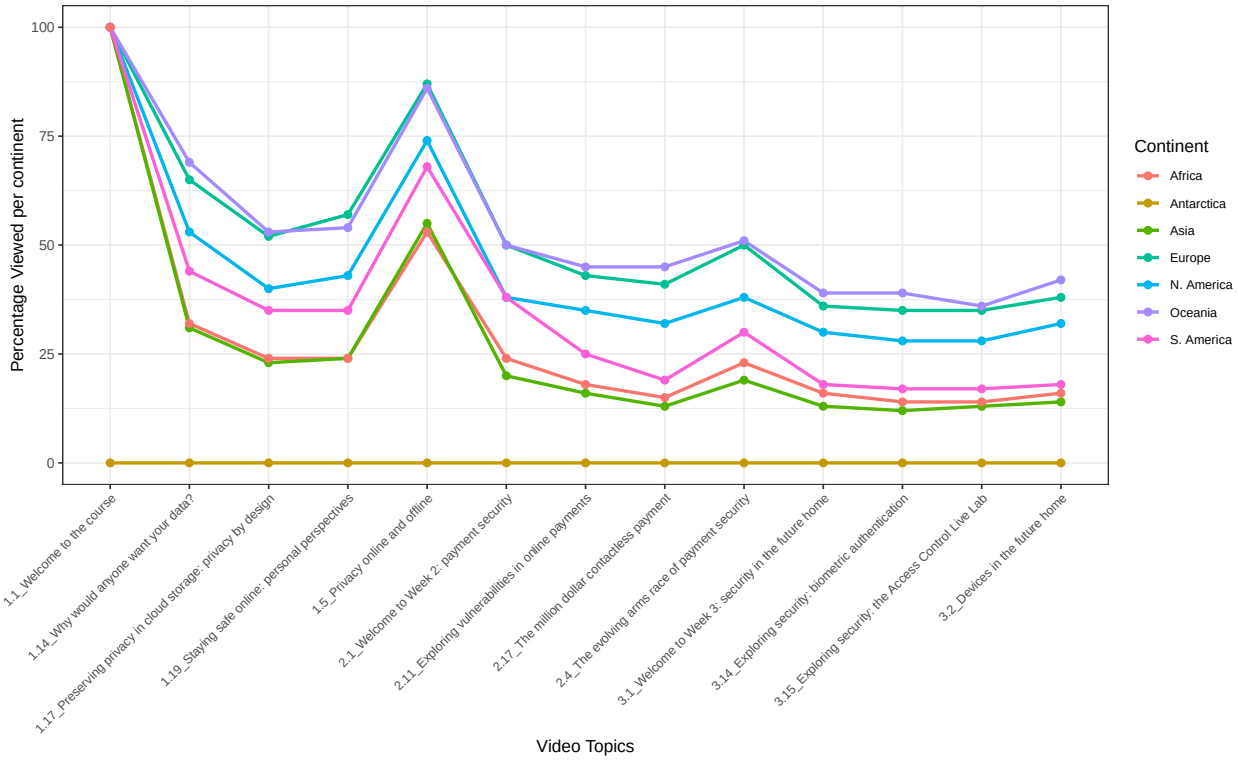
Worldwide views of videos using all runs as bars and absolute values

Figure XX: Total continental viewers on video topics



Actual views from continent using all runs and percentage

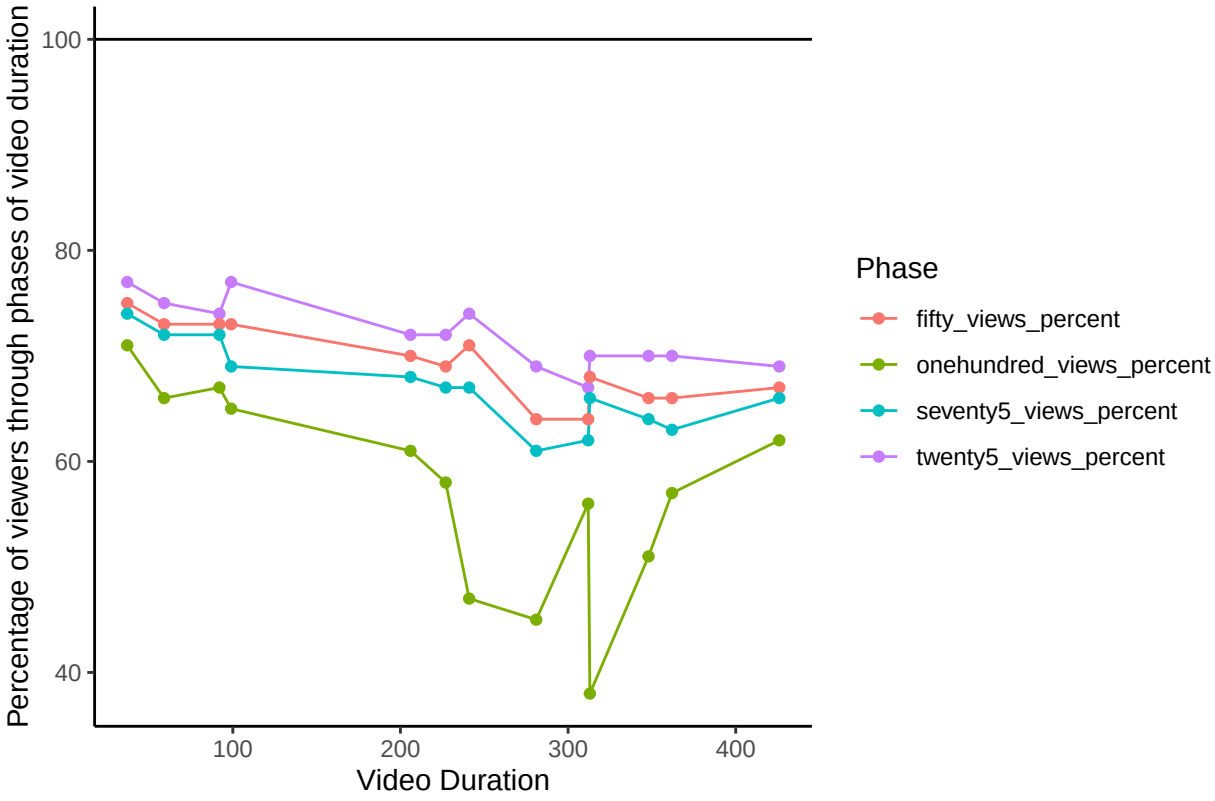
Figure XX: Total percentage of viewers for Video Topics



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

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



Evaluation

evaluate results (assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this model is deficient. Another option is to test the model(s) on test applications in the real application, if time and budget constraints permit.assesses other data mining results generated. Data mining results involve models that are necessarily related to the original business objectives and all other findings that are not necessarily related to the original business objectives, but might also unveil additional challenges, information, or hints for future directions.Summarize assessment results in terms of business success criteria, including a final statement regarding whether the project already meets the initial business objectives.After assessing models with respect to business success criteria, the generated models that meet the selected criteria become the approved models.???) think about history of data is it still relevant

Preview of videos allowed on website - does not influence the results.

Review process (resulting models appear to be satisfactory and to satisfy business needs. It is now appropriate to do a more thorough review of the data mining engagement in order to determine if there is any important factor or task that has somehow been overlooked.Summarize the process review and highlight activities that have been missed and those that should be repeated)

determine next steps (Depending on the results of the assessment and the process review, the project team decides how to proceed. The team decides whether to finish this project and move on to deployment, initiate further iterations, or set up new data mining projects. This task includes analyses of remaining resources and budget, which may influence the decisions.List the potential further actions, along with the reasons for and against each option.Describe the decision as to how to proceed, along with the rationale.)

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