CSC8631 Report

Selina So

23/11/2021

1. Business Understanding

1.1 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 (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???))

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, 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, 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 retaining 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 (https://www.futurelearn.com/courses/cyber-security). The course consist of a combination of videos, articles, exercises, discussions, quizzes and tests. 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). Therefore in this study we will examine the data from the *Cyber Security Safety at Home, Online, in Life* course to understand the factors which could influence the learners' retention rate. *Insights into the relationship of engagements across different continents. Is the course able to reach and retain a broad range of learners from across the world, from the material provided in the course. We aim to understand the areas which appear successful and areas that don't. With this insight, informed decisions could be made on areas where improvements could be made (reword).*

(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)

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

1.2.1 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 utilized for expert knowledge and as stakeholders for the study will be the University educators (Dr Matthew Forshaw and Joe Matthews) and teaching experts include?
- 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) need to include?
- The softwares used will be R, RStudio and ggplot2. include?

1.2.2 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)

There are legal obligations and privacy policies, such as GDPR (Data Protection) and the *Data Protection Act (2018)*, to consider before using the data. It is assumed that the data provided by FutureLearn have been agreed and approved for use in this study. To comply with the legal and ethical standards, we will ensure any identifying data observed will be anonymised to reduce the likelihood that an individual could be identified.

1.3 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' 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 the University of Newcastle to modify the course content to achieve optimise learner engagement and retention.

1.4 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.????)

We will utilize the CRISP-DM process to understand the data: We will perform initial investigation of the data, identify potential trends and form hypotheses which could provide interesting insights for Newcastle University and FutureLearn. Simple descriptive statistical and visualization techniques provide first insights into the data.

```
Data description may help identify interesting segments in the data.


If there are potential trends, then further in-depth analysis (or a description and summarization of t
```

Data mining problem type - Data description and summarization aims at the concise description of characteristics of the data, typically in elementary and aggregated form. This gives the user an overview of the structure of the data. Sometimes, data description and summarization alone can be an objective of a data mining project. For instance, a retailer might be interested in the turnover of all outlets broken down by categories. Changes and differences in a previous period could be summarized and highlighted. This kind of problem would be at the lower end of the scale of data mining problems. In almost all data mining projects, however, data description and summarization is a subordinate goal in the process, typically in its early stages. At the beginning of a data mining process, the user often knows neither the precise goal of the analysis nor the precise nature of the data. Initial exploratory data analysis can help users to understand the nature of the data and to form potential hypotheses for hidden information. Simple descriptive statistical and visualization techniques provide first insights into the data. For example, the distribution of customers by age and geographic regions suggests which parts of a customer group need to be addressed by further marketing strategies. Data description and summarization typically occurs in combination with other data mining problem types. For instance, data description may lead to the postulation of interesting segments in the data. Once segments are identified and defined, a description and summarization of these segments is useful. It is advisable to carry out data description and summarization before any other data mining problem type is addressed. In this document, this is reflected in the fact that data description and summarization is a task in the data understanding phase. Summarization also plays an important role in the presentation of final results. The outcomes of the other data mining problem types (e.g., concept descriptions or prediction models) may also

be considered summarizations of data, but on a higher conceptual level. Many reporting systems, statistical packages, OLAP, and EIS systems can cover data description and summarization but do usually not provide any methods to perform more advanced modeling. If data description and summarization is considered a stand-alone problem type and no further modeling is required, then these tools may be appropriate to carry out data mining engagements.

1.4.1 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.???)

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

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

2. Data Understanding (Initial Observation)

2.1 Collect initial data (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.)

Loaded data in R and cached.

Will consider integration of data later (joining) in data prep phase.

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. The 53 csv datasets are located in the *'data'* folder. There are 7 runs of data, each representing different timeframes of when the data were collected, between mid 2016 - mid 2018. Runs 3 to 7 consists of the following datafiles: *'Archetype-survey-responses', 'Enrolments', 'Leaving-survey-responses', 'Question-response', 'Step-activity', 'Weekly-sentiment-survey-response, 'Team-members', 'Video-stats'*. Run 1 consists of most datafiles expect *'Team-members', 'Video-stats'*, and run 2 consists of most datafiles except *'Video-stats'*.

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.)

2.2 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)

2.3 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 properties of interesting attributes in detail (e.g., basic statistics, interesting sub-populations: Identify characteristics of sub-populations/ Form suppositions for future analysis Consider and evaluate information and findings in the data descriptions report Form a hypothesis and identify actions Transform the hypothesis into a data mining goal, if possible Clarify data mining goals or make them more precise. A "blind" search is not necessarily useless, but a more directed search toward business objectives is preferable. Perform basic analysis to verify the hypothesis)

2.4 Verify data quality (Examine the quality of the data, addressing questions such as: Is the data complete (does it cover all the cases required)? Is it correct, or does it contain errors and, if there are errors, how common are they? Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they? List the results of the data quality verification; if quality problems exist, list possible solutions. Solutions to data quality problems generally depend heavily on both data and business knowledge. Identify special values and catalog their meaning Review keys, attributes Check coverage (e.g., whether all possible values are represented) Check keys Verify that the meanings of attributes and contained values fit together Identify missing attributes and blank fields Establish the meaning of missing data Check for attributes with different values that have similar meanings (e.g., low fat, diet) Check spelling and format of values (e.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter) Check for deviations, and decide whether a deviation is "noise" or may indicate an interesting phenomenon Check for plausibility of values, (e.g., all fields having the same or nearly the same values)Noise and inconsistencies between sources Check consistencies and redundancies between different sources Plan for dealing with noise

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

ADD HERE: Short description of other data files and why they were discarded (e.g. too many missing data points; not clear what data means without further background information, hence not being used for further analysis).

ADD HERE: The data files called '_survey' contained information about learner ID, leaving date, and learners' feedback, which was given as what is assumed to be pre-set selection of feedback options. These data would have been useful to combine with the data from the video stats files. However, when we looked at the 'feedback words' as they provide a direct source of feedback that could be analysed using Natural Language Processing (NLP), we found that the data quality isn't good enough as there isn't enough diversity within the data to draw conclusions about the student's engagement with the course. The pre-selection only allowed learners to select from a very narrow range of options, which were too high level to provide meaningful insight into the sentiments of the learners.

As the study is focussed on the use of video material, the data files with the title containing '*video.stats*' will be used. 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 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 each individual analysis.

The data sets are mostly complete with no visible missing data.

ADD HERE: 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 first column of each data set describes the step number of each video. The second column is a string which the video title. These two columns will be combined to allow quick reference to the order of which the videos appears throughout the course. Columns 3-28 consist of continuous data.

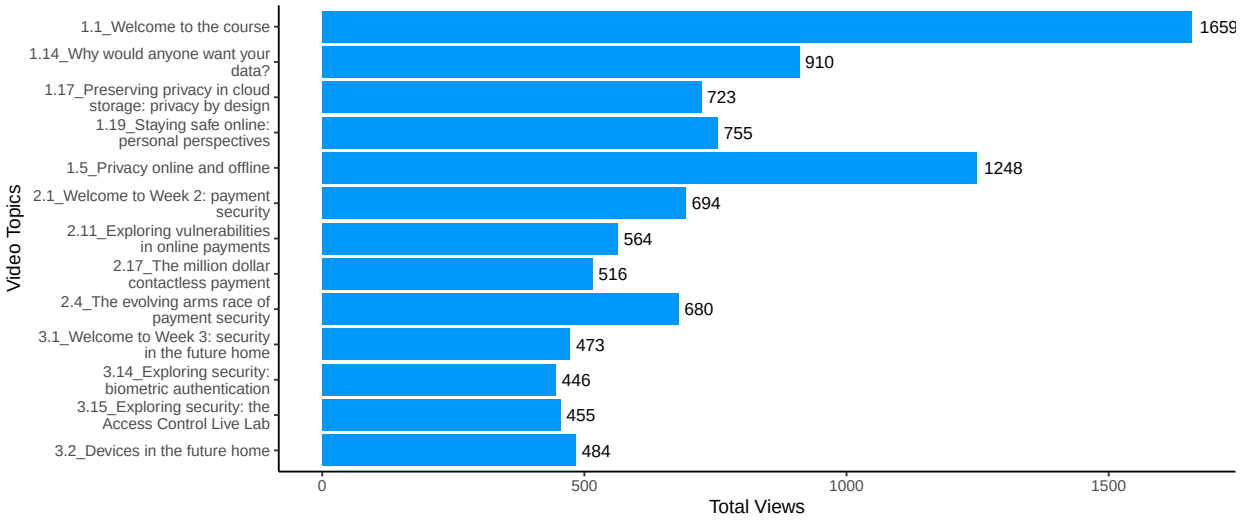
As the study is interested in the number of views across the continent and the drop out, therefore the columns relating to viewing in HD and different devices will be removed. Other remaining columns will remain as they may contain relevant information for the study.

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 data set. Based on enrolment time stamps in data set cyber-security-3_enrolments.csv, 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.

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

[QUESTION: Why do you add the description of scatterplots here and directly underneath the questions? DO you really need the scatterplots here? Or is the overview figure above and the description of column titles enough to formulate your questions?]

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?

3. Data Preparation

3.1 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)

3.2 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.)

3.3 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.)

3.4 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.)

3.5 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)

3.6 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. 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.

[DELETE THIS PARAGRAPH]The total number of viewings from each continent for each video is calculated. This is by taking the total number of learners (by taking the '*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 (?)

[REPLACE WITH THIS] The total number of viewings from each continent for each video is calculated. This is calculated for each video (i.e. each row) by taking the total 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 total views of the first video from African viewers, which is calculated as (*Total Viewed*[row 2]/100)x*africa_views_percentage*[row 2].

The percentage viewed from each continent throughout the course is calculated by [ADD DESCRIPTION OF HOW YOU CALCULATED IT].These values were then combined with the original '*step_title*' and '*video_duration*' columns. Values are rounded to full numbers [HOW MANY DECIMAL POINTS INSTEAD OF FULL NUMBERS, e.g values are given as numbers with no decimal points] and '*NaN*' are replaced with 0, as these result from the viewings across Antarctica (*antarctica_views_percentage*), which were 0 across all runs.

Run 3 data prep Info

All runs data prep info

4. Modelling

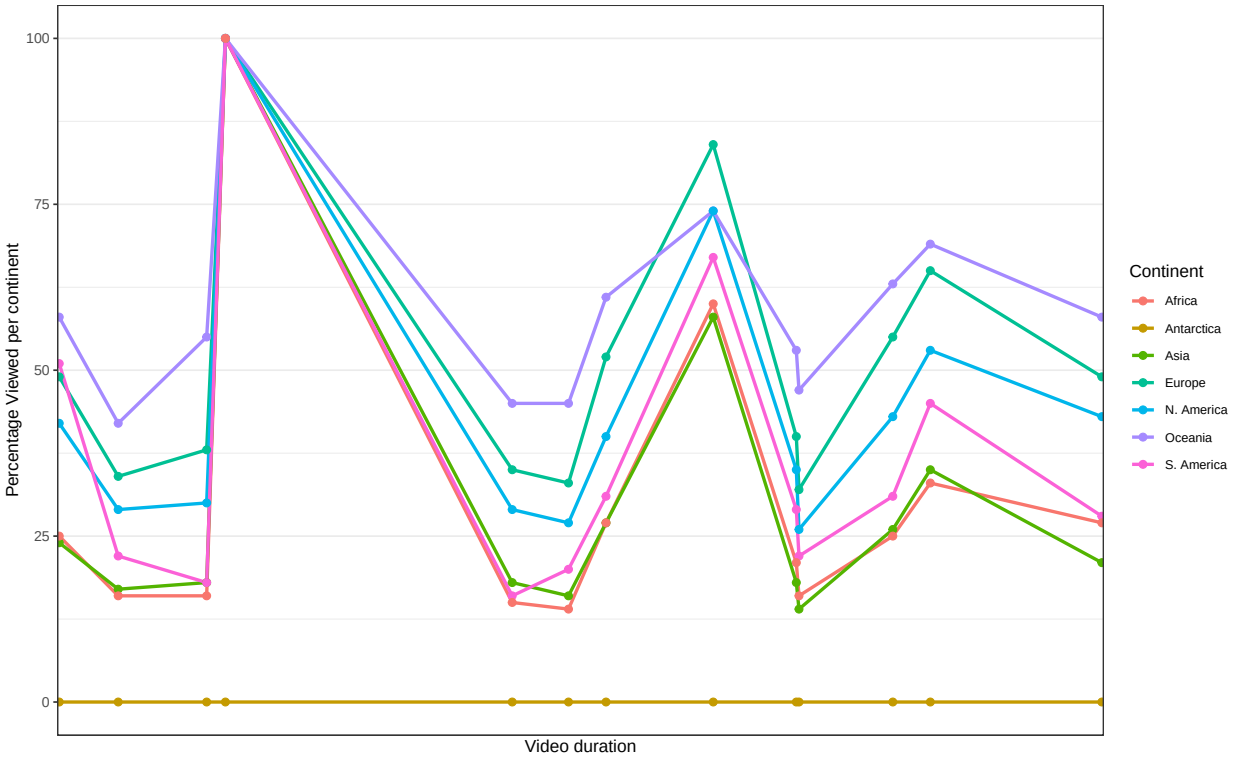
4.1 Assess model (r interprets the models according to his domain knowledge, the data mining success criteria, and the desired test design. The data mining engineer judges the success of the application of modeling and discovery techniques technically; he contacts business analysts and domain experts 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

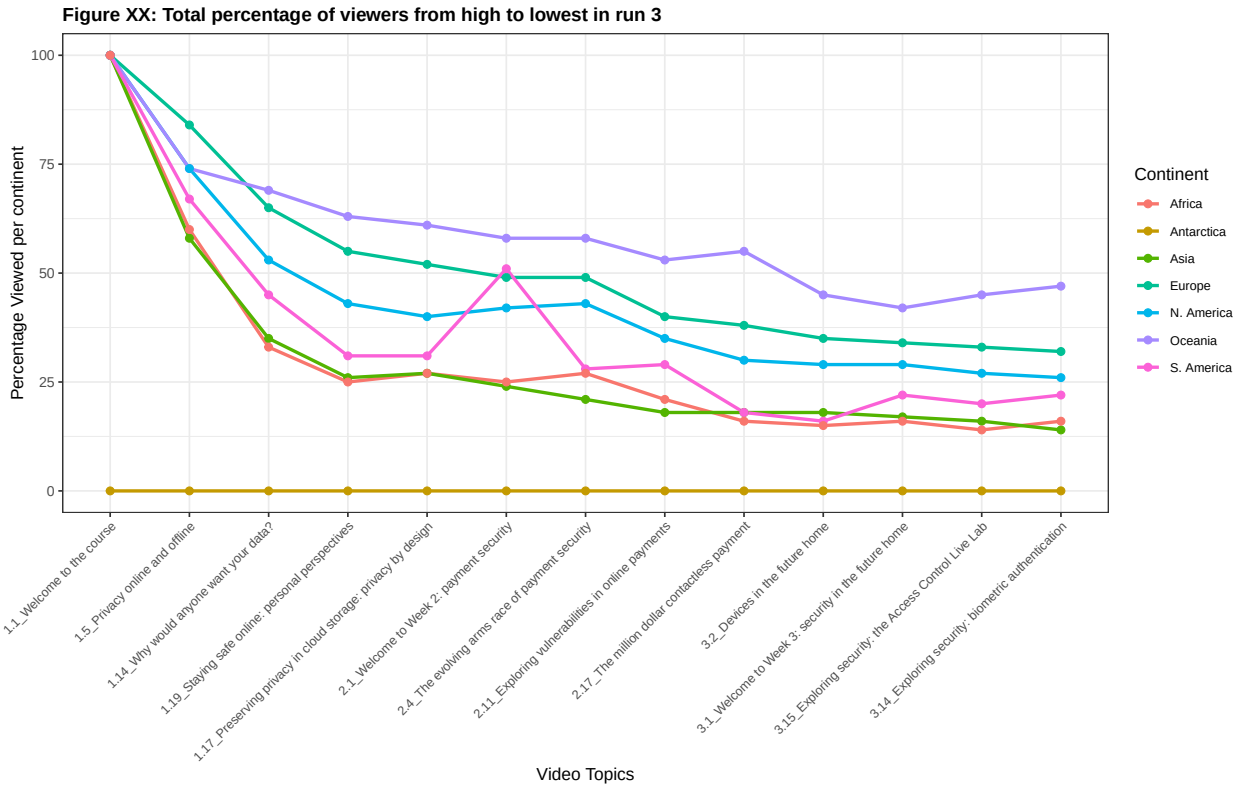
Question 1. Does the duration of the videos have an impact on the viewing rates across different continents?

Figure XX: Percentage of viewers for Video duration



no correlation

line drop out of continent vs topic

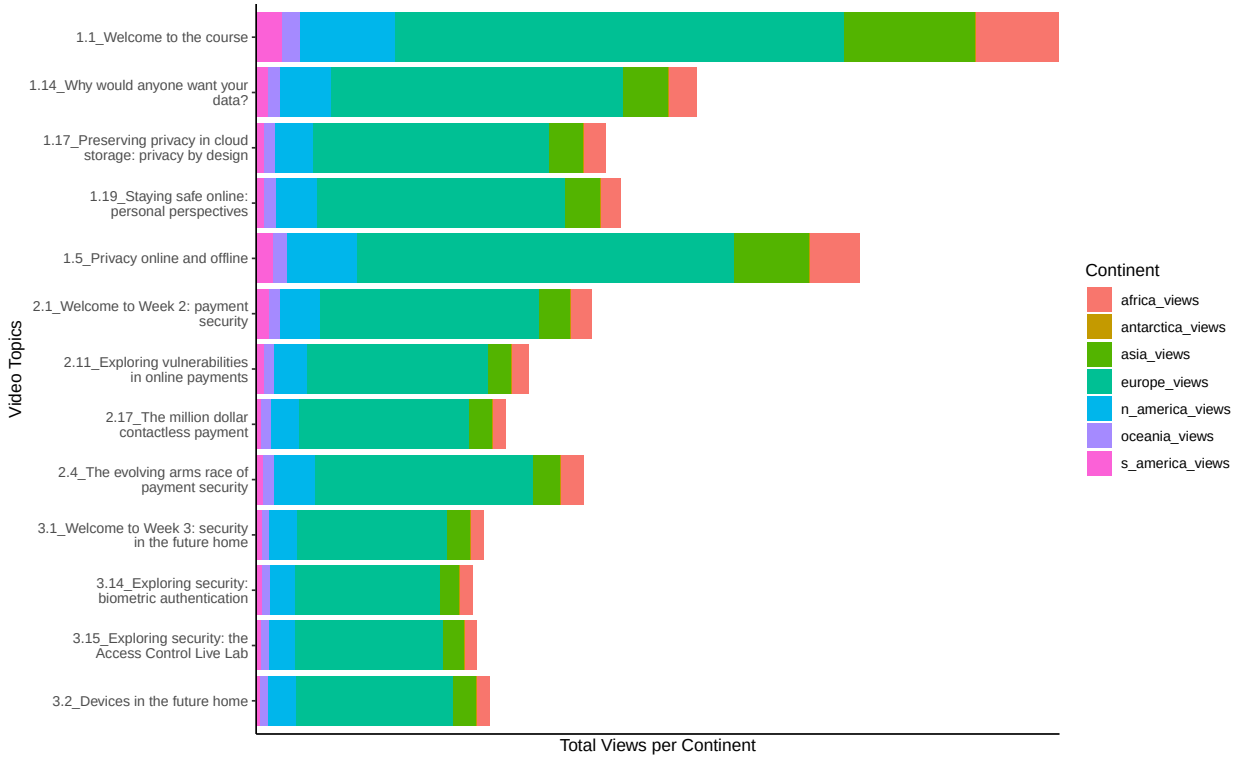


Question 2. Does the content of the videos have an impact on the viewing rates across different continents? using absolute values

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



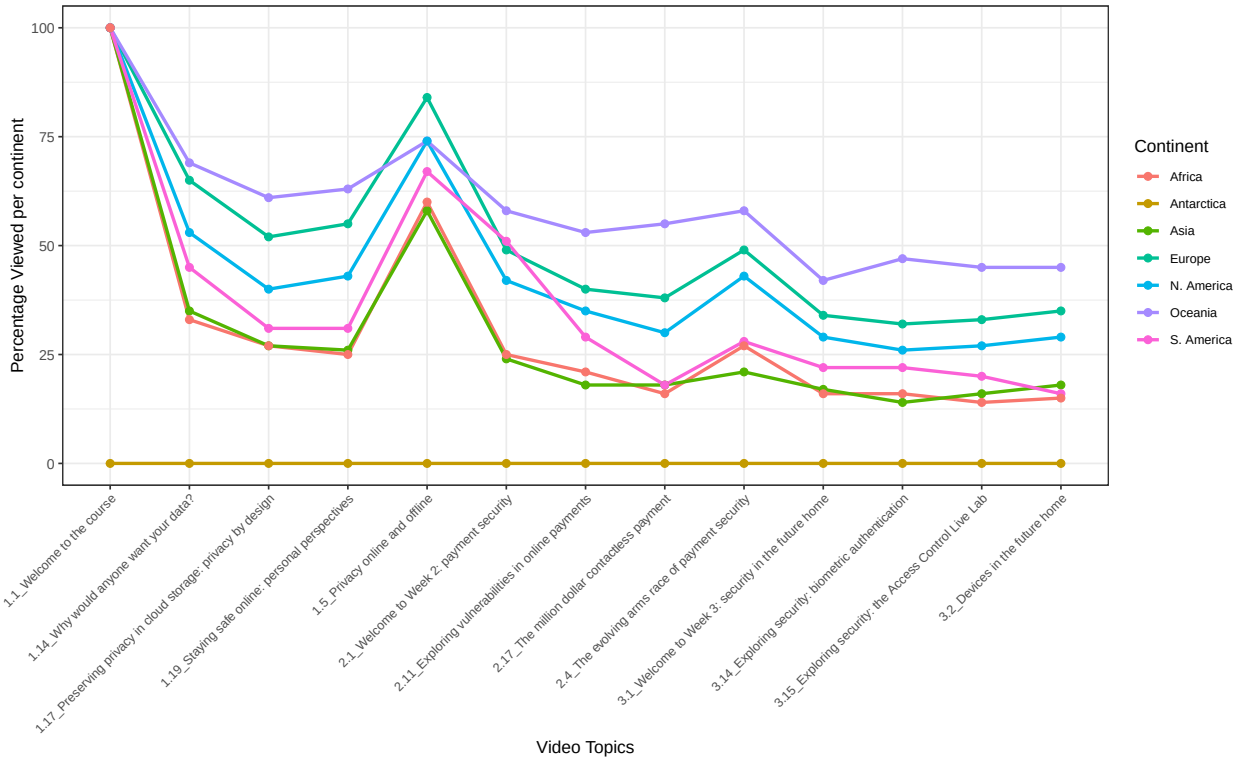
There are no obvious correlations to be seen, therefore the data which allows comparison of the views within each continent throughout the course rather than the viewings from each continent within each video will be assessed in the further analysis. This will show the drop out rate from each continent throughout the course, rather than the relative number of viewings from each continent for each video.

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 to be 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 [DO YOU MEAN ACTUAL, OR ABSOLUTE, OR WOULD IT BE BETTER TO CALL IT SOMETHING LIKE 'Comparison of views per video for each continent'] 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



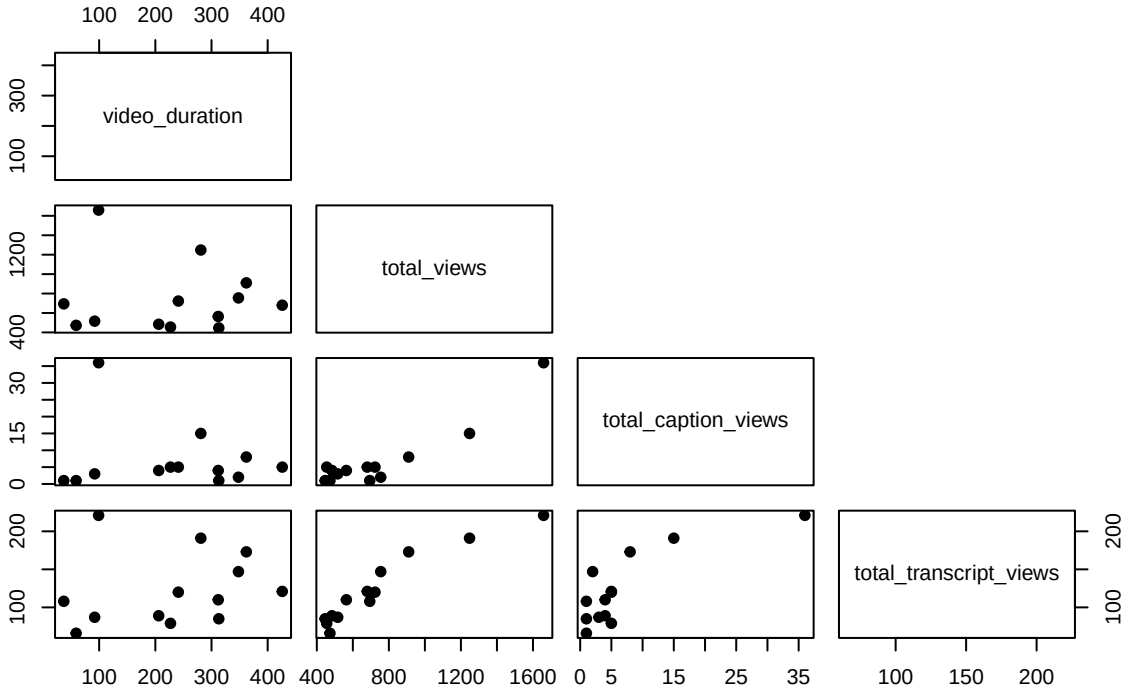
The graph shows that for certain continents (Europe, North America and Oceania), there appears to be more views, therefore more engagements, on the topics within week 2 block of the course. On the other hand, there were more engagements from the learners of Africa and Asia continent, then a steady drop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

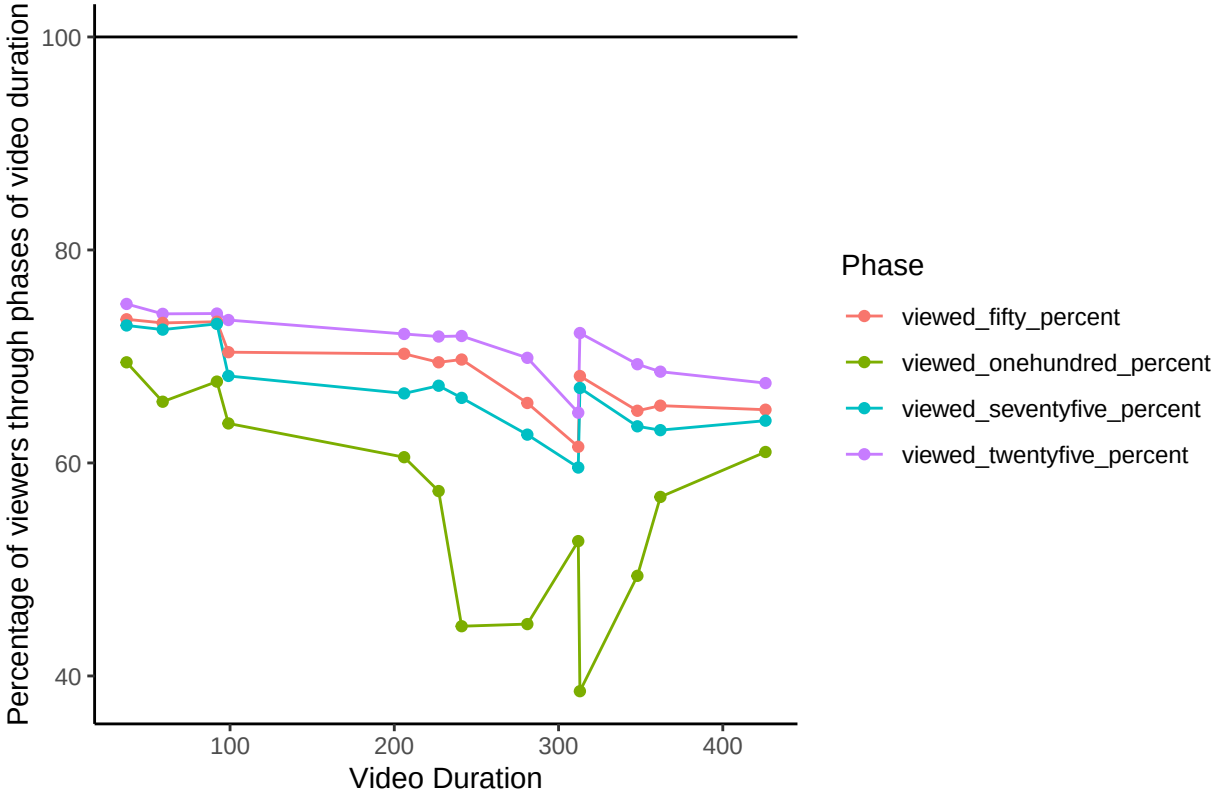
Percentage viewed whole duration of videos/ Number of viewers at 100 percent duration of the videos

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



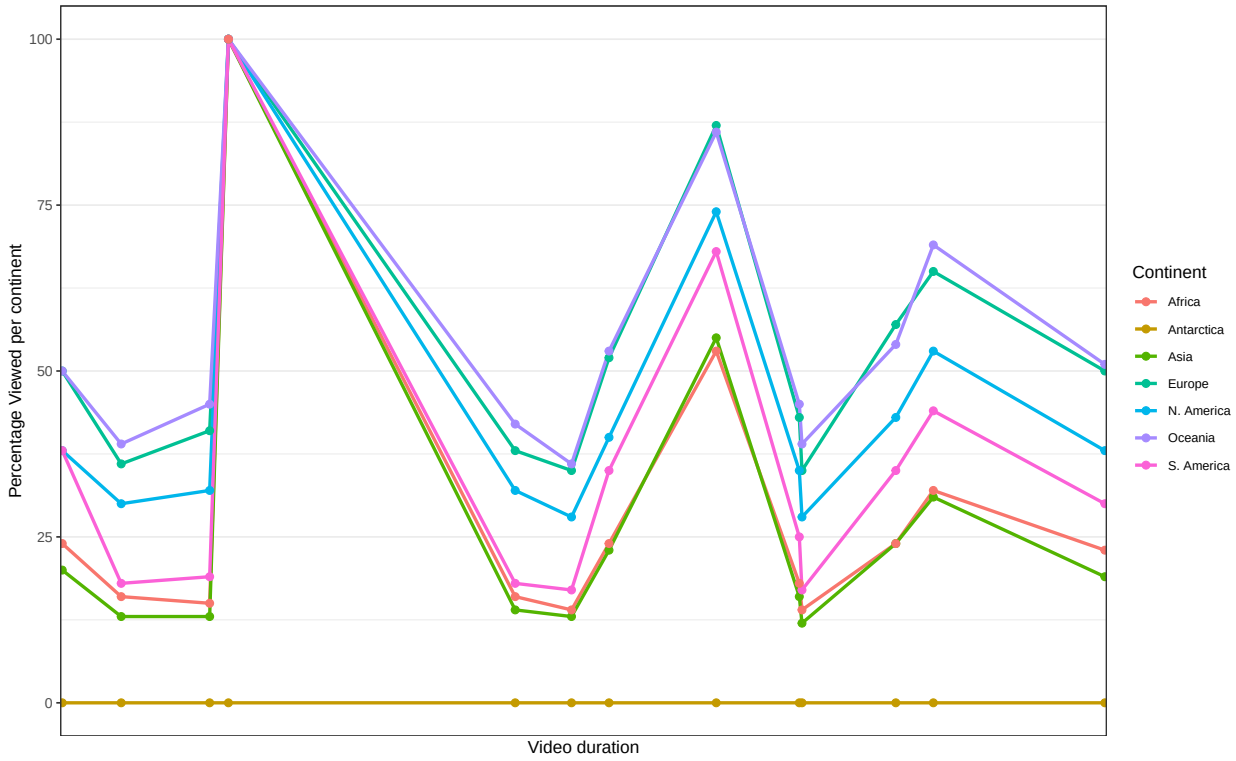
There is potentially a pattern observed on the relationship between the video duration and the number of viewers who have stayed for the whole duration of the videos. However the data is too noisy due to the relatively low number of learners for some continents, and all the runs will need to be included to enable more reliable statements to be made.

All runs

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

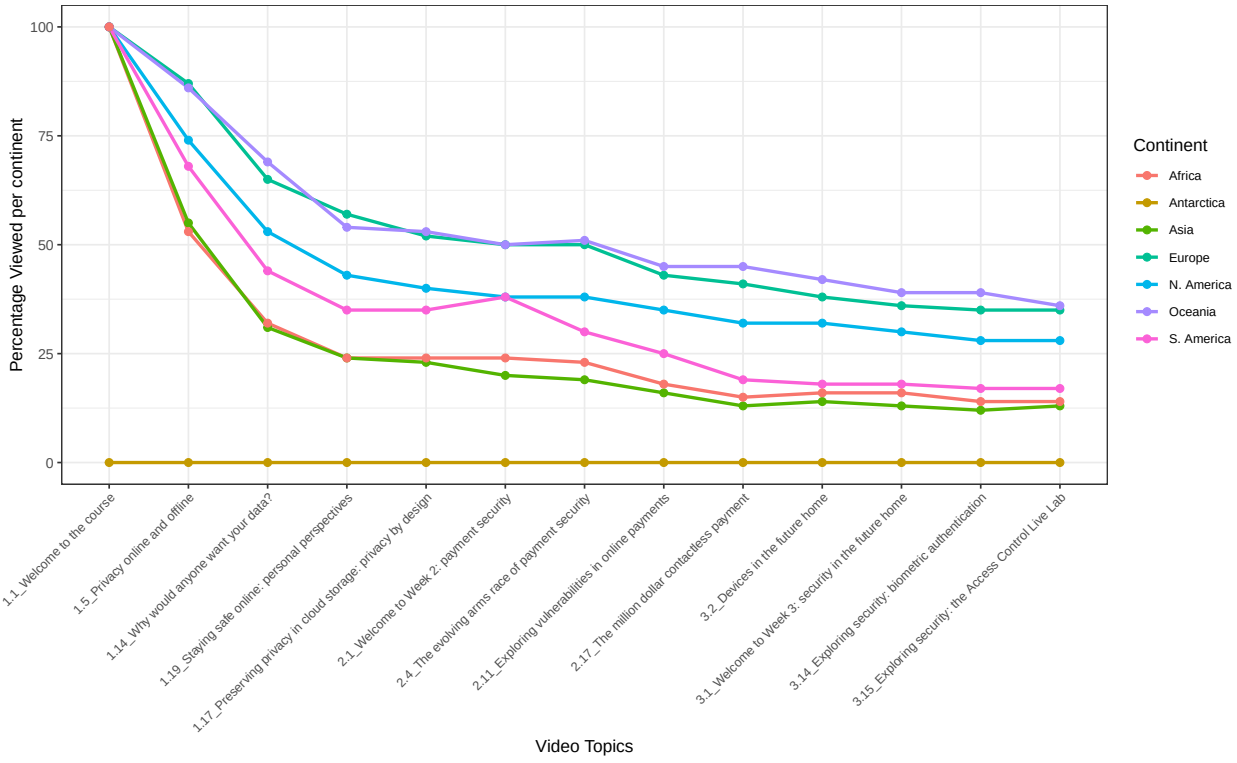
prove with all runs

Figure XX: Percentage of viewers for Video duration



line drop out of continent vs topic for all runs

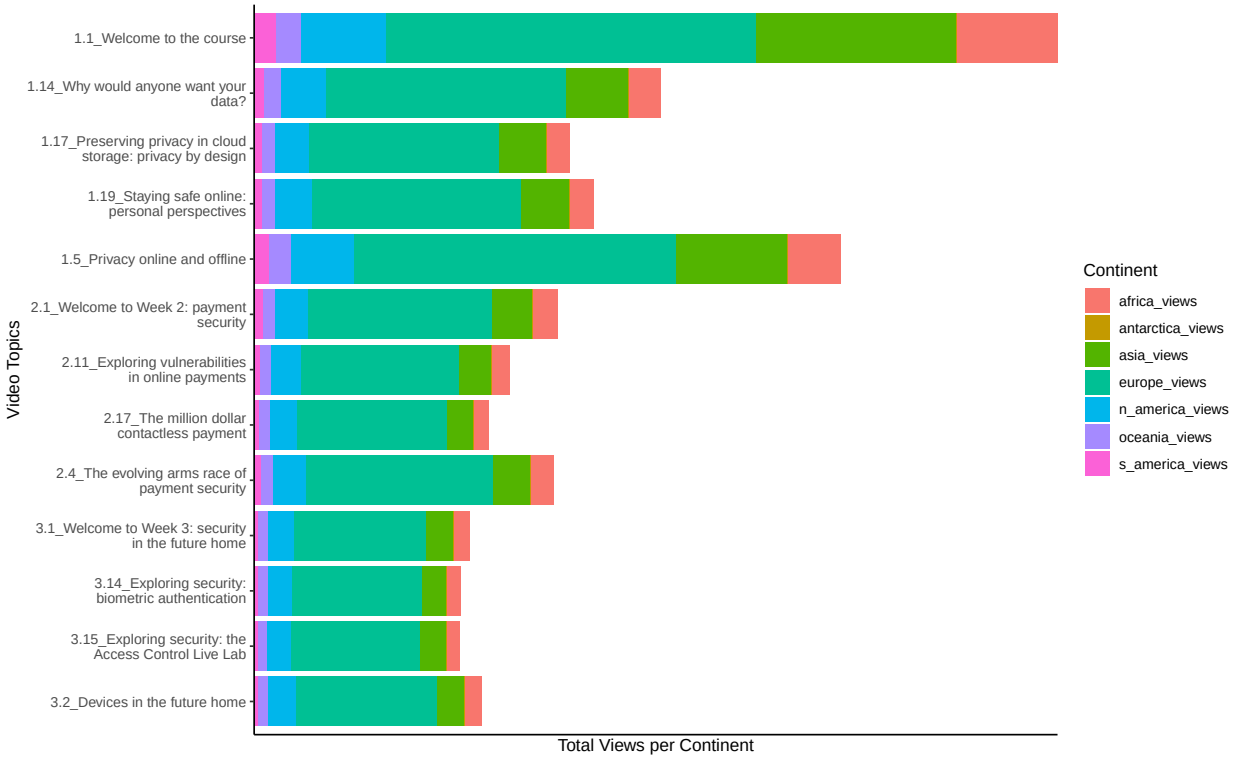
Figure XX: Total percentage of viewers from high to lowest in all runs



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

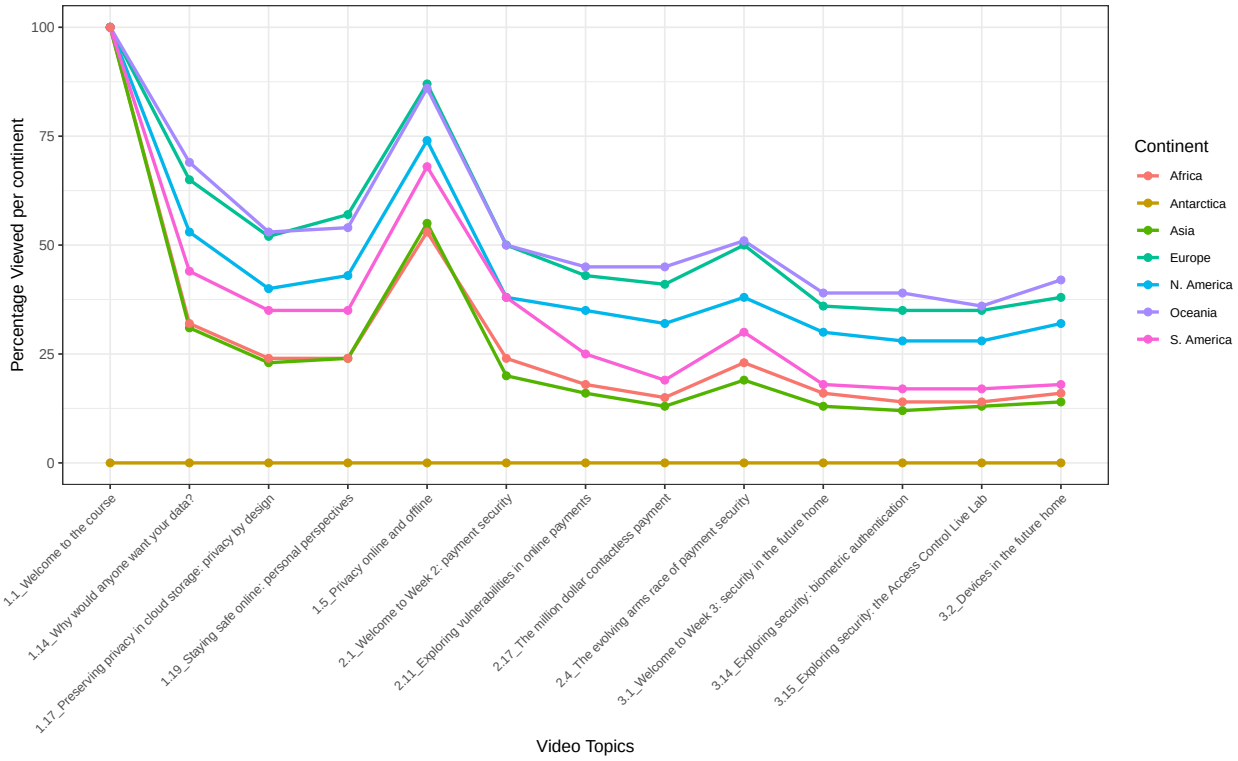
Worldwide views of videos using all runs as bars and absolute values

Figure XX: Total continental viewers on video topics



Actual views from continent using all runs and percentage

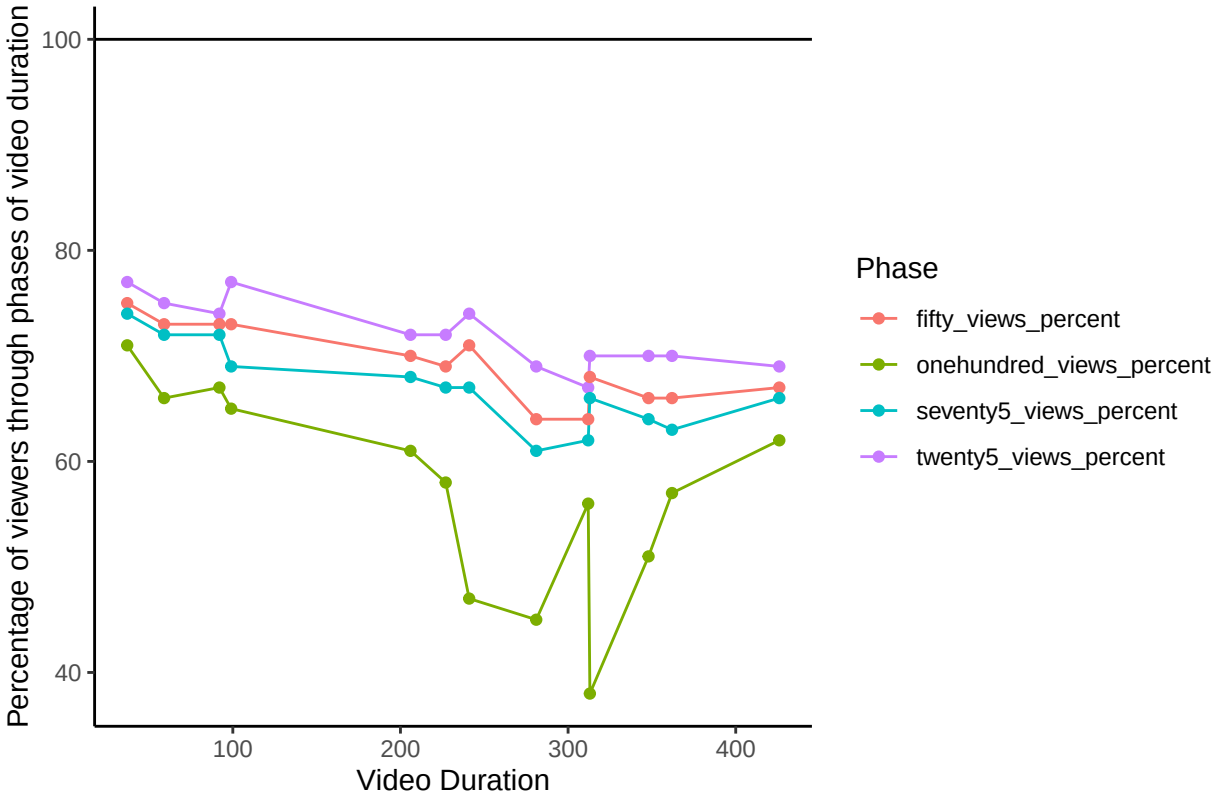
Figure XX: Total percentage of viewers for Video Topics



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



5. Evaluation

5.1 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.

Concept description aims at an understandable description of concepts or classes. The purpose is not to develop complete models with high prediction accuracy, but to gain insights. For instance, a company may be interested in learning more about its loyal and disloyal customers. From a concept description of these concepts (loyal and disloyal customers) the company might infer what could be done to keep customers loyal or to transform disloyal customers to loyal customers. Concept description has a close connection to both segmentation and classification. Segmentation may lead to an enumeration of objects belonging to a concept or class without providing any understandable description. Typically, segmentation is carried out before concept description is performed. Some techniques—conceptual clustering techniques, for example—perform segmentation and concept description at the same time.

Given more time, we could do further analysis to gain insights about the learners' from continents which have stayed engaged.

5.2 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)

5.3 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 (do we need or is it for above?)

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.

6. References