# CSC8631 Report

Selina So

23/11/2021

# 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, 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). 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).*

## 1.2 Assess situation (e resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and project plan.?? This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and in developing the project plan.)

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 provides information on the learners and the course. The remaining 7 pdf files provides 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, 2018, futurelearn site). The 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 were collected throughout the course, between mid 2016 - mid 2018. All runs consists of the following datafiles: *'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 not containing the *'Team-members', 'Video-stats'* datafiles, and run 2 not containing the *'Video-stats'* datafile.

### 1.2.1 Inventory of Resources

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

#### 1.2.1.1 Software Sources

- R
- RStudio
- Git

#### 1.2.1.2 Sources of Data and Knowledge

- The CRISP-DM methodology (Cross-Industrie Standard Process for Data Mining) will be applied to structure the project life cycle (link the CRISP-DM guide).
- 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 utilized for expert domain knowledge and technical support (as well as to provide stakeholder guidance from a business perspective - add?)

- 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: a R package to automate project file structure
- RMarkdown: a R package to produce the report
- ggplot2: a R package to produce data visualizations
- Git: will be 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.

**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 identifying data observed 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 of the course.
- There were no descriptions for the data, therefore assumptions will be made as to what the data means.

**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*. (If time allows, then the *Modelling* phase will be included too - include?)

## 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 the Newcastle University and FutureLearn to potentially modify and improve the course content to achieve optimise learner engagement and retention. (This is achieved by identifying correlations between the factors and the learners????)

## 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 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??), identify potential trends to 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 interesting 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 (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 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 being 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 will provide insight into the nature and quality of available data, and will enable fast decisions on the next steps in the analysis.) include?? personnel??*

## 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. 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.)

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 were measured several months apart from each other, between mid 2016 - mid 2018. Below is a brief description of each set of 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)).* :

- *'Archetype-survey-responses'*: The data files consists of two sets of Id-related columns, a datetime column and a categorical column *'archetype'*. As no descriptions was provided, therefore it is difficult to deduce the real meaning of the archetype data without further information. In addition, the datafiles for runs 1 to 3 are either empty or incomplete.
- *'Enrolments'*: The data files contains information of learners' ID, enrollment and unenrollment datetime. In addition, categorical data on the gender, country, age, education and employment sta-

tus and detected country. There are information on datetime of *'fully_participated_at'* and *'purchased_statement_at'*, however without further information it is unclear what these columns mean. There are many *'Unknown'* in the data files.

- *'Leaving-survey-responses'*: The data files contains information about learner ID, leaving date, and learners' feedback, which was given as what is assumed to be pre-set selection of feedback options. It contains information on the last step completed when response was provided, however are incomplete as many rows contain missing data. The datafiles 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 corrected. There is a *'response'* column which contains a selection of numerical values and a *'cloze_response'* which appears to be empty, therefore one cannot interpret what this column means. 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.
- *'Step-activity'*: This data file contains information on the learner_id and the datetime of when they had first visit and last completed a particular step. There are missing data under the *'last_completed_at'* column.
- *'Weekly-sentiment-survey-response'*: the data files contain an ID column, with the datetime of the individuals' responses. It is unclear what does 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 run 1 to 4 data files are empty, run 5 contains only 1 incomplete data and the remaining runs (6 and 7) contains a mixture of incomplete and unstructured text data for the free-text *'reason'* column.
- *'Team-members'* (not included in run 1): the data files contain Ids and the team role and user role of the individuals. The *'first_name'* and *'last_name'* columns have been anonymised to remove the names of the individuals.
- *'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, devices and features used to view the videos, how long have individuals watched the videos and the percentage of viewers from different continents.

.

### 2.1.1 Initial data file selection

The initial data file(s) of interest was(were) 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 constructive information for the course provider. The *'reason'* column was interesting as they contained free text therefore provided a direct source of feedback that could be analysed using Natural Language Processing (NLP). However, after closer observation, 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 contains pre-selected responses, which indicates that the learners were only allowed to select from a very narrow range of opinions, therefore it 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 can be assumed that the responses were from 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 use 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 utilize 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 are available. There appears to be no missing values from the data, therefore the data files for '*video.stats*' are very complete.

### 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)

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, therefore one can deduce the topic of the video. The remaining columns (columns 3 - 28) contain numerical data, of which some contains percentage values. Care will have to be taken when merging the runs together for the columns containing percentage values. The weighted average percentages will be calculated to account for the relative frequency of some factors in the dataset (reword!!), therefore the weighted average percentage will be more accurate than calculating the average percentage.

The dataset contains information on the video topics and the 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 satisfies the business objective requirements of the project.

*The data surrounding the percentage of devices (and features) used for the videos will not be selected for this study (due to time constraints, etc) (Do I include the data on using captions...) a combination of columns will be selected for each individual analysis.*

```
summary(run3)
```

```
##  step_position      title           video_duration  total_views
##  Min.   :1.100   Length:13          Min.   : 37    Min.   : 446
##  1st Qu.:1.190   Class :character   1st Qu.: 99    1st Qu.: 484
##  Median :2.110   Mode  :character   Median :241    Median : 680
##  Mean   :2.113                      Mean   :231    Mean   : 739
##  3rd Qu.:3.100                      3rd Qu.:313    3rd Qu.: 755
##  Max.   :3.200                      Max.   :426    Max.   :1659
##  total_downloads  total_caption_views total_transcript_views   viewed_hd
##  Min.   : 34.00   Min.   : 1.000      Min.   : 66.0          Min.   :  4.00
##  1st Qu.: 42.00   1st Qu.: 2.000      1st Qu.: 87.0          1st Qu.:  8.00
##  Median : 50.00   Median : 4.000      Median :110.0          Median : 13.00
##  Mean   : 58.15   Mean   : 6.923      Mean   :122.8          Mean   : 50.31
##  3rd Qu.: 63.00   3rd Qu.: 5.000      3rd Qu.:147.0          3rd Qu.: 28.00
##  Max.   :113.00   Max.   :36.000      Max.   :221.0          Max.   :434.00
##  viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
##  Min.   :70.39       Min.   :66.31      Min.   :64.72
##  1st Qu.:72.85       1st Qu.:71.92      1st Qu.:69.27
##  Median :73.72       Median :73.76      Median :71.92
##  Mean   :74.26       Mean   :72.95      Mean   :71.11
##  3rd Qu.:75.48       3rd Qu.:74.84      3rd Qu.:73.42
##  Max.   :78.45       Max.   :75.64      Max.   :74.93
##  viewed_fifty_percent viewed_seventyfive_percent viewed_ninetyfive_percent
##  Min.   :61.52        Min.   :59.57              Min.   :56.38
##  1st Qu.:65.38        1st Qu.:63.44              1st Qu.:61.59
##  Median :69.45        Median :66.53              Median :62.94
##  Mean   :68.48        Mean   :66.64              Mean   :64.25
##  3rd Qu.:70.40        3rd Qu.:68.17              3rd Qu.:66.43
##  Max.   :73.49        Max.   :73.06              Max.   :72.09
##  viewed_onehundred_percent console_device_percentage desktop_device_percentage
##  Min.   :38.57             Min.   :0.0000            Min.   :77.35
##  1st Qu.:49.40             1st Qu.:0.1300            1st Qu.:79.11
##  Median :57.36             Median :0.1500            Median :80.32
##  Mean   :56.34             Mean   :0.1508            Mean   :80.06
##  3rd Qu.:63.71             3rd Qu.:0.2100            3rd Qu.:80.99
##  Max.   :69.45             Max.   :0.2200            Max.   :82.29
##  mobile_device_percentage tv_device_percentage tablet_device_percentage
```

```
##  Min.   : 6.200          Min.   :0.000000     Min.   : 7.72
##  1st Qu.: 7.020          1st Qu.:0.000000     1st Qu.:10.55
##  Median : 8.710          Median :0.000000     Median :10.95
##  Mean   : 8.791          Mean   :0.004615     Mean   :10.52
##  3rd Qu.: 9.850          3rd Qu.:0.000000     3rd Qu.:11.17
##  Max.   :13.260          Max.   :0.060000     Max.   :11.91
##  unknown_device_percentage europe_views_percentage oceania_views_percentage
##  Min.   :0                 Min.   :55.15           Min.   :2.240
##  1st Qu.:0                 1st Qu.:64.90           1st Qu.:3.170
##  Median :0                 Median :65.60           Median :3.240
##  Mean   :0                 Mean   :64.73           Mean   :3.265
##  3rd Qu.:0                 3rd Qu.:66.25           3rd Qu.:3.550
##  Max.   :0                 Max.   :67.25           Max.   :4.070
##  asia_views_percentage north_america_views_percentage
##  Min.   : 8.24         Min.   :10.65
##  1st Qu.: 9.11         1st Qu.:11.21
##  Median : 9.51         Median :11.43
##  Mean   :10.03         Mean   :11.45
##  3rd Qu.: 9.92         3rd Qu.:11.67
##  Max.   :16.09         Max.   :12.21
##  south_america_views_percentage africa_views_percentage
##  Min.   :1.650                  Min.   : 5.170
##  1st Qu.:2.120                  1st Qu.: 5.560
##  Median :2.330                  Median : 6.200
##  Mean   :2.424                  Mean   : 6.445
##  3rd Qu.:2.660                  3rd Qu.: 6.380
##  Max.   :3.750                  Max.   :10.310
##  antarctica_views_percentage
##  Min.   :0
##  1st Qu.:0
##  Median :0
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
```

**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)**

These two columns will be combined to allow quick reference to the order of which the videos appears throughout the course.

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

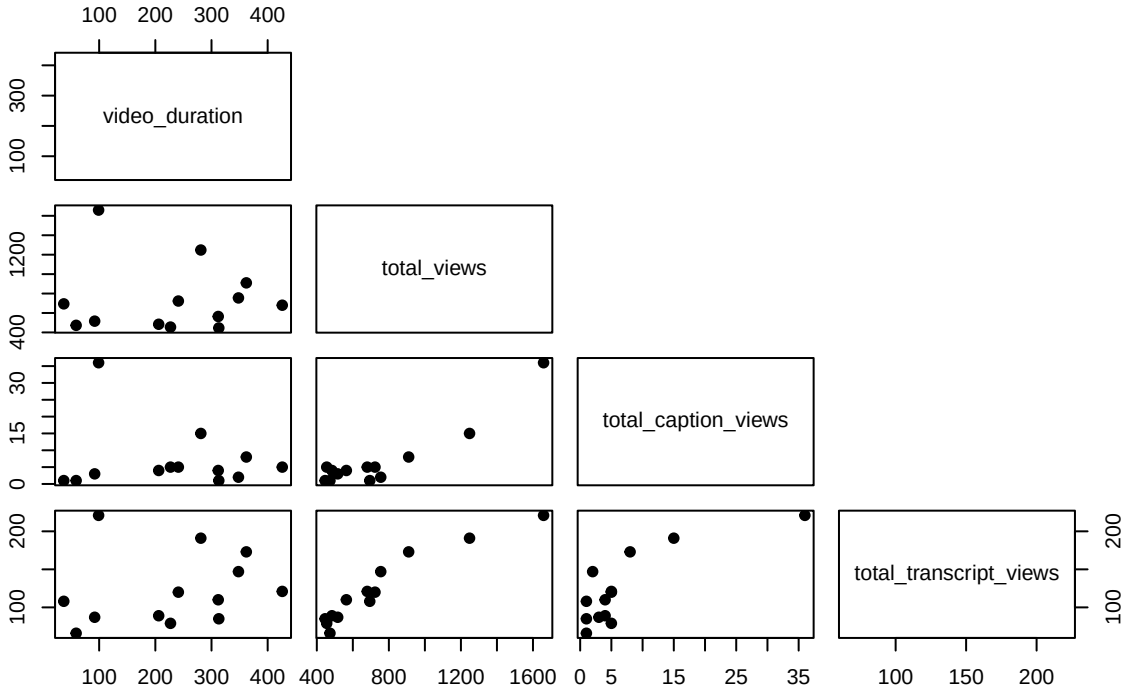
**Total number of viewing and features used through the duration of videos (DELETE??)**

The scatterplot matrix 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?]

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.

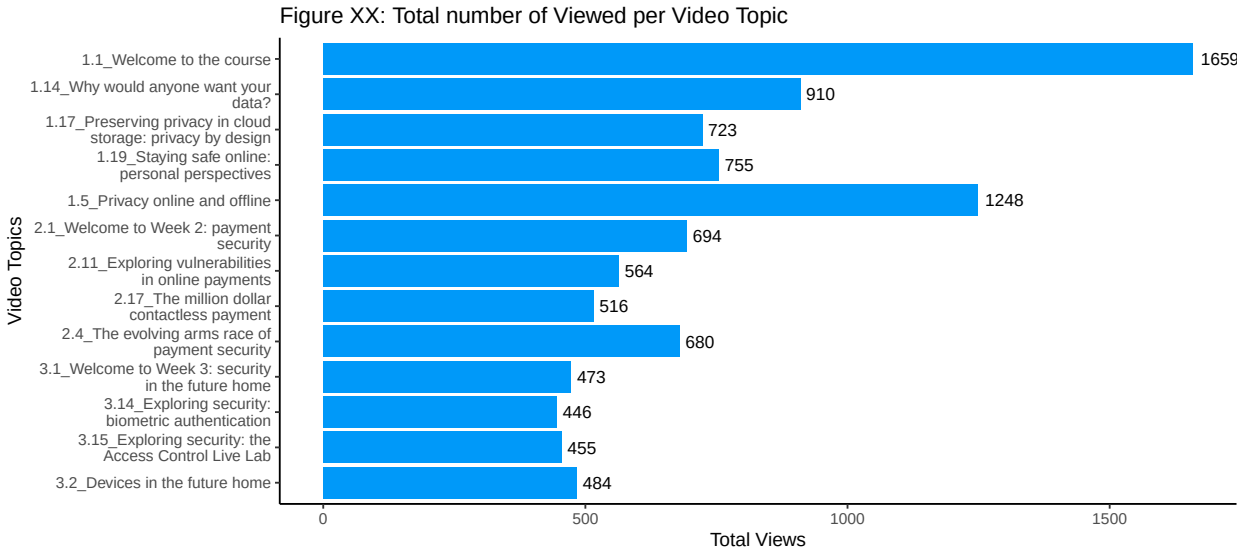
## Figure XX: Total Views and Features used through Video duration




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.

### Number of viewers for each video topic

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



#### 2.3.1 Hypothesis

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?

### 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 from the other runs could be performed as ease.

Columns *'unknown_device_percentage'* and *'antarctica_views_percentage'* contain values of 0. This can be assumed that the data provider has found no learners using unknown devices and there are no learners from the Antarctica, which is reasonable as according to the *World Population Review*, there is roughly 1000-4000 seasonal residents in the Antarctica (worldpopulationreview.com/continents/antarctica-population). Therefore we can assume that the values of 0 are accurate under the *'antarctica_views_percentage'* column.

## 3. Data Preparation - Q1 Does the duration of the videos have an impact on the viewing rates across different continents?

### Run 3 data prep

### 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.)

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.

**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.

[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.

**All runs data prep info**

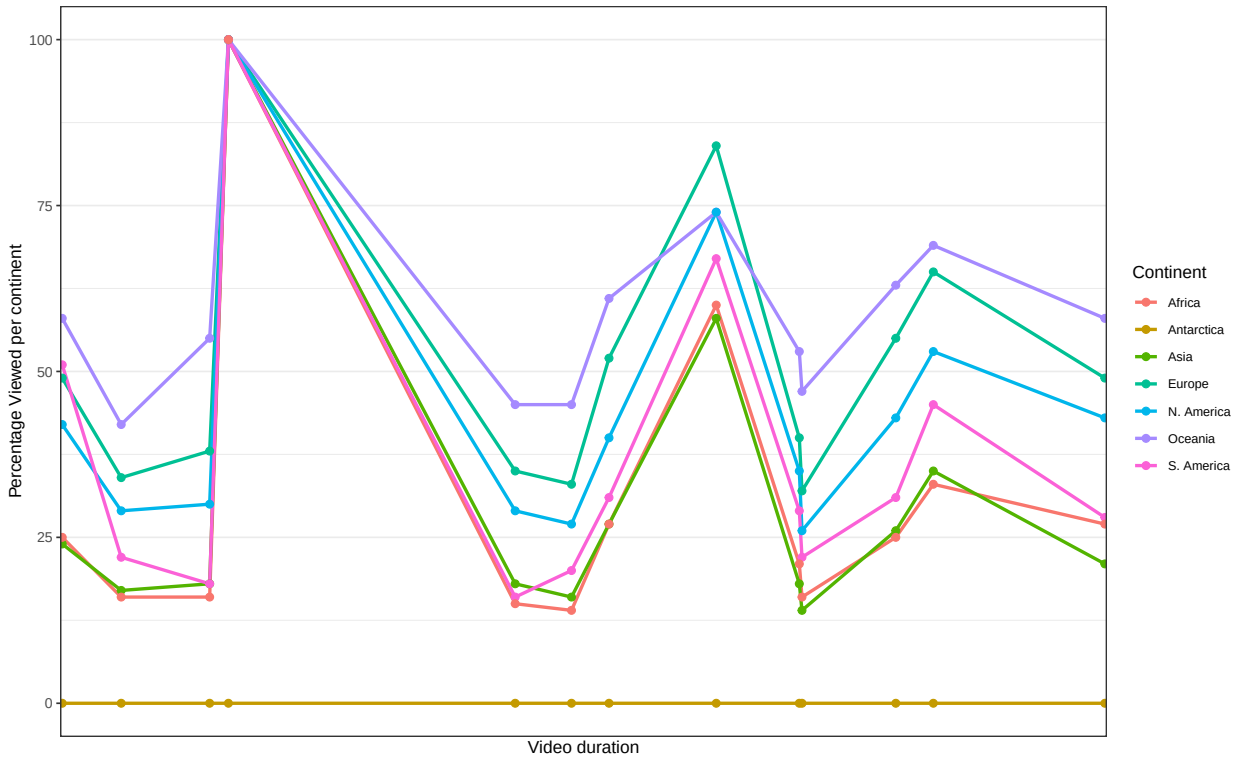
## 4. Modelling - Question 1. Does the duration of the videos have an impact on the viewing rates across different continents?

**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 modelling**

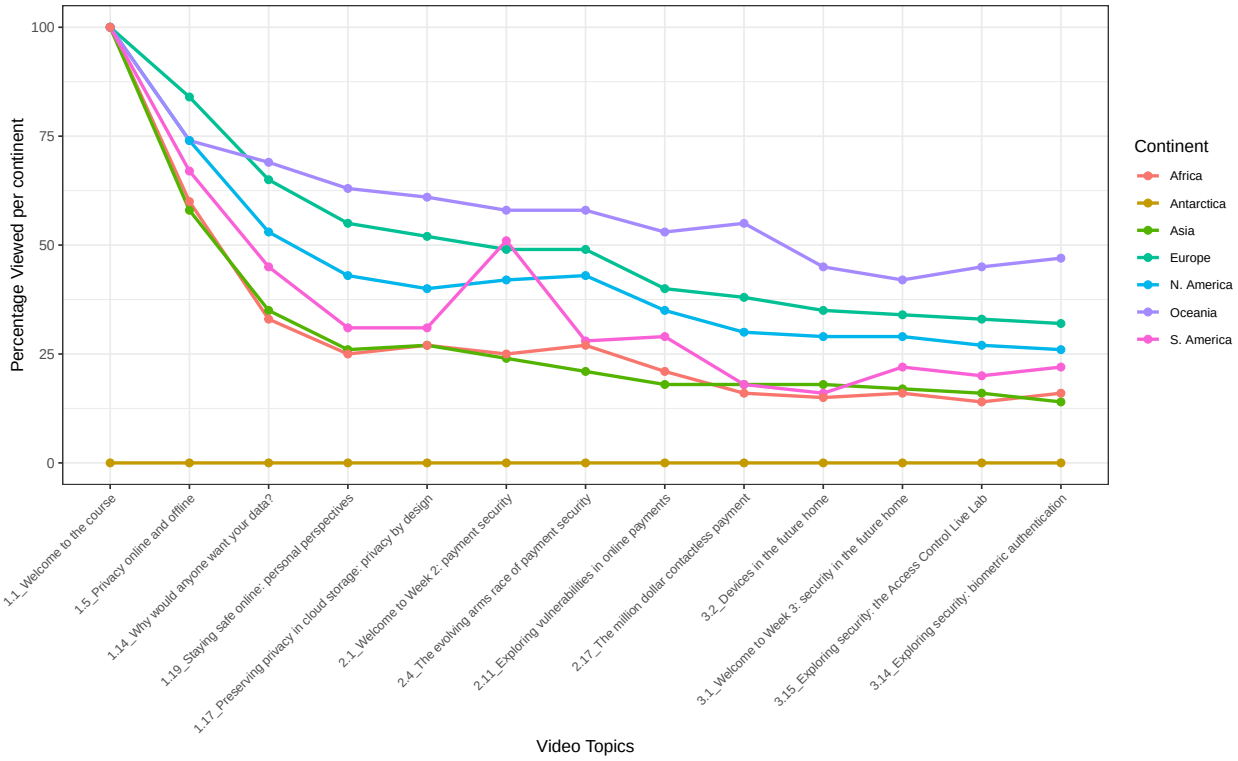
**Figure XX: Percentage of viewers for Video duration**




no correlation

### line drop out of continent vs topic

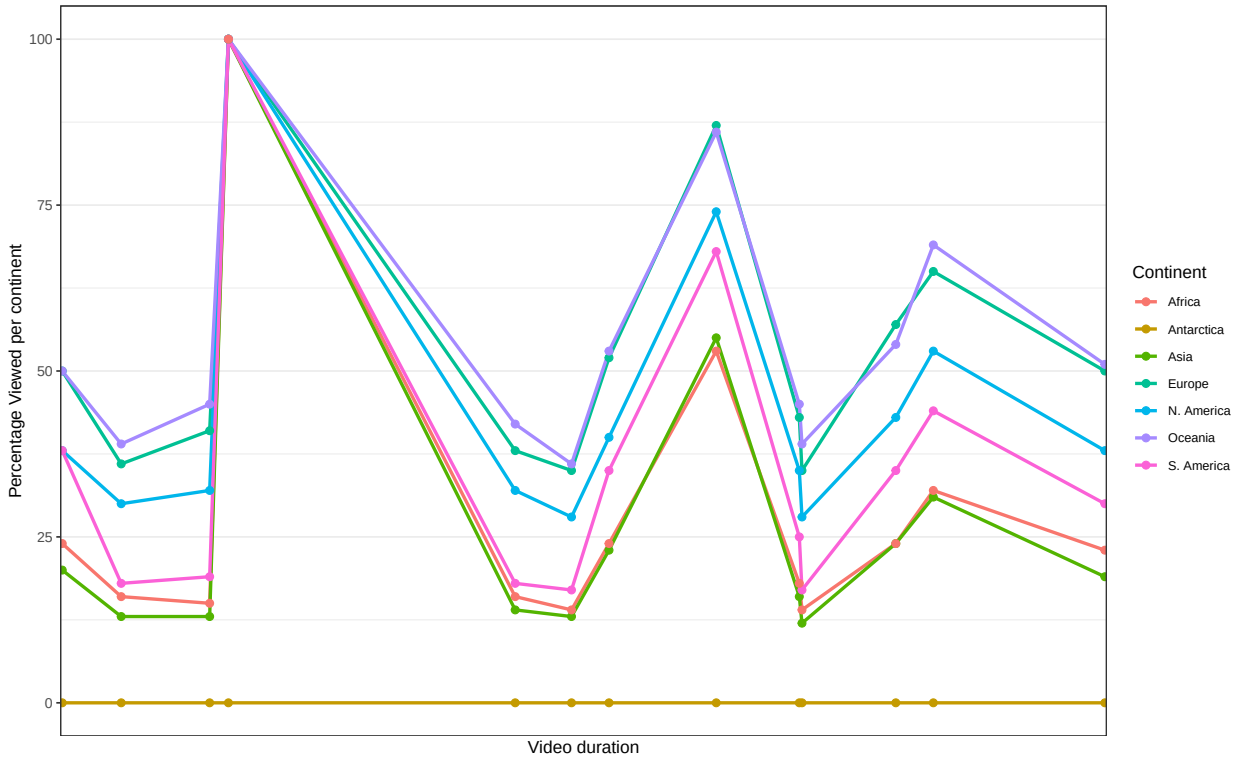
**Figure XX: Total percentage of viewers from high to lowest in run 3**




## All runs modelling

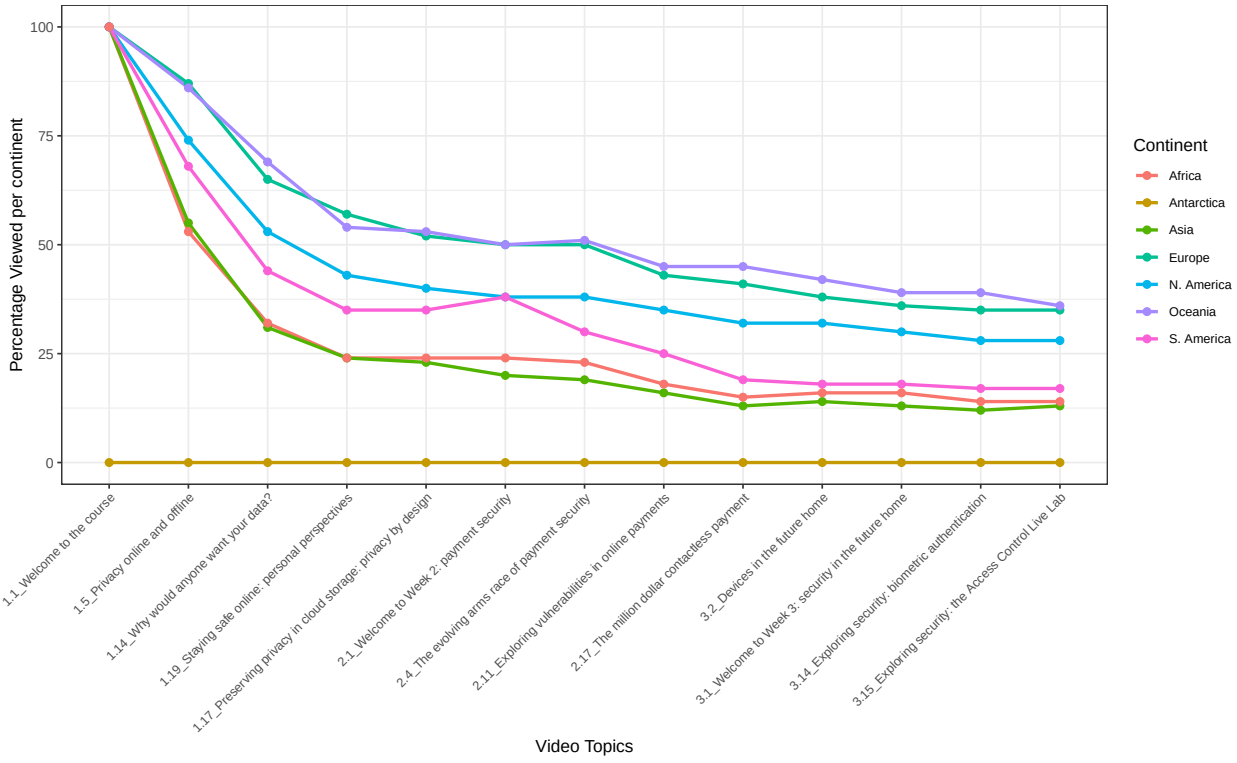
prove with all runs

**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**



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

### Run3 Data prep

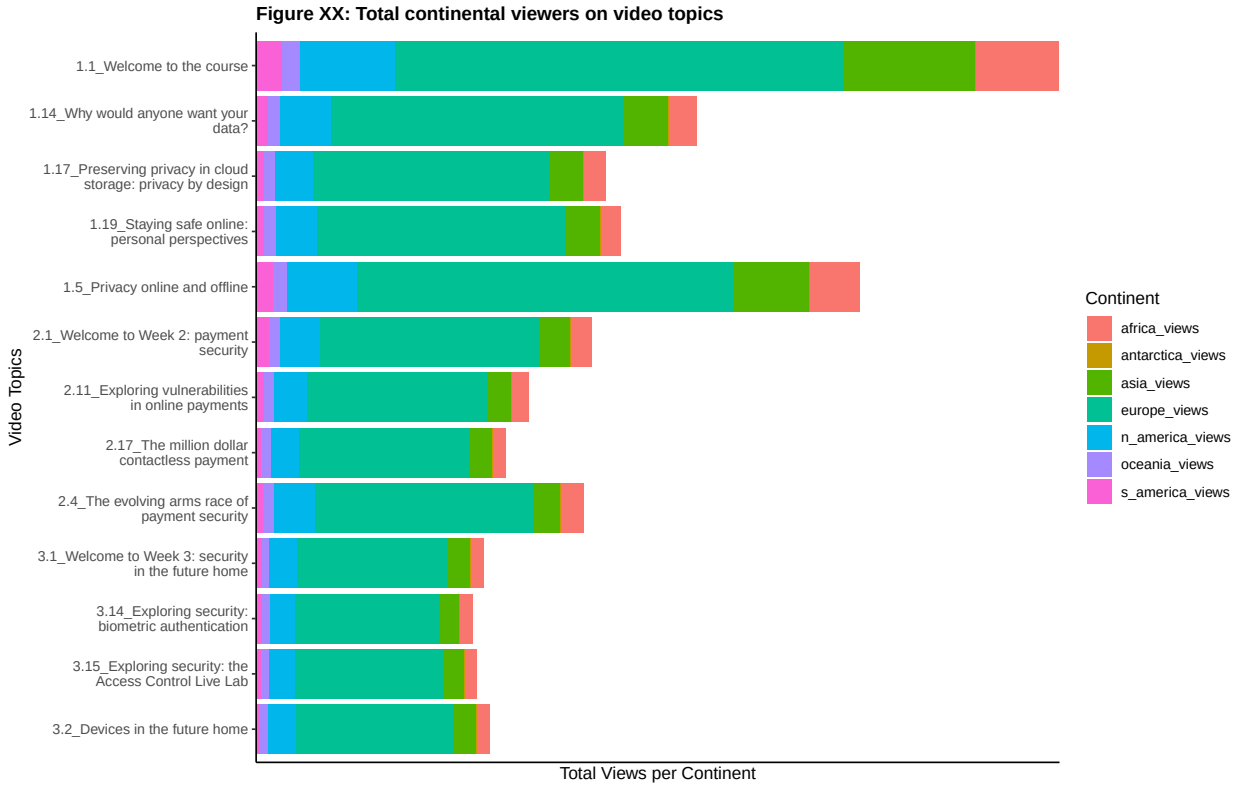
### All run data prep

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

### Run 3 modelling

#### Worldwide views of videos

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




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.

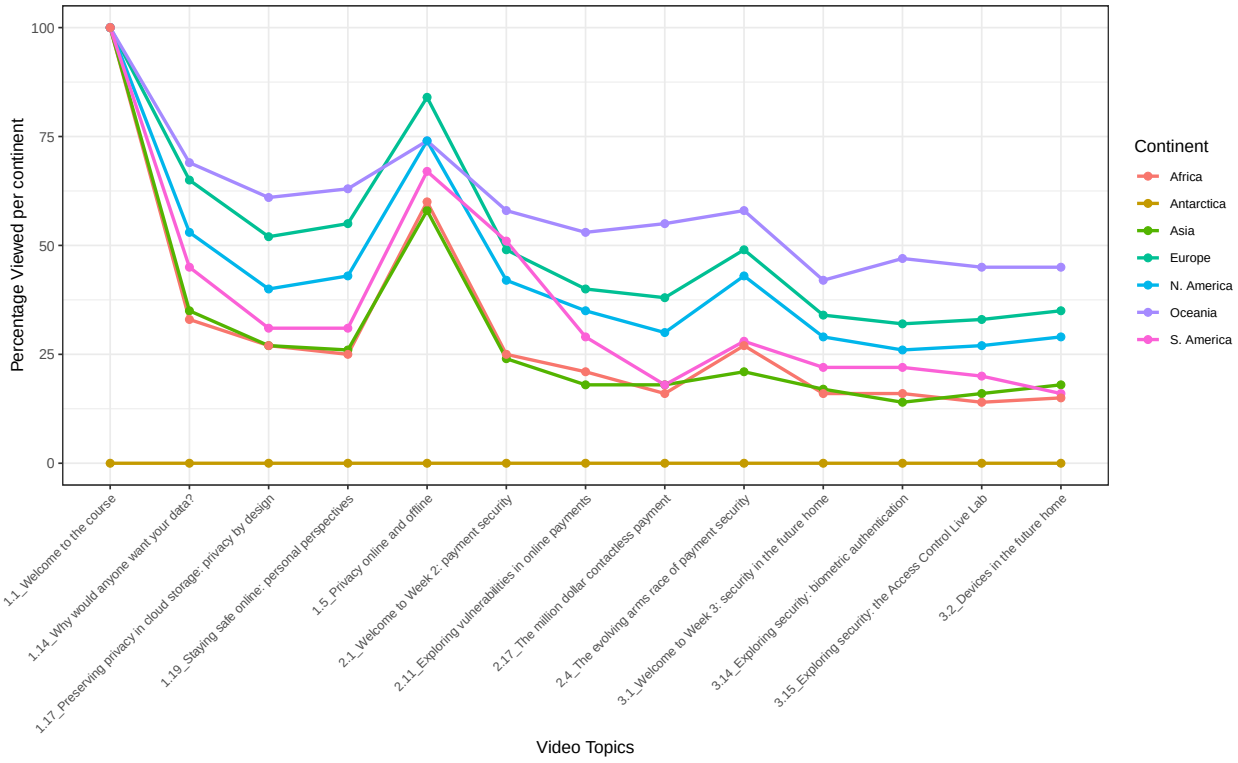
In addition, the relative views from each continent appears to be stable. There appears to be some outliers from the far left column of plots. It is unlikely to be related to the duration of the videos because the outliers

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**

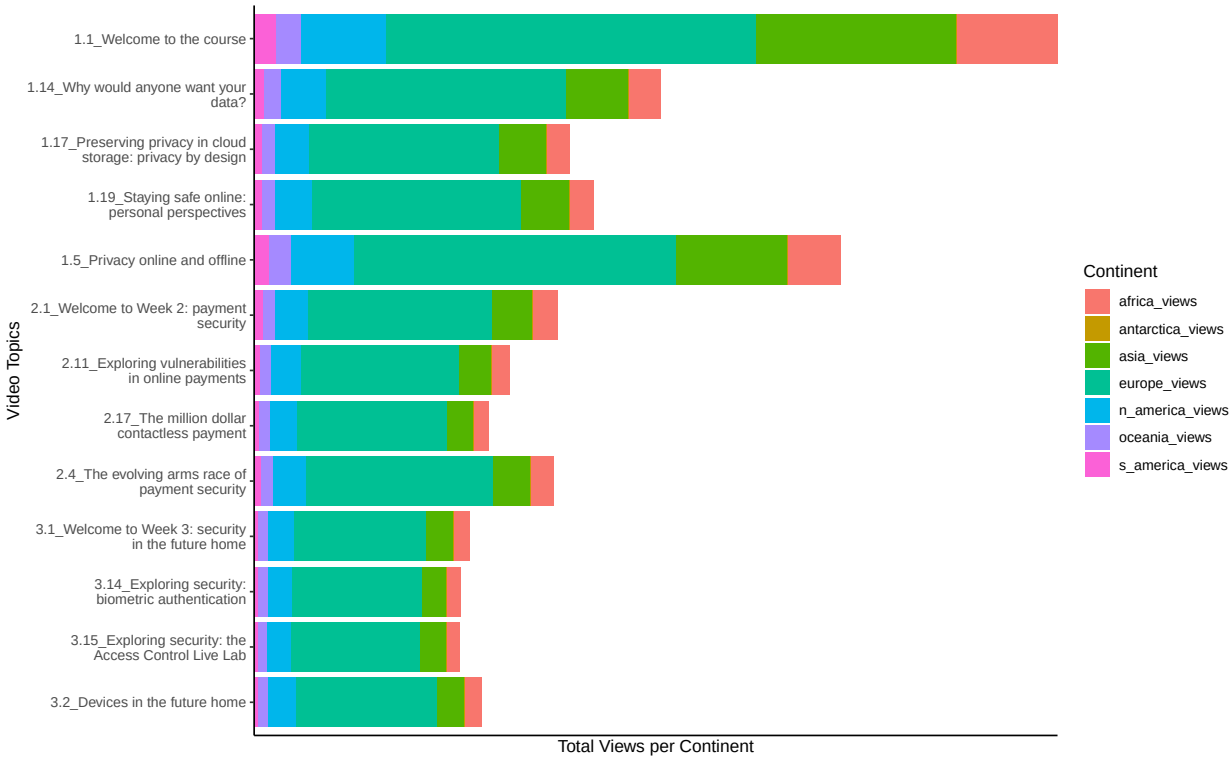



The graph shows that for certain continents (Europe, North America and Oceania), there appears to be more views, therefore more engagements, on the topics within week 2 block of the course. On the other hand, there were more engagements from the learners of Africa and Asia continent, then a steady drop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

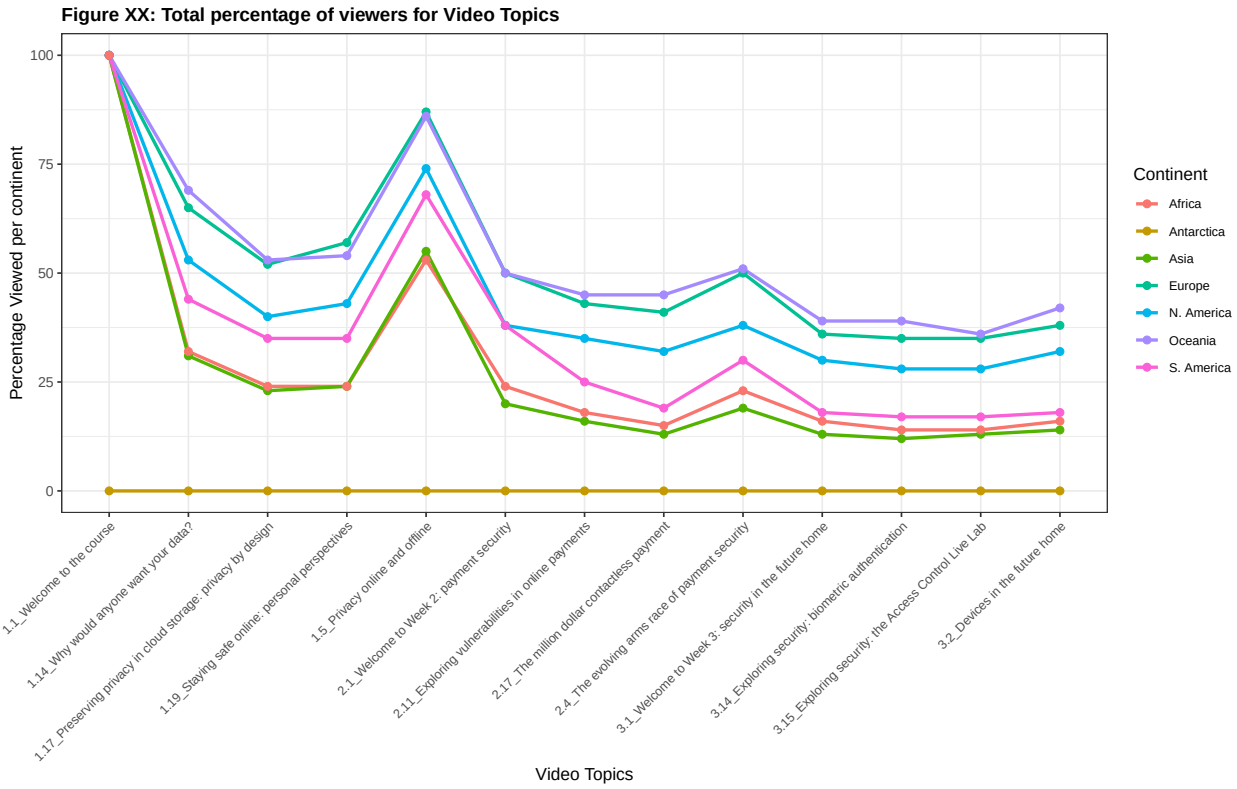
# All run modelling

## 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




# 7. Data Preparation -Question 3 Is there a correlation between duration of videos and drop out rate of the learners?

## Run 3 Data prep

## All run data prep

# 8. Modelling -Question 3 Is there a correlation between duration of videos and drop out rate of the learners?

## Run 3 Modelling

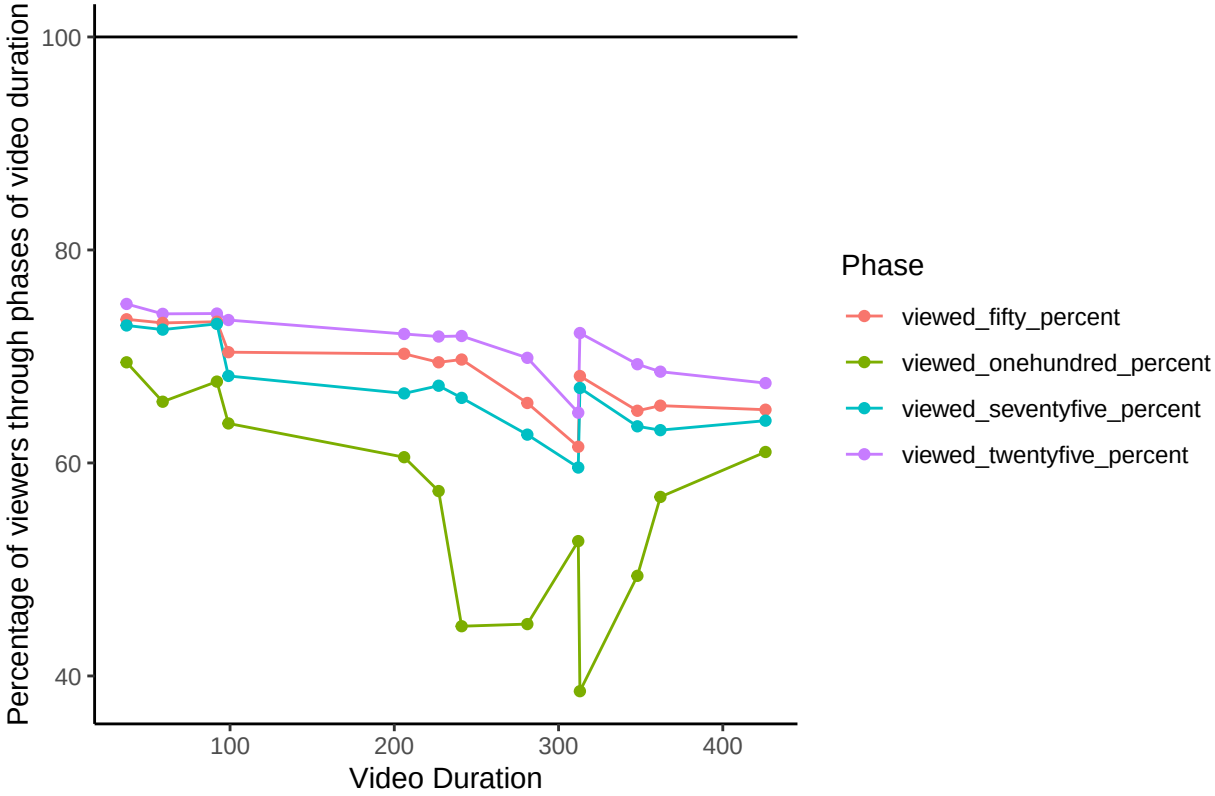
### Percentage viewed whole duration of videos/ Number of viewers at 100 percent duration of the videos

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**

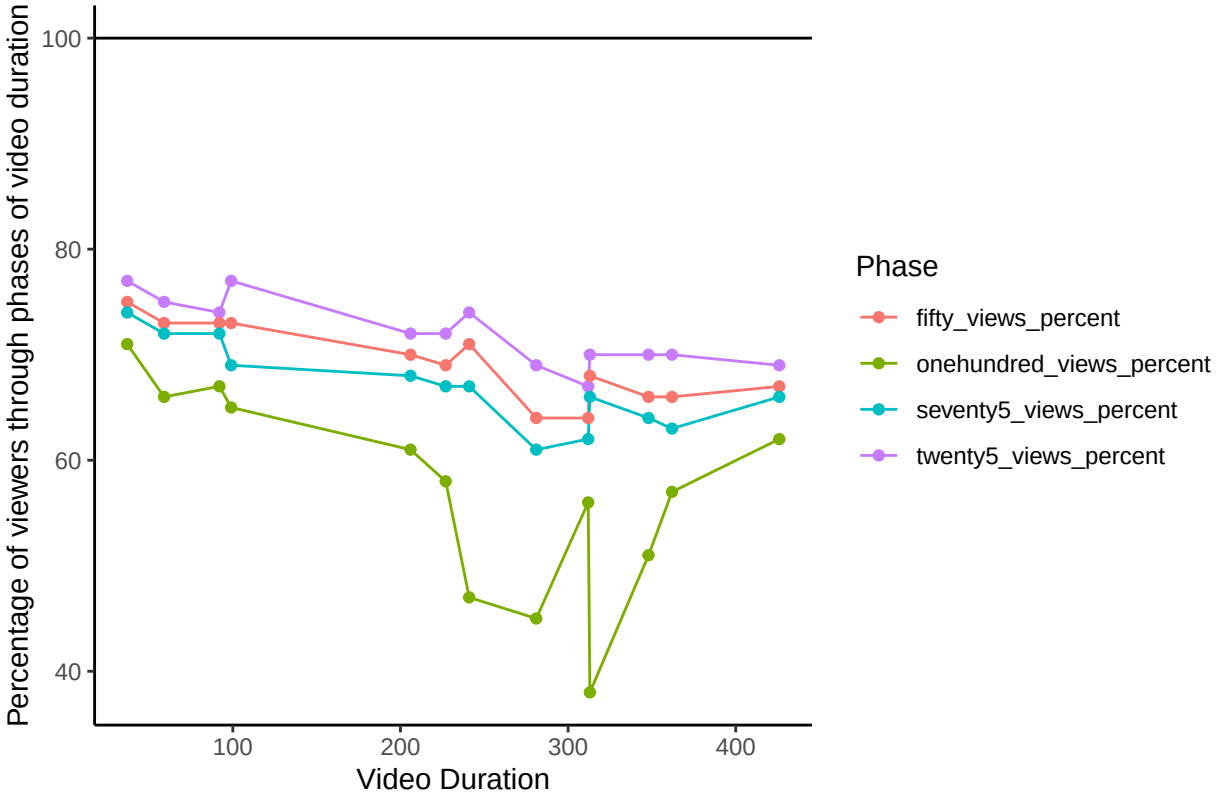



There is potentially a pattern observed on the relationship between the video duration and the number of viewers who have stayed for the whole duration of the videos. However the data is too noisy due to the relatively low number of learners for some continents, and all the runs will need to be included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

### All runs modelling

**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