CSC8631 Assignment Report

D Walmsley, C1053068

15/11/2021

Introduction

The purpose of this report is to analyse data collected from an online course ran by Futurelearn (www.futurelearn.com). The course is ran in collaboration with Newcastle University. The report will demonstrate findings from a data set across 7 runs provided by the the university. Using the CRISP-DM methodology to provide a clear structure to the analysis. This report will be broken into sections that are the key steps of the CRISP DM process. The sections are *Business Understanding, Data Understanding, Data preparation, Modelling, Evaluation and Development.*

I.Business Understanding

Newcastle University is a Russell group university offering high quality courses to students from around the world. Working with online partner Future Learn, the university are offering an online cyber security course. The course provider must show awareness of factors impacting on cohort recruitment such as demographic, age, and gender. The provider must review key performance indicators and analyse how appealing the course is to specific user groups. Course retention and understanding of students motive to enroll must be evaluated. In addition the provider must review barriers and challenges that sub groups of the cohort have experienced and how this may have prevented them completing the course. This will allow the provider to address issues in future and develop a sustainable product. Academic outcomes of students are of interest to the course providers as this allows them to identify what resources are having the greatest impact on learning, but also allows them to identify areas where improved resources would impact on student outcomes. Finally the university must consider how it could keep up with changes in new technology and how students access the content provided. Accessibility is a key factor and how the provider can ensure a course or product is more accessible to a wider audience, encouraging a greater enrolment globally.

The data files from 7 runs or repetitions of the course have been acquired. These files will be analysed to determine key links and trends within the data. The strength of the data collected will determine the strength of potential outcomes of this project. The data from the repetitions or reps needs to have similarities to ensure that files and scenarios are comparable. Clear identifiable relationships between the data files will help construct stronger lines of enquiry. This will lead to outputs that will have a greater impact on informing future decision making within the process. From the third course rep information is included about video resources used within the course material and how and where the videos were accessed. Although this video data was not accessible for reps one and two, this may be a way to compare if video files have had an impact since their introduction in rep three of the course. R scripts will be written within R studio software package to ensure that any work undertaken can be easily reproduced.

This project will look to meet the aims and goals set out in the business objectives. By exploring the data, the intention of this project is to identify possible links between the introduction of video resources and the completion of the course. Is it possible to identify a groups of students who are more likely to complete the course and is there a clear reason for this? Are there clear indications why certain students fully complete the course or actually drop out of the course? By using the data to identify clear strengths and weaknesses of the course, this should inform decision making on what resources need strengthening or where to better support students in order to improve retention and academic outcomes.

R scripts will be used in order to clean and collate data. The use of R language allows the data not only cleaned, but also means that the data can be displayed in clear visual formats that support what large datasets are saying. The R scripts will be accessed through R studio software that allows the scripts to be stored within a Project Template structure. This means managing and accessing scripts is both organised and accessible. R studio then allows these scripts, visual data representations and charts to be merged together into a report. This can be exported as a PDF file format that can be easily distributed.A key success criteria of this project is reproducibility and the CrispDM methodology gives a clear structure to the process and reproduction of the work taken place within the project. This methodology will be structured through the Project Template file structure to ensure that other users could replicate the work that had been undertaken.

A Gitlog will be included as part of the project to show version control techniques and ensure that changes to the work are recorded at regular intervals in the event of errors occurring. Version control allows all work to be rolled back to staged points in time.

II. Data Understanding

There are 53 .csv files provided by the university, spread across seven repetitions of the course. The files all start with the prefix cyber.security then the number of the repetition of the course (cyber.security.1, cyber.security.2 etc). These are then split into eight possible datasets in the final five repetitions of the course and only seven possible datasets from the first two repetitions. This is because there is no "videostats" files until repetition 3 onwards. The files provided by the university are as follows:

Archetype survey responses
Enrolments
Leaving survey responses
Question responses
Step activity
Team members
Video stats
Weekly sentiment survey responses

Describe data:

Examine the data and document its surface properties like data format, number of records, or field identities.

Unfortunately nine of the data files included no data within the files, this is a weakness in the data collection process. These files were:

* Cyber.security.1_archetype.survey.responses.csv
* Cyber.security.1_leaving.survey.responses.csv
* Cyber.security.1_weekly.sentiment.survey.responses.csv
* Cyber.security.2_archetype.survey.responses.csv
* Cyber.security.2_leaving.survey.responses.csv
* Cyber.security.2_weekly.sentiment.survey.responses.csv
* Cyber.security.3_leaving.survey.responses.csv
* Cyber.security.3_weekly.sentiment.survey.responses.csv
* Cyber.security.4_weekly.sentiment.survey.responses.csv

Looking at the enrolments data, the enrolment files were combined to see the total number of students that had enrolled over the seven repetitions.

```r
glimpse(combined_enrolments)
```

```
## Rows: 37,296
## Columns: 14
## $ learner_id                 <chr> "160d6600-ea0e-4568-bfa9-5d7cd5b8e61b", "4dc22~
## $ enrolled_at                <chr> "2016-08-10 14:28:49 UTC", "2016-05-24 17:34:3~
## $ unenrolled_at              <chr> "", "2018-10-30 20:20:51 UTC", "", "", "", "",~
## $ role                       <chr> "learner", "learner", "learner", "learner", "l~
## $ fully_participated_at      <chr> "", "", "2016-09-22 16:56:03 UTC", "", "", "20~
## $ purchased_statement_at     <chr> "", "", "", "", "", "", "", "", "", "", "", ""~
## $ gender                     <chr> "Unknown", "male", "Unknown", "Unknown", "Unkn~
## $ country                    <chr> "Unknown", "PE", "Unknown", "Unknown", "Unknow~
## $ age_range                  <chr> "Unknown", "46-55", "Unknown", "Unknown", "Unk~
## $ highest_education_level    <chr> "Unknown", "university_degree", "Unknown", "Un~
## $ employment_status          <chr> "Unknown", "working_part_time", "Unknown", "Un~
## $ employment_area            <chr> "Unknown", "teaching_and_education", "Unknown"~
## $ detected_country           <chr> "GB", "PE", "NG", "UG", "IM", "NO", "GB", "GB"~
## $ rep                        <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
```

To combine the tables, the following code was used to combine the cyber security files. This script takes the original cyber.security file and renames it with a name that is easier to remember and reuse later in enrolments_yr1 for example, then adds a column that indicates the repetition or run of the course. It then combines the enrolment_yr files into one dataframe that can be used later in the project.

```r
enrolments_yr1 = cyber.security.1_enrolments %>%
  mutate(rep = "rep 1")
enrolments_yr2 = cyber.security.2_enrolments %>%
  mutate(rep = "rep 2")
enrolments_yr3 = cyber.security.3_enrolments %>%
  mutate(rep = "rep 3")
enrolments_yr4 = cyber.security.4_enrolments %>%
  mutate(rep = "rep 4")
enrolments_yr5 = cyber.security.5_enrolments %>%
  mutate(rep = "rep 5")
enrolments_yr6 = cyber.security.6_enrolments %>%
  mutate(rep = "rep 6")
enrolments_yr7 = cyber.security.7_enrolments %>%
  mutate(rep = "rep 7")

combined_enrolments = rbind(enrolments_yr1, enrolments_yr2, enrolments_yr3, enrolments_yr4, enrolments_
```

From the enrolments data every learner had a unique learner ID and these learner ids could be found in the archetype, leaving response survey, question response survey, step activity, team members and weekly sentiment survey tables. These relational data tables make combing the files and looking for trends easier and show the individuals learning journey as a whole. Within the enrolment data there is a number of blank cells, as well as cells that contain "Unknown". There are examples of this within the columns gender, age range, highest education level and employment area.

```r
sum(combined_enrolments$gender == "Unknown")
```

```
## [1] 33137
```

```r
sum(combined_enrolments$age_range == "Unknown")
```

```
## [1] 33268
```

```r
sum(combined_enrolments$highest_education_level == "Unknown")
```

```
## [1] 33161
```

```r
sum(combined_enrolments$employment_area == "Unknown")
```

```
## [1] 34090
```

There are also a number of cells that were empty or blank cells. These cells appeared in the purchased statement, unenrolled and fully participated column. T

```r
sum(combined_enrolments$purchased_statement_at == "")
```

```
## [1] 37007
```

```r
sum(combined_enrolments$unenrolled_at == "")
```

```
## [1] 33105
```

```r
sum(combined_enrolments$fully_participated_at == "")
```

```
## [1] 35142
```

The question response data shows each users responses to the staged quizzes throughout the course. It also contains whether their responses were correct and what answers were provided to the question. This data can be linked to other data files using the individual learner ids that are provided. As for the enrolment data files the same process was completed to create a large file to check for missing data. In following this process the data would know have the rep or repetition column added to the end.

```r
glimpse(qresponse_all)
```

```
## Rows: 176,463
## Columns: 11
## $ learner_id      <chr> "77454a73-6b8b-46a2-8dee-35f36b6c4fc1", "77454a73-6b8b~
## $ quiz_question   <chr> "1.7.1", "1.7.1", "1.7.1", "1.7.1", "1.7.1", "1.7.1", ~
## $ question_type   <chr> "MultipleChoice", "MultipleChoice", "MultipleChoice", ~
## $ week_number     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ step_number     <int> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ~
## $ question_number <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ response        <chr> "1,2", "1,2,3", "1,2,3", "1,2", "2,3", "1,2,3", "1,2,3~
## $ cloze_response  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ submitted_at    <chr> "2016-07-06 10:37:05 UTC", "2016-07-06 10:57:05 UTC", ~
## $ correct         <chr> "false", "true", "true", "false", "false", "true", "tr~
## $ rep             <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", ~
```

This data contains a large amount of NA responses in the close response column.

```
  sum(is.na(qresponse_all$cloze_response))
```

```
## [1] 176463
```

By running the code you can see the qresponse_all table contains 176,463 rows of data and in the cloze_response column there is 176,463 columns containing NA.

The introduction of video stats begins in repetition 3 of the course. Within the data files are the title of each video, the video duration, total number of views, the type of device it has been watched on, whether the file was downloaded, the percentage of the video a user has watched as well as the continent where it was watched.

```
glimpse(combined_videostats)
```

```
## Rows: 65
## Columns: 29
## $ step_position                  <dbl> 1.10, 1.14, 1.17, 1.19, 1.50, 2.10, 2.1~
## $ title                          <chr> "Welcome to the course", "Why would any~
## $ video_duration                 <int> 99, 362, 241, 348, 281, 37, 312, 92, 42~
## $ total_views                    <int> 1659, 910, 723, 755, 1248, 694, 564, 51~
## $ total_downloads                <int> 113, 77, 63, 62, 100, 48, 53, 42, 50, 3~
## $ total_caption_views            <int> 36, 8, 5, 2, 15, 1, 4, 3, 5, 1, 1, 5, 4~
## $ total_transcript_views         <int> 221, 173, 120, 147, 191, 108, 110, 87, ~
## $ viewed_hd                      <int> 58, 28, 16, 10, 41, 13, 434, 7, 16, 6, ~
## $ viewed_five_percent            <dbl> 76.97, 72.53, 73.72, 72.85, 78.45, 76.3~
## $ viewed_ten_percent             <dbl> 75.35, 70.88, 73.86, 71.92, 75.64, 75.0~
## $ viewed_twentyfive_percent      <dbl> 73.42, 68.57, 71.92, 69.27, 69.87, 74.9~
## $ viewed_fifty_percent           <dbl> 70.40, 65.38, 69.71, 64.90, 65.63, 73.4~
## $ viewed_seventyfive_percent     <dbl> 68.17, 63.08, 66.11, 63.44, 62.66, 72.9~
## $ viewed_ninetyfive_percent      <dbl> 66.43, 61.54, 61.83, 61.59, 59.05, 71.1~
## $ viewed_onehundred_percent      <dbl> 63.71, 56.81, 44.67, 49.40, 44.87, 69.4~
## $ console_device_percentage      <dbl> 0.06, 0.11, 0.14, 0.13, 0.00, 0.14, 0.1~
## $ desktop_device_percentage      <dbl> 78.60, 79.23, 79.67, 78.54, 80.37, 79.1~
## $ mobile_device_percentage       <dbl> 13.26, 10.33, 8.71, 9.40, 11.38, 9.37, ~
## $ tv_device_percentage           <dbl> 0.06, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0~
## $ tablet_device_percentage       <dbl> 7.72, 10.11, 11.07, 11.39, 7.93, 10.95,~
## $ unknown_device_percentage      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ europe_views_percentage        <dbl> 55.15, 65.38, 66.25, 67.15, 61.62, 64.2~
## $ oceania_views_percentage       <dbl> 2.29, 2.86, 3.18, 3.18, 2.24, 3.17, 3.5~
## $ asia_views_percentage          <dbl> 16.09, 10.22, 9.82, 9.27, 12.34, 9.37, ~
## $ north_america_views_percentage <dbl> 11.63, 11.32, 10.65, 10.99, 11.38, 11.6~
## $ south_america_views_percentage <dbl> 3.07, 2.53, 2.21, 2.12, 2.72, 3.75, 2.6~
## $ africa_views_percentage        <dbl> 10.31, 6.26, 6.36, 5.56, 8.17, 6.20, 6.~
## $ antarctica_views_percentage    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ rep                            <chr> "rep 3", "rep 3", "rep 3", "rep 3", "re~
```

The video stats files contain a large proportion of numerical data that could be further investigated later in the project.

```
sum(combined_videostats == " ")
```

```
## [1] 0
```

There were no empty cells from the 5 reps of video stats

III. Data Preparation

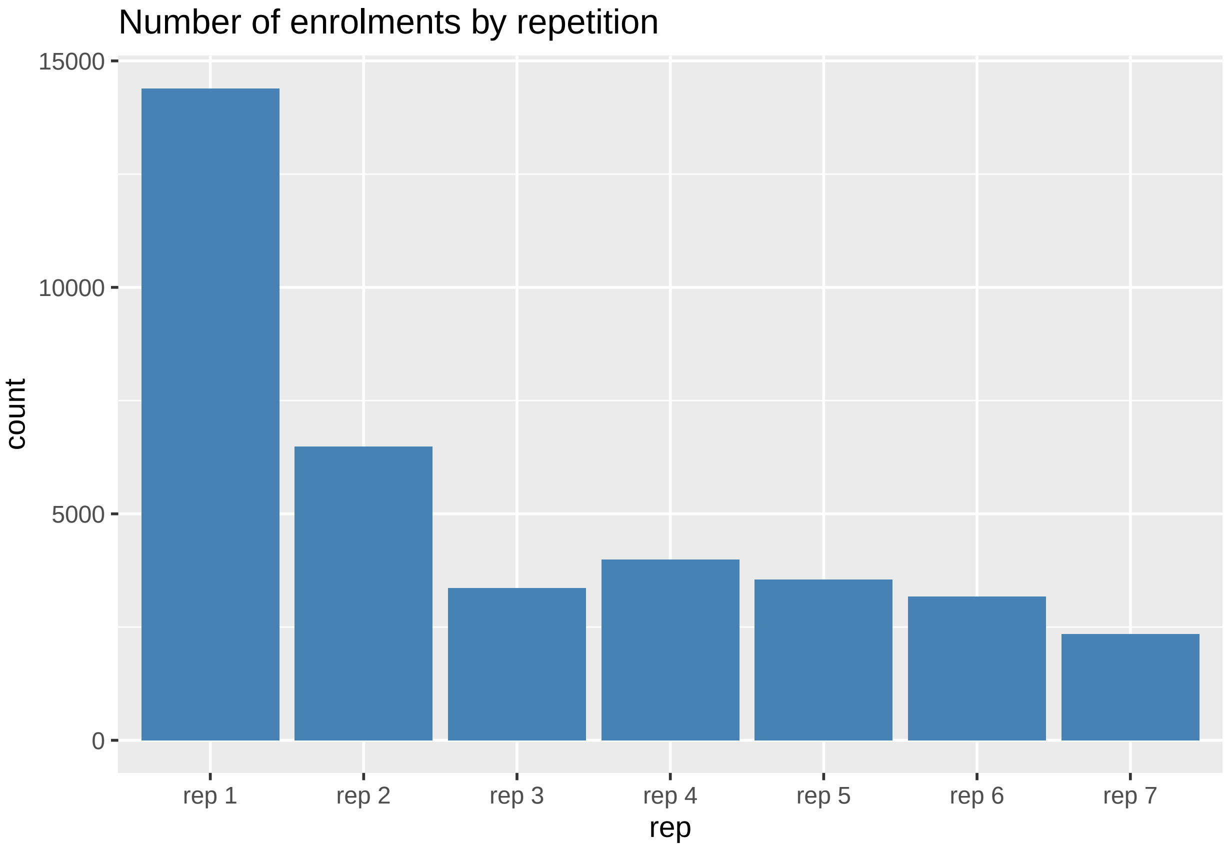
Firstly information was collated on the number of enrolments each year.The enrolments tables has been grouped in the table combined_enrolments.

```
glimpse(combined_enrolments)
```

```
## Rows: 37,296
## Columns: 14
## $ learner_id              <chr> "160d6600-ea0e-4568-bfa9-5d7cd5b8e61b", "4dc22~
## $ enrolled_at             <chr> "2016-08-10 14:28:49 UTC", "2016-05-24 17:34:3~
## $ unenrolled_at           <chr> "", "2018-10-30 20:20:51 UTC", "", "", "", "",~
## $ role                    <chr> "learner", "learner", "learner", "learner", "l~
## $ fully_participated_at   <chr> "", "", "2016-09-22 16:56:03 UTC", "", "", "20~
## $ purchased_statement_at  <chr> "", "", "", "", "", "", "", "", "", "", "", ""~
## $ gender                  <chr> "Unknown", "male", "Unknown", "Unknown", "Unkn~
## $ country                 <chr> "Unknown", "PE", "Unknown", "Unknown", "Unknow~
## $ age_range               <chr> "Unknown", "46-55", "Unknown", "Unknown", "Unk~
## $ highest_education_level <chr> "Unknown", "university_degree", "Unknown", "Un~
## $ employment_status       <chr> "Unknown", "working_part_time", "Unknown", "Un~
## $ employment_area         <chr> "Unknown", "teaching_and_education", "Unknown"~
## $ detected_country        <chr> "GB", "PE", "NG", "UG", "IM", "NO", "GB", "GB"~
## $ rep                     <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
```

To identify if there was a trend in the number of enrolments in each year, the number of enrolments over each year is displayed in the below bar chart.

```
ggplot(combined_enrolments, aes(x=rep, fill=rep))+
  geom_bar(fill = "steelblue")+
  theme_grey()+
  ggtitle("Number of enrolments by repetition")
```



Enrolments clearly decline from repetition 1 to 7, however this graph is simply based on the number of learners enrolled on the course at each repetition. It would be helpful to see the number of people who complete the course each repetition. To do this there is a column within the enrolments data which indicates the date a student has fully participated in the course

```
glimpse(combined_enrolments$fully_participated_at)
```

```
##  chr [1:37296] "" "" "2016-09-22 16:56:03 UTC" "" "" ...
```

This column does however contain blank cells.

```
sum(combined_enrolments$fully_participated_at == "")
```

```
## [1] 35142
```

From the 37,396 enrolments over the seven repetitions, there is 35,142 blank cells in the fully participated column. To remove the blank cells from the column a new table was created using the following code:

```
enrolments_fully_part = select(combined_enrolments, learner_id, rep, gender, age_range, fully_participa
  na_if("")%>%
  na.omit(combined_enrolments$fully_participated_at)
```

```
glimpse(enrolments_fully_part)
```

```
## Rows: 2,151
## Columns: 9
## $ learner_id               <chr> "ecdd37db-0c75-496e-bff2-230553d0e38c", "25cc3~
## $ rep                      <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
## $ gender                   <chr> "Unknown", "Unknown", "Unknown", "Unknown", "m~
## $ age_range                <chr> "Unknown", "Unknown", "Unknown", "Unknown", "3~
## $ fully_participated_at    <chr> "2016-09-22 16:56:03 UTC", "2016-10-25 12:44:1~
## $ highest_education_level <chr> "Unknown", "Unknown", "Unknown", "Unknown", "s~
## $ employment_status        <chr> "Unknown", "Unknown", "Unknown", "Unknown", "w~
## $ employment_area          <chr> "Unknown", "Unknown", "Unknown", "Unknown", "a~
## $ detected_country         <chr> "NG", "NO", "GB", "GB", "IT", "UA", "KE", "GB"~
```

This leaves 2151 learners that have fully completed the course. It was then possible to create a new table where the learner id of the 2151 students was used to link this table with the question responses of these learners. This table can be seen below:

```
glimpse(fully_participated_learners_qrespones)
```

```
## Rows: 71,132
## Columns: 19
## $ learner_id               <chr> "ecdd37db-0c75-496e-bff2-230553d0e38c", "ecdd3~
## $ rep.x                    <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
## $ gender                   <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ age_range                <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ fully_participated_at    <chr> "2016-09-22 16:56:03 UTC", "2016-09-22 16:56:0~
## $ highest_education_level <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ employment_status        <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ employment_area          <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ detected_country         <chr> "NG", "NG", "NG", "NG", "NG", "NG", "NG", "NG"~
## $ quiz_question            <chr> "1.7.1", "1.7.2", "1.7.2", "1.7.3", "1.7.4", "~
## $ question_type            <chr> "MultipleChoice", "MultipleChoice", "MultipleC~
## $ week_number              <int> 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 2, 2, 2, 2~
## $ step_number              <int> 7, 7, 7, 7, 7, 7, 7, 7, 11, 11, 11, 11, 8, 8, ~
## $ question_number          <int> 1, 2, 2, 3, 4, 5, 6, 6, 1, 1, 2, 3, 1, 1, 1, 2~
## $ response                 <chr> "1,2,3", "1", "2", "1,2,3,4,5", "2", "2", "2",~
## $ cloze_response           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ submitted_at             <chr> "2016-09-13 01:23:52 UTC", "2016-09-13 01:25:1~
## $ correct                  <chr> "true", "false", "true", "true", "true", "true~
## $ rep.y                    <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
```

There are still columns in this table which aren't helpful when focusing on the the students responses to questions. As demonstrated earlier in this report the "cloze_response" column contains NA in every row so therefore can clearly be removed. The submitted at column also holds no real relevance to how a student responds to a question. Also we can see in the merging of the the two tables the repetition or rep column has been duplicated as it was in both tables, so this can also be removed. The following code will be added to remove the columns

```
fully_participated_learners_qrespones = left_join(enrolments_fully_part, qresponse_all, by = "learner_i
  fully_participated_learners_qrespones = select(fully_participated_learners_qrespones, -cloze_response
  group_by(learner_id)
```

```
glimpse(fully_participated_learners_qrespones_clean)
```

```
## Rows: 71,132
## Columns: 16
## Groups: learner_id [2,145]
## $ learner_id              <chr> "ecdd37db-0c75-496e-bff2-230553d0e38c", "ecdd3~
## $ rep.x                   <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
## $ gender                  <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ age_range               <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ fully_participated_at   <chr> "2016-09-22 16:56:03 UTC", "2016-09-22 16:56:0~
## $ highest_education_level <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ employment_status       <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ employment_area         <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ detected_country        <chr> "NG", "NG", "NG", "NG", "NG", "NG", "NG", "NG"~
## $ quiz_question           <chr> "1.7.1", "1.7.2", "1.7.2", "1.7.3", "1.7.4", "~
## $ question_type           <chr> "MultipleChoice", "MultipleChoice", "MultipleC~
## $ week_number             <int> 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 2, 2, 2, 2~
## $ step_number             <int> 7, 7, 7, 7, 7, 7, 7, 7, 11, 11, 11, 11, 8, 8, ~
## $ question_number         <int> 1, 2, 2, 3, 4, 5, 6, 6, 1, 1, 2, 3, 1, 1, 1, 2~
## $ response                <chr> "1,2,3", "1", "2", "1,2,3,4,5", "2", "2", "2",~
## $ correct                 <chr> "true", "false", "true", "true", "true", "true~
```

The final data files that were combined were the video stats files. This was to look for trends in how the video files are watched over the 5 reps in that the video files were included. The combined video stats table is grouped by repetition but could also be linked to other tables using the rep column as a primary key.

```
glimpse(combined_videostats)
```

```
## Rows: 65
## Columns: 29
## $ step_position                <dbl> 1.10, 1.14, 1.17, 1.19, 1.50, 2.10, 2.1~
## $ title                        <chr> "Welcome to the course", "Why would any~
## $ video_duration               <int> 99, 362, 241, 348, 281, 37, 312, 92, 42~
## $ total_views                  <int> 1659, 910, 723, 755, 1248, 694, 564, 51~
## $ total_downloads              <int> 113, 77, 63, 62, 100, 48, 53, 42, 50, 3~
## $ total_caption_views          <int> 36, 8, 5, 2, 15, 1, 4, 3, 5, 1, 1, 5, 4~
## $ total_transcript_views       <int> 221, 173, 120, 147, 191, 108, 110, 87, ~
## $ viewed_hd                    <int> 58, 28, 16, 10, 41, 13, 434, 7, 16, 6, ~
## $ viewed_five_percent          <dbl> 76.97, 72.53, 73.72, 72.85, 78.45, 76.3~
## $ viewed_ten_percent           <dbl> 75.35, 70.88, 73.86, 71.92, 75.64, 75.0~
## $ viewed_twentyfive_percent    <dbl> 73.42, 68.57, 71.92, 69.27, 69.87, 74.9~
## $ viewed_fifty_percent         <dbl> 70.40, 65.38, 69.71, 64.90, 65.63, 73.4~
## $ viewed_seventyfive_percent   <dbl> 68.17, 63.08, 66.11, 63.44, 62.66, 72.9~
## $ viewed_ninetyfive_percent    <dbl> 66.43, 61.54, 61.83, 61.59, 59.05, 71.1~
## $ viewed_onehundred_percent    <dbl> 63.71, 56.81, 44.67, 49.40, 44.87, 69.4~
## $ console_device_percentage    <dbl> 0.06, 0.11, 0.14, 0.13, 0.00, 0.14, 0.1~
## $ desktop_device_percentage    <dbl> 78.60, 79.23, 79.67, 78.54, 80.37, 79.1~
## $ mobile_device_percentage     <dbl> 13.26, 10.33, 8.71, 9.40, 11.38, 9.37, ~
## $ tv_device_percentage         <dbl> 0.06, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0~
## $ tablet_device_percentage     <dbl> 7.72, 10.11, 11.07, 11.39, 7.93, 10.95,~
## $ unknown_device_percentage    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ europe_views_percentage      <dbl> 55.15, 65.38, 66.25, 67.15, 61.62, 64.2~
## $ oceania_views_percentage       <dbl> 2.29, 2.86, 3.18, 3.18, 2.24, 3.17, 3.5~
## $ asia_views_percentage          <dbl> 16.09, 10.22, 9.82, 9.27, 12.34, 9.37, ~
## $ north_america_views_percentage <dbl> 11.63, 11.32, 10.65, 10.99, 11.38, 11.6~
## $ south_america_views_percentage <dbl> 3.07, 2.53, 2.21, 2.12, 2.72, 3.75, 2.6~
## $ africa_views_percentage        <dbl> 10.31, 6.26, 6.36, 5.56, 8.17, 6.20, 6.~
## $ antarctica_views_percentage    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ rep                            <chr> "rep 3", "rep 3", "rep 3", "rep 3", "re~
```

The mean number of views for each viseo clip was added to the combined video stats table to give statistical analysis and a new table was created.

```
glimpse(combined_videostats_mean)
```

```
## Rows: 65
## Columns: 30
## $ title                          <chr> "Devices in the future home", "Devices ~
## $ step_position                  <dbl> 3.20, 3.20, 3.20, 3.20, 3.20, 3.14, 3.1~
## $ video_duration                 <int> 206, 206, 206, 206, 206, 313, 313, 313,~
## $ total_views                    <int> 484, 557, 590, 275, 227, 446, 500, 544,~
## $ total_downloads                <int> 34, 60, 58, 25, 12, 42, 53, 56, 20, 12,~
## $ total_caption_views            <int> 4, 3, 1, 2, 2, 1, 8, 2, 3, 3, 5, 2, 2, ~
## $ total_transcript_views         <int> 89, 121, 125, 49, 51, 85, 110, 122, 40,~
## $ viewed_hd                      <int> 8, 6, 6, 3, 3, 13, 12, 8, 4, 3, 4, 5, 6~
## $ viewed_five_percent            <dbl> 74.59, 73.97, 77.63, 76.00, 70.48, 73.3~
## $ viewed_ten_percent             <dbl> 73.76, 71.10, 77.46, 74.18, 70.48, 72.6~
## $ viewed_twentyfive_percent      <dbl> 72.11, 70.02, 75.25, 72.00, 67.40, 72.2~
## $ viewed_fifty_percent           <dbl> 70.25, 67.86, 74.24, 71.27, 66.52, 68.1~
## $ viewed_seventyfive_percent     <dbl> 66.53, 66.97, 71.02, 67.64, 65.64, 67.0~
## $ viewed_ninetyfive_percent      <dbl> 65.29, 64.45, 69.83, 65.45, 64.76, 62.7~
## $ viewed_onehundred_percent      <dbl> 60.54, 61.04, 62.71, 61.09, 56.83, 38.5~
## $ console_device_percentage      <dbl> 0.21, 0.00, 0.00, 0.00, 0.00, 0.22, 0.0~
## $ desktop_device_percentage      <dbl> 80.99, 80.61, 74.41, 74.55, 79.74, 82.2~
## $ mobile_device_percentage       <dbl> 7.02, 10.05, 10.17, 13.45, 11.89, 6.50,~
## $ tv_device_percentage           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ tablet_device_percentage       <dbl> 11.16, 8.80, 15.25, 12.00, 7.93, 10.76,~
## $ unknown_device_percentage      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ europe_views_percentage        <dbl> 66.12, 56.37, 62.88, 56.36, 66.96, 66.3~
## $ oceania_views_percentage       <dbl> 3.51, 4.13, 5.25, 5.82, 5.29, 4.04, 4.0~
## $ asia_views_percentage          <dbl> 9.92, 14.72, 7.63, 21.45, 11.01, 8.52, ~
## $ north_america_views_percentage <dbl> 11.57, 9.52, 16.78, 8.00, 9.25, 11.21, ~
## $ south_america_views_percentage <dbl> 1.65, 2.51, 1.02, 1.45, 1.76, 2.47, 2.2~
## $ africa_views_percentage        <dbl> 5.17, 12.03, 4.58, 6.91, 4.85, 6.05, 11~
## $ antarctica_views_percentage    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ rep                            <chr> "rep 3", "rep 4", "rep 5", "rep 6", "re~
## $ average_total_views            <dbl> 426.6, 426.6, 426.6, 426.6, 426.6, 388.~
```

IV. Modeling

What is widely regarded as data science's most exciting work is also often the shortest phase of the project.

Here you'll likely build and assess various models based on several different modeling techniques. This phase has four tasks:

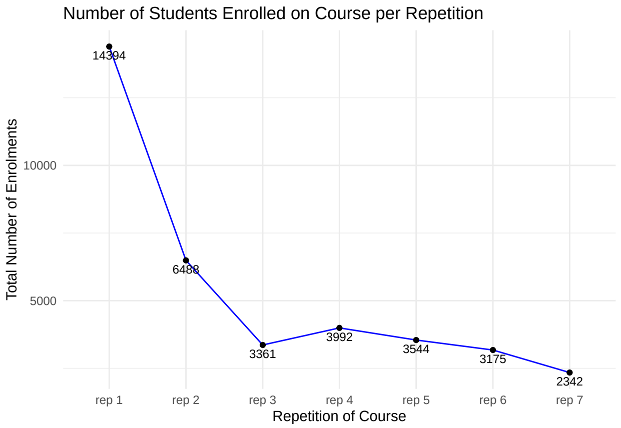
Firstly a comparison of total students enrolled over the seven repetitions and the students who have clearly completed the course.

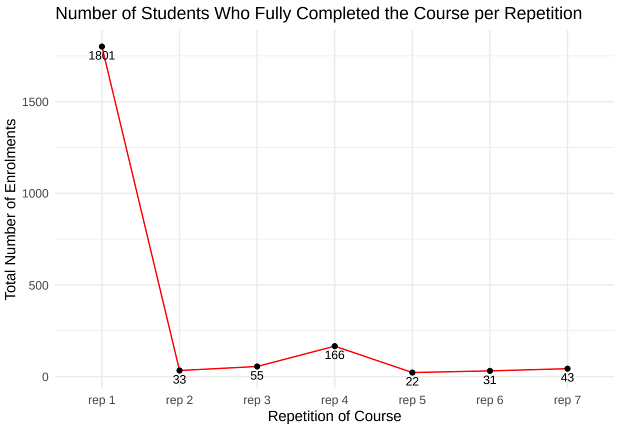
```
## # A tibble: 7 x 6
##   rep       n previous difference percentagechange percent2dp
##   <chr> <int>    <int>      <int>            <dbl> <chr>
## 1 rep 1 14394       NA         NA             NA   <NA>
## 2 rep 2  6488    14394      -7906            -54.9 -54.93%
## 3 rep 3  3361     6488      -3127            -48.2 -48.20%
## 4 rep 4  3992     3361        631             18.8 18.77%
## 5 rep 5  3544     3992       -448            -11.2 -11.22%
## 6 rep 6  3175     3544       -369            -10.4 -10.41%
## 7 rep 7  2342     3175       -833            -26.2 -26.24%
```

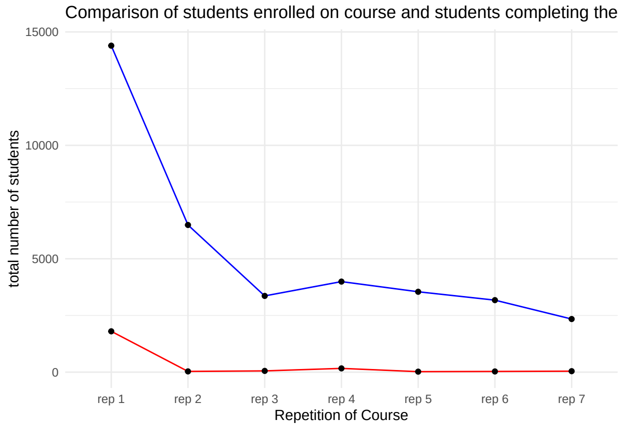
The table above was used to calculate the percentage increase or decrease in enrolments when compared to the following year. In the first two years that the course was ran the number of enrolments decreased by close to half when compared to the previous year, although there was an 18% increase in the fourth repetition. Could this potentially be linked to the introduction of supporting videos in the third repetition of the course.

After cleaning the data to show only students who had completed the course, the graph below shows that there was again a huge decline (-98%) in the students who fully completed the course in repetition two compared to the previous year. However after the introduction of the video materials in repetition three, there was a clear increase in the number of students who managed to fully complete the course. In repetition three there is an increase of 66.67% of students fully completing the course and a 201.82% increase in repetition four based on the previous years.

```
## # A tibble: 7 x 6
##   rep       n previous difference percentagechange percent2dp
##   <chr> <int>    <int>      <int>            <dbl> <chr>
## 1 rep 1  1801       NA         NA             NA   <NA>
## 2 rep 2    33     1801      -1768            -98.2 -98.17%
## 3 rep 3    55       33         22             66.7 66.67%
## 4 rep 4   166       55        111            202.  201.82%
## 5 rep 5    22      166       -144            -86.7 -86.75%
## 6 rep 6    31       22          9             40.9 40.91%
## 7 rep 7    43       31         12             38.7 38.71%
```







What was clear from the data was the percentage of students fully participating in the course was very low. The table below show the percentage rates for each repetition of the course.

```
## # A tibble: 7 x 5
## # Groups:   rep [7]
##   rep     n.x   n.y percentage.rate percentage_fully_participated
##   <chr> <int> <int>           <dbl> <chr>
## 1 rep 1 14394  1801          12.5   12.51%
## 2 rep 2  6488    33           0.509 0.51%
## 3 rep 3  3361    55           1.64  1.64%
## 4 rep 4  3992   166           4.16  4.16%
## 5 rep 5  3544    22           0.621 0.62%
## 6 rep 6  3175    31           0.976 0.98%
## 7 rep 7  2342    43           1.84  1.84%
```

In year 1 only 12.51 percent of students fully complete the course. This was the highest fully completed rate for the seven repetitions of the course. With a mean pass rate of

```
mean(compare_enrolments_percent_full$percentage.rate)%>%
  scales::percent(compare_enrolments_percent_full$percentage.rate ,accuracy = 0.01, scale = 1, prefix =
```

```
## [1] "3.18%"
```

The quality of the data makes it difficult to gather an accurate mean pass rate for the course. To ensure that there is more accurate approach only the students who have a fully completed enrolment form, with no unknowns or NA's will be used. Below is a table containing all the students details that enrolled on the course and fully completed.

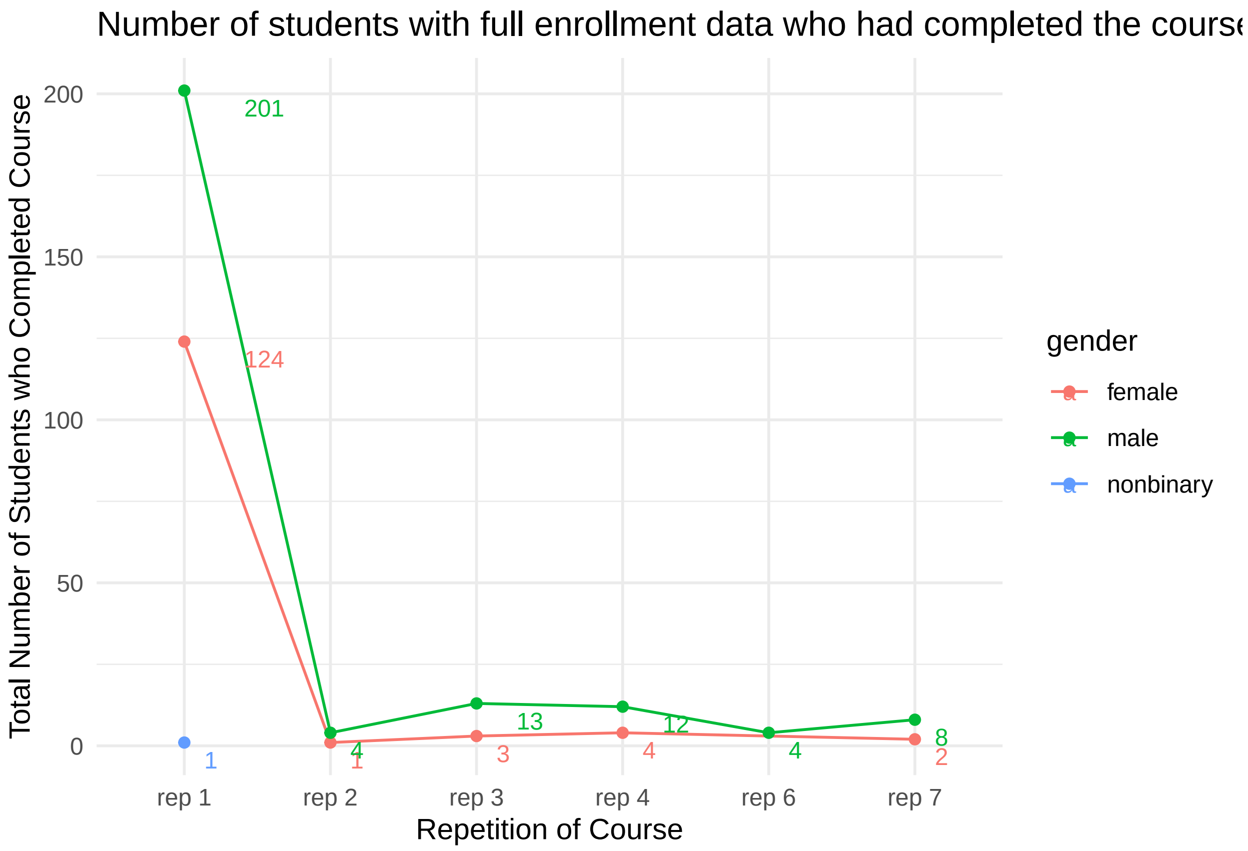
```
## Rows: 2,151
## Columns: 9
## $ learner_id              <chr> "ecdd37db-0c75-496e-bff2-230553d0e38c", "25cc3~
## $ rep                     <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
## $ gender                  <chr> "Unknown", "Unknown", "Unknown", "Unknown", "m~
## $ age_range               <chr> "Unknown", "Unknown", "Unknown", "Unknown", "3~
## $ fully_participated_at   <chr> "2016-09-22 16:56:03 UTC", "2016-10-25 12:44:1~
## $ highest_education_level <chr> "Unknown", "Unknown", "Unknown", "Unknown", "s~
## $ employment_status       <chr> "Unknown", "Unknown", "Unknown", "Unknown", "w~
## $ employment_area         <chr> "Unknown", "Unknown", "Unknown", "Unknown", "a~
## $ detected_country        <chr> "NG", "NO", "GB", "GB", "IT", "UA", "KE", "GB"~
```

The number of fields containing missing data is:

```
## [1] 8305
```

When the missing data was removed there was 377 students where all data had been collected at the point of entry. What is also noticeable at this point is there was no fully complete enrolment data in repetition five of the course. This added to the low numbers in repetitions two, three, four, six and seven indicate that the enrollments process needs strengthening.

```
## Rows: 377
## Columns: 9
## $ learner_id              <chr> "7a44b170-73f8-4863-8687-4f97934c8b0b", "5f50e~
## $ rep                     <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
## $ gender                  <chr> "male", "male", "female", "female", "female", ~
## $ age_range               <chr> "36-45", "26-35", "46-55", ">65", "36-45", "56~
## $ fully_participated_at   <chr> "2016-10-06 04:24:57 UTC", "2018-10-08 10:10:4~
## $ highest_education_level <chr> "secondary", "university_degree", "secondary",~
## $ employment_status       <chr> "working_full_time", "looking_for_work", "work~
## $ employment_area         <chr> "accountancy_banking_and_finance", "it_and_inf~
## $ detected_country        <chr> "IT", "RO", "GB", "AU", "IE", "GB", "GB", "ID"~
```



```
Rep <- c("rep1","rep2","rep3","rep4","rep6","rep7")
Male <- c(201, 4, 13, 12, 4, 8)
Female <-c(124, 1, 3, 4, 4, 2)
```

```
NonBinary <-c(1, 0, 0, 0, 0, 0)
total <-c(Male + Female +NonBinary)
Passrate_male<-c(Male/total)
Passrate_female<-c(Female/total)
Passrate_NonBinary<-c(NonBinary/total)
Male_passrate_percentage =  scales::percent(Male/total, accuracy = 0.01, scale = 100, prefix = "", suff
Female_passrate_percentage =  scales::percent(Female/total, accuracy = 0.01, scale = 100, prefix = "", s
Nonbinary_passrate_percentage =  scales::percent(NonBinary/total, accuracy = 0.01, scale = 100, prefix =

gender_all_enrolment_data_passrate <- data.frame(Male, Female, NonBinary, total, Passrate_male, Passrate
```

```
##   Male Female NonBinary total Passrate_male Passrate_female Passrate_NonBinary
## 1  201    124         1   326     0.6165644       0.3803681        0.003067485
## 2    4      1         0     5     0.8000000       0.2000000        0.000000000
## 3   13      3         0    16     0.8125000       0.1875000        0.000000000
## 4   12      4         0    16     0.7500000       0.2500000        0.000000000
## 5    4      4         0     8     0.5000000       0.5000000        0.000000000
## 6    8      2         0    10     0.8000000       0.2000000        0.000000000
##   Male_passrate_percentage Female_passrate_percentage
## 1                   61.66%                     38.04%
## 2                   80.00%                     20.00%
## 3                   81.25%                     18.75%
## 4                   75.00%                     25.00%
## 5                   50.00%                     50.00%
## 6                   80.00%                     20.00%
##   Nonbinary_passrate_percentage
## 1                         0.31%
## 2                         0.00%
## 3                         0.00%
## 4                         0.00%
## 5                         0.00%
## 6                         0.00%
```

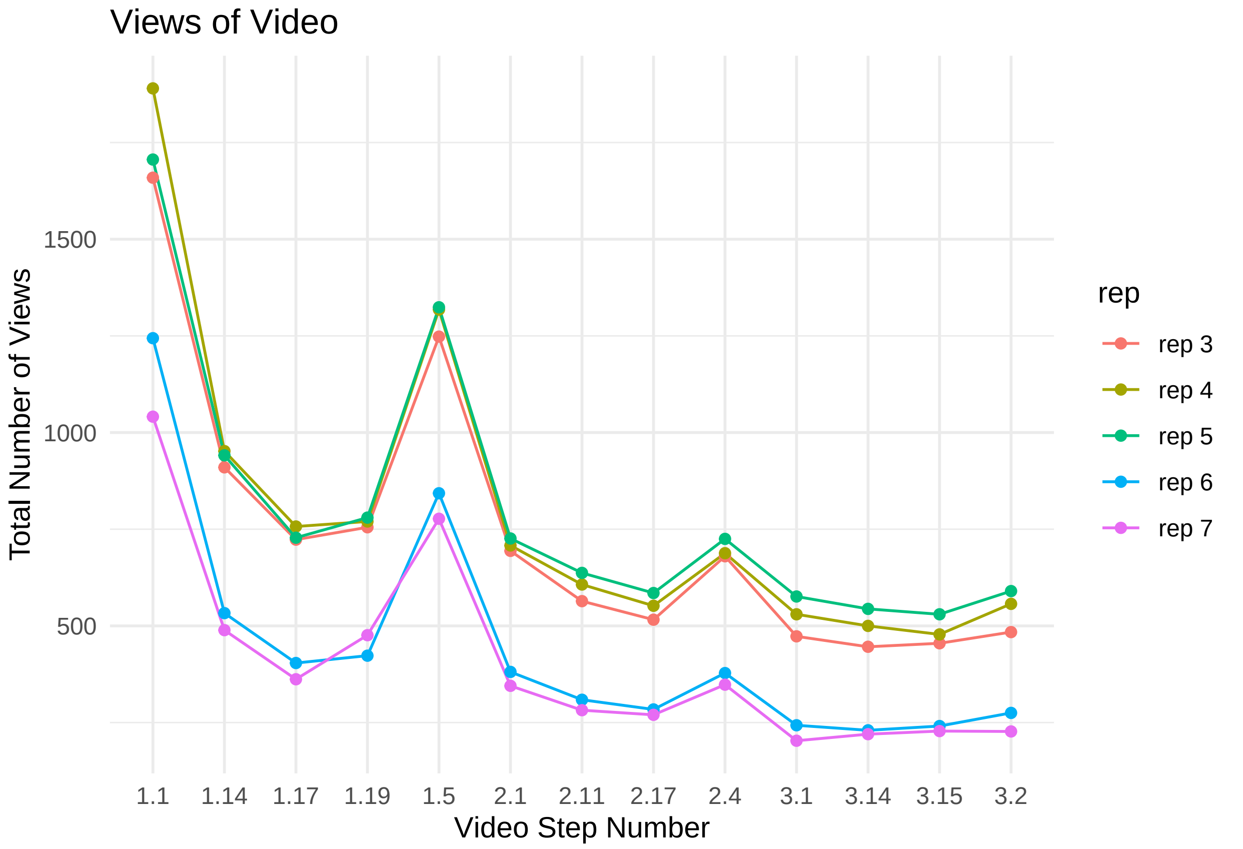
It is now possible to calculate the mean pass rate using only the data from students who had fully completed the course with fully completed enrolment data. The mean scores below show a far better pass rate than before.

```
## [1] "71.32%"
```

```
## [1] "28.63%"
```

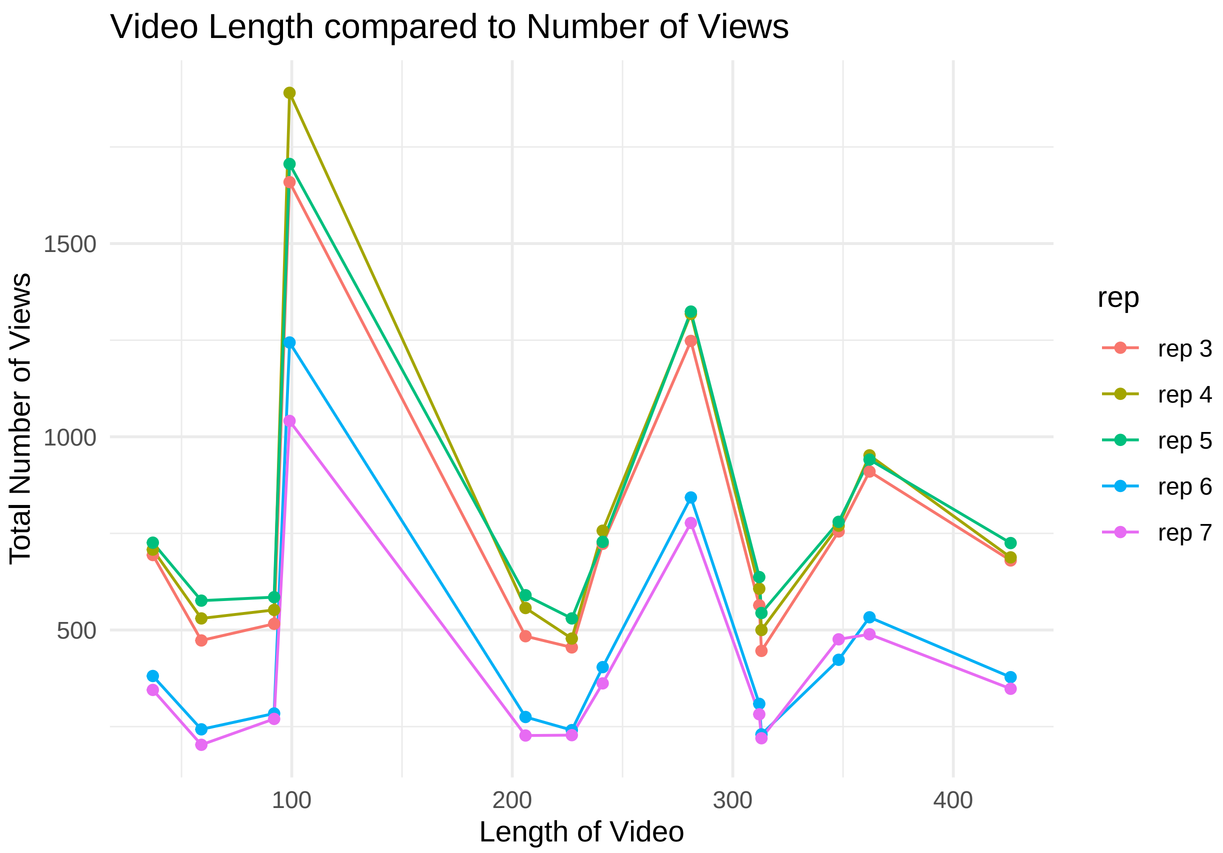
```
## [1] "0.05%"
```

Data was provided at repetition three on the video materials attached to the course. Below is a graph showing the amount of views each video had on each repetition.



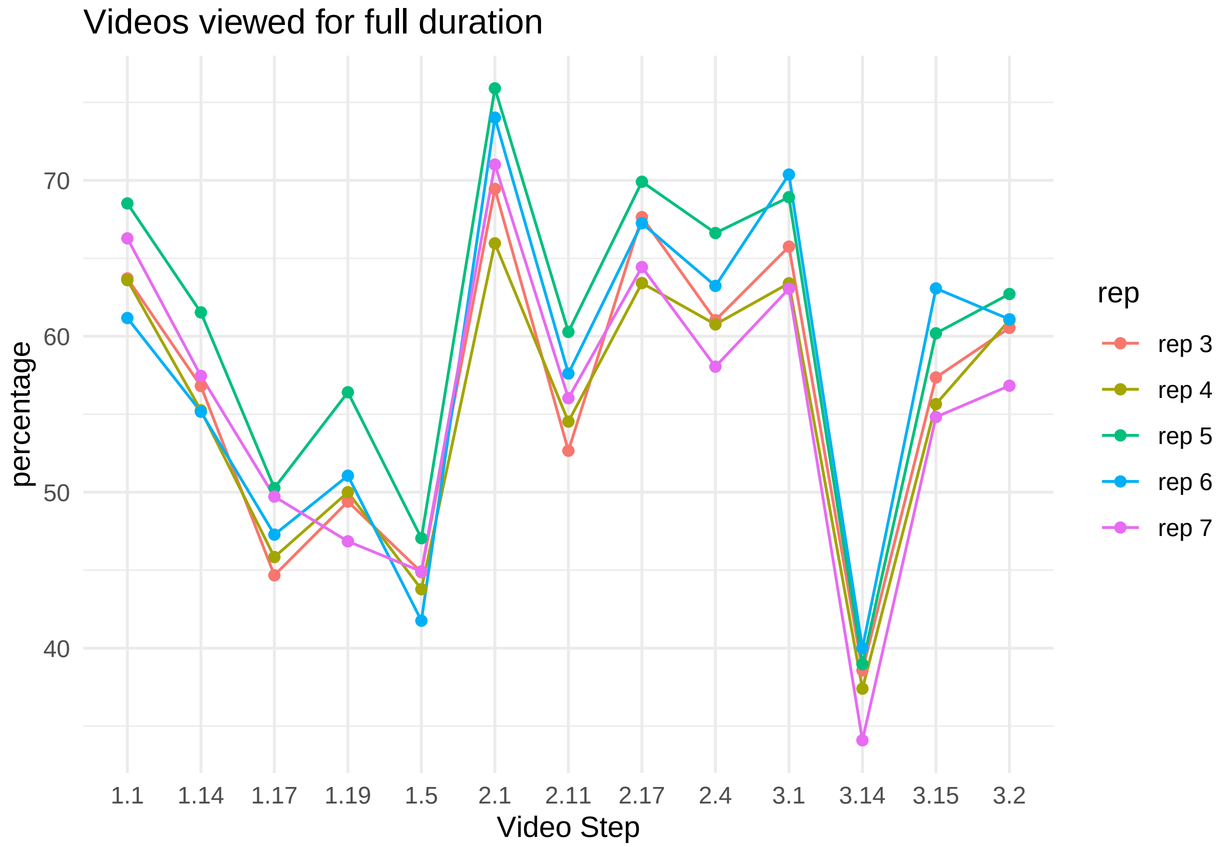
What is clear from the graph is that the videos follow an almost identical trend each time the course is ran. As shown earlier in the report there is a decline in the number of enrolments each year with a slight rise at repetition 4. The video data follows that trend with repetitions 4 and 5 having more total video views than other repetitions.

Where there are declines in the viewing figures, could this be in relation to the length of the video clips.



The above graph shows there is no correlation between video length and number of views. The above graph shows to clear spikes one at one of the shorter video lengths and one at a longer video lengths. The lines like the previous graph follow an almost identical trend.

The video stats data also shows how many of the videos were watched fully or viewed for 100% of the duration. Below is a graph showing the videos viewed to 100% of the duration.



From the above graph the video at step position 2.1 (Welcome to Week 2: payment security) has the highest percentage of people viewing the full video, with 3 of the 5 repetitions having over 70% of veiwers viewing it for the full duration. The video at step 3.14 (Exploring security: biometric authentication) has the poorest viewing percentage for the full duration. As in all 5 reps the video was only viewed fully by 40% or less of the total views.

V. Evaluation

This project initially set out to find out about learners enrolling on the course and the success and retention rates on the course. As well as looking for links to how the supporting video materials impacted on learning and if any improvements were required. The first thing that was clear was a decline in enrolments from the first year of the course being run. Finding a reasoning for this was difficult because of the gaps in the data. Across the 7 repetitions of the course there was only 377 learners that had full enrollment data as well as a fully completed the course date on their records. This made gaining an accurate conclusion on the completion of the course by any demographic difficult but did signal that a far more robust enrolment system must be introduced. The report was able to look at completion rates on the course by gender, once the data was cleaned. These result show that there was a higher rate of males completing the course than other genders. This could be further investigated to look at links between gender and age or gender and education level to test if there was a trend in the type of person that is most likely to enrol on the course. This further investigation would further all the university to target a wider audience but also look into why a certain demographic are more likely to be successful on the course.

The report also looked at the use of the video materials looking for a correlation between video length and total views. There was no clear relationship between video length and total views. The year on year trends were almost identical with certain videos being watched far more than others. This was supported when looking at which videos were being watched for 100% of the duration. What is clear is that certain videos are receiving more views than others and that this isn't linked to duration. Further investigation would be required to find out certain videos are being viewed more and for longer than others. This would allow the university to improve the quality of its content and ensure that the all the video material being put online is getting a higher number of viewing figures.

It would also be advantages to the university to look at links between the vidoe footage and question responses. This would give the university an insight into the impact that each video is having on student attainment. This would give tangible evidence on the effectiveness of the video material.

VI. Deployment

The analysis is reproducible, new course data could be loaded to the directory and ran through the project again. This would be reliant on the new data following the same name format as previous files. Cyber.security.8_video_stats for example would follow the same format as previous video files allowing any previous coding to be added using a pipeline.