Data Management and Exploratory Data Analysis

Buisness understanding

Business objectives

This report is focusing on Learning Analytics a fast developing domain of Data Science. In Learning Analytics organizations can extract context and better understanding of their programs. The goal is to derive suitable business results while improving employee productivity. The discoveries made could allow learning organizations to pick and choose learning models that works best, resulting to their organizations increase profit while lowering the time spent in developing the learning material. The organization in question provides an online course on cyber security, they have conducted 7 runs for this course. Further discussion on the course as a whole and differences between some of the runs will commence. The main objective is to produce findings derived from the data collected as well as further investigation on some of the most noticeable and interesting topics.

Assess situation

Learning Analytics refers to the collection and analysis of data about learners and their environments. Therefore, it is significant to create tools that provide and in depth understanding of the learners background undertaking the course. The scope of this analysis is to improve the learning outcomes of the course both from a business standpoint and how its best received by the learners.

Determining data mining goals

In educational data mining it is important to classify learner performance and engagement. Furthermore, it is important to assess the material provided and the levels of engagement they provided. Some of the data mining goals at this stage are:

1. Learner background information.
2. Extraction of learner course completion across all runs.
3. Assessing the course material that provide a sense of performance.
4. Analysis on the data that provide further context in terms of learner engagement.
5. Various useful information extracted by the data mining process

In addition, a successful Data Mining process will be derived by:

- The reliability of the data.
- The importance of the results in terms of context, clarity and usefulness.
- The reliability of the processes conducted to extract results.
- The appropriate method used to provide a coherent discussion,

Project plan

The tools and techniques that are going to be used include Git version control, ProjectTemplate for R and RMarkdown. Git version control is a system that allows to keep track of the changes made to the code over time. As such, collaboration is made possible because specific changes on the work are tracked and tagged by various contributors. The work for this project will be conducted by the author of this report, allowing for further expansion of the work with other collaborators in the future. ProjectTemplate in R allows for the automation of new statistical analysis projects. In addition, it provides a directory structure that automates data loading, preprocessing, library importing and testing. Finally, RMarkdown is a versatile open source markup language and can be used to format plain text while enabling the user to directly post their work on web pages if they must. For data transformation and visualization dplyr and ggplot2 will be utilized within ProjectTemplate.

The project plan consists of the tools and techniques mentioned above to dwell into the data provided from the organization and produce a coherent structured report. The project plan includes the data understanding divided by some appropriate exploration and discussions. Next,is data preparation where the cleaning and construction of the data will take place. Finally, the modelling where in depth analysis will commence. The project plan will be laid out in detail in the next chapters of this report.

Data Understading

Collection of initial data

The data are collected and loaded in the ProjectTemplate environment which has been set as a working directory using Git Bash. With the data loaded comes 7 pdf files containing the format of the material of each run commenced.

Data Description

The data that will be utilized consists of 53 csv files. The runs does not share the same number of csv files it noticeable that inclusions and exclusions made when each run commenced. This will be further explored in the data exploration.

The distinct csv files across all runs are as follows:

1. Archetype Survey responses
2. Enrollments
3. Leaving survey responses
4. Question responses
5. Step Activity
6. Weekly sentiment survey responses
7. Team members
8. Video stats

Decription of the data sets content

The archetype data set consist of the learners and an archetype answer regarding their general behavior. The enrollment data set consists of enrolled learners and 13 variables that provide numerous information such as enroll dates, fully completion dates, gender and more. Leaving survey responses includes of course the reason the learners gave for abandoning the course. The question responses data set consists of detailed learner performance regarding quizzes provided throughout the course. The step activity includes step progression for each week as well as started and completion dates for each step in the course. The weekly sentiment

responses include weekly feedback from the learners regarding weekly progress. The team members data set consists of description regarding the roles of the teaching team. Lastly, The video stats data set consist of various statistics regarding the video material on the course for example views, percentages watched, devices, geolocation and more.

Data exploration

By exploring the data several questions surfaced for a portion of the data sets provided. Although, all data sets where explored and considered for further analysis. The data sets that was deemed appropriate for the outlined goals and business objectives includes the enrollment, question responses, step activity, and video stats. The reasoning will be discussed in detail on the the data selection and construction chapter of this report.

Enrollement data set (Run 1)

The enrollment data set consists of the learners as our observations number and each learner has 13 variables.

```
head(cyber.security.1_enrolments)
```

```
## # A tibble: 6 x 13
##   learner_id  enrolled_at  unenrolled_at role  fully_participa~ purchased_state~
##   <chr>       <chr>        <chr>         <chr> <chr>            <chr>
## 1 160d6600-e~ 2016-08-10 ~ ""            lear~ ""               ""
## 2 4dc22fed-6~ 2016-05-24 ~ "2018-10-30 ~ lear~ ""               ""
## 3 ecdd37db-0~ 2016-05-19 ~ ""            lear~ "2016-09-22 16:~ ""
## 4 988964c9-7~ 2016-05-19 ~ ""            lear~ ""               ""
## 5 f1493366-1~ 2016-09-19 ~ ""            lear~ ""               ""
## 6 25cc3b46-a~ 2016-08-30 ~ ""            lear~ "2016-10-25 12:~ ""
## # ... with 7 more variables: gender <chr>, country <chr>, age_range <chr>,
## #   highest_education_level <chr>, employment_status <chr>,
## #   employment_area <chr>, detected_country <chr>
```

All 13 variable are stored as characters it noticeable that numerous rows are empty and some are stored as Unknown.

Question response data set (Run 1)

The question response data set consists of the learners as our observations number and for each learner there are 10 variables

```
head(cyber.security.1_question.response)
```

```
## # A tibble: 6 x 10
##   learner_id quiz_question question_type week_number step_number question_number
##   <chr>      <chr>         <chr>               <int>       <int>           <int>
## 1 77454a73-~ 1.7.1         MultipleChoi~           1           7               1
## 2 77454a73-~ 1.7.1         MultipleChoi~           1           7               1
## 3 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
## 4 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
## 5 a4fa6f89-~ 1.7.1         MultipleChoi~           1           7               1
```

```
## 6 f27eec8c-~ 1.7.1         MultipleChoi~           1           7               1
## # ... with 4 more variables: response <chr>, cloze_response <lgl>,
## #   submitted_at <chr>, correct <chr>
```

Some variables are stored as characters namely learner_id, quiz_question, question_type response, submitted_at, correct. The rest are stored as integers expect cloze_response that is logical and is filled with NAs. It is advisable to further investigate regarding inconsistencies.

Step activity data set (Run 1)

The step activity data set consists 143092 observations with 6 variables.

```
head(cyber.security.1_step.activity)
```

```
## # A tibble: 6 x 6
##   learner_id     step week_number step_number first_visited_at  last_completed_~
##   <chr>         <dbl>       <int>       <int> <chr>             <chr>
## 1 77454a73-6b8~   101           1           1 2016-08-02 13:45~ ""
## 2 c1a75ae7-c76~   101           1           1 2016-08-02 15:40~ ""
## 3 a4fa6f89-a59~   101           1           1 2016-08-03 07:14~ ""
## 4 60b56cea-ad2~   101           1           1 2016-08-03 08:45~ "2016-08-03 08:~
## 5 05a815ce-3c4~   101           1           1 2016-08-03 13:24~ "2016-08-03 13:~
## 6 5553e67a-178~   101           1           1 2016-08-03 14:49~ "2016-08-04 15:~
```

Some variables are stored as characters namely learner_id, first_visited_at and last_completed_at the others are stored as integers except the step variable which is double.

Video stats data set (Run 3)

The video stats data set consists of 13 observation which refer to the video step position over the 3 weeks of the course and 28 variables. It's important to mention that the video stats data is not present in the first 2 runs.

```
head(cyber.security.3_video.stats)
```

```
## # A tibble: 6 x 28
##   step_position title                video_duration total_views total_downloads
##           <dbl> <chr>                         <int>       <int>           <int>
## 1          1.1  Welcome to the course            99        1659             113
## 2          1.14 Why would anyone wan~           362         910              77
## 3          1.17 Preserving privacy i~           241         723              63
## 4          1.19 Staying safe online:~           348         755              62
## 5          1.5  Privacy online and o~           281        1248             100
## 6          2.1  Welcome to Week 2: p~            37         694              48
## # ... with 23 more variables: total_caption_views <int>,
## #   total_transcript_views <int>, viewed_hd <int>, viewed_five_percent <dbl>,
## #   viewed_ten_percent <dbl>, viewed_twentyfive_percent <dbl>,
## #   viewed_fifty_percent <dbl>, viewed_seventyfive_percent <dbl>,
## #   viewed_ninetyfive_percent <dbl>, viewed_onehundred_percent <dbl>,
## #   console_device_percentage <dbl>, desktop_device_percentage <dbl>,
## #   mobile_device_percentage <dbl>, tv_device_percentage <dbl>, ...
```

In the stat video we have variables stored as characters for the description of the videos,integers for counts and doubles for percentages.

Leaving survey responses (Run 5)

In the leaving survey response data set there are 174 learners with 8 variables.

```
head(cyber.security.5_leaving.survey.responses)
```

```
## # A tibble: 6 x 8
##      id learner_id   left_at  leaving_reason   last_completed_~ last_completed_~
##   <int> <chr>        <chr>    <chr>            <chr>                       <dbl>
## 1 34003 8853543d-b9~ 2018-01~ I prefer not to~ ""                           NA
## 2 38604 b170480c-7e~ 2018-02~ The course requ~ ""                           NA
## 3 39016 92b8485e-b2~ 2018-02~ Other            ""                           NA
## 4 39241 f4cae359-09~ 2018-02~ Other            ""                           NA
## 5 39604 3429e915-a6~ 2018-02~ I prefer not to~ ""                           NA
## 6 39657 a04b5f29-c2~ 2018-02~ Other            "2018-02-05 18:~             1.2
## # ... with 2 more variables: last_completed_week_number <int>,
## #   last_completed_step_number <int>
```

By reviewing the data it is noticeable that learners that left the course are assigned new IDs regarding their leaving survey that are stored as integers. Furthermore, The last_completed step variables are stored as doubles while the last_completed_step_number are stored as integers. NAs can be spotted instantly further investigation will be required to assess the quality of the set.

Weekly survey responses data set (Run 6)

The weekly survey responses consist of 80 observations (learners) and 5 variables.

```
head(cyber.security.6_weekly.sentiment.survey.responses)
```

```
## # A tibble: 6 x 5
##      id responded_at            week_number experience_rating reason
##   <int> <chr>                         <int>             <int> <chr>
## 1 16810 2018-06-11 18:28:04 UTC           1                 3 ""
## 2 17388 2018-06-12 16:38:12 UTC           2                 3 "It was good i ha~
## 3 17505 2018-06-12 20:24:38 UTC           2                 3 "Nice Content !!"
## 4 17807 2018-06-13 12:05:35 UTC           1                 3 ""
## 5 17819 2018-06-13 12:25:05 UTC           1                 3 ""
## 6 18277 2018-06-14 10:31:48 UTC           1                 3 ""
```

Similarly with the leaving responses they are assigned a survey ID, variables are stored as characters for the responded_at and the reason. The week_number and the experience_rating are stored as integers. It noticeable that some responses are missing for the rating reasons.

Verification of data quality

From the exploration in most data sets inconsistencies were noticed. Specifically, same type of variables where stored as different types, multiple empty rows where discovered, some rows where labeled Unknown. In addition, multiple data sets contain NAs, some columns were empty and in some of the responses surveys the characters contained symbols.In the question response data set repetition of entries were discovered for some of the questions, the assumption is that the learners were able to resubmit the quiz question if they made a mistake so the focus is shifted to the quiz quality rather than learner performance. In the step

activities data set a mistake was discovered regarding how the steps number variable were recorded and stored, appropriate transformation will take place. This report is going to cover specific topics and will be using a portion of the data sets available to explore and perform an in depth analysis. After the selection of relevant data sets and rationale for inclusion and exclusion, the cleaning process will be documented where is applicable.

Data preparation

Selection of data

In this chapter an explanation will be given for the selection of the data that was previous mentioned in data exploration. Each data set will be discussed and goals will be outlined for the production of the analysis. Furthermore, steps taken for cleaning, construction, integration and reformation will be discussed. It is important to state that the data preparation takes into account all the goals and methods outlined above. Namely, general statistics, engagement, performance and various elements that will be derived from the analysis.

Enrollment

In the enrollment data set it is possible to extract a count of the people enrolled and their background information. In addition, calculation of the percentage of the learners that fully participated in the course for each run is possible. This will show a scale of popularity among the runs and completion rates. Also, some generic statistics about the learners will be generated.

Analysis goals:

1. Learners information
2. Course Engagement rates

Questions responses

The question response data set contain variables regarding quiz performance for each learner. Quizzes are a satisfactory way to access learner performance regarding the teaching material provided.

Analysis goals:

1. Success rate based on each week.
2. Correlation and failure statistics regarding the material.

Step activity

Step activity is a particularly useful data set because it contains details about each step across the three weeks of the course. It is possible to derive results that would provide vital information on the teaching material and the reception it received from the learners.

Analysis goals:

1. Percentage of steps completed each week.
2. Completion ratio for each step type.
3. Most popular step type discussion.

Video stats

The video stats data set enables the analysis to dwell deeper into this specific step type, getting context and statistics of how well this step type was received by the learners as well as how it was conducted by the teaching team. Several assumptions could be taken beforehand that will be discussed in the analysis section.

Analysis goals:

1. Percentages of video views for each step.
2. Percentages of viewers filtered by device utilized
3. Percentages of viewers from each continent.

Data cleaning and data construction

Enrollement cleaning

By checking for NAs in the enrollment data set it returns 13 NAs.By running the code:

```
cyber.security.1_enrolments= cyber.security.1_enrolments[rowSums(is.na(cyber.security.1_enrolments)) ==
sum(is.na(cyber.security.1_enrolments))
```

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

All NAs from the data set were omitted. in addition where applicable its possible to remove only rows that are fully NA using:

```
cyber.security.1_enrolments[rowSums(is.na(cyber.security.1_enrolments)) != ncol(cyber.security.1_enrolm
```

```
## # A tibble: 14,382 x 13
##    learner_id  enrolled_at unenrolled_at role  fully_participa~ purchased_state~
##    <chr>       <chr>       <chr>         <chr> <chr>            <chr>
##  1 160d6600-e~ 2016-08-10~ ""            lear~ ""               ""
##  2 4dc22fed-6~ 2016-05-24~ "2018-10-30 ~ lear~ ""               ""
##  3 ecdd37db-0~ 2016-05-19~ ""            lear~ "2016-09-22 16:~ ""
##  4 988964c9-7~ 2016-05-19~ ""            lear~ ""               ""
##  5 f1493366-1~ 2016-09-19~ ""            lear~ ""               ""
##  6 25cc3b46-a~ 2016-08-30~ ""            lear~ "2016-10-25 12:~ ""
##  7 9c23a086-f~ 2016-06-22~ ""            lear~ "2016-10-10 11:~ ""
##  8 8851dc49-0~ 2016-08-07~ ""            lear~ "2018-10-17 18:~ ""
##  9 a59b0a12-a~ 2016-08-02~ "2018-10-17 ~ lear~ ""               ""
## 10 198c1017-5~ 2016-09-09~ ""            lear~ ""               ""
## # ... with 14,372 more rows, and 7 more variables: gender <chr>, country <chr>,
## #   age_range <chr>, highest_education_level <chr>, employment_status <chr>,
## #   employment_area <chr>, detected_country <chr>
```

Next checking for duplicates in the learner_id column:

```
cyber.security.1_enrolments$learner_id[duplicated(cyber.security.1_enrolments$learner_id)]
```

```
## character(0)
```

Enrolment data constuction for analysis

For the enrollment a new data frame is constructed containing the learners that fully participated in the course. Furthermore, this new subset is filtered by gender, education and employment status the resulting data sets will be utilized in the analysis section.

Question response cleaning

In the question response data set the relevant columns for the analysis conducted are the question number and the correct column. These two columns have been checked and they do not indicate any problems.

Question response data constuction for analysis

For the question response data set three data frames were constructed filtered by the correct answers for the questions for each week respectively. This was utilized to compare performance in terms of success. The results are stored inside a new set and visualized in the analysis section.

Step activity Cleaning

The step activity has an issue on the step number variable. Specifically is not possible to distinguish the 1.1 step from 1.10, to fix that problem multiplication by 100 took place so now 101 refers to step 1.1 and 110 to step 1.10. The data set does not contain any NA rows. In terms of quality the step activity can be considered appropriate for the work intended.

Step activity data construction for analysis

The step activity data sets are divided into three sets for each weak, the goal is to derive completion percentages for all steps in each week. Then data frame were constructed for each step type. The resulted percentages are stored in tables and utilized in the analysis.

###Video stats cleaning In video stats data set the last column is empty and thus it's omitted from the analysis. In general the set does not contain any NA rows or other inconsistencies similar to what were discovered.

Video stats data construction for analysis

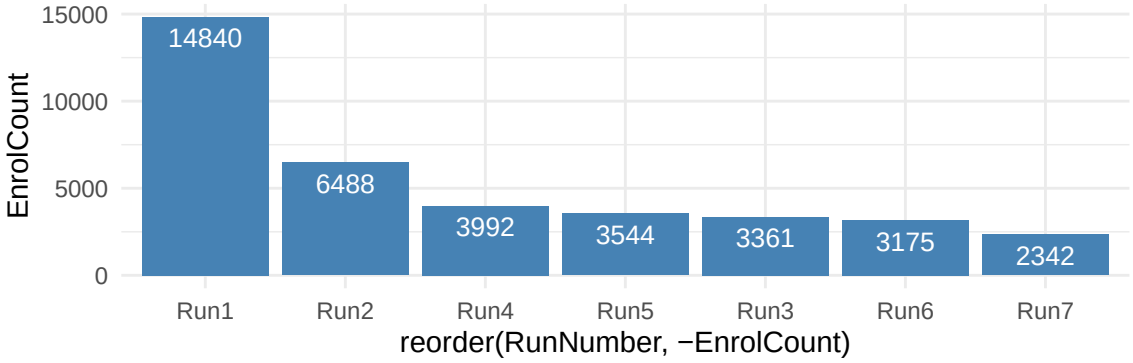
For the analysis the total_views column is needed alongside the video_duration and the step_position to investigate engagement e.g. the relation of the video lengths and the views for each step. Also, the columns regarding the devices used for viewing will be kept to provide an understanding of the viewing circumstances. Lastly, the columns regarding the location will be used and the results will be discussed. Three new data frames were constructed for the views, devices and continents.

Analysis

Enrollements

The analysis starts with the enrollment data set. The first goal is to make a comparison regarding the enrolled learners throughout the runs.

```
EnrollCountPlot
```

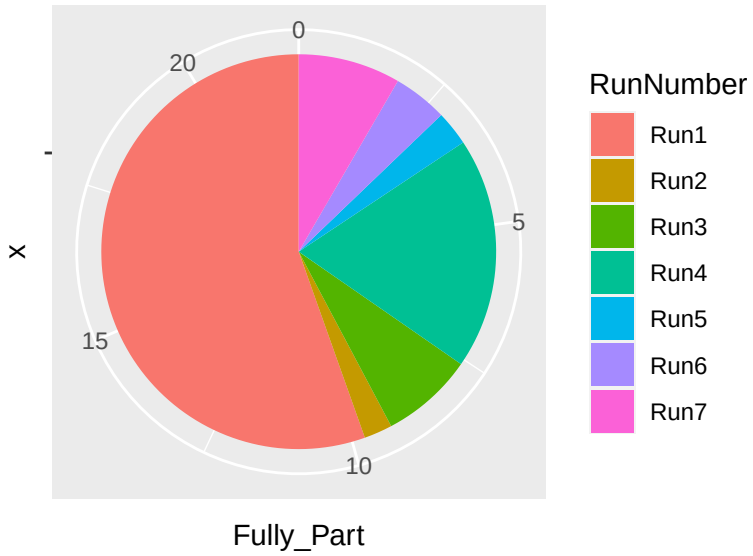


The plot depicts all enrollments made in each run the most popular one being the first with 14840 enrolled learners after comes the second run with 6488. The remaining runs were similar in terms of the volume of the people enrolled. The most popular run was the first by a huge margin. Next, the fully participation percentages was computed, meaning how many enrolled learners fully participated throughout the course. In the table below the calculated percentages are stored and visualized using a pie chart.

```
Participation_percent
```

```
##   RunNumber Fully_Part
## 1      Run1 12.1495957
## 2      Run2  0.5086313
## 3      Run3  1.6661708
## 4      Run4  4.1583166
## 5      Run5  0.6207675
## 6      Run6  0.9763780
## 7      Run7  1.8360376
```

```
pie1
```



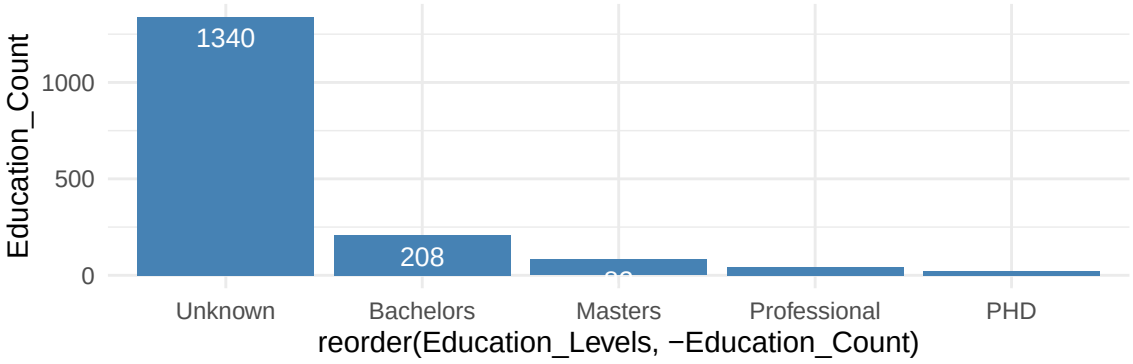
It is noticeable that the first run had 12% completion making it the run with the larger completion rate, taking into consideration that had the most enrolled learners by a significant margin suggesting that the first run its suitable for further investigation regarding the enrollments data set.

On that 12% subset from the first run some investigation took place regarding genders, education and employment.

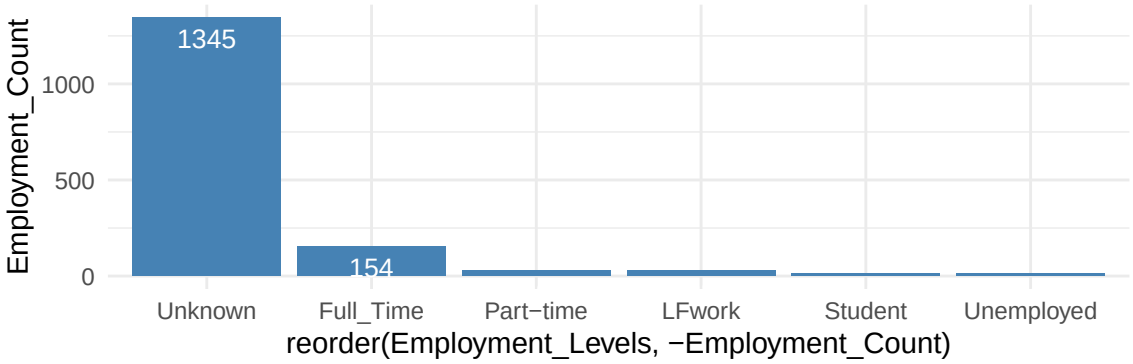
```
Run1Gender
```

```
##    Gender Gender_count
## 1    Male          263
## 2  Female          190
## 3 Unknown         1345
```

```
Education_Plot
```




```
Employment_Plot
```



From the table and the plots produced even though the sample is relatively small the unknown values are dominating the analysis. No concrete assumptions can be made regarding the learners background only some categorization and indication of the genders, employment and education can be assumed.

Enrollements discussion

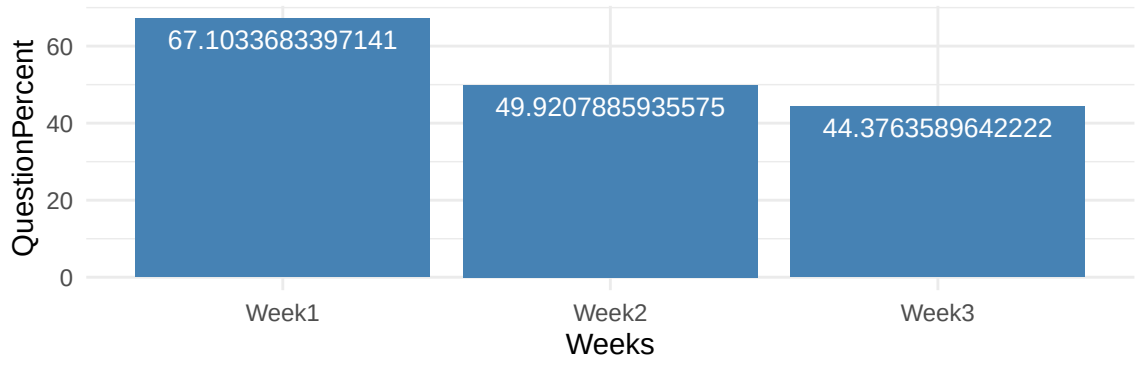
The most popular run was the first one that also had the highest completion rate regarding the full course, perhaps indicating it captured the learners interest the most. Unfortunately, the data set contains a significant amount of unknown values making further investigation highly speculative. It would be advisable to keep a better track about the learners in future runs, their background could assist on improving the course for learners that share similar characteristics.

Question response

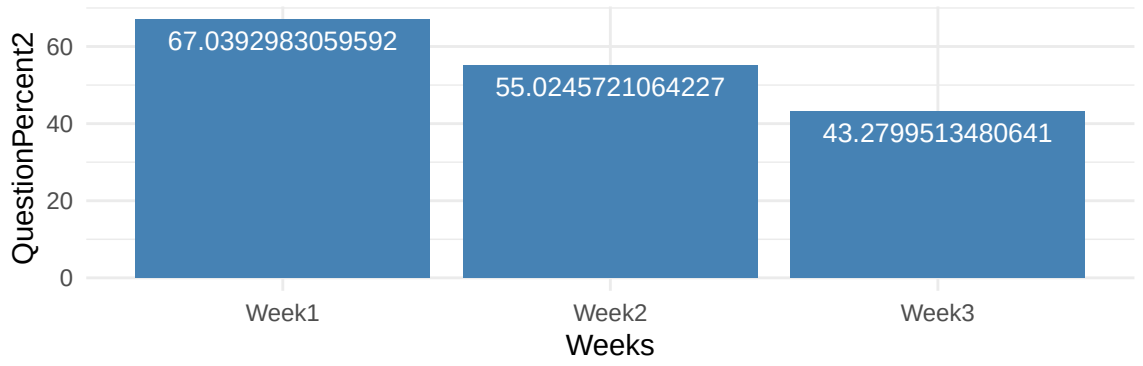
For the analysis only the runs 1, 2 and 3 will be analyzed. The reasons for this decision is the difference that was noticed in the enrollment analysis indicating similar popularity in the last 4 runs of the course. Furthermore, the first run is a very indicative for the performance of the course, on the second run and onward differences in the steps were noticed. Indicating changes that occurred in the course structure after the first run. This will be further analyzed with the step activity investigation. The third run is also important because video stats are introduced in the data which are a part of the analysis.

Regarding, the analysis that took place for the quiz responses each run was divided by 3 weeks and for each week the successful answers where measured to produce a percentage of completion.

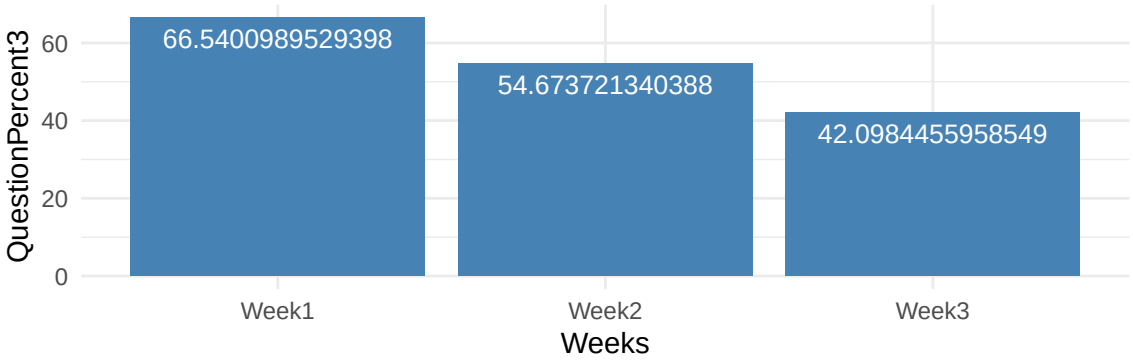
```
QuestionsPlot1
```




```
QuestionsPlot2
```




```
QuestionsPlot3
```

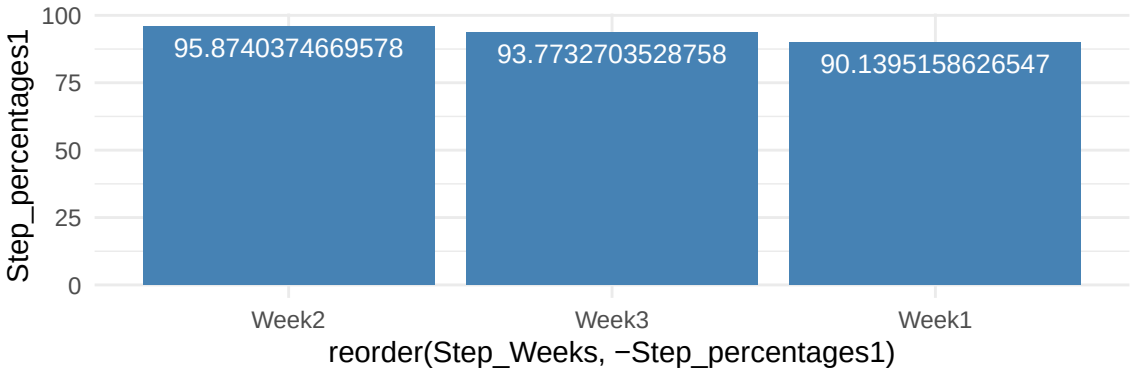


It is quite clear that similar results were produced on all runs. Meaning the changes occurred after the first run did not affect the successful quiz ratio. The first week was the most successful on all three runs. This could indicate that the material were more advanced from week to week or the structure of the teaching types somehow affected the quiz performance of the learners.

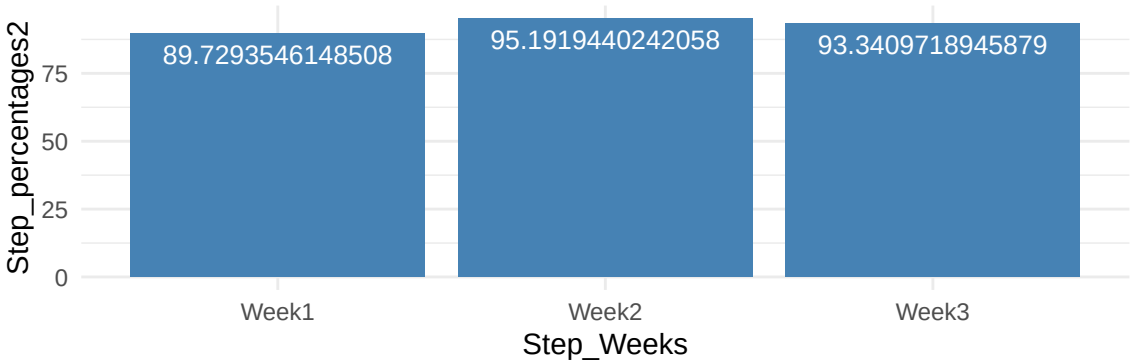
Step activity

In the step activity data the focus is on the completion rates for the steps. The analysis were conducted on the first three runs similarly with the question responses. Firstly, the data were divided to 3 weeks for each run to calculate the completion percentages and compare each week. Each plot its depicted below starting with the first run through run number three.

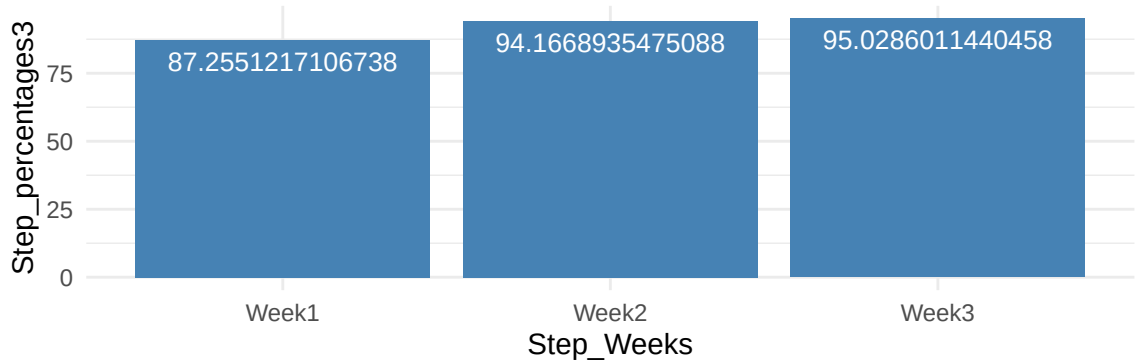
```
Step1Plot1
```




```
Step2Plot1
```

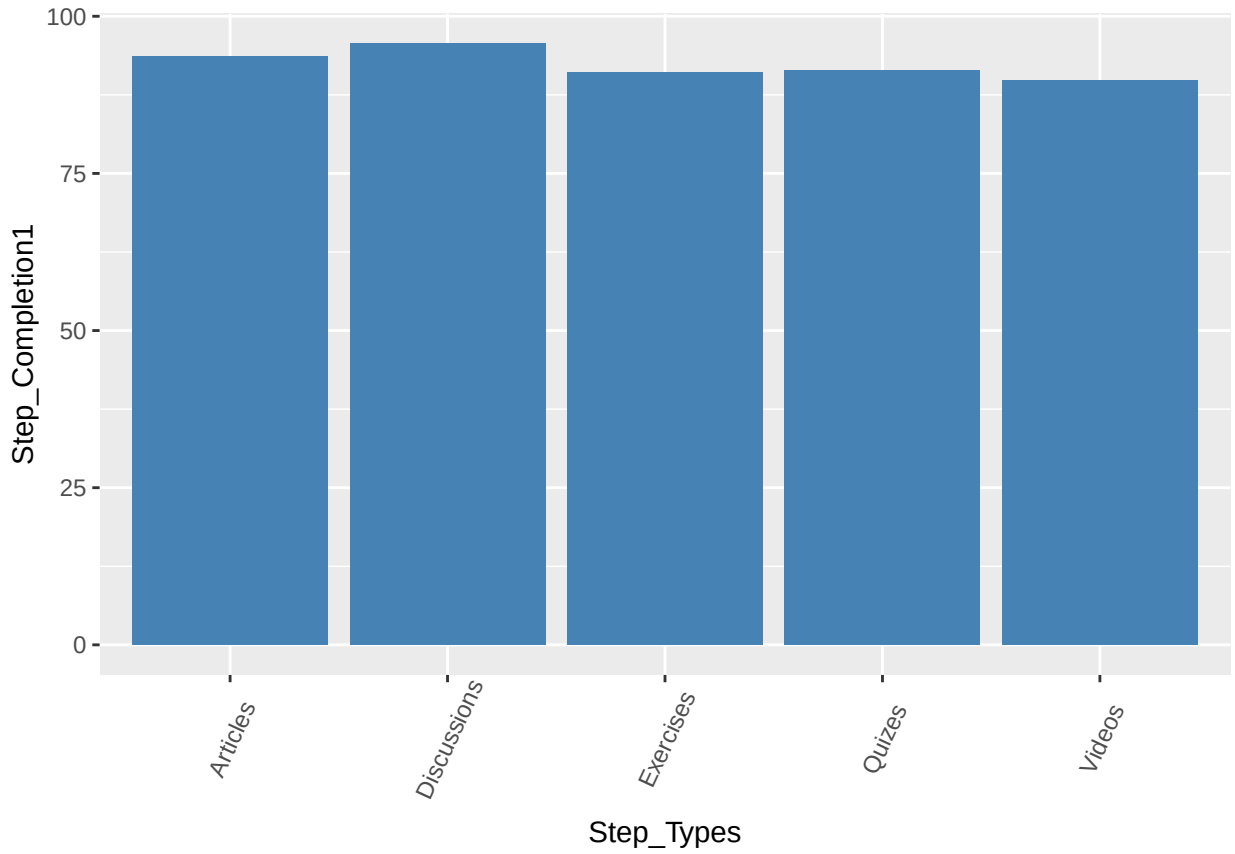


```
Step3Plot1
```

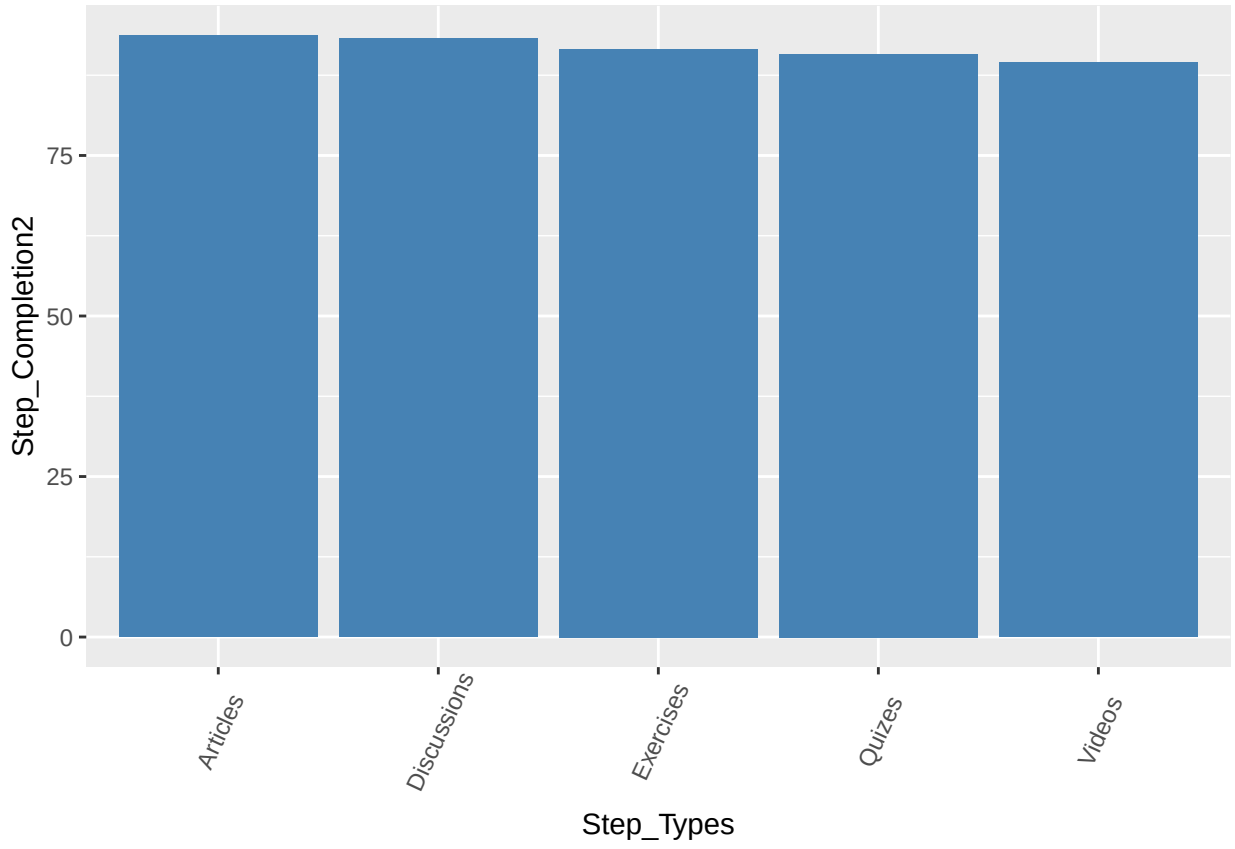


The completion rates are acceptable most of the people that started the steps completed them. It is noticeable that the first week has the lowest completion rates for all steps combined. The assumption that could be made is that the material of the first week wasn't adequate enough to sustain the engagement of the learners. The learning material contain different step types, namely Videos, Articles, Discussions, Quizzes and Exercises. This is not apparent for all types from the data sets with the exclusion of quizzes and videos that have their own sets. The pdf files that accompanied the data have a detailed structure of the course for each run. The goal was to compare each step type with the other in terms of completion. The following plots were produced for each run respectively:

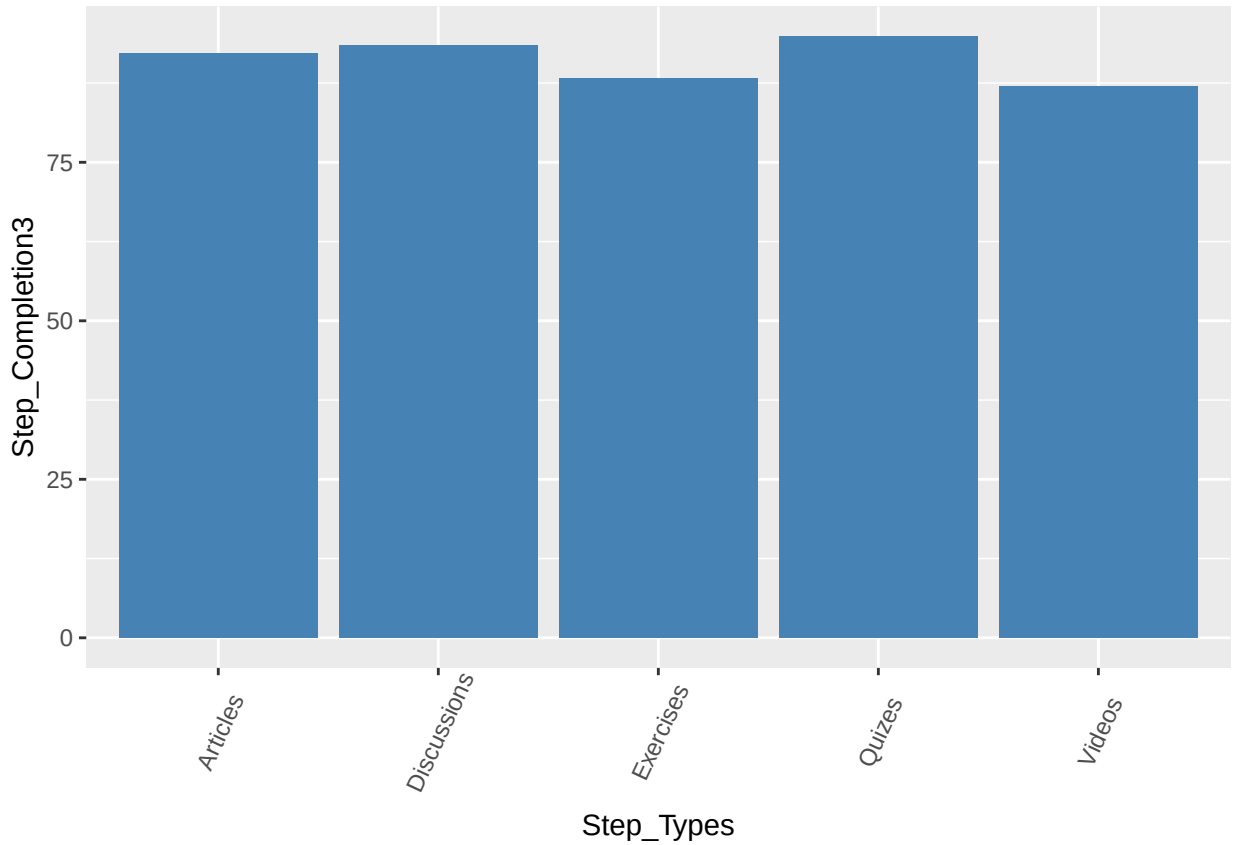
```
Step1Plot2
```



```
Step2Plot2
```




```
Step3Plot2
```



Again as expected the completion rate is significant with minor changes across runs. Although, the types are not of equal size. The following table depicts the step number counts for the first run:

```
table1
```

```
##         Types Count
## 1      Videos    12
## 2    Articles    32
## 3 Discussions     3
## 4      Quizes     8
## 5   Exercises     4
```

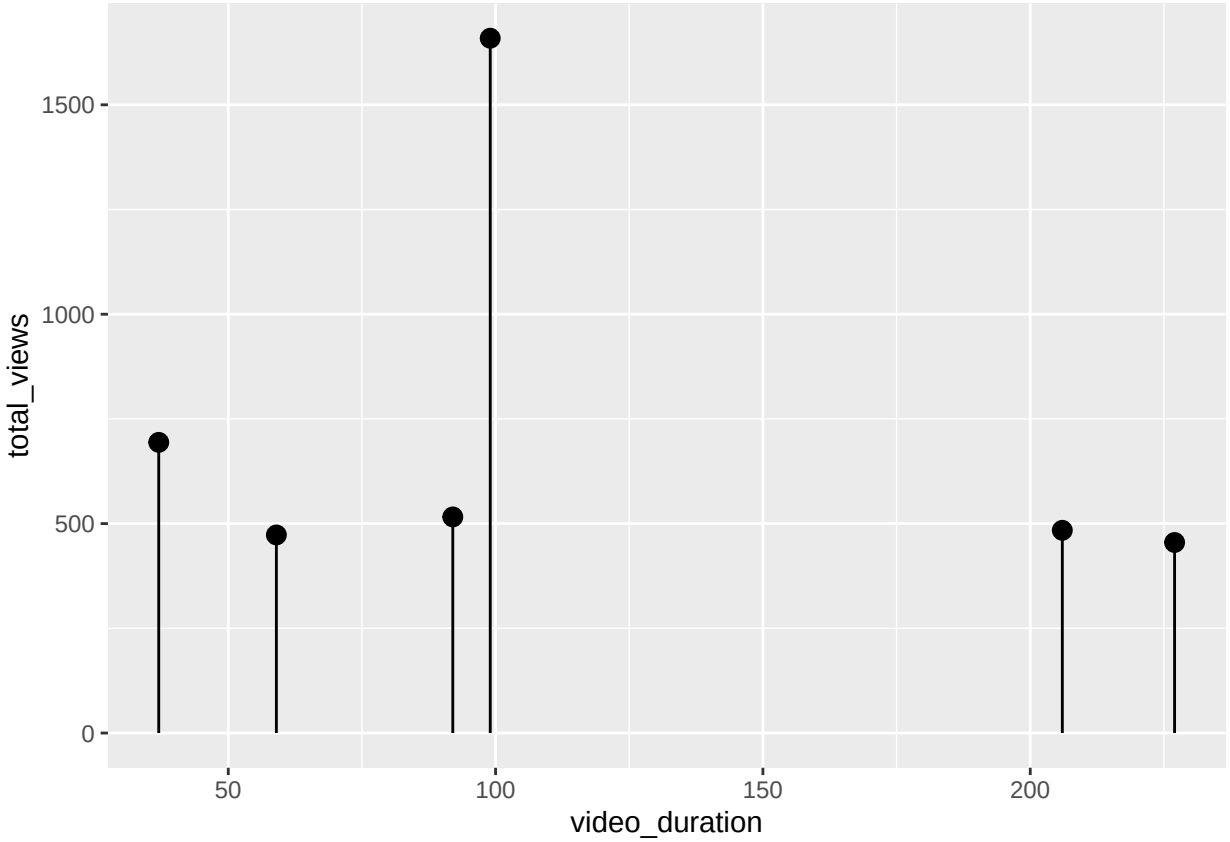
The discussions, quizzes and exercises are a fraction of the contents of the learning materials and cannot be indicative about the engagement although,they are important in the purpose they serve. It is noticeable that the articles are dominating the course in terms of what materials are provided by the teaching team. Despite almost appearing 3 times more compared to the videos their completion percentages are very similar. The articles are crucial for the course due to the majority of the material provided being in article form. Although, videos could provide a better alternative or at the very least the material could be equally divided on both articles and videos.

Video stats

The video stats data set contains various information about the videos viewed throughout the course by the learners. The first investigation focus on what may affected viewership. Firstly, the data set is filtered based on the column viewed_ninetyfive_percent that column contains viewing percentages for each video

that 95% of it was watched. This column was chosen with the assumption that 95% percent of a video is sufficient to be considered as full viewing a video. Some videos could contain blank sections in the end or some additional material beyond the focus of the video. It is logical to assume that some learners skipped these sections by choosing this criteria the learners are not falsely classified in another viewing category. On average 64.2% watched all the videos in their full duration. Next, this videos are examined based on their views and duration investigating if view counts are affected by duration, the plot below was produced.

```
VideoPlot1
```

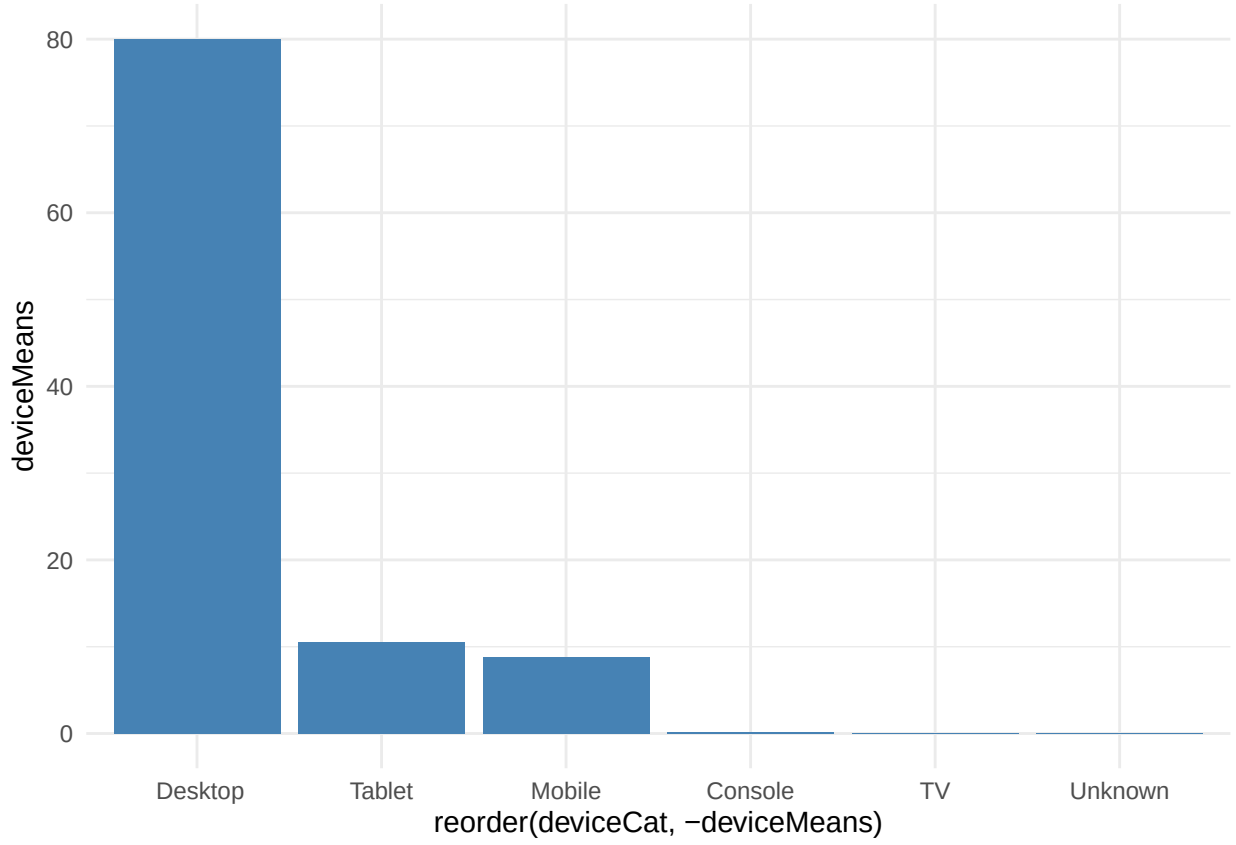


The average video length is 231 seconds with that in mind the plot seems to suggest the videos that were watched in their entirety were below average in length highly suggesting that shorter videos increase learner engagement. As discussed above in the step activity analysis, videos seems to be a popular step type, thus the introduction of more videos replacing articles with an average duration below 200 seconds could increase learner engagement and potentially performance in future runs of the course. Furthermore, devices used to view the course's videos are investigated to provide some background information about the learners a summary of their means alongside a plot are as follows:

```
colMeans(Video_devices3)
```

```
## console_device_percentage desktop_device_percentage  mobile_device_percentage
##               0.150769231              80.057692308               8.790769231
##      tv_device_percentage  tablet_device_percentage unknown_device_percentage
##               0.004615385              10.516923077               0.000000000
```

```
VideoPlot2
```

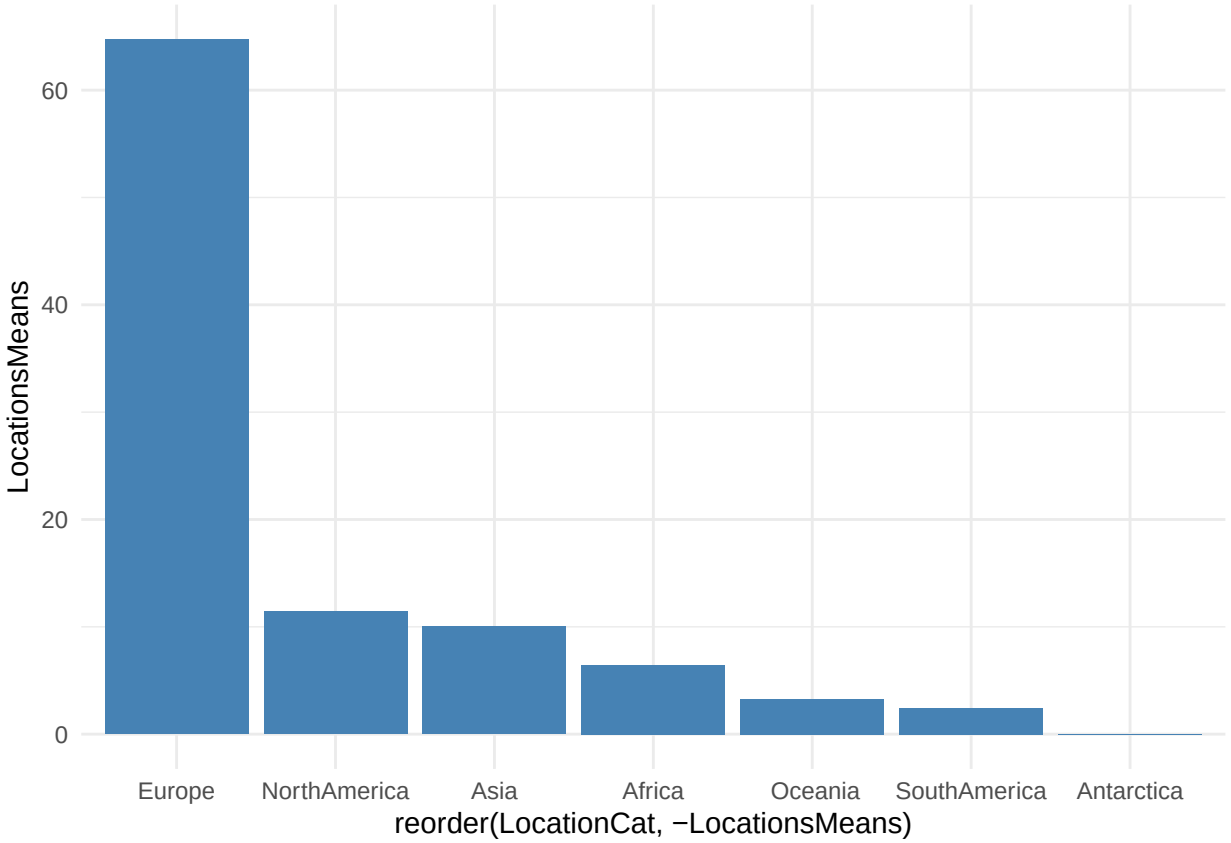


As expected, the most used device is the desktop computer followed by tablets and mobile devices. It is advisable for the teaching team to test beforehand the material of the course on a variety of devices in order to ensure an acceptable experience for all learners. As we noticed in the enrollment analysis there was a vast variety of learners enrolled is it safe to assume their circumstances differ. The teaching team should take appropriate action to provide equal and fair service for everyone within reason. Specifically, in the context of the videos this argument strengthens due to the fact that learners accessed the material from different continents.

```
colMeans(video_location3)
```

```
##        europe_views_percentage       oceania_views_percentage
##                      64.730769                       3.265385
##          asia_views_percentage north_america_views_percentage
##                      10.031538                      11.448462
## south_america_views_percentage        africa_views_percentage
##                       2.423846                       6.445385
##    antarctica_views_percentage
##                       0.000000
```

```
VideoPlot3
```



The plot above depicts the view percentages per continent, Europe was the dominant one in terms of viewership with an average of 64% across all videos watched the remaining percentage is allocated across the remaining continents.

Findings and Conclusion

From the enrollments data set the count of the enrolled users was computed for all runs. The most popular run was the first one with a staggering number of 14840 unique learners enrolled next was the second run with 6488 a significant decrease, the following runs kept on a decreasing having around 3000 learners enrolled on average. For each run the percentage of full participated learners was measured. with 12% the first run was the most successful second was the fourth run and the rest had an average of 1 percent for fully completion of the course. The investigation of the learner background wasn't fruitful due to the overwhelming Unknown variables stored in the data set. Although, a generic idea about the background can be derived looking at the various education and employment backgrounds. The course managed to reach a wealthy variety of learners.

From the question response data set by dividing the quizzes based on the week they commenced comparison of success for each question is possible. From the plots constructed decrease in terms of success can be noticed from week to week in all three runs that were investigated. This indicates that the material on week 2 and 3 didn't perform sufficient enough. After analyzing the steps it was noticed that the majority of the learners completed the steps in all weeks, although they didn't perform well in the quizzes. The quizzes as discussed in the report allowed the learners to resubmit their answers, thus shifting the focus of the investigation towards the understanding of the material. A better interpretation of the structure was obtained from the step activity analysis.

The step activity data set similarly with the question responses was divided by week for the first three runs. The general step completion for each weak was calculated showing high and similar completion rates on all

runs. Meaning that the majority of the learners that started a step finished it as mentioned above. After that an investigation commenced on the step types. The course heavily relies on articles to provide the materials needed to complete the course. The comparing of the results showed that even though the supporting videos are not essential for the completion of the course, were well received by the learners showing similar completion rates. It is therefore possible that the introduction of more videos replacing the articles would boost performance and engagement.

The videos stats were investigated in terms of views and duration, it was noticed that the duration of the videos affects the percentage of a video completion suggesting that videos below 200 seconds are more suitable. Furthermore, three type of devices used the most to view the material namely desktop computer, tablet and mobile. The desktop computer was the most commonly used assumptions can be made as to why, thus a suggestion of testing of the course material in different devices is made. Lastly, locations of where the videos viewed were analyzed. 64% of learners are located in Europe and the rest are allocated in other continents of the world. This is an indication of the variety of circumstances of the learners engaged in the course. The variety of countries and continents suggesting different time zones and that could affect any investigation on dates that were given in the data sets. Hence, it is quite important to have a healthy background data set for the learners.