CSC8631 Project

Luke Battle

22/11/2020

Introduction/Business Understanding

In this report, we will use the CRISP-DM method to analyse a data set. This will involve us running through several key steps; *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evalutaion* and *Deployment*.

Business Understanding

The first step in performing our CRISP-DM analysis is to gain an understanding of the business motivation for the analysis, with the primary goal of this analysis being to enhance Newcastle University's online resources. Specifcally, for the past several years, Newcastle University has ran a free online course called *Cyber Security: Safety at Home, Online, in life*. Several data sets have been compiled each time the course has been ran, detailing several facets of the user experience throughout the course. Newcastle University seeks to utilise this data to enhance the quality of material taught outside of the classroom, and ultimately promote learner engagement with the course and Newcastle's wider online resources. To achieve this, different elements of the collected data will be examined over consecutive runs, with the aim of identifying any trends both within and between them that may indicate an area where user engagement could be improved.

Both of these elements will be explored and examined both individually and in relation to each other, with the ultimate aim of providing feedback to Newcastle University on how it can enhance the user experience of the course. This analysis will be constructed using CRISP-DM methodology and, as such, will be structured to reflect this.

Data Understanding

Next, before beginning our analysis it is important to fully assess the data, so that we can make informed decisions about the portions of the data that should constitute the foundations of our analysis. In total, Newcastle University has provided 53 csv files that describe a range of features of user interaction with the MOOC over 7 runs of the course. These runs span from September 2016 to September 2018. For every run there is a file detailing "Archetype-survey-reponses", "Enrolments", "Leaving Survey Response", "Question Response", "Step Activity" and "Weekly Sentiment Survey Response". In addition, from run 2 to 7 a file on the "Team Members" is included, and from run 3 to 7 there is also a file on the "Video Stats". However, the data quality within some of these files is quite poor. For instance, "Leaving Survey Response" (**NOTE THAT THIS HAS BEEN USED IN LATER RUNS**), "Archetype Survey Response" and "Weekly Sentiment Survey Response" are blank except for their column headers, so they will not feature in this analysis. The "enrolments" data set does contain data, however the quality of it is poor. For 88.02% of the individuals in the data set, the only data available is the detected country of origin and enrollment date. These individuals are missing useful demographic indicators such as "gender, age range, education level and employment status. Due to the poor quality of the data this will not be used in the analysis. Out of the remaining data sets, video stats and question response may be a good place to start in assessing the efficacy of the MOOC course.

Describe Data

The video stats files are included for course runs 3 to 7. The files each have dimensions:

```
## [1] 13 28
```

With the following variables:

```
##  [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"
## [29] "Run"
```

Where each individual in the file is given by the video that the variables relate to. Which are the following:

```
##  [1] "Welcome to the course"
##  [2] "Why would anyone want your data?"
##  [3] "Preserving privacy in cloud storage: privacy by design"
##  [4] "Staying safe online: personal perspectives"
##  [5] "Privacy online and offline"
##  [6] "Welcome to Week 2: payment security"
##  [7] "Exploring vulnerabilities in online payments"
##  [8] "The million dollar contactless payment"
##  [9] "The evolving arms race of payment security"
## [10] "Welcome to Week 3: security in the future home"
## [11] "Exploring security: biometric authentication"
## [12] "Exploring security: the Access Control Live Lab"
## [13] "Devices in the future home"
```

An additional aspect of the video stats files that is important to consider for our analysis is whether the videos used in the course are the same for each run. If they are not, we could not fairly compare the efficacy of each video against each other when considering the results across runs. However, by examining the "duration" variable in each file we can see that this is identical for each video in every run, so we can assume for the purposes of this analysis that the videos have not been changed between runs.

For the question response data set, we can compute the same high level analysis. However, unlike the video stats files this file has been prepared for all 7 runs of the course, and is not in exactly the same format. Each file has 10 columns, however the rows in each data file is dependent on how many learners took part in the corresponding run, and so varies between files. In total, all 7 files have 176,463 rows.

Data Quality

We can now assess the data quality of each of these data sets using the dlookr package to see if there is any missing data in each file. Starting with the video stats file from the third run of the course, we calculate the following:

```
diagnose(video_file_raw)
```

```
## # A tibble: 28 x 6
##    variables       types  missing_count missing_percent unique_count unique_rate
##    <chr>           <chr>          <int>           <dbl>        <int>       <dbl>
##  1 step_position   numer~             0               0           13       1
##  2 title           chara~             0               0           13       1
##  3 video_duration  integ~             0               0           13       1
##  4 total_views     integ~             0               0           13       1
##  5 total_downloads integ~             0               0           12       0.923
##  6 total_caption_~ integ~             0               0            8       0.615
##  7 total_transcri~ integ~             0               0           13       1
##  8 viewed_hd       integ~             0               0           11       0.846
##  9 viewed_five_pe~ numer~             0               0           13       1
## 10 viewed_ten_per~ numer~             0               0           13       1
## # ... with 18 more rows
```

The above table shows that there is no missing data for this video stats file, and analysis on the rest of the files for video stats gives the same result. Now, we can apply the same process to the question file from the first run.

```
diagnose(question_file_raw)
```

```
## # A tibble: 10 x 6
##    variables      types   missing_count missing_percent unique_count unique_rate
##    <chr>          <chr>           <int>           <dbl>        <int>       <dbl>
##  1 learner_id     charac~             0               0         3410   0.0443
##  2 quiz_question  charac~             0               0           22   0.000286
##  3 question_type  charac~             0               0            1   0.0000130
##  4 week_number    integer             0               0            3   0.0000390
##  5 step_number    integer             0               0            5   0.0000649
##  6 question_numb~ integer             0               0            9   0.000117
##  7 response       charac~             0               0           32   0.000416
##  8 cloze_response logical         77002             100            1   0.0000130
##  9 submitted_at   charac~             0               0        75103   0.975
## 10 correct        charac~             0               0            2   0.0000260
```

The results indicate that all of the data for the "cloze_response" variable is missing, so we can consider removing this variable when we prepare our data files for analysis. Aside from this variable, the remainder of the data is intact. Analysis on this file for the rest of the course runs yields the same result, in that all data for "cloze_response" is absent whilst the rest of the data is of good quality.

Data Preparation

In order to prepare the data for analysis it would be helpful to compile all of the files for a specific aspect of the user experience, such as video stats, into one file. This will allow us to easily perform analysis to see

how the video stats changed over runs. Additionally, as we have assumed that the videos are unchanged between runs, this will give us more data to assess video performance against each other. In the previous section, we verified that the video stats data files have no issues with regards to missing data, and so there is no need to take any further cleaning steps. Additionally, as previously verified, each video file is in the same format, with the variables in the same order. Therefore we can bind each data frame under the preceding one. However, it would be useful for our subsequent analysis for us to include an identifier of what iteration of the course the data corresponds to. This has been added into the data frame as the variable, "Run".

We now have a a data frame featuring all of the available data for the video stats from runs 3 to 7 of the course. However, this includes 29 variables. It would therefore be useful to refine this data set into the data that we would like to specifically assess and visualise in our subsequent analysis. Before performing a deeper analysis on the video stats data that we have prepared, it would be useful to first motivate this analysis. One way of doing this would be to look for a relationship between the viewing statistics per run and percentage of questions answered correctly in the question response data set. More specifically, does the amount of people who viewed the videos in each course iteration suggest anything about how many questions will be answered correctly? To investigate this, the variables that will be of most use are "Run" and the percentage of learners that viewed a set proportion of the video, ranging from 5% to 100%. We will calculate the mean average over all videos in a run of the course, for each proportion of the video viewed. In order to effectively visualise this, we can then convert this data frame to a long format, as below:

```
head(viewing_avg_long)
```

```
## # A tibble: 6 x 3
##     Run viewed percentage
##   <dbl>  <dbl>      <dbl>
## 1     3      5       74.3
## 2     4      5       73.5
## 3     5      5       79.0
## 4     6      5       78.7
## 5     7      5       74.7
## 6     3     10       73.0
```

For the question respnse files, we can facilitate easier cross-run analysis by compiling all files into one, and adding an identifier labeling the run that the data pertains to. This can be performed in the same fashion as with the video stats data frame, however we established in the previous section that the variable in the question response data, "cloze response", is empty across all files. We should therefore remove this variable from our final data set to clean it. In order to calculate the percentage of questions answered correctly per run, it would be useful to transform the "correct" variable into a more calculation-friendly format. We will add a new column, "correct_binary", where if a question has been answered correctly this variable will contain a 1. If this question has been answered incorrectly it will contain a 0. This will enable us to easily calculate the percentage of questions answered correctly within each run of the course.

Once we have motivated our further analysis of the video stats, we can begin evaluating the effectiveness of each video in order to seek areas of improvement. Before beginning this analysis, we should prepare the data in the appropriate format. We will use similar fields that were used in our previous data construction, but we will use the field "step_position" instead of "Run" as an identifier for each unique video. "Step_position" is a code that is assigned to each video in the course, so each step_position will correspond to a video title, it simply allows for a more concise way of representing the same data. We can then take a mean average over every run for each video , which will inform us of how much of each video was watched by a percentage of users. To make visualisation of this data set easier, we will then convert to long format to give us the below:

```
head(step_avg_long)
```

```
## # A tibble: 6 x 4
```

```
## # Groups:   step_position [6]
##   step_position video_duration viewed percentage
##           <dbl>          <int>  <dbl>      <dbl>
## 1           1.1             99      5       81.2
## 2           1.14           362      5       74.4
## 3           1.17           241      5       77.8
## 4           1.19           348      5       74.4
## 5           1.5            281      5       79.3
## 6           2.1             37      5       78.2
```

In order to judge more specifically which videos lost their audience the most, we will add a new variable to this data set, "percentage_drop_off", that calculates the difference in the percentage who viewed 5% of the video and the percentage who viewed 95% of the video. Why we have chosen these values will be discussed later in our analysis. Note that this data set is in wide form for ease of calculation of the "percentage_drop_off" variable. Thus, giving us the below data frame:

```
head(step_avg_wide)
```

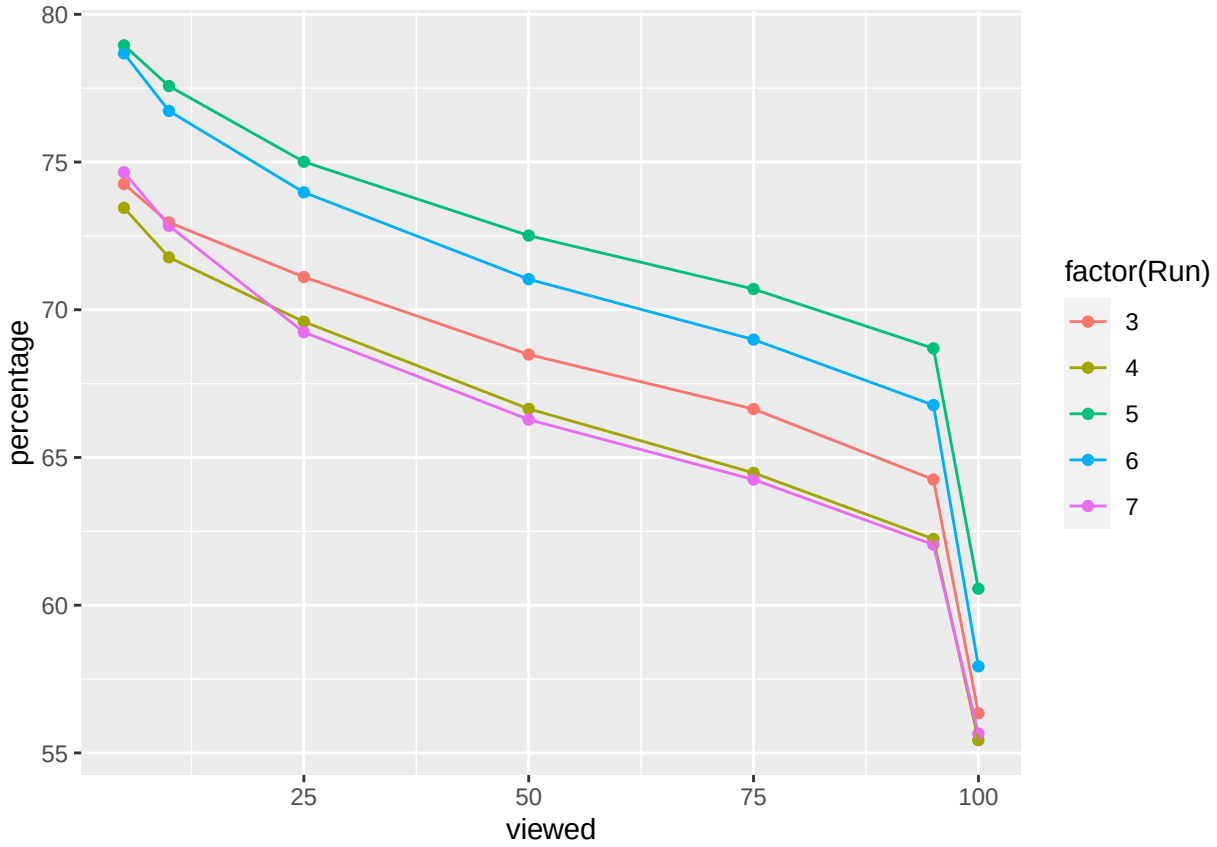
```
## # A tibble: 6 x 9
##   step_position viewed_5 viewed_10 viewed_25 viewed_50 viewed_75 viewed_95
##           <dbl>    <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1           1.1      81.2      79.4      76.9      72.6      69.5      67.8
## 2           1.14     74.4      72.9      69.8      66.0      63.1      61.8
## 3           1.17     77.8      76.4      73.9      70.6      67.5      63.9
## 4           1.19     74.4      73.1      69.7      65.4      63.4      62.1
## 5           1.5      79.3      75.8      68.5      63.1      60.3      57.6
## 6           2.1      78.2      77.4      76.6      75.1      74.1      72.7
## # ... with 2 more variables: viewed_100 <dbl>, percentage_drop_off <dbl>
```

Modelling

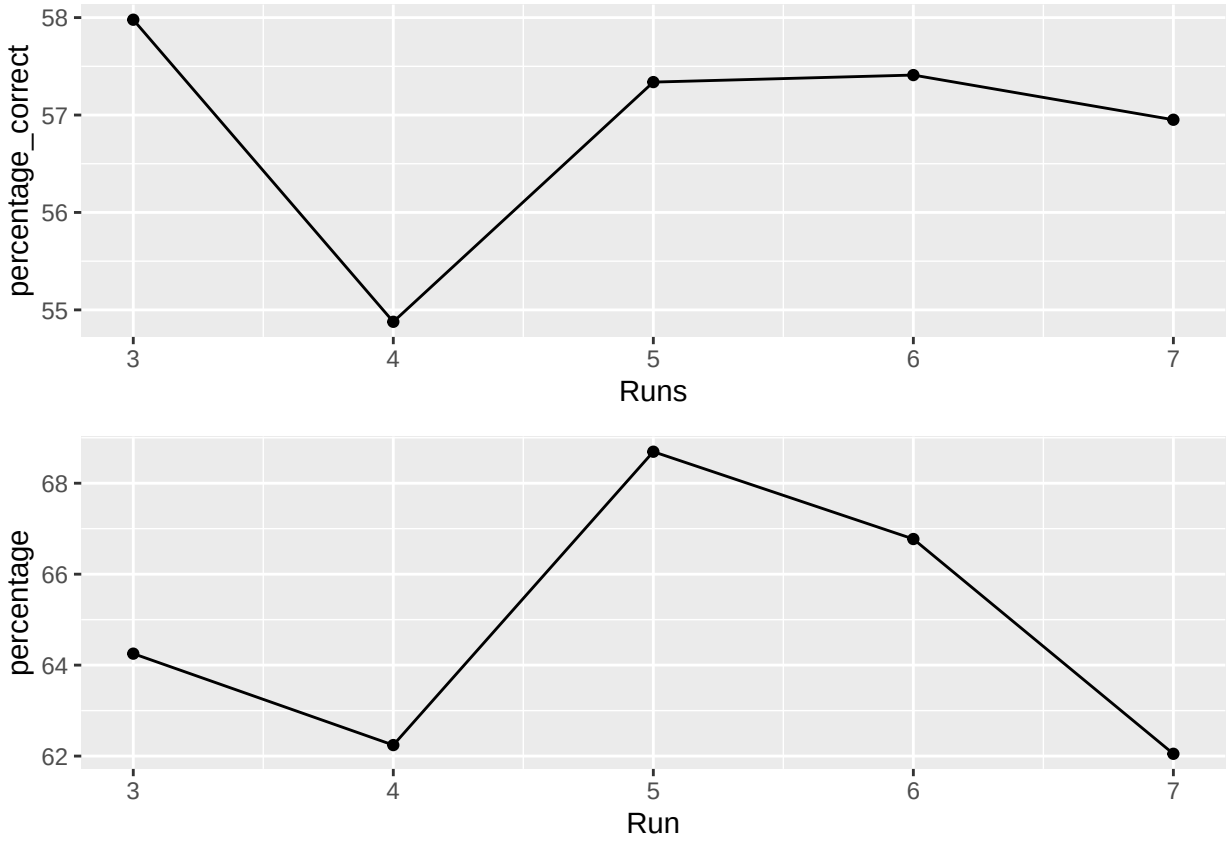
Now that both data sets are appropriately formatted, we will begin our analysis of the data. First of all, we will investigate if there is a relationship between the percentage of people who watched the videos in the course, and their results in the multiple choice quiz. To examine this, we will calculate the percentage of questions answered correctly for the each run of the course.

Run	% Correct
1	54.78
2	58.67
3	57.98
4	54.88
5	57.34
6	57.41
7	56.95

To compare this to the percentage of people who watched the videos, we will first take an average over each run of the percentage of people that watched the video. To be fully clear in our analysis, we will also split these average out into percentage of the video viewed. This gives the following plot:

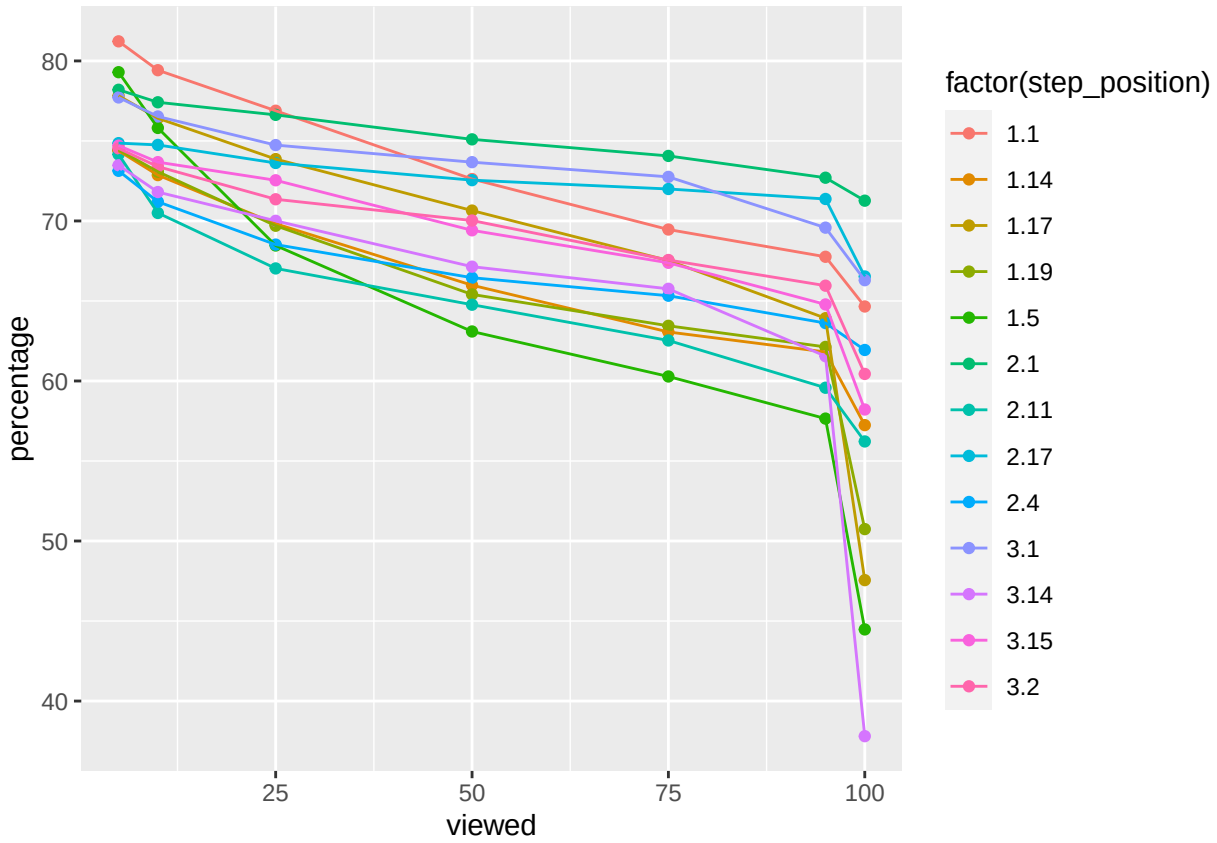


In the above plot, we can observe a steady decline of people watching the video as it progresses up until 95% of the video. A steep decline can then be seen after this with the amount of people who viewed 100% of the video. This intuitively makes sense as many viewers may end the video when there are only a few seconds left when it is obvious that the the important information in the video has been given. This means that the percentage of people that viewed the entire video is not the best metric by which we can judge how many people watched the video. The percentage of users that watched 95% of the video would be a more suitable metric, as they have still effectively watched the entirety of the video just may have not watched the final few seconds. We can now visualise the percentage of users that viewed 95% of all of the videos for each run of the course, and compare this to a plot of how many questions were answered correctly for each course iteration:

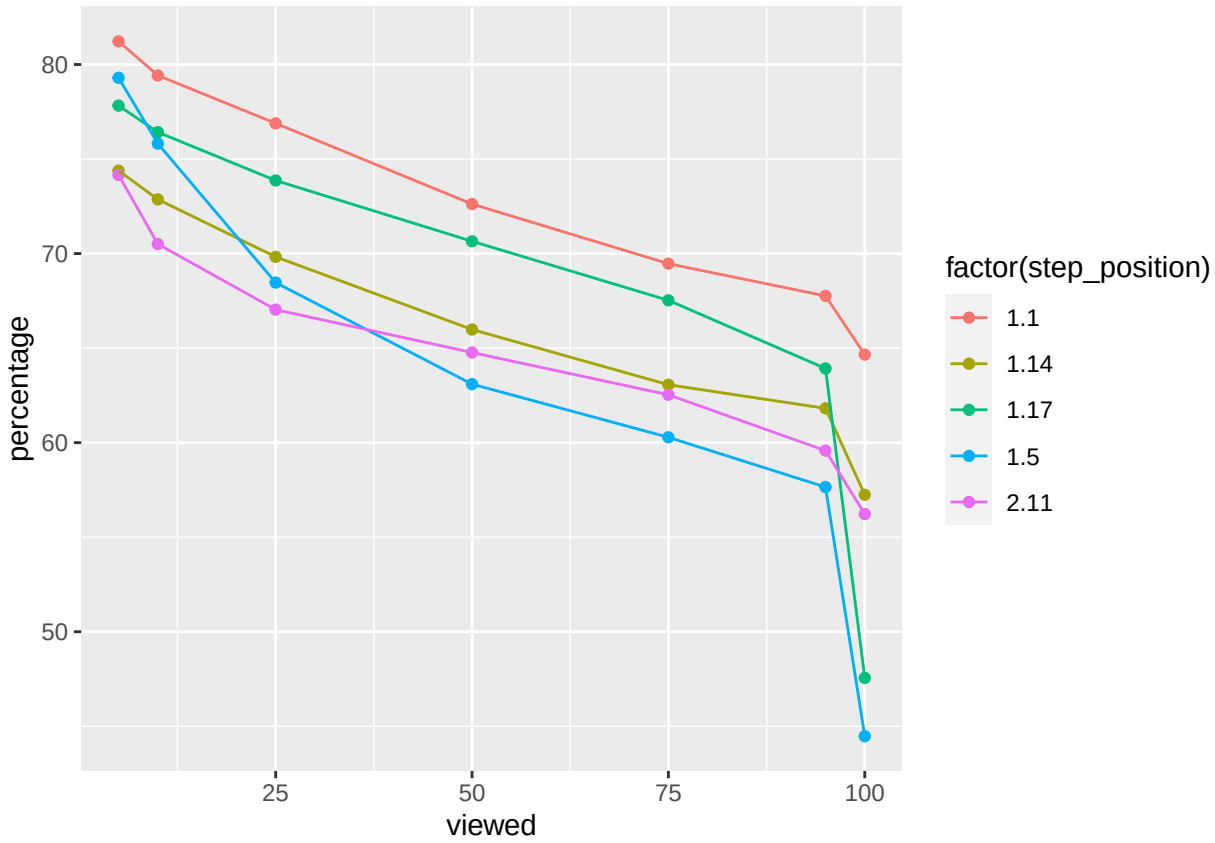


From this plot we can see clearly that in the fourth run of the MOOC course, there was a drop both in the percentage of questions answered correctly and the amount of people who viewed the course videos. Additionally, in the fifth run both the percentage of people who watched 95% of the videos increased along with the percentage of questions answered correctly. In the sixth run, both the video viewership and questions answered correctly stay relatively high. However, in the seventh run the questions answered correctly stays high whilst the percentage of people who fully watched the videos plummets. In the third run, we can also observe that the highest percentage of questions are answered correctly whilst the percentage of users that watched 95% of the videos remained quite low. These observations would suggest that there is a loose relationship between these two, as we see a similar trend specifically from run 4 to 6. Having said this, rather than causing more questions to be answered correctly, a higher percentage of users that viewed the videos may indicate more interested users in that run who may be more interested in the course and thus excel more in the questions on average.

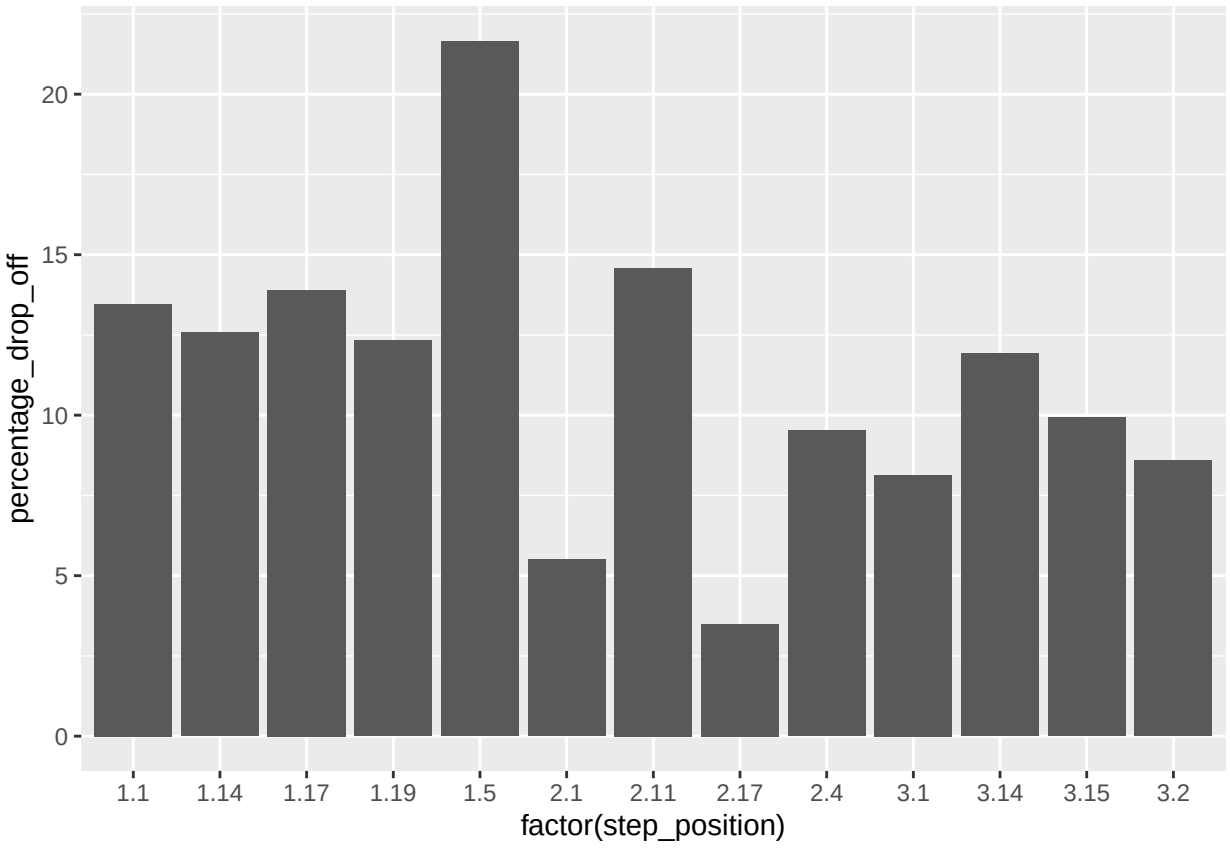
Now that we have demonstrated there is some relationship between the use of the video and outcome of the multiple choice quiz, we should look further into how the videos might be improved. In order to assess this, we will assess the average percentage of users who viewed the videos for varying lengths of time, for each video in the course. As previously noted, we are assuming that the videos have remained unaltered over runs due to the identical duration, so these average will be taken for each video over every run. Plotting the results of this gives us the following plot:



As this plot appears cluttered, it would be useful to only focus on a selection of the videos that had the highest drop off in audience as the video progressed. We will therefore filter the above plot to show only the top 5 videos with the highest % drop in audience to produce the below plot:

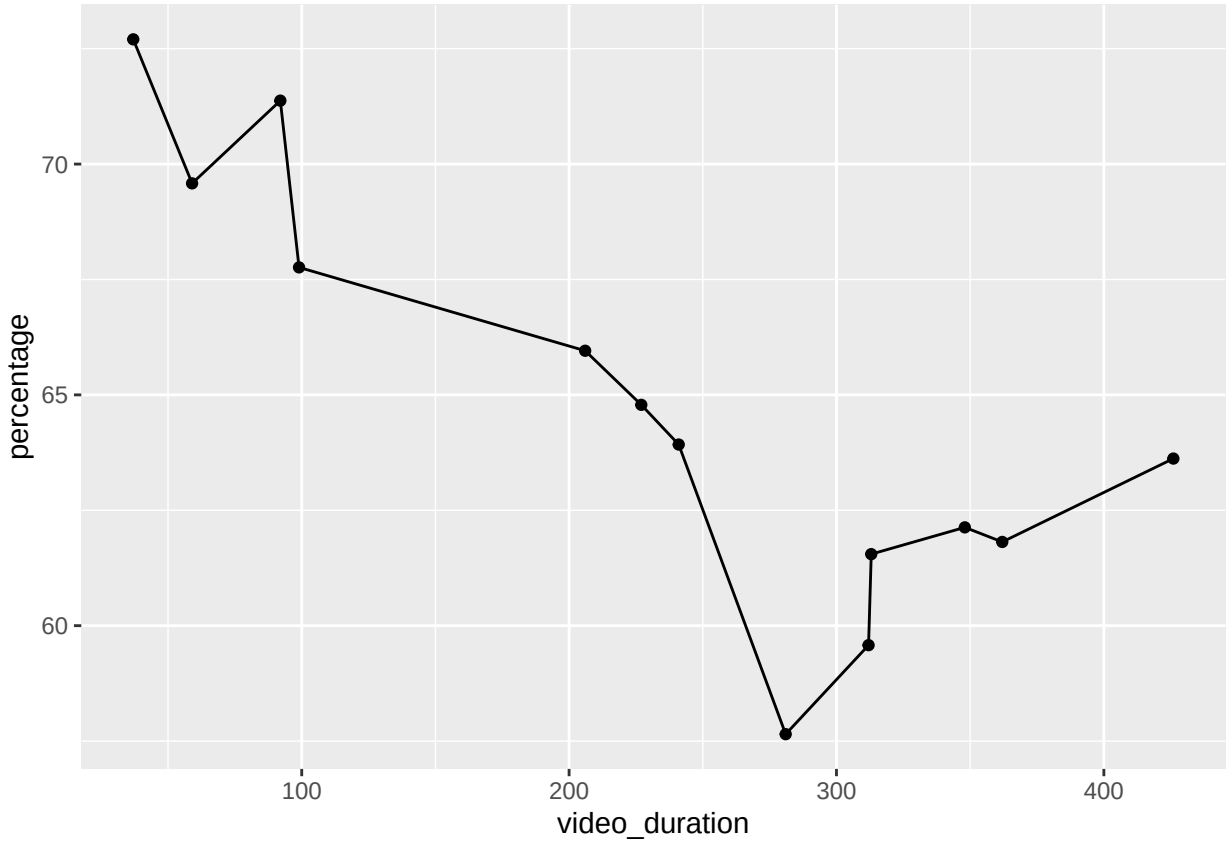


This plot indicates that video with step position 1.5 loses the highest percentage of it's audience. This corresponds to the video called "Privacy online and offline". What is particularly notable with this video, is the steep drop in audience from 5% to 25% of the video. Other videos in the above two plots have a fairly consistent gradient of audience decline as the video progresses. This video displays a significantly higher drop in audience that any other video on the course over all runs from 5% to 95% of the video, as shown in the below column chart:



Again, this shows that the video with step position 1.5, "Privacy online and offline", loses over 20% of the audience from 5% of th video to 95% of it. Comparatively, the video with step position 2.11, "Exploring vulnerabilities in online payments" has slightly less than a 15% loss in viewers. In contrast to these videos, we can observe that the video with step position 2.17, "The million dollar contactless payment", has the highest viewer retention and loses less than 5% of the audience.

Whilst many factors could contribute to how engaged a learner is with a video, an aspect from the video stats data that we could investigate as a driver of this is the duration of the video. To visualise this, we can create a plot of video duration against the average percentage viewed over the course runs. Moreover, we will plot this for the percentage of people that watched 95% of the video, due to the aforementioned drop in viewers who watched all 100% of it.



The above plot shows a clear negative correlation between the length of the video and the percentage of people who watched 95% of it. Interestingly, this correlation is particularly prevalent in the first 300 seconds. After 300 seconds, the percentage of the retained learner throughout the video increased again, but not in the vicinity of retention shown in the videos with small duration.

Evaluation

Deployment