

CSC 8631 Report

Harvey Yuan 0077439

24/11/2021

Aim and Objective

The aim of this analysis is to explore and see if there is any correlation between the time spent watching the course videos and the scores the student receives on the multiple choice questions, collectively as a group. This will allow the course providers to be able to judge if more or less parts of the course should use videos as a medium to teach students.

In addition to this, this analysis will look to find the students that are having/had the hardest time with the multiple choice questions and look to provide extra support and help to intervene at the earliest stage possible by reviewing the number of times the student has gotten the incorrect score as part of the question response.

Technologies and Libraries Used

Throughout this project, there were a number of different libraries which were used with the main library being “ProjectTemplate”.

R version 4.1.2

R is a language used for statistical computing and graphics.

RStudio 2021.09.01 Build 372

RStudio is an IDE that is used for R. RStudio includes an in-built console and terminal that allows for direct code executing and also a pane which allows for files, plots and packages management. Different libraries can be installed through this pane.

ProjectTemplate version 0.10.2

ProjectTemplate (PT) is a system that allows the user to automate medial parts of a data analysis project such as the organisation of the project files and processing data (Darke, n.d.).

This project utilises PT heavily to organise and manage the project files throughout this analysis. The “data” folder stores the raw data that will be analysed as part of this project.

PT also stores pre-process scripts, which will run as soon as the project has been loaded by PT which is stored in the “munge” folder. Similar to the munge folder, the “src” folder stores scripts that can be manually run.

dplyr version 1.0.7

The dplyr library allows for the user to use a set of verbs to write code for common data manipulation steps. It allows the user to use familiar words when scripting analysis and pre-processing scripts (such as filter), both for easy understanding of the code and proof reading (Hadley Wickham, n.d.).

readr version 2.0.1

The readr library provides for an easy way to read the csv files that are being used as part of this analysis. As well as being easier to for users to understand and read the data, it also makes analysis more reproducible as base R functions inherit behaviours from the operating system and environment variables, as such, importing code from one environment to another using readr will work without issue (Grolemund, 2016).

ggplot2 version 3.3.5

ggplot is a system for creating graphics (where the data is a selected data frame). The data is either provided by the user or created in script, which allows for said user to iteratively add new layers, components and functionality. ggplot will be used to demonstrate the analysis as part of this investigation.

Data Understanding and Preparation

As part of this project, there were a number of different data sets which accompanied each run of the FutureLearn session ranging from Enrolments to Weekly Survey Responses of the students. However, as the aim and objective of this analysis is to identify if there is a correlation for time spent watching the videos with the scores students receive, only some of the data provided will be useful to this analysis. Upon inspection of the whole, raw data set, there was a number of different data sets which were not pertinent to this investigation. Within some data sets, columns would not be populated, whereas others were void of data. In some instances, the run did not include a data set such as video stats and team members from the first run.

With this in mind, it was decided that the “Enrolments”, “Question Response” and “Video Stats” from run 3 to 7 would be used for the analysis as it was only these runs which had the complete data that would be relevant. As such, there would be 5 data sets for Enrolments, 5 for Video Stats and 5 for Question Responses. These data sets were chosen for their representation of the knowledge of the participants. The question response data would allow for the analysis of the ratio of correct answers to false answers and the video stats data set would allow for analysis of the percentages of the course videos watched. Finally, the enrolments data will allow for the analysis to learn how many people are interacting with the videos and multiple choice questions.

With this being the raw FutureLearn data, some data manipulation and transformation was required to to ready it for analysis. In the first instance, all data sets were combined into one data set according to their categorisation producing three data frames, one for question response, video stats and enrolments. This will allow easier analysis of the runs as a whole. The column names of the data sets were also slightly adjusted for easier human reading.

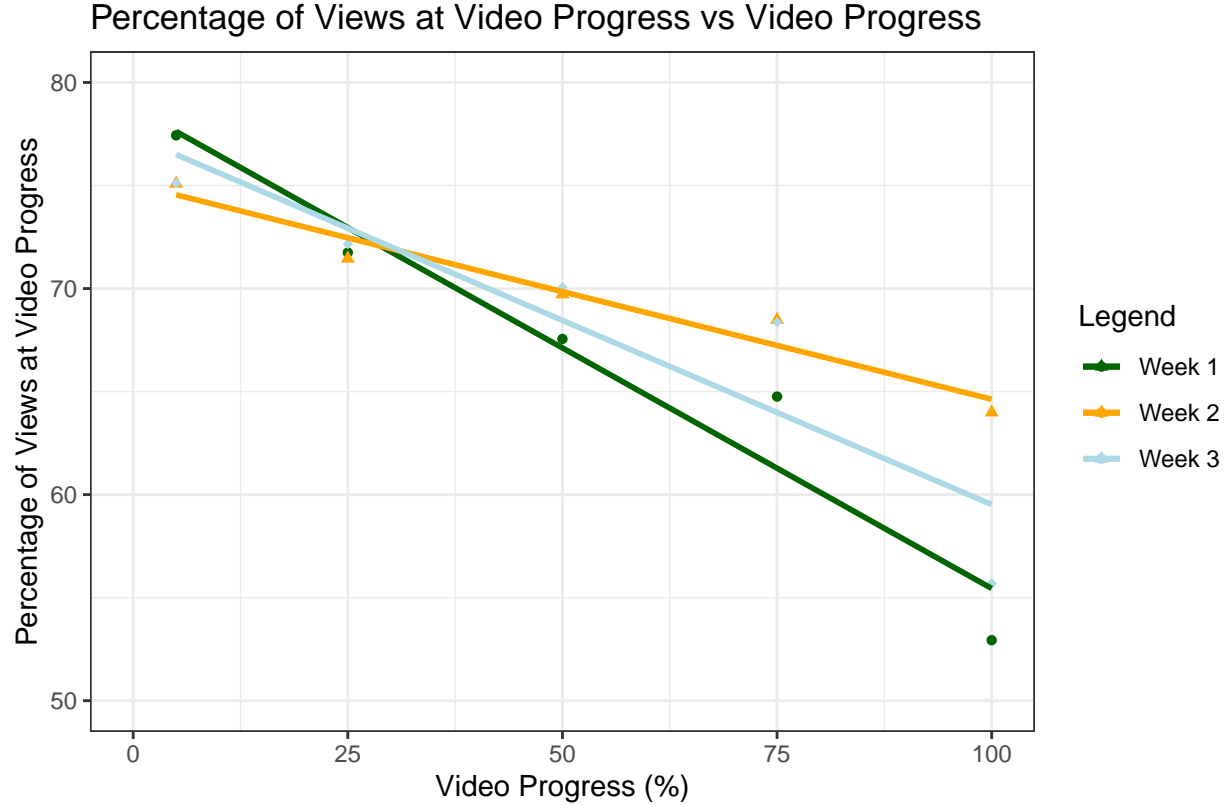
From here, individual video stats were broken down into each step and either summed or averaged depending on the type of data that the column was populated with. This creates a combined version of the video stats data set of all 5 runs as a collective. This was then filtered into individual data frames for each week to be analysed against the question response. As the question response data didn’t require columns to be either summed or averaged, there was no need to sum or average the combined data and just filter by week. With the question response data being broken down into their weeks, the comparison between video stats should be relatively straight forward.

As such, there are three main branches of data that is used are part of this project, the enrolment, video stats and question response data.

In addition to the data frames, a number of different values and variables were created to aide in the process of the analysis such as “Total_no_students” (total number of students enrolled) and “sum_qr_w1” (total number of responses received for week 1 across the question response data frame which is a combination of all 5 runs).

Analysis and Results

As the aim of this analysis is to explore if there is any correlation between time spent watching the course videos and the scores that the student receive against the multiple choice questions, it was decided that, in the first instance, a graph would be created to compare the percentage of video completed by the students against the portion of the video by week. This will allow for visual representation of how much the videos the students are completing and for comparison for the ratio of students that receive a correct mark.



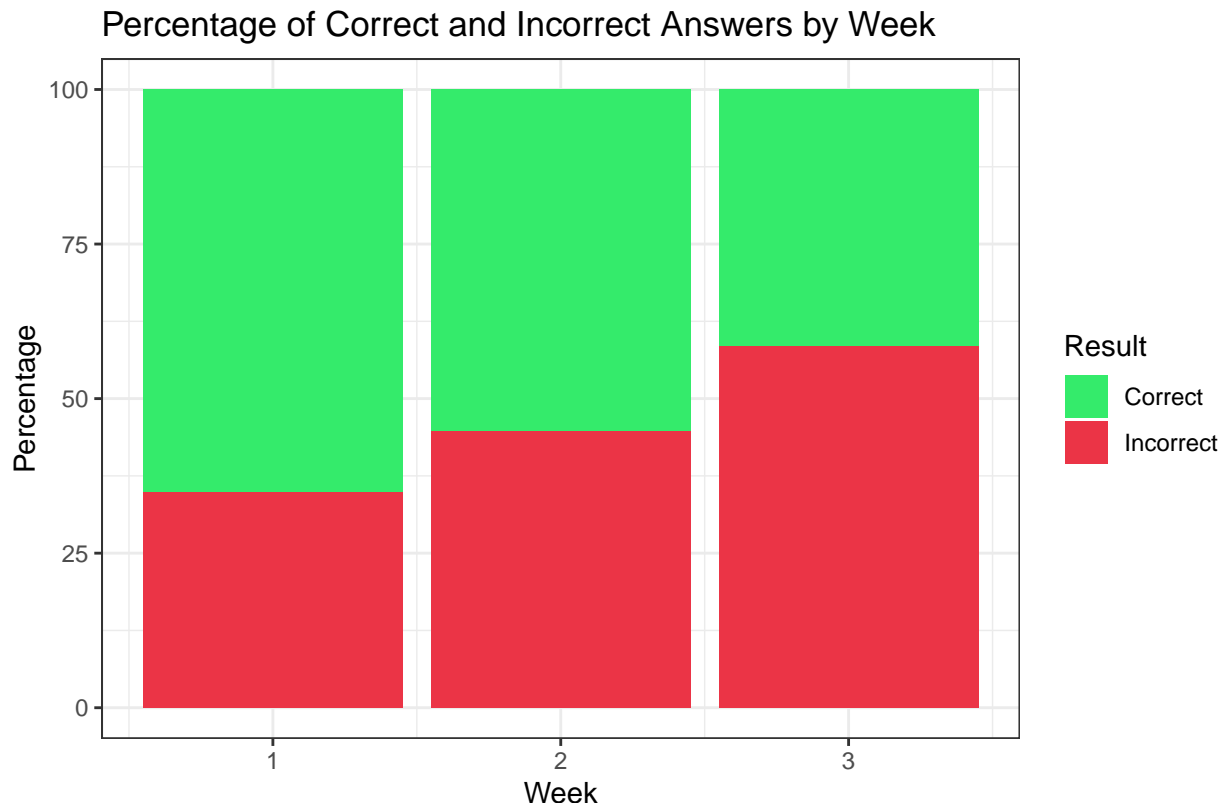
Graph 1

The plot above shows that the mean of the percentage watched by students compared to the video progress, which is indicative of how much of the students are watching the videos to completion. From the trend lines on the graph, it is evident that the videos from week 2 were watched to completion the most in comparison to week 1 and 3, ranking rank 3 and 1 respectively. This is based off the assumption that the video stats data provided by FutureLearn does not contain students that watched the videos more than once. By assuming the videos were not watched multiple times by the users, the assumption of a higher portion of the students are watching the videos which allows for more straight forward comparison when relating this data to the data from the question responses. This graph also allows us to see that there was a relatively large drop off in video completions, approximately 12.5%, in week 1 and 3 from 75% to 100%.

Table 1: Input Data

Video Progress	Percentage Views W1	Percentage Views W2	Percentage Views W3
5	77.4356	75.0890	75.1095
25	71.7496	71.4530	72.1620
50	67.5520	69.7180	70.0655
75	64.7572	68.4805	68.3655
100	52.9348	63.9895	55.6910

The input data reaffirms and shows that there is has been a drop off in video completion in week 1 and 3 whereas the video completion stats remain high for week 2. What this graph and data doesn't show, however, is the number of individual students that watched the videos. In the above paragraph, the was an assumption made that is the videos were not watched multiple times by the users. This assumption was made so that the we allow for each view to be a unique individual.



Graph 2

Table 2: Questions Response Data

Week	Total Responses	Correct	Correct (%)	Incorrect	Incorrect (%)
1	39551	25729	65.05272	13822	34.94728
2	17850	9857	55.22129	7993	44.77871
3	19597	8118	41.42471	11479	58.57529

The above stacked bar chart shows the percentage of correct and incorrect answer against the week progression. From this graph, the data shows an increasing amount of incorrect results as the weeks progress. This could be a result of the course content becoming increasing more difficult as the weeks progress. The percentage of incorrect answers across the three weeks seem to increase approximately 12% each week, however with the limited amount of data, it will be difficult to tell at which point this value will plateau as it is unreasonable to assume that the incorrect answers will just continue to increase at a rate of 12%.

However, comparing this data to the previous plot, the data implies that the percentage of views at the video completion does not affect the results of the multiple choice question. There are a number of reasons as to why this may be the case, one of which is that the videos are too difficult to understand. It is possible that within the course, the foundations of the course were not explained in a way in which a beginner will be able to pick up easily, as a result making the course content and videos difficult to understand.

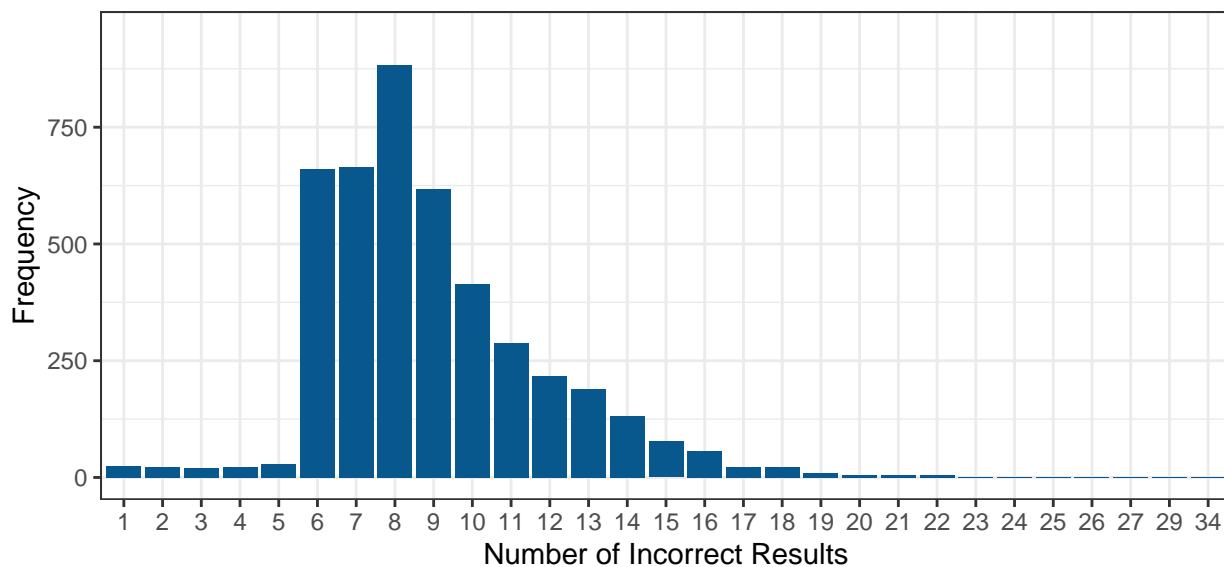
Another reason as to why this data implies that the percentage of views at video completion doesn't affect quiz outcome, is that although the videos are in the same weeks as the multiple choice questions, it does not cover what is asked during the multiple choice questions and is covered in another, non-video, part of the week. Without access to the actual course material, it will be difficult to determine if the content from the videos teach the students what they are being quizzed on which will affect this investigation.

Table 3: Most Incorrect Responses by Individuals

Week	Number Incorrect
1	34
	29
	27
	27
	26
2	16
	16
	15
	15
	15
3	44
	42
	42
	39
	38

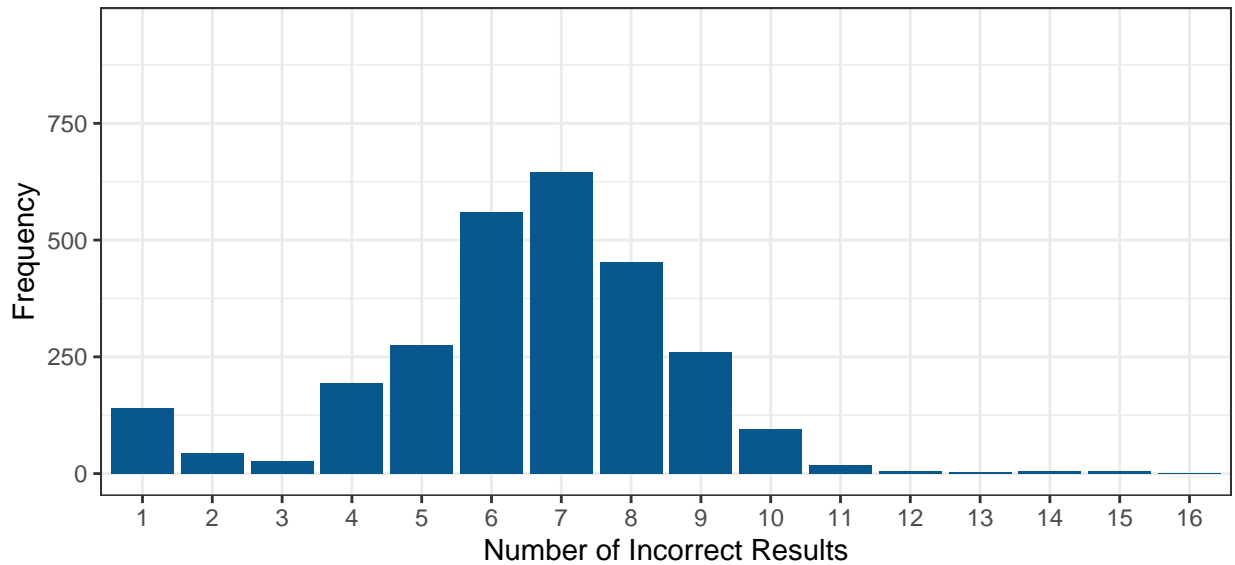
The above table shows the highest number of incorrect responses by individuals per week. This data shows us that in week 1 and 3, individuals that were repeatedly getting incorrect values were getting questions incorrect much more frequently than in week 2. This may be an indication that the higher percentage of completion of videos as shown in Graph 1 has a positive effect on the individual question responses.

Frequency of Incorrect Results Week 1



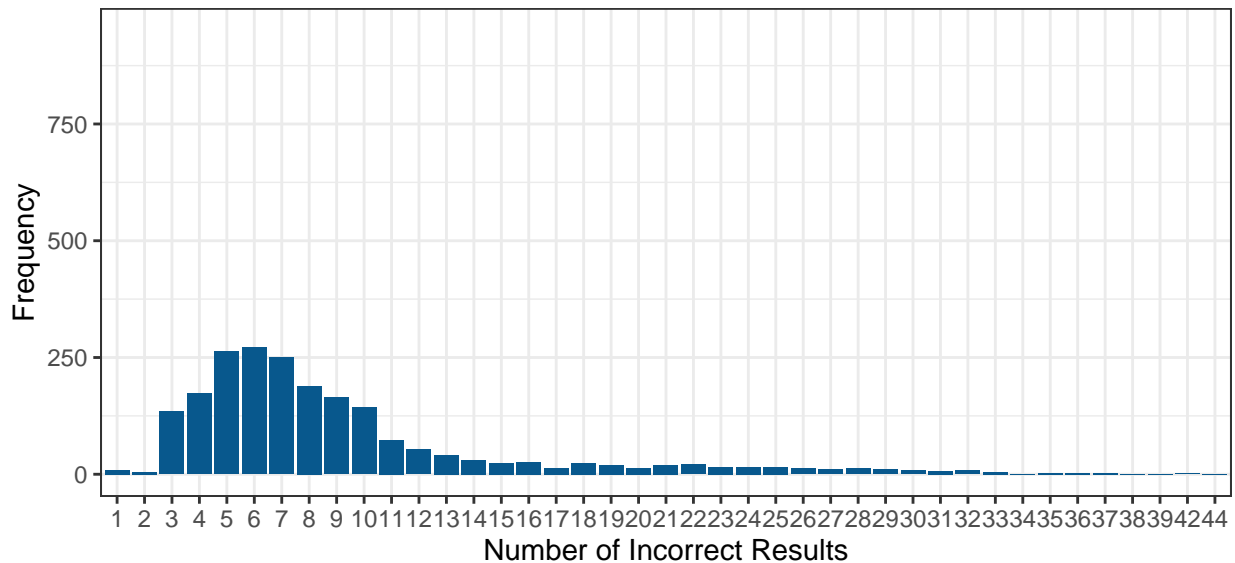
Graph 3

Frequency of Incorrect Results Week 2



Graph 4

Frequency of Incorrect Results Week 3



Graph 5

Graphs 3, 4 and 5 visualises the range of incorrect results and the frequency of which they occur. Assuming each individual attempts the multiple choice question until they get the correct answer, we are able to see, on average, the number of attempt the individuals take. With week 1 and 3, we see a positively skewed bell curve which indicates that a large proportion of the users are receiving the correct mark before the median of results. This indicates that a small number of students are receiving a large number of the incorrect scores which could be affecting Graph 2 as the small group of people are receiving a large number of incorrect answers.

From all three plots, the most frequent value of the number of incorrect results sits around 6-8. This suggest that the most of the students get the correct multiple choice answer relatively quickly in comparison to the students that fall in the end of the plot.

In Graph 3, most of the students receive a correct score in the multiple choice questions before the median value which corroborates with the Graph 2 plot showing week 1 data with a low number of incorrect scores with very few students with a large number of incorrect scores.

It could be argued that the small range of number of incorrect results in Graph 4 is a direct correlation to the percentage of view at video completion in Graph 1 as there is a higher number of people watching the videos to completion which results in a smaller number of students getting a large number of incorrect results. However, much like the results from the data in Graph 2, it is difficult to determine if the course content from the videos is having a direct impact to the scores.

The data in Graph 5 visualises that the peak of the data is not as high as the previous two plots, with the frequency slowly and gradually decreasing which suggest that the number of students getting incorrect values is more spread out. This confirms the week 3 column in Graph 2 as it shows a larger proportion of students with incorrect answers which increases the ratio of correct to incorrect answers in the column.

Summary

With the aim of this investigation to explore if there is any correlation between time spent watching the course videos and the scores received by students on the multiple choice questions, the video stat data sets and the question response data were analysed and compared to explore this. The first graph shows the percentage of views at video progress vs the video progress itself which allowed for the visualisation of the time spent watching the videos by the students. This showed that the students watched, to completion, week 2 videos the most which the videos and week 1 and 3 were comparatively less. The second graph created was to visualise the ratio of incorrect and correct results the students received between the 3 weeks. The results in this stacked bar chart showed an increasing trend of incorrect values by approximately 12% each week. The data found here would suggest that the video completions did not have an effect on the results, however, it is possible that the video content is not directly linked to the multiple choice questions. It is also possible that the students are building up their foundations of the course and therefore will take time to get to grips with the content as it becomes increasingly hard.

A table listing the value of most incorrect results showed that in week 1 and 3, the top 5 results were much higher than that of week 2. This data suggest that the video completion stats from Graph 1 has a positive effect on the students, however, without direct access to the videos and questions, it would be difficult to determine.

Lastly, 3 bar graphs were created to show the frequency of students with the number of incorrect results. The all three graphs on a whole, they showed a positively skewed plot which indicates most students completed the multiple choice questions before the median value of attempts. Graph 3 and 5 had a much larger range of incorrect answers, as shown by the graphs and table 3, whereas Graph 4 has a much lower range. This indicates that most students were getting the correct answer earlier than they were in week 1 and 3. This data complies with the original investigation of exploring the time spent watching videos with question output as the week 2 videos were watched to completion the most and the results suggest the students performed the best in the week 2 quiz.

Bibliography

Darke, P. (n.d.). Reproducible data science techniques in actuarial work. Retrieved from <https://philipdarke.com/reproducible-actuarial-work/exercise1> (Last accessed 27th of November 2021)

Hadley Wickham, R. F. (n.d.). dplyr. Retrieved from dplyr part of the tidyverse 1.0.7: <https://dplyr.tidyverse.org/> (Last accessed 27th of November 2021)

Grolemund, H. W. (2016, December). Data Import. Retrieved from R for Data Science: <https://r4ds.had.co.nz/data-import.html> (Last accessed 26th of November 2021)