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 mange 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.

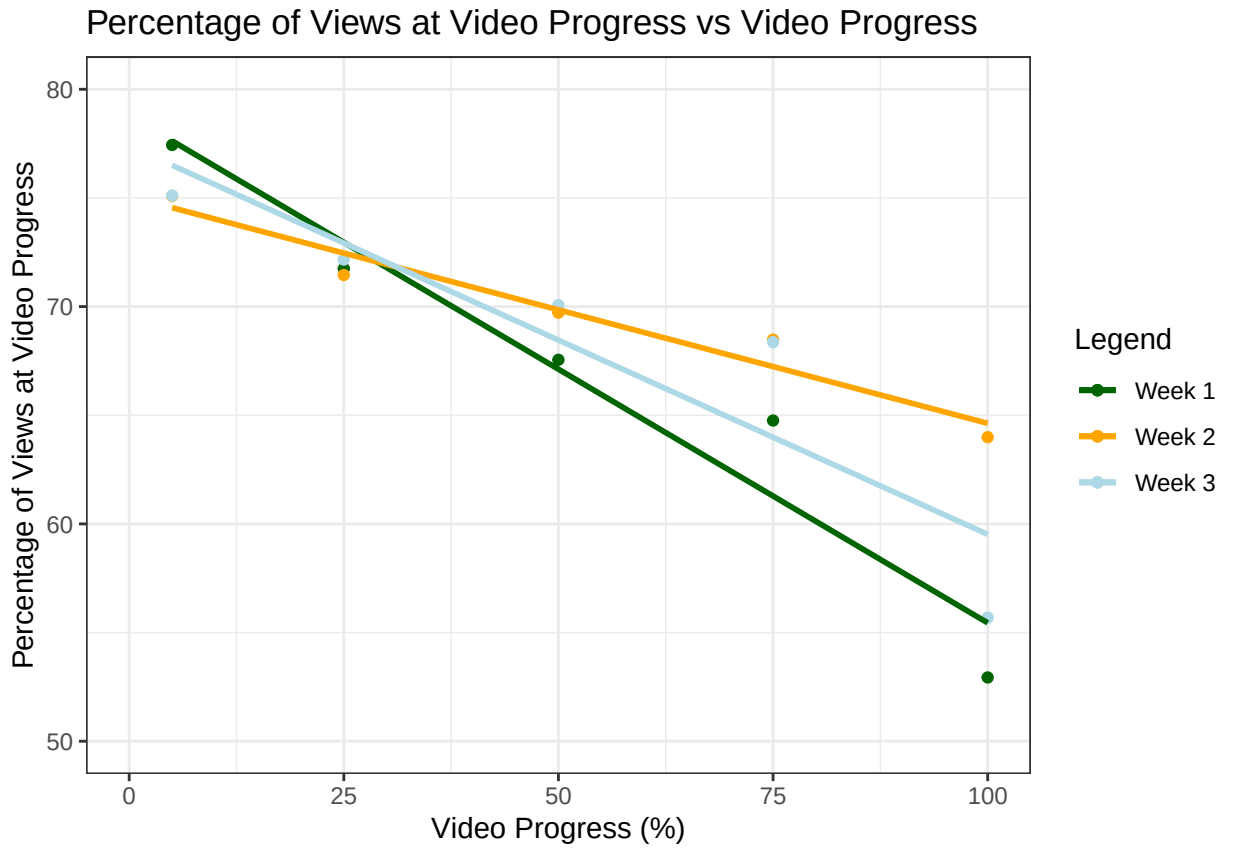
From here, individual video stats were broken down into each step and either summed or averaged depending on the type of data that the column was populated with. This creates a combined version of the video stats data set of all 5 runs as a collective. This was then filtered into individual data frames 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

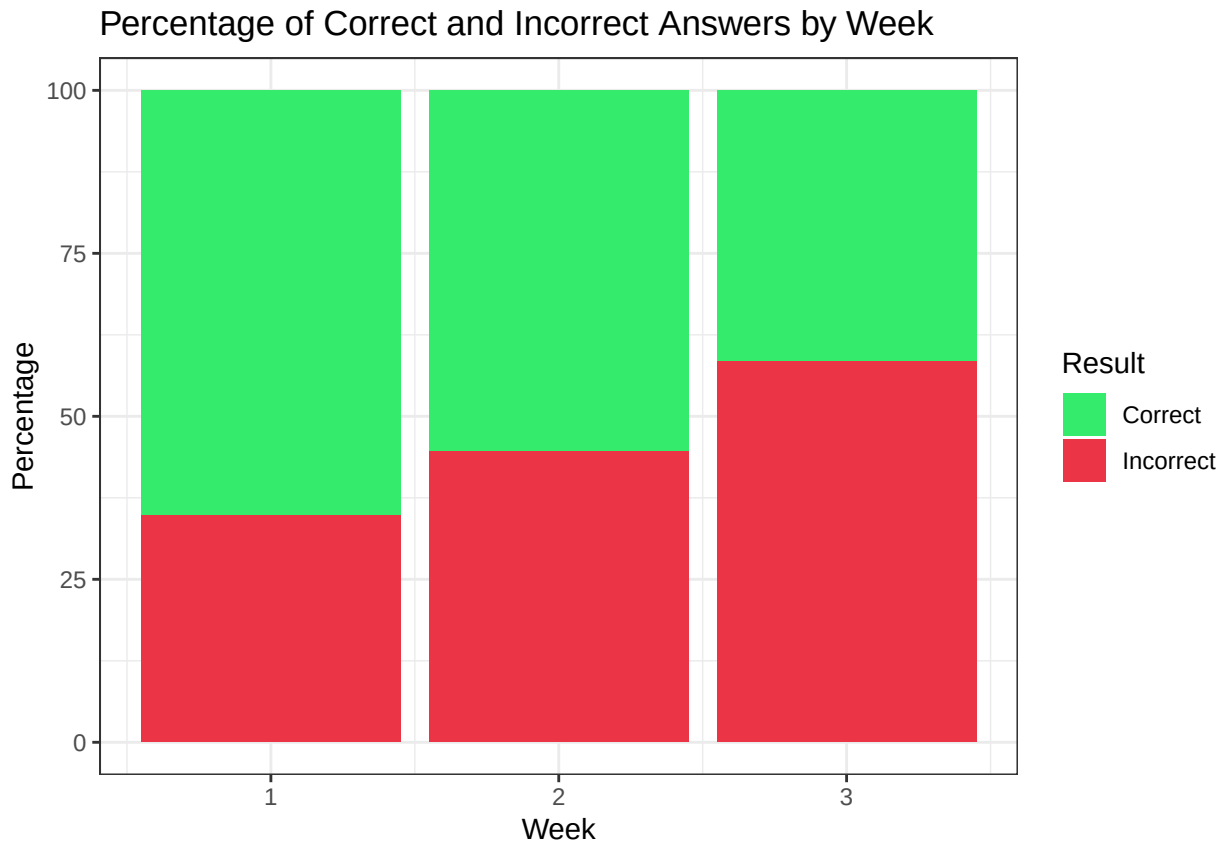
As the aim of this analysis is to explore is there is any correlation between time spent watching the course videos and the scores that the student receive against the multiple choice questions, it was decided that, in the first instance, a graph would be created to compare the percentage of video completed 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.



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%.

```
##   Video Progress Percentage Views W1 Percentage Views W2 Percentage Views W3
## 1              5              77.4356              75.0890              75.1095
## 2             25              71.7496              71.4530              72.1620
## 3             50              67.5520              69.7180              70.0655
## 4             75              64.7572              68.4805              68.3655
## 5            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.




```
##   Week Total Responses Correct Correct (%) Incorrect Incorrect (%)
## 1    1           39551   25729    65.05272     13822      34.94728
## 2    2           17850    9857    55.22129      7993      44.77871
## 3    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.

However, comparing this data to the previous plot, the data shows that the percentage of views at the video progression does not affect the results of the multiple choice question.

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)