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.

Libraries Used

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

ProjectTemplate

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

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

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).

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.

With this being the raw FutureLearn data, some data manipulation and transformation was required to to ready it for analysis.

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)