Name: Shubham Rajesh

Student Number: 210462188

Summary for Exploratory data analysis of the survey for Future Learn’s Cyber Security course

Future Learn provided the following data for the analysis of the increase in the dropout rate for their Cyber Security course, Archetype survey response, enrolments information, leaving survey response, students’ questions response, students’ step activity, and team members managed the course, and video statistics. Only the leaving survey response data is used for analysis for this report. It is assumed that there were four intakes (i.e., November, February, June, September) for this course in the ‘period of late 2017 and 2018. This data is from the students who registered in the Cyber Security online course were asked to fill out a survey for dropping out during the course. Around 400 students’ records were analysed for patterns and to establish when and why they dropped out. The key objective was to discover what was the reason for the increase in the dropout rate for the Cyber security course; hence this data was chosen.

As a result, these two questions were included in my analysis:

- What were the most and least common reasons for students dropping out of the Cyber Security course? Week wise analysis for the most common leaving reason?
- Which week did the most and least participants leave the Cyber Security course? Also, what was the most last common step completed by students before dropping out of the course?

The interpretations were expressed and aimed to achieve the two business goals that had been developed through visualisation. Two CRISP-DM cycles were employed since there were two business objectives. Data from late 2017 and 2018 was chosen because it was consistent over four, five, six, and seven runs in 2017 and 2018. The data was cleaned and rearranged in order to construct the visualisation and conclude the study. Appropriate graphs were drawn, and assumptions were made that are mentioned in the main study as part of data analysis. Also, the report was generated and presented for the Deployment stage. R markdown was used to construct the report. Formatting possibilities were possible as R markdown was used to write the report which made it easier to develop a structured format. Once the data analysis was complete and crisp DM steps were completed, we could conclude that the majority of students chose "Other", and the second most popular answer was "I don't have enough time" so it means students didn’t feel they should invest their time in this course. Also, by observing the last completed steps analysis, it can be speculated that most people completed the last step (i.e. Step 3.2) but didn't complete the course, so the "Other" leaving reason can't be neglected. Therefore, the major reason for the increase in the dropping out rate in the third week can be concluded that learners found the quiz hard and could not clear it. And for the first week, it can be concluded that most students who left the course completed Step1.2 (i.e. Why are you here?) So, most people understood the course outcomes are irrelevant to what they wanted and then decided not to continue with the course.

Reproducibility and CRISP-DM were the two main approaches employed in the preparation of this report and data analysis project. For repeatability, we utilised Project Template, which helps with the process. Project configuration, package loading, data loading, data munging for cleaning, and data analysis may all be handled using it. It generates a directory in which we may store our files as well as the settings we'll need to finish our project. It has a standard setup and loads r script files in the working directory automatically. It isn't entirely automated, and some user participation is required, but only in tiny ways. Instead of having to run each r script manually, we can import data files from the data directory and automate the initial data munging when the project is performed to run each of the r scripts manually.