Project Summary and Critical Reflection

Project summary

The data sets for carrying out this project provided the opportunity to focus on various ares that could probably produce some interesting findings. The area in which this report focuses on is participants engagement over the seven runs of the course. The reports focuses more on some specific runs (runs 4-7), because these runs had similar number of participants, but different completion rates. Moreover, these were the most recent runs, so the results form their analysis may be more up to date compered to results generated from run 1. However, some comparative analysis that includes all runs, is carried out when there are sufficient data, in order to identify patterns that occur in all runs.

The analysis behind the report, is carried out using CRISP DM methodology and consists of three cycles. The first cycle focuses on some general statistics on the number of people that participate compared to the number of participants that actually complete the course. This cycle investigates whether differences observed between participation and completion are connected with the gender or age range of the participants. This first cycle analyses primarily runs 4-7, because they have quite similar number of participants but different completion rates. First of all for the chosen runs (4-7), gender does not seem to affect initial engagement. This means that the proportion of men and women who decide to participate in the course are quite similar. However when it comes down to the completion of the course, it is observed that for all runs, regardless of whether the completion rate was higher, there were more male participants completing the course. So this is an indication that female participants stop engaging at some point of the course. Now, regarding age range, it is found that in runs where the completion rate was higher there were more people that belonged in the 18-25 and 26-35 age ranges, when compared to run 5 which had very low completion rate, and had more registered participants that belonged the age range of 46-55, 56-65 and >65. So there is an indication that when younger people participate completion rate is higher. When looking at the age ranges of people who completed the course it was found the higher completion percentage was in people who belong in the 26-35 and 36-45 age ranges. This was a bit surprising because the percentage of the people in that range, who registered for the course, was less when compared to other age ranges. However, here it should be mentioned that for both results above the data that were used, included many empty entries. This means that the analysis carried out was based only on a sample of the participants, and this makes the results more of an assumption rather than a certain conclusion.

The second cycle focuses on how the engagement varies between the weeks of the course as well as between each step and investigates the reasons for the variation. First of all it was found as the weeks pass there is a decline in engagement. However the decline in engagement during week one is much faster, when compared to the other weeks, where the decline in much slower and the engagement is more stable. So it seems that participants who start week 2 generally tend to stay until the end of the course and engage in the third week as well. Moreover, it is found that for all runs, except run 7, there is not a significance difference in step engagement between the people who completed the step and the ones that did not. So this means that in general when people engaged with the steps they also finished them. However, there were some exceptions where it was observed that people had engaged with a step but did not completed it. These steps were identified to be the quizzes and the test. After further analysis it was found that there was one quiz where engagement was high. The difference between this quiz and the other ones was the number of questions that the quiz contained. So by analyzing the relationship between engagement and length of the assessment it was found that in general participants seem to prefer shorter quizzes and tend to engage more with them.

Finally the third cycle focuses on engagement with the videos. This analysis was based again on runs 4-7. First off all it was found that there was a very steep decline in the last 5% of the videos. This was observed

for all videos in all runs, so it means that participants stopped engaging with the videos when they had watched up to 95% of them. Moreover it was found that for some videos the dropout rate was much lower compered to others, and engagement was very stable up to 95% of the duration. A main reason for this difference was video duration. By analysis the relationship between engagement and duration it was found the the two have a strong negative correlation, meaning that in general the longer the video the lower the engagement. However there were some outliers in this pattern meaning that some shorter videos had higher dropout rate when compared with longer ones. This could be an indication that the content of these videos were of great interest among the participants.

Critical Reflection

CRISP DM methodology was applied throughout the project. In general it was very useful as it created a framework for the analysis. Moreover it was very helpful because it acted as a guide for the steps that need to be followed in order to plan, organize and execute the analysis. One of the most helpful parts was the first step which had to do with the business understanding because that is how I was able to formulate some interesting questions in order to carry out further analysis. However a limitation identified with this framework, is that sometimes it is difficult to understand the business and the problem, because the problem and the aim of the analysis is not clarified, so the analysis could focus on an area that the business is less interested in when compared to other areas. Moreover a second limitation is that the general background is not taken into account. For example, for this specific project, the problem with people not registering for the course, could be that there are other organisations that provide similar course, and people prefer those. Moreover people not completing the course could be due to increased working hours, that leave people with less time to engage with the course. However this more general understanding is not part of CRISP DM, so it could be the case that some factors that affect an organisation may be disregarded. Moving on to the other tools introduced in the course I found project template very useful because it helped me organised my code much better than I would have if I was not using this. Also the fact that the data could be accessed without loading each file in R was extremely useful. Moreover, ggplots were very handy, because they had more functionalities and more advanced plots could be produced. These plots were also quite appealing for creating a nicer report.

In order to apply all the above techniques and find out the focus area for my analysis, some assumptions were made about the data. The first assumption was that the business was interested on the engagement on the course and wanted to find out which factors may affect engagement and which section of the course were more popular and why. This was a major assumption because if the organisation was not interested in that then the analysis is not that useful. The assumption that a specific participant can only register once for the course is also made, because if a participant has registered more than once, then this participant may have completed the course once, but may have not completed it many other times. This could lead in having participants not completing the course even if this is not true. It is also possible that if this participant had decided to do the course more than once, their engagement behavior may be similar over the registartions and again this would lead to some inaccurate results. Moreover, another assumption about the data sets, was that they included all participants and were complete. This was a major assumption as well, because the certainty of some results were based on the fact that the analysis was performed in the whole population of participants and not just a sample of them. So if the data are incomplete and there are more participants that are not registered, this could probably make the findings of the analysis less accurate. One more assumption is that the details that some participants had provided, like the gender and age, were true. Because especially for these data, that were available only for a sample of the clients, if they were not correct then the results would be even less accurate. The final assumption made about the data, is that these are the only data that the company has collected about the courses. If this is not true and there are more data that could explain what affects engagement, there may be other causes that are the main factors for affecting engagement and make the results of the currect analysis less strong.