Critical Reflection

Critical Reflection

My project 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, for some parts of the analysis I decided to include all runs, because I wanted to see if there are patterns that can be found in all runs.

I decided to have three cycles of CRISP DM in my analysis. The first cycle investigates whether differences observed between participation and completion are connected with the gender or age range of the participants. The results show that there are approximately equal numbers of males and females that register for the course, but a larger proportion of males that finishes it. When it comes to age it seems that runs where there are more younger participants, have higher rate of completion, and the highest rate of completion is among the people who belong in the 26-35 and 36-45 age ranges. 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 after week 1 there is more stabilization in engagement. Moreover I found that there are some steps that people start but do not finish and these are the quizzes and the test. It was also found that it in general participants seem to prefer shorter quizzes and tend to engage more with them. Finally the third cycle focuses on video engagement, and the main results are that participants tend not to watch the last five percent of the videos and in general prefer shorter videos. However there were some outliers of some videos that had long duration but high engagement. Now in order to reach these conclusions I made some assumptions about the data which I discuss below.

CRISP DM methodology was applied throughout the project. In general I found it very useful as it created a framework for my analysis and it made the whole process less chaotic . Moreover it was very helpful because it acted as a guide for the steps that I needed to be follow 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. This step was also the step that enabled me to look the data carefully, and get a first impression of how they can be used in order to formulate my questions and provide the business with useful insights. 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 when moving to the data understanding part I had to make some assumption about what some data mean, because this was not clarified. These assumptions are discussed below, but the fact that I interpreted the data myself could lead to making some inaccurate assumptions. Another 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. However there were some point on the project that I was not sure were to put some code chunks, because what I was doing sometimes was not manipulate the raw data, but more creating new data frames from the data. So at these points I

was not sure if I needed to put these code chunks in the source section or the munge section. In the end I decided to include them in the munge section, in order to make the source file shorter. Also the fact that the data could be accessed without loading each file in R was extremely useful because it saved me so much time. 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. Finally, interestingly enough I found version control quite useful. I did not expect that as it was something that I had never came across before and did not know exactly how it was working. However adding and committing my changes regularly helped me keep track of my progress. Also, even though I had some problem in the beginning with uploading my repository to git, after I managed to do it, I felt more secure that my work was not going to get lost if something happens to my computer.

In order to apply all the above techniques and find out the focus area for my analysis, some assumptions were made about the data and based on these assumptions I formulated my questions and also based my conclusions. 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 registrations 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 on the csv files, 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. Moreover, when using the video statistics I assumed that the numbers under the 9 - 15 columns were the percentage of people that had watched up to that point of the videos. It seemed to be a rational assumption, but if these numbers represent something else then my analysis could be less accurate. Also there were some data sets that were empty, so I assumed that these would not be useful for my analysis. However this was just an assumption and in reality, if present, they could give a different point of view as to how engagement is affected. 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 current analysis less strong.