CSC8631 Critical Reflection

D Walmsley, C1053068

21/12/2021

Daniel Walmsley

Critical Refelction

This purpose of this project was to create a data analysis pipeline that followed the CRISP-DM methodology. It was designed to follow a project template directory guide in order to help ensure that the project was fully reproducible in the future. R scripts and packages were used to complete the analysis and the positives and challenges faced will be discussed in this report.

The first stage of the Crisp DM process is *Business Understanding*. This section for was fairly straight forward for an educational setting as they are providing a service to their learners. The key principles for the university are student enrolment, retention and outcomes, as well as the quality of resources they provide and the impact the resources has on learning. Although gaining this business understanding was straight forward, it was difficult to apply this to a find a clear line of enquiry further in the Crisp-DM process.

The next stage of the Crisp DM process is *Data Understanding*. This stage was more difficult as there was a large number of data files provided. Looking through the files for relational data or data that could help in the investagtion was vital in this stage. There were files with large numbers of records and others with no records at all. Taking notes and assessing all files for each repetition of the course was vital here. Drawing out where relationships between files were and how the files could be linked with key indicators. Identifying potential numerical data that could be used in future calculations were also identified. R Studio was able to show brief summaries of the data files using the glimpse command and this made note taking and further investigation an easier process than looking at each csv file one by one.

The third stage of the Crisp DM process is *Data Preparation*. Here the data is cleaned and linked to create new data frames where certain data is removed or tables are joined. What was difficult in this phase of the project was understanding the structure and the commands to clean and link data. As the new data files were created they were saved as part of the munge files but what was produced was not always returning the desired outcomes. This lead to a greater understanding of the R code but meant a lot of unnecessary tables were created due to the trail and error nature of the work. When an error code was returned or a table didn't look as intended, it was difficult to know what parts of the scripts to change. Once certain commands were used successfully, they became more frequent and heavily relied on. This had both positive and negative implications for the report. Positive in that it ensured reproducibility but negative in that a wider range of commands were not used to show a wider range of what can be achieved when using R scripts and R studio.

Files added and moved to the munge folder resulted in numerous problems and made it difficult to keep a structure to the report and analysis. This issue was down to files not being created and saved in a numerical file structure. If repeating the process a more structured system would be used to name scripts. Scripts should be given a name that demonstrates what they create or contain, being easily understandable but also simple to use in other further scripts if required. These file names should be appropriately numbered in the order the are expected to be ran in order to avoid errors encountered in this project. Not having a structure

has been the cause of a number of errors during this analysis project. The project template and CRISP-DM systems are built for structure and efficiency, organising scripts appropriately would have strengthened the process.

The next stage of the Crisp DM process is *Data modelling*. This section helps to create visual representations of threads of investigation and their results. As in the *Data Preparation* section of this report, issues revolved around creating scripts that produced exactly what was trying to be displayed. There was a number of issues here. Firstly, the tables produced in the previous sections had to create a table that contained not just the data needed to be displayed, but that the data was in th required formatting. Column headers needed to be in the correct place, variables had to be identified. Identifying what data was to be used for each axis could be problematic depending on whether the data was continuous or discrete data. All of these things can be done in different ways but understanding the correct way to do this was often more challenging, than creating the script. Understanding the structure of a command was also a challenge. Commands in ggplot for example, certain commands follow a specific order. Adding appropriate data labels was difficult depending on the type of data used. A common error was using the count function which in a visual as ggplot would return the number of the variable and not the count of the numerical value of the y axis.

Again a challenge throughout was not understanding the error codes returned in R studio. So when a issue with a code was returned, it was difficult to directly address the issue using the error code and often a case of try again. This was a time consuming part of the project. Formatting data into percentages and with 2.d.p was also difficult and although there is a number of different solutions to this, this project has been heavily reliant on a finding a command that was successful. This command isn't the simplest command but once it had worked once was able to be replicated and adapted where necessary.

Looking reflectively the impact of trying to produce visualisations that were successful impacted on the success of the enquiry. The enquiry throughout this project would have been more in-depth but lost this due to the ability to use the coding language.

The final stage of the Crisp DM process is *Evaluation stage*. This section helped to focus on what had actually been discovered from the analysis but also what could be targeted if the process was to be done again. It also allowed me to look back at the business understanding section and reassess what the university would want to know from the data. Once completing the evaluation section it was clear to see what the next steps required should of been but also raised the issues talked about throughout this critical reflection.

If this process was to be completed again, clear improvements could be made from this report. Gaining a better understanding of R would greatly benefit the analysis and threads investigated, with also an assessment of how much time is valid per script or visualisation. Ensuring productivity as well as quality. Following a more structured method would help when trying to reproduce the analysis.