# CSC8631 Design and Implementation Report

## Selina So A3032929, 3rd December 2021

**Description of events:** This assignment was focussed on the investigation of data from the '*Cyber Security Safety at Home, Online, in Life'* online course, which is a course delivered by Newcastle University on the FutureLearn platform. The key focus of the assignment for module CSC8631 was reproducibility and this was achieved by using a set of tools and techniques to complete the assignment. The tools consisted of R coding language to perform the analysis, ProjectTemplate (an R package) to create a template project directory, Git to track my changes and RMarkdown to create my report. The recommended methodology to use was the CRISP-DM (Cross-Industry Standard Process for Data Mining) to structure the lifecycle of my project, in addition there was a requirement to run through each key phase of CRISP-DM at least twice. We were instructed to focus on 4 out of 6 of the CRISP-DM phases (Business Understanding, Data Understanding, Data Preparation and Evaluation). We were provided with 53 data files to explore for the analysis. However, it was clear that due to the time constraints of the assignment, it was only realistically possible for me to select a subset of data files to analyse.

My interest within Data Science is predominately on Natural Language Processing (NLP). I was very keen to work with the 'Weekly-sentiment-survey-responses' data files. However, after closer examination of the data files and through discussing with other stakeholders, I found that there was not enough information to meet the business objectives of the project. The data also contained a lot of noise, which would require more time to learn how to tidy before use, and the data appeared biased towards positive feedback. Therefore, I decided there was not much value in analysing these data files.

Instead, I decided to select the data files titled 'Video-stats', as due to the time constraints and as a beginner in R, this data file was the most appropriate data file to use. This was because the data within these files was complete and of good quality, with no missing data. This meant that I would not have to spend too much time cleaning the data, and instead was able to focus on learning how to use the tools and techniques. The 'Video-stats' data files had the same dimensions (i.e., same number of columns, rows and column names) across all runs, therefore merging the data files together would be less problematic compared to the other sets of data files.

I followed CRISP-DM to help formulate my key data analysis questions by performing initial data analysis to understand my data and identify areas of particular interest which may draw interesting insights for the business (i.e., the course provider and designers). Then I investigated the data further for each question by going through two cycles of data analysis: In the first cycle, I prepared and analysed only the run 3 data to further narrow down the areas of interest. I then combined the data from all seven runs to perform a meaningful analysis. I developed a set of data visualizations to identify valuable insights within the data. Based on the insights gained from the visualizations, I was able to evaluate the data and findings.

Although the requirement was to only demonstrate the use of 4 phases within CRISP-DM, I was able to include some aspects of the remaining 2 phases (Modelling and Deployment). The visualisations can be either classed as Evaluation, or Modelling. To increase the readability of the report, I decided to include the visualisations under the Modelling phase of CRISP-DM. Recording a presentation about my investigation can be part of the 'Deployment' phase.

**Reflection:**

**Feelings:** For the assignment, I decided to use the 'video-stats' data files as they looked the most manageable in terms of data cleaning requirements. However I had still under-estimated the time it would take for me to learn how to prepare the data for analysis in R. The other tools that were used as part of this project (Git, CRISP-DM, ProjectTemplate, RMarkdown, GitHub) were also entirely new to me and I had to learn them from scratch, which meant I occasionally felt overwhelmed by the assignment.

Therefore, I am very pleased with being able to complete my assignment on time and with some interesting results, despite the fact that some of the analyses did not show any interesting behaviour, and some questions could not be answered based on the available data. Most importantly, I am most pleased with being able to analyse my data, construct my report and do version control using R, Git, GitHub, ProjectTemplate and RMarkdown, having started with very little R experience and no Git/GitHub experience. Also, I appreciate having the opportunity to learn on an interesting research topic with up-to-date applications, and to get into the habit of using industry best practices. I feel more confident about using these tools again for future projects.

**Evaluation - what didn't go well:** At the start of the report write up, I followed CRISP-DM rigidly to ensure that I had captured all the areas of CRISP-DM on my report. However very soon, I struggled to structure my report so that it would reflect the structure of my investigation. I found CRISP-DM was only partially helpful in finding a structure for the report that reflected the multiple 'branches' of analysis that I introduced by working on three different questions. I decided to exclude the irrelevant areas and reorganised certain headings, which made it easier for my report to flow naturally and increase its readability.

As I was not familiar with how to build my project using R and ProjectTemplate I ended up building numerous files within my 'munge' folder, each consisted of small scripts of code. It was difficult for me to remember which code was used for which purpose. I tried to tidy my codes within the 'munge' folder, however I ran into issues which stopped my report from knitting. I spent some time learning how to revert back to my previous changes using Git, so although it made me quite nervous at the time, I learned something new about Git at the end of it.

**Evaluation - what went well:** Git was a saviour tool for my project because it helped me recover and revert back to my previous version of my project when I had encountered a problem from trying to tidy my 'munge' folder. The ProjectTemplate package provided a very useful template to structure my directory and automate certain tasks. Communicating and working with others throughout my project greatly helped me in completing my project. For example, at the beginning of the project I spent a long time deciding on the data files to use, and through communicating with others, I was able to gain confidence to pursue with the most sensible data file given the constraints surrounding the project.

**Analysis:** I found the ProjectTemplate package very useful, and I realised that I saved a lot of time and effort from running all my codes every time I opened the project, because ProjectTemplate automatically ran them. However, the automation feature only works well if the files within the directory are numbered in the correct order. This made it difficult for me to tidy my files in the 'munge' folder because I had lost track of which codes had to run in a particular order. Hence the issue I had with knitting my project after trying to tidy my 'munge' folder. Thankfully Git enabled me to revert to the previous version of my project, so I was able to continue with my project. Git has really shown the importance of version control and how it could help ensure one can always recover changes that were made.

**Conclusion:** I learnt that for the next project, I will write all the relevant scripts required to pre-process the data for a particular graph or model in one file within the 'munge' folder. This will enable me to refer back to certain codes with ease. Also, I will continue with the habit of regularly committing using Git, so all my changes are recorded. I will also continue collaborating with others, as communicating my thoughts and ideas with others had benefitted me greatly and provided me with the confidence to pursue my investigation in the direction that I felt were logical.

**Action Plan:** For my visualizations, I had considered colour accessibility. However, due to time constraints and being a beginner in R, I was unable to apply these features to my visualizations. For my future projects, I will take time to apply these features to ensure my report is accessible to all readers. In addition, I will spend less time reading on methods and instead start coding as soon as a plan of action is decided. I will also work on developing my coding skills in R to write optimised codes for the project, and to undertake more complex analyses in shorter periods of time.