CSC8631 – Critical Reflection

**Summarise Work Carried Out**

**No more than two pages**

**Consider merits and limitations of CRISP-DM and tools and techniques introduced in the course**

**PARTICULARLY INTERESTED IN: State assumptions made about the data and how they motivated design decisions**

In this project I have used the CRISP-DM methodology to produce a data analysis pipeline in such a way that that the analysis is fully reproducible and would be with additional data. Tools such as *ggplot, rmarkdown, Git* and *ProjectTemplate* have significantly aided this process and in this brief report the merits and limitations of these, along with CRISP-DM, will be discussed as I critically reflect on the project as a whole.

Following the CRISP-DM process allowed for an intuitive approach to the construction of my data analysis pipeline. The first step of this process is gaining business understanding, followed by data understanding, then undertaking data preparation, modelling, evaluation and deployment. The extent to which CRISP-DM could be applied in the business understanding stage of this process was limited due to the only information on the customer in the scenario being obtained in a brief project outline. Much of this stage in the CRISP-DM process is focused on gaining a deep understanding of the business, the key persons and departments, and liaising with each of these to identify problem areas where a data mining project could prove fruitful. As none of this was applicable in this project, this section of my project was brief as I listed the benefits of learning analytics and why Newcastle University may want to apply it to their online resources. Having said this, it did provide a good justification for the analysis.

The next step in CRISP-DM is to understand the data. A large swathe of the CRISP-DM process for this step requires detailing the data mining process, however this is made irrelevant in this project due to the data being given to us. Additionally, understanding the data was made difficult due to the granular nature of it, in that for each run of the course there were various categories of data file that were computed, with the same format preserved between runs. For this project, basic assessment of the data would be made easier by first engaging in the data preparation stage of CRISP-DM and create a “master file” for each category, before engaging in some basic data diagnostics. This would save having to examine each file one-by-one. Therefore, in this case some initial data preparation was useful before examining the data, with the assumption that the files in each respective category were of the same format and would continue to be as they are produced in the future.

For the data preparation stage in the CRISP-DM process, a major driver in its construction was the requirement of reproducibility such that the analysis can be easily replicated with the same data and performed on future data. There are several assumptions that were made in the design of this. Firstly, the data file format would have to be the same as they are for this analysis, and more specifically have the exact same column names. Much of the data preparation in the munge file is based on specific column names. This was to account for if the variables were ordered differently, they would still be prepared correctly, but must have the same variables name. Secondly, as the course run corresponding to the file is taken from the name, and this drives much of the subsequent analysis it is critical that the filename remain in same exact format and filetype. For the video stats file, this filename format would be “cyber-security-“, followed by the run number, then “\_video-stats.csv”. I assumed this would be the same each time in order to minimise the manual interaction with the file, as this allowed me to automate picking the course run number for each file. The alternative was for the file user to input the first run of the files uploaded manually, which would then lead to more assumptions being made in the data files, such as that the files that the analysis is ran on have to be corresponding to sequential course runs. *Dyplr* was used quite heavily in the data preparation phase of my data analysis pipeline, particularly the pivot\_longer() function for transforming my data into long form and the bind\_rows() function for compiling individual files into a data frame without growing an object incrementally (which could end up being very computationally intensive if the data analysis pipeline was used for many files). The use of the pipe in *dplyr* was also useful in keeping my code more concise and readable. ProjectTemplate was incredibly useful for this phase of the analysis, with the “munge” folder along with the “cache” allowing for a fast, seamless loading of the project, without the need to run-through any code. Both this and the *dplyr* package have proved invaluable in ensuring the reproducibility of this phase in my data analysis pipeline. I also found the steps for this phase in CRISP-DM to be useful, providing the roadmap for an efficient workflow by cleaning, constructing then formatting the data. While intuitive, I found the prompts from the CRISP-DM guide helpful in this case.

As modelling was not the focus of this project, this section in CRISP-DM was of little use. However, *ggplot2* and *rmarkdown* were hugely effective in both the production of my plots and insertion of them into my project report. As *ggplot* enables the user to create plot objects, this means that the plots can be input into a document with minimal code, enabled by *rmarkdown’s* ability to source from different files. Additionally, I found *ggplot* to be significantly easier to use than the native r plotting, allowing for plots to be built in layers. As well as this, the stat\_summary() function facilitated me being able to plot averages from across my data without the need to transform the data into smaller data frames. In calculating these averages, I assumed that the videos used in the course remained unchanged for each run. This assumption was motivated by the conclusion that the duration of the videos does not change for each run, but impacted my design decisions in that I opted to take averages of audience retention over course runs. In the context of my analysis this means that the resulting values carry more weight as they come from a larger data set, but would not be appropriate if the videos had changed.

I found the evaluation phase of CRISP-DM useful in ensuring that my evaluation is thorough. It was helpful to have a checklist on hand that documents all the many facets of a data analysis pipeline to consider upon its review. Considering things such as future data pipeline development, how results can be enhanced, possible failures and ranking project outcomes allowed for a thorough evaluation of the project as a whole.

A second assumption made,

In modelling, mention ggplot

dpylr