# Critical Reflection

## Introduction

The project focuses on Learning Analytics the data sets were generated and collected from a cyber security online course. The process of data mining uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) model that include descriptions for tasks within different phases of the project. A portion of those phases were appropriate for the project undertaken. In addition, the project was created within the ProjectTemplate framework and Git Version Control was used to make potential collaboration possible. The strategy of undertaking this project will be discussed as well as the technologies used.

## Reflection

To begin with, the first step was to review and familiarize with the technologies available and required for the project. ProjectTemplate creates a directory for your computer with a predefined structure that one need to familiarize with in order to fully exploit the frameworks capabilities. By automatically creating all directories the use of Git version control is just a single console line that quickly enables you to start working without having to think about directories and structure. Next. the data were collected and loaded. The data were multiple csv files with similar names quickly it became apparent that the course had 7 runs. Differences on each run was easy to spot suggesting that some sort of data selection should take place. After reviewing all data sets in terms of context some assumptions were noted for what information each data set can provide. The initial logic was to observe the data as two entities the learners and the teaching material.

1. What data can provide background information for the learners and how these information can shape the analysis?
2. What data can provide statistics regarding the teaching material in terms of performance and engagement?

At this point the CRISP-DM methodology was used to guide each step of the project processes. The Business understanding is a crucial phase of the project it is needed to define the business model and the business objectives. Information about Learning Analytics used to derive some of the relevant objectives. CRISP-DM has an access situation step that partially can be utilized for this project. For example, there is no information regarding costs, risks cannot be derived because there is not sensitive information the data were pre-processed to not provide these information. The importance of having this steps and planning ahead is quite clear although some did not apply to this project.

For the data understanding and data preparation default R functions as well as dplyr and tidyverse libraries were used. Dplyr introduced an easier way for data handling it took a lot of trial and error but the result was worth it. I was getting results in a quick and reliable pace. Although, a handful of function felt much easier done in default R. As the pre-processing scripts were written additions and various changes as well as creation of new files occurred everything needed to be added and commited on git. As first time using git bash very quickly I understood the importance of keeping log entries for each change on a personal level and as a business policy.

Data cleaning and data construction were made reproducible within the ProjectTemplate the data were loading automatically and new data frames derived from data cleaning were ready to be used in the analysis.

The automation saved a lot of time and reduced back-tracing. While the functions created could be used again with minor changes in future projects. After data construction and defining analysis goals, visualization was essential to discuss findings in the analysis. For this ggplot2 was used similarly with the dplyr a lot of trial and error was needed to learn to utilize it properly. The results produced were worth it, it is a powerful tool that i will be using in future projects.

The analysis consists of 4 main data sets the enrollments, step responses, step activity and video stats. For the enrollment the analysis were partially fruitful, some findings where produced regarding learner background such as learner count across runs and full participation rates. Although the investigation for shared background could no be complete due to the majority of the data being unknown. In the step responses percentages for the completion of the quiz based on each week were produced ranking the first week on each run as the most successful. High completion rates where found even though the performance on the quizzes were mediocre. The article step type is the core of the teaching material and the videos shows similar completion rates. In the step activity analysis weeks were ranked in terms of step completion and each step type were categorized and compared. Finally, in video stats an investigation commenced about video views and duration the results indicate thats video duration affects viewing.

# Conclussion

The analysis conducted followed the CRISP-DM model in the relevant for this project requirements. The analysis had an exploratory approach and each finding led to another using logic and data understanding. Numerous question emerged as result were produced for other objectives, instead of following each emerged question I decided to keep a general and cohernat throughtout investigation.