critical reflection on csc\_8631 project

Here, on the basis of experience of getting used to with the tools and techniques that were employed for the analysis done on the EDA Project, the report is developed. The project was based on how well the learners get hold of the course materials of the Cyber security course which was made available in online sessions during september,2017. There were many observations made like enrollments leaving response, question response etc. From these observations, I selected the statistical observations generated on Viewing of the Video content for my Analysis.

As instructed, the Analysis report is prepared by help of the CRISP-DM methodology. Here, the CRISP-DM methodology is followed for the data analysis part which is similar to the AGILE methodology which is used by software developers for the software development projects. The CRISP-DM suggests the steps that are needed to be followed for processing the analysis that are to be made with the chosen dataset. The process involves about understanding the business requirements and then deciding what modelling needs to be used for the analysis process. Preparing the dataset based on the requirements, cleaning up of the data. Evaluating the analysis models that was used and then deployment is to be completed. Once the analysis is completed it can be finally used by the business ends.

With the use of Project Template the analysis report is made. As guided through lectures, Project Template is one of the key features for using the reproducibility of code as per the CRISP-DM methodology. The Project Template creates the sub directories where we can load the data files, do some pre-processing techniques such as cleaning and arranging the data, modelling and analysing the data, the global configurations of loading the libraries and instructing whether we need to catch the cache files or load the libraries or do the pre-processing step everything can be handled, we can also make use of some functions which needs to be used for analysis process, for running the analysis multiple times without any delay the cache files can be stored in the CACHE sub-directory which is useful, if there are any graphical summaries, the plots can be stored as a pdf, the use of r-markdown for making report which would be generated and can be stored in the REPORTS folder. Also we can clone our Project Template work on the GitHub, creating the repository and making different commits at different working times.

The Project Template seems to be very useful as we can easily generate the report from any local machines without setting up the working directories. I felt CRISP-DM as a good approach for my analysis. I have followed the CRISP-DM methodology in making a exploratory analysis which I feel like this approach helped me to proceed in a step by step manner rather than getting confused with what approach and steps I need to use. Every step in CRISP DM works with feedback from previous milestones. As the initial analysis, business understanding of data is assumed with or without prior knowledge and data is defined with those foundations. As the data is pre-processed, it is easy to understand and modify any specific understanding of data. With the diversification in use of directories to pre-process the data and then exploit data by diving deep into graphical - numerical summaries helped in building an error proof analysis. Further, building any model on that data is very efficient as by this step, information about the data became very clear. Possibility of modifying previous steps after data modelling helped in running the model in a loop by reducing errors and improving model efficiency. Alongside of providing steps to follow while analysing the data, this methodology also helped in generating an understandable report if the flow is adhered to. Before deploying model and reporting, the flow allowed to run through all the steps. CRISP DM methodology would not be possible to follow if it were not due to project template. Project template is unique in handling multiple loops. Use of repositories for every step played a key role in following CRISP DM methodology and delivering a report after analysing all possibilities.

The Project Template is legitimately excellent tool to use as I can split and store the required R-codes in different sub-directories instead of keeping it in a single R-code. It was easy to pre-process the data I loaded, add some function and was to easy to call all these files from a single analysis file using the load project command after setting up the working directory.

In beginning of the analysis as the data is pre-processed, it is easy to understand and modify the data. With the pre-processed data it is easy to do some graphical / numerical summaries for building the analysis report. In the analysis performed on video stats, objective was to explore the data, visualize the increasing and decreasing patterns of stats given and predict the results. Post arranging and going through the data I tried to identify the drop percentage in the viewership co-relating the loaded data files. Applying the loop concept and applying mathematics I came to know about the rise and downfall of the Course with help of the 1 year apart and half year surveyed data files. Post it I used the R-tools to co-relate the different numerical data columns and predict the results for different questions related to Content and its viewership. Further to this the data is then graphically represented using the ggplot2 package which has lots of functionality that helps in understanding the data and what we can interpret form this. Usage of ggplot2, dplyr package seemed a little difficult so it took a bit of my time in understanding what those packages do and how well it to could be presented.

Reproducibility is a key nature for any analysis. The Project Template plays a key role in making the analysis reproducible by looping back to any step and whenever required. The analysis of the report is not completely reproducible as some manual interventions are required to change the variable names, binding up of the datasets as a single data frame when there appears a new datafile. Although in an overall sense of using project template and CRISP DM methodology improved performance and helped in maintaining reproducibility of analysis.