Critical Reflection of CSC8631 Data Management and Exploratory Data Analysis

By Ambreen Masud

03 December, 2021

"Meditated. Got my morning coffee in hand. Ready for the workday. Feeling great. Right, sit down. Open R. My packages aren't loading. What's going on? There's a lock file. What does that mean? Stack overflow, I need your help please. Okay, found a resolution. It's not working. Sigh. Okay, google do your thing. [Two hours and many attempts later]. Nothing is working. Frustrated. Thankful for the practical today. Run to the USB. The demonstrators don't know what's going on. Panic. Breathe. Wait for Joe. He'll know. Joe said to change the location of the project. Okay, it seems to be working. Relief. Wait. Can't get my head around how to configurate Git Bash with this new location. Stress. Right, just start again. Copy and paste is your friend. [Reset]. Perfect, it's working. Oh gosh, it's 7pm. It's okay. Have dinner, then restart again. [After dinner and procrastination]. Open R again. Do some EDA. Starting to enjoy it. Great. Coding errors? No problem, thanks Google. Feeling more productive. Okay, let's add a commit. A fatal error in my configure file! Why? Feeling the stress now. I can't cope. Close the laptop. Tomorrow's a new day." – Diary excerpt of Ambreen Masud

Perhaps the excerpt above is an over-exaggeration. However, it seemed that there were many tears and joys during this project. From problems with fatal errors on Git Bash, changing university laptops in the midst of the project and R not corresponding between my H drive and C drive, packages not loading because of said issue, and errors in coding or perhaps not quite getting the intended outcome; it has been a bit of a rollercoaster of emotions. However, when I reflect and assess how much I have learnt in the process, I realise that I thoroughly enjoyed this project and the learning curve that came with it. The problems provided quick learning, especially considering I only had basic R knowledge to begin with, and there was a lot of room for creativity, which also added to the enjoyment. Furthermore, the extension allowed for some mental space in between other deadlines. Therefore, it was possible to develop my knowledge of the tools used throughout this project in the timeline given.

Version Control - At the beginning of the project, it seemed that version control was just another unnecessary tool to make life that little bit more difficult. Particularly when coding, the use of statistics, the use of version control and github, RMarkdown, Project Template and doing an iterative exploratory data analysis were all realtively new to me. However, I now realise how necessary version control is. It really is easy to get lost in the analysis and the creative process, that ones code can easily become a mess of words and numbers. Commenting on the code just really is not enough and as someone who is a big believer of collaborative working, I can really see the value of making commits to not only remind yourself as to where you are in your project but also, to support others to follow and join on that journey with you. Despite this, there were times when I was truly frustrated with Git Bash. Mostly when it brought up fatal errors that I still do not quite understand why they arose. Especially, when restarting my laptop seemed to remove said errors.

Project Template - Project Template is a fantastic package, which encourages organisation. The hard work was done for me by R and I simply needed to ensure that the right documents were in the right folders. It took quite a few sessions to move away from the typical manner of coding that I was doing before, which was simply coding everything in one file. However, by the end of it, I became quite comfortable with organising files based on what I was assessing and uploading pre-processing code into my munge folder, making it much easier to manage the project. Furthermore, Dplyr has become a tool that I will continue to use. It is fantastic for someone who is just getting comfortable with coding. Sometimes, it can be a little arduous waiting for everything to load in Project Template. However, it is still much quicker than loading all of the code manually.

RMarkdown - Rmarkdown, whilst being fairly intuitive to use, does not have the same user-friendly interface that I am used to with Microsoft Word. I understand why it is appealing to use as it does make it much easier to input code and simply output the results, without displaying the code itself. However, it can bring up errors when knitting the document together, when the code itself has no errors, and is very slow in knitting large documents together. Although, I would be interested in trying a HTML output next time instead of a PDF output, just to see how different that is.

CRISP-DM - CRISP-DM is a good tool for planning and organising data analysis and by following the model, it also allows for someone else to follow the same process, thus allowing for reproducibility. Furthermore, it is natural to go back and forth between business understanding and data understanding, and data understanding and data preparation. However, for this project, there were oftentimes when I was in the deployment stage, or so I thought, and I noticed a small error so would need to go back to the data preparation stage to make some edits. This suggests that there is also a link between deployment and data preparation, something that the CRISP-DM process does not consider. It would be interesting to incorporate the modeling and evaluation stage in more depth, although I presume that will come with the evolution of my learning during this Masters.

Moving Forward - Whilst this project has been an emotional journey, the results of it and the learning has been gratifying. Two of the biggest lessons I have learnt from this project are:

1. To persist when I have a problem, or seek another way to do things if I have tried multiple times
2. To speak to my peers and any other person that may be able to support me with technology or coding concerns. If one person does not have an answer, then someone else may be able to assist, and this is also part of the learning process.

Following on from this project, I have already set myself a mini project over the Christmas holidays where I will use the video statistics data (which I did not quite get around to in this project), as I would like to continue developing the tools that I have developed in this project, whilst also developing my statistical knowledge. I would also like to improve my knowledge of version control, so will go back over the carpentry workshop material from the first two practical sessions of this module. This is to develop both personally and professionally, as it would be a great tool to have in my toolbelt.