CSC8631 – Data Management and Analysis - Assignment

**Summary**

Prior to starting this project all the course materials were reviewed, and relevant notes written. It was my opinion that having an overview would put me in a better position for starting the assignment. However, this just overloaded me with information and wasted time. To start the project the development environment was set up as per the assignment outlines, particularly:

* R Studio
* R Markdown
* Git at the bash prompt
* GitHub
* ProjectTemplate

This was reasonably straight forward, the VLE dataset was then imported into the data using the munge steps of ProjectTemplate, this was my first introduction to R. The CSV files were copied to the data directory and the munge steps set up. This also involved loading libraries from the Tidyverse and deal with mismatched casts and so forth.

Given the lack of direction in the assignment I began and put a significant amount of effort into reviewing the data and writing a report attempting to find a hypothesis to work with. This would have put me more onto familiar ground of having a requirement. As part of this work investigation into the data types and relationships between tables was carried out which was valuable later. However, this period in the project was very messy and caused significant rework later in the project.

After I failed to find a suitable hypothesis and could not particularly step to iteration 2 of the project, I revisited the CRISP-DM notes and this resulted in a complete rewrite of the report, concentrating on data exploration and preparation. The decision to essentially start again on the report was taken seven days before the deadline and created a lot of stress and impact on other units on the course. However, the new structure following a basic CRISP-DM outline was easier to write and the initial report I had written was rehashed into iteration 1 of the final report. This also resulted in a significant task to restructure and comment the R code which has been written to that point. I felt significant time was wasted on the initial report and associated visualisations, but I did learn R.

Iteration 2 of the assignment was much more enjoyable, particularly as I felt I had the skills to complete the task to hand. I transformed the data I had imported and explored into what I thought were logical data sets and began to visualise them. However, due to poor approach originally, I had missed any kind of data quality work so I reworked and re-applied a minimum level of type and missing data checking. I also found that a couple of the datasets I had originally disregarded (in fact not mentioned) were of use after I transformed the data.

I then began to visualise and investigate my transformed datasets and quickly realised that it would be two large a task to complete them all in either a 20-page report, or the timescales to hand. I decided to concentrate on investigating the first point of interest that lept out at me, which I did in the data quality issues regarding steps as outlined in my report. It was during this work that I realised that my failure to consider the filenames and a cohort idea earlier in the project had had a significant impact on the work.

Shambolic.

**Assumptions**

I made the mistake in the first instance of assuming that my usual ways of working, i.e. reading and making notes would stand me well. In this instance it was a time sink which left me confused by the time I started the assignment. Given my already limited time and not feeling that I had the grounding to complete the work I then made assumptions about data I could miss out, particularly handling the file name numbering and the archetype and team role datasets. Which then needed to be revisited anyway.

I assumed that I could structure the data into something close to a relational database that I am used to working with previously. I think I did this but I’m not sure that was the objective.

**Reflection**

This unit has been horrific to complete, and I am not confident that I have passed. I have completed multiple rewrites of the report. However, I am confident I could do similar work in a professional setting now, and have had plenty practice with R, Apply, GGPlot2 and DPlyr. However, I am worried I have committed a lot of time away from my next unit, to still not pass this one.

From a teaching point of view, I found that I didn’t have the pre-requisite knowledge for the work. Resources provided marked “RECAP” about R and Stats which I’ve never done before instantly put me on the back foot. Even at the end of this work I’ve no idea what sort of analysis I should have been doing, only graphical? or what the difference between a continuous and discrete variable is.

I found my delivery and management of the time at each stage difficult, it turns out that I had a belief that data science is about the production of graphs and some reasonably obvious findings, but it is more about the management of the data and the narrative of insight. I feel that the “learn as you go” approach to this unit has resulted in a piece of work that I’m not particularly happy with and would do very differently now that I’m at the end. I also spent too long worrying about visualisations and trying to graph everything I possibly could.

I spent a lot of time on this project trying not to use the video view data as a continuous data source as it had been pointed out in the introduction to the assignment that it had been done before. In retrospect if I wanted to do continuous variable visualisation I should have gone for the obvious choice. Instead, I’ve just produced a lot of categorical variable bar plots which isn’t what I was after.

I have given serious consideration as to whether data science is for me over the course of this work.