CSC8631: Data Management and Exploratory Data Analysis

Reflective Learning

Description

For this project, I was performing some data management and exploratory data analysis of a datset. This involved building the pipeline and using best practice to ensure reproducibility, version control and literate programming.

This was the first time I have used any of the languages or practices before and was not quite prepared for the amount of time needed to commit to it. This is impacted by the fact that I am an apprentice with a relatively new role, as manager of a team of data scientists, which is currently taking a lot of effort.

Feelings

My initial feelings were very positive – largely enthusiasm. I have found the academic and theoretical side of the programme very interesting so was excited to actually do some meaningful work.

Due to time management issues, I found the best way to manage this workload was to bulk the time, i.e. blast away for several hours and then come back in a few days. However, this in practice was very difficult to manage. I found the tasks of exploring data interesting, but exploring ggplot at 11pm until 4am most weekends was challenging.

This was heightened coming up to the deadline as the some debugging required to export R Markdown to PDF. I have heard this is a common thing, especially the first time you use it.

I felt under pressure and having to prioritise results over thoughtful exploration. I originally had some ambitious plans to explore the data alongside external open-source datasets but in the end, I did not have the time.

Evaluation and Analysis

I was misguided in thinking that it would be feasible to manage my time the way I did. In retrospect, I believe ‘little and often’ is a good rule here. This is especially the case when you are trying to debug some code (overflow tables in Markdown PDF) at 6am after having stayed up all night.

I have, in my opinion, underachieved for this task although I have learned an enormous amount. I am happy with the work that I have done but believe with a little more time could have produced a better analysis with a cleaner presentation.

One aspect I am happy with is the version control and how beneficial that was. I did have to roll back once or twice after messing something up. It really is a good idea! I can see on collaborative projects that this is essential.

Overall this was a great experience but I have some lessons learned.

Conclusion

I believe I could have reached out to my peers a little more or even to the teaching staff. For me, the biggest pressure was my day job (50-60 hours per week at the moment). Rushing your way through EDA is not the way to do it.

I also could have spent a bit more time reading up on GGPlot beforehand to understand some of the requirements around class types. I lost quite a lot of time becoming frustrated with time series plots that did not work.

Action Plan

There are several takeaways, and things to improve for similar kind of work, which are:

* When working to a deadline, try to do little and often rather than a lot in a shorter period - the author spent a lot of time debugging code into the **wee hours**;
* Formatting issues in R Markdown can be very time consuming - indeed there are imperfections in this report – make sure I am happy with the setup before creating a document;
* Read up on plotting types and statistical methods
* Do not use spaces in directory names; and,
* Keep doing it – try some Kaggle competitions it’s a lot of fun!