Introduction

In October 2025, I began the third module of my course, Deciphering Big Data. The module focused on breaking down the concepts, theories, and mechanisms used to appropriately handle big data. This included various data collection techniques, the data management pipeline, database modelling and construction and the normalisation of data, and considerations which must be made in relation to compliance and adherence to the legal parameters surrounding data.

This personal reflection will delve into how I have developed my technical and professional skills over the course of the module to acquire and handle big datasets and store them effectively, and ethically.

My ePorfolio containing all relevant exercises referenced below is linked here: https://liawilliams.github.io/deciphering-big-data.html.

Development of Data Collection Skills

My favourite activity of the module was the web scraping task during unit 3. This was a data collection method which previously intimidated me, as I perceived it as an advanced technical skill. Using the Guardian Jobs page as my site of choice and the BeautifulSoup method, I inspected its HTML structure to correctly identify the tags and classes containing the data-related job postings I wanted to scrape.

Although web scraping was fun, it was not without its challenges and frustrations. My biggest challenge stemmed from my inexperience with HTML prior to this activity, and I had to spend some time to understand its structure to identify the correct selectors to include in my code. I was able to overcome this challenge and became confident enough to include other details which should be scraped, such as the company and location of the job.

Overall, I really enjoyed this activity as it was rewarding to understand how data is extracted from web pages. Possessing the ability to scrape data from the web unlocks many possibilities for data analysis and research. I look forward to applying these skills in future data science and analysis projects within modules to come.

Data Cleaning and Storage

I completed the data cleaning exercise for units 4 and 5, using Python and its various libraries to clean UNICEF's Child Labour questionnaire dataset. I followed along with the tutorial from Kazil and Jarmul's (2016) 'Data Wrangling with Python' textbook. I replaced acronym column headers with the equivalent full questions, and detected duplicates using NumPy and the fuzzywuzzy libraries.

My main challenge came from my limited prior knowledge of the techniques that I was instructed to use, such as zipping data to replace the headers and using the fuzzywuzzy library to detect similarities between strings. I subsequently acquired skills in SQL and developed relational databases to store data adequately – this was completely unfamiliar to me prior to this module. It also created a 'golden thread'

between acquiring raw data, cleaning its contents, and storing it effectively before further processing. Emotionally, practising the application of my newfound knowledge in this area on real tasks and examples improved my confidence and self-trust, as the ‘learning-by-doing’ method is evidenced to do (Reese, 2011).

Group Work and Collaboration Experience

Unit 6 introduced the first opportunity to work in a group since starting my course. As I have professional experience as a Project Manager, I volunteered to lead my group and proposed the initial idea of designing a database for a B2B retail organisation. I was comfortable setting up weekly update meetings to align on progress and ensure we were on track to meet the deadline before setting up a group chat to communicate in between calls. I also ensured that every member had tasks and report sections which they were responsible for and understood their roles.

I knew that one of the key benefits of group work is the opportunity to share ideas with others and perspectives which I would not otherwise think of, such as the relationships that existed between entities. However, I do experience emotions of slight worry when participating in group work, especially because it involves relinquishing some control. Through frequent communication, we were able to submit the proposal on time, and to a distinction-level standard. This built my trust in others, both in academic and professional settings, and I feel more comfortable with the idea of working in a team in future modules.

I also participated in the seminars hosted by the module tutor; however, it was difficult to join all of them due to existing work commitments. Seminar 4 stands out to me for this reason – I was able to collaborate with my peers and exchange ideas on the considerations made when building a database for a fictional organisation, Dreamhome Property Management. I participated in a discussion on the database development process, and which stakeholders would be involved. Initially I thought that high-level ranking executives such as the CEO would be too senior a stakeholder to consider, however feedback from my peers and tutors aided the realisation that they are individuals key to the success of the database creation.

I usually feel anxious when exchanging ideas in a group. However, I felt encouraged to exchange ideas with my peers and tutor in this seminar as I could receive feedback, as well as hold constructive discussions on the reasoning behind my ideas and improve my critical thinking abilities in relation to databases. These are known benefits of collaborating in a group, in addition to developing learning communities and improving motivation to complete assignments (Laal and Ghodsi, 2012).

Impact on Academic and Professional Development

I previously believed that much of the effort would stem from statistical analysis and data visualisation. Now I believe that data cleaning is the most significant, and arguably, important part to ensure that any subsequent work (e.g. visualisation, machine learning model development etc) is accurate.

My knowledge of the wider architecture of data, including different database models, types of databases, and normalisation, was limited prior to this module. Terms such as SQL and database management systems (DBMS) were abstract concepts in my mind, and I had very little understanding of what data normalisation was. Activities completed across the module, including the normalisation and data build task in unit 7, plus the executive summary submitted during unit 11, has turned those abstract concepts into personal areas of strength. I also now consider how data is extracted, managed, stored, and processed rather than solely focusing on data analysis and subsequent model building.

After this module, my focus will be to deepen my capabilities in cleaning complex, messy data using Python, become proficient executing queries in SQL, and develop a strong understanding of building non-relational databases using MongoDB, to make me a well-rounded data professional.

References

Kazil, J. and Jarmul, K. (2016) *Data wrangling with Python: tips and tools to make your life easier*. Sebastopol: O'Reilly

Laal, M. and Ghodsi, S.M. (2012) 'Benefits of collaborative learning', *Procedia - Social and Behavioral Sciences*, 31, pp. 486–490. Available at: https://doi.org/10.1016/j.sbspro.2011.12.091.

Reese, H.W. (2011) 'The learning-by-doing principle', *Behavioral Development Bulletin*, 17(1), pp. 1–19. Available at: https://doi.org/10.1037/h0100597.