Reflection for Assessment 0

Harry Clarke October 2025

I aimed to find a dataset with code in the healthcare sector and found an analysis of the scikit-learn diabetes dataset on the XGBoosting website [1]. Unfortunately I was unable to achieve much communication with my group member so I had to do the exploration on my own. There was a silver lining to this - I have very limited experience with R, Python and Github (and data science generally!) so it gave me a great opportunity to gain some skills with each of these. I also found some great resources for resolving questions about coding and understanding data science jargon.

Due to the complications with communication, I had to get started on the project later than I would have liked. Then, as I was working alone with limited time, I wasn't able to get the project to the standard I would have wanted. I learnt a lot about exploring datasets, but I would have liked to have more experience with group work in a data context.

Initially I struggled to find a good source, and I knew I wanted to work with data around healthcare / science. I got some help from Daniel, who showed me the XGBoost datasets with code. The XGBoost website [1] is very useful for machine learning (XGBoosting / eXtreme gradient boosting) focused code used on a range of datasets. It is also useful as a resource for finding lots of 'nice' datasets (nice enough to be used for examples!). The example I selected was using the scikit diabetes dataset, a synthetic dataset with predictor and target variables, designed to be convenient for regression on many variables. Because it was designed for this, all of the data were numerical, with well labelled predictors and no missing datapoints.

Since I don't yet have the background for machine learning / XGBoosting, I decided to move away from the code that my source used and instead take an opportunity to build up a little bit of familiarity with the pandas module in python. I have a little bit of experience with arrays in python and the pandas DataFrame is close enough that I felt I had a little bit of intuition for how to work with it. The pandas website guidance was useful so I'll definitely use it more in future.

Until now I haven't had time to conduct an R analysis. I'm submitting now but will be starting the R analysis after this reflection, which I hope to upload to the github page soon.

References

[1] XGBoost for the diabetes dataset, XGBoosting. Available at: `https://xgboosting.com/xgboost-for-the-diabetes-dataset/` (Accessed: 07 October 2025).