CMP3749M Big Data Assessment Item 1

# Task 1 – PySpark Analysis of Nuclear Plants dataset

## Part 1

Firstly, the required libraries and modules were imported, and a spark session was created. A function was created named ‘clean\_data’ which takes in a data frame for cleaning; checking for duplicate rows and missing/null values, dropping the row that they occur on. The dataset is read into a spark data frame and then cleaned using the function previously mentioned, it was found that there were no duplicate or missing/null values in the dataset, with the number of samples being 996.

Incomplete datasets can have negative effects on the reliability of ML predictions. Dropping rows with missing/null or duplicated values is only one way of handling this task, other ways such as imputing data, or data prediction using machine learning (ML) could have been used in the analysis of the dataset. Imputation refers to filling missing values with substituted data (Li et al., 2015), with simple strategies such as replacing the values with the, mean, median or mode of that column/row, although this method is easy to implement it can introduce bias into the applications of the data (Jäger et al., 2021). More sophisticated ML methods such as k-nearest neighbours can be used for imputation, which can predict qualitative, and quantitative data by changing the parameter of the distance metric but can be slow as it must go through the whole dataset (Batista and Monard, 2003), which can be a huge drawback when dealing with billions of rows and columns within big datasets; due to the computational cost of processing the entire dataset.

## Part 2

## Part 3

# Task 2 – MapReduce for Marge Travel dataset

## Part 1

## Part 2

## Part 3

# Task 3 – Big Data Tools and Technology Appraisal

# References

Li, P., Stuart, E.A. and Allison, D.B. (2015) Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA,* 314(18) 1966-1967. Available from <https://doi.org/10.1001/jama.2015.15281> [accessed 06/01/2024].

Jäger, S., Allhorn, A. and Bießmann, F. (2021) *A Benchmark for Data Imputation Methods.* Frontiers Media SA. Available from <https://doi.org/10.3389/fdata.2021.693674> [accessed 06/01/2024].

Batista, G.E.A.P.A. and Monard, M.C. (2003) An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence,* 17(5-6) 519-533. Available from <https://doi.org/10.1080/713827181> [accessed 06/01/2024].