**Data Science Post Block Assignment 3: Task B**

Difedile Rasenyalo [28294882] & Isabel de Waal [20805055]

1. **Introduction** 
   1. **Background**

Many of the recent studies have focused on the imputation of categorical variables

* 1. **Objectives**

**Task A:** The objective was to evaluate the effectiveness of a baseline imputation versus that of a Naïve Bayes imputation on the performance of classification models. For this study we also decided to look at the influence of the proportion of missing values on the effectiveness of the imputation method.

1. **Methodology**

In this study, we leveraged R Studio and Python as the primary tools for imputation, modelling and evaluation. The methodology involved several key steps:

**Data Exploration:** The first phase of the analysis involved exploring the two provided datasets to gain a comprehensive understanding of their structure, variables, and content. This exploration aimed to identify any potential data inconsistencies, missing values, or outliers that could affect the quality of the analysis. This was done primarily by calculating the count, cardinality and % missing values of the dataset.

**Data Pre-processing:** Following the data exploration phase, the datasets underwent thorough cleaning and pre-processing. This step involved handling missing values,.

**Modelling:** K-Nearest Neighbours

**Visualization:** We

1. **Results and Discussion**

After pre-processing, the optimal number of clusters were to be selected for the K-means clustering.

According to [*A comparison of* imputation methods using machine learning models ] the type of dataset is an important factor when deciding on a imputation method. They found that k-NN worked better for some datasets than others. Conversely [Comparison of Performance of Data Imputation Methods for Numeric Dataset] found that the type of dataset did not have an influence on imputation performance. They did however find that the k-NN imputation generally outperformed the mean imputation. These and other studies [ref numbers], have found that the proportion of missing data also does not influence the imputation performance. This is in contrast to what we found. This could be because…

1. **Conclusion**

In conclusion, the analysis of the

1. **References**