**CCT College Dublin**

**Assessment Cover Page**

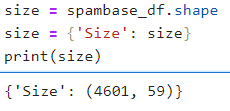
|  |  |
| --- | --- |
| **Module Title:** | Data Preparation |
| **Assessment Title:** | CA1 – Data Preparation |
| **Lecturer Name:** | David McQuaid |
| **Student Full Name:** | Angelo Luis Carrinho |
| **Student Number:** | Sba24084 |
| **Assessment Due Date:** | 03/11/2024 |
| **Date of Submission:** | 24/10/2024 |

**Declaration**

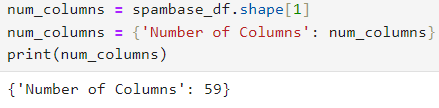
|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Characterisation of the data set:**

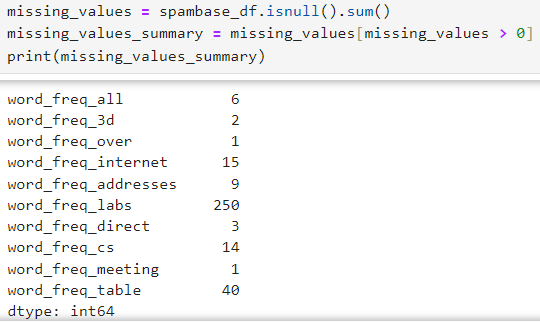
**Size**



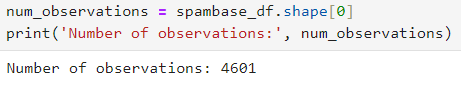
**Number of attributes;**



**has/does not have missing values**



**Number of observations**



**What these Characterisation mean in the context of this data**

Dataset have 4601 obervations and 59 variables, Dataset is of considerable large size enough to carry transactional information.

341 missing values in total among attributes indicate that some fields of the data points are blank, and hence this invariably might lead to great variation on analytics or predictions made over it.

Before performing further analysis, it is very essential to handle the missing values as they may have otherwise affected our final results.

* Application of Data preparation/evaluation methods
* (Cleaning, renaming, etc)
* EDA visualizations (plural), including a clear and concise explanation of your rationale for what you are doing with the data and why you are doing it in the context of this data.
* Use PCA to establish the minimum number of features needed for retaining 99.5% variance in the data and then implement PCA to dimensionally reduce the data to the number of features that you have discovered. Include a clear and concise explanation of your rationale for what you are doing with the data and why you are doing it in the context of this data.
* Explain **in your own words** what the “Curse of Dimensionality ” is and how it may affect your analysis going forward in the context of this problem.
* Testing your dimensionally reduced dataset practically and discuss your results in the context of this data.

Conclusions, Findings of data set and references (HARVARD style).