**CCT College Dublin**

**Assessment Cover Page**

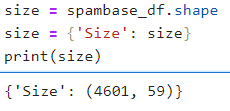
|  |  |
| --- | --- |
| **Module Title:** | Data Preparation |
| **Assessment Title:** | CA1 – Data Preparation |
| **Lecturer Name:** | David McQuaid |
| **Student Full Name:** | Angelo Luis Carrinho |
| **Student Number:** | Sba24084 |
| **Assessment Due Date:** | 03/11/2024 |
| **Date of Submission:** | 27/10/2024 |

**Declaration**

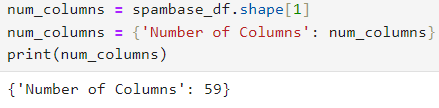
|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and 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:**

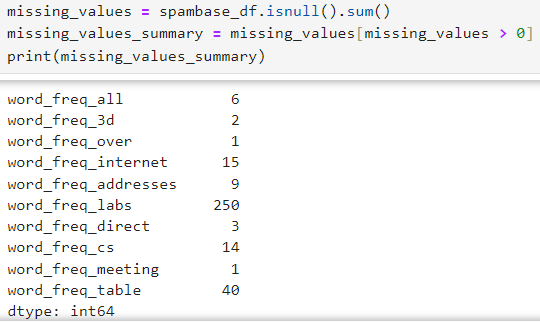
**Size**



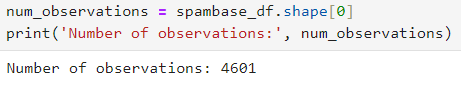
**Number of attributes;**



**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.

**Data Preparation**

**Data filter:** Separate spam messages from non-spam messages.

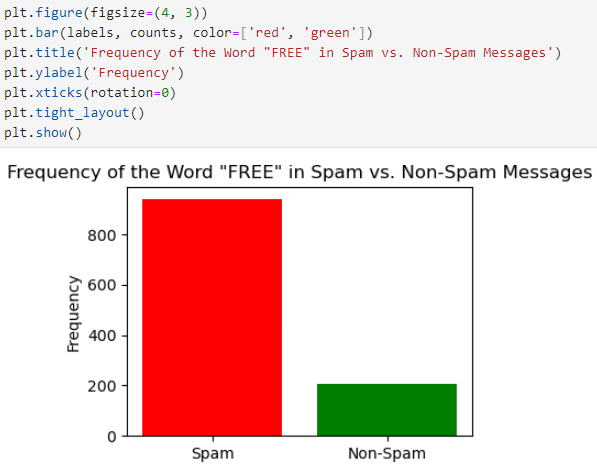


**Counting Frequencies:** Calculating How Often “FREE” Appears in Spam Messages



**Visualizations**

Frequency of the Word "FREE" in Spam vs. Non-Spam Messages



**Rationale:**

This visualization allows us to directly compare how often "FREE" appears in spam messages versus non-spam messages.

It provides information on whether this word is a common indicator of spam.

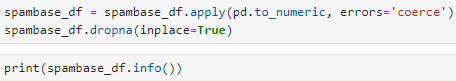
Percentage of Messages Containing the Word "FREE"



**Justification:**

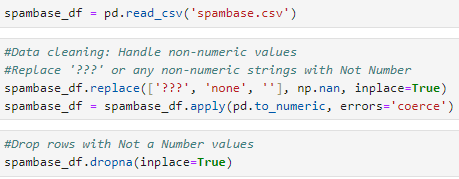
This graph provides information about the frequency distribution of the word “FREE” in spam messages. It shows the trend, helping us understand what percentage frequency the word "FREE" appears in spam messages.

* 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.



**Explanation**

**Loading the Data and Cleaning:** With this, i first load the dataset & do some cleaning by replacing nonumeric values as well as dropping missing entries.

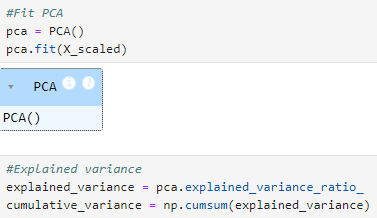


**Standardization :**

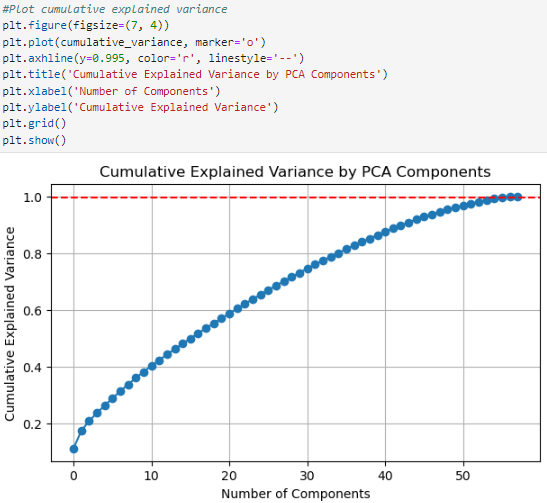
(Scale the features to get a better PCA performance)



Fit PCA to The data in X and calculate explained variance

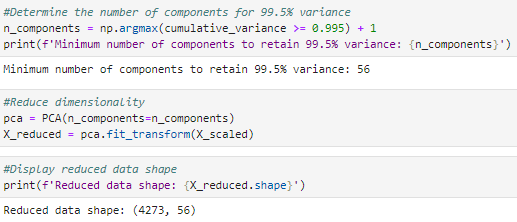


Plot of cumulative explained variance to determine the number of components needed:



**Dimensionality Reduction:** Reduces the dataset into number of components below a certain level as defined keeping 99.5% variance, in order to lessen highly correlated predictors / features from data set which are used for prediction model building.

In this way, we will be able to truncate the data set so that if there is any chance of overfitting it may help and at same time does not compromise with important aspects for building an analysis or model.



* 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.

The curse of dimensionality refers back to the demanding situations that stand up while coping with excessive-dimensional statistics.

As the number of functions (or dimensions) will increase, the spatial quantity grows exponentially and facts factors come to be greater sparse.

This sparsity can result in issues which includes overfitting (in which the version learns the noise as opposed to the underlying patterns) and problems in visualizing and deciphering the data.

In the context of unsolicited mail evaluation, because of this if there are too many features, the version may also battle to generalize properly, that may cause terrible overall performance on unseen records.

Reducing dimensionality therefore simplifies the information at the same time as preserving important information, thereby improving the accuracy and interpretability of the model.

* Testing your dimensionally reduced dataset practically and discuss your results in the context of this data.

**Only numeric columns are selected for PCA, as PCA requires numerical input.**

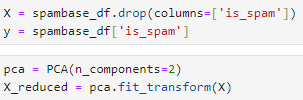
Impute Missing Values with 0:



**Perform PCA**

PCA is done to reduce the data set to two important additives.

Creating a DataFrame with Core Components



**A new DataFrame pca\_df is created containing the 2 fundamental components**



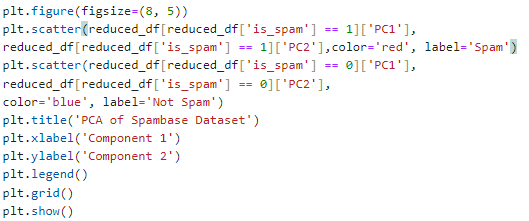
**The target variable is\_spam is passed to the PCA DataFrame to color the graph**

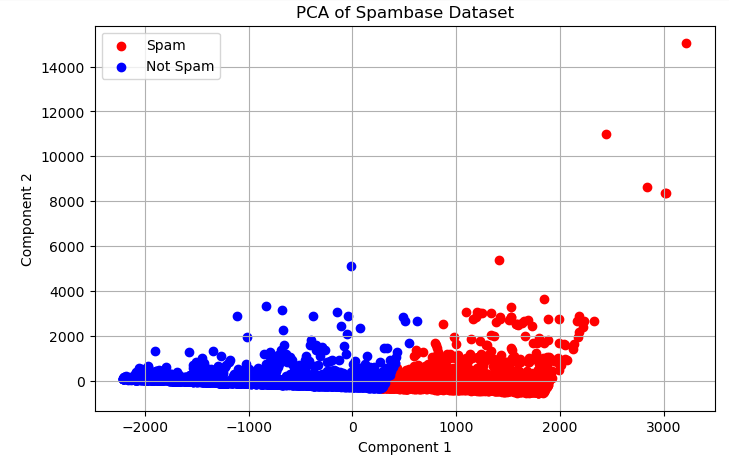
The target variable is\_spam is delivered to the PCA DataFrame for coloring the plot



**A scatterplot is created using Plotly to visualize the main components, colored via the is\_spam variable**

This visualization allows us to understand how information points are distributed within the reduced dimensional, space and whether spam mail and unsolicited emails are separable based on available components.





**References**

(Foundation, 2024)

(https://www.markdownguide.org, 2024)