## **Data Wrangling in Capstone Project Data:**

The Spambase data set was acquired from UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/spambase>) to use in my capstone project.

### Data Exploration

The Spambase dataset contains:

* Number of Instances: 4601
* Number of attributes: 58
* Number of missing data points: None
* The last column of 'spambase.data' named ‘spam’ denotes whether the email was considered spam (1) or not spam (0).
* Most of the attributes indicate whether a particular word or character was frequently occuring in the email. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. The definitions of the attributes are described in table below:

**Table 1 : Attributes in Spambase Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Column Number** | **Attribute name** | **Attribute Type** | **Attribute Description** |
| 1 to 48 | word\_freq\_WORD | continuous real [0,100] | percentage of words in the email that match WORD |
| 49 to 54 | char\_freq\_CHAR | continuous real [0,100] | percentage of characters in the email that match CHAR |
| 55 | capital\_run\_length\_average | continuous real [0,100] | average length of uninterrupted sequences of capital letters |
| 56 | capital\_run\_length\_longest | continuous integer [1,...] | length of longest uninterrupted sequence of capital letters |
| 57 | capital\_run\_length\_total | continuous integer [1,...] | total number of capital letters in the email |
| 58 | spam | nominal {0,1} | denotes whether the email was considered spam (1) or not (0) |

Outcome or dependent variable will be ‘spam’ and all other attributes from column 1 to 57 will be independent variables. Below is the summary of spam variable:

### Data Wrangling Steps:

The data wrangling steps includes as described below. The input file ‘Spam names.csv’ contains the name of attributes separated by commas for spambase data set .

1. Change the name of the below attributes which have special characters in their name as below:
   1. char\_freq\_; to char\_freq\_semic
   2. char\_freq\_( to char\_freq\_openp
   3. char\_freq\_[ to char\_freq\_openb
   4. char\_freq\_! to char\_freq\_excl
   5. char\_freq\_$ to char\_freq\_dollar
   6. char\_freq\_# to char\_freq\_pound
2. Since the column names were not present in the data, the column names were added to the data after above step.
3. Find and remove any missing values. There were no missing values in this data set.