# **Exploratory Analysis of Spambase Dataset**

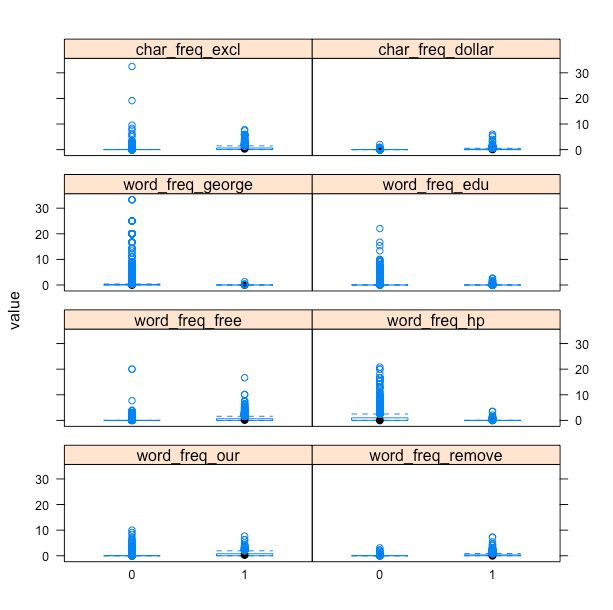
As per the source of dataset some of the words such as ‘george’, ‘hp’, ‘edu’ are associated with no-spam emails and other words such as ‘free’, ‘our’, ‘dollar’ seems to be associated with spam emails. In my exploratory analysis I decided to create visual summary statistics and frequency distribution plot in Spambase data set for some variables as listed below:

* word\_freq\_our
* word\_freq\_remove
* word\_freq\_free
* word\_freq\_hp
* word\_freq\_george
* word\_freq\_edu
* char\_freq\_excl
* char\_freq\_dollar
* capital\_run\_length\_average
* capital\_run\_length\_longest
* Spam (outcome variable with value 0 (no-spam) and 1 (spam email))

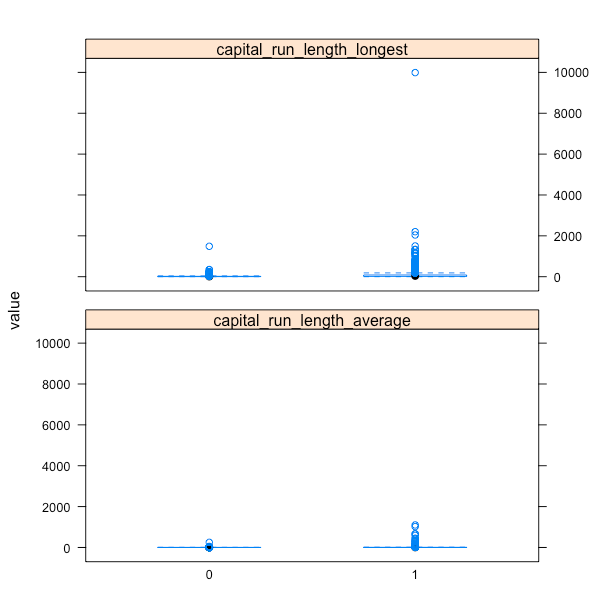
**Observe the correlations between different response variables with outcome variable and frequency distribution.**

1. Boxplots were used to get summary statistics for above variables. The box plots observation in **Figure 1** clearly shows which word has more frequency in spam(1) or no-spam(0) data. For example, word ‘george’ as expected has shown high frequency in no-spam emails and ‘char\_freq\_dollar’ has shown high frequency in spam emails.
2. **Figure 2** describes the correlations and frequency observation between ‘capital\_run\_lenth\_longest’ and ‘capital\_run\_length\_average’ and outcome variable. The visualization is showing the higher frequency of both of these variables in spam emails.
3. Another observation is that boxplots were collapsed because the Spambase data set values are highly concentrated at zero.
4. **Figure 3 and 4** are showing the density of selected response variables in the Spambase data which correlated with the boxplot observations

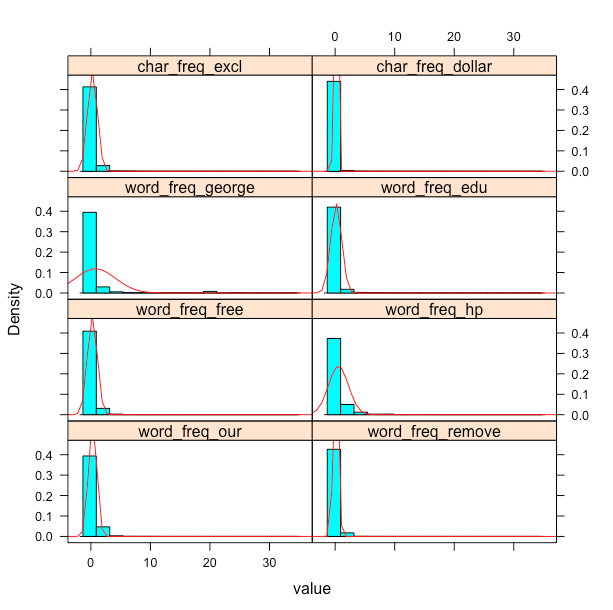
**Figure 1: Correlations and frequency observation between response variables and outcome variable**



**Figure 2: Correlations and frequency observation between ‘capital\_run\_lenth\_longest’ and ‘capital\_run\_length\_average’ and outcome variable**



**Figure 3. Density of selected response variables**



**Figure 4: Density of ‘capital\_run\_lenth\_longest’ and ‘capital\_run\_length\_average’**

