IS580 – Knowledge, Discovery and Mining

Assignment 1

Due Date: 21 March 2018

Strict Deadline: 24 March 2018

In this assignment, you are expected to characterize and understand a dataset by using descriptive statistics and visualization techniques as well as to handle the problems present in the data set such as missing value, noise and redundancy by using some preprocessing techniques. You will use the data files "assignment1.csv" , "assignment1_missing1.csv" and "assignment1_missing2.csv".

Deliverables:

1. "yourID_assignment1.Rmd" file which includes your comments on questions and R chunks related to the questions.
2. Data Files
    a. "yourID_ assignment1.csv" including transformed attributes
    b. "yourID_ assignment1_missing1.csv"
    c. "yourID_ assignment1_missing2.csv"

Data set description is as follows:

```
A1:   b, a.
A2:   continuous.
A3:   continuous.
A4:   u, y, l, t.
A5:   g, p, gg.
A6:   c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.
A7:   v, h, bb, j, n, z, dd, ff, o.
A8:   continuous.
A9:   t, f.
A10:  t, f.
A11:  continuous.
A12:  t, f.
A13:  g, p, s.
A14:  continuous.
A15:  continuous.
A16:  +,- (target variable)
```

A. Consider the data file "assignment1.csv".
    1. Summarize the data with appropriate descriptive statistics. Briefly interpret the results in text.
    2. Visualize the distributions of the numeric attributes. Comment whether they are Gaussian or not. Give your reasons and support them with appropriate visualization techniques.
    3. Determine the appropriate descriptive location measure for each of the numerical attributes.

4. Find out if there are outliers and make suggestions on how to deal with them.
5. Visualize the distributions of numeric attributes grouped by the target variable using density plots. Comment on the results.
6. Show if there is any correlation or association between the attributes.
7. Find and apply the best normalization methods for the attributes *A14* and *A15*. Save normalized versions as new attributes.
8. Which discretization methods are suitable for the attributes *A3* and *A11?* Why? Discretize the columns with the methods you choose. Save discretized versions as new attributes.
9. Save your final data file as "yourID_ assignment1.csv"

B. Consider the data file "assignment1_missing1.csv". This data file has 10% missing data in *A2* attribute.
   1. Listwise delete the data objects that are missing. Recalculate the descriptive statistics for all attributes. Comment on the results by comparing them to those calculated in A.
   2. Fill the missing values with Amelia (5 imputations are usually sufficient). Interpret the imputed data using its diagnostic tools and comment on it.
   3. Replace the missing values with the average of the corresponding imputations. Save your data file as "yourID_ assignment1_missing1.csv"

C. Consider the data file "assignment1_missing2.csv". This data file has 20% missing data in *A8* attribute.
   1. Listwise delete the data objects that are missing. Recalculate the descriptive statistics for all attributes. Comment on the results by comparing them to those calculated in A.
   2. Fill the missing values with appropriate location measure. Recalculate the descriptive statistics and visualize the distribution. Comment on the results by comparing them with the original stats and distribution you obtained in A.
   3. Fill the missing values with Amelia. Interpret the imputed data using its diagnostic tools and comment on it. Replace the missing values with the average of the corresponding imputations. Recalculate the descriptive statistics and visualize the distribution. Comment on the results by comparing them with the original stats and distribution you obtained in A and with those you obtained in C.2. Save your data file as "yourID_ assignment1_missing2.csv".

NOTE: To improve performance of the imputation you may need to transform some variables.

Important Details:

- Use ggplot2 package, qnorm() function and beanplot package while generating plots.