PESIT Department of Computer Science and Engineering

Course: Data Analysis
Semester: 2014 Spring (January – May)
Instructor: BNR (Dr. B. Narsing Rao)

Assignment: 02
Topic: Exploratory Data Analysis
Due by: Midnight on Tuesday, January 21, 2014
Method: Send zip archive (.zip, .rar, etc.) by email to bnrao@pes.edu
The name of the zip archive should be: DA-A02-your USN-your name
(USN must be upper case and your name should be in mixed case)
The zip archive should contain the following (see below for details):

1. PDF report including summary measures, your observations and explanations (named DA-A02-USN-Name.pdf)
2. Source file containing R functions (named DA-A02-USN-Name.R) used

Criminological Investigation

At the scene of a crime, the glass left can be used as evidence. For more information, including a case study, see:

http://www.evidencemagazine.com/index.php?option=com_content&task=view&id=386

The supplied file **glass.csv** has information about 214 samples of glass, the attributes of which are explained below:

1. RI: refractive index
2. Na: Sodium (unit of measurement: weight percent in corresponding oxide, as are attributes 3-9)
3. Mg: Magnesium
4. Al: Aluminum
5. Si: Silicon
6. K: Potassium
7. Ca: Calcium
8. Ba: Barium
9. Fe: Iron
10. Type of glass
    a. BWF : building_windows_float_processed
    b. BWNF : building_windows_non_float_processed
    c. VWF : vehicle_windows_float_processed
    d. CON : containers
    e. TBL : tableware
    f. HL : headlamps

Note that attributes 1 through 9 are numeric and the tenth attribute is nominal. The first step in analyzing the data is to do some exploratory data analysis, which is the purpose of this assignment.

You should produce a report (in PDF format) that summarizes the main characteristics of this dataset in the following format:

Section 1 : Summary Measures

For each numeric attribute, show the following in a table: Mean, Standard Deviation, type of Distribution (symmetric, skewed, bimodal etc. based on your observation of the histogram)

Section 2 : Box Plots

Show boxplots of each numeric attribute as a function of type. Do you find anything unusual about some of the plots (such as the one for Iron, i.e. Fe)? If so, how would you explain them?

Note: you can explore the data further by using the subset function in R to select rows from. For example, the following command selects from a data frame called **glass** only those rows for which the value of the Type attribute is equal to CON:

```
subset(glass,glass$Type %in% c("CON"))
```

After selection, you can do a histogram on various attributes to understand any unusual characteristics of the data.

Section 3: Prediction

The job of the forensic investigator is to determine the type of glass involved in the crime based on fragments found at a crime scene. In this case, which attributes, in your opinion are most useful and which the least useful in predicting the type? Explain your reasoning.