Data Mining
IT 270
Homework 4

Dr. Alexander Pelaez
alexander.pelaez@hofstra.edu

$19^{th} April, 2023$

Instructions

Please make sure if you use R you copy and paste it into Word using Courier Font (makes it easier to Read). For each of the problems that are looking for a response (not just a calculation), be sure to explain and interpret the results. If you are not sure. . . ASK PLEASE. Please start each question on a new page and clearly label that start of each problem (Maybe slightly larger font, bold face , underline. . .) anything that will help me find the problem you are working on.

Please remember if you submit a word document, you must copy your R Code and results (generally just copy what the output in the console is and it will include your code). Don't provide me R code with original data output please - it makes it too long. When you copy and paste it to word please make sure the R Code / Output is in courier font (10 pt).

1 Sustainability & Energy

Files Needed:

- Code: `HASD_doc.pdf`
- Data: `thads2013n.txt`

Notes: This is a pretty large dataset so cut it down as appropriate.

The Department of Energy has launched new initiatives around sustainability. The aim is to identify different groups of houses to identify the factors that should lead to reduced energy. However, the Director of the department is finding it difficult because they are using Excel , and so you have been asked to assist. Using the dataset `thads2013n.txt` which has a tremendous amount of information, your goal is to answer the following questions below. (The pdf file in the assignment sections has the definitions of the data set - YOU WILL NEED IT).

Your objective:

(a) Describe the dataset in terms of rows, columns, types of data and any outliers and missing data . . . the usual.

(b) Clean the data - describe what you did to clean the data

(c) Create a set number of groups of "housing" observations.

  (i) Determine which variables will you then cluster on. Remember we are focused mainly on the energy cost (UTILITY VARIABLE).

  (ii) Conduct cluster analyses using two agglomerative methods and a k-means cluster. How many clusters do you settle on using each method. Why? Provide the necessary charts.

  (iii) Define how you value or discern each cluster.

  (iv) Name your clusters.

(d) Create a new variable in your dataset which identifies which observation is within which cluster (k-means only), then provide measures for each cluster on three variables (UTILITY, TOTALSAL, ZINC2 - know what they are for your assessment, don't just give me the variable names).

(e) Conduct an analysis between each group, i.e. each cluster to determine if there is a statistically significant difference in the UTILITY, TOTALSAL and ZINC2)

Note: when you write this up, don't just give me the variables names, and numbers, provide a nice conclusion or set of statements for each of these items. Pretend its a report for someone.