PROG8430 – Data Analysis, Modeling and Algorithms

Assignment 3

Unsupervised Learning: K-Means Clustering

DUE BEFORE FEBRUARY 23, 2023; 10PM

1. Submission Guidelines

All assignments must be submitted via the econestoga course website before the due date into the assignment folder.
You may make multiple submissions, but only the most current submission will be graded.

SUBMISSIONS

In the Assignment 3 Folder submit:

1. Your *.Rmd file. This file must have all output already run *and* your comments and answers to the questions. If you do not include your output, I will *not* be running your code to generate it. I may, however run the code to verify the results.
2. The *.pdf or *.doc file that is produced from your code.

DO NOT PUT THE DOCUMENTS IN TO A ZIP FILE!

PLEASE NOTE: The marks on the assignment are generally awarded 50% for the actual R code and calculations and 50% for interpretation and demonstration that you understand what you have done.

EXAMPLES: The example output provided is simply to demonstrate what a typical submission might look like. You can use it as a basis, but your submission must be in your own words. Submissions that simply "cut and paste" my example commentary will be marked 0.

All variables in your code must abide by the naming convention [variable_name]_[intials]. For example, my variable for State would be State_DM.

THIS IS AN INDIVIDUAL ASSIGNMENT. UNAUTHORIZED COLLABORATION IS AN ACADEMIC OFFENSE AS IS DIRECT 'CUTTING AND PASTING' FROM OTHER SOURCES. Please see the Conestoga College Academic Integrity Policy for details.

Remember the discussion forums on eConestoga are a great place to ask questions.

2. Grading

This assignment will be marked out of 15 and is worth 5% of your total grade in the course.

Assignments submitted after 10pm will be reduced 20%. Assignments received after 8:00am the morning after the due date will receive a mark of 0%.

Assignments which do not follow the submission instructions may have marks deducted.

3. Data

Each student will be using one dataset:

PROG8430-23W-Assign03.txt

4. Background

The data summarizes the expenses of randomly selected participants. Each column represents the percentage of income devoted each expense category. The data dictionary is in the Appendix.

Your task is to use k-means clustering to segment these participants in to distinct clusters.

All of your charts, tables and graphs should be properly labelled.

5. Assignment Tasks

Nbr	Description	Marks
1	Data Transformation 1. Rename all variables with your initials appended (just as was done in assignment 1) 2. Standardize **all** of the variables using either of the two functions demonstrated in class. Describe why you chose the method you did.	2
2	Descriptive Data Analysis 1. Create graphical summaries of the data (as demonstrated in class: boxplots, histograms or density plots) and comment on any observations you make.	1
3	Clustering Using the K-Means procedure as demonstrated in class, create clusters with k=2,3,4,5,6,7. You will be using only two variables as your centroids (House and Food) 1. Create segmentation/cluster schemes for k=2,3,4,5,6,7. 2. Create the WSS plots as demonstrated in class and select a suitable k value based on the "elbow". [NOTE – Use the code that I provided to do this. Using other functions will yield different results.]	4

4	Evaluation of Clusters 1. Based on the "k" chosen above, create a scatter plot showing the clusters and colour-coded datapoints for each of "k-1", "k", "k+1". For example, if you think the "elbow" is at k=4 create the charts for k=3, k=4 and k=5. 2. Based on the WSS plot (3.2) and the charts (4.1) choose one set of clusters that best describes the data. 3. Create summary tables for the segmentation/clustering scheme (selected in step 4.2). 4. Create suitable descriptive names for each cluster. 5. Suggest possible uses for this clustering scheme.	 2 1 1 1 1
5	Professionalism and Clarity	2

APPENDIX ONE: DATA DICTIONARY

Name	Description
Food	Percentage of income spent on Food.
Enter	Percentage of income spent on Entertainment.
Edu	Percentage of income spent on Education.
Trans	Percentage of income spent on Transportation.
Work	Percentage of income spent on Work Related Expenses.
House	Percentage of income spent on Housing.
Oth	Percentage of income spent on Other Expenses.