PROG8430 – Data Analysis, Modeling and Algorithms

Assignment 5

Unsupervised Learning: K-Means Clustering

DUE BEFORE AUGUST 18, 2021; 10PM

1. Submission Guidelines

All assignments must be submitted via the econestoga course website before the due date in to the assignment folder.
You may make multiple submissions, but only the most current submission will be graded.
Late assignments will receive a penalty of 20%.

SUBMISSIONS
In the Assignment 4 Folder submit:

1. Your R Code
2. Your report in Word, following the template from previous lectures.

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

You may only use the 'R' packages discussed and demonstrated in class:

1. ggplot2
2. cluster
3. factoextra
4. dplyr

THIS IS AN INDIVIDUAL ASSIGNMENT. UNAUTHORIZED COLLABORATION IS AN ACADEMIC OFFENSE. Please see the Conestoga College Academic Integrity Policy for details.

2. Grading

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

3. Data

Each student will be using one dataset:

PROG8430_Clst_21S.Rdata

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 reviewers in to distinct clusters.

Your work should follow the format of the sample report used previously.

5. Assignment Tasks

Nbr	Description	Marks
1	Data Transformation 1. Standardize **all** of the variables using either of the two functions demonstrated in class. Describe why you chose the method you did.	1
2	Descriptive Data Analysis 1. Create graphical summaries of the data (as demonstrated in class: boxplots or histograms) 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. You will be using only two variables as your centroids (Hous and Entr) 1. Create segmentation/cluster schemes for k=2,3,4,5,6. 2. Create the WSS plots as demonstrated in class and select a suitable k value based on the “elbow”. [NOTE – It is easiest to create this in Excel or some other spreadsheet program]	2 2
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 2 1 1
5	Professionalism and Clarity	2

APPENDIX ONE: DATA DICTIONARY

Name	**Description**
Food	Percentage of income spent on Food.
Entr	Percentage of income spent on Entertainment.
Educ	Percentage of income spent on Education.
Trans	Percentage of income spent on Transportation.
Work	Percentage of income spent on Work Related Expenses.
Hous	Percentage of income spent on Housing.
Other	Percentage of income spent on Other Expenses.