

PROG8435 – Data Analysis, Modeling and Algorithms

Assignment 3

Unsupervised Learning: K-Means Clustering

DUE BEFORE MARCH 17, 2024; 11:59 PM

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 *.R file. This file must have all source code *and* your comments for the 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]_[initials]. 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.

NOTE: No packages other than those installed with the initial installation of R may be used in this assignment.

2. Grading

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

Assignments submitted after 11:59pm 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:

PROG8435-23W-Assign03.txt

4. Background

The data set refers to clients of a wholesale food distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

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 and Descriptive Analysis <ol style="list-style-type: none">1. Rename all variables with your initials appended (just as was done in assignment 1)2. Create graphical summaries of the data (as demonstrated in class: boxplots or histograms) and comment on any observations you make.3. Standardize all of the variables using either of the two functions demonstrated in class. Describe why you chose the method you did.	4
3	Clustering <p>Using the K-Means procedure as demonstrated in class, create clusters with k=2,3,4,5,6,7.</p> <p>You will be using only two variables as your centroids (Milk and Froz).</p> <ol style="list-style-type: none">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	<p>Evaluation of Clusters</p> <ol style="list-style-type: none"> 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=5 create the charts for k=4, k=5 and k=6. 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. 	<p>2</p> <p>1</p> <p>1</p> <p>1</p> <p>1</p>
5	Professionalism and Clarity	1

APPENDIX ONE: DATA DICTIONARY

Name	Description
Frsh	FRESH: annual spending (m.u.) on fresh products (Continuous);
Milk	MILK: annual spending (m.u.) on milk products (Continuous);
Groc	GROCERY: annual spending (m.u.)on grocery products (Continuous);
Froz	FROZEN: annual spending (m.u.)on frozen products (Continuous)
Clen	DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
Deli	DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);