SURV 616/686
Homework Assignment #05
65 points

The data are from the following repository of machine learning datasets:

http://archive.ics.uci.edu/ml/datasets/Wholesale+customers

Problem Statement

Your client is a wholesale distributor of food, beverages, and cleaning supplies. The firm is located in Portugal. The task is to identify groups of customers that may share common patterns of spending on goods from the distributor. This will aid the firm in planning its marketing. You will use K-means to identify these groups of customers.

Data

You are provided with following file: “Wholesale customers data.csv”. The data dictionary is given below.

Data Dictionary

Here is the description of all the variables:

Variable	Definition
`Fresh`	Annual spending (m.u.) on fresh products (Continuous)
`Milk`	Annual spending (m.u.) on milk products (Continuous)
`Grocery`	Annual spending (m.u.) on grocery products (Continuous)
`Frozen`	Annual spending (m.u.) on frozen products (Continuous)
`Detergents_Paper`	Annual spending (m.u.) on detergents and paper products (Continuous)
`Delicatessen`	Annual spending (m.u.) on delicatessen products (Continuous)
`Channel`	Channel - Horeca (Hotel/Restaurant/Cafe) or Retail channel (Nominal)
`Region`	Customer’s Region -- Lisbon, Oporto or Other (Nominal)

1a. [10 points] **Exploratory Analysis**. Read the data into R. Report on each variable using a summary, a figure, or a table as appropriate. You can ignore the channel (CHANNEL) and region (REGION) for this exercise.

1b. [5 points] **Feature Engineering.** What pre-processing steps are necessary before applying K-means? Are there any transformations of the data to consider for this problem? Explain your choices.

2. The first task will be to create groups using the variables FROZEN and FRESH only. For this task, ignore the other variables in the dataset.

2a. [5 points] **Hyperparameter Selection**. What is the K you choose for this problem? How do you justify that choice?

2b. [10 points] Graph the K-means clustering of the cases based upon the FROZEN and FRESH variables.

3. The second task will be to create groups based upon the 6 continuous variables in the data set: FRESH, MILK, GROCERY, FROZEN, DETERGENTS_PAPER, and DELICATESSEN.

3a. [5 points] **Hyperparameter Selection**. What is the K you choose for this problem? How do you justify that choice?

3b. [20 points] List the values of the 6 variables at each center. Write a one sentence description of each group. How is each unique?

4. [10 points] Create a two-way table showing how the groups from Question #2 and #3 intersect. Give a brief description of how the two sets of clusters relate