**Assignment #8: Clustering Using R**

|  |
| --- |
| **Submission Instructions**  **Due Date:** **Friday, April 20, 2018, before class**.   * Submit the following **five** files through Canvas>Assignments>To-Dos:  1. The completed, working **R script** that produced the analysis for the **15 cluster** scenario. 2. The **three output files**: “ClusteringOutput.txt” “ClusteringPlots.pdf” and “ClusterContents.csv” for the **15 cluster** scenario. 3. The completed **answer sheet** provided on the last page.  * If you do not follow the instructions, your assignment will be counted late.   + Late Assignment policy: Same as before.   **Evaluation**  Your submission will be graded based on the correctness of the completed answer sheet, with other files as supporting documents. |

**Before you start**

For this assignment, you’ll be working with the **Jeans.csv** file and the **Clustering.r** script (which we used in ICA #13). This file has data from 689 stores that sell four different types of jeans: leisure, fashion, stretch, and original. The marketing division of the company wants to identify groups of stores that sell a similar mix of product so that they can roll out promotions specific to those stores.

The data file contains the following fields:

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| **StoreID** | Store identification number |
| **Fashion** | The number of pairs of “fashion” style jeans sold last month |
| **Leisure** | The number of pairs of “leisure” style jeans sold last month |
| **Stretch** | The number of pairs of “stretch” style jeans sold last month |
| **Original** | The number of pairs of “original” style jeans sold last month |
| **TotalSold** | The total number of jeans sold last month |

**Guidelines**

1. You’ll need to modify the **Clustering.r** script from ICA #9 with the following information to perform the analysis:

* Set the input filename (INPUT\_FILENAME) to the store’s dataset (i.e., “Jeans.csv”).
* **Set the number of clusters to create (NUM\_CLUSTER) to 5**.
* Set the variable list (VAR\_LIST) to use the Fashion, Leisure, Stretch, and Original variables by changing it to the following:

VAR\_LIST <- c("Fashion","Leisure","Stretch","Original")

1. Once you finish modifying the script, you can set the working directory and run the script.
2. Based on your script output, answer Questions 1-5 in the answer sheet at the end of this document.
3. Now rerun the script, this time with **15 clusters**. Then answer Questions 6-11 in the answer sheet at the end of this document.

Answer Sheet for Assignment: Clustering Using R

Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Fill in the answersheet below based on the output from R/RStudio:*

|  |  |  |
| --- | --- | --- |
|  | **Question** | **Answer** |
| **5 clusters**  Based on your script output with 5 clusters, answer Questions 1-7 below. | | |
| 1 | Which cluster is the largest (write the number of the cluster)?  How many stores are in the largest cluster (i.e. what is the cluster size)? |  |
| 2 | Describe the average sales of cluster 1 for each type of jeans (compared to the overall population average across all stores)? (write one or two sentences) |  |
| 3 | In which of the 5 clusters of stores do fashion jeans sell the best, on average? |  |
| 4 | What is the range of withinss errors (i.e. within-cluster SSE) for the 5 clusters? | Lowest: \_\_\_\_\_\_\_\_\_\_  Highest: \_\_\_\_\_\_\_\_\_\_ |
| 5 | What is the *average* betweenss error (i.e. average between-cluster SSE) for all 5 clusters? |  |
| **15 clusters**  Now rerun the script, this time with 15 clusters. Then answer the following questions: | | |
| 6 | Describe the average sales of cluster 1 for each type of jeans (compared to the overall average across all stores)? (write one or two sentences) |  |
| 7 | In which of the 15 clusters of stores do fashion jeans sell the best, on average? |  |
| 8 | What is the range of withinss errors for the 15 clusters? | Lowest: \_\_\_\_\_\_\_\_\_\_  Highest: \_\_\_\_\_\_\_\_\_\_ |
| 9 | What is the *average* betweenss error for all 15 clusters? |  |
| **5 Clusters versus 15 Clusters** | | |
| 10 | Which scenario (5 clusters or 15 clusters) produces clusters with better cohesion? |  |
| 11 | Which scenario (5 clusters or 15 clusters) produces clusters with better separation? |  |