**Assignment #8: Clustering Using R**

(Due Monday, April 24, 2017 at 9:00 am)

**What to submit**

Submit the following 5 files through Blackboard before deadline.

* The completed, working R script that produced the analysis for the **20 cluster** scenario.
* The **three** output files “ClusteringOutput.txt” “ClusteringPlots.pdf” and “ClusterContents.csv” for the 20 cluster scenario.
* The completed answer sheet provided on the last page.

**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 #12). 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 their 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 #11 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 1-7 in the answer sheet at the end of this document.**
3. **Now rerun the script, this time with 20 clusters. Then answer Questions 8-14 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)? |  |
| 2 | How many stores are in the largest cluster (i.e. what is the cluster size)? |  |
| 3 | Describe the sales of cluster 1 for each type of jeans (compared to the overall population average across all stores)? (write one or two sentences) |  |
| 4 | Describe the sales of cluster 5 for each type of jeans (compared to the overall population average across all stores)? (write one or two sentences) |  |
| 5 | In which of the five clusters of stores do original jeans sell the best? |  |
| 6 | What is the range of withinss errors (i.e. within-cluster SSE) for the five clusters? | Lowest: \_\_\_\_\_\_\_\_\_\_Highest: \_\_\_\_\_\_\_\_\_\_ |
| 7 | What is the *average* betweenss error (i.e. average between-cluster SSE) for all five clusters? |  |
| **20 clusters** Now rerun the script, this time with 20 clusters. Then answer the following questions: |
| 8 | Describe the sales of cluster 15 for each type of jeans (compared to the overall average across all stores)? (write one or two sentences) |  |
| 9 | Describe the sales of cluster 20 for each type of jeans (compared to the overall average across all stores)? (write one or two sentences) |  |
| 10 | What is the range of withinss errors for the 20 clusters? | Lowest: \_\_\_\_\_\_\_\_\_\_Highest: \_\_\_\_\_\_\_\_\_\_ |
| 11 | What is the *average* betweenss error for all 20 clusters? |  |
| **5 Clusters versus 20 Clusters** |
| 12 | Which scenario (5 clusters or 20 clusters) produces clusters with better cohesion? |   |
| 13 | Which scenario (5 clusters or 20 clusters) produces clusters with better separation? |  |
|  14 | Besides cohesion and separation, what other advantage does the 5 cluster scenario have over the 20 cluster scenario? (write one or two sentences) |  |