**Assignment #3: Clustering Using Python**

|  |
| --- |
| **Submission Instructions**   * Submit the completed **answer sheet** provided on the last page. * If you do not follow the instructions, your assignment will be counted late. |

**Before you start**

For this exercise, you’ll be working with the **Jeans.csv** file. 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** |
| **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 |

**Guidelines**

1. Set k=**5** and run the clustering analysis. Based on your script output, answer Questions 1-5 in the answer sheet at the end of this document.
2. Next 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 Python

*Fill in the answersheet below based on the output from Python:*

|  |  |  |
| --- | --- | --- |
|  | **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)? | Cluster 1  Cluster size= 199 |
| 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) | The average sales for Leisure jeans in Cluster 1 are 0.004976182, which is higher than the overall population average of -1.86062E-16 for Leisure jeans. This suggests that the stores in Cluster 1 have a higher demand for Leisure jeans compared to the overall population.  On the other hand, the average sales for Fashion, Stretch, and original jeans in Cluster 1 are -0.008559494, -0.001512571, and -0.007102159, respectively. These values are lower than the overall population average for each of these types of jeans. This implies that stores in Cluster 1 have a relatively lower demand for Fashion, Stretch, and Original jeans compared to the overall population.  Average of Fashion leisure stretch original for cluster 1:  -0.008559494 0.004976182  -0.001512571 -0.007102159 |
| 3 | In which of the 5 clusters of stores do fashion jeans sell the best, on average? | Fashion jeans sell the best, on average, in Cluster 3 with an average of 0.004662995. |
| 4 | What is the withinss errors (i.e. within-cluster SSE) for the 5 clusters? | 1099.1389839909696 |
| 5 | What is the *Silhouette Score* for all 5 clusters? | 0.2786042590675735 |
| **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) | The average sales of cluster 1 for Fashion jeans are lower than the overall population average, indicating that the stores in cluster 1 sell fewer Fashion jeans compared to the overall population. On the other hand, the average sales for Leisure and Stretch jeans are higher in cluster 1, suggesting that stores in this cluster sell more of these types of jeans compared to the overall population. The average sales of Original jeans are also higher in cluster 1, indicating that stores in this cluster sell more Original jeans compared to the overall population.  Average of Fashion leisure stretch original for cluster 1:  -0.8074  -0.13527  0.41554  0.229649 |
| 7 | In which of the 15 clusters of stores do fashion jeans sell the best, on average? | fashion jeans sell the best in cluster 8 with an average sale of 2.018858083 |
| 8 | What is the withinss errors for the 15 clusters? | 542.7128421689519 |
| 9 | What is the *Silhouette Score* for all 15 clusters? | 0.24021491818975896 |
| **5 Clusters versus 15 Clusters** | | |
| 10 | Which scenario (5 clusters or 15 clusters) produces clusters with better cohesion? | The within-cluster SSE for 5 clusters is 1099.13898, while for 15 clusters, it is 306.77530. This means that the within-cluster SSE is lower for 15 clusters than for 5 clusters, indicating that the clusters in the 15-cluster scenario have better cohesion. |
| 11 | Which scenario (5 clusters or 15 clusters) produces better clusters that are denser and more separated? | Based on the within-cluster SSE and Silhouette Score, the 15-cluster scenario produced denser and more separated clusters. The lower within-cluster SSE indicates that the data points within each cluster are closer to each other, leading to denser clusters. Additionally, the higher Silhouette Score indicates a better separation between the clusters. Therefore, the 15-cluster scenario is likely to produce better denser and more separated clusters. |