**Clustering \_ practice**

A clothing chain stores sells four different types of workwear jeans and would like to analyze the sales across different stores. The **sales.csv** data set shows the number of pairs of each type sold at stores over a specific time period. Each row represents an individual store. There are in total 689 stores in this data set. Details for columns are as follow.

* **StoreID** - Idenfication number of the store
* **Fashion** – Number of pairs of fashion jeans sold at the store
* **Leisure** – Number of pairs of leisure jeans sold at the store
* **Stretch** – Number of pairs of stretch jeans sold at the store
* **Original** – Number of pairs of original jeans sold at the store
* **SalesTotal** – Total number of pairs of jeans sold (the sum of all four types sold)

Analyze the data and answer the following questions:

Perform k-means clustering using all of the variables **except StoreID and SalesTotal**.

1. Use Silhouette analysis to compare clustering formed with n =2 to n=10. What is the best option suggested by the Silhouette score plot?
2. Choose n=6, which cluster has the highest average SalesTotal? How many stores are in this cluster? What is the average SalesTotal of this cluster?

1. The company is considering closing those stores with low total sales. Which cluster has the lowest average SalesTotal? **Fill out the following table for this cluster.**

|  |  |
| --- | --- |
| The cluster with the lowest average total sales |  |
| Total number of stores |  |
| Average sales of Leisure |  |

1. The profit per unit by different types of workwear is given in the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type | Fashion | Leisure | Stretch | Original |
| Profit | 27.7 | 22.8 | 31.2 | 24.6 |

Which cluster has the highest average profit per store? What is the average profit accordingly? For this cluster, which type of workwear average sales is the higher than all other clusters?

1. Compare the clusters formed with 7 clusters and the one with 6 clusters. Which cluster(s) in n=6 will remain unchanged when we form 7 clusters instead? Which cluster(s) will have more than 15% of the data points being sent to another segment?