Lab Tasks

Area: Clustering (K-Means)

1. You can download the data from here:

<https://drive.google.com/open?id=1akEkN49ZD5Mzf16oGKKzfEQQ12-DQYJn>

|  |  |
| --- | --- |
| **Wholesale customers Data Set**  **Abstract**: The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 440 | **Area:** | Business |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 8 | **Date Donated** | 2014-03-31 |
| **Associated Tasks:** | Classification, Clustering | **Missing Values?** | N/A | **Number of Web Hits:** | 249357 |

**Source:**

Margarida G. M. S. Cardoso, margarida.cardoso **'@'** iscte.pt, ISCTE-IUL, Lisbon, Portugal

**Attribute Information:**

1) FRESH: annual spending (m.u.) on fresh products (Continuous);   
2) MILK: annual spending (m.u.) on milk products (Continuous);   
3) GROCERY: annual spending (m.u.)on grocery products (Continuous);   
4) FROZEN: annual spending (m.u.)on frozen products (Continuous)   
5) DETERGENTS\_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)   
6) DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);   
7) CHANNEL: customersâ€™ Channel - Horeca (Hotel/Restaurant/CafÃ©) or Retail channel (Nominal)   
8) REGION: customersâ€™ Region â€“ Lisnon, Oporto or Other (Nominal)

1. Upload the data into environment
2. Observe the data ( in R-through summary and str)
3. We’ll need to drop the Channel and Region variables.  These are two ID fields and are not useful in clustering. So drop it.
4. Set some seed value
5. Apply the k-mean on dataset, with k=5
6. Observe the #number samples in each cluster

1 2 3 4 5

23 10 104 223 80

1. Observe the center of cluster against each features, like

Fresh Milk Grocery Frozen Detergents\_Paper Delicassen

1 49296.087 4983.783 5590.304 8285.783 962.2609 2543.6957

2 21263.700 37443.300 46710.600 6287.200 21699.4000 8743.3000

3 21200.058 3886.423 5138.933 4119.856 1131.5192 1690.3365

4 6052.812 3266.314 4092.054 2459.682 1214.1300 990.6099

5 4738.762 11609.013 18198.775 1515.388 8003.7750 1603.8000

1. Interpret the cluster results
2. Measure homogeneity of each cluster (SSE)
3. Measure total SSE
4. Measure the heterogeneity of cluster
5. Elbow measure: run the algorithm 100 time for k=2 to 20.
6. Draw the plot and observe elbow point.