Paresh Patil

Unsupervised Learning: K-Means Clustering

**Description**

**Data Transformation:**

1. Standardize all the variables using either of the two functions demonstrated in class. Describe why you chose the method you did.

Food Entr Educ Tran Work

Min. :0.0760 Min. :0.01000 Min. :0.0100 Min. :0.01100 Min. :0.01400

1st Qu.:0.1230 1st Qu.:0.02200 1st Qu.:0.0975 1st Qu.:0.02300 1st Qu.:0.03875

Median :0.1500 Median :0.03300 Median :0.1650 Median :0.05250 Median :0.05850

Mean :0.1506 Mean :0.05561 Mean :0.1419 Mean :0.09909 Mean :0.05931

3rd Qu.:0.1780 3rd Qu.:0.09525 3rd Qu.:0.2013 3rd Qu.:0.17650 3rd Qu.:0.08100

Max. :0.2350 Max. :0.13500 Max. :0.2520 Max. :0.26500 Max. :0.10400

Hous Othr Foodnormn Entrnormn Educnormn

Min. :0.3110 Min. :0.00900 Min. :-2.01849 Min. :-1.1836 Min. :-1.7538

1st Qu.:0.4010 1st Qu.:0.02300 standardize 1st Qu.:-0.74639 1st Qu.:-0.8721 1st Qu.:-0.5900

Median :0.4830 Median :0.0320 Median :-0.01561 Median :-0.5867 Median : 0.3078

Mean :0.4549 Mean :0.03869 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000

3rd Qu.:0.5020 3rd Qu.:0.05150 3rd Qu.: 0.74224 3rd Qu.: 1.0289 3rd Qu.: 0.7899

Max. :0.5390 Max. :0.09100 Max. : 2.28499 Max. : 2.0605 Max. : 1.4649

Trannormn Worknormn Housnormn Othrnormn

Min. :-1.0610 Min. :-1.98951 Min. :-2.3608 Min. :-1.4529

1st Qu.:-0.9165 1st Qu.:-0.90273 1st Qu.:-0.8840 1st Qu.:-0.7677

Median :-0.5611 Median :-0.03549 Median : 0.4615 Median :-0.3272

Mean : 0.0000 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000

3rd Qu.: 0.9324 3rd Qu.: 0.95249 3rd Qu.: 0.7732 3rd Qu.: 0.6271

Max. : 1.9984 Max. : 1.96243 Max. : 1.3803 Max. : 2.5602

Let’s observe the data before standardization, data is disturbed, and this is due to the scale difference between the values in each column. To perform clustering, it we must standardize the data, and I am using z=x-mean/SD method, because if we standardize data with this method, each column will have means 0 and standard deviation 1 which will eliminate discrepancy caused by range.

**Descriptive Analysis:**

1. Create graphical summaries of the data (as demonstrated in class: boxplots or histograms) and comment on any observations you make.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

In above food histogram the left one is before standardization and right one is after standardization. You can observe the range is changing which depicts that it’s eliminating the scale differences. Similarly, in below work graph.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Box plot before standardization

Chart, box and whisker chart

Description automatically generated

Box plot after standardization

Chart, box and whisker chart

Description automatically generated

The boxplot before the standardization data is unscaled, the variance is high whereas box plot after standardization show the median near to 0 and data is scaled.

**Clustering:** Using the K-Means procedure as demonstrated in class, create clusters with k=3,4,5,6,7. You will be using only two variables as your centroids (Food and Work).

1. Create segmentation/cluster schemes for k=3,4,5,6,7.

* K-means clustering with 3 clusters of sizes 297, 146, 157

Cluster means:

Foodnormn Worknormn

1 -0.4643371 0.9433245

2 -0.4837000 -1.0204720

3 1.3282059 -0.8355317

* K-means clustering with 4 clusters of sizes 195, 146, 102, 157

Cluster means:

Foodnormn Worknormn

1 -0.005059243 0.9446097

2 -0.483699975 -1.0204720

3 -1.342368385 0.9408677

4 1.328205886 -0.8355317

* K-means clustering with 5 clusters of sizes 191, 103, 51, 106, 149

Cluster means:

Foodnormn Worknormn

1 -0.0001620176 0.9626065

2 -1.3357938364 0.9345858

3 -1.2855830408 -1.2740275

4 -0.0084585230 -0.8391385

5 1.3696580472 -0.8469537

* K-means clustering with 6 clusters of sizes 93, 105, 102, 100, 149, 51

Cluster means:

Foodnormn Worknormn

1 0.08945429 0.5908199

2 -0.07901957 1.2293360

3 -1.34236838 0.9408677

4 -0.01614933 -0.8882354

5 1.36965805 -0.8469537

6 -1.28558304 -1.2740275

* K-means clustering with 7 clusters of sizes 93, 102, 97, 74, 105, 80, 49

Cluster means:

Foodnormn Worknormn

1 0.08945429 0.5908199

2 -1.34236838 0.9408677

3 -0.06081087 -0.8940117

4 1.07544084 -0.9273513

5 -0.07901957 1.2293360

6 1.59244547 -0.7709945

7 -1.30980048 -1.2851501

1. Create the WSS plots as demonstrated in class and select a suitable k value based on the “elbow”. [NOTE – It is easiest to create this chart in Excel or some other spreadsheet program]

Table

Description automatically generated

Chart, line chart

Description automatically generated

In above elbow chart the point of inflection is between k=3 and k=4 but if you observe the Total\_SS is significantly decreasing up to cluster k=5 and after that it is decreasing steadily. Let’s take cluster k=5 and observe the data using scatter plot.

**Evaluation of Cluster:**

1. Based on the “k” chosen above, create a scatter plot showing the clusters and colour-coded datapoints for each of “k-1”, “k”, “k+1”. For example, if you think the “elbow” is at k=4 create the charts for k=3, k=4 and k=5.

K=4

Chart, scatter chart

Description automatically generated

K=5

Chart, scatter chart

Description automatically generated

K=6

Chart, scatter chart

Description automatically generated

1. Based on the WSS plot (3.2) and the charts (4.1) choose one set of clusters that best describes the data

Ans:- Based on the 3.2 elbow chart where we choose and the 4.1 scatter plots the cluster k=5 is best describing the data.

1. Create summary tables for the segmentation/clustering scheme (selected in step 4.2).

**cluster5** Food Entr Educ Tran Work Hous Othr N

1 0.201 0.0305 0.148 0.0493 0.0400 0.500 0.0306 149

2 0.151 0.0197 0.210 0.0202 0.0812 0.499 0.0201 191

3 0.101 0.0977 0.0262 0.223 0.0806 0.403 0.0685 103

4 0.103 0.116 0.0261 0.198 0.0303 0.456 0.0697 51

5 0.150 0.0854 0.179 0.143 0.0402 0.362 0.0396 106

1. Create suitable descriptive names for each cluster.

* Cluster-1 🡪Parents(Spending most on housing)
* Cluster-2🡪Students(Spending most on Education)
* Cluster-3🡪Employees(using most transportation)
* Cluster-4🡪Children’s
* Clusteer-5🡪Min\_Housing

1. Suggest possible uses for this clustering scheme

* Parents are spending their most income on **Food** and **Housing** and very less on **Other** category.
* Similarly, students are spending maximum on **Education** and **Housing** and very less on **Entr** category compared to other groups.
* Furthermore, Employees are spending mostly on **working**(tax).
* However, Children’s are spending approximately same on **Food, Entertainment,** and **other** category.