Clustering with WEKA

Harvey Alférez, PhD

k-means be like:

Case Study

- A local BMW dealership that is interested in increasing sales.
- The dealership has stored all its past sales information and information about each person who purchased a BMW, looked at a BMW, and browsed the BMW showroom floor.
- The dealership wants to increase future sales and employ data mining to accomplish this.

Clustering

- Question: “What age groups like the silver BMW M5?”
    - The data can be mined to compare the age of the purchaser of past cars and the colors bought in the past. From this data, it could be found whether certain age groups (22-30 year olds, for example) have a higher propensity to order a certain color of BMW M5s (75 percent buy blue).
    - Similarly, it can be shown that a different age group (55-62, for example) tend to order silver BMWs (65 percent buy silver, 20 percent buy gray).
    - The data, when mined, will tend to cluster around certain age groups and certain colors, allowing the user to quickly determine patterns in the data.

BMW Dataset

```
@attribute Dealership numeric
@attribute Showroom numeric
@attribute ComputerSearch numeric
@attribute M5 numeric
@attribute 3Series numeric
@attribute Z4 numeric
@attribute Financing numeric
@attribute Purchase numeric

@data

1,0,0,0,0,0,0,0
1,1,1,0,0,0,1,0
...
```

Clustering

- Clustering allows a user to make groups of data to determine patterns from the data.
- Clustering has its advantages when the data set is defined and a general pattern needs to be determined from the data.
- You can create a specific number of groups, depending on your business needs.
- A major disadvantage of using clustering is that the user is required to know ahead of time how many groups he wants to create.
- However, for the average user, clustering can be the most useful data mining method you can use! It can quickly take your entire set of data and turn it into groups, from which you can quickly make some conclusions.

Clustering - Overview of the Math

1. Every attribute in the data set should be normalized, whereby each value is divided by the difference between the high value and the low value in the data set for that attribute.

2. Given the number of desired clusters, randomly select that number of samples from the data set to serve as our initial test cluster centers. For example, if you want to have three clusters, you would randomly select three rows of data from the data set.

3. Compute the distance from each data sample to the cluster center (our randomly selected data row), using the least-squares method of distance calculation.

4. Assign each data row into a cluster, based on the minimum distance to each cluster center.

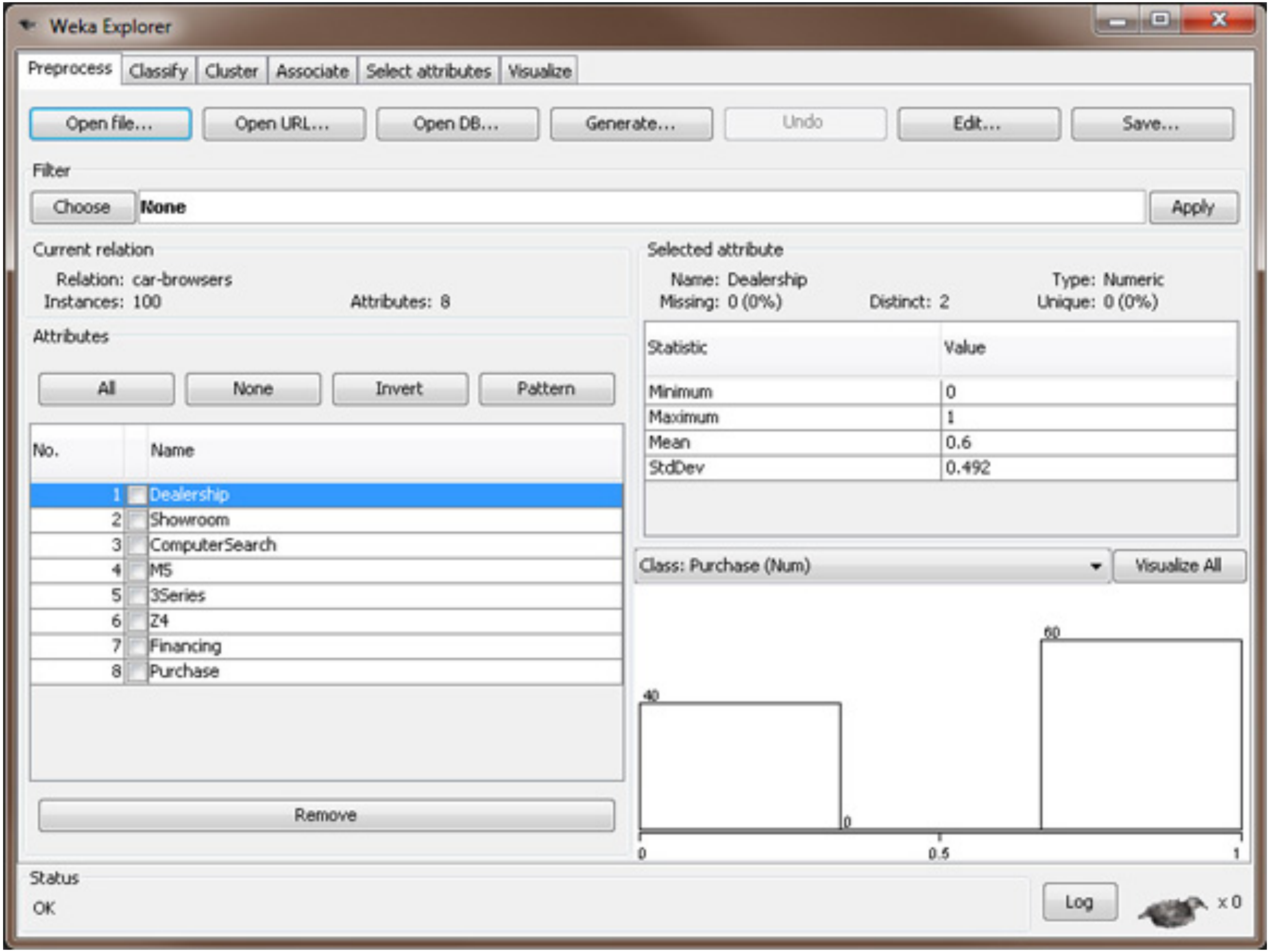
5. Compute the centroid, which is the average of each column of data using only the members of each cluster.

6. Calculate the distance from each data sample to the centroids you just created. If the clusters and cluster members don't change, you are complete and your clusters are created. If they change, you need to start over by going back to step 3, and continuing again and again until they don't change clusters.

Example

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Clustering in Weka



Filter: None

Current relation

Relation: car-browsers
Instances: 100
Attributes: 8

Selected attribute

Name: Dealership
Type: Numeric
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

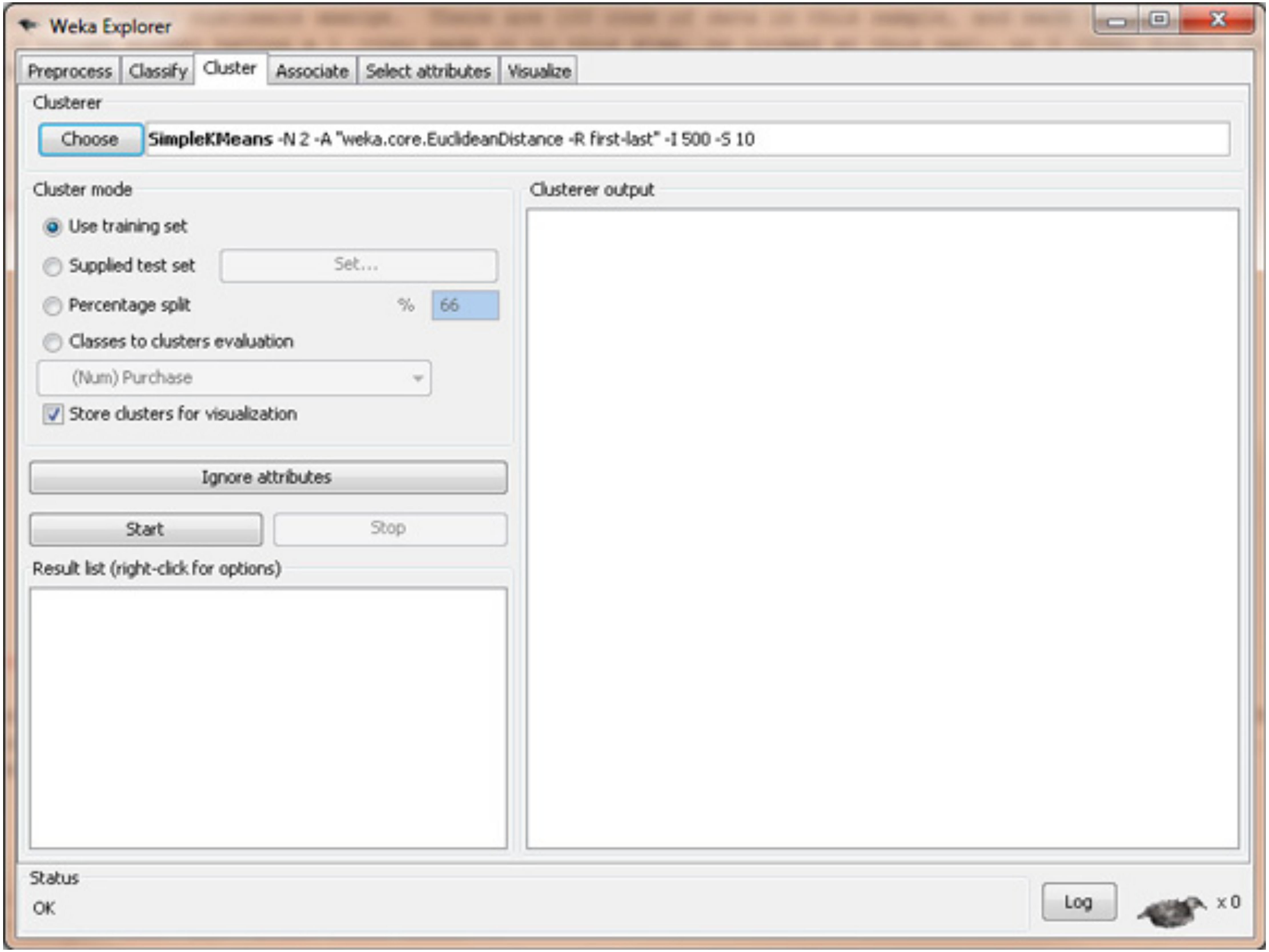
Statistic	Value
Minimum	0
Maximum	1
Mean	0.6
StdDev	0.492

No.		Name
1	□	Dealership
2	□	Showroom
3	□	ComputerSearch
4	□	M5
5	□	3Series
6	□	Z4
7	□	Financing
8	□	Purchase

Class: Purchase (Num)

Status
OK

Clustering in Weka



Clusterer

Choose **SimpleKMeans** -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

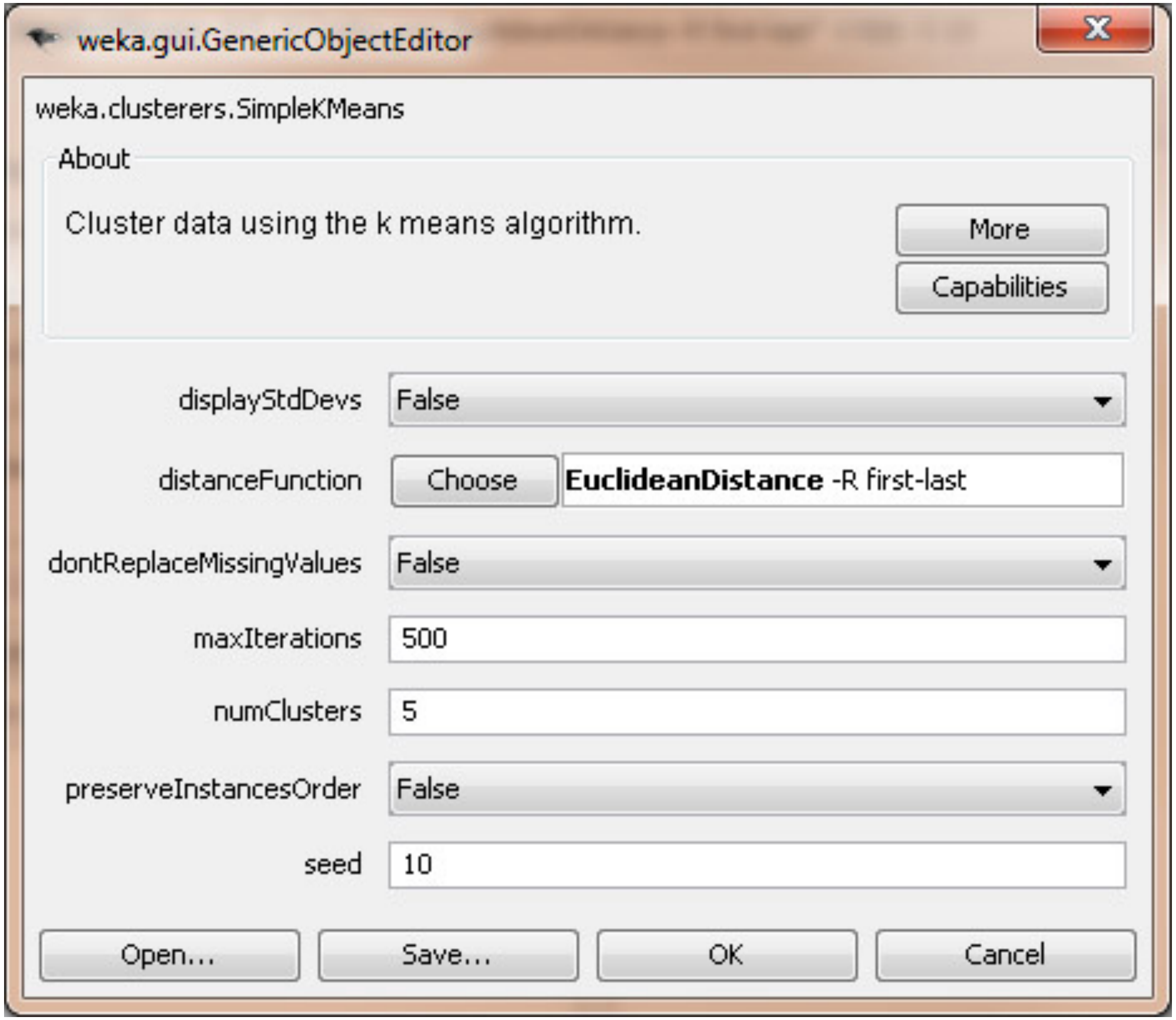
Cluster mode

- ● Use training set
- ○ Supplied test set Set...
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Num) Purchase
- ☑ Store clusters for visualization

Status

OK

Clustering in Weka



weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

displayStdDevs: False

distanceFunction: **EuclideanDistance** -R first-last

dontReplaceMissingValues: False

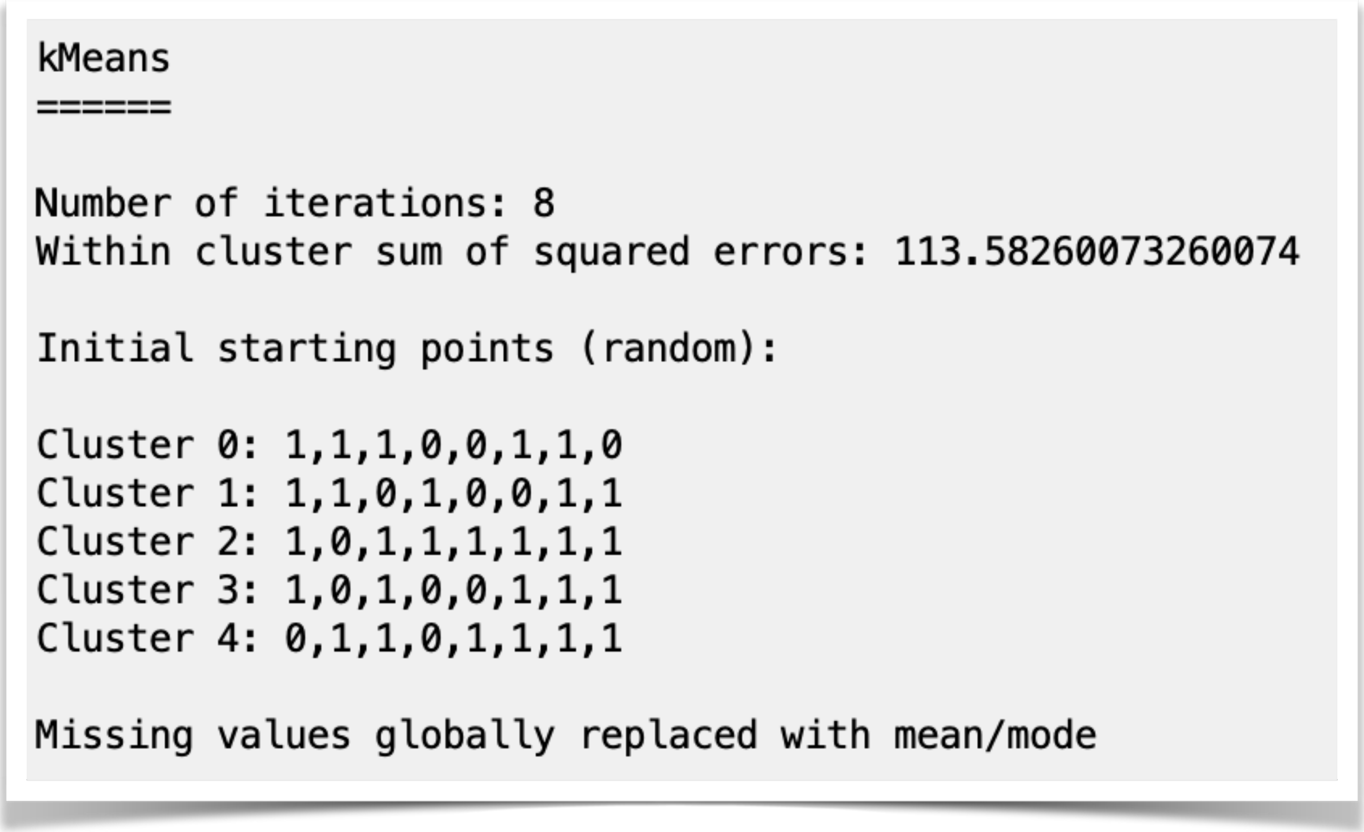
maxIterations: 500

numClusters: 5

preserveInstancesOrder: False

seed: 10

Clustering in Weka

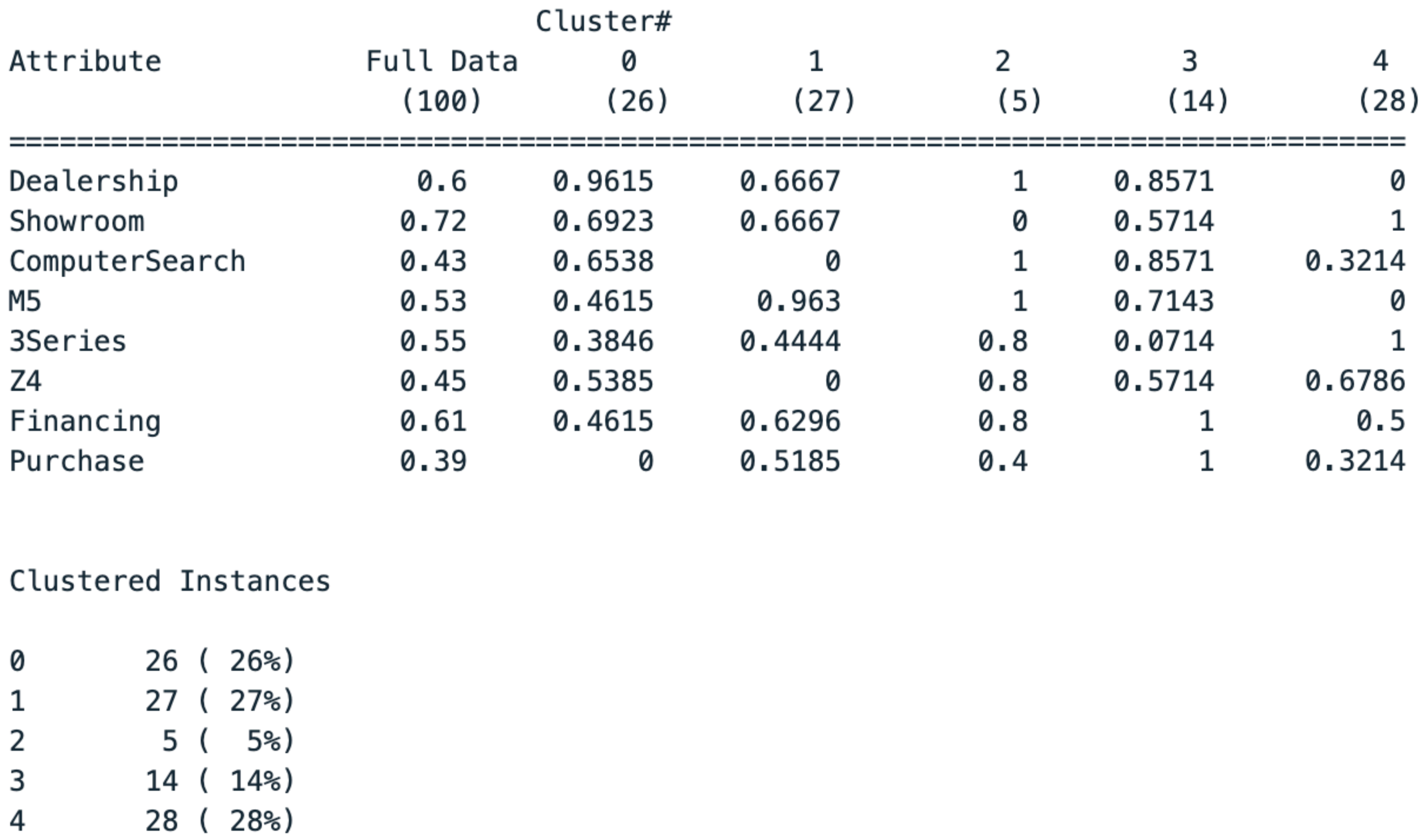


```
kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 113.58260073260074

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,1,1,1
Cluster 4: 0,1,1,0,1,1,1,1

Missing values globally replaced with mean/mode
```

Sum of Squared Errors (SSE): SSE quantifies how far each data point is from the centroid of its assigned cluster. For each data point in a cluster, you calculate the squared Euclidean distance between the data point and the centroid of its cluster and sum these squared distances across all data points in all clusters.

Within-Cluster Sum of Squared Errors (WCSS): It is the sum of the SSE values for each cluster. In other words, WCSS measures how tightly the data points are clustered within their respective clusters. A lower WCSS indicates that the data points in each cluster are closer to each other, suggesting a better clustering solution.

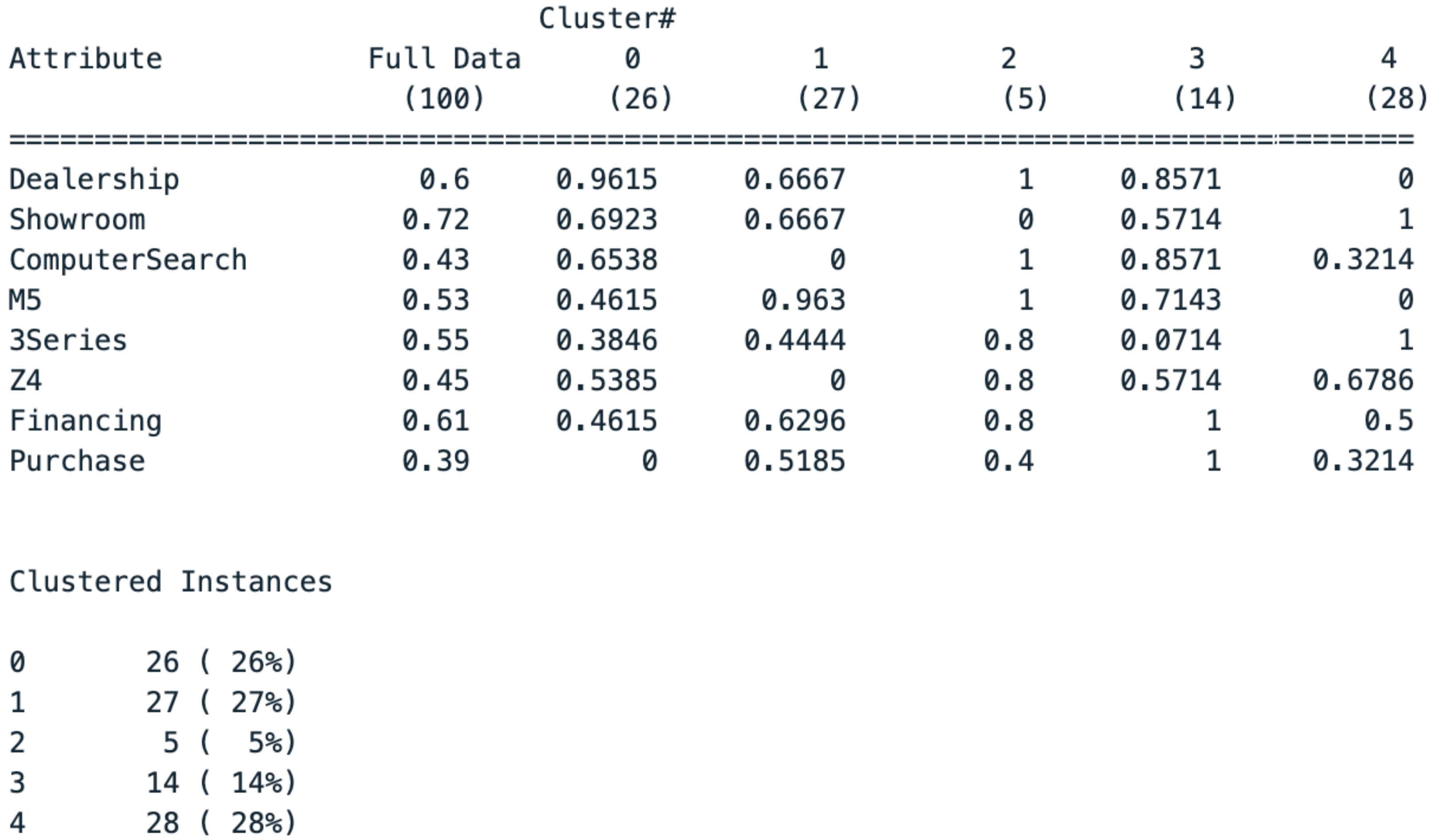
Clustering in Weka



```
                           Cluster#
Attribute         Full Data        0          1          2          3          4
                    (100)        (26)       (27)        (5)       (14)       (28)
==================================================================================
Dealership              0.6   0.9615     0.6667          1     0.8571          0
Showroom               0.72   0.6923     0.6667          0     0.5714          1
ComputerSearch         0.43   0.6538          0          1     0.8571     0.3214
M5                     0.53   0.4615      0.963          1     0.7143          0
3Series                0.55   0.3846     0.4444        0.8     0.0714          1
Z4                     0.45   0.5385          0        0.8     0.5714     0.6786
Financing              0.61   0.4615     0.6296        0.8          1        0.5
Purchase               0.39        0     0.5185        0.4          1     0.3214


Clustered Instances

0       26 ( 26%)
1       27 ( 27%)
2        5 (  5%)
3       14 ( 14%)
4       28 ( 28%)
```

Findings

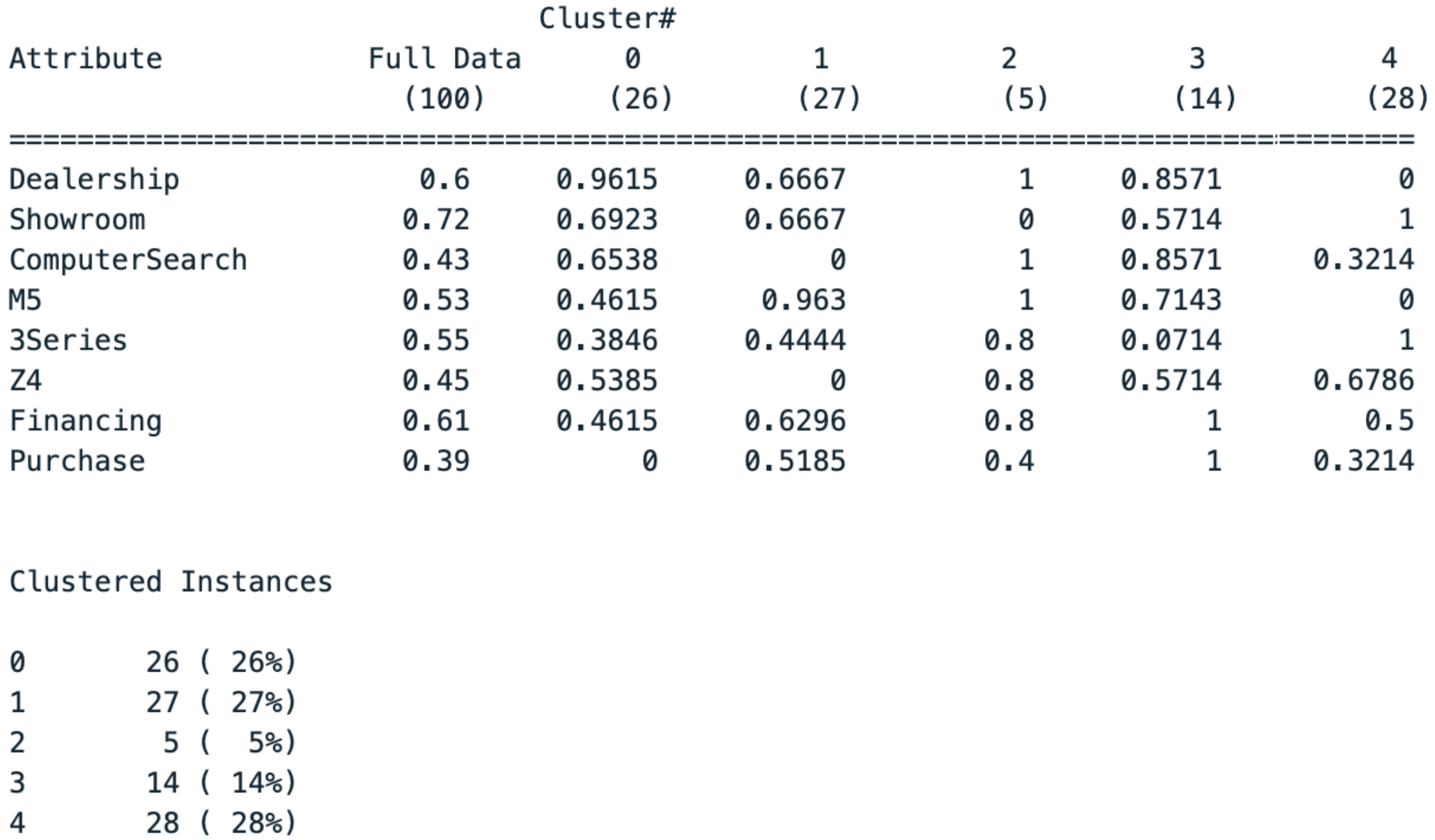
Cluster 0— This group we can call the “Dreamers,” as they appear to wander around the dealership, looking at cars parked outside on the lots, but trail off when it comes to coming into the dealership, and worst of all, they don’t purchase anything.



```
                              Cluster#
Attribute          Full Data         0          1          2          3          4
                       (100)      (26)       (27)        (5)       (14)       (28)
==================================================================================
Dealership               0.6    0.9615     0.6667          1     0.8571          0
Showroom                0.72    0.6923     0.6667          0     0.5714          1
ComputerSearch          0.43    0.6538          0          1     0.8571     0.3214
M5                      0.53    0.4615      0.963          1     0.7143          0
3Series                 0.55    0.3846     0.4444        0.8     0.0714          1
Z4                      0.45    0.5385          0        0.8     0.5714     0.6786
Financing               0.61    0.4615     0.6296        0.8          1        0.5
Purchase                0.39         0     0.5185        0.4          1     0.3214


Clustered Instances

0       26 ( 26%)
1       27 ( 27%)
2        5 (  5%)
3       14 ( 14%)
4       28 ( 28%)
```

Findings

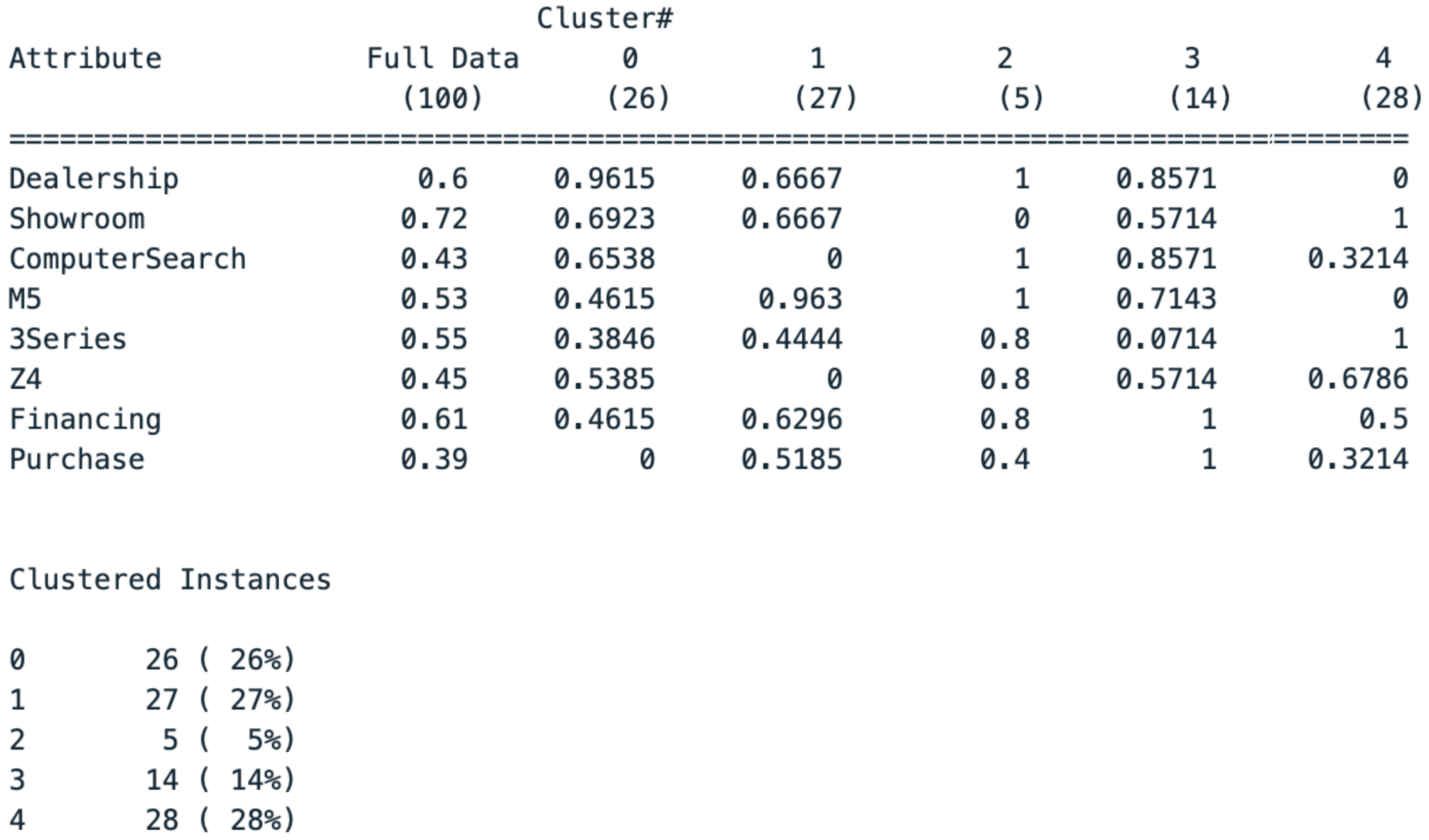
Cluster 1— We'll call this group the "M5 Lovers" because they tend to walk straight to the M5s, ignoring the 3-series cars and the Z4. However, they don't have a high purchase rate — only 52 percent. This is a potential problem and could be a focus for improvement for the dealership, perhaps by sending more salespeople to the M5 section.



```
                                      Cluster#
Attribute               Full Data          0          1          2          3          4
                            (100)       (26)       (27)        (5)       (14)       (28)
=====================================================================================
Dealership                    0.6     0.9615     0.6667          1     0.8571          0
Showroom                     0.72     0.6923     0.6667          0     0.5714          1
ComputerSearch               0.43     0.6538          0          1     0.8571     0.3214
M5                           0.53     0.4615      0.963          1     0.7143          0
3Series                      0.55     0.3846     0.4444        0.8     0.0714          1
Z4                           0.45     0.5385          0        0.8     0.5714     0.6786
Financing                    0.61     0.4615     0.6296        0.8          1        0.5
Purchase                     0.39          0     0.5185        0.4          1     0.3214


Clustered Instances

0       26 ( 26%)
1       27 ( 27%)
2        5 (  5%)
3       14 ( 14%)
4       28 ( 28%)
```

Findings

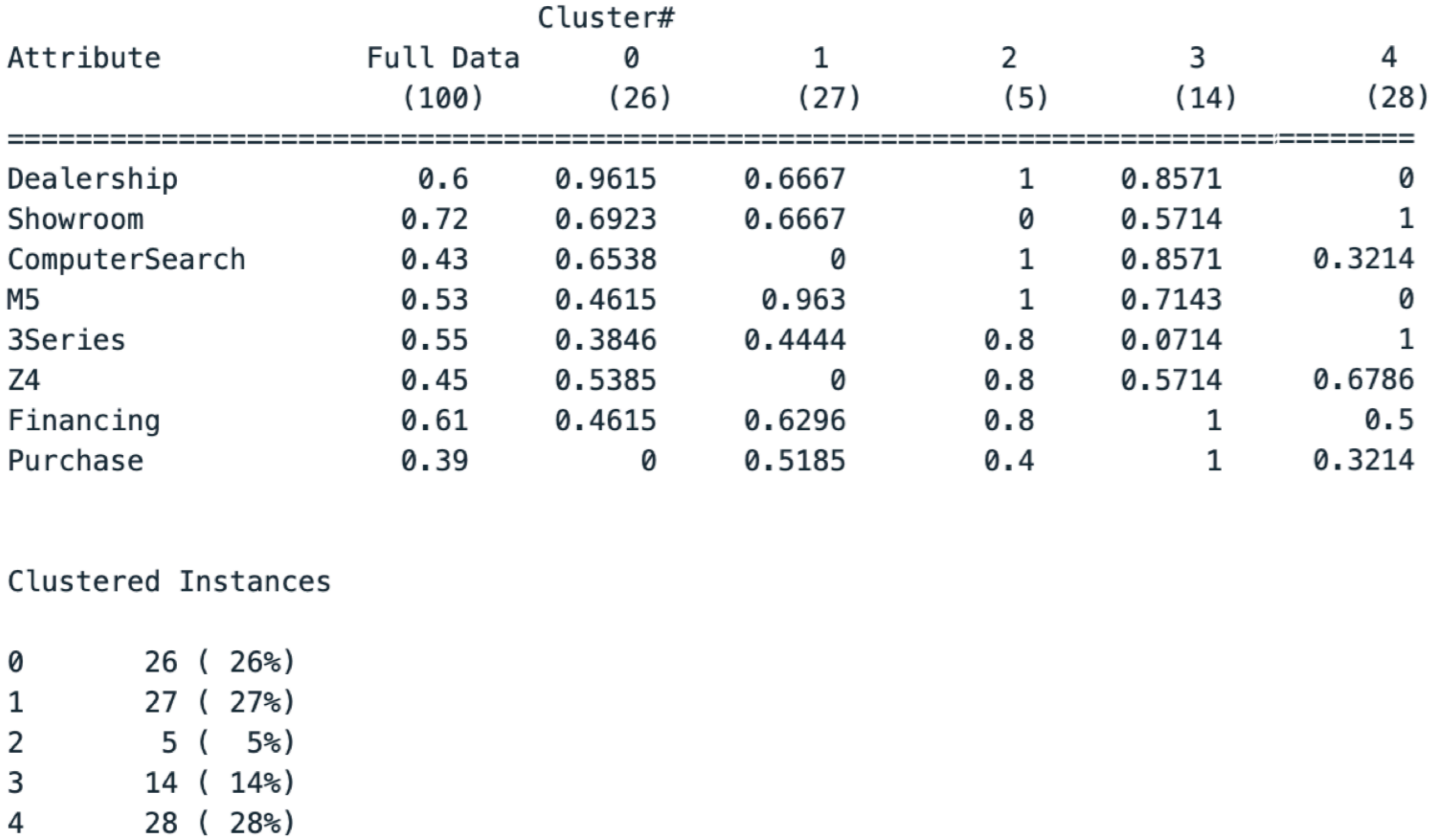
Cluster 2— This group is so small we can call them the “Throw-Aways” because they aren’t statistically relevant, and we can’t draw any good conclusions from their behavior. (This happens sometimes with clusters and may indicate that you should reduce the number of clusters you’ve created).



```
                              Cluster#
Attribute            Full Data         0          1          2          3          4
                       (100)        (26)       (27)        (5)       (14)       (28)
===================================================================================
Dealership               0.6     0.9615     0.6667          1     0.8571          0
Showroom                0.72     0.6923     0.6667          0     0.5714          1
ComputerSearch          0.43     0.6538          0          1     0.8571     0.3214
M5                      0.53     0.4615      0.963          1     0.7143          0
3Series                 0.55     0.3846     0.4444        0.8     0.0714          1
Z4                      0.45     0.5385          0        0.8     0.5714     0.6786
Financing               0.61     0.4615     0.6296        0.8          1        0.5
Purchase                0.39          0     0.5185        0.4          1     0.3214


Clustered Instances

0       26 ( 26%)
1       27 ( 27%)
2        5 (  5%)
3       14 ( 14%)
4       28 ( 28%)
```

Findings

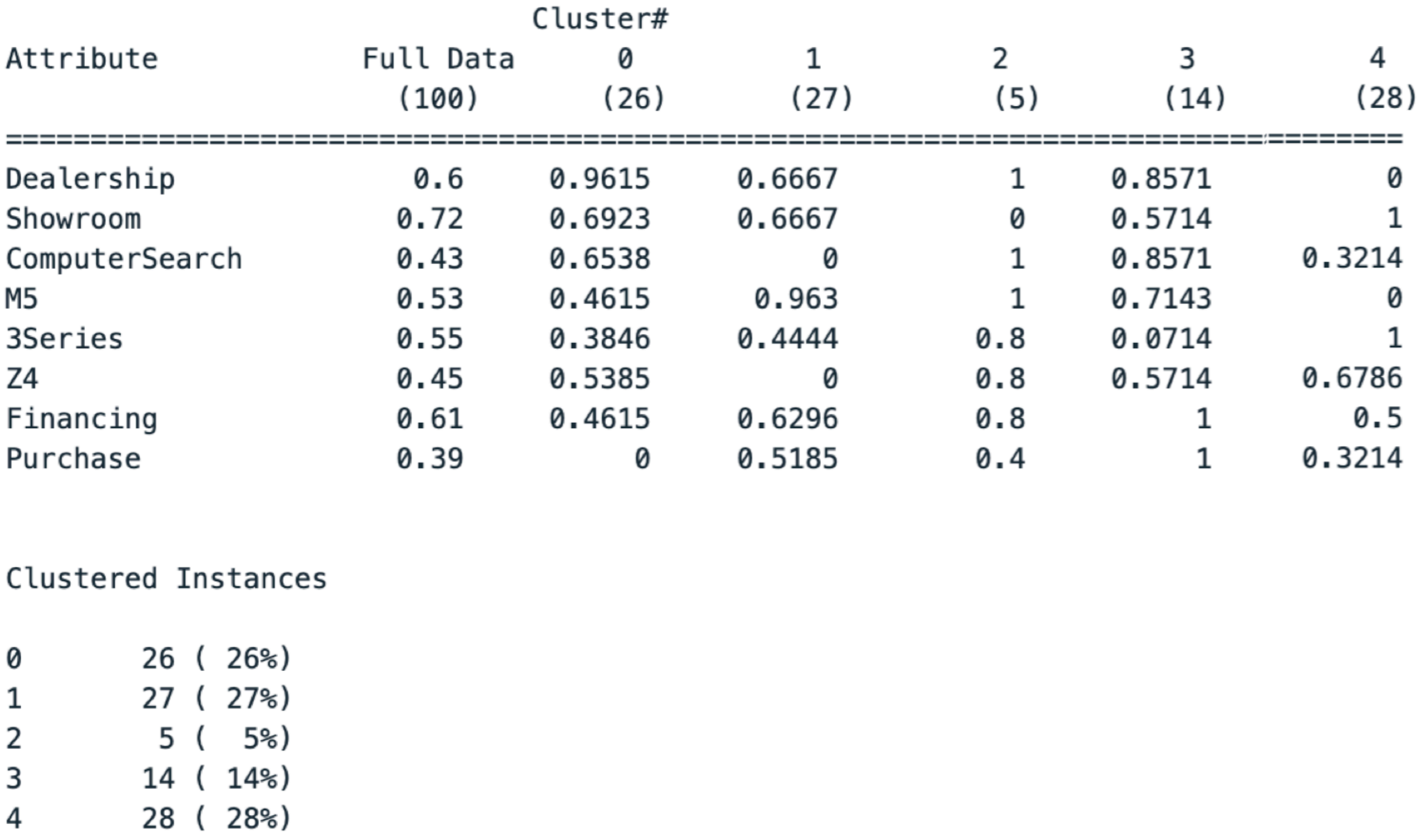
Cluster 3— This group we'll call the "BMW Babies" because they always end up purchasing a car and always end up financing it. Here's where the data shows us some interesting things: it appears they walk around the lot looking at cars, then turn to the computer search available at the dealership. Ultimately, they tend to buy M5s or Z4s (but never 3-series). This cluster tells the dealership that it should consider making its search computers more prominent around the lots (outdoor search computers?), and perhaps making the M5 or Z4 much more prominent in the search results. Once the customer has made up his mind to purchase the vehicle, he always qualifies for financing and completes the purchase.



```
                         Cluster#
Attribute            Full Data        0          1          2          3          4
                        (100)       (26)       (27)        (5)       (14)       (28)
=================================================================================
Dealership                0.6     0.9615     0.6667          1     0.8571          0
Showroom                 0.72     0.6923     0.6667          0     0.5714          1
ComputerSearch           0.43     0.6538          0          1     0.8571     0.3214
M5                       0.53     0.4615      0.963          1     0.7143          0
3Series                  0.55     0.3846     0.4444        0.8     0.0714          1
Z4                       0.45     0.5385          0        0.8     0.5714     0.6786
Financing                0.61     0.4615     0.6296        0.8          1        0.5
Purchase                 0.39          0     0.5185        0.4          1     0.3214


Clustered Instances

0       26 ( 26%)
1       27 ( 27%)
2        5 (  5%)
3       14 ( 14%)
4       28 ( 28%)
```

Findings

Cluster 4— This group we'll call the "Starting Out With BMW" because they always look at the 3-series and never look at the much more expensive M5. They walk right into the showroom, choosing not to walk around the lot and tend to ignore the computer search terminals. While 50 percent get to the financing stage, only 32 percent ultimately finish the transaction. The dealership could draw the conclusion that these customers looking to buy their first BMWs know exactly what kind of car they want (the 3-series entry-level model) and are hoping to qualify for financing to be able to afford it. The dealership could possibly increase sales to this group by relaxing their financing standards or by reducing the 3-series prices.



```
                               Cluster#
Attribute           Full Data       0          1          2          3          4
                      (100)       (26)       (27)        (5)       (14)       (28)
=====================================================================================
Dealership              0.6     0.9615     0.6667          1     0.8571          0
Showroom               0.72     0.6923     0.6667          0     0.5714          1
ComputerSearch         0.43     0.6538          0          1     0.8571     0.3214
M5                     0.53     0.4615      0.963          1     0.7143          0
3Series                0.55     0.3846     0.4444        0.8     0.0714          1
Z4                     0.45     0.5385          0        0.8     0.5714     0.6786
Financing              0.61     0.4615     0.6296        0.8          1        0.5
Purchase               0.39          0     0.5185        0.4          1     0.3214


Clustered Instances

0       26 ( 26%)
1       27 ( 27%)
2        5 (  5%)
3       14 ( 14%)
4       28 ( 28%)
```