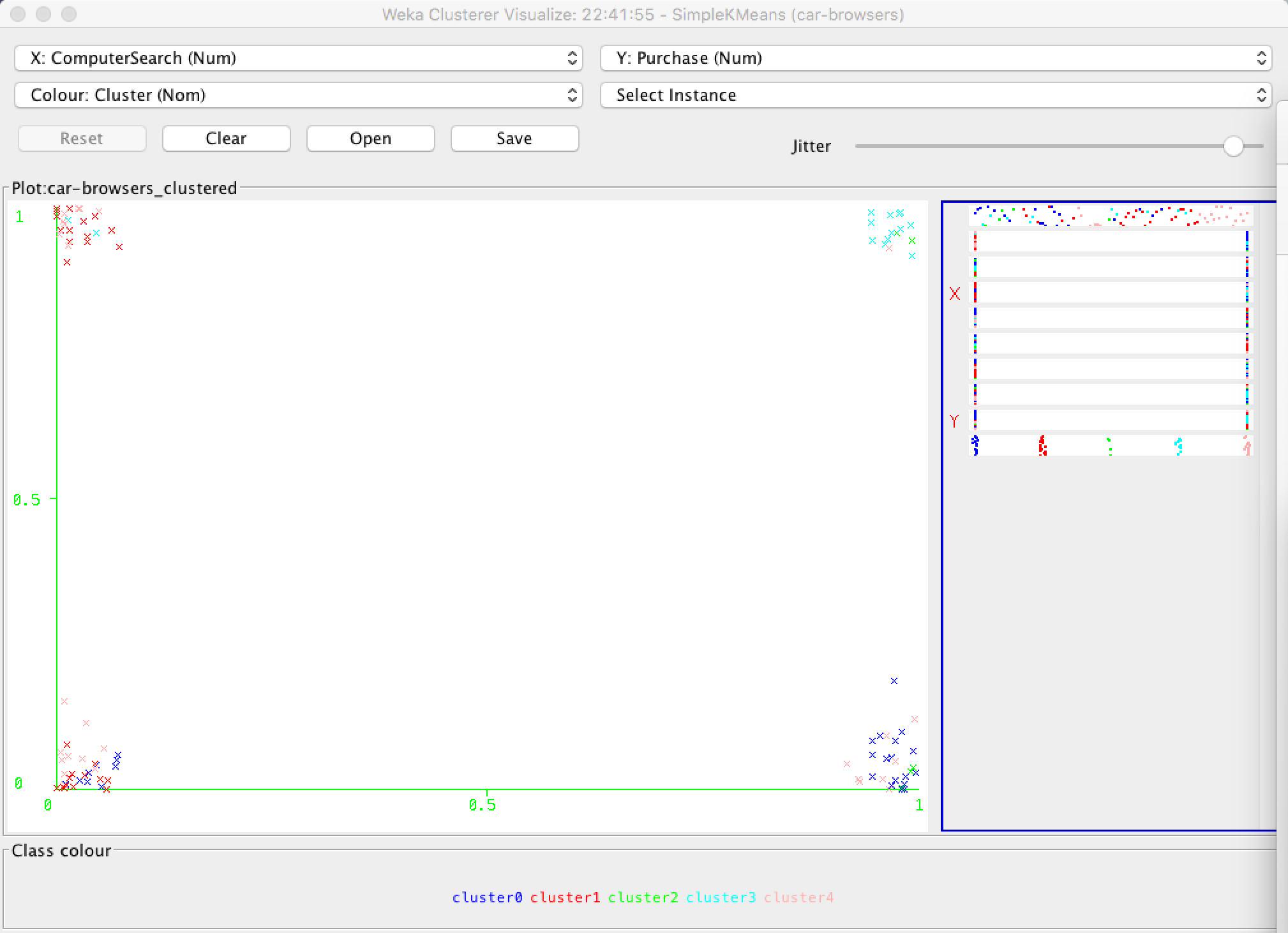
COMP 3503

Assignment #8

Lei Xie

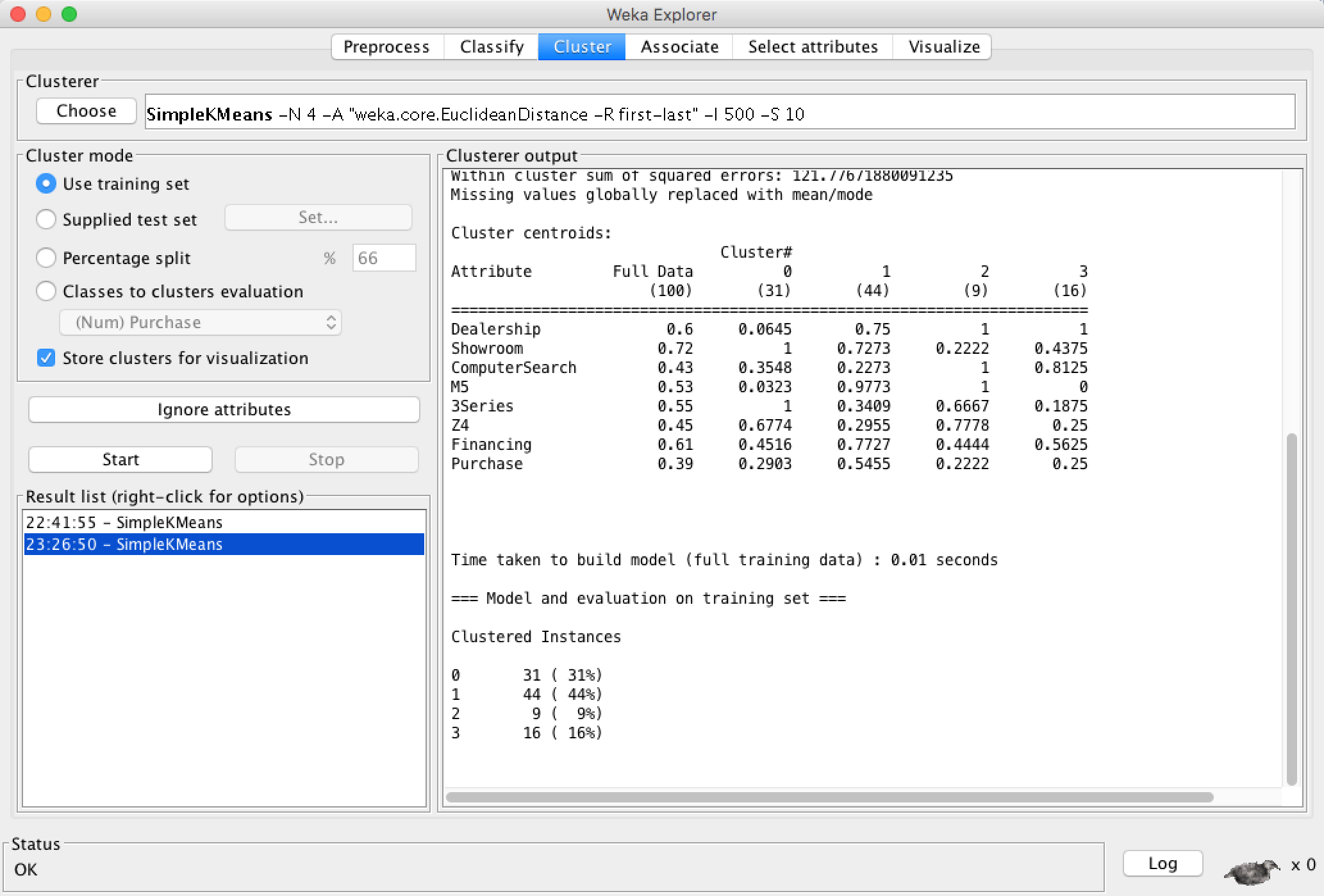
100123800

Part 1.



I choose ComputerSearch as my X, and Purchase as my Y.

The visual graph shows that, when X = 1, Y = 1, almost cluster 3. When X = 1, Y = 0, almost cluster 0, 4. Which match our conclusion, the “BMW Babies” likes to do computer search and most of them will purchase the car in the end. And for the cluster 0, they only do some looking outside and computer search but never purchase, cluster 4 always know what are they looking for, weather search or not, will not affect o their purchase.



we can analysis that

Cluster 0:

This group is like “Starting out with BMW” which means they know exactly what they are looking for is 3 Series, not interested in dealership, but all of the will visit the showroom half purchase rate than financing, which means they do not have to much money.

Cluster 1:

This group is super fan of “M5” and have high financing rate, average purchase rate. Few of them using computer searching, because they know what they are looking for is “M5”.

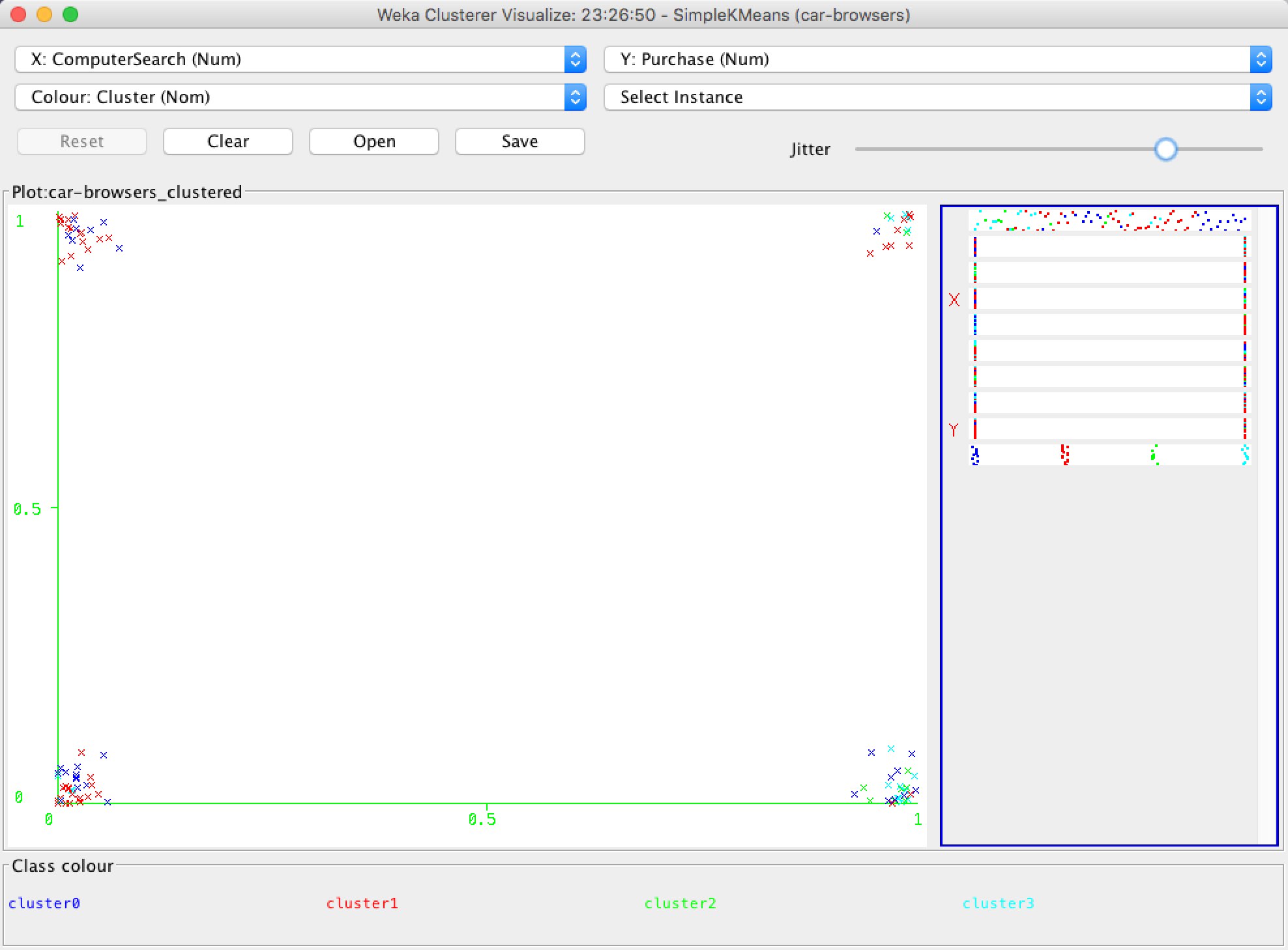
Cluster 2:

This group likes all the types of BMW, especially “M5”. The interesting thing is they have low purchase rate and average financing, they like the dealership but not the fan of showroom.

Cluster 3:

This group do not like BMW too much, and no one interested in “M5” but have the average financing rate and low purchase rate, they like to do a lot of computer search, and like dealership very much.

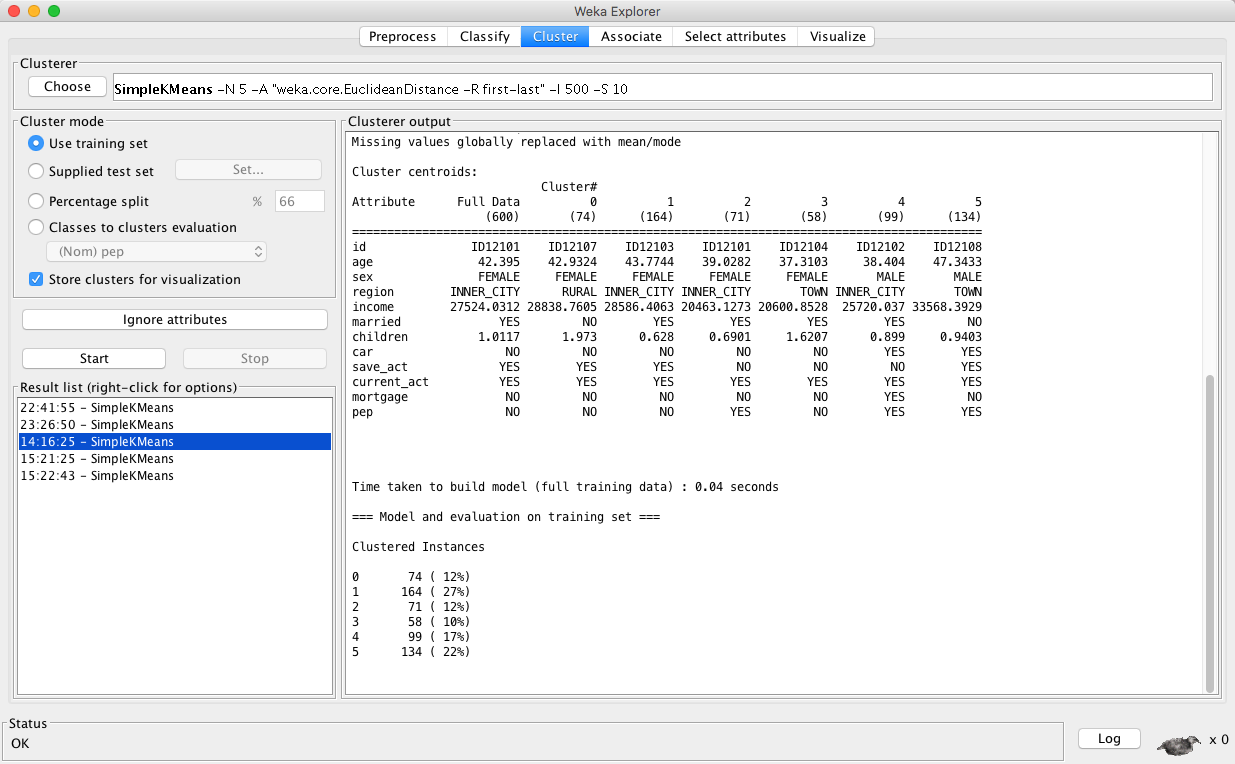
The visualize graph:



As we can see the difference between two visualize graphs with 5 and 4 clusters. The fewer clusters do no makes for better grouping, because the spots display in the graph was not clear as the graph with 5 clusters, and hard to notice the relationship.

Part 2.

The best number if clusters for this bank data is 6.



because, it clearly divides 4 female group and 2 male group, including Rural, Inner\_city and Town area. With different amount of income weather married or not, and also have clearly visualize graph with X: Cluster Y: instance number Color: sex.

Cluster 0:

“Single mom in rural area”, with 42.9 average age, have medium income around 28838.

They have 2 children with save act and current act but no mortgage or pep. No car.

Cluster 1:

“Married woman in inner city with save act”, with 43.7 average age, have medium income around 28386 similar with Cluster 0. Have current act but no mortgage or pep. No car.

Cluster 2:

“Married woman in inner city do not save”, with 39 average age, have low income around 20463, have current act and pep, but no mortgage, and have similar children with Cluster 1. No car.

Cluster 3:

“Married woman in town”, with 37 average age, have low income around 20600 similar with cluster 2, have 1.6 children, no car, save act mortgage or pep but have current act.

Cluster 4:

“Married man in inner city”, with 38.4 average age, have medium income around 25720, with 0.899 children car but do not do save. They have current act, mortgage and pep.

Cluster 5:

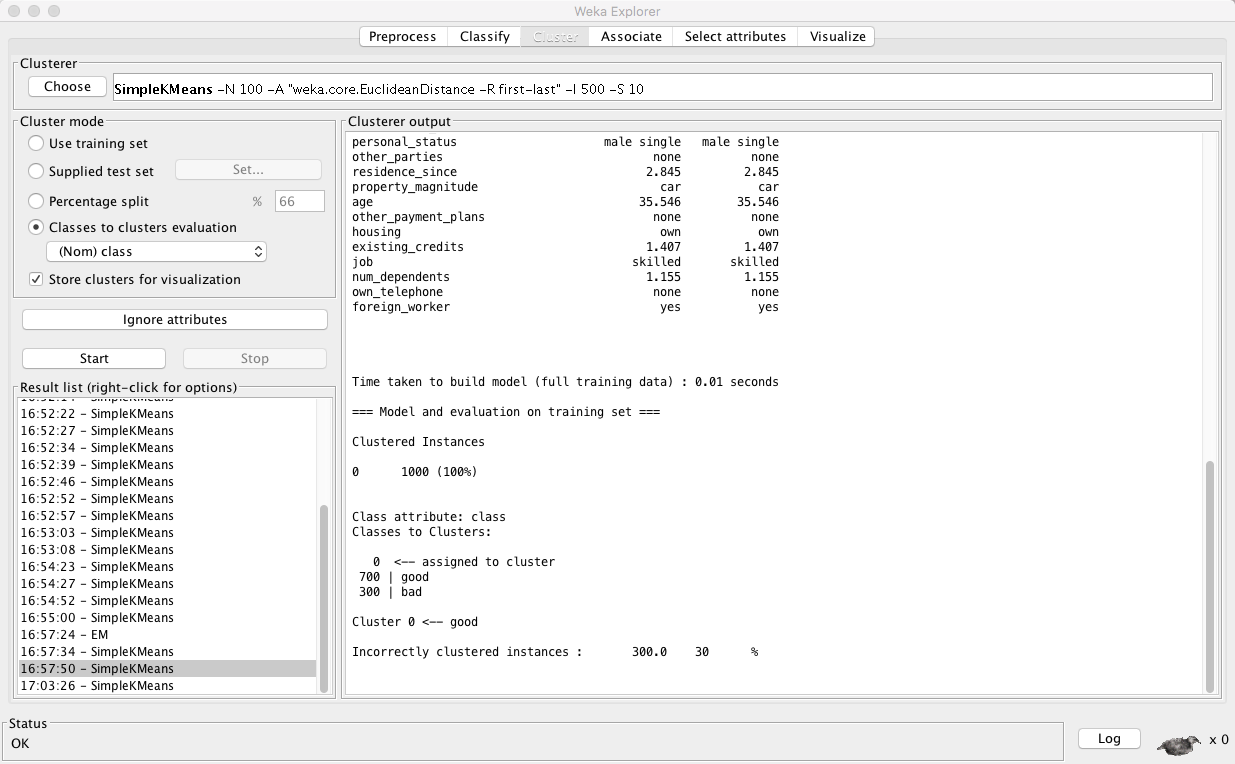
“Single man in town”, with 47.3 the highest average age, and also highest income around 33568, have 0.94 children, and car of course, but no mortgage, They have save act, current act and pep as well.

Part 3

K-Means:

I am using k-Means for this time, and I set the number of cluster as 1, due to the hint of this problem, and default seed 10.

For “Classes to clusters evaluation” choose “(Nom)class” get the smallest possible percentage is 30%



EM models:

I am using EM models for this time, and I set the number of cluster as 1 as well, and others as default.

For “Classes to clusters evaluation” choose “(Nom)class” get the smallest possible percentage is 30%

