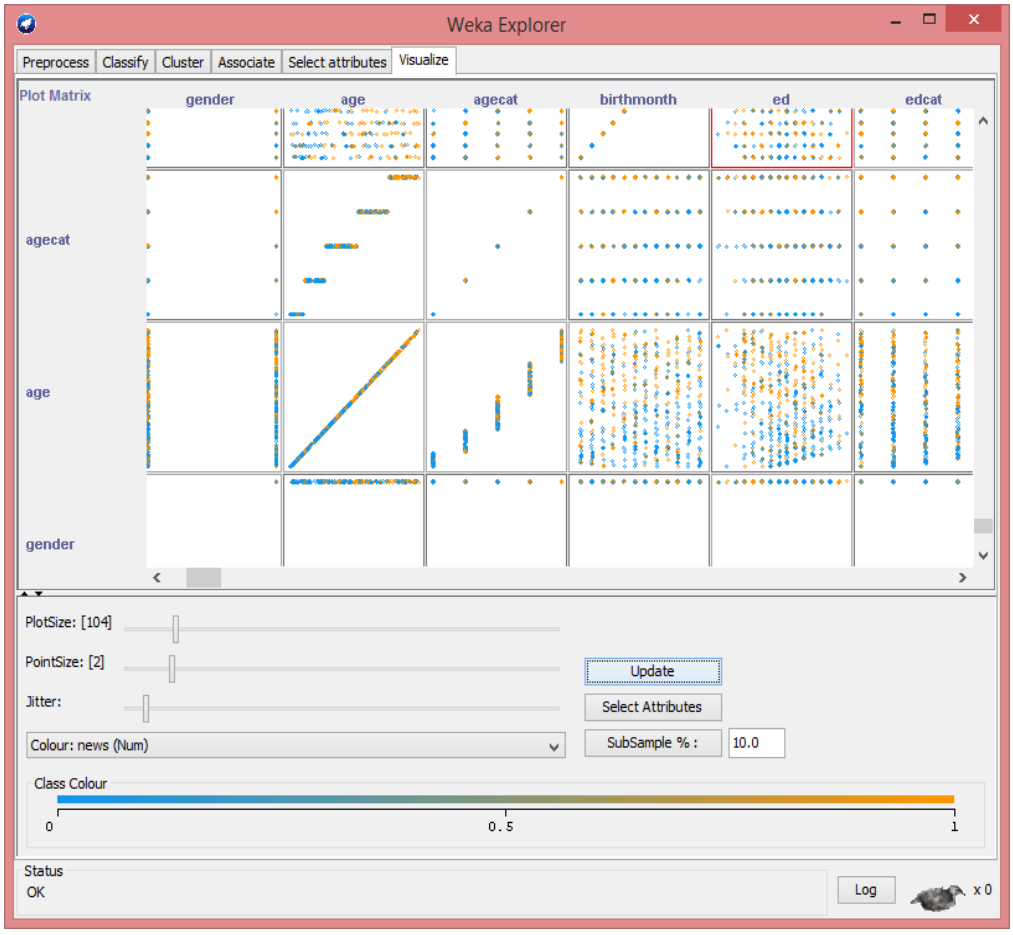
Lab Exercise Four
Clustering with WEKA Explorer

1. Fire up WEKA to get the GUI Chooser panel. Select Explorer from the four choices on the right side.

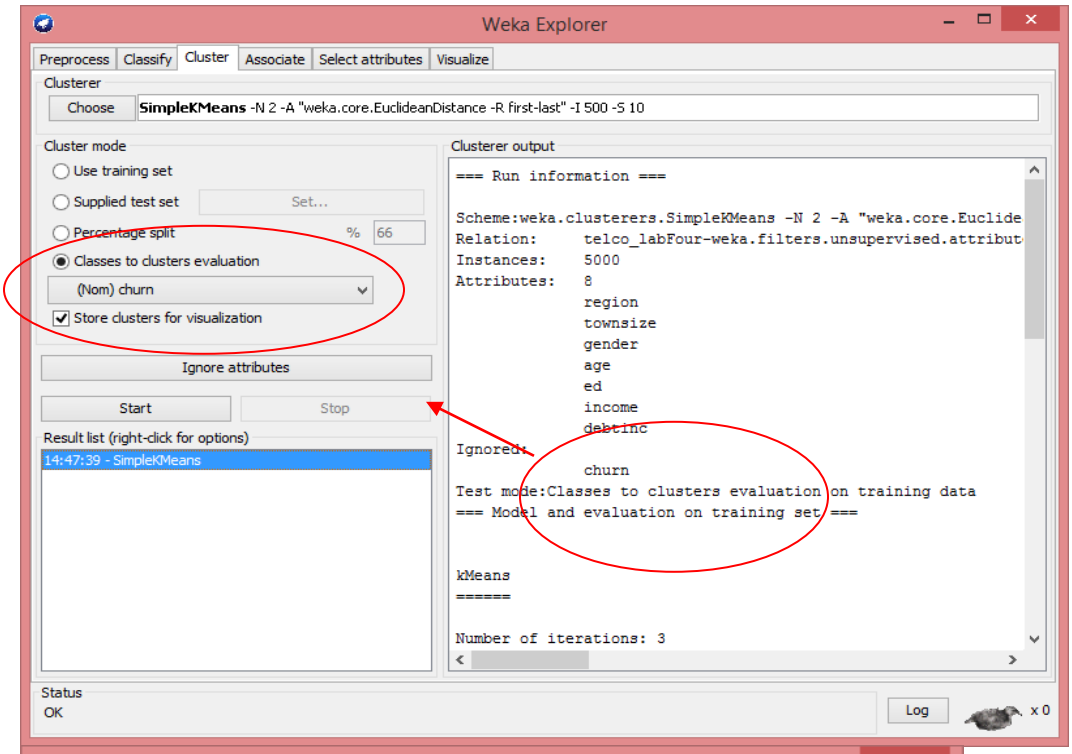
2. We are on ***Preprocess*** now. Click the ***Open file*** button to bring up a standard dialog through which you can select a file. Choose the **telco_labFour.csv** file.

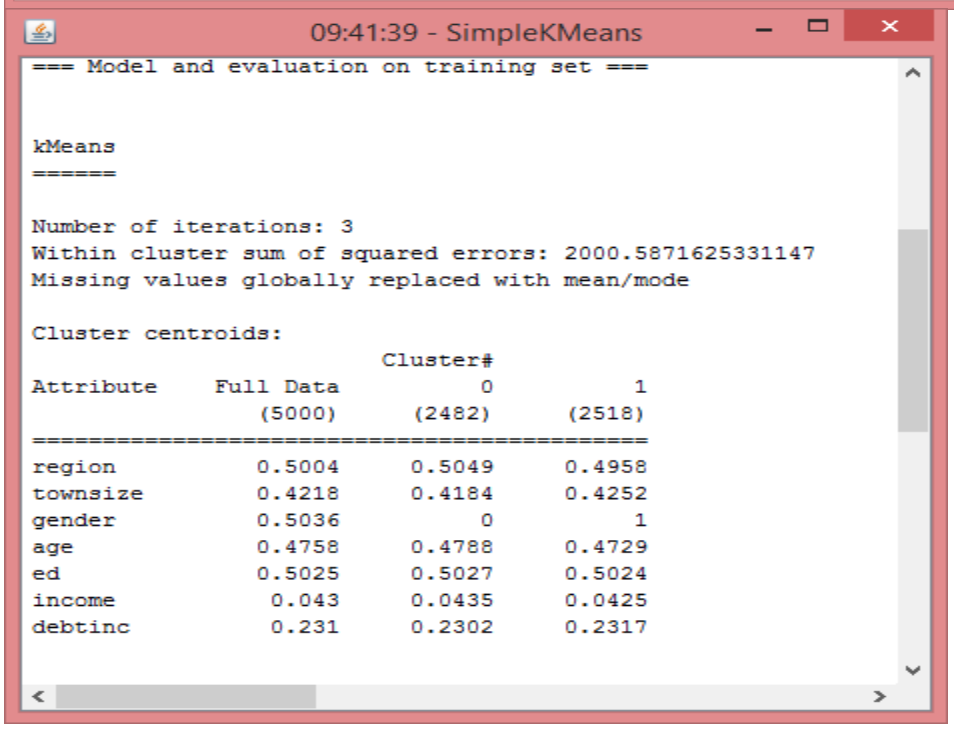
3. You could remove irrelevant attributes manually, lick **custIds**. To remove redundant attributes, we could find the correlation from Visualization of the data set under Visualize Tab. ***age*** and ***agecat*** are correlated. One of them should be removed. We keep ***age*** for clustering purpose; also ed (removing edcat), then we have 8 attributes left.

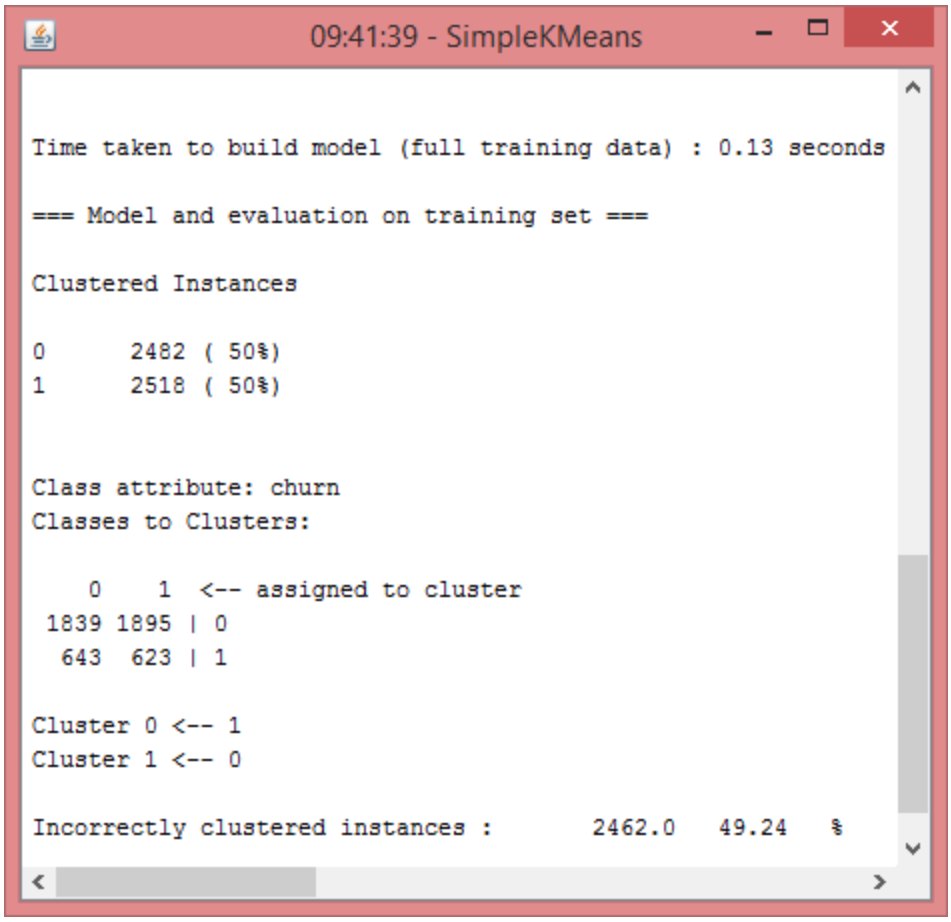


4. Before we do clustering with Weka, we need to normalize your numeric data values. Since we have the class label, we would like to set it to nominal before normalization. This information will be used to evaluate the clustering performance.

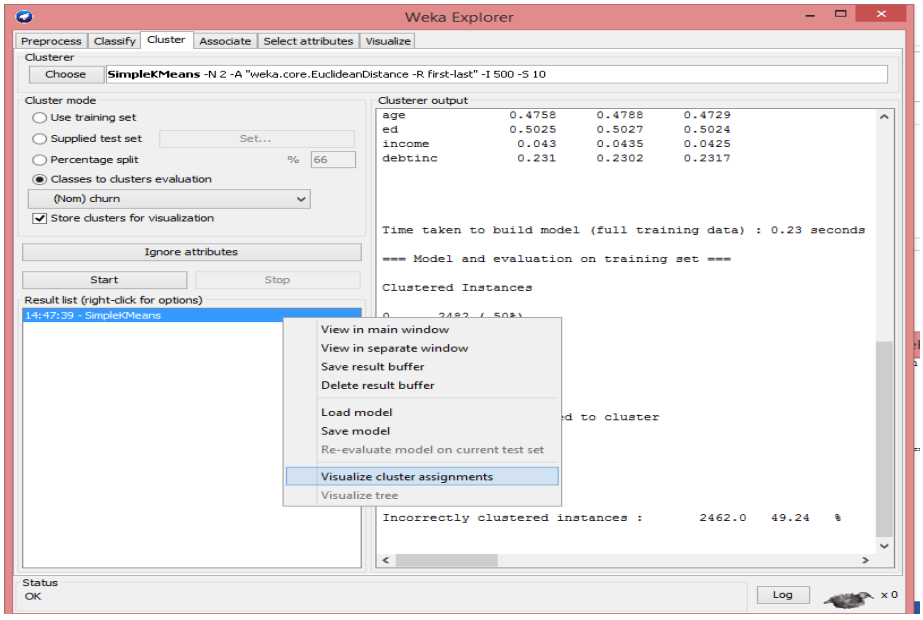
5. To perform clustering on the data set, click ***Cluster*** tab and choose ***SimpleKMeans*** algorithm. We set **k = 2** for this data set. Choose *Classes to clusters evaluation* and select the last attribute as class label. Check *Store clusters for visualization*. Click ***Ignore attributes*** and select the last attribute *churn*. Then click Start.

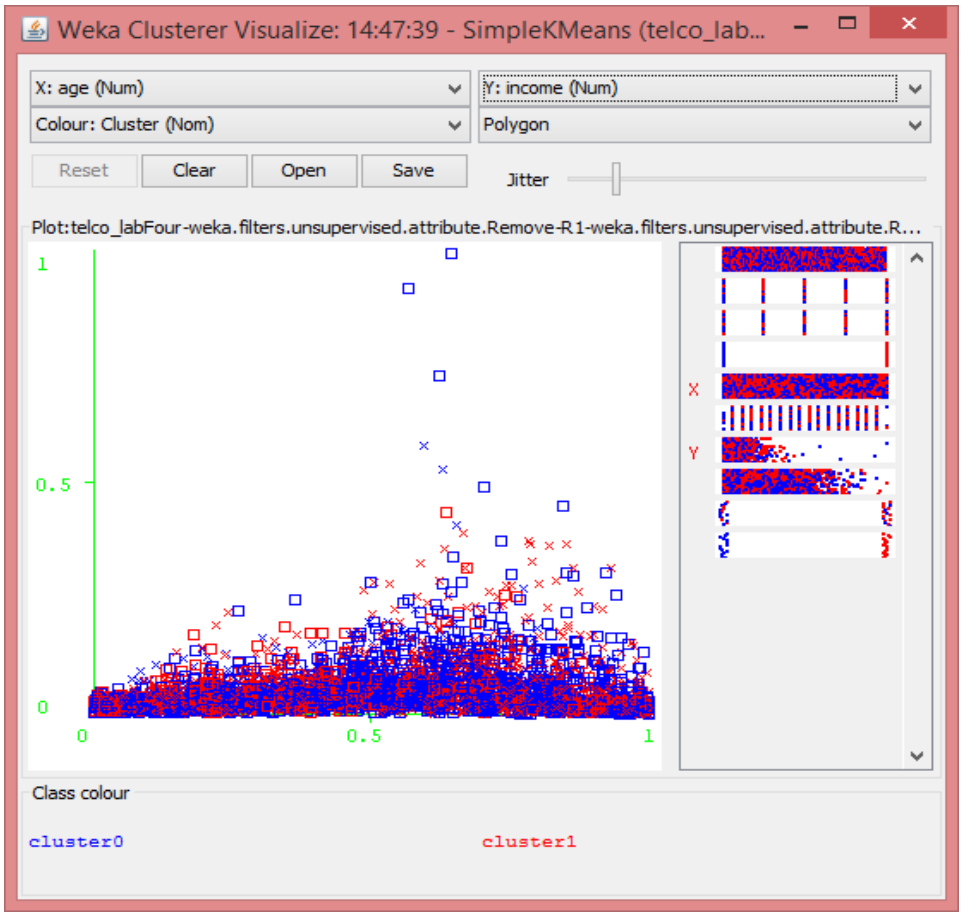




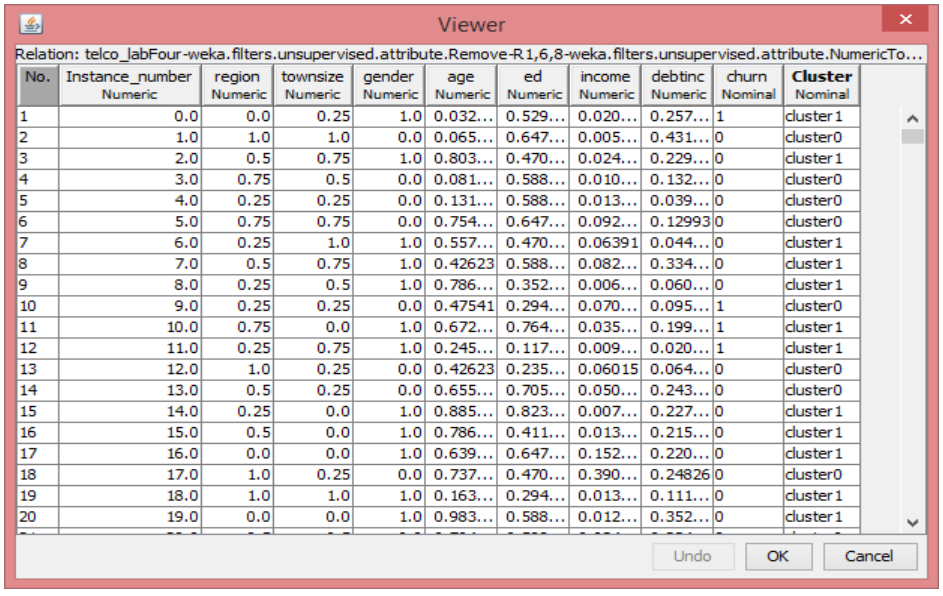


6. You could visualize the clustering results by right-clicking the result list and choose visualize clusters assignments. You could select different combination of two attributes as X and Y.



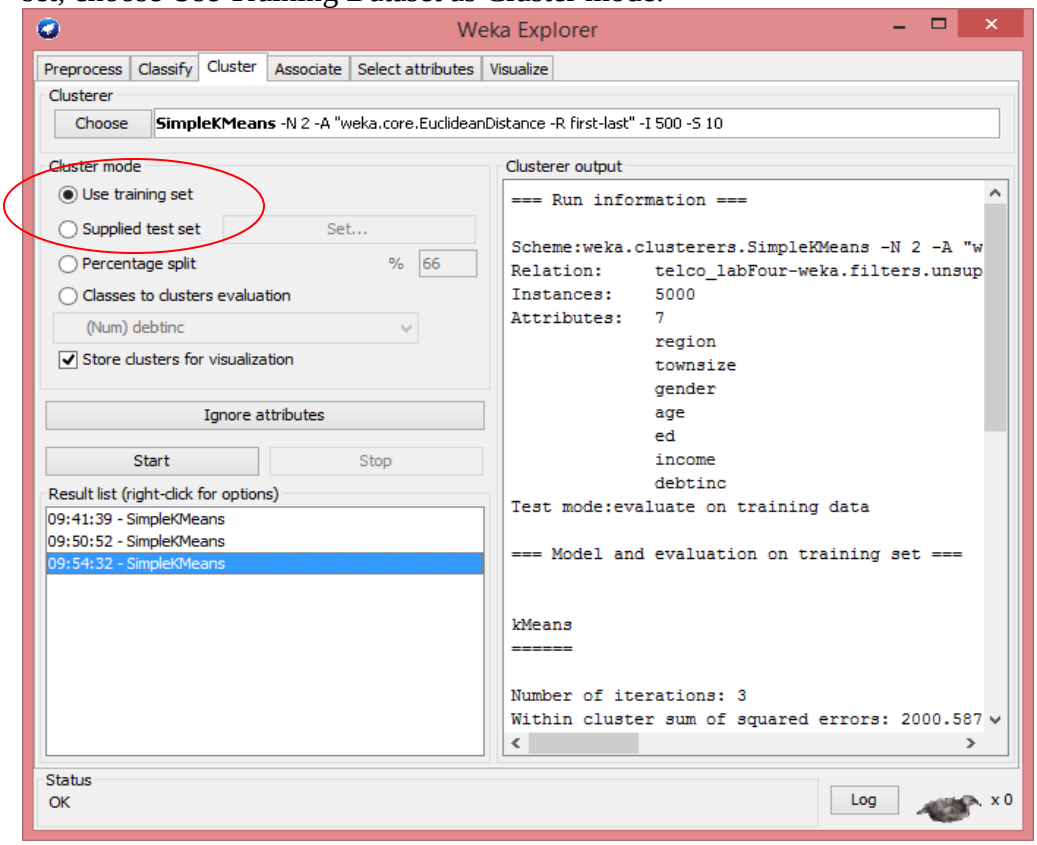


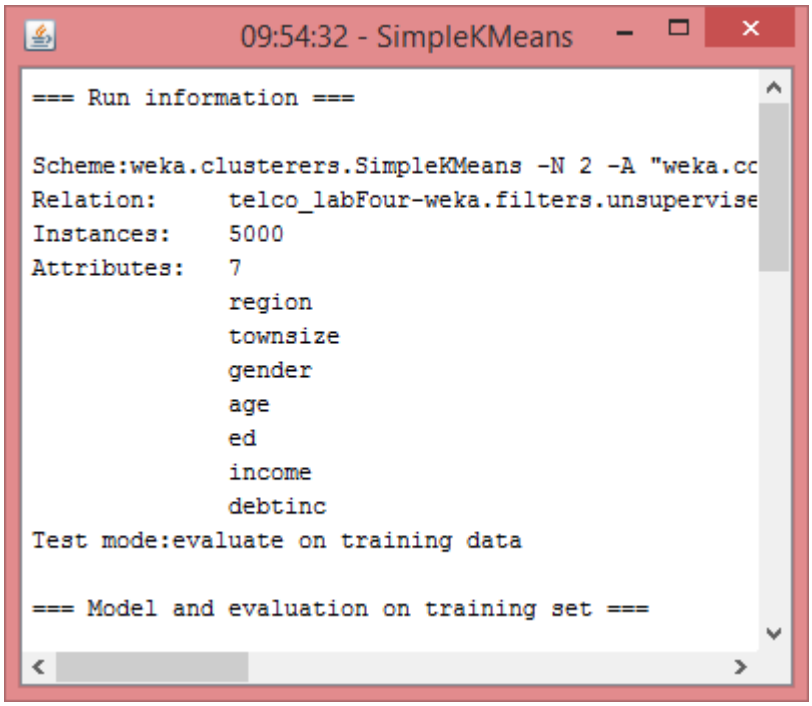
7. You could save the clustering results by clicking Save button on the Visualization panel. The results are saved in a .arff file. You could use Weka to open it and view the results.

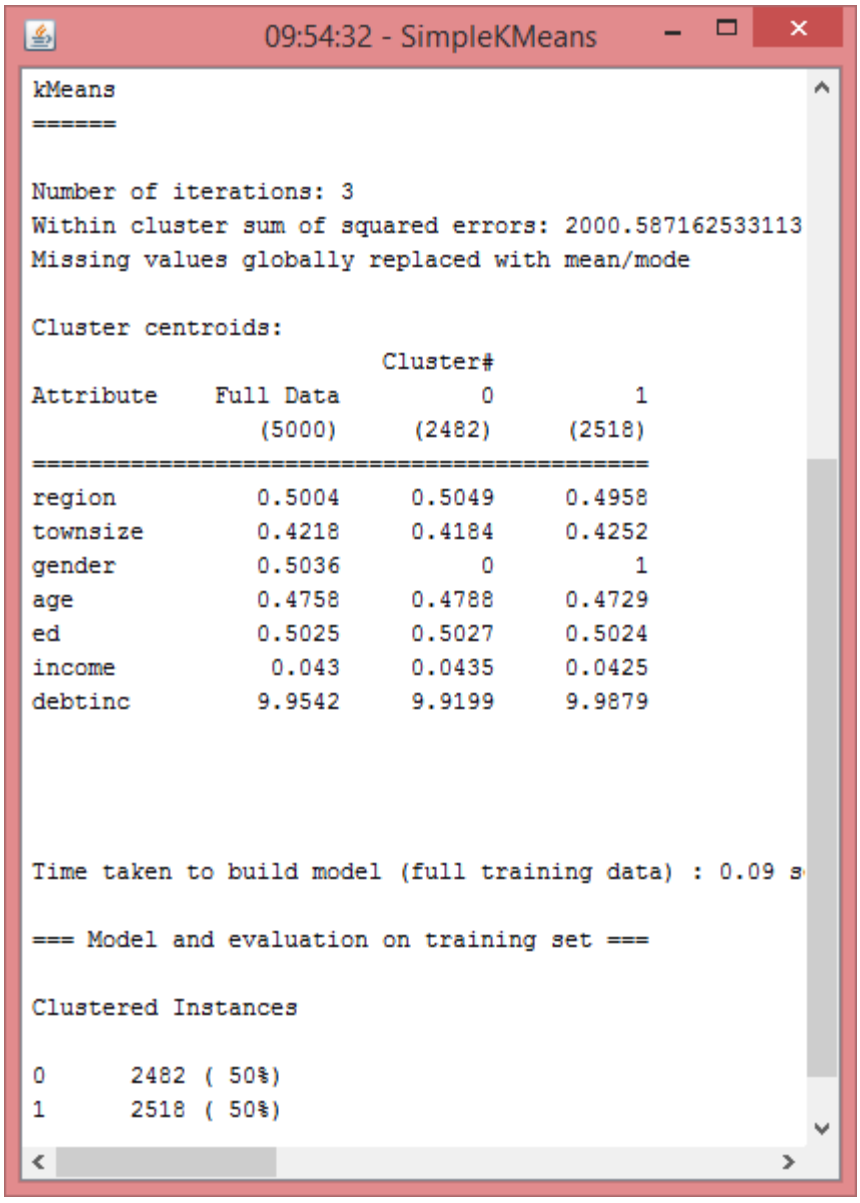


No.	Instance_number Numeric	region Numeric	townsize Numeric	gender Numeric	age Numeric	ed Numeric	income Numeric	debtinc Numeric	churn Nominal	Cluster Nominal
1	0.0	0.0	0.25	1.0	0.032...	0.529...	0.020...	0.257...	1	cluster1
2	1.0	1.0	1.0	0.0	0.065...	0.647...	0.005...	0.431...	0	cluster0
3	2.0	0.5	0.75	1.0	0.803...	0.470...	0.024...	0.229...	0	cluster1
4	3.0	0.75	0.5	0.0	0.081...	0.588...	0.010...	0.132...	0	cluster0
5	4.0	0.25	0.25	0.0	0.131...	0.588...	0.013...	0.039...	0	cluster0
6	5.0	0.75	0.75	0.0	0.754...	0.647...	0.092...	0.12993	0	cluster0
7	6.0	0.25	1.0	1.0	0.557...	0.470...	0.06391	0.044...	0	cluster1
8	7.0	0.5	0.75	1.0	0.42623	0.588...	0.082...	0.334...	0	cluster1
9	8.0	0.25	0.5	1.0	0.786...	0.352...	0.006...	0.060...	0	cluster1
10	9.0	0.25	0.25	0.0	0.47541	0.294...	0.070...	0.095...	1	cluster0
11	10.0	0.75	0.0	1.0	0.672...	0.764...	0.035...	0.199...	1	cluster1
12	11.0	0.25	0.75	1.0	0.245...	0.117...	0.009...	0.020...	1	cluster1
13	12.0	1.0	0.25	0.0	0.42623	0.235...	0.06015	0.064...	0	cluster0
14	13.0	0.5	0.25	0.0	0.655...	0.705...	0.050...	0.243...	0	cluster0
15	14.0	0.25	0.0	1.0	0.885...	0.823...	0.007...	0.227...	0	cluster1
16	15.0	0.5	0.0	1.0	0.786...	0.411...	0.013...	0.215...	0	cluster1
17	16.0	0.0	0.0	1.0	0.639...	0.647...	0.152...	0.220...	0	cluster1
18	17.0	1.0	0.25	0.0	0.737...	0.470...	0.390...	0.24826	0	cluster0
19	18.0	1.0	1.0	1.0	0.163...	0.294...	0.013...	0.111...	0	cluster1
20	19.0	0.0	0.0	1.0	0.983...	0.588...	0.012...	0.352...	0	cluster1

8. If the data set has no class labels, then when you perform clustering on the data set, choose Use Training Dataset as Cluster mode.






09:54:32 - SimpleKMeans
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 2000.587162533113
Missing values globally replaced with mean/mode
Cluster centroids:
Cluster#
Attribute Full Data 0 1
(5000) (2482) (2518)
region 0.5004 0.5049 0.4958
townsize 0.4218 0.4184 0.4252
gender 0.5036 0 1
age 0.4758 0.4788 0.4729
ed 0.5025 0.5027 0.5024
income 0.043 0.0435 0.0425
debtinc 9.9542 9.9199 9.9879
Time taken to build model (full training data) : 0.09 s
=== Model and evaluation on training set ===
Clustered Instances
0 2482 (50%)
1 2518 (50%)