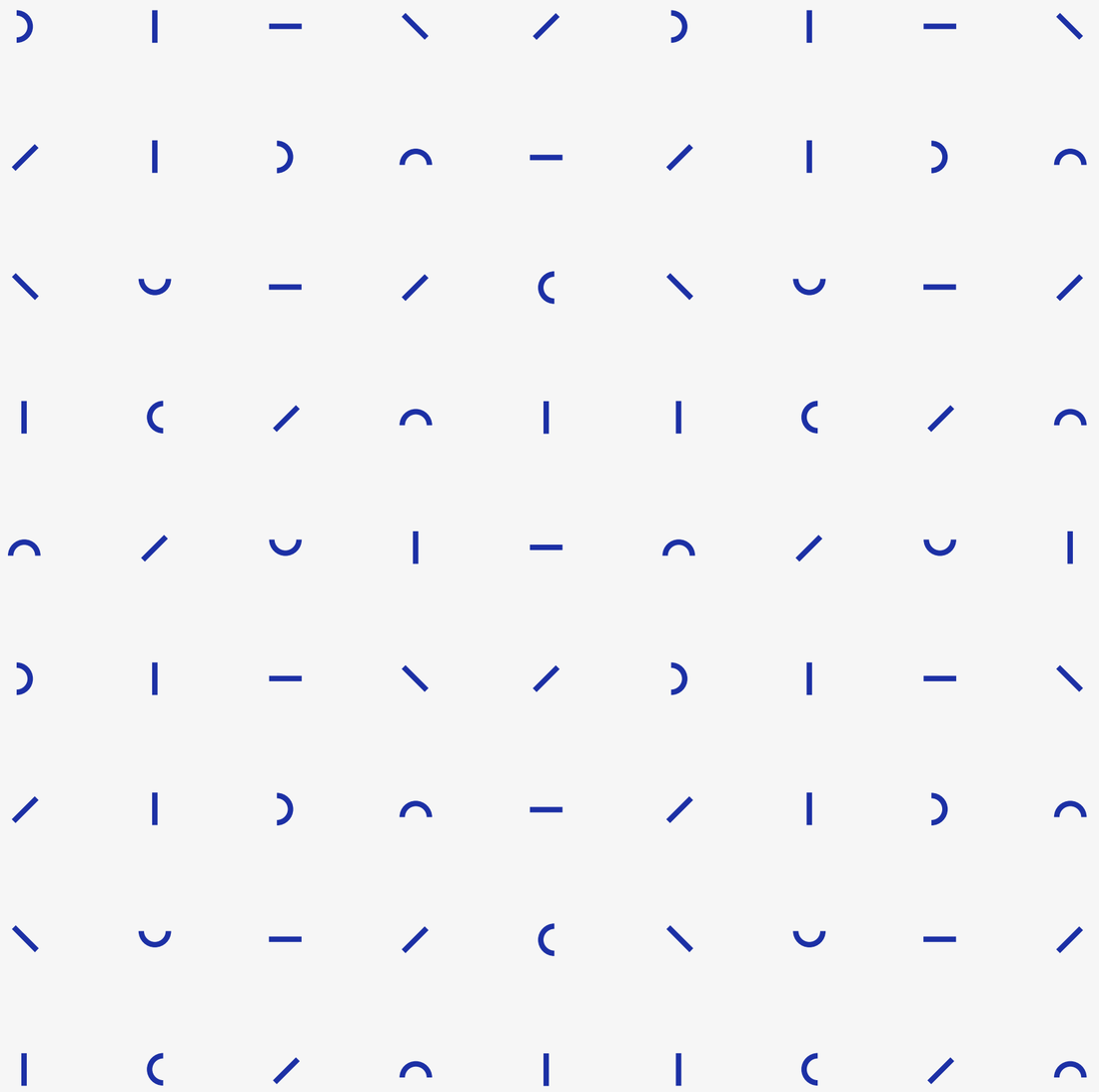
Machine Learning

Bank Data Clustering in WEKA

AURA AKSHA KARISMA



K-Means Clustering in WEKA

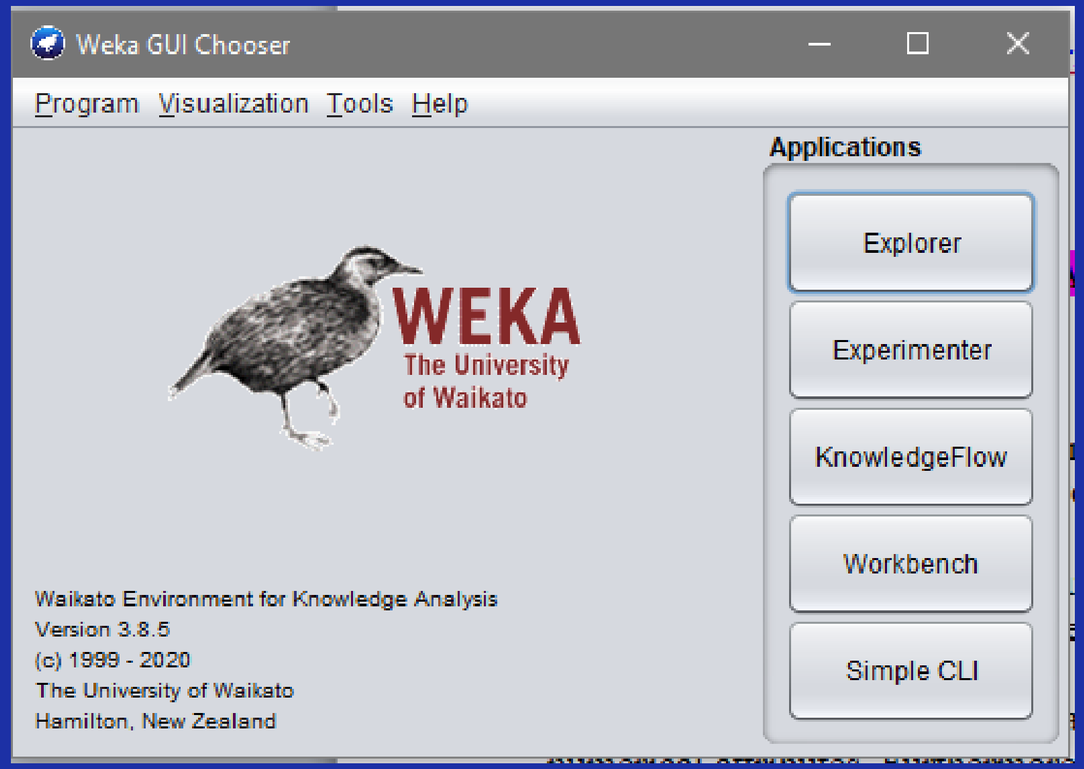
DePaul University used k-means algorithm to cluster bank customer from dataset 'bank-data.csv' to illustrate the implementation of k-means clustering in WEKA for its Web Data Mining class.

In this dataset, there are:
600 instances, 11 attributes.

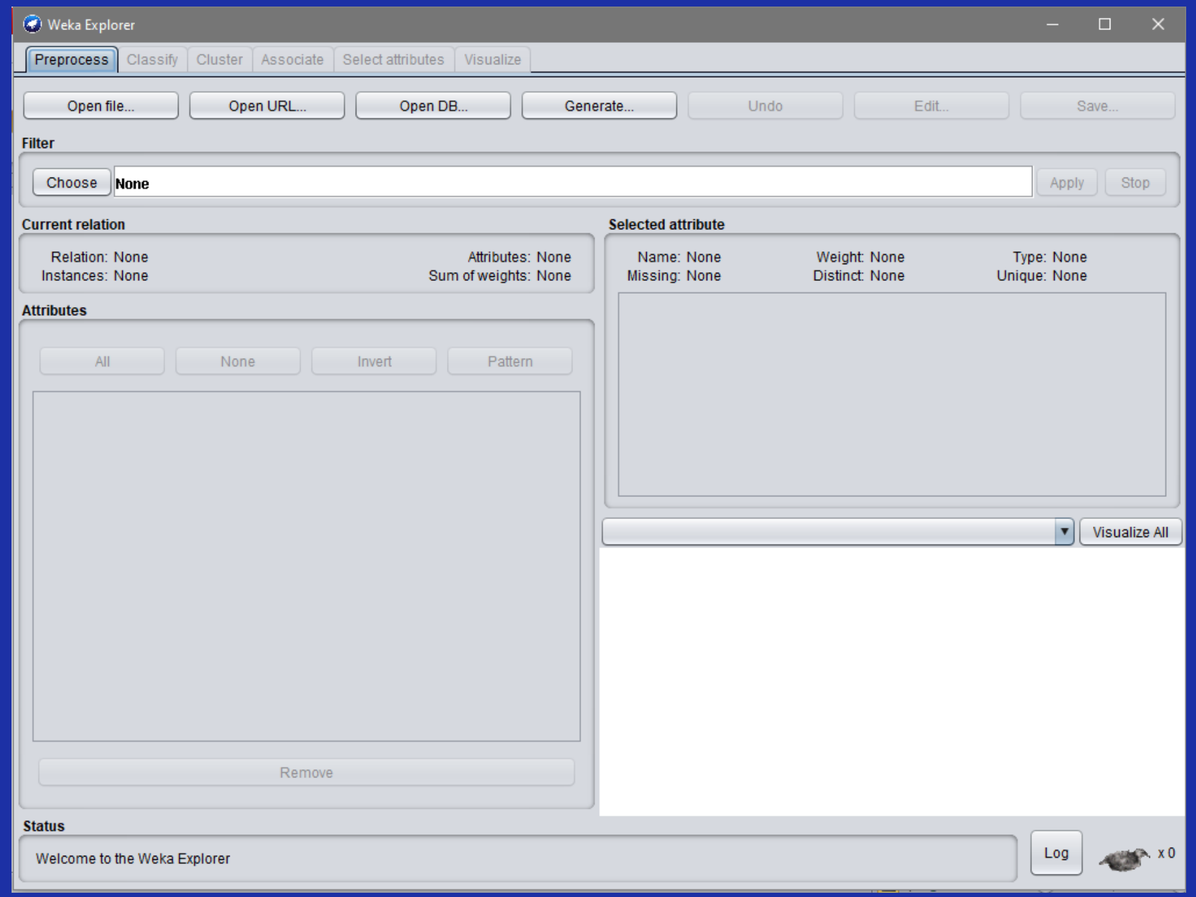
id	a unique identification number
age	age of customer in years (numeric)
sex	MALE / FEMALE
region	inner_city/rural/suburban/town
income	income of customer (numeric)
married	is the customer married (YES/NO)
children	number of children (numeric)
car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
mortgage	does the customer have a mortgage (YES/NO)
pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)

bank-data.csv attributes

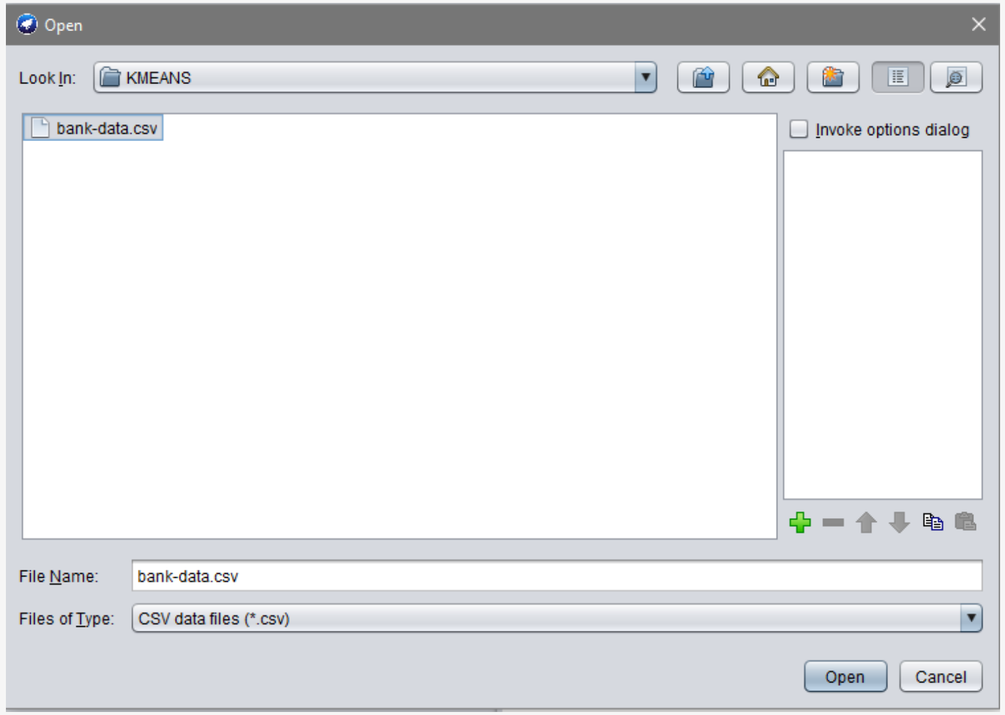
1 Load the Data



1. Open WEKA
2. Choose 'Explorer'



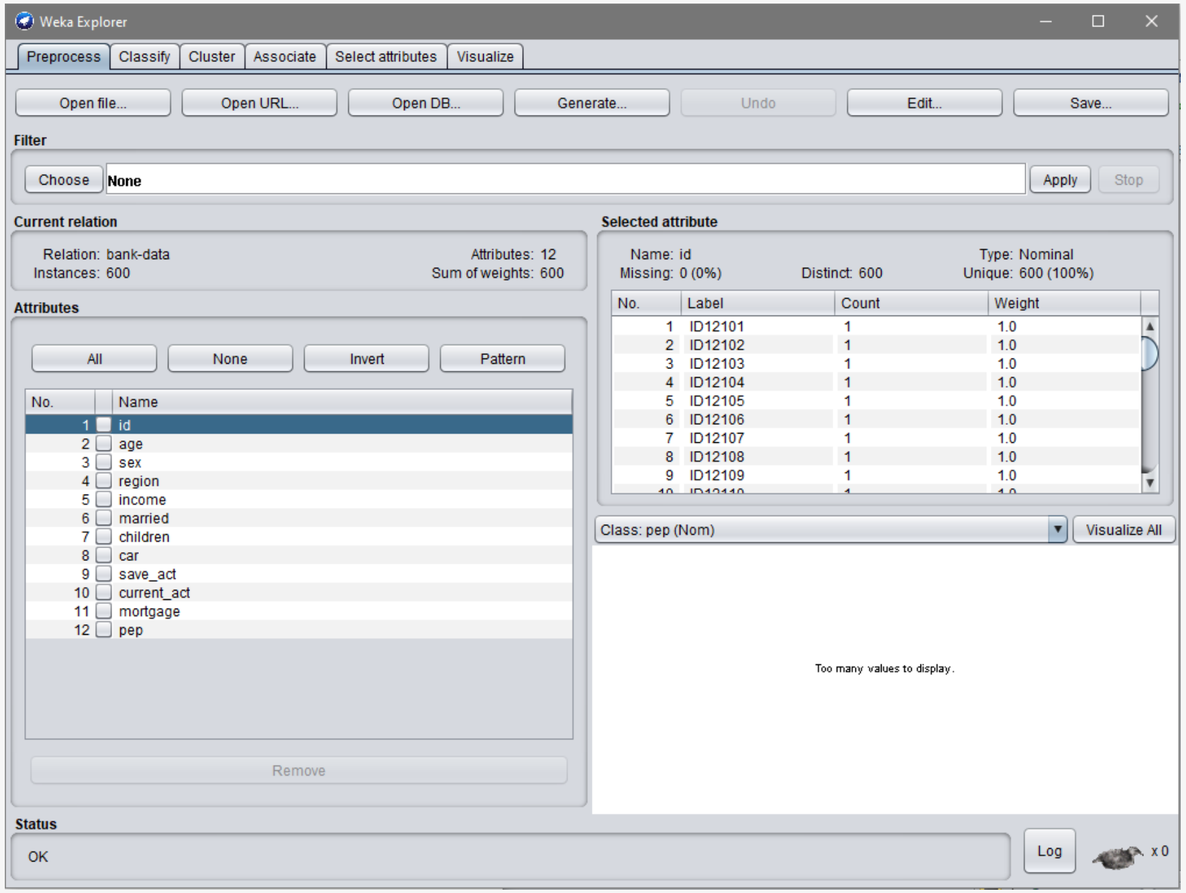
1. Open File



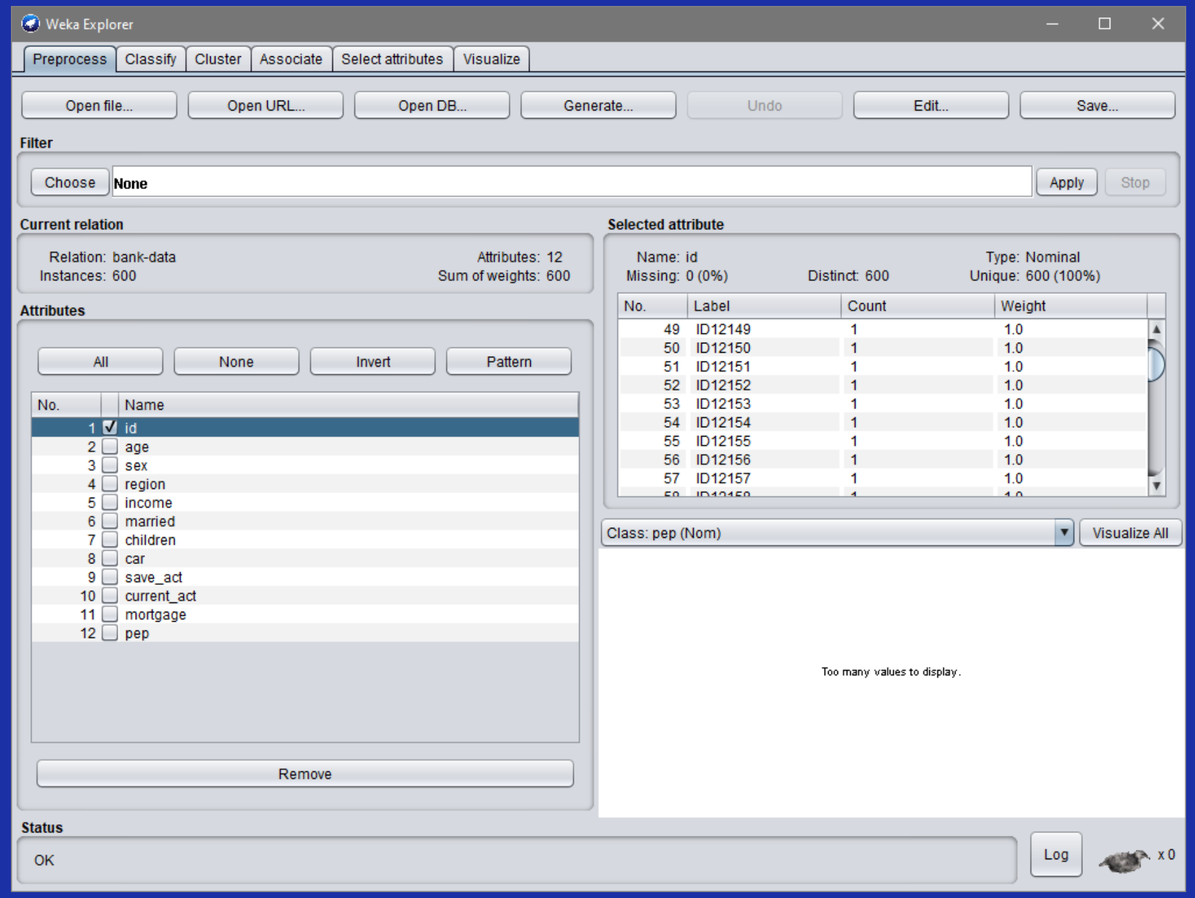
A pop-up will appear,

1. Go to the directory with the bank data
2. Open

The display once bank-data.csv is loaded:



1 Preprocessing



In this case, the author assumed that the data has been preprocessed.

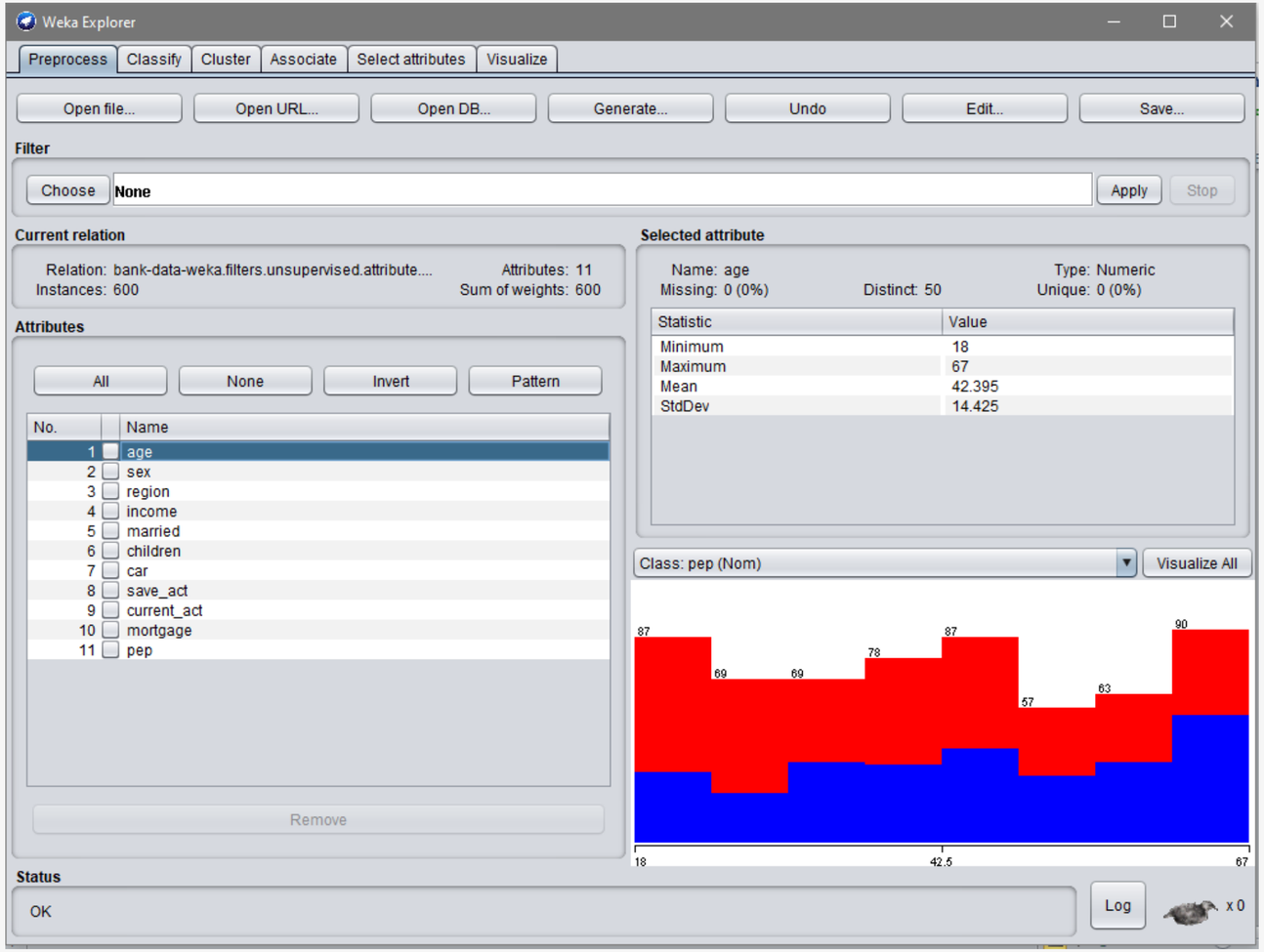
1. The ID field has been removed.
2. The "children" attribute has been converted to categorical

>> To do the first step:

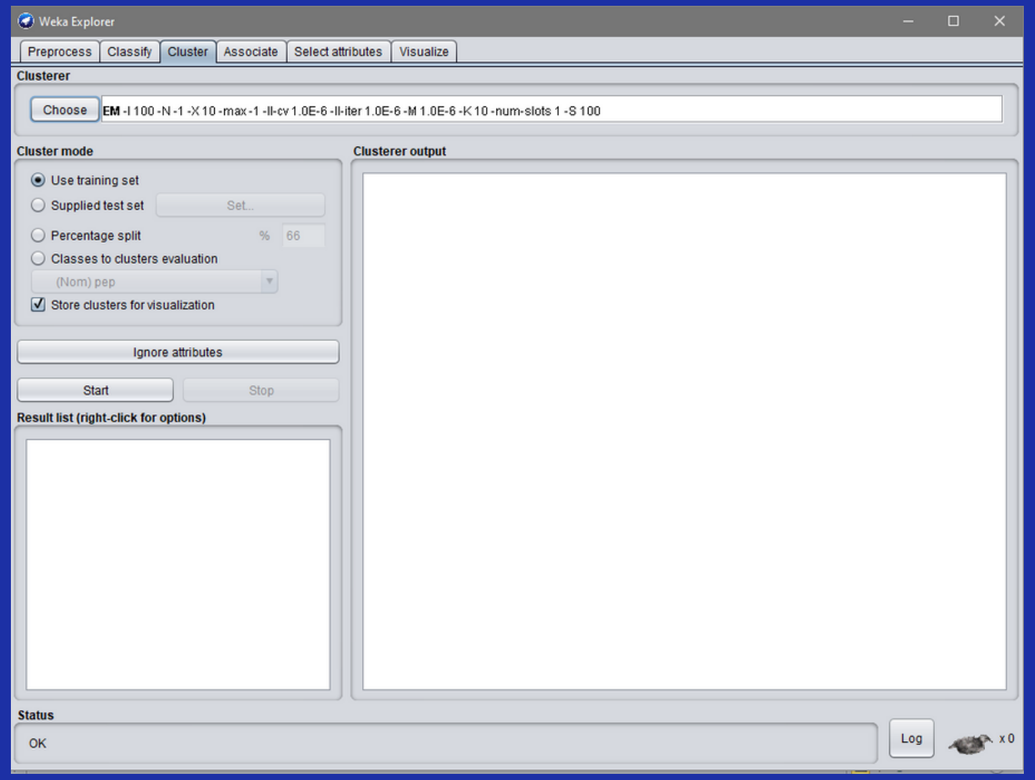
1. Select the 'id' attribute.
2. Click 'Remove'

>> The second step is skipped in our case since it is stated to be unnecessary for clustering.

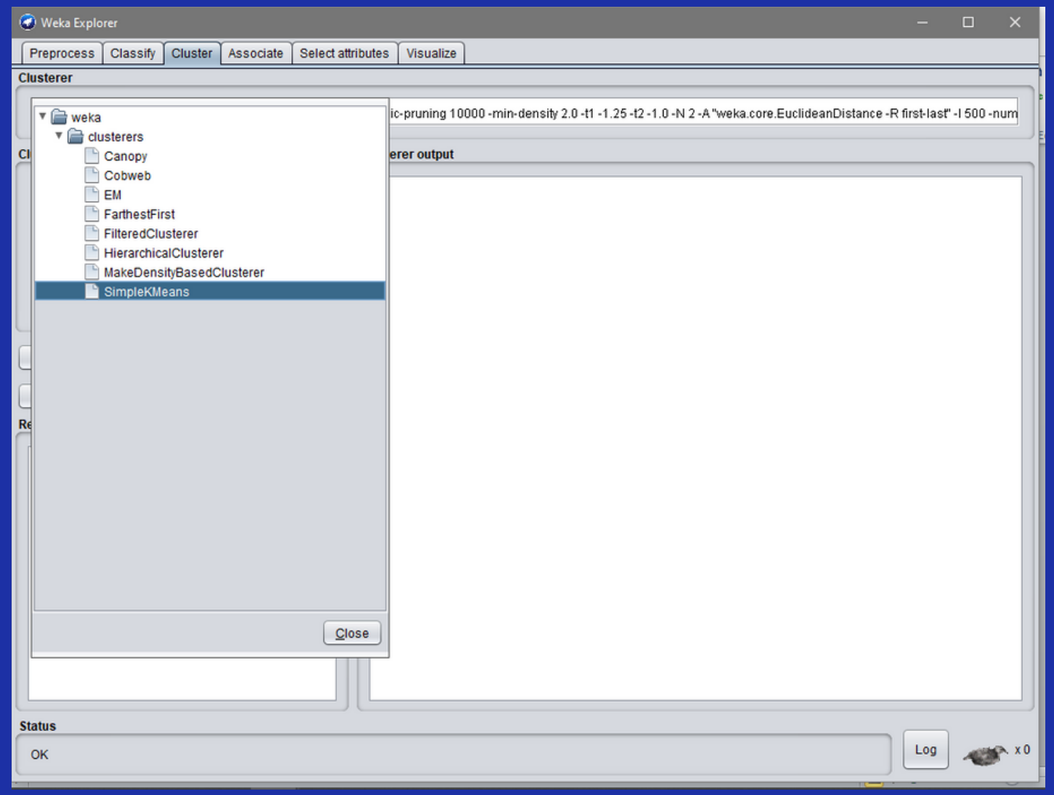
On the right is the display after the 'id' attribute is removed.



3 Cluster

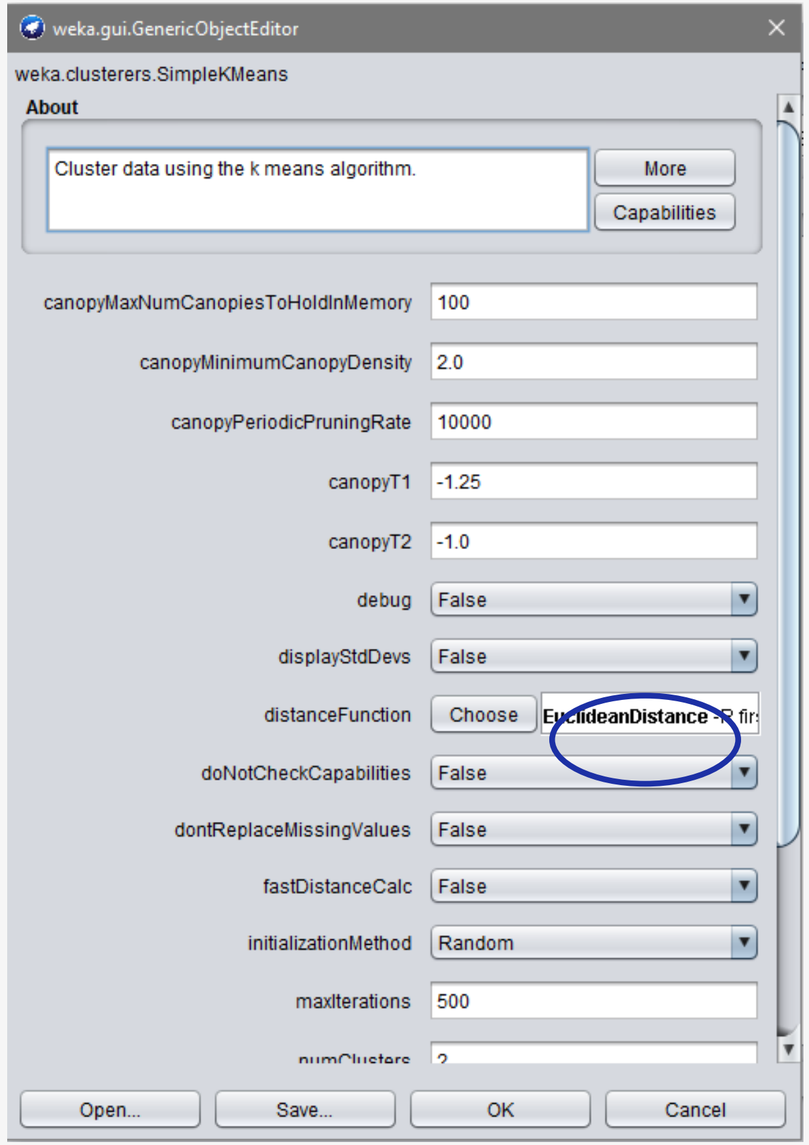


1. Go to 'Cluster' tab
2. Click 'Choose'



3. Select 'SimpleKMeans'

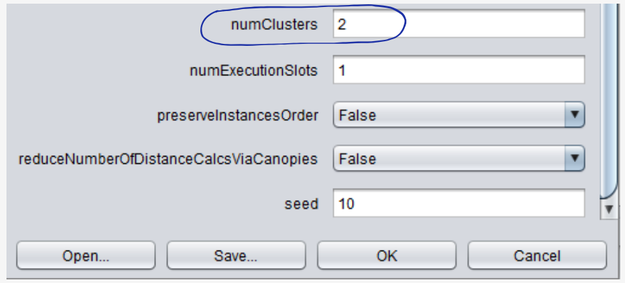
WEKA SimpleKMeans algorithm automatically handles categorical and numerical attributes. It also automatically normalizes numerical attributes.

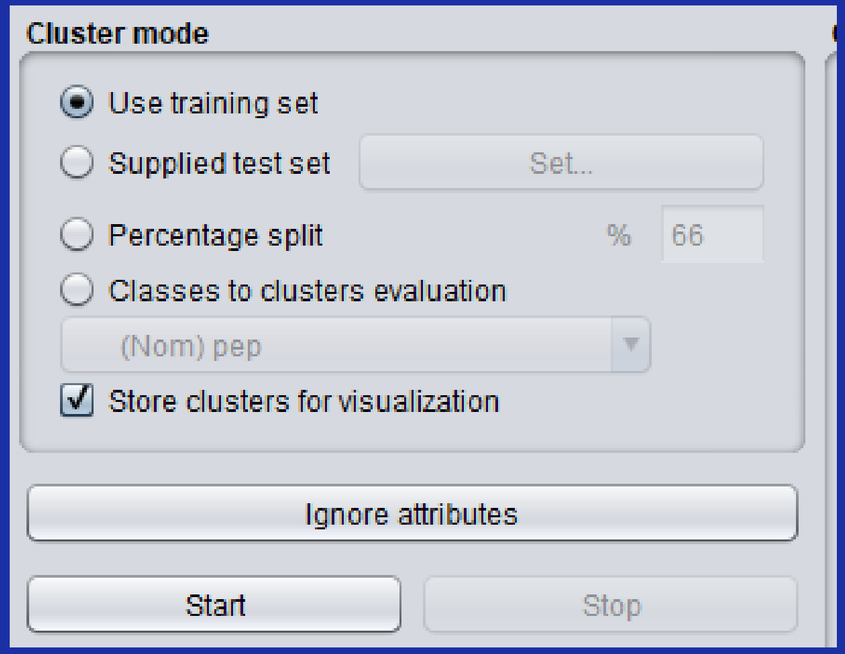


Click on the text box to the right of the "Choose" button to edit clustering parameter. A pop-up (on the right) will appear.

It can be seen that the WEKA SimpleKMeans algorithm uses Euclidean distance measure to compute distances between instances and clusters.

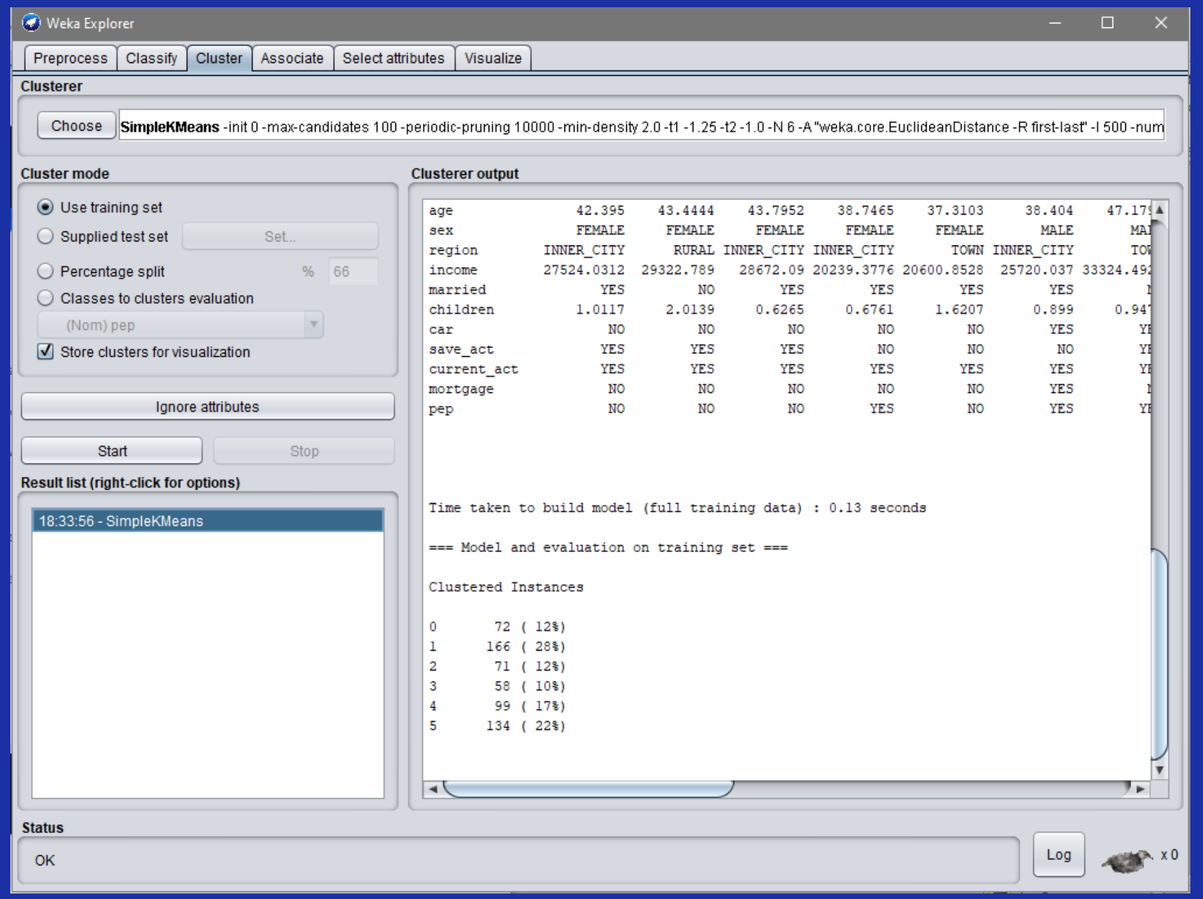
Then scroll down until you see 'numClusters'. Change it from 2 to 6. The 'seed' value is used in generating a random number for making the initial assignment of instances to clusters.

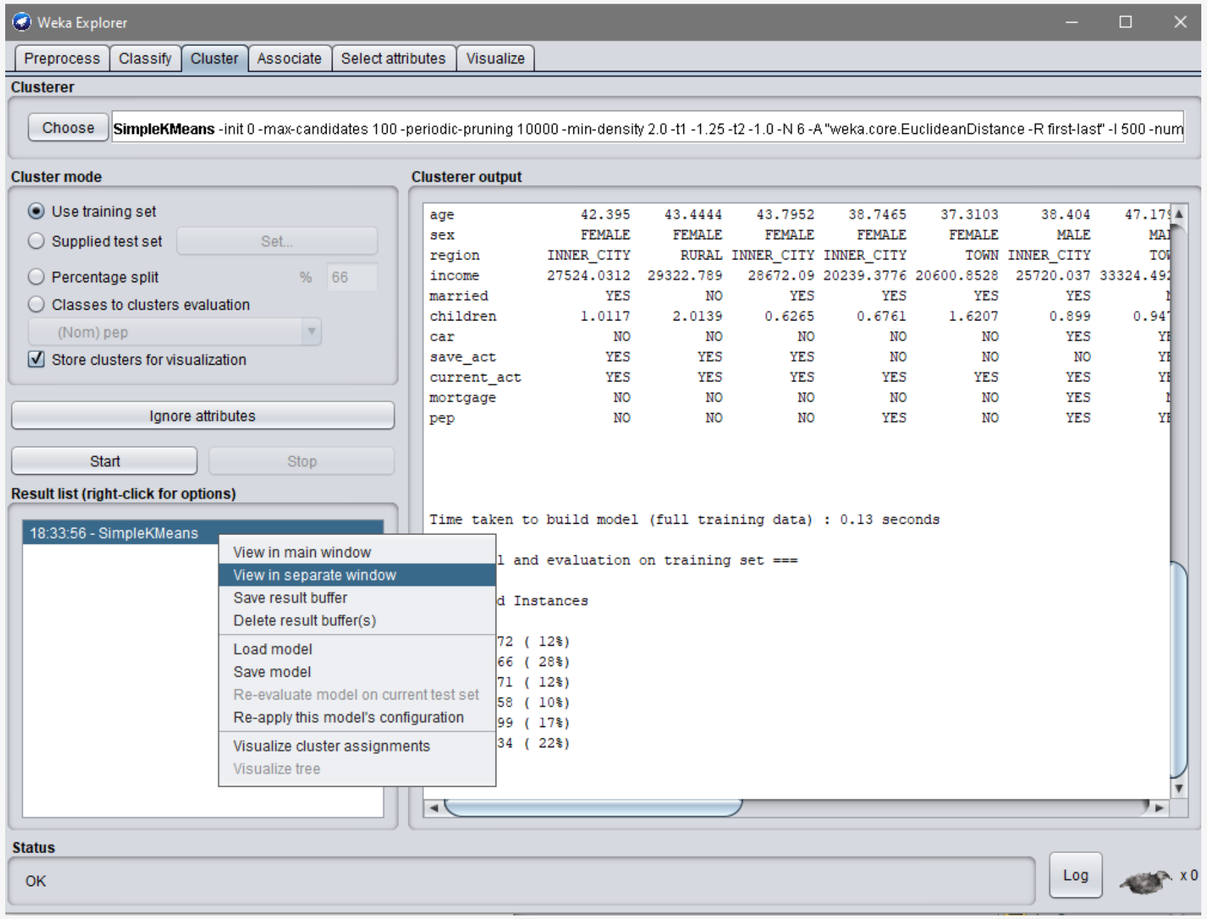




1. In the Cluster mode, make sure 'Use training set' is selected.
2. Start

The display after the clustering is started.

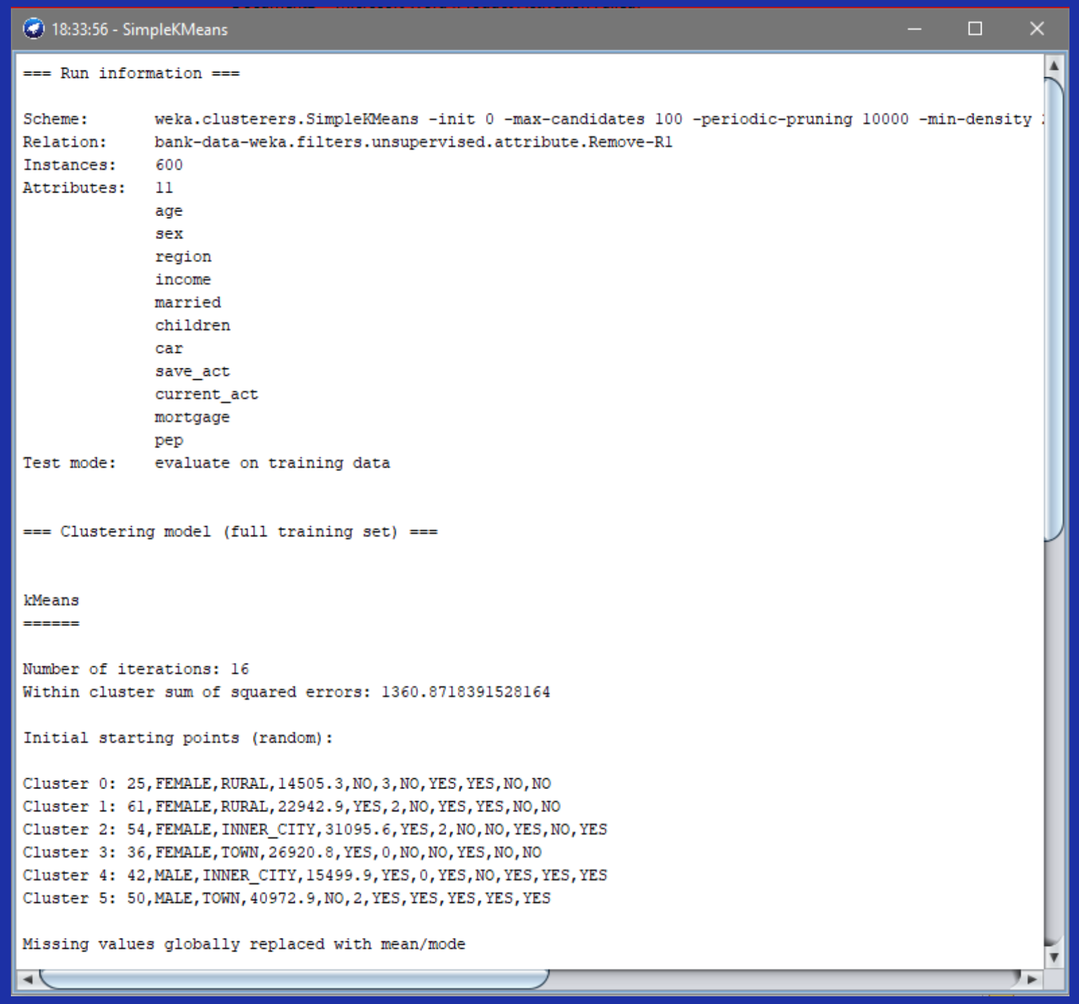




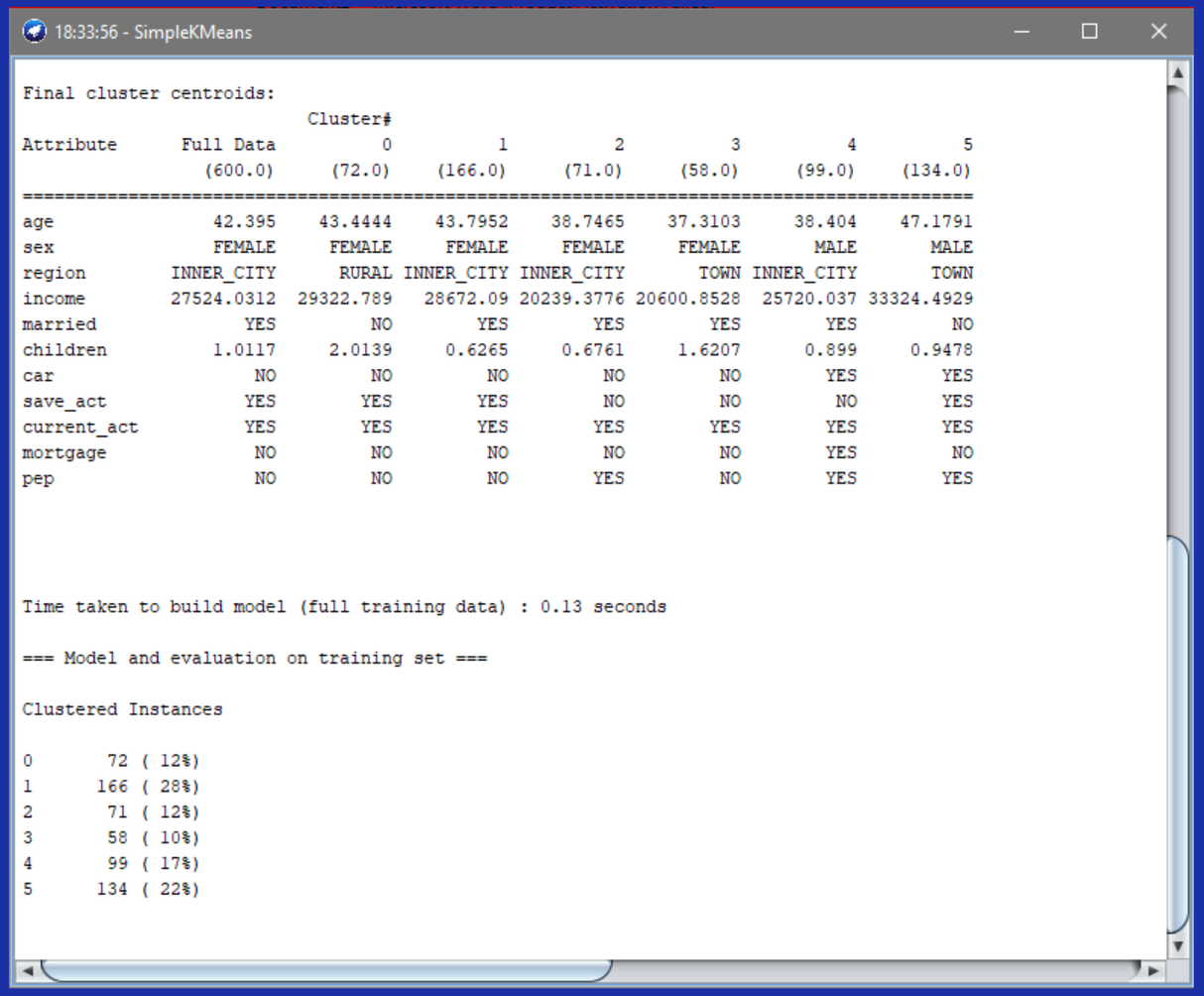
1. Right click the result set in the "Result list" panel.
2. Select "View in separate window".

The Result

It shows the centroid (value representing the clusters) of each cluster as well as statistics.



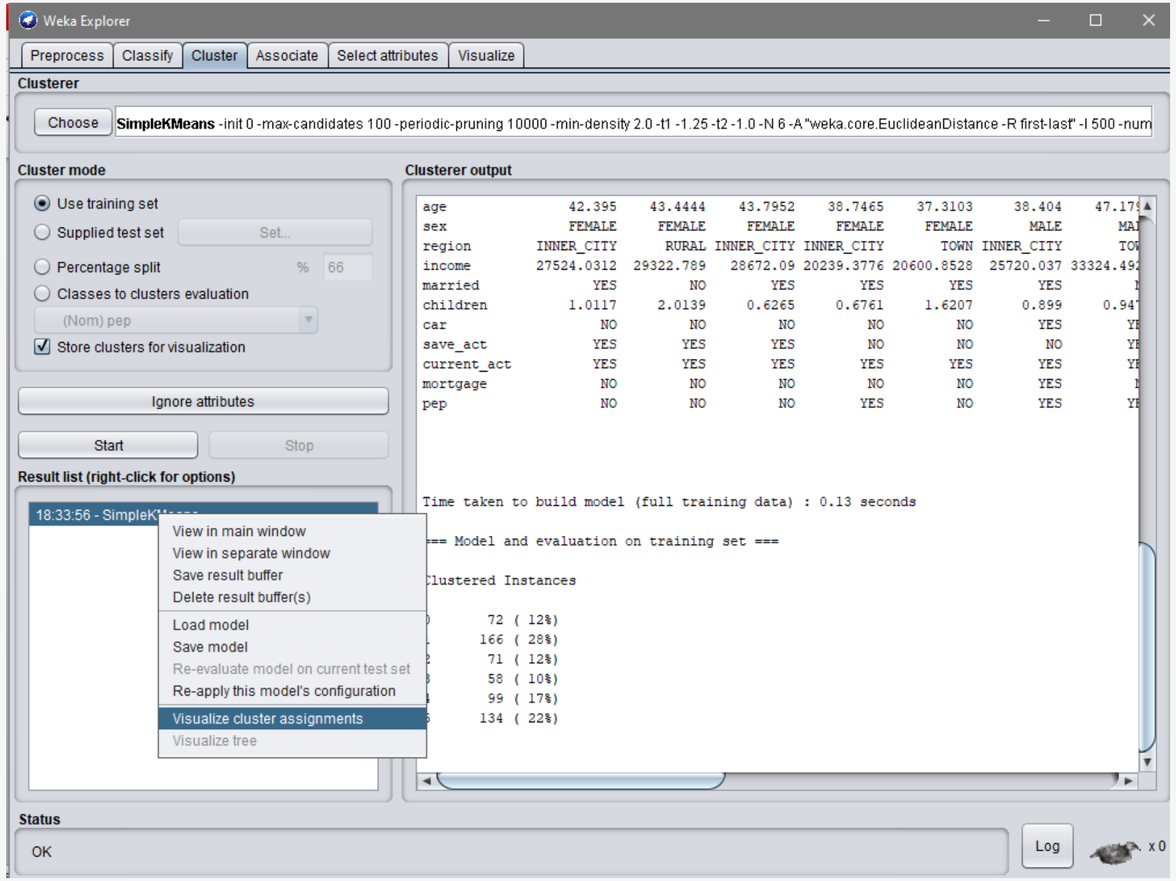
```
18:33:56 - SimpleKMeans

=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density
Relation:     bank-data-weka.filters.unsupervised.attribute.Remove-R1
Instances:    600
Attributes:   11
              age
              sex
              region
              income
              married
              children
              car
              save_act
              current_act
              mortgage
              pep
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 16
Within cluster sum of squared errors: 1360.8718391528164

Initial starting points (random):

Cluster 0: 25,FEMALE,RURAL,14505.3,NO,3,NO,YES,YES,NO,NO
Cluster 1: 61,FEMALE,RURAL,22942.9,YES,2,NO,YES,YES,NO,NO
Cluster 2: 54,FEMALE,INNER_CITY,31095.6,YES,2,NO,NO,YES,NO,YES
Cluster 3: 36,FEMALE,TOWN,26920.8,YES,0,NO,NO,YES,NO,NO
Cluster 4: 42,MALE,INNER_CITY,15499.9,YES,0,YES,NO,YES,YES,YES
Cluster 5: 50,MALE,TOWN,40972.9,NO,2,YES,YES,YES,YES,YES

Missing values globally replaced with mean/mode
```



```
18:33:56 - SimpleKMeans

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2          3          4          5
                (600.0)     (72.0)    (166.0)     (71.0)     (58.0)     (99.0)    (134.0)
=============================================================================================
age               42.395    43.4444    43.7952    38.7465    37.3103     38.404    47.1791
sex               FEMALE     FEMALE     FEMALE     FEMALE     FEMALE       MALE       MALE
region        INNER_CITY      RURAL INNER_CITY INNER_CITY       TOWN INNER_CITY       TOWN
income        27524.0312  29322.789   28672.09 20239.3776 20600.8528  25720.037 33324.4929
married              YES         NO        YES        YES        YES        YES         NO
children          1.0117     2.0139     0.6265     0.6761     1.6207      0.899     0.9478
car                  NO         NO         NO         NO         NO        YES        YES
save_act             YES        YES        YES         NO         NO         NO        YES
current_act          YES        YES        YES        YES        YES        YES        YES
mortgage              NO         NO         NO         NO         NO        YES         NO
pep                   NO         NO         NO        YES         NO        YES        YES



Time taken to build model (full training data) : 0.13 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       72 ( 12%)
1      166 ( 28%)
2       71 ( 12%)
3       58 ( 10%)
4       99 ( 17%)
5      134 ( 22%)
```

4 Visualization

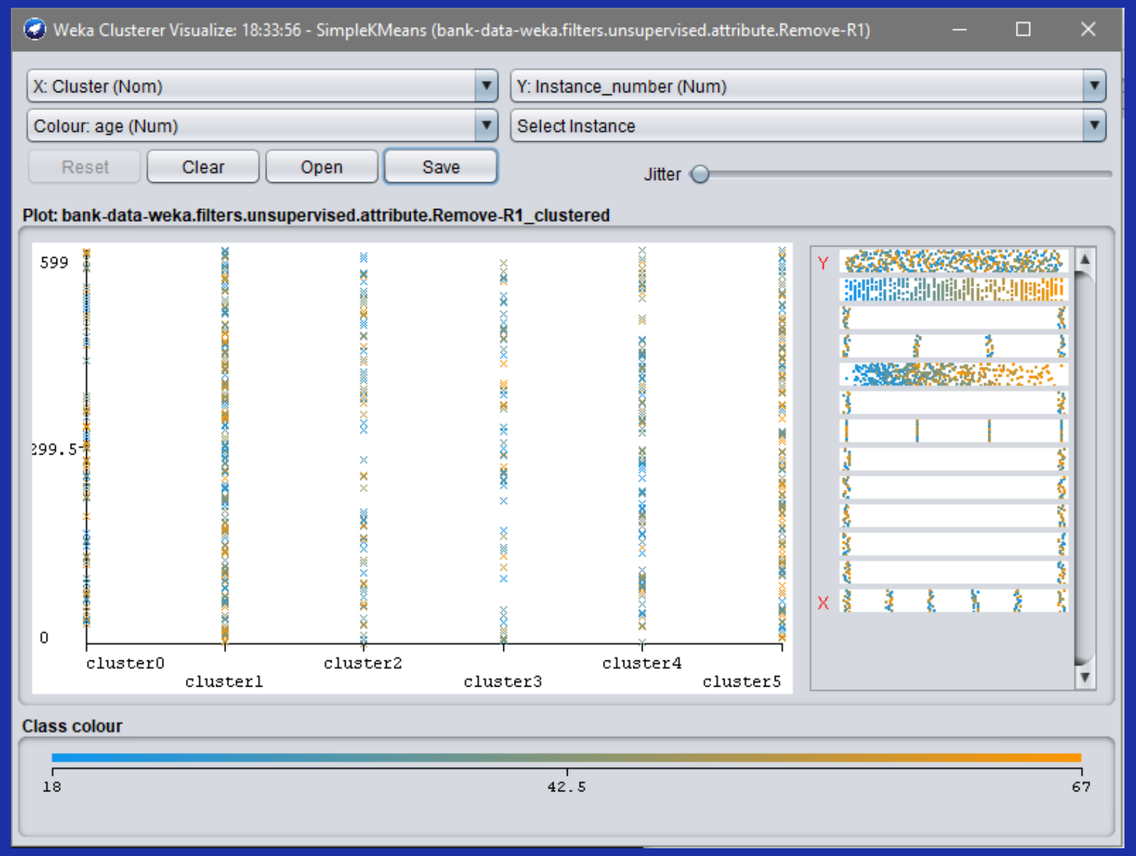
Another way to understand the characteristics of each cluster.



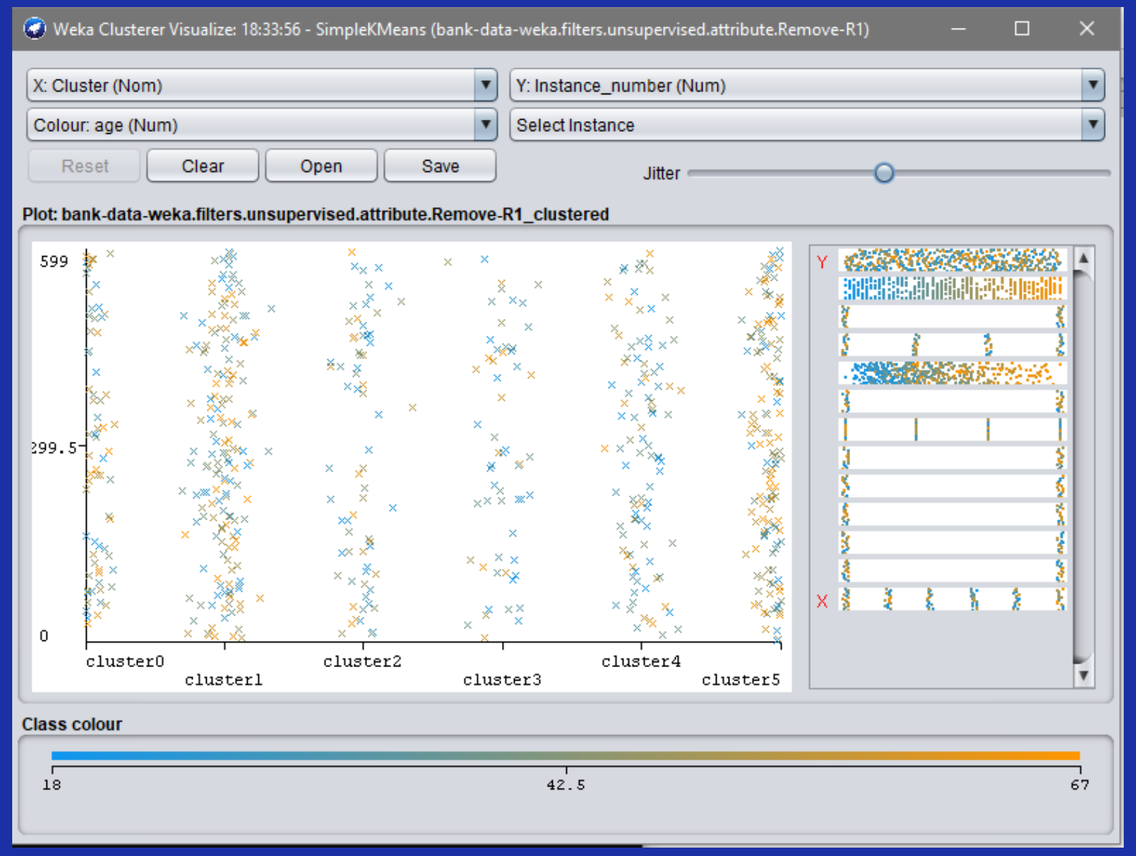
1. Right click the result set in the "Result list" panel.
2. Select "Visualize cluster assignments".

You can choose the cluster number and any of the other attributes for each of the three different dimensions available (x-axis, y-axis, and color) to see different relationships within clusters.

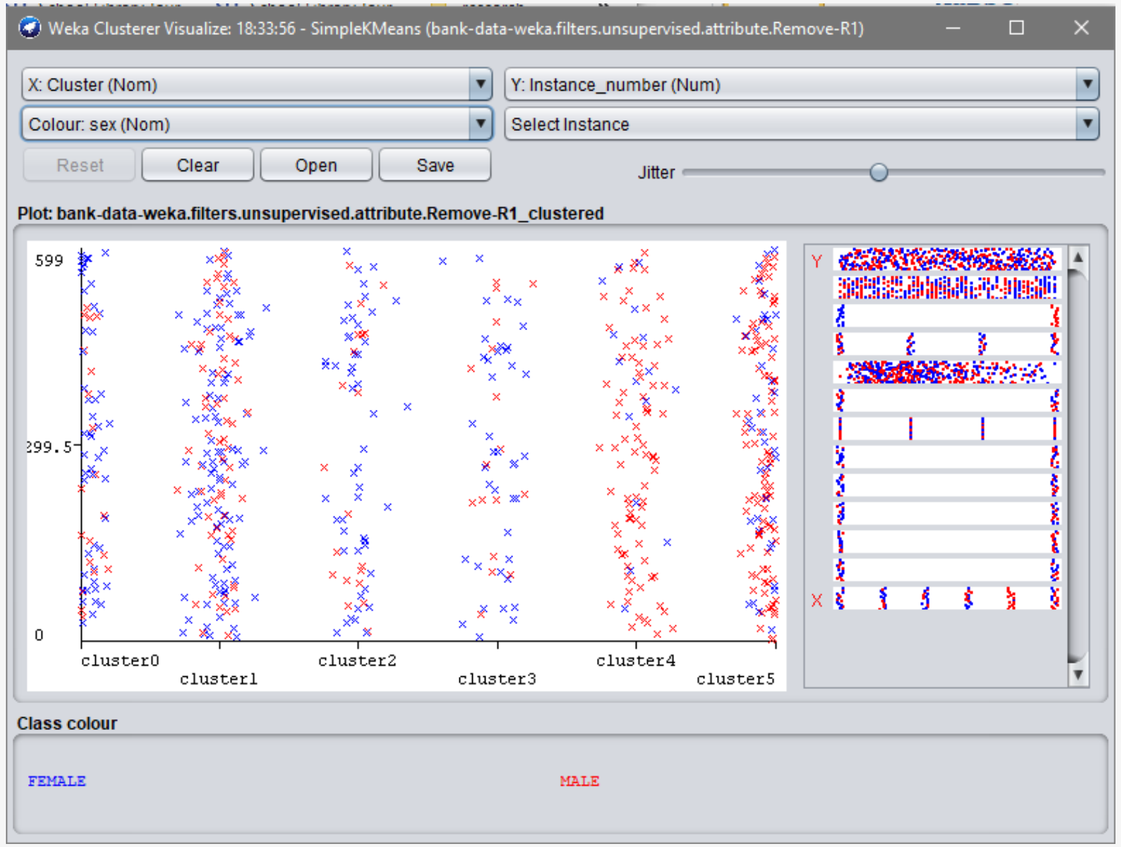
1.a. Age within each clusters



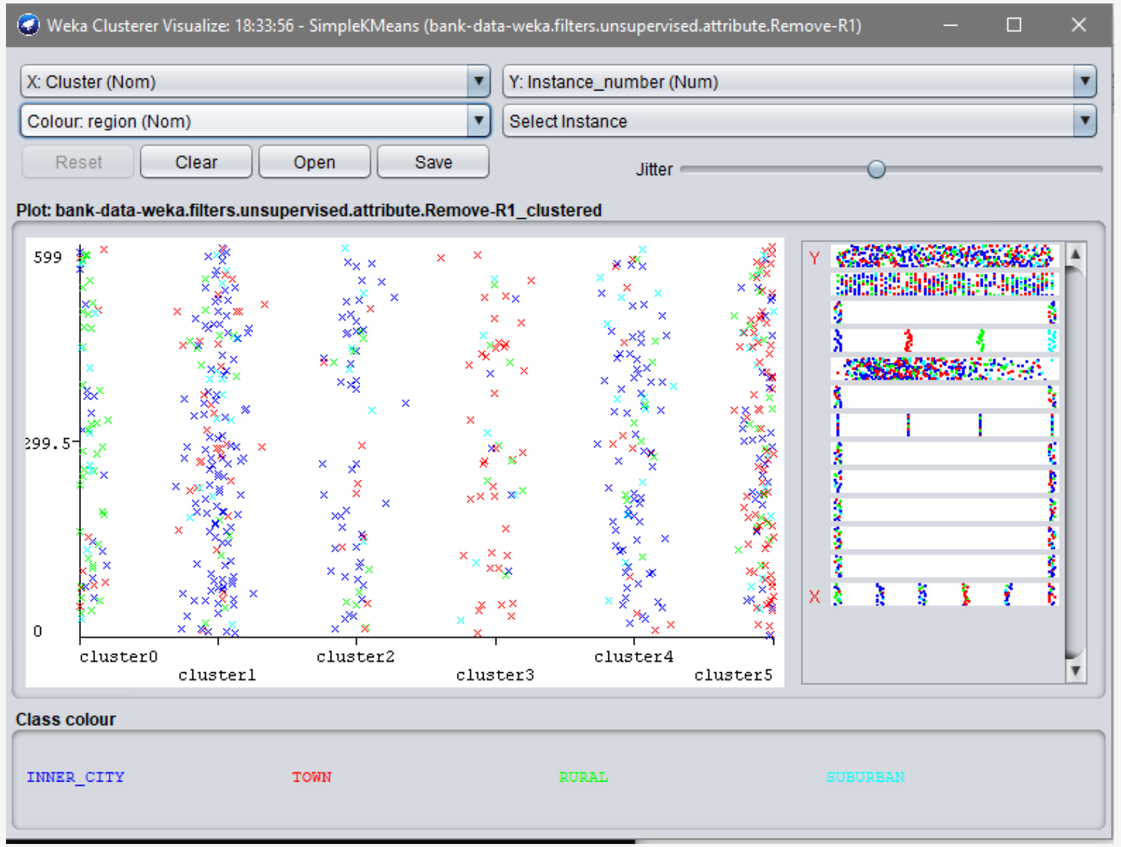
1.b. Age within each clusters with jitters



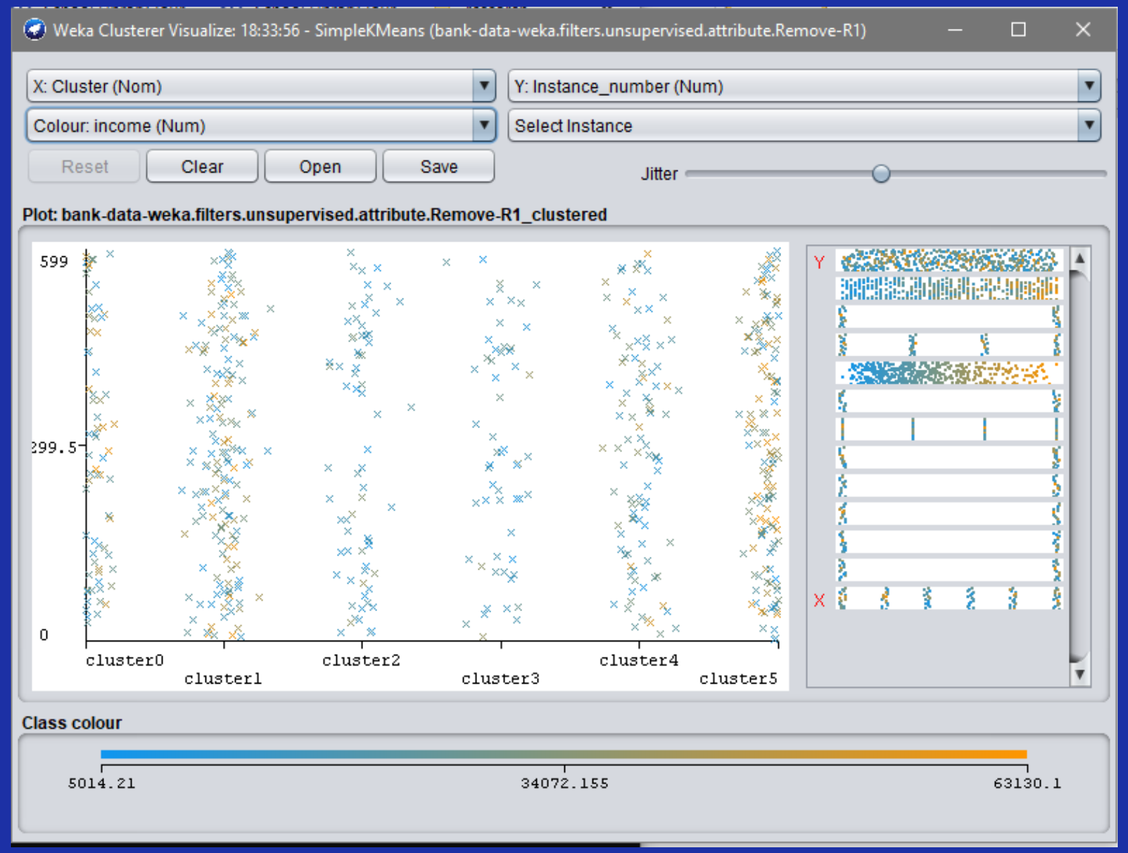
2. Sex attribute within the clusters



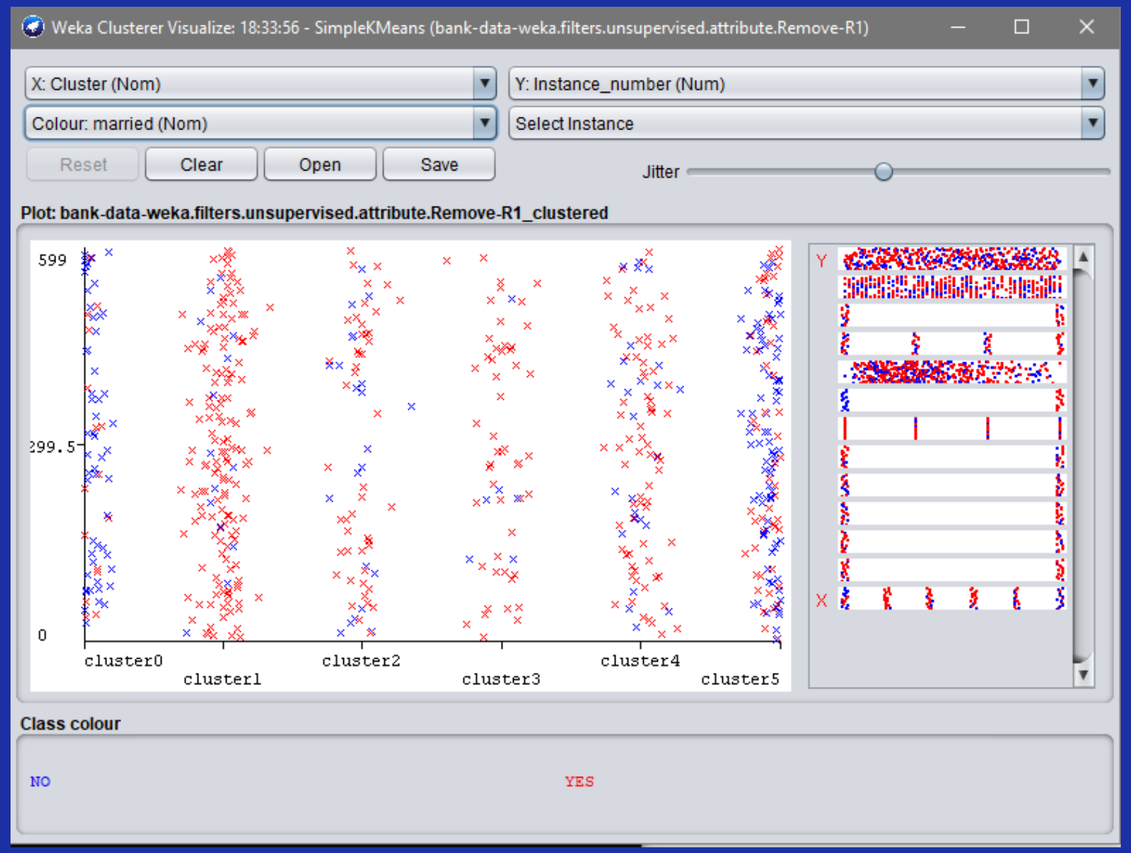
3. Region attribute within the clusters



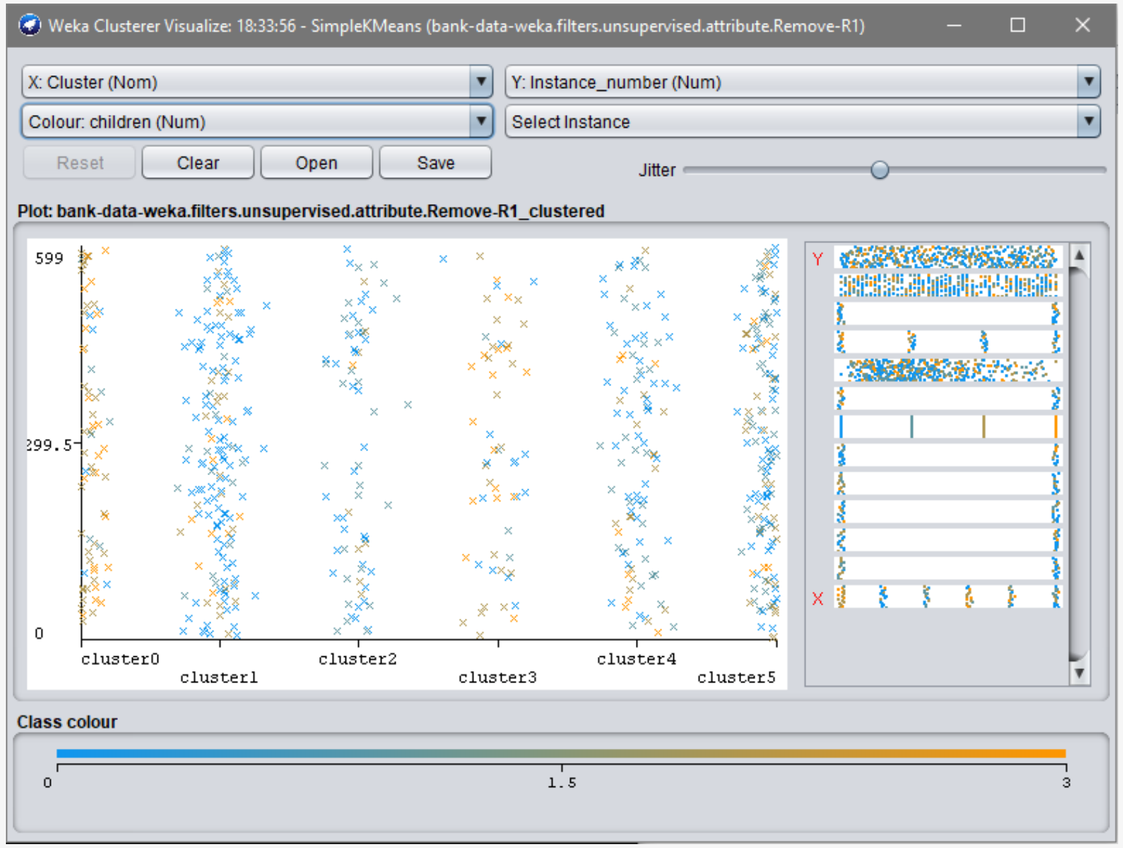
4. Income attribute within clusters



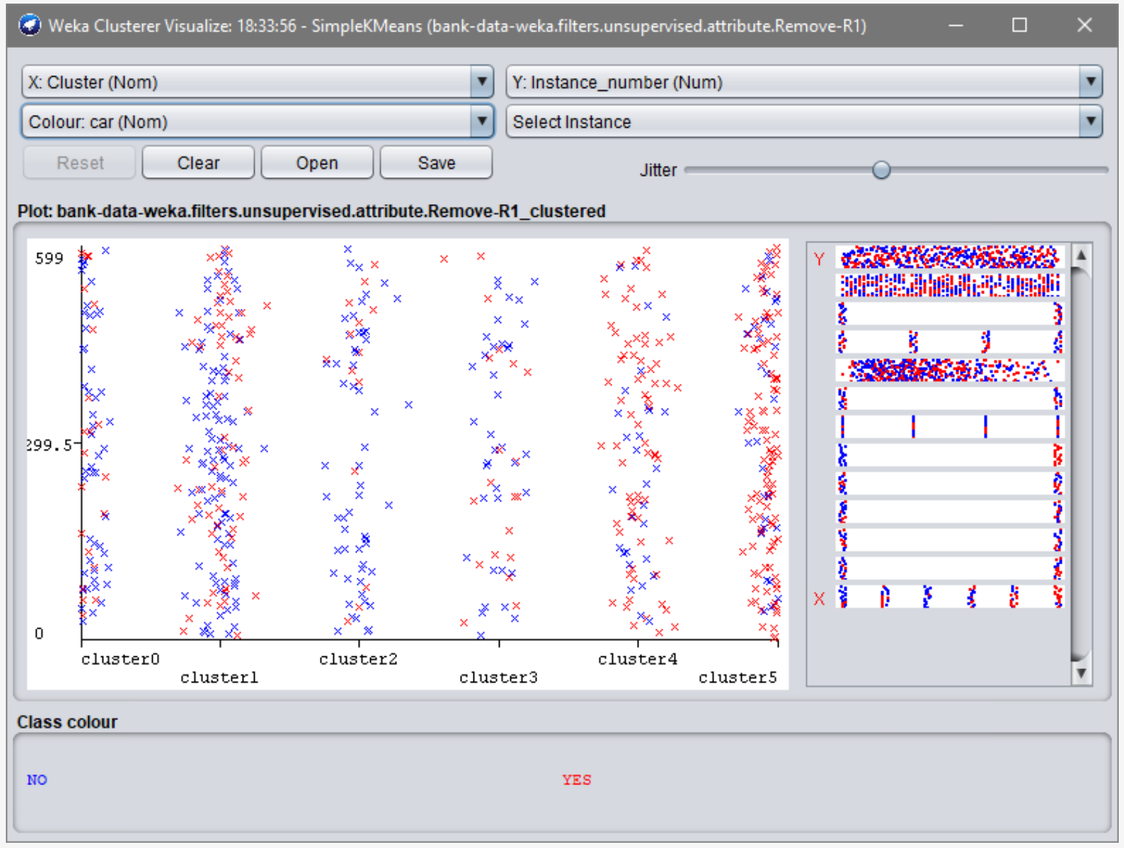
5. Married attribute within clusters



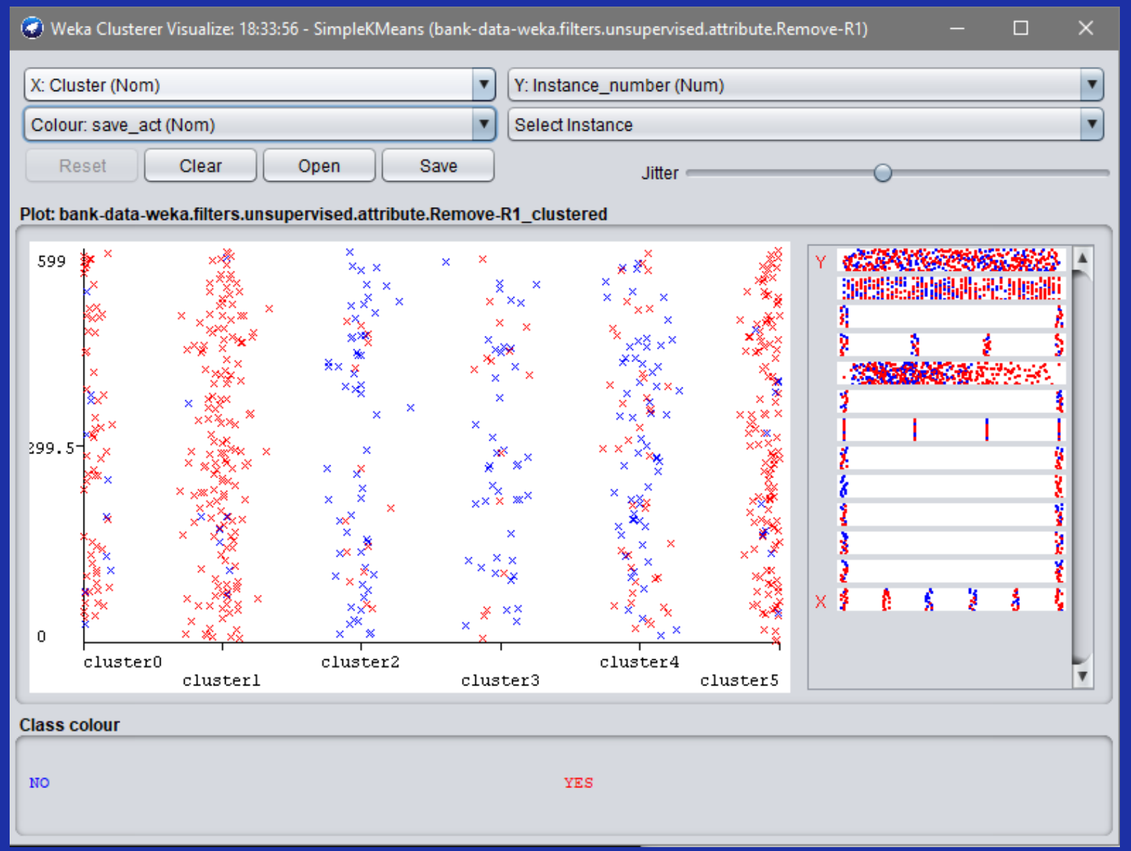
6. Children attribute within the clusters



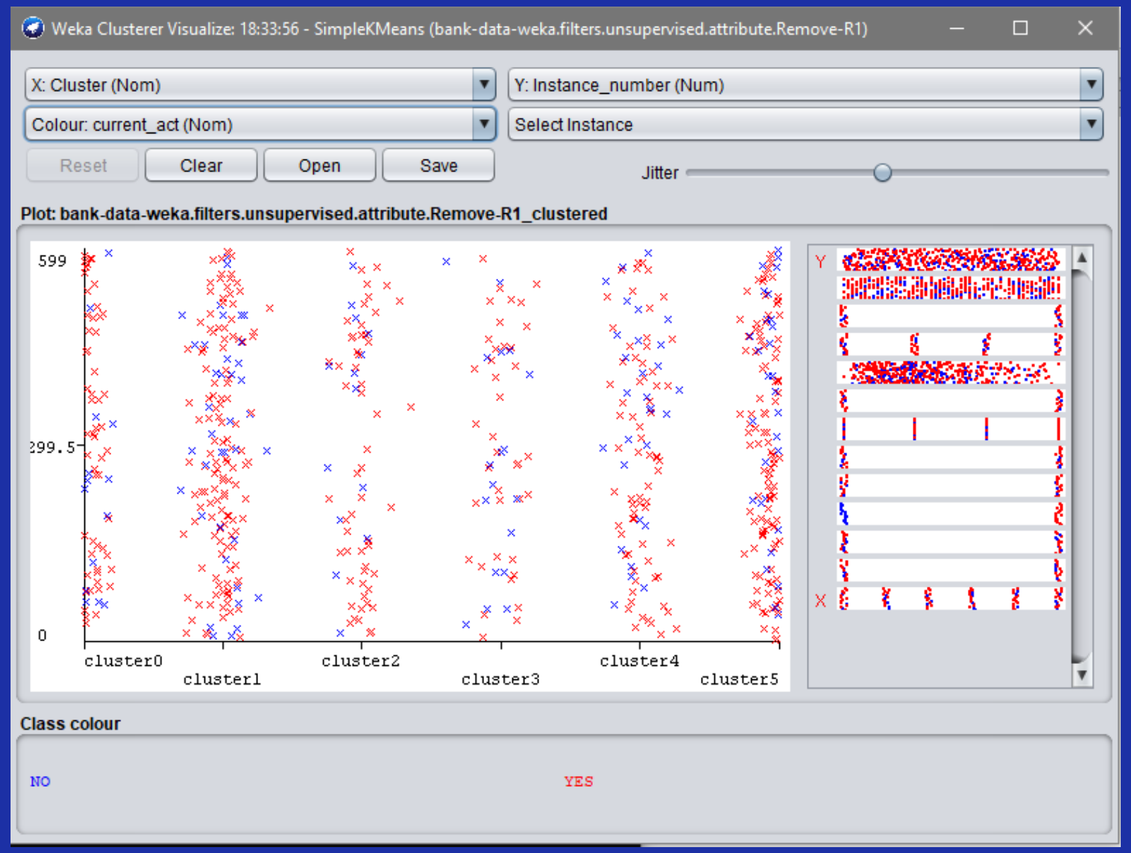
7. Car attribute within the clusters



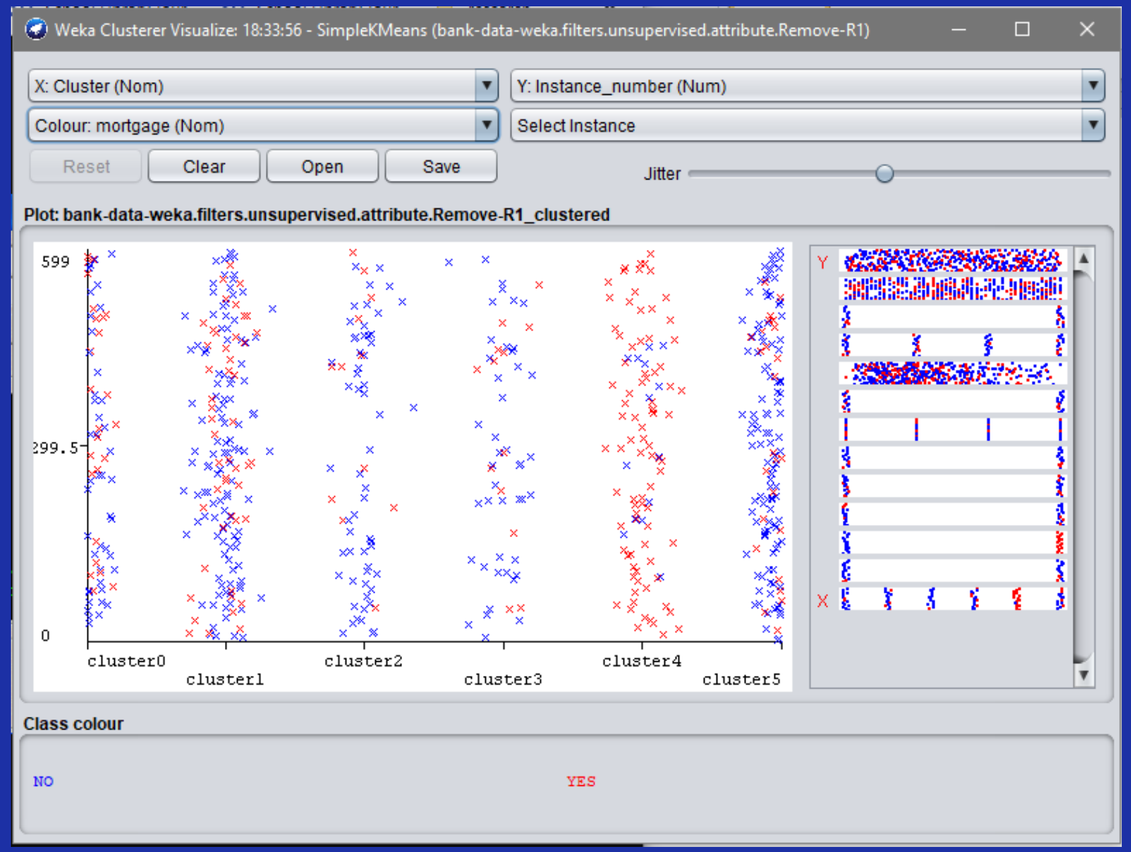
8. Save act attribute within the clusters



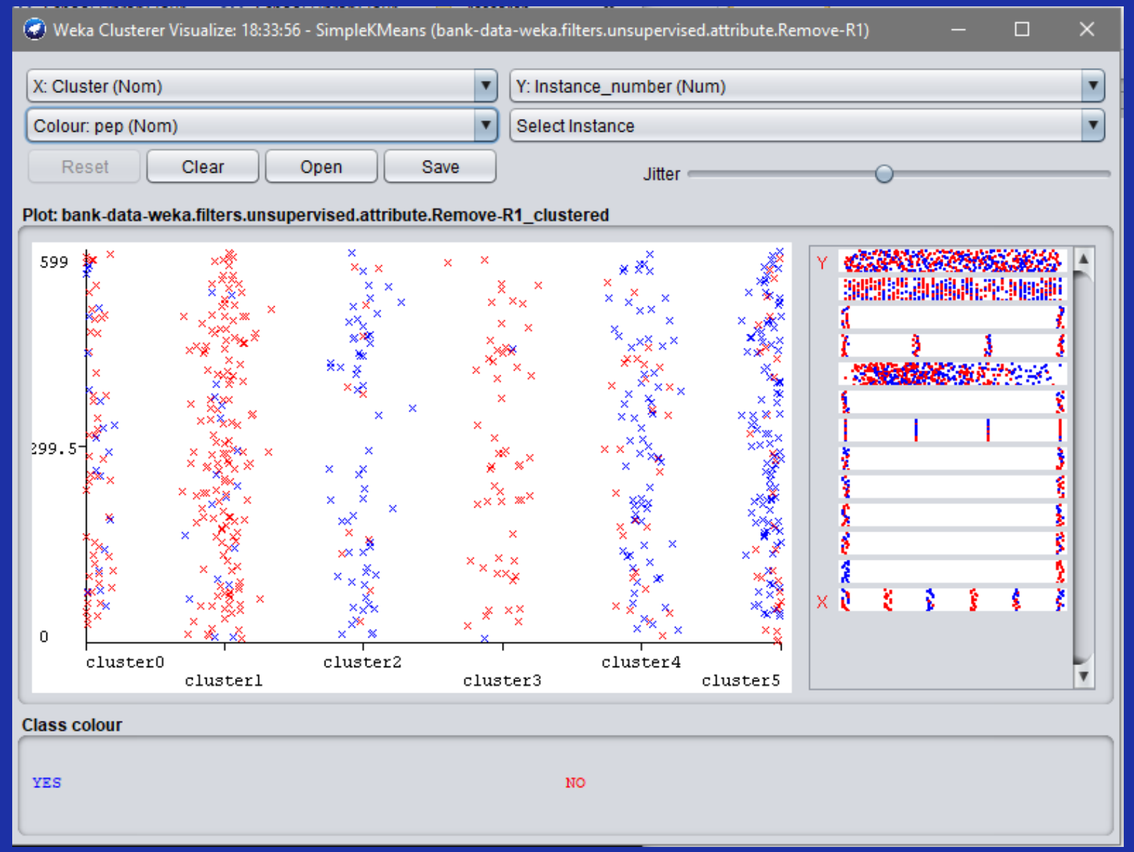
9. Current act attribute within the clusters



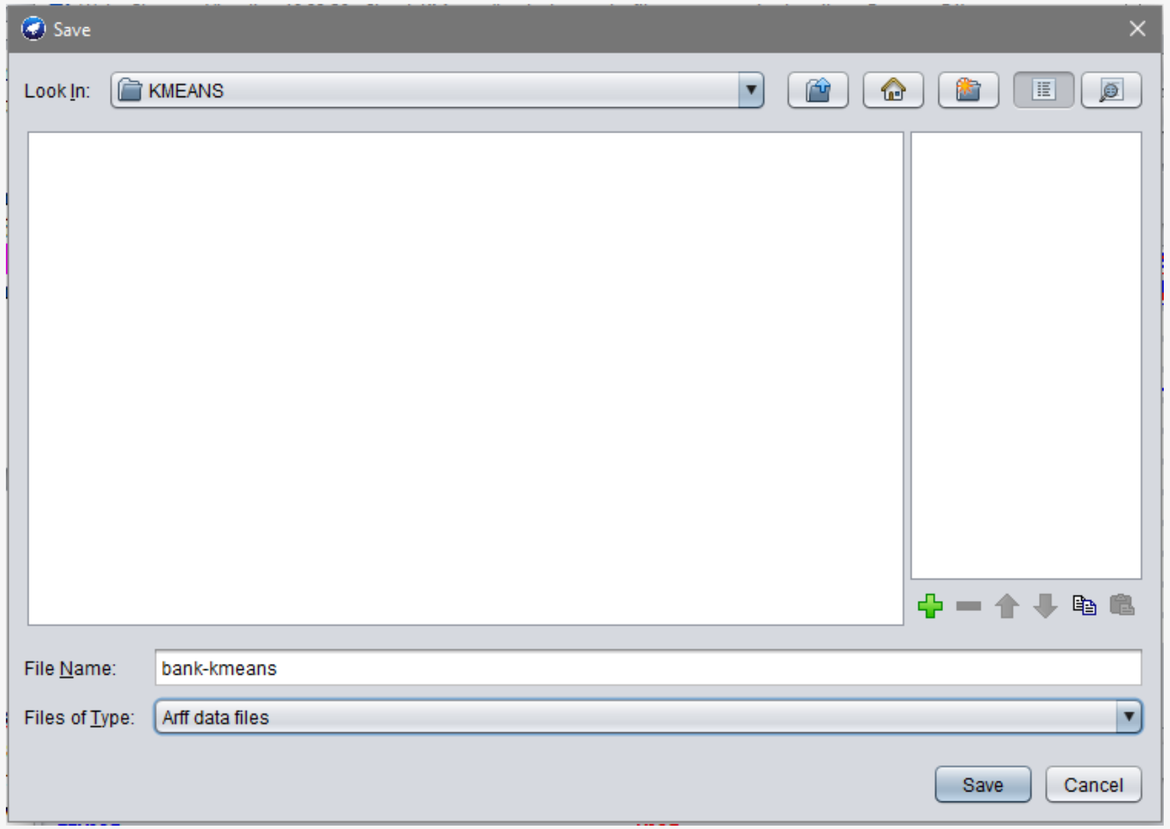
10. Mortgage attribute within the clusters



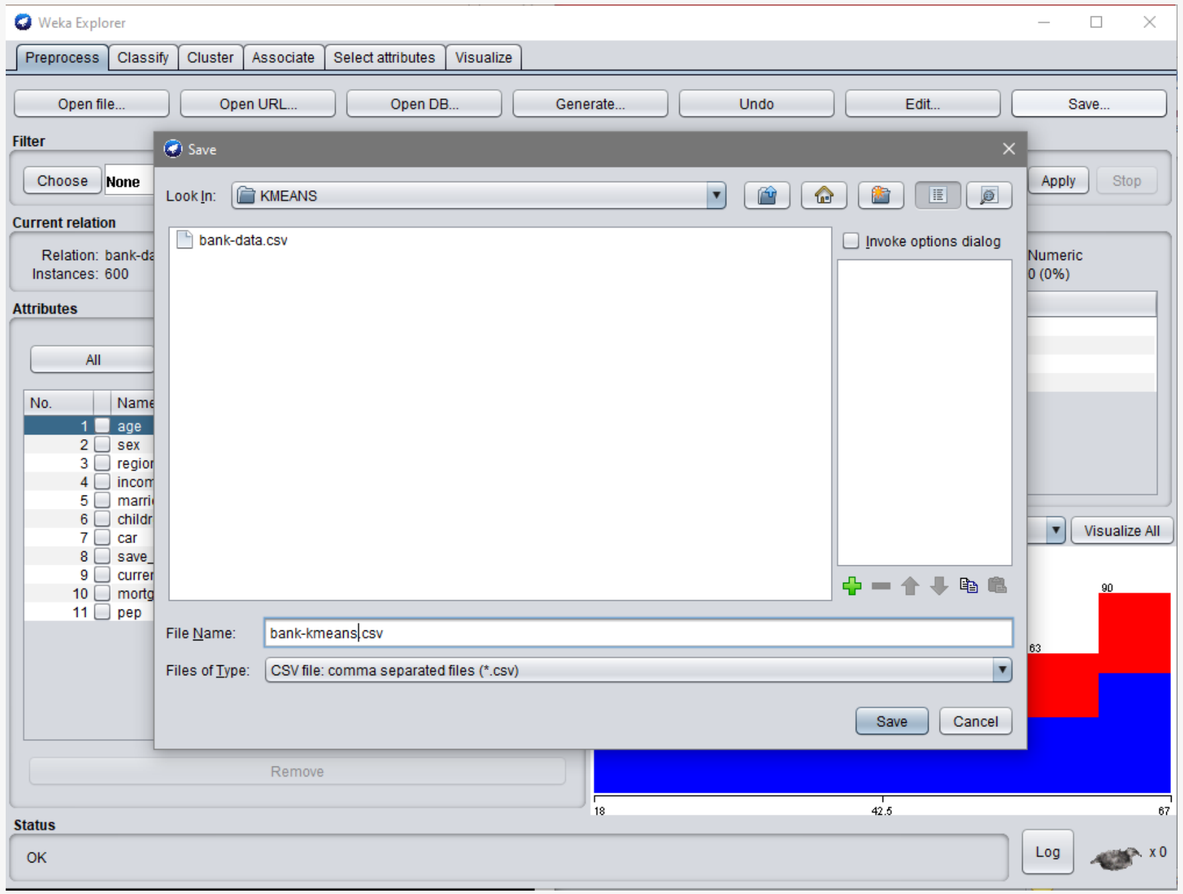
11. PEP attribute within the clusters



5 Save



1. Click the "Save" button in the visualization window.
2. Save the result as "bank-kmeans.arff".



1. To save it as csv file, go to Preprocessing tab and click the "Save" button.
2. Save the result as "bank-kmeans.csv".

Thank You

"DATA MINING WITH WEKA"

BAMSHAD MOBASHER

Mobasher, Bamshad. “K-Means Clustering in WEKA.” DePaul University College of Computing and Digital Media, facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/k-means.html. Accessed 9 May 2021.