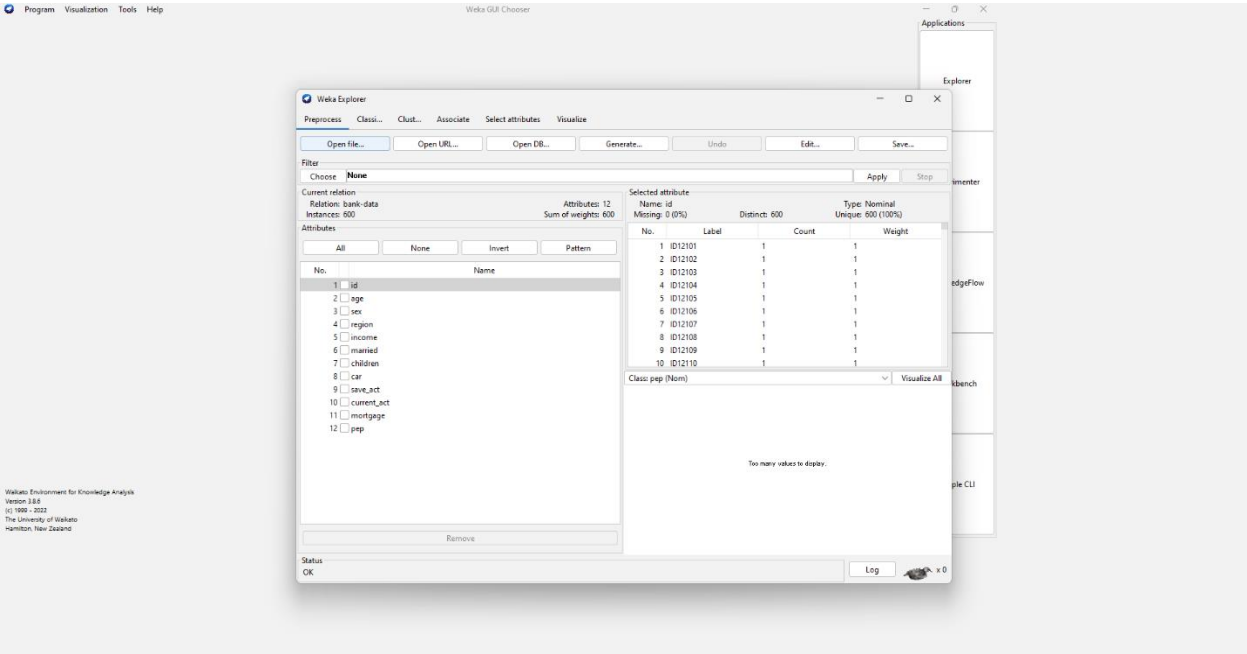
DESCRIPTION

The bank data set contains information about customers of a bank, with attributes like age, job, marital status, education, housing, loan status, and more.

The goal is to use k-means clustering on this customer data to identify segments or groups of similar customers.

Specifically, I am clustering the customers based on their attributes like demographics, finances, loan status, etc. to find clusters representing distinct customer segments.

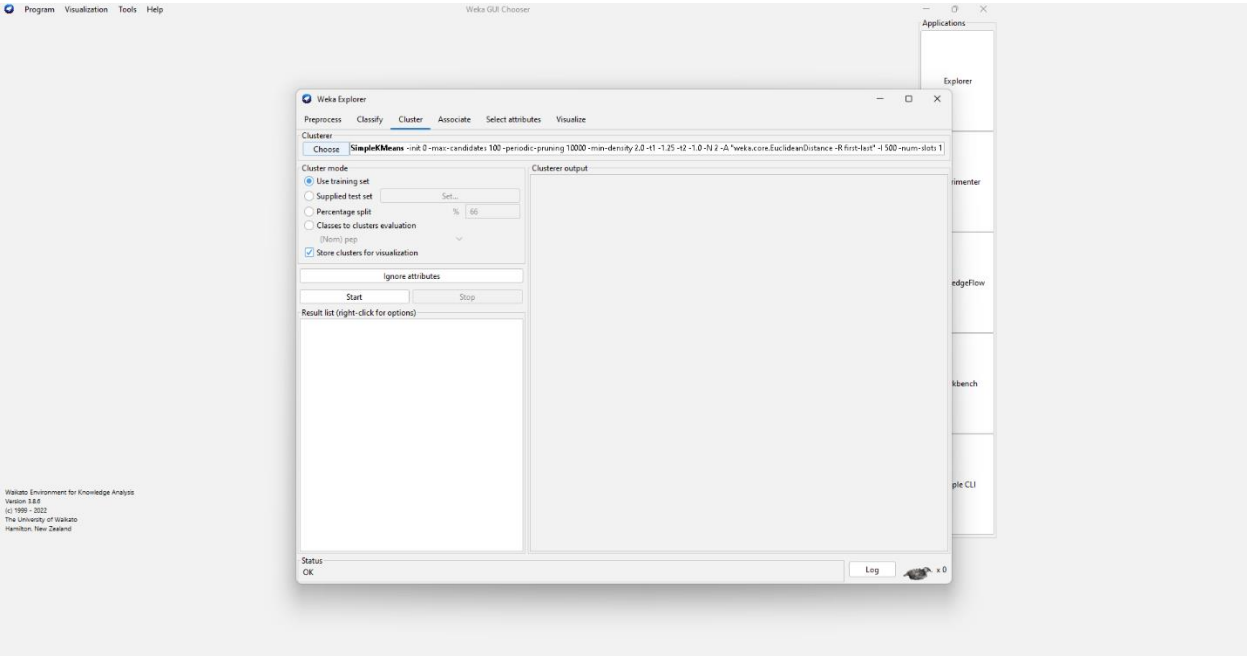
Video link: https://youtu.be/5wZLqI_bwCQ
Data Source: http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/k-means.html



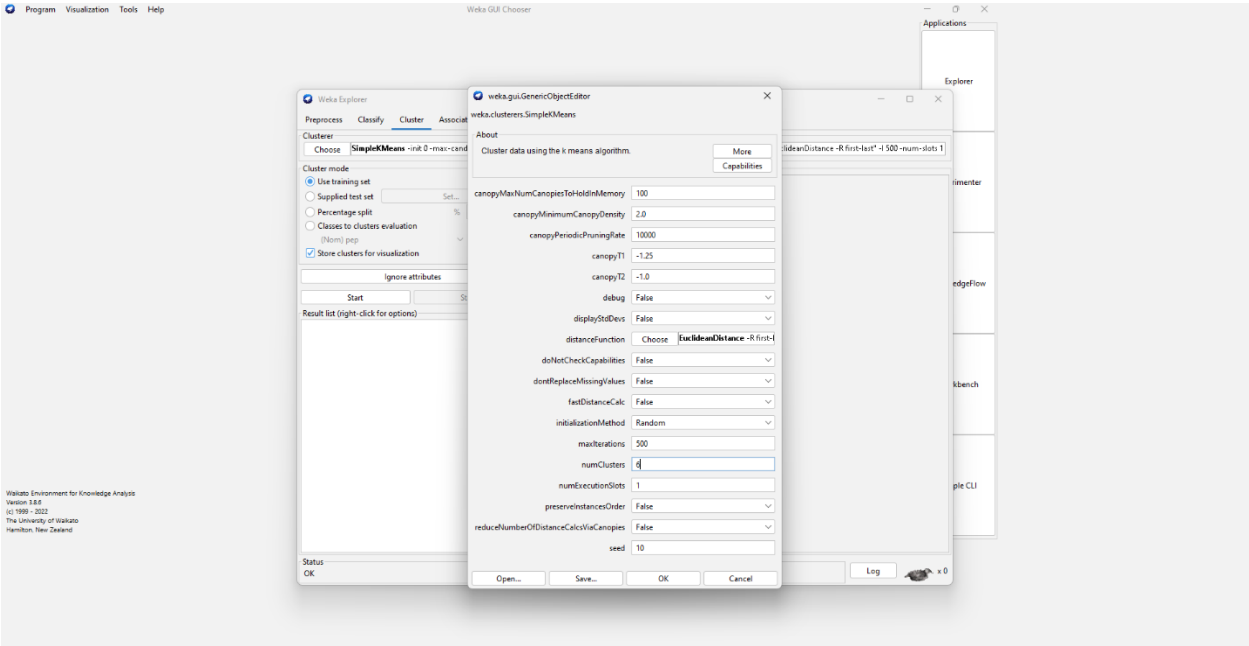
Some key points about the data and clustering goal:

- It contains 600 customers with details on their age, gender, education, job, housing, family status, income, etc.
- Also includes their response on signing up for a personal equity plan (PEP) product.
- We want to cluster customers into groups with similar characteristics using k-means.
- The number of clusters k is set to 6 based on domain knowledge and intuition.
- Attributes like age, income, location, marital status, etc. will drive the clustering to find distinct segments.
- Analyzing the cluster centroids will help characterize the key attributes of each customer segment.
- Visualizing clusters by attributes can also provide insights into the segments.

PROCEDURE



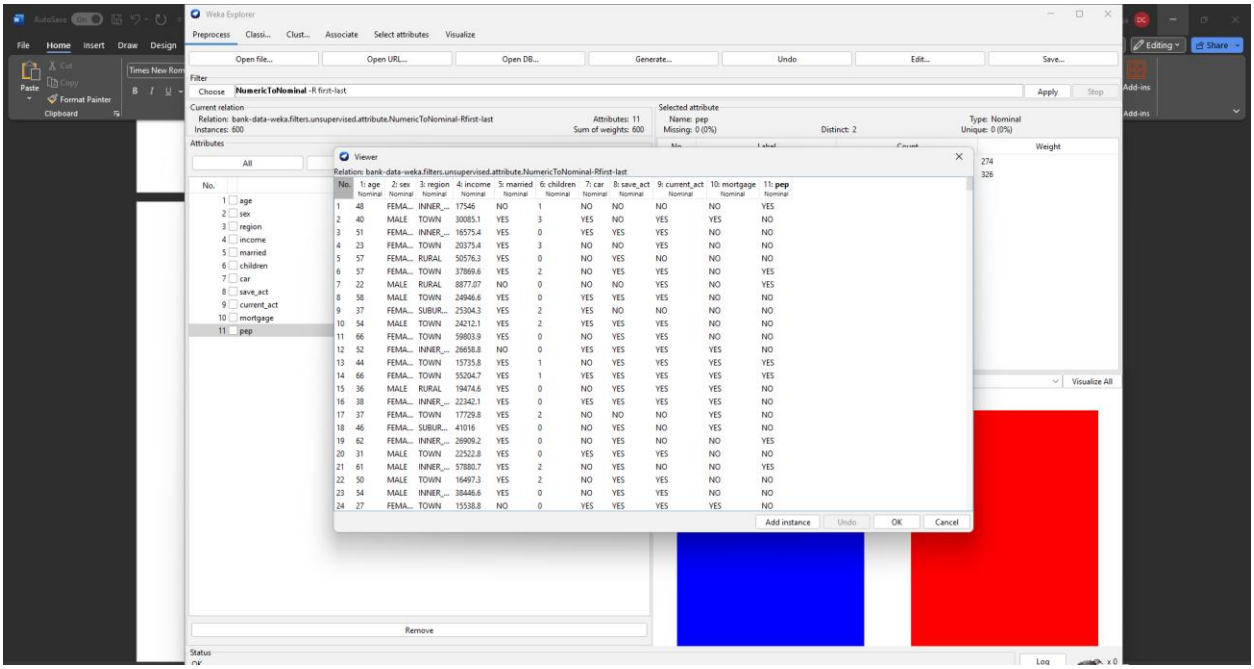
1. I clicked on the "Preprocess" tab in WEKA Explorer. This opened the data preprocessing tools.
2. I clicked the "Open file..." button and selected my bank-data.csv file containing the raw data. This loaded the data set into the Preprocess panel.
3. In the Preprocess panel on the right, I clicked the "Choose" button under Filters to bring up the Filter Selector window.



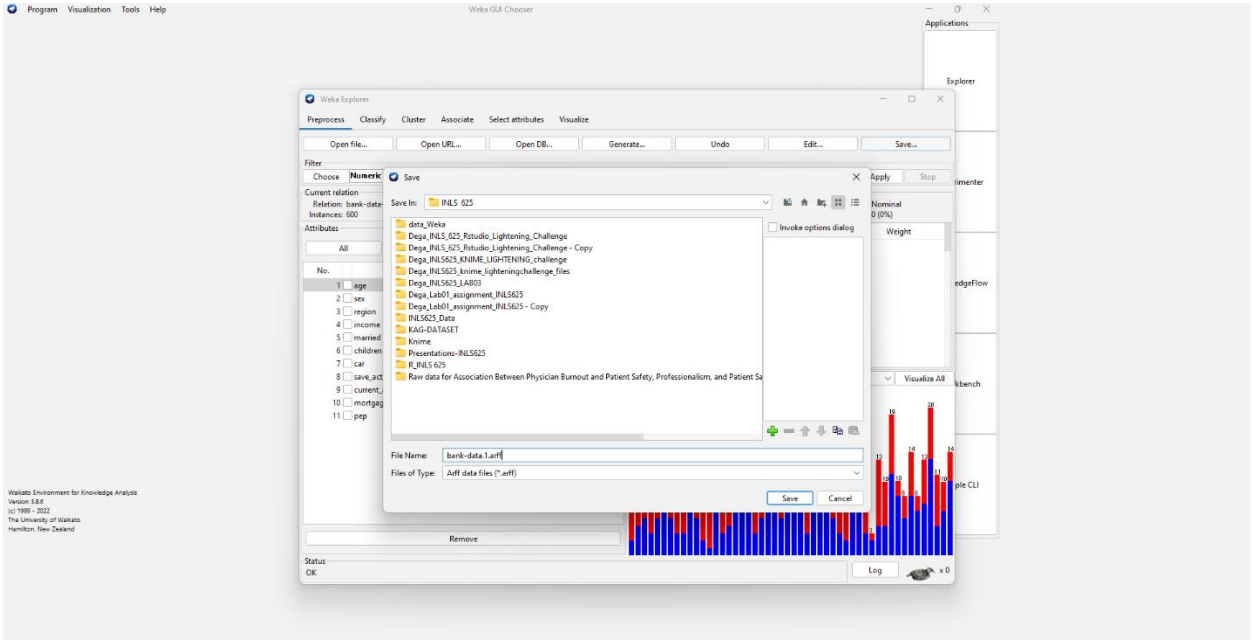
4. In the Filter selector window, I selected:

- Filter type: Unsupervised
- Filter: NumericToNominal

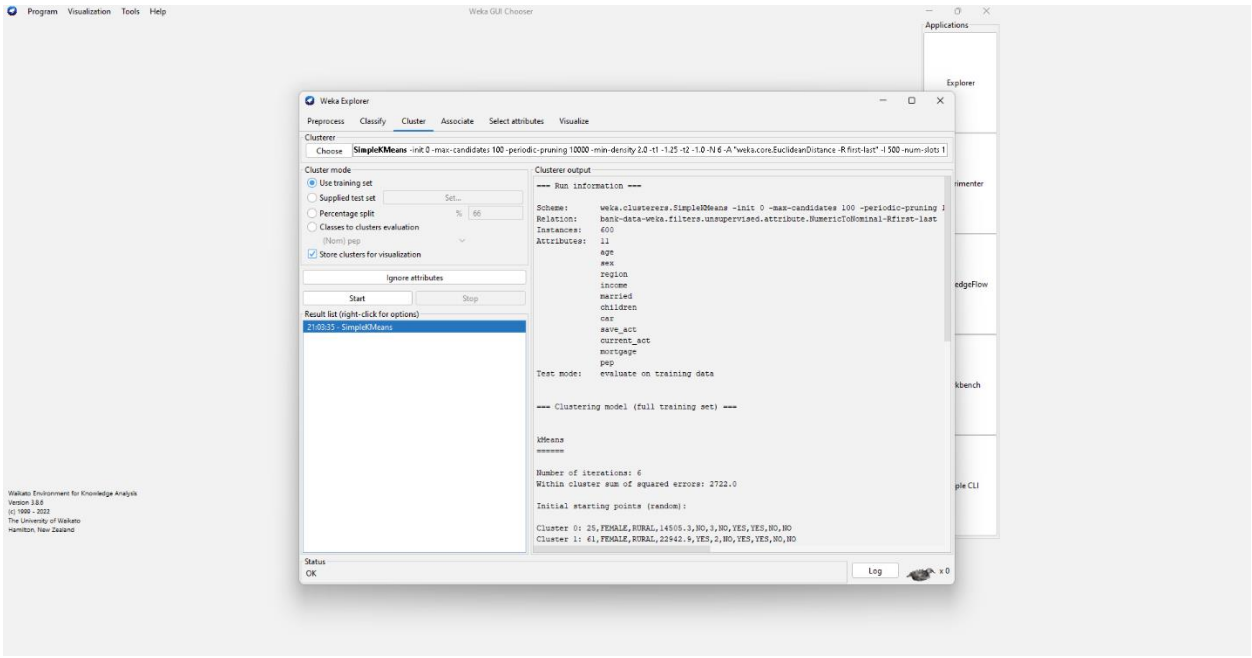
5. I clicked OK to close the Filter selector window and add the NumericToNominal filter to the Preprocess panel.
6. I clicked on the text field next to the NumericToNominal filter name to bring up the filter options.
7. For the "children" column, I set the "Attribute index" parameter to the column number for children, in this case 10.

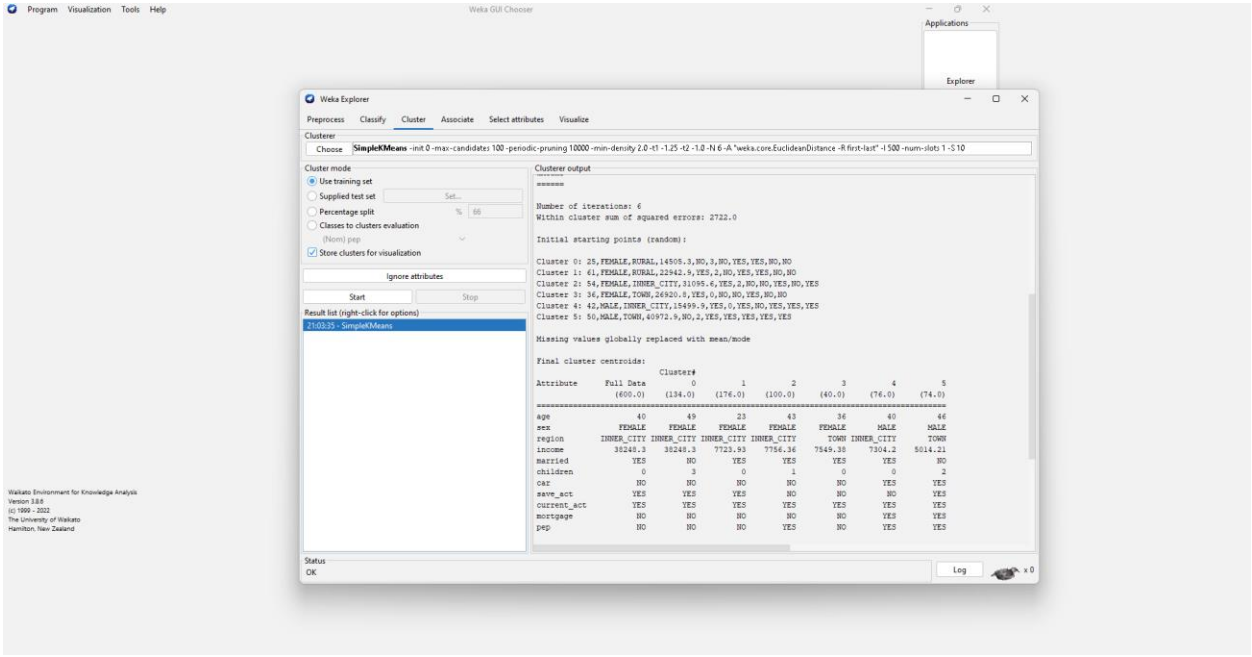


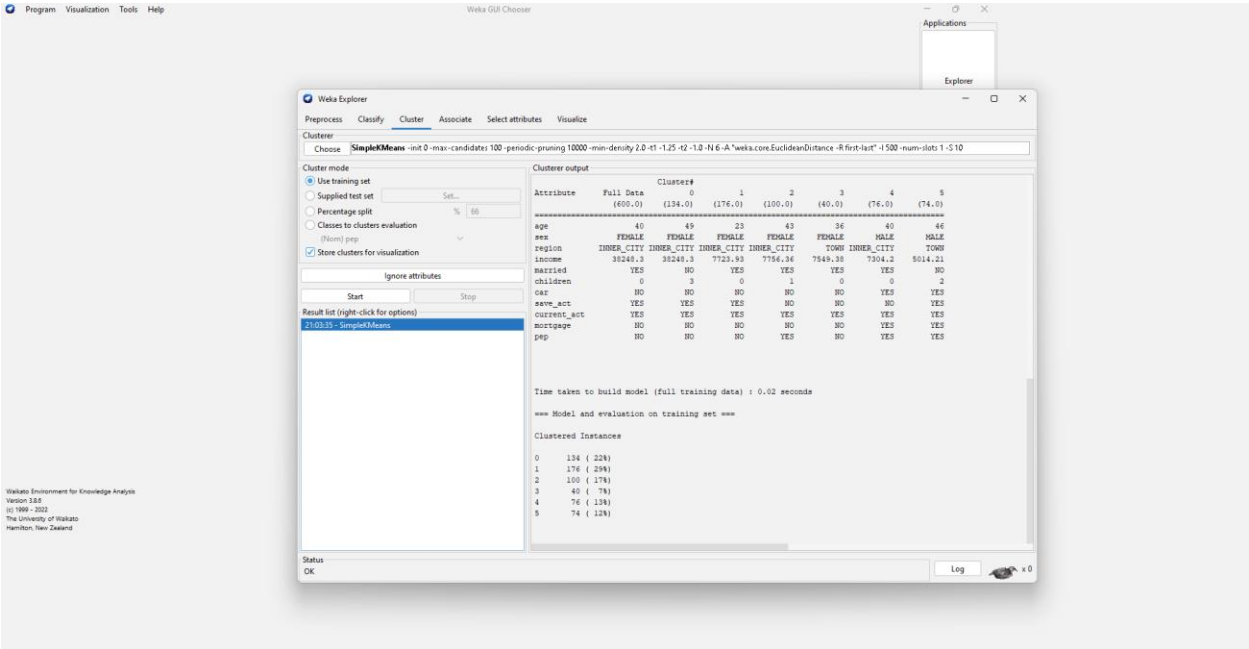
8. I clicked Apply to apply the NumericToNominal filter to the children column and convert it from numeric to nominal categories.
9. The converted data set with the changed children column now appears in the Preprocess output window. I saved this preprocessed data to use for clustering.



10. By applying the NumericToNominal filter, I was able to successfully convert the numeric "children" column into nominal values required for the clustering algorithm.


Program Visualization Tools Help
Weka GUI Chooser
Applications
Explorer
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Clusterer
Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 6 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1
Cluster mode
Use training set
Supplied test set Set...
Percentage split % 66
Classes to clusters evaluation
(Nom) pep
Store clusters for visualization
Ignore attributes
Start Stop
Result list (right-click for options)
21:03:35 - SimpleKMeans
Clusterer output
=== Run information ===
Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning
Relation: bank-data-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
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: 6
Within cluster sum of squared errors: 2722.0
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
Status
OK
Log
x 0
Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

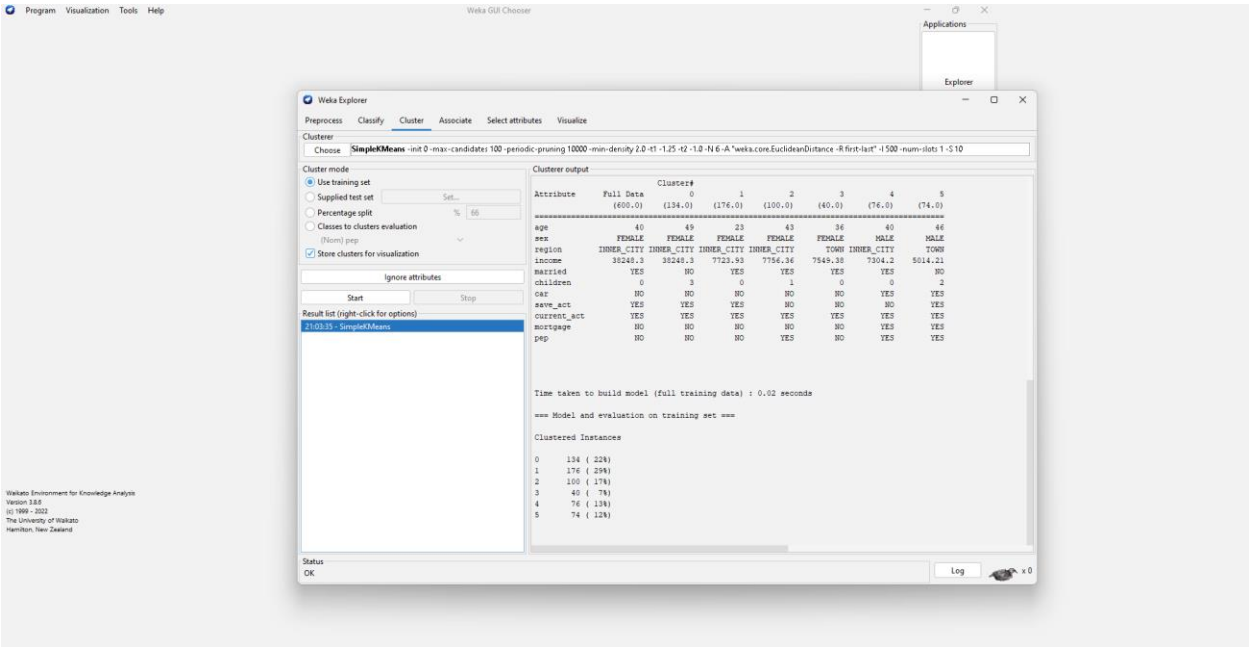

Program Visualization Tools Help
Weka GUI Chooser
Applications
Explorer
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Clusterer
Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 6 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Cluster mode
Use training set
Supplied test set Set...
Percentage split % 66
Classes to clusters evaluation
(Nom) pep
Store clusters for visualization
Ignore attributes
Start Stop
Result list (right-click for options)
21:03:35 - SimpleKMeans
Clusterer output
======
Number of iterations: 6
Within cluster sum of squared errors: 2722.0
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,26920.9,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
Final cluster centroids:
Attribute Full Data 0 1 2 3 4 5
(600.0) (134.0) (176.0) (100.0) (40.0) (76.0) (74.0)
age 40 49 23 43 36 40 46
sex FEMALE FEMALE FEMALE FEMALE FEMALE MALE MALE
region INNER_CITY INNER_CITY INNER_CITY INNER_CITY TOWN INNER_CITY TOWN
income 38248.3 38248.3 7723.93 7756.36 7549.38 7304.2 5014.21
married YES NO YES YES YES YES NO
children 0 3 0 1 0 0 2
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 YES
pep NO NO NO YES NO YES YES
Status
OK
Log
x 0
Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand



INTERPRETATION:

The centroids represent the mean (average) value for each attribute within a cluster. Looking at cluster 0 for example:

- The centroid age is 49 - meaning the average age of instances assigned to cluster 0 is 49.
- The centroid for the region is INNER_CITY - meaning most instances in cluster 0 are from the inner-city region.



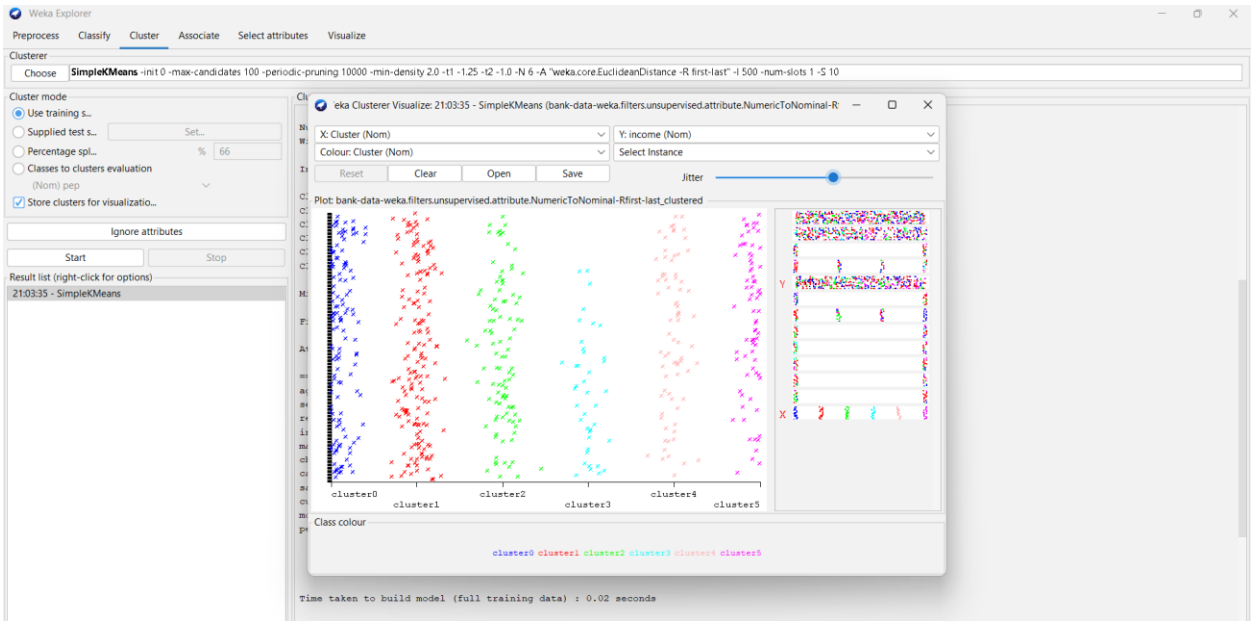
- The income centroid is 38,248.3 - this represents the mean income level for cluster 0 members.

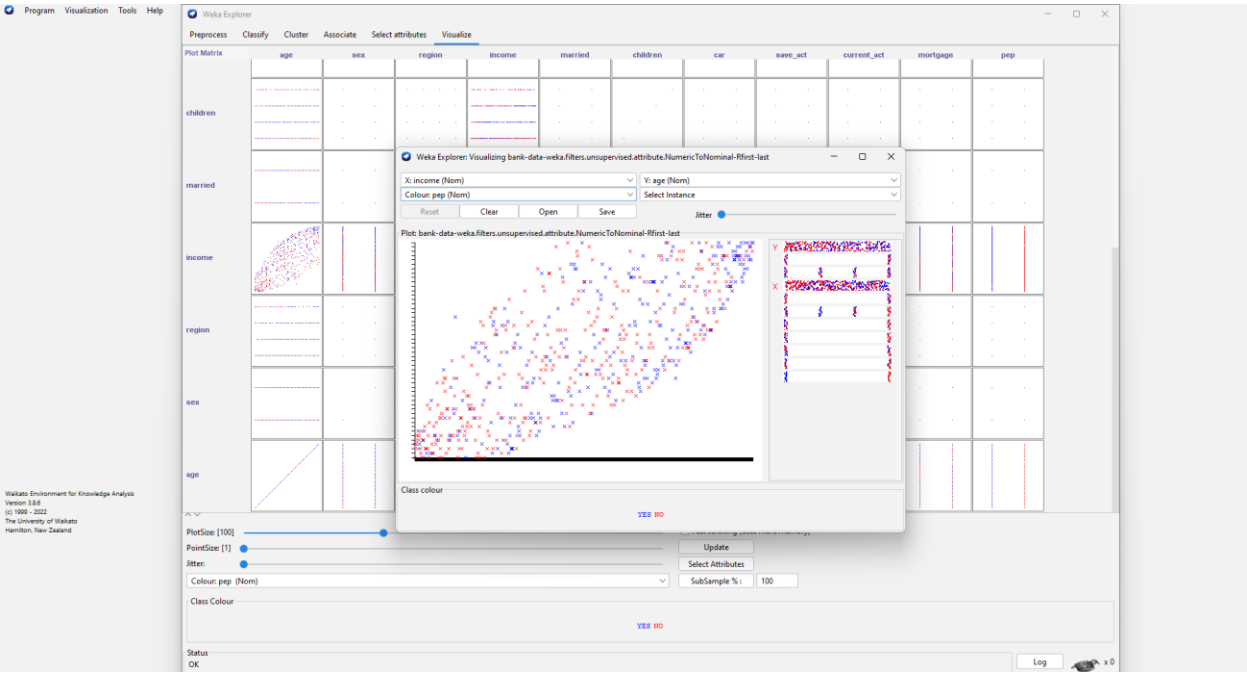
And so on for the other attributes.

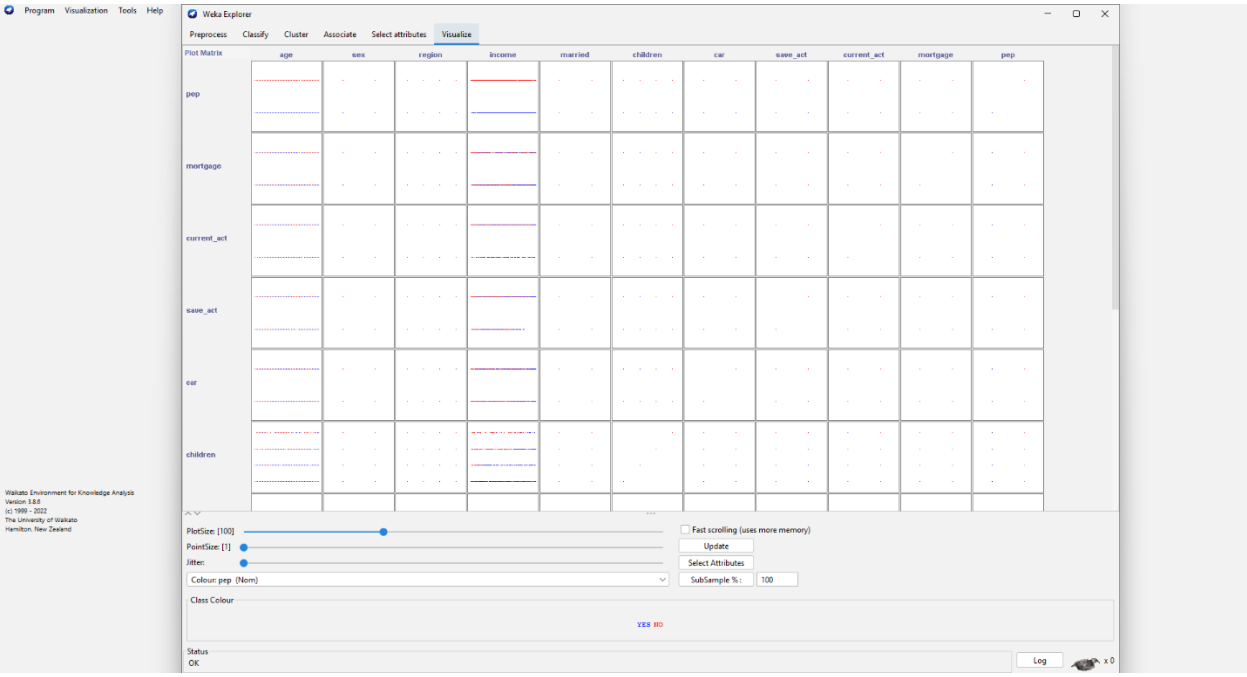
Examining the centroids allows you to characterize each cluster by the dominant attribute values:

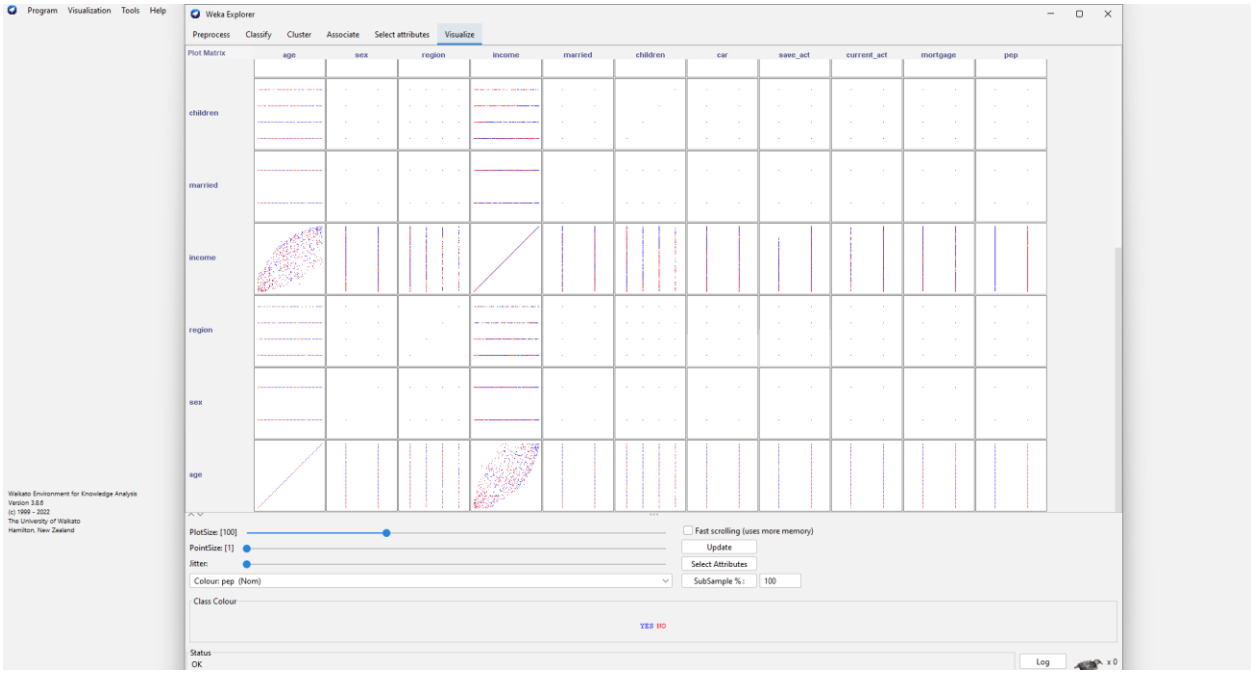
- Cluster 0: Middle-aged inner-city females, average income, no car, has savings, mortgage, or PEP.
- Cluster 1: Younger inner-city females, low income, no car or mortgage, but has savings.
- Cluster 2: Middle-aged inner-city females, average income, no car, has mortgage and PEP but no savings.
- Cluster 3: Middle-aged small-town females, average income, no car or mortgage, no savings.
- Cluster 4: Middle-aged inner-city males, low income, has car, no savings, has mortgage and PEP.
- Cluster 5: Middle-aged small-town males, high income, has car and mortgage, has savings and PEP.

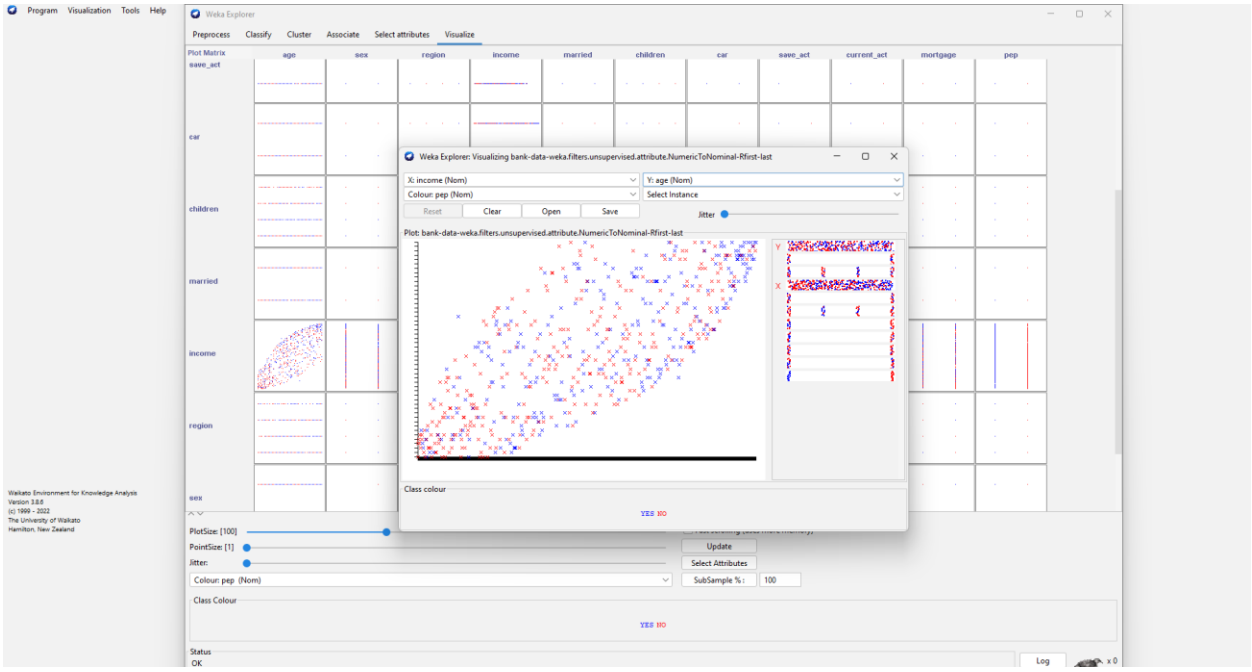
The percentages show the relative size of each cluster.




Program Visualization Tools Help
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Plot Matrix
age sex region income married children car save_act current_act mortgage pep
Weka Explorer: Visualizing bank-data-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
X: income (Nom)
Y: age (Nom)
Colour: pep (Nom)
Select Instance
Reset Clear Open Save
Jitter
Plot: bank-data-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Class colour
YES NO
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
Colour: pep (Nom)
SubSample %:
100
Class Colour
YES NO
Status
OK
Log
x 0


Program Visualization Tools Help
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Plot Matrix
age sex region income married children car save_act current_act mortgage pep
PlotSize: [100]
PointSize: [1]
Jitter:
Fast scrolling (uses more memory)
Update
Select Attributes
Colour: pep (Nom)
SubSample %:
100
Class Colour
YES NO
Status
OK
Log
x 0


Program Visualization Tools Help
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Plot Matrix
age sex region income married children car save_act current_act mortgage pep
children
married
income
region
sex
age
Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
PlotSize: [100]
PointSize: [1]
Jitter:
Fast scrolling (uses more memory)
Update
Select Attributes
Colour: pep (Nom)
SubSample %:
100
Class Colour
YES NO
Status
OK
Log
x 0


Program Visualization Tools Help
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Plot Matrix
age sex region income married children car save_act current_act mortgage pep
save_act
car
children
married
income
region
sex
Weka Explorer: Visualizing bank-data-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
X: income (Nom)
Y: age (Nom)
Colour: pep (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: bank-data-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Class colour
YES NO
Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
Colour: pep (Nom)
SubSample %:
100
Class Colour
YES NO
Status
OK
Log
x 0

Dega_INLS625_Lightning Challenge: WEKA