Task 3 A:

* In this run we have ignored some of the attributes as we won’t be clustering the data on the basis of age, sex, race, marital status, education number, relationship, work class, and fnlwght.

We can see that sum of squared errors comes out to be **8216.252621932737**

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: clusteringdata

Instances: 5000

Attributes: 15

Education

Occupation

CapitalGain

CapitalLoss

HoursPerWeek

NativeCountry

Class

Ignored:

Age

WorkClass

Fnlwght

EducationNumber

MaritalStatus

Relationship

Race

Sex

Test mode: evaluate on training data

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

kMeans

Number of iterations: 5

**Within cluster sum of squared errors: 8216.252621932737**

Initial starting points (random):

Cluster 0: ' HS-grad',' Sales',0,0,40,' United-States',' <=50K'

Cluster 1: ' Masters',' Prof-specialty',0,0,35,' United-States',' <=50K'

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute Full Data 0 1

(5000.0) (3053.0) (1947.0)

==============================================================================

Education HS-grad HS-grad Some-college

Occupation Prof-specialty Sales Prof-specialty

CapitalGain 1033.6402 873.1425 1285.3092

CapitalLoss 93.6968 47.2558 166.5187

HoursPerWeek 40.519 42.5441 37.3436

NativeCountry United-States United-States United-States

Class <=50K <=50K <=50K

Time taken to build model (full training data) : 0.02 seconds

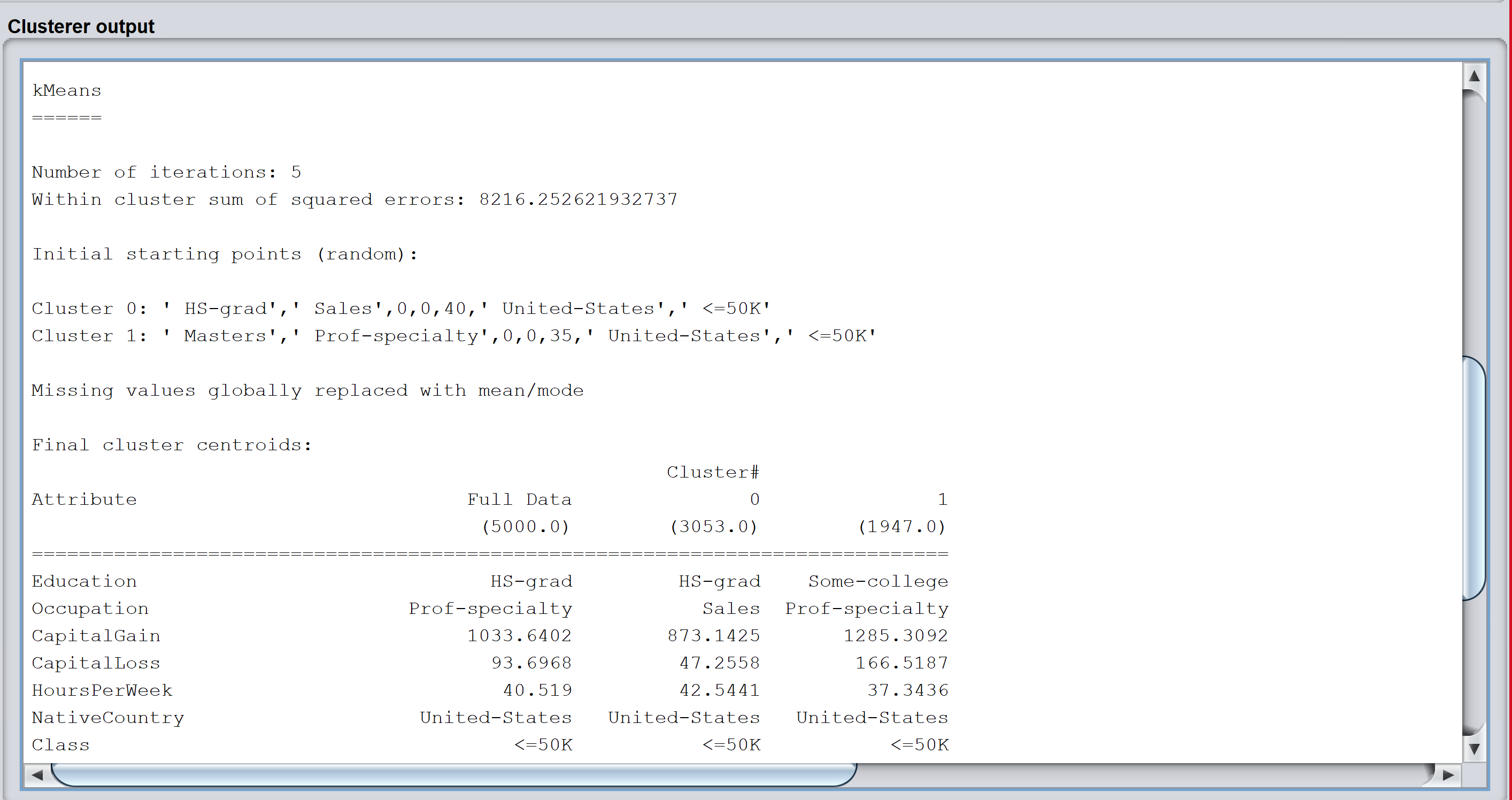
=== Model and evaluation on training set ===

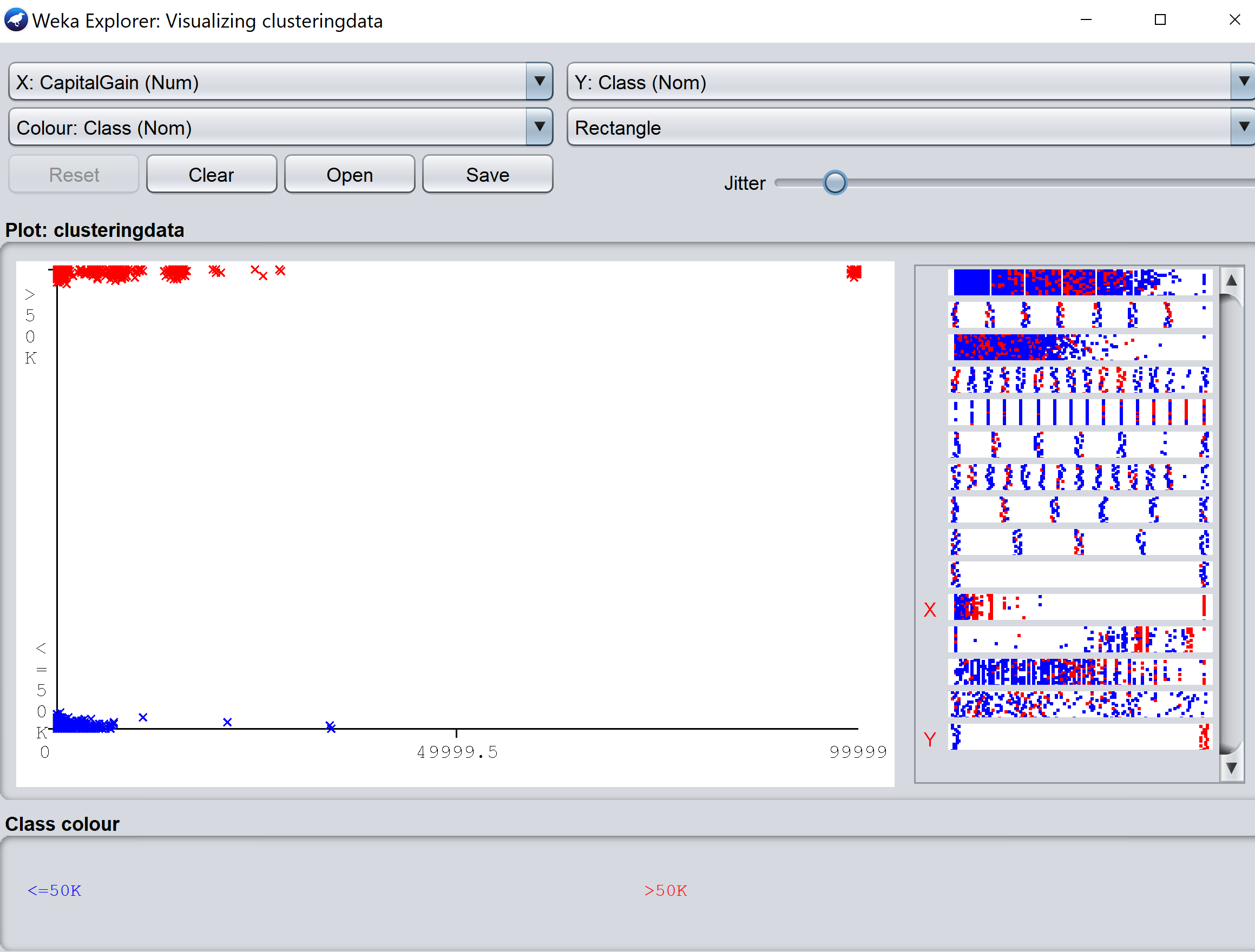
Clustered Instances

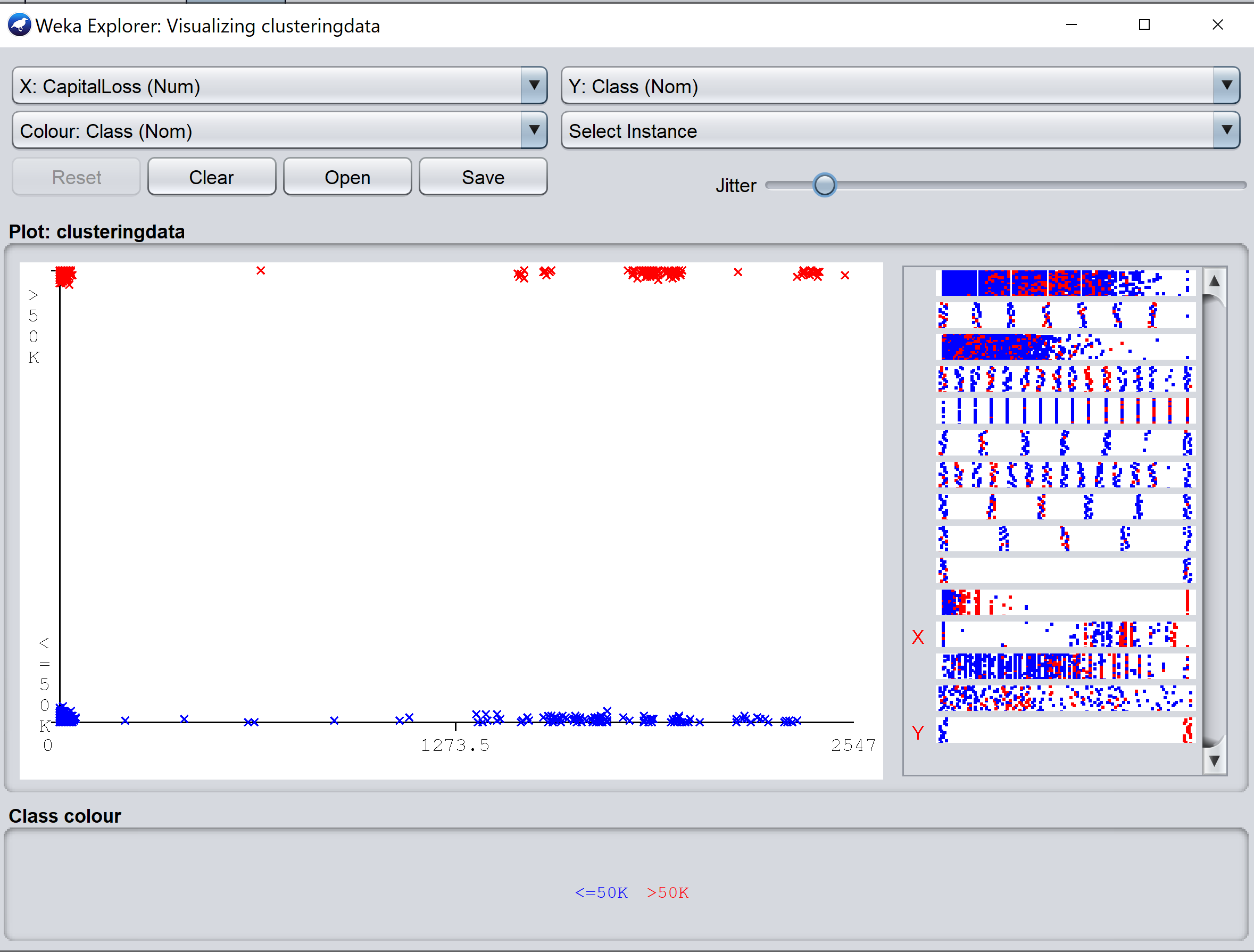
0 3053 ( 61%)

1 1947 ( 39%)

Below is the WEKA screenshot for your reference:







* Similarly, in this run we ignored same attributes as mentioned before and kept k = 3. In this case we get sum of squared error is **7204.9859420768125**

We can see that squared error is reduced when value of k is 3.

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: clusteringdata

Instances: 5000

Attributes: 15

Education

Occupation

CapitalGain

CapitalLoss

HoursPerWeek

NativeCountry

Class

Ignored:

Age

WorkClass

Fnlwght

EducationNumber

MaritalStatus

Relationship

Race

Sex

Test mode: evaluate on training data

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

kMeans

======

Number of iterations: 5

**Within cluster sum of squared errors: 7204.9859420768125**

Initial starting points (random):

Cluster 0: ' HS-grad',' Sales',0,0,40,' United-States',' <=50K'

Cluster 1: ' Masters',' Prof-specialty',0,0,35,' United-States',' <=50K'

Cluster 2: ' Bachelors',' Prof-specialty',0,1485,40,' United-States',' >50K'

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute Full Data 0 1 2

(5000.0) (2581.0) (1502.0) (917.0)

==============================================================================

Education HS-grad HS-grad Some-college Bachelors

Occupation Prof-specialty Sales Prof-specialty Prof-specialty

CapitalGain 1033.6402 386.186 208.2224 4207.9662

CapitalLoss 93.6968 51.2445 40.3449 300.5714

HoursPerWeek 40.519 41.3952 34.5985 47.7503

NativeCountry United-States United-States United-States United-States

Class <=50K <=50K <=50K >50K

Time taken to build model (full training data) : 0.03 seconds

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

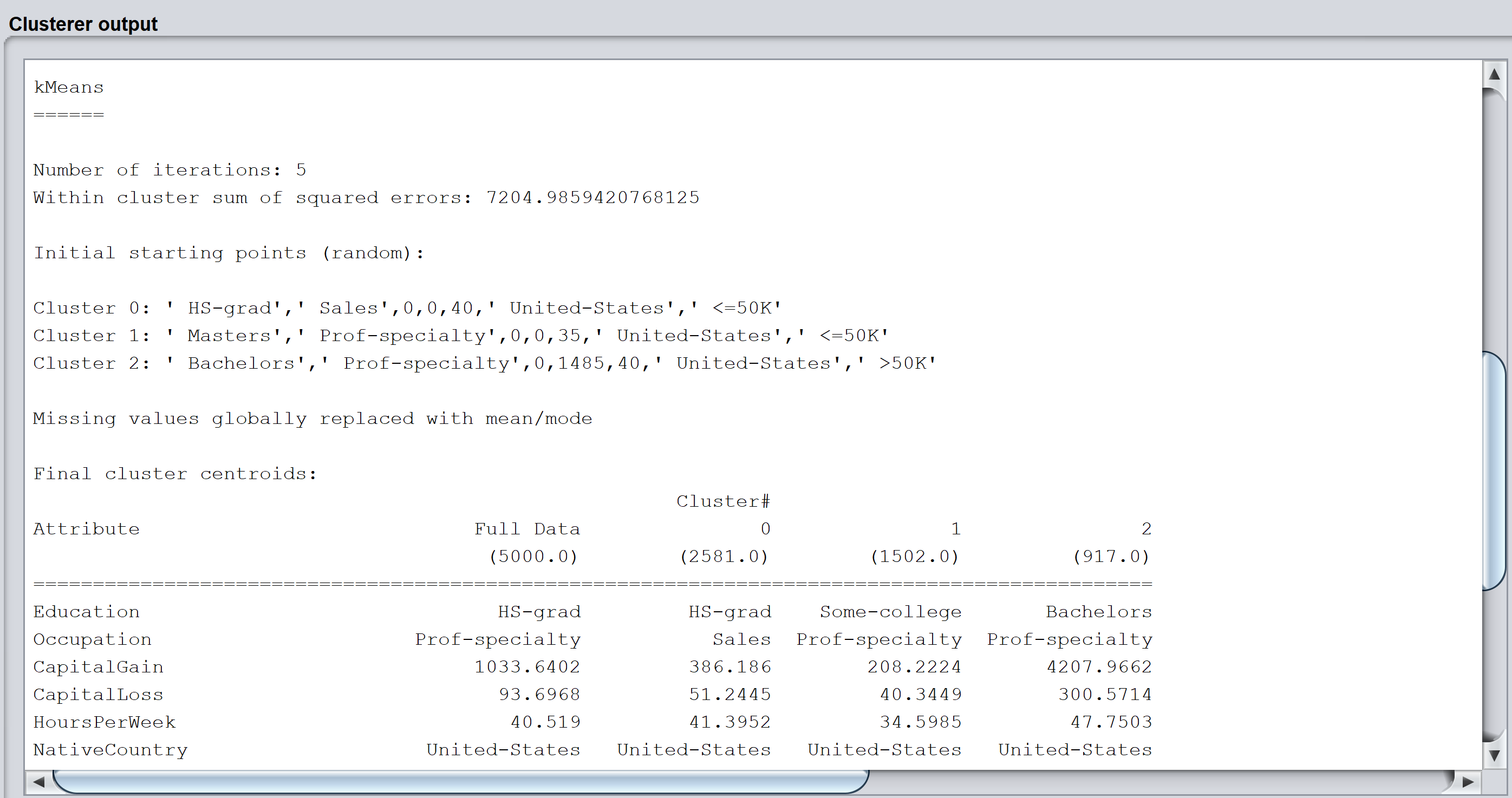
Clustered Instances

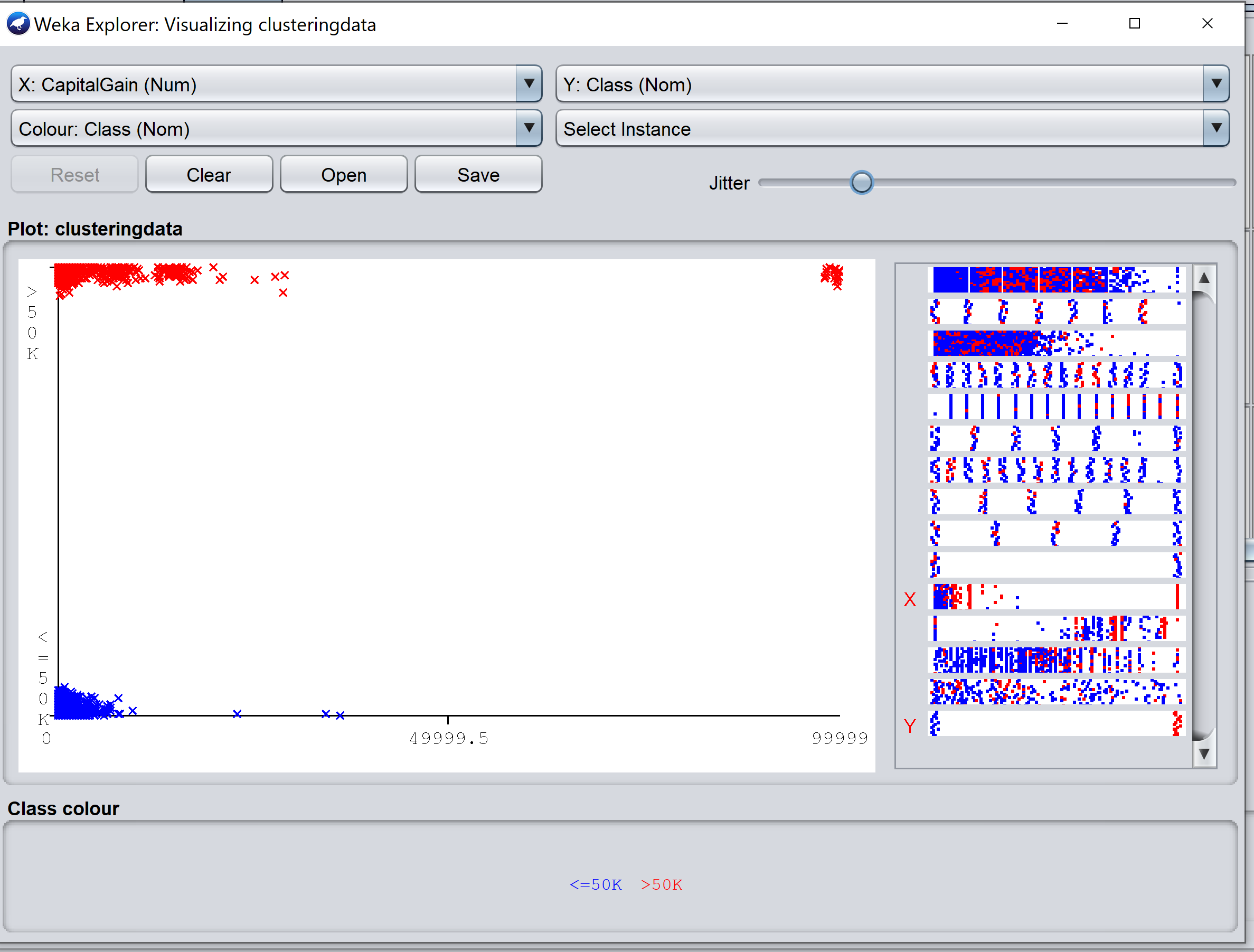
0 2581 ( 52%)

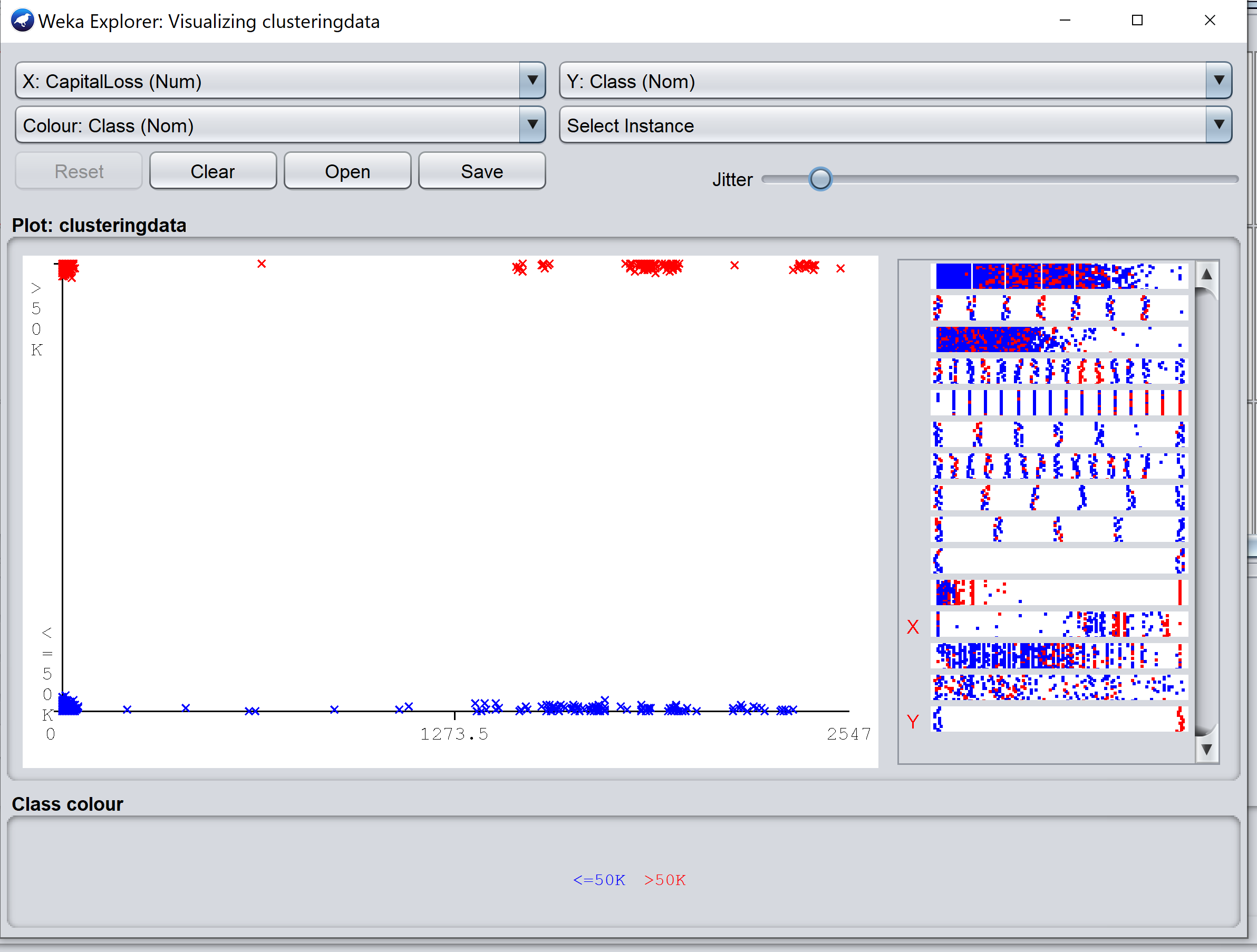
1 1502 ( 30%)

2 917 ( 18%)

Below is the WEKA screenshot for your reference.







* In this run, we did not ignore any attributes and kept number of clusters as 2. We could see that the sum of squared errors is **16413.867960259013.**

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: clusteringdata

Instances: 5000

Attributes: 15

Age

WorkClass

Fnlwght

Education

EducationNumber

MaritalStatus

Occupation

Relationship

Race

Sex

CapitalGain

CapitalLoss

HoursPerWeek

NativeCountry

Class

Test mode: evaluate on training data

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

kMeans

======

Number of iterations: 8

**Within cluster sum of squared errors: 16413.867960259013**

Initial starting points (random):

Cluster 0: 57,' Private',103809,' HS-grad',9,' Never-married',' Sales',' Unmarried',' White',' Female',0,0,40,' United-States',' <=50K'

Cluster 1: 29,' Local-gov',175262,' Masters',14,' Married-civ-spouse',' Prof-specialty',' Other-relative',' White',' Male',0,0,35,' United-States',' <=50K'

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute Full Data 0 1

(5000.0) (2443.0) (2557.0)

=====================================================================================

Age 38.6002 33.088 43.8666

WorkClass Private Private Private

Fnlwght 190768.399 194244.9668 187446.8287

Education HS-grad HS-grad HS-grad

EducationNumber 10.0796 9.7257 10.4177

MaritalStatus Married-civ-spouse Never-married Married-civ-spouse

Occupation Prof-specialty Adm-clerical Prof-specialty

Relationship Husband Not-in-family Husband

Race White White White

Sex Male Female Male

CapitalGain 1033.6402 311.9677 1723.1381

CapitalLoss 93.6968 46.9885 138.3226

HoursPerWeek 40.519 37.1347 43.7524

NativeCountry United-States United-States United-States

Class <=50K <=50K <=50K

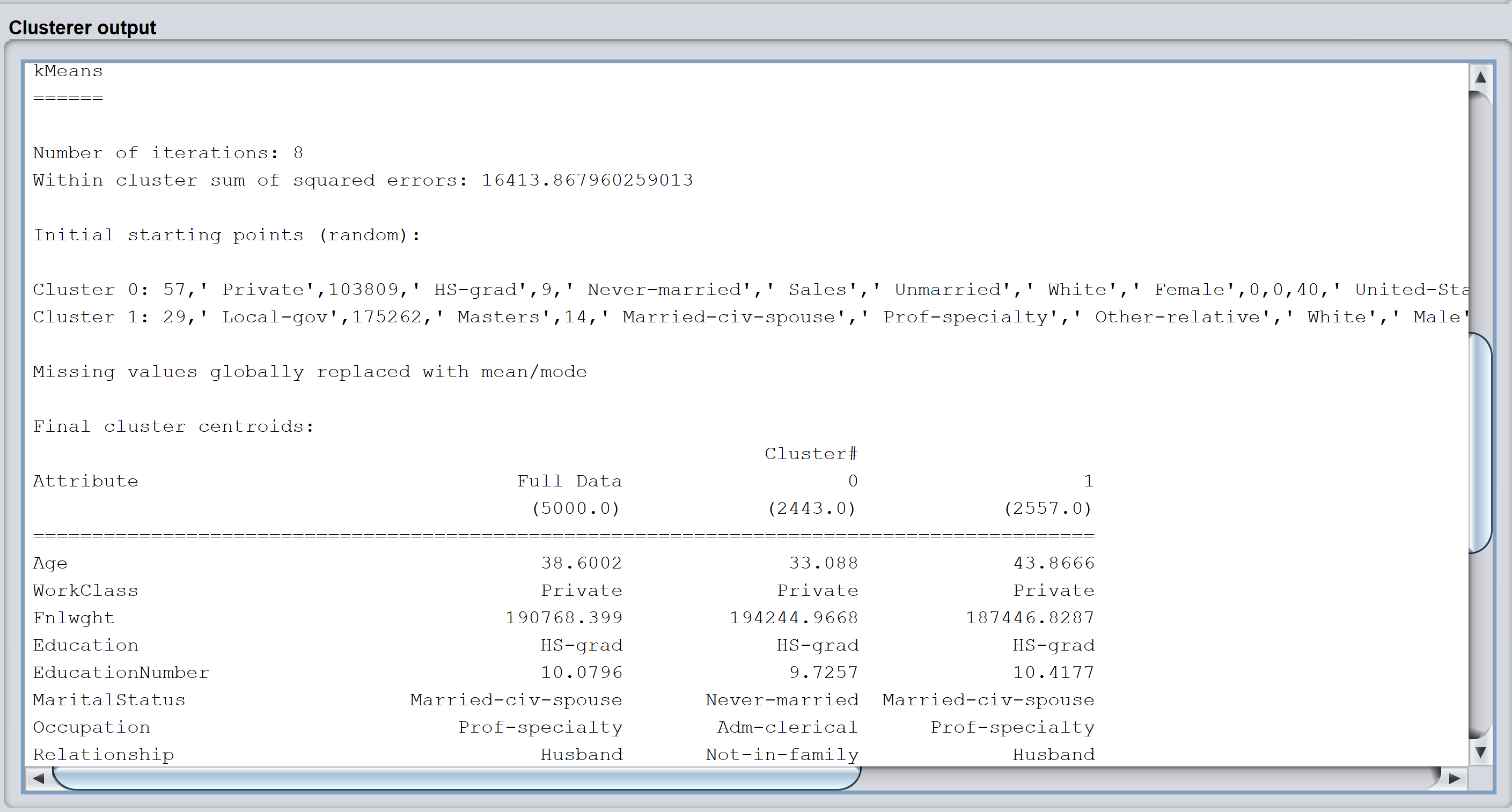
Time taken to build model (full training data) : 0.09 seconds

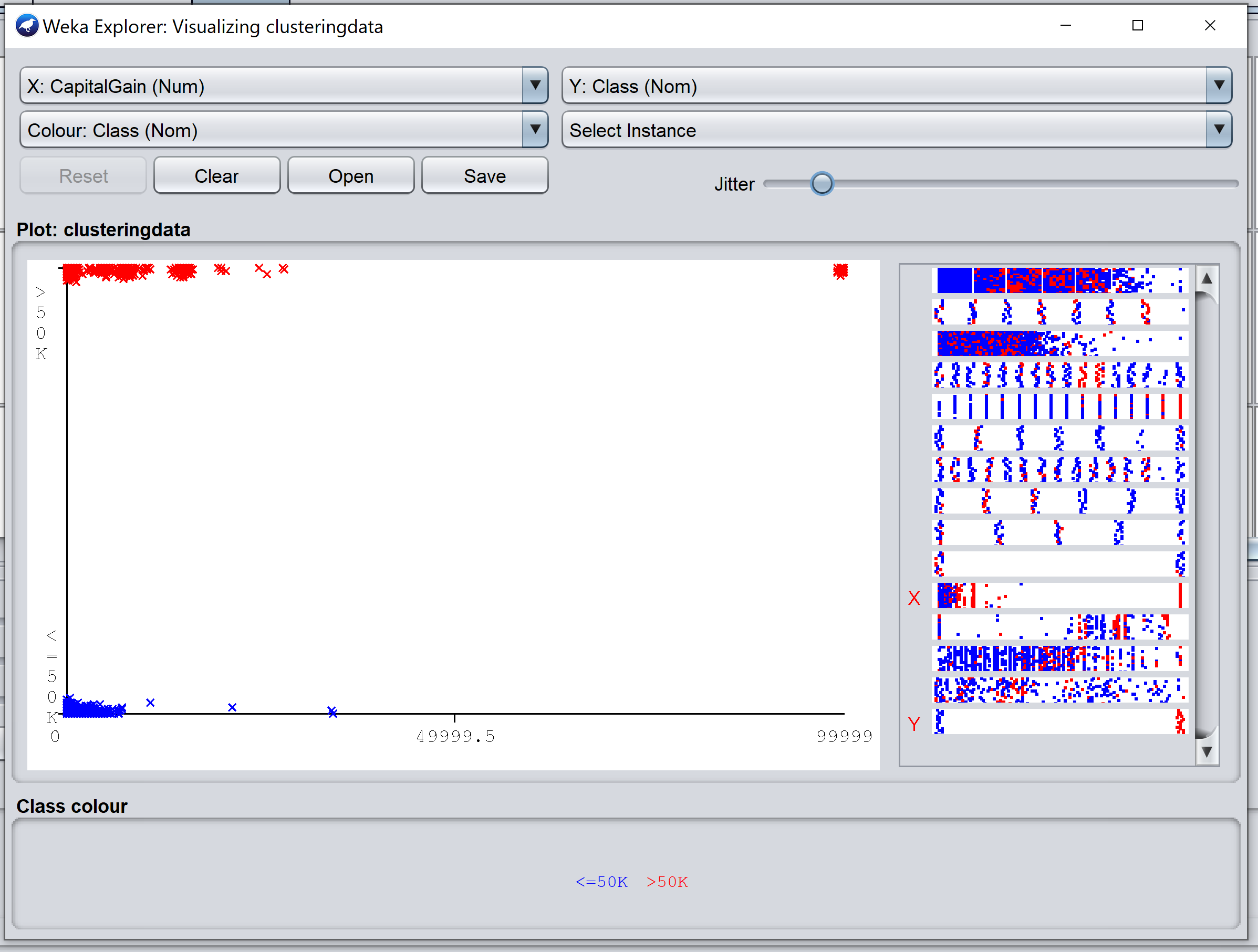
=== Model and evaluation on training set ===

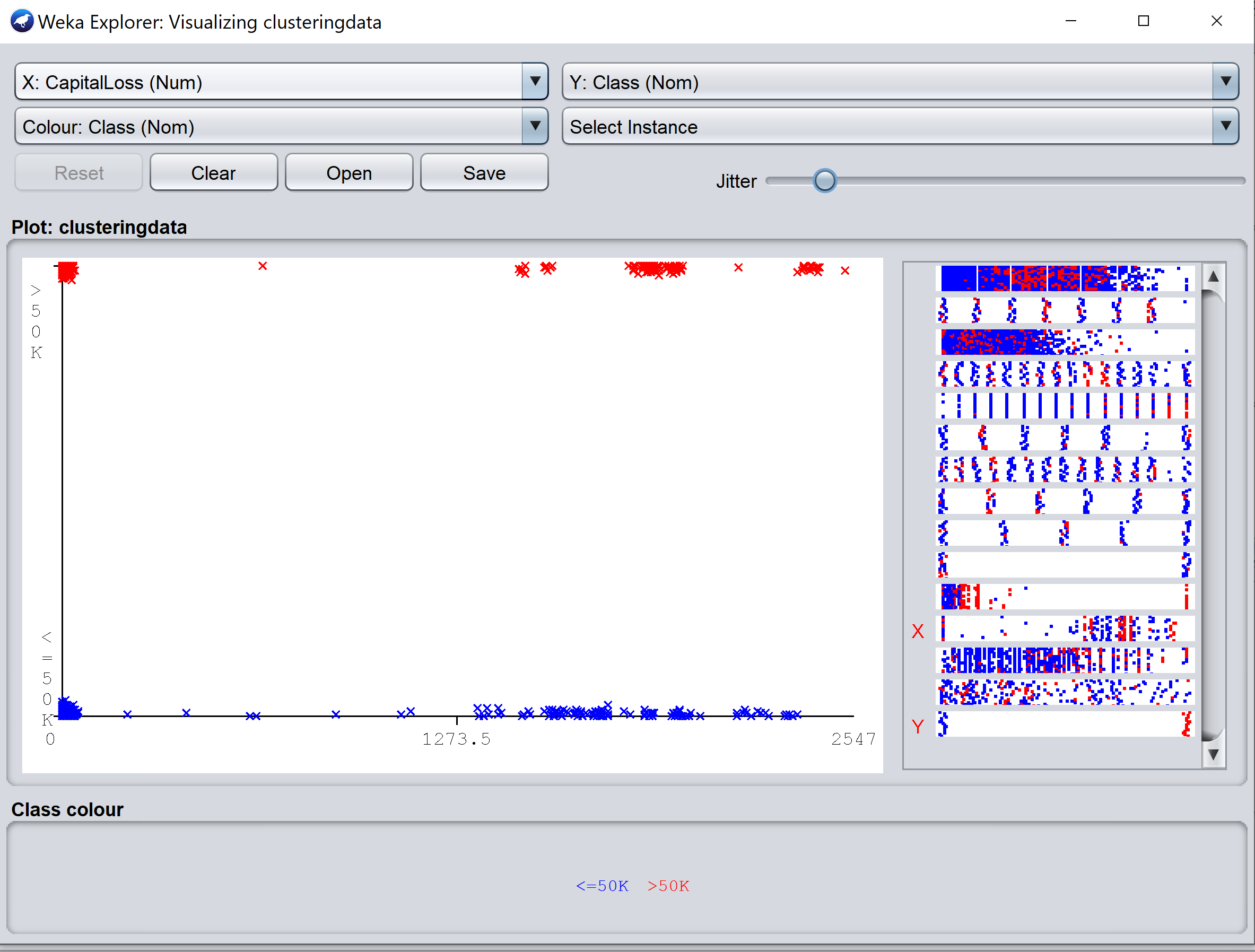
Clustered Instances

0 2443 ( 49%)

1 2557 ( 51%)







* In this run we did not ignore any attributes and kept number of clusters 3. We can see that sum of squared error is **14902.428042538753.**

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: clusteringdata

Instances: 5000

Attributes: 15

Age

WorkClass

Fnlwght

Education

EducationNumber

MaritalStatus

Occupation

Relationship

Race

Sex

CapitalGain

CapitalLoss

HoursPerWeek

NativeCountry

Class

Test mode: evaluate on training data

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

kMeans

======

Number of iterations: 11

**Within cluster sum of squared errors: 14902.428042538753**

Initial starting points (random):

Cluster 0: 57,' Private',103809,' HS-grad',9,' Never-married',' Sales',' Unmarried',' White',' Female',0,0,40,' United-States',' <=50K'

Cluster 1: 29,' Local-gov',175262,' Masters',14,' Married-civ-spouse',' Prof-specialty',' Other-relative',' White',' Male',0,0,35,' United-States',' <=50K'

Cluster 2: 48,' Private',202467,' Bachelors',13,' Married-civ-spouse',' Prof-specialty',' Husband',' White',' Male',0,1485,40,' United-States',' >50K'

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute Full Data 0 1 2

(5000.0) (2200.0) (1654.0) (1146.0)

==============================================================================

Age 38.6002 33.9686 41.2128 43.7208

WorkClass Private Private Private Private

Fnlwght 190768.399 191313.4286 193100.0586 186356.8543

Education HS-grad HS-grad Some-college Bachelors

EducationNumber 10.0796 9.7982 9.0145 12.1571

MaritalStatus Married-civ-spouse Never-married Married-civ-spouse Married-civ-spouse

Occupation Prof-specialty Adm-clerical Craft-repair Exec-managerial

Relationship Husband Not-in-family Husband Husband

Race White White White White

Sex Male Female Male Male

CapitalGain 1033.6402 276.4777 388.5133 3418.2801

CapitalLoss 93.6968 51.6845 61.6524 220.5977

HoursPerWeek 40.519 37.3077 41.1959 45.7068

NativeCountry United-States United-States United-States United-States

Class <=50K <=50K <=50K >50K

Time taken to build model (full training data) : 0.08 seconds

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

Clustered Instances

0 2200 ( 44%)

1 1654 ( 33%)

2 1146 ( 23%)

