National Institute Of Technology – Calicut

Data Mining

(Data pre-processing assignment)

Que 03.

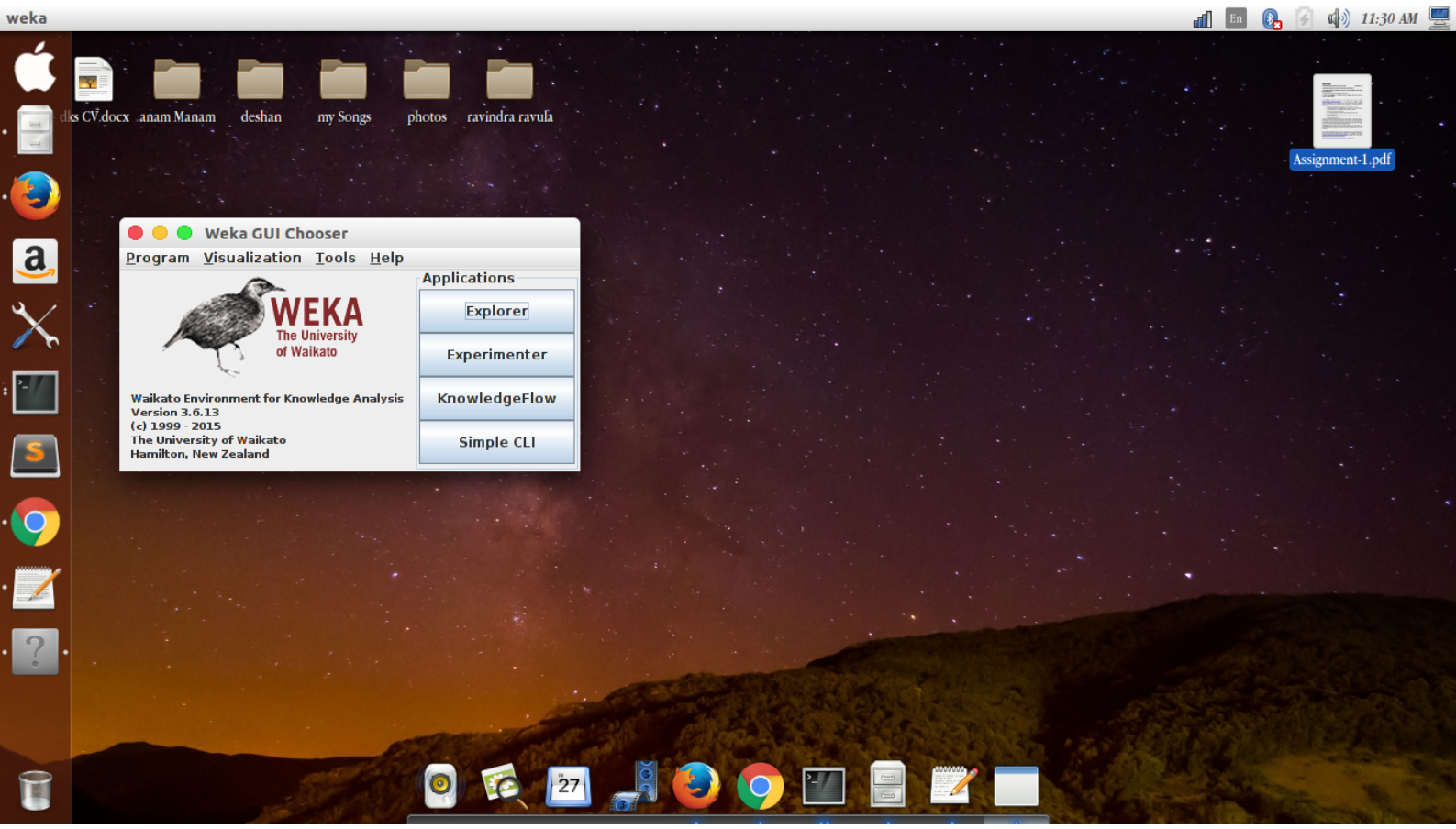
Data Pre-processing Using Weka

T.G. Deshan K. Sumanathilaka
B150413CS
Computer Science and Engineering
(B.Tech)

01.The Data set Which was downloaded from the link given.

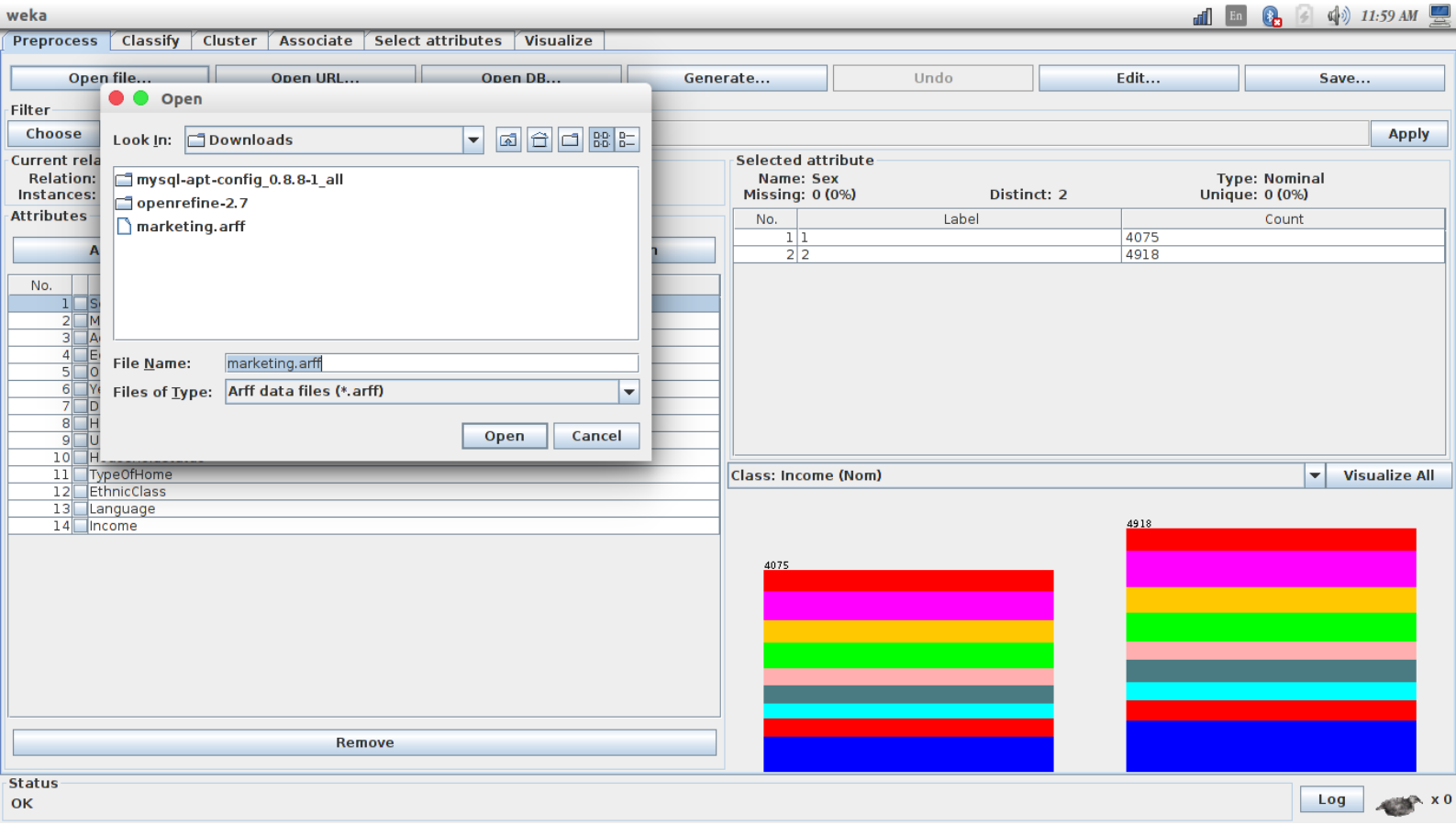
(https://sci2s.ugr.es/keel/dataset/data/missing/marketing.zip.)

02. Weka


weka
11:30 AM
ks CV.docx
anam Manam
deshan
my Songs
photos
ravindra ravula
Assignment-1.pdf
Weka GUI Chooser
Program Visualization Tools Help
WEKA
The University of Waikato
Applications
Explorer
Experimenter
KnowledgeFlow
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.6.13
(c) 1999 - 2015
The University of Waikato
Hamilton, New Zealand

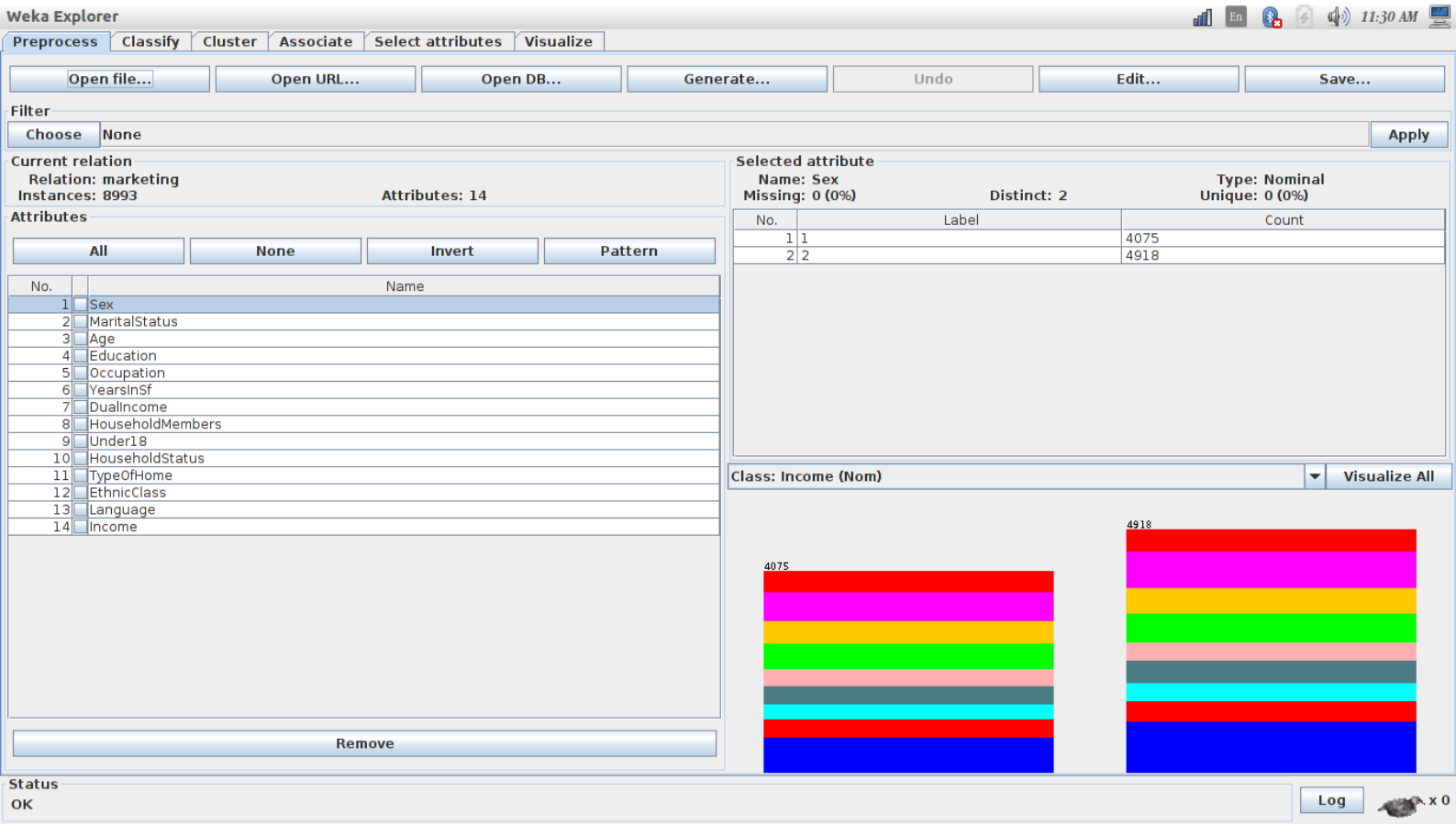
03.

Loading the data set to Weka Application.



04.

Data Set was Loaded to the Weka Application.



05. Data Pre-processing .

Step 01.

Preprocess missing values with the various options provided by WEKA

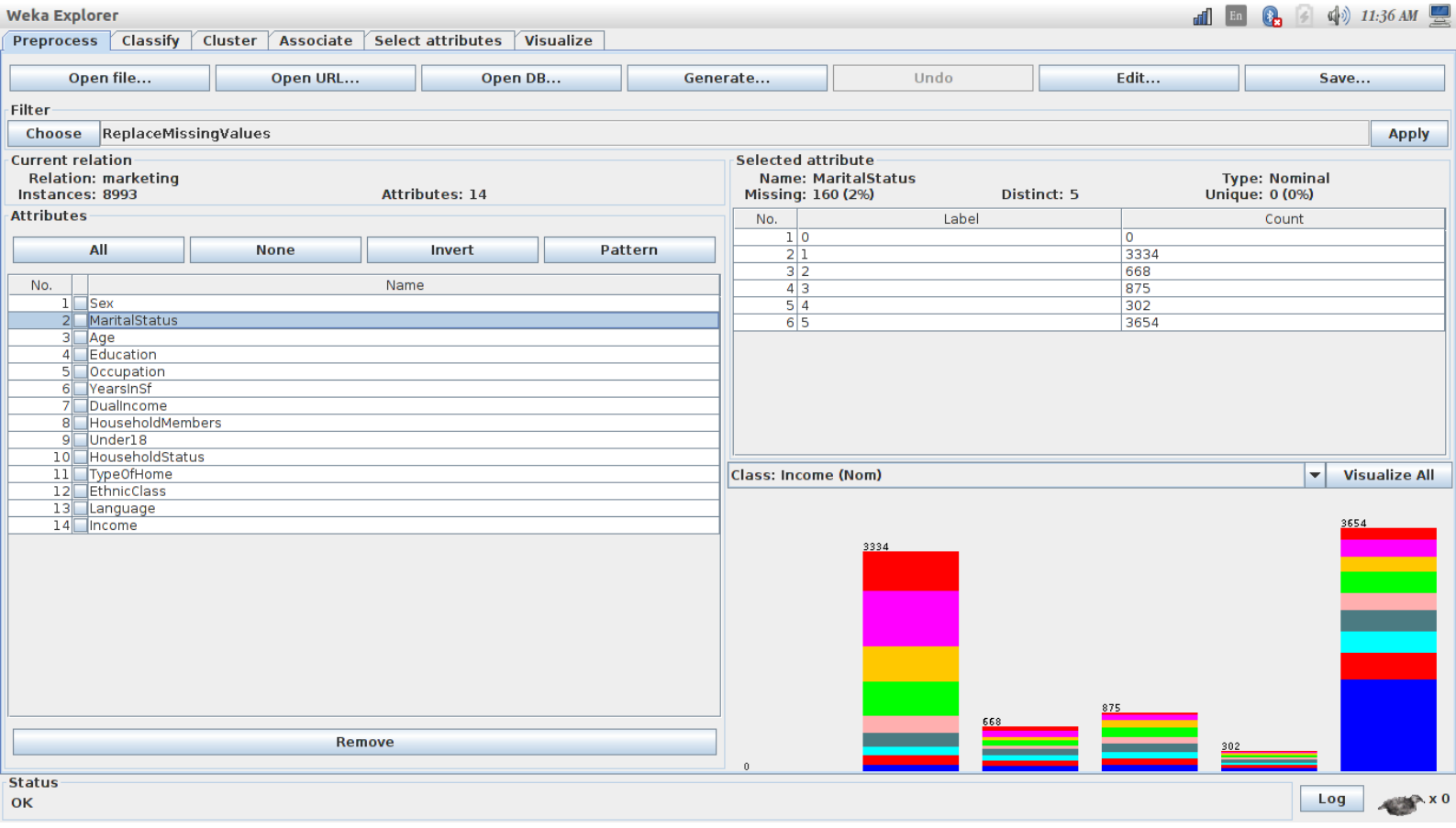
a) REPLACE MISSING VALUES

Before Applying The Filter :

1)Marital Status

Missing Values : 160 (02%)

2) Instances : 8993

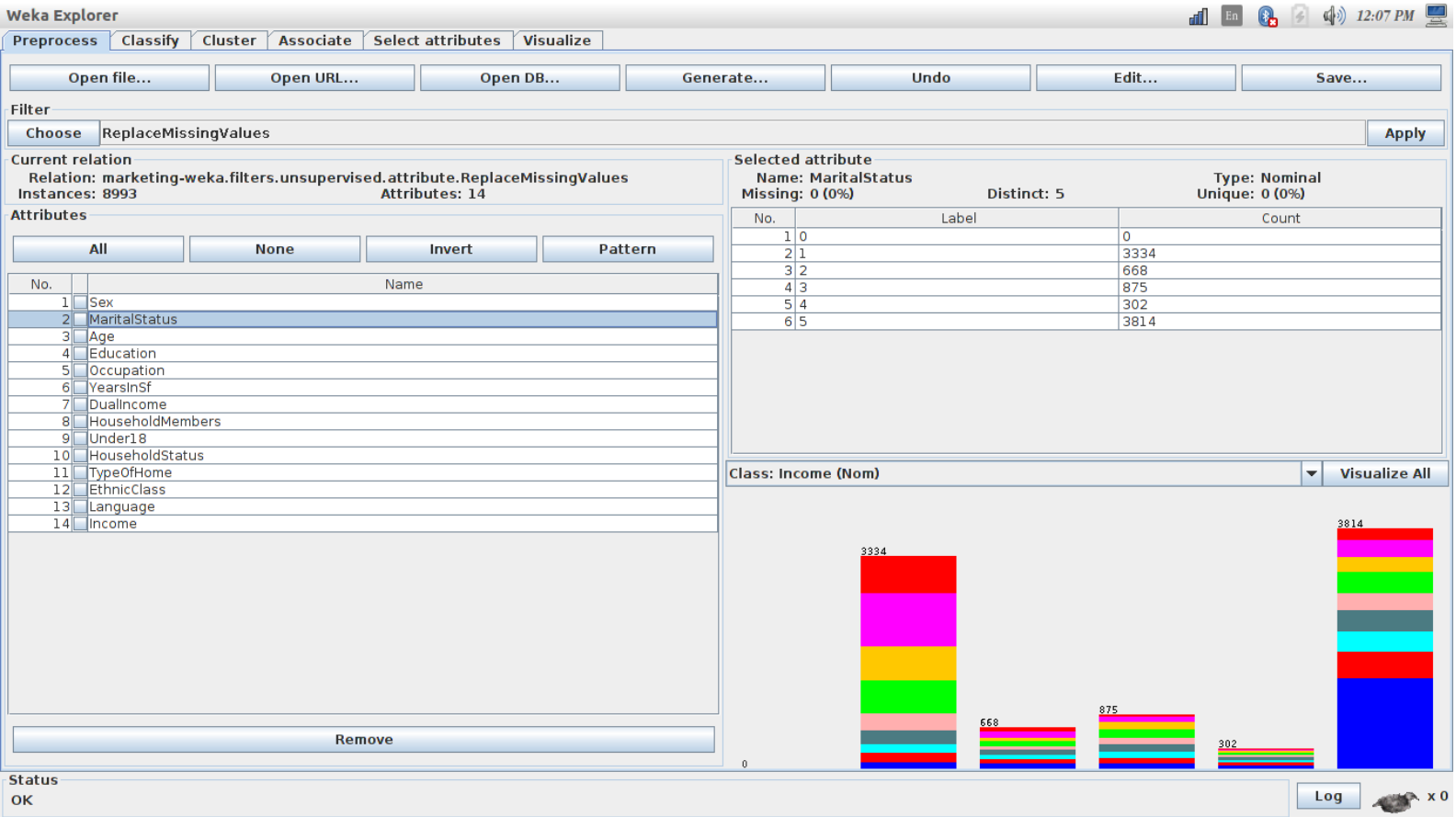


After Applying The Filter :

1)Marital Status

Missing Values : 0 (0%)

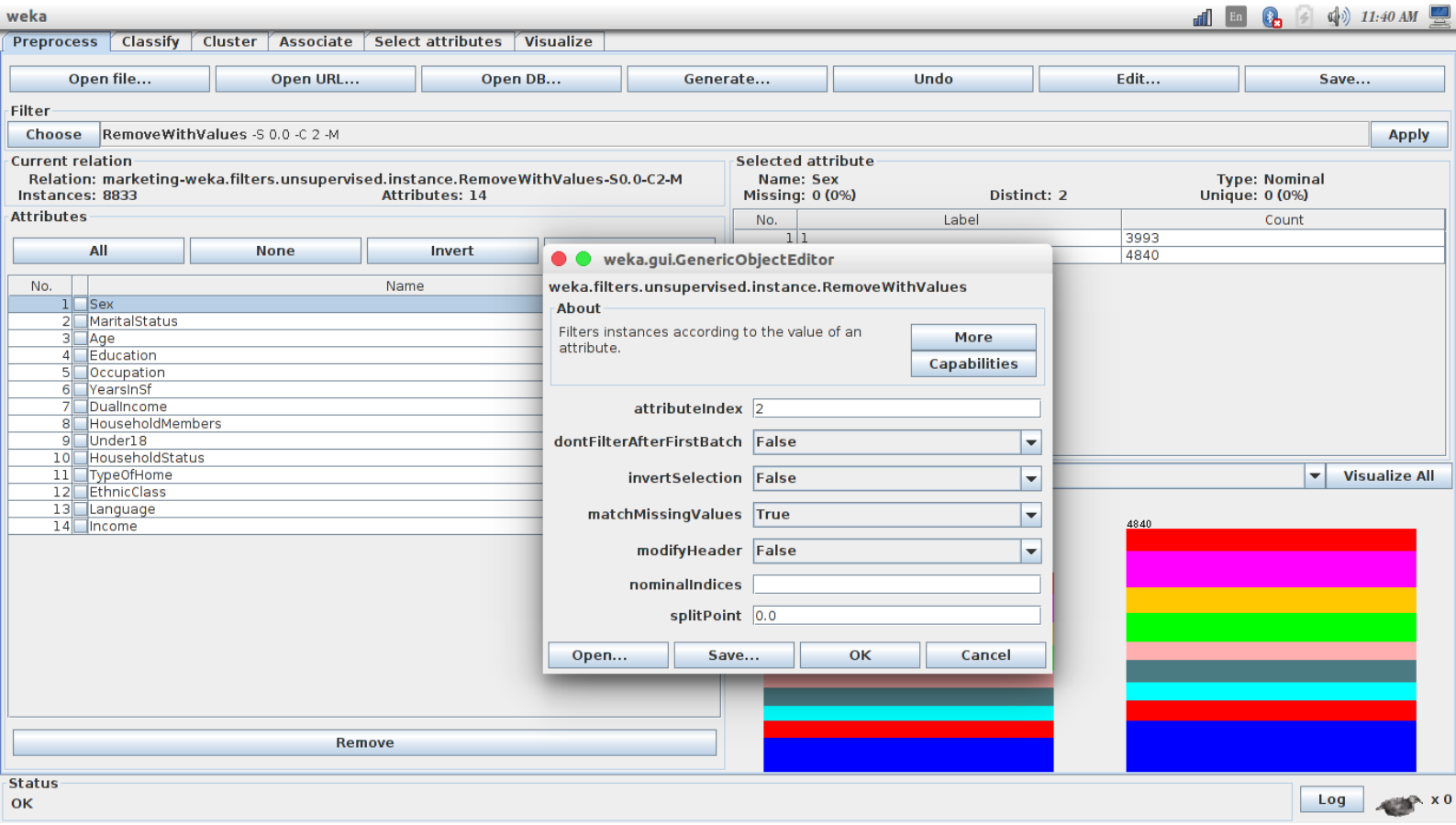
2) Instances : 8993



b) REMOVE WITH VALUES

eg:

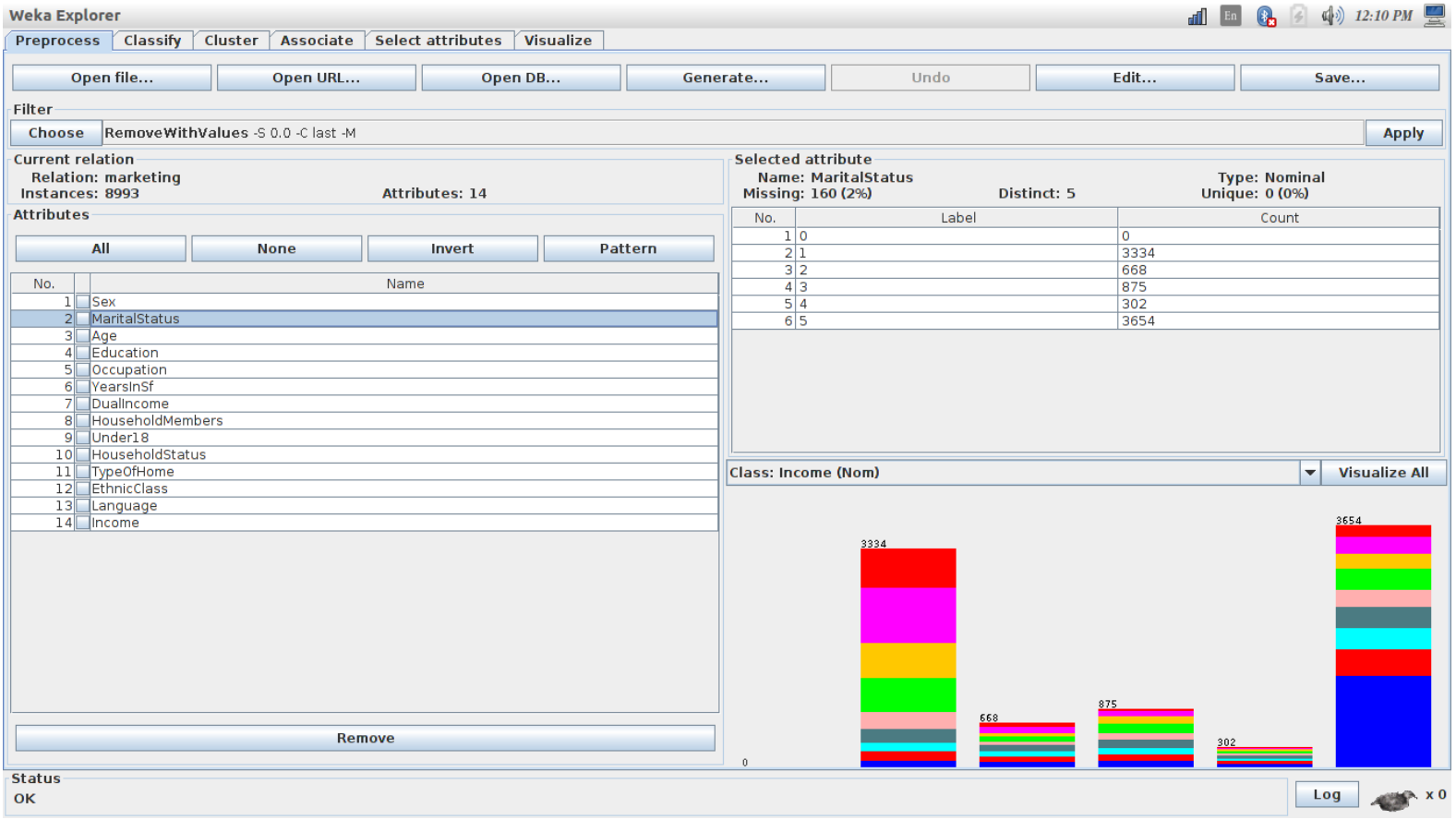
Applying the Remove with Value to 2 . marital Status



Before Applying The Filter :

1)Marital Status
Missing Values : 160 (2%)

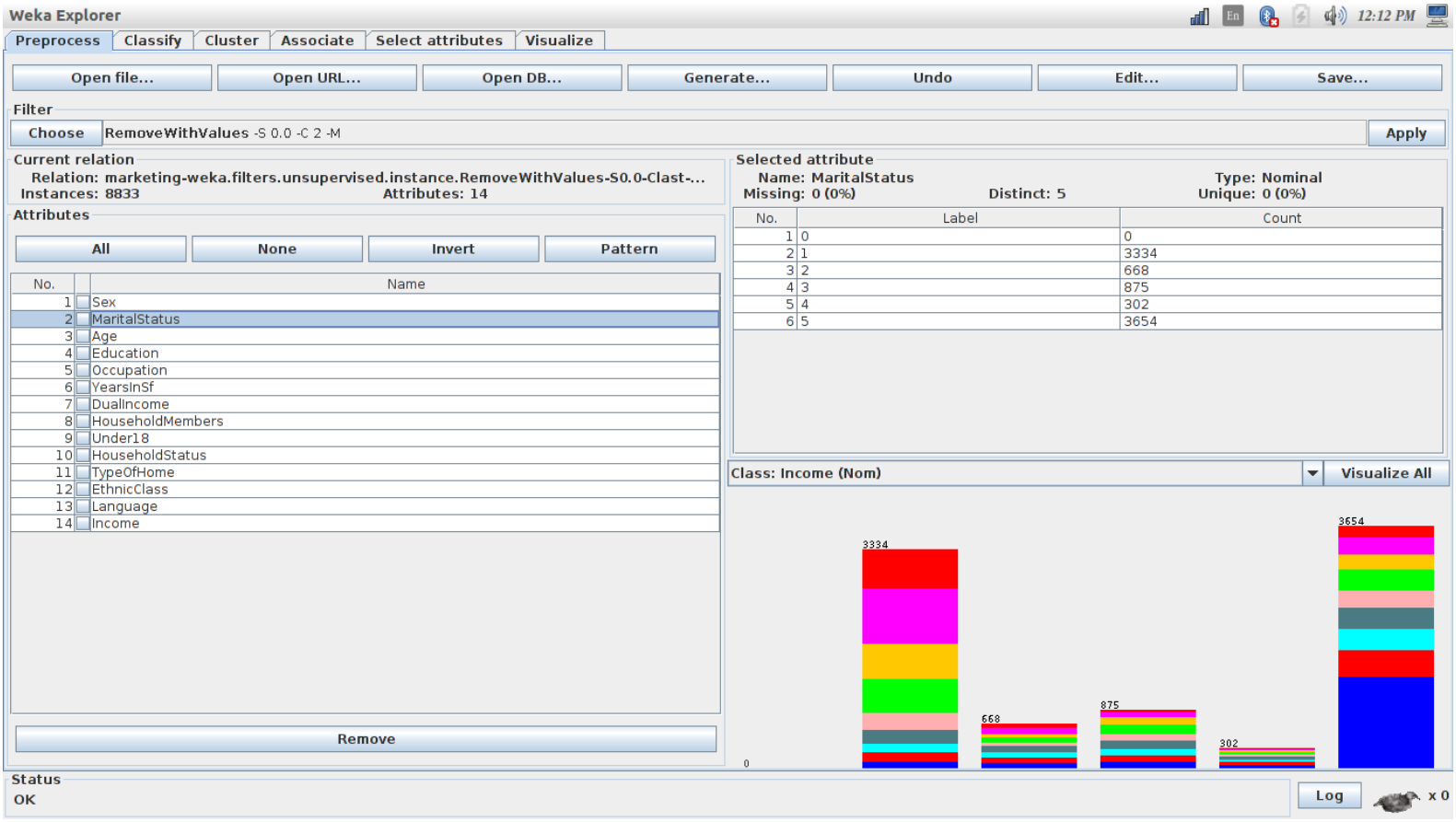
2) Instances : 8993



After Applying The Filter :

1)Marital Status
Missing Values : 0 (0%)

2) Instances : 8833



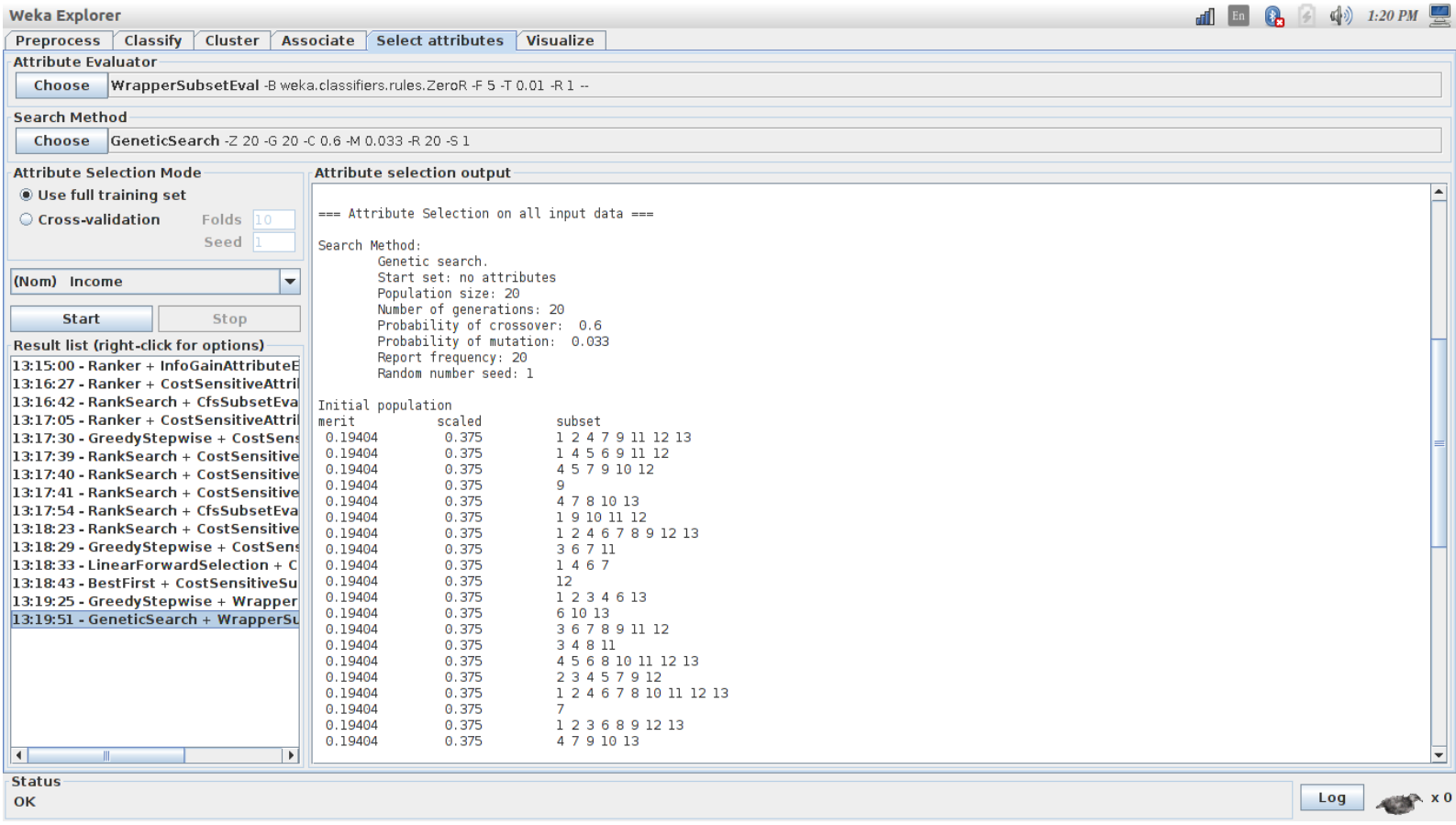
(Proceed with the option relplace missing Values)

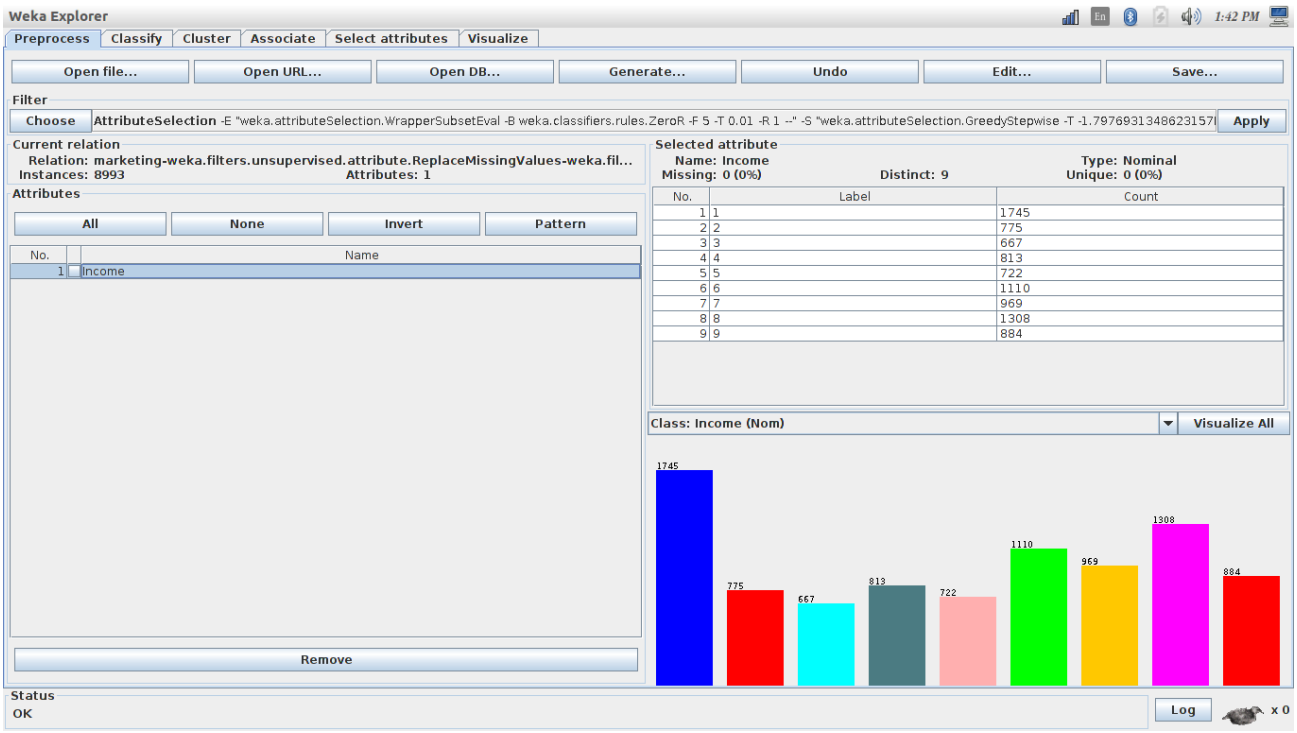
Step 02.

Demonstrate Attribute Filter option of WEKA

01) Wrapper Subset Eval

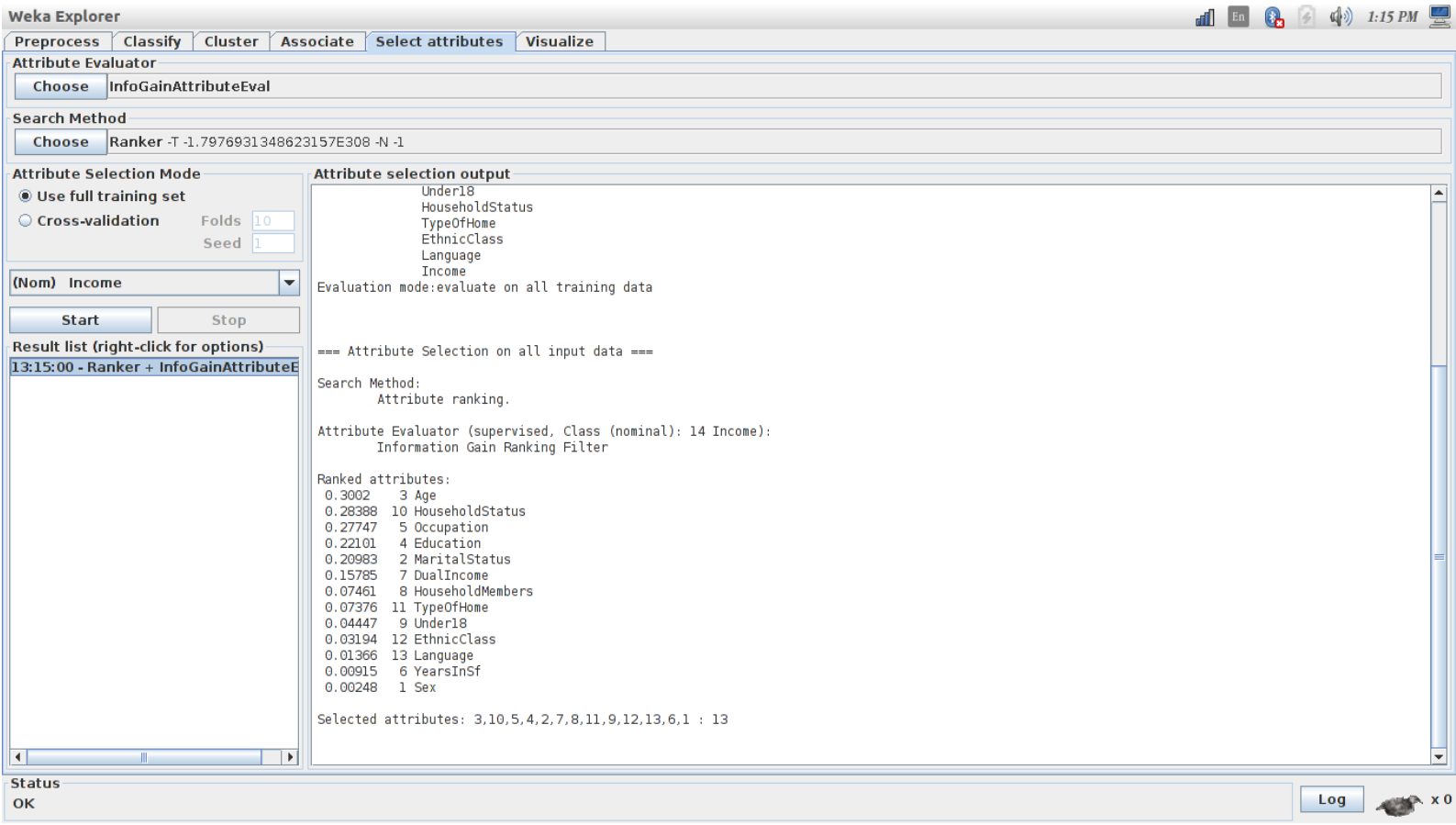
(Evaluates attribute sets by using a learning scheme.)

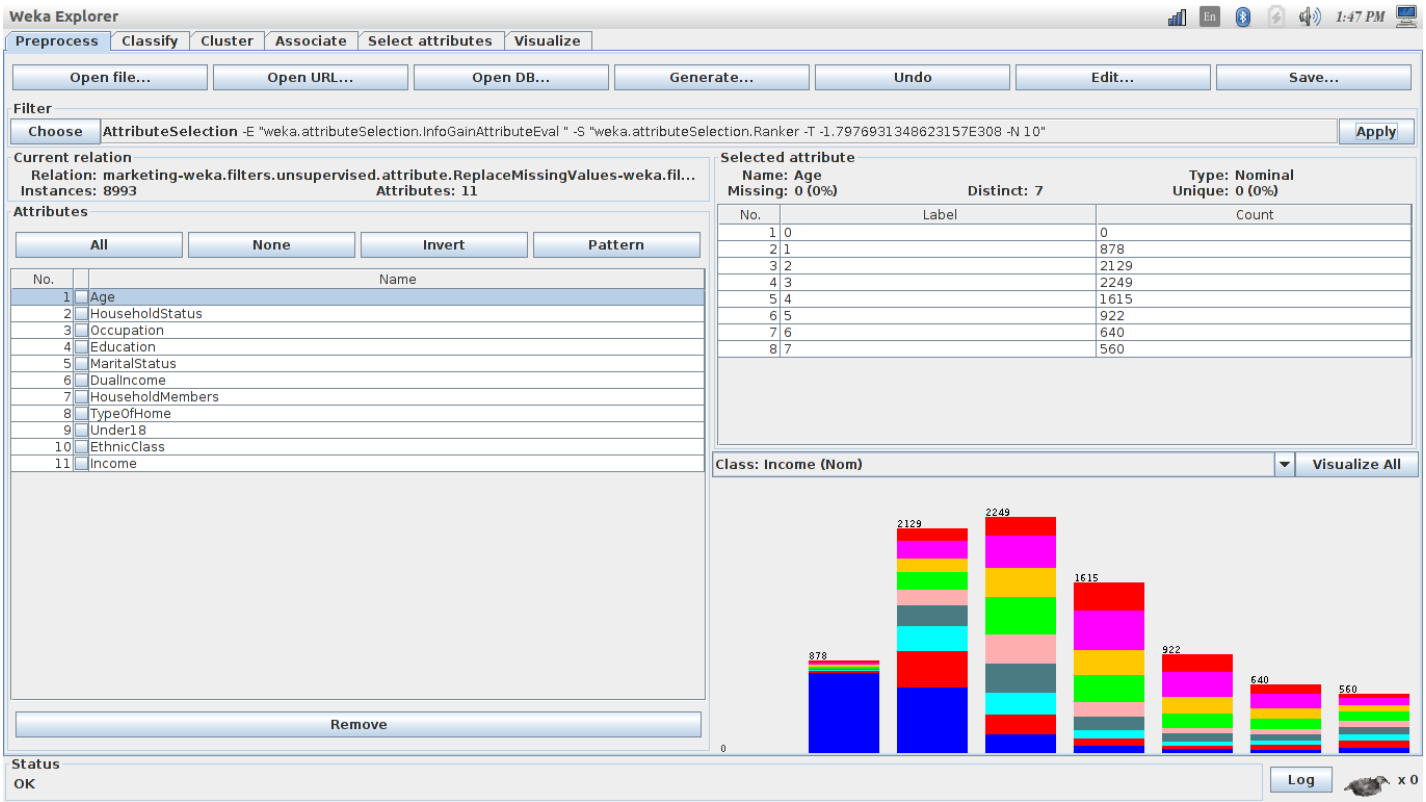




02) Info Gain attribute Eval

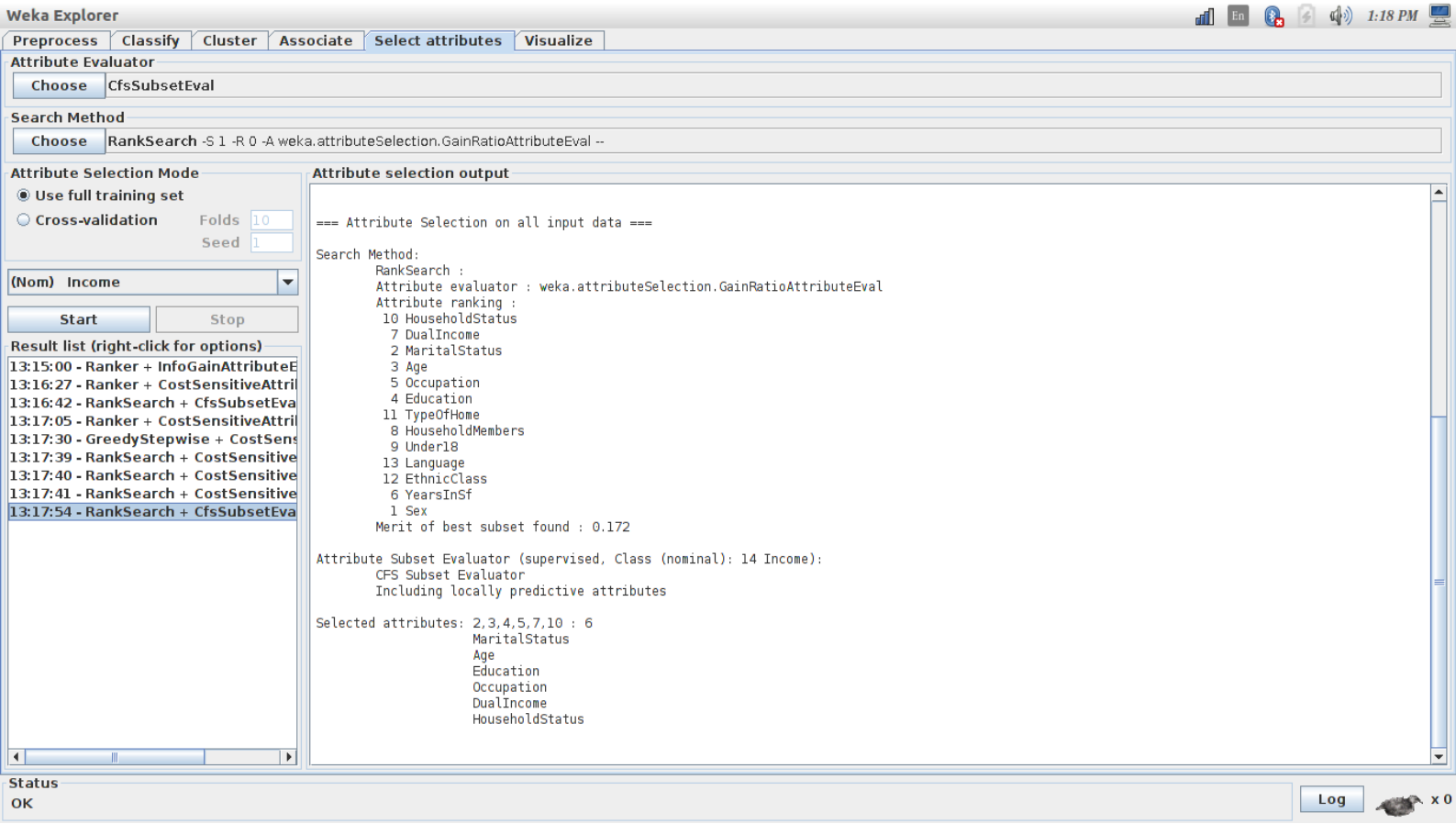
(Evaluates the worth of an attribute by measuring the information gain with respect to the class.)

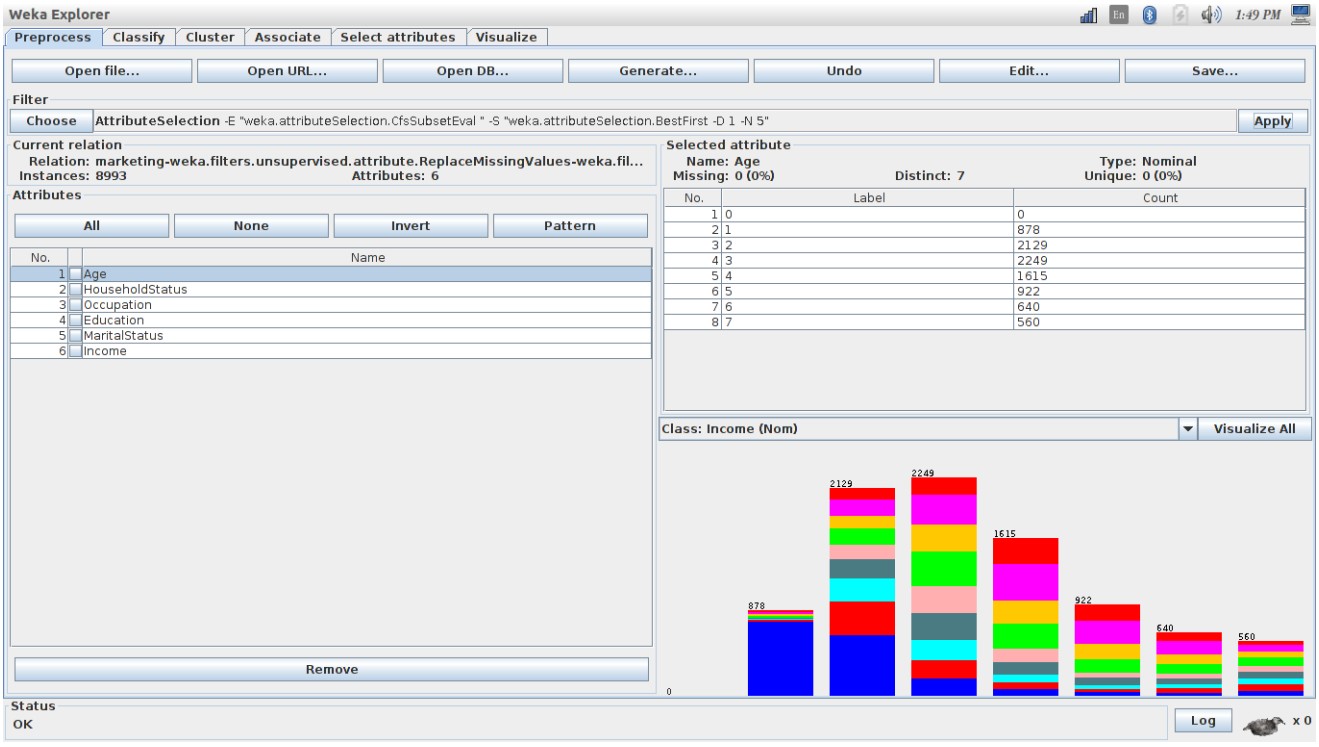




03) Cfs Subset Eval

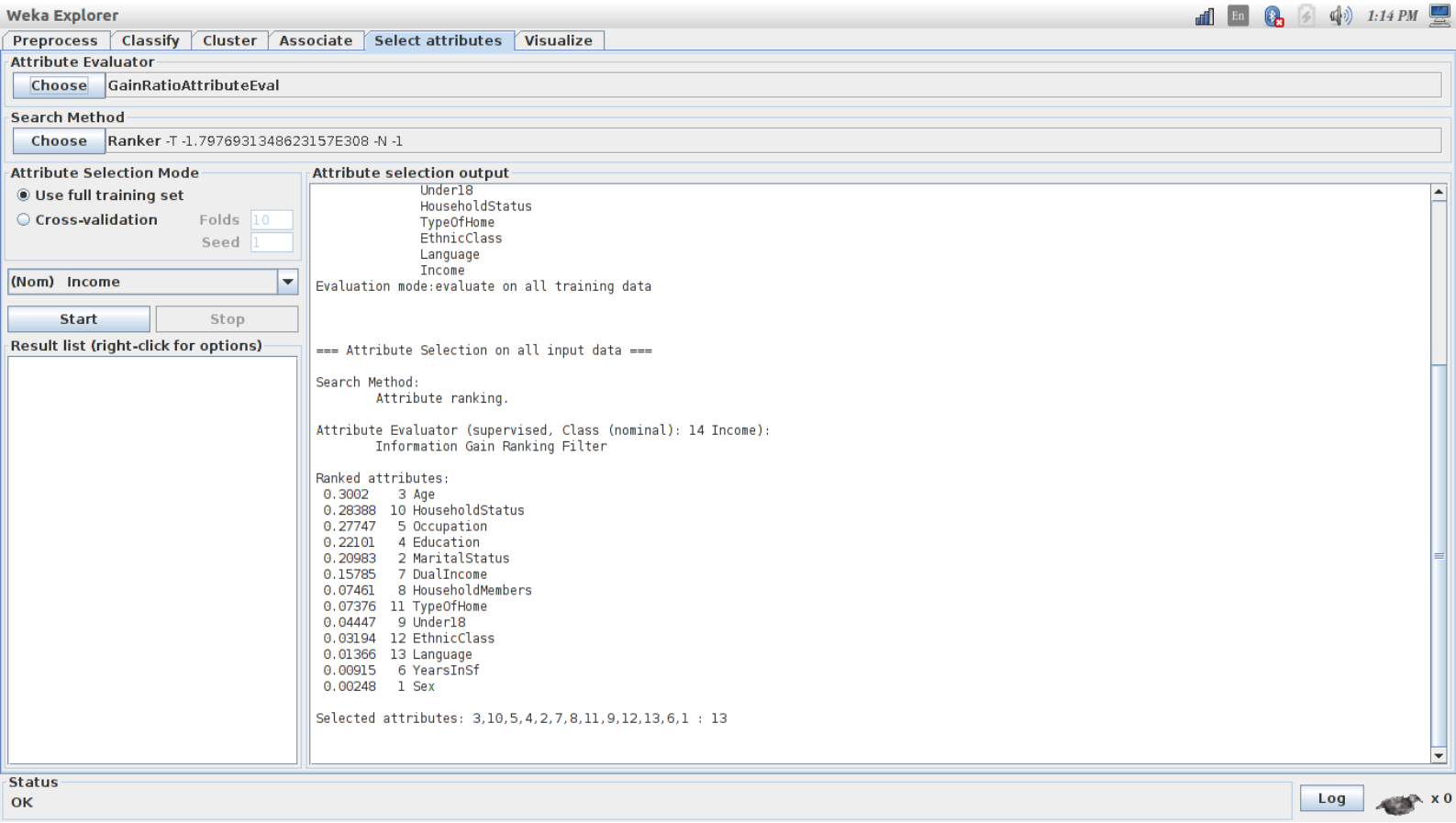
(Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.)

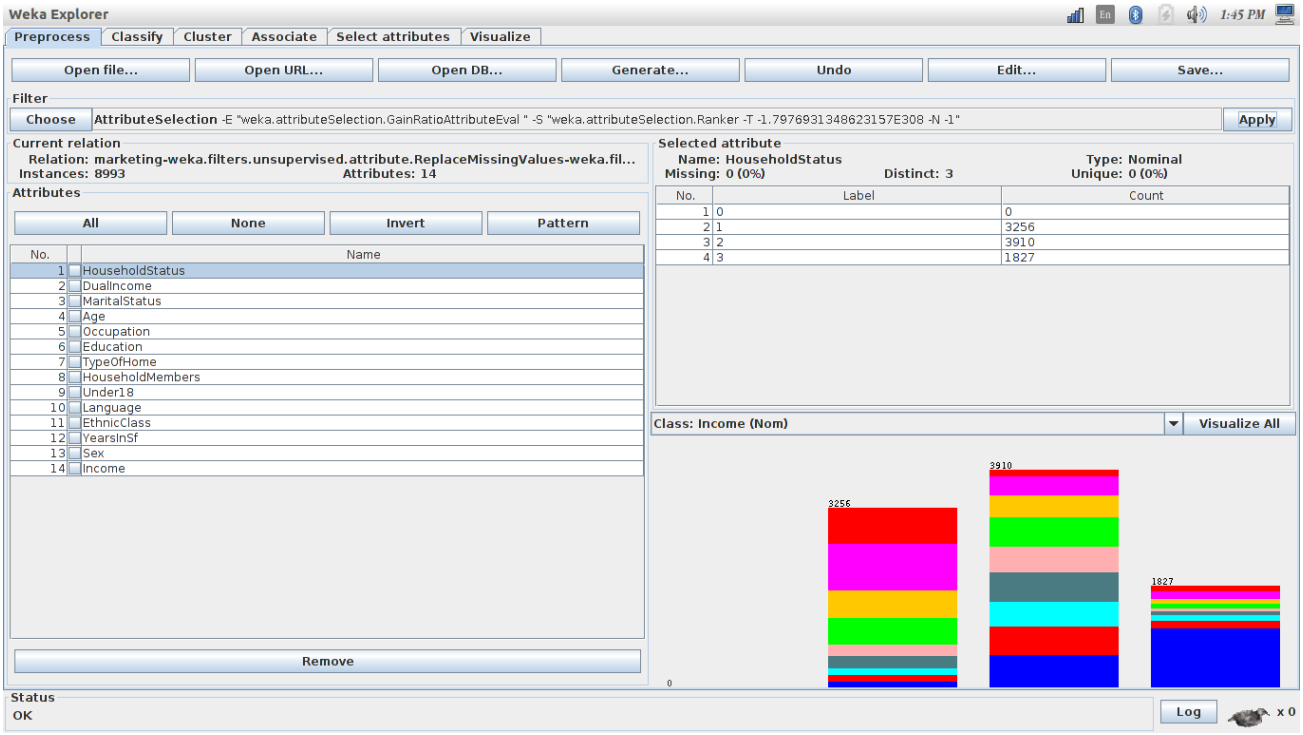




04) Gain Ratio Attribute Level

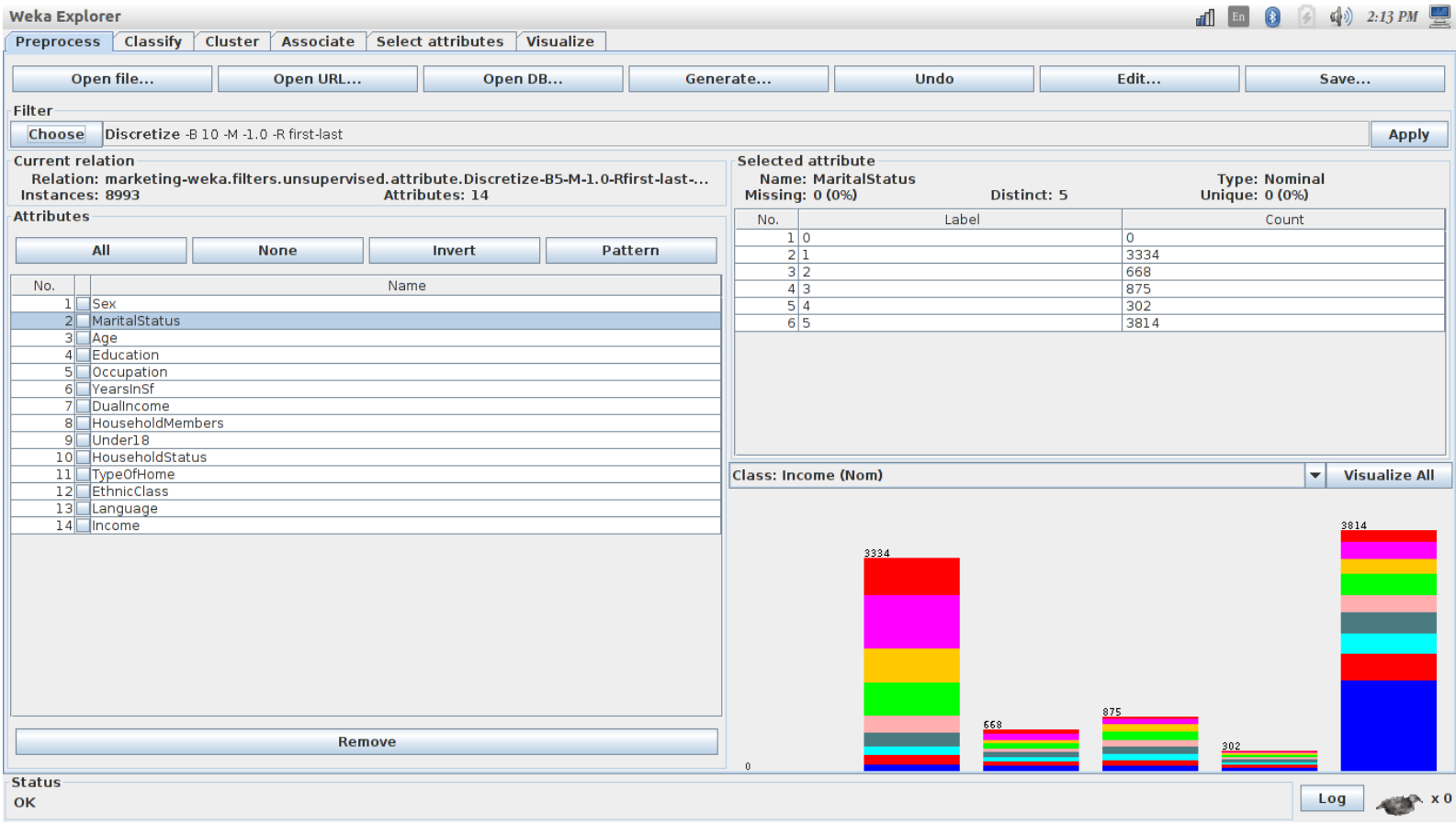
(Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.)



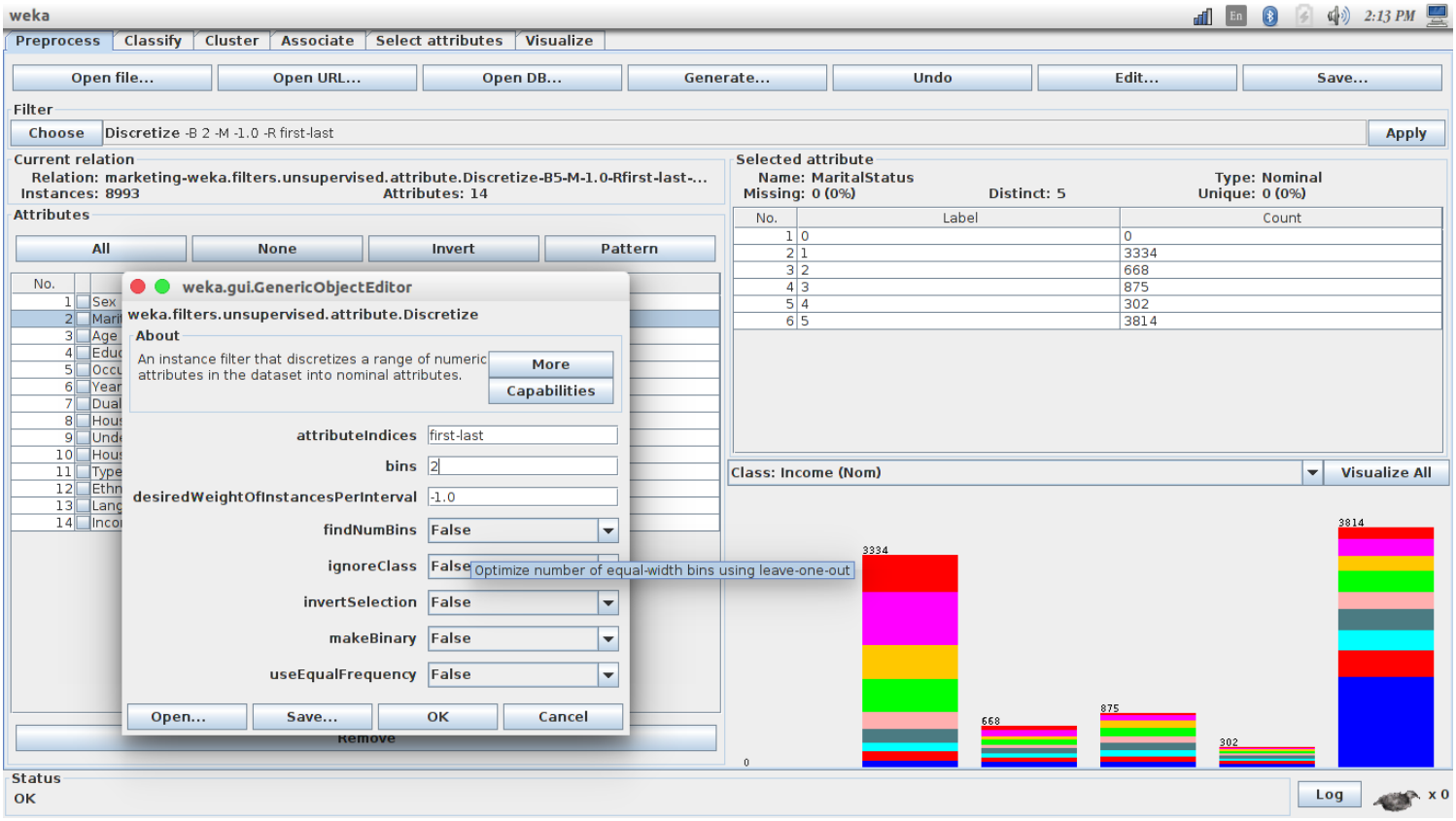


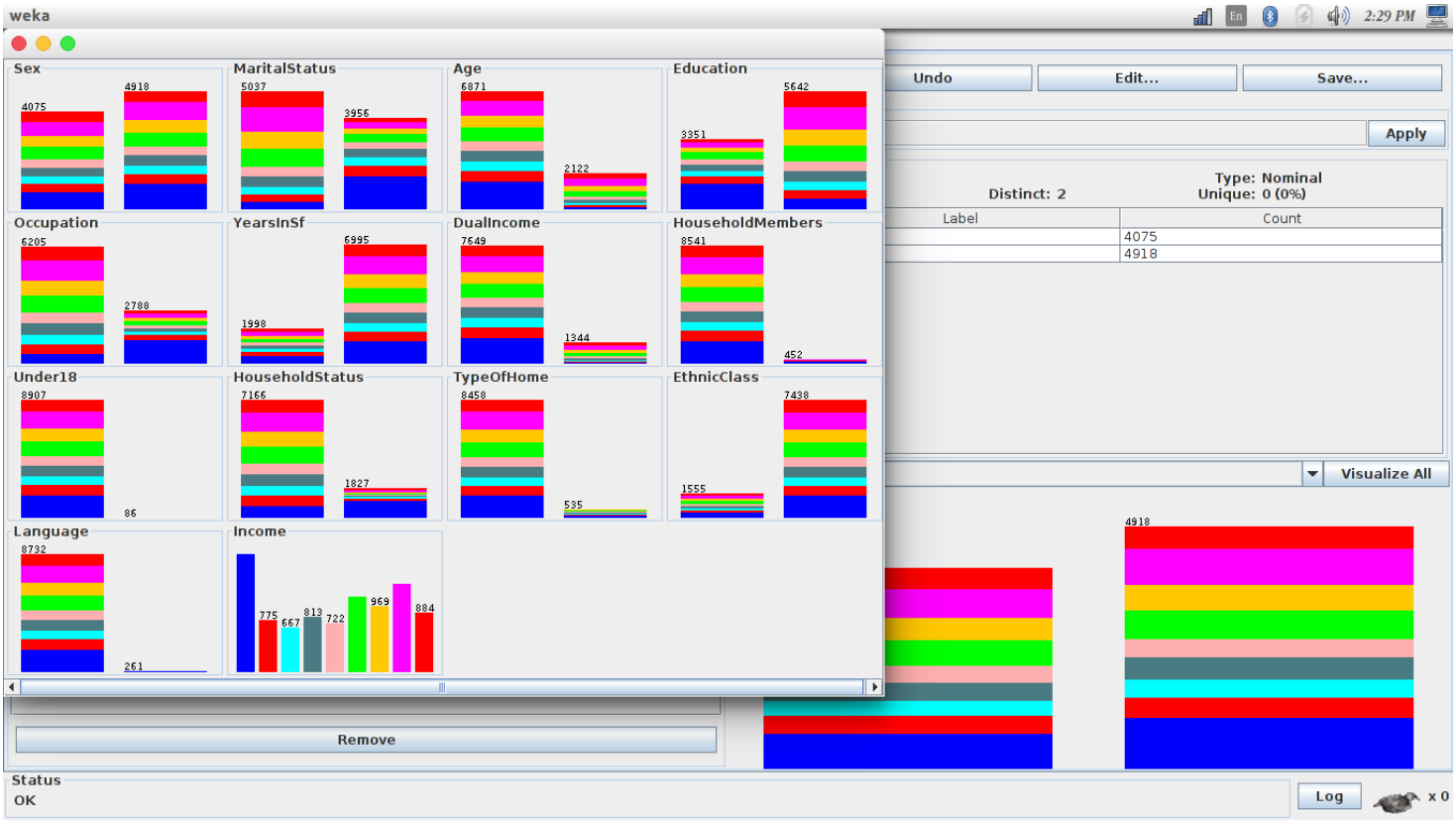
Step 03.

How Discretization can be done with WEKA for the above data

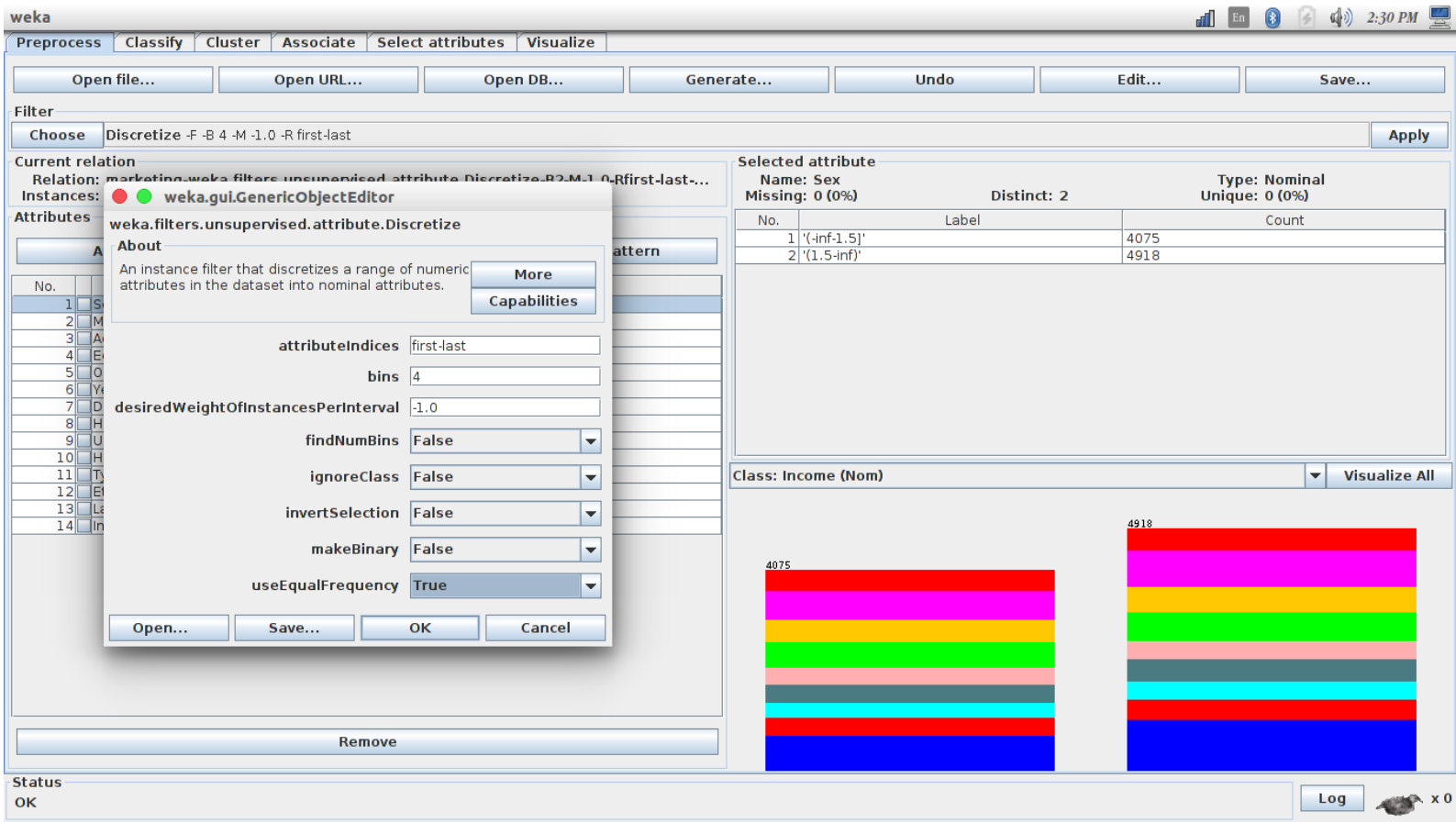


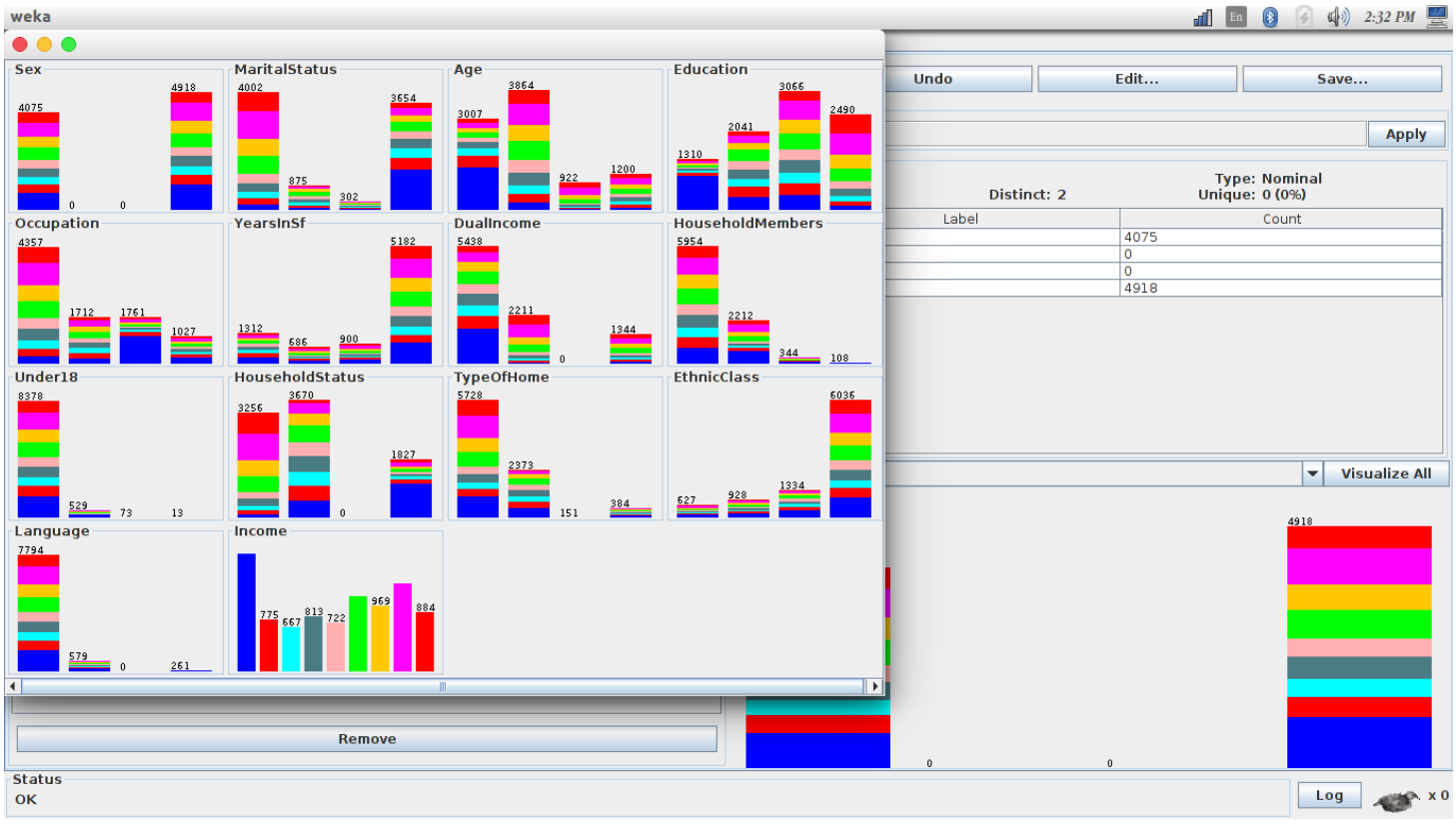
2 Bin .Discretize





4 Bin . Discretize
(Equal Frequency)

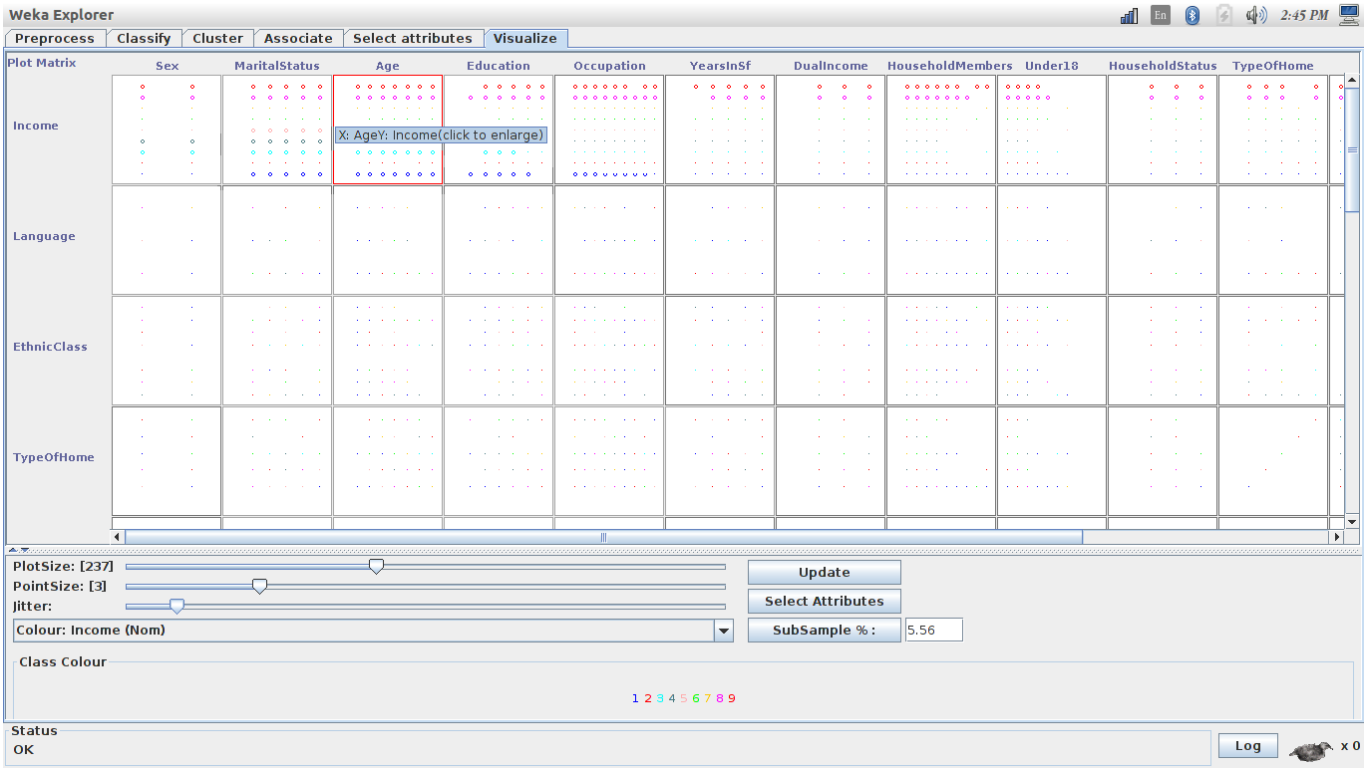


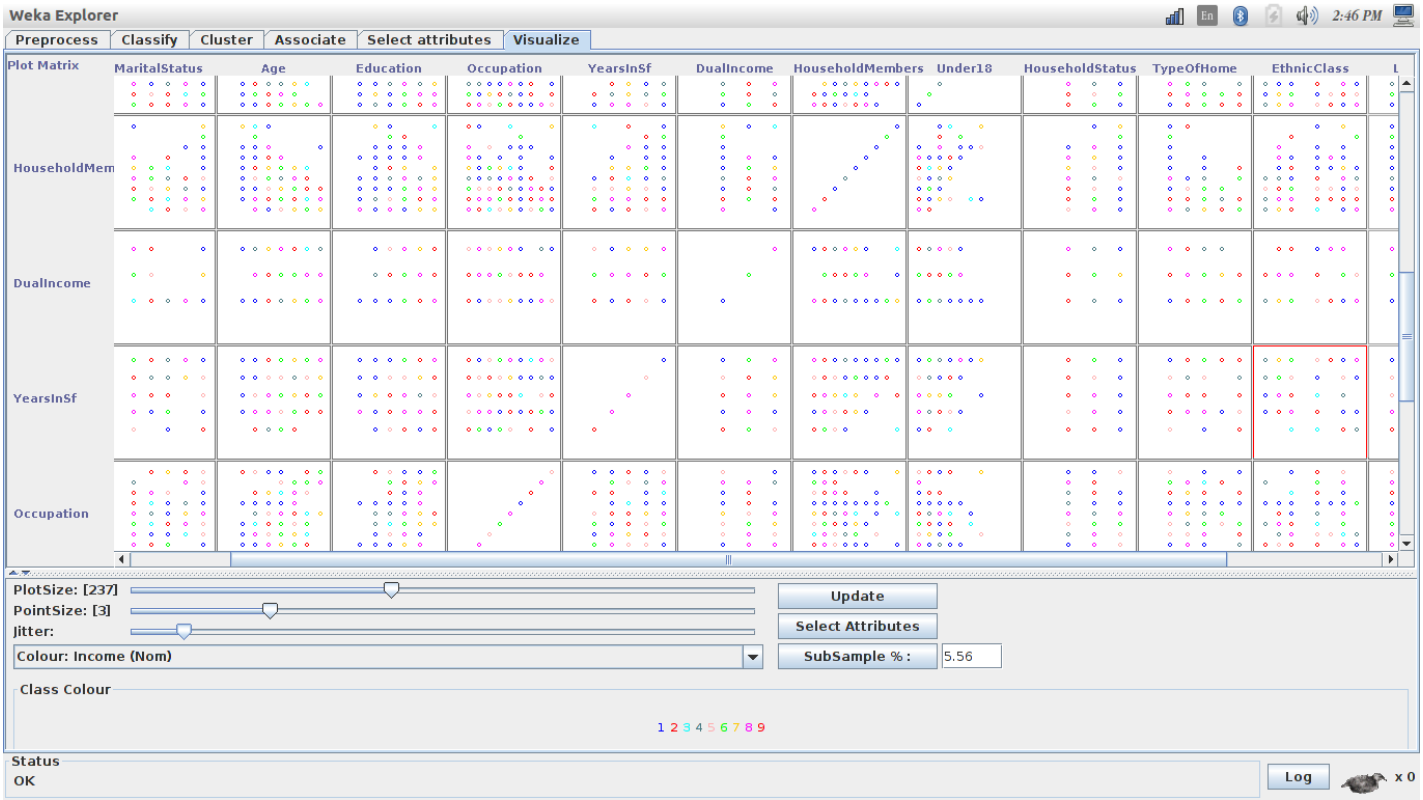


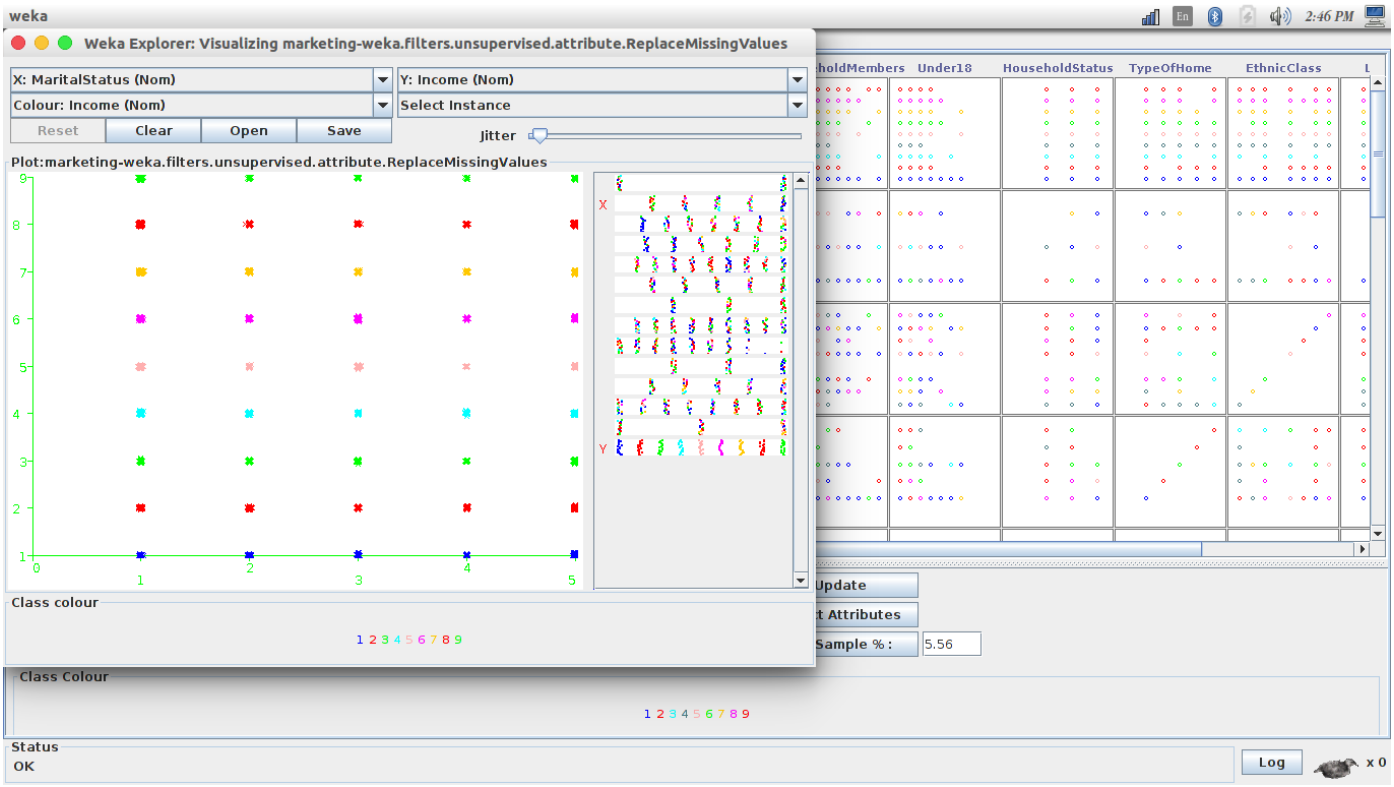
Step 04.

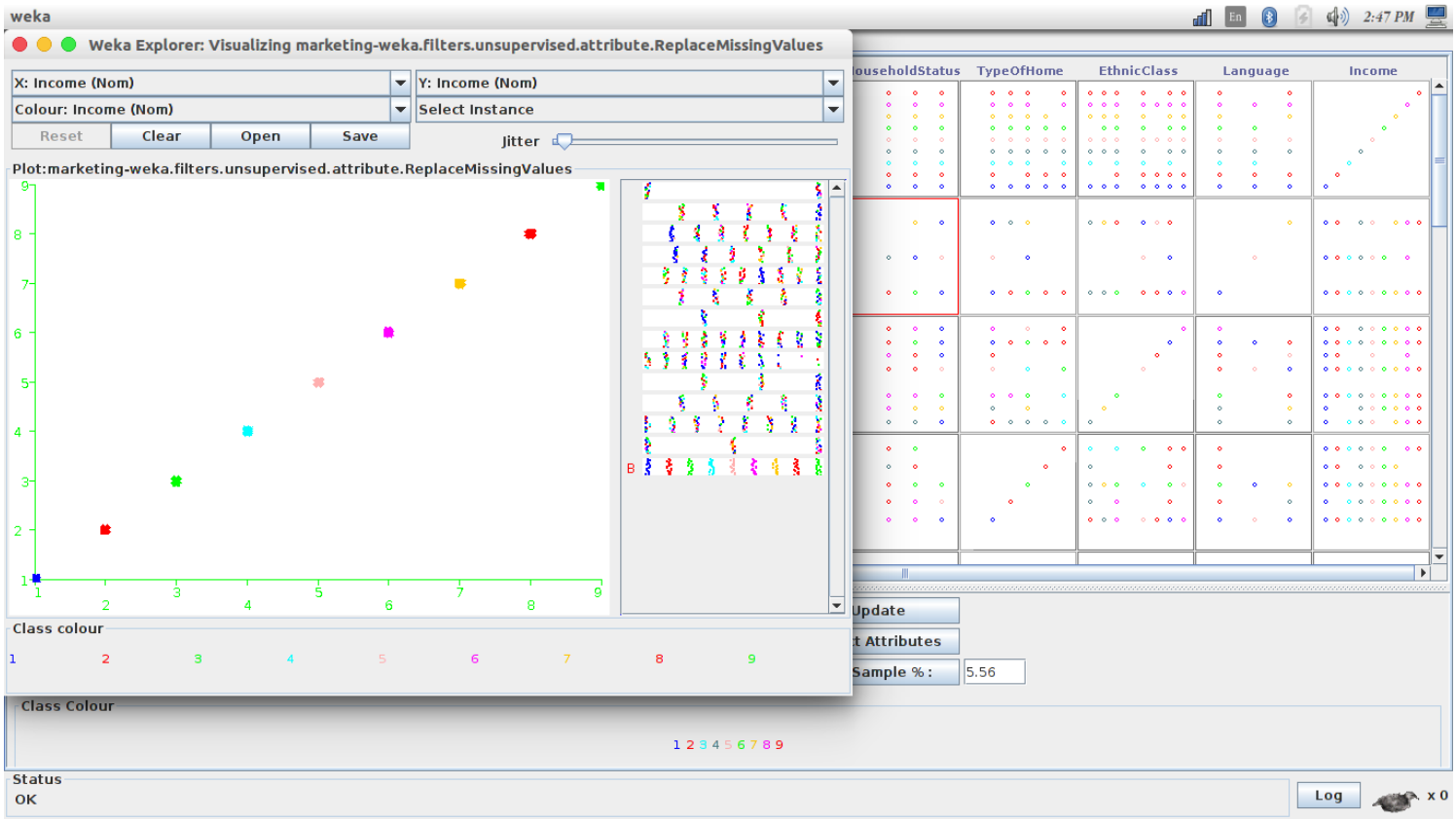
Illustrate various Visualization techniques supported by WEKA for the above data

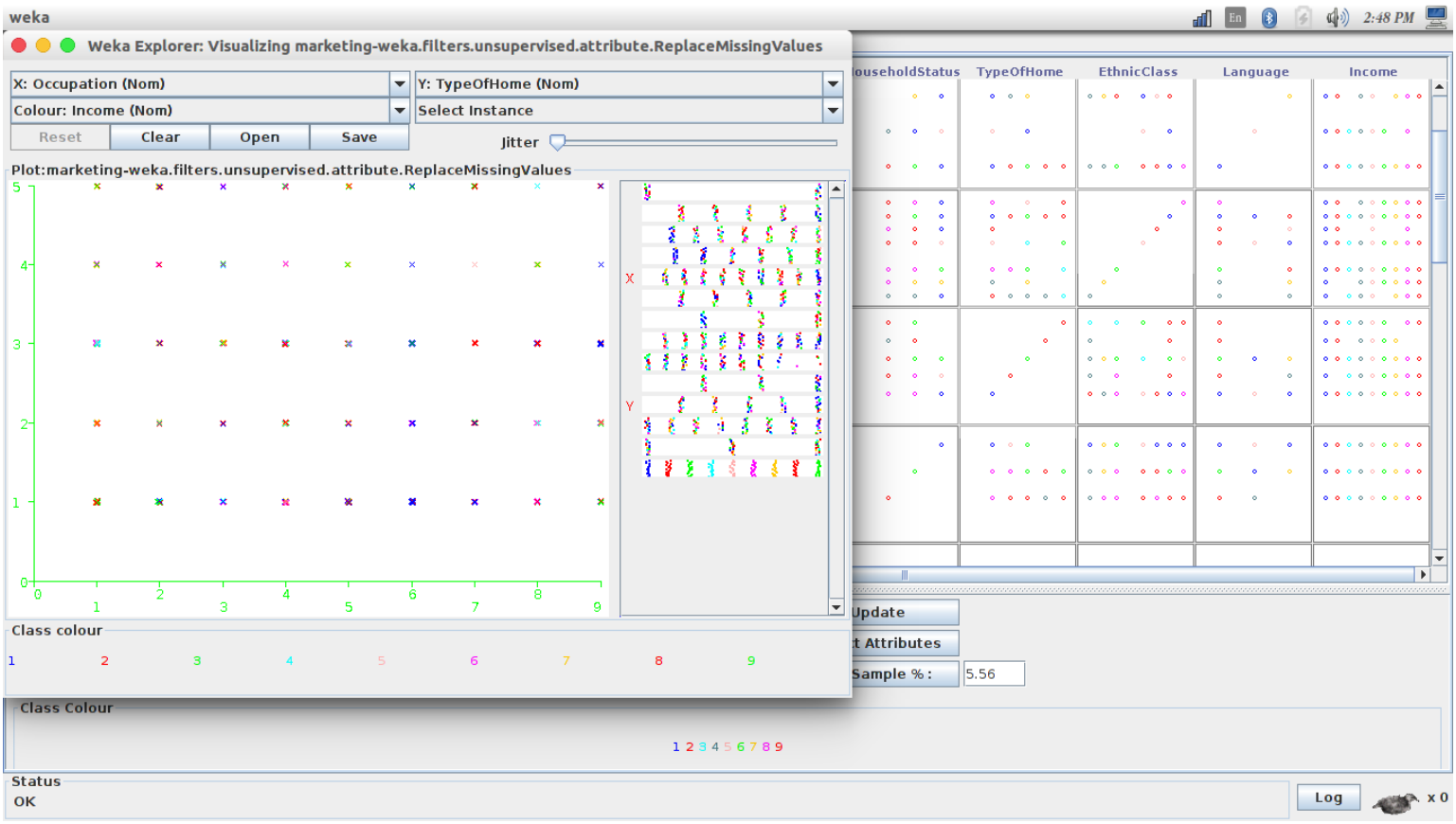
1.BOX PLOT



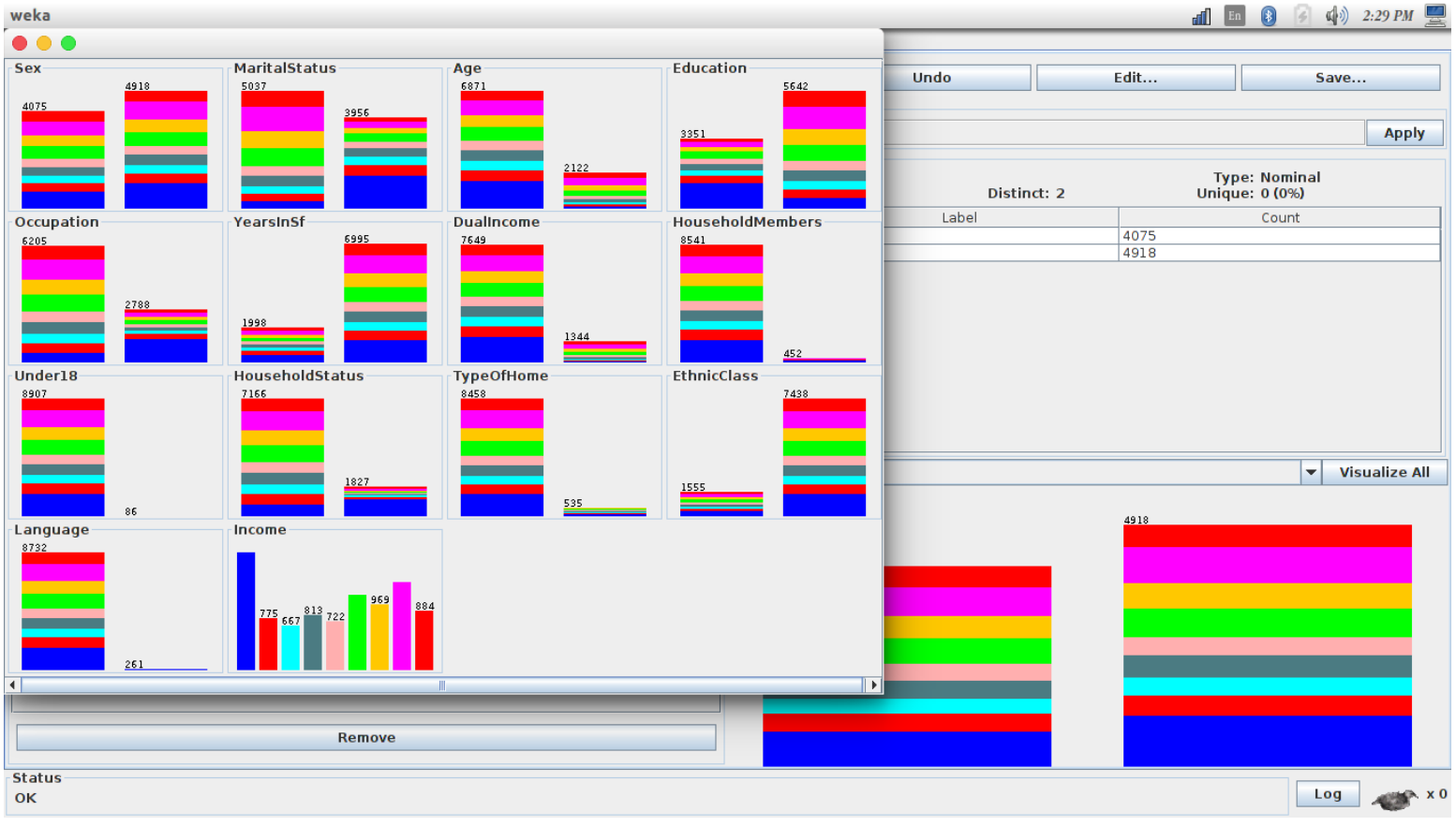



weka
2:46 PM
Weka Explorer: Visualizing marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues
X: MaritalStatus (Nom)
Y: Income (Nom)
Colour: Income (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues
HouseholdStatus
TypeOfHome
EthnicClass
Update
Attributes
Sample % :
5.56
Class colour
Class Colour
Status
OK
Log
x 0


weka
2:47 PM
Weka Explorer: Visualizing marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues
X: Income (Nom)
Y: Income (Nom)
Colour: Income (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues
HouseholdStatus
TypeOfHome
EthnicClass
Language
Income
Update
Attributes
Sample % :
5.56
Class colour
Class Colour
Status
OK
Log
x 0

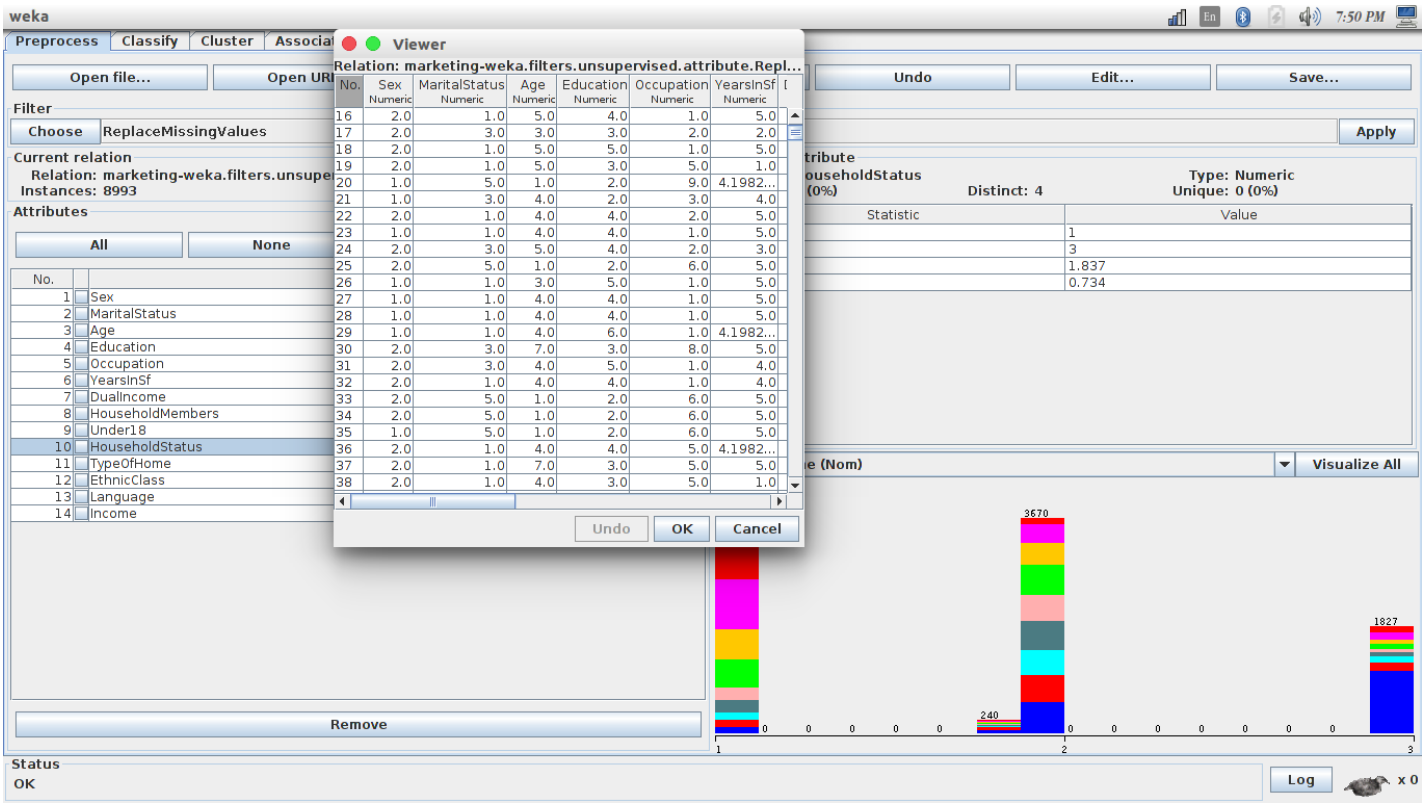

weka
Weka Explorer: Visualizing marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues
X: Occupation (Nom)
Y: TypeOfHome (Nom)
Colour: Income (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues
Class colour
HouseholdStatus
TypeOfHome
EthnicClass
Language
Income
Update
Attributes
Sample % :
5.56
Class Colour
Status
OK
Log
x 0

2.HISTOGRAM

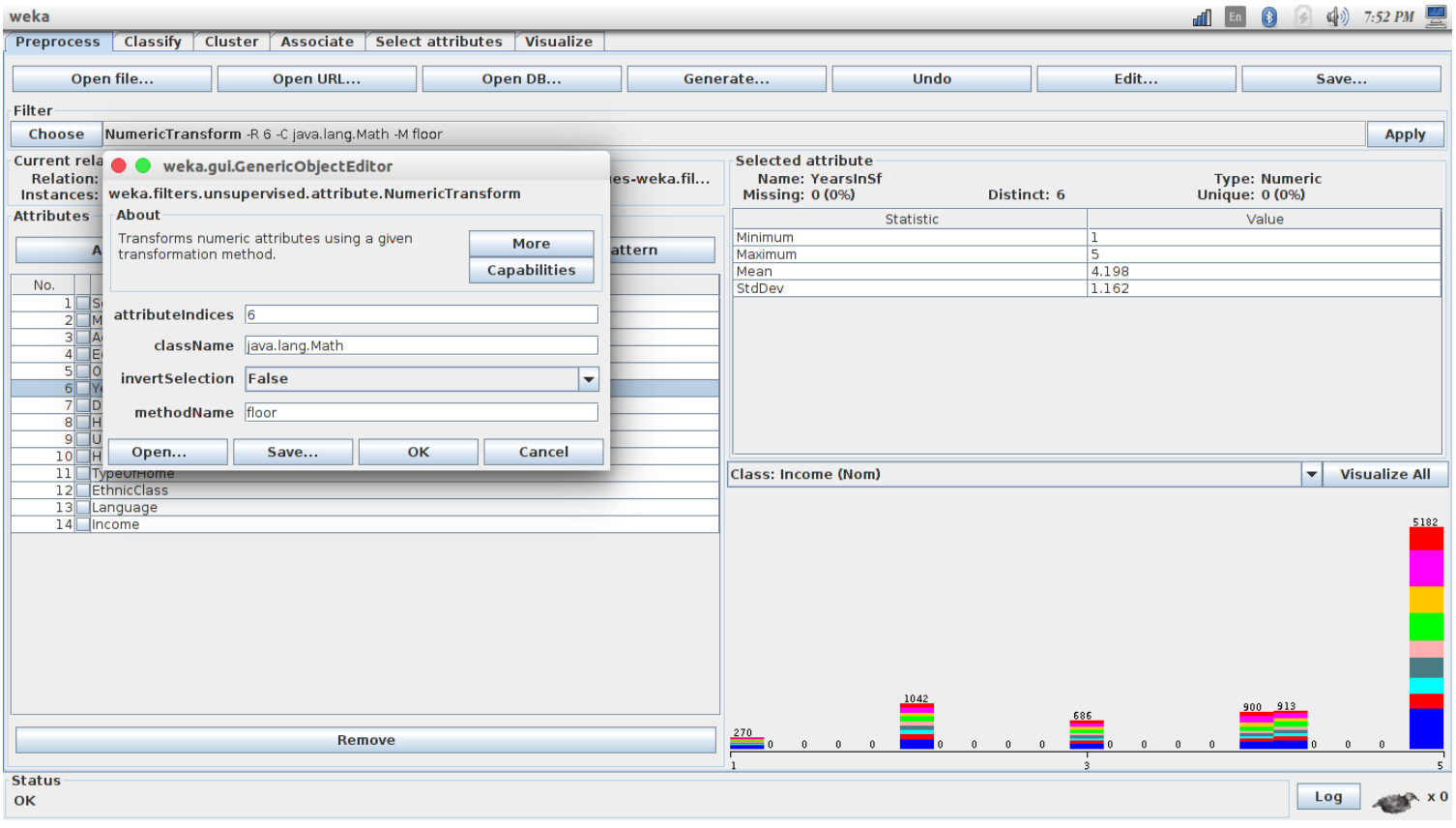

weka
Sex
MaritalStatus
Age
Education
Occupation
YearsInSf
DualIncome
HouseholdMembers
Under18
HouseholdStatus
TypeOfHome
EthnicClass
Language
Income
Undo
Edit...
Save...
Apply
Distinct: 2
Type: Nominal
Unique: 0 (0%)
Label
Count
4075
4918
Visualize All
Remove
Status
OK
Log
x 0

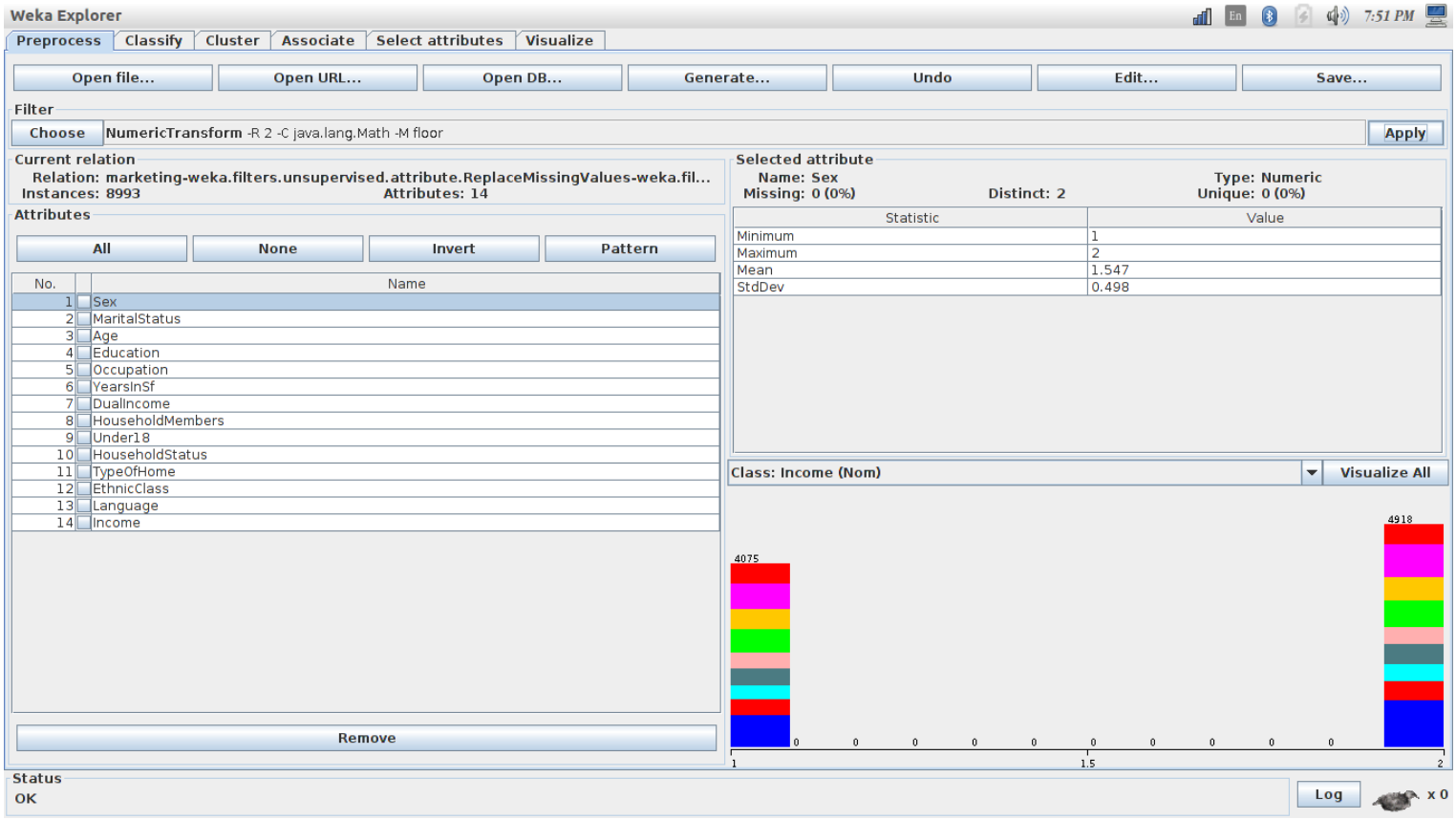
Step 05.

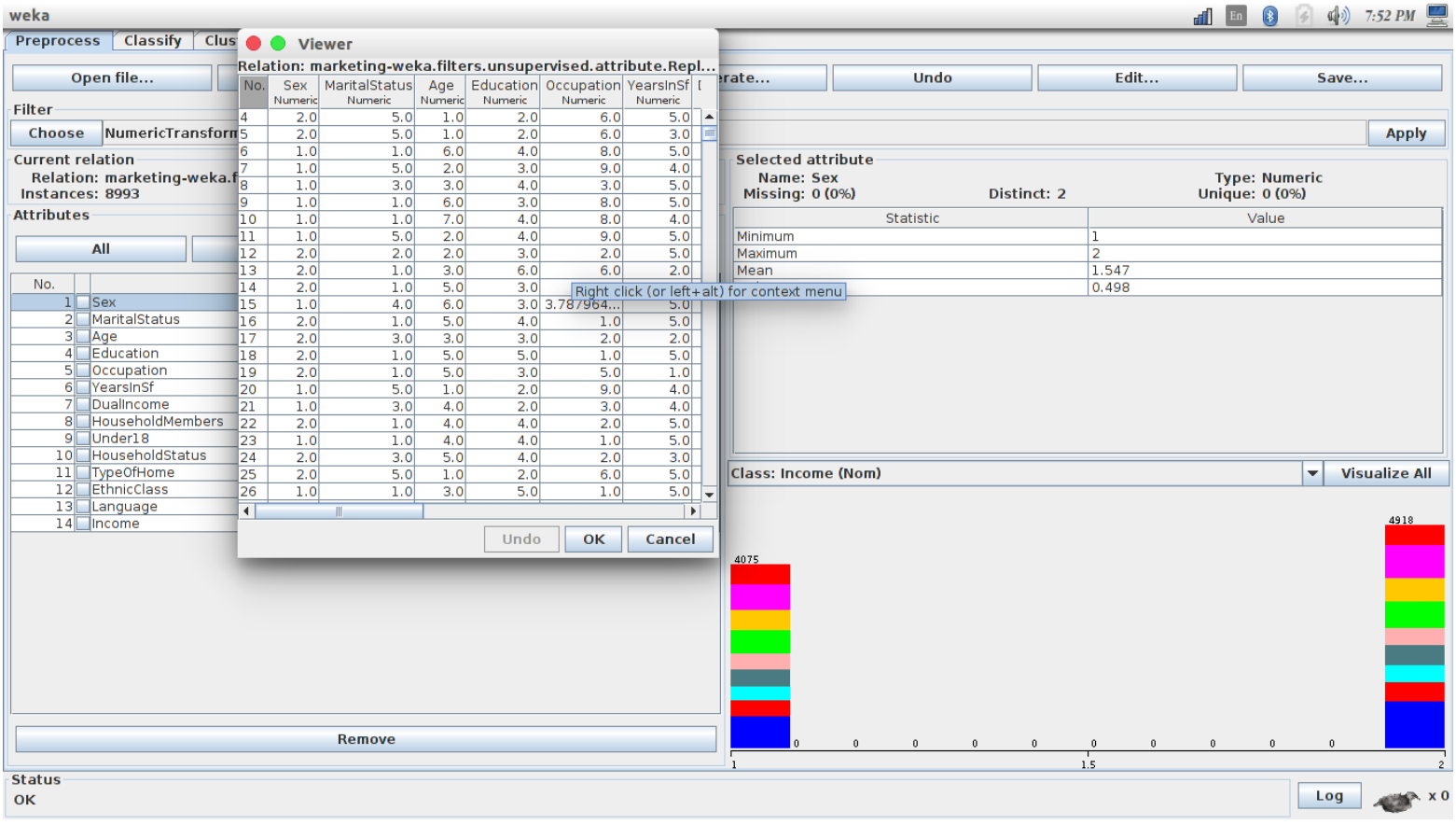
Show how data transformation is supported by WEKA.



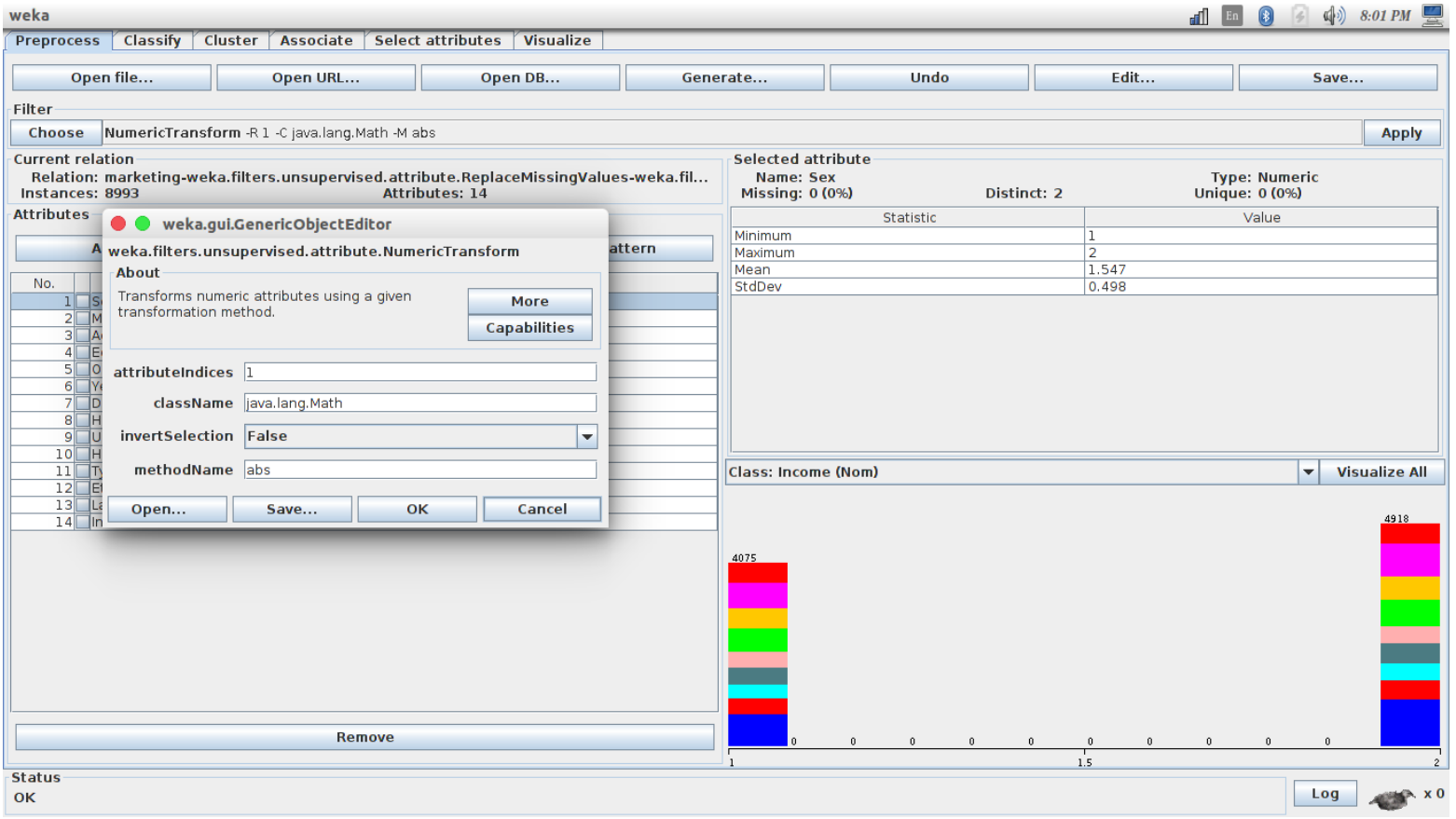
1) Floor



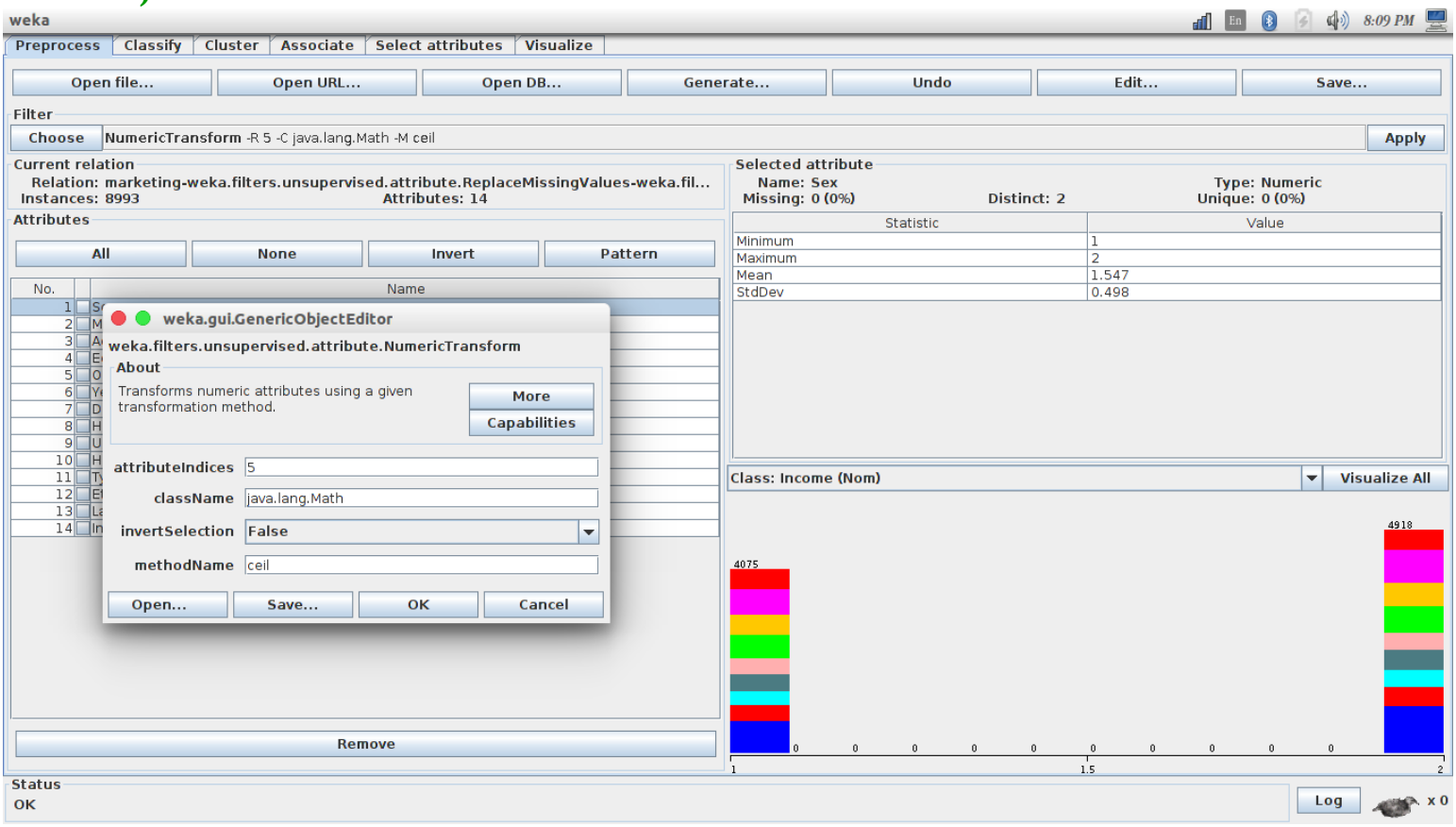

Weka Explorer
7:51 PM
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
NumericTransform -R 2 -C java.lang.Math -M floor
Apply
Current relation
Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...
Instances: 8993
Attributes: 14
Attributes
All
None
Invert
Pattern
No.
Name
1 Sex
2 MaritalStatus
3 Age
4 Education
5 Occupation
6 YearsInSf
7 DualIncome
8 HouseholdMembers
9 Under18
10 HouseholdStatus
11 TypeOfHome
12 EthnicClass
13 Language
14 Income
Remove
Selected attribute
Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum 1
Maximum 2
Mean 1.547
StdDev 0.498
Class: Income (Nom)
Visualize All
4918
4075
0 0 0 0 0 0 0 0 0 0
1
1.5
2
Status
OK
Log
x 0


weka
7:52 PM
Preprocess
Classify
Open file...
Undo
Edit...
Save...
Filter
Choose
NumericTransform
Apply
Current relation
Relation: marketing-weka.f
Instances: 8993
Attributes
All
No.
1 Sex
2 MaritalStatus
3 Age
4 Education
5 Occupation
6 YearsInSf
7 DualIncome
8 HouseholdMembers
9 Under18
10 HouseholdStatus
11 TypeOfHome
12 EthnicClass
13 Language
14 Income
Remove
Viewer
Relation: marketing-weka.filters.unsupervised.attribute.Repl...
No. Sex Numeric MaritalStatus Numeric Age Numeric Education Numeric Occupation Numeric YearsInSf Numeric
4 2.0 5.0 1.0 2.0 6.0 5.0
5 2.0 5.0 1.0 2.0 6.0 3.0
6 1.0 1.0 6.0 4.0 8.0 5.0
7 1.0 5.0 2.0 3.0 9.0 4.0
8 1.0 3.0 3.0 4.0 3.0 5.0
9 1.0 1.0 6.0 3.0 8.0 5.0
10 1.0 1.0 7.0 4.0 8.0 4.0
11 1.0 5.0 2.0 4.0 9.0 5.0
12 2.0 2.0 2.0 3.0 2.0 5.0
13 2.0 1.0 3.0 6.0 6.0 2.0
14 2.0 1.0 5.0 3.0
15 1.0 4.0 6.0 3.0 3.787964... 5.0
16 2.0 1.0 5.0 4.0 1.0 5.0
17 2.0 3.0 3.0 3.0 2.0 2.0
18 2.0 1.0 5.0 5.0 1.0 5.0
19 2.0 1.0 5.0 3.0 5.0 1.0
20 1.0 5.0 1.0 2.0 9.0 4.0
21 1.0 3.0 4.0 2.0 3.0 4.0
22 2.0 1.0 4.0 4.0 2.0 5.0
23 1.0 1.0 4.0 4.0 1.0 5.0
24 2.0 3.0 5.0 4.0 2.0 3.0
25 2.0 5.0 1.0 2.0 6.0 5.0
26 1.0 1.0 3.0 5.0 1.0 5.0
Right click (or left+alt) for context menu
Undo
OK
Cancel
Selected attribute
Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum 1
Maximum 2
Mean 1.547
0.498
Class: Income (Nom)
Visualize All
4918
4075
0 0 0 0 0 0 0 0 0 0
1
1.5
2
Status
OK
Log
x 0

2) abs


weka
8:01 PM
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
NumericTransform -R 1 -C java.lang.Math -M abs
Apply
Current relation
Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...
Instances: 8993
Attributes: 14
Attributes
Selected attribute
Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum 1
Maximum 2
Mean 1.547
StdDev 0.498
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.NumericTransform
About
Transforms numeric attributes using a given transformation method.
More
Capabilities
attributeIndices 1
className java.lang.Math
invertSelection False
methodName abs
Open...
Save...
OK
Cancel
Class: Income (Nom)
Visualize All
4918
4075
Remove
Status
OK
Log
x 0

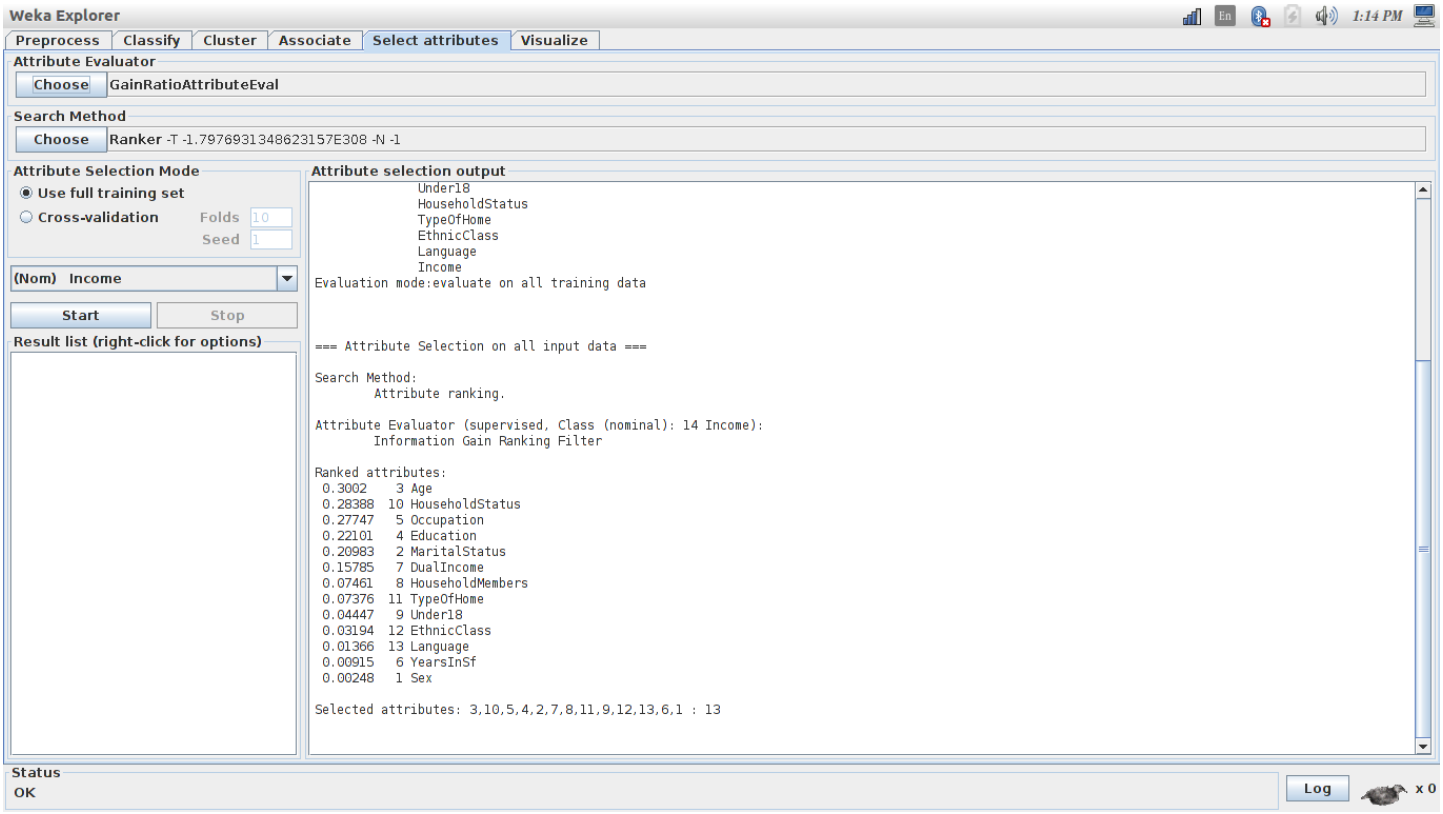
3) Ceil

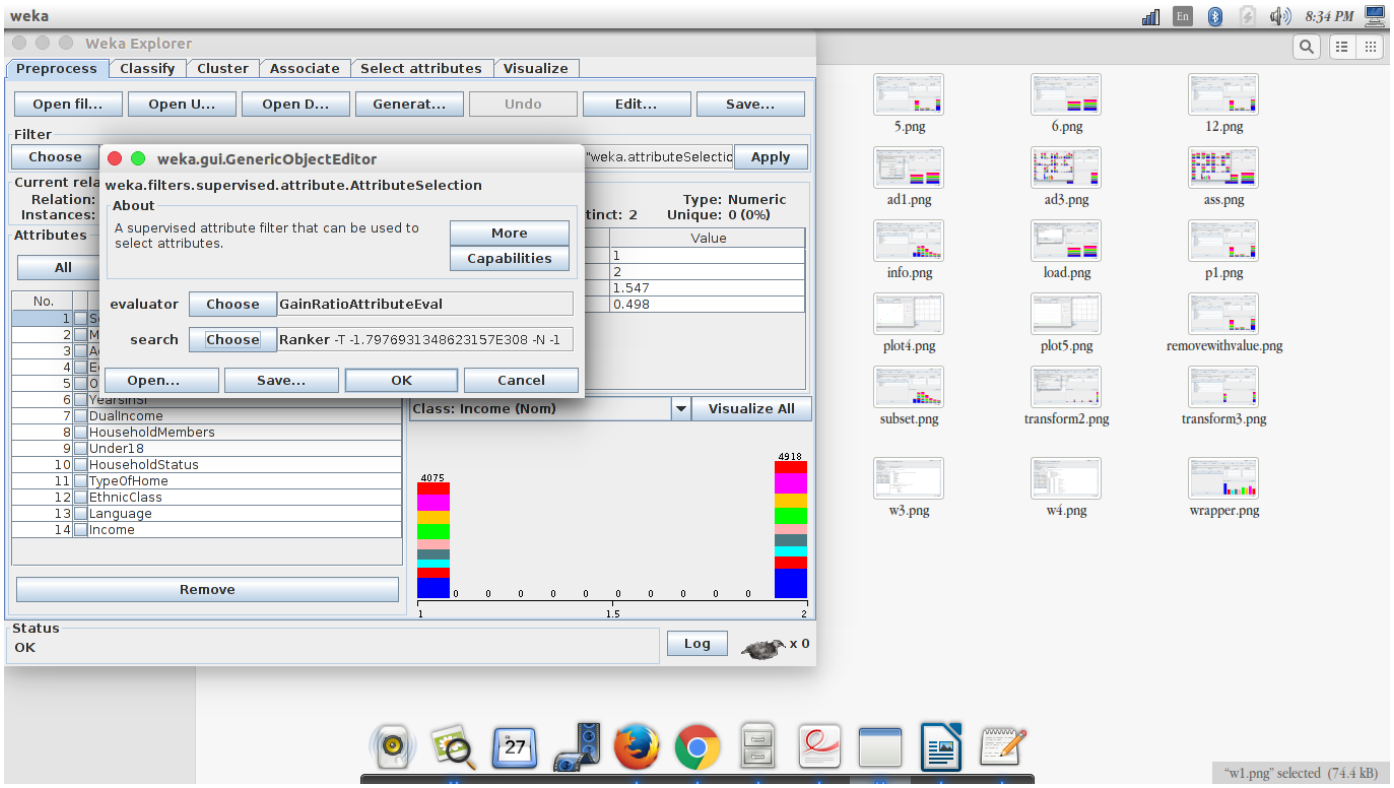

weka
8:09 PM
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
NumericTransform -R 5 -C java.lang.Math -M ceil
Apply
Current relation
Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...
Instances: 8993
Attributes: 14
Attributes
Selected attribute
Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum 1
Maximum 2
Mean 1.547
StdDev 0.498
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.NumericTransform
About
Transforms numeric attributes using a given transformation method.
More
Capabilities
attributeIndices 5
className java.lang.Math
invertSelection False
methodName ceil
Open...
Save...
OK
Cancel
Class: Income (Nom)
Visualize All
4918
4075
Remove
Status
OK
Log
x 0

Step 06.

Demonstrate attribute selection feature of WEKA for the above data.

Fast Attribute Selection Using Ranking





Weka Explorer

1:45 PM

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

Choose | AttributeSelection -E "weka.attributeSelection.GainRatioAttributeEval " -S "weka.attributeSelection.Ranker -T -1.797693134862315 7E308 -N -1" | Apply

Current relation

Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...
Instances: 8993 Attributes: 14

Attributes

All | None | Invert | Pattern

No.		Name
1	☐	HouseholdStatus
2	☐	DualIncome
3	☐	MaritalStatus
4	☐	Age
5	☐	Occupation
6	☐	Education
7	☐	TypeOfHome
8	☐	HouseholdMembers
9	☐	Under18
10	☐	Language
11	☐	EthnicClass
12	☐	YearsInSf
13	☐	Sex
14	☐	Income

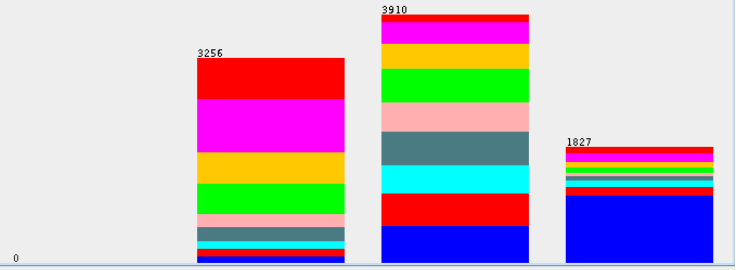
Remove

Selected attribute

Name: HouseholdStatus Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	3256
3	2	3910
4	3	1827

Class: Income (Nom) | Visualize All



Status

OK

Log | x 0