Data Preprocessing in WEKA

This exercise illustrates some of the basic data preprocessing operations that can be performed using WEKA. The sample data set used for this example is the "bank data" available in comma-separated format (bank-data.csv).

The data contains the following fields

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)

Loading the Data

In addition to the native ARFF data file format, WEKA has the capability to read in ".csv" format files. This is fortunate since many databases or spreadsheet applications can save or export data into flat files in this format. As can be seen in the sample data file, the first row contains the attribute names (separated by commas) followed by each data row with attribute values listed in the same order (also separated by commas). In fact, once loaded into WEKA, the data set can be saved into ARFF format.

In this example, we load the data set into WEKA, perform a series of operations using WEKA's preprocessing filters. While all of these operations can be performed from the command line, we use the GUI interface for WEKA Explorer.

Initially (in the Preprocess tab) click "open" and navigate to the directory containing the data file (.csv or .arff). In this case we will open the above data file. This is shown in Figure p1.
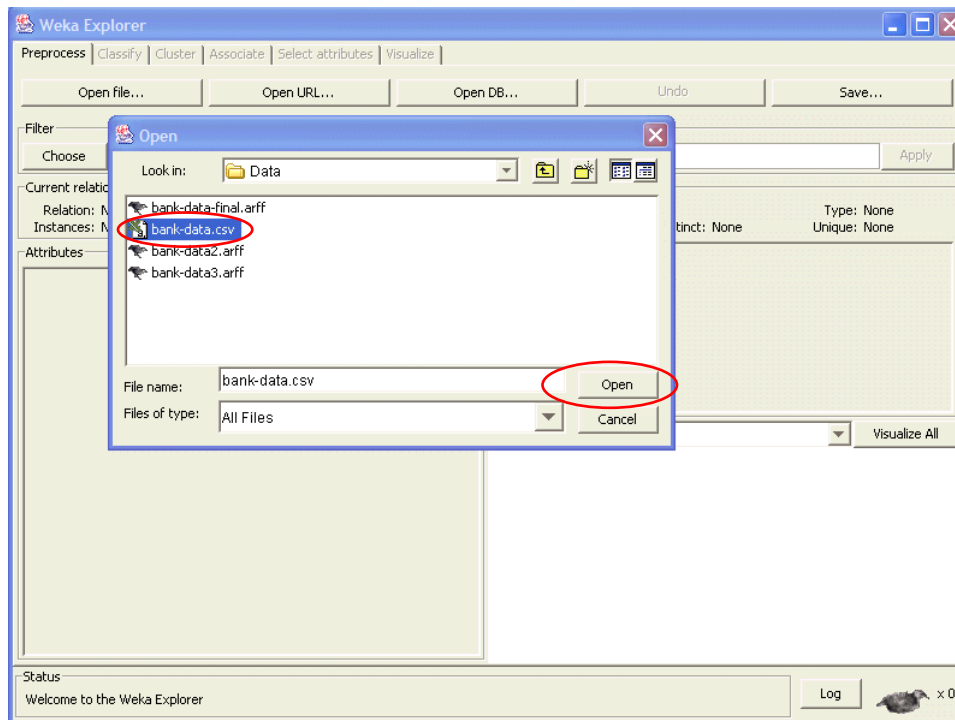
Figure p1

Once the data is loaded, WEKA will recognize the attributes and during the scan of the data will compute some basic statistics on each attribute. The left panel in Figure p2 shows the list of recognized attributes, while the top panels indicate the names of the base relation (or table) and the current working relation (which are the same initially).
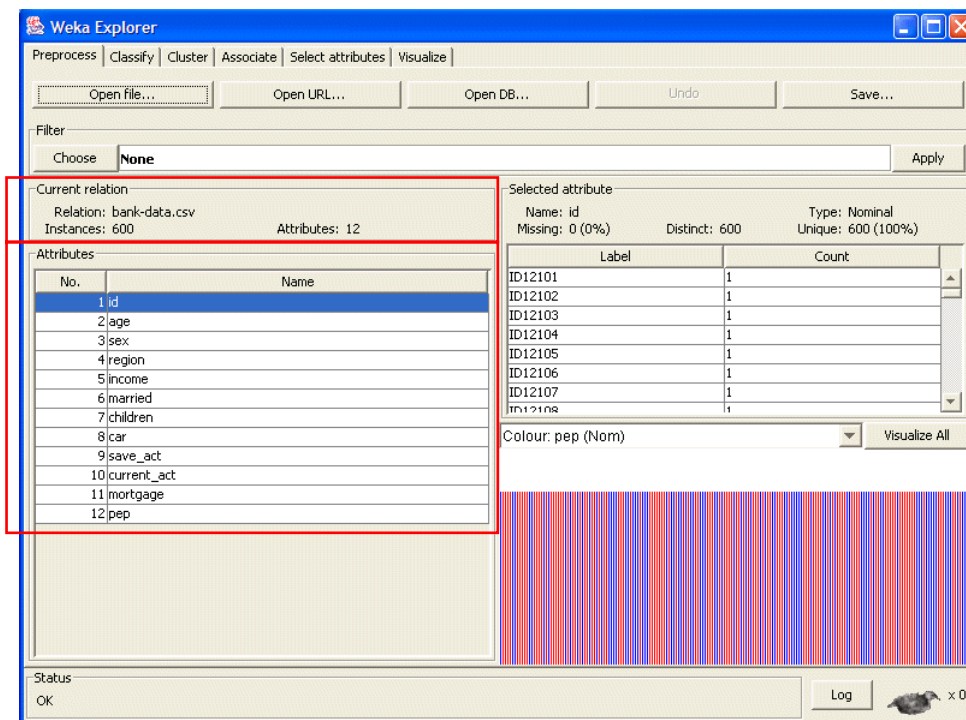

Figure p2

Clicking on any attribute in the left panel will show the basic statistics on that attribute. For categorical attributes, the frequency for each attribute value is shown, while for continuous

attributes we can obtain min, max, mean, standard deviation, etc. As an example, see Figures p3 and p4 below which show the results of selecting the "age" and "married" attributes, respectively.
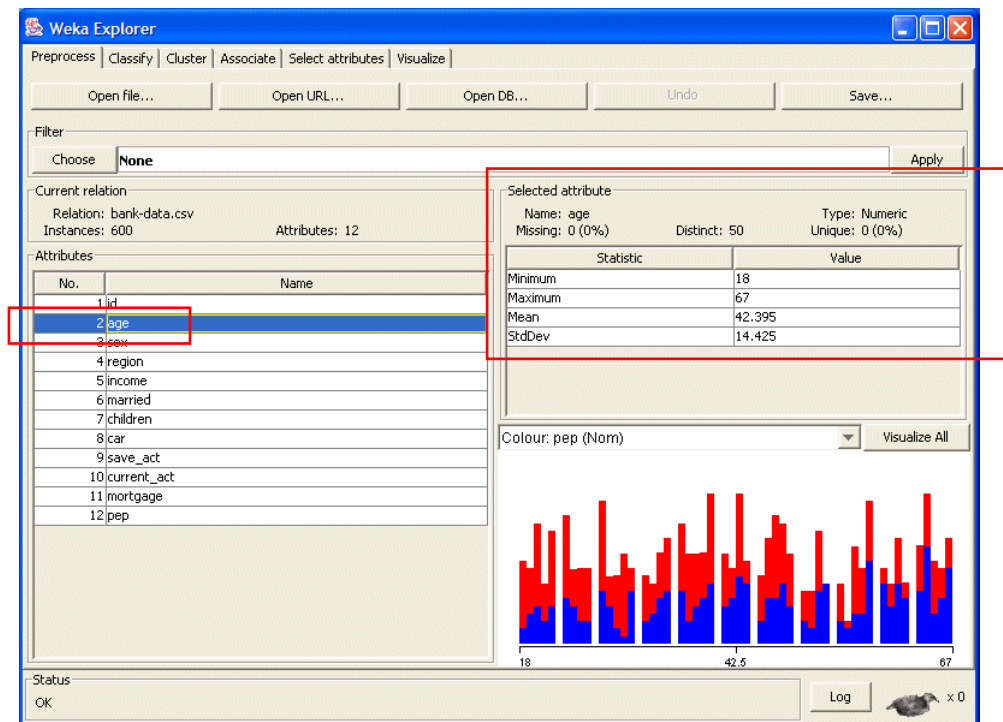
Figure p3
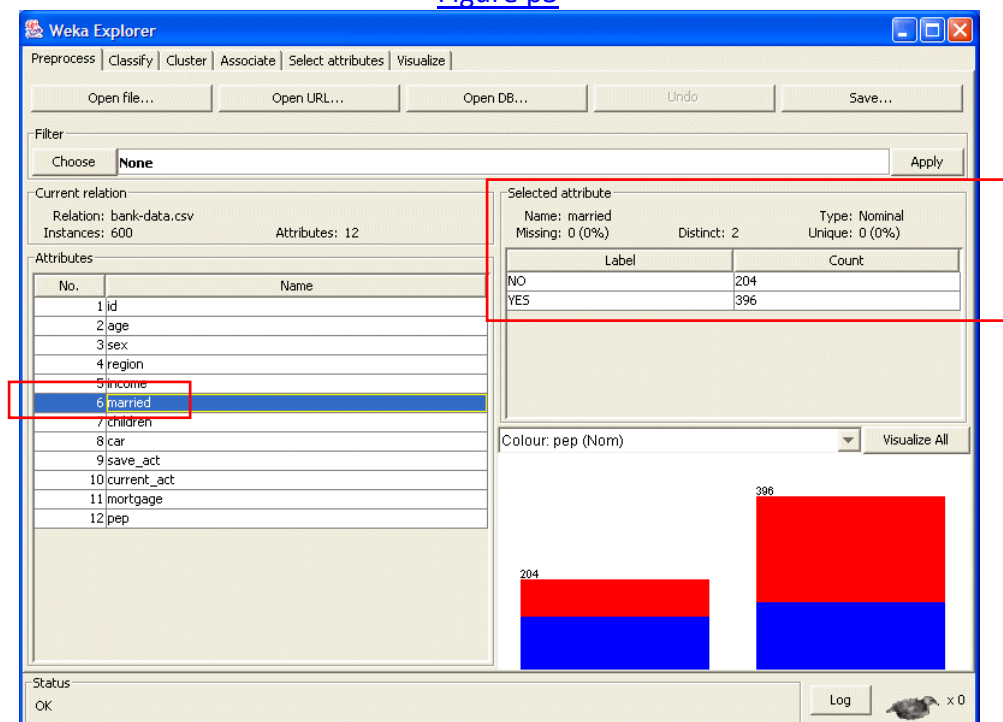
Figure p4

Note that the visualization in the right bottom panel is a form of cross-tabulation across two attributes. For example, in Figure p4 above, the default visualization panel cross-tabulates

"married" with the "pep" attribute (by default the second attribute is the last column of the data file). You can select another attribute using the drop down list.

Selecting or Filtering Attributes

In our sample data file, each record is uniquely identified by a customer id (the "id" attribute). We need to remove this attribute before the data mining step. We can do this by (1) simply select the attribute and click on "Remove button" as shown in Figure p5 (WEKA 3.6.2) or
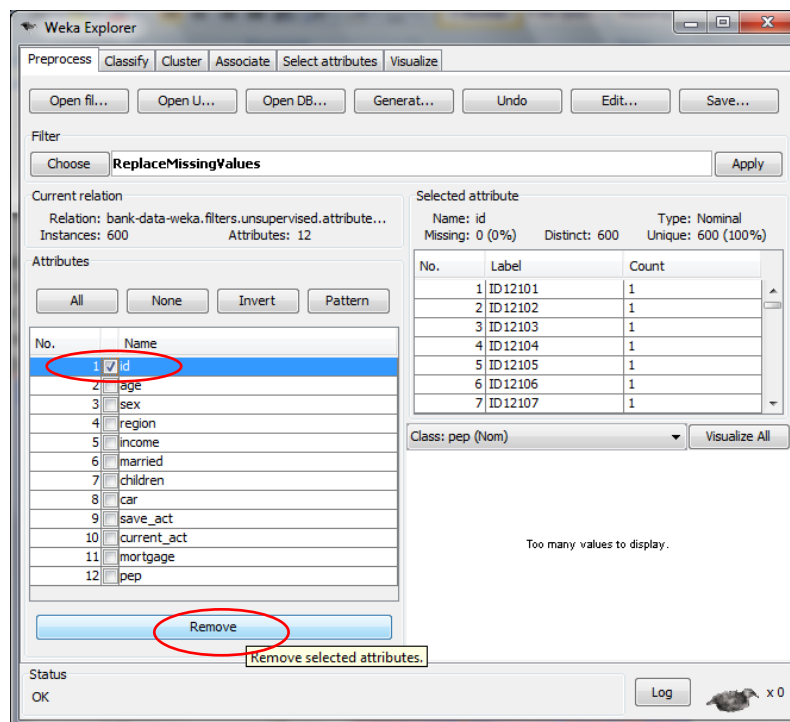
Figure p5

(2) using the Attribute filters in WEKA. In the "Filter" panel, click on the "Choose" button. This will show a popup window with a list available filters. Scroll down the list and select the "weka.filters.unsupervised.attribute.Remove" filter as shown in Figure p6.
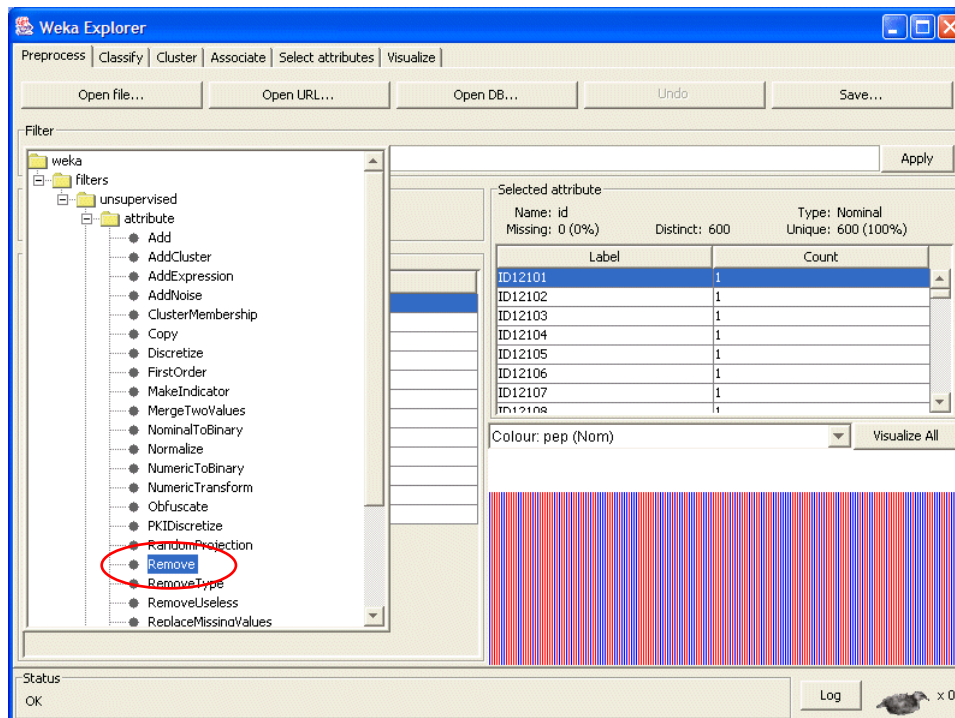
Figure p6

Next, click on text box immediately to the right of the "Choose" button. In the resulting dialog box enter the index of the attribute to be filtered out (this can be a range or a list separated by commas). In this case, we enter 1 which is the index of the "id" attribute (see the left panel). Make sure that the "invertSelection" option is set to false (otherwise everything except attribute 1 will be filtered). Then click "OK" (See Figure p7). Now, in the filter box you will see "Remove -R 1" (see Figure p8).
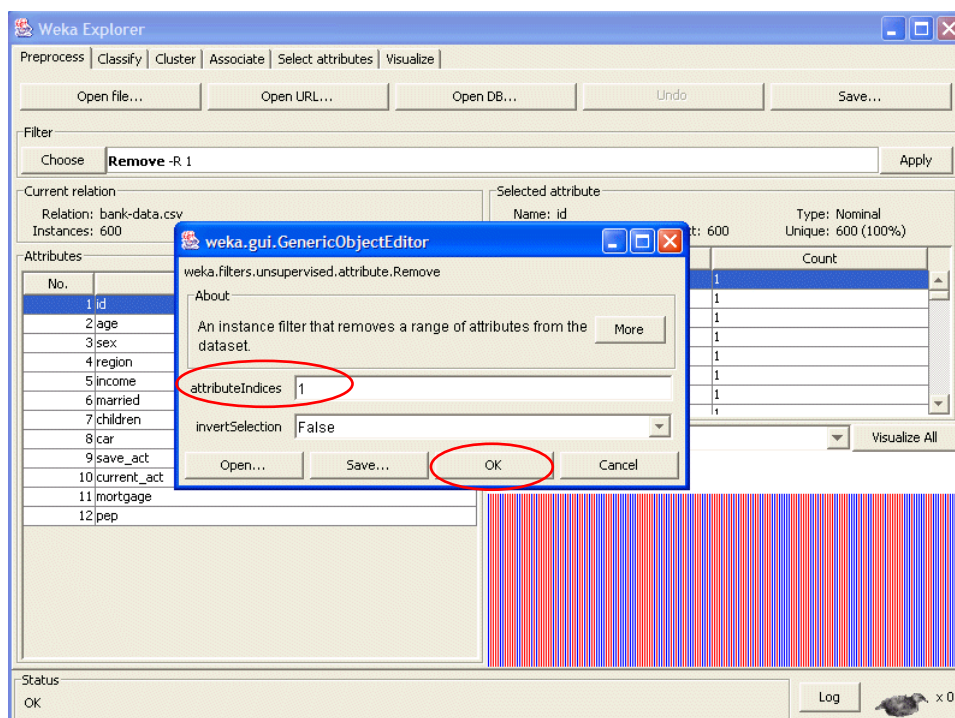
Figure p7
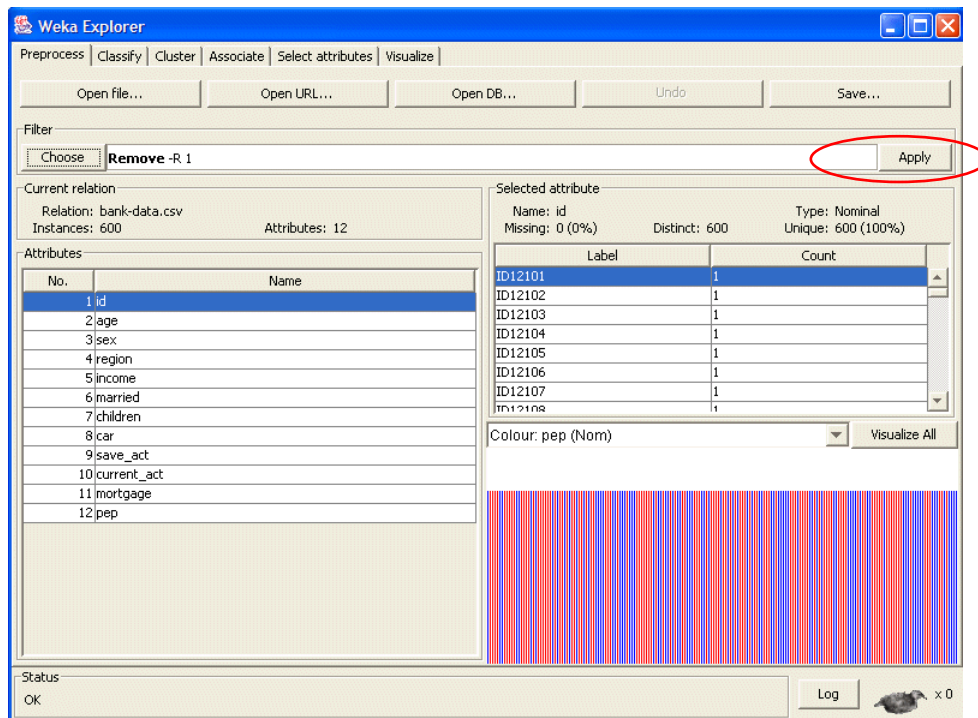
Figure p8

Click the "Apply" button to apply this filter to the data. This will remove the "id" attribute and create a new working relation (whose name now includes the details of the filter that was applied). The result is depicted in Figure p9:
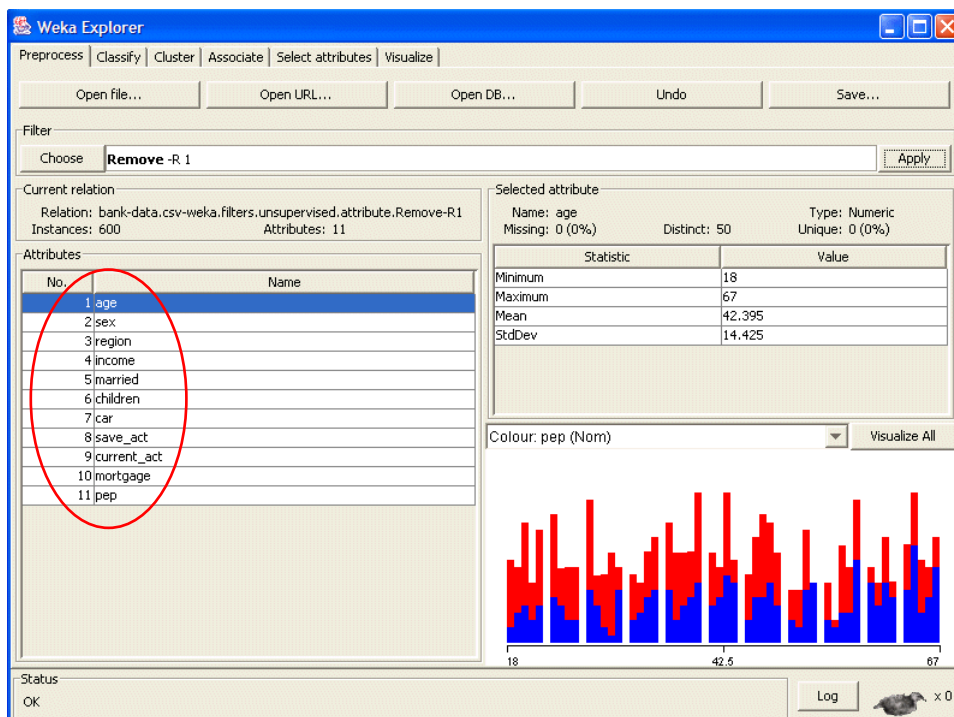
Figure p9

It is possible now to apply additional filters to the new working relation. In this example, however, we will save our intermediate results as separate data files and treat each step as a separate WEKA session. To save the new working relation as an ARFF file, click on save

button in the top panel. Here, as shown in the "save" dialog box (see Figure p10), we will save the new relation in the file "bank-data-R1.arff".
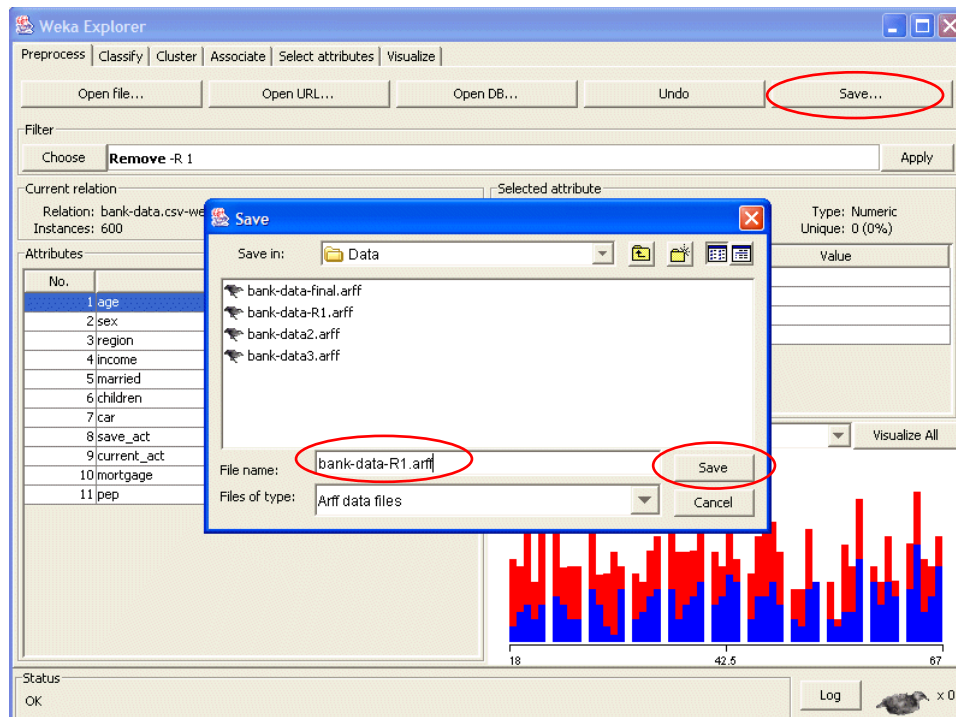
Figure p10

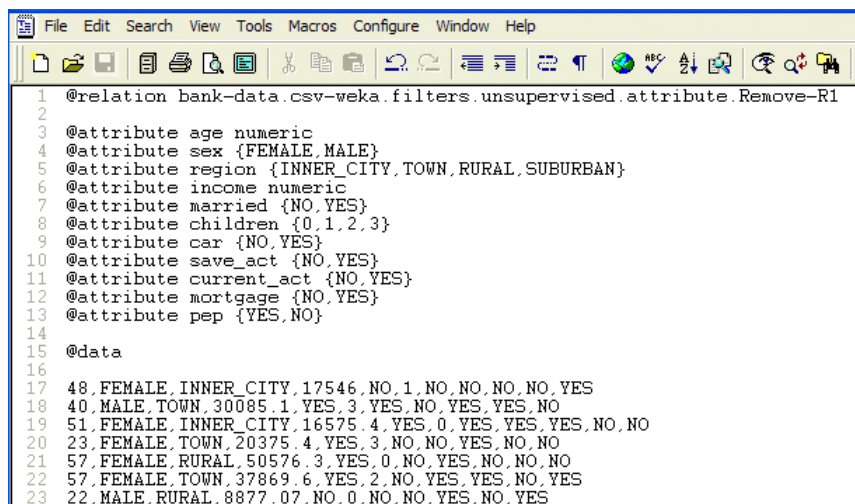Figure p11 shows the top portion of the new generated ARFF file (in text editor).

Figure p11

Note that in the new data set, the "id" attribute and all the corresponding values in the records have been removed. Also, note that WEKA has automatically determined the correct types and values associated with the attributes, as listed in the Attributes section of the ARFF file.

Discretization

Some techniques, such as association rule mining, can only be performed on categorical data. This requires performing discretization on numeric or continuous attributes. There are 3 such attributes in this data set: "age", "income", and "children". In the case of the "children" attribute the range of possible values are only 0, 1, 2, and 3. In this case, we have opted for keeping all of these values in the data. This means we can simply discretize by removing the keyword "numeric" as the type for the "children" attribute in the ARFF file, and replacing it with the set of discrete values. We do this directly in our text editor as seen in Figure p12. In this case, we have saved the resulting relation in a separate file "bank-data2.arff".


```
File  Edit  Search  View  Tools  Macros  Configure  Window  Help

 1   @relation bank-data.csv-weka.filters.unsupervised.attribute.Remove-R1
 2
 3   @attribute age numeric
 4   @attribute sex {FEMALE,MALE}
 5   @attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
 6   @attribute income numeric
 7   @attribute married {NO,YES}
 8   @attribute children {0,1,2,3}
 9   @attribute car {NO,YES}
10   @attribute save_act {NO,YES}
11   @attribute current_act {NO,YES}
12   @attribute mortgage {NO,YES}
13   @attribute pep {YES,NO}
14
15   @data
16
17   48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES
18   40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO
19   51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO
20   23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO
21   57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO
22   57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES
23   22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES
```
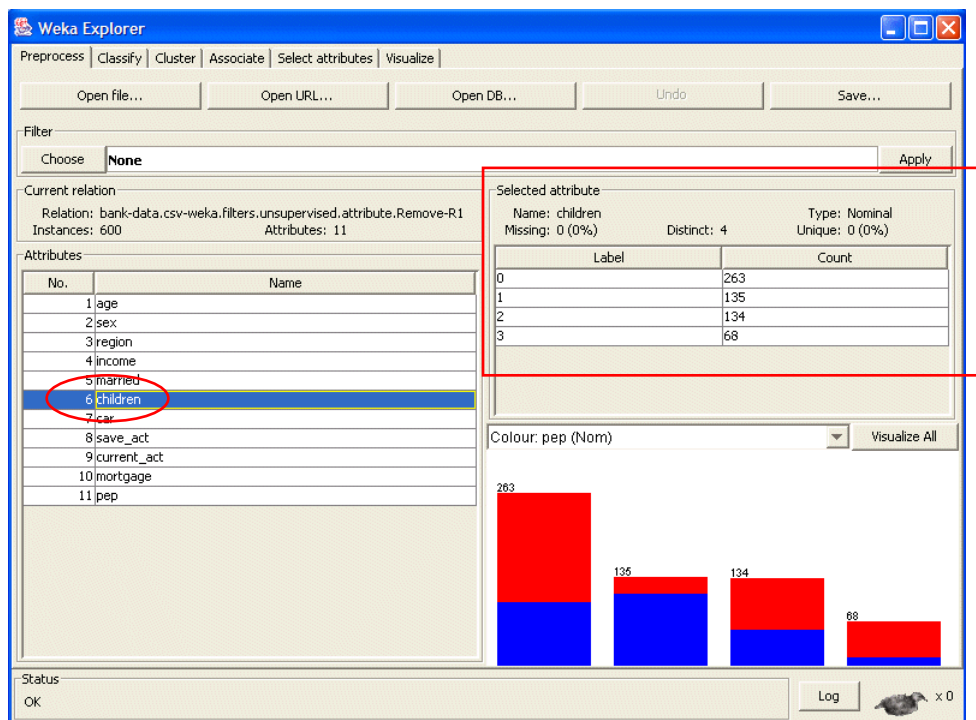Figure p12

We will rely on WEKA to perform discretization on the "age" and "income" attributes. In this example, we divide each of these into 3 bins (intervals). The WEKA discretization filter, can divide the ranges blindly, or used various statistical techniques to automatically determine the best way of partitioning the data. In this case, we will perform simple binning.

First we will load our filtered data set into WEKA by opening the file "bank-data2.arff". The "open" dialog box in depicted in Figure p13.
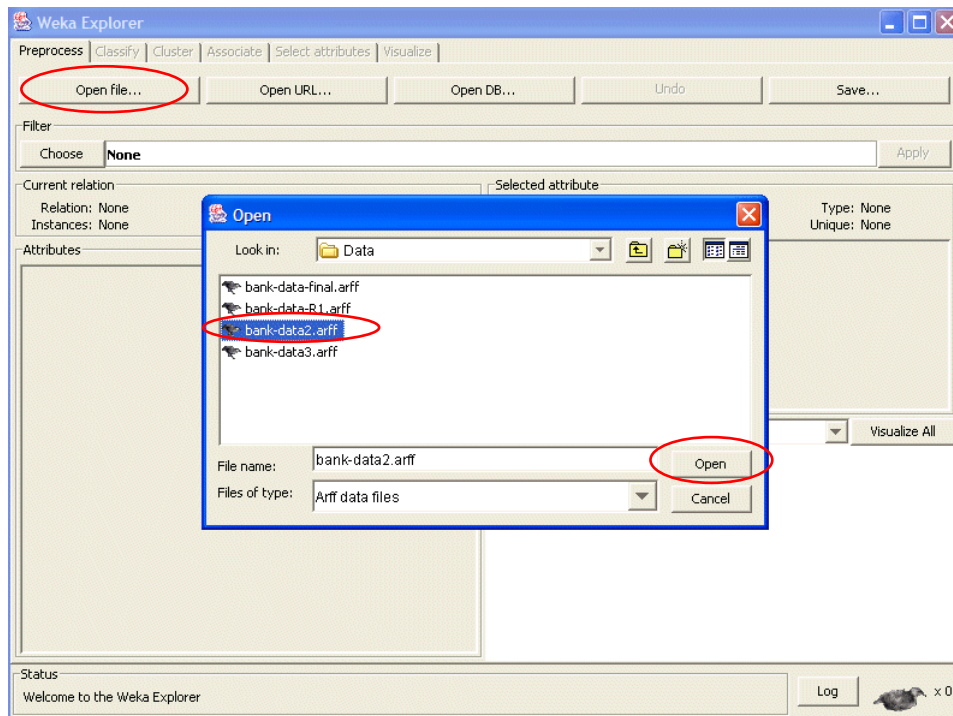
Figure p13

If we select the "children" attribute in this new data set, we see that it is now a categorical attribute with four possible discrete values. This is depicted in Figure p14.

Figure p14

Now, once again we activate the Filter dialog box, but this time, we will select "weka.filters.unsupervised.attribute.Discretize" from the list (see Figure p15).
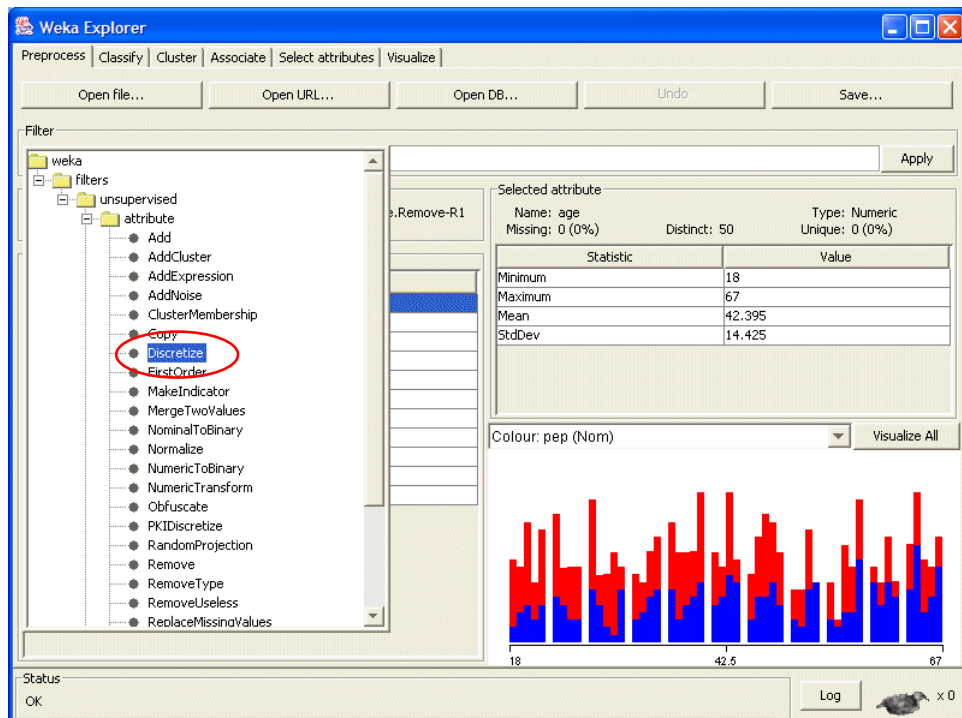
Figure p15

Next, to change the defaults for this filter, click on the box immediately to the right of the "Choose" button. This will open the Discretize Filter dialog box. We enter the index for the attributes to be discretized. In this case we enter 1 corresponding to attribute "age". We also enter 3 as the number of bins (note that it is possible to discretize more than one attribute at the same time (by using a list of attribute indices). Since we are doing simple binning, all of the other available options are set to "false". The dialog box is depicted in Figure p16. Clicking on "More" will give you detail of each parameter.
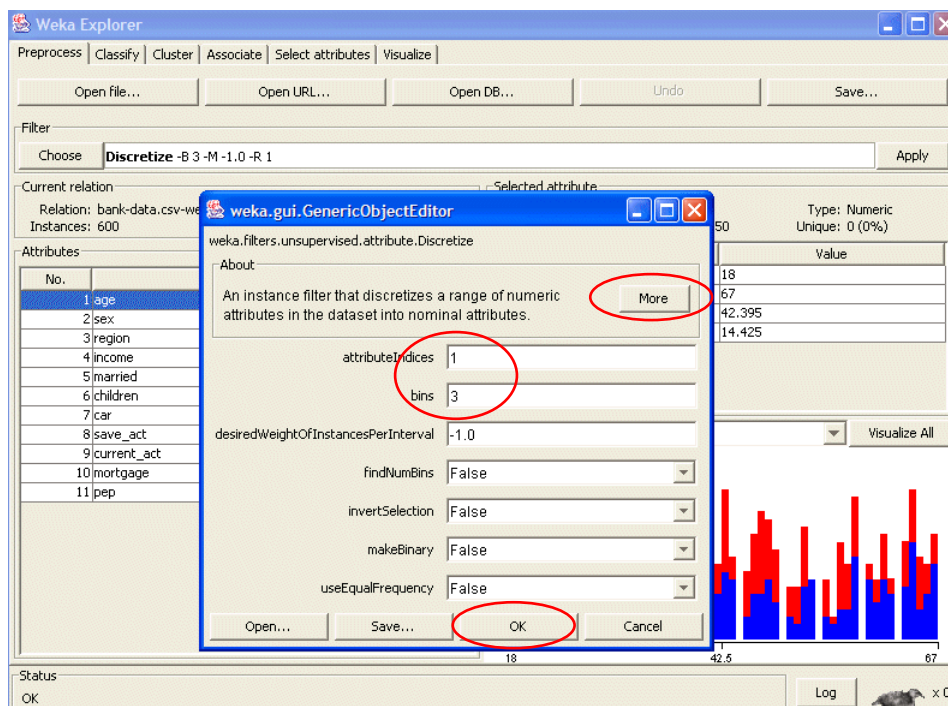
Figure p16

Click "Apply" in the Filter panel. This will result in a new working relation with the selected attribute partitioned into 3 bins (see Figure p17). To examine the results, we save the new working relation in the file "bank-data3.arff" as depicted in Figure p18.

Figure p17

Figure p18

Let us now examine the new data set using our text editor. The top portion of the data is shown in Figure p18. You can observe that WEKA has assigned its own labels to each of the value ranges for the discretized attribute. For example, the lower range in the "age" attribute is labeled "(-inf-34.333333]" (enclosed in single quotes and escape characters), while the middle range is labeled "(34.333333-50.666667]", and so on. These labels now also appear in the data records where the original age value was in the corresponding range.

Next, we apply the same process to discretize the "income" attribute into 3 bins. Again, Weka automatically performs the binning and replaces the values in the "income" column with the appropriate automatically generated labels. We save the new file into "bank-data3.arff", replacing the older version.

Clearly, the WEKA labels, while readable, leave much to be desired as far as naming conventions go. We will thus use the global search/replace functions in text editor to replace these labels with more succinct and readable ones.

Replace all of the WEKA-assigned labels of "age" and "income" attributes. Note that the attribute section (the top part) of the arff file must be adjusted accordingly.

Figure p19 shows the final result of the transformation and the newly assigned labels for these attribute values.
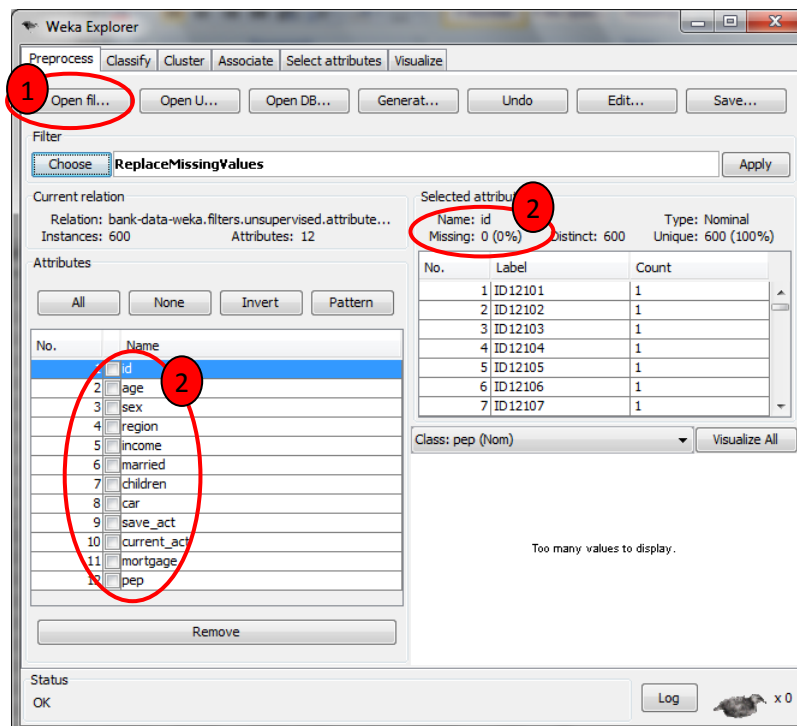

Figure p19

We now also change the relation name in the ARFF file to "bank-data-final" and save the file as "bank-data-final.arff".
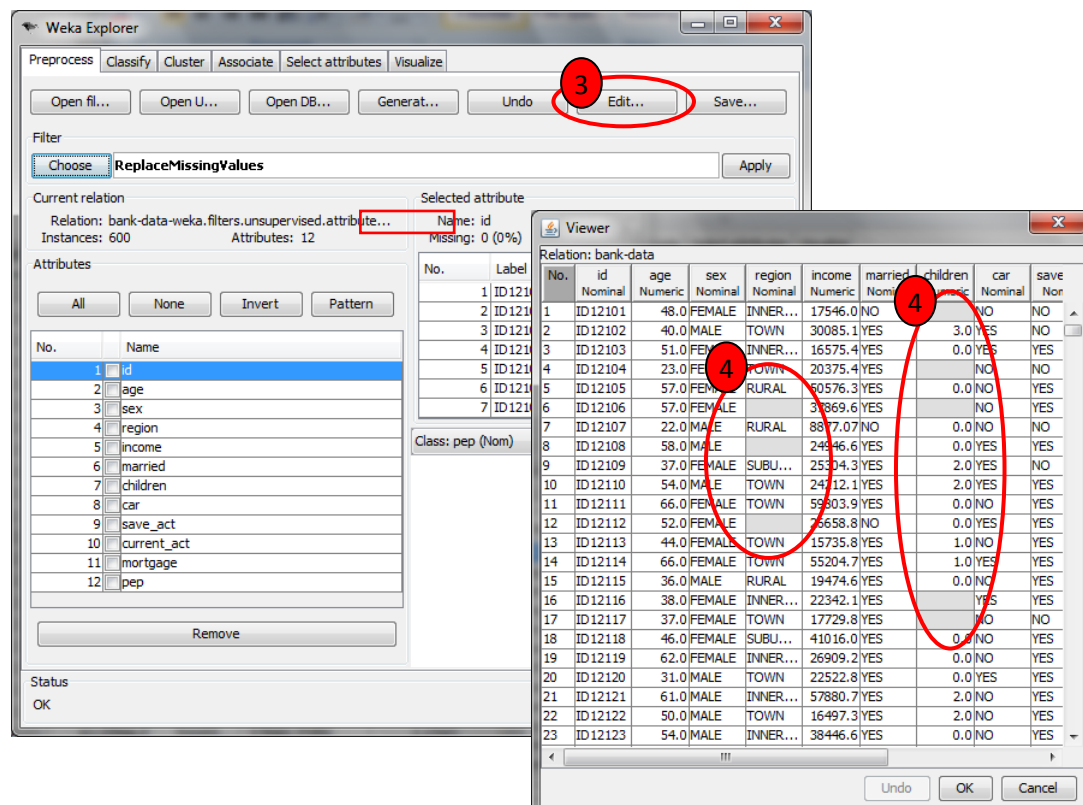
You may try with different number of bins. There is also a parameter for equal-frequency binning. Check it out.
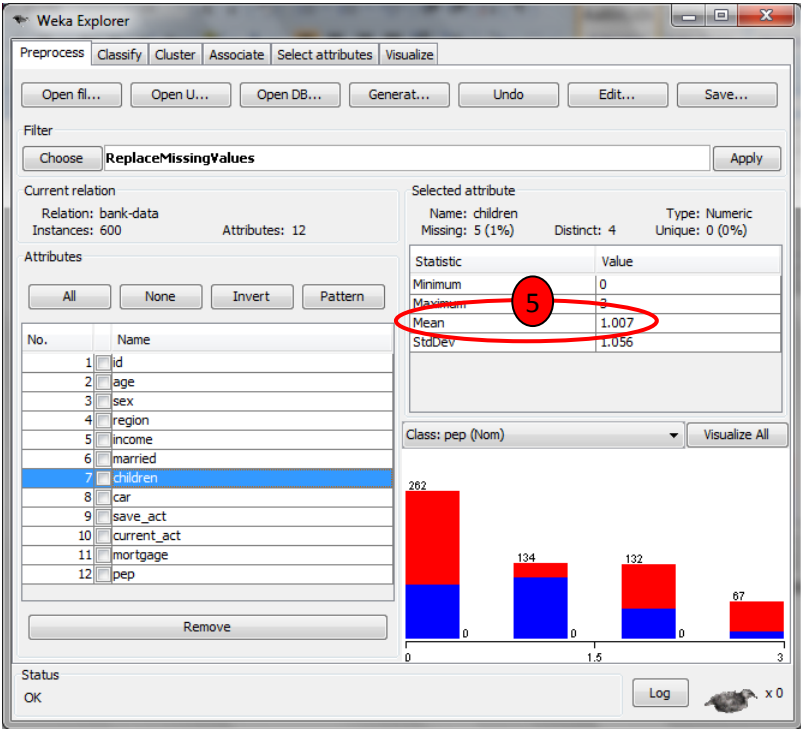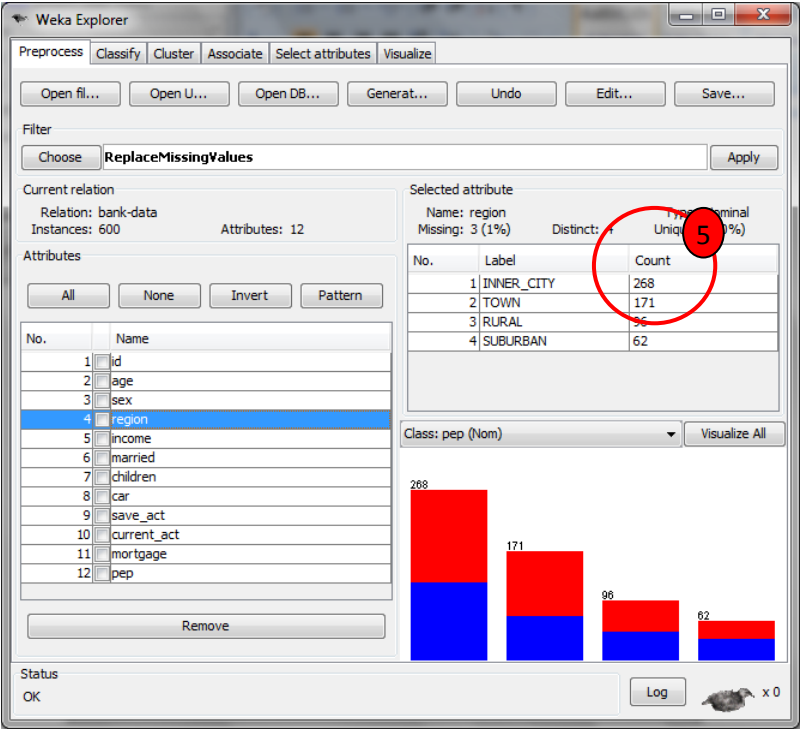
Missing Values

1. Open file "bank-data.arff"
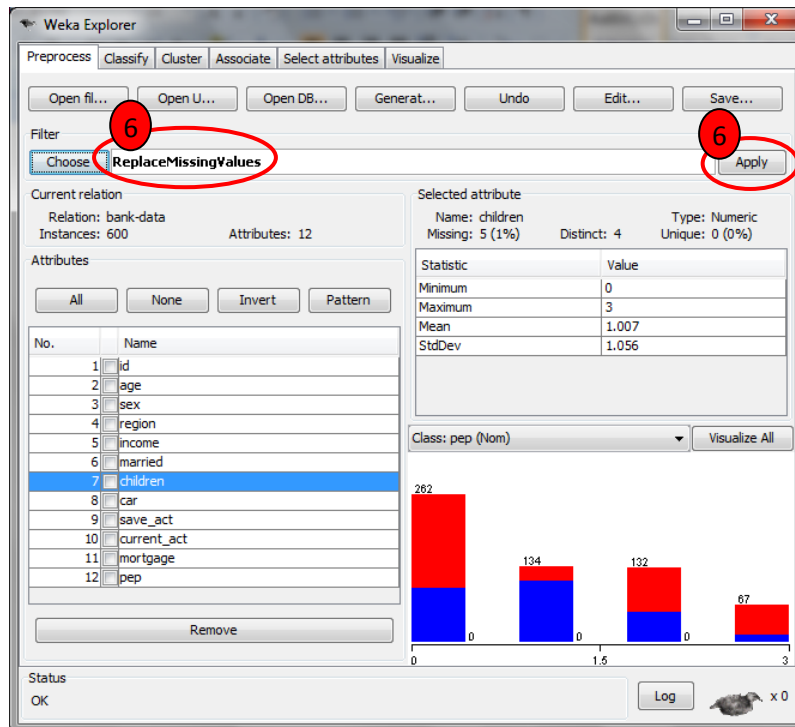2. Check if there is any missing values in any attribute.

3. Edit data to make some missing values.
4. Delete some data in "region"(Nominal) and "children"(Numeric) attributes. Click on "OK" button when finish.

5. Make note of Label that has Max Count in "region" and Mean of "children" attributes.

6. Choose "ReplaceMissingValues" filter
(weka.filters.unsupervised.attribute.ReplaceMissingValues). Then, click on Apply button.

7. Look into the data. How did those missing values get replaced ?

8. Edit "bank-data.arff" with text editor. Make some data missing by replacing them with '?'.
(Try with nominal and numeric attributes). Save to "bank-data-missing.arff".

9. Load "bank-data-missing.arff" into WEKA, observe the data and attribute information.

10. Replace missing values by the same procedure you had done before.