ECLT5810/SEEM5750

Lab Class 1: Weka - installation, feature engineering, decision trees, explain Assignment 1

Shuaiyi Li
sli@se.cuhk.edu.hk

Arrangement

- Weka Demo (25 min)
- Practice & QA & Break (20 min)
- Weka Demo & Assignment 1 (25 min)
- Practice & QA (20 min)

What is Weka?

Weka is an open source machine learning software that can be accessed through a graphical user interface, standard terminal applications, or a Java API. It is widely used for teaching, research, and industrial applications, contains a lot of built-in tools for standard machine learning tasks.

Here is the official website: https://www.cs.waikato.ac.nz/ml/weka/

Weka installation

Please follow the instruction here to install the stable version (3.8) of Weka

https://waikato.github.io/weka-wiki/downloading_weka/

It provides different versions to suit different OS. Please select the one you are using.

Dataset

We will use the Bank Marketing Data Set.

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

It contains 4521 samples and 16 input variables. The target **y** is the client subscribed a term deposit or not. In machine learning terminology, it is a binary classification problem.

Dataset

Here is the information of the 16 input variables:

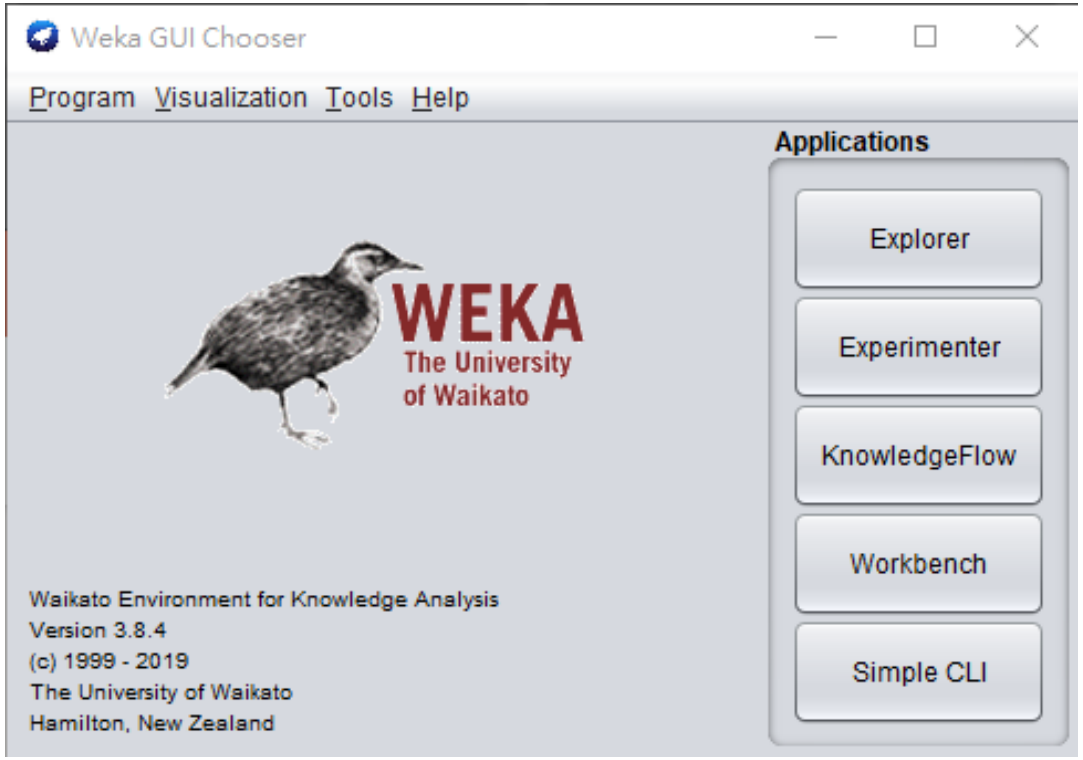
- 1 - age (numeric)
- 2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student",
- "blue-collar","self-employed","retired","technician","services")
- 3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
- 4 - education (categorical: "unknown","secondary","primary","tertiary")
- 5 - default: has credit in default? (binary: "yes","no")
- 6 - balance: average yearly balance, in euros (numeric)
- 7 - housing: has housing loan? (binary: "yes","no")
- 8 - loan: has personal loan? (binary: "yes","no")

Dataset

- 9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
- 10 - day: last contact day of the month (numeric)
- 11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- 12 - duration: last contact duration, in seconds (numeric)
- # other attributes:
- 13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
- 15 - previous: number of contacts performed before this campaign and for this client (numeric)
- 16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

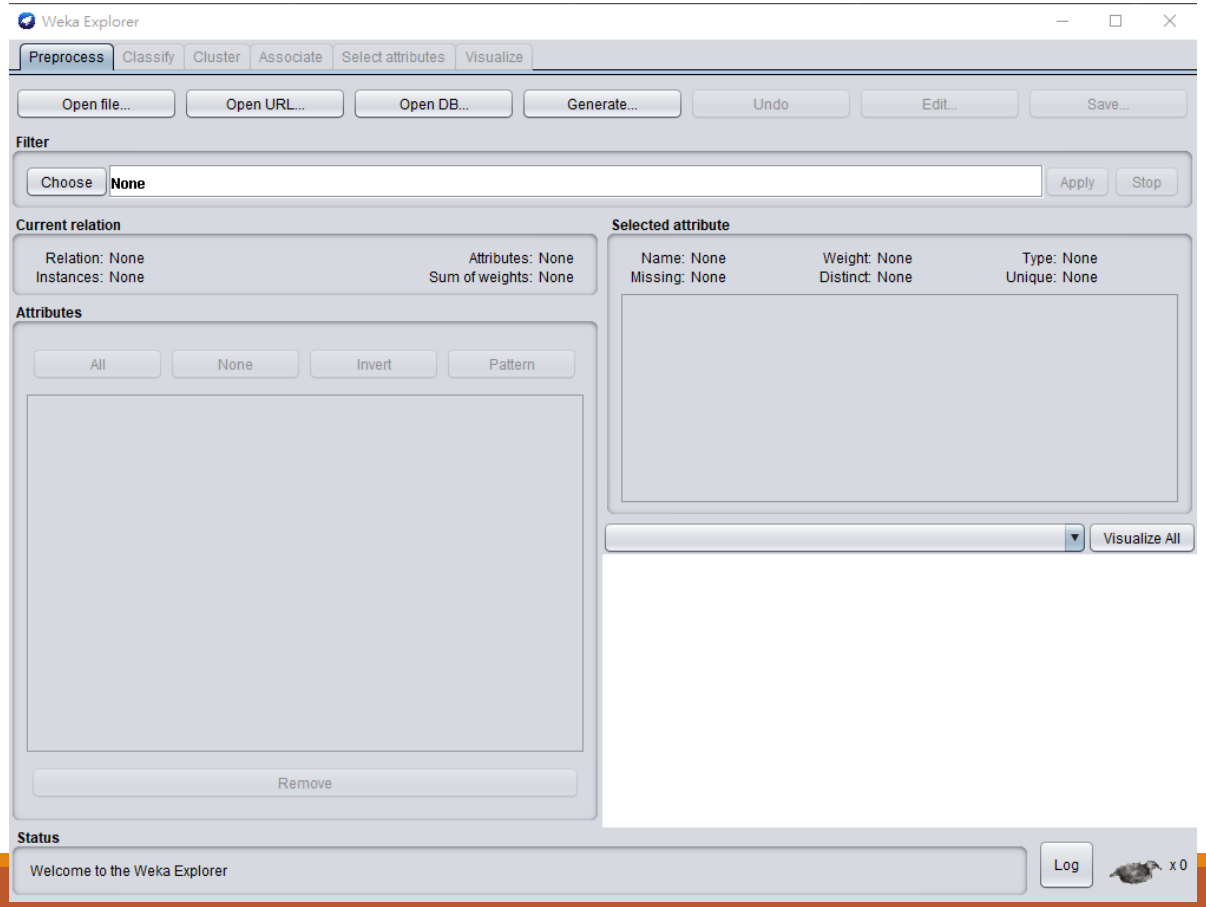
Weka GUI Chooser

If you open the Weka software, first is the Weka GUI Chooser like the following.



Explorer

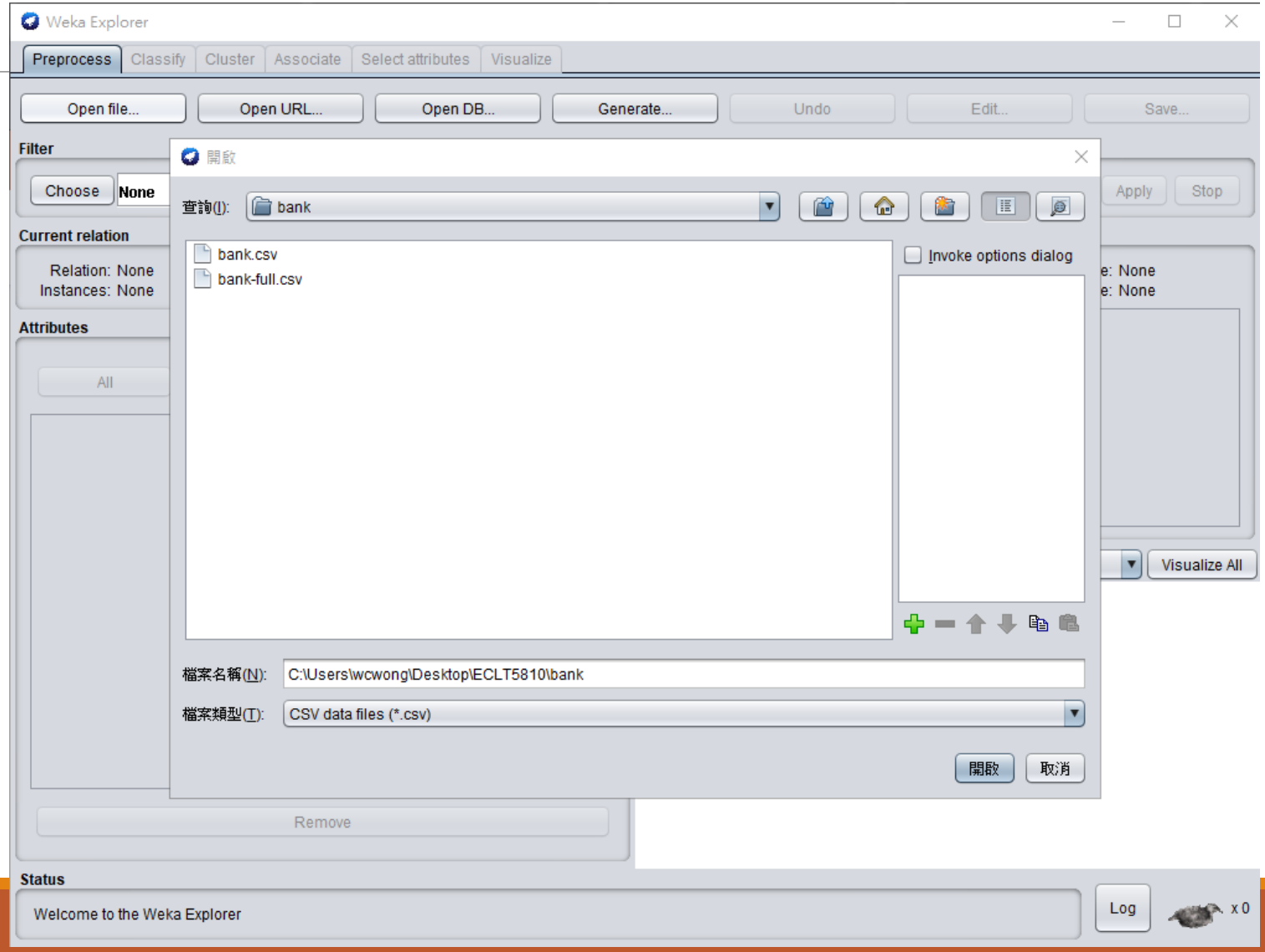
Then, select the Explorer, you will enter to this interface.



Explorer

Click **Open file**, then open the bank.csv saved in your computer.

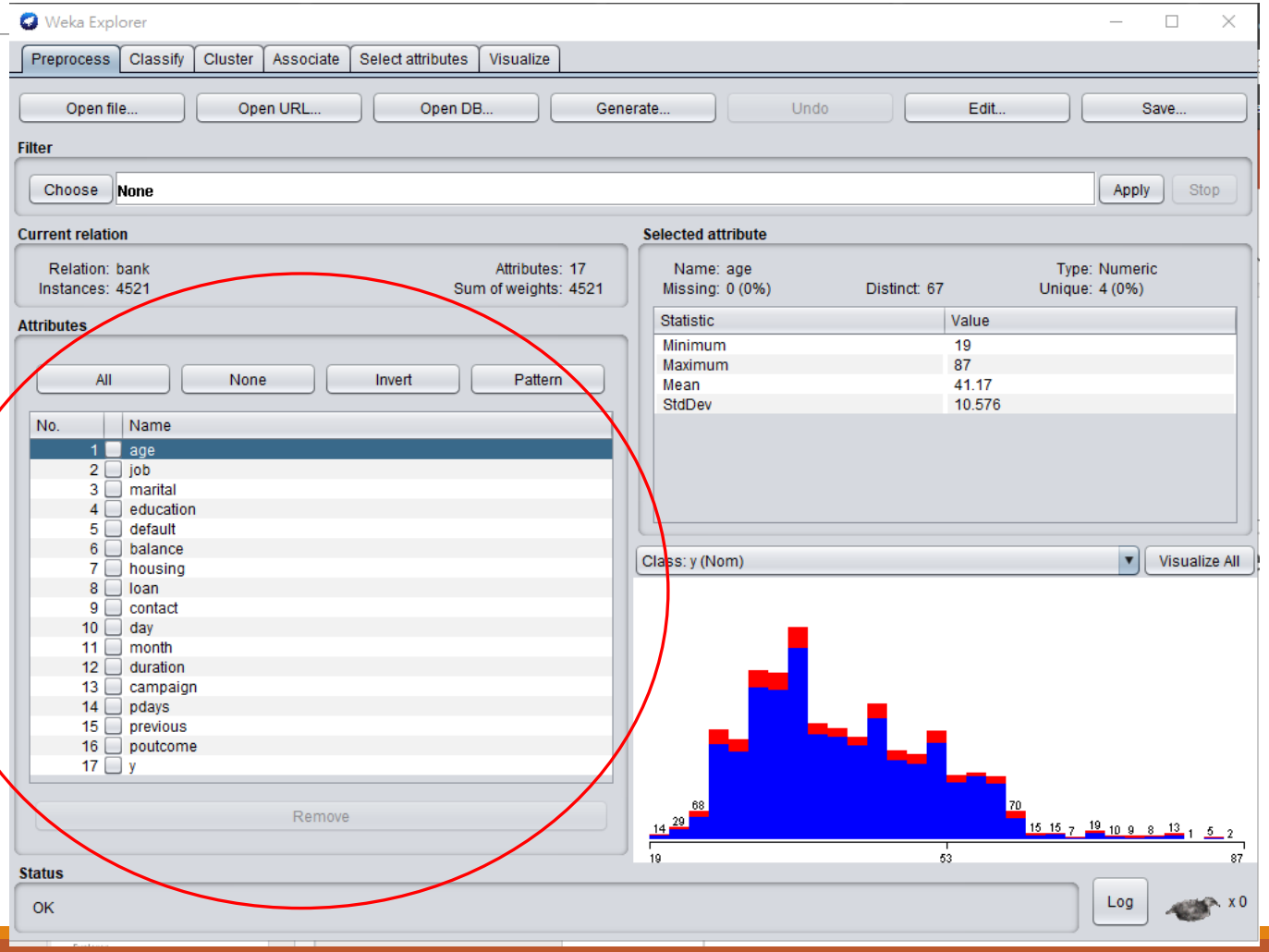
Please remember to change to **CSV data files(*.csv)** in file type.



Explorer

Now, you can see the data is loaded into Explorer.

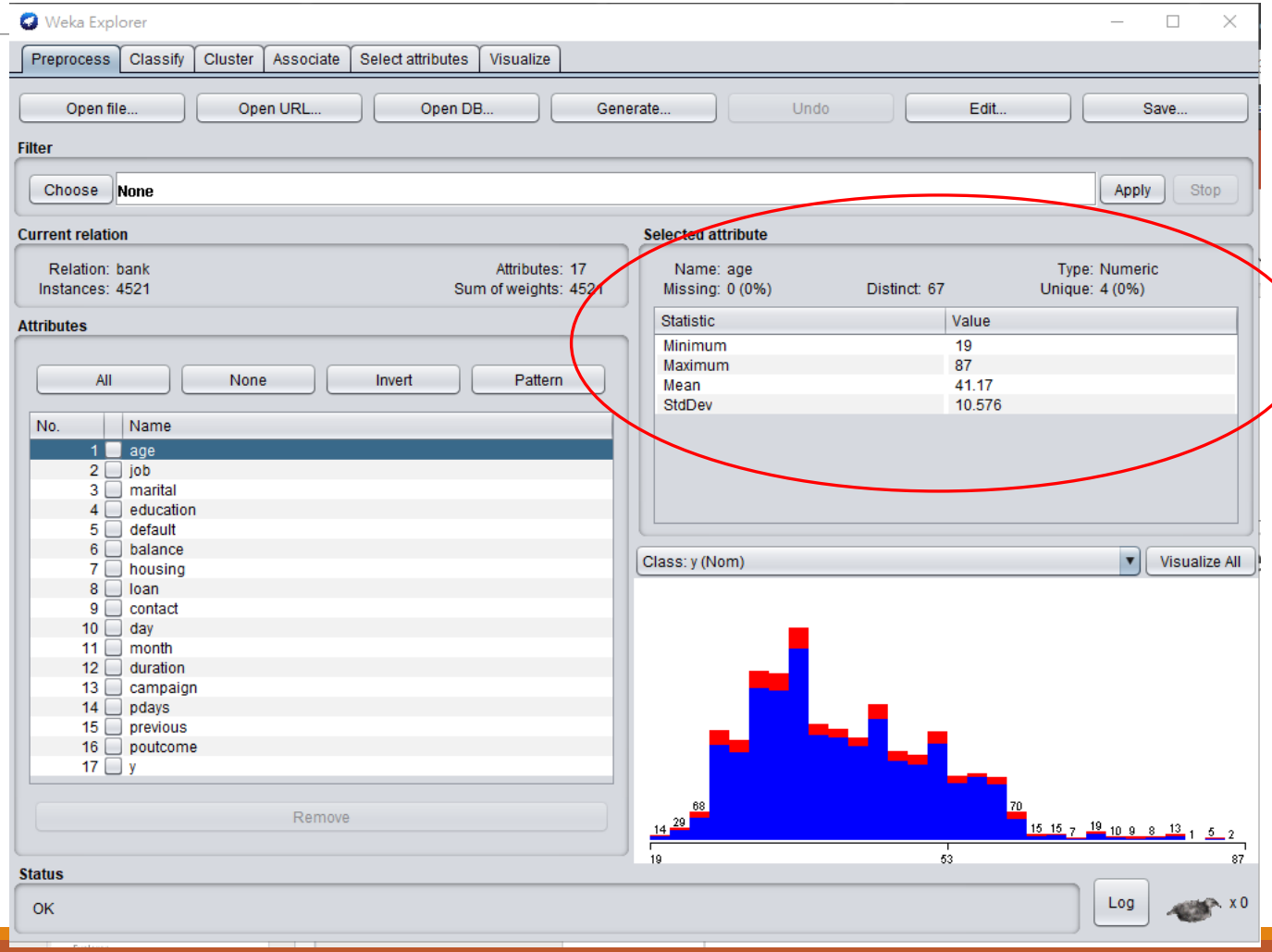
You can check out each variable by click on it in this panel.



Explorer

The statistics for each variable are also shown here.

For example, the maximum and minimum value of age is 87 and 19 respectively.



Feature Engineering

Feature engineering mostly contains two components.

- The variable transformation and,
- The variable selection

Variable transformation can be applied to the inputs for improving the precision of the predictive models.

Variable selection is useful when you want to make an initial selection of inputs or eliminate irrelevant inputs. It can also help identify non-linear relationships between the inputs and the target.

Variable Transformation

There are several variable transformation methods that can be applied to the input variables such that the precision of the predictive models can be improved.

However, we cannot know which variable transformation methods will produce the most accurate models.

Therefore, it is a good idea to try a number of different variable transformation methods techniques on the data and in turn create many different models to test it.

Variable Transformation

In Weka, it provides filters for variable transformation.

- Supervised Filters: That can be applied but require user control or make use of the class information in some way. Such as rebalancing instances for a class.
- Unsupervised Filters: That can be applied in an undirected manner. For example, discretize the numerical attributes or rescale all values in the attribution to the range 0 to 1.

In this tutorial, we will use the Unsupervised Filters.

Variable Transformation

Under these two filters, there are two groups:

- Attribute Filters: Apply an operation on attributes or one attribute at a time.
- Instance Filters: Apply an operation on instances or one instance at a time.

In this tutorial, we will mostly deal with the Attribute Filters.

Variable Transformation

Discretize Numerical Attributes

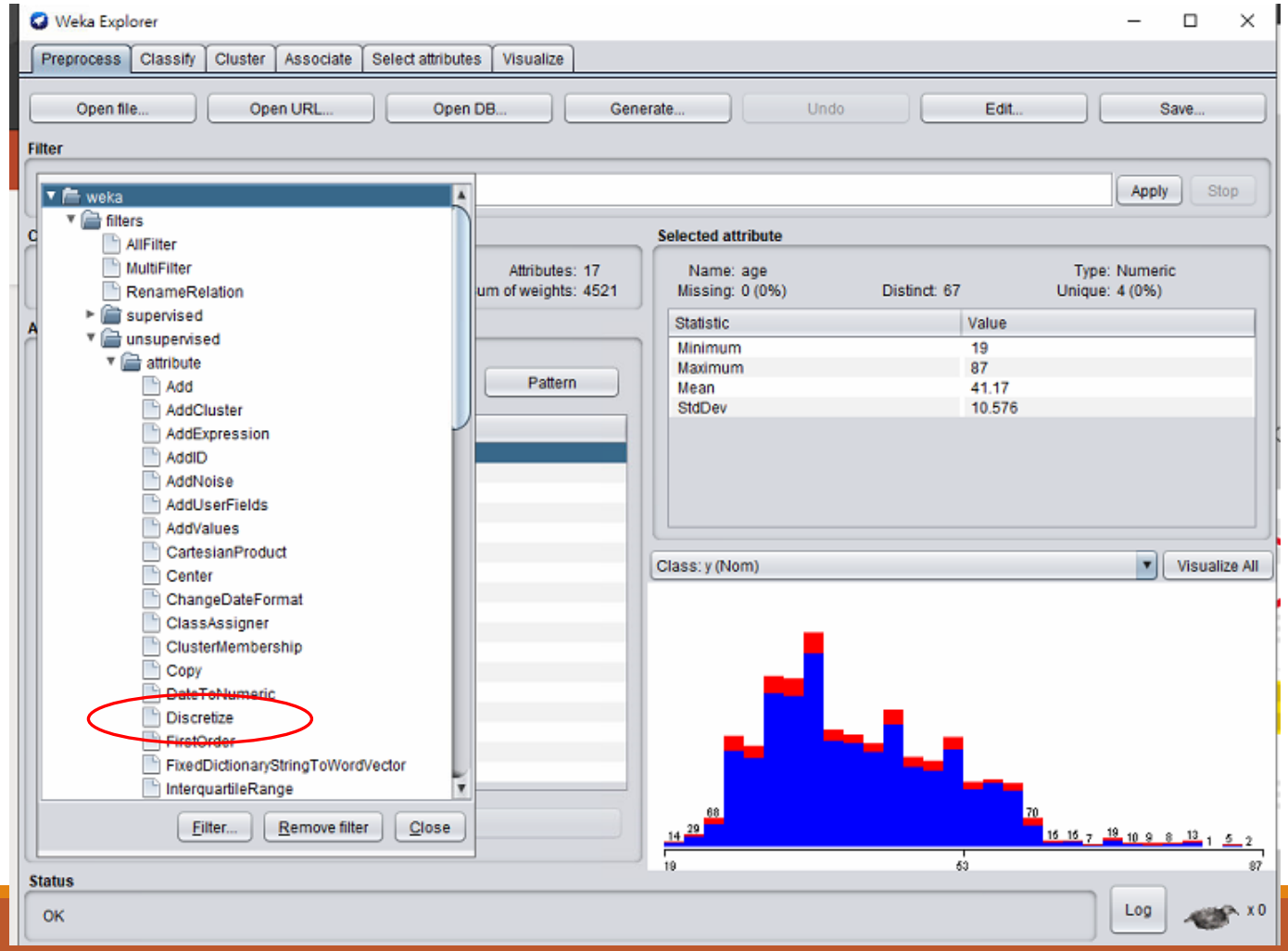
Some machine learning algorithms prefer to work with discrete attributes rather than real-valued attributes.

For example, decision tree algorithms can choose split points in real-valued attributes but are much cleaner when split points are chosen between bins or predefined groups.

Discrete attributes are those that describe a category, called nominal attributes. Those attributes that describe a category that where there is a meaning in the order for the categories are called ordinal attributes. The process of converting a real-valued attribute into an ordinal attribute or bins is called discretization.

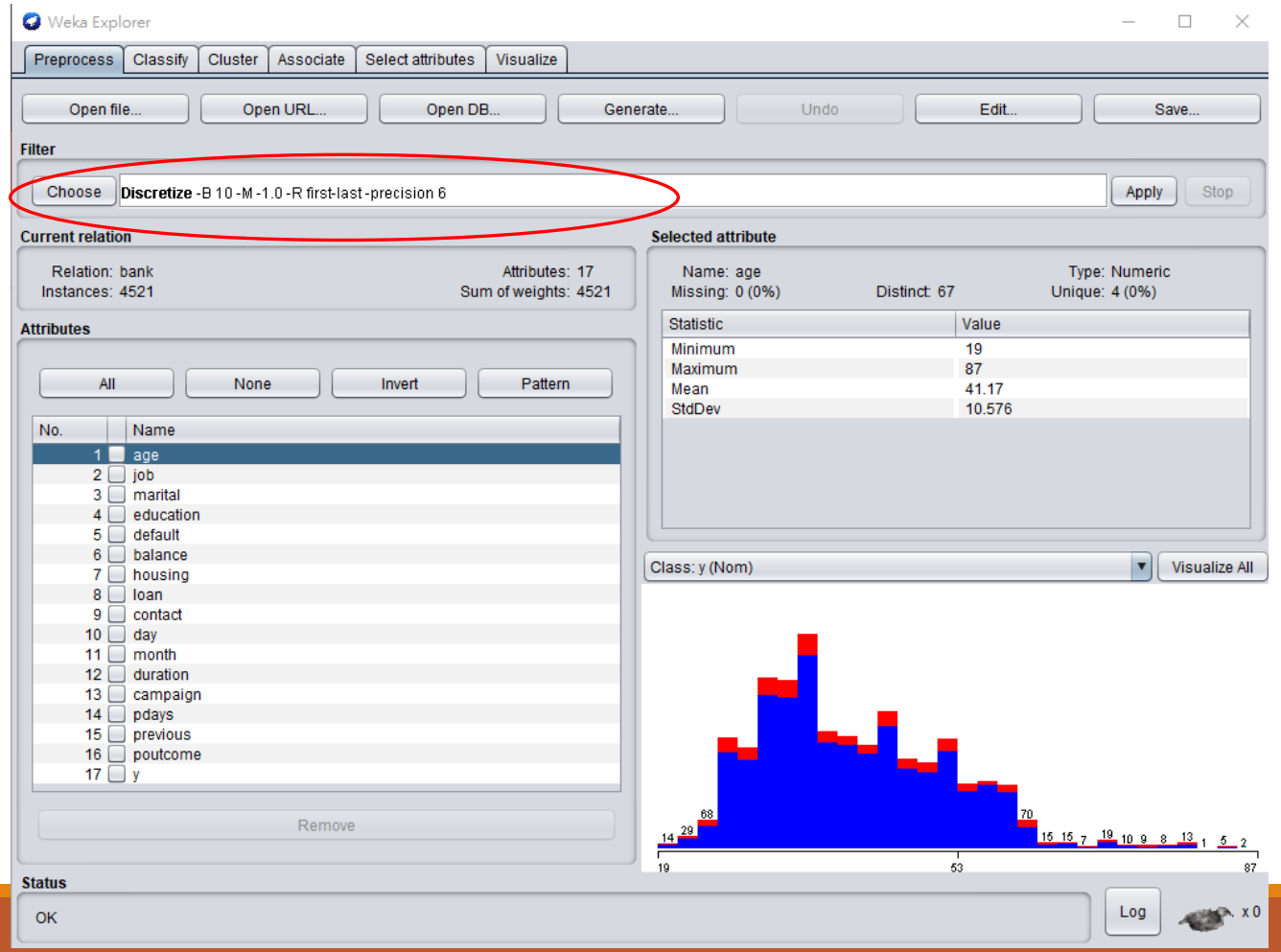
Variable Transformation

Click Choose, under filters->unsupervised->attributes, select Discretize



Variable Transformation

Click on the text near Choose, you can configure the setting of the method



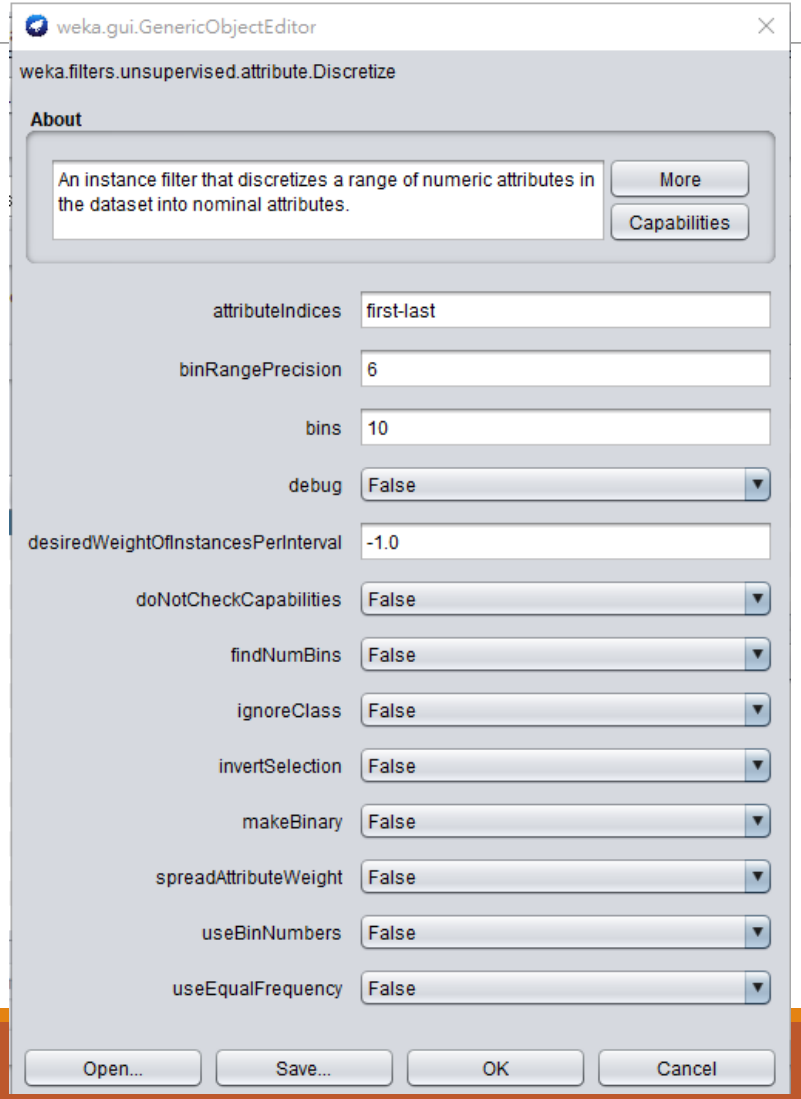
Variable Transformation

Here is the configuration of Discretize

attributesIndices means specify range of attributes to act on

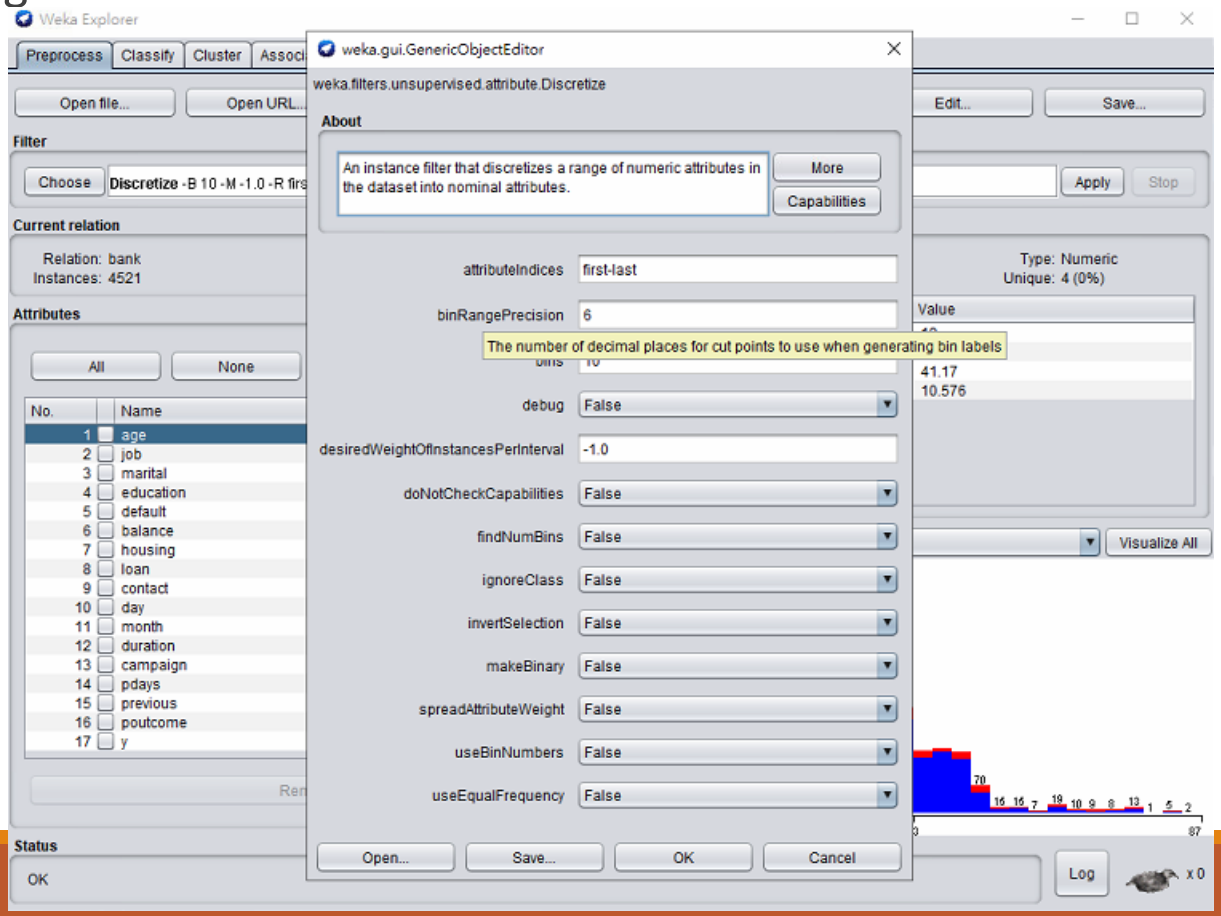
binRangePrecison means the number of decimal places for cut points to use

bins means numbers of bin



Variable Transformation

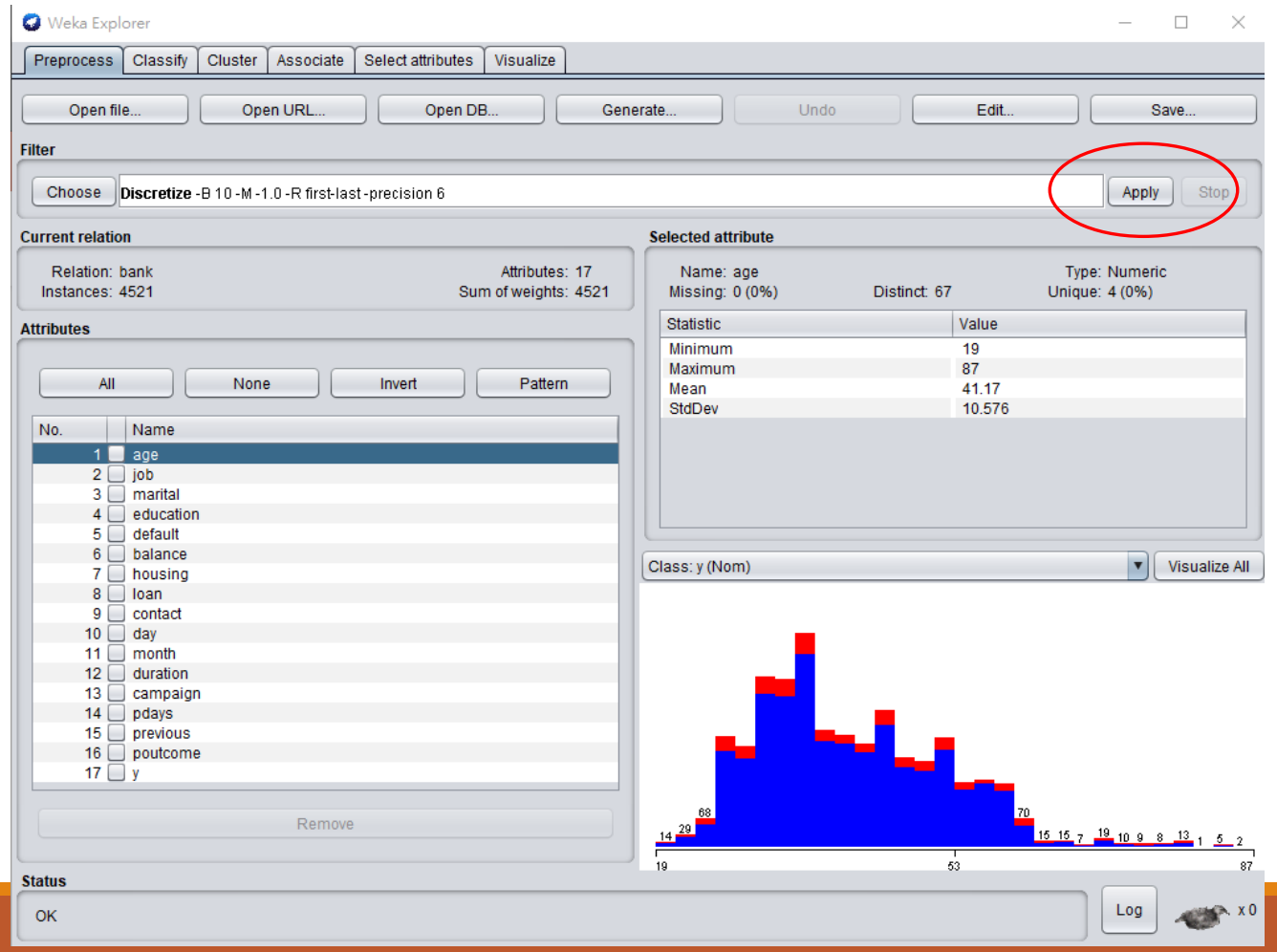
You can move the mouse cursor on the configuration' s attributes to see the meaning of it.



Variable Transformation

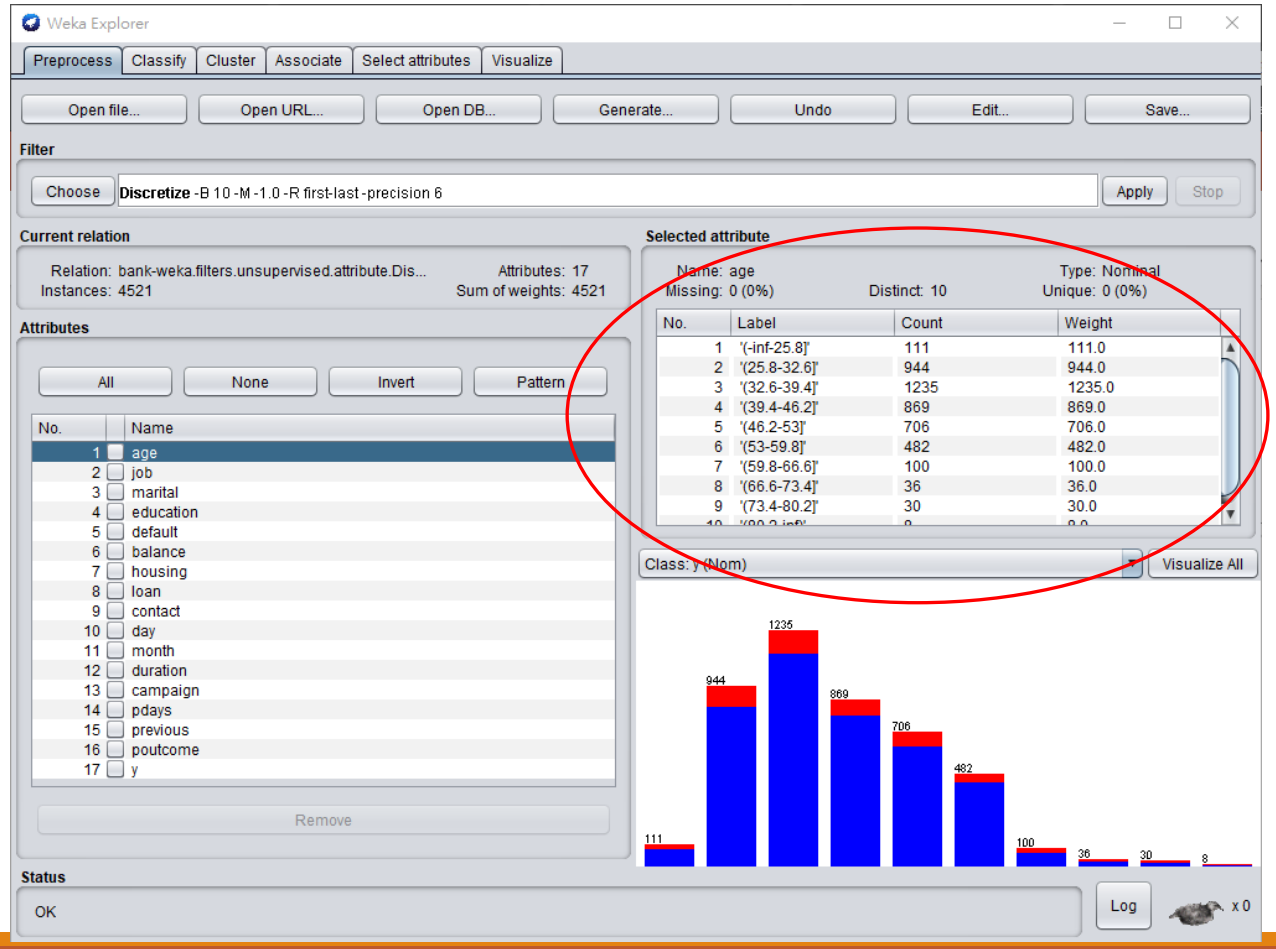
Let's leave the setting as default and click OK

Then, click apply.



Variable Transformation

Now, you can see the attribute age is discretized to 10 bins and you can see the range of each bin in this panel



Variable Transformation

Convert Nominal Attributes to Dummy Variables

Some machine learning algorithms prefer to use real valued inputs and do not support nominal or ordinal attributes.

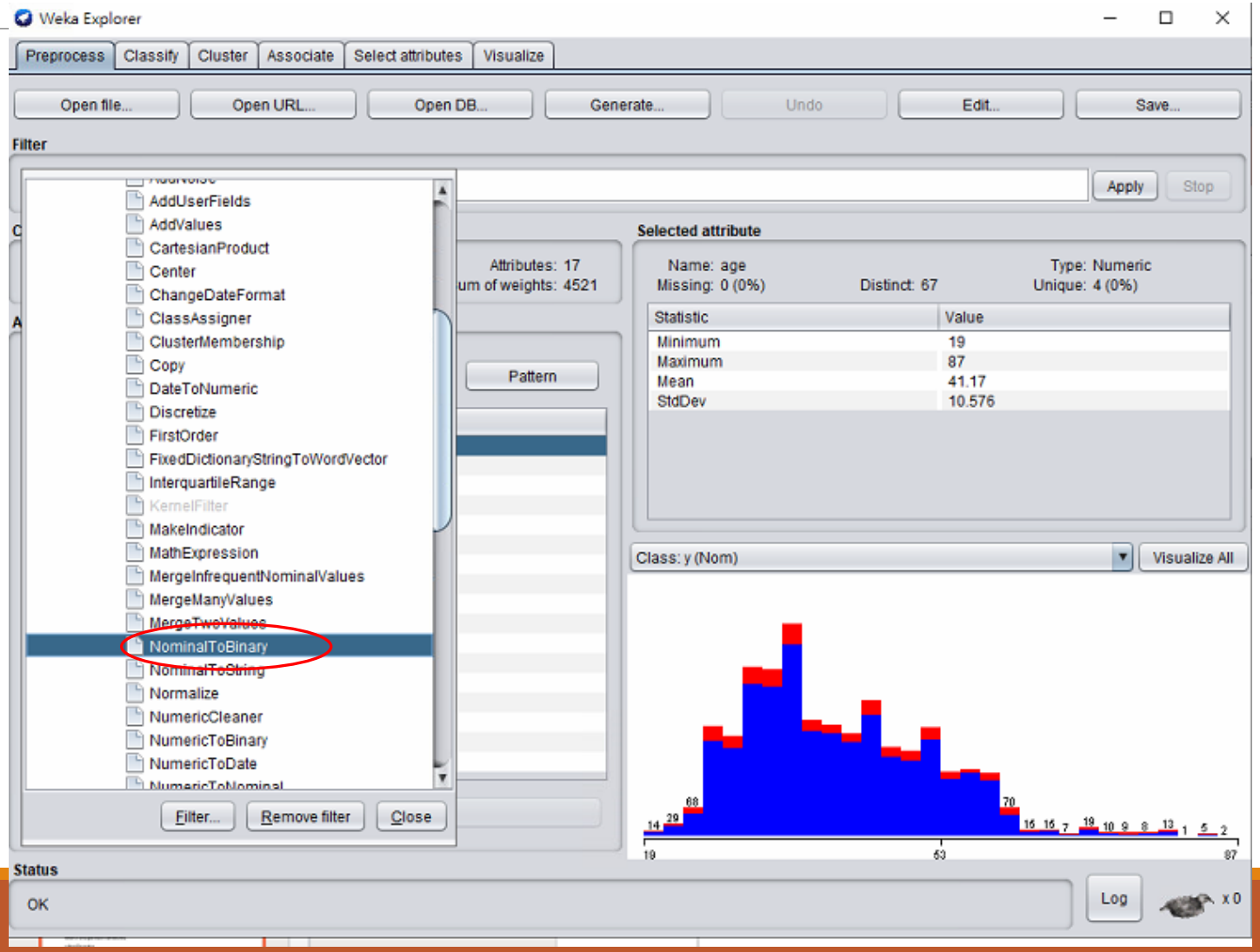
Nominal attributes can be converted to real values. This is done by creating one new binary attribute for each category. For a given instance that has a category for that value, the binary attribute is set to 1 and the binary attributes for the other categories is set to 0. This process is called creating dummy variables.

Variable Transformation

Click Choose, under

filters->unsupervised->attributes,

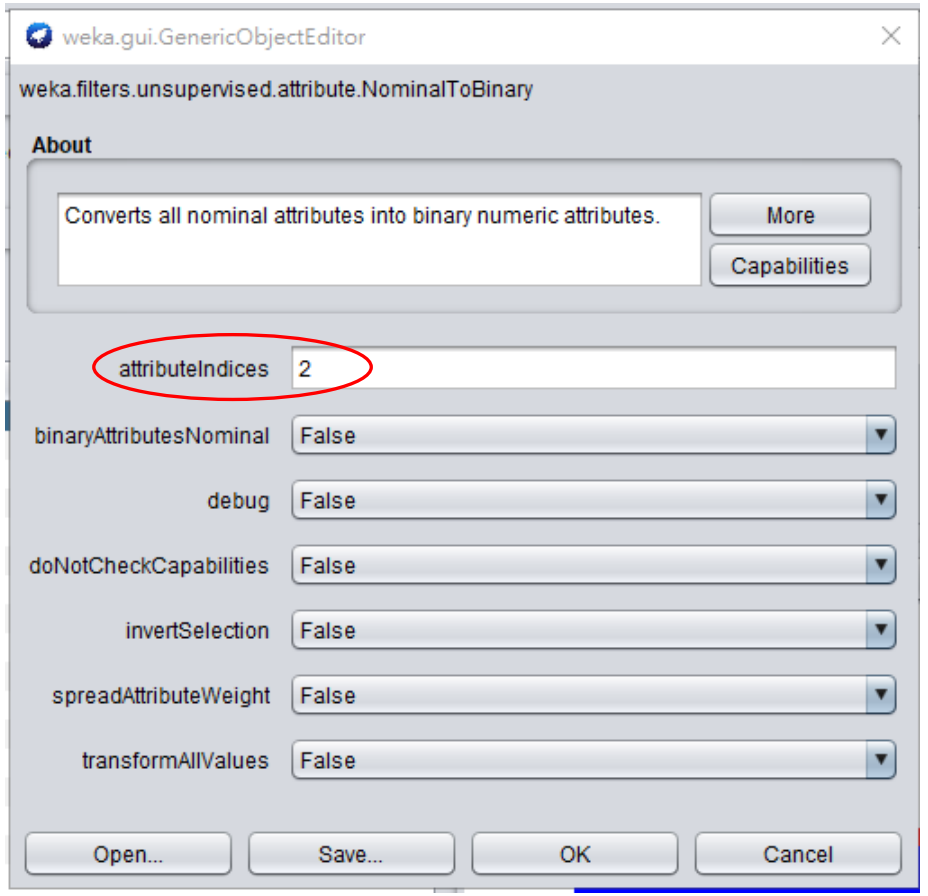
select NominalToBinary



Variable Transformation

In the configuration, set the attributeIndices to 2 as this time we only want to transform the job attirbute, then Click OK

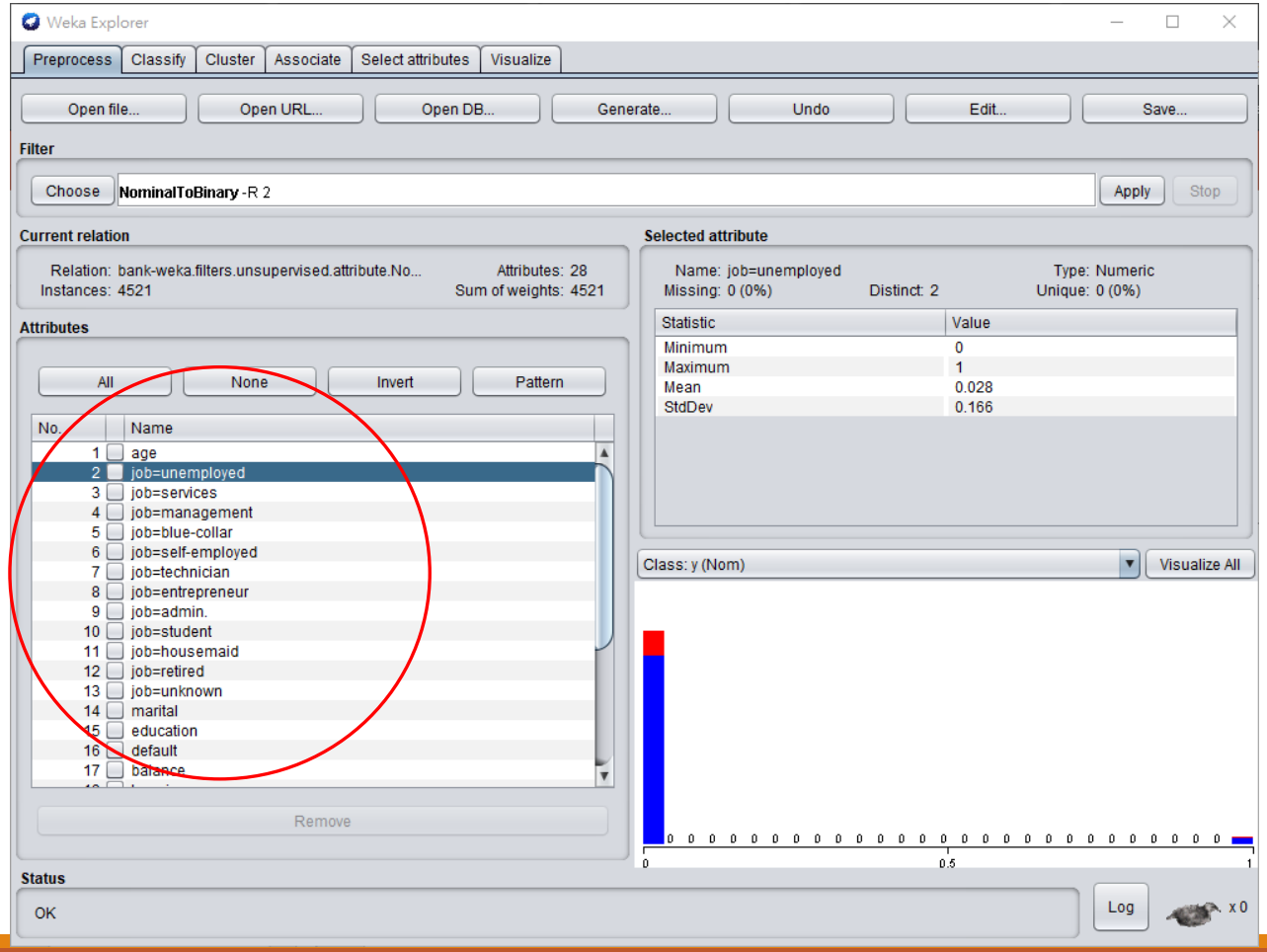
Click Apply



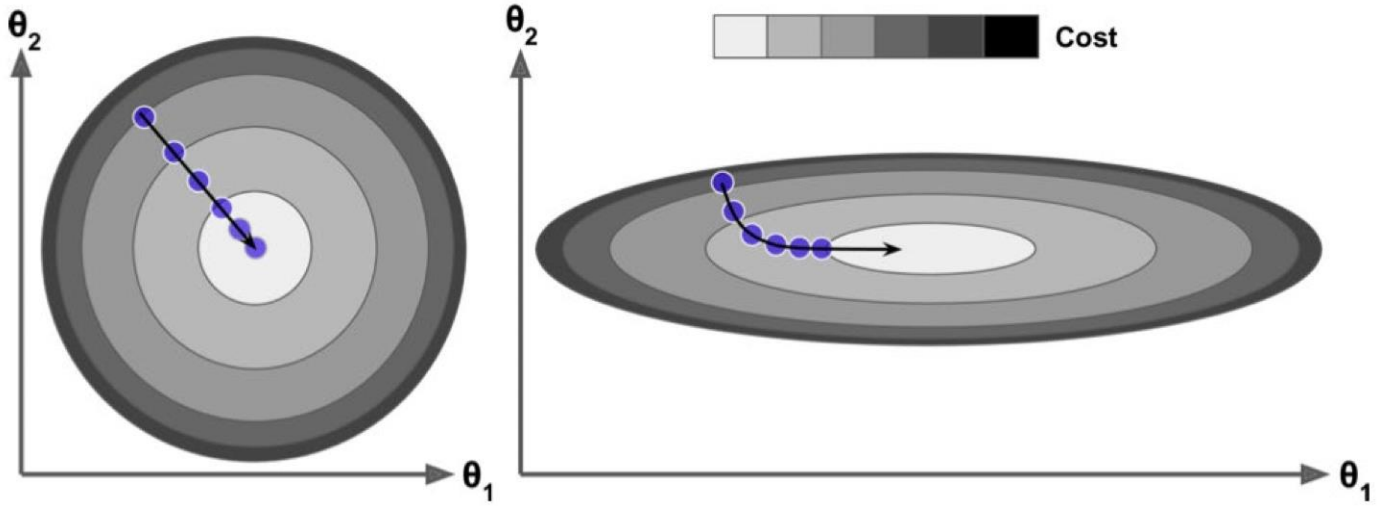
Variable Transformation

Now, you can see the job attribute is split into 12 new attributes where each of it denotes an instance belongs to that occupation or not

For example, job=unemployed, 0 denotes not unemployed, 1 denotes unempolyed. The count of 0 is 4393 and the count of 1 is 128



Gradient Descent with and without feature scaling



The MSE cost function for a Linear Regression model has the shape of a bowl, but it can be an elongated bowl if the features have very different scales.

The figure above shows Gradient Descent on a training set where features 1 and 2 have the same scale (on the left), and on a training set where feature 1 has much smaller values than feature 2 (on the right).

Variable Transformation

Standardization

Standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1.

Standardization assumes that your data has a Normal distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Normal.

Variable Transformation

The formula is given below.

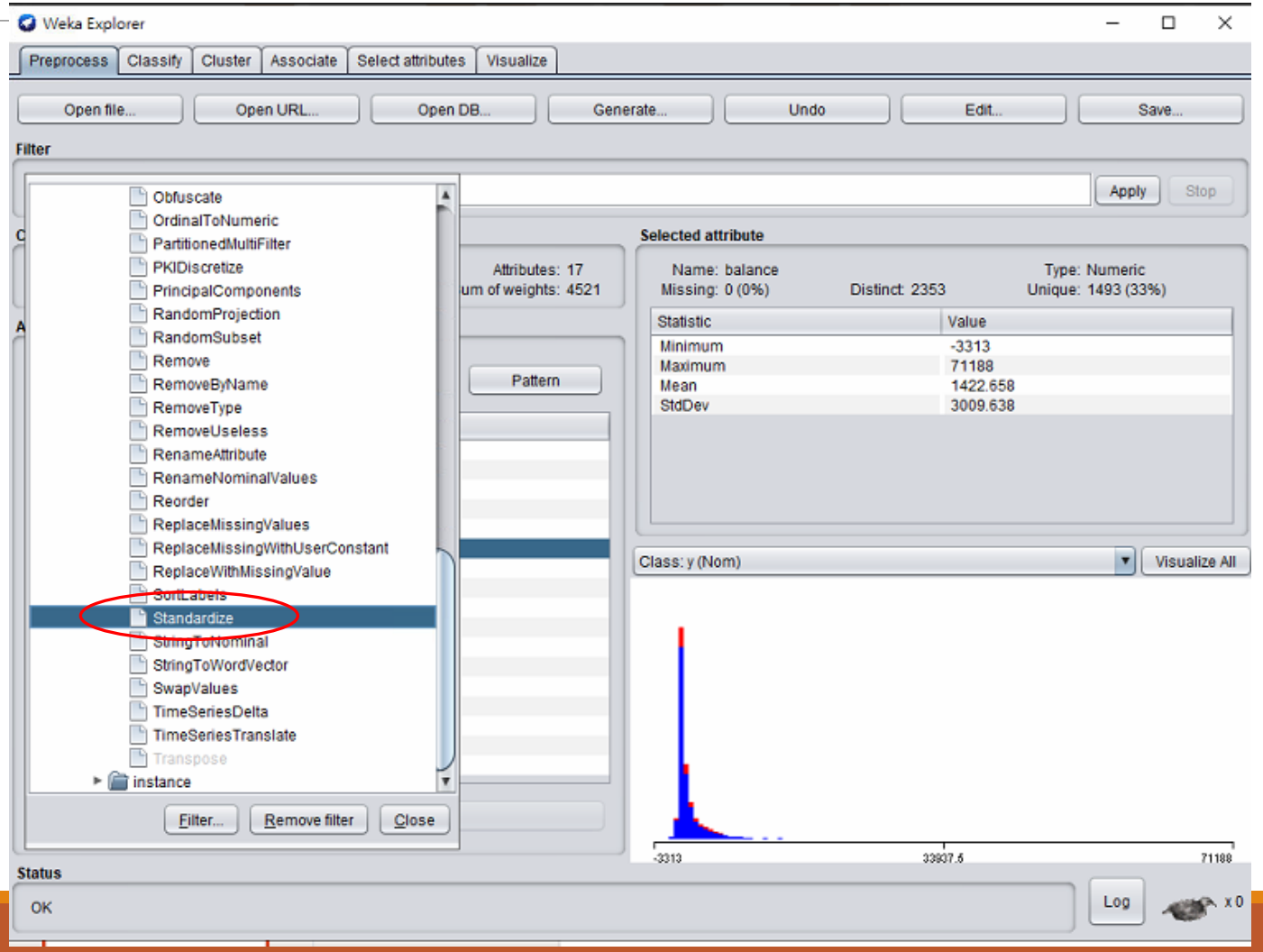
$$z = \frac{x_i - \mu}{\sigma}$$

Variable Transformation

Click Choose, under

filters->unsupervised->attributes,

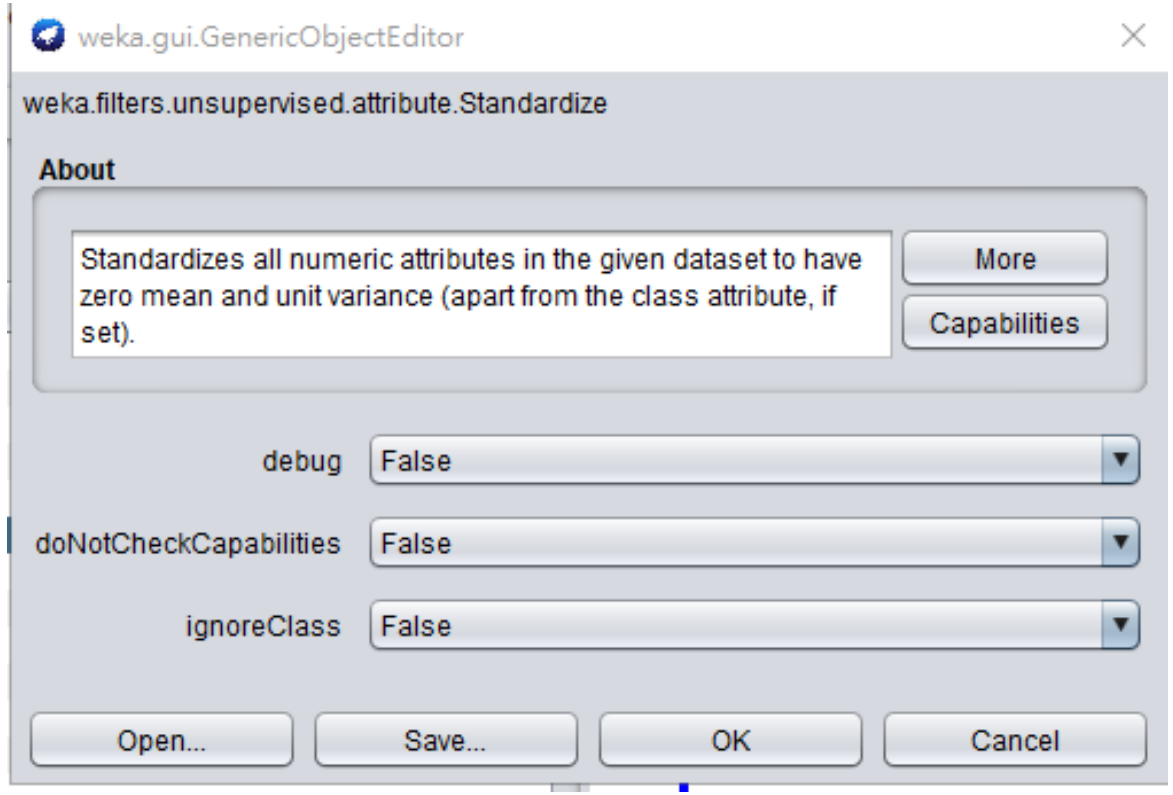
select Standardize



Variable Transformation

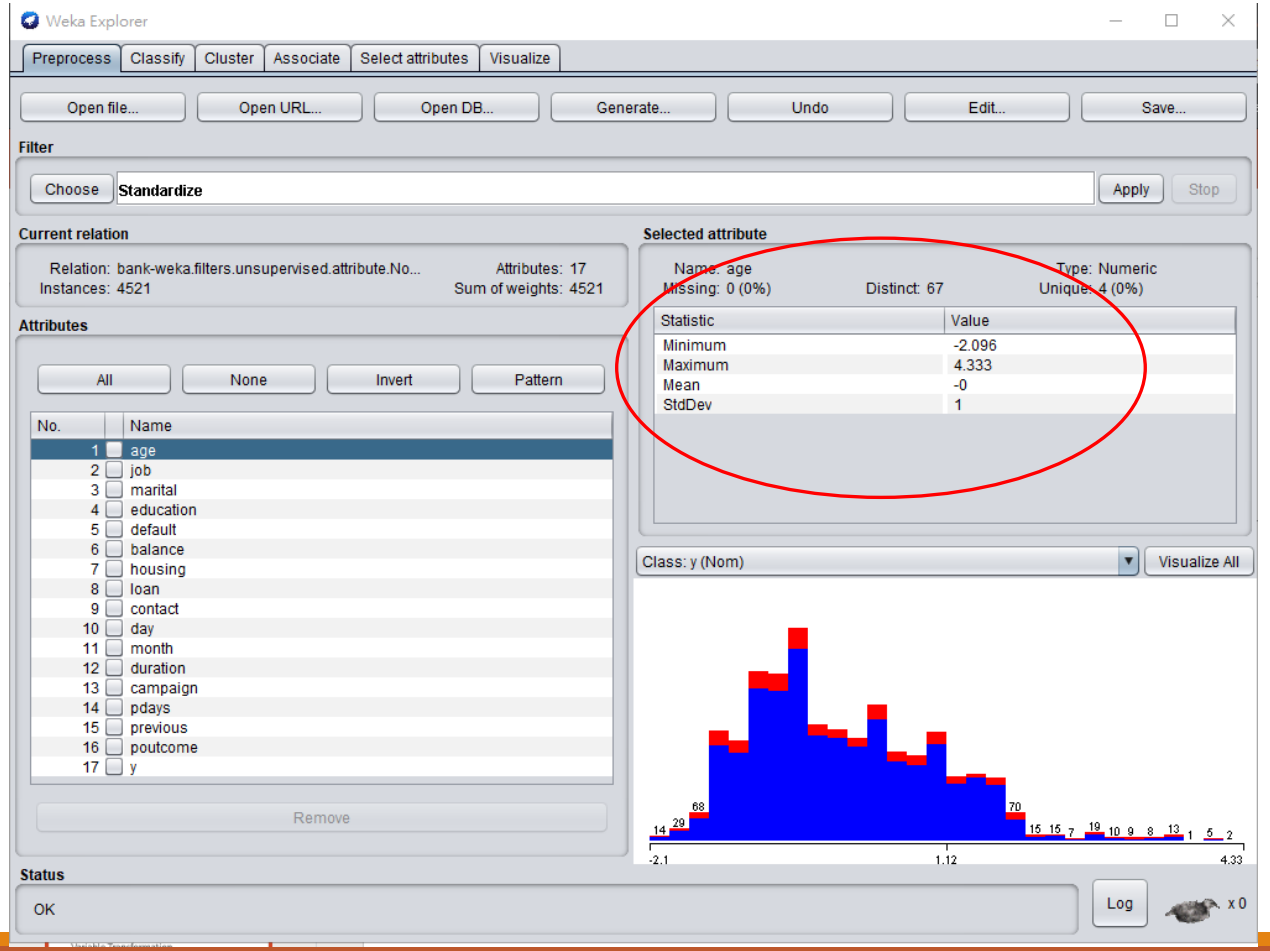
No setting is required in the configuration, simply click OK

Then, click Apply



Variable Transformation

Now, you can see the age attribute is rescaled. The mean is 0 and standard deviation is 1.



Variable Transformation

Normalization

Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.

Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Normal distribution. Also, some machine learning algorithms are sensitive so the scale of the data, rescale the data into a range of 0 to 1 can lower the effect of scale.

Variable Transformation

This is also called min-max normalization, the formula is given below.

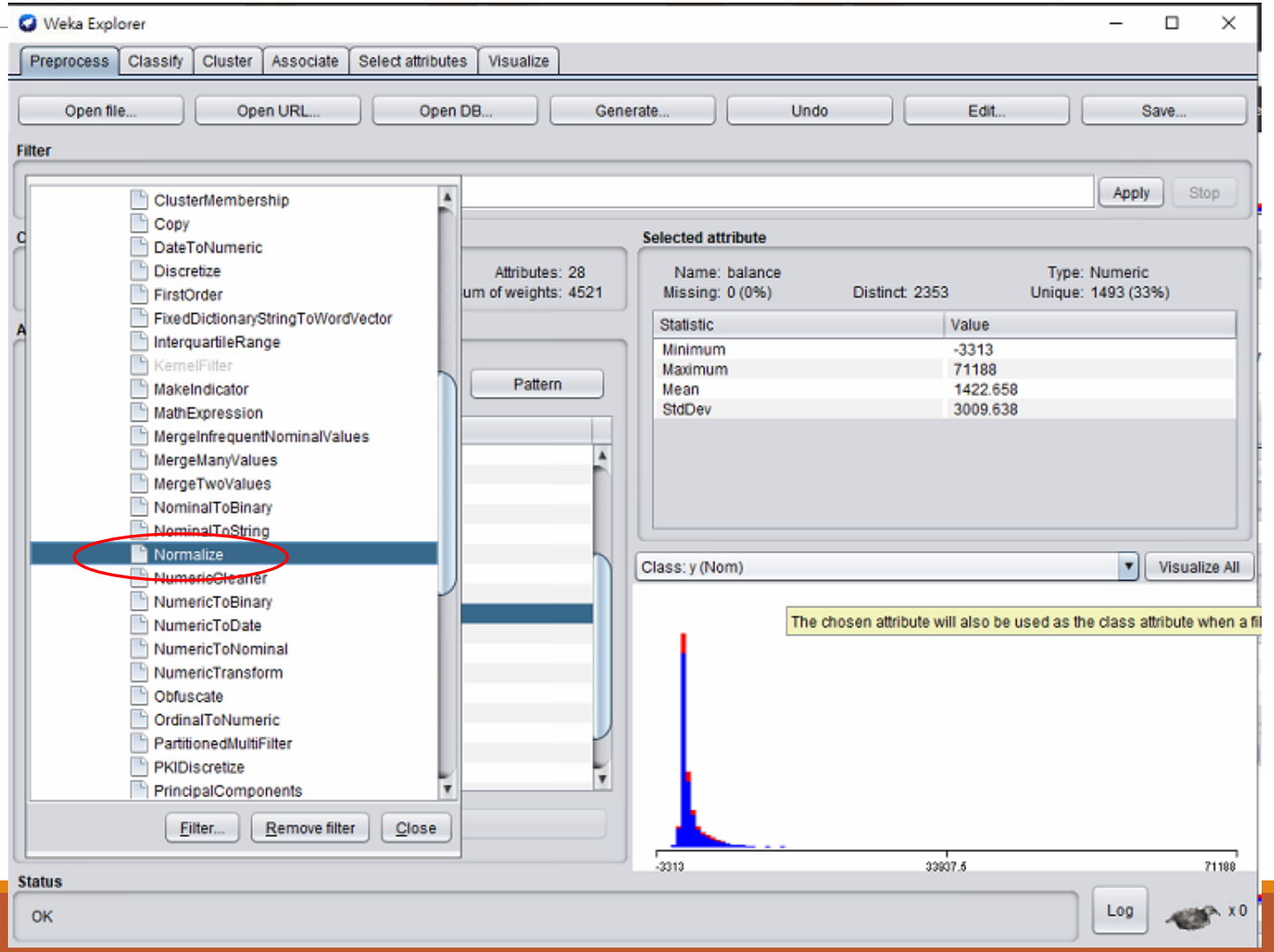
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Variable Transformation

Click Choose, under

filters->unsupervised->attributes,

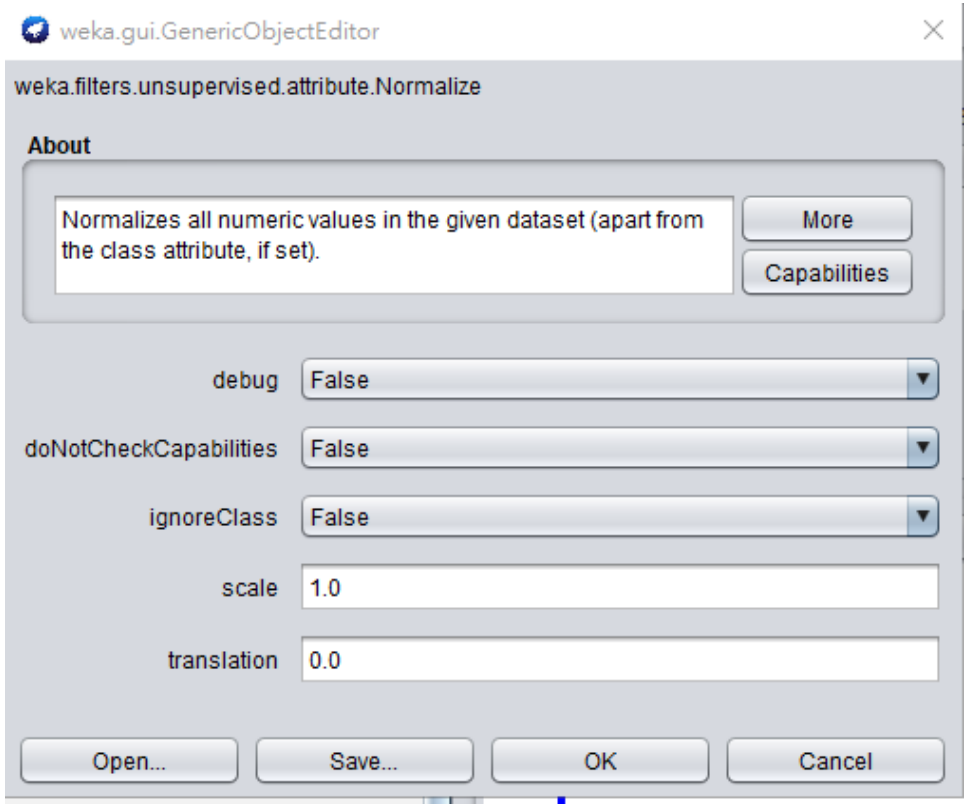
select Normalize



Variable Transformation

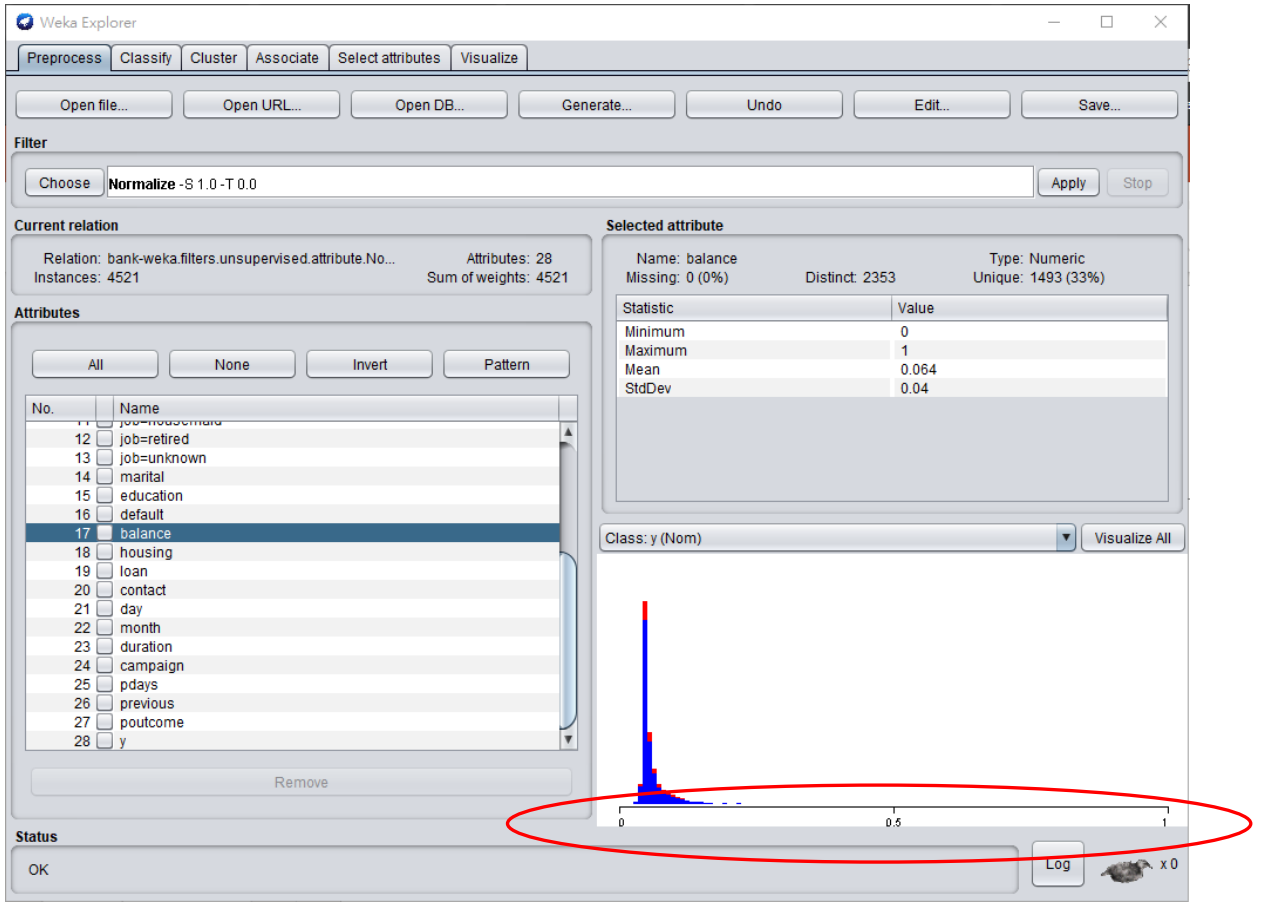
In the configuration, the scale and translation mean the maximum and minimum value after normalization where the default value is 1 and 0 respectively. We leave it as default. Click OK

Then, click Apply



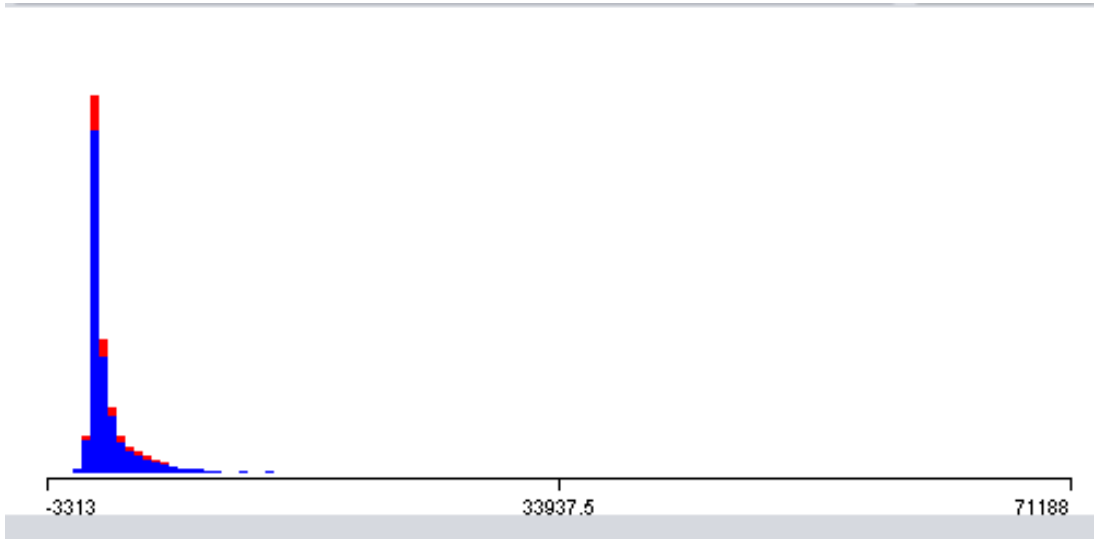
Variable Transformation

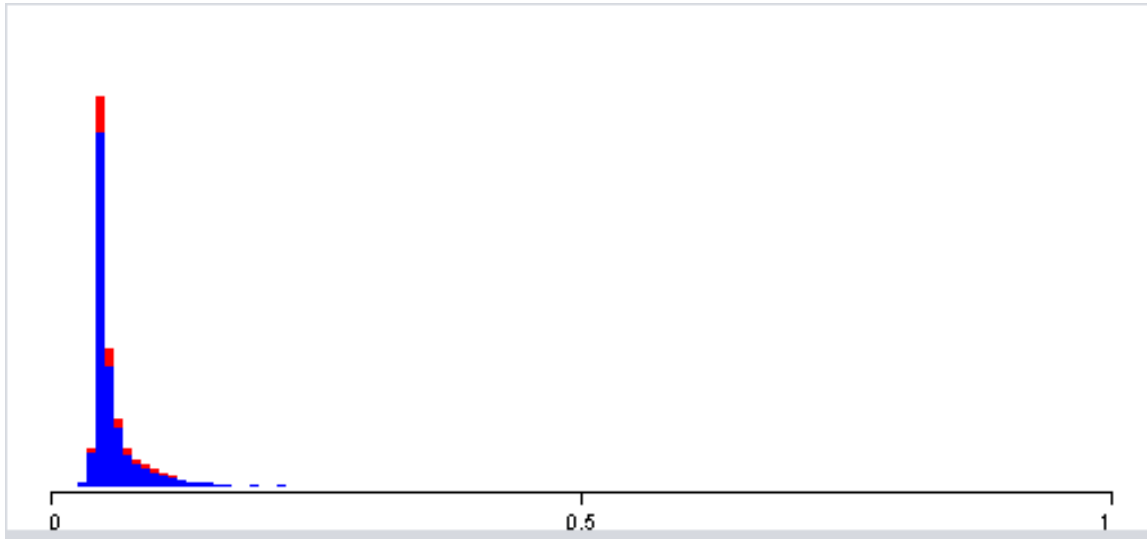
Now, you can see the balance attribute is rescale into the range of 0 to 1



Variable Transformation

This normalization does not affect the distribution of the data. For example, look at the graph of balance attribution, the shape is the same after performing the normalization.





What is Decision Tree?

A decision tree is decision making tool using a tree-like graph or model of decisions and their possible consequences such as event outcomes, resource costs, and utility.

All the conditional control statements used in the decision tree can be displayed for easily understand the logic behind it.

What is Decision Tree?

A decision tree is a flowchart-like structure contains these components:

- each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails)
- each branch represents the outcome of the test
- each leaf node represents a class label (decision taken after computing all attributes).

The paths from root to leaf represent classification rules.

What is Decision Tree?

This is an example of a decision tree for the target variable response. This variable has two labels: 1 for response and 0 for no response.

Each node determine which attribute should be used for splitting the dataset based on the information gain. In this example, Node 1 uses Income as splitting attribute, <$25k go to Node 2 and >= $25k go to Node 3.

There are 4 leaf nodes (Node 4-7) for determine the predicted label.