

VNUHCM – UNIVERSITY OF SCIENCE



FACULTY OF INFORMATION TECHNOLOGY



LAB 01

SUBJECT: DATA MINING

TEACHER: LÊ HOÀI BẮC

NGUYỄN THỊ THU HẰNG

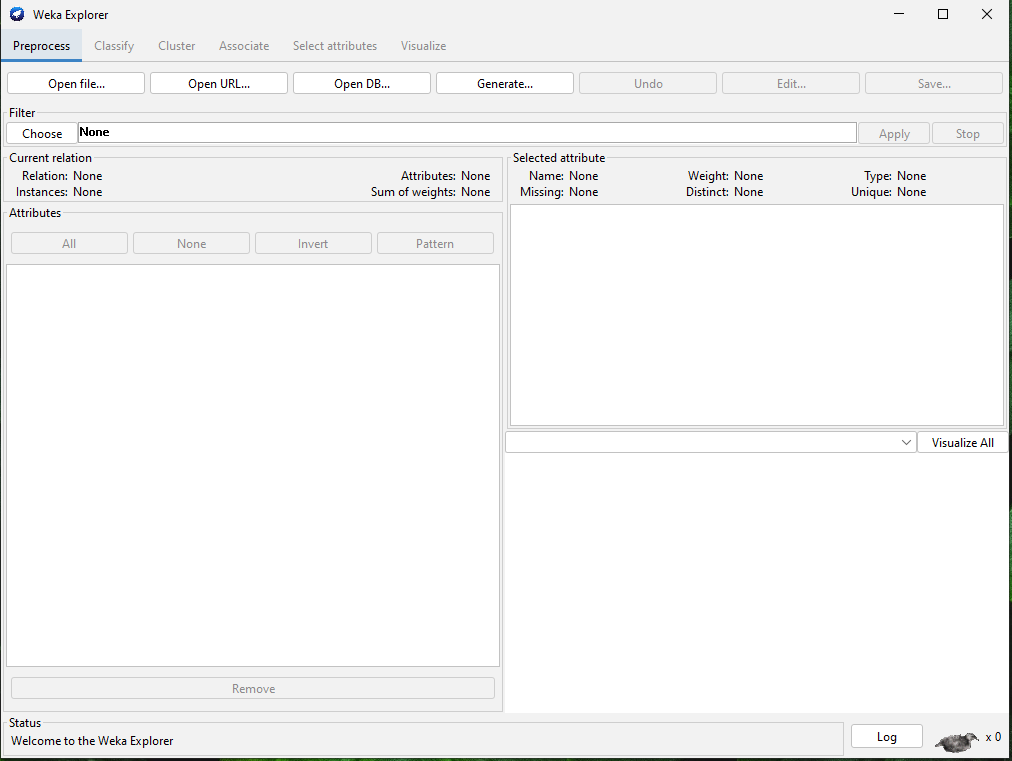
NGUYỄN BẢO LONG

HỒ CHÍ MINH – 2023

**I/ Information:**

|  |  |
| --- | --- |
| Full Name | MSSV |
| Bùi Duy Bảo | 20127444 |
| Nguyễn Thái Bảo | 20127448 |

**II/ Task:**

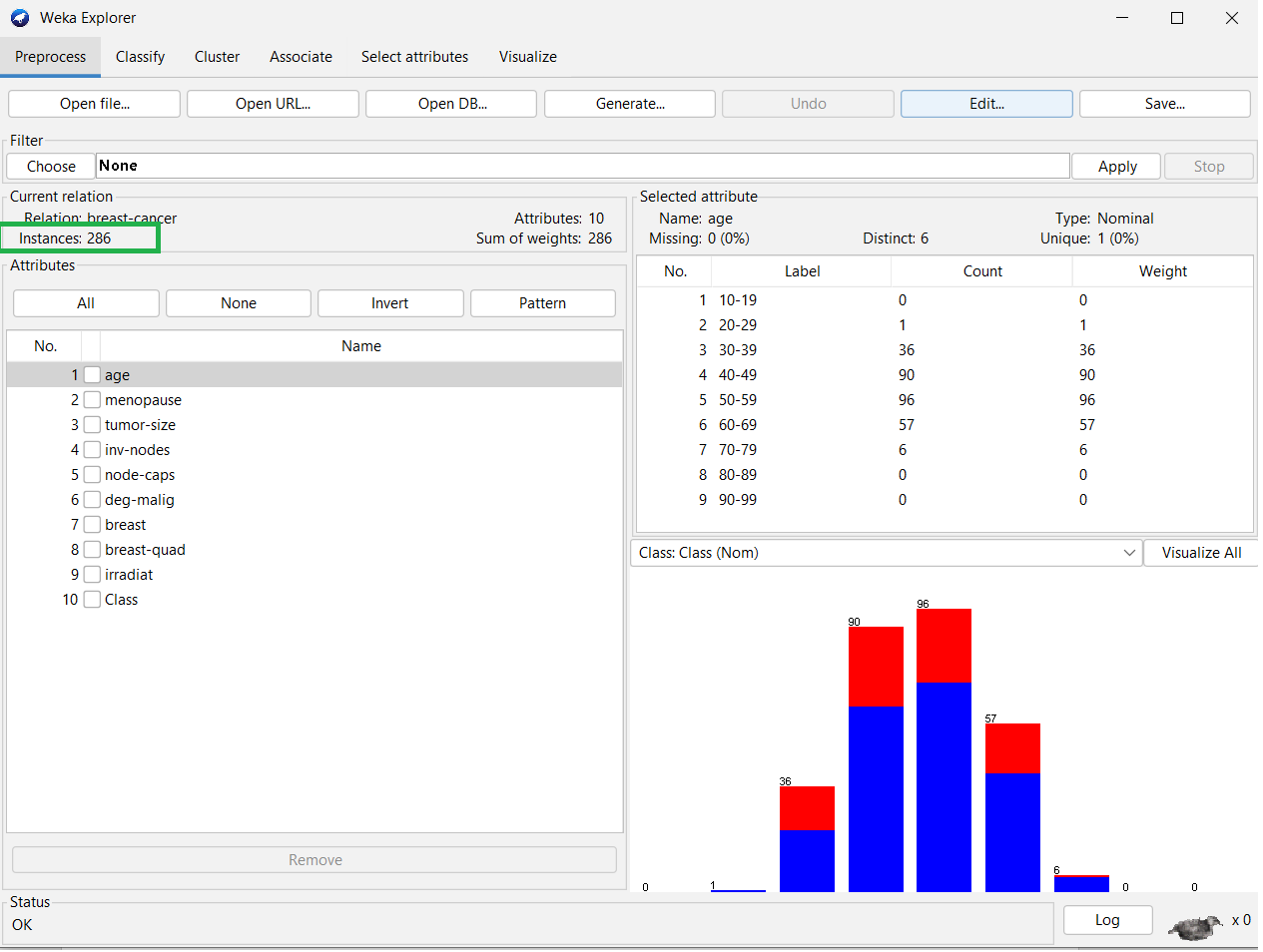


3.2 Getting Acquainted With WEKA

3.2.1 Exploring Breast Cancer data set

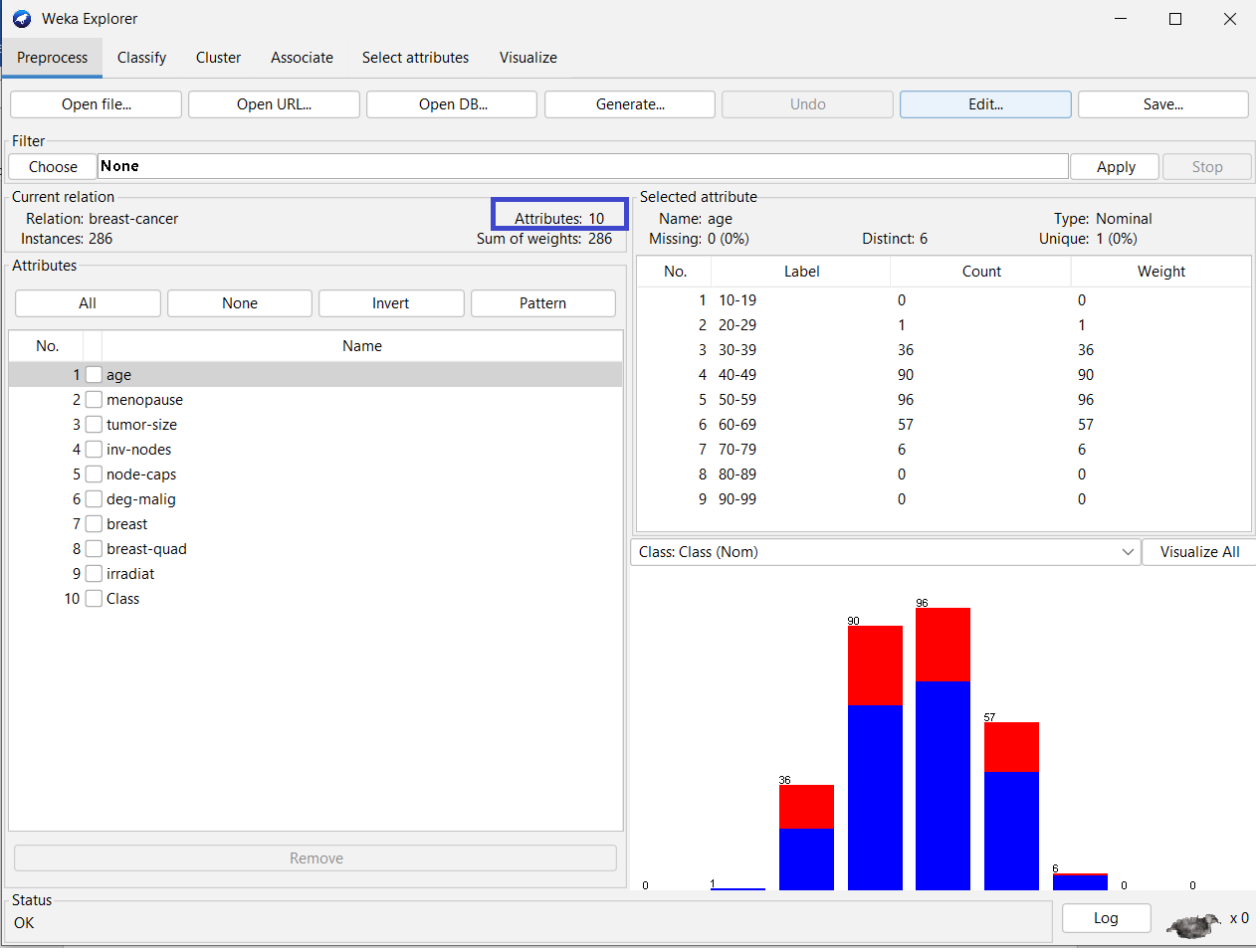
– How many instances does this data set have?

This data set has 286 instances.

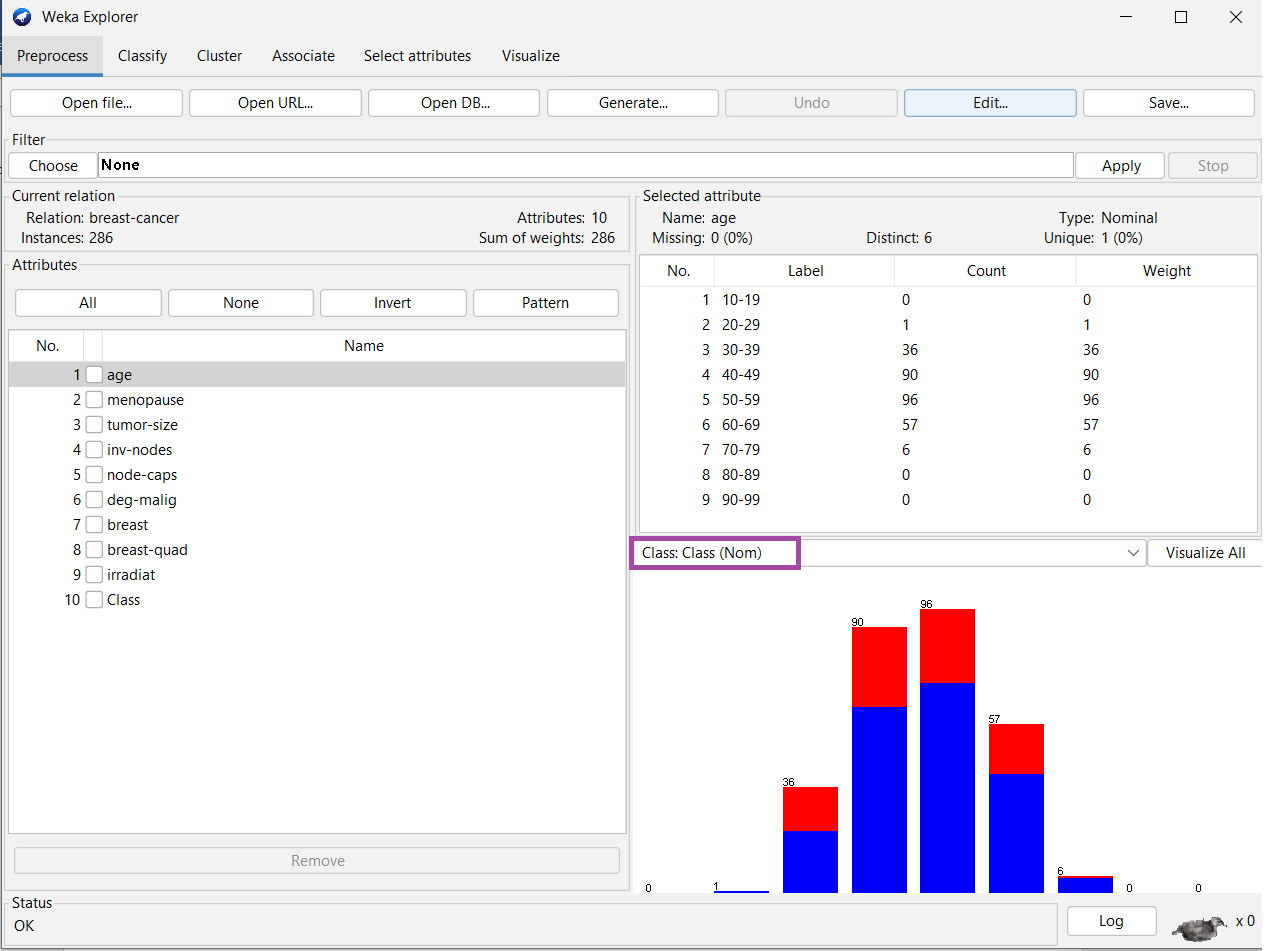


– How many attributes does this data set have?

This data set has 10 attributes



– Which attribute is used for the label? Can it be changed? How?

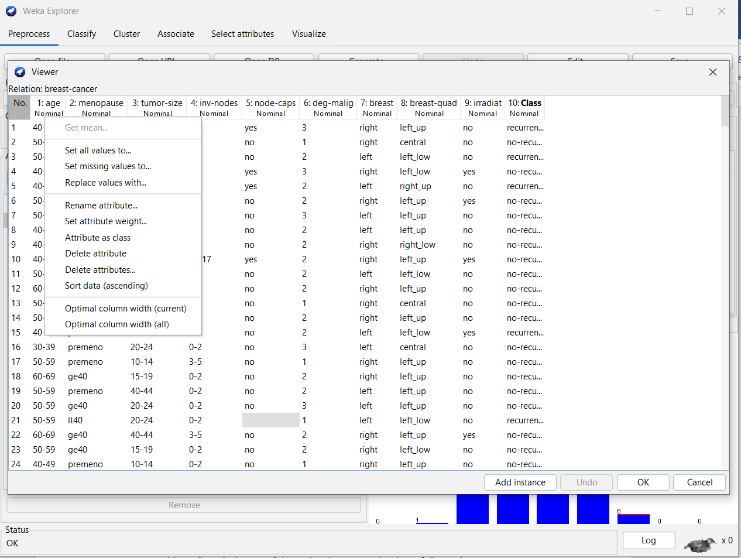
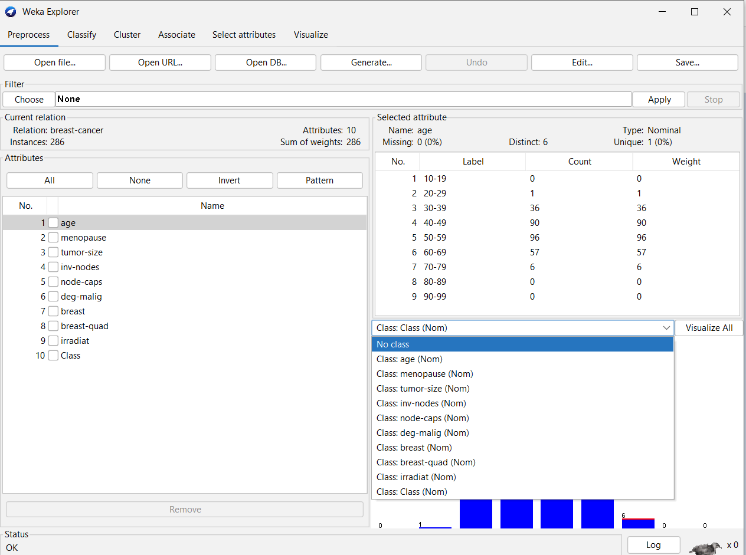


Attribute "Class" which is used to classify the data. The Class attribute distinguishes between benign (non-cancerous) and malignant (cancerous) tumors in breast cancer patients.

It is possible to change the Class attribute in WEKA by editing the ARFF file that contains the dataset.

C1: Edit -> Choose attribute is used for label -> Attribute as class

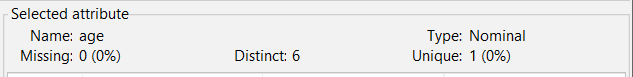
C2: Choose in main screen

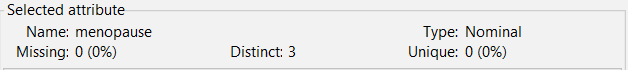


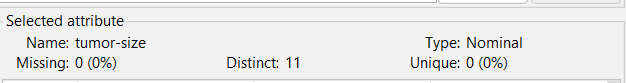
– What is the meaning of each attribute?

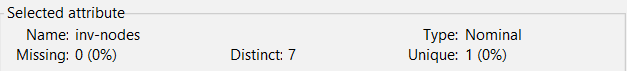
1. Class: This attribute is the class label, which indicates the diagnosis of the breast cancer patient. It has two possible values: "benign" for non-cancerous tumors and "malignant" for cancerous tumors.
2. Age: This attribute is the age of the patient in years at the time of diagnosis.
3. Menopause: This attribute indicates whether the patient has experienced menopause or not at the time of diagnosis. It has three possible values: "premeno" for premenopausal patients, "lt40" for patients with menopause at age less than 40, and "ge40" for patients with menopause at age greater than or equal to 40.
4. Tumor-size: This attribute is the size of the tumor in millimeters at the time of diagnosis.
5. Inv-nodes: This attribute is the number of axillary lymph nodes that contain cancer at the time of diagnosis.
6. Nodecaps: This attribute indicates whether there is evidence of cancer cells in the lymph node capsules or not. It has two possible values: "yes" for cancer cells detected and "no" for cancer cells not detected.
7. Deg-malig: This attribute is the degree of malignancy of the tumor, which is determined by examining the tumor cells under a microscope. It has three possible values: 1 for mild malignancy, 2 for moderate malignancy, and 3 for severe malignancy.
8. Breast: This attribute indicates which breast the tumor was detected in. It has two possible values: "left" for the left breast and "right" for the right breast.
9. Breast-quad: This attribute is the location of the tumor within the breast. It is divided into four quadrants: "left-up", "left-low", "right-up", and "right-low".
10. Irradiat: This attribute indicates whether the patient received radiation therapy after surgery or not. It has two possible values: "yes" for radiation therapy received and "no" for radiation therapy not received.

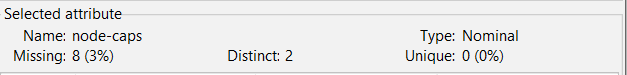
– Let’s investigate the missing value status in each attribute and describe in general ways to solve the problem of missing values.

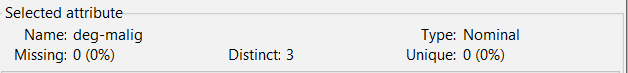


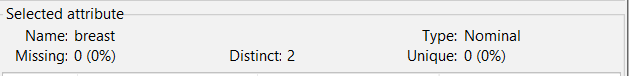


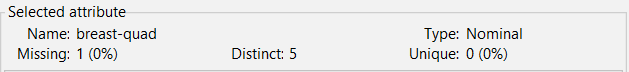


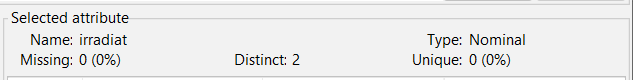


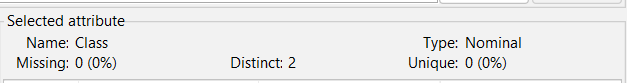












There are several ways to handle missing values in a dataset, including:

* Delete rows containing missing values: This is the simplest method, but it can lead to loss of important data and degrade the quality of the data file
* Use mean: If the attribute is numeric, you can use the average of the attribute to fill in the missing values. This method can help retain the distribution of the original data, but it can be affected by outliers.
* Use the median: Similar to the mean, if the attribute is numeric, you can use the median of the attribute to fill in the missing values. This method can help retain the distribution of the original data, while minimizing the effect of outliers.
* Use the mode: If the attribute is categorical, you can use the attribute's most common value to fill in the missing values. This method can help to retain the categorization of the original data.
* Use predictive models: You can use predictive models to predict missing values based on other attributes.

– Let’s propose solutions to the problem of missing values in the specific attribute.

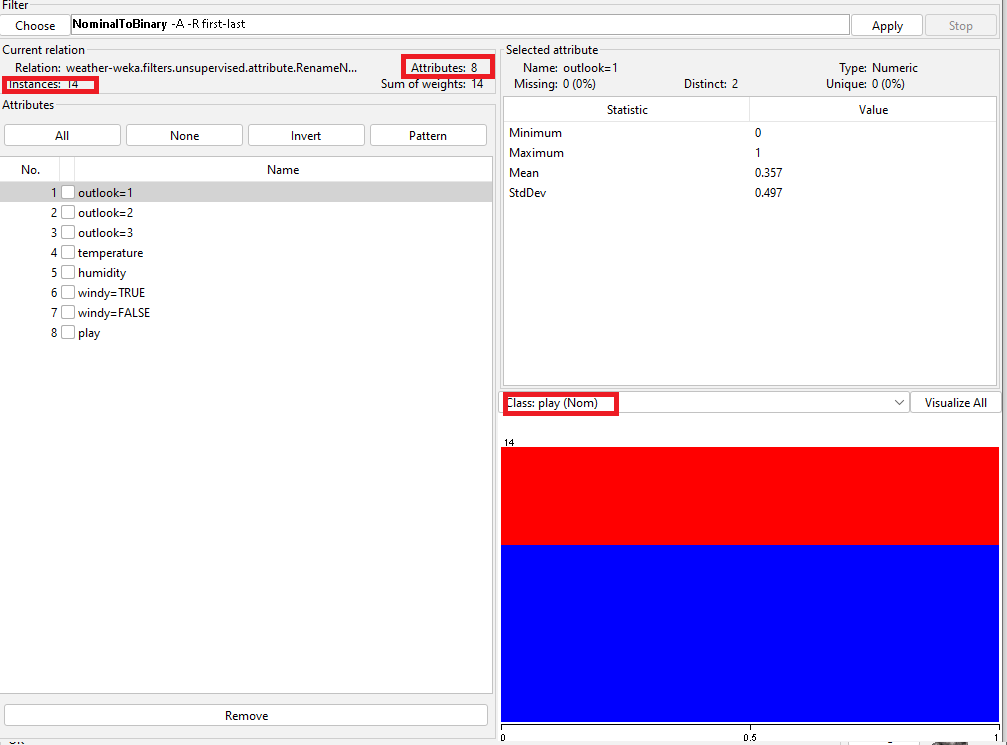
Node-caps: Using the Mode

Breast-quad:Using the Mode

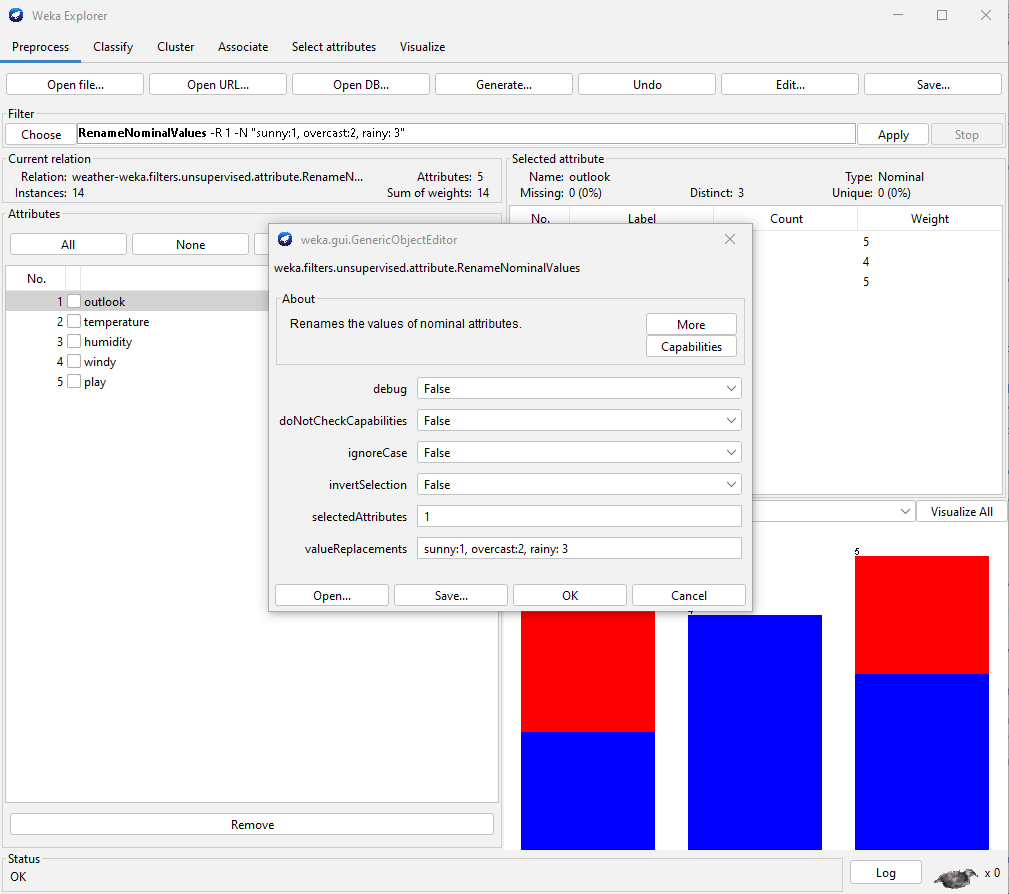
– Let’s explain the meaning of the chart in the WEKA Explorer. Setting the title for it and describing its legend.

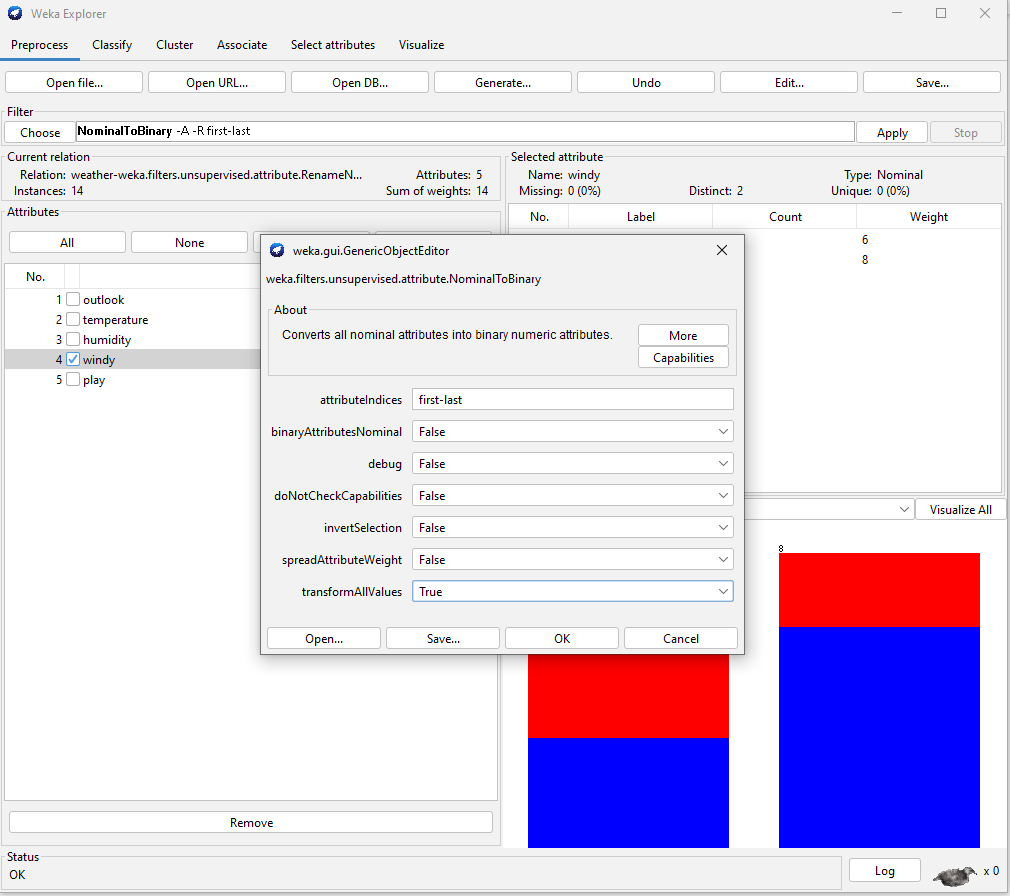
3.2.2 Exploring Weather data set

How many attributes does this data set have? How many samples? Which attributes have data type categorical? Which attributes have a data type that is numerical? Which attribute is used for the label?

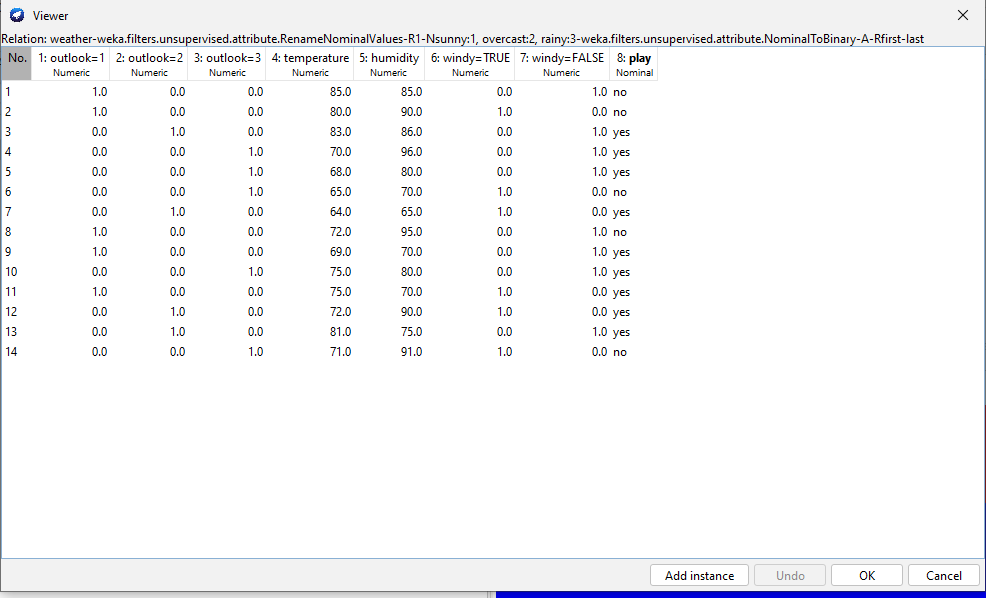


* Number of attributes: 8
* Number of samples: 14
* Attributes with categorical data type: None. All attributes have numerical data type.
* Attributes with numerical data type: All attributes.
* Label attribute: The last attribute (i.e., attribute 8), which is "play", is used as the label attribute.



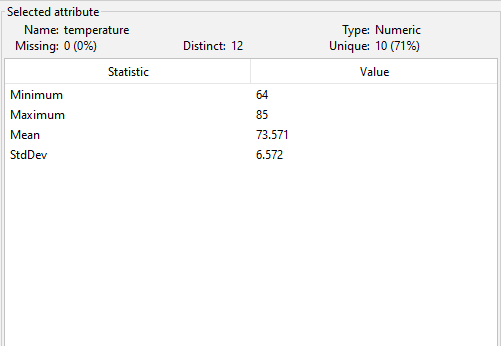


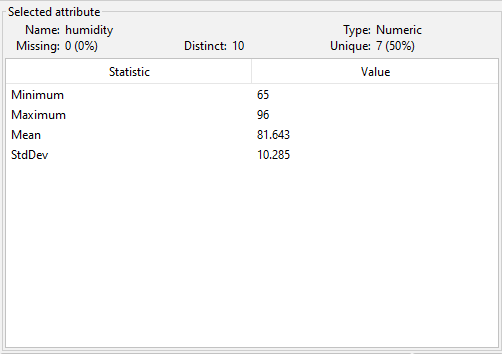
Note that the "weather.numeric.arff" dataset is a modified version of the well-known "weather" dataset, where the original nominal values have been replaced with numerical values. Therefore, all attributes in this dataset are numerical, even though they may represent categories such as "sunny", "overcast", and "rainy" in the original dataset. how to replace



Let’s list five-number summary of two attributes temperature and humidity. Does WEKA provide these values?

Yes, Weka provides five-number summary of the attributes in a dataset, including temperature and humidity



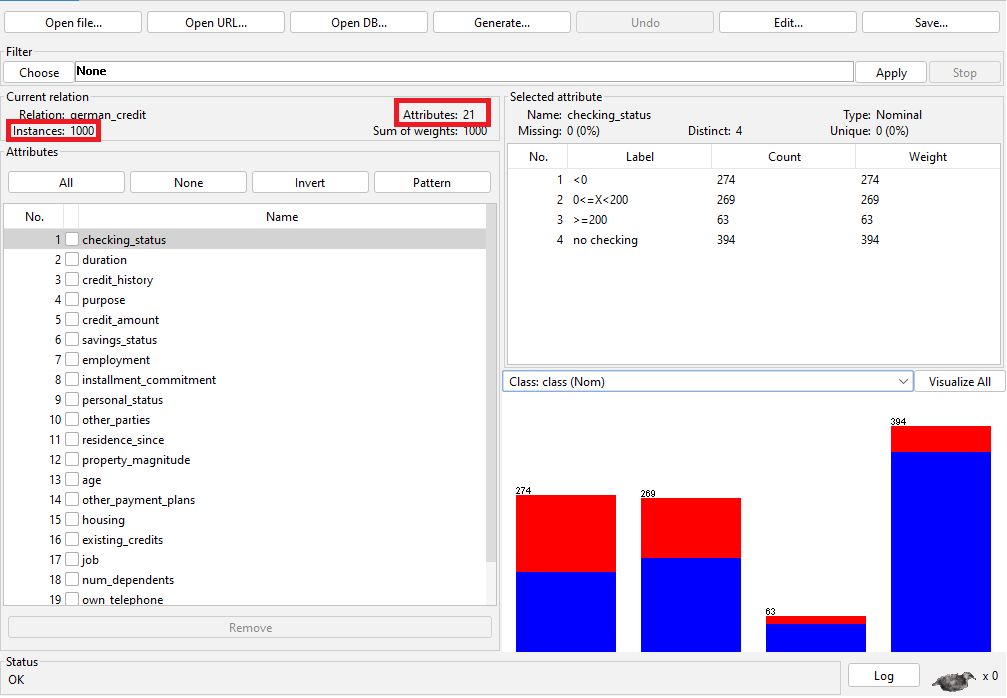


Let’s explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.

Let’s move to the Visualize tag. What’s the name of this chart? Do you think there are any pairs of different attributes that have correlated?

3.2.3 Exploring Credit in Germany data set

– What is the content of the comments section in credit-g.arff (when opened with any text editor) about? How many samples does the data set have? How many attributes? Describe any five attributes (must have both discrete and continuous attributes).



The content of the comments section in credit-g.arff is that the data was collected from a German bank and that the goal is to predict whether an applicant is a good or bad credit risk based on a set of attributes.

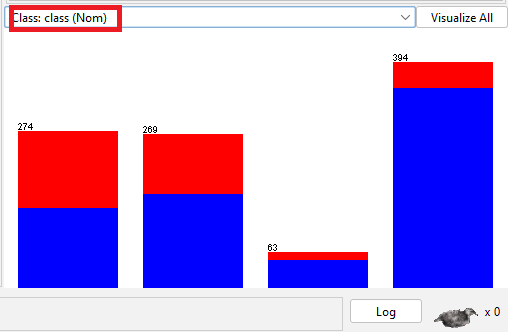
The dataset contains 1000 samples or instances.

There are 21 attributes in total

Here are brief descriptions of five attributes, including both discrete and continuous types:

1. Age: a continuous attribute representing the age of the applicant in years.
2. Checking\_account: a categorical attribute with four possible values indicating the status of the applicant's checking account: "no checking", "less than 0 DM", "0 to 200 DM", or "greater than 200 DM".
3. Credit\_history: a categorical attribute with five possible values indicating the credit history of the applicant: "no credit history", "all loans at bank paid back", "existing loans paid back", "delay in paying off previous loans", or "critical account/other credits existing".
4. Credit\_amount: a continuous attribute representing the amount of the requested credit in DM.
5. Employment\_duration: a categorical attribute with five possible values indicating the length of the applicant's current employment: "unemployed", "less than 1 year", "1 to 4 years", "4 to 7 years", or "greater than 7 years".

– Which attribute is used for the label?



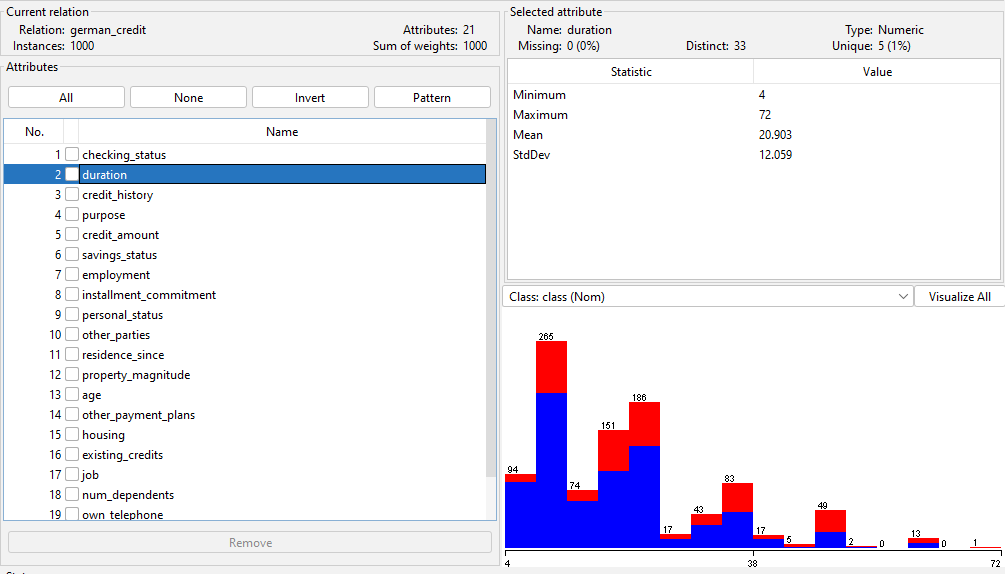
The attribute "class" is used for the label

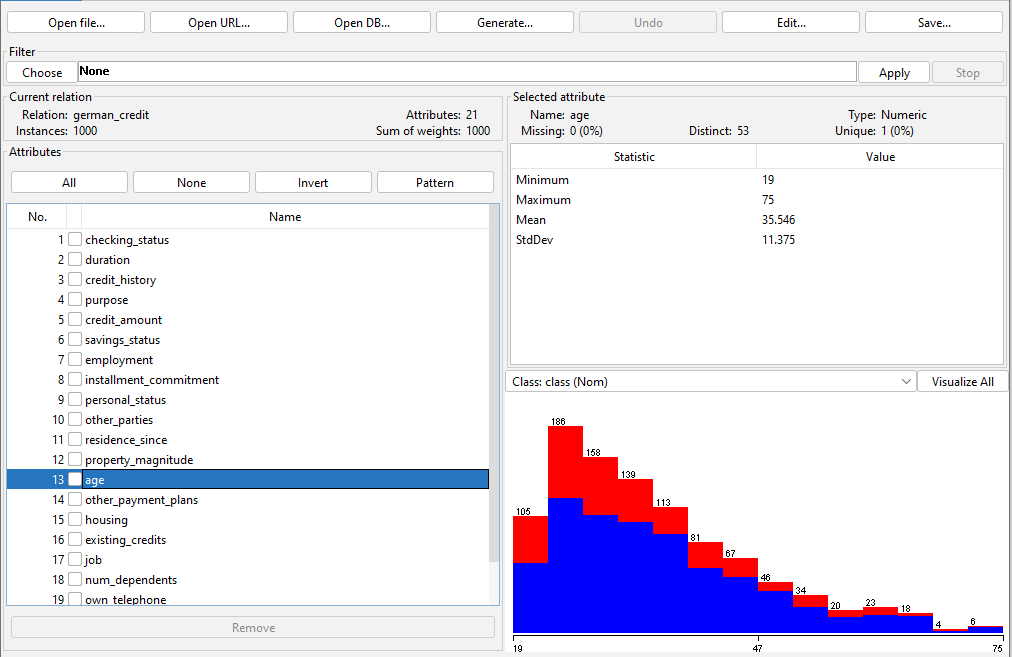
– Let’s describe the distribution of continuous attributes? (Left skewed or right skewed ?)

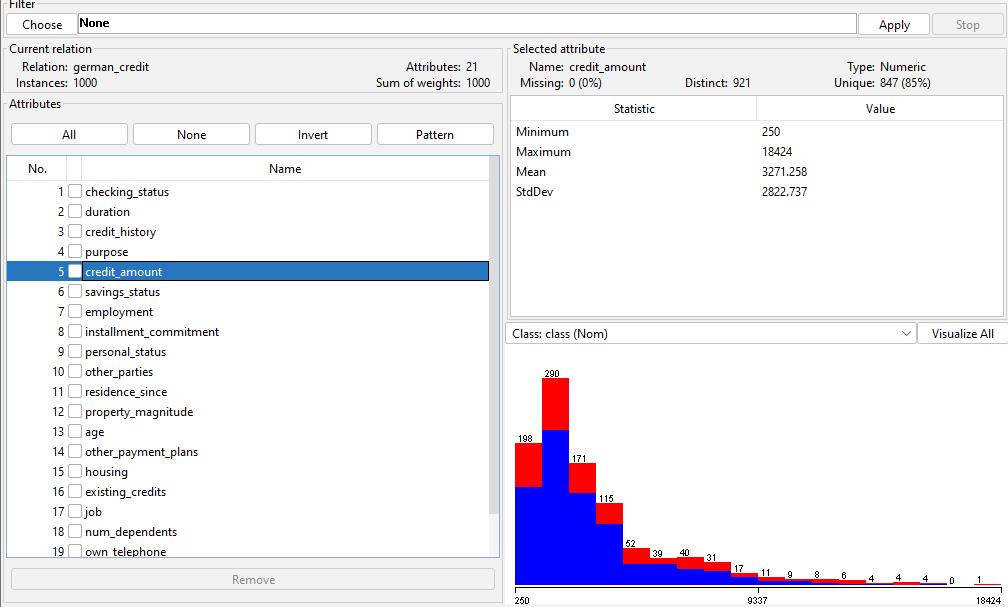
A distribution with a positive skew is said to be right-skewed, meaning that the tail of the distribution extends more to the right, while a distribution with a negative skew is left-skewed, meaning that the tail of the distribution extends more to the left.

Here are the skewness values for the continuous attributes in the **credit-g.arff** dataset:

* Age: right-skewed
* Credit\_amount: right-skewed
* Duration: right-skewed







– Let’s explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.

– Let’s move to the Select attributes tag. Describe all of the options for attribute selection. – Which options should be used to select the 5 attributes with the highest correlation?(Step-by-step description, with step-by-step photos and final results)

To select the 5 attributes with the highest correlation in WEKA, we can use the "AttributeSelection" filter. Here are the steps:

1. Open WEKA and load the dataset you want to analyze.
2. Click on the "Select attribute”
3. In the "Attribute Evaluator", select "CorrelationAttributeEval"
4. In the "Search Method", select the "Ranker"
5. Set the number of attributes to 5
6. Click on the "Start" button to apply the filter and wait for WEKA to generate the output.

