HO CHI MINH UNIVERSITY OF SCIENCE

VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

REPORT LAB 01

DATA PREPROCESSING - DATA EXPLORATION

Course: CSC14004 – DATA MINING AND APPLICATIONS

Lecturer:	Prof. Dr. Le Hoai Bac	
Teaching Assistants:	Ms. Nguyen Thi Thu Hang	
	Mr. Nguyen Bao Long	
Group students:	Dang Tien Dat	20127458
	Pham Thi Anh Phat	20127680

(2022 - 2023)

CONTENTS

I. GROUP INFORMATION

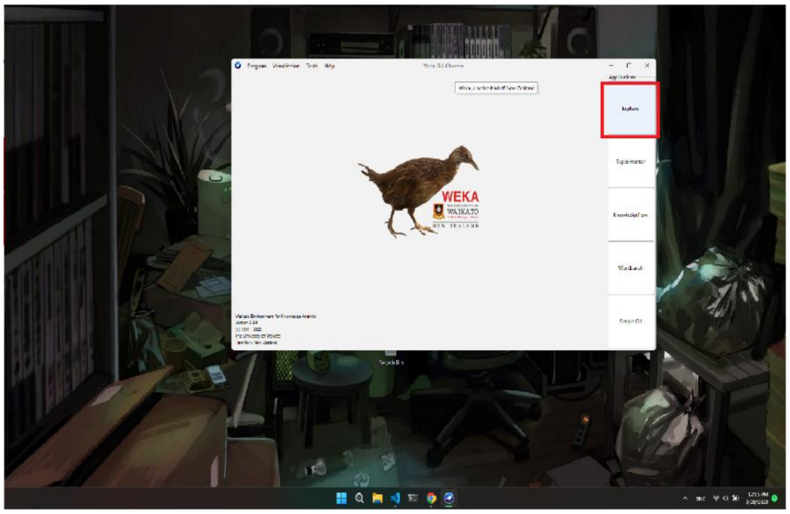
ID	Full Name	Contribution rate (%)
20127458	Dang Tien Dat	50
20127680	Pham Thi Anh Phat	50

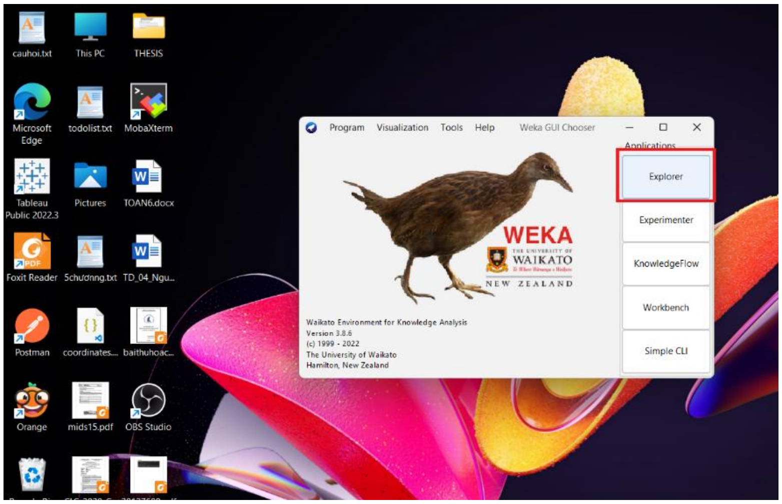
- We definitely completed all questions and requirements. 👏 👏

II. REQUIREMENTS

a. Install WEKA

a. Requirement 1: capture a screen that contains the ”**Explorer**” function in your desktop background.

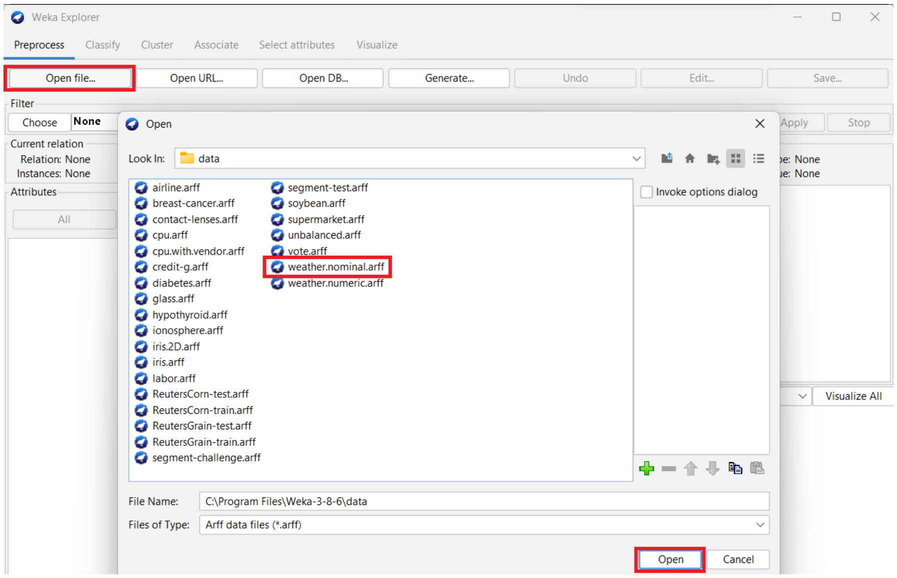




Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

b. Requirement 2: Explain the meaning of **Current Relation**, **Attributes**, and **Selected Attribute** in **Preprocess** tag. Briefly explain the meaning of the other tags in WEKA Explorer.

After load available dataset



Filter: None

Current relation
Relation: None
Instances: None

Attributes

Look In: data

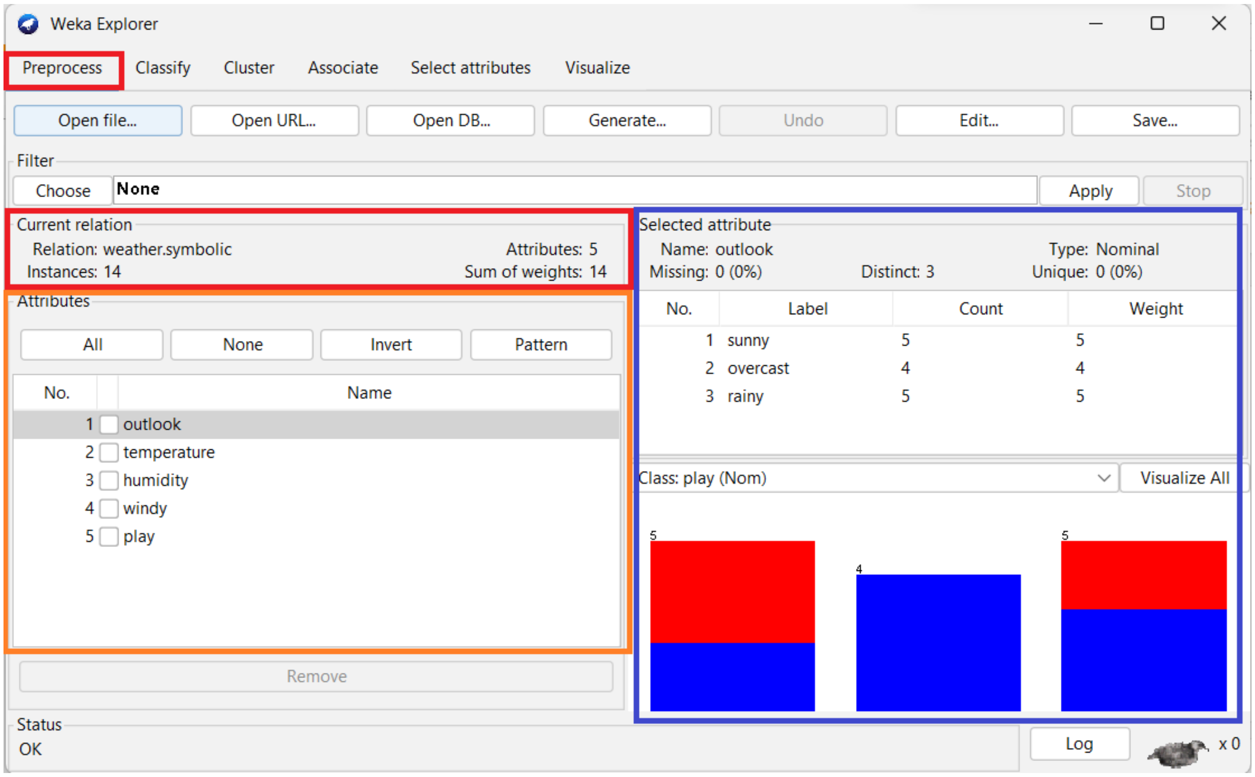
airline.arff
breast-cancer.arff
contact-lenses.arff
cpu.arff
cpu.with.vendor.arff
credit-g.arff
diabetes.arff
glass.arff
hypothyroid.arff
ionosphere.arff
iris.2D.arff
iris.arff
labor.arff
ReutersCorn-test.arff
ReutersCorn-train.arff
ReutersGrain-test.arff
ReutersGrain-train.arff
segment-challenge.arff
segment-test.arff
soybean.arff
supermarket.arff
unbalanced.arff
vote.arff
weather.nominal.arff
weather.numeric.arff

File Name: C:\Program Files\Weka-3-8-6\data

Files of Type: Arff data files (*.arff)

The screen displays:

1. PREPROCESS TAB



Filter: None

Current relation
Relation: weather.symbolic   Attributes: 5
Instances: 14   Sum of weights: 14

Attributes

No.	Name
1	□ outlook
2	□ temperature
3	□ humidity
4	□ windy
5	□ play

Selected attribute
Name: outlook   Type: Nominal
Missing: 0 (0%)   Distinct: 3   Unique: 0 (0%)

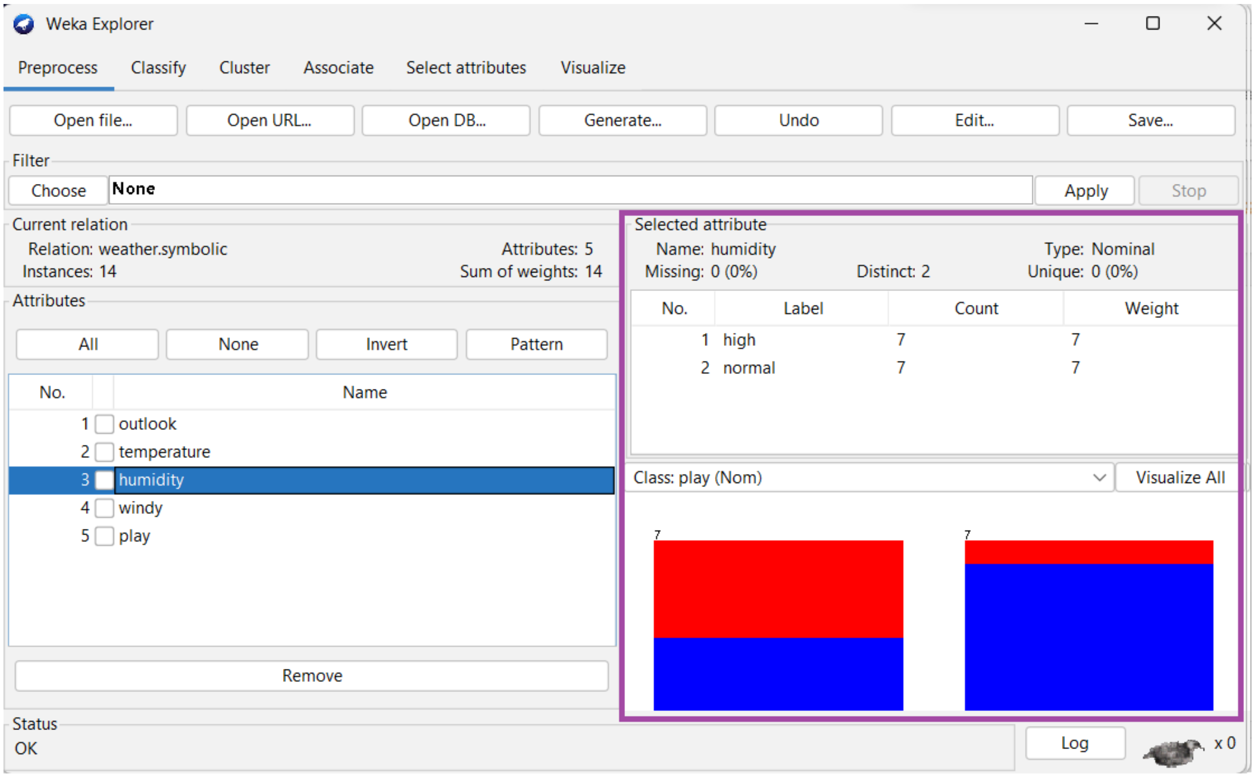
No.	Label	Count	Weight
1	sunny	5	5
2	overcast	4	4
3	rainy	5	5

Class: play (Nom)

Status
OK

- **Current relation:** shows the name of the of the database currently loaded as well as number of instances, number of attributes and sum of weights.

- **Attributes:** displays all fields in the database. The weather database has 5 attributes: outlook, temperature, humidity, windy, play. When clicking one of these attributes, **more details** on the attribute itself are displayed on the **Selected attribute**. Further more, we can remove feature by click on **Remove** button.



None

Current relation
Relation: weather.symbolic
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

No.	Name
1	□ outlook
2	□ temperature
3	□ humidity
4	□ windy
5	□ play

Selected attribute
Name: humidity
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

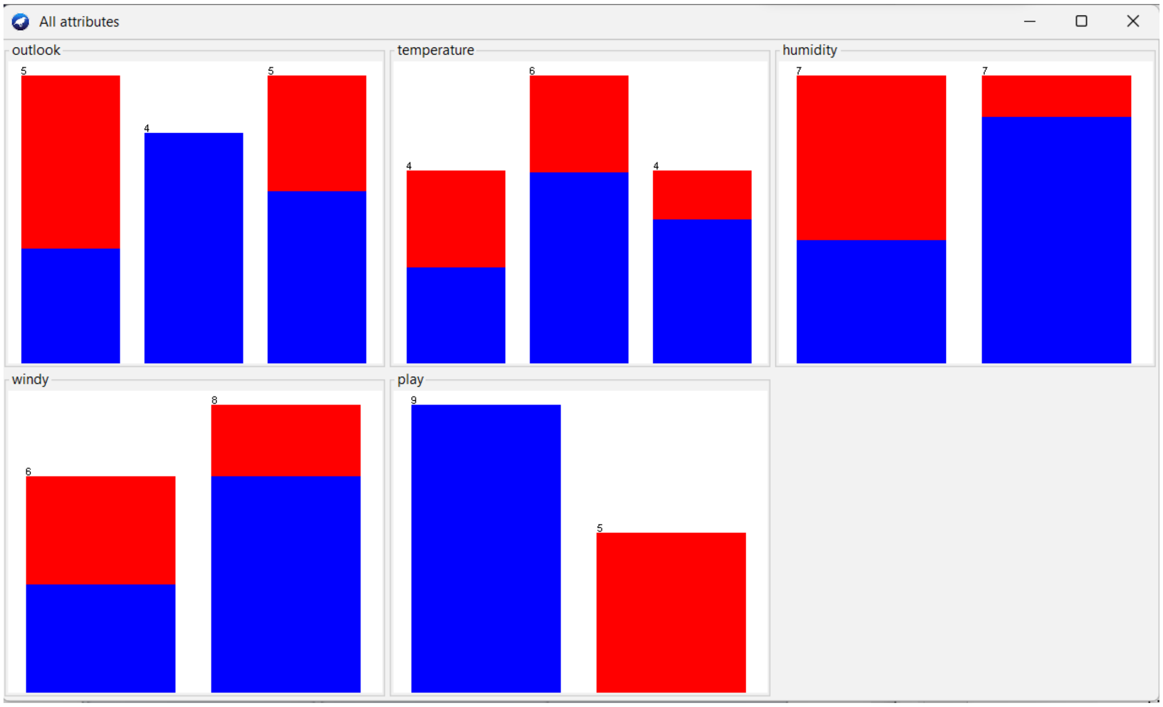
No.	Label	Count	Weight
1	high	7	7
2	normal	7	7

Class: play (Nom)

Status
OK

x 0

- **Selected attribute** displays:
  c. Name and type of chosen attribute.
  d. Number of missing value, distinct value and unique value of this attribute.
  e. *For example:* type of humidity is Nominal, no missing value, 2 distinct values with no unique value.
  f. The table below shows the nominal values for this humidity field as high and normal, it also shows count and weight for each value.
  g. The bottom subwindow visualize representation of the class values.
  h. If you click on **Visualize All**, all features in one single window as shown here:

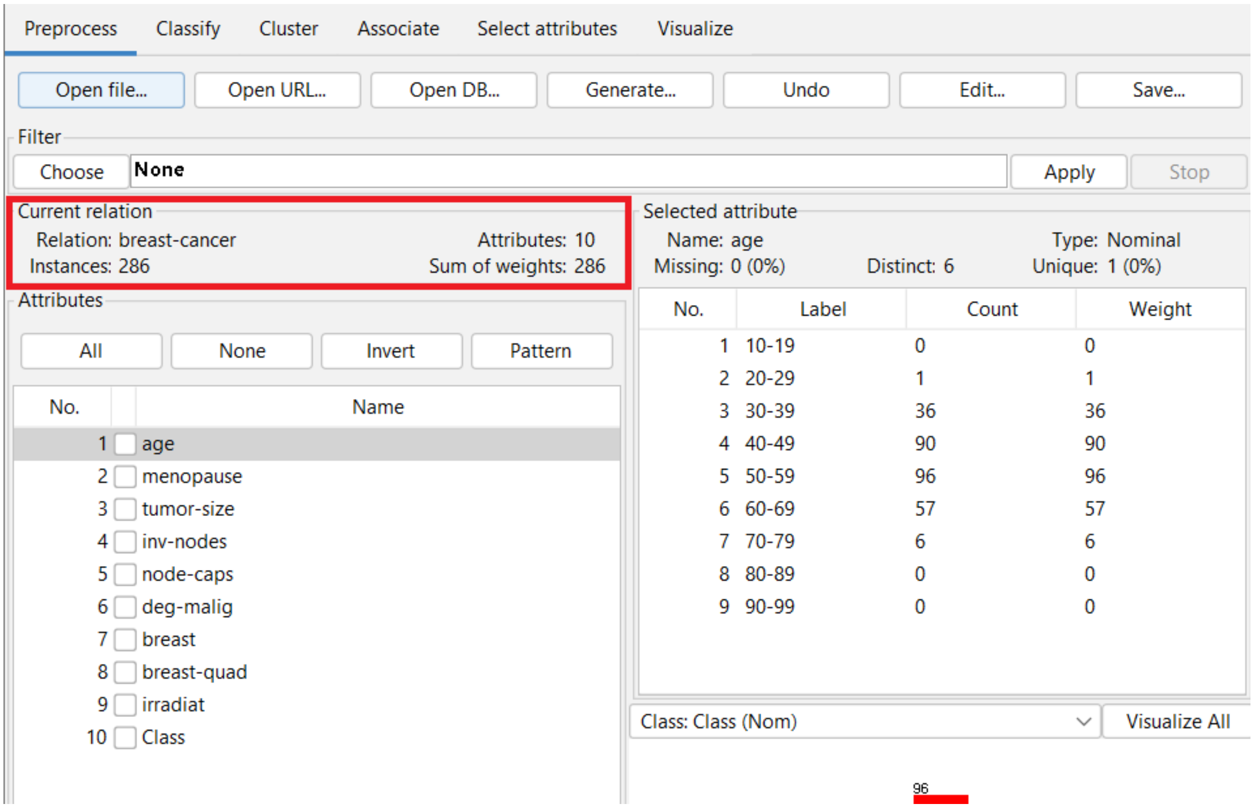


2. **ANOTHER TABS IN WEKA EXPLORER:**
    a. **CLASSIFY TAB:** provides you several both supervised and unsupervised machine learning algorithms for the classification of your data, such as Linear Regression, Logistic Regression, Support Vector Machines, Decision Trees, RandomTree, RandomForest, NaiveBayes, and so on.
    b. **CLUSTER TAB:** provides several clustering algorithms such as SimpleKMeans, FilteredClusterer, HierarchicalClusterer, and so on.
    c. **ASSOCIATE TAB:** several associate algorithms: Apriori, FilteredAssociator and FPGrowth.
    d. **SELECT ATTRIBUTES:** allows you feature selections based on several algorithms such as ClassifierSubsetEval, PrinicipalComponents, etc.
    e. **VISUALIZE:** visualize your processed data for analysis**.**

b. Getting Acquainted With WEKA

❖ Exploring Breast Cancer data set: **breast cancer.arff**

a. How many instances does this data set have? How many attributes does this data set have?



Filter: None

Current relation

Relation: breast-cancer   Attributes: 10
Instances: 286   Sum of weights: 286

Selected attribute

Name: age   Type: Nominal
Missing: 0 (0%)   Distinct: 6   Unique: 1 (0%)

Attributes

No.	Name
1	□ age
2	□ menopause
3	□ tumor-size
4	□ inv-nodes
5	□ node-caps
6	□ deg-malig
7	□ breast
8	□ breast-quad
9	□ irradiat
10	□ Class

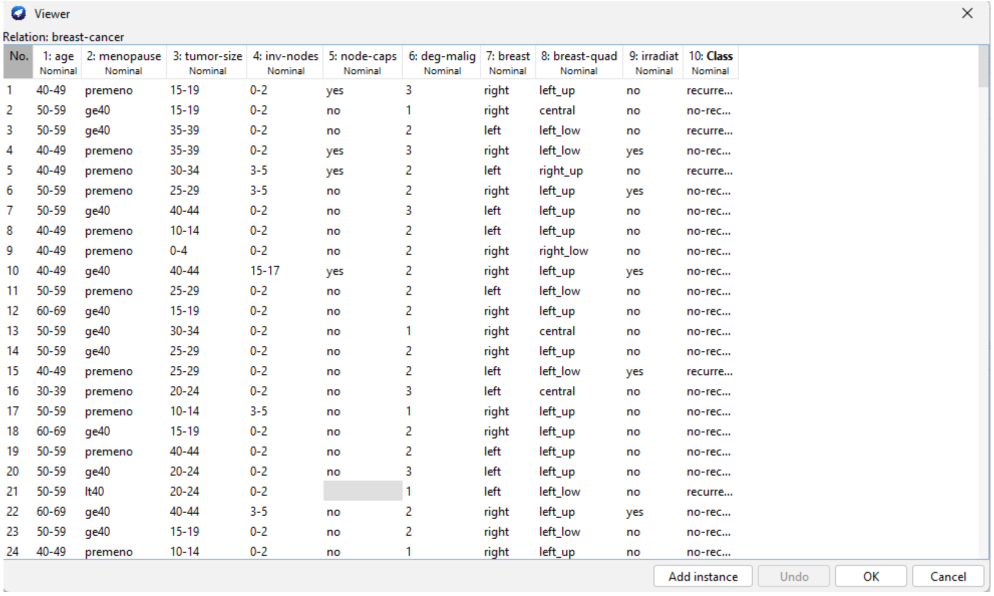
No.	Label	Count	Weight
1	10-19	0	0
2	20-29	1	1
3	30-39	36	36
4	40-49	90	90
5	50-59	96	96
6	60-69	57	57
7	70-79	6	6
8	80-89	0	0
9	90-99	0	0

Class: Class (Nom)

⇨ Data set has **286 instances** with **10 attributes**. The information is in the red box.

b. <u>Which attribute is used for the label? Can it be changed? How?</u>

⇨ The **label** is **Class** attribute.



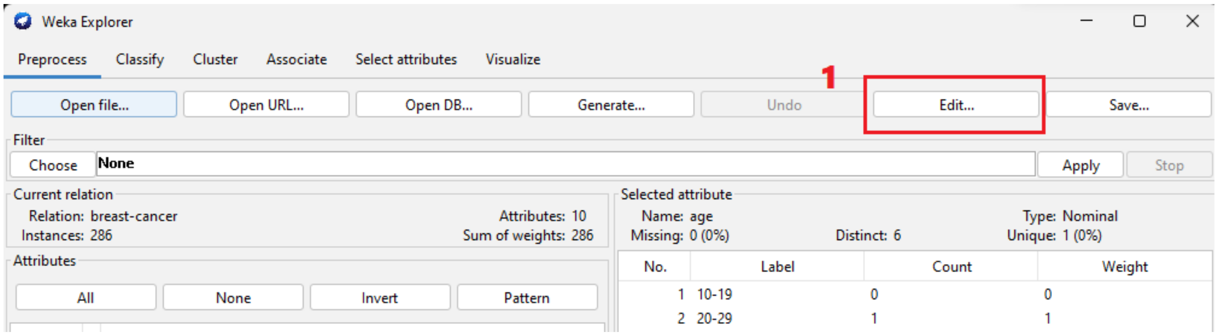
Relation: breast-cancer

No.	1: age Nominal	2: menopause Nominal	3: tumor-size Nominal	4: inv-nodes Nominal	5: node-caps Nominal	6: deg-malig Nominal	7: breast Nominal	8: breast-quad Nominal	9: irradiat Nominal	10: **Class** Nominal
1	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurre...
2	50-59	ge40	15-19	0-2	no	1	right	central	no	no-rec...
3	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurre...
4	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-rec...
5	40-49	premeno	30-34	3-5	yes	2	left	right_up	no	recurre...
6	50-59	premeno	25-29	3-5	no	2	right	left_up	yes	no-rec...
7	50-59	ge40	40-44	0-2	no	3	left	left_up	no	no-rec...
8	40-49	premeno	10-14	0-2	no	2	left	left_up	no	no-rec...
9	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-rec...
10	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-rec...
11	50-59	premeno	25-29	0-2	no	2	left	left_low	no	no-rec...
12	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-rec...
13	50-59	ge40	30-34	0-2	no	1	right	central	no	no-rec...
14	50-59	ge40	25-29	0-2	no	2	right	left_up	no	no-rec...
15	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurre...
16	30-39	premeno	20-24	0-2	no	3	left	central	no	no-rec...
17	50-59	premeno	10-14	3-5	no	1	right	left_up	no	no-rec...
18	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-rec...
19	50-59	premeno	40-44	0-2	no	2	left	left_up	no	no-rec...
20	50-59	ge40	20-24	0-2	no	3	left	left_up	no	no-rec...
21	50-59	lt40	20-24	0-2		1	left	left_low	no	recurre...
22	60-69	ge40	40-44	3-5	no	2	right	left_up	yes	no-rec...
23	50-59	ge40	15-19	0-2	no	2	right	left_low	no	no-rec...
24	40-49	premeno	10-14	0-2	no	1	right	left_up	no	no-rec...

We see that the **highlighted Class attribute** tells us that it is used as the label.

Absolutely yes, we **can change label** by following these steps:

- **Step 1:** Click on the "Edit" button in the top-left corner of the screen.



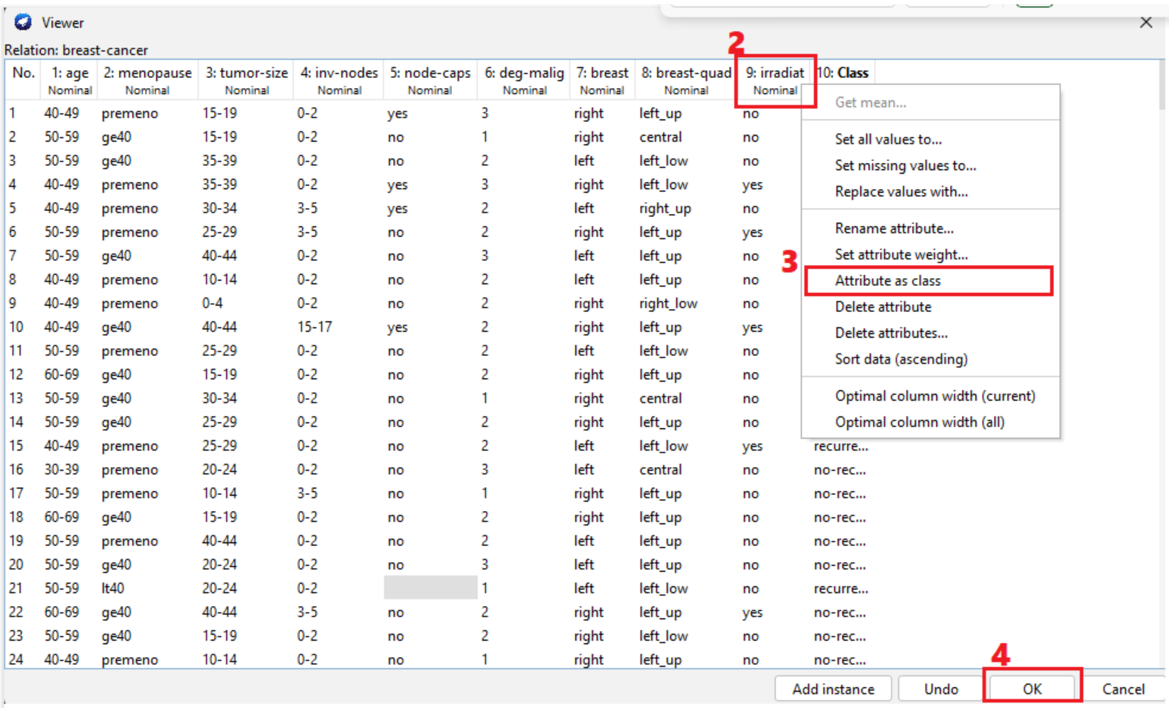
Filter: None

Current relation
Relation: breast-cancer
Instances: 286
Attributes: 10
Sum of weights: 286

Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 6
Type: Nominal
Unique: 1 (0%)

No.	Label	Count	Weight
1	10-19	0	0
2	20-29	1	1

- **Step 2:** Select the attribute that want to use as the label by clicking on its name in the **Attributes** section. And right-click to attribute that we want.
- **Step 3:** Click on the **"Nominal**" button to change the attribute type to nominal if it is not already a nominal attribute. Click on the "**Set as class**" button to set the selected attribute as the label attribute.
- **Step 4:** Save your modified dataset using the "**OK**" button.



Relation: breast-cancer

No.	1: age Nominal	2: menopause Nominal	3: tumor-size Nominal	4: inv-nodes Nominal	5: node-caps Nominal	6: deg-malig Nominal	7: breast Nominal	8: breast-quad Nominal	9: irradiat Nominal	10: Class
1	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	
2	50-59	ge40	15-19	0-2	no	1	right	central	no	
3	50-59	ge40	35-39	0-2	no	2	left	left_low	no	
4	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	
5	40-49	premeno	30-34	3-5	yes	2	left	right_up	no	
6	50-59	premeno	25-29	3-5	no	2	right	left_up	yes	
7	50-59	ge40	40-44	0-2	no	3	left	left_up	no	
8	40-49	premeno	10-14	0-2	no	2	left	left_up	no	
9	40-49	premeno	0-4	0-2	no	2	right	right_low	no	
10	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	
11	50-59	premeno	25-29	0-2	no	2	left	left_low	no	
12	60-69	ge40	15-19	0-2	no	2	right	left_up	no	
13	50-59	ge40	30-34	0-2	no	1	right	central	no	
14	50-59	ge40	25-29	0-2	no	2	right	left_up	no	
15	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurre...
16	30-39	premeno	20-24	0-2	no	3	left	central	no	no-rec...
17	50-59	premeno	10-14	3-5	no	1	right	left_up	no	no-rec...
18	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-rec...
19	50-59	premeno	40-44	0-2	no	2	left	left_up	no	no-rec...
20	50-59	ge40	20-24	0-2	no	3	left	left_up	no	no-rec...
21	50-59	lt40	20-24	0-2		1	left	left_low	no	recurre...
22	60-69	ge40	40-44	3-5	no	2	right	left_up	yes	no-rec...
23	50-59	ge40	15-19	0-2	no	2	right	left_low	no	no-rec...
24	40-49	premeno	10-14	0-2	no	1	right	left_up	no	no-rec...

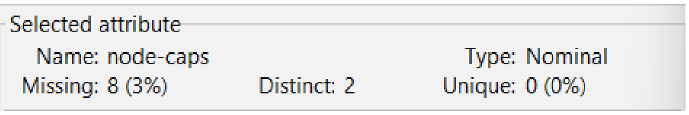
c. <u>What is the meaning of each attribute?</u>

- **Class** - the target variable or label that indicates whether the tumor is benign or malignant.

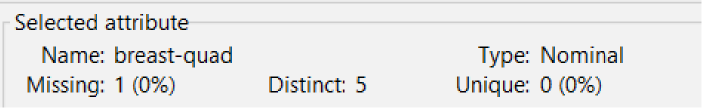
- **Age** - the age of the patient at the time of diagnosis.
- **Menopause** - whether the patient has gone through menopause or not.
- **Tumor size** - the size of the tumor in millimeters.
- **Inv-nodes** - the number of axillary lymph nodes that contain metastatic cancer cells.
- **Node-caps** - whether the cancer cells have spread to the lymph node capsule or not.
- **Deg-malig** - the degree of malignancy of the tumor on a scale of 1 to 3.
- **Breast** - which breast the tumor was found in (left or right).
- **Breast-quad** - the quadrant of the breast in which the tumor was found.
- **Irradiat** - whether the patient received radiation therapy after surgery or not.

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

- There are 2 attributes have missing values: node-caps attribute(3%), breast-quad (1 instance)



Selected attribute
Name: node-caps Type: Nominal
Missing: 8 (3%) Distinct: 2 Unique: 0 (0%)



Selected attribute
Name: breast-quad Type: Nominal
Missing: 1 (0%) Distinct: 5 Unique: 0 (0%)

- To solve the problem of missing values in a dataset, there are several general approaches that you can take using Weka:

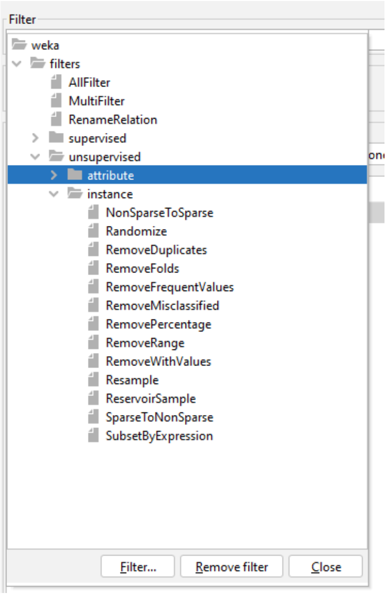
1. **Delete the instances or attributes that contain missing values**. This approach is simple but can result in a loss of information if there are many missing values in the dataset.
2. **Replace the missing values with a fixed value**, such as the mean, median, or mode of the attribute values. This approach can work well if the missing values are random and not correlated with the target variable, but can introduce bias if the missing values are related to other variables in the dataset.
3. **Use an imputation algorithm to estimate the missing values** based on the values of other attributes in the dataset. Weka provides several built-in imputation algorithms, such as k-nearest neighbors imputation and expectation-maximization imputation. These algorithms can be useful for

handling missing values when there is a complex relationship between the missing values and other variables in the dataset.

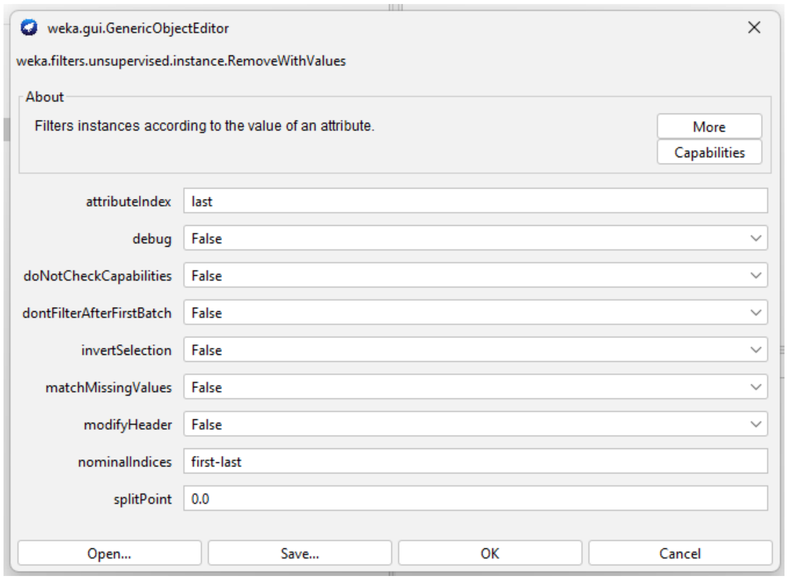
e. <u>Let’s propose solutions to the problem of missing values in the specific attribute.</u>

- There is a solution for **missing values in node-caps** attribute.

Solution: Filter → unsupervised → instance → RemoveWithValues



Click **filter button**, the screen shows:



weka.filters.unsupervised.instance.RemoveWithValues

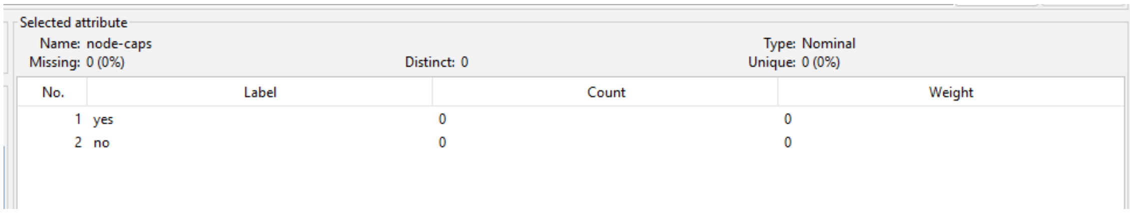
About

Filters instances according to the value of an attribute.

attributeIndex	last
debug	False
doNotCheckCapabilities	False
dontFilterAfterFirstBatch	False
invertSelection	False
matchMissingValues	False
modifyHeader	False
nominalIndices	first-last
splitPoint	0.0

Click **matchMissingValues → True → OK → Apply**

- Result:



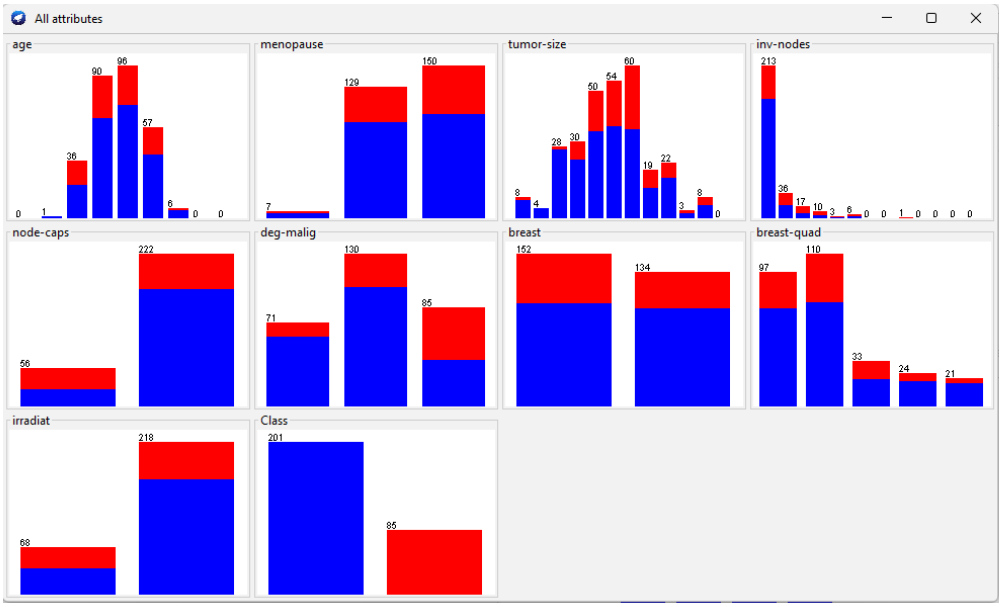
Selected attribute

Name: node-caps Type: Nominal

Missing: 0 (0%) Distinct: 0 Unique: 0 (0%)

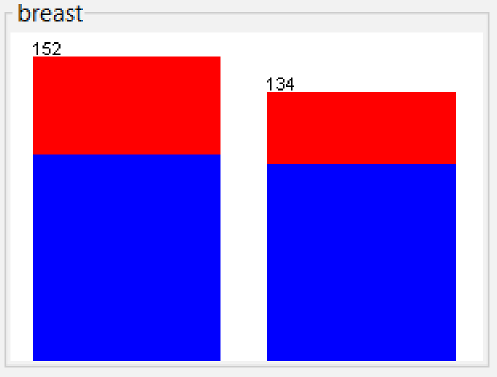
No.	Label	Count	Weight
1	yes	0	0
2	no	0	0

f. <u>Let’s explain the meaning of the chart in the WEKA Explorer. Setting the title for it and describing its legend</u>.



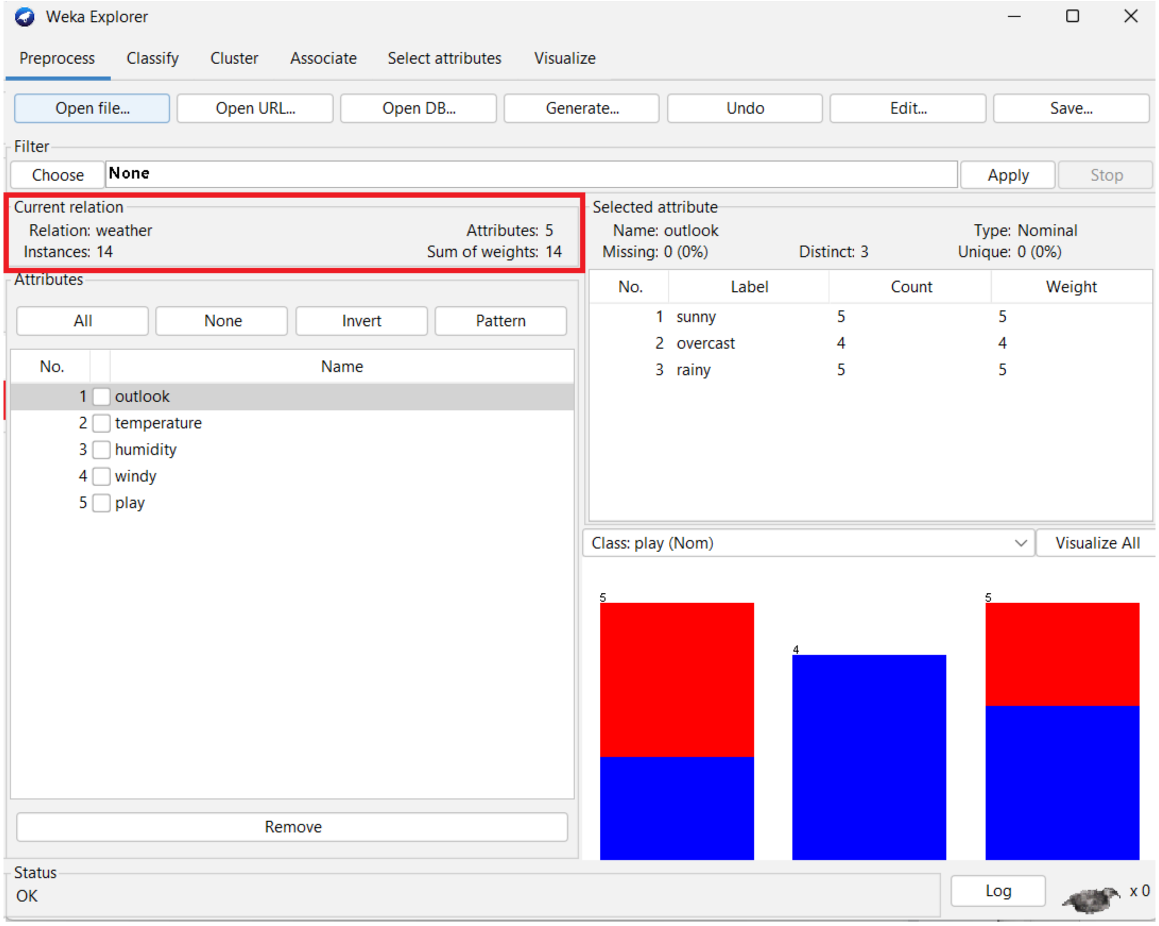
- We can quickly get an overview of the distribution of all attributes in the dataset and the breakdown of distributions by class, red color is **recurrence-events** label, blue is **no-recurrence-events** label.

- *For example*, in **breast** chart. There are 152 instances having left breast and 134 instances having right breast.



❖ Exploring Weather data set: **weather.numeric.arff**

a. 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?



Filter: **None**

Current relation
Relation: weather Attributes: 5
Instances: 14 Sum of weights: 14

Selected attribute
Name: outlook Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5
2	overcast	4	4
3	rainy	5	5

Attributes

No.		Name
1	□	outlook
2	□	temperature
3	□	humidity
4	□	windy
5	□	play

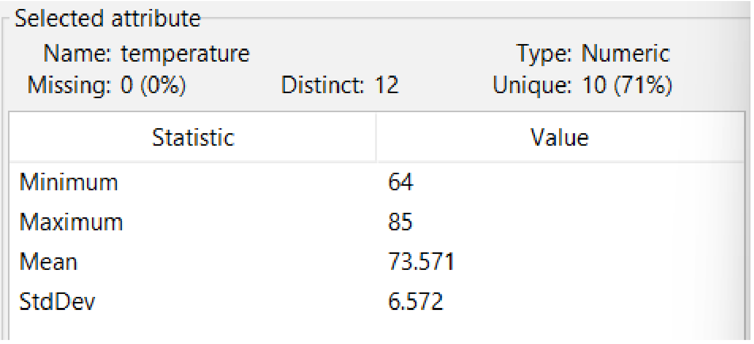
Class: play (Nom)

Status
OK

- Data set has **14 samples** with **5 attributes, 3 categorical attributes** (outlook, windy, play) and **2 numeric attributes** (temperature, humidity).
- **Play** attribute is the **label** with 2 values: yes, no.

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

- Five-number summary of numeric attribute are: min, max, mean, StdDev and median.
- WEKA just provide four values from **Selected attribute** in **Preprocess** tab.



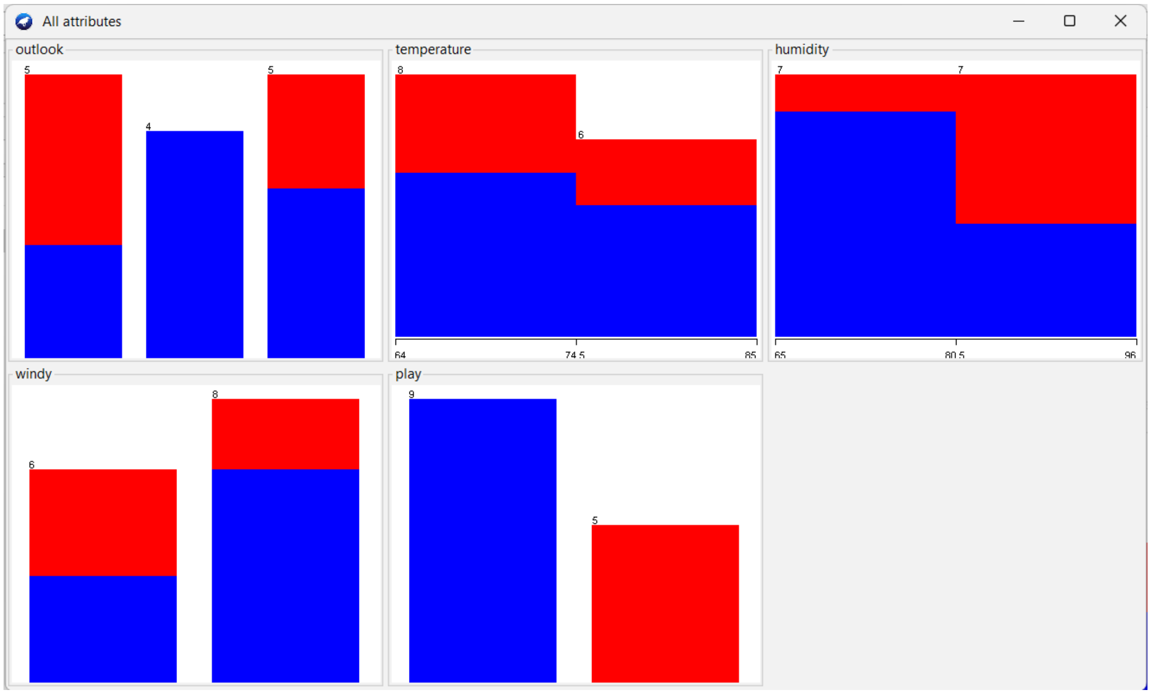
Selected attribute

Name: temperature	Type: Numeric

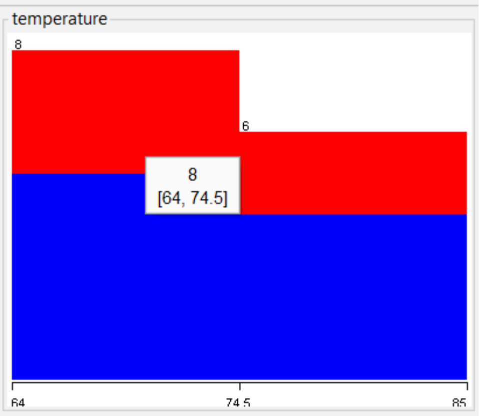
Missing: 0 (0%)	Distinct: 12	Unique: 10 (71%)

Statistic	Value
Minimum	64
Maximum	85
Mean	73.571
StdDev	6.572

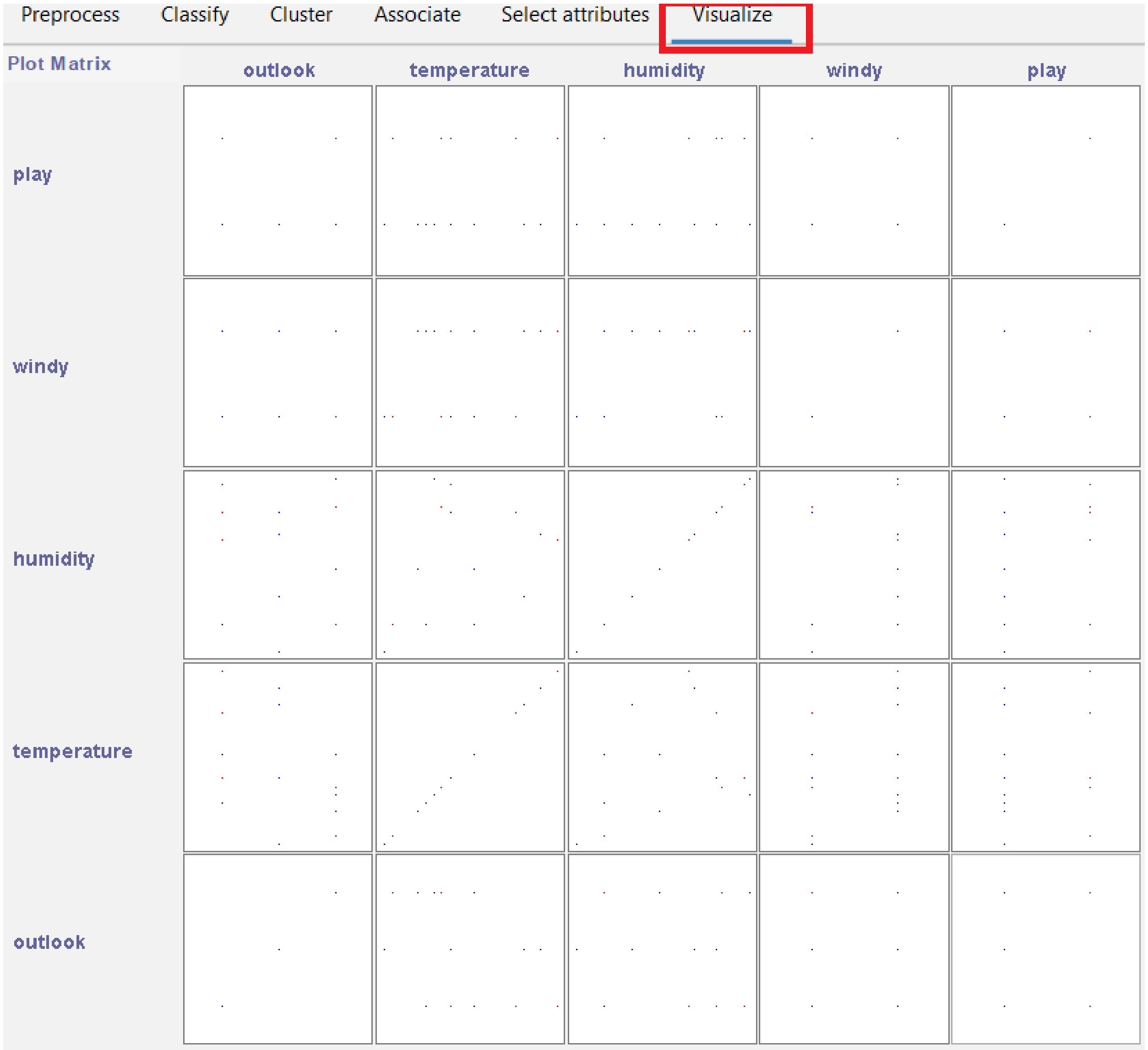
c. <u>Let's explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.</u>



- We can quickly get an overview of the distribution of all attributes in the dataset and the breakdown of distributions by class, red color is **no** label, blue is **yes** label.
- *For example*, in **temperature** chart. There are 8 instances having temperatures in range[64, 74.5] and 6 instances having temperature in range[74.5, 85].



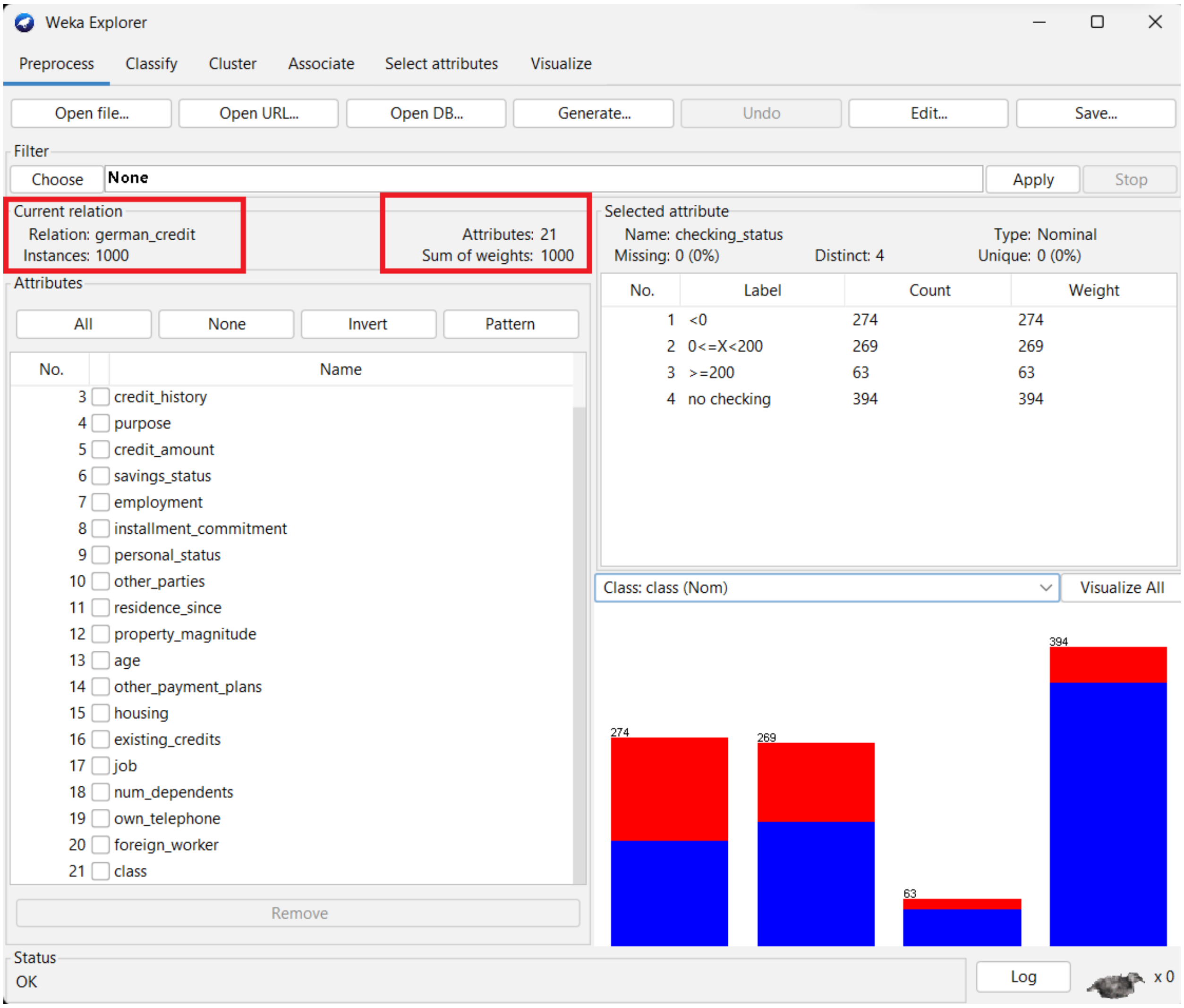
d. In **Visualize** tag. What's the name of this chart? Do you think there are any pairs of different attributes that have correlated?



- This is plot matrix.
- Of course, there are pairs of different attributes that have correlated. And to know more detail how correlative between them, you can click here.

- Exploring Credit in Germany data set: **credit-g.arff**

a. 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).



Filter: None

Current relation
Relation: german_credit
Instances: 1000

Attributes: 21
Sum of weights: 1000

Selected attribute
Name: checking_status
Type: Nominal
Missing: 0 (0%)
Distinct: 4
Unique: 0 (0%)

No.	Label	Count	Weight
1	<0	274	274
2	0<=X<200	269	269
3	>=200	63	63
4	no checking	394	394

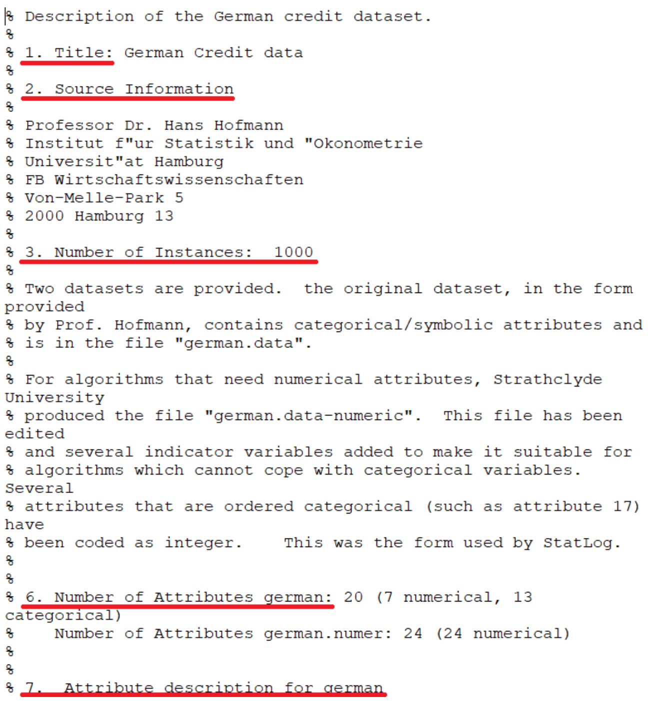
Attributes

No.	Name
3	□ credit_history
4	□ purpose
5	□ credit_amount
6	□ savings_status
7	□ employment
8	□ installment_commitment
9	□ personal_status
10	□ other_parties
11	□ residence_since
12	□ property_magnitude
13	□ age
14	□ other_payment_plans
15	□ housing
16	□ existing_credits
17	□ job
18	□ num_dependents
19	□ own_telephone
20	□ foreign_worker
21	□ class

Class: class (Nom)

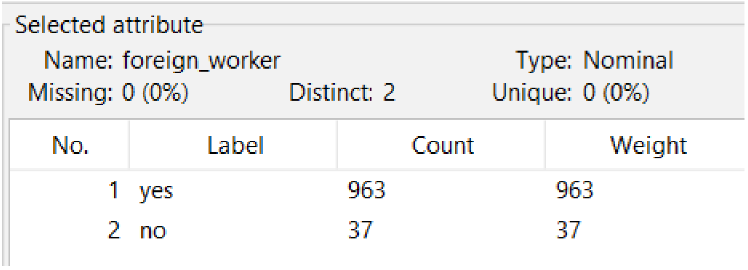
Status
OK

- Content of comments have many sections:
  1. Title
  2. Source Information
  3. Number of instances
  4. Number of attributes german
  5. Number of attributes german.numer
  6. Attribute description for german
  7. Cost matrix



```
% Description of the German credit dataset.
%
% 1. Title: German Credit data
%
% 2. Source Information
%
% Professor Dr. Hans Hofmann
% Institut f"ur Statistik und "Okonometrie
% Universit"at Hamburg
% FB Wirtschaftswissenschaften
% Von-Melle-Park 5
% 2000 Hamburg 13
%
% 3. Number of Instances:  1000
%
% Two datasets are provided.  the original dataset, in the form
provided
% by Prof. Hofmann, contains categorical/symbolic attributes and
% is in the file "german.data".
%
% For algorithms that need numerical attributes, Strathclyde
University
% produced the file "german.data-numeric".  This file has been
edited
% and several indicator variables added to make it suitable for
% algorithms which cannot cope with categorical variables.
Several
% attributes that are ordered categorical (such as attribute 17)
have
% been coded as integer.    This was the form used by StatLog.
%
%
% 6. Number of Attributes german: 20 (7 numerical, 13
categorical)
%     Number of Attributes german.numer: 24 (24 numerical)
%
%
% 7.  Attribute description for german
```

- Data set has **1000 samples** with **21 attributes.**
- Describe 5 attributes:
  - **Foreign_worker:** Binary attribute with 2 values yes or no. Means whether worker is foreign or not. In the dataset, it has 963 yes and 37 no.



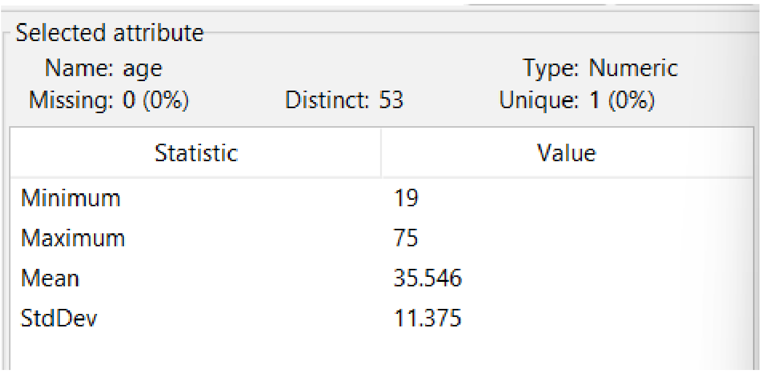
Selected attribute

Name: foreign_worker   Type: Nominal

Missing: 0 (0%)   Distinct: 2   Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	963	963
2	no	37	37

- **Age**: Numeric attribute and has 53 distinct values, 1 unique value and no missing values. Min: 19, Max: 75, Mean: 35.546, StdDev: 11.375.



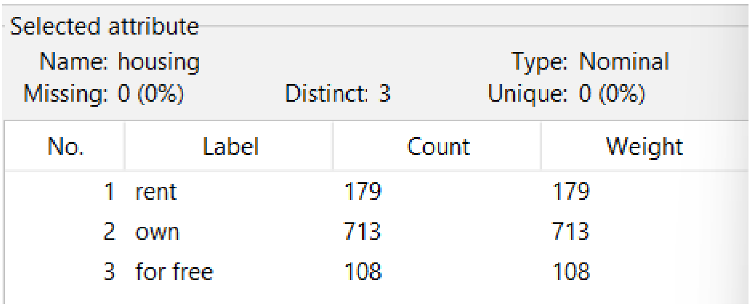
Selected attribute

Name: age Type: Numeric

Missing: 0 (0%) Distinct: 53 Unique: 1 (0%)

Statistic	Value
Minimum	19
Maximum	75
Mean	35.546
StdDev	11.375

- **Housing:** Type of accommodation. This is Nominal attribute with 3 distinct values: rent, own or for free. No missing value. Own housing is majority.



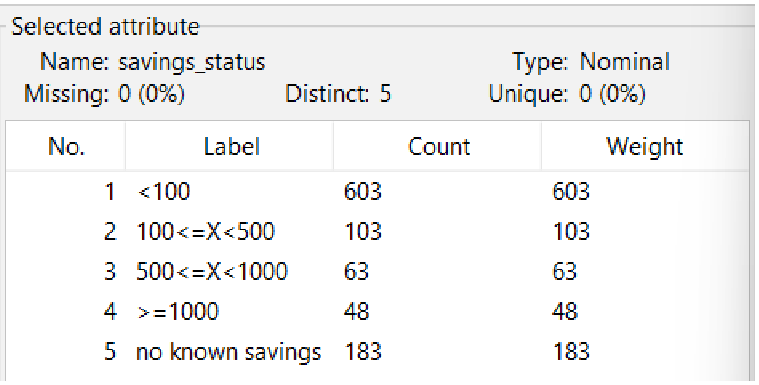
Selected attribute

Name: housing Type: Nominal

Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	rent	179	179
2	own	713	713
3	for free	108	108

- **Savings_status:** status of savings with 5 distinct values, no missing or unique value.
    - Saving < 100: 603 instances.
    - 100 <= saving < 500: 103 instances.



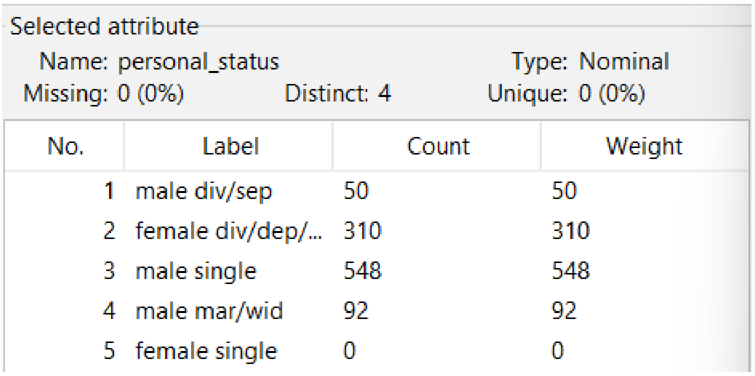
Selected attribute

Name: savings_status Type: Nominal

Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	<100	603	603
2	100<=X<500	103	103
3	500<=X<1000	63	63
4	>=1000	48	48
5	no known savings	183	183

- **Personal_status:** male or female and div/sep or single. It has 4 distinct values and male single is majority.



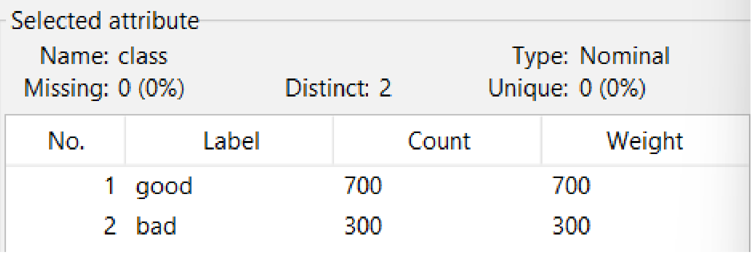
Selected attribute

Name: personal_status Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	male div/sep	50	50
2	female div/dep/...	310	310
3	male single	548	548
4	male mar/wid	92	92
5	female single	0	0

b. Which attribute is used for the label?

- **Class is the label** with 2 distinct values: good (700 instances) and bad (300 instances).



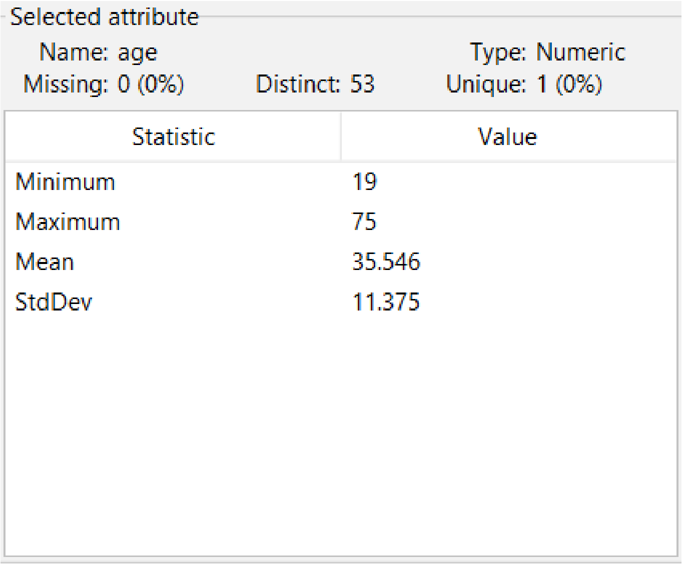
Selected attribute

Name: class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	good	700	700
2	bad	300	300

c. Distribution of continuous attributes?(Left skewed or right skewed?)

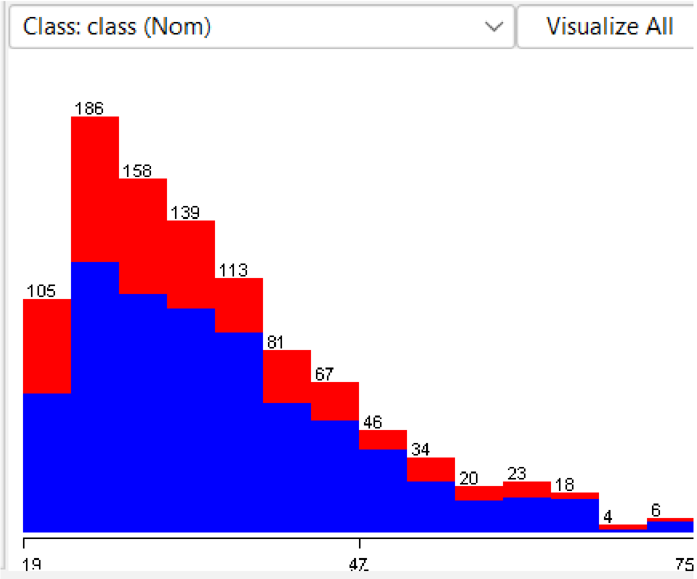
- **age**: This attribute has a **right-skewed distribution**, with a more instances towards lower age (around 18-25) and a longer tail towards the higher age (around 40 - 70).



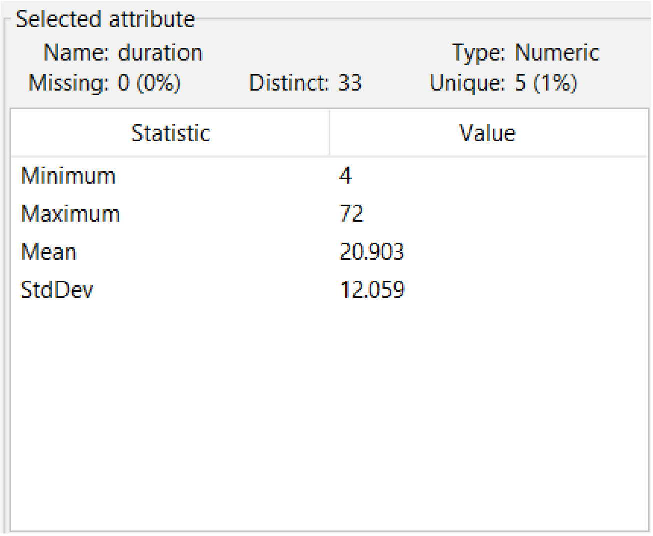
Selected attribute

Name: age Type: Numeric
Missing: 0 (0%) Distinct: 53 Unique: 1 (0%)

Statistic	Value
Minimum	19
Maximum	75
Mean	35.546
StdDev	11.375



- **duration:** This attribute is slightly right-skewed, with most instances having shorter loan durations and a few instances having longer loan durations.

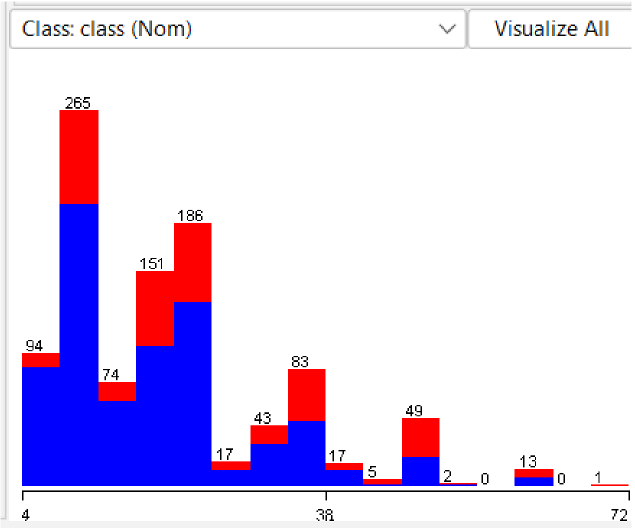


Selected attribute

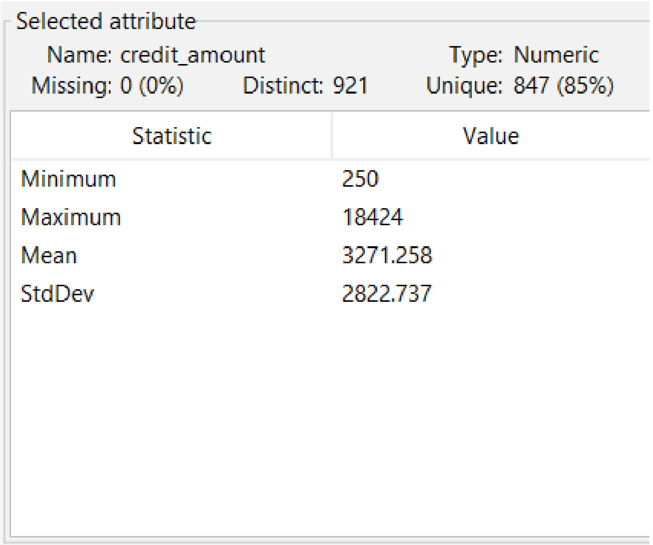
Name: duration Type: Numeric

Missing: 0 (0%) Distinct: 33 Unique: 5 (1%)

Statistic	Value
Minimum	4
Maximum	72
Mean	20.903
StdDev	12.059



- **credit amount:** This attribute is highly right-skewed, with most instances having relatively low credit amounts and a few instances having very high credit amounts.

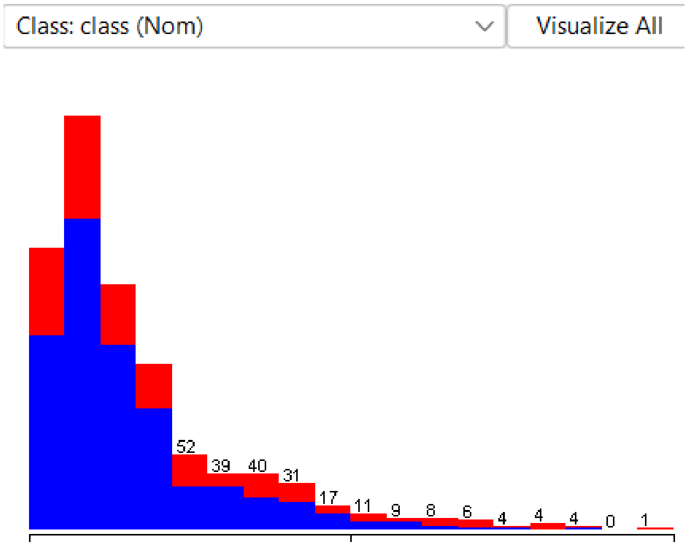


Selected attribute

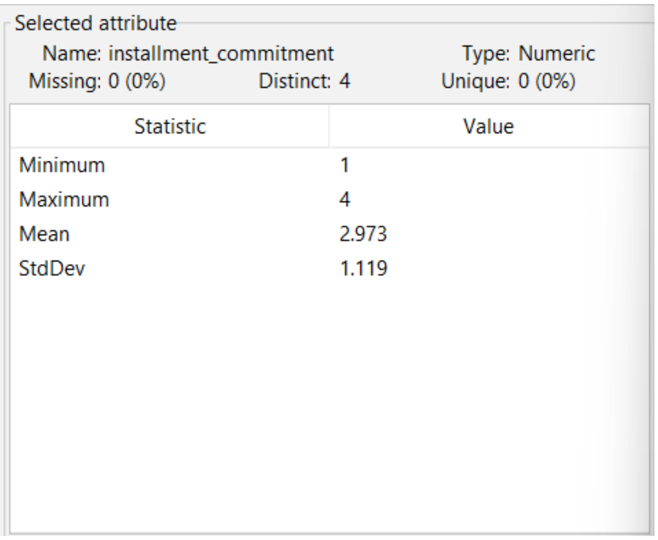
Name: credit_amount Type: Numeric

Missing: 0 (0%) Distinct: 921 Unique: 847 (85%)

Statistic	Value
Minimum	250
Maximum	18424
Mean	3271.258
StdDev	2822.737



- **installment_commitment**: This attribute is left-skewed, with most instances having moderate to high installment commitments.

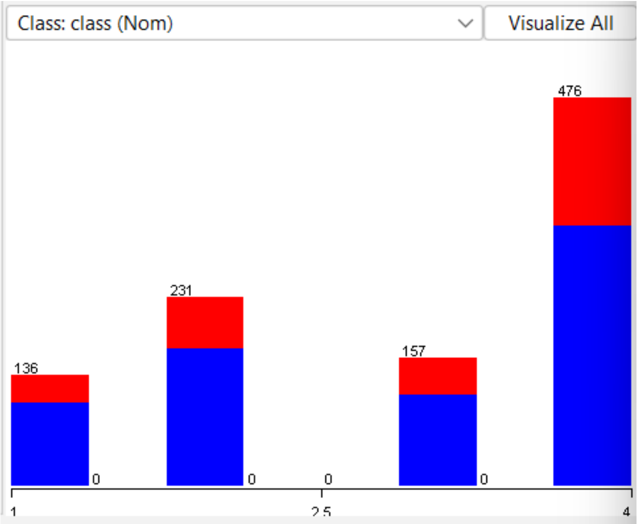


Selected attribute

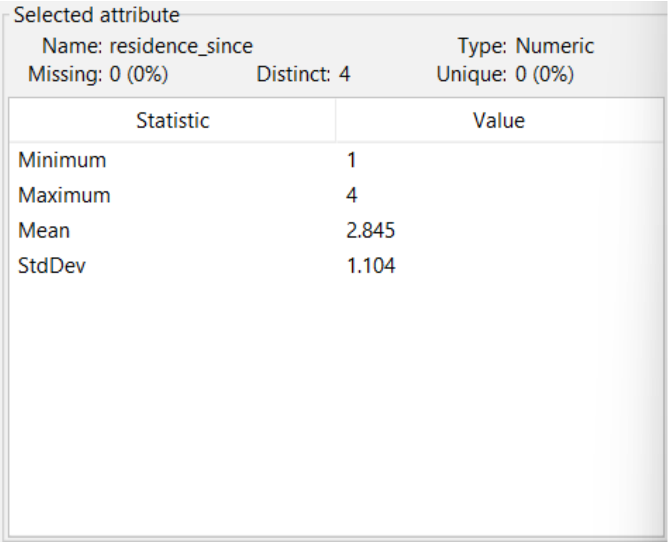
Name: installment_commitment　Type: Numeric

Missing: 0 (0%)　Distinct: 4　Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	4
Mean	2.973
StdDev	1.119



- **resident_since: sightly left-skewed**

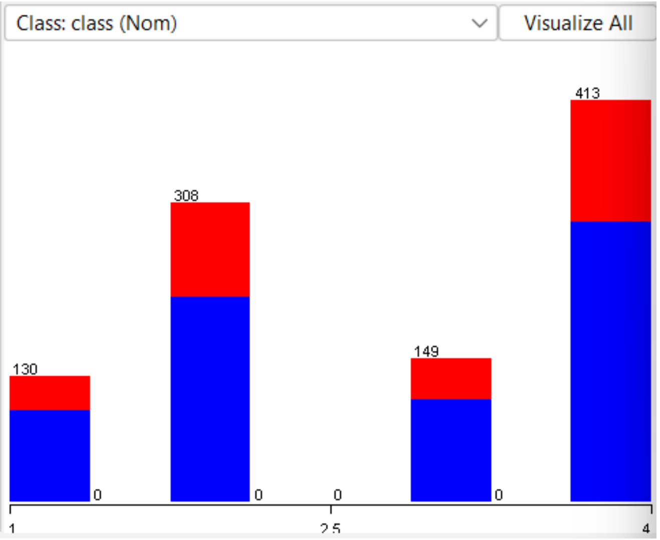


Selected attribute

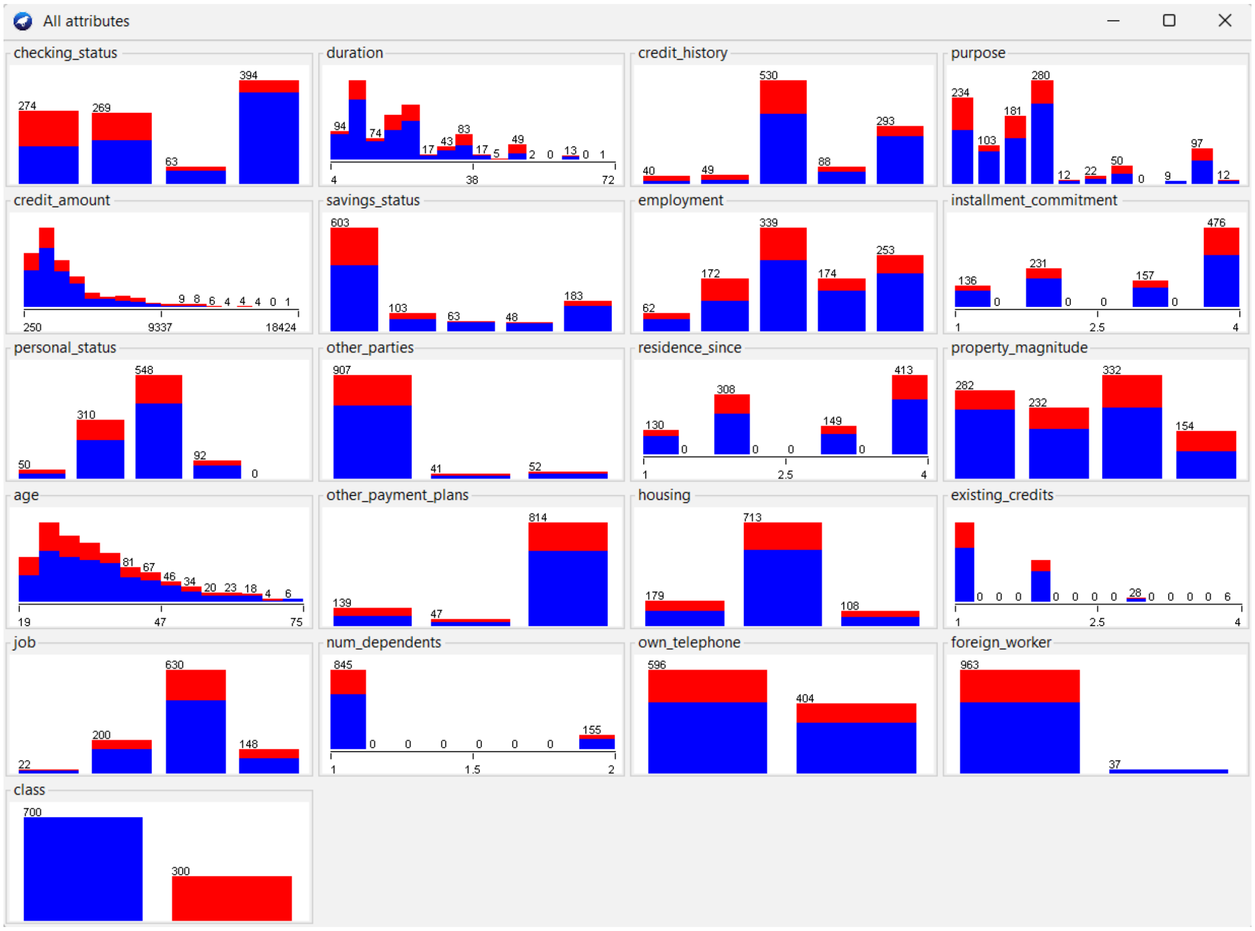
Name: residence_since　Type: Numeric

Missing: 0 (0%)　Distinct: 4　Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	4
Mean	2.845
StdDev	1.104

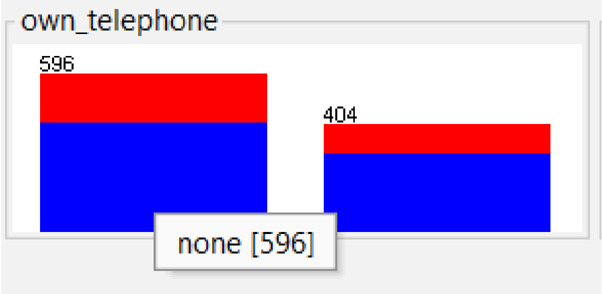


d. <u>Explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.</u>



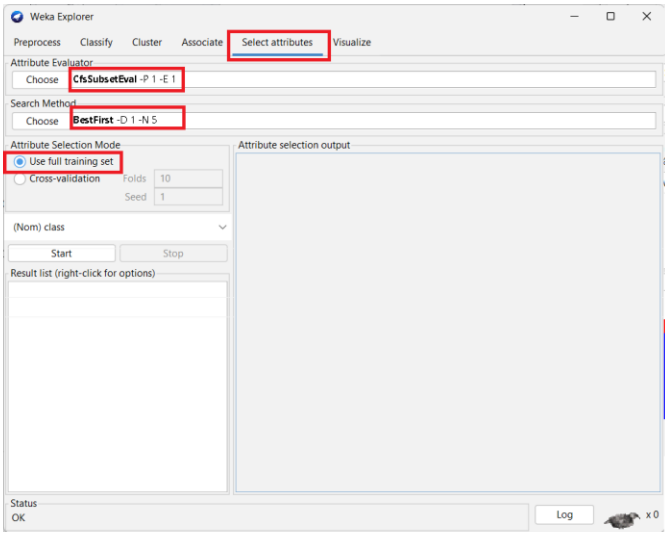
- We can quickly get an overview of the distribution of all attributes in the dataset and the breakdown of distributions by class, red color is **bad** label, blue is **good** label.

- <u>For example,</u> bar chart shows 596 instances have none own_telephone, 404 instances have own_telephone.



e. <u>**Select attributes** tag. Describe all of the options for attribute selection.</u>

- When click on Select Tab, you will see the next screen. It has **Attribute Evaluator, Search Method, Attribute Selection Mode,** you will find several options by click *Choose* button.



Attribute Evaluator: **CfsSubsetEval** -P 1 -E 1

Search Method: **BestFirst** -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) class

Result list (right-click for options)

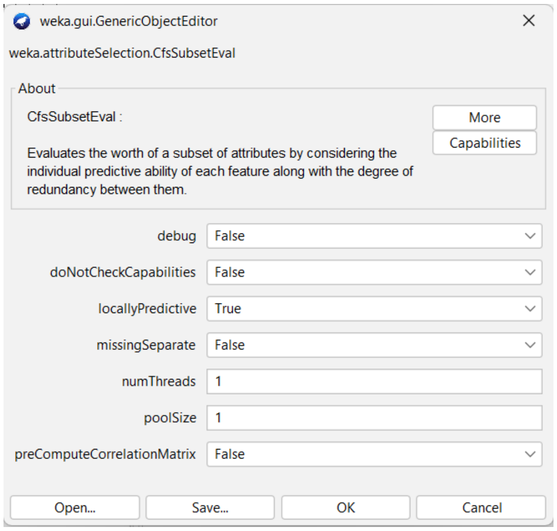
Attribute selection output

Status

OK

- ▪ **Attribute Evaluator.**

When click **Choose**, it shows all options for attribute selection.



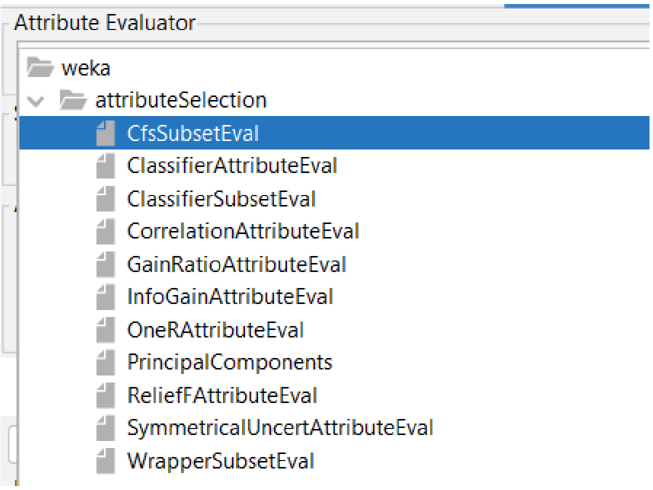
weka.attributeSelection.CfsSubsetEval

CfsSubsetEval :

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.

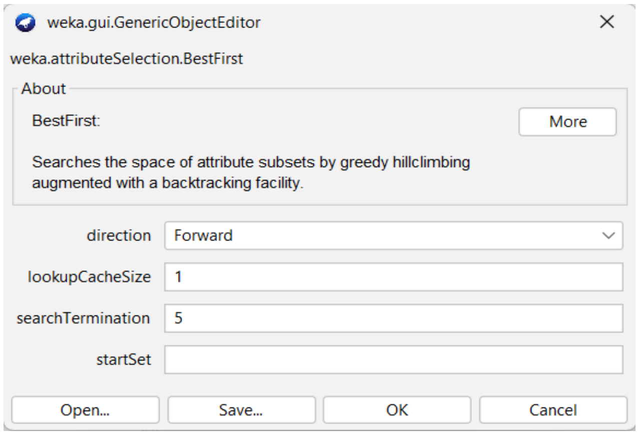
debug	False
doNotCheckCapabilities	False
locallyPredictive	True
missingSeparate	False
numThreads	1
poolSize	1
preComputeCorrelationMatrix	False

If ***left-click in CfsSubsetEval***, you can modify some parameters.



- **Search Method**

When click on **Choose**, it shows all search method.



weka.gui.GenericObjectEditor

weka.attributeSelection.BestFirst

About

BestFirst:

Searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility.

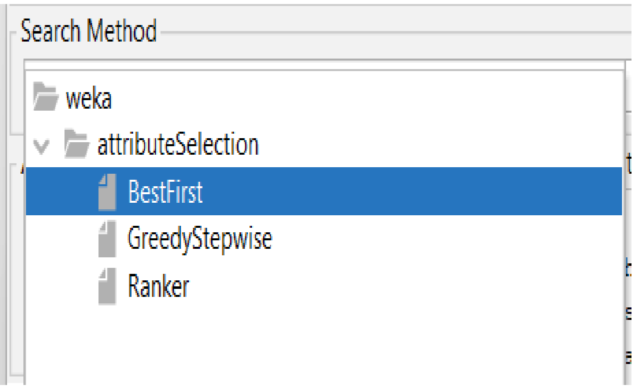
direction: Forward

lookupCacheSize: 1

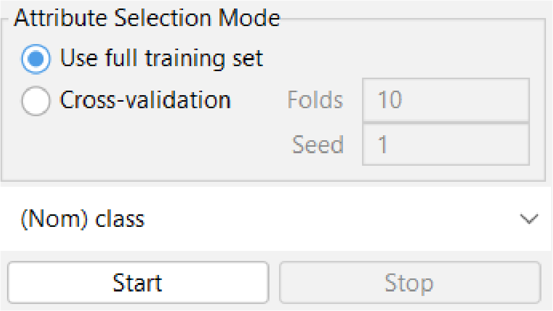
searchTermination: 5

startSet:

left-click on BestFirst to modify parameters.



- **Attribute Selection Mode** has 2 options: use full training set or cross-validation.



Attribute Selection Mode

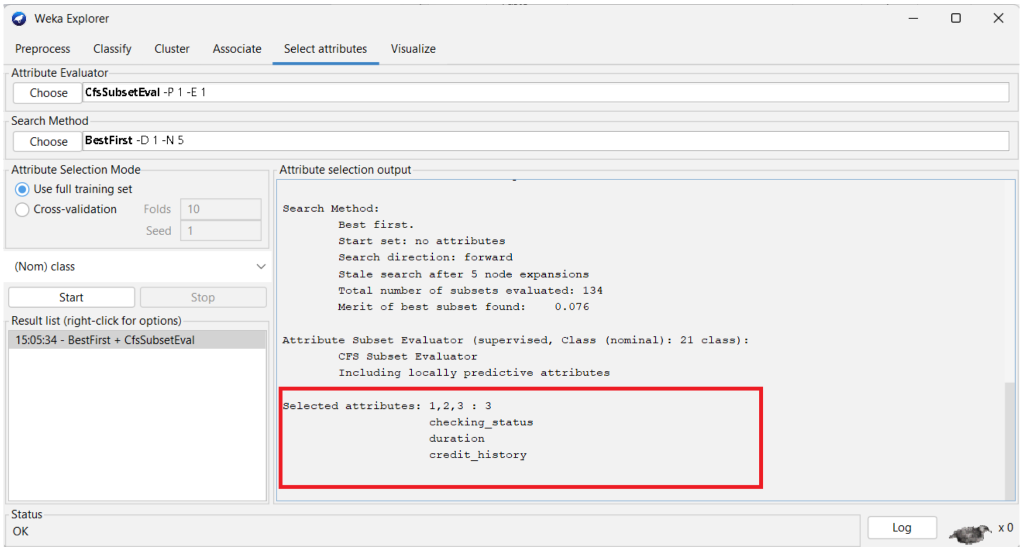
● Use full training set

○ Cross-validation Folds 10

Seed 1

(Nom) class

When click **Start** button to process the dataset. You will see the following output.



Attribute Evaluator: Choose **CfsSubsetEval** -P 1 -E 1

Search Method: Choose **BestFirst** -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) class

Result list (right-click for options)

15:05:34 - BestFirst + CfsSubsetEval

Attribute selection output

```
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 134
        Merit of best subset found:    0.076

Attribute Subset Evaluator (supervised, Class (nominal): 21 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,3 : 3
                     checking_status
                     duration
                     credit_history
```

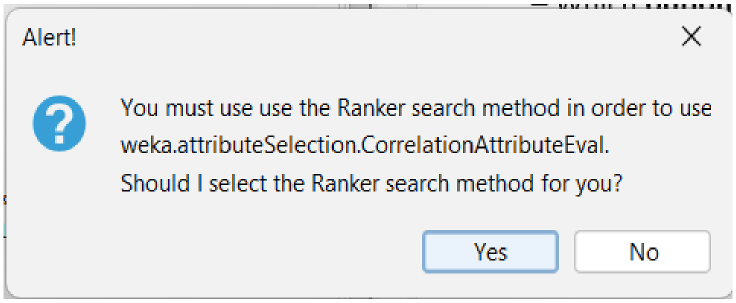
Status

OK

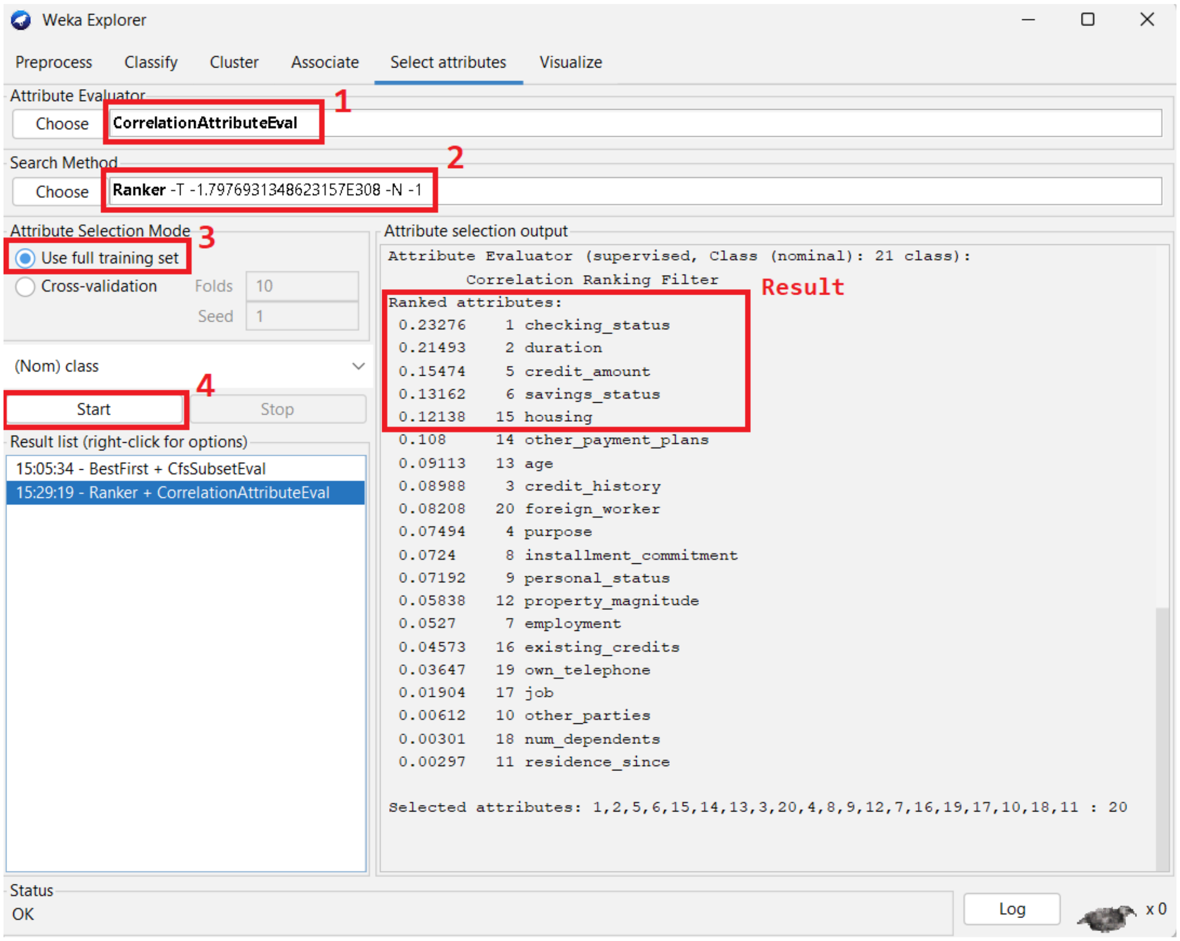
- In the **Attribute selection output** subwindow, you will get result is the list of **Selected** attributes.

f. <u>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).</u>

- **Step 1:** To calculate correlation, we should choose **CorrelationAttributeEval** in attribute evaluator, the Alert will show as below, click **Yes** to select the **Ranker** search method.


Alert!

You must use use the Ranker search method in order to use weka.attributeSelection.CorrelationAttributeEval.
Should I select the Ranker search method for you?



1 Attribute Evaluator: CorrelationAttributeEval

2 Search Method: Ranker -T -1.7976931348623157E308 -N -1

3 Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) class

4 Start

Result list (right-click for options)
15:05:34 - BestFirst + CfsSubsetEval
15:29:19 - Ranker + CorrelationAttributeEval

Attribute selection output

Result

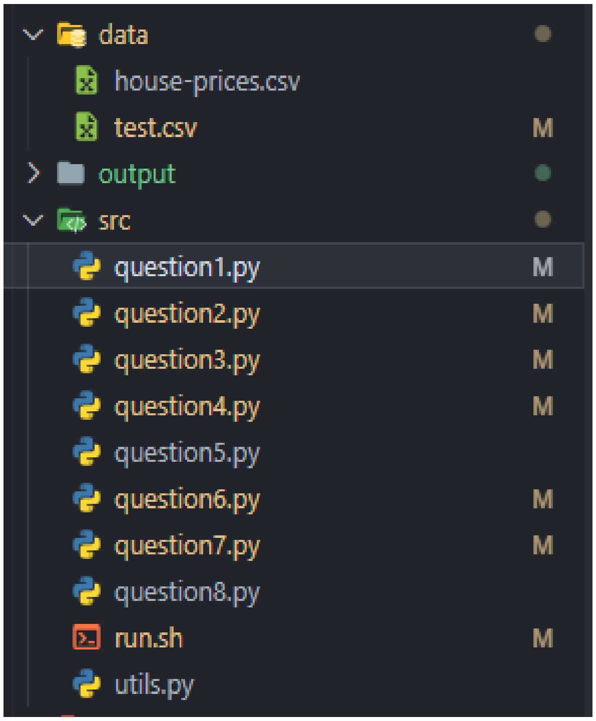
```
Attribute Evaluator (supervised, Class (nominal): 21 class):
        Correlation Ranking Filter
Ranked attributes:
 0.23276    1 checking_status
 0.21493    2 duration
 0.15474    5 credit_amount
 0.13162    6 savings_status
 0.12138   15 housing
 0.108     14 other_payment_plans
 0.09113   13 age
 0.08988    3 credit_history
 0.08208   20 foreign_worker
 0.07494    4 purpose
 0.0724     8 installment_commitment
 0.07192    9 personal_status
 0.05838   12 property_magnitude
 0.0527     7 employment
 0.04573   16 existing_credits
 0.03647   19 own_telephone
 0.01904   17 job
 0.00612   10 other_parties
 0.00301   18 num_dependents
 0.00297   11 residence_since

Selected attributes: 1,2,5,6,15,14,13,3,20,4,8,9,12,7,16,19,17,10,18,11 : 20
```

Status
OK

- **Step 2:** Choose **Ranker** by click Yes on Alert, because it is recommendation if we use CorrelationAttributeEval.
- **Step 3:** Choose **Use full training set** in attribute selection mode.
- **Step 4:** Click on **Start** button to process the data.
- **Step 5:** Get the result from Attribute Selection Output as the figure. So five **attributes with highest correlation** are **checking_status, duration, credit_amount, savings_status, housing.**

c. Preprocessing Data in Python (5 points)

Directory structure:



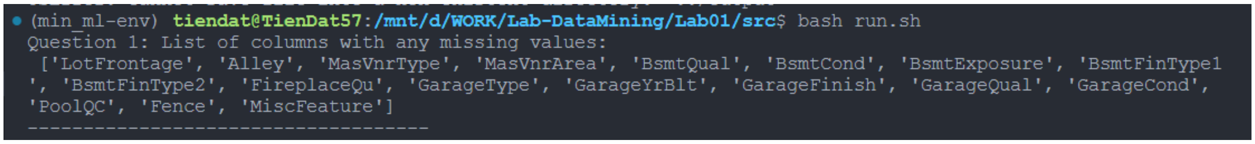
```
data
    house-prices.csv
    test.csv          M
output
src
    question1.py      M
    question2.py      M
    question3.py      M
    question4.py      M
    question5.py
    question6.py      M
    question7.py      M
    question8.py
    run.sh            M
    utils.py
```

- We have a data folder (containing input and output data for programs) at the same level as the src folder (containing program files). In there:
- **data folder**:
  - `house-prices.csv:` file input after the program is ready.
  - `test.csv`: file using to test during programming.
- **src folder:**
  - The files corresponding to the question (`question1.py`, ...)
  - `utils.py`: This file contains functions that are reused many times.
  - `run.sh`: file contains scripts when we execute it will execute the commands contained in the file (Here this file will run all the questions in one go with bash run.sh command).

1. Extract columns with missing values:

- **Describe**: The requirement for extracting columns with missing values is to identify which columns have missing data and determine the appropriate way to handle them.
- **Input**: Read file **house-prices.csv**
- **Run:** `python3 question1.py house-prices.csv`

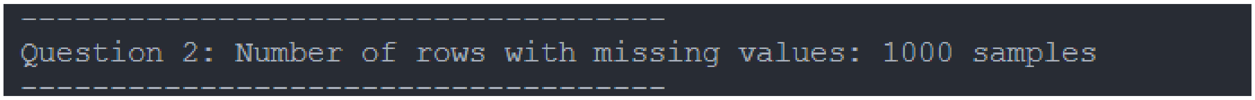
- **Output:** List of columns with any missing values



```
● (min_ml-env) tiendat@TienDat57:/mnt/d/WORK/Lab-DataMining/Lab01/src$ bash run.sh
 Question 1: List of columns with any missing values:
  ['LotFrontage', 'Alley', 'MasVnrType', 'MasVnrArea', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1
 ', 'BsmtFinType2', 'FireplaceQu', 'GarageType', 'GarageYrBlt', 'GarageFinish', 'GarageQual', 'GarageCond',
 'PoolQC', 'Fence', 'MiscFeature']
 ------------------------------------
```

2. **Count the number of lines with missing data:**

- **Describe:** Counting the number of lines with missing data involves scanning each row in a data table and identifying which rows have one or more missing values. We can do this by iterating over each row in the data table and checking for missing values, which can be represented in different ways. The count of lines with missing data can be used as a quality metric for data analysis, and can inform decisions on data cleaning, imputation, or removal.
- **Input:** Read file **house-prices.csv**
- **Run:** `python3 question2.py house-prices.csv`
- **Output:** Number of rows with missing values



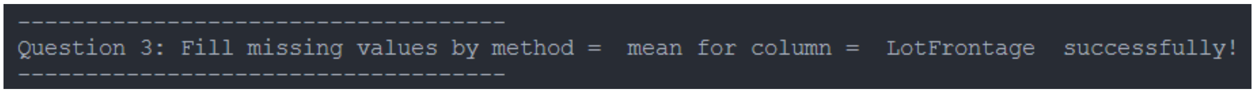
```
------------------------------------
Question 2: Number of rows with missing values: 1000 samples
------------------------------------
```

3. **Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute).**

- **Describe:** Filling in missing values involves replacing missing values with estimated values based on the available data.
- **Input:** Read file **house-prices.csv**
- **Run:**

```
python3 question3.py house-prices.csv --method=mean --columns=LotFrontage --out=fill_nan_values.csv
```

- **Output: After full nan values → data save to csv** `data/fill_nan_values.csv`



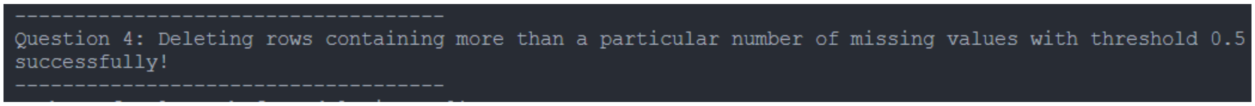
```
------------------------------------
Question 3: Fill missing values by method =  mean for column =  LotFrontage  successfully!
------------------------------------
```

4. **Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes)**

- **Input:** Read file **house-prices.csv**
- **Run:**

```
python3 question4.py house-prices.csv --threshold=0.5 --out=del_rows.csv
```

- **Output:**

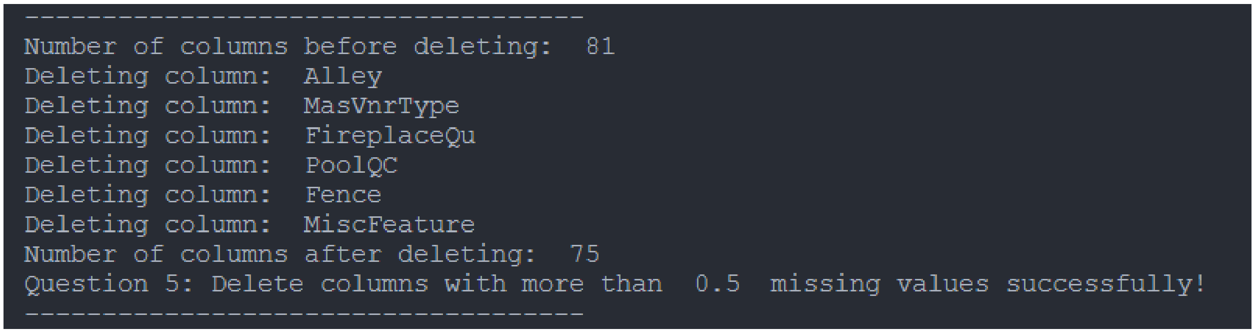


```
-----------------------------------
Question 4: Deleting rows containing more than a particular number of missing values with threshold 0.5
successfully!
-----------------------------------
```

5. **Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples).**
   - **Input:** Read file **house-prices.csv**
   - **Run:**

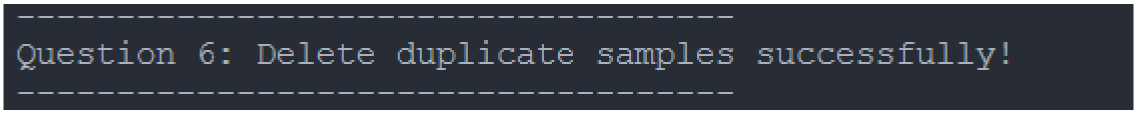
```
python3 question5.py house-prices.csv --threshold=0.5 --out=del_cols.csv
```

- **Output:**



```
------------------------------------
Number of columns before deleting:  81
Deleting column:  Alley
Deleting column:  MasVnrType
Deleting column:  FireplaceQu
Deleting column:  PoolQC
Deleting column:  Fence
Deleting column:  MiscFeature
Number of columns after deleting:  75
Question 5: Delete columns with more than  0.5  missing values successfully!
------------------------------------
```

6. **Delete duplicate samples**
   - **Input:** Read file **house-prices.csv**
   - **Run:** `python3 question6.py house-prices.csv --out=del_dup.csv`
   - **Output:**

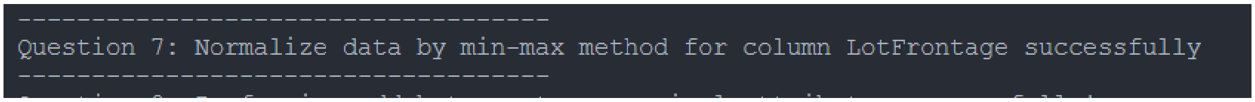


```
------------------------------------
Question 6: Delete duplicate samples successfully!
------------------------------------
```

7. **Normalize a numeric attribute using min-max and Z-score methods**
   - **Input:** Read file **house-prices.csv**
   - **Run:**

```
python3 question7.py house-prices.csv --method=min-max --columns=LotFrontage --out=normalize.csv
```

- **Output:**



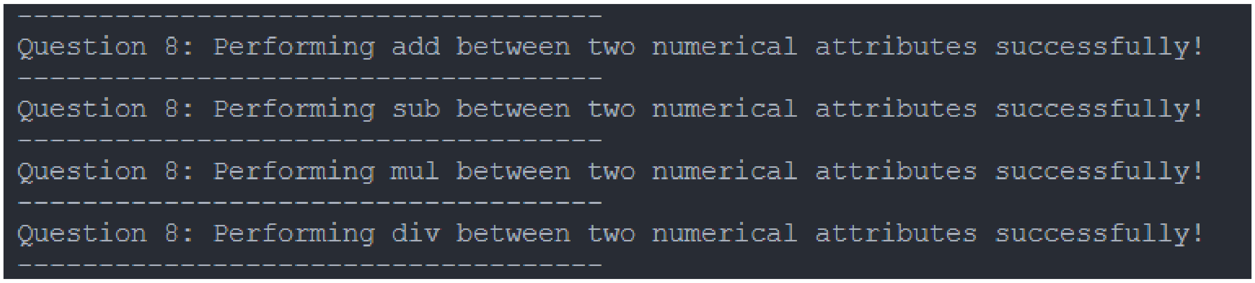
```
------------------------------------
Question 7: Normalize data by min-max method for column LotFrontage successfully
------------------------------------
```

8. **Performing addition, subtraction, multiplication, and division between two numerical attributes.**

- **Input:** Read file **house-prices.csv**
- **Run:**

```
python3 question8.py house-prices.csv --method=add --columns=LotFrontage,1stFlrSF --out=add.csv

python3 question8.py house-prices.csv --method=sub --columns=LotFrontage,1stFlrSF --out=sub.csv

python3 question8.py house-prices.csv --method=mul --columns=LotFrontage,1stFlrSF --out=mul.csv

python3 question8.py house-prices.csv --method=div --columns=LotFrontage,1stFlrSF --out=div.csv
```

- **Output:**



```
------------------------------------
Question 8: Performing add between two numerical attributes successfully!
------------------------------------
Question 8: Performing sub between two numerical attributes successfully!
------------------------------------
Question 8: Performing mul between two numerical attributes successfully!
------------------------------------
Question 8: Performing div between two numerical attributes successfully!
------------------------------------
```

III. REFERENCES

[1]. Handle missing values

[2]. weka_feature_selection

[3]. skewed

[4]. Weka_tutorial

[5]. Lecture slides

[6]. passing args

[7]. discrete-vs-continuous-data

[8]. calculate correlation