UNIVERSITY OF SCIENCE, VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY

FACULTY OF INFORMATION TECHNOLOGY

# LAB 01 REPORT

## DATA PREPROCESSING AND DATA EXPLORATION

### COURSE NAME: DATA MINING AND APPLICATION

| Instructors: | Team member: |
|---|---|
| Prof. Le Hoai Bac | 20127028 – Vo Van Hoang |
| Nguyen Thi Thu Hang | 20127054 – Ngo Van Trung Nguyen |

HO CHI MINH CITY, MARCH, 2023

# TABLE OF CONTENT

# I. General information

## 1. Student information

| Name | Class | Student ID | Email |
|---|---|---|---|
| Vo Van Hoang | 20 KHDL | 20127028 | 20127028@student.hcmus.edu.vn |
| Ngo Van Trung Nguyen | 20 KHDL | 20127054 | 20127054@student.hcmus.edu.vn |

## 2. Member contribution rate

| Name | Responsibility | Detail | Completed rate |
|---|---|---|---|
| Hoang | Writing Report | Try to present as clear as possible | 100% |
| | Install WEKA part | Requirement 1 | 100% |
| | | Requirement 2 | 100% |
| | Getting Acquainted With WEKA part | Exploring Breast Cancer data set | 100% |
| | | Exploring Weather data set | 100% |
| | | Exploring Credit in Germany data set | 100% |
| Nguyen | Preprocessing Data in Python part | Extract columns with missing values | 100% |
| | | Count the number of lines with missing data | 100% |
| | | Fill in the missing value using mean, median and mode | 100% |
| | | Deleting rows containing more than a particular number of missing values | 100% |
| | | Deleting columns containing more than a particular number of missing values | 100% |
| | | Delete duplicate samples | 100% |
| | | Normalize a numeric attribute using min-max and Z-score methods | 100% |
| | | Performing addition, subtraction, multiplication, and division | 100% |

**** In general:***

***% Completed project** (100%) = **Hoang's work**(50%) + **Nguyen's work**(50%), so that we share the tasks fairly equally.*

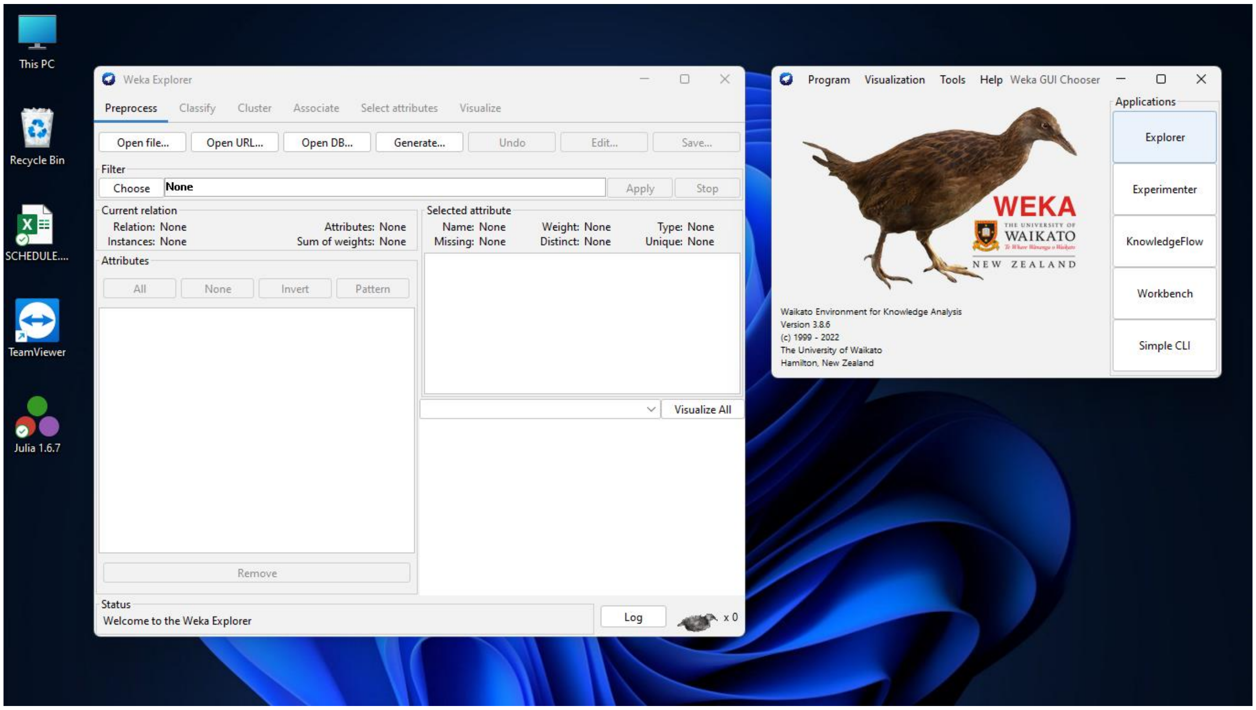
## 3. Questions or requirements that have not been completed

- We completely finished all the tasks on time.

## II. Install WEKA

### 1. Requirement 1

"After installing, you capture a screen that contains the "Explorer" function in your desktop background."



Filter: None

Current relation
Relation: None
Attributes: None
Instances: None
Sum of weights: None

Attributes

Selected attribute
Name: None
Weight: None
Type: None
Missing: None
Distinct: None
Unique: None

Status
Welcome to the Weka Explorer

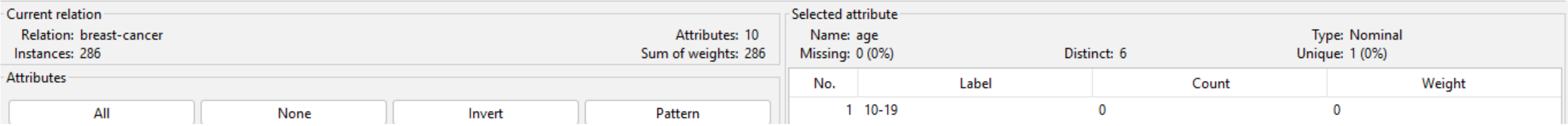
Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

*Picture 1. The display of "Explorer" function of WEKA*

### 2. Requirement 2

"Students open any data set (with extended part .arff). Explain the meaning of Current Relation, Attributes, and Selected Attribute in Preprocess tag. Briefly explain the meaning of the other tags in WEKA Explorer."

→ We open breast cancer.arff dataset, and we can see the picture bellow:



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

Attributes

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

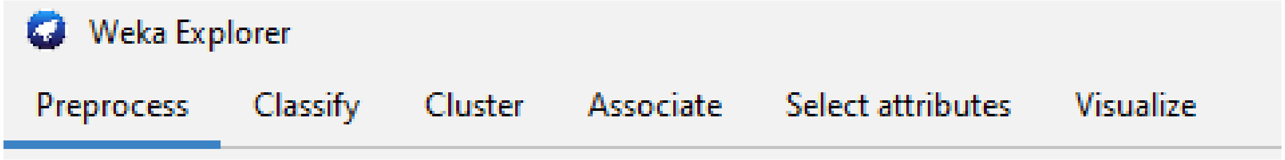
| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | 10-19 | 0 | 0 |

*Picture 2. The picture of PreProcess tag*

a. Explain the meaning of Current Relation, Attributes, and Selected Attribute in Preprocess tag

| Name need to explain | Explanation |
|---|---|
| *Current relation* | Information of the current table, including: name of the table, the number of attributes, sum of weights and the number of samples. |
| *Attributes* | Present the attributes of the table, allowing us to choose the attributes which we need to explore,... |
| *Selected attributes* | Information of the selected attribute in the Attributes group (attribute name, data type, percentage of missing data,...). Besides, it also shows other information about the max, min, average,...of the values in that attribute. |

b. Explain the meaning of the other tags in WEKA Explorer



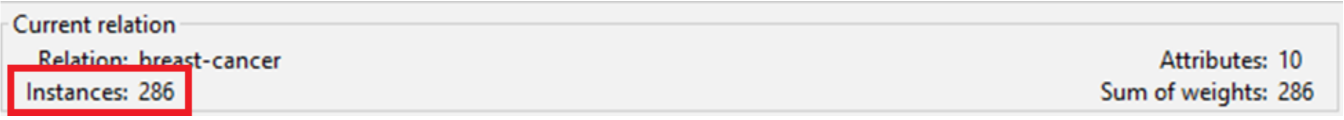
*Picture 3. The picture of the other tags in WEKA Explorer*

| Name of tag | Explanation |
|---|---|
| *Preprocess* | Select and preprocess the data to work with. |
| *Classify* | Data classification |
| *Cluster* | Clustering data |
| *Associate* | Mining association rules of data |
| *Select attributes* | Select relevant and important attributes of the data |
| *Visualize* | Display chart of the data(data visualization) |

# III. Getting Acquainted With WEKA

## 1. Exploring Breast Cancer data set

a. How many instances does this data set have?



Current relation

Relation: breast-cancer

Instances: 286

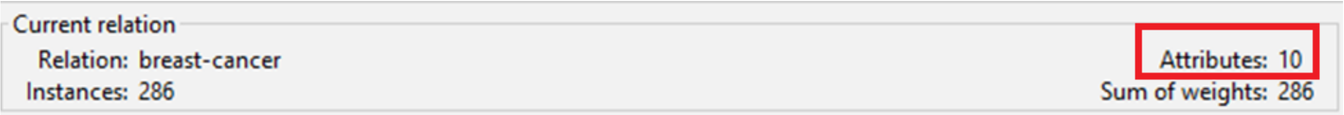
Attributes: 10

Sum of weights: 286

*Picture 4. Information about number of intances*

➔ This data set has 286 intances.

b. How many attributes does this data set have?



Current relation

Relation: breast-cancer

Instances: 286

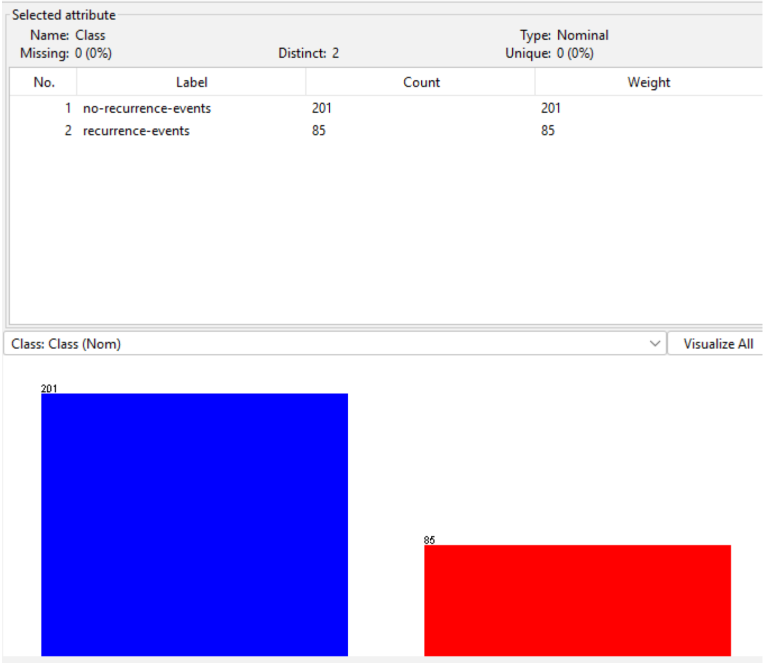
Attributes: 10

Sum of weights: 286

*Picture 5. Information about number of attributes*

➔ This data set has 10 attributes.

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



Selected attribute

Name: Class

Type: Nominal

Missing: 0 (0%)

Distinct: 2

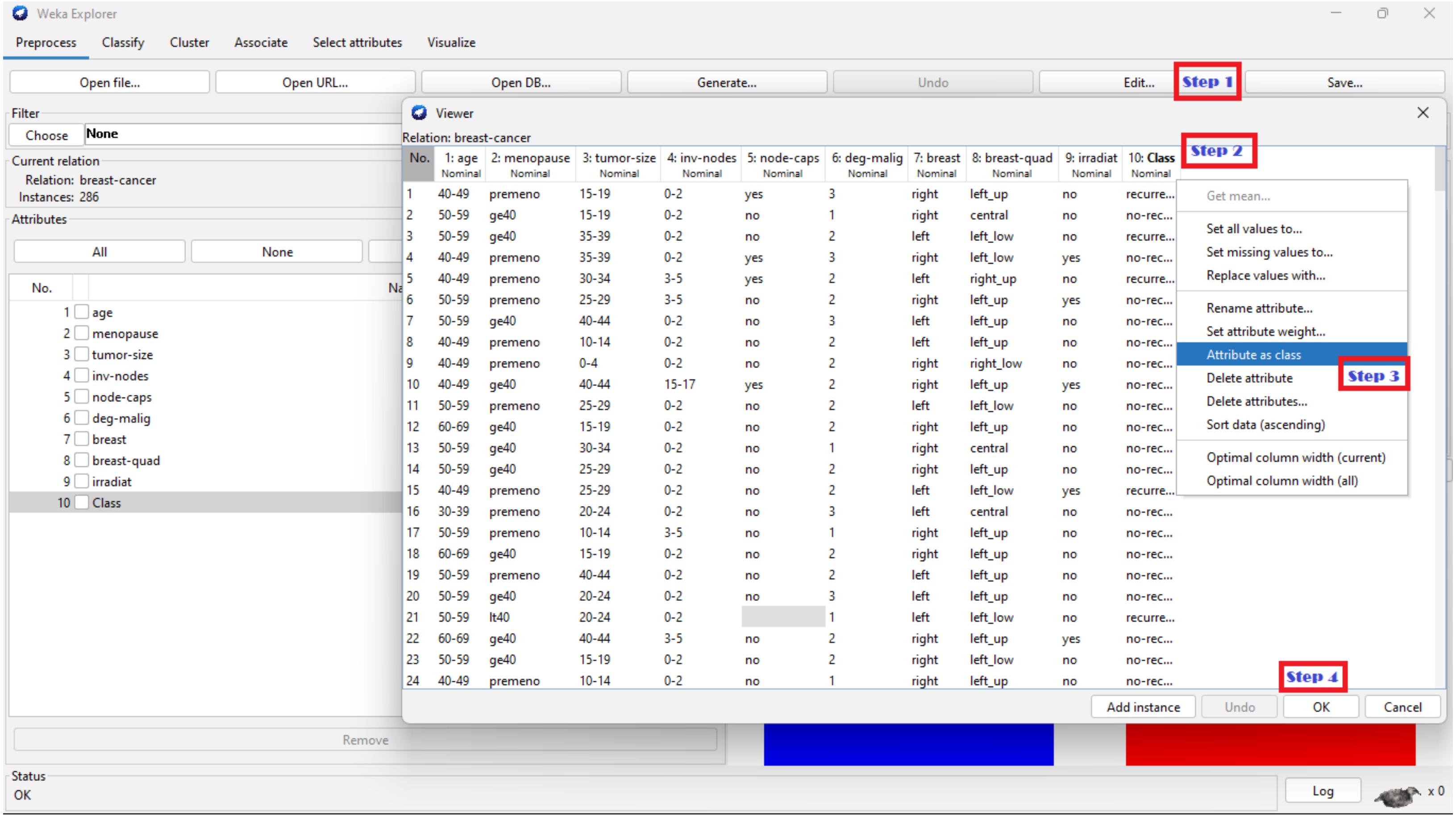
Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | no-recurrence-events | 201 | 201 |
| 2 | recurrence-events | 85 | 85 |

Class: Class (Nom)

*Picture 6. Picture of class attribute*

- ➔ Every data has a class attribute, in this data set, the attribute is used for the label named "Class" which includes 2 values: *no-recurrence-events* and *recurrence-events*.
- ➔ We can change the class attribute by clicking "**Edit**" in tab Explorer, choose a attribute  we want it as an class attribute, after that, click right mouse on this attribute and choose "**Attribute as class**". Finally, clicking "**OK**".



Relation: breast-cancer
Instances: 286

Relation: breast-cancer

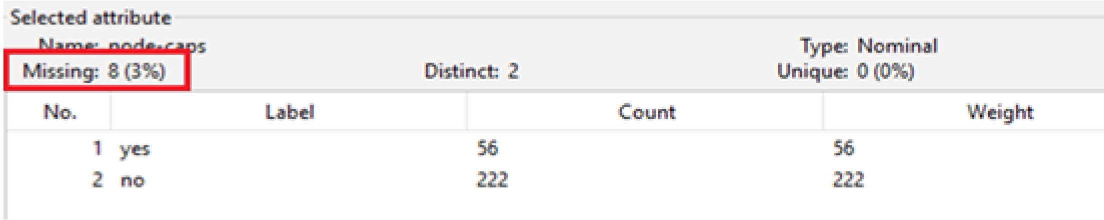
| No. | 1: age<br>Nominal | 2: menopause<br>Nominal | 3: tumor-size<br>Nominal | 4: inv-nodes<br>Nominal | 5: node-caps<br>Nominal | 6: deg-malig<br>Nominal | 7: breast<br>Nominal | 8: breast-quad<br>Nominal | 9: irradiat<br>Nominal | 10: **Class**<br>Nominal |
|---|---|---|---|---|---|---|---|---|---|---|
| 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... |

*Picture 7. Sumarize the method to change class attribute (label attribute)*

d. What is the meaning of each attribute?

| Name of attribute | Meaning |
|---|---|
| age | Patient's age |
| menopause | Indicate the number of patients before menopause and after menopause |
| tumor-size | The size of the tumor |
| inv-nodes | The number of axillary lymph nodes containing metastatic breast cancer that is visible on histological examination |
| node-caps | Indicates whether the tumor can penetrate the capsule and invade the tissues |
| deg-malig | Degree of malignancy |
| breast | Number of breast cancer left, right |
| breast-quad | Parts of the breast |
| irradiat | Possible to have radiation therapy or not |
| Class | No recurrence and recurrence |

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

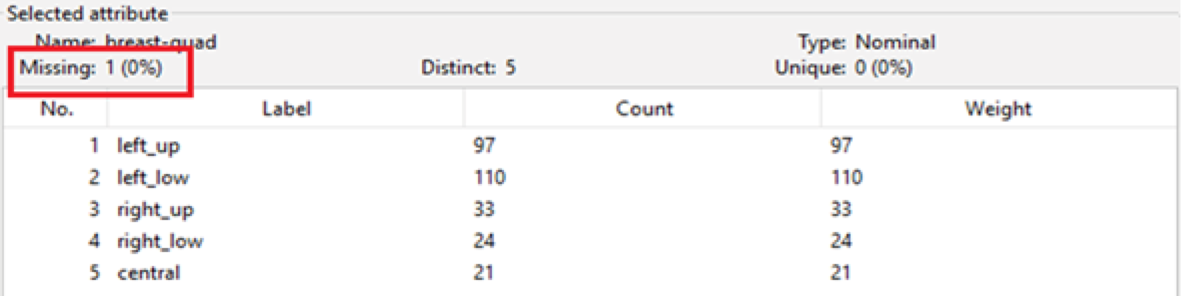


Selected attribute
Name: node-caps
Missing: 8 (3%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | yes | 56 | 56 |
| 2 | no | 222 | 222 |

*Picture 8. The number and rate of missing value of "node-caps" attribute*

→"**node-caps**" attribute has **8** missing values which occupies about **3%** in this data set



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

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | left_up | 97 | 97 |
| 2 | left_low | 110 | 110 |
| 3 | right_up | 33 | 33 |
| 4 | right_low | 24 | 24 |
| 5 | central | 21 | 21 |

*Picture 9. The number and rate of missing value of "breast-quad" attribute*

→"**breast-quad**" attribute has only **1** missing value which occupies approximately **0%** in this data set.

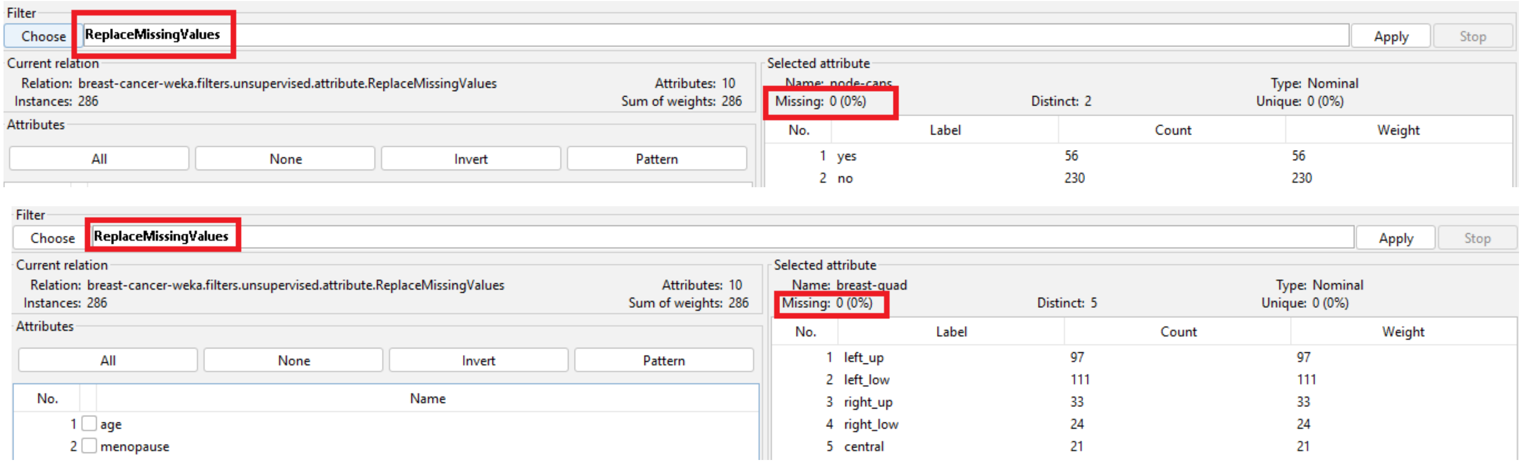
→In general, there are a variety of ways do eliminate missing values, for example:

- When the number of missing values is not many for the data (such as: only 2 missing values out of 1000 rows data) or the missing values are not necessary to the data set, so we can delete this data column/ attributes.
- We can also handle missing values problem by filling the missing values by average value/ median with numeric attribute or mode value by nominal attribute. Furthermore, we can eliminate missing intances, filling NULL(unknown) in missing positions.
- Last but not least, we can predict the most probable value for the missing and use models such as regression, Bayesian-based models or decisive tree, KNN to determine. These models can be trained and use other attributes of the data set.

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

→In specific, with WEKA, we can use Filter **ReplaceMissingValues,** Filter **ReplaceMissing-WithUserConstant** :

- Filter **ReplaceMissingValues:** used to replace missing values with the mean (for numeric attributes) and mode (for discrete attributes) so as to solve the missing values troubles.



ReplaceMissingValues

Relation: breast-cancer-weka.filters.unsupervised.attribute.ReplaceMissingValues
Instances: 286
Attributes: 10
Sum of weights: 286

Name: node-caps
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | yes | 56 | 56 |
| 2 | no | 230 | 230 |

ReplaceMissingValues

Relation: breast-cancer-weka.filters.unsupervised.attribute.ReplaceMissingValues
Instances: 286
Attributes: 10
Sum of weights: 286

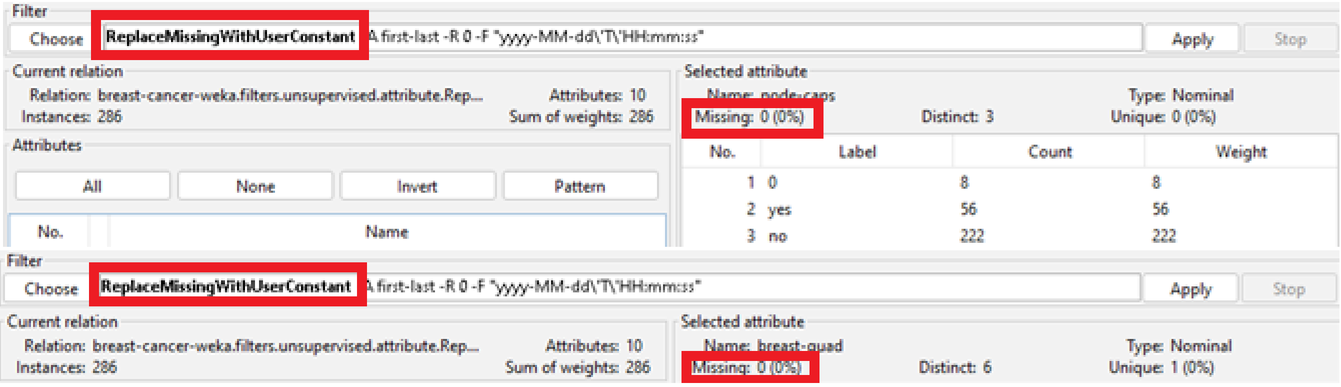
| No. | Name |
|---|---|
| 1 | age |
| 2 | menopause |

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

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | left_up | 97 | 97 |
| 2 | left_low | 111 | 111 |
| 3 | right_up | 33 | 33 |
| 4 | right_low | 24 | 24 |
| 5 | central | 21 | 21 |

*Picture 10. Example of using ReplaceMissingValues*

- Filter **ReplaceMissingWithUserConstant**: used to replace missing values with constant values filled by user.



Filter

ReplaceMissingWithUserConstant -A first-last -R 0 -F "yyyy-MM-dd\'T\'HH:mm:ss"

Current relation

Relation: breast-cancer-weka.filters.unsupervised.attribute.Rep... Attributes: 10

Instances: 286 Sum of weights: 286

Selected attribute

Name: node-caps Type: Nominal

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

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | 0 | 8 | 8 |
| 2 | yes | 56 | 56 |
| 3 | no | 222 | 222 |

Filter

ReplaceMissingWithUserConstant -A first-last -R 0 -F "yyyy-MM-dd\'T\'HH:mm:ss"

Current relation

Relation: breast-cancer-weka.filters.unsupervised.attribute.Rep... Attributes: 10

Instances: 286 Sum of weights: 286

Selected attribute

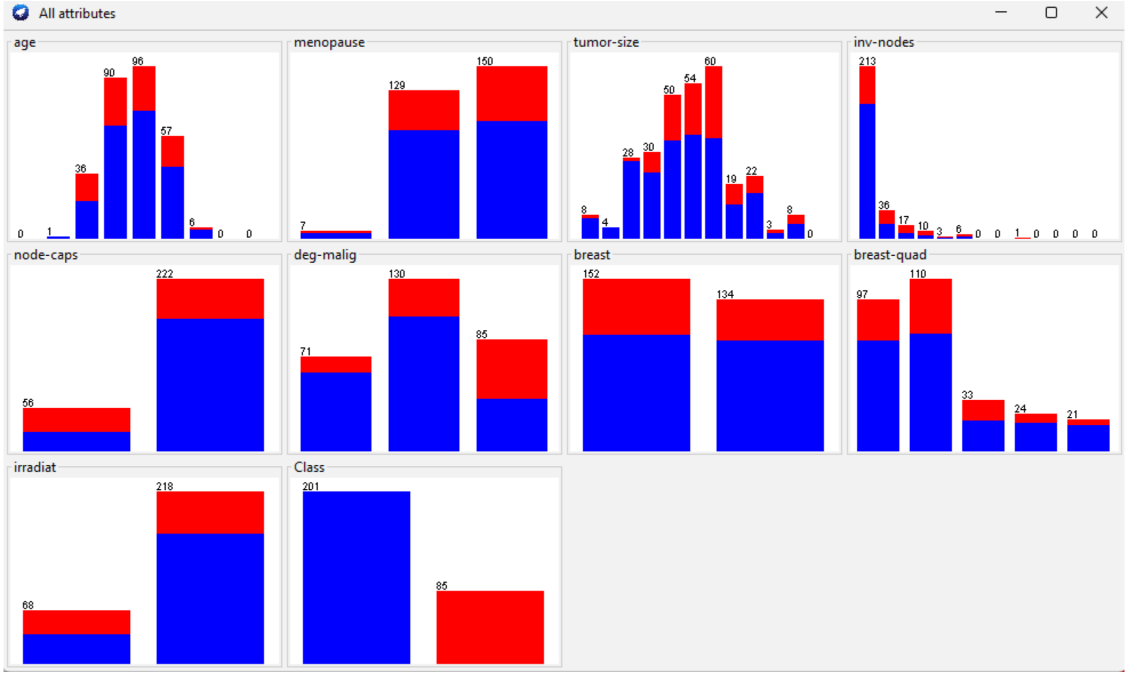
Name: breast-quad Type: Nominal

Missing: 0 (0%) Distinct: 6 Unique: 1 (0%)

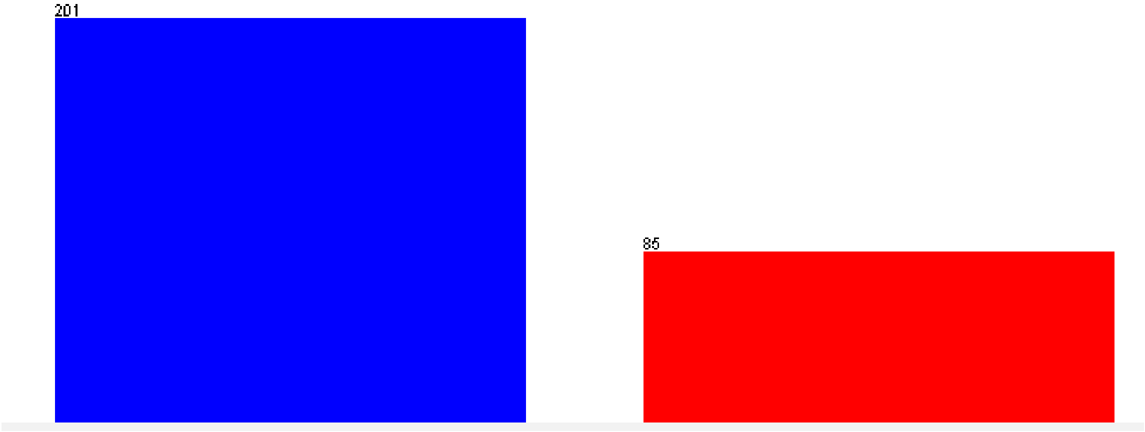
*Picture 11. Example of using ReplaceMissingWithUserConstant*

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

- The chart in WEKA Explorer shows the distribution of the attribute's values
- The chart in the right corner shows the number of samples according to each label of each attribute.
- The columns are "Label" displayed in the Selected Attribute frame, the column height is the size of "Count".
- For numeric attributes, this graph will be a histogram, dividing the value domain [min, max] into many subdomains [$a_x$, $b_x$] with approximately the same size. For each subdomain, we count the number of samples whose attribute values are in the domain and present it as a column chart.
- For a discrete attribute (nominal), for each attribute value, count the number of samples with that value and also represent it as a bar chart.
- The chart also shows the relative number of samples for each label. Each column will contain many colors stacked, each color corresponds to a branch (for example, here will be blue and red).



*Picture 12. Charts of all the attributes of Breast cancer data set*



*Picture 13. The chart shows the distribution of the number of samples of the* ***class*** *attribute*

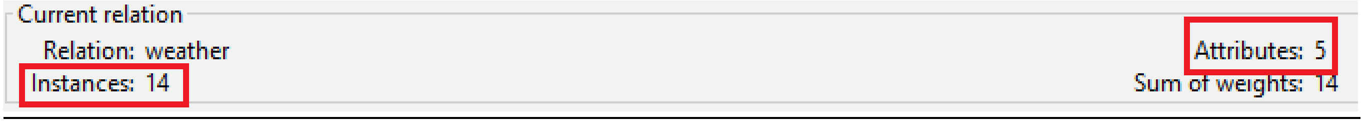
- The legend of its is ***class*** attribute: blue color corresponds to the number of samples labeled "*no-recurrence-events*", and red color corresponds to the number of samples labeled "*recurrence-events*".

→Therefore, our team reckon that the title for chart in WEKA Explorer should be ***"The posibility of patients to be recurrence-events or no-recurrence-events"*** in general.

## 2. Exploring Weather data set

"Second, you will load the data file namely weather.numeric.arff into the WEKA explorer. After successful, let's look at the Explorer site to answer questions or perform requirements in the followings:"

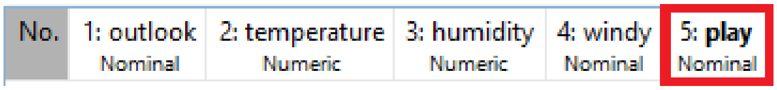
a. How many attributes does this data set have? How many samples? Which attributes have data type categorical? Which attributes have a data type that is numerical? Which attribute is used for the label?



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

*Picture 14. Picture about the number of attributes and samples weather data set*

- ➔ This data set has 5 attributes và 14 samples
- ➔ These attributes are devided into 2 types: **Numeric & Categorical**
- ■ **Numeric:** *Temperature, Humidity*
- ■ **Categorical:** *Outlook, Play, Windy*
- ➔ Thuộc tính dùng làm lớp là "**play**" có 2 giá trị *Yes* và *No*.
- ➔ Attribute "**play**" is used for the label which has 2 values **Yes** and **No**



| No. | 1: outlook | 2: temperature | 3: humidity | 4: windy | 5: play |
|---|---|---|---|---|---|
|  | Nominal | Numeric | Numeric | Nominal | Nominal |

*Picture 15. "Play" is used as class attribute*

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

- ■ **Five-number summary includes:**
- ■ Highest value in the dataset.
- ■ Third quartile (Q3) - greater than 75% of the values in the dataset
- ■ Median or second quartile (Q2) - splits the dataset in half.
- ■ First quartile (Q1) - greater than 25% of the values.
- ■ Lowest value in the dataset.

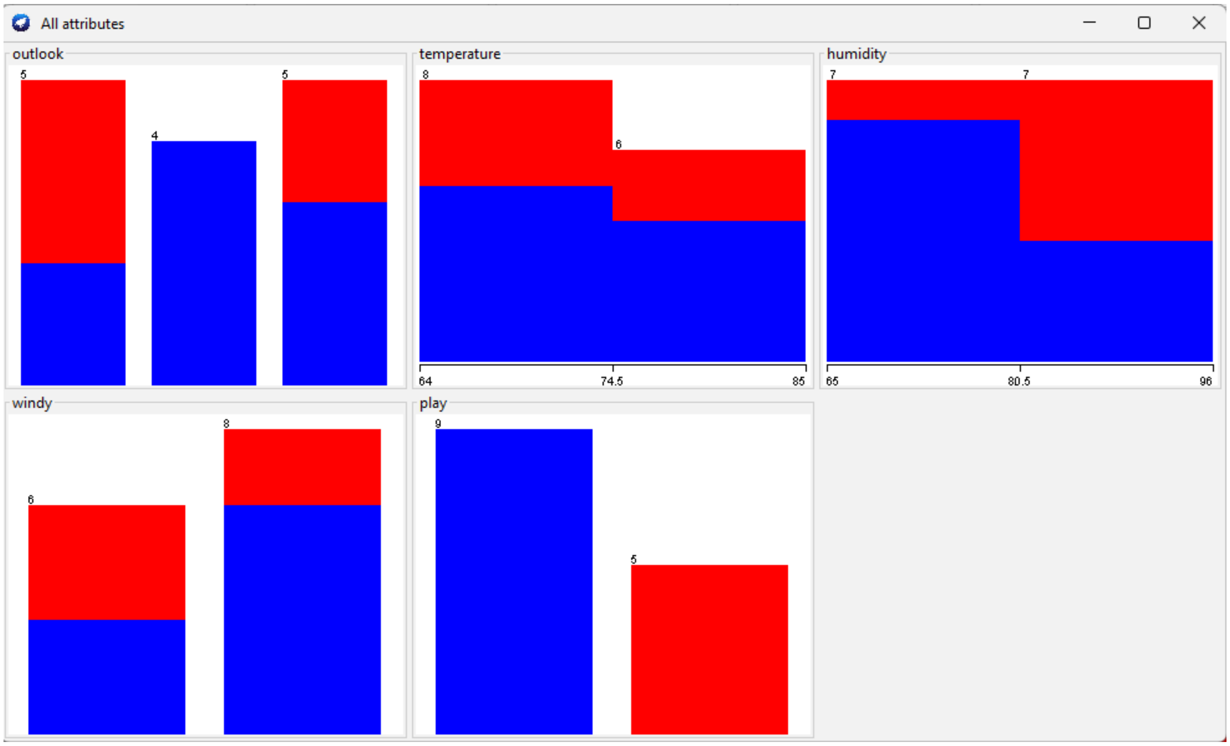
| **Attribute** | **Min** | **1st quartile (Q1)** | **Median** | **3rd quartile (Q3)** | **Max** |
|---|---|---|---|---|---|
| *Temperature* | 64 | 69 | 74.5 | 80 | 85 |
| *Humidity* | 65 | 70 | 80.5 | 90 | 96 |

- In WEKA has:

| **Attribute** | **Min** | **1st quartile (Q1)** | **Median** | **3rd quartile (Q3)** | **Max** |
|---|---|---|---|---|---|
| *Temperature* | 64 | X | X | X | 85 |
| *Humidity* | 65 | X | X | X | 96 |

→WEKA only provides us 2 values: **Min** and **Max**, it lacks of **First quartile**, **Median** and **Third quartile**.

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

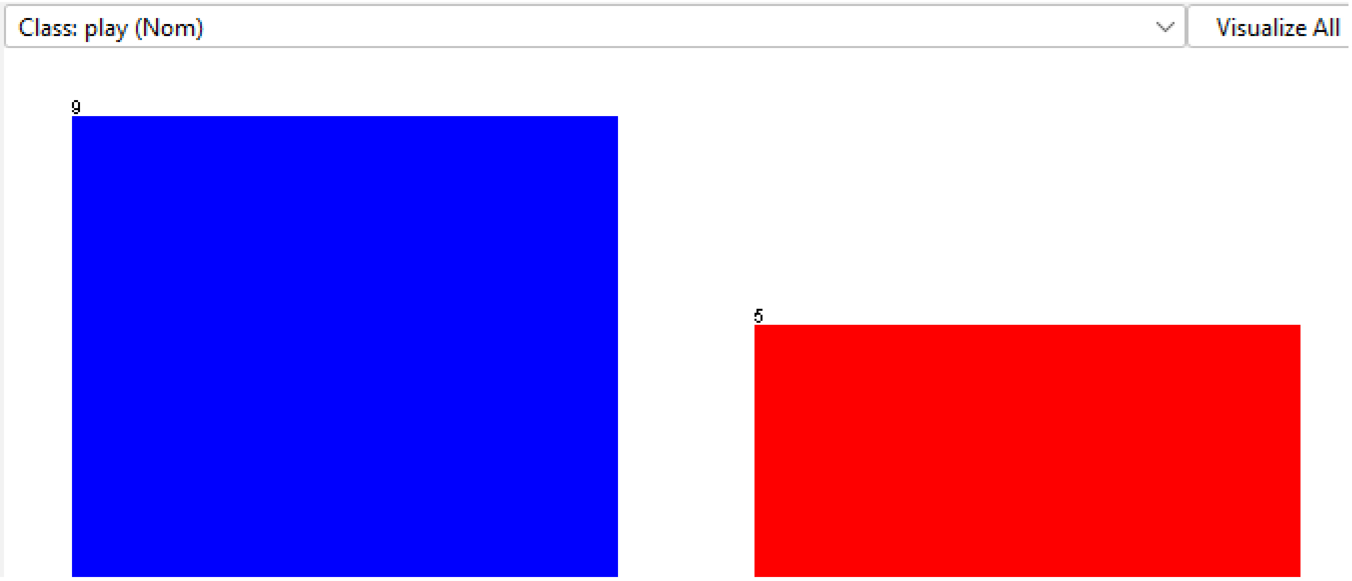


*Picture 16. Chart displays all the attributes of weather data set*

- Outlook:
- *In the **sunny** label, the number of values that satisfy the **yes** label of the class is more than the number of values that satisfy the **no** label, while the opposite is true for the **rainy** label.*
- *In the **overcast** label, all values satisfy the **yes** label of the class.*
- Temperature:
- *During this time, the label **yes** is always more than the **no**.*
- Humidity:
- *We can see that from 65 to 80.5, almost the values are **yes**, but in contrast, from 80.5 to 96 the **no** is more than **yes**.*

- Windy:
- *We notice that in the **False** label, the number of values that satisfy the **yes** label of the class is significantly more than the number of values that satisfy the **no** label of the class, and in the **True** label, the number of values that satisfy the 2 labels of the class is equal.*
- Play:
- *Because this is the "class" attribute of the data, through the graph below we can see the number of values distributed in the two labels **yes** and **no** specifically (the number of values that satisfy the yes label is 9, while Satisfactory value for label no is 5).*

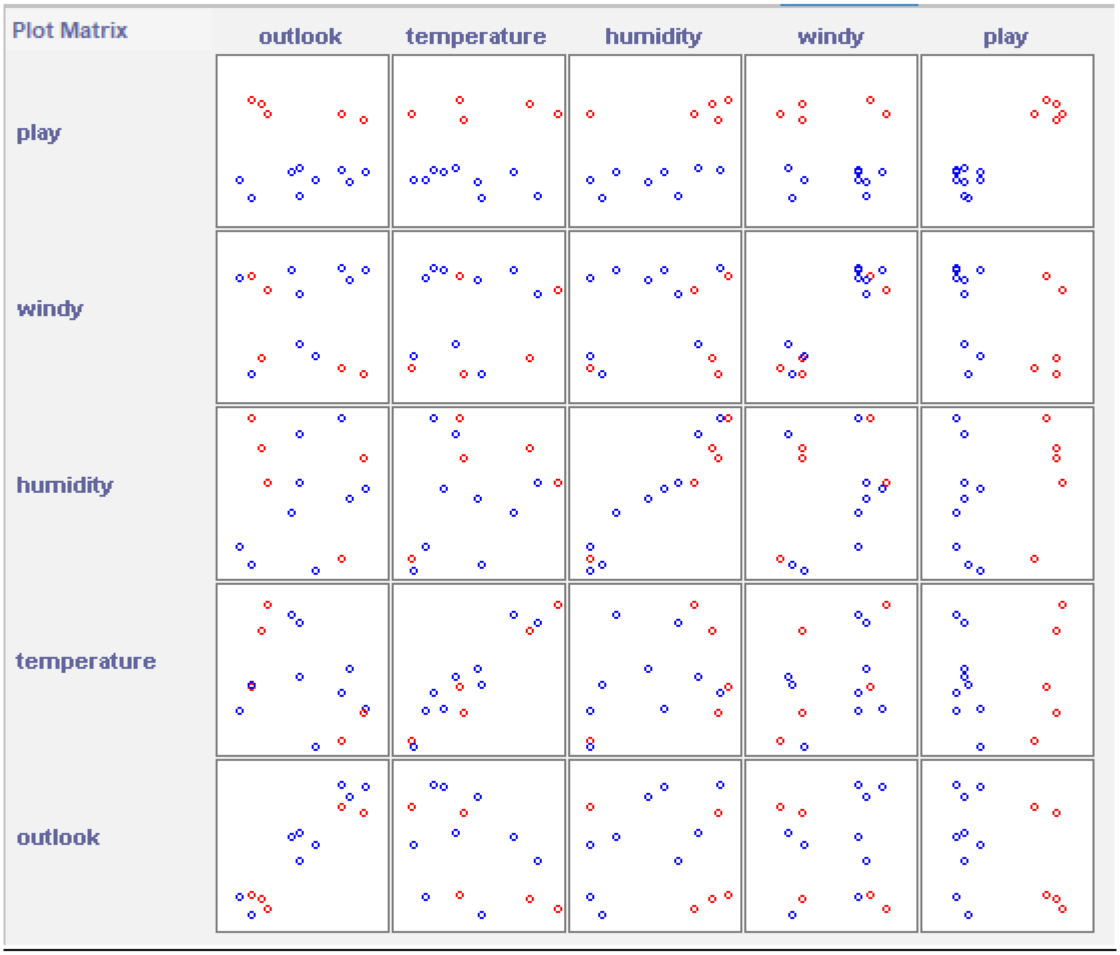
→ The meaning of all charts in the WEKA Explorer: because "*Class*" is a label attribute, all of the charts are histogram, with only two columns displaying the density of *Class* value (attribute for the label). Besides that, other charts are stacked histograms, with two stacks differentiated by two "Class" categories and bars displaying the density of value of those attributes. On all charts, the blue column represents **yes** class and the red column represents **no** class.



*Picture 17. Chart of Class attribute (play attribute) of Weather data set*

- ➔ *This is the class attribute (play attribute) which is the legend of  its, there is 9 for **yes** and 5 for **no,** which has the value of distribution left – skewed.*
- ➔ *We can tittle for it: "The chart displays the possibility whether we should go out to play or not".*

d. 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?



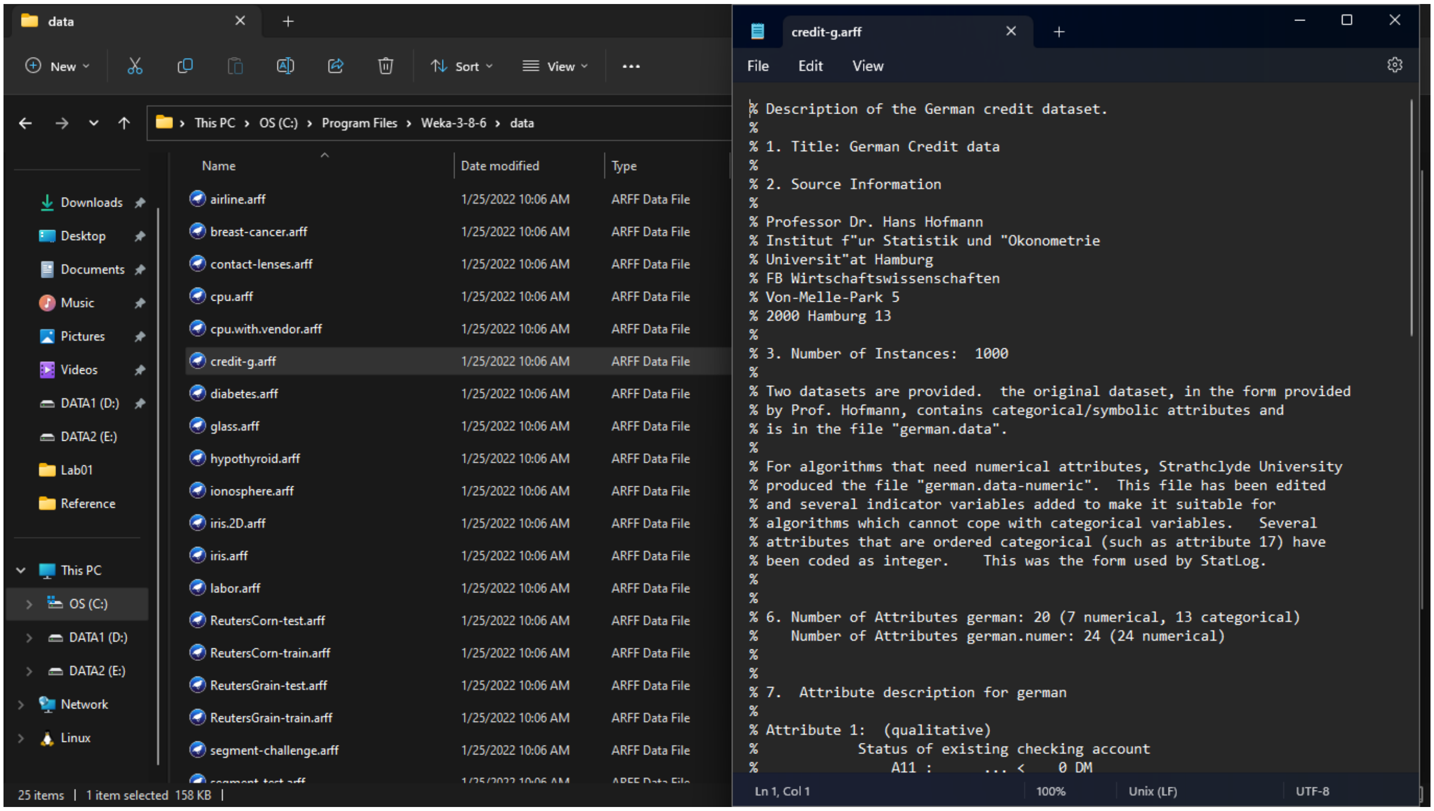
*Picture 18. Scatter plot of the attributes in the Weather dataset*

- ➔ The name of this chart is "scatter plot".
- ➔ For us, pair (temperature, humidity), (humidity, windy), (outlook, play) is fairly correlated.

### 3. Exploring Credit in Germany data set

"Similarly, you will also load the data file namely credit-g.arff into the WEKA explorer. After successful, let's look at the Explorer site to answer questions or perform requirements in the followings"

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).



| Name | Date modified | Type |
|---|---|---|
| airline.arff | 1/25/2022 10:06 AM | ARFF Data File |
| breast-cancer.arff | 1/25/2022 10:06 AM | ARFF Data File |
| contact-lenses.arff | 1/25/2022 10:06 AM | ARFF Data File |
| cpu.arff | 1/25/2022 10:06 AM | ARFF Data File |
| cpu.with.vendor.arff | 1/25/2022 10:06 AM | ARFF Data File |
| credit-g.arff | 1/25/2022 10:06 AM | ARFF Data File |
| diabetes.arff | 1/25/2022 10:06 AM | ARFF Data File |
| glass.arff | 1/25/2022 10:06 AM | ARFF Data File |
| hypothyroid.arff | 1/25/2022 10:06 AM | ARFF Data File |
| ionosphere.arff | 1/25/2022 10:06 AM | ARFF Data File |
| iris.2D.arff | 1/25/2022 10:06 AM | ARFF Data File |
| iris.arff | 1/25/2022 10:06 AM | ARFF Data File |
| labor.arff | 1/25/2022 10:06 AM | ARFF Data File |
| ReutersCorn-test.arff | 1/25/2022 10:06 AM | ARFF Data File |
| ReutersCorn-train.arff | 1/25/2022 10:06 AM | ARFF Data File |
| ReutersGrain-test.arff | 1/25/2022 10:06 AM | ARFF Data File |
| ReutersGrain-train.arff | 1/25/2022 10:06 AM | ARFF Data File |
| segment-challenge.arff | 1/25/2022 10:06 AM | ARFF Data File |

```
% 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
%
% Attribute 1:  (qualitative)
%            Status of existing checking account
%                A11 :      ... <    0 DM
```

*Picture 19. Content of Credit in Germany data set when opening with Notepad*

→The content of the notes when opening the file with notepad is some basic information about the dataset which we can view information: dataset name, source information, number of instances, number of attributes, parameters of attributes (equivalent to the selected attribute frame in weka) and cost matrix.

→The data set has 1000 samples.

→It also has 21 attributes.

→Five attributes that we describe:

| Name of Attribute | Description | Data type |
|---|---|---|
| *duration* | Loan term (calculate by month) | Continuous attribute |
| *purpose* | Customer credit card usage's goals. | Discrete attribute |
| *personal_status* | Indicate the status (gender, marriage) of clients | Discrete attribute |
| *age* | The age of customers | Continuous attribute |
| *job* | Occupation status | Discrete attribute |

**Insights:**

- *Duration*
- It is a continuous attribute because its value can be display form 1 to 12 continuously, max value of this column is the min value of the next column.
- *Purpose*
- It is a discrete attribute because the targets of customers are different, such as:

Buy new car

Own radio or TV

Pay for education

Spend on business

Buy used car

Buy domestic appliances

Pay for vacation

Spend on other aspects

Purchase furniture or equipment

Pay for repairs

Pay for retrainning

- *personal_status*
- This attribute is adiscrete attribute clearly because the gender and status of mariage of each client is different, as:

male : divorced/separated

female : divorced/separated/married

male : single

male : married/widowed

female : single

- *Age*
- This value is continuous from 19 to 75, max value of this column is the min value of the next column.
- *Job*
- This value is so independent because it shows different status of job, such as:

unemployed/ unskilled - non-resident

unskilled - resident

skilled employee / official

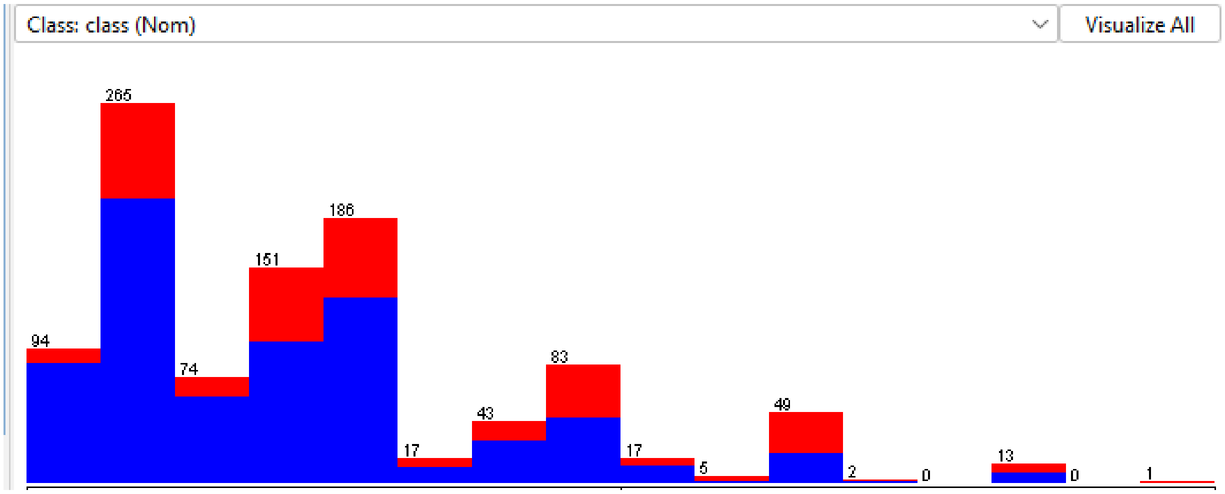
highly qualified employee/ management/ self-employed

b. <u>Which attribute is used for the label?</u>

➔ Attribute is used for the label is “class”, which can be also called class attribute.

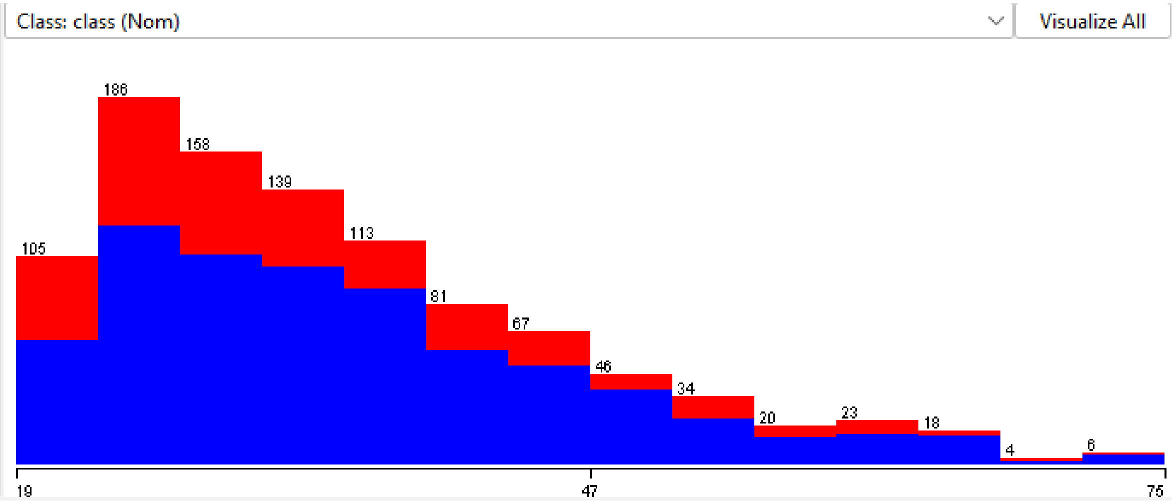
c. <u>Let’s describe the distribution of continuous attributes?(Left skewed or right skewed ?)</u>

➔ Continuous attributes are: *duration*, *age*



*Picture 20. Chart of duration distribution*

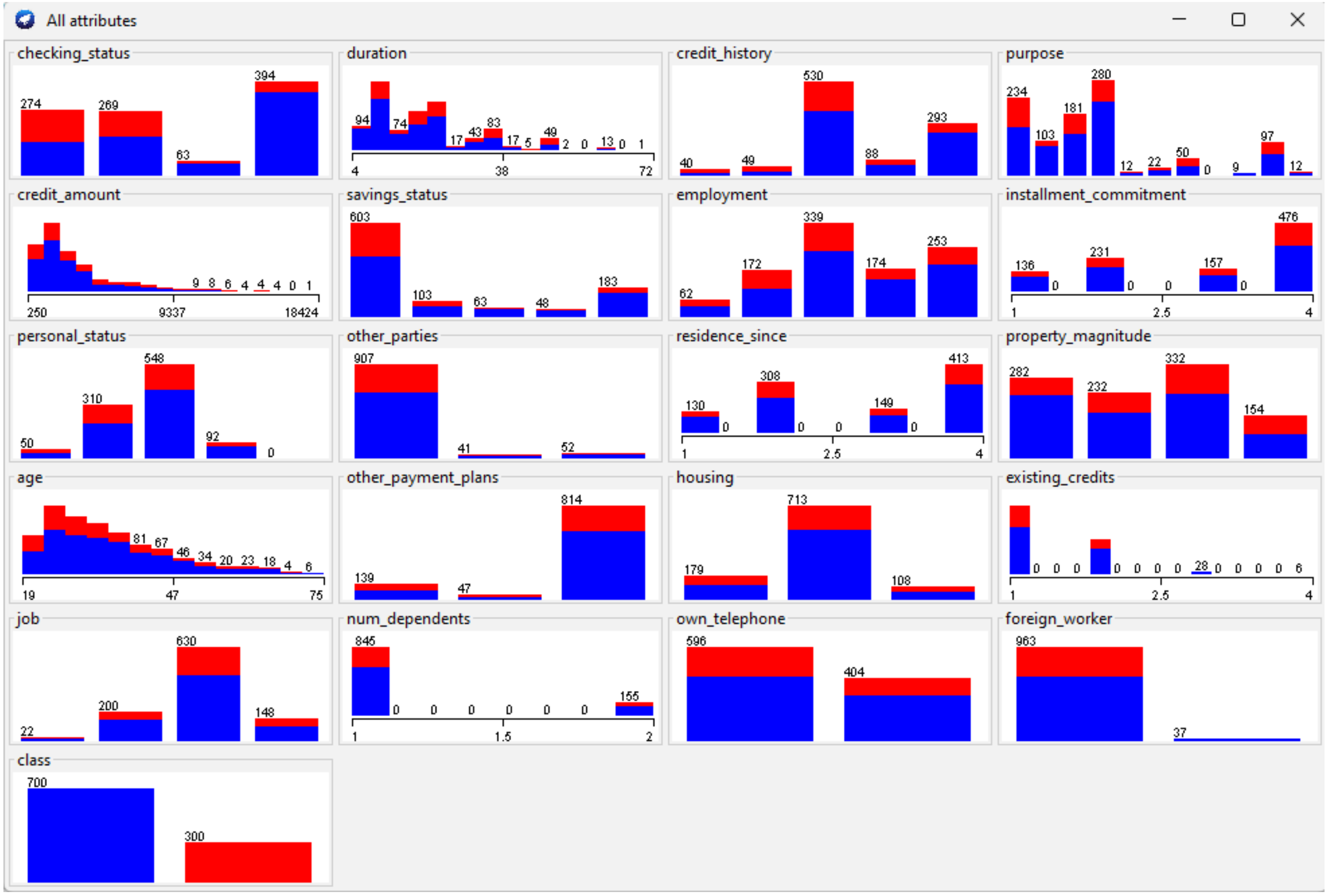
- *Minimum: 4, Maximun: 72, Mean: 20.903, Standard deviation:12.059, left - skewed*



*Picture 21. Chart of age distribution*

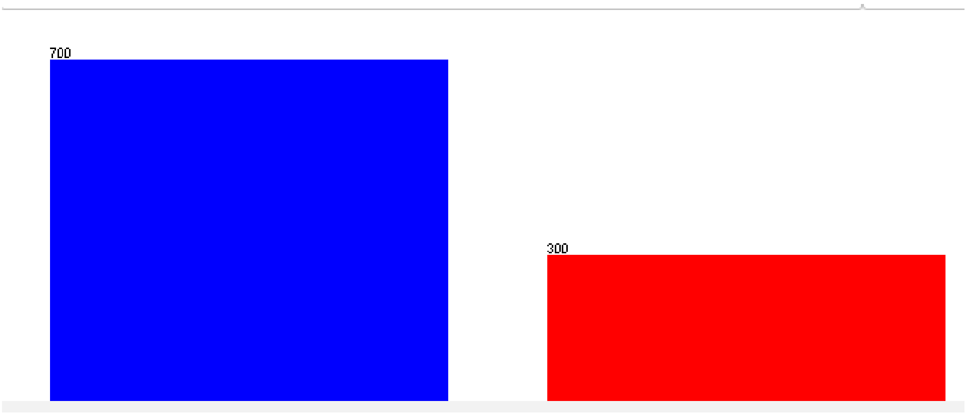
- *Minimum: 19, Maximun: 75, Mean: 35.546, Standard deviation:11.375, left - skewed*

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



*Picture 22. Charts displays all the attribute of credit card in Germany of the WEKA Explorer*

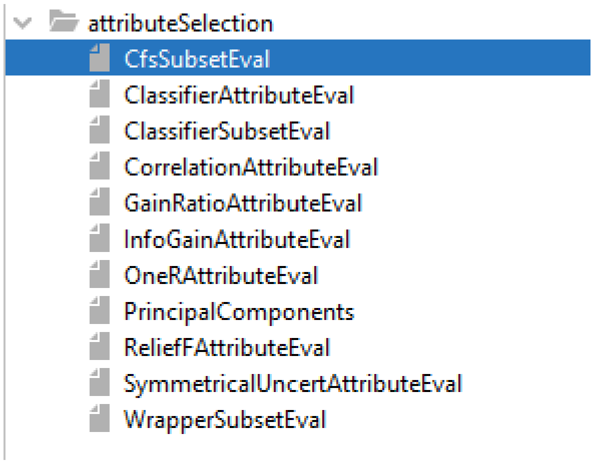
→ The meaning of all charts in the WEKA Explorer: because “*Class*” is a label attribute, all of the charts are histogram, with only two columns displaying the density of *Class* value (attribute for the label). Besides that, other charts are stacked histograms, with two stacks differentiated by two “*Class*” categories and bars displaying the density of value of those attributes. On all charts, the blue column represents **good** class and the red column represents **bad** class.



*Picture 23. Chart of class attribute of Credit in Germany data set*

- ➔ We can find out that the class attribute is the attribute for the label named "Class", which has the values distribution left - skewed
- ➔ We can call that *"The chart shows the ability to pay for credit card in German is good or bad"*

e. <u>Let's move to the Select attributes tag. Describe all of the options for attribute selection.</u>



attributeSelection
- CfsSubsetEval
- ClassifierAttributeEval
- ClassifierSubsetEval
- CorrelationAttributeEval
- GainRatioAttributeEval
- InfoGainAttributeEval
- OneRAttributeEval
- PrincipalComponents
- ReliefFAttributeEval
- SymmetricalUncertAttributeEval
- WrapperSubsetEval

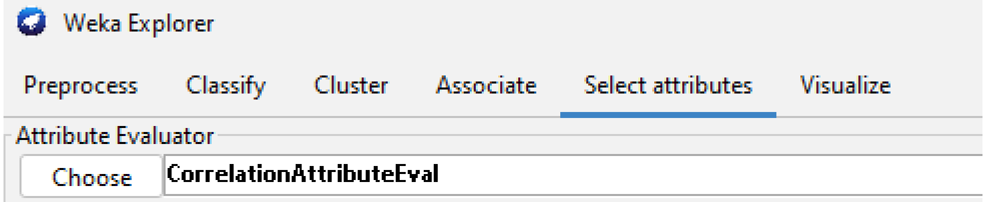
*Picture 24. All options of the attribute selection*

| Name of option | Description |
|---|---|
| ***CfsSubsetEval*** | Evaluate the value of a subset of attributes by looking at the individual predictability of each feature along with the degree of redundancy between them. |
| ***ClassifierAttributeEval*** | Evaluate the value of a subset of attributes using a user-specified classifier. |
| ***ClassifierSubsetEval*** | Evaluate attribute subsets on separate training data or pause test sets. |

| | |
|---|---|
| ***CorrelationAttributeEval*** | Evaluates the value of an attribute by measuring the correlation between it and the class. |
| ***GainRatioAttributeEval*** | Evaluate the value of an attribute by measuring the gain relative to the class. |
| ***InfoGainAttributeEval*** | Evaluates the value of an attribute by measuring acquired class-related information. |
| ***OneRAttributeEval*** | Evaluate the value of an attribute using the OneR classifier. |
| ***PrincipalComponents*** | Perform analysis and transformation of key components of data. |
| ***ReliefFAttributeEval*** | Evaluates the value of an attribute by repeatedly sampling an object and considering the value of the given attribute for the nearest object of the same and different classes. |
| ***SymmetricalUncertAttributeEval*** | Evaluate the value of an attribute by measuring the symmetric measurement uncertainty with respect to the class. |
| ***WrapperSubsetEval*** | Evaluate attribute sets using a learning schema. |

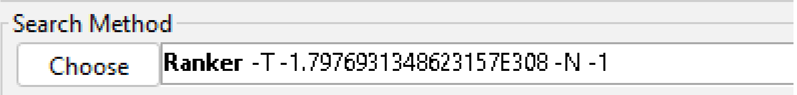
f. <u>Which options should be used to select the 5 attributes with the highest correlation?(Step-by-step description, with step-by-step photos and final results)</u>

- We use the **CorrelationAttributeEval** filter to select the attributes highest correlation with the class attribute.
- Step-by-step description:
- <u>Step 1:</u> In the Select attributes tab in the *Attribute Evaluator* section, select **Correlation-AttributeEval**



Attribute Evaluator: CorrelationAttributeEval

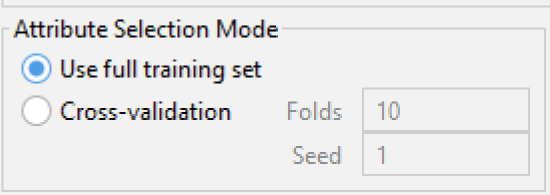
Then *Search Method* will be automatically selected as **Ranker**



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

*Picture 25. First step demo*

- Step 2: After that, in the *Attribute Selection Mode*, choose **Use full training set**



Attribute Selection Mode

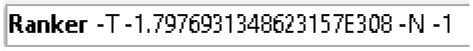
● Use full training set

○ Cross-validation Folds 10

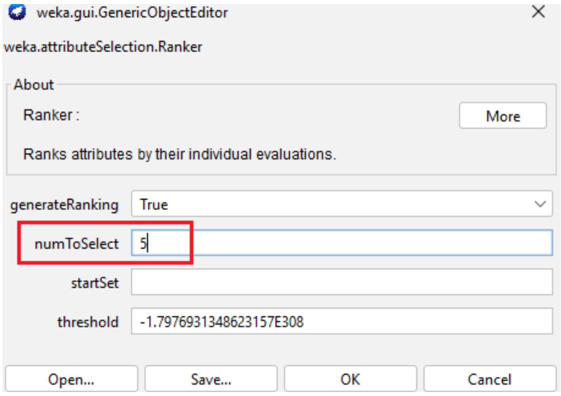
Seed 1

*Picture 26. Second step demo*

- Step 3: Click to the content **Ranker** bellow, a table will display and choose **numToSelect** is 5 and click **Start**



**Ranker** -T -1.7976931348623157E308 -N -1



weka.gui.GenericObjectEditor

weka.attributeSelection.Ranker

About

Ranker :

Ranks attributes by their individual evaluations.

generateRanking True

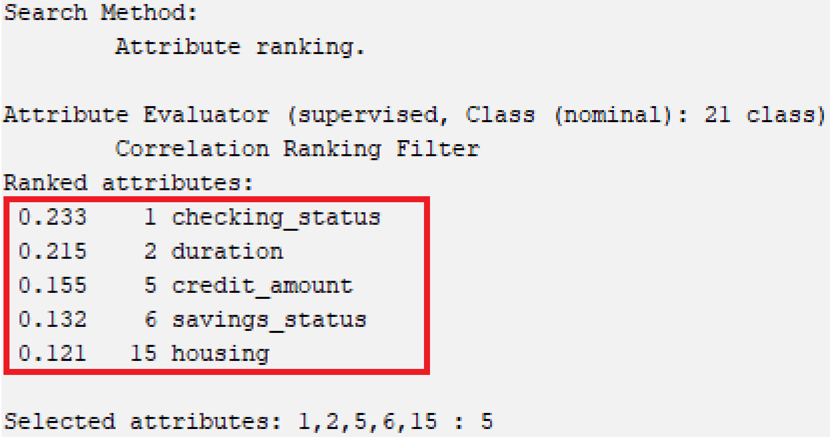
numToSelect 5

startSet

threshold -1.7976931348623157E308

*Picture 27. Third step demo*

→The final result we get is the attributes sorted by correlation with the class



```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Correlation Ranking Filter
Ranked attributes:
 0.233     1 checking_status
 0.215     2 duration
 0.155     5 credit_amount
 0.132     6 savings_status
 0.121    15 housing

Selected attributes: 1,2,5,6,15 : 5
```

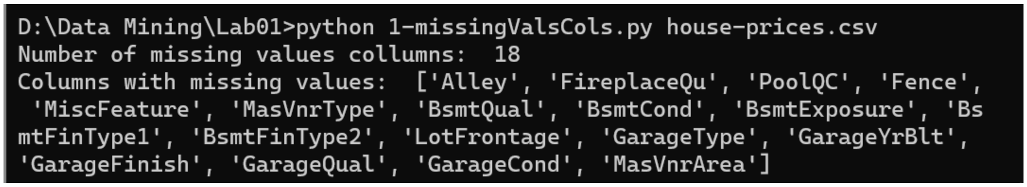
*Picture 28. Result*

In conclusion, 5 attributes with the highest correlation are: ***checking_status, duration, credit_amount, savings_status, housing.***

## IV. Preprocessing Data in Python

### 1. Extract columns with missing values

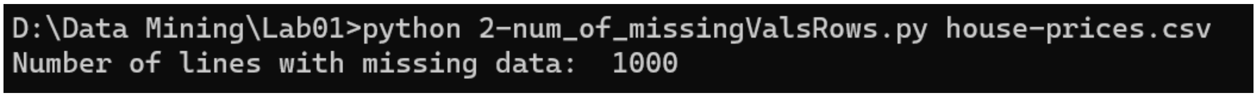
- Command line arguments: python 1-missingValsCols.py <FileIn>
- Output: number of missing values collumns and a list of them.
- Example: python 1-missingValsCols.py house-prices.csv



```
D:\Data Mining\Lab01>python 1-missingValsCols.py house-prices.csv
Number of missing values collumns:  18
Columns with missing values:  ['Alley', 'FireplaceQu', 'PoolQC', 'Fence',
 'MiscFeature', 'MasVnrType', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'Bs
mtFinType1', 'BsmtFinType2', 'LotFrontage', 'GarageType', 'GarageYrBlt',
'GarageFinish', 'GarageQual', 'GarageCond', 'MasVnrArea']
```

### 2. Count the number of lines with missing data

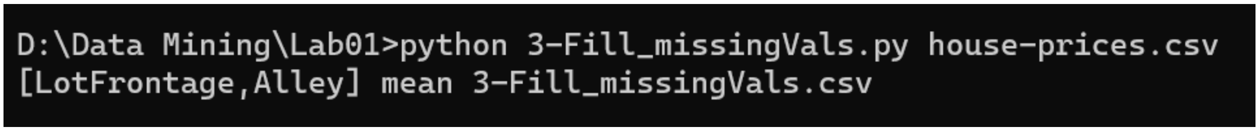
- Command line arguments: python 2-num_of_missingValsRows.py <FileIn>
- Output: number of lines with missing data.
- Example: python 2-num_of_missingValsRows.py house-prices.csv



```
D:\Data Mining\Lab01>python 2-num_of_missingValsRows.py house-prices.csv
Number of lines with missing data:  1000
```

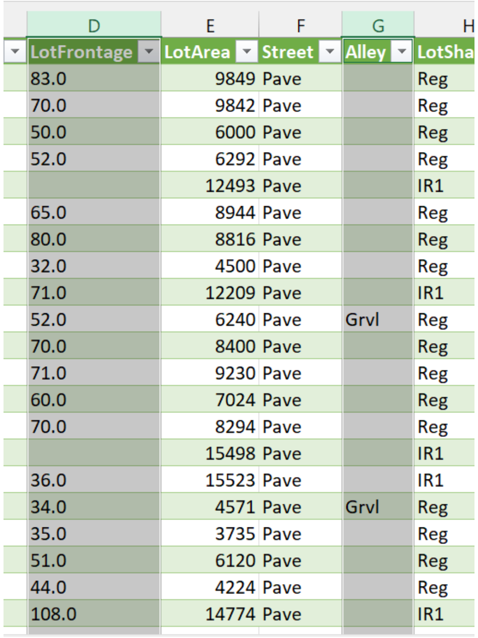
### 3. Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute)

- Command line arguments: python 3-Fill_missingVals.py <FileIn> <[collumn name]> <method> <FileOut>
- Output: Written in FileOut.
- Example 1: (mean method for numerical and mode for categorical): python 3-Fill_missingVals.py house-prices.csv [LotFrontage,Alley] mean 3-Fill_missingVals.csv



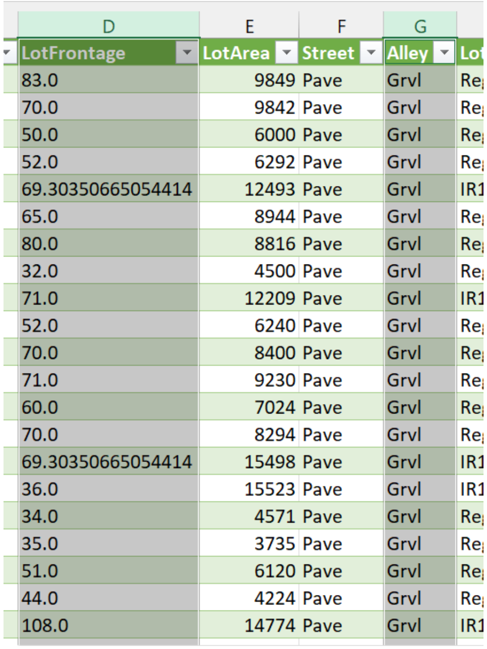
```
D:\Data Mining\Lab01>python 3-Fill_missingVals.py house-prices.csv
[LotFrontage,Alley] mean 3-Fill_missingVals.csv
```

***Before:***



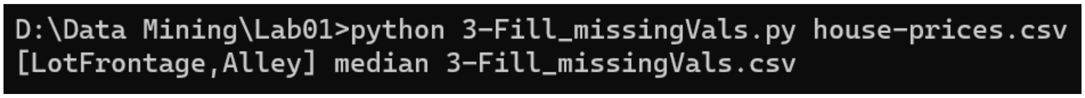
| D | E | F | G | H |
|---|---|---|---|---|
| LotFrontage | LotArea | Street | Alley | LotSha |
| 83.0 | 9849 | Pave | | Reg |
| 70.0 | 9842 | Pave | | Reg |
| 50.0 | 6000 | Pave | | Reg |
| 52.0 | 6292 | Pave | | Reg |
| | 12493 | Pave | | IR1 |
| 65.0 | 8944 | Pave | | Reg |
| 80.0 | 8816 | Pave | | Reg |
| 32.0 | 4500 | Pave | | Reg |
| 71.0 | 12209 | Pave | | IR1 |
| 52.0 | 6240 | Pave | Grvl | Reg |
| 70.0 | 8400 | Pave | | Reg |
| 71.0 | 9230 | Pave | | Reg |
| 60.0 | 7024 | Pave | | Reg |
| 70.0 | 8294 | Pave | | Reg |
| | 15498 | Pave | | IR1 |
| 36.0 | 15523 | Pave | | IR1 |
| 34.0 | 4571 | Pave | Grvl | Reg |
| 35.0 | 3735 | Pave | | Reg |
| 51.0 | 6120 | Pave | | Reg |
| 44.0 | 4224 | Pave | | Reg |
| 108.0 | 14774 | Pave | | IR1 |

***After:***



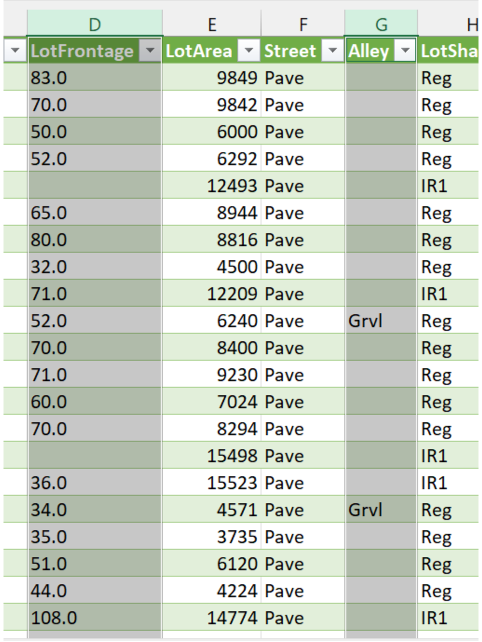
| D | E | F | G | |
|---|---|---|---|---|
| LotFrontage | LotArea | Street | Alley | Lot |
| 83.0 | 9849 | Pave | Grvl | Re |
| 70.0 | 9842 | Pave | Grvl | Re |
| 50.0 | 6000 | Pave | Grvl | Re |
| 52.0 | 6292 | Pave | Grvl | Re |
| 69.30350665054414 | 12493 | Pave | Grvl | IR1 |
| 65.0 | 8944 | Pave | Grvl | Re |
| 80.0 | 8816 | Pave | Grvl | Re |
| 32.0 | 4500 | Pave | Grvl | Re |
| 71.0 | 12209 | Pave | Grvl | IR1 |
| 52.0 | 6240 | Pave | Grvl | Re |
| 70.0 | 8400 | Pave | Grvl | Re |
| 71.0 | 9230 | Pave | Grvl | Re |
| 60.0 | 7024 | Pave | Grvl | Re |
| 70.0 | 8294 | Pave | Grvl | Re |
| 69.30350665054414 | 15498 | Pave | Grvl | IR1 |
| 36.0 | 15523 | Pave | Grvl | IR1 |
| 34.0 | 4571 | Pave | Grvl | Re |
| 35.0 | 3735 | Pave | Grvl | Re |
| 51.0 | 6120 | Pave | Grvl | Re |
| 44.0 | 4224 | Pave | Grvl | Re |
| 108.0 | 14774 | Pave | Grvl | IR1 |

- <u>Example 2:</u> (median method for numerical and mode for categorical): python 3-Fill_missingVals.py house-prices.csv [LotFrontage,Alley] median 3-Fill_missingVals.csv



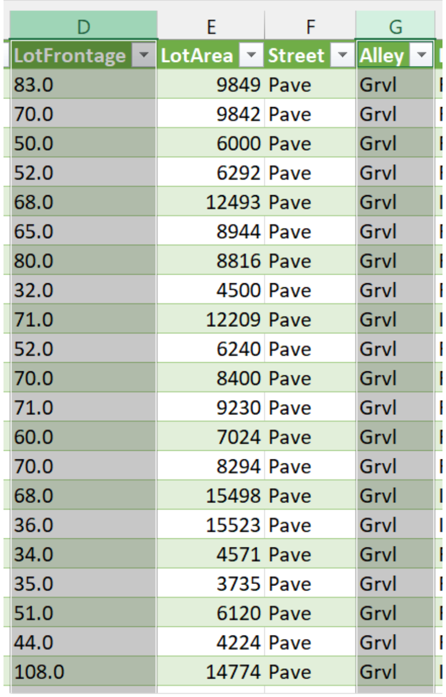
```
D:\Data Mining\Lab01>python 3-Fill_missingVals.py house-prices.csv
[LotFrontage,Alley] median 3-Fill_missingVals.csv
```

***Before:***



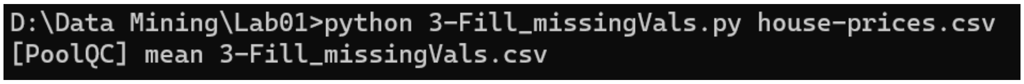
| D | E | F | G | H |
|---|---|---|---|---|
| LotFrontage | LotArea | Street | Alley | LotSha |
| 83.0 | 9849 | Pave | | Reg |
| 70.0 | 9842 | Pave | | Reg |
| 50.0 | 6000 | Pave | | Reg |
| 52.0 | 6292 | Pave | | Reg |
| | 12493 | Pave | | IR1 |
| 65.0 | 8944 | Pave | | Reg |
| 80.0 | 8816 | Pave | | Reg |
| 32.0 | 4500 | Pave | | Reg |
| 71.0 | 12209 | Pave | | IR1 |
| 52.0 | 6240 | Pave | Grvl | Reg |
| 70.0 | 8400 | Pave | | Reg |
| 71.0 | 9230 | Pave | | Reg |
| 60.0 | 7024 | Pave | | Reg |
| 70.0 | 8294 | Pave | | Reg |
| | 15498 | Pave | | IR1 |
| 36.0 | 15523 | Pave | | IR1 |
| 34.0 | 4571 | Pave | Grvl | Reg |
| 35.0 | 3735 | Pave | | Reg |
| 51.0 | 6120 | Pave | | Reg |
| 44.0 | 4224 | Pave | | Reg |
| 108.0 | 14774 | Pave | | IR1 |

***After:***



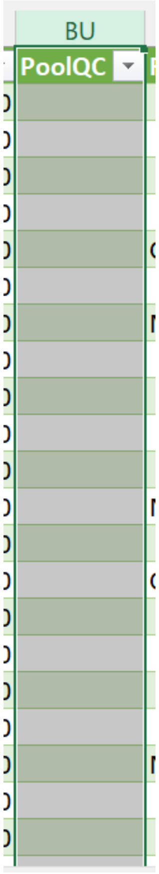
| D | E | F | G |
|---|---|---|---|
| LotFrontage | LotArea | Street | Alley |
| 83.0 | 9849 | Pave | Grvl |
| 70.0 | 9842 | Pave | Grvl |
| 50.0 | 6000 | Pave | Grvl |
| 52.0 | 6292 | Pave | Grvl |
| 68.0 | 12493 | Pave | Grvl |
| 65.0 | 8944 | Pave | Grvl |
| 80.0 | 8816 | Pave | Grvl |
| 32.0 | 4500 | Pave | Grvl |
| 71.0 | 12209 | Pave | Grvl |
| 52.0 | 6240 | Pave | Grvl |
| 70.0 | 8400 | Pave | Grvl |
| 71.0 | 9230 | Pave | Grvl |
| 60.0 | 7024 | Pave | Grvl |
| 70.0 | 8294 | Pave | Grvl |
| 68.0 | 15498 | Pave | Grvl |
| 36.0 | 15523 | Pave | Grvl |
| 34.0 | 4571 | Pave | Grvl |
| 35.0 | 3735 | Pave | Grvl |
| 51.0 | 6120 | Pave | Grvl |
| 44.0 | 4224 | Pave | Grvl |
| 108.0 | 14774 | Pave | Grvl |

- Example 3: (PoolQC: all missing values): python 3-Fill_missingVals.py house-prices.csv [PoolQC] mean 3-Fill_missingVals.csv
- Output: PoolQC remains the same.

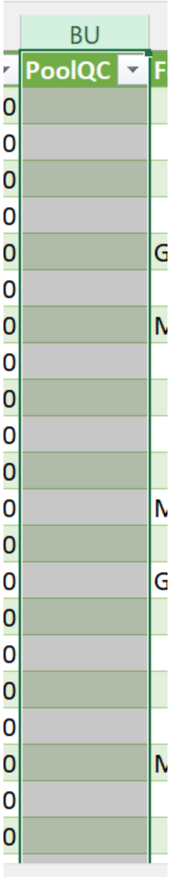


```
D:\Data Mining\Lab01>python 3-Fill_missingVals.py house-prices.csv
[PoolQC] mean 3-Fill_missingVals.csv
```

***Before:***

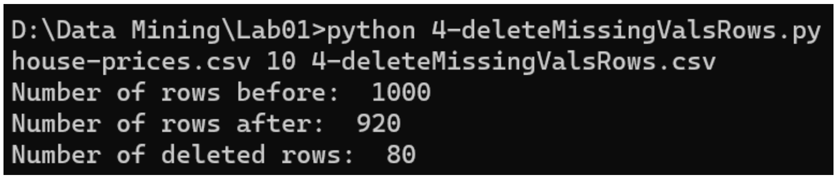


***After:***

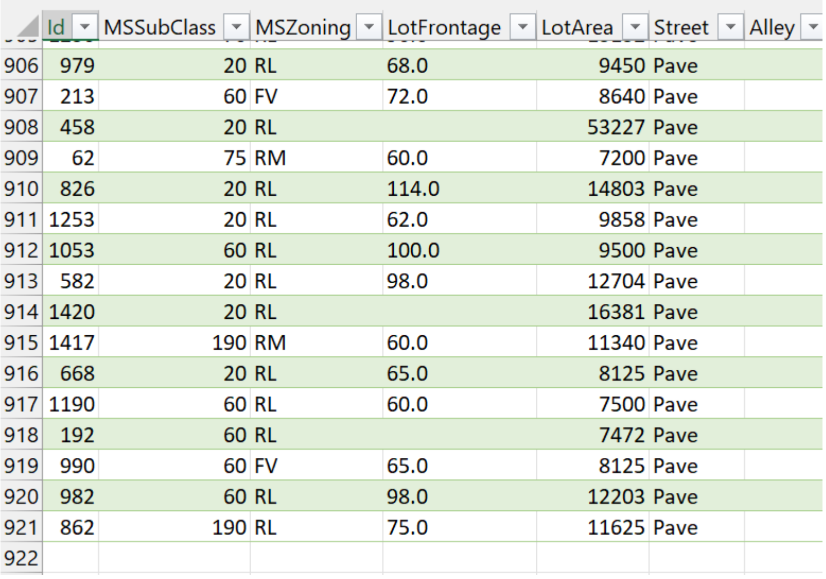


### 4. Deleting rows containing more than a particular number of missing values

- Command line arguments: python 4-deleteMissingValsRows.py <FileIn> <Percentage> <FileOut>
- Output: number of rows before, number of rows after, Number of deleted rows and write the result in FileOut.
- Example 1: python 4-deleteMissingValsRows.py house-prices.csv 10 4-deleteMissingValsRows.csv

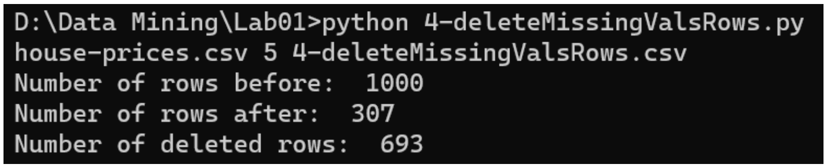


```
D:\Data Mining\Lab01>python 4-deleteMissingValsRows.py
house-prices.csv 10 4-deleteMissingValsRows.csv
Number of rows before:  1000
Number of rows after:  920
Number of deleted rows:  80
```

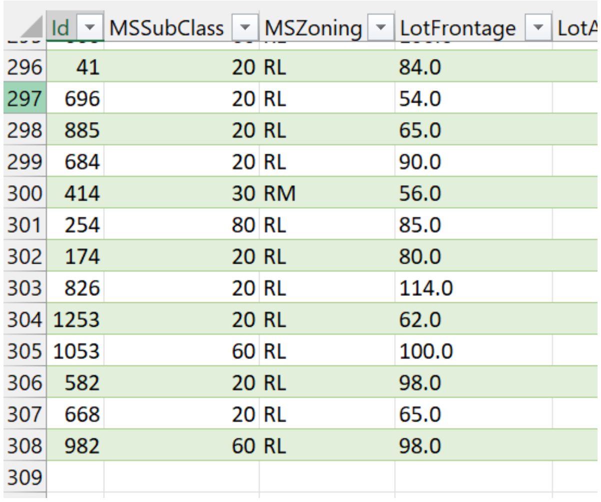


| | Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley |
|---|---|---|---|---|---|---|---|
| 906 | 979 | 20 | RL | 68.0 | 9450 | Pave | |
| 907 | 213 | 60 | FV | 72.0 | 8640 | Pave | |
| 908 | 458 | 20 | RL | | 53227 | Pave | |
| 909 | 62 | 75 | RM | 60.0 | 7200 | Pave | |
| 910 | 826 | 20 | RL | 114.0 | 14803 | Pave | |
| 911 | 1253 | 20 | RL | 62.0 | 9858 | Pave | |
| 912 | 1053 | 60 | RL | 100.0 | 9500 | Pave | |
| 913 | 582 | 20 | RL | 98.0 | 12704 | Pave | |
| 914 | 1420 | 20 | RL | | 16381 | Pave | |
| 915 | 1417 | 190 | RM | 60.0 | 11340 | Pave | |
| 916 | 668 | 20 | RL | 65.0 | 8125 | Pave | |
| 917 | 1190 | 60 | RL | 60.0 | 7500 | Pave | |
| 918 | 192 | 60 | RL | | 7472 | Pave | |
| 919 | 990 | 60 | FV | 65.0 | 8125 | Pave | |
| 920 | 982 | 60 | RL | 98.0 | 12203 | Pave | |
| 921 | 862 | 190 | RL | 75.0 | 11625 | Pave | |
| 922 | | | | | | | |

- Example 2: python 4-deleteMissingValsRows.py house-prices.csv 5 4-deleteMissingValsRows.csv

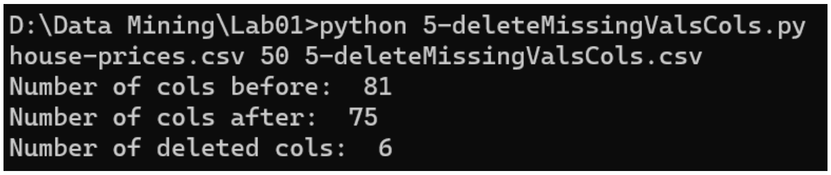


```
D:\Data Mining\Lab01>python 4-deleteMissingValsRows.py
house-prices.csv 5 4-deleteMissingValsRows.csv
Number of rows before:  1000
Number of rows after:  307
Number of deleted rows:  693
```

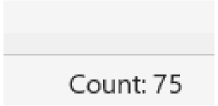


| | Id | MSSubClass | MSZoning | LotFrontage | LotA |
|---|---|---|---|---|---|
| 296 | 41 | 20 | RL | 84.0 | |
| 297 | 696 | 20 | RL | 54.0 | |
| 298 | 885 | 20 | RL | 65.0 | |
| 299 | 684 | 20 | RL | 90.0 | |
| 300 | 414 | 30 | RM | 56.0 | |
| 301 | 254 | 80 | RL | 85.0 | |
| 302 | 174 | 20 | RL | 80.0 | |
| 303 | 826 | 20 | RL | 114.0 | |
| 304 | 1253 | 20 | RL | 62.0 | |
| 305 | 1053 | 60 | RL | 100.0 | |
| 306 | 582 | 20 | RL | 98.0 | |
| 307 | 668 | 20 | RL | 65.0 | |
| 308 | 982 | 60 | RL | 98.0 | |
| 309 | | | | | |

5. **Deleting columns containing more than a particular number of missing values**
   - Command line arguments: python 5-deleteMissingValsCols.py <FileIn> <Percentage> <FileOut>
   - Output: number of cols before, number of cols after, Number of deleted cols and write the result in FileOut.
   - Example 1: python 5-deleteMissingValsCols.py house-prices.csv 50 5-deleteMissingValsCols.csv

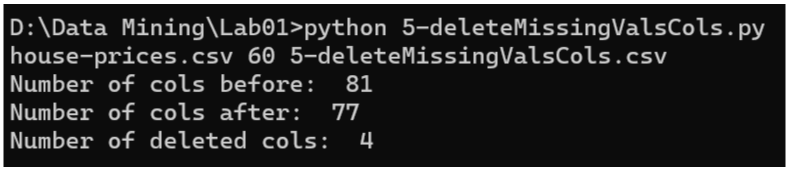


```
D:\Data Mining\Lab01>python 5-deleteMissingValsCols.py
house-prices.csv 50 5-deleteMissingValsCols.csv
Number of cols before:  81
Number of cols after:  75
Number of deleted cols:  6
```

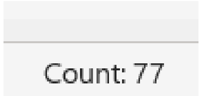


Count: 75

   - Example 2: python 5-deleteMissingValsCols.py house-prices.csv 60 5-deleteMissingValsCols.csv



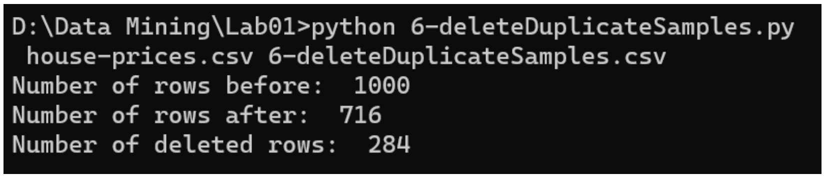
```
D:\Data Mining\Lab01>python 5-deleteMissingValsCols.py
house-prices.csv 60 5-deleteMissingValsCols.csv
Number of cols before:  81
Number of cols after:  77
Number of deleted cols:  4
```



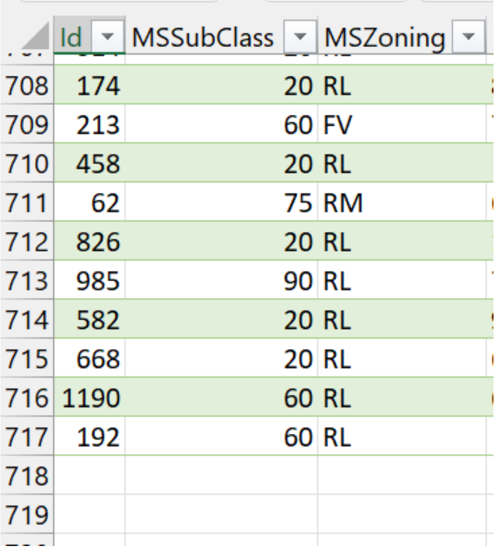
Count: 77

## 6. Delete duplicate samples

- Command line arguments: python 6-deleteDuplicateSamples.py <FileIn> <FileOut>
- Output: number of rows before, number of rows after, Number of deleted rows and write the result in FileOut.
- Example: python 6-deleteDuplicateSamples.py house-prices.csv 6-deleteDuplicateSamples.csv



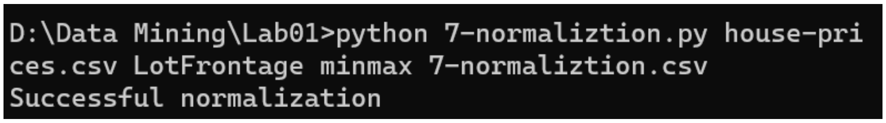
```
D:\Data Mining\Lab01>python 6-deleteDuplicateSamples.py
 house-prices.csv 6-deleteDuplicateSamples.csv
Number of rows before:  1000
Number of rows after:  716
Number of deleted rows:  284
```



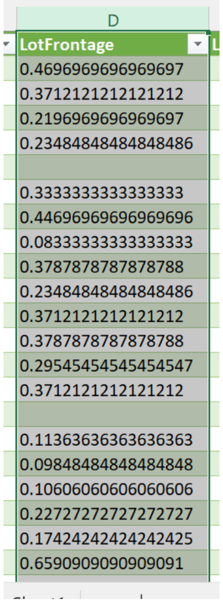
| | Id | MSSubClass | MSZoning |
|---|---|---|---|
| 708 | 174 | 20 | RL |
| 709 | 213 | 60 | FV |
| 710 | 458 | 20 | RL |
| 711 | 62 | 75 | RM |
| 712 | 826 | 20 | RL |
| 713 | 985 | 90 | RL |
| 714 | 582 | 20 | RL |
| 715 | 668 | 20 | RL |
| 716 | 1190 | 60 | RL |
| 717 | 192 | 60 | RL |
| 718 | | | |
| 719 | | | |

## 7. Normalize a numeric attribute using min-max and Z-score methods

- Command line arguments: python 7-normaliztion.py <FileIn> <attribute> <method> <FileOut>
- Output: written to FileOut.
- Example 1: python 7-normaliztion.py house-prices.csv LotFrontage minmax 7-normaliztion.csv

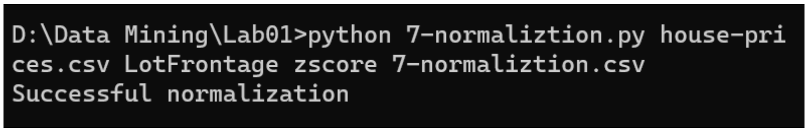


```
D:\Data Mining\Lab01>python 7-normaliztion.py house-pri
ces.csv LotFrontage minmax 7-normaliztion.csv
Successful normalization
```

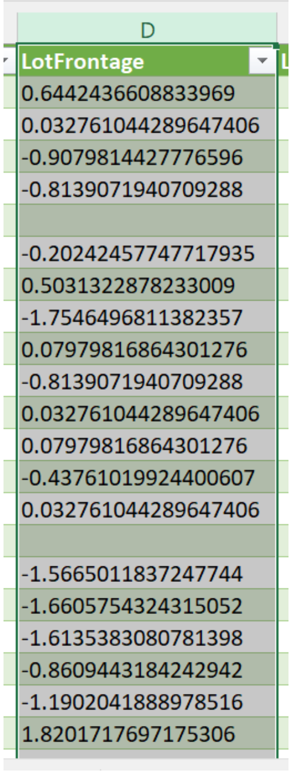


| LotFrontage |
|---|
| 0.4696969696969697 |
| 0.3712121212121212 |
| 0.2196969696969697 |
| 0.23484848484848486 |
| |
| 0.3333333333333333 |
| 0.44696969696969696 |
| 0.08333333333333333 |
| 0.3787878787878788 |
| 0.23484848484848486 |
| 0.3712121212121212 |
| 0.3787878787878788 |
| 0.29545454545454547 |
| 0.3712121212121212 |
| |
| 0.11363636363636363 |
| 0.09848484848484848 |
| 0.10606060606060606 |
| 0.22727272727272727 |
| 0.17424242424242425 |
| 0.6590909090909091 |

- Example 2: python 7-normaliztion.py house-prices.csv LotFrontage zscore 7-normaliztion.csv

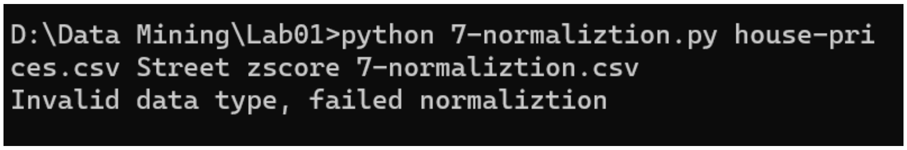


```
D:\Data Mining\Lab01>python 7-normaliztion.py house-pri
ces.csv LotFrontage zscore 7-normaliztion.csv
Successful normalization
```



| D |
|---|
| LotFrontage |
| 0.6442436608833969 |
| 0.032761044289647406 |
| -0.9079814427776596 |
| -0.8139071940709288 |
| |
| -0.20242457747717935 |
| 0.5031322878233009 |
| -1.7546496811382357 |
| 0.07979816864301276 |
| -0.8139071940709288 |
| 0.032761044289647406 |
| 0.07979816864301276 |
| -0.43761019924400607 |
| 0.032761044289647406 |
| |
| -1.5665011837247744 |
| -1.6605754324315052 |
| -1.6135383080781398 |
| -0.8609443184242942 |
| -1.1902041888978516 |
| 1.820171769717530 6 |

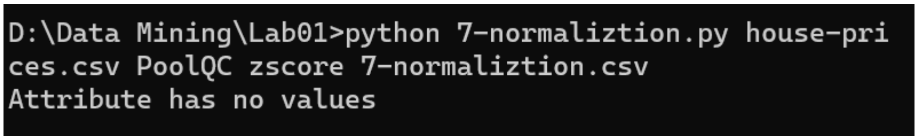
- Example 3: python 7-normaliztion.py house-prices.csv Street zscore 7-normaliztion.csv



```
D:\Data Mining\Lab01>python 7-normaliztion.py house-pri
ces.csv Street zscore 7-normaliztion.csv
Invalid data type, failed normaliztion
```

→Because *Street* is str.

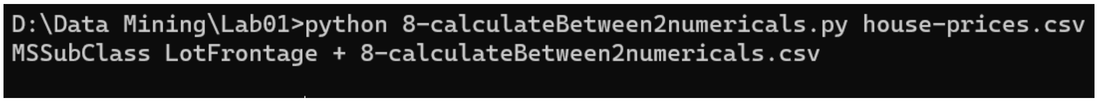
- Example 4: python 7-normaliztion.py house-prices.csv PoolQC zscore 7-normaliztion.csv



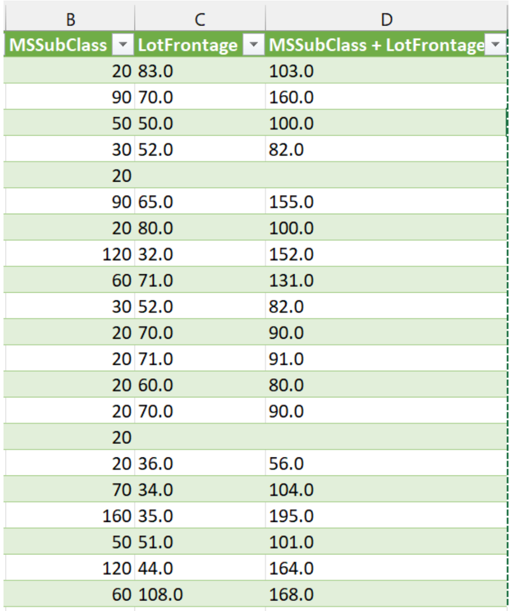
```
D:\Data Mining\Lab01>python 7-normaliztion.py house-pri
ces.csv PoolQC zscore 7-normaliztion.csv
Attribute has no values
```

→ Because *PoolQC* has no values.

8. **Performing addition, subtraction, multiplication, and division between two numerical attributes**
    - Command line arguments: python 8-calculateBetween2numericals.py <FileIn> <attribute1> <attribute2> <method> <FileOut>
    - Output: written to FileOut.
    - Example 1: python 8-calculateBetween2numericals.py house-prices.csv MSSubClass LotFrontage + 8-calculateBetween2numericals.csv

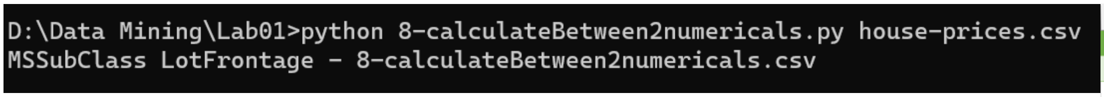


```
D:\Data Mining\Lab01>python 8-calculateBetween2numericals.py house-prices.csv
MSSubClass LotFrontage + 8-calculateBetween2numericals.csv
```

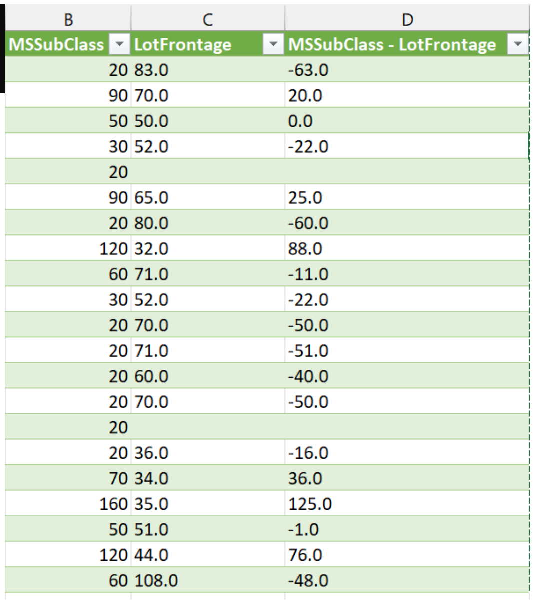


| B | C | D |
|---|---|---|
| MSSubClass | LotFrontage | MSSubClass + LotFrontage |
| 20 | 83.0 | 103.0 |
| 90 | 70.0 | 160.0 |
| 50 | 50.0 | 100.0 |
| 30 | 52.0 | 82.0 |
| 20 | | |
| 90 | 65.0 | 155.0 |
| 20 | 80.0 | 100.0 |
| 120 | 32.0 | 152.0 |
| 60 | 71.0 | 131.0 |
| 30 | 52.0 | 82.0 |
| 20 | 70.0 | 90.0 |
| 20 | 71.0 | 91.0 |
| 20 | 60.0 | 80.0 |
| 20 | 70.0 | 90.0 |
| 20 | | |
| 20 | 36.0 | 56.0 |
| 70 | 34.0 | 104.0 |
| 160 | 35.0 | 195.0 |
| 50 | 51.0 | 101.0 |
| 120 | 44.0 | 164.0 |
| 60 | 108.0 | 168.0 |

- Example 2: python 8-calculateBetween2numericals.py house-prices.csv MSSubClass LotFrontage - 8-calculateBetween2numericals.csv

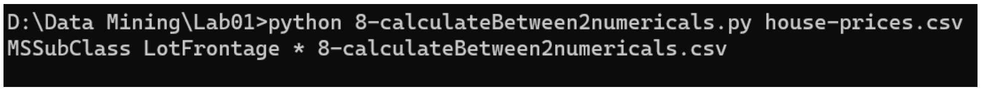


```
D:\Data Mining\Lab01>python 8-calculateBetween2numericals.py house-prices.csv
MSSubClass LotFrontage - 8-calculateBetween2numericals.csv
```

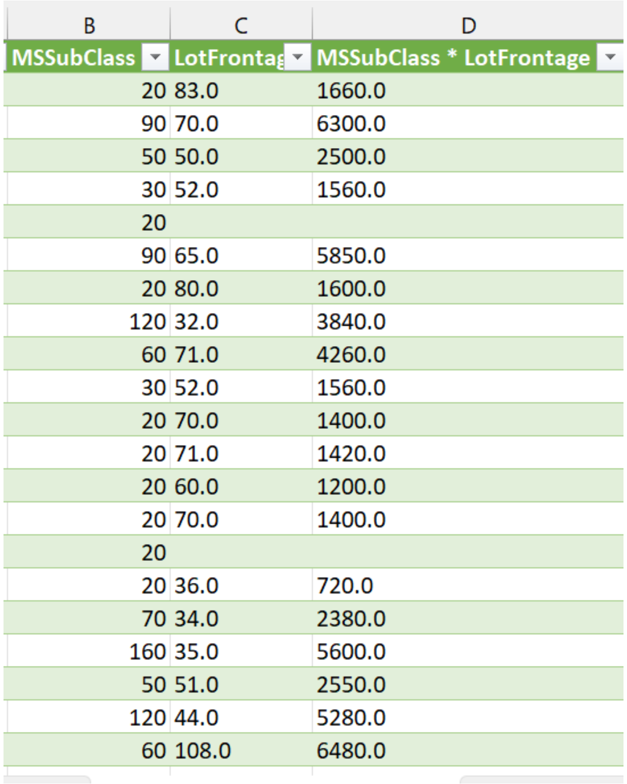


| B | C | D |
|---|---|---|
| MSSubClass | LotFrontage | MSSubClass - LotFrontage |
| 20 | 83.0 | -63.0 |
| 90 | 70.0 | 20.0 |
| 50 | 50.0 | 0.0 |
| 30 | 52.0 | -22.0 |
| 20 | | |
| 90 | 65.0 | 25.0 |
| 20 | 80.0 | -60.0 |
| 120 | 32.0 | 88.0 |
| 60 | 71.0 | -11.0 |
| 30 | 52.0 | -22.0 |
| 20 | 70.0 | -50.0 |
| 20 | 71.0 | -51.0 |
| 20 | 60.0 | -40.0 |
| 20 | 70.0 | -50.0 |
| 20 | | |
| 20 | 36.0 | -16.0 |
| 70 | 34.0 | 36.0 |
| 160 | 35.0 | 125.0 |
| 50 | 51.0 | -1.0 |
| 120 | 44.0 | 76.0 |
| 60 | 108.0 | -48.0 |

- Example 3: python 8-calculateBetween2numericals.py house-prices.csv MSSubClass LotFrontage * 8-calculateBetween2numericals.csv

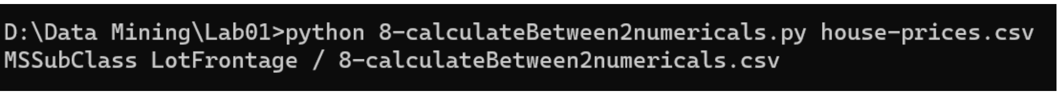


```
D:\Data Mining\Lab01>python 8-calculateBetween2numericals.py house-prices.csv
MSSubClass LotFrontage * 8-calculateBetween2numericals.csv
```

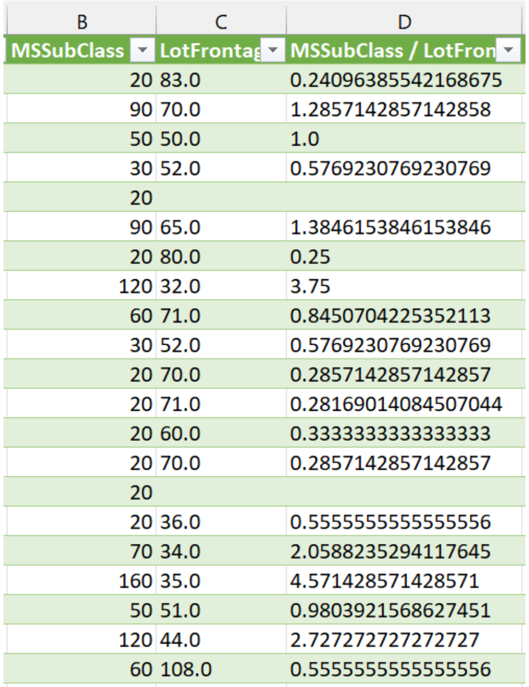


| B | C | D |
|---|---|---|
| MSSubClass | LotFrontag | MSSubClass * LotFrontage |
| 20 | 83.0 | 1660.0 |
| 90 | 70.0 | 6300.0 |
| 50 | 50.0 | 2500.0 |
| 30 | 52.0 | 1560.0 |
| 20 | | |
| 90 | 65.0 | 5850.0 |
| 20 | 80.0 | 1600.0 |
| 120 | 32.0 | 3840.0 |
| 60 | 71.0 | 4260.0 |
| 30 | 52.0 | 1560.0 |
| 20 | 70.0 | 1400.0 |
| 20 | 71.0 | 1420.0 |
| 20 | 60.0 | 1200.0 |
| 20 | 70.0 | 1400.0 |
| 20 | | |
| 20 | 36.0 | 720.0 |
| 70 | 34.0 | 2380.0 |
| 160 | 35.0 | 5600.0 |
| 50 | 51.0 | 2550.0 |
| 120 | 44.0 | 5280.0 |
| 60 | 108.0 | 6480.0 |

- Example 4: python 8-calculateBetween2numericals.py house-prices.csv MSSubClass LotFrontage / 8-calculateBetween2numericals.csv

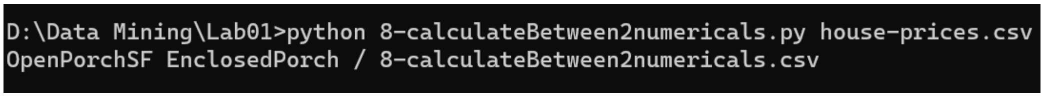


```
D:\Data Mining\Lab01>python 8-calculateBetween2numericals.py house-prices.csv
MSSubClass LotFrontage / 8-calculateBetween2numericals.csv
```

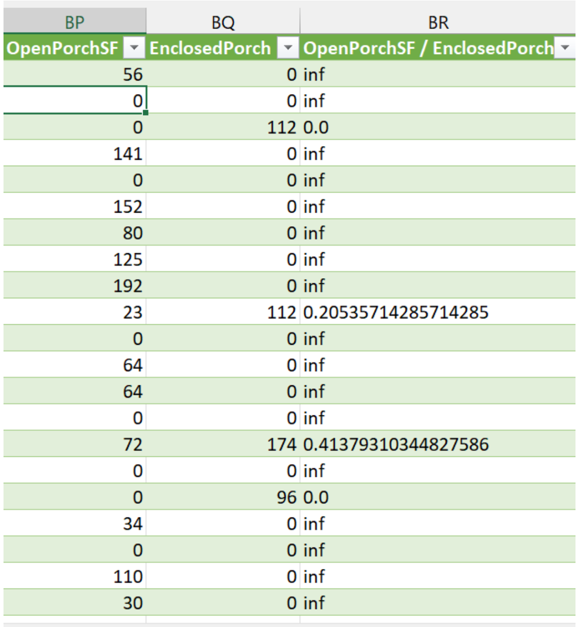


| B | C | D |
|---|---|---|
| MSSubClass | LotFrontag | MSSubClass / LotFron |
| 20 | 83.0 | 0.24096385542168675 |
| 90 | 70.0 | 1.2857142857142858 |
| 50 | 50.0 | 1.0 |
| 30 | 52.0 | 0.5769230769230769 |
| 20 | | |
| 90 | 65.0 | 1.3846153846153846 |
| 20 | 80.0 | 0.25 |
| 120 | 32.0 | 3.75 |
| 60 | 71.0 | 0.8450704225352113 |
| 30 | 52.0 | 0.5769230769230769 |
| 20 | 70.0 | 0.2857142857142857 |
| 20 | 71.0 | 0.28169014084507044 |
| 20 | 60.0 | 0.3333333333333333 |
| 20 | 70.0 | 0.2857142857142857 |
| 20 | | |
| 20 | 36.0 | 0.5555555555555556 |
| 70 | 34.0 | 2.0588235294117645 |
| 160 | 35.0 | 4.571428571428571 |
| 50 | 51.0 | 0.9803921568627451 |
| 120 | 44.0 | 2.7272727272727 |
| 60 | 108.0 | 0.5555555555555556 |

- Example 5: python 8-calculateBetween2numericals.py house-prices.csv OpenPorchSF EnclosedPorch / 8-calculateBetween2numericals.csv



```
D:\Data Mining\Lab01>python 8-calculateBetween2numericals.py house-prices.csv
OpenPorchSF EnclosedPorch / 8-calculateBetween2numericals.csv
```



| BP | BQ | BR |
|---|---|---|
| OpenPorchSF | EnclosedPorch | OpenPorchSF / EnclosedPorch |
| 56 | 0 | inf |
| 0 | 0 | inf |
| 0 | 112 | 0.0 |
| 141 | 0 | inf |
| 0 | 0 | inf |
| 152 | 0 | inf |
| 80 | 0 | inf |
| 125 | 0 | inf |
| 192 | 0 | inf |
| 23 | 112 | 0.20535714285714285 |
| 0 | 0 | inf |
| 64 | 0 | inf |
| 64 | 0 | inf |
| 0 | 0 | inf |
| 72 | 174 | 0.41379310344827586 |
| 0 | 0 | inf |
| 0 | 96 | 0.0 |
| 34 | 0 | inf |
| 0 | 0 | inf |
| 110 | 0 | inf |
| 30 | 0 | inf |

## V. References

- Slides on Moodles of Professor. Le Hoai Bac
- Weka description of Lecturer. Nguyen Thi Thu Hang
- https://www.youtube.com/watch?v=m7kpIBGEdkI&t=2s

## THANK YOU TEACHER FOR READING OUR REPORT