Information of the group

1. Link github: https://github.com/Quanhcmus/Lab1_data_mining

2. Group member information

ID	FULL NAME	EMAIL
20120554	Nguyễn Minh Quân	20120554@student.hcmus.edu.vn
20120587	Nguyễn Hoàng Thịnh	20120587@student.hcmus.edu.vn

3. The contribution rate of each member

   - 20120554:
     - 3.1: Install WEKA (Requirement 1) (100%)
     - 3.2.1: Exploring Breast Cancer data set (100%)
     - 3.2.2: Exploring Weather data set (100%)
     - 3.3.5: Deleting columns containing more than a particular number of missing values (100%)
     - 3.3.6: Delete duplicate samples. (100%)
     - 3.3.7: Normalize a numeric attribute using min-max and Z-score methods. (100%)
     - 3.3.8: Performing addition, subtraction, multiplication, and division between two numerical attributes (100%)
   - 20120587:
     - 3.1 : Install WEKA (Requirement 1+2) (100%)
     - 3.2.3: Exploring Credit in Germany data set (0%)
     - 3.3.1: Extract columns with missing values (0%)
     - 3.3.2: Count the number of lines with missing data (0%)
     - 3.3.3: Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute). (0%)
     - 3.3.4: Deleting rows containing more than a particular number of missing values (0%)

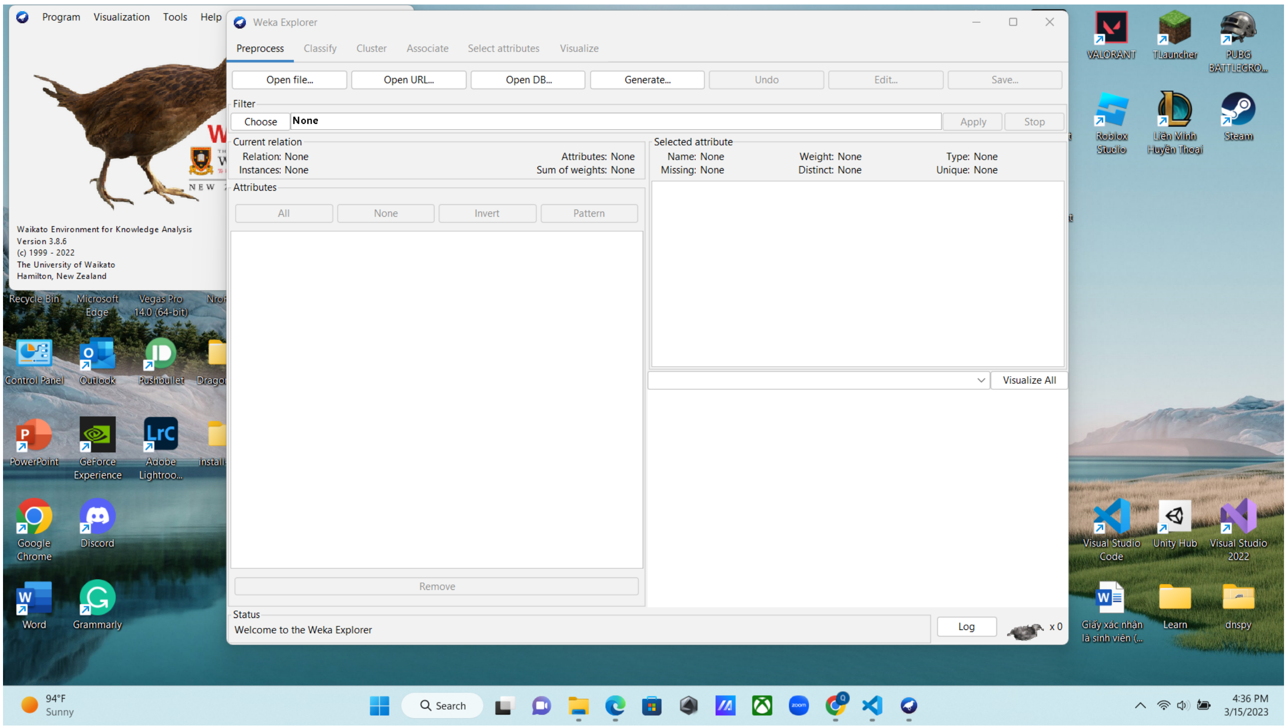
   Total task completed: 100%

Preprocessing and data mining

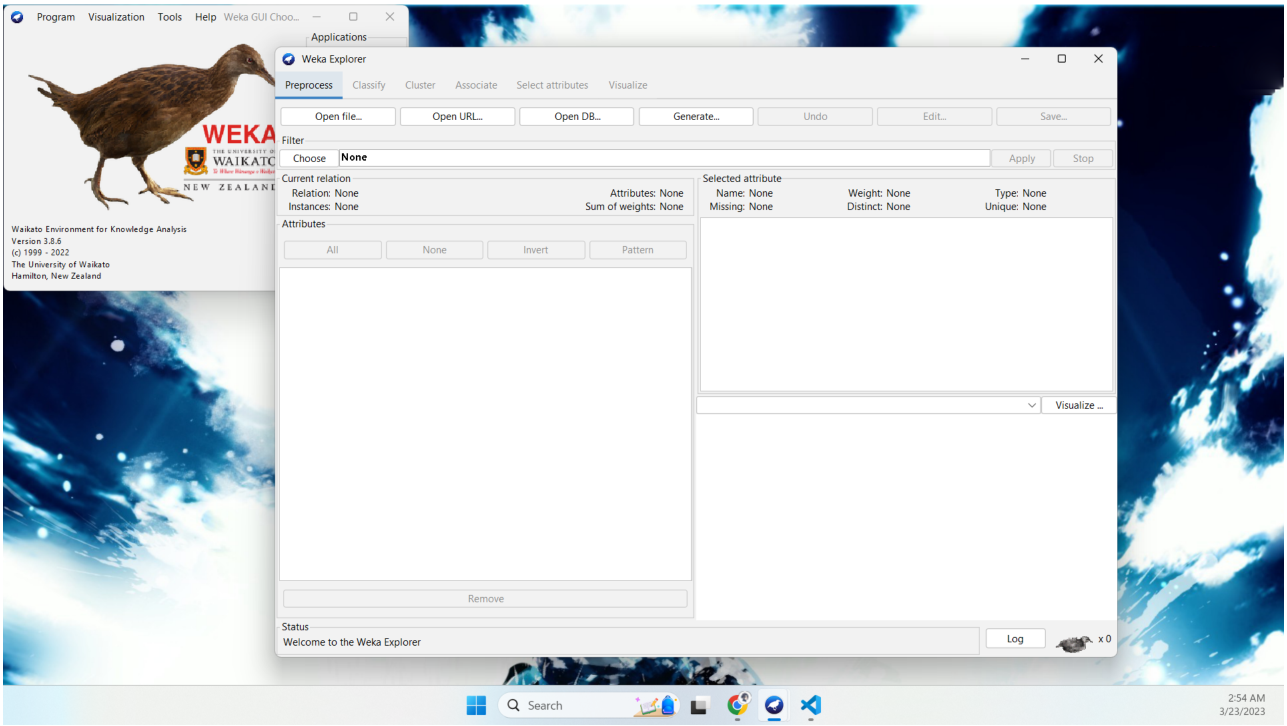
Install WEKA (0.5 points)

Requirement 1: Capturing a screen.

- 20120554



- 20120587



Requirement 2: Explaining the meaning.

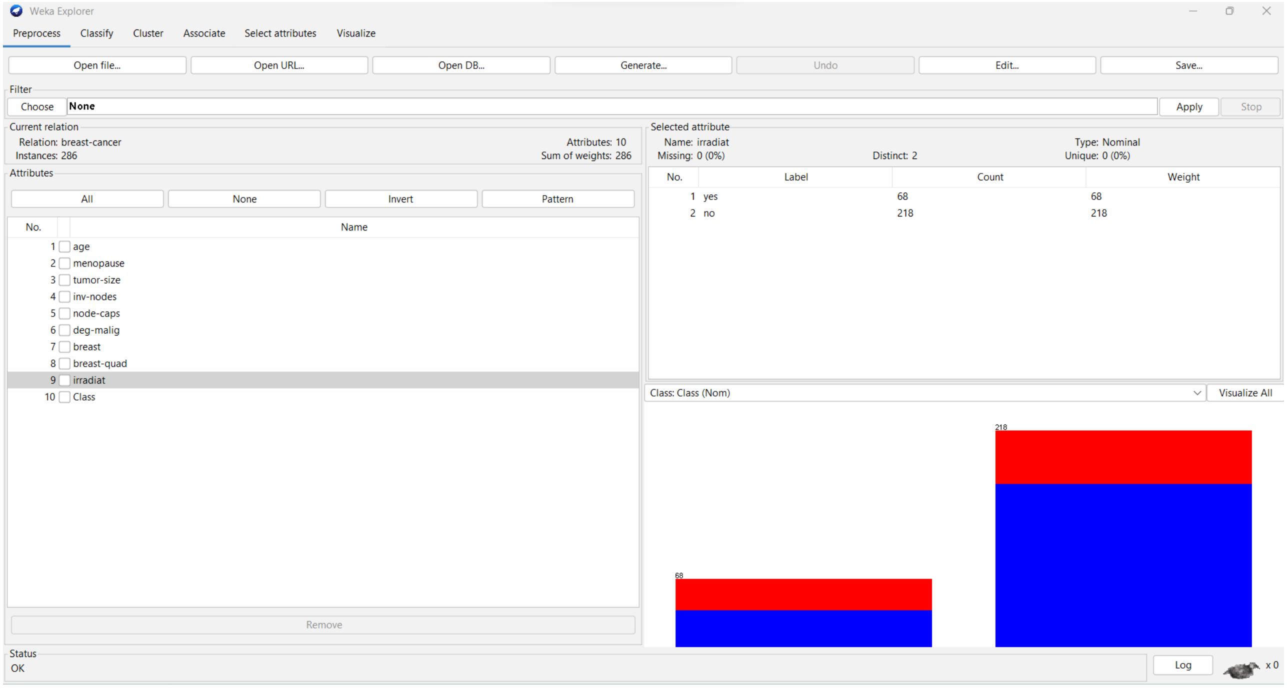
- **Preprocess** tag:
  - Current Relation:
    - Relation: Refers to the name of the current dataset being processed.
    - Attributes: The number of attributes.
    - Instances: The number of rows.
    - Sum of weights: The total weight of instances.
  - Attributes: Specify a subset of attributes that should be used for subsequent processing.
    - All: Selects all attributes.
    - None: Unselects all attributes.

        - Invert: Inverts the current attribute selection.
        - Pattern: Selects all attributes that match a reg. expression.
    - Selected Attributes: Specify a single attribute that should be used for subsequent processing and have some information like Name, Type, Missing, Distinct, Unique.
- **Classify** tag: Uses machine learning algorithm to predict the class label of a data instance based on its input attributes.
- **Cluster** tag: Uses machine learning algorithm to group data instances into clusters based on their similarity.
- **Associate** tag: Uses machine learning algorithm to discover association rules from data.
- **Select attributes** tag: Specify the set of attributes to be used for a particular task, such as classification or clustering. This tag is typically used in conjunction with other tags, such as the Classifier or Cluster tags.
- **Visualize** tag: Specify a visualization method for the results of an analysis and help users understand the patterns and relationships in the data

Getting Acquainted with WEKA (4.5 points)

Exploring Breast Cancer data set

- Load the data file **breast cancer.arff**



Filter: None

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

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

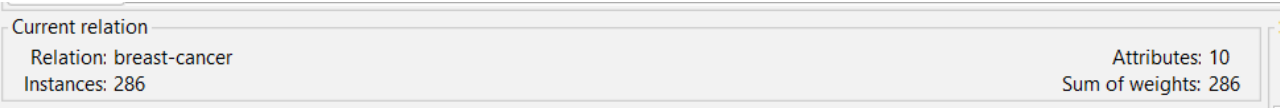
Selected attribute
Name: irradiat Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	68	68
2	no	218	218

Class: Class (Nom)

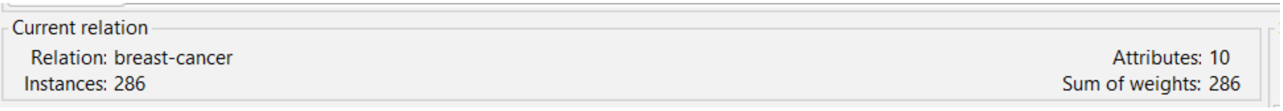
Status
OK

- **How many instances does this data set have?** There are 286 instances in this dataset



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

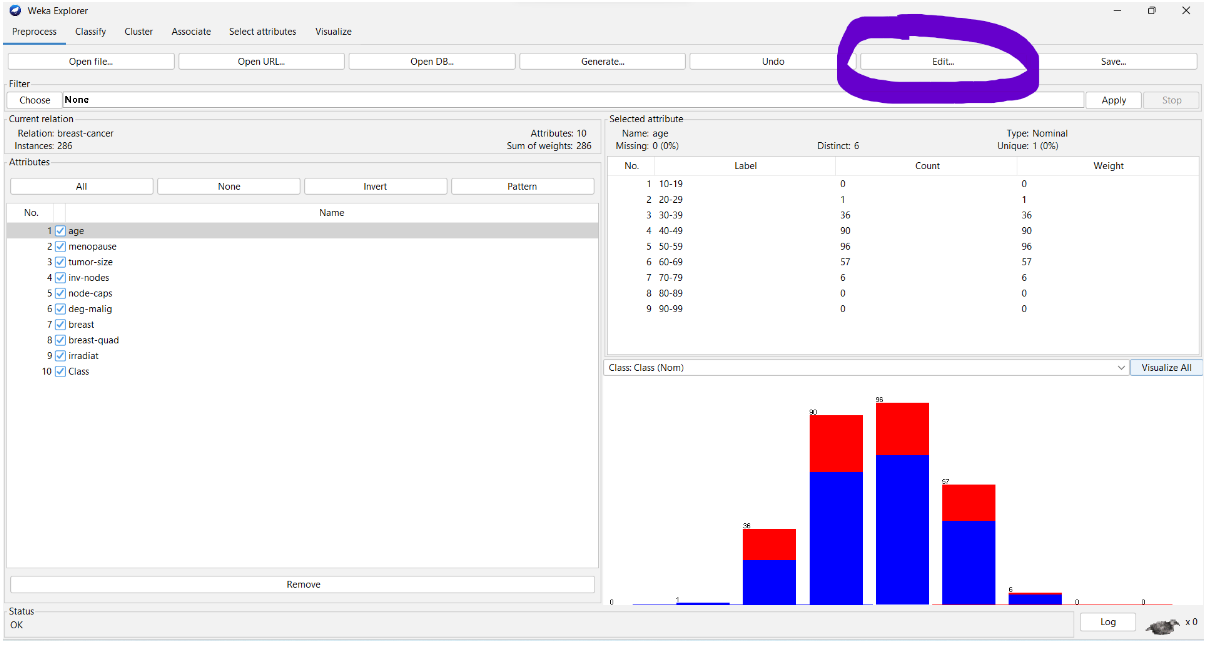
- **How many attributes does this data set have?** There are 10 attributes in this dataset



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

- **Which attribute is used for the label? Can it be changed? How?** *Class* is attributes used for the label, we can change by following these step:

- Click *edit* button



Filter: None

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

Attributes

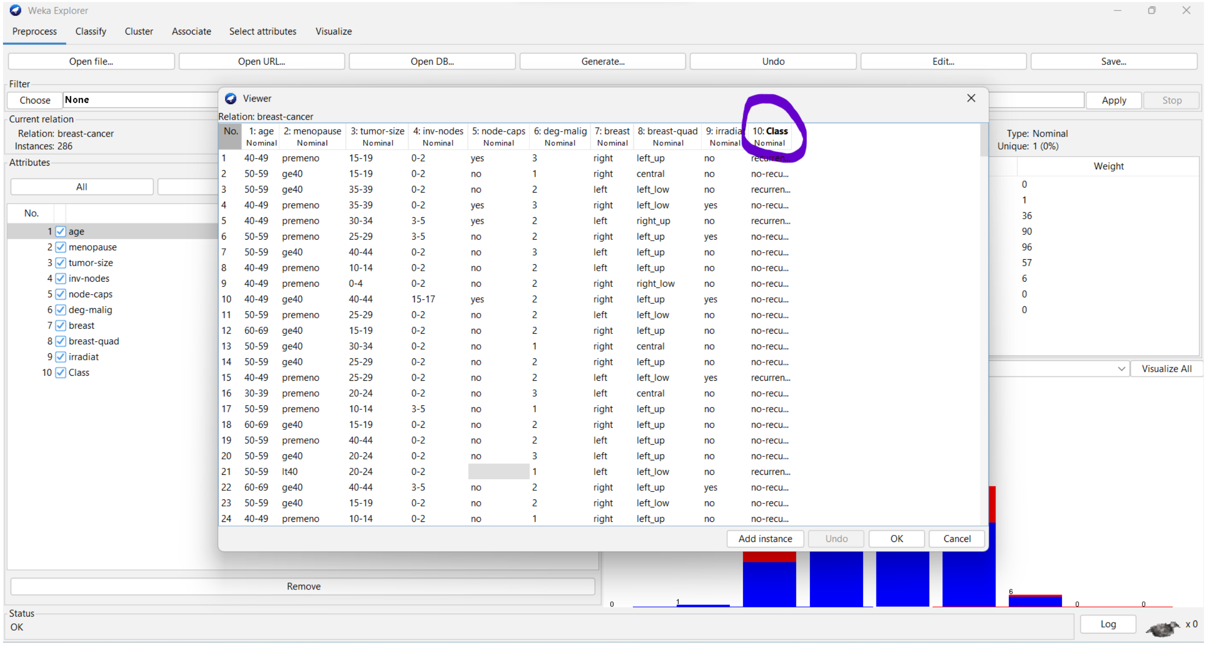
No.	Name
1	☑ age
2	☑ menopause
3	☑ tumor-size
4	☑ inv-nodes
5	☑ node-caps
6	☑ deg-malig
7	☑ breast
8	☑ breast-quad
9	☑ irradiat
10	☑ Class

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
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

Status
OK

- Choose the label *class*



Viewer
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	recurren...
2	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recu...
3	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurren...
4	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recu...
5	40-49	premeno	30-34	3-5	yes	2	left	right_up	no	recurren...
6	50-59	premeno	25-29	3-5	no	2	right	left_up	yes	no-recu...
7	50-59	ge40	40-44	0-2	no	3	left	left_up	no	no-recu...
8	40-49	premeno	10-14	0-2	no	2	left	left_up	no	no-recu...
9	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-recu...
10	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-recu...
11	50-59	premeno	25-29	0-2	no	2	left	left_low	no	no-recu...
12	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recu...
13	50-59	ge40	30-34	0-2	no	1	right	central	no	no-recu...
14	50-59	ge40	25-29	0-2	no	2	right	left_up	no	no-recu...
15	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurren...
16	30-39	premeno	20-24	0-2	no	3	left	central	no	no-recu...
17	50-59	premeno	10-14	3-5	no	1	right	left_up	no	no-recu...
18	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recu...
19	50-59	premeno	40-44	0-2	no	2	left	left_up	no	no-recu...
20	50-59	ge40	20-24	0-2	no	3	left	left_up	no	no-recu...
21	50-59	lt40	20-24	0-2		1	left	left_low	no	recurren...
22	60-69	ge40	40-44	3-5	no	2	right	left_up	yes	no-recu...
23	50-59	ge40	15-19	0-2	no	2	right	left_low	no	no-recu...
24	40-49	premeno	10-14	0-2	no	1	right	left_up	no	no-recu...

Filter: None

Current relation
Relation: breast-cancer
Instances: 286

Attributes

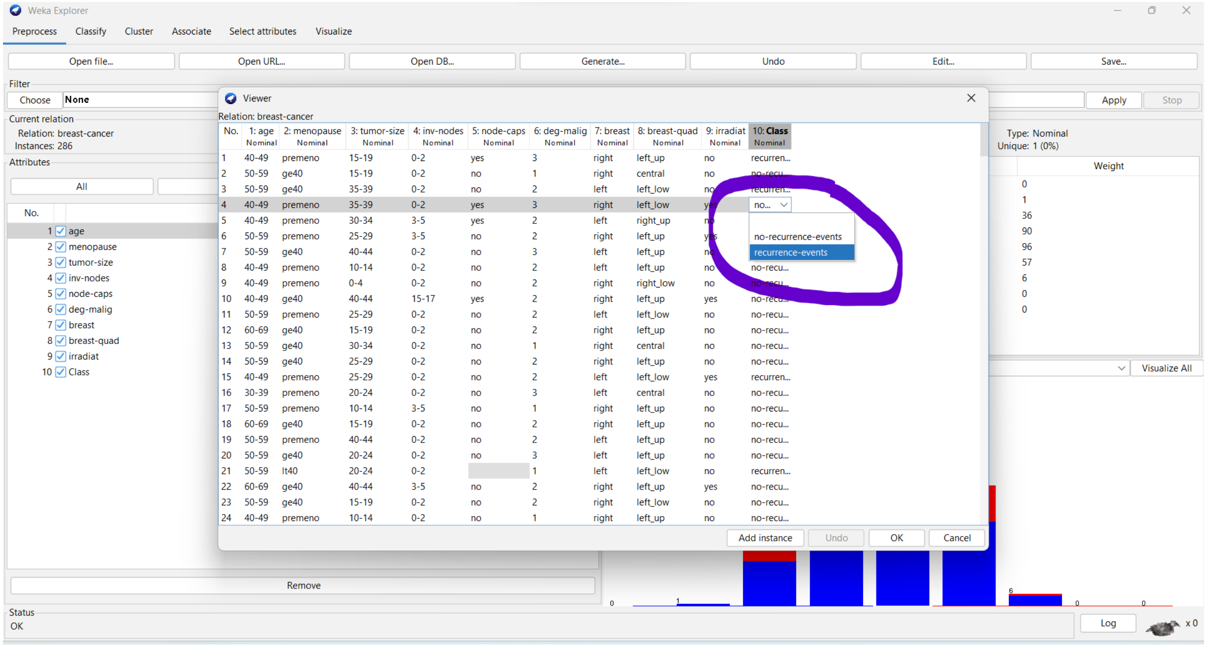
No.	Name
1	☑ age
2	☑ menopause
3	☑ tumor-size
4	☑ inv-nodes
5	☑ node-caps
6	☑ deg-malig
7	☑ breast
8	☑ breast-quad
9	☑ irradiat
10	☑ Class

Type: Nominal
Unique: 1 (0%)

Weight
0
1
36
90
96
57
6
0
0

Status
OK

- Select the cell which you want change

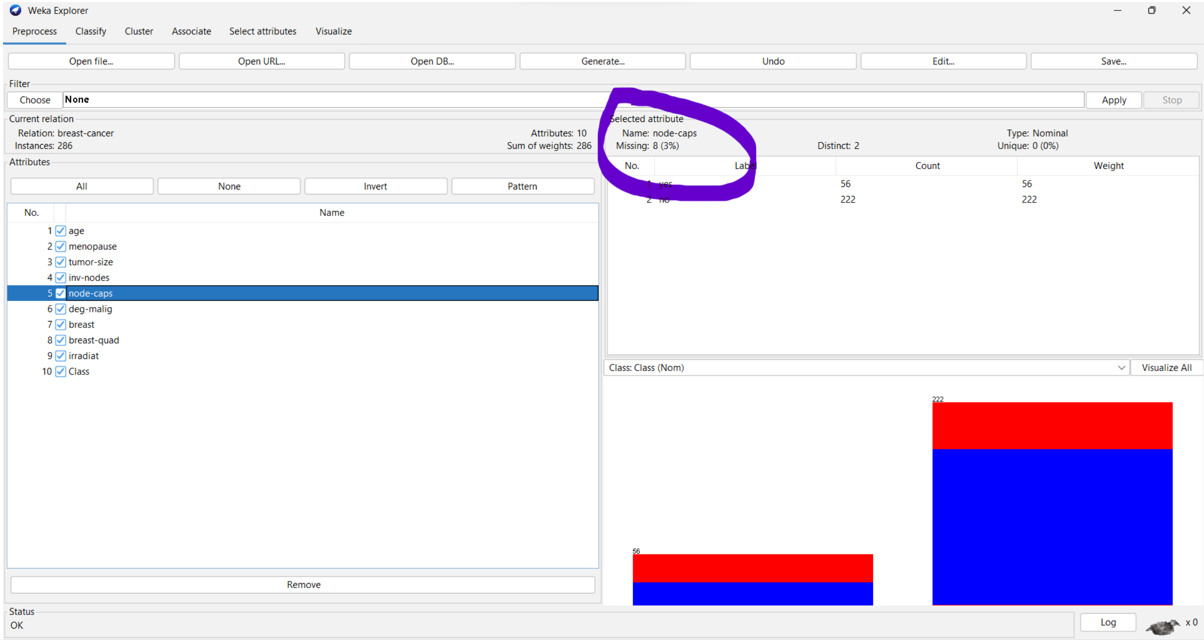


Relation: breast-cancer

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	recurren...
2	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recu...
3	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurren...
4	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no...
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	no-recu...
9	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-recu...
10	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-recu...
11	50-59	premeno	25-29	0-2	no	2	left	left_low	no	no-recu...
12	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recu...
13	50-59	ge40	30-34	0-2	no	1	right	central	no	no-recu...
14	50-59	ge40	25-29	0-2	no	2	right	left_up	no	no-recu...
15	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurren...
16	30-39	premeno	20-24	0-2	no	3	left	central	no	no-recu...
17	50-59	premeno	10-14	3-5	no	1	right	left_up	no	no-recu...
18	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recu...
19	50-59	premeno	40-44	0-2	no	2	left	left_up	no	no-recu...
20	50-59	ge40	20-24	0-2	no	3	left	left_up	no	no-recu...
21	50-59	lt40	20-24	0-2		1	left	left_low	no	recurren...
22	60-69	ge40	40-44	3-5	no	2	right	left_up	yes	no-recu...
23	50-59	ge40	15-19	0-2	no	2	right	left_low	no	no-recu...
24	40-49	premeno	10-14	0-2	no	1	right	left_up	no	no-recu...

- **What is the meaning of each attribute?**
  - **Age**: Patient's age include (10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.)
  - **Menopause**: when your periods stop due to lower hormone levels. (lt40, ge40, premeno.)
  - **Tumor-size**: often measured in centimeters (cm) or inches. (0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.)
  - **Inv-nodes**: the number (range 0 - 39) of axillary lymph nodes that contain metastatic breast cancer visible on histological examination. (02,3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.)
  - **Node-caps**: if the cancer does metastasise to a. lymph node, although outside the original site of. the tumor it may remain “contained” by the cap- sule of the lymph node. (yes, no.)
  - **Deg-malig**: the degree of malignancy of the tumor, which is also known as the tumor grade. (1, 2, 3.)
  - **Breast**: the breast location where the tumor was found. (left, right.)
  - **Breast.quad**: the quadrant of the breast where the tumor was found. (left-up, left-low, right-up, right-low, central.)
  - **Irradiat**: whether or not the patient received radiation therapy as part of their treatment for breast cancer. (yes, no.)
  - **Class**: indicates whether or not a patient experienced a recurrence of breast cancer after their initial treatment. (yes, no.)
- **Let’s investigate the missing value status in each attribute and describe in general ways to solve the problem of missing values.**

- **Node-caps**: 8 missing values



Filter: None

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

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

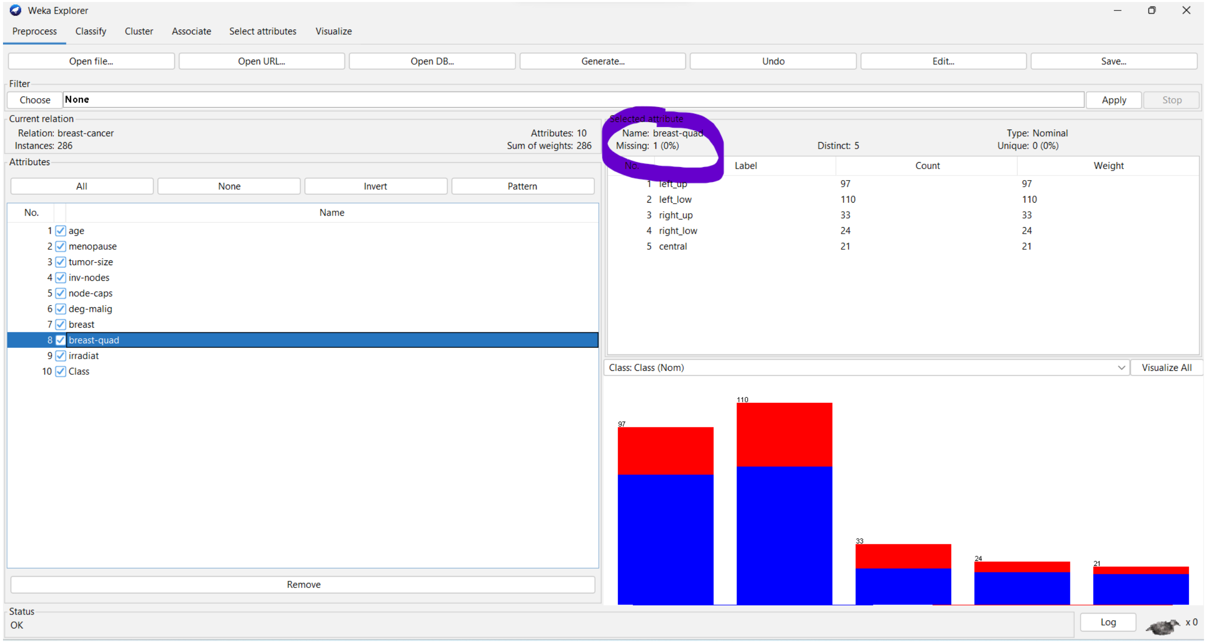
Attributes

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

Class: Class (Nom)

Status
OK

- **Breast.quad**: 1 missing values



Filter: None

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

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

Attributes

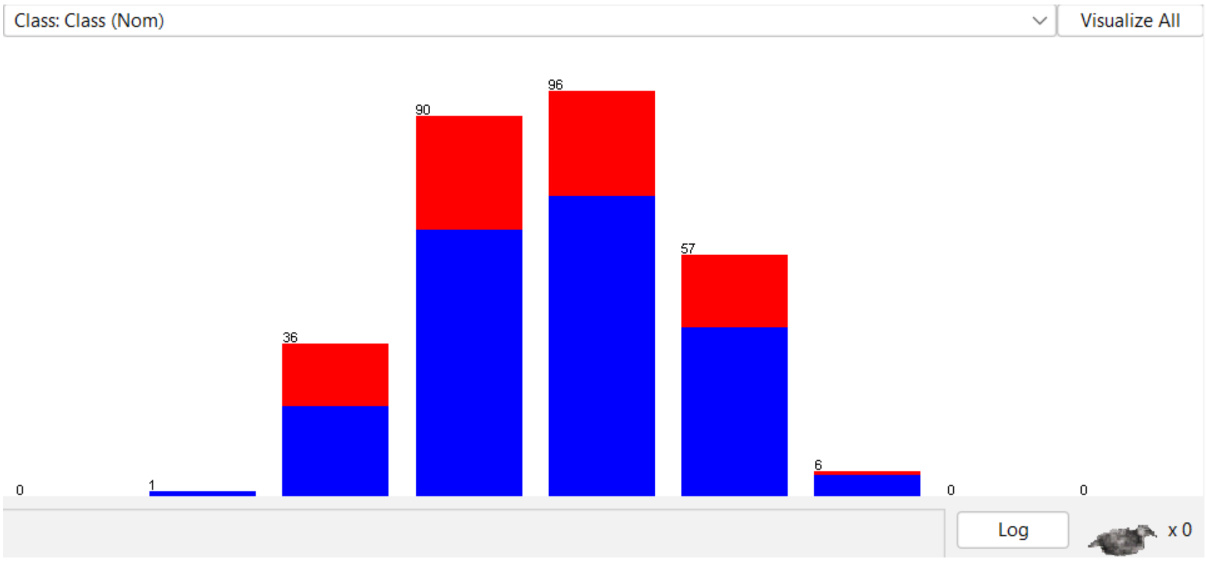
No.	Name
1	☑ age
2	☑ menopause
3	☑ tumor-size
4	☑ inv-nodes
5	☑ node-caps
6	☑ deg-malig
7	☑ breast
8	☑ breast-quad
9	☑ irradiat
10	☑ Class

Class: Class (Nom)

Status
OK

- we can handle missing values by replaced with the property's mean

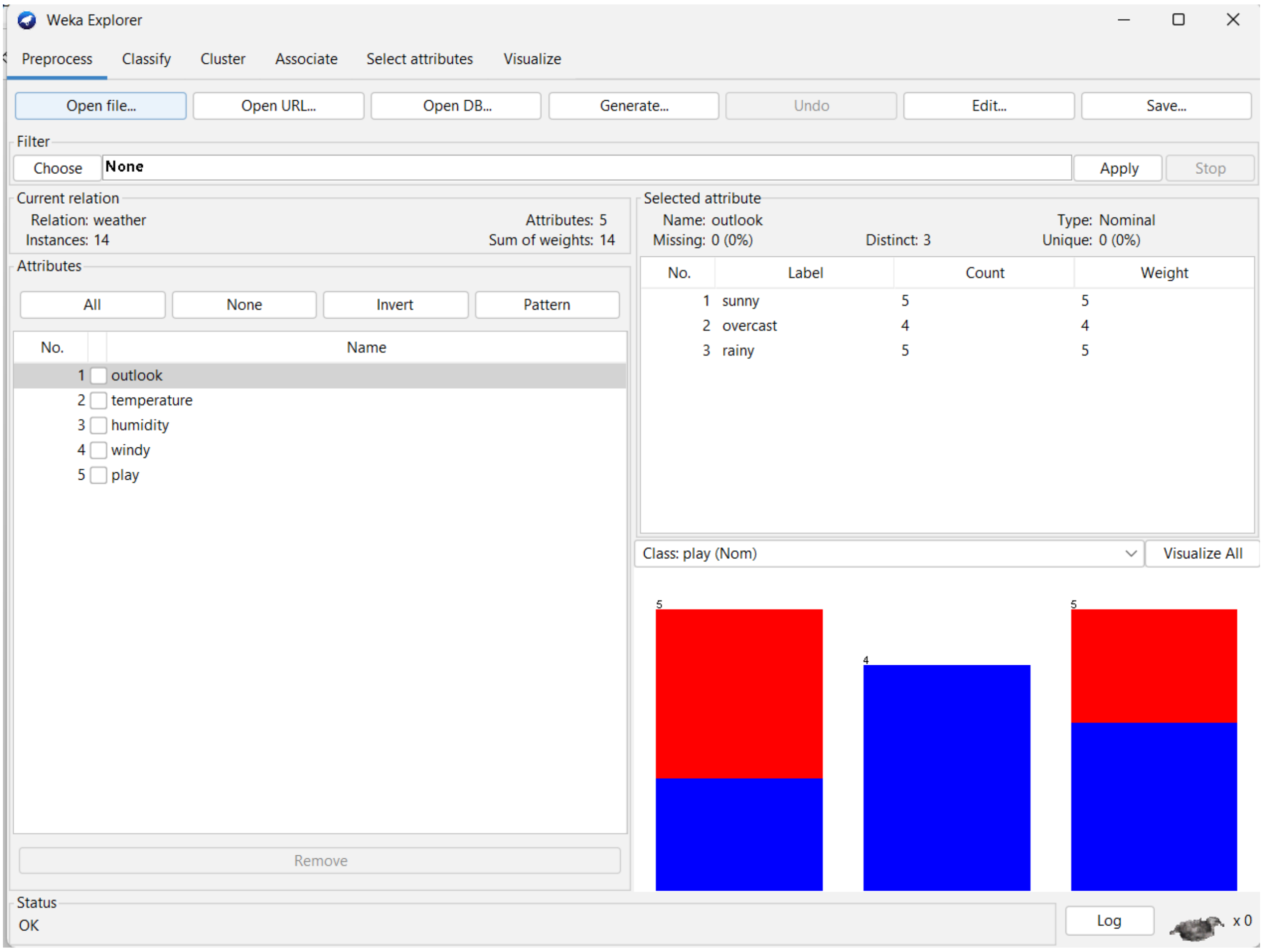
- **Let's propose solutions to the problem of missing values in the specific attribute.**
  - In **Node-caps** attributes we can replace with the most likely value infer from a Bayesian formula, decision tree or EM algorithm
  - In **Breast.quad** attributes we can replace with the property's mean
- **Let's explain the meaning of the chart in the WEKA Explorer. Setting the title for it and describing its legend.**

- 

- We can setting title chart is stacked bar chart
- Red represents the patients recurrence-events
- Blue represents the patients no-recurrence-events

Exploring Weather data set

- Load the data file **weather.numeric.arff**



Filter: None

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

Attributes

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

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

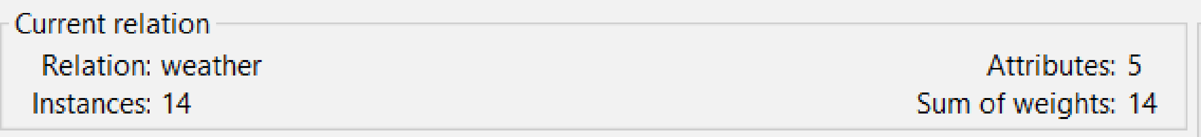
No.	Label	Count	Weight
1	sunny	5	5
2	overcast	4	4
3	rainy	5	5

Class: play (Nom)

Status
OK

- **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?**

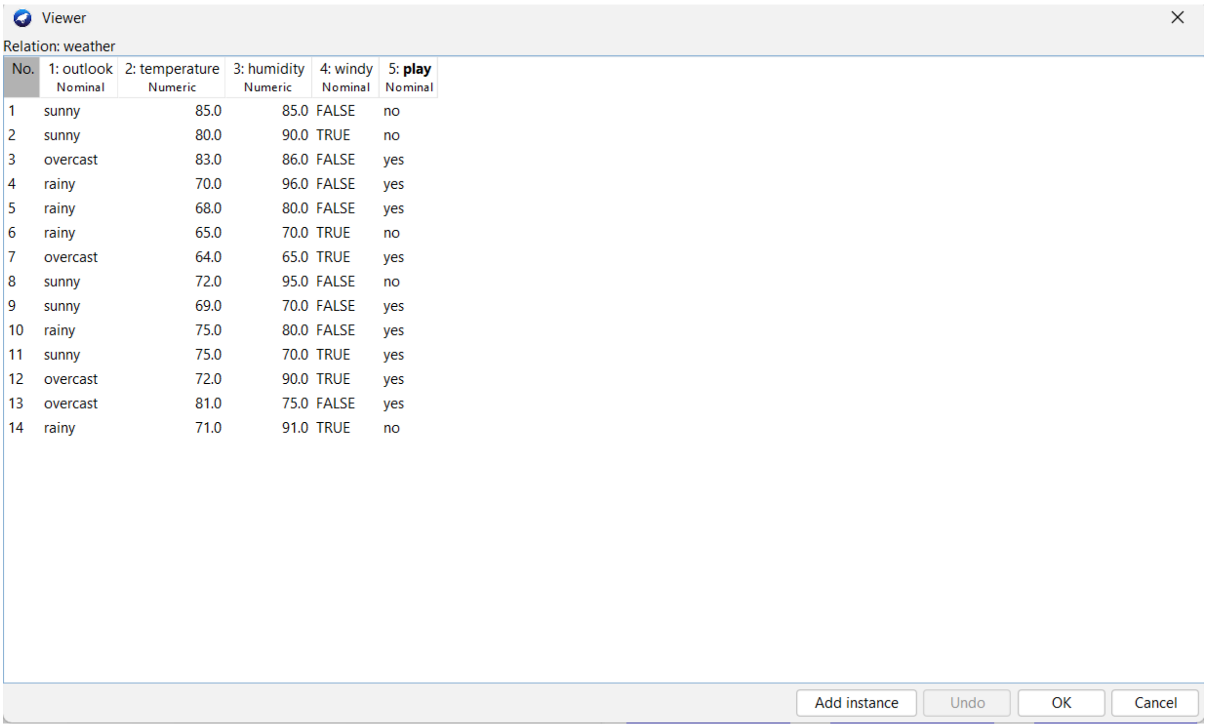
- There are **5** attributes, **14** samples in this data set



Current relation

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

- Attributes have data type categorical is **outlook**, **windy** and **play**
- Attributes have data type numerical is **temperature** and **humudity**
- Attributes used for the label is **play**



Relation: weather

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

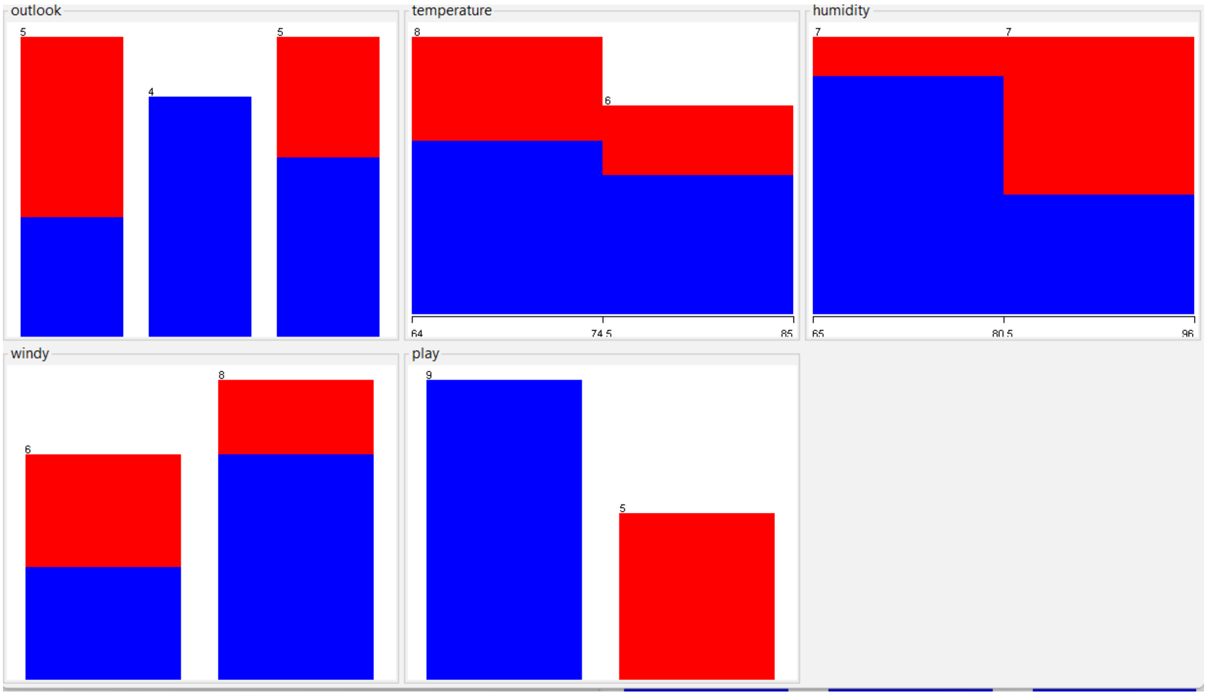
- **Let’s list five-number summary of two attributes *temperature* and *humidity*. Does WEKA provide these values?**

	Min	Q1	Median	Q3	Max
temperature	64	69.25	72	78.75	85
humidity	65	71.25	82.5	90	96

  - WEKA don't provide these values

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

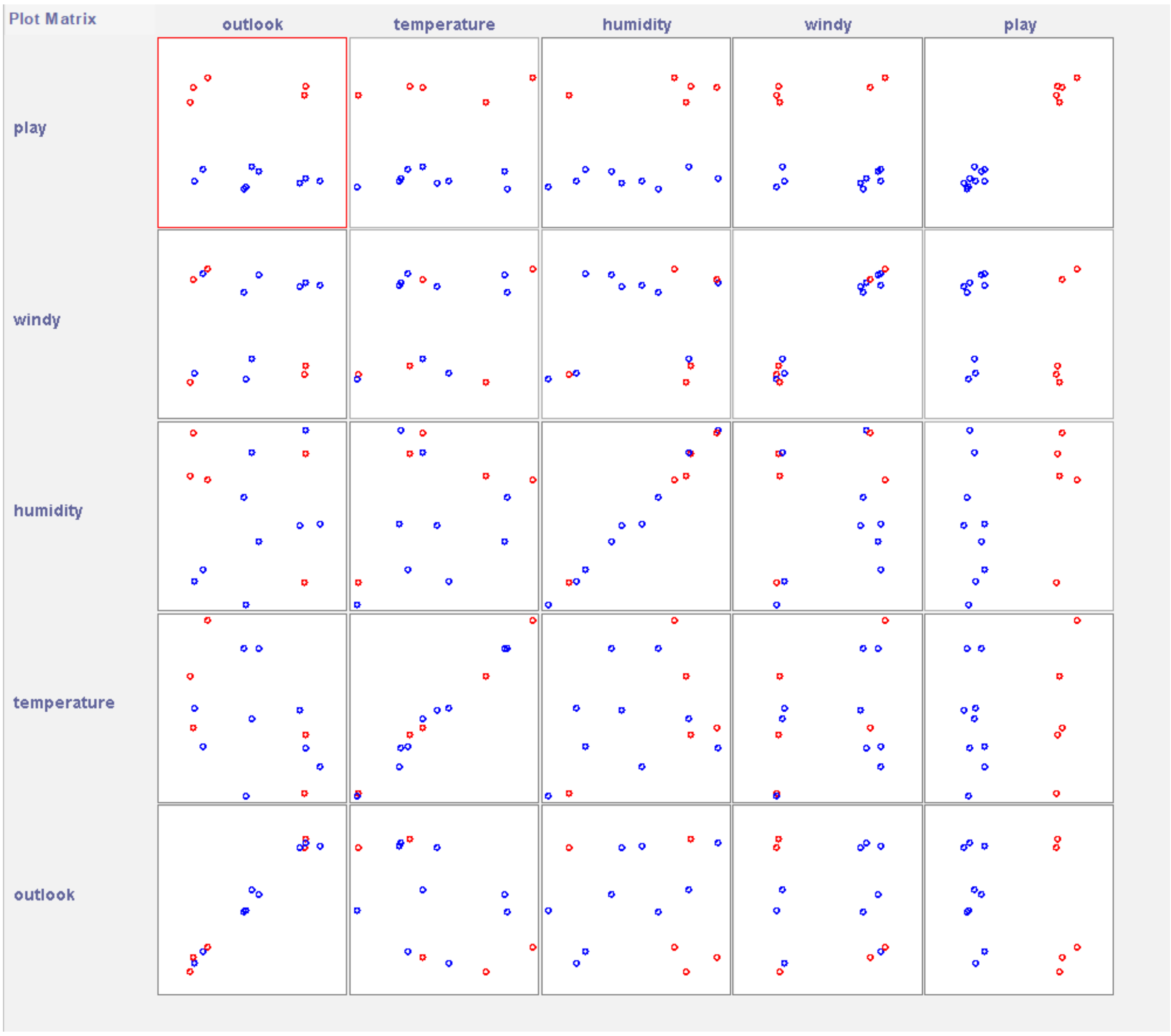
- All chart



- The graph shows the data distribution of the attributeThe graph shows the data distribution of the attribute
- The title for it could be "Distribution chart"
- Blue is no
- Red is yes

- **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?**

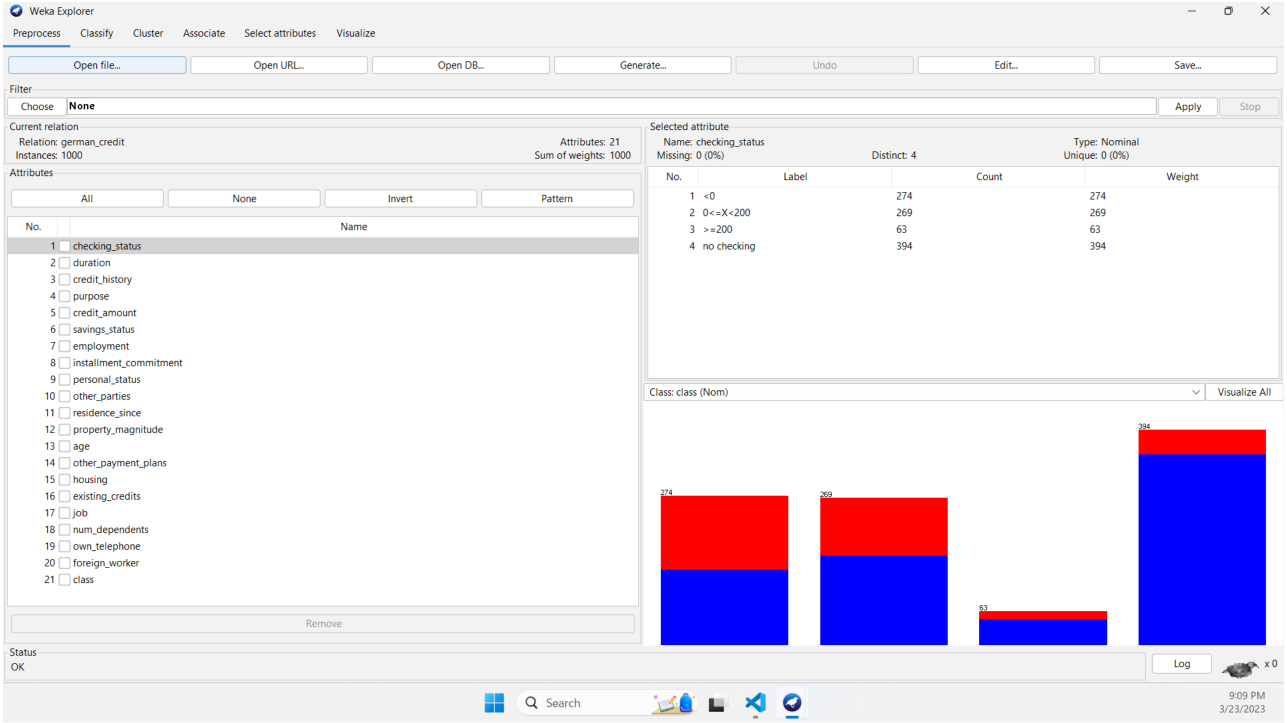
- Chart in Visualize tag



- The name of this chart is scatter plot of attributes
- We think humidity and play are correlated

Exploring Credit in Germany data set

- Load the data file **credit-g.arf**



Current relation
Relation: german_credit
Instances: 1000
Attributes: 21
Sum of weights: 1000

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

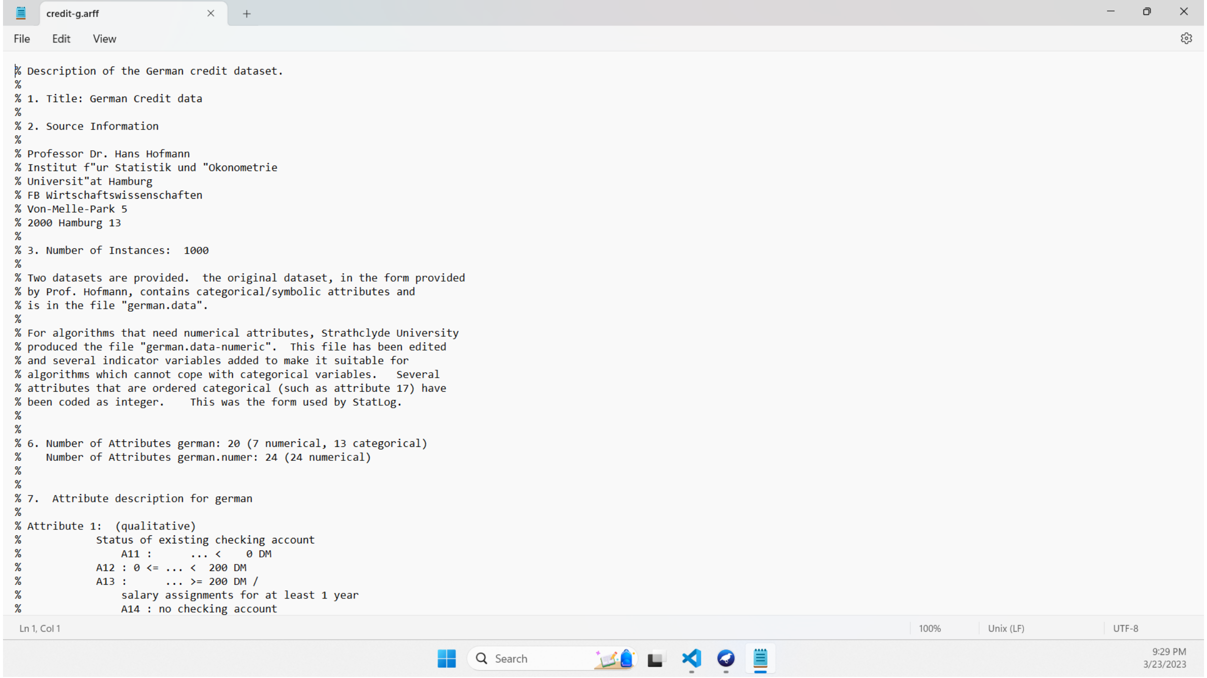
No.	Label	Count	Weight
1	<0	274	274
2	0<=X<200	269	269
3	>=200	63	63
4	no checking	394	394

Attributes

No.	Name
1	checking_status
2	duration
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

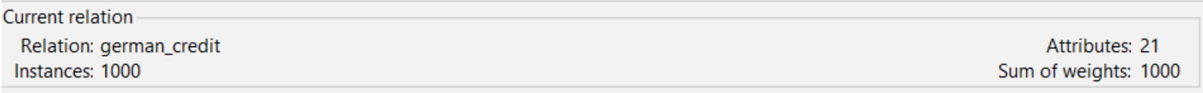
Status
OK

- **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 fiveattributes (must have both discrete and continuous attributes).**
  - The content of the comments section (when opened with any text editor): Description of the German credit dataset. Included: Title, Source Information, Number of Instances,etc.



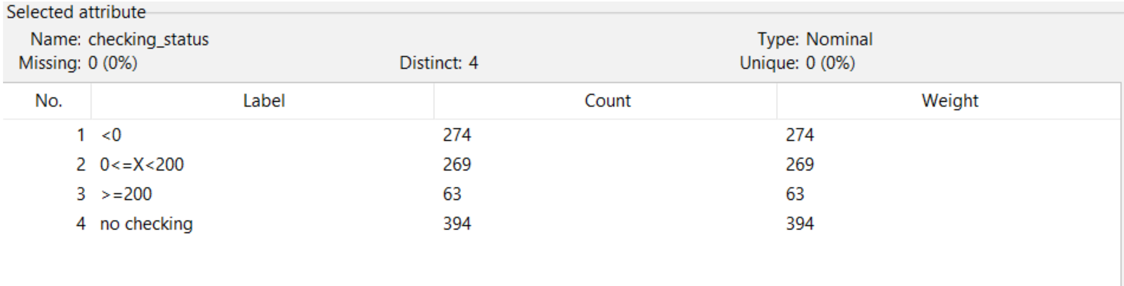
```
% Description of the German credit dataset.
%
% 1. Title: German Credit data
%
% 2. Source Information
%
% Professor Dr. Hans Hofmann
% Institut f"ur Statistik und "Okonometrie
% Universit"at Hamburg
% FB Wirtschaftswissenschaften
% Von-Melle-Park 5
% 2000 Hamburg 13
%
% 3. Number of Instances:  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
%            A12 : 0 <= ... <  200 DM
%            A13 :      ... >= 200 DM /
%                salary assignments for at least 1 year
%                A14 : no checking account
```

  - The data set have 1000 samples
  - The data set have 21 attributes



Current relation
Relation: german_credit
Instances: 1000
Attributes: 21
Sum of weights: 1000

  - Describe any five attributes:
    - checking_status (Discrete attribute): Status of existing checking account. Missing: 0%, Distinct: 4, Type: Nominal, Unique: 0%.

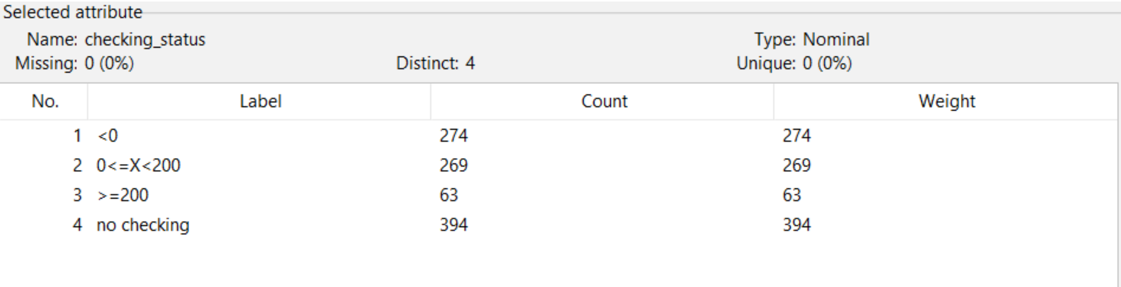


Selected attribute

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

- duration (Continuous attribute): Duration in month. Missing: 0%, Distinct: 33, Type: Numeric, Unique: 5(1%).

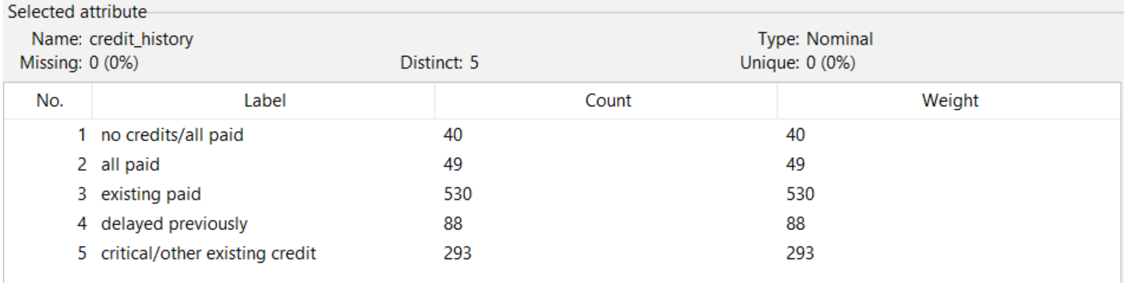


Selected attribute

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

- credit_history (Discrete attribute): Credit history. Missing: 0%, Distinct: 5, Type: Nominal, Unique: 0%.

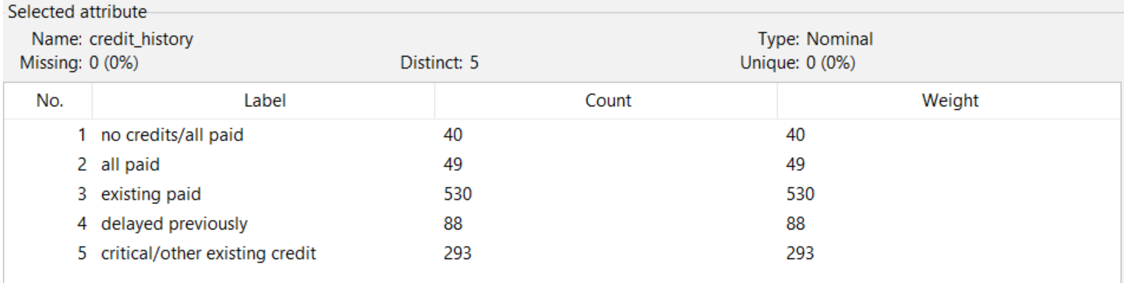


Selected attribute

Name: credit_history Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	no credits/all paid	40	40
2	all paid	49	49
3	existing paid	530	530
4	delayed previously	88	88
5	critical/other existing credit	293	293

- purpose (Discrete attribute): Purpose. Missing: 0%, Distince: 10, Type: Nominal, Unique: 0%.

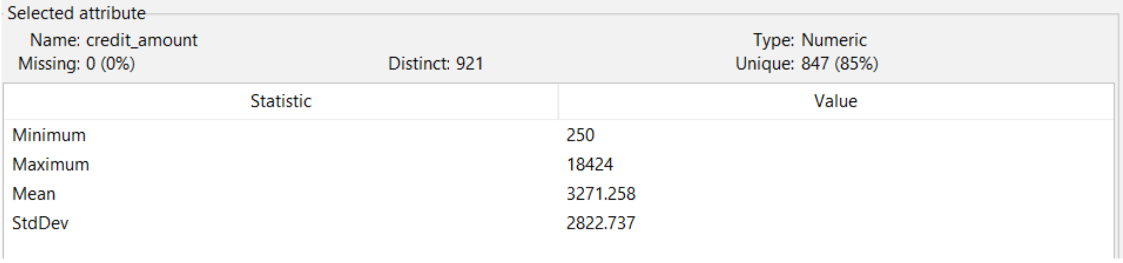


Selected attribute

Name: credit_history Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	no credits/all paid	40	40
2	all paid	49	49
3	existing paid	530	530
4	delayed previously	88	88
5	critical/other existing credit	293	293

- credit_amount (Continuous attribute): Credit amount. Missing: 0%, Distinct: 921, Type: Numeric, Unique: 847(85%).

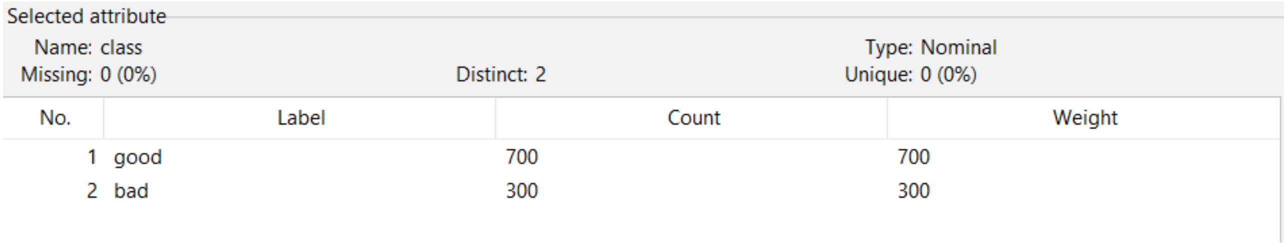


Selected attribute

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

- **Which attribute is used for the label? =>class** is used for the label.



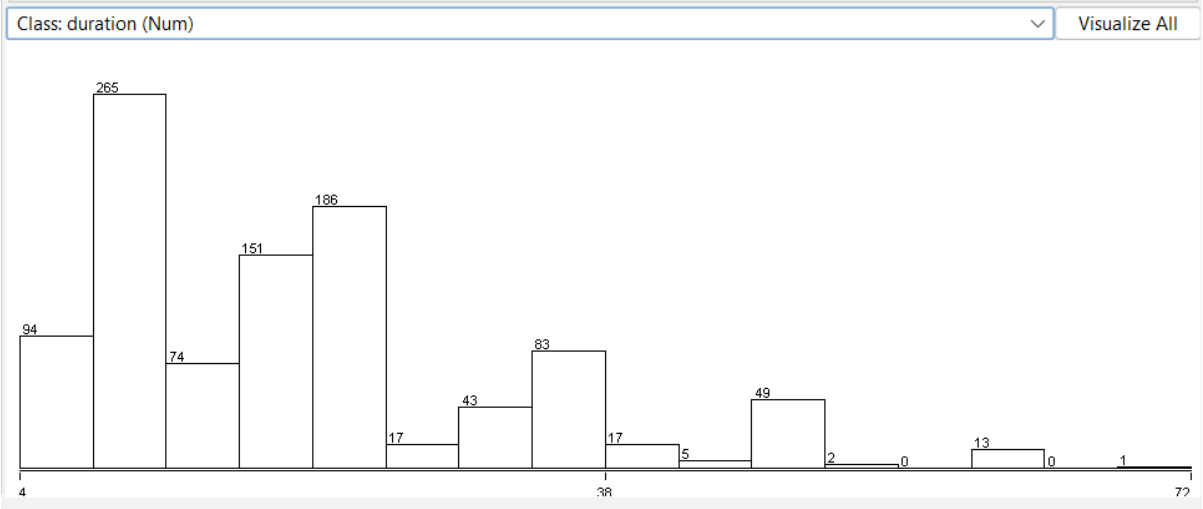
Selected attribute

Name: class Type: Nominal

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

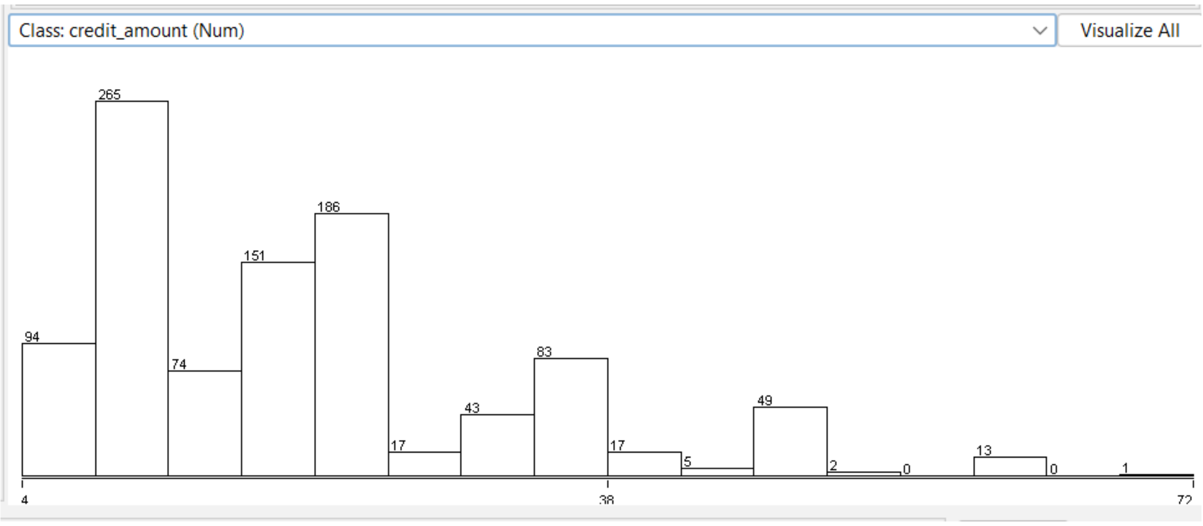
No.	Label	Count	Weight
1	good	700	700
2	bad	300	300

- **Let's describe the distribution of continuous attributes? (Left skewed or right skewed ?) =>** Contrinous attributes: **duration**, **credit_amount**, **installment_commitment**, **residence_since**, **age**, **existing_credits**, **num_dependents**.
  - **duration**: Right skewed.



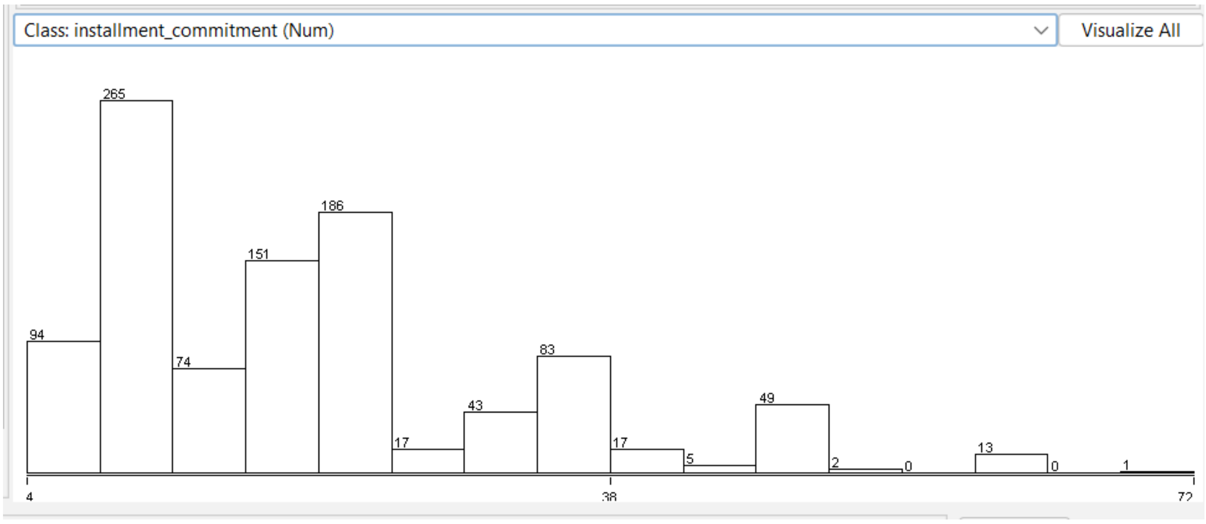
Class: duration (Num)

  - **credit_amount**: Right skewed.

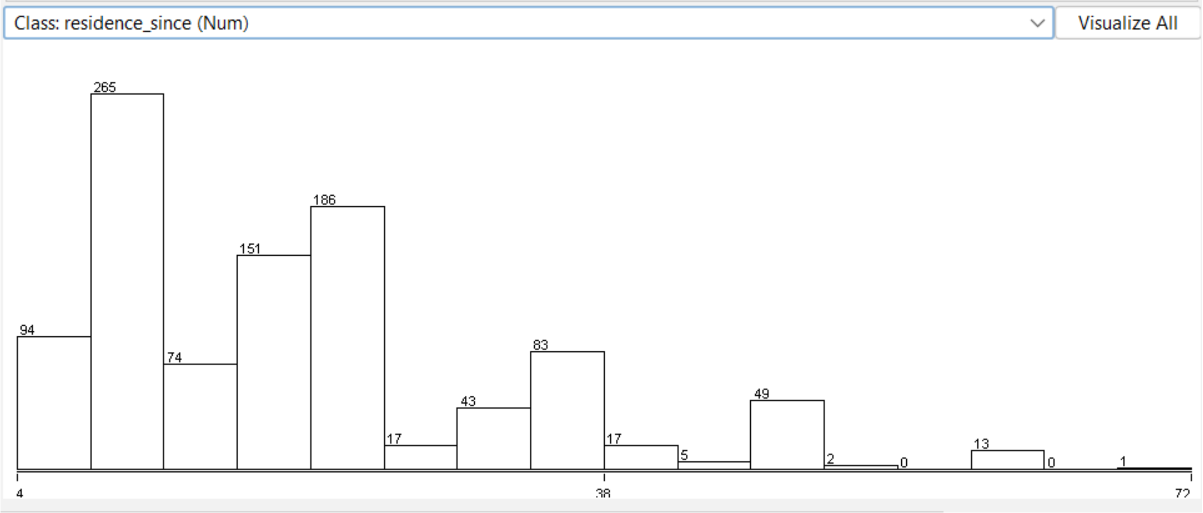


Class: credit_amount (Num)

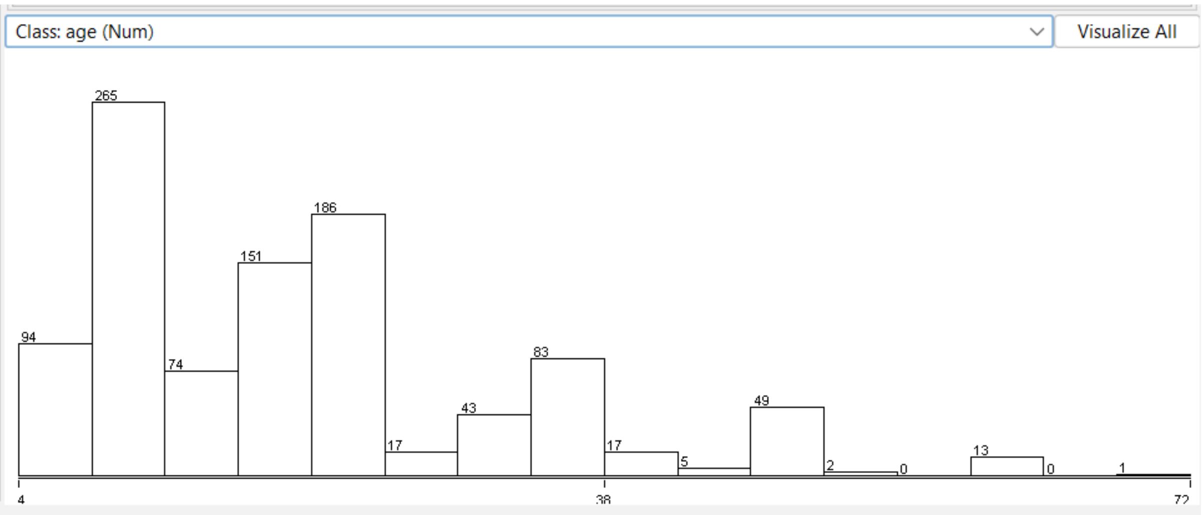
- **installment_commitment**: Right skewed.



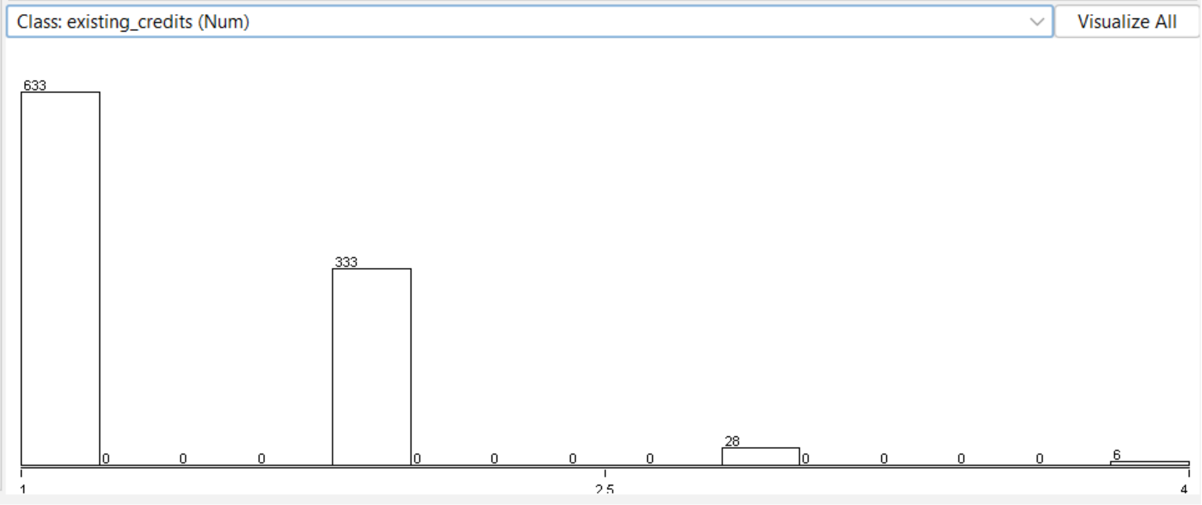
- **residence_since**: Right skewed.



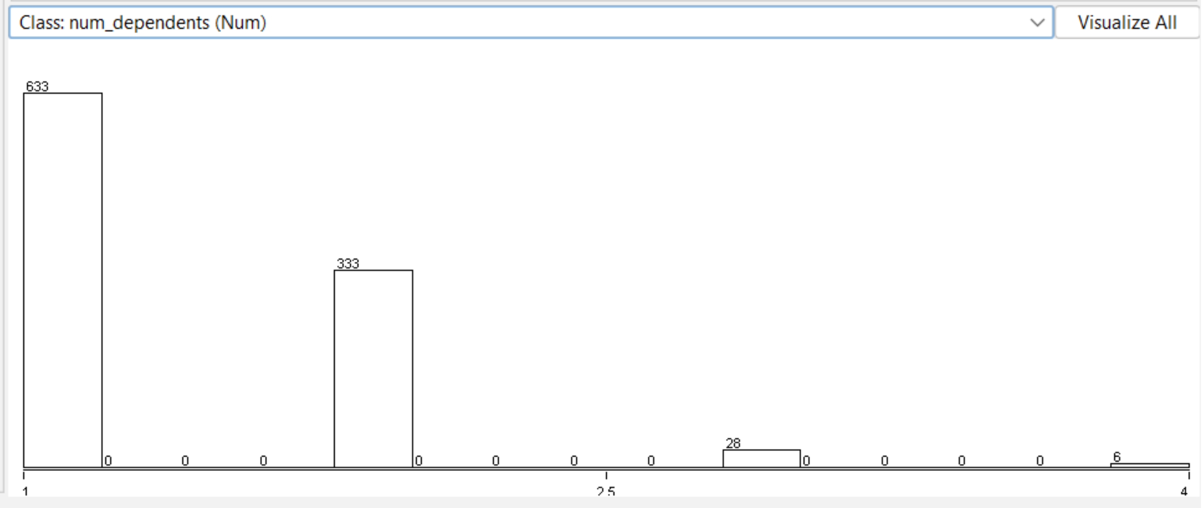
- **age**: Right skewed.



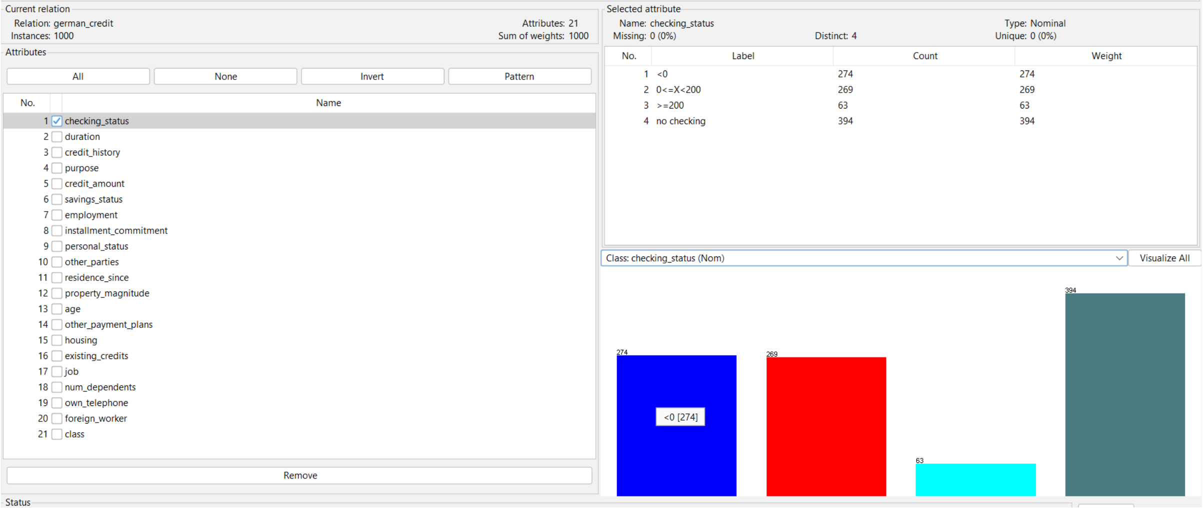
- **existing_credits**: Right skewed.



- **num_dependents**: Right skewed.



- **Let’s explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend. =>** Having 4 charts in the WEKA Explorer:
  - The first chart:



Current relation
Relation: german_credit
Instances: 1000
Attributes: 21
Sum of weights: 1000

Attributes

All None Invert Pattern

No.	Name
1	☑ checking_status
2	□ duration
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

Remove

Status

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

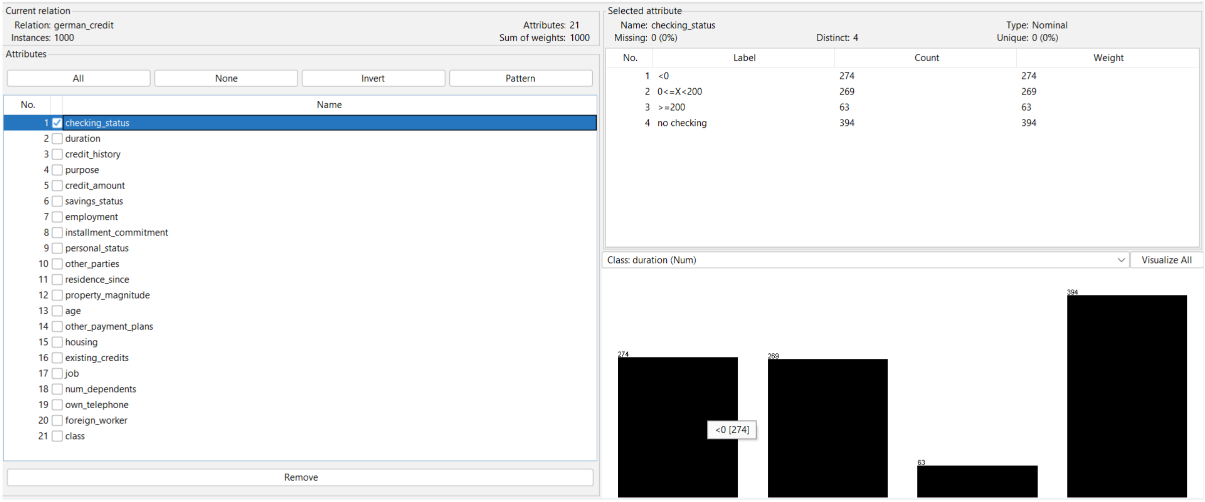
No.	Label	Count	Weight
1	<0	274	274
2	0<=X<200	269	269
3	>=200	63	63
4	no checking	394	394

Class: checking_status (Nom)

Visualize All

    - Explaining: It shows the bar chart with different colors correspond to each label at nominal attributes.
    - Setting the title for it: Own chart of nominal attribute.
    - Describing its legend: Each color corresponds to each label of nominal attribute.

- The second chart:



Current relation
Relation: german_credit
Instances: 1000
Attributes: 21
Sum of weights: 1000

Attributes

No.	Name
1	☑ checking_status
2	□ duration
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

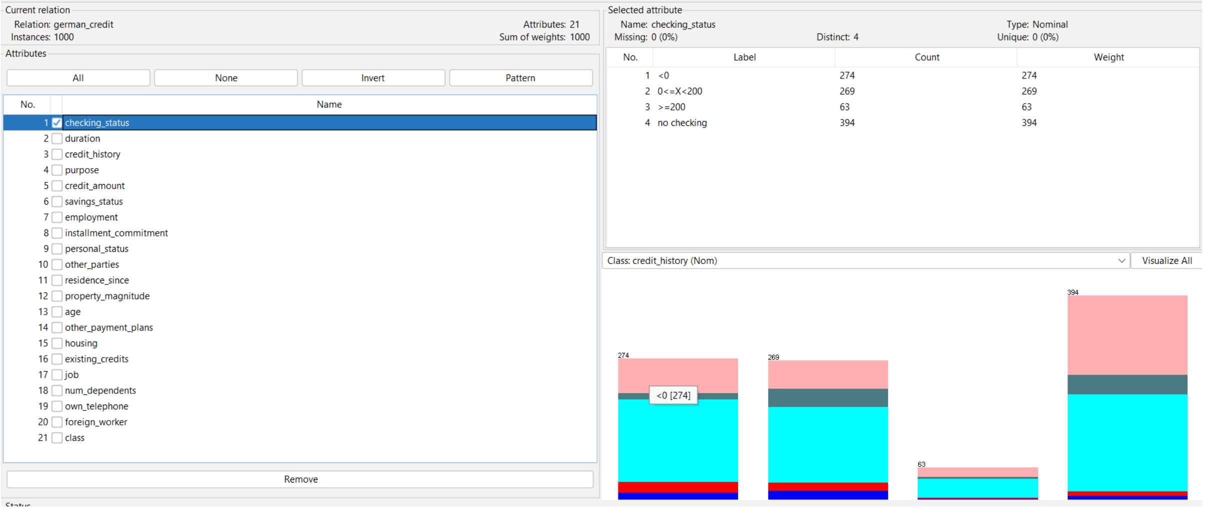
Selected attribute
Name: checking_status
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

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

Class: duration (Num)

  - Explaining: It shows the bar chart with one color (Black) about the relationship from nominal attribute (no attribute) to numeric attribute and because the numeric attribute don't have label so it is only one color.
  - Setting the title for it: Chart from nominal(no class) to numeric.
  - Describing its legend: Black shows value of numeric attribute.

- The third chart:



Current relation
Relation: german_credit
Instances: 1000
Attributes: 21
Sum of weights: 1000

Attributes

No.	Name
1	☑ checking_status
2	□ duration
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

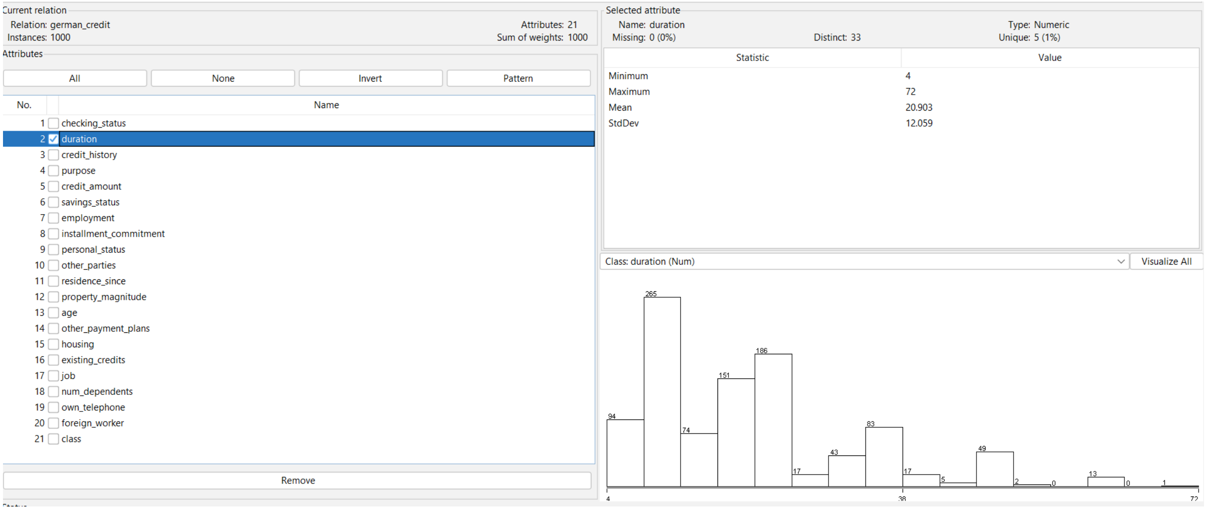
Selected attribute
Name: checking_status
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

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

Class: credit_history (Nom)

  - Explaining: It shows the stacked bar chart about the relationship from nominal attribute to different nominal attribute.
  - Setting the title for it: Chart from nominal to nominal.
  - Describing its legend: Each color corresponds to labels of the nominal attribute be chosen in "Attributes" (On the left)

- The fourth chart:



Current relation

Relation: german_credit
Instances: 1000
Attributes: 21
Sum of weights: 1000

Attributes

No.		Name
1	☐	checking_status
2	☑	duration
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

Selected attribute

Name: duration
Missing: 0 (0%)
Distinct: 33
Type: Numeric
Unique: 5 (1%)

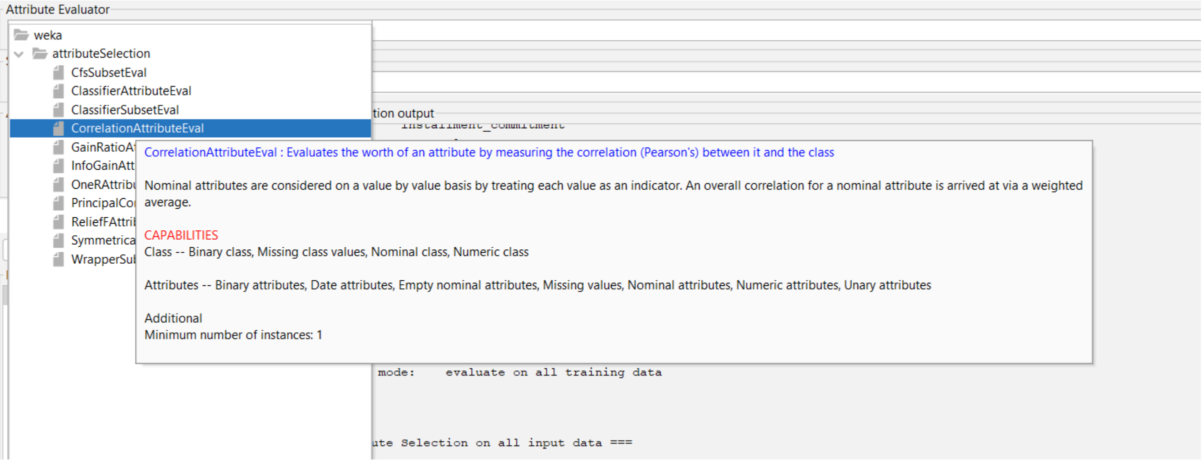
Statistic	Value
Minimum	4
Maximum	72
Mean	20.903
StdDev	12.059

Class: duration (Num)

  - Explaining: It shows the bar chart having only the range of value in numeric attribute.
  - Setting the title for it: Chart for numeric
  - Describing its legend: Don't have legend.

- **Let’s move to the Select attributes tag. Describe all of the options for attribute selection. =>** Having 11 options for Attribute Evaluator and 3 options for Search Method.
  - Attribute Evaluator:
    - **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.
    - **ClassifierAttributeEval**: Evaluates the worth of an attribute by using a user-specified classifier.
    - **ClassifierSubsetEval**: Evaluates attribute subsets on training data or a separate hold out testing set.
    - **CorrelationAttributeEval**: Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class.
    - **GainRadioAttributeEval**: Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.
    - **InfoGainAttributeEval**: Evaluates the worth of an attribute by measuring the information gain with respect to the class.
    - **OneRAttributeEval**: Evaluates the worth of an attribute by using the OneR classifier.
    - **PrincipalComponents**: Performs a principal components analysis and transformation of the data.
    - **ReliefFAttributeEval**: Evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.
    - **SymmetricalUncertAttributeEval**: Evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class.
    - **WrapperSubsetEval**: Evaluates attribute sets by using a learning scheme.
  - Search Method:
    - **BestFirst**: Searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility.
    - **GreedyStepwise**: Performs a greedy forward or backward search through the space of attribute subsets.
    - **Ranker**: Ranks attributes by their individual evaluations.

- **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) =>** The options should be used are **CorrelationAttributeEval**, **InfoGainAttributeEval**, **ReliefFAttributeEval** or **PrincipalComponents**, etc. - For example with **CorrelationAttributeEval**:
    - Step 1: Click on the "Choose" button in "Attribute Evalutor" and click on "CorrelationAttributeEval".



CorrelationAttributeEval : Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class

Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.

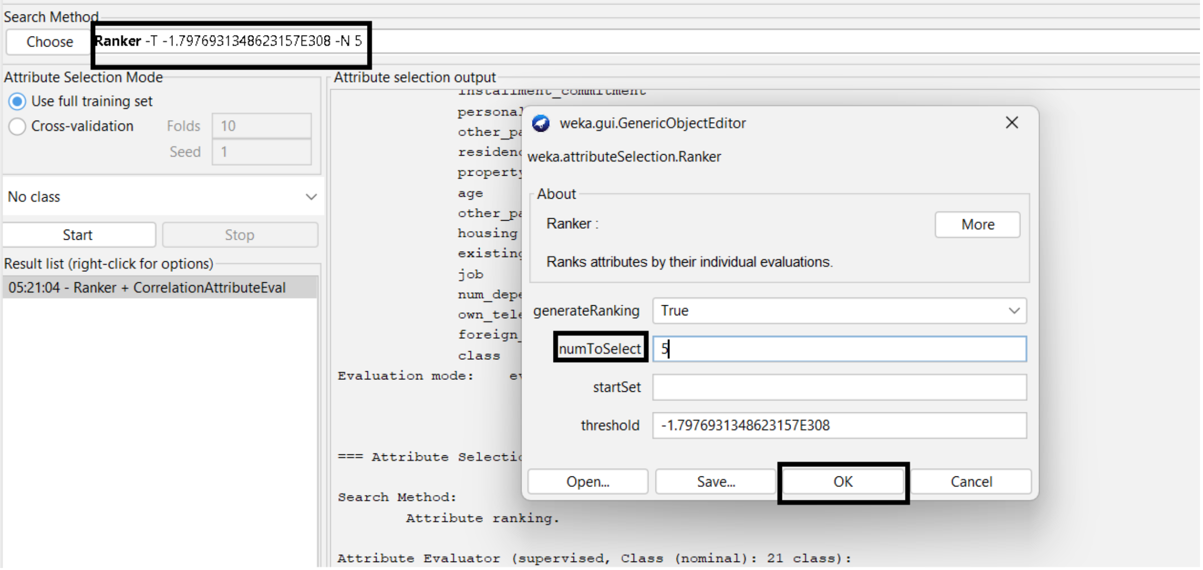
CAPABILITIES
Class -- Binary class, Missing class values, Nominal class, Numeric class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Unary attributes

Additional
Minimum number of instances: 1

```
mode:    evaluate on all training data
```

    - Step 2: Click on the "Ranker" in "Search Method" (Ranker is defaulted to choose when choosing "CorrelationAttributeEval") and set numToSelect to 5, after that click on the "OK" button.



Ranker -T -1.7976931348623157E308 -N 5

○ Cross-validation Folds 10 Seed 1

05:21:04 - Ranker + CorrelationAttributeEval

weka.attributeSelection.Ranker

Ranker :

Ranks attributes by their individual evaluations.

generateRanking: True

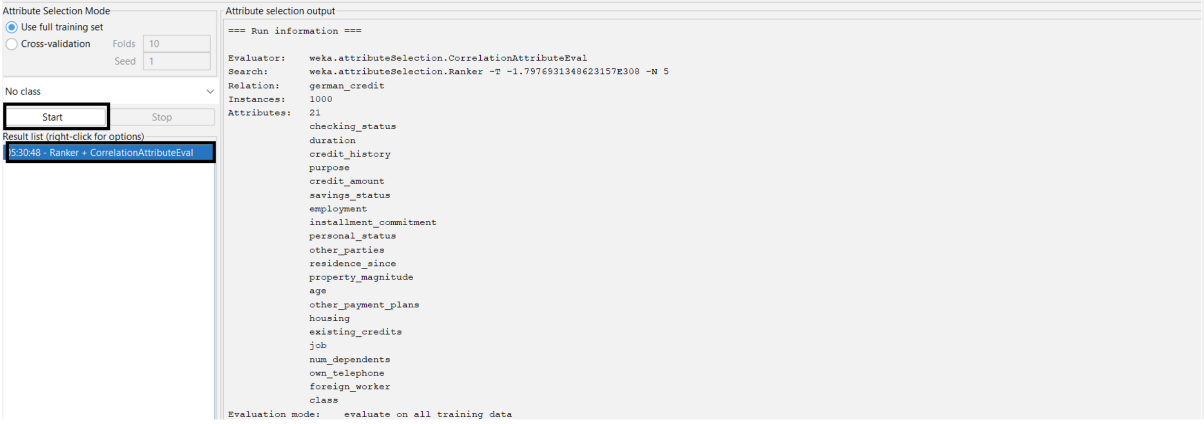
numToSelect: 5

startSet:

threshold: -1.7976931348623157E308

```
Evaluation mode:    e
=== Attribute Selecti
Search Method:
        Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 21 class):
```

    - Step 3: Click on the "Start" button to run the filter. The output will display the 5 attributes with the highest correlation. (Note: Choose the right list in result lists)

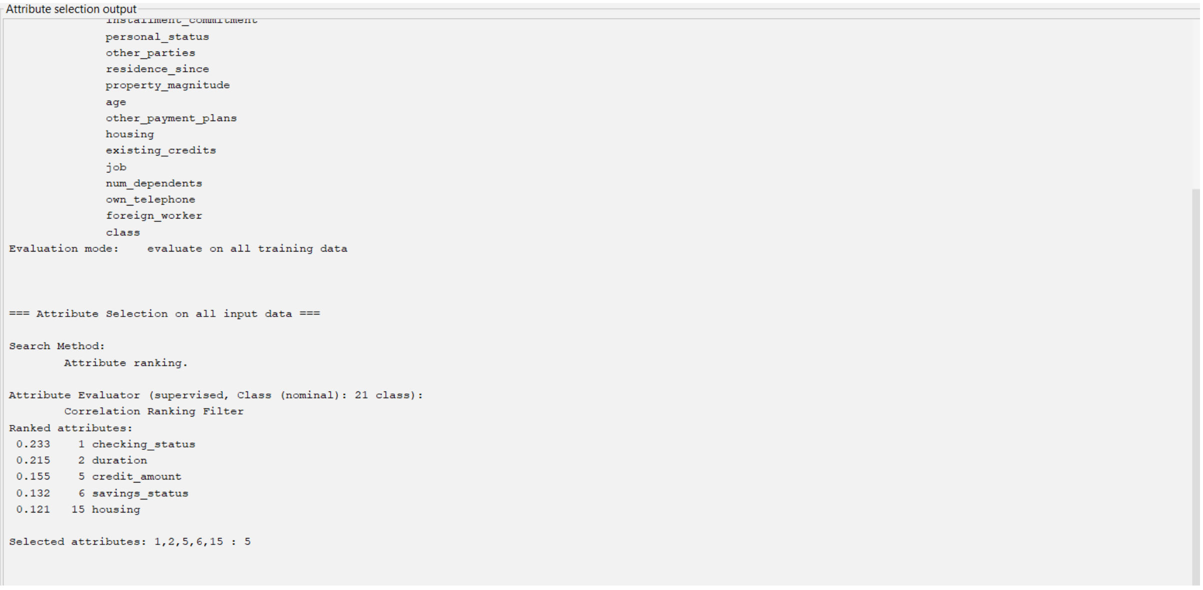


)5:30:48 - Ranker + CorrelationAttributeEval

```
=== Run information ===

Evaluator:    weka.attributeSelection.CorrelationAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 5
Relation:     german_credit
Instances:    1000
Attributes:   21
              checking_status
              duration
              credit_history
              purpose
              credit_amount
              savings_status
              employment
              installment_commitment
              personal_status
              other_parties
              residence_since
              property_magnitude
              age
              other_payment_plans
              housing
              existing_credits
              job
              num_dependents
              own_telephone
              foreign_worker
              class
Evaluation mode:    evaluate on all training data
```

- The final result:



```
             installment_commitment
             personal_status
             other_parties
             residence_since
             property_magnitude
             age
             other_payment_plans
             housing
             existing_credits
             job
             num_dependents
             own_telephone
             foreign_worker
             class
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

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
```

Preprocessing Data in Python (5 points)

The program must have the following functions (0.5 points for each function):

Extract columns with missing values

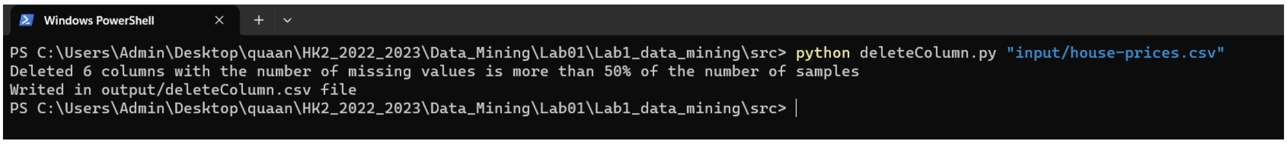
Count the number of lines with missing data.

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

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).

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).

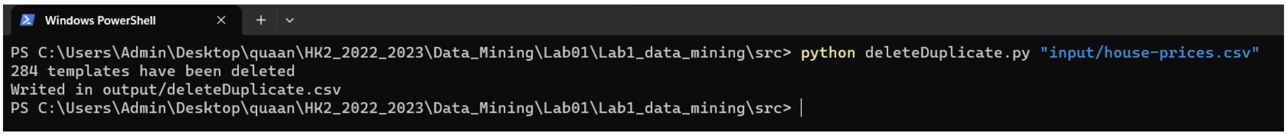
- Run file **deleteColumn.py**:



```
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> python deleteColumn.py "input/house-prices.csv"
Deleted 6 columns with the number of missing values is more than 50% of the number of samples
Writed in output/deleteColumn.csv file
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> |
```

Delete duplicate samples.

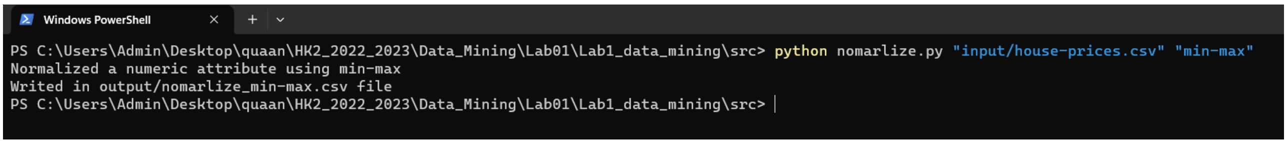
- Run file **deleteDuplicate.py**:



```
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> python deleteDuplicate.py "input/house-prices.csv"
284 templates have been deleted
Writed in output/deleteDuplicate.csv
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> |
```

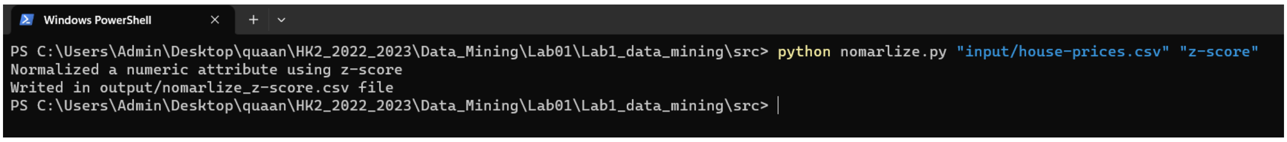
Normalize a numeric attribute using min-max and Z-score methods

- Run file **nomarlize.py** with min-max score:



```
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> python nomarlize.py "input/house-prices.csv" "min-max"
Normalized a numeric attribute using min-max
Writed in output/nomarlize_min-max.csv file
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> |
```

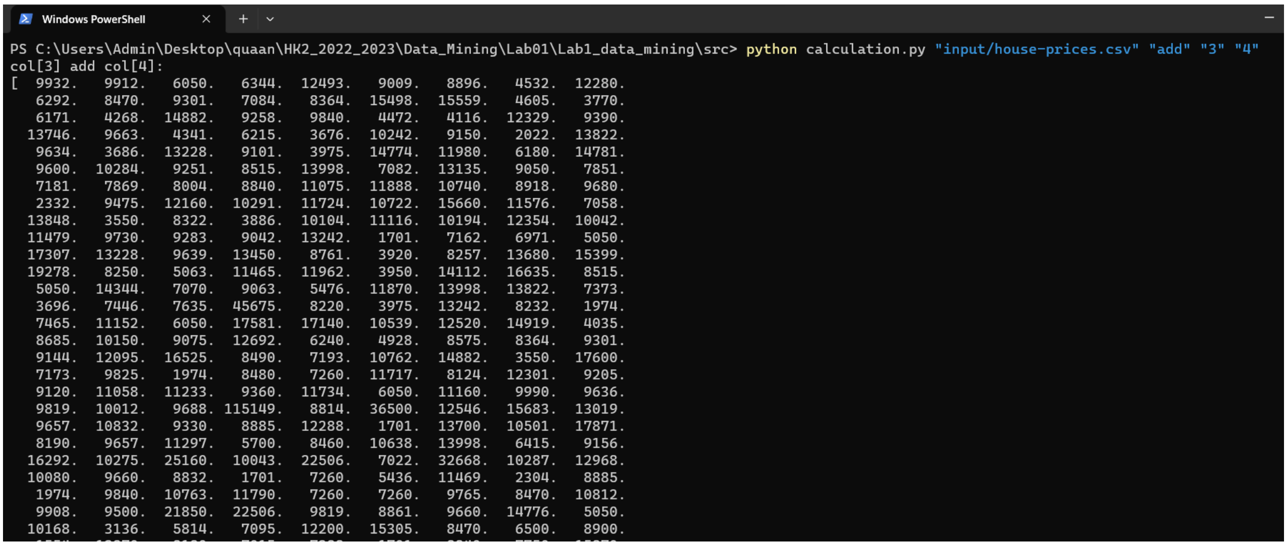
- Run file **nomarlize.py** with z-score:



```
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> python nomarlize.py "input/house-prices.csv" "z-score"
Normalized a numeric attribute using z-score
Writed in output/nomarlize_z-score.csv file
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> |
```

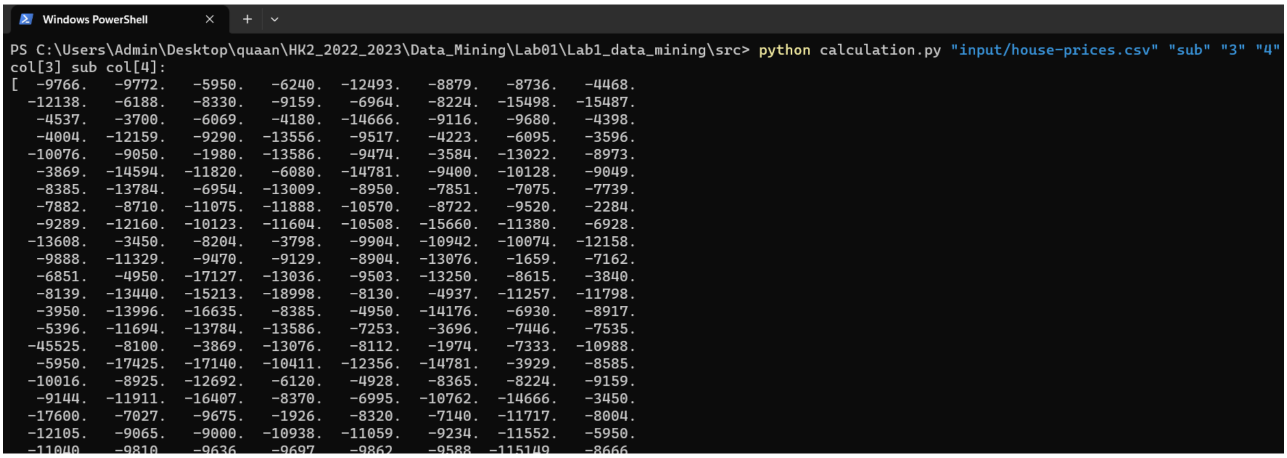
Performing addition, subtraction, multiplication, and division between two numerical attributes

- Run file **calculation** with add column 3 and 4:



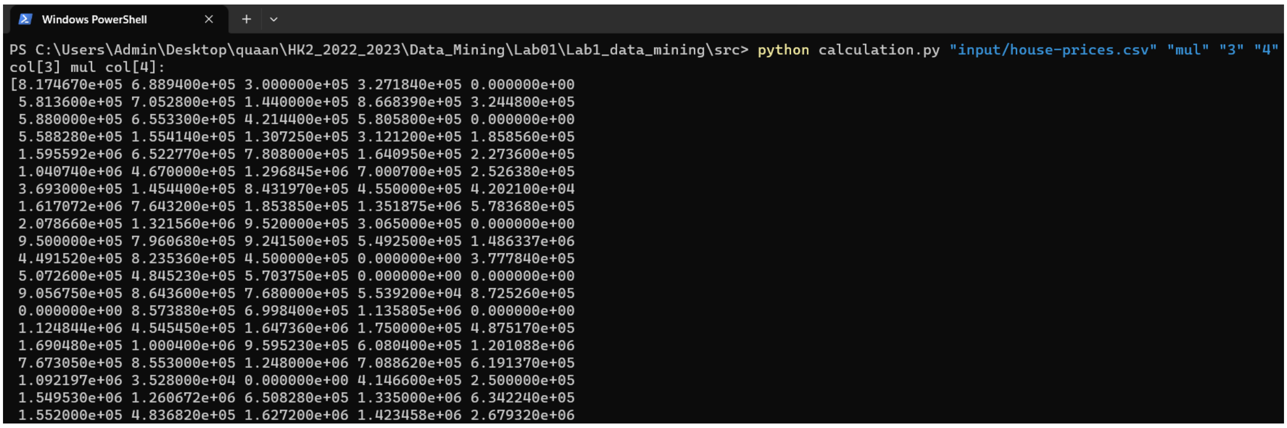
```
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> python calculation.py "input/house-prices.csv" "add" "3" "4"
col[3] add col[4]:
[  9932.   9912.   6050.   6344.  12493.   9009.   8896.   4532.  12280.
   6292.   8470.   9301.   7084.   8364.  15498.  15559.   4605.   3770.
   6171.   4268.  14882.   9258.   9840.   4472.   4116.  12329.   9390.
  13746.   9663.   4341.   6215.   3676.  10242.   9150.   2022.  13822.
   9634.   3686.  13228.   9101.   3975.  14774.  11980.   6180.  14781.
   9600.  10284.   9251.   8515.  13998.   7082.  13135.   9050.   7851.
   7181.   7869.   8004.   8840.  11075.  11888.  10740.   8918.   9680.
   2332.   9475.  12160.  10291.  11724.  10722.  15660.  11576.   7058.
  13848.   3550.   8322.   3886.  10104.  11116.  10194.  12354.  10042.
  11479.   9730.   9283.   9042.  13242.   1701.   7162.   6971.   5050.
  17307.  13228.   9639.  13450.   8761.   3920.   8257.  13680.  15399.
  19278.   8250.   5063.  11465.  11962.   3950.  14112.  16635.   8515.
   5050.  14344.   7070.   9063.   5476.  11870.  13998.  13822.   7373.
   3696.   7446.   7635.  45675.   8220.   3975.  13242.   8232.   1974.
   7465.  11152.   6050.  17581.  17140.  10539.  12520.  14919.   4035.
   8685.  10150.   9075.  12692.   6240.   4928.   8575.   8364.   9301.
   9144.  12095.  16525.   8490.   7193.  10762.  14882.   3550.  17600.
   7173.   9825.   1974.   8480.   7260.  11717.   8124.  12301.   9205.
   9120.  11058.  11233.   9360.  11734.   6050.  11160.   9990.   9636.
   9819.  10012.   9688. 115149.   8814.  36500.  12546.  15683.  13019.
   9657.  10832.   9330.   8885.  12288.   1701.  13700.  10501.  17871.
   8190.   9657.  11297.   5700.   8460.  10638.  13998.   6415.   9156.
  16292.  10275.  25160.  10043.  22506.   7022.  32668.  10287.  12968.
  10080.   9660.   8832.   1701.   7260.   5436.  11469.   2304.   8885.
   1974.   9840.  10763.  11790.   7260.   7260.   9765.   8470.  10812.
   9908.   9500.  21850.  22506.   9819.   8861.   9660.  14776.   5050.
  10168.   3136.   5814.   7095.  12200.  15305.   8470.   6500.   8900.
```

- Run file **calculation** with sub column 3 and 4:



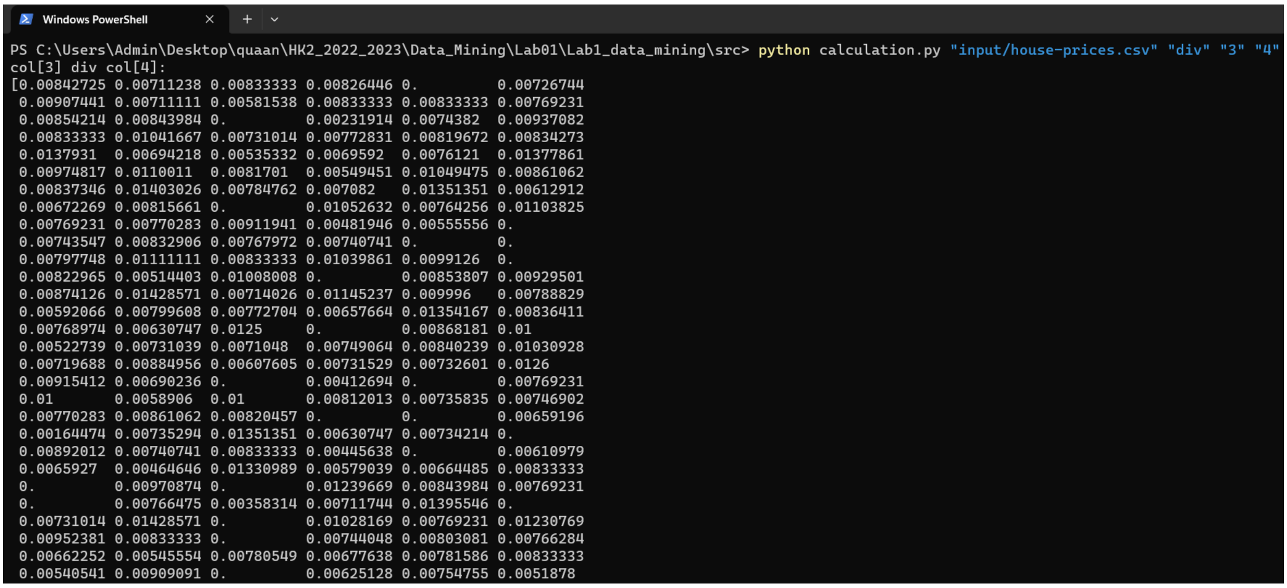
```
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> python calculation.py "input/house-prices.csv" "sub" "3" "4"
col[3] sub col[4]:
[  -9766.   -9772.   -5950.   -6240.  -12493.   -8879.   -8736.   -4468.
  -12138.   -6188.   -8330.   -9159.   -6964.   -8224.  -15498.  -15487.
   -4537.   -3700.   -6069.   -4180.  -14666.   -9116.   -9680.   -4398.
   -4004.  -12159.   -9290.  -13556.   -9517.   -4223.   -6095.   -3596.
  -10076.   -9050.   -1980.  -13586.   -9474.   -3584.  -13022.   -8973.
   -3869.  -14594.  -11820.   -6080.  -14781.   -9400.  -10128.   -9049.
   -8385.  -13784.   -6954.  -13009.   -8950.   -7851.   -7075.   -7739.
   -7882.   -8710.  -11075.  -11888.  -10570.   -8722.   -9520.   -2284.
   -9289.  -12160.  -10123.  -11604.  -10508.  -15660.  -11380.   -6928.
  -13608.   -3450.   -8204.   -3798.   -9904.  -10942.  -10074.  -12158.
   -9888.  -11329.   -9470.   -9129.   -8904.  -13076.   -1659.   -7162.
   -6851.   -4950.  -17127.  -13036.   -9503.  -13250.   -8615.   -3840.
   -8139.  -13440.  -15213.  -18998.   -8130.   -4937.  -11257.  -11798.
   -3950.  -13996.  -16635.   -8385.   -4950.  -14176.   -6930.   -8917.
   -5396.  -11694.  -13784.  -13586.   -7253.   -3696.   -7446.   -7535.
  -45525.   -8100.   -3869.  -13076.   -8112.   -1974.   -7333.  -10988.
   -5950.  -17425.  -17140.  -10411.  -12356.  -14781.   -3929.   -8585.
  -10016.   -8925.  -12692.   -6120.   -4928.   -8365.   -8224.   -9159.
   -9144.  -11911.  -16407.   -8370.   -6995.  -10762.  -14666.   -3450.
  -17600.   -7027.   -9675.   -1926.   -8320.   -7140.  -11717.   -8004.
  -12105.   -9065.   -9000.  -10938.  -11059.   -9234.  -11552.   -5950.
  -11040.   -9810.   -9636.   -9697.   -9862.   -9588. -115149.   -8666.
```

- Run file **calculation** with mul column 3 and 4:



```
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> python calculation.py "input/house-prices.csv" "mul" "3" "4"
col[3] mul col[4]:
[8.174670e+05 6.889400e+05 3.000000e+05 3.271840e+05 0.000000e+00
 5.813600e+05 7.052800e+05 1.440000e+05 8.668390e+05 3.244800e+05
 5.880000e+05 6.553300e+05 4.214400e+05 5.805800e+05 0.000000e+00
 5.588280e+05 1.554140e+05 1.307250e+05 3.121200e+05 1.858560e+05
 1.595592e+06 6.522770e+05 7.808000e+05 1.640950e+05 2.273600e+05
 1.040740e+06 4.670000e+05 1.296845e+06 7.000700e+05 2.526380e+05
 3.693000e+05 1.454400e+05 8.431970e+05 4.550000e+05 4.202100e+04
 1.617072e+06 7.643200e+05 1.853850e+05 1.351875e+06 5.783680e+05
 2.078660e+05 1.321560e+06 9.520000e+05 3.065000e+05 0.000000e+00
 9.500000e+05 7.960680e+05 9.241500e+05 5.492500e+05 1.486337e+06
 4.491520e+05 8.235360e+05 4.500000e+05 0.000000e+00 3.777840e+05
 5.072600e+05 4.845230e+05 5.703750e+05 0.000000e+00 0.000000e+00
 9.056750e+05 8.643600e+05 7.680000e+05 5.539200e+04 8.725260e+05
 0.000000e+00 8.573880e+05 6.998400e+05 1.135805e+06 0.000000e+00
 1.124844e+06 4.545450e+05 1.647360e+06 1.750000e+05 4.875170e+05
 1.690480e+05 1.000400e+06 9.595230e+05 6.080400e+05 1.201088e+06
 7.673050e+05 8.553000e+05 1.248000e+06 7.088620e+05 6.191370e+05
 1.092197e+06 3.528000e+04 0.000000e+00 4.146600e+05 2.500000e+05
 1.549530e+06 1.260672e+06 6.508280e+05 1.335000e+06 6.342240e+05
 1.552000e+05 4.836820e+05 1.627200e+06 1.423458e+06 2.679320e+06
```

- Run file **calculation** with div column 3 and 4:



```
PS C:\Users\Admin\Desktop\quaan\HK2_2022_2023\Data_Mining\Lab01\Lab1_data_mining\src> python calculation.py "input/house-prices.csv" "div" "3" "4"
col[3] div col[4]:
[0.00842725 0.00711238 0.00833333 0.00826446 0.         0.00726744
 0.00907441 0.00711111 0.00581538 0.00833333 0.00833333 0.00769231
 0.00854214 0.00843984 0.         0.00231914 0.0074382  0.00937082
 0.00833333 0.01041667 0.00731014 0.00772831 0.00819672 0.00834273
 0.0137931  0.00694218 0.00535332 0.0069592  0.0076121  0.01377861
 0.00974817 0.0110011  0.0081701  0.00549451 0.01049475 0.00861062
 0.00837346 0.01403026 0.00784762 0.007082   0.01351351 0.00612912
 0.00672269 0.00815661 0.         0.01052632 0.00764256 0.01103825
 0.00769231 0.00770283 0.00911941 0.00481946 0.00555556 0.
 0.00743547 0.00832906 0.00767972 0.00740741 0.         0.
 0.00797748 0.01111111 0.00833333 0.01039861 0.0099126  0.
 0.00822965 0.00514403 0.01008008 0.         0.00853807 0.00929501
 0.00874126 0.01428571 0.00714026 0.01145237 0.009996   0.00788829
 0.00592066 0.00799608 0.00772704 0.00657664 0.01354167 0.00836411
 0.00768974 0.00630747 0.0125     0.         0.00868181 0.01
 0.00522739 0.00731039 0.0071048  0.00749064 0.00840239 0.01030928
 0.00719688 0.00884956 0.00607605 0.00731529 0.00732601 0.0126
 0.00915412 0.00690236 0.         0.00412694 0.         0.00769231
 0.01       0.0058906  0.01       0.00812013 0.00735835 0.00746902
 0.00770283 0.00861062 0.00820457 0.         0.         0.00659196
 0.00164474 0.00735294 0.01351351 0.00630747 0.00734214 0.
 0.00892012 0.00740741 0.00833333 0.00445638 0.         0.00610979
 0.0065927  0.00464646 0.01330989 0.00579039 0.00664485 0.00833333
 0.         0.00970874 0.         0.01239669 0.00843984 0.00769231
 0.         0.00766475 0.00358314 0.00711744 0.01395546 0.
 0.00731014 0.01428571 0.         0.01028169 0.00769231 0.01230769
 0.00952381 0.00833333 0.         0.00744048 0.00803081 0.00766284
 0.00662252 0.00545554 0.00780549 0.00677638 0.00781586 0.00833333
 0.00540541 0.00909091 0.         0.00625128 0.00754755 0.0051878
```