



LAB 01: DATA PREPROCESSING AND DATA EXPLORATION

Class: 20KHDL

Students:

Hồ Minh Thanh Tài – 20127068

Nguyễn Minh Tuấn – 20127092

Instructors:

Teacher Lê Hoài Bắc

Teacher Nguyễn Thị Thu Hằng

Academic Year: 2022-2023

The contribution rate of each member:

	Name	ID	Contribution rate (%)
1	Hồ Minh Thanh Tài	20127068	50
2	Nguyễn Minh Tuấn	20127092	50

The questions that have not been completed: None

Contents

1	Install WEKA	3
1.1	Requirement 1	3
1.2	Requirement 2:	3
2.	GETTING ACQUAINTED WITH WEKA.....	5
2.1	Exploring Breast Cancer Data Set	5
2.2	Exploring Weather Data Set.....	9
2.3	Exploring Credit In Germany Data Set	13
3	PREPROCESSING DATA IN PYTHON.....	18
3.1	Extract columns with missing values	18
3.2	Count the number of lines with missing data	18
3.3	Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute).....	19
3.3.1	Using mean	19
3.3.2	Using median	20
3.3.3	Using mode	21
3.4	Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes).	22
3.5	Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples).....	23
3.6	Delete duplicate samples.	24
3.7	Normalize a numeric attribute using min-max and Z-score methods.	25
3.7.1	min-max method.....	25
3.7.2	z-score method.....	26
3.8	Performing addition, subtraction, multiplication, and division between two numerical attributes.	27
3.8.1	addition.....	27
3.8.2	subtraction.....	28
3.8.3	multiplication.....	29
3.8.4	division	30
4.	REFERENCE	31

1 Install WEKA

1.1 Requirement 1: After installing, you capture a screen that contains the "Explorer" function in your desktop background.

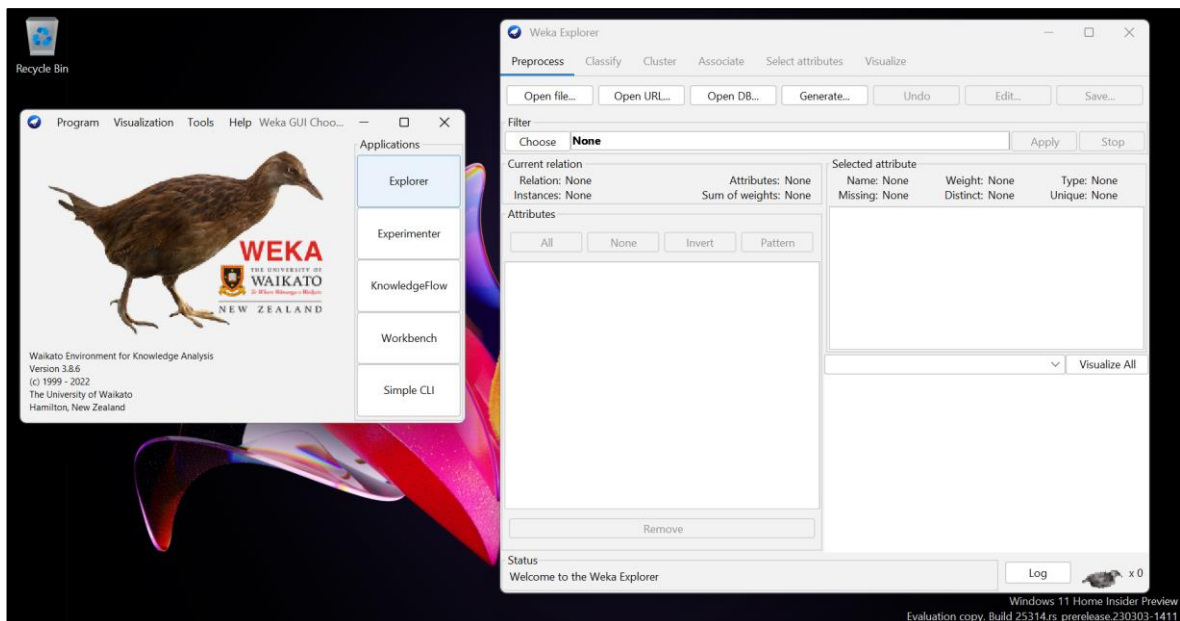


Image 1: A screen that contains the "Explorer" function in your desktop background.

1.2 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.

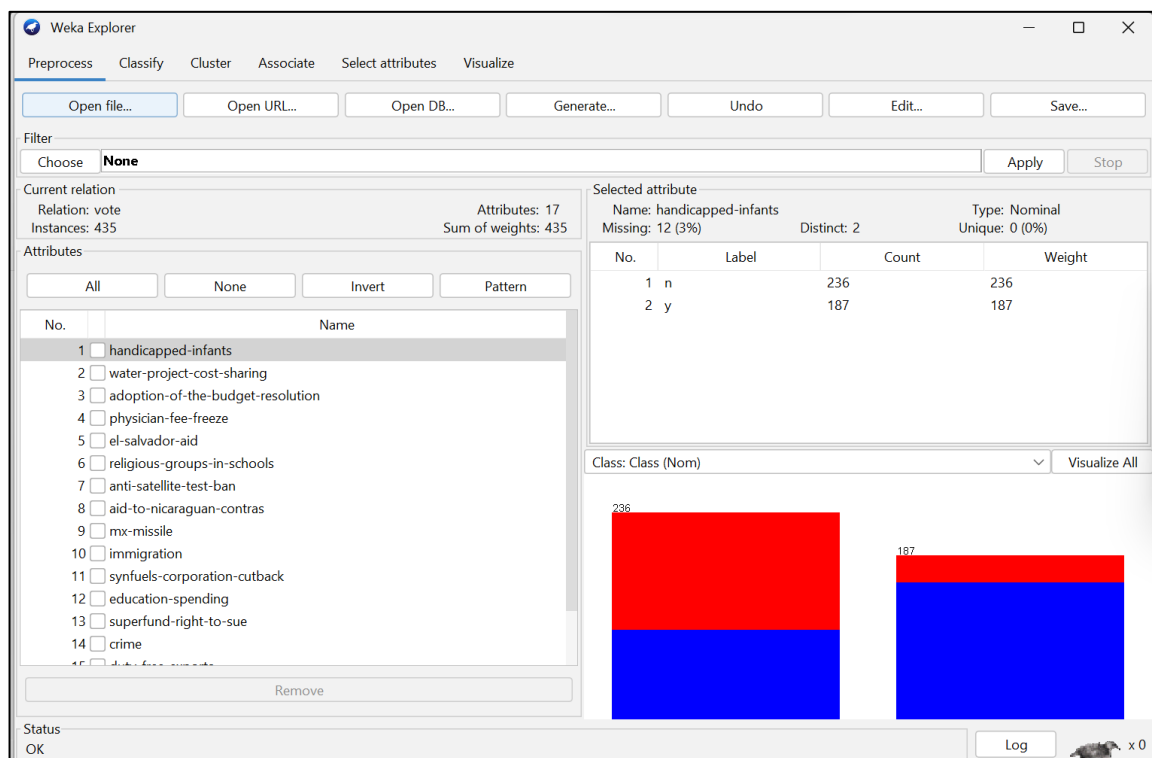


Image 2: Students open any data set – vote.arff

The meaning of Current Relation, Attributes, and Selected attribute in Preprocess tag.

In Preprocess Tag:

- **Current Relation:** Information about the data set, including its relation, instances (the number of rows in the table), the number of characteristics it contains, and the weighted average of those attributes, is now related.
- **Attributes:** the number of attributes. When a user selects an attribute, information about that attribute will be shown on the right side.
- **Selected Attribute:** The selected characteristics' information includes the following: name, missing rate, distinct, type and unique. Moreover, it displays the count and weight for each nominal value as a percentage.

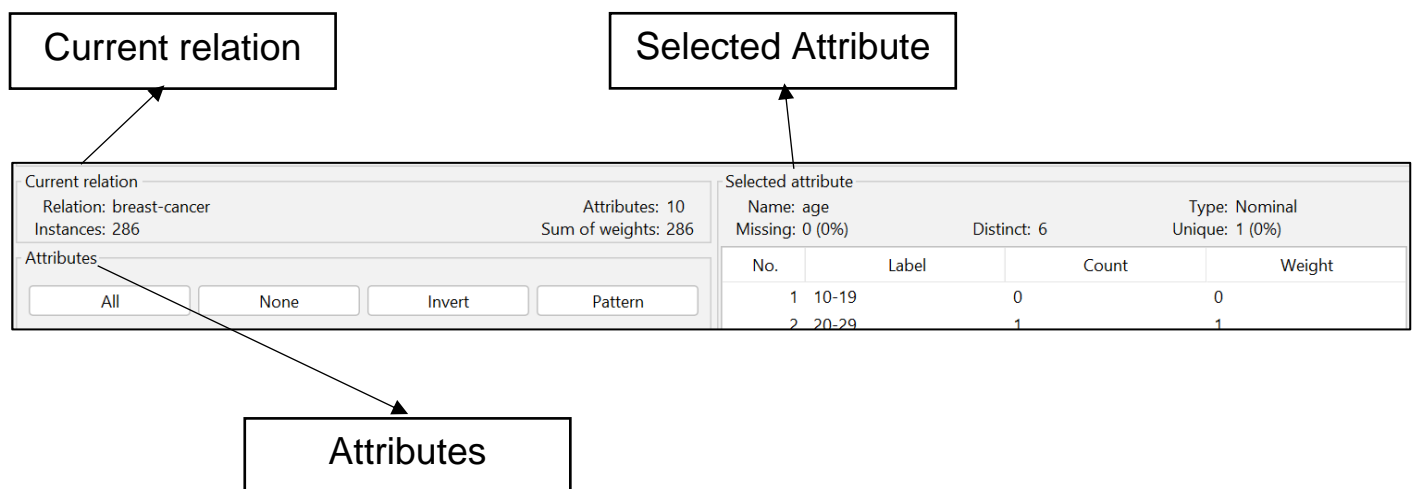


Image 3: Preproces Tag

Briefly explain the meaning of the other tags in WEKA Explorer.

Other Tags In WEKA Explorer include: classify, cluster, associate, select attributes and visualize.

- **Classify:** provide many machine learning algorithms to classify data.
- **Cluster:** provide many algorithms of clustering.
- **Associate:** contains Apriori, FilteredAssociator and FPGrowth.
- **Select attributes:** allow user to choose features based on several algorithms such as ClassifierSubsetEval, PrincipalComponents, etc.
- **Visualize:** allow user to visualize processed data for analysis.

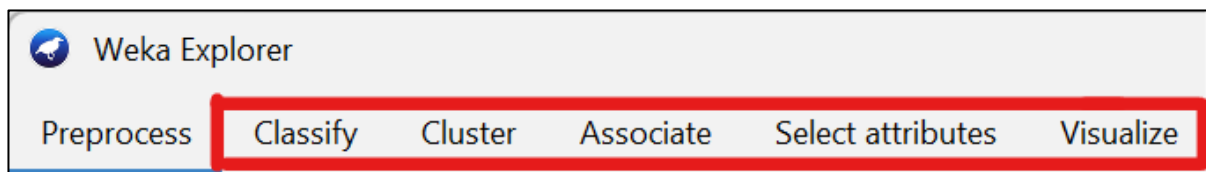


Image 4: Other Tags in Weka Explorer

2. GETTING ACQUAINTED WITH WEKA

2.1 Exploring Breast Cancer Data Set

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

– How many instances does this data set have?

→ This data set have 286 instances.

– How many attributes does this data set have?

→ This data set have 10 attributes.

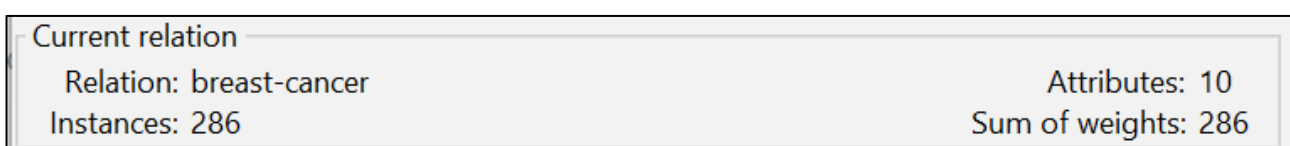


Image 5: Current relation of data set

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

- ➔ Class. Since the "Class" property indicates the result of cancer prediction. This will enable the data to be merged with all other elements to establish whether a victim's issue was a cancer recurrence occurrence or not.
- ➔ It can be changed. If the aforementioned characteristic is necessary for breast cancer to be connected to another ailment, then this label may be changed to another.

– What is the meaning of each attribute?

1. **age:** Patient's age at the time of the diagnosis.
2. **menopause:** At the time of diagnosis, the patient must be either pre- or postmenopausal.
3. **tumor-size:** the largest diameter of the removed tumor, measured in millimeters.
4. **inv-nodes:** the number (between 0 and 39) of axillary lymph nodes that exhibit histological evidence of metastatic breast cancer.
5. **node-caps:** If the cancer does spread to a lymph node, even if it is no longer at the site of the tumor, it may still be "contained" by the lymph node's capsule. But, over time and a more severe illness, the tumor can take the place of the lymph node and eventually break through the capsule, allowing it to infiltrate the tissues close by.
6. **deg-malig:** the tumor's histological grade (grades vary from 1-3). Grade 1 tumors are primarily made up of neoplastic cells that nonetheless largely retain their normal properties. Grade 3 tumors primarily consist of extremely aberrant cells;
7. **breast:** It is evident that breast cancer can affect either breast;
8. **breast-quad:** the breast may be divided into four quadrants, using the nipple as a central point;
9. **irradiat:** radiation therapy is a medical procedure that uses high-energy x-rays to kill cancer cells. The breast can be divided into four quadrants, with the nipple serving as the center point.

10. Class: The cancer "coming back" after at least a year of remission is what constitutes a recurrence event. It is possible for this recurrence to occur locally (in the same organ), regionally (in a nearby organ), or distantly (in another part of the body).

No.		Name
1	<input type="checkbox"/>	age
2	<input type="checkbox"/>	menopause
3	<input type="checkbox"/>	tumor-size
4	<input type="checkbox"/>	inv-nodes
5	<input type="checkbox"/>	node-caps
6	<input type="checkbox"/>	deg-malig
7	<input type="checkbox"/>	breast
8	<input type="checkbox"/>	breast-quad
9	<input type="checkbox"/>	irradiat
10	<input type="checkbox"/>	Class

Image 6: Attributes of the data set "breast cancer"

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

The properties **node-caps** and **breast-quad** are missing their values.

General solutions to the missing values issue include:

1. Subtract missing values from the data in your data set: Eliminating instances with one or more absent values is a simple technique to handle missing data.
2. Impute absent values: Instead of deleting instances with missing values, the data set might substitute the missing values with alternative values. Imputing missing values is what is done in this situation.

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

The most popular option is to investigate the continuous variable as a Mean-based workaround in the case of missing values. However, given the bulk of the variables in this data set are nominal, using the Mode instead of the other options was the best course of action.



Image 7: Node-caps attribute

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

Patient's age at the time of the diagnosis is divided into 9 parts. Through the chart, we can see the probability of recurrent or non-recurrent of breast cancer of each age group.

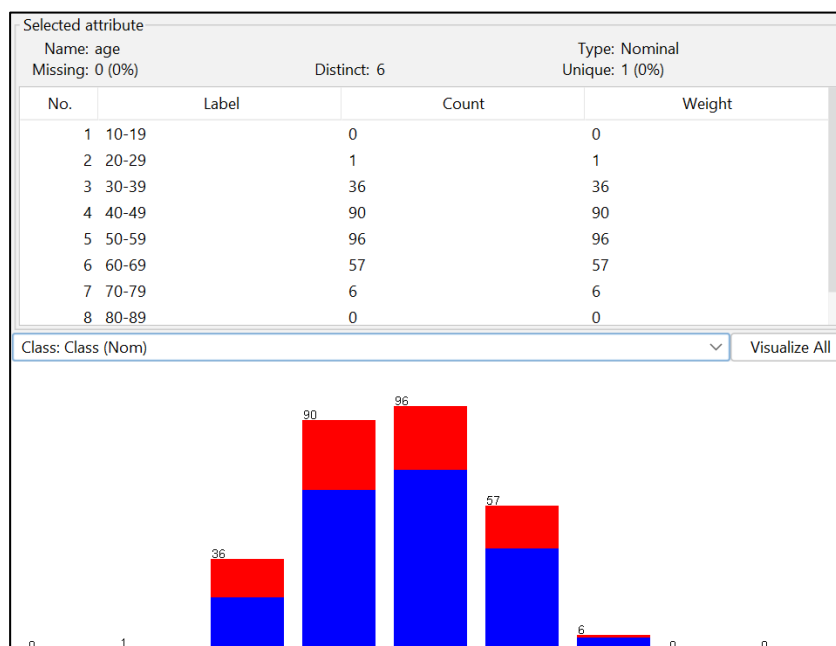


Image 8: Age attribute

2.2 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:

– How many attributes does this data set have?

→ This data set have 5 attributes

– How many samples?

→ 14 samples

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

Image 9: Current relation of the data set.

Which attributes have data type categorical?

→ Attributes have data type categorical: outlook, windy, play.

Which attributes have a data type that is numerical?

→ Attributes have a data type that is numerical: temperature, humidity.

Which attribute is used for the label?

→ Attribute is used for the label: play.

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

Weka offers several summaries of numerical attributes, although the presentation only includes a handful of the following figures;

Selected attribute		
Name: temperature		Type: Numeric
Missing: 0 (0%)	Distinct: 12	Unique: 10 (71%)
Statistic	Value	
Minimum	64	
Maximum	85	
Mean	73.571	
StdDev	6.572	

Image 10: Summary of Temperature

Selected attribute		
Name: humidity		Type: Numeric
Missing: 0 (0%)	Distinct: 10	Unique: 7 (50%)
Statistic	Value	
Minimum	65	
Maximum	96	
Mean	81.643	
StdDev	10.285	

Image 11: Summary of Humidity

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

The chart uses categorical numbers to show the distribution and tendency of each property. The two numerical values, which include temperature and humidity, display the distribution of each bin for the continuous range of data.

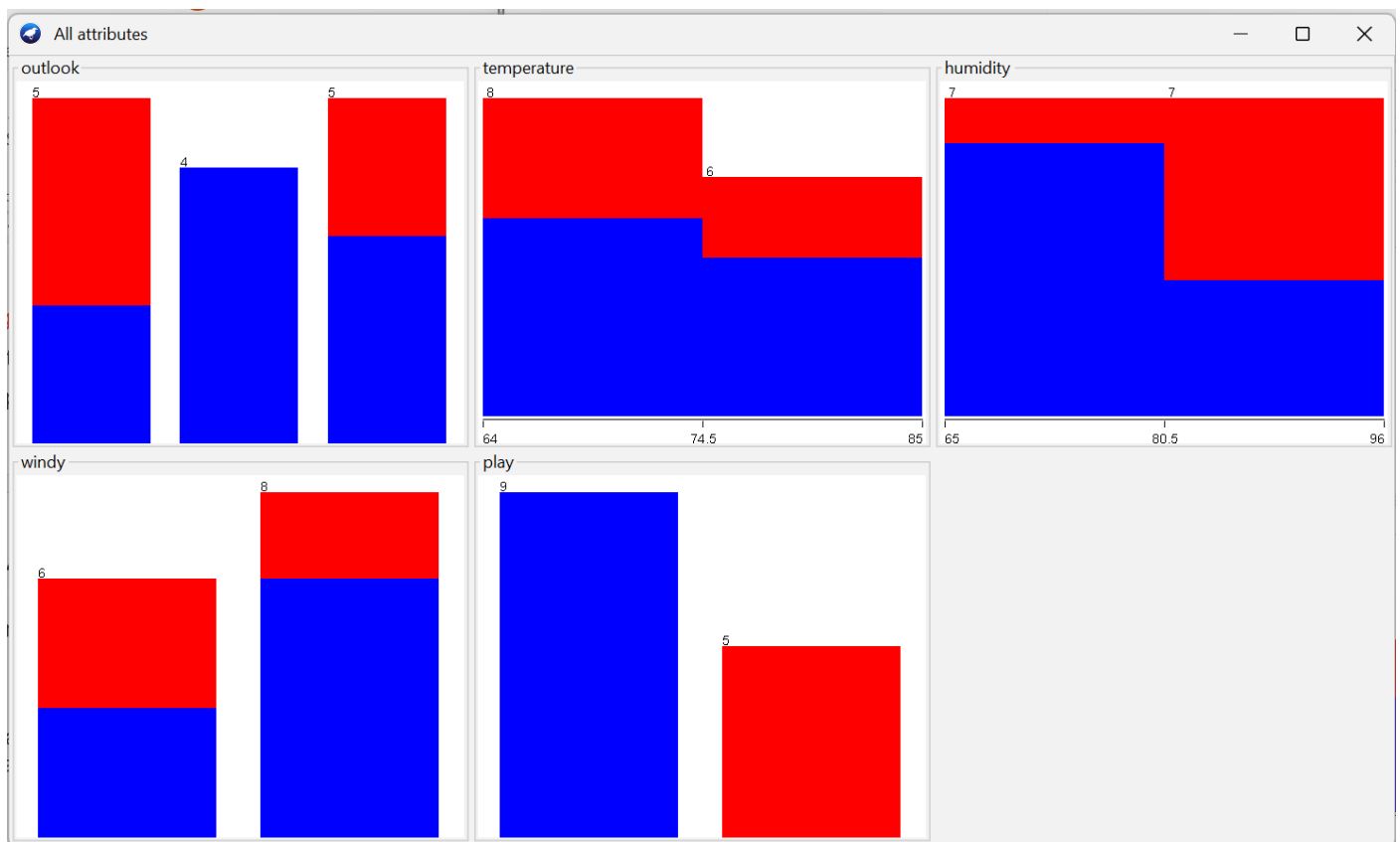


Image 12: Weather data set's charts

– 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?

The name of the chart in the Visualize Tag might be Matrix Plot.

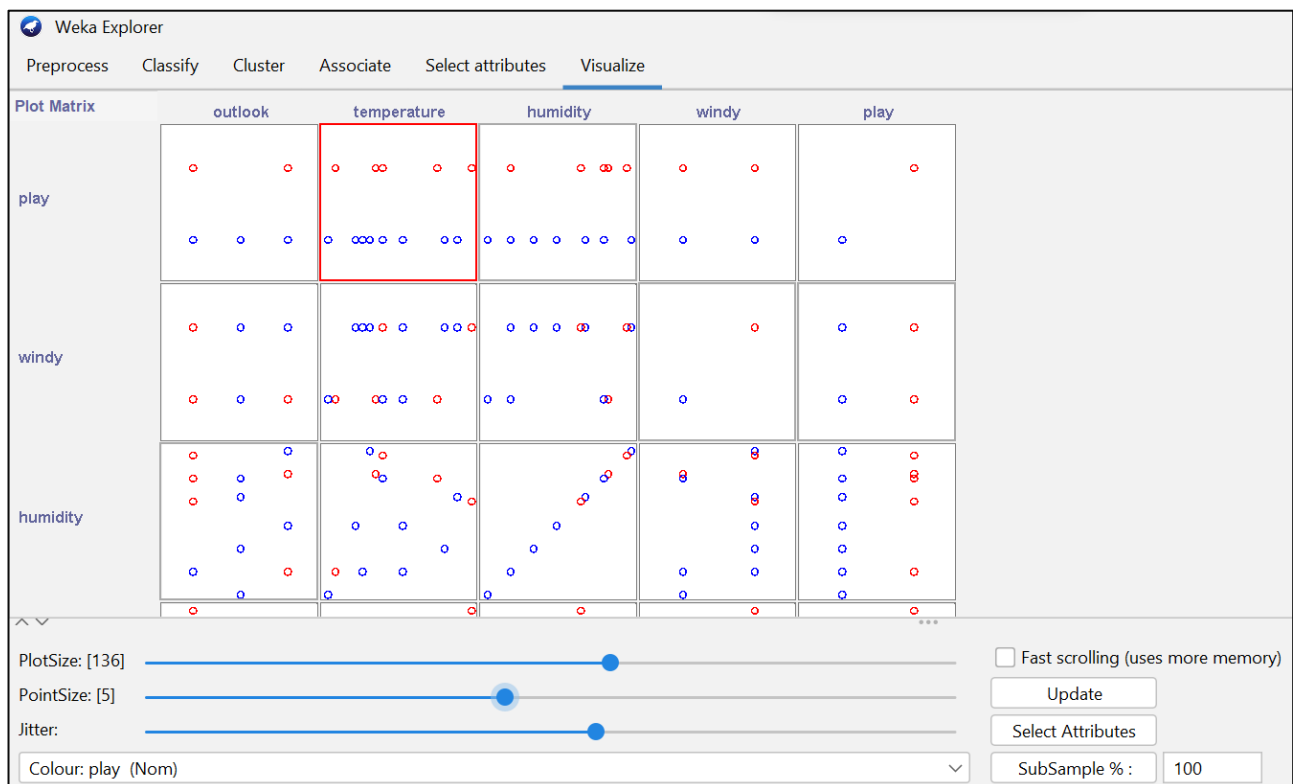


Image 13: Visualize "Weather data set"

Thereafter, no more attribute pairs from this table are related because it displays every relationship between each pair of attributes in full.

2.3 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:

– What is the content of the comments section in credit-g.arff (when opened with any text editor) about?

The information in the dataset was described in the file's comment.

For algorithms that need numerical properties, Strathclyde University generated the file "german.data-numeric." To make it suitable for algorithms that cannot handle category variables, this file has been changed and a lot of indicator variables added. Attribute 17 is one of the ordered categorical qualities that has been coded as an integer. Similar to this form was used by StatLog.

– How many samples does the data set have?

→ The data set have 1000 samples.

– How many attributes?

→ This data set have 21 attributes.

– Describe any five attributes (must have both discrete and continuous attributes).

1. checking-status: Status of existing checking account.
2. purpose:
3. duration: Installment rate in percentage of disposable income
4. credit amount:
5. installment_commitment: Installment rate in percentage of disposable income.

– Which attribute is used for the label?

Attribute is used for the label is: Class.

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

Continuous attributes: credit amount, installment_commitment, residence_since, age, existing_credits, num_dependents.

Let's look at how the continuous attribute "age" is distributed. The distribution showed left-skewed data in perspective.

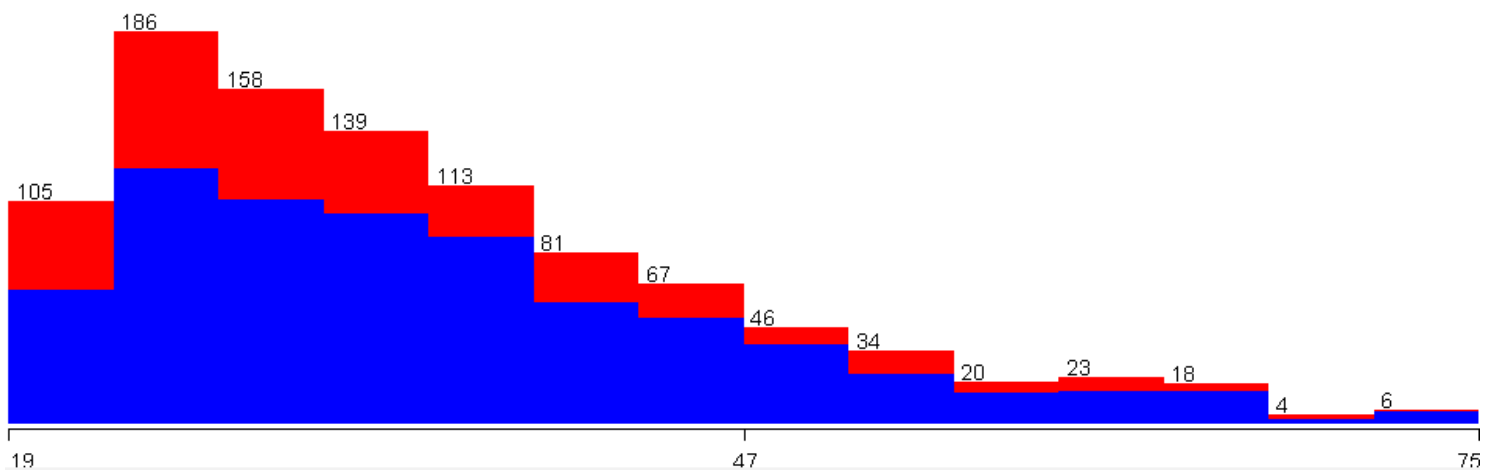


Image 14: Age chart

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

The group of bar charts below displays the distributions of each attribute along with their connections to the label one. It is clear that this dataset's distribution deviates significantly from the norm and is generally biased to the left. Moreover, a number of characteristics, some of which exhibit the Bernoulli distribution, result in an unbalanced distribution.

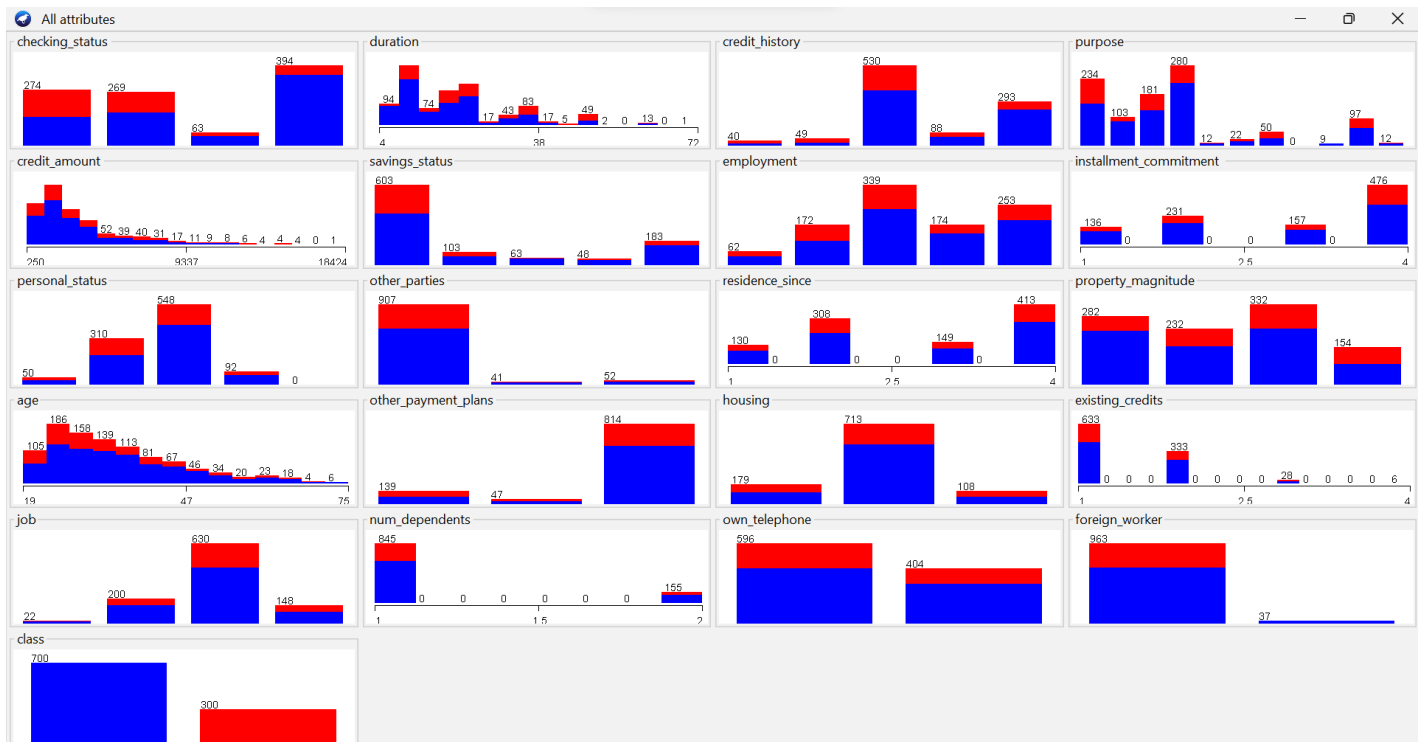


Image 15: All attribute distribution and its relationship to label one

– Let's move to Select attributes tag. Describe all of the options for attribute selection.

In the Select Attribute, which would process the feature selections measure with several principal options.

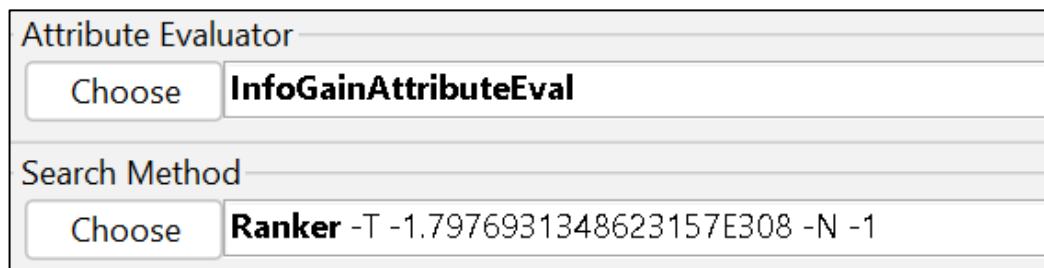
Attribute Evaluator: An algorithm, such as PCA, would be used to categorize the attributes.

Search Method: The technique that would be applied to the search, such as BestFirst.

Attribute Selection Mode: The options that apply Cross Validation for the feature selection operation or training specifically on the training set are referred to as attribute selection mode.

– 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)

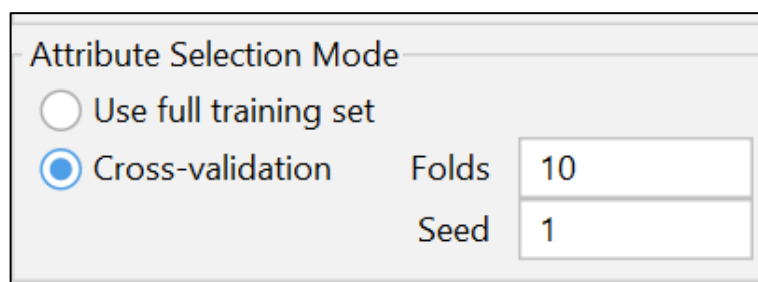
First, we would select the Ranker for the Method and the CorrelationAttributeEval for the Evaluator.



The screenshot shows a software interface with two sections. The top section is titled 'Attribute Evaluator' and contains a 'Choose' button followed by the text 'InfoGainAttributeEval'. The bottom section is titled 'Search Method' and contains a 'Choose' button followed by the text 'Ranker -T -1.7976931348623157E308 -N -1'.

Image 16: Option for Evaluator and Method

Second, in order to eliminate bias in the process and increase practical accuracy, we use the CrossValidation selection mode (10 folds per step).



The screenshot shows a software interface titled 'Attribute Selection Mode'. It has two radio buttons: 'Use full training set' (unselected) and 'Cross-validation' (selected). To the right of the 'Cross-validation' option, there are two input fields: 'Folds' with the value '10' and 'Seed' with the value '1'.

Image 17: Attribute selection mode

Lastly, we select the label attribute as the comparison's anchor attribute (class). then select "start" to obtain the outcome.

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===
```

average merit	average rank	attribute
0.095 +- 0.006	1 +- 0	1 checking_status
0.044 +- 0.003	2.1 +- 0.3	3 credit_history
0.033 +- 0.006	3.5 +- 0.92	2 duration
0.029 +- 0.004	3.9 +- 0.83	6 savings_status
0.026 +- 0.002	4.5 +- 0.5	4 purpose
0.019 +- 0.002	6.2 +- 0.4	5 credit_amount
0.017 +- 0.002	7.1 +- 0.7	12 property_magnitude
0.013 +- 0.001	8.5 +- 0.92	7 employment
0.013 +- 0.001	9 +- 0.45	15 housing
0.009 +- 0.002	10.4 +- 1.28	14 other_payment_plans
0.007 +- 0.001	11.7 +- 1.1	9 personal_status
0.006 +- 0.002	12 +- 1	20 foreign_worker
0.005 +- 0.001	13.1 +- 0.83	10 other_parties
0.002 +- 0.001	14.6 +- 0.66	17 job
0.006 +- 0.007	14.7 +- 4.8	13 age
0.001 +- 0	15.2 +- 0.6	19 own_telephone
0 +- 0	16.6 +- 0.92	18 num_dependents
0 +- 0	17.6 +- 0.66	8 installment_commitment
0 +- 0	18.6 +- 0.66	11 residence_since
0 +- 0	19.7 +- 0.9	16 existing_credits

Image 18: result table

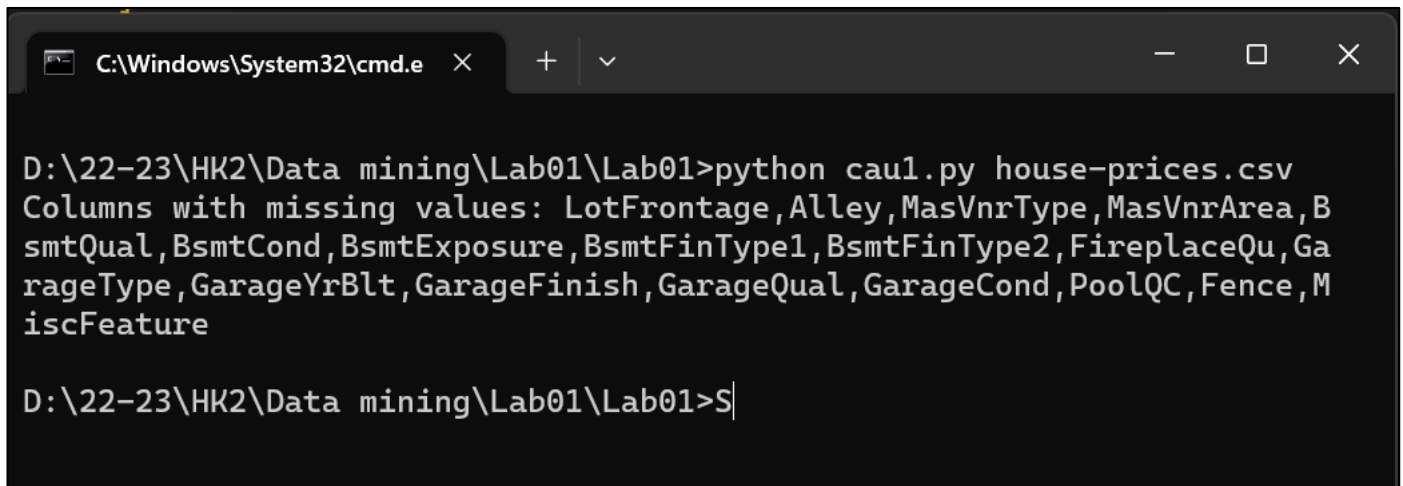
Conclusion: The top 5 positions are obtained from the results above. "checking_status," "credit_history," "duration," "savings_status," and "purpose" were the attributes with the highest relationship.

3 PREPROCESSING DATA IN PYTHON

The program must have the following functions.

3.1 Extract columns with missing values

Command line: python cau1.py house-prices.csv



```
C:\Windows\System32\cmd.e  X  +  v  -  □  X

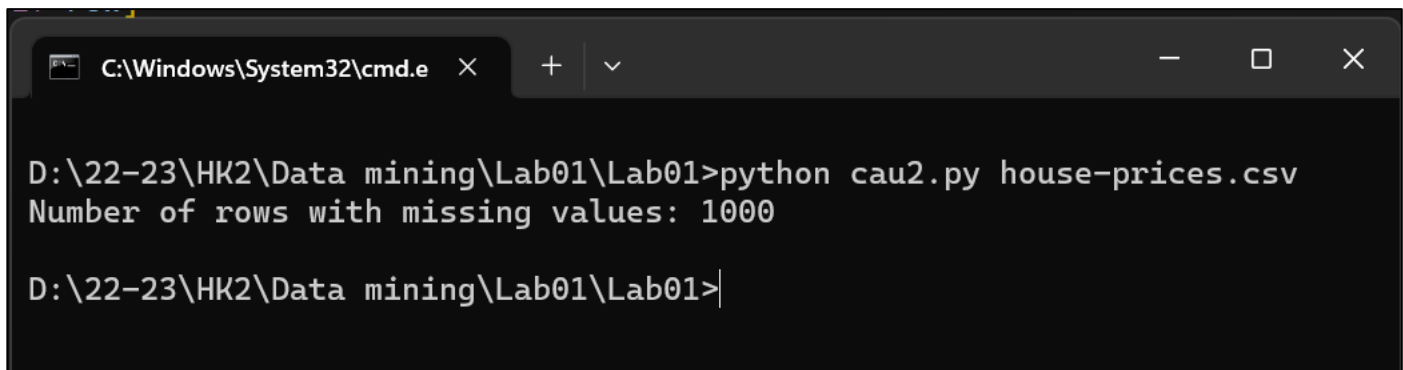
D:\22-23\HK2\Data mining\Lab01\Lab01>python cau1.py house-prices.csv
Columns with missing values: LotFrontage,Alley,MasVnrType,MasVnrArea,B
smtQual,BsmtCond,BsmtExposure,BsmtFinType1,BsmtFinType2,FireplaceQu,Ga
rageType,GarageYrBlt,GarageFinish,GarageQual,GarageCond,PoolQC,Fence,M
iscFeature

D:\22-23\HK2\Data mining\Lab01\Lab01>S|
```

Image 19: command line and results of extracting columns with missing values

3.2 Count the number of lines with missing data.

Command line: python cau2.py house-prices.csv



```
C:\Windows\System32\cmd.e  X  +  v  -  □  X

D:\22-23\HK2\Data mining\Lab01\Lab01>python cau2.py house-prices.csv
Number of rows with missing values: 1000

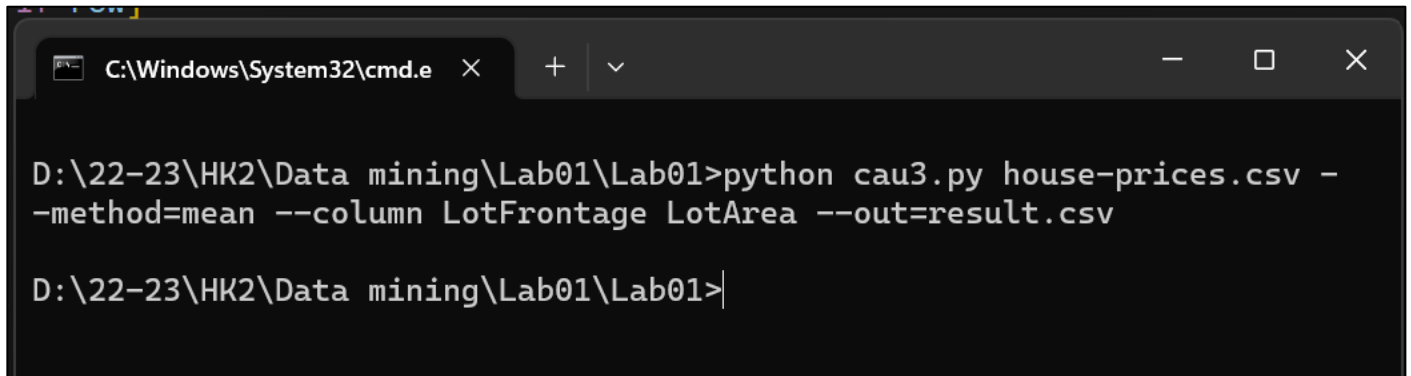
D:\22-23\HK2\Data mining\Lab01\Lab01>|
```

Image 20: command line and results of counting the number of lines with missing data

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

3.3.1 Using mean

Command line: python cau3.py house-prices.csv --method=mean --column LotFrontage LotArea --out=result.csv

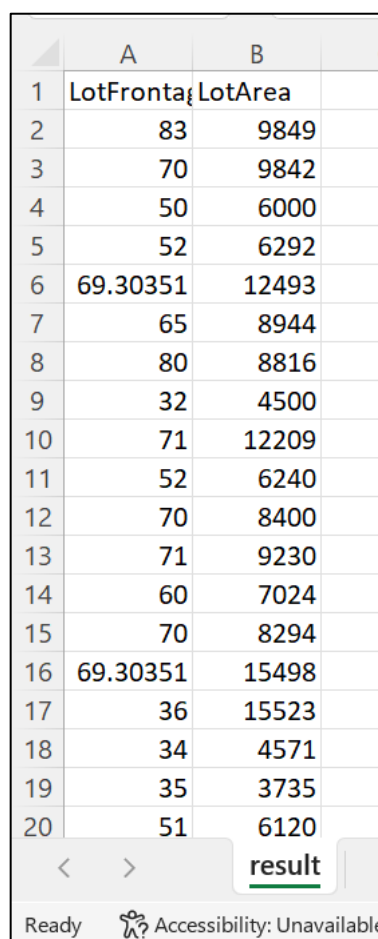


```
C:\Windows\System32\cmd.e x + v

D:\22-23\HK2\Data mining\Lab01\Lab01>python cau3.py house-prices.csv -
-method=mean --column LotFrontage LotArea --out=result.csv

D:\22-23\HK2\Data mining\Lab01\Lab01>
```

Image 21: command line to use mean value

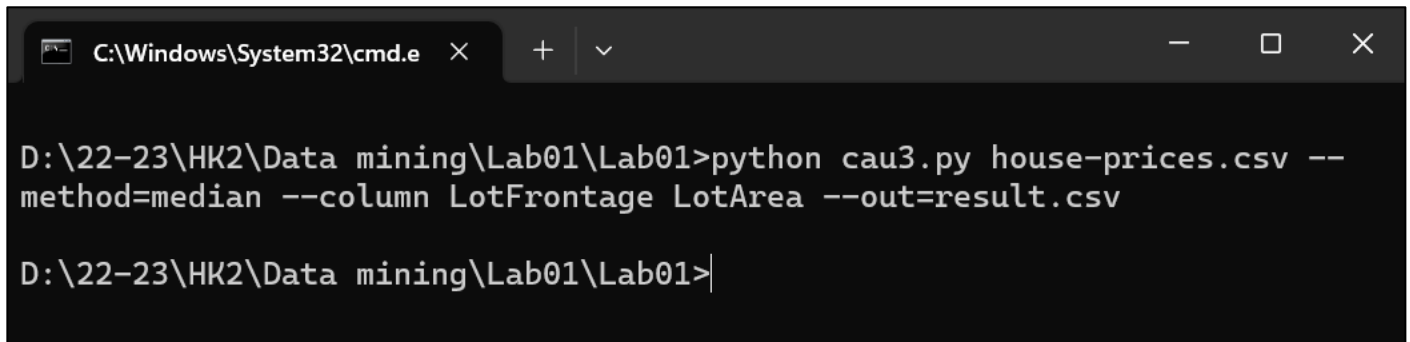


	A	B	C
1	LotFrontage	LotArea	
2	83	9849	
3	70	9842	
4	50	6000	
5	52	6292	
6	69.30351	12493	
7	65	8944	
8	80	8816	
9	32	4500	
10	71	12209	
11	52	6240	
12	70	8400	
13	71	9230	
14	60	7024	
15	70	8294	
16	69.30351	15498	
17	36	15523	
18	34	4571	
19	35	3735	
20	51	6120	

Image 22: result when using mean value

3.3.2 Using median

Command line: python cau3.py house-prices.csv --method=median --column LotFrontage LotArea --out=result.csv



```
C:\Windows\System32\cmd.e x + v - □ ×  
  
D:\22-23\HK2\Data mining\Lab01\Lab01>python cau3.py house-prices.csv --  
method=median --column LotFrontage LotArea --out=result.csv  
  
D:\22-23\HK2\Data mining\Lab01\Lab01>
```

Image 23: command line to use median value

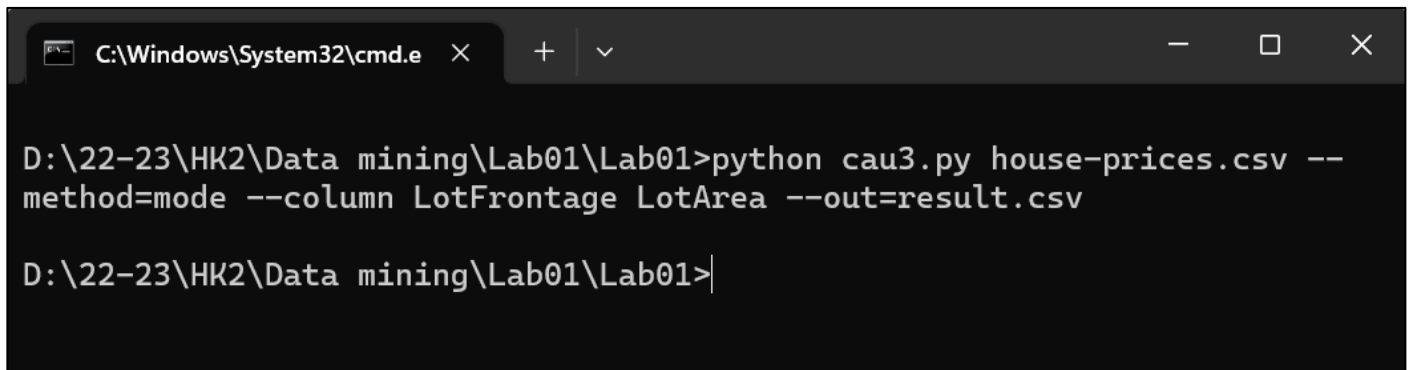
	A	B
1	LotFrontage	LotArea
2	83	9849
3	70	9842
4	50	6000
5	52	6292
6	68	12493
7	65	8944
8	80	8816
9	32	4500
10	71	12209
11	52	6240
12	70	8400
13	71	9230
14	60	7024
15	70	8294
16	68	15498
17	36	15523
18	34	4571
19	35	3735
20	51	6120

< > result

Image 24: result when using median value

3.3.3 Using mode

Command line: python cau3.py house-prices.csv --method=mode --column LotFrontage LotArea --out=result.csv



A screenshot of a Windows command prompt window. The title bar shows the path 'C:\Windows\System32\cmd.e' and standard window controls. The command prompt displays the following text:

```
D:\22-23\HK2\Data mining\Lab01\Lab01>python cau3.py house-prices.csv --method=mode --column LotFrontage LotArea --out=result.csv
```

The cursor is positioned at the end of the second line, ready for another command.

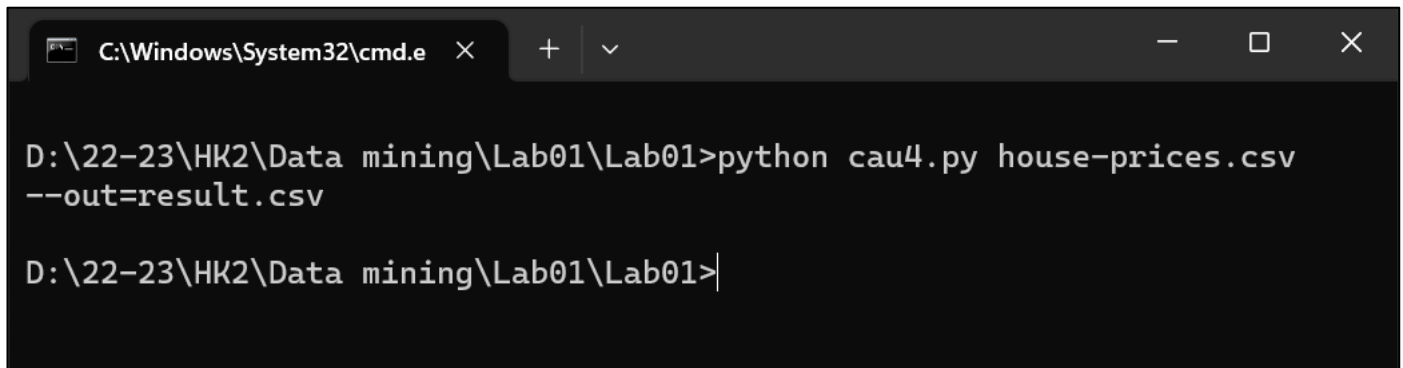
Image 25: command line to use mode value

	A	B
1	LotFrontage	LotArea
2	83	9849
3	70	9842
4	50	6000
5	52	6292
6		12493
7	65	8944
8	80	8816
9	32	4500
10	71	12209
11	52	6240
12	70	8400
13	71	9230
14	60	7024
15	70	8294
16		15498
17	36	15523
18	34	4571
19	35	3735
20	51	6120

Image 26: result when using mode value

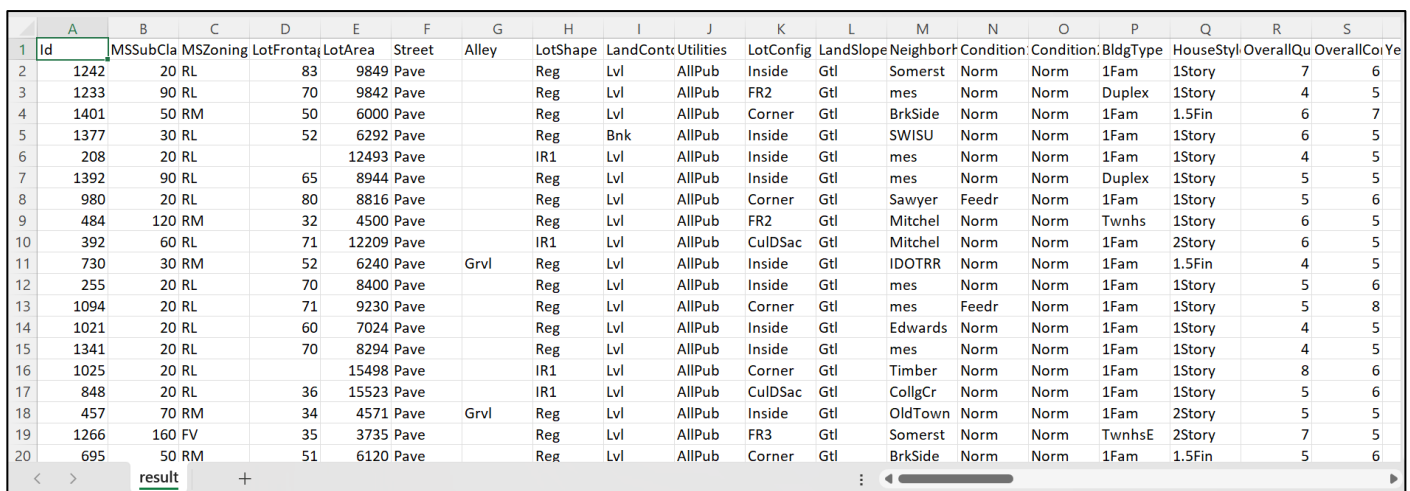
3.4 Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes).

Command line: python cau4.py house-prices.csv --out=result.csv



```
C:\Windows\System32\cmd.e X + v
D:\22-23\HK2\Data mining\Lab01\Lab01>python cau4.py house-prices.csv
--out=result.csv
D:\22-23\HK2\Data mining\Lab01\Lab01>
```

Image 27: command line to delete rows containing more than a particular number of missing values

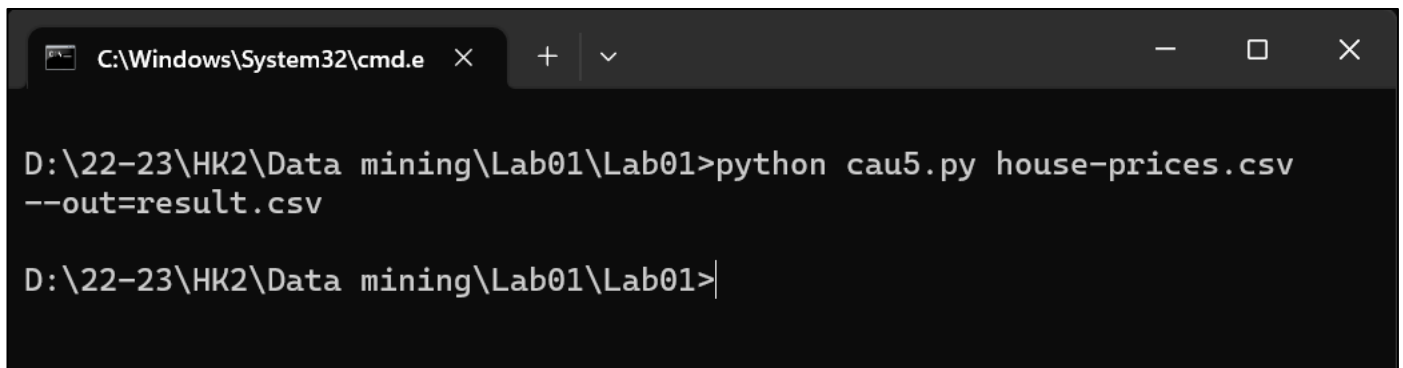


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Id	MSSubCla	MSZoning	LotFrontaj	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighbort	Condition	Condition	BldgType	HouseStyl	OverallQu	OverallCo
2	1242	20	RL	83	9849	Pave		Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6
3	1233	90	RL	70	9842	Pave		Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5
4	1401	50	RM	50	6000	Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7
5	1377	30	RL	52	6292	Pave		Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5
6	208	20	RL		12493	Pave		IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5
7	1392	90	RL	65	8944	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5
8	980	20	RL	80	8816	Pave		Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6
9	484	120	RM	32	4500	Pave		Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5
10	392	60	RL	71	12209	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5
11	730	30	RM	52	6240	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5
12	255	20	RL	70	8400	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6
13	1094	20	RL	71	9230	Pave		Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8
14	1021	20	RL	60	7024	Pave		Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5
15	1341	20	RL	70	8294	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5
16	1025	20	RL		15498	Pave		IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6
17	848	20	RL	36	15523	Pave		IR1	Lvl	AllPub	CulDSac	Gtl	CollCr	Norm	Norm	1Fam	1Story	5	6
18	457	70	RM	34	4571	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5
19	1266	160	FV	35	3735	Pave		Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5
20	695	50	RM	51	6120	Pave		Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6

Image 28: result when using function to delete rows containing more than a particular number of missing values

3.5 Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples).

Command line: python cau5.py house-prices.csv --out=result.csv

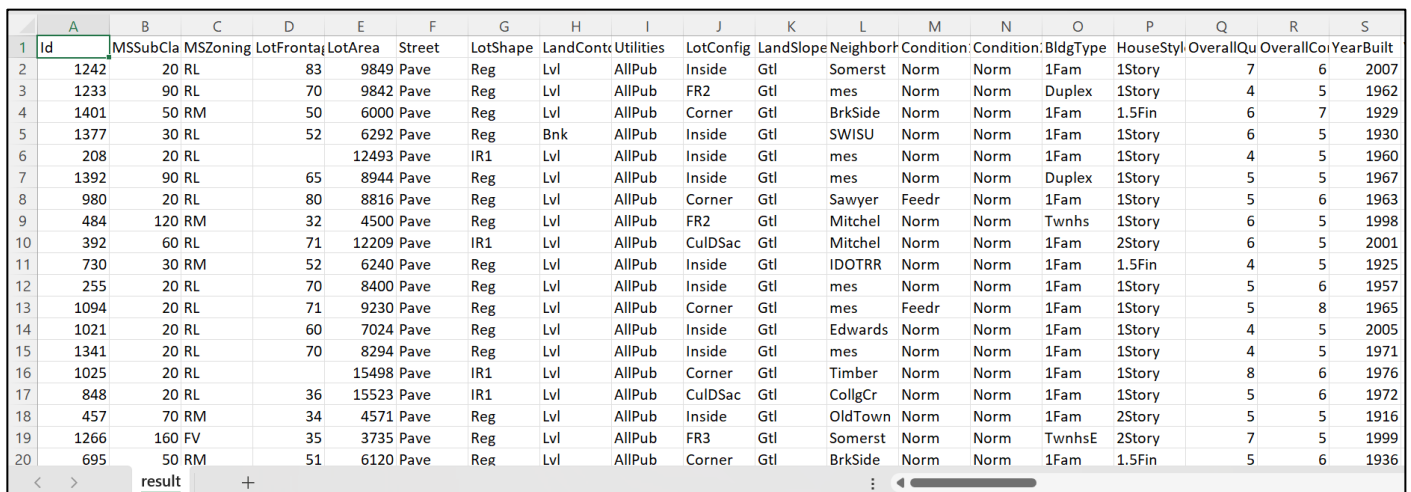


```
C:\Windows\System32\cmd.e X + v

D:\22-23\HK2\Data mining\Lab01\Lab01>python cau5.py house-prices.csv
--out=result.csv

D:\22-23\HK2\Data mining\Lab01\Lab01>
```

Image 29: command line to delete columns containing more than a particular number of missing values

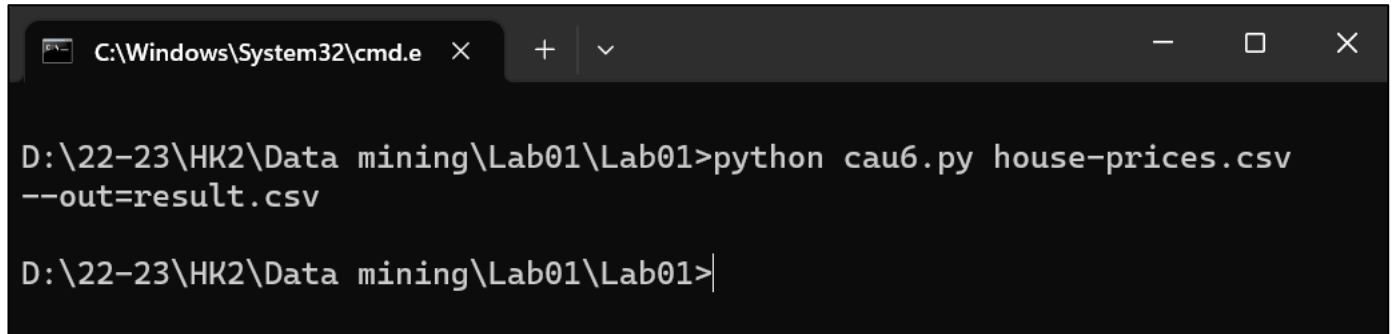


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Id	MSSubCla	MSZoning	LotFrontaj	LotArea	Street	LotShape	LandContx	Utilities	LotConfig	LandSlope	Neighbort	Condition	Condition	BldgType	HouseStyl	OverallQu	OverallCoi	YearBuilt
2	1242	20	RL	83	9849	Pave	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	7	6	2007
3	1233	90	RL	70	9842	Pave	Reg	Lvl	AllPub	FR2	Gtl	mes	Norm	Norm	Duplex	1Story	4	5	1962
4	1401	50	RM	50	6000	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	6	7	1929
5	1377	30	RL	52	6292	Pave	Reg	Bnk	AllPub	Inside	Gtl	SWISU	Norm	Norm	1Fam	1Story	6	5	1930
6	208	20	RL		12493	Pave	IR1	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1960
7	1392	90	RL	65	8944	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	Duplex	1Story	5	5	1967
8	980	20	RL	80	8816	Pave	Reg	Lvl	AllPub	Corner	Gtl	Sawyer	Feedr	Norm	1Fam	1Story	5	6	1963
9	484	120	RM	32	4500	Pave	Reg	Lvl	AllPub	FR2	Gtl	Mitchel	Norm	Norm	Twnhs	1Story	6	5	1998
10	392	60	RL	71	12209	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	Mitchel	Norm	Norm	1Fam	2Story	6	5	2001
11	730	30	RM	52	6240	Pave	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Fin	4	5	1925
12	255	20	RL	70	8400	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	6	1957
13	1094	20	RL	71	9230	Pave	Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8	1965
14	1021	20	RL	60	7024	Pave	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	4	5	2005
15	1341	20	RL	70	8294	Pave	Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	4	5	1971
16	1025	20	RL		15498	Pave	IR1	Lvl	AllPub	Corner	Gtl	Timber	Norm	Norm	1Fam	1Story	8	6	1976
17	848	20	RL	36	15523	Pave	IR1	Lvl	AllPub	CulDSac	Gtl	CollgCr	Norm	Norm	1Fam	1Story	5	6	1972
18	457	70	RM	34	4571	Pave	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	2Story	5	5	1916
19	1266	160	FV	35	3735	Pave	Reg	Lvl	AllPub	FR3	Gtl	Somerst	Norm	Norm	TwnhsE	2Story	7	5	1999
20	695	50	RM	51	6120	Pave	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Fin	5	6	1936

Image 28: result when using function to delete columns containing more than a particular number of missing values

3.6 Delete duplicate samples.

Command line: python cau6.py house-prices.csv --out=result.csv

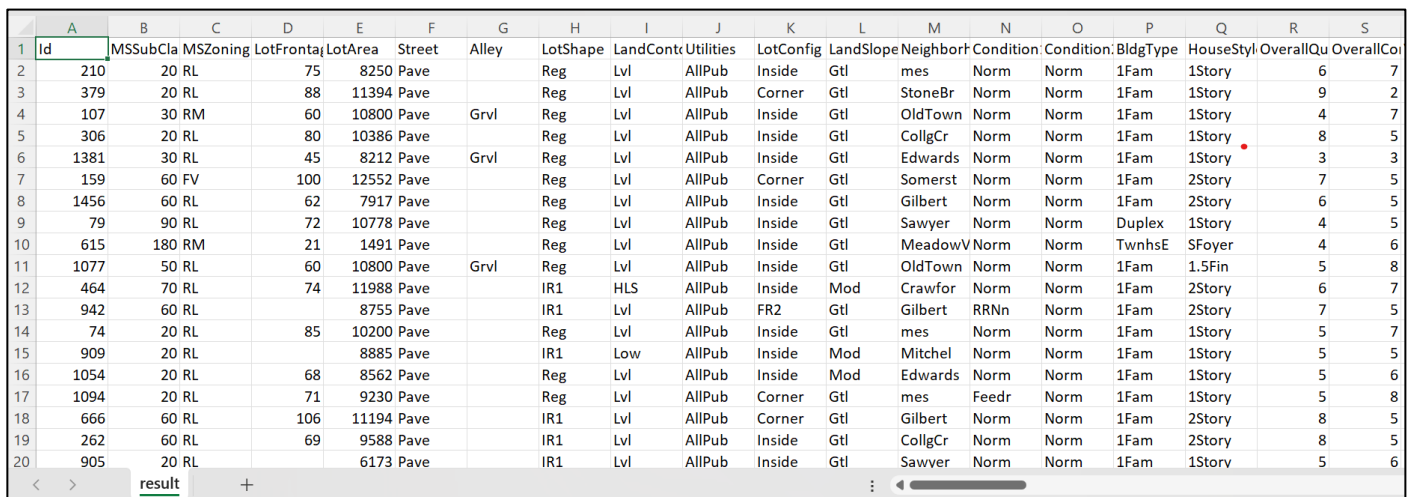


```
C:\Windows\System32\cmd.e X + v - □ X

D:\22-23\HK2\Data mining\Lab01\Lab01>python cau6.py house-prices.csv
--out=result.csv

D:\22-23\HK2\Data mining\Lab01\Lab01>
```

Image 29: command line to delete duplicate samples



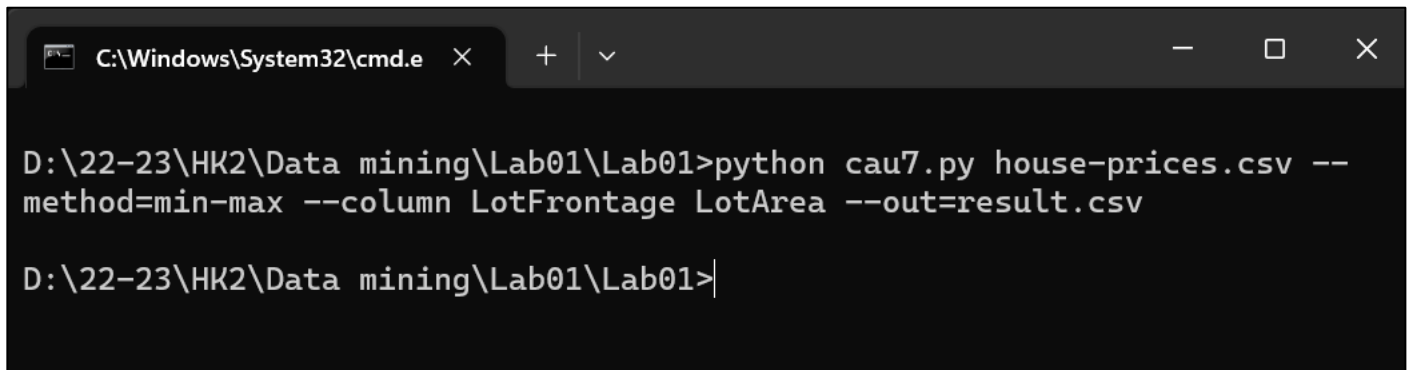
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Id	MSSubCla	MSZoning	LotFrontaj	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborf	Condition	Condition	BldgType	HouseStyl	OverallQu	OverallCoi
2	210	20 RL		75	8250	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	6	7
3	379	20 RL		88	11394	Pave		Reg	Lvl	AllPub	Corner	Gtl	StoneBr	Norm	Norm	1Fam	1Story	9	2
4	107	30 RM		60	10800	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	1Story	4	7
5	306	20 RL		80	10386	Pave		Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	8	5
6	1381	30 RL		45	8212	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Story	3	3
7	159	60 FV		100	12552	Pave		Reg	Lvl	AllPub	Corner	Gtl	Somerst	Norm	Norm	1Fam	2Story	7	5
8	1456	60 RL		62	7917	Pave		Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	6	5
9	79	90 RL		72	10778	Pave		Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	Duplex	1Story	4	5
10	615	180 RM		21	1491	Pave		Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	SFoyer	4	6
11	1077	50 RL		60	10800	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam	1.5Fin	5	8
12	464	70 RL		74	11988	Pave		IR1	HLS	AllPub	Inside	Mod	Crawfor	Norm	Norm	1Fam	2Story	6	7
13	942	60 RL			8755	Pave		IR1	Lvl	AllPub	FR2	Gtl	Gilbert	RRNn	Norm	1Fam	2Story	7	5
14	74	20 RL		85	10200	Pave		Reg	Lvl	AllPub	Inside	Gtl	mes	Norm	Norm	1Fam	1Story	5	7
15	909	20 RL			8885	Pave		IR1	Low	AllPub	Inside	Mod	Mitchel	Norm	Norm	1Fam	1Story	5	5
16	1054	20 RL		68	8562	Pave		Reg	Lvl	AllPub	Inside	Mod	Edwards	Norm	Norm	1Fam	1Story	5	6
17	1094	20 RL		71	9230	Pave		Reg	Lvl	AllPub	Corner	Gtl	mes	Feedr	Norm	1Fam	1Story	5	8
18	666	60 RL		106	11194	Pave		IR1	Lvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	1Fam	2Story	8	5
19	262	60 RL		69	9588	Pave		IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	8	5
20	905	20 RL			6173	Pave		IR1	Lvl	AllPub	Inside	Gtl	Sawver	Norm	Norm	1Fam	1Story	5	6

Image 30: result when deleting duplicate samples

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

3.7.1 min-max method

Command line: python cau7.py house-prices.csv --method=min-max --column LotFrontage LotArea --out=result.csv



```
C:\Windows\System32\cmd.e x + v - □ X

D:\22-23\HK2\Data mining\Lab01\Lab01>python cau7.py house-prices.csv --method=min-max --column LotFrontage LotArea --out=result.csv

D:\22-23\HK2\Data mining\Lab01\Lab01>
```

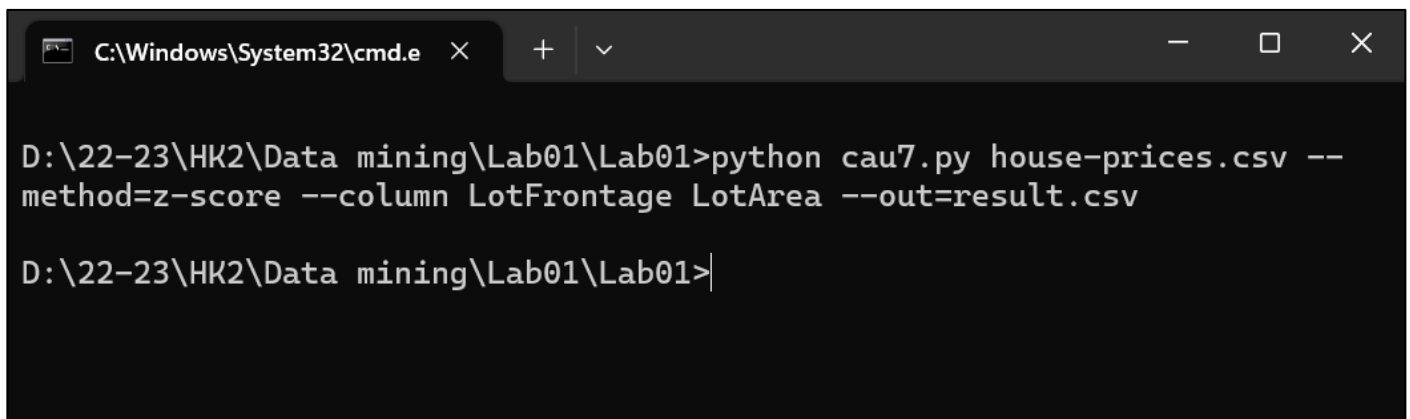
Image 31: command line to normalize a numeric attribute using min-max methods

	A	B
1	LotFrontage	LotArea
2	0.542484	0.039164
3	0.457516	0.039131
4	0.326797	0.021158
5	0.339869	0.022524
6	0.339869	0.051533
7	0.424837	0.03493
8	0.522876	0.034332
9	0.20915	0.014141
10	0.464052	0.050204
11	0.339869	0.022281
12	0.457516	0.032386
13	0.464052	0.036268
14	0.392157	0.025949
15	0.457516	0.03189
16	0.457516	0.06559
17	0.235294	0.065707
18	0.222222	0.014474
19	0.228758	0.010563
20	0.333333	0.02172

Image 32: result when normalizing a numeric attribute using min-max methods

3.7.2 z-score method

Command line: `python cau7.py house-prices.csv --method=z-score --column LotFrontage LotArea --out=result.csv`



```
C:\Windows\System32\cmd.e x + v  
D:\22-23\HK2\Data mining\Lab01\Lab01>python cau7.py house-prices.csv --  
method=z-score --column LotFrontage LotArea --out=result.csv  
D:\22-23\HK2\Data mining\Lab01\Lab01>
```

Image 33: command line to normalize a numeric attribute using z-score methods

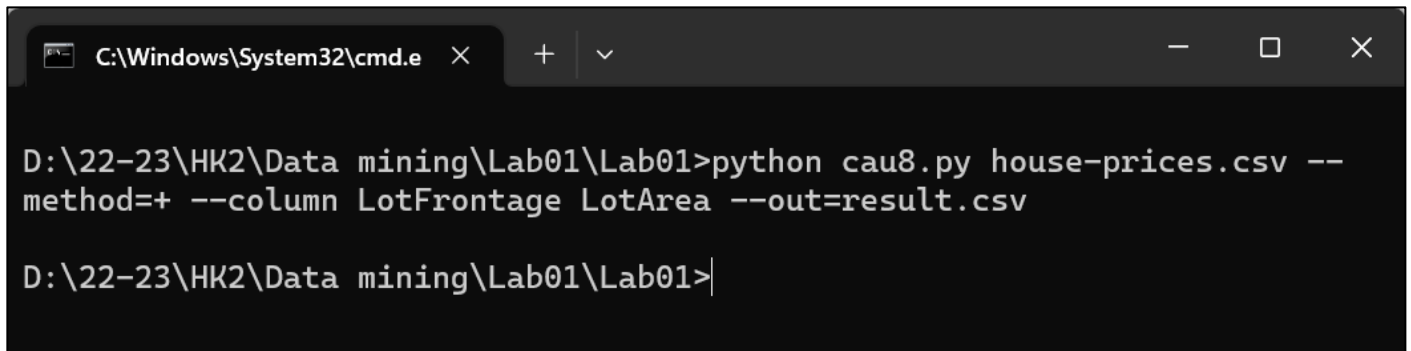
	A	B
1	LotFrontage	LotArea
2	0.788584657	-0.03953
3	0.389472279	-0.04029
4	-0.224546764	-0.45877
5	-0.16314486	-0.42697
6		0.248468
7	0.235967518	-0.1381
8	0.6964818	-0.15205
9	-0.777163903	-0.62216
10	0.420173231	0.217534
11	-0.16314486	-0.43263
12	0.389472279	-0.19736
13	0.420173231	-0.10695
14	0.082462757	-0.34724
15	0.389472279	-0.2089
16		0.575784
17	-0.654360094	0.578508
18	-0.715761998	-0.61443

Image 34: result when normalizing a numeric attribute using z-score methods

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

3.8.1 addition

Command line: python cau8.py house-prices.csv --method=+ --column LotFrontage LotArea --out=result.csv



```
C:\Windows\System32\cmd.e x + v
D:\22-23\HK2\Data mining\Lab01\Lab01>python cau8.py house-prices.csv --method=+ --column LotFrontage LotArea --out=result.csv
D:\22-23\HK2\Data mining\Lab01\Lab01>
```

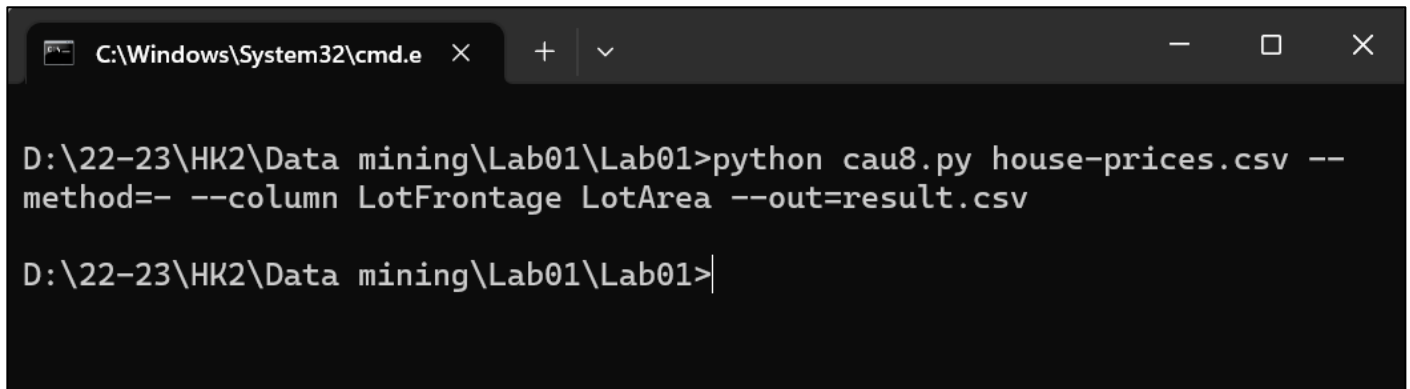
Image 35: command line to perform addition

	A	B	C
1	LotFrontage	LotArea	result
2	83	9849	9932
3	70	9842	9912
4	50	6000	6050
5	52	6292	6344
6		12493	
7	65	8944	9009
8	80	8816	8896
9	32	4500	4532
10	71	12209	12280
11	52	6240	6292
12	70	8400	8470
13	71	9230	9301
14	60	7024	7084
15	70	8294	8364
16		15498	
17	36	15523	15559
18	34	4571	4605
19	35	3735	3770
20	51	6120	6171

Image 36: result when performing addition

3.8.2 subtraction

Command line: python cau8.py house-prices.csv --method=- --column LotFrontage LotArea --out=result.csv



```
C:\Windows\System32\cmd.e x + v  
  
D:\22-23\HK2\Data mining\Lab01\Lab01>python cau8.py house-prices.csv --  
method=- --column LotFrontage LotArea --out=result.csv  
  
D:\22-23\HK2\Data mining\Lab01\Lab01>
```

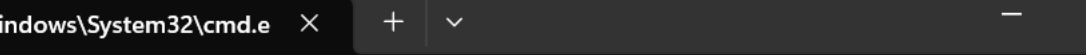
Image 37: command line to perform subtraction

	A	B	C
1	LotFrontage	LotArea	result
2	83	9849	-9766
3	70	9842	-9772
4	50	6000	-5950
5	52	6292	-6240
6		12493	
7	65	8944	-8879
8	80	8816	-8736
9	32	4500	-4468
10	71	12209	-12138
11	52	6240	-6188
12	70	8400	-8330
13	71	9230	-9159
14	60	7024	-6964
15	70	8294	-8224
16		15498	
17	36	15523	-15487
18	34	4571	-4537
19	35	3735	-3700
20	51	6120	-6069

Image 38: result when performing subtraction

3.8.3 multiplication

Command line: python cau8.py house-prices.csv --method=* --column LotFrontage LotArea --out=result.csv



The screenshot shows a Windows command prompt window with the title bar "C:\Windows\System32\cmd.e". The command prompt displays the following text:

```
D:\22-23\HK2\Data mining\Lab01\Lab01>python cau8.py house-prices.csv -  
-method=* --column LotFrontage LotArea --out=result.csv  
  
D:\22-23\HK2\Data mining\Lab01\Lab01>
```

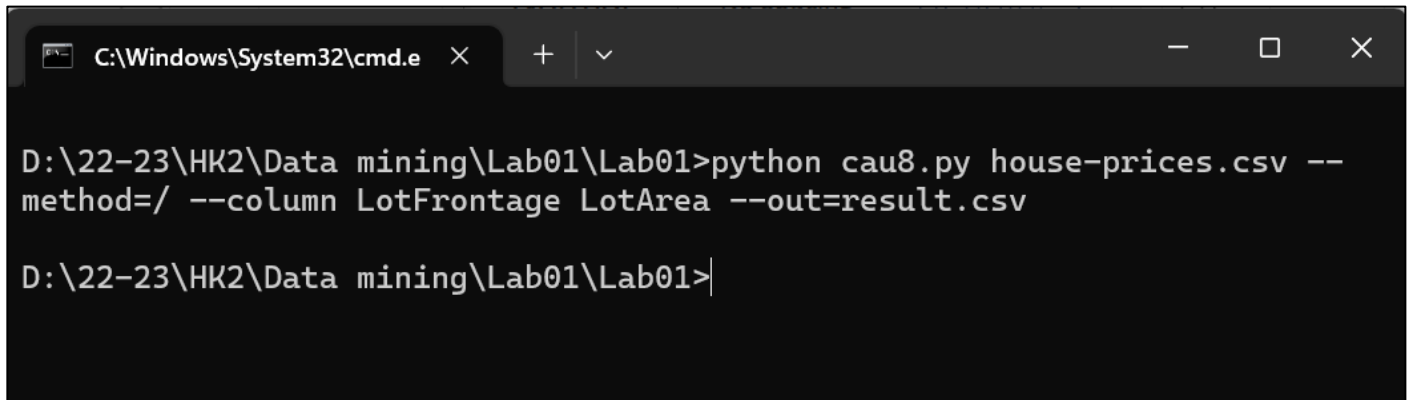
Image 39: command line to perform multiplication

	A	B	C
1	LotFrontal	LotArea	result
2	83	9849	817467
3	70	9842	688940
4	50	6000	300000
5	52	6292	327184
6		12493	
7	65	8944	581360
8	80	8816	705280
9	32	4500	144000
10	71	12209	866839
11	52	6240	324480
12	70	8400	588000
13	71	9230	655330
14	60	7024	421440
15	70	8294	580580
16		15498	
17	36	15523	558828
18	34	4571	155414

Image 40: result when performing multiplication

3.8.4 division

Command line: python cau8.py house-prices.csv --method=/ --column LotFrontage LotArea --out=result.csv



```
C:\Windows\System32\cmd.e x + v
D:\22-23\HK2\Data mining\Lab01\Lab01>python cau8.py house-prices.csv --
method=/ --column LotFrontage LotArea --out=result.csv
D:\22-23\HK2\Data mining\Lab01\Lab01>
```

Image 41: command line to perform division

	A	B	C
1	LotFrontage	LotArea	result
2	83	9849	0.008427
3	70	9842	0.007112
4	50	6000	0.008333
5	52	6292	0.008264
6		12493	
7	65	8944	0.007267
8	80	8816	0.009074
9	32	4500	0.007111
10	71	12209	0.005815
11	52	6240	0.008333
12	70	8400	0.008333
13	71	9230	0.007692
14	60	7024	0.008542
15	70	8294	0.00844
16		15498	
17	36	15523	0.002319
18	34	4571	0.007438

Image 42: result when performing division

4. REFERENCE

[1] Lecture slides

[2] <https://www.cs.waikato.ac.nz/ml/weka/>

[3] Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 2: Data Preprocessing

[4] Using machine learning techniques to predict the recurrence of breast cancer – LinkedIn

<https://www.linkedin.com/pulse/using-machine-learning-techniques-predict-recurrence-breast-alva>

[5] German Credit Risk Analysis – Hemang Goswami

https://rstudio-pubs-static.s3.amazonaws.com/368878_a277debe13724075a69189451dadd751.html