DATA MINING

ASSIGNMENT-1

VISUALIZATION AND ANALYSIS OF THE GIVEN DATASET USING WEKA

Prof: Dr Elizabeth D Diaz

TEAM MEMBERS

Harshini Kandimalla – 1001960046

Pratik Antoni Patekar – 1001937948

Pratik Dhanraj Chavan – 1001963580

Introduction:

Weka (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. Weka is open-source software issued under the **GNU General Public License**.

Weka supports **deep learning**!

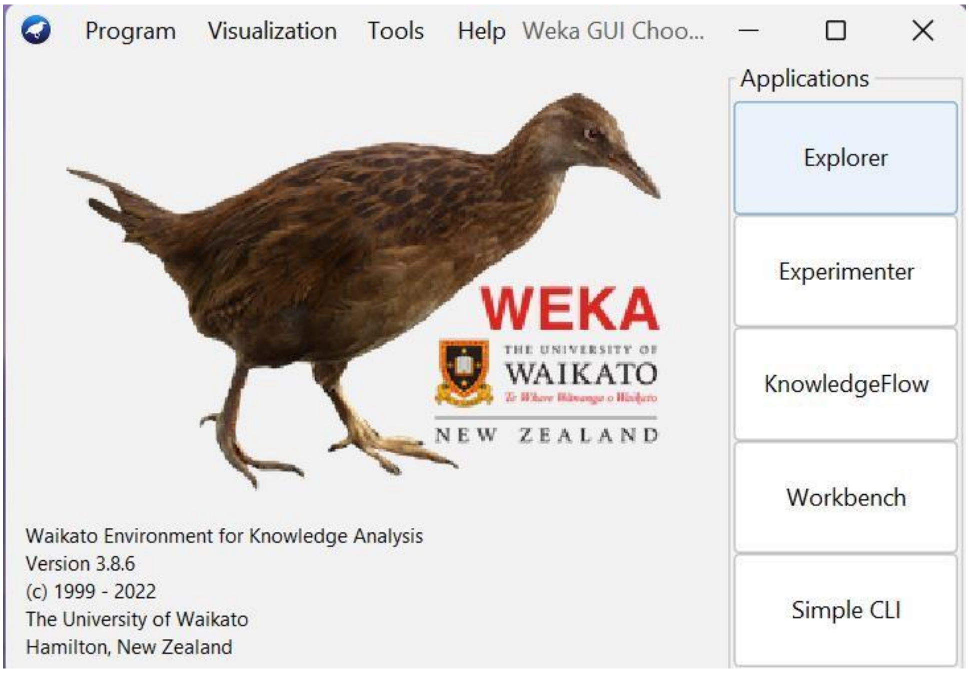
In this, we have implemented visualization technique on Housing dataset.

Retrieving Data:

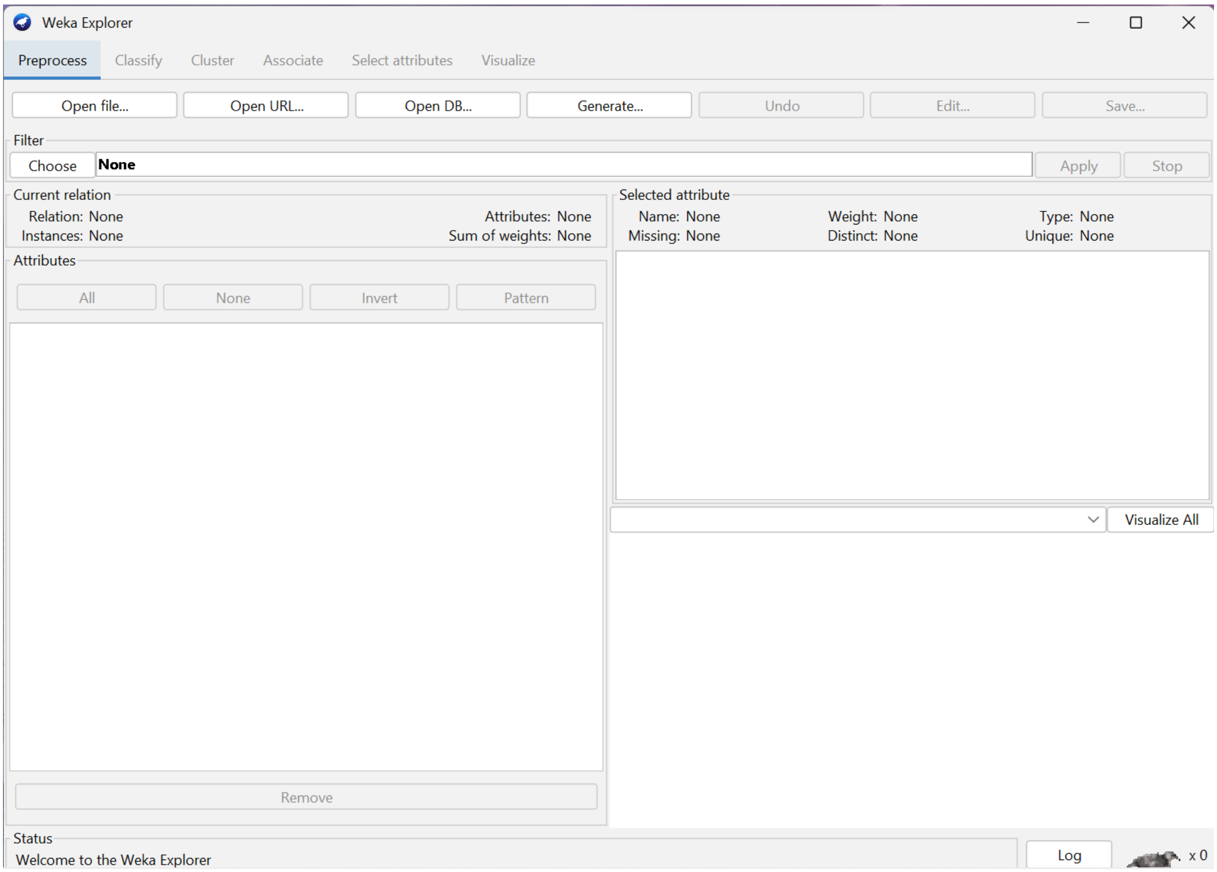
Data must be loaded into weka. Dataset can be either csv file or arfftype. If we csv file, it can be uploaded directly or can be converted into arff type and then load and likewise.

1. Start the weka GUI chooser

2. Launch the weeka explorer by clicking on explorer button



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



Current relation

Relation: None

Instances: None

Attributes: None

Sum of weights: None

Selected attribute

Name: None

Missing: None

Weight: None

Distinct: None

Type: None

Unique: None

Attributes

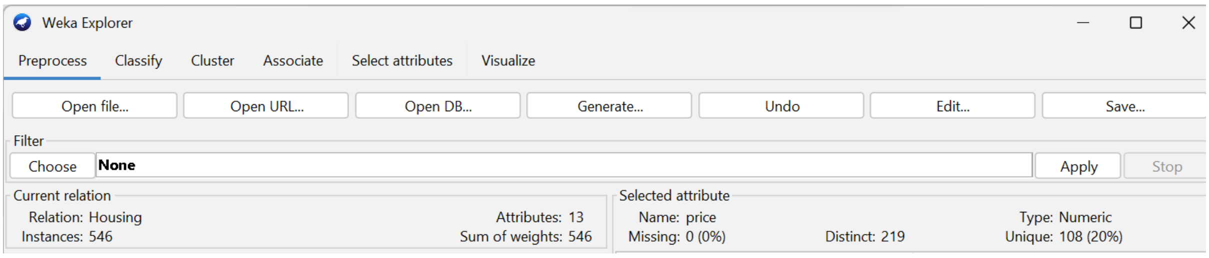
Status

Welcome to the Weka Explorer

3. Now click on the open file to and it navigates to your directory and you can select the dataset to be used either in csv form or arff type.

Glimpse of Data:

Once the file is opened in weka, we can check the data by clicking on the edit button in preprocessing tab.



Current relation

Relation: Housing

Instances: 546

Attributes: 13

Sum of weights: 546

Selected attribute

Name: price

Missing: 0 (0%)

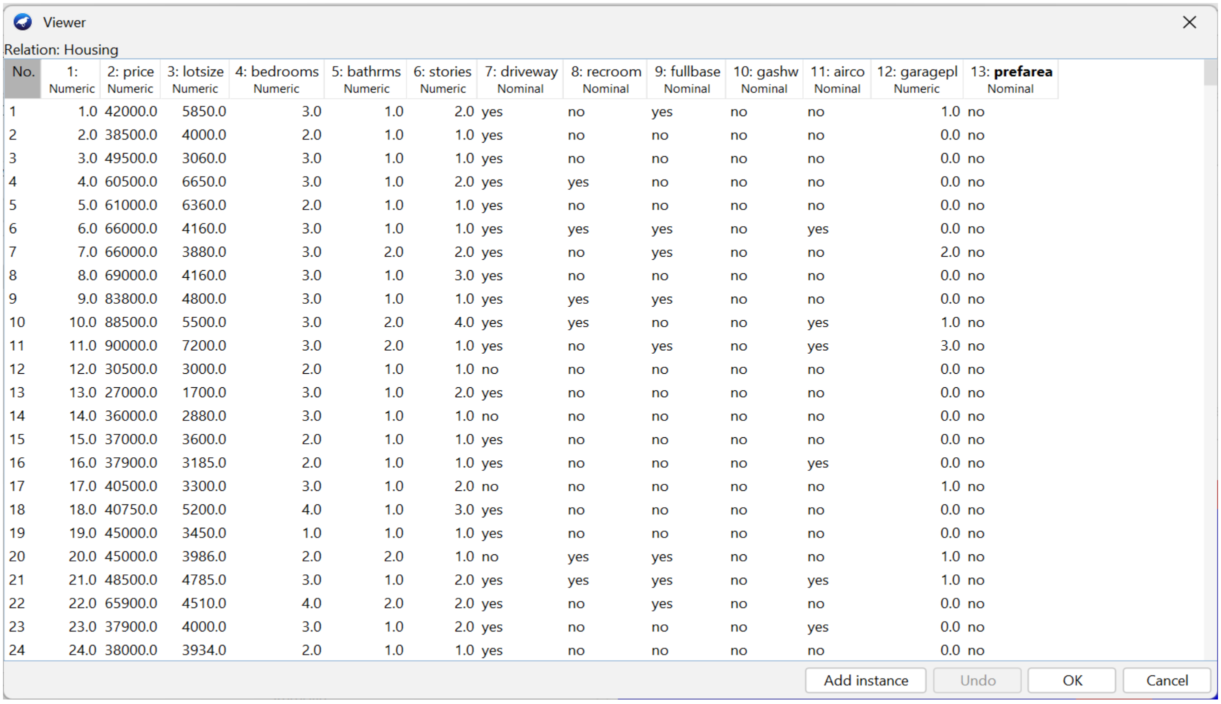
Distinct: 219

Type: Numeric

Unique: 108 (20%)

From Current relation in above figure, we can get the details of attributes in the file and no of values it contains.

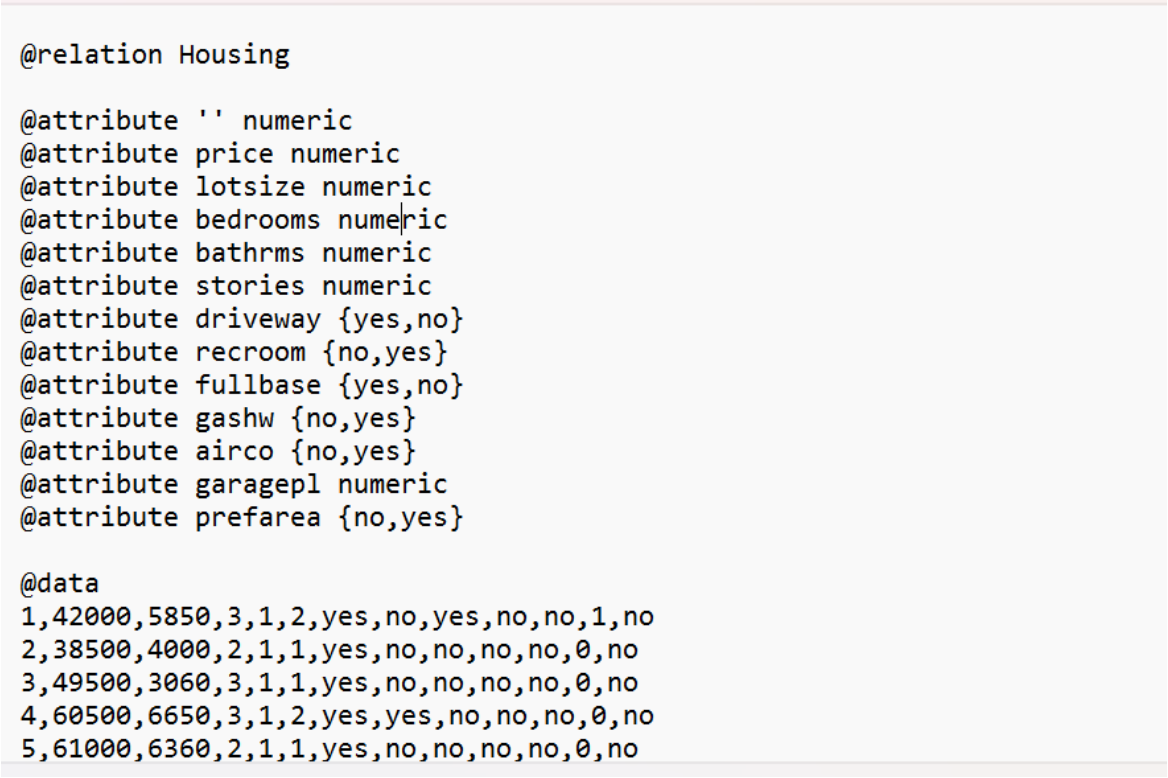
This is the dataset of Housing we are performing exploratory analysis in .csv file format.



Relation: Housing

No.	1:	2: price	3: lotsize	4: bedrooms	5: bathrms	6: stories	7: driveway	8: recroom	9: fullbase	10: gashw	11: airco	12: garagepl	13: **prefarea**
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Nominal
1	1.0	42000.0	5850.0	3.0	1.0	2.0	yes	no	yes	no	no	1.0	no
2	2.0	38500.0	4000.0	2.0	1.0	1.0	yes	no	no	no	no	0.0	no
3	3.0	49500.0	3060.0	3.0	1.0	1.0	yes	no	no	no	no	0.0	no
4	4.0	60500.0	6650.0	3.0	1.0	2.0	yes	yes	no	no	no	0.0	no
5	5.0	61000.0	6360.0	2.0	1.0	1.0	yes	no	no	no	no	0.0	no
6	6.0	66000.0	4160.0	3.0	1.0	1.0	yes	yes	yes	no	yes	0.0	no
7	7.0	66000.0	3880.0	3.0	2.0	2.0	yes	no	yes	no	no	2.0	no
8	8.0	69000.0	4160.0	3.0	1.0	3.0	yes	no	no	no	no	0.0	no
9	9.0	83800.0	4800.0	3.0	1.0	1.0	yes	yes	yes	no	no	0.0	no
10	10.0	88500.0	5500.0	3.0	2.0	4.0	yes	yes	no	no	yes	1.0	no
11	11.0	90000.0	7200.0	3.0	2.0	1.0	yes	no	yes	no	yes	3.0	no
12	12.0	30500.0	3000.0	2.0	1.0	1.0	no	no	no	no	no	0.0	no
13	13.0	27000.0	1700.0	3.0	1.0	2.0	yes	no	no	no	no	0.0	no
14	14.0	36000.0	2880.0	3.0	1.0	1.0	no	no	no	no	no	0.0	no
15	15.0	37000.0	3600.0	2.0	1.0	1.0	yes	no	no	no	no	0.0	no
16	16.0	37900.0	3185.0	2.0	1.0	1.0	yes	no	no	no	yes	0.0	no
17	17.0	40500.0	3300.0	3.0	1.0	2.0	no	no	no	no	no	1.0	no
18	18.0	40750.0	5200.0	4.0	1.0	3.0	yes	no	no	no	no	0.0	no
19	19.0	45000.0	3450.0	1.0	1.0	1.0	yes	no	no	no	no	0.0	no
20	20.0	45000.0	3986.0	2.0	2.0	1.0	no	yes	yes	no	no	1.0	no
21	21.0	48500.0	4785.0	3.0	1.0	2.0	yes	yes	yes	no	yes	1.0	no
22	22.0	65900.0	4510.0	4.0	2.0	2.0	yes	no	yes	no	no	0.0	no
23	23.0	37900.0	4000.0	3.0	1.0	2.0	yes	no	no	no	yes	0.0	no
24	24.0	38000.0	3934.0	2.0	1.0	1.0	yes	no	no	no	no	0.0	no

Now, we are converting the .csv file to arff file and we can load data into weka

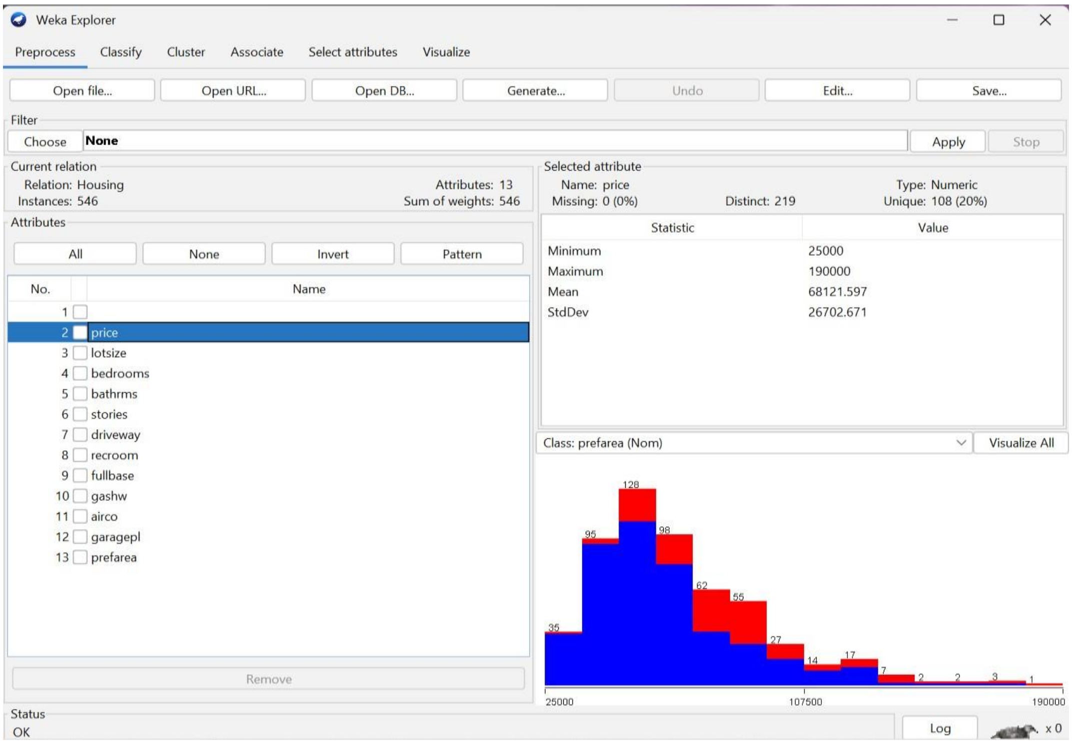


```
@relation Housing

@attribute '' numeric
@attribute price numeric
@attribute lotsize numeric
@attribute bedrooms numeric
@attribute bathrms numeric
@attribute stories numeric
@attribute driveway {yes,no}
@attribute recroom {no,yes}
@attribute fullbase {yes,no}
@attribute gashw {no,yes}
@attribute airco {no,yes}
@attribute garagepl numeric
@attribute prefarea {no,yes}

@data
1,42000,5850,3,1,2,yes,no,yes,no,no,1,no
2,38500,4000,2,1,1,yes,no,no,no,no,0,no
3,49500,3060,3,1,1,yes,no,no,no,no,0,no
4,60500,6650,3,1,2,yes,yes,no,no,no,0,no
5,61000,6360,2,1,1,yes,no,no,no,no,0,no
```

The arff type dataset is opened in the weka explorer where we can analyse each attribute in detail. For example, below we can see the details of attribute price such as the minimum value, maximum value, mean and standard deviation.



Filter: None

Current relation
Relation: Housing
Attributes: 13
Instances: 546
Sum of weights: 546

Selected attribute
Name: price
Type: Numeric
Missing: 0 (0%)
Distinct: 219
Unique: 108 (20%)

Statistic	Value
Minimum	25000
Maximum	190000
Mean	68121.597
StdDev	26702.671

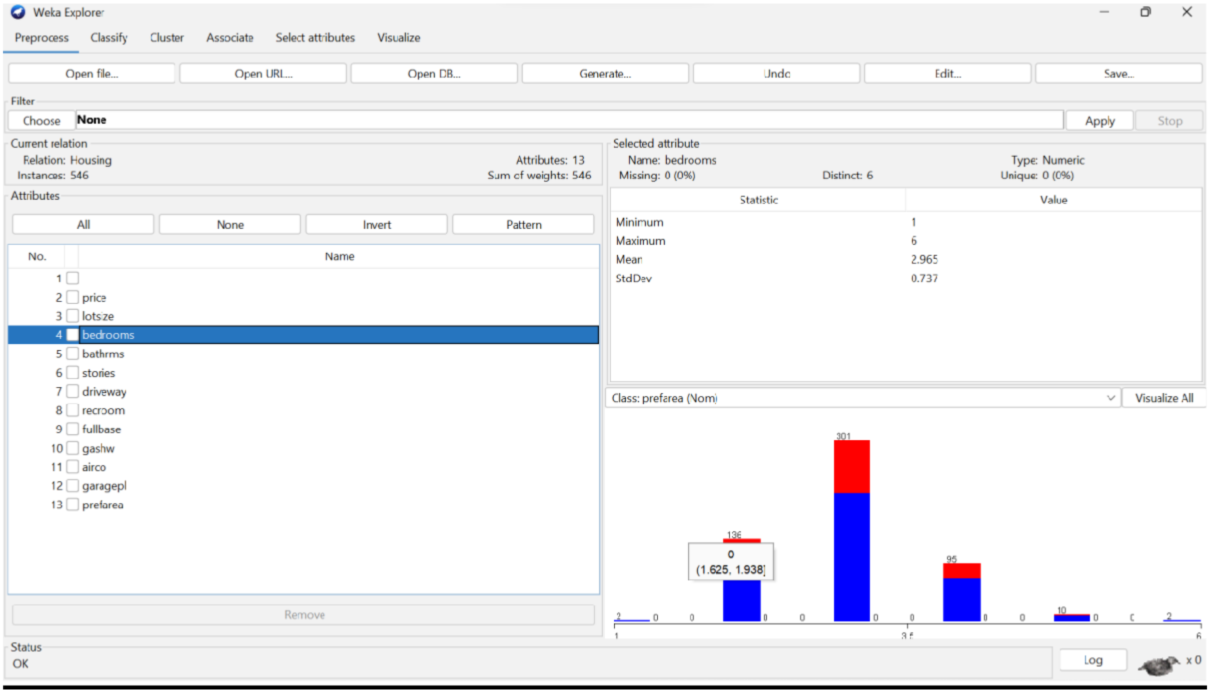
Attributes

No.	Name
1	
2	price
3	lotsize
4	bedrooms
5	bathrms
6	stories
7	driveway
8	recroom
9	fullbase
10	gashw
11	airco
12	garagepl
13	prefarea

Class: prefarea (Nom)

Status
OK

The following graph is about attribute bedrooms:



Filter: None

Current relation
Relation: Housing
Attributes: 13
Instances: 546
Sum of weights: 546

Selected attribute
Name: bedrooms
Type: Numeric
Missing: 0 (0%)
Distinct: 6
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	6
Mean	2.965
StdDev	0.737

Attributes

No.	Name
1	
2	price
3	lotsize
4	bedrooms
5	bathrms
6	stories
7	driveway
8	recroom
9	fullbase
10	gashw
11	airco
12	garagepl
13	prefarea

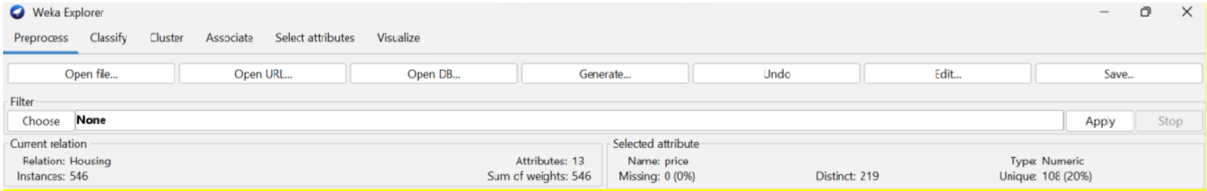
Class: prefarea (Nom)

Status
OK

Missing Data:

Once we load the data into explorer, on the right side in the selected attribute we can see missing from this we will know if there is any data missing and if there are any missing values, we click the button filter then click choose in that go to filter then unsupervised and then attribute in that we will find replace missing values you need to select that.

But in the dataset, we are using there is no missing data.



Filter: None

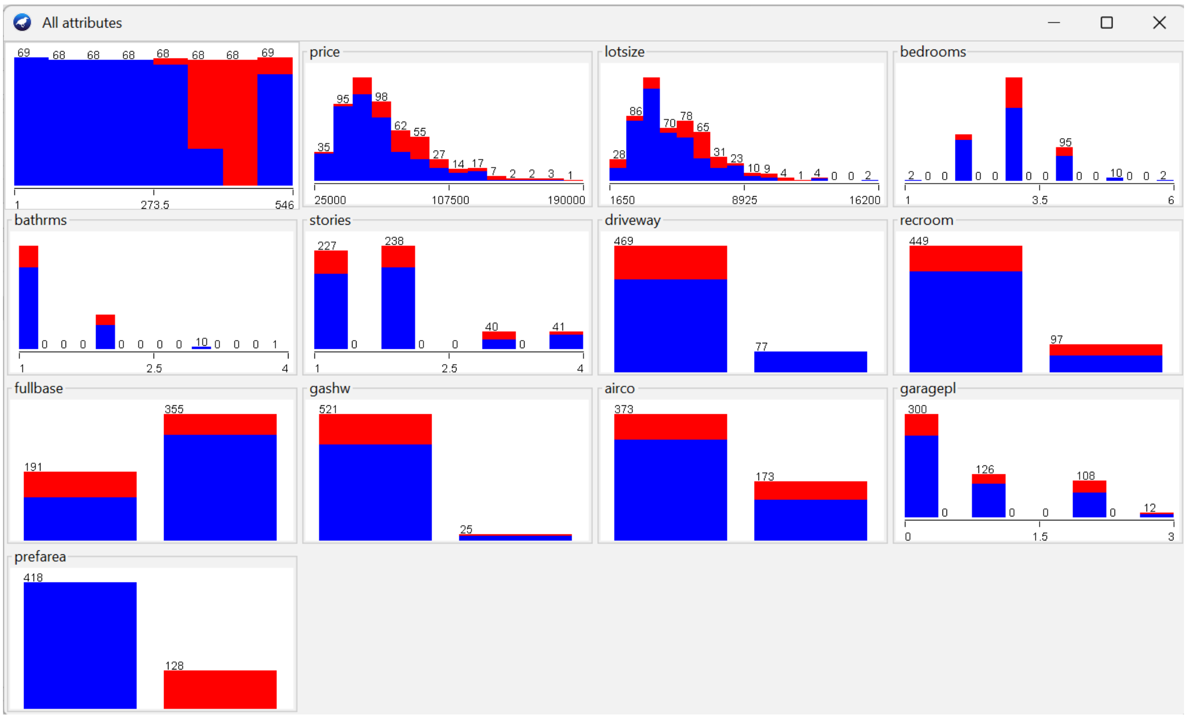
Relation: Housing
Instances: 546
Attributes: 13
Sum of weights: 546

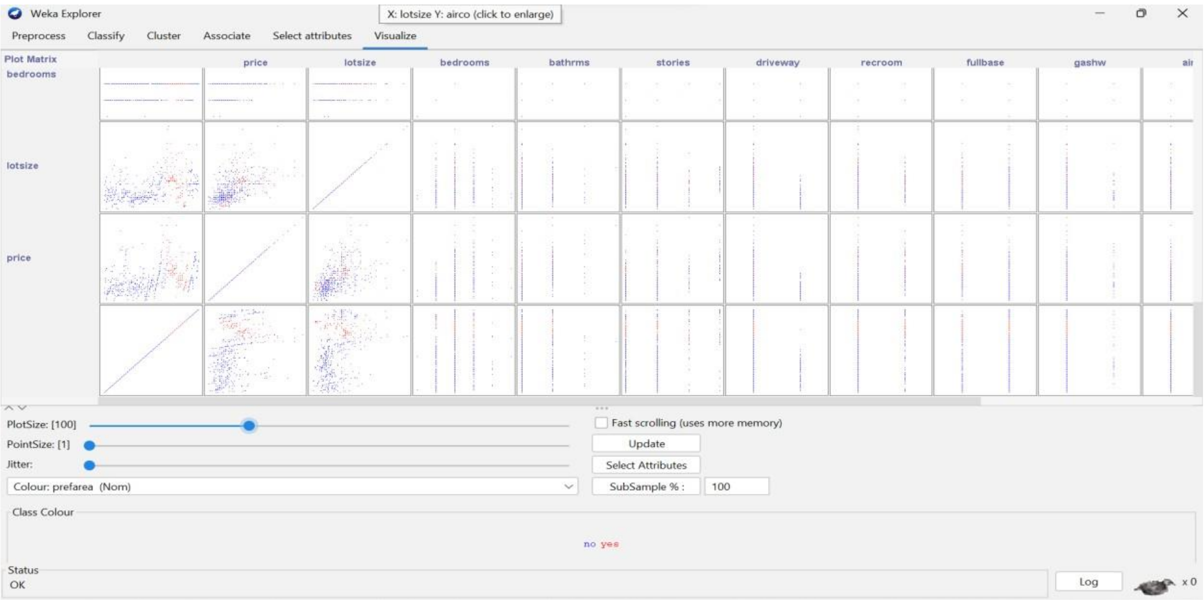
Name: price
Missing: 0 (0%)
Distinct: 219
Type: Numeric
Unique: 108 (20%)

Visualization of all attributes:

The data set on an attribute with respect to other attributes are show below. This can be seen by clicking on the tab visualize all.

- The X-axis and Y-axis represent the attribute





PlotSize: [100]

PointSize: [1]

Jitter:

Colour: prefarea (Nom)

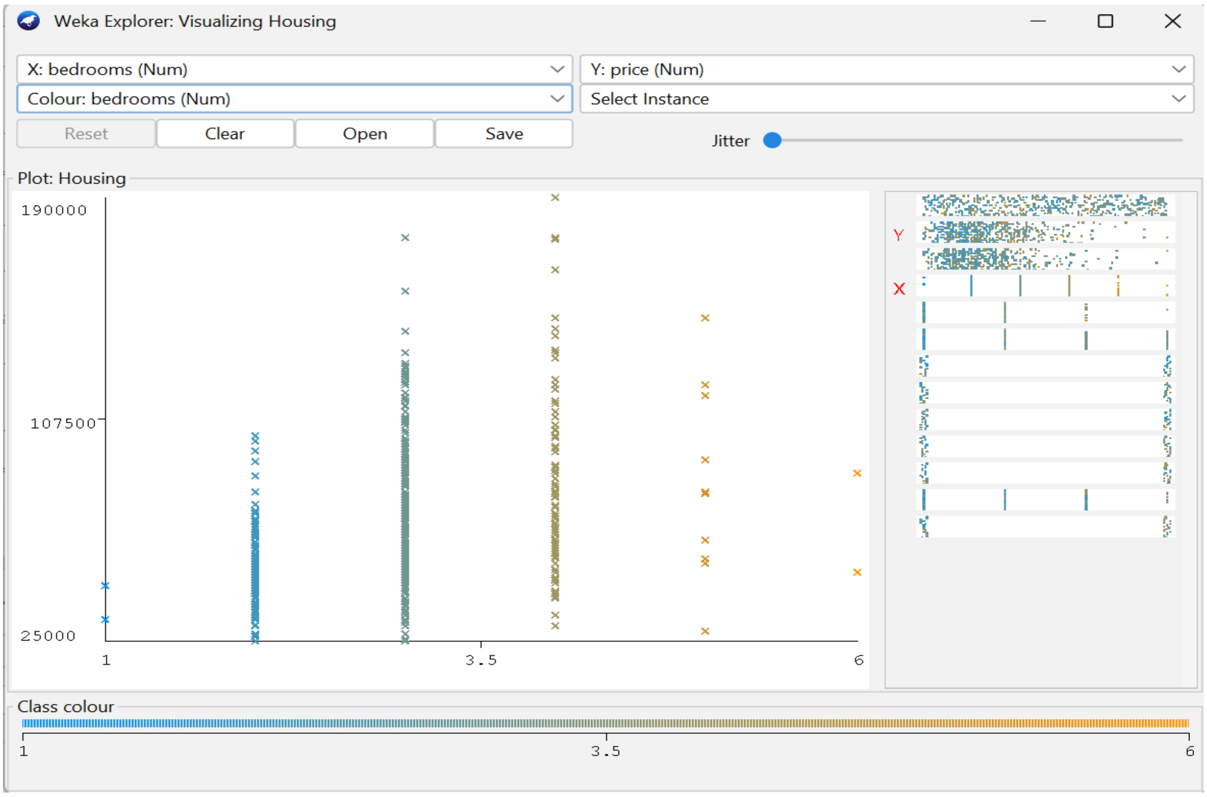
SubSample % : 100

Status

OK

Visualization

1. The following graph shows bedroom over X-axis and Price over Y-axis



X: bedrooms (Num)

Y: price (Num)

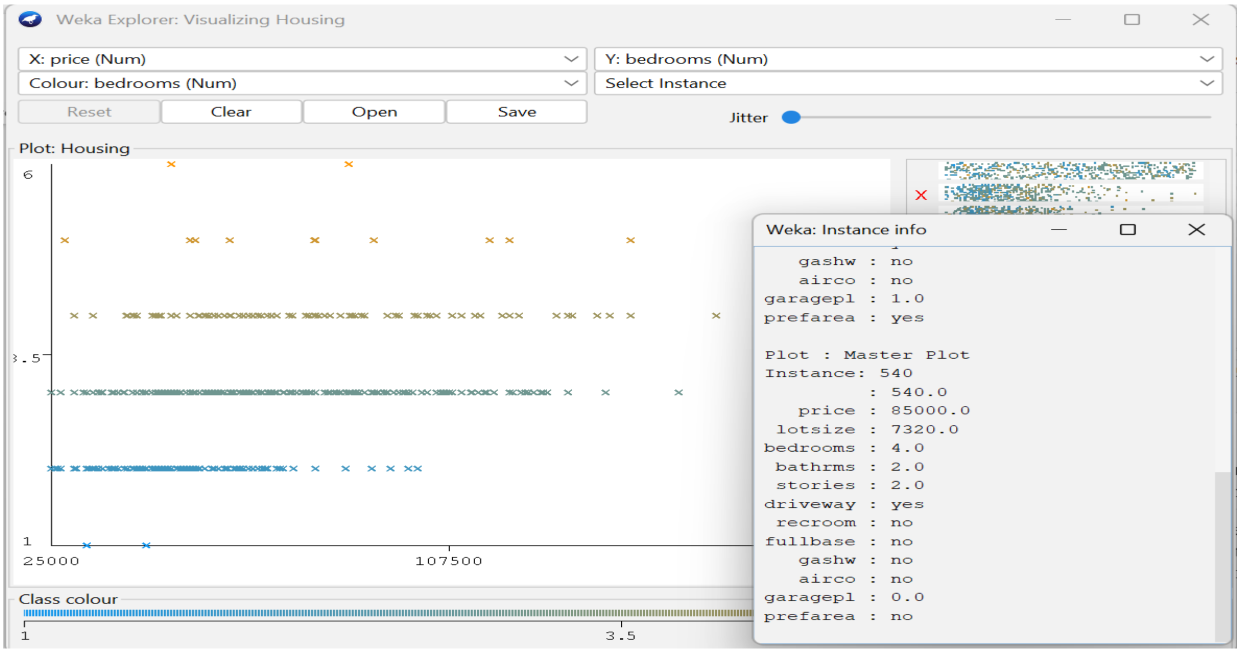
Colour: bedrooms (Num)

Select Instance

Price vs Bedrooms

The above plot shows no of bedrooms for a certain price. There are a lot of flats which can have over the pricing range between 25000 and 107500 with 2, 3, 4 and 5 rooms, but we have only two six bedrooms within this price. All the 2 bedrooms and most of 3 bedrooms are below 107500. From the plot we can say that we can get same no of bedrooms for all prices, but the size of each bedroom may not be same they can differ upon pricing. If we spend more money the size of the room might be large

1. Click on the instance represented by ‘x’ in the plot. It will give the instance details.
2. The X and Y-axis attributes can be changed in panel. We can view different plots by changing attributes in X and Y-axis.
3. Sometimes the points may overlap. The darker the spots we can say that they have multiple instances.


Weka Explorer: Visualizing Housing

X: price (Num)

Y: bedrooms (Num)

Colour: bedrooms (Num)

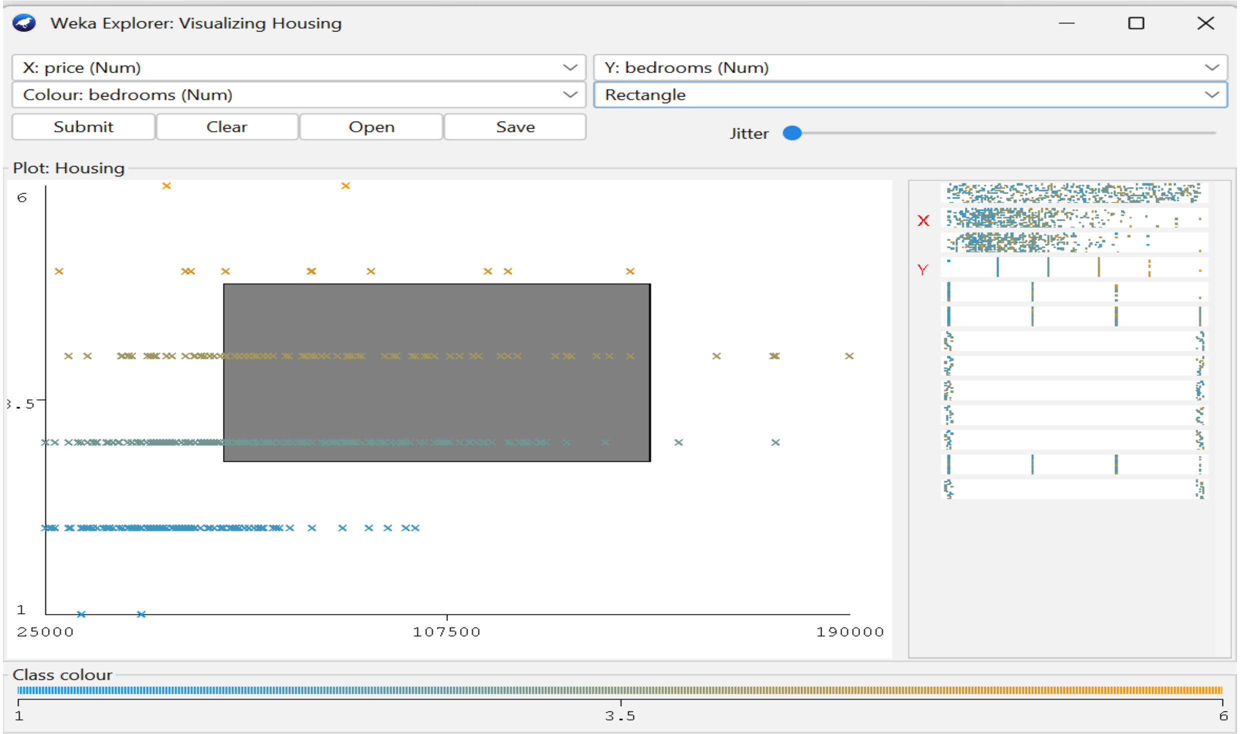
Select Instance

Weka: Instance info

```
   gashw : no
   airco : no
garagepl : 1.0
prefarea : yes

Plot : Master Plot
Instance: 540
         : 540.0
   price : 85000.0
 lotsize : 7320.0
bedrooms : 4.0
 bathrms : 2.0
 stories : 2.0
driveway : yes
 recroom : no
fullbase : no
   gashw : no
   airco : no
garagepl : 0.0
prefarea : no
```

To get the clearer view of the data, we can select an instance from dropdown. From below plot you can see we selected the rectangle, and we can get clear view of the data present in rectangle.


Weka Explorer: Visualizing Housing

X: price (Num)

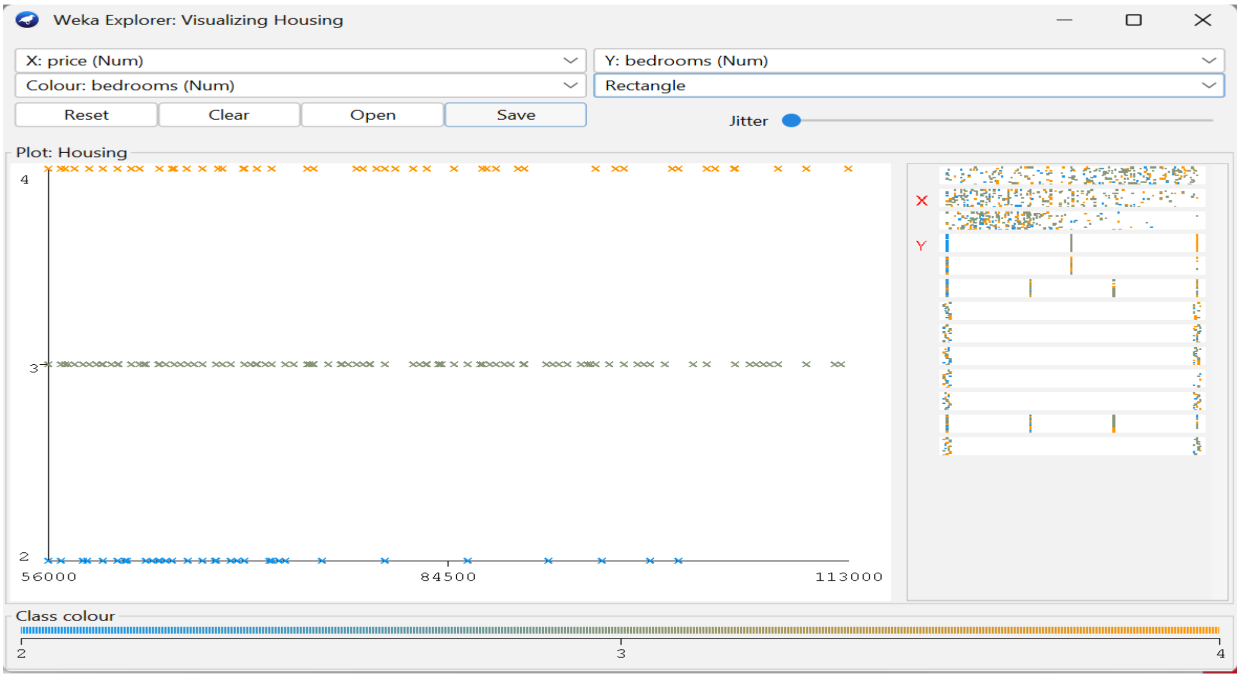
Y: bedrooms (Num)

Colour: bedrooms (Num)

Rectangle

From above graph after selecting rectangle, you can click submit, then only the selected data will be displayed and the rest is cleared.

This is the elaborated graph of points selected in the rectangular box



Weka Explorer: Visualizing Housing

X: price (Num)

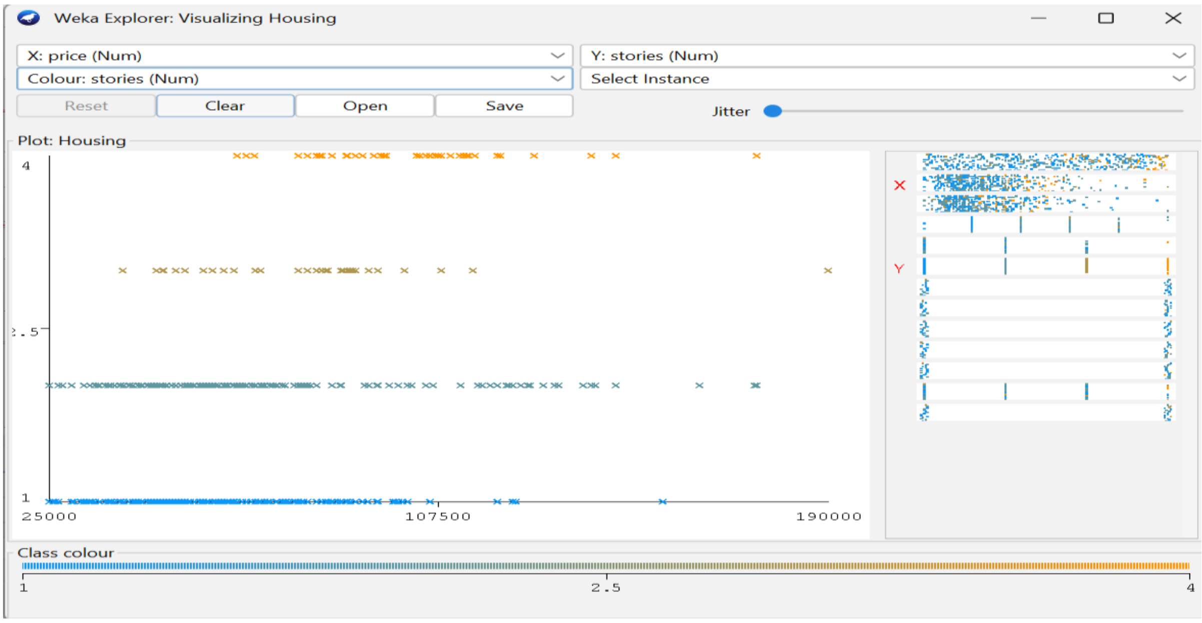
Y: bedrooms (Num)

Colour: bedrooms (Num)

Rectangle

Plot: Housing

2.The following graph shows plot between the price and stories.



Weka Explorer: Visualizing Housing

X: price (Num)

Y: stories (Num)

Colour: stories (Num)

Select Instance

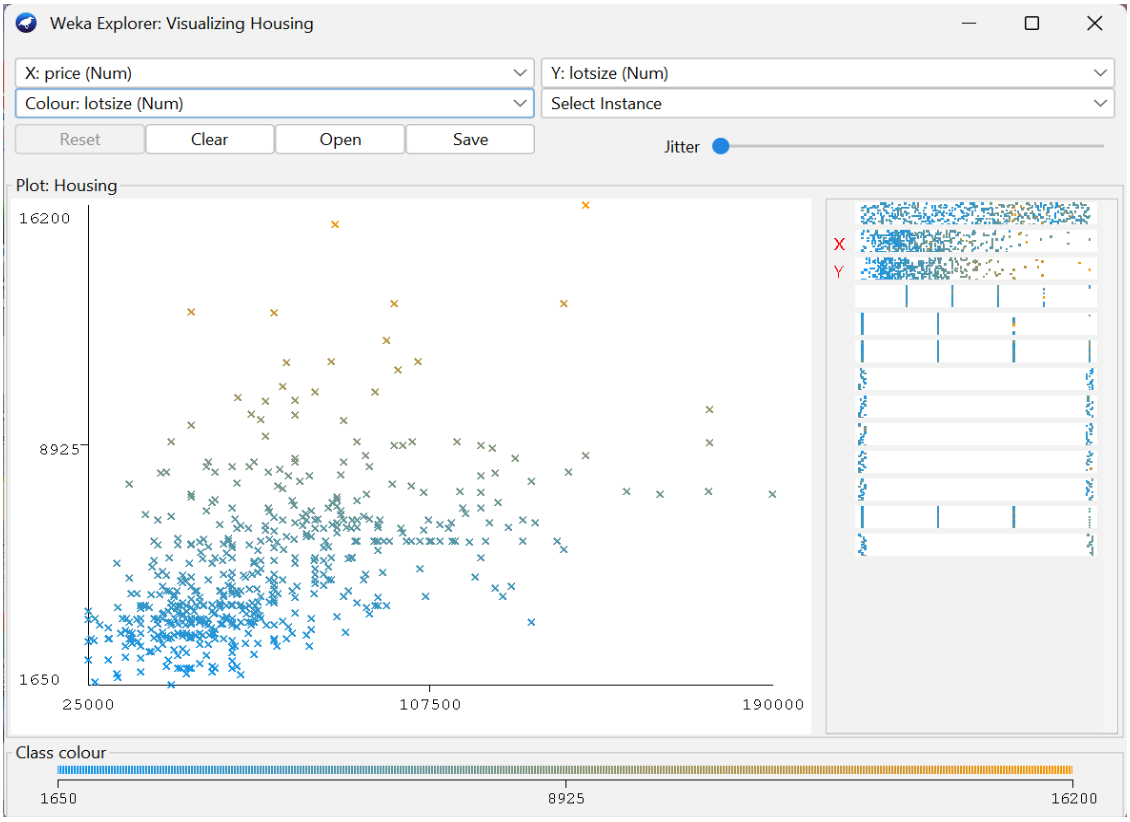
Plot: Housing

Plot vs Stories

From the above plot we have price over X-axis and stories on Y-axis. There are most of the 2,3 and 4 stories building in the range between 25000 and 107500, and very few in the range of 190000. Even though we get more stories for the same price, the size of the every building may not be similar.

3. The below graph is price over X-axis and lotsize over Y-axis:

From the plot we can observe that the highest price i.e., 190000 has a lotsize around 7400 and the highest lotesize i.e., 16200 is around 142000. The most of lotsize between 1650 and 8925 has the pricing range between 25000 and 107500. From this we can say that all the Lotsize with in the same range have similar price and we get only few bigger lotsize with that similar price.



Price vs Lotsize

References:

- https://www.cs.waikato.ac.nz/ml/weka/
- https://cs.ccsu.edu/~markov/weka-tutorial.pdf
- https://www.softwaretestinghelp.com/weka-explorer-tutorial/

Observation of other Teammates:

Weka is a very efficient tool to perform data mining tasks and can experiment with new datasets. The interface and functionalities are very easy to use and have different analyses to choose. Almost every data can be loaded directly which reduces our work and worry about coding. Even if there is any missing data it can be seen in interface and can be replaced directly from the interface. It has many built-in features which can be modified as necessary.