Housing-dataset by Weka Report

Student Details:

- Kuldip Rameshbhai Savaliya - 1001832000
- Shivani Manojkumar Panchiwala - 1001982478
- Meghaben Ghanshyambhai Patel – 1002006777

Introduction:

Tool

Weka is a collection of machine learning algorithms for data mining tasks. In addition to data preparation, classification, regression, clustering, and association rules mining, it contains visualization tools and analysis tools. Python & R are ruling this era of Data Science.

For my first datamining Assignment, I looked into another data science tool called Weka. Weka was developed at the University of Waikato and has been around for quite a long time now.

Dataset

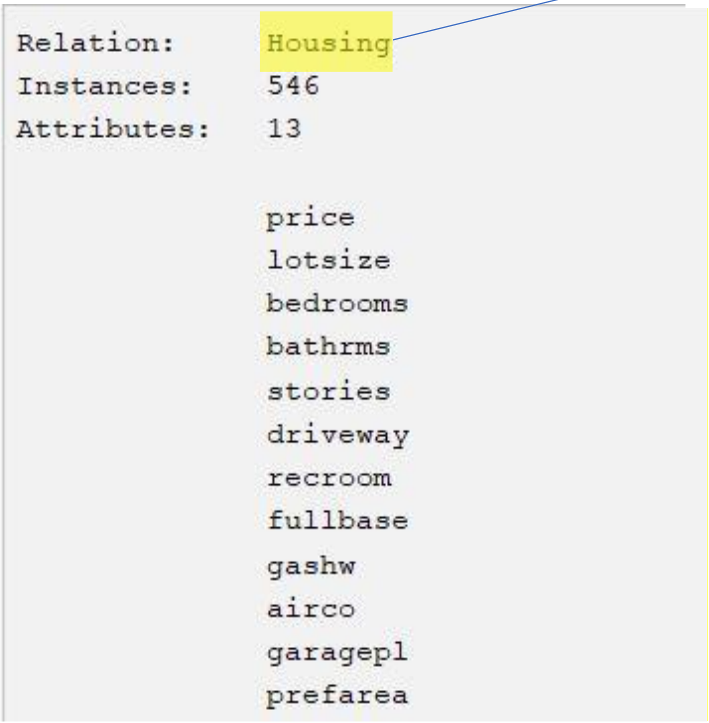
The dataset used in this report is Housing-dataset. There are 12 attributes in this dataset, of which 6 are numeric types. The dataset gives us housing information and they allow us to determine if the number of bedrooms, bathrooms, air conditioners, and all these utilities affect the overall price.
The goal of the project is to present the visualization of the data and extract relevant information from it.

Retrieving the Data:

- First, the dataset given is in CSV file format.
- But Weka prefers to load data in the ARFF format.
- ARFF means Attribute-Relation File Format.
- For that Weka provides a handy tool 'Weka Experimenter' to load CSV files and save and convert them in ARFF.
- ARFF format uses a header which provide metadata about the types of data in the columns.
- I can now load housing. arff data file directly into 'Weka Explorer'.

Glimpse of Data:

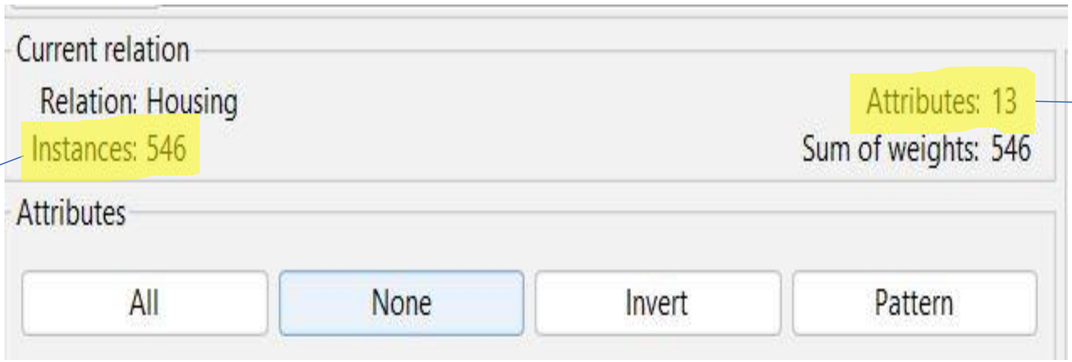
Task 1–a: Print the details of dataframe.



Dataset Name

```
Relation:      Housing
Instances:     546
Attributes:    13

               price
               lotsize
               bedrooms
               bathrms
               stories
               driveway
               recroom
               fullbase
               gashw
               airco
               garagepl
               prefarea
```

Task 1–b: Find the numbers of rows and colums in dataset.



Current relation

Relation: Housing

Attributes: 13

Instances: 546

Sum of weights: 546

Attributes

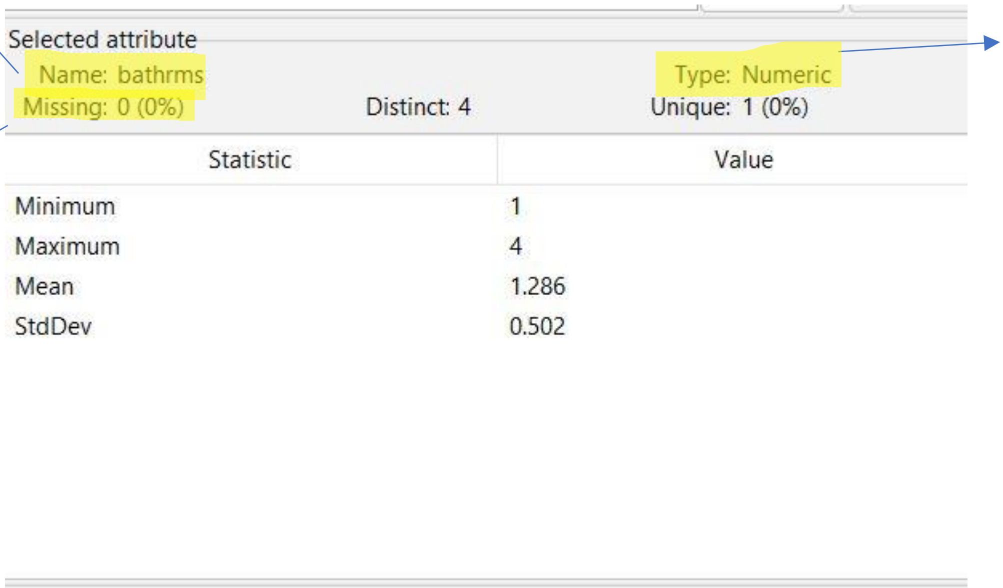
No. of columns

No. of rows

Check for missing data:

Task 1–c: Print descriptive details for ‘bathrms’ columns of the dataset.

Attribute Name



Selected attribute

Name: bathrms

Missing: 0 (0%)   Distinct: 4   Type: Numeric   Unique: 1 (0%)

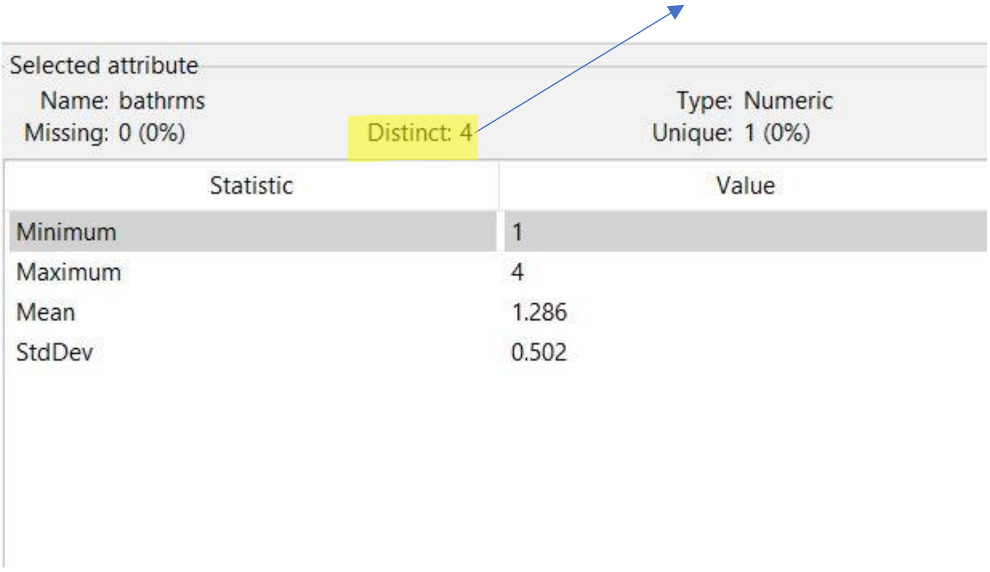
Statistic	Value
Minimum	1
Maximum	4
Mean	1.286
StdDev	0.502

Datatype

No. of missing values

Task 1-d(i): Show all the distinct values of ‘bathrms’.

No. of distinct values



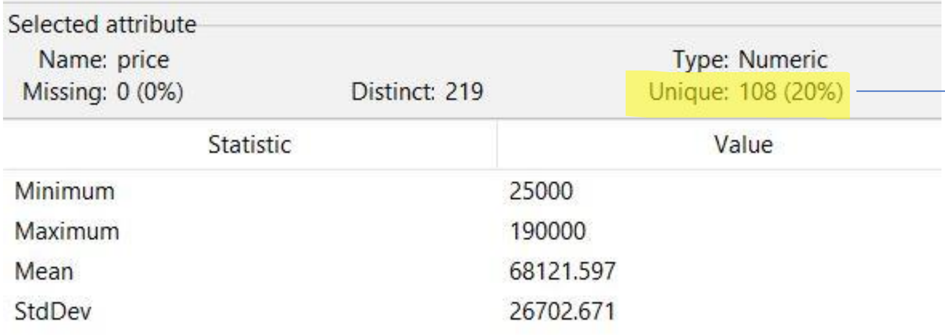
Selected attribute

Name: bathrms

Missing: 0 (0%)   Distinct: 4   Type: Numeric   Unique: 1 (0%)

Statistic	Value
Minimum	1
Maximum	4
Mean	1.286
StdDev	0.502

Task 1-d(ii): Find percentages of ‘price’ for unique values.



Selected attribute

Name: price | Type: Numeric

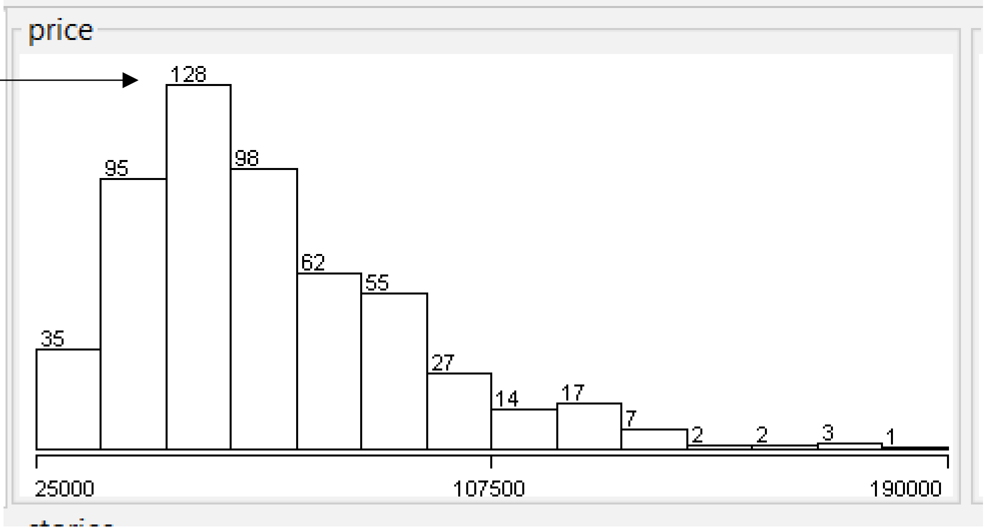
Missing: 0 (0%) | Distinct: 219 | Unique: 108 (20%)

Statistic	Value
Minimum	25000
Maximum	190000
Mean	68121.597
StdDev	26702.671

Percentages of Unique values

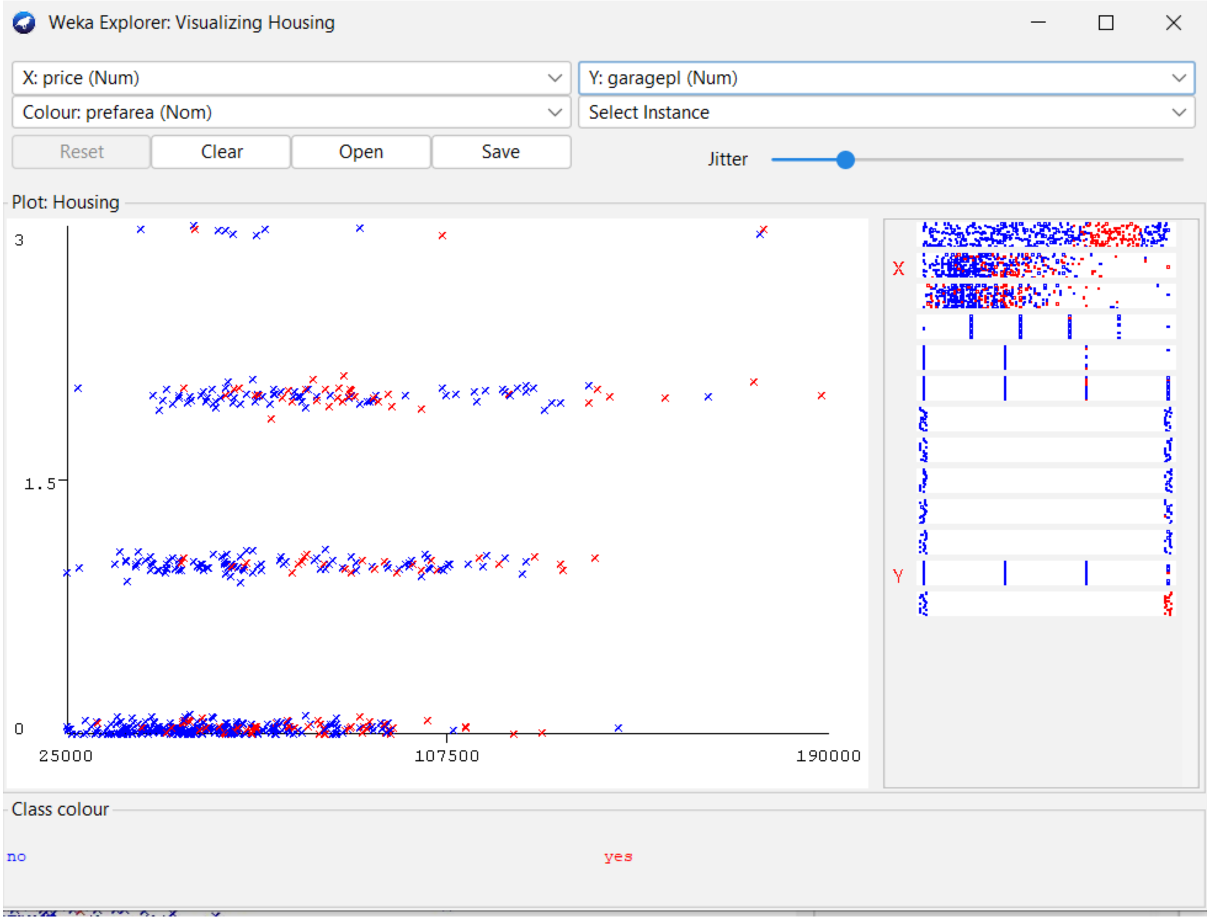
Data Exploration:

Task 2-a: Find out the price with largest number of records.



- The highest number of records lies between the range of 25000 to 107500. Approximately, 128 records lie on the price of 54000$.

Task 2-b: Find out the garagepl based on price of a house.



Weka Explorer: Visualizing Housing

X: price (Num)

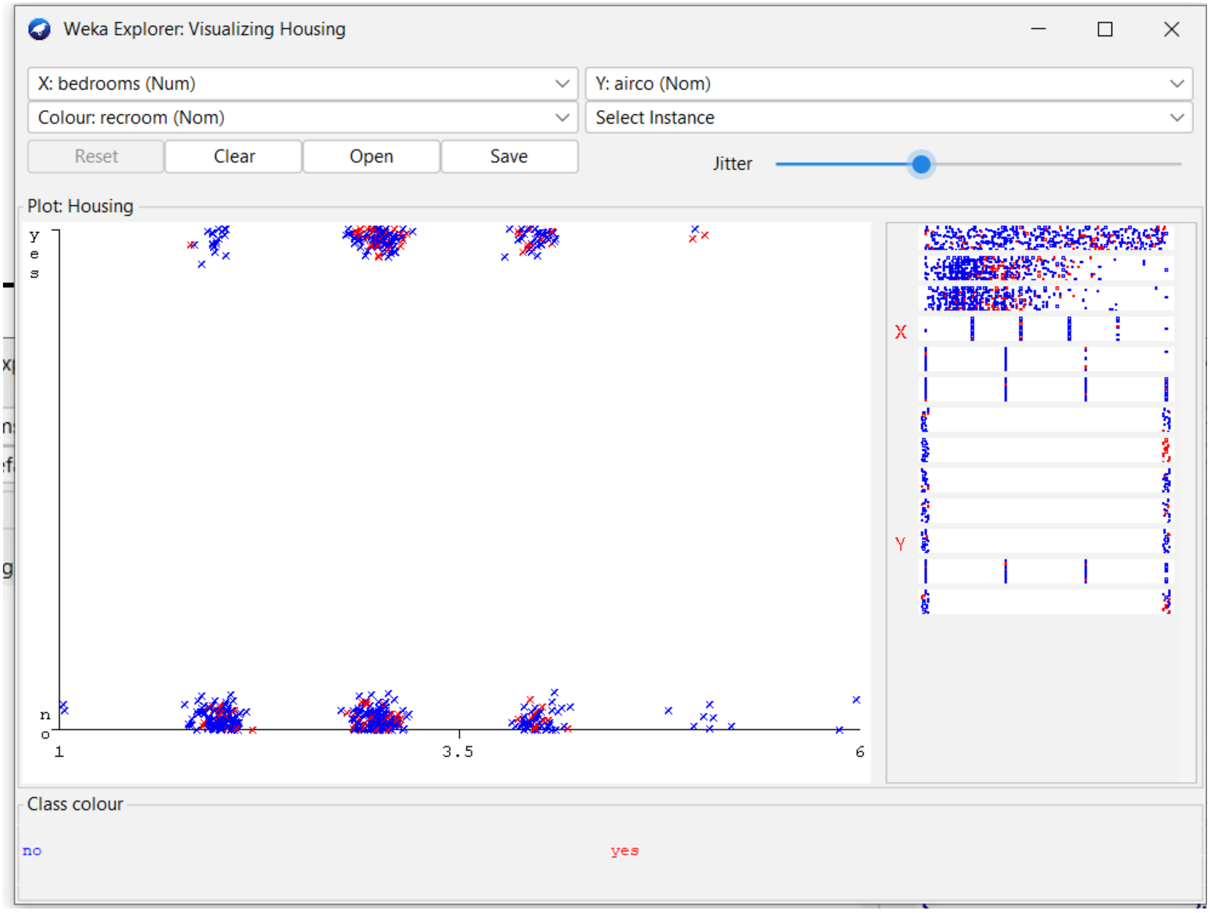
Y: garagepl (Num)

Colour: prefarea (Nom)

Select Instance

- X axis – Price (Num)
  Y axis – garagepl (Num)
- The above graph shows, as the price of house increases there are more no. of houses equipped with the garages.
- There are certain exceptions since we see from the graph that even though prices are along line of 150000$ the houses are not equipped with any garages.

Task 2-c: Outliers of houses with bedrooms without airconditioning.



X: bedrooms (Num)

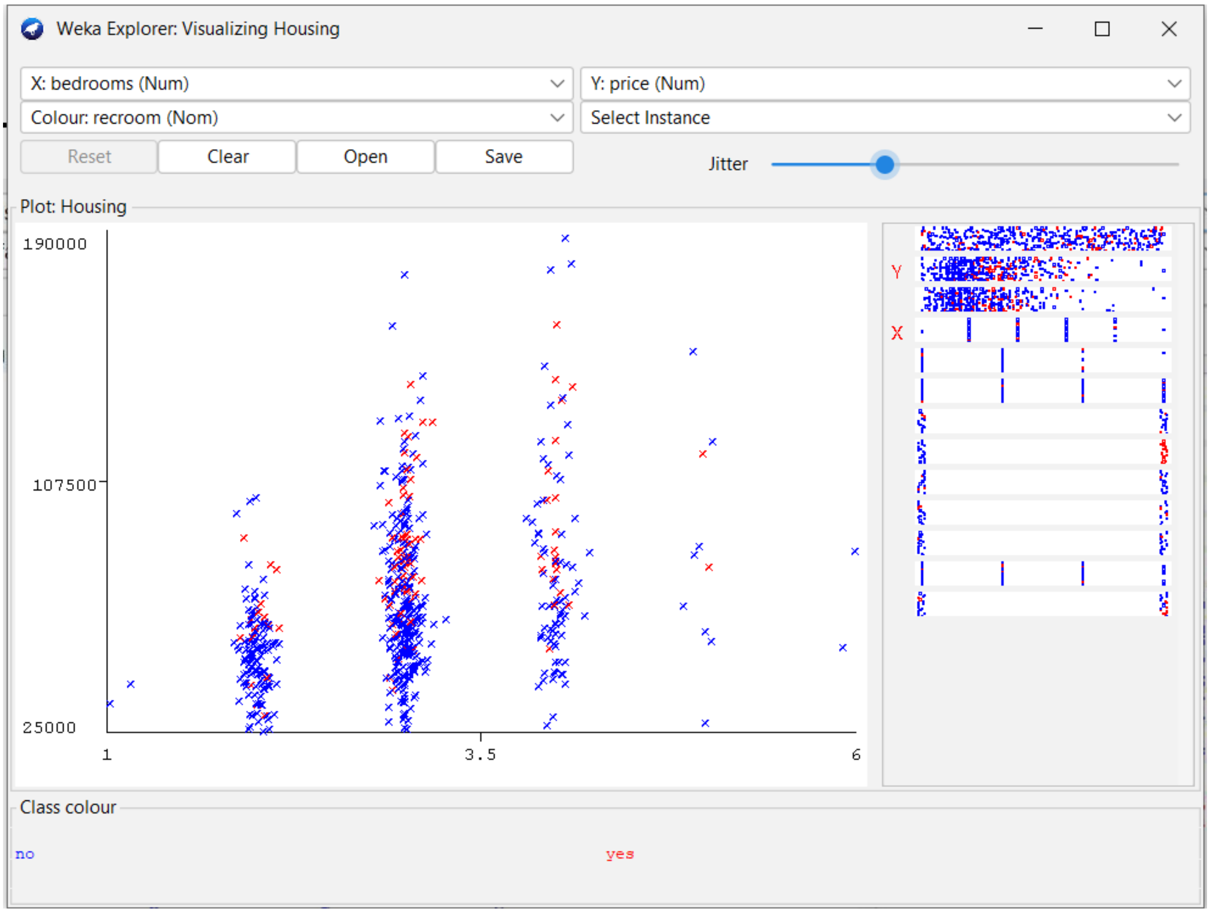
Y: airco (Nom)

Colour: recroom (Nom)

Select Instance

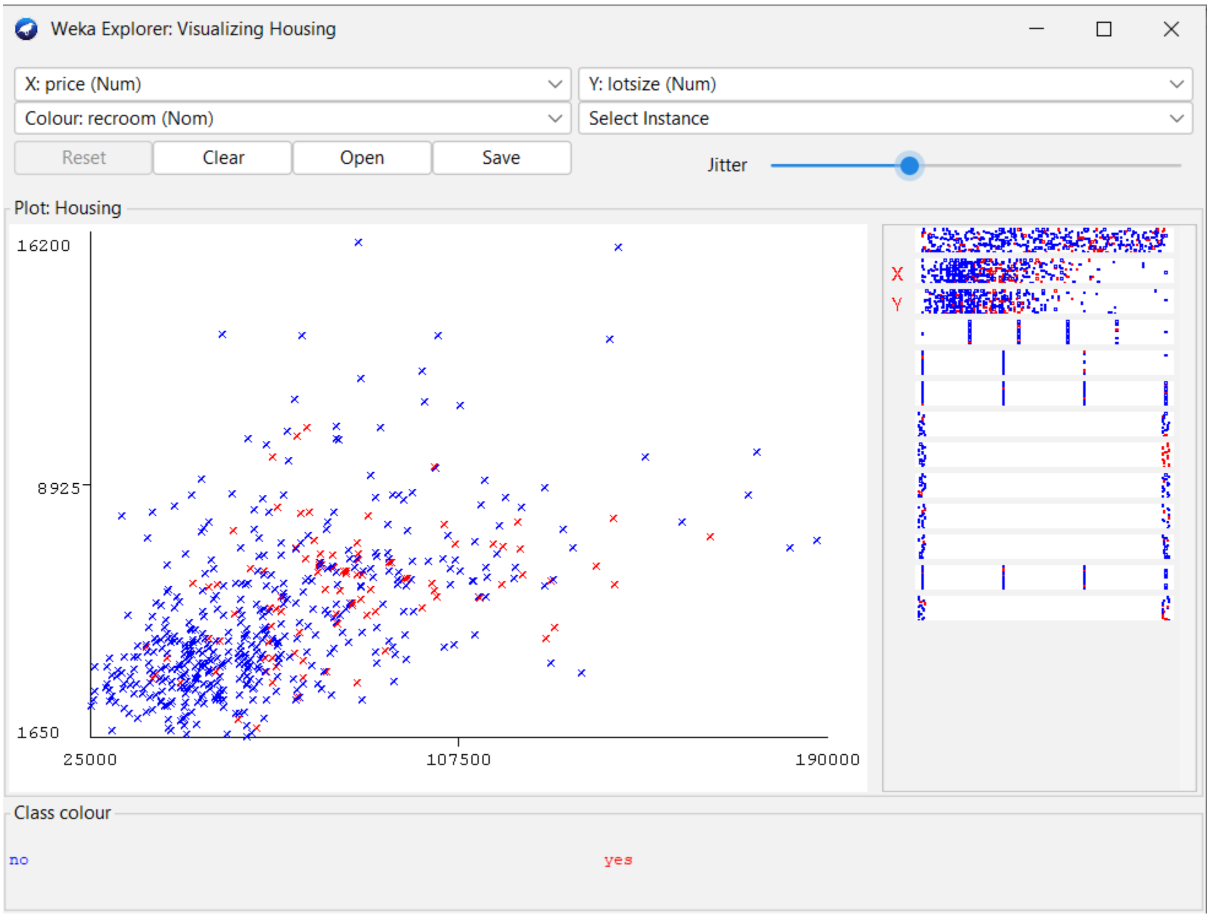
- X axis – bedrooms (Num)
  Y axis – Airconditioning (Nom)
- We can see that there are plenty of houses where it does not matter if the house has a lot of rooms, it will not be equipped with air conditioning.
- For example, the graph shows an element where houses equipped with 6 bedrooms still have no air conditioning.

Task 3-a: Show the bedrooms distribution against each price.



- X axis - bedrooms (Num)
  Y axis – price (Num)
- From the above graph, we can see that if the demand for bedrooms increases the price of the house increases respectively. However, if the customer targets the budget of 90000$ to 120000$, he can have options of houses with 2,3, and 4 bedrooms respectively.

Task 3-b: Visualizing the size of lot according to price.



Weka Explorer: Visualizing Housing

X: price (Num)

Y: lotsize (Num)

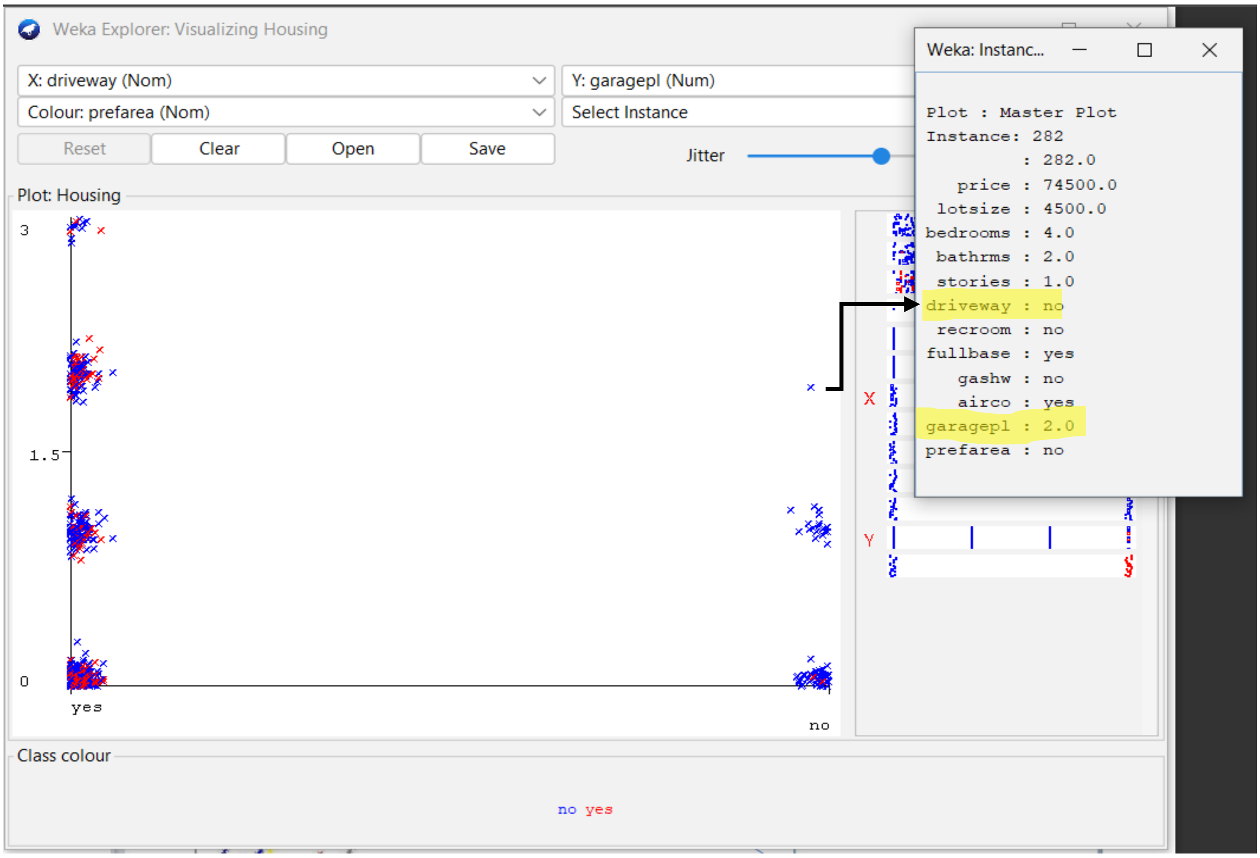
Colour: recroom (Nom)

Select Instance

- X axis – price (Num)
  Y axis – lotsize (Num)
- We can see that if a house's price rises, the size of the lot also rises. In the case of a house priced between 25000 and 107500, the lot size will range between 1650 and 8925. Additionally, lower prices and smaller lot sizes are more common than one with a high price and a big lot size. Thus, people prefer houses that are more affordable to those that are more expensive.

Task 4: Some Additional Insights.

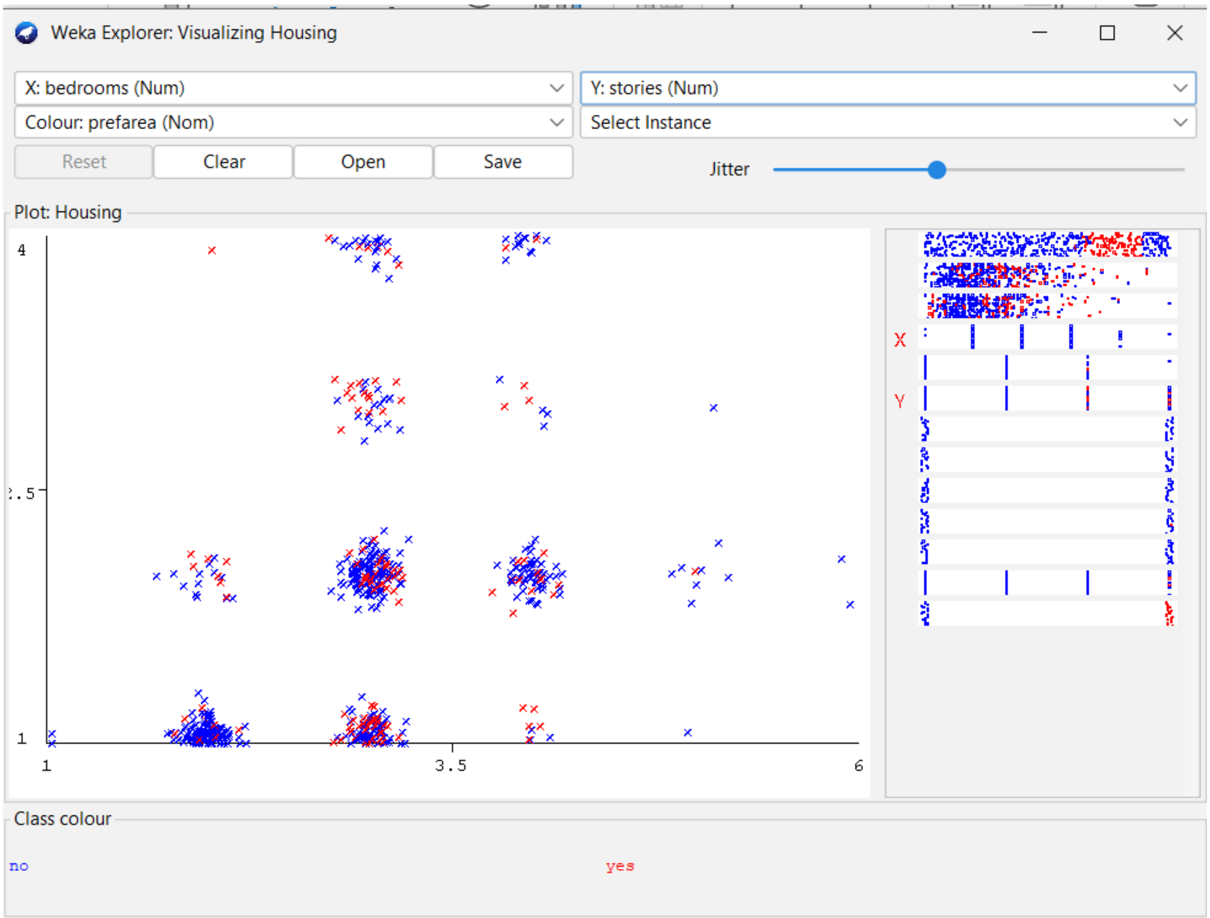
4(i): Relationship between the driveway and garages.



```
Plot : Master Plot
Instance: 282
         : 282.0
   price : 74500.0
 lotsize : 4500.0
bedrooms : 4.0
 bathrms : 2.0
 stories : 1.0
driveway : no
 recroom : no
fullbase : yes
   gashw : no
  airco : yes
garagepl : 2.0
prefarea : no
```

- X axis (driveway)
  Y axis (garagepl)
- From the above graph, we can see that irrespective of the number of garages the house is equipped with a driveway. However, there are certain outliers to this.
- For example, a customer expects that if he buys a house with 2 garages, he should have a driveway for one of his cars which is not the case here. (Look at the pointer)

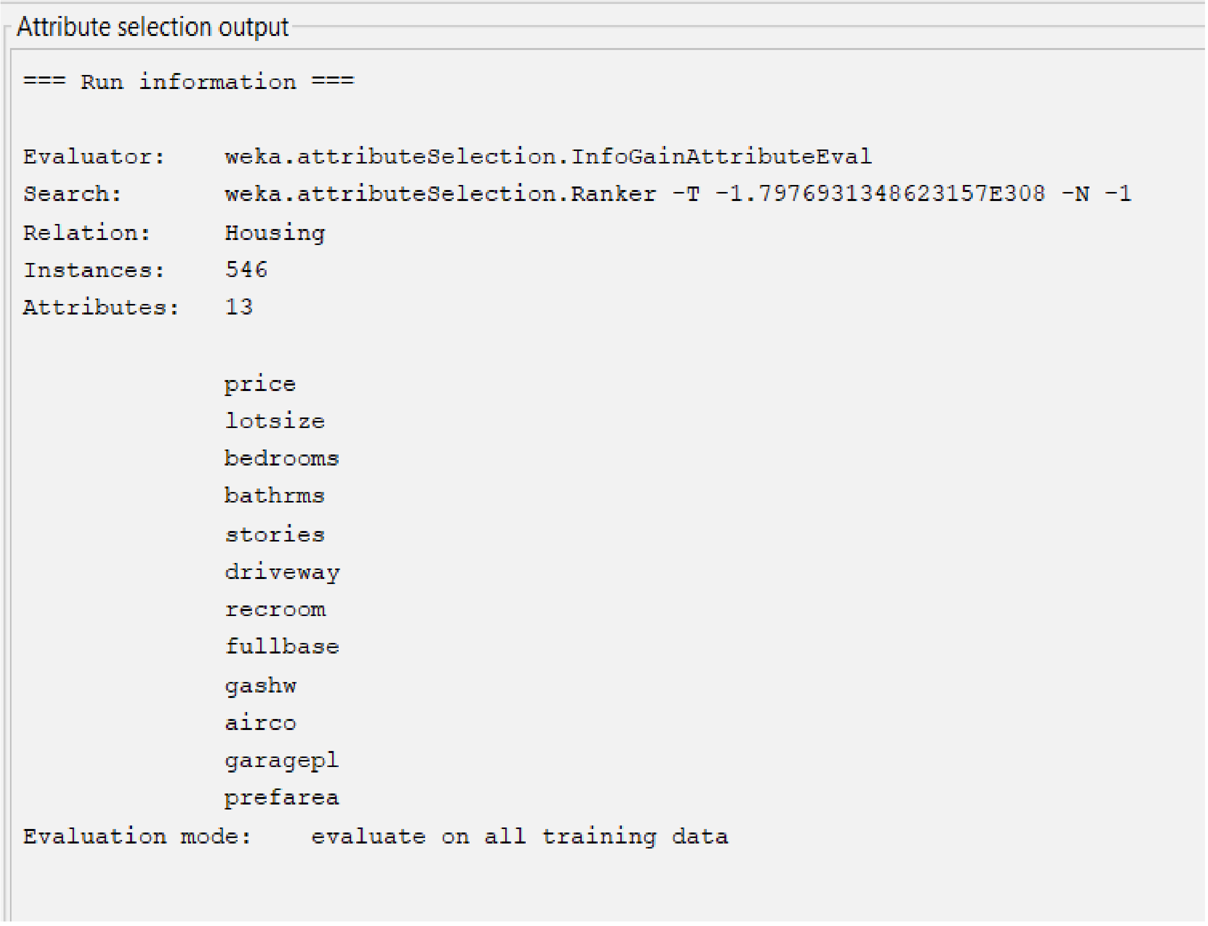
4(II): Relationship between the no. of bedrooms Vs No. of stories.



- X axis – bedrooms (Num)
  Y axis – stories (Num)
- Interestingly, the graph above depicts those houses with 3 and 4 bedrooms respectively can be built in designs of 2-storey,3-storey, and 4-storey. However, 2 bedrooms house construction is limited to 2-storey only.

Feature Selection using the InfoGainAttributeEval.

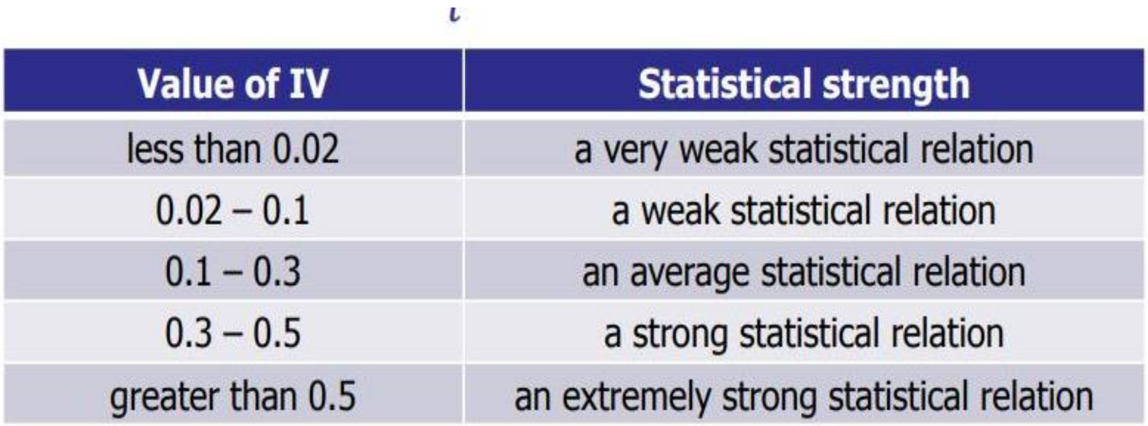
We can select the feature based on attribute evaluator and I select InfoGainAttributeeEval, and I select Ranker method for feature selection of housing dataset.



Attribute selection output

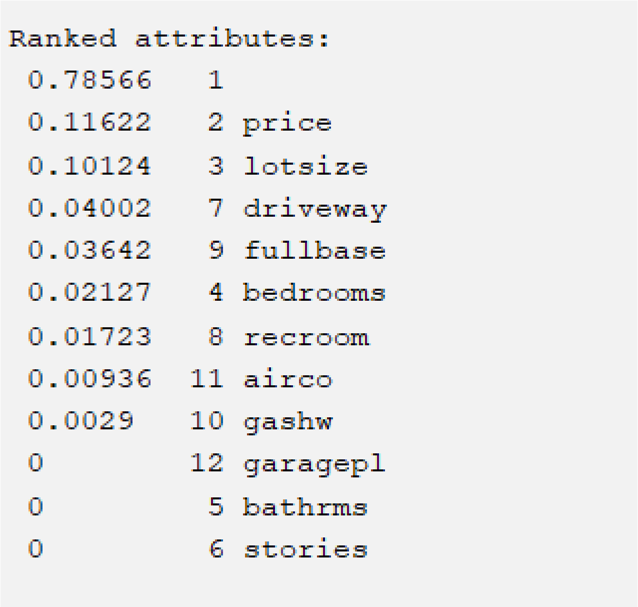
```
=== Run information ===

Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     Housing
Instances:    546
Attributes:   13

              price
              lotsize
              bedrooms
              bathrms
              stories
              driveway
              recroom
              fullbase
              gashw
              airco
              garagepl
              prefarea
Evaluation mode:    evaluate on all training data
```

The below table shows the measure for the effiecency of correlation between the attributes.



Value of IV	Statistical strength
less than 0.02	a very weak statistical relation
0.02 – 0.1	a weak statistical relation
0.1 – 0.3	an average statistical relation
0.3 – 0.5	a strong statistical relation
greater than 0.5	an extremely strong statistical relation

Based on the above-mentioned table the results of my analysis are show in the following order.



```
Ranked attributes:
 0.78566    1
 0.11622    2 price
 0.10124    3 lotsize
 0.04002    7 driveway
 0.03642    9 fullbase
 0.02127    4 bedrooms
 0.01723    8 recroom
 0.00936   11 airco
 0.0029    10 gashw
 0         12 garagepl
 0          5 bathrms
 0          6 stories
```

Attribute price and lot size have an average statistical strength towards correlation. The rest attributes have reasonably weak performance.

Team Work:

- First, I learn the basic concepts of the weka tool and after that, I analyze the given housing dataset for my assignment and I get information about the given dataset such as no. of columns, attribute name, and attribute types. After that, I learn how to convert .csv files into. arff format because weka prefers to load files into arff format. Then, I performed the given task 1 and task 2 and discussed with other team members. I discussed the dataset with my team member, and we get some important insights for visualizations.
- In task 2, Shivani was quick in figuring out the outliers regarding the graph of bedrooms Vs air conditioning.
- In task 3, Megha helped me figure out the pattern in the graph of price Vs lot size using the jitter for proper visualization.
- In task 4, I performed the visualization between driveway and garages and found significant insights which I later discussed with my team members, and they were supportive of it.