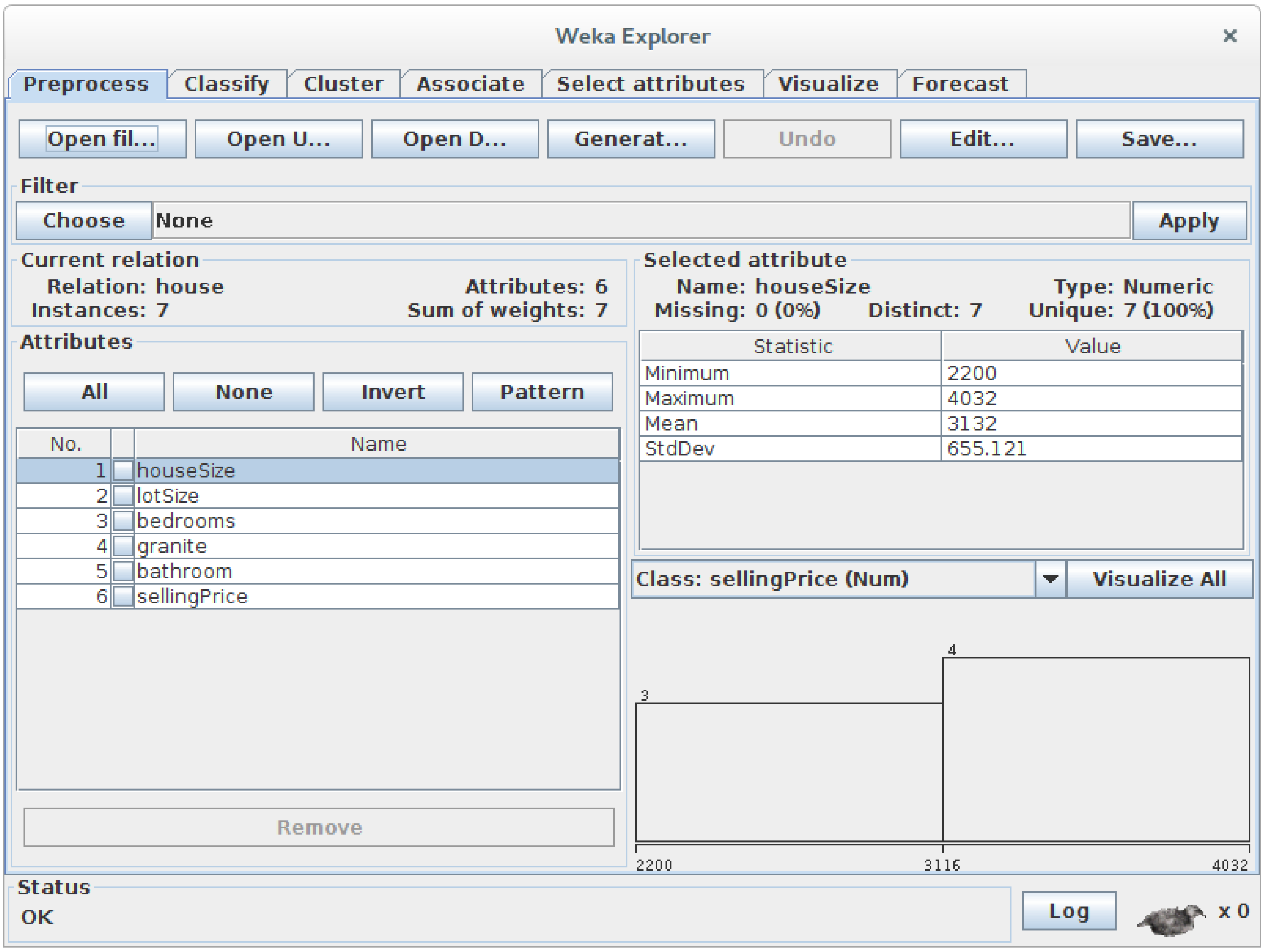
Step1. Open the data/house.arff Dataset

Click the "***Open file…***" button to open a data set and double click on the "***data***" directory.

Weka provides a number of small common machine learning datasets that you can use to practice on.

Select the "**house.*arff***" file to load the house dataset.



Filter: None

Current relation
Relation: house
Instances: 7
Attributes: 6
Sum of weights: 7

Attributes

No.		Name
1	□	houseSize
2	□	lotSize
3	□	bedrooms
4	□	granite
5	□	bathroom
6	□	sellingPrice

Selected attribute
Name: houseSize
Missing: 0 (0%)
Distinct: 7
Type: Numeric
Unique: 7 (100%)

Statistic	Value
Minimum	2200
Maximum	4032
Mean	3132
StdDev	655.121

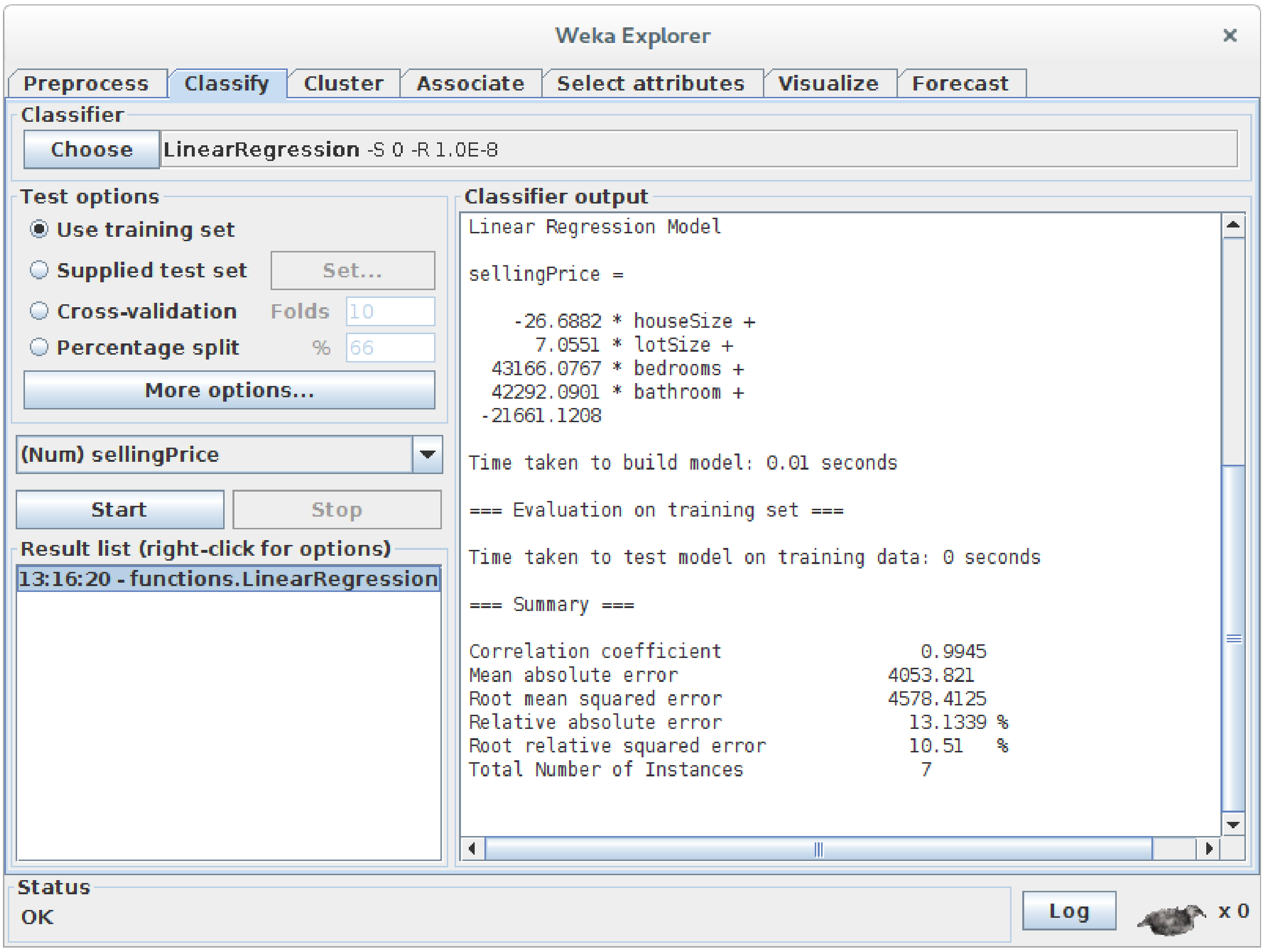
Class: sellingPrice (Num)

Status
OK

Step2. Creating the regression model with WEKA

To create the model, click on the **Classify** tab. The first step is to select the model we want to build, so WEKA knows how to work with the data, and how to create the appropriate model:

1. Click the **Choose** button, then expand the **functions** branch
2. Select the **LinearRegression** leaf with using **training set** test option.
3. Click Start to create a model.



Classifier: LinearRegression -S 0 -R 1.0E-8

- ● Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ○ Percentage split % 66

(Num) sellingPrice

13:16:20 - functions.LinearRegression

```
Linear Regression Model

sellingPrice =

    -26.6882 * houseSize +
      7.0551 * lotSize +
  43166.0767 * bedrooms +
  42292.0901 * bathroom +
 -21661.1208

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient                  0.9945
Mean absolute error                   4053.821
Root mean squared error               4578.4125
Relative absolute error                 13.1339 %
Root relative squared error             10.51   %
Total Number of Instances                7
```

Status
OK

Step3: Interpreting the regression model

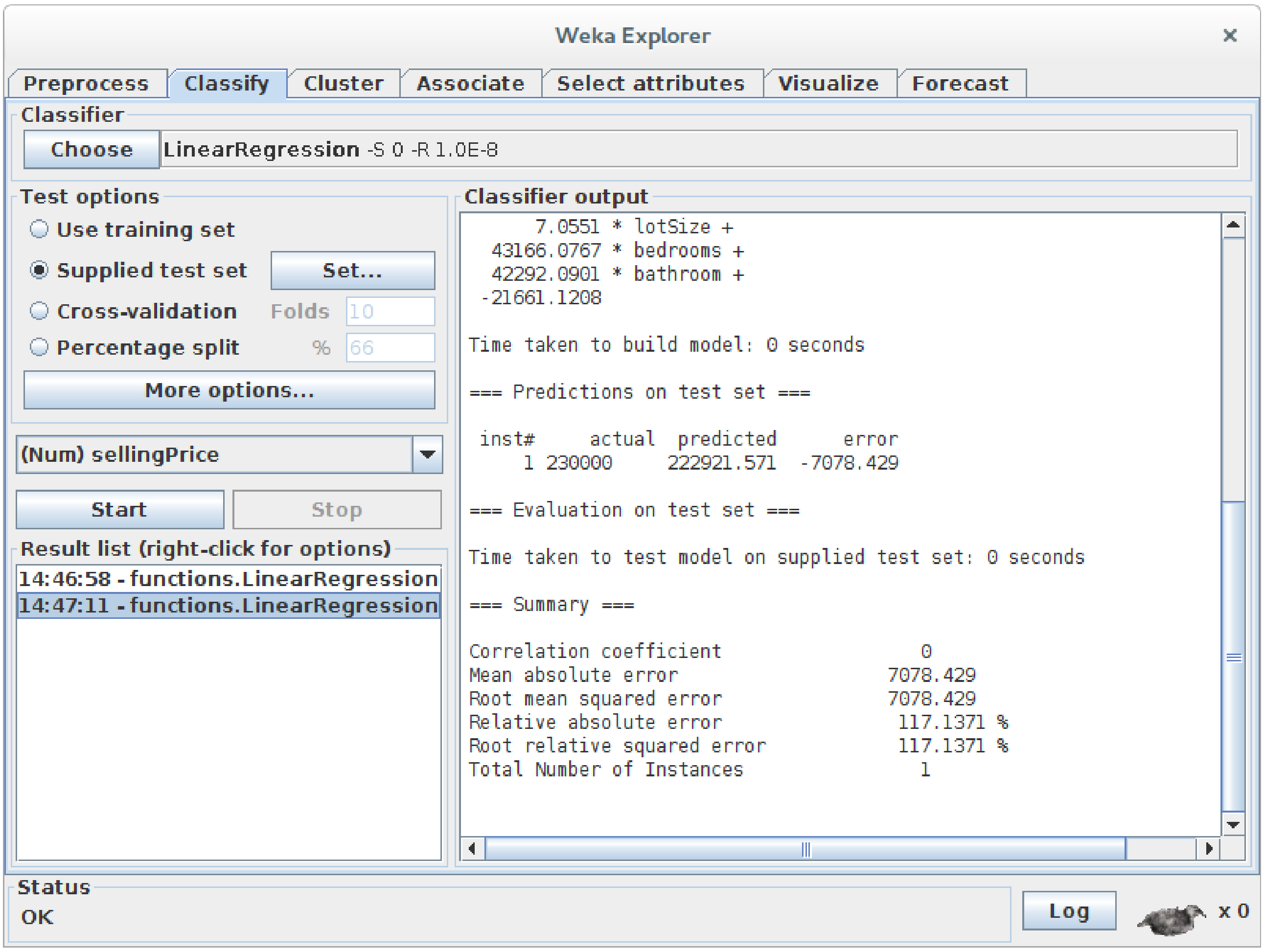
Regression output

```
sellingPrice=(-26.6882*houseSize)+(7.0551*lotSize)+(43166.0767*bedrooms)+
(42292.0901*bathroom)-21661.1208
```

House value using regression model

```
SellingPrice=(-26.6882*2983)+(7.0551*9365)+(43166.0767*5)+(42292.0901*1)        -
21661.1208
```

sellingPrice = 222,921



LinearRegression -S 0 -R 1.0E-8

○ Use training set
● Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66

(Num) sellingPrice

14:46:58 - functions.LinearRegression
14:47:11 - functions.LinearRegression

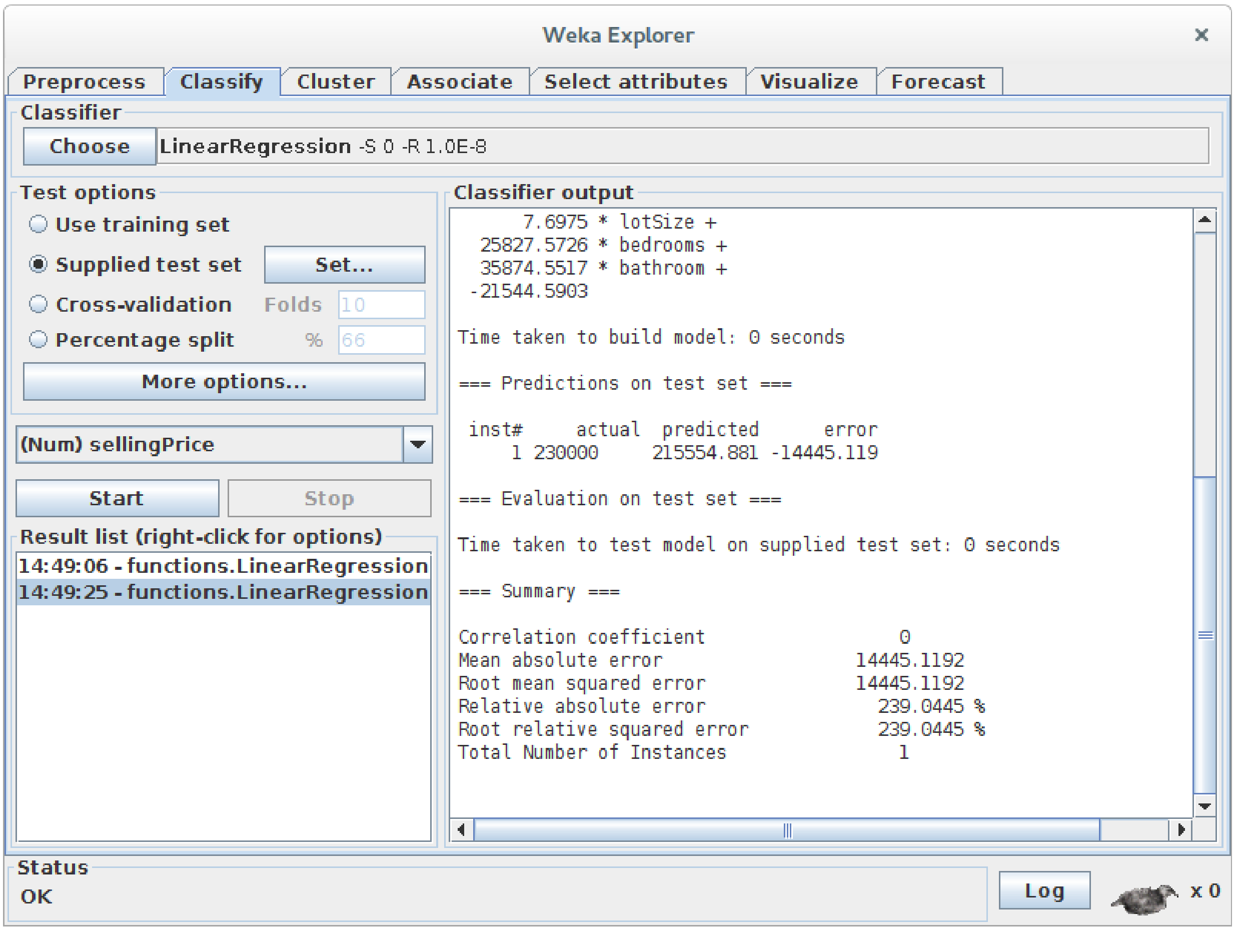
```
      7.0551 * lotSize +
  43166.0767 * bedrooms +
  42292.0901 * bathroom +
 -21661.1208

Time taken to build model: 0 seconds

=== Predictions on test set ===

 inst#     actual  predicted      error
     1 230000      222921.571  -7078.429

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correlation coefficient                  0
Mean absolute error                   7078.429
Root mean squared error               7078.429
Relative absolute error                117.1371 %
Root relative squared error            117.1371 %
Total Number of Instances                1
```

Status
OK

Step4: Interpret the patterns and conclusions that our model tells us

- **Granite doesn't matter :**It will throw out and ignore columns that don't help in creating a good model. So this regression model is telling us that granite in your kitchen doesn't affect the house's value.
- **Bathrooms do matter:** Since we use a simple 0 or 1 value for an upgraded bathroom, we can use the coefficient from the regression model to determine the value of an upgraded bathroom on the house value. The model tells us it adds $42,292 to the house value.
- **Bigger houses reduce the value:** Model is telling us that the bigger our house is, the lower the selling price? This can be seen by the negative coefficient in front of the `houseSize` variable. The model is telling us that every additional square foot of the house reduces its

price by $26? That doesn't make any sense at all. How should we interpret this? The house size, unfortunately, isn't an independent variable because it's related to the bedrooms variable, which makes sense, since bigger houses tend to have more bedrooms. So our model isn't perfect. But we can fix this. On the **Preprocess** tab, you can remove columns from the data set. Remove the **houseSize** column and create another model. How does it affect the price of my house? How does this new model make more sense? (My amended house value: $215,554).



Classifier: LinearRegression -S 0 -R 1.0E-8

Test options

○ Use training set

● Supplied test set

○ Cross-validation Folds 10

○ Percentage split % 66

(Num) sellingPrice

Result list (right-click for options)

14:49:06 - functions.LinearRegression

14:49:25 - functions.LinearRegression

Classifier output

```
      7.6975 * lotSize +
  25827.5726 * bedrooms +
  35874.5517 * bathroom +
 -21544.5903

Time taken to build model: 0 seconds

=== Predictions on test set ===

 inst#     actual  predicted      error
     1 230000     215554.881 -14445.119

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correlation coefficient                  0
Mean absolute error                  14445.1192
Root mean squared error              14445.1192
Relative absolute error                239.0445 %
Root relative squared error            239.0445 %
Total Number of Instances                1
```

Status

OK