Machine Learning and Weka

Tong Liu, 10/12/2018

Dip. Informatica, Mura Anteo Zamboni 7

BOLOGNA BUSINESS SCHOOL
Alma Mater Studiorum Università di Bologna

Outline

- Overview: machine learning (ML) in AI
- What is ML ?
- Supervised ML vs. Unsupervised ML
- Types of prediction problems (Supervised ML)
- Learning workflow (supervised & unsupervised)
- Classical ML methods vs. Deep Learning (DL)
- Get started with Weka
- Demo

Overview: ML in AI

Artificial Intelligence

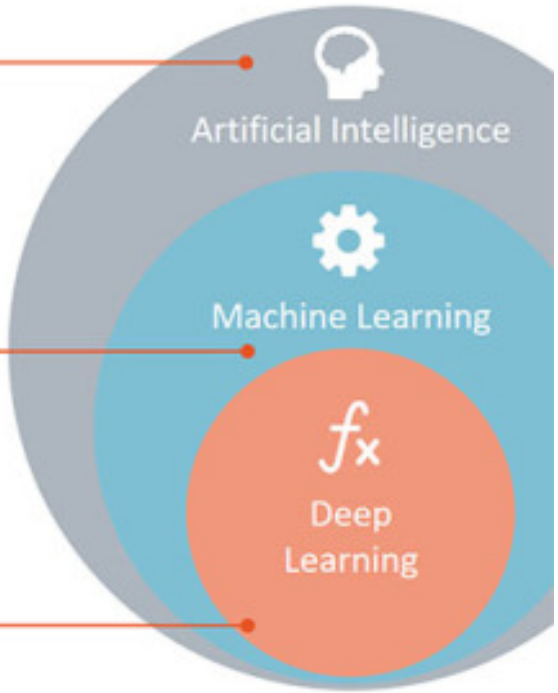
Any technique which enables computers to mimic human behavior.

Machine Learning

Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

Deep Learning

Subset of ML which make the computation of multi-layer neural networks feasible.



[Rapidminer]

What is the ML ?

- Alan Turing (1912-1954)
- “what we want is a **machine** that can learn from *experience*.”

[The Imitation Game]

Experience for machines ?

[The Imitation Game]

Dataset - a kind of experience

- A dataset with attributes (a.k.a features) that can be processed by a computer.

Size of House	Lot Size (acre)	# of Bedrooms	# of Bathrooms	Price of House
950	2.5	2	1	$127,325
1,535	1.5	2	2	$156,570
1,605	2.25	3	1.5	$158,895
1,905	2.5	2	1.5	$200,025
2,057	2.25	3	2	$230,384
2,227	2.75	3	2	$233,835
3,150	1	4	2	$261,420
3,620	3	4	3	$433,500

House prices dataset

Dataset - supervised vs. unsupervised

- Supervised ML: data are labelled and an interest is given - learn with a purpose
- Unsupervised ML: no specific interest is given - exploratory analysis

Size of House	Lot Size (acre)	# of Bedrooms	# of Bathrooms	Price of House
950	2.5	2	1	$127,325
1,535	1.5	2	2	$156,570
1,605	2.25	3	1.5	$158,895
1,905	2.5	2	1.5	$200,025
2,057	2.25	3	2	$230,384
2,227	2.75	3	2	$233,835
3,150	1	4	2	$261,420
3,620	3	4	3	$433,500

House prices dataset

Types of Prediction problems (supervised learning)

- Regression: the output variable takes continuous values.
- Classification: the output variable takes class labels.

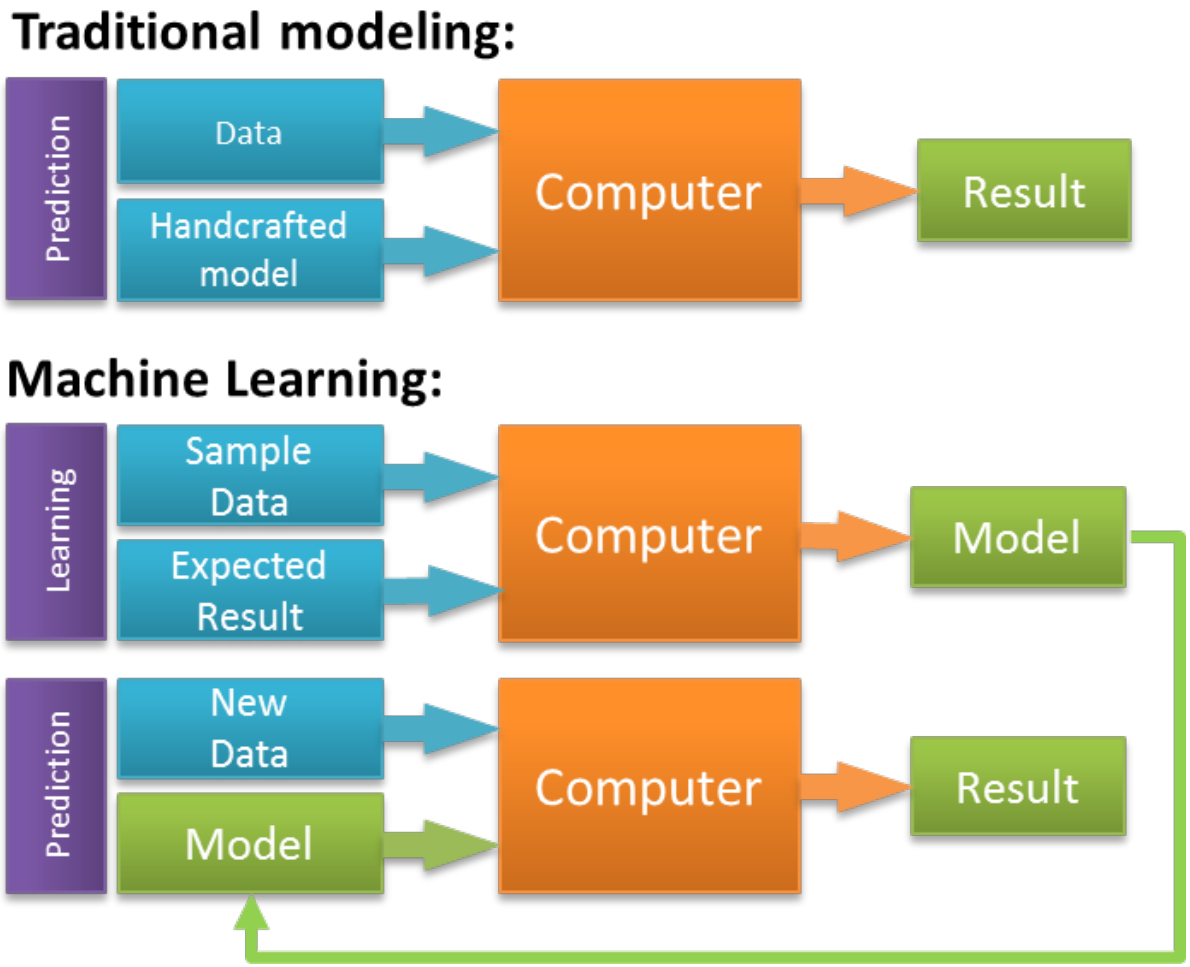
Size of House	Lot Size (acre)	# of Bedrooms	# of Bathrooms	Price of House
950	2.5	2	1	$127,325
1,535	1.5	2	2	$156,570
1,605	2.25	3	1.5	$158,895
1,905	2.5	2	1.5	$200,025
2,057	2.25	3	2	$230,384
2,227	2.75	3	2	$233,835
3,150	1	4	2	$261,420
3,620	3	4	3	$433,500

House prices dataset

How we work with dataset?

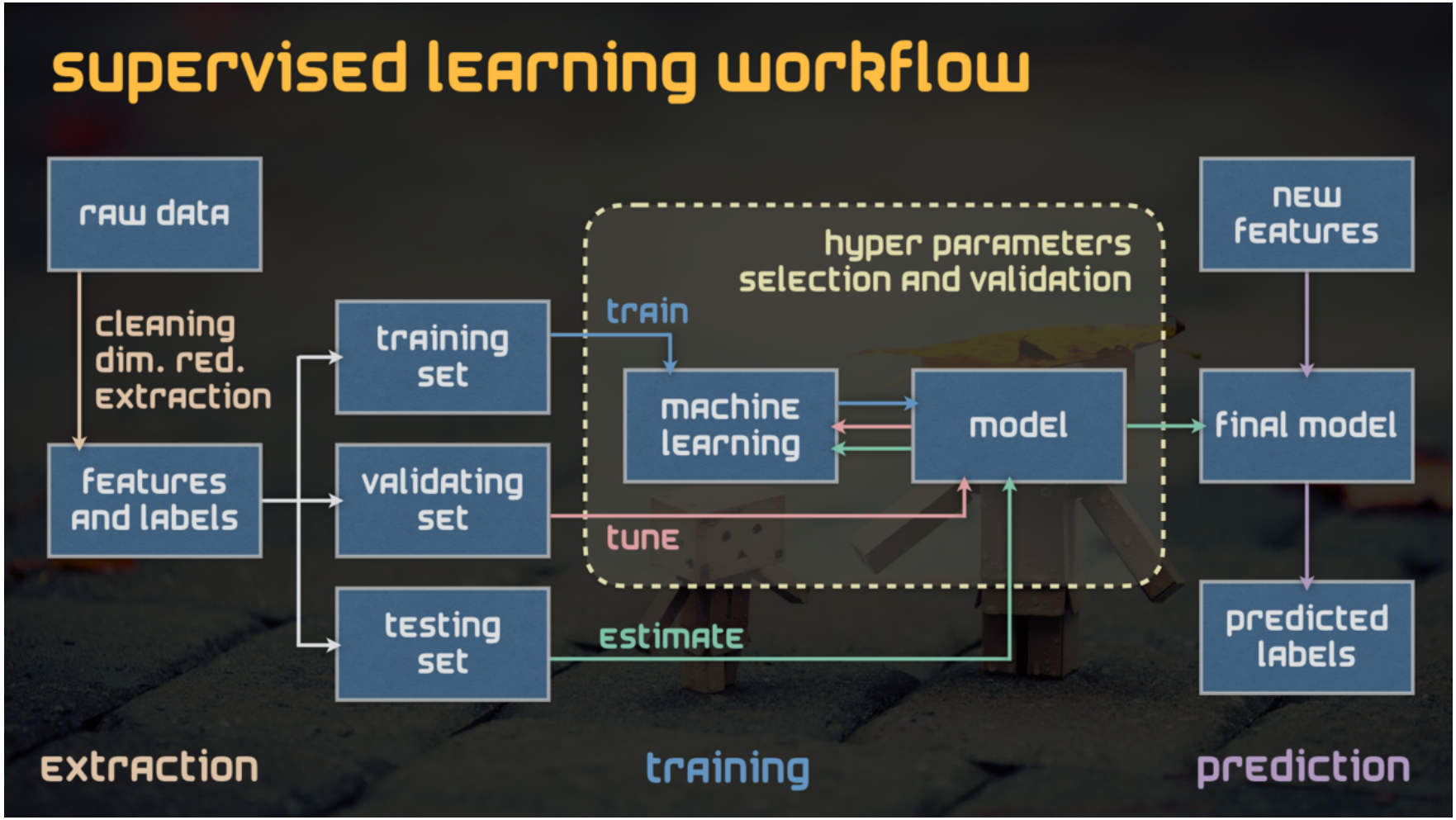
Constraint Programming

Automatic Learning



[Zeiss]

Supervised ML workflow



[BidMotion]

Classical ML vs. Deep Learning

CLASSIC MACHINE LEARNING

How do you engineer the best features?

FEATURES

- Roundness of face;
- Distance between eyes;
- Nose width;
- Eye socket depth;
- Cheek bone structure;
- Jaw line length…

CLASSIFIER ALGORITHMS

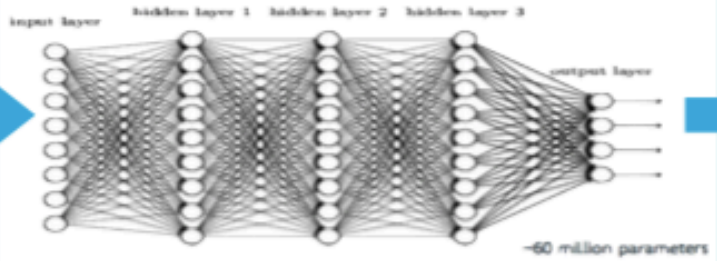
- SVM;
- Random Forest;
- Naïve Bayes;
- Decision Tree;
- Regression
- …many more…

Jim Carrey

DEEP LEARNING

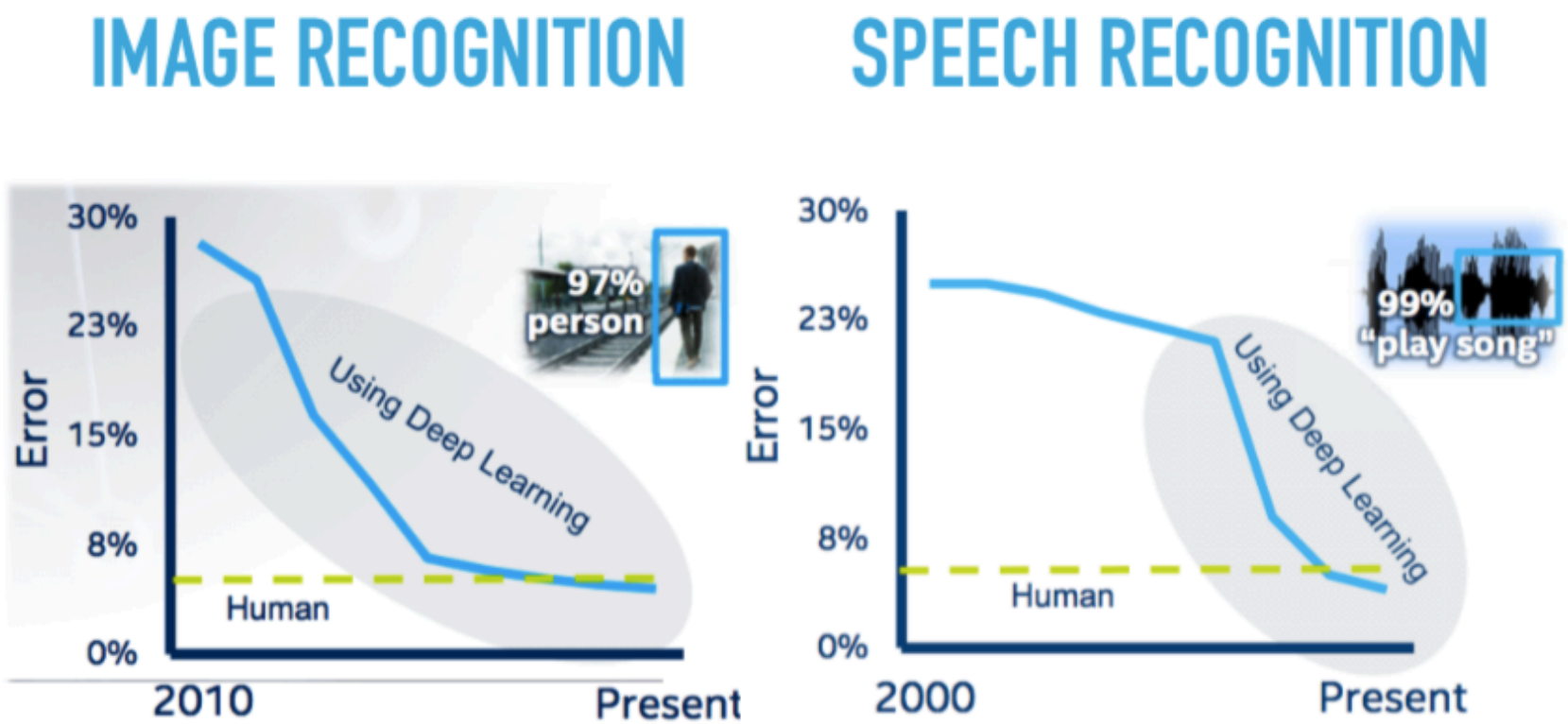
How do you guide the model to find the best features?

DEEP NEURAL NETWORK



Jim Carrey

Deep Learning breakthroughs

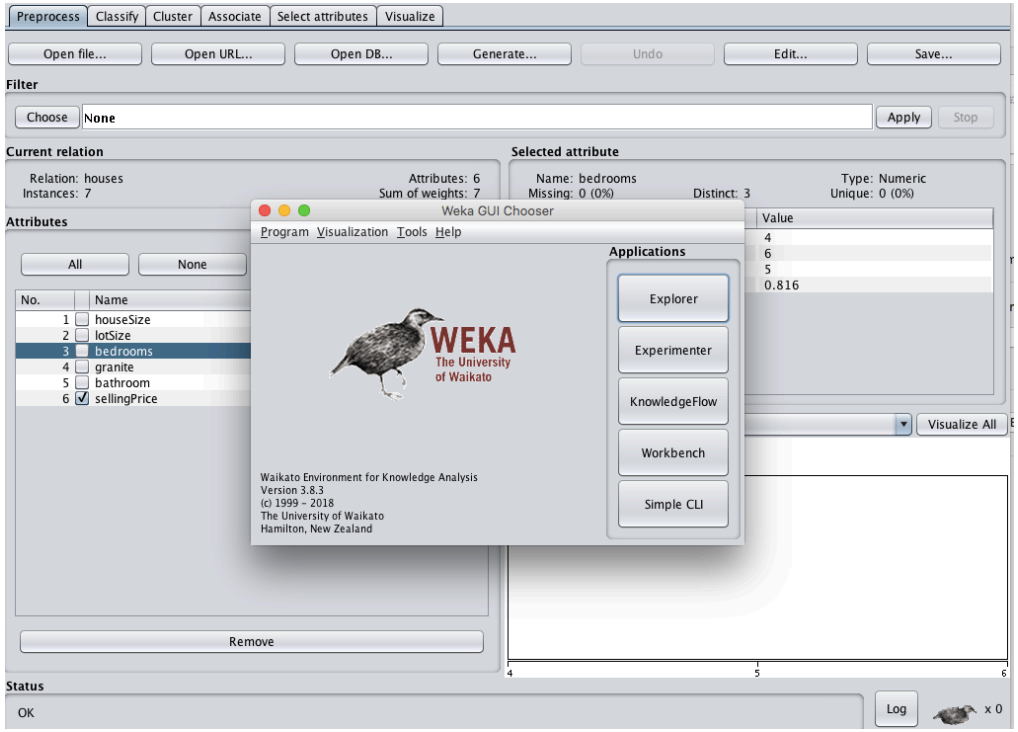


MACHINES ABLE TO MEET OR EXCEED HUMAN IMAGE & SPEECH RECOGNITION (TO SOME EXTEND...)

A simple example of prediction

WEKA is developed by the University of Waikato (New Zealand) under the GNU General Public License (GPL).

It is written in the JavaTM object-oriented programming language and provides a GUI for interacting with data files and producing visual results.



Collected data

House size (square feet)	Lot size	Bedrooms	Granite	Upgraded bathroom?	Selling price
3529	9191	6	0	0	205000
3247	10061	5	1	1	224900
4032	10150	5	0	1	197900
2397	14156	4	1	0	189900
2200	9600	4	0	1	195000
3536	19994	6	1	1	325000
2983	9365	5	0	1	230000
3198	9669	5	1	1	????

Collected data

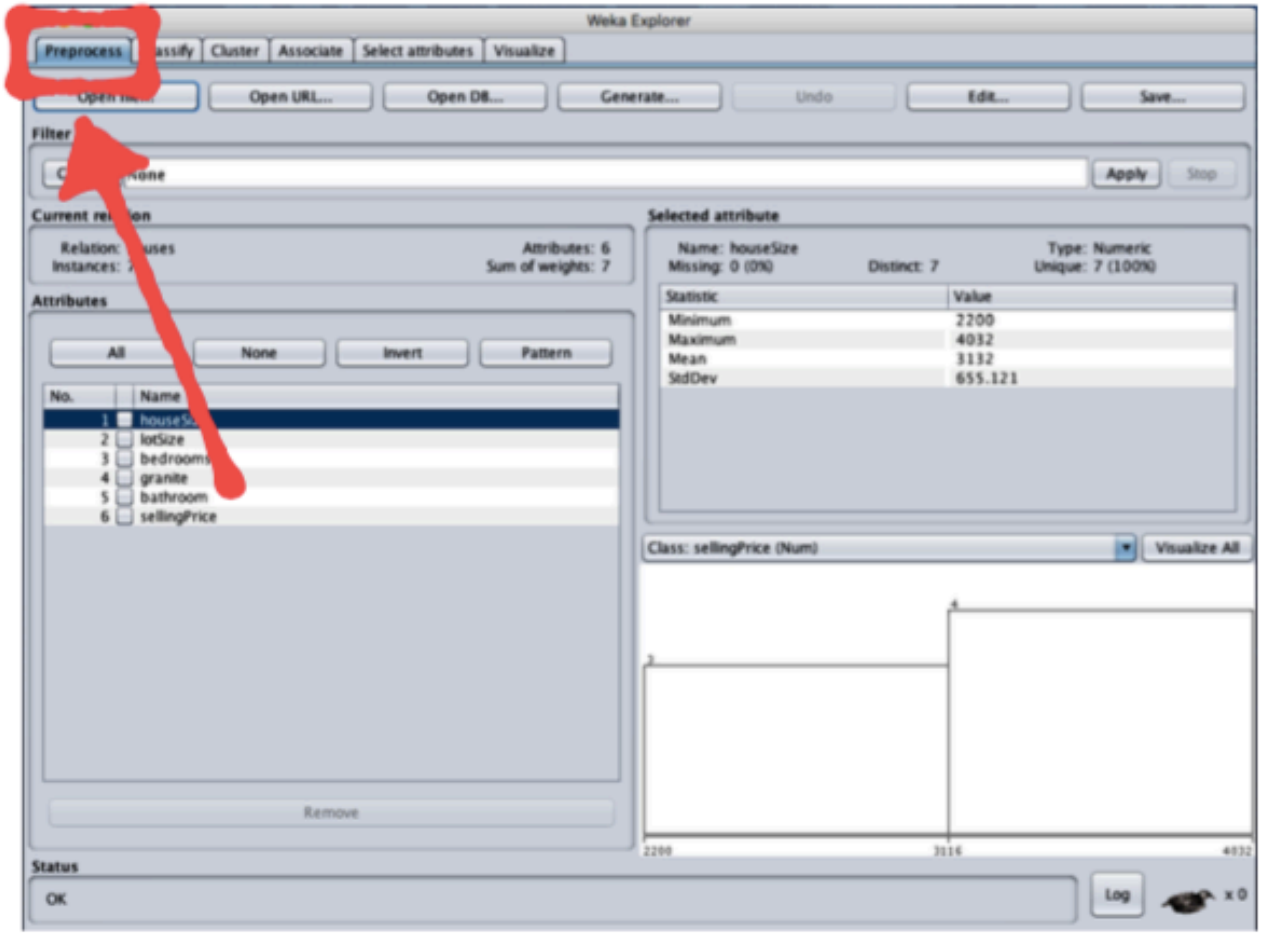
Training Data

House size (square feet)	Lot size	Bedrooms	Granite	Upgraded bathroom?	Selling price
3529	9191	6	0	0	205000
3247	10061	5	1	1	224900
4032	10150	5	0	1	197900
2397	14156	4	1	0	189900
2200	9600	4	0	1	195000
3536	19994	6	1	1	325000
2983	9365	5	0	1	230000
3198	9669	5	1	1	????

Test Data

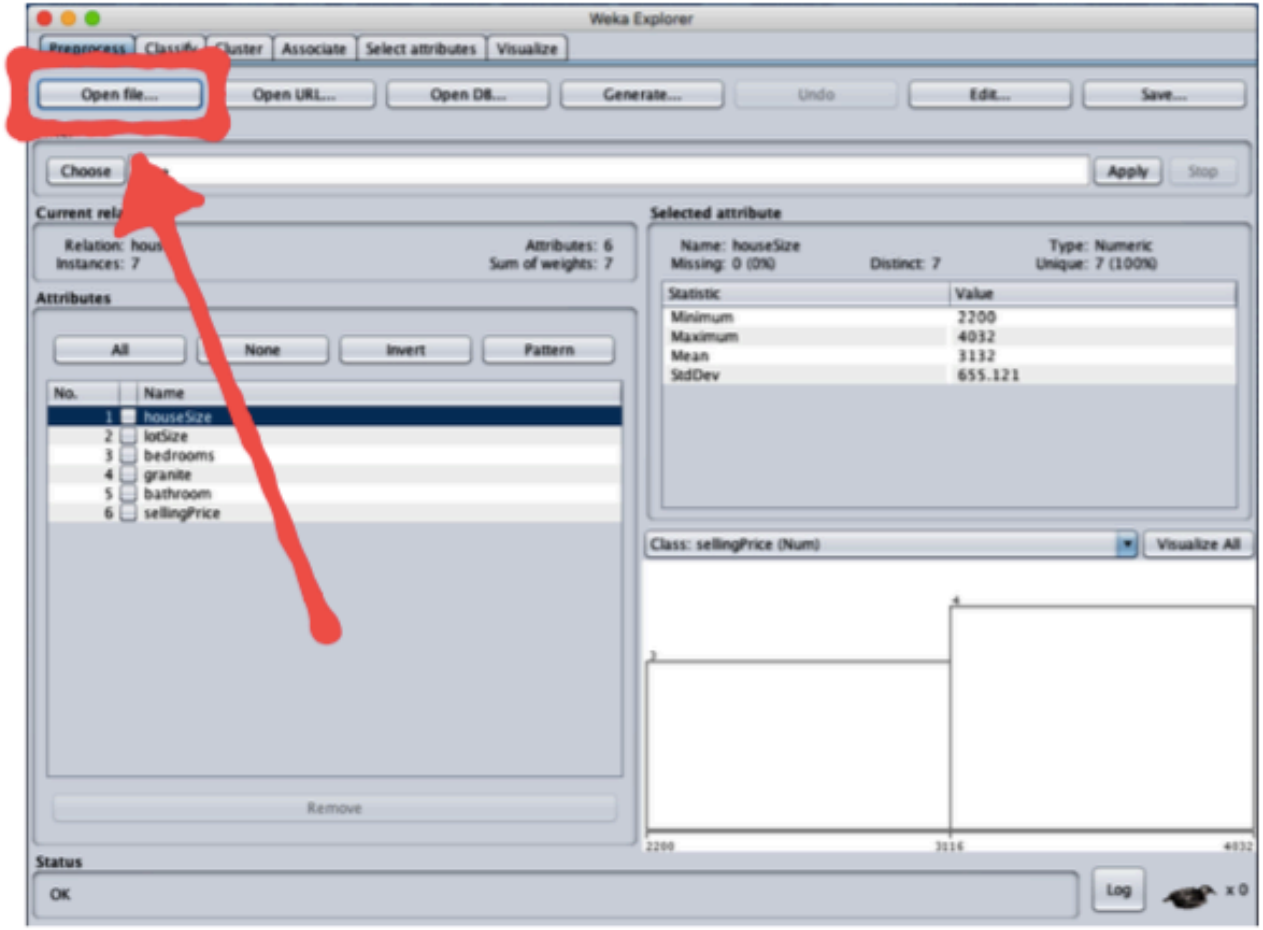
Loading dataset

- Preprocess Tab



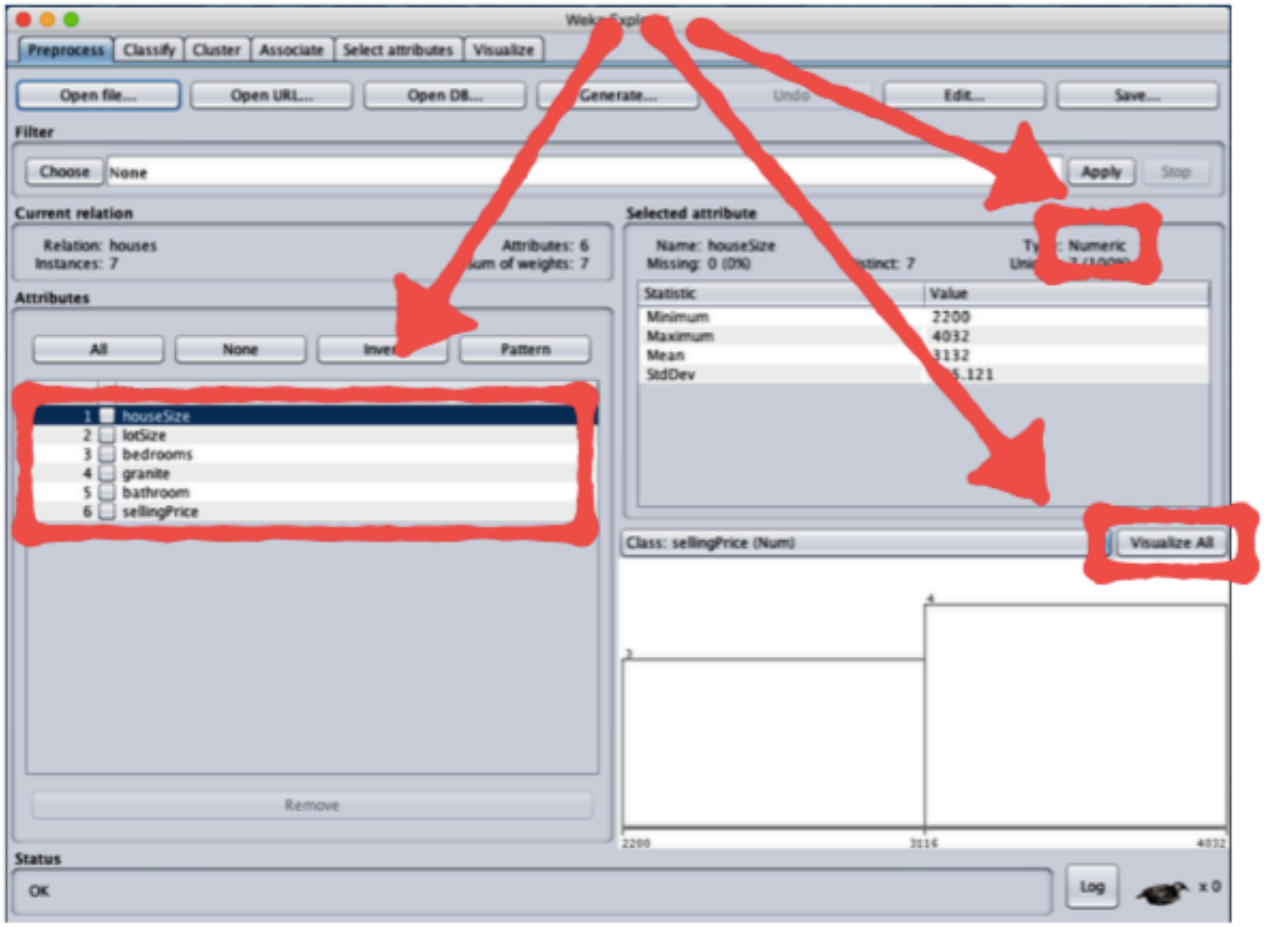
Loading dataset

- Preprocess Tab
- Open File:
  - CSV format
  - XLS format
  - ARFF format



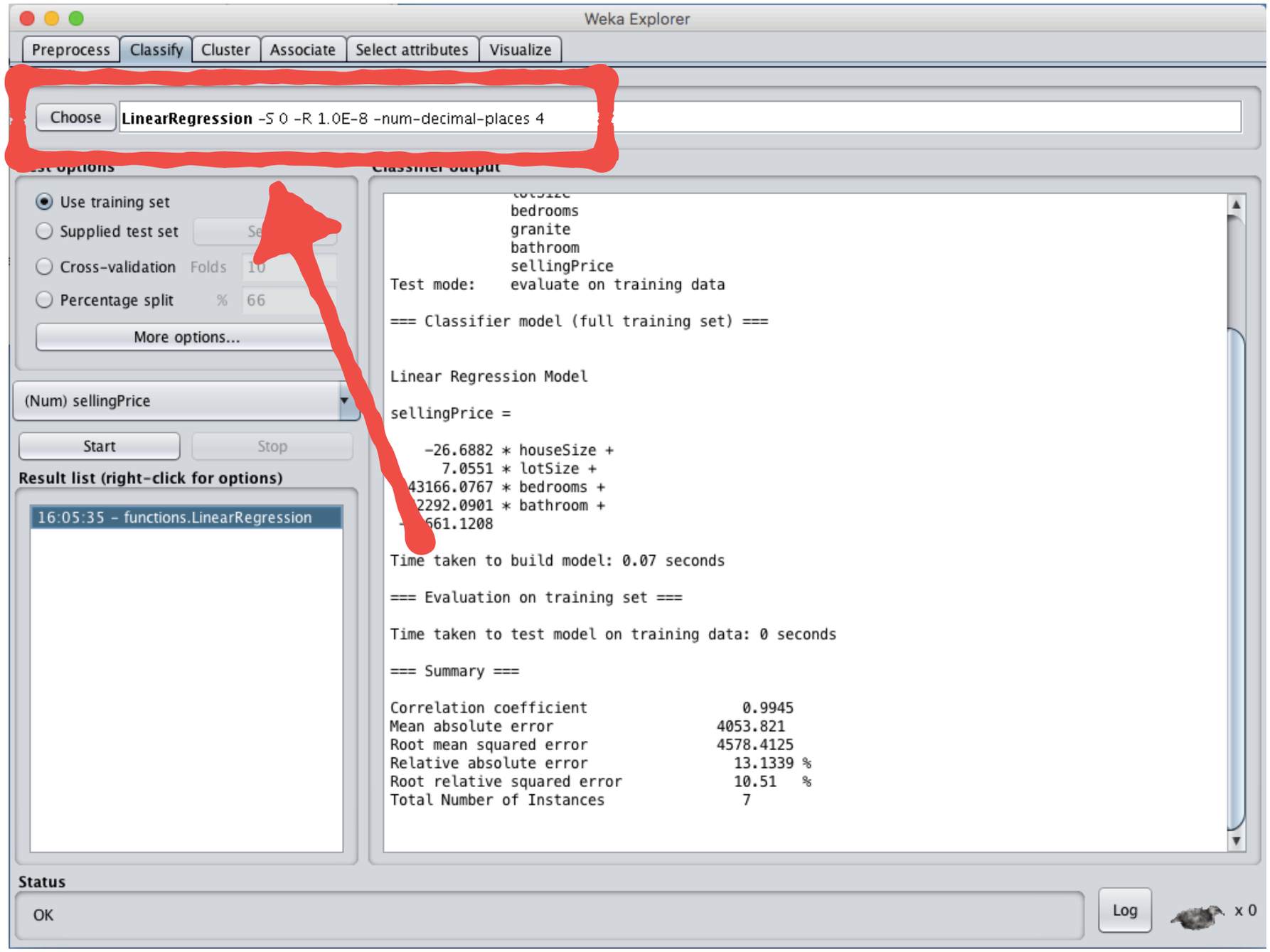
Loading dataset

- **Numerical *variables***
  data with value representable with numbers
- **Visualize All**
  shows all graphics at once



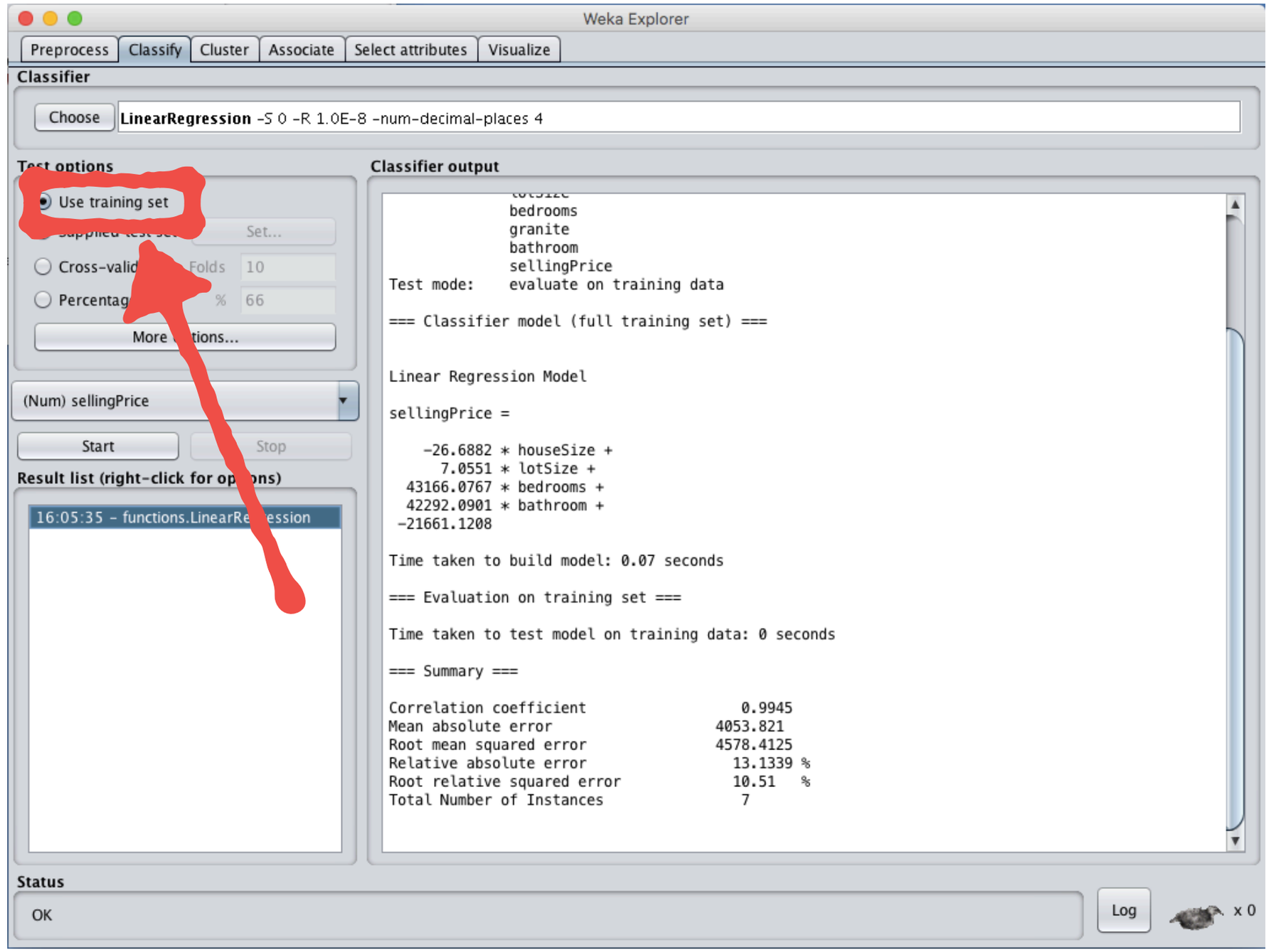
CLASSIFY THE DATA WITH WEKA
LINEAR REGRESSION MODEL

- Classify Tab



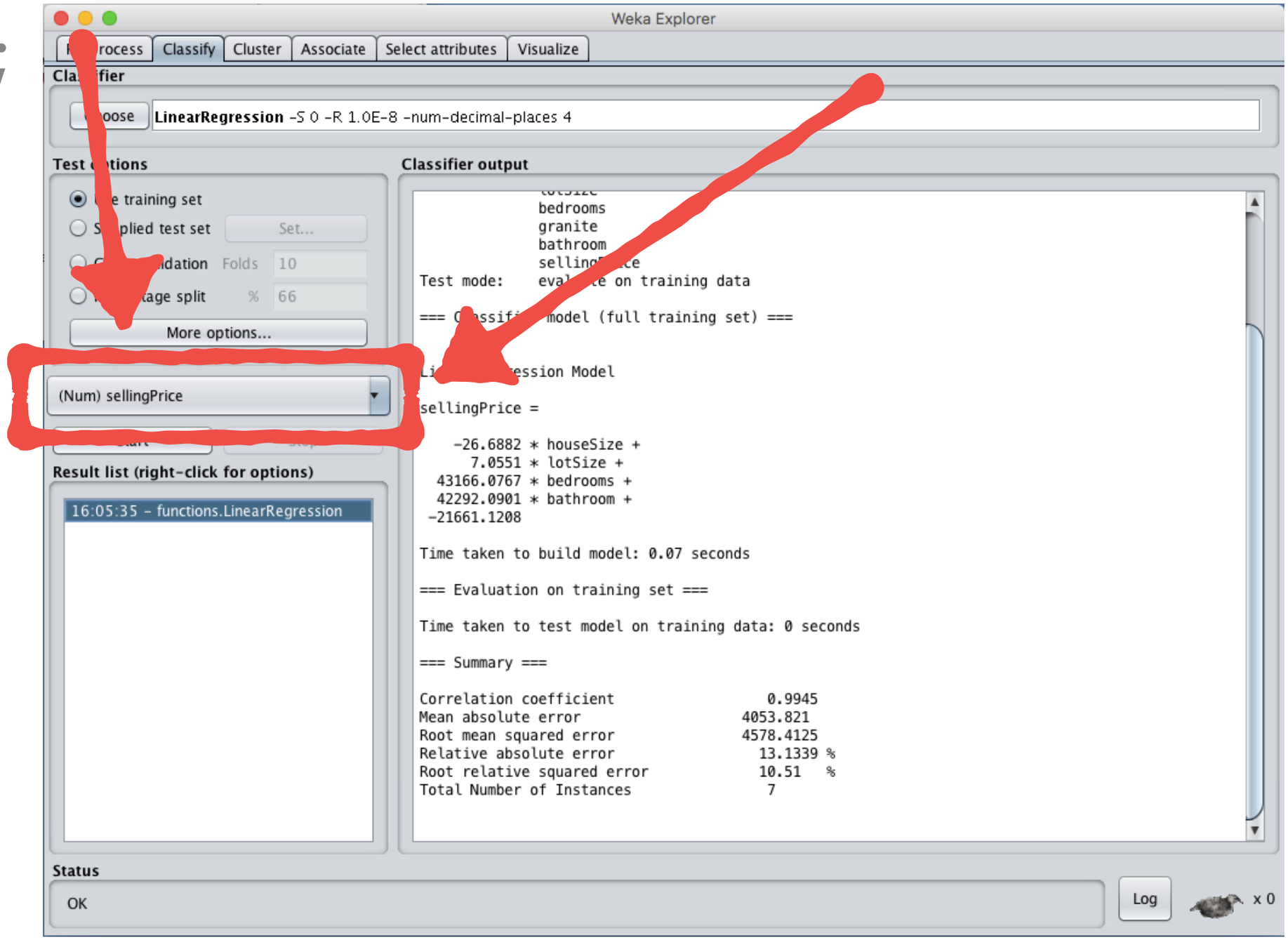
CLASSIFY THE DATA WITH WEKA
LINEAR REGRESSION MODEL

- Classify Tab
- Use training set



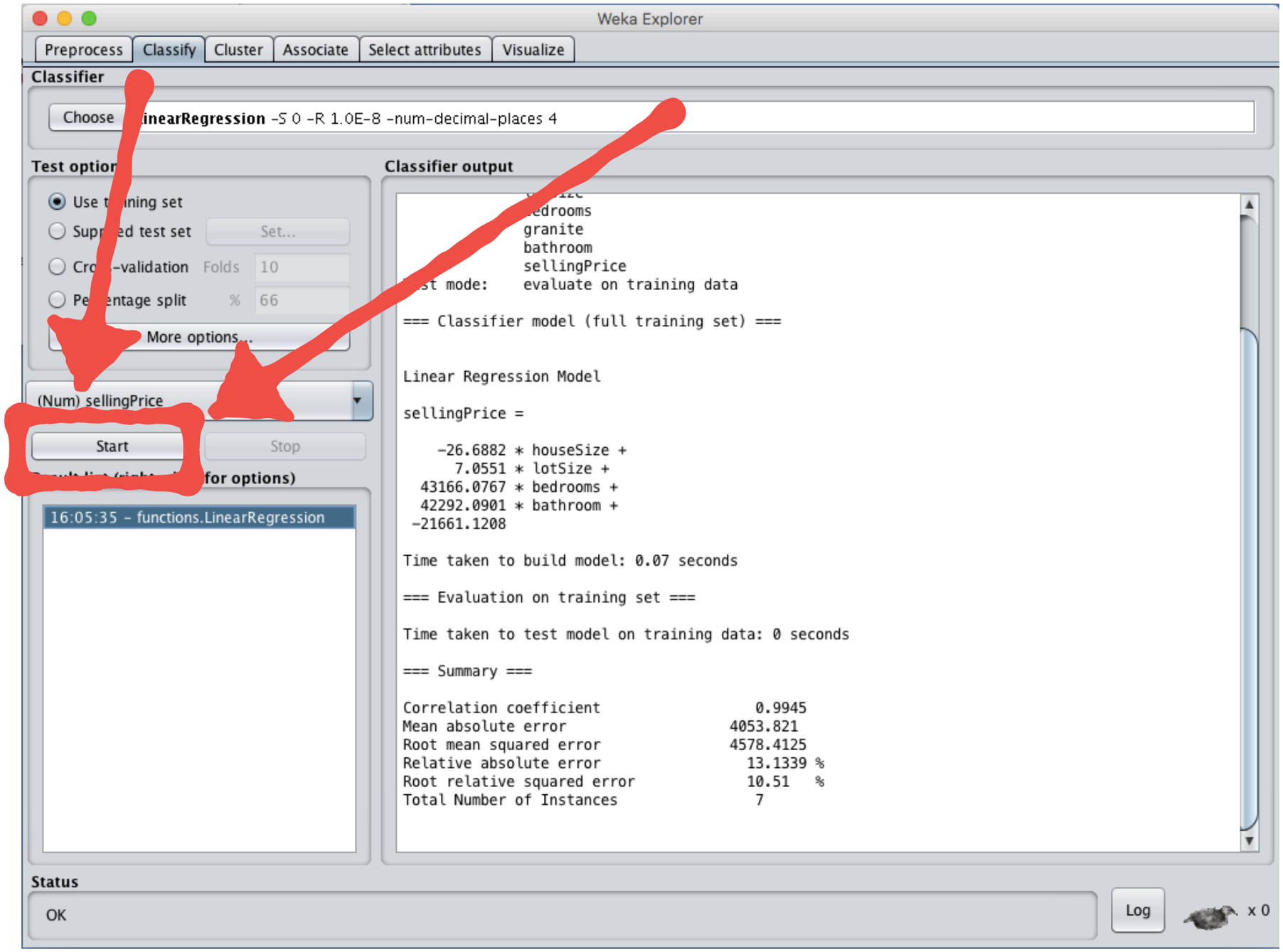
CLASSIFY THE DATA WITH WEKA
LINEAR REGRESSION MODEL

- Classify Tab;
- Use training set;
- Class = sellingPrice;



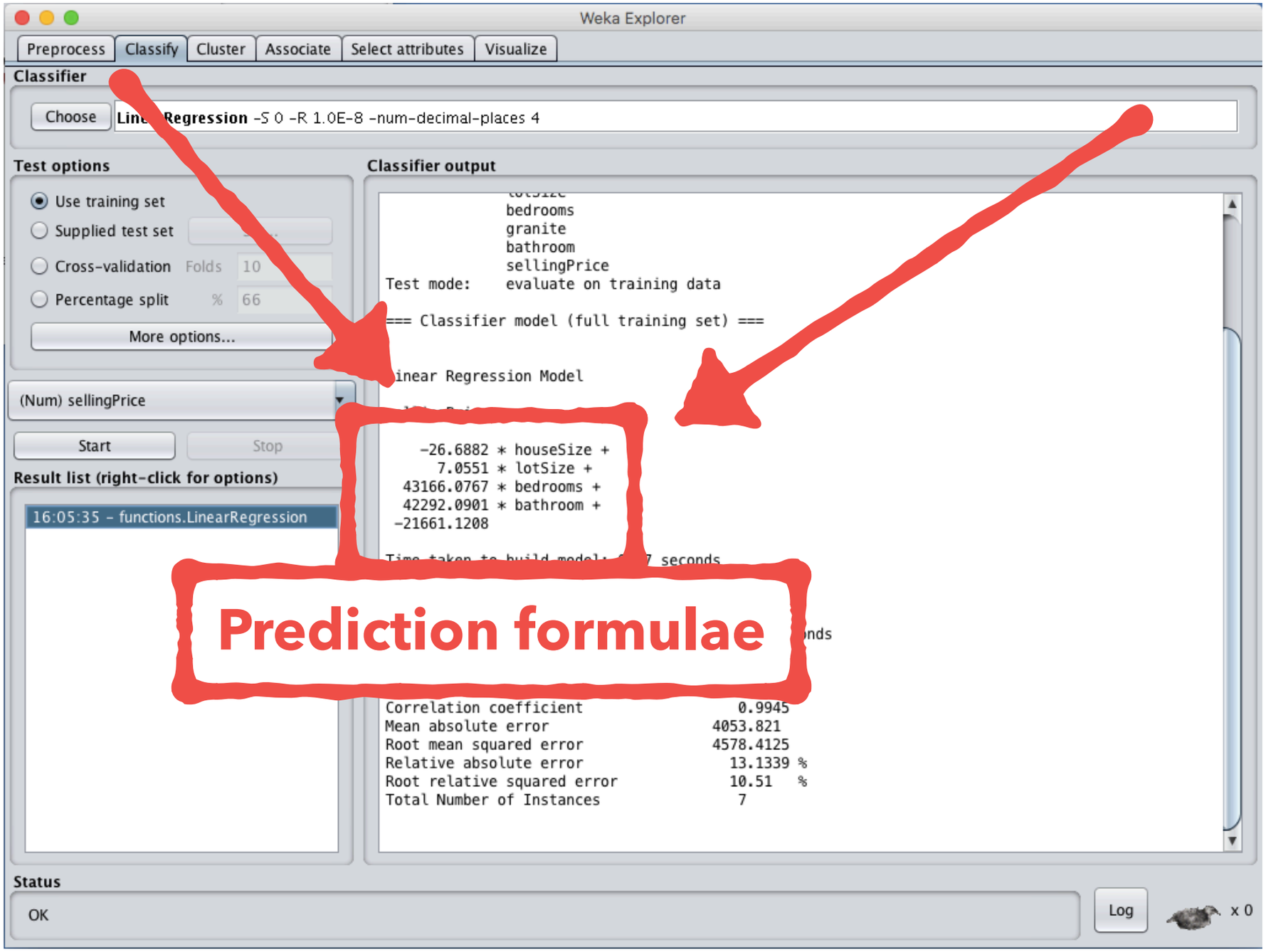
CLASSIFY THE DATA WITH WEKA
LINEAR REGRESSION MODEL

- Classify Tab;
- Use training set;
- Class = sellingPrice;
- Start building the model



CLASSIFY THE DATA WITH WEKA
LINEAR REGRESSION MODEL

- Classify Tab;
- Use training set;
- Class = sellingPrice;
- Start building the model



FINAL PREDICTION

sellingPrice =

- 26,68 * [**houseSize = 3198**]

+ 7,05 * [**lotSize = 9669**]

+ 43.166,07 * [**bedrooms = 5**]

+ 42.292,09 * [**bathroom = 1**]

- 21.661,12 = **219.328,25**

ANOTHER EXAMPLE OF CLASSIFICATION
CAR DEALERSHIP

The dealership is **starting a promotional campaign**, whereby it is **trying to push a two-year extended warranty** to its past customers.

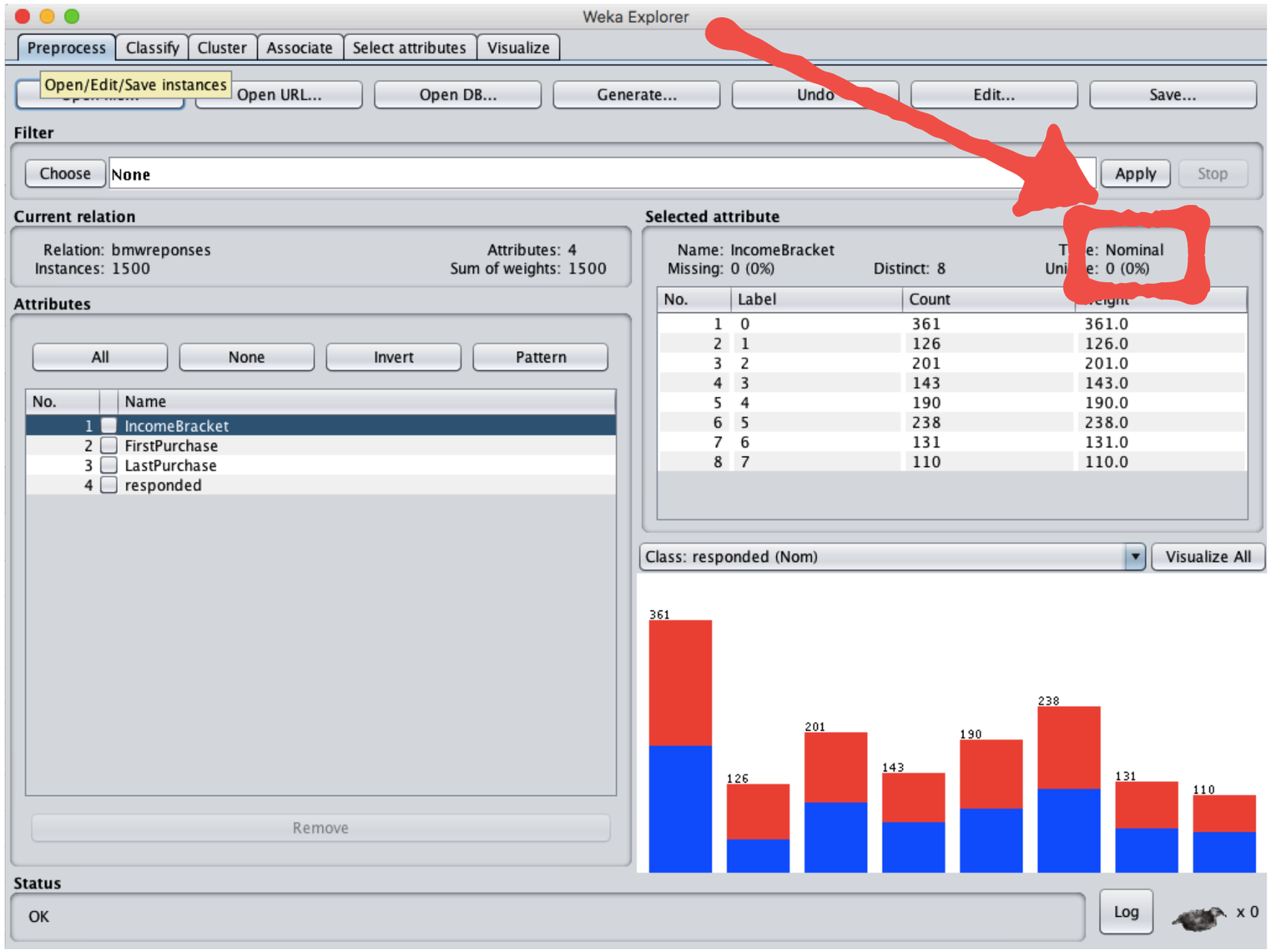
The dealership **has done this before** and has gathered **4,500 data points from past** sales of extended warranties.

The attributes in the data set are:

- Income bracket [0=$0-$30k, 1=$31k-$40k, 2=$41k-$60k, 3=$61k-$75k, 4=$76k-$100k, 5=$101k-$150k, 6=$151k-$500k, 7=$501k+]
- Year/month first car bought
- Year/month most recent car bought
- **Whether they responded or not** to the extended warranty offer in the past

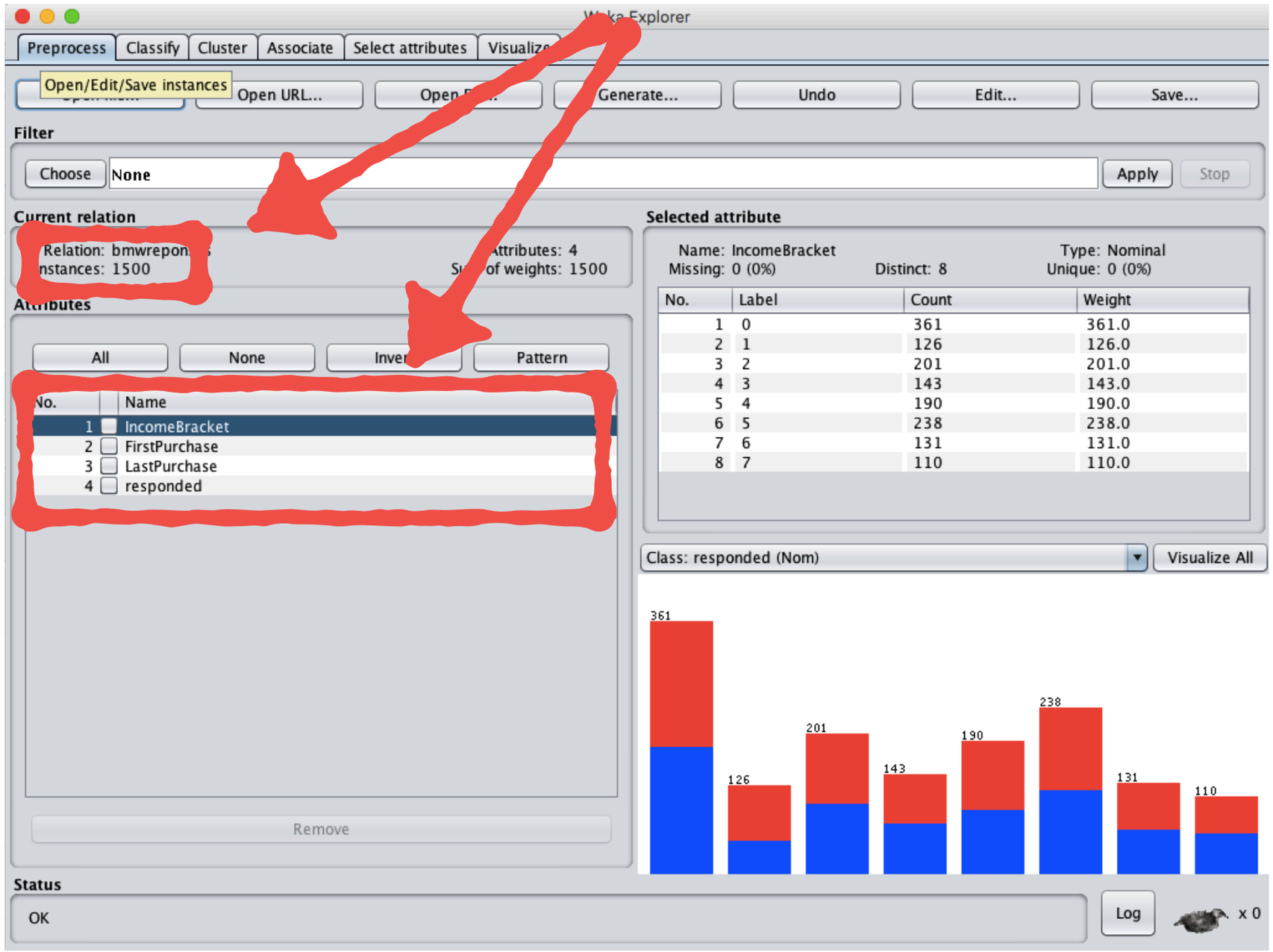
PREPROCESS THE DATA WITH WEKA
LOAD THE DATASET FOR TRAINING

- **Nominal *variables*** – *labelled data*



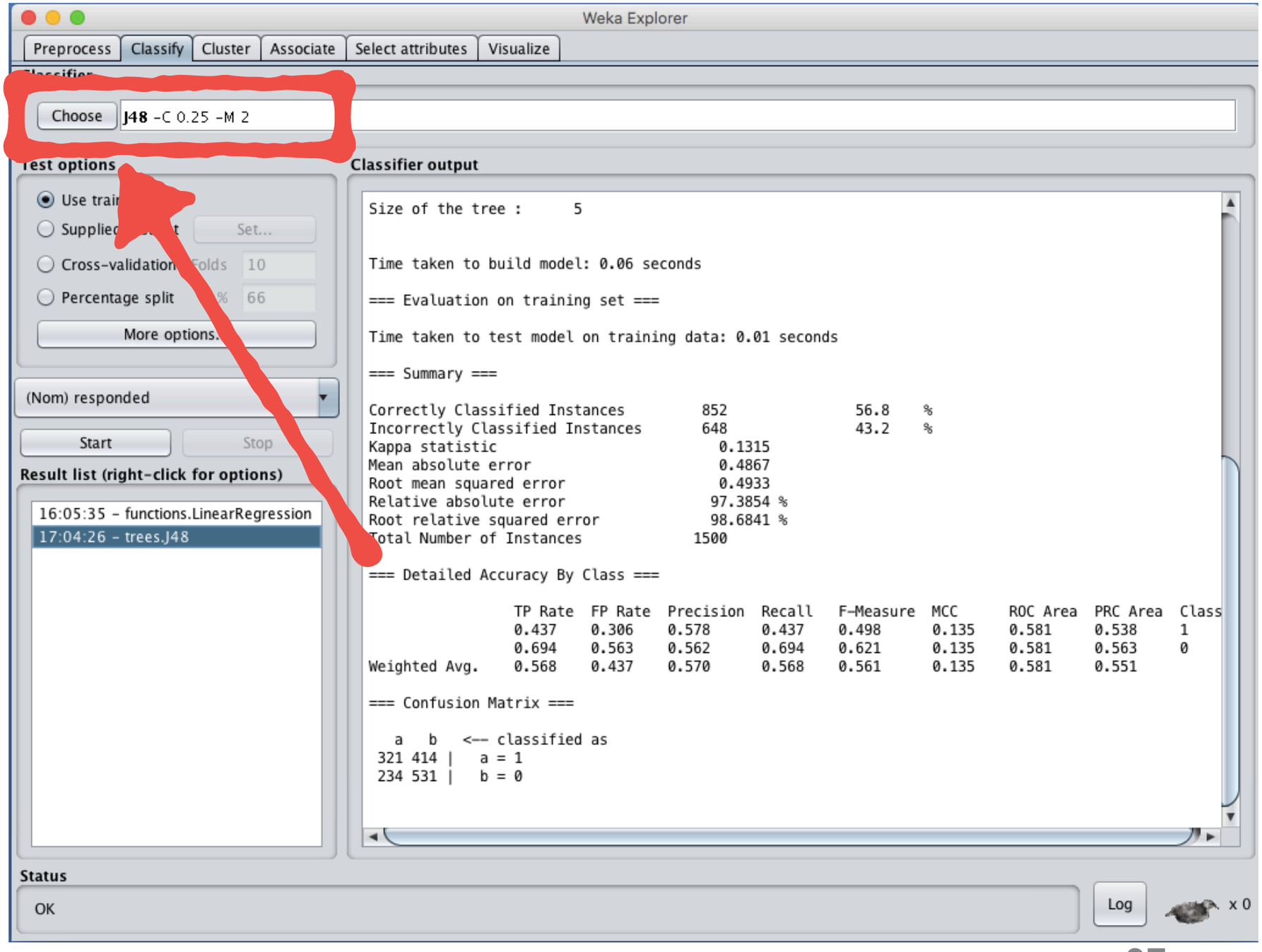
PREPROCESS THE DATA WITH WEKA
LOAD THE DATASET FOR TRAINING

- **Nominal *variables*** – *labelled data*
- **1500 instances**



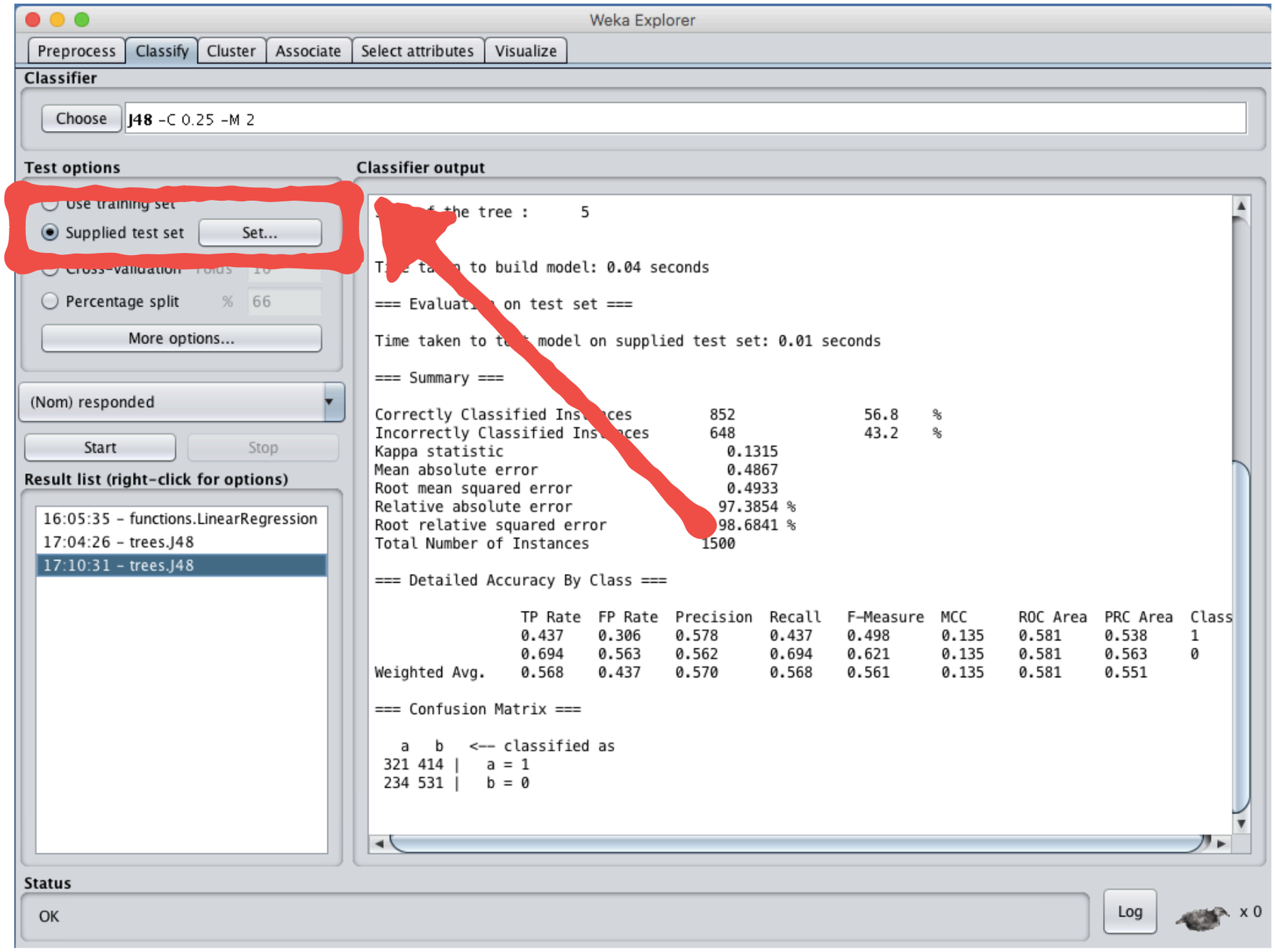
CLASSIFY THE DATA WITH WEKA
DECISION TREE MODEL TRAINING

- Classify Tab;
- Use training set;
- Class = responded;
- Start building the model



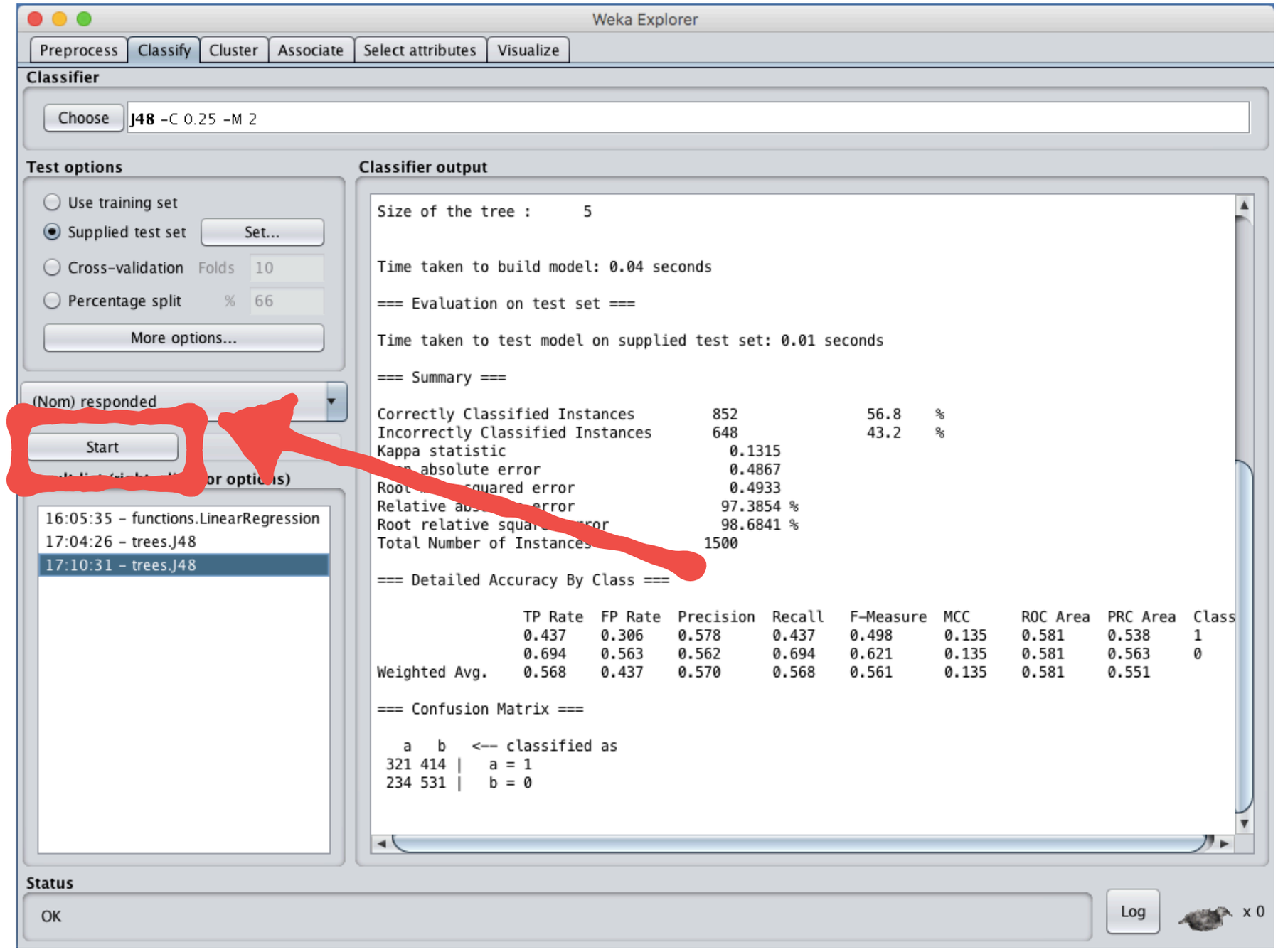
CLASSIFY THE DATA WITH WEKA
DECISION TREE MODEL TESTING

- Supplied test set;



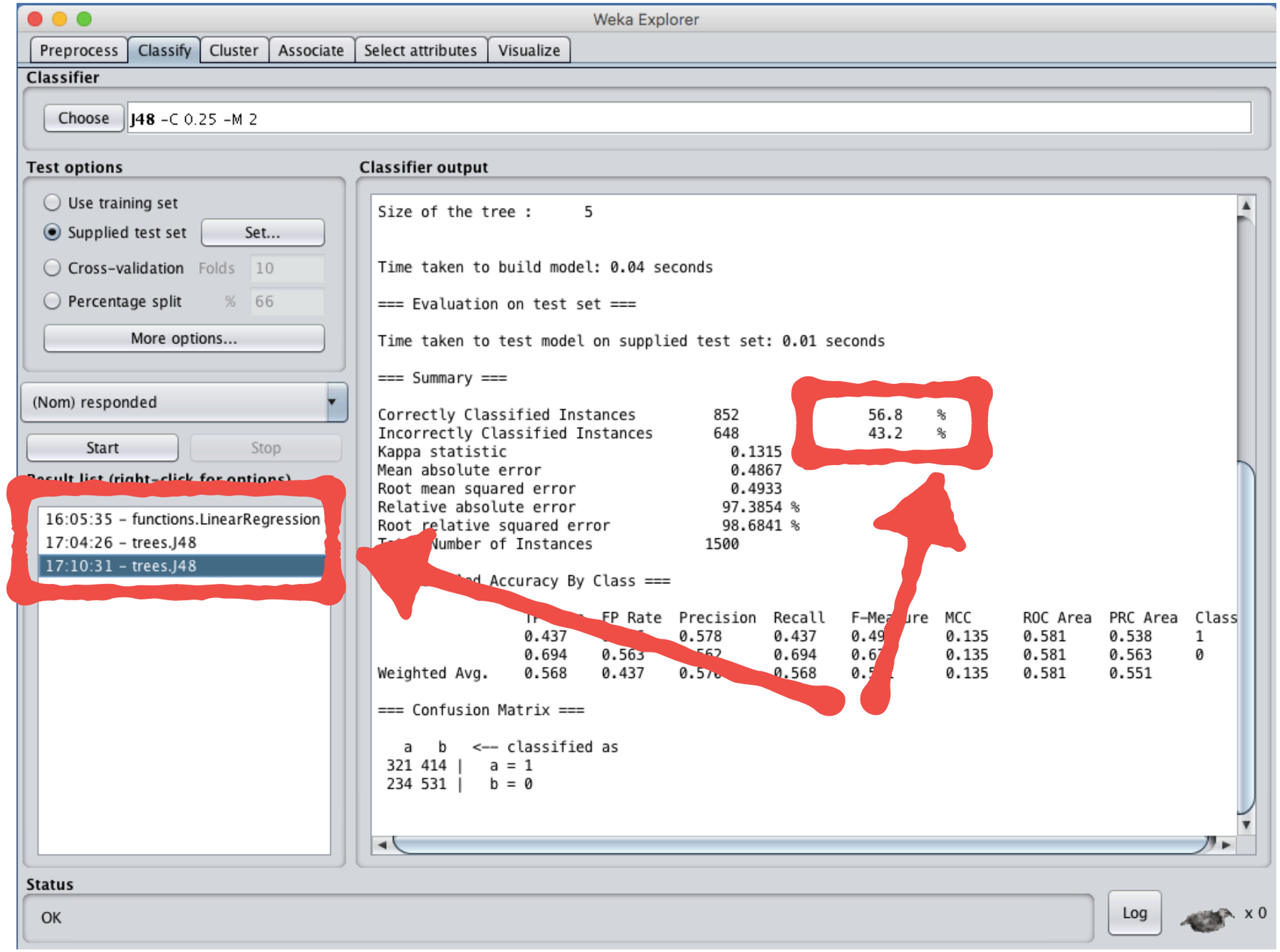
CLASSIFY THE DATA WITH WEKA
DECISION TREE MODEL TESTING

- Supplied test set;
- Start testing the model;



CLASSIFY THE DATA WITH WEKA
DECISION TREE MODEL TESTING

- Supplied test set;
- Start testing the model;
- Compare models accuracy between train and test.



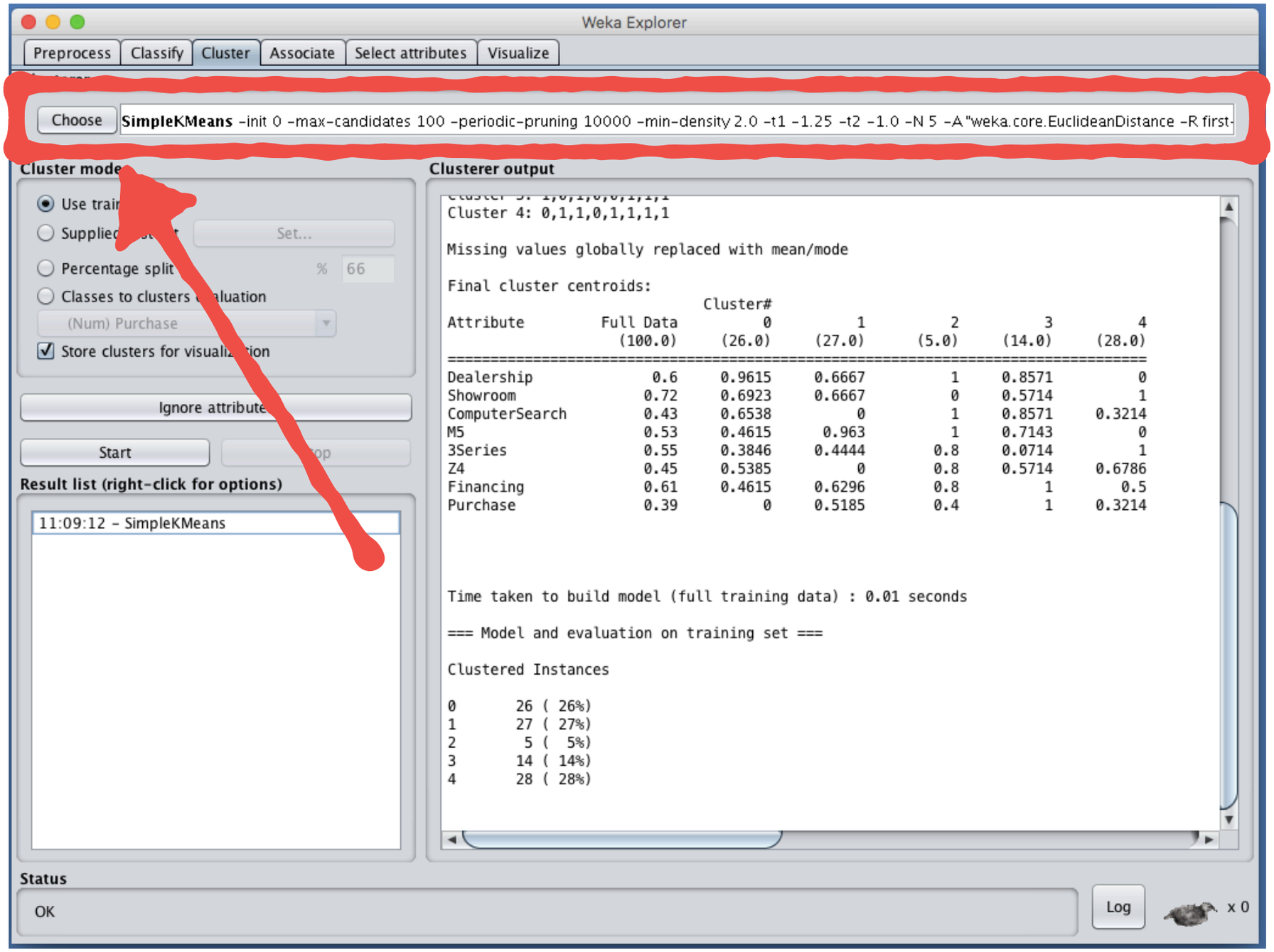
AN EXAMPLE OF CLUSTERING
CAR DEALERSHIP BEHAVIOUR ANALYSIS

The dealership has **kept track of how people walk through the dealership** and the showroom, what cars they look at, and how often they ultimately make purchases.

They are hoping to **mine this data by finding patterns** in the data and by using clusters to determine **if certain behaviours in their customers emerge**.

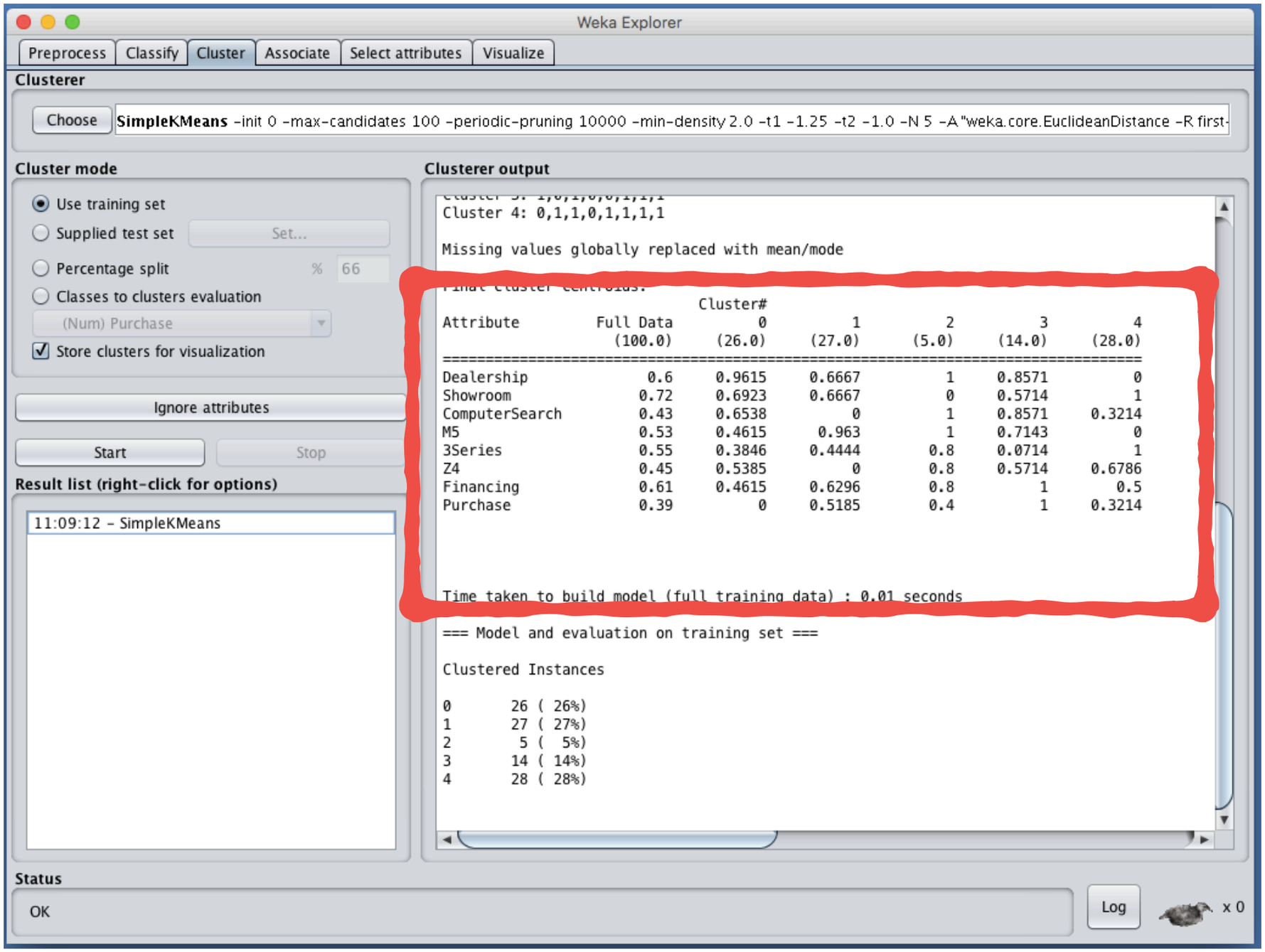
CLUSTERING THE DATA WITH WEKA
K-MEANS BEHAVIOUR ANALYSIS

- Cluster Tab;
- Use training set;
- No Class;
- Start



CLUSTERING THE DATA WITH WEKA
K-MEANS BEHAVIOUR ANALYSIS

- Cluster Tab;
- Use training set;
- No Class;
- Start;
- Evaluate patterns.



```
                                Cluster#
Attribute          Full Data          0          1          2          3          4
                     (100.0)     (26.0)     (27.0)      (5.0)     (14.0)     (28.0)
==================================================================================
Dealership               0.6     0.9615     0.6667          1     0.8571          0
Showroom                0.72     0.6923     0.6667          0     0.5714          1
ComputerSearch          0.43     0.6538          0          1     0.8571     0.3214
M5                      0.53     0.4615      0.963          1     0.7143          0
3Series                 0.55     0.3846     0.4444        0.8     0.0714          1
Z4                      0.45     0.5385          0        0.8     0.5714     0.6786
Financing               0.61     0.4615     0.6296        0.8          1        0.5
Purchase                0.39          0     0.5185        0.4          1     0.3214


Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       26 ( 26%)
1       27 ( 27%)
2        5 (  5%)
3       14 ( 14%)
4       28 ( 28%)
```

EXAMPLE CONCLUSION
K-MEANS BEHAVIOUR ANALYSIS

- **Cluster 0**– This group we can call the "Dreamers," as they appear to wander around the dealership, looking at cars parked outside on the lots, but trail off when it comes to coming into the dealership, and worst of all, they don't purchase anything.
- **Cluster 1**– We'll call this group the "M5 Lovers" because they tend to walk straight to the M5s, ignoring the 3-series cars and the Z4. However, they don't have a high purchase rate – only 52 percent. This is a potential problem and could be a focus for improvement for the dealership, perhaps by sending more salespeople to the M5 section.
- **Cluster 2**– This group is so small we can call them the "Throw-Aways" because they aren't statistically relevent, and we can't draw any good conclusions from their behaviour. (This happens sometimes with clusters and may indicate that you should reduce the number of clusters you've created).
- **Cluster 3**– This group we'll call the "BMW Babies" because they always end up purchasing a car and always end up financing it. Here's where the data shows us some interesting things: It appears they walk around the lot looking at cars, then turn to the computer search available at the dealership. Ultimately, they tend to buy M5s or Z4s (but never 3-series). This cluster tells the dealership that it should consider making its search computers more prominent around the lots (outdoor search computers?), and perhaps making the M5 or Z4 much more prominent in the search results. Once the customer has made up his mind to purchase the vehicle, he always qualifies for financing and completes the purchase.
- **Cluster 4**– This group we'll call the "Starting Out With BMW" because they always look at the 3-series and never look at the much more expensive M5. They walk right into the showroom, choosing not to walk around the lot and tend to ignore the computer search terminals. While 50 percent get to the financing stage, only 32 percent ultimately finish the transaction. The dealership could draw the conclusion that these customers looking to buy their first BMWs know exactly what kind of car they want (the 3-series entry-level model) and are hoping to qualify for financing to be able to afford it. The dealership could possibly increase sales to this group by relaxing their financing standards or by reducing the 3-series prices.

References

- MOOC: https://www.youtube.com/user/WekaMOOC
- The Weka workbench: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf
- Weka tutorial by Google: https://www.youtube.com/watch?v=TF1yh5PKaqI
- dataset: https://www.kaggle.com/datasets
- dataset: https://archive.ics.uci.edu/ml/datasets.html

Slide courtesy of Dr. Stefano Pio Zingaro: https://saltgz.github.io