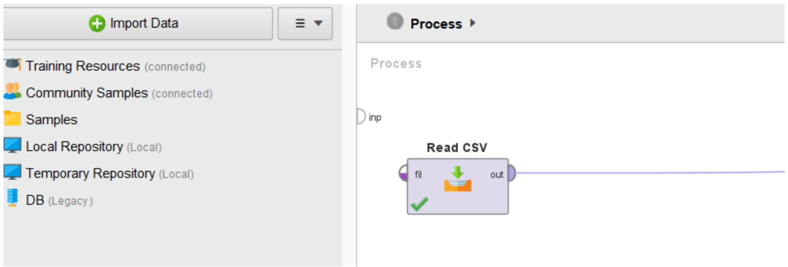
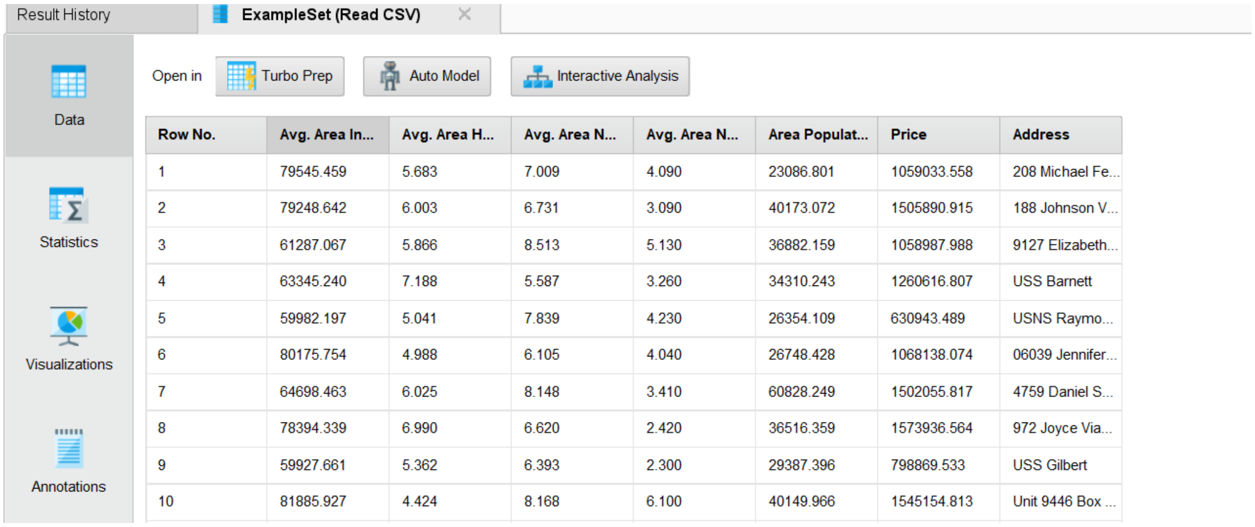
Rapidminer USA-Housing

1. Import *.csv file, Get csv file from local, click **import Configuration Wizard** , connect [readcsv]--res and [Run]


Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)
Process
Process
inp
Read CSV
fil
out

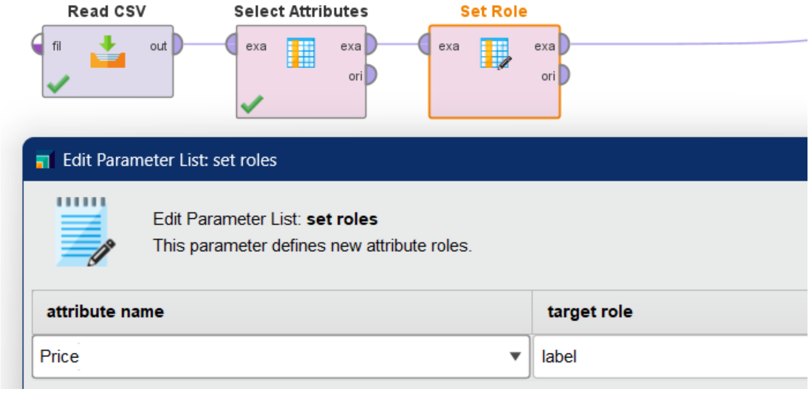


Row No.	Avg. Area In...	Avg. Area H...	Avg. Area N...	Avg. Area N...	Area Populat...	Price	Address
1	79545.459	5.683	7.009	4.090	23086.801	1059033.558	208 Michael Fe...
2	79248.642	6.003	6.731	3.090	40173.072	1505890.915	188 Johnson V...
3	61287.067	5.866	8.513	5.130	36882.159	1058987.988	9127 Elizabeth...
4	63345.240	7.188	5.587	3.260	34310.243	1260616.807	USS Barnett
5	59982.197	5.041	7.839	4.230	26354.109	630943.489	USNS Raymo...
6	80175.754	4.988	6.105	4.040	26748.428	1068138.074	06039 Jennifer...
7	64698.463	6.025	8.148	3.410	60828.249	1502055.817	4759 Daniel S...
8	78394.339	6.990	6.620	2.420	36516.359	1573936.564	972 Joyce Via...
9	59927.661	5.362	6.393	2.300	29387.396	798869.533	USS Gilbert
10	81885.927	4.424	8.168	6.100	40149.966	1545154.813	Unit 9446 Box ...

Step 2. Operators: type and move [Select Attributes] to **design** pane.

Parameters: attribute filter type: **a subset** , click [select Attributes], the dialog window opens, choose every items beside **Address** [Apply]

Step 3. Operators: type, move [Set Role] to **design** pane. We set dependent value: **price => label**



Edit Parameter List: set roles

Edit Parameter List: **set roles**
This parameter defines new attribute roles.

attribute name	target role
Price	label

Step 4. Operators: type and move **[Split Data**] to **design** pane.

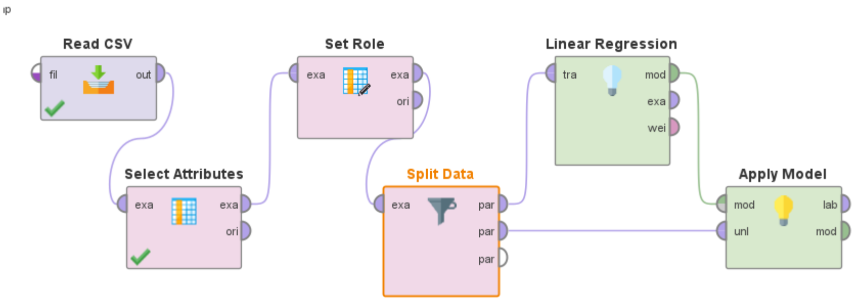
 Parameters: Partition, click Edit Enumertation, click add entry 0.8, then add entry: 0.2 [ok]

Step 5. Operators: type and move **[Linear Regression]** to **design** pane.

  [Split Data] par ----- tra **[Linear Regression]**

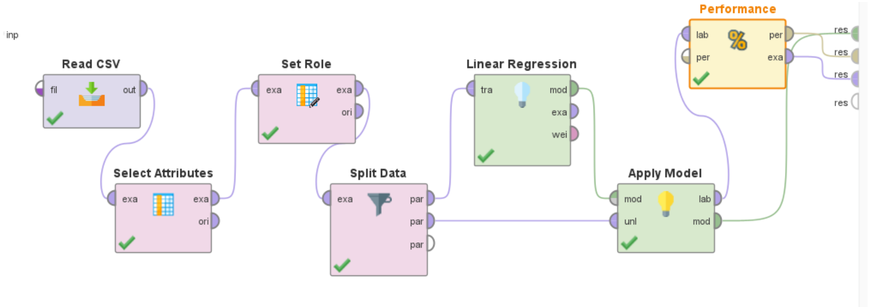
Step 6: Operators: type and move **[Apply Mode**] to **design** pane.

Connection as following

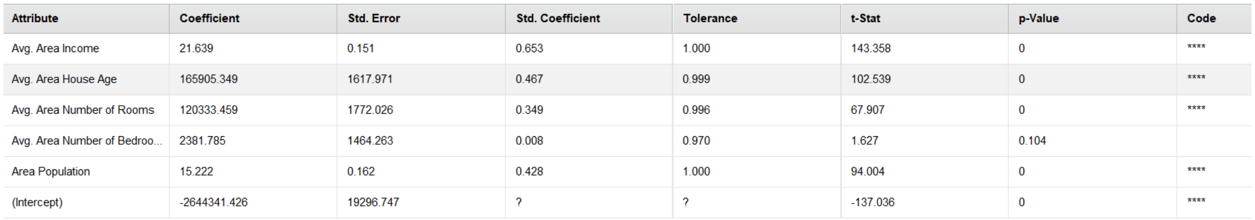


Step 7: Operators: type and move **[Performance (Regression**)] to **design** pane.

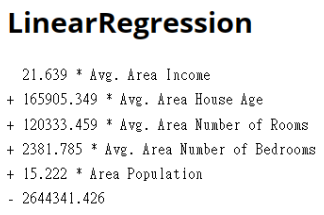
Parameters: select: root mean squared error(default), squared correlation



Data for LinearRegression(Linear Regression)



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Avg. Area Income	21.639	0.151	0.653	1.000	143.358	0	****
Avg. Area House Age	165905.349	1617.971	0.467	0.999	102.539	0	****
Avg. Area Number of Rooms	120333.459	1772.026	0.349	0.996	67.907	0	****
Avg. Area Number of Bedroo...	2381.785	1464.263	0.008	0.970	1.627	0.104	
Area Population	15.222	0.162	0.428	1.000	94.004	0	****
(Intercept)	-2644341.426	19296.747	?	?	-137.036	0	****



LinearRegression

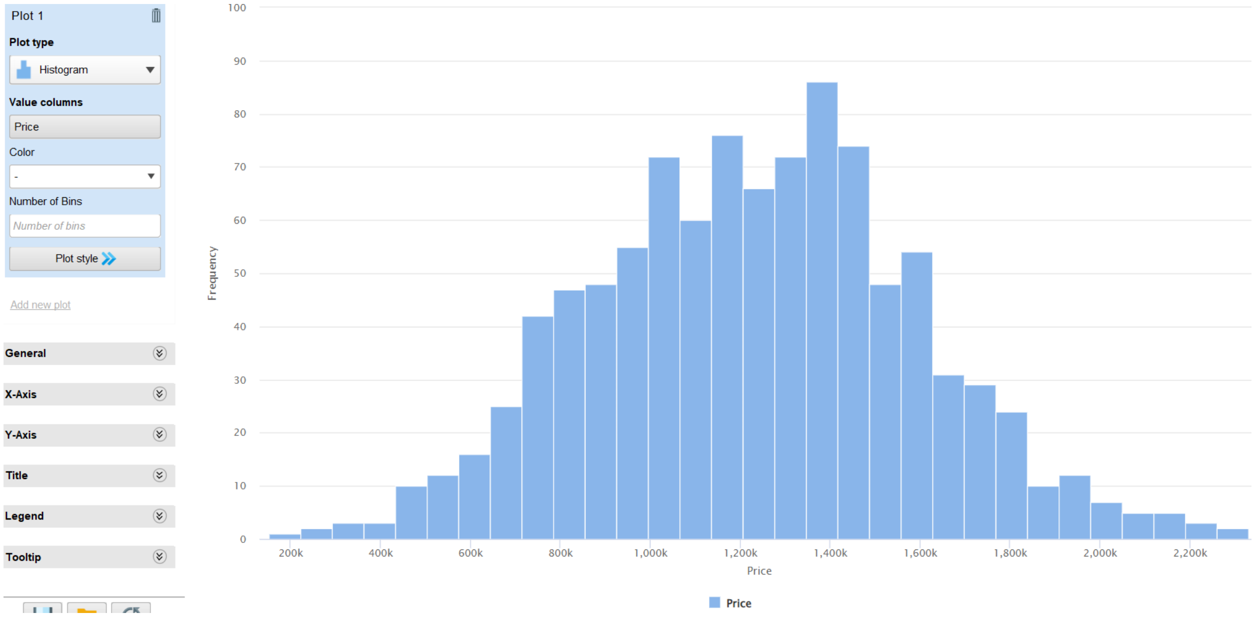
```
  21.639 * Avg. Area Income
+ 165905.349 * Avg. Area House Age
+ 120333.459 * Avg. Area Number of Rooms
+ 2381.785 * Avg. Area Number of Bedrooms
+ 15.222 * Area Population
- 2644341.426
```

PerformanceVector

Root_mean_squared_error : 101010.879 +/- 0.000

Square correlation: 0.921

ExampleSet(Apple Model) ==Visualizations



Plot 1

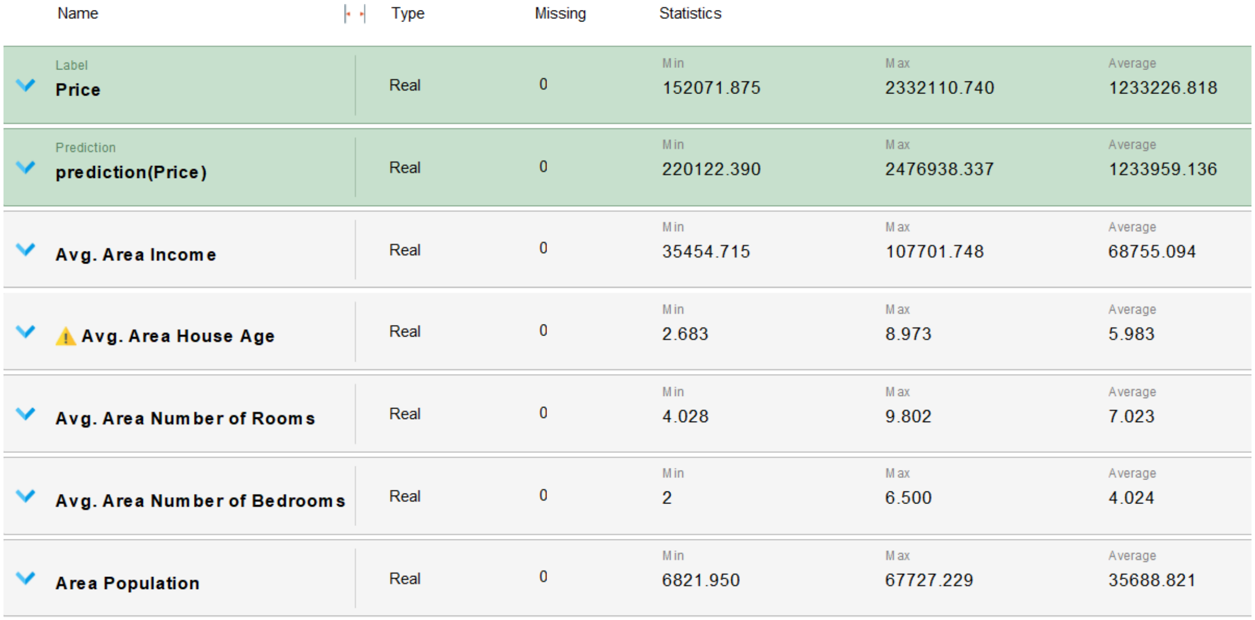
Plot type: Histogram

Value columns: Price

Color: -

Number of Bins: Number of bins

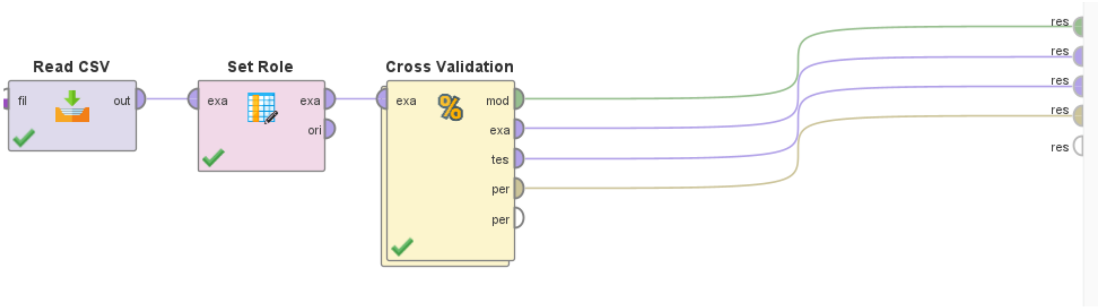
- Data



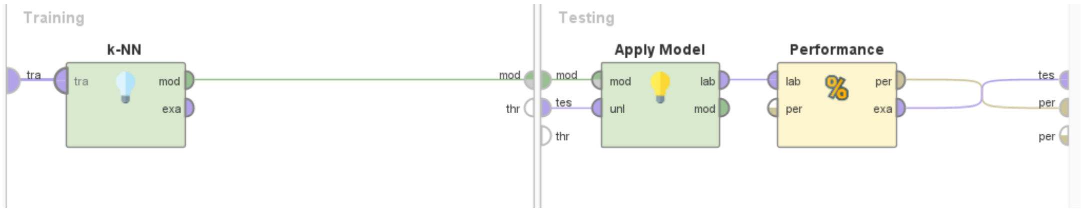
Name	Type	Missing	Min	Max	Average
Label **Price**	Real	0	152071.875	2332110.740	1233226.818
Prediction **prediction(Price)**	Real	0	220122.390	2476938.337	1233959.136
Avg. Area Income	Real	0	35454.715	107701.748	68755.094
⚠ **Avg. Area House Age**	Real	0	2.683	8.973	5.983
Avg. Area Number of Rooms	Real	0	4.028	9.802	7.023
Avg. Area Number of Bedrooms	Real	0	2	6.500	4.024
Area Population	Real	0	6821.950	67727.229	35688.821

Comparing the train data price and the predicated price, the house price will grow, but it also has a lot of conditions to relate to the house price. It is not simple.

Knn with Cross Validation (USA-Housing)



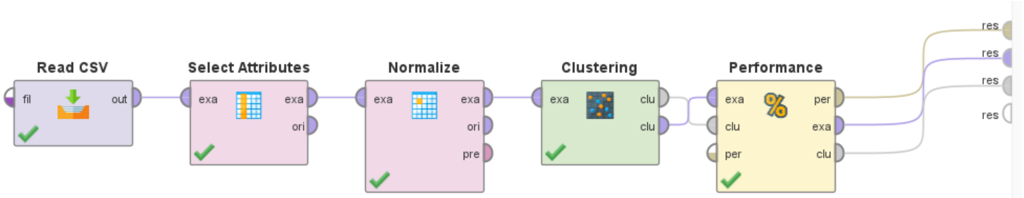
Inside Cross Validation



1. Set role ➔ price, Label

2. Apply do not do anything, Performance ➔ Performance

KMean (USA-housing , This file is not good to use Kmean, just try for fun)



Select Attributes - Select every field, except [Address]
Normalize: The files need to be clustered
Clustering number 3
Performance: Performance(Cluster Distance Performance)