

The screenshot displays the Rapidminer Process Designer interface. At the top, a 'Process' tab is active, showing a workflow diagram. The workflow consists of the following steps:

- Retrieve Boston Housing Data**: The first step, which outputs data to the 'Set Role' step.
- Set Role**: A step that assigns roles to the data, outputting to the 'Select Attributes' step.
- Select Attributes**: A step that selects specific attributes from the data, outputting to the 'Linear Regression' step.
- Linear Regression**: A modeling step, highlighted with a green border, which performs a linear regression analysis. It has a 'mod' output and a 'test' output.
- Apply Model**: A step that applies the trained model to new data, highlighted with a green border. It receives input from the 'Linear Regression' step and outputs results.

The workflow is connected to a 'Retrieve Boston Housing Data' step and an 'Apply Model' step. The 'Linear Regression' step is highlighted with a green border. The 'Apply Model' step is also highlighted with a green border. The workflow is connected to a 'Retrieve Boston Housing Data' step and an 'Apply Model' step.

To leverage the Wisdom of Crowds, you must be a member of the Rapidminer Community!

[Join the community](#)

intercept=-28.811

predictio... ↑	RM	CRIM	ZN	INDUS	CHAS
20.806	6	0.200	0	7	0

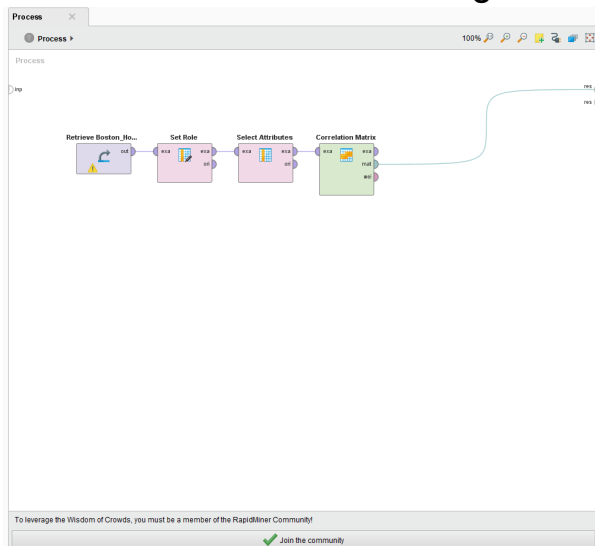
$$y = \text{intercept} + \text{RM} + \text{CRIM} + \text{CHAS}$$

$$y = -28.881 + 8.278(6) - 0.261(0.2) + 3.763(0)$$

$$y = 20.7348$$

### Search for possible multicollinearity

- b) There are several variables that measure levels of industrialization, which are expected to be positively correlated. These include INDUS, NOX (pollution), and TAX. **We expect a positive relationship between NOX (nitric oxides concentration, a pollutant), INDUS (proportion of non-retail business acres per town) and TAX (tax rate), because areas that have a high proportion of non-retail businesses tend to have higher taxes and more pollution.** These 3 predictors are likely to measure the same thing.

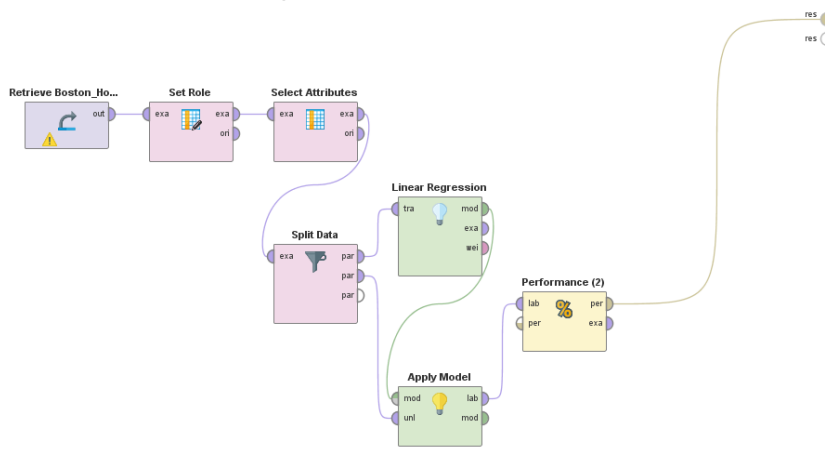


Attributes	INDUS	NOX	TAX
INDUS	1	0.764	0.721
NOX	0.764	1	0.668
TAX	0.721	0.668	1

- NOX and INDUS pair high correlated
- remove INDUS

**Reduce the number of predictors & Propose Your Best Model**

- c) Use a feature selection mechanism to reduce the remaining predictors (from previous step). Run each model separately using the training/testing datasets. Then, give the best model in terms of regression equation. What is RMSE ?



RMSE = 5.236