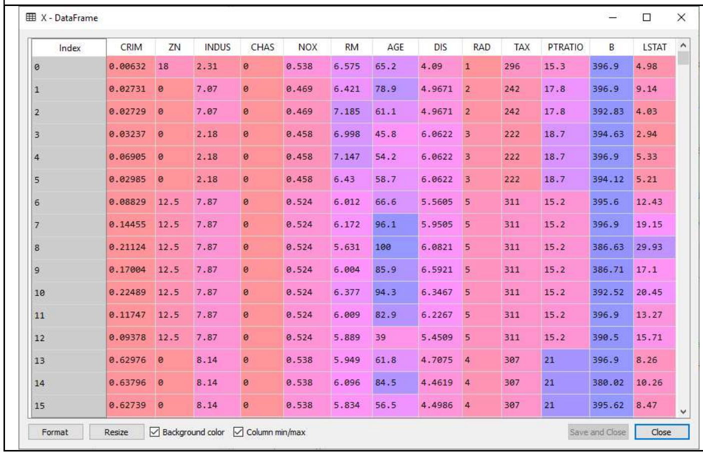
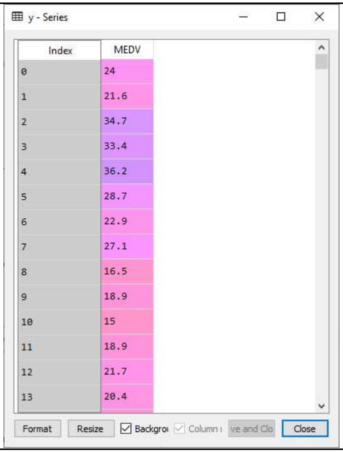
Feature Selection

Here we are performing linear regression, on a dataset with 13 features and 1 target.

Features	Target
	

Index	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98
1	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14
2	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03
3	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94
4	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33
5	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21
6	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43
7	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15
8	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93
9	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1
10	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45
11	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27
12	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71
13	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26
14	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26
15	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47

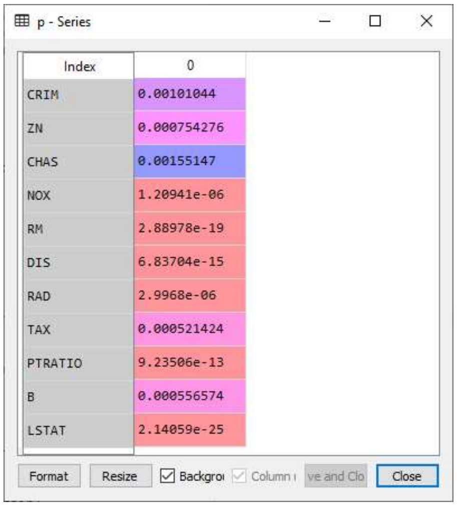
Index	MEDV
0	24
1	21.6
2	34.7
3	33.4
4	36.2
5	28.7
6	22.9
7	27.1
8	16.5
9	18.9
10	15
11	18.9
12	21.7
13	20.4

To perfume linear regression we present 4 robust methods for feature selection:

- ***Backward Elimination:***

This method is based on feeding all the possible features to the model at first. Then, based on the performance of the model features the worst performing features are removed. The criteria here is the pvalue. Features with pvalue > 0.05 are removed.

- pvalues for all features:



Index	0
CRIM	0.00101044
ZN	0.000754276
CHAS	0.00155147
NOX	1.20941e-06
RM	2.88978e-19
DIS	6.83704e-15
RAD	2.9968e-06
TAX	0.000521424
PTRATIO	9.23506e-13
B	0.000556574
LSTAT	2.14059e-25

For this particular dataset 2 features have pvalue above 0.05

Features	**pvalue**
AGE	0.958229309205725
INDUS	0.7379887092915007

Based on the Backward Elimination method the selected features are: **['CRIM', 'ZN', 'CHAS', 'NOX', 'RM', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT']**

- ***Recursive Feature Elimination:***

This is a recursive method that remove features given a desired number of features. The method takes is the model 'Linear Regression' and the desired number of feature. For this case we try to optimize the number of features based on the score of the model given a number features.

For this particular case the Optimum number of features is 10

Score with 10 features is 0.663581

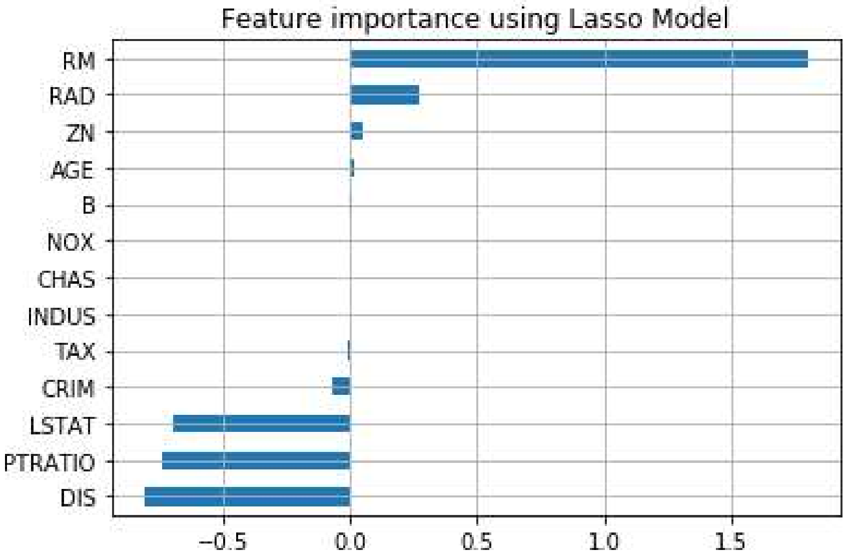
Based on the Recursive Feature Elimination method the selected features are: **['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'DIS', 'RAD', 'PTRATIO', 'LSTAT']**

- ***Embedded Method (Lasso) without threshold:***

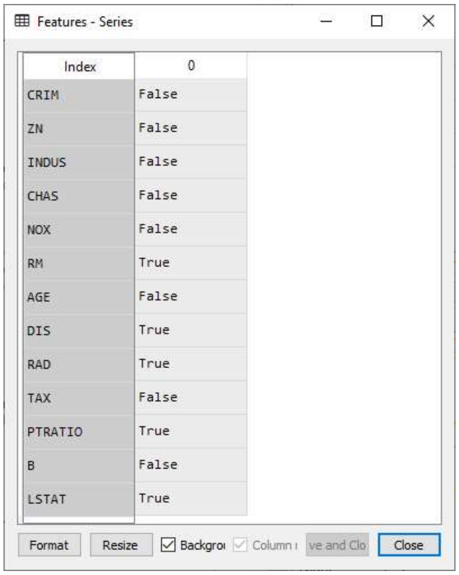
This a regularization method that aims at extracting the features which contributes the most to the training. Here we use Lasso Regularization. Features with Lasso coefficient = 0 are removed form the model.

Based on the Embedded method (Lasso) (without threshold) the selected features are: **['CRIM', 'ZN', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT']**

Lasso model picked 10 features and removed 3 features



- ***<u>Embedded Method (Lasso) with threshold:</u>***



Index	0
CRIM	False
ZN	False
INDUS	False
CHAS	False
NOX	False
RM	True
AGE	False
DIS	True
RAD	True
TAX	False
PTRATIO	True
B	False
LSTAT	True

The model selected 5 important features and removed 8 less important features

Based on the Embedded method (Lasso) (with threshold of 0.2) the selected features are: **['RM', 'DIS', 'RAD', 'PTRATIO', 'LSTAT']**