main

November 4, 2024

0.1 1) Formulate question

House prices are determined by a multitude of variables, but which one(s) have the biggest effect?

Here, we have data that can give us a clue as to what variables influence house prices, in addition to creating to a model for predicting house prices for those in other locations (e.g. New York, St. Louis, Atlanta).

0.2 2) Gather data

0.2.1 Background info:

The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978. Used in Belsley, Kuh & Welsch, 'Regression diagnostics …', Wiley, 1980. N.B. Various transformations are used in the table on pages 244-261 of the latter.

Variables (in order):

- **CRIM**: Per capita crime rate by town
- **ZN**: Proportion of residential land zoned for lots over 25,000 sq.ft.
- **INDUS**: Proportion of non-retail business acres per town
- **CHAS**: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- **NOX**: Nitric oxides concentration (parts per 10 million)
- **RM**: Average number of rooms per dwelling
- **AGE**: Proportion of owner-occupied units built prior to 1940
- **DIS**: Weighted distances to five Boston employment centres
- **RAD**: Index of accessibility to radial highways
- **TAX**: Full-value property-tax rate per $10,000
- **PTRATIO**: Pupil-teacher ratio by town
- **B**: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- **LSTAT**: % lower status of the population
- **MEDV**: Median value of owner-occupied homes in $1000's (target variable)

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	\
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	

	PTRATIO	B	LSTAT	MEDV
0	15.3	396.90	4.98	24.0
1	17.8	396.90	9.14	21.6
2	17.8	392.83	4.03	34.7
3	18.7	394.63	2.94	33.4
4	18.7	396.90	5.33	36.2

0.3 3) Clean/preprocess data

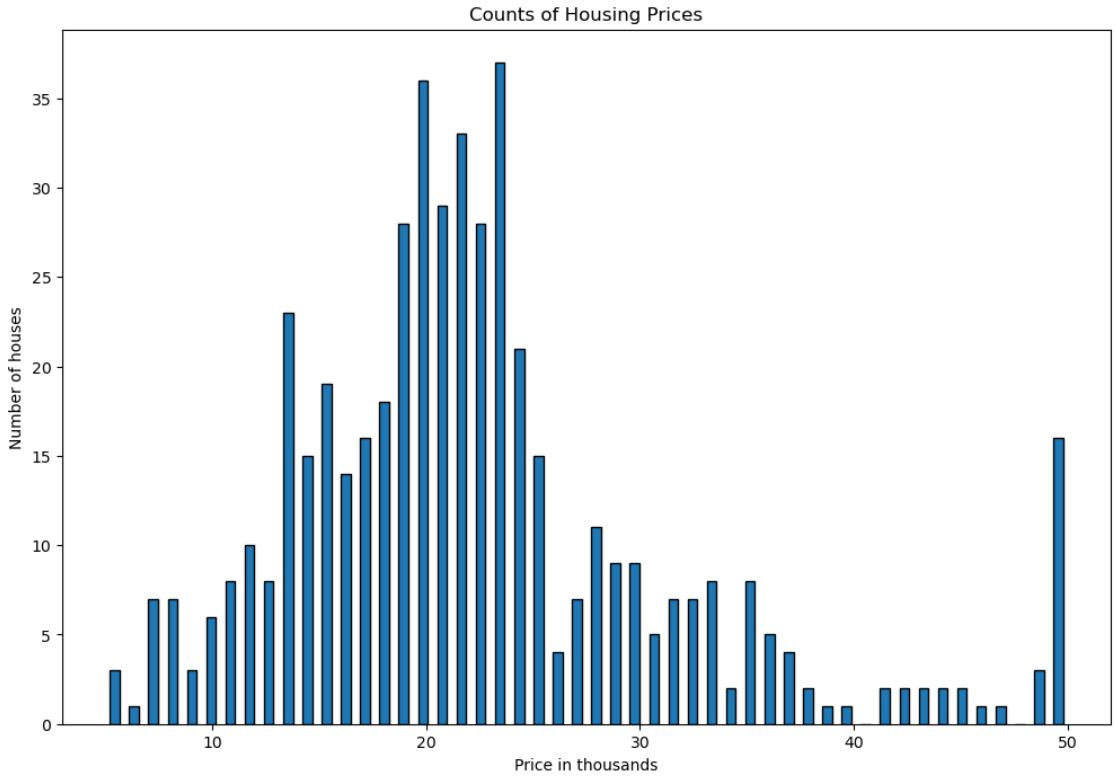
(506, 14)

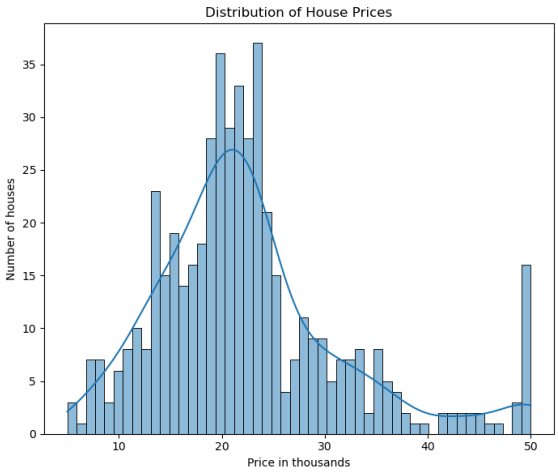
	CRIM	ZN	INDUS	CHAS	NOX	RM	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	

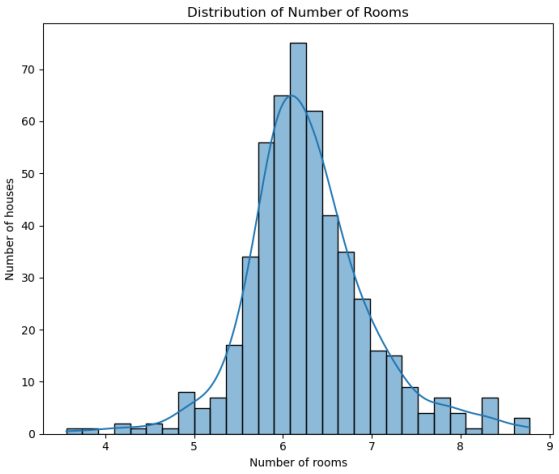
	AGE	DIS	RAD	TAX	PTRATIO	B	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	
std	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	
min	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	
25%	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	
50%	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	
75%	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	
max	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	

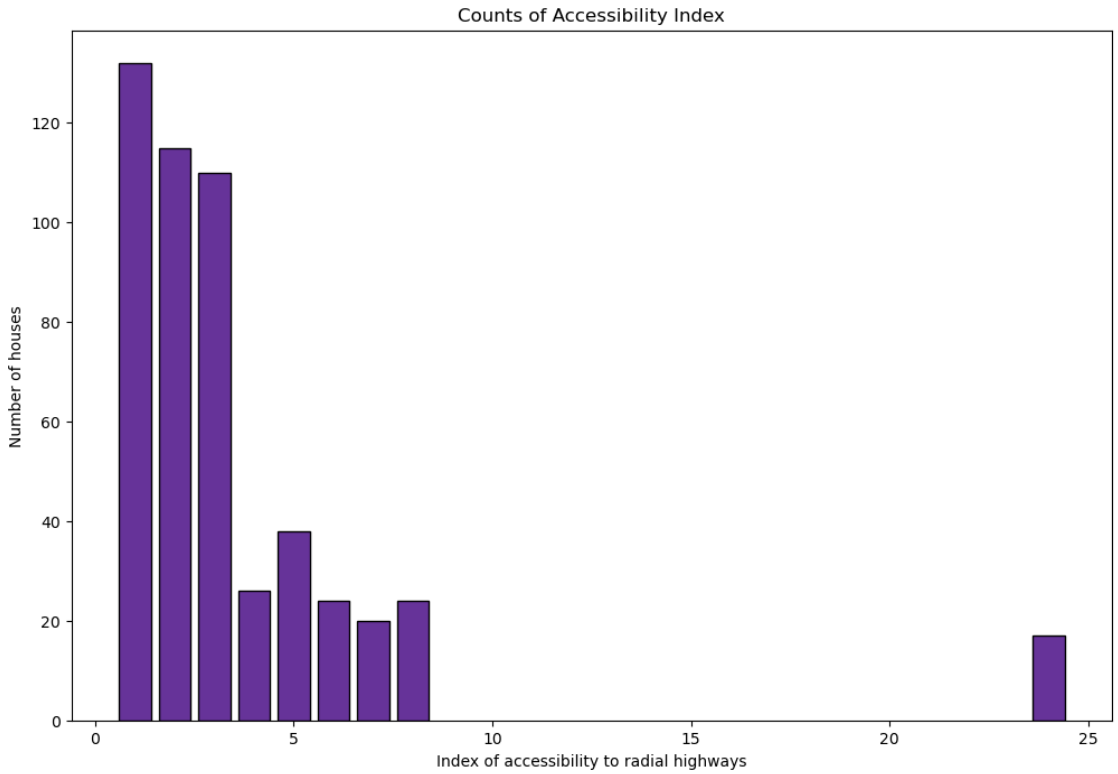
	LSTAT	MEDV
count	506.000000	506.000000
mean	12.653063	22.532806
std	7.141062	9.197104
min	1.730000	5.000000
25%	6.950000	17.025000
50%	11.360000	21.200000
75%	16.955000	25.000000
max	37.970000	50.000000

1 4) Visualize data









1.1 Descriptive Statistics

	CRIM	ZN	INDUS	CHAS	NOX	RM	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	

	AGE	DIS	RAD	TAX	PTRATIO	B	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	
std	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	
min	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	
25%	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	
50%	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	
75%	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	
max	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	

	LSTAT	MEDV
count	506.000000	506.000000
mean	12.653063	22.532806
std	7.141062	9.197104
min	1.730000	5.000000
25%	6.950000	17.025000
50%	11.360000	21.200000
75%	16.955000	25.000000
max	37.970000	50.000000

1.2 Inferential Statistics

1.2.1 Correlation

$$\rho_{XY} = Corr(XY)$$

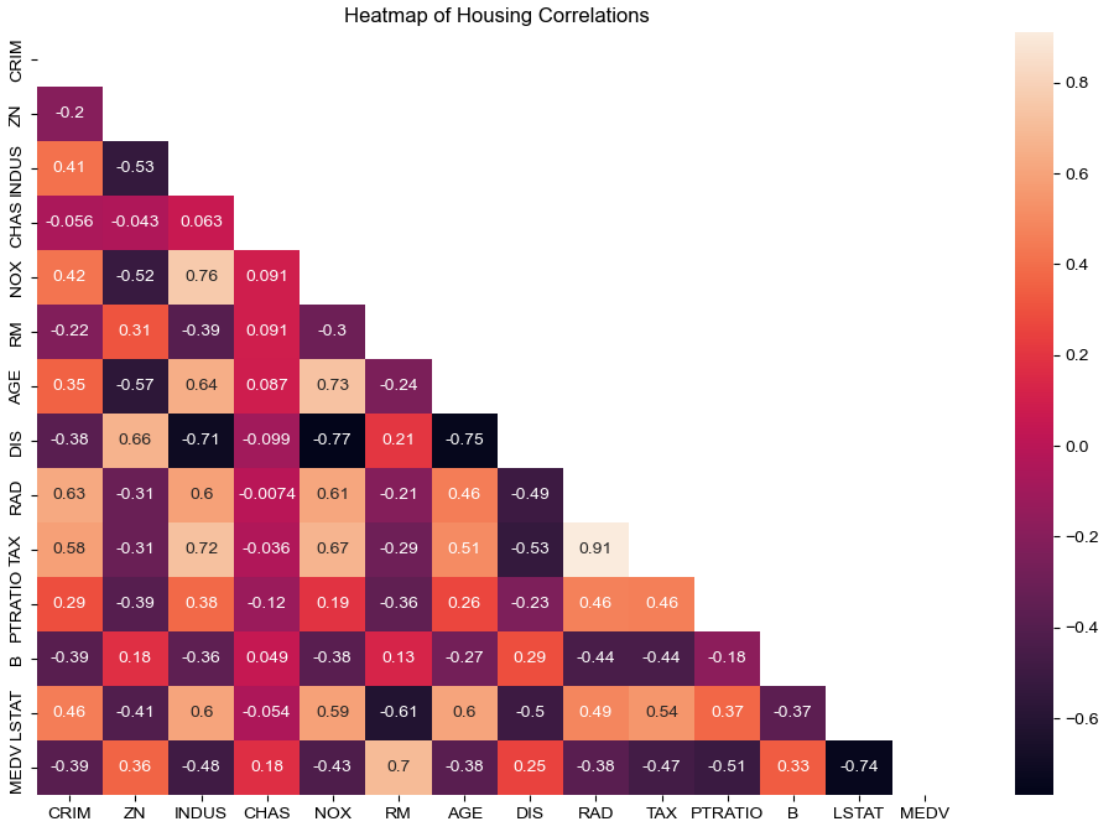
$$-1.0 \leq \rho_{XY} \leq +1.0$$

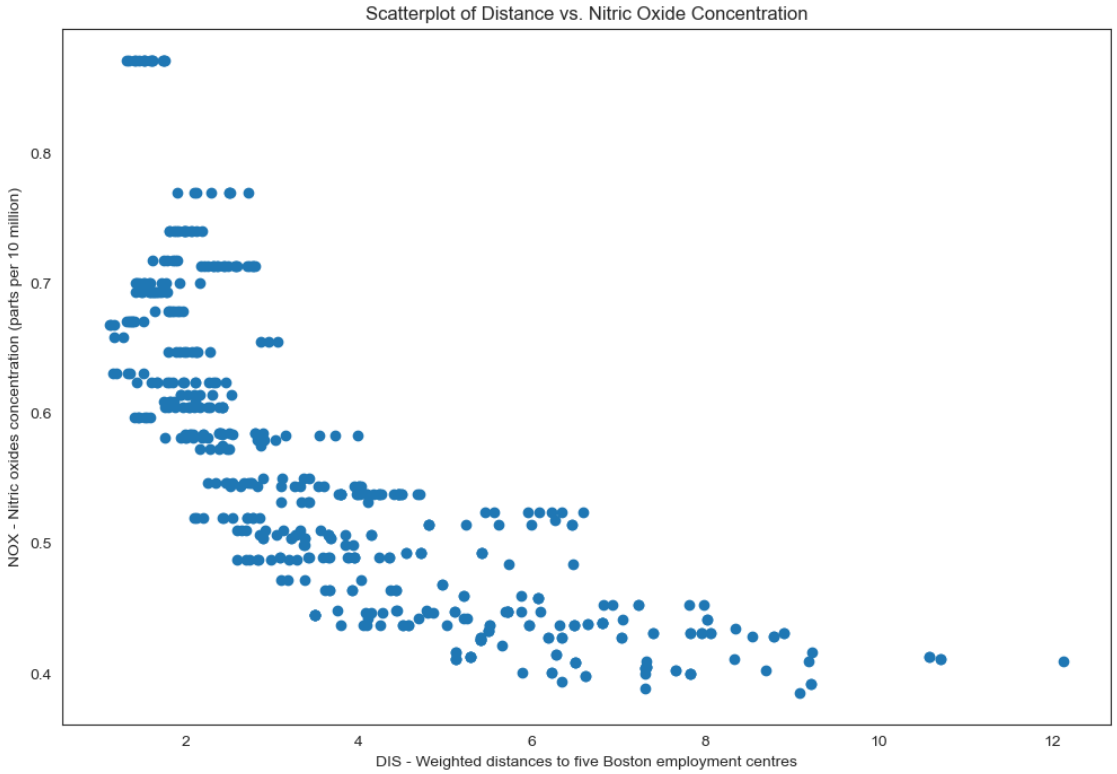
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	\
CRIM	1.000000	-0.200469	0.406583	-0.055892	0.420972	-0.219247	0.352734	
ZN	-0.200469	1.000000	-0.533828	-0.042697	-0.516604	0.311991	-0.569537	
INDUS	0.406583	-0.533828	1.000000	0.062938	0.763651	-0.391676	0.644779	
CHAS	-0.055892	-0.042697	0.062938	1.000000	0.091203	0.091251	0.086518	
NOX	0.420972	-0.516604	0.763651	0.091203	1.000000	-0.302188	0.731470	
RM	-0.219247	0.311991	-0.391676	0.091251	-0.302188	1.000000	-0.240265	
AGE	0.352734	-0.569537	0.644779	0.086518	0.731470	-0.240265	1.000000	
DIS	-0.379670	0.664408	-0.708027	-0.099176	-0.769230	0.205246	-0.747881	
RAD	0.625505	-0.311948	0.595129	-0.007368	0.611441	-0.209847	0.456022	
TAX	0.582764	-0.314563	0.720760	-0.035587	0.668023	-0.292048	0.506456	
PTRATIO	0.289946	-0.391679	0.383248	-0.121515	0.188933	-0.355501	0.261515	
B	-0.385064	0.175520	-0.356977	0.048788	-0.380051	0.128069	-0.273534	
LSTAT	0.455621	-0.412995	0.603800	-0.053929	0.590879	-0.613808	0.602339	
MEDV	-0.388305	0.360445	-0.483725	0.175260	-0.427321	0.695360	-0.376955	

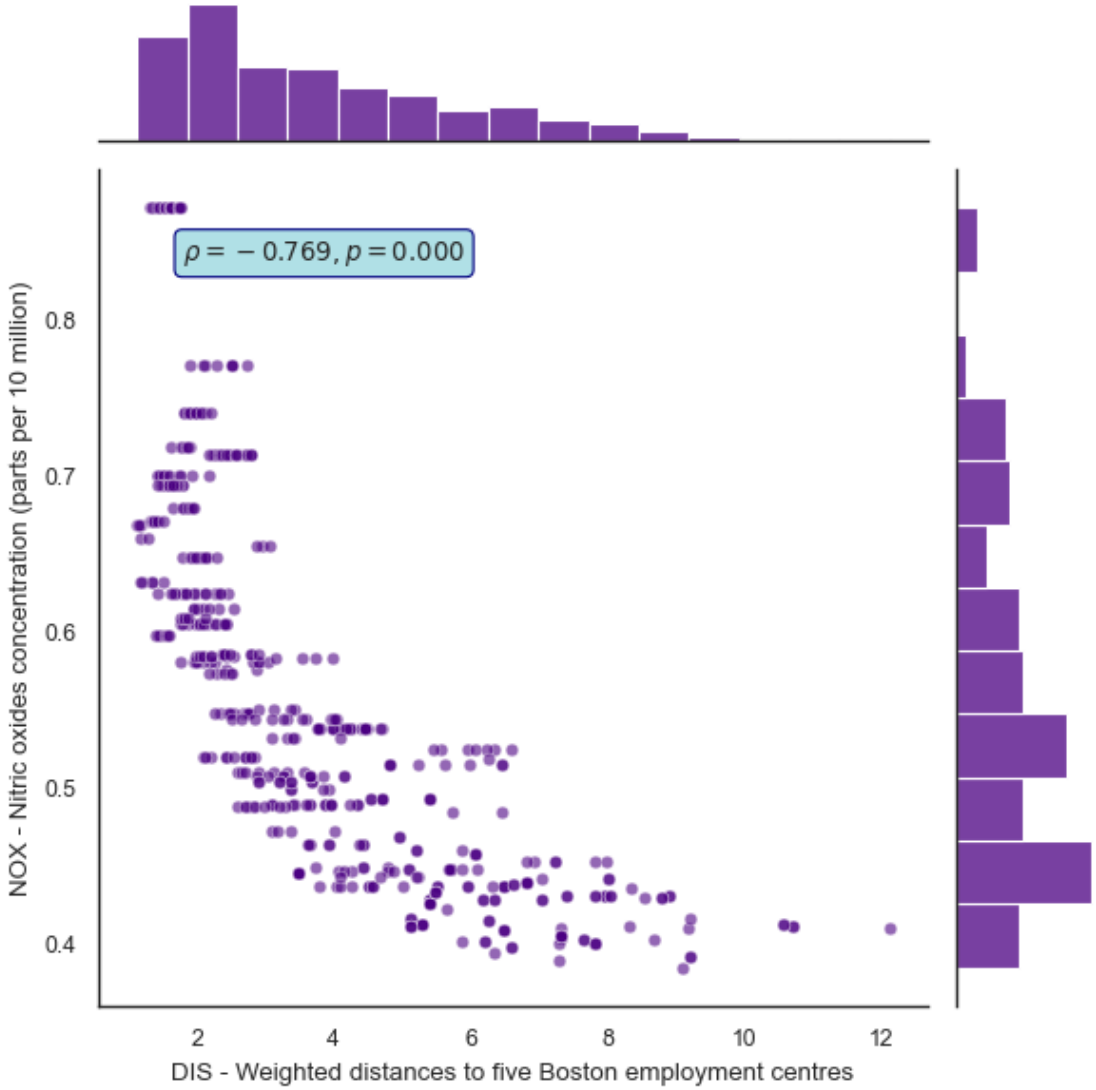
	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	-0.379670	0.625505	0.582764	0.289946	-0.385064	0.455621	-0.388305
ZN	0.664408	-0.311948	-0.314563	-0.391679	0.175520	-0.412995	0.360445
INDUS	-0.708027	0.595129	0.720760	0.383248	-0.356977	0.603800	-0.483725
CHAS	-0.099176	-0.007368	-0.035587	-0.121515	0.048788	-0.053929	0.175260
NOX	-0.769230	0.611441	0.668023	0.188933	-0.380051	0.590879	-0.427321
RM	0.205246	-0.209847	-0.292048	-0.355501	0.128069	-0.613808	0.695360
AGE	-0.747881	0.456022	0.506456	0.261515	-0.273534	0.602339	-0.376955
DIS	1.000000	-0.494588	-0.534432	-0.232471	0.291512	-0.496996	0.249929
RAD	-0.494588	1.000000	0.910228	0.464741	-0.444413	0.488676	-0.381626

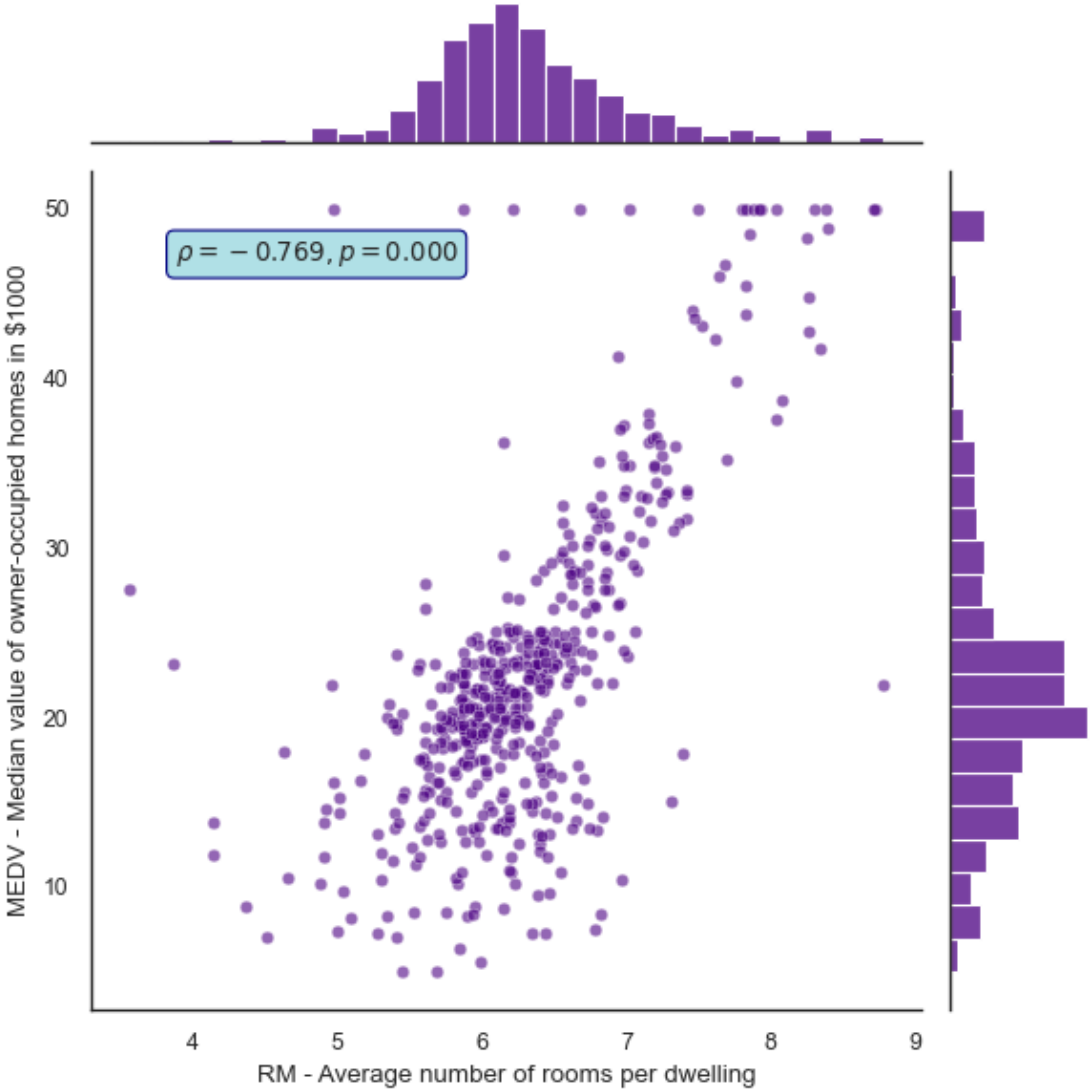
```
TAX     -0.534432  0.910228  1.000000  0.460853 -0.441808  0.543993 -0.468536
PTRATIO -0.232471  0.464741  0.460853  1.000000 -0.177383  0.374044 -0.507787
B        0.291512 -0.444413 -0.441808 -0.177383  1.000000 -0.366087  0.333461
LSTAT   -0.496996  0.488676  0.543993  0.374044 -0.366087  1.000000 -0.737663
MEDV     0.249929 -0.381626 -0.468536 -0.507787  0.333461 -0.737663  1.000000
```

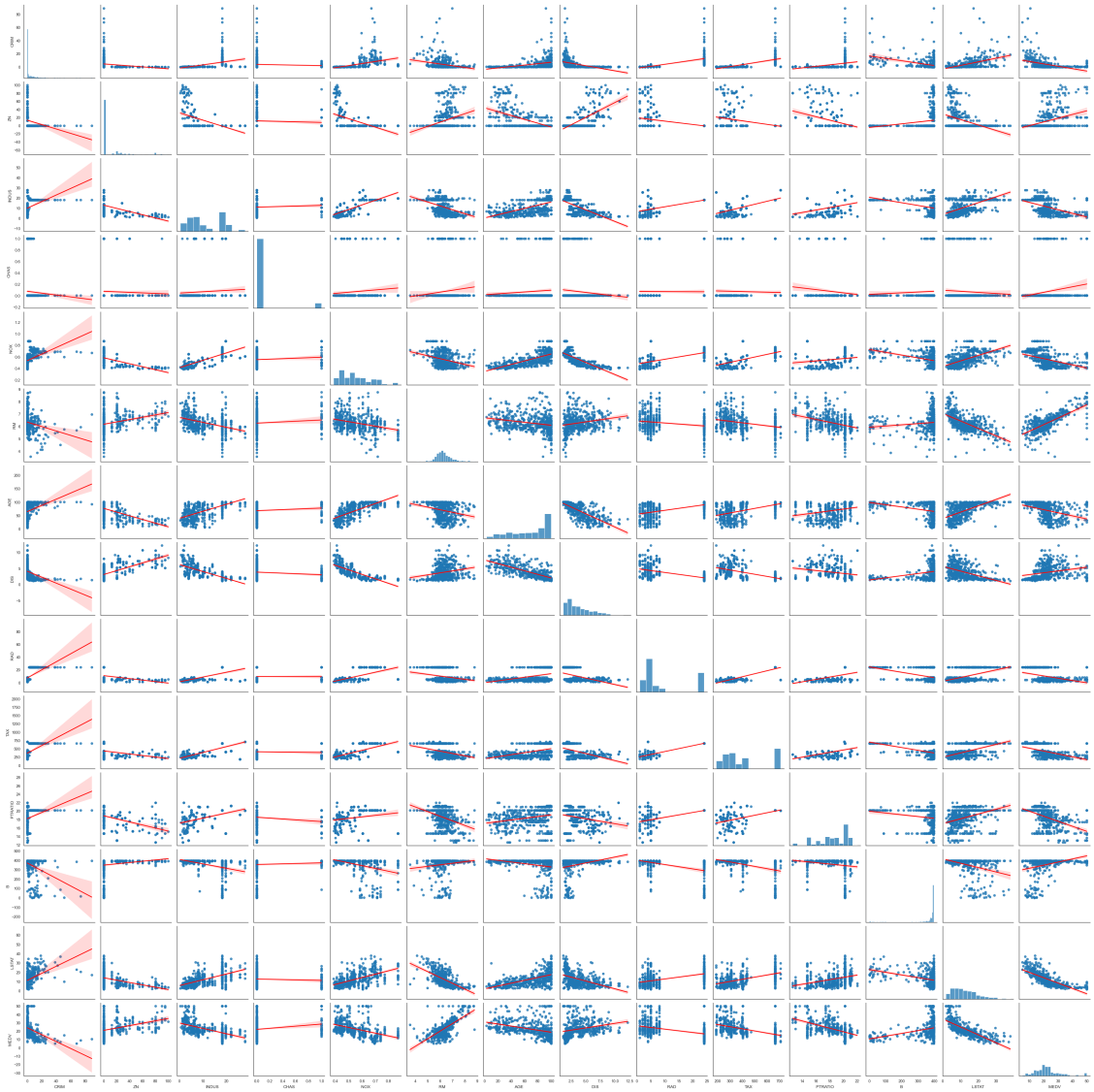
```
array([[1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 1., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 1., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 1., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 1., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 1.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1.]])
```




Scatterplot of Distance vs. Nitric Oxide Concentration
NOX - Nitric oxides concentration (parts per 10 million)
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
12
DIS - Weighted distances to five Boston employment centres


$\rho = -0.769, p = 0.000$
NOX - Nitric oxides concentration (parts per 10 million)
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
12
DIS - Weighted distances to five Boston employment centres


$\rho = -0.769, p = 0.000$
MEDV - Median value of owner-occupied homes in $1000
RM - Average number of rooms per dwelling
50
40
30
20
10
4
5
6
7
8
9



```
CPU times: user 52.7 s, sys: 721 ms, total: 53.4 s
Wall time: 56.8 s
```

1.3 5) Train & build algorithm: Multiple Regression

1.3.1 Training & testing sets

1.3.2 Training Results

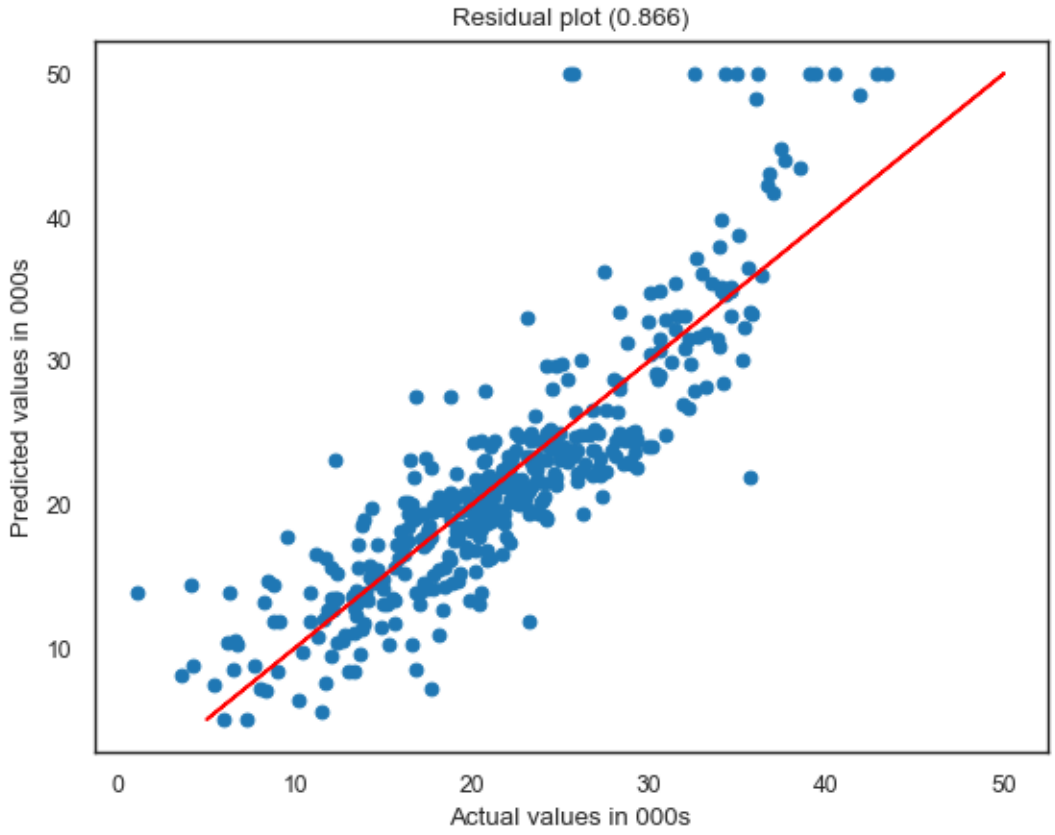
```
R-squared:  0.750121534530608
         Coefficients
CRIM         -0.128181
ZN            0.063198
INDUS        -0.007576
CHAS          1.974515
```

```
NOX        -16.271989
RM           3.108456
AGE          0.016292
DIS         -1.483014
RAD          0.303988
TAX         -0.012082
PTRATIO     -0.820306
B            0.011419
LSTAT       -0.581626
Intercept:  36.53305138282434
```

```
[21.02958601 12.21844467 13.74785342 20.7351517  23.41262356 13.91896524
 28.93270221 15.93275264 15.22218031 22.25484624 26.38641515 29.18733455
 24.23140501 18.09942968 16.40587835 17.4186263  15.65122947 21.28912789
 34.31299511 29.88616576 20.90278802 13.75023401 16.19813658 29.16630099
 13.32233024 22.40099547 24.18543936 31.58822487 33.13516905  6.5126739
 35.3907224  24.21142064 17.27480742 24.18719251 28.20666734 34.62364626
  6.55528144  4.26819403 28.27658413 12.64000902 18.06198741 20.13023426
  6.15645739 14.150693   36.70774402 25.83392494 23.2145325  26.285604
 12.76642228 20.19773103 34.98291717 20.44191405 11.6176045  16.42015377
 24.42945242 10.40597467 14.67714741 25.7886378  11.26165698 12.11852939
 19.16397498 19.3141205  32.26664779 22.71058476 25.62230749  8.50310476
 21.00910913  6.73731392 27.8998401  20.8354717  23.94878492 25.67341126
 27.1039239  14.96205256 11.98747357 23.08232446 20.20617473 18.47488856
 22.79404142 20.68474847 22.67708381 19.28454019  8.03109049 32.76450859
 21.70516314 17.43825991 23.45620372 23.59614125 28.88717675 22.33159132
 30.60589895 34.55494214 20.76321628 31.88444619 16.61196204 23.68580044
 21.98266123 32.00201434 28.00999611 33.88393763 25.93358549 22.12832754
 20.79474916 23.33606368 33.54486859 28.88319431 30.53450432 17.74512067
 21.12450337 30.93400461 30.03614104 35.58242779 24.02203797 17.41690607
 23.19785465 30.0323026  21.61242878 22.40964155 13.94600108 15.66214472
 13.79029063 13.42167315 18.83088089 26.7708844  13.19122389 14.192501
 33.2175314  17.72004308 24.63203065 36.12003775 16.83521193 23.9000005
 19.61054474 20.50795972 40.43756108 19.08059797 23.5558946  22.08334885
  6.22827016 18.44728152 17.62379748 21.04549499 23.7227903  27.51457375
 22.57721739 14.13506525 21.14006129 30.32051076 22.05794048 19.11203863
 21.78245049 23.17316973 20.46521844 17.74423582 17.75048976 21.66429287
 18.40325907 14.38435756 20.57128808 36.63940352 10.19041295 34.06472665
 34.3195163  20.40239203 25.37374616  9.49044096 12.3299658  17.01868025
 15.56477034 42.88189691  5.46249378 18.3714866  23.42040409 15.81363683
 17.08144397 18.09935134 16.82557692 11.43328366  8.34173792 24.7168947
 27.22056849 11.74609003  5.99313665 24.04158889 22.31331274 18.78749453
 30.14551898 21.01013218 31.47459335 16.46570057 12.24740394 17.8262914
 25.3889327  32.92917504 22.12393049 15.31210338 13.38060836 34.91151246
 21.84693348 19.84635301 21.42436658 39.02990277 29.23304426 14.28190644
  8.94474091 20.22604989 14.63275313 36.27993935 19.64984827  1.08767142
 18.66415262 35.82218783 34.09308555 18.76537952 35.19869636  8.27271481
 13.54719007 33.98102401 23.18633639 20.87766448 31.98890322 20.00439746
```

```
 24.12500577 13.89454264 26.0992943  16.57055932 20.67688049 23.26754119
 17.27974991 18.9538199  11.83457117 22.25843441 20.66810392 24.53080672
 16.28431195 12.02178999 12.319154   25.86776284 16.1331981  30.5438815
 37.36066103 24.13473736 18.30792779 39.35764324 20.5206877  16.7709602
  8.81387707 15.51358248 15.67443213 13.31856854 26.10316649 28.33403738
 20.95491778 23.02781695 24.74978082 16.17189654 24.72074601 17.130221
 17.35384925 22.42287729 25.72976869 14.84788476  8.78237945 16.62259055
 20.40110222  7.30306583 27.08071196 26.78112732 36.03947745 19.68774042
 19.7901588  13.83221844 16.70130292 23.68253606 31.17748679 34.01072305
 16.8202101  17.69880495 24.10972302 19.99635376 17.05885432 22.23195591
 21.84584444 28.62317427 27.2516681  28.2549724  14.43864505 19.21053553
 24.88791041 29.97555608 21.82499918 12.30210693 26.13817796 25.12144059
 14.9799101  32.21007934 24.38395172 36.99893418 13.00240533 22.09355257
 24.70002878 19.20357309 32.24421081 38.44437756 21.02915426 21.62466551
 20.86358618 30.60879061 27.56113489 20.03161867 13.58585317 11.97713884
 18.45564721 26.92662003 23.29650531 34.15401892 33.97583104 35.71066748
 35.67322524 18.75783227 13.63453125  7.72417361 27.48611925 16.40223191
 22.51313663 17.51087673 23.82126731 26.52379347 31.47467634 41.82307398
 17.49171313 17.26751731 10.88394911 10.88385116 26.93451054 21.02281041
 20.23959401 18.7305654  23.02417741 20.50210243 25.40644204 24.54465458
 21.47837821 20.97271167 21.24927726 27.35298617 20.71214127  9.09339883
 20.17694343 28.91367004 16.32102777 43.39813352 32.5656775  25.9060061
 18.85339203 24.91329906 13.42728208 30.46417809 29.14062649 20.55075919
 21.67906512 28.095632    4.16467731 13.80650636 18.80826148 26.81898184
 20.16240575 21.93250508 22.89661432 28.26532198 32.59880179 29.28804669
 19.94830572 21.9477766  30.93090651 14.92630564 24.60451111 20.37848006
 20.15868929 28.69894714 16.65338476 20.3648217  28.13568886 19.19445466
 20.70719308 25.47953329 19.67485932 34.58480536  3.62264799 37.63427821
 23.08044691 11.12352083 20.78751336 11.71352368 22.76714085 25.0269165
 13.45932025 15.88345795 32.51022857 24.84599397 19.89193201 19.16746008
 22.40634226 28.57061195]
50     19.7
367    23.1
34     13.5
78     21.2
172    23.1
       …
320    23.8
15     19.9
484    20.6
125    21.4
265    22.8
Name: MEDV, Length: 404, dtype: float64
```

1.3.3 Regression: Actual values vs. predicted



1.3.4 Assessing the model (e.g. coefficient p-Values)

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.742
Method:	Least Squares	F-statistic:	90.06
Date:	Mon, 04 Nov 2024	Prob (F-statistic):	1.12e-108
Time:	11:19:50	Log-Likelihood:	-1170.5
No. Observations:	404	AIC:	2369.
Df Residuals:	390	BIC:	2425.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	36.5331	5.428	6.730	0.000	25.861	47.205
CRIM	-0.1282	0.032	-4.005	0.000	-0.191	-0.065
ZN	0.0632	0.014	4.420	0.000	0.035	0.091

```
INDUS         -0.0076      0.063     -0.119      0.905      -0.132       0.117
CHAS           1.9745      0.924      2.138      0.033       0.159       3.790
NOX          -16.2720      3.965     -4.104      0.000     -24.067      -8.477
RM             3.1085      0.449      6.926      0.000       2.226       3.991
AGE            0.0163      0.015      1.123      0.262      -0.012       0.045
DIS           -1.4830      0.214     -6.920      0.000      -1.904      -1.062
RAD            0.3040      0.067      4.514      0.000       0.172       0.436
TAX           -0.0121      0.004     -3.208      0.001      -0.019      -0.005
PTRATIO       -0.8203      0.141     -5.826      0.000      -1.097      -0.543
B              0.0114      0.003      4.239      0.000       0.006       0.017
LSTAT         -0.5816      0.053    -11.016      0.000      -0.685      -0.478
==============================================================================
Omnibus:                      141.305   Durbin-Watson:                   2.125
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              651.065
Skew:                           1.454   Prob(JB):                    4.20e-142
Kurtosis:                       8.497   Cond. No.                     1.54e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.54e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
50     -1.329586
367    10.881555
34     -0.247853
78      0.464848
172    -0.312624
         …
320    -1.045994
15      0.008068
484     1.432540
125    -1.006342
265    -5.770612
Length: 404, dtype: float64
19.921197403247984
19.230858879373056
```

1.3.5 A. Checking that there's a linear relationship

1.3.6 B. Showing that the residuals are independent using the Durbin-Watson statistic

```
2.124548455902406
```

Because the Durbin-Watson test returns a value between 1.5 and 2.5, autocorrelation is likely not a cause for concern.

Therefore, the residuals are independent.

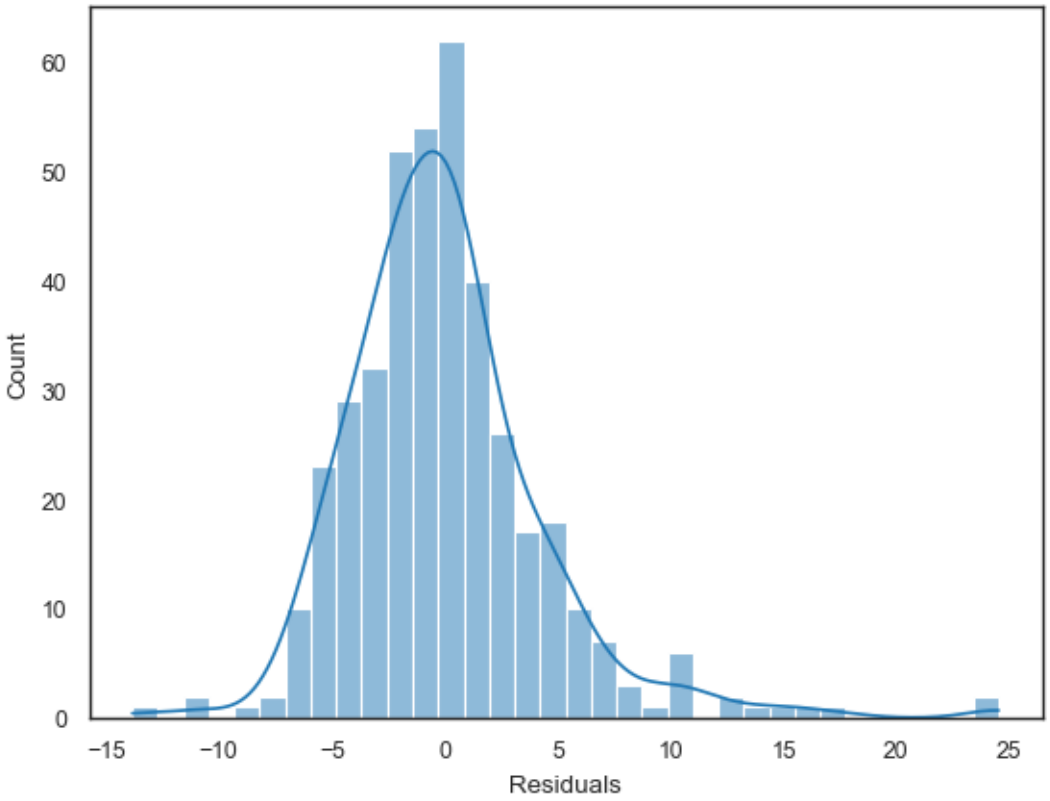
1.3.7 C. Showing that the residuals display homoscedasticity (i.e. constant variance) using the Breusch-Pagan Test

```
Lagrange multiplier statistic    6.254941e+01
p-value                          1.830725e-08
f-value                          5.495618e+00
f p-value                        3.498179e-09
dtype: float64
```

Since the p-value is less than our significance less (0.05), our regression is homoscedastic.

Source: https://www.statology.org/breusch-pagan-test/

1.3.8 D. Showing that the residuals display normality



1.3.9 E. Checking for multicollinearity using the variance influence factor (VIF)

```
CRIM       597.55
ZN           1.71
INDUS        2.33
CHAS         3.94
NOX          1.08
RM           4.41
AGE          1.84
DIS          3.33
RAD          4.22
TAX          7.31
PTRATIO      8.51
B            1.84
LSTAT        1.34
dtype: float64
```

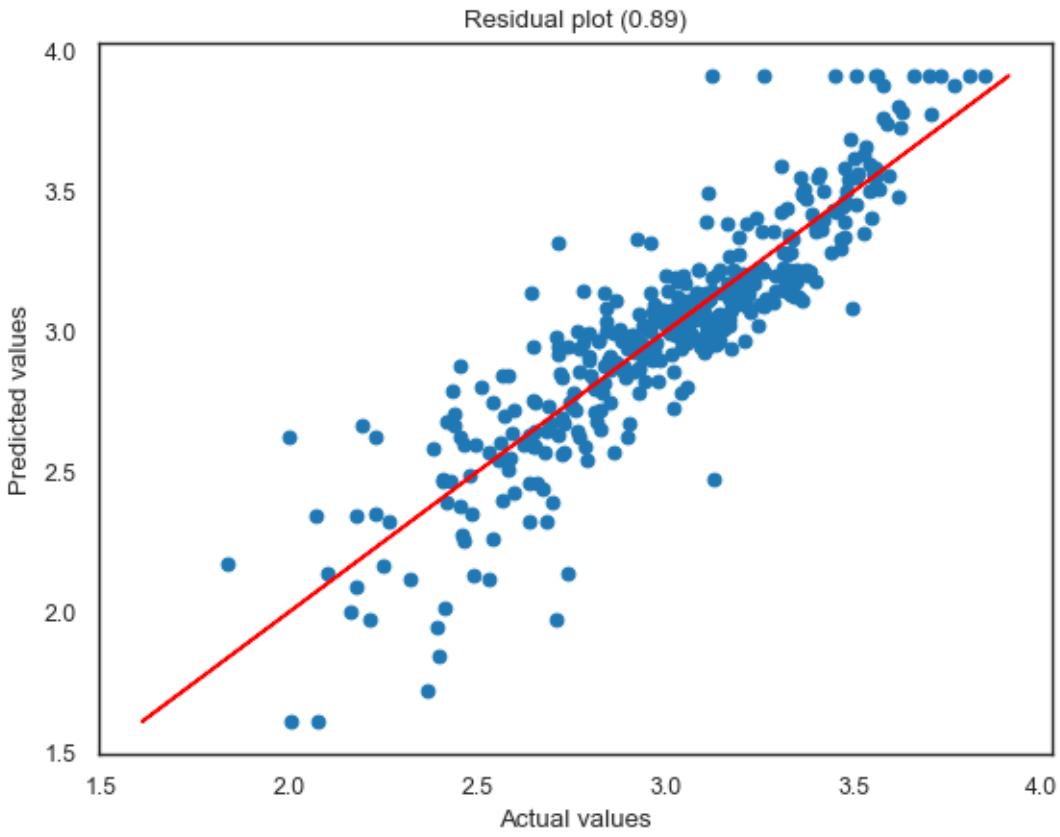
The cutoff for multicollinearity is 10, and none of the features meet that cutoff.

Therefore, we can say that there is no multicollinearity.

1.4 Transforming Regression with Log Prices

```
R-squared: 0.7918657661852815
         Coefficients
CRIM        -0.010702
ZN           0.001461
CHAS         0.086449
NOX         -0.616448
RM           0.076133
DIS         -0.052692
RAD          0.013743
TAX         -0.000590
PTRATIO     -0.033481
B            0.000518
LSTAT       -0.030271
Intercept: 4.03592171504836
```

1.4.1 Regression: Actual values vs. predicted



1.4.2 Assessing the model (e.g. coefficient p-Values)

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.792
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	135.6
Date:	Mon, 04 Nov 2024	Prob (F-statistic):	3.68e-126
Time:	11:19:51	Log-Likelihood:	110.76
No. Observations:	404	AIC:	-197.5
Df Residuals:	392	BIC:	-149.5
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0359	0.226	17.819	0.000	3.591	4.481
CRIM	-0.0107	0.001	-8.002	0.000	-0.013	-0.008
ZN	0.0015	0.001	2.465	0.014	0.000	0.003
CHAS	0.0864	0.038	2.251	0.025	0.011	0.162
NOX	-0.6164	0.155	-3.990	0.000	-0.920	-0.313
RM	0.0761	0.018	4.155	0.000	0.040	0.112
DIS	-0.0527	0.008	-6.376	0.000	-0.069	-0.036
RAD	0.0137	0.003	5.060	0.000	0.008	0.019
TAX	-0.0006	0.000	-4.098	0.000	-0.001	-0.000
PTRATIO	-0.0335	0.006	-5.770	0.000	-0.045	-0.022
B	0.0005	0.000	4.611	0.000	0.000	0.001
LSTAT	-0.0303	0.002	-14.706	0.000	-0.034	-0.026

Omnibus:	30.564	Durbin-Watson:	2.072
Prob(Omnibus):	0.000	Jarque-Bera (JB):	117.705
Skew:	0.117	Prob(JB):	2.76e-26
Kurtosis:	5.634	Cond. No.	1.50e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.5e+04. This might indicate that there are strong multicollinearity or other numerical problems.

```
50    -0.056143
367    0.498215
34    -0.033868
78     0.043520
172    0.033242
         …
320   -0.041251
15    -0.033156
```

```
484    0.074891
125    0.008542
265   -0.214594
Length: 404, dtype: float64
0.03487337082354599
```

1.4.3 A. Checking that there's a linear relationship

1.4.4 B. Showing that the residuals are independent using the Durbin-Watson statistic

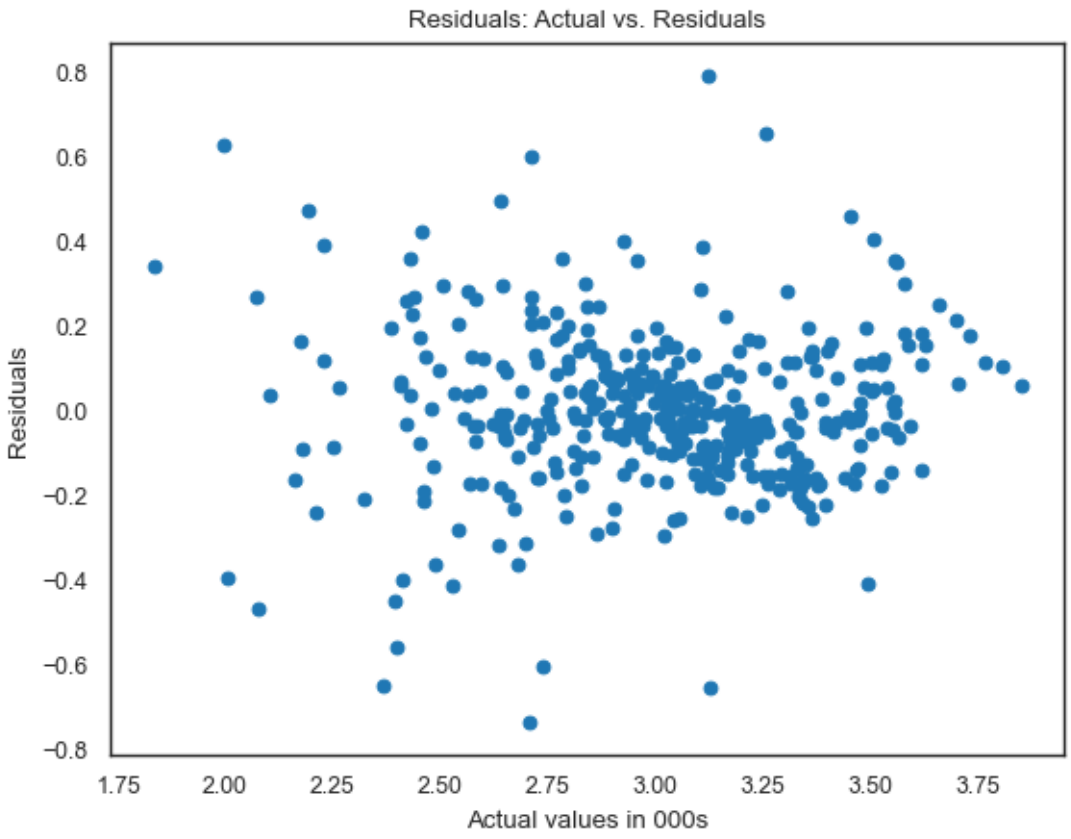
```
2.0716128816841124
```

Because the Durbin-Watson test returns a value between 1.5 and 2.5, autocorrelation is likely not a cause for concern.

Therefore, the residuals are independent.

1.4.5 C. Showing that the residuals display homoscedasticity (i.e. constant variance) using the Breusch-Pagan Test

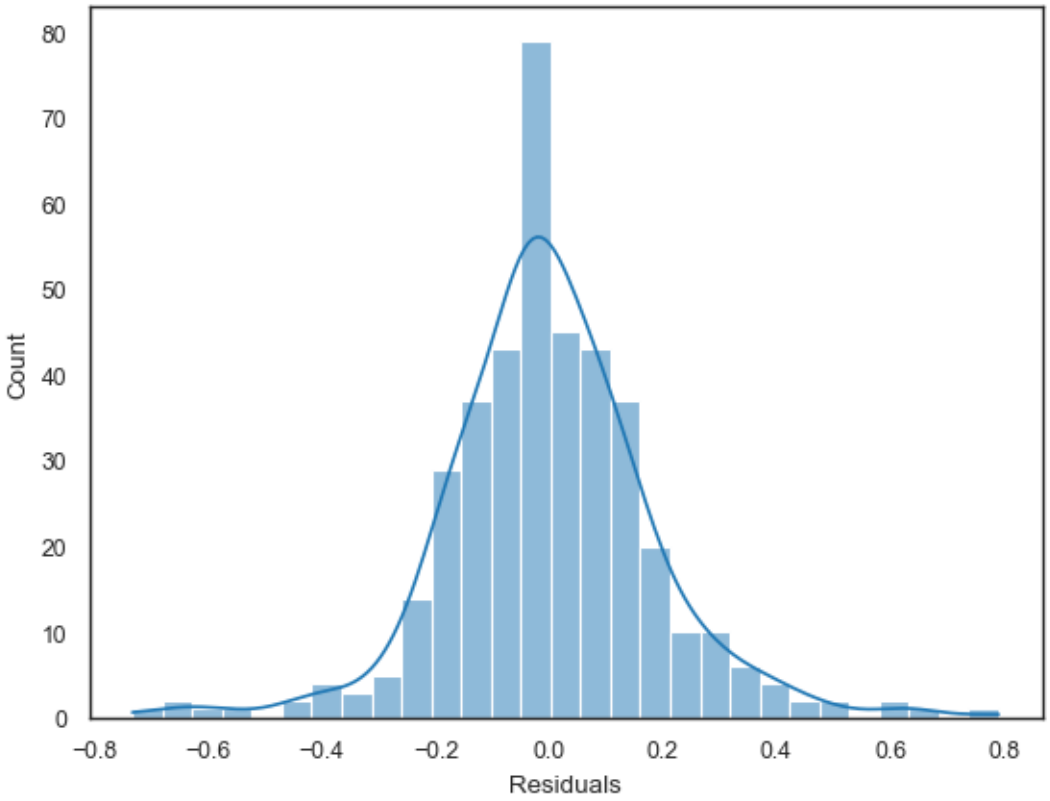
```
Lagrange multiplier statistic    5.792609e+01
p-value                          2.245454e-08
f-value                          5.964839e+00
f p-value                        5.194101e-09
dtype: float64
```



Since the p-value is less than our significance less (0.05), our regression is homoscedastic.

Source: https://www.statology.org/breusch-pagan-test/

1.4.6 D. Showing that the residuals display normality



1.4.7 E. Checking for multicollinearity using the variance influence factor (VIF)

```
CRIM       594.28
ZN           1.71
CHAS         2.29
NOX          1.07
RM           3.83
DIS          1.75
RAD          3.59
TAX          6.79
PTRATIO      7.11
B            1.79
LSTAT        1.33
dtype: float64
```

The cutoff for multicollinearity is 10, and none of the features meet that cutoff.

Therefore, we can say that there is no multicollinearity.

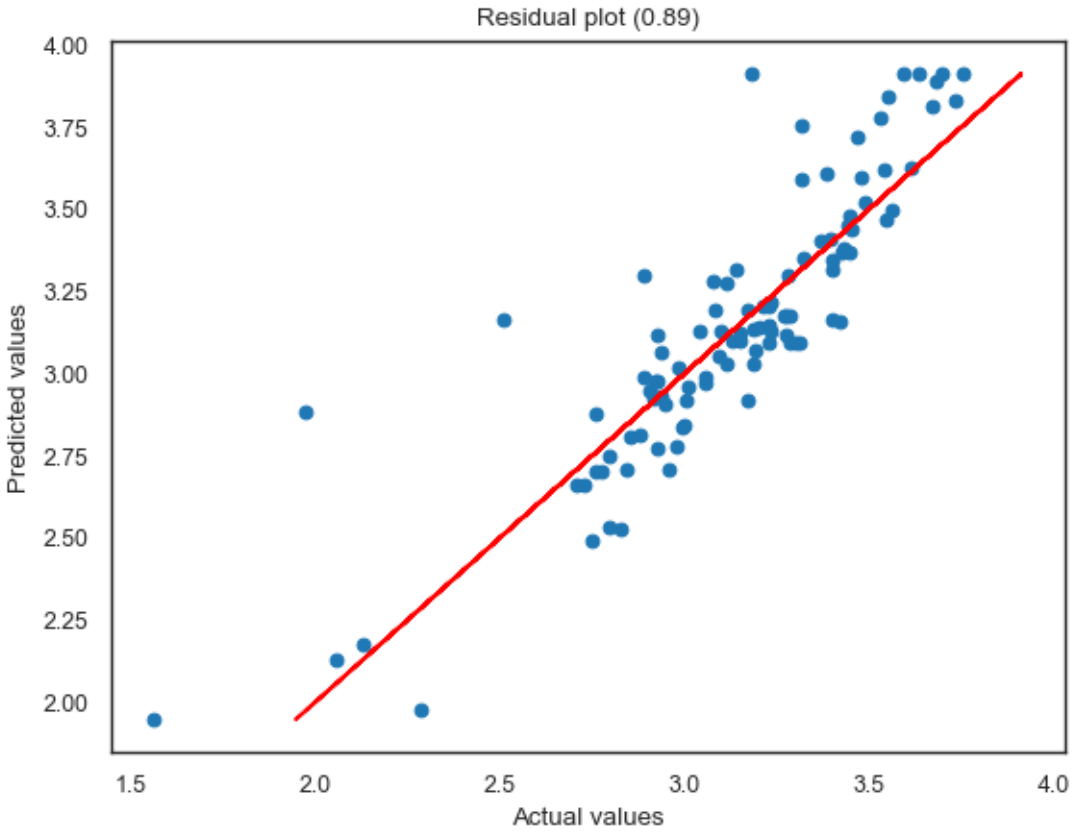
1.4.8 Test Results

```
R-squared: 0.7490934185196063
         Coefficients
CRIM        -0.010702
ZN           0.001461
CHAS         0.086449
NOX         -0.616448
RM           0.076133
DIS         -0.052692
RAD          0.013743
TAX         -0.000590
PTRATIO     -0.033481
B            0.000518
LSTAT       -0.030271
Intercept: 4.03592171504836
```

1.4.9 How well does the training set fit the test set? Do new data points (i.e. test set) have similar coefficients, MSE, R-Squared, etc. as the training set?

Metric	Training	Testing
R^2…	0.7919	0.7491
MSE…	407.3670	0.0419
RMSE…	20.1833	0.2046

1.4.10 Regression: Actual values vs. predicted



OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.802
Model:	OLS	Adj. R-squared:	0.777
Method:	Least Squares	F-statistic:	33.08
Date:	Mon, 04 Nov 2024	Prob (F-statistic):	6.22e-27
Time:	11:19:53	Log-Likelihood:	29.106
No. Observations:	102	AIC:	-34.21
Df Residuals:	90	BIC:	-2.712
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.3113	0.473	9.122	0.000	3.372	5.250
CRIM	-0.0071	0.005	-1.318	0.191	-0.018	0.004
ZN	-0.0007	0.001	-0.516	0.607	-0.003	0.002
CHAS	0.1604	0.073	2.193	0.031	0.015	0.306
NOX	-1.1155	0.358	-3.115	0.002	-1.827	-0.404

```
RM             0.1501      0.037      4.053      0.000       0.077       0.224
DIS           -0.0474      0.017     -2.849      0.005      -0.080      -0.014
RAD            0.0095      0.007      1.374      0.173      -0.004       0.023
TAX           -0.0003      0.000     -0.853      0.396      -0.001       0.000
PTRATIO       -0.0471      0.011     -4.195      0.000      -0.069      -0.025
B             -0.0004      0.000     -1.283      0.203      -0.001       0.000
LSTAT         -0.0250      0.005     -4.862      0.000      -0.035      -0.015
==============================================================================
Omnibus:                       28.336   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               50.440
Skew:                           1.156   Prob(JB):                     1.11e-11
Kurtosis:                       5.554   Cond. No.                     1.44e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.44e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
305    0.018684
193   -0.002554
65    -0.166738
349    0.195032
151    0.166675
         …
208    0.020968
174   -0.132241
108   -0.060968
242   -0.068103
102   -0.283282
Length: 102, dtype: float64
Mean absolute error: 0.03749988129807514
```

1.4.11 A. Checking that there's a linear relationship

1.4.12 B. Showing that the residuals are independent using the Durbin-Watson statistic

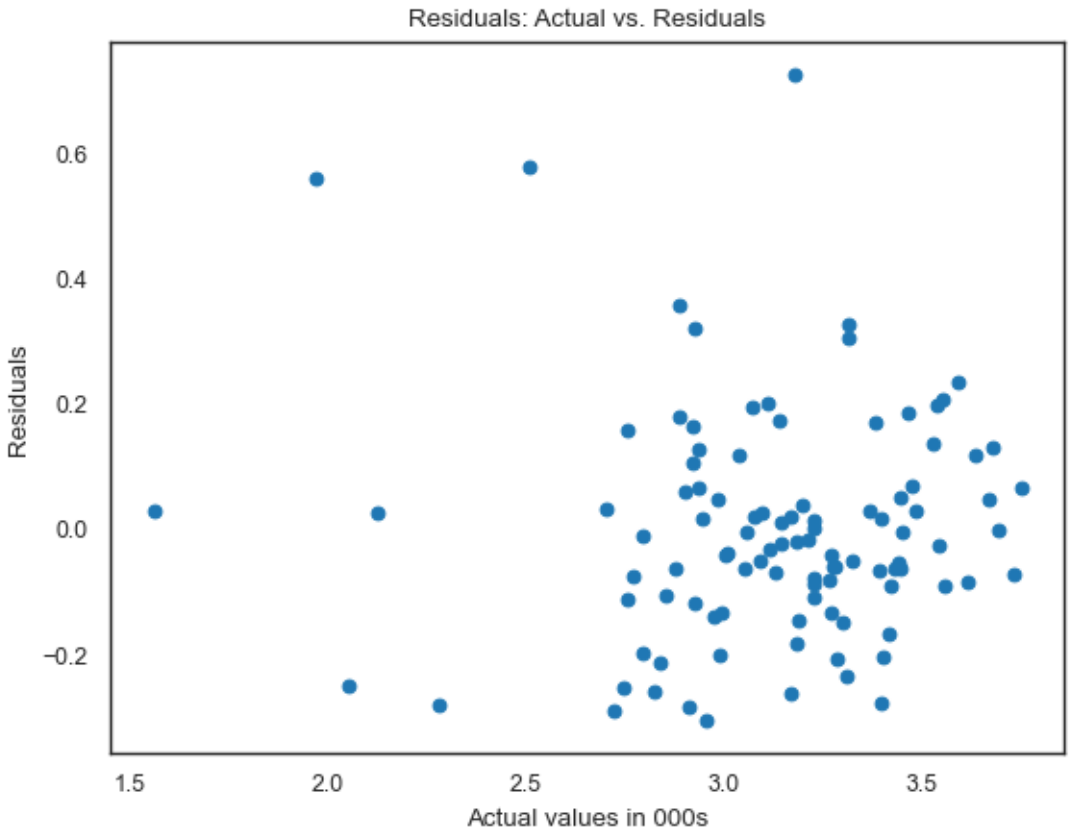
```
1.9925897546011744
```

Because the Durbin-Watson test returns a value between 1.5 and 2.5, autocorrelation is likely not a cause for concern.

Therefore, the residuals are independent.

1.4.13 C. Showing that the residuals display homoscedasticity (i.e. constant variance) using the Breusch-Pagan Test

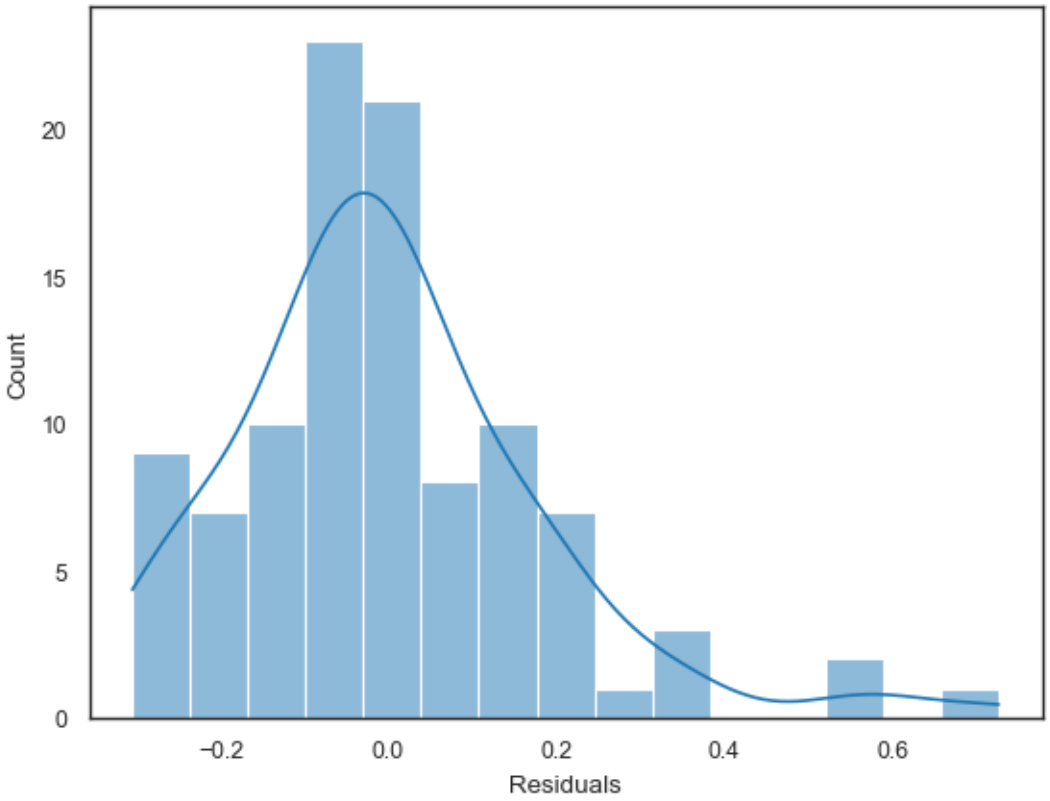
```
Lagrange multiplier statistic    23.792137
p-value                           0.013639
f-value                           2.489046
f p-value                         0.008936
dtype: float64
```



Since the p-value is less than our significance less (0.05), our regression is homoscedastic.

Source: https://www.statology.org/breusch-pagan-test/

1.4.14 D. Showing that the residuals display normality



1.4.15 E. Checking for multicollinearity using the variance influence factor (VIF)

```
CRIM       607.66
ZN           2.97
CHAS         2.23
NOX          1.05
RM           3.97
DIS          2.31
RAD          2.99
TAX          7.50
PTRATIO      8.29
B            1.68
LSTAT        1.63
dtype: float64
```

The cutoff for multicollinearity is 10, and none of the features meet that cutoff.

Therefore, we can say that there is no multicollinearity.

1.5 6) Evaluation: How well does the model do?

```
The prediction interval is 34999.98140857814 <= 35000 <= 35000.01859142186
```