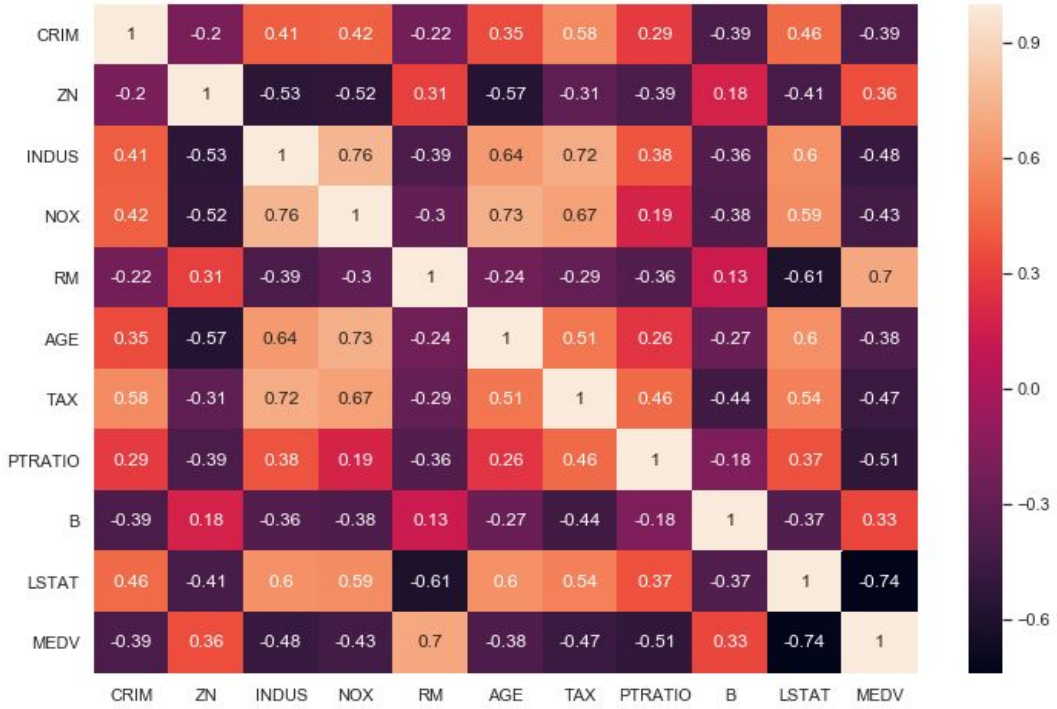
Regression Analysis of Boston House Prices

Purpose

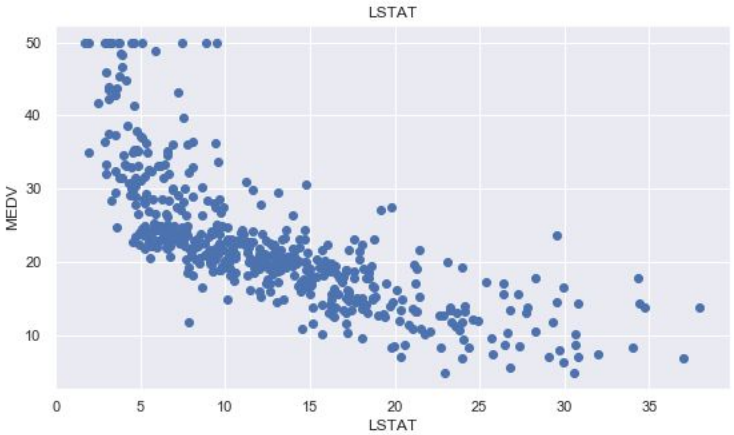
To determine which variables have a high correlation with median home values

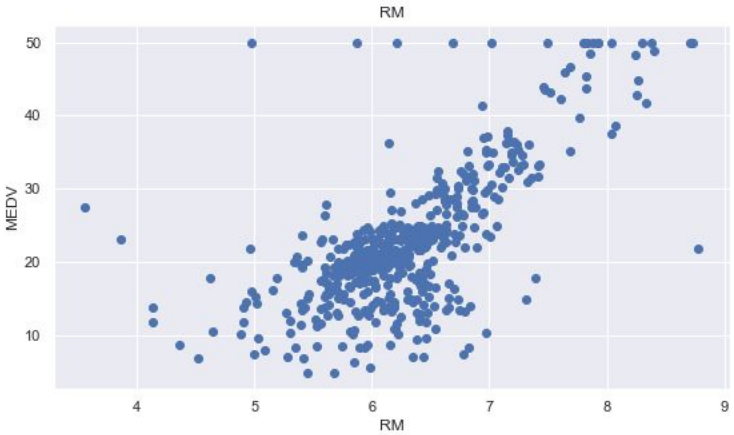
Creating a Model

- Choose predictors and target variable
  - Target: median home values (MEDV)
- Check for normality


CRIM 1 -0.2 0.41 0.42 -0.22 0.35 0.58 0.29 -0.39 0.46 -0.39
ZN -0.2 1 -0.53 -0.52 0.31 -0.57 -0.31 -0.39 0.18 -0.41 0.36
INDUS 0.41 -0.53 1 0.76 -0.39 0.64 0.72 0.38 -0.36 0.6 -0.48
NOX 0.42 -0.52 0.76 1 -0.3 0.73 0.67 0.19 -0.38 0.59 -0.43
RM -0.22 0.31 -0.39 -0.3 1 -0.24 -0.29 -0.36 0.13 -0.61 0.7
AGE 0.35 -0.57 0.64 0.73 -0.24 1 0.51 0.26 -0.27 0.6 -0.38
TAX 0.58 -0.31 0.72 0.67 -0.29 0.51 1 0.46 -0.44 0.54 -0.47
PTRATIO 0.29 -0.39 0.38 0.19 -0.36 0.26 0.46 1 -0.18 0.37 -0.51
B -0.39 0.18 -0.36 -0.38 0.13 -0.27 -0.44 -0.18 1 -0.37 0.33
LSTAT 0.46 -0.41 0.6 0.59 -0.61 0.6 0.54 0.37 -0.37 1 -0.74
MEDV -0.39 0.36 -0.48 -0.43 0.7 -0.38 -0.47 -0.51 0.33 -0.74 1
CRIM ZN INDUS NOX RM AGE TAX PTRATIO B LSTAT MEDV
0.9
0.6
0.3
0.0
−0.3
−0.6

Base Model Predictors





Base Model Results

Dep. Variable:	MEDV	**R-squared:**	0.639
Model:	OLS	**Adj. R-squared:**	0.637
Method:	Least Squares	**F-statistic:**	444.3
Date:	Wed, 02 Oct 2019	**Prob (F-statistic):**	7.01e-112
Time:	14:30:58	**Log-Likelihood:**	-1582.8
No. Observations:	506	**AIC:**	3172.
Df Residuals:	503	**BIC:**	3184.
Df Model:	2		
Covariance Type:	nonrobust		

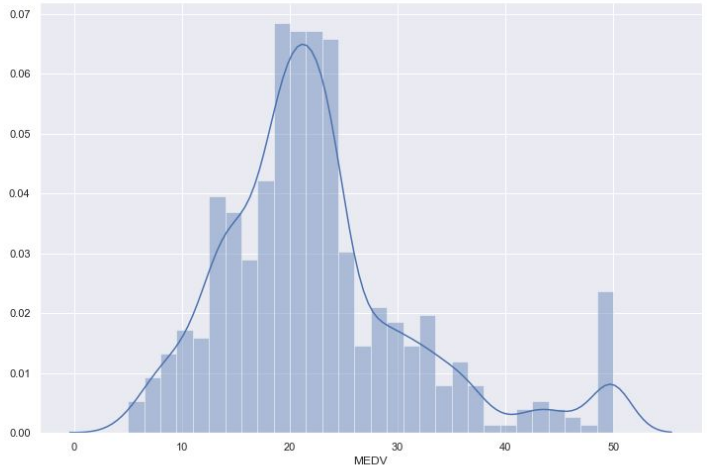
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.3583	3.173	-0.428	0.669	-7.592	4.875
RM	5.0948	0.444	11.463	0.000	4.222	5.968
LSTAT	-0.6424	0.044	-14.689	0.000	-0.728	-0.556

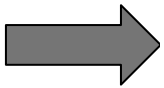
Omnibus:	145.712	**Durbin-Watson:**	0.834
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	457.690
Skew:	1.343	**Prob(JB):**	4.11e-100
Kurtosis:	6.807	**Cond. No.**	202.

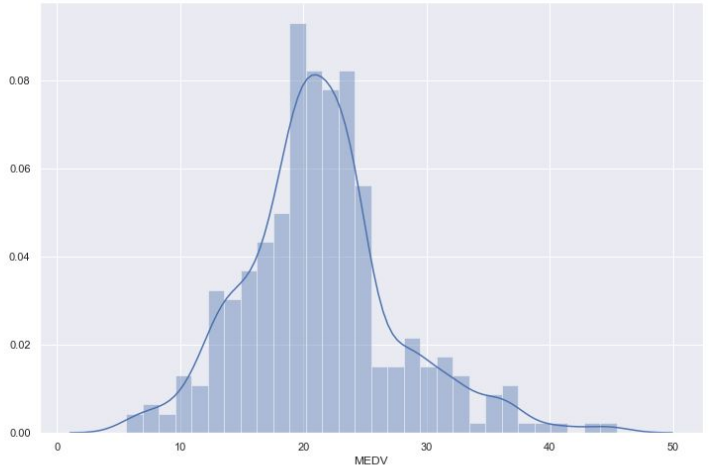
Not bad, but we can do better!

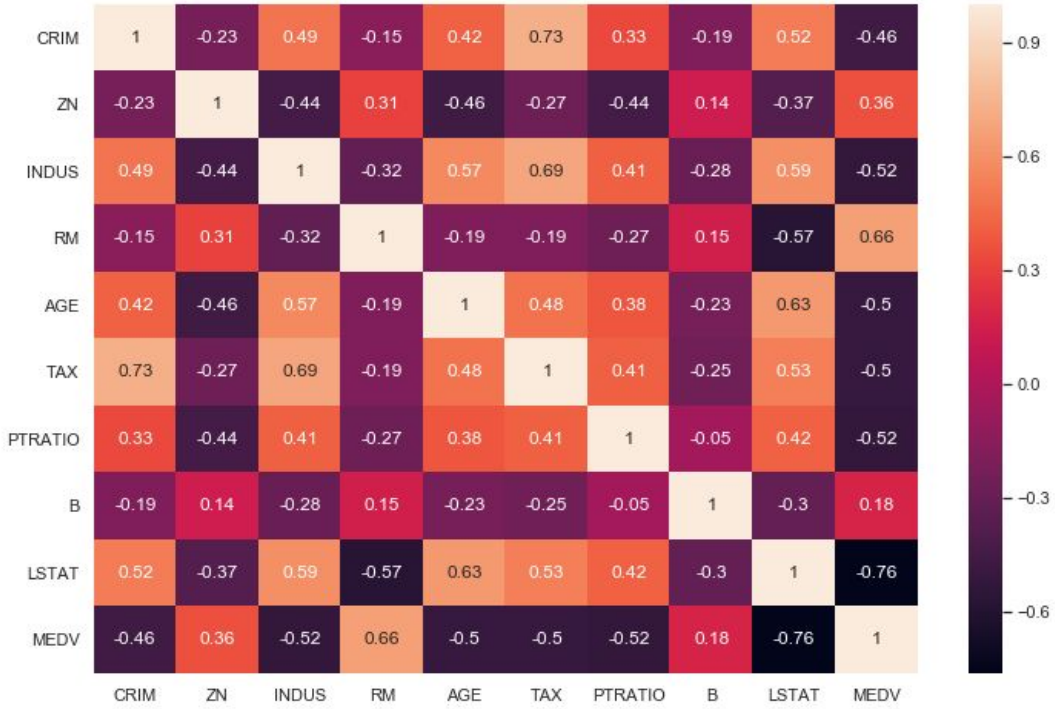
Normalization

Removed all values with std > 0.25








CRIM 1 -0.23 0.49 -0.15 0.42 0.73 0.33 -0.19 0.52 -0.46
ZN -0.23 1 -0.44 0.31 -0.46 -0.27 -0.44 0.14 -0.37 0.36
INDUS 0.49 -0.44 1 -0.32 0.57 0.69 0.41 -0.28 0.59 -0.52
RM -0.15 0.31 -0.32 1 -0.19 -0.19 -0.27 0.15 -0.57 0.66
AGE 0.42 -0.46 0.57 -0.19 1 0.48 0.38 -0.23 0.63 -0.5
TAX 0.73 -0.27 0.69 -0.19 0.48 1 0.41 -0.25 0.53 -0.5
PTRATIO 0.33 -0.44 0.41 -0.27 0.38 0.41 1 -0.05 0.42 -0.52
B -0.19 0.14 -0.28 0.15 -0.23 -0.25 -0.05 1 -0.3 0.18
LSTAT 0.52 -0.37 0.59 -0.57 0.63 0.53 0.42 -0.3 1 -0.76
MEDV -0.46 0.36 -0.52 0.66 -0.5 -0.5 -0.52 0.18 -0.76 1
CRIM ZN INDUS RM AGE TAX PTRATIO B LSTAT MEDV
0.9
0.6
0.3
0.0
−0.3
−0.6

Final Results

Dep. Variable:	MEDV	**R-squared:**	0.719
Model:	OLS	**Adj. R-squared:**	0.712
Method:	Least Squares	**F-statistic:**	108.4
Date:	Wed, 02 Oct 2019	**Prob (F-statistic):**	1.14e-88
Time:	13:50:09	**Log-Likelihood:**	-910.79
No. Observations:	348	**AIC:**	1840.
Df Residuals:	339	**BIC:**	1874.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.6055	4.067	2.853	0.005	3.605	19.606
RM	5.0552	0.501	10.080	0.000	4.069	6.042
LSTAT	-0.3639	0.055	-6.607	0.000	-0.472	-0.256
TAX	-0.0028	0.002	-1.404	0.161	-0.007	0.001
CRIM	-0.1201	0.061	-1.962	0.051	-0.241	0.000
ZN	-0.0287	0.017	-1.739	0.083	-0.061	0.004
INDUS	0.0112	0.044	0.257	0.797	-0.075	0.097
AGE	-0.0218	0.010	-2.252	0.025	-0.041	-0.003
PTRATIO	-0.7573	0.121	-6.239	0.000	-0.996	-0.519

Recommendations

- Number of rooms is the greatest positive predictor for home values
- Crime rate and low socioeconomic status are negative predictors