Regression Analysis of Boston House Prices
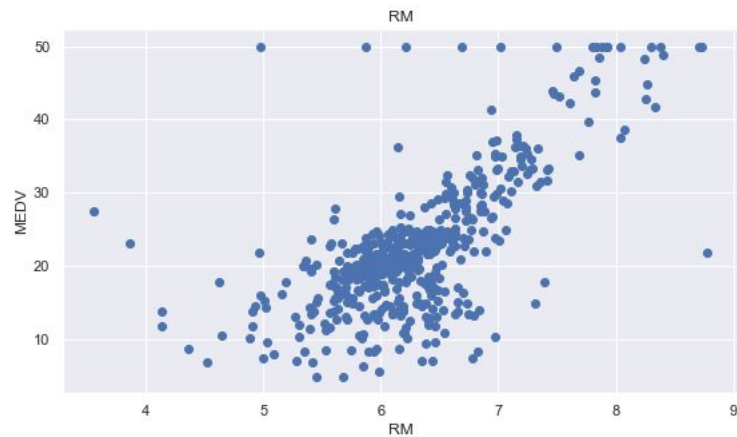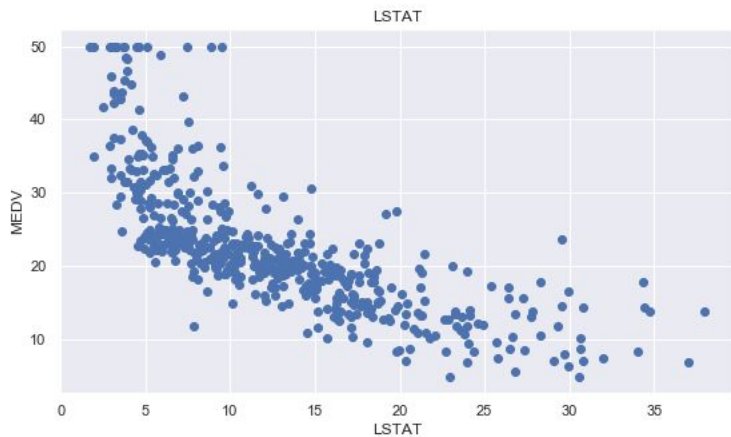
Ahmed Safdar

Purpose

- Determine which factors predict a higher home value for a construction company.
- Also determine which factors predict a lower value.

Creating a Model

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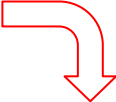- Check for normality

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

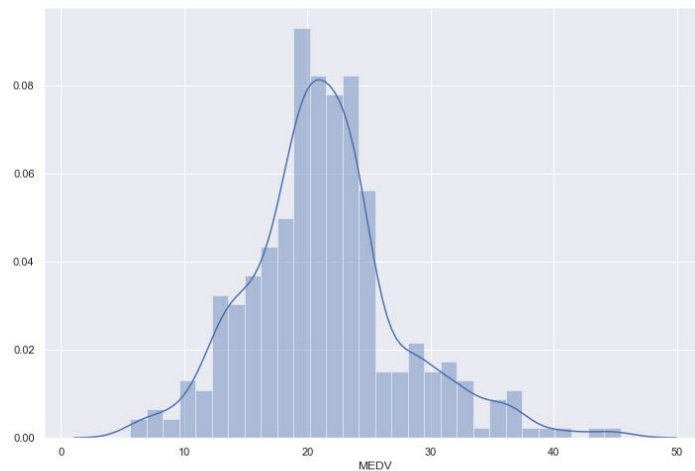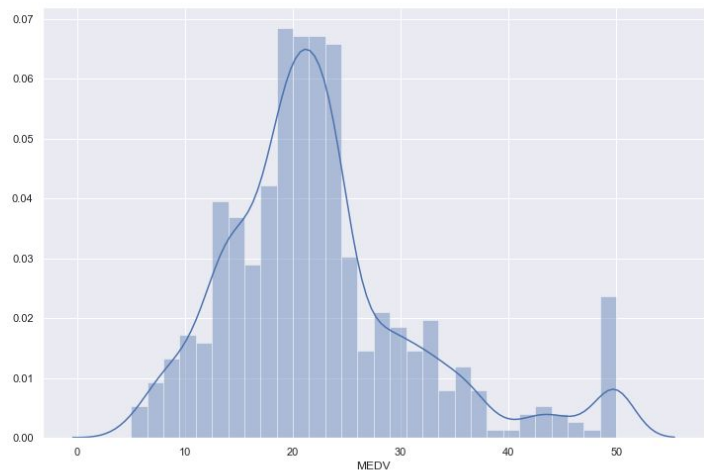| Omnibus:	145.712	Durbin-Watson:	0.834
Prob(Omnibus):	0.000	Jarque-Bera (JB):	457.690
Skew:	1.343	Prob(JB):	4.11e-100
Kurtosis:	6.807	Cond. No.	202.

This means that 64% of the variance in our target variable (median values) can be explained by our predictors.
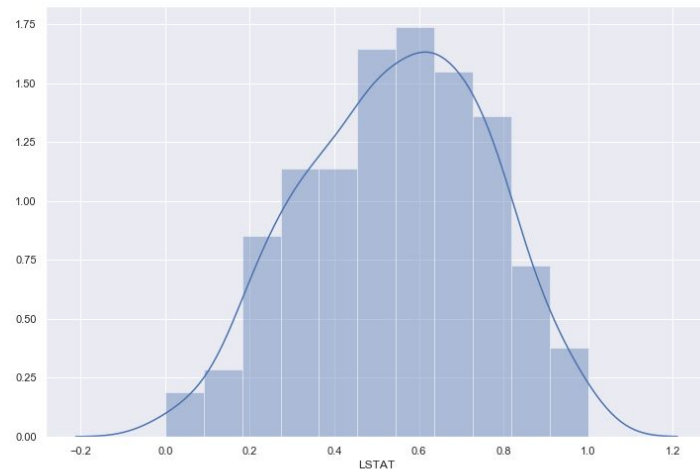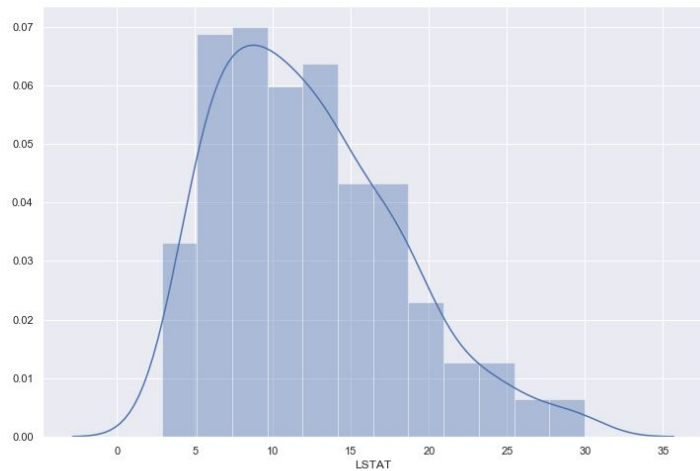
Not bad, but we can do better!

Improving Our Model

Removed all values with std > 2.5

Improving Our Model (contd.)

	CRIM	ZN	INDUS	RM	AGE	TAX	PTRATIO	B	LSTAT	MEDV
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

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

In general, a higher R-squared means that the model is a better fit.

Coefficients are the values that multiply the predictor variables.

Recommendations

- Number of rooms is the greatest positive predictor for home values
- Crime rate, low socioeconomic status, and a high student-to-teacher ratio are negative predictors