MSDS 422 Assignment 2

In order to predict the market value of homes in Boston for a real estate firm, regression models were employed. In order to determine which regression model works best for this case, it is important to compare each method by comparing the predictions made on a sample data set. This data contained a list of 506 houses with the following attributes:

Variable Name	*Description*
neighborhood	Name of the Boston neighborhood (location of the census tract)
mv	Median value of homes in thousands of 1970 dollars
nox	Air pollution (nitrogen oxide concentration)
crim	Crime rate
zn	Percent of land zoned for lots
indus	Percent of business that is industrial or nonretail
chas	On the Charles River (1) or not (0)
rooms	Average number of rooms per home
age	Percentage of homes built before 1940
dis	Weighted distance to employment centers
rad	Accessibility to radial highways
tax	Tax rate
ptratio	Pupil/teacher ratio in public schools
lstat	Percentage of population of lower socio-economic status

Some preliminary statistics and graphs were used in order to receive a better understanding of the data. The following is the descriptive statistics of each attribute:

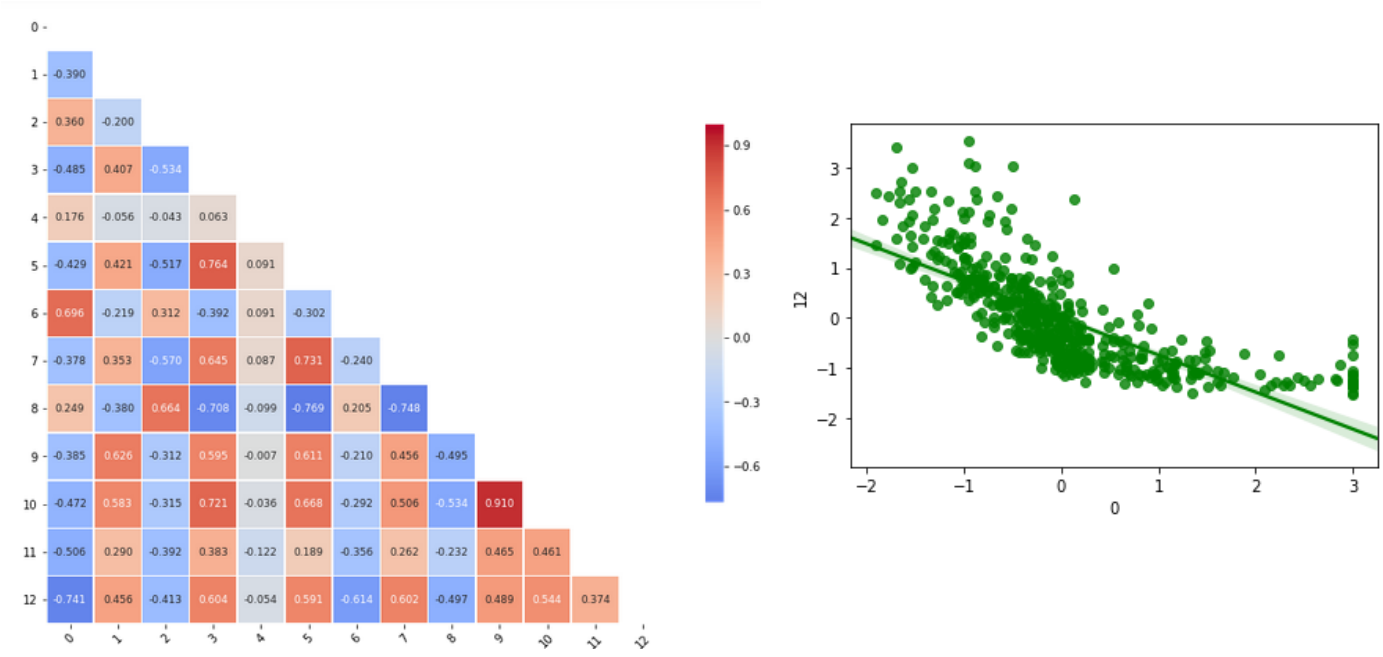
Descriptive statistics of the boston DataFrame:

	crim	zn	indus	chas	nox	rooms	\
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%	3.677082	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	lstat	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	68.574901	3.795043	9.549407	408.237154	18.455534	12.653063	
std	28.148861	2.105710	8.707259	168.537116	2.164946	7.141062	
min	2.900000	1.129600	1.000000	187.000000	12.600000	1.730000	
25%	45.025000	2.100175	4.000000	279.000000	17.400000	6.950000	
50%	77.500000	3.207450	5.000000	330.000000	19.050000	11.360000	
75%	94.075000	5.188425	24.000000	666.000000	20.200000	16.955000	
max	100.000000	12.126500	24.000000	711.000000	22.000000	37.970000	

	mv
count	506.000000
mean	22.528854
std	9.182176
min	5.000000
25%	17.025000
50%	21.200000
75%	25.000000
max	50.000000

One can notice that these attributes are on different scales, so it will be important use scaling techniques on the data in order to get the data on a common scale without distorting it. Each variable was then plotted against the attribute of interest, median value of homes in order to view correlations. As seen by this heat graph and scatter plot, median value(attribute #12) is most correlated with crime rate (attribute #0).



Next, multiple regression methods were compared using the scaled data and having median value as the response variable. Linear regression, Lasso regression, Ridge regression and ElasticNet Regression were all compared using the Root Mean Square Error (RMSE). Also known as prediction error, this value is an indicator of how far off the regression line is from the test data set. The lower the RMSE value, the closer the regression fits the test data. The results of each regression are the following:

- Linear : 0.5619360369051359
- Lasso : 0.9620739310707962
- Ridge : 0.561424623010174
- ElasticNet : 0.8412368223631603

As seen, by the results, Linear and Ridge regressions have the lowest RMSE values and fit the test datasets the best.

When recommending to the real estate firm model to choose, I would advise them to choose either linear or ridge regression. This is due to the fact that they have very similar RMSE values and are both lower than the other two models. If they were to decide between the two, I would recommend to obtain more data if possible and further compare the two. The more data present, the more accurate the model will be and it potentially differentiate the two.

Note: Code and output pdf and python notebook submitted as well.