Boston Housing Price Prediction

Aditi Patil and Masha Volkova

Dataset Overview

Boston Housing Dataset

- Derived from the U.S. Census Service
- Housing in the area of Boston, MA
- 506 entries

Link:

https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html)

Why Housing Price Prediction?

- Purchasing home- big decision
- Important to consider the impact of different characteristics of the houses such as location, proximity to businesses and employment, environmental factors such as pollution, and social factors such as crime rate.
- Improve real estate efficiency through housing price prediction by targeting the optimal houses to buy and sell at a given time.

Objective

- Build Ridge and Lasso regression models to predict the median value of Boston houses using the Boston Housing Dataset
- Goal: Study the impact of high correlation among predictor variables
- Goal: Compare the performance of Ridge and Lasso regression models
    - Programming language: Python

Variables

The predictor variables:

`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

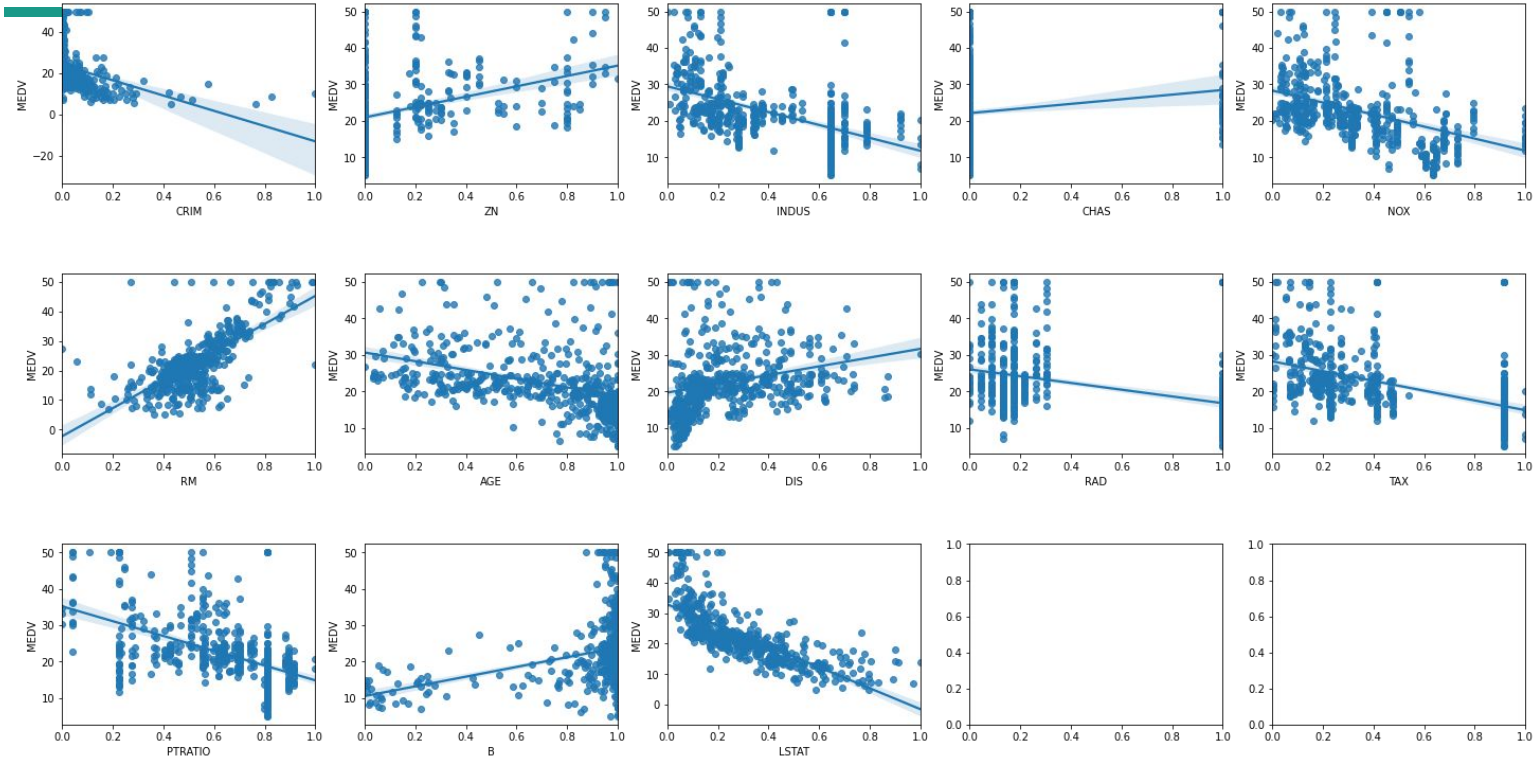
The response variable:

`MEDV` - Median value of owner-occupied homes in $1000's

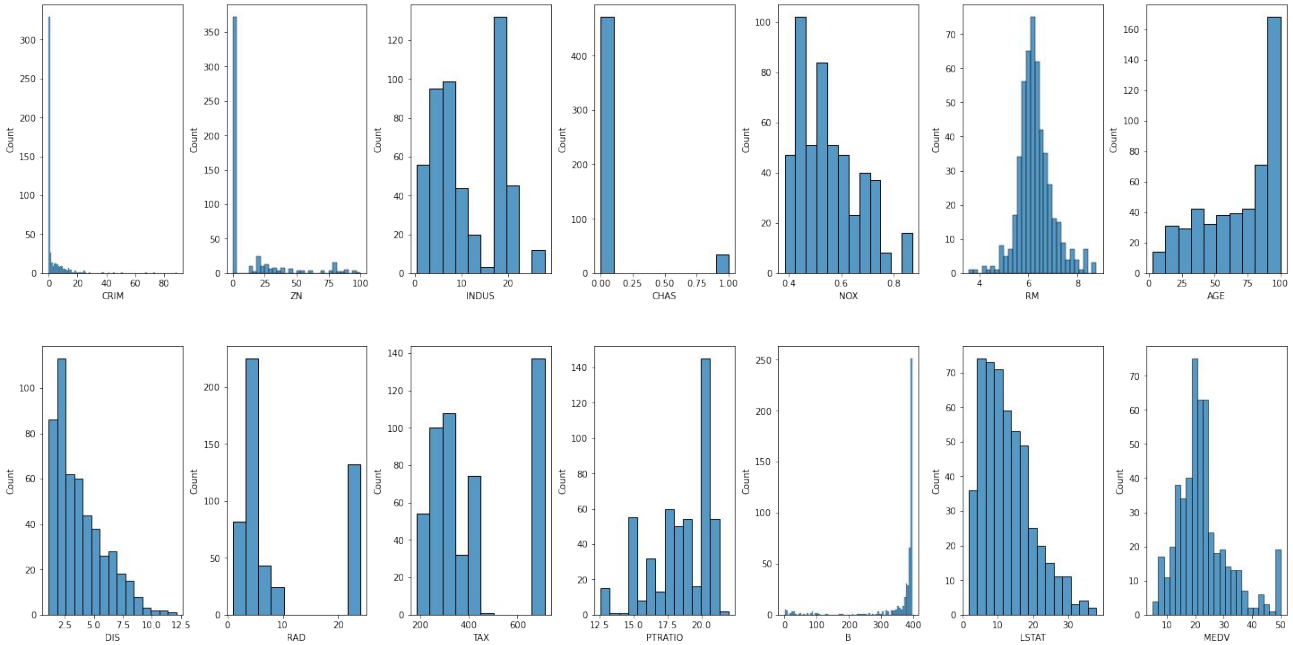
Data Preprocessing

- Since Ridge and Lasso Regression were going to be used, the columns were scaled
  - This is done to ensure that the regularization term has equal scaling effects on each of the coefficients
- We dropped the missing/null values
- Dropped the duplicates
- To handle categorical attributes “CHAS” and “RAD” we added dummy variable columns

Plots of Predictors vs Response

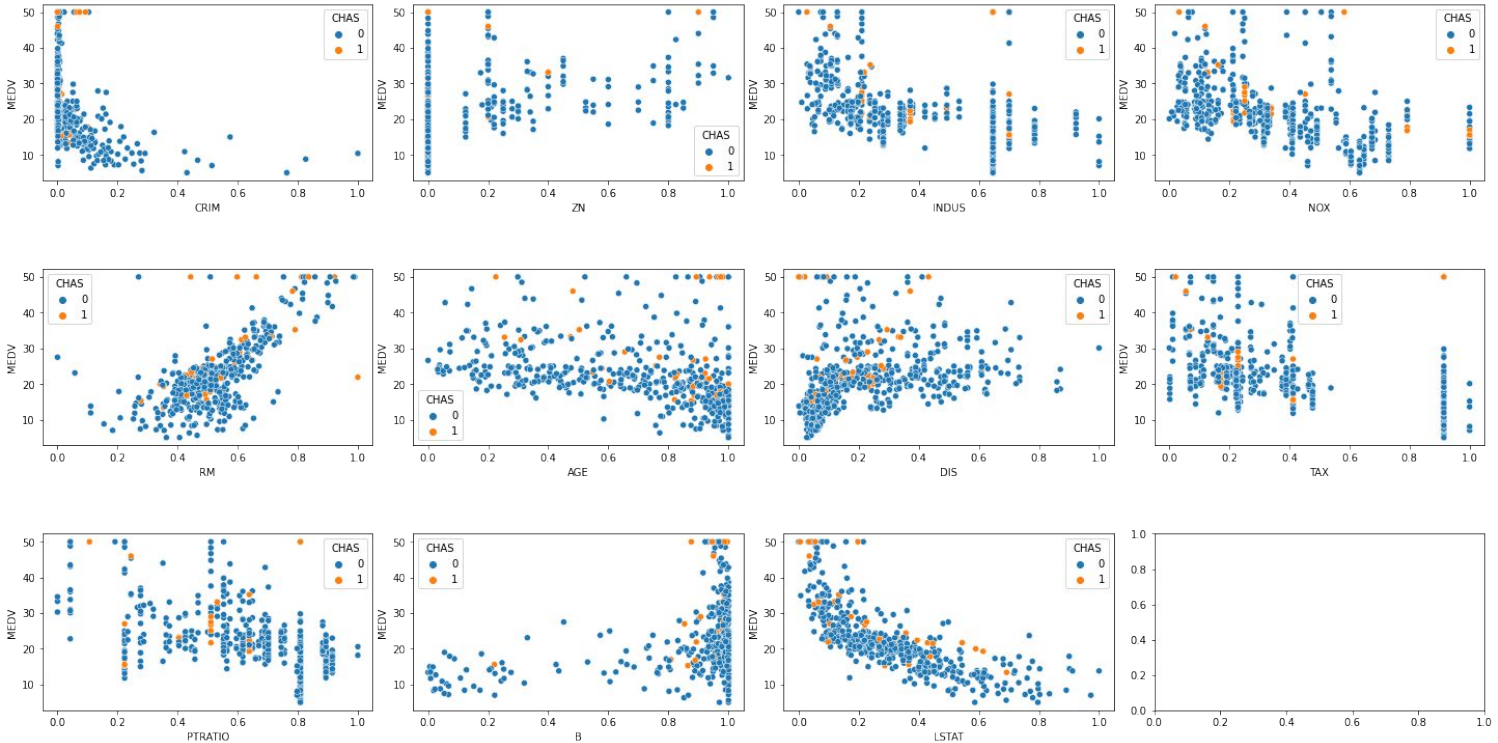


Plots of Distributions

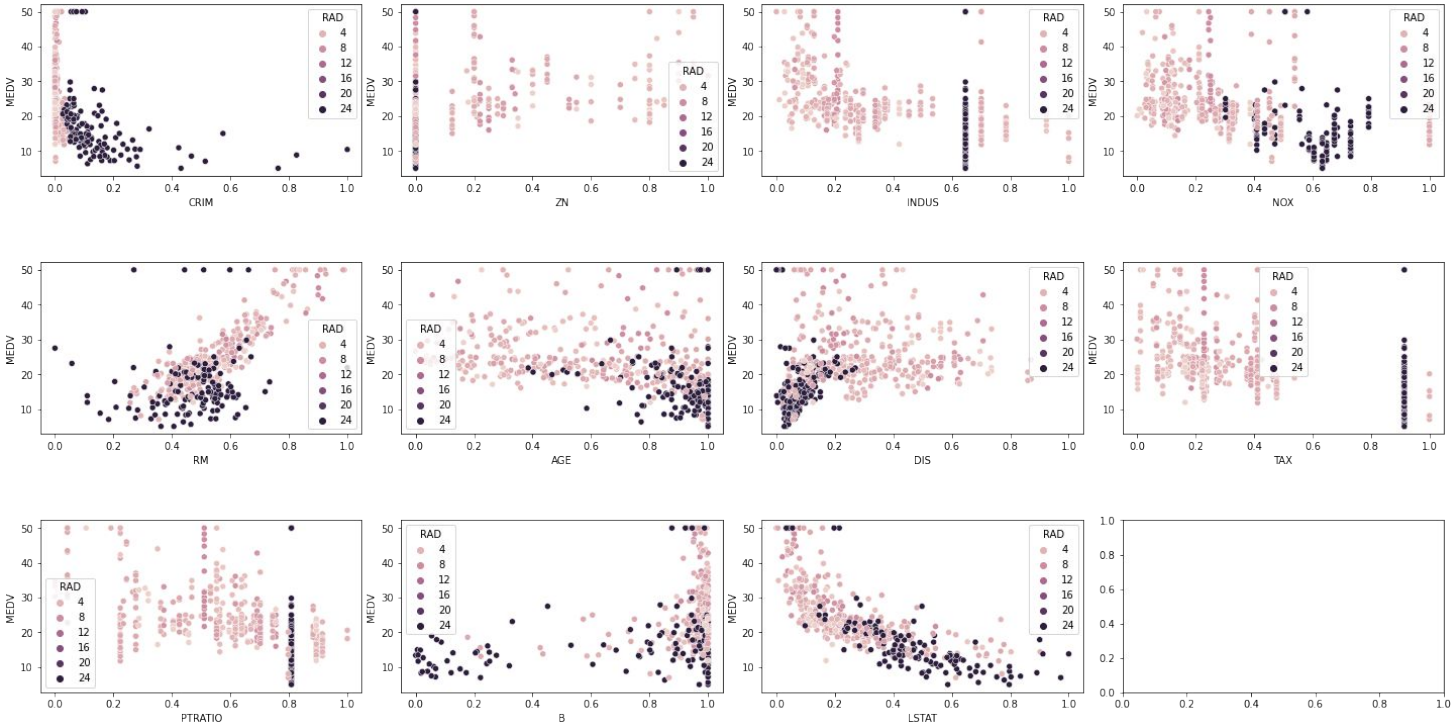


Correlation Matrix

“CHAS” vs the other predictors with respect to response



“RAD” vs the other predictors with respect to response



Process

We did the following twice : 1) all the variables 2) dropping “RAD”

- We split the dataset into training and test sets where the test size was 20%
- Trained the data with Ridge and Lasso regression (separately) with 10-fold cross validation to find the optimal alpha value (same as lambda from lecture slides)
- Trained the data with regular Ridge and Lasso regression using the optimal alpha value
- Predicted housing price values using test set and obtained mean squared error and r^2 values

Finding the optimal lambda and fitting the model

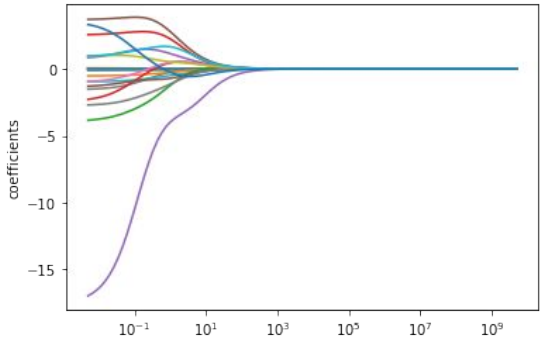
```
# Split data into training and test sets
X_train, X_test , y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=1)

# Find the optimal a value using cross validation
ridgecv = RidgeCV(alphas = lambdas, scoring = 'neg_mean_squared_error', normalize = True)
ridgecv.fit(X_train, y_train)
print("λ value: ", ridgecv.alpha_)

# Using the optimal a to apply ridge regression model
ridge = Ridge(alpha = ridgecv.alpha_, normalize = True)
ridge.fit(X_train, y_train) # Fit a ridge regression on the training data
pred = ridge.predict(X_test)  # Use this model to predict the test data
results_ridge =  pd.Series(ridge.coef_, index = X.columns)
```

Code to make plot of lambda vs coefficients :

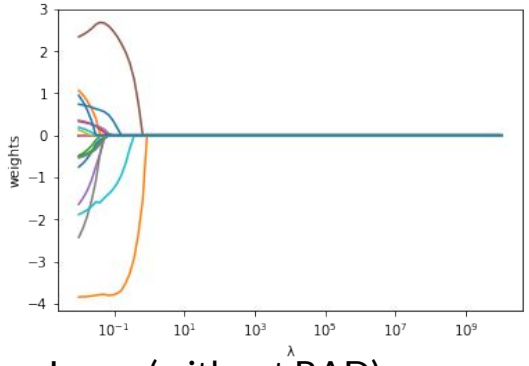
```
# For each a value fit the model and plot λ vs coefficients
for λ in lambdas:
    ridge.set_params(alpha = λ)
    ridge.fit(X, y)
    coefs.append(ridge.coef_)

np.shape(coefs)

ax = plt.gca()
ax.plot(lambdas, coefs)
ax.set_xscale('log')
plt.axis('tight')
plt.xlabel('λ')
plt.ylabel('coefficients')
```

Summary: Lambda values vs Coefficient

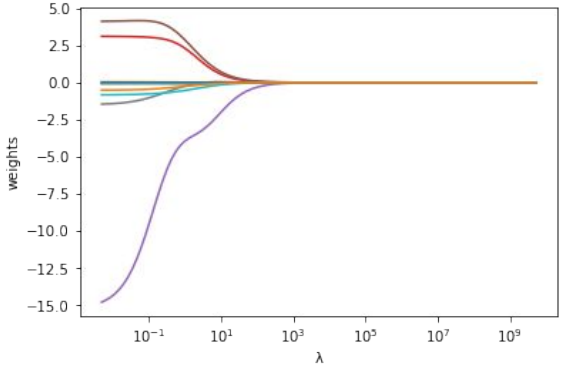
Ridge with high correlation (with RAD)



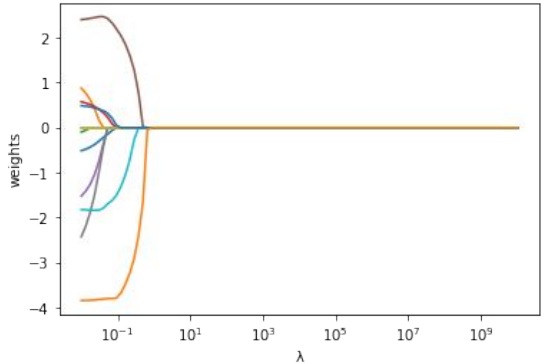
Lasso with high correlation (with RAD)



Ridge (without RAD)



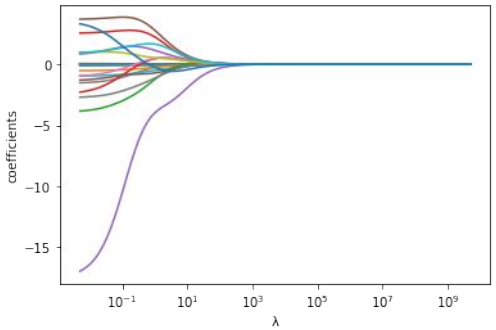
Lasso (without RAD)



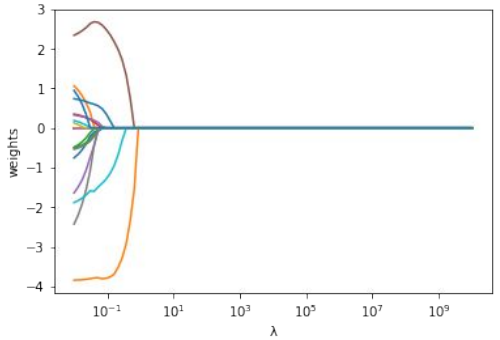
Why does the lambda behaviour make sense?

Ridge Regression	$\underset{\theta}{argmin}\ SSE + \lambda \sum_{i=1}^{K} \theta_i^2$
Lasso Regression	$\underset{\theta}{argmin}\ SSE + \lambda \sum_{i=1}^{K} \lvert\theta_i\rvert$

Ridge (with RAD)



Lasso with (with RAD)

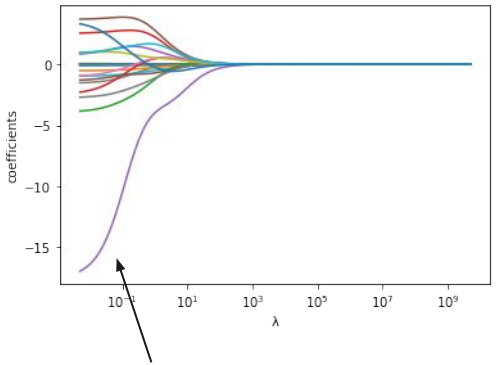


- Ridge and Lasso regression regularize the linear regression by imposing a penalty on the size of coefficients
  - When the λ's were large, the penalty term is very large so the β is driven to zero
  - When the λ's were small, the penalty is not as large, the β computed is allowed to be large

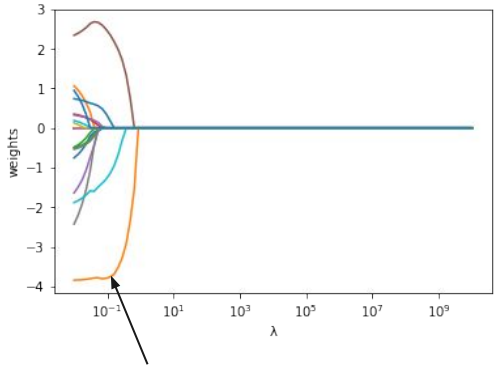
Why does the lambda behaviour make sense? (continued)

Ridge Regression	$\underset{\theta}{argmin}\ SSE + \lambda \sum_{i=1}^{K} \theta_i^2$
Lasso Regression	$\underset{\theta}{argmin}\ SSE + \lambda \sum_{i=1}^{K} \lvert\theta_i\rvert$

Ridge (with RAD)



Lasso (with RAD)



- From the figure above, Lasso drives coefficients to zero much quicker than the Ridge
  - This is a consequence of the shape of the L1 norm decision boundary which allows for coefficients to be set to zero.
- The magnitude of the coefficients in Ridge is much higher for certain variables than in Lasso

Summary:

Coefficients

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

	Ridge	Ridge (RAD dropped)	Lasso	Lasso (RAD dropped)
CRIM	-0.091985	-0.075512	-0.101130	-0.057496
ZN	0.048255	0.043540	0.062845	0.039761
INDUS	-0.076575	-0.056472	-0.043202	-0.016730
CHAS	1.440306	2.481992	1.304383	2.172716
NOX	-12.647999	-14.910269	-16.340509	-13.408384
RM	3.362380	3.480742	3.179704	3.487117
AGE	0.004999	-0.001513	0.008150	-0.000000
DIS	-1.023857	-1.313067	-1.329288	-1.188105
TAX	-0.002798	0.001932	-0.005735	0.000000
PTRATIO	-0.835468	-0.865079	-0.888832	-0.839858
B	0.008660	0.006361	0.008599	0.005386
LSTAT	-0.477832	-0.511279	-0.528166	-0.533956
RAD_1	-3.587521	NaN	-3.136548	NaN
RAD_2	-0.598936	NaN	-0.245179	NaN
RAD_3	0.700142	NaN	1.300276	NaN
RAD_4	-0.835709	NaN	-0.006810	NaN
RAD_5	-0.675505	NaN	-0.000000	NaN
RAD_6	-2.944628	NaN	-2.220763	NaN
RAD_7	0.681946	NaN	1.580950	NaN
RAD_8	0.725187	NaN	1.377779	NaN
RAD_24	2.345582	NaN	4.408863	NaN

Summary: R^2 values and Mean Squared Errors on Test

Mean squared error

Ridge	23.110737
Ridge (RAD dropped)	24.834684
Lasso	22.448442
Lasso (RAD dropped)	24.902874

R^2

Ridge	0.736638
Ridge (RAD dropped)	0.748706
Lasso	0.744185
Lasso (RAD dropped)	0.748016

Impacts of correlation (Multicollinearity)

- Coefficients- sensitive to small changes in model, coefficient estimates can be impacted significantly by which predictor variables are included in the model
- **Multicollinearity**- is the occurrence of high intercorrelations between two or more predictor variables in a multiple regression model.
    - Weakens regression model because precision of estimated coefficients is reduced
    - Higher than 0.7- candidate for removal

- Lasso and Ridge- can handle collinearity, but only to a certain extent
- Standardization of variables- reduces collinearity

Future improvements

- Check for numerical interactions in predictor variables, especially with NOX, by adding interaction terms to model matrix (check if lasso drives their coefficients to zero)
- Dealing with outliers
- Dropping variables with moderately high correlation such as “DIS” and “NOX” and checking if our models perform better in terms of both R^2 and MSE