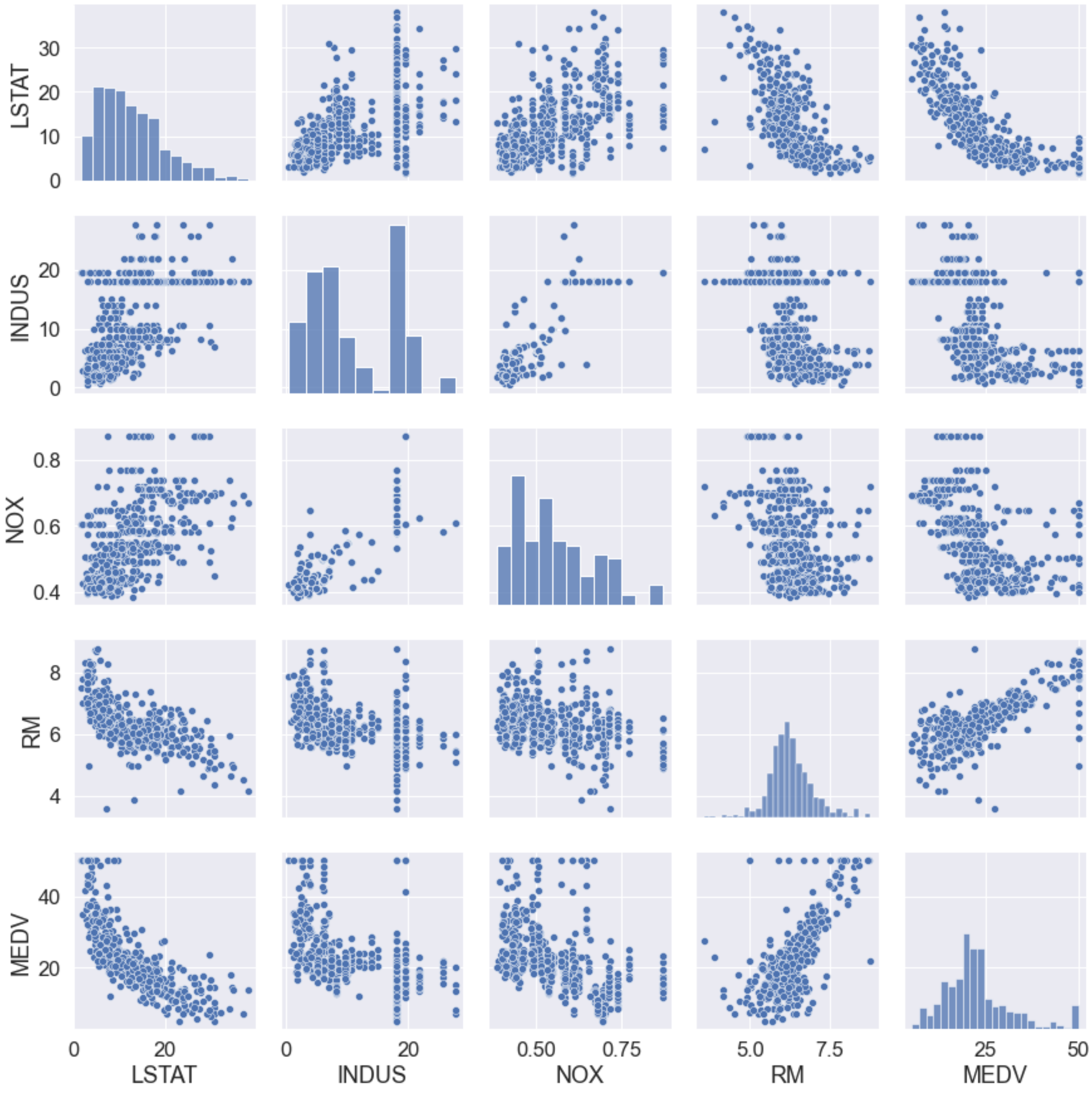
Out[1]:

Click here to toggle on/off the raw code.

Out[98]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LS
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	

Part 1: Exploratory Data Analysis



Data set size and statistic value

```
Number of Columns : 14
Number of Rows : 506
```

Attributes datatype

Out[101]:

	float	int
CRIM	506.0	0.0
ZN	506.0	0.0
INDUS	506.0	0.0
CHAS	0.0	506.0
NOX	506.0	0.0
RM	506.0	0.0
AGE	506.0	0.0
DIS	506.0	0.0
RAD	0.0	506.0
TAX	0.0	506.0
PTRATIO	506.0	0.0
B	506.0	0.0
LSTAT	506.0	0.0
MEDV	506.0	0.0

Statistical Summaries

Out[102]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
count	506.000000	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	68.574901
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000

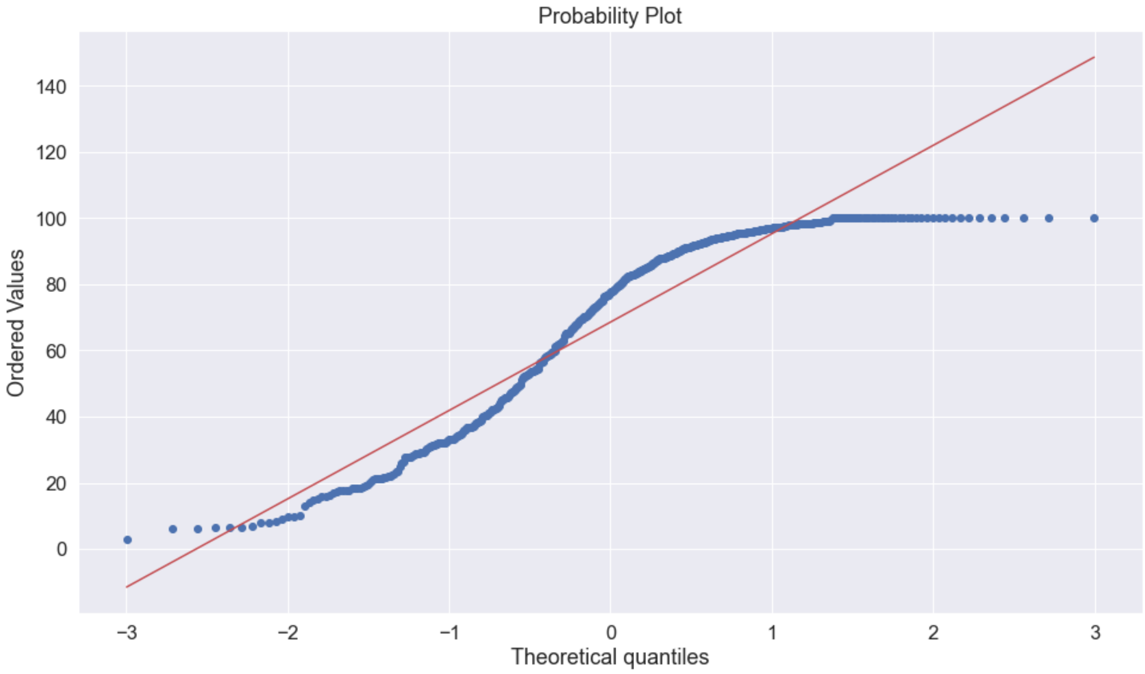
More Statistical Summaries

Out[103]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTI
Boundary for 10 percentile	0.038195	0.0	2.91	0.0	0.427	5.5935	26.95	1.62830	3.0	233.0	
Boundary for 20 percentile	0.064170	0.0	4.39	0.0	0.442	5.8370	37.80	1.95120	4.0	273.0	
Boundary for 30 percentile	0.099245	0.0	5.96	0.0	0.472	5.9505	52.40	2.25965	4.0	289.0	
Boundary for 40 percentile	0.150380	0.0	7.38	0.0	0.507	6.0860	65.40	2.64030	5.0	307.0	
Boundary for 50 percentile	0.256510	0.0	9.69	0.0	0.538	6.2085	77.50	3.20745	5.0	330.0	
Boundary for 60 percentile	0.550070	0.0	12.83	0.0	0.575	6.3760	85.90	3.87500	5.0	398.0	
Boundary for 70 percentile	1.728440	0.0	18.10	0.0	0.605	6.5025	91.80	4.54040	8.0	437.0	
Boundary for 80 percentile	5.581070	20.0	18.10	0.0	0.668	6.7500	95.60	5.61500	24.0	666.0	
Boundary for 90 percentile	10.753000	42.5	19.58	0.0	0.713	7.1515	98.80	6.81660	24.0	666.0	
Boundary for 100 percentile	88.976200	100.0	27.74	1.0	0.871	8.7800	100.00	12.12650	24.0	711.0	

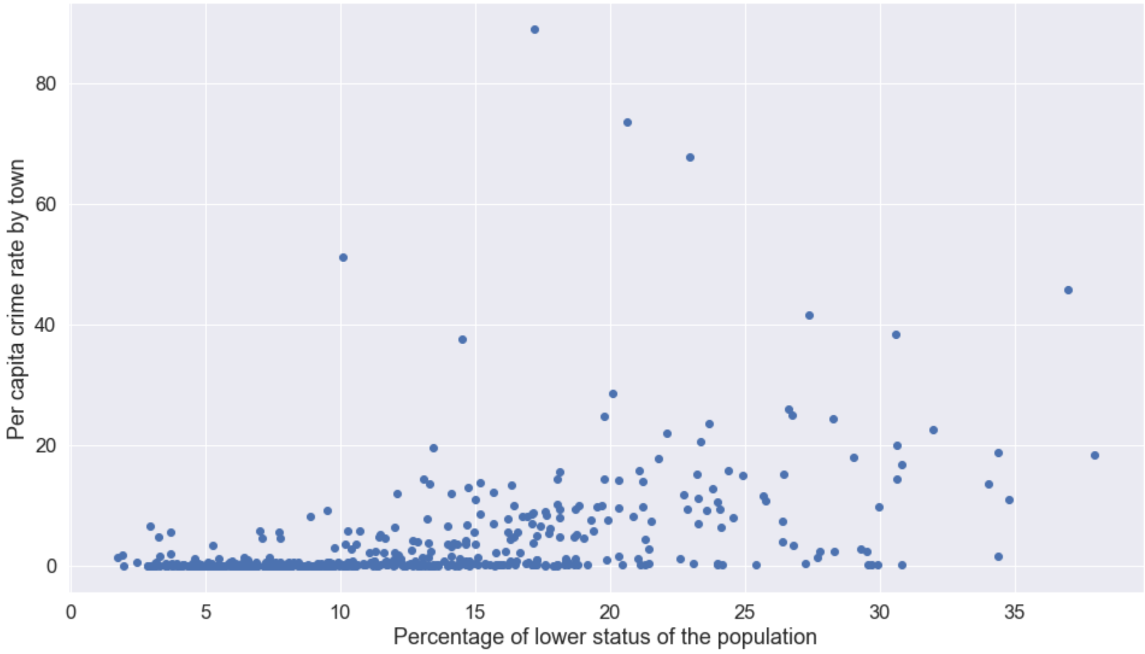
Q-Q plot

showing if the data is gaussian

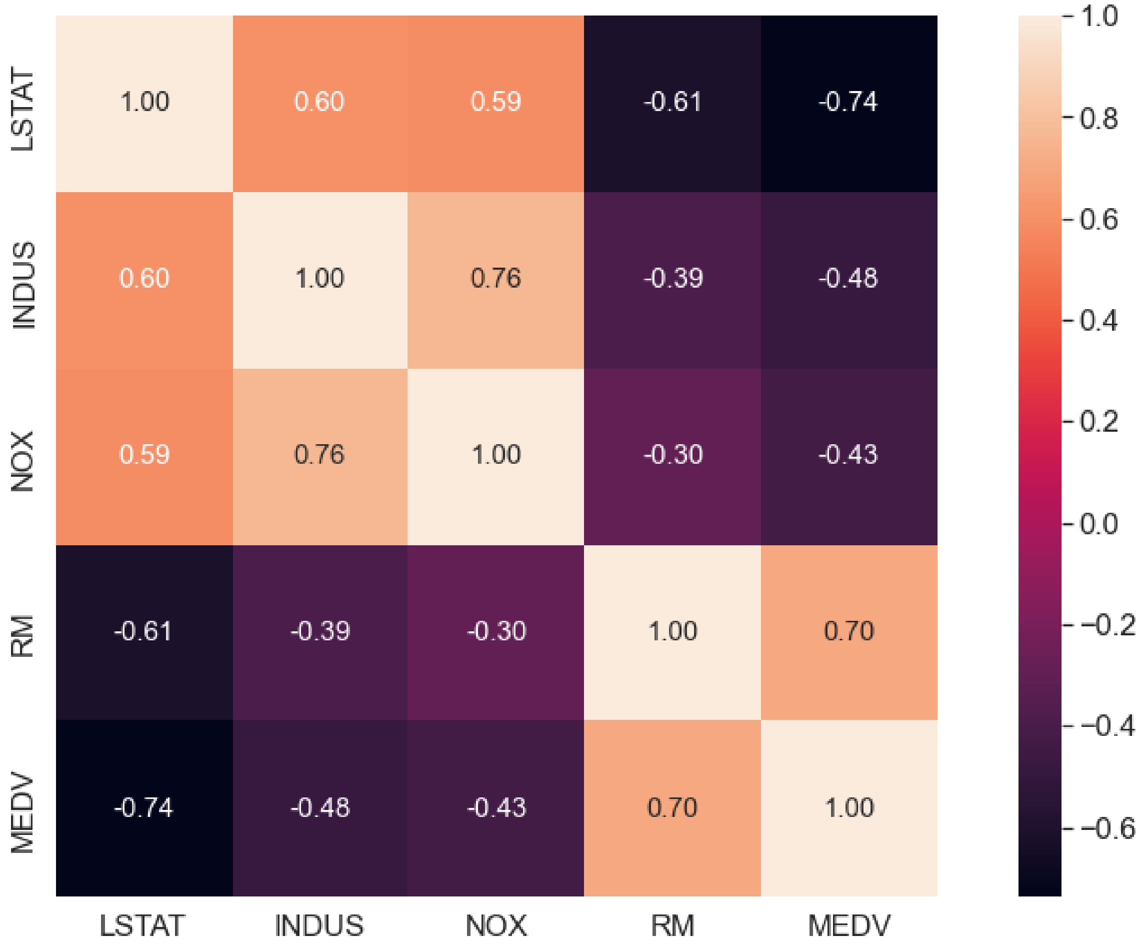


```
AGE does not look Gaussian (reject H0)
```

Cross ploting LSTAT - CRIM



Heat map plot



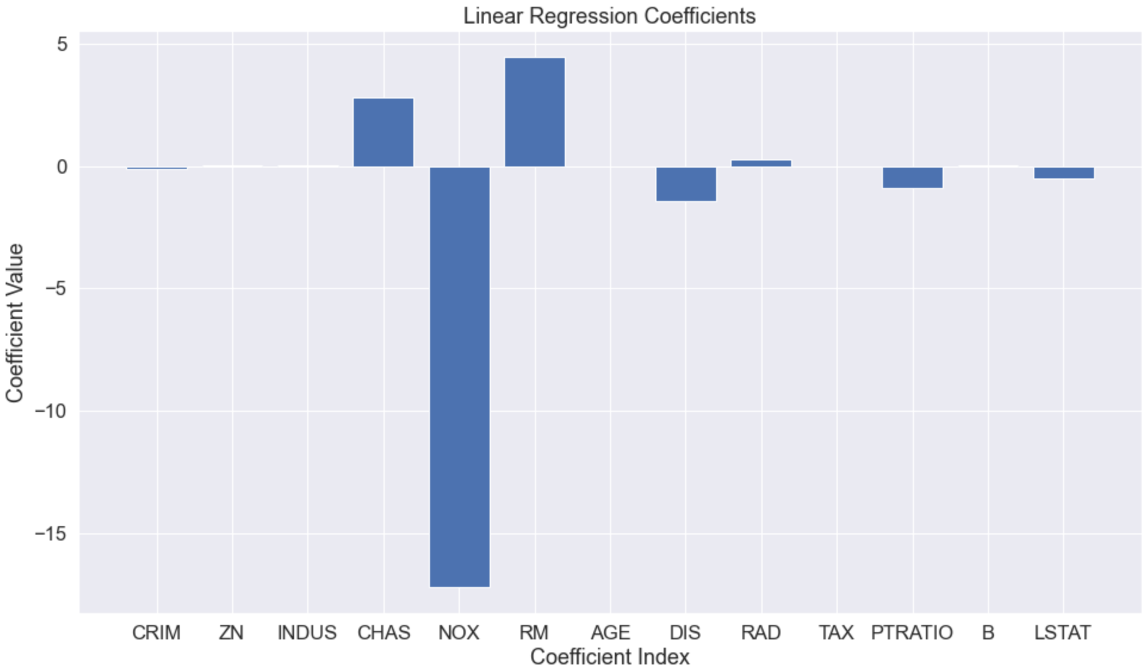
Create Train test split for the next part

Use random_state = 42. Use 80% of the data for the training set. Use the same split for all models.

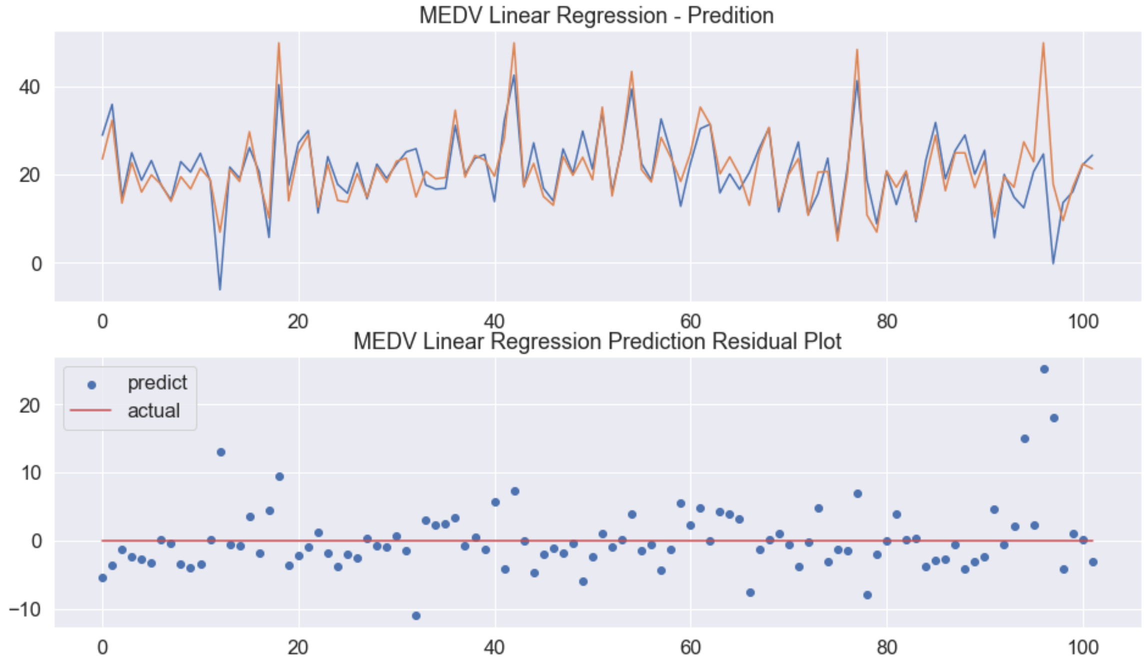
Part 2: Linear regression

train and fit LinearRegression model

```
Intercept:  30.24675099392408
MSE : 24.29111947497371
R-Squared : 0.6333247469014311
```



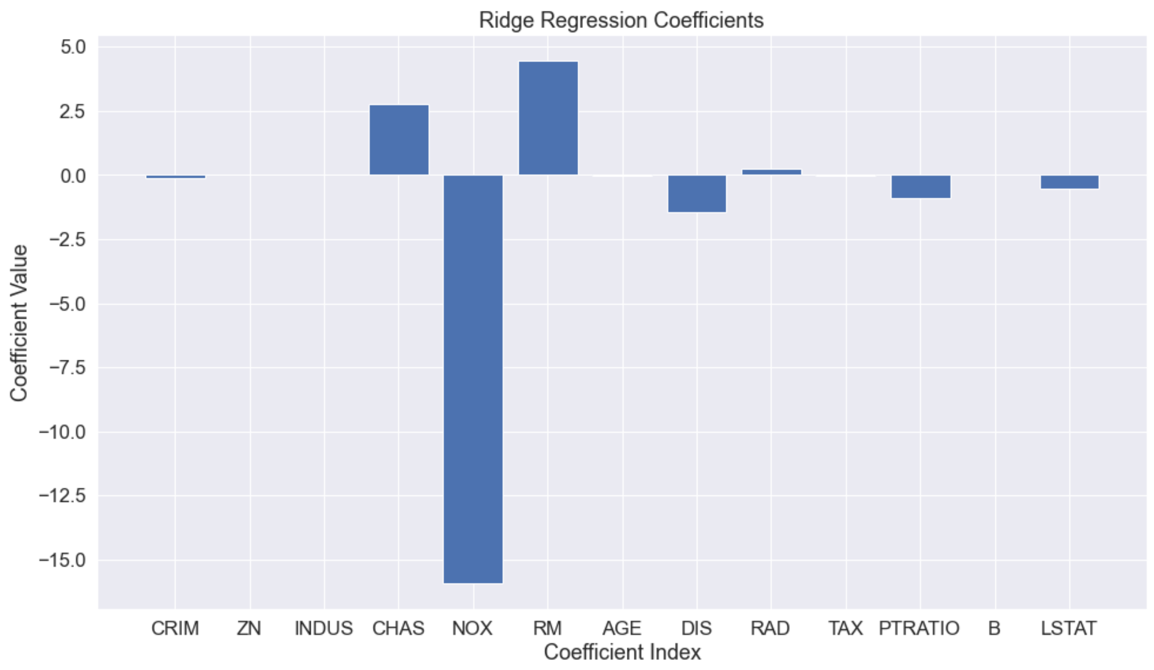
Ploting prediction



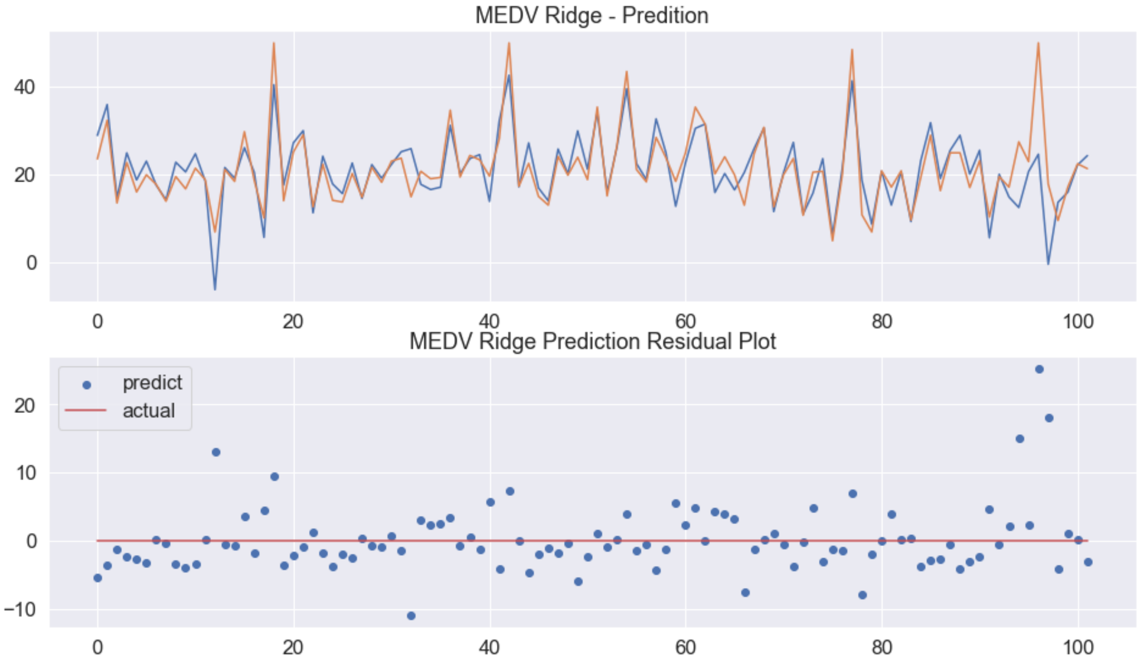
Part 3.1: Ridge regression

finding best params using Gridsearch

```
Best params:  {'alpha': 0.1}
Best Estimator:  Ridge(alpha=0.1)
Intercept:  29.366271272576704
MSE : 24.301025500192758
R-Squared : 0.63326467382235
```



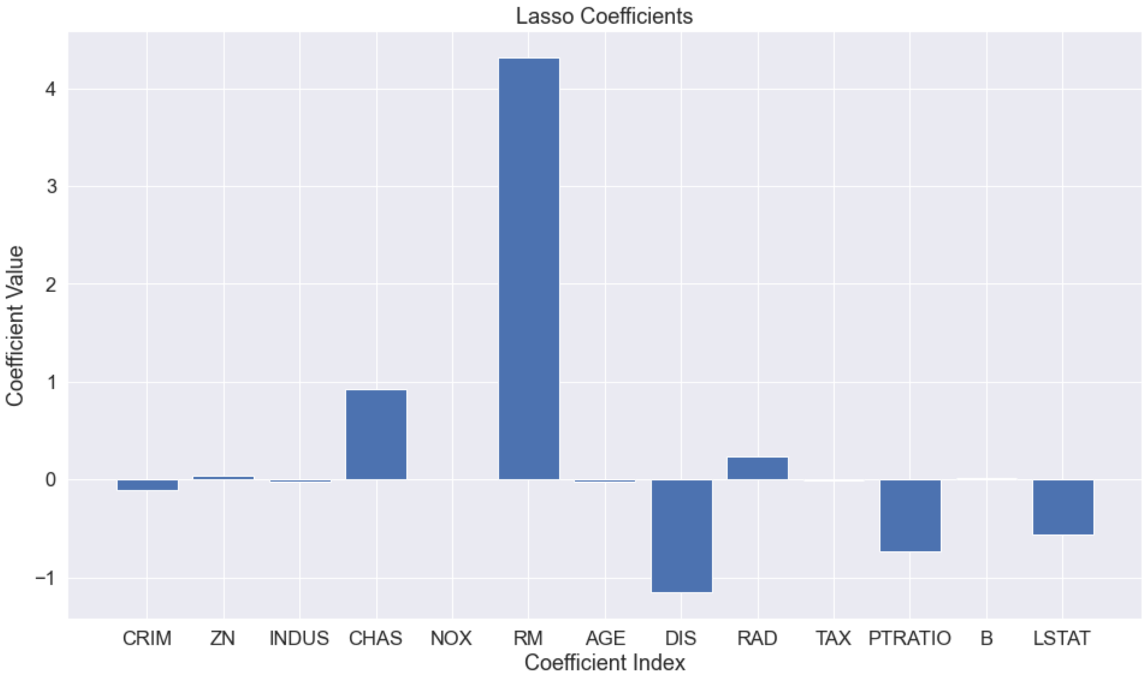
Plot Ridge Prediction and Residual using subplot

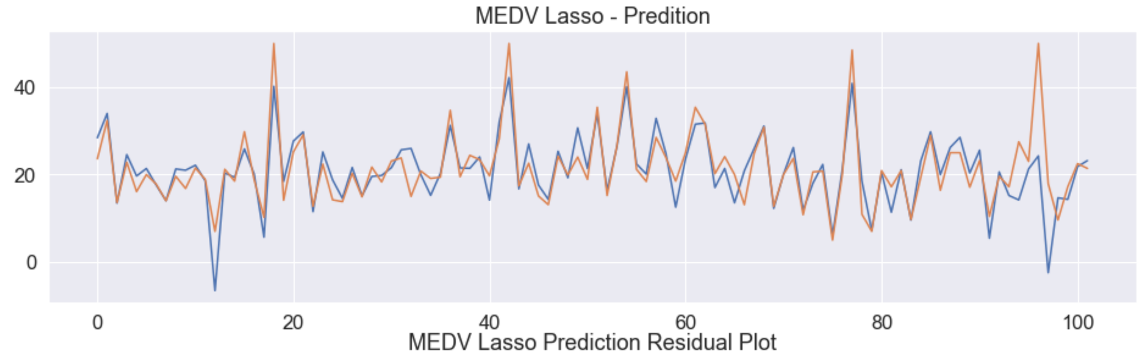


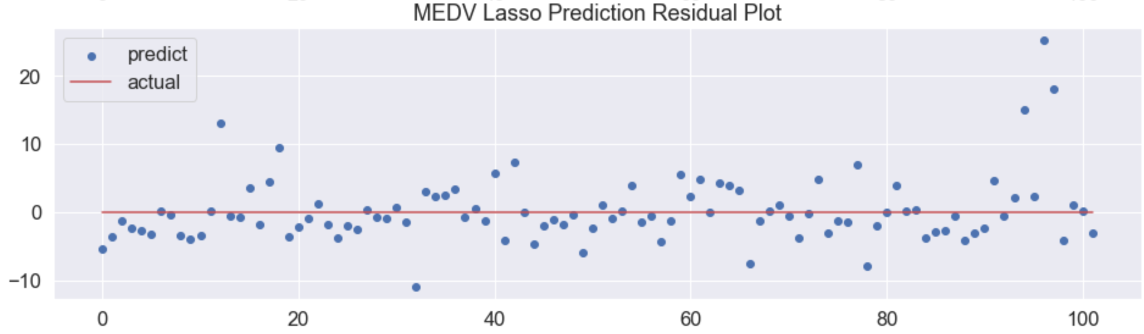
Part 3.2: LASSO regression

using gird search to find best lasso model

```
Best params:  {'alpha': 0.1}
Best Estimator:  Lasso(alpha=0.1)
Intercept:  19.859769480417444
MSE : 25.155593753934173
R-Squared : 0.6201889701292777
```







Part 4: Conclusions

From the results of the linear regression models, it is evident that Ridge and Lasso have different regularization effects on the model coefficients. By analyzing the coefficients plot, we can observe the impact of the regularization on the feature "NOX". In the plain linear regression (LR) model, there is a substantial negative coefficient for the "NOX" feature, while in the Ridge model, the coefficient value is smaller, and in the Lasso model, the coefficient is effectively zero. This reduction in the magnitude of the "NOX" coefficient in Ridge and Lasso models helps to reduce the model's dependence on a single feature, reducing the risk of overfitting. However, this comes at a cost of reduced accuracy, with a slight decrease in accuracy in the Ridge model and a significant drop in accuracy in the Lasso model.

Part 5: Appendix

```
My name is Saranpat Prasertthum
My NetID is: 655667271
I hereby certify that I have read the University policy on Academic Integr
ity and that I am not in violation.
```