Boston Housing

Author: ZHAO Cheng

Date: 2024/09/06

Data Analysis Report

Data Exploration

Data Size

Total number of entries: 514

Total number of columns: 14

Total number of data points: 7196

Total number of missing values: 10

Percentage of missing values: 0.13896609227348528

Random Sample of Data

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
158	3.54	0.0	19.58	1	0.87	6.15	82.6	1.75	5.0	403.0	14.7	88.01	15.02	15.6
118	0.13	0.0	10.01	0	0.55	6.18	72.5	2.73	6.0	432.0	17.8	393.30	12.04	21.2
447	22.05	0.0	18.10	0	0.74	5.82	92.4	1.87	24.0	666.0	20.2	391.45	22.11	10.5
108	0.17	0.0	8.56	0	0.52	5.84	91.9	2.21	5.0	384.0	20.9	395.67	18.66	19.5
184	0.07	0.0	2.46	0	0.49	6.14	62.2	2.60	3.0	193.0	17.8	396.90	9.45	36.2

Data Statistics

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	514.00	512.00	514.00	514.00	514.00	514.00	514.00	514.00	509.00	511.00	514.00	514.00	514.00	514.00
mean	3.58	11.07	11.14	0.07	0.55	6.28	68.62	3.81	9.59	408.62	18.46	357.19	12.68	22.50
std	8.54	23.01	6.86	0.25	0.12	0.70	28.13	2.11	8.74	168.78	2.16	90.68	7.15	9.18
min	0.01	0.00	0.46	0.00	0.38	3.56	2.90	1.13	1.00	187.00	12.60	0.32	1.73	5.00
25%	0.08	0.00	5.19	0.00	0.45	5.88	45.02	2.10	4.00	279.00	17.40	375.61	7.04	17.02
50%	0.25	0.00	9.69	0.00	0.54	6.20	77.50	3.22	5.00	330.00	19.00	391.48	11.40	21.10
75%	3.65	12.50	18.10	0.00	0.62	6.62	94.10	5.21	24.00	666.00	20.20	396.24	16.96	25.00
max	88.98	100.00	27.74	1.00	0.87	8.78	100.00	12.13	24.00	711.00	22.00	396.90	37.97	50.00

Data Analysis Report

Data Preprocessing

Remove Duplicates

Number of duplicate rows: 8

Duplicates removed.

Handle Missing Values

Rows with missing values:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
22	0.85204	0.0	8.14	0	0.538	5.965	89.2	4.0123	NaN	307.0	21.0	392.53	13.83	19.6
33	1.38799	0.0	8.14	0	0.538	5.950	82.0	3.9900	4.0	NaN	21.0	232.60	27.71	13.2
62	0.17171	NaN	5.13	0	0.453	5.966	93.4	6.8185	8.0	NaN	19.7	378.08	14.44	16.0
100	0.08187	0.0	2.89	0	0.445	7.820	36.9	3.4952	NaN	276.0	18.0	393.53	3.57	43.8
114	0.12329	0.0	10.01	0	0.547	5.913	92.9	2.3534	6.0	NaN	17.8	394.95	16.21	18.8
150	2.36862	0.0	19.58	0	0.871	4.926	95.7	1.4608	NaN	403.0	14.7	391.71	29.53	14.6
193	0.09068	45.0	3.44	0	0.437	6.951	21.5	6.4798	NaN	398.0	15.2	377.68	5.10	37.0
200	0.04011	NaN	1.52	0	0.404	7.287	34.1	7.3090	2.0	329.0	12.6	396.90	4.08	33.3
238	0.44791	0.0	6.20	1	0.507	6.726	66.5	3.6519	NaN	307.0	17.4	360.20	8.05	29.0

Percentage of rows with missing values:1.78%

Rows with at least 2 missing values removed.

Missing values handled using front method.

Negative values removed.

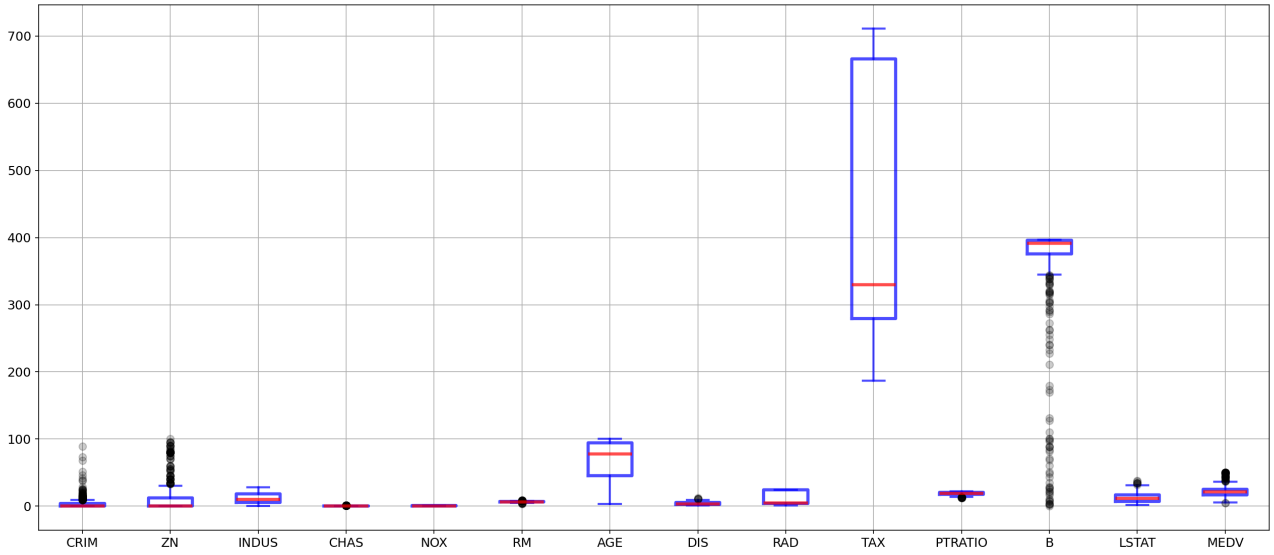
Handle Outliers

Negative values removed.

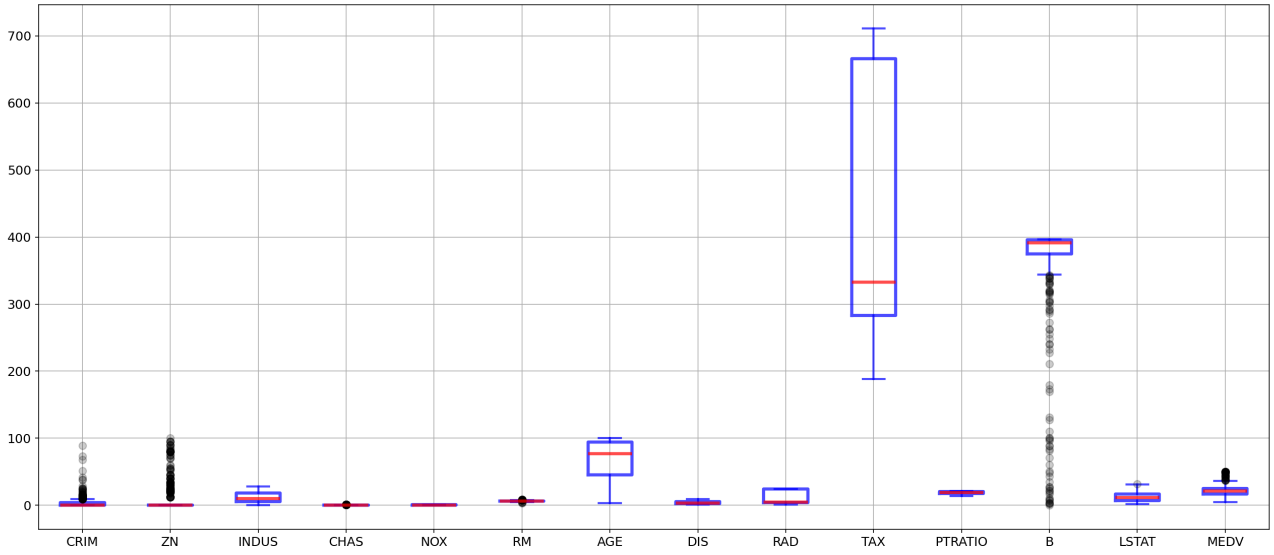
Outliers removed using IQR method.

Data Analysis Report

Before:



After:

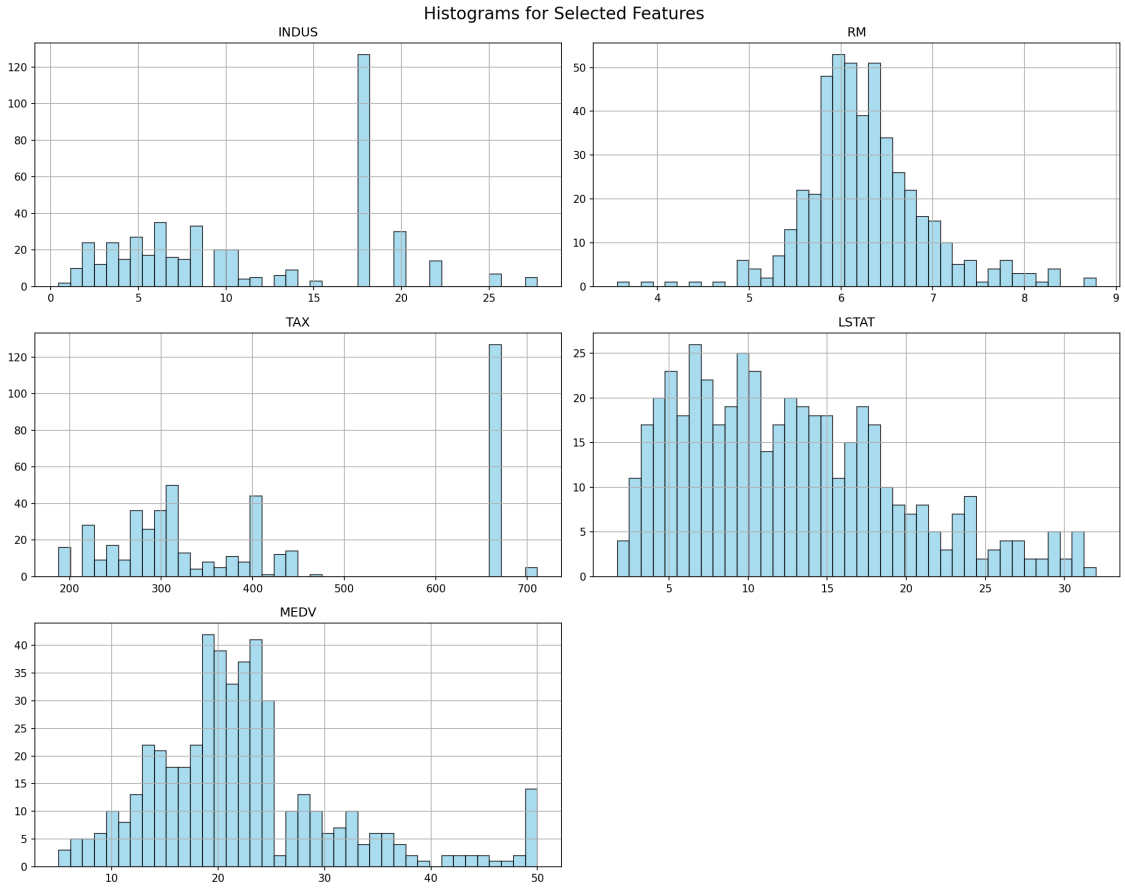


Data Analysis Report

Feature Engineering

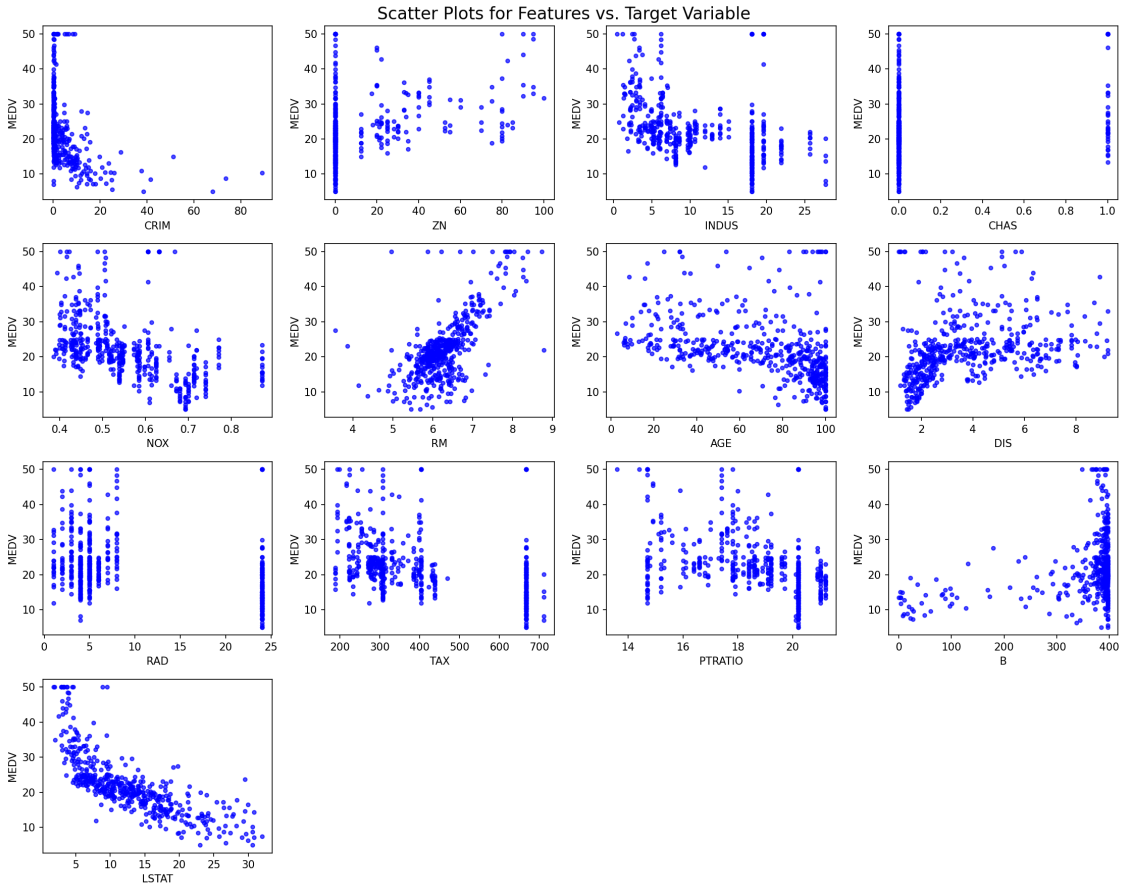
Visualize selected features.

Feature Histograms



Data Analysis Report

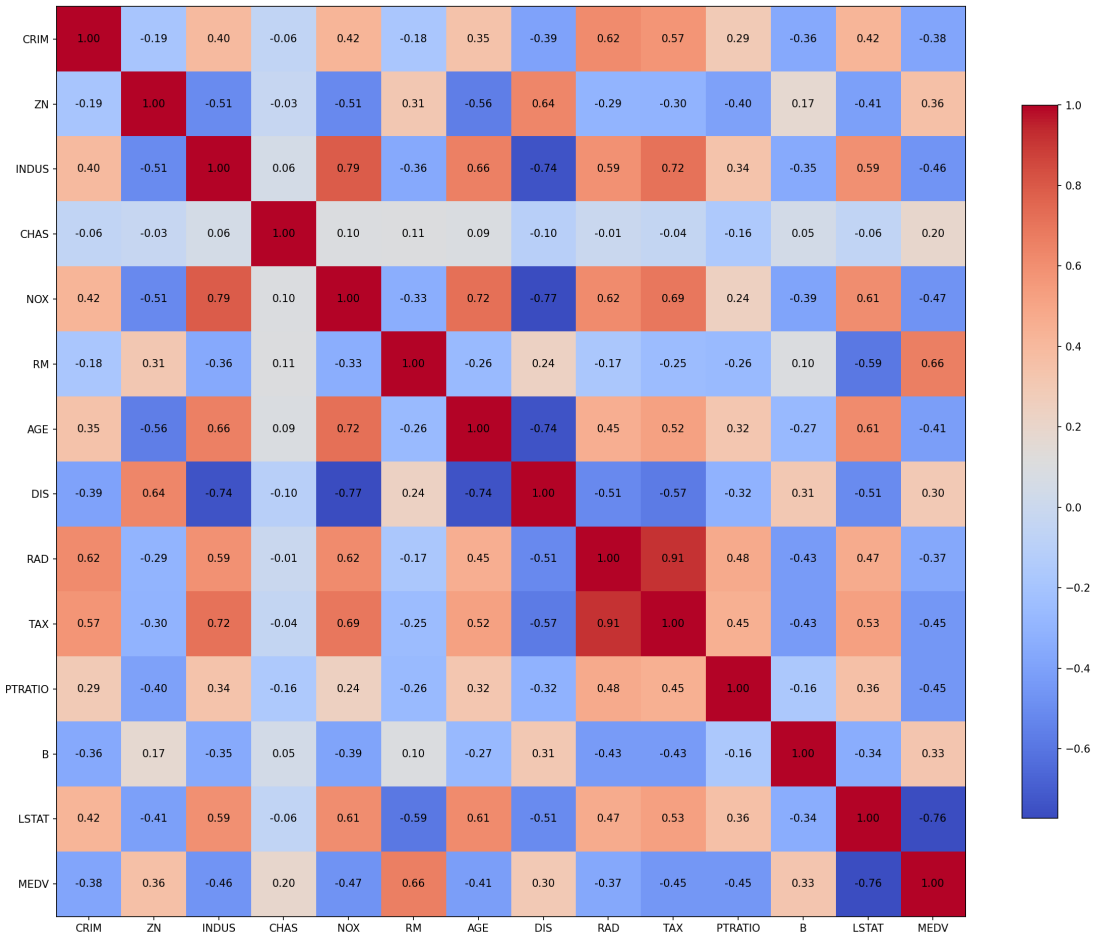
Feature Scatter Plots



Data Analysis Report

Select Features

Correlation Matrix



Select features based on correlation threshold.

Correlation threshold: 0.4

Selected Features: ['INDUS', 'NOX', 'RM', 'AGE', 'TAX', 'PTRATIO', 'LSTAT', 'MEDV']

Data Analysis Report

Model Building and Evaluation

Train Model

Training data size: 384 samples.

Testing data size: 96 samples.

Problem Type: regression

Method: decision_tree

Model: DecisionTreeRegressor()

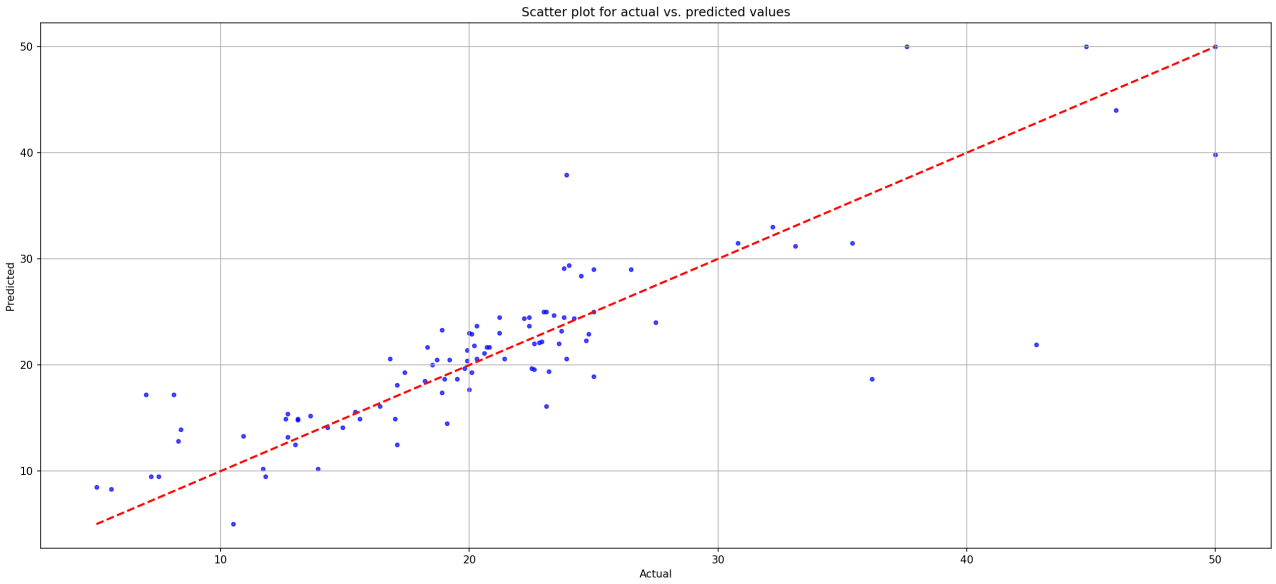
Model Evaluation

Mean Squared Error (MSE): 21.12

R^2 Score: 0.73

Data Analysis Report

Scatter plot for actual vs. predicted values



Line plot for actual vs. predicted values