Boston House Price Prediction

Presented by:
Julius Hermanto

Workflow

Objectives

Data Understanding

Descriptive Statistics

Exploratory Data Analysis

Data Preparation

Modeling

Evaluation

Conclusion

Objectives

01 Identify the variables that influence increasing house prices through exploratory data analysis.

02 Predict house prices using machine learning models.

Data Understanding

Data Understanding

- Dataset has 506 rows of data with 14 columns:
  - 13 columns are numerical data
  - 1 column is categorical data

- Based on the dataset source, variable **MEDV** is suggested to be filtered due to data censoring (50.00)

Variables	Description
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

Descriptive Statistics

Descriptive Statistics

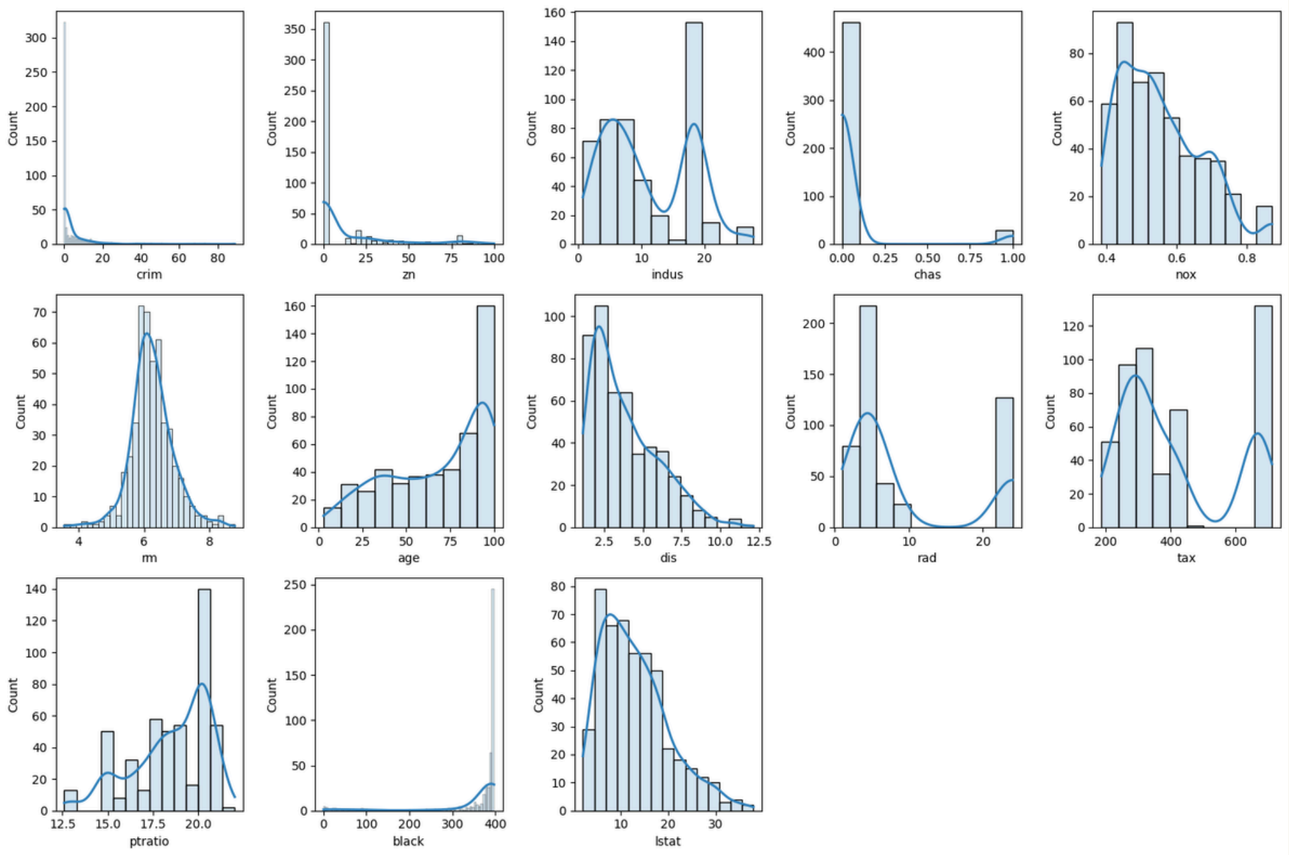
Based on categorical data, most houses sold or on sale do not have tract bounds to Charles River.

Based on numerical data here are as follows:

- Median crime rate: **0.256**
- Median proportion of non-retail businesses per town: **9.69**
- Median nitrous oxide concentration (per 10 million): **0.53**
- Median average number of rooms per house: **6 rooms**
- Median house age: **77 years old**
- Median weighted distance to five Boston employment centers: **3 miles**
- Median accessibility to radial highways: **5 highways**
- Median tax rates (per $10,000): **330**
- Median Pupil-Teacher ratio: **19**
- Median percentage of lower status population: **11,36%**
- Median value of owner-occupied homes: **$21,200**

Exploratory Data Analysis

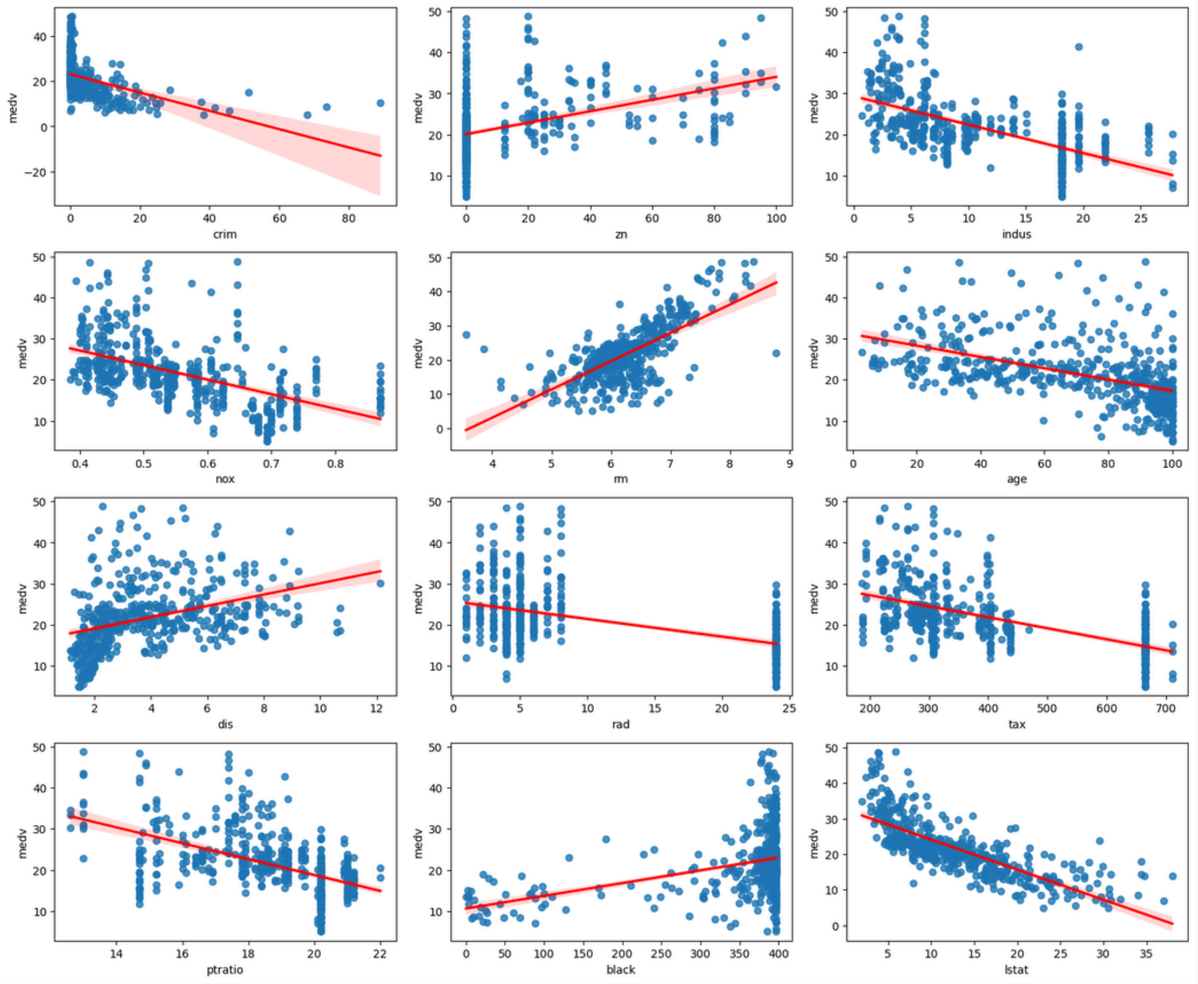
Univariate Analysis



- **CHAS** is a categorical value since it has discrete values of 0 and 1.
- **CRIM**, **ZN**, and **B** have extreme values/outliers that skew the most.
- We have a mixture of data skewness indicating some not correlating with the target variable.
- **RAD** and **TAX** have bimodal skew indicating a clear gap in the data.

Bivariate Analysis

- **CHAS** is a categorical value since it has discrete values of 0 and 1.
- **CRIM**, **ZN**, and **B** have extreme values/outliers that skew the most.
- We have a mixture of data skewness indicating some not correlating with the target variable.
- **RAD** and **TAX** have bimodal skew indicating a clear gap in the data.



Bivariate Analysis

Based on individual feature correlation to the target, the summary is as follows:

- **Environmental Factors**: Lower crime rates, NOx concentrations, and newer homes are linked to higher house values.
- **Infrastructure and Accessibility**: Greater distance from employment centers and higher accessibility to radial highways tend to lower house values.
- **Economic and Social Factors**: Higher industrial proportion, property tax rates, and pupil-teacher ratios negatively impact house values, while more rooms in a house generally increase its value.
- **Demographic Influence**: A higher proportion of lower-status individuals decreases house values, while the proportion of Black residents shows a weak positive correlation due to outliers.

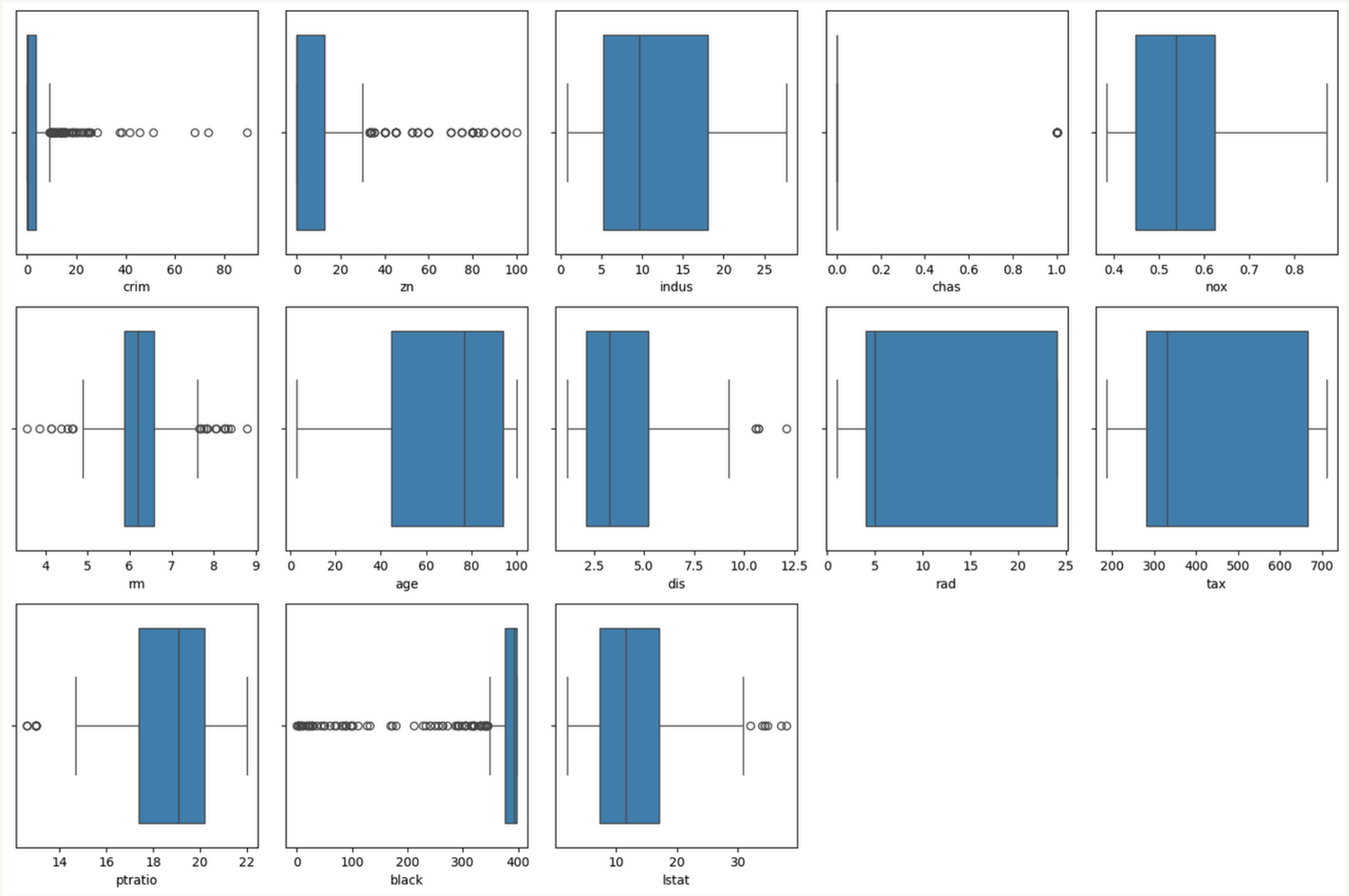
Data Preparation

Outlier Handling

Based on our data understanding earlier, we remove data with **MEDV** value of >= 50.00

```
df = df[~(df['medv'] >= 50.0)]
```

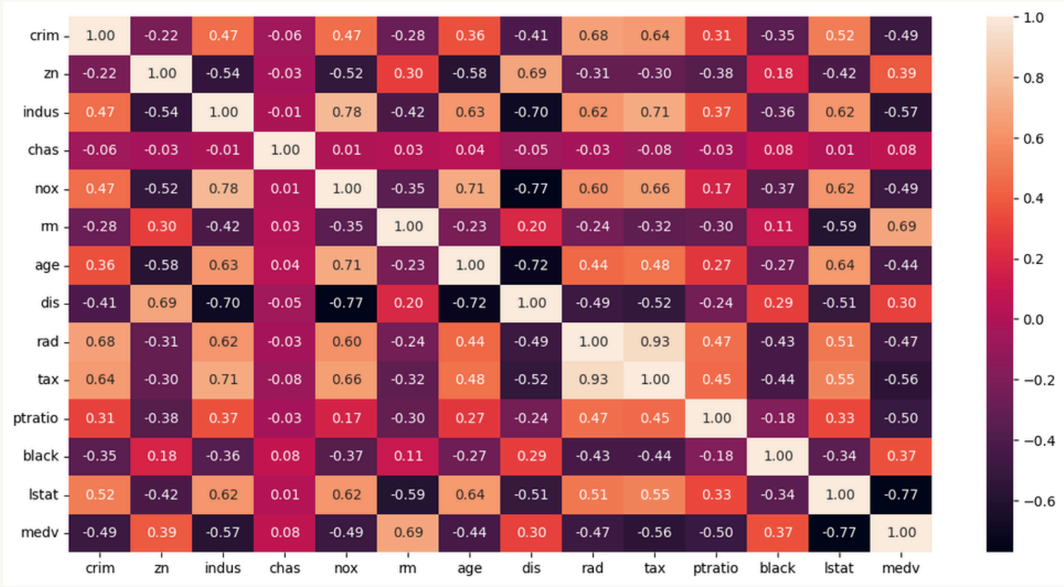
There seems to be outliers but due to the fact we only have 490 data after removing data, we do not handle outliers.



Multicollinearity Check

- **NOX**, **DIS**, **RAD**, and **TAX** have high VIF score > 4 (correlation to other independent features)
- A correlation matrix is generated, and it is found that **RAD** and **TAX** have a high correlation (> 0.8).
- We drop **TAX** since it has a lower correlation to the target variable out of the two.

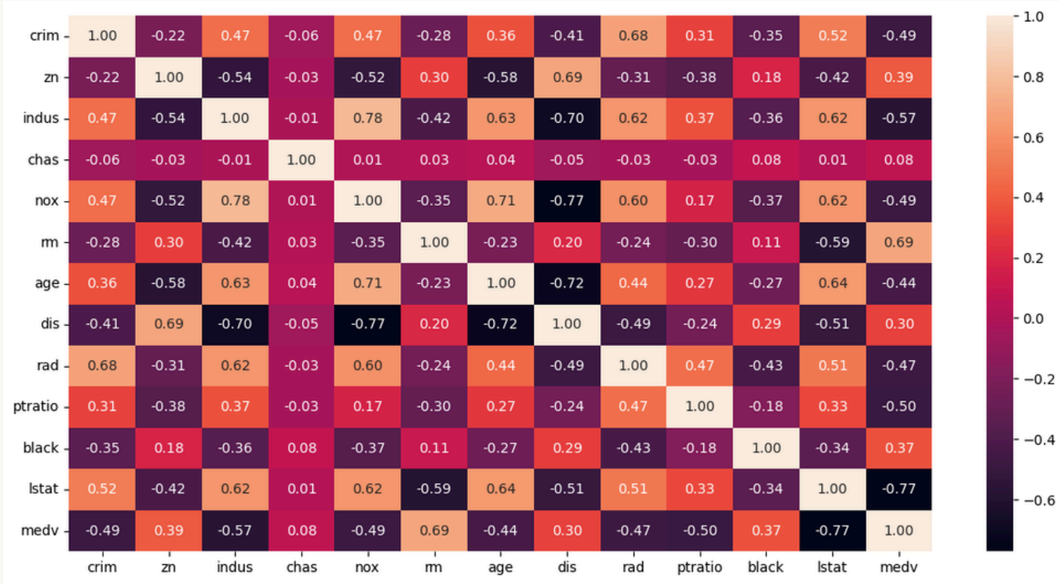
	feature	vif_score
1	crim	2.079488
2	zn	2.418546
3	indus	3.811354
4	chas	1.049900
5	nox	4.654454
6	rm	1.923483
7	age	3.101301
8	dis	4.071004
9	rad	8.739623
10	tax	10.040523
11	ptratio	1.721366
12	black	1.301664
13	lstat	3.108848



Multicollinearity Check (2)

- After checking the VIF score a second time, we found that **NOX** and **DIS** still have scores above 4.
- Generating a second correlation matrix found that **INDUS** and **NOX** are highly correlated. Even though the correlation score is not > 0.8 we still need to drop to lower the VIF score.
- We drop **INDUS** since it has a lower correlation to the target variable out of the two.

	feature	vif_score
1	crim	2.078939
2	zn	2.284336
3	indus	3.428515
4	chas	1.022502
5	nox	4.587722
6	rm	1.907334
7	age	3.101228
8	dis	4.068802
9	rad	3.021291
10	ptratio	1.708722
11	black	1.299347
12	lstat	3.108655



Multicollinearity Check (3)

	feature	vif_score
1	crim	2.062715
2	zn	2.281093
3	chas	1.020817
4	rm	1.744360
5	age	2.854964
6	dis	3.098843
7	rad	2.528489
8	ptratio	1.500509
9	black	1.283745
10	lstat	3.070016

A final check of the VIF scores shows us there is no high correlation between independent variables anymore. (All < 4)

Modeling

Ridge Regression

```
Training model: Ridge(alpha=0.01, random_state=42)
R2: 0.6575      RMSE: 5.12899   MAE: 3.59306    MAPE: 19.13%

Training model: Ridge(alpha=0.1, random_state=42)
R2: 0.6575      RMSE: 5.12875   MAE: 3.59294    MAPE: 19.13%

Training model: Ridge(alpha=1, random_state=42)
R2: 0.6578      RMSE: 5.12674   MAE: 3.59190    MAPE: 19.11%

Training model: Ridge(alpha=10, random_state=42)
R2: 0.6575      RMSE: 5.12915   MAE: 3.59224    MAPE: 18.99%
```

	feature	coefficient
0	intercept	17.774394
1	crim	-0.127534
2	zn	0.035543
3	chas	1.483492
4	rm	3.800898
5	age	-0.017288
6	dis	-0.846056
7	rad	0.011880
8	ptratio	-0.702780
9	black	0.011536
10	lstat	-0.514617

After integrating multiple alphas to test our **Root Mean Squared Error**, we found that an alpha of **1** represents the best results of the lowest error prediction of **5.126%** for Ridge Regression.

We then implemented an alpha of 1 to calculate coefficients resulting in an intercept of **17.774**.

Meaning, without other independent variables, the median value of owner-occupied houses (**MEDV**) is $17,774

Lasso Regression

```
Training model: Lasso(alpha=0.01, random_state=42)
R2: 0.6591      RMSE: 5.11701   MAE: 3.58546    MAPE: 19.09%

Training model: Lasso(alpha=0.1, random_state=42)
R2: 0.6688      RMSE: 5.04391   MAE: 3.53026    MAPE: 18.77%

Training model: Lasso(alpha=1, random_state=42)
R2: 0.5930      RMSE: 5.59081   MAE: 3.81569    MAPE: 18.69%

Training model: Lasso(alpha=10, random_state=42)
R2: 0.5187      RMSE: 6.07996   MAE: 4.45429    MAPE: 22.78%
```

	feature	coefficient
0	intercept	19.264422
1	crim	-0.127534
2	zn	0.035543
3	chas	1.483492
4	rm	3.800898
5	age	-0.017288
6	dis	-0.846056
7	rad	0.011880
8	ptratio	-0.702780
9	black	0.011536
10	lstat	-0.514617

After integrating multiple alphas to test our **Root Mean Squared Error**, we found that an alpha of **0.1** represents the best results of the lowest error prediction of **5.043%** for Lasso Regression.

We then implemented an alpha of 1 to calculate coefficients resulting in an intercept of **19.264**.

Meaning, without other independent variables, the median value of owner-occupied houses (**MEDV**) is $19,264

Evaluation

Mean Absolute Percentage Errors

Ridge Regression

```
Training error:
R2: 0.7703      RMSE: 3.72191   MAE: 2.70116    MAPE: 13.82%

Validation error:
R2: 0.6578      RMSE: 5.12674   MAE: 3.59190    MAPE: 19.11%

Testing error:
R2: 0.6944      RMSE: 3.95087   MAE: 2.94015    MAPE: 18.10%
```

- Training error is quite high (> 10%) indicating the model may be **underfit**
- Testing error MAPE is over Training error MAPE, indicating the model is also **overfit**

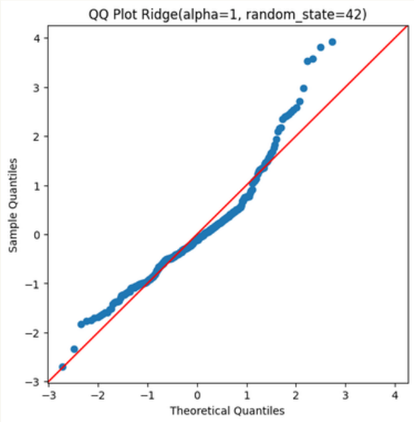
Lasso Regression

```
Training error:
R2: 0.7674      RMSE: 3.74483   MAE: 2.74087    MAPE: 13.92%

Validation error:
R2: 0.6688      RMSE: 5.04391   MAE: 3.53026    MAPE: 18.77%

Testing error:
R2: 0.6874      RMSE: 3.99614   MAE: 2.99914    MAPE: 18.29%
```

- Training error is quite high (> 10%) indicating the model may be **underfit**
- Testing error MAPE is over Training error MAPE, indicating the model is also **overfit**

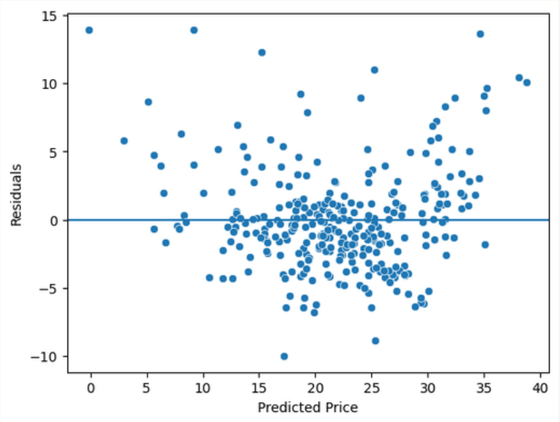
Diagnositic Study

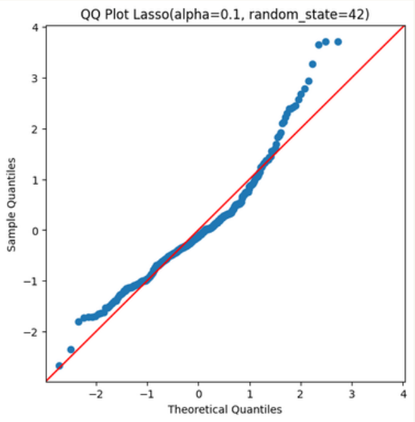
Ridge Regression



- Linearity Assumption: Invalid (U or V shape)
- Constant Variance (Homoscedasticity): Invalid (non-constant lower limit of convex)
- Uncorrelated Residuals: Valid (random cluster, no pattern)
- Normality of Residuals: Invalid (deviations from the normal line)

Lasso Regression





- Linearity Assumption: Invalid (U or V shape)
- Constant Variance (Homoscedasticity): Invalid (non-constant lower limit of convex)
- Uncorrelated Residuals: Valid (random cluster, no pattern)
- Normality of Residuals: Invalid (deviations from the normal line)

Conclusion

Conclusion

1. The dataset may be too small and may not be varied enough to describe the model. Indicating the test result indicates overfitting compared to the training result.
2. The model may not be appropriate for predicting MEDV.

Thank you!