

kelompok 10

FINAL PROJECT

date : 22/12/2023



Outlines

Key points for discussion

01
Objectives

03
Model Choices

02
EDA

04
**Best Model and
Recommendation**

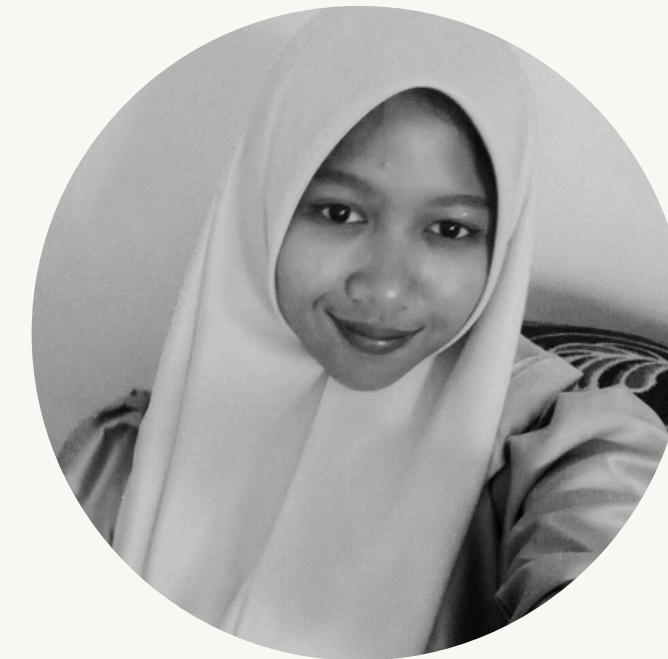
kelompok 10

We are



01

Balqis Dwian F. Z.



02

Eva Carla



03

Naila Selvira B

Analisis Model Regresi Linear pada Dataset Boston Housing: Menemukan Model Terbaik untuk Prediksi Harga Rumah

Final Project Kelompok 10



01 Problem Statement

Dalam rangka memahami faktor-faktor yang mempengaruhi harga rumah, kami memiliki dataset Housing yang berisi informasi beragam tentang rumah-rumah di suatu daerah. Tujuan kami adalah untuk menganalisis dataset ini guna mengidentifikasi korelasi antar variabel serta menemukan model prediktif yang signifikan dan akurat untuk memprediksi harga rumah. Dengan demikian, proyek ini akan membantu konsumen maupun pengusaha perumahan di Boston untuk melakukan pengembangan.

02 Objective

- Mencari korelasi antarvariabel
- Menentukan top 5 rumah dengan pajak terbesar
- Model mana yang signifikan dan bagus ?
- Bagaimana prediksi harga rumah dengan model terbaik ?





About Data

Data utama yang kami gunakan yaitu **Boston Housing Data**. Data ini menunjukkan perumahan Boston yang terdiri dari harga rumah di berbagai tempat di Boston. Dataset perumahan Boston dikumpulkan pada tahun 1978 dan masing-masing dari 506 entri mewakili data agregat tentang 14 fitur untuk rumah dari berbagai pinggiran kota di Boston, Massachusetts. Selain harga, dataset tersebut juga memberikan informasi seperti Crime (CRIM), wilayah usaha non retail di dalam kota (INDUS), umur pemilik rumah (AGE), dan masih banyak atribut lain yang dapat dilihat pada gambar sebagai berikut.

Variable	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
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in \$1000's

head : menampilkan data teratas

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	NA	36.2
6	0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7

datasetnya terdiri
dari 506 baris dan 14
kolom

tail : menampilkan data terbawah

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
	<dbl>	<dbl>	<dbl>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
501	0.22438	0	9.69	0	0.585	6.027	79.70000	2.4982	6	391	19.2	396.90	14.33000	16.8
502	0.06263	0	11.93	0	0.573	6.593	69.10000	2.4786	1	273	21.0	391.99	12.71543	22.4
503	0.04527	0	11.93	0	0.573	6.120	76.70000	2.2875	1	273	21.0	396.90	9.08000	20.6
504	0.06076	0	11.93	0	0.573	6.976	91.00000	2.1675	1	273	21.0	396.90	5.64000	23.9
505	0.10959	0	11.93	0	0.573	6.794	89.30000	2.3889	1	273	21.0	393.45	6.48000	22.0
506	0.04741	0	11.93	0	0.573	6.030	68.51852	2.5050	1	273	21.0	396.90	7.88000	11.9





—

EDA

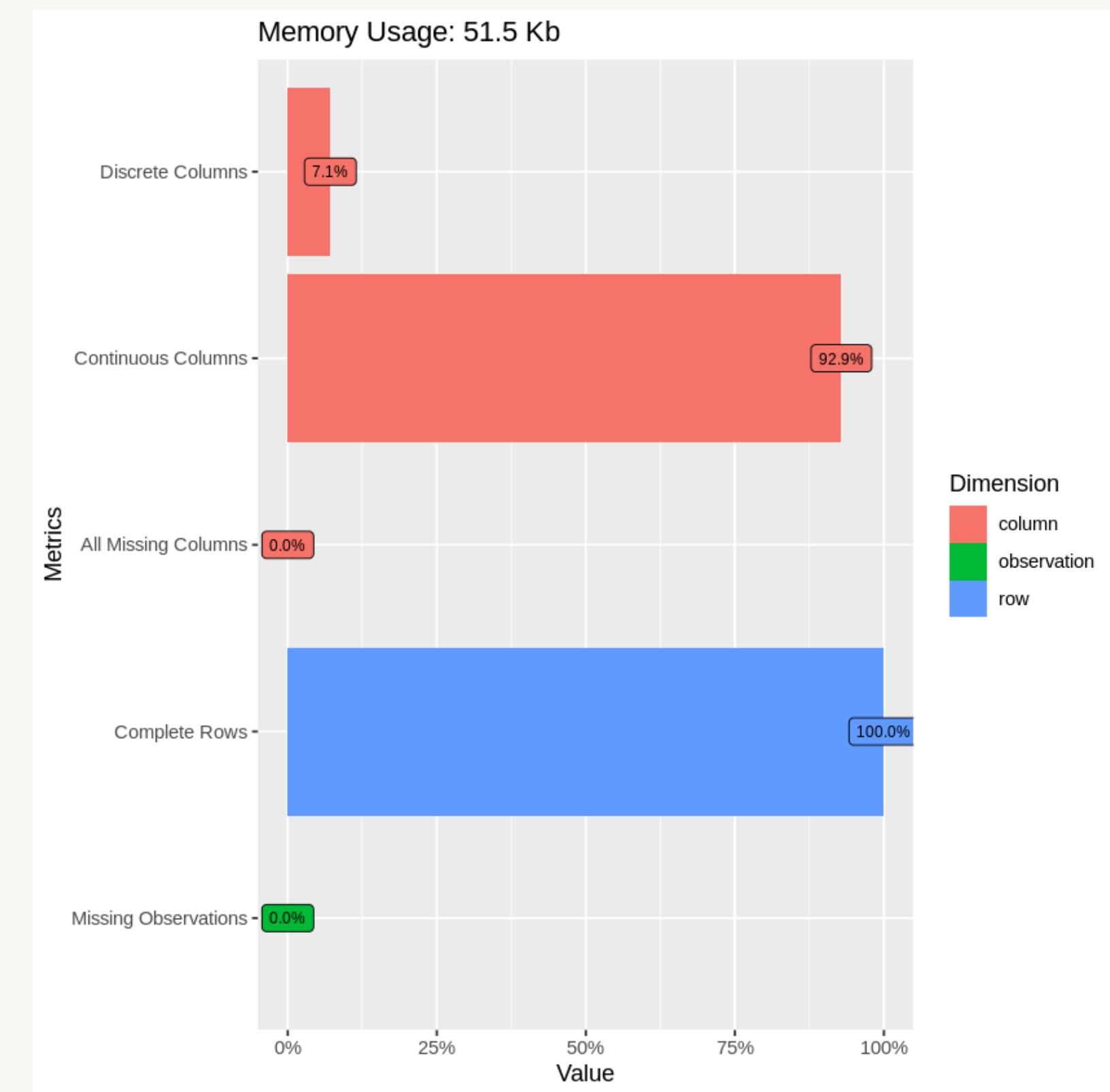
Exploratory Data Analysis

Missing Value

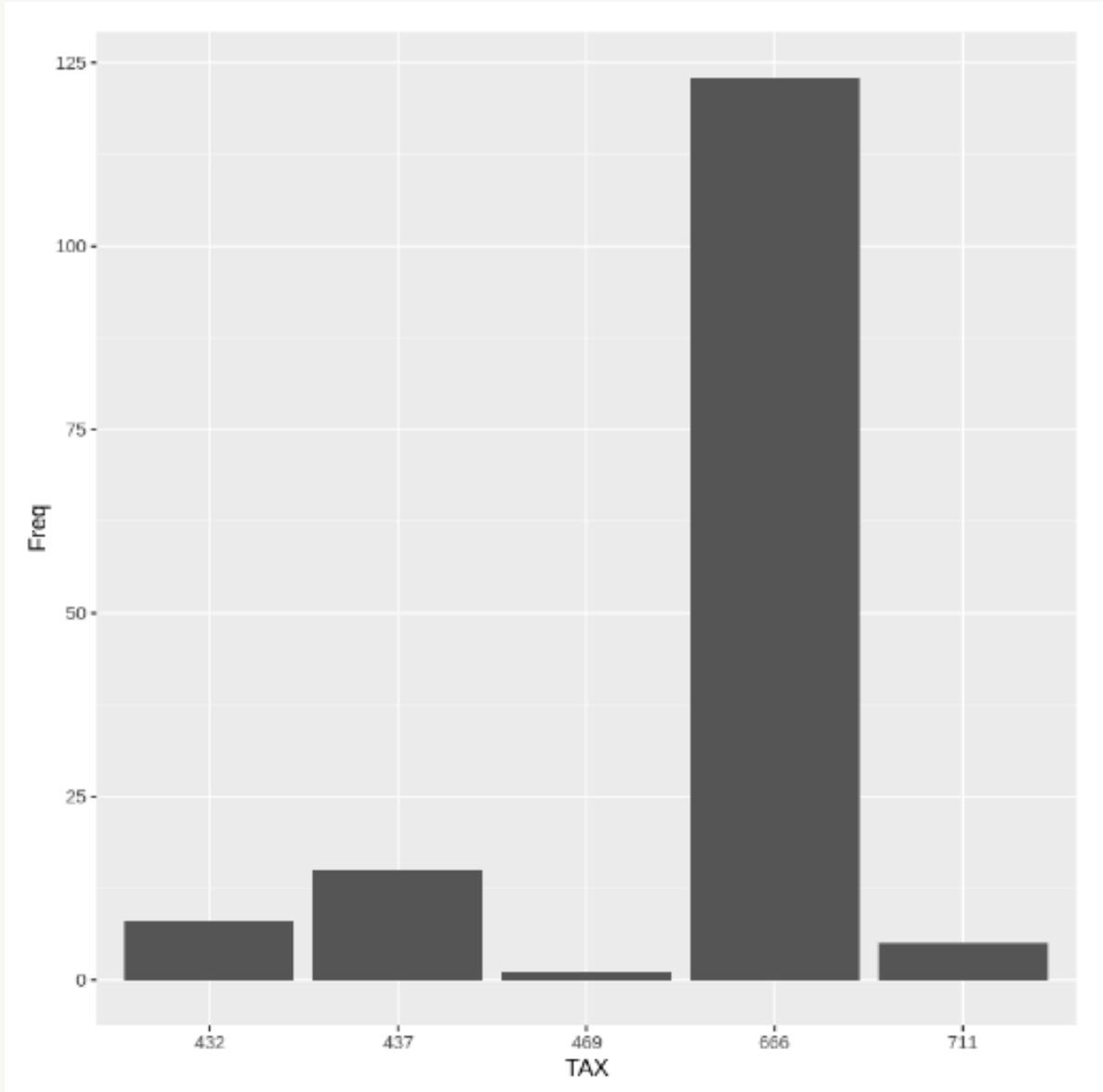
missing_values	<dbl>
CRIM	20
ZN	20
INDUS	20
CHAS	20
NOX	0
RM	0
AGE	20
DIS	0
RAD	0
TAX	0
PTRATIO	0
B	0
LSTAT	20
MEDV	0

Terlihat jelas pada gambar sebelah kiri masih banyak missing values pada dataset yang mana $> 20\%$ sehingga kami akan memasukkan nilai missing value menggunakan nilai mean.

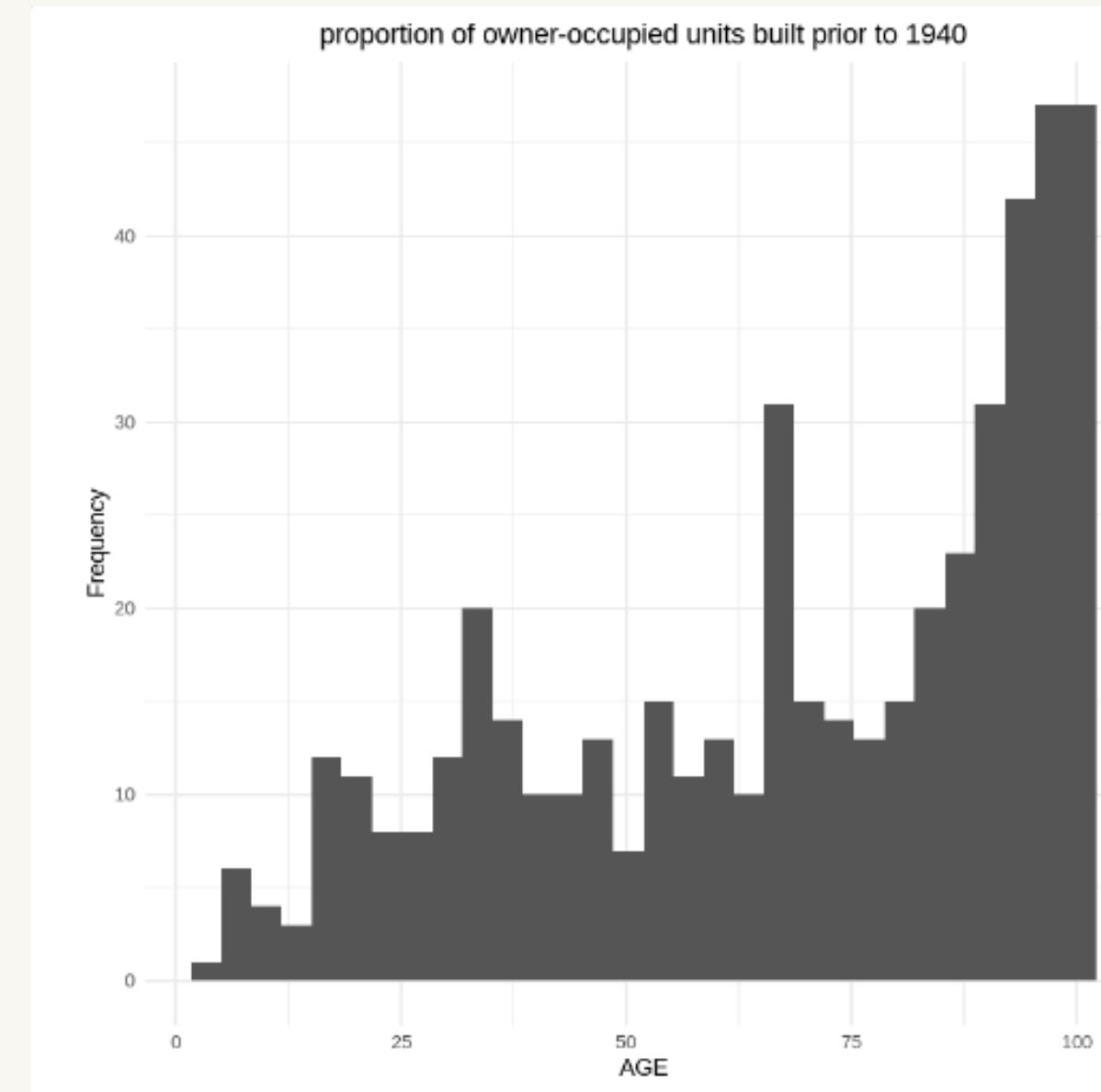
Visualisasi Metrics Value



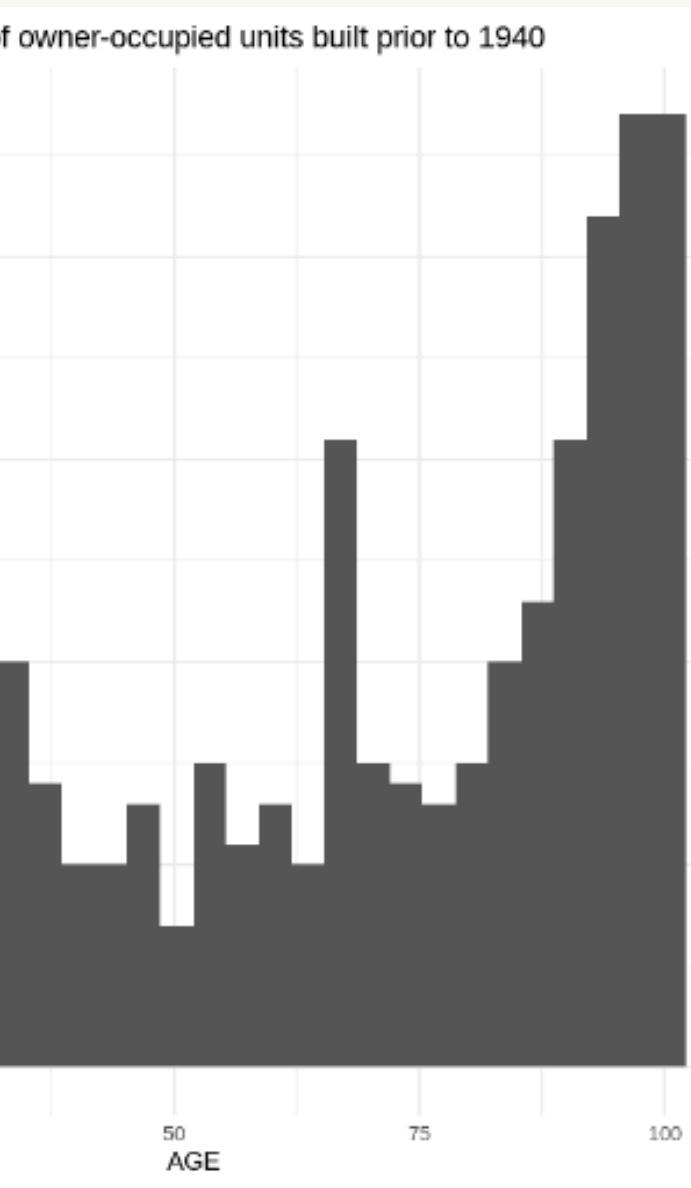
Data Insights Through Exploration



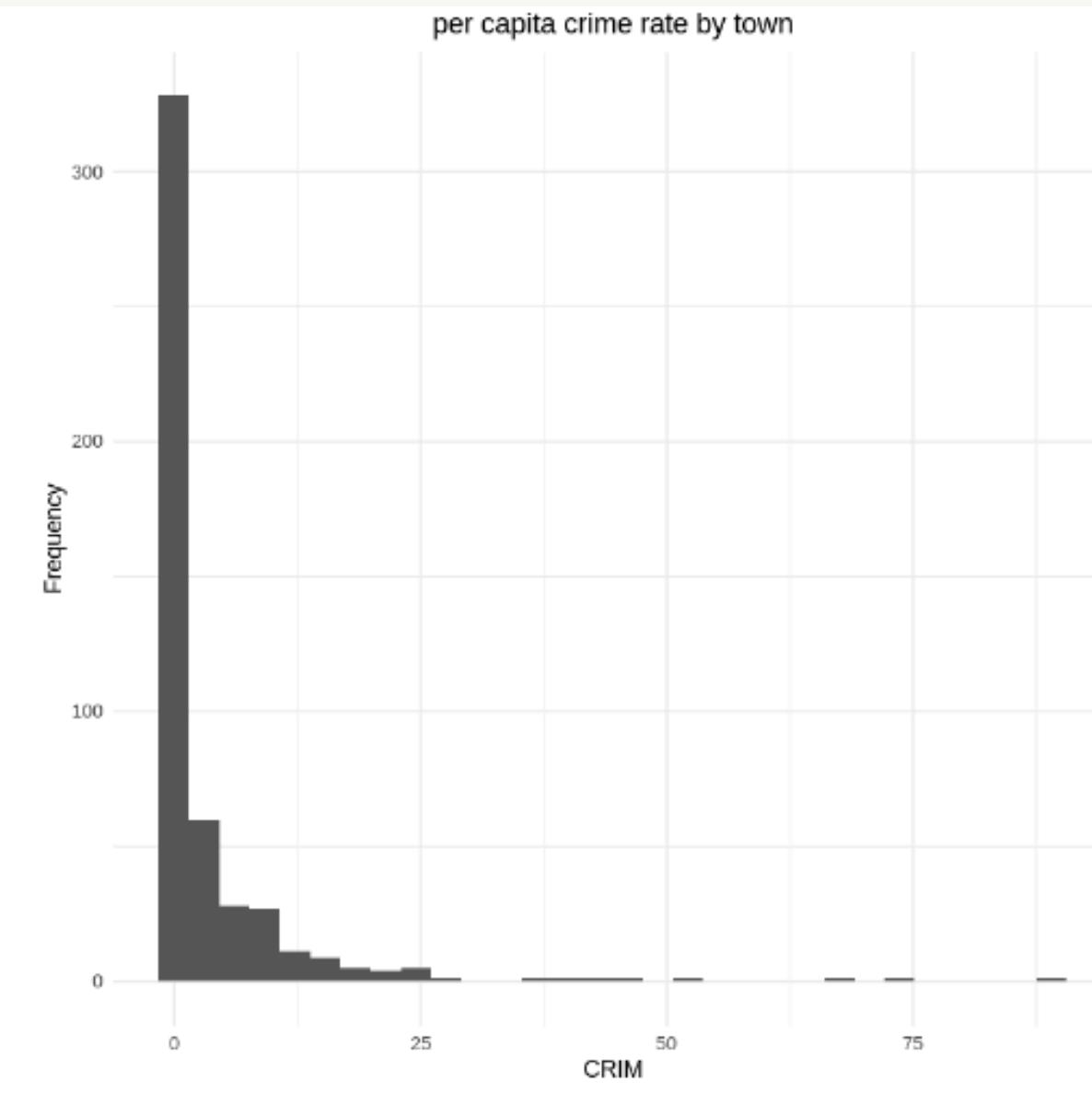
rumah dengan 5 TAX tertinggi



frekuensi
rumah



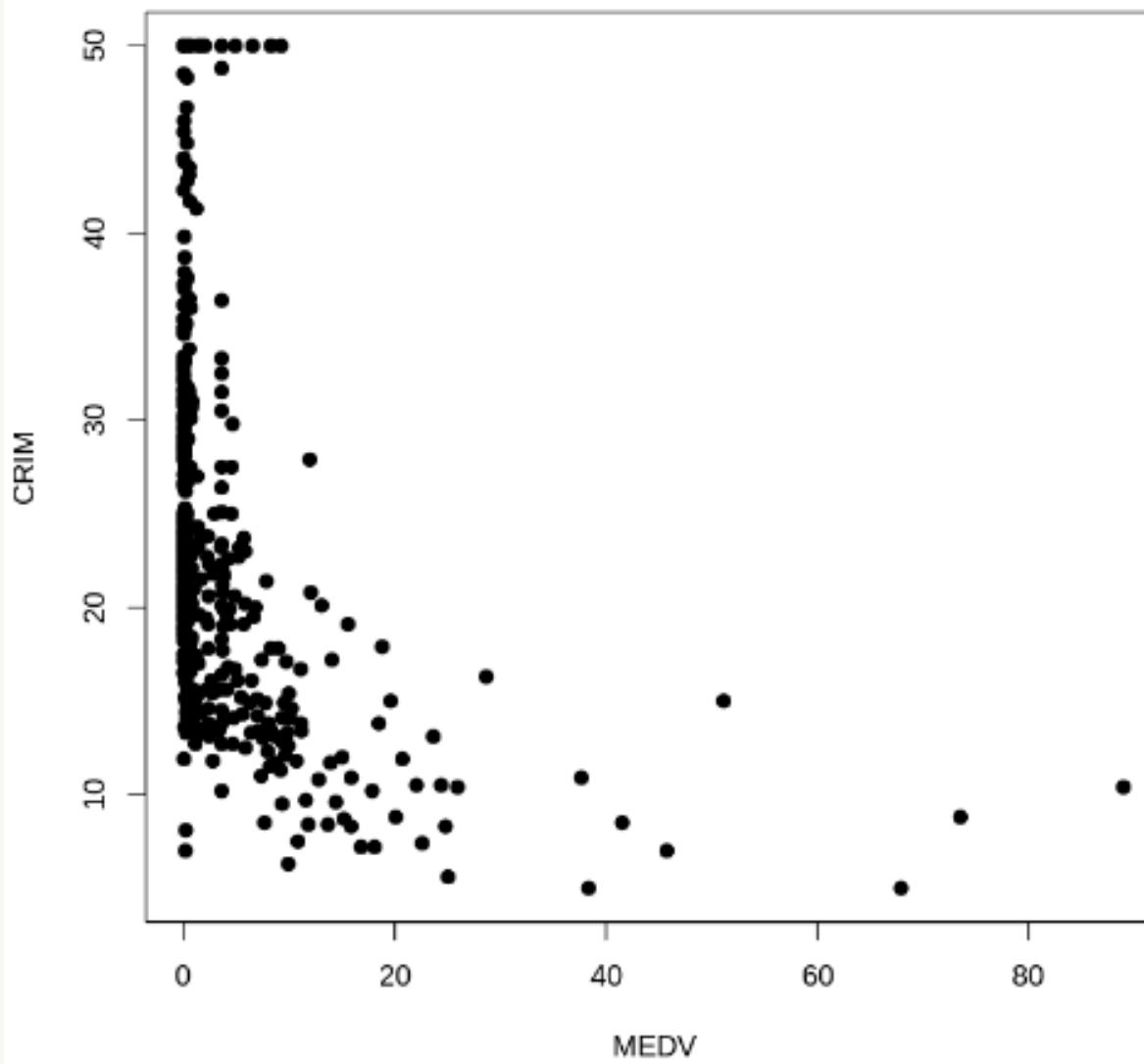
umur pemilik



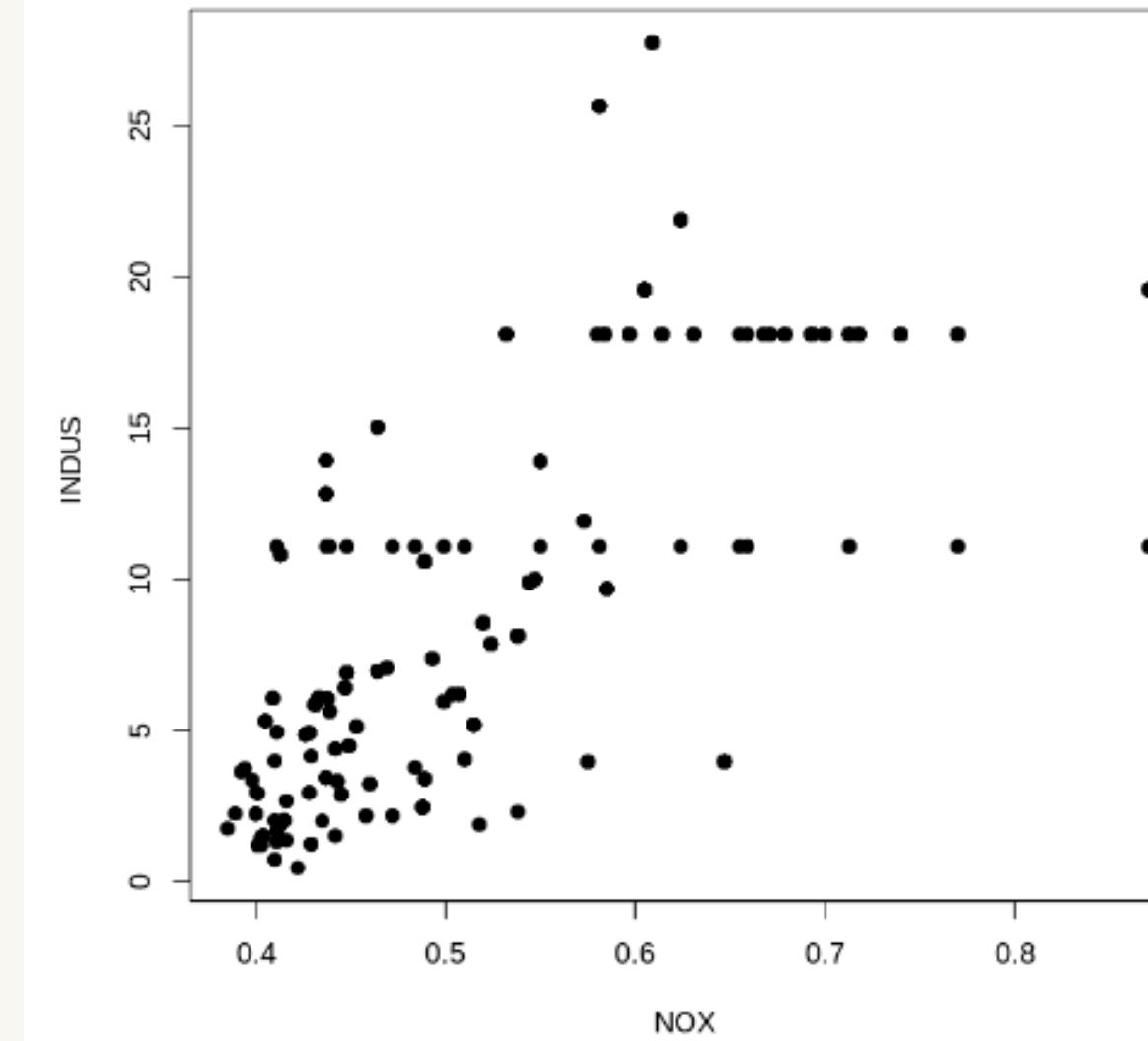
frekuensi tingkat kriminalitas

Data Insights Through Exploration

Scatterplot to show the relation between CRIM and MEDV



Scatterplot to show the relation between NOX and INDUS



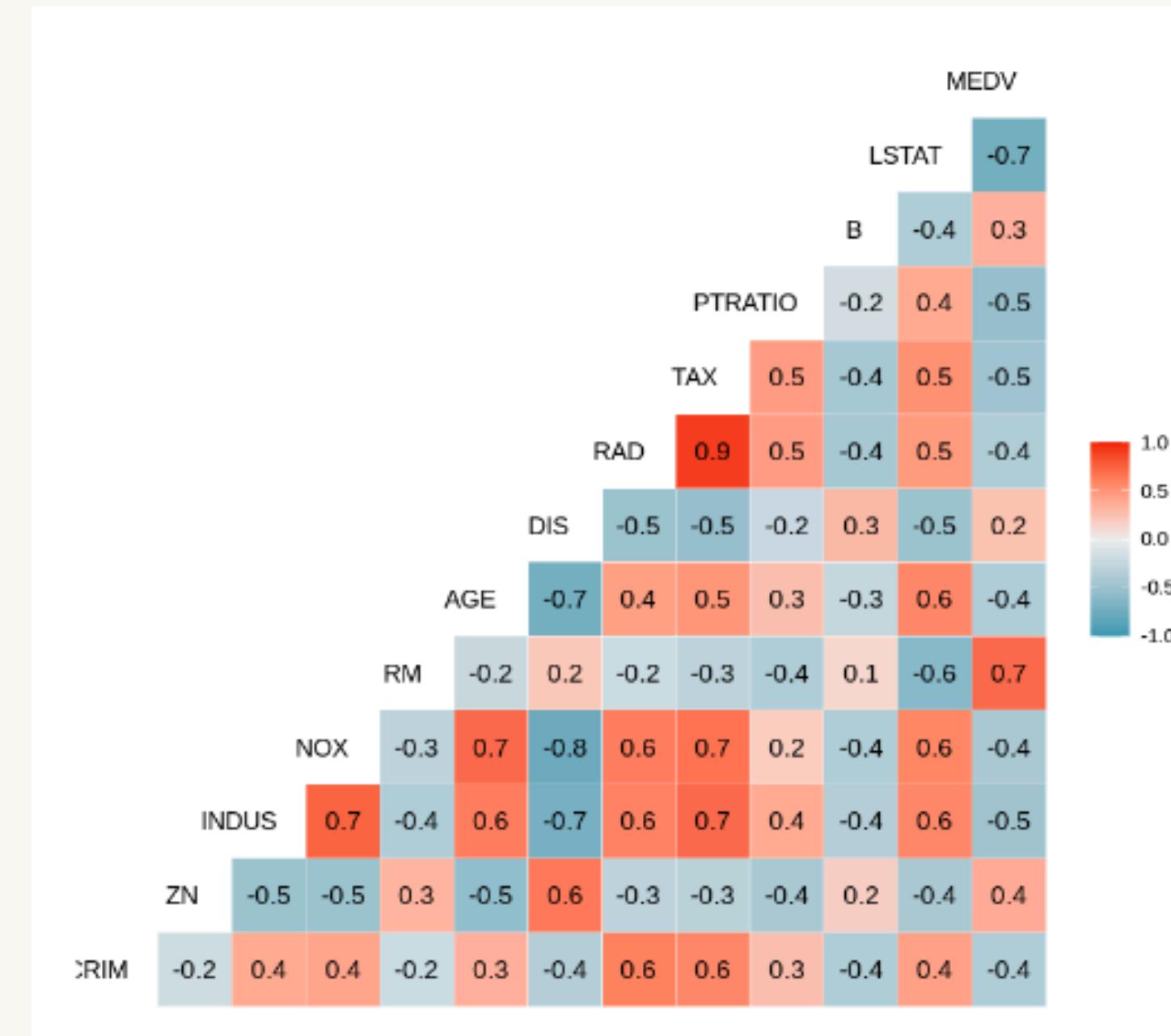
CRIM per capita crime rate by town

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

INDUS proportion of non-retail business acres per town.

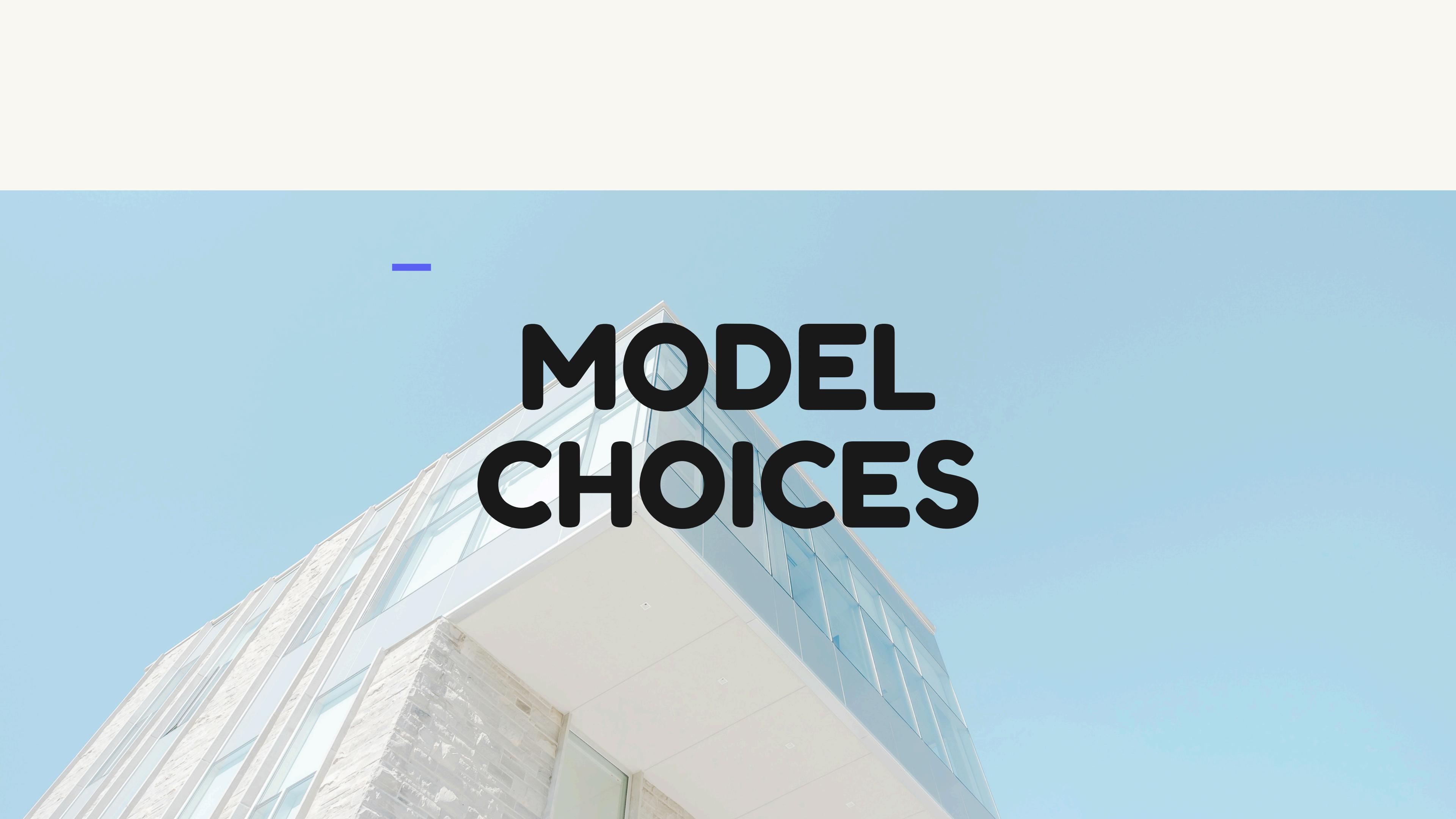
NOX nitric oxides concentration (parts per 10 million)

Correlation



Hasil Analisis dari Output di atas :

- Korelasi hubungan linear paling kuat diperoleh oleh `TAX` sebesar 0.9 terhadap `RAD`
- Korelasi yang paling tidak berhubungan diperoleh oleh Variable `NOX` sebesar -0.8 terhadap `DIS`

A photograph of a modern building's corner, featuring a glass facade with a grid pattern and a light-colored stone or brick base. The building is set against a clear blue sky. In the upper right quadrant of the image, large, bold, black sans-serif text is overlaid.

**MODEL
CHOICES**

Regresi linear dengan 1 prediktor

```
Call:  
lm(formula = MEDV ~ RM, data = df)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-24.255 -2.487  0.108  3.117 39.867  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -36.8574    2.6983 -13.66 <2e-16 ***  
RM            9.4547    0.4259  22.20 <2e-16 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 6.558 on 484 degrees of freedom  
Multiple R-squared:  0.5045,   Adjusted R-squared:  0.5035  
F-statistic: 492.8 on 1 and 484 DF,  p-value: < 2.2e-16
```

- Dari model diperoleh nilai Intercept = -36.8574 dan koefisien RM = 9.4547
Sehingga diperoleh formula model sebagai berikut.
- $MEDV = -36.8574 + 9.4547 \cdot RM$
- Variansi variabel target MEDV dijelaskan sebesar 50.35% oleh variable RM.

Regresi linear dengan seluruh prediktor

```
Call:  
lm(formula = MEDV ~ ., data = df)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-16.7626 -2.8534 -0.5878  1.7728 28.0120  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 30.310808  5.314841  5.703 2.08e-08 ***  
CRIM        -0.107352  0.033081 -3.245 0.001257 **  
ZN          0.038322  0.013991  2.739 0.006396 **  
INDUS       -0.033622  0.061754 -0.544 0.586394  
CHAS1        2.969330  0.885392  3.354 0.000862 ***  
NOX         -16.004708  3.889637 -4.115 4.57e-05 ***  
RM           4.487547  0.433634 10.349 < 2e-16 ***  
AGE         -0.013400  0.013246 -1.012 0.312212  
DIS          -1.458987  0.202662 -7.199 2.40e-12 ***  
RAD          0.261734  0.067651  3.869 0.000125 ***  
TAX          -0.010153  0.003813 -2.663 0.008012 **  
PTRATIO     -0.921395  0.136109 -6.770 3.85e-11 ***  
B            0.010677  0.002818  3.789 0.000171 ***  
LSTAT       -0.443971  0.051303 -8.654 < 2e-16 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 4.818 on 472 degrees of freedom  
Multiple R-squared:  0.7392,   Adjusted R-squared:  0.732  
F-statistic: 102.9 on 13 and 472 DF,  p-value: < 2.2e-16
```

- Berdasarkan output terlihat mayoritas prediktor dari hasil nilai Pr(>|t|) banyak yang dibawah 0.05 yang berarti prediktor signifikan
- Variansi atribut target MEDV dapat dijelaskan oleh variansi CRIM sampai LSTAT sebesar 73.20% . Model ini lebih efisien karena hasil yang diperoleh lebih baik dari pemodelan dengan 1 prediktor.

Regresi linear model stepwise

```
Call:  
lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +  
    TAX + PTRATIO + B + LSTAT, data = df)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-16.5944 -2.8705 -0.5188  1.8591 27.9794  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 30.744417  5.297174  5.804 1.19e-08 ***  
CRIM        -0.107709  0.033031 -3.261 0.001191 **  
ZN          0.040587  0.013841  2.932 0.003527 **  
CHAS1       2.883220  0.879164  3.279 0.001116 **  
NOX        -17.483763  3.657140 -4.781 2.33e-06 ***  
RM          4.435654  0.424171 10.457 < 2e-16 ***  
DIS        -1.374601  0.187935 -7.314 1.11e-12 ***  
RAD         0.276743  0.064956  4.260 2.46e-05 ***  
TAX        -0.011182  0.003442 -3.249 0.001242 **  
PTRATIO     -0.942039  0.134348 -7.012 8.14e-12 ***  
B           0.010537  0.002805  3.757 0.000194 ***  
LSTAT      -0.460114  0.048991 -9.392 < 2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.815 on 474 degrees of freedom  
Multiple R-squared:  0.7384,   Adjusted R-squared:  0.7324  
F-statistic: 121.7 on 11 and 474 DF,  p-value: < 2.2e-16
```

Variansi atribut target MEDV dapat dijelaskan menggunakan model stepwise backward sebesar 73.24% . Model ini sedikit lebih efisien karena hasil yang diperoleh sedikit lebih baik dari pemodelan keseluruhan prediktor

Comparation

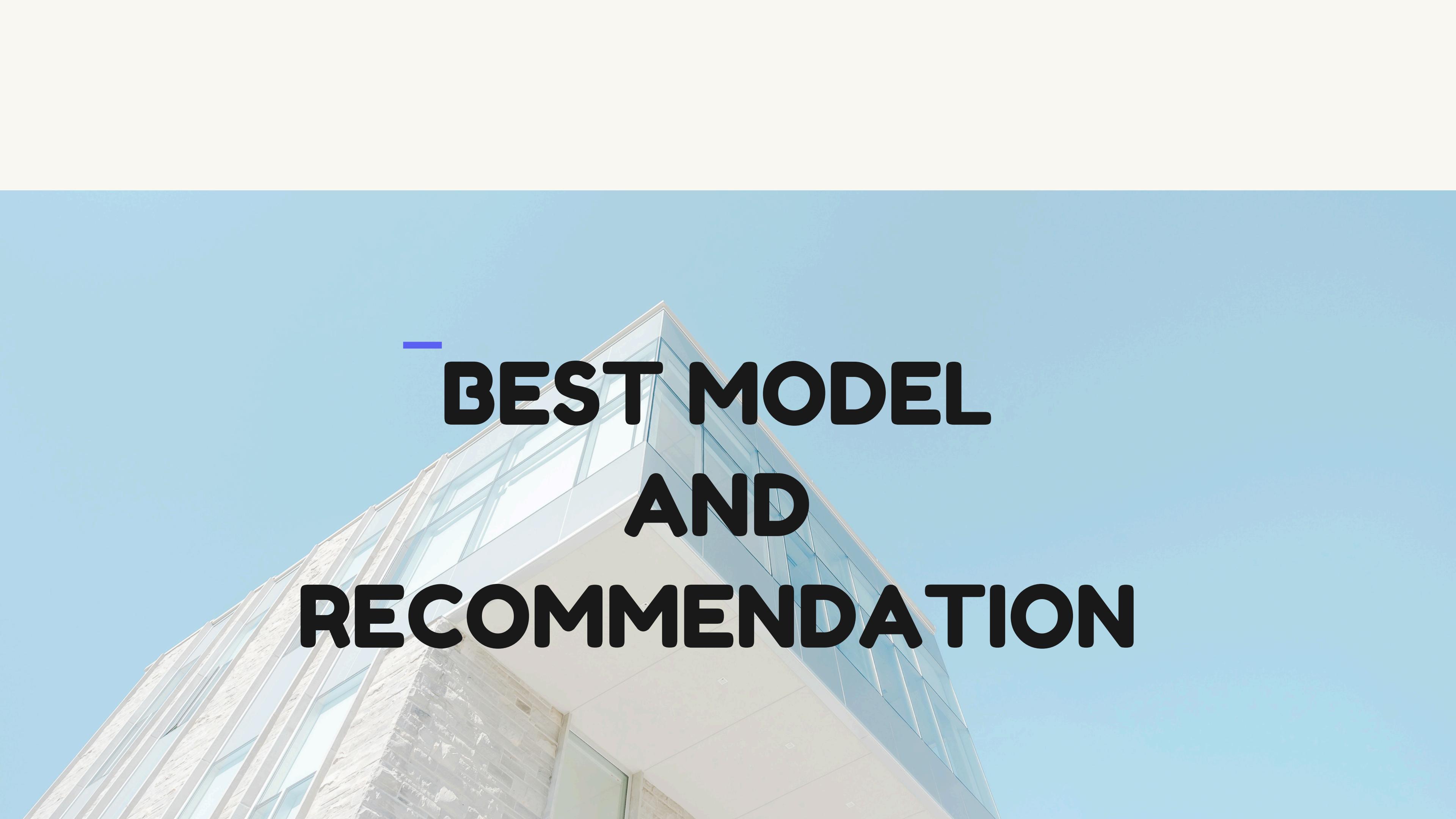
```
[ ] summary(model1)$r.squared  
0.504515940564308
```

```
[ ] summary(model2)$adj.r.squared  
0.731982398817667
```

```
[ ] summary(stepwise)$adj.r.squared  
0.732365792171517
```

- terlihat dari hasil output di atas nilai R-Squared yang paling tinggi terdapat pada **model stepwise** oleh karena itu model ini akan digunakan untuk tahapan selanjutnya

notes : R-squared yang tinggi menunjukkan seberapa baik variabel prediktor dapat menjelaskan variasi dalam variabel respons.

A photograph of a modern building's corner, featuring a glass facade with a grid pattern and a light-colored stone or brick base. The building is set against a clear, light blue sky.

**BEST MODEL
AND
RECOMMENDATION**

Cross-Validation

Train Set

```
Call:  
lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +  
    TAX + PTRATIO + B + LSTAT, data = df_train)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-15.9234 -2.8494 -0.5083  1.8474 26.0741  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 32.496672   5.831973   5.572 4.80e-08 ***  
CRIM        -0.146477   0.039290  -3.728 0.000223 ***  
ZN          0.046775   0.015837   2.954 0.003339 **  
CHAS1       3.558480   0.995346   3.575 0.000396 ***  
NOX        -16.803922   4.076823  -4.122 4.63e-05 ***  
RM          3.938991   0.480590   8.196 3.95e-15 ***  
DIS        -1.259223   0.206278  -6.105 2.57e-09 ***  
RAD         0.274786   0.071682   3.833 0.000148 ***  
TAX        -0.010386   0.003707  -2.802 0.005350 **  
PTRATIO    -0.928800   0.149767  -6.202 1.47e-09 ***  
B           0.010193   0.003223   3.163 0.001691 **  
LSTAT      -0.448701   0.054721  -8.200 3.85e-15 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.739 on 376 degrees of freedom  
Multiple R-squared:  0.7291,    Adjusted R-squared:  0.7212  
F-statistic: 92.02 on 11 and 376 DF,  p-value: < 2.2e-16
```

Test Set

```
Call:  
lm(formula = MEDV ~ NOX + RM + AGE + DIS + RAD + TAX + PTRATIO +  
    B + LSTAT, data = df_test)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-7.7529 -2.7579 -0.2956  2.0460 28.3056  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 33.252845  12.845404   2.589 0.011271 *  
NOX        -20.330343   8.603769  -2.363 0.020333 *  
RM          5.890276   0.908864   6.481 5.10e-09 ***  
AGE        -0.056298   0.030894  -1.822 0.071803 .  
DIS        -2.394567   0.500978  -4.780 6.98e-06 ***  
RAD         0.359849   0.150730   2.387 0.019109 *  
TAX        -0.018182   0.008767  -2.074 0.041006 *  
PTRATIO   -1.131147   0.290839  -3.889 0.000195 ***  
B           0.017614   0.005524   3.188 0.001981 **  
LSTAT      -0.383094   0.115487  -3.317 0.001323 **  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.827 on 88 degrees of freedom  
Multiple R-squared:  0.8106,    Adjusted R-squared:  0.7912  
F-statistic: 41.85 on 9 and 88 DF,  p-value: < 2.2e-16
```

Hasil Analisis :

- Nilai R-Squared dari stepwise_train sebesar 72.91% dan nilai R-Squared dari stepwise_test sebesar 81.06%
- Nilai AIC terendah yang diperoleh dari stepwise_train sebesar 1219.18 dan nilai AIC terendah yang diperoleh dari stepwise_test sebesar 318

Prediction

Real

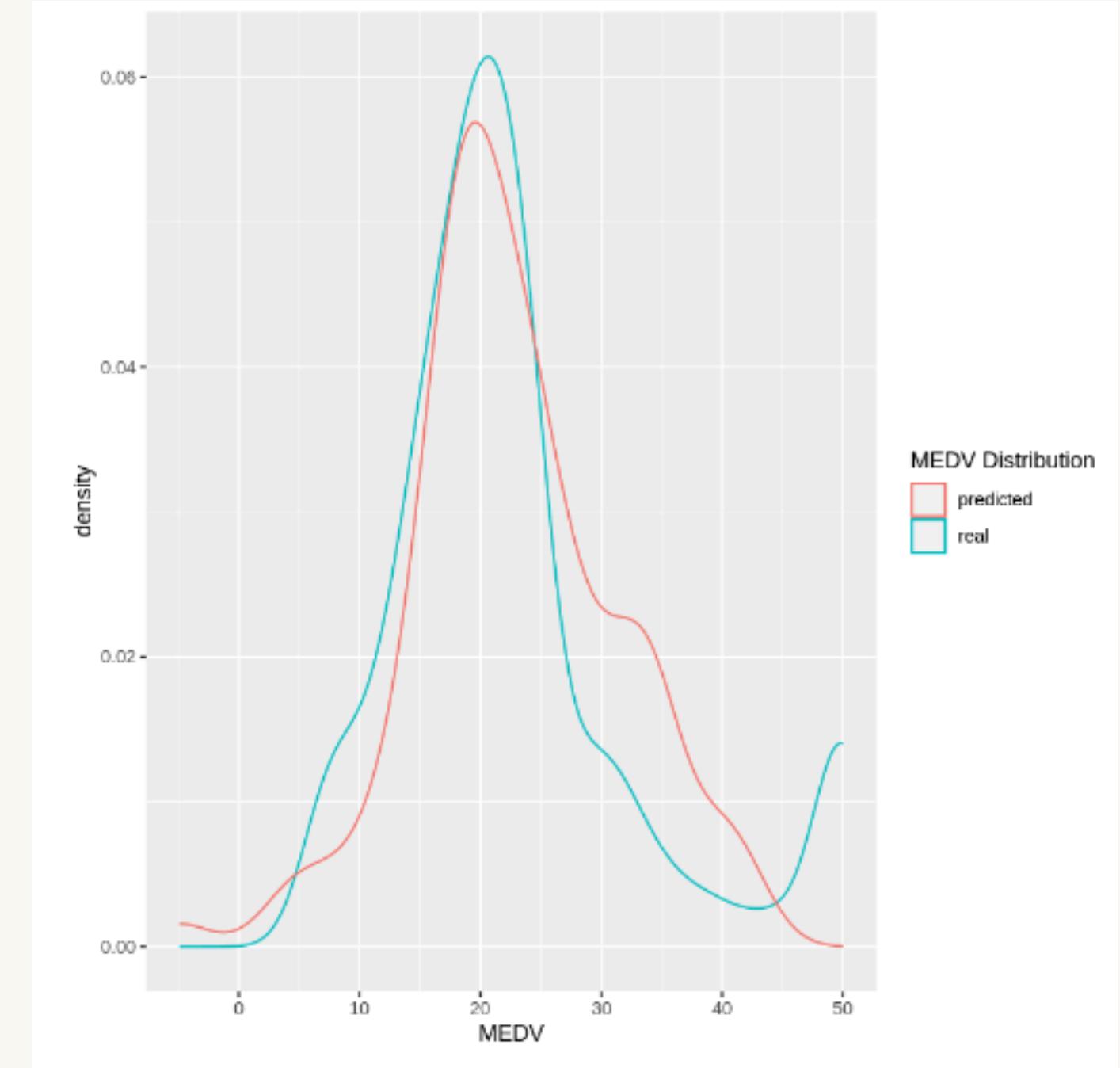
	fit	lwr	upr
1	29.84691	28.78465	30.90917
6	25.31990	24.16421	26.47559
18	17.33987	16.30528	18.37445
20	18.45738	17.33163	19.58312
22	17.97654	16.95281	19.00028

Prediction

A matrix: 5 × 3 of type dbl

	fit	lwr	upr
1	29.84691	20.467752	39.22607
6	25.31990	15.929701	34.71010
18	17.33987	7.963802	26.71593
20	18.45738	9.070818	27.84394
22	17.97654	8.601672	27.35142

Plot



Terlihat dari hasil plot ternyata tidak sampai 80% cocok dalam melakukan prediksi

Comparation

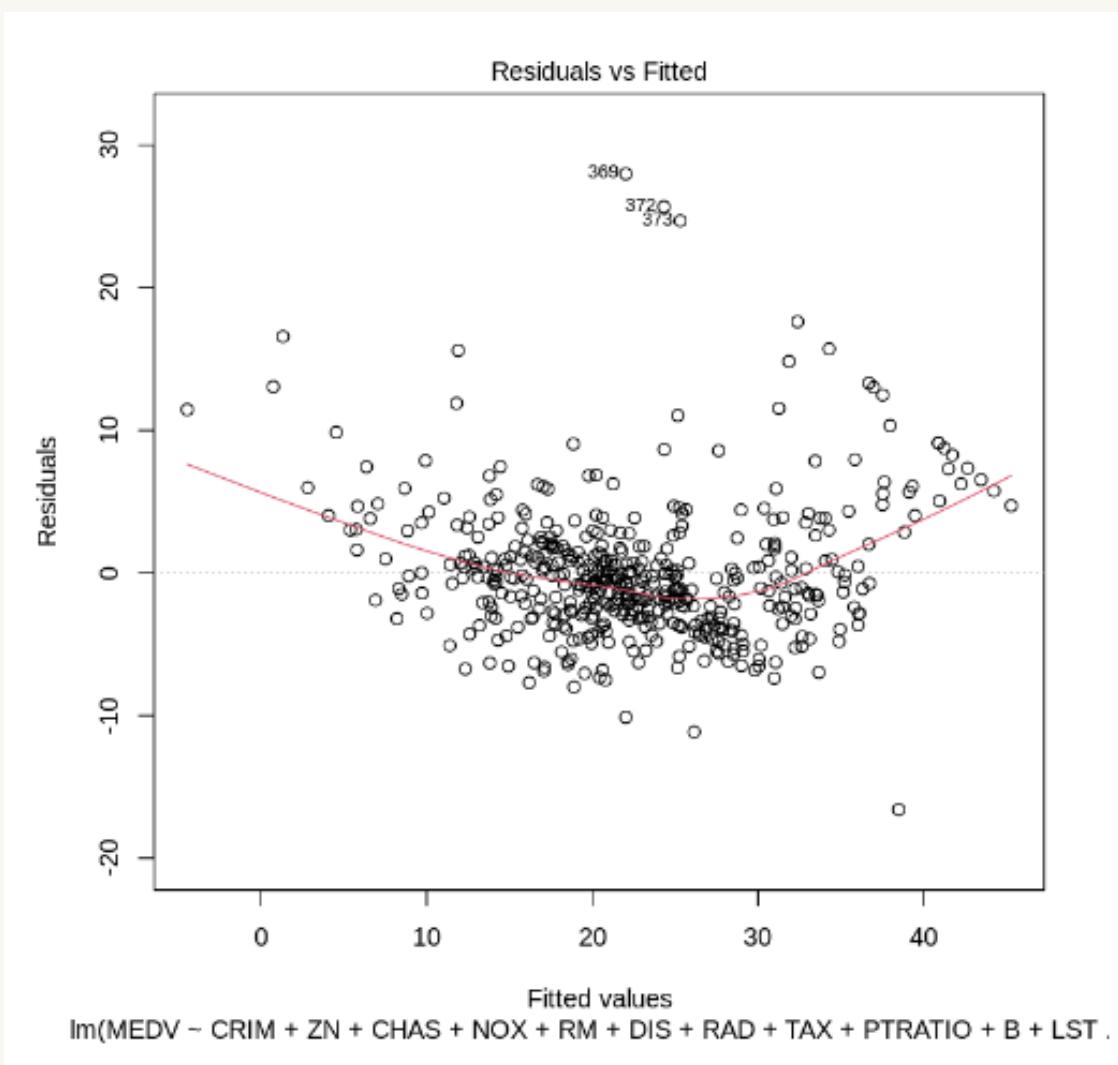
RMSE_train	RMSE_test
<dbl>	<dbl>
4.691222	4.998986

Hasil Analisis :

- Nilai RMSE seberapa besar kesalahan rata-rata dari model dalam memprediksi variabel respons.
- RMSE_train: Ini menunjukkan bahwa rata-rata kesalahan prediksi dari model pertama sebesar 4.691222. Ini artinya, perbedaan antara nilai prediksi dan nilai aktual pada data sekitar 4.69.
- RMSE_test: Sedangkan nilai RMSE kedua, 4.998986, menunjukkan bahwa rata-rata kesalahan prediksi dari model kedua adalah sekitar 4.99.
- Semakin rendah nilai RMSE, semakin baik performa prediksi model. Oleh karena itu, model pertama memiliki tingkat kesalahan yang lebih rendah dibandingkan dengan model kedua, yang berarti model pertama cenderung lebih baik dalam melakukan prediksi terhadap variabel respons. Namun, penting untuk mempertimbangkan konteks data spesifik dan tujuan analisis sebelum mengambil kesimpulan final terkait performa model.

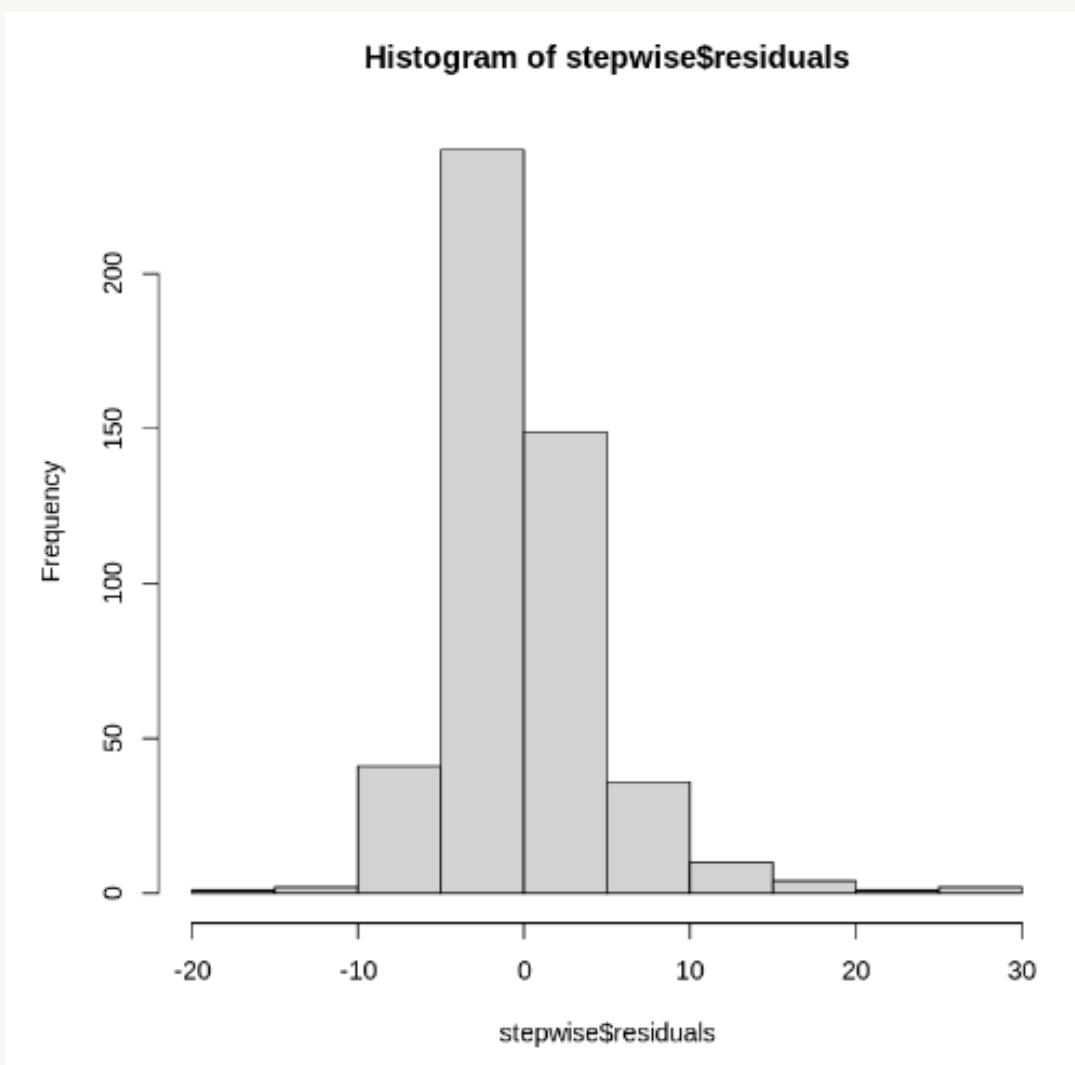
Assumptions

linearitas



Model stepwise memenuhi asumsi linearity karena sebaran nilai residual "bounce randomly" atau sebaran datanya berada disekitar nilai 0.

normalitas



Shapiro-Wilk hypothesis test:

H₀ : berdistribusi normal
H₁ : tidak berdistribusi normal
Kondisi yang diharapkan: H₀

Dari hasil output yang diperoleh :

p-value < 0.05 sehingga tolak H₀ yang berarti model tidak berdistribusi normal atau asumsi Normality of Residuals tidak terpenuhi

```
Shapiro-Wilk normality test
data: stepwise$residuals
W = 0.89403, p-value < 2.2e-16
```

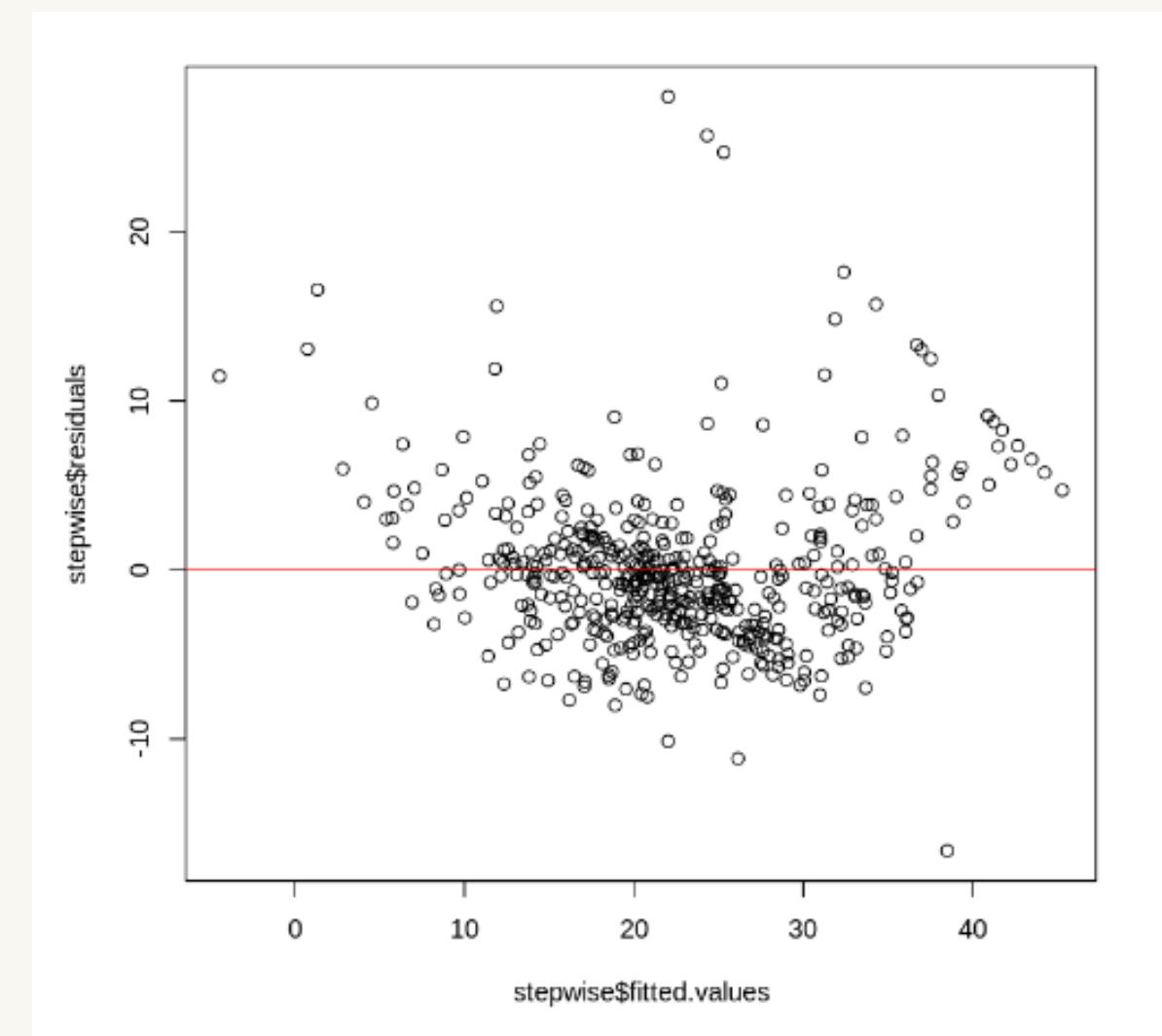
studentized Breusch-Pagan test

```
data: stepwise
BP = 58.104, df = 11, p-value = 2.082e-08
```

Breusch-Pagan hypothesis test:

H₀ : homoscedasticity
H₁ : heteroscedasticity
Kondisi yang diharapkan: H₀

Dari hasil output yang diperoleh : p-value < 0.05 sehingga tolak H₀ yang berarti model mempunyai error yang menyebar tidak konstan atau heteroscedasticity dan asumsinya tidak terpenuhi



Multicollinearity

linearitas

	A data.frame: 11 × 1
vif_values	
<dbl>	
CRIM	1.701916
ZN	2.133868
CHAS	1.054347
NOX	3.725858
RM	1.840166
DIS	3.247334
RAD	6.569621
TAX	6.963005
PTRATIO	1.777505
B	1.361923
LSTAT	2.510595

Uji VIF (Variance Inflation Factor)

nilai VIF > 10 : terjadi multicollinearity

nilai VIF < 10 : tidak terjadi multicollinearity Kondisi yang diharapkan: VIF < 10

Dari hasil output yang diperoleh :

- tidak ada multicollinearity pada model stepwise
- Dari hasil multicollinearity tidak ada nilai yang lebih dari 10, sehingga tidak ditemukan kondisi korelasi antar prediktor yang kuat. tentu saja kita tidak mengharapkan prediktor kita redundan pada sebuah model yang kita buat

CONCLUSION

Model stepwise menjadi model yang cukup baik karena memiliki nilai R-squared paling tinggi senilai 0.732 dibandingkan model lain. Pada pengujian asumsi, model ini hanya berhasil melewati pengujian linearity dan multicollinearity sedangkan pada pengujian lainnya gagal. Jadi dari hasil asumsi tersebut untuk model stepwise memang cukup baik dibandingkan dengan model lainnya namun tidak disarankan untuk 100% terpaku terhadap hasil dalam memprediksi harga rumah terkait mungkin jika dilanjutkan lebih dalam bisa saja memperoleh hasil yang cocok untuk memprediksi harga rumah.

—



Google Colab



Thank you!

**Feel free to call or message for
any questions or clarifications.**