Trabalho - Análise de Dados com R

Introdução

Este documento apresenta uma análise exploratória e modelagem preditiva utilizando o dataset `boston.csv`, conforme as instruções fornecidas. O objetivo é realizar:

1. Análise exploratória dos dados.
2. Ajuste de um modelo de regressão linear.
3. Diagnóstico do modelo.
4. Geração de um relatório em PDF.

Instruções

As etapas seguidas neste trabalho foram:

1. **Análise Exploratória de Dados (EDA):**
   - Inspecionar os dados.
   - Criar gráficos para entender relações e distribuições.
2. **Preparação dos Dados:**
   - Tratar valores ausentes e transformar variáveis, se necessário.
3. **Construção do Modelo:**
   - Ajustar um modelo de regressão linear múltipla.
   - Aplicar técnicas de seleção de variáveis.
4. **Avaliação e Diagnóstico:**
   - Verificar normalidade dos resíduos, multicolinearidade e pontos influentes.
5. **Relatório Final:** Compilado em formato PDF.

Código R

```
# Carregar pacotes necessários
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
```

```
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(skimr)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(broom)
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(rmarkdown)

# 1. Carregar e explorar os dados
boston <- read.csv("boston.csv")

# Resumo estatístico
summary(boston)
```

```
##       CRIM                ZN             INDUS            CHAS
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
```

```
##       NOX               RM              AGE              DIS        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       RAD              TAX           PTRATIO            B        
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      LSTAT            MEDV      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

```
skim(boston)
```

Table 1: Data summary

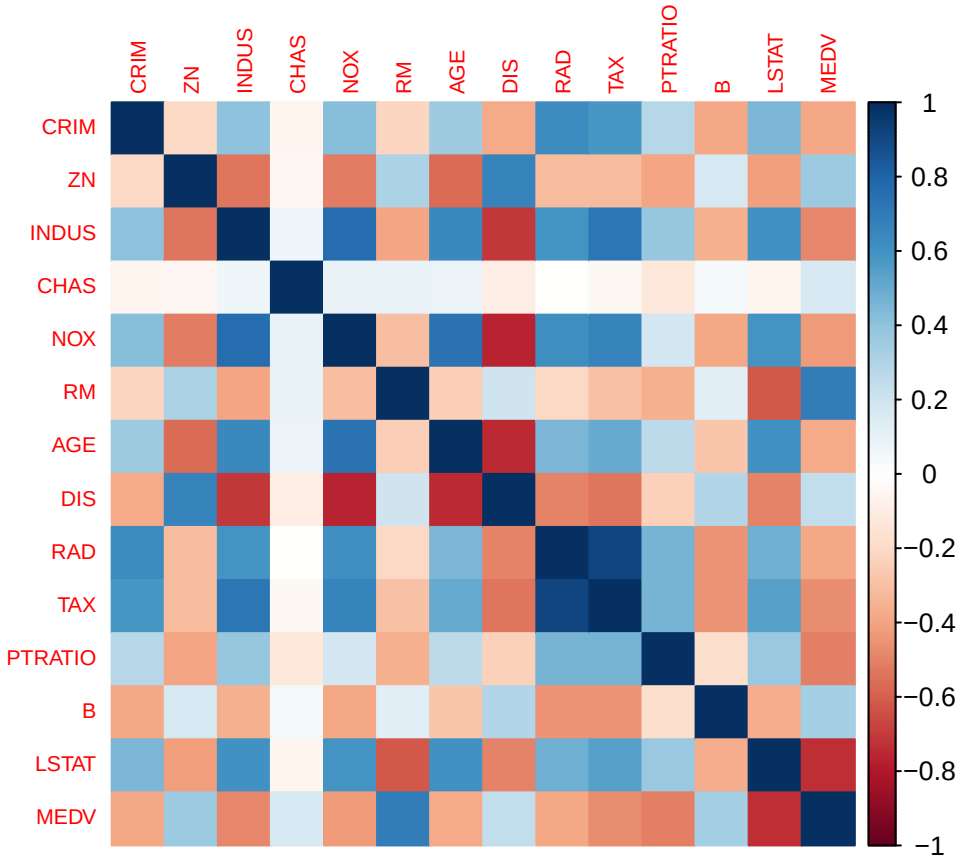
Name	boston
Number of rows	506
Number of columns	14
Column type frequency:	
numeric	14
Group variables	None

Variable type: numeric

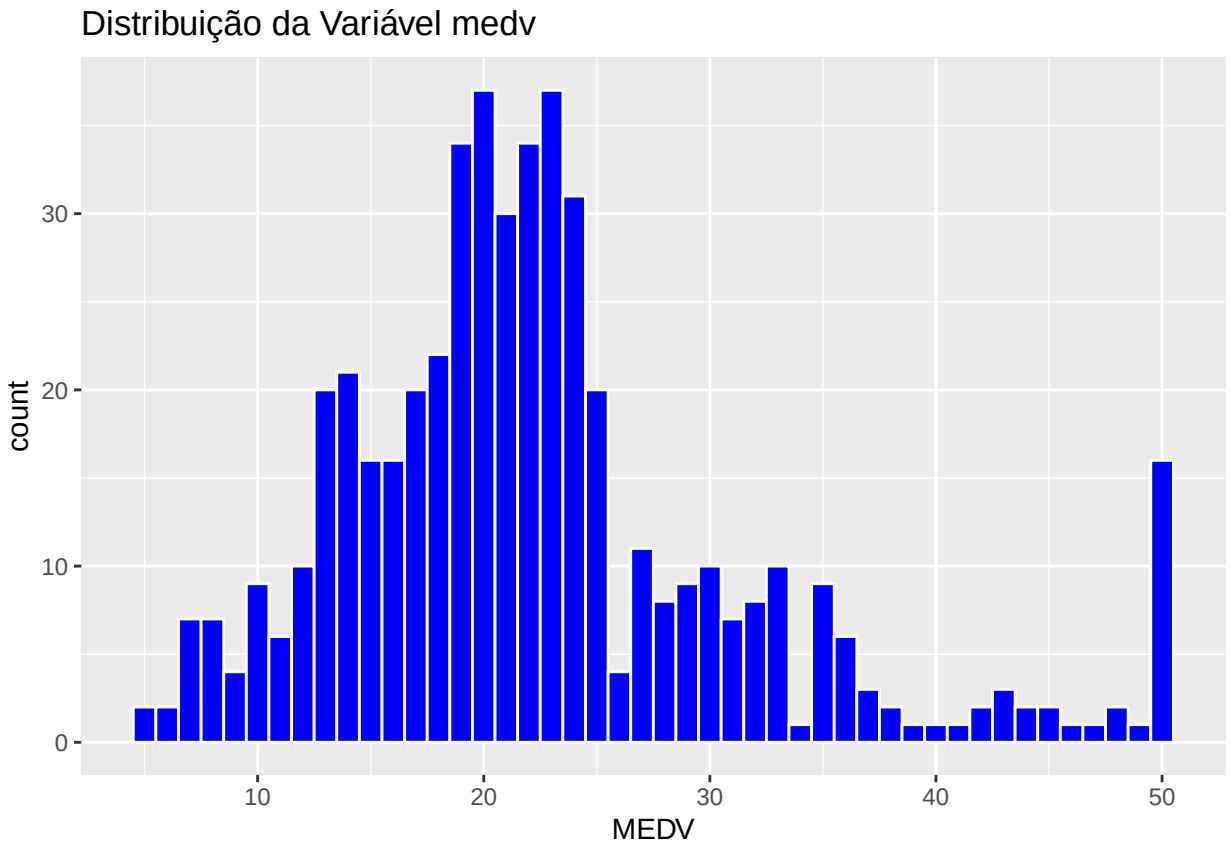
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
CRIM	0	1	3.61	8.60	0.01	0.08	0.26	3.68	88.98	
ZN	0	1	11.36	23.32	0.00	0.00	0.00	12.50	100.00	
INDUS	0	1	11.14	6.86	0.46	5.19	9.69	18.10	27.74	
CHAS	0	1	0.07	0.25	0.00	0.00	0.00	0.00	1.00	
NOX	0	1	0.55	0.12	0.38	0.45	0.54	0.62	0.87	
RM	0	1	6.28	0.70	3.56	5.89	6.21	6.62	8.78	
AGE	0	1	68.57	28.15	2.90	45.02	77.50	94.07	100.00	
DIS	0	1	3.80	2.11	1.13	2.10	3.21	5.19	12.13	
RAD	0	1	9.55	8.71	1.00	4.00	5.00	24.00	24.00	
TAX	0	1	408.24	168.54	187.00	279.00	330.00	666.00	711.00	
PTRATIO	0	1	18.46	2.16	12.60	17.40	19.05	20.20	22.00	
B	0	1	356.67	91.29	0.32	375.38	391.44	396.22	396.90	
LSTAT	0	1	12.65	7.14	1.73	6.95	11.36	16.96	37.97	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
MEDV	0	1	22.53	9.20	5.00	17.02	21.20	25.00	50.00	

```
# Mapa de calor para correlação
cor_matrix <- cor(boston)
corrplot(cor_matrix, method = "color", tl.cex = 0.7)
```

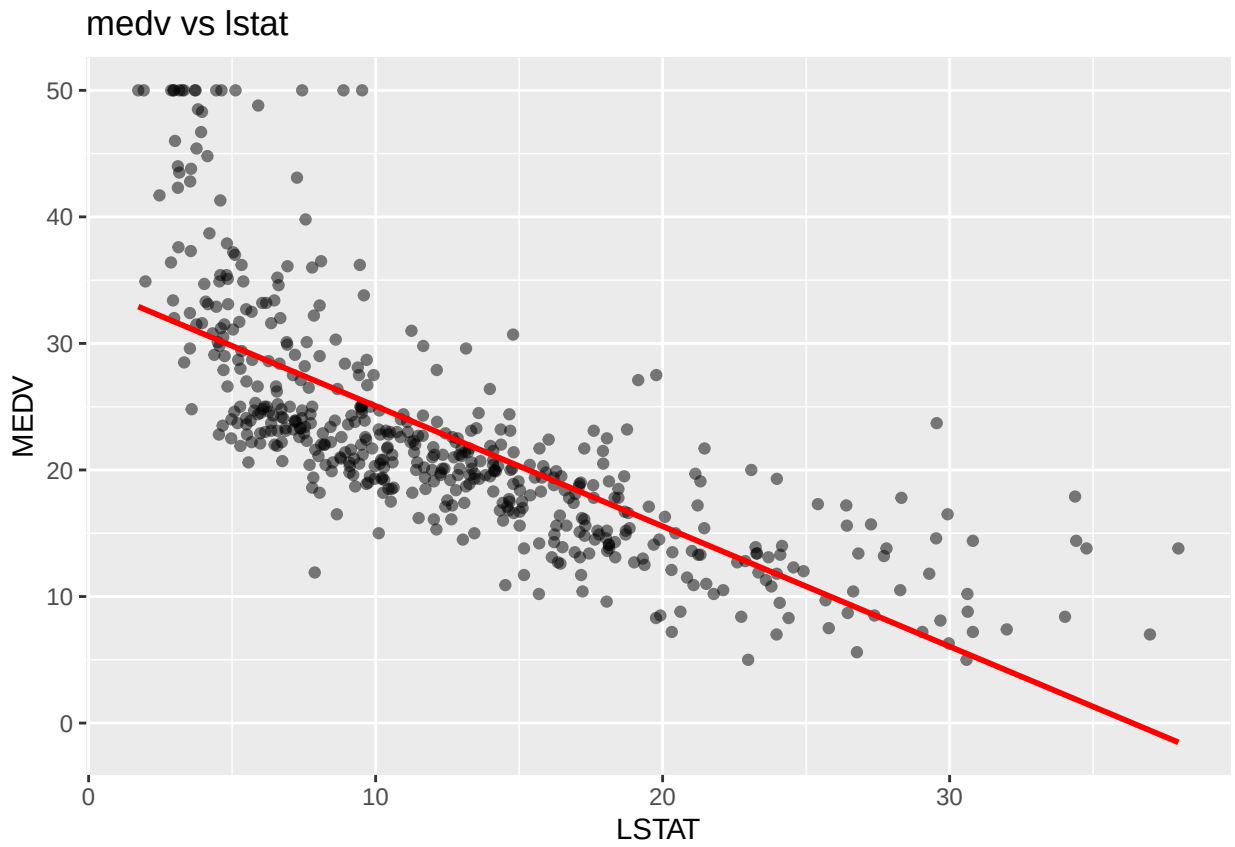



```
# Gráficos exploratórios
ggplot(boston, aes(x = MEDV)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "white") +
  ggtitle("Distribuição da Variável medv")
```




```
ggplot(boston, aes(x = LSTAT, y = MEDV)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  ggtitle("medv vs lstat")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
# 2. Preparação dos dados
# Verificar valores ausentes
sum(is.na(boston))
```

```
## [1] 0
```

```r
# Verificar e ajustar distribuições (exemplo de transformação log)
boston$LSTAT_log <- log(boston$LSTAT + 1)

# 3. Modelo de Regressão Linear
modelo <- lm(MEDV ~ ., data = boston)
summary(modelo)
```

```
##
## Call:
## lm(formula = MEDV ~ ., data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.230  -2.618  -0.281   1.764  25.534
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.454745   5.157653  12.885  < 2e-16 ***
## CRIM         -0.150206   0.029197  -5.145 3.88e-07 ***
```

```
## ZN            0.014013   0.012407   1.129 0.259288
## INDUS         0.008504   0.054237   0.157 0.875478
## CHAS          2.092358   0.761393   2.748 0.006215 **
## NOX         -16.315436   3.370512  -4.841 1.74e-06 ***
## RM            2.461826   0.385537   6.385 3.97e-10 ***
## AGE           0.026950   0.011856   2.273 0.023447 *
## DIS          -1.157665   0.177898  -6.507 1.89e-10 ***
## RAD           0.293110   0.058515   5.009 7.64e-07 ***
## TAX          -0.010800   0.003319  -3.254 0.001215 **
## PTRATIO      -0.837517   0.115770  -7.234 1.81e-12 ***
## B             0.008046   0.002371   3.394 0.000746 ***
## LSTAT         0.462933   0.094253   4.912 1.23e-06 ***
## LSTAT_log   -15.880762   1.334009 -11.905  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.184 on 491 degrees of freedom
## Multiple R-squared:  0.7987, Adjusted R-squared:  0.793
## F-statistic: 139.2 on 14 and 491 DF,  p-value: < 2.2e-16
```

```r
# Seleção de variáveis (Stepwise)
modelo_step <- step(modelo, direction = "both")
```

```
## Start:  AIC=1463.33
## MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT + LSTAT_log
##
##             Df Sum of Sq     RSS    AIC
## - INDUS      1      0.43  8597.8 1461.4
## - ZN         1     22.33  8619.7 1462.6
## <none>                    8597.3 1463.3
## - AGE        1     90.48  8687.8 1466.6
## - CHAS       1    132.23  8729.6 1469.0
## - TAX        1    185.45  8782.8 1472.1
## - B          1    201.65  8799.0 1473.1
## - NOX        1    410.29  9007.6 1484.9
## - LSTAT      1    422.40  9019.7 1485.6
## - RAD        1    439.34  9036.7 1486.5
## - CRIM       1    463.44  9060.8 1487.9
## - RM         1    713.94  9311.3 1501.7
## - DIS        1    741.49  9338.8 1503.2
## - PTRATIO    1    916.38  9513.7 1512.6
## - LSTAT_log  1   2481.46 11078.8 1589.6
##
## Step:  AIC=1461.36
## MEDV ~ CRIM + ZN + CHAS + NOX + RM + AGE + DIS + RAD + TAX +
##     PTRATIO + B + LSTAT + LSTAT_log
##
##             Df Sum of Sq     RSS    AIC
## - ZN         1     21.93  8619.7 1460.6
## <none>                    8597.8 1461.4
## + INDUS      1      0.43  8597.3 1463.3
## - AGE        1     90.54  8688.3 1464.7
## - CHAS       1    135.16  8732.9 1467.2
```

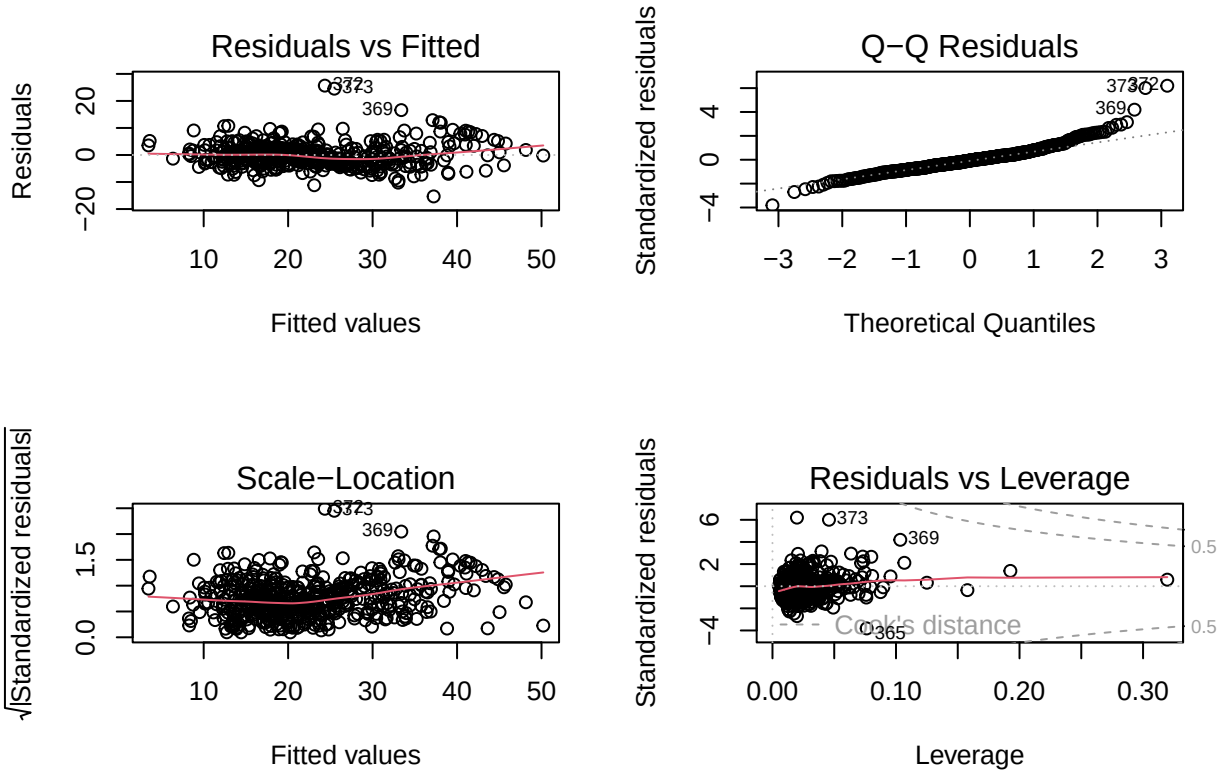
```
## - B          1    201.25  8799.0 1471.1
## - TAX        1    219.95  8817.7 1472.1
## - LSTAT      1    425.00  9022.8 1483.8
## - NOX        1    433.61  9031.4 1484.2
## - CRIM       1    465.25  9063.0 1486.0
## - RAD        1    468.05  9065.8 1486.2
## - RM         1    716.92  9314.7 1499.9
## - DIS        1    782.01  9379.8 1503.4
## - PTRATIO    1    927.09  9524.8 1511.2
## - LSTAT_log  1   2483.55 11081.3 1587.8
##
## Step:  AIC=1460.64
## MEDV ~ CRIM + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO +
##     B + LSTAT + LSTAT_log
##
##             Df Sum of Sq     RSS    AIC
## <none>                    8619.7 1460.6
## + ZN         1     21.93  8597.8 1461.4
## + INDUS      1      0.02  8619.7 1462.6
## - AGE        1     84.40  8704.1 1463.6
## - CHAS       1    133.97  8753.7 1466.5
## - TAX        1    201.09  8820.8 1470.3
## - B          1    201.92  8821.6 1470.4
## - NOX        1    446.46  9066.1 1484.2
## - CRIM       1    454.50  9074.2 1484.7
## - RAD        1    454.78  9074.5 1484.7
## - LSTAT      1    489.57  9109.3 1486.6
## - RM         1    746.44  9366.1 1500.7
## - DIS        1    842.23  9461.9 1505.8
## - PTRATIO    1   1117.80  9737.5 1520.3
## - LSTAT_log  1   2716.61 11336.3 1597.3
```

```
summary(modelo_step)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + CHAS + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT + LSTAT_log, data = boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3296  -2.5537  -0.2686   1.7453  25.6568
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.185356   5.104546  13.162  < 2e-16 ***
## CRIM         -0.148348   0.029096  -5.099 4.89e-07 ***
## CHAS          2.095084   0.756855   2.768 0.005850 **
## NOX         -16.386144   3.242715  -5.053 6.13e-07 ***
## RM            2.495638   0.381951   6.534 1.60e-10 ***
## AGE           0.025953   0.011813   2.197 0.028482 *
## DIS          -1.074593   0.154828  -6.941 1.24e-11 ***
## RAD           0.285466   0.055973   5.100 4.85e-07 ***
## TAX          -0.009904   0.002920  -3.391 0.000751 ***
```

```
## PTRATIO      -0.874011   0.109309  -7.996 9.23e-15 ***
## B             0.008047   0.002368   3.398 0.000733 ***
## LSTAT         0.486207   0.091883   5.292 1.83e-07 ***
## LSTAT_log   -16.211090   1.300531 -12.465  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.181 on 493 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.7933
## F-statistic: 162.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

```r
# 4. Diagnósticos do modelo
par(mfrow = c(2, 2))
plot(modelo_step)
```




```r
# Multicolinearidade
vif_model <- vif(modelo_step)
vif_model
```

```
##      CRIM      CHAS       NOX        RM       AGE       DIS       RAD       TAX
##  1.809149  1.067381  4.078159  2.080173  3.193491  3.070039  6.860612  6.997195
##   PTRATIO         B     LSTAT LSTAT_log
##  1.617541  1.349638 12.434880 14.194443
```

```r
# Resíduos e normalidade
shapiro.test(residuals(modelo_step))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo_step)
## W = 0.93079, p-value = 1.523e-14
```

```r
qqnorm(residuals(modelo_step))
qqline(residuals(modelo_step))

# Pontos influentes
influence <- cooks.distance(modelo_step)
ggplot(data.frame(obs = 1:length(influence), influence), aes(x = obs, y = influence)) +
  geom_bar(stat = "identity") +
  ggtitle("Cooks Distance para Pontos Influentes")
```