Final Project

Rujia Xie

2025-05-15

1. Exploratory Data Analysis

```r
#Load in data downloaded from Kaggle
house_price <- read.csv("Boston-house-price-data.csv")
names(house_price)
```

```
##  [1] "CRIM"    "ZN"      "INDUS"   "CHAS"    "NOX"     "RM"      "AGE"
##  [8] "DIS"     "RAD"     "TAX"     "PTRATIO" "B"       "LSTAT"   "MEDV"
```

This is a dataset called "Boston-house-price-data," which contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. Each observation in the Boston Housing dataset represents a single census tract (neighborhood) in the Boston area. The dataset includes various attributes describing the socioeconomic, environmental, and housing characteristics of these tracts, along with the median value of owner-occupied homes.

All variables in the dataset in order are:

- 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)ˆ2 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

For this project, I will use MEDV as the target/outcome/dependent variable, and CRIM, RM, DIS, and RAD as the predictor/independent variables.

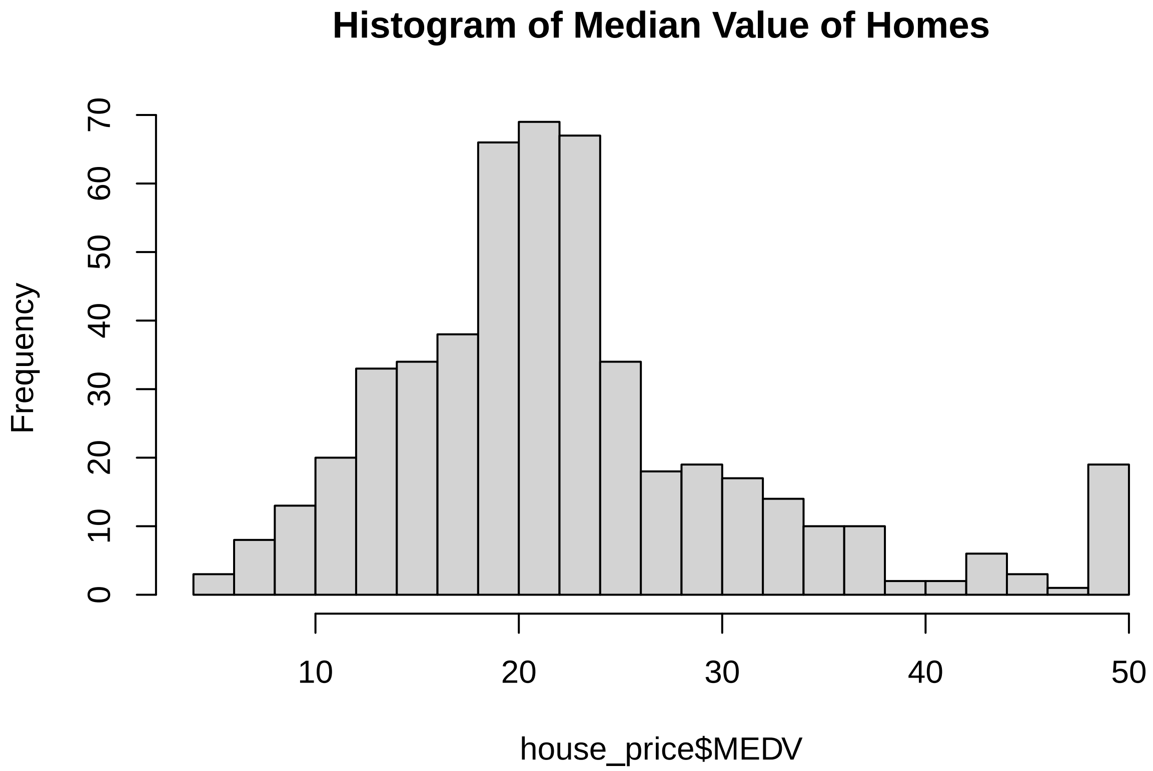
1.1 Summary statistics of variables

```
describe(house_price)
```

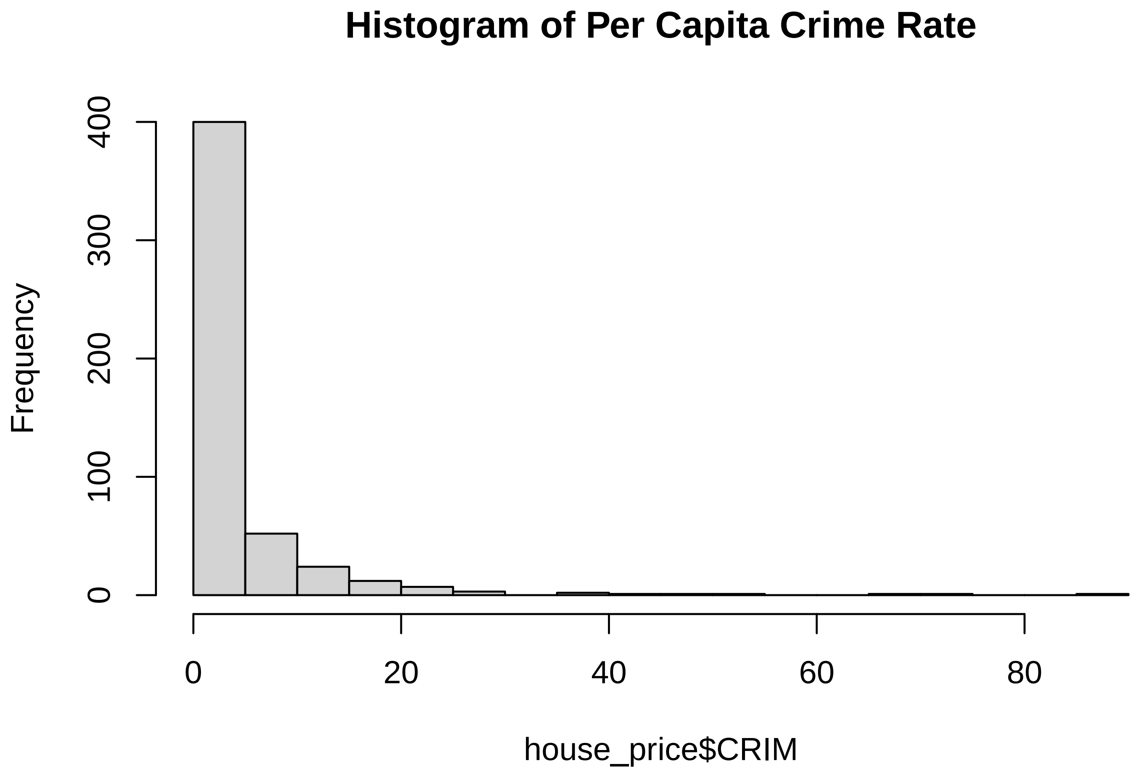
```
##         vars   n   mean     sd median trimmed    mad    min    max  range  skew
## CRIM       1 506   3.61   8.60   0.26    1.68   0.33   0.01  88.98  88.97  5.19
## ZN         2 506  11.36  23.32   0.00    5.08   0.00   0.00 100.00 100.00  2.21
## INDUS      3 506  11.14   6.86   9.69   10.93   9.37   0.46  27.74  27.28  0.29
## CHAS       4 506   0.07   0.25   0.00    0.00   0.00   0.00   1.00   1.00  3.39
## NOX        5 506   0.55   0.12   0.54    0.55   0.13   0.38   0.87   0.49  0.72
## RM         6 506   6.28   0.70   6.21    6.25   0.51   3.56   8.78   5.22  0.40
## AGE        7 506  68.57  28.15  77.50   71.20  28.98   2.90 100.00  97.10 -0.60
## DIS        8 506   3.80   2.11   3.21    3.54   1.91   1.13  12.13  11.00  1.01
## RAD        9 506   9.55   8.71   5.00    8.73   2.97   1.00  24.00  23.00  1.00
## TAX       10 506 408.24 168.54 330.00  400.04 108.23 187.00 711.00 524.00  0.67
## PTRATIO   11 506  18.46   2.16  19.05   18.66   1.70  12.60  22.00   9.40 -0.80
## B         12 506 356.67  91.29 391.44  383.17   8.09   0.32 396.90 396.58 -2.87
## LSTAT     13 506  12.65   7.14  11.36   11.90   7.11   1.73  37.97  36.24  0.90
## MEDV      14 506  22.53   9.20  21.20   21.56   5.93   5.00  50.00  45.00  1.10
##         kurtosis   se
## CRIM       36.60 0.38
## ZN          3.95 1.04
## INDUS      -1.24 0.30
## CHAS        9.48 0.01
## NOX        -0.09 0.01
## RM          1.84 0.03
## AGE        -0.98 1.25
## DIS         0.46 0.09
## RAD        -0.88 0.39
## TAX        -1.15 7.49
## PTRATIO    -0.30 0.10
## B           7.10 4.06
## LSTAT       0.46 0.32
## MEDV        1.45 0.41
```

1.2 Visualization of distributions and relationships

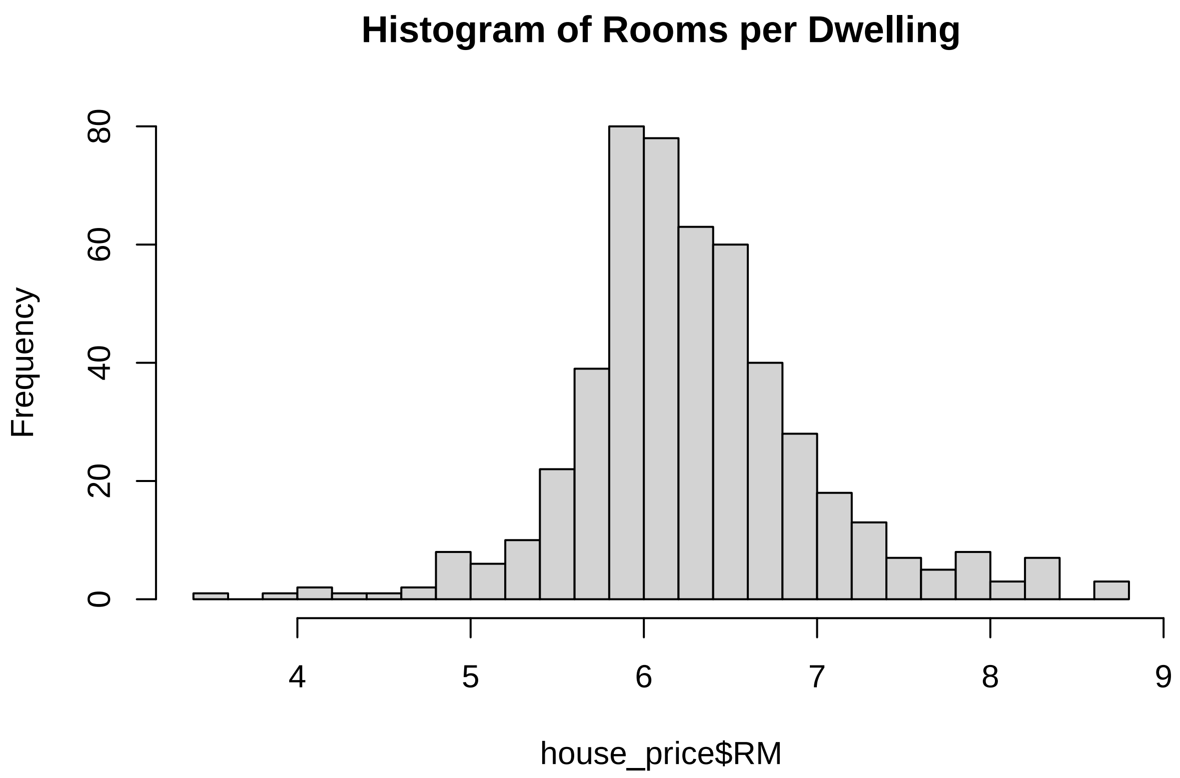
```
hist(house_price$MEDV, breaks = 30, main = "Histogram of Median Value of Homes")
```



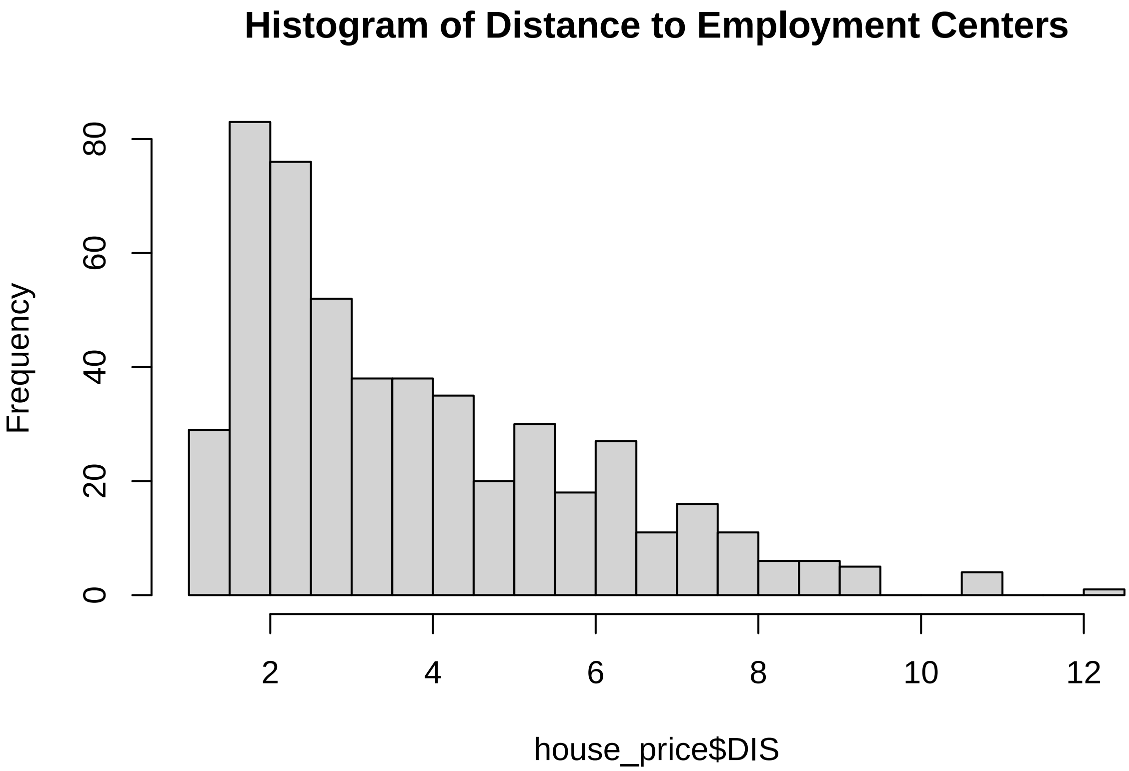
```r
hist(house_price$CRIM, breaks = 30, main = "Histogram of Per Capita Crime Rate")
```



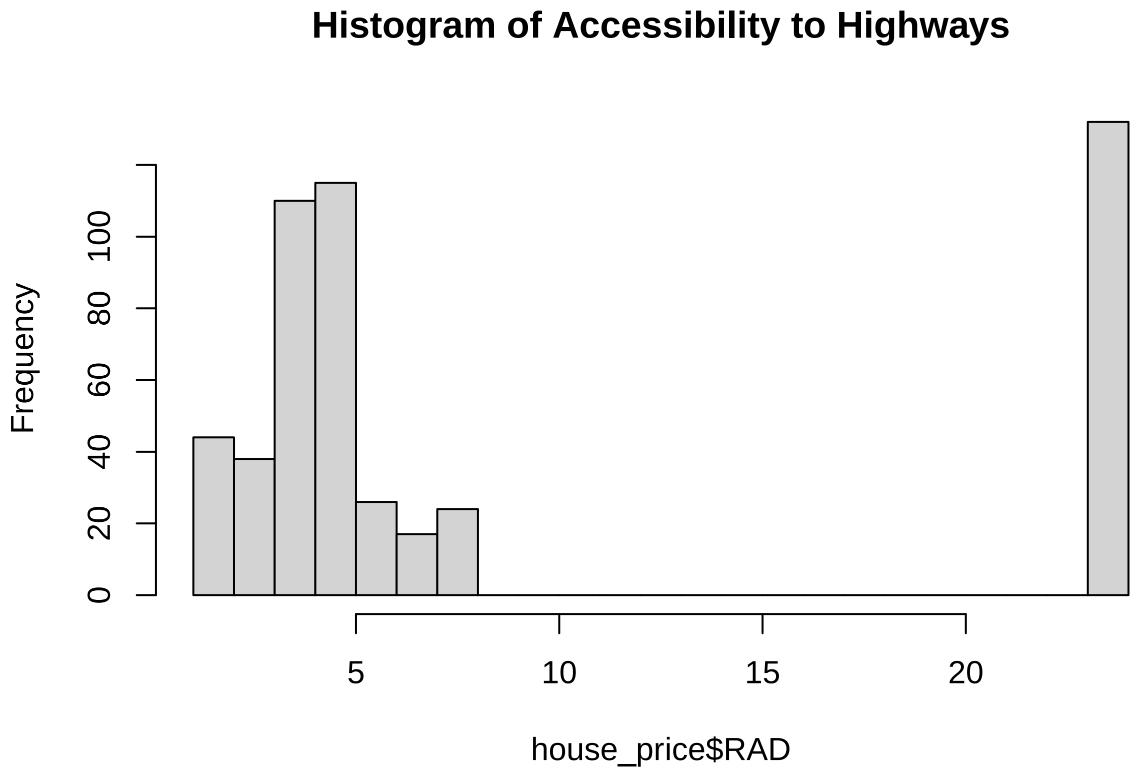
```
hist(house_price$RM, breaks = 30, main = "Histogram of Rooms per Dwelling")
```



```
hist(house_price$DIS, breaks = 30, main = "Histogram of Distance to Employment Centers")
```



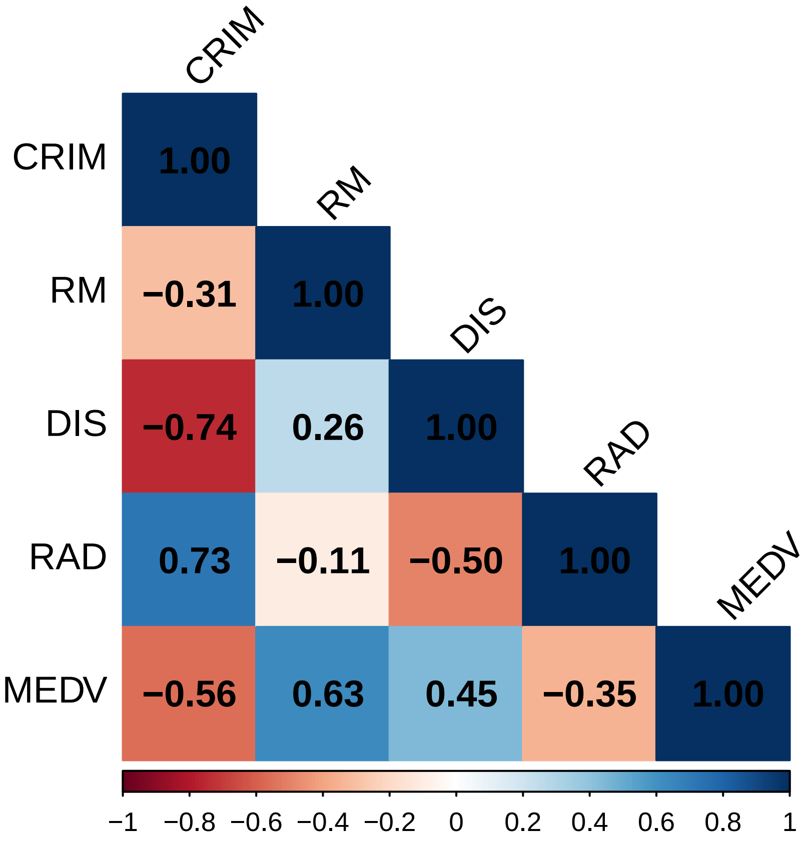
```r
hist(house_price$RAD, breaks = 30, main = "Histogram of Accessibility to Highways")
```



In the following correlation matrix, a correlation coefficient was shown in each cell.A red filling in the cell indicated a negative correlation, and a blue filling indicated a positive correlation. The P value for each correlation was also calculated.Correlations with p>.05 were crossed out in the matrix.

```
house_subset <- house_price [c(1, 6, 8, 9, 14)]
house_correlation <-cor(house_subset, method="spearman",
                           use="pairwise.complete.obs")


pmatrix <- cor.mtest(house_subset, conf.level = .95)
corrplot(house_correlation, p.mat = pmatrix$p, sig.level = .05,
         insig = "pch", pch.cex = 3, pch.col = "red", type = "lower",
         method="color", addCoef.col = "black", number.cex = 1.2,
         tl.cex = 1.2, tl.srt = 45, tl.col = "black")
```



1.3 Identification of missing values and outliers

```
sum(is.na(house_price))
```

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

1.4 Data cleaning and preprocessing steps

```
house_price$CHAS <- factor(house_price$CHAS)
```

2. Regression Assumptions Verification

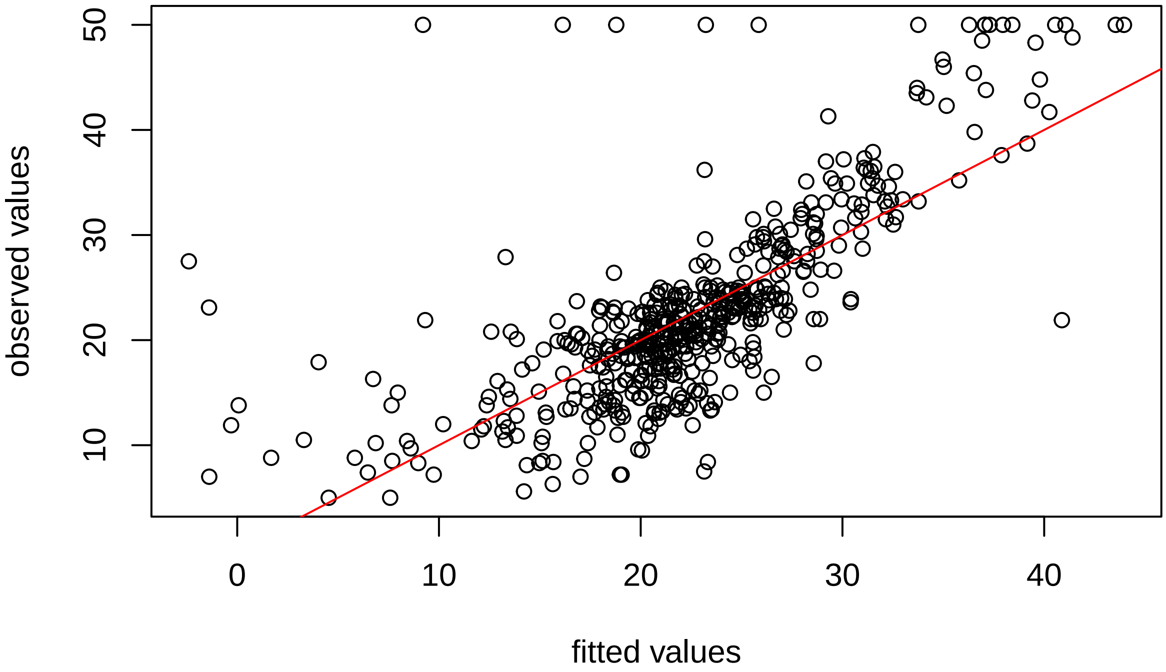
```
# Run the fitted linear regression model
house_model <- lm(MEDV ~ CRIM + RM + DIS + RAD,
           data = house_price)
summary(house_model)
```

```
##
```

```
## Call:
## lm(formula = MEDV ~ CRIM + RM + DIS + RAD, data = house_price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.977  -3.011  -0.581   2.391  40.792
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.86226    2.64775 -10.145  < 2e-16 ***
## CRIM         -0.16749    0.04123  -4.062 5.64e-05 ***
## RM            8.26162    0.40350  20.475  < 2e-16 ***
## DIS          -0.08018    0.15126  -0.530   0.5963
## RAD          -0.16930    0.04317  -3.921   0.0001 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.151 on 501 degrees of freedom
## Multiple R-squared:  0.5563, Adjusted R-squared:  0.5527
## F-statistic:   157 on 4 and 501 DF,  p-value: < 2.2e-16
```

2.1 Linearity assessment

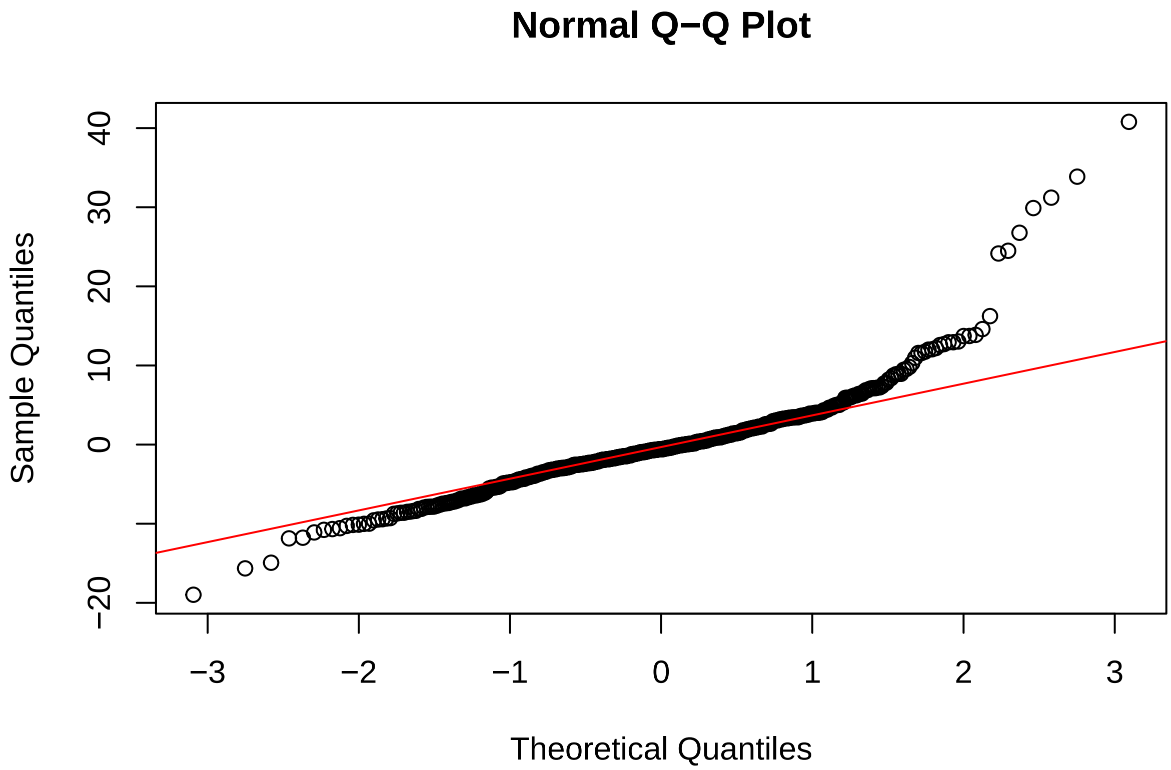
```r
plot(house_model$fitted.values, house_price$MEDV,
     xlab = "fitted values", ylab = "observed values")
abline(a = 0, b = 1, col = "red")
```



Majority of the observations follow a linear trend.

2.2 Normality of residuals

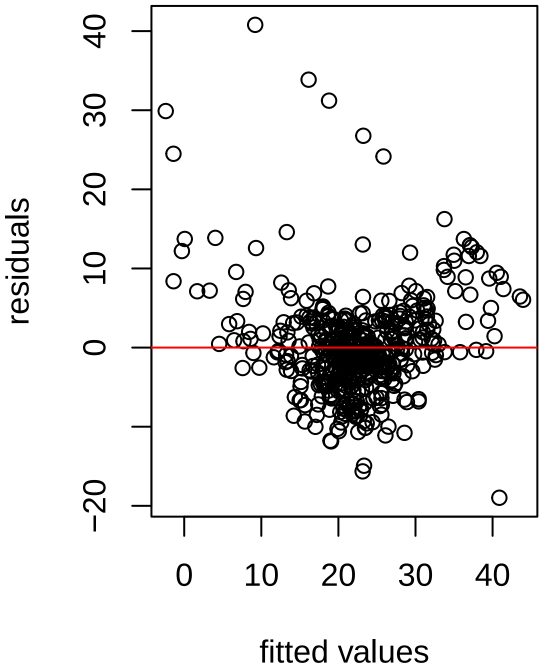
```
qqnorm(house_model$residuals)
qqline(house_model$residuals, col = "red")
```



Residuals of most observations follow a normal distribution, although outliers are present at both ends.

2.3 Check homoscedasticity (constant variance of residuals)

```
par(mfrow = c(1,2))
plot(house_model$fitted.values, house_model$residuals,
     xlab = "fitted values", ylab = "residuals")
abline(h = 0, col = "red")
```



Heteroscedasticity is present in the residuals, since they are not randomly scattered around the red line in the plot.

2.4 check independence of observations

Because each row represents a unique census tract (neighborhood) in the Boston area, there are no repeated observations from the same subject. Thus, each observation is independent of the others.

2.5 Check multicollinearity

```
vif(house_model)
```

```
##     CRIM       RM      DIS      RAD
## 1.678828 1.072855 1.354230 1.886375
```

All VIFs shown above are near 1, so multicollinearity is not present in the model.

3. Assumption Violation Handling

3.1 Apply appropriate transformations when assumptions are violated and document your approach to each violation

According to last section, the homoscedasticity assumption is violated, so we need to use heteroskedasticity-Consistent (HC) Standard Errors to address heteroscedasticity.

```
sandwich1 <- coeftest(house_model, vcov = vcovHC(house_model, type = 'HC3'))
sandwich1
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -26.862256   4.238145 -6.3382 5.188e-10 ***
## CRIM         -0.167485   0.039007 -4.2937 2.111e-05 ***
## RM            8.261620   0.689705 11.9785 < 2.2e-16 ***
## DIS          -0.080179   0.110215 -0.7275  0.467271
## RAD          -0.169299   0.053668 -3.1545  0.001704 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

3.2 Compare models before and after corrections

This is the model summary before correction:

```
Call:
lm(formula = MEDV ~ CRIM + RM + DIS + RAD, data = house_price)

Residuals:
   Min     1Q Median     3Q    Max
-18.98  -3.01  -0.58   2.39  40.79

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.8623     2.6478  -10.15  < 2e-16 ***
CRIM         -0.1675     0.0412   -4.06 0.000056 ***
RM            8.2616     0.4035   20.48  < 2e-16 ***
DIS          -0.0802     0.1513   -0.53   0.5963
RAD          -0.1693     0.0432   -3.92   0.0001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.15 on 501 degrees of freedom
Multiple R-squared:  0.556, Adjusted R-squared:  0.553
F-statistic:  157 on 4 and 501 DF,  p-value: <2e-16
```

This is the model summary after correction:

```
t test of coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.8623     4.2381   -6.34  5.2e-10 ***
CRIM         -0.1675     0.0390   -4.29  2.1e-05 ***
RM            8.2616     0.6897   11.98  < 2e-16 ***
DIS          -0.0802     0.1102   -0.73   0.4673
RAD          -0.1693     0.0537   -3.15   0.0017 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In other words, after correction, RAD becomes less significant, and RM's standard error increases from 0.40 to 0.69.

4. Variable Selection & Hypothesis Testing

4.1 Implement at least two different variable selection techniques

```r
# Forward selection
null <- lm(MEDV ~ 1, data = house_price)
full <- lm(MEDV ~ ., data = house_price)
n <- nrow(house_price)

forward_model <- stats::step(null,
                      scope = list(lower = null, upper = full),
                      direction = "forward",
                      k = log(n))
```

```
## Start:  AIC=2250.74
## MEDV ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + LSTAT    1   23243.9 19472 1859.5
## + RM       1   20654.4 22062 1922.6
## + PTRATIO  1   11014.3 31702 2106.1
## + INDUS    1    9995.2 32721 2122.1
## + TAX      1    9377.3 33339 2131.6
## + NOX      1    7800.1 34916 2154.9
## + CRIM     1    6440.8 36276 2174.3
## + RAD      1    6221.1 36495 2177.3
## + AGE      1    6069.8 36647 2179.4
## + ZN       1    5549.7 37167 2186.6
## + B        1    4749.9 37966 2197.3
## + DIS      1    2668.2 40048 2224.3
## + CHAS     1    1312.1 41404 2241.2
## <none>                 42716 2250.7
##
## Step:  AIC=1859.46
## MEDV ~ LSTAT
##
##           Df Sum of Sq   RSS    AIC
## + RM       1    4033.1 15439 1748.3
## + PTRATIO  1    2670.1 16802 1791.1
## + CHAS     1     786.3 18686 1844.8
## + DIS      1     772.4 18700 1845.2
## + AGE      1     304.3 19168 1857.7
## + TAX      1     274.4 19198 1858.5
## <none>                 19472 1859.5
## + B        1     198.3 19274 1860.5
## + ZN       1     160.3 19312 1861.5
## + CRIM     1     146.9 19325 1861.9
## + INDUS    1      98.7 19374 1863.1
## + RAD      1      25.1 19447 1865.0
## + NOX      1       4.8 19468 1865.6
##
## Step:  AIC=1748.26
## MEDV ~ LSTAT + RM
##
##           Df Sum of Sq   RSS    AIC
## + PTRATIO  1   1711.32 13728 1695.0
## + CHAS     1    548.53 14891 1736.2
## + B        1    512.31 14927 1737.4
## + TAX      1    425.16 15014 1740.3
## + DIS      1    351.15 15088 1742.8
## + CRIM     1    311.42 15128 1744.2
## <none>                 15439 1748.3
## + RAD      1    180.45 15259 1748.5
## + INDUS    1     61.09 15378 1752.5
## + ZN       1     56.56 15383 1752.6
## + AGE      1     20.18 15419 1753.8
## + NOX      1     14.90 15424 1754.0
##
## Step:  AIC=1695.04
## MEDV ~ LSTAT + RM + PTRATIO
##
##         Df Sum of Sq   RSS    AIC
## + DIS    1    499.08 13229 1682.5
## + B      1    389.68 13338 1686.7
## + CHAS   1    377.96 13350 1687.1
## <none>               13728 1695.0
## + CRIM   1    122.52 13606 1696.7
## + AGE    1     66.24 13662 1698.8
## + TAX    1     44.36 13684 1699.6
## + NOX    1     24.81 13703 1700.3
## + ZN     1     14.96 13713 1700.7
## + RAD    1      6.07 13722 1701.0
## + INDUS  1      0.83 13727 1701.2
##
## Step:  AIC=1682.53
## MEDV ~ LSTAT + RM + PTRATIO + DIS
##
##         Df Sum of Sq   RSS    AIC
## + NOX    1    759.56 12469 1658.8
## + B      1    502.64 12726 1669.2
## + CHAS   1    267.43 12962 1678.4
## + INDUS  1    242.65 12986 1679.4
## + TAX    1    240.34 12989 1679.5
## + CRIM   1    233.54 12995 1679.7
## <none>               13229 1682.5
## + ZN     1    144.81 13084 1683.2
## + AGE    1     61.36 13168 1686.4
## + RAD    1     22.40 13206 1687.9
##
## Step:  AIC=1658.83
## MEDV ~ LSTAT + RM + PTRATIO + DIS + NOX
##
##         Df Sum of Sq   RSS    AIC
## + CHAS   1    328.27 12141 1651.6
## + B      1    311.83 12158 1652.2
## <none>               12469 1658.8
## + ZN     1    151.71 12318 1658.9
## + CRIM   1    141.43 12328 1659.3
## + RAD    1     53.48 12416 1662.9
## + INDUS  1     17.10 12452 1664.4
## + TAX    1     10.50 12459 1664.6
## + AGE    1      0.25 12469 1665.0
##
## Step:  AIC=1651.56
## MEDV ~ LSTAT + RM + PTRATIO + DIS + NOX + CHAS
##
##         Df Sum of Sq   RSS    AIC
## + B      1   272.837 11868 1646.3
## + ZN     1   164.406 11977 1650.9
## <none>               12141 1651.6
## + CRIM   1   116.330 12025 1652.9
## + RAD    1    58.556 12082 1655.3
## + INDUS  1    26.274 12115 1656.7
## + TAX    1     4.187 12137 1657.6
## + AGE    1     2.331 12139 1657.7
##
## Step:  AIC=1646.28
## MEDV ~ LSTAT + RM + PTRATIO + DIS + NOX + CHAS + B
##
##         Df Sum of Sq   RSS    AIC
## + ZN     1   189.936 11678 1644.3
## <none>               11868 1646.3
## + RAD    1   144.320 11724 1646.3
## + CRIM   1    55.633 11813 1650.1
## + INDUS  1    15.584 11853 1651.8
## + AGE    1     9.446 11859 1652.1
## + TAX    1     2.703 11866 1652.4
##
## Step:  AIC=1644.35
## MEDV ~ LSTAT + RM + PTRATIO + DIS + NOX + CHAS + B + ZN
##
##         Df Sum of Sq   RSS    AIC
## <none>               11678 1644.3
## + CRIM   1    94.712 11584 1646.5
## + RAD    1    93.614 11585 1646.5
## + INDUS  1    16.048 11662 1649.9
## + TAX     1     3.952 11674 1650.4
## + AGE     1     1.491 11677 1650.5
```

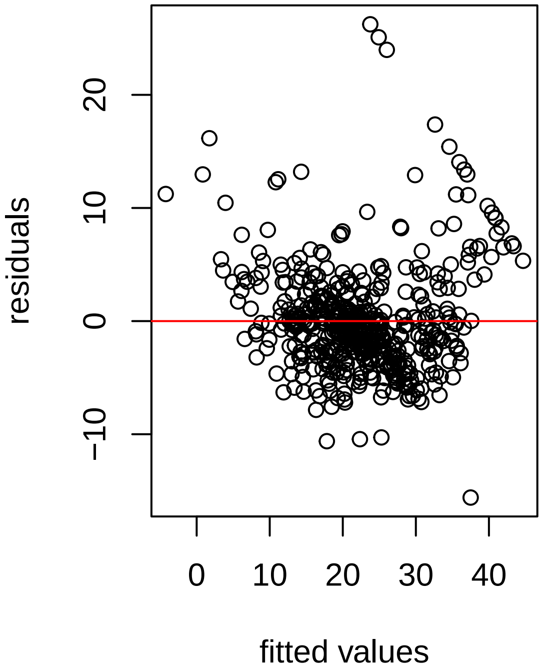
```
# Backward selection
backward_model <- stats::step(full,
                      scope = list(lower = null, upper = full),
                      direction = "backward",
                      k = log(n))
```

```
## Start:  AIC=1648.81
## MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## - AGE      1      0.06 11079 1642.6
## - INDUS    1      2.52 11081 1642.7
## <none>                 11079 1648.8
## - CHAS     1    218.97 11298 1652.5
## - TAX      1    242.26 11321 1653.5
## - CRIM     1    243.22 11322 1653.6
## - ZN       1    257.49 11336 1654.2
## - B        1    270.63 11349 1654.8
## - RAD      1    479.15 11558 1664.0
## - NOX      1    487.16 11566 1664.4
## - PTRATIO  1   1194.23 12273 1694.4
## - DIS      1   1232.41 12311 1696.0
## - RM       1   1871.32 12950 1721.6
## - LSTAT    1   2410.84 13490 1742.2
##
## Step:  AIC=1642.59
## MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + DIS + RAD + TAX +
##     PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## - INDUS    1      2.52 11081 1636.5
## <none>                 11079 1642.6
## - CHAS     1    219.91 11299 1646.3
## - TAX      1    242.24 11321 1647.3
## - CRIM     1    243.20 11322 1647.3
## - ZN       1    260.32 11339 1648.1
## - B        1    272.26 11351 1648.7
## - RAD      1    481.09 11560 1657.9
## - NOX      1    520.87 11600 1659.6
## - PTRATIO  1   1200.23 12279 1688.4
## - DIS      1   1352.26 12431 1694.6
## - RM       1   1959.55 13038 1718.8
## - LSTAT    1   2718.88 13798 1747.4
##
## Step:  AIC=1636.48
## MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO +
##     B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1636.5
## - CHAS     1    227.21 11309 1640.5
## - CRIM     1    245.37 11327 1641.3
## - ZN       1    257.82 11339 1641.9
## - B        1    270.82 11352 1642.5
## - TAX      1    273.62 11355 1642.6
## - RAD      1    500.92 11582 1652.6
## - NOX      1    541.91 11623 1654.4
## - PTRATIO  1   1206.45 12288 1682.5
## - DIS      1   1448.94 12530 1692.4
## - RM       1   1963.66 13045 1712.8
## - LSTAT    1   2723.48 13805 1741.5
```

4.2 Perform hypothesis tests on coefficients

I will use the coefficients in the back selection model as the example.

```
par(mfrow = c(1,2))
plot(backward_model$fitted.values, backward_model$residuals,
     xlab = "fitted values", ylab = "residuals")
abline(h = 0, col = "red")
```



As we can see from the plot, heteroscedasticity is present in the model, which prompts us to use the robust standard errors for hypothesis testing of coefficients.

```
sandwich2 <- coeftest(backward_model, vcov = vcovHC(backward_model, type = 'HC3'))
sandwich2
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  36.3411450   8.1120182  4.4799 9.287e-06 ***
## CRIM         -0.1084133   0.0338469 -3.2030 0.0014476 **
## ZN            0.0458449   0.0138083  3.3201 0.0009664 ***
## CHAS1         2.7187163   1.3455538  2.0205 0.0438687 *
## NOX         -17.3760234   3.4185429 -5.0829 5.284e-07 ***
## RM            3.8015788   0.8218324  4.6257 4.774e-06 ***
## DIS          -1.4927115   0.2277446 -6.5543 1.409e-10 ***
## RAD           0.2996085   0.0635015  4.7181 3.103e-06 ***
## TAX          -0.0117780   0.0029349 -4.0130 6.924e-05 ***
## PTRATIO      -0.9465246   0.1152760 -8.2109 1.929e-15 ***
## B             0.0092908   0.0027637  3.3618 0.0008344 ***
## LSTAT        -0.5225535   0.0882650 -5.9203 6.020e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see from the summary above, all coefficients have a p-value smaller than 0.05. Therefore, we reject the null hypothesis that the coefficient is 0 for all of the coefficients in the backward_model.

4.3 Assess model performance with metrics (R^2, adjusted R^2, RMSE, etc.)

```
summary(forward_model)
```

```
##
## Call:
## lm(formula = MEDV ~ LSTAT + RM + PTRATIO + DIS + NOX + CHAS +
##     B + ZN, data = house_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6996  -2.7925  -0.5477   1.7005  27.6510
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.316950   4.870856   6.224 1.03e-09 ***
## LSTAT        -0.543125   0.047652 -11.398  < 2e-16 ***
## RM            4.116082   0.408594  10.074  < 2e-16 ***
## PTRATIO      -0.881851   0.115718  -7.621 1.29e-13 ***
## DIS          -1.382714   0.187604  -7.370 7.15e-13 ***
## NOX         -16.687428   3.228873  -5.168 3.43e-07 ***
## CHAS1         3.111062   0.870076   3.576 0.000384 ***
## B             0.009404   0.002639   3.563 0.000401 ***
## ZN            0.037808   0.013298   2.843 0.004652 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.847 on 497 degrees of freedom
## Multiple R-squared:  0.7266, Adjusted R-squared:  0.7222
## F-statistic: 165.1 on 8 and 497 DF,  p-value: < 2.2e-16
```

```
summary(backward_model)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT, data = house_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## CRIM         -0.108413   0.032779  -3.307 0.001010 **
## ZN            0.045845   0.013523   3.390 0.000754 ***
## CHAS1         2.718716   0.854240   3.183 0.001551 **
## NOX         -17.376023   3.535243  -4.915 1.21e-06 ***
## RM            3.801579   0.406316   9.356  < 2e-16 ***
## DIS          -1.492711   0.185731  -8.037 6.84e-15 ***
## RAD           0.299608   0.063402   4.726 3.00e-06 ***
## TAX          -0.011778   0.003372  -3.493 0.000521 ***
## PTRATIO      -0.946525   0.129066  -7.334 9.24e-13 ***
## B             0.009291   0.002674   3.475 0.000557 ***
## LSTAT        -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

```
AIC(forward_model)
```

```
## [1] 3044.275
```

```
AIC(backward_model)
```

```
## [1] 3023.726
```

```
BIC(forward_model)
```

```
## [1] 3086.54
```

```
BIC(backward_model)
```

```
## [1] 3078.671
```

Based on the statistics above, backward_model should be preferred, as it has a higher r-squared and adjusted r-squared as well as a lower AIC and BIC than forward_model.

4.4 Validate your model using appropriate cross-validation techniques

```
control <- trainControl(method = "cv", number = 10)
model_cv <- train(MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
    TAX + PTRATIO + B + LSTAT, data = house_price,
                  method = "lm",
                  trControl = control)
print(model_cv)
```

```
## Linear Regression
##
## 506 samples
##  11 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 456, 455, 455, 456, 455, 456, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.784618  0.7407132  3.3649
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

For the backward_model, the 10-fold cross-validated RMSE is 4.764, which means the model predicts median housing values (MEDV) with an average error of about $4,764. An R^2 of 0.743 suggests that about 74.3% of the variance in MEDV is explained by the predictors. The model generalizes relatively well to new data.

5. Feature Impact Analysis

5.1 Quantify and interpret the impact of each feature on the target

From ##4.2, we have the following:

```
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.34115    8.11202    4.48  9.3e-06 ***
CRIM         -0.10841    0.03385   -3.20  0.00145 **
ZN            0.04584    0.01381    3.32  0.00097 ***
CHAS1         2.71872    1.34555    2.02  0.04387 *
```

```
NOX         -17.37602    3.41854   -5.08  5.3e-07 ***
RM            3.80158    0.82183    4.63  4.8e-06 ***
DIS          -1.49271    0.22774   -6.55  1.4e-10 ***
RAD           0.29961    0.06350    4.72  3.1e-06 ***
TAX          -0.01178    0.00293   -4.01  6.9e-05 ***
PTRATIO      -0.94652    0.11528   -8.21  1.9e-15 ***
B             0.00929    0.00276    3.36  0.00083 ***
LSTAT        -0.52255    0.08827   -5.92  6.0e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Due to the statistical significance of all variables in the backward_model, I will pick two variables to interpret.

1) The estimate for NOX is -17.376, indicating that holding everything else constant, a 1-unit increase in nitric oxides concentration (parts per 10 million) is associated with a $17,376 decrease in the expected median value of owner-occupied homes.
2) The estimate for CHAS1 is 2.719, indicating that holding everything else constant, being a census tract bounding the Charles River is associated with a $2,719 increase in the expected median value of owner-occupied homes.

5.2 Provide confidence intervals for significant coefficients

Due to the statistical significance of all variables in the backward_model, I will calculate the confidence interval for two of the significant coefficients.

```
t_star = qt((1-0.95)/2, df = 494, lower = F)
b1 = sandwich2["NOX", "Estimate"]
se1 = sandwich2["NOX", "Std. Error"]
lb1 = b1 - t_star*se1
ub1 = b1 + t_star*se1
lb1
```

```
## [1] -24.0927
```

```
ub1
```

```
## [1] -10.65935
```

```
b2 = sandwich2["CHAS1", "Estimate"]
se2 = sandwich2["CHAS1", "Std. Error"]
lb2 = b2 - t_star*se2
ub2 = b2 + t_star*se2
lb2
```

```
## [1] 0.0750021
```

```
ub2
```

```
## [1] 5.362431
```

5.3 Explain the practical significance of your findings in the context of the dataset

In the first half of the project, I investigated the relationship between the target variable median value of owner-occupied homes and the predictors per capita crime rate, average number of rooms per dwelling, weighted distances to five Boston employment centres, and index of accessibility to radial highways. I found that crime rate and accessibility to highway are significantly negatively associated with home values, while number of rooms is significantly positively associated with home values. In the second half of the project, I employed forward selection and backward selection to pick out the best model to predict housing prices in Boston.

Deliverables GitHub Repository containing:

- All code (well-documented Rmd files)
- README.md with clear instructions on how to run your analysis
- Data folder (or instructions for accessing the data)
- Requirements.txt or environment.yml file

Final Report (PDF) containing:

- Introduction: dataset description and problem statement
- Methodology: techniques used and justification
- Results: findings from your analysis
- Discussion: interpretation of results and limitations
- Conclusion: summary and potential future work
- References: cite all sources used

Evaluation Criteria

Your project will be evaluated based on:

- Correctness of statistical analysis and procedures
- Proper handling of regression assumptions
- Quality of variable selection and hypothesis testing
- Clarity of interpretation and insights
- Organization and documentation of code
- Professional presentation of findings