Final Project

Rujia Xie

May 14, 2025

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)^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

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 Identification of missing values and outliers

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

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

1.2 Data cleaning and preprocessing steps

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

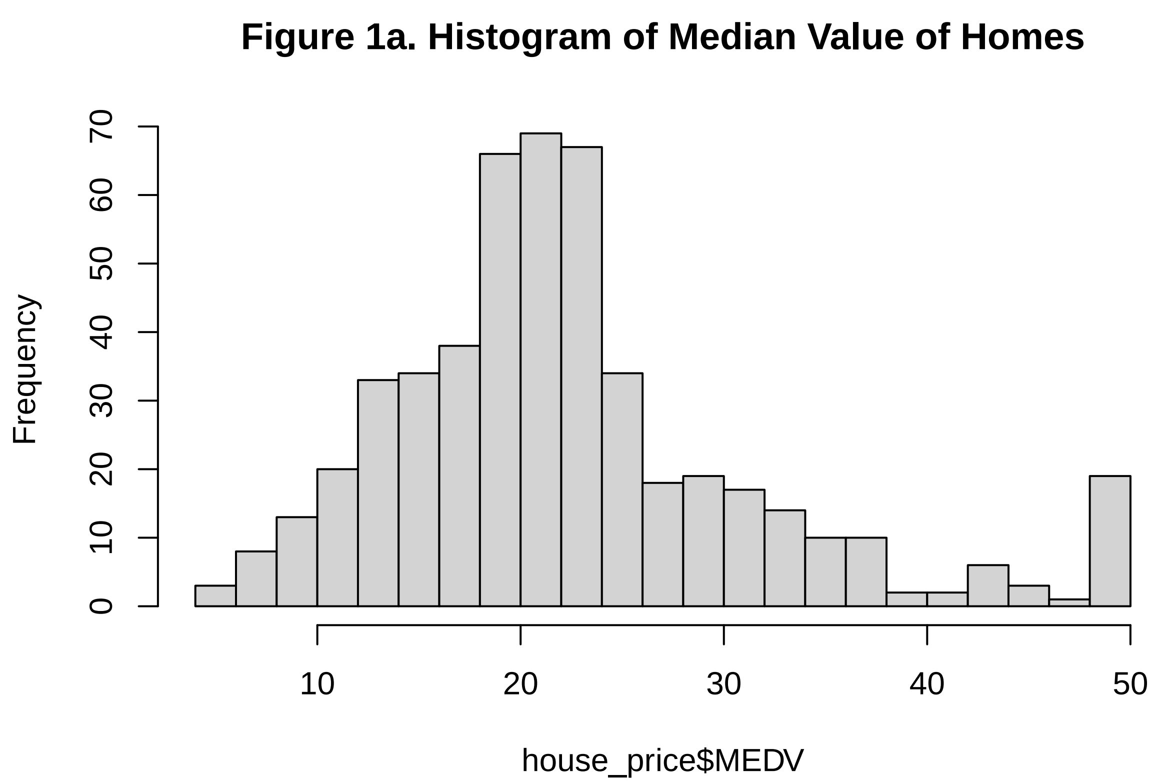
1.3 Summary statistics of variables

```
describe(house_price)
```

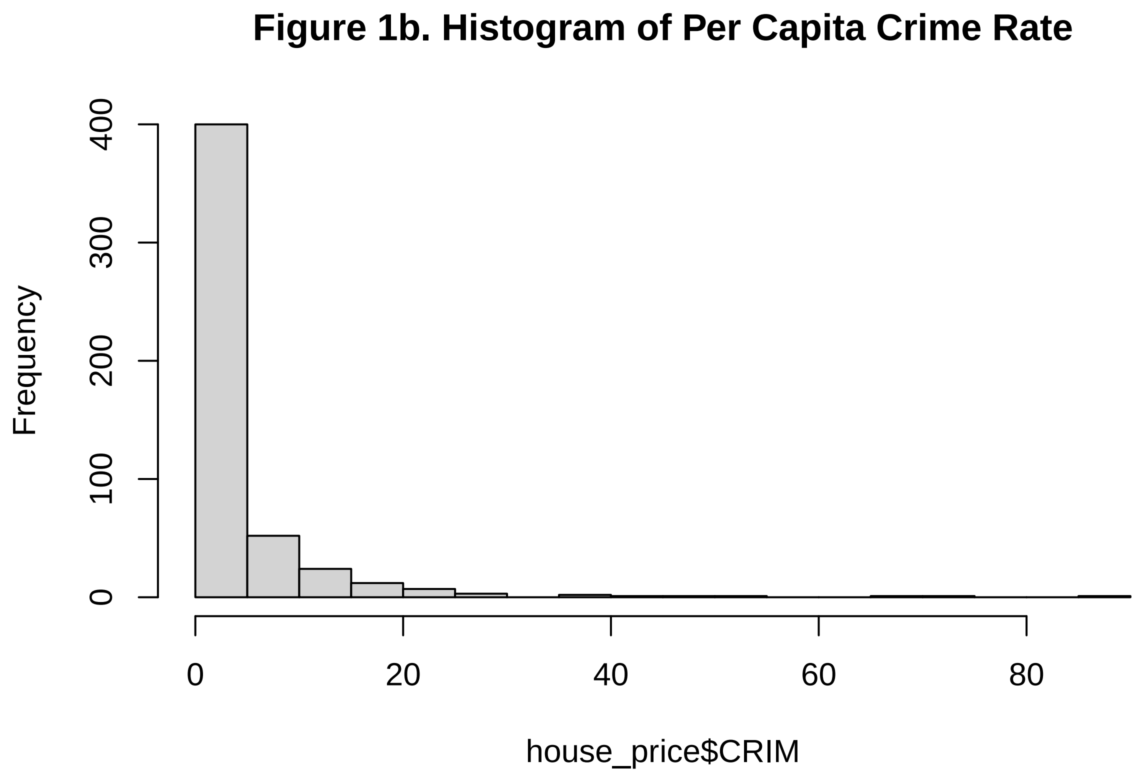
```
##         vars   n   mean     sd median trimmed    mad    min    max  range  skew
## CRIM       1 506   3.61   8.60   0.26    1.68   0.33   0.01  88.98  88.97  5.19
## ZN         2 506  11.36  23.32   0.00    5.08   0.00   0.00 100.00 100.00  2.21
## INDUS      3 506  11.14   6.86   9.69   10.93   9.37   0.46  27.74  27.28  0.29
## CHAS*      4 506   1.07   0.25   1.00    1.00   0.00   1.00   2.00   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
```

1.4 Visualization of distributions and relationships

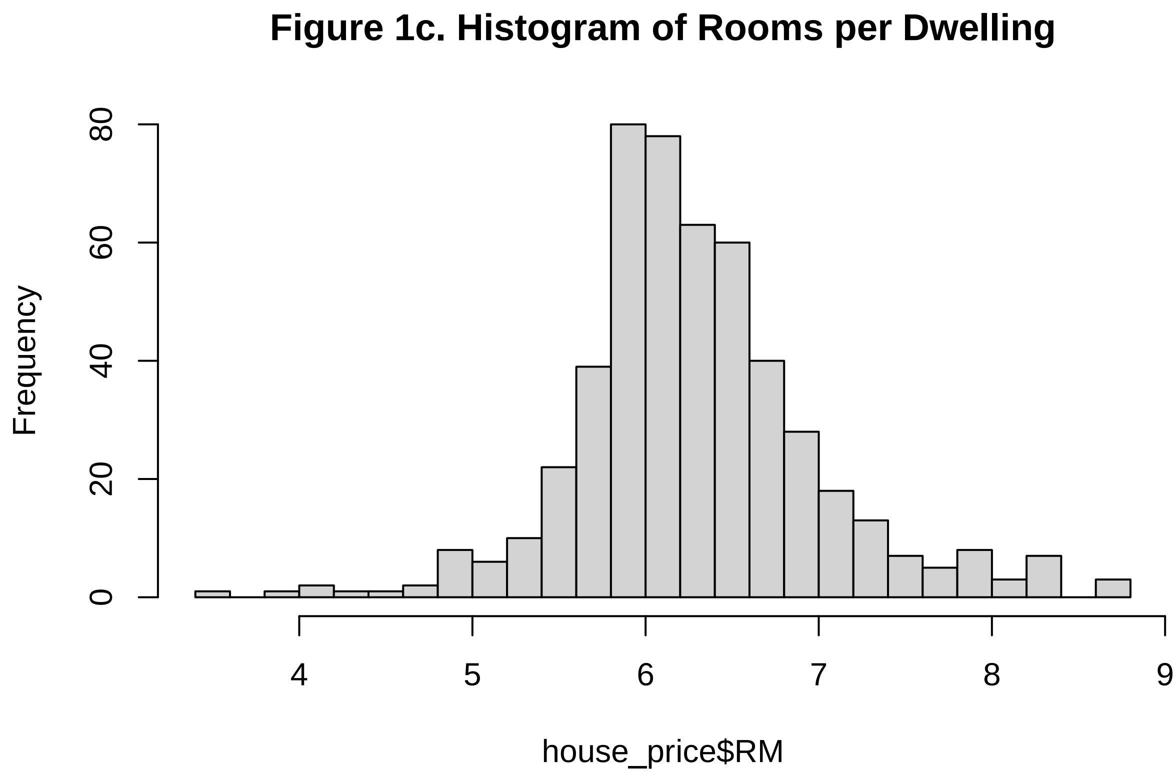
```r
hist(house_price$MEDV, breaks = 30, main = "Figure 1a. Histogram of Median Value of Homes")
```




```r
hist(house_price$CRIM, breaks = 30, main = "Figure 1b. Histogram of Per Capita Crime Rate")
```

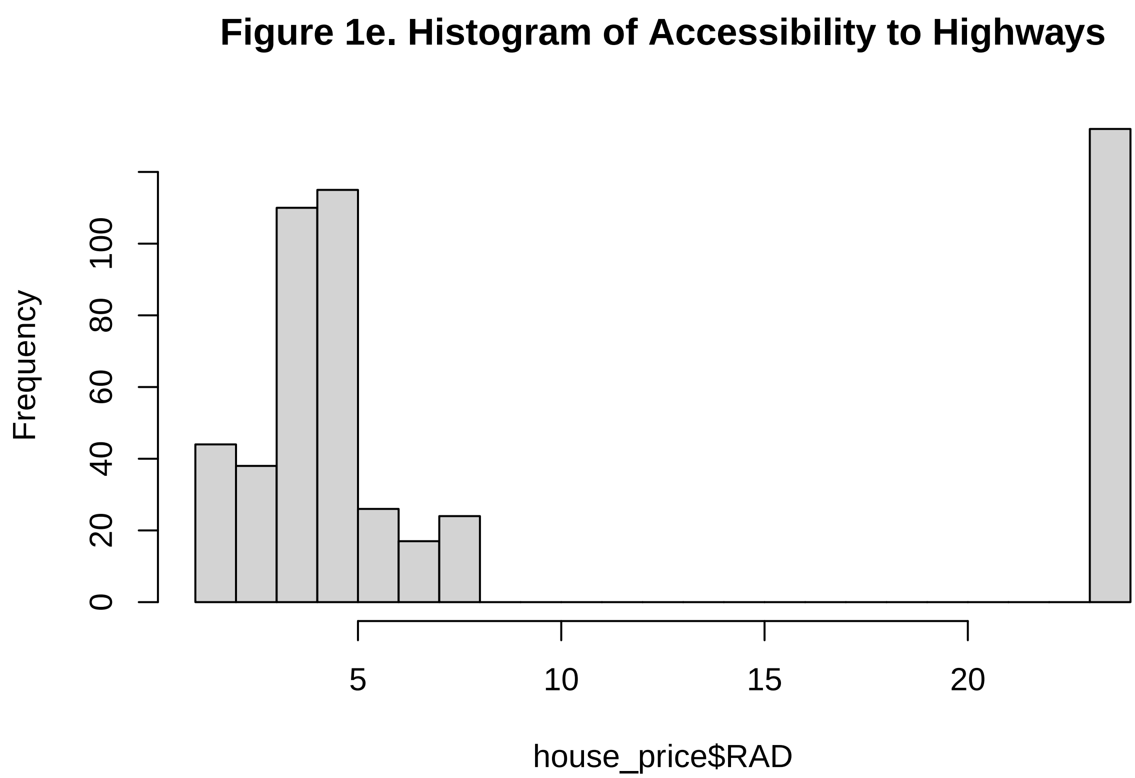



```
hist(house_price$RM, breaks = 30, main = "Figure 1c. Histogram of Rooms per Dwelling")
```




```
hist(house_price$DIS, breaks = 30, main = "Figure 1d. Histogram of Distance to Employment Centers")
```

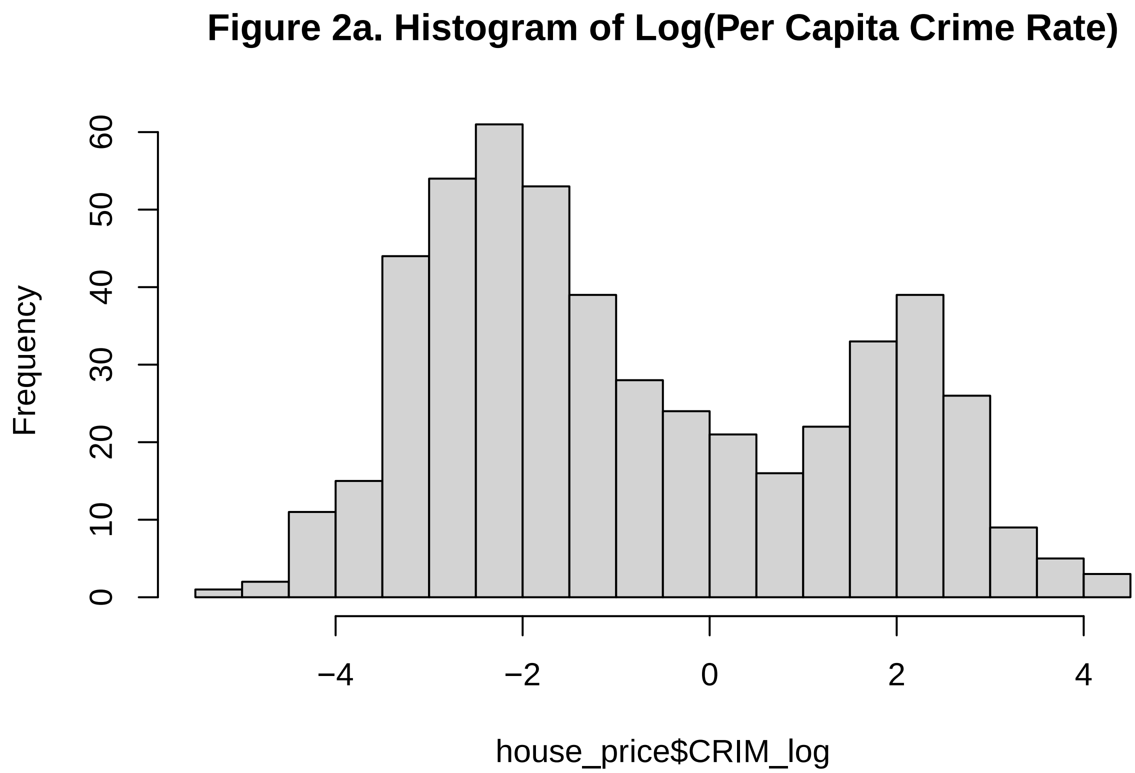

```r
hist(house_price$RAD, breaks = 30, main = "Figure 1e. Histogram of Accessibility to Highways")
```



Figures 1b and 1d indicate that CRIM and DIS are skewed to the right, which requires log transformation. Figure 1e presents the bimodal distribution of RAD, so it will be converted to a categorical variable with two levels: "low" (when $RAD \leq 15$) and "high" (when $RAD > 15$). Figures 2a-c present the distribution of the three transformed variables.

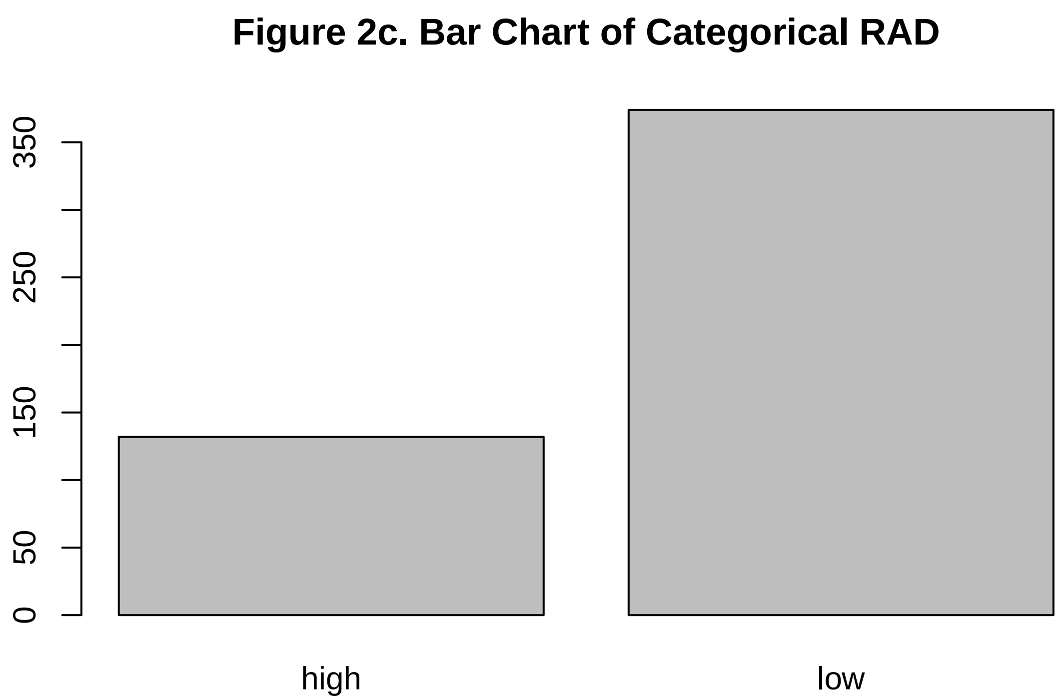
```
house_price$RAD_cat <- ifelse(house_price$RAD <= 15, "low", "high")
house_price$CRIM_log <- log(house_price$CRIM)
house_price$DIS_log <- log(house_price$DIS)

hist(house_price$CRIM_log, breaks = 30, main = "Figure 2a. Histogram of Log(Per Capita Crime Rate)")
```




```
hist(log(house_price$DIS), breaks = 30, main = "Figure 2b. Histogram of Log(Distance to Employment Cent
```

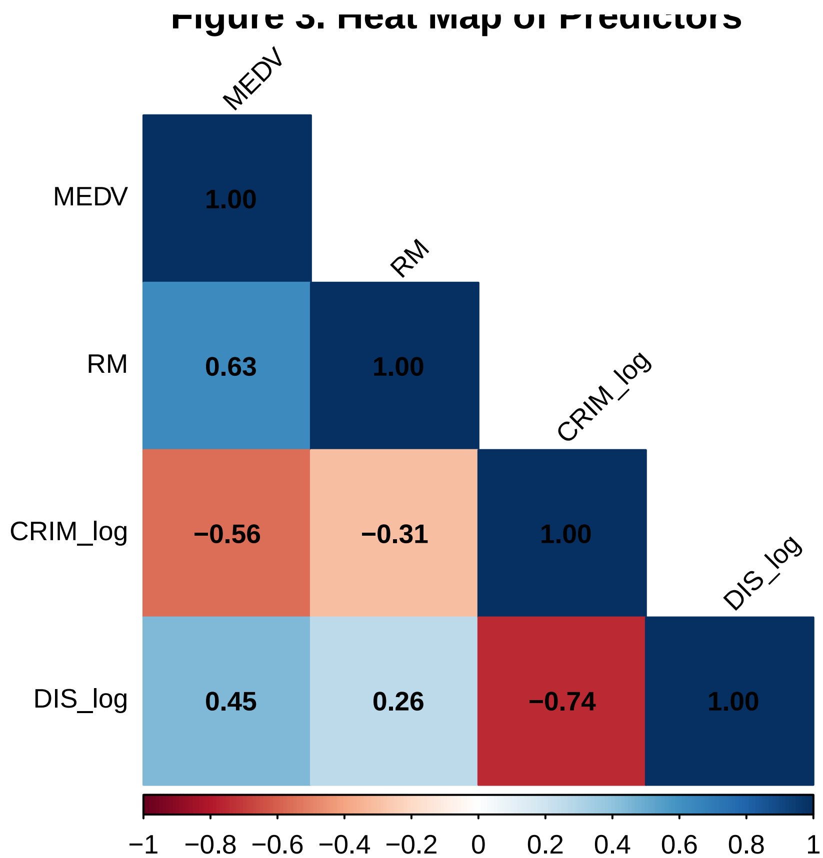

```
barplot(table(house_price$RAD_cat), main = "Figure 2c. Bar Chart of Categorical RAD")
```



Since there are multiple predictors in the model, a heat map was drawn to check for multicollinearity among the outcome and continuous predictors. 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(14, 6, 16, 17 )]
house_correlation <-cor(house_subset, method="spearman",
                           use="pairwise.complete.obs")


pmatrix <- cor.mtest(house_subset, conf.level = .95)
corrplot(title = "Figure 3. Heat Map of Predictors",
         house_correlation, p.mat = pmatrix$p, sig.level = .05,
         insig = "pch", pch.cex = 1.5, pch.col = "red", type = "lower",
         method="color", addCoef.col = "black", number.cex = 0.8,
         tl.cex = 0.8, tl.srt = 45, tl.col = "black")
```



As shown in Figure 3, CRIM_log and DIS_log are significantly correlated with a coefficient of -0.74. Thus, CRIM_log was removed from the model.

2. Regression Assumptions Verification

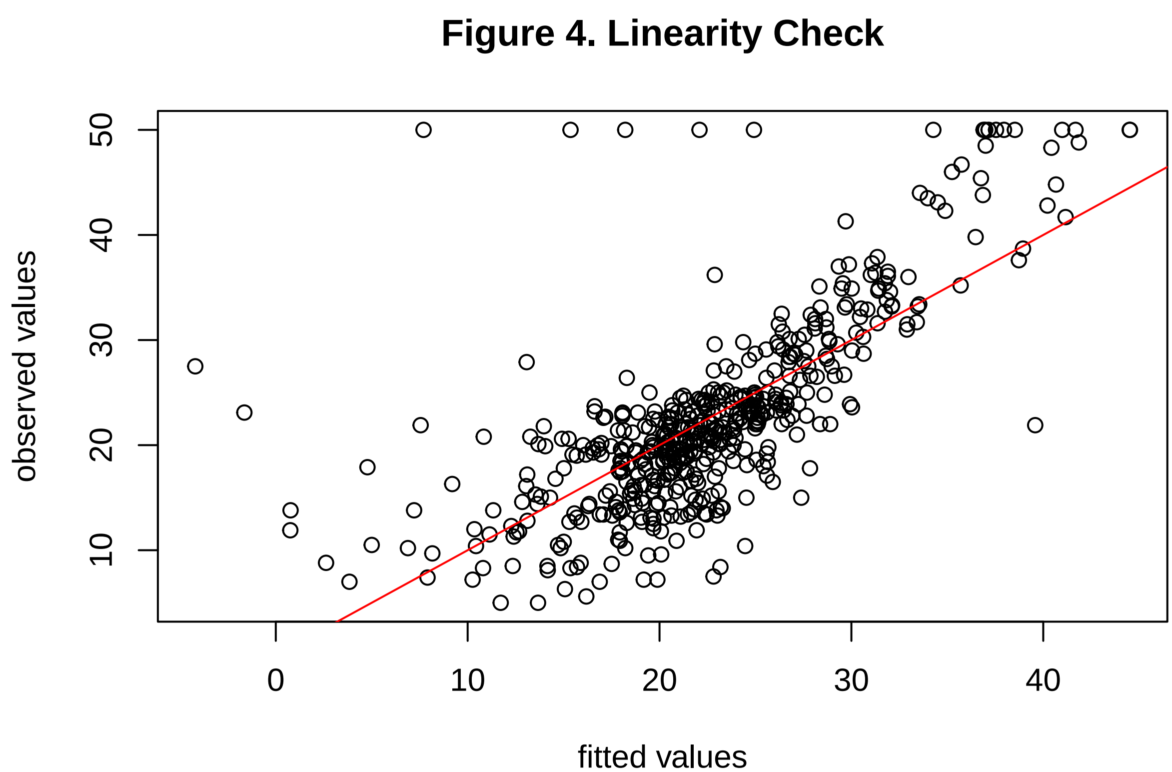
```r
# Run the fitted linear regression model
house_model <- lm(MEDV ~ RM + DIS_log + RAD_cat,
           data = house_price)
summary(house_model)
```

```
##
## Call:
## lm(formula = MEDV ~ RM + DIS_log + RAD_cat, data = house_price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.681  -3.055  -0.520   2.548  42.298
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.9432     2.4928 -13.616  < 2e-16 ***
## RM            8.3992     0.4096  20.507  < 2e-16 ***
## DIS_log      -0.3439     0.6266  -0.549    0.583
```

```
## RAD_catlow    5.5449     0.7625   7.272 1.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.22 on 502 degrees of freedom
## Multiple R-squared:  0.5453, Adjusted R-squared:  0.5426
## F-statistic: 200.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

2.1 Linearity assessment

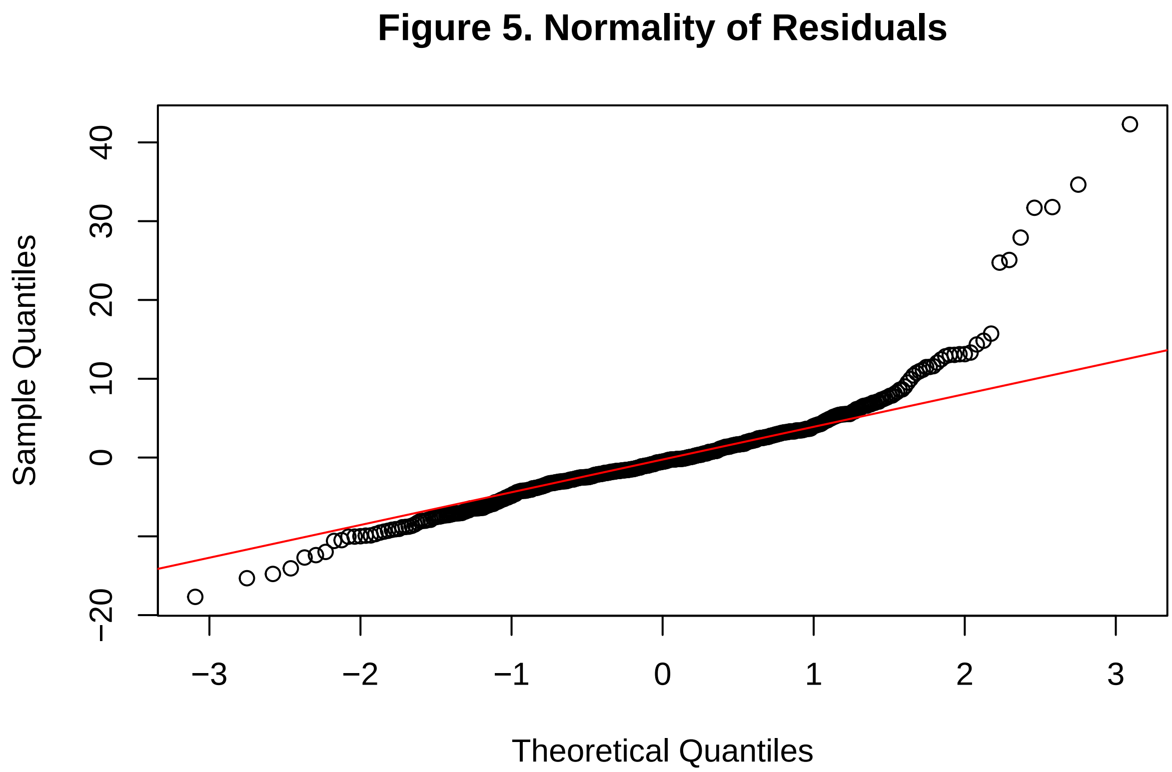
```r
plot(house_model$fitted.values, house_price$MEDV,
     xlab = "fitted values", ylab = "observed values",
     main = "Figure 4. Linearity Check")
abline(a = 0, b = 1, col = "red")
```



Majority of the observations follow a linear trend.

2.2 Normality of residuals

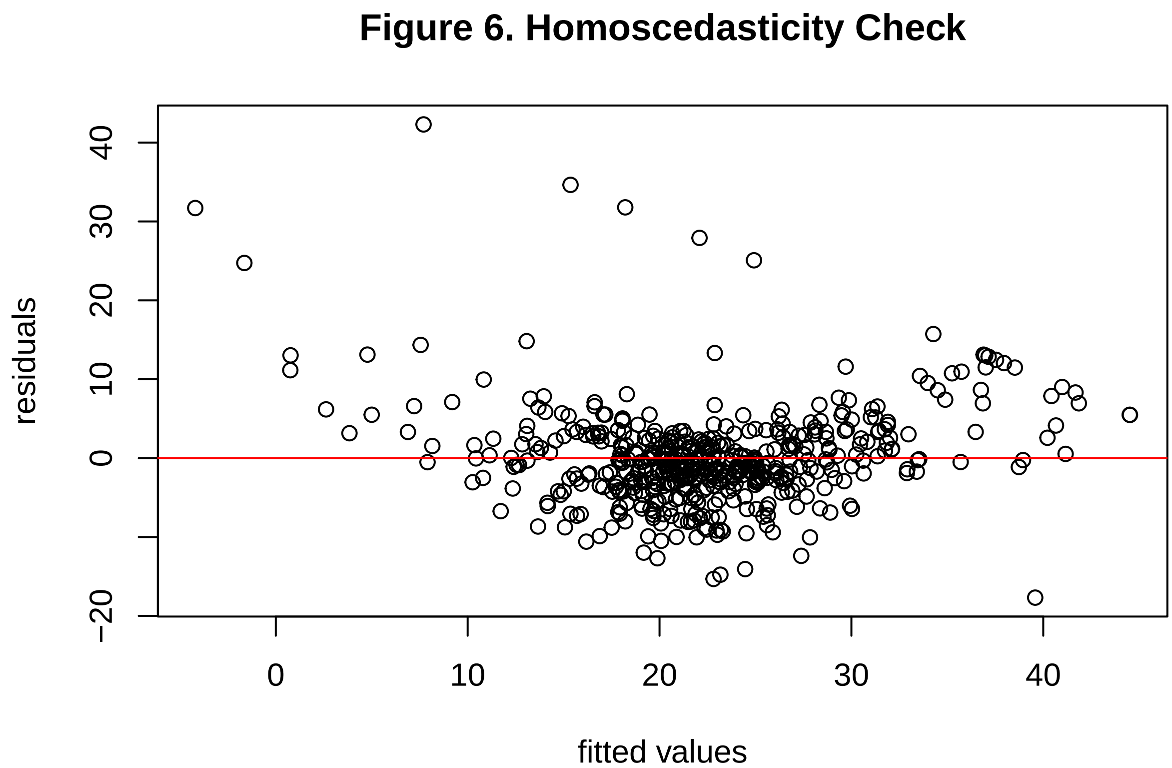
```r
qqnorm(house_model$residuals, main = "Figure 5. Normality of Residuals")
qqline(house_model$residuals, col = "red")
```



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

2.3 Check homoscedasticity (constant variance of residuals)

```
plot(house_model$fitted.values, house_model$residuals,
     xlab = "fitted values", ylab = "residuals",
     main = "Figure 6. Homoscedasticity Check")
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)
```

```
##       RM  DIS_log  RAD_cat
## 1.080962 1.492040 1.466173
```

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 heteroscedasticity-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) -33.94324    4.29995 -7.8939 1.855e-14 ***
## RM            8.39924    0.68418 12.2764 < 2.2e-16 ***
## DIS_log      -0.34386    0.59922 -0.5739    0.5663
## RAD_catlow    5.54493    0.78050  7.1043 4.167e-12 ***
## ---
## 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:

```
summary(house_model)
```

```
##
## Call:
## lm(formula = MEDV ~ RM + DIS_log + RAD_cat, data = house_price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.681  -3.055  -0.520   2.548  42.298
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.9432     2.4928 -13.616  < 2e-16 ***
## RM            8.3992     0.4096  20.507  < 2e-16 ***
## DIS_log      -0.3439     0.6266  -0.549    0.583
## RAD_catlow    5.5449     0.7625   7.272 1.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.22 on 502 degrees of freedom
## Multiple R-squared:  0.5453, Adjusted R-squared:  0.5426
## F-statistic: 200.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

This is the model summary after correction:

```
sandwich1
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -33.94324    4.29995 -7.8939 1.855e-14 ***
## RM            8.39924    0.68418 12.2764 < 2.2e-16 ***
## DIS_log      -0.34386    0.59922 -0.5739    0.5663
## RAD_catlow    5.54493    0.78050  7.1043 4.167e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In other words, after correction, RAD becomes less significant (but still significant). RM's standard error (SE) increases from 0.41 to 0.68, while DIS_log's SE decreases from 0.63 to 0.60.

4. Variable Selection & Hypothesis Testing

4.1 Implement at least two different variable selection techniques

Step-wise regression was employed to find the model that accurately predicts the median value of owner-occupied homes (MEDV) in Boston, MA. Both forward and backward selection were conducted. We subset the house_price dataset so that the raw forms of the three transformed variables will not be included in the model.

```
# Forward selection
null <- lm(MEDV ~ 1, data = house_price[-c(1, 8, 9)])
full <- lm(MEDV ~ ., data = house_price[-c(1, 8, 9)])
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
## + CRIM_log  1    8816.2 33900 2140.0
## + NOX       1    7800.1 34916 2154.9
## + RAD_cat   1    6708.6 36008 2170.5
## + AGE       1    6069.8 36647 2179.4
## + ZN        1    5549.7 37167 2186.6
## + B         1    4749.9 37966 2197.3
## + DIS_log   1    3650.0 39066 2211.8
```

```
## + 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
## + DIS_log   1     946.4 18526 1840.5
## + CHAS      1     786.3 18686 1844.8
## + 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
## + INDUS     1      98.7 19374 1863.1
## + RAD_cat   1      54.2 19418 1864.3
## + NOX       1       4.8 19468 1865.6
## + CRIM_log  1       4.4 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
## + DIS_log   1    497.64 14942 1737.9
## + TAX       1    425.16 15014 1740.3
## + RAD_cat   1    226.46 15213 1747.0
## <none>                  15439 1748.3
## + INDUS     1     61.09 15378 1752.5
## + ZN        1     56.56 15383 1752.6
## + CRIM_log  1     35.56 15404 1753.3
## + 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_log   1    618.57 13109 1677.9
## + B         1    389.68 13338 1686.7
## + CHAS      1    377.96 13350 1687.1
## <none>                  13728 1695.0
## + AGE       1     66.24 13662 1698.8
## + TAX       1     44.36 13684 1699.6
## + NOX       1     24.81 13703 1700.3
## + CRIM_log  1     17.38 13711 1700.6
## + ZN        1     14.96 13713 1700.7
## + RAD_cat   1      1.85 13726 1701.2
## + INDUS     1      0.83 13727 1701.2
##
```

```
## Step:  AIC=1677.93
## MEDV ~ LSTAT + RM + PTRATIO + DIS_log
##
##            Df Sum of Sq   RSS    AIC
## + NOX       1   1273.51 11836 1632.5
## + B         1    543.66 12566 1662.7
## + INDUS     1    411.14 12698 1668.0
## + TAX       1    408.23 12701 1668.2
## + CHAS      1    259.24 12850 1674.0
## + CRIM_log  1    186.37 12923 1676.9
## <none>                  13109 1677.9
## + AGE       1    117.41 12992 1679.6
## + ZN        1     95.21 13014 1680.5
## + RAD_cat   1     80.12 13029 1681.1
##
## Step:  AIC=1632.45
## MEDV ~ LSTAT + RM + PTRATIO + DIS_log + NOX
##
##            Df Sum of Sq   RSS    AIC
## + CHAS      1    334.18 11502 1624.2
## + B         1    309.86 11526 1625.2
## <none>                  11836 1632.5
## + INDUS     1     75.47 11760 1635.4
## + TAX       1     62.33 11774 1636.0
## + ZN        1     60.08 11776 1636.1
## + AGE       1     19.04 11817 1637.9
## + RAD_cat   1      7.97 11828 1638.3
## + CRIM_log  1      6.61 11829 1638.4
##
## Step:  AIC=1624.19
## MEDV ~ LSTAT + RM + PTRATIO + DIS_log + NOX + CHAS
##
##            Df Sum of Sq   RSS    AIC
## + B         1   270.108 11232 1618.4
## <none>                  11502 1624.2
## + INDUS     1    95.074 11407 1626.2
## + ZN        1    72.904 11429 1627.2
## + TAX       1    43.377 11458 1628.5
## + AGE       1    30.690 11471 1629.1
## + RAD_cat   1    12.148 11490 1629.9
## + CRIM_log  1     7.664 11494 1630.1
##
## Step:  AIC=1618.39
## MEDV ~ LSTAT + RM + PTRATIO + DIS_log + NOX + CHAS + B
##
##            Df Sum of Sq   RSS    AIC
## <none>                  11232 1618.4
## + ZN        1    91.341 11140 1620.5
## + INDUS     1    73.478 11158 1621.3
## + RAD_cat   1    61.356 11170 1621.8
## + CRIM_log  1    54.923 11177 1622.1
## + AGE       1    50.965 11181 1622.3
## + TAX       1     8.756 11223 1624.2
```

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

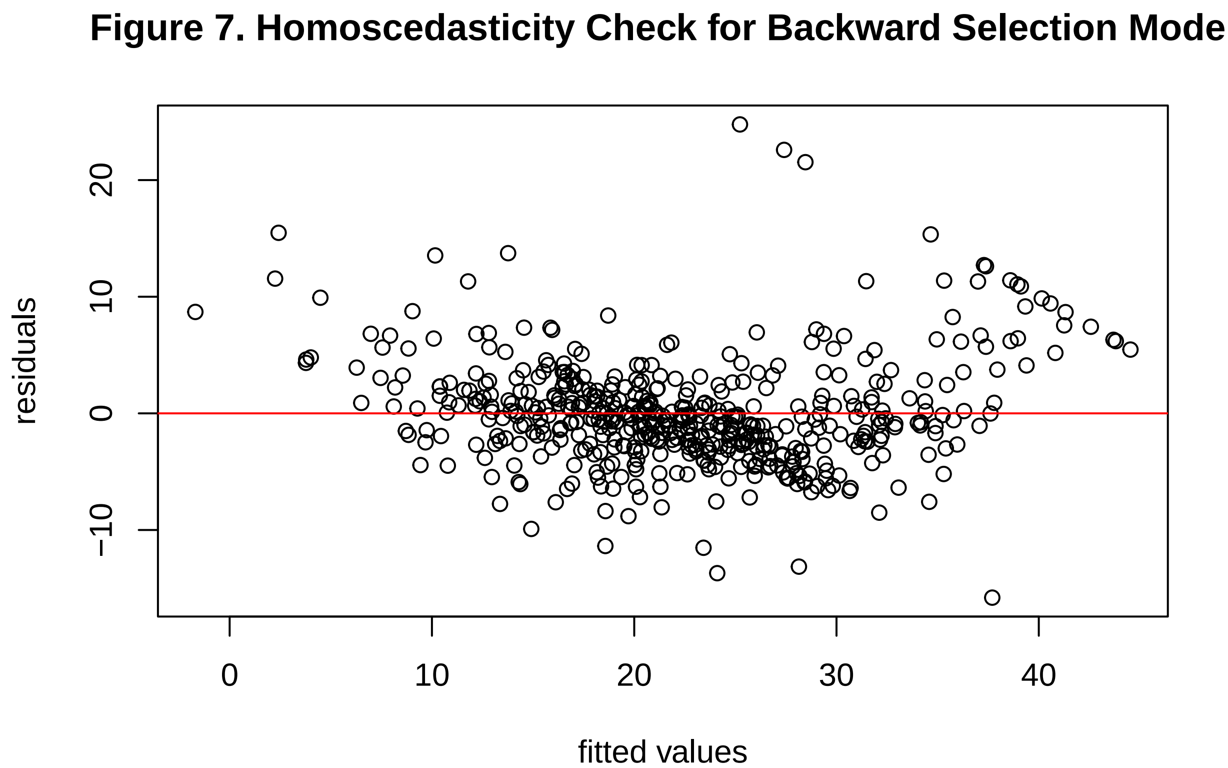
```
## Start:  AIC=1633.14
## MEDV ~ ZN + INDUS + CHAS + NOX + RM + AGE + TAX + PTRATIO + B +
##     LSTAT + RAD_cat + CRIM_log + DIS_log
##
##            Df Sum of Sq   RSS    AIC
## - INDUS     1      8.91 10750 1627.3
## - AGE       1     27.79 10769 1628.2
## - CRIM_log  1     57.90 10799 1629.6
## - RAD_cat   1    101.01 10842 1631.7
## - ZN        1    112.82 10854 1632.2
## <none>                  10741 1633.1
## - TAX       1    221.28 10962 1637.2
## - CHAS      1    278.14 11019 1639.8
## - B         1    330.88 11072 1642.3
## - NOX       1    742.61 11484 1660.8
## - PTRATIO   1   1018.43 11759 1672.8
## - DIS_log   1   1680.82 12422 1700.5
## - RM        1   2103.89 12845 1717.4
## - LSTAT     1   2858.32 13599 1746.3
##
## Step:  AIC=1627.34
## MEDV ~ ZN + CHAS + NOX + RM + AGE + TAX + PTRATIO + B + LSTAT +
##     RAD_cat + CRIM_log + DIS_log
##
##            Df Sum of Sq   RSS    AIC
## - AGE       1     27.18 10777 1622.4
## - CRIM_log  1     57.24 10807 1623.8
## - RAD_cat   1    120.66 10870 1626.8
## - ZN        1    122.02 10872 1626.8
## <none>                  10750 1627.3
## - CHAS      1    271.16 11021 1633.7
## - TAX       1    300.77 11051 1635.1
## - B         1    334.03 11084 1636.6
## - NOX       1    811.45 11561 1657.9
## - PTRATIO   1   1064.41 11814 1668.9
## - DIS_log   1   1724.89 12475 1696.4
## - RM        1   2152.16 12902 1713.5
## - LSTAT     1   2887.27 13637 1741.5
##
## Step:  AIC=1622.39
## MEDV ~ ZN + CHAS + NOX + RM + TAX + PTRATIO + B + LSTAT + RAD_cat +
##     CRIM_log + DIS_log
##
##            Df Sum of Sq   RSS    AIC
## - CRIM_log  1      50.5 10828 1618.5
## <none>                  10777 1622.4
## - RAD_cat   1     136.1 10913 1622.5
```

```
## - ZN         1      140.4 10917 1622.7
## - CHAS       1      263.5 11040 1628.4
## - TAX        1      301.3 11078 1630.1
## - B          1      322.7 11100 1631.1
## - NOX        1      891.9 11669 1656.4
## - PTRATIO    1     1107.6 11885 1665.7
## - DIS_log    1     1787.6 12565 1693.8
## - RM         1     2149.3 12926 1708.2
## - LSTAT      1     3399.0 14176 1754.9
##
## Step:  AIC=1618.53
## MEDV ~ ZN + CHAS + NOX + RM + TAX + PTRATIO + B + LSTAT + RAD_cat +
##     DIS_log
##
##           Df Sum of Sq   RSS    AIC
## - ZN       1     104.4 10932 1617.2
## <none>                 10828 1618.5
## - CHAS     1     269.2 11097 1624.7
## - RAD_cat  1     277.9 11105 1625.1
## - TAX      1     281.0 11108 1625.3
## - B        1     293.9 11121 1625.8
## - NOX      1     842.4 11670 1650.2
## - PTRATIO  1    1162.1 11990 1663.9
## - DIS_log  1    1879.1 12707 1693.3
## - RM       1    2176.7 13004 1705.0
## - LSTAT    1    3360.3 14188 1749.1
##
## Step:  AIC=1617.15
## MEDV ~ CHAS + NOX + RM + TAX + PTRATIO + B + LSTAT + RAD_cat +
##     DIS_log
##
##           Df Sum of Sq   RSS    AIC
## <none>                 10932 1617.2
## - TAX      1     238.4 11170 1621.8
## - CHAS     1     261.2 11193 1622.9
## - RAD_cat  1     291.0 11223 1624.2
## - B        1     297.0 11229 1624.5
## - NOX      1     965.3 11897 1653.8
## - PTRATIO  1    1675.8 12608 1683.1
## - DIS_log  1    1801.1 12733 1688.1
## - RM       1    2285.7 13218 1707.0
## - LSTAT    1    3344.5 14276 1746.0
```

4.2 Perform hypothesis tests on coefficients

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

```r
plot(backward_model$fitted.values, backward_model$residuals,
     xlab = "fitted values", ylab = "residuals",
     main = "Figure 7. Homoscedasticity Check for Backward Selection Model")
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)  48.8179345   9.6956248  5.0350 6.697e-07 ***
## CHAS1         2.9056484   1.2285482  2.3651 0.0184088 *
## NOX         -24.2413924   4.1899195 -5.7856 1.281e-08 ***
## RM            4.0349518   0.7827948  5.1545 3.678e-07 ***
## TAX          -0.0107518   0.0026806 -4.0110 6.978e-05 ***
## PTRATIO      -1.0434531   0.1137522 -9.1730 < 2.2e-16 ***
## B             0.0096793   0.0029097  3.3266 0.0009443 ***
## LSTAT        -0.5752277   0.0911076 -6.3137 6.055e-10 ***
## RAD_catlow   -4.3153415   1.1635475 -3.7088 0.0002318 ***
## DIS_log      -6.5110159   1.0414524 -6.2519 8.754e-10 ***
## ---
## 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 (R², adjusted R², RMSE, etc.)

```r
summary(forward_model)
```

```
##
## Call:
## lm(formula = MEDV ~ LSTAT + RM + PTRATIO + DIS_log + NOX + CHAS +
##     B, data = house_price[-c(1, 8, 9)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0323  -2.7068  -0.5117   1.9253  25.5948
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.922872   5.027419   7.742 5.50e-14 ***
## LSTAT        -0.569960   0.047014 -12.123  < 2e-16 ***
## RM            4.253108   0.394770  10.774  < 2e-16 ***
## PTRATIO      -0.982716   0.108549  -9.053  < 2e-16 ***
## DIS_log      -6.370602   0.725305  -8.783  < 2e-16 ***
## NOX         -24.519459   3.489253  -7.027 6.97e-12 ***
## CHAS1         3.076277   0.851425   3.613 0.000333 ***
## B             0.008931   0.002581   3.461 0.000585 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.749 on 498 degrees of freedom
## Multiple R-squared:  0.7371, Adjusted R-squared:  0.7334
## F-statistic: 199.4 on 7 and 498 DF,  p-value: < 2.2e-16
```

```r
summary(backward_model)
```

```
##
## Call:
## lm(formula = MEDV ~ CHAS + NOX + RM + TAX + PTRATIO + B + LSTAT +
##     RAD_cat + DIS_log, data = house_price[-c(1, 8, 9)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8003  -2.7006  -0.4858   2.0582  24.7727
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.817935   5.760455   8.475 2.71e-16 ***
## CHAS1         2.905648   0.844026   3.443 0.000625 ***
## NOX         -24.241392   3.662909  -6.618 9.46e-11 ***
## RM            4.034952   0.396222  10.184  < 2e-16 ***
## TAX          -0.010752   0.003269  -3.289 0.001078 **
## PTRATIO      -1.043453   0.119667  -8.720  < 2e-16 ***
## B             0.009679   0.002637   3.671 0.000268 ***
## LSTAT        -0.575228   0.046696 -12.319  < 2e-16 ***
## RAD_catlow   -4.315342   1.187658  -3.633 0.000309 ***
```

```
## DIS_log      -6.511016   0.720253  -9.040  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.695 on 496 degrees of freedom
## Multiple R-squared:  0.7441, Adjusted R-squared:  0.7394
## F-statistic: 160.2 on 9 and 496 DF,  p-value: < 2.2e-16
```

```r
AIC(forward_model)
```

```
## [1] 3022.542
```

```r
AIC(backward_model)
```

```
## [1] 3012.855
```

```r
BIC(forward_model)
```

```
## [1] 3060.581
```

```r
BIC(backward_model)
```

```
## [1] 3059.347
```

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

```r
control <- trainControl(method = "cv", number = 10)
model_backward_cv <- train(MEDV ~ CHAS + NOX + RM + TAX + PTRATIO + B + LSTAT +
    RAD_cat + DIS_log, data = house_price[-c(1, 8, 9)],
                  method = "lm",
                  trControl = control)
print(model_backward_cv)
```

```
## Linear Regression
##
## 506 samples
##   9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 456, 455, 456, 456, 455, 455, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.759766  0.7369261  3.409284
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

For the backward_model, the 10-fold cross-validated RMSE is 4.729, which means the model predicts median housing values (MEDV) with an average error of about \$4,729. An R^2 of 0.738 suggests that about 73.8% 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

For the linear regression model generated for the first objective, we have the following coefficients and interpretations:

- Intercept: -33.9432 ($p < 0.001$)

  The expected median value of owner-occupied homes when RM = 0, DIS_log = 0, and RAD_cat = "high" is -\$33,943.2, which serves as a baseline value.

- RM: 8.3992 ($p < 0.001$)

  Holding DIS_log and RAD_cat constant, a 1-unit increase in the average number of rooms per dwelling is associated with a \$8,399.2 increase in the expected median value of owner-occupied homes. This association is statistically significant.

- DIS_log: -0.3439 ($p = 0.57$)

  The log of weighted distance to five Boston employment centers shows a non-significant effect on the expected median value of owner-occupied homes.

- RAD_cat: 5.5449 ($p < 0.001$)

  Holding RM and DIS_log constant, a low index of accessibility to radial highways ($RAD \leq 15$) is associated with a \$5,544.9 increase in the expected median value of owner-occupied homes compared to a high index of accessibility to radial highways ($RAD > 15$). This association is statistically significant.

For the predictive model generated for the second objective, we have the following coefficients and interpretations:

- Intercept: 48.8179 ($p < 0.001$)

  The expected median value of owner-occupied homes (in \$1,000) when all predictors = 0, which serves as a baseline value.

- CHAS: 2.9056 ($p = 0.02$)

  Holding everything else constant, being near the Charles River (CHAS = 1) is associated with a \$2,905.6 increase in the expected median value of owner-occupied homes. This association is statistically significant.

- NOX: -24.2414 ($p < 0.001$)

  Holding everything else constant, a 1-unit increase in the Nitric Oxide concentration (parts per 10 million) is associated with a \$24,241.4 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

- RM: 4.0350 ($p < 0.001$)

  Holding everything else constant, a 1-unit increase in the average number of rooms per dwelling is associated with a \$4,035.0 increase in the expected median value of owner-occupied homes. This association is statistically significant.

- TAX: -0.0108 (p < 0.001)

  Holding everything else constant, a 1-unit increase in the full-value property-tax rate per $10,000 is associated with a $10.8 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

- PTRATIO: -1.0435 (p < 0.001)

  Holding everything else constant, a 1-unit increase in the pupil-teacher ratio is associated with a $1,043.5 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

- B: 0.0097 (p < 0.001)

  Holding everything else constant, a 1-unit increase in the parabolically transformed value of African American resident proportion (the formula is 1000(Bk - 0.63)^2, where Bk is the proportion of African American residents) is associated with a $9.7 increase in the expected median value of owner-occupied homes. This association is statistically significant.

- LSTAT: -0.5752 (p < 0.001)

  Holding everything else constant, a 1-unit increase in the percentage of lower status of the population is associated with a $575.2.0 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

- RAD_cat: -4.3153 (p < 0.001)

  Holding everything else constant, a low index of accessibility to radial highways ($RAD \leq 15$) is associated with a $4,315.3 decrease in the expected median value of owner-occupied homes compared to a high index of accessibility to radial highways ($RAD > 15$). This association is statistically significant.

- DIS_log: -6.5110 (p < 0.001)

  Holding everything else constant, a 1% increase in the weighted distance to five Boston employment centers is associated with a $65.11 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

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.

```r
t_star = qt((1-0.95)/2, df = 496, lower = F)
b1 = sandwich2["NOX", "Estimate"]
se1 = sandwich2["NOX", "Std. Error"]
lb1 = b1 - t_star*se1
ub1 = b1 + t_star*se1
lb1
```

```
## [1] -32.47357
```

```r
ub1
```

```
## [1] -16.00921
```

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

```
## [1] 0.4918481
```

```r
ub2
```

```
## [1] 5.319449
```

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

This project does two things. First, it investigated the relationship between the median value of owner-occupied homes (MEDV) and the average number of rooms per dwelling (RM), the weighted distance to five Boston employment centers (DIS), and the index of accessibility to radial highways (RAD). Then, the project developed a predictive model that estimates the median value of owner-occupied homes (MEDV) in Boston, MA, using step-wise regression.

According to the first model, the median value of homes is positively associated with the number of rooms and negatively associated with accessibility. In the more comprehensive predictive model, the median value of homes is higher near the Charles River, with more rooms, and slightly with a higher parabolically transformed value of African American resident proportion. Conversely, home values decline with increased NOx pollution, higher property taxes, greater pupil-teacher ratios, a higher percentage of lower-status residents, and greater (log-transformed) distances from employment centers. While these findings offer insights into Boston's housing dynamics for policymakers and real estate professionals, it's critical that such models are applied thoughtfully to avoid reinforcing systemic inequities.

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