DANA4810-Project

Patricia Tating, Milkah Nyaingo, Bemba Munkhchuluun & Maryam Gadimova

2024-04-05

# STUDY OBJECTIVES

1. What factors have the most significant impact on the median value of homes in the Boston area?
2. Which predictive model most accurately forecasts housing prices based on the available dataset features?

# VARIABLES

* CRIM - Per capita crime rate by town - Numerical
* ZN - Proportion of residential land zoned for lots over 25,000 sq.ft. - Numerical
* INDUS - Proportion of non-retail business acres per town - Numerical
* CHAS - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) - Categorical
* NOX - Nitric oxides concentration (parts per 10 million) - Numerical
* RM - Average number of rooms per dwelling - Numerical
* AGE - Proportion of owner-occupied units built prior to 1940 - Numerical
* DIS - Weighted distances to five Boston employment centres - Numerical
* RAD - Index of accessibility to radial highways - Numerical
* TAX - Full-value property-tax rate per $10,000 - Numerical
* PTRATIO - Pupil-teacher ratio by town - Numerical
* B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town - Numerical
* LSTAT - % lower status of the population - Numerical
* PRICE - Median value of owner-occupied homes in $1000’s - Numerical

# READING THE DATA

bhouse <- read.csv("Boston\_housing.csv", header=TRUE)  
View(bhouse)

**Splitting data in two datasets**

set.seed(100) # setting seed to reproduce results of random sampling  
trainingRowIndex <- sample(1:nrow(bhouse), 0.90\*nrow(bhouse)) # row indices for training data  
training <- bhouse[trainingRowIndex, ] # model training data  
testing <- bhouse[-trainingRowIndex, ] # test data

dim(training)

## [1] 455 14

dim(testing)

## [1] 51 14

# EXPLORATORY DATA ANALYSIS

**Plots and Summaries**

sum(is.na(training))

## [1] 0

# Remove rows with missing values  
training <- na.omit(training)

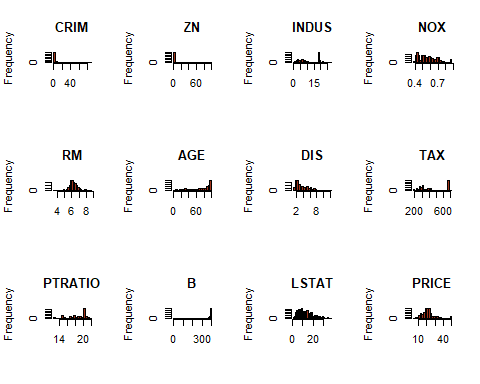
# Summary statistics for numerical variables  
summary(training)

## CRIM ZN INDUS CHAS   
## Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000   
## 1st Qu.: 0.08254 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000   
## Median : 0.25915 Median : 0.00 Median : 9.90 Median :0.00000   
## Mean : 3.62385 Mean : 11.08 Mean :11.35 Mean :0.06593   
## 3rd Qu.: 3.69503 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. :0.3890 Min. :3.561 Min. : 2.90 Min. : 1.130   
## 1st Qu.:0.4490 1st Qu.:5.885 1st Qu.: 45.25 1st Qu.: 2.102   
## Median :0.5380 Median :6.193 Median : 77.70 Median : 3.103   
## Mean :0.5576 Mean :6.270 Mean : 68.75 Mean : 3.763   
## 3rd Qu.:0.6275 3rd Qu.:6.594 3rd Qu.: 94.10 3rd Qu.: 5.165   
## 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.:281.0 1st Qu.:17.00 1st Qu.:375.43   
## Median : 5.000 Median :337.0 Median :19.10 Median :391.98   
## Mean : 9.571 Mean :411.5 Mean :18.44 Mean :356.97   
## 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.36   
## Max. :24.000 Max. :711.0 Max. :22.00 Max. :396.90   
## LSTAT PRICE   
## Min. : 1.73 Min. : 5.00   
## 1st Qu.: 7.19 1st Qu.:16.75   
## Median :11.50 Median :21.00   
## Mean :12.81 Mean :22.19   
## 3rd Qu.:17.11 3rd Qu.:24.75   
## Max. :37.97 Max. :50.00

str(training)

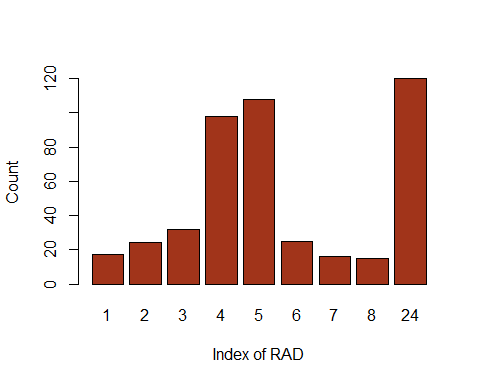
## 'data.frame': 455 obs. of 14 variables:  
## $ CRIM : num 0.0345 0.0453 3.8497 0.1008 0.2391 ...  
## $ ZN : num 82.5 0 0 0 0 0 0 0 0 0 ...  
## $ INDUS : num 2.03 11.93 18.1 10.01 9.69 ...  
## $ CHAS : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ NOX : num 0.415 0.573 0.77 0.547 0.585 0.58 0.489 0.609 0.659 0.58 ...  
## $ RM : num 6.16 6.12 6.39 6.71 6.02 ...  
## $ AGE : num 38.4 76.7 91 81.6 65.3 75 22.3 98.8 100 56.7 ...  
## $ DIS : num 6.27 2.29 2.51 2.68 2.41 ...  
## $ RAD : int 2 1 24 6 6 24 4 4 24 24 ...  
## $ TAX : int 348 273 666 432 391 666 277 711 666 666 ...  
## $ PTRATIO: num 14.7 21 20.2 17.8 19.2 20.2 18.6 20.1 20.2 20.2 ...  
## $ B : num 394 397 391 396 397 ...  
## $ LSTAT : num 7.43 9.08 13.27 10.16 12.92 ...  
## $ PRICE : num 24.1 20.6 21.7 22.8 21.2 23.2 22.6 13.6 11.9 20.1 ...

# Identify numerical columns by excluding categorical ones  
numerical\_columns <- names(training)[!names(training) %in% c("RAD", "CHAS")]  
  
# Set up the plotting area to display multiple histograms  
par(mfrow = c(3, 4))  
  
# Loop through only the numerical columns  
for(i in numerical\_columns) {  
 hist(training[[i]], main = i, xlab = "", col = "#a1341a", breaks = 30)  
}

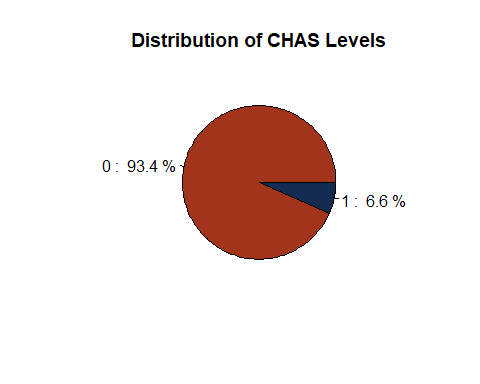


# Reset the plotting area  
par(mfrow = c(1, 1))

Rad\_levels <- table(training$RAD)  
barplot(Rad\_levels, xlab = "Index of RAD", ylab = "Count",col = "#a1341a")



CHAS\_levels <- table(training$CHAS)  
percentages <- round(100 \* CHAS\_levels / sum(CHAS\_levels), 1)  
  
# Create pie chart  
pie(CHAS\_levels,   
 main = "Distribution of CHAS Levels",   
 labels = paste(names(CHAS\_levels), ": ", percentages, "%"),   
 col = c("#a1341a","#152a51"))



**Checking for Multicollinearity**

cor(training)

## CRIM ZN INDUS CHAS NOX RM  
## CRIM 1.00000000 -0.19257571 0.38444551 -0.04707634 0.40657279 -0.2039928  
## ZN -0.19257571 1.00000000 -0.54117925 -0.05437532 -0.51135612 0.3596844  
## INDUS 0.38444551 -0.54117925 1.00000000 0.05415386 0.75811109 -0.3792452  
## CHAS -0.04707634 -0.05437532 0.05415386 1.00000000 0.08992628 0.1186172  
## NOX 0.40657279 -0.51135612 0.75811109 0.08992628 1.00000000 -0.3005324  
## RM -0.20399283 0.35968435 -0.37924520 0.11861718 -0.30053241 1.0000000  
## AGE 0.33956777 -0.56391554 0.65666966 0.10508326 0.73580406 -0.2719321  
## DIS -0.36905673 0.65238706 -0.71324191 -0.09752029 -0.77304099 0.2344420  
## RAD 0.60752087 -0.30268440 0.57721478 0.01708287 0.60564575 -0.1923162  
## TAX 0.56261217 -0.30925043 0.70564316 -0.02191669 0.66041740 -0.2697208  
## PTRATIO 0.28144659 -0.41406670 0.39524962 -0.09771645 0.19761793 -0.3660427  
## B -0.36677853 0.17288061 -0.33991171 0.04686768 -0.36516566 0.1069338  
## LSTAT 0.44469890 -0.42650485 0.60474826 -0.04216762 0.59600474 -0.6154606  
## PRICE -0.39344899 0.40434936 -0.48633438 0.19736800 -0.43211802 0.6905533  
## AGE DIS RAD TAX PTRATIO B  
## CRIM 0.3395678 -0.36905673 0.60752087 0.56261217 0.28144659 -0.36677853  
## ZN -0.5639155 0.65238706 -0.30268440 -0.30925043 -0.41406670 0.17288061  
## INDUS 0.6566697 -0.71324191 0.57721478 0.70564316 0.39524962 -0.33991171  
## CHAS 0.1050833 -0.09752029 0.01708287 -0.02191669 -0.09771645 0.04686768  
## NOX 0.7358041 -0.77304099 0.60564575 0.66041740 0.19761793 -0.36516566  
## RM -0.2719321 0.23444204 -0.19231616 -0.26972076 -0.36604269 0.10693378  
## AGE 1.0000000 -0.75695624 0.44678032 0.50071955 0.26307846 -0.26169432  
## DIS -0.7569562 1.00000000 -0.48840618 -0.53322472 -0.25661944 0.28427801  
## RAD 0.4467803 -0.48840618 1.00000000 0.90674551 0.46520417 -0.42487512  
## TAX 0.5007196 -0.53322472 0.90674551 1.00000000 0.45984796 -0.42245800  
## PTRATIO 0.2630785 -0.25661944 0.46520417 0.45984796 1.00000000 -0.16908761  
## B -0.2616943 0.28427801 -0.42487512 -0.42245800 -0.16908761 1.00000000  
## LSTAT 0.6304324 -0.51837084 0.48457836 0.53892776 0.36947022 -0.35842695  
## PRICE -0.4064507 0.28974073 -0.39066704 -0.47090000 -0.52627665 0.33829673  
## LSTAT PRICE  
## CRIM 0.44469890 -0.3934490  
## ZN -0.42650485 0.4043494  
## INDUS 0.60474826 -0.4863344  
## CHAS -0.04216762 0.1973680  
## NOX 0.59600474 -0.4321180  
## RM -0.61546061 0.6905533  
## AGE 0.63043244 -0.4064507  
## DIS -0.51837084 0.2897407  
## RAD 0.48457836 -0.3906670  
## TAX 0.53892776 -0.4709000  
## PTRATIO 0.36947022 -0.5262767  
## B -0.35842695 0.3382967  
## LSTAT 1.00000000 -0.7446804  
## PRICE -0.74468044 1.0000000

options(repos = c(CRAN = "https://cloud.r-project.org"))  
install.packages("corrplot")

## Installing package into 'C:/Users/patri/AppData/Local/R/win-library/4.3'  
## (as 'lib' is unspecified)

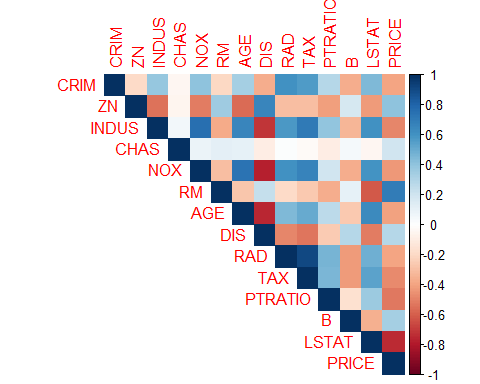
## package 'corrplot' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\patri\AppData\Local\Temp\RtmpyGVW02\downloaded\_packages

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.3.3

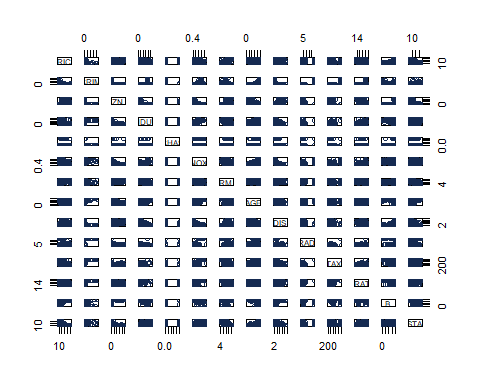
## corrplot 0.92 loaded

correlation\_matrix <- cor(training)  
corrplot(correlation\_matrix, method = "color", type = "upper")



# MODEL BUILDING AND VARIABLE SELECTION

pairs(PRICE~CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+B+LSTAT, data=training, col = "#152a51")



**Full model** (All 13 candidate explanatory variables)

fit1 <- lm(PRICE~CRIM+ZN+INDUS+factor(CHAS)+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+B+LSTAT, data=training)  
summary(fit1)

##   
## Call:  
## lm(formula = PRICE ~ CRIM + ZN + INDUS + factor(CHAS) + NOX +   
## RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT, data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.7158 -2.6163 -0.4263 1.5291 26.4620   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.579e+01 5.257e+00 6.808 3.24e-11 \*\*\*  
## CRIM -1.046e-01 3.170e-02 -3.298 0.001053 \*\*   
## ZN 4.896e-02 1.379e-02 3.550 0.000427 \*\*\*  
## INDUS 4.403e-02 6.039e-02 0.729 0.466270   
## factor(CHAS)1 3.529e+00 8.866e-01 3.980 8.06e-05 \*\*\*  
## NOX -1.539e+01 3.794e+00 -4.055 5.93e-05 \*\*\*  
## RM 3.617e+00 4.443e-01 8.142 4.03e-15 \*\*\*  
## AGE -6.861e-04 1.364e-02 -0.050 0.959896   
## DIS -1.351e+00 2.034e-01 -6.643 9.09e-11 \*\*\*  
## RAD 2.730e-01 6.541e-02 4.174 3.61e-05 \*\*\*  
## TAX -1.147e-02 3.673e-03 -3.124 0.001904 \*\*   
## PTRATIO -9.820e-01 1.304e-01 -7.533 2.84e-13 \*\*\*  
## B 9.883e-03 2.660e-03 3.715 0.000229 \*\*\*  
## LSTAT -5.297e-01 5.216e-02 -10.156 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.525 on 441 degrees of freedom  
## Multiple R-squared: 0.7523, Adjusted R-squared: 0.745   
## F-statistic: 103 on 13 and 441 DF, p-value: < 2.2e-16

step(fit1, direction ="both")

## Start: AIC=1387.45  
## PRICE ~ CRIM + ZN + INDUS + factor(CHAS) + NOX + RM + AGE + DIS +   
## RAD + TAX + PTRATIO + B + LSTAT  
##   
## Df Sum of Sq RSS AIC  
## - AGE 1 0.05 9028.1 1385.5  
## - INDUS 1 10.89 9038.9 1386.0  
## <none> 9028.0 1387.5  
## - TAX 1 199.74 9227.8 1395.4  
## - CRIM 1 222.64 9250.7 1396.5  
## - ZN 1 258.00 9286.0 1398.3  
## - B 1 282.59 9310.6 1399.5  
## - factor(CHAS) 1 324.26 9352.3 1401.5  
## - NOX 1 336.59 9364.6 1402.1  
## - RAD 1 356.66 9384.7 1403.1  
## - DIS 1 903.35 9931.4 1428.8  
## - PTRATIO 1 1161.54 10189.6 1440.5  
## - RM 1 1357.02 10385.1 1449.2  
## - LSTAT 1 2111.60 11139.6 1481.1  
##   
## Step: AIC=1385.45  
## PRICE ~ CRIM + ZN + INDUS + factor(CHAS) + NOX + RM + DIS + RAD +   
## TAX + PTRATIO + B + LSTAT  
##   
## Df Sum of Sq RSS AIC  
## - INDUS 1 10.86 9038.9 1384.0  
## <none> 9028.1 1385.5  
## + AGE 1 0.05 9028.0 1387.5  
## - TAX 1 199.95 9228.0 1393.4  
## - CRIM 1 222.61 9250.7 1394.5  
## - ZN 1 262.27 9290.4 1396.5  
## - B 1 283.70 9311.8 1397.5  
## - factor(CHAS) 1 325.33 9353.4 1399.6  
## - RAD 1 359.18 9387.3 1401.2  
## - NOX 1 361.64 9389.7 1401.3  
## - DIS 1 994.88 10023.0 1431.0  
## - PTRATIO 1 1166.04 10194.1 1438.7  
## - RM 1 1407.91 10436.0 1449.4  
## - LSTAT 1 2464.11 11492.2 1493.2  
##   
## Step: AIC=1384  
## PRICE ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS + RAD + TAX +   
## PTRATIO + B + LSTAT  
##   
## Df Sum of Sq RSS AIC  
## <none> 9038.9 1384.0  
## + INDUS 1 10.86 9028.1 1385.5  
## + AGE 1 0.03 9038.9 1386.0  
## - TAX 1 199.04 9238.0 1391.9  
## - CRIM 1 226.37 9265.3 1393.2  
## - ZN 1 253.27 9292.2 1394.6  
## - B 1 280.58 9319.5 1395.9  
## - factor(CHAS) 1 336.80 9375.7 1398.6  
## - RAD 1 353.49 9392.4 1399.5  
## - NOX 1 355.43 9394.4 1399.5  
## - DIS 1 1089.00 10127.9 1433.8  
## - PTRATIO 1 1158.27 10197.2 1436.9  
## - RM 1 1398.81 10437.8 1447.5  
## - LSTAT 1 2455.07 11494.0 1491.3

##   
## Call:  
## lm(formula = PRICE ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS +   
## RAD + TAX + PTRATIO + B + LSTAT, data = training)  
##   
## Coefficients:  
## (Intercept) CRIM ZN factor(CHAS)1 NOX   
## 35.492058 -0.105361 0.047840 3.576005 -14.678729   
## RM DIS RAD TAX PTRATIO   
## 3.596424 -1.377626 0.259541 -0.010317 -0.968359   
## B LSTAT   
## 0.009813 -0.527412

The model from stepwise analysis shows insignificant variables are INDUS and AGE. We will build our base model from the remaining significant variables. We have also eliminated the multicollinearity between INDUS~DIS and AGE~DIS after the stepwise process. We decided to keep NOX~DIS as we know both variables are significant in determining the house prices.

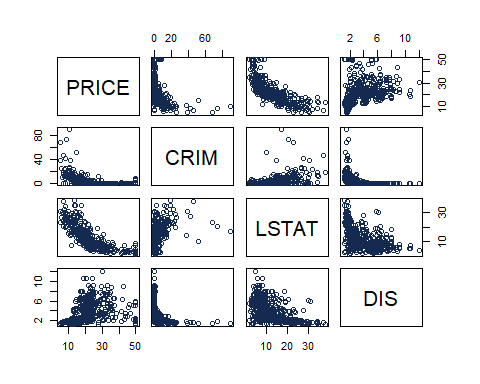
**Base Model** (Stepwise selection model without INDUS & AGE)

fit2 <- lm(PRICE~CRIM+ZN+factor(CHAS)+NOX+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT, data=training)  
summary(fit2)

##   
## Call:  
## lm(formula = PRICE ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS +   
## RAD + TAX + PTRATIO + B + LSTAT, data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.7018 -2.5879 -0.4098 1.6336 26.4848   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35.492058 5.218803 6.801 3.38e-11 \*\*\*  
## CRIM -0.105361 0.031632 -3.331 0.000938 \*\*\*  
## ZN 0.047840 0.013579 3.523 0.000471 \*\*\*  
## factor(CHAS)1 3.576005 0.880172 4.063 5.73e-05 \*\*\*  
## NOX -14.678729 3.516977 -4.174 3.61e-05 \*\*\*  
## RM 3.596424 0.434359 8.280 1.47e-15 \*\*\*  
## DIS -1.377626 0.188571 -7.306 1.29e-12 \*\*\*  
## RAD 0.259541 0.062355 4.162 3.79e-05 \*\*\*  
## TAX -0.010317 0.003303 -3.123 0.001905 \*\*   
## PTRATIO -0.968359 0.128525 -7.534 2.79e-13 \*\*\*  
## B 0.009813 0.002646 3.708 0.000235 \*\*\*  
## LSTAT -0.527412 0.048081 -10.969 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.517 on 443 degrees of freedom  
## Multiple R-squared: 0.752, Adjusted R-squared: 0.7459   
## F-statistic: 122.1 on 11 and 443 DF, p-value: < 2.2e-16

**Higher Order Model** (Based on Pair plot matrix => CRIM^2, LSTAT^2, DIS^2)

subset\_data <- training[, c("PRICE","CRIM", "LSTAT", "DIS")]  
  
pairs(subset\_data, col = "#152a51")



fit3 <- lm(PRICE~CRIM+ZN+factor(CHAS)+NOX+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT+I(CRIM^2)+I(LSTAT^2)+I(DIS^2), data=training)  
summary(fit3)

##   
## Call:  
## lm(formula = PRICE ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS +   
## RAD + TAX + PTRATIO + B + LSTAT + I(CRIM^2) + I(LSTAT^2) +   
## I(DIS^2), data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.3073 -2.3513 -0.2618 2.0662 21.6693   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 51.227465 5.157437 9.933 < 2e-16 \*\*\*  
## CRIM -0.401353 0.088757 -4.522 7.89e-06 \*\*\*  
## ZN 0.012128 0.012510 0.969 0.332845   
## factor(CHAS)1 3.287802 0.773578 4.250 2.61e-05 \*\*\*  
## NOX -16.616794 3.418791 -4.860 1.63e-06 \*\*\*  
## RM 2.975862 0.386283 7.704 8.83e-14 \*\*\*  
## DIS -3.180355 0.526716 -6.038 3.32e-09 \*\*\*  
## RAD 0.350872 0.062388 5.624 3.32e-08 \*\*\*  
## TAX -0.010647 0.002955 -3.603 0.000350 \*\*\*  
## PTRATIO -0.884670 0.114223 -7.745 6.64e-14 \*\*\*  
## B 0.007915 0.002346 3.374 0.000807 \*\*\*  
## LSTAT -1.688744 0.116939 -14.441 < 2e-16 \*\*\*  
## I(CRIM^2) 0.003284 0.001144 2.869 0.004310 \*\*   
## I(LSTAT^2) 0.034016 0.003219 10.568 < 2e-16 \*\*\*  
## I(DIS^2) 0.171713 0.045859 3.744 0.000205 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.962 on 440 degrees of freedom  
## Multiple R-squared: 0.8105, Adjusted R-squared: 0.8044   
## F-statistic: 134.4 on 14 and 440 DF, p-value: < 2.2e-16

**Interaction Model** (Based on theory, we know that AGE is significant variable, so we fitted it in an interaction model)

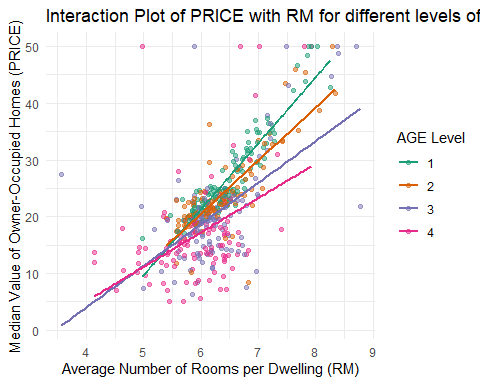
fit4 <- lm(PRICE~CRIM+factor(CHAS)+NOX+AGE+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT+I(CRIM^2)+I(LSTAT^2)+I(DIS^2)+AGE:RM+RM:LSTAT, data=training)  
summary(fit4)

##   
## Call:  
## lm(formula = PRICE ~ CRIM + factor(CHAS) + NOX + AGE + RM + DIS +   
## RAD + TAX + PTRATIO + B + LSTAT + I(CRIM^2) + I(LSTAT^2) +   
## I(DIS^2) + AGE:RM + RM:LSTAT, data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.0303 -2.1890 -0.2408 1.7957 22.2355   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.532893 7.662493 1.244 0.214129   
## CRIM -0.445600 0.084000 -5.305 1.79e-07 \*\*\*  
## factor(CHAS)1 2.797230 0.736067 3.800 0.000165 \*\*\*  
## NOX -17.272308 3.217559 -5.368 1.29e-07 \*\*\*  
## AGE 0.265883 0.075282 3.532 0.000456 \*\*\*  
## RM 8.973771 0.949395 9.452 < 2e-16 \*\*\*  
## DIS -3.081608 0.529931 -5.815 1.17e-08 \*\*\*  
## RAD 0.390600 0.059139 6.605 1.16e-10 \*\*\*  
## TAX -0.011709 0.002772 -4.225 2.91e-05 \*\*\*  
## PTRATIO -0.820349 0.103236 -7.946 1.64e-14 \*\*\*  
## B 0.004531 0.002274 1.992 0.046972 \*   
## LSTAT 0.541831 0.421272 1.286 0.199061   
## I(CRIM^2) 0.003766 0.001082 3.481 0.000550 \*\*\*  
## I(LSTAT^2) 0.012654 0.004548 2.782 0.005631 \*\*   
## I(DIS^2) 0.177555 0.043557 4.076 5.43e-05 \*\*\*  
## AGE:RM -0.040595 0.011670 -3.479 0.000555 \*\*\*  
## RM:LSTAT -0.266672 0.050977 -5.231 2.61e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.737 on 438 degrees of freedom  
## Multiple R-squared: 0.8321, Adjusted R-squared: 0.826   
## F-statistic: 135.7 on 16 and 438 DF, p-value: < 2.2e-16

*Interaction Plot for AGE & RM*

library(ggplot2)  
  
# We use 'cut' to create discrete bins for AGE, here I've used quartiles as an example  
training$AGE\_group <- cut(training$AGE, breaks=quantile(training$AGE, probs=0:4/4), include.lowest=TRUE, labels=FALSE)  
  
# Plot RM vs PRICE for different levels of AGE  
ggplot(training, aes(x = RM, y = PRICE, color = as.factor(AGE\_group))) +  
 geom\_point(alpha = 0.5) + # Actual data points  
 geom\_smooth(method = "lm", se = FALSE, aes(group = AGE\_group)) + # Add a regression line for each AGE group  
 theme\_minimal() +  
 labs(title = "Interaction Plot of PRICE with RM for different levels of AGE",  
 x = "Average Number of Rooms per Dwelling (RM)",  
 y = "Median Value of Owner-Occupied Homes (PRICE)",  
 color = "AGE Level") +  
 scale\_color\_brewer(palette = "Dark2") # Use a color palette that's visually distinct

## `geom\_smooth()` using formula = 'y ~ x'



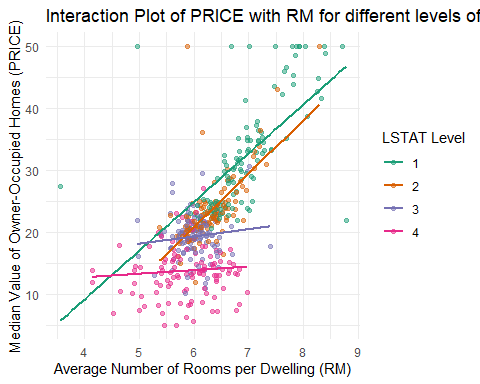
# Print the plot  
ggsave("interaction\_plot\_RM\_AGE.png", width = 10, height = 8)

## `geom\_smooth()` using formula = 'y ~ x'

*Interaction Plot for RM & LSTAT*

# We use 'cut' to create discrete bins for LSTAT, here I've used quartiles as an example  
training$LSTAT\_group <- cut(training$LSTAT, breaks=quantile(training$LSTAT, probs=0:4/4), include.lowest=TRUE, labels=FALSE)  
  
# Now we plot RM vs PRICE for different levels of LSTAT  
ggplot(training, aes(x = RM, y = PRICE, color = as.factor(LSTAT\_group))) +  
 geom\_point(alpha = 0.5) + # Actual data points  
 geom\_smooth(method = "lm", se = FALSE, aes(group = LSTAT\_group)) + # Add a regression line for each LSTAT group  
 theme\_minimal() +  
 labs(title = "Interaction Plot of PRICE with RM for different levels of LSTAT",  
 x = "Average Number of Rooms per Dwelling (RM)",  
 y = "Median Value of Owner-Occupied Homes (PRICE)",  
 color = "LSTAT Level") +  
 scale\_color\_brewer(palette = "Dark2") # Use a color palette that's visually distinct

## `geom\_smooth()` using formula = 'y ~ x'



# Print the plot  
ggsave("interaction\_plot\_RM\_LSTAT.png", width = 10, height = 8)

## `geom\_smooth()` using formula = 'y ~ x'

**Final Model** (We choose the simpler model => we removed LSTAT^2 & B)

fit5 <- lm(PRICE~CRIM+factor(CHAS)+NOX+AGE+RM+DIS+RAD+TAX+PTRATIO+LSTAT+I(CRIM^2)+I(DIS^2)+AGE:RM+RM:LSTAT, data=training)  
summary(fit5)

##   
## Call:  
## lm(formula = PRICE ~ CRIM + factor(CHAS) + NOX + AGE + RM + DIS +   
## RAD + TAX + PTRATIO + LSTAT + I(CRIM^2) + I(DIS^2) + AGE:RM +   
## RM:LSTAT, data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.7702 -2.0475 -0.2859 1.7568 24.6899   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.681998 6.891463 0.244 0.807291   
## CRIM -0.445477 0.082537 -5.397 1.11e-07 \*\*\*  
## factor(CHAS)1 2.750470 0.742440 3.705 0.000239 \*\*\*  
## NOX -18.195196 3.235163 -5.624 3.32e-08 \*\*\*  
## AGE 0.266761 0.076019 3.509 0.000496 \*\*\*  
## RM 10.345240 0.823817 12.558 < 2e-16 \*\*\*  
## DIS -3.298068 0.529625 -6.227 1.11e-09 \*\*\*  
## RAD 0.397454 0.059674 6.660 8.17e-11 \*\*\*  
## TAX -0.012605 0.002785 -4.526 7.75e-06 \*\*\*  
## PTRATIO -0.812305 0.104127 -7.801 4.50e-14 \*\*\*  
## LSTAT 1.597633 0.221524 7.212 2.42e-12 \*\*\*  
## I(CRIM^2) 0.003718 0.001070 3.474 0.000563 \*\*\*  
## I(DIS^2) 0.197398 0.043431 4.545 7.10e-06 \*\*\*  
## AGE:RM -0.041663 0.011773 -3.539 0.000445 \*\*\*  
## RM:LSTAT -0.375426 0.037019 -10.141 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.774 on 440 degrees of freedom  
## Multiple R-squared: 0.8281, Adjusted R-squared: 0.8226   
## F-statistic: 151.4 on 14 and 440 DF, p-value: < 2.2e-16

# MODEL COMPARISON

A Test for Comparing Nested Models

**Model Comparison Fit2 ~ Fit1**

* Reduced Model: fit2 <- lm(PRICE~CRIM+ZN+factor(CHAS)+NOX+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT, data=training)
* Full Model: fit1 ~ lm(PRICE~CRIM+ZN+factor(CHAS)+NOX+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT+AGE+INDUS, data=training)

Hypothesis: Reduced Model (Fit2): β12=β13=0  
Full Model (Fit1): At least one of the parameters, β12 or β13 differs from 0

The following outputs are for the full, reduced models, and partial F test respectively:

anova(fit2, fit1)

## Analysis of Variance Table  
##   
## Model 1: PRICE ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS + RAD + TAX +   
## PTRATIO + B + LSTAT  
## Model 2: PRICE ~ CRIM + ZN + INDUS + factor(CHAS) + NOX + RM + AGE + DIS +   
## RAD + TAX + PTRATIO + B + LSTAT  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 443 9038.9   
## 2 441 9028.0 2 10.914 0.2666 0.7661

p\_value (0.77) is greater than α (0.05) so we fail to reject null hypothesis, reduced model (fit2) is significant.

At 5% level of significance, we have sufficient evidence that reduced model contribute to the prediction of price

**Model Comparison Fit2 ~ Fit3**

* Reduced Model: fit2 <- lm(PRICE~CRIM+ZN+factor(CHAS)+NOX+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT, data=training)
* Full Model: fit3 <- lm(PRICE~CRIM+ZN+factor(CHAS)+NOX+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT+I(CRIM2)+ I(LSTAT2)+ I(DIS2), data=training)

Hypothesis: Reduced Model (Fit2): β12=β13=β14=0  
Full Model (Fit3): At least one of the parameters, β12,β13 or β14 differs from 0

The following outputs are for the full, reduced models, and partial F test respectively:

anova(fit2, fit3)

## Analysis of Variance Table  
##   
## Model 1: PRICE ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS + RAD + TAX +   
## PTRATIO + B + LSTAT  
## Model 2: PRICE ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS + RAD + TAX +   
## PTRATIO + B + LSTAT + I(CRIM^2) + I(LSTAT^2) + I(DIS^2)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 443 9038.9   
## 2 440 6908.1 3 2130.8 45.24 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p\_value (2.2e-16) is smaller than α (0.05) so we reject null hypothesis, full model (fit3) is significant.

At 5% level of significance, we do not have sufficient evidence that reduced model contribute to the prediction of price

**Model Comparison Fit3 ~ Fit4**

* Reduced Model: fit3 <- lm(PRICE~CRIM+ZN+factor(CHAS)+NOX+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT+I(CRIM2)+ I(LSTAT2)+ I(DIS2), data=training)
* Full Model: fit4 <- lm(PRICE~CRIM+factor(CHAS)+NOX+AGE+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT+I(CRIM2)+ I(LSTAT2)+ I(DIS2)+AGE:RM+RM:LSTAT, data=training)

Hypothesis: Reduced Model (Fit3): β15=β16=β17=0  
Full Model (Fit4): At least one of the parameters, β15, β16, β17=0 differs from 0

The following outputs are for the full, reduced models, and partial F test respectively:

anova(fit3, fit4)

## Analysis of Variance Table  
##   
## Model 1: PRICE ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS + RAD + TAX +   
## PTRATIO + B + LSTAT + I(CRIM^2) + I(LSTAT^2) + I(DIS^2)  
## Model 2: PRICE ~ CRIM + factor(CHAS) + NOX + AGE + RM + DIS + RAD + TAX +   
## PTRATIO + B + LSTAT + I(CRIM^2) + I(LSTAT^2) + I(DIS^2) +   
## AGE:RM + RM:LSTAT  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 440 6908.1   
## 2 438 6118.4 2 789.73 28.267 2.842e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p\_value (2.842e-12) is smaller than α (0.05) so we reject null hypothesis, full model (fit4) is significant.

At 5% level of significance, we do not have sufficient evidence that reduced model contribute to the prediction of price

**Model Comparison Fit4 ~ Fit5**

* Reduced Model: fit5 <- lm(PRICE~CRIM+factor(CHAS)+NOX+AGE+RM+DIS+RAD+TAX+PTRATIO+LSTAT+I(CRIM2)+I(DIS2)+AGE:RM+RM:LSTAT, data=training)
* Full Model: fit4 <- lm(PRICE~CRIM+factor(CHAS)+NOX+AGE+RM+DIS+RAD+TAX+PTRATIO+B+LSTAT+I(CRIM2)+I(LSTAT2)+I(DIS2)+AGE:RM+RM:LSTAT, data=training)

Hypothesis: Reduced Model (Fit5): β15=β16=β17=0  
Full Model (Fit4): At least one of the parameters, β15, β16, β17=0 differs from 0

The following outputs are for the full, reduced models, and partial F test respectively:

anova(fit5, fit4)

## Analysis of Variance Table  
##   
## Model 1: PRICE ~ CRIM + factor(CHAS) + NOX + AGE + RM + DIS + RAD + TAX +   
## PTRATIO + LSTAT + I(CRIM^2) + I(DIS^2) + AGE:RM + RM:LSTAT  
## Model 2: PRICE ~ CRIM + factor(CHAS) + NOX + AGE + RM + DIS + RAD + TAX +   
## PTRATIO + B + LSTAT + I(CRIM^2) + I(LSTAT^2) + I(DIS^2) +   
## AGE:RM + RM:LSTAT  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 440 6267.4   
## 2 438 6118.4 2 149.03 5.3345 0.005141 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p\_value (0.005141) is smaller than α (0.05) so we fail to reject null hypothesis, reduced model (fit5) is significant.

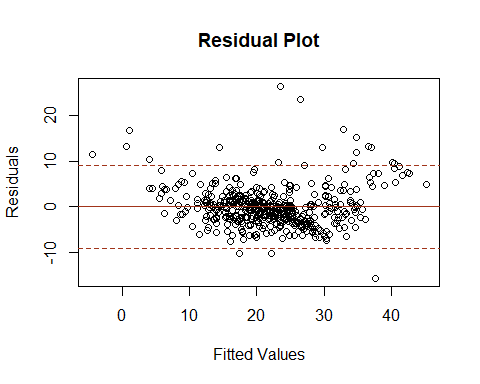
At 5% level of significance, we have sufficient evidence that reduced model (fit5) contribute to the prediction of price

# RESIDUAL ANALYSIS

**Residual Plot**

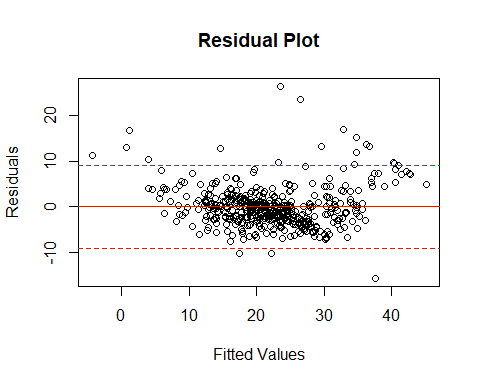
*Residual Plot for Fitted Model 1*

sd = summary(fit1)$sigma  
  
plot(fit1$fitted.values, fit1$residuals,  
 xlab = "Fitted Values",  
 ylab = "Residuals",  
 main = "Residual Plot")  
abline(h = 0, col = "#a1341a")  
abline(h = c(2, -2) \* sd, lty = "dashed", col = "#a1341a")



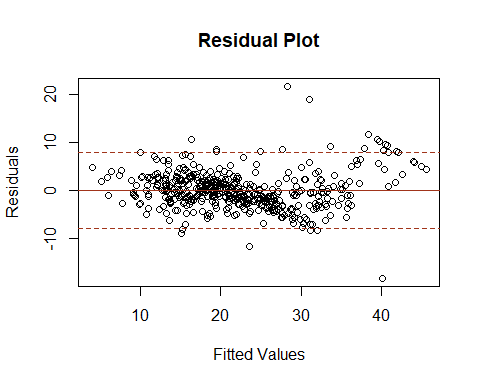
*Residual Plot for Fitted Model 2*

sd = summary(fit2)$sigma  
  
plot(fit2$fitted.values, fit2$residuals,  
 xlab = "Fitted Values",  
 ylab = "Residuals",  
 main = "Residual Plot")  
abline(h = 0, col = "#a1341a")  
abline(h = c(2, -2) \* sd, lty = "dashed", col = "#a1341a")



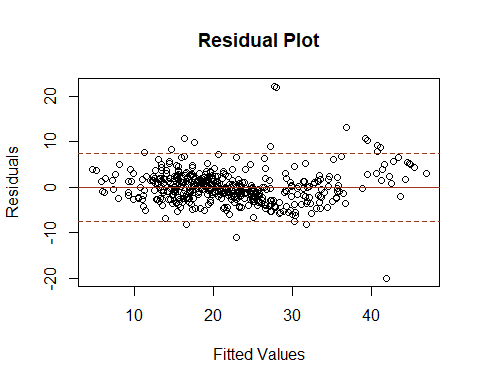
*Residual Plot for Fitted Model 3*

sd = summary(fit3)$sigma  
  
plot(fit3$fitted.values, fit3$residuals,  
 xlab = "Fitted Values",  
 ylab = "Residuals",  
 main = "Residual Plot")  
abline(h = 0, col = "#a1341a")  
abline(h = c(2, -2) \* sd, lty = "dashed", col = "#a1341a")



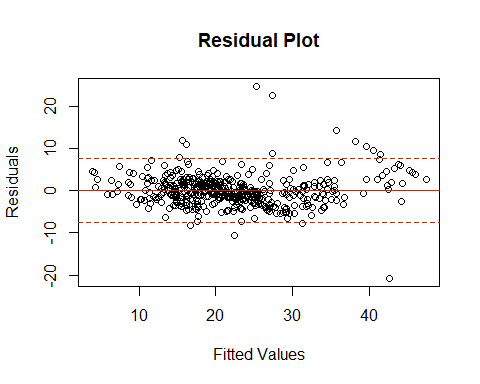
*Residual Plot for Fitted Model 4*

sd = summary(fit4)$sigma  
  
plot(fit4$fitted.values, fit4$residuals,  
 xlab = "Fitted Values",  
 ylab = "Residuals",  
 main = "Residual Plot")  
abline(h = 0, col = "#a1341a")  
abline(h = c(2, -2) \* sd, lty = "dashed", col = "#a1341a")



*Residual Plot for Fitted Model 5*

sd = summary(fit5)$sigma  
  
plot(fit5$fitted.values, fit5$residuals,  
 xlab = "Fitted Values",  
 ylab = "Residuals",  
 main = "Residual Plot")  
abline(h = 0, col = "#a1341a")  
abline(h = c(2, -2) \* sd, lty = "dashed", col = "#a1341a")



*Estimation of σ^2* s = 3.973 (from Residual standard error of lm function)

INTERPRETATION: We expect the model to provide predictions of House Pricing to be within about ±2s = ±2(3.973) = ±7.946 every unit of pricing.Residual plot shows we have constant variance and no pattern.

# PREDICTION EQUATION

summary(fit5)

##   
## Call:  
## lm(formula = PRICE ~ CRIM + factor(CHAS) + NOX + AGE + RM + DIS +   
## RAD + TAX + PTRATIO + LSTAT + I(CRIM^2) + I(DIS^2) + AGE:RM +   
## RM:LSTAT, data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.7702 -2.0475 -0.2859 1.7568 24.6899   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.681998 6.891463 0.244 0.807291   
## CRIM -0.445477 0.082537 -5.397 1.11e-07 \*\*\*  
## factor(CHAS)1 2.750470 0.742440 3.705 0.000239 \*\*\*  
## NOX -18.195196 3.235163 -5.624 3.32e-08 \*\*\*  
## AGE 0.266761 0.076019 3.509 0.000496 \*\*\*  
## RM 10.345240 0.823817 12.558 < 2e-16 \*\*\*  
## DIS -3.298068 0.529625 -6.227 1.11e-09 \*\*\*  
## RAD 0.397454 0.059674 6.660 8.17e-11 \*\*\*  
## TAX -0.012605 0.002785 -4.526 7.75e-06 \*\*\*  
## PTRATIO -0.812305 0.104127 -7.801 4.50e-14 \*\*\*  
## LSTAT 1.597633 0.221524 7.212 2.42e-12 \*\*\*  
## I(CRIM^2) 0.003718 0.001070 3.474 0.000563 \*\*\*  
## I(DIS^2) 0.197398 0.043431 4.545 7.10e-06 \*\*\*  
## AGE:RM -0.041663 0.011773 -3.539 0.000445 \*\*\*  
## RM:LSTAT -0.375426 0.037019 -10.141 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.774 on 440 degrees of freedom  
## Multiple R-squared: 0.8281, Adjusted R-squared: 0.8226   
## F-statistic: 151.4 on 14 and 440 DF, p-value: < 2.2e-16

**Predicted Housing Prices = 1.68 – 0.45(CRIM) + 2.75(CHAS:Yes) – 18.20(NOX) + 0.27(AGE) + 10.35(RM) – 3.30(DIS) + 0.40(RAD) – 0.01(TAX) – 0.81(PTRATIO) + 1.60(LSTAT) + 0.004 (CRIM2) + 0.20(DIS2) – 0.04(AGE:RM) – 0.38(RM:LSTAT)**

INTERPRETATION:

* CRIM (-0.46): For every one-unit increase in the per capita crime rate, the predicted housing price decreases by 0.46 units, assuming all other variables remain constant.
* CHAS:Yes (2.75): If the house borders the Charles River (CHAS is “Yes”), the predicted housing price is 2.75 units higher than if it doesn’t border the river, all else being equal.
* NOX (-18.20): For each one-unit increase in nitric oxides concentration, the predicted housing price decreases by 18.20 units, keeping all other variables constant.
* AGE (0.27): For every additional year in the age of the home, the predicted housing price increases by 0.27 units, assuming other variables stay the same.
* RM (10.36): Each additional room in the house is associated with an increase of 10.36 units in the predicted housing price, with all other factors held constant.
* DIS (-3.30): For every one-unit increase in the weighted distances to five Boston employment centers, the predicted housing price decreases by 3.30 units, assuming all else is constant.
* RAD (0.40): Each additional index of accessibility to radial highways is associated with an increase of 0.40 units in the predicted housing price, with all other variables held constant.
* TAX (-0.01): For every one-unit increase in the full-value property-tax rate per $10,000, the predicted housing price decreases by 0.01 units, assuming all other variables remain constant.
* PTRATIO (-0.81): For each one-unit increase in the pupil-teacher ratio by town, the predicted housing price decreases by 0.81 units, other factors being equal.
* LSTAT (1.60): For each one-unit increase in the percentage of the lower status of the population, the predicted housing price increases by 1.60 units, holding other variables constant.
* CRIM^2 (0.004): This term represents the non-linear effect of the crime rate. As CRIM increases, the impact on housing prices changes by an additional 0.004 units for each unit increase in CRIM, adjusted for the effect of CRIM itself.
* DIS^2 (0.20): Similar to CRIM^2, this represents the non-linear effect of the weighted distances to employment centers. The effect on housing prices changes by 0.20 units for each unit increase in DIS, considering the linear effect of DIS.
* AGE : RM (-0.04): This interaction term suggests that the effect of the age of the home on the predicted housing prices changes by -0.04 units for each additional room, implying that the positive effect of additional rooms diminishes as the house gets older.
* RM : LSTAT (-0.38): This interaction term indicates that the effect of the number of rooms (RM) on housing prices is reduced by 0.38 units for each additional unit of LSTAT. This suggests that the positive impact of having more rooms is less in areas with a higher proportion of lower status population.

# TESTING UTILITY OF MODEL

**Anova F-Test**

H0: β1 = β2 = … βi = 0 - model is not adequate in predicting y

Ha: At least one βi != 0 - model is adequate in predicting y

F-statistic: 151.4 (F-statistic on fit5) P-value: < 2.2e-16

INTERPRETATION: At 1% level of significance, we reject the H0. The data provide strong evidence to conclude that the model is useful in predicting the House Prices.

**Testing Individual Parameter**

H0: βi = 0 - there is no relationship

Ha : βi != 0 - there is a relationship

INTERPRETATION: Each variable contributes significantly to the prediction of housing prices, with different degrees of impact. The test statistics and its associated p-value provide evidence to reject the null hypothesis of no effect.

**Multiple Coefficients of Determination: R^2a**

summary(fit5)$adj.r.squared

## [1] 0.8225846

INTERPRETATION: R2a = 0.8225846 which means after adjusting for sample size and the number of model parameters, about 82% of the total sample variation in Price (y) is explained by the model; the remainder is explained by random error.

# MODEL EVALUATION

**Fit 3**

actual\_y <- testing$PRICE  
y\_hat\_fit3 <- predict(fit3, testing, type = "response")  
  
absolute\_difference\_fit3 <- abs(actual\_y - y\_hat\_fit3)  
cbind(actual\_y, y\_hat\_fit3, absolute\_difference\_fit3)

## actual\_y y\_hat\_fit3 absolute\_difference\_fit3  
## 5 36.2 29.140788 7.05921224  
## 24 14.5 11.599452 2.90054794  
## 29 18.4 17.132212 1.26778829  
## 51 19.7 18.722770 0.97723044  
## 54 23.4 23.594351 0.19435056  
## 59 23.3 23.324327 0.02432741  
## 60 19.6 20.622943 1.02294259  
## 62 16.0 16.041573 0.04157302  
## 86 26.6 28.326050 1.72604952  
## 94 25.0 29.366619 4.36661930  
## 101 27.5 24.447659 3.05234142  
## 105 20.1 20.823392 0.72339166  
## 109 19.8 21.784991 1.98499122  
## 153 15.3 21.538490 6.23848954  
## 161 27.0 36.273645 9.27364536  
## 176 29.4 32.906186 3.50618565  
## 180 37.2 34.988736 2.21126424  
## 181 39.8 34.331100 5.46890005  
## 187 50.0 37.499841 12.50015866  
## 213 22.4 21.021814 1.37818645  
## 215 23.7 14.453706 9.24629399  
## 226 50.0 40.918660 9.08134006  
## 227 37.6 40.432877 2.83287657  
## 232 31.7 34.556472 2.85647155  
## 234 48.3 38.892357 9.40764255  
## 236 24.0 24.237446 0.23744607  
## 238 31.5 34.262701 2.76270142  
## 240 23.3 27.658662 4.35866210  
## 248 20.5 19.690884 0.80911597  
## 256 20.9 21.335955 0.43595518  
## 257 44.0 37.213834 6.78616646  
## 259 36.0 36.252484 0.25248369  
## 275 32.4 38.280770 5.88077036  
## 277 33.2 35.104540 1.90454028  
## 279 29.1 29.205179 0.10517898  
## 287 20.1 17.850167 2.24983305  
## 317 17.8 15.143982 2.65601848  
## 350 26.6 24.208585 2.39141530  
## 355 18.2 17.769797 0.43020306  
## 368 23.1 11.031960 12.06803953  
## 372 50.0 25.684564 24.31543619  
## 376 15.0 21.392607 6.39260748  
## 385 8.8 7.545254 1.25474643  
## 386 7.2 12.154793 4.95479342  
## 410 27.5 17.080181 10.41981869  
## 414 16.3 8.157142 8.14285794  
## 422 14.2 17.018185 2.81818533  
## 423 20.8 16.640421 4.15957946  
## 437 9.6 11.650680 2.05067958  
## 446 11.8 15.105360 3.30535987  
## 484 21.8 20.945411 0.85458874

mean(absolute\_difference\_fit3)

## [1] 4.143922

Fit 3 - Mean Absolute Difference (Actual Y - Y\_hat): 4.14

**Fit 4**

y\_hat\_fit4 <- predict(fit4, testing, type = "response")  
  
absolute\_difference\_fit4 <- abs(actual\_y - y\_hat\_fit4)  
cbind(actual\_y, y\_hat\_fit4, absolute\_difference\_fit4)

## actual\_y y\_hat\_fit4 absolute\_difference\_fit4  
## 5 36.2 30.37561 5.82438694  
## 24 14.5 12.79995 1.70004682  
## 29 18.4 17.52477 0.87523068  
## 51 19.7 18.67182 1.02818395  
## 54 23.4 22.19425 1.20575206  
## 59 23.3 22.50432 0.79568202  
## 60 19.6 20.14888 0.54888142  
## 62 16.0 17.46615 1.46615206  
## 86 26.6 28.29967 1.69966776  
## 94 25.0 27.56189 2.56188668  
## 101 27.5 24.51475 2.98524925  
## 105 20.1 21.22624 1.12624074  
## 109 19.8 21.91458 2.11457709  
## 153 15.3 21.71383 6.41382580  
## 161 27.0 34.47045 7.47044647  
## 176 29.4 32.11502 2.71501803  
## 180 37.2 35.17701 2.02299027  
## 181 39.8 34.80026 4.99973616  
## 187 50.0 39.98053 10.01946864  
## 213 22.4 20.57851 1.82148897  
## 215 23.7 10.26168 13.43831942  
## 226 50.0 43.12247 6.87753030  
## 227 37.6 41.62036 4.02035693  
## 232 31.7 35.29958 3.59958128  
## 234 48.3 41.52218 6.77782103  
## 236 24.0 24.21335 0.21335256  
## 238 31.5 35.08890 3.58889897  
## 240 23.3 27.66083 4.36083282  
## 248 20.5 20.46716 0.03284425  
## 256 20.9 19.31920 1.58079503  
## 257 44.0 38.60567 5.39433243  
## 259 36.0 35.17641 0.82358585  
## 275 32.4 36.99450 4.59450181  
## 277 33.2 35.59576 2.39575940  
## 279 29.1 28.38768 0.71232398  
## 287 20.1 17.67491 2.42508556  
## 317 17.8 15.68375 2.11624912  
## 350 26.6 25.27013 1.32986612  
## 355 18.2 15.62566 2.57434196  
## 368 23.1 15.95927 7.14072990  
## 372 50.0 25.79630 24.20369943  
## 376 15.0 19.65743 4.65742532  
## 385 8.8 12.30858 3.50858276  
## 386 7.2 11.55749 4.35749230  
## 410 27.5 15.04231 12.45768773  
## 414 16.3 11.35657 4.94342623  
## 422 14.2 17.96989 3.76989456  
## 423 20.8 17.90395 2.89604726  
## 437 9.6 11.89782 2.29781502  
## 446 11.8 13.50600 1.70599853  
## 484 21.8 20.26071 1.53929313

mean(absolute\_difference\_fit4)

## [1] 3.916262

Fit 4 - Mean Absolute Difference (Actual Y - Y\_hat): 3.92

**Fit 5**

y\_hat\_fit5 <- predict(fit5, testing, type = "response")  
  
absolute\_difference\_fit5 <- abs(actual\_y - y\_hat\_fit5)  
cbind(actual\_y, y\_hat\_fit5, absolute\_difference\_fit5)

## actual\_y y\_hat\_fit5 absolute\_difference\_fit5  
## 5 36.2 30.253718 5.94628216  
## 24 14.5 12.892244 1.60775572  
## 29 18.4 17.505907 0.89409285  
## 51 19.7 19.130923 0.56907657  
## 54 23.4 22.141047 1.25895302  
## 59 23.3 22.268757 1.03124348  
## 60 19.6 20.080499 0.48049900  
## 62 16.0 17.796854 1.79685429  
## 86 26.6 28.026473 1.42647278  
## 94 25.0 27.117771 2.11777057  
## 101 27.5 24.458720 3.04128003  
## 105 20.1 21.335585 1.23558485  
## 109 19.8 21.945538 2.14553775  
## 153 15.3 21.762551 6.46255061  
## 161 27.0 33.665848 6.66584753  
## 176 29.4 31.690056 2.29005641  
## 180 37.2 34.977444 2.22255614  
## 181 39.8 35.159635 4.64036454  
## 187 50.0 40.398393 9.60160678  
## 213 22.4 21.088238 1.31176206  
## 215 23.7 9.917166 13.78283404  
## 226 50.0 44.024538 5.97546242  
## 227 37.6 41.781800 4.18179974  
## 232 31.7 35.258215 3.55821491  
## 234 48.3 42.062857 6.23714312  
## 236 24.0 24.370422 0.37042210  
## 238 31.5 34.883399 3.38339860  
## 240 23.3 27.600680 4.30068031  
## 248 20.5 20.585960 0.08595985  
## 256 20.9 19.591576 1.30842427  
## 257 44.0 38.616340 5.38365974  
## 259 36.0 35.161285 0.83871498  
## 275 32.4 36.309389 3.90938938  
## 277 33.2 35.654071 2.45407139  
## 279 29.1 28.275591 0.82440878  
## 287 20.1 18.667761 1.43223930  
## 317 17.8 15.885628 1.91437233  
## 350 26.6 25.378871 1.22112877  
## 355 18.2 15.871189 2.32881119  
## 368 23.1 17.570274 5.52972587  
## 372 50.0 25.624480 24.37552011  
## 376 15.0 19.418576 4.41857604  
## 385 8.8 13.521036 4.72103557  
## 386 7.2 10.318375 3.11837489  
## 410 27.5 15.221557 12.27844276  
## 414 16.3 12.847086 3.45291397  
## 422 14.2 18.310708 4.11070839  
## 423 20.8 18.548048 2.25195205  
## 437 9.6 13.158106 3.55810624  
## 446 11.8 13.639871 1.83987059  
## 484 21.8 20.102511 1.69748926

mean(absolute\_difference\_fit5)

## [1] 3.756667

Fit 5 - Mean Absolute Difference (Actual Y - Y\_hat): 3.76

# ESTIMATION AND PREDICTION

**Prediction Interval**

new\_data <- data.frame(CRIM=0.63, CHAS=factor(0), NOX=0.52, AGE=60, RM=5, DIS=4.5, RAD=4, TAX=307, PTRATIO=21, LSTAT=5.5)

pred <- predict(fit5, newdata = new\_data, interval = "prediction", level = 0.95)  
pred

## fit lwr upr  
## 1 15.45467 7.89914 23.0102

INTERPRETATION: We are 95% confidence that the true value of the response variable will fall within the range given by the lower and upper bounds of the prediction interval. In this case, the predicted value of the response variable is 15.45(’000 dollars), and the 95% prediction interval for this prediction is [7.90, 23.01] (’000 dollars).

**Confidence Interval**

conf <- predict(fit5, newdata = new\_data, interval = "confidence", level = 0.95)  
conf

## fit lwr upr  
## 1 15.45467 14.01744 16.8919

INTERPRETATION: We are 95% confident that the mean value of the response variable falls within the range given by the lower and upper bounds of the confidence interval. In this case, the predicted value of the response variable is 15.45(’000 dollars), and the 95% confidence interval for this prediction is [14.02, 16.89] (’000 dollars).