ANUSHKA\_INDOLIKE\_TASK\_2\_HD

ANUSHKA SHARMA

2025-02-16

library(readxl)  
HousingData <- read\_excel("~/Downloads/HousingData.xlsx")  
View(HousingData)

# Convert to numeric  
HousingData$LSTAT <- as.numeric(as.character(HousingData$LSTAT))

## Warning: NAs introduced by coercion

HousingData$RM <- as.numeric(as.character(HousingData$RM))  
HousingData$PTRATIO <- as.numeric(as.character(HousingData$PTRATIO))  
HousingData$MEDV <- as.numeric(as.character(HousingData$MEDV))  
  
# Check structure again  
str(HousingData)

## tibble [506 × 14] (S3: tbl\_df/tbl/data.frame)  
## $ CRIM : chr [1:506] "6.3200000000000001E-3" "2.7310000000000001E-2" "2.7289999999999998E-2" "3.2370000000000003E-2" ...  
## $ ZN : chr [1:506] "18" "0" "0" "0" ...  
## $ INDUS : chr [1:506] "2.31" "7.07" "7.07" "2.1800000000000002" ...  
## $ CHAS : chr [1:506] "0" "0" "0" "0" ...  
## $ NOX : num [1:506] 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...  
## $ RM : num [1:506] 6.58 6.42 7.18 7 7.15 ...  
## $ AGE : chr [1:506] "65.2" "78.900000000000006" "61.1" "45.8" ...  
## $ DIS : num [1:506] 4.09 4.97 4.97 6.06 6.06 ...  
## $ RAD : num [1:506] 1 2 2 3 3 3 5 5 5 5 ...  
## $ TAX : num [1:506] 296 242 242 222 222 222 311 311 311 311 ...  
## $ PTRATIO: num [1:506] 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...  
## $ B : num [1:506] 397 397 393 395 397 ...  
## $ LSTAT : num [1:506] 4.98 9.14 4.03 2.94 NA ...  
## $ MEDV : num [1:506] 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...

# Load necessary libraries  
library(tidyverse) # For data manipulation and visualization

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

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(caret) # For model training

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

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(ggplot2) # For plotting  
  
# View basic details of the dataset  
glimpse(HousingData)

## Rows: 506  
## Columns: 14  
## $ CRIM <chr> "6.3200000000000001E-3", "2.7310000000000001E-2", "2.728999999…  
## $ ZN <chr> "18", "0", "0", "0", "0", "0", "12.5", "12.5", "12.5", "12.5",…  
## $ INDUS <chr> "2.31", "7.07", "7.07", "2.1800000000000002", "2.1800000000000…  
## $ CHAS <chr> "0", "0", "0", "0", "0", "0", "NA", "0", "0", "NA", "0", "0", …  
## $ NOX <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524, 0.524,…  
## $ RM <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172, 5.631,…  
## $ AGE <chr> "65.2", "78.900000000000006", "61.1", "45.8", "54.2", "58.7", …  
## $ DIS <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605, 5.9505…  
## $ RAD <dbl> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4,…  
## $ TAX <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311, 31…  
## $ PTRATIO <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2, 15…  
## $ B <dbl> 396.90, 396.90, 392.83, 394.63, 396.90, 394.12, 395.60, 396.90…  
## $ LSTAT <dbl> 4.98, 9.14, 4.03, 2.94, NA, 5.21, 12.43, 19.15, 29.93, 17.10, …  
## $ MEDV <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, 18.9, 15…

summary(HousingData)

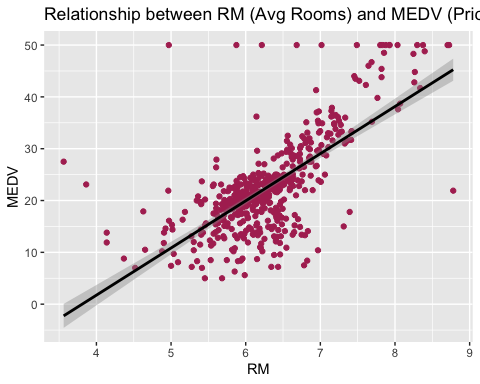
## CRIM ZN INDUS CHAS   
## Length:506 Length:506 Length:506 Length:506   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## NOX RM AGE DIS   
## Min. :0.3850 Min. :3.561 Length:506 Min. : 1.130   
## 1st Qu.:0.4490 1st Qu.:5.886 Class :character 1st Qu.: 2.100   
## Median :0.5380 Median :6.208 Mode :character Median : 3.207   
## Mean :0.5547 Mean :6.285 Mean : 3.795   
## 3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 5.188   
## Max. :0.8710 Max. :8.780 Max. :12.127   
##   
## RAD TAX PTRATIO B   
## Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 0.32   
## 1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.:375.38   
## Median : 5.000 Median :330.0 Median :19.05 Median :391.44   
## Mean : 9.549 Mean :408.2 Mean :18.46 Mean :356.67   
## 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.23   
## Max. :24.000 Max. :711.0 Max. :22.00 Max. :396.90   
##   
## LSTAT MEDV   
## Min. : 1.730 Min. : 5.00   
## 1st Qu.: 7.125 1st Qu.:17.02   
## Median :11.430 Median :21.20   
## Mean :12.715 Mean :22.53   
## 3rd Qu.:16.955 3rd Qu.:25.00   
## Max. :37.970 Max. :50.00   
## NA's :20

# Check for missing values  
missing\_values <- colSums(is.na(HousingData))  
print(missing\_values)

## CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX   
## 0 0 0 0 0 0 0 0 0 0   
## PTRATIO B LSTAT MEDV   
## 0 0 20 0

# If there are missing values, handle them (e.g., remove or impute)  
housing\_data <- na.omit(HousingData)  
  
# Visualizing the relationship between predictors and price  
ggplot(HousingData, aes(x = RM, y = MEDV)) +   
 geom\_point(color = "maroon") +   
 geom\_smooth(method = "lm", color = "black") +   
 ggtitle("Relationship between RM (Avg Rooms) and MEDV (Price)")

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



#Possible Insights:  
#The strong upward trend indicates that larger houses generally have higher prices.  
#The regression line captures the overall trend well, though there are a few outliers.  
#This insight can be useful for real estate pricing models and investment decisions.

# Split data into training (80%) and testing (20%) sets  
set.seed(123) # For reproducibility  
split\_index <- createDataPartition(HousingData$MEDV, p = 0.8, list = FALSE)  
train\_data <- HousingData[split\_index, ]  
test\_data <- HousingData[-split\_index, ]  
  
# Build a linear regression model  
model <- lm(MEDV ~ RM + LSTAT + PTRATIO, data = train\_data) # Using key features  
  
# View model summary  
summary(model)

##   
## Call:  
## lm(formula = MEDV ~ RM + LSTAT + PTRATIO, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.329 -3.079 -0.741 1.817 29.035   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.82353 4.34526 5.022 7.82e-07 \*\*\*  
## RM 4.05315 0.47094 8.606 < 2e-16 \*\*\*  
## LSTAT -0.58046 0.04908 -11.826 < 2e-16 \*\*\*  
## PTRATIO -0.94604 0.13441 -7.038 8.97e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.27 on 385 degrees of freedom  
## (18 observations deleted due to missingness)  
## Multiple R-squared: 0.6715, Adjusted R-squared: 0.6689   
## F-statistic: 262.3 on 3 and 385 DF, p-value: < 2.2e-16

# Predict on test data  
predictions <- predict(model, test\_data)  
  
# Model evaluation - Compute RMSE and R-squared  
rmse <- sqrt(mean((test\_data$MEDV - predictions)^2))  
r\_squared <- cor(test\_data$MEDV, predictions)^2  
  
print(paste("RMSE: ", round(rmse, 2)))

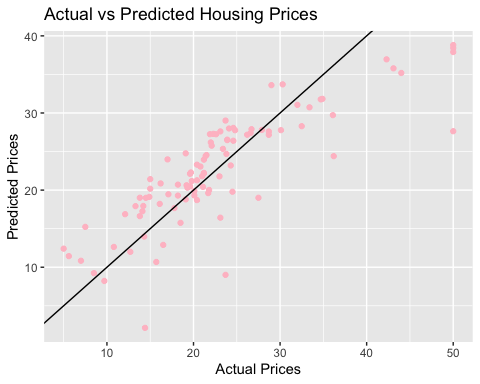
## [1] "RMSE: NA"

print(paste("R-squared: ", round(r\_squared, 2)))

## [1] "R-squared: NA"

# Plot actual vs predicted values  
ggplot(data = test\_data, aes(x = MEDV, y = predictions)) +  
 geom\_point(color = "pink") +  
 geom\_abline(intercept = 0, slope = 1, color = "black") +  
 ggtitle("Actual vs Predicted Housing Prices") +  
 xlab("Actual Prices") +  
 ylab("Predicted Prices")

## Warning: Removed 2 rows containing missing values or values outside the scale range  
## (`geom\_point()`).



#This is an Actual vs. Predicted Housing Prices plot! 📊🏡  
  
#Analysis:  
#The black diagonal line represents the ideal scenario where actual prices match predicted prices exactly.  
#Most points are close to this line, indicating that the model is making reasonably good predictions.  
#However, there are some deviations, especially at higher prices, suggesting the model might not be capturing all the complexities of the data.  
  
#Possible Next Steps:  
#✅ Evaluate model performance using R², RMSE, and MAE to quantify accuracy.  
#✅ Try feature engineering or non-linear models if prediction errors are high.  
#✅ Check for potential outliers or missing data that might affect accuracy.