R Mini Project

Ruthvik Akula
70572200028

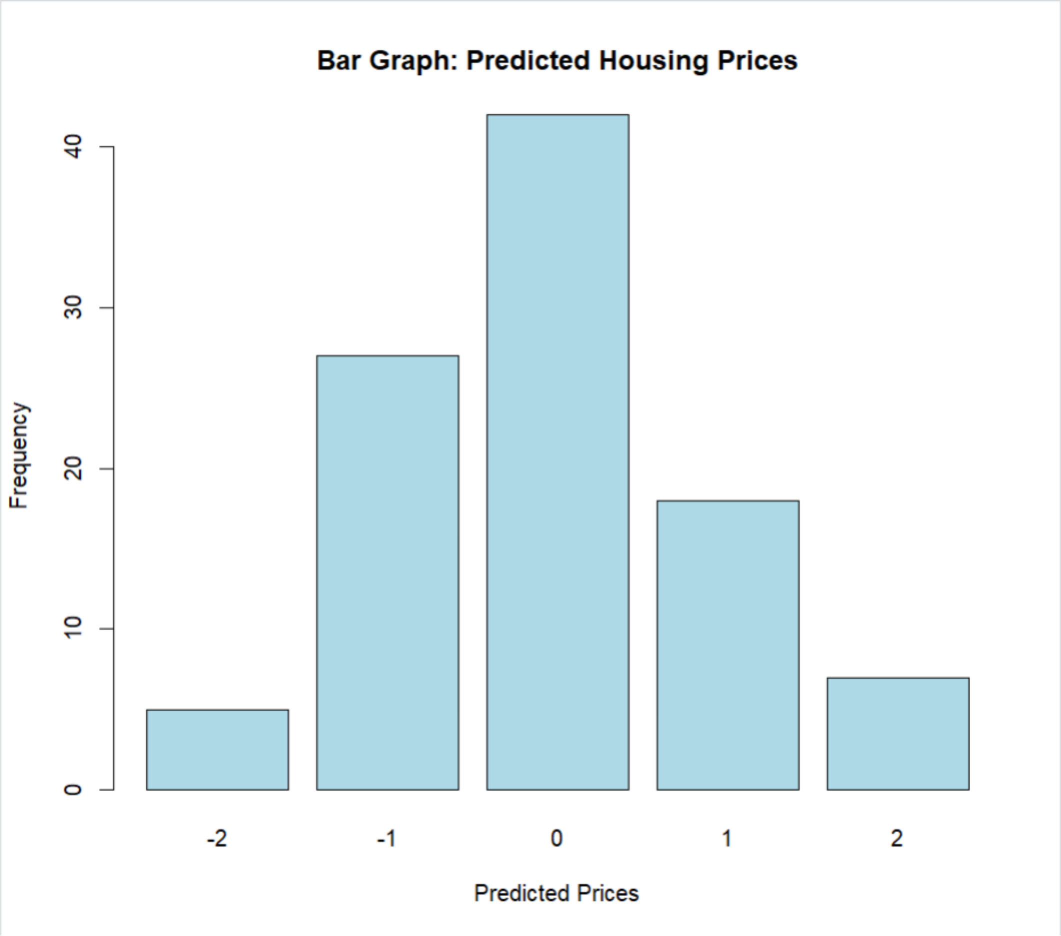
# Aim- R project: Preprocessing, Visualization and Prediction model of House prices using "Boston Housing Data" dataset.

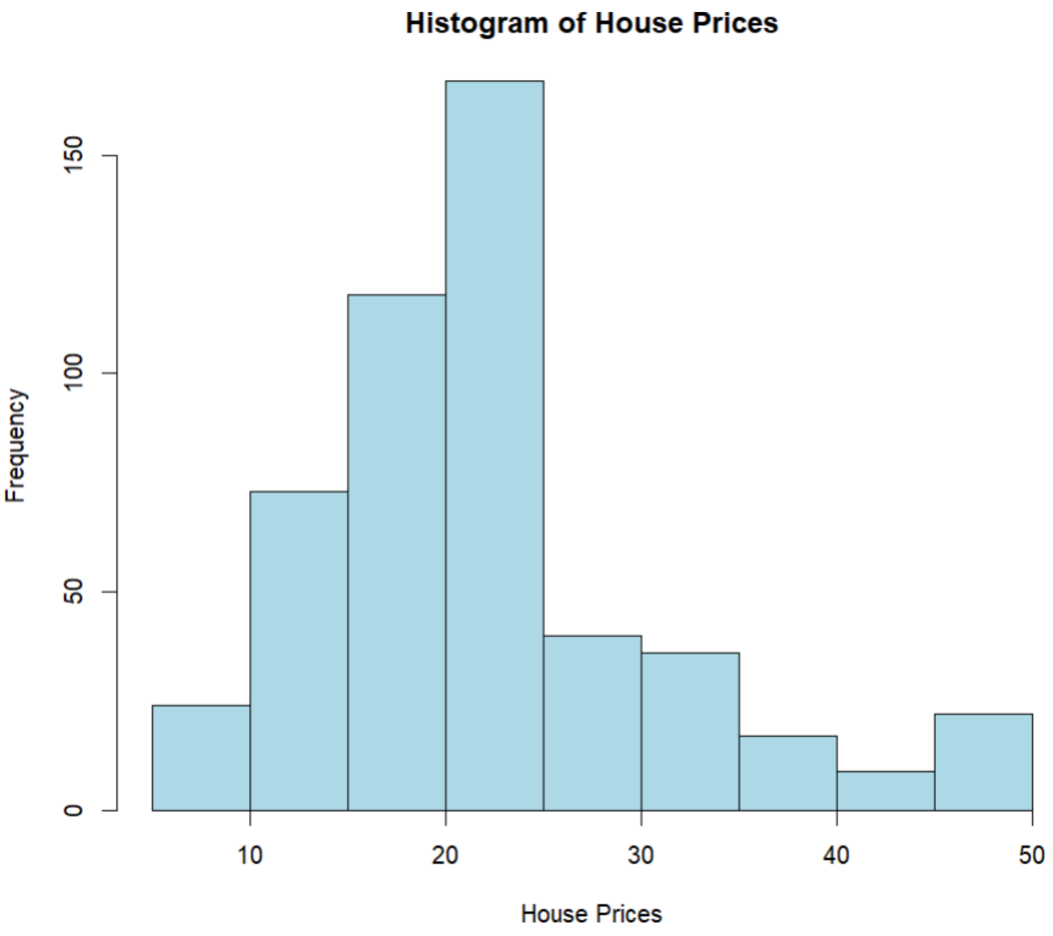
CODE:

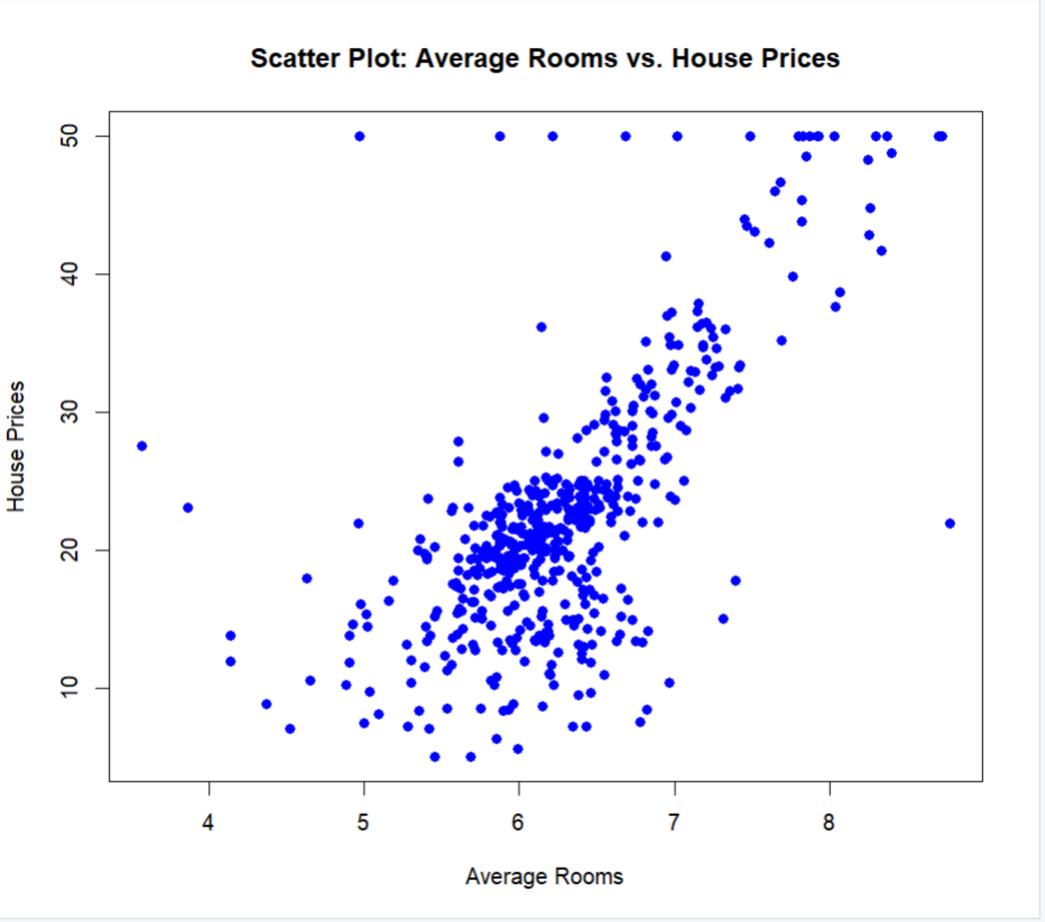
```
# Load necessary libraries

library(MASS)

library(ggplot2)

library(caret)

library(e1071)


# Load the Boston Housing dataset

data(Boston)

boston <- Boston


# Display the first few rows of the dataset

head(boston)


# Data preprocessing

# Check for missing values

missing_values <- colSums(is.na(boston))

print("Missing Values:")

print(missing_values)


# Check summary statistics

summary(boston)


# Feature scaling (optional)

# You can use other scaling techniques based on your preference
```

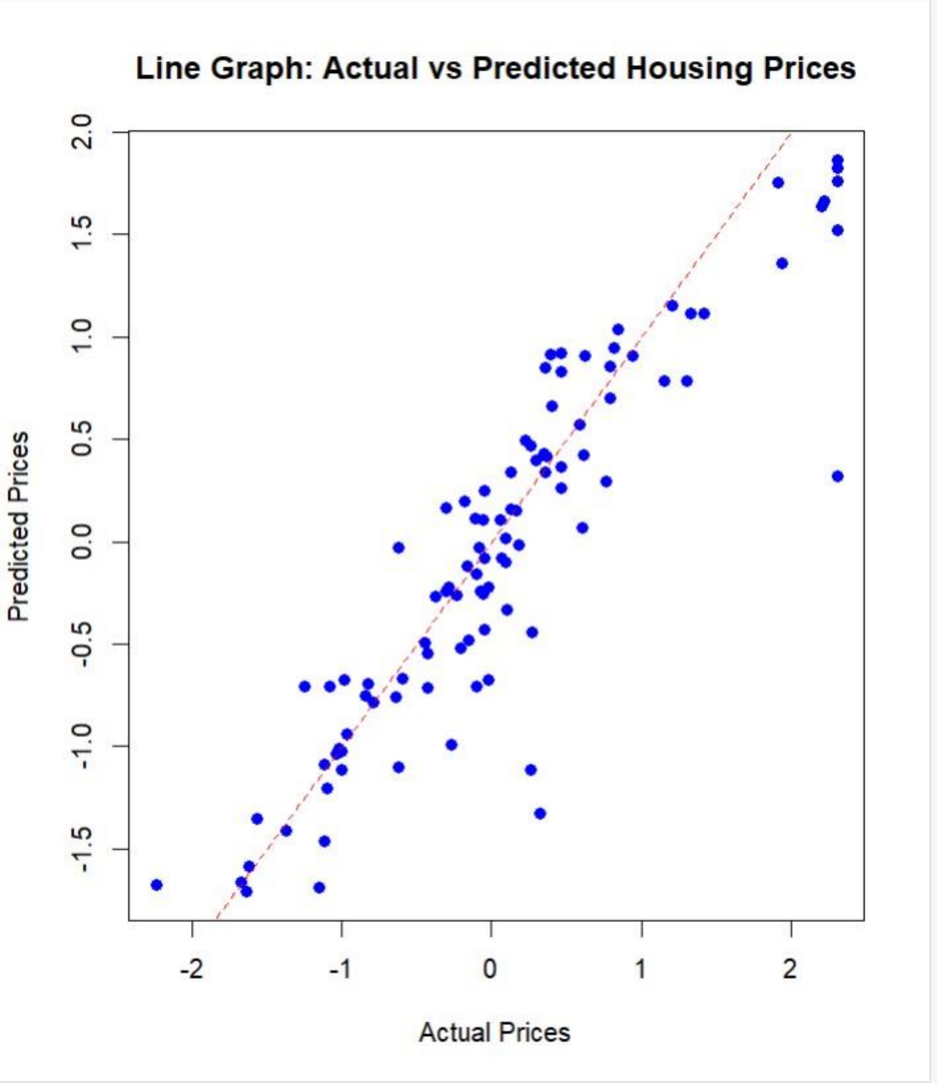
```
boston_scaled <- as.data.frame(scale(boston))

# Split the data into training and testing sets
set.seed(42)
splitIndex <- createDataPartition(boston$medv, p = 0.8, list = FALSE)
train_data <- boston_scaled[splitIndex, ]
test_data <- boston_scaled[-splitIndex, ]

# Additional preprocessing operations using caret
preprocess_params <- preProcess(train_data, method = c("center", "scale", "zv", "knnImpute", "YeoJohnson"))
train_data <- predict(preprocess_params, train_data)
test_data <- predict(preprocess_params, test_data)

# Model development
model <- lm(medv ~ ., data = train_data)

# Make predictions on the test set
predictions <- predict(model, newdata = test_data)

# Model evaluation
mse <- mean((test_data$medv - predictions)^2)
r_squared <- 1 - (mse / var(test_data$medv))

cat("Mean Squared Error:", mse, "\n")
cat("R-squared:", r_squared, "\n")

# Visualization
# Scatter plot for the relationship between 'rm' and 'medv'
plot(boston$rm, boston$medv, main = "Scatter Plot: Average Rooms vs. House Prices",
     xlab = "Average Rooms", ylab = "House Prices", col = "blue", pch = 16)


# Display histograms for selected numeric variables
hist(boston$medv, main = "Histogram of House Prices", xlab = "House Prices", col = "lightblue")

ggplot() +
  geom_point(aes(x = test_data$medv, y = predictions), color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(x = "Actual Prices (Scaled)", y = "Predicted Prices (Scaled)", title = "Actual vs Predicted Prices")

# Line Graph
plot(test_data$medv, predictions,
     main = "Line Graph: Actual vs Predicted Housing Prices",
     xlab = "Actual Prices",
     ylab = "Predicted Prices",
     col = "blue",
     pch = 16)

# Add a diagonal line for reference
abline(a = 0, b = 1, col = "red", lty = 2)
```

OUTPUT :#graphs


Bar Graph: Predicted Housing Prices
Frequency
0
10
20
30
40
-2
-1
0
1
2
Predicted Prices


Histogram of House Prices
Frequency
0
50
100
150
10
20
30
40
50
House Prices


Scatter Plot: Average Rooms vs. House Prices
House Prices
50
40
30
20
10
4
5
6
7
8
Average Rooms


Line Graph: Actual vs Predicted Housing Prices
Predicted Prices
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2
-1
0
1
2
Actual Prices