Name: Farouq Fayez Faky Darir

ID: 0366675

Section Number: 02

Contents

Table of Figure

Question 1:

Part a:

```
> # Q1a: Load dataset and display the first 6 rows (5 marks)
> # Load the Boston dataset and display the first 6 rows to inspect initial data
> head(Boston)  # Outputs the first 6 rows of the dataset for preview
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

Figure 1:Code and output for loading dataset and displaying 6 rows

Part b:

```
> # Q1b: Check for missing values and handle them appropriately (5 marks)
> # Check for missing values in each column of the dataset
> missing_values <- colSums(is.na(Boston))  # Calculate sum of NA values per column
> print("Missing values in each column:")
[1] "Missing values in each column:"
> print(missing_values)  # Display results; Boston has no missing values
   crim      zn   indus    chas     nox      rm     age     dis     rad     tax ptratio   black   lstat    medv
      0       0       0       0       0       0       0       0       0       0       0       0       0       0
```

Figure 2: Code and Output for missing values

As seen in the above figure there are no missing values in the Boston housing dataset

Part c:

```
> # Q1c: Generate summary statistics for all numerical variables (5 marks)
> # Generate summary statistics (min, max, median, mean, quartiles) for all variables
> summary(Boston)  # Provides a statistical overview of the dataset
      crim                zn             indus            chas              nox               rm             age              dis              rad        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000  
      tax           ptratio          black            lstat            medv      
 Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
 Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20  
 Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00  
```

Figure 3:Code and Output for summary statistics

Part d:

First visual

```
# First visualization: Histogram of medv
hist(Boston$medv, main="Distribution of Median House Value",
     xlab="Median Value ($1000s)", ylab="Frequency", col="lightblue", border="black")
# Explanation: This histogram illustrates the distribution of median house values, showing a right-skewed pattern.
```

Figure 4: Code for Distribution of Median House Value

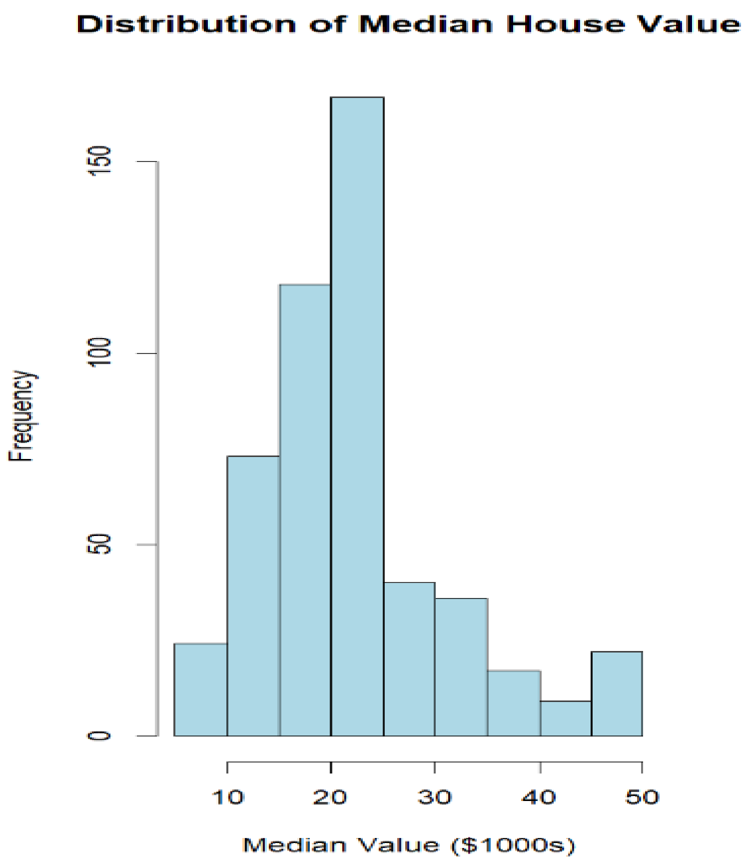

Figure 5: Output histogram of distribution of median house value

Interpretation:

To begin with it can be seen from the above figure that the distribution is right skewed showing that majority homes are lower prices but there are neighbourhoods with a higher median home values leading to stretching the tails to the right side. Secondly, it can be seen that most house values had a median values from, 20k-25k USD which may indicate a socioeconomic concentration that's suggest where a majority of the population lives in similarly values homes

Second visual:

```
# Second visualization: Scatterplot of medv vs rm
plot(Boston$rm, Boston$medv, main="House Value vs Average Rooms",
     xlab="Average Rooms", ylab="Median Value ($1000s)", pch=19, col="darkgreen")
```

Figure 6: Code for Scatter plot of House Value vs Average Room

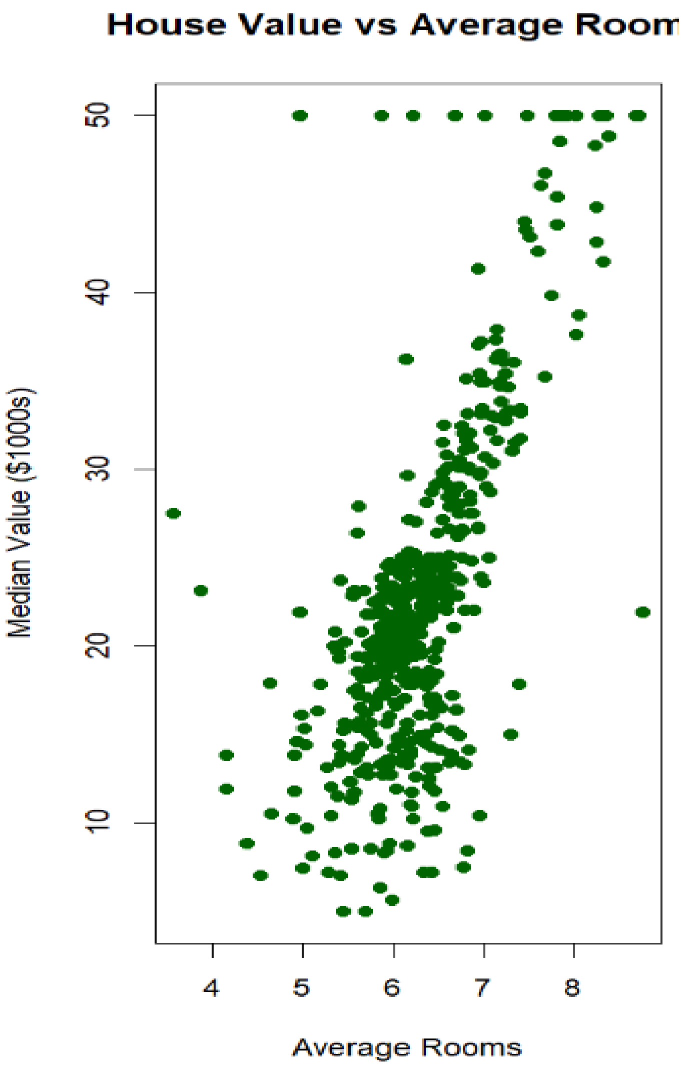

Figure 7: Output for Scatter plot of House Value vs Average Room

Interpretation:

This scatter plot illustrates the relationship between the average number of rooms per dwelling (rm) and the median house value (medv) for the Boston Housing dataset and each point represent a suburb. The scatter plot is positively correlated since as the number of average rooms increase the house median value increase which could be to factors like with more rooms comes more spacious houses.

Third Visual

```
# Additional effort: Check for outliers using boxplot
boxplot(Boston$medv, main="Boxplot of Median House Value", ylab="Median Value ($1000s)")
# Comments: Identifies outliers (e.g., medv > 1.5*IQR above Q3) for potential removal if needed.
```

Figure 8:Code for Boxplot of median house value

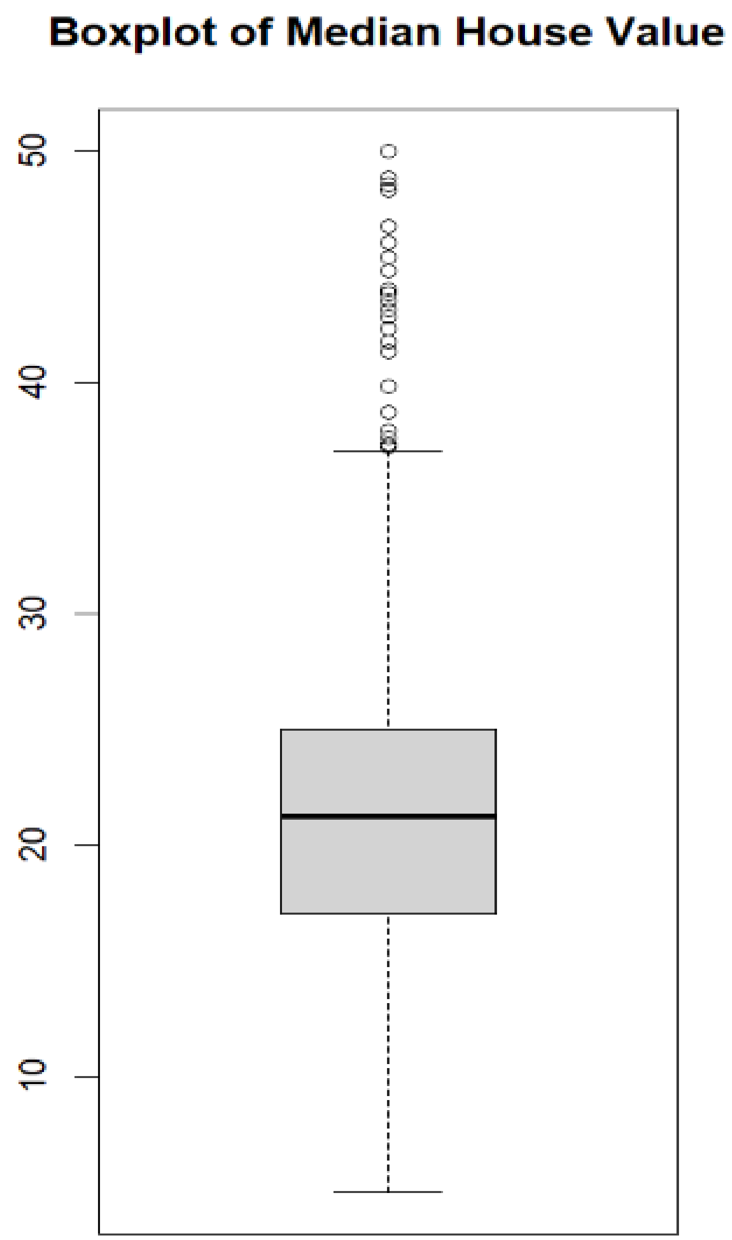

Figure 9: Boxplot of Median House Value

Interpretation:

The above figure represents the boxplot of for the median House value which displays a median value of around 22, minimum of around 3, maximum of around 37, lower quartile of 18 and upper quartile of 25. The diagram also shows small circle; each circle represents an outlier detected.

Question 2:

Part a:

```
# Q2a: Identify two groups in the dataset (5 marks)
# Identify two groups based on pupil-teacher ratio (ptratio) split at median
median_ptratio <- median(Boston$ptratio)  # Calculate median ptratio (~19.05)
low_ptratio <- Boston$medv[Boston$ptratio <= median_ptratio]  # Group 1: Lower ptratio areas
high_ptratio <- Boston$medv[Boston$ptratio > median_ptratio]  # Group 2: Higher ptratio areas
# Explanation of variable 1: low_ptratio represents areas with better student-teacher ratios (below median).
# Explanation of variable 2: high_ptratio represents areas with higher student-teacher ratios (above median).
print(head(low_ptratio))  # Sample output for low_ptratio group
print(head(high_ptratio))  # Sample output for high_ptratio group
```

Figure 10: Code to divide dataset into 2 groups

```
print(head(low_ptratio))  # Sample output for low_ptratio group
.] 24.0 21.6 34.7 33.4 36.2 28.7
print(head(high_ptratio))  # Sample output for high_ptratio group
.] 20.4 18.2 19.9 23.1 17.5 20.2
```

Figure 11:Output to dividing dataset into groups

Explanation:

Using the Boston housing dataset two groups were identified based on the pupil to teacher ratio (ptratio) which represented the number of students per the number pf students per teacher in each suburban area so the dataset was split at the median of ptratios which was 19.05 to create two distinct groups.

Variable 1: Low pupil-teacher ratio group (low_ptratio)

This group included suburbs where the pupil-teacher ration is less than or equal to 19.05 which represented areas with relatively more fewer students per teacher which could indicate higher educational quality which can affect housing desirability and might increase prices.

Variable 2: High Pupil-Teacher Ratio group (high_ptratio)

This group includes suburbs where the pupil-teacher ratio is more than 19.05 which represents areas with relatively more students per teacher which ,may indicate a less favourable educational environment which can impact the median house prices negatively.

These 2 groups will be used later to test the hypothesis .

Part b:

Null Hypothesis (H0):

There is no difference or negligible difference in the mean median house value (medv) between the low pupil-teacher ratio and the high pupil-teacher ratio which leads to assumption that pupil-teacher ratio doesn't affect the median house value

Alternative Hypothesis (H1):

There is a significant difference in the mean median house value (medv) between low pupil-teacher ratio group and high pupil-teacher ratio group which suggest that pupil to teacher ratio impacts significantly the house prices

Part c:

```
# Q2c: Perform an appropriate statistical test (5 marks)
# Perform a two-sample t-test to compare means of the two groups
t_test_result <- t.test(low_ptratio, high_ptratio, var.equal=FALSE)  # Welch t-test for unequal variances
print(t_test_result)  # Output includes t-statistic, degrees of freedom, and p-value
```

Figure 12: Code for statistical test

```
> # Q2c: Perform an appropriate statistical test (5 marks)
> # Perform a two-sample t-test to compare means of the two groups
> t_test_result <- t.test(low_ptratio, high_ptratio, var.equal=FALSE)  # Welch t-test for unequal variances
> print(t_test_result)  # Output includes t-statistic, degrees of freedom, and p-value

	Welch Two Sample t-test

data:  low_ptratio and high_ptratio
t = 12.328, df = 484.54, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.435316 10.254803
sample estimates:
mean of x mean of y 
 26.95534  18.11028 
```

Figure 13: Output for Statistical Test

Part d:

Interpret the Results of the Test

The test conducted to test the Null hypothesis and alternative hypothesis was a Welch two Sample t-test which compares the means of the medv(median house values) between low_ptratio and high_ptratio groups in the Boston housing dataset, this test is good because it doesn’t assume equal variance between the two groups which helps in dealing the scenario of potential difference in variability of house values across suburbs with differing. The results were as followed:

- t-value: 12.328
- Degree of freedom: 484.54
- p-value: <2.2e-16
- 95% confidence interval : 7.435 to 10.254
- Sample mean: 26.955 (low_ptratio) and 18.110(high_ptrtaio)

The test statistic with a value of 12.328 indicates a substantial difference between the 2 groups means relative to standard error and with the degree of freedom being at 484.54 it represents the sample size which was adjusted for an unequal variance. In addition, to the p-value was <2.2e-16 which was extremely small and far below 0.05 which led to the rejection for the Null hypothesis and supporting the aforementioned Alternate hypothesis and with the 95% confidence interval being from 7.43 to 10.25 this difference in interval confirms the true difference between the two groups is positive and significant .Lastly the sample means of both groups illustrated that suburbs with lower pupil to teacher ratio will have a higher average house value than higher pupil to teacher ratio by around 8,850 USD (26.9k-18.1k).

Question 3

Part a:

Response variable:

The median house price (medv) will be chosen as the response variable since it is the primary outcome variable in this dataset which represents the housing marks trends, it ranged from 5 thousand to 50,000 dollars in a continuous data type nature which makes it perfect for regression analysis. The goal of is to predict the medv based on characteristics of suburbs and factors.

Equation:

The regression model is a multiple linear regression model which is expressed as:

$$medv = \beta0 + \beta1 * rm + \beta2 * lstat + \beta3 * crim + \beta4 * ptratio + \varepsilon$$

the following predictors are as followed, rm is the average number of rooms in the house which typically influence medv positively meaning that a higher rm is expected to increase medv. Secondly, is the lstat which represents the percentage of lower status population which reflects the socioeconomic status of the area which could be affected by factors like income levels, safety of area, usually higher lstat would negatively impact medv and lead to lower median house prices. Moreover, the crim is the crime rate of the area which may drastically cause a negative impact on medv and then lastly is ptratio which is the pupil-teacher ratio. All these four predictors would help the model capture various factors like rooms, socioeconomic stats, safety and educational aspects to provide a framework to help the model to explain the variations of the median house value (medv) across the dataset.

Then there are 6 components in the model, the first one is β0 which represents the y-intercept estimated when all the predictors are zero, it just acts as a baseline because in the context of number of rooms(rm) a house with zero rooms is unrealistic. On the other hand, β1,β2,β3,β4 are the coefficients for rm, lstat, crim and ptratio respectively which quantity their individual effects on the median house price (medv) and the last one is the ε which represents the error terms which calculates for random variation and unmodeled factors.

Part b:

```
# Q3b: Fit a multiple linear regression model (5 marks)
# Fit the regression model with selected predictors
model <- lm(medv ~ rm + lstat + crim + ptratio, data=Boston)
# Comments: Predictors chosen for their potential impact on house value; rm (positive), lstat/crim/ptratio (negative influence).
print(summary(model))  # Display model summary for screenshot
```

Figure 14: Code to train multiple linear regression

Part c:

```
> # Q3c: Display the summary of the regression model (5 marks)
> # Display detailed summary including coefficients, R-squared, and p-values
> summary(model)  # Provides statistical significance and model fit

Call:
lm(formula = medv ~ rm + lstat + crim + ptratio, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.5492  -3.1454  -0.9357   1.6894  30.1563 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.92330    3.97564   4.257 2.48e-05 ***
rm           4.61862    0.42716  10.812  < 2e-16 ***
lstat       -0.53431    0.04564 -11.708  < 2e-16 ***
crim        -0.06544    0.03081  -2.124   0.0342 *  
ptratio     -0.88969    0.11883  -7.487 3.19e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.211 on 501 degrees of freedom
Multiple R-squared:  0.6815,	Adjusted R-squared:  0.6789 
F-statistic:   268 on 4 and 501 DF,  p-value: < 2.2e-16
```

Figure 15: Summery of the regression model

Part d:

Interpret the Coefficients of the Model

- **Intercept (β0 = 16.92330):**

  The positive value of 16.9 thousand is statistically significant with p= 2.48e-05 it indicated that model has a valid baseline, but its practical interpretation is limited due to unrealistic conditions.

- **rm (β1 = 4.61862):**

  this coefficient indicates that for each additional room per dwelling the median house value (medv) increases by approximately 4.62 thousand dollars with all the other factors held constant. This positive coefficient has high statistical significance also due to p-value being equal to <2e-16. This suggest that larger houses with more rooms are valued higher in Bostone de to more space which makes rm (number of rooms) a key factor for house prices.

- **lstat (β2 = -0.53431):** For each 1% increase in the lower-status population, the median house value decreases by about $534, holding other predictors constant. The p-value (< 2e-16) indicates high significance, implying that areas with higher proportions of lower-status residents tend to have lower house values, likely due to socioeconomic factors.

- **crim (β3 = -0.06544):** For each unit increase in per capita crime rate, the median house value decreases by about $65, holding other predictors constant. The p-value (0.0342) is below 0.05, indicating statistical significance, though less strong than rm or lstat. This suggests higher crime rates reduce house values, as safety is a key factor for buyers.

- **ptratio (β4 = -0.88969):** For each unit increase in the pupil-teacher ratio, the median house value decreases by about $890, holding other predictors constant. The p-value (3.19e-13) shows high significance, indicating that higher pupil-teacher ratios (suggesting lower-quality schools) negatively impact house values.

Question 4

Part a:

```
# Q4a: Create a new data frame with hypothetical values (10 marks)
# Create a new data frame with hypothetical predictor values
new_data <- data.frame(
  rm = c(6.5, 7.0, 5.5),        # Average rooms for different houses
  lstat = c(10, 5, 15),          # % lower status population
  crim = c(0.5, 1.0, 2.0),       # Crime rate per capita
  ptratio = c(15, 18, 20)        # Pupil-teacher ratio
)
# Comments: Represents three hypothetical housing scenarios with varied characteristics.
print(new_data)  # Display new data frame for screenshot
```

Figure 16: Code to create a new dataframe

```
> new_data <- data.frame(
+   rm = c(6.5, 7.0, 5.5),        # Average rooms for different houses
+   lstat = c(10, 5, 15),          # % lower status population
+   crim = c(0.5, 1.0, 2.0),       # Crime rate per capita
+   ptratio = c(15, 18, 20)        # Pupil-teacher ratio
+ )
> # Comments: Represents three hypothetical housing scenarios with varied characteristics.
> print(new_data)  # Display new data frame for screenshot
   rm lstat crim ptratio
1 6.5    10  0.5      15
2 7.0     5  1.0      18
3 5.5    15  2.0      20
>
```

Figure 17: Output to new dataframe

A new data frame named new_data is created with 3 rows where each row represents a hypothetical Boston suburbs with assigned values for the predictors used in the regression model, these values were chosen since they are realistic based on the dataset's ranges from Q1 part c summary(Boston) which represented various which will lets us test how the model predicts house prices for new scenarios.

- **House 1:** 6.5 rooms, 10% lower-status population, 0.5 crime rate, 15 students per teacher.
- **House 2:** 7 rooms, 5% lower-status population, 1.0 crime rate, 18 students per teacher.
- **House 3:** 5.5 rooms, 15% lower-status population, 2.0 crime rate, 20 students per teacher.

Part b:

```
# Q4b: Use the regression model to predict the response variable (10 marks)
# Predict medv for the new data points
predictions <- predict(model, newdata=new_data)
print("Predicted Median House Values ($1000s):")
print(predictions)  # Display predictions for screenshot
# Comments: Applies the fitted model to estimate medv for new data.
```

Figure 18: Code to using regression model to predict

```
> # Q4b: Use the regression model to predict the response variable (10 marks)
> # Predict medv for the new data points
> predictions <- predict(model, newdata=new_data)
> print("Predicted Median House Values ($1000s):")
[1] "Predicted Median House Values ($1000s):"
> print(predictions)  # Display predictions for screenshot
       1        2        3 
28.22316 30.50225 16.38637 
> # Comments: Applies the fitted model to estimate medv for new data.
>
```

Figure 19: Output of regression model predictions

The predicted values for each scenario was as followed

- **House 1:** Predicted price is $28,223.
- **House 2:** Predicted price is $30,502.
- **House 3:** Predicted price is $16,386.

Question 5:

Part a:

```
# Q5a: Calculate the R-squared value of the model (5 marks)
# Extract R-squared and adjusted R-squared from the model summary
r_squared <- summary(model)$r.squared
adj_r_squared <- summary(model)$adj.r.squared
cat("R-squared:", r_squared, "\nAdjusted R-squared:", adj_r_squared, "\n")
# Explanation: R-squared (e.g., ~0.6-0.7) shows the proportion of variance in medv explained by the model.
```

Figure 20: Code to calculate R-squared

```
> # Q5a: Calculate the R-squared value of the model (5 marks)
> # Extract R-squared and adjusted R-squared from the model summary
> r_squared <- summary(model)$r.squared
> adj_r_squared <- summary(model)$adj.r.squared
> cat("R-squared:", r_squared, "\nAdjusted R-squared:", adj_r_squared, "\n")
R-squared: 0.6814923
Adjusted R-squared: 0.6789494
```

Figure 21: R-squared output

R-squared which is also known as the coefficient of determination which represents the proportion of variance in the response variable that is explained by the predictor variables with the attained value of 0.6815 means that the model accounts for 68.15% of the variability in medv (median house price) due to rm, lstat, crim and ptrtaio which means that the remaining 32% are due to factors that are not in the model like location or house age

On the other hand, Adjusted R-squared is similar to R-squared but is used to adjust the R-squared value for the number of the predictions in the model by penalizing the addition of irrelevant variables sop with a value of 0.6789 means that 67.89% of the variance in medv is explained by the model. The small defence between 0.6815 and 0.6789 which is 0.0026 suggests that the 4 predictors are relevant since the penalty added to them is very minimal.

Part b:

```
# Q5b: Plot the residuals vs fitted values (6 marks)
# Plot residuals against fitted values to check for patterns
plot(fitted(model), residuals(model),
     main="Residuals vs Fitted Values",
     xlab="Fitted Values ($1000s)", ylab="Residuals", pch=19, col="blue")
abline(h=0, col="red")  # Add a horizontal line at zero for reference
# Comments: Visualizes residual distribution to assess model assumptions.
# Explanation: Random scatter around zero suggests linearity and homoscedasticity; patterns indicate issues.
```

Figure 22:Code to plot residuals vs fitted values

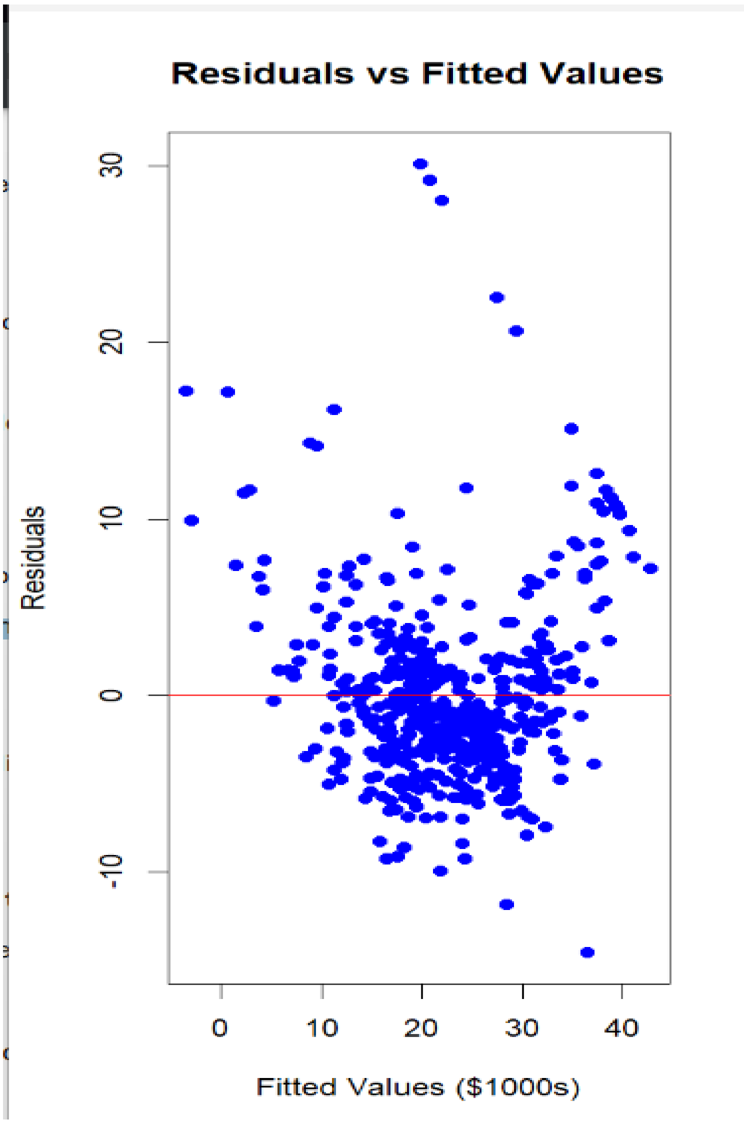

Figure 23: Residual vs Fitted values graph

Explanation:

The residual vs fitted values plot used to assess the fit of the regression model that is predicting the median house value based on the 4 predictors rm, lstat, crim, ptratio and the plot will displays the residuals which is the difference between the observed and predicted medv values and will be on the y axis meanwhile on the x axis is the predicted values. Blue points represent the data points and the red horizontal line at zero indicates the ideal residual value where prediction perfectly match observation. Moreover the residuals scattering randomly around zero shows that there is good linearity but with a increased variability at higher variability which could means there is mild heteroscedasticity.

Part c:

A regression model needs certain conditions (assumptions) to be reliable. We check these using the residuals vs. fitted values plot and other ideas from the code. Here are the four main assumptions and whether they're likely met:

1. **Linearity:**

   The linearity assumptions require a linear relationship between the predictors and response variable so since residuals are randomly scattered at zero with no clear curvilinear pattern this indicates that the linear model captures the linear relationship between medv and the predictors effectively for most data points. Therefore, the linearity assumption is satisfied as the random scatter in the residual plot support the use of the linear model for the vast majority of the data with only few minor deviations from a linear pattern exists at the high end and do not substantially influence the fit overall.

2. **Independence:**

   The independence assumption states that observations are independent from one another. This is a design assumption based on the Boston housing datasets which compromises data from distinct suburbs with no apparent temporal specified, support this. The residual plot shows no clustering or patterns that would suggest dependence and with the large sample size around 506 observations will help reduce the likelihood of a systematic dependencies making the assumption satisfied due to the lack of clustering in the residual plot which provide no evidence of spatial or temporal trend allowing observations to be treated as independent.

3. **Homoscedasticity (Equal Variance):**

   Homoscedasticity requires the variance of the residuals to be constant across all the fitted values, so the residual vs fitted plot have shown relatively consistent spread of residuals for most fitted values which supports the assumption but there has been a trends where variability increased at higher fitted values here residuals are more than 30,000 dollars which could be due to outliers which leads to mild heteroscedasticity. As a result, the Homoscedasticity assumption is almost satisfied which suggest that the models are generally robust.

4. **No Perfect Multicollinearity:**

   The absence of perfect multicollinearity will enable the model to produce unique and unstable coefficient making it essential. The model's convergence and significant coefficients with all p value being less than 0.05 suggests no sever multi-collinearity and the small difference between R-squared and adjusted R-squares suggests the there ius very low multicollinearity. Therefore, the multi-collinearity assumptions are satisfied proving the model's estimations are reliable.