Statistical analysis of Boston House Price

Name: Moon Karmakar

Course Title: Statistics for Business

Course Code: MGT7177

Student Id: 40389123

Date: 3rd August 2023

Word count: 2067

Contents

1.0 INTRODUCTION AND BACKGROUND:

The housing market plays a pivotal role in the world of real estate, necessitating that real estate agencies understand the factors influencing house prices for accurate property valuation and sales. This statistical analysis examines the Boston House Price dataset, providing valuable insights into the average characteristics and economic factors of Boston neighborhoods' houses. It is crucial to note that these data are neighborhood-level aggregates and do not represent individual property prices(Glaeser et al., n.d.).

This study's primary objective is to identify influential factors affecting Boston property prices. A thorough analysis will reveal relationships between numerous independent variables and the dependent variable, i.e., the sale price of the property. Through this study, real estate professionals can gain a deeper comprehension of the significant factors influencing property values, allowing for more precise valuations and strategic decisions ("ANALYSIS AND PREDICTION OF REAL ESTATE PRICES: A CASE OF THE BOSTON HOUSING MARKET," 2018).

To achieve this objective, descriptive statistics will be utilized to gain a comprehensive understanding of the dataset. In addition to addressing any potential data quality issues, the investigation procedure will also involve handling any potential data quality issues. Then, hypotheses will be developed to assess the relationships between five selected independent variables and the house sale price. These hypotheses will be supported and justified by existing literature, and findings will be compared to previous studies in the field (Sanyal et al., 2022).

In addition, regression models will be developed to investigate anticipated correlations and to generate predictions on a test dataset in order to evaluate the model's accuracy. This analysis will provide valuable insights into the factors influencing Boston's housing market, paving the way for improved property valuation and informed real estate industry decision-making (Association & 1896, n.d.).

2.0 METHODOLOGY

Data Preprocessing: During this phase, the dimensions of the dataset were examined; 333 rows and 14 columns were discovered. Afterwards, a data summary containing statistical information for each variable was generated (Famili et al., n.d.).

```
> dim(data)
[1] 333  14
```

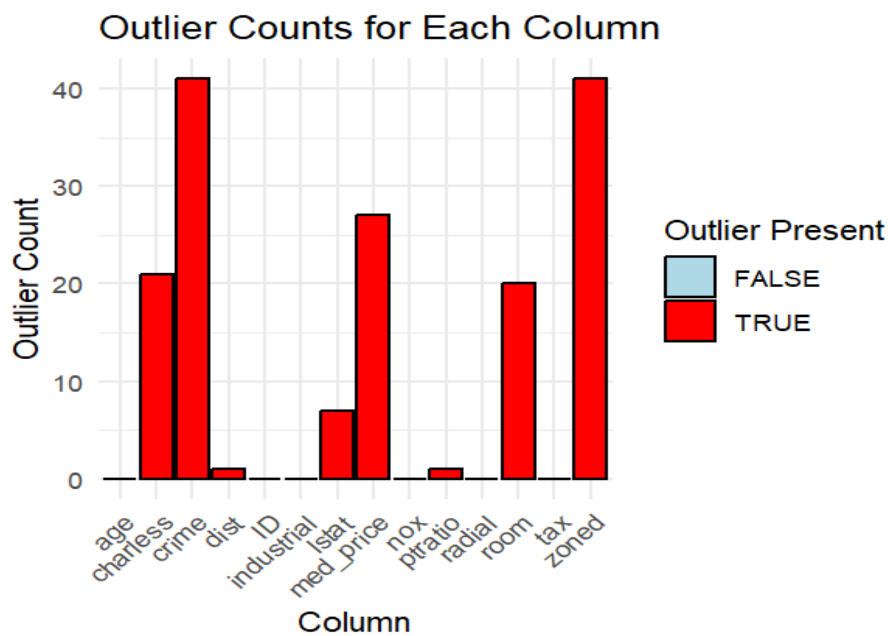
```
> summary(data)
       ID          crime             zoned          industrial       charless
 Min.   :  1   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.74   Min.   :-1.00000
 1st Qu.:123   1st Qu.: 0.07896   1st Qu.:  0.00   1st Qu.: 5.13   1st Qu.: 0.00000
 Median :244   Median : 0.26169   Median :  0.00   Median : 9.90   Median : 0.00000
 Mean   :251   Mean   : 3.36034   Mean   : 10.95   Mean   :11.29   Mean   : 0.05105
 3rd Qu.:377   3rd Qu.: 3.67822   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.: 0.00000
 Max.   :506   Max.   :73.53410   Max.   :100.00   Max.   :27.74   Max.   : 1.00000
                                  NA's   :8
      nox             room            age              dist            radial
 Min.   :0.3850   Min.   :3.561   Min.   :  6.00   Min.   : 1.130   Min.   : 1.000
 1st Qu.:0.4530   1st Qu.:5.884   1st Qu.: 45.40   1st Qu.: 2.122   1st Qu.: 4.000
 Median :0.5380   Median :6.202   Median : 76.70   Median : 3.092   Median : 5.000
 Mean   :0.5571   Mean   :6.266   Mean   : 68.29   Mean   : 3.710   Mean   : 9.634
 3rd Qu.:0.6310   3rd Qu.:6.595   3rd Qu.: 93.80   3rd Qu.: 5.117   3rd Qu.:24.000
 Max.   :0.8710   Max.   :8.725   Max.   :120.00   Max.   :10.710   Max.   :24.000

      tax           ptratio          lstat          med_price
 Min.   :188.0   Min.   :12.60   Min.   : 1.73   Min.   :15.00
 1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 7.18   1st Qu.:27.40
 Median :330.0   Median :19.00   Median :10.97   Median :31.60
 Mean   :409.3   Mean   :18.45   Mean   :12.52   Mean   :32.77
 3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.42   3rd Qu.:35.00
 Max.   :711.0   Max.   :21.20   Max.   :37.97   Max.   :60.00
```

.

1. Data Quality Check: Through arranging the data by 'ID' and performing a thorough examination of the dataset, it was determined that eight missing values existed in the data.
   Missing Value Handling: Rows with missing values were removed using the "na.omit" function to ensure data completeness.

```
> cat("Number of Missing Values:", num_missing_values, "\n")
Number of Missing Values: 8
```

2. Outlier Detection: Boxplots were used to detect outliers in each column. The Interquartile Range (IQR) method was applied to calculate the lower and upper bounds, and the number of outliers for each variable was counted.



Outlier Treatment: Outliers in specific variables ("crime," "dist," "lstat," "zoned," "ptratio," and "room") were treated by replacing them with the boundaries determined by the IQR.

```
> print(outlier_counts)
        ID      crime      zoned industrial   charless        nox       room        age       dist
         0         41         41          0         21          0         20          0          1
    radial        tax    ptratio      lstat  med_price
         0          0          1          7         27
```

3. Data Transformation: The "age" column was transformed to handle extreme values. Ages below 17 were replaced with 17, and ages above 100 were replaced with 100.

```
> cat("Rows and Columns after removing outliers:", dim(data)[1], "rows,", dim(data)[2], "columns")
Rows and Columns after removing outliers: 325 rows, 14 columns
```

Data Encoding: The "charless" column was converted into a variable whose values were encoded as 0 or 1. This encoding method facilitates data representation and analysis, allowing the investigation of relationships and patterns involving this categorical feature.

Data visualization was conducted using a pairs plot, which allowed us to explore the relationships between variables. Each unique 'ID' was represented by a distinct color, which facilitated the examination of patterns and associations between various variables

4. A correlation analysis was performed to calculate the Pearson correlation coefficients between variables. The resulting correlation matrix was then used to create a heatmap, which visually represents the strength and direction of the correlations between the variables (Automatica & 1980, n.d.).

5. Hypothesis Testing: A t-test was performed to assess the significance of the binary variable "charless" on the median house price ("med_price"). The p-value from the t-test result was reported as 0.0075, indicating a statistically significant relationship between "charless" and "med_price."

```
> cat("P-value:", p_value)
P-value: 0.007527152
> print(correlation_matrix)
                     ID        crime       zoned  industrial      charless         nox        room
ID          1.0000000000  0.640467604 -0.17050032  0.41953862  0.0002692455  0.43676508 -0.09664584
crime       0.6404676044  1.000000000 -0.36833367  0.60688861 -0.0059501463  0.66980611 -0.32793147
zoned      -0.1705003228 -0.368333671  1.00000000 -0.57699080 -0.0425939305 -0.54092054  0.38672950
industrial  0.4195386186  0.606888614 -0.57699080  1.00000000  0.0446190621  0.74972484 -0.46820553
charless    0.0002692455 -0.005950146 -0.04259393  0.04461906  1.0000000000  0.08632688  0.13753405
nox         0.4367650771  0.669806108 -0.54092054  0.74972484  0.0863268824  1.00000000 -0.35906807
room       -0.0966458390 -0.327931469  0.38672950 -0.46820553  0.1375340549 -0.35906807  1.00000000
age         0.2574432640  0.529006464 -0.58304809  0.64980523  0.0704150285  0.74457062 -0.27726770
dist       -0.3509618308 -0.541737231  0.67503688 -0.70371644 -0.1052345334 -0.77191162  0.28040014
radial      0.7110507282  0.923012868 -0.33781893  0.56489743  0.0054473699  0.61008154 -0.27575850
tax         0.6876272891  0.860069953 -0.37998374  0.70632519 -0.0329808086  0.66889838 -0.36834232
ptratio     0.3040499010  0.415286222 -0.42423464  0.38695123 -0.1354455779  0.18776376 -0.37067012
lstat       0.2822878542  0.600562998 -0.43231947  0.62374905 -0.0614724800  0.61294080 -0.63903229
med_price  -0.2096811123 -0.417915606  0.36246081 -0.46933197  0.2281664164 -0.40952118  0.69618328
                   age       dist      radial         tax    ptratio       lstat  med_price
ID          0.25744326 -0.3509618  0.71105073  0.68762729  0.3040499  0.28228785 -0.2096811
crime       0.52900646 -0.5417372  0.92301287  0.86006995  0.4152862  0.60056300 -0.4179156
zoned      -0.58304809  0.6750369 -0.33781893 -0.37998374 -0.4242346 -0.43231947  0.3624608
industrial  0.64980523 -0.7037164  0.56489743  0.70632519  0.3869512  0.62374905 -0.4693320
charless    0.07041503 -0.1052345  0.00544737 -0.03298081 -0.1354456 -0.06147248  0.2281664
nox         0.74457062 -0.7719116  0.61008154  0.66889838  0.1877638  0.61294080 -0.4095212
room       -0.27726770  0.2804001 -0.27575850 -0.36834232 -0.3706701 -0.63903229  0.6961833
age         1.00000000 -0.7771736  0.45166810  0.51914123  0.2641628  0.60174890 -0.3634492
dist       -0.77717356  1.0000000 -0.47759219 -0.53139570 -0.2262953 -0.51194506  0.2420433
radial      0.45166810 -0.4775922  1.00000000  0.90106371  0.4680611  0.48829971 -0.3429203
tax         0.51914123 -0.5313957  0.90106371  1.00000000  0.4647141  0.55261205 -0.4420428
ptratio     0.26416279 -0.2262953  0.46806110  0.46471405  1.0000000  0.37526493 -0.4775877
lstat       0.60174890 -0.5119451  0.48829971  0.55261205  0.3752649  1.00000000 -0.7502923
med_price  -0.36344921  0.2420433 -0.34292033 -0.44204281 -0.4775877 -0.75029231  1.0000000
>
```

3.0 HYPOTHESES

H0 :- No relationship between room and med_price -- Null

H1 :- Relationship between crime and med_price -- Positive

H2 :- Relationship between room and med_price -- Positive

H3 :- Relationship between industrial and med_price -- Positive

H4 :- Relationship between age and med_price -- Positive

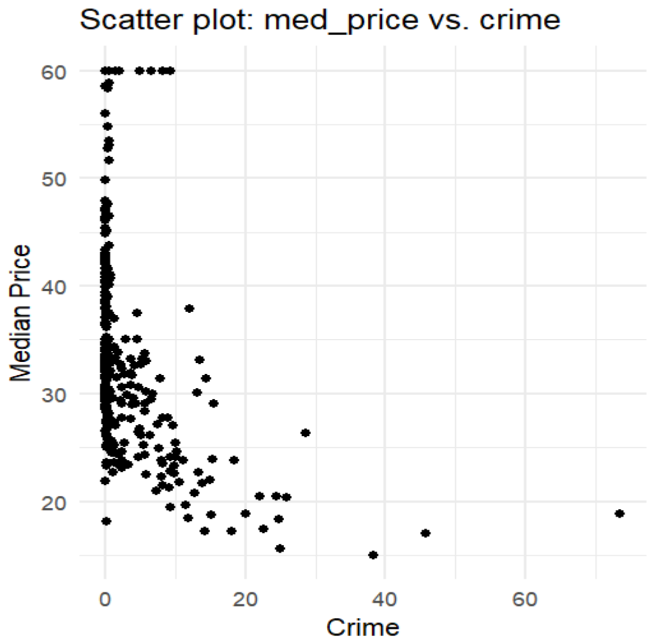
H5 :- Relationship between lstat and med_price -- Negative

These hypotheses are based on the visualizations and exploratory data analysis performed in the code, which provides insights into potential relationships between various features and the target variable (median housing price) in the Boston housing dataset (*Enterprise Knowledge Management: The Data Quality Approach - David Loshin - Google Books*, n.d.).

4.0 VISUALIZATIONS

To visualize the correlations between the independent variables and the dependent variable (med_price), scatter plots and box plots are used to conduct an initial data exploration.

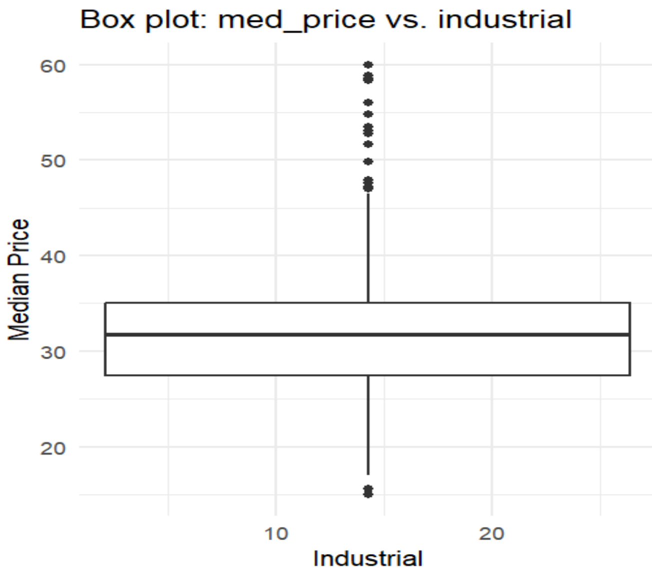
V1: Scatter plot



The scatter plot depicts the relationship between the two variables graphically. Each pixel on the graph represents a distinct community, and its position is determined by the town's crime rate (x-axis) and median home price (y-axis). Concurrently, it illustrates the relationship between the median property price and the crime rate.

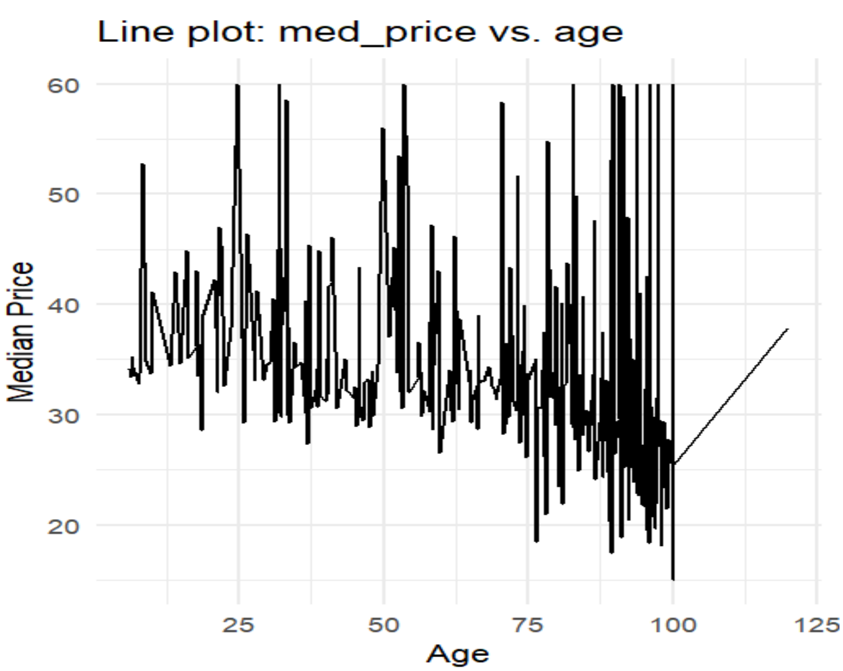
By analyzing this scatter plot, it is possible to determine whether the crime rate and the median house price exhibit any discernible pattern or trend. If the points on the graph exhibit a distinct pattern, such as movement in a particular direction, this indicates a possible relationship between the crime rate and home prices. If, on the other hand, the points are dispersed indiscriminately without any obvious pattern, this indicates a diminished or nonexistent relationship between the two variables.

V2: Box Plot



The box plot shows how the median house price varies across different groups of the "Industrial" variable. Each box on the map shows the interquartile range (IQR) of the median house price, and the straight line inside each box shows the median value. The edges go to the minimum and highest non-outlier values within 1.5 times the IQR. Any data points outside of this range are shown as single points, could be outliers.

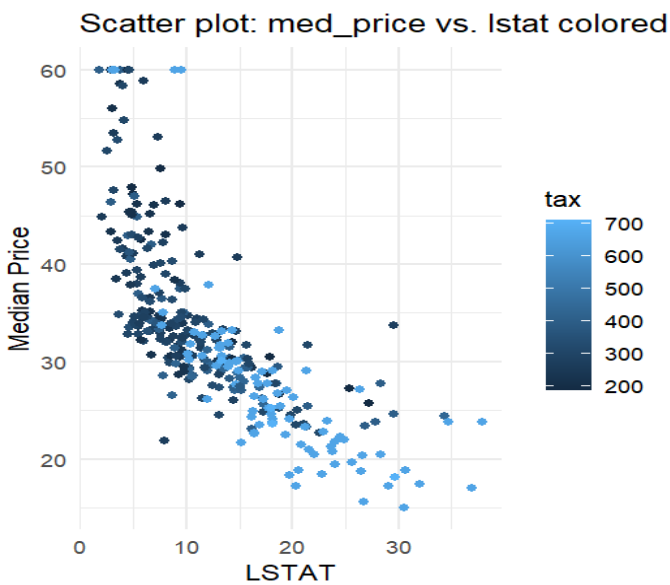
V3: Line Plot



In this line plot, the x-axis is labeled "Age," and it shows how many owner-occupied units were built before 1940. The y-axis is labeled "Median Price," and it shows how much the average house costs.

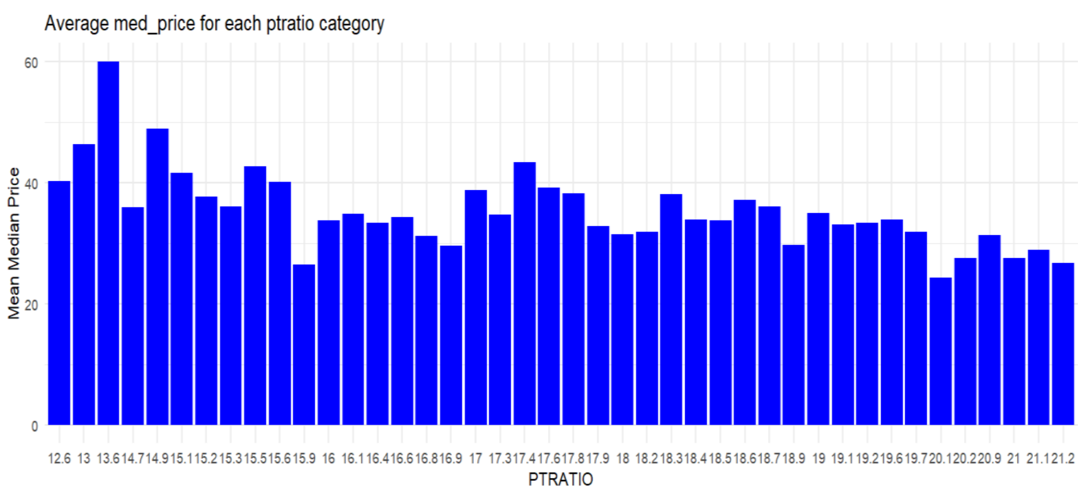
The line links the points with different "Age" numbers to show how the median house price has changed over time at different amounts of building occupancy.

V4: Scatter Plot



The scatter plot relates the "Median Price" of houses (y-axis) to the "% lower status of the population" (x-axis). Also, the colors of the data points are based on the "full-value property-tax rate per $10,000" (tax) from the collection "data." The "Tax" variable is shown by the color of each data point, which shows that each town has a different tax amount.

V5: Bar Plot



The bar plot shows the "Median Price" (y-axis) of houses for each "PTRATIO" (x-axis) group. Each bar shows the mean median price, which shows that we are looking at how the pupil-teacher ratio (PTRATIO) changes the average median house price (Fan et al., 2018).

5.0 FEATURE SELECTION:

The data set was separated into training and testing sets in order to construct and evaluate models for predicting Boston house prices. Initially, for linear regression modeling, all relevant independent variables, including crime rate, zoning information, industry share, nitric oxide percentage, and room count, were considered (Ho et al., 2021).

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared, and Adjusted R-squared were used to measure how well the models worked. These measurements were used to test how well the models could predict the future and how well they could explain the differences in median house prices.

In a later study, a group of the most important factors, including "industrial," "nox" (percentage of nitric oxides), "room," "age," and "lstat" (% of the people with a lower socioeconomic level), were looked at. With these five factors, a new linear regression model was trained, and its performance was measured.

6.0 RESULT AND DISCUSSION

Regression model and accuracy measures:

The correlation between dependent and independent variables determines the predictive power of linear regression. Model precision and dependability are enhanced by significant correlations, ideally with coefficients of 0.5 or higher. The inclusion of highly correlated variables improves predictive accuracy (Automatica & 1980, n.d.).

Split the data: When employing supervised learning algorithms, it is required to divide the data into two sets: the training set and the testing set. The training set is used to train the model by learning from its observations, whereas the testing set is used to assess the predictive performance of the model. This data partitioning improves the model's precision and prevents overfitting, ensuring that the model generalizes well to new, unseen data (*Bhalla, D. (2017)Splitting Data into Training and... - Google Scholar*, n.d.)

Single Linear Regression: The objective of simple regression analysis is to determine the influence of a predictor variable on a particular outcome. In contrast, correlation studies evaluate the strength and direction of the association between variables (Zou et al., 2003).

We will test our hypothesis as simple linear models-

```
Residuals:
     Min       1Q   Median       3Q      Max 
-13.6339  -3.4403  -0.8438   2.1123  25.5578 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.06353    5.41841   2.042   0.0422 *  
room         4.77862    0.77243   6.186 2.44e-09 ***
crime        0.02403    0.15148   0.159   0.8741    
industrial   0.00797    0.07756   0.103   0.9182    
age          0.02341    0.01812   1.292   0.1976    
lstat       -0.80358    0.08644  -9.296  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.599 on 255 degrees of freedom
Multiple R-squared:  0.6326,	Adjusted R-squared:  0.6253 
F-statistic: 87.79 on 5 and 255 DF,  p-value: < 2.2e-16
```

In this investigation, R-squared value of 0.20 to 0.30 or higher are typical for the majority of hypotheses indicating a positive relationship. We reject the relationship between Istat and med_price as it have a negative value.

Multiple Linear Regression: The primary objective of multiple regression analysis is to explore the correlations among more than two variables, with a focus on identifying cause-and-effect relationships. This analysis allows us to utilize these relationships to make predictions about the outcome (Eberly, 2007) .

In our study, we developed three distinct regression models, each incorporating different sets of related variables. We then proceeded to assess the accuracy of the predicted prices by analyzing the regression summary of these models (Kumara Swamy et al., 2017).

To determine the significance of the relationships between variables, we examined the p-values in the regression summary. A p-value below 0.05 indicates a statistically significant relationship between the variables (Whitley & Ball, 2002). This level of significance provides us with confidence in the observed results, as it suggests that in only 5% of cases, we would draw incorrect conclusions due to chance. This level of risk or margin of error (5 out of 100) is considered acceptable for our research work (Tamhane & Gou, 2021).

EVALUATION OF ACCURACY FOR HYPOTHESIS MODEL 1

```
Call:
lm(formula = med_price ~ room, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.010  -2.920  -0.034   2.748  40.344 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -30.6178     4.3565  -7.028 1.85e-11 ***
room         10.1155     0.6907  14.645  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.779 on 259 degrees of freedom
Multiple R-squared:  0.453,	Adjusted R-squared:  0.4509 
F-statistic: 214.5 on 1 and 259 DF,  p-value: < 2.2e-16
```

*P- value of all the variables in the model is < 0.05, hence we can accept.

```
    RMSE Rsquared      MAE 
5.889455 0.611005 3.976721 
```

Accuracy of mode 1: 61.10%

Assumption Check:

Independent Error:

```
	Durbin-Watson test

data:  lin_model
DW = 0.67337, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is not 0
```

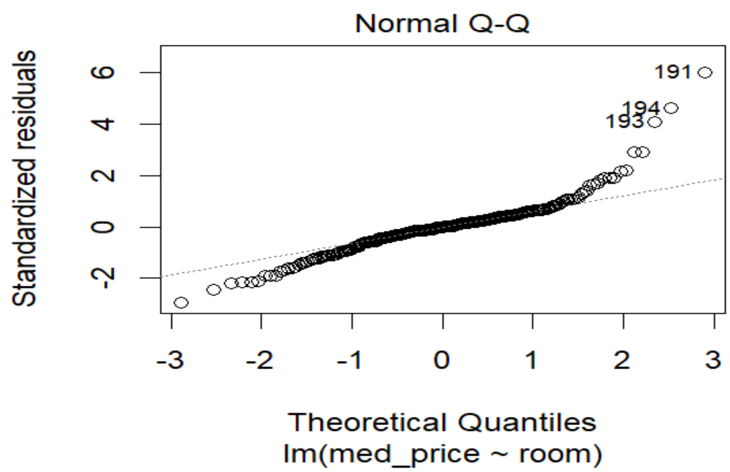
The Durbin-Watson test statistic of 0.67337 and a p-value less than 2.2e-16 suggest strong positive autocorrelation in the residuals, rejecting the null hypothesis that there is no autocorrelation.

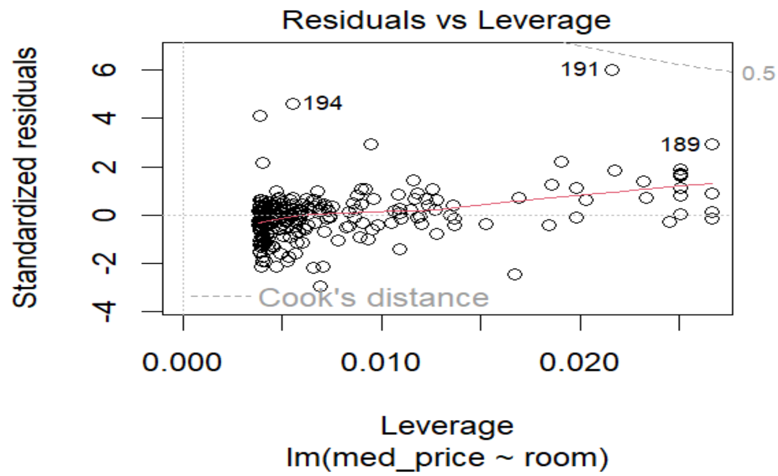
No Multicollinearity:

```
> print(vif_price_model)
   zoned  charless       room       dist      crime industrial        age      lstat 
2.145550  1.032984   1.968353   3.615567   2.026093   2.789866   3.091499   2.948575 
```

*The largest vif is not greater than 10, no areas of concerns.

Heteroscedasticity Assumption:





Based on the analysis we can further improve this model.

EVALUATION OF ACCURACY FOR HYPOTHESIS MODEL 2

```
Call:
lm(formula = med_price ~ room + crime + industrial + age + lstat,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-13.6339  -3.4403  -0.8438   2.1123  25.5578

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.06353    5.41841   2.042   0.0422 *
room         4.77862    0.77243   6.186 2.44e-09 ***
crime        0.02403    0.15148   0.159   0.8741
industrial   0.00797    0.07756   0.103   0.9182
age          0.02341    0.01812   1.292   0.1976
lstat       -0.80358    0.08644  -9.296  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.599 on 255 degrees of freedom
Multiple R-squared:  0.6326,    Adjusted R-squared:  0.6253
F-statistic: 87.79 on 5 and 255 DF,  p-value: < 2.2e-16
```

R-squared = 0.6326 indicates that approximately 63.26 percent of the response variable's variability is explained by the predictors. The F-statistic (87.79, p <2.2e-16) demonstrates that the model is extremely significant. However, "crime," "industrial," and "age" have no effect on the response, whereas "room" and "lstat" have a significant effect.

```
> print(accuracy_Hyp_model)
     RMSE  Rsquared       MAE
5.1769892 0.6969093 3.9258802
```

Accuracy is 69.69%

Assumption Check:

No Multicollinearity

```
print(vif_Hyp_model)
    room     crime industrial       age     lstat
1.833088  1.977597   2.394476  2.049350  2.894346
```

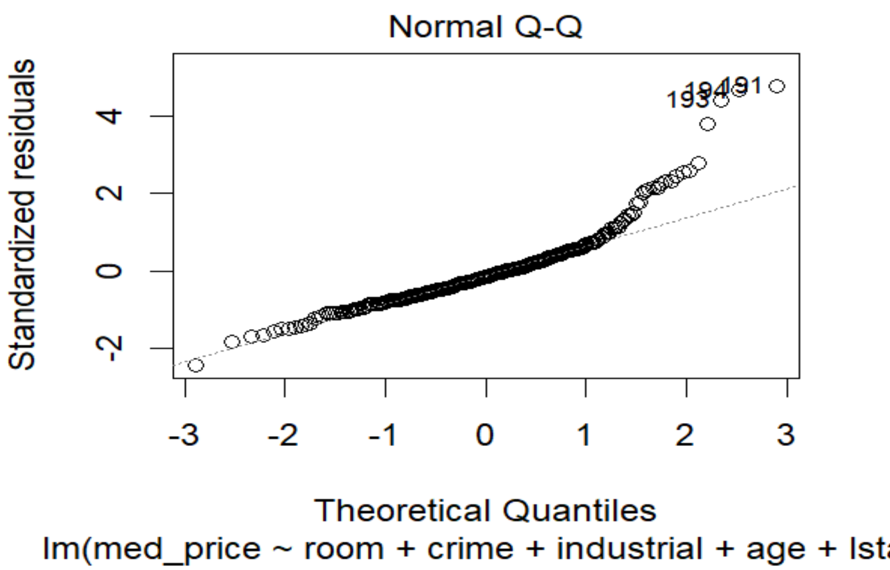
The largest vif is not greater than 10, no areas of concerns.

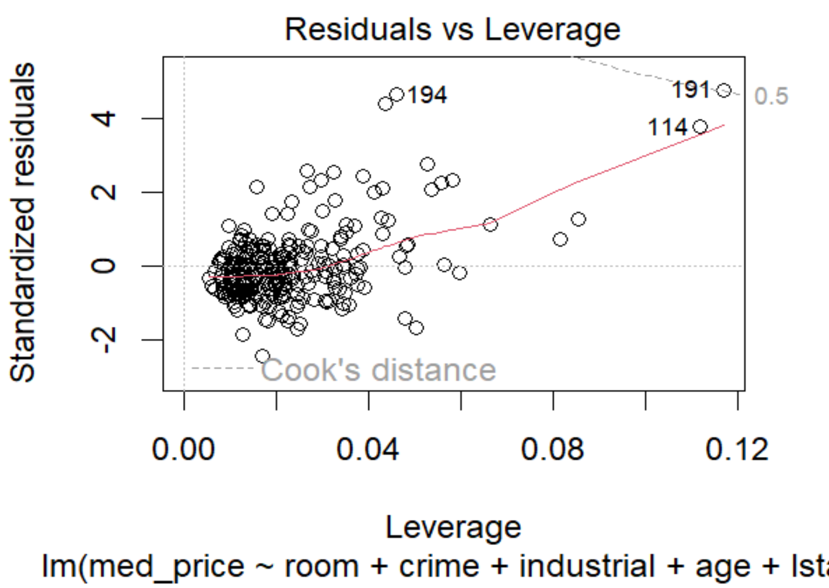
Independent Error:

```
        Durbin-Watson test

data:  Hyp_model
DW = 0.8874, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The Durbin-Watson test statistic of 0.8874 and a p-value less than 2.2e-16 indicate strong positive autocorrelation in the residuals, supporting the alternative hypothesis that true autocorrelation is greater than 0.

Heteroscedasticity Assumption:





Based on the analysis, second model is comparatively better.

EVALUATION OF ACCURACY FOR HYPOTHESIS MODEL 3

```
> print(accuracy_model2)
     RMSE  Rsquared       MAE 
5.7457779 0.6265963 4.0805264 
```

The accuracy is 62.65%

```
Residuals:
     Min       1Q   Median       3Q      Max 
-14.1551  -3.1086  -0.6961   1.9248  24.4787 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.81317    5.60212   4.251 3.00e-05 ***
zoned        0.11432    0.04071   2.808  0.00538 ** 
charless     4.51164    1.37362   3.284  0.00117 ** 
room         4.10075    0.74871   5.477 1.05e-07 ***
dist        -1.50628    0.30725  -4.902 1.70e-06 ***
crime       -0.10359    0.14342  -0.722  0.47078    
industrial  -0.09623    0.07831  -1.229  0.22029    
age         -0.01936    0.02082  -0.930  0.35319    
lstat       -0.76080    0.08161  -9.322  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.238 on 252 degrees of freedom
Multiple R-squared:  0.6823,	Adjusted R-squared:  0.6722 
F-statistic: 67.64 on 8 and 252 DF,  p-value: < 2.2e-16
```

The p-value of 3.00e-05 (***), along with the low p-values (< 0.05) for 'zoned', 'charless', 'room', and 'dist' ()** suggests a significant relationship between these predictor variables and the outcome. This indicates a statistical confidence in their impact on the outcome.

However, 'crime', 'industrial', 'age', and 'lstat' have p-values greater than 0.05, indicating that they are not statistically significant at the 5% significance level.

Assumptions check:

No Multicollinearity:

```
> print(vif_model2)
   zoned charless     room     dist    crime industrial      age    lstat 
2.145550 1.032984 1.968353 3.615567 2.026093   2.789866 3.091499 2.948575 
```

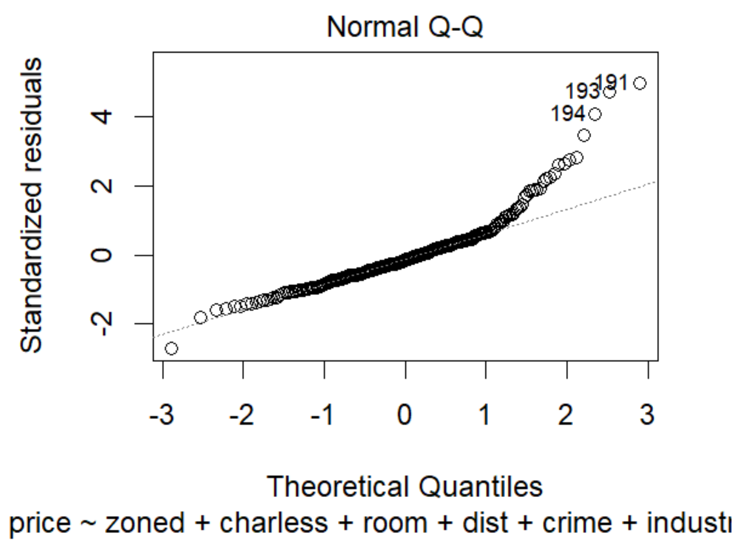
The largest vif is not greater than 10, no areas of concerns.

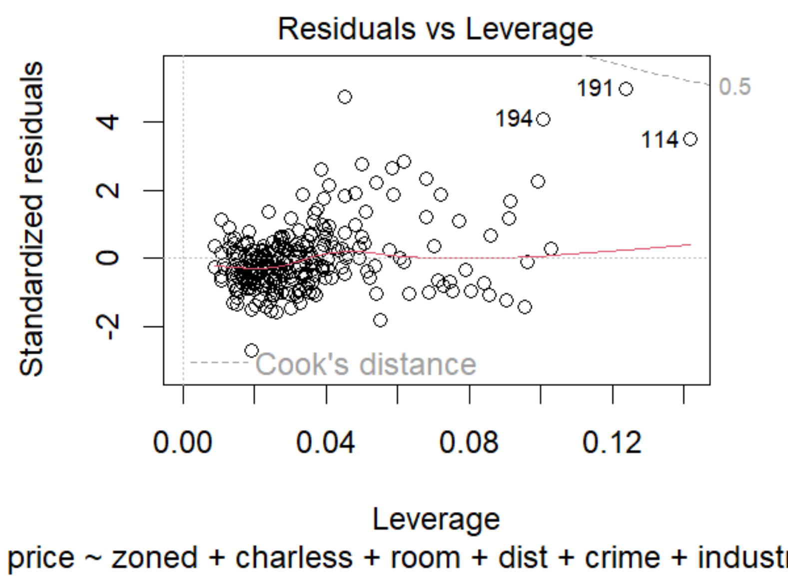
Independent error:

```
        Durbin-Watson test

data:  price_model
DW = 1.0851, p-value = 5.159e-15
alternative hypothesis: true autocorrelation is greater than 0
```

The Durbin-Watson test statistic of 1.0851 and a p-value of 5.159e-15 indicate strong positive autocorrelation in the residuals, supporting the alternative hypothesis that true autocorrelation is greater than 0.

Heteroscedasticity Assumption:





Based on the analysis we can further improve this model.

7.0 CONCLUSION

An optimal model is not always synonymous with a robust model. This data set has a high loss rate and few processing features. To improve the accuracy of forecasts, we must interpolate absent values and work on feature engineering, including the extension and selection of features. We compare model performance and employ combination forecasting to develop pertinent models, thereby enhancing the output (Oakden-Rayner et al., n.d.).

Model 2 demonstrates the highest level of predictive accuracy (69.69%) among the three models. However, all models require additional enhancements to address the issue of residual autocorrelation. In addition, Model 1 contains non-statistically significant variables, which may need to be reconsidered or refined in future iterations of the model.

8.0 REFECTIVE COMMENTARY:

I have reviewed the fundamentals of statistics and applied them to analyze real-world data throughout this session. The modules have helped me learn and improve my R skills, and I now understand how the characteristics of a home affect its price. This dataset inspired me to experiment with new software packages and methodologies, in addition to what I learned in class. This module will enhance my technical skills and make me more resilient.

REFERENCE

ANALYSIS AND PREDICTION OF REAL ESTATE PRICES: A CASE OF THE BOSTON HOUSING MARKET. (2018). *Issues In Information Systems*. https://doi.org/10.48009/2_iis_2018_109-118

Association, H. W.-P. of the A. S., & 1896, undefined. (n.d.). Real Estate Values in Boston. *Taylor & Francis*. Retrieved August 3, 2023, from https://www.tandfonline.com/doi/pdf/10.1080/15225437.1896.10504082

Automatica, K. G.-, & 1980, undefined. (n.d.). Correlation methods. *Elsevier*. Retrieved August 3, 2023, from https://www.sciencedirect.com/science/article/pii/000510988090076X

Bhalla, D. (2017)Splitting data into training and... - Google Scholar. (n.d.). Retrieved August 2, 2023, from https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Bhalla%2C+D.+%282017%29Splitting+data+into+training+and+test+sets+with+R&btnG=

Eberly, L. E. (2007). Multiple linear regression. *Methods in Molecular Biology (Clifton, N.J.)*, *404*, 165–187. https://doi.org/10.1007/978-1-59745-530-5_9

Enterprise Knowledge Management: The Data Quality Approach - David Loshin - Google Books. (n.d.). Retrieved August 3, 2023, from https://books.google.co.uk/books?hl=en&lr=&id=3BXTfCtR8zsC&oi=fnd&pg=PR13&dq=coupon+acceptance+data+quality&ots=s0hcKbdIj5&sig=BVF_jcd23vlzkyFQgS4QxoaoTHs&redir_esc=y#v=onepage&q=coupon%20acceptance%20data%20quality&f=false

Famili, A., Shen, W., … R. W.-I. data, & 1997, undefined. (n.d.). Data preprocessing and intelligent data analysis. *Content.Iospress.ComA Famili, WM Shen, R Weber, E SimoudisIntelligent Data Analysis, 1997•content.Iospress.Com*. Retrieved August 2, 2023, from https://content.iospress.com/articles/intelligent-data-analysis/ida1-1-02

Fan, C., Cui, Z., & Zhong, X. (2018). House prices prediction with machine learning algorithms. *ACM International Conference Proceeding Series*, 6–10. https://doi.org/10.1145/3195106.3195133

Glaeser, E. L., Schuetz, J., & Ward, B. (n.d.). *P B-2 0 0 6-1 | | F e b r u a r y 2 0 0 6 Regulation and the Rise of Housing Prices in Greater Boston*.

Ho, W., Tang, B., Research, S. W.-J. of P., & 2021, undefined. (2021). Predicting property prices with machine learning algorithms. *Taylor & Francis*, *38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Kumara Swamy, M., Krishna Reddy, P., & Bhalla, S. (2017). Association rule based approach to improve diversity of query recommendations. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10439 LNCS*, 340–350. https://doi.org/10.1007/978-3-319-64471-4_27/COVER

Oakden-Rayner, L., Carneiro, G., reports, T. B.-S., & 2017, undefined. (n.d.). Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework. *Nature.Com*. Retrieved August 3, 2023, from https://www.nature.com/articles/s41598-017-01931-W

Sanyal, S., Kumar Biswas, S., Das, D., Chakraborty, M., & Purkayastha, B. (2022). Boston House Price Prediction Using Regression Models. *2022 2nd International Conference on Intelligent Technologies, CONIT 2022*. https://doi.org/10.1109/CONIT55038.2022.9848309

Tamhane, A. C., & Gou, J. (2021). Multiple Test Procedures Based on p-Values. *Handbook of Multiple Comparisons*, 11–34. https://doi.org/10.1201/9780429030888-2/MULTIPLE-TEST-PROCEDURES-BASED-VALUES-AJIT-TAMHANE-JIANGTAO-GOU

Whitley, E., & Ball, J. (2002). Statistics review 3: Hypothesis testing and P values. *Critical Care*, *6*(3), 222–225. https://doi.org/10.1186/CC1493

Zou, K., Tuncali, K., Radiology, S. S.-, & 2003, undefined. (2003). Correlation and simple linear regression. *Pubs.Rsna.OrgKH Zou, K Tuncali, SG SilvermanRadiology, 2003•pubs.Rsna.Org*, *227*(3), 617–622. https://doi.org/10.1148/radiol.2273011499

APPENDIX: R CODE

```
library("readxl")

library("dplyr")

library("tidyverse")

library("ggplot2")

library("corrplot")


data <- read_excel("C:/Users/moon/Downloads/boston_housing (1).xlsx")

View(data)

library(readxl)


dim(data)

summary(data)


# Load the required library

library(RColorBrewer)


# Generate colors for each unique 'ID' in the dataset

colors <- brewer.pal(n = nlevels(as.factor(data$ID)), name = "Set1")


# Create a pairs plot with distinct colors for each 'ID'

pairs(data, col = colors[as.numeric(as.factor(data$ID))])

# Data Quality Check


# Sort the dataframe by 'ID'

sorted_data <- arrange(data, ID)


# Count the number of missing values
```

```
num_missing_values <- sum(is.na(sorted_data))

# Display the count of missing values
cat("Number of Missing Values:", num_missing_values, "\n")

# Remove rows with NA values
data <- na.omit(data)
sum(is.na(data)) #there are 8 Missing Values

#age column
data$age[data$age <= 17] <- 17
data$age[data$age >= 100] <- 100

# Convert 'charless' to a binary variable (0 or 1)
data$charless <- ifelse(data$charless == -1, 1, data$charless)
# Outliers in each column - Boxplots

# Load necessary libraries
library(ggplot2)
library(gridExtra)

# Initialize an empty list for plots
plots <- list()

# Initialize a vector to store outlier counts for each column
outlier_counts <- numeric(length(data))
names(outlier_counts) <- names(data)

# Loop through each column in the data to calculate outlier counts
for (i in names(data)) {
```

```r
  Q1 <- quantile(data[[i]], 0.25, na.rm = TRUE)
  Q3 <- quantile(data[[i]], 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  outlier_counts[i] <- sum(data[[i]] < lower_bound | data[[i]] > upper_bound, na.rm = TRUE)
}

# Create a bar plot to visualize outlier counts for each column
library(ggplot2)

outlier_counts_df <- data.frame(column = names(outlier_counts), count = outlier_counts)

ggplot(outlier_counts_df, aes(x = column, y = count, fill = count > 0)) +
  geom_col(color = "black") +
  labs(title = "Outlier Counts for Each Column",
       x = "Column",
       y = "Outlier Count",
       fill = "Outlier Present") +
  scale_fill_manual(values = c("lightblue", "red")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Arrange all plots in a grid
grid_plot <- gridExtra::grid.arrange(grobs = plots, ncol = 7, top = "Outliers in Data")

# Print the outlier counts
print(outlier_counts)

# List of variables to treat
```

```r
vars_to_treat <- c("crime", "dist", "lstat", "zoned", "ptratio", "room")

# Loop through each variable to treat and replace outliers with IQR boundaries
for (var in vars_to_treat) {
  Q1 <- quantile(data[[var]], 0.25, na.rm = TRUE)
  Q3 <- quantile(data[[var]], 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR

  data[[var]] <- pmin(pmax(data[[var]], lower_bound), upper_bound)
}

# Displaying  the dimensions of the dataset after removing outliers
cat("Rows and Columns after removing outliers:", dim(data)[1], "rows,", dim(data)[2], "columns")

# Compute the correlation matrix
correlation_matrix <- cor(data, use = "complete.obs", method = "pearson")

# Create a heatmap for the correlation matrix
library(ggplot2)
library(reshape2)

cor_data <- melt(correlation_matrix)
ggplot(cor_data, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient(low = "blue", high = "red") +
  theme_minimal() +
  labs(title = "Correlation Heatmap")
```

```
# t-test for the binary variable "charless"
t_test_result <- t.test(med_price ~ charless, data = data)

# print the p-value from the t-test result
p_value <- t_test_result$p.value
cat("P-value:", p_value)

# Compute the correlation matrix
correlation_matrix <- cor(data, use = "complete.obs", method = "pearson")

# Print the correlation matrix
print(correlation_matrix)

#visualizations
# Load the necessary libraries
library(ggplot2)

# Scatter plot: med_price vs. crime
ggplot(data, aes(x = crime, y = med_price)) +
  geom_point() +
  labs(title = "Scatter plot: med_price vs. crime", x = "Crime", y = "Median Price") +
  theme_minimal()

# Scatter plot: med_price vs. room
ggplot(data, aes(x = industrial, y = med_price)) +
  geom_point() +
  labs(title = "Scatter plot: med_price vs. industrial", x = "Number of industrial", y = "Median Price") +
  theme_minimal()

# Box plot: med_price vs. industrial
```

```
ggplot(data, aes(x = industrial, y = med_price)) +
  geom_boxplot() +
  labs(title = "Box plot: med_price vs. industrial", x = "Industrial", y = "Median Price") +
  theme_minimal()

# Line plot: med_price vs. age
ggplot(data, aes(x = age, y = med_price, group = 1)) +
  geom_line() +
  labs(title = "Line plot: med_price vs. age", x = "Age", y = "Median Price") +
  theme_minimal()

# Scatter plot: med_price vs. lstat colored by tax
ggplot(data, aes(x = lstat, y = med_price, color = tax)) +
  geom_point() +
  labs(title = "Scatter plot: med_price vs. lstat colored by tax", x = "LSTAT", y = "Median Price") +
  theme_minimal()

# Bar plot: Average med_price for each ptratio category
ggplot(data, aes(x = factor(ptratio), y = med_price)) +
  geom_bar(stat = "summary", fun = "mean", fill = "blue") +
  labs(title = "Average med_price for each ptratio category", x = "PTRATIO", y = "Mean Median Price") +
  theme_minimal()



###



#Building linear models
#Singe linear regression
# Load the caret package
```

```r
library(lattice)

library(caret)

set.seed(40389123)

index <- createDataPartition(data$med_price, list=FALSE, p=0.8, times=1)

train <- data[index,]

test <- data[-index,]


#### B### Singe linear regression


# Build the linear model 1 (lin_model)

lin_model <- lm(med_price ~ room, data = train)


# Display summary of the linear model (lin_model)

model_summary <- summary(lin_model)

print(model_summary)


# Make predictions using the linear model (lin_model) on the test data

lin_model_predictions <- predict(lin_model, newdata = test)


#Hypothesis model 2

# Create the linear model with updated variables

Hyp_model <- lm(med_price ~ room + crime + industrial + age + lstat, data = train)


# Display the summary of the linear model

summary(Hyp_model)



# Make predictions using the linear model on the test data

Hyp_model_predictions <- predict(Hyp_model, newdata = test)
```

#Multiple Leniar model-3

Build Model 2 - Multiple Linear Regression with adding additional Positive variables

price_model <- lm(med_price ~ zoned + charless + room + dist+crime + industrial+ age + lstat , data = train)

Display summary of Model 3

model_summary <- summary(price_model)

print(model_summary)

Make predictions using Model 3 on the test data

price_predictions <- predict(med_price_model, newdata = test)

#Evaluation of model

Load the 'car' package for evaluation

library(car)

Calculate the post-resampling performance of the model

library(caret)

library(ggplot2)

library(lattice)

Calculate the post-resampling performance of the model

post_resample_perf <- postResample(lin_model_predictions, test$med_price)

print(post_resample_perf)

Calculate the root mean squared error (RMSE) for the model

rmse <- sqrt(mean((lin_model_predictions - test$med_price)^2, na.rm = TRUE))

print(rmse)

Calculate the post-resampling performance of Model 1

```
accuracy_lin_model <- postResample(lin_model_predictions, test$med_price)
print(accuracy_lin_model)

# Calculate the Variance Inflation Factors (VIF) for Model 1
# Load the car package
library(car)
library(carData)



# Calculate the Variance Inflation Factors (VIF) for Model 1
vif_price_model <- car::vif(price_model)
print(vif_price_model)

# Create a diagnostic plot for Model 1
plot(lin_model)

# Load the 'lmtest' package for hypothesis testing
library(lmtest)

# Perform the Durbin-Watson test for Model 1 with lag = 1
dw_test_result <- dwtest(lin_model, alternative = "two.sided")
print(dw_test_result)

# Calculate Cook's distance for Model 1
cook_distance <- cooks.distance(lin_model)
print(sum(cook_distance > 1))

## Evaluation of accuracy for Hypothesis Model 2 ###

# Calculate the post-resampling performance of Hypothesis Model 1
```

```
accuracy_Hyp_model <- postResample(Hyp_model_predictions, test$med_price)
print(accuracy_Hyp_model)

# Calculate the Variance Inflation Factors (VIF) for Hypothesis Model 2
vif_Hyp_model <- car::vif(Hyp_model)
print(vif_Hyp_model)

# Create a diagnostic plot for Hypothesis Model 2
plot(Hyp_model)

# Load the 'lmtest' package for hypothesis testing
library(lmtest)

# Durbin-Watson test for autocorrelation in Hyp_mode2
dw_test_result <- dwtest(Hyp_model)
print(dw_test_result)

# Calculate Cook's distance for Hyp_model
cook_distance <- cooks.distance(Hyp_model)
print(sum(cook_distance > 1))



#evaluation model 3

# Calculate the post-resampling performance of Model
accuracy_model2 <- postResample(price_predictions, test$med_price)
print(accuracy_model2)

# Calculate the Variance Inflation Factors (VIF) for Model
```

```
vif_model2 <- car::vif(price_model)
print(vif_model2)

# Create a diagnostic plot for Model
plot(price_model)

# Load the 'lmtest' package for hypothesis testing
library(lmtest)

# Perform the Durbin-Watson test for Model 3
dw_test_result_model2 <- dwtest(price_model)
print(dw_test_result_model2)

# Calculate Cook's distance for Model 3
cook_distance_model2 <- cooks.distance(price_model)
print(sum(cook_distance_model2 > 1))
```