BTRY 6020 Final Report

Rujia Xie

May 14, 2025

1. Introduction

1.1 Dataset Description

This paper uses the “Boston-house-price-data” from Kaggle.com (Arun Kumar, n.d.). The Boston house price data was originally published by Harrison, D. and Rubinfeld, D.L. “Hedonic prices and the demand for clean air”, J. Environ. Economics & Management, vol.5, 81-102, 1978 (Harrison and Rubinfeld 1978).

The dataset comprises information collected by the U.S. Census Bureau concerning housing in Boston, Massachusetts. Each observation in the Boston Housing dataset represents a single census tract (neighborhood) in the Boston area. The dataset includes various attributes describing these tracts’ socioeconomic, environmental, and housing characteristics, along with the median value of owner-occupied homes. All variables in the dataset in order are:

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

1.2 Problem Statement

There are two objectives of this project. The first objective is to understand the relationship between the median value of owner-occupied homes (MEDV) and per capita crime rates (CRIM), average number of rooms per dwelling (RM), weighted distances to five Boston employment centers (DIS), and index of accessibility to radial highways (RAD). The second objective is to develop a predictive model that accurately estimates the median value of owner-occupied homes (MEDV) in Boston, MA.

By leveraging multiple linear regression techniques, this project aims to use the model to identify and quantify the impact of various factors on housing prices. Combined, this project hopes to:

a. Investigate which factors significantly impact housing prices in Boston, MA, and assess the strength and direction (positive or negative) of these correlations.

b. Employ forward and backward selection for variable selection, and evaluate the models' predictive performance using diverse metrics and validation techniques.

c. Interpret the results from the optimal model to provide actionable insights into the housing market dynamics of Boston for stakeholders, including policymakers, real estate professionals, and potential homeowners, facilitating informed decision-making in the housing sector.

2. Methodology

2.1 Data Preprocessing

Before formal analysis, data preprocessing was conducted. No missing values were found in the dataset. All variables in this dataset are continuous except CHAS (Charles River dummy variable, 1 if tract bounds river, 0 otherwise), which was converted to a categorical variable using the 'as.factor' function.

2.2 Exploratory Analysis

To investigate the first objective of the project, exploratory data analysis and multiple linear regression were conducted to analyze the relationship between five pre-specified variables.

The outcome variable is MEDV (the median value of owner-occupied homes), and the predictor variables are CRIM (per capita crime rates), RM (average number of rooms per dwelling), DIS (weighted distances to five Boston employment centers), and RAD (the index of accessibility to radial highways). These four predictors were selected because the author is interested in studying these four variables, and existing literature has indicated the impacts of crime rates and accessibility to highways on house prices ["Measuring the Impact of Crime on House Prices: Applied Economics: Vol 33, No 15" (n.d.)](Levkovich, Rouwendal, and Marwijk 2016).

In the exploratory analysis, summary statistics of all variables in the dataset were calculated, and histograms of the five selected variables were plotted (Figures 1a-1e).

```
##         vars   n  mean    sd median trimmed   mad  min    max  range  skew
## CRIM       1 506  3.61  8.60   0.26    1.68  0.33 0.01  88.98  88.97  5.19
## ZN         2 506 11.36 23.32   0.00    5.08  0.00 0.00 100.00 100.00  2.21
## INDUS      3 506 11.14  6.86   9.69   10.93  9.37 0.46  27.74  27.28  0.29
## CHAS*      4 506  1.07  0.25   1.00    1.00  0.00 1.00   2.00   1.00  3.39
## NOX        5 506  0.55  0.12   0.54    0.55  0.13 0.38   0.87   0.49  0.72
## RM         6 506  6.28  0.70   6.21    6.25  0.51 3.56   8.78   5.22  0.40
## AGE        7 506 68.57 28.15  77.50   71.20 28.98 2.90 100.00  97.10 -0.60
## DIS        8 506   3.80   2.11   3.21    3.54   1.91   1.13  12.13  11.00  1.01
## RAD        9 506   9.55   8.71   5.00    8.73   2.97   1.00  24.00  23.00  1.00
## TAX       10 506 408.24 168.54 330.00  400.04 108.23 187.00 711.00 524.00  0.67
## PTRATIO   11 506  18.46   2.16  19.05   18.66   1.70  12.60  22.00   9.40 -0.80
## B         12 506 356.67  91.29 391.44  383.17   8.09   0.32 396.90 396.58 -2.87
## LSTAT     13 506  12.65   7.14  11.36   11.90   7.11   1.73  37.97  36.24  0.90
## MEDV      14 506  22.53   9.20  21.20   21.56   5.93   5.00  50.00  45.00  1.10
##          kurtosis   se
## CRIM       36.60 0.38
## ZN          3.95 1.04
## INDUS      -1.24 0.30
## CHAS*       9.48 0.01
## NOX        -0.09 0.01
## RM          1.84 0.03
## AGE        -0.98 1.25
## DIS         0.46 0.09
## RAD        -0.88 0.39
## TAX        -1.15 7.49
## PTRATIO    -0.30 0.10
## B           7.10 4.06
## LSTAT       0.46 0.32
## MEDV        1.45 0.41
```

Figure 1a. Histogram of Median Value of Homes

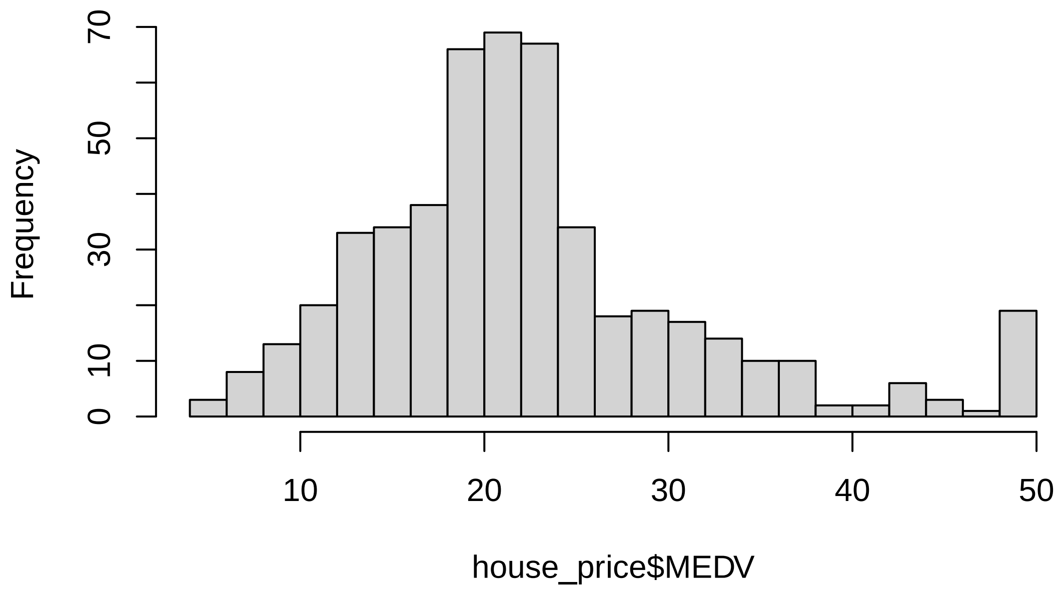


Figure 1b. Histogram of Per Capita Crime Rate

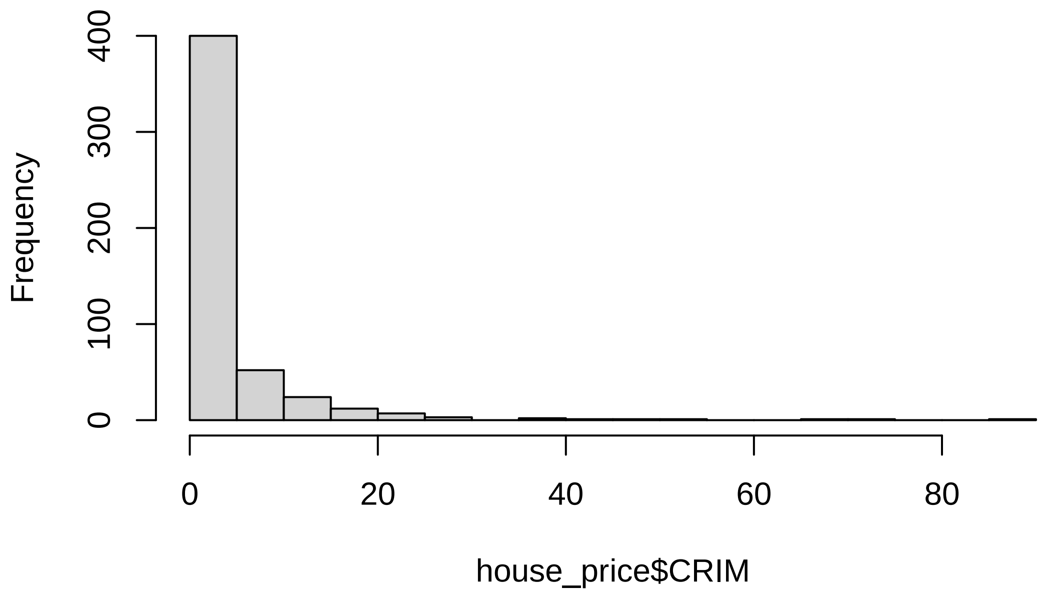


Figure 1c. Histogram of Rooms per Dwelling

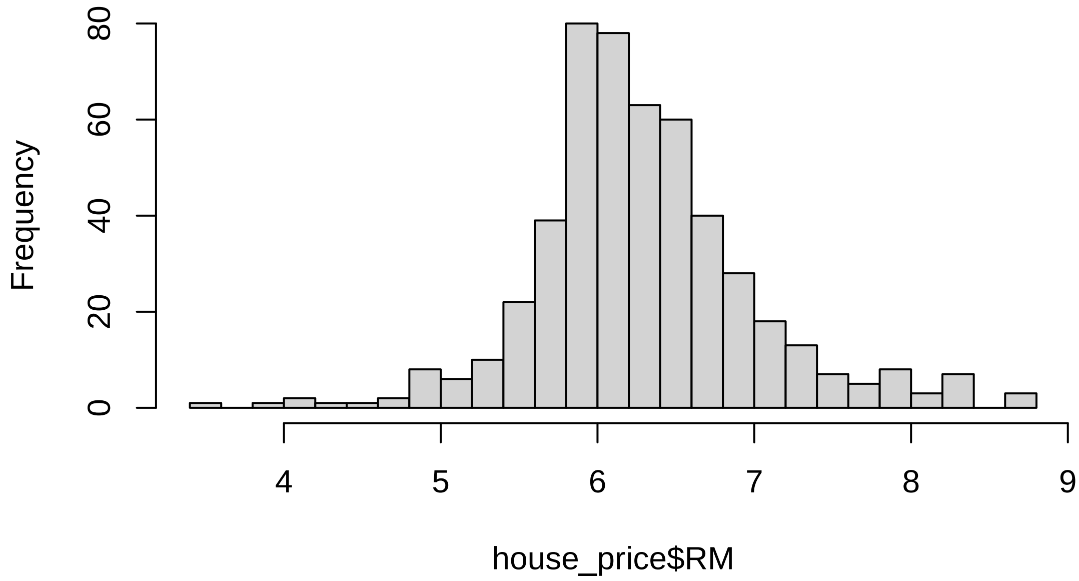


Figure 1d. Histogram of Distance to Employment Centers

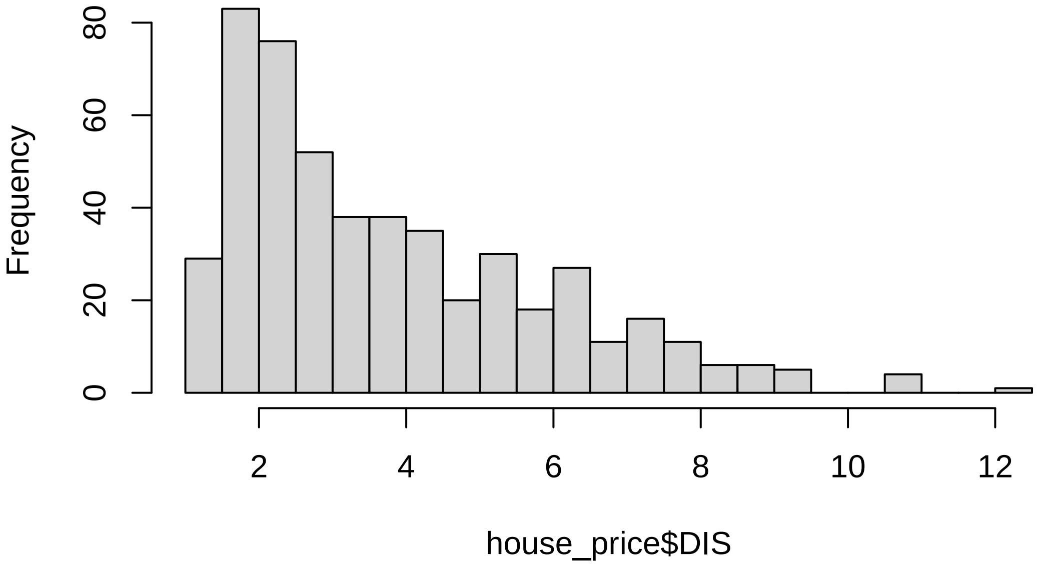
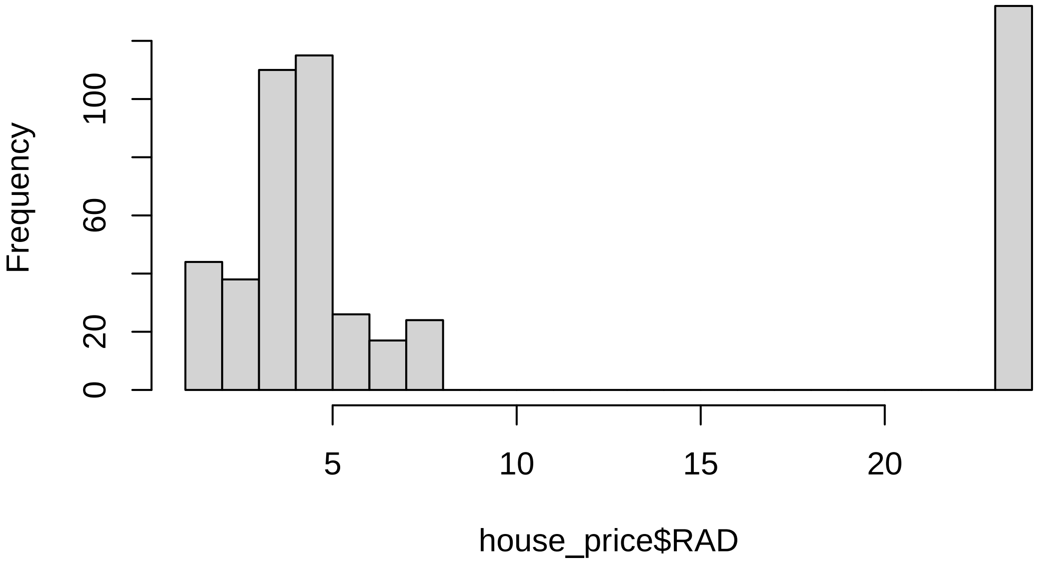


Figure 1e. Histogram of Accessibility to Highways



Figures 1b and 1d indicate that CRIM and DIS are skewed to the right, which requires log transformation. Figure 1e presents the bimodal distribution of RAD, so it will be converted to a categorical variable with two levels: “low” (when $RAD \leq 15$) and “high” (when $RAD > 15$). Figures 2a-c present the distribution of the three transformed variables.

Figure 2a. Histogram of Log(Per Capita Crime Rate)

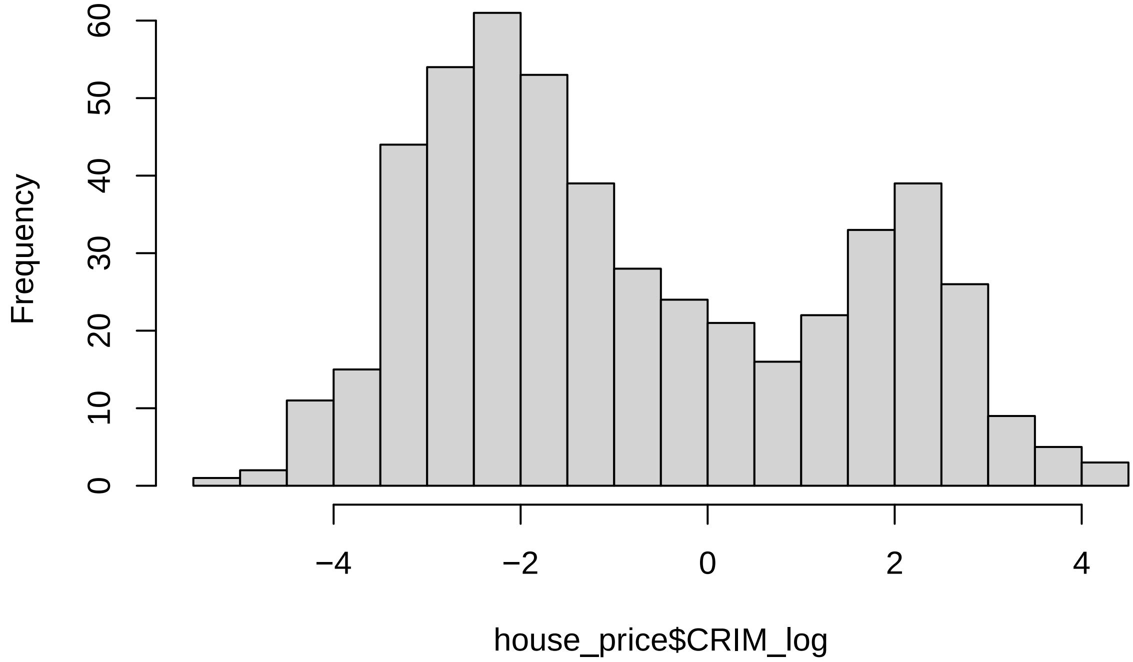


Figure 2b. Histogram of Log(Distance to Employment Centers)

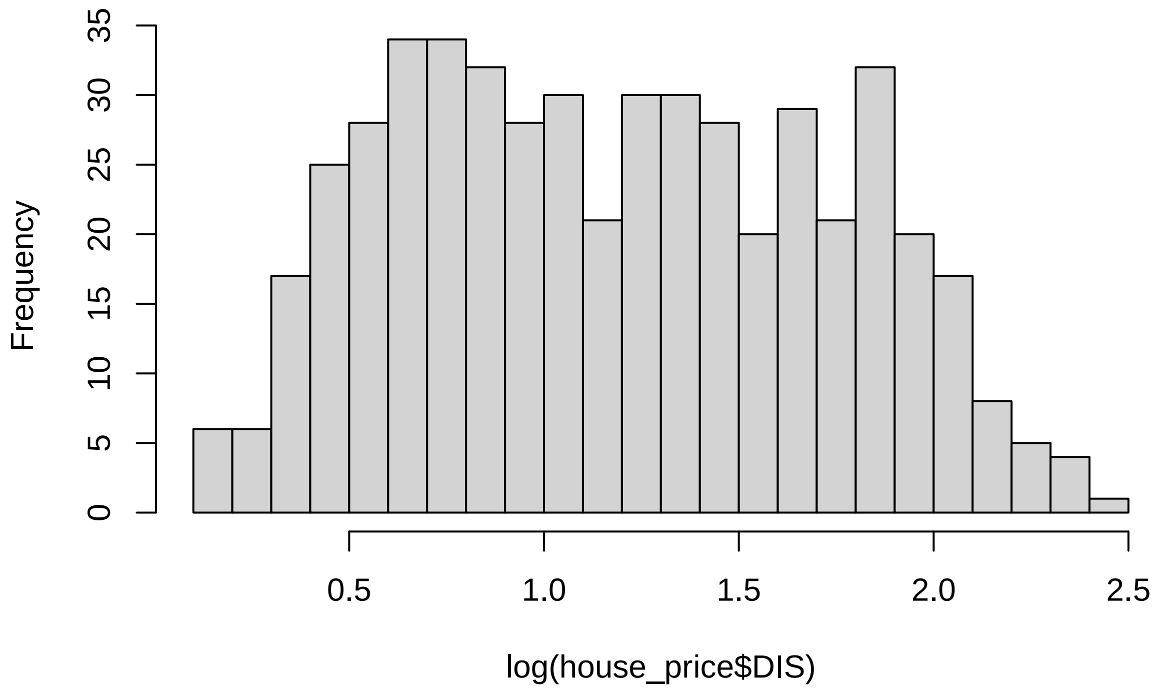
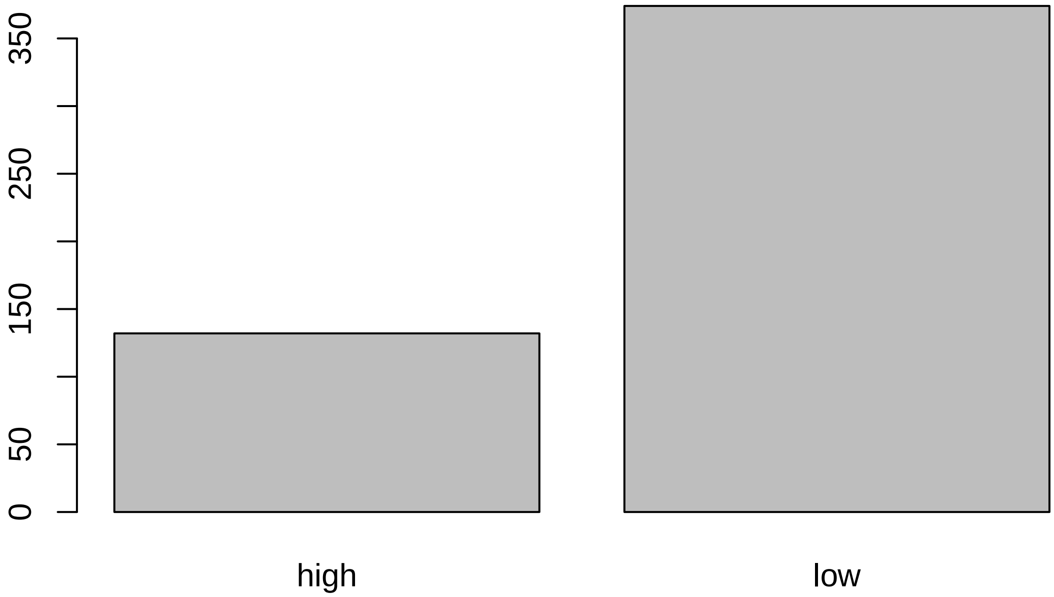


Figure 2c. Bar Chart of Categorical RAD



Since there are multiple predictors in the model, a heat map was drawn to check for multicollinearity among the outcome and continuous predictors. A red filling in the cell indicated a negative correlation, and a blue filling indicated a positive correlation. The P value for each correlation was also calculated. Correlations with p>.05 were crossed out in the matrix.

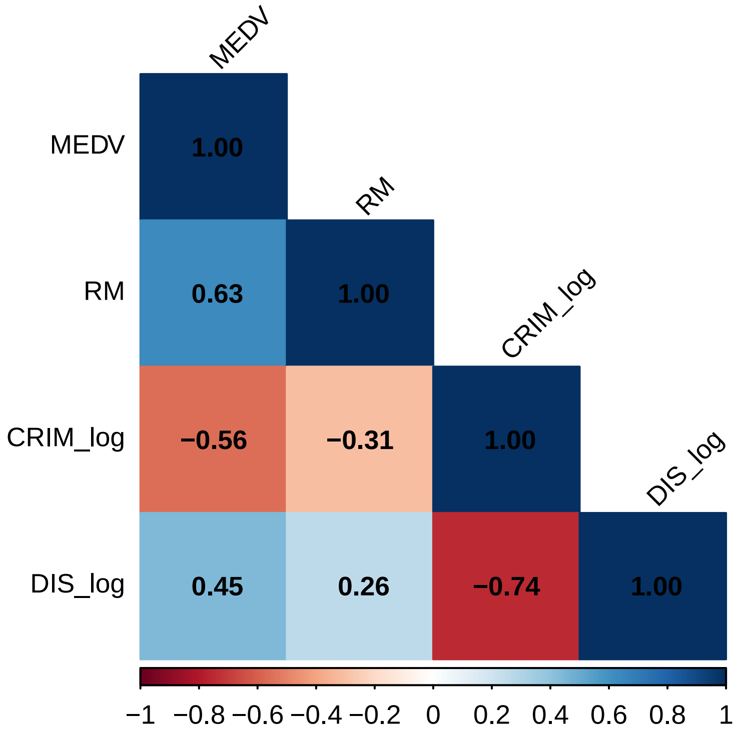


Figure 3: Heat Map of Predictors

As shown in Figure 3, CRIM_log and DIS_log are significantly correlated with a coefficient of -0.74. Thus, CRIM_log was removed from the model.

2.3 Multiple Linear Regression

In the multiple linear regression model, a model was fit to quantify the relationship between MEDV and RM, DIS_log, and RAD_cat.

```r
house_model <- lm(MEDV ~ RM + DIS_log + RAD_cat,
           data = house_price)
summary(house_model)
```

```
##
## Call:
## lm(formula = MEDV ~ RM + DIS_log + RAD_cat, data = house_price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.681  -3.055  -0.520   2.548  42.298
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.9432     2.4928 -13.616  < 2e-16 ***
## RM            8.3992     0.4096  20.507  < 2e-16 ***
## DIS_log      -0.3439     0.6266  -0.549    0.583
## RAD_catlow    5.5449     0.7625   7.272 1.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

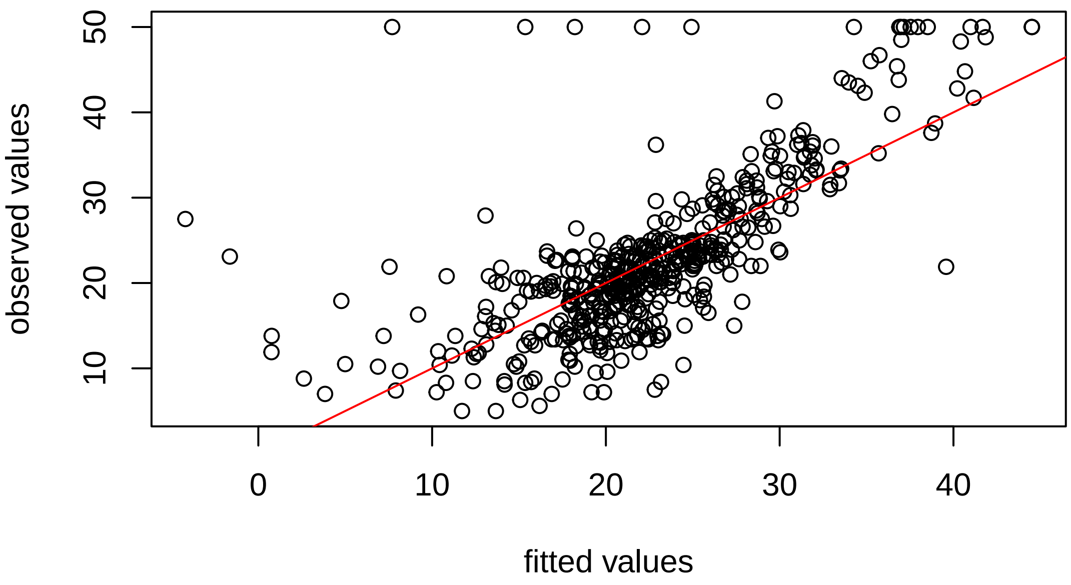
```
##
## Residual standard error: 6.22 on 502 degrees of freedom
## Multiple R-squared:  0.5453, Adjusted R-squared:  0.5426
## F-statistic: 200.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

Next, all assumptions of the linear regression model were checked, including linearity, normality of residuals, homoscedasticity, independence of observations, and multicollinearity, as detailed below.

2.3.1 Linearity Assessment

As shown in Figure 4, the majority of the observations follow a linear trend.

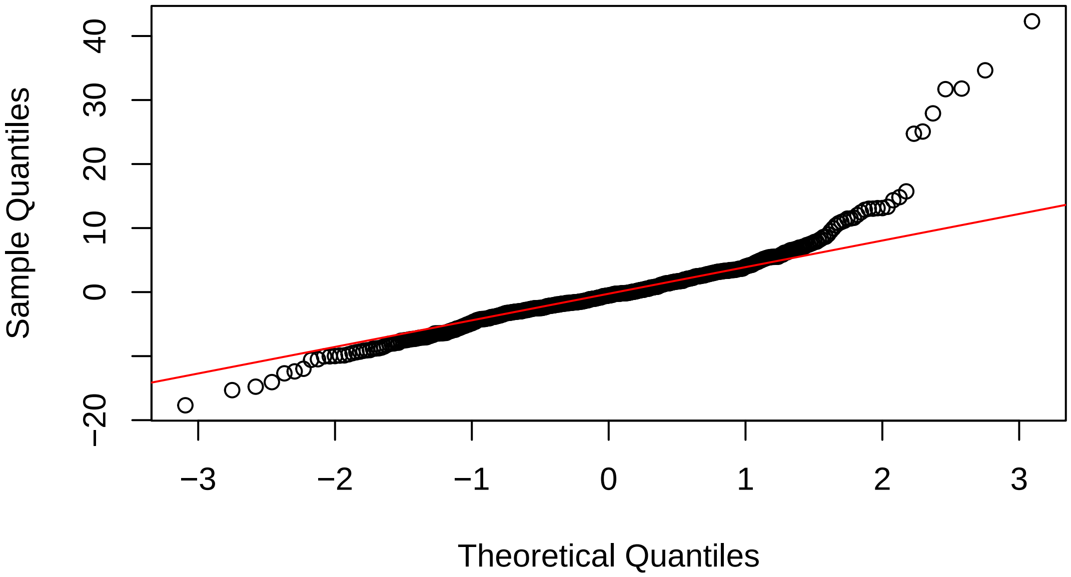
Figure 4. Linearity Check



2.3.2 Normality of Residuals

As shown in Figure 5, the residuals of most observations follow a normal distribution, although outliers are present at both ends.

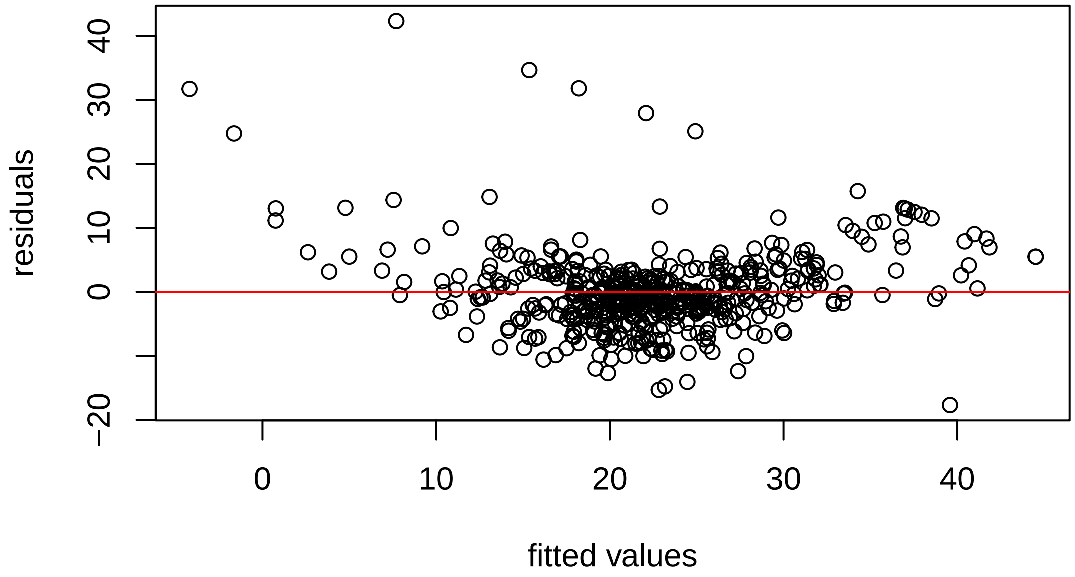
Figure 5. Normality of Residuals



2.3.3 Homoscedasticity Check

As shown in Figure 6, heteroscedasticity is present in the residuals, since they are not randomly scattered around the red line in the plot.

Figure 6. Homoscedasticity Check



2.3.4 Independence of Observations

Because each row represents a unique census tract (neighborhood) in the Boston area, there are no repeated observations from the same subject. Thus, each observation is independent of the others.

2.3.5 Multicollinearity Check

The VIFs of all variables are near 1, so multicollinearity is not present in the model.

```
vif(house_model)
```

```
##       RM  DIS_log  RAD_cat
## 1.080962 1.492040 1.466173
```

2.3.6 Assumption Violation Handling

According to 2.3.3, the homoscedasticity assumption is violated, so we used heteroscedasticity-consistent (HC) Standard Errors to address heteroscedasticity. This step strengthens the validity of p-values for hypothesis tests on coefficients.

```
sandwich1 <- coeftest(house_model, vcov = vcovHC(house_model, type = 'HC3'))
```

2.4 Step-wise Regression for Predictive Model

To investigate the second objective of the project, step-wise regression was employed to find the model that accurately predicts the median value of owner-occupied homes (MEDV) in Boston, MA. Both forward and backward selection were conducted. We subset the house_price dataset so that the raw forms of the three transformed variables will not be included in the model.

This is the summary of the model from forward selection.

```
##
## Call:
## lm(formula = MEDV ~ LSTAT + RM + PTRATIO + DIS_log + NOX + CHAS +
##     B, data = house_price[-c(1, 8, 9)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0323  -2.7068  -0.5117   1.9253  25.5948
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.922872   5.027419   7.742 5.50e-14 ***
## LSTAT        -0.569960   0.047014 -12.123  < 2e-16 ***
## RM            4.253108   0.394770  10.774  < 2e-16 ***
## PTRATIO      -0.982716   0.108549  -9.053  < 2e-16 ***
## DIS_log      -6.370602   0.725305  -8.783  < 2e-16 ***
## NOX         -24.519459   3.489253  -7.027 6.97e-12 ***
## CHAS1         3.076277   0.851425   3.613 0.000333 ***
## B             0.008931   0.002581   3.461 0.000585 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.749 on 498 degrees of freedom
## Multiple R-squared:  0.7371, Adjusted R-squared:  0.7334
## F-statistic: 199.4 on 7 and 498 DF,  p-value: < 2.2e-16
```

This is the summary of the model from backward selection.

```
##
## Call:
## lm(formula = MEDV ~ CHAS + NOX + RM + TAX + PTRATIO + B + LSTAT +
##     RAD_cat + DIS_log, data = house_price[-c(1, 8, 9)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8003  -2.7006  -0.4858   2.0582  24.7727
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.817935   5.760455   8.475 2.71e-16 ***
## CHAS1         2.905648   0.844026   3.443 0.000625 ***
## NOX         -24.241392   3.662909  -6.618 9.46e-11 ***
## RM            4.034952   0.396222  10.184  < 2e-16 ***
## TAX          -0.010752   0.003269  -3.289 0.001078 **
## PTRATIO      -1.043453   0.119667  -8.720  < 2e-16 ***
## B             0.009679   0.002637   3.671 0.000268 ***
## LSTAT        -0.575228   0.046696 -12.319  < 2e-16 ***
## RAD_catlow   -4.315342   1.187658  -3.633 0.000309 ***
## DIS_log      -6.511016   0.720253  -9.040  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.695 on 496 degrees of freedom
## Multiple R-squared:  0.7441, Adjusted R-squared:  0.7394
## F-statistic: 160.2 on 9 and 496 DF,  p-value: < 2.2e-16
```

2.5 Model Comparison with Metrics and Cross-Validation

To compare the two models from forward and backward selection, the AIC and BIC values of each model were calculated.

```
AIC(forward_model)
```

```
## [1] 3022.542
```

```
AIC(backward_model)
```

```
## [1] 3012.855
```

```
BIC(forward_model)
```

```
## [1] 3060.581
```

```
BIC(backward_model)
```

```
## [1] 3059.347
```

To assess the generalizability of the two models, 10-fold cross-validation was conducted. This is the cross-validation result for the forward-selection model.

```
## Linear Regression
##
## 506 samples
##   7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 455, 456, 454, 456, 455, 455, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.882039  0.7270138  3.475384
##
## Tuning parameter ’intercept’ was held constant at a value of TRUE
```

This is the cross-validation result for the backward-selection model.

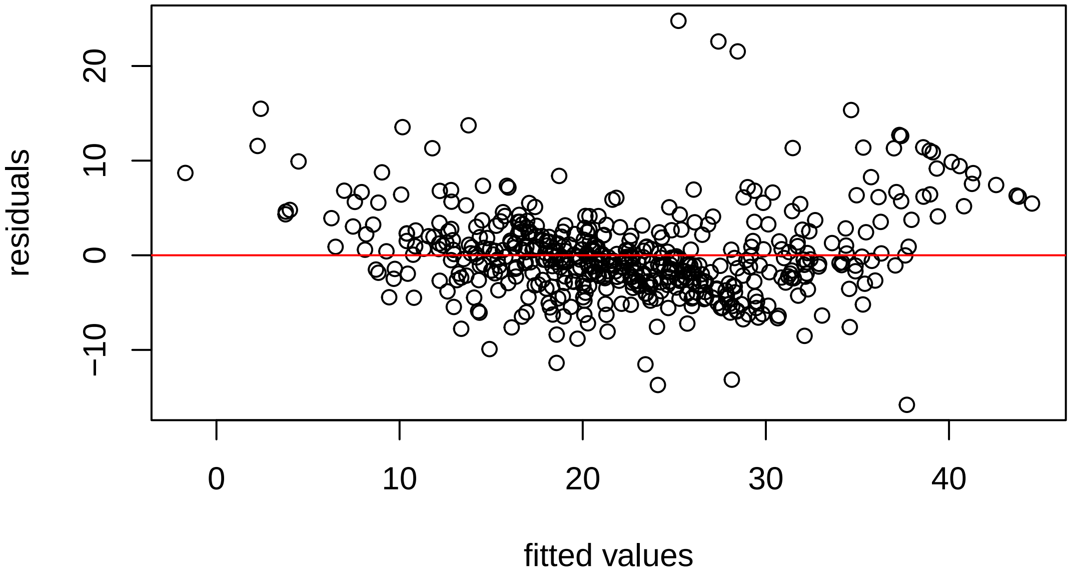
```
## Linear Regression
##
## 506 samples
##   9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 457, 455, 456, 455, 456, 456, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.745823  0.7407939  3.400113
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

2.6 Final Model and Coefficients

Based on the statistics above, the backward selection model was preferred, as it has a higher R-squared and adjusted R-squared as well as a lower AIC and BIC in the model summary. It also has a higher R-squared and lower RMSE in the cross-validation than the forward selection model.

To appropriately interpret the coefficients of the backward selection model, the homoscedasticity assumption was checked again.

Figure 7. Homoscedasticity Check for Final Model



As we can see from Figure 7, heteroscedasticity is present in the model, which prompted us to calculate the robust standard errors for the final model. This step strengthens the validity of p-values for hypothesis tests on coefficients.

```
sandwich2 <- coeftest(backward_model, vcov = vcovHC(backward_model, type = 'HC3'))
```

3. Results

3.1 Linear Regression Model

For the first objective, the summary of the model using robust standard errors is shown below.

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -33.94324    4.29995 -7.8939 1.855e-14 ***
## RM            8.39924    0.68418 12.2764 < 2.2e-16 ***
## DIS_log      -0.34386    0.59922 -0.5739    0.5663
## RAD_catlow    5.54493    0.78050  7.1043 4.167e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the model, MEDV (the median value of owner-occupied homes) is the outcome variable, and RM (average number of rooms per dwelling), DIS_log (log of weighted distances to five Boston employment centers), and RAD_cat (the categorical index of accessibility to radial highways) are the predictors.

3.2 Predictive Model for Boston House Prices

For the second objective, the summary of the final predictive model using robust standard errors is shown below.

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  48.8179345   9.6956248  5.0350 6.697e-07 ***
## CHAS1         2.9056484   1.2285482  2.3651 0.0184088 *
## NOX         -24.2413924   4.1899195 -5.7856 1.281e-08 ***
## RM            4.0349518   0.7827948  5.1545 3.678e-07 ***
## TAX          -0.0107518   0.0026806 -4.0110 6.978e-05 ***
## PTRATIO      -1.0434531   0.1137522 -9.1730 < 2.2e-16 ***
## B             0.0096793   0.0029097  3.3266 0.0009443 ***
## LSTAT        -0.5752277   0.0911076 -6.3137 6.055e-10 ***
## RAD_catlow   -4.3153415   1.1635475 -3.7088 0.0002318 ***
## DIS_log      -6.5110159   1.0414524 -6.2519 8.754e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the model, MEDV (the median value of owner-occupied homes) is the outcome variable, and the nine predictors included in the model are:

- CHAS (Charles River dummy variable, 1 if tract bounds river, 0 otherwise)

- NOX (Nitric Oxide concentration, parts per 10 million)
- RM (average number of rooms per dwelling)
- TAX (full-value property-tax rate per $10,000)
- PTRATIO (pupil-teacher ratio by town)
- B (1000(Bk - 0.63)^2, where Bk is the proportion of Black residents by town)
- LSTAT (percentage of lower status of the population).
- RAD_cat (categorical index of accessibility to radial highways)
- DIS_log (log of weighted distances to five Boston employment centers)

4. Discussion

4.1 Interpretation of Results

For the linear regression model generated for the first objective, we have the following coefficients and interpretations:

- Intercept: -33.9432 ($p < 0.001$)

  The expected median value of owner-occupied homes when RM = 0, DIS_log = 0, and RAD_cat = "high" is -$33,943.2, which serves as a baseline value.

- RM: 8.3992 ($p < 0.001$)

  Holding DIS_log and RAD_cat constant, a 1-unit increase in the average number of rooms per dwelling is associated with a $8,399.2 increase in the expected median value of owner-occupied homes. This association is statistically significant.

- DIS_log: -0.3439 ($p = 0.57$)

  The log of weighted distance to five Boston employment centers shows a non-significant effect on the expected median value of owner-occupied homes.

- RAD_cat: 5.5449 ($p < 0.001$)

  Holding RM and DIS_log constant, a low index of accessibility to radial highways ($RAD \leq 15$) is associated with a $5,544.9 increase in the expected median value of owner-occupied homes compared to a high index of accessibility to radial highways (RAD > 15). This association is statistically significant.

For the predictive model generated for the second objective, we have the following coefficients and interpretations:

- Intercept: 48.8179 ($p < 0.001$)

  The expected median value of owner-occupied homes (in $1,000) when all predictors = 0, which serves as a baseline value.

- CHAS: 2.9056 ($p = 0.02$)

  Holding everything else constant, being near the Charles River (CHAS = 1) is associated with a $2,905.6 increase in the expected median value of owner-occupied homes. This association is statistically significant.

- NOX: -24.2414 ($p < 0.001$)

  Holding everything else constant, a 1-unit increase in the Nitric Oxide concentration (parts per 10 million) is associated with a $24,241.4 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

- RM: 4.0350 ($p < 0.001$)

  Holding everything else constant, a 1-unit increase in the average number of rooms per dwelling is associated with a $4,035.0 increase in the expected median value of owner-occupied homes. This association is statistically significant.

- TAX: -0.0108 ($p < 0.001$)

  Holding everything else constant, a 1-unit increase in the full-value property-tax rate per $10,000 is associated with a $10.8 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

- PTRATIO: -1.0435 ($p < 0.001$)

  Holding everything else constant, a 1-unit increase in the pupil-teacher ratio is associated with a $1,043.5 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

- B: 0.0097 ($p < 0.001$)

  Holding everything else constant, a 1-unit increase in the parabolically transformed value of African American resident proportion (the formula is 1000(Bk - 0.63)^2, where Bk is the proportion of African American residents) is associated with a $9.7 increase in the expected median value of owner-occupied homes. This association is statistically significant.

- LSTAT: -0.5752 ($p < 0.001$)

  Holding everything else constant, a 1-unit increase in the percentage of lower status of the population is associated with a $575.2.0 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

- RAD_cat: -4.3153 ($p < 0.001$)

  Holding everything else constant, a low index of accessibility to radial highways ($RAD \leq 15$) is associated with a $4,315.3 decrease in the expected median value of owner-occupied homes compared to a high index of accessibility to radial highways ($RAD > 15$). This association is statistically significant.

- DIS_log: -6.5110 ($p < 0.001$)

  Holding everything else constant, a 1% increase in the weighted distance to five Boston employment centers is associated with a $65.11 decrease in the expected median value of owner-occupied homes. This association is statistically significant.

These two models provide informative insights into factors impacting the house prices in Boston. Besides the information provided by the coefficients and interpretations, I would like to note down additional observations and comments.

First, it's interesting to see how the coefficients of RAD_cat and DIS_log changed directions and significance in the two models. While DIS_log is non-significant and RAD_cat shows a significantly positive association with the median value of owner-occupied homes in the first model, both the former and the latter have a significantly negative association in the predictive model. This switch indicates that context matters in terms of interpreting the coefficients, which is why it's essential to keep other covariates in mind when making conclusions.

In addition, it's important to pause and reflect on the implications of including certain variables in predictive models. One variable in the second model caught my eye - B, defined by the authors as "1000(Bk - 0.63)^2, where Bk is the proportion of Black residents by town." The paper that the dataset was from used it to conduct a hedonic pricing model for houses in Boston, MA. Hedonic pricing is a "technique that involves modeling the prices of goods and services by features both internal and external to the good or service" (Carlisle 2020). In other words, it "postulates that this kind of modeling is possible, and that marginal changes in an identified feature will have a corresponding effect in the price of the underlying" (Carlisle 2020). By including the proportion of Black residents as a covariate, the original model captures the effects of racial demographics on home prices. While this may reveal patterns of racial discrimination in housing markets, its uncritical inclusion in predictive modeling risks perpetuating the very inequalities it reflects. If used in forecasting or policy tools without a critical lens, such variables could reinforce systemic racism by implicitly endorsing the historical devaluation of neighborhoods with more Black residents.

4.2 Limitations

We have several limitations. First, step-wise regression was employed to find the predictive model, but we recognize that a branch and bound procedure can sometimes outperform the step-wise regression as it "prunes the search tree" by eliminating entire groups of suboptimal models at a time without computing individual RSS values, which makes it much more efficient than exhaustively fitting all 2p models, as would be the case with forward and backward selection. Also, branch and bound can find the best model according to the chosen criterion and not just the locally good one - it ensures optimality within the scope/models you are searching. We picked step-wise regression because there are a large number of variables in our house price dataset, making branch and bound too computationally intensive. Forward or backward selection is fast and scalable since it's a step-by-step model in which the best local option is chosen at each step, so it's useful for our project settings.

Second, this dataset was published in 1978, which is almost half a decade ago. Thus, the data can be outdated and may not be as generalizable to modern days. Also, as discussed before, the uncritical inclusion in predictive modeling of variables like B can sustain systemic racism by endorsing the historical discriminatory practices in housing values.

5. Conclusion

This project does two things. First, it investigated the relationship between the median value of owner-occupied homes (MEDV) and the average number of rooms per dwelling (RM), the weighted distance to five Boston employment centers (DIS), and the index of accessibility to radial highways (RAD). Then, the project developed a predictive model that estimates the median value of owner-occupied homes (MEDV) in Boston, MA, using step-wise regression.

According to the first model, the median value of homes is positively associated with the number of rooms and negatively associated with accessibility. In the more comprehensive predictive model, the median value of homes is higher near the Charles River, with more rooms, and slightly with a higher parabolically transformed value of African American resident proportion. Conversely, home values decline with increased NOx pollution, higher property taxes, greater pupil-teacher ratios, a higher percentage of lower-status residents, and greater (log-transformed) distances from employment centers. While these findings offer insights into Boston's housing dynamics for policymakers and real estate professionals, it's critical that such models are applied thoughtfully to avoid reinforcing systemic inequities.

6. References

Arun Kumar. n.d. "Boston House Price Data." https://www.kaggle.com/datasets/arunjathari/bostonhousepricedata.

Carlisle, M. 2020. "Racist Data Destruction?" https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8.

Harrison, David, and Daniel L Rubinfeld. 1978. "Hedonic Housing Prices and the Demand for Clean Air." *Journal of Environmental Economics and Management* 5 (1): 81–102. https://doi.org/10.1016/0095-0696(78)90006-2.

Levkovich, Or, Jan Rouwendal, and Ramona van Marwijk. 2016. "The Effects of Highway Development on Housing Prices." *Transportation* 43 (2): 379–405. https://doi.org/10.1007/s11116-015-9580-7.

"Measuring the Impact of Crime on House Prices: Applied Economics: Vol 33, No 15." n.d. https://www.tandfonline.com/doi/abs/10.1080/00036840110021735?casa_token=CPQLRe1LLwEAAAAA:JvZ6h2bW2BxPXT1Hjj19vBtnhZtTyS_SqdXbRkyoxbhA8aWyVRyRT4xMEz0YZvBLurQRKfLdsogG.