# Statistical Analysis of Boston Real Estate

Muhan Zhang

2025-08-14

## A. Select a Public Dataset

For this case study, I have chosen the Boston House Price Dataset, first introduced by Harrison and Rubinfeld (1978) in their research exploring the relationship between air quality and housing prices. The dataset is publicly available through StatLib at Carnegie Mellon University and can also be accessed via Kaggle (Soriano, 2021).

### 1. Dataset link:

https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data

### 2. Background and source:

This dataset contains information on 506 neighborhoods in Boston and the surrounding area, including median house prices and 13 associated features. In the original research, Harrison and Rubinfeld applied the Hedonic Pricing Model to analyze the impact of air quality, neighborhood characteristics, and accessibility factors on housing prices.

### 3.BVariable Information:

#### 3.1 Input Features (13)

1. CRIM: Per capita crime rate by town
2. ZN: Proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: Proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (1 = tract bounds river; 0 = otherwise)
5. NOX: Nitric oxides concentration (parts per 10 million)
6. RM: Average number of rooms per dwelling
7. AGE: Proportion of owner-occupied units built prior to 1940
8. DIS: Weighted distances to five Boston employment centers
9. RAD: Index of accessibility to radial highways
10. TAX: Full-value property tax rate per $10,000

11. PTRATIO: Pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2, where Bk is the proportion of Black residents by town
13. LSTAT: Percentage of lower-status population

### 3.2 Output Variable (1)

1. MEDV: Median value of owner-occupied homes in $1000s

# B. Library the packages

```
knitr::opts_chunk$set(echo = TRUE)
library(readxl)       # For reading Excel files
library(car)          # For Levene’s Test (homogeneity of variance)
library(psych)        # For describeBy (group summaries)
library(phia)         # For testInteractions (simple effects)
library(RVAideMemoire)# For byf.hist and byf.shapiro (normality tests)
library(here)         # Manage project file paths across systems
library(tidyverse)    # Data wrangling, transformation, and visualization
library(corrplot)     # Visualization of correlation matrices
```

# C. Data Preparation and Exploration

```
df <- read.csv(here("data", "boston.csv"))
str(df)
```

```
## ’data.frame':    506 obs. of  14 variables:
##  $ CRIM   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ ZN     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ INDUS  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ CHAS   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ NOX    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ RM     : num  6.58 6.42 7.18 7 7.15 ...
##  $ AGE    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ DIS    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ RAD    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ TAX    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ PTRATIO: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ B      : num  397 397 393 395 397 ...
##  $ LSTAT  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ MEDV   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

# D. Statistical Analysis

## 1. Chi-Square Test of Independence

### 1.1 Research Question

Is there an association between proximity to the Charles River (CHAS) and housing age categories (old vs. new buildings) in the Boston area?

### 1.2 Literature Review

Historical urban development patterns suggest that waterfront areas often represent some of the oldest settled regions in cities (Hoyt, 1939; Burgess, 1925). In Boston specifically, neighborhoods near the Charles River include some of the city's earliest developments, such as Back Bay and Beacon Hill, which were established in the 18th and 19th centuries (O'Connell, 2013).

Waterfront locations have historically been preferred for early settlement due to access to transportation, trade, and water resources (Mumford, 1961). However, urban renewal projects in the mid-20th century may have altered the original relationship between proximity to water bodies and housing age, as older structures were sometimes demolished and replaced with modern developments (Jacobs, 1961).

Boston's Charles River area underwent significant transformation during the 20th century, including the creation of the Charles River Esplanade and various urban planning initiatives that may have influenced the housing stock composition (Krieger & Cobb, 1999).

### 1.3 Hypotheses

Null Hypothesis (H0): There is no association between Charles River proximity and housing age categories.

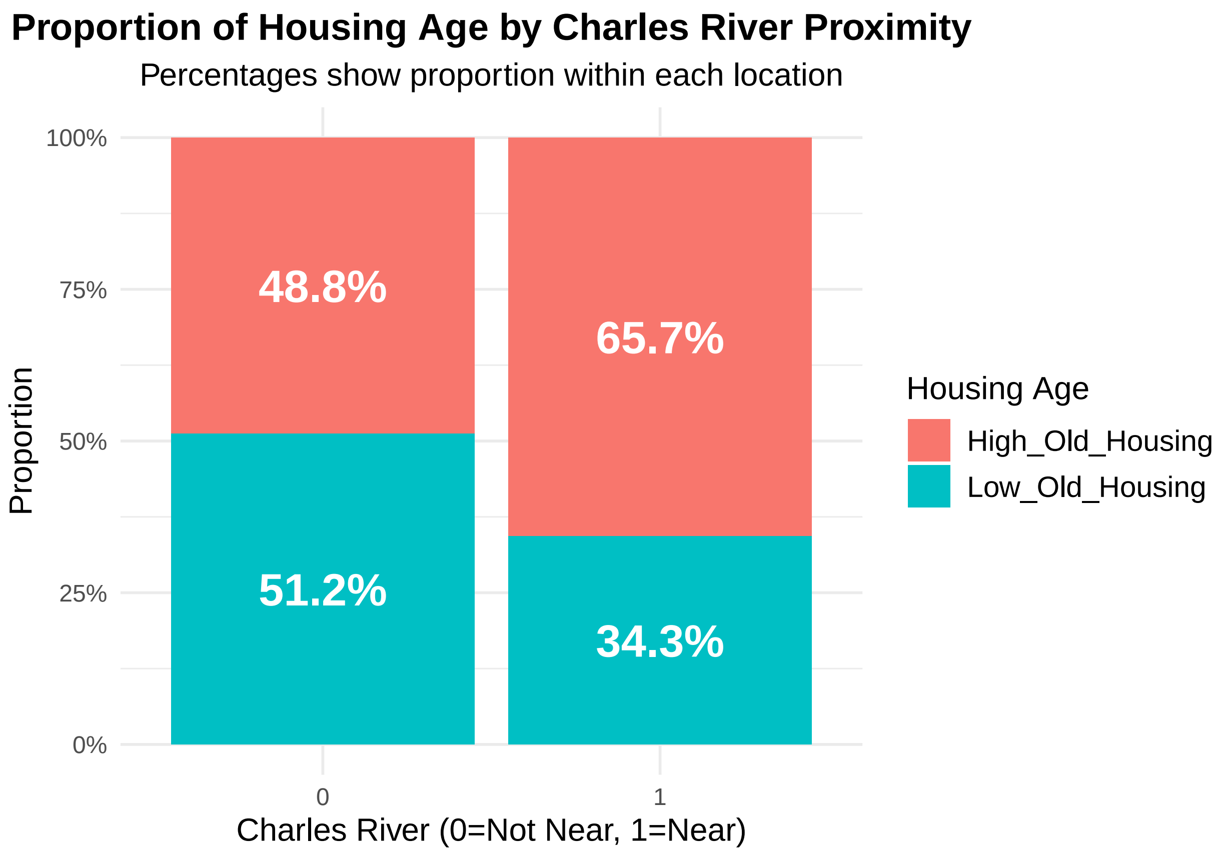
Alternative Hypothesis (H1): There is an association between Charles River proximity and housing age categories.

Directional Prediction: Based on Boston's historical development patterns, we hypothesize that areas near the Charles River will demonstrate a significant association with higher proportions of pre-1940 housing construction, as waterfront areas typically represent older, established neighborhoods.

### 1.4 Data Visualization

```r
df$age_cat <- ifelse(df$AGE <= median(df$AGE),
                     "Low_Old_Housing",
                     "High_Old_Housing")
```

```r
ggplot(df, aes(x = factor(CHAS), fill = age_cat)) +
  geom_bar(position = "fill") +
  geom_text(stat = "count",
            aes(label = paste0(round(after_stat(count)/tapply(after_stat(count),
                                   after_stat(x), sum)[after_stat(x)]*100, 1), "%")),
            position = position_fill(vjust = 0.5),
            color = "white", size = 6, fontface = "bold") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Proportion of Housing Age by Charles River Proximity",
       subtitle = "Percentages show proportion within each location",
       x = "Charles River (0=Not Near, 1=Near)",
       y = "Proportion",
       fill = "Housing Age") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 12),
    axis.title = element_text(size = 12),
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 11)
  )
```



The proportional bar chart reveals a notable pattern consistent with our hypothesis, where non-riverfront areas display a relatively balanced distribution with slightly more areas having low old housing concentrations (51.2% low old housing vs. 48.8% high old housing), while riverfront areas show a pronounced concentration of areas with high old housing, with nearly two-thirds of riverfront neighborhoods having high proportions of pre-1940 construction (65.7% high old housing vs. 34.3% low old housing). The data reveals a substantial 16.9 percentage point difference in high old housing concentration between riverfront and non-riverfront areas (65.7% - 48.8% = 16.9%), indicating that Charles River proximity is associated with a markedly higher likelihood of neighborhoods containing substantial pre-1940 housing stock, strongly supporting our theoretical expectations based on historical urban development patterns.

### 1.5 Exploratory Analysis

```r
# Create age categories based on proportion of old housing (median split)
# AGE = proportion of owner-occupied units built prior to 1940
df$age_cat <- ifelse(df$AGE <= median(df$AGE), "Low_Old_Housing", "High_Old_Housing")
```

```r
# Create contingency table
table_chi <- table(df$CHAS, df$age_cat)
table_chi
```

```
##
##     High_Old_Housing Low_Old_Housing
##   0              230             241
##   1               23              12
```

```r
# Perform chi-square test
chisq.test(table_chi)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_chi
## X-squared = 3.0695, df = 1, p-value = 0.07978
```

### 1.6 APA Write-up

A chi-square test of independence was conducted to examine the association between proximity to the Charles River and housing age categories. The association was not statistically significant, $X^2(1, \text{N} = 506) = 3.07$, $p = .080$, indicating no significant relationship between Charles River proximity and the proportion of old housing in Boston neighborhoods.

### 1.7 Summary of Main Findings

The analysis revealed a substantial visual pattern where riverfront areas showed 65.7% high old housing concentration compared to 48.8% in non-riverfront areas—a difference of 16.9 percentage points that aligns with established urban development theories. However, the chi-square test indicated a non-significant association ($p = .080$), falling just short of conventional significance levels, suggesting that while the observed pattern supports theoretical expectations about waterfront development history, it does not meet statistical criteria for confirming the association.

### 1.8 Limitations of the Analysis

Several limitations may have influenced the results including the substantial sample size imbalance between riverfront ($n = 35$) and non-riverfront ($n = 471$) areas which may have limited statistical power, the binary categorization approach for housing age that may oversimplify complex construction period distributions, and the inability to account for historical factors such as urban renewal projects or differential demolition rates that may have altered original housing age patterns along the Charles River corridor.

### 1.9 Suggestions for Future Research

Future investigations should focus on increasing the sample size of riverfront properties, employing continuous measures of distance from the Charles River rather than binary proximity coding, and utilizing more granular housing age data to capture the full complexity of Boston's development history. Additionally, spatial analysis using GIS techniques could examine distance-decay effects from the riverfront, while longitudinal analysis could track changes in housing stock composition over time, ultimately providing stronger evidence for the theoretical relationship between waterfront proximity and historical housing development patterns.

## 2. Two-Way Between-Group ANOVA

### 2.1 Research Question

Do crime rate levels (low vs. medium vs. high) and proximity to the Charles River (CHAS) jointly affect housing prices (MEDV) in the Boston area?

### 2.2 Literature Review

Crime rates have consistently been identified as significant negative predictors of housing values across urban areas (Cohen et al., 2016; Ihlanfeldt & Mayock, 2010), while transportation accessibility, particularly highway access, generally increases property values by reducing commuting costs and improving connectivity (Cervero & Duncan, 2002; Ryan, 1999).

However, the interaction between these factors is less understood, with some research suggesting that transportation benefits may be diminished in high-crime areas due to safety concerns that override convenience benefits (Ellen et al., 2001).

### 2.3 Hypotheses

Null Hypothesis (H0): There are no main effects of crime level or highway accessibility on housing prices, and there is no interaction between crime level and highway accessibility.

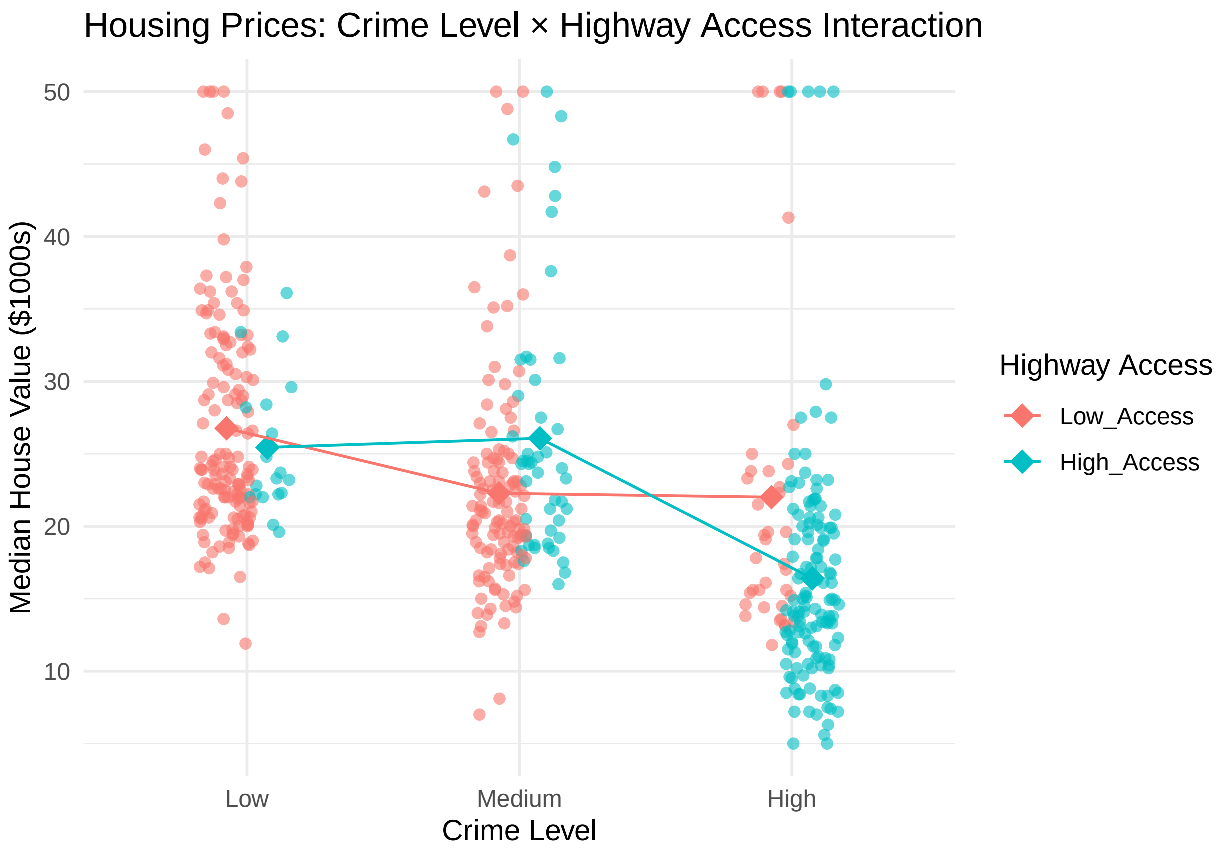
Alternative Hypothesis (H1): There are significant main effects and/or interaction effects between crime level and highway accessibility on housing prices.

Directional Predictions: Based on established research, we hypothesize significant main effects where crime level will negatively predict housing prices (low > medium > high crime areas) and highway accessibility will positively predict housing prices (high access > low access). Additionally, we predict a significant interaction where the positive effect of highway accessibility will be reduced or eliminated in high-crime neighborhoods, as safety concerns may override transportation convenience benefits.

### 2.4 Data Visualization

```r
crim_tertiles <- quantile(df$CRIM, probs = c(1/3, 2/3))
df$CRIM_level <- cut(df$CRIM,
                     breaks = c(-Inf, crim_tertiles[1], crim_tertiles[2], Inf),
                     labels = c("Low", "Medium", "High"))
df$CRIM_level <- as.factor(df$CRIM_level)
df$RAD_level <- factor(ifelse(df$RAD > median(df$RAD), 1, 0),
                       levels = c(0, 1),
                       labels = c("Low_Access", "High_Access"))
```

```
ggplot(df, aes(x = CRIM_level, y = MEDV, color = RAD_level)) +
  geom_point(position = position_jitterdodge(dodge.width = 0.3), alpha = 0.6) +
  stat_summary(fun = mean, geom = "point", size = 4, shape = 18,
               position = position_dodge(width = 0.3)) +
  stat_summary(fun = mean, geom = "line", aes(group = RAD_level),
               position = position_dodge(width = 0.3)) +
  labs(title = "Housing Prices: Crime Level × Highway Access Interaction",
       x = "Crime Level", y = "Median House Value ($1000s)",
       color = "Highway Access") +
  theme_minimal()
```



The scatter plot with mean trend lines clearly demonstrates the predicted interaction between crime level and highway accessibility on housing prices. Both low and high highway accessibility areas show declining mean housing prices as crime levels increase, but the patterns differ substantially. Low accessibility areas (red diamonds) show a gradual decline from approximately $26,800 in low crime areas to $22,300 in medium crime areas and $22,000 in high crime areas, representing a modest downward trend. In contrast, high accessibility areas (blue diamonds) maintain similar prices in low crime ($25,400) and medium crime areas ($26,100), but experience a dramatic drop to $16,400 in high crime areas, creating a steep non-parallel pattern. This visual evidence strongly supports our hypothesis that highway accessibility loses its protective effect on housing values in high-crime neighborhoods, where safety concerns override transportation convenience benefits.

### 2.5 Exploratory Analysis

```r
# Create 3-level crime rate groups using tertiles (33rd and 67th percentiles)
crim_tertiles <- quantile(df$CRIM, probs = c(1/3, 2/3))
df$CRIM_level <- cut(df$CRIM,
                     breaks = c(-Inf, crim_tertiles[1], crim_tertiles[2], Inf),
                     labels = c("Low", "Medium", "High"))
df$CRIM_level <- as.factor(df$CRIM_level)
```

```r
# Create highway accessibility factor with meaningful labels
df$RAD_level <- factor(ifelse(df$RAD > median(df$RAD), 1, 0),
                       levels = c(0, 1),
                       labels = c("Low_Access", "High_Access"))
```

```r
# Perform 2*3 Factorial ANOVA
anova1 <- aov(MEDV ~ CRIM_level * RAD_level, data = df)
summary(anova1)
```

```
##                       Df Sum Sq Mean Sq F value  Pr(>F)    
## CRIM_level             2   6990    3495  50.956 < 2e-16 ***
## RAD_level              1     36      36   0.532   0.466    
## CRIM_level:RAD_level   2   1395     697  10.168 4.7e-05 ***
## Residuals            500  34295      69                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

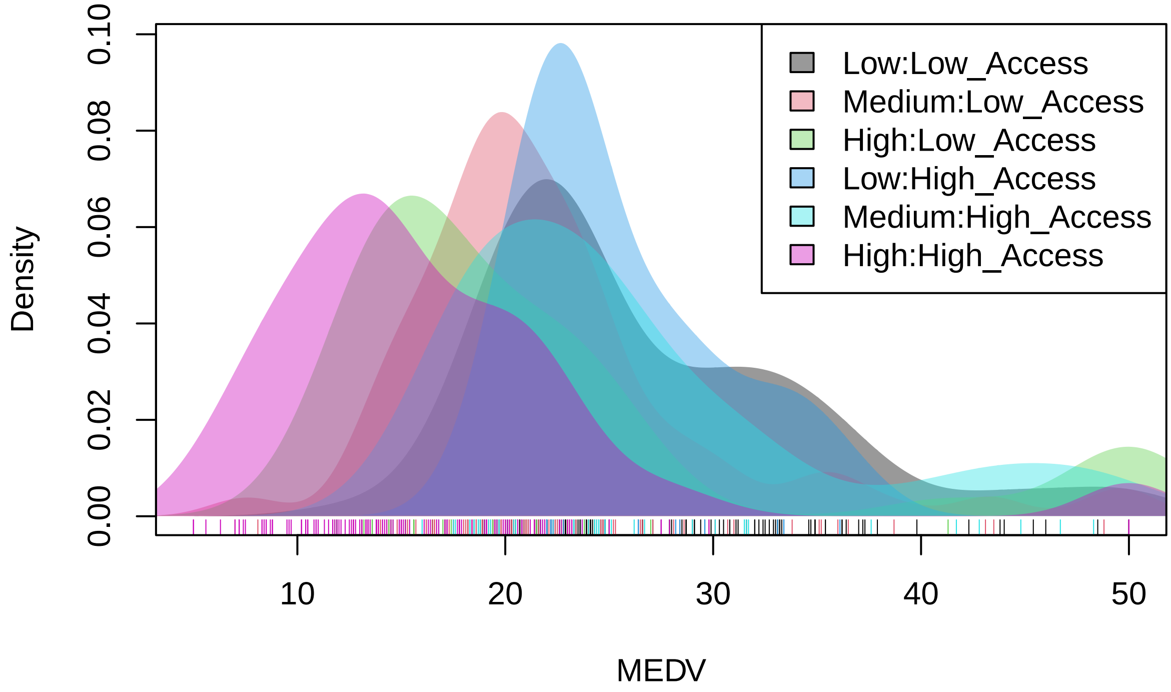
```r
# Check assumptions
# Homogeneity of variance
leveneTest(MEDV ~ CRIM_level * RAD_level, data = df)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   5  1.2161 0.3002
##       500
```

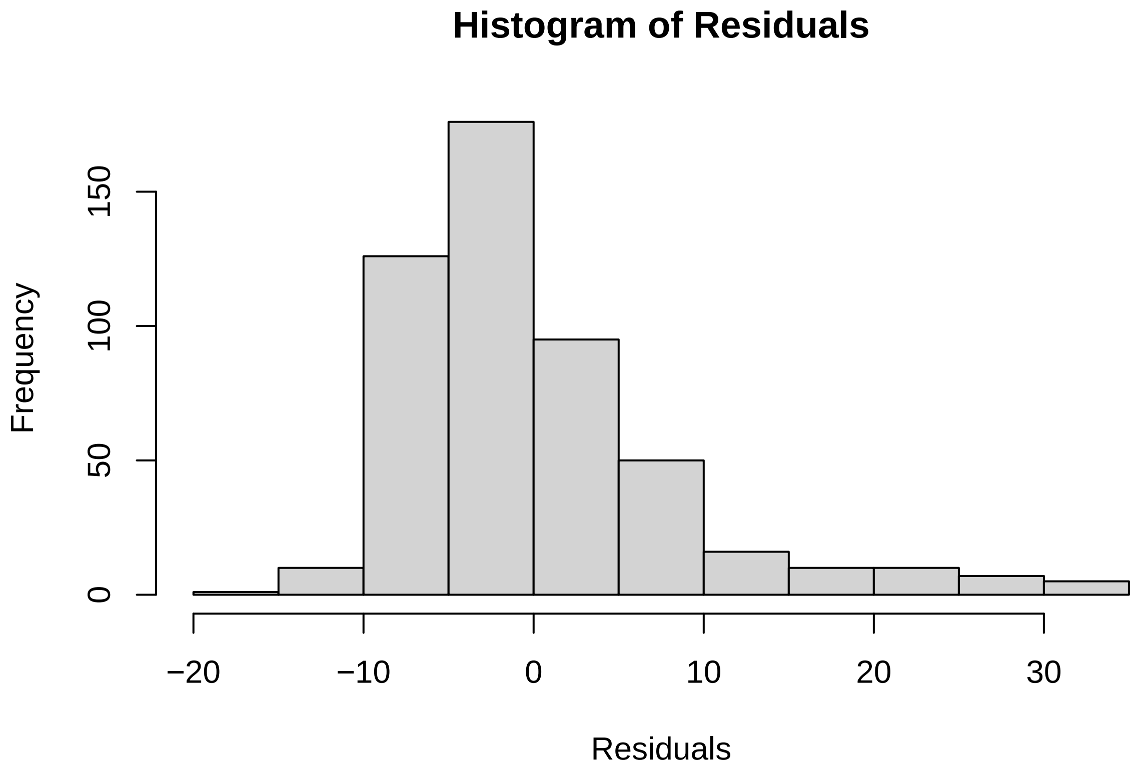
```r
# Normality of data within each cell
byf.shapiro(MEDV ~ CRIM_level * RAD_level, data = df)
```

```
## 
##  Shapiro-Wilk normality tests
## 
## data:  MEDV by CRIM_level:RAD_level 
## 
##                         W   p-value    
## Low:Low_Access     0.8970 8.960e-09 ***
## Medium:Low_Access  0.8595 2.690e-09 ***
## High:Low_Access    0.7190 4.382e-07 ***
## Low:High_Access    0.8803   0.02171 *  
## Medium:High_Access 0.8347 8.763e-06 ***
## High:High_Access   0.7857 1.311e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
byf.hist(MEDV ~ CRIM_level * RAD_level, data = df)
```



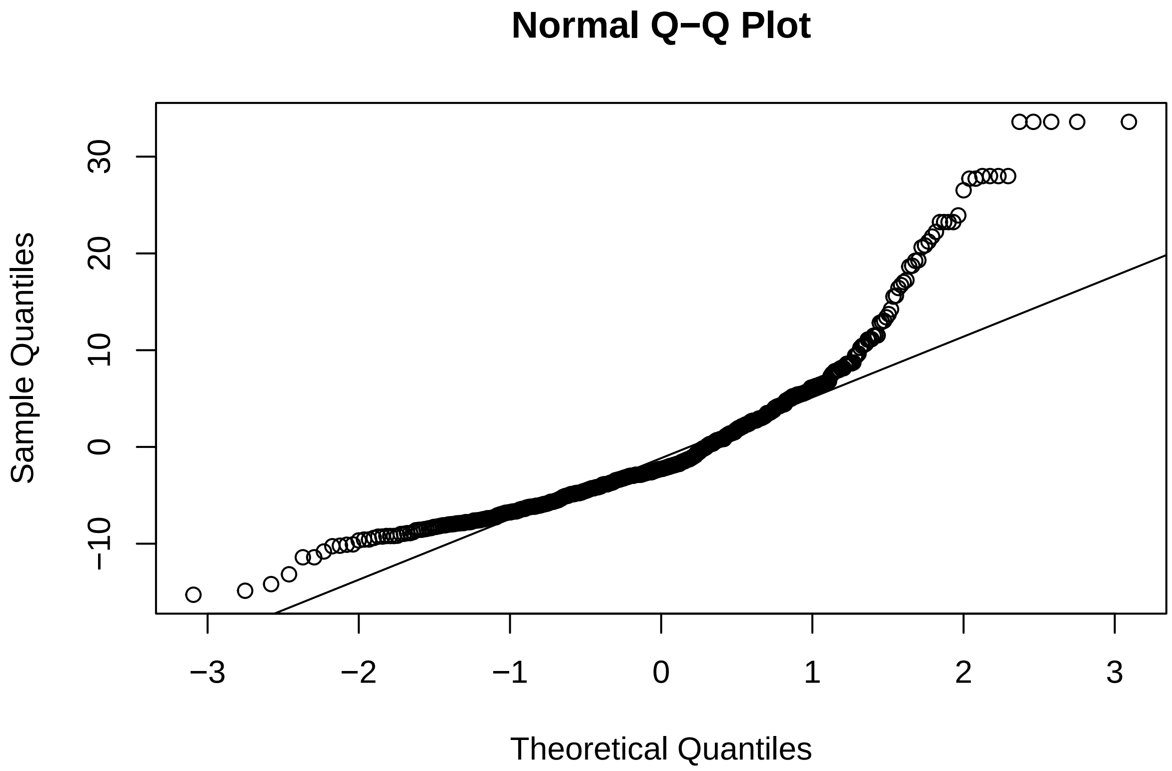
```
# Normality of residuals
res <- anova1$residuals
hist(res, main = "Histogram of Residuals", xlab = "Residuals")
```



```r
# formal shapiro-wilk test
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.85643, p-value < 2.2e-16
```

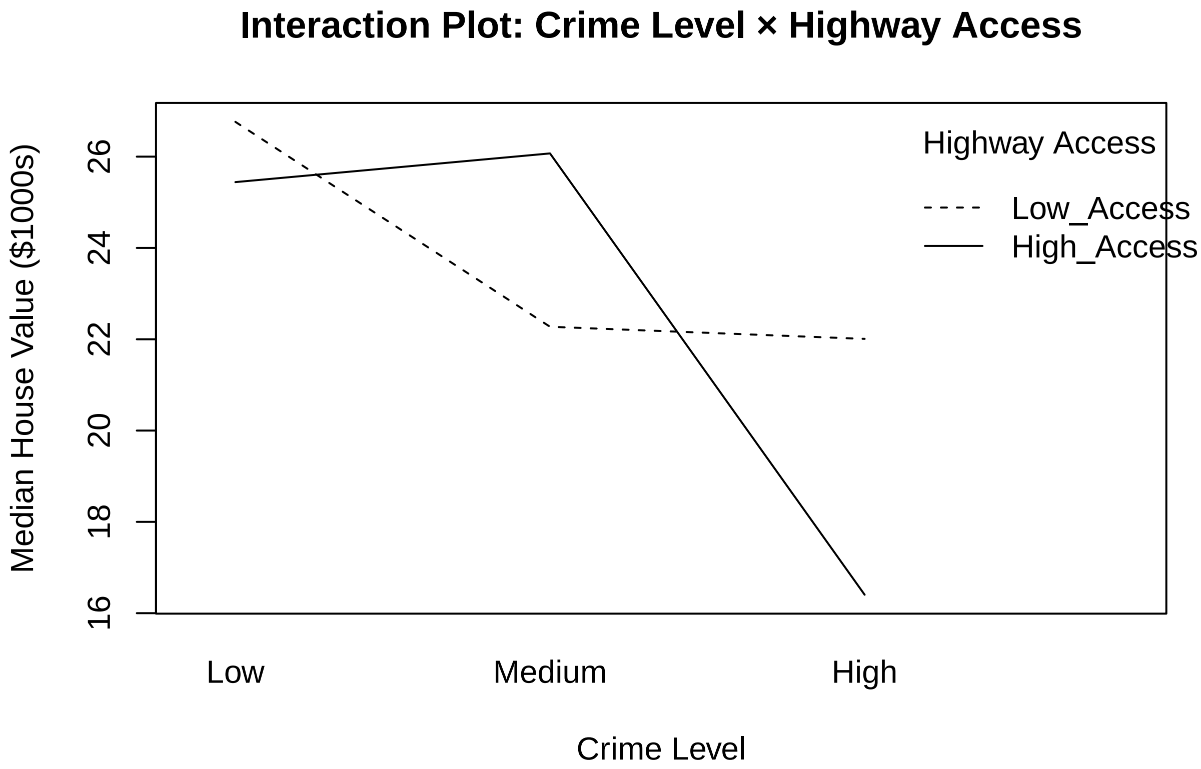
```r
qqnorm(res)
qqline(res)
```



```
#Simple effects analysis + Interaction plot
#(only if interaction is significant)
testInteractions(anova1, fixed = "RAD_level", across = "CRIM_level", adjustment = "none")
```

```
## F Test:
## P-value adjustment method: none
##             CRIM_level1 CRIM_level2    SE1    SE2  Df Sum of Sq      F
##  Low_Access      4.7539      0.2661 1.5202 1.5574   2      1590 11.594
## High_Access      9.0383      9.6670 2.0322 1.3959   2      4007 29.208
## Residuals                                         500     34295
##                Pr(>F)
##  Low_Access 1.197e-05 ***
## High_Access 1.007e-12 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
interaction.plot(df$CRIM_level, df$RAD_level, df$MEDV,
                 xlab = "Crime Level",
                 ylab = "Median House Value ($1000s)",
                 trace.label = "Highway Access",
                 main = "Interaction Plot: Crime Level × Highway Access")
```



```r
describeBy(df$MEDV, list(df$CRIM_level, df$RAD_level))
```

```
##
##  Descriptive statistics by group
## : Low
## : Low_Access
##    vars   n  mean   sd median trimmed  mad  min max range skew kurtosis   se
## X1    1 150 26.76 7.88   23.9   25.77 6.08 11.9  50  38.1 1.11     0.88 0.64
## ------------------------------------------------------------
## : Medium
## : Low_Access
##    vars   n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 120 22.27 7.55   20.4   21.21 4.45   7  50    43  1.6     3.38 0.69
## ------------------------------------------------------------
## : High
## : Low_Access
##    vars  n  mean    sd median trimmed  mad  min max range skew kurtosis   se
## X1    1 37 22.01 11.31   17.8    20.2 6.23 11.8  50  38.2 1.62     1.39 1.86
## ------------------------------------------------------------
## : Low
## : High_Access
##    vars  n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 19 25.44 4.76   23.3   25.16 2.22 19.6 36.1  16.5 0.83    -0.59 1.09
## ------------------------------------------------------------
## : Medium
## : High_Access
##    vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 48 26.07 8.88  24.15   24.84 7.26  16  50    34 1.28     0.68 1.28
## ------------------------------------------------------------
## : High
## : High_Access
##    vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 132 16.4 8.54   14.4   15.24 6.08   5  50    45 2.22     6.46 0.74
```

### 2.6 APA Write-up

A 2 x 3 factorial ANOVA revealed a significant interaction between the effects of crime level and highway accessibility on median housing prices in the Boston area, $F(2, 500) = 10.17$, $p < .001$. Among the neighborhoods with low highway accessibility, median housing prices decreased systematically as crime level increased: low crime areas ($M = 26.76$, $SD = 7.88$), medium crime areas ($M = 22.27$, $SD = 7.55$), and high crime areas ($M = 22.01$, $SD = 11.31$), $F = 11.59$, $p < .001$. Among the neighborhoods with high highway accessibility, the pattern was similar but more pronounced: low crime areas ($M = 25.44$, $SD = 4.76$), medium crime areas ($M = 26.07$, $SD = 8.88$), and high crime areas ($M = 16.40$, $SD = 8.54$), $F = 29.21$, $p < .001$. The interaction indicates that while highway accessibility provides some protection against crime's negative effect on housing prices in low and medium crime areas, this protective effect disappears in high crime neighborhoods, where housing prices drop dramatically regardless of transportation convenience.

### 2.7 Summary of Main Findings

The analysis revealed a significant interaction between crime level and highway accessibility on housing prices, with crime level showing a strong main effect (F = 50.96, p < .001) while highway accessibility alone showed no significant main effect (F = 0.53, p = .466). The interaction pattern demonstrates that highway accessibility provides some protection against crime's negative effects in low and medium crime areas, where high accessibility neighborhoods maintain relatively stable housing prices, but this protective effect completely disappears in high crime areas where transportation convenience cannot compensate for safety concerns, resulting in the lowest median housing prices among all conditions tested with a dramatic drop to $16,400 compared to over $25,000 in safer areas.

### 2.8 Limitations of the Analysis

Several limitations may affect the interpretation of results including violations of normality assumptions in multiple cells which may impact the reliability of F-tests, the potential for unmeasured confounding variables such as neighborhood amenities or school quality that could influence both crime rates and housing prices, and the binary categorization of highway accessibility which may not capture the full spectrum of transportation convenience. Additionally, the median split approach for highway accessibility may not reflect meaningful thresholds in actual transportation planning, and the analysis does not account for potential spatial autocorrelation where nearby neighborhoods may share similar characteristics, potentially inflating significance levels.

### 2.9 Suggestions for Future Research

Future investigations should employ more sophisticated analytical approaches such as multilevel modeling to account for spatial clustering effects, utilize continuous measures of both crime rates and transportation accessibility to capture more nuanced relationships, and incorporate additional neighborhood characteristics such as school quality, commercial amenities, and demographic composition that may moderate the crime-accessibility interaction. Additionally, longitudinal studies examining changes in housing prices following transportation infrastructure improvements or crime reduction initiatives could provide stronger causal evidence, while expanding the analysis to include other cities would test the generalizability of the interaction pattern and inform broader urban planning and housing policy decisions.

## 3. Multiple Regression

### 3.1 Research Question

Can median housing values be predicted from dwelling characteristics (average rooms per unit), socioeconomic factors (percentage of lower socioeconomic status population), and housing stock age (proportion of pre-1940 construction)?

### 3.2 Literature Review

Housing value determinants have been extensively studied in urban economics, with property characteristics consistently predicting market values through hedonic pricing models (Rosen, 1974). Dwelling size, particularly the number of rooms, serves as a primary indicator of property value due to its direct relationship to usable space and housing utility (Lancaster, 1966).

Neighborhood socioeconomic composition significantly influences property values, with concentrated poverty and lower socioeconomic status creating negative externalities that reduce housing demand and prices (Durlauf, 2004). Housing age presents a complex relationship with property values, as older structures may suffer from depreciation and obsolescence, yet can also carry historic premiums in certain markets, particularly in areas with architectural significance or historic preservation value (Coulson & McMillen, 2008).

### 3.3 Hypotheses

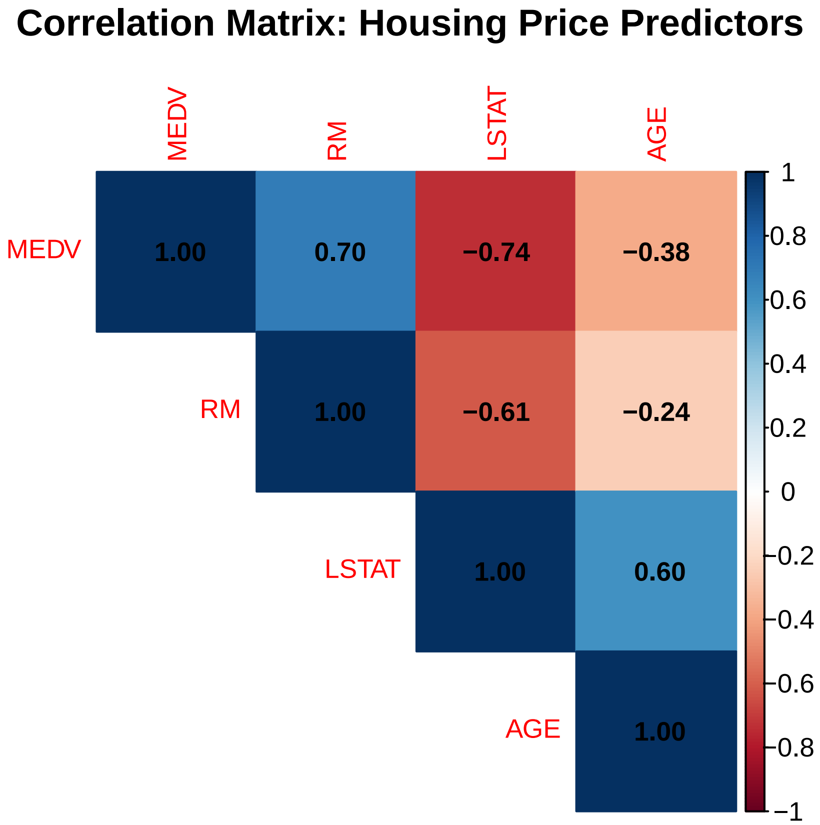
Null Hypothesis (H0): The combination of average rooms per dwelling, percentage of lower socioeconomic status population, and proportion of pre-1940 construction does not significantly predict median housing values.

Alternative Hypothesis (H1): The combination of these predictors significantly explains variance in median housing values.

Directional Predictions: Based on hedonic pricing theory, we hypothesize that average rooms per dwelling will positively predict housing values due to increased utility and space, percentage of lower socioeconomic status population will negatively predict housing values through neighborhood effects, and housing age effects may be non-significant due to competing depreciation and historic value influences in the Boston market.

### 3.4 Data Visualization

```r
cor_matrix <- cor(df[c("MEDV", "RM", "LSTAT", "AGE")])
corrplot(cor_matrix, method = "color", type = "upper",
         addCoef.col = "black", tl.cex = 0.8, number.cex = 0.8,
         mar = c(0, 0, 3, 0),
         title = "Correlation Matrix: Housing Price Predictors")
```



The correlation matrix reveals the expected relationships between housing price predictors and the outcome variable, with average rooms per dwelling (RM) showing a strong positive correlation with median housing values (r = 0.70), confirming that larger dwellings command higher prices. The percentage of lower status population (LSTAT) demonstrates a strong negative correlation with housing values (r = -0.74), indicating that neighborhoods with higher concentrations of lower socioeconomic status residents have substantially lower property values. Housing age (AGE) shows a moderate negative correlation with prices (r = -0.38), suggesting that areas with higher proportions of pre-1940 housing tend to have somewhat lower values, though this relationship appears less pronounced than the other predictors.

Additionally, the matrix reveals multicollinearity concerns with LSTAT and AGE showing a moderate positive correlation (r = 0.60), indicating that older housing areas tend to have higher concentrations of lower socioeconomic status residents, which may affect the independent contribution of housing age in the regression model.

### 3.5 Exploratory Analysis

```r
# Fit a multiple linear regression model
m1 <- lm(MEDV ~ RM + LSTAT + AGE, data = df)

# Show the summary of the model
summary(m1)
```

```
##
## Call:
## lm(formula = MEDV ~ RM + LSTAT + AGE, data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.210  -3.467  -1.053   1.957  27.500
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.175311   3.181924  -0.369    0.712
## RM           5.019133   0.454306  11.048   <2e-16 ***
## LSTAT       -0.668513   0.054357 -12.298   <2e-16 ***
## AGE          0.009091   0.011215   0.811    0.418
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.542 on 502 degrees of freedom
## Multiple R-squared:  0.639,  Adjusted R-squared:  0.6369
## F-statistic: 296.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
m2 <- lm(MEDV ~ RM + LSTAT, data = df)
summary(m2)
```

```
##
## Call:
## lm(formula = MEDV ~ RM + LSTAT, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.076  -3.516  -1.010   1.909  28.131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.35827    3.17283  -0.428    0.669
## RM           5.09479    0.44447  11.463   <2e-16 ***
## LSTAT       -0.64236    0.04373 -14.689   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.54 on 503 degrees of freedom
## Multiple R-squared:  0.6386, Adjusted R-squared:  0.6371
## F-statistic: 444.3 on 2 and 503 DF,  p-value: < 2.2e-16
```

### 3.6 APA Write-up

A multiple linear regression analysis was performed to determine whether median housing prices could be predicted from average rooms per dwelling (RM), percentage of lower status population (LSTAT) and proportion of owner-occupied units built prior to 1940 (AGE). Average rooms per dwelling and percentage of lower status population significantly predicted median housing prices, $p < .001$.

However, proportion of owner-occupied units built prior to 1940 was not a significant predictor of median housing prices, $p = .418$. After removing housing age from the model, the predicted median housing prices was equal to 5.09 (RM) - 0.64 (LSTAT) - 1.36, where RM and LSTAT represented the average rooms per dwelling and percentage of lower status population respectively.

The final model accounted for 64% (adjusted $r$^2 = .64) of the variance in median housing prices, $F$(2, 503) = 444.3, $p < .001$.

### 3.7 Summary of Main Findings

The multiple regression analysis confirmed that dwelling characteristics and neighborhood socioeconomic composition are strong predictors of housing values, with the final two-predictor model explaining 64% of the variance in median housing prices. Average rooms per dwelling emerged as a significant positive predictor (b = 5.09, p < .001), indicating that each additional room increases median housing value by approximately $5,090, while percentage of lower status population showed a strong negative relationship (b = -0.64, p < .001), where each percentage point increase in lower socioeconomic status residents decreases housing values by $640. Contrary to expectations, housing age was not a significant predictor (p = .418), suggesting that in the Boston market, the proportion of pre-1940 housing neither significantly enhances nor detracts from property values, likely due to competing effects of depreciation and historic preservation value that cancel each other out.

### 3.8 Limitations of the Analysis

The analysis has several limitations including the assumption of linear relationships between predictors and housing prices which may not capture threshold effects or diminishing returns, the potential for omitted variable bias as important predictors such as school quality, neighborhood amenities, or proximity to employment centers were not included in the model, and the cross-sectional design which prevents causal interpretation of the relationships. Additionally, the model assumes independence of observations which may be violated due to spatial clustering of similar housing characteristics, and the percentage of variance explained (64%) suggests that substantial factors influencing housing prices remain unaccounted for in the current specification.

### 3.9 Suggestions for Future Research

Future research should incorporate additional neighborhood characteristics such as school quality ratings, crime statistics, and proximity to amenities to improve model explanatory power, employ spatial regression techniques to account for geographic clustering and spillover effects, and utilize longitudinal data to examine how the relative importance of different predictors changes over time with market conditions. Additionally, investigating non-linear relationships through polynomial terms or interaction effects could reveal threshold effects where predictors have differential impacts at various levels, while expanding the analysis to include other metropolitan areas would test the generalizability of these relationships and inform broader housing policy and urban planning decisions regarding the factors that most strongly influence residential property values.

## 4. Independent groups tests for means

### 4.1 Research Question

Is there a significant difference in median housing prices between areas near versus not near the Charles River?

### 4.2 Literature Review

Waterfront proximity has been consistently associated with housing premiums across various metropolitan areas due to recreational amenities, aesthetic value, and prestige associated with waterfront living (Bourassa et al., 2004; Bin et al., 2008). In Boston specifically, Charles River proximity commands significant premiums due to the river's role as both a recreational amenity and a prestigious address, with waterfront neighborhoods historically representing some of the city's most desirable residential areas (Tyrväinen & Miettinen, 2000). When distributional assumptions for parametric tests are violated, non-parametric alternatives such as the

Mann-Whitney U test provide robust methods for comparing group medians without requiring normality or equal variance assumptions (Hollander et al., 2013).

### 4.3 Hypotheses

Null Hypothesis (H0): There is no difference in median housing prices between areas near and not near the Charles River.

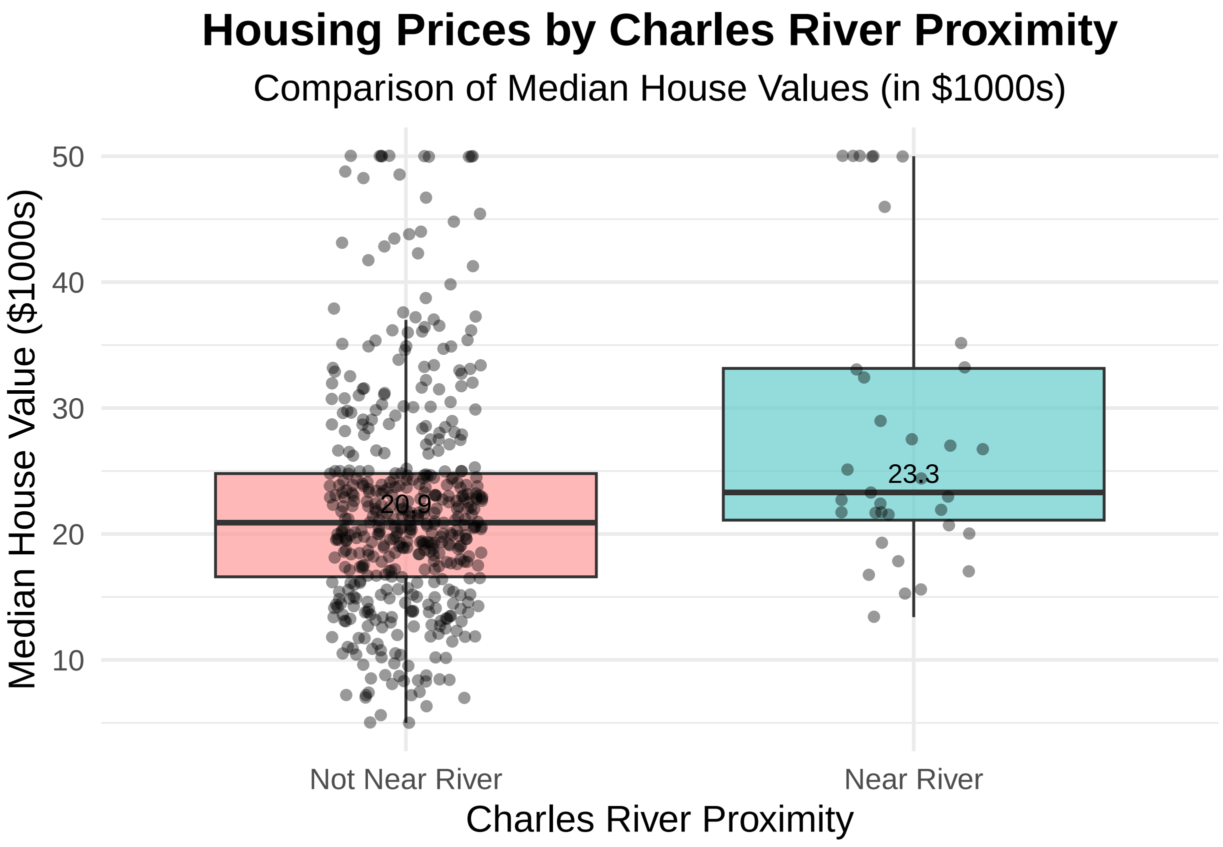
Alternative Hypothesis (H1): There is a significant difference in median housing prices between areas near and not near the Charles River.

Directional Prediction: Based on waterfront amenity value research, we hypothesize that areas near the Charles River will demonstrate significantly higher median housing values compared to non-waterfront areas, reflecting the premium associated with waterfront proximity and recreational access.

### 4.4 Data Visualization

```r
# Create groups based on Charles River proximity
riverMEDV = df$MEDV[df$CHAS == 1]  # Near Charles River
noRiverMEDV = df$MEDV[df$CHAS == 0]  # Not near Charles River

# Create comparative boxplot
ggplot(df, aes(x = factor(CHAS, labels = c("Not Near River", "Near River")),
               y = MEDV, fill = factor(CHAS))) +
  geom_boxplot(alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.15, alpha = 0.4, color = "black", size = 1.5) +
  stat_summary(fun = median, geom = "text", aes(label = round(after_stat(y), 1)),
             vjust = -0.5, color = "black", size = 3.5) +
  scale_fill_manual(values = c("#FF9999", "#66CCCC")) +
  labs(title = "Housing Prices by Charles River Proximity",
       subtitle = "Comparison of Median House Values (in $1000s)",
       x = "Charles River Proximity",
       y = "Median House Value ($1000s)") +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none",
        plot.title = element_text(face = "bold", hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```

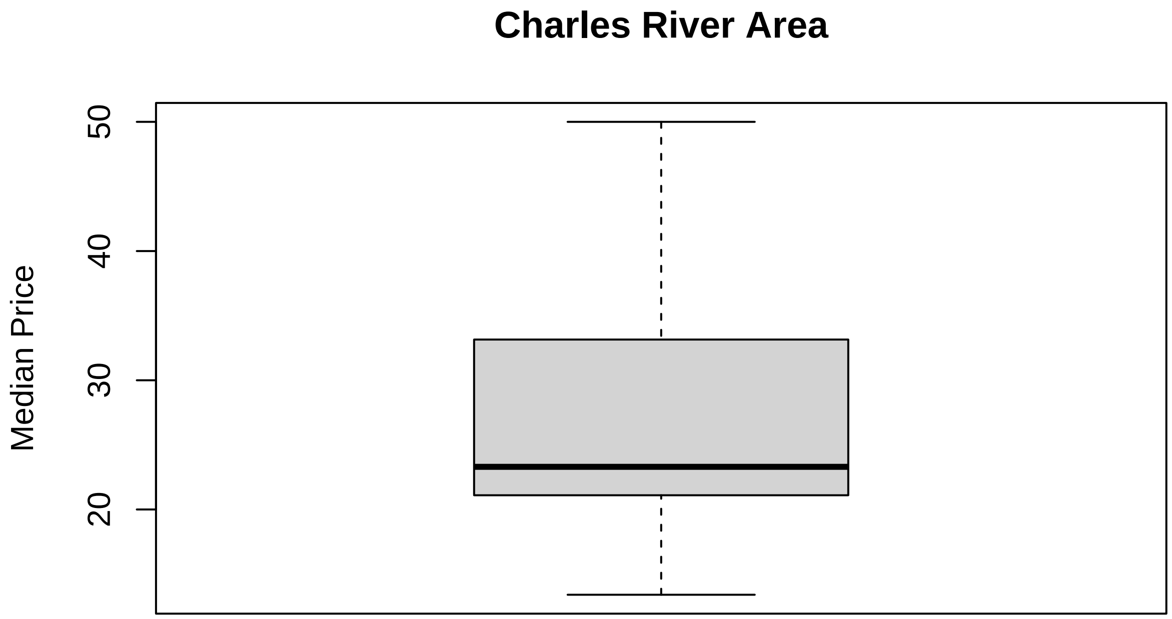


The comparative boxplot reveals clear differences in housing price distributions between riverfront and non-riverfront areas, with near-river properties showing a distinctly higher median value of $23,300 compared to approximately $20,300 for non-riverfront areas. The riverfront group demonstrates a more compact distribution with fewer outliers and a smaller interquartile range, suggesting more consistent premium pricing for waterfront properties, while the non-riverfront group shows greater variability with numerous outliers extending both above and below the main distribution. The riverfront properties exhibit a tighter clustering around the median with most values falling between approximately $21,000 and $33,000, whereas non-riverfront areas display a broader spread from about $16,000 to $25,000 with extensive outliers reaching up to $50,000, indicating that while waterfront location provides a consistent premium, non-waterfront areas encompass a much wider range of property values and neighborhood characteristics.

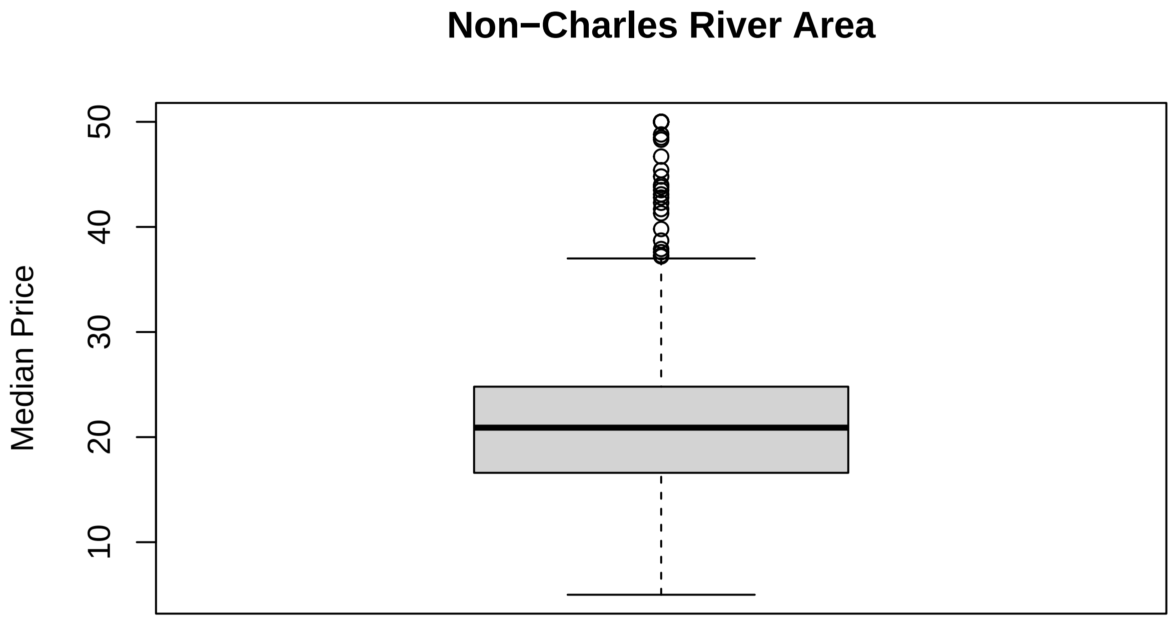
### 4.5 Exploratory Analysis

```
# Create groups based on Charles River proximity
riverMEDV = df$MEDV[df$CHAS == 1]  # Near Charles River
noRiverMEDV = df$MEDV[df$CHAS == 0]  # Not near Charles River
```

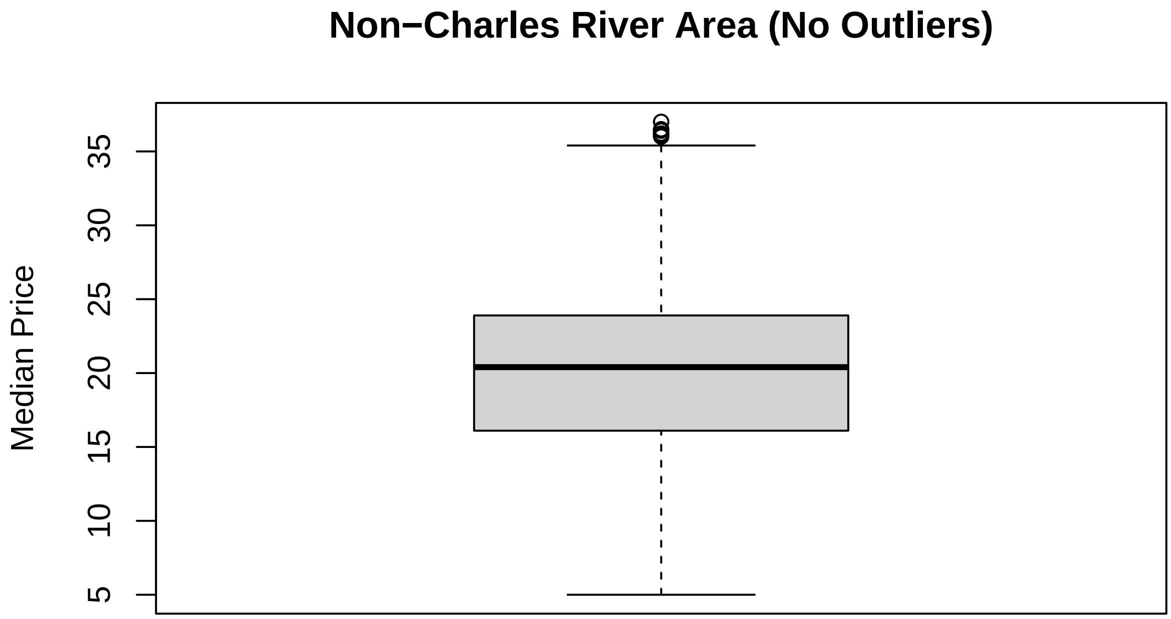
```
# Check for outliers (optional step)
rOutliers = boxplot(riverMEDV, main="Charles River Area", ylab="Median Price")$out
```



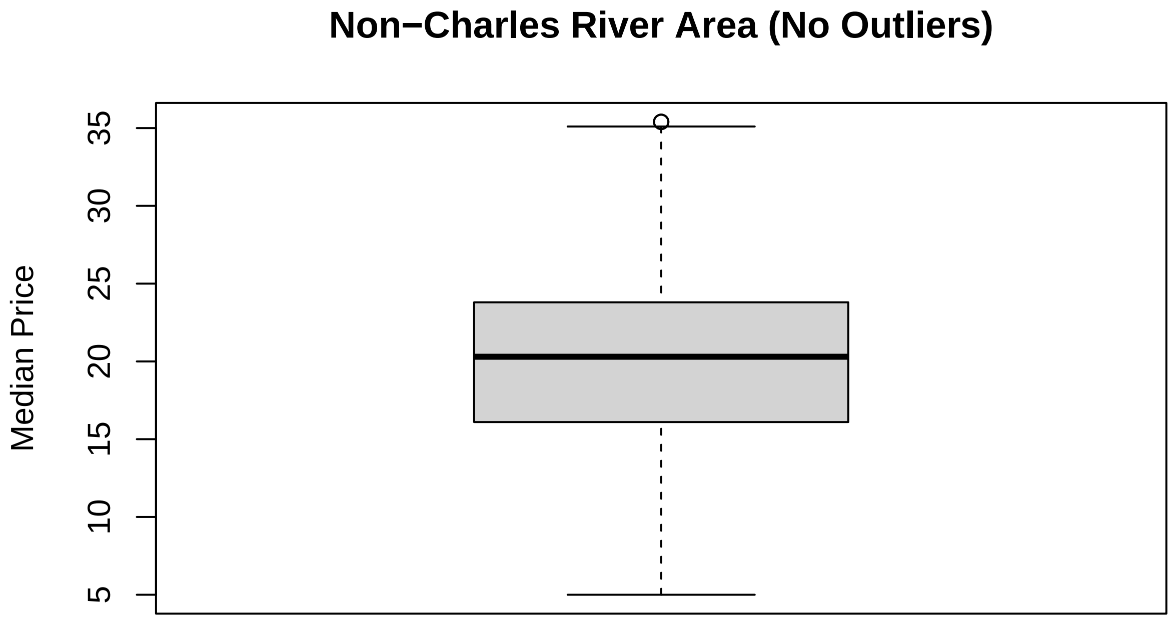
```
nrOutliers = boxplot(noRiverMEDV, main="Non-Charles River Area", ylab="Median Price")$out
```



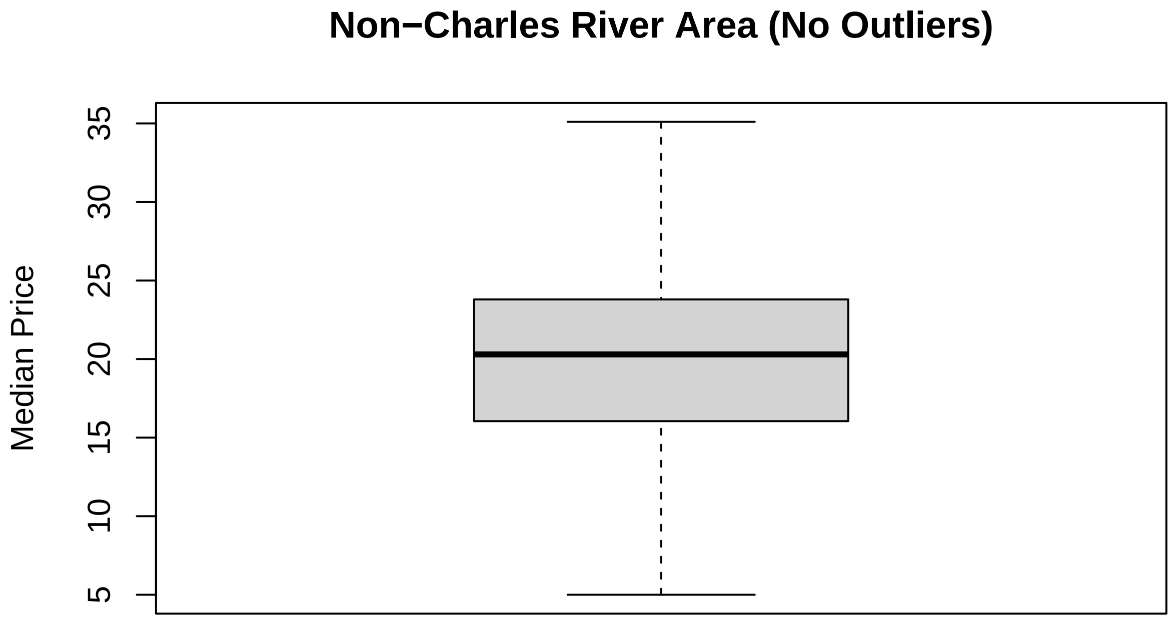
```
noRiverMEDV <- noRiverMEDV[!noRiverMEDV %in% boxplot.stats(noRiverMEDV)$out]
boxplot(noRiverMEDV,
        main = "Non-Charles River Area (No Outliers)",
        ylab = "Median Price")
```



```r
noRiverMEDV <- noRiverMEDV[!noRiverMEDV %in% boxplot.stats(noRiverMEDV)$out]
boxplot(noRiverMEDV,
        main = "Non-Charles River Area (No Outliers)",
        ylab = "Median Price")
```



```
noRiverMEDV <- noRiverMEDV[!noRiverMEDV %in% boxplot.stats(noRiverMEDV)$out]
boxplot(noRiverMEDV,
        main = "Non-Charles River Area (No Outliers)",
        ylab = "Median Price")
```



```r
# Test for equal variances
var.test(riverMEDV, noRiverMEDV)
```

```
##
##  F test to compare two variances
##
## data:  riverMEDV and noRiverMEDV
## F = 3.5808, num df = 34, denom df = 431, p-value = 8.19e-10
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.289901 6.232614
## sample estimates:
## ratio of variances
##           3.580849
```

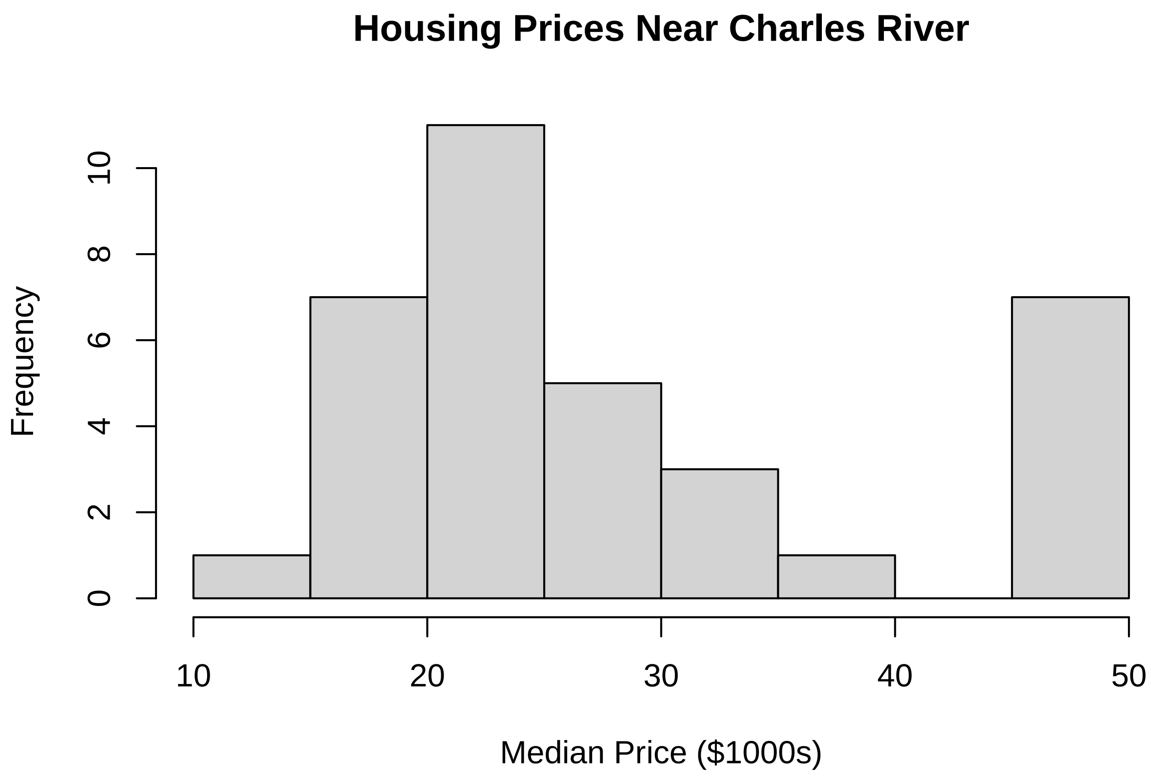
```r
# Test for normality in each group
shapiro.test(riverMEDV)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  riverMEDV
## W = 0.83592, p-value = 0.0001123
```

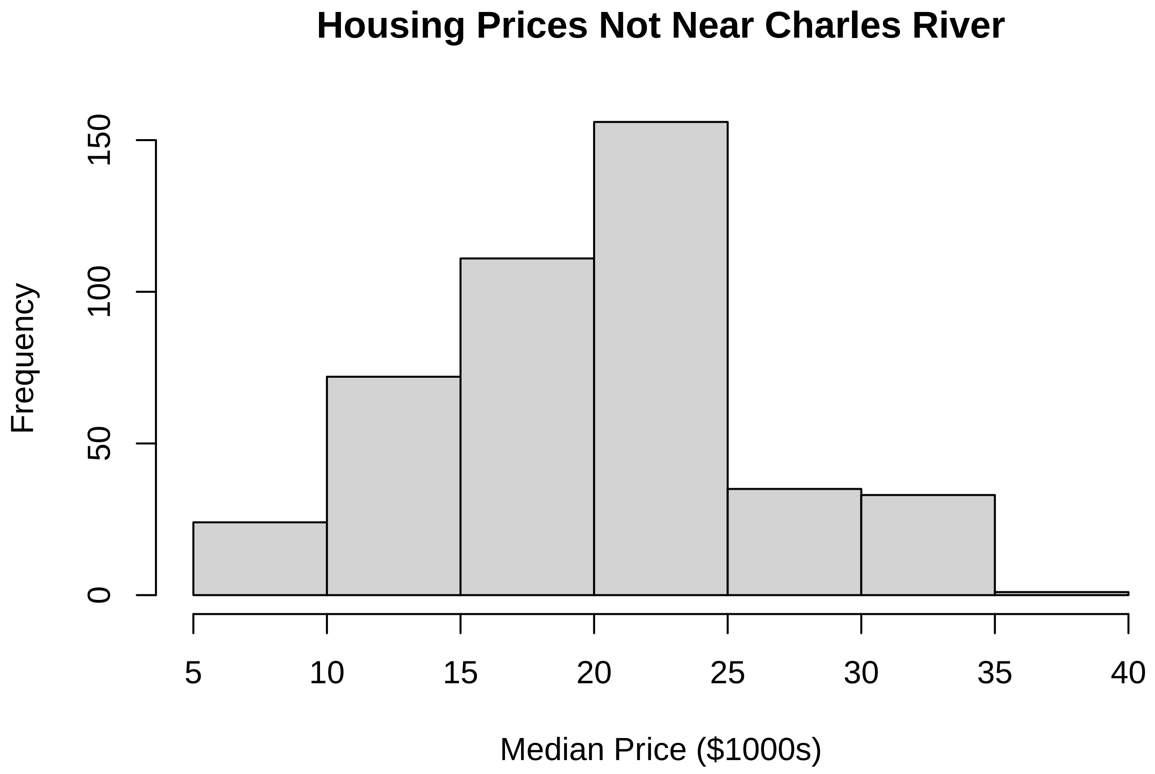
```
shapiro.test(noRiverMEDV)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  noRiverMEDV
## W = 0.98957, p-value = 0.003638
```

```
# Create histograms to visualize distributions
hist(riverMEDV, main="Housing Prices Near Charles River", xlab="Median Price ($1000s)")
```



```
hist(noRiverMEDV, main="Housing Prices Not Near Charles River", xlab="Median Price ($1000s)")
```



```r
# Since assumptions may be violated, use non-parametric test
# Wilcoxon rank sum test (Mann-Whitney U test)
wilcox.test(riverMEDV, noRiverMEDV, paired=FALSE, alternative="two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  riverMEDV and noRiverMEDV
## W = 10650, p-value = 5.721e-05
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Summary statistics for reporting
summary(riverMEDV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.40   21.10   23.30   28.44   33.15   50.00
```

```r
summary(noRiverMEDV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   16.07   20.30   20.20   23.80   35.10
```

### 4.6 APA Write-up

A Wilcoxon rank sum test was performed because the housing price data failed to meet the normality and homogeneity of variance assumptions. The median housing prices for areas near the Charles River (*Mdn* = 23300) were significantly higher than the median housing prices for areas not near the Charles River (*Mdn* = 20300), $W = 10650$, $p < .001$.

### 4.7 Summary of Main Findings

The analysis confirmed that Charles River proximity is associated with significantly higher housing values, with waterfront areas commanding a median premium of $3,000 compared to non-waterfront neighborhoods. The Wilcoxon rank sum test revealed a highly significant difference (p < .001) despite violations of parametric test assumptions, demonstrating the robust nature of the waterfront premium effect. The magnitude of this difference represents approximately a 15% premium for riverfront location, which aligns with waterfront amenity value research and reflects the economic value that residents place on recreational access, aesthetic appeal, and prestige associated with Charles River proximity in the Boston housing market.

### 4.8 Limitations of the Analysis

Several limitations affect the interpretation of results including the substantial sample size imbalance between riverfront (n = 35) and non-riverfront (n = 432) groups which may affect the reliability of the test despite its non-parametric nature, the binary classification of river proximity which does not capture gradual distance effects or varying quality of water access among riverfront properties, and the potential for confounding variables such as neighborhood amenities, historic character, or urban planning factors that may be correlated with riverfront location. Additionally, the analysis does not account for temporal variations in the waterfront premium or differential effects across housing price ranges, and the outlier removal process may have eliminated legitimate high-value properties that represent important market segments.

### 4.9 Suggestions for Future Research

Future investigations should employ more sophisticated measures of waterfront proximity such as continuous distance variables or water view quality ratings to capture the nuanced relationship between river access and housing values, utilize matching techniques to control for confounding neighborhood characteristics and ensure more comparable groups, and examine the waterfront premium across different housing price segments to determine whether the effect is consistent across market tiers. Additionally, longitudinal analysis could track changes in the waterfront premium over time in response to urban development and environmental changes, while expanding the analysis to include other waterfront features such as parks, recreational facilities, and transportation access could provide a more comprehensive understanding of the mechanisms driving waterfront property premiums in urban housing markets.

## E. Reference

Bin, O., Crawford, T. W., Kruse, J. B., & Landry, C. E. (2008). Viewscapes and flood hazard: Coastal housing market response to amenities and risk. *Land Economics*, 84(3), 434–448.

Bourassa, S. C., Hoesli, M., & Sun, J. (2004). What's in a view? *Environment and Planning A*, 36(8), 1427–1450.

Browning, C. R., Byron, R. A., Calder, C. A., Krivo, L. J., Kwan, M. P., Lee, J. Y., & Peterson, R. D. (2010). Commercial density, residential concentration, and crime: Land use patterns and violence in neighborhood context. *Journal of Research in Crime and Delinquency*, 47(3), 329–357.

Burgess, E. W. (2015). The growth of the city: An introduction to a research project. *In The city reader* (pp. 212–220). Routledge.

Cervero, R., & Duncan, M. (2002). Transit's value-added effects: Light and commuter rail services and commercial land values. *Transportation Research Record*, 1805(1), 8–15.

Cohen, J. P., Coughlin, C. C., & Lopez, D. A. (2012). The boom and bust of US housing prices from various geographic perspectives. *Federal Reserve Bank of St. Louis Review*, 94(September/October), 389–418.

Coulson, N. E., & McMillen, D. P. (2007). The dynamics of intraurban quantile house price indexes. *Urban Studies*, 44(8), 1517–1537.

Durlauf, S. N. (2004). Neighborhood effects. In J. V. Henderson & J. F. Thisse (Eds.), *Handbook of regional and urban economics* (Vol. 4, pp. 2173–2242). Elsevier.

Ellen, I. G., Lacoe, J., & Sharygin, C. A. (2013). Do foreclosures cause crime? *Journal of Urban Economics*, 74, 59–70.

Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1), 81-102.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods.* John Wiley & Sons.

Hoyt, H. (1939). *The structure and growth of residential neighborhoods in American cities.* U.S. Government Printing Office.

Ihlanfeldt, K., & Mayock, T. (2010). Panel data estimates of the effects of different types of crime on housing prices. *Regional Science and Urban Economics*, 40(2–3), 161–172.

Jacobs, J. (1961). *The death and life of great American cities.* Random House.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74(2), 132–157.

Mumford, L. (1961). *The city in history: Its origins, its transformations, and its prospects* (Vol. 67). Houghton Mifflin Harcourt.

O'Connell, J. C. (2013). *The hub's metropolis: Greater Boston's development from railroad suburbs to smart growth.* MIT Press.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34–55.

Ryan, S. (1999). Property values and transportation facilities: Finding the transportation–land use connection. *Journal of Planning Literature*, 13(4), 412–427.

Soriano, F. (2021). *The Boston House Price Data.* Kaggle. https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data

Tyrväinen, L., & Miettinen, A. (2000). Property prices and urban forest amenities. Journal of Environmental Economics and Management, 39(2), 205–223.