STATS 101A - An Analysis of Housing Prices in Boston

Isabelle Tandradinata, Samuel Reade, Maureen Widjaja, Celine Nugroho, Claire Nabours, Madalyn Do

Introduction

In this project, we aim to construct a predictive model of housing prices to explore how various housing characteristics impact the price of houses. The dataset we used was obtained from Kaggle and describes housing prices located in the Northeastern United States in 1970. It contains 545 observations and we selected 6 variables to analyze:

- **Price** (numeric continuous): Measure of the house cost in an undisclosed unit
- **Area** (numeric continuous): Measures the house size in an undisclosed unit
- **Bedrooms** (numeric discrete): The count of bedrooms per house
- **Bathrooms** (numeric discrete): The count of bathrooms per house
- **Mainroad** (categorical): Indicates whether the house is located on a main road
- **Furnishing status** (categorical): Indicates whether the house is unfurnished

Price will be our response variable and Area, Bedrooms, Bathrooms, Mainroad, and Furnishing Status are our predictor variables. Multiple linear regression was used to model this relationship since the initial scatterplot matrix indicated some linear correlation between the variables. The analysis for this study was conducted using R.

After conducting exploratory data analysis to identify correlations and distributions, we will first fit a full multiple linear regression model. Then we will apply necessary transformations and select key variables, performing diagnostic checks throughout the process. Finally, we will fit a final MLR model, interpret the final coefficients, and discuss real-world applications.

Data Description

We begin by reviewing the summary statistics in Table 1.

Table 1 presents each variable's mean and standard deviation. We can see that price and area have high variability, while bedrooms and bathrooms are more consistent. Furthermore, houses in this data set average about 3 bedrooms and 1 bathroom. Since mainroad and furnishing status are categorical they lack mean and standard deviation values.

Variable	Mean	Standard Deviation
Price	4766729	1870440
Area	5150.541	2170.141
Bedrooms	2.965138	0.7380639
Bathrooms	1.286239	0.5024696
Mainroad	N/A	N/A
Furnishing Status	N/A	N/A

Table 1: Means and standard deviations

Observing the visual distribution of each variable in Figure 1, we can see that price, area, and bathrooms are right-skewed. Bedrooms is unimodal, peaking at 3 which is consistent with Table 1. Also, mainroad and furnishing status are binary, with most homes being near a main road and furnished.

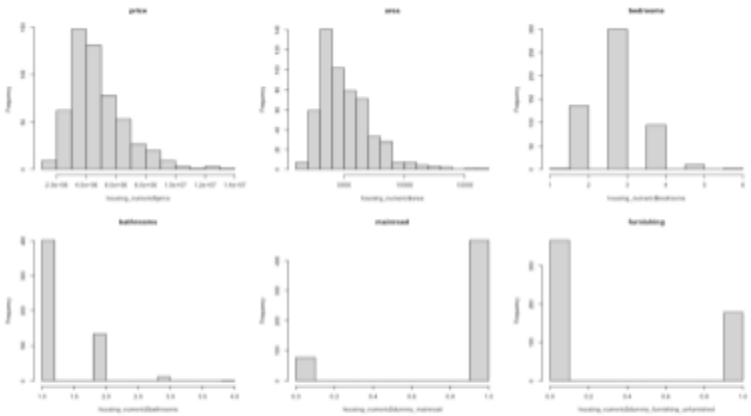

Figure 1: Variable distributions

The scatter plot matrix in Figure 2 highlights key relationships in the dataset. Price shows a positive linear correlation with area, though non-constant variance suggests a need for transformation. Bathrooms moderately correlate with price, while bedrooms lack a clear trend. Mainroad shows that prices are higher if they are next to a mainroad. Furnishing status shows higher prices for furnished homes. The correlation matrix in Table 2 shows the possibility of multicollinearity concerns. It indicates that bedrooms and bathrooms are somewhat correlated, potentially affecting regression stability. Thus, careful variable selection and transformations are essential for meaningful insights. Given the linear relationship between price and area and the impact of non-continuous variables on price, we begin our analysis with a multiple linear regression model.

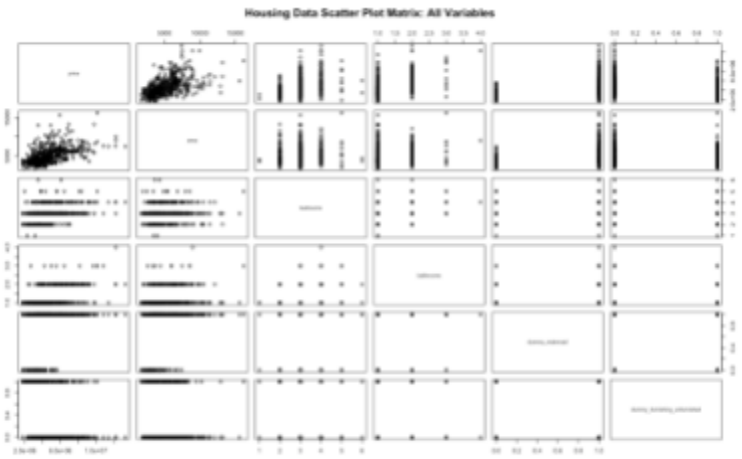

Figure 2: Scatter Plot Matrix

	price	area	bedrooms	bathrooms	dummy_mainroad	dummy_furnishing_unfurnished
price	1.0000000	0.5359973	0.36649483	0.51754534	0.29689849	-0.2805874
area	0.5359973	1.0000000	0.15185849	0.19381953	0.28887411	-0.1422782
bedrooms	0.3664948	0.1518585	1.00000000	0.37393824	-0.01203324	-0.1262520
bathrooms	0.5175453	0.1938195	0.37393824	1.00000000	0.04239762	-0.1321875
dummy_mainroad	0.2968985	0.2888741	-0.01203324	0.04239762	1.00000000	-0.1331233
dummy_furnishing_unfurnished	-0.2805874	-0.1422782	-0.12625202	-0.13210748	-0.13312327	1.0000000

Table 2: Correlation Matrix

Results & Interpretation

```
Call:
lm(formula = price ~ area + bedrooms + bathrooms + mainroad +
    furnishing, data = housing)

Residuals:
     Min       1Q   Median       3Q      Max
-3290347  -793032   -69105   521887  6123170

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -917414.61  278800.32  -3.291  0.00107 **
area            325.16      27.24  11.937  < 2e-16 ***
bedrooms     402834.34   81218.18   4.960 9.47e-07 ***
bathrooms   1340360.48  119945.28  11.175  < 2e-16 ***
mainroad     837627.02  166317.25   5.036 6.48e-07 ***
furnishing   551845.23  120291.14   4.588 5.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1285000 on 539 degrees of freedom
Multiple R-squared:  0.532,     Adjusted R-squared:  0.5277
F-statistic: 122.6 on 5 and 539 DF,  p-value: < 2.2e-16
```

Figure 3: Original model summary in R

Our initial model is a full model with all untransformed variables.

From the output, we see that all variables are significant with increases in area, bedrooms, bathrooms, presence of a mainroad, and a furnished status to increase the price of a house. The model's R^2 indicates that 53.20% of the variance in Price is explained by the explanatory variables in our model. The overall F-test had a p-value of < 2.2e-16, which is smaller than 0.05, thus we conclude at least 1 explanatory variable significantly contributes to predicting the response variable. Therefore, the regression model is statistically significant.

Based on Figure 3, the full multiple linear regression equation is:

$$\widehat{Price} = \hat{\beta}_0 + \hat{\beta}_1(area) + \hat{\beta}_2(bedrooms) + \hat{\beta}_3(bathrooms) + \hat{\beta}_4(mainroad) + \hat{\beta}_5(furnishingstatus)$$

The standardized residual plot for area shows that the points are randomly distributed around the x-axis, meaning there's constant variance but there is slight clustering on the left side.

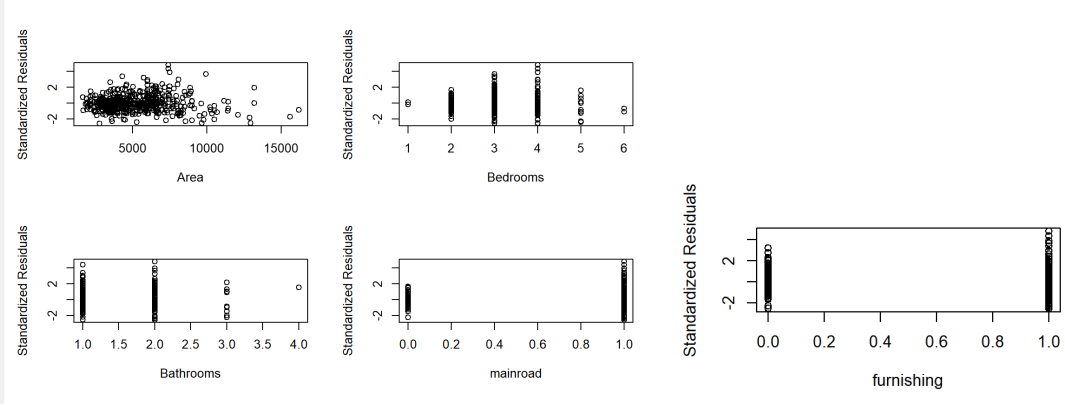

Figure 4: Standardised residual plots

Since bedrooms, bathrooms, mainroad, and furnishing are all categorical variables they should be evenly distributed across categories to be valid which is indicated in the plots. Since all the standardized residual plots for each variable are valid it suggests that the model is likely valid for the given predictors. There is an outlier in the bathrooms plot, so we should use other plots to check model validity.

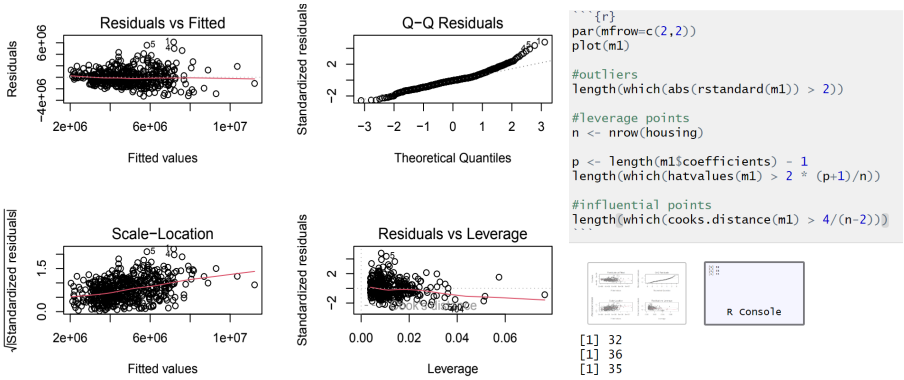

Figure 5: Diagnostic plots

In Figure 5, the Residual vs Fitted plot shows that the error term looks mostly scattered along the horizontal red line. This indicates the average of the error term is around 0, but there is slight clustering on the left side which we will analyze further. In the Q-Q residual plot, most of the data points are aligned with the dotted line, indicating that the error term is normally distributed. However, the left end of the data points are below the dotted line and the right endpoints go above the dotted line, indicating that the data is skewed. The Scale-Location plot looks mostly random, but we can see that the data points are more clustered on the left, indicating that the variance of the error term is not constant for all data points. Finally, the Residuals vs. Leverage plot confirms outliers, with 32 outliers, 36 high-leverage points, and 35 influential points. The existence of these points can cause points in the model to deviate, causing skewness of the fitted model.

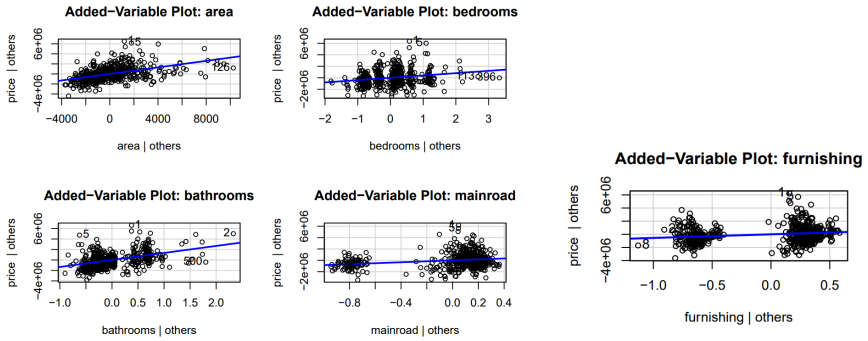

Figure 6: Added variable plots

Additionally, we examine added variable plots to examine the significance of each variable to the price of housing in Boston. The plot shows that area and bathroom have a strong, positive linear relationship with price. While bedrooms, mainroad, and furnishing also share a positive linear relationship with price, their linear relationship with price is rather weak as the trend line looks flatter in comparison to area and bathrooms'. It suggests that some type of transformation may be needed for each of these variables to improve the model's linearity and overall fit.

To better satisfy the model assumptions, we decided to attempt transformation in 2 ways: Box-Cox method for all numeric predictors and the response variable (simultaneously), and using the Box-Cox method for all numeric predictors before transforming the response variable with an inverse response plot. We did not include the categorical variables in any of the transformations as the Box-Cox method is not appropriate for categorical data.

```
bcPower Transformations to Multinormality
           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
price         0.0047          0.0       -0.1569       0.1664
area         -0.1366          0.0       -0.3006       0.0273
bedrooms      0.3561          0.5        0.1027       0.6096
bathrooms    -4.4977         -4.5       -5.0063      -3.9891

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

Figure 7: Box-Cox for numeric predictors and the response variable

The first attempt in which we applied the Box-Cox method for numeric predictors and the response variable suggested log transformations for price and area, while bedrooms and bathrooms were to be raised to the powers 0.5 and -4.5 respectively.

```
bcPower Transformations to Multinormality
           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
area         -0.1197          0.0       -0.2987       0.0592
bedrooms      0.4256          0.5        0.1688       0.6825
bathrooms    -4.5034         -4.5       -5.0114      -3.9953

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

	LRT <dbl>	df <int>	pval <chr>
LR test, lambda = (0 0 0)	423.3768	3	< 2.22e-16

	LRT <dbl>	df <int>	pval <chr>
LR test, lambda = (1 1 1)	843.995	3	< 2.22e-16

lambda <dbl>	RSS <dbl>
0.1183518	4.407001e+14
-1.0000000	1.020065e+15
0.0000000	4.413127e+14
1.0000000	4.733394e+14

Figure 8: Box-Cox for numeric predictors (left), Inverse response plot for response variable (right)

Meanwhile, the second attempt in which we applied the Box-Cox method to just the numeric predictors and later transformed the response variable with an inverse response plot showed similar results to the first attempt. While the inverse response plot suggested that $\lambda = 0.1183518$ would produce the smallest RSS, we chose $\lambda = 0$ to transform the response logarithmically since the latter produces a comparably small RSS. Hence, the transformed model is as follows:

$$\widehat{\log(price)} = \hat{\beta}_0 + \hat{\beta}_1 log(area) + \hat{\beta}_2 bedrooms^{0.5} + \hat{\beta}_3 bathrooms^{-4.5} + \hat{\beta}_4 mainroad + \hat{\beta}_5 furnishing$$

```
Call:
lm(formula = log(price) ~ log(area) + I(bedrooms^0.5) + I(bathrooms^-4.5) +
    mainroad + furnishing, data = housing)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77003 -0.15134  0.00712  0.13446  0.70196

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       11.44445    0.24935  45.897  < 2e-16 ***
log(area)          0.38657    0.02869  13.475  < 2e-16 ***
I(bedrooms^0.5)    0.31965    0.05319   6.010 3.42e-09 ***
I(bathrooms^-4.5) -0.26555    0.02706  -9.813  < 2e-16 ***
mainroad           0.16326    0.03211   5.084 5.12e-07 ***
furnishing         0.15082    0.02292   6.579 1.12e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2447 on 539 degrees of freedom
Multiple R-squared:  0.5716,    Adjusted R-squared:  0.5676
F-statistic: 143.8 on 5 and 539 DF,  p-value: < 2.2e-16
```

Figure 9: Transformed output

The output in Figure 9 indicates that area, the number of bedrooms, the location on a main road, and the presence of furnishing have a positive effect on price, while the number of bathrooms has a negative effect on price. All the predictors are significant with p-values < 0.05, with 57.16% variability explained by the model, given by the model's R^2 which is 0.5716.

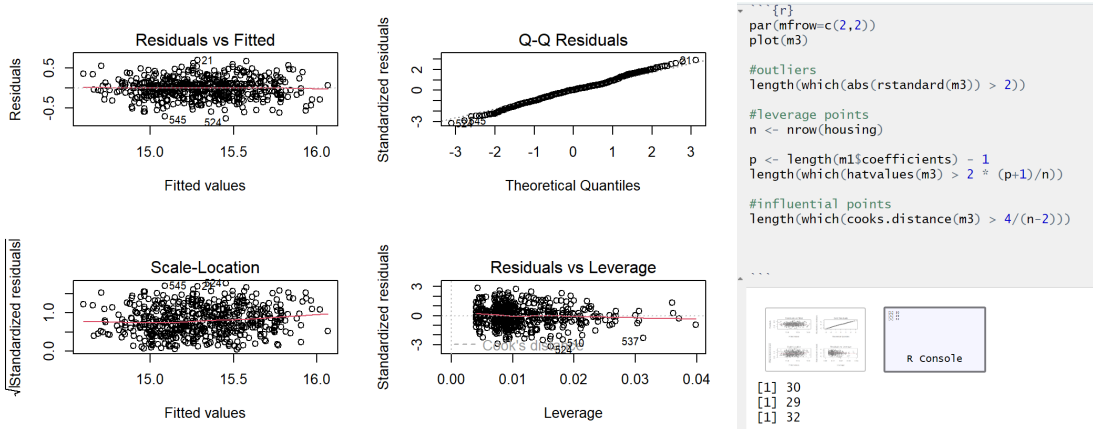

Figure 10: Transformed model diagnostic plots

After transforming our original model, we can see that the overall fit of our model has significantly improved. In our original model, the normality and constant variance of errors were slightly violated. Our new model's Residuals vs Fitted plot now shows more randomness, with a more constant variance of the error term. Q-Q residuals plot confirms the normality of the error term as well. The Scale-Location plot shows the errors now have constant variance as the errors are shown randomly distributed across the plot and the red line becoming more horizontal. Additionally, the transformed model has fewer outliers, leverage points, and influential points than the original.

Since the bathrooms variable has a negative coefficient, which is not expected, we decided to do a partial F-test with a reduced model without said variable. The test results showed sufficient evidence to choose the full model over the reduced model, so we proceeded with the full transformed model.

```
Analysis of Variance Table

Model 1: log(price) ~ log(area) + I(bedrooms^0.5) + mainroad + furnishing
Model 2: log(price) ~ log(area) + I(bedrooms^0.5) + I(bathrooms^-4.5) +
    mainroad + furnishing
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    540 38.046
2    539 32.279  1    5.7669 96.296 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: Partial F-test to evaluate model without bathrooms variable

We also evaluated VIF values for the transformed variables. Seeing that they were all under 5, there was no suggested multicollinearity and we did not proceed with variable selection.

log(area)	I(bedrooms^0.5)	I(bathrooms^-4.5)	mainroad	furnishing
1.185985	1.176399	1.190920	1.138705	1.051867

Figure 12: VIF for transformed variables

Thus, the final model is as follows:

$$\log(\hat{price}) = 11.44 + 0.39log(area) + 0.32bedrooms^{0.5} - 0.27bathrooms^{-4.5} + 0.16mainroad + 0.15furnishing$$

Discussion

The goal of our project is to explore the relationship between various real estate attributes and the price of the properties. Initially, we began with our full, untransformed, model. We investigated the variables individually and with respect to each other, as well as if this model was a good predictor for price. After, using analysis methods, we determined we should transform some of the predictors to improve the prediction capability for price. This decision was influenced mainly by deviations from the model assumptions. To transform our variables we used the Box-Cox method and Inverse response plot. Once we had our ideal transformations we investigated the model again and found that it was improved. We were concerned that one of our variables was negative when we assumed it should be positive, so we ran a partial F-test. The F-test disproved that the full model was not ideal. We also checked for multicollinearity with VIF.

In the context of the real world, the majority of our findings make sense. All the variables we assume should raise the price of a house have positive coefficients. However, the variable bathrooms has a negative coefficient which is slightly concerning. In the two articles below one can view how our model and findings reflect the real world. After doing the partial F-test and checking VIF there was no evidence of multicollinearity or to remove it. This issue would need further analysis in the future. The primary limitation of this model is the number of outliers still present in the final model. Additionally, the numeric discrete and categorical variables used to fit the model may pose an issue, as linear regression is mainly designed for numeric continuous variables. Other transformations or adding more predictors from the original dataset could address these limitations.

Article 1: *https://www.opendoor.com/articles/factors-that-influence-home-value*
Article 2: *https://apadala-90574.medium.com/determinants-of-housing-price-bdefc783cf6b*