**Exploring Housing Values in the Suburbs of Boston**

Group 8

Athiya Rastogi/ Hongbo Huang/ Sankalp Sharma/ Usama Tariq/ Gaonyatswe Ntebogang Mmusi

**Abstract**

Boston has seen a dramatic rise in housing prices over the years (Case & Shiller, 2004). The goal of this paper is to study variables that contribute the most to housing prices in the suburbs of Boston. Factors such as the percentage of the lower status population in town influence these prices greatly as compared to the other factors (Thomas, Moye, Henderson, & Horton, 2017). We have used R for statistical analysis to create a regression model that suggests the relationship between housing values in the suburbs of Boston and relevant factors that influence it.

**Background and Significance**

Purchasing a home is often times the biggest purchase in most people's lives. Understanding the real estate market and the factors that affect housing prices help one make an informed decision when buying a house. Several studies have been done to understand the variables which have the most impact on real estate values. The amount that the government spends on educational institutions, and property tax rates has shown a positive correlation to increased housing values (Schwartz & Voicu, 2003). School characteristics such as teacher-pupil ratio and average school enrollment also show a positive impact on real estate values (Schwartz & Voicu, 2003). Other factors such as increased government spending in an area have shown positive signs for real estate values (Schwartz & Voicu, 2003). Crime rates are linked to decreasing real estate values in an area (Schwartz & Voicu, 2003). Besides, some studies have shown that housing prices initially increase as the distance from the business center increases, and then decrease (Osland & Pryce, 2012). And the concentration of nitric oxides in Boston is higher in the area closer to major employment centers and radial highways. The higher the concentration of nitrogen oxides in one area, the more serious the air pollution is and the lower the house prices in the vicinity will be (Harrison & Rubinfeld, 1978). We are proposing this model to predict housing values in the suburbs of Boston. This will allow us to understand the predictor variables which have the most effect on housing values, which in turn will allow parties interested in these housing markets to make an informed decision about purchasing, investing or further developing housing in the suburbs of Boston.

**Exploratory Data Analysis**

This data set contains 506 observations of 13 predictors that may affect the median house value of houses in the suburbs of Boston.

**Per capita crime rate by town (x1)**

This variable represents the number of crime for each person in town. Per capita crime rate is calculated by dividing the total number of crimes by total population. A higher crime rate by a town is linked with lower property values but this could change with changing in policing techniques employed in the neighbourhood (Schwartz & Voicu, 2003). **Median: 0. 2561 Mean: 3.61325**

**Proportion of residential land zoned for lots over 25,000 sq.ft. (x2)**

This refers to the parts of residential land that is dedicated to development in a lot size that goes beyond 25,000 sq.ft. **Median: 0 Mean:11.36**

**Proportion of non-retail business acres per town (x3)**

Refers to space for non-retail businesses in each town. **Median: 9.69 Mean: 11.14**

**Charles River dummy variable (x4)**

Charles River dummy indicates proximity to the Charles River. 1 if the parcel of land is in close proximity to the River, 0 if the land is not in close proximity to the river. Having a view of water or living near it can be a nice benefit of a real estate dwelling, but in a study conducted on dwellings in Lyon France, it has not shown a significant positive effect on real estate in close proximity to water (Kryvobokov, 2010). Few houses are close to Charles River. **Media: 0 Mean: 0.069**

**Nitric oxide concentration (parts per 10 million) (x5)**

This represents contamination per unit volume of water. **Median: 0.5380 Mean: 0.5547**

**Average number of rooms per dwelling (x6)**

This is the standard amount of rooms per house or apartment in Boston. **Median: 6.208 Mean: 6.825**

**Proportion of owner occupied units built prior to 1940 (x7)**

This represents a fraction of the houses and apartments developed prior to 1940 that is used as the owner's primary residence. **Median: 77.50 Mean: 68.57**

**Weighted distances to five Boston employment centers (x8)**

This is a spatial analysis of the weighted distances to five Boston employment centers is the measured range to the 5 Boston employment district/zone. The travel times to these employment centers may have a positive effect on the surrounding values of residential real estate (Kryvobokov, 2010).  **Median: 3.207 Mean: 3.795**

**Index of accessibility to radial highways (x9)**

This is the arrangement of accessibility to branching highways. Radial highway development is important in suburbs as cars are a pervasive part of the lifestyle in America. Having easier access to radial highways impacts the value of real estate positively but there is a double entendre here;  being too close to a major radial highway where homeowners will be affected by noise pollution and traffic intensity will have negative impacts on the value of real estate (Levkovich, Rouwendal & Marwijk, 2015). **Median: 5.00 Mean: 9.549**

**Full-value property-tax rate per $10,000 (x10)**

The property-tax amount that each homeowner pays for each $10,000 their dwelling is valued at. A higher property-tax rate could indicate higher-property values as in a study conducted on neighbourhoods in New Jersey (Schwartz & Voicu, 2003). **Median: 330.0 Mean: 408.2**

**Pupil-teacher ratio by town (x11)**

The ratio or scale of the student to teacher in each town. A higher number indicates more students per teachers and a lower number indicates fewer students per teachers. A lower ratio for the pupil-teacher ratio is better for both the teachers and students and provides for a more inclusive classroom for student development (Doherty, 1966). The average mean is

**Median: 19.05 Mean: 18.46**

**1000(B - 0:63)^2 where B is the proportion of African Americans by town (x12)**

This value ranges from 0.32 to 396.9 for our data set so that means the value of B is in the range of 0.647888543 to 1.26. A higher value of this predictor variable indicates that more African Americans reside in the area and a lower value in this predictor variable indicates that fewer African Americans reside in the area. **Median: 391.44  Mean:356.67**

**A numeric vector of percentage values of lower status population (x13)**

This variable is a percentage between 0% to 100%. It represents the percentage of the population that is considered lower in socioeconomic status. A higher percentage indicates that more people in the sampled area are poorer, and a lower percentage in these predictor variables means that there are fewer poorer people in the sampled area. **Median :11.36 Mean :12.65**

## y  
## x1 -0.3883046  
## x2 0.3604453  
## x3 -0.4837252  
## x4 0.1752602  
## x5 -0.4273208  
## x6 0.6953599  
## x7 -0.3769546  
## x8 0.2499287  
## x9 -0.3816262  
## x10 -0.4685359  
## x11 -0.5077867  
## x12 0.3334608  
## x13 -0.7376627  
## y 1.0000000

This is the last column of the correlation matrix. It shows how each predictor variable is correlated with the response variable (medv.).

**Model**

**Model Selection**

The data set has been randomly divided into 70% training and 30% validation. Our initial model consisted of all 13 predictor variables. Then, we did backwards elimination of the full model with a significance level of 0.05. This removed 4 predictor variables, x3, x7,x9 and x10 from the full model. Next, we used Akaike’s Information Criterion to remove predictor values which are insignificant. This also removed 2 predictor variables, x7 and x3, from the full model. After, we applied the variation inflation factors on the reduced model. With this, we were able to remove the variables x9 and x10 which were deemed to have high collinearity. This left us with a reduced model with 9 remaining predictor variables. The final model is Y ~ x1 + x2 + x4 + x5 + x6 + x8 + x11 + x12 + x13. To validate this model, we tested the final model against the testing set. We determined the mean squared prediction error and the mean squared residuals. The mean squared prediction error (22.892) was close to the mean squared residuals (23.354) giving a good indication of the predictive ability of the model.

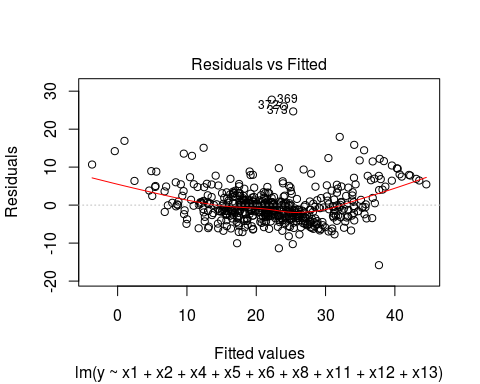
**Summary of Final Model**

model\_reduced <- lm(y ~ x1 + x2 + x4 + x5 + x6 + x8 + x11 + x12 + x13, data = boston\_housing\_train)  
summary(model\_reduced)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x4 + x5 + x6 + x8 + x11 + x12 + x13,   
## data = boston\_housing\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.803 -2.832 -0.625 1.454 27.766   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 29.507997 4.872538 6.056 2.76e-09 \*\*\*  
## x1 -0.061174 0.030377 -2.014 0.044567 \*   
## x2 0.042032 0.013422 3.131 0.001842 \*\*   
## x4 3.029924 0.868349 3.489 0.000527 \*\*\*  
## x5 -16.088513 3.232702 -4.977 8.93e-07 \*\*\*  
## x6 4.149667 0.407685 10.179 < 2e-16 \*\*\*  
## x8 -1.431665 0.188603 -7.591 1.59e-13 \*\*\*  
## x11 -0.838640 0.117342 -7.147 3.19e-12 \*\*\*  
## x12 0.008292 0.002688 3.084 0.002153 \*\*   
## x13 -0.525004 0.048351 -10.858 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.833 on 496 degrees of freedom  
## Multiple R-squared: 0.7288, Adjusted R-squared: 0.7239   
## F-statistic: 148.1 on 9 and 496 DF, p-value: < 2.2e-16

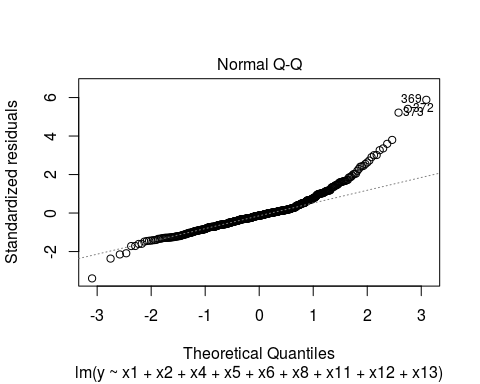
**Model Diagnostics**

**Residuals vs. Fitted Models**



The points in the Residual vs. Fitted plot are randomly dispersed around the horizontal line with some outliers, which suggests heteroscedasticity (non-constant variance).

**Normal Q-Q Plot**

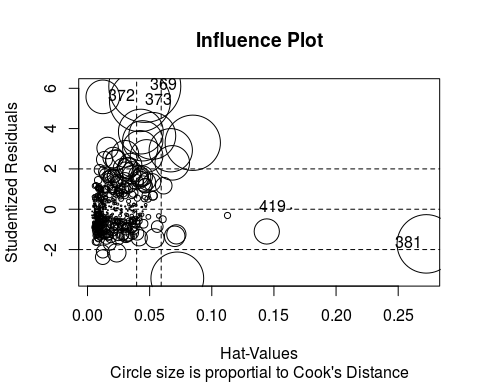


The normal q-q plot of residuals shows that the errors follow a line and are therefore normally distributed. Although most points fit the regression line, there are 3 outliers; these are the 369th, 372nd, 373rd observations.

**Y Outliers**

We have 3 Y outliers - 366th, 369th and 373rd observations based on our analysis of the data set. The t critical value was: 3.925493, these outliers had residuals all greater than that, as indicated by the influence plot below.

**Influence Plot**



According to our analysis, Cook’s distance is the highest at 369th observation (0.1671809)

and 372nd observation (0.0364751).  Using the formula for the 50th percentile : (4/n).

4/506 = 0.007. The cook’s distance for 369,372,373 and 381 are larger than 0.007 so we can conclude that these data points are considered to be a strong influence.

## StudRes Hat CookD  
## 369 6.09311683 0.04608405 1.671809e-01  
## 372 5.57433529 0.01229686 3.647451e-02  
## 373 5.36011650 0.04217171 1.197996e-01  
## 381 -1.71914481 0.27261992 1.103347e-01  
## 419 0.04641776 0.16390887 4.232449e-05

**Variance Influence Factors (VIF)**

VIF

## x1 x2 x4 x5 x6 x8 x11 x12   
## 1.476281 2.119038 1.051879 3.034316 1.774261 3.410535 1.395503 1.302455   
## x13   
## 2.577927

VIFbar

## [1] 2.0158

The variance influence factors for our final model shows us that there is low collinearity between our variables. The range of the factors is 1 < VIF < 5 which means they are moderately correlated. The average variance influence factors is 2.0158.

**Validation**  
MSPE

## [1] 22.89247

MS\_res

## [1] 23.35401

The mean squared prediction error and mean squared residuals are close in value. This indicates that the final model selected is appropriate for predicting housing values in the suburbs of Boston.

**Discussion/Conclusion**

The goal of this study is to create a regression model to predict the housing values in the suburbs of Boston based on the data set. We found that proportion of non-retail business acres per town (x3), proportion of owner occupied units built prior to 1940 (x7), index of accessibility to radial highways (x9) and full -value property tax rate per $10,000 (x10) were not very significant in building a model that accurately predicted housing prices in the suburbs of Boston. These findings impact people looking to invest in real estate for future profits or buying them from the purpose of living in them in the long run. It’s helpful for them to make an educated decision and figure out if the investment is worth their money and if it has the potential to grow in the future based on the predictor variables used. The limitation of this study is the data set is from 1978 (Harrison & Rubinfeld, 1978), which may not be very relevant today. Also, the predictor variables in the data set may not be enough to describe the current demographics. For future research, the predictor variables in the data set can be modified to fit the current demographics or more predictor variables can be added to enhance the accuracy of the regression model and to reduce any bias.

**References**

Case, K. E., & Shiller, R. J. (2004, January 23). Is There a Bubble in the Housing Market? Retrieved from <https://muse.jhu.edu/article/51240/summary>

Doherty, J. (1966). Pupil-Teacher Ratio in Head Start Centers. *Childhood Education,43*(1), 7-8. doi:10.1080/00094056.1966.10728006

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air.*Journal of Environmental Economics and Management,5*(1), 81-102. doi:10.1016/0095-0696(78)90006-2

Kryvobokov, M. (2010). Is it worth identifying service employment (sub)centres when modelling apartment prices? *Journal of Property Research,27*(4), 337-356. doi:10.1080/09599916.2010.517852

Levkovich, O., Rouwendal, J., & Marwijk, R. V. (2015). The effects of highway development on housing prices. *Transportation,43*(2), 379-405. doi:10.1007/s11116-015-9580-7

Osland, L., & Pryce, G. (2012). Housing Prices and Multiple Employment Nodes: Is the Relationship Nonmonotonic? Housing Studies, 27(8), 1182-1208. doi:10.1080/02673037.2012.728571

Schwartz, A. E., Susin, S., & Voicu, I. (2003). Has Falling Crime Driven New York City's Real Estate Boom? *Journal of Housing Research*. Retrieved March 28, 2019.

Thomas, M. E., Moye, R., Henderson, L., & Horton, H. D. (2017). Separate and Unequal: The Impact of Socioeconomic Status, Segregation, and the Great Recession on Racial Disparities in Housing Values. *Sociology of Race and Ethnicity,* *4*(2), 229-244. doi:10.1177/2332649217711457