STAC67 Case Study: Model For Predicting House Prices (Group 6)

Ethan Anada 1003171907, Joseph Dong 1003349272, Subagan Kamaleswaran 1002163712, Shengji Lin 1001704300, Yi Chao Wang (unresponsive)

# Abstract

There are many influential factors that help determine the estimated house price in a particular neighborhood. In order to model the value of Boston’s suburban houses, we have obtained a data-set of 506 observations with 13 factors that could potentially affect a house’s value (Harrison & Rubinfield 1978). These factors include looking into the effects of crime rates, number of rooms per home. Furthermore, we used R in combination with modeling techniques to create a linear model of Boston’s house prices.

# Background and Significance

Housing prices play an important role in deciding when to buy or sell houses. For home buyers, the ability to predict the housing prices can help make a decision on whether to over or under bid on a house. On the other hand, home sellers can use the housing price model to predict the how much money they can make by selling their house. Those that are in the real estate business would also be interested in a model of house prices in order to deliver the best reults for their clients, or in order to make a profit on their own properties.

Due to its significance it is essential to create a model to predict the prices of houses by considering influential factors such as crime rate in a specific town. A study done by Claudio Frischtak and Benjamin R. Mandel called Crime, House Prices, and Inequality: The Effect of UPPs in Rio, looks at the relationship between crime rate and housing prices. The experiment was conducted in low-income communities where the crime rate was reduced, and consequently, the prices of nearby residential real estate increased (Frischtak et al., 2012). Nevertheless, the experiment shows crime rates in an area can affect the prices of houses.

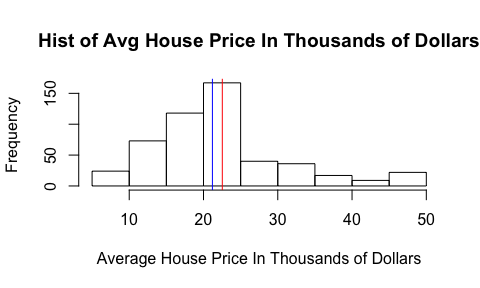
However, besides just perusing the literature and using our own hypotheses on influential parameters, we were also able to use statistical tests to decide on influential parameters and also to eliminate parameters that tracked each other too closely. By applying a combination of these techniques we should be able achieve our goal of creating a model for predicting the price of a house.

# Exploratory Data Analysis

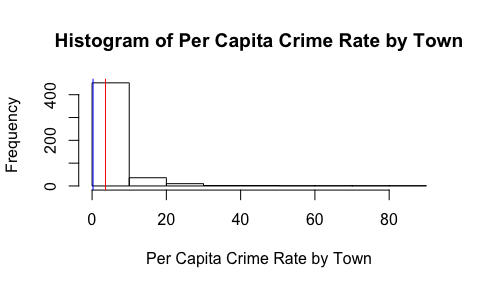
Our data contains information regarding the price of homes in suburbs of Boston. It contains 506 entries, each corresponding to a community. Each entry contains 13 parameters that may or may not affect the price of the median house price of a community, and a corresponding label for the average price of a house in thousands of dollars.

For the following histograms a red line indicates mean, and a blue line indicates median.

## 0. Average House Price in Thousands of Dollars (1978) (y)

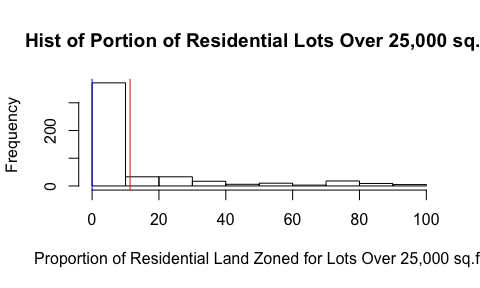
 The response variable, mean house price is 22.53 thousand dollars, median is 21.2 thousand dollars.

## 1. per capita crime rate by town (x1)



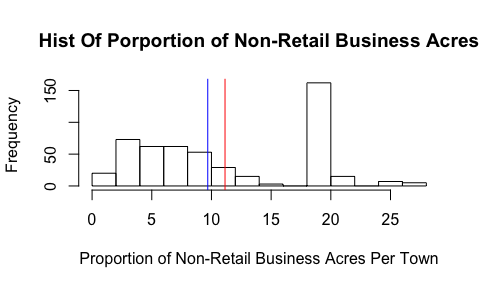
The mean of the crime rate per capita by town is 3.61, and the median is 0.25. Which leads to the data being right-skewed showing that crime tends to be concentrated. We expect that (x1) will be negatively correlated with house price, since in general people find crime undesirable. (Frischtak et al. 2012)

## 2. proportion of residential land zoned for lots over 25,000 sq.ft. (x2)



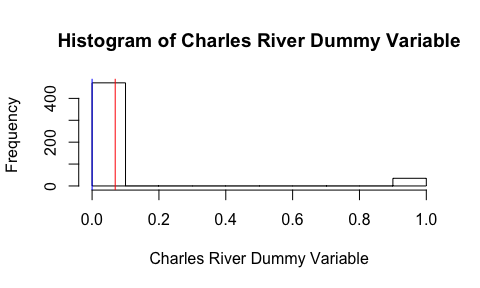
The mean is 11.36, and the median is 0. Therefore, majority of communities don’t have large lot homes but some communities have large lots. Consequently, the parameter is right skewed. Also, larger the lot the more expensive the houses.

## 3. proportion of non-retail business acres per town (x3)



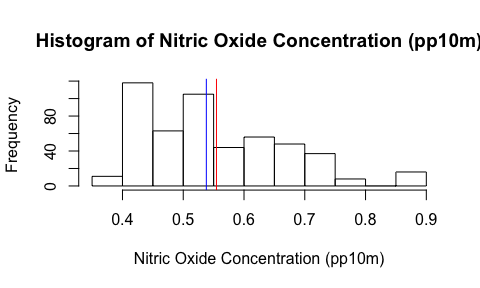
The proportion of the non-retail business acres per towns means that the surrounding area in the community are residences or recreation area. This may lead to calmer surroundings which could positively affect the house value of the surroundings. The mean is 11.14, and median is 9.69

## 4. Charles River dummy variable (x4)



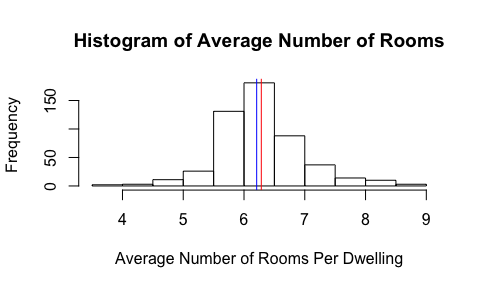
The Charles River dummy variables (x4) is 1 whenever the house is near the river, and 0 otherwise. In general, waterfront property is desirable, so we expect this this to increase house prices. The mean of this parameter is 0.07, and the median is 0.

## 5. nitric oxide concentration (parts per 10 million) (x5)



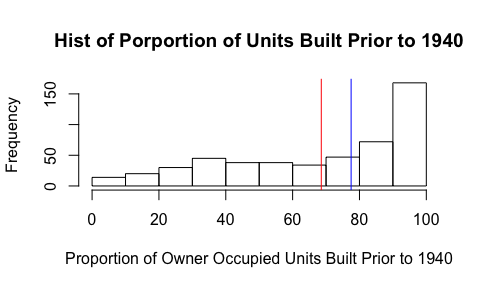
Nitric oxide can be very dangerous. According to J. S. Beckman, and W. H. Koppenol’s paper, “Nitric oxide, superoxide, and peroxynitrite: the good, the bad, and ugly”, it is mentioned the toxicity of nitric oxide can be greatly enchanced when reacting with superoxide (Beckman & Koppenol, 1996). As the ppm levels increase, the value of the variable increases and results in a lower house price. The mean concentration of nitric oxide is 0.5547, and the median is 0.538.

## 6. average number of rooms per dwelling (x6)



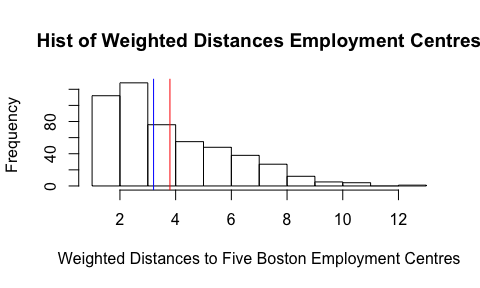
In general the number of rooms in a house is very strongly correlated to the size of the house. So we expect that this parameter is positively correlated to the price of the house. The mean rooms per house is 6.23, and the median of 6.20.

## 7. proportion of owner occupied units built prior to 1940 (x7)



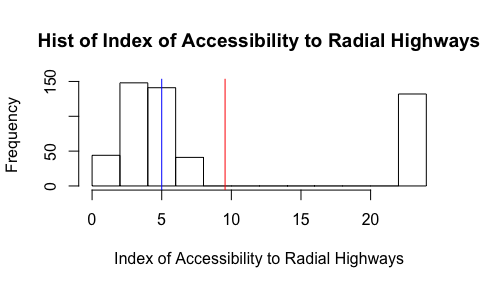
There are pros and cons to owning an old house but mostly it is undesirable to own an old house due to the age of its components which will require more maintenance, so we believe that this parameter is negatively correlated to house price. The mean proportion of old houses built prior to 1940 is 68.57, and the median is 77.5.

## 8. weighted distances to five Boston employment centres (x8)



In general it is desirable to live close to your place of employment, since it will reduce commute time, which in turn increases your leisure time. So we expect that this parameter is negatively correlated to house price. The mean is 3.98, and the median of 3.20.

## 9. index of accessibility to radial highways (x9)



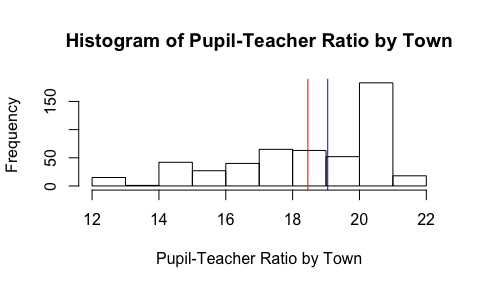
The index of accessibility to radial highways is a value that measure the farness of a house to a radial highway. This parameter is high when a house is far from a radial highway, and low when it is close. We expect that the convenience of being close to a radial highway will increase the price of a house. The mean is 9.549, and the median is 5.

## 10. full-value property-tax rate per 10,000 (x10)



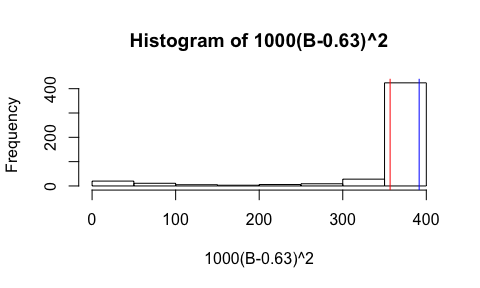
The mean of full-value property-tax rate per 10,000 is 408.2 and the median of 330.0. We expect that property tax will be strongly correlated to house price, since in real life property tax rates are often calculated as a percentage of the value of the property.

## 11. pupil-teacher ratio by town (x11)



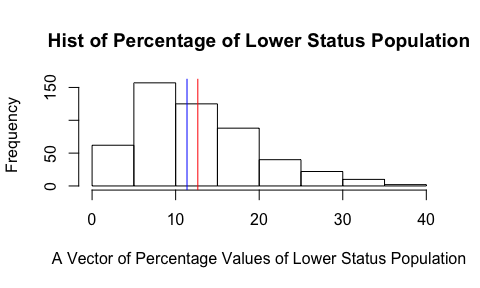
In general a lower pupil to teacher ratio indicates a well funded school, so we think that this parameter will be negatively correlated to the price of a house. The mean amount of students that each teacher has to teach in this dataset is 18.46, and the median is 19.05.

## 12. 1000(B-0.63)^2 where B is the proportion of African Americans by town (x12)



Note that the function 1000(B-0.63)^2 in the range of B in [0, 1] is high when B is near 0 and middle value when B is near 1, but is low when B is near 0.63. This can be interpreted as (x12) being higher when a community is homogenous in ethnicity. We theorize that homogenous communities would have higher house prices due to reduced racial tensions making a safer neighborhood. The mean is 356.674, and the median is 391.44.

## 13. a numeric vector of percentage values of lower status population (x13).



Almost by definition impoverished people cannot afford expensive housing, so it is expected that communities with more impoverished people will have lower house prices. The mean of this parameter is 12.65, and the median is 11.36.

# Model

In order to begin the model selection process, we decided to use the stepAIC algorithm provided by an r library in order to automatically remove some variables from a full model ().

The results of the algorithm recommended that we remove and from our model, we decided do so since it lowered our AIC and also improved our . We also decided to transform by natural log, since we found that it also increases our .

We narrowed down parameters for our model, by doing a cycle of checking for multicollinearity for each parameter by computing variance inflation factor (VIF). This was done to remove parameters with high VIF, then we used hypothesis testing to check if the parameters were significant. The process was repeated until we obtained low VIFs, and all parameters were significant.

Computing VIF values shows that has high VIF which is indicative of serious multicollinearity, so we decide to remove from our model.

Looking at the significance of parameters by looking at the summary of the new model, we see that the coefficient of is probably plausibly 0, so we decide to remove to simplify our model.

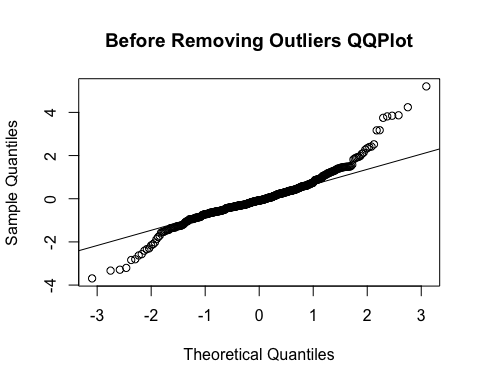
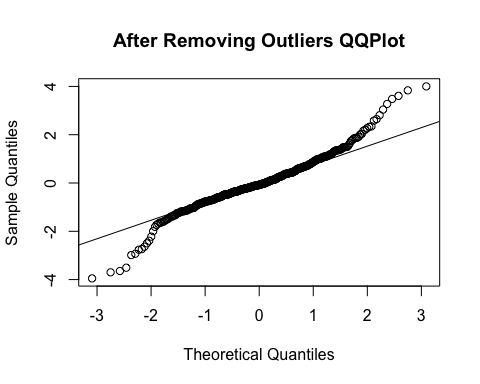
When our algroithm terminates after a few iterations, we arrive at a final model which is

##   
## Call:  
## lm(formula = log(y) ~ x1 + x4 + x6 + x8 + x11 + x12 + x13, data = housingData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73467 -0.10434 -0.01549 0.08468 1.02176   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.2404102 0.1703437 19.023 < 2e-16 \*\*\*  
## x1 -0.0086059 0.0012377 -6.953 1.13e-11 \*\*\*  
## x4 0.1042281 0.0358575 2.907 0.003815 \*\*   
## x6 0.1150341 0.0166811 6.896 1.63e-11 \*\*\*  
## x8 -0.0197258 0.0051015 -3.867 0.000125 \*\*\*  
## x11 -0.0325010 0.0045973 -7.070 5.28e-12 \*\*\*  
## x12 0.0004629 0.0001098 4.215 2.97e-05 \*\*\*  
## x13 -0.0312488 0.0019401 -16.107 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2002 on 498 degrees of freedom  
## Multiple R-squared: 0.7635, Adjusted R-squared: 0.7601   
## F-statistic: 229.6 on 7 and 498 DF, p-value: < 2.2e-16

To ensure the legitimacy of our model, we decided to check if the AIC of our model has improved from the full model, and also use a validation strategy to detect overfitting. The final model AIC was , it was significantly lower than the AIC of the full model .

The validation strategy used was partitioning the data-set randomly into 75% training data and 25% test data. The validation result were and . Both of the values were similar to our final model’s and . Therefore, our final model is a valid model for the data-set

Finally, we examined outliers, leverage points and influential points of the data-set. We used studentized deleted residuals to determine outliers and removed them to normalize the errors.

  Then we tried to find the leverage point and the influential points in our data-set, using the Cooke Distance and outlying X Observation Diagnostics.

The influential Points were the rows of the data-set, and there were 66 outlying X variables in our dataset using looser guideline and 0 outlying X variables in our dataset using a more strict guideline. Both of these sets of points were found after the outlying Y variable was removed from the data-set. We found no good reason to remove these points our training data.

# Discussion and Conclusions

The goal of this study was to create a model for house prices in Boston’s suburbs. Throughout this report, we used different statistical methods along with R to create a linear model for the housing prices. Through creating the model, we discovered 7 of the 13 parameters were significant to the model. The 7 important factors consisted of crime rate by town, Charles River, average rooms per house, weighted distances to 5 Boston employment centers, pupil-teacher ratio, percentage value of lower status population and finally, 1000(B-0.63)^2 where B is the proportion of African Americans by town. These findings are significant to the field as it raises awareness to what factors people look at when choosing a home to live in. As a result, realtors would be able to use the model to predict the price of a suitable home for their clients. Moreover, by understanding the various factors that can influence the pricing of houses, more studies can be conducted to understand why people prefer a certain attribute over another. This data can also be used to market certain houses and make it appealing to a certain demographic. This would benefit contractors as they would be able to strategically find locations to buy land and build housing in. The limitations of our model is that it is made from data collected in 1978. As a result, the model may be outdated and the parameters that affected the prices of a home in 1978 may not be as significant in 2019. Nevertheless, this model does not take market conditions into consideration. Therefore, the model we create whether it is with new or old data would only be accurate for a short amount of time.

# References

Beckman,, J. S., & Koppenol, W. H. (1996). Nitric oxide, superoxide, and peroxynitrite: The good, the bad, and ugly. American Journal of Physiology Cell Physiology, 271(5). Retrieved March 30, 2018, from <https://doi.org/10.1152/ajpcell.1996.271.5.C1424>.

Frischtak, C., & Mandel, B. R. (2012). Crime, House Prices, and Inequality: The Effect of UPPs in Rio (542nd ed.). Federal Reserve Bank of New York Staff Reports. Retrieved March 25, 2019, from <https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr542.pdf>.

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>