STAC67 Case Study: A model for predicting housing values in Boston

Revon Villava-Rayen (1002987501)

Sizhe Gao ()

Alexander Nguyen ()

Gopi Santhiran (1001371534)

Yu Heng Wu (1003475330)

Group 13

March 20 2019

# Abstract

The housing market of the city of Boston is a volatile market, as with any world reknowned city. There are many attempts to create a formula to predict housing prices; using predictor variables such as per capita crime rate by town, or average number of rooms per dwelling. Using R statistical analysis, this study aims to examine and determine which model best represents the correlation between housing prices and these influential factors.

# Background and Significance

Boston is the captial and largest city in Massachusetts. Founded in 1630 Boston is one of the oldest cities in America and the city played an important role in American History. As Boston is a world reknowned city, the housing market in Boston is expensive, with finding and owning property is an important financial goal for many that live within or near Boston.

As Boston currently faces a “growing housing challenge … likely because younger residents are marrying later … many of them are overloaded with college debt” (Bluestone). It is clear that buying a House is an expensive financial burden that less and less people are able to afford. Especially when a house must be able to fit the needs of the buyer as a living space.

We propose this model with the given predictor variables to accurately predict housing values within Boston. By being able to purchase a house at an optimal time, it minimizes the financial burden of the purchase and also allows you to maximize any capital gains that may result from the changing value of the purchased house. Through this process, it is possible to maximize the value of your purchase and minimize the cost of buying a house which leads to more people being able to afford housing in the city of Boston.

# Exploratory Data Analysis

This Data set contains 506 observations on 13 predictors variables, which are per capita crime rate by town, proportion of residential land zoned for lots over 25,000 sq. ft., proportion of non-retail business acres per town Charles River dummy variable, nitric oxide concentration (parts per 10 million), average number of rooms per dwelling proportion of owner occupied units built prior to 1940, weighted distances to five Boston employment centres index of accessibility to radial highways, full-value property-tax rate per 10,000, pupil-teacher ratio by town, 1000(B â 0.63)^2 where B is the proportion of African Americans by town, a numeric vector of percentage values of lower status population.

Because of the rather long names, we will assign shorter names in capitals listed below:

### CRIME - per capita crime rate by town

A crime rate describes the number of crimes per 100,000 of the population. The mean and mode of the crime rate is 3.613 and 0.01501 respectively

### ZL - proportion of residential land zoned for lots over 25,000 sq. ft.

The land set aside for residential buildings that is over 25,000 sq ft.The mean and mode of the residential land zoned for lots over 25,000 sq. ft. is 11.363 and 0 and respectively

### NR\_PROP - proportion of non-retail business acres per town

The mean and mode of the non-retail buisness acres per town is 11.136 and 18.1 respectively

### CHR\_V - Charles River dummy variable

The Charles River is a 129 km long river in eastern Massachueetts.The mean and mode of the Charles River dummy variable is 0.069 and 0 respectively

### NOX - nitric oxide concentration (parts per 10 million)

The mean and mode of nitric oxide concentration is 0.554 and 0.538 respectively

### ROOM - average number of rooms per dwelling

The mean and mode of number of rooms per dwelling is 6.284 and 5.713 respectively

### AGE - proportion of owner occupied units built prior to 1940

The mean and mode of the proporition of owner occuiped units built prior to 1940 is 68.574 and 100 respectively

### DIS - weighted distances to five Boston employment centres

The mean and mode weighted distances to five Boston employment centres is 3.795 and 3.4952 respectively

### HWY - index of accessibility to radial highways

The mean and mode of index of accessibility to radial highways is 9.549 and 24 respectively

### TAX - full-value property-tax rate per 10,000

The mean and mode of full-value propert-tax per rate 10,000 is 408.237 and 666 respectively

### PT\_RATIO - pupil-teacher ratio by town

The mean and mode of pupil-teacher ratio by town is 18.455 and 20.2 respectively

### B - 1000(B â 0.63)^2 where B is the proportion of African Americans by town

The mean and mode of the proportion of African Americans by town is 356.674 and 396.9 respectively

### L\_PER - a numeric vector of percentage values of lower status population

The mean and mode the proportion of the numberic vector of percentage values of lower status population is 12.653 and 8.05 respectively

# Model

# Co-Lineartiy

## CRIME ZL NR\_PROP CHR\_V NOX ROOM AGE DIS HWY  
## CRIME 1.000 -0.200 0.407 -0.056 0.421 -0.219 0.353 -0.380 0.626  
## ZL -0.200 1.000 -0.534 -0.043 -0.517 0.312 -0.570 0.664 -0.312  
## NR\_PROP 0.407 -0.534 1.000 0.063 0.764 -0.392 0.645 -0.708 0.595  
## CHR\_V -0.056 -0.043 0.063 1.000 0.091 0.091 0.087 -0.099 -0.007  
## NOX 0.421 -0.517 0.764 0.091 1.000 -0.302 0.731 -0.769 0.611  
## ROOM -0.219 0.312 -0.392 0.091 -0.302 1.000 -0.240 0.205 -0.210  
## AGE 0.353 -0.570 0.645 0.087 0.731 -0.240 1.000 -0.748 0.456  
## DIS -0.380 0.664 -0.708 -0.099 -0.769 0.205 -0.748 1.000 -0.495  
## HWY 0.626 -0.312 0.595 -0.007 0.611 -0.210 0.456 -0.495 1.000  
## TAX 0.583 -0.315 0.721 -0.036 0.668 -0.292 0.506 -0.534 0.910  
## PT\_RATIO 0.290 -0.392 0.383 -0.122 0.189 -0.356 0.262 -0.232 0.465  
## B -0.385 0.176 -0.357 0.049 -0.380 0.128 -0.274 0.292 -0.444  
## L\_PER 0.456 -0.413 0.604 -0.054 0.591 -0.614 0.602 -0.497 0.489  
## MEDV -0.388 0.360 -0.484 0.175 -0.427 0.695 -0.377 0.250 -0.382  
## TAX PT\_RATIO B L\_PER MEDV  
## CRIME 0.583 0.290 -0.385 0.456 -0.388  
## ZL -0.315 -0.392 0.176 -0.413 0.360  
## NR\_PROP 0.721 0.383 -0.357 0.604 -0.484  
## CHR\_V -0.036 -0.122 0.049 -0.054 0.175  
## NOX 0.668 0.189 -0.380 0.591 -0.427  
## ROOM -0.292 -0.356 0.128 -0.614 0.695  
## AGE 0.506 0.262 -0.274 0.602 -0.377  
## DIS -0.534 -0.232 0.292 -0.497 0.250  
## HWY 0.910 0.465 -0.444 0.489 -0.382  
## TAX 1.000 0.461 -0.442 0.544 -0.469  
## PT\_RATIO 0.461 1.000 -0.177 0.374 -0.508  
## B -0.442 -0.177 1.000 -0.366 0.333  
## L\_PER 0.544 0.374 -0.366 1.000 -0.738  
## MEDV -0.469 -0.508 0.333 -0.738 1.000

The R code shown below displays the step forward and backwards process as well as the final model we will be using.

##   
## Call:  
## lm(formula = MEDV ~ CRIME + CHR\_V + NOX + ROOM + AGE + DIS +   
## TAX + PT\_RATIO + B + L\_PER, data = mod\_housing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.7713 -1.7942 -0.3075 1.6178 12.0434   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -13.977535 4.703496 -2.972 0.00317 \*\*   
## CRIME 1.339359 0.439198 3.050 0.00247 \*\*   
## CHR\_V 0.910221 0.630738 1.443 0.14990   
## NOX -8.527803 3.489057 -2.444 0.01502 \*   
## ROOM 9.376350 0.368828 25.422 < 2e-16 \*\*\*  
## AGE -0.046949 0.009518 -4.933 1.27e-06 \*\*\*  
## DIS -0.852483 0.126584 -6.735 6.95e-11 \*\*\*  
## TAX -0.012875 0.002683 -4.798 2.40e-06 \*\*\*  
## PT\_RATIO -0.661513 0.088148 -7.505 5.35e-13 \*\*\*  
## B 0.015816 0.004780 3.309 0.00104 \*\*   
## L\_PER -0.087802 0.047735 -1.839 0.06673 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.064 on 343 degrees of freedom  
## Multiple R-squared: 0.8718, Adjusted R-squared: 0.8681   
## F-statistic: 233.4 on 10 and 343 DF, p-value: < 2.2e-16

## Start: AIC=805.64  
## MEDV ~ CRIME + ZL + NR\_PROP + CHR\_V + NOX + ROOM + AGE + DIS +   
## HWY + TAX + PT\_RATIO + B + L\_PER  
##   
## Df Sum of Sq RSS AIC  
## - NR\_PROP 1 8.6 3193.0 804.59  
## - CHR\_V 1 14.5 3198.8 805.24  
## - HWY 1 15.2 3199.6 805.32  
## <none> 3184.4 805.64  
## - ZL 1 22.8 3207.1 806.16  
## - L\_PER 1 36.6 3221.0 807.69  
## - NOX 1 61.2 3245.6 810.38  
## - CRIME 1 81.0 3265.4 812.53  
## - B 1 102.2 3286.5 814.81  
## - AGE 1 206.2 3390.5 825.84  
## - TAX 1 245.7 3430.1 829.95  
## - DIS 1 387.3 3571.6 844.26  
## - PT\_RATIO 1 418.8 3603.2 847.38  
## - ROOM 1 5507.9 8692.2 1159.11  
##   
## Step: AIC=804.59  
## MEDV ~ CRIME + ZL + CHR\_V + NOX + ROOM + AGE + DIS + HWY + TAX +   
## PT\_RATIO + B + L\_PER  
##   
## Df Sum of Sq RSS AIC  
## - HWY 1 12.2 3205.1 803.94  
## - CHR\_V 1 17.2 3210.1 804.49  
## <none> 3193.0 804.59  
## - ZL 1 21.1 3214.0 804.92  
## + NR\_PROP 1 8.6 3184.4 805.64  
## - L\_PER 1 33.7 3226.7 806.31  
## - NOX 1 54.4 3247.3 808.57  
## - CRIME 1 86.8 3279.8 812.09  
## - B 1 102.2 3295.2 813.75  
## - AGE 1 209.7 3402.6 825.10  
## - TAX 1 239.2 3432.1 828.16  
## - PT\_RATIO 1 410.3 3603.3 845.39  
## - DIS 1 430.6 3623.5 847.37  
## - ROOM 1 5510.3 8703.2 1157.56  
##   
## Step: AIC=803.94  
## MEDV ~ CRIME + ZL + CHR\_V + NOX + ROOM + AGE + DIS + TAX + PT\_RATIO +   
## B + L\_PER  
##   
## Df Sum of Sq RSS AIC  
## - ZL 1 15.3 3220.4 803.62  
## <none> 3205.1 803.94  
## - CHR\_V 1 19.9 3225.0 804.13  
## + HWY 1 12.2 3193.0 804.59  
## + NR\_PROP 1 5.6 3199.6 805.32  
## - L\_PER 1 35.5 3240.6 805.83  
## - NOX 1 53.0 3258.2 807.75  
## - CRIME 1 89.2 3294.3 811.65  
## - B 1 103.7 3308.9 813.21  
## - AGE 1 207.2 3412.3 824.11  
## - TAX 1 227.0 3432.1 826.16  
## - PT\_RATIO 1 412.3 3617.5 844.78  
## - DIS 1 419.3 3624.4 845.46  
## - ROOM 1 5695.6 8900.8 1163.51  
##   
## Step: AIC=803.62  
## MEDV ~ CRIME + CHR\_V + NOX + ROOM + AGE + DIS + TAX + PT\_RATIO +   
## B + L\_PER  
##   
## Df Sum of Sq RSS AIC  
## <none> 3220.4 803.62  
## - CHR\_V 1 19.6 3240.0 803.76  
## + ZL 1 15.3 3205.1 803.94  
## + HWY 1 6.4 3214.0 804.92  
## + NR\_PROP 1 5.0 3215.4 805.07  
## - L\_PER 1 31.8 3252.2 805.09  
## - NOX 1 56.1 3276.5 807.73  
## - CRIME 1 87.3 3307.7 811.09  
## - B 1 102.8 3323.2 812.74  
## - TAX 1 216.1 3436.5 824.61  
## - AGE 1 228.5 3448.9 825.88  
## - DIS 1 425.8 3646.2 845.58  
## - PT\_RATIO 1 528.8 3749.2 855.44  
## - ROOM 1 6067.9 9288.3 1176.59

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## MEDV ~ CRIME + ZL + NR\_PROP + CHR\_V + NOX + ROOM + AGE + DIS +   
## HWY + TAX + PT\_RATIO + B + L\_PER  
##   
## Final Model:  
## MEDV ~ CRIME + CHR\_V + NOX + ROOM + AGE + DIS + TAX + PT\_RATIO +   
## B + L\_PER  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 340 3184.365 805.6358  
## 2 - NR\_PROP 1 8.590616 341 3192.956 804.5895  
## 3 - HWY 1 12.182496 342 3205.139 803.9376  
## 4 - ZL 1 15.259703 343 3220.398 803.6190

## [1] 803.619

## [1] 832.5803

# Citations

Bluestone, B. & Huessy, J. (2017). The Greater Boston Housing Report Card 2017 Ideas from the Urban Core - Responsive Development - as a Model for Regional Growth, 68