STAC67 Case Study: Describing the behavior of Housing Values

Group 15: Yiu-meng Choi 1003488923, Dwight Sampson 1002094627, Yuting Wang 1003340333

April 1, 2019

## 

## Background and Significance

The housing market has for a long time been an important topic of discussion among homebuyers and real estate owners. According to RBC’s annual homeownership poll, 50 per cent of those aged between 18 and 34 claimed they were likely to purchase within the next year (Powell, 2018). However, there is still the obstacle of being able to afford a house for potential buyers. It is not much of a secret to many that house pricing has gone up drastically from many years ago. In fact, in the United States, 41 states have seen median housing price rise from 2007 to 2017 (Digg, 2018). Currently, the median home value in the United States is $226,300, which is up 7.2 per cent from the previous year (“United States Home Prices & Values”, 2019). There are several factors which may be attributed to house pricing. Here we propose a model to accurately predict the median value of homes in Boston suburbs. If the median pricing of houses can be predicted, then that would give helpful insights to potential buyers, and for real estate investors. It may also be used to help understand drastic shifts in house pricing, known as housing bubbles, which can damage the economy (Kenton, 2018).

# 

# Exploratory Data Analysis

## Per capita crime rate by town is the amount of crime that happens for each person in town.

This is important as areas that higher crime rates tend to be viewed as undesirable places to live. Proportion of residential land zoned for lots over 25 000 sq.ft(PRL) This is the fraction of residential land that has a lot size of over 25000 sq.ft. The average size of an American home in 2015 was 2687 sq.ft (Geffner), so it stand to reason that any lot significantly larger than 2687 sq.ft is massive. More land a property has the more value the property obtains.

## Proportion of non-retail business acres per town (PNB)

This is the amount of land that is currently occupied by businesses that are not in retail. This may be important as one considerations home buyers often make when purchasing a home is whether a home is close to amenities like groceries (Williams, 2014).

## Charles River dummy variable (CRD)

This is an indicator variable determining whether or not a home is near the Charles river. Zero indicates the home is not near the Charles River, while one indicates otherwise.

## Nitric oxide concentration (parts per 10 million) (NOC)

The higher concentration of nitric oxide represents more air pollution in the area. This is important as high air pollution levels can cause health problems.

## Average Number of rooms per dwelling (NRB)

The average number of rooms may affect the home value since adding rooms can meet more demands of people.

## Proportion of owner occupied units built prior to 1940 (POO)

The proportion of owner-occupied units built prior to 1940 represents the percentage of the old houses that built before 1940. This is an essential measurement since the houses’ age may influence their price. (Hardle and Simar, 2007)

## Weighted distances to five Boston employment centres (WBE)

The weighted distances to five Boston employment centres measure the distance between the living places and the five main working places in Boston. This may be important as most people like to live close to their place of work.

## Index of accessibility to radial highways (RAD)

Index of accessibility to radial highways is the normal of radial highways that are close are in the town. Radial highways are highways that lead to urban centres.

## Full-value property-tax rate per 10000(FPT)

Full-value property-tax rate per 10000 is the amount of tax you have to pay per $10000 of house worth. This might have an effect on housing value because there may be more accessible public services in the area

## Pupil-teacher ratio by town (PTR)

The number of students per teacher in the town. This is important to measure because education can be a factor into housing value

## 1000(B - 0.63)2 where B is the proportion of African Americans by town (PAA)

The proportion of african americans by town is the number of african americans over the the number of people in the town meaning that the higher the proportion of african americans the lower the housing value

## A numeric vector of percentage values of lower status population (PLP)

The percentage of lower status population the percentage of population that is poor in the area

#determine any correlation between variables  
varCor0 <- cor(bostonhousing)  
varCor0

## PCC PRL PNB CRD NOC  
## PCC 1.00000000 -0.20046922 0.40658341 -0.055891582 0.42097171  
## PRL -0.20046922 1.00000000 -0.53382819 -0.042696719 -0.51660371  
## PNB 0.40658341 -0.53382819 1.00000000 0.062938027 0.76365145  
## CRD -0.05589158 -0.04269672 0.06293803 1.000000000 0.09120281  
## NOC 0.42097171 -0.51660371 0.76365145 0.091202807 1.00000000  
## NRB -0.21924670 0.31199059 -0.39167585 0.091251225 -0.30218819  
## POO 0.35273425 -0.56953734 0.64477851 0.086517774 0.73147010  
## WBE -0.37967009 0.66440822 -0.70802699 -0.099175780 -0.76923011  
## RAD 0.62550515 -0.31194783 0.59512927 -0.007368241 0.61144056  
## FPT 0.58276431 -0.31456332 0.72076018 -0.035586518 0.66802320  
## PTR 0.28994558 -0.39167855 0.38324756 -0.121515174 0.18893268  
## PAA -0.38506394 0.17552032 -0.35697654 0.048788485 -0.38005064  
## PLP 0.45562148 -0.41299457 0.60379972 -0.053929298 0.59087892  
## Y -0.38830461 0.36044534 -0.48372516 0.175260177 -0.42732077  
## NRB POO WBE RAD FPT  
## PCC -0.21924670 0.35273425 -0.37967009 0.625505145 0.58276431  
## PRL 0.31199059 -0.56953734 0.66440822 -0.311947826 -0.31456332  
## PNB -0.39167585 0.64477851 -0.70802699 0.595129275 0.72076018  
## CRD 0.09125123 0.08651777 -0.09917578 -0.007368241 -0.03558652  
## NOC -0.30218819 0.73147010 -0.76923011 0.611440563 0.66802320  
## NRB 1.00000000 -0.24026493 0.20524621 -0.209846668 -0.29204783  
## POO -0.24026493 1.00000000 -0.74788054 0.456022452 0.50645559  
## WBE 0.20524621 -0.74788054 1.00000000 -0.494587930 -0.53443158  
## RAD -0.20984667 0.45602245 -0.49458793 1.000000000 0.91022819  
## FPT -0.29204783 0.50645559 -0.53443158 0.910228189 1.00000000  
## PTR -0.35550149 0.26151501 -0.23247054 0.464741179 0.46085304  
## PAA 0.12806864 -0.27353398 0.29151167 -0.444412816 -0.44180801  
## PLP -0.61380827 0.60233853 -0.49699583 0.488676335 0.54399341  
## Y 0.69535995 -0.37695457 0.24992873 -0.381626231 -0.46853593  
## PTR PAA PLP Y  
## PCC 0.2899456 -0.38506394 0.4556215 -0.3883046  
## PRL -0.3916785 0.17552032 -0.4129946 0.3604453  
## PNB 0.3832476 -0.35697654 0.6037997 -0.4837252  
## CRD -0.1215152 0.04878848 -0.0539293 0.1752602  
## NOC 0.1889327 -0.38005064 0.5908789 -0.4273208  
## NRB -0.3555015 0.12806864 -0.6138083 0.6953599  
## POO 0.2615150 -0.27353398 0.6023385 -0.3769546  
## WBE -0.2324705 0.29151167 -0.4969958 0.2499287  
## RAD 0.4647412 -0.44441282 0.4886763 -0.3816262  
## FPT 0.4608530 -0.44180801 0.5439934 -0.4685359  
## PTR 1.0000000 -0.17738330 0.3740443 -0.5077867  
## PAA -0.1773833 1.00000000 -0.3660869 0.3334608  
## PLP 0.3740443 -0.36608690 1.0000000 -0.7376627  
## Y -0.5077867 0.33346082 -0.7376627 1.0000000

# double braket means extremely strong correlation  
  
# moderate correlation between (PCC and RAD), (PRL and WBE), (PNB and PLP), (PNB and POO), (NOC and RAd),  
# (NOC and FPT), (NRB and PLP)  
  
#strong correlation between (PNB and FPT), (PNB and WBE), (PNB and NOC), (NOC and POO), (NOC and WBE), (POO and WBE),  
# ((RAD and FPT))

# Model Building

The model has a lower residual vs leverage spread when the Y value is logged

##   
## Call:  
## lm(formula = log(Y) ~ PRL + CRD + NOC + NRB + POO + WBE + RAD +   
## PTR + PAA + PLP, data = bostonhousingsample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.81719 -0.10093 -0.01290 0.09542 0.80457   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.241e+00 3.033e-01 10.686 < 2e-16 \*\*\*  
## PRL 3.315e-04 8.392e-04 0.395 0.69323   
## CRD 1.314e-01 5.013e-02 2.622 0.00930 \*\*   
## NOC -5.880e-01 1.883e-01 -3.123 0.00201 \*\*   
## NRB 1.623e-01 2.575e-02 6.301 1.39e-09 \*\*\*  
## POO -8.370e-04 7.558e-04 -1.107 0.26924   
## WBE -5.108e-02 1.131e-02 -4.518 9.75e-06 \*\*\*  
## RAD 9.641e-05 2.086e-03 0.046 0.96317   
## PTR -3.731e-02 7.353e-03 -5.074 7.78e-07 \*\*\*  
## PAA 9.986e-04 1.668e-04 5.988 7.62e-09 \*\*\*  
## PLP -2.540e-02 3.075e-03 -8.259 9.65e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1949 on 242 degrees of freedom  
## Multiple R-squared: 0.7737, Adjusted R-squared: 0.7643   
## F-statistic: 82.73 on 10 and 242 DF, p-value: < 2.2e-16

Loss in R2 and R2 adj is negligable, thus we must use other citeriea to determine the best model. According to C criterion and AIC criterion the model [1]: (log(Y) ~ PRL + CRD + NOC + NRB + POO + WBE + RAD + PTR + PAA + PLP) is the best, followed by model [3]: (log(Y) ~ CRD + NOC + NRB + POO + WBE + RAD + PTR + PAA + PLP).

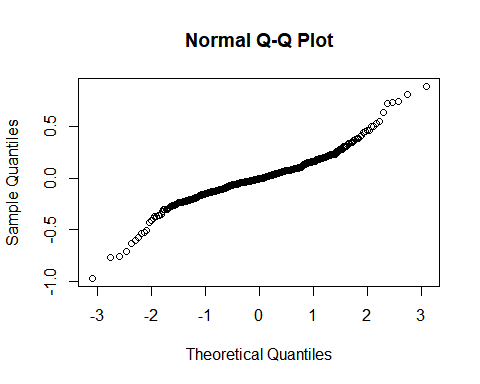
# Model Validation

## [1] 0.05031588

## [1] 0.03798076

this model is valid because MSPE is about equal to MS\_res

# Model Diagnostic



The model created follows the normal distribution between the middle quantiles while having opposed heavy tails.

## Outlying Y observations

### Statistical test

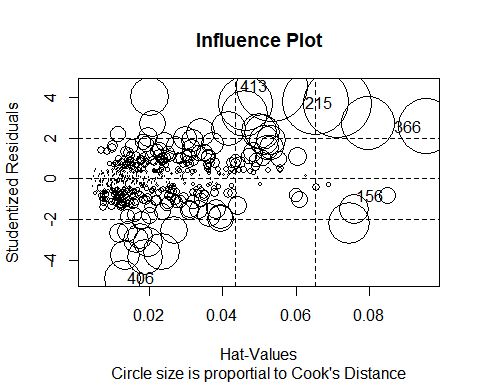
## rstudent unadjusted p-value Bonferonni p  
## 406 -4.903406 1.2801e-06 0.00064771  
## 413 4.557486 6.5329e-06 0.00330560  
## 372 4.062949 5.6355e-05 0.02851600

According to the test there exist 3 outlying Y observations in the dataset which is the 406,412 and 372 observations

## Outlying X observations

There is 1 observation which is observation 366

## Influence



## StudRes Hat CookD  
## 156 -0.8379476 0.08499490 0.005932964  
## 215 3.7128605 0.07132046 0.093820528  
## 366 2.5660303 0.09546397 0.062470113  
## 406 -4.9034065 0.01272225 0.026913294  
## 413 4.5574857 0.05345778 0.102546363

According to cooks distance the only outliers would be observations 413,215,366,156,406. The largest cook’s distance is observation 413 where it had a cooks distance of 0.1. The DFFITS value of this observation is 1.0830805902.

## 

When compare the residuals against the fitted values we can see that there are only 3 extreme outliers which are observations 413,372,406 and that the point are centered around a line at 0;

## Plots

# 

# 

# References

Digg. (2018, December 4). How Housing Prices Have Changed 10 Years After The Real Estate Bubble. Retrieved March 29, 2019, from <http://digg.com/2018/housing-price-change-2007-to-2017>

Hardle, W. Simar, L. (2007). Applied Multivariate Statistical Analysis. (2nd ed.). Retrieved from <https://books.google.ca/books?id=QWKOZIZdnhMC&pg=PA33&lpg=PA33&dq=Weighted+distances+to+five+Boston+employment+centres&source=bl&ots=7kx_UcfKSb&sig=ACfU3U33zYIRikw9I5tb8N_DF6uQFmkDlA&hl=en&sa=X&ved=2ahUKEwiw2dSCjZ_hAhVk64MKHXCTCOEQ6AEwAnoECAkQAQ#v=onepage&q=Weighted%20distances%20to%20five%20Boston%20employment%20centres&f=false>

Kenton, W. (2019, March 27). Housing Bubble. Retrieved March 29, 2019, from <https://www.investopedia.com/terms/h/housing_bubble.asp>

Powell, N. (2018, November 05). Millennial housing crisis? Turns out, it’s real and worse than you thought. Retrieved March 29, 2019, from <https://business.financialpost.com/real-estate/millennial-housing-crisis-turns-out-its-real-and-worse-than-you-thought>

Ristorcelli, S. (2017, March 08). The Dos and Don’ts of Touring a Model Home. Retrieved March 20, 2019, from <https://www.newhomesource.com/guide/articles/how-much-square-footage-fits-your-family>

Williams, G. (2014, July 16). When House Hunting, How to Assess a Neighborhood. Retrieved March 20, 2019, from <https://loans.usnews.com/when-house-hunting-how-to-assess-a-neighborhood>

Zillow, Inc. (n.d.). United States Home Prices & Home Values. Retrieved March 29, 2019, from <https://www.zillow.com/home-values/>