Statistical Analysis to identify key parameters affecting housing prices in Boston

Team Members:

Aditya Aghi

Arjun Berry

Jayanti Trivedi

Tsou-Wei Chu

Table of Contents

1.Introduction …….3

1.1 Project Motivation ……..……..……3

1.2 Variable Description ……..……..……..……3

2.Analysis ……...4

2.1 Hypothesis Testing ……..……...4

2.2 Data Displays ……..……...5

2.3 Model Building Process ……..……...6

2.4 Interpretation of the final model ……..……..7

2.5 Recommendations ……..……8

2.6 Limitation ……...……8

3.Conclusion ……..9

Appendix ……..……...9

References ……..…….. 14

1. INTRODUCTION

1.1 Project Motivation

Our team conducted statistical analysis for identifying key parameters affecting housing prices in Boston. Since boston is amongst the top 30 economically powerful cities in the world, it has plenty of job opportunities. Moreover research indicates that approximately 200,000 new jobs are expected in the city during 2015-2018.

In this plethora of growth opportunities, it is also possible that some of the recent graduates from our MSBA batch get a chance to move there for work purpose. High demand /heavy movement of population to the city for work will certainly give an upward push to the housing prices.

Considering the practical significance of the scenario, our team decided to take forward this project. For moving forward, we took data set from StatLib library - Carnegie Mellon University, comprising 506 rows and 14 columns. Out of 14 variables, we ended up with one response variable and few selected predictors in our final model.

Below are the listed variables which we thought were significant, and later tested through multiple iterations to conclude a final model. Our final model includes PTRATIO, LSTAT and DIS as three major predictors.

1.2 Variables Description

CRIM	Per capita crime rate by town
DIS	Weighted distances to five Boston employment centres
LSTAT	% lower status of the population
NOX	Nitric oxides concentration (parts per 10 million)
TAX	Full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town

2. ANALYSIS

Our primary goal was to conduct statistical analysis, to identify the key parameters affecting housing prices in Boston. To perform this, we needed to have an understanding of the variables.

From the dataset of variables[1], we had to select the input variables, that we thought could potentially have an impact on our output variable which was the median house prices in Boston. To ensure this, from our dataset we started with regressing the per capita crime rate with the median house prices in Boston. However, we wanted to know how our model could be faired; hence, we added other variables to it like weighted distances to five Boston Employment Centers, in the regression model.

Our approach was to have an iterative format to find the best predictors. From our analysis, we found that some of the predictors(per capita Crime Rate) didn't contribute much change to the median value, in the presence of other variables[2].

To test and analyze this, we built our initial hypothesis by considering some of the predictors to observe and test how our output variable changes[3]. Iterating through multiple models, we were able to find the best model, we thought might help us estimate the better median value of House Prices in Boston with minimized error. This led us to find the best model having weighted distances to five Boston employment centers, pupil to teacher ratio and percentage of lower status of the population, which predicted the median value of House prices in Boston[4].
Hence, our final hypothesis was as stated below -

2.1 Hypothesis Testing

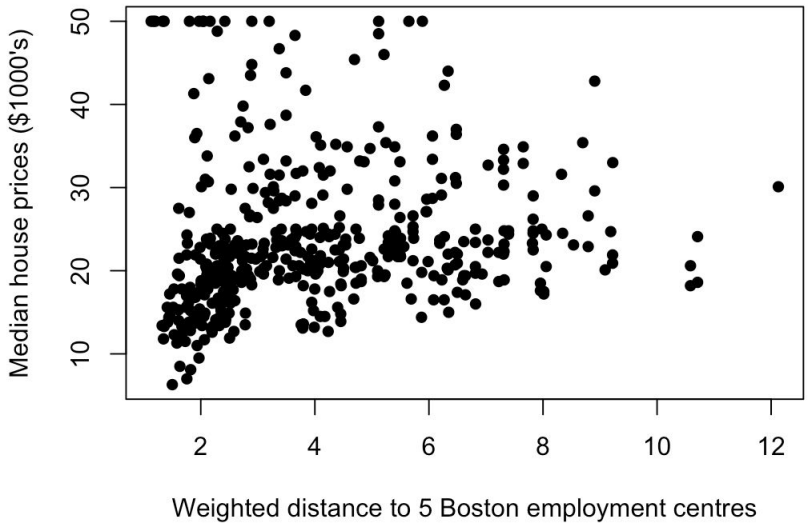
2.1.1 Null Hypothesis: There's no relation between median value of house price and weighted distances to five Boston employment centres, % lower status of the population and pupil-teacher ratio by town.

2.1.2 Alternative Hypothesis: There is a relation between the median value of house price with weighted distances to five Boston employment centers, %lower status of the population and pupil-teacher ratio by town.

We tested this hypothesis, and found that there indeed is a relationship between the above predictors and our response variable. Hence, we rejected the Null Hypothesis and accepted the Alternative hypothesis.

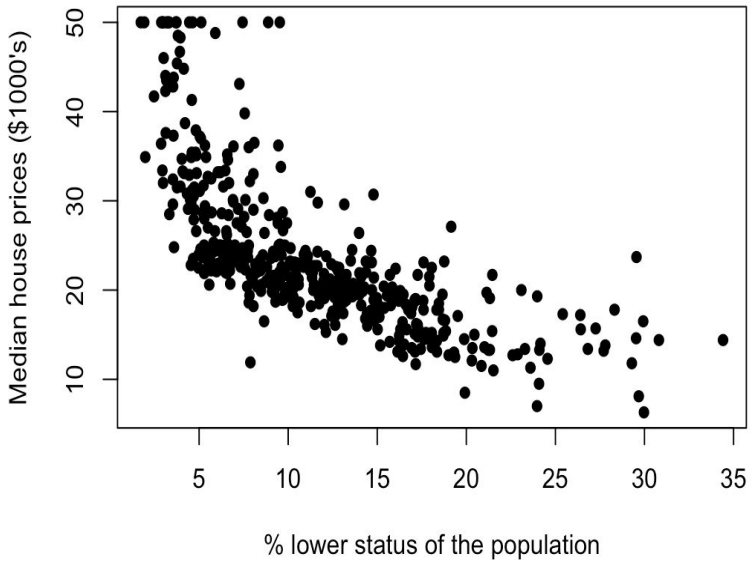
2.2 Data Displays

Plots between outcome variable and each predictor variables:

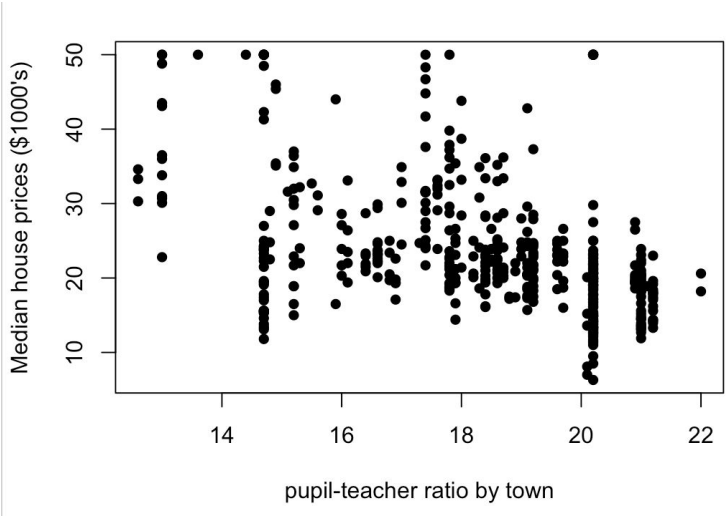


As we can see in the above plot, the spots are highly aggregated near 2 to 3 weighted distances; therefore, we can infer that most people are living near the employment center. Even with a weighted distance of 2, the house prices are in the upper range as high as 50 thousand dollars.

Moreover, as the distance goes farther, there are fewer spots with high house prices, which means that the houses that are far away from 5 Boston employment centres don't cost much to people. As a result, our general assumption is that there's negative relationship between the median house price and the weighted distance to 5 Boston employment centers.



For above scatter plot, we found that there's negative linear relationship between median house prices and percentage lower status of the population. For instance, if the lower status of the population is 5%, the median house price will go up to 50 thousand dollars.



From above graph, we see that median house prices are in the higher bracket of $20-50K, when the pupil to teacher ratio is lesser than 10. In addition to this, we see that the points go lower and lower when pupil to teacher ratio increases. This could hint that the median house prices, decrease when the pupil teacher ratio is increases. We see a few outliers from 17 and 20 people to teacher ratio, where the price also increases.

2.3 Model Building Process

To find the key parameters affecting the house prices in Boston, we made several models with various predictors. We performed multiple regressions to find the best fit model which will take into account the relationship that the key predictors have on housing prices in Boston and give us the overall model. The final key predictors are: -

- Weighted distances to 5 Boston employment centres
- % lower status of the population
- Pupil-teacher ratio by town

The relationship that we obtained from performing multiple regression analysis is as below: -

Estimated Median value of houses ($1000's) =

59.09

+

(-0.86)(weighted distances to five Boston employment centres)

+

(-1.01)(% lower status of the population)

+

(-1.11)(pupil-teacher ratio by town)

2.4 Interpretation of the Final Model

As per our model, all the final three key predictors have a negative relationship with the price of houses. Whenever each of these predictors go high, the price of house will go down. Our model explains 60% of variation in the prices of houses in Boston.

Interpretations of individual variables and the relationship that they have with house prices: -

2.5.1 Weighted distances to five Boston employment centres

The negative relationship is explained numerically by the coefficient of the variable in model. When the weighted distance to 5 Boston employment centres increase by 1 unit, the price of house decreases by $860, over and above the presence of other variables in the model. The result for this particular variable is according to our thinking that the properties or houses that are farther from the employment centres are lower in price. The houses that are very near to employment centres hold the highest price. This is also seen and realized in real life.

2.5.2 % lower status of the population

The coefficient of this variables takes a negative value representing inverse relationship between price of houses and percent lower status of the population. When the percent of lower status of the population increases by 1% in an area, the price of house goes down by $1010, over and above the presence of other predictors in the model.

2.5.3 Pupil-teacher ratio by town

As seen in the model, the coefficient is negative for pupil-teacher ratio by town. This shows that house prices are inversely related to pupil-teacher ratio. The numerical interpretation of this variable is that when the pupil-teacher ratio goes up by 1 in a town, the price of the houses go down by $1110, over and above the presence of other predictors in the model. This basically means if in a town, number of teachers are few as compared to pupil in the town, the prices will be low as compared to the area which will have more number of teachers as compared to pupils.

2.6 Recommendations

Depending upon an individual's situation, there are multiple ways by which a person can opt for economical housing in Boston. There is certainly a tradeoff between the choice of factors that one has to make to buy the house in Boston at an optimal price best suited for one's need. For example, for a bachelor best optimal deal would be to buy a house which is closer to the the business centers though it may have high pupil to teacher ratio.

Likewise for a person with kids, optimal choice would be to opt for housing where there is low pupil to teacher ratio while distance from economic centers can be high.

Hence, in the end it will boil down to the individual's situation, needs and this model then can be used as a guiding tool to estimate housing prices and negotiate with suppliers.

2.7 Limitations

There are some limitations in our predictive model. First, some variables have moderate correlation with each other. For example, the percentage lower status of the population is moderately correlated to weighted distances to five Boston employment centres. Likewise, the percentage lower status of the population is moderately correlated to pupil-teacher ratio by town[4].Therefore, we have to be a little cautious while predicting the estimated prices of houses by this model.

The second point is that the residual of the percentage lower status of the population has a pattern in the plot. Additionally, there are few outliers in the residual plot[5]. Thus, it violates the standard residual assumption, which is scatter distribution in the plot. As a result, to move forward, we will refer to statistics experts.

3. CONCLUSION

In the light of exploratory analysis/team discussions, we built our initial hypothesis considering few variables including crime rate (CRIM), Weighted distances to five Boston employment centres (DIS) and % lower status of the population as the key predictors (LSTAT).

As we moved further in the model building process, our detailed statistical analysis showed that there is indeed a negative relation of response variable (Median value of owner-occupied homes in $1000's) with each of the predictors including DIS, LSTAT and PTRATIO (pupil-teacher) ratio by town.

Some key takeaways from the model are:

- Every single unit increase in weighted distances to five Boston employment centres will decrease the estimated median value of the house by $860
- Every single unit % increase in lower status of the population will decrease the estimated median value of the house by $1010
- Every single unit% increase in pupil-teacher ratio by town will decrease the estimated median value of the house by $1110

APPENDIX

1. **Variable information**

Variable Names	Definition
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment

	centres
RAD	index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

2. **Null Hypothesis:** There's no relation between median value of housing prices in Boston and per capita crime rate by town, weighted distances to five Boston employment centres, % lower status of the population

Alternate Hypothesis : There's a relation between median value of house price in 1000s and per capita crime rate by town, weighted distances to five Boston employment centres and % lower status of the population.

Residuals:

Min	1Q	Median	3Q	Max
-17.464	-3.820	-1.475	2.082	23.106

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	40.36216	1.07179	37.659	< 2e-16 ***
CRIM	-0.20210	0.13411	-1.507	0.133
DIS	-0.90700	0.16002	-5.668	2.59e-08 ***
LSTAT	-1.10620	0.05323	-20.780	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.043 on 448 degrees of freedom
Multiple R-squared: 0.5325, Adjusted R-squared: 0.5294
F-statistic: 170.1 on 3 and 448 DF, p-value: < 2.2e-16

Having a p-value larger than 0.05, from CRIM field, makes it insignificant for the model.

3. **Coefficients and statistics of predictors in the final model**

Residuals:

Min	1Q	Median	3Q	Max
-13.793	-3.293	-0.832	1.901	24.015

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	59.09371	2.36415	24.996	< 2e-16 ***
DIS	-0.85806	0.13904	-6.171	1.52e-09 ***
LSTAT	-1.01363	0.04897	-20.701	< 2e-16 ***
PTRATIO	-1.11118	0.12558	-8.849	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

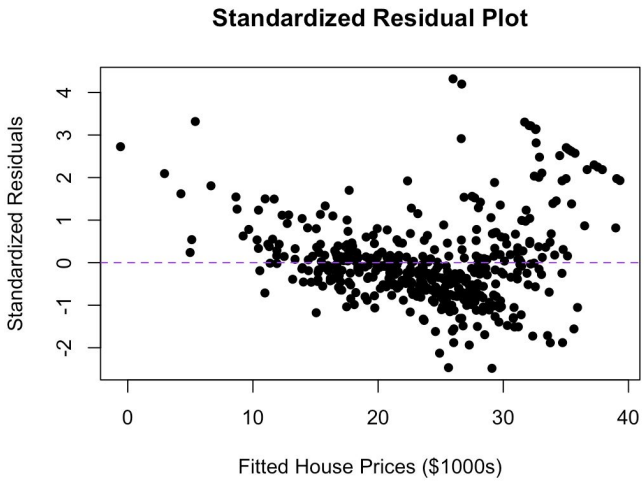
Residual standard error: **5.589** on 448 degrees of freedom
Multiple R-squared: 0.6001, Adjusted R-squared: 0.5974
F-statistic: 224.1 on 3 and 448 DF, p-value: < 2.2e-16

As for error, in our final model, we found that it was as low as **5.589**, as shown above. This supports our model for better accuracy.

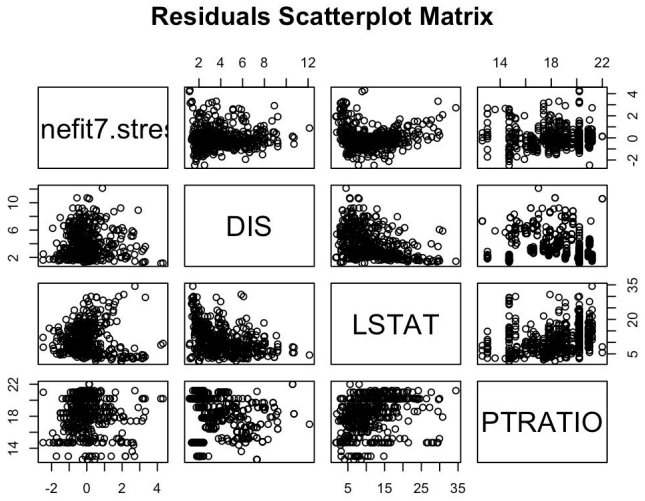
4. **Collinearity:**

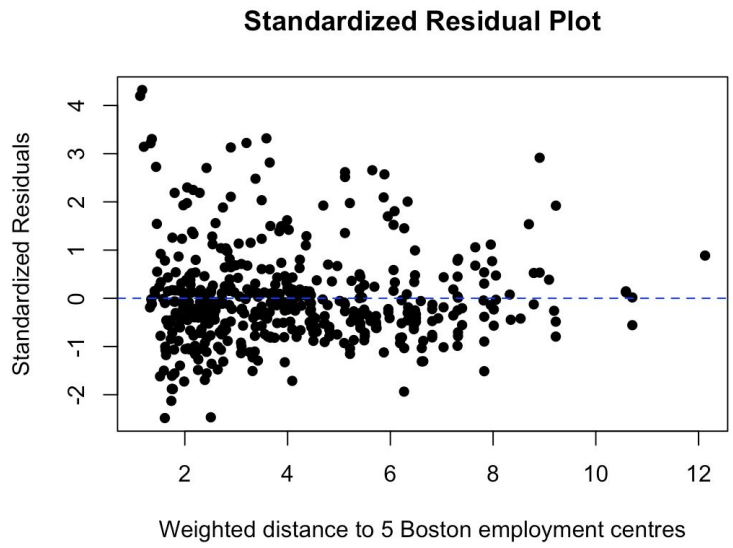
The correlation between the percentage lower status of the population and the weighted distances to five Boston employment centres is moderate as the correlation value is -0.423724689. Furthermore, the percentage lower status of the population has moderate correlation of 0.303043086 to pupil-teacher ratio by town. We are considering this collinearity as moderate because it lies between .2 and .7.

5.

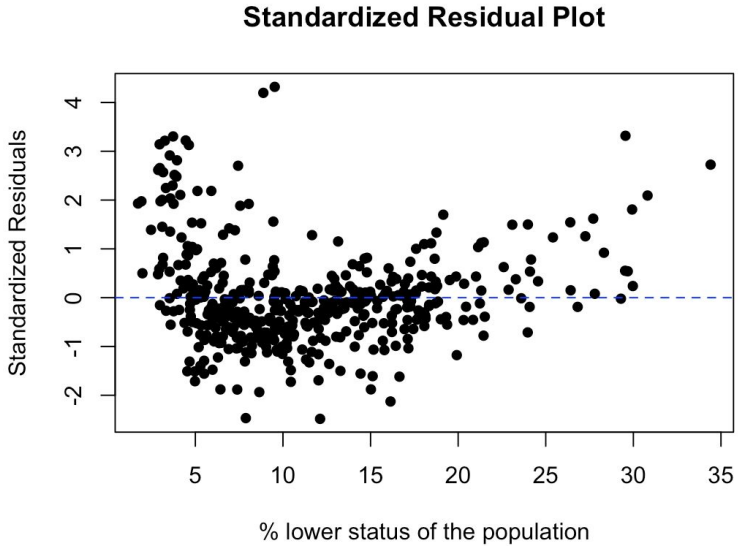


In the total residual plot, we see that without a couple of outliers, majority of the residuals here are pretty much aligned towards the mean.

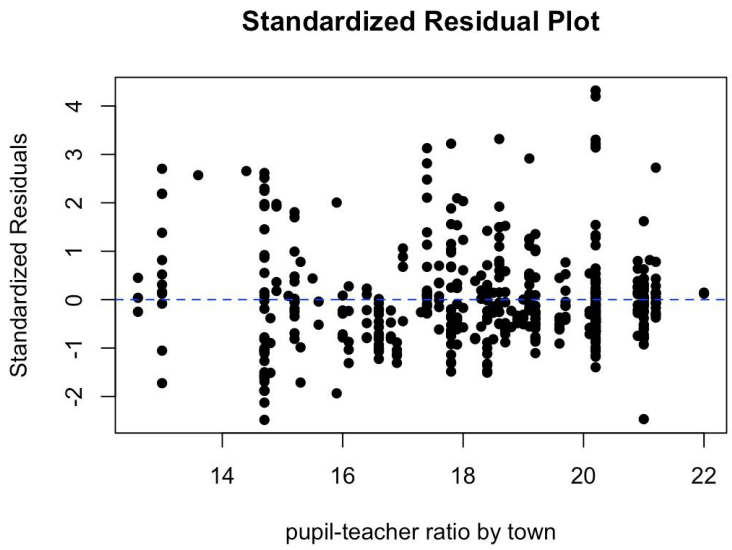




In this plot we see that there is some variation above the mean in positive direction, which broadens the range. There are a few instances of outliers as well.



In this plot, the outliers are taking the residuals to unit 3 or 4 due to the deviation.



This residual plot is appropriately scattered along the mean, apart from a few outliers.

REFERENCES

1. **Title:** Boston Housing Data
2. **Origin:** This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.
3. **Creator:** Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.
4. **Date:** July 7, 1993
5. **Past Usage:**
    - Used in Belsley, Kuh & Welsch, 'Regression diagnostics ...', Wiley, 1980. N.B. Various transformations are used in the table on pages 244-261.
    - Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning.In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.