Boston Housing Price Analysis

By: Renee Ge, Seowoo Kim, Taeim Kwon, Yun Jin Park, Luyang You

Introduction:

Boston is now the second-most expensive city in the country for renters according to a report issued by Zumper(Oct. 27,2022), a national platform that connects renters with new properties. Data from the report shows that Boston jumped ahead of San Francisco over the past month when it comes to one bedroom median rent prices. Therefore we are interested in which factors make such an increase in Boston housing price. Our ultimate purpose is developing the best model for predicting Boston housing prices.

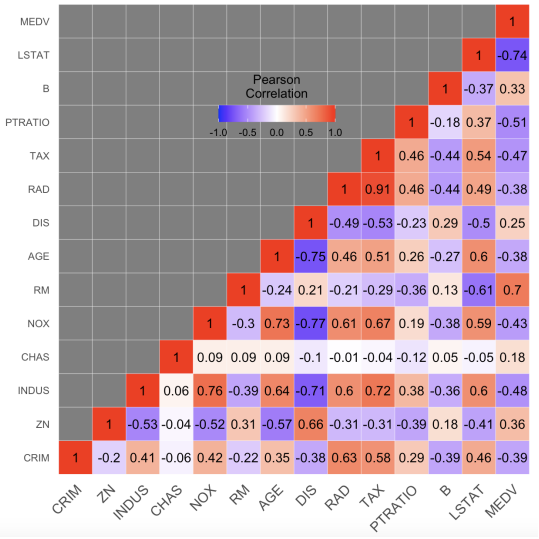
Analysis Plan:

We will start with some exploratory data analyses to examine our outcome and covariates more. We will then fit our full model with all covariates and check the linear model assumptions, examine outliers, and check for collinearity among our variables. We will make any transformations/perform any measures necessary to address uncovered issues. Finally, we will use backwards stepwise regression with a 0.05 significance level to select our final model.

Exploratory Data Analysis:

Looking at the distributions of each variable we found that our outcome variable MEDV seems slightly skewed. Within our predictors we also found that AGE, DIS, and LSTAT had somewhat skewed distributions as well.

In an examination of the correlation between all our variables we found that the highest correlation within our predictors existed between TAX and RAD. We also saw that the DIS variable had somewhat strong correlation as well with the AGE, NOX, and INDUS variables. This could indicate collinearity issues and we will examine this more in depth later on.



Full Model:

Our full model is shown as below:

$$MEDV = \beta_0 + \beta_1 CRIM + \beta_2 ZN + \beta_3 INDUS + \beta_4 CHAS + \beta_5 NOX + \beta_6 RM + \beta_7 AGE + \beta_8 DIS + \beta_9 RAD + \beta_{10} TAX + \beta_{11} PTRATIO + \beta_{12} B + \beta_{13} LSTAT + \varepsilon$$

The adjusted R^2 for the full model is 0.7338. In Fig. 1, most of the parameter estimates are statistically significant except AGE and INDUS variables, which have p-values greater than 0.05.

Since we used least squares estimation to fit our model, we checked five assumptions with the result shown in Fig. 2. First of all, the existence assumption holds true because we are considering a finite number of subjects (n=506) to fit our model. Secondly, although there might be correlation between samples since it's spatial data, we will assume that independence assumption holds true. Third, we concluded that linearity assumption and homogeneity assumption is violated in our current model due to a clear pattern shown in the residual plots of Fig. 2. Lastly, in the QQ plot of Fig. 2, the curve is concave upwards, meaning that the data is positively-skewed and the residuals do not follow normal distribution.

The result of the Kolmogorov-Smirov test (Fig. 3) also supports that Gaussian errors assumption is also violated.

In addition to evaluating the full model, we calculated Cook's distance to assess which data point is influential in fitting our model. According to Fig. 4, observations 365, 369, 373 were shown to be the three most influential observations in our dataset. Fig. 5 shows further analysis on these three data points based on their values of each variable. Since the RM value of 365th observation(8.780) deviates largely from the mean of the RM (6.285), we decided to remove 365th observation from our data.

Transformations:

To find a model that fits our data better, we conducted some tests to transform or remove variables from our full model.

As previously seen in the diagnostics, the residuals of the data are not normally distributed. So, we performed the box-cox transformation and the selected lambda is 0 (Fig. 8). Therefore, a log transformation was performed on the median value of owner-occupied homes in $1000's (MEDV). Comparing the residual plots of the raw data and the transformed data, the distribution of the transformed data residuals is more evenly spread out (Fig. 9).

We used the variance inflation factor (VIF) to check the collinearity among the predictors of the dataset. The VIF value of the index of accessibility to radial highways (RAD) is 7.48 and the VIF value of the full-value property-tax rate per $10,000 (TAX) is 9.01 which are both not close to 1 and greater than 5 which indicates high collinearity (Fig. 6). We decided to remove only the predictor TAX and found the new VIF value of RAD is 2.83 where there is no longer collinearity between the remaining predictors (Fig. 7).

Next, we examined all the remaining predictors individually through observing the trend of scatterplots between the predictors and the log of dependent variable MEDV. We found that only the weighted distances to five Boston employment centers (DIS) had a log relation with the transformed dependent variable, so a log transformation was performed on predictor DIS (Fig. 10). Other than this particular predictor, we did not find any other predictors that are in need of transformations.

Model Selection:

In order to achieve the optimal reduced model, backward elimination method with significance level of 0.05 was selected to perform on all the remaining and transformed predictors. Three predictors including proportion of residential land zoned for lots over 25,000 sq.ft. (ZN), proportion of non-retail business acres per town (INDUS) and proportion of owner-occupied units built prior to 1940 (AGE) were eliminated along the process (Fig. 11).

New Model:

Our new model is shown as follows:

$$log(MEDV) = \beta_0 + \beta_1 CRIM + \beta_2 CHA + \beta_3 NOX + \beta_4 RM + \beta_5 log(DIS) + \beta_6 RAD + \beta_7 PTRATIO + \beta_8 B + \beta_9 LSTAT + \varepsilon$$

The adjusted R^2 for this model was 0.7959, which is higher than the adjusted R^2 for the full model (0.7338). The table of the parameter estimates (Fig. 12) shows that all the parameter estimates are statistically significant as their p-values are less than 0.05, thereby suggesting the new model was improved compared with the full model.

Revisiting Assumption Checking:

The assumptions were revisited for the new model as linearity, homogeneity, and gaussian errors assumptions were what we were concerned about from the full model. It was found that the points from the residual plot (Fig. 13) got more randomly spreaded out after the box-cox transformation. That is, we could say linearity and homogeneity assumptions would hold true. The Q-Q plot for the new model (Fig. 13) seems to roughly follow the normal distribution compared with the one for the full model (Fig. 2) that has few points greatly off of the line at the right top. In addition, the sample size (n = 505) was large enough to suggest that the gaussian errors assumption holds true.

Discussion:

We found from the new model that on average, the log of median values of homes in Boston was positively associated with the four covariates: CHAS, RM, RAD, and B. This could be interpreted meaningfully such that being closer to Chales River, having a larger number of rooms per dwelling, being more accessible to radial highways, and having more population homogeneity had a positive impact on the mean log of median values of homes in Boston. However, on average, there was a negative association found between the log of median values of homes in Boston and the six covariates: CRIM, NOX, DIS, TAX, PTRATIO, and LSTAT. That is, being distant from the five Boston employment centers and having more crimes, more air pollution, higher tax rate, higher pupil-teacher ratio, and proportion of lower status led to decreased mean log of median values of homes in Boston.

One of our future directions is to perform polynomial regression analysis to fit the model better. Also, it should be noted that our data is from the 1970s, which means the data is somewhat outdated and may not be relevant to Boston these days. Therefore, more accurate predictions would be able to be made with more recent data.

References:

Zumper National Rent Report(Oct.2022),Zumper

Appendix:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	36.45949	5.10346	7.14	<.0001
CRIM	1	-0.10801	0.03286	-3.29	0.0011
ZN	1	0.04642	0.01373	3.38	0.0008
INDUS	1	0.02056	0.06150	0.33	0.7383
CHAS	1	2.68673	0.86158	3.12	0.0019
NOX	1	-17.76661	3.81974	-4.65	<.0001
RM	1	3.80987	0.41793	9.12	<.0001
AGE	1	0.00069222	0.01321	0.05	0.9582
DIS	1	-1.47557	0.19945	-7.40	<.0001
RAD	1	0.30605	0.06635	4.61	<.0001
TAX	1	-0.01233	0.00376	-3.28	0.0011
PTRATIO	1	-0.95275	0.13083	-7.28	<.0001
B	1	0.00931	0.00269	3.47	0.0006
LSTAT	1	-0.52476	0.05072	-10.35	<.0001

Fig. 1: Table of Parameter Estimates for the Full Model.

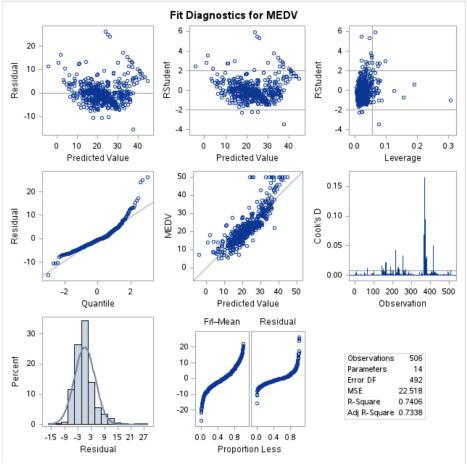


Fig. 2: Regression Diagnostic Plots for the Full Model.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  model$residuals
D = 0.34781, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Fig. 3: Result of Kolmogorov-Smirnov test

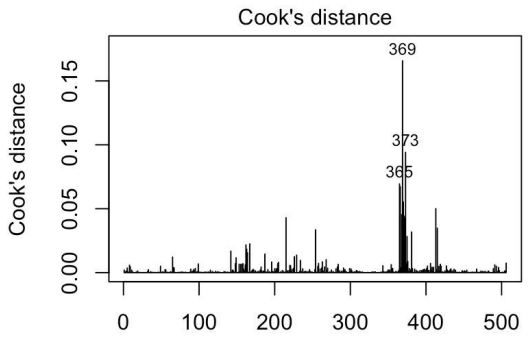


Fig. 4: Cook's distance

Variable Names	365th Obs	369th Obs	373th Obs	Comment
CRIM	-0.016	0.149	0.541	
ZN	-0.487	-0.487	-0.487	
INDUS	1.015	1.015	1.015	
CHAS	3.665	-0.272	3.665	Categorical
NOX	1.409	0.658	0.978	
RM	3.552	-1.871	-0.583	365th Obs > 3 SD
AGE	0.509	1.116	0.747	
DIS	-0.898	-1.169	-1.266	
RAD	1.660	1.660	1.660	
TAX	1.529	1.529	1.529	
PTRATIO	0.806	0.806	0.806	
B	-0.023	0.206	-0.096	
LSTAT	-1.031	-1.315	-0.528	
MEDV	0.127	2.987	2.987	369th Obs >2 SD 373th Obs >2 SD

Fig. 5: Table showing how much the values of observation 365, 369, 373 deviate from the mean of each variable.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	36.45949	5.10346	7.14	<.0001	.	0
CRIM	per capita crime rate by town	1	-0.10801	0.03286	-3.29	0.0011	0.55798	1.79219
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.	1	0.04642	0.01373	3.38	0.0008	0.43502	2.29876
INDUS	proportion of non-retail business acres per town	1	0.02056	0.06150	0.33	0.7383	0.25053	3.99160
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)	1	2.68673	0.86158	3.12	0.0019	0.93110	1.07400
NOX	nitric oxides concentration (parts per 10 million)	1	-17.76661	3.81974	-4.65	<.0001	0.22760	4.39372
RM	average number of rooms per dwelling	1	3.80987	0.41793	9.12	<.0001	0.51713	1.93374
AGE	proportion of owner-occupied units built prior to 1940	1	0.00069222	0.01321	0.05	0.9582	0.32249	3.10083
DIS	weighted distances to five Boston employment centres	1	-1.47557	0.19945	-7.40	<.0001	0.25278	3.95594
RAD	index of accessibility to radial highways	1	0.30605	0.06635	4.61	<.0001	0.13361	7.48450
TAX	full-value property-tax rate per $10,000	1	-0.01233	0.00376	-3.28	0.0011	0.11101	9.00855
PTRATIO	pupil-teacher ratio by town	1	-0.95275	0.13083	-7.28	<.0001	0.55584	1.79908
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town	1	0.00931	0.00269	3.47	0.0006	0.74155	1.34852
LSTAT	% lower status of the population	1	-0.52476	0.05072	-10.35	<.0001	0.33996	2.94149

Fig. 6: Variance Inflation Factor value table under Full Model.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	34.62864	5.12280	6.76	<.0001	.	0
CRIM	1	-0.10673	0.03319	-3.22	0.0014	0.55805	1.79194
ZN	1	0.03637	0.01351	2.69	0.0074	0.45783	2.18424
INDUS	1	-0.06778	0.05583	-1.21	0.2253	0.30998	3.22602
CHAS	1	3.02923	0.86365	3.51	0.0005	0.94498	1.05822
NOX	1	-18.70121	3.84662	-4.86	<.0001	0.22887	4.36927
RM	1	3.91169	0.42088	9.29	<.0001	0.52000	1.92307
AGE	1	-0.00060540	0.01333	-0.05	0.9638	0.32278	3.09804
DIS	1	-1.48830	0.20138	-7.39	<.0001	0.25288	3.95445
RAD	1	0.13458	0.04125	3.26	0.0012	0.35242	2.83749
PTRATIO	1	-0.98513	0.13174	-7.48	<.0001	0.55902	1.78884
B	1	0.00955	0.00271	3.52	0.0005	0.74208	1.34756
LSTAT	1	-0.52221	0.05121	-10.20	<.0001	0.34004	2.94080

Fig. 7: Variance Inflation Factor value table after removing predictor TAX.

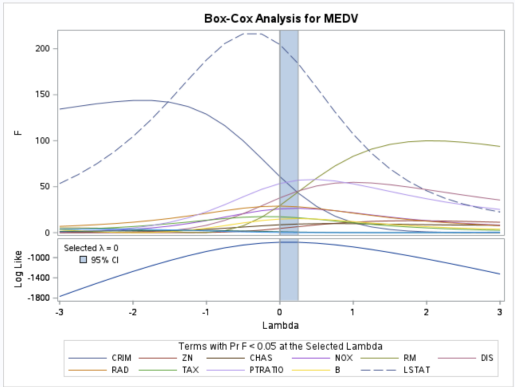


Fig. 8: Box-Cox Transformation (λ=0)

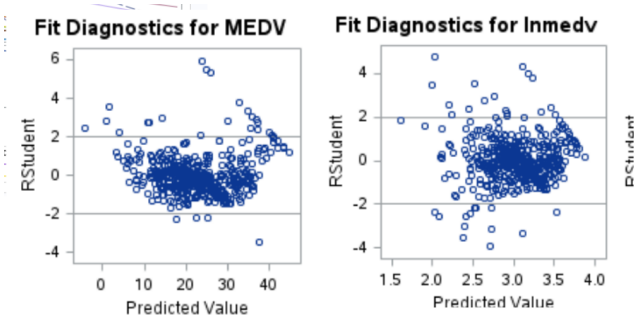


Fig. 9: Residual plot before log transformation (right); residual plot after log transformation (left).

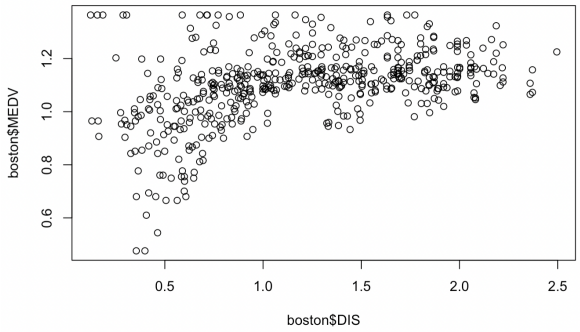


Fig. 10: Scatter plot of predictor DIS vs log(MEDV).

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	INDUS	12	0.0001	0.8010	12.1679	0.17	0.6821
2	AGE	11	0.0001	0.8008	10.5323	0.37	0.5460
3	ZN	10	0.0009	0.7999	10.7256	2.20	0.1387

Fig. 11: Summary of backward elimination method (significance level = 0.05).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	1	4.33156	0.20992	20.63	<.0001	14.54911
CRIM	1	-0.01147	0.00129	-8.88	<.0001	2.69219
CHAS	1	0.11475	0.03372	3.40	0.0007	0.39561
NOX	1	-0.98294	0.14642	-6.71	<.0001	1.53996
RM	1	0.10088	0.01588	6.35	<.0001	1.37988
DIS	1	-0.24644	0.02890	-8.53	<.0001	2.48512
RAD	1	0.01446	0.00248	5.83	<.0001	1.16189
TAX	1	-0.00059945	0.00012981	-4.62	<.0001	0.72863
PTRATIO	1	-0.03897	0.00472	-8.26	<.0001	2.33164
B	1	0.00038863	0.00010456	3.72	0.0002	0.47209
LSTAT	1	-0.02935	0.00186	-15.80	<.0001	8.53181

Fig. 12: Table of Parameter Estimates for the New Model.

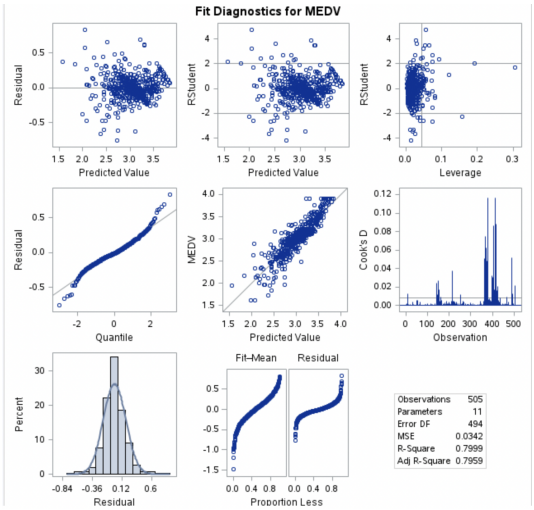


Fig. 13: Regression Diagnostic Plots for the New Model.