Boston Housing Prices:

A case study

Motivation

To conduct a linear regression analysis on the Boston House Pricing dataset

Anish Chakrabarty (171024)
Rahul Singh (171111)
Soumyadip Fadikar (171154)
Sourav Nandi (171157)

☙ *Introduction*

This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass.The dataset has information on 14 variables/attributes on 506 observations. The variables will be denoted as the terms expressed in the brackets. The data description is given as:-

- crim(x_1) =per capita crime rate by town.
- zn(x_2) = proportion of residential land zoned for lots over 25,000 sq.ft.
- Indus (x_3)= proportion of non-retail business acres per town.
- Chas (x_4)= Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- nox (x_5)= nitrogen oxides concentration (parts per 10 million).
- Rm (x_6) =average number of rooms per dwelling.
- age(x_7) =proportion of owner-occupied units built prior to 1940.
- dis(x_8) =weighted mean of distances to five Boston employment centres.
- rad(x_9)=index of accessibility to radial highways.
- tax(x_{10})=full-value property-tax rate per \$10,000.
- ptratio(x_{11})=pupil-teacher ratio by town.
- black(x_{12})=*1000(Bk - 0.63)^2* where *Bk* is the proportion of blacks by town.
- lstat(x_{13})=lower status of the population (percent).
- medv(y)=median value of owner-occupied homes in \$1000s.

Of these, medv is the response variable while the other 13 variables are possible predictors. The goal of this analysis is to fit a regression model that best explains the variation in medv. First, the individual variables will be analyzed to assess if the data is skewed in a way that would affect the accuracy of the model. Some variables may require transformations to better fit the model. Next, a full regression model will be fitted using the transformations as appropriate, with 13 individual predictor variables. Some of the predictor variables may do a good job of explaining this variation, while others may not. Furthermore, some predictor variables may be highly correlated, which causes the issue of multicollinearity in which we cannot be certain which of the correlated predictors is most responsible for variation in the response. To address this, variance inflation factors will be computed to quantify the effect of multicollinearity and eliminate predictors as necessary keeping in mind the ultimate goal of finding the best predictors.

After eliminating variables that are weak predictors, a residual analysis will be performed using quantile plots, scatterplots and histograms to check that the model satisfies the assumptions of linear regression, namely, that the errors are normally distributed and have constant variance. This will also serve the purpose of identifying potential outliers which are affecting the model's accuracy. Of particular interest is identifying the outliers that have a large effect on the regression. These influential observations will be identified and removed using Cook's Distance. From the model thus obtained, interesting inferences can be drawn about the data and conclusions can be made about the most important factors that explain the variation in housing prices.

Here the variable of interest is y. We want to implement a linear regression model to predict the value of y based on the values of other x_i's.

The linear model to be implemented is given as,

$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}+\beta_6x_{i6}+\beta_7x_{i7}+\beta_8x_{i8}+\beta_9x_{i9}+\beta_{10}x_{i10}+\beta_{11}x_{i11}+\beta_{12}x_{i12}+\beta_{13}x_{1i13}+\varepsilon_i$ for $i=1,2,....,n=506$

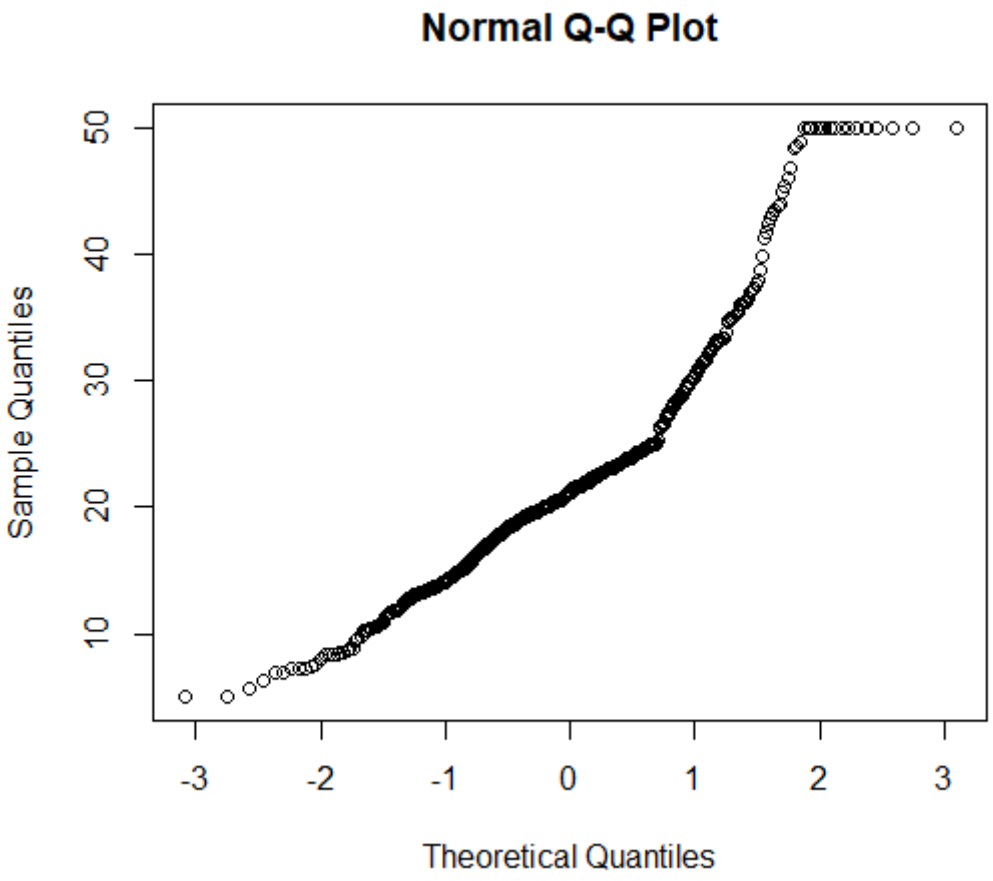
where β_i's represent the linear regression coefficient corresponding to the ith regressor.

The assumption for a Multiple Linear Regression(MLR) model is given as,

$$\varepsilon_i \sim N(0,\sigma^2) \implies y_i \sim N(\beta_0+\sum_{j=1}^{13}\beta j\ xij\ ,\sigma^2)$$

So before fitting the MLR model we first test the assumption of normality of the response variable. We use a qq-plot for this. We first plot the qq-plot for all the observations.

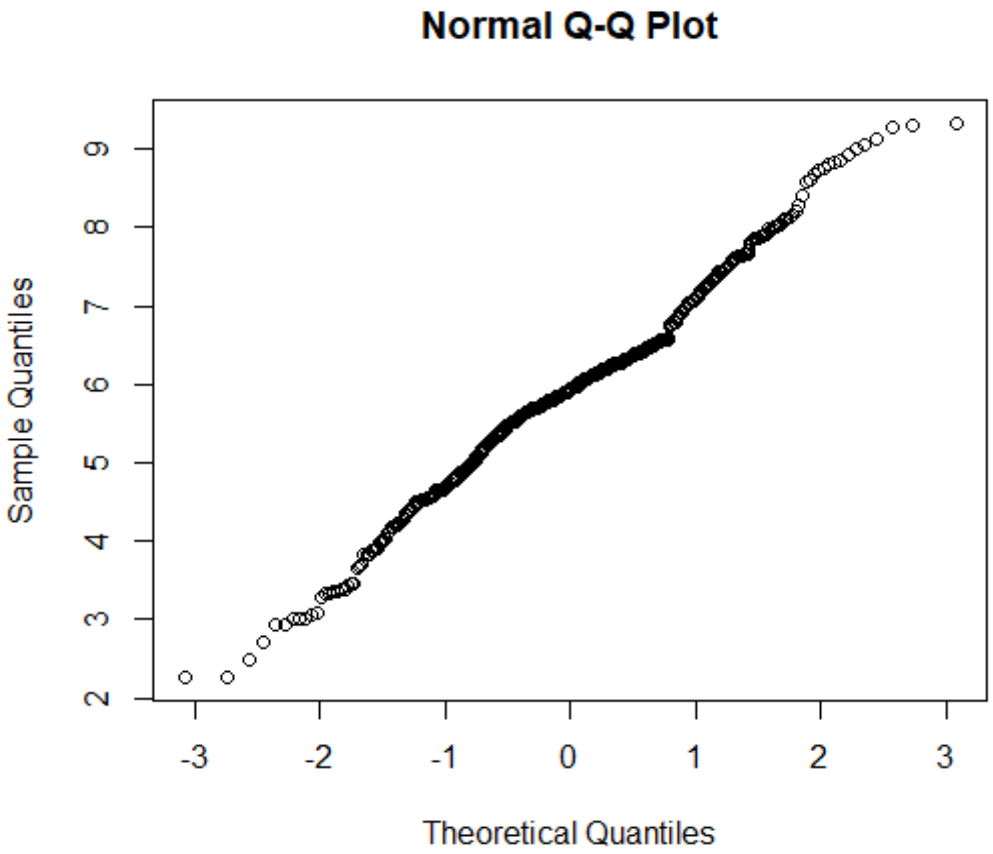
The following graph is obtained.



We observe that there is possible censoring in the data. The maximum of medv is 50 and this value is reported for 16 observations. This is causing a heavy right tail and thus the deviation from normality is appearing to be visually significant. We remove this 506 observations from the data set and fit a model on the remaining 490 observations. Again the visual deviation from normality appears significant and therefore we apply the one parameter Box-Cox Transformation which is given as follows:-

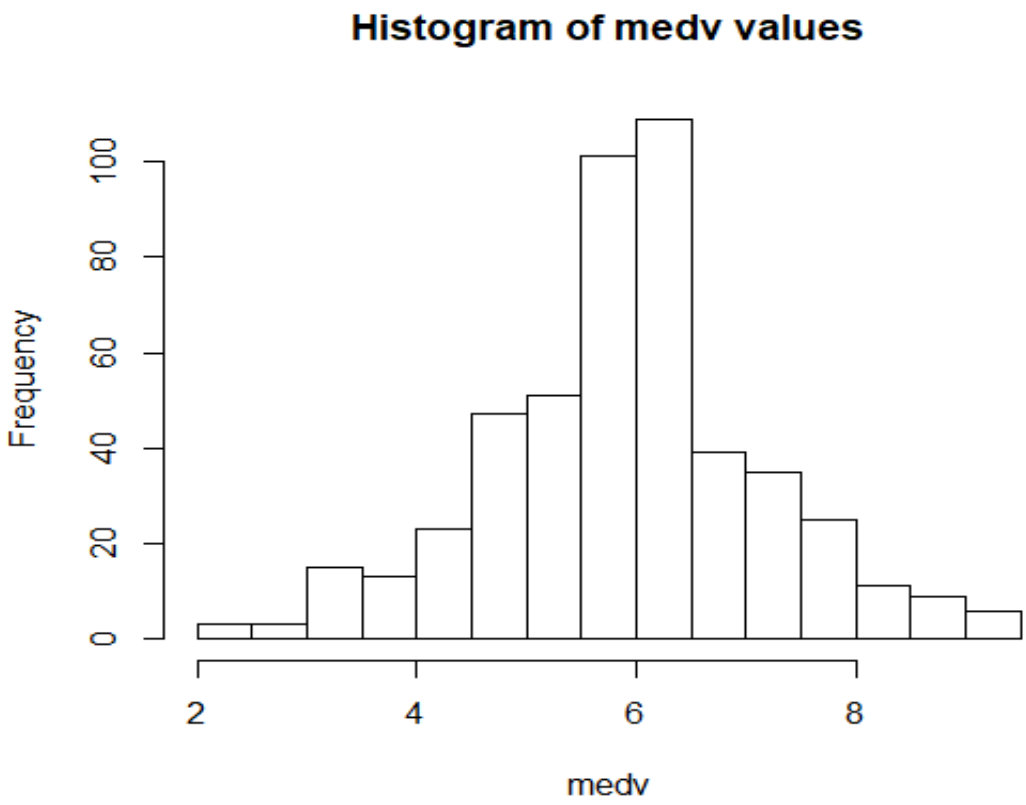
$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$

The estimated value λ=0.4
After applying the Box-Cox Transformation we again plot the qq-plot. We obtain the following graph.



In this graph, we have enough visual evidence that the observed medv values can be considered to have been originated from a normal distribution.

The histogram of medv is obtained as,



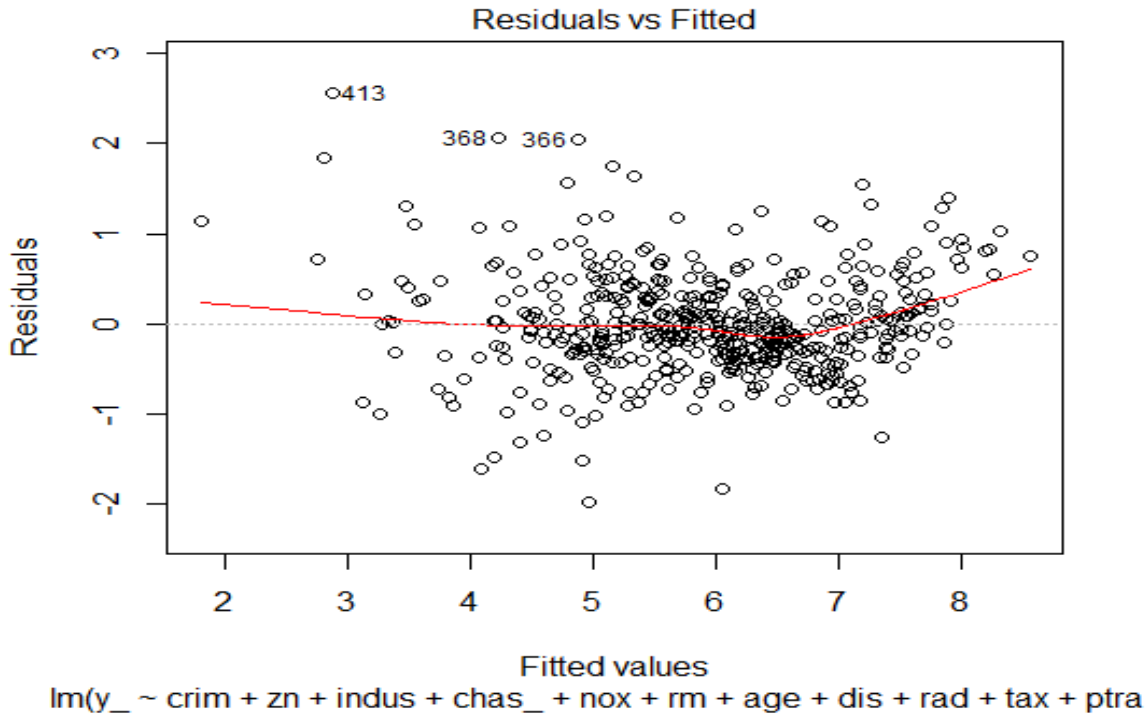
This histogram further justifies that the observed medv values can considered to have been originated from a Normal distribution. From now on, we denote the transformed output variable as $medv_{0.4}$(or $y_{0.4}$)

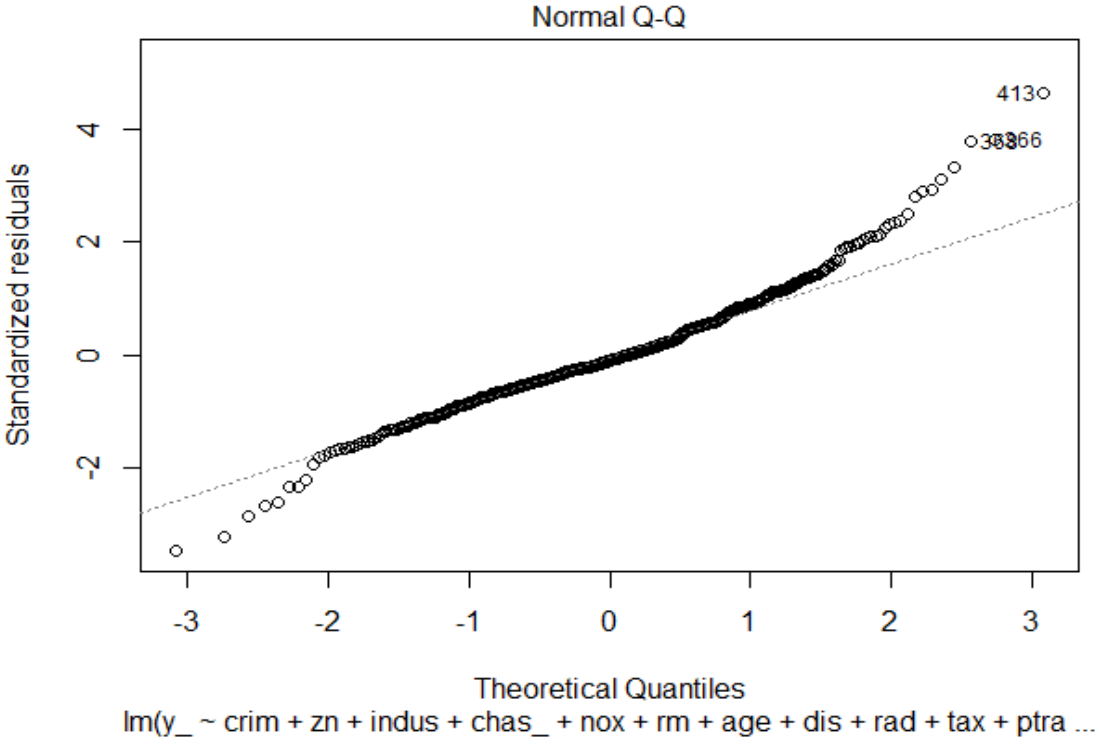
Next, we want to check whether the error term can be considered to be homoscedastic.

We fit the MLR model using all the regressors and obtain the residuals and plot the residuals to obtain the following graph:-

From the graph, it is observed that the residuals are scattered randomly against the fitted values and there is no observable pattern. Thus we can consider the residuals to be homoscedastic.

The qqplot of the standardised residuals is given below:





Now, before tackling the problems of multicollinearity and variable selection, we fit a linear regression model to obtain a general sense of the implementation of the model.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.5200386	0.6227103	13.682	< 2e-16
crim	-0.0259649	0.0039426	-6.586	1.20e-10
zn	0.0041339	0.0017019	2.429	0.0155
indus	-0.0007761	0.0075221	-0.103	0.9179
chas_1	0.1224234	0.1120042	1.093	0.2749
nox	-2.0853598	0.4616207	-4.517	7.90e-06
rm	0.4287159	0.0539660	7.944	1.42e-14
age	-0.0025406	0.0016085	-1.579	0.1149
dis	-0.1626478	0.0242602	-6.704	5.74e-11
rad	0.0408931	0.0080162	5.101	4.88e-07
tax	-0.0021928	0.0004529	-4.842	1.74e-06
ptratio	-0.1252955	0.0159113	-7.875	2.33e-14
black	0.0012700	0.0003218	3.947	9.11e-05
lstat	-0.0700554	0.0064207	-10.911	< 2e-16

From this primary model we observe that the hypothesis H_{0j}: $\beta_j=0$ vs. $H_{1j:}$ Not H_{oj} is getting accepted for the regressors indus,chas and age at 5% level of significance.

❖ *Influential Point Detection:*

We start with the problem of outlier detection. For improving the model we identify the outliers using Cook's Distance and replacing the outliers by the mean value of y.

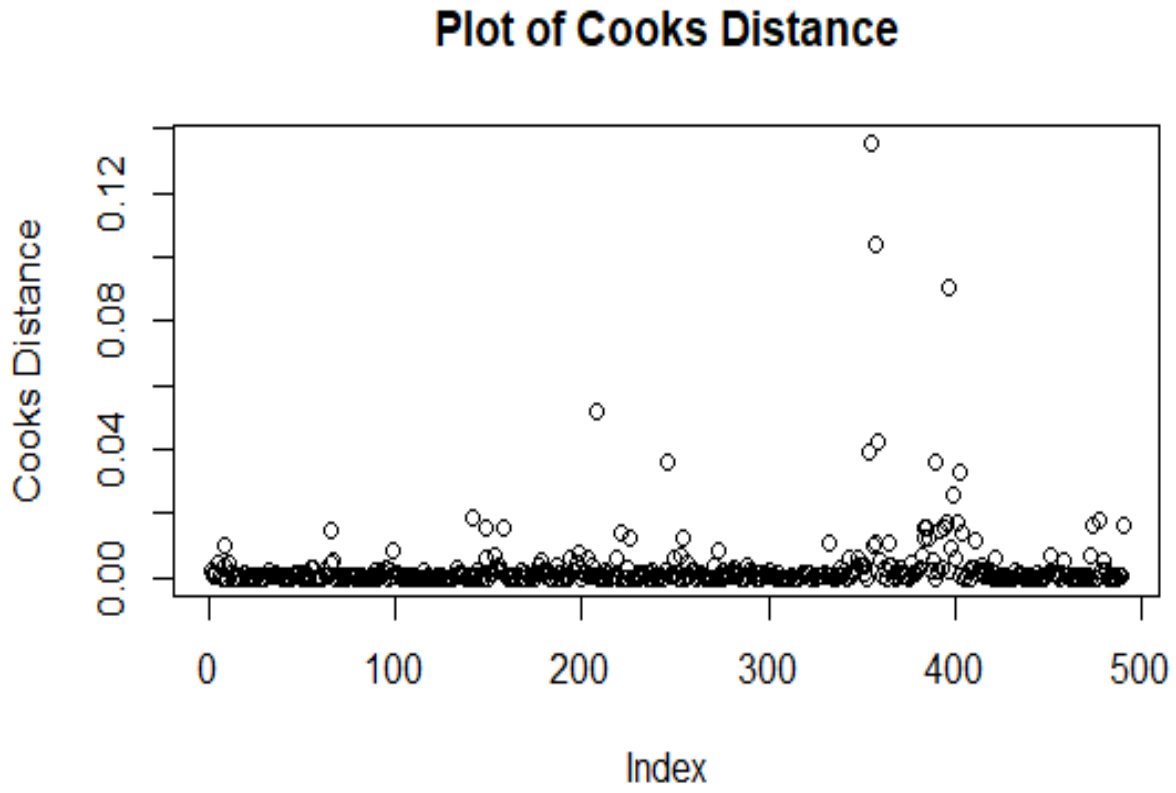
Cook's distance or **Cook's *D*** is a commonly used estimate of the influence of a data point when per forming a least-squares regression analysis.[1] In a practical ordinary least squares analysis, Cook's distance can be used in several ways, to indicate influential data points that are particularly worth checking for validity.

$D_i=((y_i-\hat{y_i})*h_{ii})/(p*MSE*(1-h_{ii})^2)$

Where, $\hat{y_i}$denotes the predicted value of the i^{th} observation and

h_{ii} denotes the i^{th} diagonal element in the hat matrix=$X(X'X)^{-1}X'$

p=Number of regressors.



There are different opinions regarding what cut-off values to use for spotting highly influential points. A simple operational guideline of Di>1 has been used. Thus, we observe that there are no outliers in the dataset when we are fitting the linear regression model.

❖ *Multicollinearity:*

Now, we obtain the Variance Inflation factor(VIF) for the regressors to judge whether multicollinearity is present in the data or not.[0]

$VIF_j = 1/(1-R_j^2)$

where R_j=The correlation coefficient between X_j and the predicted value of $X_j = \hat{X}_j$ which is obtained from the subset regression:-

$X_{ij} = \delta_0 + \delta_1 X_{i1} + \ldots\ldots + \delta_{(j-1)} X_{i(j-1)} + \delta_{(j+1)} X_{i(j+1)+\ldots+} \delta_p X_{ip}$ where i=1,2,….n

The VIF values are obtained as,

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
1.79	2.29	3.99	1.06	4.40	1.8	3.11	3.97	7.36	8.79	1.71	1.34	3.13

From the VIF values it is observed that the variables rad and tax are involved in multicollinearity because the VIF values of these variables is greater than 5.

We also observe that the correlation coefficient of rad and tax=0.908

We drop the variable tax because the VIF value of tax is observed to greater than that of rad.
We do not continue to the Variation Decomposition Method because only two variables are involve d.The VIF values of the regressors after dropping the tax variable are obtained as,

crim	zn	indus	chas	nox	rm	age	dis	rad	ptratio	black	lstat
1.79	2.17	3.30	1.04	4.37	1.87	3.11	3.97	2.85	1.70	1.34	3.13

We observe that after dropping the variable tax from the set of regressors the VIF values of the every regressor turns out to be <5. Thus, the presence of multicollinearity has been removed from the model by dropping the variable "tax".

We do not go on to check for the presence of autocorrelation in the data because the data points in the dataset do not vary over time but over space in the city of Boston and the data points have been collected randomly over the localities of Boston City.

❖ *Variable Selection:*

We now move on to the problem of variable selection. We want to select a smaller set of predictor variables from the larger set because even though the coefficients of the variables retained in the smaller set may be a little biased the variance of these estimates is smaller than those in the larger set of variables. Thus, through such variable selection we seek for a trade off between the biasedness and the variances of the estimates of the variables retained.

Here, we use the Backward Elimination method for Variable Selection Method because of its faster convergence to the best possible subset. At α=0.05 we obtain the following variables are removed from the model.

```
Backward Elimination Method
---------------------------

Candidate Terms:

1 . crim
2 . zn
3 . indus
4 . chas_
5 . nox
6 . rm
7 . age
8 . dis
9 . rad
10 . ptratio
11 . black
12 . lstat

We are eliminating variables based on p value...

× zn

Backward Elimination: Step 1

 Variable zn Removed
```

Model Summary

R	0.887	RMSE	0.581
R-Squared	0.787	Coef. Var	9.828
Adj. R-Squared	0.782	MSE	0.338
Pred R-Squared	0.771	MAE	0.427

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	8.233	0.634		12.989	0.000	6.987	9.478
crim	-0.025	0.004	-0.176	-6.265	0.000	-0.033	-0.017
indus	-0.016	0.007	-0.087	-2.264	0.024	-0.030	-0.002
chas_1	0.191	0.114	0.036	1.686	0.093	-0.032	0.414
nox	-2.270	0.471	-0.213	-4.818	0.000	-3.196	-1.344
rm	0.455	0.055	0.239	8.340	0.000	0.348	0.563
age	-0.003	0.002	-0.066	-1.774	0.077	-0.006	0.000
dis	-0.149	0.022	-0.253	-6.672	0.000	-0.193	-0.105
rad	0.011	0.005	0.078	2.190	0.029	0.001	0.021
ptratio	-0.138	0.016	-0.233	-8.865	0.000	-0.168	-0.107
black	0.001	0.000	0.098	3.993	0.000	0.001	0.002
lstat	-0.070	0.007	-0.397	-10.621	0.000	-0.083	-0.057

✕ chas_

Backward Elimination: Step 2

Variable chas_ Removed

Model Summary

R	0.886	RMSE	0.582
R-Squared	0.786	Coef. Var	9.847
Adj. R-Squared	0.781	MSE	0.339
Pred R-Squared	0.770	MAE	0.430

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	8.236	0.635		12.969	0.000	6.988	9.484
crim	-0.026	0.004	-0.179	-6.356	0.000	-0.033	-0.018
indus	-0.016	0.007	-0.086	-2.236	0.026	-0.030	-0.002
nox	-2.230	0.471	-0.209	-4.729	0.000	-3.156	-1.303
rm	0.459	0.055	0.241	8.393	0.000	0.351	0.566
age	-0.003	0.002	-0.064	-1.730	0.084	-0.006	0.000
dis	-0.150	0.022	-0.255	-6.703	0.000	-0.194	-0.106
rad	0.011	0.005	0.077	2.165	0.031	0.001	0.021
ptratio	-0.140	0.015	-0.237	-9.043	0.000	-0.171	-0.110
black	0.001	0.000	0.100	4.066	0.000	0.001	0.002
lstat	-0.070	0.007	-0.397	-10.603	0.000	-0.083	-0.057

✕ age

Backward Elimination: Step 3

Variable age Removed

Model Summary

R	0.886	RMSE	0.584
R-Squared	0.784	Coef. Var	9.867
Adj. R-Squared	0.780	MSE	0.341
Pred R-Squared	0.771	MAE	0.432

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	8.349	0.633		13.190	0.000	7.105	9.593
crim	-0.026	0.004	-0.179	-6.329	0.000	-0.033	-0.018
indus	-0.016	0.007	-0.085	-2.206	0.028	-0.029	-0.002
nox	-2.462	0.453	-0.231	-5.436	0.000	-3.352	-1.572
rm	0.440	0.054	0.231	8.196	0.000	0.335	0.546
dis	-0.136	0.021	-0.230	-6.512	0.000	-0.177	-0.095
rad	0.012	0.005	0.084	2.378	0.018	0.002	0.022
ptratio	-0.143	0.015	-0.243	-9.311	0.000	-0.174	-0.113
black	0.001	0.000	0.097	3.964	0.000	0.001	0.002
lstat	-0.074	0.006	-0.421	-12.075	0.000	-0.086	-0.062

No more variables satisfy the condition of p value = 0.05

Variables Removed:
✕ zn
✕ chas_
✕ age

Final Model Output

Model Summary

R	0.886	RMSE	0.584
R-Squared	0.784	Coef. Var	9.867
Adj. R-Squared	0.780	MSE	0.341
Pred R-Squared	0.771	MAE	0.432

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	zn	0.7868	0.7819	12.7315	872.7919	0.5813
2	chas_	0.7856	0.7811	13.5773	873.6961	0.5824
3	age	0.7842	0.7802	14.5871	874.7490	0.5836

We then implement a model by removing the above variables and the following summary is obtained,

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.3490378	0.6329838	13.190	< 2e-16
crim	-0.0255062	0.0040301	-6.329	5.67e-10
indus	-0.0155256	0.0070368	-2.206	0.0278
nox	-2.4617605	0.4528934	-5.436	8.70e-08
rm	0.4401863	0.0537054	8.196	2.28e-15
dis	-0.1358914	0.0208691	-6.512	1.88e-10
rad	0.0120561	0.0050698	2.378	0.0178
ptratio	-0.1434209	0.0154030	-9.311	< 2e-16
black	0.0013072	0.0003298	3.964	8.49e-05
lstat	-0.0739630	0.0061254	-12.075	< 2e-16

Residual standard error: 0.5836 on 480 degrees of freedom
Multiple R-squared: 0.7842, Adjusted R-squared: 0.7802
F-statistic: 193.9 on 9 and 480 DF, p-value: < 2.2e-16

The hypothesis H_{0j}: $\beta_j=0$ against H_{1j}: Not H_{0j} gets rejected for all the regressors. We can conclude this for all the regressors observing the p-values which is observed to be$>\alpha=0.05$

Also the hypothesis H_0: The linear regression model is not an appropriate model vs. H_1: Not H_0 gets rejected. We can conclude this after observing the p-value of the implemented MLR model which is observed to be $<\alpha=0.05$

Thus the model which we have implemented is an appropriate model.

The adj. R^2 for the model just after dropping tax turned out to be 0.7823 while the adj. R^2 after three other variables ,namely zn,age and chas , has turned out to be 0.7802. Thus the reduction in the value of the adj. R^2 turns out to be very small but the removal of these three variables leads to the reduction in the standard error of the other variables, thus smaller confidence intervals can be provided for the estimates of the coefficients at the same level of confidence.

From the model, the first thing that can be interpreted is that the average number of rooms is positively correlated with house price. This makes sense and should be expected. Second, there is also a positive correlation if the house is next to the Charles River. It is reasonable that more people would want to live closer to the river for the great view on offer and that this should raise the house prices. Similarly, negative correlations with crime rate and pupil-teacher ratios are also to be expected. People would prefer to live in areas that have less crime and where there is a low pupil-teacher ratio. An increased demand for houses in such areas would drive house prices up based on those factors. What is most interesting for the purpose of this report is how distance to the main employment centers and nitrogen oxide levels influence the house prices. There is a negative correlation for both, and this will be explored further in the conclusion.

❖ *Conclusion*

The goal of this report was to determine the neighborhood attributes that best explained variation in house pricing. Various statistical techniques were used to eliminate predictors and extraneous observations. In examining the final model, one finds – quite reasonably – that house prices are higher in areas with lower crime and lower pupil-teacher ratios. House prices also tend to be higher closer to the Charles River, and houses with more rooms are pricier. This report is interested in the neighborhood attributes of houses, so the number of rooms is not an important predictor. The most interesting factors to consider are nitrogen oxide levels and distance to the main employment centers. On the one hand, people would want to live close to their place of employment. Yet it is reasonable to suggest that pollution levels are higher as one moves closer to these main employment centers. Most importantly, when talking of pollution, it is not just nitrogen oxide levels that are higher, but also noise pollution levels. The regression model that was fitted shows that higher levels of pollution decrease house prices to a greater extent than distance to employment centers. This suggests that people would prefer to live further away from their place of employment if it meant lower levels of pollution, which is an interesting point to consider. On a concluding note, it is important to note that the data for this report was collected several decades ago. In the years since, there is no doubt that pollution levels have risen and it would be interesting to examine the ways in which that affects house pricing in Boston today.

❖ *Appendix (R Code)*

```
library(MASS)
library(caret)
library(car)
library(corrplot)
library(olsrr)

View(Boston)

data=data.frame(Boston)
qqnorm(data$medv)
aa=data[which(data$medv!=50.0),]
qqnorm(aa$medv)
attach(aa)
View(aa)

#Box-Cox Transformation
y=medv
BoxCoxTrans(y)
y_=((y^0.4)-1)/(0.4)

hist(y_,ylab = "Frequency", xlab="medv", main="Histogram of medv values")
qqnorm(y_)
b=data.frame(cbind(y_,aa[,-14]))
chas_=as.factor(chas)
model=lm(y_~crim+zn+indus+chas_+nox+rm+age+dis+rad+tax+ptratio+black+lstat,data=b)
summary(model)
ycap=predict(model)
plot(model)

#VIF Calculation
vif(model)

#Cook's distance
cooksd<-cooks.distance(model)
plot(cooksd)

model1=lm(y_~crim+zn+indus+chas_+nox+rm+age+dis+rad+ptratio+black+lstat,data=b)
plot(model1)
summary(model1)

#Forward Selection Method
fd<-ols_step_forward_p(model1)
plot(fd)
ols_step_forward_p(model1,details = TRUE)

#Backward Elimination Method
bd<-ols_step_backward_p(model1)
ols_step_backward_p(model1,details = TRUE)
plot(bd)
```

```
#Stepwise Selection Method
ols_step_both_p(model1)
stp<-ols_step_both_p(model1,details = TRUE)
plot(stp)

#AIC
ols_step_forward_aic(model1)
ols_step_backward_aic(model1)

model2=lm(y_~crim+indus+nox+rm+dis+rad+ptratio+black+lstat,data=b)
plot(model2)
summary(model2)
```

❖ *Acknowledgement:*

A project is a golden opportunity for learning. We consider ourselves very lucky and honoured to have so many wonderful people lead us through this attempt. Our grateful thanks to Dr. Sharmishtha Mitra, Associate Professor, Department. of Mathematics and Statistics, IIT Kanpur. She always helped us in every possible way whenever we faced a difficulty. We also thank our classmates and seniors for their support and solidarity. Thank you all.

❖ *References*

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management 5 , 81–102.

Belsley D.A., Kuh, E. and Welsch, R.E. (1980)Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Faraway, J. Linear Models with R. Chapman & Hall/CRC Texts in Statistical Science (Book 63).

http://math.furman.edu/~dcs/courses/math47/R/library/car/html/outlier.test.html

https://en.wikipedia.org/wiki/Cook%27s_distance