ALY6015 – INTERMEDIATE ANALYTICS

BY

INSTRUCTOR: VALERIY SHEVCHENKO

FINAL PROJECT: BOSTON HOUSING PRICES

NORTHEASTERN UNIVERSITY

BY

ANITA PREKO

SWAPNIL NANDKISHOR SAKORKAR

MRIGANKSHANKAR SINGH

JUNE 28, 2019

INTRODUCTION

Housing has been one of the most important factors since we as humans understood the worth of 'Shelter' under our 3 basic necessities comprising of food and clothing to be the other two. Over the years this has changed from shelter to luxury and standard of living. The structure of the house (Interior and exterior), the room style and size, the location, if it has modern amenities like modern kitchen, bathrooms and now a days devices working on IoT. As these necessities shift towards luxury, it directly proportional to the prices of these houses.

With the chosen dataset, we have tried to use similar techniques but by comparing different datasets. Our dataset is called Boston House Price which is maintained by Carnegie Mellon University but is freely available to everyone. This dataset comprises of 506 observations in the data for 14 variables including the median price of houses in Boston. There are 12 numerical variables in our dataset and 1 categorical variable. The following screenshot can help us understand all the required variables in greater detail

1. CRIM – per capita crime rate by town
2. ZN – proportion of residential land zoned for lots over 25,000 sq.ft
3. INDUS – proportion of non-retail business acres per town
4. CHAS – Charles River dummy variable (1 if tract bounds river; else 0)
5. NOX – nitric oxides concentration (parts per 10 million)
6. RM – average number of rooms per dwelling
7. AGE – proportion of owner-occupied units built prior to 1940
8. DIS – weighted distances to five Boston employment centres
9. RAD – index of accessibility to radial highways
10. TAX – full-value property-tax rate per $10,000
11. PTRATIO – pupil-teacher ratio by town
12. B – 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT – % lower status of the population
14. MEDV – Median value of owner-occupied homes in $1000's

To analyze and statistically regress our data, we have used various factors that can provide us with the required results. Factors such as correlation, Linear Modelling, RMSE and Coefficient of determination (R^2) with the P value was used for analysis.

ANALYSIS

To begin the analysis of the Boston housing data, we have to explore the data frame after loading into using read.csv in R. Using the function str() we can display the structure of the data frame like the number of observations and variables, names and class of each column, and sample values from each column. To get more statistical information, the summary function() is used to display the summary statistics such as minimum value, maximum, median, mean, and the 1st and 3rd quartile values. Also using na.omit() function helps to eliminate any null values in the dataset.

```
Install package 'caret', ggplot, lattice
Load Library 'caret'
library(ggplot2)
library(lattice)
library(caret)
 Read the dataset "housingdata file using read.csv
housing.df <- read.csv(file.choose(), header = T)

Cleaning the data my omitting the NA values and removing the column CHAS
housing.df <- na.omit(housing.df)

create new dataset without missing data
housing.df

This Displays the housing dataframe
head(housing.df)
```

```
> head(housing.df)
     CRIM ZN INDUS CHAS   NOX    RM  AGE    DIS RAD TAX PTRATIO      B LSTAT MEDV
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90    NA 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
>
```

Display the structure of the housing.df data frame

str(housing.df)

```
> # Display the structure of the housing.df data frame
> str(housing.df)
'data.frame':	506 obs. of  14 variables:
 $ CRIM   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ ZN     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ INDUS  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ CHAS   : int  0 0 0 0 0 0 NA 0 0 NA ...
 $ NOX    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ RM     : num  6.58 6.42 7.18 7 7.15 ...
 $ AGE    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ DIS    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ RAD    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ TAX    : int  296 242 242 222 222 222 311 311 311 311 ...
 $ PTRATIO: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ B      : num  397 397 393 395 397 ...
 $ LSTAT  : num  4.98 9.14 4.03 2.94 NA ...
 $ MEDV   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
>
```

This Displays the summary statistics like minimum value, maximum value, median, mean, and the 1st and 3rd quartile values for each column in our dataset.

summary(housing.df)

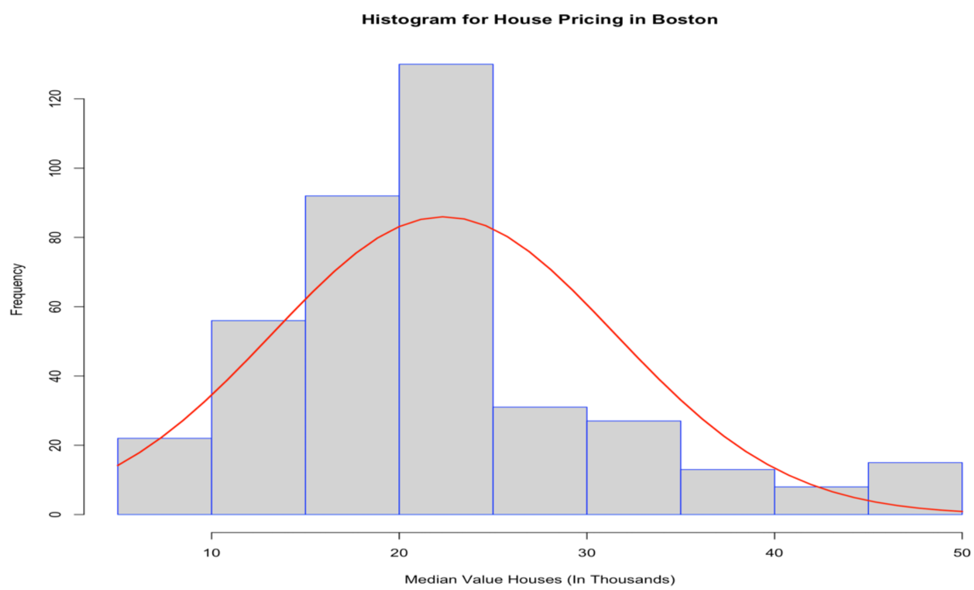
```
> #Displays the summary statistics like minimum value, maximum value, median, mean, and the 1st and 3rd
> #quartile values for each column in our dataset.
> summary(housing.df)
      CRIM               ZN              INDUS            CHAS              NOX               RM       
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561  
 1st Qu.: 0.08190   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886  
 Median : 0.25372   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208  
 Mean   : 3.61187   Mean   : 11.21   Mean   :11.08   Mean   :0.06996   Mean   :0.5547   Mean   :6.285  
 3rd Qu.: 3.56026   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780  
 NA's   :20         NA's   :20       NA's   :20      NA's   :20                                        
      AGE              DIS              RAD              TAX           PTRATIO            B              LSTAT       
 Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.730  
 1st Qu.: 45.17   1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 7.125  
 Median : 76.80   Median : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.430  
 Mean   : 68.52   Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.715  
 3rd Qu.: 93.97   3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.955  
 Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.970  
 NA's   :20                                                                                          NA's   :20      
      MEDV      
 Min.   : 5.00  
 1st Qu.:17.02  
 Median :21.20  
 Mean   :22.53  
 3rd Qu.:25.00  
 Max.   :50.00  
```

To better understand more about the data, two data visualization types were chosen which are the histogram and a boxplot. The most important variable for the dataset is the MDEV, which is the median value of the house in $1000's which is also the dependent variable in our model.

The histogram of the MEDV shows that the median value of housing price is skewed to the right, with a number of outliers to the right. Also, the data shows a normal distribution with a center around $22,000.

Code:

```
 visualize the distribution and density of the outcome, MEDV. The black curve represents the density
This calculates the mean of housing MEDV
m <- mean(housing.df$MEDV)
This calculates the standard deviation of housing MEDV
sd <-sd(housing.df$MEDV)
This creates a histogram of the MEDV
h <-hist(housing.df$MEDV,
    main="Histogram for House Pricing in Boston",
    xlab="Median Value Houses (In Thousands)",
    border="blue",
    col="lightgray",
    breaks=12
   )
This adds a curve to the histogram while maintaining the count of data instead of density
xfit<-seq(min(housing.df$MEDV),max(housing.df$MEDV),length=40)
#This gives the density of MEDV,
yfit<-dnorm(xfit,mean=m,sd=sd)
#This calculates the size of the bins
yfit <- yfit*diff(h$mids[1:2])*length(housing.df$MEDV)
#this adds a curve to the hsitogram to show normal distribution of data.
lines(xfit, yfit, col="red", lwd=2)
```

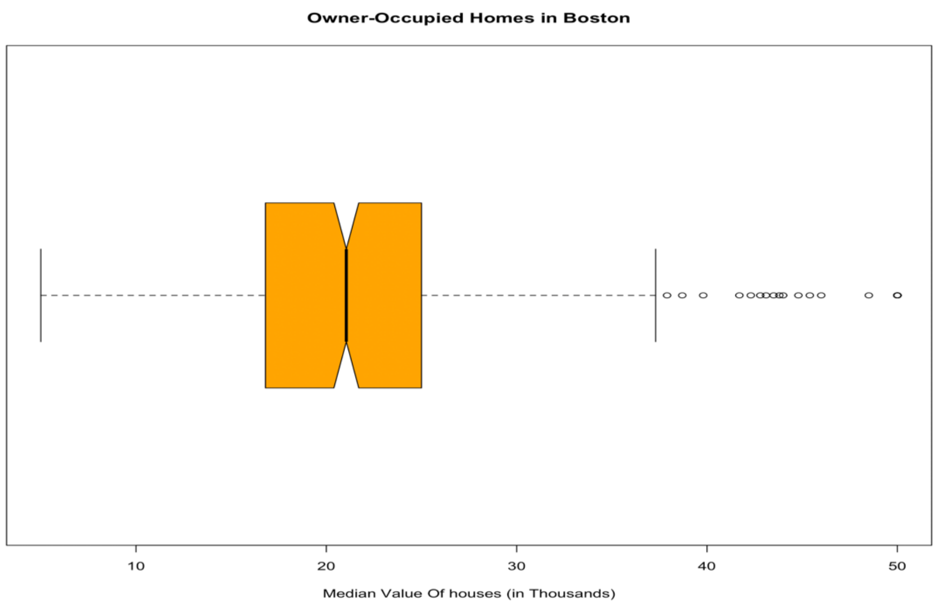


Also, the boxplot shows that there are a lot of outliers in the data.

Some of the independent variables such as INDUS, NOX, RM, PTRATIO, LSAT, MEDV, can be plotted against the dependent variable (MEDV). We see that there is strong positive or negative correlation between these variables and the outcome MEDV.

Code:

```
the boxplot is also plotted to bring an additional perspective of MEDV
boxplot(housing.df$MEDV,data=housing.df,
        main="Owner-Occupied Homes in Boston",
        xlab="Median Value Of houses (in Thousands)",
        horizontal = T,
        col = "Orange",
        notch = TRUE)
```

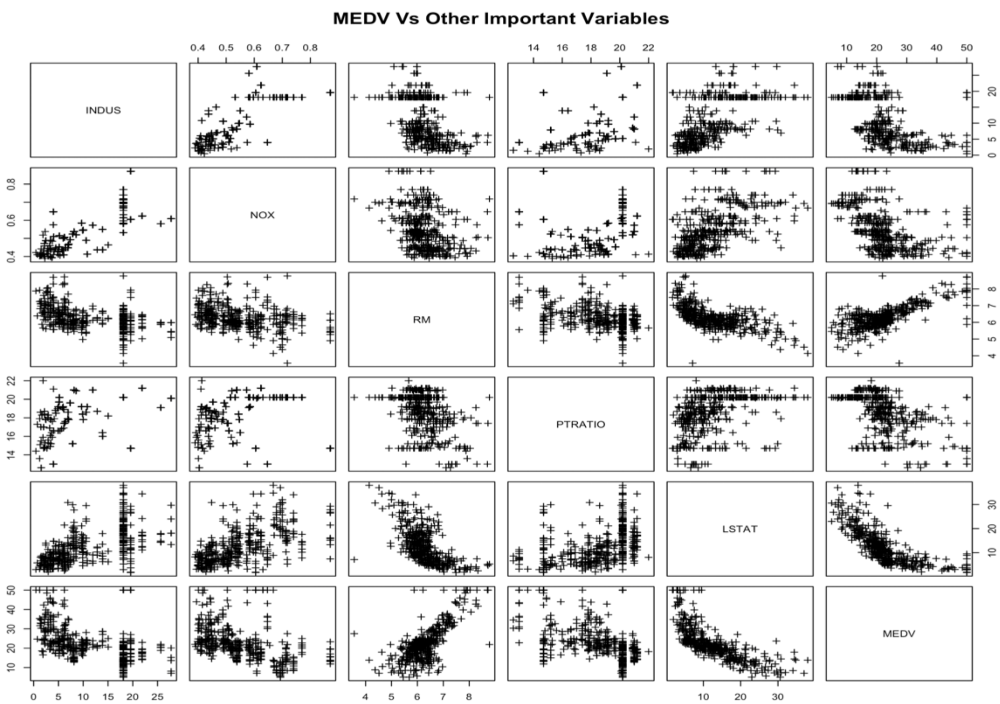


Among these independent variables which of them is correlated to the dependent variable MEDV? To answer this, we can perform the correlation function in R.

Code:

This creates a scatterplot of some of the important variables (based on intuition) with the outcome variable MEDV.

```
plot(housing.df[,c(3,5,6,11,13,14)],
    pch=3,
    main = ("MEDV Vs Other Important Variables"))
```



Correlation of each independent variable with the dependent variable

cor(housing.df,housing.df$MEDV)

```
> # Correlation of each independent variable with the dependent variable
> cor(housing.df,housing.df$MEDV)
              [,1]
CRIM    -0.3972301
ZN       0.4068215
INDUS   -0.5108292
CHAS     0.1737012
NOX     -0.4590543
RM       0.7239508
AGE     -0.4074705
DIS      0.2795469
RAD     -0.4166377
TAX     -0.5088643
PTRATIO -0.5438090
B        0.3472561
LSTAT   -0.7434496
MEDV     1.0000000
>
```

We can Infer that the number of rooms RM has the strongest positive correlation with the median value of the housing price, while the percentage of lower status population, LSTAT and the pupil-teacher ratio, PTRATIO, have strong negative correlation.

Before running a model, such as regression (predicting a continuous variable) or classification (predicting a discrete variable), on data, you almost always want to do some preprocessing. Therefore, before partitioning the data into training and test data, we scaled the independent features against MEDV using the cbind() in R.

```
We partition the data on a 7/3 ratio as training/test datasets.
set.seed(12345)

housing.df <- cbind(scale(housing.df[1:13]), housing.df[14])
Do data partitioning
inTrain <- createDataPartition(y = housing.df$MEDV, p = 0.70, list = FALSE)
Partition data into training data
training <- housing.df[inTrain,]
Partition data into testing data
testing <- housing.df[-inTrain,]
```

Once a model has been trained on a given set of data, it can now be used to make predictions on new sets of input data. We generalized the linear regression model with MEDV as the dependent variable and all the remaining variables as independent variables. We train the model with the training dataset using linear model 2, which is the log transformation of MEDV. Using the predict() function in R, we predicted the outcome (MEDV) for the testing dataset and viewed the coefficients of all the independent variables in (2 decimal places) after performing the linear regression.

```
 Perform  linear regression model with MEDV as the dependent variable
and all the remaining variables as independent variables.
set.seed(12345)
Try linear model using all features
fit.lm <- lm(log(MEDV)~.,data = training)
```

This prints the coefficients of the x variables of Boston housing dataset

data.frame(coef = round(fit.lm$coefficients,2))

```
> #This prints the coefficients of the x variables of boston housing dataset
> data.frame(coef = round(fit.lm$coefficients,2))
             coef
(Intercept)  3.02
CRIM        -0.09
ZN           0.03
INDUS        0.02
CHAS         0.03
NOX         -0.10
RM           0.08
AGE         -0.01
DIS         -0.09
RAD          0.11
TAX         -0.12
PTRATIO     -0.09
B            0.02
LSTAT       -0.17
```

set.seed(12345)

predict on test set

pred.lm <- predict(fit.lm, newdata = testing)

Root-mean squared error

rmse.lm <- sqrt(sum((exp(pred.lm) - testing$MEDV)^2)/length(testing$MEDV))

This prints the RMSE, R2 and P-value of the predicted test data

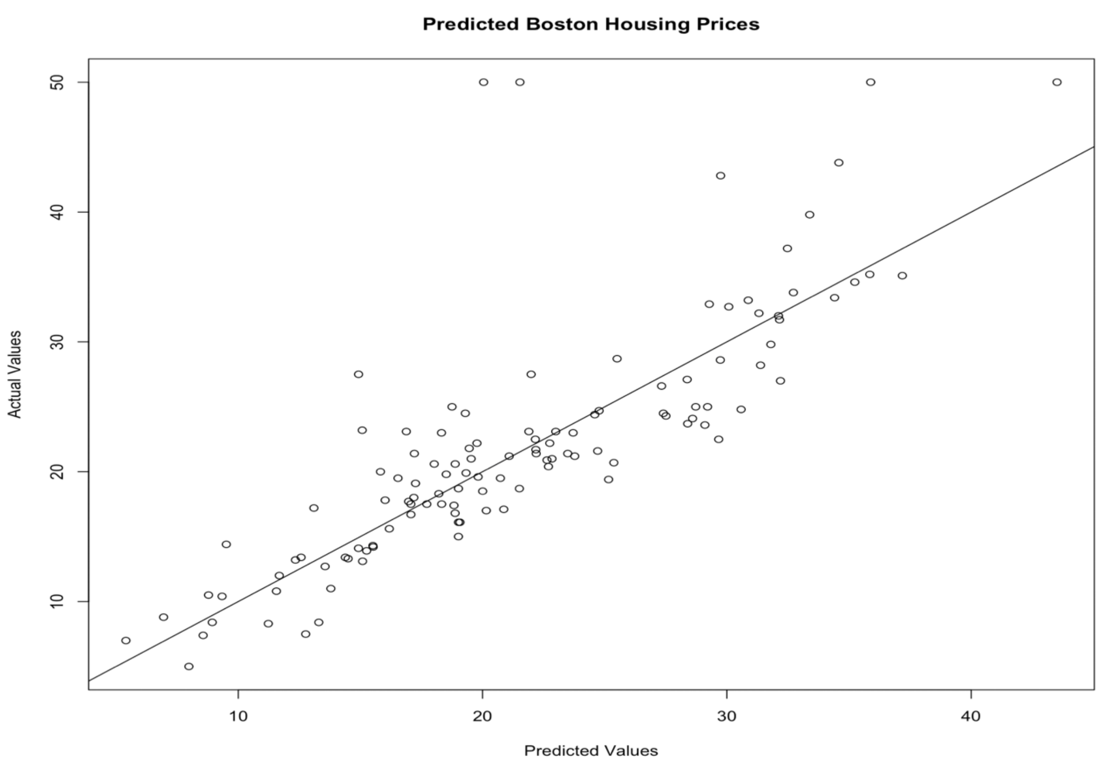
c(RMSE = rmse.lm, R2 = summary(fit.lm)$r.squared, P_value = summary(fit.lm)$coefficients[1,4])

```
> #This prints the RMSE, R2 and P-value of the predicted test data
> c(RMSE = rmse.lm, R2 = summary(fit.lm)$r.squared, P_value = summary(fit.lm)$coefficients[1,4])
     RMSE        R2   P_value
5.3381196 0.8217427 0.0000000
>
```

After predicting the data for the test data, we see that the RMSE is 5.338 and the R^2 value is 0.821 for this model. This shows that 82% of the variance in the True Value is predictable from the Prediction. As this is a very high percentage, we can call this model to be a successful model. The calculated p-value for each independent variable in the linear model that is less than 0.05 significant value shows which variables are not contributing significantly for the model, due to multicollinearity. Multicollinearity shows that some independent variables in the regression model are correlated while they should have been independent. And these predictors that are not statistically significant can be removed from the model.

Plot of predicted price vs actual price

```
plot(exp(pred.lm),testing$MEDV,
    main = "Predicted Boston Housing Prices",
    xlab = "Predicted Values",
    ylab = "Actual Values",
    abline(a = 0, b = 1))
```



Plotting the predicted price to the Actual price shows a positive correlation . Therefore, these show that the selected variables , RM, CRIM,CHAS ,NOX ,RM ,DIS , PTRATIO , RAD ,LSTAT affect the housing prices in Boston.

summary of the regression model

summary(fit.lm)

```
> #summary of the regression model
> summary(fit.lm)

Call:
lm(formula = log(MEDV) ~ ., data = training)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63135 -0.10008 -0.00679  0.10029  0.77855 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.02184    0.01066 283.513  < 2e-16 ***
CRIM        -0.09229    0.01421  -6.496 4.11e-10 ***
ZN           0.02809    0.01594   1.762  0.07929 .  
INDUS        0.02322    0.02082   1.115  0.26584    
CHAS         0.02948    0.01099   2.684  0.00774 ** 
NOX         -0.09863    0.02265  -4.355 1.91e-05 ***
RM           0.07977    0.01662   4.800 2.67e-06 ***
AGE         -0.00911    0.01860  -0.490  0.62475    
DIS         -0.09084    0.02054  -4.424 1.42e-05 ***
RAD          0.11151    0.02683   4.156 4.38e-05 ***
TAX         -0.12473    0.02899  -4.302 2.38e-05 ***
PTRATIO     -0.09238    0.01391  -6.640 1.79e-10 ***
B            0.01632    0.01315   1.242  0.21540    
LSTAT       -0.16637    0.02007  -8.290 5.87e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1762 on 263 degrees of freedom
Multiple R-squared:  0.8217,	Adjusted R-squared:  0.8129 
F-statistic: 93.26 on 13 and 263 DF,  p-value: < 2.2e-16

> 
```

The predictors that are not statistically significant can be removed from the model. These are indus- proportion of non-retail business acres per town, age-– proportion of owner-occupied units built prior to 1940, B-1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town and zn- proportion of residential land zoned for lots over 25,000 sq.ft. Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between Variables with asterisks and MEDV.

CONCLUSION

During the analysis, our objective was to evaluate various techniques and methods correlating to the highly influential variables for housing prices in Boston. In order to do so we have used linear regression model along with log transformations for enhanced precision. We observed that the coefficient of determination (R^2) is 82% which in turn denotes that the variable; number of rooms (RM) has a high correlation to the median value of the house. According to our research, no other variable has higher significance compared to (RM) rooms. As it is also directly proportional to the median value i.e. higher the number of rooms, the prices of the houses will also experience an increase. On the basis of the final model our further analysis concludes that areas with low crime rate (CRIM) and pupil-teacher ratios (PTRATIO) in the nearby schools have higher housing prices.

Another significant observation was that houses near the employment centers (DIS had high prices which may be because of easier proximity to work. On a concluding note however, this data was collected decades ago, and it would be interesting to see the influence of other factors such as nitric oxide concentration and accessibility to highways on the housing prices in the city.

REFERENCES

Boston Housing. (n.d.). Retrieved from https://www.kaggle.com/c/boston-housing

Prabhakaran, S. (n.d.). Eval(ez_write_tag([[728,90],'r_statistics_co-box-3','ezslot_4',109,'0']));Linear Regression. Retrieved from http://r-statistics.co/Linear-Regression.html