# Capstone Project – Real Estate

## Boston Housing Dataset

# Contents

# Overview of the project

- **The objective of our project was to understand the drivers behind the value of houses in Boston and arrive at data-driven recommendations on how the client can increase the value of housing.**

- **The housing dataset contains 506 observations of 14 variables.**

- **As the dependent variable MEDV (Median prices) is continuous, we had implemented the Multiple Linear regression approach.**

- **The project document is organized to demonstrate the entire process right from: Exploring the data, Cleaning the data, model building, prediction, model performance and understanding the importance of various features in influencing the housing prices.**

# Data Description

| Variable | Description |
|---|---|
| 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 oxide 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 |

# Data Exploration

```
> names(boston)
 [1] "CRIM"    "ZN"      "INDUS"   "CHAS"    "NOX"     "RM"      "AGE"     "DIS"     "RAD"
[10] "TAX"     "PTRATIO" "B"       "LSTAT"   "MEDV"
```

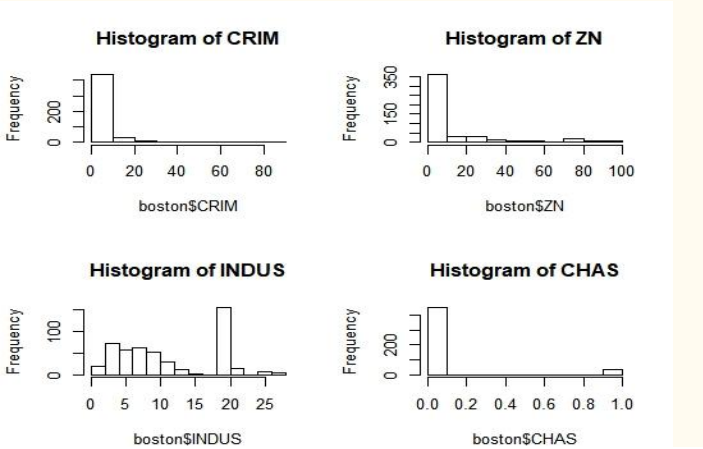
```
> head(boston)
     CRIM ZN INDUS CHAS   NOX    RM  AGE    DIS RAD TAX PTRATIO      B LSTAT MEDV
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 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
```

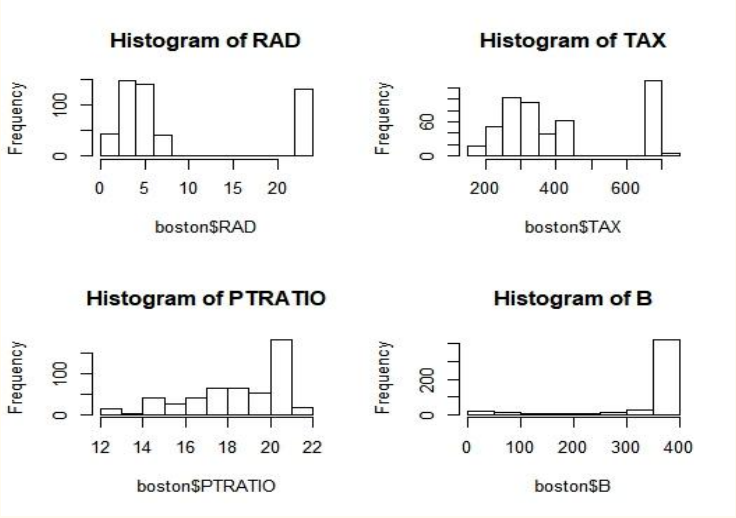
```
> str(boston)
'data.frame':   506 obs. of  14 variables:
 $ CRIM   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ ZN     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ INDUS  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ CHAS   : int  0 0 0 0 0 0 NA 0 0 NA ...
 $ NOX    : 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 ...
```

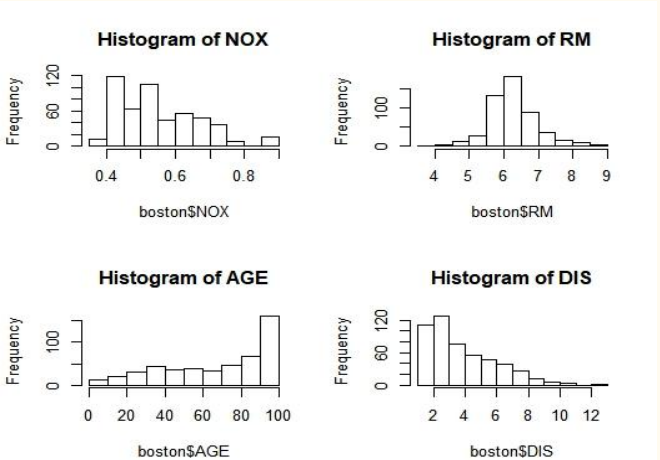
# Data Exploration – cont'd

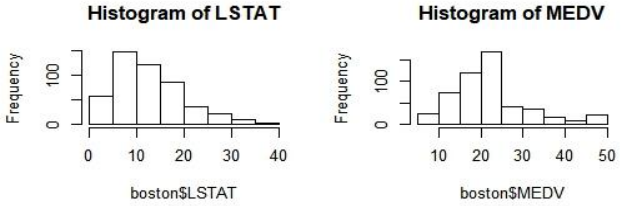
```
> summary(boston)
      CRIM                ZN              INDUS             CHAS              NOX        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850  
 1st Qu.: 0.08190   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490  
 Median : 0.25372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

# Data Exploration – Histograms

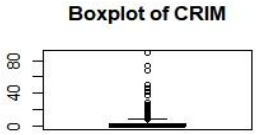


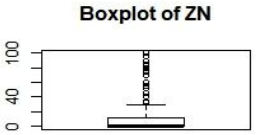



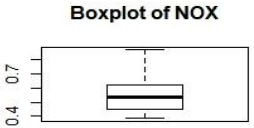



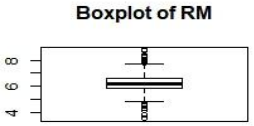



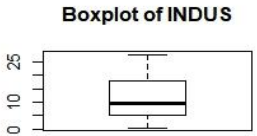
# Data Exploration – Boxplots

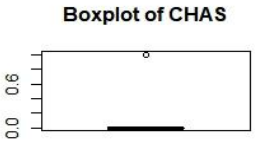


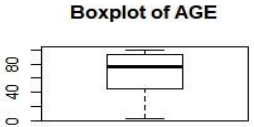



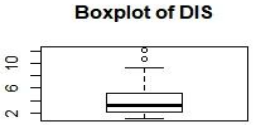



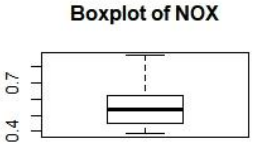



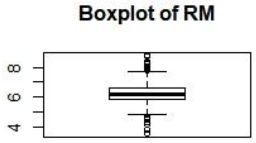



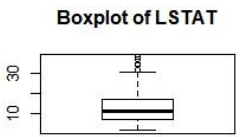



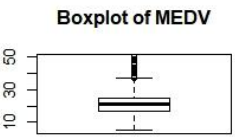



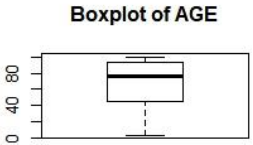



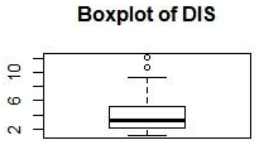



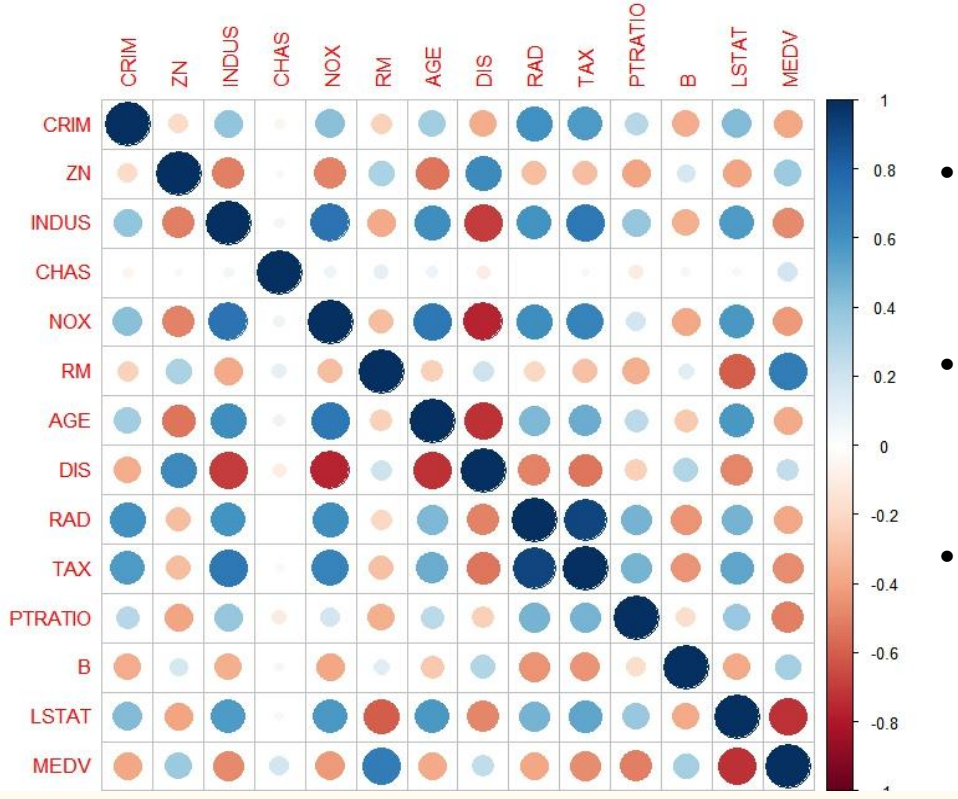



# Data Exploration – Study of Potential Variables

| CRIME RATE PER TOWN (CRIM) | |
|---|---|
| CRIM < 10 | 435 Observations |
| CRIM > 50 | 4 Observations |
| CRIM > 80 | 1 Observation |
| **AVERAGE NUMBER OF ROOMS (RM)** | |
| RM > 7 | 64 Observations |
| RM > 8 | 13 Observations |
| RM < 4 | 2 Observations |
| RM < 3 | None |
| **BOUND THE CHARLES RIVER OR NOT (CHAS)** | |
| CHAS = 1 | 32 Observations |
| CHAS = 0 | 452 Observations |
| **FULL-VALUE PROPERTY TAX RATE PER $ 10,000 (TAX)** | |
| TAX > 600 | 137 Observations |
| TAX < 600 | 69 Observations |

# Data Exploration – Correlation Analysis

```
> boston_new <- boston
> boston_cor <- cor(boston_new) #All 14 variables
> corrplot(boston_cor, method = "circle")
```




- **Positive correlations are displayed in blue and negative correlations in red colour.**
- **Colour intensity and the size of the circle are proportional to the correlation coefficients.**
- **Variables are highly correlated to itself.(example: CRIM with CRIM, ZN with ZN, etc.)**

# Data Preparation – Missing Value Treatment

```
> boston_new$CRIM[which(is.na(boston_new$CRIM))] <- median(boston_new$CRIM, na.rm= T)
> boston_new$ZN[which(is.na(boston_new$ZN))] <- median(boston_new$ZN, na.rm= T)
> boston_new$INDUS[which(is.na(boston_new$INDUS))] <- median(boston_new$INDUS, na.rm= T)
> boston_new$CHAS[which(is.na(boston_new$CHAS))] <- median(boston_new$CHAS, na.rm= T)
> boston_new$AGE[which(is.na(boston_new$AGE))] <- median(boston_new$AGE, na.rm= T)
> boston_new$LSTAT[which(is.na(boston_new$LSTAT))] <- median(boston_new$LSTAT, na.rm= T)
```

```
> summary(boston_new)
      CRIM                ZN              INDUS             CHAS              NOX        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850  
 1st Qu.: 0.08324   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490  
 Median : 0.25372   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380  
 Mean   : 3.47914   Mean   : 10.77   Mean   :11.03   Mean   :0.06719   Mean   :0.5547  
 3rd Qu.: 2.80872   3rd Qu.:  0.00   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710  
       RM              AGE              DIS              RAD              TAX        
 Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.   :187.0  
 1st Qu.:5.886   1st Qu.: 45.92   1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0  
 Median :6.208   Median : 76.80   Median : 3.207   Median : 5.000   Median :330.0  
 Mean   :6.285   Mean   : 68.85   Mean   : 3.795   Mean   : 9.549   Mean   :408.2  
 3rd Qu.:6.623   3rd Qu.: 93.58   3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0  
 Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.   :711.0  
    PTRATIO             B              LSTAT             MEDV      
 Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 7.23   1st Qu.:17.02  
 Median :19.05   Median :391.44   Median :11.43   Median :21.20  
 Mean   :18.46   Mean   :356.67   Mean   :12.66   Mean   :22.53  
 3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.57   3rd Qu.:25.00  
 Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00  
```

**NA values in variables replaced by their Median values**

# Data Preparation – Outlier Treatment

```
#OUTLIER TREATMENT USING IQR
#creating a vector containing names of variables we wish to remove the outliers if present.
variables <- c("CRIM", "ZN", "INDUS", "CHAS", "NOX", "RM", "AGE", "DIS", "RAD",
               "TAX", "PTRATIO", "B", "LSTAT", "MEDV")

#creating the object outliers to store the row id's containing  outliers for removal
outliers <-  c()

#creating a boundary for each variable (0.95 and 0.05 for 3 S.D from the mean)
#To loop through the columns specified
for(i in variables){

  #Get the min/max values(Boundaries for each variable)
  max <- quantile(boston_new[,i], 0.95, na.rm=TRUE) + (IQR(boston_new[,i],na.rm=TRUE))
  min <- quantile(boston_new[,i],0.05,na.rm=TRUE) -  (IQR(boston_new[,i],na.rm=TRUE))

  #Get row ids which contain outliers
  id <- which(boston_new[,i] < min| boston_new[,i] > max)


  #Print the number of outliers in each variable
  print(paste(i, length(id), sep =''))

  #Append the outliers list
  outliers <- c(outliers, id)

} #loop closure

#sorting the outliers
outliers <- sort(outliers)

#remove the outliers from the dataset
boston_new <- boston_new[-outliers,]
```

# Data Preparation – Outlier Treatment

- **Outlier /extreme values in data set are identified as it will change fit estimates and predictions**
- **Data has been cleaned by eliminating rows that contain the outliers.**
- **Summarized view of data after outlier treatment is shown below.**

```
> summary(boston_new)
      CRIM                ZN              INDUS        CHAS          NOX               RM       
 Min.   : 0.00632   Min.   : 0.000   Min.   : 0.46   0:414   Min.   :0.3850   Min.   :4.880  
 1st Qu.: 0.08265   1st Qu.: 0.000   1st Qu.: 5.19   1: 31   1st Qu.:0.4490   1st Qu.:5.889  
 Median : 0.22969   Median : 0.000   Median : 8.56           Median :0.5240   Median :6.195  
 Mean   : 1.84593   Mean   : 9.312   Mean   :10.71           Mean   :0.5473   Mean   :6.278  
 3rd Qu.: 1.35472   3rd Qu.: 0.000   3rd Qu.:18.10           3rd Qu.:0.6090   3rd Qu.:6.579  
 Max.   :17.86670   Max.   :80.000   Max.   :27.74           Max.   :0.8710   Max.   :8.297  

      AGE              DIS              RAD             TAX           PTRATIO          B        
 Min.   :  2.90   Min.   : 1.130   5       :109   Min.   :188.0   Min.   :12.60   Min.   : 68.95  
 1st Qu.: 45.80   1st Qu.: 2.222   4       :107   1st Qu.:277.0   1st Qu.:17.40   1st Qu.:377.51  
 Median : 76.50   Median : 3.411   24      : 90   Median :311.0   Median :18.70   Median :392.18  
 Mean   : 67.85   Mean   : 3.882   3       : 35   Mean   :388.9   Mean   :18.41   Mean   :372.98  
 3rd Qu.: 92.90   3rd Qu.: 5.231   6       : 26   3rd Qu.:432.0   3rd Qu.:20.20   3rd Qu.:396.33  
 Max.   :100.00   Max.   :10.710   8       : 22   Max.   :711.0   Max.   :22.00   Max.   :396.90  
                                   (Other): 56                                                    
     LSTAT            MEDV      
 Min.   : 1.73   Min.   : 6.30  
 1st Qu.: 7.39   1st Qu.:18.10  
 Median :11.43   Median :21.50  
 Mean   :12.15   Mean   :22.82  
 3rd Qu.:15.70   3rd Qu.:25.00  
 Max.   :34.77   Max.   :50.00  
```

# Data Preparation – cont'd

```
#Convert variables CHAS and RAD from integer to factor
boston_new$CHAS <- as.factor(boston_new$CHAS)
boston_new$RAD <- as.factor(boston_new$RAD)
levels(boston_new$RAD) <- c(1,2,3,4,5,6,7,8,24)
```

```
> str(boston_new)
'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   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ 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    : Factor w/ 9 levels "1","2","3","4",..: 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 11.43 ...
 $ MEDV   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

**CHAS and RAD are now factors.**

## Model Building: Data partition

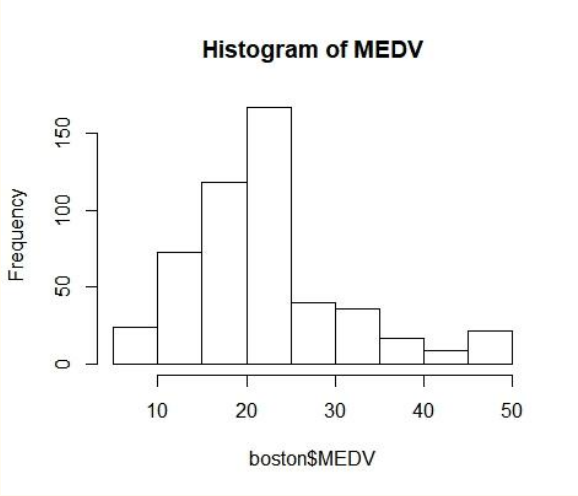
- **The dataset now contains 445 observations of 14 variables.**
- **Splitting the data set into train and test data**

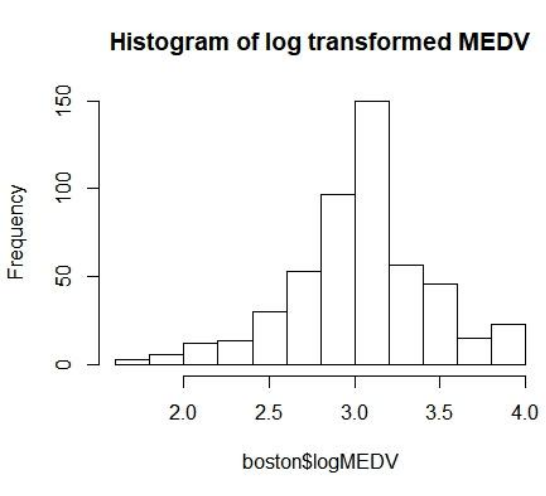
```
#Split the dataset boston_new into train and test data (70% for train and 30% for test)
set.seed(1234)
training <- sample(1:nrow(boston_new), 0.7*nrow(boston_new))
train <- boston_new[training,]
test <- boston_new[-training,]
```

- **The train dataset is containing 311 observations, while the test dataset contains 134 observations. The model will be trained**

# Model Building : Log transformation of MEDV

- **Why to log transform the MEDV variable?**
- **Log transformation is done to transform a skewed distribution to a normal distribution. This is also a common practise to tackle heteroskedasticity**
- **Based on the approaches we had taken, there was considerable improvement in model performance when MEDV was log transformed.**
- **The following histograms shows you that log transformation of MEDV has transformed its earlier skewed distribution to a normal distribution.**



# Model Building : Multiple linear regression model

- **We build the first model "fit" with log transformed MEDV and include all the independent variables.**

```
#model building with all the independant variables
fit <- lm(log(MEDV) ~., data= train)
summary(fit) #Adjusted R-squared:  0.7652

#model building by dropping insignificant variables ZN, INDUS, AGE
fit1 <- update(fit, ~. - ZN - INDUS - AGE)
summary(fit1)#Adjusted R-squared:  0.7664

#checking for multicollinearity
vif(fit1, th=5)#RAD and TAX have VIF greater than 5
```

```
> #checking for multicollinearity
> vif(fit1, th=5)#RAD and TAX have VIF greater than 5
             GVIF Df GVIF^(1/(2*Df))
CRIM     4.172004  1        2.042548
CHAS     1.096606  1        1.047190
NOX      3.907282  1        1.976685
RM       1.931618  1        1.389827
DIS      2.890218  1        1.700064
RAD     18.598792  8        1.200446
TAX      6.064420  1        2.462604
PTRATIO  1.760372  1        1.326790
B        1.230293  1        1.109186
LSTAT    2.388040  1        1.545329
```

- **Muliticollinearity present in RAD and TAX. They have VIF greater than 5.**
- **Will build the next model by dropping variables RAD and TAX**

# Model Summary

```r
#model building by dropping insignificant variables RAD - TAX)
fit2 <- update(fit1, ~. - RAD - TAX)
summary(fit2)#Adjusted R-squared:  0.7482
```

```
> summary(fit2)#Adjusted R-squared:  0.7482

Call:
lm(formula = log(MEDV) ~ CRIM + CHAS + NOX + RM + DIS + PTRATIO +
    B + LSTAT, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69688 -0.08664 -0.01595  0.08271  0.91544

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.9661804  0.2570955  11.537  < 2e-16 ***
CRIM        -0.0133970  0.0041102  -3.259  0.00124 **
CHAS1        0.1588957  0.0396347   4.009 7.69e-05 ***
NOX         -0.4567497  0.1623203  -2.814  0.00522 **
RM           0.1809727  0.0218266   8.291 3.73e-15 ***
DIS         -0.0316789  0.0075661  -4.187 3.71e-05 ***
PTRATIO     -0.0277677  0.0051388  -5.404 1.33e-07 ***
B            0.0004492  0.0001944   2.311  0.02150 *
LSTAT       -0.0247731  0.0023363 -10.604  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1763 on 302 degrees of freedom
Multiple R-squared:  0.7547,    Adjusted R-squared:  0.7482
F-statistic: 116.1 on 8 and 302 DF,  p-value: < 2.2e-16
```

# Performance of the Model

```
#Model prediction
predict_train <- predict(fit2, train) #for known data
predict_test <- predict(fit2, test) #for unseen data

#Calculating Mean square error as performance metrics for regression
mse_train <- mean((exp(predict_train) - train$MEDV)^2)
mse_train #16.81506

mse_test <- mean((exp(predict_test) - test$MEDV)^2)
mse_test # 21.24562
```

```
> #Calculating Mean square error as performance metrics for regression
> mse_train <- mean((exp(predict_train) - train$MEDV)^2)
> mse_train #16.81506
[1] 16.81506
> mse_test <- mean((exp(predict_test) - test$MEDV)^2)
> mse_test # 21.24562
[1] 21.24562
```

# Performance of the Model (cont'd)

- Comparing our model to a model without log transformed MEDV, we discovered the following:

| Performance metric | Model with log transformed MEDV | Model without log transformed MEDV |
|---|---|---|
| Mean Square Error(MSE) | 21.2456 | 24.1225 |
| Adjusted R-squared value | 0.7482 | 0.7093 |

- The MSE value and Adjusted R-squared value are better in our model compared to a model without log transformation of MEDV.

# Model Diagnostics

- To check If the model satisfies the assumptions of linear regression
- Assumption: Errors are not autocorrelated.
- We employ durbinWatsonTest from car package.

```
> durbinWatsonTest(fit2) #DW statistic is 2.067375
 lag Autocorrelation D-W Statistic p-value
   1     -0.03479875       2.067375   0.562
 Alternative hypothesis: rho != 0
```

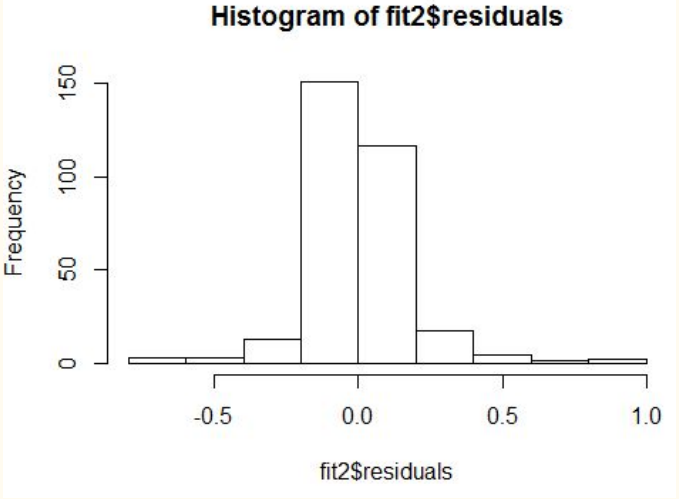
- Null hypothesis says errors are not autocorrelated. They are independent. While Alternate hypothesis says errors are autocorrelated.
- Since the p value is greater than 0.05, we satisfy Null hypothesis. In other words, we failed to reject the Null hypothesis.
- The DW statistic is close to 2. We satisfied this assumption.

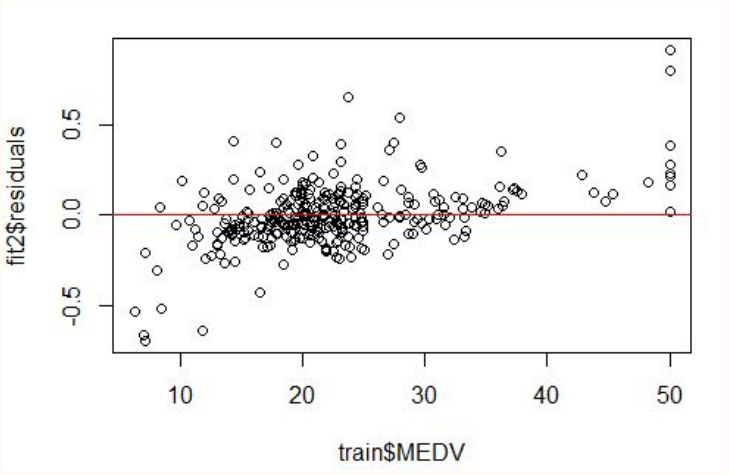
# Model Diagnostics (cont'd)

**Other assumptions have been executed below and the plots are shown below:**

- **Assumption: Errors are normally distributed.**

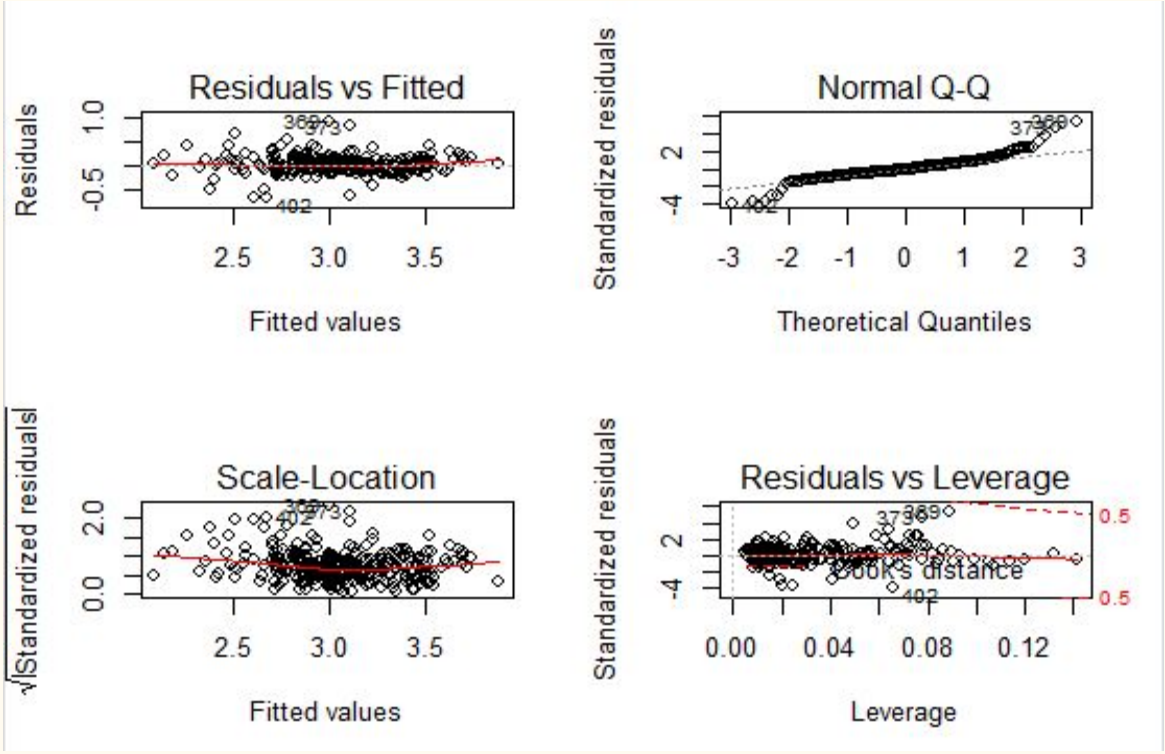
```
hist(fit2$residuals) #Errors are normally distributed

plot(train$MEDV , fit2$residuals) #Variance is constant i.e Homoscedasticity of variance
abline(h=0, col = "red")
```



# Model Diagnostics (cont'd)

```
par(mfrow = c(2,2))
plot(fit2)
```



## Final Conclusion of the Project- Insights derived and Recommendation

- Following independent variables were the most significant based on the summary of the model.

| SIGNIFICANT VARIABLES | DESCRIPTION | ESTIMATED COEFFICENTS |
|---|---|---|
| CRIM | Crime rate per town | - 0.0133970 |
| CHAS | Houses close to Charles River | 0.1588957 |
| NOX | Nitric Oxide emission | - 0.4567497 |
| RM | Average number of rooms per dwelling | 0.1809727 |
| DIS | Distance to work | - 0.0316789 |
| PTRATIO | Pupil teacher ratio by town | - 0.0277677 |
| B | Proportion of blacks by town | 0.0004492 |
| LSTAT | Percentage of lower status of population | - 0.0247731 |

- In examining the table, unit increase in any of these variables influences the MEDV value. Unit increase in CRIM (crime rate) brings down the house price by (- 0.0133970). Whereas, unit increase in number of rooms (RM), increases the value of the house by 0.18097. Similarly, we can deduce for other variables.

**Final Conclusion of the Project- Insights derived and Recommendation**

- **Insights Derived:**

1. **The factors that drive the value of houses in Boston are crime rate, distance from the Charles river, nitric oxide emission, number of rooms, distance from workplace, pupil-teacher ratio, proportion of blacks, and percentage of lower status of population.**
2. **The value of houses tend to increase when there are more rooms, and when it is located close to the Charles river. Lower crime rate and lower pupil-teacher ratio also contributed in increasing the value of houses.**
3. **The fitted regression model shows that higher levels of pollution decrease house prices to a greater extent than distance to work. Employment zones tend to have higher levels of nitrogen oxide emission. Hence, it is reasonable to think that people would prefer living farther from their workplace if it meant lower levels of pollution.**

- **Our recommendation to client (city council of Boston, MA. ):**

1. **Design houses with more number of rooms.**
2. **Locate the houses close to the Charles river and preferably distant from the industrial zones.**

# References


- **Business requirement provided by IMS PRO in CapStoneProject1_Question.pdf**

- **Boston Housing data provided IMS PRO HousingData.csv**

- **Outlier treatment using IQR rule : http://stamfordresearch.com/outlier-removal-in-r-using-iqr-rule/#comment-43**

# Thank You

**By S.Vismay Archi**