Question 1: Our objective of this historical dataset on housing in the Boston area is to investigate, summarize and visualize the statistics in R. The dataset contains 14 variables using which we have summarized the data's features which in turn will help to evaluate the values of the property and per capita crime rate. To analyze the data, we have produced some descriptive statistics on data.

We used libraries like **ggplot2, ggcorrplot, GGally, summary tools,** etc., to analyze the data imported from QMPlus as housing.csv. **setwd()** function is used to set the working directory for the dataset, and **read.csv()** function is used to load the data.

```
# Checking missing values, total number of values & Summary of the loaded data
>head(housing)
>table(is.na(housing))
>housing_summary<-dfSummary(housing)
>view(housing_summary)
```

Table 1: Summary statistics for "housing.csv" dataset.

Data Frame Summary

housing

Dimensions: 506 x 14
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	CRIM [numeric]	Mean (sd) : 3.6 (8.6) min ≤ med ≤ max: 0 ≤ 0.3 ≤ 89 IQR (CV) : 3.6 (2.4)	504 distinct values		506 (100.0%)	0 (0.0%)
2	CR01 [integer]	Min : 0 Mean : 0.5 Max : 1	0 : 253 (50.0%) 1 : 253 (50.0%)		506 (100.0%)	0 (0.0%)
3	ZN [numeric]	Mean (sd) : 11.4 (23.3) min ≤ med ≤ max: 0 ≤ 0 ≤ 100 IQR (CV) : 12.5 (2.1)	26 distinct values		506 (100.0%)	0 (0.0%)
4	INDUS [numeric]	Mean (sd) : 11.1 (6.9) min ≤ med ≤ max: 0.5 ≤ 9.7 ≤ 27.7 IQR (CV) : 12.9 (0.6)	76 distinct values		506 (100.0%)	0 (0.0%)
5	CHAS [integer]	Min : 0 Mean : 0.1 Max : 1	0 : 471 (93.1%) 1 : 35 (6.9%)		506 (100.0%)	0 (0.0%)
6	NOX [numeric]	Mean (sd) : 0.6 (0.1) min ≤ med ≤ max: 0.4 ≤ 0.5 ≤ 0.9 IQR (CV) : 0.2 (0.2)	81 distinct values		506 (100.0%)	0 (0.0%)
7	RM [numeric]	Mean (sd) : 6.3 (0.7) min ≤ med ≤ max: 3.6 ≤ 6.2 ≤ 8.8 IQR (CV) : 0.7 (0.1)	446 distinct values		506 (100.0%)	0 (0.0%)
8	AGE [numeric]	Mean (sd) : 68.6 (28.1) min ≤ med ≤ max: 2.9 ≤ 77.5 ≤ 100 IQR (CV) : 49 (0.4)	356 distinct values		506 (100.0%)	0 (0.0%)
9	DIS [numeric]	Mean (sd) : 3.8 (2.1) min ≤ med ≤ max: 1.1 ≤ 3.2 ≤ 12.1 IQR (CV) : 3.1 (0.6)	412 distinct values		506 (100.0%)	0 (0.0%)
10	RAD [integer]	Mean (sd) : 9.5 (8.7) min ≤ med ≤ max: 1 ≤ 5 ≤ 24 IQR (CV) : 20 (0.9)	1 : 20 (4.0%) 2 : 24 (4.7%) 3 : 38 (7.5%) 4 : 110 (21.7%) 5 : 115 (22.7%) 6 : 26 (5.1%) 7 : 17 (3.4%) 8 : 24 (4.7%) 24 : 132 (26.1%)		506 (100.0%)	0 (0.0%)
11	TAX [integer]	Mean (sd) : 408.2 (168.5) min ≤ med ≤ max: 187 ≤ 330 ≤ 711 IQR (CV) : 387 (0.4)	66 distinct values		506 (100.0%)	0 (0.0%)
12	PTRATIO [numeric]	Mean (sd) : 18.5 (2.2) min ≤ med ≤ max: 12.6 ≤ 19.1 ≤ 22 IQR (CV) : 2.8 (0.1)	46 distinct values		506 (100.0%)	0 (0.0%)
13	LSTAT [numeric]	Mean (sd) : 12.7 (7.1) min ≤ med ≤ max: 1.7 ≤ 11.4 ≤ 38 IQR (CV) : 10 (0.6)	455 distinct values		506 (100.0%)	0 (0.0%)
14	MEDV [numeric]	Mean (sd) : 22.5 (9.2) min ≤ med ≤ max: 5 ≤ 21.2 ≤ 50 IQR (CV) : 8 (0.4)	229 distinct values		506 (100.0%)	0 (0.0%)

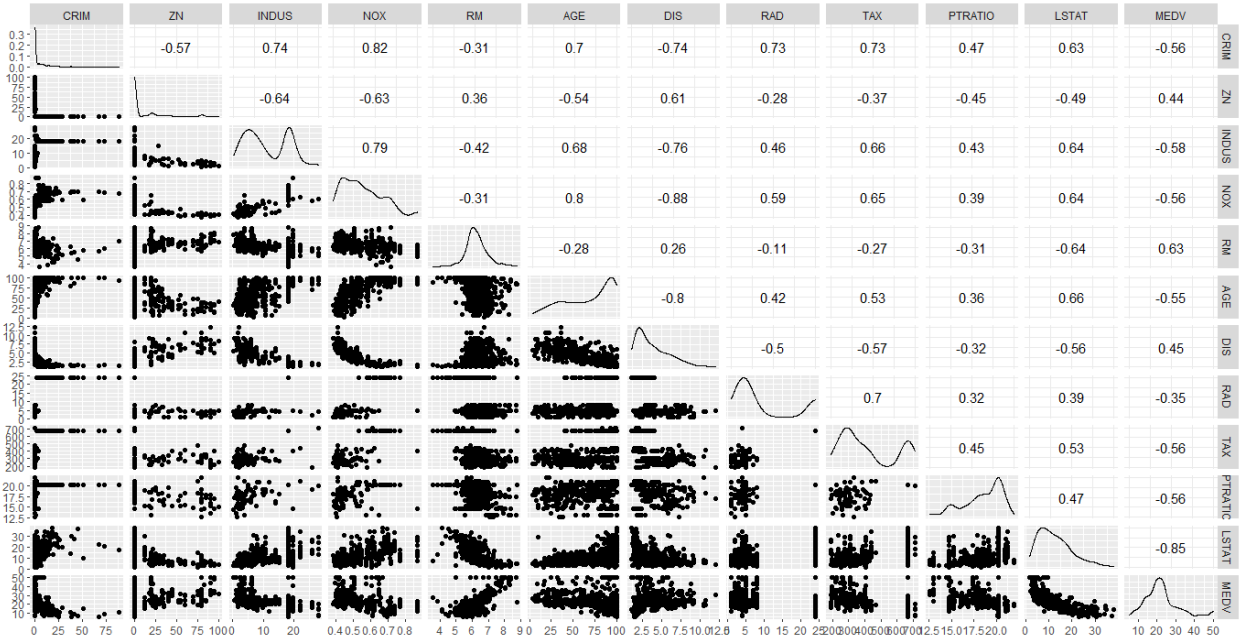
As per the initial analysis, there are 7084 observations in the dataset. There are 14 columns, of which 11 are numerical variables, and 3 are numeric, categorical variables. There are no missing values in the data set, and no cleaning is required.

For the categorical variables with values 0 & 1, we changed them to more meaningful string values as per the variable description by using the **factor()** function.

```
# Setting CR01 and CHAS as factor since its Categorical:
>housing$CR01 = factor(housing$CR01,levels=c(0,1), labels=("LessThanCrMedian","MoreThanCrMedian"))
>housing$CHAS = factor(housing$CHAS,levels = c(0,1), labels = c('FarfromRiver','ClosetoRiver'))
```

After changing the values of the categorical variable it will be useful to see the correlation using **ggpairs()** function between all the variables that are numerical by filtering them and storing them in the new dataset.

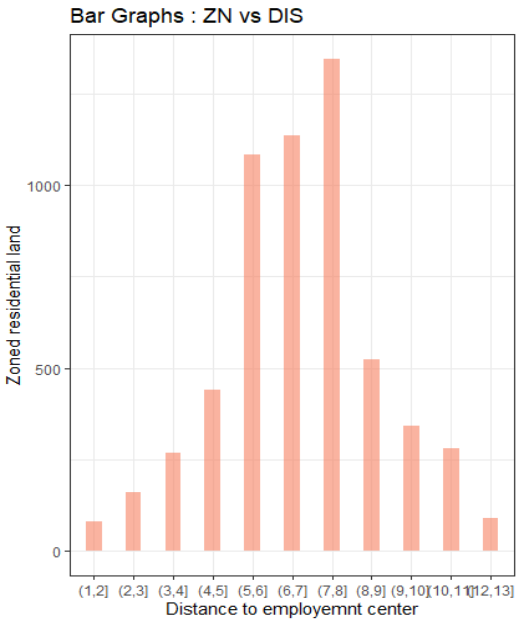
Plot 2: Correlation statistics for the numerical dataset.



It is evident from the above plot that MEDV has the weakest correlation with LSTAT, which means as the prevalence of both parameters diminishes, so does its correlation value. Also a similar trend can be observed with PTRATIO, INDUS, TAX, etc., where linear correlation is very weak. Where MEDV has a strong correlation with RM, the prevalence of both parameters soars so does its correlation value. For CRIM , the weakest correlation it has is with DIS. Similarly, it has the strongest correlation with NOX, whereas it has a strong correlation with INDUS, RAD, and TAX. Now, let's dive into exploratory analysis on the dataset:

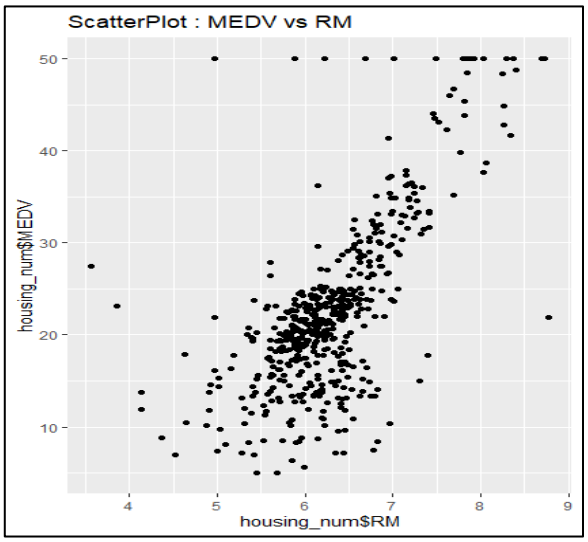
```
# Bar graph to show ZN vs DIS
>housing_num$DIS_grp = cut(housing_num$DIS, breaks = seq(1, 13, 1))
>ZN_grp1 = housing_num %>% group_by(DIS_grp) %>%
  summarise(ZN_1 = sum(ZN))
> ZN_grp1 %>%
  ggplot(aes(x=DIS_grp, y=ZN_1)) +
  geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
  xlab("Distance to employemnt center") + ylab("Zoned residential land") +  ggtitle("Bar Graphs : ZN vs DIS") + theme_bw()
```

Plot 3: Bar plot of ZN vs DIS.



The correlation between ZN and DIS is strong. The bar graph suggests that the maximum proportion of residential lands and above 25000 sq ft allotted are 5 to 8 unit distance away from the employment centers. This also indicates that these types of residential zones are neither too close nor too far from the employment center. Thus maximum population lives 5 to 8 unit distance away from the Boston employment centers.

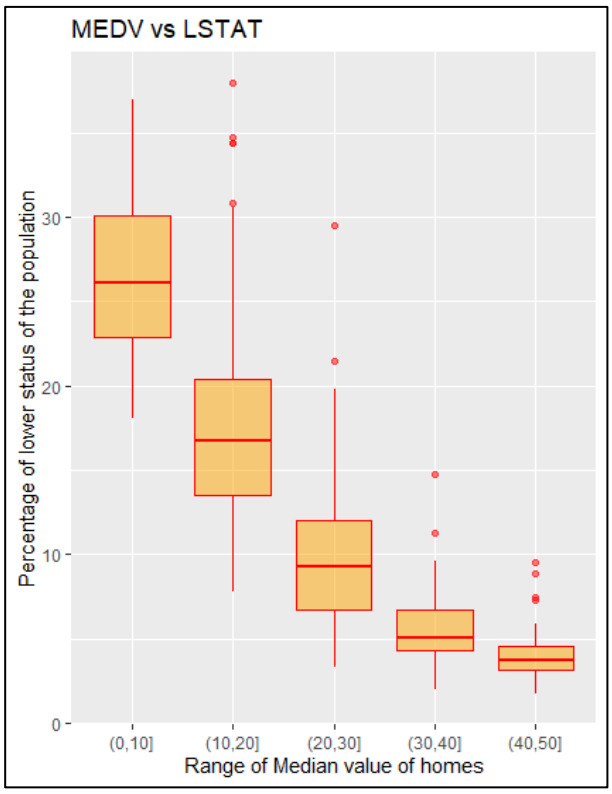
Plot 4: Scatter plot of MEDV vs RM.




```
# ScatterPlot : MEDV vs RM
>ggplot(housing_num, aes(x=housing_num$RM,
y=housing_num$MEDV))
+geom_point()+ggtitle("ScatterPlot : MEDV vs RM")
```

Plot 4 displays a positive linear relationship between MEDV and RM. The maximum density of houses has 15k$ to 25k$ value which has on an average 6 rooms. Whereas the second-highest density of houses has 26k$ to 40k$ value which has on an average 7 rooms. This positive correlation suggests that as the number of rooms increases in a house its value also increases, but there are few outliers that show that some median value of houses does not follow the trend.

Plot 5: Box plot of MEDV vs LSTAT.



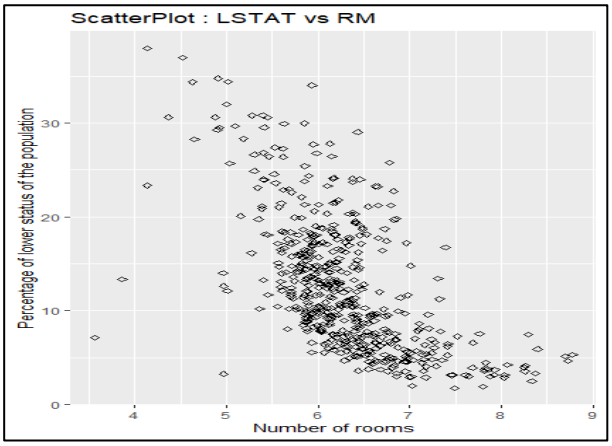

```
# Box plot for MEDV and LSTAT
>housing_num$MEDV_grp3 = cut(housing_num$MEDV,
breaks = seq(0, 51, 10))
>ggplot(housing_num, aes(x=MEDV_grp3, y=LSTAT)) +
 geom_boxplot(color="red", fill="orange", alpha=0.5)+
 scale_y_continuous(name = "Percentage of lower status
of the population") +
 scale_x_discrete(name = "Range of Median value of
homes") +
 ggtitle("MEDV vs LSTAT")
```

We grouped the median value into 5 buckets to observe the relation between LSTAT and MEDV. IT is evident from plot 5 that as the value of homes increases the percent of lower status population decreases which indicates that there is a very weak correlation between the two variables, which is -0.74. It is evident from the plot that more than 25% of the lower-status population lives in homes with MEDV 0 to $10K, whereas less than 5% of the lower-status population lives in homes with MEDV more than $40k. Maximum lower status population percent comes in 0 to $30K bracket of MEDV.

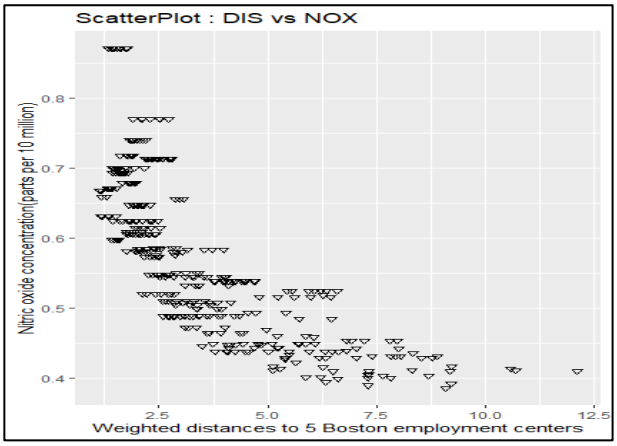
In plot 6 there is a clear negative trend which can be observed, as the correlation between LSTAT and RM is -0.61, which suggests that the number of rooms in a house is directly proportional to the status of any individual, people with good wealth can afford the house with more number of rooms. It is also noticeable that the majority of people in Boston owns 6 or 7 room houses and the 5 to 20 percent of lower status people lives there

```
# Scatter plot LSTAT vs RM
>ggplot(housing_num, aes(x=housing_num$RM, y=housing_num$LSTAT)) +
 geom_point(shape = 23) + ggtitle("ScatterPlot : LSTAT vs RM")+
 xlab("Number of rooms") + ylab("Percentage of lower status of the population")
```

Plot 6: Scatter plot of LSTAT vs RM.



Plot 7: Scatter plot of DIX vs NOX.



Plot 7 between DIS and NOX shows the negative correlation which means that the concentration of nitric oxide decreases as the distance to Boston employment centers increases. This indicates that nitric oxide concentration is maximum near the employment centers.

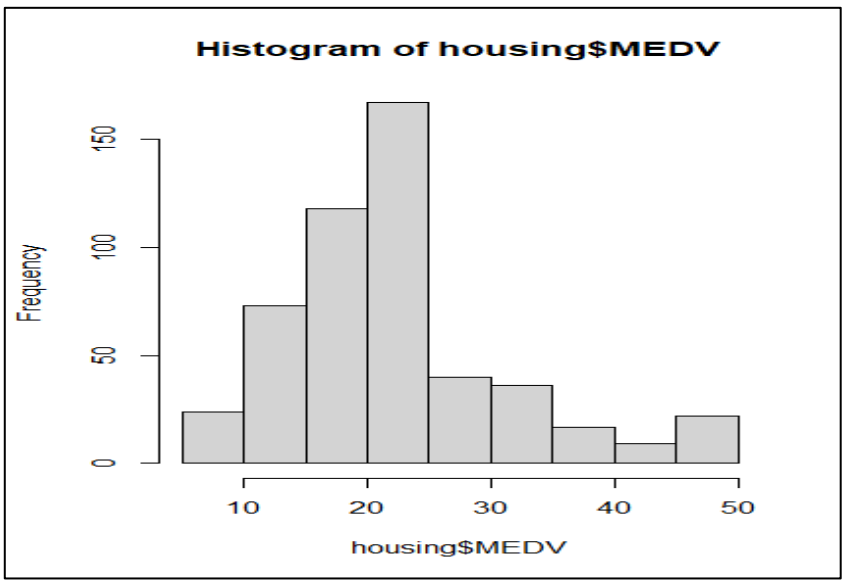
```
# ScatterPlot : DIS vs NOX
>ggplot(housing_num, aes(x=housing_num$DIS, y=housing_num$NOX)) +  geom_point(shape = 25) + ggtitle("ScatterPlot : DIS vs NOX")+  xlab("Weighted distances to 5 Boston employment centers") +  ylab("Nitric oxide concentration(parts per 10 million)")
```

Plot 7 between DIS and NOX shows the negative correlation which means that the concentration of nitric oxide decreases as the distance to Boston employment centers increases. This indicates that nitric oxide concentration is maximum near the employment centers.

Question 2: Develop a Regression model to predict the Value of MEDV

In the second part of the report, we are aiming to develop a regression model with explanatory variables capable of predicting the Median Value of Residences in a Neighborhood (MEDV). To begin with, we checked the normality of the target variable (MEDV) using the **hist()** function.

Plot 2.1: Normality Test of MEDV



As per **Plot 2.1**, we can clearly see that the target variable is almost normal with slight right skewness.

2.1: Identify key explanatory Variables:

In the first step, we will run a linear model with all the variables available in the data. We will take a summary of the linear model and graph it to show the linear nature of the data.

Linear Model 1:

```
linearModel1 <-
lm(MEDV~CRIM+CR01+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+LSTAT,data=housing)
summary(linearModel1)
plot(linearModel1,which=2)
```

After performing the above linear regression, we proceeded with selecting the most significant variables. To do that, we first looked at the p-values to determine the level of significance of each variable correlation coefficient.

Table 2.1: Results of Linear regression Model 1

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.030856   4.965070   8.868  < 2e-16 ***
CRIM         -0.122047   0.032744  -3.727 0.000216 ***
CR01          1.976489   0.667794   2.960 0.003228 **
ZN            0.048842   0.013786   3.543 0.000433 ***
INDUS         0.006749   0.061703   0.109 0.912948
CHAS          2.823384   0.863258   3.271 0.001148 **
NOX         -22.415794   4.016266  -5.581 3.95e-08 ***
RM            3.539569   0.418896   8.450 3.32e-16 ***
AGE          -0.001764   0.013350  -0.132 0.894903
DIS          -1.485572   0.200062  -7.426 4.99e-13 ***
RAD           0.249315   0.067755   3.680 0.000259 ***
TAX          -0.012054   0.003777  -3.191 0.001508 **
PTRATIO      -0.944055   0.131196  -7.196 2.33e-12 ***
LSTAT        -0.550050   0.050269 -10.942  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.761 on 492 degrees of freedom
Multiple R-squared:  0.739,     Adjusted R-squared:  0.7321
F-statistic: 107.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

As per the first linear regression model results, we can see Coefficients of all the variables are less than or more than 0 but not zero. So, we can say that this model is possible.

We can see in **Table 2.1** that the residual standard error is 4.761 and the Adjusted R-squared value is 0.7321 which is a pretty good number, but we can see that there are a few insignificant factors that are included in the model. As INDUS and AGE have p-value>0.05. So, we will drop INDUS and AGE, try to improve the performance of the model.

2.1.1: Perform Regression with selected Variables

Linear Model 2: As we have seen in **Table 2.1**, INDUS and AGE have high p-value, so we will remove these from our model. We will also remove CR01 since we already have CRIM in the regression.

```
linearModel2 <- lm(MEDV~CRIM+ZN+CHAS+NOX+RM+DIS+RAD+TAX+PTRATIO+LSTAT,data=housing)
summary(linearModel2)
plot(linearModel2,which = 1)
```

Adjusted R-square is 0.7331, Residual standard error is 4.751 and our model can predict this with 494 degrees of freedom. Our model’s performance has been improved.

Linear Model 3: Referring to **Table 2.1**, we can observe that the Coefficient of ZN and TAX are very low. It shows that these variables are not much important for MEDV. SO, we will eliminate these variables in our Linear Model. Adjusted R-square is 0.7188, Residual standard error is 4.877and our model can predict this with 497 degrees of freedom.

Linear Model 4: We performed a correlation between variables in question 1(). After examining the correlation between the variables, adding an interaction variable between NOX and DIS makes sense. This is our improved Linear equation. This results with Adjusted R-square is 0.724, Residual standard error is 4.831 and our model can predict this with 496 degrees of freedom.

```
LinearModel4 <- lm(MEDV~CRIM+CHAS+NOX+RM+DIS+RAD+PTRATIO+LSTAT+NOX*DIS,data=housing)
summary(LinearModel4)
hist(LinearModel4$residuals)
```

Linear Model 5:

LOG TRANSFORMATION

Data transformation method when we replace x to log(x) to reduce the skewness in data and make it closer to normal distribution. This model is giving the best Adjusted R-square and minimum error of all our Linear models. Adjusted R-square is 00.7707, Residual standard error is 0.1957, and degree of freedom is 496

```
LinearModel5 <-
lm(log(MEDV)~CRIM+CHAS+NOX+RM+DIS+RAD+PTRATIO+LSTAT+NOX*DIS,data=housing)
summary(LinearModel5)
plot(LinearModel5)
hist(LinearModel5$residuals)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.166722   0.216182  19.274  < 2e-16 ***
CRIM        -0.010956   0.001376  -7.960 1.18e-14 ***
CHAS         0.123297   0.035122   3.511 0.000488 ***
NOX         -0.765804   0.200977  -3.810 0.000156 ***
RM           0.096609   0.016484   5.861 8.41e-09 ***
DIS         -0.011292   0.032475  -0.348 0.728191
RAD          0.004480   0.001631   2.747 0.006225 **
PTRATIO     -0.040590   0.005182  -7.832 2.93e-14 ***
LSTAT       -0.030108   0.001942 -15.507  < 2e-16 ***
NOX:DIS     -0.079975   0.074961  -1.067 0.286540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1957 on 496 degrees of freedom
Multiple R-squared:  0.7748,    Adjusted R-squared:  0.7707
F-statistic: 189.6 on 9 and 496 DF,  p-value: < 2.2e-16
```

Table 2.1: Results of Linear regression Model 1

If we compare the summary of Linear model 1**(Table 2.1)** and Final Linear model 5**(Table 2.2),** we can see that the adjusted R-squared has been increased from 0.7321 to 0.7707.

A residual error has been decreased from 4.761 to 0.1957, less error is proof of a better model. And the degree of freedom has also been increased in our final model, it increases from 492 to 496. P-value is less than 2.2e-16. All this shows that our model has been improved by the elimination of unnecessary variables.

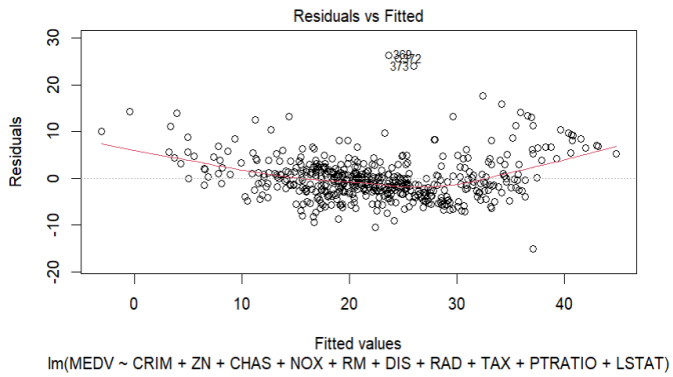
❖ **HOW GOOD IS THE RELATIONSHIP BETWEEN ALL VARIABLES AND MEDV?**

R-SQAURE=77%

R-SQUARE IS HIGH RELATIONSHIP IS GOOD.

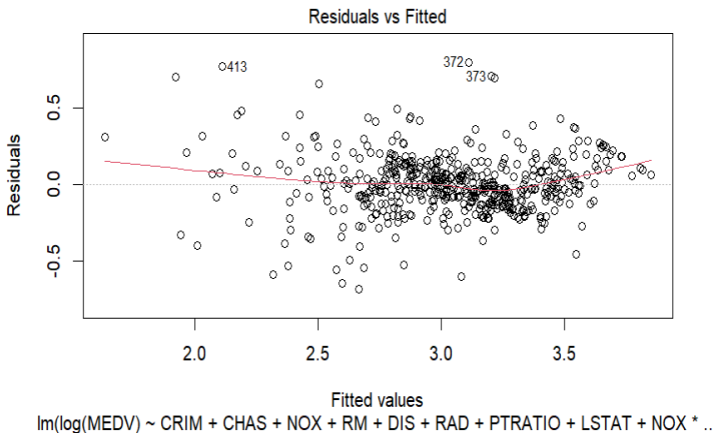
Based on Residual and Fitted Graph:

Plot 2.2: Linear Model 2 Residual graph



As we can see in **Plot 2.2** residuals are dispersed, in middle, most of the residuals are near the predicted line. However, on the right-hand side, many residuals are dispersed.
Here, we can use a linear model, but we would need to be more cautious because data have **heteroscedasticity.**

Plot 2.3: Linear Model 5 Residual graph



In the final model **Plot 2.3**, we can see that residuals are less dispersed from the predicted line as compared to Linear model 1. This shows our model's efficiency and accuracy have been increased. Our predicted values are closer to original values, this shows our model is good.

As we can see even in the above Residual graph, residuals are still spreading from the fitted line. So, **heteroscedasticity** is there. Therefore, we would need a non-linear model for this data**.**

2.2: Consider Polynomial Models

In this part, we will work on Polynomial models. Polynomial models show a much clearer relationship between dependent and independent variables. A polynomial model is built with variables that are most important for the independent variable. As we have already observed that RM and LSTAT have low p-value, so these are very important variables. We will create a polynomial model on them.

```
n=nrow(housing)
cv_res1 = vector(length=n)
for(i in 1:n){
 fiti = lm(MEDV~CRIM+CR01+CHAS+NOX+DIS+LSTAT+RAD+PTRATIO+poly(RM,2),data=housing[-i,])
 predi = predict(fiti, newdata=housing[i,])
 cv_res1[i] = housing$MEDV[i] - predi
}
# PRESS is sum of squared cross-validated residuals
PRESS1 = sum(cv_res1^2)
#other Model
cv_res2 = vector(length=n)
for(i in 1:n){
 fiti1 = lm(MEDV~CRIM+CR01+CHAS+NOX+DIS+RAD+PTRATIO+LSTAT+poly(RM,3),data=housing[-i,])
 predi1 = predict(fiti1,newdata=housing[i,])
 cv_res2[i] = housing$MEDV[i] - predi1
}
PRESS2 = sum(cv_res2^2)
```

Consider the degree of a polynomial by evaluating the press of each model. We calculate cross-validation cv_res1 with degree 2 for RM and then calculate PRESS1. Then cross-validate again with degree 3 for RM and calculate PRESS2. We observed that PRESS static value increases when we go from degree 2 polynomial to degree 3 for RM. PRESS 1 is 9622.232 and PRESS 2 is 9911.016. So, we will set RM for degree 2 and we will repeat the similar steps for LSTAT to find the best degree for the polynomial model.

As we can see in the below equation. After increasing the value of LSTAT polynomial what we observed that PRESS3 value has decreased significantly to 8831.995. So, we decided to further increase the degree of LSTAT in polynomial till 5 and we got results for each sum of squared cross-validated residuals respectively.

```
cv_res3 = vector(length=n)
for(i in 1:n){
  fiti1 = lm(MEDV~CRIM+CR01+CHAS+NOX+DIS+RAD+PTRATIO+poly(RM,2)+poly(LSTAT,2),data=housing[-i,])
  predi1 = predict(fiti1,newdata=housing[i,])
  cv_res3[i] = housing$MEDV[i] - predi1}
PRESS3 = sum(cv_res3^2)
cv_res4 = vector(length=n)
for(i in 1:n){
  fiti1 = lm(MEDV~CRIM+CR01+CHAS+NOX+DIS+RAD+PTRATIO+poly(RM,2)+poly(LSTAT,3), data=housing[-i,])
  predi1 = predict(fiti1,newdata=housing[i,])
  cv_res4[i] = housing$MEDV[i] - predi1}
PRESS4 = sum(cv_res4^2)
cv_res5 = vector(length=n)
for(i in 1:n){
  fiti1 = lm(MEDV~CRIM+CR01+CHAS+NOX+DIS+RAD+PTRATIO+poly(RM,2)+poly(LSTAT,4), data=housing[-i,])
  predi1 = predict(fiti1,newdata=housing[i,])
  cv_res5[i] = housing$MEDV[i] - predi1
PRESS5 = sum(cv_res5^2)
cv_res6 = vector(length=n)
for(i in 1:n){
  fiti1 = lm(MEDV~CRIM+CR01+CHAS+NOX+DIS+RAD+PTRATIO+poly(RM,2)+poly(LSTAT,5),data=housing[-i,])
  predi1 = predict(fiti1,newdata=housing[i,])
  cv_res6[i] = housing$MEDV[i] - predi1}
PRESS6 = sum(cv_res6^2)
```

Table 2.2: Results of the lowest sum of squared cross-validated residuals

```
> PRESS1
[1] 9622.232
> PRESS2
[1] 9911.016
> PRESS3
[1] 8831.995
> PRESS4
[1] 8781.273
> PRESS5
[1] 8627.922
> PRESS6
[1] 8396.607
```

As per **Table 2.2**, we can clearly see that sum of squared cross-validated residuals decreased every time when we increase the degree of LSTAT in the polynomial model. So, in the end, we selected degree 5 for LSTAT where we got the lowest sum of squared cross-validated residuals value.

2.2: Performance of Polynomial Model

```
finalModel=lm(MEDV~CRIM+CR01+CHAS+NOX+DIS+RAD+PTRATIO+poly(RM,2)+poly(LSTAT,5),data=housing)
```

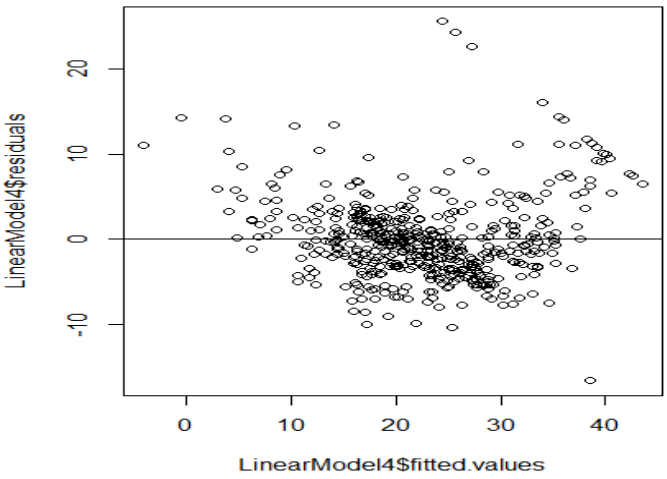
Table 2.3: Results of Polynomial model with RM degree 2 and LSTAT degree 5

```
Call:
lm(formula = MEDV ~ CRIM + CHAS + NOX + DIS + RAD + PTRATIO +
    poly(RM, 2) + poly(LSTAT, 5), data = housing)

Residuals:
     Min       1Q   Median       3Q      Max
-24.3760  -2.1640  -0.1494   1.6917  27.6777

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        50.33185    2.96761  16.960  < 2e-16 ***
CRIM               -0.15205    0.02737  -5.555 4.55e-08 ***
CHASClosetoRiver    2.30607    0.71189   3.239 0.001279 **
NOX               -18.26728    2.88376  -6.335 5.38e-10 ***
DIS                -1.00851    0.13594  -7.419 5.22e-13 ***
RAD                 0.09068    0.03300   2.748 0.006220 **
PTRATIO            -0.77565    0.10356  -7.490 3.22e-13 ***
poly(RM, 2)1       50.43292    5.73091   8.800  < 2e-16 ***
poly(RM, 2)2       40.14881    4.62564   8.680  < 2e-16 ***
poly(LSTAT, 5)1  -101.45196    6.49010 -15.632  < 2e-16 ***
poly(LSTAT, 5)2    33.56117    4.72828   7.098 4.45e-12 ***
poly(LSTAT, 5)3   -10.76325    4.27073  -2.520 0.012043 *
poly(LSTAT, 5)4    15.06007    4.10013   3.673 0.000266 ***
poly(LSTAT, 5)5   -14.11791    4.00708  -3.523 0.000466 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.944 on 492 degrees of freedom
Multiple R-squared:  0.8208,    Adjusted R-squared:  0.8161
F-statistic: 173.4 on 13 and 492 DF,  p-value: < 2.2e-16
```

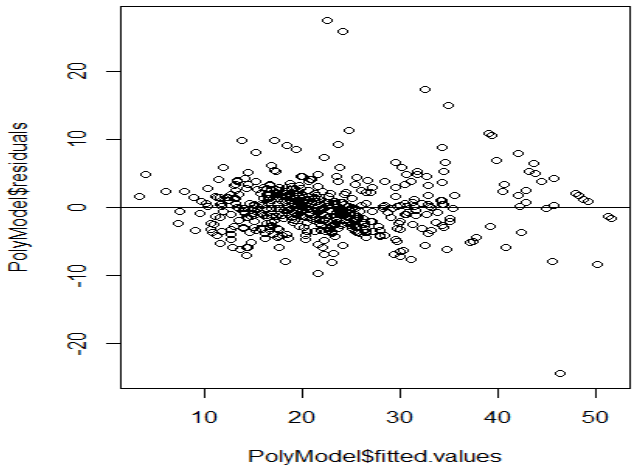
As we can see in **Table 2.3** Adjusted R-square value has been improved significantly from 0.77 in the Linear model to in the Polynomial model 0.8208.

2.2: Comparison between Linear and Polynomial Model using residual scatterplot

Plot 2.4: Linear Model Residuals



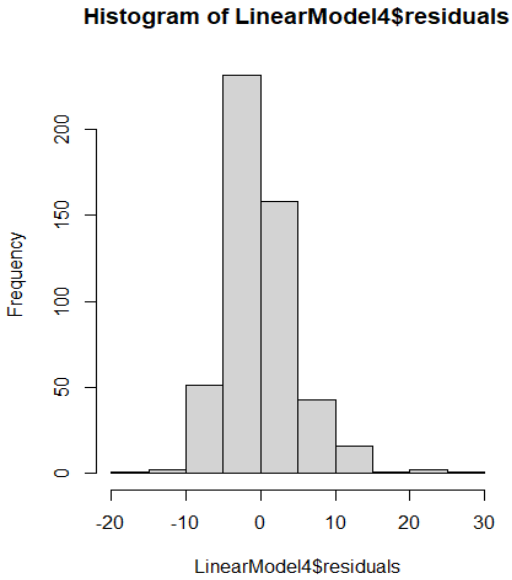
Plot 2.5: Polynomial Model Residuals



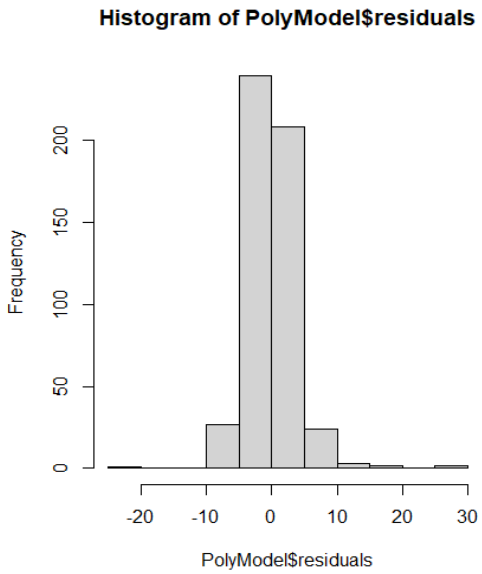
On left, we have a linear model residual scatter plot **(Plot 2.4)** and in right, we have the Polynomial model **(Plot 2.5**). In the Linear model, plot residuals are very much scattered, which shows **heteroscedasticity** in the model. Whereas in the Polynomial model scatter plot on the right we can see that residuals are more compact and closer to the fitted line.

2.2: Comparison between Linear and Polynomial Model using Histogram

Plot 2.6: Linear Model Histogram



Plot 2.7: Polynomial Model Histogram



As we can see in **Plot 2.6** Linear histogram is approximately normal but a little skewed towards the right. Whereas in Polynomial model histogram **Plot 2.7** it is not skewed at all, it is perfectly normal. It shows our polynomial model is better than the Linear model.

Question 3: In the third part of the report we were asked to create a classification model to predict whether a neighborhood has crime levels above the median value (CR01=1) or below (CR01=0), i.e. if it has a high or low crime per capita rate. To see how the other variables affect whether a neighborhood is classified with CR01=1 or 0 and identify the most significant ones, we performed a regression with CR01 as the target variable and all the other variables given in the dataset except the actual Crime Rate to observe the most significant explanatory variables.

3.1: Predictive Variable Selection:

There was no need to tamper with CR01 as it was correctly classified as a factor variable in Part 1 of the report, with level 0 corresponding to 'LessThanCRmedian' and level 1 corresponding to 'MoreThanCRmedian.'

```
> prediction1 <- glm(CR01 ~.-CRIM, family=binomial,data=housing)
> summary(prediction1)
```

After performing the above logistic regression, we selected the most significant variables. To that we first looked at the p-values to determine the level of significance of each variable's correlation coefficient. These variables proved to be the most significant in predicting CR01 with p-values<0.05 :

NOX, DIS,RAD,PTRATIO,MEDV,ZN,TAX

After selecting these explanatory variables to consider for the classification model of CR01, we visually explored the effect of the Crime rate on each variable and vice-versa.

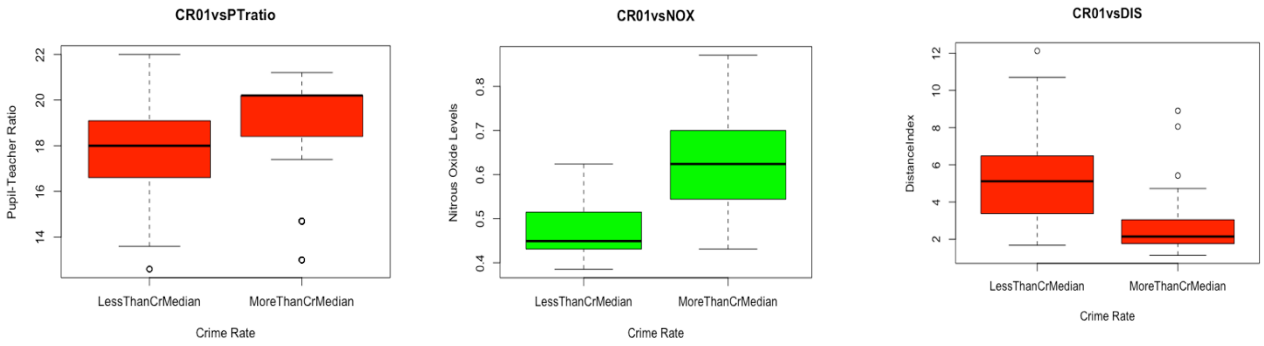
We created box plots and scatterplots that depict the differences in the marginal distributions of each variable given the level of crime rate the neighborhood has, above or below the median.

```
Call:
glm(formula = CR01 ~ . - CRIM, family = binomial, data = housing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0082  -0.1688  -0.0004   0.0027   3.4324

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -38.832427   6.086281  -6.380 1.77e-10 ***
ZN           -0.086228   0.034090  -2.529   0.0114 *
INDUS        -0.052438   0.042817  -1.225   0.2207
CHAS          0.619914   0.722150   0.858   0.3907
NOX          47.913837   7.344214   6.524 6.84e-11 ***
RM           -0.271941   0.676239  -0.402   0.6876
AGE           0.021474   0.012105   1.774   0.0761 .
DIS           0.669991   0.214618   3.122   0.0018 **
RAD           0.669240   0.151742   4.410 1.03e-05 ***
TAX          -0.006165   0.002622  -2.351   0.0187 *
PTRATIO       0.326433   0.116296   2.807   0.0050 **
LSTAT         0.053537   0.047105   1.137   0.2557
MEDV          0.147987   0.064347   2.300   0.0215 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 701.46  on 505  degrees of freedom
Residual deviance: 218.75  on 493  degrees of freedom
AIC: 244.75

Number of Fisher Scoring iterations: 9
```

3.2:Graphical Representation of Relationship of Explanatory Variables with CR01:

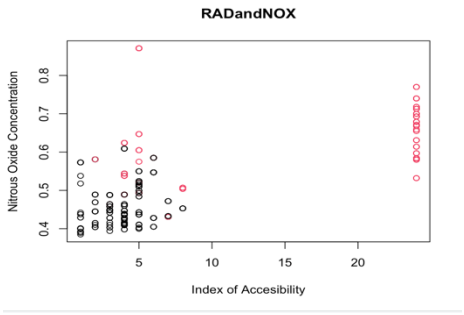
In the boxplots below, it can be observed that as Nitrous Oxide Levels(NOX) increase, so does the probability of a neighborhood having high levels of Crime, i.e., CR01=1. In the boxplot of CR01vsNOX, the mean value of NOX for areas below the crime rate median is 0.45, while the mean value of NOX for neighborhoods with more than the crime median is much higher at 0.62. NOX and CR01 are positively correlated.

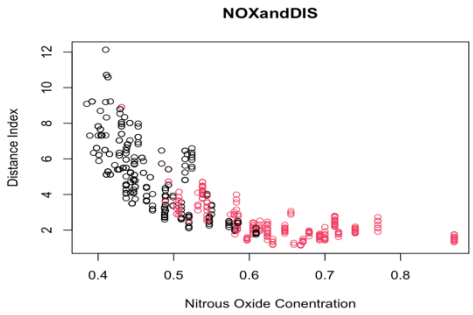
Plots 3.1: CR01=0 vs CR01=1



Pupil-Teacher ratio and CR01 also seem to be positively correlated. As PTRATIO increases, so does the probability of the neighborhood having above the median crime rate. On the other hand, the Distance index seems to be negatively correlated with CR01 as it can be observed that as DIS decreases, so does the probability of a neighborhood having above the median crime rate.

In the scatterplots below, neighborhoods above the medial level of crime rate are depicted with red dots, and areas with below the medial level of crime rate are shown with black dots.





It can be observed that neighborhoods with a high index of accessibility (RAD) tend to have crime rates above the median value. i.e., CR01=1. Furthermore, the inverse effects of DIS and NOX on Crime rate can be seen in the scatterplot above, where DIS is negatively correlated with CR01, whereas NOX is positively correlated with CR01.

```
>plot(x=housing$CR01,y=housing$DIS,col="red",main="CR01vsDIS",xlab="Crime Rate",ylab="DistanceIndex")
>plot(x=housing$CR01,y=housing$NOX,col="green",main="CR01vsNOX",xlab="Crime Rate",ylab="Nitrous Oxide Levels")
>plot(x=housing$CR01,y=housing$RAD,col="red",main="CR01vsRAD",xlab="Crime Rate",ylab="Index of accesibility")
>plot(x=housing$CR01,y=housing$PTRATIO,col="red",main="CR01vsPTratio",xlab="Crime Rate",ylab="Pupil-Teacher Ratio")
>plot(x=housing$NOX,y=housing$DIS,col=housing$CR01, main="NOXandDIS",xlab= 'Nitrous Oxide Concentration',ylab='Distance Index')
>plot(x=housing$RAD,y=housing$NOX,col=housing$CR01, main="RADandNOX",xlab= 'Index of Accesibility',ylab='Nitrous Oxide Concentration')
```

3.3:Final Classification model:

```
> prediction2 <- glm(CR01 ~RAD+NOX+DIS+PTRATIO+MEDV,family=binomial,data=housing)
> summary(prediction2)
```

The final classification model we created consists of 5 explanatory variables:

RAD,NOX,DIS,PTRATIO,MEDV

All of which are significant in capturing the probability of whether a neighborhood has above the median crime rates(CR01=1) or below the median crime rates(CR01=0)

Table 3.2: Significance of Variables

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -31.52889    4.69156  -6.720 1.81e-11 ***
RAD           0.50275    0.10700   4.699 2.62e-06 ***
NOX          38.27444    5.47234   6.994 2.67e-12 ***
DIS           0.28254    0.14262   1.981  0.04758 *
PTRATIO       0.28733    0.09531   3.015  0.00257 **
MEDV          0.08491    0.02740   3.099  0.00194 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
#set n as the number of rows in the original dataset
#define a vector of random row numbers to sample test dataset from original dataset, with around 1/3 in the test data and 2/3 in the training data:
> n=nrow(housing)
>testindex = sample(1:n, size=n/3)
# test dataset:
>test=housing[testindex,]
# training dataset:
>train=housing[-testindex,]
```

3.4:Model Performance Assessment:

3.4.1:Sampling and splitting of the Dataset:

To assess the performance of the final classification model, we proceeded in creating samples from the original dataset so that 1/3 of the initial observations will be the test data and 2/3 will be the training data. This has been done in the code above.

3.4.2: Training the Model:

We proceeded by fitting the classification model we created, on the training data ("train"), storing the trained model as prediction1fitted. We then used the fitted model to predict the probability of the observations in the test dataset('test'), having above the median crime rate, i.e. CR01=1. We stored the predictions as "test prob".

```
# fit model to training data
>prediction1fitted <- glm (CR01 ~RAD+NOX+DIS+PTRATIO+MEDV, family=binomial, data=train)
# calculate predicted probabilities for the test data
>testprob=predict (prediction1fitted, newdata=test,type="response")
```

3.4.3: ROC Threshold Selection:

-We proceeded by evaluating the model's performance given each possible threshold. We created a vector of the same length as the test set containing only "LessThanCRMedian," i.e., CR01 = 0, and we only proceeded to set those entries in the vector to "MoreThanCRMedian," i.e. CR01 = 1 when the probability given by the classification model is greater than the threshold stated in each instance of the loop (=i/10).

-We decided that 0.5 is the most appropriate threshold to use in the classification model we created. It gives a very high value for the true positive rate (0.916) and a very low value for the false positive rate (0.125). The misclassification rate is the lower out of all the thresholds at 0.107.

Table 3.3: ROC Table

	TPR	FPR	MR
0.1	1.0000000	0.42708333	0.2440476
0.2	1.0000000	0.34375000	0.1964286
0.3	1.0000000	0.27083333	0.1547619
0.4	0.9861111	0.22916667	0.1369048
0.5	0.9166667	0.12500000	0.1071429
0.6	0.8750000	0.10416667	0.1130952
0.7	0.8194444	0.05208333	0.1071429
0.8	0.7500000	0.05208333	0.1369048
0.9	0.7222222	0.00000000	0.1190476

-A high TPR indicates the extent of the ability of our classification to correctly predict if a neighborhood is classified as having crime rate above the median.

-A low FPR indicates that the probability of the classification model wrongly classifying a neighborhood as CR01=1, is low.

```
> table(test$CR01, testpred)
                testpred
                 LessThanCrMedian MoreThanCrMedian
LessThanCrMedian               84                12
MoreThanCrMedian                6                66
```

Confusion Matrix at 0.5

Ultimately, a high TPR and a low FPR characterize an accurate model.

(0.5) would lie at the top left corner of the ROC plot making it the most desirable threshold level for our final classification model with a FPR of 0.125 (=12/84+12) and a TPR of 0.917 (=66/66+6).

```
>ClassificationTable = data.frame(row.names= c("0.1","0.2","0.3","0.4","0.5", "0.6","0.7","0.8","0.9"))
for (i in 1:9) {
>testpred2=rep('LessThanCrMedian',length=length(testprob))
>testpred2[testprob>((i)/10)]='MoreThanCrMedian'
>ConfusionMatrix=table(test$CR01, testpred2)
>ClassificationTable$TPR[i]=ConfusionMatrix [2,2]/( ConfusionMatrix [2,2]+ ConfusionMatrix [2,1])
>ClassificationTable$FPR[i]=ConfusionMatrix [1,2]/( ConfusionMatrix [1,1]+ ConfusionMatrix [1,2])
>ClassificationTable$MR[i]=(ConfusionMatrix [1,2]+ ConfusionMatrix [2,1])/length(testprob)}
>view(ClassificationTable)
```

Declaration

We hereby declare that this project is our own work, written by ourselves in our own words. If quotations are included from published or unpublished sources, these are clearly indicated and acknowledged as such. We also declare that the distribution of workload, as indicated below, has been agreed among all members of the group.

Group number:

1. Student name: George Karseras
Student number: 180098895
Responsible for the following content: Contributed to Question 2 and Question 3

Signature: GK

2. Student name: Shalu Chaudhary
Student number:200614104
Responsible for the following content: Contributed to Question 1 and Question 2

Signature: Shalu Chaudhary

3. Student name: Dev Upadhyay
Student number: 200644879
Responsible for the following content: Contributed to Question 1 and Question 2

Signature: Dev Upadhyay