ANA515_Assignment 4

Sher Mayne

4/30/2021

##Business Goal - Boston is one of the cities that has a huge concentration of college students and young professionals. With its growing population, the housing market has been booming, leading to an exponential increase in the sale price of the homes in the area. The purpose of this research is to understand the Boston home prices in the 1970s based on various variables, focusing on the number of rooms, riverfront property, accessibility to highways, and distance to employment centers. My aim is the determine which different variables are most influential on the housing price. At the end of this research, we will be able to find out how each of the factors affects the home price in Boston.

##Dataset Information - The dataset I will be using is Boston Housing Dataset. The information in this dataset is collected by the U.S. Census Service which can be found on StatLib archive. You may download the file from url = http://lib.stat.cmu.edu/datasets/boston

##Cleaning and Importing Dataset - The data was cleaned manually by having all 14 variables on the same row and inputing the column names. This is to allow R to import the txt file properly.

```
library(readr)
boston_housing<-read_delim("Boston_Housing.txt", delim = " ",skip = 21)
```

```
head(boston_housing,3)
```

```
## # A tibble: 3 x 14
##   CRIM  ZN    INDUS CHAS  NOX   RM    AGE   DIS   RAD   TAX   PTRATIO     B
##   <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   <dbl>
## 1 " 0.~ " 18~ "  2~ " 0"  " 0.~ " 6.~ " 65~ " 4.~ "  1" " 29~ " 15.3~  397.
## 2 " 0.~ "  0~ "  7~ " 0"  " 0.~ " 6.~ " 78~ " 4.~ "  2" " 24~ " 17.8~  397.
## 3 " 0.~ "  0~ "  7~ " 0"  " 0.~ " 7.~ " 61~ " 4.~ "  2" " 24~ " 17.8~  393.
## # ... with 2 more variables: LSAT <chr>, MEDV <chr>
```

#Change Variables' Class Type

```
boston_housing$CRIM<-as.numeric(boston_housing$CRIM)
boston_housing$ZN<-as.numeric(boston_housing$ZN)
boston_housing$INDUS<-as.numeric(boston_housing$INDUS)
boston_housing$CHAS<-as.numeric(boston_housing$CHAS)
boston_housing$NOX<-as.numeric(boston_housing$NOX)
boston_housing$RM<-as.numeric(boston_housing$RM)
boston_housing$AGE<-as.numeric(boston_housing$AGE)
boston_housing$DIS<-as.numeric(boston_housing$DIS)
boston_housing$RAD<-as.numeric(boston_housing$RAD)
boston_housing$TAX<-as.numeric(boston_housing$TAX)
boston_housing$PTRATIO<-as.numeric(boston_housing$PTRATIO)
```

```
boston_housing$B<-as.numeric(boston_housing$B)
boston_housing$LSAT<-as.numeric(boston_housing$LSAT)
boston_housing$MEDV<-as.numeric(boston_housing$MEDV)
print(sapply(boston_housing,class))
```

```
##      CRIM        ZN     INDUS      CHAS       NOX        RM       AGE       DIS
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##       RAD       TAX   PTRATIO         B      LSAT      MEDV
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
```

#Change Column Names to Lowercase

```
library(dplyr)
```

```
boston_housing<-rename_all(boston_housing,tolower)
head(boston_housing, 3)
```

```
## # A tibble: 3 x 14
##      crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio     b
##     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>
## 1 0.00632    18  2.31     0 0.538  6.58  65.2  4.09     1   296    15.3  397.
## 2 0.0273      0  7.07     0 0.469  6.42  78.9  4.97     2   242    17.8  397.
## 3 0.0273      0  7.07     0 0.469  7.18  61.1  4.97     2   242    17.8  393.
## # ... with 2 more variables: lsat <dbl>, medv <dbl>
```

##Describing Dataset This dataset has 14 columns and 506 rows.

```
ncol(boston_housing)
nrow(boston_housing)
```

#Data Description

```
variables<-colnames(boston_housing)
details<-c("per capita crime rate by town",
           "proportion of residential land zoned for lots over 25,000 sq.ft.",
           "proportion of non-retail business acres per town",
           "Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)",
           "nitric oxides concentration (parts per 10 million)",
           "average number of rooms per dwelling",
           "proportion of owner-occupied units built prior to 1940",
           "weighted distances to five Boston employment centres",
           "index of accessibility to radial highways",
           "full-value property-tax rate per $10,000",
           "pupil-teacher ratio by town",
           "1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town",
           "% lower status of the population",
           "Median value of owner-occupied homes in $1000's")
print(data_description<-tibble(variables,details))
```

```
## # A tibble: 14 x 2
```

```
##    variables details
##    <chr>     <chr>
##  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; 0 otherwi~
##  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 lsat      % lower status of the population
## 14 medv      Median value of owner-occupied homes in $1000's
```

#Summary Statistics of Dataset

```
print(sboston_housing<-(summary(boston_housing)))
```

```
##       crim                zn              indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm             age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio            b
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##       lsat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

##Data Preparation - As mentioned above, cleaning was done before importing the data as it was not possible to import the dataset without errors. I will only be using 4 variables against the home price in my analysis.

```
colnames(boston_housing)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "b"       "lsat"    "medv"
```

```
proj_bhm<-boston_housing[c(4,6,8,9,14)]
head(proj_bhm, 3)
```

```
## # A tibble: 3 x 5
##    chas    rm   dis   rad  medv
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     0  6.58  4.09     1  24
## 2     0  6.42  4.97     2  21.6
## 3     0  7.18  4.97     2  34.7
```

#Rearranging Columns

```
proj_bhm<-proj_bhm[, c(5, 2, 1, 3, 4)]
head(proj_bhm, 3)
```

```
## # A tibble: 3 x 5
##    medv    rm  chas   dis   rad
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  24    6.58     0  4.09     1
## 2  21.6  6.42     0  4.97     2
## 3  34.7  7.18     0  4.97     2
```

##Average Home Price in Boston

```
smedv<-summary(proj_bhm$medv)
srm<-summary(proj_bhm$rm)
schas<-summary(proj_bhm$chas)
sdis<-summary(proj_bhm$dis)
srad<-summary(proj_bhm$rad)
sum_stats<-c("Min","1stQu","Median","Mean","3rdQu","Max")
print(summarydata<-tibble(sum_stats,smedv,srm,schas,sdis,srad))
```

```
## # A tibble: 6 x 6
##   sum_stats smedv    srm      schas      sdis      srad
##   <chr>     <table>  <table>  <table>    <table>   <table>
## 1 Min        5.00000 3.561000 0.00000000  1.129600  1.000000
## 2 1stQu     17.02500 5.885500 0.00000000  2.100175  4.000000
## 3 Median    21.20000 6.208500 0.00000000  3.207450  5.000000
## 4 Mean      22.53281 6.284634 0.06916996  3.795043  9.549407
## 5 3rdQu     25.00000 6.623500 0.00000000  5.188425 24.000000
## 6 Max       50.00000 8.780000 1.00000000 12.126500 24.000000
```

```
cchas<-c("non-Riverfront","Riverfront")
tibble(cchas,table(boston_housing$chas))
```

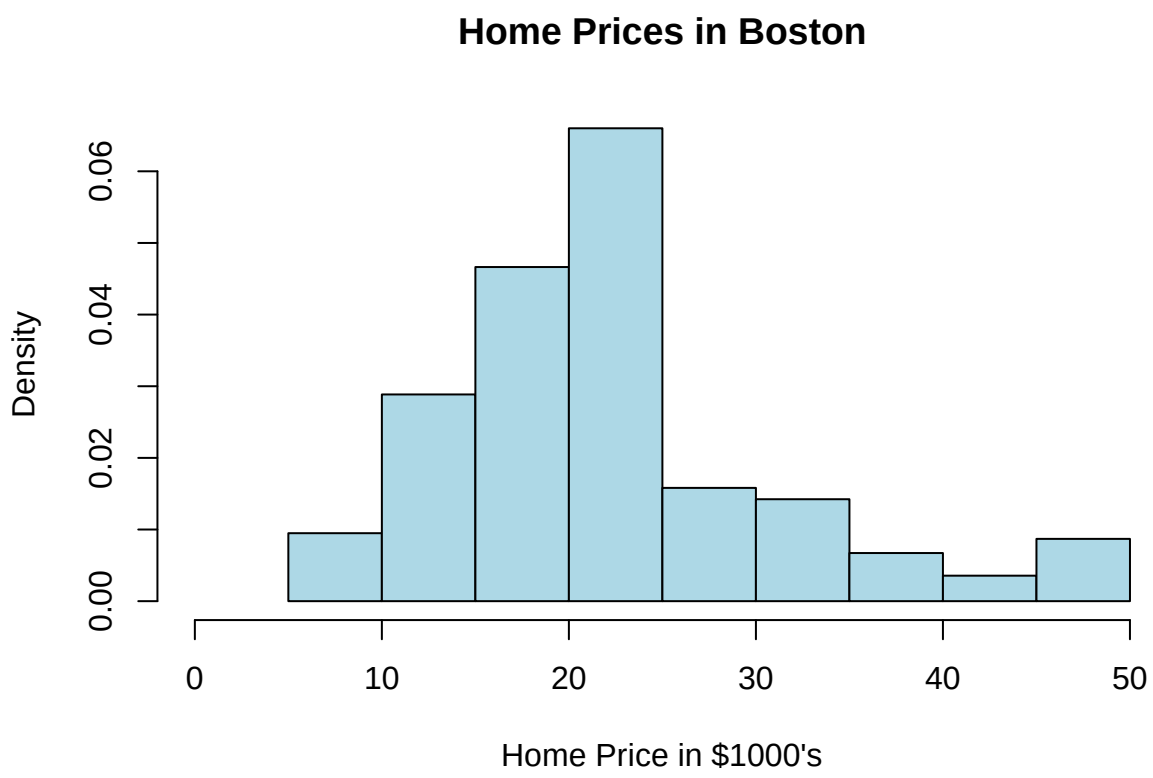
```
## # A tibble: 2 x 2
##   cchas          `table(boston_housing$chas)`
##   <chr>          <table>
## 1 non-Riverfront 471
## 2 Riverfront      35
```

As we can see from the summary table, the average price of a Boston home in the 1970s is 22.5328063, where it has 6.2846344 rooms, not by Charles River, 3.7950427 weighted distance to five Boston employment centres, and 9.5494071 of index accessibility to radial highways.

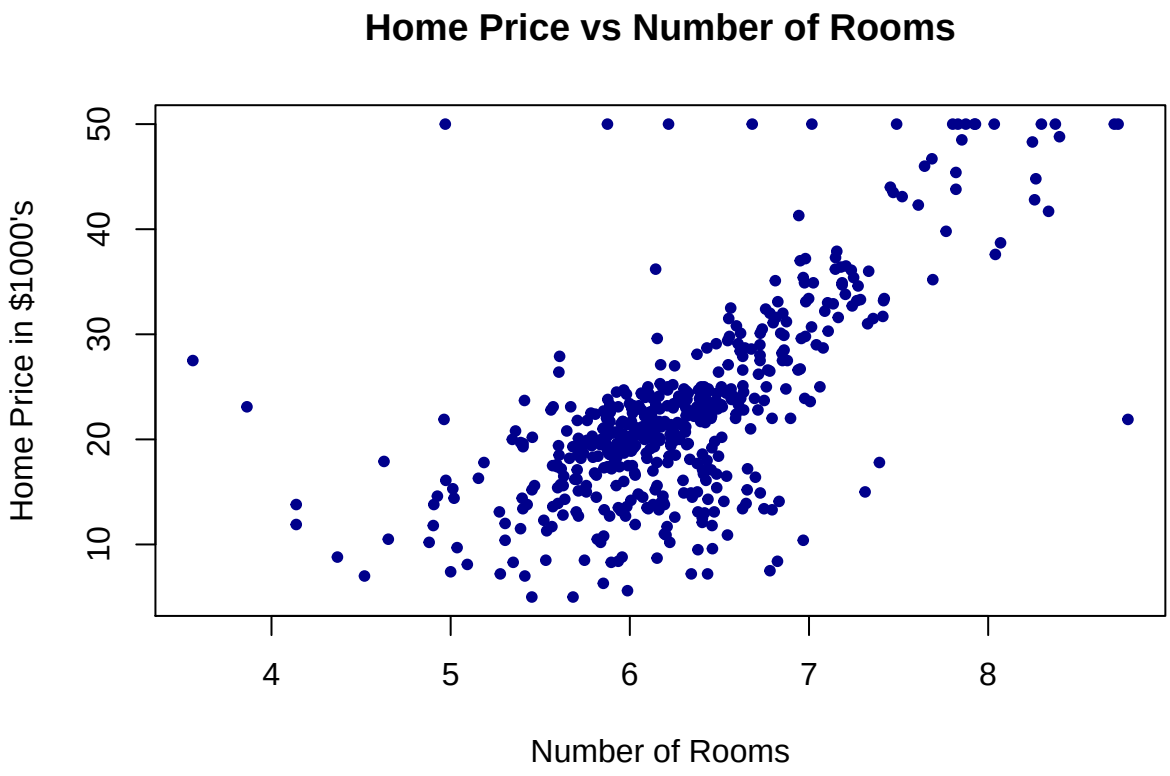
```
mean(boston_housing$rm)
mean(boston_housing$dis)
mean(boston_housing$rad)
mean(boston_housing$medv)
```

We can look deeper into the result above by plotting them out.

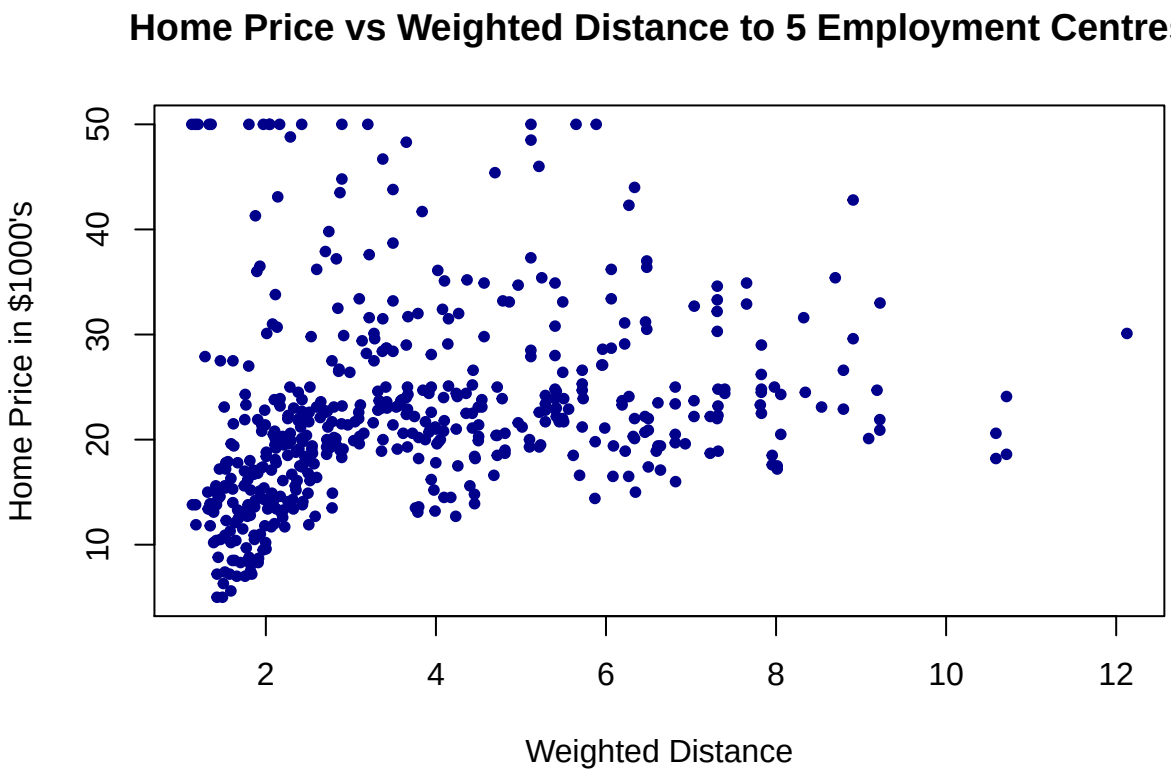
```
bhmmedv<-proj_bhm$medv
bhmrm<-proj_bhm$rm
bhmchas<-proj_bhm$chas
bhmdis<-proj_bhm$dis
bhmrad<-proj_bhm$rad
hist(bhmmedv,main="Home Prices in Boston",xlab="Home Price in $1000's",xlim=c(0,50),col=("lightblue"),f
```



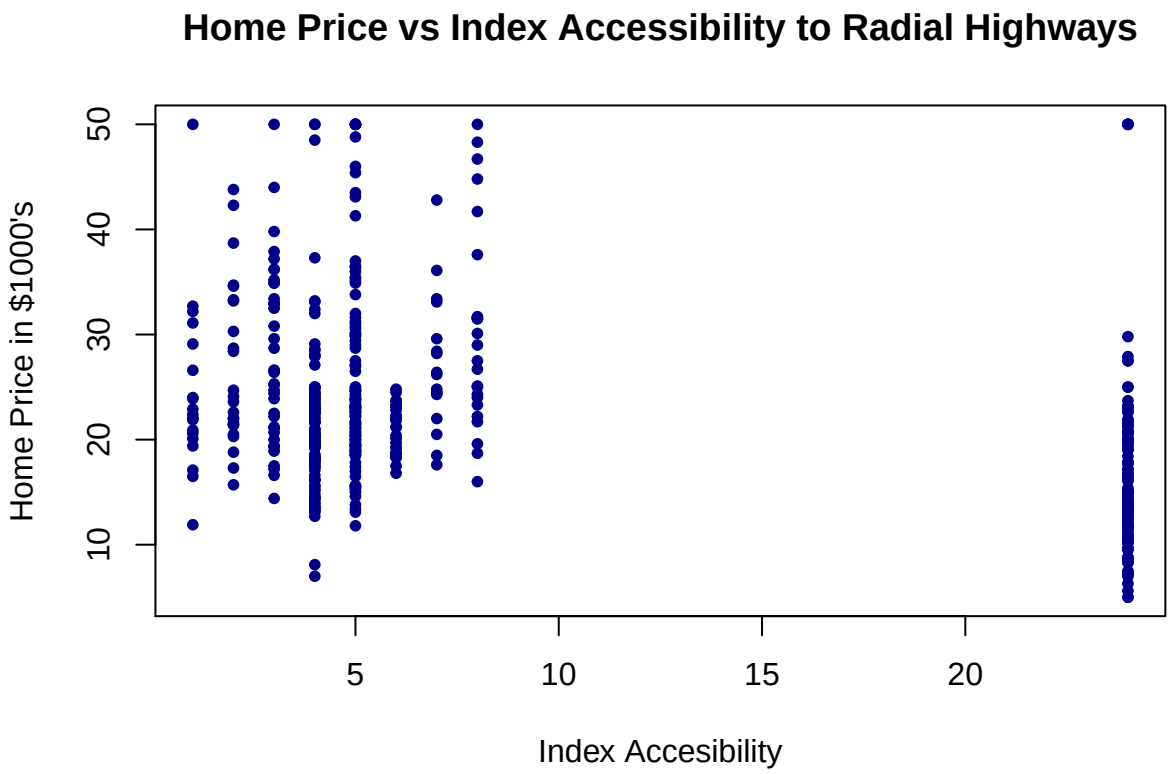
```
plot(bhmrm,bhmmedv,main="Home Price vs Number of Rooms",xlab="Number of Rooms",ylab="Home Price in $100
```



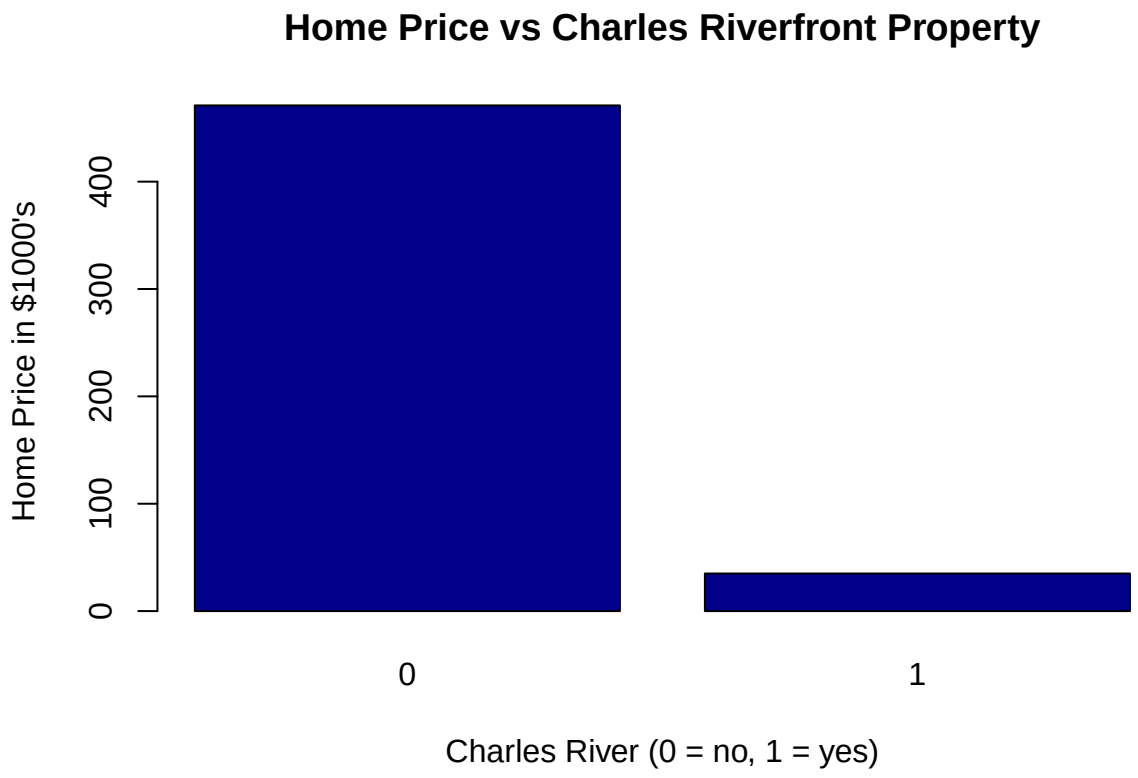

```
plot(bhmdis,bhmmedv,main="Home Price vs Weighted Distance to 5 Employment Centres",xlab="Weighted Dista
```




```
plot(bhmrad,bhmmedv,main="Home Price vs Index Accessibility to Radial Highways",xlab="Index Accesibilit
```




```
barplot(table(proj_bhm$chas),main="Home Price vs Charles Riverfront Property",xlab="Charles River (0 = n
```



##Correlation Matrix

```
print(round(cor(proj_bhm),4))
```

```
##         medv      rm    chas     dis     rad
## medv  1.0000  0.6954  0.1753  0.2499 -0.3816
## rm    0.6954  1.0000  0.0913  0.2052 -0.2098
## chas  0.1753  0.0913  1.0000 -0.0992 -0.0074
## dis   0.2499  0.2052 -0.0992  1.0000 -0.4946
## rad  -0.3816 -0.2098 -0.0074 -0.4946  1.0000
```

Based on the correlation matrix table, the variable which has the strongest association with the home price is the number of rooms with a value of 0.6954, while the variable which has the weakest association with the home price is Charles Riverfront property with a value of 0.1753. There is no evidence of multicollinearity as the correlation among the four variables of rm, chas, dis, and rad are not larger than 0.7

##Multiple Regression Model

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_regress(bhmmedv ~ bhmrm + bhmchas + bhmdis + bhmrad, data = proj_bhm)
```

```
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.745       RMSE                 6.160
## R-Squared               0.555       Coef. Var           27.336
## Adj. R-Squared          0.551       MSE                 37.940
## Pred R-Squared          0.540       MAE                  4.121
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## --------------------------------------------------------------------------
##                  Sum of
##                 Squares         DF    Mean Square       F         Sig.
## --------------------------------------------------------------------------
## Regression    23708.468          4       5927.117    156.224    0.0000
## Residual      19007.827        501         37.940
## Total         42716.295        505
## --------------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta       t        Sig       lower      upper
## ----------------------------------------------------------------------------------------------
## (Intercept)    -27.455         2.642                -10.391    0.000    -32.646    -22.264
##       bhmrm      8.264         0.404        0.631    20.446    0.000      7.470      9.058
##     bhmchas      4.239         1.093        0.117     3.878    0.000      2.091      6.387
##      bhmdis      0.053         0.152        0.012     0.348    0.728     -0.246      0.352
##      bhmrad     -0.256         0.037       -0.242    -6.999    0.000     -0.328     -0.184
## ----------------------------------------------------------------------------------------------
```

I formulated the estimated regression equation using Ordinary Least Squares function. Price = -27.455 + 8.264bhmrm + 4.239bhmchas + 0.053bhmdis - 0.256bhmrad

Let us interpret the estimated regression equation. For an increase of 1 bedroom, we expect the home price to increase by $8264. When a home is a Charles Riverfront property, the home price is $4239 higher relative to a non-waterfront property. For a 1 unit increase weighted distances to five Boston employment centres, the home price increases by $5.3. For a one unit increase in index of accessibility to radial highways, the home price decreases by $256. The slope coefficient for number of rooms, Charles Riverfront property, and index of accessibility to radial highways are statistically significant as p-value < 0.05.