Problem 1

Topic: To investigate whether R-score and/or L-score for a student can be used to accurately predict the DIEBLS score.

We now analyze this problem from 5 aspects:

1) To see whether there is a noticeable difference of R-score and L-score for the students in these two groups, we divide them into two groups, specifically, the one with INP = 1, and the other with INP = 0.

   Then we perform a Welch two sample t test on R-score and L-score for two groups:

   - <u>R-score:</u>

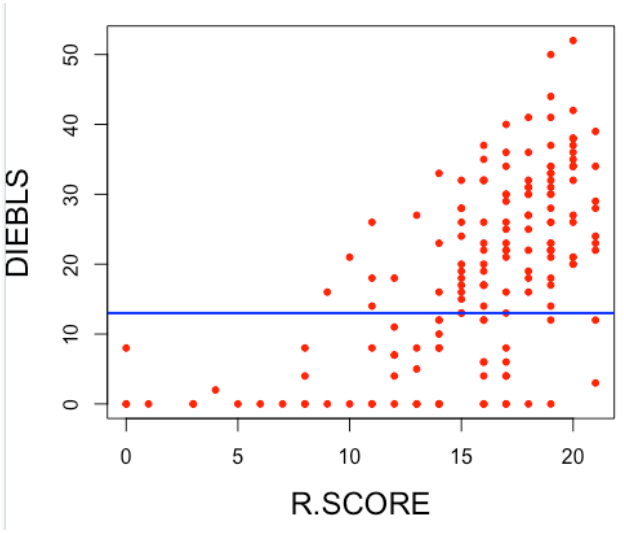
     H0: ux = uy

     H1: ux ≠ uy

```
> group1 <- filter(kread, INP == 1)
> group2 <- filter(kread, INP == 0)
> t.test(group1$R.SCORE, group2$R.SCORE)

        Welch Two Sample t-test

data:  group1$R.SCORE and group2$R.SCORE
t = -7.1754, df = 72.392, p-value = 5.113e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.676951 -3.773805
sample estimates:
mean of x mean of y
 12.28814  17.51351
```

   Here, we have a very small p-value, so we reject the null hypothesis and there is a difference in R-score in these two groups.

We can also have a visualization of R-scores of these two groups:



For those students that are assigned to group with INP = 0 (points above the blue line), meaning that these students do not need extra help, they all have an R-score of at least 9. Whereas those students who are assigned to the group with INP = 1 (points below or on the blue line) have a wide range of R-score. But at least we can tell that those students who have a passing score of DIEBLS are unlikely to have an R-score that is too low.

- <u>L-score:</u>

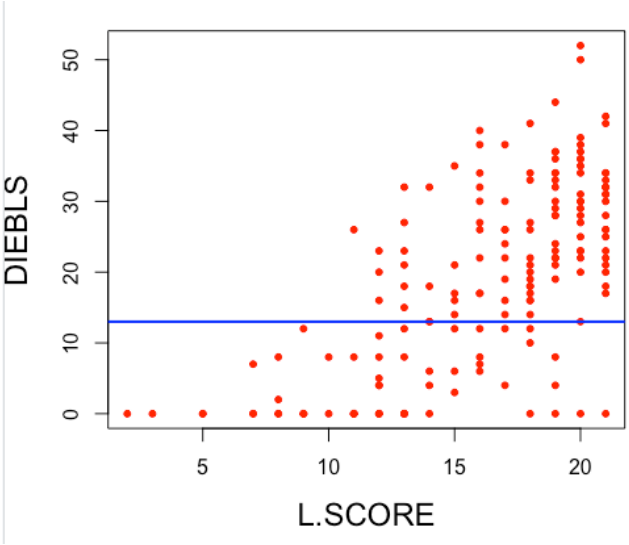
  H0: ux = uy

  H1: ux ≠ uy

```
> t.test(group1$L.SCORE, group2$L.SCORE)

        Welch Two Sample t-test

data:  group1$L.SCORE and group2$L.SCORE
t = -9.1847, df = 80.133, p-value = 3.755e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.845763 -4.407559
sample estimates:
mean of x mean of y
 12.50847  18.13514
```

We also have a very small p-value, so we reject the null hypothesis and there is a difference in L-score in these two groups.

We can also have a visualization on L-scores of these two groups:



For those students that are assigned to group with INP = 0 (points above the blue line), meaning that these students do not need extra help, they all have an L-score of at least 11. Whereas those students who are assigned to the group with INP = 1 (points below or on the blue line) have a wide range of L-score. But at least we can tell that those students who have a passing score of DIEBLS are unlikely to have an L-score that is too low.

Therefore, there is a significant difference in terms of both R-score and L-score in these two groups.

2) In order to predict DIEBLS, I want to use best subset selection to see which variables should be included. The candidates here are only R-score, L-score, and FARM, since ID# has nothing to do with DIEBLS and INP is totally dependent on DIEBLS. So we should ignore ID# and INP.

After running a forward best subset selection with BIC on R, we can see that the model we should use is the second one which only includes R-score and L-score as the covariates:

```
> regfit.full <- regsubsets(DIEBLS~.,select(kread, R.SCORE, L.SCORE, FARM,DIEBLS), nvmax= 3)
> reg.summary <- summary(regfit.full)
> reg.summary
Subset selection object
Call: regsubsets.formula(DIEBLS ~ ., select(kread, R.SCORE, L.SCORE,
    FARM, DIEBLS), nvmax = 3)
3 Variables  (and intercept)
        Forced in Forced out
R.SCORE     FALSE      FALSE
L.SCORE     FALSE      FALSE
FARMY       FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: exhaustive
         R.SCORE L.SCORE FARMY
1  ( 1 ) " "     "*"     " "
2  ( 1 ) "*"     "*"     " "
3  ( 1 ) "*"     "*"     "*"
> num_select <- which.min(reg.summary$bic)
> num_select
[1] 2
```

3) Next we want to see if our model can predict DIEBLS accurately. I then separate the data into training data and test data and fit a model we just derived on the training data and then see the test error and R-squared.

```
> select_coef <- coef(regfit.fwd,num_select)
> test.mat <- model.matrix(DIEBLS~., data = test)
> fbic_y_hat <- test.mat[,names(select_coef)] %*% select_coef
> test_err <- mean((test[, 2] - fbic_y_hat)^2)
> test_err
[1] 46.6628
```

```
> r2 <- 1- mean((test['DIEBLS'] - fbic_y_hat)^2) / mean((test['DIEBLS'] - test_mean)^2)
> r2
[1] 0.5234279
```

As we can see, we have a test error of around 46.66 and R-squared of around 0.52. Our model can only explain 52% of the data, so it does not predict quite accurately.

Because we only have a small size of data which only includes 185 students and we only have three covariates to incorporate in our model, which makes our model not very accurate.

4) For those students who have either R-score or L-score missing, I can check if they have another score together with DIEBLS that are similar to some other students who have full data. For example, ID#5 has a missing value of R-score and has an L-score of 6 and a DIEBLS score of 0. We can search for those students who have similar R-score and DIEBLS scores and use the mean of their R-scores as a prediction for the missing R-score for this student. There are no students with an L-score of 6, but we have ID# 3,4 with L-score of 5 and #ID 6,7 with L-score of 7, and they all have DIEBLS of 0. So we predict the R-score of ID#5 will be the mean of the R-scores of these four similar students, which is 9. We can do the same thing for those students who have an L-score missing.

5) One remaining question is whether FARM can be regarded as a predictor in our model for predicting DIEBLS.

   I perform an ANOVA F test with the null model incorporating only R-score and L-score as covariates and the full model incorporating R-score, L-score, and FARM as covariates:

```
> model1 <- lm(kread$DIEBLS~kread$R.SCORE + kread$L.SCORE)
> model2 <- lm(kread$DIEBLS~kread$R.SCORE + kread$L.SCORE + kread$FARM)
> anova(model1, model2)
Analysis of Variance Table

Model 1: kread$DIEBLS ~ kread$R.SCORE + kread$L.SCORE
Model 2: kread$DIEBLS ~ kread$R.SCORE + kread$L.SCORE + kread$FARM
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    167 14310
2    166 14310  1   0.13161 0.0015 0.9689
```

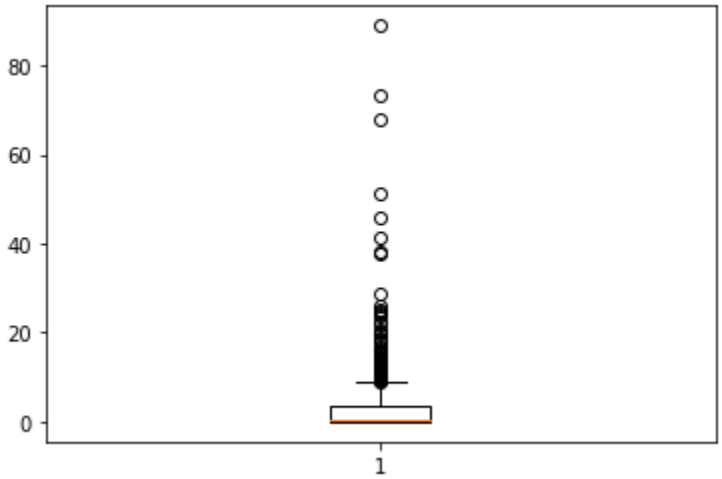
We can see that we have a p-value of 0.969, which means that we fail to reject our null model in this case. So we should only incorporate R-score and L-score in our model. Therefore, even though FARM can be seen as an indicator of readiness of reading for a student to some degree, it is not a good predictor to be included in our model.

Problem 2

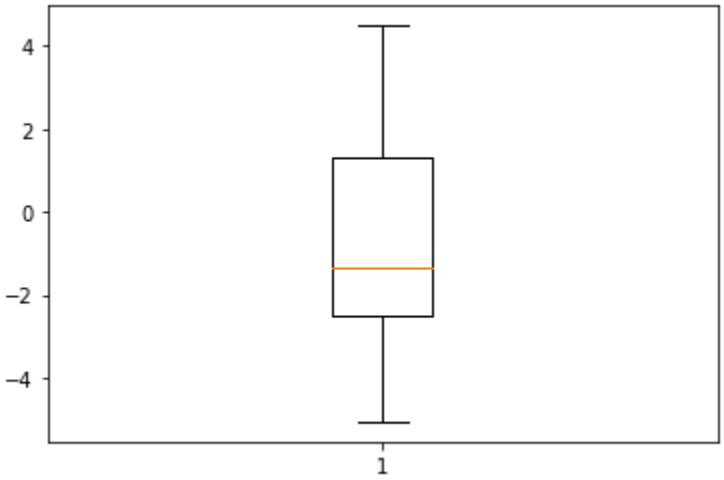
Part A

- **Per-Capira Crime Rate X1**

  Boxplot of X1:



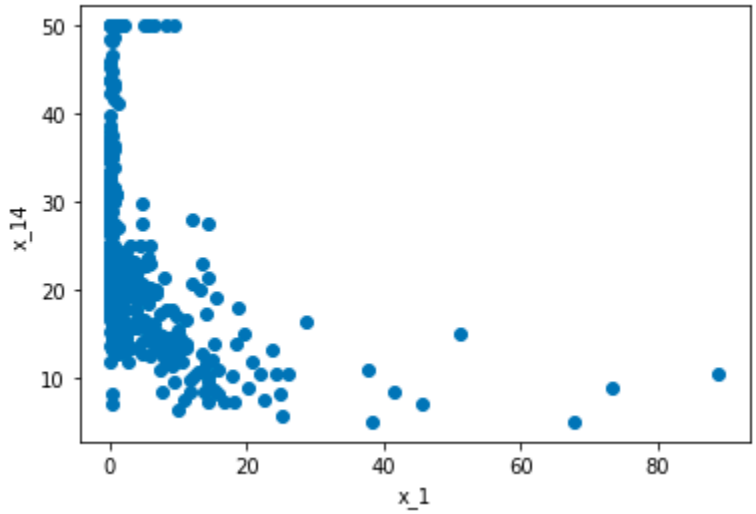
  Boxplot of log(X1):



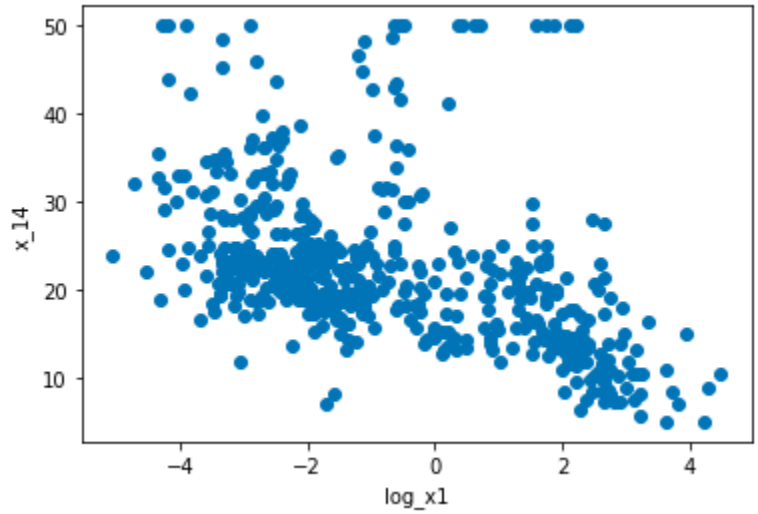
As we can see from the above two plots, boxplot of X1 is not informative at all. It has many outliers and the median is at the very downside of the plot. However, boxpolt

of log(X1) is much more informative. The median is around the middle of the dataset, and the distribution now is more symmetric than before.

Scatter plot of X1 versus X14:



Scatter plot of log(X1) versus X14:

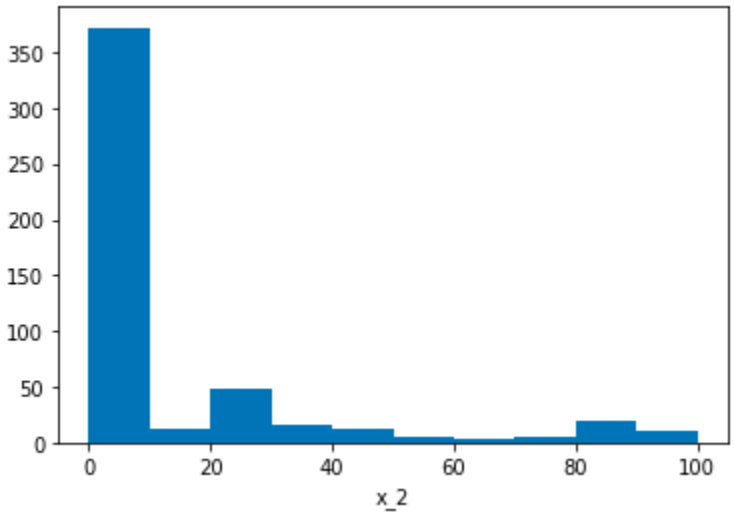


As we can see from the above two plots, the scatter plot of X1 and X14 has too many points clustered on the left of the plot, which makes it harder for us to predict the

median price. However, the distribution of points in the second plot suggests a negative relation, which gives us clues to predict the median.

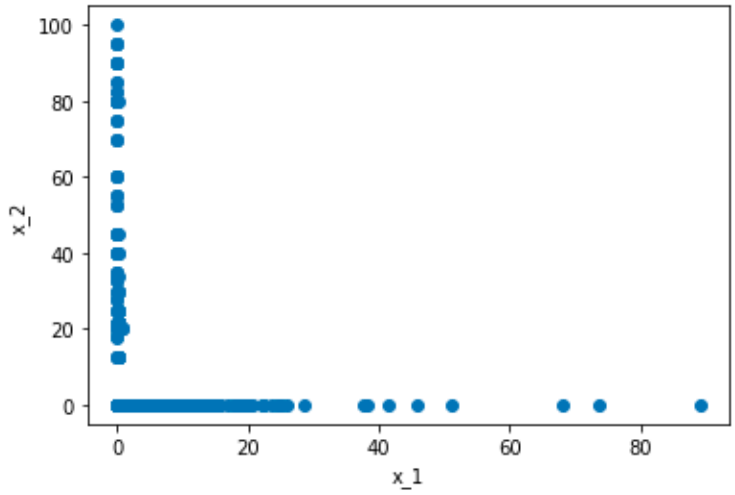
- **Proportion of Residual Area Zoned for Large Lots X2**

Histogram of X2:

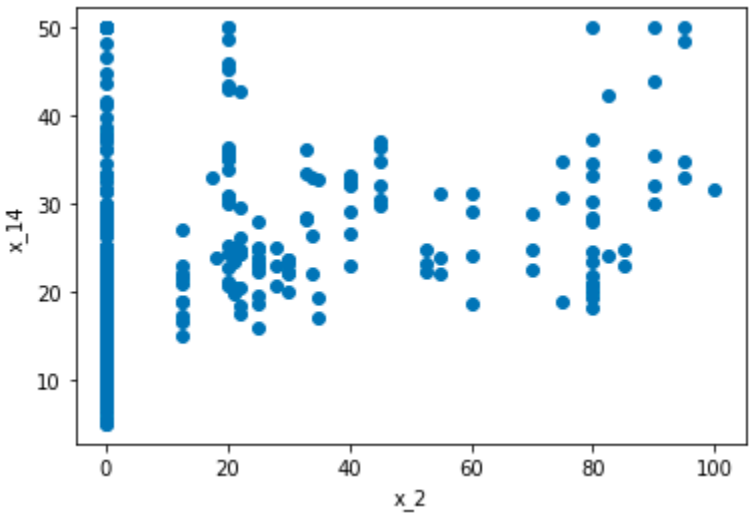


X2 has a very skewed distribution, and most of the data are clustered at around 0%-10%, meaning that very few areas are zoned for large lots.

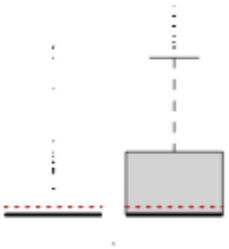
There is a strong negative relation between X1 and X2, the data with high values of X2 almost all have a zero value of X1, and many data with a zero value of X2 have a high value of X1 :



There is not a clear linear relation between X2 and X14 seeing from the scatter plot:

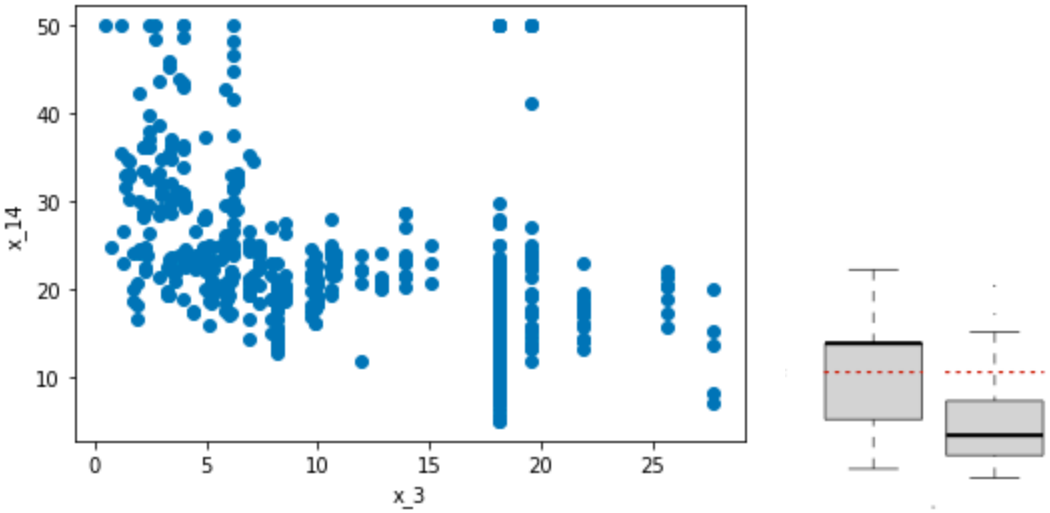


But from the box plot of X2, we can tell that X14 tend to be greater as X2 gets greater, so they actually have a positive relation:



- **Proportion of Non-retail Business Acres X3**

  From the below scatter plot of X3 and X14 and the box plot of X3, X3 tends to have a negative relation with X14, but the pattern is not so clear. But it is still reasonable to include X3 as an explanatory variable for X14 in the linear regression model.



- **Charles River Dummy Variable X4**

  From R code:

```
> nrow(subset(Boston, chas ==1))
[1] 35
```

  We can tell that 35 suburbs in this data set bound the Charles River.

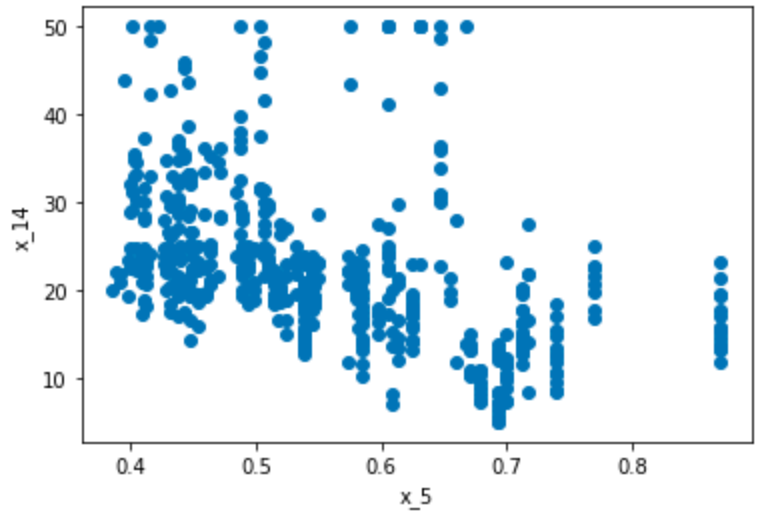
  To see if there is a relation between X4 and X14, we can separate the data into two groups, one with X4 = 1 and one with X4 = 0. Then we perform a 2 sample t test with these two groups to see if there is a significant difference in terms of X14:

```
> group1 <- filter(Boston, chas == 1)
> group2 <- filter(Boston, chas == 0)
> t.test(group1$medv, group2$medv)

        Welch Two Sample t-test

data:  group1$medv and group2$medv
t = 3.1133, df = 36.876, p-value = 0.003567
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.215483 10.476831
sample estimates:
mean of x mean of y
 28.44000  22.09384
```

Here we have a p-value smaller than 0.05, so there is a difference of X14 between the two groups, and it is reasonable to think there is a relation between X4 and X14. But whether this difference in the house price is totally due to the proximity to the river cannot be determined, since there may be other factors that make those areas with higher house prices tend to cluster together and they are coincidentally near the river.

- **Nitric Oxides Concentration X5**

  X5 seems to have a relation to X14 and the points indicate a negative association, see the plot below. And the correlation computed also suggests a negative association since the correlation has a negative value.




```
> cor(Boston$nox, Boston$medv)
[1] -0.4273208
```

- **Average Number of Rooms per Dewlling X6**

From R code, the correlation between X6 and X14 is about 0.695, which indicates they have a quite strong correlation.

```
> cor(Boston$rm, Boston$medv)
[1] 0.6953599
```

From R code, 64 of the suburbs average more than seven rooms per dwelling, and 13 of the suburbs average more than eight rooms per dwelling.

```
> nrow(subset(Boston, rm > 7))
[1] 64
> nrow(subset(Boston, rm > 8))
[1] 13
```

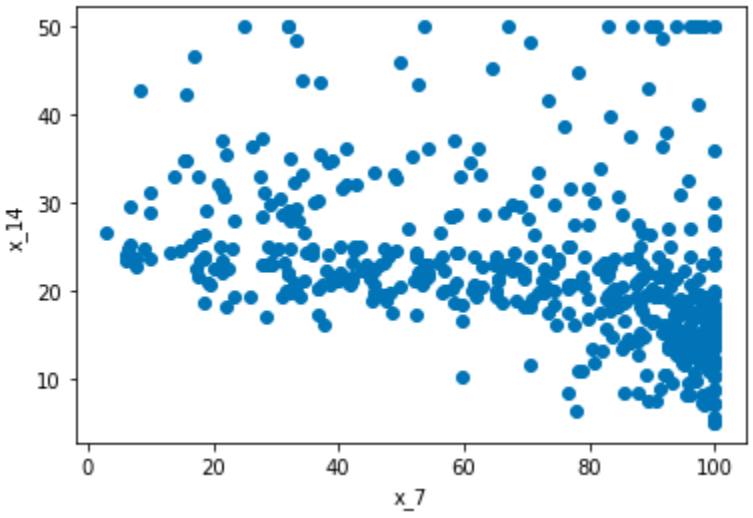
We can look at the summary of those 13 suburbs average more than eight rooms per dwelling and the summary of the overall dataset:

```
> summary(subset(Boston, rm > 8))
      crim               zn             indus             chas              nox               rm             age       
 Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000   Min.   :0.4161   Min.   :8.034   Min.   : 8.40  
 1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000   1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40  
 Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000   Median :0.5070   Median :8.297   Median :78.30  
 Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538   Mean   :0.5392   Mean   :8.349   Mean   :71.54  
 3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000   3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50  
 Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000   Max.   :0.7180   Max.   :8.780   Max.   :93.90  
      dis              rad              tax            ptratio          black            lstat           medv      
 Min.   :1.801   Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6   Min.   :2.47   Min.   :21.9  
 1st Qu.:2.288   1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5   1st Qu.:3.32   1st Qu.:41.7  
 Median :2.894   Median : 7.000   Median :307.0   Median :17.40   Median :386.9   Median :4.14   Median :48.3  
 Mean   :3.430   Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2   Mean   :4.31   Mean   :44.2  
 3rd Qu.:3.652   3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7   3rd Qu.:5.12   3rd Qu.:50.0  
 Max.   :8.907   Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9   Max.   :7.44   Max.   :50.0  

> summary(Boston)
      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.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285  
 3rd Qu.: 3.67708   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  
      age              dis              rad              tax            ptratio          black            lstat      
 Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73  
 1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95  
 Median : 77.50   Median : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36  
 Mean   : 68.57   Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65  
 3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95  
 Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97  
      medv      
 Min.   : 5.00  
 1st Qu.:17.02  
 Median :21.20  
 Mean   :22.53  
 3rd Qu.:25.00  
 Max.   :50.00  
```

From the above two tables, suburbs average more than eight rooms per dwelling generally have a lower crime rate (X1), so maybe those suburbs that are wealthier (can have more rooms per dwelling) also have a more secure living environment. Also, they have a lower tax (X10), so maybe tax is not directly related to the price of the house. They also have a lower proportion of the lower status of the population (X10) and higher median values of owner-occupied homes in $1000 (X14). These are also reasonable since these areas tend to be wealthier than the rest.

- **Proportion of Owner-Occupied Units Built Prior to 1940 X7**

  Seeing from the scatter plot of X7 and X14, there is not a visible connection, although there may be a not visible negative relation.



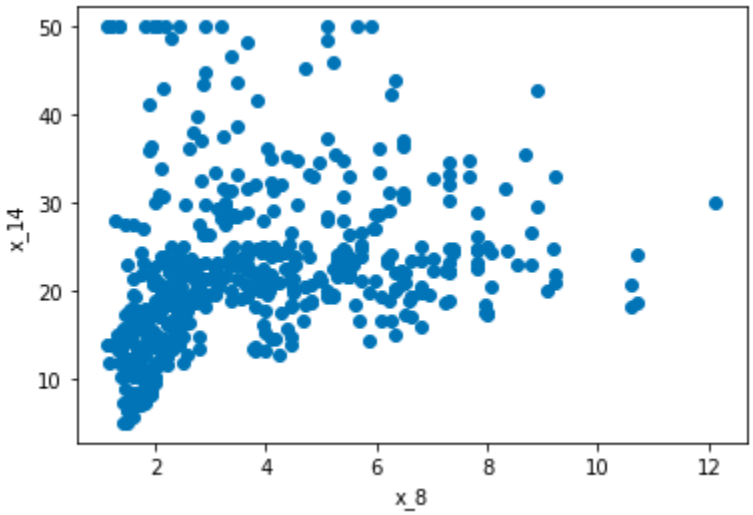
Therefore, the value of the house may not directly relate to age. Since there could be newer houses that have better facilities and environment so that they value more than those older houses. But there could also be houses that have appreciation due to their locations and historical instances. So the effect of age may be bidirectional.

- **Weighted Distance to Five Boston Employment Centers X8**

  No. There is actually a positive relation between the distances to the employment centers and house prices. Since the correlation between them is 0.249:

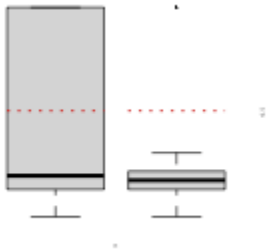
```
> cor(Boston$dis, Boston$medv)
[1] 0.2499287
```

Also, the scatter plot suggests a positive relation between them:



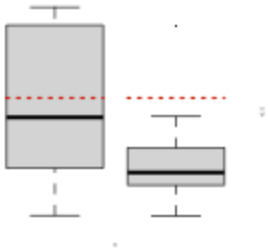
- **Index of Accessibility to Radial Highways X9**

No. As we can see from the two boxplots of X9, the median of X9 in the cheaper and more expensive houses are about the same.



- **Full-Value Property Tax X10**

Yes. As we can see from the two boxplots of X10, the subgroup with a lower value of X10 has a higher median indicating a more expensive house price, and the subgroup with a higher value of X10 has a lower median indicating a cheaper house price.

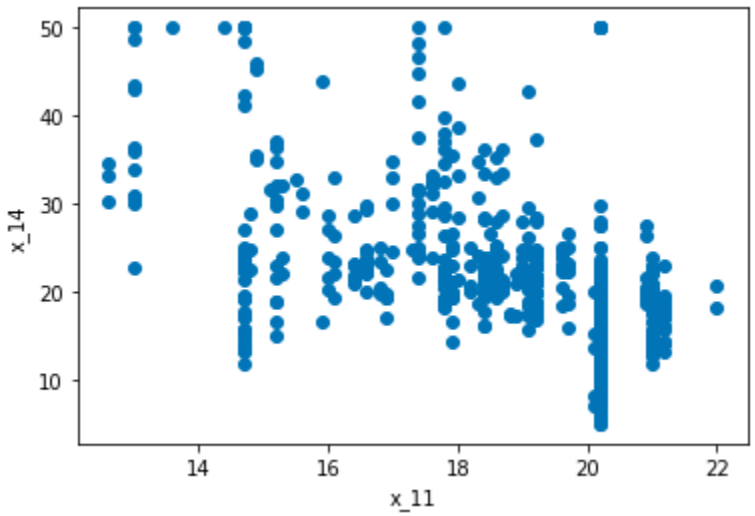


- **Pupil/Teacher Ratio X11**

  From the R code, the median pupil-teacher ratio is 19.05:

```
> median(Boston$ptratio)
[1] 19.05
```

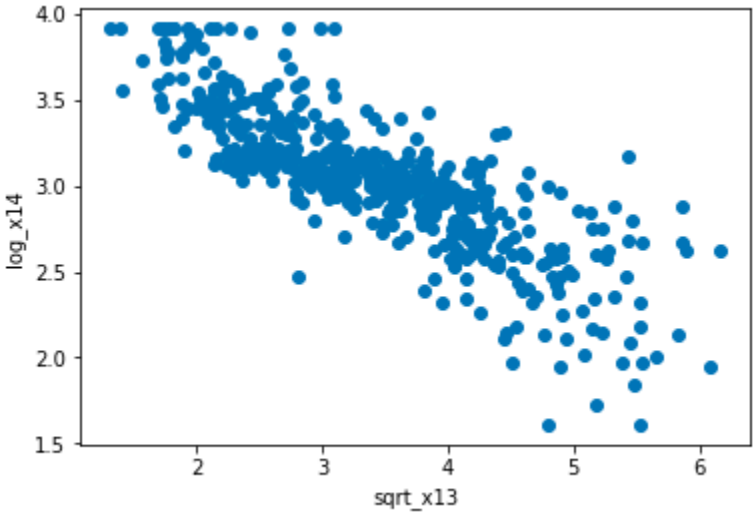
  As can be seen from the scatter plot, there is a negative relation between X11 and X14, indicating the higher the Pupil/Teacher Ratio, the lower the median of the house price tends to be, it is not very visible but we can tell the pattern here.



- **Proportion of Lower Status of the Population X13**

  There is a strong negative relation between X13 and X14, but the relation is not linear.

However, the square root of X13 has a clear linear relation with log(X14):



Part B/C

(a) Use R to make a linear model:

```
X1 = log(Boston$crim)
X2 = Boston$zn/10
X3 = log(Boston$indus)
X4 = Boston$chas
X5 = log(Boston$nox)
X6 = log(Boston$rm)
X7 = Boston$age**2.5/10000
X8 = log(Boston$dis)
X9 = log(Boston$rad)
X10 = log(Boston$tax)
X11 = exp(0.4*Boston$ptratio)/1000
X13 = Boston$lstat**(1/2)
X14 = log(Boston$medv)

lm.fit <- lm(X1~X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X13+X14)
```

We can look at the p-values of each covariate:

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.40901 -0.55398 -0.01097  0.52093  2.50886

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.68009    1.63664  -2.860  0.00442 **
X2          -0.03781    0.02204  -1.715  0.08694 .
X3           0.16057    0.08734   1.838  0.06660 .
X4           0.01277    0.14512   0.088  0.92989
X5           2.11023    0.40566   5.202  2.9e-07 ***
X6           0.26100    0.43382   0.602  0.54770
X7           0.06539    0.01880   3.478  0.00055 ***
X8          -0.30622    0.14738  -2.078  0.03825 *
X9           1.12767    0.07316  15.413  < 2e-16 ***
X10          0.53809    0.19193   2.804  0.00525 **
X11          0.02037    0.03248   0.627  0.53081
X13          0.04837    0.07836   0.617  0.53736
X14         -0.35259    0.17366  -2.030  0.04286 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7883 on 493 degrees of freedom
Multiple R-squared:  0.8702,    Adjusted R-squared:  0.8671
F-statistic: 275.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

At alpha = 0.05 significance level, we can reject the null hypothesis for X5, X7, X8, X9, X10, and X14 since they all have p-values smaller than 0.05.

(b) I tried the best subset selection with BIC, Ridge regression, and Lasso here.

- By running a forward best subset selection with BIC on the training set, the covariates that are kept as explanatory variables are X3, X5, X7, X9, X10, and X14

```
> best.fss <- var_name[idx[1:which.min(bic)]]
> best.fss
[1] "X9"  "X5"  "X7"  "X3"  "X14" "X10"
```

Fitting this new model, we can see the summary below:

```
Call:
lm(formula = y ~ X9 + X5 + X7 + X3 + X14 + X10)

Residuals:
     Min       1Q   Median       3Q      Max
-2.37502 -0.58040 -0.00766  0.52405  2.64030

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.97496    1.16654  -3.407 0.000709 ***
X9           1.15547    0.07167  16.122  < 2e-16 ***
X5           2.59624    0.34409   7.545 2.16e-13 ***
X7           0.08836    0.01625   5.437 8.50e-08 ***
X3           0.27218    0.07488   3.635 0.000307 ***
X14         -0.36335    0.11015  -3.299 0.001041 **
X10          0.45313    0.17744   2.554 0.010952 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7913 on 499 degrees of freedom
Multiple R-squared:  0.8676,    Adjusted R-squared:  0.866
F-statistic: 545.2 on 6 and 499 DF,  p-value: < 2.2e-16
```

The model has an R-squared around 0.87, which indicates the model fits the data quite well.

The test error of this model is around 0.567:

```
> regfit.fwd = regsubsets(X1~., data = new_data, nvmax= 12, method = 'forward')
> select_coef <- coef(regfit.fwd,num_select)
> test.mat <- model.matrix(X1~., data = test)
> fbic_y_hat <- test.mat[,names(select_coef)] %*% select_coef
> test_err <- mean((test[, 'X1'] - fbic_y_hat)^2)
> test_err
[1] 0.5667166
```

- Then we can also run a ridge regression on the training set, and we have a quite close test error of 0.589 and an R-squared of around 0.885:

```
> grid <- 10^ seq(4, -2, length = 100)
> mod.ridge <- cv.glmnet(train.mat, train[, 'X1'], alpha = 0, lambda = grid)
> lambda <- mod.ridge$lambda.min
> y_hat <- predict(mod.ridge, newx = test.mat, s = lambda)
> test_err <- mean((test[,'X1'] - y_hat)^2)
> test_err
[1] 0.5887895
```

```
> ridge_r2 <- 1 - mean((test[,'X1'] - ridge_y_hat)^2) / mean((test[,'X1'] - test_mean)^2)
> ridge_r2
[1] 0.8853831
```

- We can also run Lasso to fit on the training set, and we get a much higher test error of 0.66 and a lower R-square of around 0.84:

```
> lasfit = glmnet(train.mat, train[, 'X1'], alpha = 1)
> cv.out.las = cv.glmnet(train.mat, train[, 'X1'],alpha = 1)
> bestlam = cv.out.las$lambda.min
> index_min = which(cv.out.las$lambda == bestlam)
> lasso_y_hat = predict(lasfit,s=bestlam,newx=test.mat)
> test_err = mean((las.pred-test[, 'X1'])^2)
> test_err
[1] 0.6611002
> lasso_r2 <- 1- mean((test[,'X1'] - lasso_y_hat)^2) / mean((test[,'X1'] - test_mean)^2)
> lasso_r2
[1] 0.8420934
```

From the above three approaches, best subset selection with BIC and Ridge Regression perform quite equally, they have a smaller value of test error and a greater value of R-squared.

(c) The model selected by best subset selection has a good performance on this data set.

- Below are the sets of models selected by best subset selection:

```
1 subsets of each size up to 12
Selection Algorithm: exhaustive
          X2  X3  X4  X5  X6  X7  X8  X9  X10 X11 X13 X14
1  ( 1 )  " " " " " " " " " " " " "*" " " " " " " " " "
2  ( 1 )  " " " " " " "*" " " " " " " "*" " " " " " " " "
3  ( 1 )  " " " " " " "*" " " "*" " " "*" " " " " " " " "
4  ( 1 )  " " " " " " "*" " " "*" " " "*" "*" " " " " " "
5  ( 1 )  " " "*" " " "*" " " "*" " " "*" "*" " " " " " "
6  ( 1 )  " " "*" " " "*" " " "*" " " "*" "*" " " " " "*"
7  ( 1 )  "*" " " " " " "*" " " "*" "*" "*" "*" " " " " "*"
8  ( 1 )  "*" "*" " " " "*" " " "*" "*" "*" "*" " " " " "*"
9  ( 1 )  "*" "*" " " "*" " " "*" "*" "*" "*" " " " "*" "*"
10  ( 1 ) "*" "*" " " "*" "*" "*" "*" "*" "*" " " " "*" "*"
11  ( 1 ) "*" "*" " " "*" "*" "*" "*" "*" "*" "*" "*" "*"
12  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

To determine which model we should choose, we can compute cross-validation error for those models and select the one that minimizes the cross-validation error, which is model 8 and it has a CV error of around 0.607:

```
> # CV error
> regfit.best = regsubsets(X1~., data = train, nvmax= 12)
> test.mat = model.matrix(X1~., data = test)
> val.errors = rep(NA,12)
> for(i in 1:12){
+   coefi = coef(regfit.best, id = i)
+   pred = test.mat[,names(coefi)] %*% coefi
+   val.errors[i]= mean((test[,'X1']- pred)^2)
+ }
> val.errors
 [1] 1.2164517 0.6893321 0.6215957 0.6654429 0.6362676 0.6241610 0.6135167 0.6067322 0.6104166 0.6095205 0.6083400
[12] 0.6083467
> which.min(val.errors)
[1] 8
```

Model 8 includes covariate X2, X3, X5, X7, X8, X9, X10, and X14.

- We can also look at the CV error for ridge regression, which is around 0.640:

```
> ridfit = glmnet(train.mat,train[, 'X1'],alpha = 0)
> cv.out.rid=cv.glmnet(train.mat,train[, 'X1'], nfolds = 10)
> bestlam=cv.out.rid$lambda.min
> index_min = which(cv.out.rid$lambda == bestlam)
> cv.out.rid$cvm[index_min] #CV error
[1] 0.6396284
```

- And the CV error for Lasso is around 0.612:

```
> lasfit = glmnet(train.mat, train[, 'X1'], alpha = 1)
> cv.out.las = cv.glmnet(train.mat, train[, 'X1'],alpha = 1)
> bestlam = cv.out.las$lambda.min
> index_min = which(cv.out.las$lambda == bestlam)
> cv.out.las$cvm[index_min]
[1] 0.61214
```

Therefore, by comparing the CV errors of all these models, best subset selection seems to have a better performance.

(d) I did not include all the covariates in my model, only X2, X3, X5, X7, X8, X9, X10, and X14 are included. Since I am performing best subset selection here, I have a set of models each incorporates several covariates, and I use cross-validation error as an indicator to decide which model I should choose. Then the model which minimizes the cross-validation error is my desired model, so this model is not guaranteed to incorporate all the covariates presented. Since incorporating those other covariates may increase the cross-validation error. Only the most relevant ones are being selected so that the cross-validation error is minimized.