Assignment 3

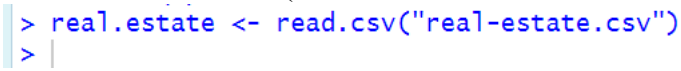
Due: Wednesday, October 9, 2019

This assignment relates to the 'Real-estate' data set. People buying or selling houses would like to know how much they can expect to get, or pay, for a property. This is also a concern for those who are making mortgage loans, or for those taxing real estate (and who are more likely to commission statistical studies than individual home-owners). The price of a house depends on its physical characteristics, including size, features, quality of construction, age, etc. It also depends on location, and current market characteristics. You are approached by a research group which has a data on a sample of residential sales in a midwestern city; the variables are described in Table below.

Variable Name	Description
Sales price	Sales price of residence (dollars)
Finished square feet	Finished area of residence (square feet)
Number of bedrooms	Total number of bedrooms in residence
Number of bathrooms	Total number of bathrooms in residence
Air conditioning	Presence or absence of air conditioning: 1 if yes; 0 otherwise
Garage size	Number of cars that garage will hold
Pool	Presence or absence of swimming pool: 1 if yes; 0 otherwise
Year built	Year property was originally constructed
Quality	1= high quality, 2 = medium, 3 = low
Lot size	Lot size (square feet)
Adjacent to highway	1 if the property is adjacent to a highway, 0 otherwise

1. Read the data into R. Call the loaded data "real.estate".

```
real.estate <- read.csv("real-estate.csv",row.names = "ID")
```



2. Answer the following sub-questions
   i) Use the "summary()" function to identify the types of variables. Which variables are categorical? Which variables are quantitative? Are there any concerns in the summary table? Explain.

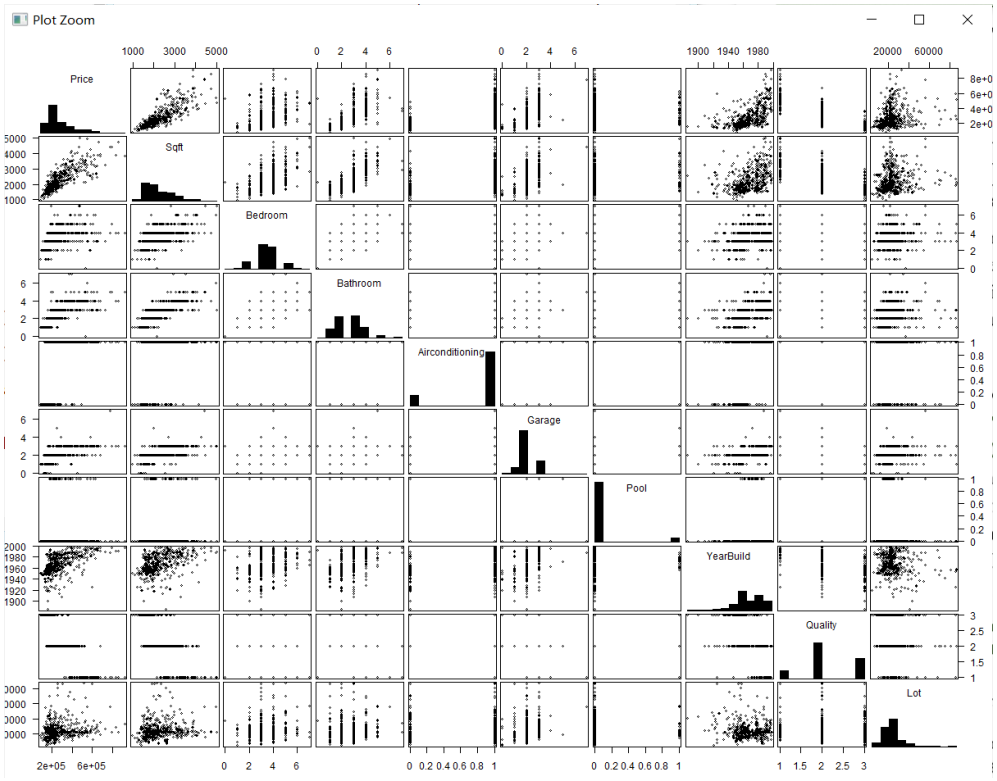
   All of the variables are quantitative. None of the variables are categorical. Yes. For the quantitative variable, the summary function will provides the minimum, maximum, quartiles, and mean. While for the categorical variable, the summary function only displays the frequencies in each category.

```
> summary(real.estate)
     Price             Sqft         Bedroom       Bathroom   Airconditioning
 Min.   : 84000   Min.   : 980   3      :202   3      :175   0:  7
 1st Qu.:180000   1st Qu.:1701   4      :179   2      :171   1: 52
 Median :229900   Median :2061   2      : 64   4      : 84   2:353
 Mean   :277894   Mean   :2261   5      : 52   1      : 71   3:106
 3rd Qu.:335000   3rd Qu.:2636   6      : 12   5      : 17   4:  2
 Max.   :920000   Max.   :5032   1      :  9   7      :  2   5:  1
                                 (Other):  4   (Other):  2   7:  1
     Garage    Pool      YearBuild    Quality      Lot         AdjHighway
 Min.   :0.0   0:486   Min.   :1885   1: 68   Min.   : 4560   0:511
 1st Qu.:2.0   1: 36   1st Qu.:1956   2:290   1st Qu.:17205   1: 11
 Median :2.0           Median :1966   3:164   Median :22200
 Mean   :2.1           Mean   :1967           Mean   :24370
 3rd Qu.:2.0           3rd Qu.:1981           3rd Qu.:26787
 Max.   :7.0           Max.   :1998           Max.   :86830

 Air
 0:  7
 1: 52
 2:353
 3:106
 4:  2
 5:  1
 7:  1
```

ii) Use the "pairs()" or "gpairs()" function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first then columns of a matrix A using A [,1:10]. Is there any interesting patterns? Which variables seem associated with the sales price? Explain.

Yes, Price is almost proportional to Sqft, Bedroom, Bathroom, Garage, Airconditon, YearBuild, Quality, Lot. But, what's interesting is it's negative proportional to the Pool. All the variables seem associated with the sales price. Because the scatterplot depicts the positive or negative proportional relationship.



iii) Use the "as.factor" function to regenerate categorical variables.

```
real.estate$Bathroom <- as.factor(real.estate$Bathroom)
real.estate$Bedroom <- as.factor(real.estate$Bedroom)
real.estate$Airconditioning <- as.factor(real.estate$Garage)
real.estate$Pool <- as.factor(real.estate$Pool)
real.estate$Quality <- as.factor(real.estate$Quality)
real.estate$AdjHighway <- as.factor(real.estate$AdjHighway)
summary(real.estate)
```

```
> summary(real.estate)
       ID              Price             Sqft          Bedroom       Bathroom
 Min.   :  1.0   Min.   : 84000   Min.   : 980   3       :202   3       :175
 1st Qu.:131.2   1st Qu.:180000   1st Qu.:1701   4       :179   2       :171
 Median :261.5   Median :229900   Median :2061   2       : 64   4       : 84
 Mean   :261.5   Mean   :277894   Mean   :2261   5       : 52   1       : 71
 3rd Qu.:391.8   3rd Qu.:335000   3rd Qu.:2636   6       : 12   5       : 17
 Max.   :522.0   Max.   :920000   Max.   :5032   1       :  9   7       :  2
                                                 (Other):  4   (Other):  2
 Airconditioning     Garage    Pool      YearBuild    Quality        Lot
 0:  7           Min.   :0.0   0:486   Min.   :1885   1: 68   Min.   : 4560
 1: 52           1st Qu.:2.0   1: 36   1st Qu.:1956   2:290   1st Qu.:17205
 2:353           Median :2.0           Median :1966   3:164   Median :22200
 3:106           Mean   :2.1           Mean   :1967           Mean   :24370
 4:  2           3rd Qu.:2.0           3rd Qu.:1981           3rd Qu.:26787
 5:  1           Max.   :7.0           Max.   :1998           Max.   :86830
 7:  1
 AdjHighway
 0:511
 1: 11
```

3. Fit the models and address the following when building that model:

i) Fit the null model and the full model

```
> null<- lm(Price~1-ID, data=real.estate)
> full <- lm(Price ~ . -ID, data=real.estate)
```

ii) Find ***the best sets*** of predictors using the stepwise procedures.

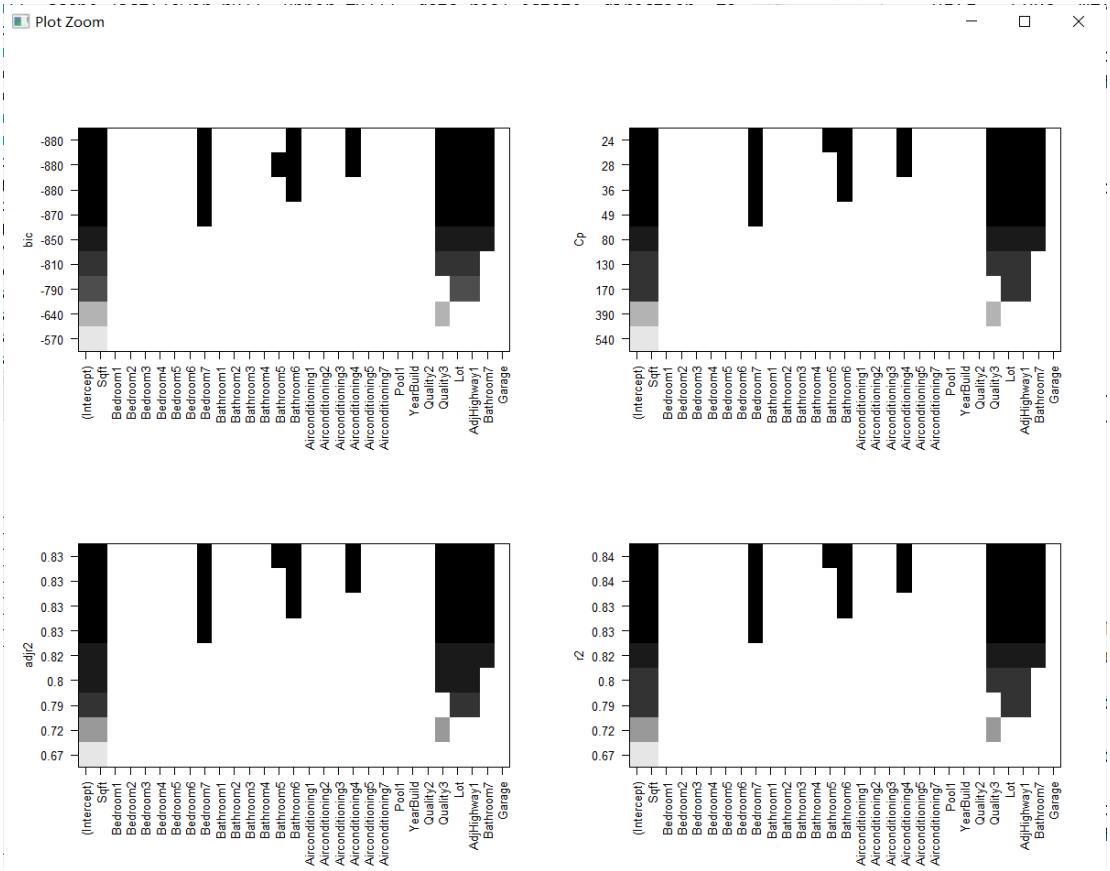
```
Step:  AIC=11427.44
Price ~ Sqft + Quality + YearBuild + Lot + Bedroom + Bathroom +
    Airconditioning + Pool + AdjHighway

                 Df  Sum of Sq        RSS   AIC
<none>                          1.5141e+12 11427
- AdjHighway      1 6.1826e+09 1.5203e+12 11428
- Pool            1 1.2707e+10 1.5268e+12 11430
- Airconditioning  6 4.7379e+10 1.5615e+12 11432
- Bathroom        6 1.0733e+11 1.6215e+12 11451
- Bedroom         6 1.1642e+11 1.6305e+12 11454
- YearBuild       1 1.4439e+11 1.6585e+12 11473
- Lot             1 1.5659e+11 1.6707e+12 11477
- Sqft            1 5.6046e+11 2.0746e+12 11590
- Quality         2 5.7862e+11 2.0927e+12 11592

Call:
lm(formula = Price ~ Sqft + Quality + YearBuild + Lot + Bedroom +
    Bathroom + Airconditioning + Pool + AdjHighway, data = real.estate)

Coefficients:
     (Intercept)             Sqft         Quality2         Quality3         YearBuild
      -2.308e+06        8.851e+01       -1.364e+05       -1.300e+05        1.304e+03
             Lot         Bedroom1         Bedroom2         Bedroom3         Bedroom4
       1.604e+00       -1.645e+05       -1.489e+05       -1.546e+05       -1.536e+05
        Bedroom5         Bedroom6         Bedroom7        Bathroom1        Bathroom2
      -1.440e+05       -1.716e+05       -3.565e+05        6.263e+04        6.842e+04
       Bathroom3        Bathroom4        Bathroom5        Bathroom6        Bathroom7
       9.183e+04        9.244e+04        1.260e+05       -1.082e+05               NA
Airconditioning1 Airconditioning2 Airconditioning3 Airconditioning4 Airconditioning5
      -3.465e+04       -3.355e+04       -8.401e+03       -7.349e+04       -4.159e+04
Airconditioning7            Pool1       AdjHighway1
       4.423e+04        2.082e+04       -2.433e+04
```

iii) Find ***the best sets*** of predictors using the best subset approach.



Bic: 7peaks; Cp:7peaks; Adjr2: 9 peaks; r2: 9 peaks
Price ~ Sqft + Quality + YearBuild + Lot + Garage + Bedroom + AdjHighway + Air + Airconditioning + Bathroom + Pool

iv) Considering the models in part (ii) and part(iii), choose the best model.

```
> anova(best1, best2)
Analysis of Variance Table

Model 1: Price ~ Sqft + Quality + YearBuild + Lot + Bathroom + Airconditioning +
    AdjHighway
Model 2: Price ~ Sqft + Quality + YearBuild + Lot + Bedroom + Bathroom +
    Airconditioning + Pool + AdjHighway
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    502 1.6383e+12
2    495 1.5141e+12  7 1.2416e+11 5.7985 1.785e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

best=lm(formula = Price ~ Sqft + Quality + YearBuild + Lot + Bedroom + Bathroom + Airconditioning + Pool + AdjHighway, data = real.estate)

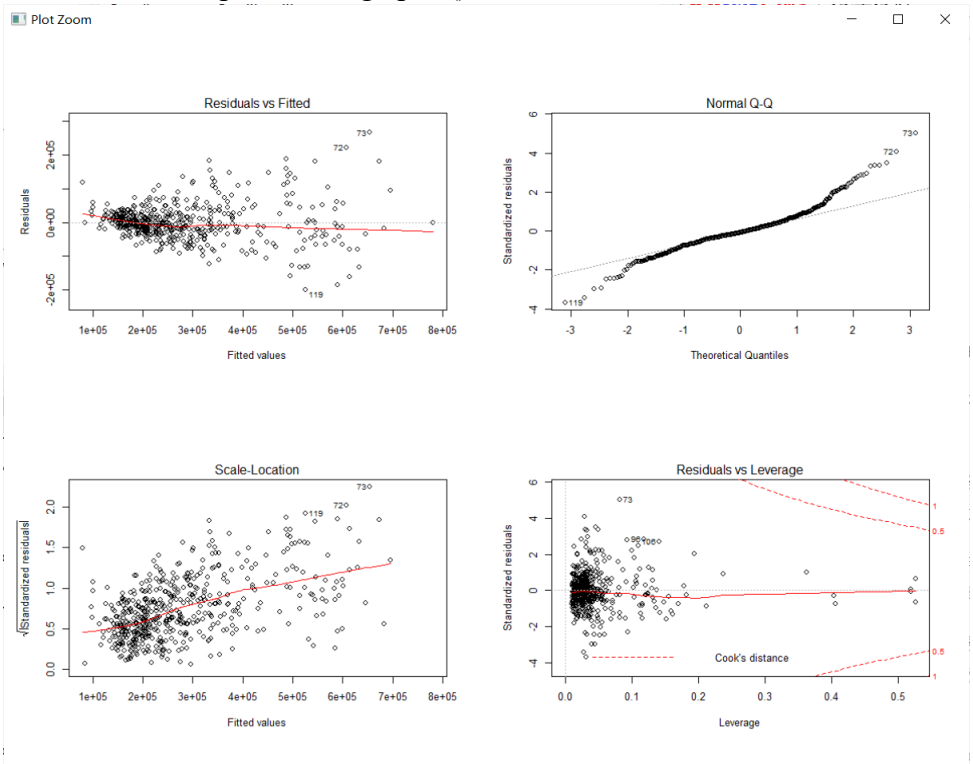
v) Interpret the coefficients of the best model ***in this context***.

The Price increases with the Sqft, the Beadroom. And the closer with the highway, the lower the Price becomes. And the price decays with the Airconditioning. The worse it is, the lower the price is. Price is Proportional to the Sqft, YearBuild, Lot, Bathroom, Pool. And it is negative proportional to the Airconditioning, Quality, Bedroom.

vi) Evaluate the best model.

```
Residual standard error: 66170 on 515 degrees of freedom
Multiple R-squared:  0.7725,    Adjusted R-squared:  0.7698
F-statistic: 291.4 on 6 and 515 DF,  p-value: < 2.2e-16
```

vii) Check the assumptions using "plot()" function.



1. The residuls are distributed with normal distribution form the top two figures.
2. But the homoscedasticity may be violated from the third plot.

viii) Continue exploring the data and provide a brief summary of what you discover.

```
shrinkage <- function(fit,k=10){

 require(bootstrap)


 # define functions

 theta.fit <- function(x,y){lsfit(x,y)}

 theta.predict <- function(fit,x){cbind(1,x)%*%fit$coef}


 # matrix of predictors

 x <- fit$model[,2:ncol(fit$model)]

 # vector of predicted values

 y <- fit$model[,1]


 results <- crossval(x,y,theta.fit,theta.predict,ngroup=k)

 r2 <- cor(y, fit$fitted.values)**2 # raw R2

 r2cv <- cor(y,results$cv.fit)**2 # cross-validated R2

 cat("Original R-square =", r2, "\n")

 cat(k, "Fold Cross-Validated R-square =", r2cv, "\n")

 cat("Change =", r2-r2cv, "\n")

}
```

```
> str(real.estate)
'data.frame':   522 obs. of  13 variables:
 $ ID             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Price          : int  360000 340000 250000 205500 275500 248000 229900 150000 195000 160000 ...
 $ Sqft           : int  3032 2058 1780 1638 2196 1966 2216 1597 1622 1976 ...
 $ Bedroom        : int  4 4 4 4 4 4 3 2 3 3 ...
 $ Bathroom       : int  4 2 3 2 3 3 2 1 2 3 ...
 $ Airconditioning: int  1 1 1 1 1 1 1 1 1 0 ...
 $ Garage         : int  2 2 2 2 2 5 2 1 2 1 ...
 $ Pool           : int  0 0 0 0 0 1 0 0 0 0 ...
 $ YearBuild      : int  1972 1976 1980 1963 1968 1972 1972 1955 1975 1918 ...
 $ Quality        : int  2 2 2 2 2 2 2 2 3 3 ...
 $ Lot            : int  22221 22912 21345 17342 21786 18902 18639 22112 14321 32358 ...
 $ AdjHighway     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Air            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

In the previous problems, we examine our model with my training data. This will be too optimistic. So here we use cross-validation to examine our model. We can see the original R-square is larger than the newer ten-fold cross-validated R-square. So we can redo our previous model selection with cross-validation. And we will reselect the best model.