STA302 Assignment 3

Manyi Luo - 1003799419

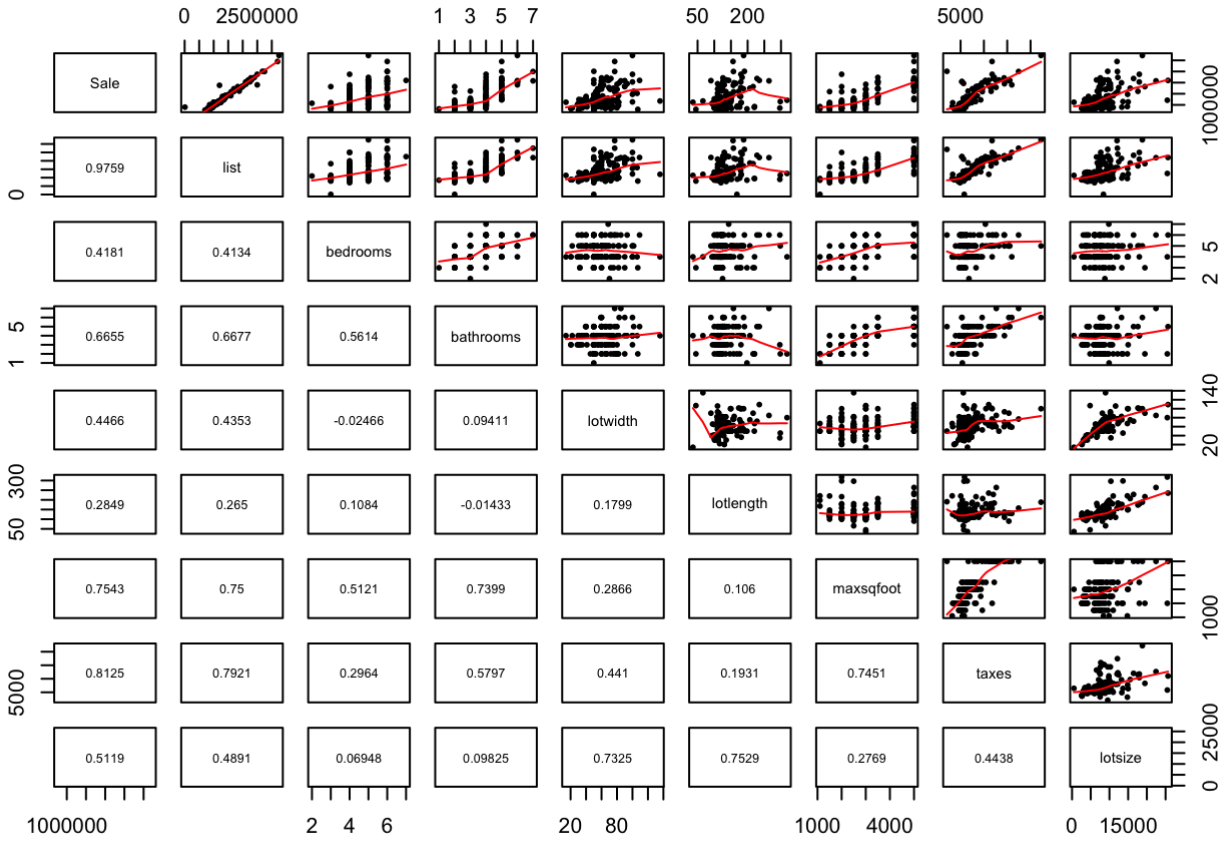
2019/11/29

Solution

Question 1.

```
## 'data.frame':    162 obs. of  11 variables:
##  $ Case_ID  : int  3 5 7 10 11 12 14 15 16 18 ...
##  $ Sale     : int  1038000 1150000 912000 1050000 1075000 1155000 860000 1110000 1128
000 1010000 ...
##  $ list     : int  1080000 1169000 919000 1089000 1100000 1150000 899000 1249000 1175
000 859000 ...
##  $ bedrooms : int  5 4 4 4 3 4 6 4 6 5 ...
##  $ bathrooms: int  4 4 3 3 3 3 3 2 4 4 ...
##  $ lotwidth : num  50 53.3 41.1 65.8 50 ...
##  $ lotlength: num  120 113.6 100.3 94.9 115 ...
##  $ maxsqfoot: int  3000 3000 2000 2500 1500 2000 2500 1500 3000 2000 ...
##  $ taxes    : int  6477 6494 5352 5706 5213 6067 5740 6032 5717 4829 ...
##  $ location : Factor w/ 2 levels "O","X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ lotsize  : num  6000 6057 4122 6242 5750 ...
```

```
## 'data.frame':    162 obs. of  9 variables:
##  $ Sale     : int  1038000 1150000 912000 1050000 1075000 1155000 860000 1110000 1128
000 1010000 ...
##  $ list     : int  1080000 1169000 919000 1089000 1100000 1150000 899000 1249000 1175
000 859000 ...
##  $ bedrooms : int  5 4 4 4 3 4 6 4 6 5 ...
##  $ bathrooms: int  4 4 3 3 3 3 3 2 4 4 ...
##  $ lotwidth : num  50 53.3 41.1 65.8 50 ...
##  $ lotlength: num  120 113.6 100.3 94.9 115 ...
##  $ maxsqfoot: int  3000 3000 2000 2500 1500 2000 2500 1500 3000 2000 ...
##  $ taxes    : int  6477 6494 5352 5706 5213 6067 5740 6032 5717 4829 ...
##  $ lotsize  : num  6000 6057 4122 6242 5750 ...
```



From the pairwise correlation and scattor plot, we can see that the sale price has the highest correlation with list. The rank of all quantitative predictors' correlation coefficients with sale price are presented below (from highest to lowest):

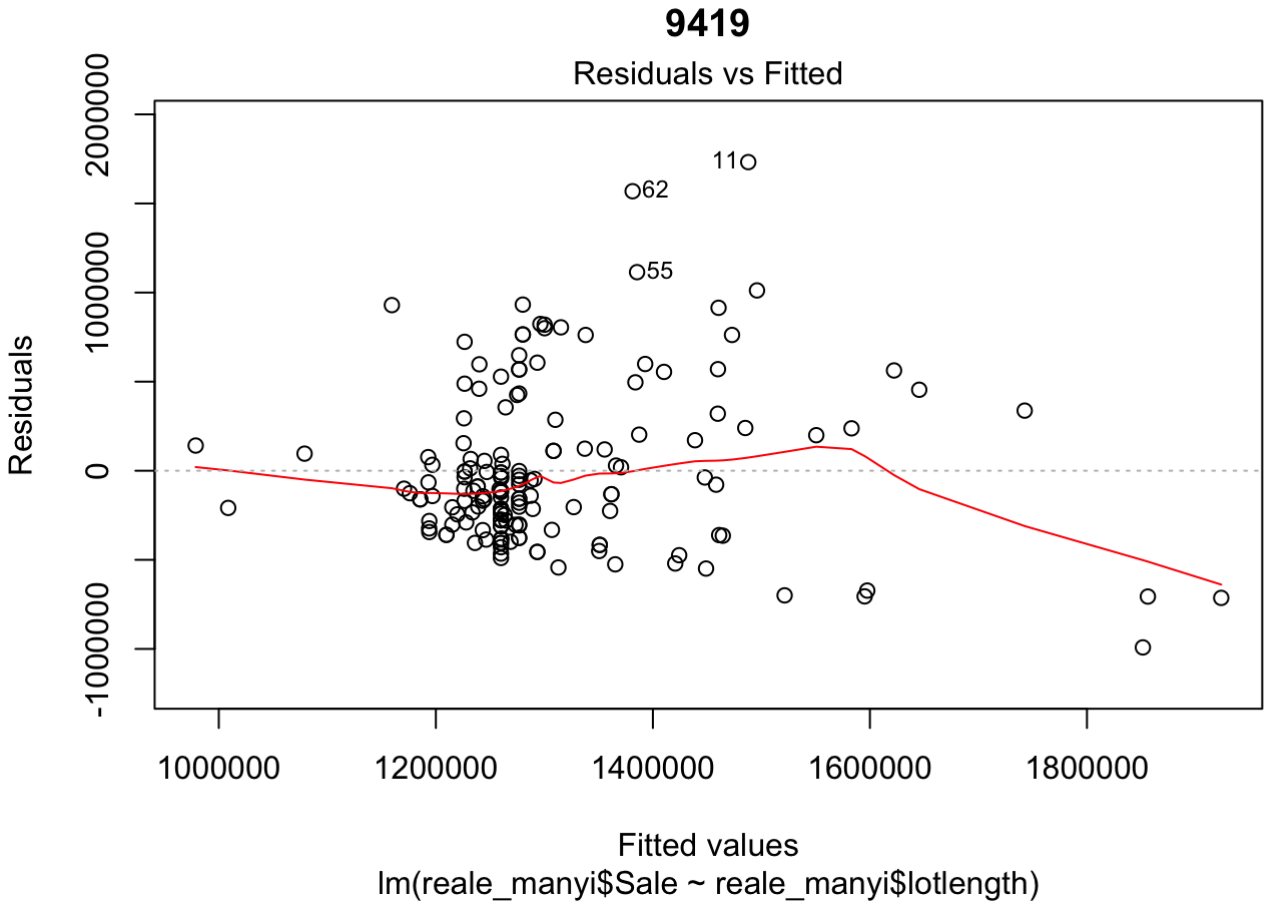
```
1. list: r = 0.9759
2. taxes: r = 0.8125
3. maxsqfoot: r = 0.7543
4. bathrooms: r = 0.6655
5. lotsize: r = 0.5109
6. lotwidth: r = 0.4466
7. bedrooms: r = 0.4181
8. lotlegth: r = 0.4181
```

Question 2.

(i). Refering back to the matrix in question 1, predictor lotlength may violate the assumption of constant variance. From the standardized residual plot presented below, the pattern of residual's variance is not randomly distributed. Its distribution is concentrated around 120000 (foot) and a trend of increasing variance can be identified when sale price is higher.

(ii).

```
##
## Call:
## lm(formula = reale_manyi$Sale ~ reale_manyi$lotlength)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -991525 -302093 -111825  192369 1732098
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            857845.8   125371.3   6.842 1.57e-10 ***
## reale_manyi$lotlength   3351.4      891.3   3.760 0.000238 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 449600 on 160 degrees of freedom
## Multiple R-squared:  0.08118,    Adjusted R-squared:  0.07544
## F-statistic: 14.14 on 1 and 160 DF,  p-value: 0.0002379
```



(iii). In order to conquer this problem, we can apply transformation to predictor or or weighted least square method to solve the unequal variance.

Question 3.

(i).

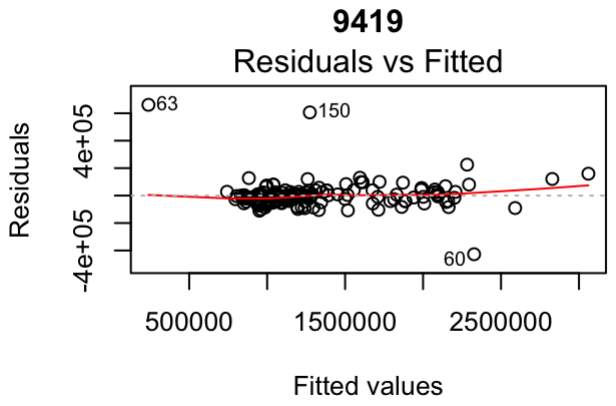
```
##
## Call:
## lm(formula = reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$lotwidth + reale_manyi$lotlength +
##     reale_manyi$maxsqfoot + reale_manyi$taxes + as.factor(reale_manyi$location) +
##     reale_manyi$lotsize)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -427339  -40483   -6274   19780  661411
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       6.779e+04  8.483e+04   0.799  0.42550
## reale_manyi$list                  7.568e-01  2.962e-02  25.553  < 2e-16
## reale_manyi$bedrooms              9.108e+03  1.049e+04   0.869  0.38648
## reale_manyi$bathrooms             9.594e+03  1.230e+04   0.780  0.43660
## reale_manyi$lotwidth             -4.092e+02  1.061e+03  -0.386  0.70037
## reale_manyi$lotlength            -1.119e+02  5.475e+02  -0.204  0.83829
## reale_manyi$maxsqfoot             3.898e+00  1.434e+01   0.272  0.78617
## reale_manyi$taxes                 1.450e+01  4.685e+00   3.094  0.00235
## as.factor(reale_manyi$location)X -1.086e+03  1.613e+04  -0.067  0.94642
## reale_manyi$lotsize               7.988e+00  7.838e+00   1.019  0.30976
##
## (Intercept)
## reale_manyi$list                 ***
## reale_manyi$bedrooms
## reale_manyi$bathrooms
## reale_manyi$lotwidth
## reale_manyi$lotlength
## reale_manyi$maxsqfoot
## reale_manyi$taxes                **
## as.factor(reale_manyi$location)X
## reale_manyi$lotsize
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97890 on 152 degrees of freedom
## Multiple R-squared:  0.9586, Adjusted R-squared:  0.9562
## F-statistic: 391.2 on 9 and 152 DF,  p-value: < 2.2e-16
```

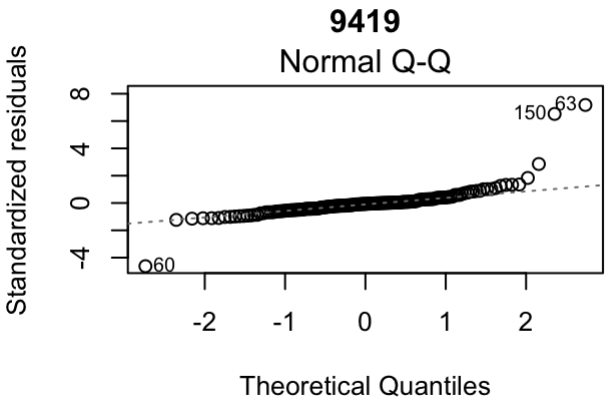
(ii). The estimated regression coefficients (Estimate) and the p-values (Pr(>|t|)) for the corresponding t-tests for these coefficients are listed below:

```
1. list: Estimate = 0.7568; Pr(>|t|) = < 2 * 10^-16 ***

2. bedrooms: Estimate = 9108; Pr(>|t|) = 0.38648

3. bathrooms: Estimate = 9594; Pr(>|t|) = 0.43660

4. lotwidth: Estimate = -409.2; Pr(>|t|) = 0.70037

5. lotlength: Estimate = -111.9; Pr(>|t|) = 0.83829

6. maxsqfoot: Estimate = 3.898; Pr(>|t|) = 0.78617

7. taxes: Estimate = 14.50; Pr(>|t|) = 0.00235 **

8. locationX: Estimate = -1086; Pr(>|t|) = 0.94642

9. lotsize: Estimate = 7.988; Pr(>|t|) = 0.30976
```

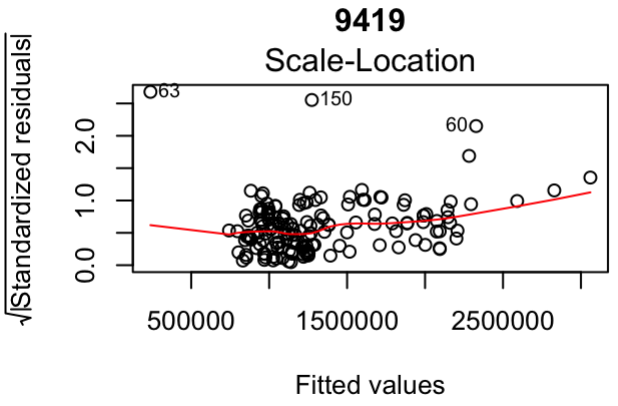
(iii). We can see that the p value of variable list and variable tax are significant, since they are less than 0.05. Their estimated regression coefficients can be interpreted as: holding all other variables constant, one unit (dollar) increase in list (price) is associated with 0.7568 units (dollars) increase in sale price; and one unit (dollar) increase in tax is associated with 14.50 units (dollars) increase in sale price.

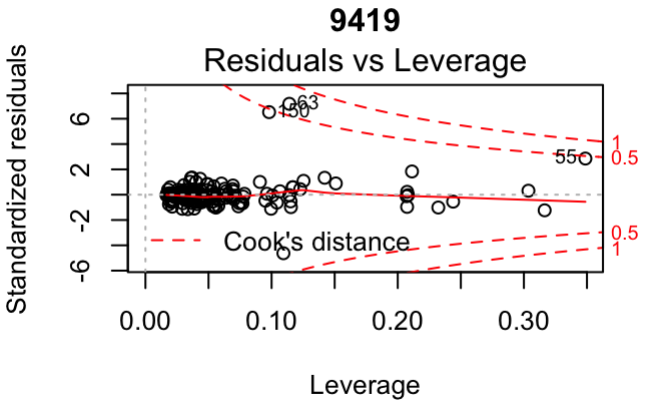
Question 4.

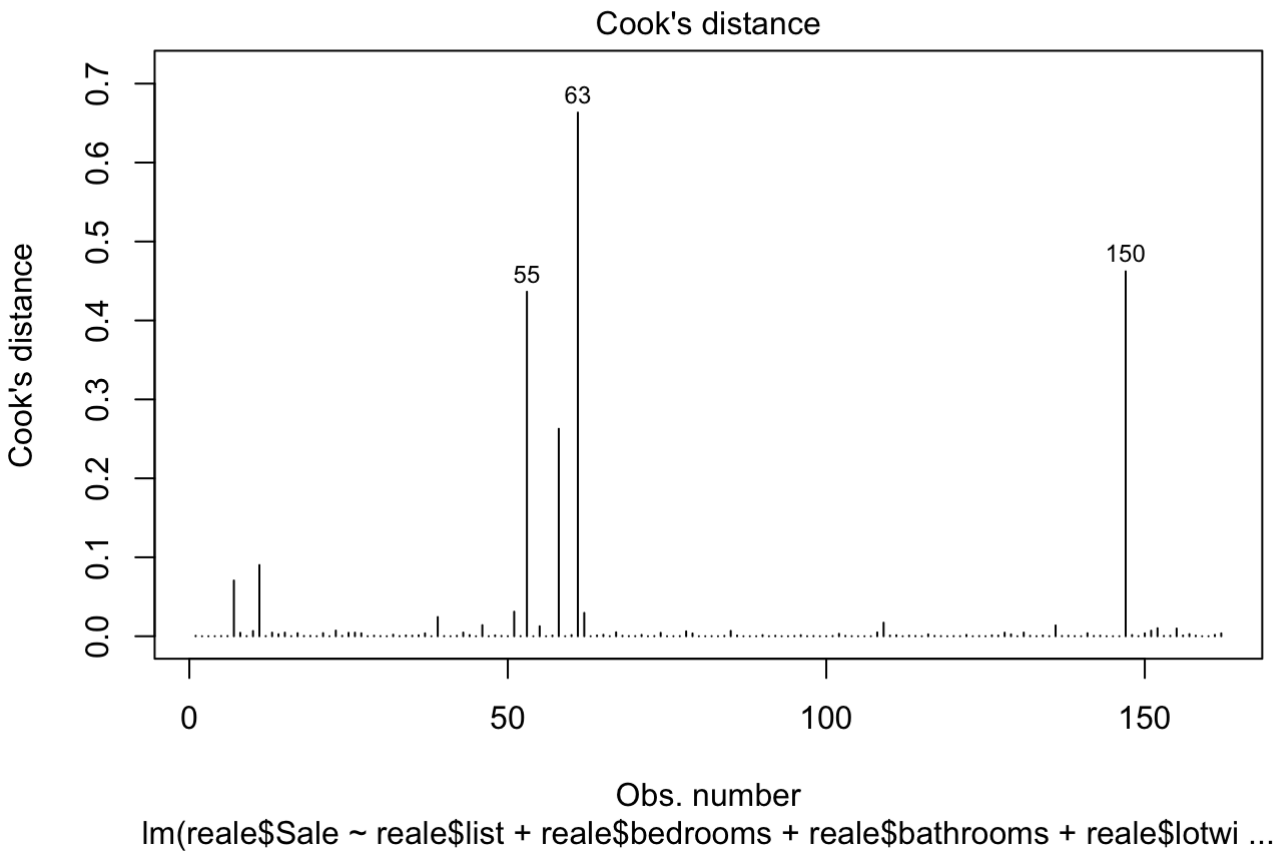




(i).







(ii). The Case ID's that may be considered as influential are: 55, 63, 150.

(iii). A point is considered as influential point if its cook distance is greater than 4/(n-k-1). In this case, the threshold value is 4/(162-9-1) = 0.026. There are 3 points with cook distance greater than 0.5 and their case_id are 53, 61, 147 respectively. Case 53 and 147 have cook distance at around 0.5 and case 61 has cook distance higher than 0.5 (around 0.65).

Question 5.

```
## Start:  AIC=3732.96
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$lotwidth + reale_manyi$lotlength +
##     reale_manyi$maxsqfoot + reale_manyi$taxes + as.factor(reale_manyi$location) +
##     reale_manyi$lotsize
##
##                                   Df  Sum of Sq        RSS    AIC
## - as.factor(reale_manyi$location)  1 4.3414e+07 1.4566e+12 3731.0
## - reale_manyi$lotlength            1 4.0049e+08 1.4569e+12 3731.0
## - reale_manyi$maxsqfoot            1 7.0777e+08 1.4573e+12 3731.0
## - reale_manyi$lotwidth             1 1.4244e+09 1.4580e+12 3731.1
## - reale_manyi$bathrooms            1 5.8302e+09 1.4624e+12 3731.6
## - reale_manyi$bedrooms             1 7.2284e+09 1.4638e+12 3731.8
## - reale_manyi$lotsize              1 9.9528e+09 1.4665e+12 3732.1
## <none>                                          1.4565e+12 3733.0
## - reale_manyi$taxes                1 9.1754e+10 1.5483e+12 3740.9
## - reale_manyi$list                 1 6.2570e+12 7.7136e+12 4001.0
##
## Step:  AIC=3730.96
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$lotwidth + reale_manyi$lotlength +
##     reale_manyi$maxsqfoot + reale_manyi$taxes + reale_manyi$lotsize
##
##                          Df  Sum of Sq        RSS    AIC
## - reale_manyi$lotlength   1 3.7519e+08 1.4570e+12 3729.0
## - reale_manyi$maxsqfoot   1 6.6588e+08 1.4573e+12 3729.0
## - reale_manyi$lotwidth    1 1.4106e+09 1.4580e+12 3729.1
## - reale_manyi$bathrooms   1 5.9660e+09 1.4626e+12 3729.6
## - reale_manyi$bedrooms    1 7.1972e+09 1.4638e+12 3729.8
## - reale_manyi$lotsize     1 9.9282e+09 1.4665e+12 3730.1
## <none>                                 1.4566e+12 3731.0
## - reale_manyi$taxes       1 9.3091e+10 1.5497e+12 3739.0
## - reale_manyi$list        1 6.2864e+12 7.7430e+12 3999.6
##
## Step:  AIC=3729
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$lotwidth + reale_manyi$maxsqfoot +
##     reale_manyi$taxes + reale_manyi$lotsize
##
##                          Df  Sum of Sq        RSS    AIC
## - reale_manyi$maxsqfoot   1 6.0072e+08 1.4576e+12 3727.1
## - reale_manyi$lotwidth    1 1.4436e+09 1.4584e+12 3727.2
## - reale_manyi$bathrooms   1 6.7364e+09 1.4637e+12 3727.8
## - reale_manyi$bedrooms    1 6.9494e+09 1.4639e+12 3727.8
## <none>                                 1.4570e+12 3729.0
## - reale_manyi$lotsize     1 4.4471e+10 1.5014e+12 3731.9
## - reale_manyi$taxes       1 9.4352e+10 1.5513e+12 3737.2
## - reale_manyi$list        1 6.2861e+12 7.7431e+12 3997.6
##
## Step:  AIC=3727.07
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$lotwidth + reale_manyi$taxes +
##     reale_manyi$lotsize
```

```
##
##                          Df  Sum of Sq        RSS    AIC
## - reale_manyi$lotwidth    1 1.3002e+09 1.4589e+12 3725.2
## - reale_manyi$bedrooms    1 8.4027e+09 1.4660e+12 3726.0
## - reale_manyi$bathrooms   1 9.3962e+09 1.4670e+12 3726.1
## <none>                                 1.4576e+12 3727.1
## - reale_manyi$lotsize     1 4.3912e+10 1.5015e+12 3729.9
## - reale_manyi$taxes       1 1.1766e+11 1.5752e+12 3737.6
## - reale_manyi$list        1 6.5242e+12 7.9817e+12 4000.5
##
## Step:  AIC=3725.22
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$taxes + reale_manyi$lotsize
##
##                          Df  Sum of Sq        RSS    AIC
## - reale_manyi$bedrooms    1 9.6765e+09 1.4685e+12 3724.3
## - reale_manyi$bathrooms   1 9.7008e+09 1.4686e+12 3724.3
## <none>                                 1.4589e+12 3725.2
## - reale_manyi$lotsize     1 5.6582e+10 1.5155e+12 3729.4
## - reale_manyi$taxes       1 1.1652e+11 1.5754e+12 3735.7
## - reale_manyi$list        1 6.5374e+12 7.9963e+12 3998.8
##
## Step:  AIC=3724.29
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bathrooms +
##     reale_manyi$taxes + reale_manyi$lotsize
##
##                          Df  Sum of Sq        RSS    AIC
## <none>                                 1.4685e+12 3724.3
## - reale_manyi$bathrooms   1 2.2847e+10 1.4914e+12 3724.8
## - reale_manyi$lotsize     1 5.6361e+10 1.5249e+12 3728.4
## - reale_manyi$taxes       1 1.1050e+11 1.5790e+12 3734.0
## - reale_manyi$list        1 6.6842e+12 8.1528e+12 4000.0
```

```
##
## Call:
## lm(formula = reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bathrooms +
##     reale_manyi$taxes + reale_manyi$lotsize)
##
## Coefficients:
##           (Intercept)       reale_manyi$list  reale_manyi$bathrooms
##             7.116e+04              7.609e-01              1.591e+04
##     reale_manyi$taxes    reale_manyi$lotsize
##             1.434e+01              5.750e+00
```

(i). By using coefficients obtained above, the final fitted model is

$$\hat{Sale} = 71160 + list * 0.7609 + bathrooms * 15910 + taxes * 14.34 + lotsize * 5.750.$$

(ii). No, the results are inconsistent with what we derived in question 3. The final model eliminates some explanatory variables comparing to the full model in question 3, for example: bedrooms, lotwidth, lotlength etc. We are using backward AIC to select model and the model with smaller AIC is more prefered than larger AIC. Therefore, when we are removing factors from the full model, we will keep eliminating if AIC keeps decreasing.

Question 6.

```
## Start:  AIC=3763.83
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$lotwidth + reale_manyi$lotlength +
##     reale_manyi$maxsqfoot + reale_manyi$taxes + as.factor(reale_manyi$location) +
##     reale_manyi$lotsize
##
##                                   Df  Sum of Sq        RSS    AIC
## - as.factor(reale_manyi$location)  1 4.3414e+07 1.4566e+12 3758.8
## - reale_manyi$lotlength            1 4.0049e+08 1.4569e+12 3758.8
## - reale_manyi$maxsqfoot            1 7.0777e+08 1.4573e+12 3758.8
## - reale_manyi$lotwidth             1 1.4244e+09 1.4580e+12 3758.9
## - reale_manyi$bathrooms            1 5.8302e+09 1.4624e+12 3759.4
## - reale_manyi$bedrooms             1 7.2284e+09 1.4638e+12 3759.5
## - reale_manyi$lotsize              1 9.9528e+09 1.4665e+12 3759.8
## <none>                                          1.4565e+12 3763.8
## - reale_manyi$taxes                1 9.1754e+10 1.5483e+12 3768.6
## - reale_manyi$list                 1 6.2570e+12 7.7136e+12 4028.8
##
## Step:  AIC=3758.75
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$lotwidth + reale_manyi$lotlength +
##     reale_manyi$maxsqfoot + reale_manyi$taxes + reale_manyi$lotsize
##
##                          Df  Sum of Sq        RSS    AIC
## - reale_manyi$lotlength   1 3.7519e+08 1.4570e+12 3753.7
## - reale_manyi$maxsqfoot   1 6.6588e+08 1.4573e+12 3753.7
## - reale_manyi$lotwidth    1 1.4106e+09 1.4580e+12 3753.8
## - reale_manyi$bathrooms   1 5.9660e+09 1.4626e+12 3754.3
## - reale_manyi$bedrooms    1 7.1972e+09 1.4638e+12 3754.5
## - reale_manyi$lotsize     1 9.9282e+09 1.4665e+12 3754.8
## <none>                                 1.4566e+12 3758.8
## - reale_manyi$taxes       1 9.3091e+10 1.5497e+12 3763.7
## - reale_manyi$list        1 6.2864e+12 7.7430e+12 4024.3
##
## Step:  AIC=3753.71
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$lotwidth + reale_manyi$maxsqfoot +
##     reale_manyi$taxes + reale_manyi$lotsize
##
##                          Df  Sum of Sq        RSS    AIC
## - reale_manyi$maxsqfoot   1 6.0072e+08 1.4576e+12 3748.7
## - reale_manyi$lotwidth    1 1.4436e+09 1.4584e+12 3748.8
## - reale_manyi$bathrooms   1 6.7364e+09 1.4637e+12 3749.4
## - reale_manyi$bedrooms    1 6.9494e+09 1.4639e+12 3749.4
## - reale_manyi$lotsize     1 4.4471e+10 1.5014e+12 3753.5
## <none>                                 1.4570e+12 3753.7
## - reale_manyi$taxes       1 9.4352e+10 1.5513e+12 3758.8
## - reale_manyi$list        1 6.2861e+12 7.7431e+12 4019.2
##
## Step:  AIC=3748.68
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$lotwidth + reale_manyi$taxes +
##     reale_manyi$lotsize
```

```
##
##                          Df  Sum of Sq        RSS    AIC
## - reale_manyi$lotwidth    1 1.3002e+09 1.4589e+12 3743.7
## - reale_manyi$bedrooms    1 8.4027e+09 1.4660e+12 3744.5
## - reale_manyi$bathrooms   1 9.3962e+09 1.4670e+12 3744.6
## - reale_manyi$lotsize     1 4.3912e+10 1.5015e+12 3748.4
## <none>                                 1.4576e+12 3748.7
## - reale_manyi$taxes       1 1.1766e+11 1.5752e+12 3756.2
## - reale_manyi$list        1 6.5242e+12 7.9817e+12 4019.1
##
## Step:  AIC=3743.74
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms +
##     reale_manyi$bathrooms + reale_manyi$taxes + reale_manyi$lotsize
##
##                          Df  Sum of Sq        RSS    AIC
## - reale_manyi$bedrooms    1 9.6765e+09 1.4685e+12 3739.7
## - reale_manyi$bathrooms   1 9.7008e+09 1.4686e+12 3739.7
## <none>                                 1.4589e+12 3743.7
## - reale_manyi$lotsize     1 5.6582e+10 1.5155e+12 3744.8
## - reale_manyi$taxes       1 1.1652e+11 1.5754e+12 3751.1
## - reale_manyi$list        1 6.5374e+12 7.9963e+12 4014.3
##
## Step:  AIC=3739.72
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bathrooms +
##     reale_manyi$taxes + reale_manyi$lotsize
##
##                          Df  Sum of Sq        RSS    AIC
## - reale_manyi$bathrooms   1 2.2847e+10 1.4914e+12 3737.1
## <none>                                 1.4685e+12 3739.7
## - reale_manyi$lotsize     1 5.6361e+10 1.5249e+12 3740.7
## - reale_manyi$taxes       1 1.1050e+11 1.5790e+12 3746.4
## - reale_manyi$list        1 6.6842e+12 8.1528e+12 4012.3
##
## Step:  AIC=3737.14
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$taxes + reale_manyi$lotsize
##
##                         Df  Sum of Sq        RSS    AIC
## - reale_manyi$lotsize    1 3.8225e+10 1.5296e+12 3736.2
## <none>                                1.4914e+12 3737.1
## - reale_manyi$taxes      1 1.3054e+11 1.6219e+12 3745.6
## - reale_manyi$list       1 9.4650e+12 1.0956e+13 4055.1
##
## Step:  AIC=3736.15
## reale_manyi$Sale ~ reale_manyi$list + reale_manyi$taxes
##
##                       Df  Sum of Sq        RSS    AIC
## <none>                               1.5296e+12 3736.2
## - reale_manyi$taxes    1 1.4758e+11 1.6772e+12 3746.0
## - reale_manyi$list     1 1.0430e+13 1.1960e+13 4064.2
```

```
##
## Call:
## lm(formula = reale_manyi$Sale ~ reale_manyi$list + reale_manyi$taxes)
##
## Coefficients:
##       (Intercept)   reale_manyi$list  reale_manyi$taxes
##         1.167e+05          7.957e-01          1.626e+01
```

(i). Similarly, by using coefficients obtained above, the final model is
$\hat{Sale} = 116700 + list * 0.7957 + taxes * 16.26.$

(ii). In this case, the results are inconsistent with question 3 and question 5. In the removing process, BIC is more strict (heavy penalty) in choosing variables comparing to AIC, which makes the results different. Therefore, question 6 produces a final fitted model with less variables, since more predictors are dropped during the process.

Appendix

```
Q1
reale <- read.csv("/Users/meow/Desktop/STA302/A3/reale_a3data.csv")
reale_manyi = subset(reale,!is.na(reale$lotwidth) & !is.na(reale$lotlength) & !is.na(rea
le$taxes))
reale_manyi$lotsize = reale_manyi$lotwidth * reale_manyi$lotlength
str(reale_manyi)

reale_manyi_quant = reale_manyi[,c(2:9, 11)]
str(reale_manyi_quant)

manyi.cor <- function(x, y, digits = 4, prefix = "", cex.cor, ...){
  usr <- par("usr");
  on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  txt1 <- format(cor(x,y), digits = digits)
  text(0.5, 0.5, paste(txt1), cex = 0.6)
  }
pairs(~ Sale + list + bedrooms + bathrooms + lotwidth + lotlength + maxsqfoot + taxes +
lotsize, data = reale_manyi_quant, lower.panel = manyi.cor, cex = 0.6, pch = 20, bg = "y
ellow", cex.labels = 0.7, font.labels = 0.7, upper.panel = panel.smooth)

Q2
model1_manyi <- lm(reale_manyi$Sale ~ reale_manyi$lotlength)
summary(model1_manyi)
plot(model1_manyi, 1)

Q3
model2_manyi <- lm(reale_manyi$Sale ~ reale_manyi$list + reale_manyi$bedrooms + reale_ma
nyi$bathrooms + reale_manyi$lotwidth + reale_manyi$lotlength + reale_manyi$maxsqfoot + r
eale_manyi$taxes + as.factor(reale_manyi$location) + reale_manyi$lotsize)
summary(model2_manyi)

Q4
reale$lotsize = reale$lotwidth * reale$lotlength
model3_manyi <- lm(reale$Sale ~ reale$list + reale$bedrooms + reale$bathrooms + reale$lo
twidth + reale$lotlength + reale$maxsqfoot + reale$taxes + as.factor(reale$location) + r
eale$lotsize)
par(mfrow=c(2,2))
plot(model3_manyi)
plot(model3_manyi, 4)

Q5
step(model2_manyi,direction = "backward")

Q6
step(model2_manyi, direction = "backward", k=log(162))
```