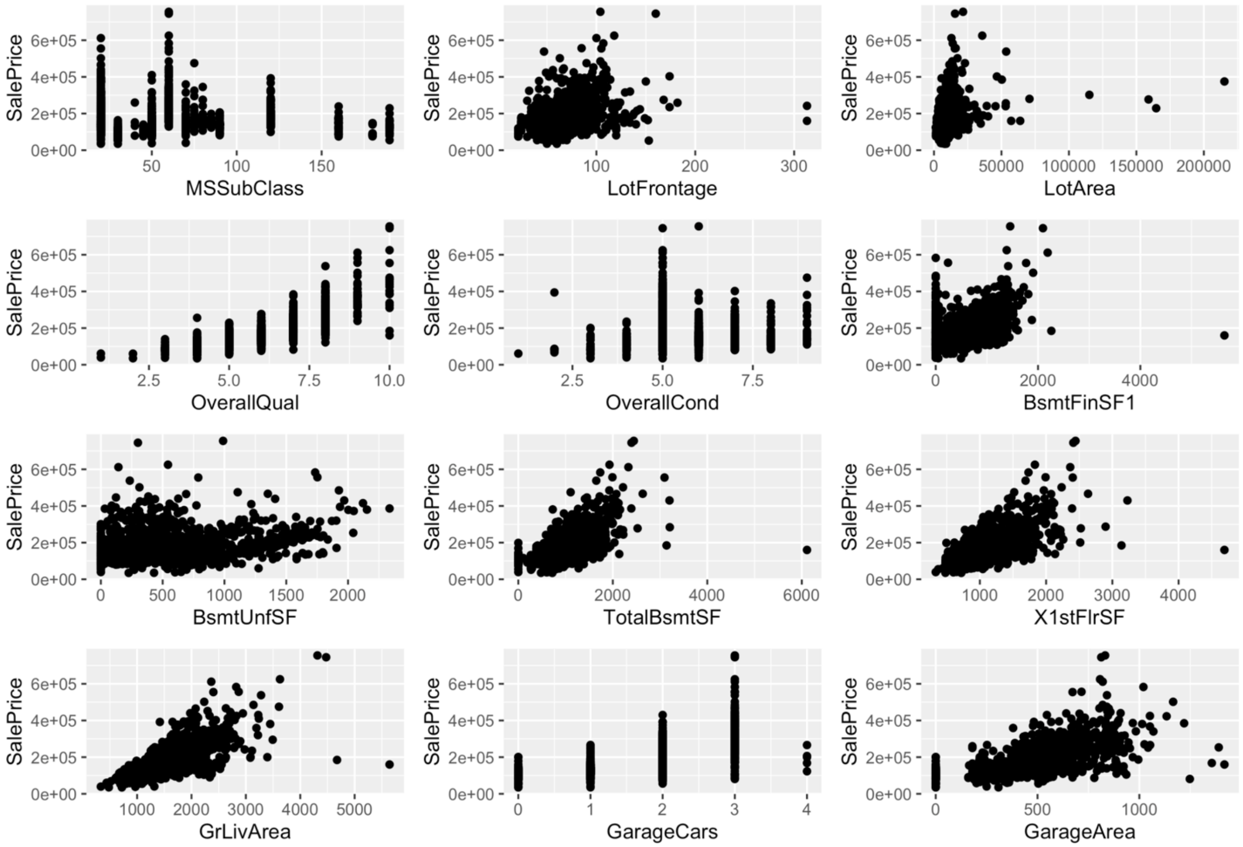
SSC 442 / Lab 1

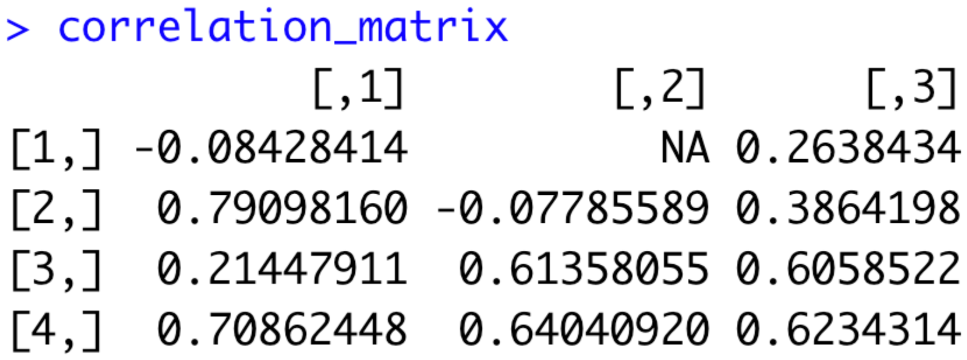
Shun Hu, Yuhan Liu, Yanfei Du, Hongji Li

Exercise 1

2.



3.



```
> correlation_matrix
            [,1]        [,2]      [,3]
[1,] -0.08428414          NA 0.2638434
[2,]  0.79098160 -0.07785589 0.3864198
[3,]  0.21447911  0.61358055 0.6058522
[4,]  0.70862448  0.64040920 0.6234314
```

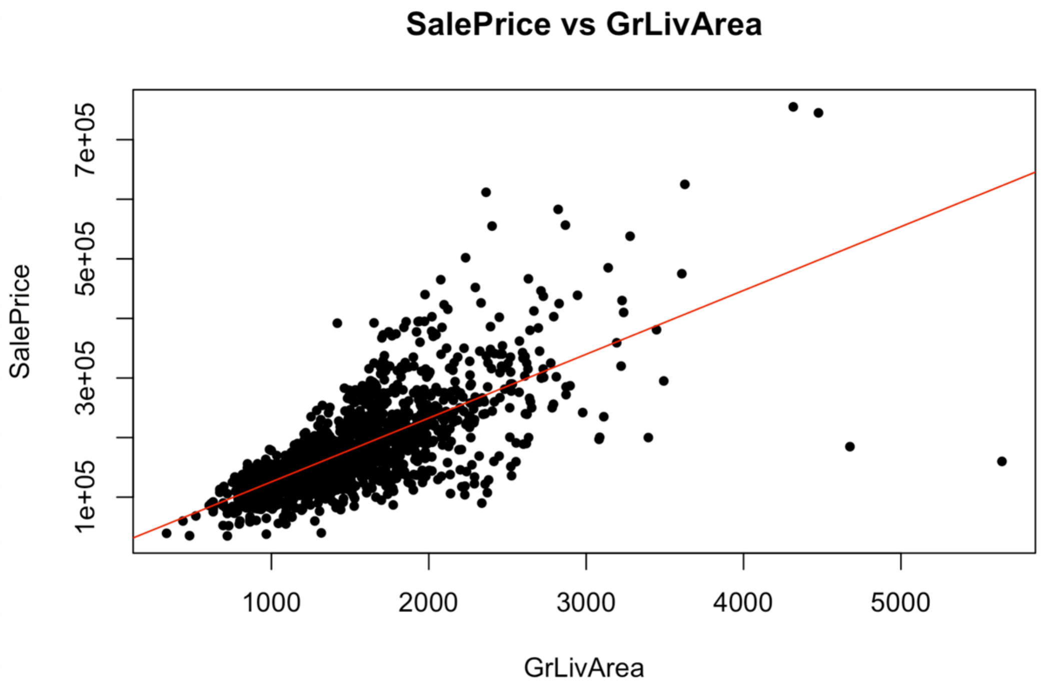
Yes, this match our prior beliefs.

MSSubClass and OverallCond have a negative correlation with SalePrice.

Correlation between LotFrontage and SalePrice is NA.

All the other variables have a positive correaltion with SalePrice.

4.

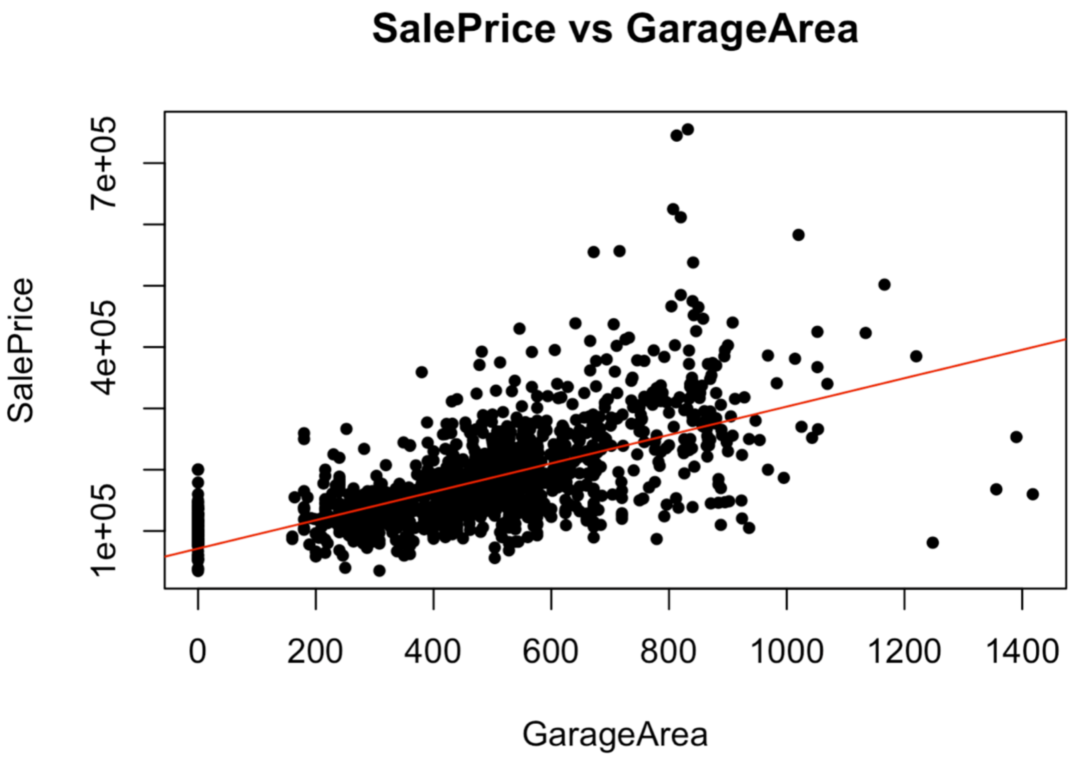


The largest outlier that is above the regression line is (4316,755000)

An increase in overall living area of 1 ft is correlated with an expected increase in sales price of $107.

Ecercise 2

1.



An increase in garage area of 1 ft is correlated with an expected increase in sales price of $ 232.

2.

Is there a relationship between the predictors and the response?

Yes, there is a relationship between the predictors and the response.

Which predictors appear to have a statistically significant relationship to the response?

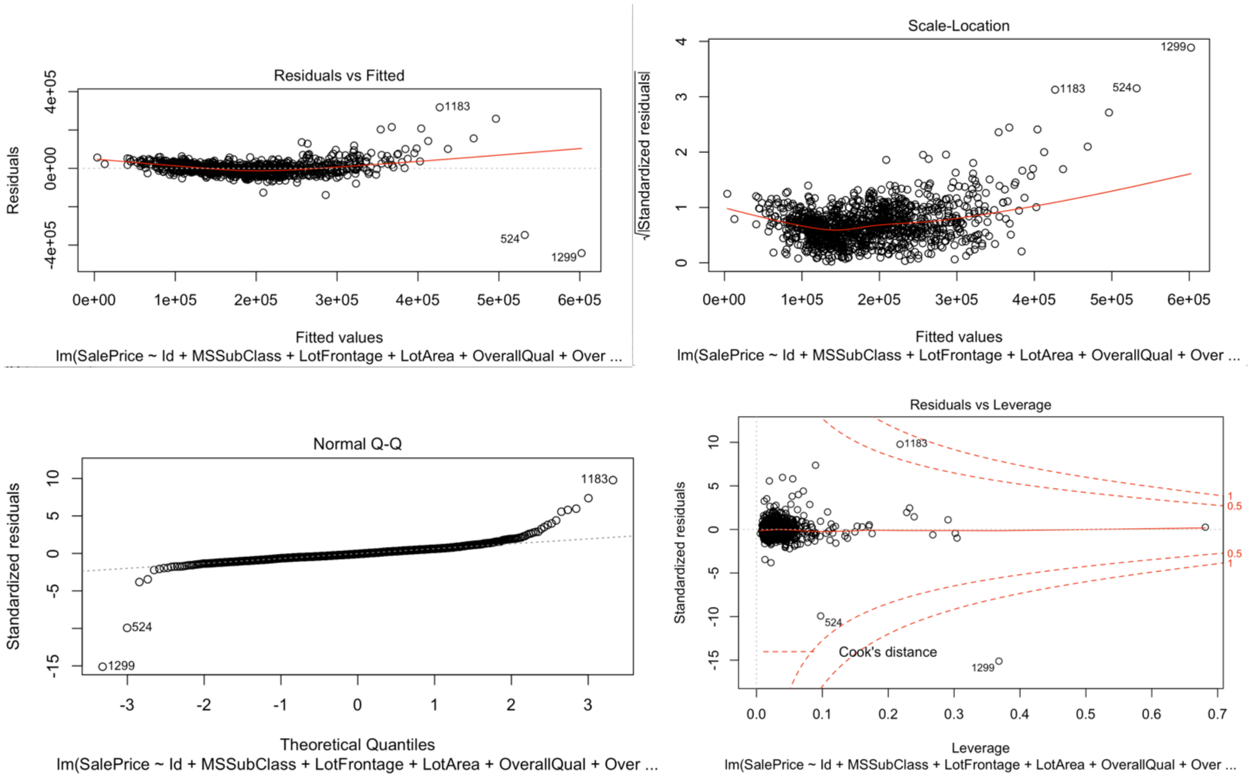
MSSubClass, LotArea, OverallQual, OverallCond, YearBuilt, MasVnrArea, BsmtFinSF1, X1stFlrSF, X2ndFlrSF, BsmtFullBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, WoodDeckSF, ScreenPorch, PoolArea have a statistically significant relationship to the response, SalePrice.

What does the coefficient for the year variable suggest?

Only YearBuilt has a statistically significant relationship to SalePrice.

An increase in YearBuilt of 1 is correlated with an expected increase in sales price of 3.164e+02

3.



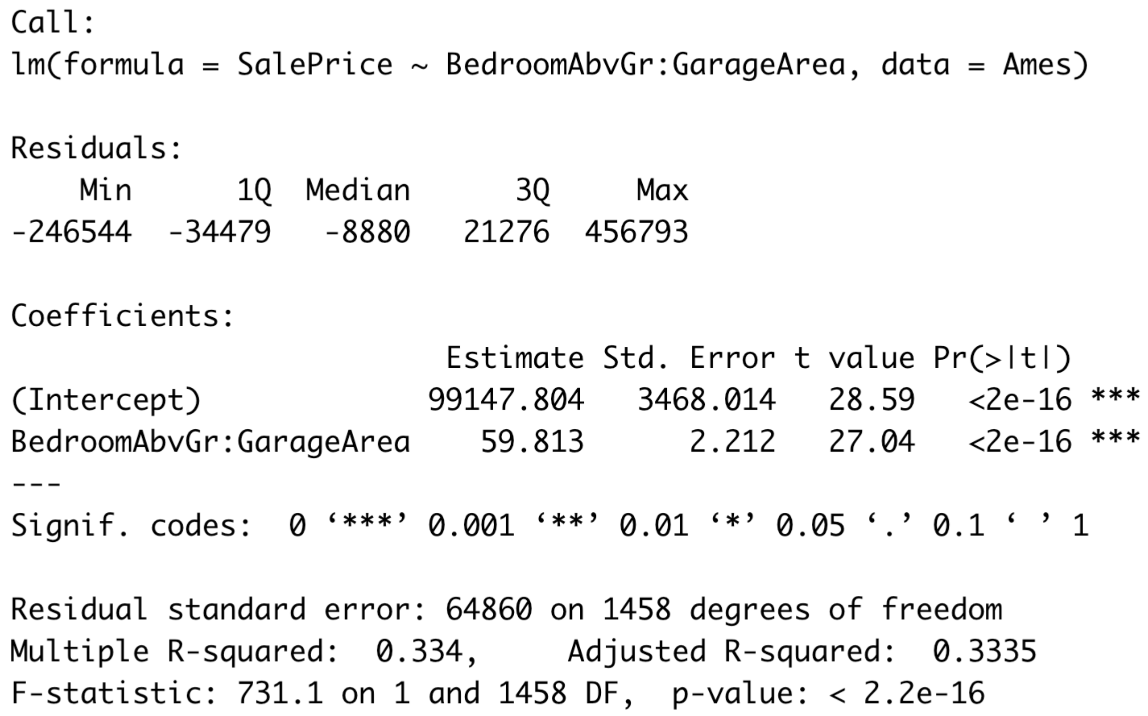
Do the residual plots suggest any unusually large outliers?

Yes, there is some unusually large outliers.

Does the leverage plot identify any observations with unusually high leverage?

Yes, there is some points with unusually high leverage.

4.

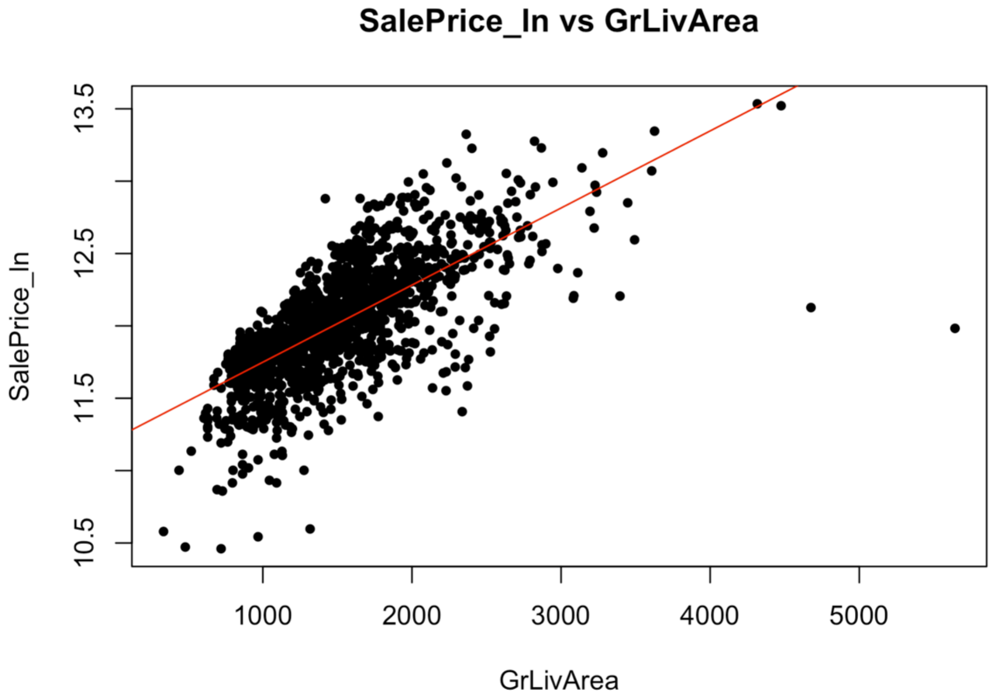


```
Call:
lm(formula = SalePrice ~ BedroomAbvGr:GarageArea, data = Ames)

Residuals:
    Min      1Q  Median      3Q     Max 
-246544  -34479   -8880   21276  456793 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             99147.804   3468.014   28.59   <2e-16 ***
BedroomAbvGr:GarageArea    59.813      2.212   27.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64860 on 1458 degrees of freedom
Multiple R-squared:  0.334,	Adjusted R-squared:  0.3335 
F-statistic: 731.1 on 1 and 1458 DF,  p-value: < 2.2e-16
```

Do any interactions appear to be statistically significant?

This interaction between “SalePrice” and “BedroomAbvGr:GarageArea” seems to appear statistically significant, because P-value is very small.

5.



```
Call:
lm(formula = SalePrice_ln ~ GrLivArea, data = Ames)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.23982 -0.14271  0.03034  0.16317  0.90636 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.122e+01  2.277e-02  492.51   <2e-16 ***
GrLivArea   5.328e-04  1.420e-05   37.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.285 on 1458 degrees of freedom
Multiple R-squared:  0.4913,	Adjusted R-squared:  0.4909 
F-statistic:  1408 on 1 and 1458 DF,  p-value: < 2.2e-16
```

We tried ln(SalePrice) and GrLivArea.

Since P-value is very small, it is statistically significant.

ln(SalePrice) and GrLivArea have a positive correlation