Lab 2
Alexandra Holland, Maura Glynn, Tyler Kramer, Jonathan Zaremba

Exercise 1

1. (also saved in Github as a .txt file)

The Ames data set contains 15 variables: ID, LotArea, OverallQual, OverallCond, YearBuilt, X1stFlrSF, X2ndFlrSF, LowQualFinSF, BsmtFullBath, BsmtHalfBath, GrLivArea, FullBath, BedroomAbvGr, BsmtFullBath, BsmtHalfBath, GarageCars, and SalePrice. I will go through and give a description of each one:
ID: This is just to identify the specific property that the other columns are referring to.
LotArea: This is a measure of the size of a given property given in square feet. This value should always be positive.
OverallQual: This is a measure of the quality of a given property on a scale of 1-10 with 10 being the best and 1 being the worst.
OverallCond: This is a measure of the condition the property is in on a scale of 1-10 with 10 being brand new and 1 being unlivable.
YearBuilt: This is the year in which the house was built.
X1stFlrSF: This is a measure of the square footage of the first floor of the house. Should always be positive.
X2ndFlrSF: This is a measure of the square footage of the second floor of the house. Cannot be negative.
LowQualFinSF: This might be a measure of the amount of low quality financing owners required to buy a given house. Never negative.
GrLivArea: This appears to be a measure of the total square footage of the house. Always positive.
FullBath: This is a measure of how many full bathrooms are in a house. No negatives.
BedroomAbvGr: This is how many bedrooms are above the height of the garage. Cannot be negative.
BsmtFullBath: This is a measure of how many full bathrooms are in the basement of a given property. Cannot be negative.
BsmtHalfBath: This is a measure of how many half-bathrooms are in the basement of a given property. Cannot be negative.
GarageCars: This is how many cars a garage can hold. No negatives.
SalePrice: This is a measure of how much a property was sold for in US Dollars. All values are positive.

2. Scatterplot matrix which includes 12 of the variables that are type=int in the dataset:
    a. ID, LotArea, OverallQual, OverallCond, YearBuilt, X1stFlrSF, X2ndFlrSF, LowQualFinSF, GrLivArea, FullBath, BedroomAbvGr, BsmtFullBath, BsmtHalfBath, GarageCars, SalePrice

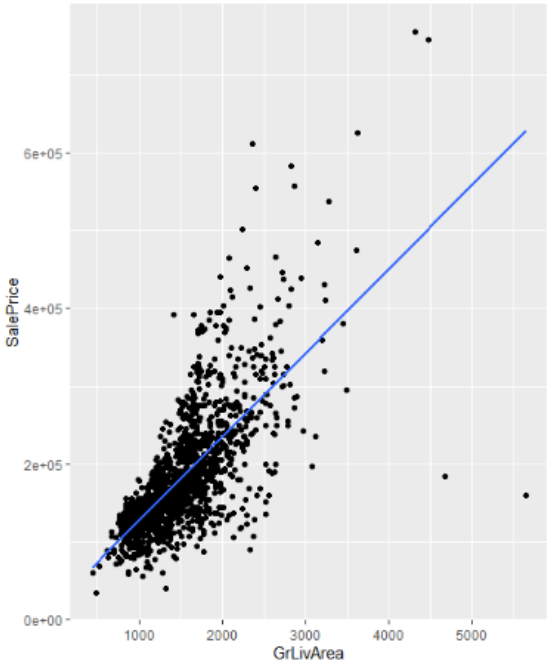
3. The correlations between the variables had a somewhat surprising result relative to our initial beliefs. For example, we expected “Overall Condition” to have a significant positive correlation with sales price, but instead we see that there is actually a slight negative correlation. Also, it’s surprising that the year a house was built has such a high positive correlation with the sales price, but this can also be believable. While old houses may be expensive due to their perceived *antique* value, it makes sense the newer houses may command a higher price. Most of the other correlations intuitively make sense, such as “Overall Quality,” which has a high positive correlation with sales price.

	LotArea	OverallQual	OverallCond	YearBuilt	X1stFlrSF
LotArea	1.000000000	0.09001631	-0.002869219	-0.005920805	0.29147788
OverallQual	0.090016311	1.00000000	-0.136232205	0.572082457	0.46742531
OverallCond	-0.002869219	-0.13623220	1.000000000	-0.403601675	-0.14717147
YearBuilt	-0.005920805	0.57208246	-0.403601675	1.000000000	0.25816141
X1stFlrSF	0.291477878	0.46742531	-0.147171474	0.258161406	1.00000000
X2ndFlrSF	0.046253072	0.28610907	0.016720474	0.017151632	-0.22371034
LowQualFinSF	0.010177287	-0.01454954	0.043216146	-0.159332561	-0.01885367
GrLivArea	0.257243272	0.58958384	-0.092217302	0.194662669	0.55462010
FullBath	0.117064147	0.55623075	-0.218473655	0.482739335	0.36605135
BedroomAbvGr	0.119746500	0.08171377	0.014984665	-0.071794474	0.10486968
GarageCars	0.137977589	0.58138144	-0.247317060	0.523349483	0.43973981
SalePrice	0.252921459	0.78722783	-0.095277741	0.507584064	0.59493527

	X2ndFlrSF	LowQualFinSF	GrLivArea	FullBath	BedroomAbvGr
LotArea	0.04625307	0.010177287	0.2572433	0.11706415	0.11974650
OverallQual	0.28610907	-0.014549544	0.5895838	0.55623075	0.08171377
OverallCond	0.01672047	0.043216146	-0.0922173	-0.21847366	0.01498466
YearBuilt	0.01715163	-0.159332561	0.1946627	0.48273934	-0.07179447
X1stFlrSF	-0.22371034	-0.018853672	0.5546201	0.36605135	0.10486968
X2ndFlrSF	1.00000000	0.045546751	0.6834407	0.41704852	0.50757441
LowQualFinSF	0.04554675	1.000000000	0.1018108	-0.02329168	0.05656100
GrLivArea	0.68344067	0.101810802	1.0000000	0.62420025	0.51231197
FullBath	0.41704852	-0.023291676	0.6242003	1.00000000	0.35791106
BedroomAbvGr	0.50757441	0.056561005	0.5123120	0.35791106	1.00000000
GarageCars	0.18590305	-0.017162953	0.4838987	0.51124364	0.11752232
SalePrice	0.31333559	-0.008364395	0.7081721	0.55655030	0.16465495

	GarageCars	SalePrice
LotArea	0.13797759	0.252921459
OverallQual	0.58138144	0.787227826
OverallCond	-0.24731706	-0.095277741
YearBuilt	0.52334948	0.507584064
X1stFlrSF	0.43973981	0.594935270
X2ndFlrSF	0.18590305	0.313335590
LowQualFinSF	-0.01716295	-0.008364395
GrLivArea	0.48389867	0.708172114
FullBath	0.51124364	0.556550302
BedroomAbvGr	0.11752232	0.164654949
GarageCars	1.00000000	0.637095406
SalePrice	0.63709541	1.000000000

4. Produce a scatterplot between SalePrice and GrLivArea using abline():



   a. The largest outlier above this regression line is the house with ID: 1299
   b. The following are features of the house:
      i. Lot Area: 63,887 sqft, Overall Quality: 10, Overall Condition: 5, Year Built: 2008, 1st Floor sq ft: 4,692, 2nd Floor sq ft: 950, Ground Living Area: 5,642, Full Baths: 2, Bedrooms Above Ground: 3, Basement Full Baths: 2, Basement Half Baths: 0, Garage Cars: 2, Sales Price: 160,000

Exercise 2

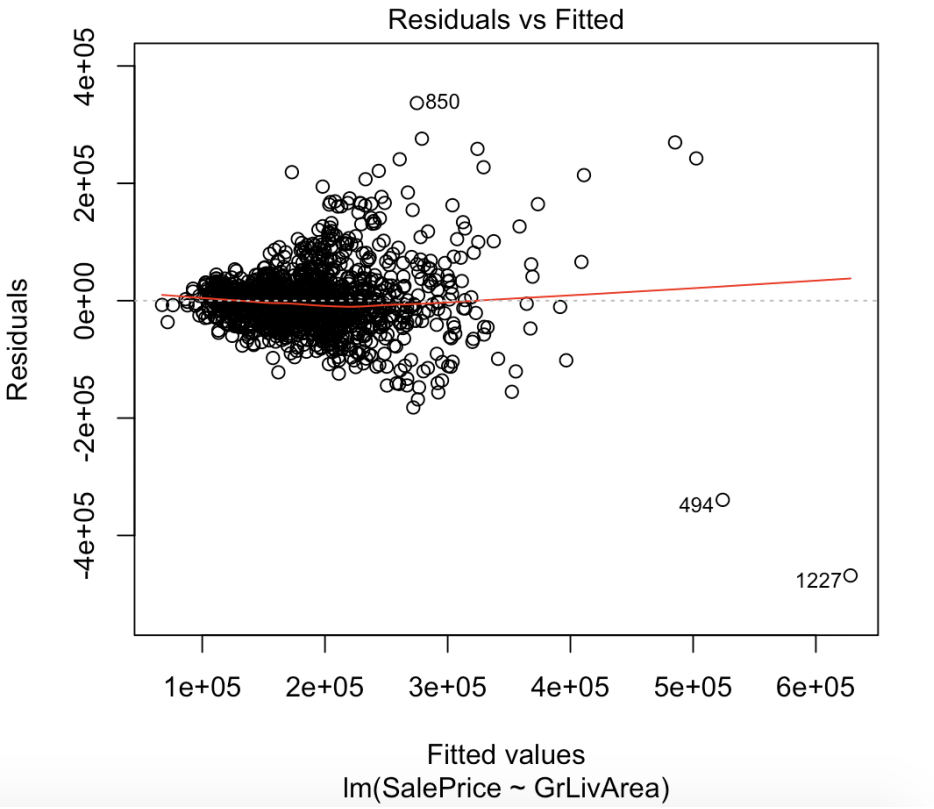
TRY ITS

- GrLivArea is a variable that holds the size of the living areas for the houses

- Regression on SalePrice of GrLivArea:

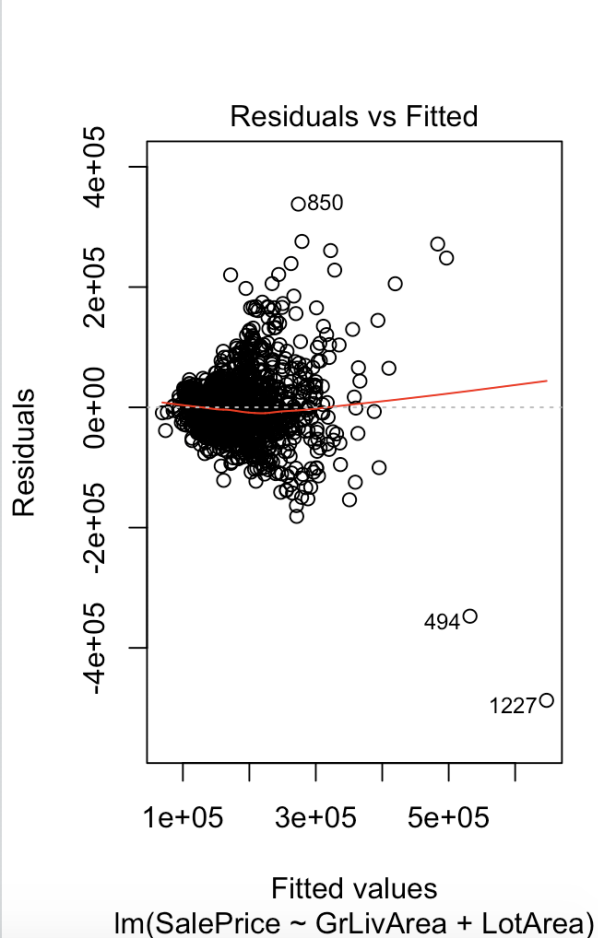
```
lm(formula = SalePrice ~ GrLivArea)

Residuals:
    Min      1Q  Median      3Q     Max
-468241  -28629   -2139   20802  336779

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20042.428   4692.086   4.272 2.07e-05 ***
GrLivArea     107.798      2.896  37.220  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55810 on 1377 degrees of freedom
Multiple R-squared:  0.5015,    Adjusted R-squared:  0.5011
F-statistic:  1385 on 1 and 1377 DF,  p-value: < 2.2e-16
```

- Plot of lm.fit:



- Regression on SalePrice of GrLivArea controlling for LotArea:

```
lm(formula = SalePrice ~ GrLivArea + LotArea)

Residuals:
    Min      1Q  Median      3Q     Max
-487232  -29020   -2436   20561  337936

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.833e+04  4.690e+03   3.908 9.76e-05 ***
GrLivArea   1.048e+02  2.982e+00  35.154  < 2e-16 ***
LotArea     5.861e-01  1.516e-01   3.867 0.000115 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55530 on 1376 degrees of freedom
Multiple R-squared:  0.5069,    Adjusted R-squared:  0.5062
F-statistic: 707.2 on 2 and 1376 DF,  p-value: < 2.2e-16
```



Controlling for Lot Area does not change the qualitative conclusions made by the previous regression but does change the quantitative results as the F statistic is lowered significantly but maintains the same p value.

EXERCISE 2

1.

```
lm(formula = SalePrice ~ GarageOutside, data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max
-150409  -44237  -13043   25098  548598

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    206402       2291   90.08   <2e-16 ***
GarageOutside  -72859       4276  -17.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71840 on 1377 degrees of freedom
Multiple R-squared:  0.1741,    Adjusted R-squared:  0.1735
F-statistic: 290.3 on 1 and 1377 DF,  p-value: < 2.2e-16
```

   a. Having an outdoor garage is estimated to have a lower Sale Price by $72,859 than a house with an indoor garage.

2.

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

Residuals:
    Min      1Q  Median      3Q     Max
-472209  -18377   -1800   16191  278363

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.612e+05  1.022e+05  -9.402  < 2e-16 ***
ID           -1.836e+00  2.365e+00  -0.777 0.437583
LotArea       5.540e-01  1.041e-01   5.321 1.21e-07 ***
OverallQual   2.128e+04  1.203e+03  17.698  < 2e-16 ***
OverallCond   7.532e+03  1.030e+03   7.314 4.39e-13 ***
YearBuilt     4.258e+02  5.253e+01   8.105 1.16e-15 ***
X1stFlrSF     7.760e+01  4.021e+00  19.297  < 2e-16 ***
X2ndFlrSF     5.638e+01  3.563e+00  15.825  < 2e-16 ***
LowQualFinSF  3.356e+01  2.520e+01   1.332 0.183169
GrLivArea            NA         NA      NA       NA
FullBath     -2.743e+03  2.775e+03  -0.989 0.323082
BedroomAbvGr -6.082e+03  1.622e+03  -3.750 0.000185 ***
BsmtFullBath  1.295e+04  2.152e+03   6.020 2.24e-09 ***
BsmtHalfBath  4.828e+03  4.282e+03   1.127 0.259746
GarageCars    1.757e+04  2.153e+03   8.159 7.62e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36740 on 1365 degrees of freedom
Multiple R-squared:  0.7859,    Adjusted R-squared:  0.7838
F-statistic: 385.3 on 13 and 1365 DF,  p-value: < 2.2e-16
```

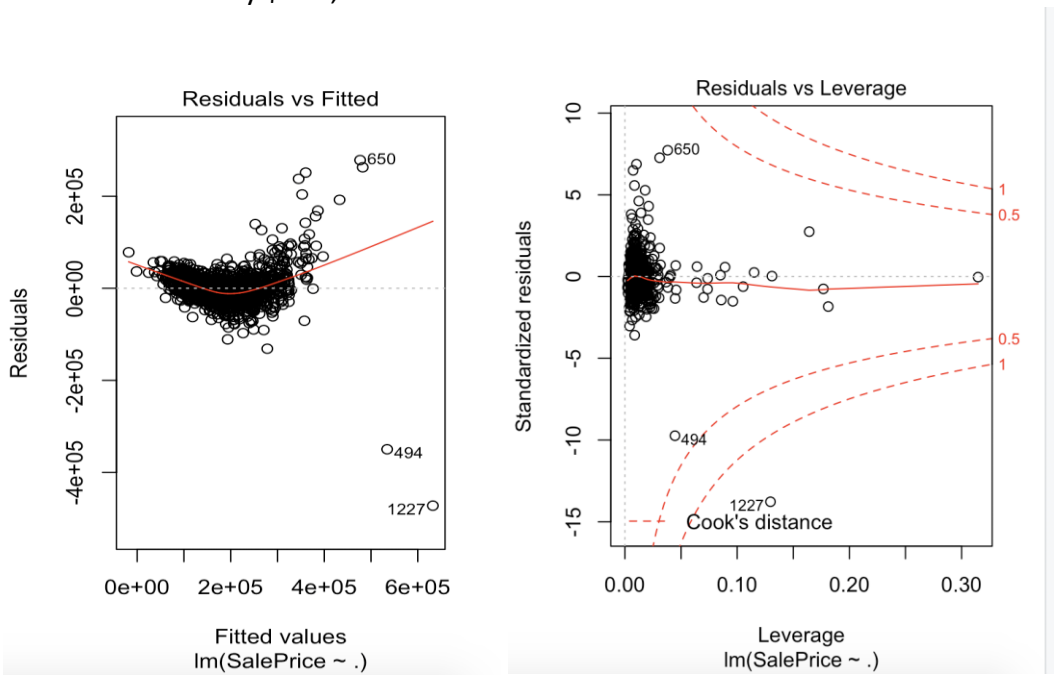
   a. By looking at the F statistic and the Multiple/adjusted R^2, there does appear to be a relationship between the predictors and the response. Adding all of the

variables in the data set, it can be seen that not all of the predictors are statistically significant but a good amount of them are indicating a relationship with SalePrice.

b. The predictors that have statistically significant relationships with SalePrice are LotArea, OverallQual, OverallCond, YearBuilt, X1stFlrSF, X2ndFlrSF, BedroomAbvGr, BsmtFullBath, and GarageCars.

c. The coefficient for the year variable suggests that for every one year newer that the house is (one year increase from when it was built), the price fo the house increases by $425,800.

3.



a. The fit seems to more or less be pretty good especially since we know that the R^2 is around 0.7. That being said, there do appear to be some large outliers towards the upper right hand corner of the graph, indicating higher residuals and fitted values than the majority than the rest of the data points. There are also outliers at the bottom right hand corner which display high fitted values but low residuals. The points centered around 2e+05 fitted value and a residual below zero seem to have larger leverage in pulling the fitted line below zero to the negative residual half. At larger fitted values, the points start to trend upward with residuals indicating high leverage in pulling up the fit line.

```
Residuals:
    Min      1Q  Median      3Q     Max
-473036  -18241   -1797   15974  279801

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.270e+06  4.169e+05  -3.046 0.002364 **
ID                   -1.826e+00  2.365e+00  -0.772 0.440243
LotArea               5.558e-01  1.042e-01   5.336 1.11e-07 ***
OverallQual           2.104e+04  1.243e+03  16.923  < 2e-16 ***
OverallCond           6.115e+04  7.023e+04   0.871 0.384076
YearBuilt             5.843e+02  2.141e+02   2.728 0.006444 **
X1stFlrSF             7.784e+01  4.034e+00  19.294  < 2e-16 ***
X2ndFlrSF             5.635e+01  3.563e+00  15.813  < 2e-16 ***
LowQualFinSF          3.346e+01  2.520e+01   1.328 0.184559
GrLivArea                    NA         NA      NA       NA
FullBath             -2.892e+03  2.782e+03  -1.040 0.298707
BedroomAbvGr         -5.965e+03  1.630e+03  -3.661 0.000261 ***
BsmtFullBath          1.306e+04  2.157e+03   6.057 1.79e-09 ***
BsmtHalfBath          5.084e+03  4.296e+03   1.183 0.236841
GarageCars            1.740e+04  2.165e+03   8.034 2.02e-15 ***
OverallCond:YearBuilt -2.746e+01  3.597e+01  -0.764 0.445286
```

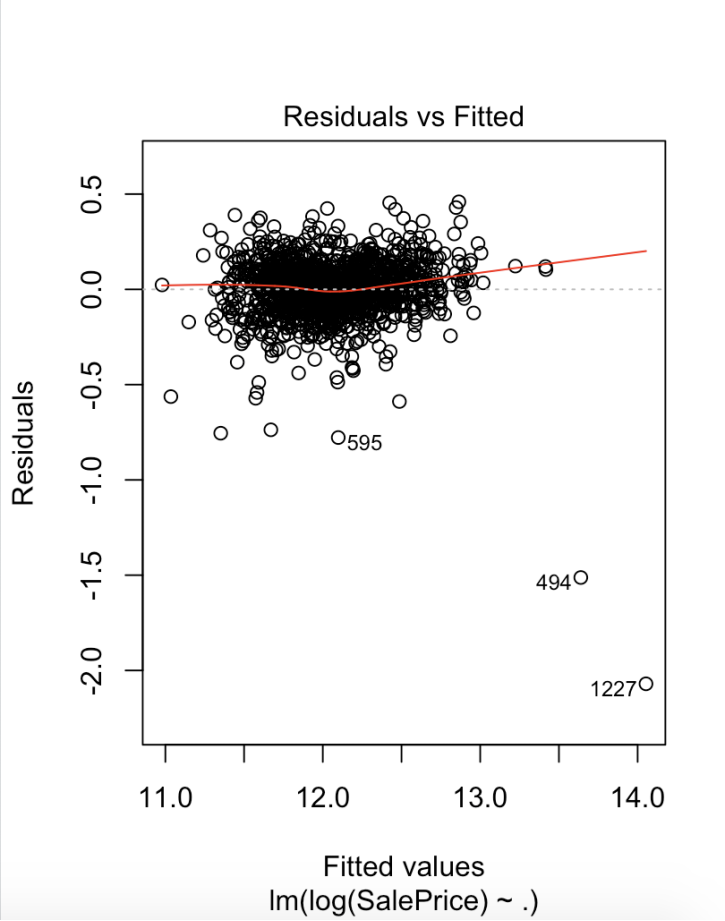
4.

```
Residuals:
    Min      1Q  Median      3Q     Max
-474142  -17873   -2011   15901  282370

Coefficients: (1 not defined because of singularities)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.225e+06  4.175e+05  -2.935 0.003392
ID                        -1.775e+00  2.364e+00  -0.751 0.452789
LotArea                    5.586e-01  1.041e-01   5.365 9.49e-08
OverallQual                2.770e+04  4.242e+03   6.528 9.37e-11
OverallCond                6.093e+04  7.019e+04   0.868 0.385465
YearBuilt                  5.420e+02  2.156e+02   2.514 0.012048
X1stFlrSF                  7.769e+01  4.033e+00  19.264  < 2e-16
X2ndFlrSF                  5.661e+01  3.565e+00  15.881  < 2e-16
LowQualFinSF               3.403e+01  2.519e+01   1.351 0.176950
GrLivArea                         NA         NA      NA       NA
FullBath                  -2.702e+03  2.783e+03  -0.971 0.331805
BedroomAbvGr              -5.940e+03  1.629e+03  -3.647 0.000275
BsmtFullBath               1.296e+04  2.156e+03   6.008 2.40e-09
BsmtHalfBath               5.119e+03  4.293e+03   1.192 0.233390
GarageCars                 1.760e+04  2.168e+03   8.120 1.03e-15
OverallCond:YearBuilt     -2.380e+01  3.602e+01  -0.661 0.508903
OverallQual:OverallCond   -1.221e+03  7.443e+02  -1.640 0.101237
Residuals:
    Min      1Q  Median      3Q     Max
-330388  -17896   -1905   16244  314131

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error t value
(Intercept)                     -1.342e+06  4.103e+05  -3.270
ID                              -1.292e+00  2.322e+00  -0.556
LotArea                          8.431e-01  1.097e-01   7.688
OverallQual                      2.689e+04  4.168e+03   6.453
OverallCond                      7.023e+04  6.893e+04   1.019
YearBuilt                        5.971e+02  2.118e+02   2.819
X1stFlrSF                        9.033e+01  4.335e+00  20.838
X2ndFlrSF                        7.471e+01  4.316e+00  17.312
LowQualFinSF                     4.501e+01  2.478e+01   1.816
GrLivArea                               NA         NA      NA
FullBath                        -4.485e+03  2.744e+03  -1.634
BedroomAbvGr                    -7.602e+03  1.616e+03  -4.704
BsmtFullBath                     1.275e+04  2.118e+03   6.022
BsmtHalfBath                     6.303e+03  4.219e+03   1.494
GarageCars                       1.604e+04  2.140e+03   7.497
OverallCond:YearBuilt           -2.825e+01  3.537e+01  -0.799
OverallQual:OverallCond         -1.246e+03  7.309e+02  -1.705
LotArea:X1stFlrSF:X2ndFlrSF     -7.506e-07  1.047e-07  -7.172
```

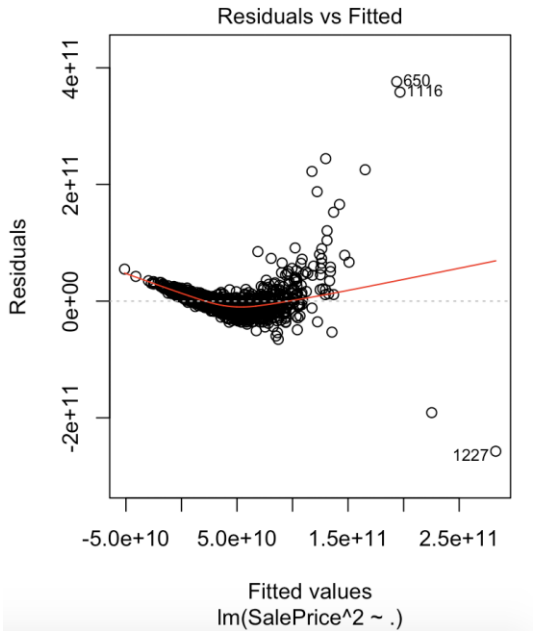
a. The only statistically significant interaction term in the three models above is that between LotArea (lot area), X1stFlrSF(Square footage of the first floor), and X2ndFlrSF(Square footage of the second floor), demonstrated through the '***' next to the predictor.

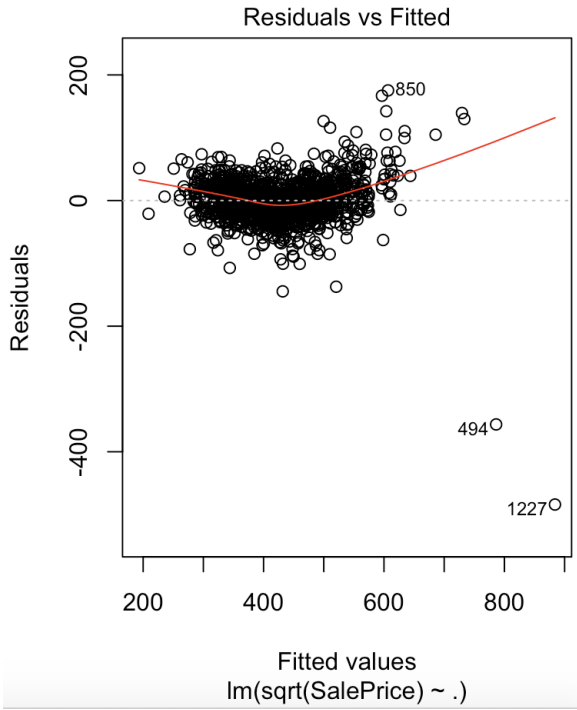
5.

a. LOG TRANSFORMATION



b. SQUARED TRANSFORMATION





c. The log and the square root transformations both brought the residual and fitted values to smaller values but the log transformation linearized the graphs the most, so it would be noteworthy to include a log transformation over the other two transformations in a model of SalePrice.