Kurtis Potier, Nick Hiller, Zane Shango
Professor Bushong
SSC 442

Lab 2 Write-Up

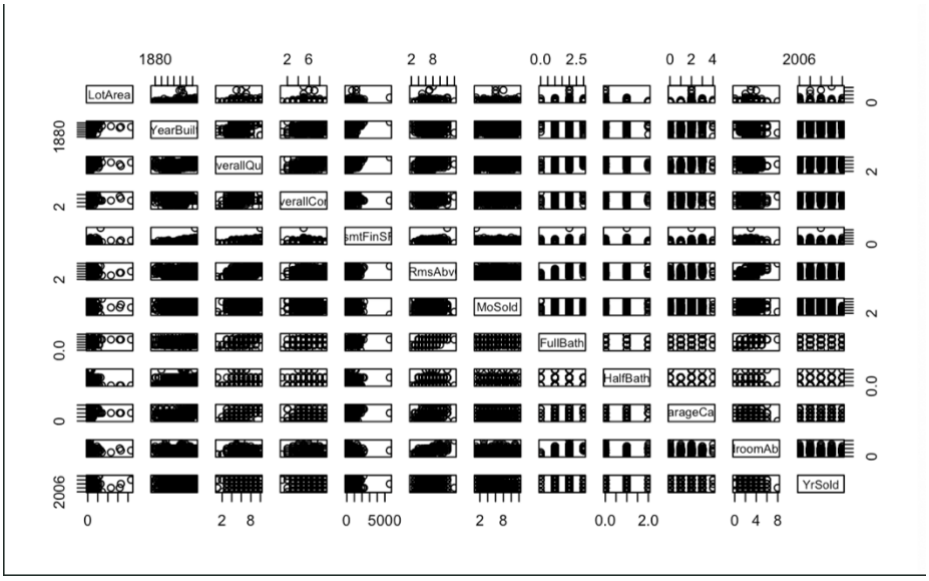
Exercise 1:

Below are the correlation results on the twelve variables we chose to go with SalePrice. We chose: LotArea, OverallQual, OverallCond, YearBuilt, BsmtFinSF1, FullBath, Halfbath, BedroomAbvGr, TotRmsAbvGrd, GarageCars, MoSold, and YrSold. Most of them matched our prior beliefs, however overall condition and sale price were negatively correlated which was surprising.

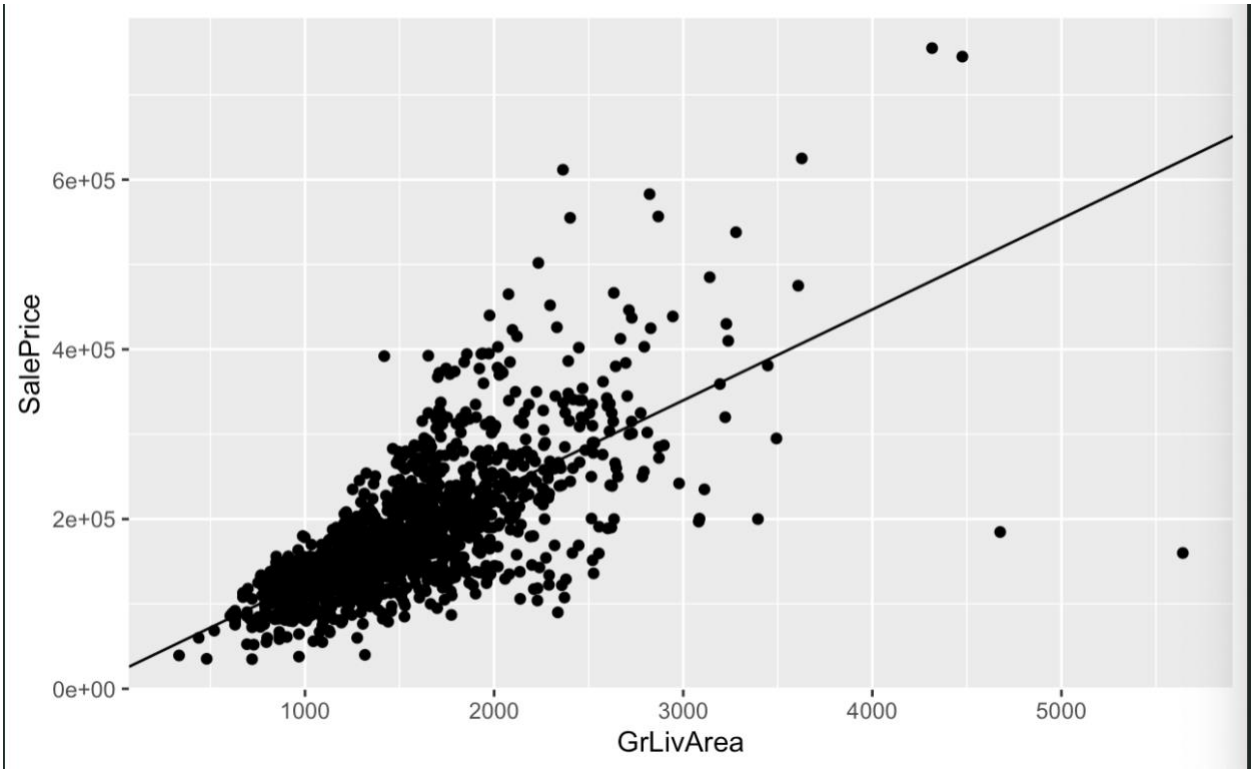
	LotArea	OverallQual	OverallCond	YearBuilt	BsmtFinSF1	FullBath	HalfBath
LotArea	1.000000000	0.10580574	-0.005636270	0.01422765	0.214103131	0.12603063	0.014259469
OverallQual	0.105805742	1.00000000	-0.091932343	0.57232277	0.239665966	0.55059971	0.273458099
OverallCond	-0.005636270	-0.09193234	1.000000000	-0.37598320	-0.046230856	-0.19414949	-0.060769327
YearBuilt	0.014227652	0.57232277	-0.375983196	1.00000000	0.249503197	0.46827079	0.242655910
BsmtFinSF1	0.214103131	0.23966597	-0.046230856	0.24950320	1.000000000	0.05854314	0.004262424
FullBath	0.126030627	0.55059971	-0.194149489	0.46827079	0.058543137	1.00000000	0.136380589
HalfBath	0.014259469	0.27345810	-0.060769327	0.24265591	0.004262424	0.13638059	1.000000000
BedroomAbvGr	0.119689908	0.10167636	0.012980060	-0.07065122	-0.107354677	0.36325198	0.226651484
TotRmsAbvGrd	0.190014778	0.42745234	-0.057583166	0.09558913	0.044315624	0.55478425	0.343414858
GarageCars	0.154870740	0.60067072	-0.185757511	0.53785009	0.224053522	0.46967204	0.219178152
MoSold	0.001204988	0.07081517	-0.003510839	0.01239847	-0.015726948	0.05587213	-0.009049888

	BedroomAbvGr	TotRmsAbvGrd	GarageCars	MoSold	YrSold	SalePrice
YrSold	-0.014261407	-0.02734671	0.043949746	-0.01361768	0.014358922	-0.01966884	-0.010268669
SalePrice	0.263843354	0.79098160	-0.077855894	0.52289733	0.386419806	0.56066376	0.284107676
LotArea	0.11968991	0.19001478	0.15487074	0.001204988	-0.01426141	0.26384335	
OverallQual	0.10167636	0.42745234	0.60067072	0.070815172	-0.02734671	0.79098160	
OverallCond	0.01298006	-0.05758317	-0.18575751	-0.003510839	0.04394975	-0.07785589	
YearBuilt	-0.07065122	0.09558913	0.53785009	0.012398471	-0.01361768	0.52289733	
BsmtFinSF1	-0.10735468	0.04431562	0.22405352	-0.015726948	0.01435892	0.38641981	
FullBath	0.36325198	0.55478425	0.46967204	0.055872129	-0.01966884	0.56066376	
HalfBath	0.22665148	0.34341486	0.21917815	-0.009049888	-0.01026867	0.28410768	
BedroomAbvGr	1.00000000	0.67661994	0.08610644	0.046543860	-0.03601389	0.16821315	
TotRmsAbvGrd	0.67661994	1.00000000	0.36228857	0.036907077	-0.03451635	0.53372316	
GarageCars	0.08610644	0.36228857	1.00000000	0.040521730	-0.03911690	0.64040920	
MoSold	0.04654386	0.03690708	0.04052173	1.000000000	-0.14572141	0.04643225	
YrSold	-0.03601389	-0.03451635	-0.03911690	-0.145721413	1.00000000	-0.02892259	
SalePrice	0.16821315	0.53372316	0.64040920	0.046432245	-0.02892259	1.00000000	

Here is the scatterplot matrix:



SalePrice and GrLivArea Scatterplot:



This is the house that is the largest outlier:

	Id <int>	MSSubClass <int>	MSZoning <fctr>	LotFrontage <int>	LotArea <int>	Street <fctr>	Alley <fctr>	LotShape <fctr>	LandContour <fctr>
899	899	20	RL	100	12919	Pave	*NA*	IR1	Lvl

1 row | 1-10 of 88 columns

Exercise 2:

Regression with SalePrice as the response to determine the value of IndoorGarage:

```
Call:
lm(formula = SalePrice ~ GarageOutside, data = ameslist)

Residuals:
    Min      1Q  Median      3Q     Max
-145200  -50986  -18100   32389  572051

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     180100       2465  73.059   <2e-16 ***
GarageOutside     2849       4591   0.621    0.535
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79460 on 1458 degrees of freedom
Multiple R-squared:  0.0002641, Adjusted R-squared:  -0.0004216
F-statistic: 0.3852 on 1 and 1458 DF,  p-value: 0.5349
```

After this test we can determine the value of an Indoor garage is $2,849; however, the p value does not indicate significance.

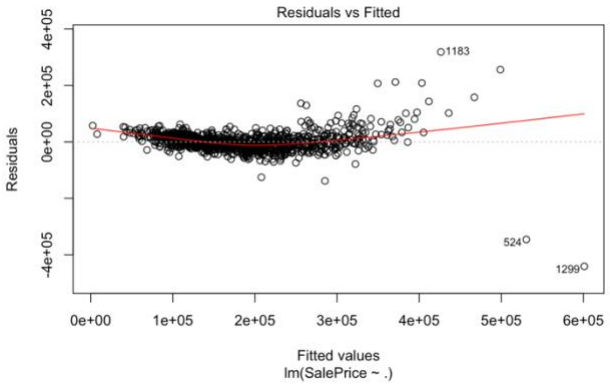
Multiple Linear Regression:

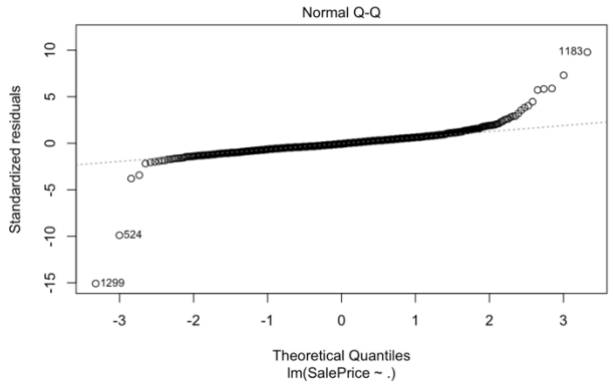
```
Call:
lm(formula = SalePrice ~ ., data = Ames)

Residuals:
    Min      1Q  Median      3Q     Max
-441034  -16641   -2397   14700  318576

Coefficients: (4 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.825e+05  1.707e+06  -0.224 0.822762
Id              -1.294e+00  2.661e+00  -0.486 0.626961
MSSubClass      -1.971e+02  3.464e+01  -5.690 1.63e-08 ***
LotFrontage     -1.198e+02  6.132e+01  -1.954 0.050952 .
LotArea          5.484e-01  1.576e-01   3.479 0.000522 ***
OverallQual      1.866e+04  1.483e+03  12.579  < 2e-16 ***
OverallCond      5.258e+03  1.369e+03   3.842 0.000129 ***
YearBuilt        3.096e+02  8.786e+01   3.524 0.000443 ***
YearRemodAdd     1.209e+02  8.676e+01   1.394 0.163608
MasVnrArea       3.145e+01  7.033e+00   4.472 8.57e-06 ***
BsmtFinSF1       1.709e+01  5.849e+00   2.923 0.003544 **
BsmtFinSF2       8.187e+00  8.784e+00   0.932 0.351496
BsmtUnfSF        4.924e+00  5.286e+00   0.931 0.351857
TotalBsmtSF             NA         NA      NA       NA
X1stFlrSF        4.597e+01  7.377e+00   6.231 6.62e-10 ***
X2ndFlrSF        4.609e+01  6.120e+00   7.530 1.07e-13 ***
LowQualFinSF     3.094e+01  2.797e+01   1.106 0.268852
GrLivArea               NA         NA      NA       NA
BsmtFullBath     8.996e+03  3.200e+03   2.811 0.005028 **
BsmtHalfBath     2.578e+03  5.089e+03   0.507 0.612535
FullBath         5.536e+03  3.548e+03   1.560 0.118951
HalfBath        -1.042e+03  3.334e+03  -0.313 0.754617
BedroomAbvGr    -1.013e+04  2.160e+03  -4.688 3.11e-06 ***
KitchenAbvGr    -2.292e+04  6.778e+03  -3.382 0.000745 ***
TotRmsAbvGrd     5.519e+03  1.490e+03   3.704 0.000223 ***
Fireplaces       4.505e+03  2.194e+03   2.053 0.040330 *
GarageYrBlt     -3.787e+01  9.122e+01  -0.415 0.678143
GarageCars       1.685e+04  3.498e+03   4.816 1.67e-06 ***
GarageArea       6.055e+00  1.214e+01   0.499 0.618184
WoodDeckSF       2.063e+01  1.004e+01   2.054 0.040180 *
OpenPorchSF     -1.817e+00  1.953e+01  -0.093 0.925868
EnclosedPorch    7.339e+00  2.064e+01   0.356 0.722247
X3SsnPorch       3.622e+01  3.763e+01   0.963 0.335910
ScreenPorch      5.820e+01  2.043e+01   2.849 0.004466 **
PoolArea        -5.744e+01  2.996e+01  -1.917 0.055457 .
MiscVal         -3.902e+00  6.995e+00  -0.558 0.577118
MoSold          -2.367e+02  4.237e+02  -0.559 0.576570
YrSold          -2.261e+02  8.485e+02  -0.267 0.789887
GarageType2Types 5.286e+03  1.882e+04   0.281 0.778821
GarageTypeAttchd -3.896e+03  2.518e+03  -1.547 0.122164
GarageTypeBasment -4.929e+03  9.567e+03  -0.515 0.606502
GarageTypeBuiltIn -7.549e+03  4.790e+03  -1.576 0.115283
GarageTypeCarPort  8.532e+03  1.346e+04   0.634 0.526439
GarageTypeDetchd          NA         NA      NA       NA
GarageOutside             NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36810 on 1080 degrees of freedom
  (339 observations deleted due to missingness)
Multiple R-squared:  0.8104,    Adjusted R-squared:  0.8033
F-statistic: 115.4 on 40 and 1080 DF,  p-value: < 2.2e-16
```

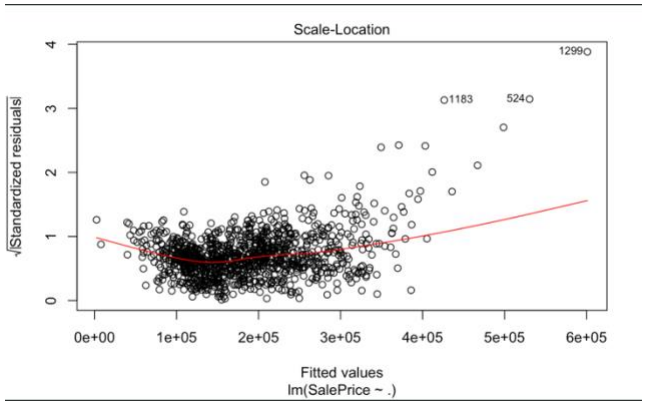
There is a relationship between the predictors and the response. The variables: MSSubClass,LotArea, OverallQual, OverallCond, YearBuilt,MasVnrArea,BsmtFinSF1, X1stFlrSF,

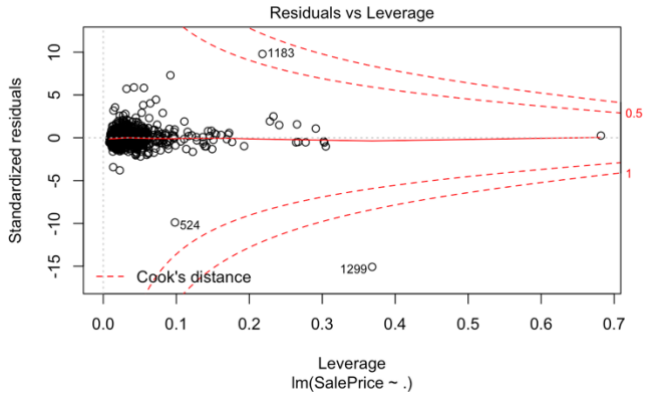
X2ndFlrSF, BsmtFullBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, and ScreenPorch appear to have a statistically significant relationship. The coefficient for the year variable suggests that the trend of housing prices was downward for the duration of the data collection.

Diagnostic Plots of Linear Regression Fit:









There are several large outliers; one observation has very high leverage, some others have moderately high leverage.

Recall that the operator : designates the interaction between two variables; and the operator * designates the interaction between the two variables plus the main effects.

Find Statistically Significant Interactions:

```
Call:
lm(formula = SalePrice ~ GrLivArea * FullBath, data = ameslist)

Residuals:
    Min      1Q  Median      3Q     Max
-405547  -26243   -1872   20913  343810

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        38613.215  11906.700   3.243  0.00121 **
GrLivArea             62.712      8.847   7.089  2.1e-12 ***
FullBath            7270.139   7018.848   1.036  0.30047
GrLivArea:FullBath    14.053      4.326   3.248  0.00119 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54680 on 1456 degrees of freedom
Multiple R-squared:  0.5272,    Adjusted R-squared:  0.5262
F-statistic: 541.2 on 3 and 1456 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = SalePrice ~ LotFrontage:LotArea, data = ameslist)

Residuals:
    Min      1Q  Median      3Q     Max
-342721  -50310  -20140   32035  549760

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.677e+05  2.779e+03  60.354   <2e-16 ***
LotFrontage:LotArea 1.675e-02  1.934e-03   8.664   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80930 on 1199 degrees of freedom
  (259 observations deleted due to missingness)
Multiple R-squared:  0.05892,   Adjusted R-squared:  0.05814
F-statistic: 75.07 on 1 and 1199 DF,  p-value: < 2.2e-16
```

Transformation of Variables:

- Baseline

```
Call:
lm(formula = SalePrice ~ LotArea, data = ameslist)

Residuals:
    Min      1Q  Median      3Q     Max
-275668  -48169  -17725   31248  553356

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.588e+05  2.915e+03   54.49   <2e-16 ***
LotArea     2.100e+00  2.011e-01   10.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76650 on 1458 degrees of freedom
Multiple R-squared:  0.06961,   Adjusted R-squared:  0.06898
F-statistic: 109.1 on 1 and 1458 DF,  p-value: < 2.2e-16
```

- Ln: The t value improved from 10 to 16 when using an ln function and the r^2 also got much better, going from 0.07 to 0.15.

```
Call:
lm(formula = SalePrice ~ log(LotArea), data = ameslist)

Residuals:
    Min      1Q  Median      3Q     Max
-145261  -47995  -16602   34118  531531

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -362527      33809  -10.72   <2e-16 ***
log(LotArea)    59648       3705   16.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73230 on 1458 degrees of freedom
Multiple R-squared:  0.1509,    Adjusted R-squared:  0.1504
F-statistic: 259.2 on 1 and 1458 DF,  p-value: < 2.2e-16
```

- Squared: There is no difference in result when the value is squared

```
Call:
lm(formula = SalePrice ~ (LotArea * LotArea), data = ameslist)

Residuals:
    Min      1Q  Median      3Q     Max
-275668  -48169  -17725   31248  553356

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.588e+05  2.915e+03   54.49   <2e-16 ***
LotArea     2.100e+00  2.011e-01   10.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76650 on 1458 degrees of freedom
Multiple R-squared:  0.06961,   Adjusted R-squared:  0.06898
F-statistic: 109.1 on 1 and 1458 DF,  p-value: < 2.2e-16
```

- Square Root: Using the square root, the r^2 value got better, going from .07 to .14 and the t value also improved from 10 to 15.

```
Call:
lm(formula = SalePrice ~ sqrt(LotArea), data = ameslist)

Residuals:
    Min      1Q  Median      3Q     Max
-263407  -46335  -16276   32787  537176

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   81166.31    6905.23   11.75   <2e-16 ***
sqrt(LotArea)  1013.33      67.33   15.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73930 on 1458 degrees of freedom
Multiple R-squared:  0.1345,    Adjusted R-squared:  0.1339
F-statistic: 226.5 on 1 and 1458 DF,  p-value: < 2.2e-16
```