Jianqi Liu
Aver Mckay
Dylan Patel

Lab 2

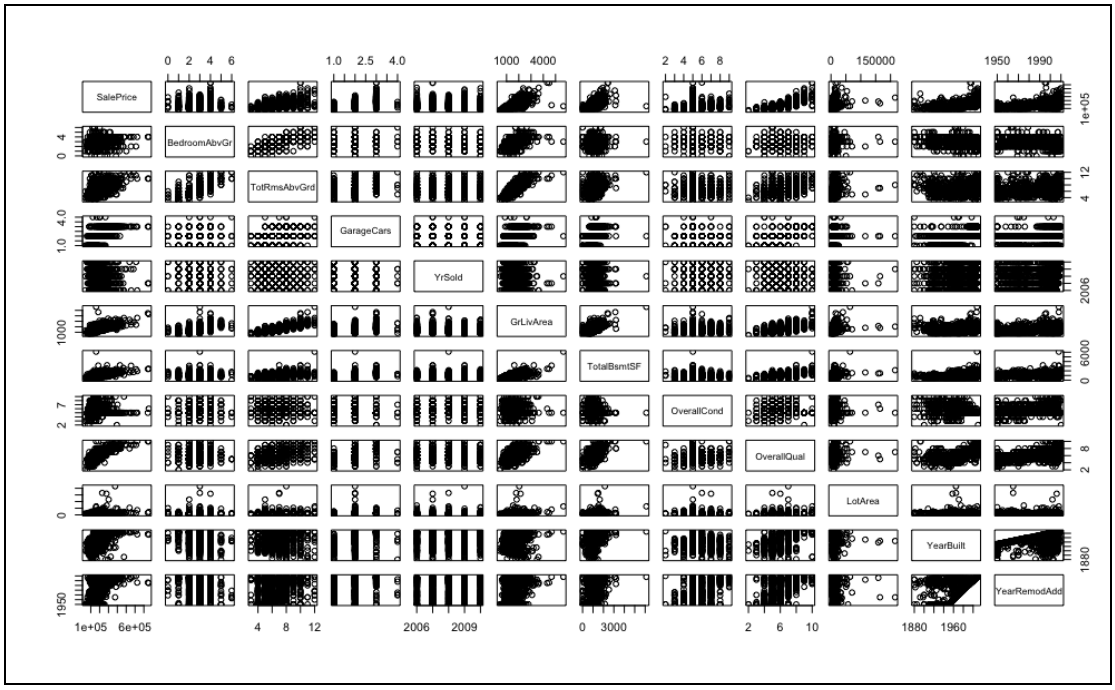
Exercise 1:

1.

```
int_type = lapply(ameslist, class)
#get name which type is int
Ames = ameslist[int_type=='integer']

names(Ames)
#check the names we want to leave
Ames <-Ames[ , !(names(Ames) %in% c("MSSubClass", "MasVnrArea", "BsmtFinSF1", "BsmtFinSF2", "BsmntUnfSF", "LowQualFinSF", "X3SsnPorch", "MiscVal"))]

names(Ames)
#save the new select data
save(Ames, file = "Ames.txt")
```

2.

```
#12 of the variables that are type = int in the data set.
features <-c("SalePrice","BedroomAbvGr","TotRmsAbvGrd","GarageCars", "YrSold", "GrLivArea", "TotalBsmtSF", "OverallCond","OverallQual","LotArea","YearBuilt","YearRemodAdd")
pairs(Ames[,features])
```



3.

A matrix of correlation between the 12 variables. The only two correlations we fond that confused us were that the data says as year sold increases the sales price decreases, along as if the condition of the house increases, the sales price decreases.

```
cor(Ames[,features])
```

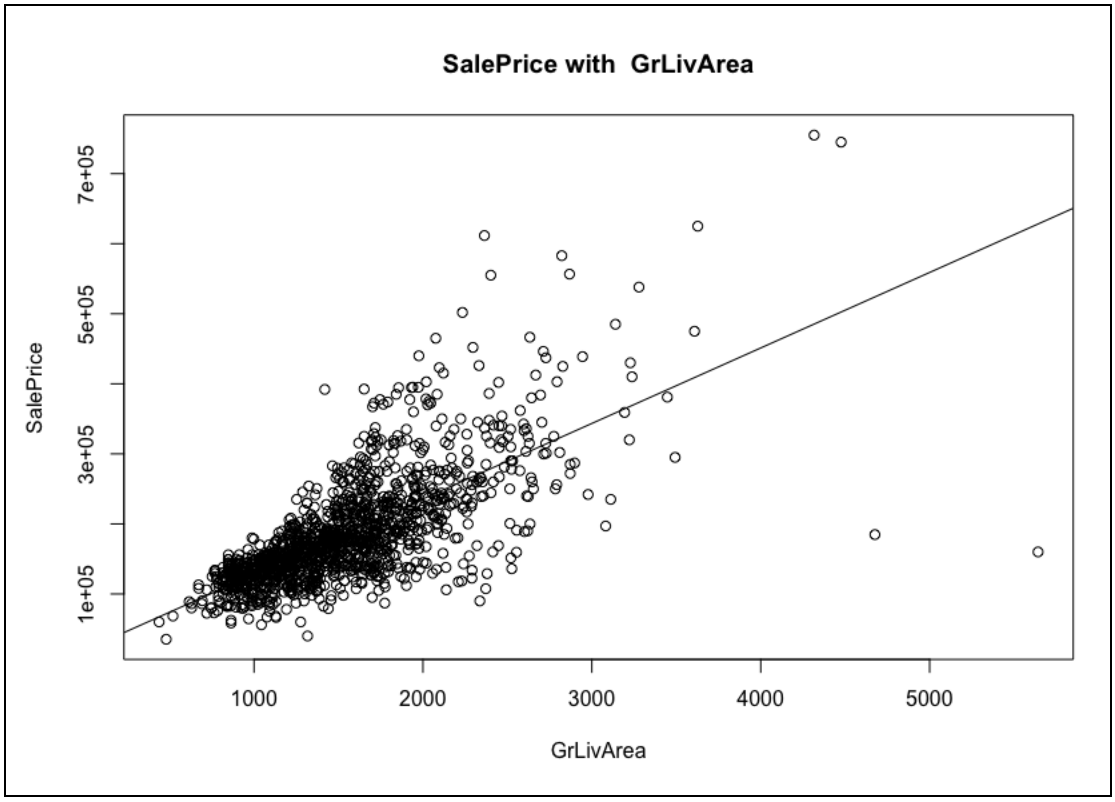
```
> cor(Ames[,features])
              SalePrice BedroomAbvGr TotRmsAbvGrd  GarageCars       YrSold
SalePrice     1.00000000   0.16465495   0.53830912  0.63709541 -0.026725513
BedroomAbvGr  0.16465495   1.00000000   0.66596277  0.11752232 -0.031268531
TotRmsAbvGrd  0.53830912   0.66596277   1.00000000  0.40017023 -0.038735406
GarageCars    0.63709541   0.11752232   0.40017023  1.00000000 -0.041069077
YrSold       -0.02672551  -0.03126853  -0.03873541 -0.04106908  1.000000000
GrLivArea     0.70817211   0.51231197   0.82097478  0.48389867 -0.040161906
TotalBsmtSF   0.60358341   0.02390892   0.26818757  0.43280422 -0.013975050
OverallCond  -0.09527774   0.01498466  -0.06857329 -0.24731706  0.050308554
OverallQual   0.78722783   0.08171377   0.42683435  0.58138144 -0.019272613
LotArea       0.25292146   0.11974650   0.18572391  0.13797759 -0.013797348
YearBuilt     0.50758406  -0.07179447   0.10153857  0.52334948 -0.006809173
YearRemodAdd  0.50543406  -0.06737458   0.17337459  0.45065920  0.039936949
               GrLivArea TotalBsmtSF  OverallCond OverallQual      LotArea
SalePrice     0.70817211  0.60358341 -0.095277741  0.78722783  0.252921459
BedroomAbvGr  0.51231197  0.02390892  0.014984665  0.08171377  0.119746500
TotRmsAbvGrd  0.82097478  0.26818757 -0.068573288  0.42683435  0.185723905
GarageCars    0.48389867  0.43280422 -0.247317060  0.58138144  0.137977589
YrSold       -0.04016191 -0.01397505  0.050308554 -0.01927261 -0.013797348
GrLivArea     1.00000000  0.44214629 -0.092217302  0.58958384  0.257243272
TotalBsmtSF   0.44214629  1.00000000 -0.182020072  0.53197707  0.252932445
OverallCond  -0.09221730 -0.18202007  1.000000000 -0.13623220 -0.002869219
OverallQual   0.58958384  0.53197707 -0.136232205  1.00000000  0.090016311
LotArea       0.25724327  0.25293245 -0.002869219  0.09001631  1.000000000
YearBuilt     0.19466267  0.37697704 -0.403601675  0.57208246 -0.005920805
YearRemodAdd  0.27886410  0.28357204  0.048339636  0.55777194  0.002763920
               YearBuilt YearRemodAdd
SalePrice     0.507584064   0.50543406
BedroomAbvGr -0.071794474  -0.06737458
TotRmsAbvGrd  0.101538568   0.17337459
GarageCars    0.523349483   0.45065920
YrSold       -0.006809173   0.03993695
GrLivArea     0.194662669   0.27886410
TotalBsmtSF   0.376977038   0.28357204
OverallCond  -0.403601675   0.04833964
OverallQual   0.572082457   0.55777194
LotArea      -0.005920805   0.00276392
YearBuilt     1.000000000   0.61805808
YearRemodAdd  0.618058076   1.00000000
```

4

Realtionship between SalePrice and GrLivArea.

The largest outlier is at x~4200, y~7e+05

```
attach(Ames)
lm.fit = lm(SalePrice ~ GrLivArea)
plot(Ames$GrLivArea, Ames$SalePrice, main = "price with living space", ylab =
"price", xlab = "living space") + abline(lm.fit)
```



Exercise 2:

1. simple linear regression

```
> summary(model1)

Call:
lm(formula = SalePrice ~ ameslist$GarageOutside)

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 ***
ameslist$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
```

2.

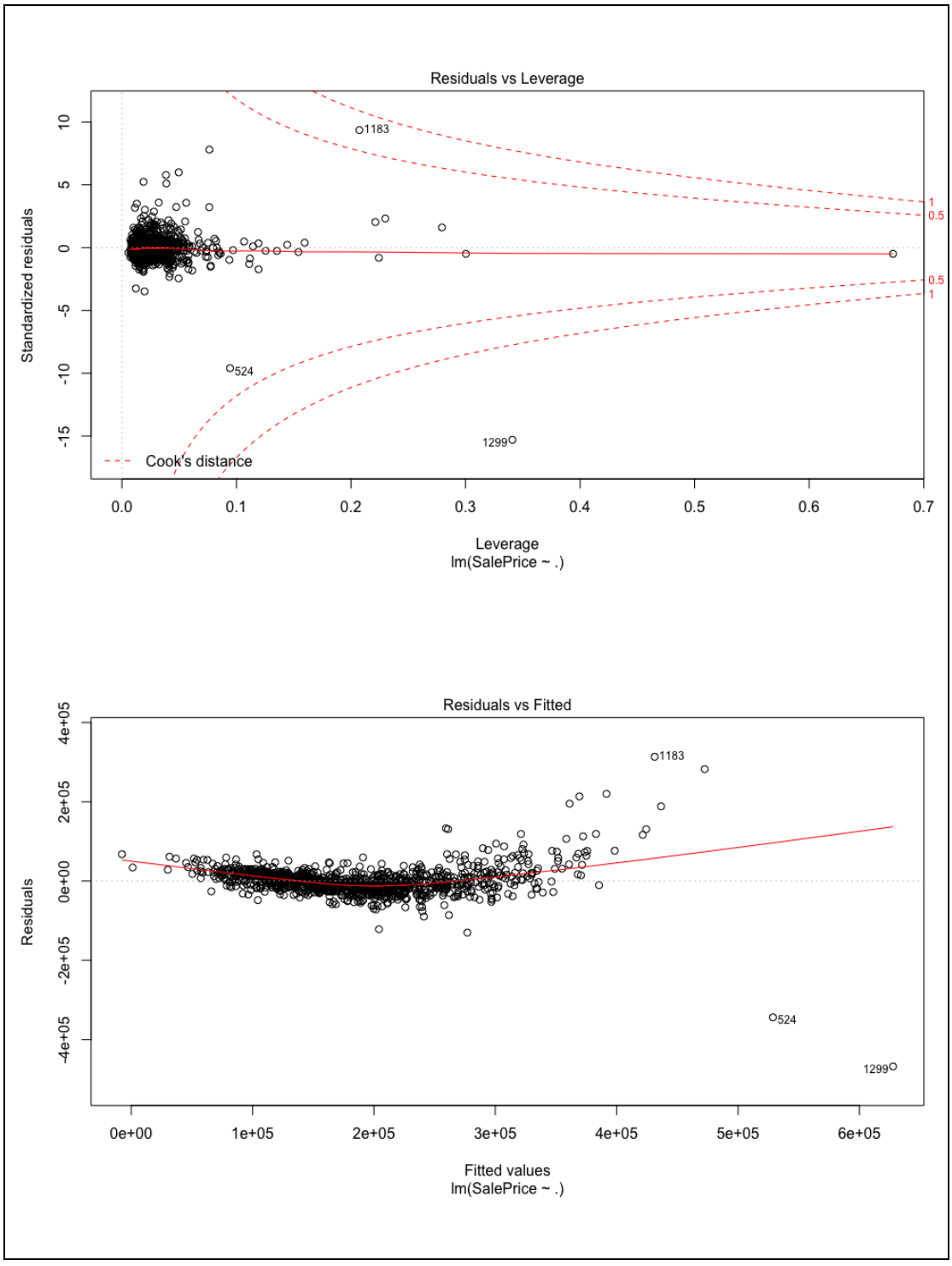
Most of the variables do have a correlation with the sales price. The statistically significant variables seem to be the lot area, overall quality, overall condition, year built, year remodeled, size of basement, number of kitchens, number of bedrooms, number of total rooms, how many cars the garage is able to fit, the deck size, and the area of the pool if it has one. The year sold variable is deemed statistically insignificant.

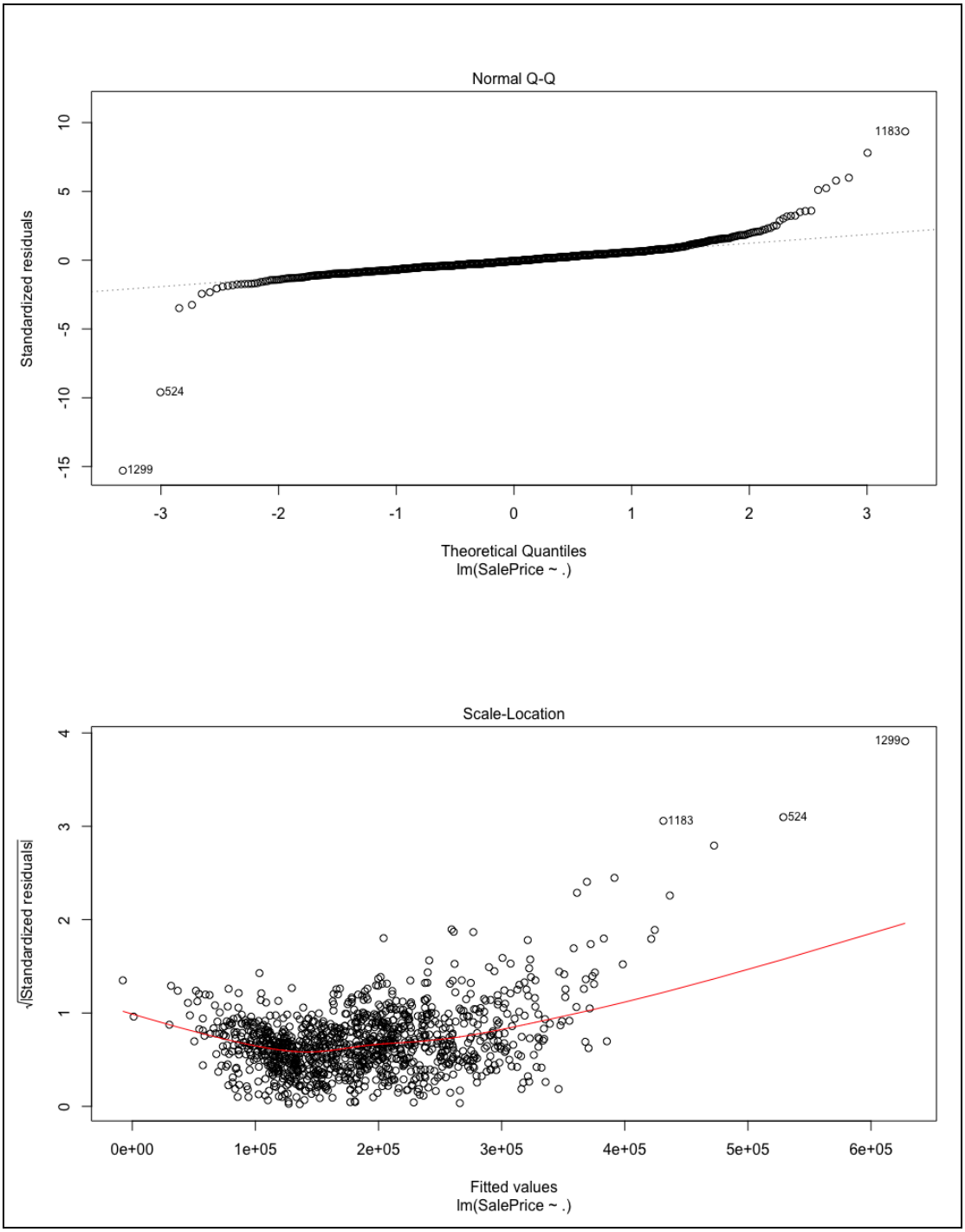
```
> #===================================2.0=================================
> model2 <- lm(SalePrice ~ ., data = Ames)
> summary(model2)

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

Residuals:
    Min      1Q  Median      3Q     Max
-467752  -16792   -2180   14737  313676

Coefficients:
                  Estimate   Std. Error t value    Pr(>|t|)
(Intercept)   -165164.5301 1735624.0125  -0.095    0.924204
Id                 -2.2838       2.7023  -0.845    0.398217
LotFrontage         4.2922      59.1989   0.073    0.942213
LotArea             0.5270       0.1598   3.297    0.001007 **
OverallQual     18663.9699    1498.8854  12.452     < 2e-16 ***
OverallCond      5610.1322    1393.0492   4.027 0.000060325 ***
YearBuilt         356.2908      88.8156   4.012 0.000064406 ***
YearRemodAdd      102.1997      88.2394   1.158    0.247031
BsmtUnfSF         -12.2739       3.9076  -3.141    0.001729 **
TotalBsmtSF        21.0194       5.8508   3.593    0.000342 ***
X1stFlrSF          26.9960      29.2112   0.924    0.355604
X2ndFlrSF          20.2645      28.6139   0.708    0.478968
GrLivArea          21.9289      28.5270   0.769    0.442233
BsmtFullBath     6911.3490    3250.1017   2.127    0.033684 *
BsmtHalfBath      508.1952    5168.8700   0.098    0.921697
FullBath         3371.4809    3586.1907   0.940    0.347359
HalfBath         -823.5253    3367.7622  -0.245    0.806865
BedroomAbvGr    -9382.9562    2177.9509  -4.308 0.000017939 ***
KitchenAbvGr   -34641.1444    6473.0758  -5.352 0.000000106 ***
TotRmsAbvGrd     6271.2513    1509.5892   4.154 0.000035169 ***
Fireplaces       3839.9791    2227.2889   1.724    0.084979 .
GarageYrBlt       -98.2759      92.7365  -1.060    0.289500
GarageCars      17600.2393    3567.7720   4.933 0.000000935 ***
GarageArea         14.1265      12.3418   1.145    0.252623
WoodDeckSF         23.0022      10.1956   2.256    0.024261 *
OpenPorchSF        -9.1742      19.7157  -0.465    0.641793
EnclosedPorch       5.5595      21.0431   0.264    0.791678
ScreenPorch        55.8002      20.5785   2.712    0.006801 **
PoolArea          -84.3404      30.1707  -2.795    0.005273 **
MoSold            -67.9712     429.5609  -0.158    0.874301
YrSold           -313.8101     863.0879  -0.364    0.716234
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37670 on 1096 degrees of freedom
  (252 observations deleted due to missingness)
Multiple R-squared:  0.8005,    Adjusted R-squared:  0.7951
F-statistic: 146.6 on 30 and 1096 DF,  p-value: < 2.2e-16
```

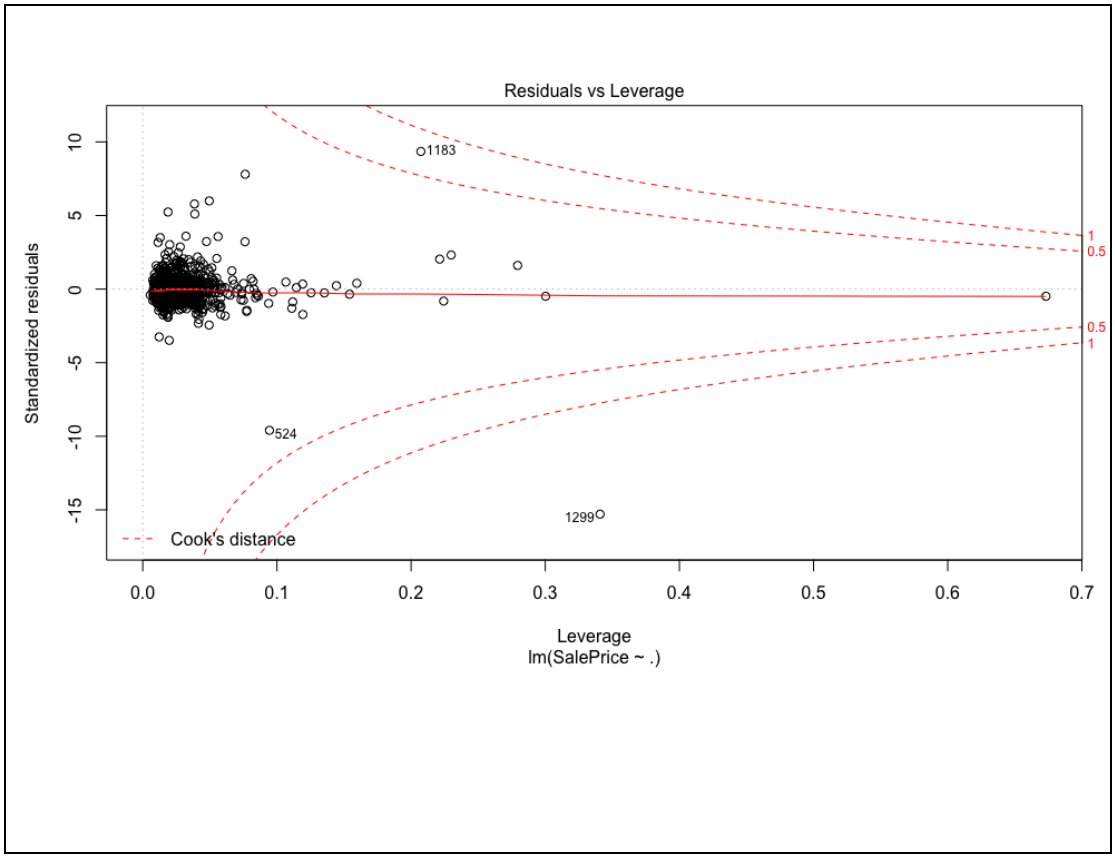
3.

Let's first look at the residual plot and the fit plot. As we can see, it is a relatively

straight line with a uniform distribution of residuals. This is a good thing because it reflects that there is a non-linear relationship. However, if we look closely, there will be a slight parabolic shape, which may reflect a slight non-linear relationship. In addition, as sales prices rose, we noticed that the data began to have larger residuals and more outliers.




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(SalePrice ~ .)
1183
524
1299
Scale-Location
√|Standardized residuals|
Fitted values
lm(SalePrice ~ .)
1299
1183
524



4.

When in relation to sales price, the overall quality, lot area, lot frontage, ground living area, and lot frontage * lot area all are deemed statistically significant.

```
> model4 <- lm(SalePrice ~ GrLivArea + OverallQual + LotArea + LotFrontage +  LotFrontage
LotArea, data=Ames)
> summary(model4)

Call:
lm(formula = SalePrice ~ GrLivArea + OverallQual + LotArea +
    LotFrontage + LotFrontage * LotArea, data = Ames)

Residuals:
    Min      1Q  Median      3Q     Max
-317851  -21973   -2033   19720  279047

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.648e+05  6.811e+03 -24.200  < 2e-16 ***
GrLivArea            4.679e+01  3.152e+00  14.843  < 2e-16 ***
OverallQual          3.510e+04  1.115e+03  31.490  < 2e-16 ***
LotArea              4.791e+00  3.639e-01  13.165  < 2e-16 ***
LotFrontage          5.247e+02  6.608e+01   7.941 4.84e-15 ***
LotArea:LotFrontage -3.120e-02  2.595e-03 -12.027  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41090 on 1121 degrees of freedom
  (252 observations deleted due to missingness)
Multiple R-squared:  0.7573,    Adjusted R-squared:  0.7562
F-statistic: 699.5 on 5 and 1121 DF,  p-value: < 2.2e-16
```

5.

Taking the log of a finished basement could be useful when calculating percent changes of a given price, however in this model keeping just the normal basement would suffice. Square rooting and squaring the data would also not make sense in this data when using quantitative units such as the number of bathrooms and basements.

```
> model5 <- lm(SaePrice ~ log(GrLivArea), data = Ames)
Error in eval(predvars, data, env) : 找不到对象'SaePrice'
> model5 <- lm(SalePrice ~ log(GrLivArea), data = Ames)
> summary(model5)

Call:
lm(formula = SalePrice ~ log(GrLivArea), data = Ames)

Residuals:
    Min      1Q  Median      3Q     Max
-255521  -31667   -2531   24797  384982

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1051582      34639  -30.36   <2e-16 ***
log(GrLivArea)   169843       4751   35.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56930 on 1377 degrees of freedom
Multiple R-squared:  0.4813,    Adjusted R-squared:  0.481
F-statistic:  1278 on 1 and 1377 DF,  p-value: < 2.2e-16

> model6 <- lm(SalePrice ~ LotArea + I(LotArea^2), data = Ames)
> summary(model6)

Call:
lm(formula = SalePrice ~ LotArea + I(LotArea^2), data = Ames)

Residuals:
    Min      1Q  Median      3Q     Max
-226736  -46888  -17016   31590  532360

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.307e+05  4.523e+03  28.900   <2e-16 ***
LotArea       5.645e+00  4.267e-01  13.228   <2e-16 ***
I(LotArea^2) -2.562e-05  2.637e-06  -9.719   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74010 on 1376 degrees of freedom
Multiple R-squared:  0.1241,    Adjusted R-squared:  0.1228
F-statistic: 97.47 on 2 and 1376 DF,  p-value: < 2.2e-16
```

```
> model7 <- lm(SalePrice ~ sqrt(LotArea), data = Ames)
> summary(model7)

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

Residuals:
    Min      1Q  Median      3Q     Max
-248965  -46975  -16724   31533  534980

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   90768.15    7110.10   12.77   <2e-16 ***
sqrt(LotArea)   954.07      68.75   13.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74040 on 1377 degrees of freedom
Multiple R-squared:  0.1227,    Adjusted R-squared:  0.1221
F-statistic: 192.6 on 1 and 1377 DF,  p-value: < 2.2e-16
```