Lab 2 – Group 1

Linear Regression and Simple Analyses

Exercise #1

Introduction:

Headers in the Original Dataset

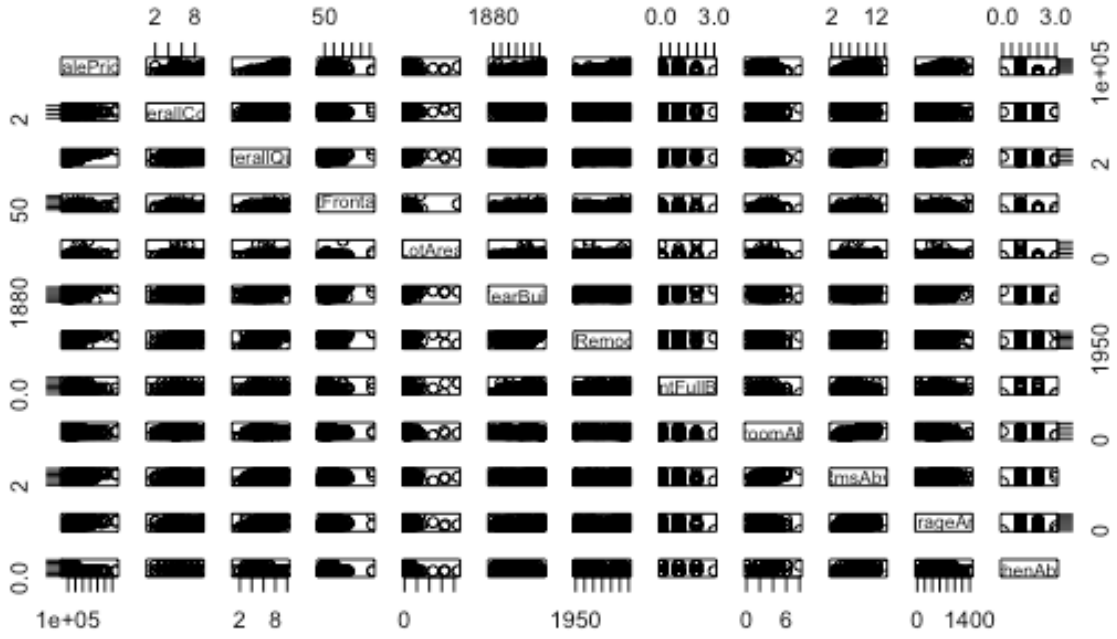
1. With Header = TRUE, the dataset formats certain variables as integers and factors, as mentioned.

2. With Header = FALSE, all the variables are marked as factors.

3. Without either Header = TRUE or Header = FALSE, the variable titles are gone and the variables are all factors.

Question 1: TXT File present in GitHub Repository.

Question 2 – Produce a scatterplot matric which includes 12 variables of type = int:



Question 3 – Produce a matrix of correlations between variables using function cor().

	Ames.SalePrice	Ames.OverallCond	Ames.OverallQual	Ames.LotFrontage
Ames.SalePrice	1.00000000	-0.07785589	0.79098160	NA
Ames.OverallCond	-0.07785589	1.00000000	-0.09193234	NA
Ames.OverallQual	0.79098160	-0.09193234	1.00000000	NA
Ames.LotFrontage	NA	NA	NA	1
Ames.LotArea	0.26384335	-0.00563627	0.10580574	NA
Ames.YearBuilt	0.52289733	-0.37598320	0.57232277	NA
Ames.YearRemodAdd	0.50710097	0.07374150	0.55068392	NA
Ames.BsmtFullBath	0.22712223	-0.05494152	0.11109779	NA
Ames.BedroomAbvGr	0.16821315	0.01298006	0.10167636	NA
Ames.TotRmsAbvGrd	0.53372316	-0.05758317	0.42745234	NA
Ames.GarageArea	0.62343144	-0.15152137	0.56202176	NA
Ames.KitchenAbvGr	-0.13590737	-0.08700086	-0.18388223	NA

	Ames.LotArea	Ames.YearBuilt	Ames.YearRemodAdd	Ames.BsmtFullBath
Ames.SalePrice	0.26384335	0.52289733	0.50710097	0.22712223
Ames.OverallCond	-0.00563627	-0.37598320	0.07374150	-0.05494152
Ames.OverallQual	0.10580574	0.57232277	0.55068392	0.11109779
Ames.LotFrontage	NA	NA	NA	NA
Ames.LotArea	1.00000000	0.01422765	0.01378843	0.15815453
Ames.YearBuilt	0.01422765	1.00000000	0.59285498	0.18759855
Ames.YearRemodAdd	0.01378843	0.59285498	1.00000000	0.11946988
Ames.BsmtFullBath	0.15815453	0.18759855	0.11946988	1.00000000
Ames.BedroomAbvGr	0.11968991	-0.07065122	-0.04058093	-0.15067281
Ames.TotRmsAbvGrd	0.19001478	0.09558913	0.19173982	-0.05327524
Ames.GarageArea	0.18040276	0.47895382	0.37159981	0.17918948
Ames.KitchenAbvGr	-0.01778387	-0.17480025	-0.14959752	-0.04150255

	Ames.BedroomAbvGr	Ames.TotRmsAbvGrd	Ames.GarageArea	Ames.KitchenAbvGr
Ames.SalePrice	0.16821315	0.53372316	0.62343144	-0.13590737
Ames.OverallCond	0.01298006	-0.05758317	-0.15152137	-0.08700086
Ames.OverallQual	0.10167636	0.42745234	0.56202176	-0.18388223
Ames.LotFrontage	NA	NA	NA	NA
Ames.LotArea	0.11968991	0.19001478	0.18040276	-0.01778387
Ames.YearBuilt	-0.07065122	0.09558913	0.47895382	-0.17480025
Ames.YearRemodAdd	-0.04058093	0.19173982	0.37159981	-0.14959752
Ames.BsmtFullBath	-0.15067281	-0.05327524	0.17918948	-0.04150255
Ames.BedroomAbvGr	1.00000000	0.67661994	0.06525253	0.19859676
Ames.TotRmsAbvGrd	0.67661994	1.00000000	0.33782212	0.25604541
Ames.GarageArea	0.06525253	0.33782212	1.00000000	-0.06443305
Ames.KitchenAbvGr	0.19859676	0.25604541	-0.06443305	1.00000000

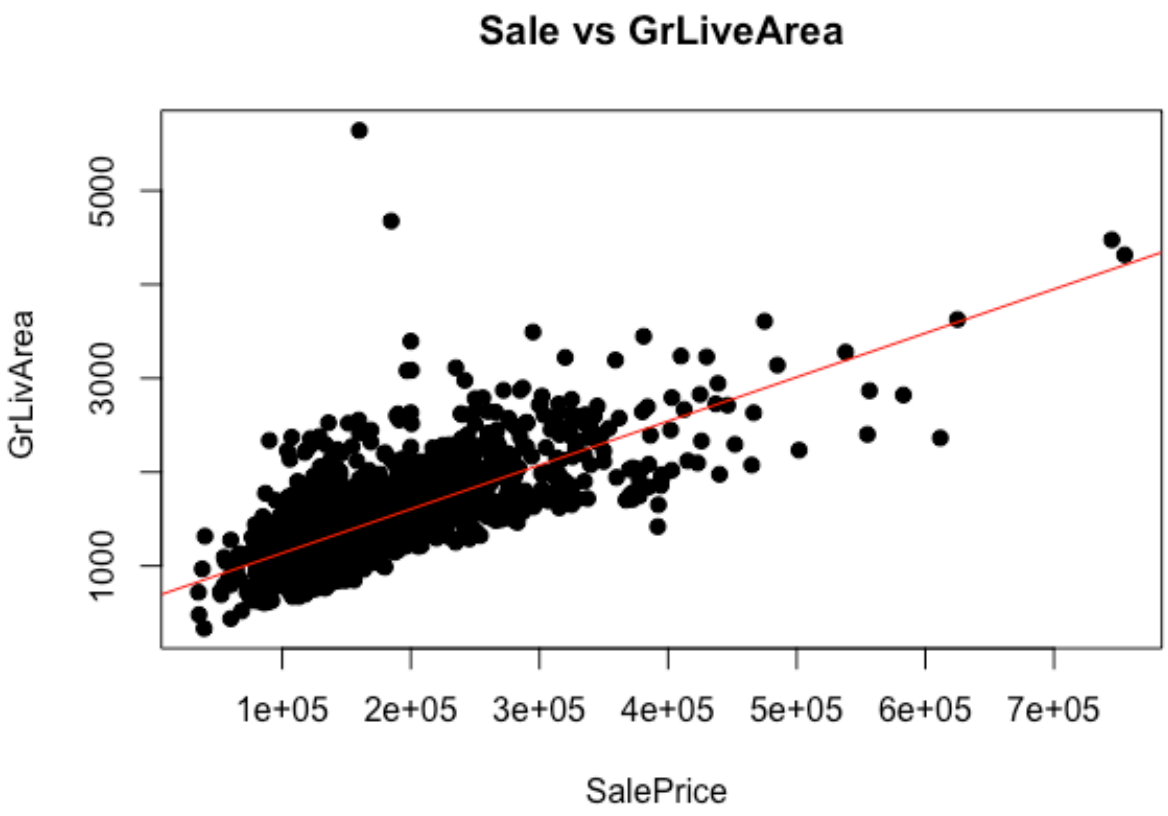
Observations and comments on the above correlation matrix:

1. The correlation matrix above does match original beliefs. We assume Sale Price (as the reference variable) to be highly correlated with the Overall Condition of the house and not much so correlated to the year the house was remodeled. Similarly, there are other variables (such as total Rooms above ground and garage area) that, from prior knowledge following the housing

industry that would be expected to correlate more positively to the growth in price of a house. Other factors, perhaps not so much.

Question 4 – Produce a scatterplot between SalePrice and GrLivArea.

Observations and comments on the above scatterplot:



1. The largest outlier that is above the regression line is a little below the 2e+05 Sale Price.
2. The largest outlier is the position row 1229 of the modified Ames dataset. Other information on this outlier can be viewed in the R-Markdown.

Exercise #2

Question 1 – Produce a simple linear regression of Sale Price to determine the value of an indoor garage:

```
lm(formula = SalePrice ~ GarageType)

Coefficients:
      (Intercept)   GarageTypeAttchd  GarageTypeBasment  GarageTypeBuiltIn
           151283              51609               9287             103468
GarageTypeCarPort   GarageTypeDetchd
           -41321             -17192
```

Question 2 – Perform a multiple linear regression:

```
Call:
lm(formula = SalePrice ~ Id + OverallQual + MasVnrArea + TotalBsmtSF +
    GrLivArea + HalfBath + Fireplaces + WoodDeckSF + ScreenPorch +
    YrSold + MSSubClass + OverallCond + BsmtFinSF1 + X1stFlrSF +
    BsmtFullBath + BedroomAbvGr + GarageYrBlt + OpenPorchSF +
    PoolArea + LotFrontage + YearBuilt + BsmtFinSF2 + X2ndFlrSF +
    BsmtHalfBath + KitchenAbvGr + GarageCars + EnclosedPorch +
    MiscVal + LotArea + YearRemodAdd + BsmtUnfSF + LowQualFinSF +
    FullBath + TotRmsAbvGrd + GarageArea + X3SsnPorch + MoSold,
    data = Ames)

Coefficients:
  (Intercept)            Id    OverallQual     MasVnrArea    TotalBsmtSF
   -3.351e+05    -1.205e+00     1.866e+04     3.141e+01     5.005e+00
    GrLivArea      HalfBath    Fireplaces    WoodDeckSF    ScreenPorch
    3.341e+01    -1.098e+03     4.372e+03     2.144e+01     5.805e+01
       YrSold    MSSubClass   OverallCond    BsmtFinSF1     X1stFlrSF
   -2.474e+02    -2.001e+02     5.239e+03     1.235e+01     1.257e+01
BsmtFullBath   BedroomAbvGr   GarageYrBlt   OpenPorchSF      PoolArea
    9.043e+03    -1.022e+04    -4.728e+01    -2.252e+00    -6.052e+01
 LotFrontage      YearBuilt    BsmtFinSF2     X2ndFlrSF   BsmtHalfBath
   -1.160e+02     3.164e+02     3.337e+00     1.322e+01     2.465e+03
KitchenAbvGr    GarageCars  EnclosedPorch       MiscVal       LotArea
   -2.202e+04     1.685e+04     7.295e+00    -3.761e+00     5.422e-01
YearRemodAdd     BsmtUnfSF   LowQualFinSF      FullBath   TotRmsAbvGrd
    1.194e+02            NA            NA     5.433e+03     5.464e+03
  GarageArea    X3SsnPorch        MoSold
    6.274e+00     3.349e+01    -2.217e+02
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-442182  -16955   -2824   15125  318183

Coefficients: (2 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.351e+05  1.701e+06  -0.197 0.843909
Id            -1.205e+00  2.658e+00  -0.453 0.650332
OverallQual    1.866e+04  1.482e+03  12.592  < 2e-16 ***
MasVnrArea     3.141e+01  7.022e+00   4.473 8.54e-06 ***
TotalBsmtSF    5.005e+00  5.277e+00   0.948 0.343173
GrLivArea      3.341e+01  2.794e+01   1.196 0.232009
HalfBath      -1.098e+03  3.321e+03  -0.331 0.740945
Fireplaces     4.372e+03  2.189e+03   1.998 0.046020 *
WoodDeckSF     2.144e+01  1.002e+01   2.139 0.032662 *
ScreenPorch    5.805e+01  2.041e+01   2.844 0.004532 **
YrSold        -2.474e+02  8.458e+02  -0.293 0.769917
MSSubClass    -2.001e+02  3.451e+01  -5.797 8.84e-09 ***
OverallCond    5.239e+03  1.368e+03   3.830 0.000135 ***
BsmtFinSF1     1.235e+01  3.949e+00   3.128 0.001810 **
X1stFlrSF      1.257e+01  2.862e+01   0.439 0.660701
BsmtFullBath   9.043e+03  3.198e+03   2.828 0.004776 **
BedroomAbvGr  -1.022e+04  2.155e+03  -4.742 2.40e-06 ***
GarageYrBlt   -4.728e+01  9.106e+01  -0.519 0.603742
OpenPorchSF   -2.252e+00  1.949e+01  -0.116 0.907998
PoolArea      -6.052e+01  2.990e+01  -2.024 0.043204 *
LotFrontage   -1.160e+02  6.126e+01  -1.894 0.058503 .
YearBuilt      3.164e+02  8.766e+01   3.610 0.000321 ***
BsmtFinSF2     3.337e+00  7.649e+00   0.436 0.662704
X2ndFlrSF      1.322e+01  2.804e+01   0.471 0.637433
BsmtHalfBath   2.465e+03  5.073e+03   0.486 0.627135
KitchenAbvGr  -2.202e+04  6.710e+03  -3.282 0.001063 **
GarageCars     1.685e+04  3.491e+03   4.827 1.58e-06 ***
EnclosedPorch  7.295e+00  2.062e+01   0.354 0.723590
MiscVal       -3.761e+00  6.960e+00  -0.540 0.589016
LotArea        5.422e-01  1.575e-01   3.442 0.000599 ***
YearRemodAdd   1.194e+02  8.668e+01   1.378 0.168607
```

```
BsmtUnfSF             NA         NA      NA       NA
LowQualFinSF          NA         NA      NA       NA
FullBath       5.433e+03  3.531e+03   1.539 0.124182
TotRmsAbvGrd   5.464e+03  1.487e+03   3.674 0.000251 ***
GarageArea     6.274e+00  1.213e+01   0.517 0.605002
X3SsnPorch     3.349e+01  3.758e+01   0.891 0.373163
MoSold        -2.217e+02  4.229e+02  -0.524 0.600188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36800 on 1085 degrees of freedom
  (339 observations deleted due to missingness)
Multiple R-squared:  0.8096,    Adjusted R-squared:  0.8034
F-statistic: 131.8 on 35 and 1085 DF,  p-value: < 2.2e-16
```

1. Is there a relationship between the predictors and the response?

   Yes, there is an overall relationship between the predictors (being all the variables in Ames) and the actual Sale Price of the homes in Ames.

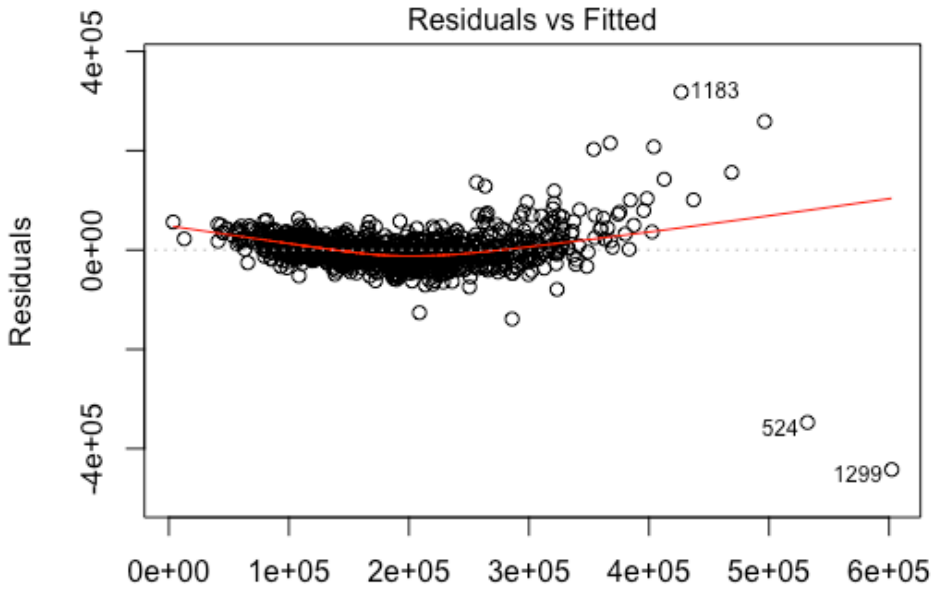
2. Which predictors appear to have a statistically significant relationship to the response?

   The predictors Fireplaces,WoodDeckSF,ScreenPorch,OverallCond, BsmtFinSF1,BsmtFullBath, PoolArea, LotFrontage, YearBuilt, KitchenAbvGr, LotArea, and TotRmsAbvGrd because they have p-values below the statistically significant level of 0.05.

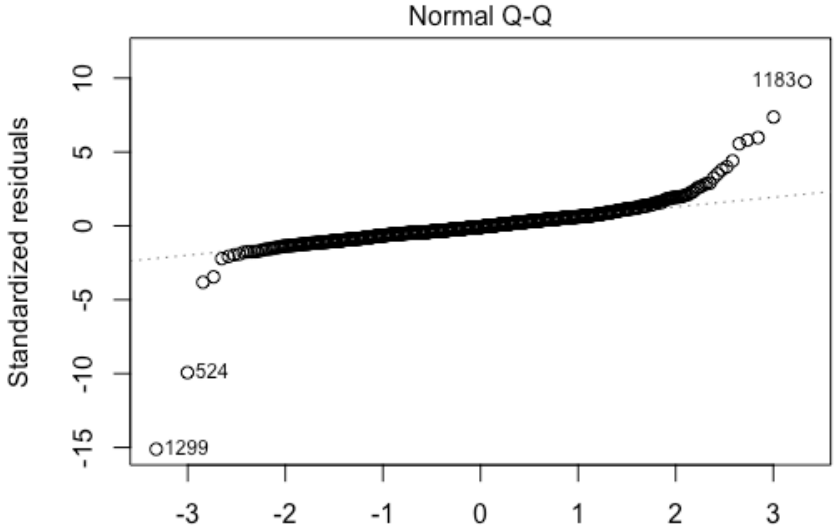
3. What does the coefficient for the year variable suggest?

   The coefficient for the year built variable (3.164e+02) suggests that for every 1 year increase in the year built of the house, the sale price goes up by 3.164e+02.

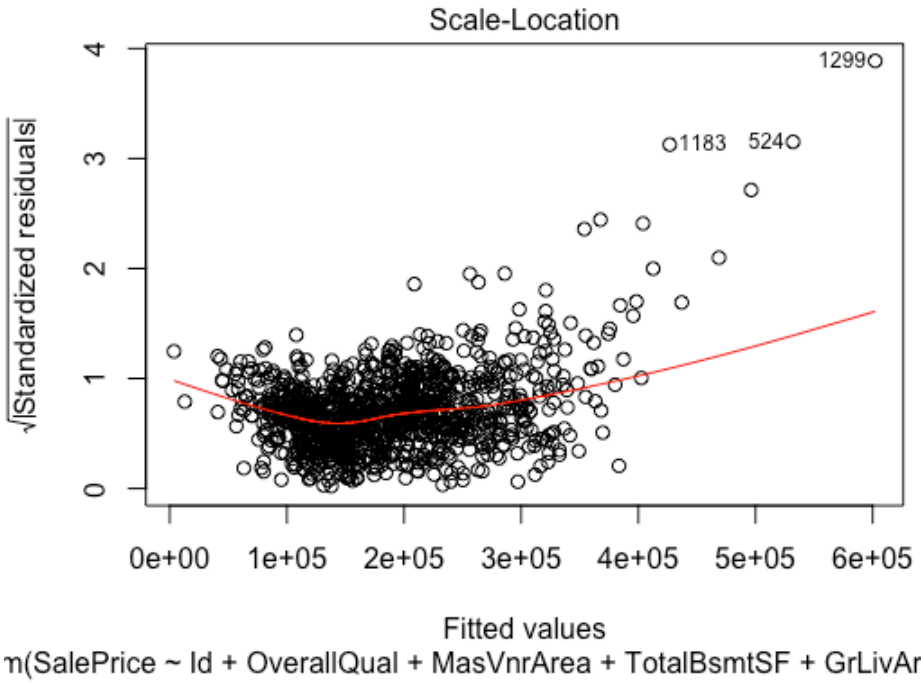
Question 3 – Produce Diagnostic plots of the linear regression fit:

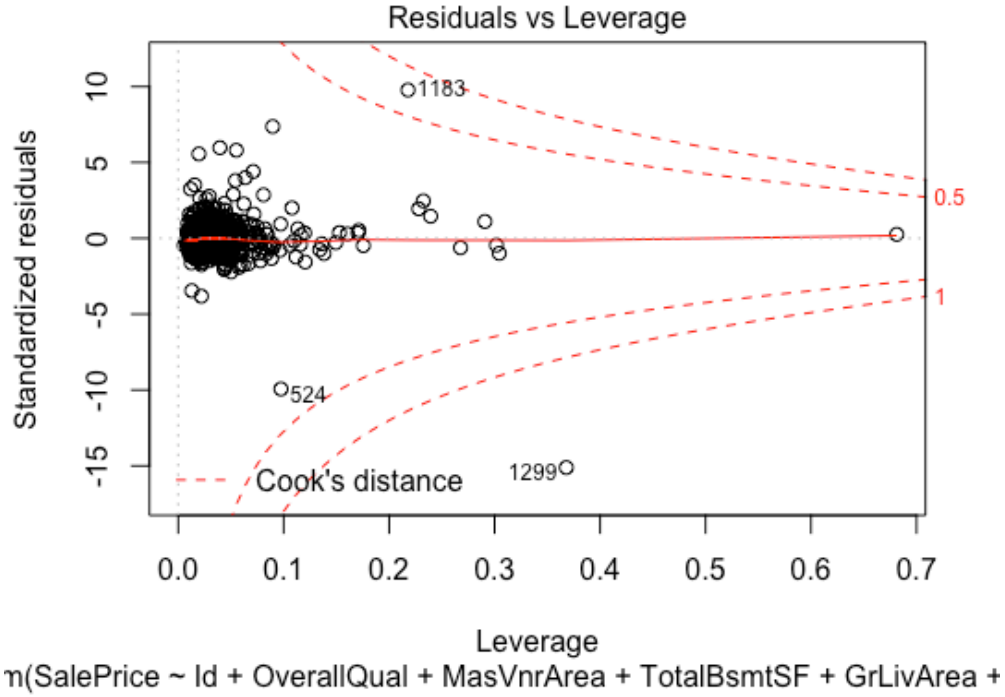


m(SalePrice ~ Id + OverallQual + MasVnrArea + TotalBsmtSF + GrLivArea +



m(SalePrice ~ Id + OverallQual + MasVnrArea + TotalBsmtSF + GrLivArea +





1. Comment on any problems you see with the fit.

The fitted values are in clusters and that could prove to be problematic for predicting sale price in the future. As well as potential outliers that could skew our analysis.

2. Do the residual plots suggest any unusually large outliers?

The residuals suggest two unusually large outliers in the dataset. One at 524 and one at 1299 in the plot.

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

Yes, the cluster of points on the leverage plot are around 0.0 to 0.1, with a few reaching 0.3. However, there is one point that reaches a leverage of 0.7.

Question 4 – Create interactions:

```
Call:
lm(formula = SalePrice ~ Id + OverallQual * OverallCond + MasVnrArea +
    TotalBsmtSF + GrLivArea + HalfBath + Fireplaces + WoodDeckSF +
    ScreenPorch + MSSubClass + BsmtFinSF1 + X1stFlrSF + BsmtFullBath +
    BedroomAbvGr + GarageYrBlt * GarageCars + OpenPorchSF + PoolArea:LotFrontage +
    YearBuilt + BsmtFinSF2 + X2ndFlrSF + BsmtHalfBath + KitchenAbvGr +
    EnclosedPorch + MiscVal + LotArea + YearRemodAdd * YrSold +
    BsmtUnfSF + LowQualFinSF + FullBath + TotRmsAbvGrd + GarageArea +
    X3SsnPorch + MoSold, data = Ames)

Residuals:
    Min      1Q  Median      3Q     Max
-380021  -15013   -2414   13426  378404
```

```
Coefficients: (2 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.911e+08  1.504e+08   1.270 0.204377
Id                      -4.998e-01  2.534e+00  -0.197 0.843686
OverallQual              1.573e+04  4.754e+03   3.308 0.000971 ***
OverallCond              5.738e+03  4.799e+03   1.196 0.232133
MasVnrArea               2.284e+01  6.743e+00   3.388 0.000730 ***
TotalBsmtSF              6.476e+00  5.121e+00   1.265 0.206262
GrLivArea                5.886e+01  2.673e+01   2.202 0.027894 *
HalfBath                 3.369e+02  3.184e+03   0.106 0.915742
Fireplaces               4.367e+03  2.084e+03   2.095 0.036396 *
WoodDeckSF               1.774e+01  9.540e+00   1.859 0.063275 .
ScreenPorch              6.577e+01  1.948e+01   3.377 0.000759 ***
MSSubClass              -1.425e+02  3.142e+01  -4.534 6.42e-06 ***
BsmtFinSF1               1.859e+01  3.828e+00   4.856 1.38e-06 ***
X1stFlrSF               -4.654e+00  2.730e+01  -0.170 0.864691
BsmtFullBath             6.856e+03  3.063e+03   2.238 0.025424 *
BedroomAbvGr            -8.679e+03  2.061e+03  -4.211 2.75e-05 ***
GarageYrBlt             -9.483e+02  1.483e+02  -6.392 2.42e-10 ***
GarageCars              -1.138e+06  1.546e+05  -7.364 3.53e-13 ***
OpenPorchSF             -6.503e+00  1.862e+01  -0.349 0.727005
YearBuilt                2.983e+02  8.450e+01   3.531 0.000432 ***
BsmtFinSF2               8.159e+00  7.306e+00   1.117 0.264346
X2ndFlrSF               -3.896e+00  2.674e+01  -0.146 0.884171
BsmtHalfBath             3.206e+03  4.834e+03   0.663 0.507301
KitchenAbvGr            -1.701e+04  6.417e+03  -2.650 0.008157 **
EnclosedPorch            1.438e+01  1.973e+01   0.729 0.466298
MiscVal                 -1.984e+00  6.587e+00  -0.301 0.763353
LotArea                  6.056e-01  1.459e-01   4.151 3.58e-05 ***
YearRemodAdd            -9.529e+04  7.580e+04  -1.257 0.208947
YrSold                  -9.463e+04  7.493e+04  -1.263 0.206883
BsmtUnfSF                       NA         NA      NA       NA
LowQualFinSF                    NA         NA      NA       NA
FullBath                 5.411e+03  3.374e+03   1.604 0.109082
TotRmsAbvGrd             2.490e+03  1.446e+03   1.723 0.085258 .
GarageArea               4.326e+00  1.158e+01   0.373 0.708877
X3SsnPorch               3.378e+01  3.574e+01   0.945 0.344810
MoSold                  -2.916e+02  4.030e+02  -0.724 0.469419
OverallQual:OverallCond  1.063e+01  8.178e+02   0.013 0.989627
GarageYrBlt:GarageCars   5.843e+02  7.838e+01   7.455 1.84e-13 ***
PoolArea:LotFrontage    -1.463e+00  1.964e-01  -7.449 1.91e-13 ***
YearRemodAdd:YrSold      4.750e+01  3.775e+01   1.258 0.208551
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 35050 on 1083 degrees of freedom
  (339 observations deleted due to missingness)
Multiple R-squared:  0.8276,    Adjusted R-squared:  0.8217
F-statistic: 140.5 on 37 and 1083 DF,  p-value: < 2.2e-16
```

1. Do any interactions appear to be statistically significant?

The interaction between garage year built and garage cars, as well as the interaction between pool area and lot frontage are statistically significant, with p-values below 0.05.

Question 5 – Transformations of variables:

1. Do any of these make sense to include in a model of SalePrice? Comment on your findings.

The log transformation produces some NaN values that aren't very helpful in interpreting the data. However, I believe taking the log of the data helps with the readability (in terms of axis scale) of the Ames dataset. On the other hand, transforming the data based on either squaring it or squaring it, doesn't have much of an impact on the relationship between variables, but would have an impact on the scale.

Bonus:

1. How might we build a model to estimate the elasticity of demand from this dataset?

We could possibly estimate an elasticity of demand model for this dataset by making a linear regression model, which is a representation of the linear relationship between a dependent variable and one or more independent variables. Then, after doing so, we could interpret the coefficients as such: x,y, and b. Using the x (independent) and y (dependent) values, we could find elasticity of demand through it equaling b (coefficient of x) multiplied by (x/y).