Lab 2: Linear Regression and Simple Analyses

Exercise 1:

```
ameslist <-
read.table("https://msudataanalytics.github.io/SSC442/Labs/data/ames.csv",
           header = TRUE,
           sep = ",")
```

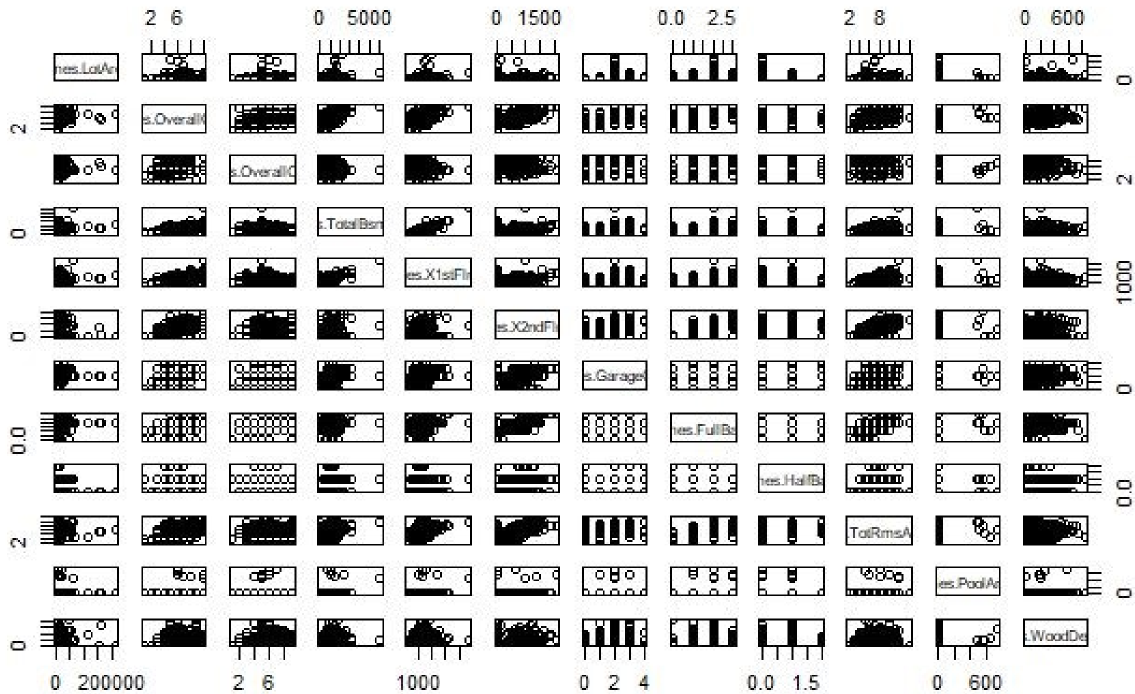
Question1

```
amesmapp <- mapply(is.integer,ameslist)
dropped_amesmapp <- subset(amesmapp, amesmapp == TRUE)
Ames <- subset(ameslist, select = amesmapp)
```

Question2

```
set1 <- data.frame(Ames$LotArea, Ames$OverallQual, Ames$OverallCond,
Ames$TotalBsmtSF, Ames$X1stFlrSF, Ames$X2ndFlrSF, Ames$GarageCars,
Ames$FullBath, Ames$HalfBath, Ames$TotRmsAbvGrd, Ames$PoolArea,
Ames$WoodDeckSF)
pairs(set1)
```

We think these 12 variables shown below are correlated with SalePrice, and these variables had been chosen for scatterplot:
LotArea, OverallQual, OverallCond, TotalBsmntSF, X1stFlrSF, X2ndFlrSf, GarageCars, FullBath, HalfBath, TotRmsAbvGrd, PoolArea, WoodDeckSf



Question3
cor(set1, Ames$SalePrice)

Results:
Ames.LotArea 0.26384335
Ames.OverallQual 0.79098160
Ames.OverallCond -0.07785589
Ames.TotalBsmtSF 0.61358055
Ames.X1stFlrSF 0.60585218
Ames.X2ndFlrSF 0.31933380
Ames.GarageCars 0.64040920
Ames.FullBath 0.56066376
Ames.HalfBath 0.28410768
Ames.TotRmsAbvGrd 0.53372316
Ames.PoolArea 0.09240355
Ames.WoodDeckSF 0.32441344

Based on what we learned, those variables, which have correlation coefficient above 0.5, are highly correlated with SalePrice. And according to the result, these variables' correlation coefficient above 0.5: OverallQual, TotalBsmtSF, X1stFirSF, GarageCars. FullBath, TotRmsAbvGrd. Therefore, we can conclude that OverallQual, TotalBsmtSF, X1stFirSF, GarageCars. FullBath, TotRmsAbvGrd match our prior belief, but others do not match due to their low correlation coefficients (lower than 0.5).

Question 4:

test <- order(Ames$GrLivArea, decreasing = TRUE)[1:2]
Ames$GrLivArea[1299] # Index of outlier
set1[1299,1:12]
The most noticeable outlier on this graph is the largest of the houses in the dataset but is not one of the most expensive. This is because it has an overall condition score of 5 and only has 2 ½ bathrooms with a 2 car garage.

Exercise 2:

attach(ameslist) # makes the function the data that is being used

lm.fit = lm(SalePrice ~ GrLivArea)
summary(lm.fit)
plot(lm.fit)

lm.fit = lm(SalePrice ~ GrLivArea + LotArea)
plot(lm.fit)

garagevalue = lm(SalePrice ~ GarageTemp)

allvariables = lm(SalePrice ~ ., ameslist)

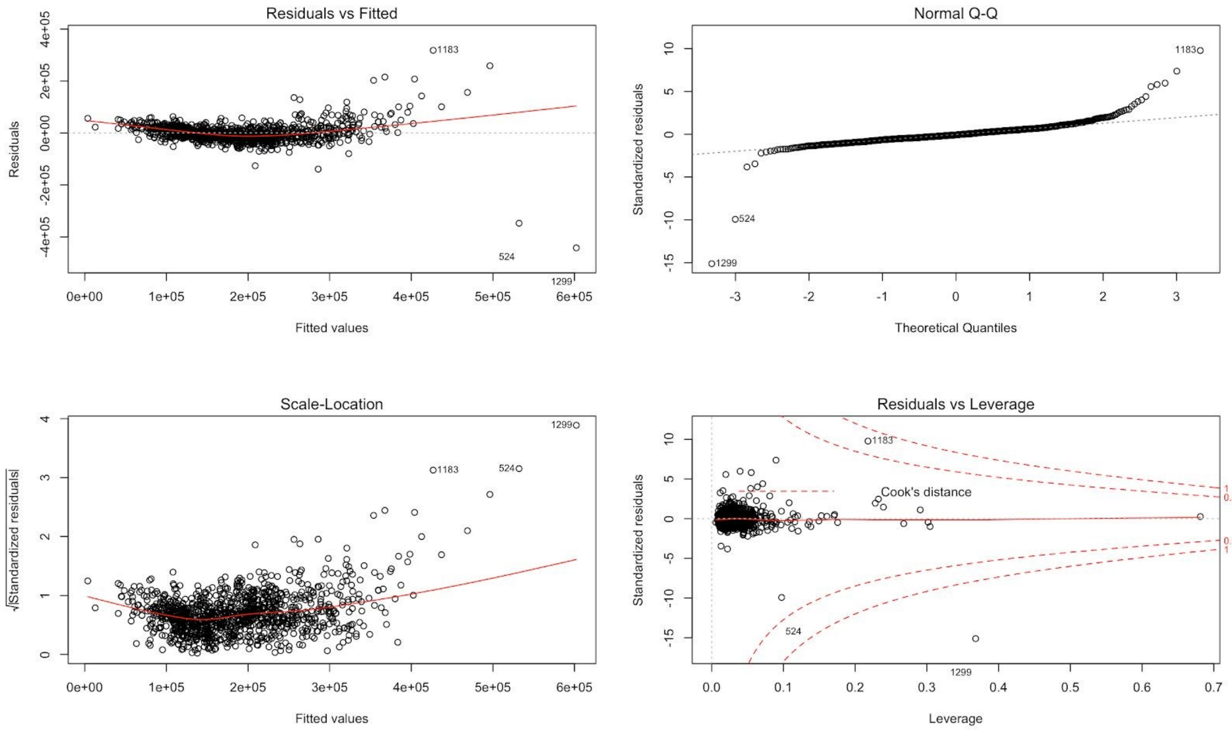
summary(allvariables)

```
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    
MSSubClass    -2.001e+02  3.451e+01  -5.797 8.84e-09 ***
LotFrontage   -1.160e+02  6.126e+01  -1.894 0.058503 .  
LotArea        5.422e-01  1.575e-01   3.442 0.000599 ***
OverallQual    1.866e+04  1.482e+03  12.592  < 2e-16 ***
OverallCond    5.239e+03  1.368e+03   3.830 0.000135 ***
YearBuilt      3.164e+02  8.766e+01   3.610 0.000321 ***
YearRemodAdd   1.194e+02  8.668e+01   1.378 0.168607    
MasVnrArea     3.141e+01  7.022e+00   4.473 8.54e-06 ***
BsmtFinSF1     1.736e+01  5.838e+00   2.973 0.003014 ** 
BsmtFinSF2     8.342e+00  8.766e+00   0.952 0.341532    
BsmtUnfSF      5.005e+00  5.277e+00   0.948 0.343173    
TotalBsmtSF           NA         NA      NA       NA    
X1stFlrSF      4.597e+01  7.360e+00   6.246 6.02e-10 ***
X2ndFlrSF      4.663e+01  6.102e+00   7.641 4.72e-14 ***
LowQualFinSF   3.341e+01  2.794e+01   1.196 0.232009    
GrLivArea             NA         NA      NA       NA    
BsmtFullBath   9.043e+03  3.198e+03   2.828 0.004776 ** 
BsmtHalfBath   2.465e+03  5.073e+03   0.486 0.627135    
FullBath       5.433e+03  3.531e+03   1.539 0.124182    
HalfBath      -1.098e+03  3.321e+03  -0.331 0.740945    
BedroomAbvGr  -1.022e+04  2.155e+03  -4.742 2.40e-06 ***
```

```
KitchenAbvGr  -2.202e+04  6.710e+03  -3.282 0.001063 **
TotRmsAbvGrd   5.464e+03  1.487e+03   3.674 0.000251 ***
Fireplaces     4.372e+03  2.189e+03   1.998 0.046020 *
GarageYrBlt   -4.728e+01  9.106e+01  -0.519 0.603742
GarageCars     1.685e+04  3.491e+03   4.827 1.58e-06 ***
GarageArea     6.274e+00  1.213e+01   0.517 0.605002
WoodDeckSF     2.144e+01  1.002e+01   2.139 0.032662 *
OpenPorchSF   -2.252e+00  1.949e+01  -0.116 0.907998
EnclosedPorch  7.295e+00  2.062e+01   0.354 0.723590
X3SsnPorch     3.349e+01  3.758e+01   0.891 0.373163
ScreenPorch    5.805e+01  2.041e+01   2.844 0.004532 **
PoolArea      -6.052e+01  2.990e+01  -2.024 0.043204 *
MiscVal       -3.761e+00  6.960e+00  -0.540 0.589016
MoSold        -2.217e+02  4.229e+02  -0.524 0.600188
YrSold        -2.474e+02  8.458e+02  -0.293 0.769917
---
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
```

2. When you use the lm() function of the SalePrice on all of the variables, the coefficients of each of them are very small in comparison to when there is only one predictor. At a 5% significance level, MSSubclass, LotArea, OverallQual, OverallCond, YearBuilt, MassVnrArea, BsmtFinFS1, X1stFlrSF, X2ndFlrSF, BsmtFullBath, BedroomAbvGrd, KitchenAbvGrd, TotRmsAbvGrd, Fireplaces, GarageCars, WoodDeckSF, ScreenPorch and PoolArea are the only variables that are significant. The YrBuilt Coefficient is 316.4, meaning that for every year that a house is in age, a house changes value by $316.

plot(allvariables)



3. Because the residuals curve is fairly straight, it shows that there aren't very many issues with the fit. The assumption of normality is not violated either.

```
lm_ex_four = lm(SalePrice ~ OverallQual * OverallCond )
lm_ex_four1 = lm(SalePrice ~ OverallQual + OverallCond +
OverallQual:OverallCond)

plot(lm_ex_four)
plot(lm_ex_four1)

summary(lm_ex_four)
summary(lm_ex_four1)
```

4. The results for using * and : are the same. In this model, R^2 has increased which suggests this model have a better fit.

```
lm_ex_five= lm(SalePrice ~ . + log(BsmtFinSF1) + sqrt(BsmtUnfSF) +
(TotalBsmtSF^2))

plot(lm_ex_five)
```

summary(lm_ex_five)

5. It makes more sense to just include those variables normally without the transformations.