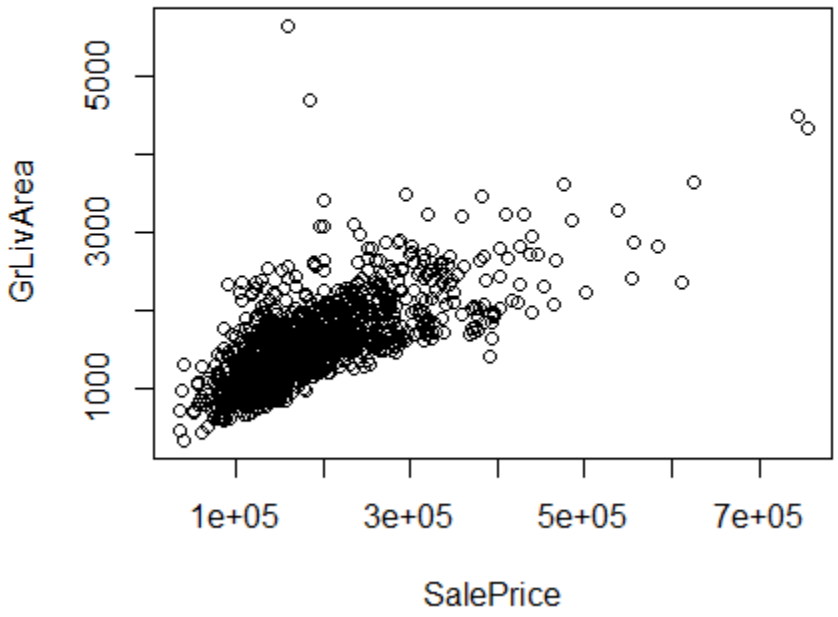
```
##Lab 2##

## Backstory and Set Up - Data Exploration and Processing ##

ameslist <- read.table("https://msudataanalytics.github.io/SSC442/Labs/data/ames.csv",

                       header = TRUE,

                       sep = ",")


ameslist <- read.table("https://msudataanalytics.github.io/SSC442/Labs/data/ames.csv",

                       header = FALSE,

                       sep = ",")

## When we specify header = FALSE, The header row is not recognized and gets moved into the 'actual' data


ameslist <- read.table("https://msudataanalytics.github.io/SSC442/Labs/data/ames.csv"

                       sep = ",")


?read.table

## Default behavior of the function: Reads a file in table format and creates a

## data frame from it, with cases corresponding to lines and variables to fields in the file.


ameslist <- read.table("https://msudataanalytics.github.io/SSC442/Labs/data/ames.csv",

                       header = TRUE ,

                       sep = ",")


names(ameslist)

#Tells us the variable names of the data


typeof(ameslist)

#Tells us that the data type is a list
```

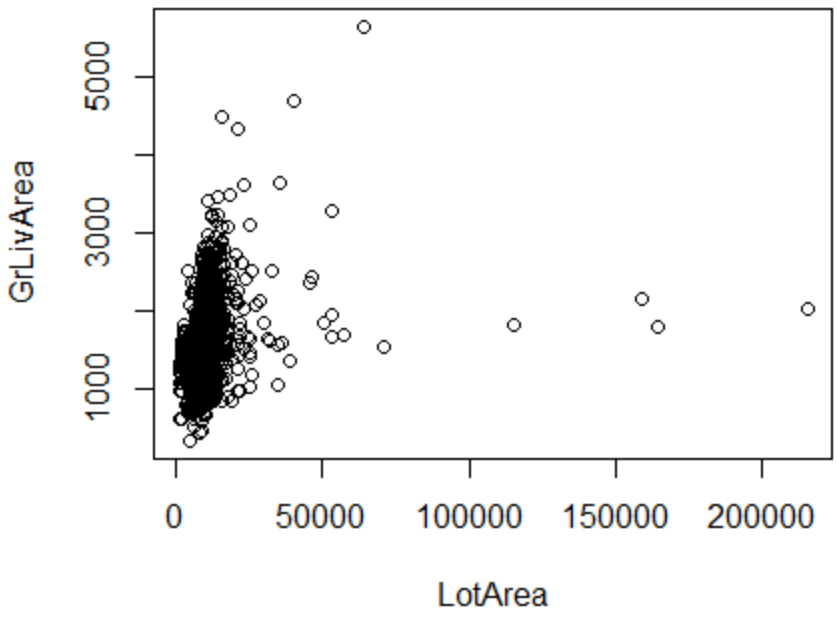
```
View(ameslist)

unique(ameslist$GarageType)

# create GarageType
GarageType = ameslist$GarageType

# create GarageTemp as model of

options(na.action='na.pass')

GarageTemp = model.matrix( ~ GarageType - 1, data=ameslist$GarageType)
help(model.matrix)

# First try it was the na.action = na.pass above, to stop model.matrix from removing NA values

ameslist <- cbind(ameslist, GarageTemp)

options(na.action='na.exclude')

ameslist$GarageOutside <- ifelse(ameslist$GarageTypeDetchd == 1 | ameslist$GarageTypeCarPort == 1,
1, 0)
GarageOutside = ameslist$GarageOutside
GarOut_NoNa = na.omit(GarageOutside)
unique(GarOut_NoNa)



options('na.action')
```

```
View(ameslist)

View(GarageTemp)


## Exercise 1 ##

# 1. Prune the data

keeps = c('Id', 'MSSubClass', 'LotFrontage', 'GarageOutside', 'LotArea', 'OverallQual', 'OverallCond',
'MasVnrArea', 'GrLivArea','BsmtFullBath','BsmtHalfBath', 'FullBath','HalfBath','BedroomAbvGr',
'KitchenAbvGr', 'TotRmsAbvGrd', 'Fireplaces','GarageCars', 'GarageArea', 'WoodDeckSF', 'PoolArea',
'MiscVal', 'SalePrice')

Ames = ameslist[keeps]


# b. Scatter plot matrix with 12 variables and SalesPrice

pairs(Ames[,11:22])


# c. create a correlation matrix with these variables

help(cor)

cor(Ames[,11:22])


# d. scatter plot between SalePrice and GrLivArea

library(tidyverse)

library(ggplot2)


p <- ggplot(data = Ames,

        mapping = aes(x = SalePrice, y = GrLivArea))

slope = cor(Ames['GrLivArea'],Ames['SalePrice'])

p + abline(slope,0) + geom_point()

help(abline)


## Building a Model ##
```

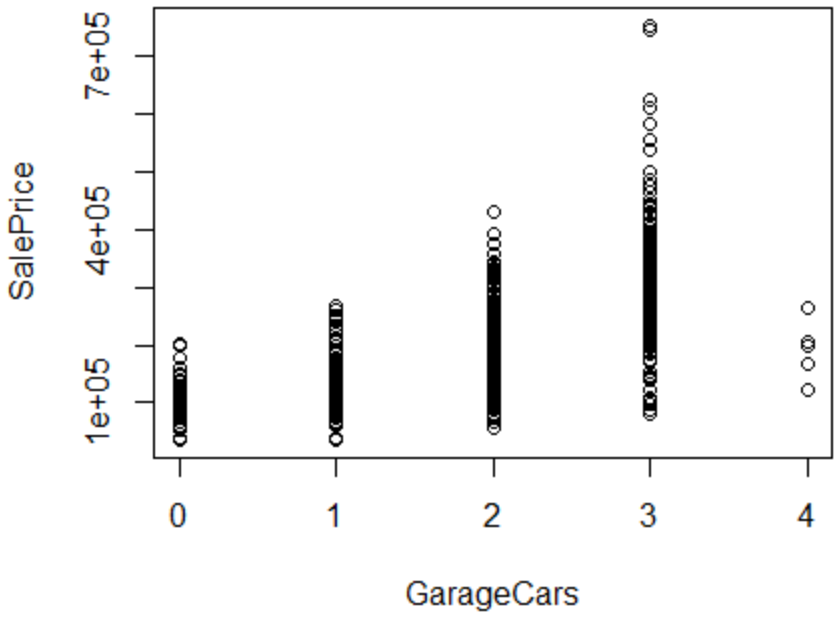
```
attach(ames)
lm.fit = lm(SalePrice ~ GrLivArea)
lm.fit

# What is GrLivArea?
# GrLivArea is the grand living area that is the total square footage.

plot(SalePrice, GrLivArea)
```



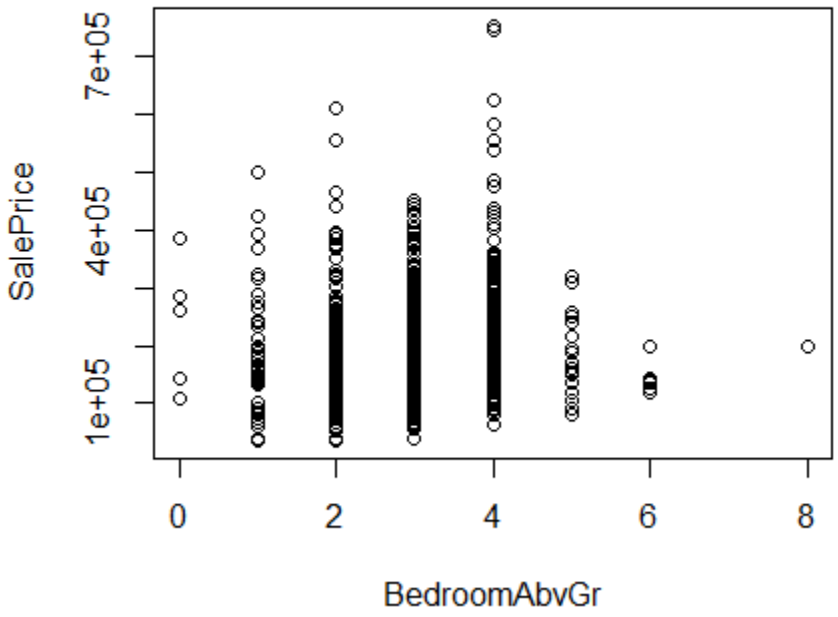

```
plot(LotArea, GrLivArea)
```



plot(GarageCars,SalePrice)



plot(BedroomAbvGr, SalePrice)



Do you suspect that some outliers have a large influence on the data?

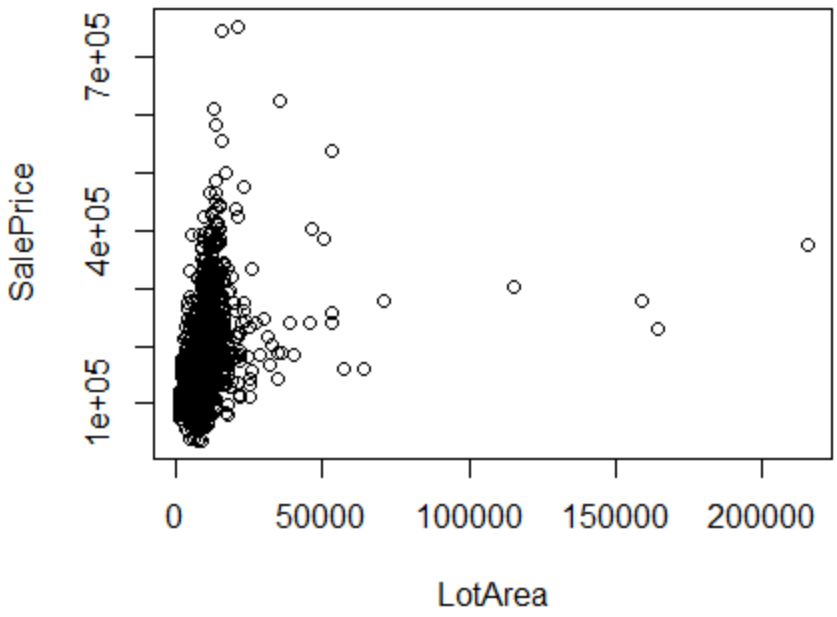
I don't think there will be outliers that have a big influence because all plots have been predictable.

But there will certainly be outliars that have some sort of effect on plots, just not big.

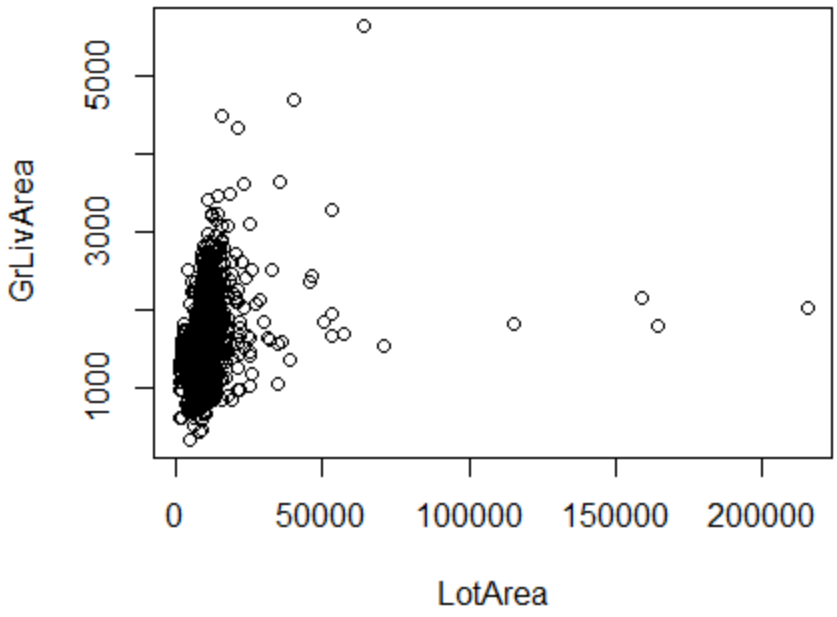
Does controlling for LotArea change the qualitative conclusions from the previous regression?

What about the quantitative results? Does the direction of the change in the quantitative results make sense to you?

plot(LotArea, SalePrice)



plot(LotArea, GrLivArea)



Yes it does to a certain extent but most LotArea are around the same range with there being several outliers in the plot.

The results make sense how they are coming out because most homes have around the same GrLivArea and LotArea.

Most the plotting on the graph has creating a general area with some homes being outliers.

exercise 2a, run regression on outside garage

attach(Ames)

garage.fit = lm(SalePrice ~ GarageOutside)

garage.fit

exercise 2b, run regression on SalePrice with all Ames

all.fit = lm(SalePrice ~ ., data = Ames[,-1])

plot(all.fit)

summary(all.fit)

#There does seem to be a relationship between the predictors and the response.

#The multiple r-squared is fairly high which indicates correlation and the estimated coefficients

#for each variable is large enough for the most part that it does seem to influence the response

#MSSubClass, GarageOutside, LotArea, OverallQual, OverallCond, MasVnrArea, GrLivArea, BsmtFullBath,

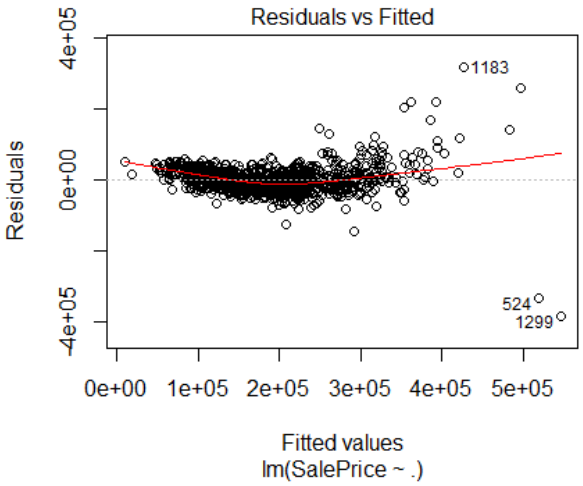
#BsmtHalfBath, FullBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, WoodDeckSF

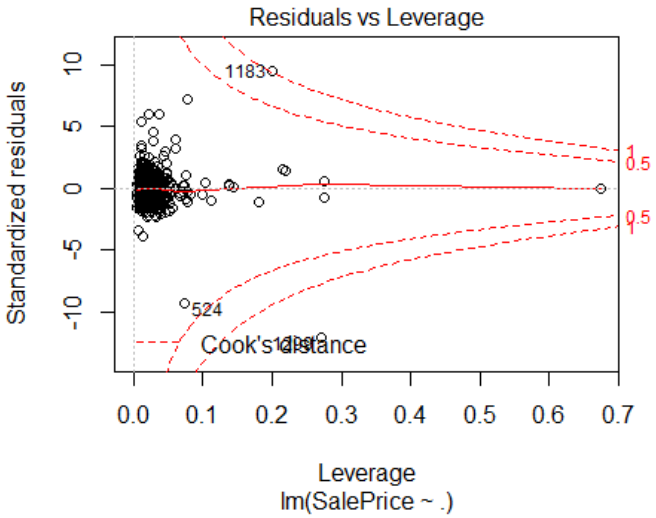
#are all important variables since they all have p-values less than .05

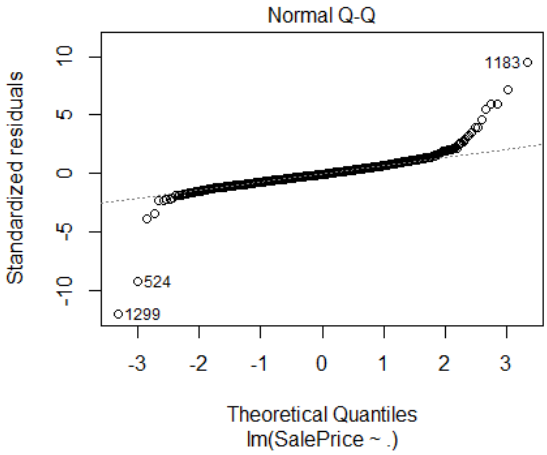
all.fit = lm(SalePrice ~ ., data = Ames[,-1])plot(all.fit)

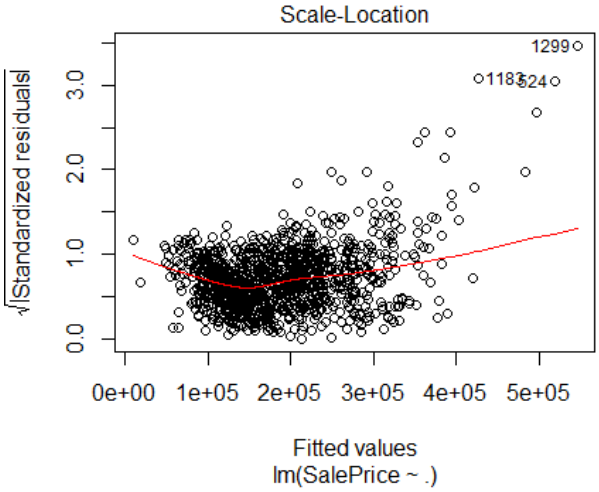
all.fit = lm(SalePrice ~ ., data = Ames[,-1])

#1183, 524, and 1299 all seem to be outliers that show up on the residual plot as well as the residuals vs. leverage one









cor(log(ameslist['LotArea']), ameslist['SalePrice'])

SalePrice

LotArea 0.3885203

> cor(log(ameslist['GrLivArea']), ameslist['SalePrice'])

SalePrice

GrLivArea 0.6951181

> cor((ameslist['GrLivArea'])**2, ameslist['SalePrice'])

SalePrice

GrLivArea 0.6522667

> cor((ameslist['GrLivArea'])**(1/2), ameslist['SalePrice'])

SalePrice

GrLivArea 0.7087645

> #log of GrLivArea has a better correlation with SalePrice than x^2 but sqrt is even better.