House Price Data Linear Regression Analysis

Ruolan Zeng(rxz171630), Zhichao Yuan(zxy180004), Yi Su(yxs173830)

0.Introduction

we use the dataset: House Prices: Advanced Regression Techniques https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

We put it on the url: https://github.com/lypf2018/HousePricesKaggle/raw/master/dataset/train.csv

There are 1460 rows and 81 columns and the last column "SalePrice" is the output variable

1.Preprocessing

1.1 Remove NA Values

```
data <- read.csv("https://github.com/lypf2018/HousePricesKaggle/raw/master/dataset/train.csv")
probNA <- function(x){
  sum(is.na(x))/nrow(data)
}
probNAcol <- apply(data, 2, probNA)
as.data.frame(probNAcol[probNAcol>0.3])
```

```
##              probNAcol[probNAcol > 0.3]
## Alley                         0.9376712
## FireplaceQu                   0.4726027
## PoolQC                        0.9952055
## Fence                         0.8075342
## MiscFeature                   0.9630137
```

Obviously, there are many NA values(percentage of NA values bigger than 30%) in above 5 columns. We decided to remove these columns then remove all other NA values.

```
data[c("Alley","FireplaceQu","PoolQC","Fence","MiscFeature")] <- NULL
data["Id"] <- NULL
data <- na.omit(data)
```

1.2 Remove Non-numeric Freatures

Since there are enough features(80 features) in our data, we decide to directly remove all non-numeric features. The numeric features also include some categorical data, which use numbers to present different categories.

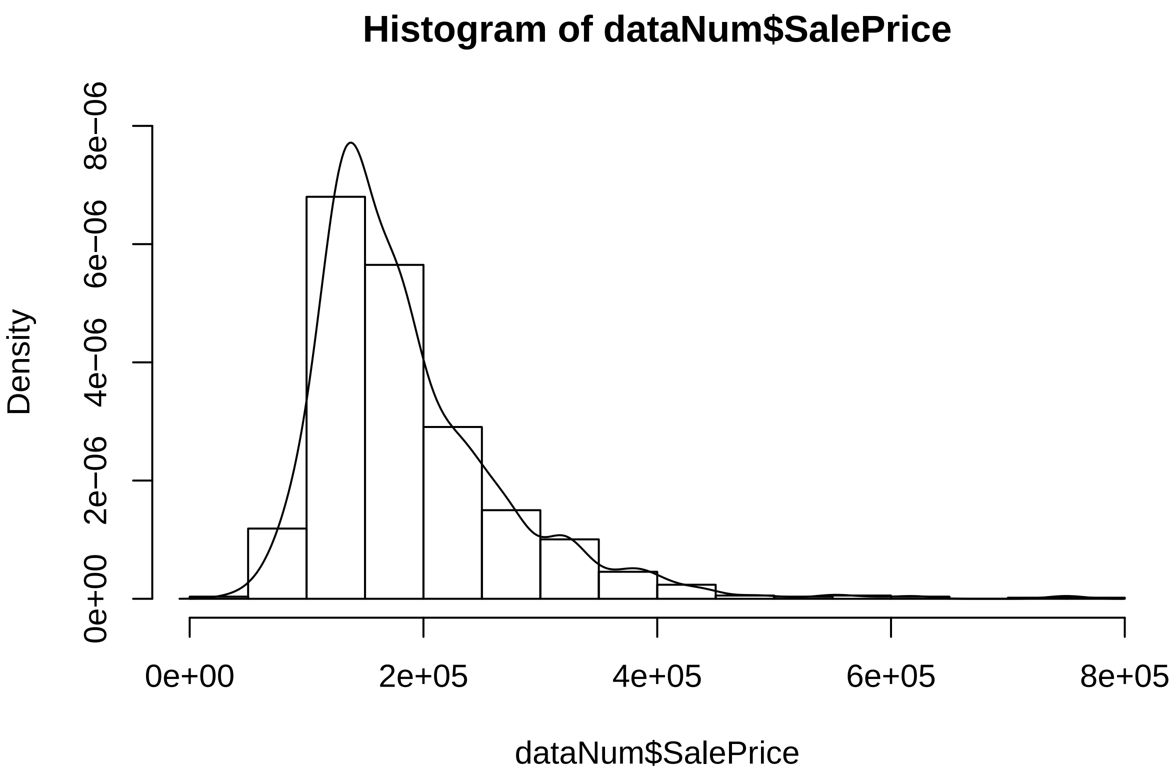
```
dataNum <- as.data.frame(data[,apply(data, 2, function(x) !any(is.na(as.numeric(x))))])
```

2. SalesPrice Analysis

```
summary(dataNum$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35311  132500  165750  187033  221000  755000
```

```
hist(dataNum$SalePrice, ylim = c(0, 8*10^-6), probability = TRUE)
lines(density(dataNum$SalePrice))
```



finding: 1. From summary of SalesPrice: Minimal house price is largger than 0, so it would not destroy our model.

2. From histogram of SalesPrice: the distribution

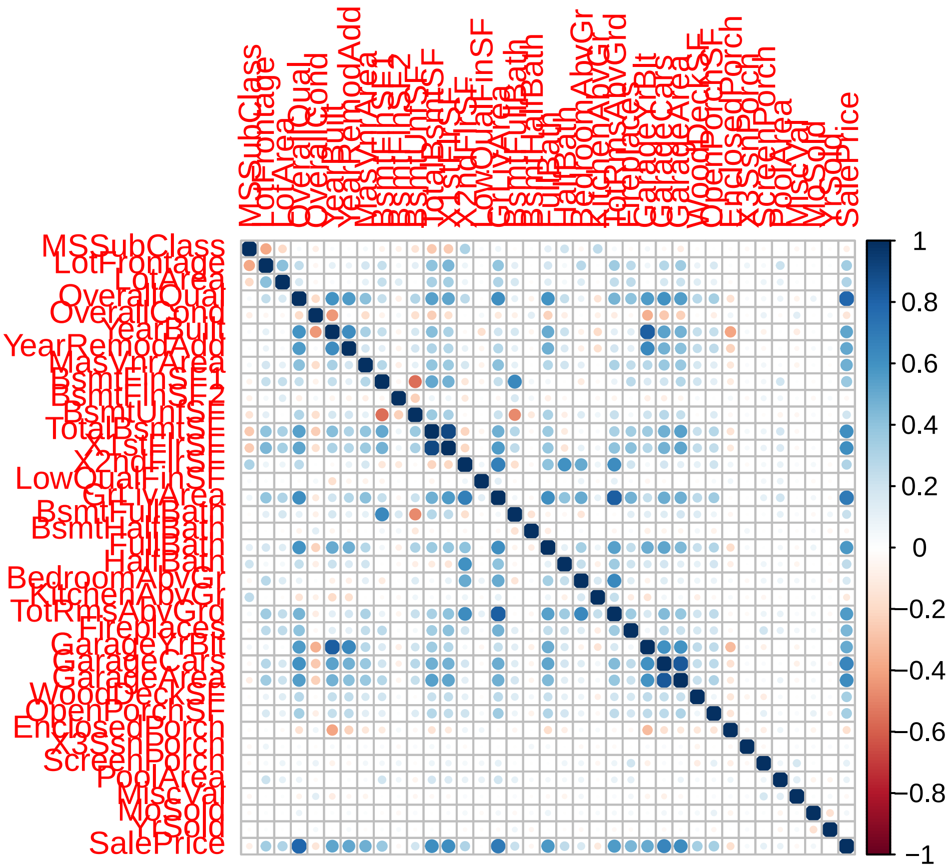
a. Deviate from the normal distribution.
b. Have appreciable positive skewness.
c. Show peakedness.

It seems that there are few extremly rich peolple bought very expensive houses, which makes the distribution has a long tail on the right, that is, positive skewness.

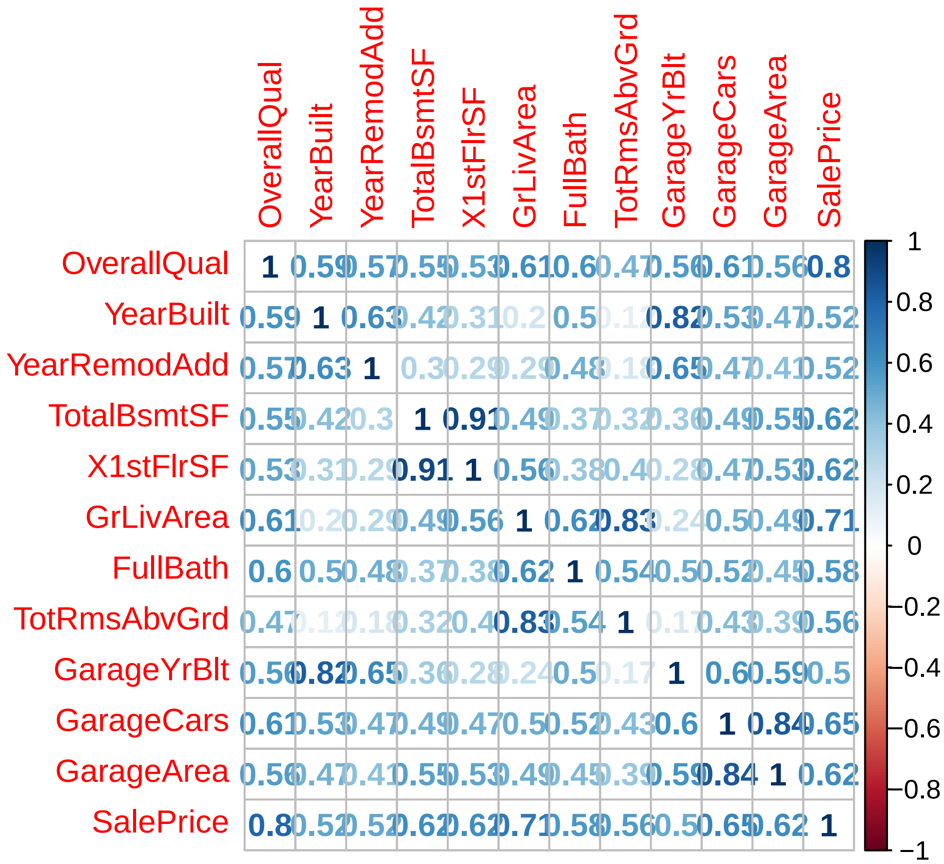
2. Correlation Analysis

2.1 Correlation Plot

```
require(MASS)
require(ISLR)
require(corrplot)
M <- cor(dataNum)
corMat <- as.data.frame(corrplot(M,method = "circle"))
```




```r
corrplot(cor(dataNum[row.names(corMat)[abs(corMat$SalePrice) > 0.50]]), method = "number")
```



finding: From the correlation plot we can see the 11 features with the strongest effect(correlation > 0.5) on SalePrice:

1.OverallQual: Rates the overall material and finish of the house. (1 very poor - 10 very excellent).

2.YearBuilt: Original construction date.

3.YearRemodAdd: Remodel date (same as construction date if no remodeling or additions).

4.TotalBsmtSF: Total square feet of basement area.

5.X1stFlrSF: First Floor square feet.

6.GrLivArea: Above grade (ground) living area square feet.

7.FullBath: Full bathrooms above grade.

8.TotRmsAbvGrd: Total rooms above grade (does not include bathrooms).

9.GarageYrBlt: Year garage was built.

10.GarageCars: Size of garage in car capacity.

11.GarageArea: Size of garage in square feet.

All of them are positive correlations.

We can see there are also some strong relations between these features. For example, correlation between TotalBsmtSF and X1stFlrSF is 0.91, correlation between GrLivArea and TotRmsAbvGrd is 0.83. That make sense since a big basement area is always together with a big area first floor, and the size of the living area will most likely be a constraint on the number of rooms above ground.

2.2 Scatter Plots

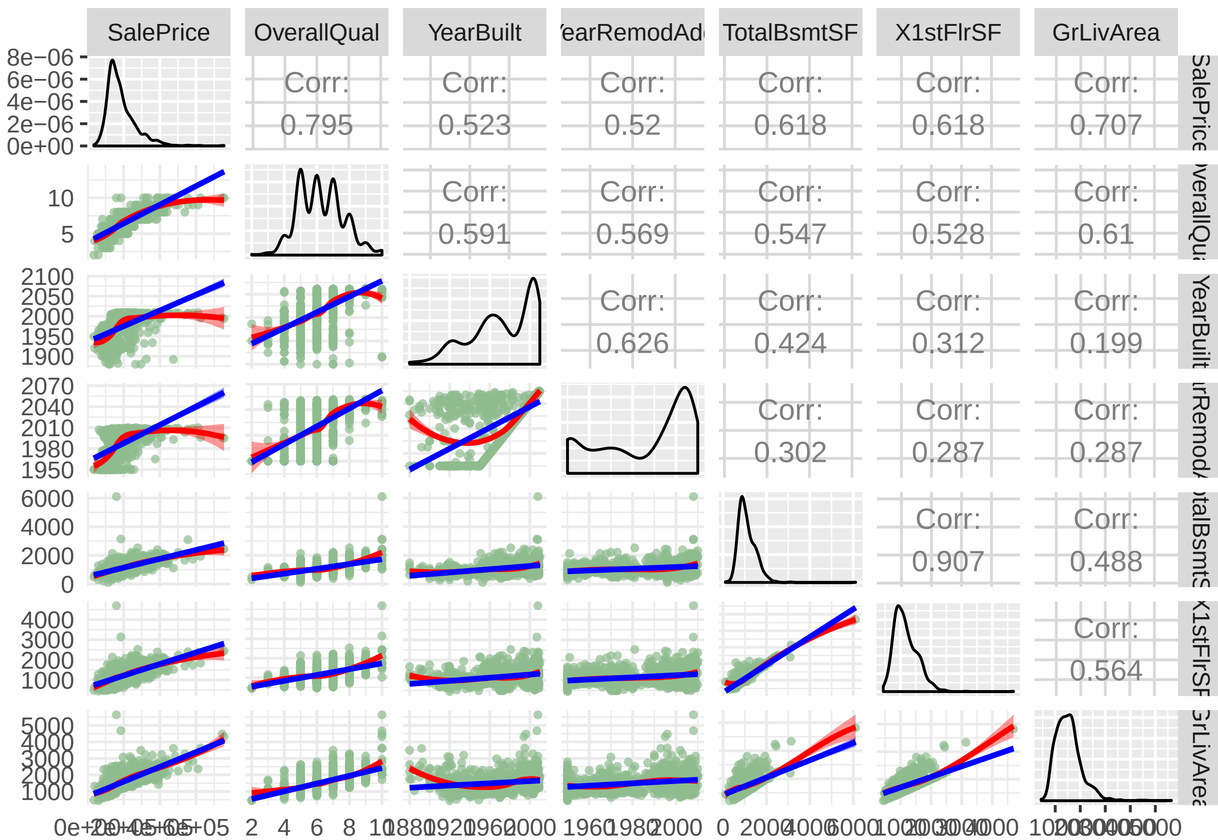
We can print a matrix of scatter plots to see what the relationships between features look like.

```r
require(GGally)
```

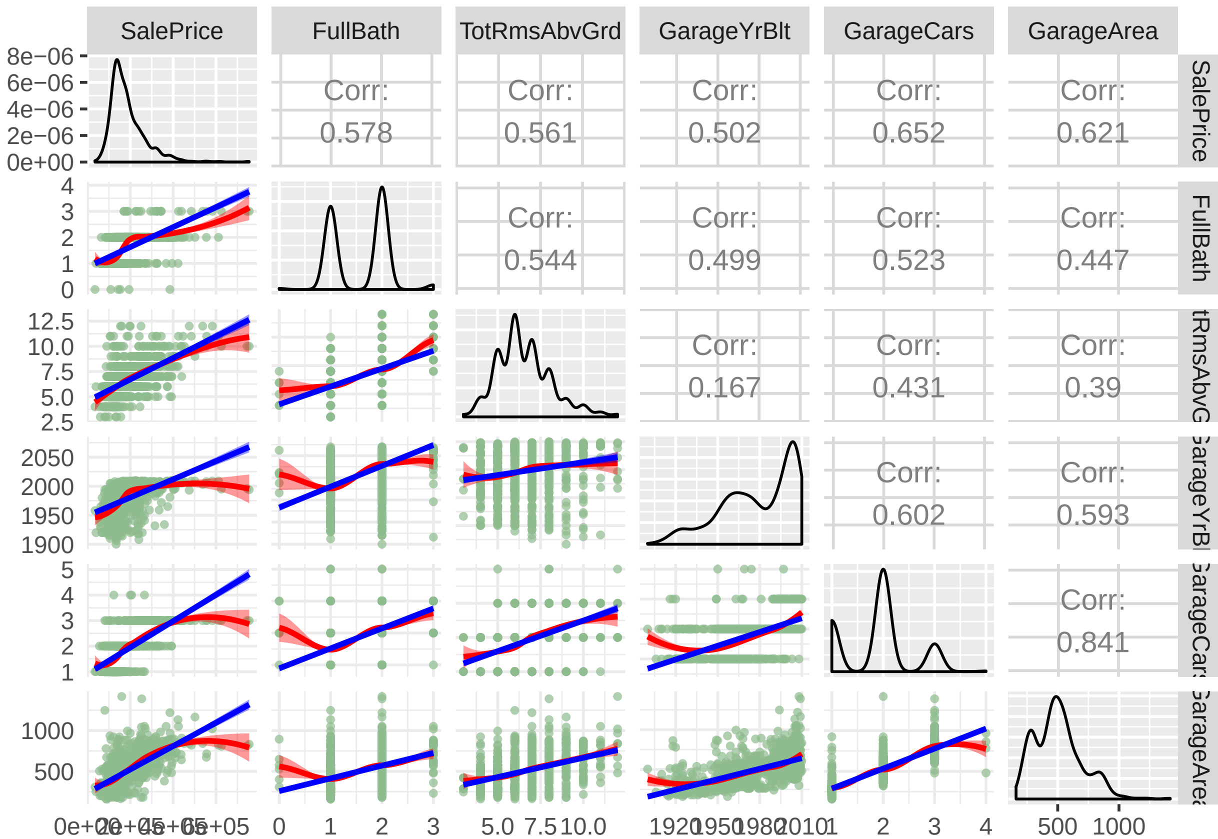
```
## Loading required package: GGally
```

```
## Loading required package: ggplot2
```

```r
corr.idx <- row.names(corMat)[abs(corMat$SalePrice) > 0.5]
lm.plt <- function(data, mapping, ...){
   plt <- ggplot(data = data, mapping = mapping) +
    geom_point(shape = 20, alpha = 0.7, color = 'darkseagreen') +
    geom_smooth(method=loess, fill="red", color="red") +
    geom_smooth(method=lm, fill="blue", color="blue") +
    theme_minimal()
  return(plt)
}
ggpairs(dataNum, corr.idx[c(12,1:6)], lower = list(continuous = lm.plt))
```




```r
ggpairs(dataNum, corr.idx[c(12,7:11)], lower = list(continuous = lm.plt))
```



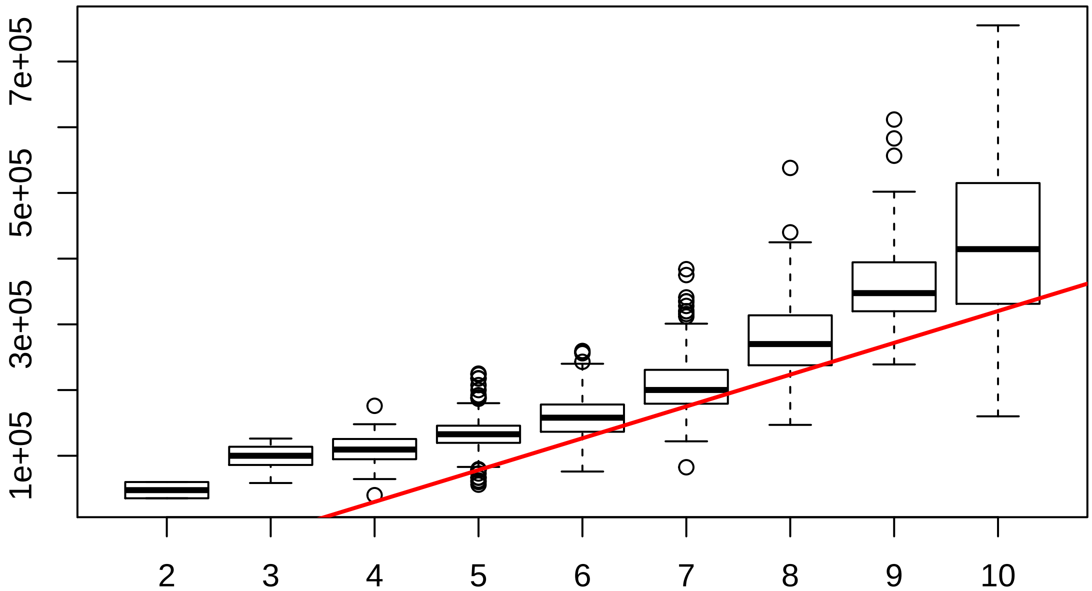
2.3 Single Feature Analysis

2.3.1 categorical feature

```
lm.fit = lm(SalePrice~OverallQual, data = dataNum)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual, data = dataNum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -208643  -31369   -1227   21325  386357
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -115355       7137  -16.16   <2e-16 ***
## OverallQual    48400       1116   43.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50420 on 1092 degrees of freedom
## Multiple R-squared:  0.6327, Adjusted R-squared:  0.6324
## F-statistic:  1881 on 1 and 1092 DF,  p-value: < 2.2e-16
```

```r
boxplot(dataNum$SalePrice~dataNum$OverallQual)
abline(lm.fit,lwd = 2, col="red")
```



finding: OverallQual:Rates the overall material and finish of the house. (1 very poor - 10 very excellent).

Frome above statistics and graph we can see:

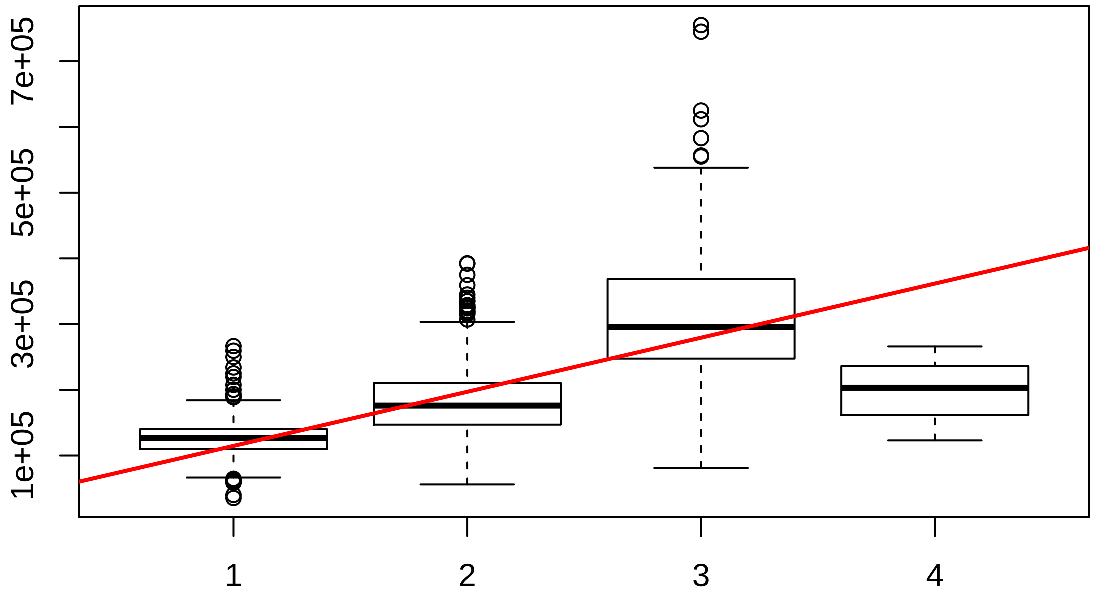
The relationship between Salesprice and OverallQual is linear relationship and OverallQual is an important factor.

When overall quality of houses are around medium quality(rate:3-7), house prices are concentrated. However, when overall quality of houses are above excellent(rate:9-10), house prices are dispersed, which means the range of house prices can be very large (more than 6e+05). This may be because some other factors like the living area, number of rooms and overall conditions.

```r
lm.fit2 = lm(SalePrice~GarageCars, data = dataNum)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = SalePrice ~ GarageCars, data = dataNum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -238662  -37060   -4546   26571  475684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    32276       5769   5.595 2.79e-08 ***
## GarageCars     82347       2897  28.424  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63080 on 1092 degrees of freedom
## Multiple R-squared:  0.4252, Adjusted R-squared:  0.4247
## F-statistic: 807.9 on 1 and 1092 DF,  p-value: < 2.2e-16
```

```r
boxplot(dataNum$SalePrice~dataNum$GarageCars)
abline(lm.fit2,lwd = 2, col="red")
```



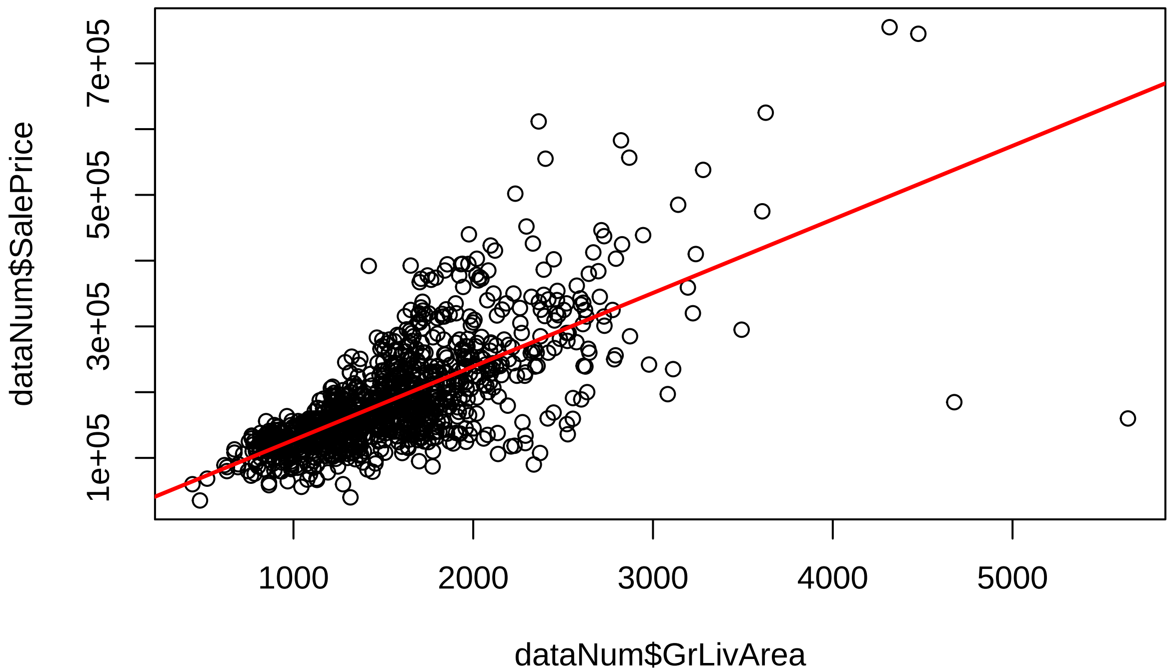
finding: When car capacity in one garage is 4, the house price decreases, which means people may think it's unnecessary to have such a big garage in a house.

2.3.2 numeric feature

```r
lm.fit3 = lm(SalePrice~GrLivArea, data = dataNum)
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea, data = dataNum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -486328  -29378   -2472   21283  331917
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15366.836   5485.443   2.801  0.00518 **
## GrLivArea     111.833      3.381  33.080  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58800 on 1092 degrees of freedom
## Multiple R-squared:  0.5005, Adjusted R-squared:  0.5001
## F-statistic:  1094 on 1 and 1092 DF,  p-value: < 2.2e-16
```

```r
plot(dataNum$GrLivArea, dataNum$SalePrice)
abline(lm.fit3,lwd = 2, col="red")
```

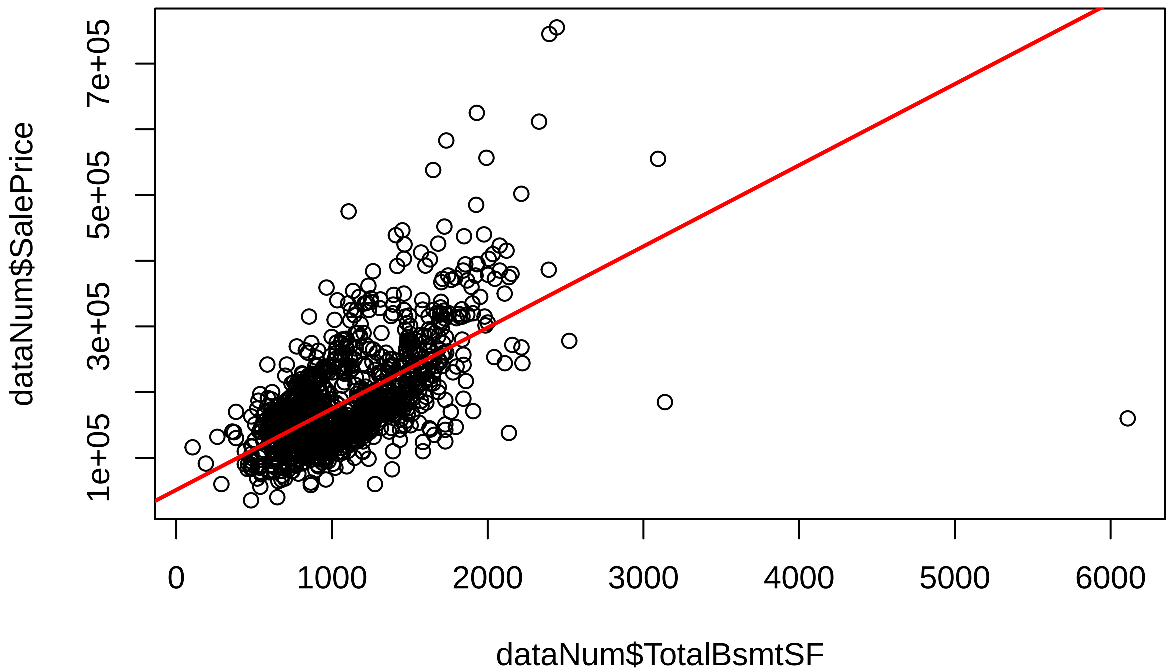


finding This distribution is almost perfect linear.

```
lm.fit4 = lm(SalePrice~TotalBsmtSF, data = dataNum)
summary(lm.fit4)
```

```
##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF, data = dataNum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -646027  -40185  -15144   34717  401874
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51192.531   5594.361   9.151   <2e-16 ***
## TotalBsmtSF   123.541      4.759  25.959   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65430 on 1092 degrees of freedom
## Multiple R-squared:  0.3816, Adjusted R-squared:  0.381
## F-statistic: 673.9 on 1 and 1092 DF,  p-value: < 2.2e-16
```

```
plot(dataNum$TotalBsmtSF, dataNum$SalePrice)
abline(lm.fit4,lwd = 2, col="red")
```



2.4 Multiple Feature Analysis

```
lm.fit5 = lm(SalePrice~OverallQual+GarageCars+GrLivArea, data = dataNum)
summary(lm.fit5)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GarageCars + GrLivArea,
##     data = dataNum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -342376  -23732   -1488   20692  291962
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.202e+05  6.046e+03  -19.88   <2e-16 ***
## OverallQual  2.921e+04  1.328e+03   22.00   <2e-16 ***
## GarageCars   2.608e+04  2.515e+03   10.37   <2e-16 ***
## GrLivArea    4.933e+01  3.162e+00   15.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42690 on 1090 degrees of freedom
## Multiple R-squared:  0.7373, Adjusted R-squared:  0.7365
## F-statistic:  1020 on 3 and 1090 DF,  p-value: < 2.2e-16
```

```
lm.fit6 = lm(SalePrice~OverallQual+GarageCars+GrLivArea+TotalBsmtSF+BsmtUnfSF+MSSubClass,
             data = dataNum)
summary(lm.fit6)
```

```
##
```

```
## Call:
## lm(formula = SalePrice ~ OverallQual + GarageCars + GrLivArea +
##     TotalBsmtSF + BsmtUnfSF + MSSubClass, data = dataNum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -482060  -18994   -1628   17414  275331
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.116e+05  5.821e+03 -19.173  < 2e-16 ***
## OverallQual  2.812e+04  1.287e+03  21.857  < 2e-16 ***
## GarageCars   2.334e+04  2.382e+03   9.799  < 2e-16 ***
## GrLivArea    4.697e+01  3.027e+00  15.515  < 2e-16 ***
## TotalBsmtSF  3.016e+01  3.921e+00   7.693 3.21e-14 ***
## BsmtUnfSF   -2.463e+01  2.919e+00  -8.437  < 2e-16 ***
## MSSubClass  -2.002e+02  3.082e+01  -6.498 1.24e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39530 on 1087 degrees of freedom
## Multiple R-squared:  0.7753, Adjusted R-squared:  0.7741
## F-statistic: 625.1 on 6 and 1087 DF,  p-value: < 2.2e-16
```

3.Compare Models

```r
anova(lm.fit,lm.fit5)
```

```
## Analysis of Variance Table
##
## Model 1: SalePrice ~ OverallQual
## Model 2: SalePrice ~ OverallQual + GarageCars + GrLivArea
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1092 2.7765e+12
## 2   1090 1.9862e+12  2 7.9034e+11 216.86 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
AIC(lm.fit,lm.fit5)
```

```
##         df      AIC
## lm.fit   3 26800.80
## lm.fit5  5 26438.33
```

finding: From above statitics we can see:

result of fuction anova : p-value is 2.2e-16

result of AIC: AIC value of model is smaller

The multiple linear regression model with three varibles(OverallQual, GarageCars, GrLivArea) is better than linear regression model with only one varible(OverallQual).

```r
anova(lm.fit5,lm.fit6)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: SalePrice ~ OverallQual + GarageCars + GrLivArea
## Model 2: SalePrice ~ OverallQual + GarageCars + GrLivArea + TotalBsmtSF +
##     BsmtUnfSF + MSSubClass
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1090 1.9862e+12
## 2   1087 1.6986e+12  3 2.8758e+11 61.345 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
AIC(lm.fit5,lm.fit6)
```

```
##         df      AIC
## lm.fit5  5 26438.33
## lm.fit6  8 26273.22
```

finding: From above statitics we can see:

result of fuction anova : p-value is 2.2e-16

result of AIC: AIC value of model is smaller

The multiple linear regression model with six varibles(OverallQual, GarageCars, GrLivArea, TotalBsmtSF, BsmtUnfSF, MSSubClass) is better than multiple linear regression model with three varibles(OverallQual, GarageCars, GrLivArea).