ECON 430 Homework 2

Gefei Zhao

2020/10/21

Question 1

The dataset train.csv contains 79 explanatory variables. The data description and csv file can be downloaded directly from kaggle(https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data). Your task, as suggested on the kaggle website, is to build a model to predict final home prices. Note, this is part of a kaggle competition which you might consider participating in later on. Before you start the parts below, identify any 10 variables of your choice and write a brief paragraph of why you selected them. These are the predictors you will use for solving the problem.

- Importing Libraries and Pre-processing Data

```
# import required libraries
library(dplyr)
library(corrplot) # correlation plot
library(car)
library(psych)
library(ggplot2)
library(olsrr)
library(leaps)
library(lmtest)
library(caret)
library(DAAG)
library(tseries)
library(stargazer) # make latex tables
library(caTools)
library(forecast)
library(multcomp)

# input data
train <- read.csv("train.csv", header = TRUE)
test <- read.csv("test.csv", header = TRUE)
```

- Choosing Variables I choose the following variables:

From my view, *space*, *location* and *physical characteristics* of houses are the most important and influential aspects of home price. As for spcace, I choose the variable of area above ground, 'GrLivArea', basement area 'TotalBsmtSF' to measure.

Besides, I use the variables 'LandContour' and 'MSZoning' to measure the influence of location of house on home price.

As for the physical characteristics of houses, I subjectively pick several important variables when choosing a house. We firstly look at the type of dwelling which is 'MSSubClass'. For a specific house, we basically

care about the overall quality and the year of the house built. Furthermore, we might focus on details of the house such as garage capacity, the number of bedroom and the heating quality.

```
# Select predictors from dataset
var=c('GrLivArea', 'TotalBsmtSF', 'LandContour', 'MSZoning', 'MSSubClass',
      'OverallQual', 'YearBuilt', 'GarageCars', 'BedroomAbvGr', 'HeatingQC', 'SalePrice')

df <- subset(train,select=var)
attach(df)  # Variables in the data frame can be accessed by simply giving their names
```

(a)

Summary Statistics

```
describe(df)
```

```
##              vars    n      mean       sd   median   trimmed      mad   min
## GrLivArea       1 1460   1515.46   525.48   1464.0   1467.67   483.33   334
## TotalBsmtSF     2 1460   1057.43   438.71    991.5   1036.70   347.67     0
## LandContour*    3 1460      3.78     0.71      4.0      4.00     0.00     1
## MSZoning*       4 1460      4.03     0.63      4.0      4.06     0.00     1
## MSSubClass      5 1460     56.90    42.30     50.0     49.15    44.48    20
## OverallQual     6 1460      6.10     1.38      6.0      6.08     1.48     1
## YearBuilt       7 1460   1971.27    30.20   1973.0   1974.13    37.06  1872
## GarageCars      8 1460      1.77     0.75      2.0      1.77     0.00     0
## BedroomAbvGr    9 1460      2.87     0.82      3.0      2.85     0.00     0
## HeatingQC*     10 1460      2.54     1.74      1.0      2.42     0.00     1
## SalePrice      11 1460 180921.20 79442.50 163000.0 170783.29 56338.80 34900
##                 max  range  skew kurtosis      se
## GrLivArea      5642   5308  1.36     4.86   13.75
## TotalBsmtSF    6110   6110  1.52    13.18   11.48
## LandContour*      4      3 -3.16     8.65    0.02
## MSZoning*         5      4 -1.73     6.25    0.02
## MSSubClass      190    170  1.40     1.56    1.11
## OverallQual      10      9  0.22     0.09    0.04
## YearBuilt      2010    138 -0.61    -0.45    0.79
## GarageCars        4      4 -0.34     0.21    0.02
## BedroomAbvGr      8      8  0.21     2.21    0.02
## HeatingQC*        5      4  0.48    -1.51    0.05
## SalePrice    755000 720100  1.88     6.50 2079.11
```

From the summary statistics, we can found that:

Sample Size: 1460

Quantitive Variables: GrLivArea, TotalBsmtSF, SalePrice, MSSubClass, OverallQual, YearBuilt, GarageCars, BedroomAbvGr

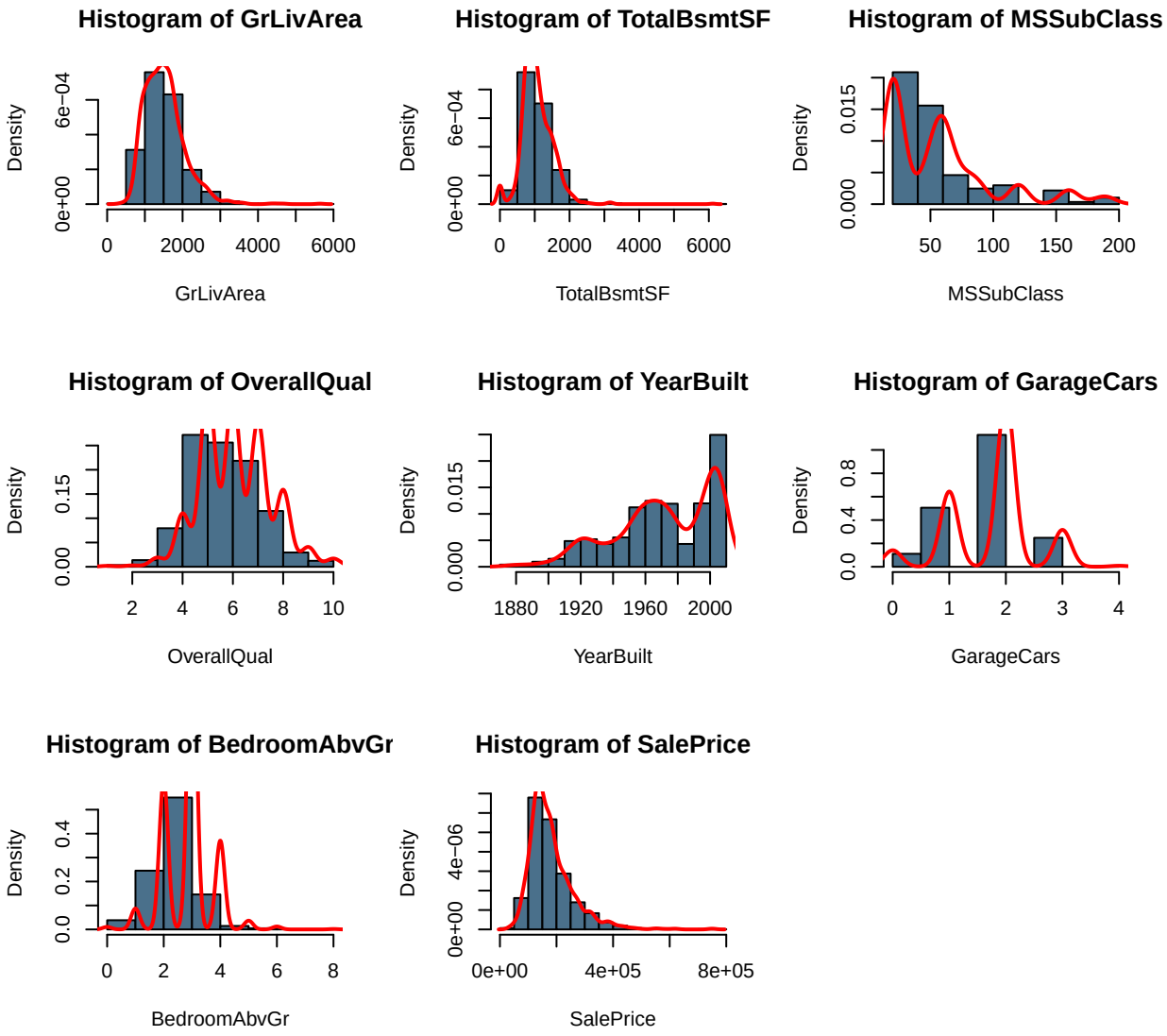
Categorical Variables: LandContour, MSZoning, HeatingQC

Although "MSSubClass, OverallQual, YearBuilt, GarageCars, BedroomAbvGr" are quantitive variables, they are discrete variables. Therefore, the density plots and qq-plots are not plausible for these variables.

Univariate Analysis

- Continuous Variables

```r
# histograms
par(mfrow = c(3,3))
n <- 1460; k <- 1 + log2(n)

hist(GrLivArea, breaks = k, col='skyblue4', ylab = 'Density', probability = T)
lines(density(GrLivArea),lwd = 2, col = 'red')
hist(TotalBsmtSF, breaks = k, col='skyblue4', ylab = 'Density', probability = T)
lines(density(TotalBsmtSF),lwd = 2, col = 'red')
hist(MSSubClass, breaks = k, col='skyblue4', ylab = 'Density', probability = T)
lines(density(MSSubClass),lwd = 2, col = 'red')
hist(OverallQual, breaks = k, col='skyblue4', ylab = 'Density', probability = T)
lines(density(OverallQual),lwd = 2, col = 'red')
hist(YearBuilt, breaks = k, col='skyblue4', ylab = 'Density', probability = T)
lines(density(YearBuilt),lwd = 2, col = 'red')
hist(GarageCars, breaks = k, col='skyblue4', ylab = 'Density', probability = T)
lines(density(GarageCars),lwd = 2, col = 'red')
hist(BedroomAbvGr, breaks = k, col='skyblue4', ylab = 'Density', probability = T)
lines(density(BedroomAbvGr),lwd = 2, col = 'red')
hist(SalePrice, breaks = k, col='skyblue4', ylab = 'Density', probability = T)
lines(density(SalePrice),lwd = 2, col = 'red')
```



Since overall quality, bedroom number, garage capacity, MSSubClass and year built are actually discrete variables, their density plots fluctuates a lot and are not normal distributed. However, we do not need to do transforms on these variables.

For else continous variables, they have the same problem that the orders of magnitude of x axis is quite large compared to the density. We should do transforms in the next questions.

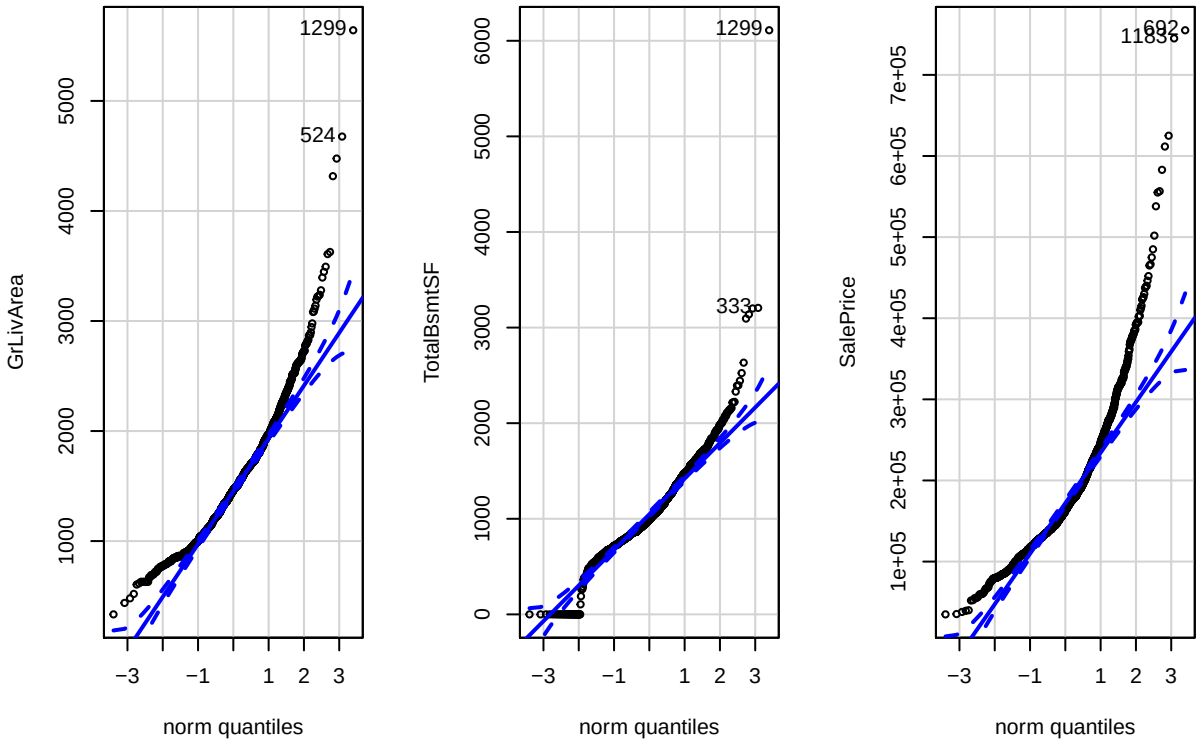
```
# qqplots
par(mfrow = c(1,3))

qqPlot(~ GrLivArea, data = df)
```

```
## [1] 1299  524
```

```
qqPlot(~ TotalBsmtSF, data = df)
```

```
## [1] 1299  333
```

```
qqPlot(~ SalePrice, data = df)
```

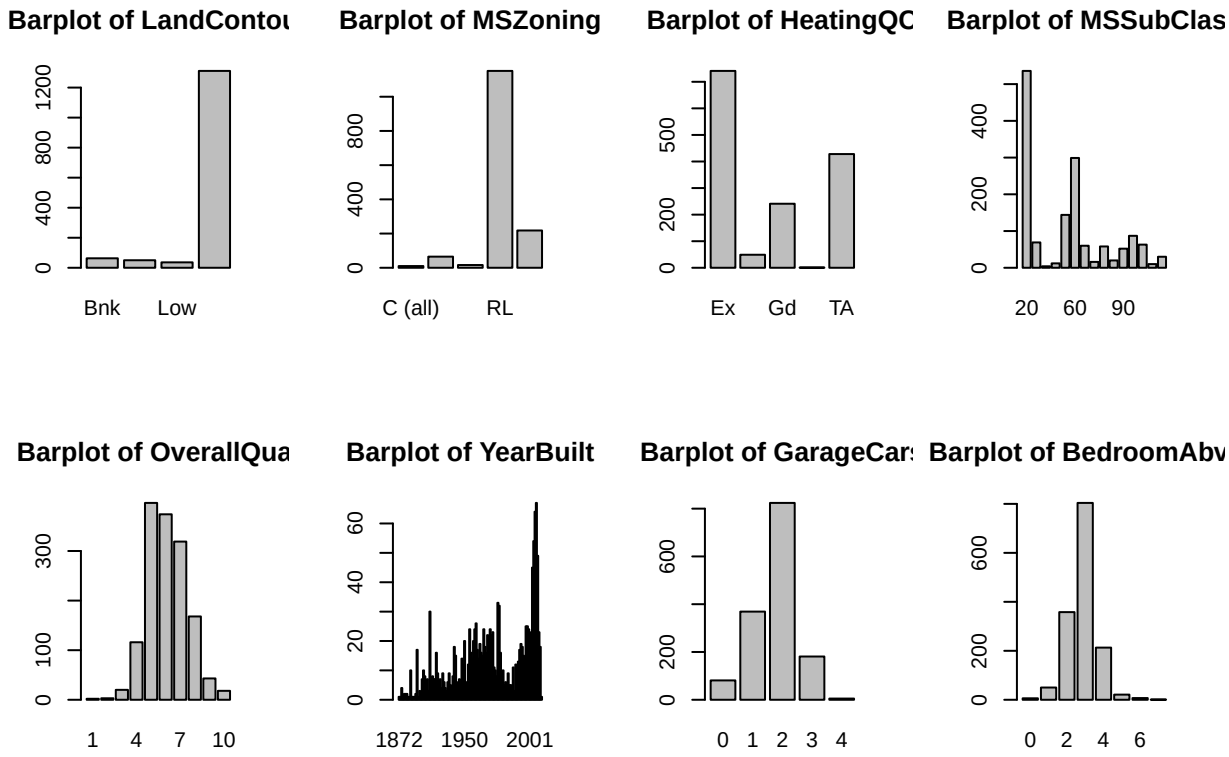



```
## [1]  692 1183
```

We found there are significant upper tails for GrLivArea and SalePrcie. And there are also some distraction in lower tail for all three variables. Therefore, all of the continous variables are not nicely normal distributed.

- Categorical Variables

```
# barplots
par(mfrow = c(2,4))
barplot(table(LandContour), main = "Barplot of LandContour")
barplot(table(MSZoning), main = "Barplot of MSZoning")
barplot(table(HeatingQC), main = "Barplot of HeatingQC")
barplot(table(MSSubClass), main = "Barplot of MSSubClass")
barplot(table(OverallQual), main = "Barplot of OverallQual")
barplot(table(YearBuilt), main = "Barplot of YearBuilt")
barplot(table(GarageCars), main = "Barplot of GarageCars")
barplot(table(BedroomAbvGr), main = "Barplot of BedroomAbvGr")
```

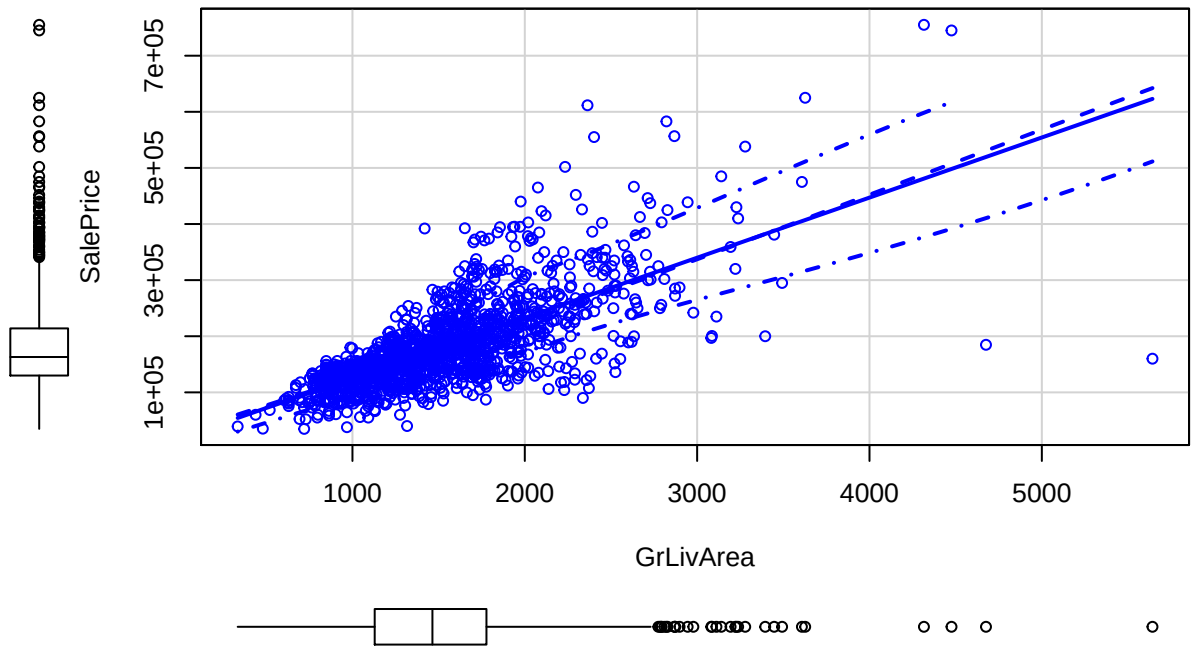


From the barplots, we can see the number of each category of different variables.

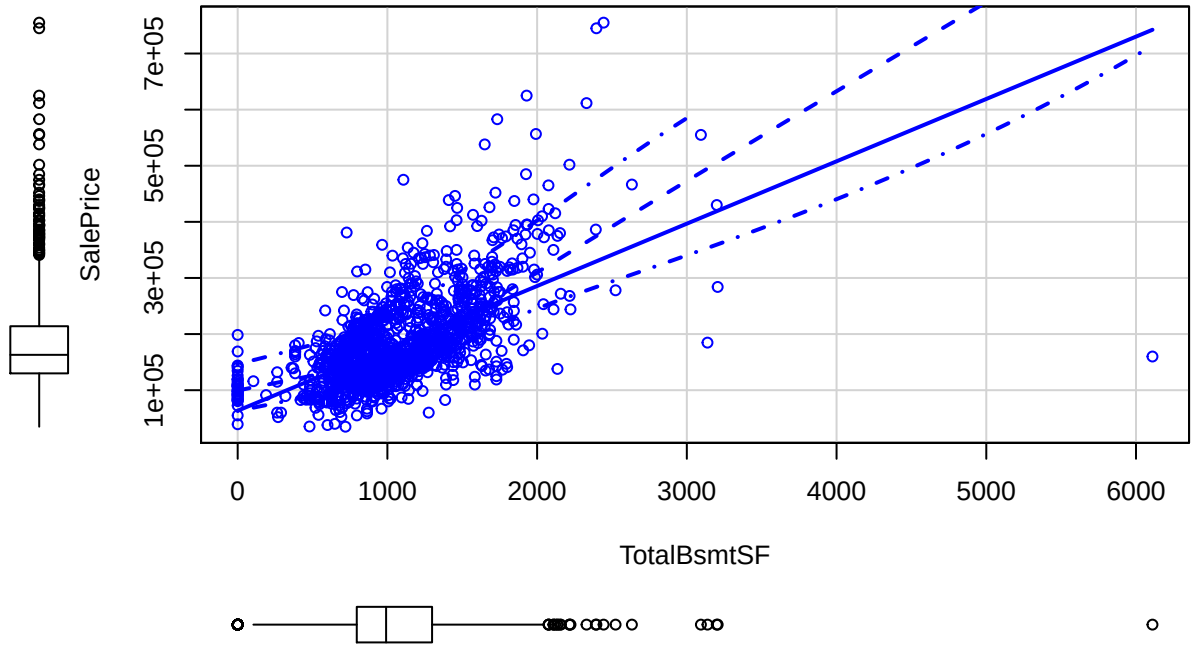
Bivariate Analysis

- Quantitive & Quantitive Variables

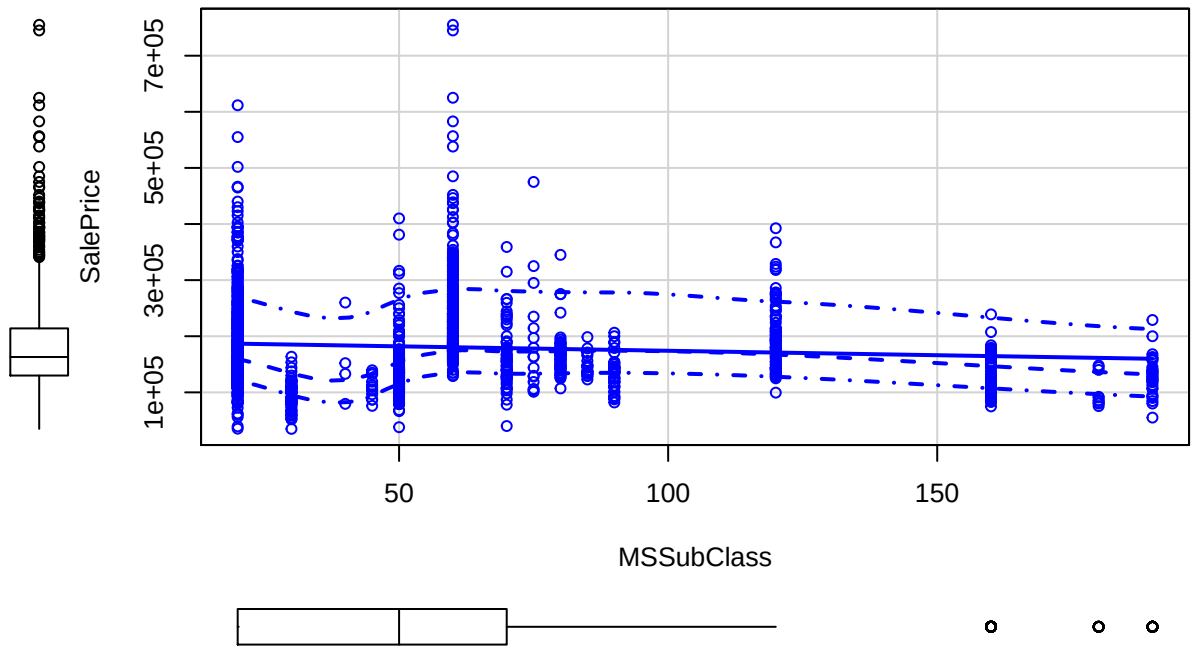
```
# scatterplots
par(mfrow=c(2,4))
scatterplot(SalePrice ~ GrLivArea)
```



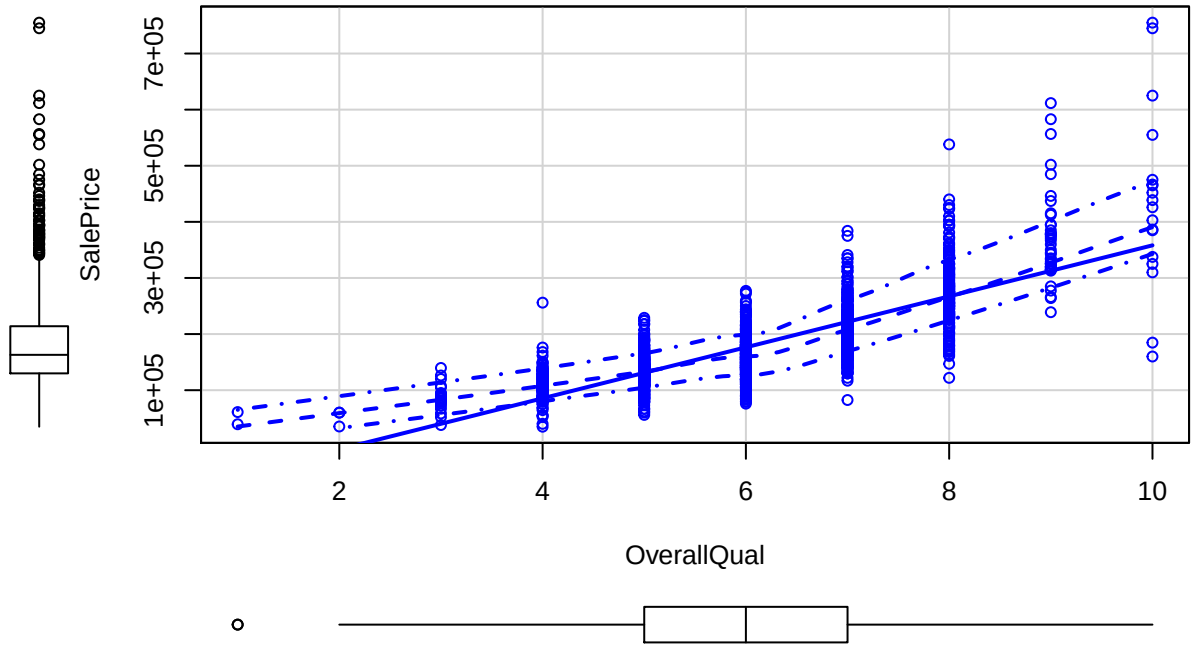

```
scatterplot(SalePrice ~ TotalBsmtSF)
```



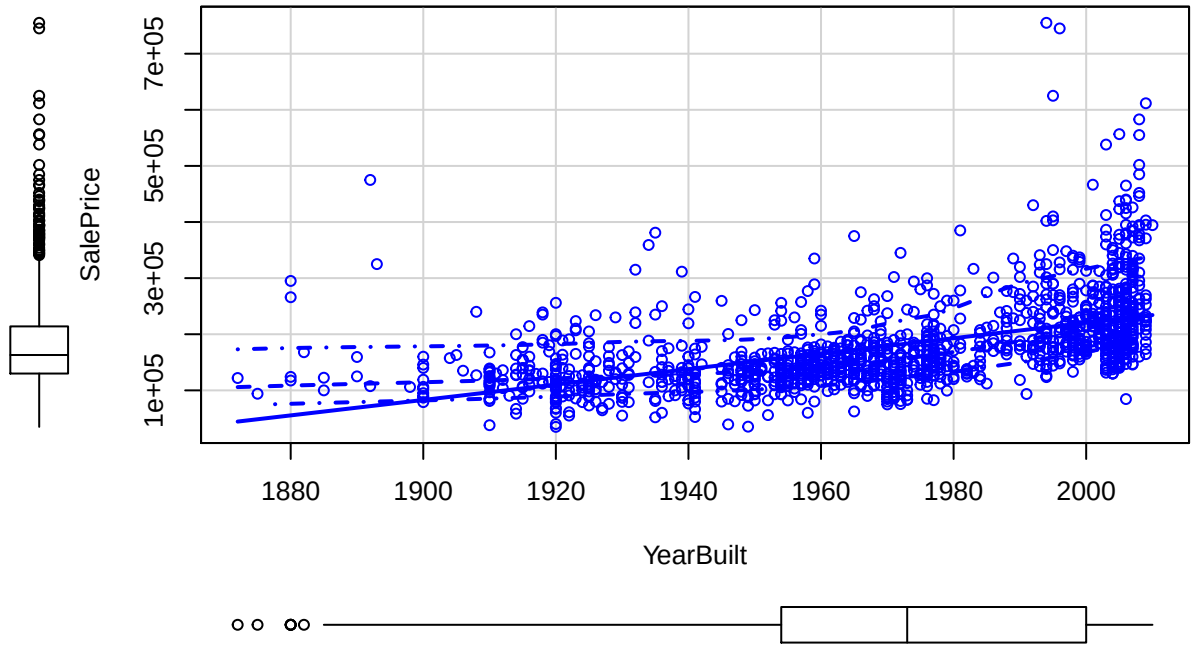

```
scatterplot(SalePrice ~ MSSubClass)
```



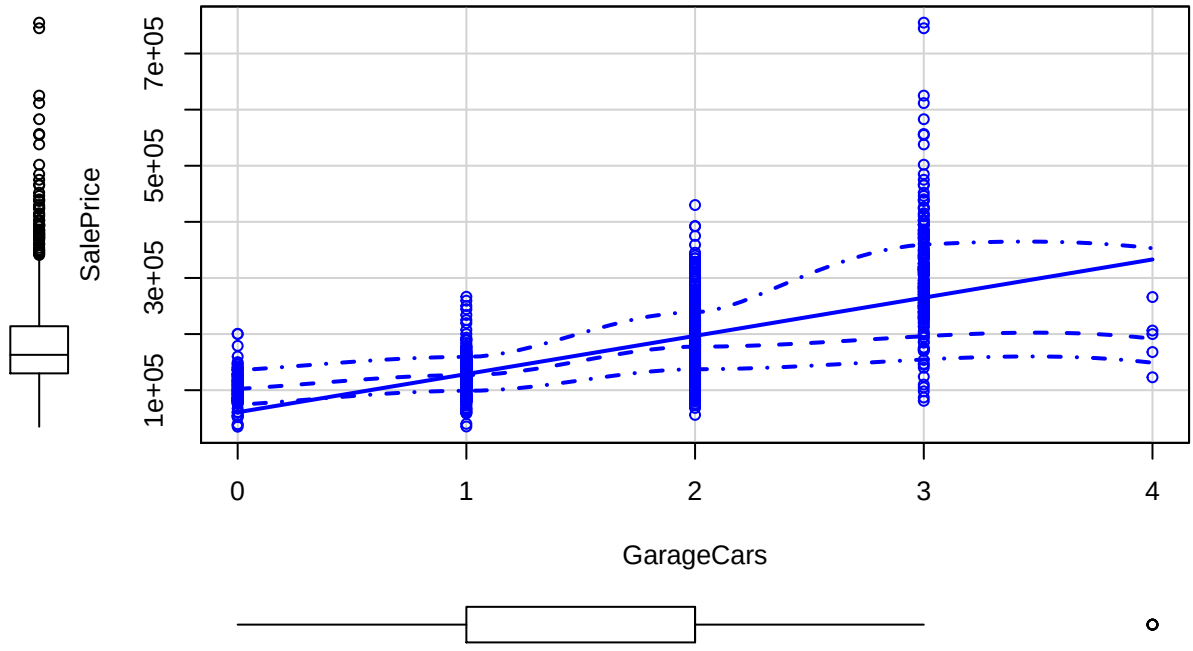

```
scatterplot(SalePrice ~ OverallQual)
```



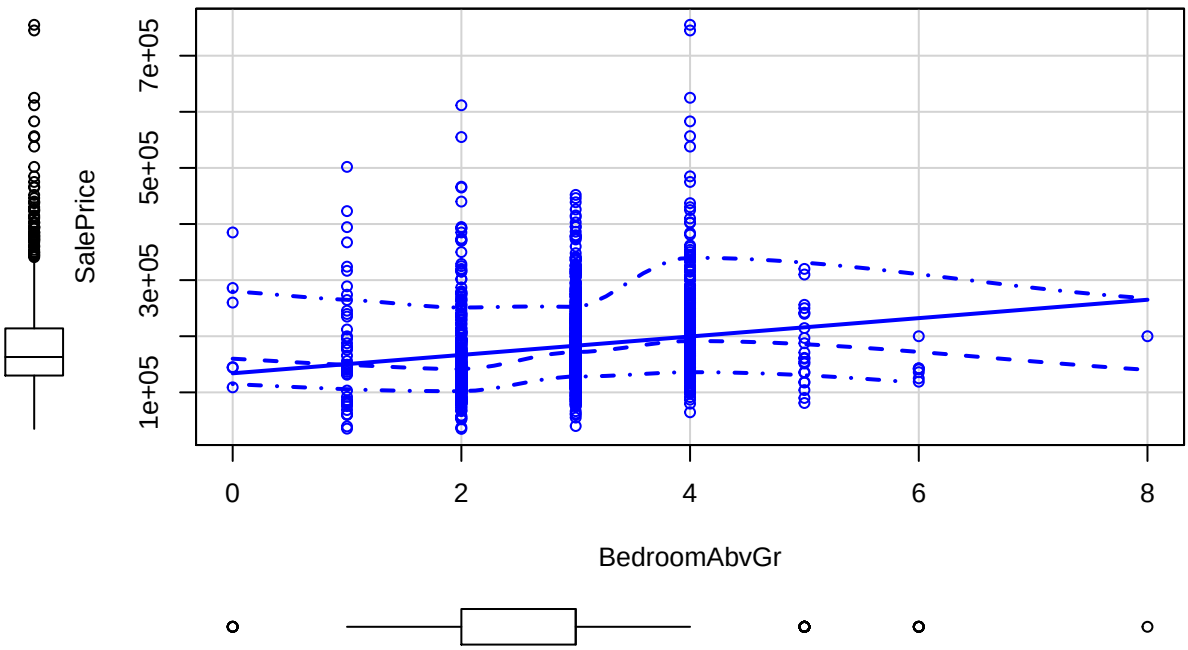

```
scatterplot(SalePrice ~ YearBuilt)
```




```
scatterplot(SalePrice ~ GarageCars)
```




```
scatterplot(SalePrice ~ BedroomAbvGr)
```



For continous variables, the points are concentrated in the lower left corner in the three plots. Combining the comments for histograms, it is necessary to do transforms on y axis because of the large orders of magnitude.

MSSubClass: There are no significant linear relationship between type of dwelling and saleprice from the scatterplot.

OverallQual: The spread and mean increase as the overall quality improved. Obviously, as the overall quality increases, the home price increases.

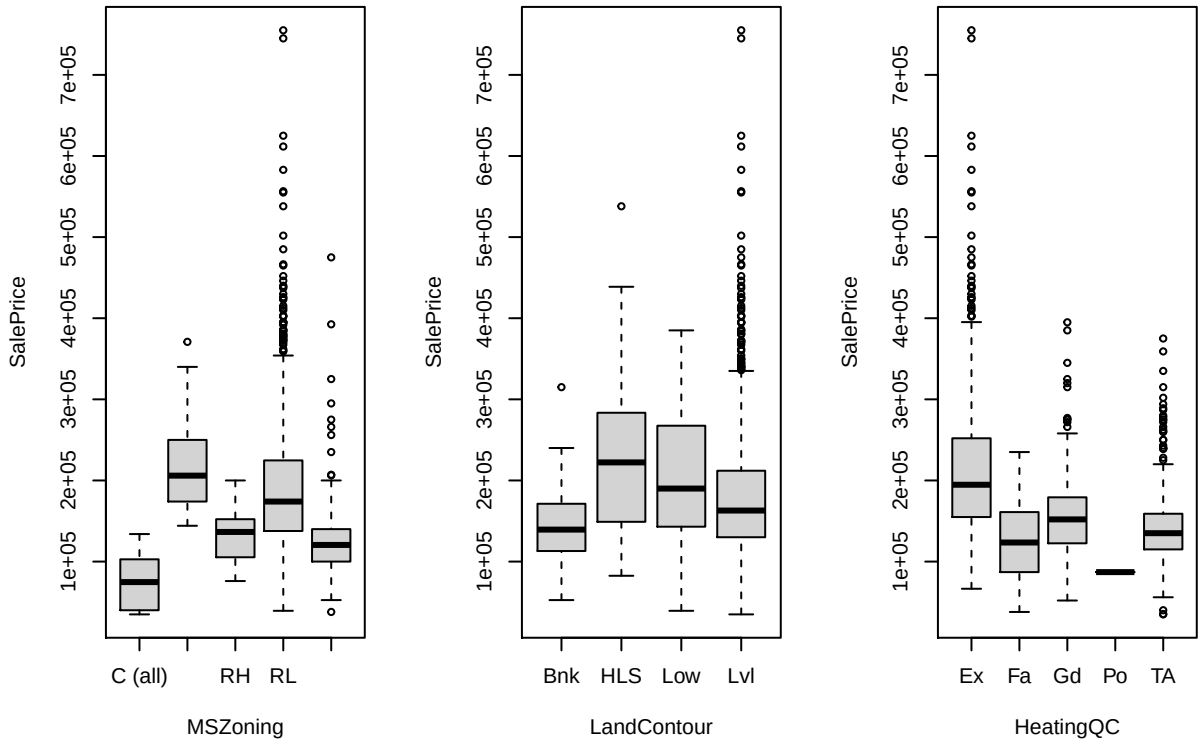
YearBuilt: The data shows that the homeprice raises with slow growth as the year built becomes latest.

GarageCars: There is a signficant increasing associated with an additional car capacity in the garage.

BedroomAbvGr: Home prices raises slowly as the number of bedroom increases.

- Categorical & Quantitive

```
# boxplots
par(mfrow = c(1,3))
boxplot(SalePrice ~ MSZoning)
boxplot(SalePrice ~ LandContour)
boxplot(SalePrice ~ HeatingQC)
```



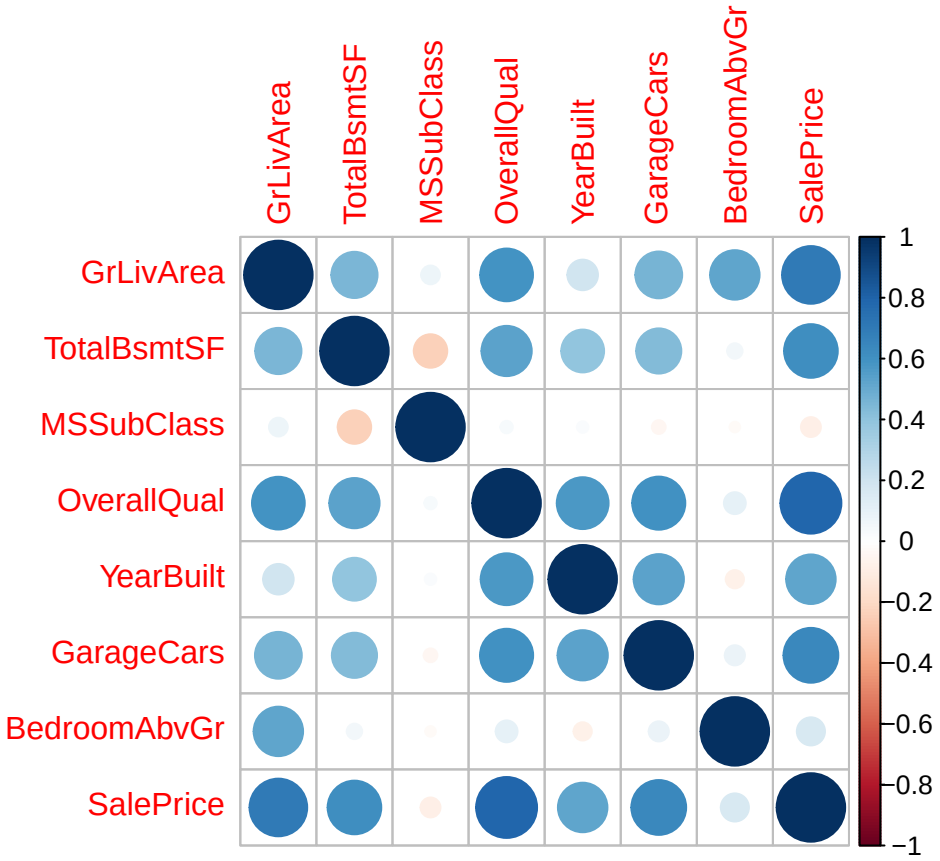
MSZoning: The homeprice of floating village residential and low density residential are the most expensive types. It is make sense because people prefer to live in low density community instead of crowded area.

LandContour: HillSide houses are the most expensive category and the banked houses are the least worthy type.

HeatingQC: The sale prices increases as the heating quality increases.

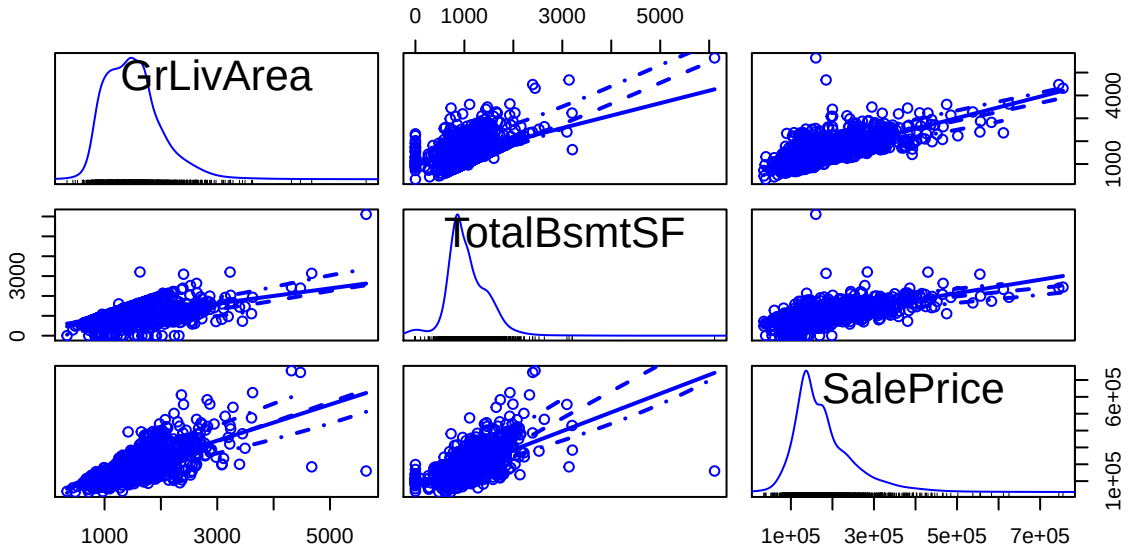
- Entire Model

```
# correlation plot
var_quantitive=c('GrLivArea', 'TotalBsmtSF', 'MSSubClass', 'OverallQual', 'YearBuilt', 'GarageCars', 'B
df_quantitive <- subset(train,select=var_quantitive)
corrplot(cor(df_quantitive))
```



The corvariance between most of preditors and dependent variables are positively strong. However, there are quite significant colinearity among some predictors. We would solve this problem in the later question.

```
# scatterplot
scatterplotMatrix(~ GrLivArea + TotalBsmtSF + SalePrice)
```

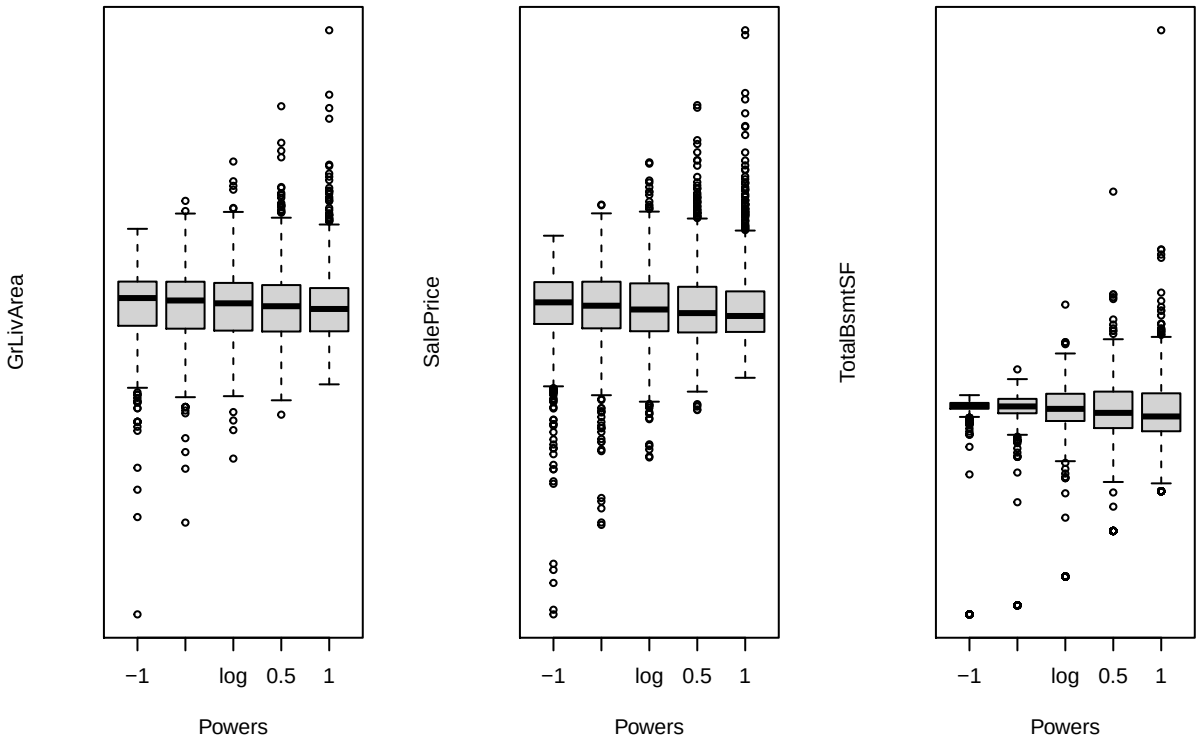


The scatterplot of the continous variables shows that there exit some relationship but the data points are too concentrated.

(b)

```
par(mfrow = c(1,3))
symbox(GrLivArea)
symbox(SalePrice)
symbox(TotalBsmtSF)
```

```
## Warning in symbox.default(TotalBsmtSF): start set to 61.1
```



The outliers of GrLivArea and SalePrice are more even in log transformation. The transformation of TotalBsmtSF can not be decided from the symbox plot because the outliers spread of log or 0.5 power transformation are not even.

```
summary(powerTransform(GrLivArea), data = df, family = 'bcPower')
```

```
## bcPower Transformation to Normality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## GrLivArea    0.0063           0       -0.113       0.1256
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                              LRT df    pval
## LR test, lambda = (0) 0.0107245  1 0.91752
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 275.9499  1 < 2.22e-16
```

```
summary(powerTransform(SalePrice), data = df, family = 'bcPower')
```

```
## bcPower Transformation to Normality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## SalePrice   -0.0769           0      -0.1681       0.0142
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df     pval
```

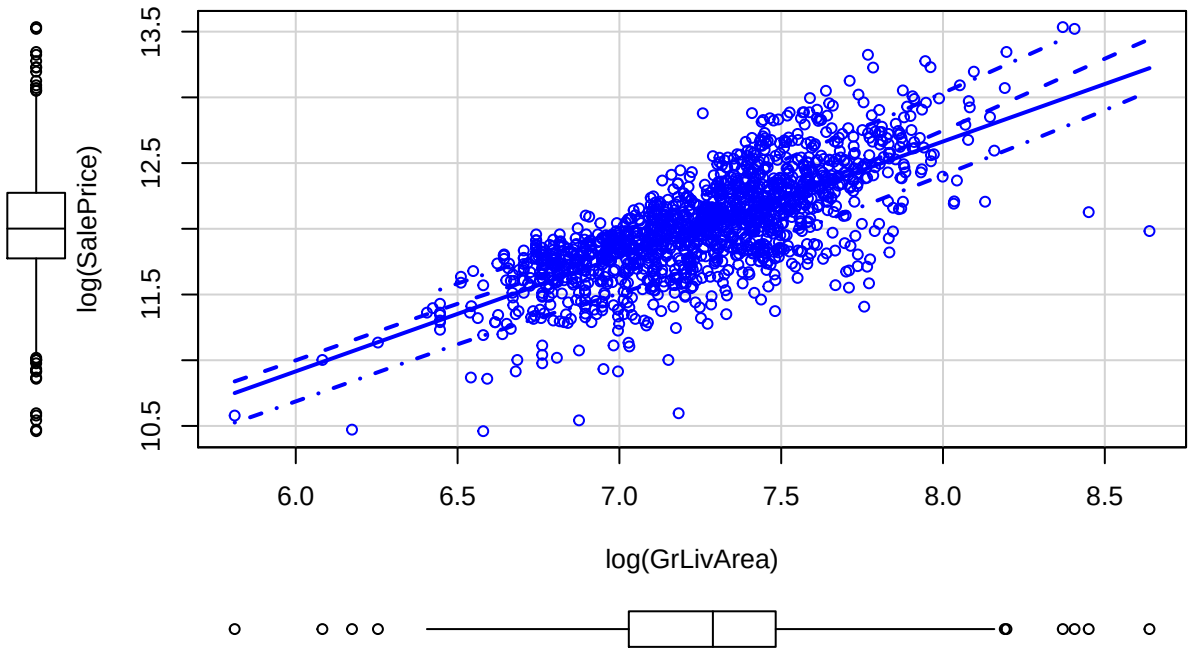
```
## LR test, lambda = (0) 2.723665  1 0.098871
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 517.8114  1 < 2.22e-16
```

```
# Since there are 0 in the datasets of TotalBsmSF, we cannot use Box-Cox transformation.
# We can use Yeo-Johnson transformation to deal with non-positive data.
summary(powerTransform(cbind(TotalBsmtSF) ~ 1, data = df, family = 'yjPower'))
```

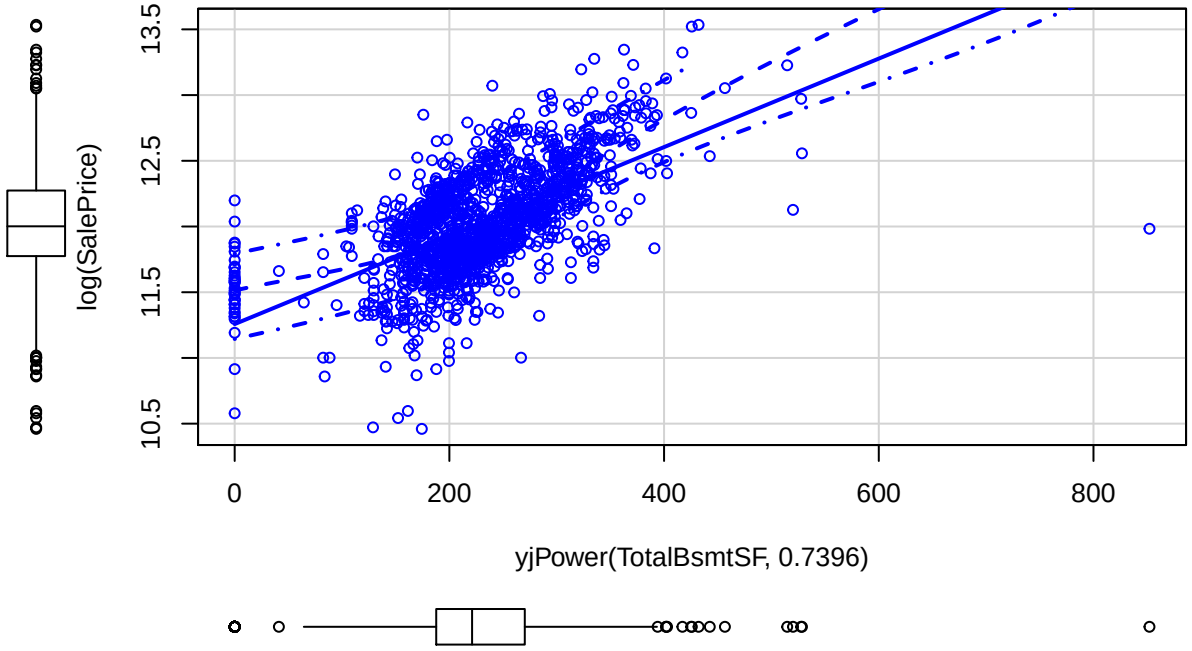
```
## yjPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.7396        0.74       0.6965       0.7827
##
##  Likelihood ratio test that transformation parameter is equal to 0
##                             LRT df       pval
## LR test, lambda = (0) 2458.861  1 < 2.22e-16
```

The results shows that we should do log tranformations for 'GrLivArea' and 'SalePrice' because it fail to reject null hypothesis(need a log transformation) and reject the null hypothesis(no need to do a transformation). And we should do power transformation for 'TotalBsmSF' and 'GarageArea' because it reject the null hypothesis.

```
# scatterplot after transformation
scatterplot(log(SalePrice) ~ log(GrLivArea))
```




```
scatterplot(log(SalePrice) ~ yjPower(TotalBsmtSF, 0.7396))
```



After transformation, the data points are more spread out instead of concentrating in the left-down corner and the relationship between home price and its predictors are more linear and clear.

(c)

```
reg.mod <- lm(log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF, 0.7396) + LandContour
              + MSZoning + MSSubClass + OverallQual + YearBuilt +  + GarageCars + BedroomAbvGr
              + HeatingQC, data = df)

summary(reg.mod)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF,
##     0.7396) + LandContour + MSZoning + MSSubClass + OverallQual +
##     YearBuilt + +GarageCars + BedroomAbvGr + HeatingQC, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48614 -0.08122  0.00843  0.09121  0.49396
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.961e+00  4.393e-01  11.294  < 2e-16 ***
## log(GrLivArea)                 4.519e-01  2.113e-02  21.384  < 2e-16 ***
## yjPower(TotalBsmtSF, 0.7396)  6.672e-04  7.199e-05   9.268  < 2e-16 ***
## LandContourHLS                 1.270e-01  2.977e-02   4.265 2.12e-05 ***
## LandContourLow                 1.629e-01  3.294e-02   4.944 8.53e-07 ***
## LandContourLvl                 4.874e-02  2.046e-02   2.382 0.017347 *
```

```
## MSZoningFV                    4.912e-01  5.444e-02   9.024  < 2e-16 ***
## MSZoningRH                    4.438e-01  6.275e-02   7.073 2.36e-12 ***
## MSZoningRL                    4.977e-01  5.002e-02   9.950  < 2e-16 ***
## MSZoningRM                    3.868e-01  5.047e-02   7.663 3.32e-14 ***
## MSSubClass                   -4.923e-04  1.091e-04  -4.512 6.93e-06 ***
## OverallQual                   9.034e-02  4.955e-03  18.232  < 2e-16 ***
## YearBuilt                     1.288e-03  2.055e-04   6.267 4.87e-10 ***
## GarageCars                    7.634e-02  7.455e-03  10.240  < 2e-16 ***
## BedroomAbvGr                 -2.445e-02  6.508e-03  -3.757 0.000179 ***
## HeatingQCFa                  -1.282e-01  2.390e-02  -5.364 9.48e-08 ***
## HeatingQCGd                  -3.498e-02  1.226e-02  -2.853 0.004397 **
## HeatingQCPo                  -1.938e-01  1.543e-01  -1.256 0.209279
## HeatingQCTA                  -5.999e-02  1.074e-02  -5.587 2.76e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1535 on 1441 degrees of freedom
## Multiple R-squared:  0.8542, Adjusted R-squared:  0.8524
## F-statistic:   469 on 18 and 1441 DF,  p-value: < 2.2e-16
```

- Statisitical Significance

  We can find from the regression results that the coefficients of all ten variables are statistical significant. Although one category of HeatingQC is not significant, it is plausible to ignore the unsignificance. Besides, adjusted R^2 is large, p-value is quite small and degrees of freedom is 1441. Therefore, our model is statistical significant.

- Economic Significance and interpretation (Holding else constant)

Intercept: With other variables equal to zero or at the bottom line, the land price or some compulsive expenditures would be 4.961% of te house price on average.

$\beta_{log(GrLivArea)}$: It make sense that as living area above ground increases, home price increases. Additional 1% of total square feet of living area is associated with 0.4519% increases in home price on average.

$\beta_{TotalBsmtSF}$: The estimate of coefficient indicates that square foot of basement increases, home price increase. Additional 1 square feet of basement is associated with 0.06672% increases in home price on average.

$\beta_{LandContour}$: The most worthy flatness of property is depression, and then is hillside, the third worthy one is level, the least worthy one is the base group banked. Compared banked property, depression property is worth more by 16.29%. Compared banked property, hill side property is worth more by 12.7%. Compared to banked property, flat property is worth more by 4.874%. However, the results are kind of not economic significant because we can see that the hillside houses are the most expensive type from the statisitc summary and common sense.

$\beta_{MSZoning}$: The baseline of MSZoning is the least worthy Commercial classification. People prefer to live in low density area instead of high density or crowed commercial area. The sale price of floating village is expensive than commercial property by 49.12%. The sale price of high density residential is expensive than commercial property by 44.38%. The sale price of low density residential is expensive than commercial property by 49.77%. The sale price of medium density residential is expensive than commercial property by 38.68%.

$\beta_{MSSubClass}$: I think it is not economical significant because it hard to explain.

$\beta_{OverallQual}$: People always prefer good quality houses. As the overall quality increases 1 level, the home price would increases by 9.034%.

$\beta_{YearBuilt}$: People also prefer to live in newly built houses because the equipments, furnitures and functionals are much better. Additional 1 year later of the houses built is associated with an increase of home price by

0.1288% on average.

$\beta_{GarageCars}$: The capacity of garage is another factors for people. It is really convenient to have more capacity of cars. Therefore, additional 1 car capacity in garage is associated with increases of home price by 7.634% on average.

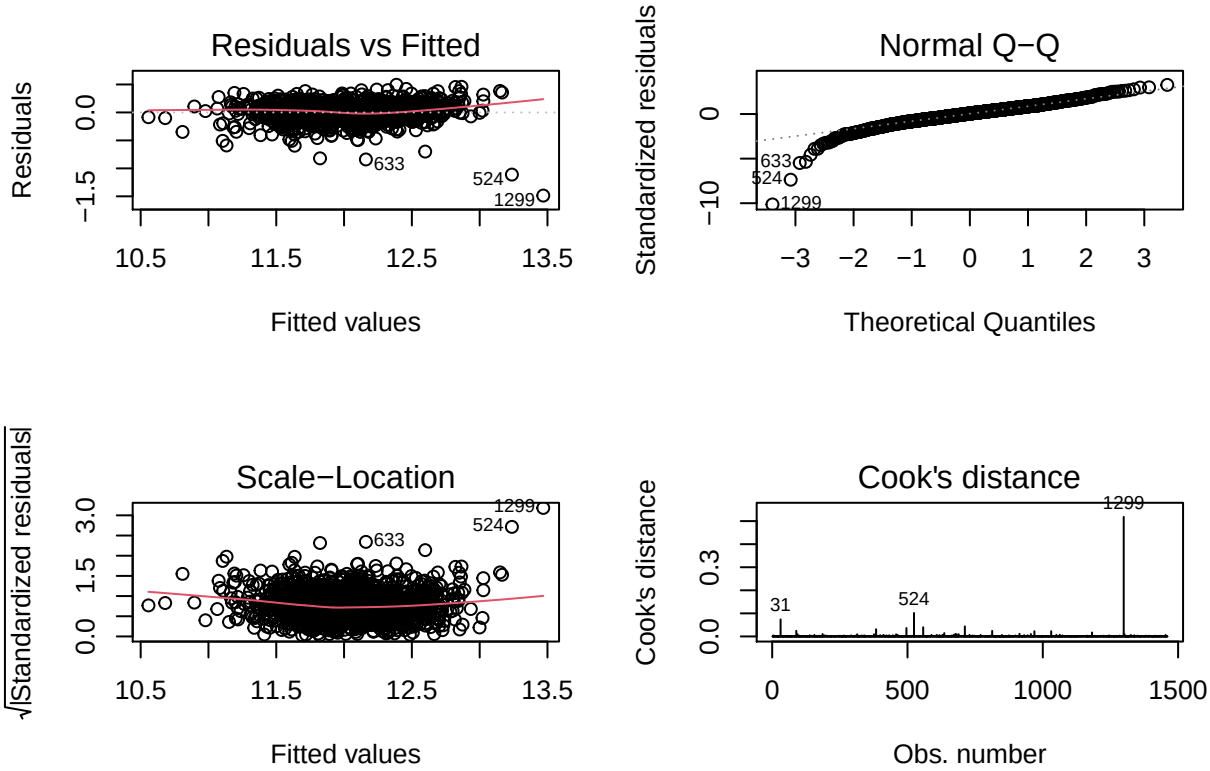
$\beta_{BedroomAbvGr}$: The coefficient is not economic significant because it indicate more bedroom would cause less home price which is not plausible in really life.

$\beta_{HeatingQC}$: It is economic significant because people would prefer excellent heating quality. Compared to excellent heating quality, the sale price of house with fair heating quality would decrease by 12.82%. The sale price of house with good heating quality would less expensive than houses with excellent heating quality by 3.498%. Samely, the sale price of house with average quality would decrease by 2.999%. The house with poor heating quality is not statistic significant but it is economic significant, because it mean the sale price would decreases by 19.38%.

(d)

```
par(mfrow = c(2,2))
plot(reg.mod, 1:4)
```

```
## Warning: not plotting observations with leverage one:
##   326
```



From the above plots, we can figure out four outliers which are 633, 524, 1299 and 31. Specifically, the outliers are which residuals smaller than -0.5.

```
position = which(abs(reg.mod$resid) > 0.8)
position
```

```
##  411  524  633 1299
##  411  524  633 1299
```

So we can remove the outliers using above functions.

```
reg.mod_RemoveOutliers = lm(log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF, 0.7396) + LandContour
              + MSZoning + MSSubClass + OverallQual + YearBuilt +  + GarageCars + BedroomAbvGr
              + HeatingQC, data = df, subset = abs(reg.mod$resid) <= 0.8)

summary(reg.mod_RemoveOutliers)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF,
##     0.7396) + LandContour + MSZoning + MSSubClass + OverallQual +
##     YearBuilt + +GarageCars + BedroomAbvGr + HeatingQC, data = df,
##     subset = abs(reg.mod$resid) <= 0.8)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72152 -0.08344  0.00754  0.09140  0.46937
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.630e+00  4.047e-01  11.439  < 2e-16 ***
## log(GrLivArea)                4.772e-01  1.953e-02  24.435  < 2e-16 ***
## yjPower(TotalBsmtSF, 0.7396)  8.939e-04  6.817e-05  13.112  < 2e-16 ***
## LandContourHLS                7.573e-02  2.764e-02   2.740  0.00623 **
## LandContourLow                1.056e-01  3.058e-02   3.453  0.00057 ***
## LandContourLvl                1.097e-03  1.916e-02   0.057  0.95435
## MSZoningFV                    4.897e-01  5.008e-02   9.779  < 2e-16 ***
## MSZoningRH                    4.335e-01  5.773e-02   7.510 1.03e-13 ***
## MSZoningRL                    4.945e-01  4.602e-02  10.747  < 2e-16 ***
## MSZoningRM                    3.864e-01  4.643e-02   8.322  < 2e-16 ***
## MSSubClass                   -4.591e-04  1.005e-04  -4.569 5.32e-06 ***
## OverallQual                   8.804e-02  4.566e-03  19.283  < 2e-16 ***
## YearBuilt                     1.383e-03  1.891e-04   7.311 4.37e-13 ***
## GarageCars                    6.735e-02  6.887e-03   9.779  < 2e-16 ***
## BedroomAbvGr                 -2.991e-02  6.001e-03  -4.984 6.98e-07 ***
## HeatingQCFa                  -1.190e-01  2.200e-02  -5.412 7.28e-08 ***
## HeatingQCGd                  -3.017e-02  1.128e-02  -2.673  0.00759 **
## HeatingQCPo                  -1.697e-01  1.419e-01  -1.196  0.23199
## HeatingQCTA                  -5.352e-02  9.895e-03  -5.409 7.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1412 on 1437 degrees of freedom
## Multiple R-squared:  0.8762, Adjusted R-squared:  0.8746
## F-statistic: 564.8 on 18 and 1437 DF,  p-value: < 2.2e-16
```

```
compareCoefs(reg.mod, reg.mod_RemoveOutliers)
```

```
## Calls:
## 1: lm(formula = log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF,
```

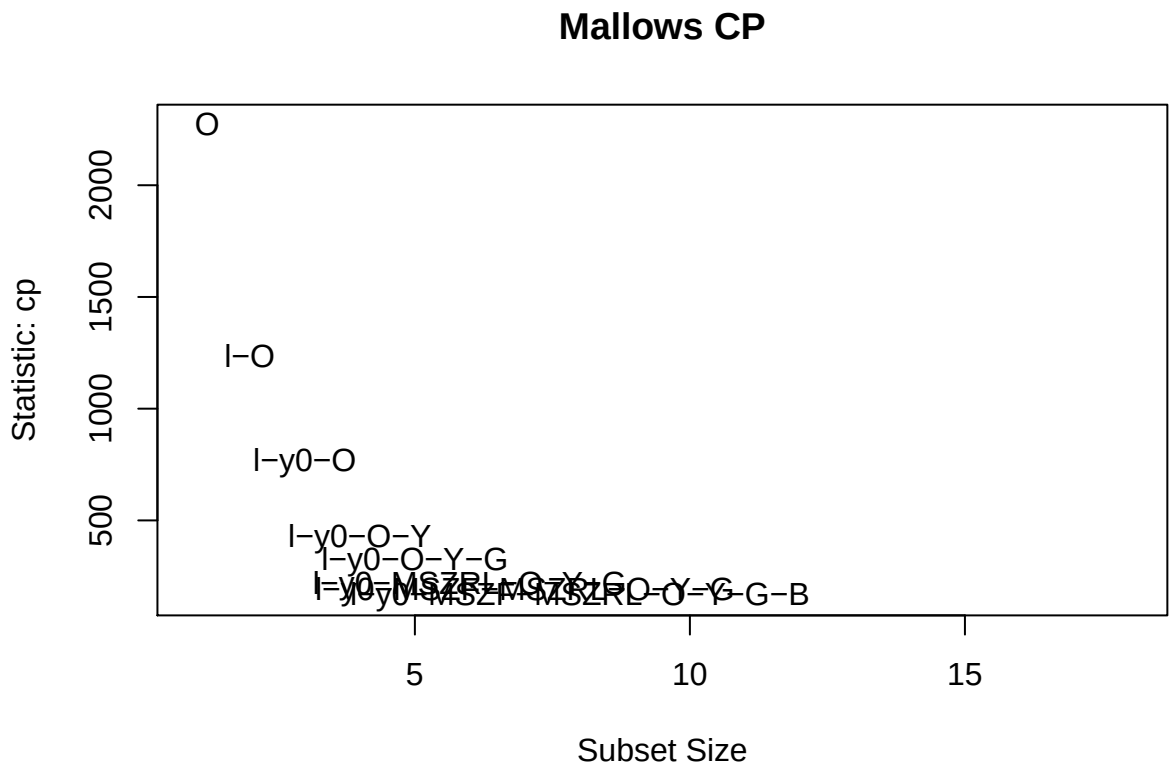
```
##   0.7396) + LandContour + MSZoning + MSSubClass + OverallQual + YearBuilt +
##   +GarageCars + BedroomAbvGr + HeatingQC, data = df)
## 2: lm(formula = log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF,
##   0.7396) + LandContour + MSZoning + MSSubClass + OverallQual + YearBuilt +
##   +GarageCars + BedroomAbvGr + HeatingQC, data = df, subset =
##   abs(reg.mod$resid) <= 0.8)
##
##                                Model 1   Model 2
## (Intercept)                      4.961     4.630
## SE                               0.439     0.405
##
## log(GrLivArea)                  0.4519    0.4772
## SE                              0.0211    0.0195
##
## yjPower(TotalBsmtSF, 0.7396)  6.67e-04  8.94e-04
## SE                            7.20e-05  6.82e-05
##
## LandContourHLS                  0.1270    0.0757
## SE                              0.0298    0.0276
##
## LandContourLow                  0.1629    0.1056
## SE                              0.0329    0.0306
##
## LandContourLvl                  0.0487    0.0011
## SE                              0.0205    0.0192
##
## MSZoningFV                      0.4912    0.4897
## SE                              0.0544    0.0501
##
## MSZoningRH                      0.4438    0.4335
## SE                              0.0628    0.0577
##
## MSZoningRL                       0.498     0.494
## SE                               0.050     0.046
##
## MSZoningRM                      0.3868    0.3864
## SE                              0.0505    0.0464
##
## MSSubClass                    -0.000492 -0.000459
## SE                             0.000109  0.000100
##
## OverallQual                     0.09034   0.08804
## SE                              0.00496   0.00457
##
## YearBuilt                      0.001288  0.001383
## SE                             0.000205  0.000189
##
## GarageCars                      0.07634   0.06735
## SE                              0.00745   0.00689
##
## BedroomAbvGr                   -0.02445  -0.02991
## SE                              0.00651   0.00600
##
## HeatingQCFa                     -0.1282   -0.1190
```

```
## SE                              0.0239    0.0220
##
## HeatingQCGd                    -0.0350   -0.0302
## SE                              0.0123    0.0113
##
## HeatingQCPo                     -0.194    -0.170
## SE                               0.154     0.142
##
## HeatingQCTA                   -0.05999  -0.05352
## SE                             0.01074   0.00989
##
```

After removing the outliers, the adjusted R^2 is larger and the standard error of all estimate coefficients are smaller.

(e)

```
# use Mallow CP to decide remain which variables
ss <- regsubsets(log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF, 0.7396) + LandContour
              + MSZoning + MSSubClass + OverallQual + YearBuilt + GarageCars + BedroomAbvGr + HeatingQC
              method = c("exhaustive"), nbest = 1, data = df, subset = abs(reg.mod$resid) <= 0.8)

subsets(ss, statistic = "cp",legend = F, main = "Mallows CP")
```




```
##                              Abbreviation
## log(GrLivArea)                          l
```

```
## yjPower(TotalBsmtSF, 0.7396)           y0
## LandContourHLS                        LCH
## LandContourLow                    LndCntrLw
## LandContourLvl                    LndCntrLv
## MSZoningFV                             MSZF
## MSZoningRH                            MSZRH
## MSZoningRL                            MSZRL
## MSZoningRM                            MSZRM
## MSSubClass                              MSS
## OverallQual                               O
## YearBuilt                                 Y
## GarageCars                                G
## BedroomAbvGr                              B
## HeatingQCFa                            HQCF
## HeatingQCGd                            HQCG
## HeatingQCPo                            HQCP
## HeatingQCTA                            HQCT
```

The Mallow CP indicate that we should use log(GrLivArea), yjPower(TotalBsmtSF, 0.7396), MSZoning, OverallQual, YearBuilt, GarageCars and BedroomAbvGr as our predictors.

```
# test for multicollinearity
vif(reg.mod_RemoveOutliers)
```

```
##               log(GrLivArea) yjPower(TotalBsmtSF, 0.7396)
##                       3.0438                       1.6872
##               LandContourHLS               LandContourLow
##                       1.8512                       1.6472
##               LandContourLvl                  MSZoningFV
##                       2.4333                       7.8136
##                  MSZoningRH                  MSZoningRL
##                       2.6458                      25.8620
##                  MSZoningRM                  MSSubClass
##                      20.0480                       1.3211
##                 OverallQual                   YearBuilt
##                       2.8851                       2.3838
##                  GarageCars                 BedroomAbvGr
##                       1.9340                       1.7542
##                 HeatingQCFa                 HeatingQCGd
##                       1.1495                       1.2850
##                 HeatingQCPo                 HeatingQCTA
##                       1.0103                       1.4805
```

Since there are multicolinearity in MSZoningFV, MSZoningRL and MSZoningRM(VIFs > 5), we should remove these in our model.

All in all, we should remove MSZoning as well.

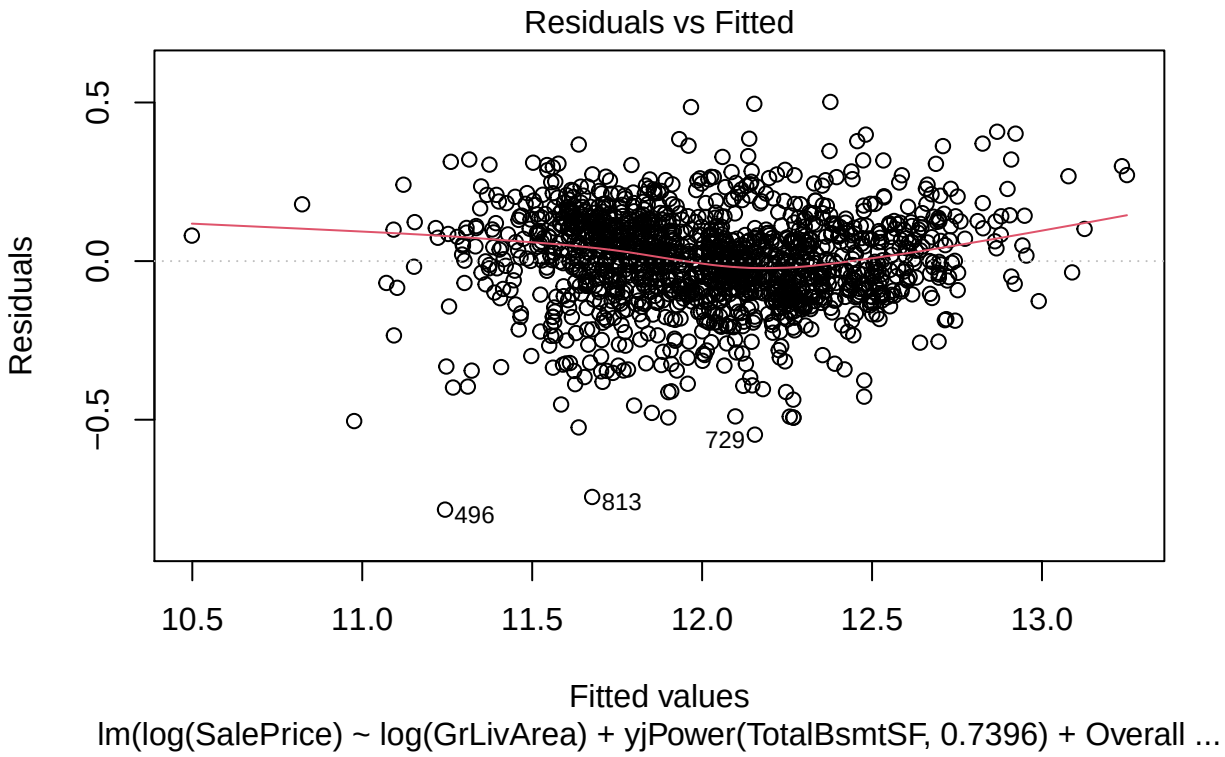
```
# re-estimate the model
reg.mod2 <- lm(log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF, 0.7396)
               + OverallQual + YearBuilt + GarageCars + BedroomAbvGr,
               data = df, subset = abs(reg.mod$resid) <= 0.5)
summary(reg.mod2)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF,
```

```
##     0.7396) + OverallQual + YearBuilt + GarageCars + BedroomAbvGr,
##     data = df, subset = abs(reg.mod$resid) <= 0.5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78339 -0.08728  0.00491  0.09352  0.50166
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.321e+00  3.713e-01   8.945  < 2e-16 ***
## log(GrLivArea)                4.738e-01  2.000e-02  23.689  < 2e-16 ***
## yjPower(TotalBsmtSF, 0.7396)  1.111e-03  6.728e-05  16.513  < 2e-16 ***
## OverallQual                   8.932e-02  4.668e-03  19.137  < 2e-16 ***
## YearBuilt                     2.239e-03  1.722e-04  13.003  < 2e-16 ***
## GarageCars                    6.660e-02  7.227e-03   9.215  < 2e-16 ***
## BedroomAbvGr                 -2.230e-02  6.166e-03  -3.617 0.000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1495 on 1444 degrees of freedom
## Multiple R-squared:  0.8567, Adjusted R-squared:  0.8561
## F-statistic:  1438 on 6 and 1444 DF,  p-value: < 2.2e-16
```

(f)

```
plot(reg.mod2, 1)
```



Although there are some outliers, the spread of residual is same and the mean of residul is almost zero.

(g)

```
AIC(reg.mod, reg.mod2)
```

```
## Warning in AIC.default(reg.mod, reg.mod2): models are not all fitted to the same
## number of observations
```

```
##          df       AIC
## reg.mod  20 -1308.559
## reg.mod2  8 -1388.289
```

```
BIC(reg.mod, reg.mod2)
```

```
## Warning in BIC.default(reg.mod, reg.mod2): models are not all fitted to the same
## number of observations
```

```
##          df       BIC
## reg.mod  20 -1202.835
## reg.mod2  8 -1346.049
```

According to the AIC and BIC, reg,mod2 is better because it has smaller AIC and BIC value.

(h)

```
resettest(reg.mod2, power = 2)
```

```
##
##  RESET test
##
## data:  reg.mod2
## RESET = 27.605, df1 = 1, df2 = 1443, p-value = 1.711e-07
```

The RESET test indicate that the model need quadratic terms to improve our model.

```
model_reset2 <- lm(log(SalePrice) ~ (log(GrLivArea) + yjPower(TotalBsmtSF, 0.7396)
                                     + OverallQual + YearBuilt + GarageCars + BedroomAbvGr)^2
                   , data = df, subset = abs(reg.mod$resid) <= 0.8)
summary(model_reset2)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ (log(GrLivArea) + yjPower(TotalBsmtSF,
##     0.7396) + OverallQual + YearBuilt + GarageCars + BedroomAbvGr)^2,
##     data = df, subset = abs(reg.mod$resid) <= 0.8)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92490 -0.07545  0.01045  0.09069  0.53267
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 1.503e+00  8.900e+00   0.169
## log(GrLivArea)                             -2.357e-01  1.489e+00  -0.158
## yjPower(TotalBsmtSF, 0.7396)               -9.197e-03  7.691e-03  -1.196
## OverallQual                                 1.217e+00  3.262e-01   3.731
## YearBuilt                                   4.348e-03  4.659e-03   0.933
## GarageCars                                 -2.283e-01  5.667e-01  -0.403
## BedroomAbvGr                                1.209e-01  5.769e-01   0.210
## log(GrLivArea):yjPower(TotalBsmtSF, 0.7396)  2.252e-04  3.412e-04   0.660
## log(GrLivArea):OverallQual                  4.223e-02  1.583e-02   2.669
## log(GrLivArea):YearBuilt                    1.882e-04  7.829e-04   0.240
## log(GrLivArea):GarageCars                   3.372e-02  2.933e-02   1.150
## log(GrLivArea):BedroomAbvGr                -8.700e-03  1.675e-02  -0.520
## yjPower(TotalBsmtSF, 0.7396):OverallQual    1.211e-04  6.894e-05   1.756
## yjPower(TotalBsmtSF, 0.7396):YearBuilt      4.871e-06  3.652e-06   1.334
## yjPower(TotalBsmtSF, 0.7396):GarageCars    -4.388e-04  1.232e-04  -3.562
## yjPower(TotalBsmtSF, 0.7396):BedroomAbvGr  -2.986e-04  9.664e-05  -3.090
## OverallQual:YearBuilt                      -7.529e-04  1.538e-04  -4.895
## OverallQual:GarageCars                      1.024e-02  6.651e-03   1.540
## OverallQual:BedroomAbvGr                   -1.373e-03  6.089e-03  -0.225
## YearBuilt:GarageCars                        5.022e-05  2.603e-04   0.193
## YearBuilt:BedroomAbvGr                      3.054e-06  2.779e-04   0.011
## GarageCars:BedroomAbvGr                    -2.460e-03  9.405e-03  -0.262
##                                            Pr(>|t|)
## (Intercept)                                0.865926
## log(GrLivArea)                             0.874263
## yjPower(TotalBsmtSF, 0.7396)               0.231976
```

```
## OverallQual                                    0.000198 ***
## YearBuilt                                      0.350834
## GarageCars                                     0.687128
## BedroomAbvGr                                   0.834076
## log(GrLivArea):yjPower(TotalBsmtSF, 0.7396) 0.509283
## log(GrLivArea):OverallQual                     0.007702 **
## log(GrLivArea):YearBuilt                       0.810061
## log(GrLivArea):GarageCars                      0.250539
## log(GrLivArea):BedroomAbvGr                    0.603476
## yjPower(TotalBsmtSF, 0.7396):OverallQual    0.079224 .
## yjPower(TotalBsmtSF, 0.7396):YearBuilt      0.182441
## yjPower(TotalBsmtSF, 0.7396):GarageCars     0.000380 ***
## yjPower(TotalBsmtSF, 0.7396):BedroomAbvGr   0.002041 **
## OverallQual:YearBuilt                          1.1e-06 ***
## OverallQual:GarageCars                         0.123785
## OverallQual:BedroomAbvGr                       0.821648
## YearBuilt:GarageCars                           0.847032
## YearBuilt:BedroomAbvGr                         0.991233
## GarageCars:BedroomAbvGr                        0.793729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1505 on 1434 degrees of freedom
## Multiple R-squared:  0.8596, Adjusted R-squared:  0.8575
## F-statistic:   418 on 21 and 1434 DF,  p-value: < 2.2e-16
```

The regression results indicate that the quadratic term of overall quality and an interaction term of overall quality and year built is statistical significant. They are also economic significant because there is diminishing effect on overall quality and the overall quality and year built is related to some degree.

```
# re-estimate the model
reg.mod3 <- lm(log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF, 0.7396) + OverallQual
               + I(OverallQual^2) + YearBuilt + OverallQual:YearBuilt + GarageCars + BedroomAbvGr,
               data = df, subset = abs(reg.mod$resid) <= 0.8)
summary(reg.mod3)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF,
##     0.7396) + OverallQual + I(OverallQual^2) + YearBuilt + OverallQual:YearBuilt +
##     GarageCars + BedroomAbvGr, data = df, subset = abs(reg.mod$resid) <=
##     0.8)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92894 -0.08348  0.00915  0.09406  0.51537
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -4.129e+00  1.529e+00  -2.700  0.00701 **
## log(GrLivArea)                 4.733e-01  2.051e-02  23.069  < 2e-16 ***
## yjPower(TotalBsmtSF, 0.7396)  1.107e-03  6.926e-05  15.980  < 2e-16 ***
## OverallQual                    1.219e+00  2.347e-01   5.193 2.37e-07 ***
## I(OverallQual^2)               9.629e-03  1.863e-03   5.170 2.67e-07 ***
## YearBuilt                      6.200e-03  7.898e-04   7.850 8.05e-15 ***
```

```
## GarageCars                    6.695e-02  7.423e-03   9.020  < 2e-16 ***
## BedroomAbvGr                 -2.067e-02  6.402e-03  -3.229  0.00127 **
## OverallQual:YearBuilt        -6.331e-04  1.255e-04  -5.045 5.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1533 on 1447 degrees of freedom
## Multiple R-squared:  0.853,  Adjusted R-squared:  0.8522
## F-statistic:  1049 on 8 and 1447 DF,  p-value: < 2.2e-16
```

```
compareCoefs(reg.mod, reg.mod2, reg.mod3)
```

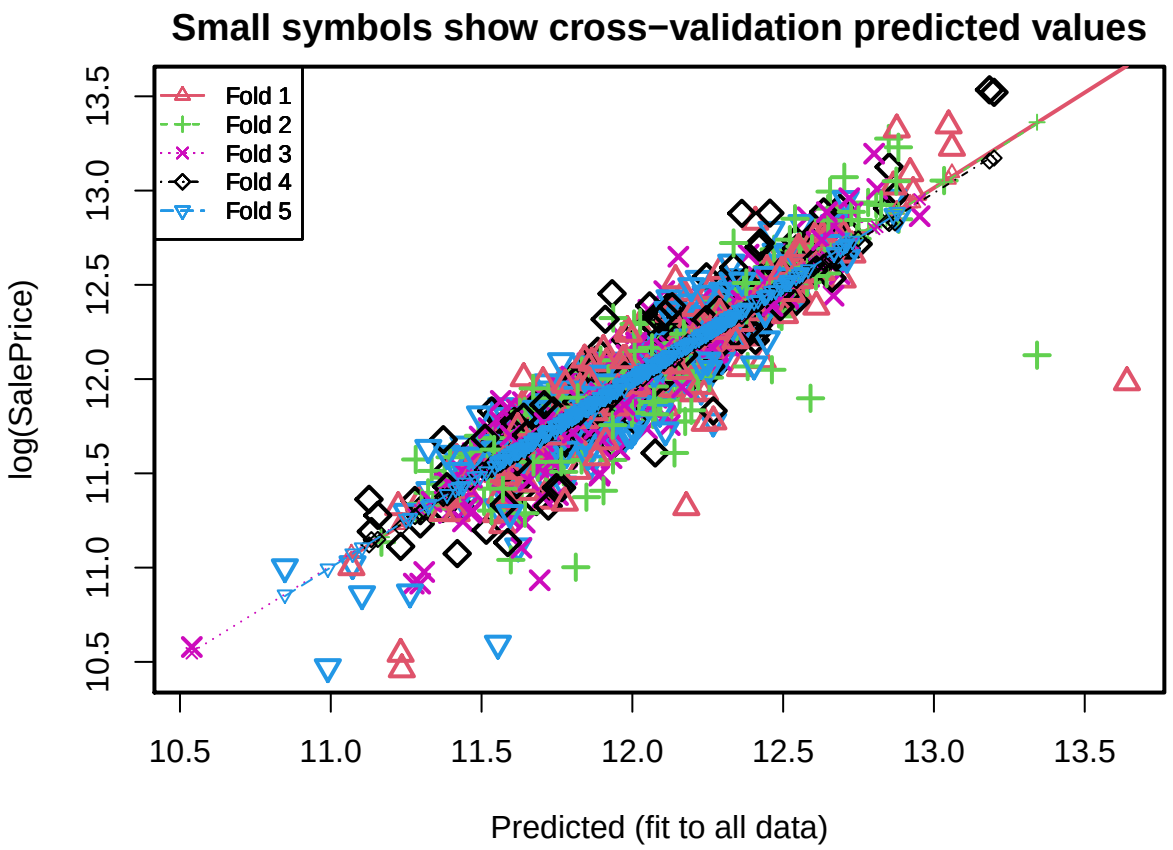
```
## Calls:
## 1: lm(formula = log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF,
##   0.7396) + LandContour + MSZoning + MSSubClass + OverallQual + YearBuilt +
##   +GarageCars + BedroomAbvGr + HeatingQC, data = df)
## 2: lm(formula = log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF,
##   0.7396) + OverallQual + YearBuilt + GarageCars + BedroomAbvGr, data = df,
##   subset = abs(reg.mod$resid) <= 0.5)
## 3: lm(formula = log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF,
##   0.7396) + OverallQual + I(OverallQual^2) + YearBuilt +
##   OverallQual:YearBuilt + GarageCars + BedroomAbvGr, data = df, subset =
##   abs(reg.mod$resid) <= 0.8)
##
##                               Model 1  Model 2   Model 3
## (Intercept)                     4.961    3.321    -4.129
## SE                              0.439    0.371     1.529
##
## log(GrLivArea)                 0.4519   0.4738    0.4733
## SE                             0.0211   0.0200    0.0205
##
## yjPower(TotalBsmtSF, 0.7396)  6.67e-04 1.11e-03  1.11e-03
## SE                            7.20e-05 6.73e-05  6.93e-05
##
## LandContourHLS                 0.1270
## SE                             0.0298
##
## LandContourLow                 0.1629
## SE                             0.0329
##
## LandContourLvl                 0.0487
## SE                             0.0205
##
## MSZoningFV                     0.4912
## SE                             0.0544
##
## MSZoningRH                     0.4438
## SE                             0.0628
##
## MSZoningRL                      0.498
## SE                              0.050
##
## MSZoningRM                     0.3868
## SE                             0.0505
##
```

```
## MSSubClass                    -0.000492
## SE                             0.000109
##
## OverallQual                     0.09034  0.08932   1.21890
## SE                              0.00496  0.00467   0.23473
##
## YearBuilt                      0.001288 0.002239  0.006200
## SE                             0.000205 0.000172  0.000790
##
## GarageCars                      0.07634  0.06660   0.06695
## SE                              0.00745  0.00723   0.00742
##
## BedroomAbvGr                   -0.02445 -0.02230  -0.02067
## SE                              0.00651  0.00617   0.00640
##
## HeatingQCFa                    -0.1282
## SE                              0.0239
##
## HeatingQCGd                    -0.0350
## SE                              0.0123
##
## HeatingQCPo                     -0.194
## SE                               0.154
##
## HeatingQCTA                    -0.0600
## SE                              0.0107
##
## I(OverallQual^2)                                    0.00963
## SE                                                  0.00186
##
## OverallQual:YearBuilt                              -0.000633
## SE                                                  0.000125
##
```

Compared to previous two model, the model with quadratic terms are kind of not economic significant. The estimate coefficient of $OverallQual^2$ is positive, however, it should be negative because of the diminishing marginal effect of overall quality. And the intercept is negative, which is also irrational because there do exist some fixed cost on house sale. Among the three model, I prefer the model 2.

(i)

```
cvResults <- suppressWarnings(CVlm(data = df, form.lm = reg.mod2, m=5,
                                   dots = FALSE, seed = 1, legend.pos = "topleft",
                                   printit = FALSE))
```



The fit don't vary too much with respect the slope and level. It indicates the model 2 fits good.

```
train_control <- trainControl(method = "cv", number = 5, savePredictions = TRUE, returnResamp = "all")
train(log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF, 0.7396) + OverallQual
      + YearBuilt + GarageCars + BedroomAbvGr, data = df, subset = abs(reg.mod$resid) <= 0.8,
      trControl = train_control, method = "lm")
```

```
## Linear Regression
##
## 1460 samples
##    6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1165, 1165, 1165, 1164, 1165
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1550539  0.8477172  0.1153473
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

R^2 is large and RMSE is small, our model is good with 5-fold cv test.

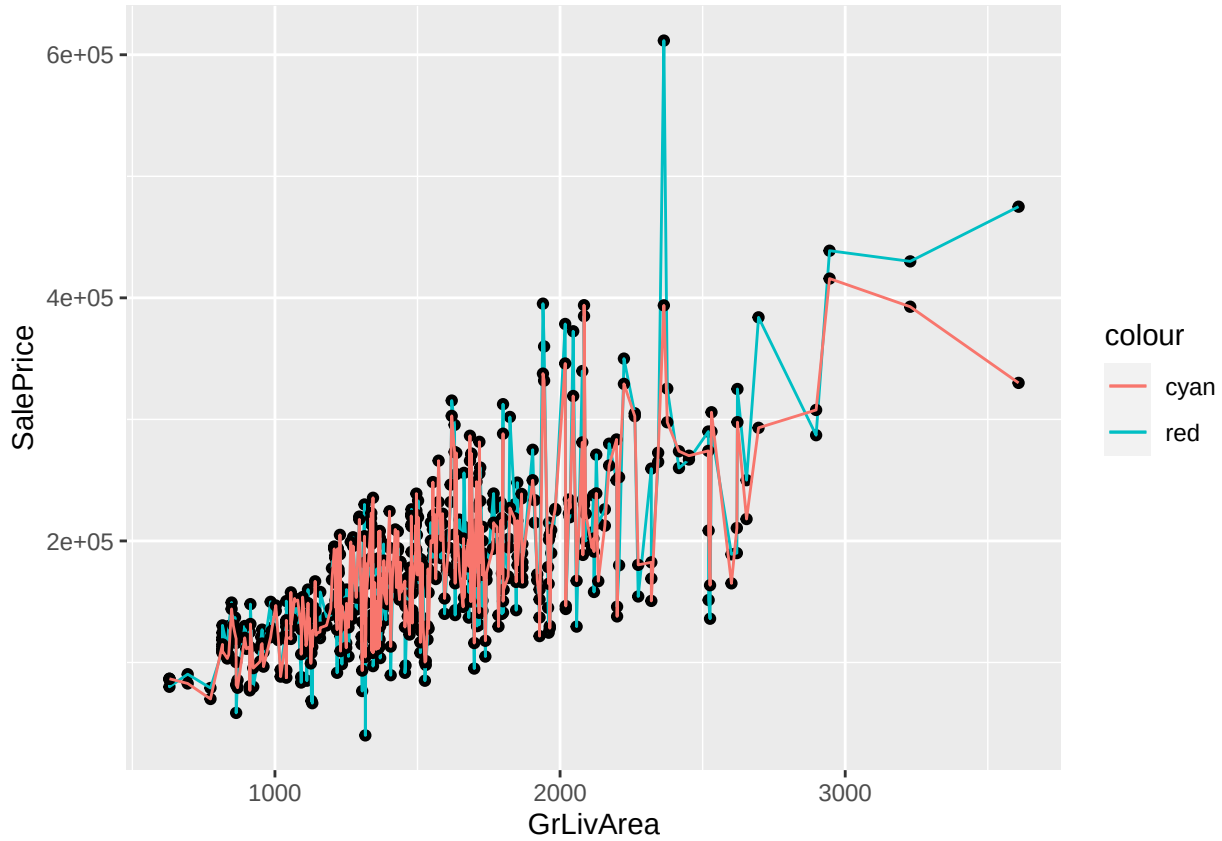
Next, we should do prediction to test the model using dataset "testing.csv". However, there are no data about saleprice which is the important dependent variables y. Therefore, I divided the training dataset to two datasets as new training and testing data sets.

```
set.seed(123)
split = sample.split(train, SplitRatio = 0.8)

training_set = subset(train, split == TRUE)
test_set = subset(train, split == FALSE)

# train model
reg.mod2_train <- lm(log(SalePrice) ~ log(GrLivArea) + yjPower(TotalBsmtSF, 0.7396)
                + OverallQual + YearBuilt + GarageCars + BedroomAbvGr,
               data = training_set, subset = abs(reg.mod$resid) <= 0.8)

# test model
predict.test <- predict(reg.mod2_train, newdata = test_set)
accuracy(exp(predict.test), test_set$SalePrice)
```

```
##                 ME     RMSE      MAE       MPE     MAPE
## Test set 569.0149 28419.88 19245.39 -2.486041 11.81183
```

The average sale pice in dataset is $180921.20. It is plausible for our dataset to have a RMSE of $32706.32 on average.

```
result <- as.data.frame(cbind(test_set$GrLivArea, test_set$SalePrice, exp(predict.test)))
names(result) <- c("GrLivArea", "SalePrice", "SalePrice.hat")

ggplot(result, aes(x=GrLivArea)) + geom_point(aes(y=SalePrice)) + geom_line(aes(y=SalePrice, color="red
```



From the plot, we can figure out the predicted sale price is close to the real sale price, although the real data is more fluctuate than predicted data. Therefore, the model 2 performs good.

Question 2

Assume a healthcare insurance company hired you as a consultant to develop an econometric model to estimate the number of doctor visits a patient has over a 3 month period. The rational behind this study is that patients with a higher number of doctors visits wold pose a higher liability in terms of insurance expenses, and therefore, this may be mitigated via a higher insurance premium. The panel data are from the German Health Care Usage Dataset, and consist of 7,293 individuals across varying numbers of periods with a total of 27,326 observations.

```
#import data
health <- read.csv("german_healthcare_usage.csv", header = TRUE)
attach(health)
```

(a)

The number of people visiting doctors last three month is a reflection of their health. So, the basic indicators for health are people's health satisfaction, degree of handicap and their age.

Specifically, I purpose age has a positive quadractic term. It is because that when young people growing up, their immune systems getting stronger so that they seldom visit doctors and the aged people would more frequently visit doctors as their ages increase. And the numer of person go to hosipital lase year also can indirectly reflect their health conditions.

Besides, people who go to hospital more frequently last year might also have a insurance so that they can pay by the insurance. Therefore, the number of hospital visiting is associated with whether a person has public insurance. An interaction is necessary to add in our model.

The health conditions might differ from group, such as different type of occupations(blue collar, white collar and selfemployed) and gender(female and male). We can add these indicator variables into our model to figure out the relationship among groups.

```
reg.model <- lm(DOCVIS ~ AGE + I(AGE^2) + HOSPVIS + NEWHSAT + HANDPER + PUBLIC + HOSPVIS:PUBLIC + FEMALE
                + BLUEC + WHITEC, data = health)
summary(reg.model)
```

```
##
## Call:
## lm(formula = DOCVIS ~ AGE + I(AGE^2) + HOSPVIS + NEWHSAT + HANDPER +
##     PUBLIC + HOSPVIS:PUBLIC + FEMALE + BLUEC + WHITEC, data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.801  -2.326  -0.876   0.779 112.121
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.4862384  0.5409701  17.536  < 2e-16 ***
## AGE            -0.1185023  0.0242336  -4.890 1.01e-06 ***
## I(AGE^2)        0.0015498  0.0002756   5.624 1.89e-08 ***
## HOSPVIS         0.2617937  0.1050988   2.491 0.012747 *
## NEWHSAT        -0.7790718  0.0147657 -52.762  < 2e-16 ***
## HANDPER         0.0296172  0.0017858  16.585  < 2e-16 ***
## PUBLIC          0.3082253  0.1077906   2.859 0.004247 **
## FEMALE          0.9417299  0.0665076  14.160  < 2e-16 ***
## BLUEC          -0.0268950  0.0880348  -0.306 0.759985
```

```
## WHITEC         -0.0109302  0.0798304  -0.137 0.891096
## HOSPVIS:PUBLIC  0.3859590  0.1116121   3.458 0.000545 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.192 on 27312 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.1673, Adjusted R-squared:  0.167
## F-statistic: 548.8 on 10 and 27312 DF,  p-value: < 2.2e-16
```

Although the ajusted R^2 is not good, the significance of coefficients, p-value and degree of freedom perform good.

As health satisfication increases, the number of visiting doctors decreases. As the degree of handicap increases, the number of visiting doctors increases.

Just as I purposed, age needs an quadratic term and the coefficient is positive. There is no significant difference between different type of occupation but do have difference between men and women. The regression indicates females visit doctors more frequently than males by 0.94 times.

Additional 1 time visiting the hospital last year is associated with an increase of the number of visiting doctors last three month by 0.26 times on average. And people with public insurance would visit doctors more frequently in last three month by 0.308 times compared to people without public health insurance. Addtional 1 time of people with an insurance go to hosptial last year is associated with 0.39 times more visiting doctors last three month compared to people wo don't have public insurance.

(b)

i.

```
health$POLICY[health$YEAR < 1987] = 0
health$POLICY[health$YEAR >= 1987] = 1
```

$$DOCVIS = \beta_1 + \delta_1 FEMALE + \delta_2 POLICY + \gamma FEMALE * POLCY + e$$

γ is the difference-in-difference estimator.

```
did.mod1 <- lm(DOCVIS ~ FEMALE + POLICY + FEMALE:POLICY, data = health)
summary(did.mod1)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + POLICY + FEMALE:POLICY, data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -3.916  -2.641  -1.641   0.385 118.359
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.64095    0.07329  36.033   <2e-16 ***
## FEMALE         1.27464    0.10583  12.045   <2e-16 ***
## POLICY        -0.02622    0.09613  -0.273    0.785
## FEMALE:POLICY -0.18852    0.13889  -1.357    0.175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 5.66 on 27321 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.01064,    Adjusted R-squared:  0.01053
## F-statistic: 97.92 on 3 and 27321 DF,  p-value: < 2.2e-16
```

The coefficients of POLICY and FEMALE:POLICY are not statistical significant. The policy did not work for women.

ii.

$$DOCVIS = \beta_1 + \delta_1 UNEMPOLY + \delta_2 POLICY + \gamma UNEMPOLY * POLICY + e$$

```
did.mod2 <- lm(DOCVIS ~ UNEMPLOY + POLICY + UNEMPLOY:POLICY, data = health)
summary(did.mod2)
```

```
##
## Call:
## lm(formula = DOCVIS ~ UNEMPLOY + POLICY + UNEMPLOY:POLICY, data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -4.306  -2.722  -1.722   0.325 116.694
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.67495    0.06564  40.754   <2e-16 ***
## UNEMPLOY         1.63062    0.11030  14.783   <2e-16 ***
## POLICY           0.04671    0.08474   0.551   0.5815
## UNEMPLOY:POLICY -0.25872    0.14741  -1.755   0.0793 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.647 on 27321 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.01508,    Adjusted R-squared:  0.01497
## F-statistic: 139.4 on 3 and 27321 DF,  p-value: < 2.2e-16
```

The coefficients of POLICY and UNEMPOLY:POLICY are not statistical significant. The policy did not work for unemployed.

(c)

$$H_0 : \beta_{female} > 0 \quad H_1 : \beta_{female} \leq 0$$

```
mod <- lm(DOCVIS ~ FEMALE, data = health)
summary(mod)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE, data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
##  -3.791  -2.626  -1.626   0.374 118.374
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.62571    0.04743   55.36   <2e-16 ***
## FEMALE       1.16538    0.06854   17.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.66 on 27323 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.01047,    Adjusted R-squared:  0.01043
## F-statistic: 289.1 on 1 and 27323 DF,  p-value: < 2.2e-16
```

```
linearHypothesis(mod, "FEMALE = 1.16538")
```

```
## Linear hypothesis test
##
## Hypothesis:
## FEMALE = 1.16538
##
## Model 1: restricted model
## Model 2: DOCVIS ~ FEMALE
##
##   Res.Df    RSS Df  Sum of Sq  F Pr(>F)
## 1  27324 875310
## 2  27323 875310  1 1.0291e-07  0      1
```

We fail to reject null hypothesis so that women go to doctors more times than men on average.

(d)

I want to test the number of doctor visiting last 2 month if a person's degree of health satisfication decreases by 1 level and the number of the hospital visiting last year increases by 2 times.

```
mod2 <- lm(DOCVIS ~ NEWHSAT + HOSPVIS, data= health)

summary(glht(mod2, linfct=c( "-1*NEWHSAT + 2*HOSPVIS = 0")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = DOCVIS ~ NEWHSAT + HOSPVIS, data = health)
##
## Linear Hypotheses:
##                                  Estimate Std. Error t value Pr(>|t|)
## -1 * NEWHSAT + 2 * HOSPVIS == 0  2.15147    0.07219    29.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

We estimate that the number of doctor visiting would increases 2.15 times on average.