Final Project

Chang Cong, Hanlin Liu, Qiushi Wang

2019/6/3

```
library(knitr)
library(ggplot2)
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(scales)
library(Rmisc)
library(ggrepel)
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:randomForest':
##
##     outlier

## The following objects are masked from 'package:scales':
##
##     alpha, rescale

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(xgboost)
```

```
##
## Attaching package: 'xgboost'

## The following object is masked from 'package:dplyr':
##
##     slice
```

```
library(ggplot2)
library(readr)
```

```
##
## Attaching package: 'readr'
```

```
## The following object is masked from 'package:scales':
##
##     col_factor
```

```
library(gplots)
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(repr)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
setwd("~/Desktop/ChangCongCapstone/dataset")
library(readr)
data <- read.csv("train.csv",stringsAsFactors = F)
set.seed(111)
```

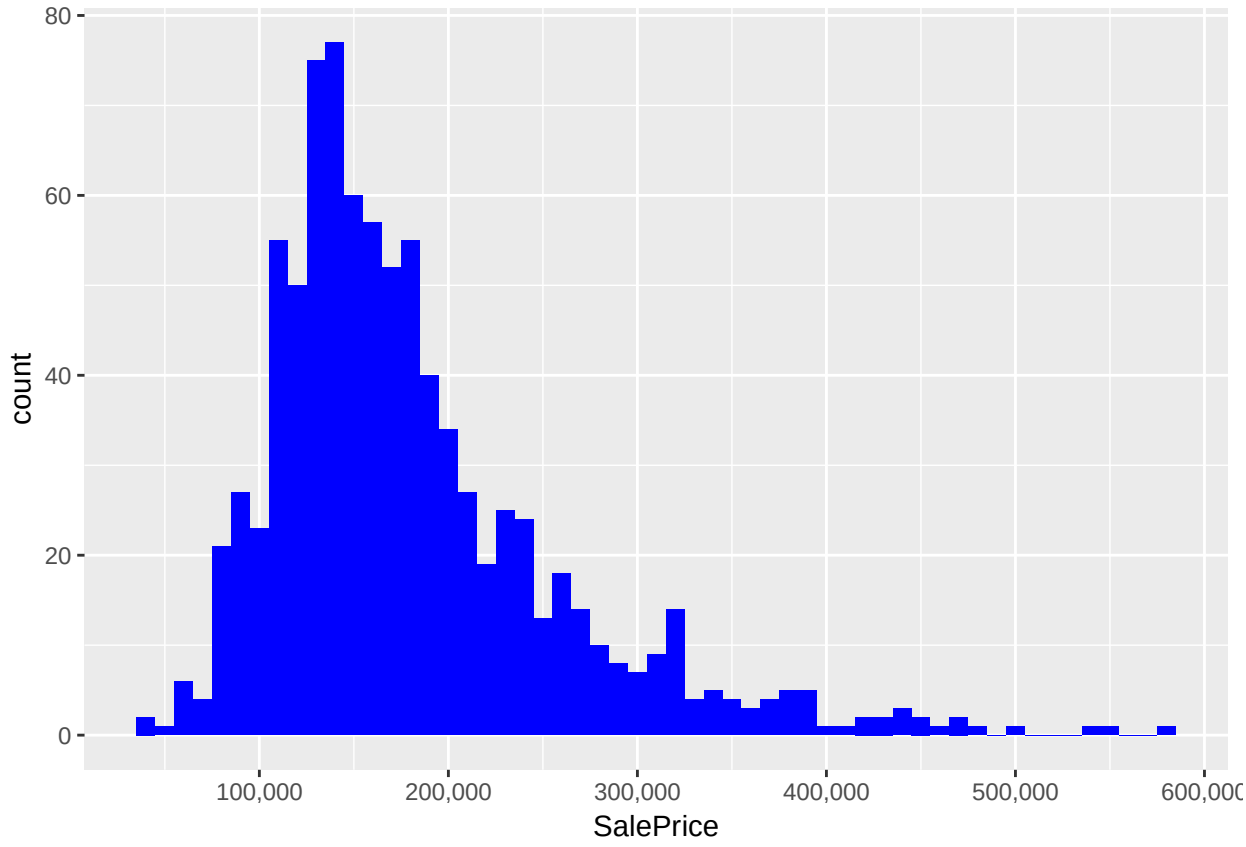
```
var_name <- names(data)
select_var <- c('MSZoning','Utilities','Neighborhood','BldgType','HouseStyle','OverallQual','OverallCon
                'BsmtQual','BsmtCond','TotalBsmtSF','Heating','HeatingQC',
                'CentralAir','Electrical','GrLivArea','BedroomAbvGr','KitchenAbvGr','KitchenQual','TotR
               'GarageArea','GarageQual','GarageCond','OpenPorchSF','PoolArea',
                'Fence','MoSold','YrSold','SaleType','SaleCondition','SalePrice')
select_train <- data[,select_var]
select_train$logPrice <- log(select_train$SalePrice)

train.index <- sample(row.names(select_train), 0.6*dim(data)[1])
valid.index <- setdiff(row.names(select_train), train.index)
train.df <- select_train[train.index, ]
valid.df <- select_train[valid.index, ]
write.csv(train.df, file = "trainselected.csv")
write.csv(valid.df, file = "validselected.csv")
```

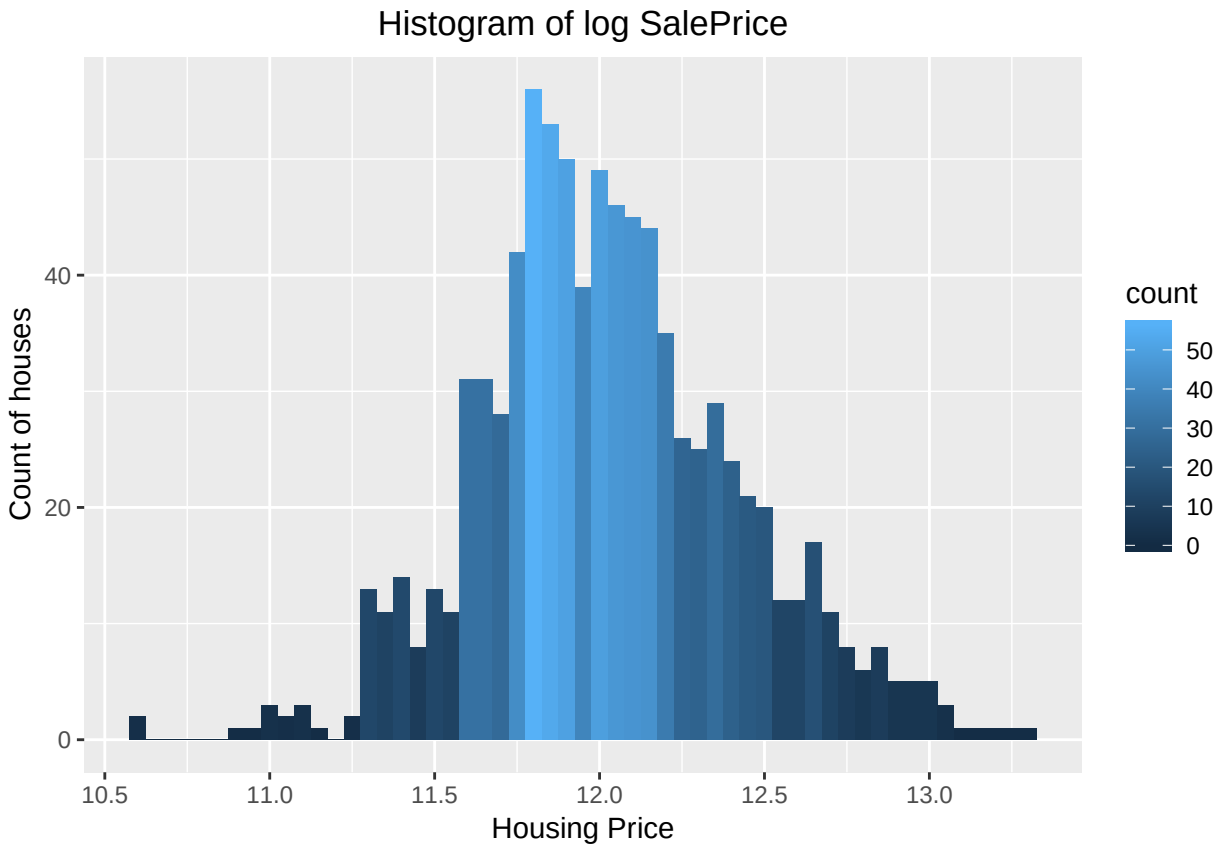
```
summary(select_train$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   39300  130000  161875  181177  213625  582933
```

```
ggplot(data=select_train[!is.na(select_train$SalePrice),], aes(x=SalePrice)) +
        geom_histogram(fill="blue", binwidth = 10000) +
        scale_x_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```




```
ggplot(select_train, aes(x = logPrice, fill = ..count..)) +
  geom_histogram(binwidth = 0.05) +
  ggtitle("Histogram of log SalePrice") +
  ylab("Count of houses") +
  xlab("Housing Price") +
  theme(plot.title = element_text(hjust = 0.5))
```

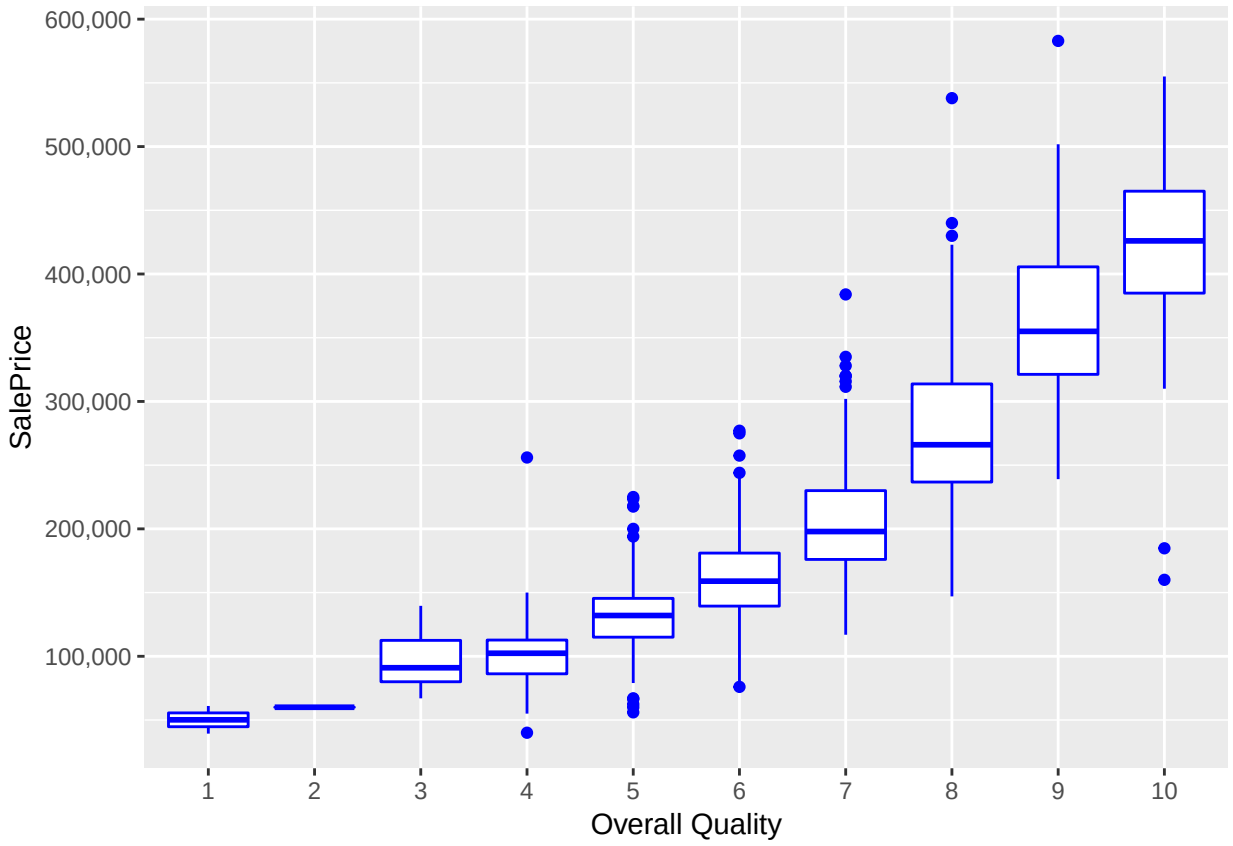



```
cor_numVar <- cor(select_train[,-37], use="pairwise.complete.obs")
```

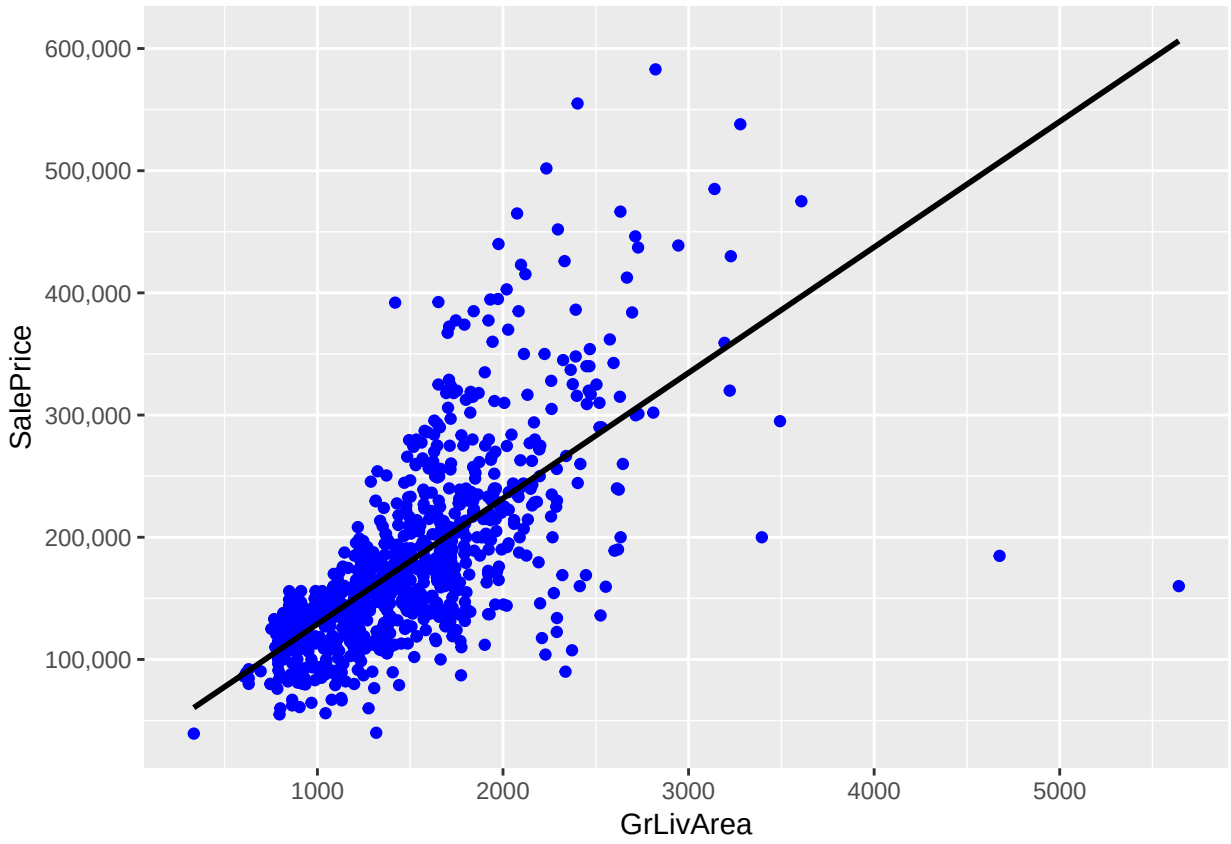
```
## Warning in cor(select_train[, -37], use = "pairwise.complete.obs"): the
## standard deviation is zero
```

```
#sort on decreasing correlations with SalePrice
cor_sorted <- as.matrix(sort(cor_numVar[,'SalePrice'], decreasing = TRUE))
 #select only high corelations
CorHigh <- names(which(apply(cor_sorted, 1, function(x) abs(x)>0.6)))
cor_numVar <- cor_numVar[CorHigh, CorHigh]
corrplot.mixed(cor_numVar, tl.col="black", tl.pos = "lt")
```

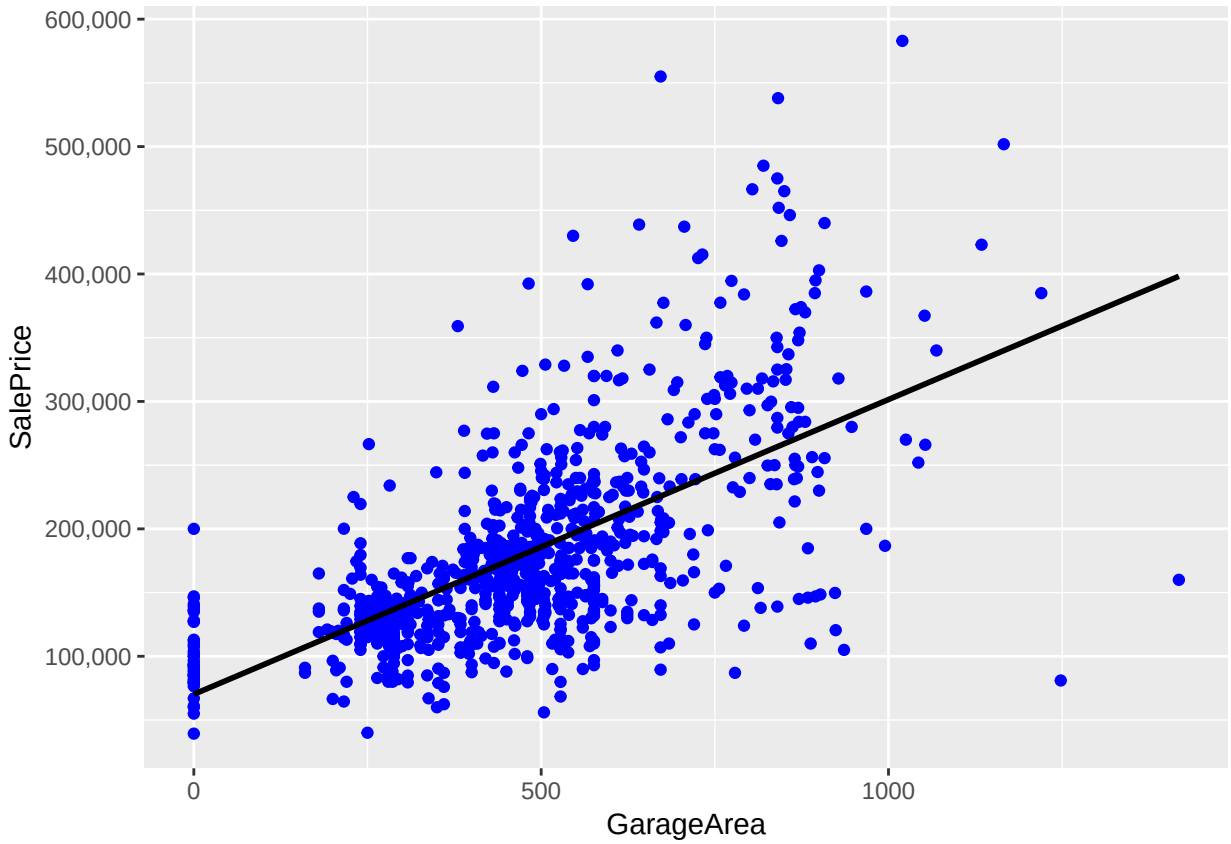

```
ggplot(data=select_train[!is.na(select_train$SalePrice),], aes(x=factor(OverallQual), y=SalePrice))+
        geom_boxplot(col='blue') + labs(x='Overall Quality') +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```




```
ggplot(data=select_train[!is.na(select_train$SalePrice),], aes(x=GrLivArea, y=SalePrice))+
        geom_point(col='blue') + geom_smooth(method = "lm", se=FALSE, color="black", aes(group=1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```




```
ggplot(data=select_train[!is.na(select_train$SalePrice),], aes(x=GarageArea, y=SalePrice))+
        geom_point(col='blue') + geom_smooth(method = "lm", se=FALSE, color="black", aes(group=1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```



Linear Regression

```
linreg <- lm(logPrice~.-SalePrice, data = train.df)
summary(linreg)
```

```
##
## Call:
## lm(formula = logPrice ~ . - SalePrice, data = train.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74567 -0.07247  0.00325  0.07960  0.66460
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.674e+01  1.128e+01   1.483   0.1386
## MSZoning      -2.016e-02  1.017e-02  -1.981   0.0481 *
## Utilities             NA         NA      NA       NA
## Neighborhood  -2.124e-03  1.361e-03  -1.560   0.1193
## BldgType      -1.760e-02  8.668e-03  -2.031   0.0428 *
## HouseStyle     5.357e-03  6.118e-03   0.876   0.3816
## OverallQual    8.740e-02  9.062e-03   9.644  < 2e-16 ***
## OverallCond    4.730e-02  7.808e-03   6.057 2.76e-09 ***
## YearBuilt      3.673e-03  4.320e-04   8.502 2.27e-16 ***
## ExterQual     -2.633e-02  1.605e-02  -1.640   0.1016
## ExterCond     -9.844e-03  1.717e-02  -0.573   0.5666
## BsmtQual       1.197e-02  1.003e-02   1.193   0.2335
```

```
## BsmtCond      -2.167e-02  1.281e-02  -1.692   0.0913 .
## TotalBsmtSF    4.028e-05  2.073e-05   1.943   0.0526 .
## Heating        1.854e-03  3.929e-02   0.047   0.9624
## HeatingQC     -1.074e-02  9.601e-03  -1.119   0.2639
## CentralAir    -4.734e-02  4.480e-02  -1.057   0.2911
## Electrical    -1.936e-02  1.532e-02  -1.264   0.2069
## GrLivArea      2.278e-04  3.239e-05   7.031 6.93e-12 ***
## BedroomAbvGr  -1.915e-02  1.331e-02  -1.439   0.1507
## KitchenAbvGr  -2.887e-02  4.140e-02  -0.697   0.4859
## KitchenQual    2.623e-03  1.276e-02   0.206   0.8371
## TotRmsAbvGrd   1.012e-02  9.773e-03   1.036   0.3008
## Functional    -1.026e-02  1.062e-02  -0.965   0.3348
## Fireplaces     4.161e-02  1.799e-02   2.313   0.0211 *
## FireplaceQu    1.921e-02  9.734e-03   1.974   0.0490 *
## GarageArea     2.004e-04  4.899e-05   4.091 5.01e-05 ***
## GarageQual    -2.257e-02  1.730e-02  -1.304   0.1927
## GarageCond     3.254e-02  2.089e-02   1.557   0.1200
## OpenPorchSF    2.049e-04  1.204e-04   1.702   0.0894 .
## PoolArea      -7.022e-04  1.637e-04  -4.290 2.15e-05 ***
## Fence         -6.679e-03  1.045e-02  -0.639   0.5229
## MoSold        -6.620e-04  2.669e-03  -0.248   0.8042
## YrSold        -6.519e-03  5.606e-03  -1.163   0.2454
## SaleType       9.607e-03  1.045e-02   0.919   0.3583
## SaleCondition  1.259e-02  9.062e-03   1.389   0.1655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.159 on 490 degrees of freedom
## Multiple R-squared:  0.8462, Adjusted R-squared:  0.8356
## F-statistic: 79.31 on 34 and 490 DF,  p-value: < 2.2e-16
```

```
backward<-stepAIC(linreg,direction='backward',trace=FALSE)
summary(backward)
```

```
##
## Call:
## lm(formula = logPrice ~ MSZoning + BldgType + OverallQual + OverallCond +
##     YearBuilt + BsmtCond + TotalBsmtSF + HeatingQC + GrLivArea +
##     Fireplaces + FireplaceQu + GarageArea + GarageQual + GarageCond +
##     OpenPorchSF + PoolArea + YrSold + SaleCondition, data = train.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74021 -0.07817  0.00518  0.08375  0.66728
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.191e+01  1.077e+01   2.034  0.04248 *
## MSZoning      -2.082e-02  9.784e-03  -2.128  0.03383 *
## BldgType      -1.902e-02  7.178e-03  -2.649  0.00831 **
## OverallQual    9.450e-02  8.214e-03  11.505  < 2e-16 ***
## OverallCond    5.000e-02  7.253e-03   6.893 1.63e-11 ***
## YearBuilt      3.864e-03  3.719e-04  10.389  < 2e-16 ***
## BsmtCond      -2.344e-02  1.157e-02  -2.026  0.04324 *
```

```
## TotalBsmtSF    4.092e-05  1.977e-05   2.069  0.03902 *
## HeatingQC     -1.800e-02  8.778e-03  -2.051  0.04081 *
## GrLivArea      2.197e-04  1.999e-05  10.990  < 2e-16 ***
## Fireplaces     4.602e-02  1.724e-02   2.670  0.00783 **
## FireplaceQu    1.816e-02  9.465e-03   1.918  0.05563 .
## GarageArea     2.188e-04  4.638e-05   4.716 3.11e-06 ***
## GarageQual    -2.435e-02  1.651e-02  -1.476  0.14069
## GarageCond     3.183e-02  2.016e-02   1.579  0.11498
## OpenPorchSF    2.227e-04  1.178e-04   1.891  0.05918 .
## PoolArea      -7.191e-04  1.566e-04  -4.592 5.54e-06 ***
## YrSold        -9.373e-03  5.352e-03  -1.751  0.08049 .
## SaleCondition  1.397e-02  8.542e-03   1.635  0.10267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1587 on 506 degrees of freedom
## Multiple R-squared:  0.8418, Adjusted R-squared:  0.8362
## F-statistic: 149.6 on 18 and 506 DF,  p-value: < 2.2e-16
```

```
library(forecast)
library(ModelMetrics)
```

```
##
## Attaching package: 'ModelMetrics'

## The following objects are masked from 'package:caret':
##
##     confusionMatrix, precision, recall, sensitivity, specificity

## The following object is masked from 'package:base':
##
##     kappa
```

```
#use predict() to make prediction on a new set
pred1 <- predict(backward,valid.df,type = "response")
residuals <-valid.df$logPrice - pred1
linreg_pred <- data.frame("Predicted" = pred1, "Actual" = valid.df$logPrice, "Residual" = residuals)
accuracy(pred1, valid.df$logPrice)
```

```
##                    ME      RMSE       MAE         MPE      MAPE
## Test set -0.009858649 0.1668785 0.1084131 -0.09887042 0.9071223
```

```
rmse(pred1, valid.df$logPrice)
```

```
## [1] 0.1668785
```