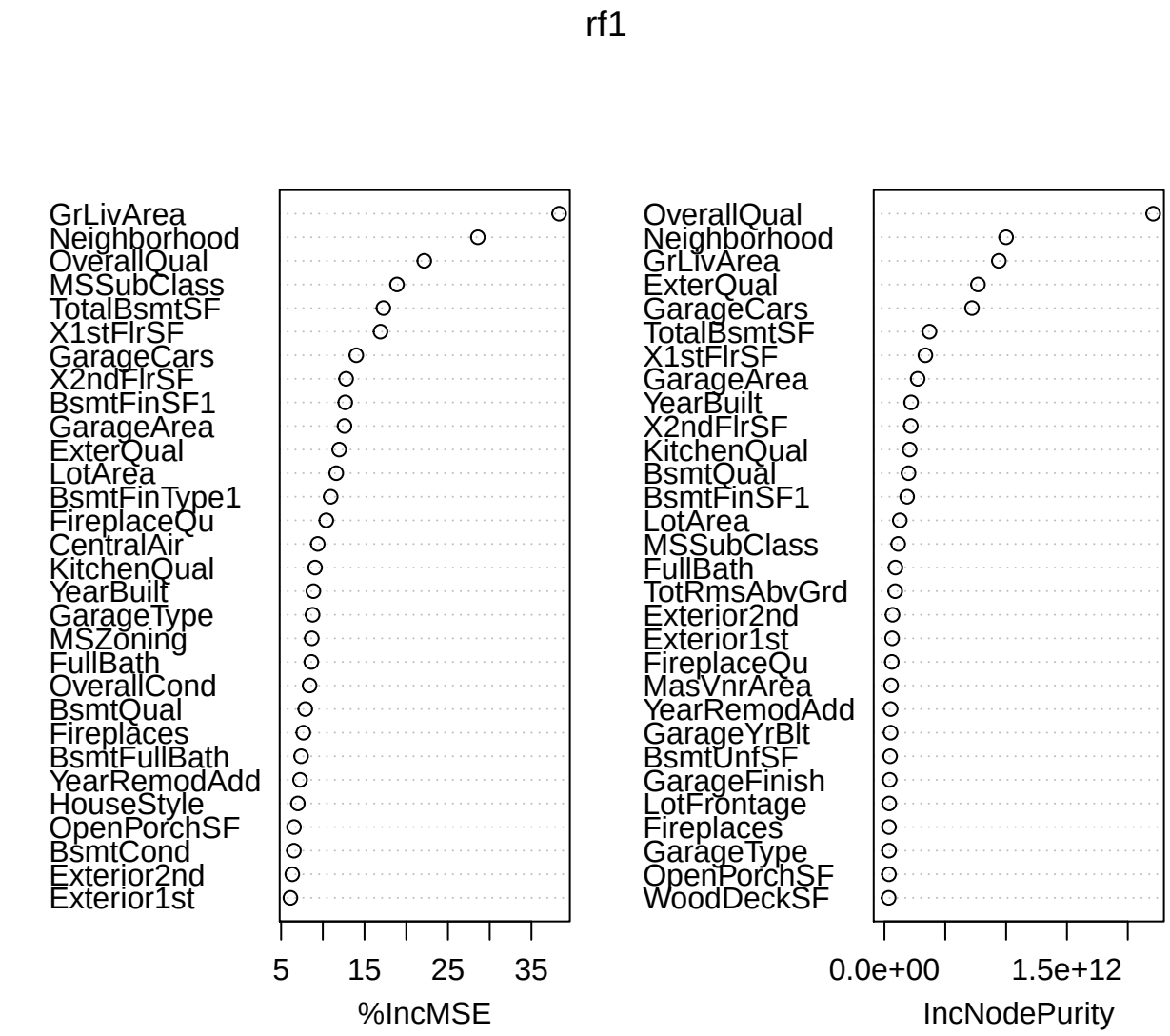
# Project_01

*Group_1*

*March 21, 2017*

```
rf1 = randomForest(x = train[, 1:79], y = train$SalePrice, importance = TRUE,
    data = train)

varImpPlot(rf1)
```




```
# testpredrf1 = predict(rf1, test) testpredrf1 = cbind(1461:2919,
# testpredrf1) colnames(testpredrf1) = c('Id', 'SalePrice')
# write.csv(testpredrf1, file = 'testpredrf1.csv', row.names = FALSE)

# sum(sapply(train[, 1:79], class) == sapply(test[, 1:79], class))
```

```
mse = function(linmod) sum(linmod$residuals^2)/linmod$df.residual
```

Created a function to calculate the MSE for our linear models.

```
# Linear model using important variables from RF selection IncNodePurity
lm2 = lm(SalePrice ~ OverallQual + Neighborhood + GrLivArea + ExterQual + GarageCars,
    data = train)
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: SalePrice
##                Df     Sum Sq    Mean Sq  F value    Pr(>F)
## OverallQual     1 5.7609e+12 5.7609e+12 4745.132 < 2.2e-16 ***
## Neighborhood   24 8.0705e+11 3.3627e+10   27.698 < 2.2e-16 ***
## GrLivArea       1 6.7724e+11 6.7724e+11  557.820 < 2.2e-16 ***
## ExterQual       3 1.4241e+11 4.7471e+10   39.100 < 2.2e-16 ***
## GarageCars      1 8.5355e+10 8.5355e+10   70.304 < 2.2e-16 ***
## Residuals    1429 1.7349e+12 1.2141e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mse(lm2)
```

```
## [1] 1214075108
```

```
sqrt(mse(lm2))
```

```
## [1] 34843.58
```

```
summary(lm2)[8]
```

```
## $r.squared
## [1] 0.8115845
```

```
summary(lm2)[9]
```

```
## $adj.r.squared
## [1] 0.8076289
```

```
Sales.xval = rep(0, nrow(train))
xvs = rep(1:10, length = nrow(train))
xvs = sample(xvs)
for (i in 1:10) {
    xvstest = train[xvs == i, ]
    xvstrain = train[xvs != i, ]
    glub = lm(SalePrice ~ OverallQual + Neighborhood + GrLivArea + ExterQual +
        GarageCars, data = xvstrain)
    Sales.xval[xvs == i] = predict(glub, xvstest)
    if (i == 10)
        print(sum((train$SalePrice - Sales.xval)^2)/glub$df.residual)
}
```

```
## [1] 1451958954
```

```
testpred2 = predict(lm2, test)
length(testpred2)
```

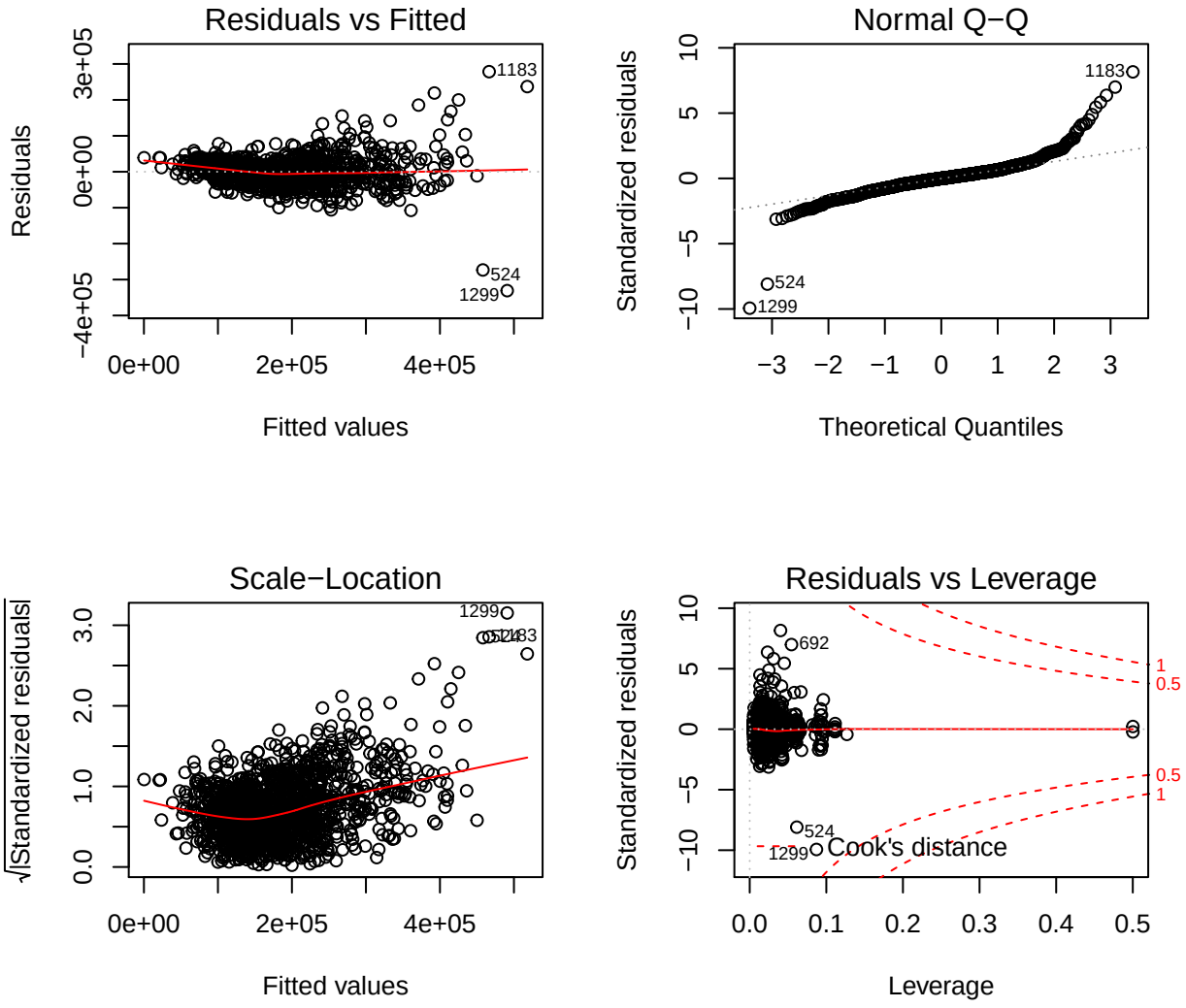
```
## [1] 1459
```

```r
length(1461:2919)
```

```
## [1] 1459
```

```r
testpred2 = cbind(1461:2919, testpred2)
colnames(testpred2) = c("Id", "SalePrice")
# write.csv(testpred2, file = 'testpred2.csv', row.names = FALSE)

par(mfrow = c(2, 2))
plot(lm2)
```




After performing a Random Forest variable selection, the IncNodePurity variable importance plot showed the top five variables of OverallQual, Neighborhood, GrLivArea, ExterQual, GarageCars. These make sense intuitively as good predictors for the sale price of a home as stated above. We ran a linear model, called lm2, and calculated the root MSE as 34843.58. Although this linear model violates the assumptions of normality and nonconstant variance as seen in the diagnostic plots, we accept it as a baseline for futher models. We submitted an entry to Kaggle, and ranked 1769 out of 2055. 86% of entries have a lower root mean squared logged error. Additionally, the linear model's $R^2 = 0.812$, and the adjusted $R^2 = 0.808$.