# dong_chris_housing

*Chris Dong*

*September 20, 2017*

## Loading the data and any packages

```
options("max.print"=5)
library(tidyverse)
library(magrittr)
library(leaps)
house <- read_csv("housing.txt")
names(house) <- tolower(names(house))
house$mssubclass <- factor(house$mssubclass)


house %>% sapply(function(x) sum(is.na(x))) %>% sort(decreasing = T)
```

```
##      poolqc miscfeature       alley       fence fireplacequ
##        1453        1406        1369        1179         690
##  [ reached getOption("max.print") -- omitted 76 entries ]
```

```
house$bsmtfintype1[which(is.na(house$bsmtfintype1))] <- 0
house$bsmtfintype2[which(is.na(house$bsmtfintype2))] <- 0
house$masvnrarea <- as.numeric(house$masvnrarea)
house$masvnrarea[which(is.na(house$masvnrarea))] <- 0

house$garageyrblt <- (house$garageyrblt > house$yearbuilt) * 1
house$garageyrblt[is.na(house$garageyrblt)] <- 0
```

Impute the NA in `lotfrontage`, `electrical` with K-Nearest Neighbors

```
#install.packages('VIM')
library(VIM)

k = round(sqrt(1460*.8) / 2)

house$lotfrontage <- kNN(house, variable = "lotfrontage",  k = k)$lotfrontage
house$electrical <- kNN(house, variable = "electrical",  k = k)$electrical

house[is.na(house)] <- "None"
```

Split the data into either numeric or categorical. When doing a linear model on only numerical variables, we find that two variables, `totalbsmtsf` and `grlivarea` are perfectly collinear with other variables so we remove them. If VIF is higher than 2, there is some collinearity problem.

```
#install.packages("fmsb")
library(car)
```

```
##
## Attaching package: 'car'
```

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

```
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
house$remodel <- T
house[house$yearbuilt == house$yearremodadd,]$remodel <- F
house %<>% select(-yearremodadd)

house$soldminusbuilt <- (house$yrsold - house$yearbuilt)
house %<>% select(-yrsold,-yearbuilt)


house$porcharea <- with(house, openporchsf + enclosedporch +
    `3ssnporch` + screenporch)


house %<>% select(-id)



house$lotshape <- (house$lotshape == 'Reg') *1

house_by_neighborhood <- house %>% group_by(neighborhood) %>% summarise(avgprc = median(saleprice)) %>%

house_by_neighborhood
```

```
## # A tibble: 25 x 2
##    neighborhood avgprc
##           <chr>  <dbl>
##  1      NridgHt 315000
##  2      NoRidge 301500
##  3      StoneBr 278000
##  4       Timber 228475
##  5      Somerst 225500
##  6      Veenker 218000
##  7      Crawfor 200624
##  8      ClearCr 200250
##  9      CollgCr 197200
## 10      Blmngtn 191000
## # ... with 15 more rows
```

```
library(Hmisc)
```

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

```
## Loading required package: survival
```

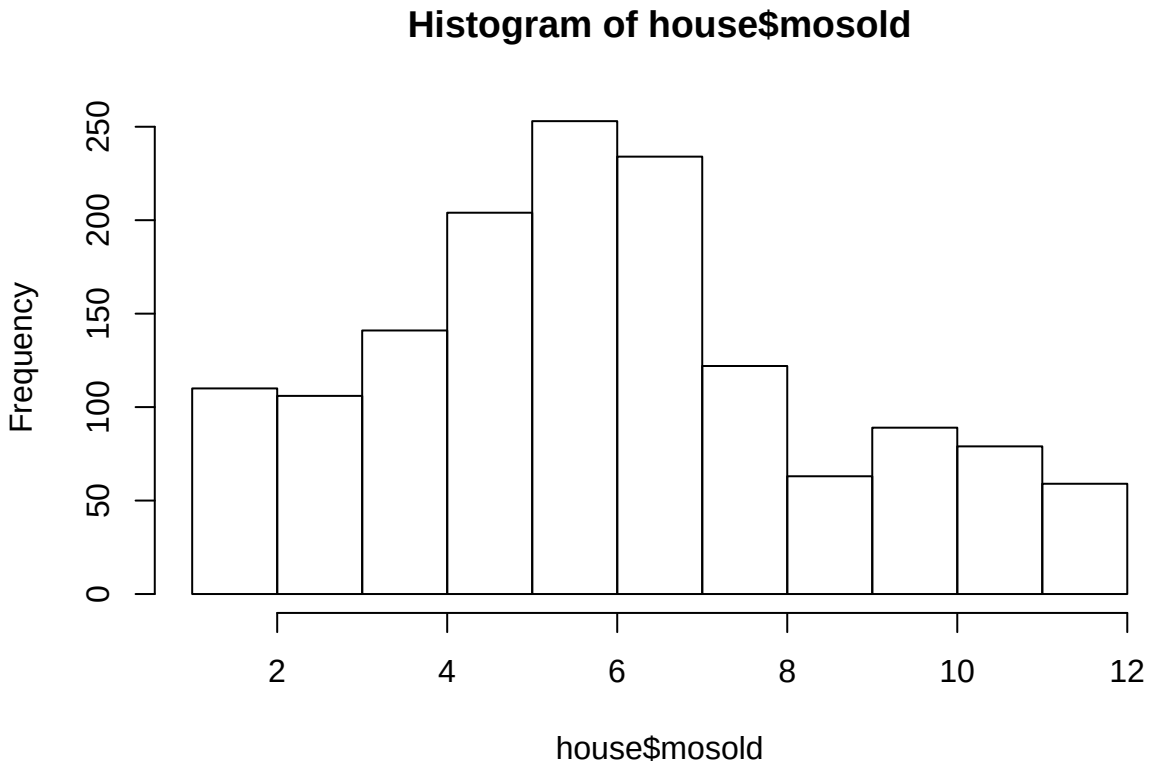
```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
house_by_neighborhood$pricecategory <- as.numeric(factor(cut2(house_by_neighborhood$avgprc, quantile(hou
       labels = 1:4))

house_by_neighborhood <- house_by_neighborhood[,-2]

house %<>% left_join(house_by_neighborhood, by = "neighborhood") %>% select(-neighborhood)

hist(house$mosold, bins=12)
```




```
house$summertime <- (house$mosold %in% 4:7) * 1

house %<>% select(-mosold, -landcontour, -alley)
house %<>% select(-lotshape)
```

```
house$lotconfig <- (house$lotconfig == "Inside")  * 1

house %<>% select(-lotconfig)

fullmodel <- lm(saleprice~sqrt(lotfrontage)+porcharea+.,data = house)
summary(fullmodel)
```

```
##
## Call:
## lm(formula = saleprice ~ sqrt(lotfrontage) + porcharea + ., data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -190224   -9965       0   10029  190224
##
## Coefficients: (10 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -9.771e+05  1.555e+05  -6.284 4.57e-10 ***
##  [ reached getOption("max.print") -- omitted 238 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23570 on 1231 degrees of freedom
## Multiple R-squared:  0.9257, Adjusted R-squared:  0.912
## F-statistic: 67.28 on 228 and 1231 DF,  p-value: < 2.2e-16
```

```
house$condition1 <- relevel(factor(house$condition1), ref = "Norm")

house$condition2 <- relevel(factor(house$condition2), ref = "Norm")

house %<>% select(-roofstyle)
house %<>% select(-exterior2nd)



table(house$bldgtype)
```

```
##
##   1Fam 2fmCon Duplex  Twnhs TwnhsE
##   1220     31     52     43    114
```

```
house %>% group_by(housestyle) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 8 x 2
##   housestyle avgprc
##        <chr>  <dbl>
## 1     2.5Fin 194000
## 2     2Story 190000
## 3       SLvl 164500
## 4     1Story 154750
## 5     SFoyer 135960
## 6     2.5Unf 133900
## 7     1.5Fin 132000
## 8     1.5Unf 111250
```

```
summary(lm(saleprice~housestyle,data=house))
```

```
##
## Call:
## lm(formula = saleprice ~ housestyle, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170052  -45502  -16519   26556  544948
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        143117       6134  23.332  < 2e-16 ***
##  [ reached getOption("max.print") -- omitted 7 rows ]
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76120 on 1452 degrees of freedom
## Multiple R-squared:  0.08631,    Adjusted R-squared:  0.08191
## F-statistic:  19.6 on 7 and 1452 DF,  p-value: < 2.2e-16
```

```
house <- house %>% select(-`1stflrsf`, -`2ndflrsf`, -lowqualfinsf,
    -totalbsmtsf, -openporchsf, -enclosedporch, - `3ssnporch`,
    - screenporch, -garagearea)

fullmodel <- lm(saleprice~sqrt(lotfrontage)+.,data = house)
summary(fullmodel)
```

```
##
## Call:
## lm(formula = saleprice ~ sqrt(lotfrontage) + ., data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197401  -10704       0   10023  197401
##
## Coefficients: (6 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -9.942e+05  1.491e+05  -6.670 3.82e-11 ***
##  [ reached getOption("max.print") -- omitted 209 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23790 on 1256 degrees of freedom
## Multiple R-squared:  0.9228, Adjusted R-squared:  0.9103
## F-statistic: 73.94 on 203 and 1256 DF,  p-value: < 2.2e-16
```

```
house %>% group_by(salecondition) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 6 x 2
##   salecondition avgprc
##           <chr>  <dbl>
## 1       Partial 244600
## 2        Normal 160000
## 3        Alloca 148145
## 4        Family 140500
## 5       Abnorml 130000
## 6       AdjLand 104000
```

```
house$salecondition <- (house$salecondition == "Normal") * 1

house %>% group_by(saletype) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 9 x 2
##   saletype avgprc
##      <chr>  <dbl>
## 1      Con 269600
## 2      New 247453
## 3      CWD 188750
## 4       WD 158000
```

```
## 5    ConLw 144000
## 6    ConLD 140000
## 7      COD 139000
## 8    ConLI 125000
## 9      Oth 116050
```

```
house$newtype <- (house$saletype == 'New') * 1

house <- house %>% select(-saletype)

house$miscfeature <- (house$miscfeature != 'None') * 1
house %<>% select(-miscval)
house %<>% select(-miscfeature)

house$paveddrive <- (house$paveddrive == 'Y') * 1
house %<>% select(-paveddrive)


house$poolqc <- (house$poolqc !="None")*1
house$fence <- (house$fence !="None")*1

house$garagecond <-  as.numeric(factor(house$garagecond,
    levels = c("None","Po","Fa","TA","Gd","Ex"), labels = 0:5))
house$garagequal <-  as.numeric(factor(house$garagequal,
    levels = c("None","Po","Fa","TA","Gd","Ex"), labels = 0:5))

glimpse(house)
```

```
## Observations: 1,460
## Variables: 62
## $ mssubclass     <fctr> 60, 20, 60, 70, 60, 50, 20, 60, 50, 190, 20, 6...
## $ mszoning       <chr> "RL", "RL", "RL", "RL", "RL", "RL", "RL", "RL",...
## $ lotfrontage    <int> 65, 80, 68, 60, 84, 85, 75, 71, 51, 50, 70, 85,...
## $ lotarea        <int> 8450, 9600, 11250, 9550, 14260, 14115, 10084, 1...
## $ street         <chr> "Pave", "Pave", "Pave", "Pave", "Pave", "Pave",...
## $ utilities      <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub...
## $ landslope      <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl"...
## $ condition1     <fctr> Norm, Feedr, Norm, Norm, Norm, Norm, Norm, Pos...
## $ condition2     <fctr> Norm, Norm, Norm, Norm, Norm, Norm, Norm, Norm...
## $ bldgtype       <chr> "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam",...
## $ housestyle     <chr> "2Story", "1Story", "2Story", "2Story", "2Story...
## $ overallqual    <int> 7, 6, 7, 7, 8, 5, 8, 7, 7, 5, 5, 9, 5, 7, 6, 7,...
## $ overallcond    <int> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8,...
## $ roofmatl       <chr> "CompShg", "CompShg", "CompShg", "CompShg", "Co...
## $ exterior1st    <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Sdng", "Vi...
## $ masvnrtype     <chr> "BrkFace", "None", "BrkFace", "None", "BrkFace"...
## $ masvnrarea     <dbl> 196, 0, 162, 0, 350, 0, 186, 240, 0, 0, 0, 286,...
## $ exterqual      <chr> "Gd", "TA", "Gd", "TA", "Gd", "TA", "Gd", "TA",...
## $ extercond      <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA",...
## $ foundation     <chr> "PConc", "CBlock", "PConc", "BrkTil", "PConc", ...
## $ bsmtqual       <chr> "Gd", "Gd", "Gd", "TA", "Gd", "Gd", "Ex", "Gd",...
## $ bsmtcond       <chr> "TA", "TA", "TA", "Gd", "TA", "TA", "TA", "TA",...
## $ bsmtexposure   <chr> "No", "Gd", "Mn", "No", "Av", "No", "Av", "Mn",...
## $ bsmtfintype1   <chr> "GLQ", "ALQ", "GLQ", "ALQ", "GLQ", "GLQ", "GLQ"...
## $ bsmtfinsf1     <int> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851...
```

```
## $ bsmtfintype2   <chr> "Unf", "Unf", "Unf", "Unf", "Unf", "Unf", "Unf"...
## $ bsmtfinsf2     <int> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0...
## $ bsmtunfsf      <int> 150, 284, 434, 540, 490, 64, 317, 216, 952, 140...
## $ heating        <chr> "GasA", "GasA", "GasA", "GasA", "GasA", "GasA",...
## $ heatingqc      <chr> "Ex", "Ex", "Ex", "Gd", "Ex", "Ex", "Ex", "Ex",...
## $ centralair     <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y...
## $ electrical     <chr> "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr", "S...
## $ grlivarea      <int> 1710, 1262, 1786, 1717, 2198, 1362, 1694, 2090,...
## $ bsmtfullbath   <int> 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0,...
## $ bsmthalfbath   <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ fullbath       <int> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1,...
## $ halfbath       <int> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,...
## $ bedroomabvgr   <int> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2,...
## $ kitchenabvgr   <int> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1,...
## $ kitchenqual    <chr> "Gd", "TA", "Gd", "Gd", "Gd", "TA", "Gd", "TA",...
## $ totrmsabvgrd   <int> 8, 6, 6, 7, 9, 5, 7, 7, 8, 5, 5, 11, 4, 7, 5, 5...
## $ functional     <chr> "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ"...
## $ fireplaces     <int> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0,...
## $ fireplacequ    <chr> "None", "TA", "TA", "Gd", "TA", "None", "Gd", "...
## $ garagetype     <chr> "Attchd", "Attchd", "Attchd", "Detchd", "Attchd...
## $ garageyrblt    <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,...
## $ garagefinish   <chr> "RFn", "RFn", "RFn", "Unf", "RFn", "Unf", "RFn"...
## $ garagecars     <int> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2,...
## $ garagequal     <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 3, 5, 4, 4, 4, 4, 4, 4,...
## $ garagecond     <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,...
## $ wooddecksf     <int> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147,...
## $ poolarea       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ poolqc         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ fence          <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,...
## $ salecondition  <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1,...
## $ saleprice      <int> 208500, 181500, 223500, 140000, 250000, 143000,...
## $ remodel        <lgl> FALSE, FALSE, TRUE, TRUE, FALSE, TRUE, TRUE, FA...
## $ soldminusbuilt <int> 5, 31, 7, 91, 8, 16, 3, 36, 77, 69, 43, 1, 46, ...
## $ porcharea      <int> 61, 0, 42, 307, 84, 350, 57, 432, 205, 4, 0, 21...
## $ pricecategory  <dbl> 3, 4, 3, 4, 4, 2, 4, 3, 1, 1, 2, 4, 2, 3, 2, 1,...
## $ summertime     <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1,...
## $ newtype        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0,...
```

```
house %<>% select(-fence,-poolqc,-garagecond)
```

```
house %>% group_by(garagefinish) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 4 x 2
##   garagefinish avgprc
##          <chr>  <dbl>
## 1          Fin 215000
## 2          RFn 190000
## 3          Unf 135000
## 4         None 100000
```

```
house$garagefinish <-(house$garagefinish == "Fin") *1
house %<>% select(-garagefinish)
```

```
#!diagnosticsoff
house %<>% select(-garageyrblt)

house$garagetype <- relevel(factor(house$garagetype), ref = "None")



house$fireplacequ <-  as.numeric(factor(house$fireplacequ,
    levels = c("None","Po","Fa","TA","Gd","Ex"), labels = 0:5))
cor(house$saleprice,house$fireplacequ); cor(house$saleprice,house$fireplaces)
```

```
## [1] 0.5204376
```

```
## [1] 0.4669288
```

```
house %<>% select(-fireplacequ, -fireplaces)

house$functional <- (house$functional == "Typ") * 1

house$kitchenqual <-  as.numeric(factor(house$kitchenqual,
    levels = c("Po","Fa","TA","Gd","Ex"), labels = 1:5))

cor(house$totrmsabvgrd ,house$saleprice);cor(house$grlivarea ,house$saleprice)
```

```
## [1] 0.5337232
```

```
## [1] 0.7086245
```

```
house %<>% select(-totrmsabvgrd)

fullmodel <- lm(saleprice~sqrt(lotfrontage)+.,data = house)
summary(fullmodel)
```

```
##
## Call:
## lm(formula = saleprice ~ sqrt(lotfrontage) + ., data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -201143  -10696       0   11118  201143
##
## Coefficients: (3 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -7.615e+05  6.021e+04 -12.648  < 2e-16 ***
##  [ reached getOption("max.print") -- omitted 157 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24530 on 1305 degrees of freedom
## Multiple R-squared:  0.9147, Adjusted R-squared:  0.9047
## F-statistic: 90.91 on 154 and 1305 DF,  p-value: < 2.2e-16
```

```
table(house$fullbath)
```

```
##
##   0   1   2   3
##   9 650 768  33
```

```
house$bath <- house$fullbath + house$halfbath + house$bsmtfullbath + house$bsmthalfbath
house %<>% select(-fullbath,-halfbath, -bsmthalfbath, -bsmtfullbath)

house %<>% select(-bath)
```

```
house %>% group_by(electrical) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 5 x 2
##   electrical avgprc
##        <chr>  <dbl>
## 1      SBrkr 170000
## 2      FuseA 121250
## 3      FuseF 115000
## 4      FuseP  82000
## 5        Mix  67000
```

```
house$electrical <- (house$electrical == "SBrkr") * 1

house %<>% select(-electrical, -centralair)

house$heatingqc <- as.numeric(factor(house$heatingqc,
                  levels = c("Po","Fa","TA","Gd","Ex"), labels = 1:5))

table(house$heatingqc)
```

```
##
##   1   2   3   4   5
##   1  49 428 241 741
```

```
house$heatingqc <- (house$heatingqc == 5) * 1

house %<>% select(-heating)

table(house$bsmtfintype1)
```

```
##
##   0 ALQ BLQ GLQ LwQ
##  37 220 148 418  74
##  [ reached getOption("max.print") -- omitted 2 entries ]
```

```
house$bsmtfintype1 <- as.numeric(factor(house$bsmtfintype1,
      levels = c("0","Unf","LwQ","Rec","BLQ","ALQ","GLQ"),
      labels = 0:6))
house$bsmtfintype2 <- as.numeric(factor(house$bsmtfintype2,
      levels = c("0","Unf","LwQ","Rec","BLQ","ALQ","GLQ"),
      labels = 0:6))
house$bsmtfintype1 <- house$bsmtfintype1 + house$bsmtfintype2
house %<>% select(-bsmtfintype1, -bsmtfintype2)

house$bsmtexposure <- relevel(factor(house$bsmtexposure), ref = "None")

table(house$bsmtexposure)
```

```
##
## None   Av   Gd   Mn   No
##   38  221  134  114  953
```

```
house %>% group_by(bsmtexposure) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 5 x 2
##   bsmtexposure avgprc
##         <fctr>  <dbl>
## 1           Gd 226975
## 2           Av 185850
## 3           Mn 182450
## 4           No 154000
## 5         None 104025
```

```
house$bsmtexposure <- (house$bsmtexposure == "Gd") * 1

house %>% group_by(bsmtcond) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 5 x 2
##   bsmtcond avgprc
##      <chr>  <dbl>
## 1       Gd 193879
## 2       TA 165000
## 3       Fa 118500
## 4     None 101800
## 5       Po  64000
```

```
table(house$bsmtcond)
```

```
##
##   Fa   Gd None   Po   TA
##   45   65   37    2 1311
```

```
house$bsmtcond <- as.numeric(factor(house$bsmtcond,
      levels = c("None","Po","Fa","TA","Gd","Ex"),
      labels = 0:5))

house$bsmtqual <- as.numeric(factor(house$bsmtqual,
      levels = c("None","Po","Fa","TA","Gd","Ex"),
      labels = 0:5))
cor(house$bsmtcond,house$bsmtqual)
```

```
## [1] 0.6337134
```

```
cor(house$bsmtcond,house$saleprice);cor(house$bsmtqual,house$saleprice)
```

```
## [1] 0.2126072
```

```
## [1] 0.5852072
```

```
house %<>% select(-bsmtcond)

house %<>% select(-bsmtqual)

table(house$foundation)
```

```
##
## BrkTil CBlock  PConc   Slab  Stone   Wood
##    146    634    647     24      6      3
```

```
house %>% group_by(foundation) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 6 x 2
##   foundation avgprc
##        <chr>  <dbl>
## 1      PConc 205000
## 2       Wood 164000
## 3     CBlock 141500
## 4      Stone 126500
## 5     BrkTil 125250
## 6       Slab 104150
```

```
house$foundation <- (house$foundation == "PConc")*1



house$extercond <- as.numeric(factor(house$extercond,
      levels = c("Po","Fa","TA","Gd","Ex"),
      labels = 1:5))
house$exterqual <- as.numeric(factor(house$exterqual,
      levels = c("Po","Fa","TA","Gd","Ex"),
      labels = 1:5))
cor(house$extercond,house$exterqual)
```

```
## [1] 0.00918398
```

```
house$masvnrtype <- relevel(factor(house$masvnrtype), ref = "None")

table(house$masvnrtype)
```

```
##
##    None  BrkCmn BrkFace   Stone
##     872      15     445     128
```

```
house$masvnrtype <- (house$masvnrtype != "None") * 1

house_by_exterior <- house %>%  group_by(exterior1st) %>% summarise(avgprc = median(saleprice)) %>% arra


house_by_exterior$exteriorcategory <- as.numeric(factor(cut2(house_by_exterior$avgprc, quantile(house_by
       labels = 1:4))

house_by_exterior <- house_by_exterior[,-2]

house %<>% left_join(house_by_exterior, by = "exterior1st") %>% select(-exterior1st)

house %<>% select(-exteriorcategory)


table(house$housestyle)
```

```
##
## 1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf
##    154     14    726      8     11
##  [ reached getOption("max.print") -- omitted 3 entries ]
```

```
house %>%  group_by(housestyle) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 8 x 2
```

```
##   housestyle avgprc
##        <chr>  <dbl>
## 1     2.5Fin 194000
## 2     2Story 190000
## 3       SLvl 164500
## 4     1Story 154750
## 5     SFoyer 135960
## 6     2.5Unf 133900
## 7     1.5Fin 132000
## 8     1.5Unf 111250
```

```
house$housestyle <- (house$housestyle == "2Story" | house$housestyle == "2.5Fin")*1


table(house$bldgtype)
```

```
##
##   1Fam 2fmCon Duplex  Twnhs TwnhsE
##   1220     31     52     43    114
```

```
house %>%  group_by(bldgtype) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 5 x 2
##   bldgtype avgprc
##      <chr>  <dbl>
## 1   TwnhsE 172200
## 2     1Fam 167900
## 3    Twnhs 137500
## 4   Duplex 135980
## 5   2fmCon 127500
```

```
house$bldgtype <- (house$bldgtype == "1Fam" | house$bldgtype == "2FmCon") * 1
house %<>% select(-bldgtype)

table(house$landslope)
```

```
##
##  Gtl  Mod  Sev
## 1382   65   13
```

```
house$landslope <- (house$landslope == "Gtl") * 1
house %<>% select(-landslope)

table(house$utilities)
```

```
##
## AllPub NoSeWa
##   1459      1
```

```
house %<>% select(-utilities, -street)

house %>%  group_by(mszoning) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 5 x 2
##   mszoning avgprc
##      <chr>  <dbl>
## 1       FV 205950
## 2       RL 174000
```

```
## 3       RH 136500
## 4       RM 120500
## 5  C (all)  74700
```

```
table(house$mszoning)
```

```
##
## C (all)      FV      RH      RL      RM
##      10      65      16    1151     218
```

```
house$mszoning <- relevel(factor(house$mszoning), ref = "RL")

house %<>% select(-mszoning)

house %>%  group_by(mssubclass) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 15 x 2
##    mssubclass avgprc
##         <fctr>  <dbl>
##  1         60 215200
##  2        120 192000
##  3         80 166500
##  4         75 163500
##  5         20 159250
##  6         70 156000
##  7        160 146000
##  8         40 142500
##  9         85 140750
## 10         90 135980
## 11         50 132000
## 12        190 128250
## 13         45 107500
## 14         30  99900
## 15        180  88500
```

```
house %<>% select(-mssubclass, -lotfrontage, -porcharea, -extercond,-foundation)


house %>%  group_by(condition1) %>% summarise(avgprc = median(saleprice)) %>% arrange(desc(avgprc))
```

```
## # A tibble: 9 x 2
##   condition1 avgprc
##       <fctr>  <dbl>
## 1       RRNn 214000
## 2       PosA 212500
## 3       PosN 200000
## 4       RRNe 190750
## 5       RRAn 171495
## 6       Norm 166500
## 7       RRAe 142500
## 8      Feedr 140000
## 9     Artery 119550
```

```
house$condition1 <- (house$condition1 == "Artery" | house$condition1 =="Feedr"|
  house$condition1 == "RRAe")*1
house$condition2 <- (house$condition2 == "PosN") * 1
```

```r
cor(house$garagequal, house$garagecars)
```

```
## [1] 0.5766224
```

```r
house %<>% select(-garagequal)

fullmodel <- lm(saleprice~.,data = house)
summary(fullmodel)
```

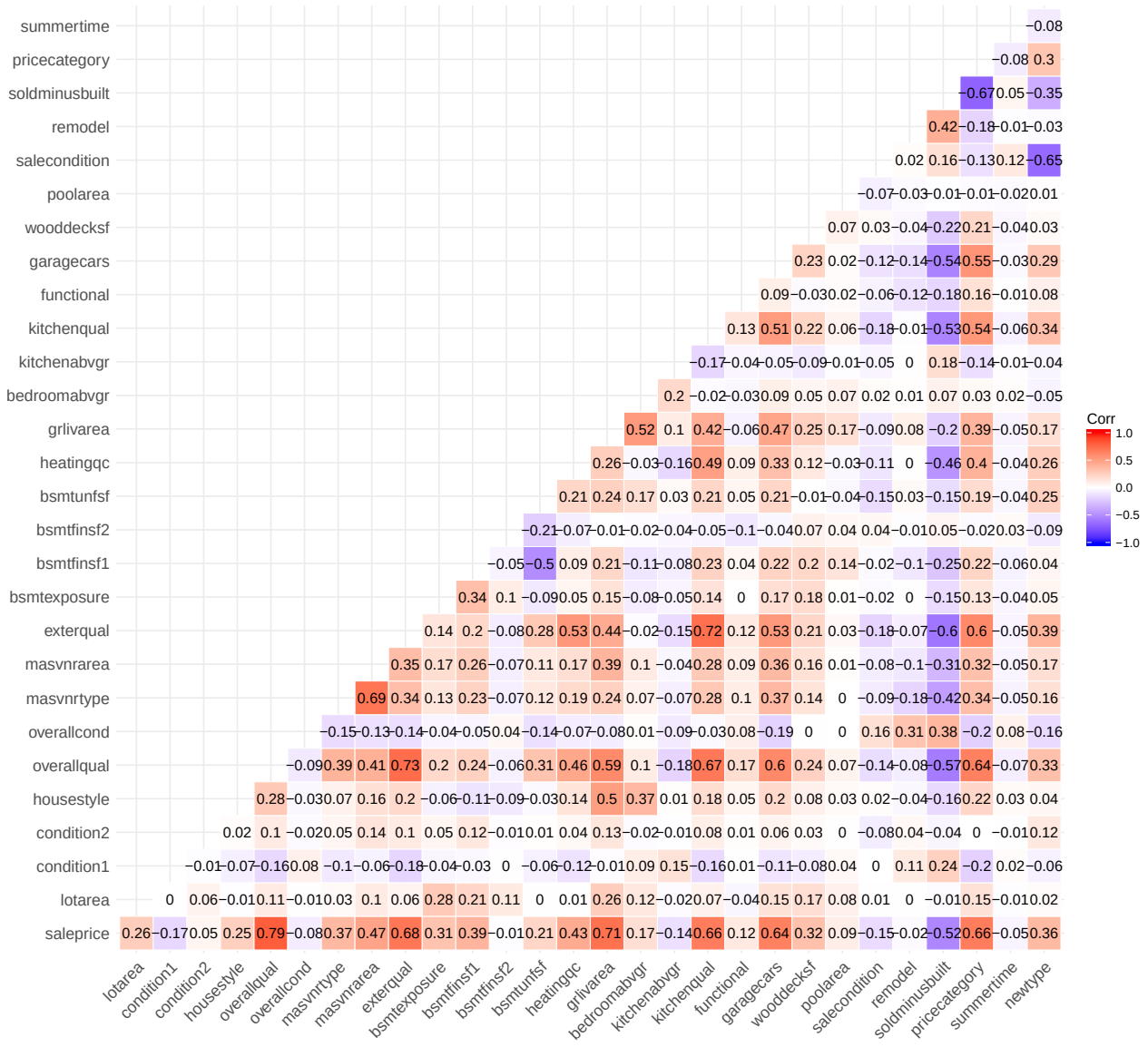
```
##
## Call:
## lm(formula = saleprice ~ ., data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197807  -13366     209   12887  202564
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.724e+05  3.366e+04 -22.948  < 2e-16 ***
##  [ reached getOption("max.print") -- omitted 41 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27030 on 1418 degrees of freedom
## Multiple R-squared:  0.8875, Adjusted R-squared:  0.8842
## F-statistic: 272.7 on 41 and 1418 DF,  p-value: < 2.2e-16
```

Checking multicollinearity. Looks good. For the generalized variance inflation factor (normalized by the degree of freedom), everything except one is less than 2.

```r
vif(fullmodel)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## lotarea        1.285538  1        1.133816
##  [ reached getOption("max.print") -- omitted 29 rows ]
```

```r
house$remodel <- as.numeric(house$remodel)

house_numeric <- house[,sapply(house,function(x) is.numeric(x))]

house_numeric %<>% select(saleprice, everything())
#install.packages("ggcorrplot")

library(ggcorrplot)

cor_matrix <- cor(house_numeric)

ggcorrplot(cor_matrix, type = "lower", outline.col = "white",
           lab = T, insig = "blank")
```




Interestingly, `soldminusbuilt` which is `yrsold` - `yearbuilt` becomes insignificant in this smaller model with only the best predictors

```
bestpredictors <- names(house_numeric)[sapply(house_numeric, function(x) abs(cor(house_numeric$saleprice

bestpredictors <- bestpredictors[-6]

bestmodel <- lm(saleprice~overallqual + exterqual + grlivarea + kitchenqual +
     garagecars + pricecategory, data = house)

summary(bestmodel)
```

```
##
## Call:
## lm(formula = saleprice ~ overallqual + exterqual + grlivarea +
##     kitchenqual + garagecars + pricecategory, data = house)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -317854  -20377   -1463   17249  294682
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.365e+05  6.394e+03 -21.348  < 2e-16 ***
##  [ reached getOption("max.print") -- omitted 6 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37390 on 1453 degrees of freedom
## Multiple R-squared:  0.7794, Adjusted R-squared:  0.7785
## F-statistic: 855.5 on 6 and 1453 DF,  p-value: < 2.2e-16
```

Subset with only best predictors

```
housesubset <- house %>% select(bestpredictors)
```

So, 6 variables capture 78% of the variation in sale price for our model.

Checking assumptions.

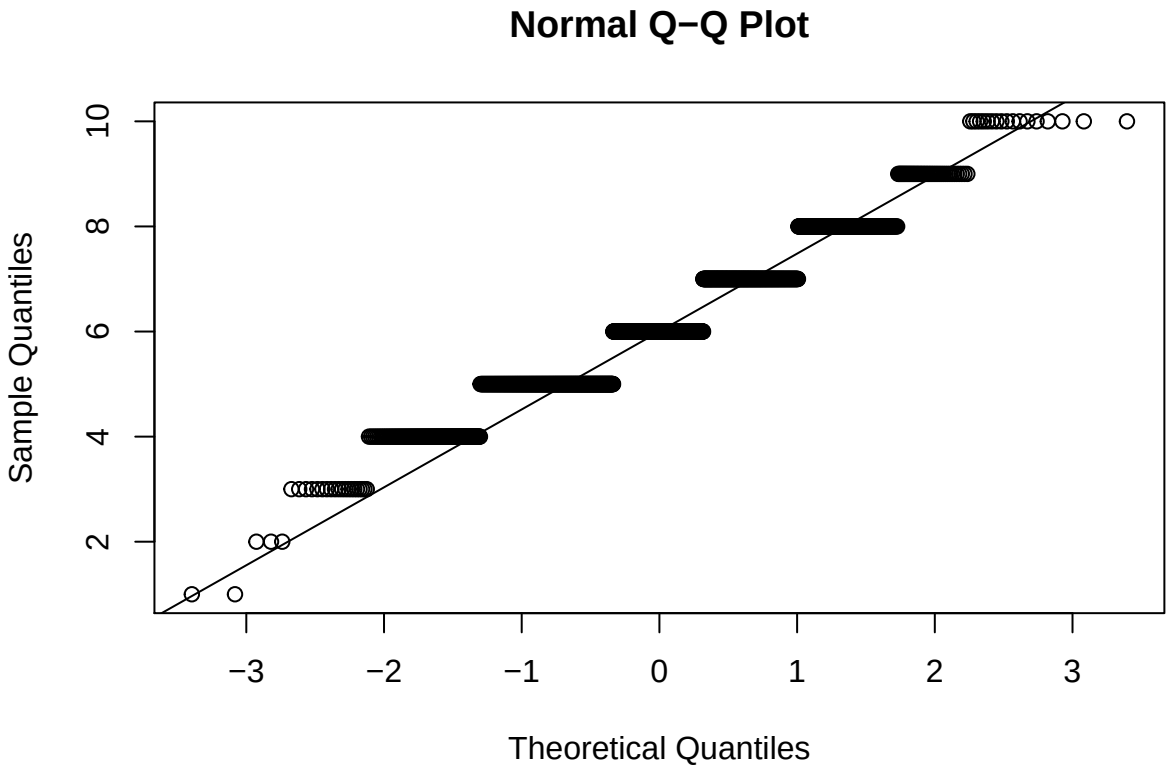
```
cor(housesubset)
```

```
##               overallqual exterqual grlivarea kitchenqual garagecars
##               pricecategory
##  [ reached getOption("max.print") -- omitted 6 rows ]
```

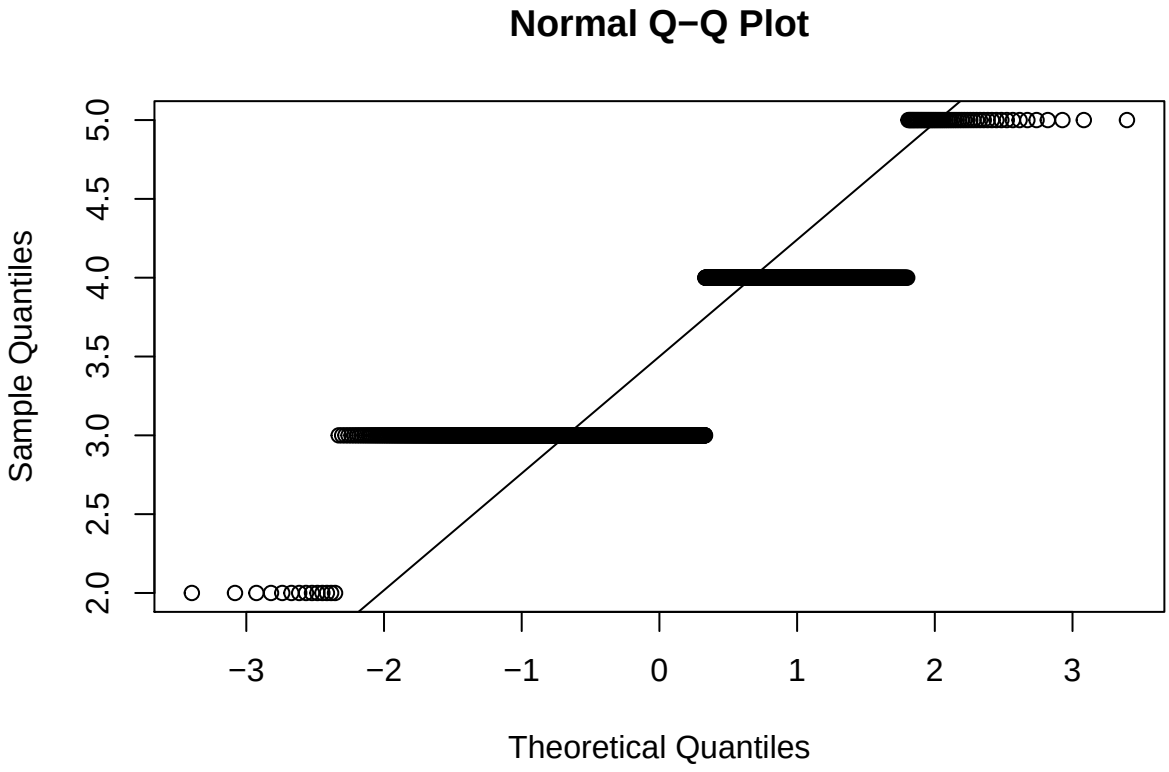
```
vif(bestmodel)
```

```
##   overallqual     exterqual     grlivarea   kitchenqual    garagecars
##      3.243341      2.759469      1.592611      2.326244      1.773987
## pricecategory
##      1.929410
```

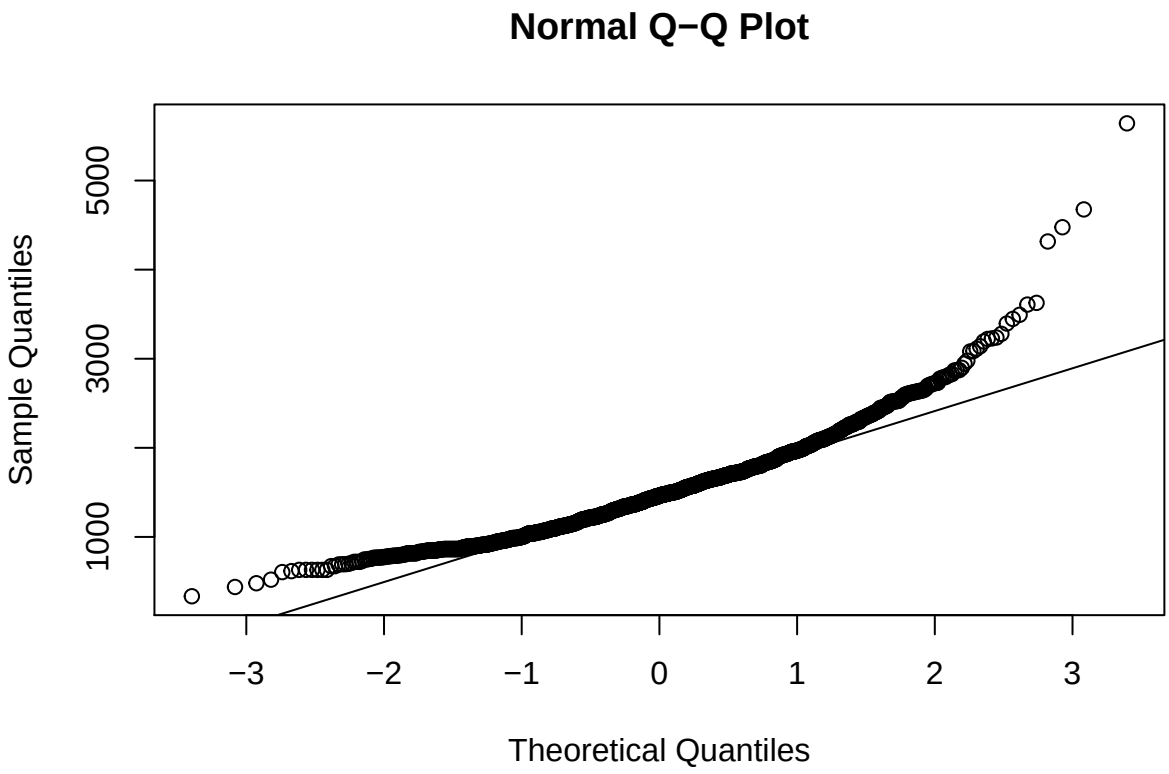
```
qqnorm(housesubset$overallqual)
qqline(housesubset$overallqual)
```



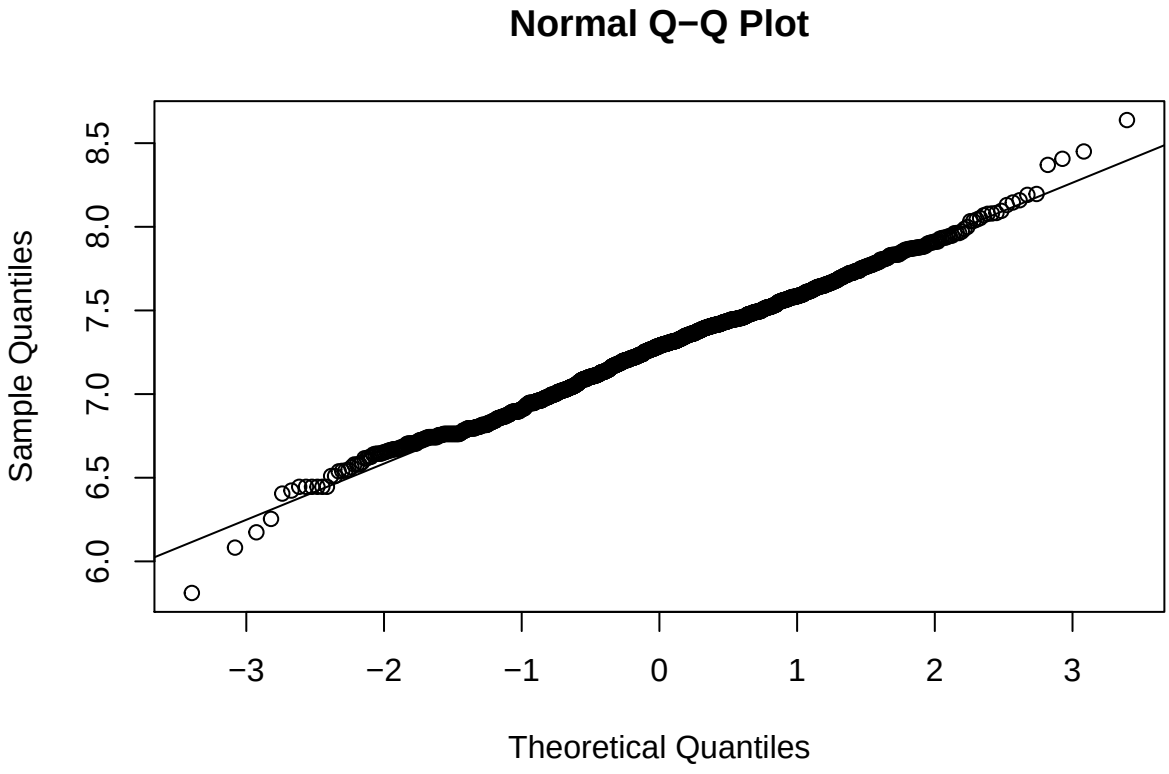

```
qqnorm(housesubset$exterqual)
qqline(housesubset$exterqual)
```



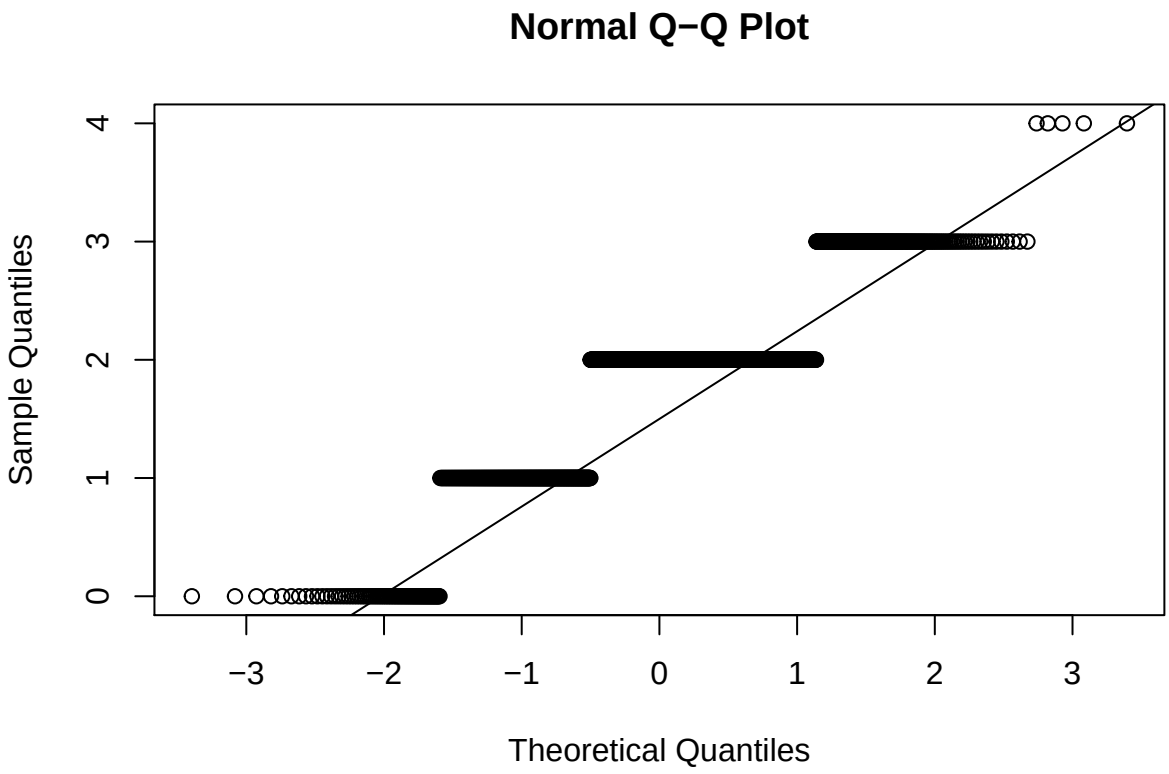

```
qqnorm(housesubset$grlivarea)
qqline(housesubset$grlivarea)
```



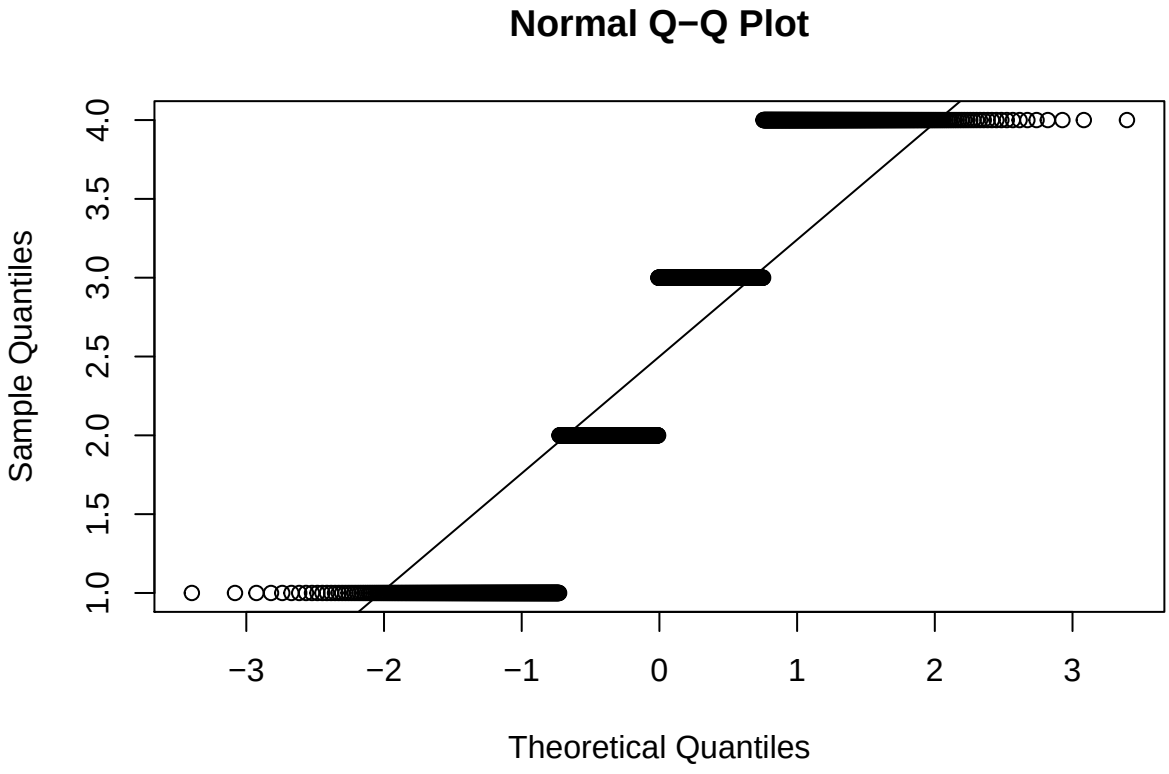

```
qqnorm(log(housesubset$grlivarea))
qqline(log(housesubset$grlivarea))
```



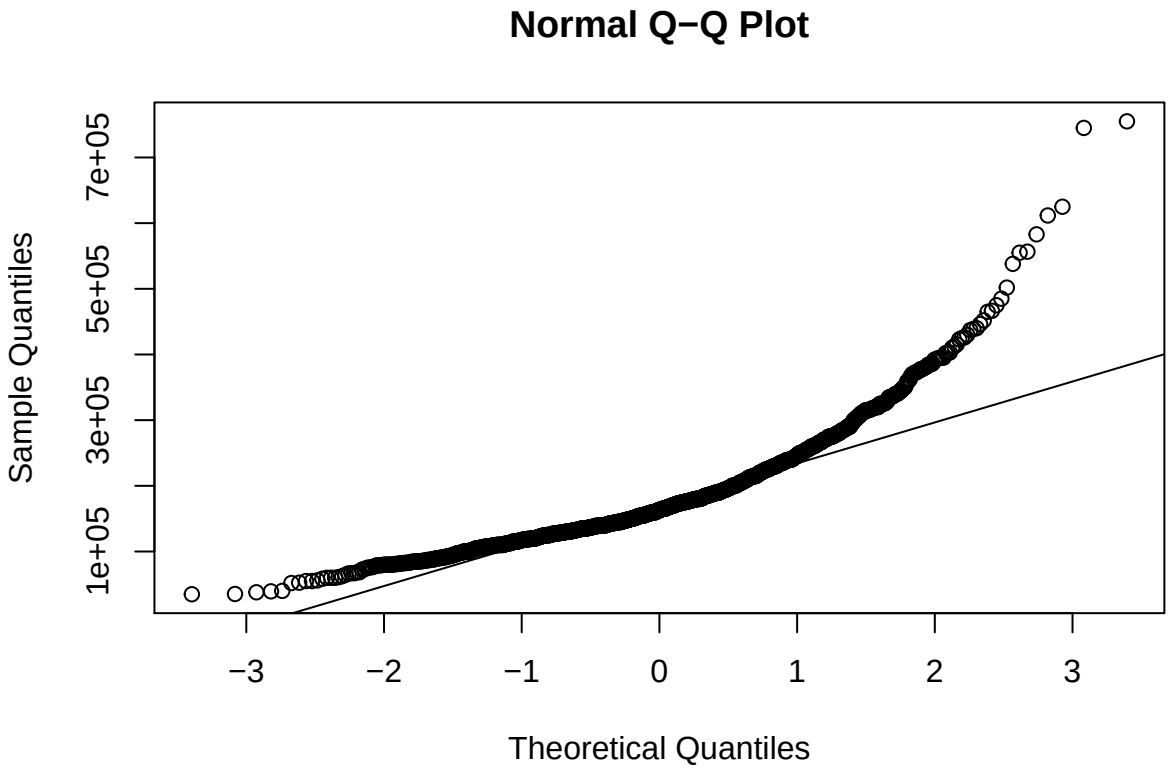

```
qqnorm(housesubset$garagecars)
qqline(housesubset$garagecars)
```



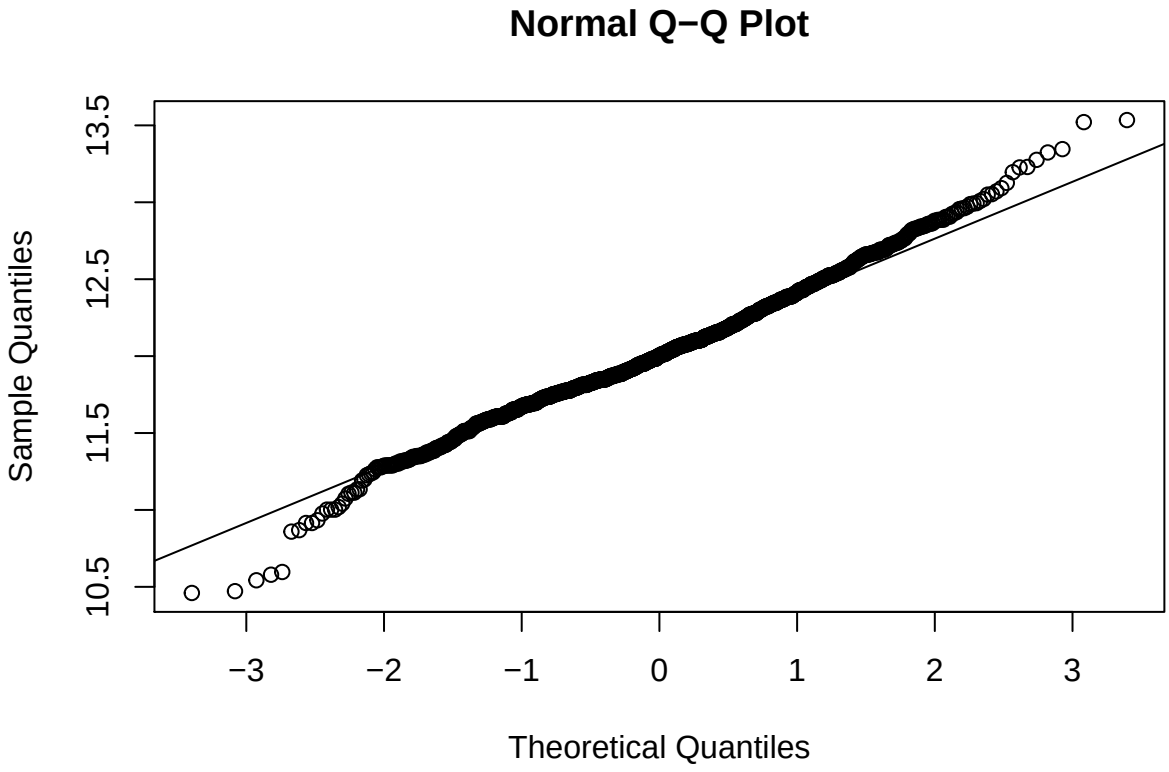

```
qqnorm(housesubset$pricecategory)
qqline(housesubset$pricecategory)
```




```
qqnorm(house$saleprice)
qqline(house$saleprice)
```




```
qqnorm(log(house$saleprice))
qqline(log(house$saleprice))
```




```
bestmodel2 <- lm(log(saleprice)~overallqual  + exterqual +  log(grlivarea) +
    kitchenqual + garagecars + pricecategory, data = house)
```

```
summary(bestmodel2)
```

```
##
## Call:
## lm(formula = log(saleprice) ~ overallqual + exterqual + log(grlivarea) +
##     kitchenqual + garagecars + pricecategory, data = house)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99069 -0.08304  0.00866  0.10361  0.54294
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.055655   0.111219  72.431  < 2e-16 ***
##  [ reached getOption("max.print") -- omitted 6 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1649 on 1453 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8296
## F-statistic:  1185 on 6 and 1453 DF,  p-value: < 2.2e-16
```

`exterqual` becomes insignificant once we take the log of the response variable

```
bestmodel3 <- lm(log(saleprice)~overallqual  +  log(grlivarea) +
    kitchenqual + garagecars + pricecategory, data = house)
summary(bestmodel3)
```

```
##
## Call:
## lm(formula = log(saleprice) ~ overallqual + log(grlivarea) +
##     kitchenqual + garagecars + pricecategory, data = house)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98631 -0.08375  0.00866  0.10300  0.54504
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.082140   0.109886  73.550   <2e-16 ***
##  [ reached getOption("max.print") -- omitted 5 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1649 on 1454 degrees of freedom
## Multiple R-squared:  0.8301, Adjusted R-squared:  0.8295
## F-statistic:  1420 on 5 and 1454 DF,  p-value: < 2.2e-16
```

Check for influence points

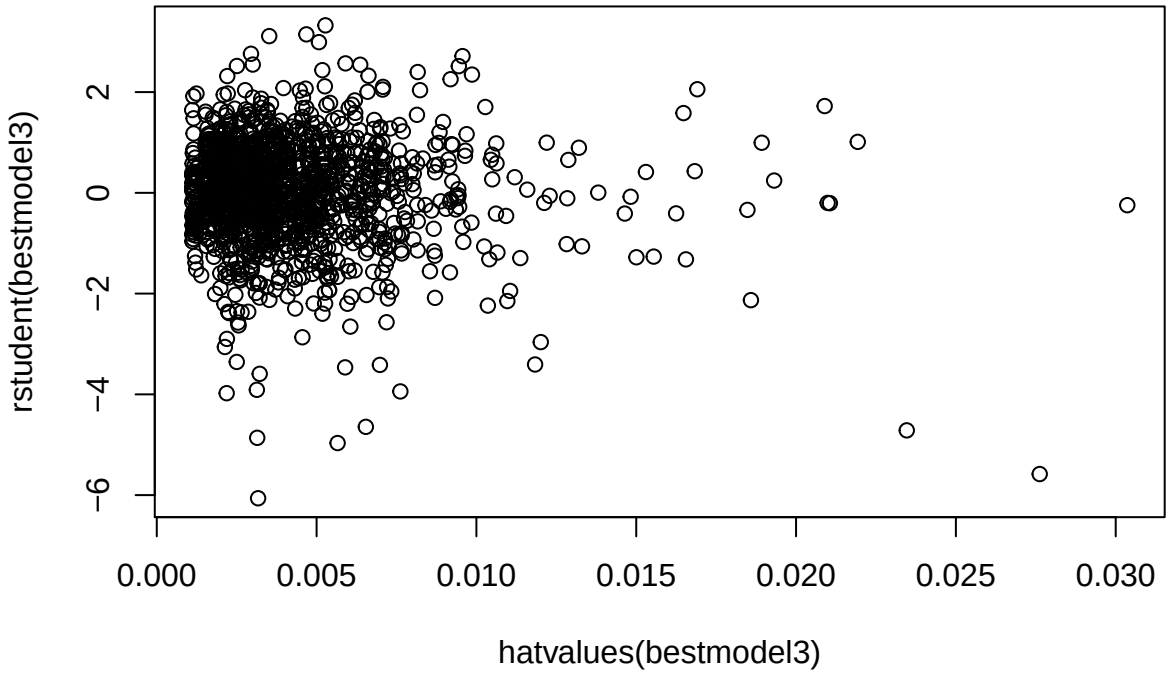
```
infm <- influence.measures(bestmodel3)
which(apply(infm$is.inf,1,any)) #influential observations
```

```
##  4 16 29 30 31
##  4 16 29 30 31
##  [ reached getOption("max.print") -- omitted 94 entries ]
```

```
summary(infm)
```

```
## Potentially influential observations of
##   lm(formula = log(saleprice) ~ overallqual + log(grlivarea) +      kitchenqual + garagecars + price
##
##      dfb.1_ dfb.ovrl dfb.lg() dfb.ktch dfb.grgc dfb.prcc dffit   cov.r
##      cook.d hat
##  [ reached getOption("max.print") -- omitted 99 rows ]
```

```
plot(rstudent(bestmodel3) ~ hatvalues(bestmodel3))
```



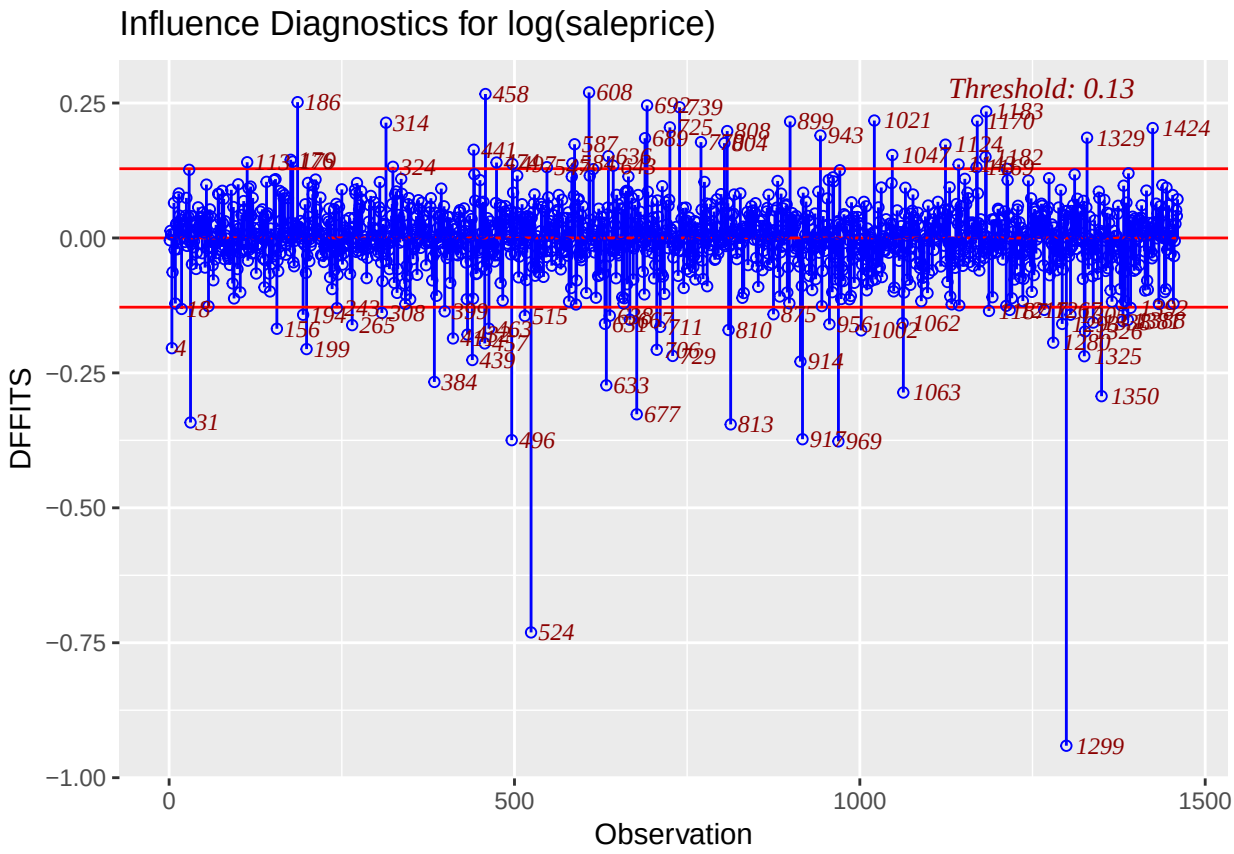

```
#install.packages("olsrr")
```

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
threshold <- 2*sqrt(5/1460)
ols_dffits_plot(bestmodel3)
```




```
help("ols_dffits_plot")
```

Let's examine Observation # 1299, and 524

```
house[1299,] %>% View()
house[542,] %>% View()

bestmodel4 <- lm(log(saleprice)~overallqual  +  log(grlivarea) +
    kitchenqual + garagecars + pricecategory, data = house[c(-1299,-542),])
summary(bestmodel4)
```

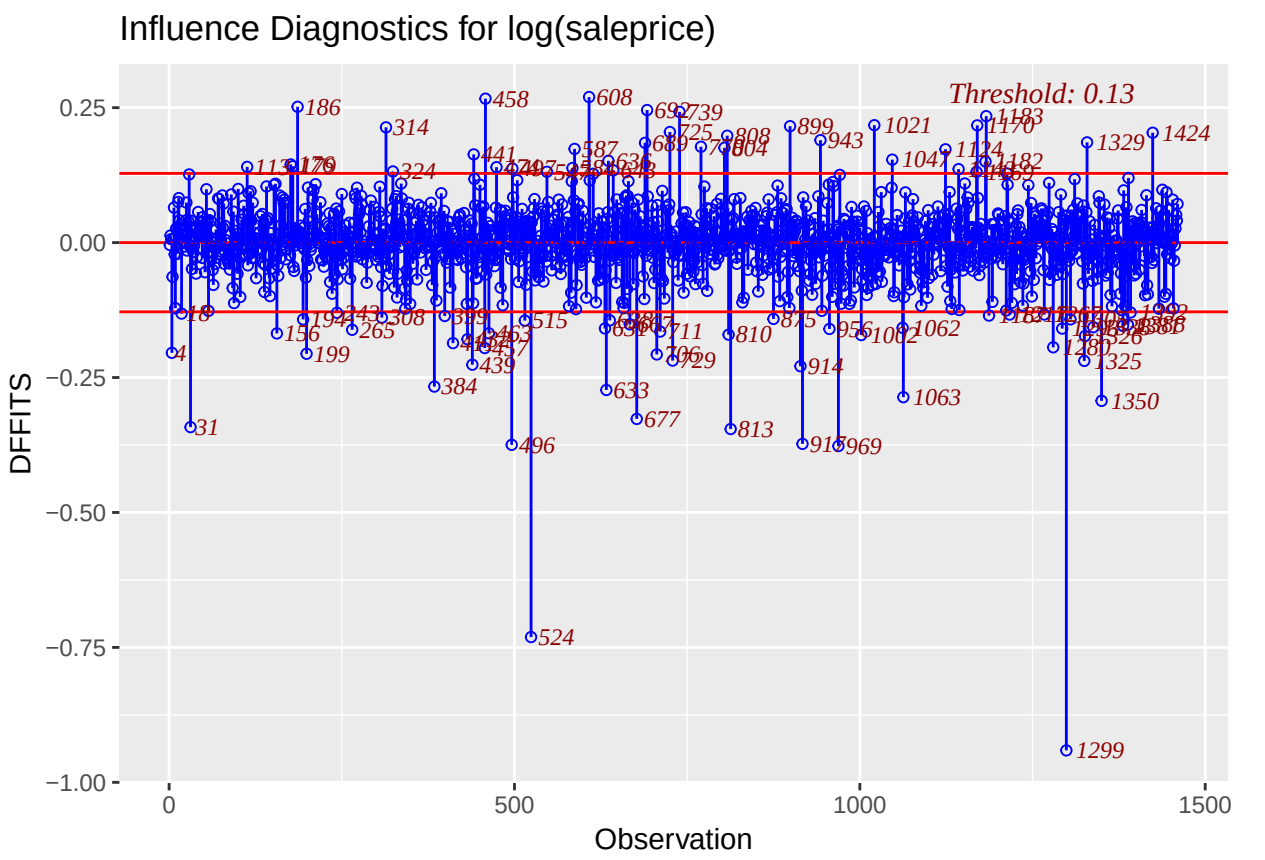
```
##
## Call:
## lm(formula = log(saleprice) ~ overallqual + log(grlivarea) +
##     kitchenqual + garagecars + pricecategory, data = house[c(-1299,
##     -542), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98783 -0.08482  0.00710  0.10281  0.54125
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.013851   0.109393  73.257   <2e-16 ***
##  [ reached getOption("max.print") -- omitted 5 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1632 on 1452 degrees of freedom
## Multiple R-squared:  0.8337, Adjusted R-squared:  0.8331
## F-statistic:  1456 on 5 and 1452 DF,  p-value: < 2.2e-16
```

By just removing two points, our Adjusted R-squared went from 0.8294849 to 0.8331439

Let's see what happens if we simply remove the observations.

```
influence <- ols_dffits_plot(bestmodel3)
```




```
influenceindex <- unlist(influence$outliers[1])

bestmodelnoinfluence <- lm(log(saleprice)~overallqual  +  log(grlivarea) +
    kitchenqual + garagecars + pricecategory, data = house[-influenceindex,])
summary(bestmodelnoinfluence)
```

```
##
## Call:
## lm(formula = log(saleprice) ~ overallqual + log(grlivarea) +
##     kitchenqual + garagecars + pricecategory, data = house[-influenceindex,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42359 -0.08062  0.00367  0.08874  0.39582
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.223182   0.093062  88.362   <2e-16 ***
```

```
##  [ reached getOption("max.print") -- omitted 5 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1292 on 1368 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8713
## F-statistic:  1860 on 5 and 1368 DF,  p-value: < 2.2e-16
```

We see that our Adjusted R-squared went from 0.8331439 to 0.8713259 after removing ALL the influence points.

```
#house[influenceindex, ]$saleprice <- NA
#kNN(house, variable = "saleprice",  k = k)$saleprice
```

Check MSPE

```
set.seed(888)
train <- sample(nrow(house), nrow(house) * .8)
test <- (-train)

trainhouse <- house[train,]
testhouse <- house[test,]

trainmodel <- lm(log(saleprice)~overallqual  +  log(grlivarea) +
    kitchenqual + garagecars + pricecategory, data = trainhouse)
exp(predict(trainmodel, testhouse)) - testhouse$saleprice
```

```
##          1          2          3          4          5
##  -5512.680 111684.272  37783.261  -9117.124  -5211.532
##  [ reached getOption("max.print") -- omitted 287 entries ]
```