EDA_1

Dan Watson

December 6, 2016

Overall thoughts from EDA

Lot.Frontage may be important, may be same as lot size- that variable not reviewed here. Basements seem to matter- possible interaction term. Square-footage seems to matter, do we have total sqft instead of separated by levels? If so, probably just as valuable. Pools and fences probably aren't important, we can expand upon this if necessary, but doesn't seem worth going much further with these.

Prep

Load libraries

```
library(dplyr)
```

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

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

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

```
library(tidyr)
library(ggplot2)
library(car)
```

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

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

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(sandwich)
```

Set WD and read data

```r
housing_data <- read.csv("AmesHousing_data_2010.csv")
```

Variable Analysis

Variable: Lot.Frontage

Meaning: Linear feet of street connected to property.

```r
typeof(housing_data$Lot.Frontage)
```

```
## [1] "integer"
```

```r
length(housing_data$Lot.Frontage)
```

```
## [1] 341
```

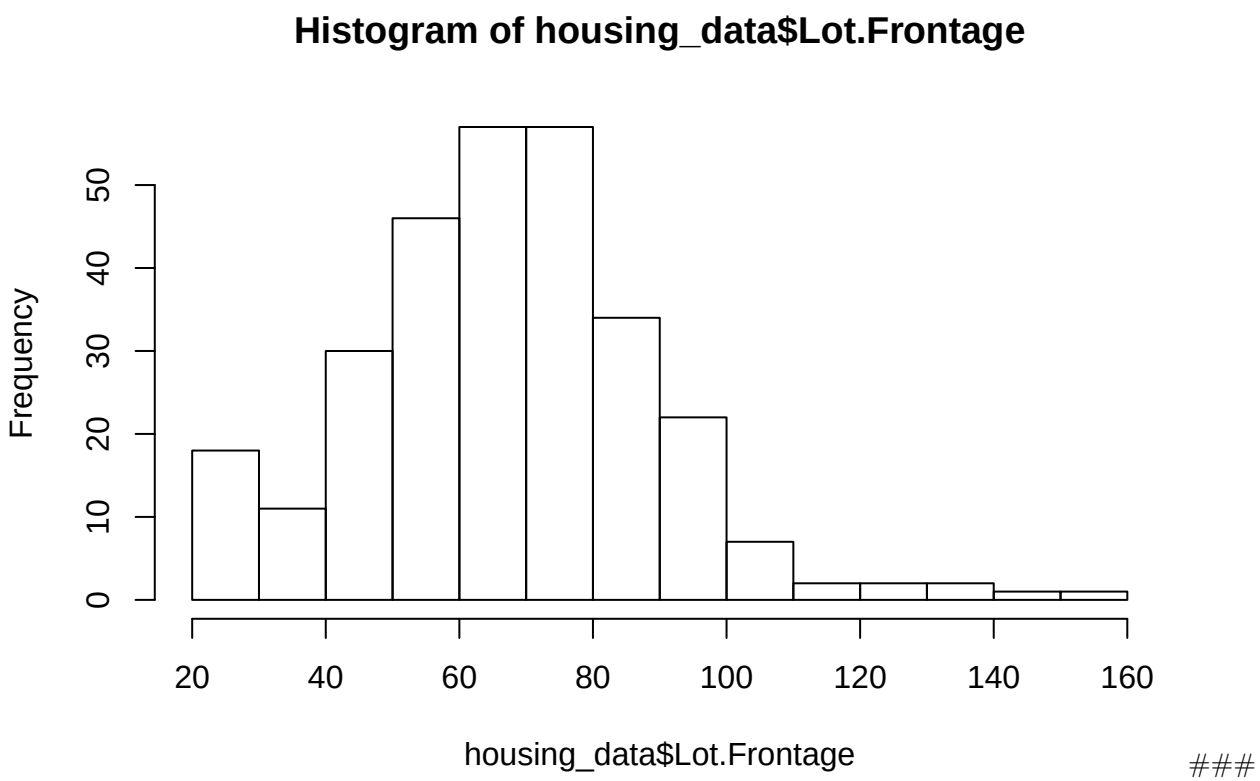
No missing values.

```r
table(housing_data$Lot.Frontage)
```

```
##
##  21  24  25  26  30  31  35  36  38  39  40  41  43  44  45  47  48  50
##   7   6   1   1   3   1   2   2   1   3   2   2   3   4   1   2   2  16
##  51  52  53  54  55  56  57  58  59  60  61  63  64  65  66  67  68  69
##   2   4   3   1   3   3   2   2   1  25   1   3   5   9   5   3  11   2
##  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87
```

```
##  18   1   4   6   4  10   4   2   6   1  19   3   2   3   1  10   1   3
##  88  89  90  92  93  94  95  96  98 100 102 105 107 108 110 119 120 124
##   5   2   4   3   2   5   1   1   3   7   1   2   1   1   2   1   1   1
## 129 137 140 141 152
##   1   1   1   1   1
```

```
hist(housing_data$Lot.Frontage)
```



###

Not terribly skewed- slight R tail, could transform, but probably not necessary.

```
summary(housing_data$Lot.Frontage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   21.00   56.00   70.00   68.79   80.00  152.00      51
```

51 NAs. Mean 68.79, median 70. Range 21-152

```
sum(is.na(housing_data$Lot.Frontage))
```

```
## [1] 51
```

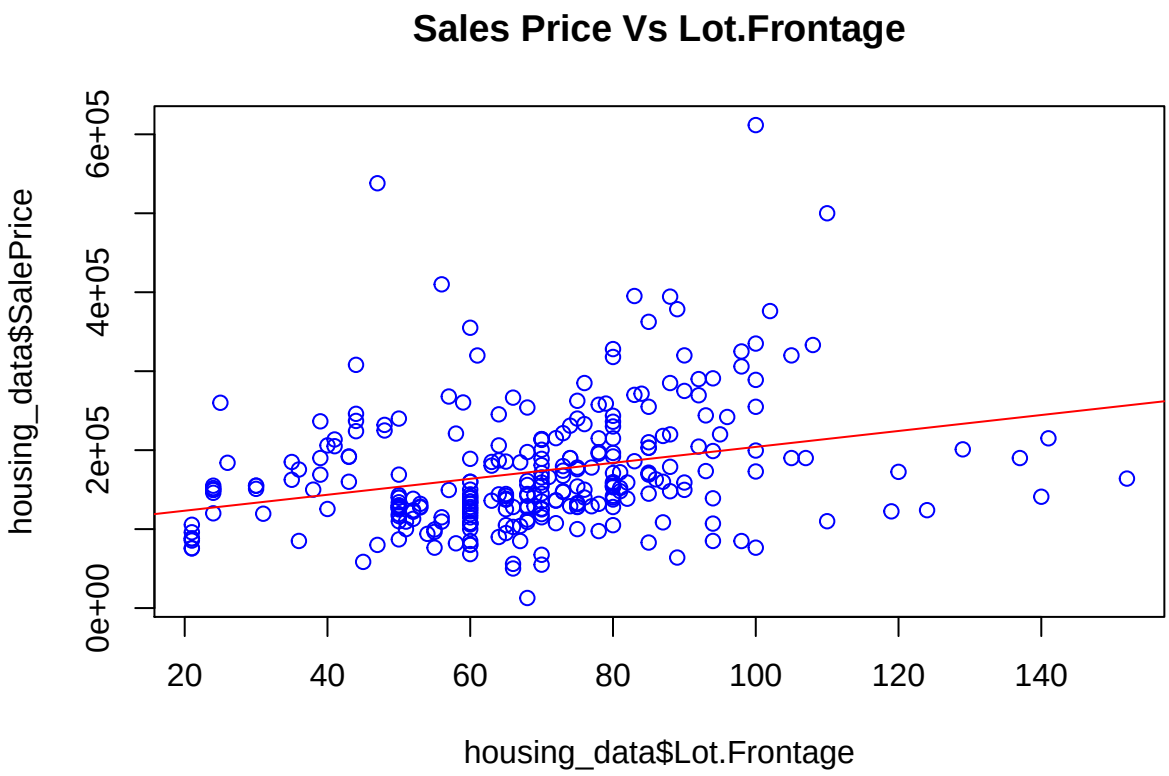
Double checked NA calculation, should we use complete cases?

```
cc<-complete.cases(housing_data)
df2<-housing_data[cc, ]
```

There are no complete cases at this point, seems that a lot of variables have no values- such as alley. Can see if complete cases of our final variables makes sense to use.

Plot relationship

```
plot(housing_data$Lot.Frontage, housing_data$SalePrice, main = "Sales Price Vs Lot.Frontage", col ="blu
abline(lm(SalePrice~Lot.Frontage, data= housing_data), col = "red")
```



###

Does not seem to be a tight linear fit between the variables.

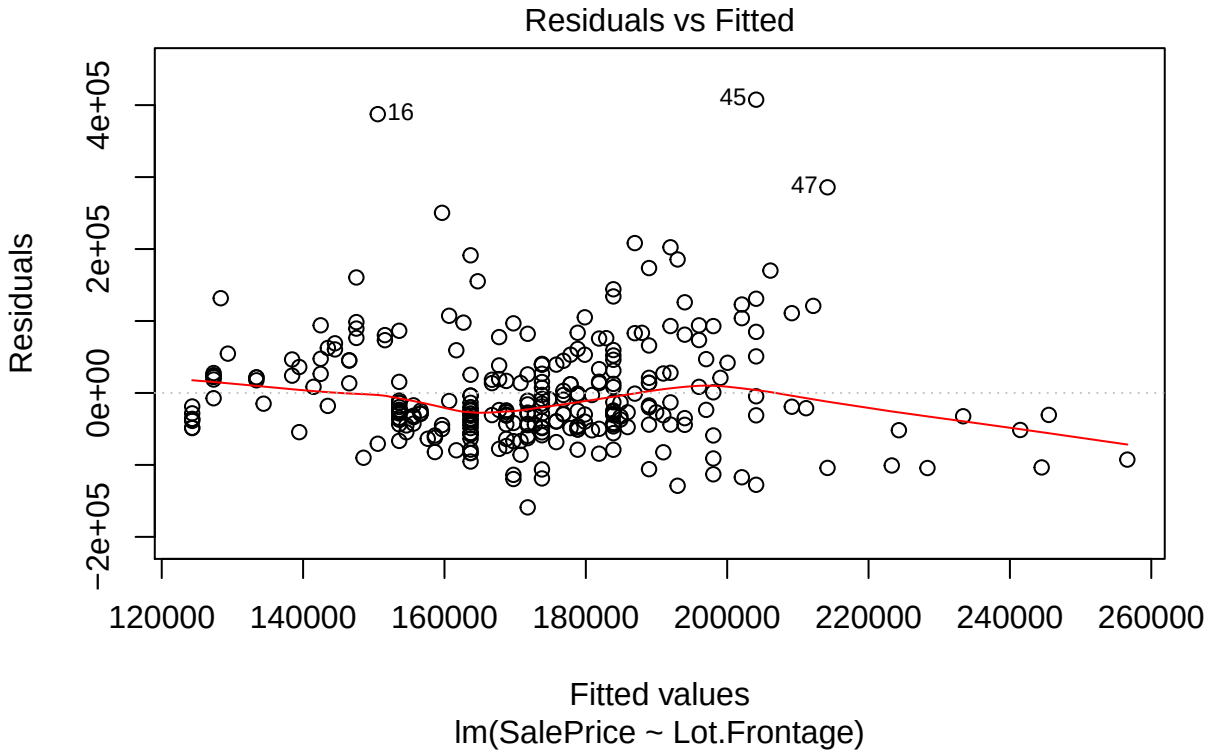
Simple linear model

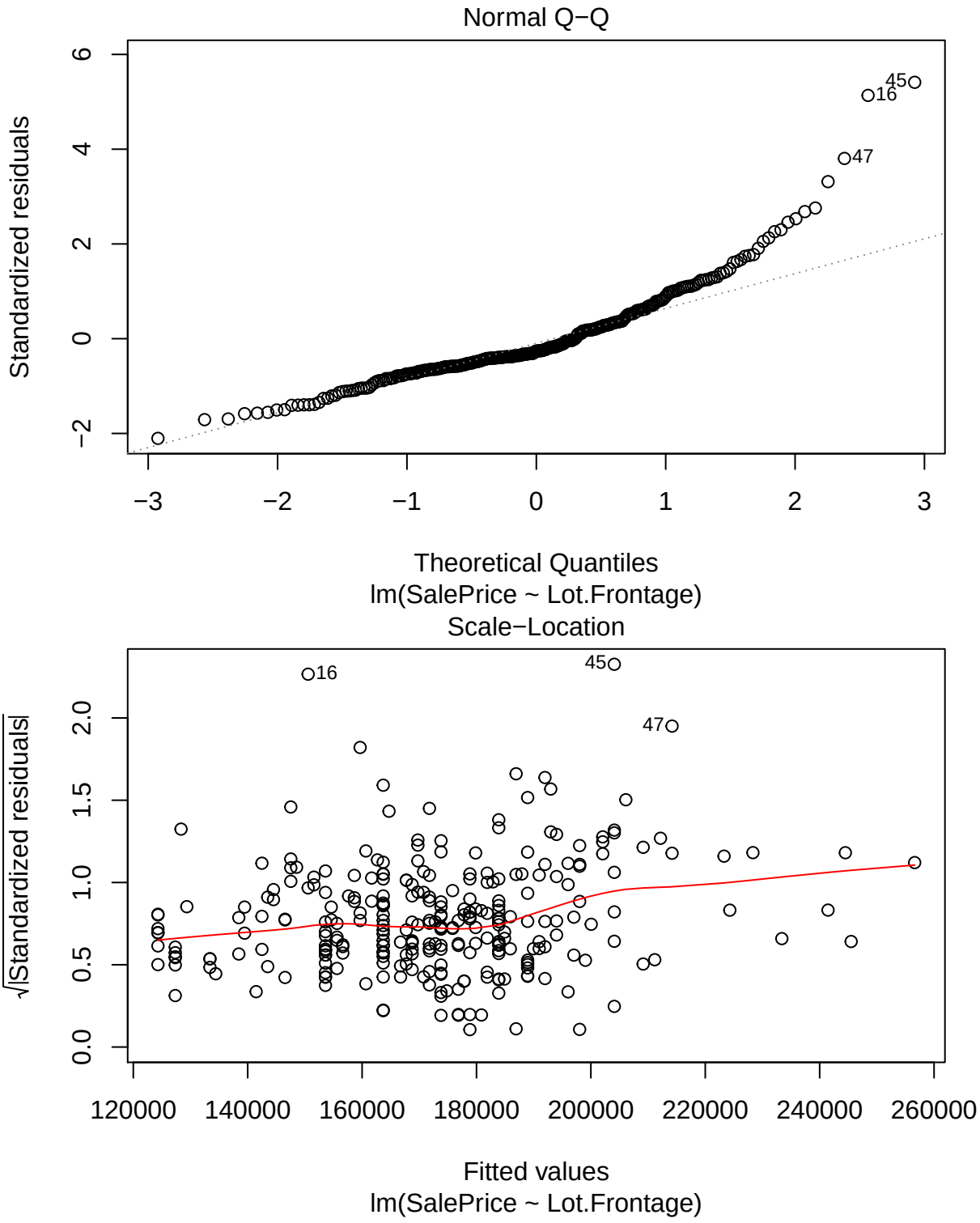
```
ModelFrontage <- lm(SalePrice~Lot.Frontage, data= housing_data)
summary(ModelFrontage)
```

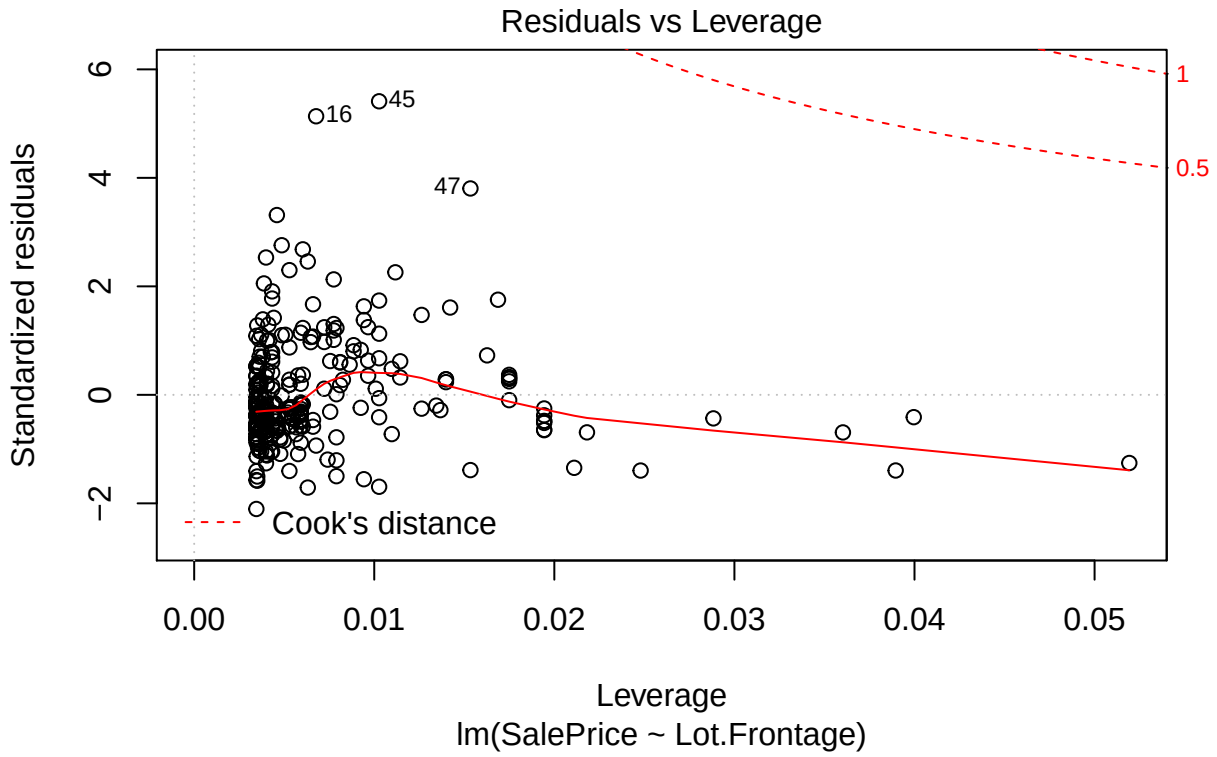
```
##
## Call:
## lm(formula = SalePrice ~ Lot.Frontage, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158987  -44566  -19821   30260  407562
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  103097.6    14478.3   7.121 8.59e-12 ***
```

```
## Lot.Frontage   1010.0      200.3   5.042 8.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75710 on 288 degrees of freedom
##   (51 observations deleted due to missingness)
## Multiple R-squared:  0.08111,    Adjusted R-squared:  0.07792
## F-statistic: 25.42 on 1 and 288 DF,  p-value: 8.163e-07
```

```
plot(ModelFrontage)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SalePrice ~ Lot.Frontage)
Scale−Location
√|Standardized residuals|
Fitted values
lm(SalePrice ~ Lot.Frontage)
16
45
47



Statistically significant relationship. QQ plot looks good until tail. Seems like we have heteroskedasticity.

```
coeftest(ModelFrontage, vcov= vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  103097.55   15064.61  6.8437 4.644e-11 ***
## Lot.Frontage   1009.97     230.19  4.3876 1.610e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appears significant, T value 4.3876, P 1.610e-05

Variable: BsmtFinSF1

Meaning: SQFT of the basement type1

If we keep this variable, should probably be interaction term as this variable by itself won't provide much context.

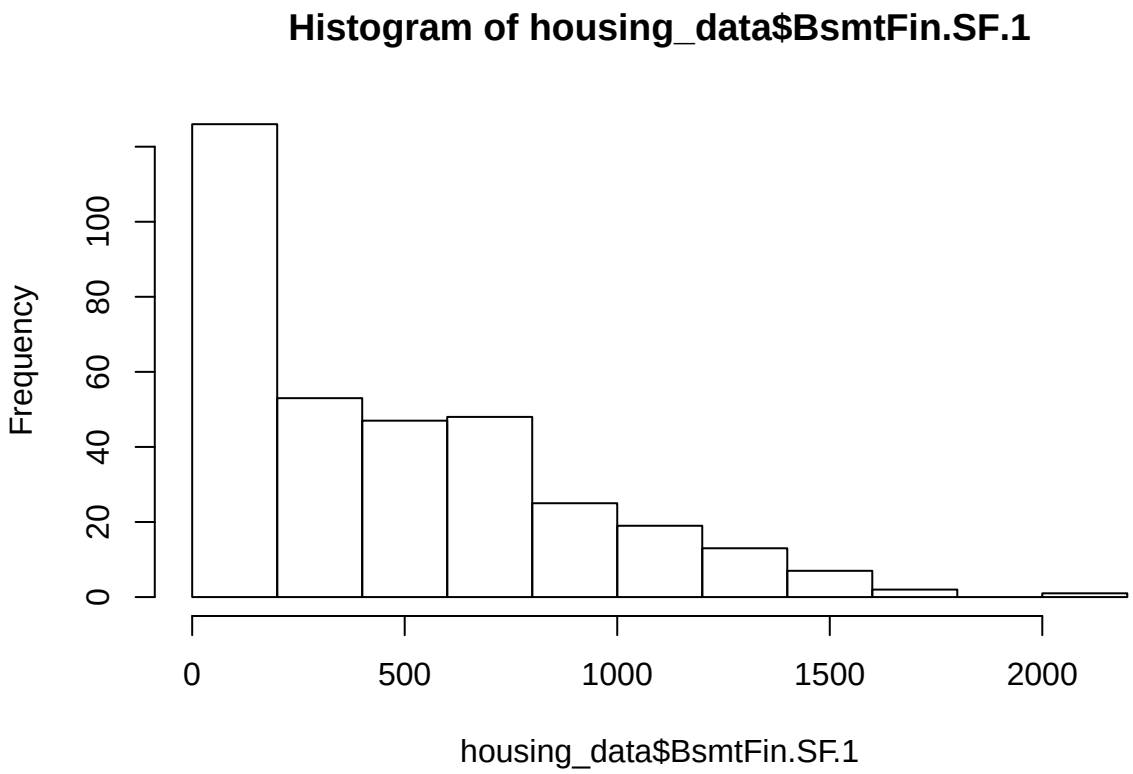
```
typeof(housing_data$BsmtFin.SF.1)
```

```
## [1] "integer"
```

```
length(housing_data$BsmtFin.SF.1)
```

```
## [1] 341
```

```
hist(housing_data$BsmtFin.SF.1)
```



###

Note the huge number of 0s for basements

```
table(housing_data$BsmtFin.SF.1)
```

```
##
##   0   2  16  24  28  36  42  48  49  72  73  76  78  94 104
##  92   1   1   5   1   1   1   1   1   1   1   1   1   1   2
## 108 110 120 126 129 133 150 156 176 182 188 192 198 203 207
##   1   1   1   1   1   1   1   1   2   1   2   1   1   1   1
## 218 222 234 238 240 244 247 248 250 263 264 267 274 275 276
##   1   1   2   1   1   1   1   1   2   1   1   1   1   1   1
## 280 283 288 294 300 308 311 320 326 329 334 338 339 340 343
##   1   1   1   1   3   1   1   1   1   1   1   1   1   1   1
## 348 353 354 360 368 370 371 375 376 378 383 394 402 422 426
##   1   2   1   2   3   1   1   1   1   2   1   1   1   1   1
## 427 432 438 442 443 448 450 452 456 458 466 468 469 474 476
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 480 481 483 484 504 505 506 507 510 514 524 528 532 533 539
##   2   1   1   1   1   1   1   1   1   1   1   1   1   2   1
## 540 543 553 569 578 579 588 593 600 602 604 615 616 623 624
```

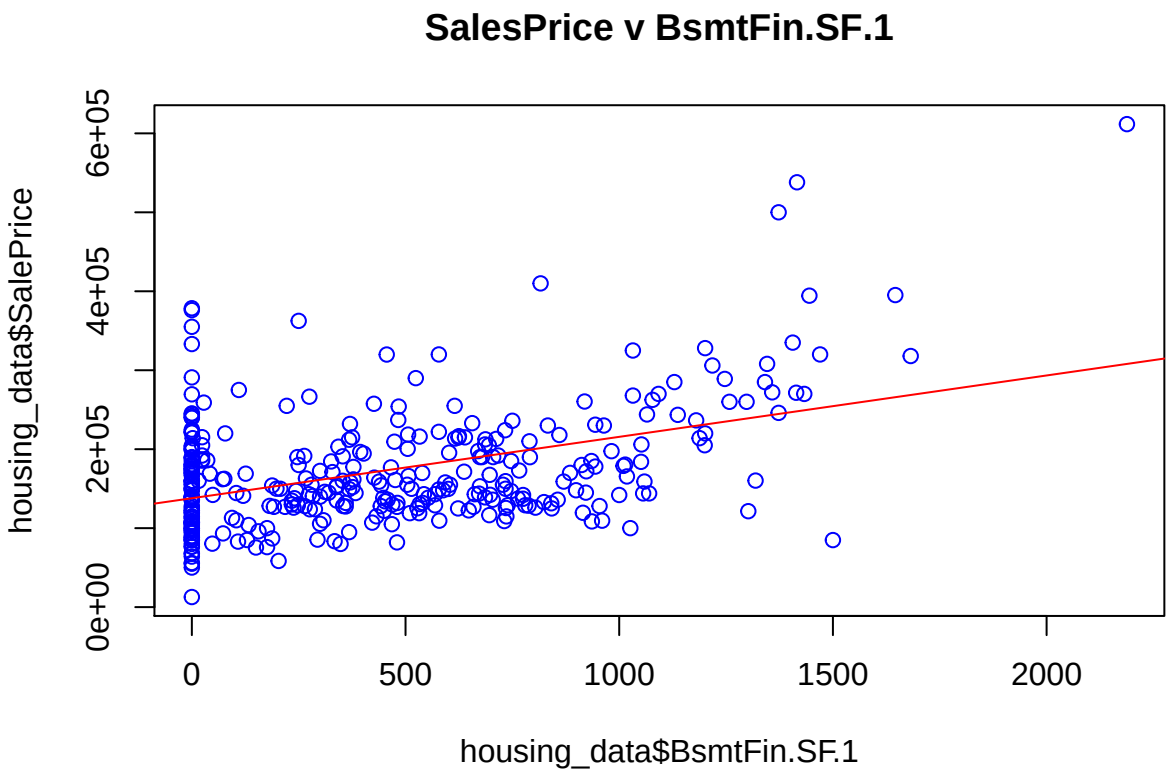
```
##    1    1    1    2    3    1    1    1    1    1    1    1    1    1    1
##  625  637  639  648  656  659  662  670  672  674  679  686  687  695  696
##    1    1    1    1    1    1    1    1    1    2    1    2    1    1    1
##  697  698  704  705  712  717  728  731  733  734  735  736  739  747  750
##    1    1    1    1    1    1    1    1    1    2    1    1    1    2    1
##  763  766  775  780  788  790  791  804  816  824  833  841  842  856  860
##    1    1    2    1    1    1    1    1    1    1    1    1    1    1    1
##  870  885  899  912  915  919  922  923  935  936  944  954  960  964  982
##    1    1    1    1    1    1    1    1    1    1    2    1    1    1    1
## 1000 1010 1014 1018 1026 1032 1051 1052 1056 1059 1065 1070 1078 1092 1129
##    1    1    1    1    1    2    1    1    1    1    1    1    1    1    1
## 1137 1180 1188 1200 1201 1218 1247 1258 1298 1302 1319 1341 1346 1358 1373
##    1    1    1    1    2    1    1    1    1    1    1    1    1    1    2
## 1406 1414 1416 1433 1445 1470 1500 1646 1682 2188
##    1    1    1    1    1    1    1    1    1    1
```

92 properties have no basement, but concerned about values even through 104sqft- maybe a 10x10 room is considered a basement but what about 49sqft- 7x7, is that a basement?24sqft? Where is the line?

```
summary(housing_data$BsmtFin.SF.1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0   368.0   446.2   728.0  2188.0
```

```
plot(housing_data$BsmtFin.SF.1, housing_data$SalePrice, col="blue", main = "SalesPrice v BsmtFin.SF.1")
abline(lm(SalePrice~ BsmtFin.SF.1, data= housing_data), col ="red")
```



###

There seems to be some relationship, not tightly fitted. I don't think it makes sense at this time to do an

analysis of this as a linear model until we decide if we want to explore basements at all due to the limited number of responses. If kept, we should create an interaction term with sqft of types.

Variable: BsmtFinSF2

Meaning: Similar to the variable above, this is the sqft for the second type of basement. Example- if a home has a half finished and half unfinished basement, this will give the sqft of the second type of basement.

```
typeof(housing_data$BsmtFin.SF.2)
```

```
## [1] "integer"
```

```
length(housing_data$BsmtFin.SF.2)
```

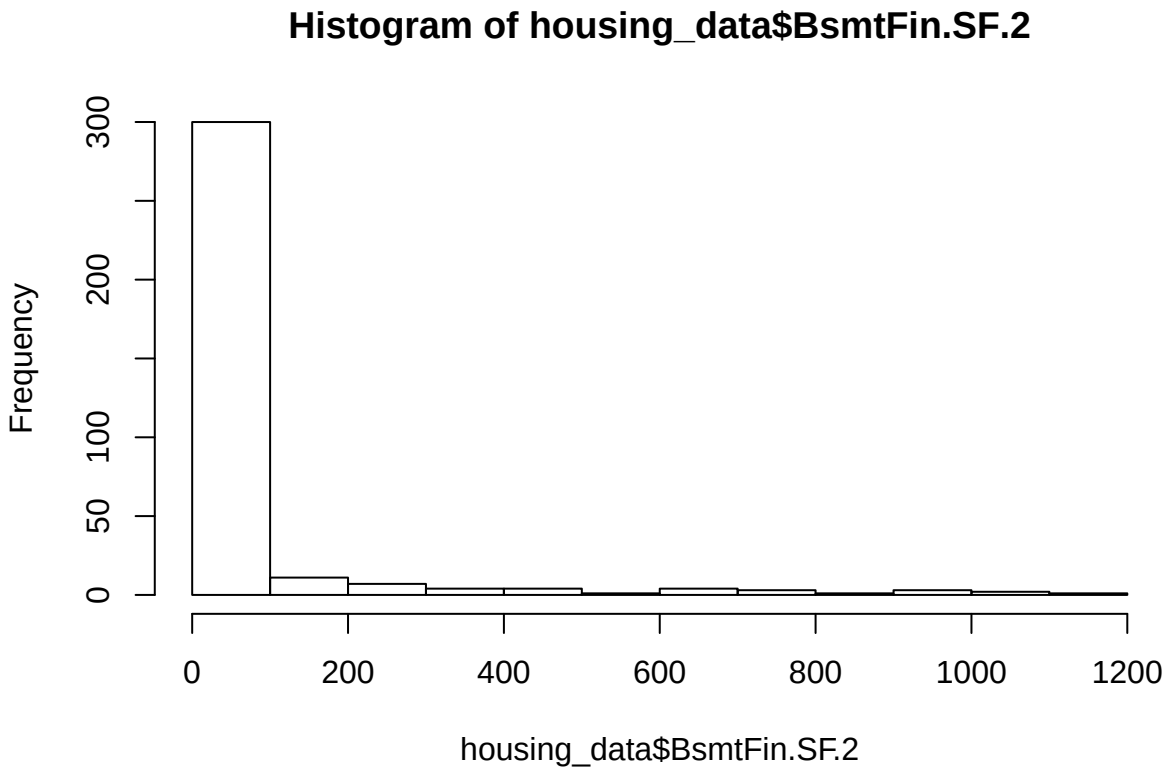
```
## [1] 341
```

```
table(housing_data$BsmtFin.SF.2)
```

```
##
##    0   12   42   46   78   81  117  119  121  127  132  144  159  162  163
##  295    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##  168  174  232  240  252  258  263  284  290  334  350  362  387  453  474
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##  480  486  590  668  684  688  692  712  713  859  906  972  981 1029 1073
##    1    1    1    1    1    1    1    2    1    1    1    1    1    1    1
## 1120
##    1
```

We see the same issue here with the size. Maybe the small sizes make sense if one room unfinished or a small area where water heater/furnace located unfinished?

```
hist(housing_data$BsmtFin.SF.2)
```



###

The vast majority of houses don't have two different types of basements in the same home.

Considering the size issues in BsmtFin.SF.1, are there any basements where basement 1 is smaller than basement2?

```
sum(housing_data$BsmtFin.SF.2> housing_data$BsmtFin.SF.1)
```

```
## [1] 23
```

we have 23 basements where the second type is larger than the first. lets explore total bsmt size for these.

```
df3<-housing_data$Total.Bsmt.SF[housing_data$BsmtFin.SF.2> housing_data$BsmtFin.SF.1]
table(df3)
```

```
## df3
##  536  630  663  720  816  833  864  894  900  926  972 1026 1060 1063 1078
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1086 1208 1231 1246 1347 1488 1517 1657
##    1    1    1    1    1    1    1    1
```

All of these sizes make sense, so maybe we don't have an issues

```r
table(housing_data$Total.Bsmt.SF)
```

```
##
##    0  346  360  381  384  423  458  480  483  525  528  530  533  536  540
##    7    1    1    1    3    1    1    1    1    2    1    1    1    1    1
##  546  559  572  576  600  608  622  624  629  630  650  660  662  663  672
##    4    1    1    3    5    1    1    2    1    1    1    1    1    2    2
##  676  678  686  689  696  698  707  709  710  715  720  725  728  735  738
##    1    1    1    2    1    1    1    1    1    1    1    1    2    1    1
##  741  744  747  750  756  763  764  765  774  777  780  782  788  789  796
##    1    2    2    1    6    1    1    1    1    1    3    1    1    1    1
##  804  806  814  816  817  827  831  832  833  835  836  840  847  848  855
##    2    1    1    3    1    1    1    1    1    1    1    3    1    2    2
##  856  858  859  860  864  870  876  878  882  884  888  894  900  910  912
##    2    1    1    1    8    1    1    1    3    1    1    3    1    1    4
##  918  923  926  928  930  936  941  945  946  948  950  956  960  967  972
##    1    1    2    2    1    4    1    1    1    1    1    1    1    1    2
##  975  980  982  988  991  994  996 1004 1008 1010 1012 1026 1027 1029 1032
##    1    1    1    2    1    1    1    1    3    1    1    2    1    1    2
## 1039 1040 1049 1050 1053 1055 1056 1057 1060 1063 1067 1068 1069 1078 1080
##    1    4    1    1    1    1    3    1    1    1    1    1    1    2    2
## 1086 1105 1107 1108 1109 1116 1117 1121 1124 1140 1143 1144 1145 1152 1156
##    1    1    1    1    1    1    1    1    1    1    1    1    1    4    1
## 1161 1168 1172 1175 1187 1188 1191 1194 1195 1196 1206 1208 1209 1212 1214
##    1    1    1    1    1    2    1    1    1    1    1    1    1    2    1
## 1216 1218 1222 1224 1226 1231 1232 1243 1244 1246 1250 1256 1268 1280 1300
##    2    1    1    1    2    1    1    1    1    1    1    1    2    2    1
## 1306 1314 1319 1324 1328 1329 1332 1336 1338 1344 1347 1358 1370 1390 1392
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1393 1395 1398 1405 1414 1420 1422 1430 1433 1463 1468 1470 1473 1480 1488
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    2
## 1492 1501 1508 1510 1517 1528 1541 1542 1544 1560 1566 1581 1590 1594 1595
##    1    1    1    1    2    1    1    1    1    1    1    1    1    1    2
## 1604 1610 1629 1642 1649 1650 1657 1671 1673 1679 1694 1698 1704 1720 1728
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1752 1768 1822 1829 1832 1844 1856 1902 1930 1947 1958 1978 2002 2033 2110
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 2330 2846
##    1    1
```

Looking at total, the smallest bsmt is 346 sqft, so it is more of a classification issue than a problem in the data.

Overall recommendation, consider interaction terms if we want to consider effect of a basement.

Variable: BSMT Unf SF

Meaning: Total unfinished sqft of basement

```
typeof(housing_data$Bsmt.Unf.SF)
```

```
## [1] "integer"
```

```
length(housing_data$Bsmt.Unf.SF)
```

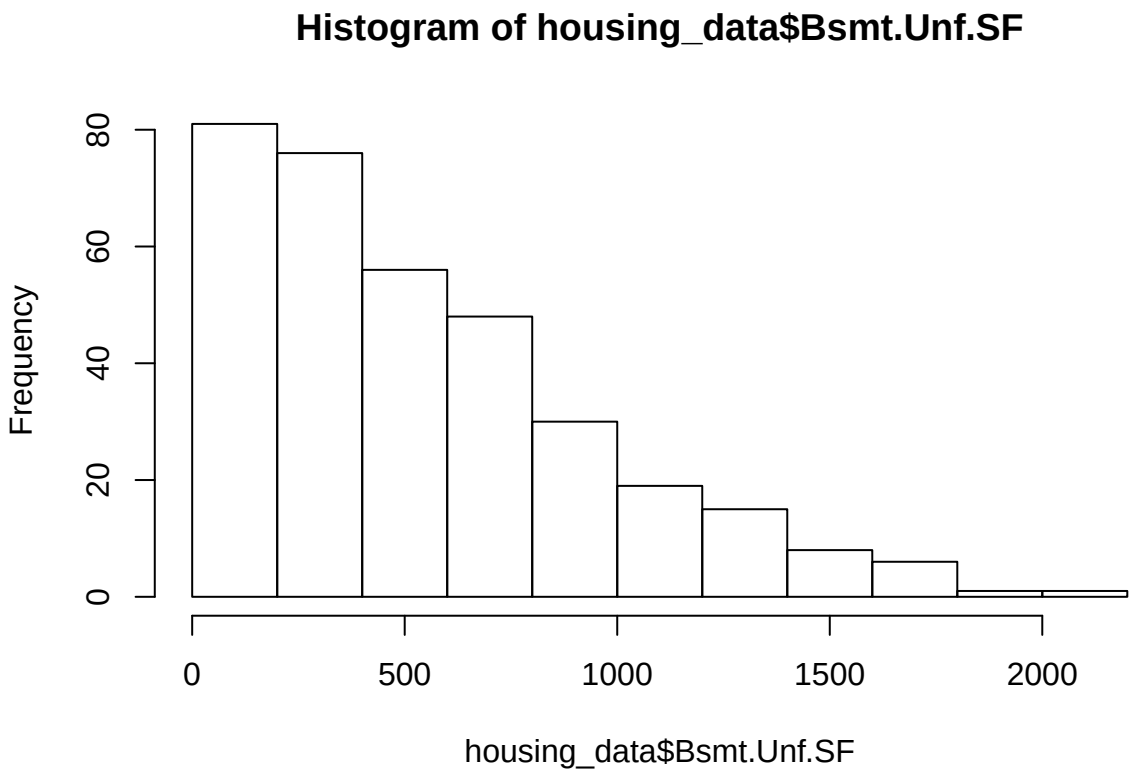
```
## [1] 341
```

```
table(housing_data$Bsmt.Unf.SF)
```

```
##
##    0   30   36   54   58   60   61   70   74   76   80   86   89   90   92
##   28    1    1    2    1    1    1    1    1    1    1    1    1    1    1
##  105  120  125  126  132  134  136  137  138  140  142  143  144  150  153
##    1    1    1    1    1    1    1    1    1    1    1    1    1    2    1
##  161  162  164  166  172  173  174  175  176  179  180  186  188  189  190
##    1    2    1    1    1    1    1    1    1    1    1    2    1    1    2
##  192  193  198  200  203  204  208  210  212  216  218  221  223  224  225
##    1    1    1    1    1    1    1    1    1    1    1    1    2    2    1
##  228  232  233  234  235  237  247  252  253  254  261  270  278  281  284
##    1    1    1    2    2    1    1    2    1    1    2    2    1    2    1
##  286  290  292  296  299  300  308  311  312  315  316  320  321  324  326
##    1    1    2    1    1    2    1    1    1    1    1    1    1    3    1
##  327  328  340  341  344  346  350  354  357  360  370  378  380  381  384
##    1    1    1    1    1    1    2    1    1    1    1    1    1    3    1
##  386  388  392  395  396  403  404  405  406  410  411  412  415  426  431
##    1    1    1    1    1    1    1    1    1    1    1    1    2    1    1
##  432  434  441  448  455  456  458  470  473  480  482  486  488  491  496
##    2    1    1    1    1    1    1    1    1    4    1    1    1    2    1
##  500  501  506  510  525  533  534  540  544  546  557  564  571  576  577
##    1    1    1    1    3    1    1    1    1    1    1    2    1    3    1
##  586  589  594  595  600  610  618  622  624  625  628  634  641  650  656
##    1    1    1    1    1    1    1    1    1    2    1    2    1    1    1
##  657  659  660  661  662  663  676  678  689  698  702  709  710  722  728
##    1    1    1    2    2    1    1    2    1    1    1    1    1    1    2
##  732  733  741  744  747  756  761  763  764  769  774  777  780  789  801
##    1    1    1    2    1    2    1    1    1    1    1    1    2    1    1
##  804  806  808  816  827  831  832  833  836  840  847  850  859  877  884
##    1    1    1    1    1    1    2    1    1    3    1    1    1    1    1
##  888  892  894  912  918  930  960  967  974  994 1008 1017 1026 1032 1035
##    1    1    1    2    1    1    1    1    1    1    1    1    1    1    1
## 1040 1045 1055 1093 1115 1116 1121 1128 1129 1139 1176 1194 1195 1198 1214
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1216 1217 1218 1226 1232 1250 1258 1296 1323 1324 1344 1346 1390 1393 1430
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1473 1515 1544 1560 1588 1590 1595 1604 1629 1649 1694 1768 1794 1958 2002
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
```

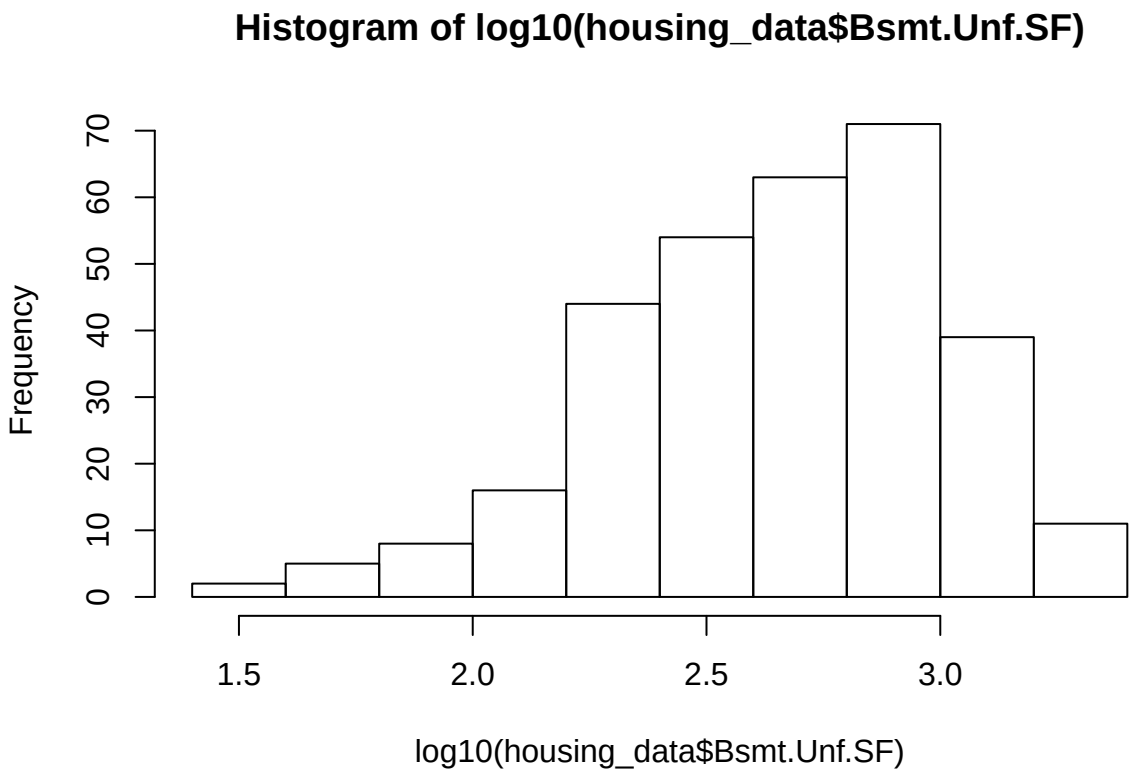
only 28 homes with 0sqft of unfinished basement, some are very small, but could be for utility room.

```
hist(housing_data$Bsmt.Unf.SF)
```



Expectedly we have a right skew to this variable.

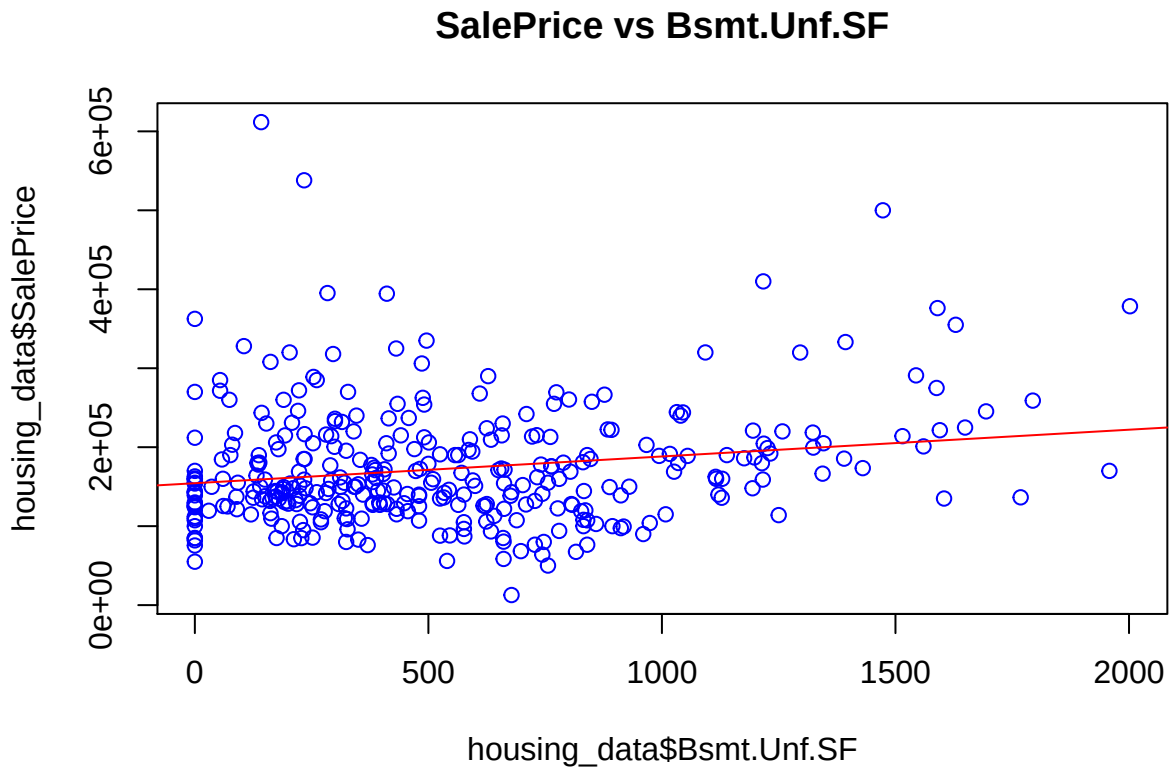
```
hist(log10(housing_data$Bsmt.Unf.SF))
```



Log

gets us closer to a nml distribution if we want to use this variable.

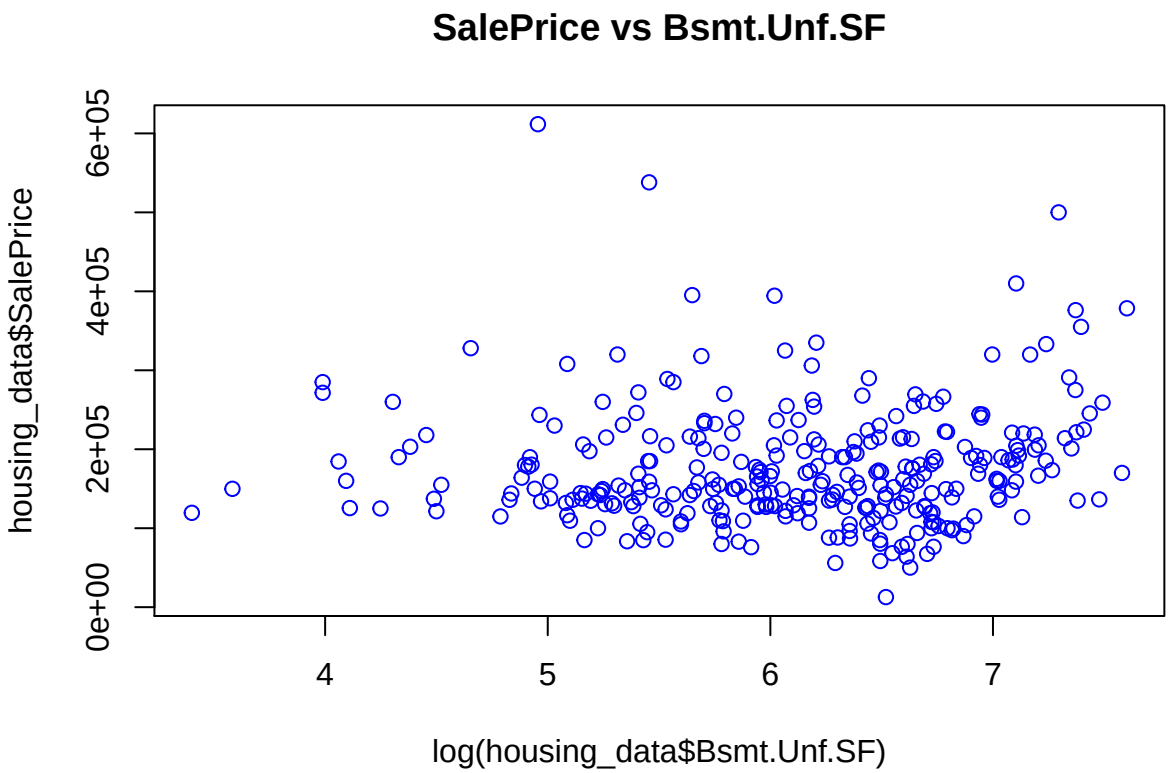
```
plot(housing_data$Bsmt.Unf.SF, housing_data$SalePrice, main = "SalePrice vs Bsmt.Unf.SF", col = "blue")
abline(lm(SalePrice~ Bsmt.Unf.SF, data= housing_data), col ="red")
```



###

Appears to be a slightly positive relationship

```
plot(log(housing_data$Bsmt.Unf.SF), housing_data$SalePrice, main = "SalePrice vs Bsmt.Unf.SF", col = "bl
```



Variable: Total.Bsmt.SF

Meaning: Total sqft of basement

```
typeof(housing_data$Total.Bsmt.SF)
```

```
## [1] "integer"
```

```
length(housing_data$Total.Bsmt.SF)
```

```
## [1] 341
```

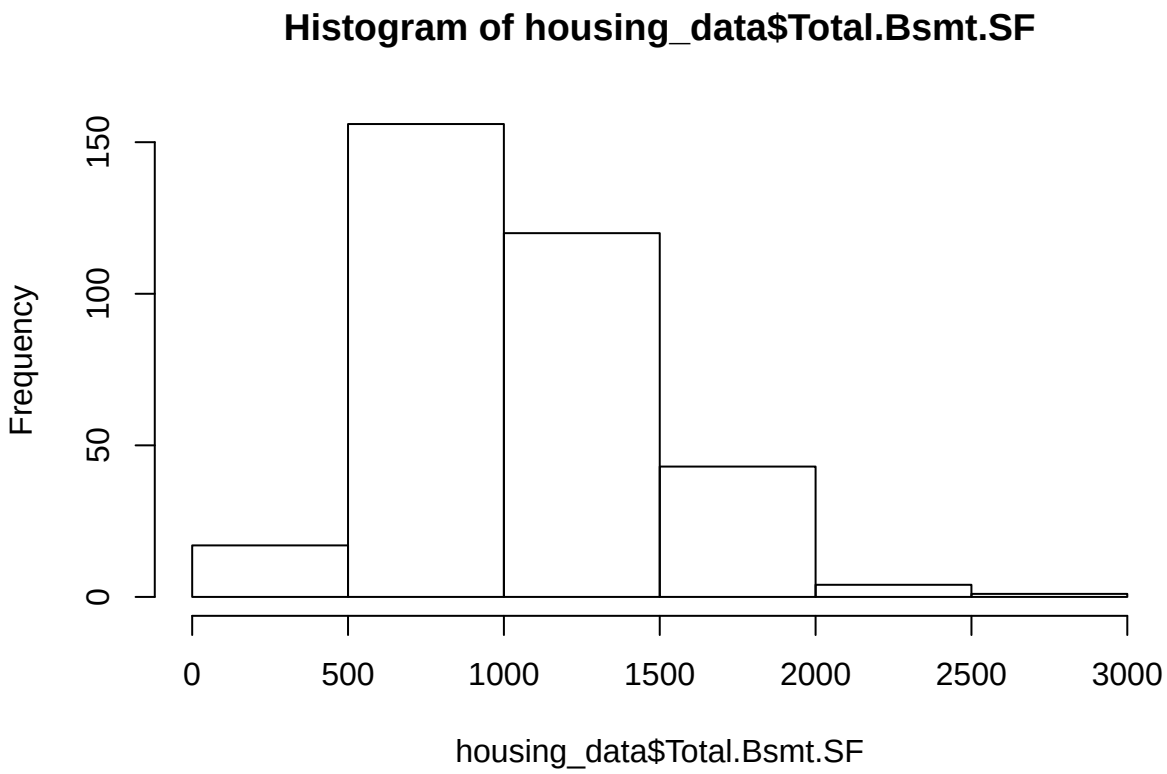
```
table(housing_data$Total.Bsmt.SF)
```

```
##
##    0  346  360  381  384  423  458  480  483  525  528  530  533  536  540
##    7    1    1    1    3    1    1    1    1    2    1    1    1    1    1
##  546  559  572  576  600  608  622  624  629  630  650  660  662  663  672
##    4    1    1    3    5    1    1    2    1    1    1    1    1    2    2
##  676  678  686  689  696  698  707  709  710  715  720  725  728  735  738
```

```
##    1    1    1    2    1    1    1    1    1    1    1    1    2    1    1
##  741  744  747  750  756  763  764  765  774  777  780  782  788  789  796
##    1    2    2    1    6    1    1    1    1    1    3    1    1    1    1
##  804  806  814  816  817  827  831  832  833  835  836  840  847  848  855
##    2    1    1    3    1    1    1    1    1    1    1    3    1    2    2
##  856  858  859  860  864  870  876  878  882  884  888  894  900  910  912
##    2    1    1    1    8    1    1    1    3    1    1    3    1    1    4
##  918  923  926  928  930  936  941  945  946  948  950  956  960  967  972
##    1    1    2    2    1    4    1    1    1    1    1    1    1    1    2
##  975  980  982  988  991  994  996 1004 1008 1010 1012 1026 1027 1029 1032
##    1    1    1    2    1    1    1    1    3    1    1    2    1    1    2
## 1039 1040 1049 1050 1053 1055 1056 1057 1060 1063 1067 1068 1069 1078 1080
##    1    4    1    1    1    1    3    1    1    1    1    1    1    2    2
## 1086 1105 1107 1108 1109 1116 1117 1121 1124 1140 1143 1144 1145 1152 1156
##    1    1    1    1    1    1    1    1    1    1    1    1    1    4    1
## 1161 1168 1172 1175 1187 1188 1191 1194 1195 1196 1206 1208 1209 1212 1214
##    1    1    1    1    1    2    1    1    1    1    1    1    1    2    1
## 1216 1218 1222 1224 1226 1231 1232 1243 1244 1246 1250 1256 1268 1280 1300
##    2    1    1    1    2    1    1    1    1    1    1    1    2    2    1
## 1306 1314 1319 1324 1328 1329 1332 1336 1338 1344 1347 1358 1370 1390 1392
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1393 1395 1398 1405 1414 1420 1422 1430 1433 1463 1468 1470 1473 1480 1488
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    2
## 1492 1501 1508 1510 1517 1528 1541 1542 1544 1560 1566 1581 1590 1594 1595
##    1    1    1    1    2    1    1    1    1    1    1    1    1    1    2
## 1604 1610 1629 1642 1649 1650 1657 1671 1673 1679 1694 1698 1704 1720 1728
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1752 1768 1822 1829 1832 1844 1856 1902 1930 1947 1958 1978 2002 2033 2110
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 2330 2846
##    1    1
```

only 7 homes have no basement, sizes here make sense.

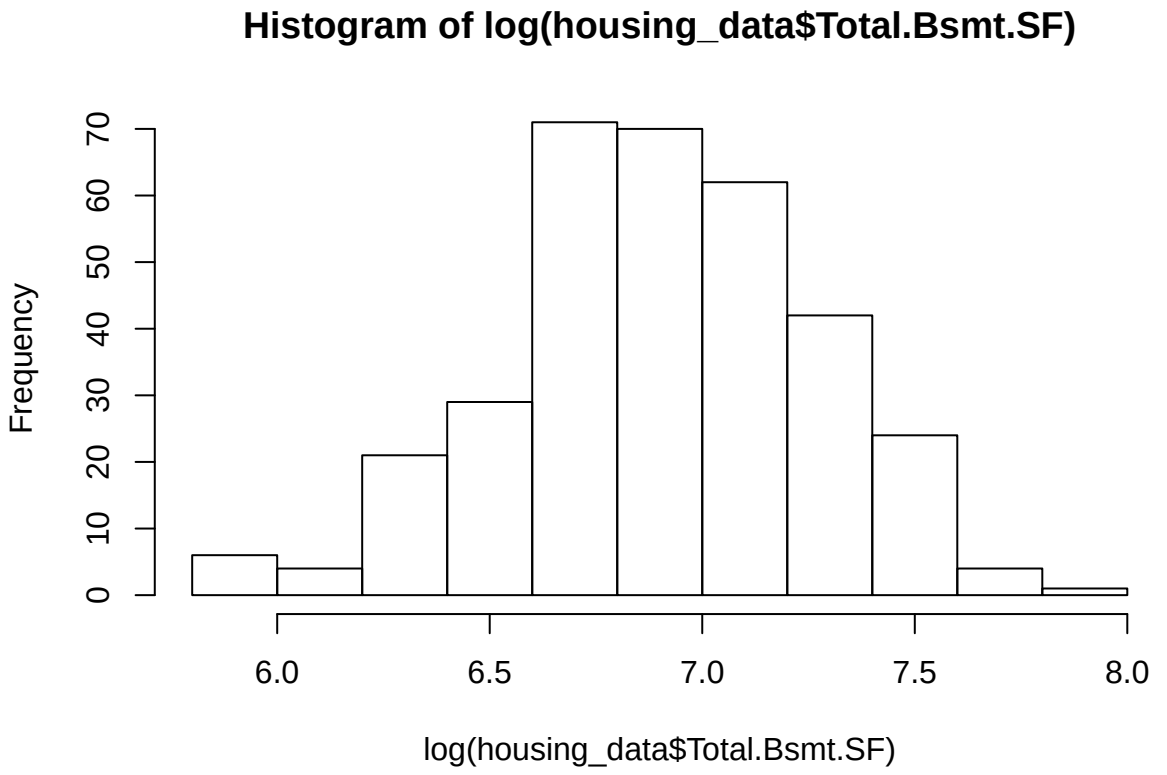
```
hist(housing_data$Total.Bsmt.SF)
```



###

Have a slight R skew, imagine this distribution closely resembles total home sqft histogram.

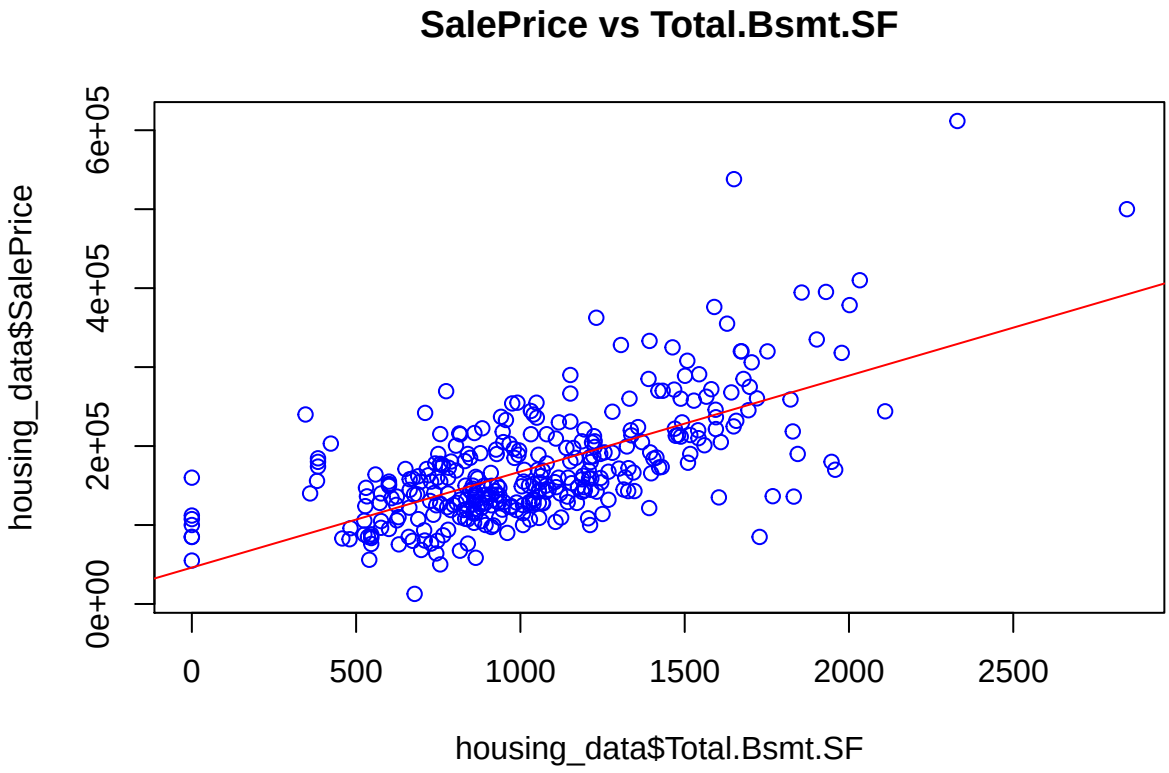
```
hist(log(housing_data$Total.Bsmt.SF))
```



###

log transformation makes this much more normal

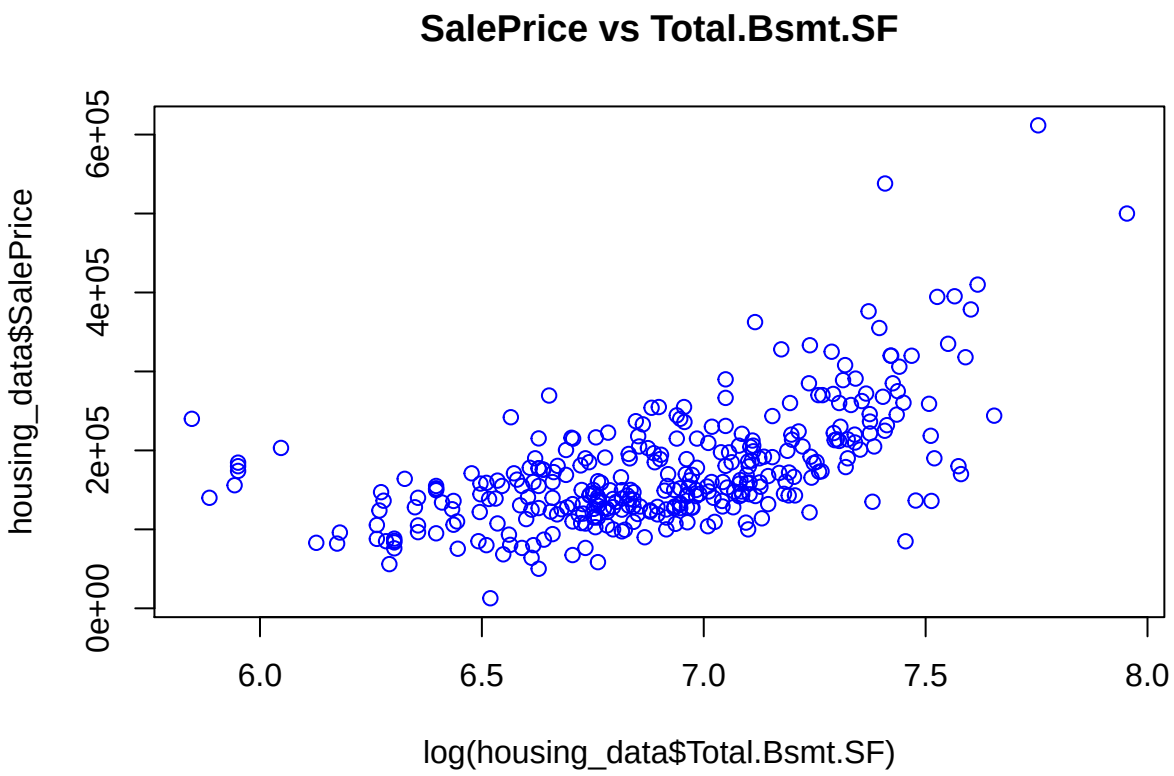
```
plot(housing_data$Total.Bsmt.SF, housing_data$SalePrice, main = "SalePrice vs Total.Bsmt.SF", col = "blu
abline(lm(SalePrice~ Total.Bsmt.SF, data = housing_data), col = "red")
```



###

Seems to be a linear relationship.

```
plot(log(housing_data$Total.Bsmt.SF), housing_data$SalePrice, main = "SalePrice vs Total.Bsmt.SF", col =
```

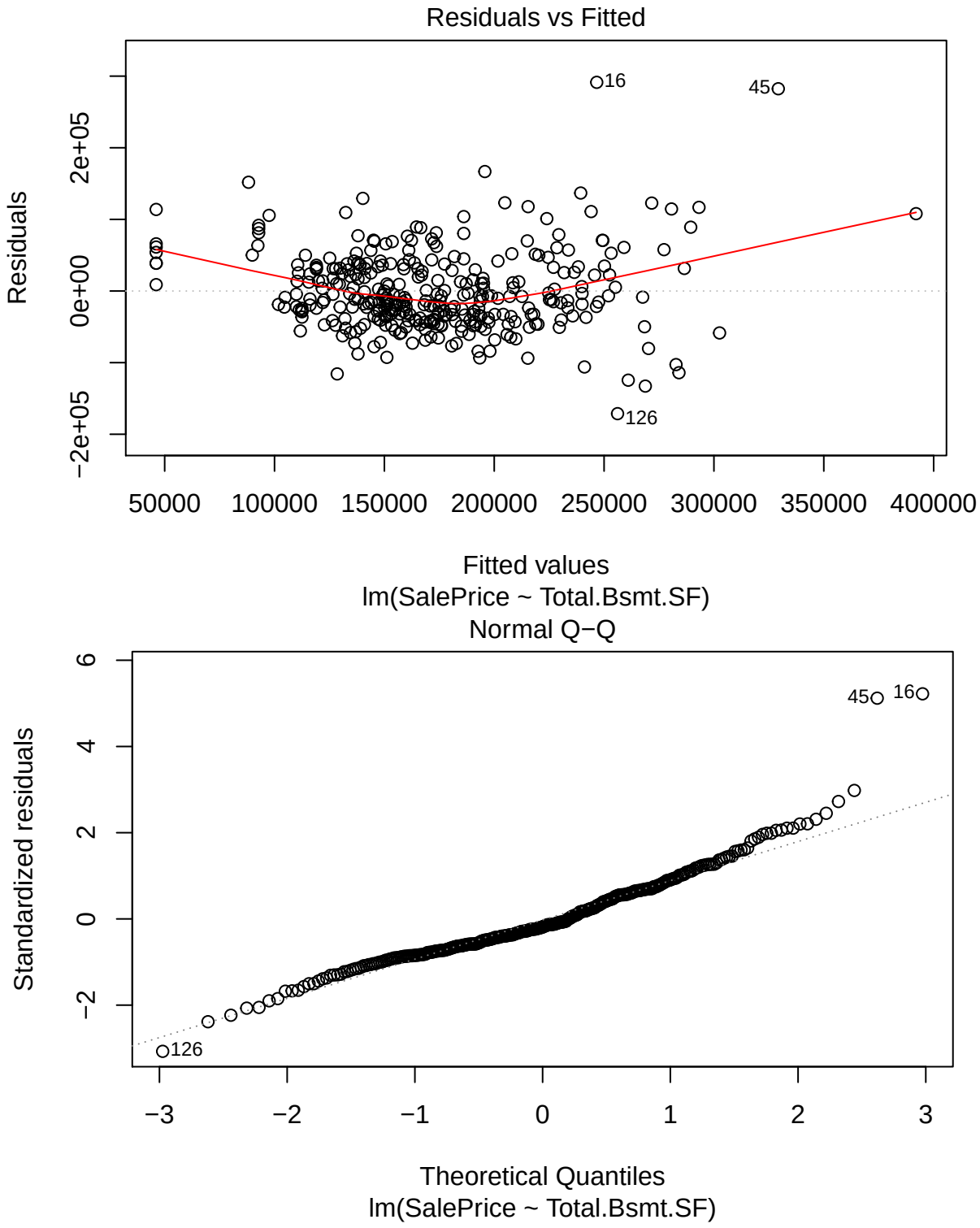


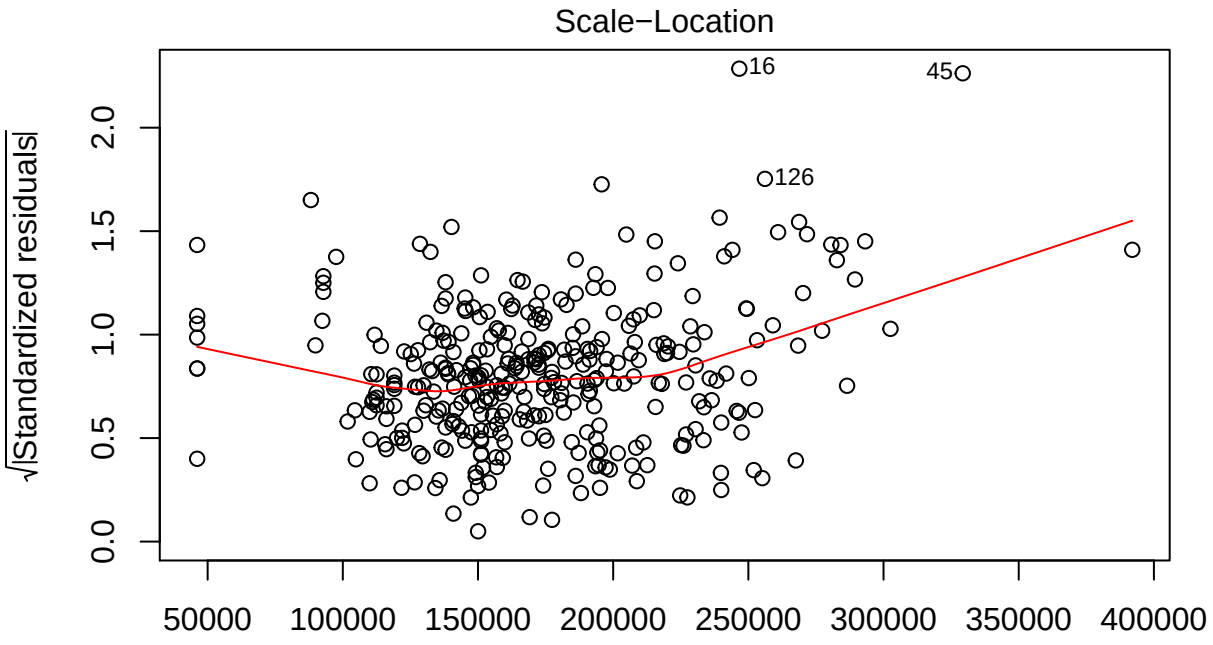

```
ModelTotBsmtSF <- lm(SalePrice ~ Total.Bsmt.SF, data =housing_data)
coeftest(ModelTotBsmtSF, vcov= vcovHC)
```

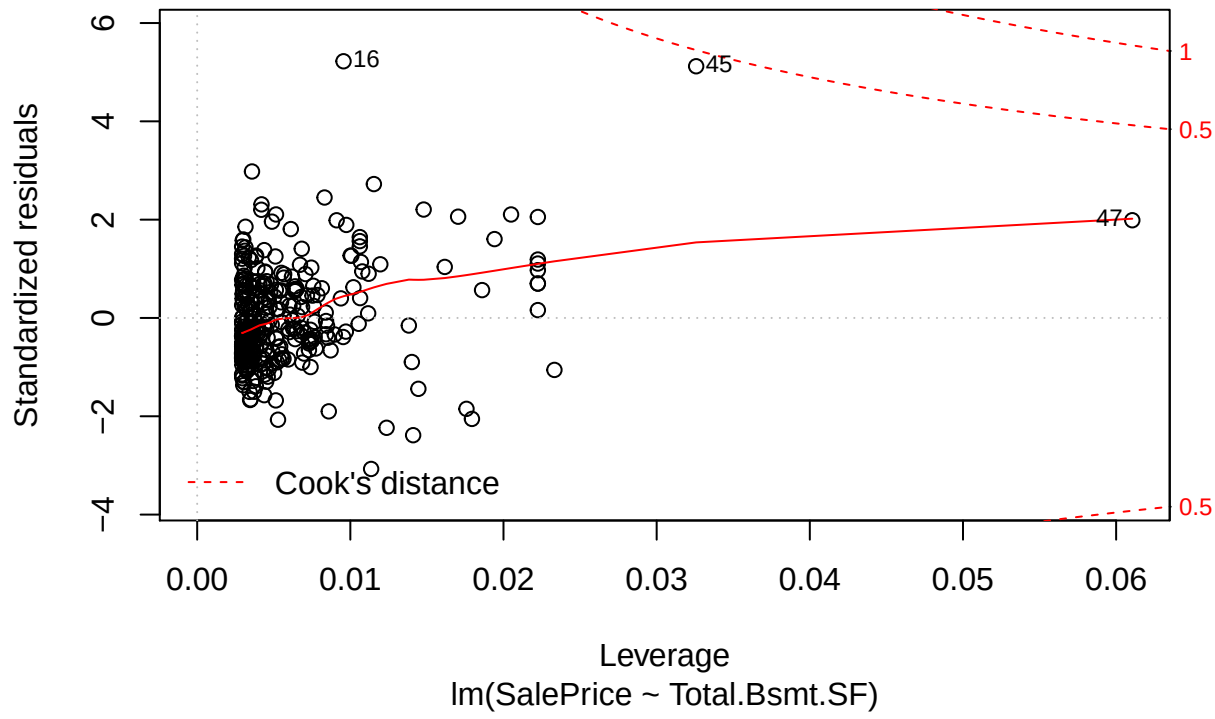
```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   46129.68   11540.90  3.9971 7.873e-05 ***
## Total.Bsmt.SF   121.53      11.94 10.1783 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

statistically significant by self.

```
plot(ModelTotBsmtSF)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SalePrice ~ Total.Bsmt.SF)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SalePrice ~ Total.Bsmt.SF)
16
45
126





Some issues with the residuals v fitted at the extremes, but not a lot of values, looks good for majority of values. QQ plot looks good, a little skew at the high end. Scale location also good in middle, but heteroskedastic at extremes. A couple values close to .5 Cook's distance, but none reach it.

Seems that bsmt sqft should be included. Possbily basement type if interaction terms give good information

---

Variable: x1st.Flr.SF

Meaning: Square footage of first floor

Note that R puts an X infront of numbers that start column names in housing_data

```
typeof(housing_data$X1st.Flr.SF)
```

```
## [1] "integer"
```

```
length(housing_data$X1st.Flr.SF)
```

```
## [1] 341
```

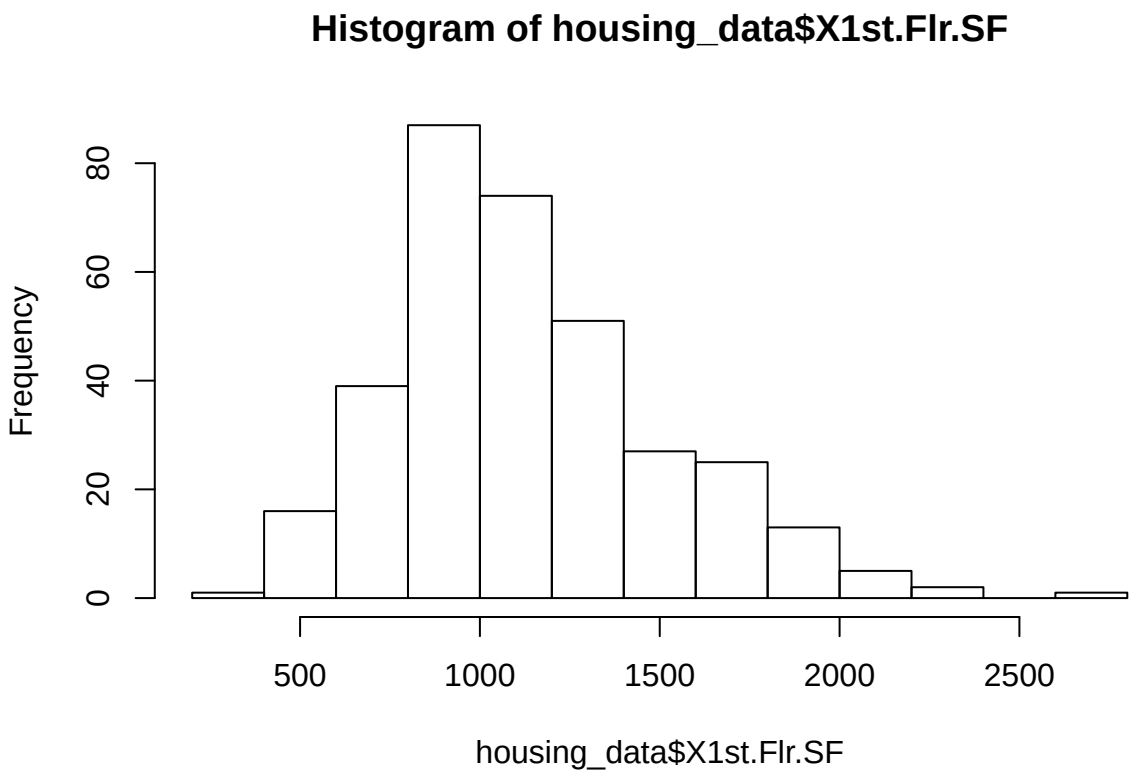
```
table(housing_data$X1st.Flr.SF)
```

```
##
##  372  483  525  530  536  540  546  548  600  608  624  630  640  662  663
##    1    1    2    1    1    1    4    1    5    1    1    1    1    1    1
##  672  689  696  698  720  725  727  741  744  747  752  756  763  764  765
##    2    1    1    1    2    2    1    1    2    1    1    2    1    2    1
##  769  774  780  788  789  792  796  798  804  814  816  824  827  831  832
##    1    2    3    1    1    2    1    1    2    1    2    1    2    2    2
##  833  835  836  840  841  847  848  855  856  858  859  860  862  864  868
##    1    1    1    2    1    1    3    2    1    1    1    1    1    5    1
##  874  876  882  884  888  892  894  896  900  902  910  912  918  923  926
##    1    1    1    1    2    2    3    2    2    1    1    3    1    1    2
##  928  930  936  941  942  945  952  955  956  960  964  965  966  972  975
##    2    3    3    1    1    1    1    1    1    1    1    1    1    1    1
##  977  980  985  988  991  992  993  996 1004 1012 1014 1019 1026 1027 1028
##    1    1    1    1    1    1    1    1    1    1    1    1    2    1    2
## 1030 1032 1034 1036 1039 1040 1044 1051 1052 1054 1055 1056 1057 1060 1061
##    1    1    2    1    1    2    2    1    2    1    1    2    1    2    1
## 1062 1063 1064 1067 1068 1069 1074 1078 1080 1086 1096 1097 1102 1105 1116
##    1    1    1    1    1    1    1    2    1    1    1    1    1    1    1
## 1117 1121 1128 1131 1143 1144 1145 1152 1155 1157 1160 1164 1169 1172 1173
##    2    1    2    1    1    1    1    2    1    2    1    1    1    1    1
## 1175 1187 1188 1191 1194 1195 1196 1206 1207 1208 1209 1212 1214 1216 1218
##    1    1    3    1    1    1    1    1    1    1    1    2    1    1    1
## 1222 1225 1226 1232 1236 1246 1251 1264 1268 1269 1280 1285 1287 1296 1298
##    2    2    2    1    1    2    1    2    1    1    2    1    2    1    1
## 1306 1314 1318 1324 1328 1329 1332 1337 1338 1341 1344 1346 1347 1358 1370
##    2    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1373 1381 1392 1395 1402 1414 1418 1422 1430 1433 1468 1478 1480 1483 1484
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
```

```
## 1488 1492 1494 1500 1502 1508 1510 1520 1526 1535 1541 1544 1560 1566 1580
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1595 1601 1604 1610 1616 1627 1645 1646 1656 1661 1664 1668 1675 1687 1690
##    1    1    1    1    1    1    1    1    1    1    1    1    1    2    1
## 1694 1698 1699 1700 1704 1720 1728 1752 1768 1801 1803 1822 1829 1832 1844
##    1    1    2    1    1    1    1    1    1    1    1    1    1    1    1
## 1856 1888 1898 1902 1929 1940 1978 2018 2048 2053 2073 2110 2207 2364 2696
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
```

smallest home has 372 sqft on first floor, this is small, but could make sense.

```
hist(housing_data$X1st.Flr.SF)
```

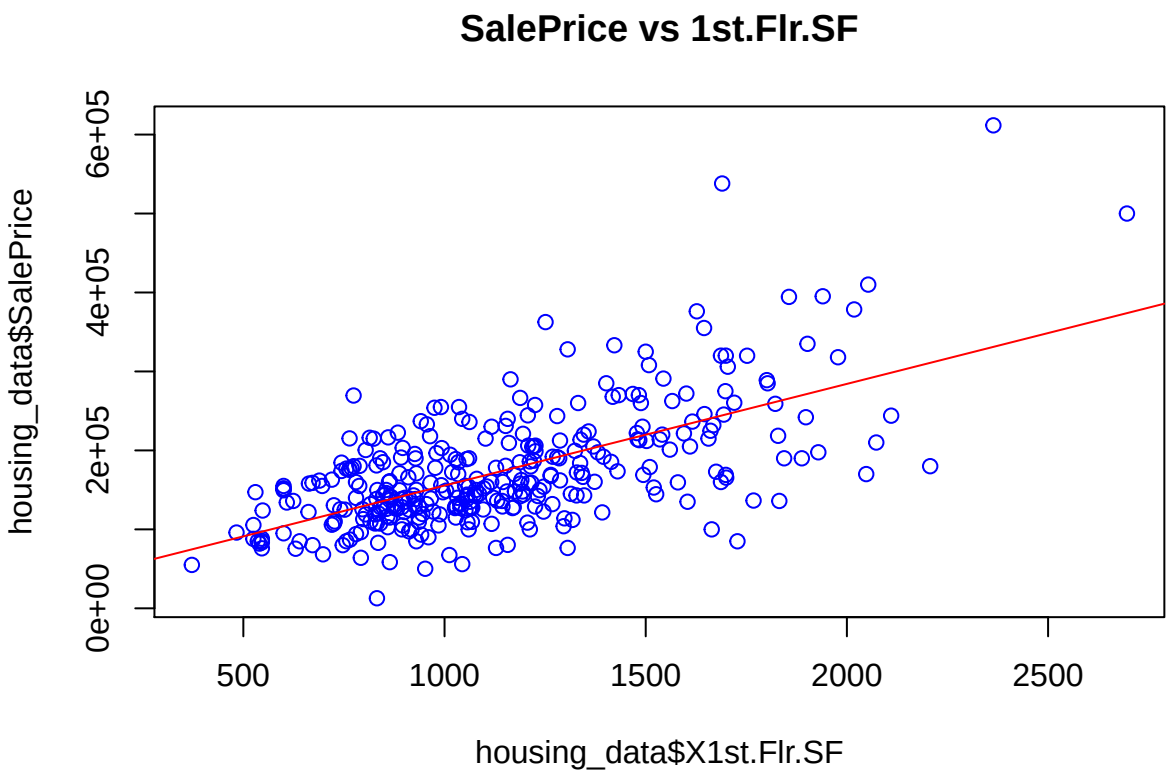


appears to be close to normally distributed, slight skew R.

```
summary(housing_data$X1st.Flr.SF)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     372     864    1060    1133    1332    2696
```

```
plot(housing_data$X1st.Flr.SF, housing_data$SalePrice, main="SalePrice vs 1st.Flr.SF", col="blue")
abline(lm(SalePrice ~ X1st.Flr.SF, data= housing_data), col="red")
```



###

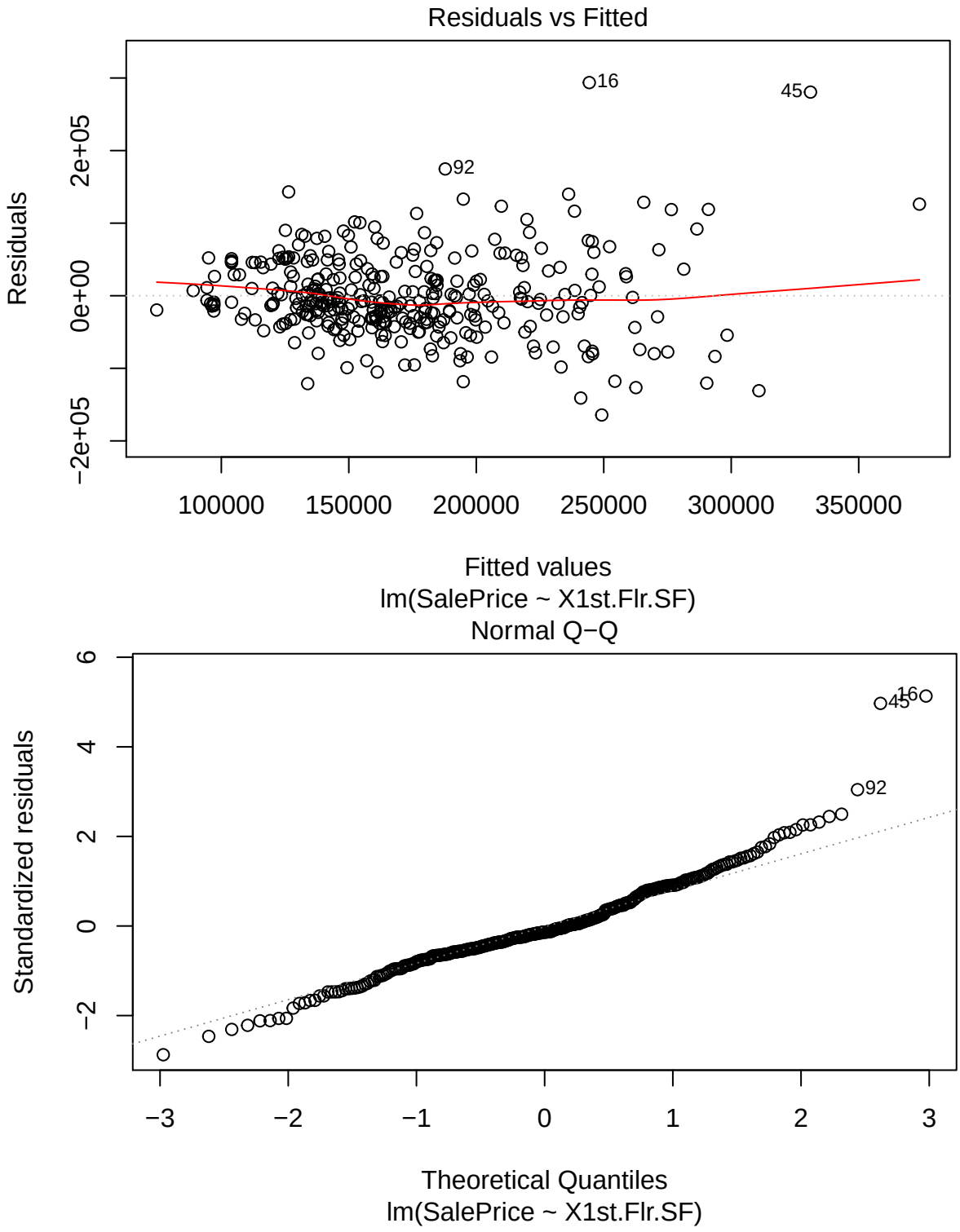
Appears to be a strong linear relationship, which makes sense intuitively.

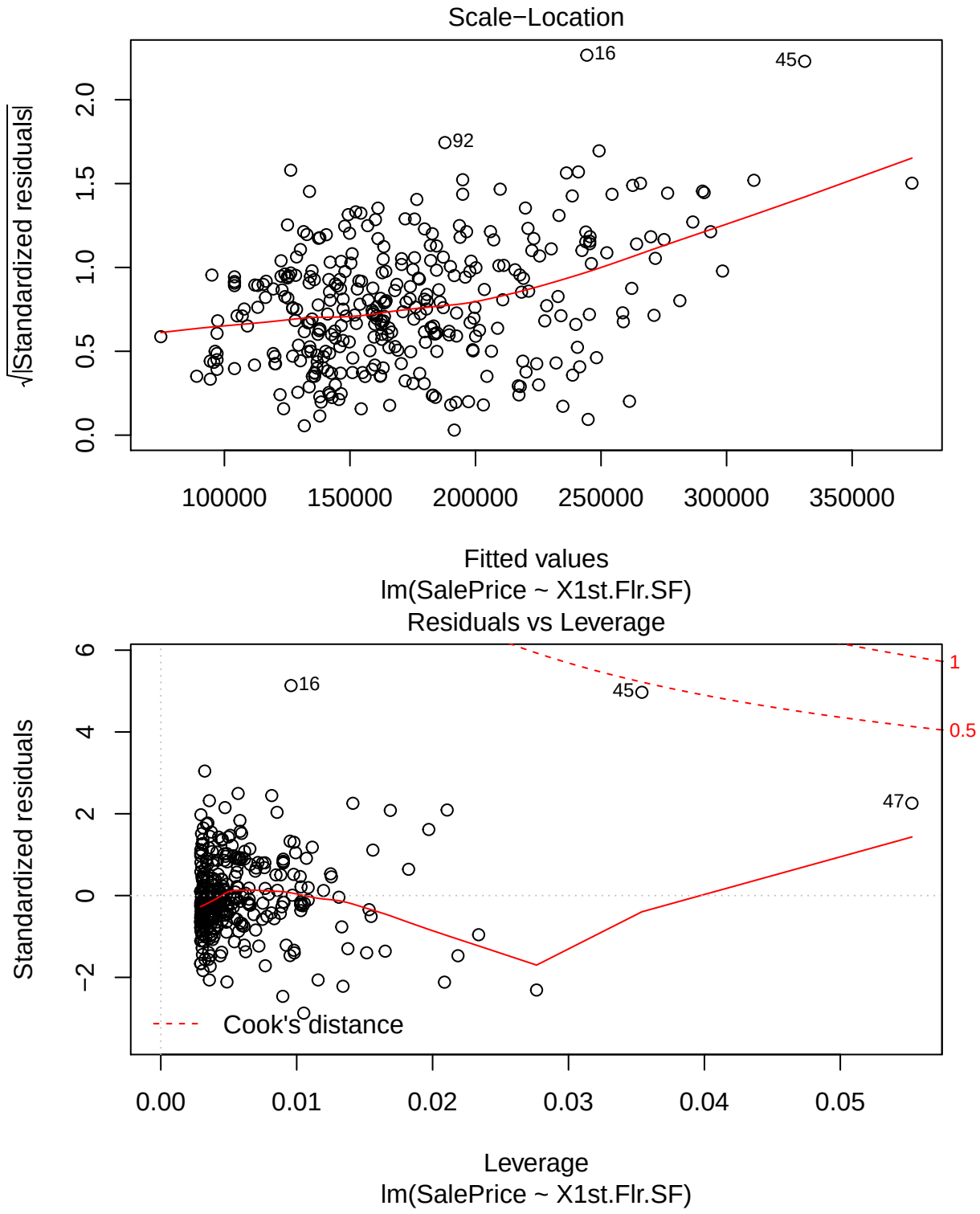
```
Model1stFlrSqft <- lm(SalePrice ~ X1st.Flr.SF, data= housing_data)
coeftest(Model1stFlrSqft, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26776.056  13792.348  1.9414  0.05304 .
## X1st.Flr.SF   128.731     13.423  9.5900  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

T 9.59 and P > 2e-16

```
plot(Model1stFlrSqft)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SalePrice ~ X1st.Flr.SF)
16
45
92
2e+05
0e+00
−2e+05
100000
150000
200000
250000
300000
350000
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SalePrice ~ X1st.Flr.SF)
6
4
2
0
−2
−3
−2
−1
0
1
2
3



Some heteroskedasticity in residuals v fitted. QQ plot good, but some variance at extremes. How do we interpret the scale location? Residual v leverage okay, #45 still close to Cook's distance. I'd imagine that this somewhat matters, but this is will closely correlated with total sqft.

Variable: x2nd.Flr.SF

Meaning: Total square footage of second floor

```
typeof(housing_data$X2nd.Flr.SF)
```

```
## [1] "integer"
```

```
length(housing_data$X2nd.Flr.SF)
```

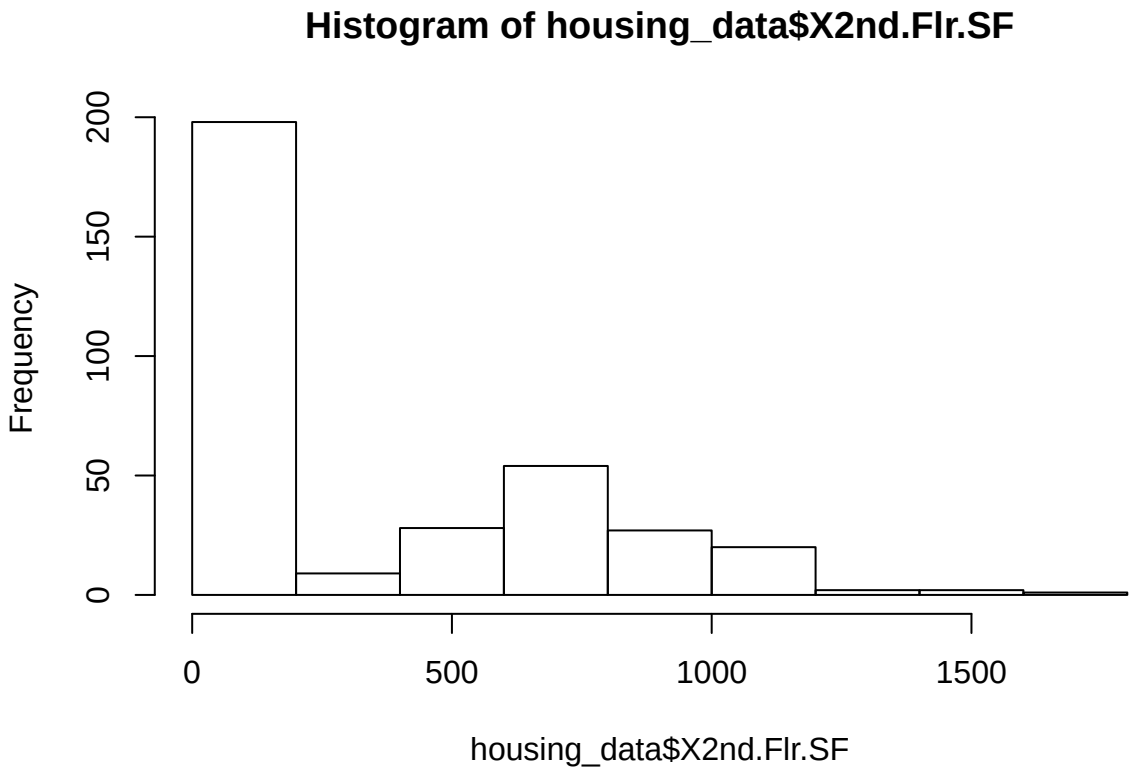
```
## [1] 341
```

```
table(housing_data$X2nd.Flr.SF)
```

```
##
##    0  180  185  192  240  252  319  328  348  358  380  408  424  430  475
##  195    1    1    1    1    1    2    1    2    1    1    1    1    1    1
##  492  498  499  504  505  524  532  537  546  550  558  563  567  576  582
##    1    1    1    1    1    1    1    1    5    1    1    1    2    1    1
##  584  600  601  602  604  606  608  614  615  622  630  636  644  645  650
##    1    3    2    1    1    1    1    1    1    1    1    3    2    1    1
##  656  662  665  672  676  678  686  687  689  700  701  703  707  714  715
##    1    1    1    1    2    1    1    1    2    1    1    1    1    1    1
##  720  730  734  744  748  754  755  756  765  772  776  780  783  788  790
##    3    1    1    1    1    1    1    2    1    1    1    2    1    1    1
##  800  804  806  808  823  828  830  838  840  860  862  864  873  878  880
##    1    1    1    1    1    1    1    1    1    1    2    1    1    1    1
##  886  887  888  892  908  912  915  942  954  956 1044 1054 1070 1074 1075
##    1    1    2    1    2    1    1    1    1    1    1    1    1    1    1
## 1080 1098 1100 1106 1111 1122 1128 1151 1152 1169 1177 1185 1194 1196 1215
##    1    1    1    1    1    1    2    1    1    1    1    1    1    1    1
## 1216 1523 1589 1788
##    1    1    1    1
```

195 homes without a second floor. Also, many have small second floor.

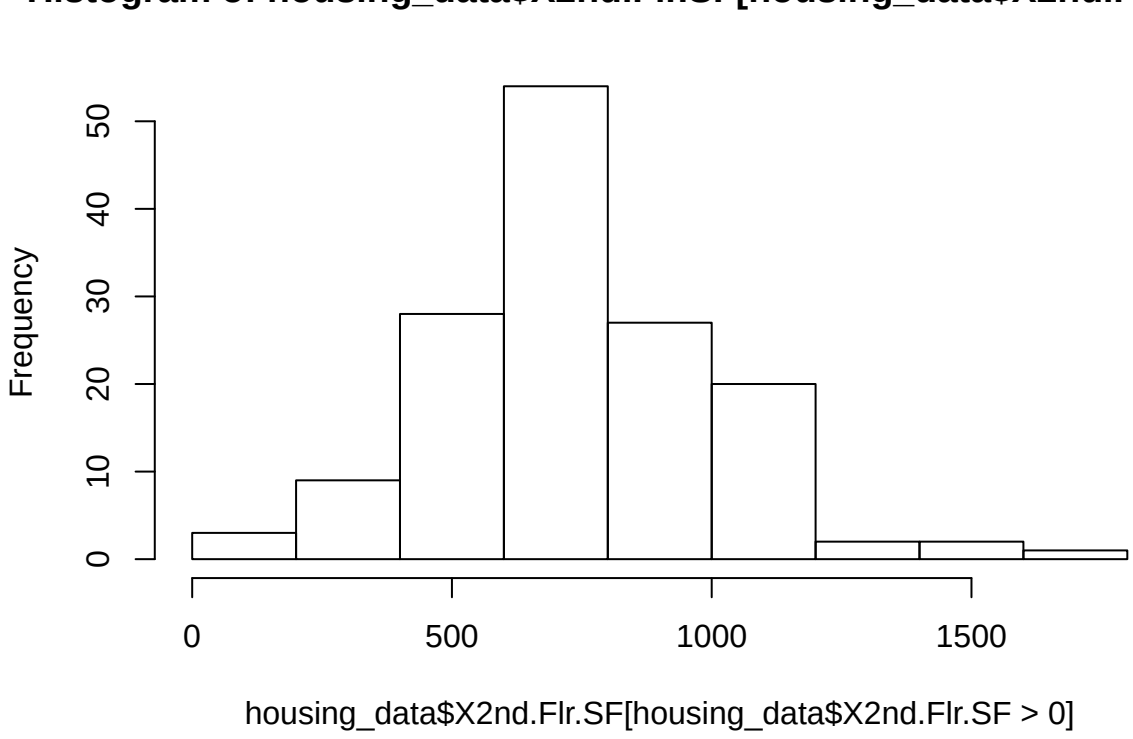
```
hist(housing_data$X2nd.Flr.SF)
```



###

Skewed due to many zeros.

```
hist(housing_data$X2nd.Flr.SF[housing_data$X2nd.Flr.SF>0])
```



```
summary(housing_data$X2nd.Flr.SF)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0     320     672    1788
```

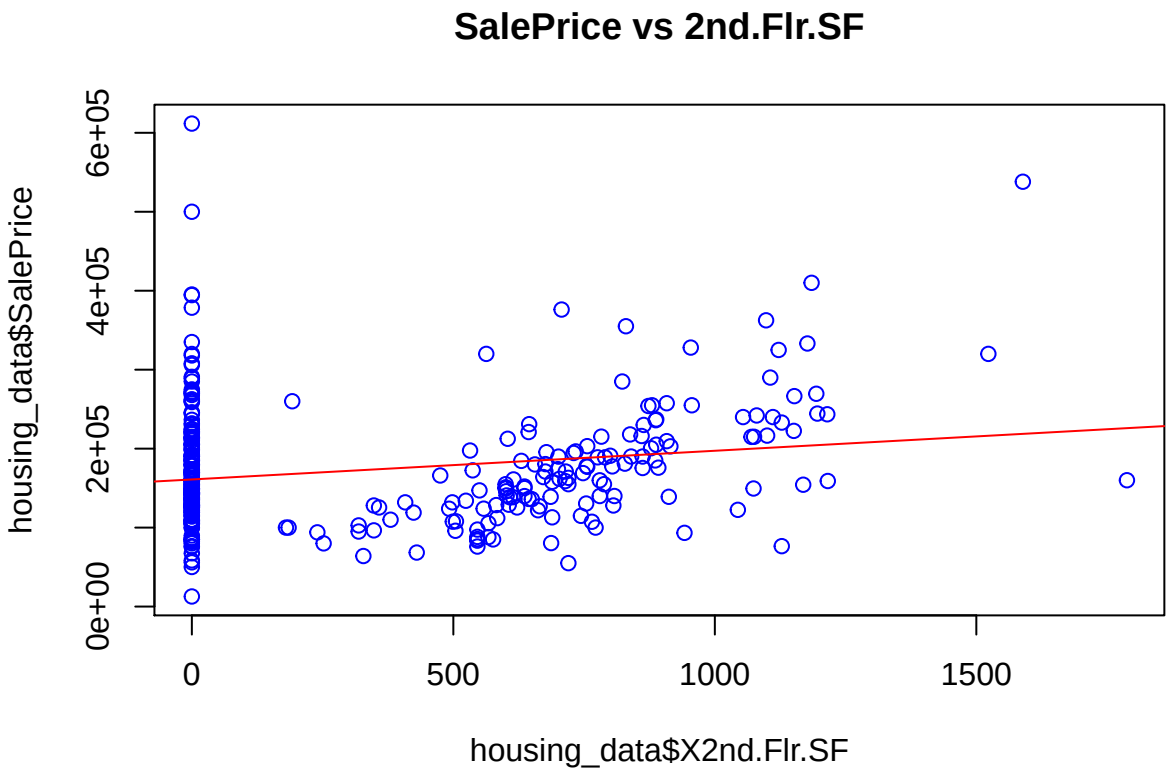
Mean 320, but median 0

```
summary(housing_data$X2nd.Flr.SF[housing_data$X2nd.Flr.SF>0])
```

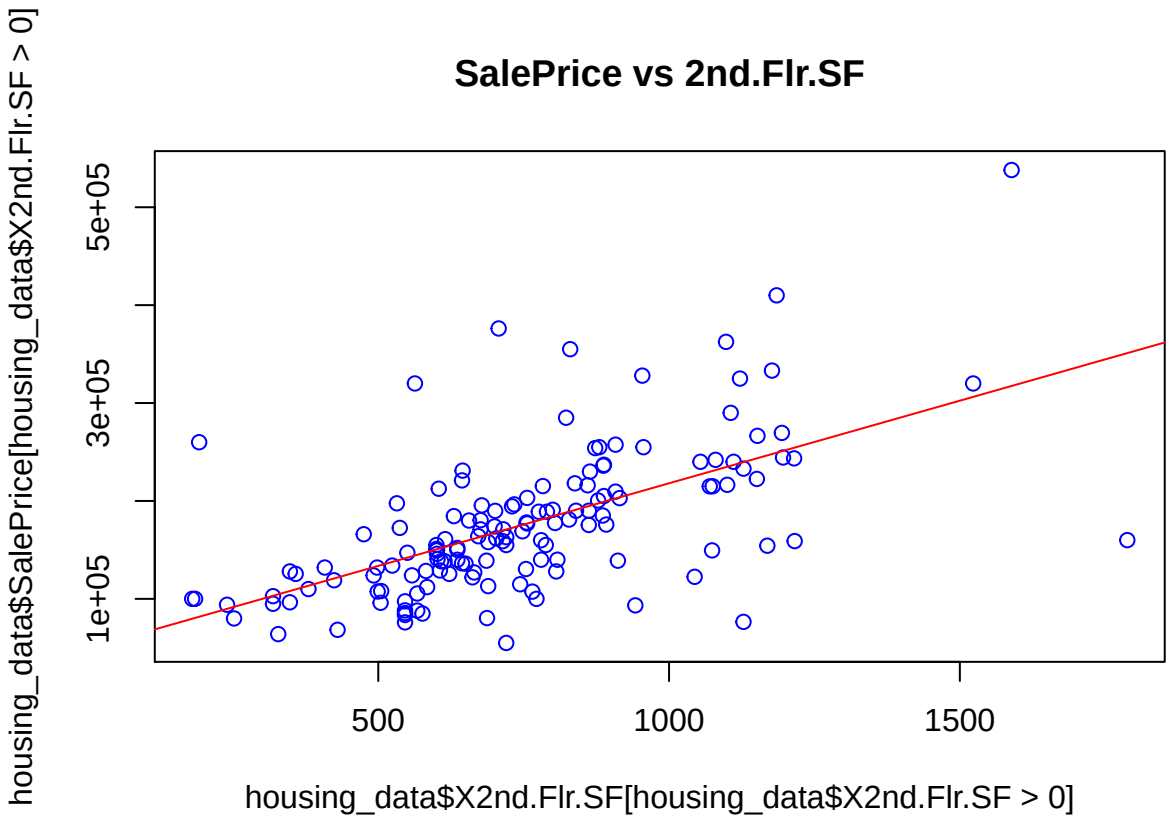
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   180.0   588.0   714.5   747.5   884.5  1788.0
```

Mean and median close for homes with a second floor, which is consistent with thoughts of being close to normally distributed.

```
plot(housing_data$X2nd.Flr.SF, housing_data$SalePrice, main="SalePrice vs 2nd.Flr.SF", col="blue")
abline(lm(SalePrice~X2nd.Flr.SF, data=housing_data), col = "red")
```




```
plot(housing_data$X2nd.Flr.SF[housing_data$X2nd.Flr.SF>0], housing_data$SalePrice[housing_data$X2nd.Flr
abline(lm(SalePrice[housing_data$X2nd.Flr.SF>0]~X2nd.Flr.SF[housing_data$X2nd.Flr.SF>0], data=housing_da
```



###

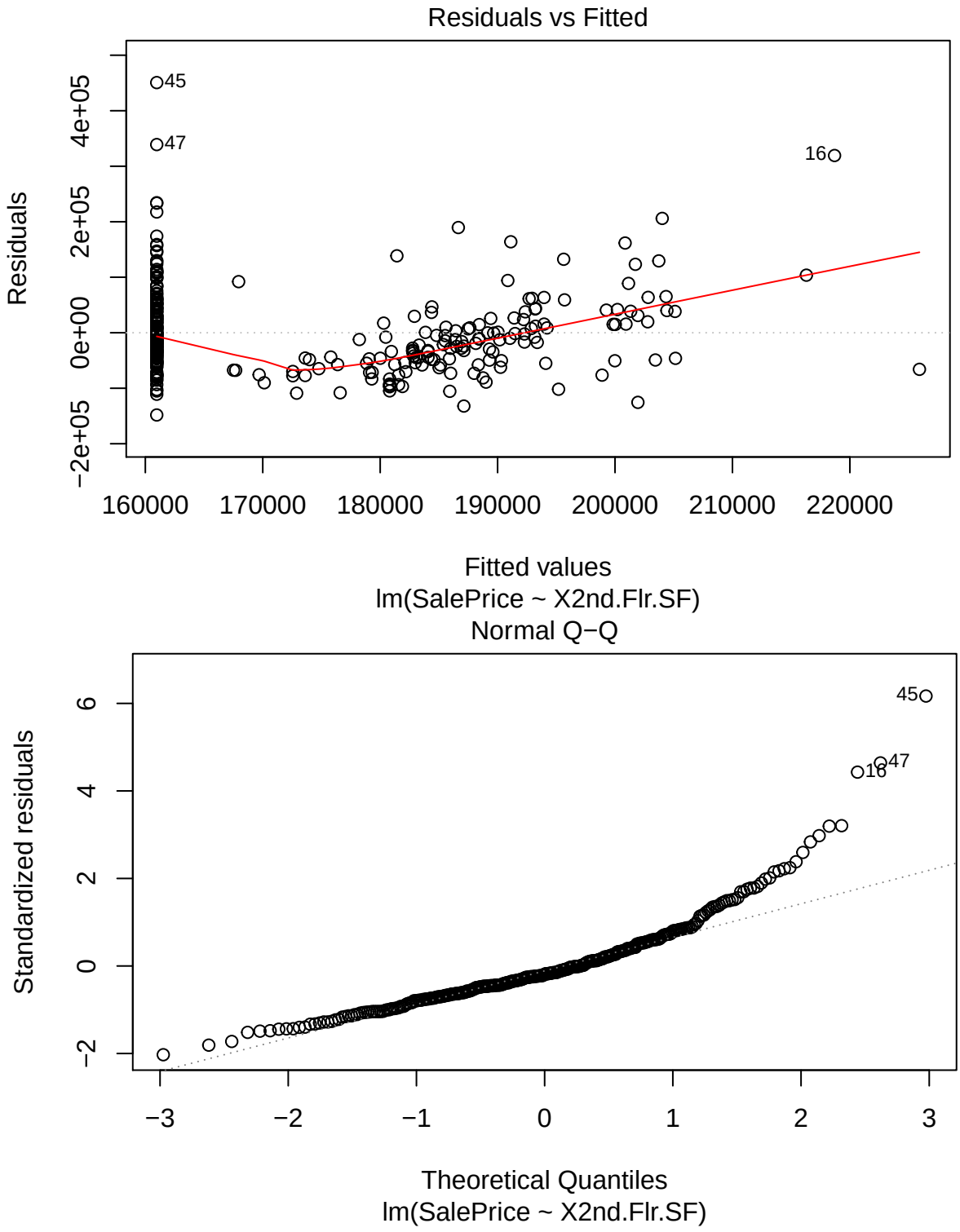
By removing the homes without a second floor, we see a somewhat linear relationship.

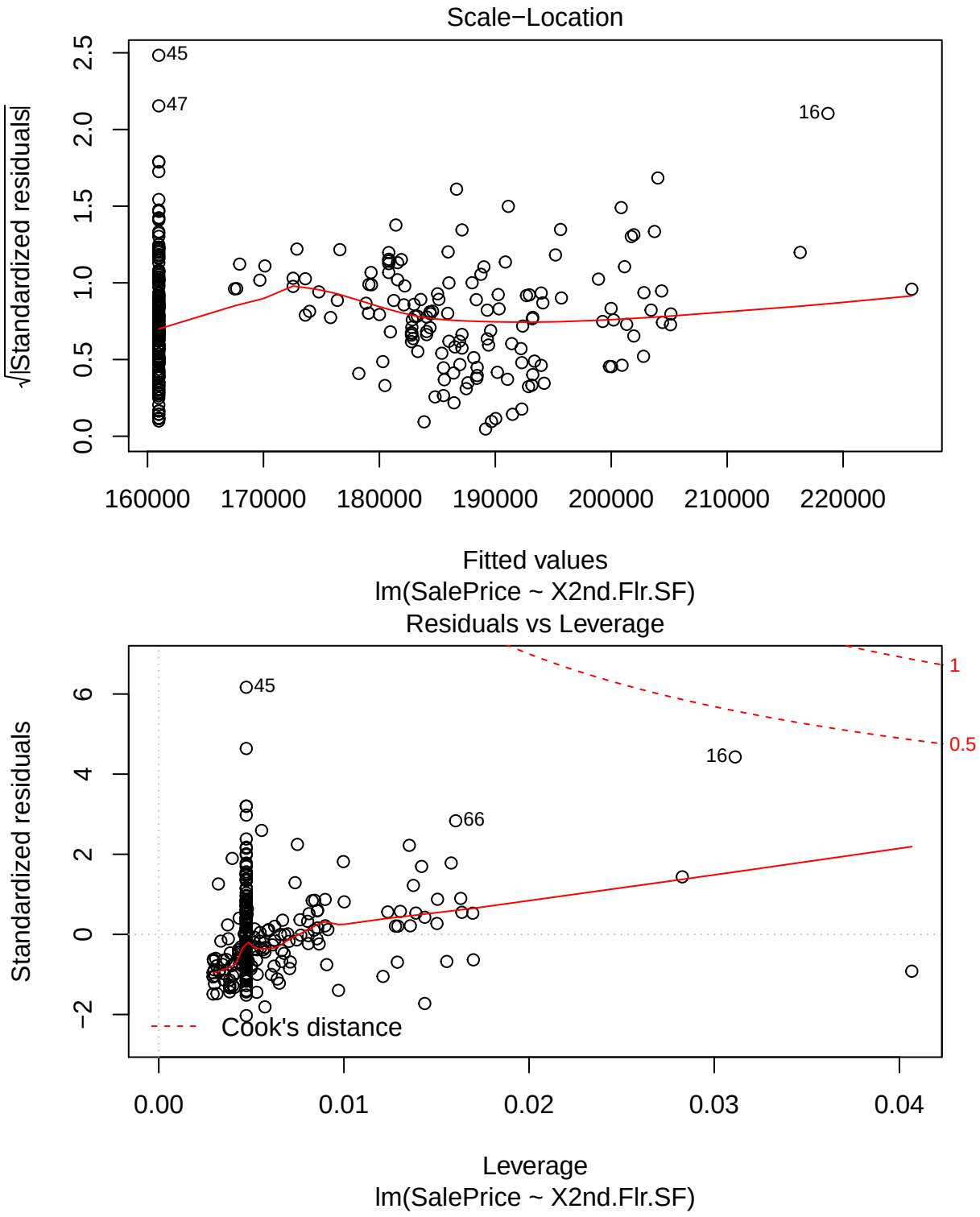
```
Model2ndFlrSqft <- lm(SalePrice~X2nd.Flr.SF, data=housing_data)
coeftest(Model2ndFlrSqft, vcov= vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 160971.460   5346.554 30.1075 < 2.2e-16 ***
## X2nd.Flr.SF     36.328     12.166  2.9861  0.003031 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Statistically significant at .01

```
plot(Model2ndFlrSqft)
```


Residuals vs Fitted
Residuals
Fitted values
lm(SalePrice ~ X2nd.Flr.SF)
45
47
16
−2e+05
0e+00
2e+05
4e+05
160000
170000
180000
190000
200000
210000
220000
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SalePrice ~ X2nd.Flr.SF)
−2
0
2
4
6
−3
−2
−1
0
1
2
3



Some issues in residuals v fitted. QQ good between -1 and 1, then skews. Scale location also issue with heteroskedasticity.

Variable: Low.Qual.Fin.SF

Meaning: Total square footage of low quality finish for all floors.

```
typeof(housing_data$Low.Qual.Fin.SF)
```

```
## [1] "integer"
```

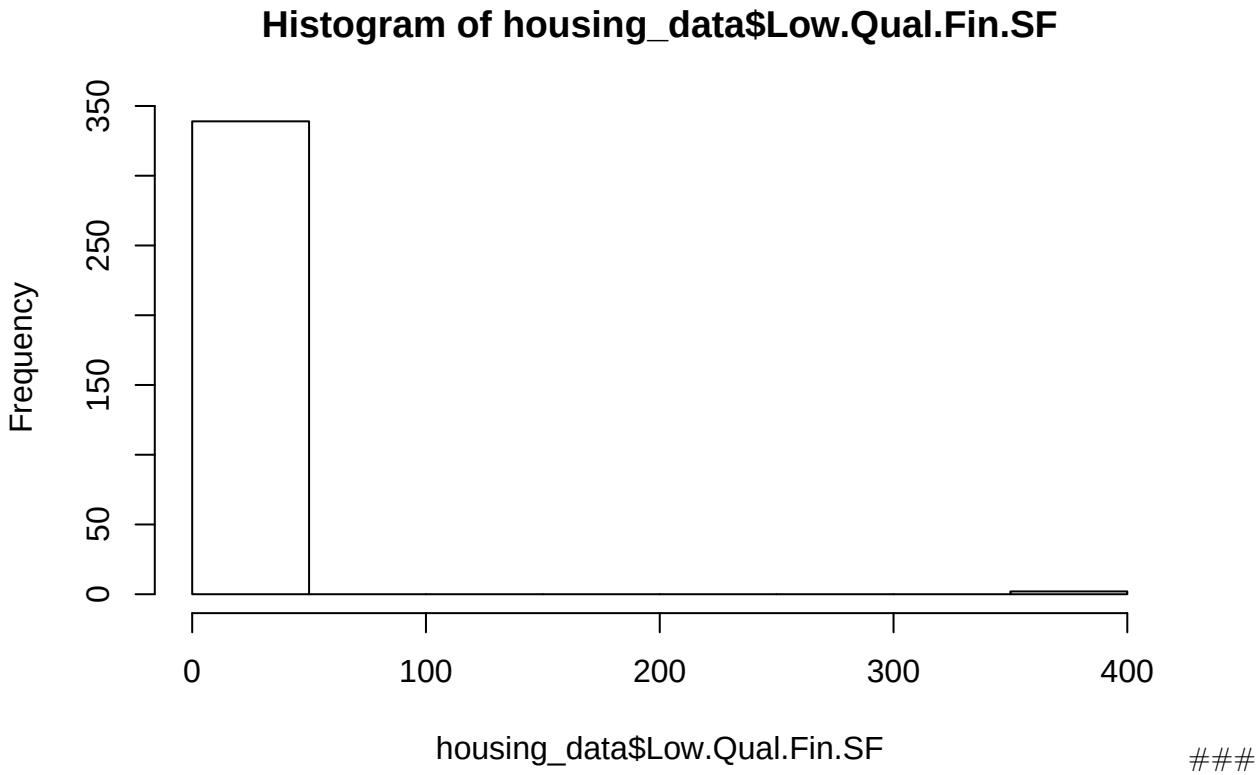
```
length(housing_data$Low.Qual.Fin.SF)
```

```
## [1] 341
```

```
table(housing_data$Low.Qual.Fin.SF)
```

```
##
##   0 362 390
## 339   1   1
```

```
hist(housing_data$Low.Qual.Fin.SF)
```



###

There are only two non-zero values. No need to go further here.

Variable: Pool Area

Meaning: Pool area in sqft, unclear if that means area around pool or just pool itself.

```
typeof(housing_data$Pool.Area)
```

```
## [1] "integer"
```

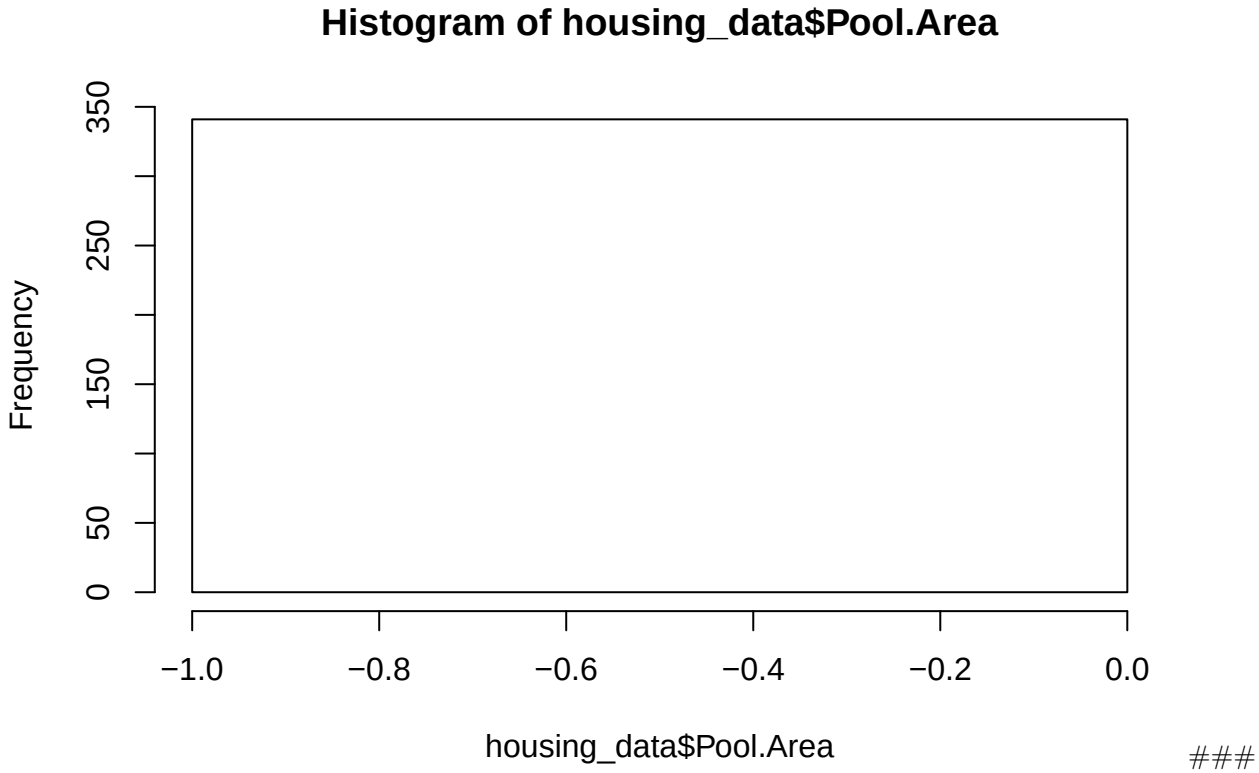
```
length(housing_data$Pool.Area)
```

```
## [1] 341
```

```
table(housing_data$Pool.Area)
```

```
##
##   0
## 341
```

```
hist(housing_data$Pool.Area)
```



###

None of these homes have pool areas, nothing to do here.

Variable: House Style

Style of the home based upon number of levels and whether or not finished.

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

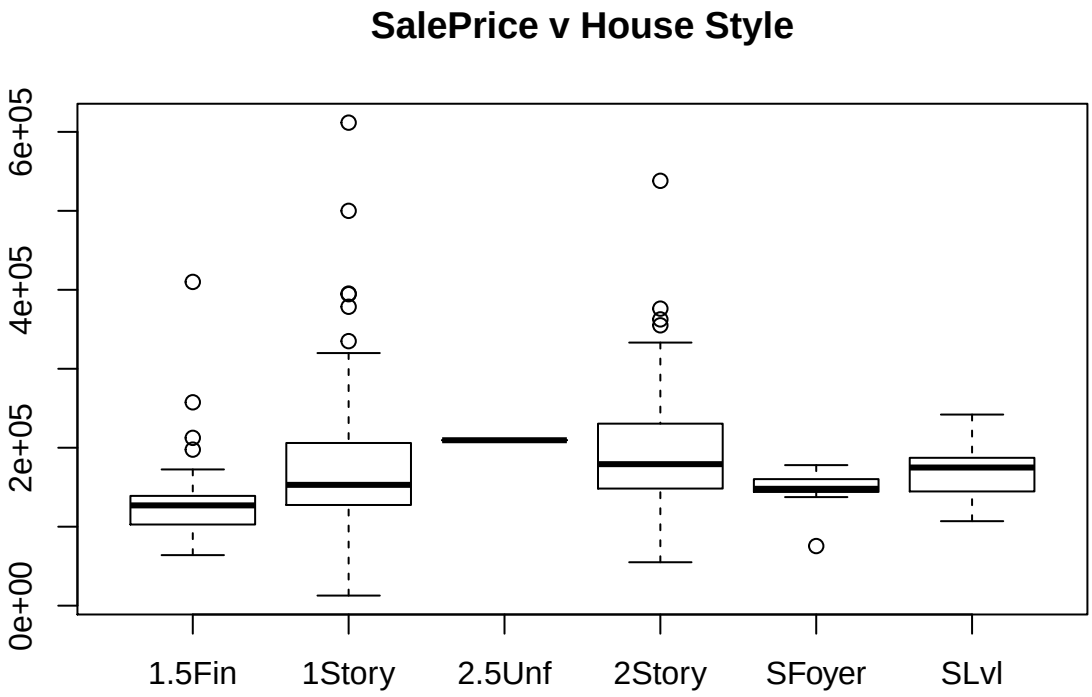
```
length(housing_data$House.Style)
```

```
## [1] 341
```

```
table(housing_data$House.Style)
```

```
##
## 1.5Fin 1Story 2.5Unf 2Story SFoyer   SLvl
##     45    178      1     96      9     12
```

```
plot(housing_data$House.Style, housing_data$SalePrice, main = "SalePrice v House Style")
```



```r
x1<- mean(housing_data$SalePrice[housing_data$House.Style=="1.5Fin"])
y1<-sd(housing_data$SalePrice[housing_data$House.Style=="1.5Fin"])

x2<-mean(housing_data$SalePrice[housing_data$House.Style=="1Story"])
y2<-sd(housing_data$SalePrice[housing_data$House.Style=="1Story"])

x3<-mean(housing_data$SalePrice[housing_data$House.Style=="2.5Unf"])
y3<-sd(housing_data$SalePrice[housing_data$House.Style=="2.5Unf"])

x4<-mean(housing_data$SalePrice[housing_data$House.Style=="2Story"])
y4<-sd(housing_data$SalePrice[housing_data$House.Style=="2Story"])

x5<-mean(housing_data$SalePrice[housing_data$House.Style=="SFoyer"])
y5<-sd(housing_data$SalePrice[housing_data$House.Style=="SFoyer"])

x6<-mean(housing_data$SalePrice[housing_data$House.Style=="SLvl"])
y6<-sd(housing_data$SalePrice[housing_data$House.Style=="SLvl"])

xMeans <-c(x1, x2, x3, x4, x5, x6)
yStdDev <- c(y1, y2, y3, y4, y5, y6)
housetypes <- c("1.5Fin", "1Story", "2.5Unf", "2Story", "SFoyer", "SLvl")
cbind(housetypes, xMeans,yStdDev)
```

```
##      housetypes xMeans             yStdDev
## [1,] "1.5Fin"   "133761.688888889" "55715.2159173703"
## [2,] "1Story"   "173285.241573034" "77392.5603013116"
## [3,] "2.5Unf"   "209500"           NA
## [4,] "2Story"   "192046.322916667" "77008.9256600537"
## [5,] "SFoyer"   "146583.333333333" "29623.6814052541"
## [6,] "SLvl"     "168877.916666667" "36572.5423245819"
```

Variable: Roof.Style

Meaning: Style of roof category

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

```r
typeof(housing_data$Roof.Style)
```

```
## [1] "integer"
```

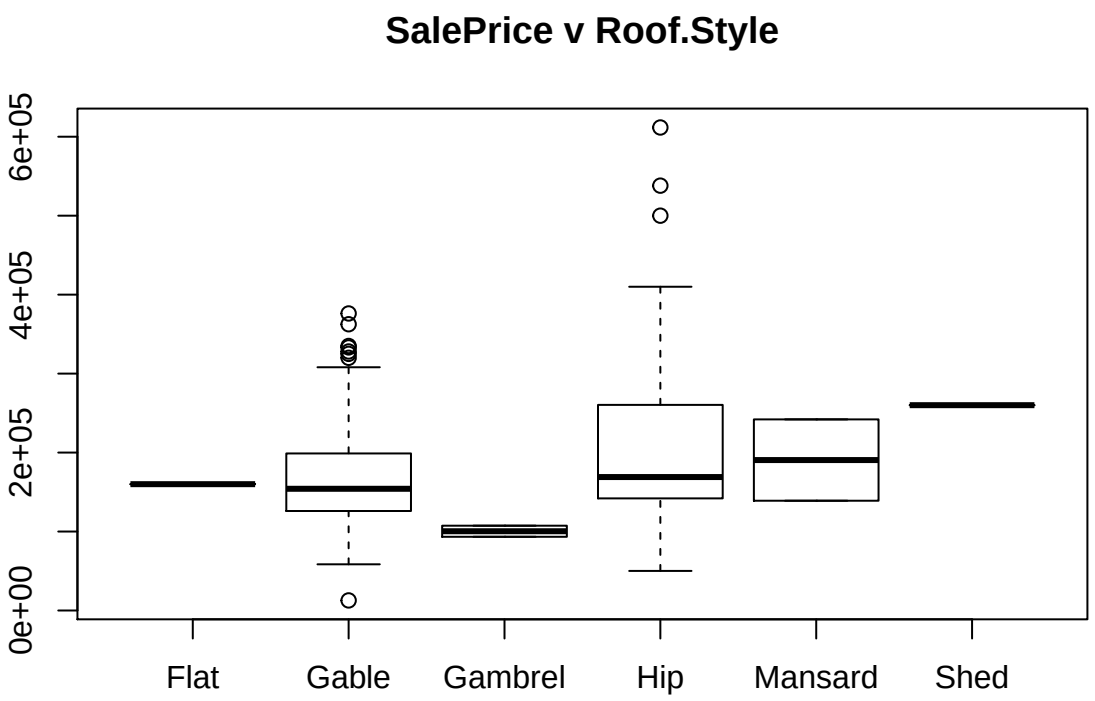
```
length(housing_data$Roof.Style)
```

```
## [1] 341
```

```
table(housing_data$Roof.Style)
```

```
##
##    Flat   Gable Gambrel     Hip Mansard    Shed
##       1     278       2      57       2       1
```

```
plot(housing_data$Roof.Style, housing_data$SalePrice, main= "SalePrice v Roof.Style")
```



Note vast majority are Gable roofs, probably have issues with low number of shed, Mansard, Gambrel, and Flat roofs.

```
xx1<- mean(housing_data$SalePrice[housing_data$Roof.Style=="Flat"])
yy1<-sd(housing_data$SalePrice[housing_data$Roof.Style=="Flat"])

xx2<-mean(housing_data$SalePrice[housing_data$Roof.Style=="Gable"])
yy2<-sd(housing_data$SalePrice[housing_data$Roof.Style=="Gable"])

xx3<-mean(housing_data$SalePrice[housing_data$Roof.Style=="Gambrel"])
yy3<-sd(housing_data$SalePrice[housing_data$Roof.Style=="Gambrel"])

xx4<-mean(housing_data$SalePrice[housing_data$Roof.Style=="Hip"])
yy4<-sd(housing_data$SalePrice[housing_data$Roof.Style=="Hip"])

xx5<-mean(housing_data$SalePrice[housing_data$Roof.Style=="Mansard"])
```

```
yy5<-sd(housing_data$SalePrice[housing_data$Roof.Style=="Mansard"])

xx6<-mean(housing_data$SalePrice[housing_data$Roof.Style=="Shed"])
yy6<-sd(housing_data$SalePrice[housing_data$Roof.Style=="Shed"])

xxMeans <-c(xx1, xx2, xx3, xx4, xx5, xx6)
yyStdDev <- c(yy1, yy2, yy3, yy4, yy5, yy6)
housetypes <- c("Flat", "Gable", "Gambrel", "Hip", "Mansard", "Shed")
cbind(housetypes, xxMeans,yyStdDev)
```

```
##      housetypes xxMeans            yyStdDev
## [1,] "Flat"     "160000"           NA
## [2,] "Gable"    "164744.737410072" "58980.2033058499"
## [3,] "Gambrel"  "100384.5"         "9921.41524682845"
## [4,] "Hip"      "211490.789473684" "119711.077663762"
## [5,] "Mansard"  "190500"           "72831.9984622144"
## [6,] "Shed"     "260000"           NA
```

Mean is about the same, some high values with Hip roof, but probably nothing statistically significant due to low number of observations for most values.

Variable: Roof.Matl

Meaning: Material of roofing material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

```
typeof(housing_data$Roof.Matl)
```

```
## [1] "integer"
```

```
length(housing_data$Roof.Matl)
```

```
## [1] 341
```

```
table(housing_data$Roof.Matl)
```

```
##
## CompShg Tar&Grv WdShake
##     337       2       2
```

They are almost all CompShg, no need to go further since no variation

---

Variable: Paved.Drive

Meaning: Type of driveway leading to home.

Y Paved

P Partial Pavement

N Dirt/Gravel

```
typeof(housing_data$Paved.Drive)
```

```
## [1] "integer"
```

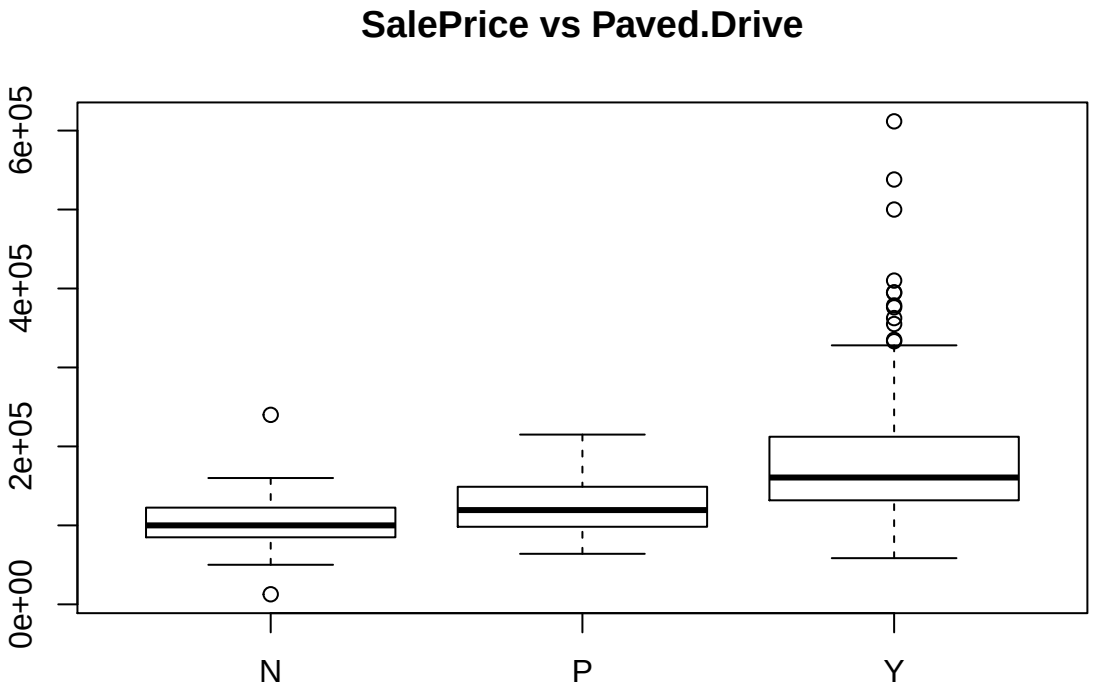
```
length(housing_data$Paved.Drive)
```

```
## [1] 341
```

```
table(housing_data$Paved.Drive)
```

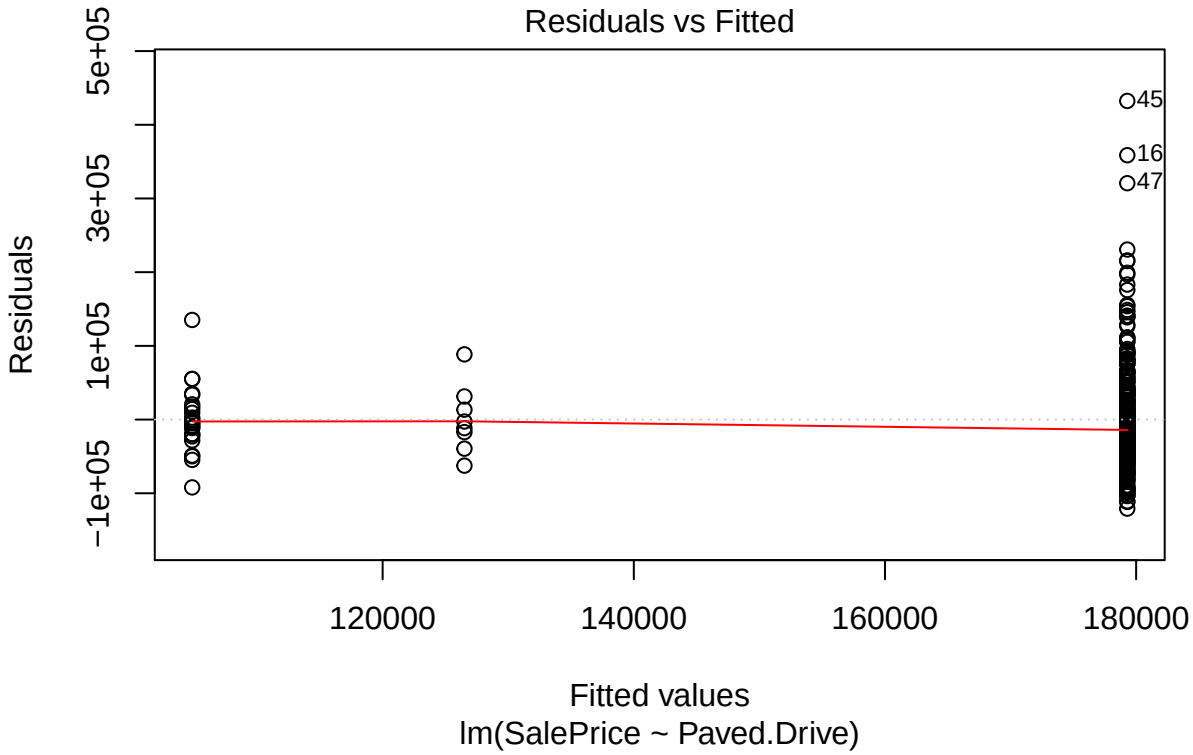
```
##
##   N   P   Y
##  25   8 308
```

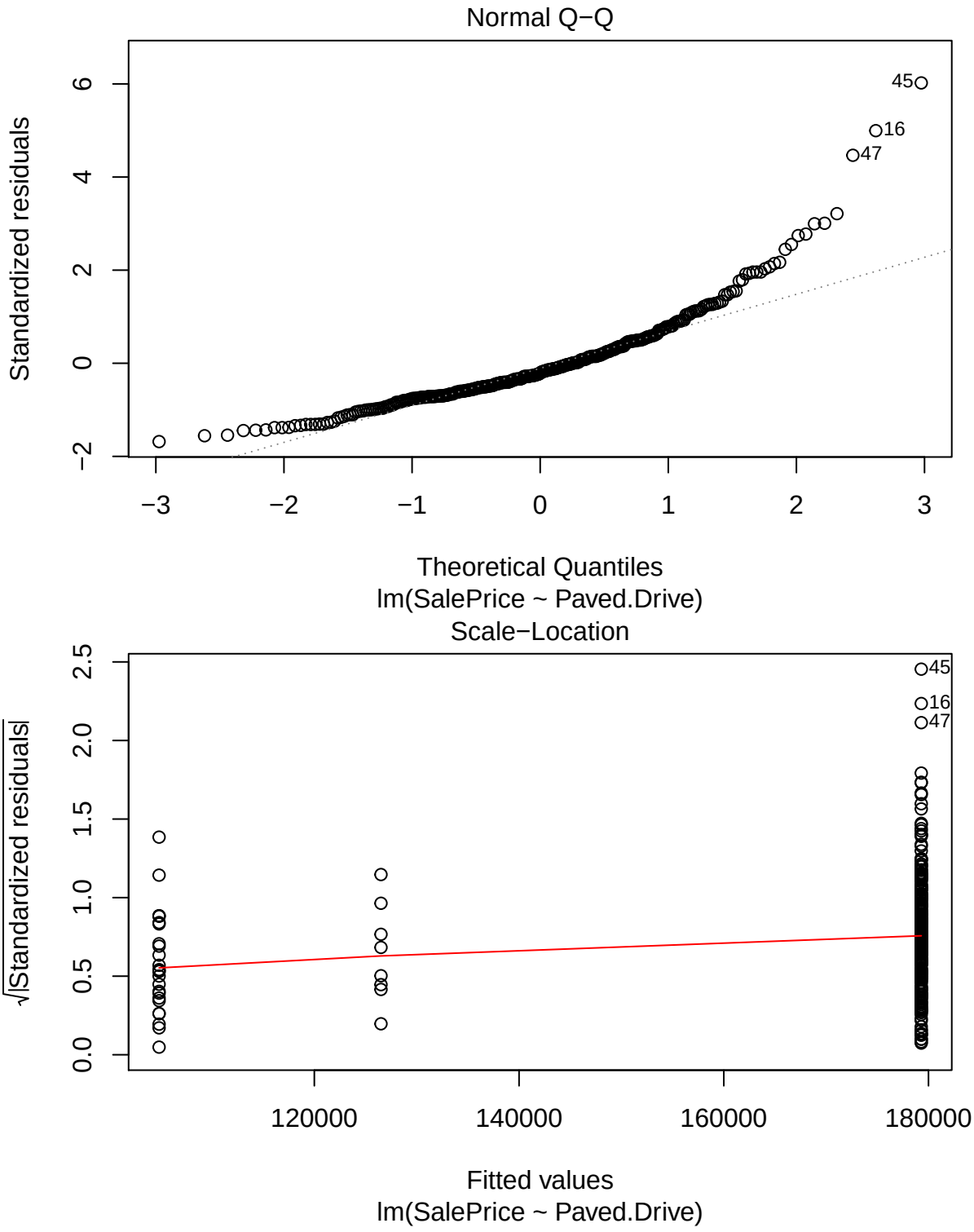
```
plot(housing_data$Paved.Drive, housing_data$SalePrice, main = "SalePrice vs Paved.Drive")
```



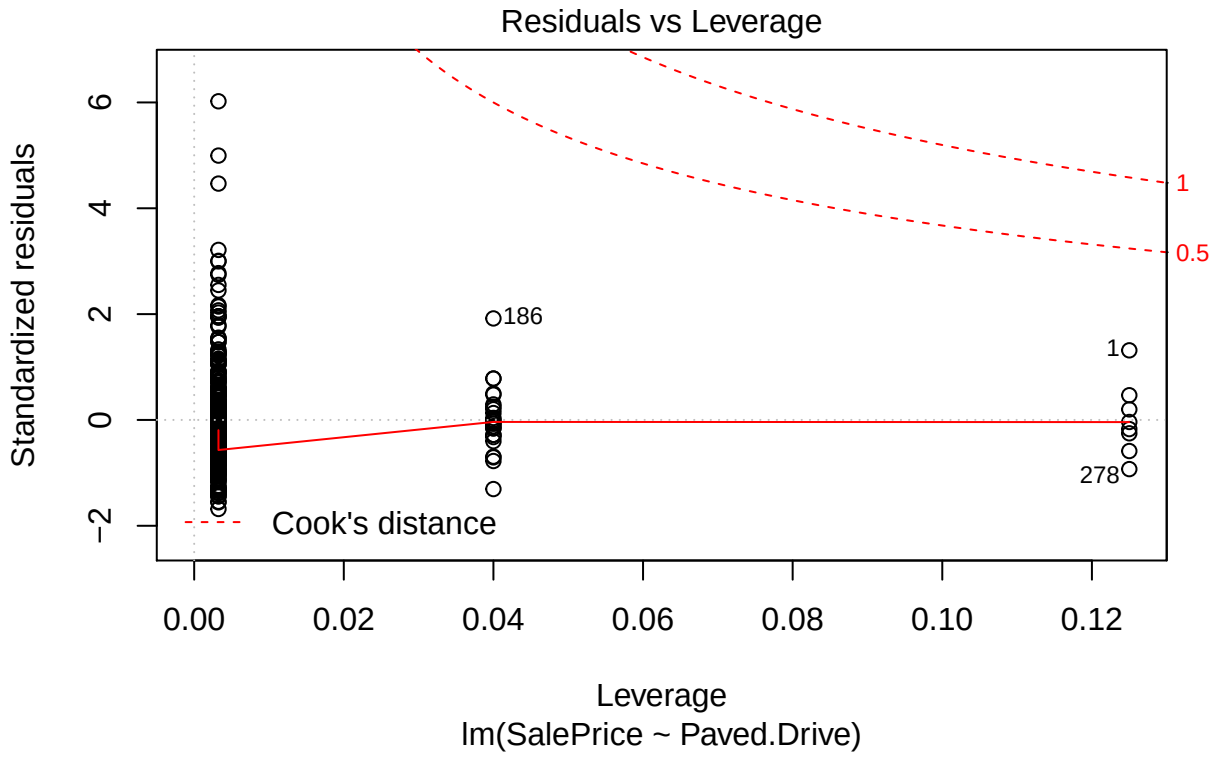
Not too surprising, the mean increases with a paved drive.

```
ModelPavedDrive<- lm(SalePrice~ Paved.Drive, data= housing_data)
plot(ModelPavedDrive)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SalePrice ~ Paved.Drive)
45
16
47
−3
−2
−1
0
1
2
3
−2
0
2
4
6

Scale−Location
√|Standardized residuals|
Fitted values
lm(SalePrice ~ Paved.Drive)
45
16
47
120000
140000
160000
180000
0.0
0.5
1.0
1.5
2.0
2.5



Not really sure how to interpret these, suggestions?

Variable: Pool QC

Meaning: Quality of pool

```
typeof(housing_data$Pool.QC)
```

```
## [1] "logical"
```

```
length(housing_data$Pool.QC)
```

```
## [1] 341
```

```
table(housing_data$Pool.QC)
```

```
## < table of extent 0 >
```

```
sum(is.na(housing_data$Pool.QC))
```

```
## [1] 341
```

no data

Variable: Fence

Meaning: Type and quality of fencing

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

```
typeof(housing_data$Fence)
```

```
## [1] "integer"
```

```
length(housing_data$Fence)
```

```
## [1] 341
```

```
table(housing_data$Fence)
```

```
##
## GdPrv  GdWo MnPrv
##    16    11    50
```

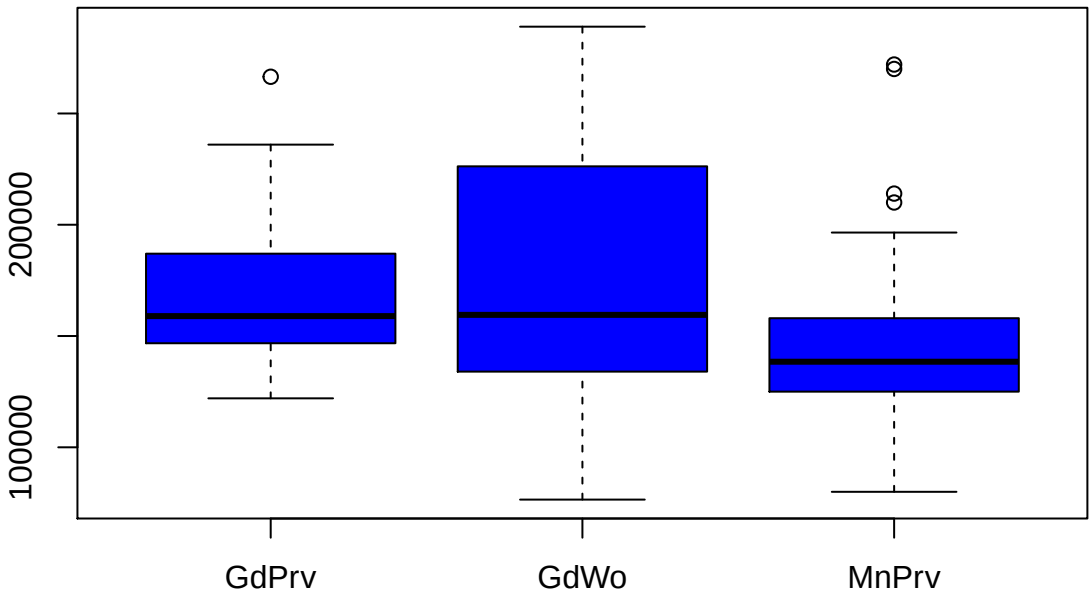
```
levels(housing_data$Fence)
```

```
## [1] "GdPrv" "GdWo"  "MnPrv"
```

```
sum(is.na(housing_data$Fence))
```

```
## [1] 264
```

```
plot(housing_data$Fence, housing_data$SalePrice, "SalePrice vs Fence Type", col="Blue")
```



Not enough data to make this meaningful.