Stats101A Final Project

Meiyi Ye

505569664

2023-03-24

Contents

Introduction

For ages, buying a home has been considered a great monumental goal for countless individuals. In our current society, a house can also be a means to financial wealth/stability. Purchasing your own house can be a sort of investment as it usually appreciates with time. Past trends indicates a stable positive growth in housing prices that can correlate to building generational wealth, an aspect of financial well being that is often overlooked. Knowing the importance of owning your own home can provide little comfort in terms of first home buyers and sellers. There are many aspects to consider when buying and selling a home. Such as the price, the location, the amount of rooms available etc. One question that can arise from this is whether a home's features can affect its price. In this project, I will be diving into the the relationship between housing price and its features.

The data I use for this project is a housing data set from kaggle.com. https://www.kaggle.com/datasets/ananthreddy/housing The housing data set has 12 variables, price, lotsize, bedrooms, bathrms, stories, driveway, recroom, fullbase, gashw, airco, garagepl, prefarea. There are a total of 546 observations.

- price: Price of a house.
- lotsize: The square feet lot size 0f a property.
- bedrooms: The number of bedrooms.
- bathrms : The number of bathrooms.
- stories: The number of stories not including basement.
- driveway: Is there a driveway?
- recroom: Is there a recreational room?
- fullbase: Is there a basement?
- gashw: Does the house use gas for water heating?
- airco: is there air conditioning?
- garagepl: The number of garage places.
- prefarea: Is the house located in a preferred area?

```
##   price lotsize bedrooms bathrms stories driveway recroom fullbase gashw airco
## 1 42000    5850        3       1       2        1       0        1     0     0
## 2 38500    4000        2       1       1        1       0        0     0     0
## 3 49500    3060        3       1       1        1       0        0     0     0
## 4 60500    6650        3       1       2        1       1        0     0     0
##   garagepl prefarea
## 1        1        0
## 2        0        0
## 3        0        0
## 4        0        0
```

I chose to use a multiple linear model for this project. Why did I choose a multiple linear model? Because it is a powerful statistical technique for exploring the relationship between one response variable and multiple explanatory variables that are supposedly related to the response variable. In this case, it was exactly what

I aimed to explore in my research, the relationship between the price of housing and the different attributes of housing, such as lot size, number of bathrooms, etc. The multiple linear model allows us to see how strong or weak a explanatory variable associates to the response variable. This method is also impressive in the way that it helps us make predictions about the dependent variable based on the values of the independent variables, for example predicting price.

This is the end of the introduction. Following this is the data description where the housing summary statistics, such as mean, standard deviation, correlation, and etc are reported. The distribution of each variable and relationships among the variables will also be presented using graphs. The third section is where the results of the multiple linear housing model will be interpreted. I will also discuss the model candidates that I compared during the search for the best fitting model. Furthermore, I will determine the best predictive model and assess the model using diagnostic tools. Lastly, I will summarize my project, discuss if my final model makes sense in the real world.

Data Description

The minimum, 1st quartile, median, mean, 3rd quartile, maximum of each variable in the housing data.

```
summary(housing)
```

```
##      price           lotsize         bedrooms        bathrms
##  Min.   : 25000   Min.   : 1650   Min.   :1.000   Min.   :1.000
##  1st Qu.: 49125   1st Qu.: 3600   1st Qu.:2.000   1st Qu.:1.000
##  Median : 62000   Median : 4600   Median :3.000   Median :1.000
##  Mean   : 68122   Mean   : 5150   Mean   :2.965   Mean   :1.286
##  3rd Qu.: 82000   3rd Qu.: 6360   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :190000   Max.   :16200   Max.   :6.000   Max.   :4.000
##     stories         driveway        recroom         fullbase
##  Min.   :1.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :2.000   Median :1.000   Median :0.0000   Median :0.0000
##  Mean   :1.808   Mean   :0.859   Mean   :0.1777   Mean   :0.3498
##  3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :4.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##      gashw             airco           garagepl         prefarea
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.04579   Mean   :0.3168   Mean   :0.6923   Mean   :0.2344
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :3.0000   Max.   :1.0000
```

The standard deviation of each variable in the housing data.

```
apply(housing, 2, sd)
```

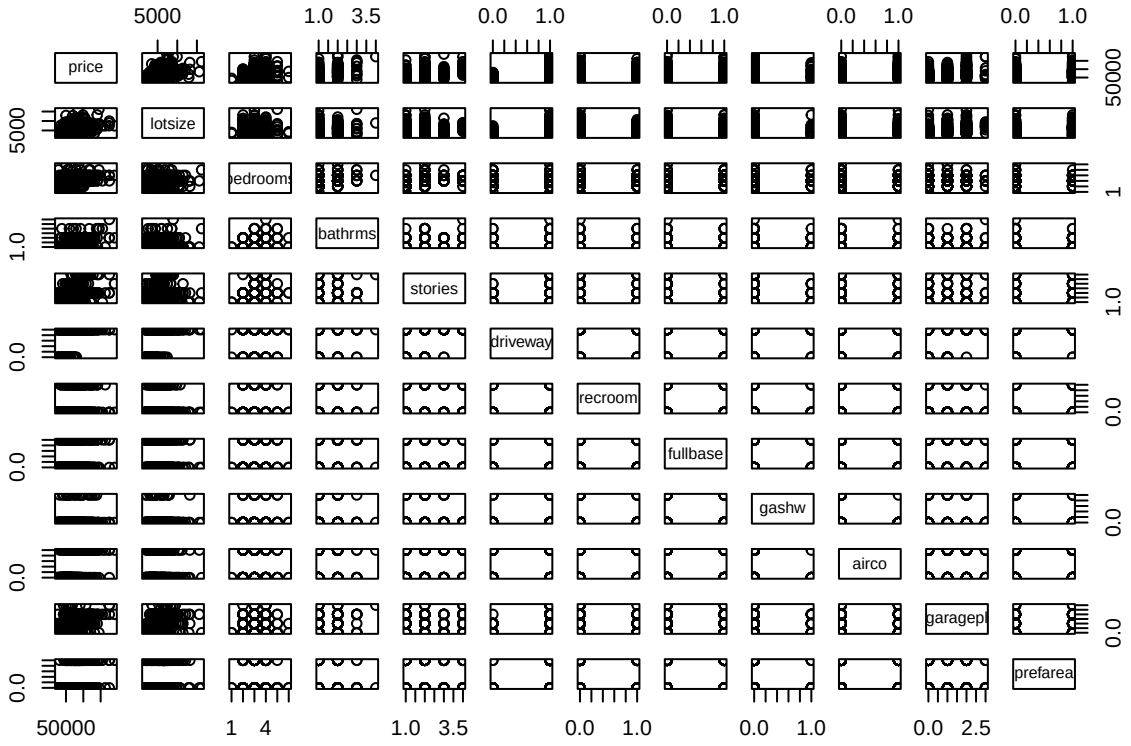
```
##        price      lotsize     bedrooms      bathrms      stories     driveway
## 2.670267e+04 2.168159e+03 7.373879e-01 5.021579e-01 8.682025e-01 3.483672e-01
##      recroom     fullbase        gashw        airco     garagepl     prefarea
## 3.825731e-01 4.773493e-01 2.092157e-01 4.656750e-01 8.613066e-01 4.240319e-01
```

This is the correlation of the response variable (price) and each explanatory variable in the housing data.

```
## Correlation between price and lotsize:  0.5357957

## Correlation between price and bedrooms:  0.3664474

## Correlation between price and bathrms:  0.5167193

## Correlation between price and stories:  0.4211902

## Correlation between price and driveway:  0.2971668

## Correlation between price and recroom:  0.2549595

## Correlation between price and fullbase:  0.1862177

## Correlation between price and gashw:  0.09283654

## Correlation between price and airco:  0.4533466

## Correlation between price and garagepl:  0.383302

## Correlation between price and prefarea:  0.3290743
```

The distribution of each variable are presented by histograms (See Figure. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) and the relationship among variables are presented by scatter plots.

```
plot(housing)
```



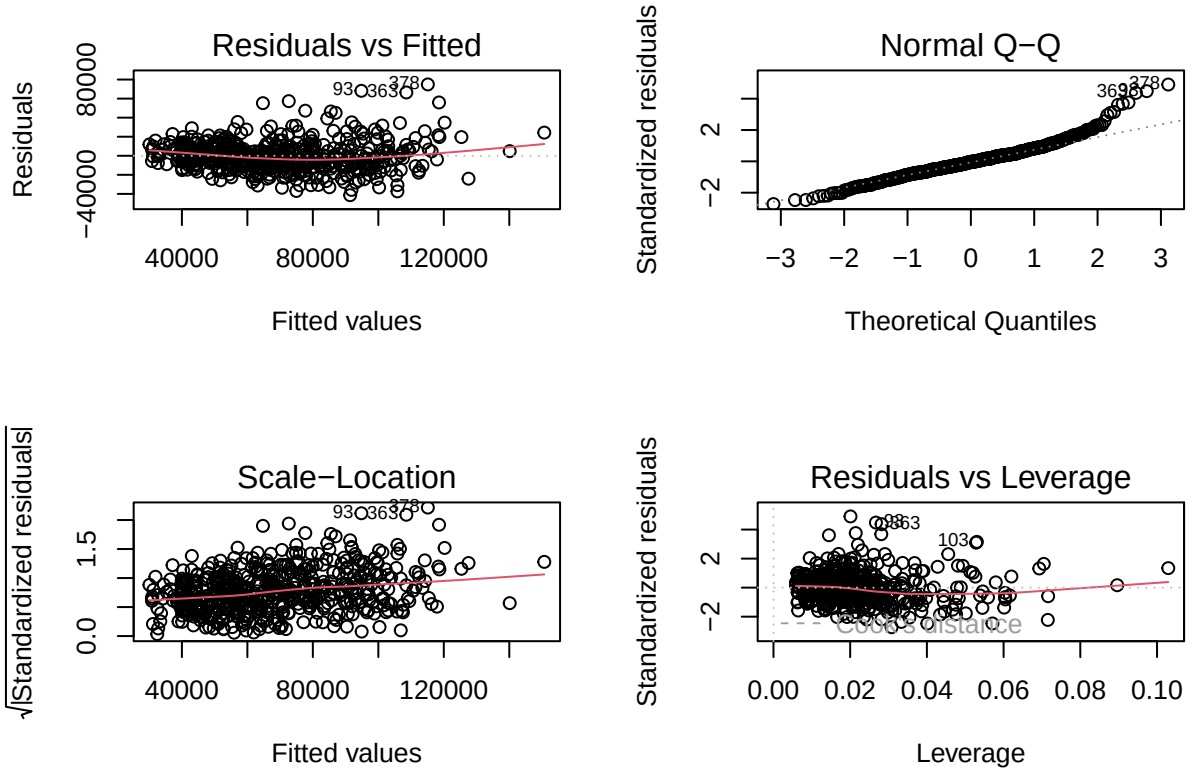
Results and Interpretation

Here, I developed a regression model with the response variable price and all of the 11 potential predictor variables.

```
model1 <- lm(price~lotsize+bedrooms+bathrms+stories+driveway+
               recroom+fullbase+gashw+airco+garagepl+prefarea, data = housing)
```

Next, I produced four diagnostic plots. In the Residual vs. Fitted plot, the constant variance of the error term (the red line) does not appear completely linear, but we want the red line as linear as possible. In the Normal Q-Q plot, the points are mostly aligned to the straight line, which implies the normality of the errors, but the plot could be improved. The plot of Residual vs. Leverage is to check the outliers or (possible) influential points, I will tackle outliers later.

```
par(mfrow=c(2,2))
plot(model1)
```

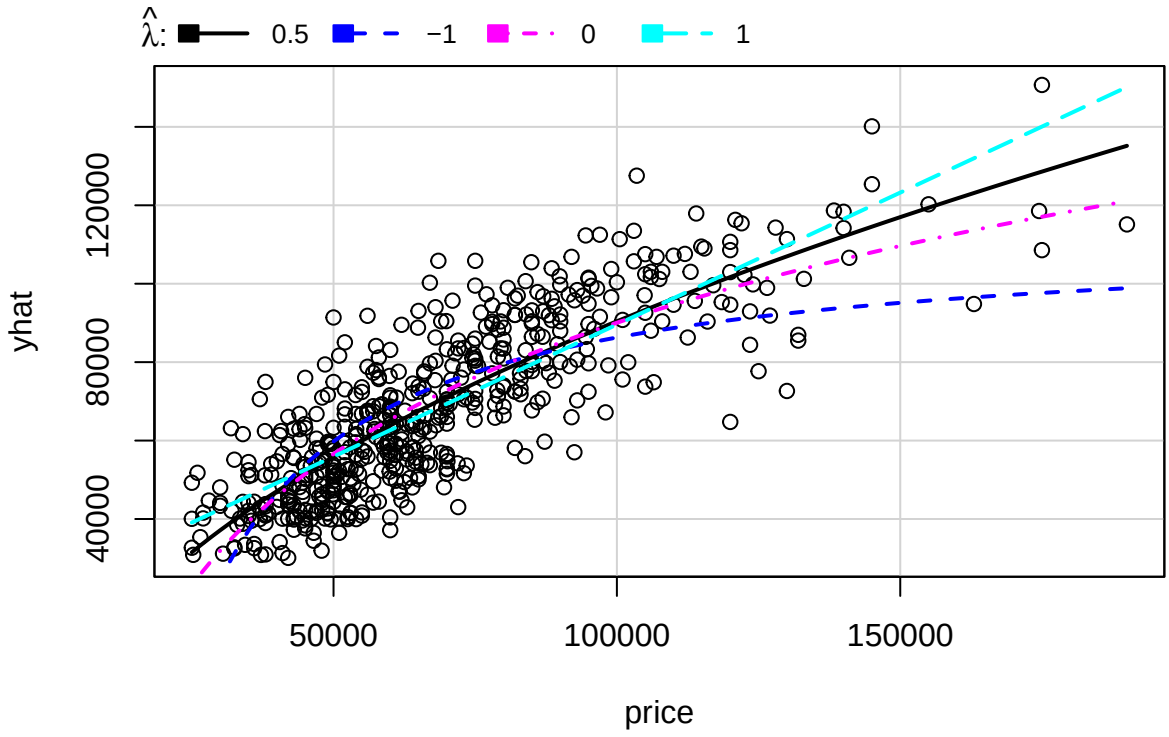


From above we observed our model diagnostic plots have some problems. To improve the non-constant variance and non-normality issues, I chose to use a power transformation method called the inverse response plot which transforms the response variable only.

```
library(car)
```

```
## Loading required package: carData
```

```
#inverse response plot
inverseResponsePlot(model1, key=TRUE)
```



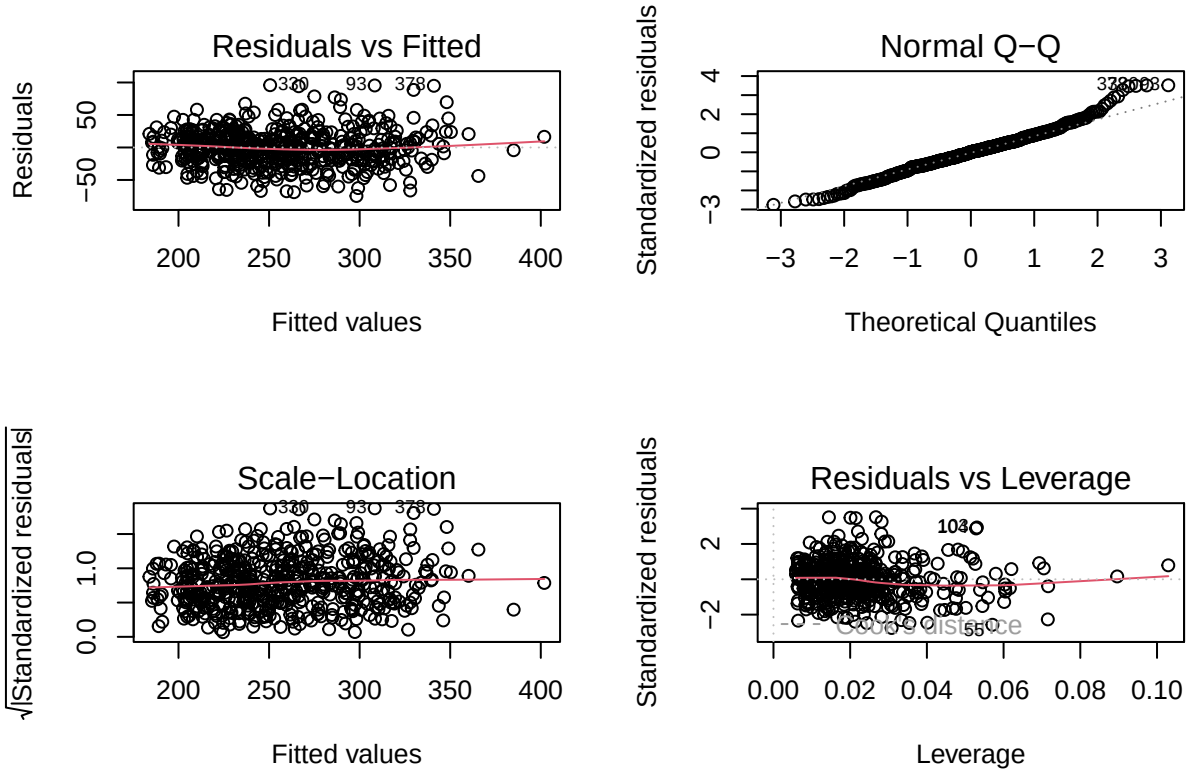

```
##       lambda          RSS
## 1  0.5040991  82242118968
## 2 -1.0000000 110181593030
## 3  0.0000000  85651130662
## 4  1.0000000  85503576138
```

After running the inverse response plot, λ equaled to 0.5, thus we transform the response variable by taking the square root. I developed a new regression model with the square root response variable price and all of the 11 potential predictor variables.

```
model2 <- lm(sqrt(price) ~ lotsize + bedrooms + bathrms + stories +
    driveway + recroom + fullbase + gashw + airco + garagepl +
    prefarea, data = housing)
```

We should check the model diagnostic plots of models every time we do a transformation or removing outliers or leverage points to see if the plots improve. So, we generated the model diagnostic plots of the new power transformation model. The results are good. In the Residual vs. Fitted plot, the constant variance of the error term appeared to be linear, which implied constant variance, which is what we want. More points on Normal Q-Q plot aligned on the line than the previous model, this implies normality of errors. We can conclude that the power transform model is a better fitting model than our original model.

```
par(mfrow=c(2,2))
plot(model2)
```



I once again try to improve my model by removing the points greater than the leverage point. However, when I checked its model diagnostic plots, the plots (See Figure. 13) looked worst than the previous model. The Residual vs. Fitted plot looked less linear and the points in the Normal Q-Q plot was less aligned with the line. Since this model was no good, I decided to just stick to my power transformation model.

```
model3 <- lm(price ~ lotsize + bedrooms + bathrms + stories +
    driveway + recroom + fullbase + gashw + airco + garagepl + #leverage point
    prefarea, data = housing[hatvalues(model2) < 0.04395604,]) #2 * ((p + 1) / n) = 0.04395604
```

Now, I want to test whether the power transformation model terms are significant, thus running the ANOVA test is perfect for our task. From its output we see that all the variables are significant because they all have p-value < 0.05. So, we can try to test the variable significance another way, use the added-variable plots check what variables are significant.

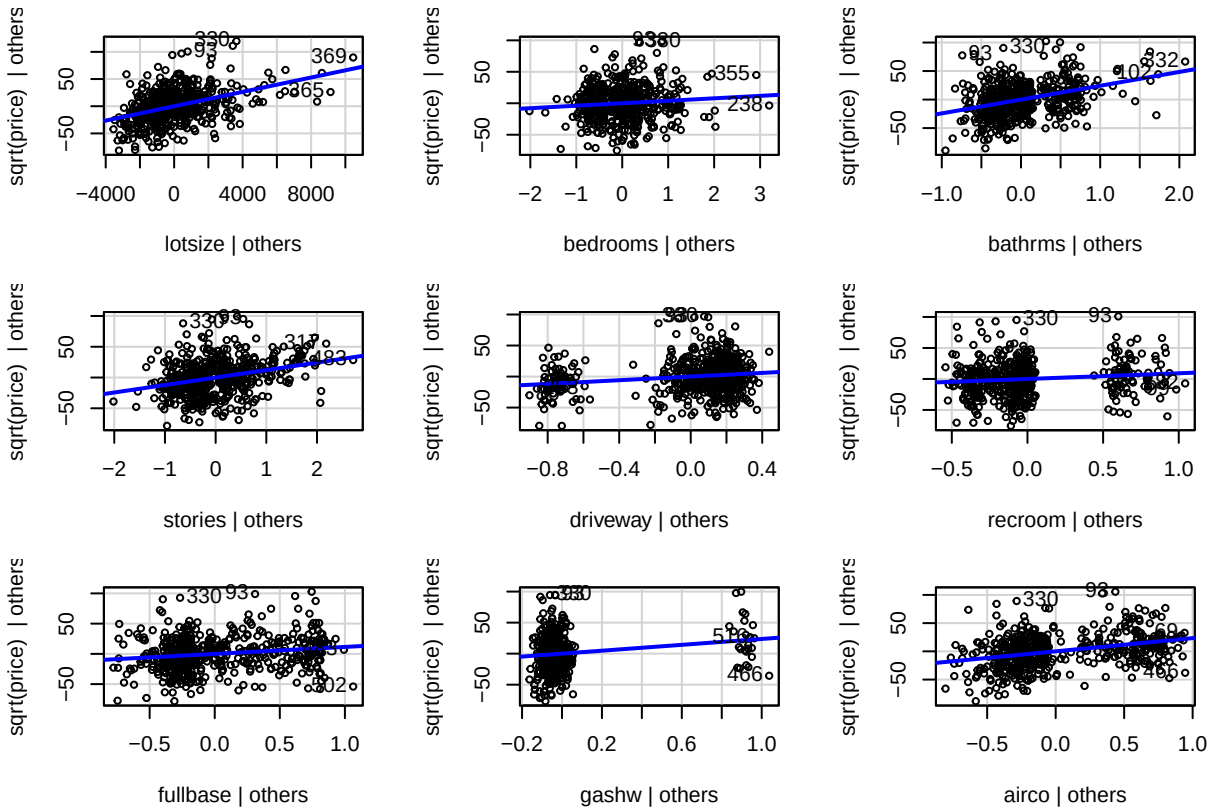
```
anova(model2)
```

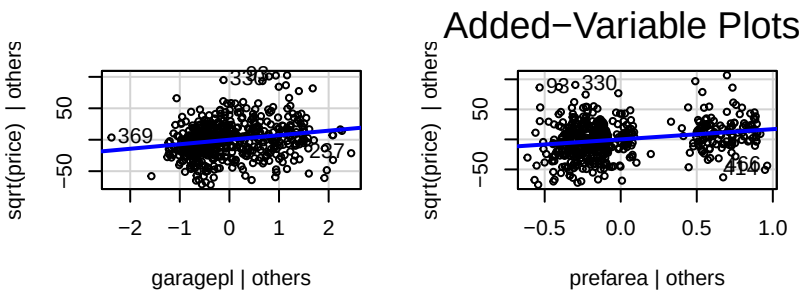
```
## Analysis of Variance Table
##
## Response: sqrt(price)
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## lotsize     1 382053  382053 504.9391 < 2.2e-16 ***
## bedrooms    1 110168  110168 145.6030 < 2.2e-16 ***
```

```
## bathrms     1 135920  135920 179.6390 < 2.2e-16 ***
## stories     1  65304   65304  86.3083 < 2.2e-16 ***
## driveway    1  29719   29719  39.2783 7.579e-10 ***
## recroom     1  26436   26436  34.9385 6.072e-09 ***
## fullbase    1  25200   25200  33.3053 1.336e-08 ***
## gashw       1   6390    6390   8.4452  0.003812 **
## airco       1  62557   62557  82.6779 < 2.2e-16 ***
## garagepl    1  16965   16965  22.4212 2.809e-06 ***
## prefarea    1  24663   24663  32.5962 1.884e-08 ***
## Residuals 534 404041     757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To check which variables contribute the most to the model we could look at the magnitude of the slope. If the slope of the variable is steep (not flat), it means it contributes and it is significant. If the slope is flat or almost flat it means it does not contributes and it is not significant. From the output of the added-variable plot of the power transformation model, I observed that variables fullbase, bedrooms, driveway, recroom, prefarea appeared to have a low magnitude.

```
avPlots(model2)
```





To test the theory that variables fullbase, bedrooms, driveway, recroom, prefarea are not significant variables, we created a reduced model which is the transformed model but with the fullbase variable excluded (See Figure. 14). Then, we use the ANOVA partial F-test to compare the the reduced model against the full model (transformed model), which resulted in the p-value < 0.05, thus we choose the full model as the better fitting model. This method is then repeated by removing variables bedrooms, driveway, recroom, prefarea from the reduced model one at a time, and compared against the full model (See Figure. 15, 16, 17, 18). They all resulted in a p-value < 0.05, which we concluded the full model was the best fitting model.

Even after we make a model valid, we can sometimes expect some improvements by selecting variables. Variable selection methods aim to choose the subset of the predictors that is the best set. In our case, we chose to do backward elimination AIC and BIC, and forward selection AIC and BIC. Backward elimination deletes a variable with the biggest p-value each round. Forward selection adds a variable with the smallest p-value each round. When we ran our backward AIC, it showed that the full transformation model was the best model. When we ran our backward BIC, it showed that the full transformation model excluding bedrooms variable was the best model. Backward AIC and forward AIC arrives at the same model, this also applies to Backward BIC and forward BIC (See Figure. 19, 20, 21, 22). To find out which model is better, we ran a summary on backward AIC and BIC then compared their adjusted R^2, 0.6802 vs. 0.6782 (See Figure. 23). The better model have the higher adjusted R^2 value. In this case, the backward AIC had the high adjusted R^2 value, thus we come to the conclusion that the full model is the best fitting model for our housing data.

```
vif(model2)
```

```
##  lotsize bedrooms  bathrms  stories driveway  recroom fullbase    gashw
## 1.321632 1.365633 1.282494 1.478584 1.163091 1.210501 1.316543 1.038246
##    airco garagepl prefarea
## 1.201397 1.200839 1.147639
```

I also checked that all variance inflation factors are less than 5, there doesn't appear to be an issue with multicollinearity.

The final model (See Figure. 24): Predicted price = 1.219e+02 + 6.602e-03(lotsize) + 3.919e+00(bedrooms) + 2.420e+01(bathrms) + 1.216e+01(stories) + 1.474e+01(driveway) + 9.162e+00(recroom) + 1.150e+01(fullbase) + 2.367e+01(gashw) + 2.341e+01(airco) + 7.188e+00(garagepl) + 1.700e+01(prefarea)

Discussion

Based on the Residual vs Fitted plot, the constant variance of the error term did not appear linear, so I performed a power transformation on the full model. Afterward, I ran the Residual vs. Fitted plot again, and this time the constant variance of the error term improved and appeared much more linear. I also attempted to remove points that are greater than the leverage point in the data, however, the Residual vs Fitted plot constant variance of the error term looked less linear than the power transformation model, thus I stuck with the power transformation model. Following that, I performed the ANOVA test and the added-variables plots on the model so I can see which variables were significant and which were not. Next, I used ANOVA partial F-test to compare the different reduced models and the full model, which resulted in the full model being the best fitting model. I also did variable selection, which once again confirmed that the full model was the best fitting model. Power transformation with the full model was the best fitting model for the housing data, the model is also confirmed by ANOVA and variable selection. Lastly, I checked the variance inflation factors, and there is no issue with multicollinearity.

My final model makes sense in real world situations. The price of a house is very much dependent on the different features of the house, the lot size, number of bedrooms, number of bathrooms, and etc. So, it makes prefect sense that all the variables would be included in my model, because every aspect of a house affect the house's price. It is very much prevalent in today's society, a home that has a big lot size, or a home in a preferred and nice area, or a home that has a multi-garage, will be priced at a higher value than a home that has a small lot size, or a home in a less-preferred area, or a home that has no garage.

A potential limitation of this analysis is that the model might only be valid for housing prices in the location where the data set was collected.

Appendix

Figure 1

```
hist(housing$price,
     density = 30, col = "blue", angle = 45)
```

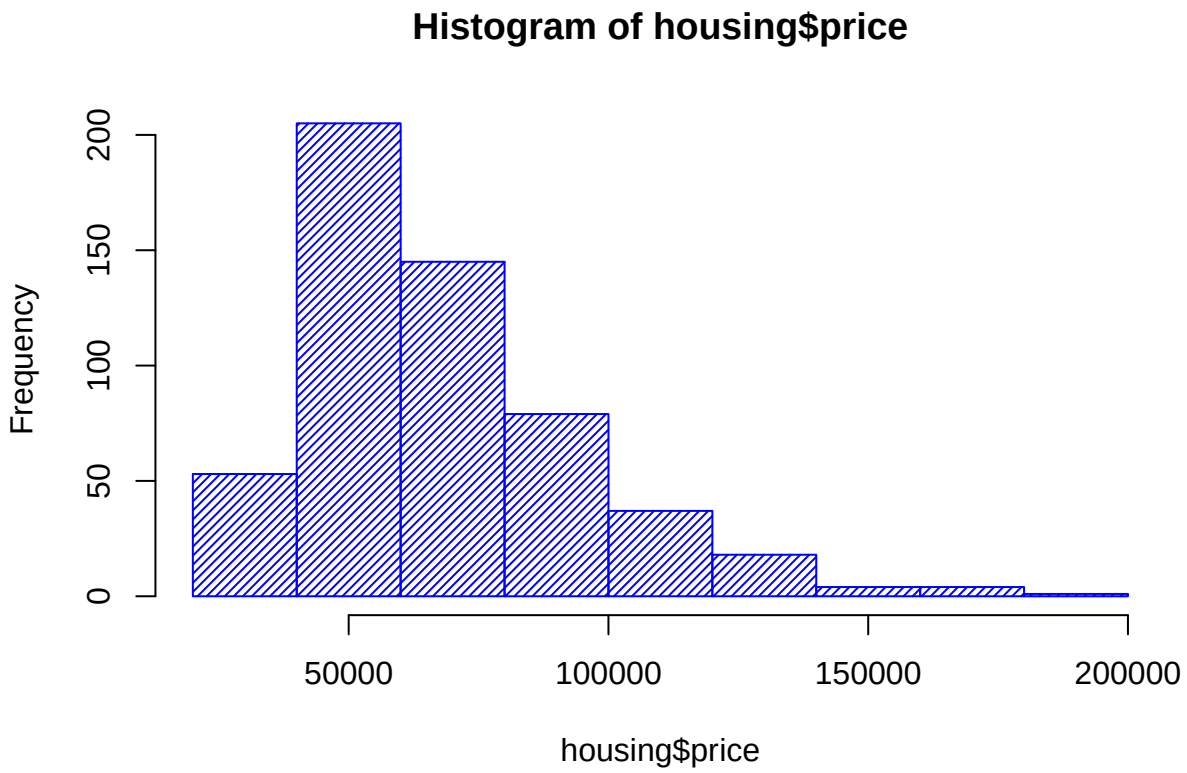


Figure 2

```
hist(housing$lotsize,
     density = 30, col = "blue", angle = 45)
```

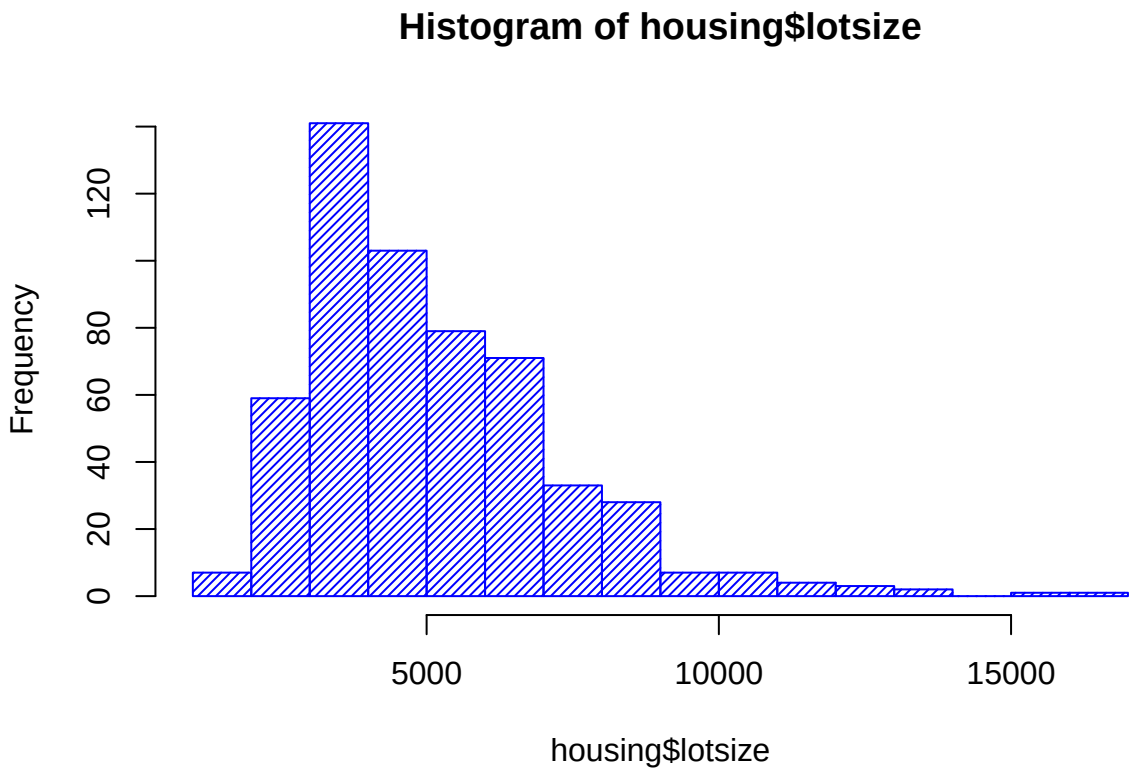


Figure 3

```
hist(housing$bedrooms,
     density = 30, col = "blue", angle = 45)
```

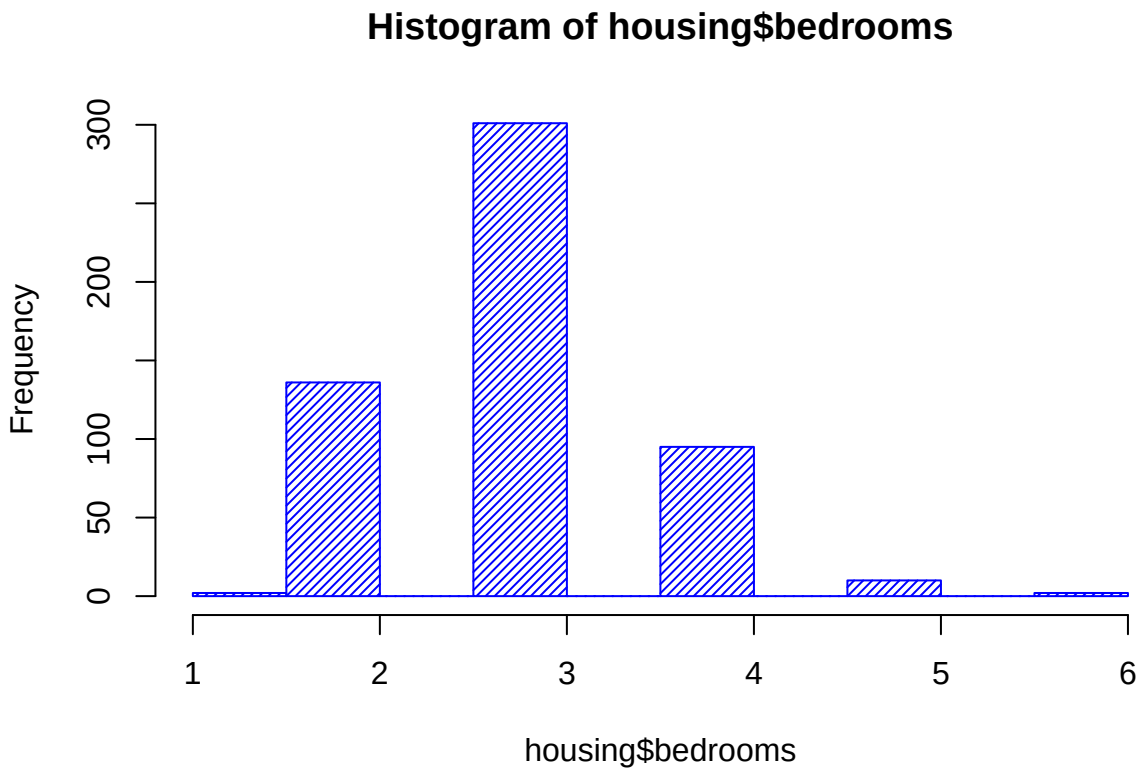


Figure 4

```
hist(housing$bathrms,
     density = 30, col = "blue", angle = 45)
```

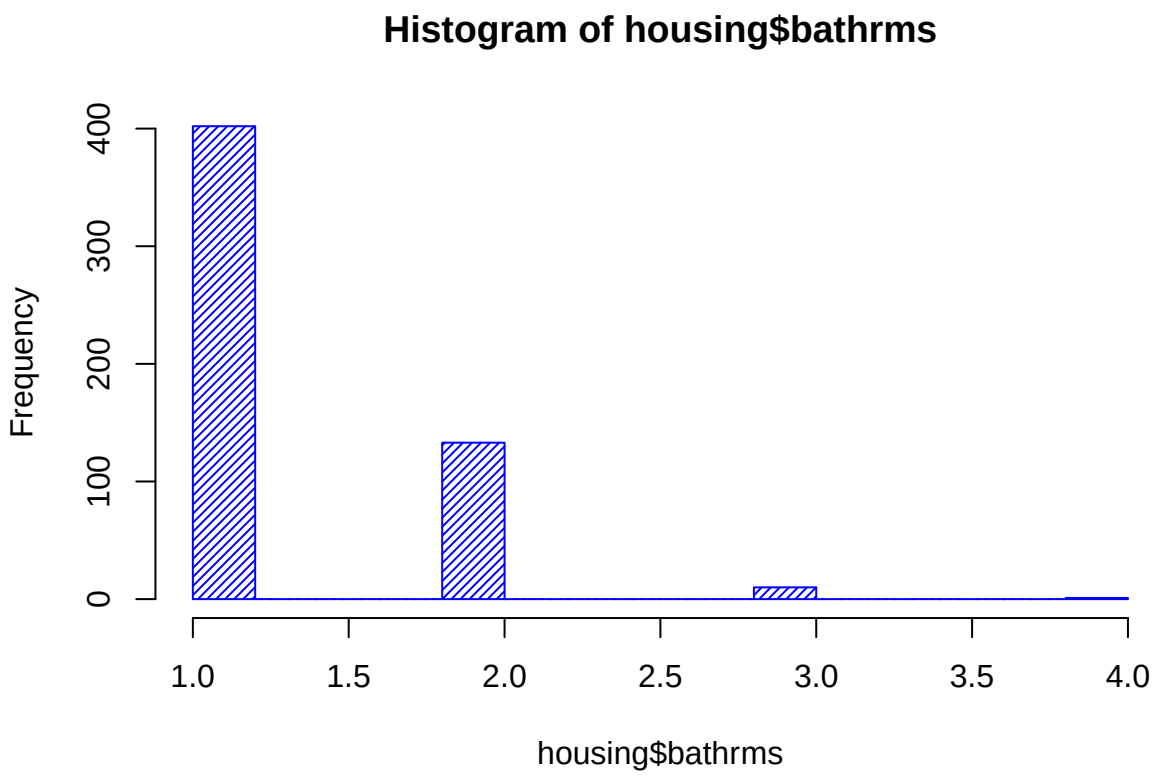


Figure 5

```
hist(housing$stories,
     density = 30, col = "blue", angle = 45)
```

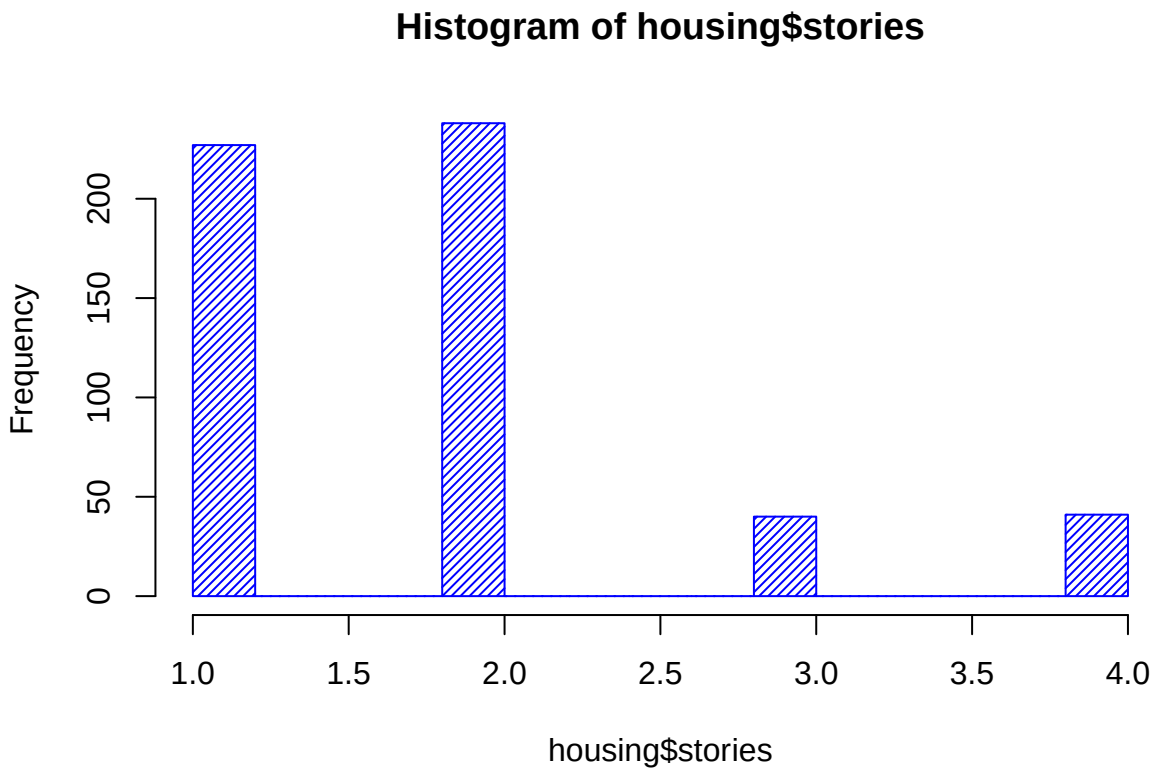


Figure 6

```
hist(housing$driveway,
     density = 30, col = "blue", angle = 45)
```

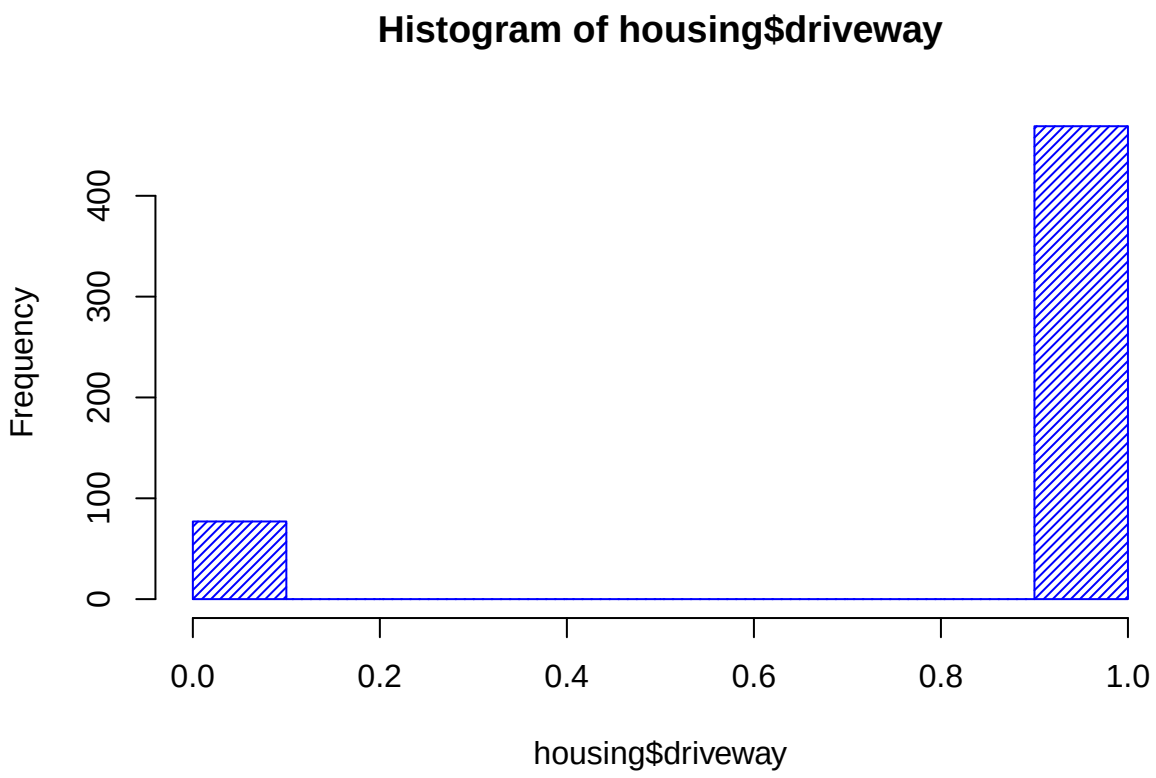


Figure 7

```
hist(housing$recroom,
     density = 30, col = "blue", angle = 45)
```

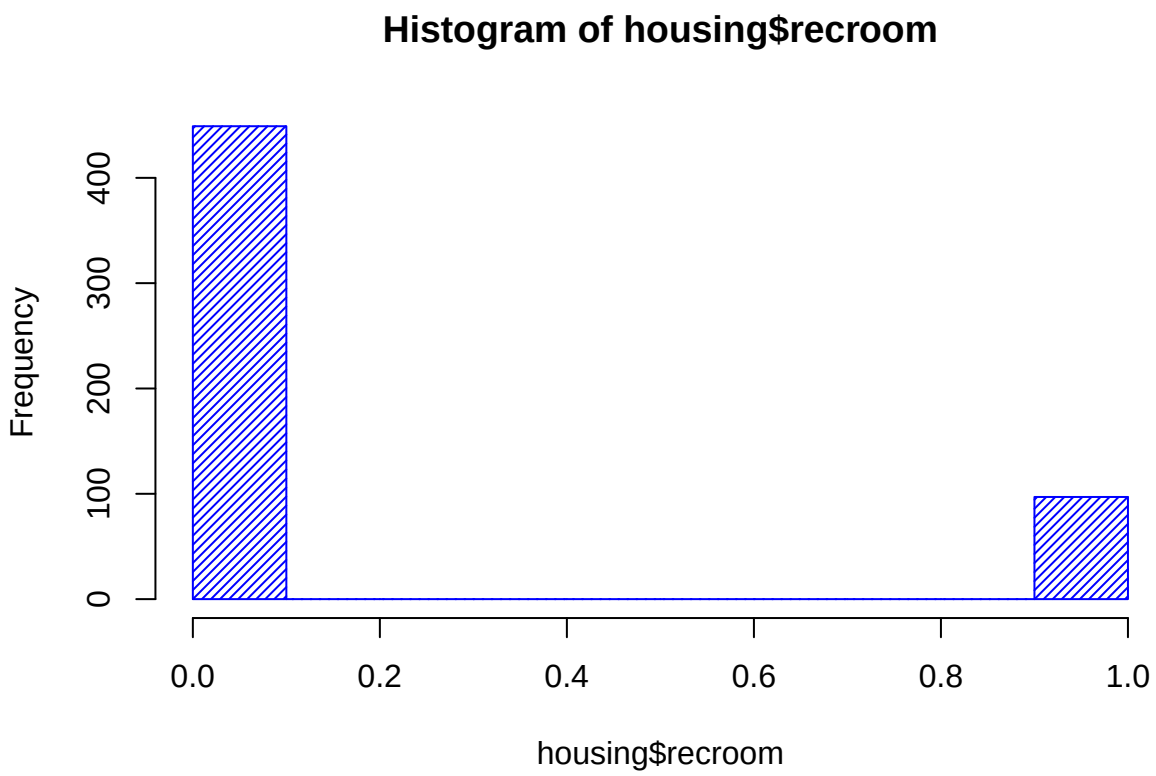


Figure 8

```
hist(housing$fullbase,
     density = 30, col = "blue", angle = 45)
```

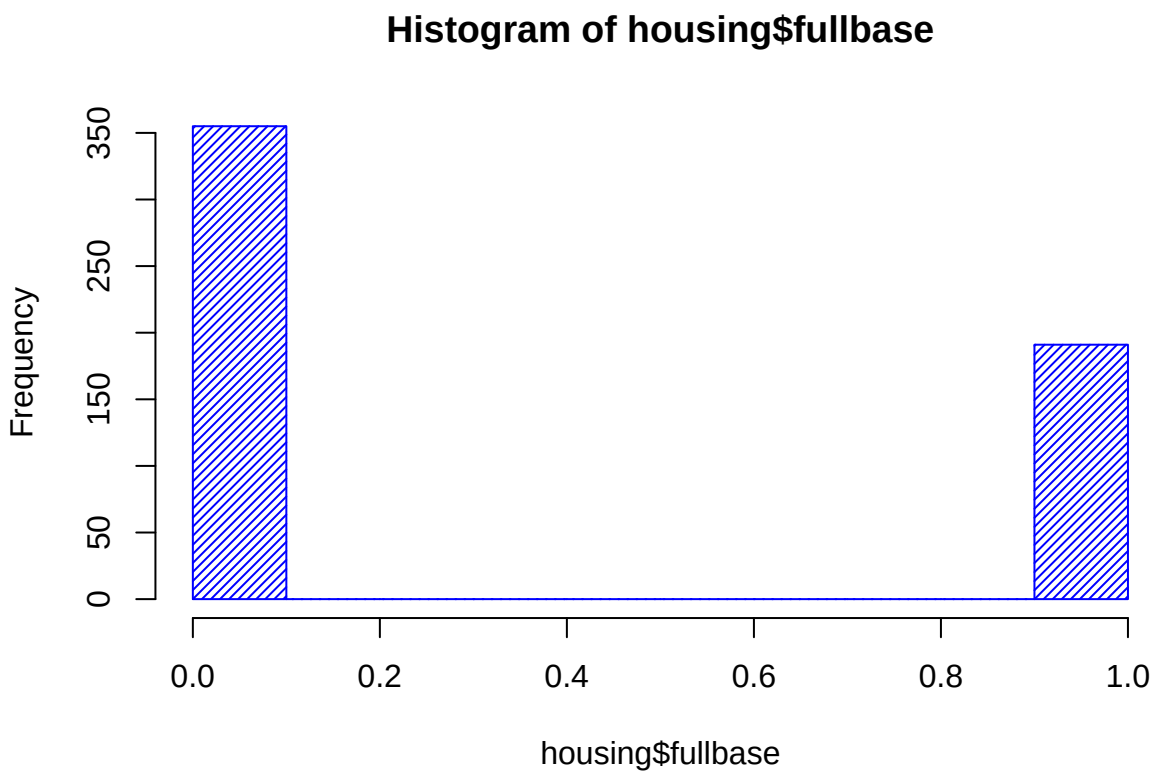


Figure 9

```
hist(housing$gashw,
     density = 30, col = "blue", angle = 45)
```

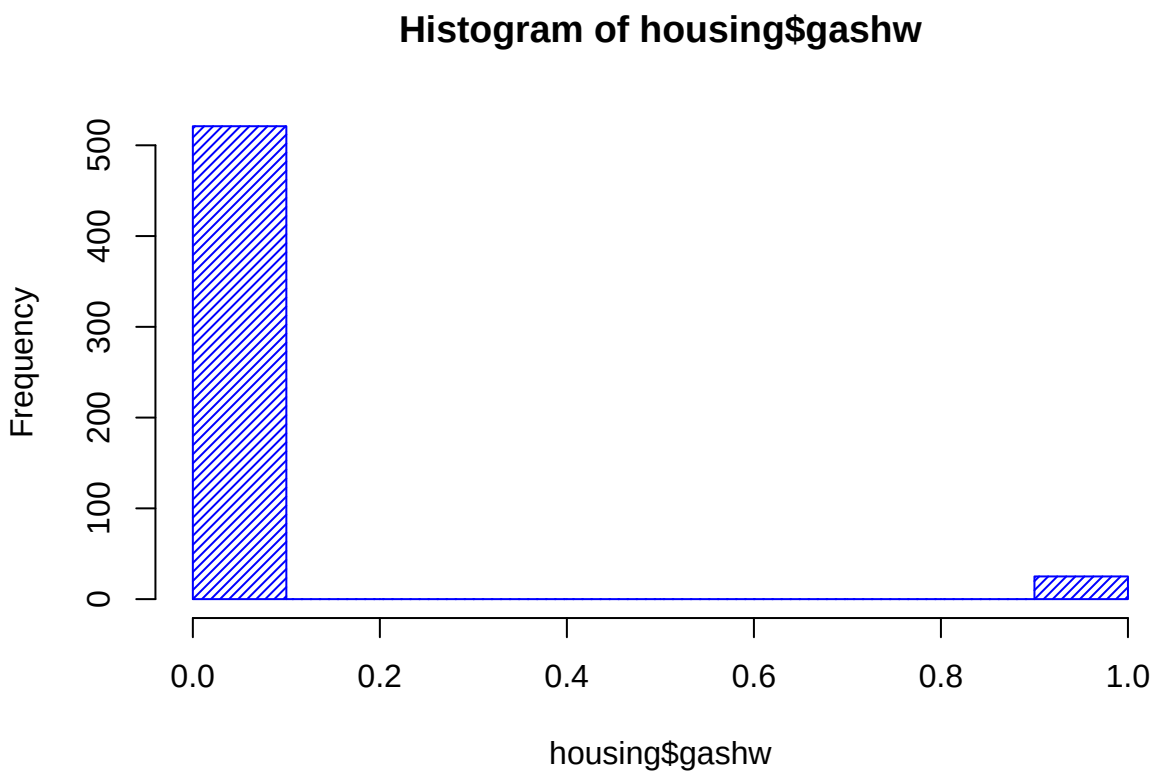


Figure 10

```
hist(housing$airco,
     density = 30, col = "blue", angle = 45)
```

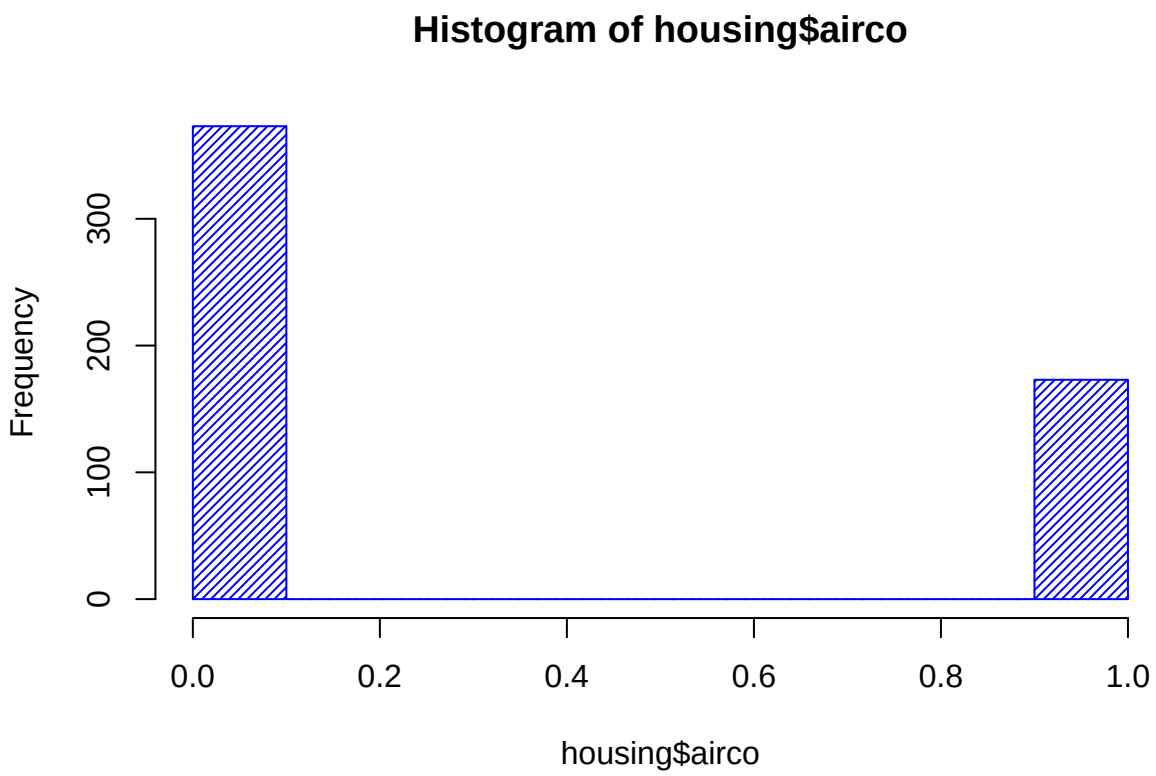


Figure 11

```
hist(housing$garagepl,
     density = 30, col = "blue", angle = 45)
```

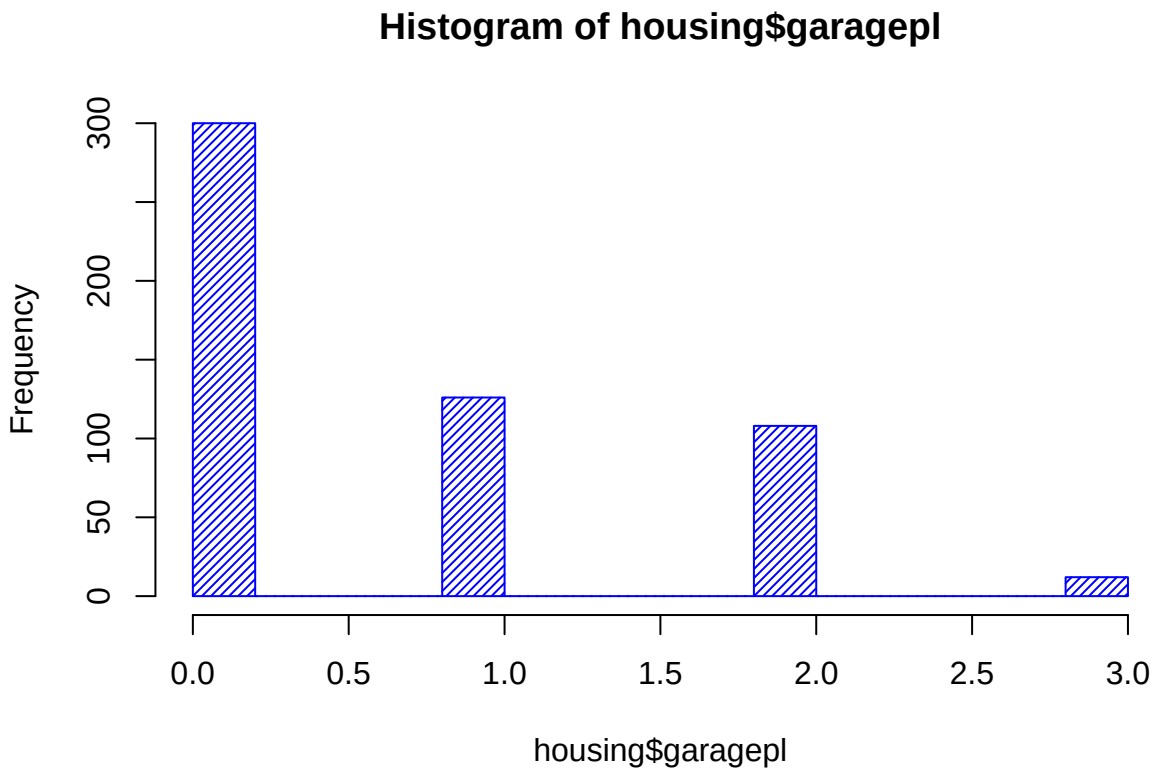


Figure 12

```
hist(housing$prefarea,
     density = 30, col = "blue", angle = 45)
```

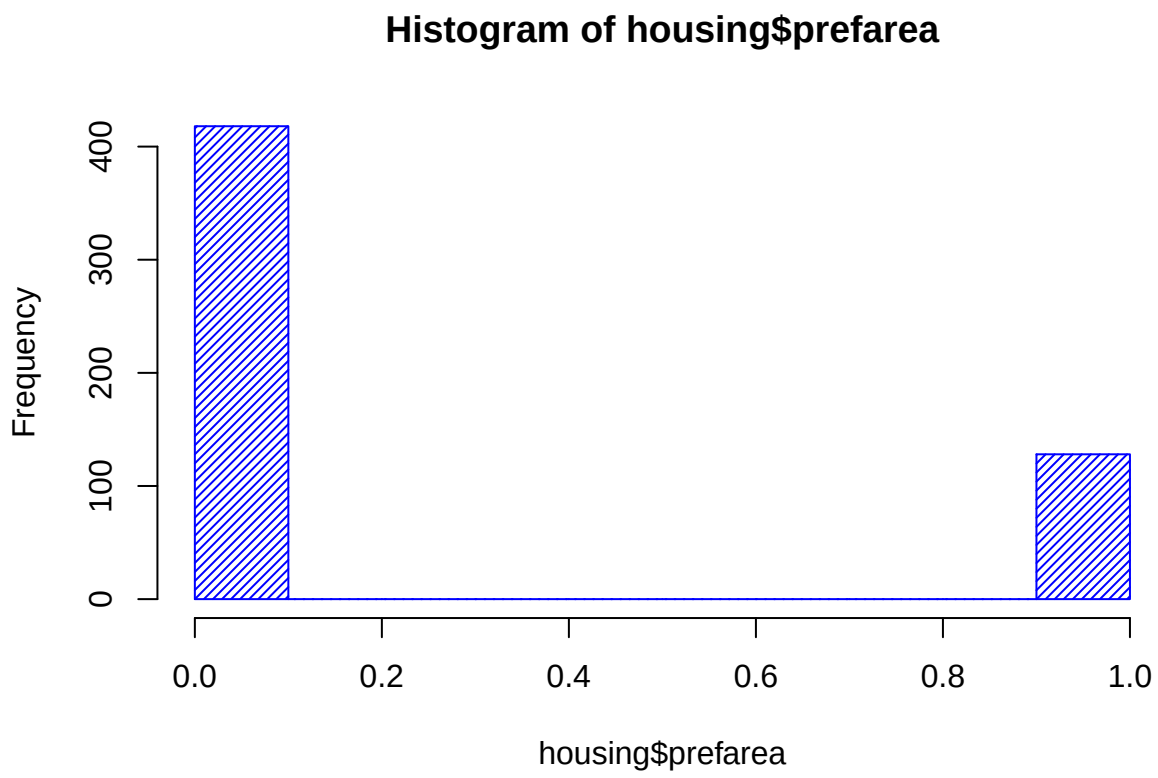


Figure 13

```
par(mfrow=c(2,2))
plot(model3)
```

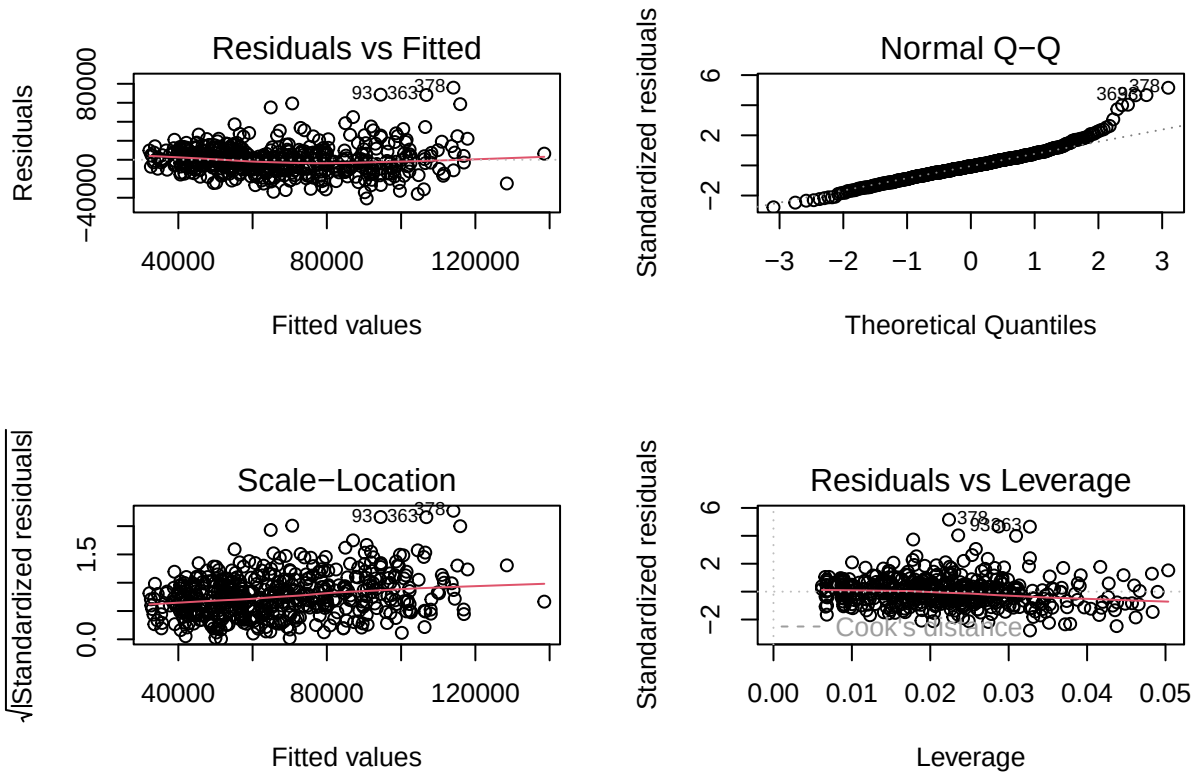


Figure 14

```
reduced_model1 <- lm(sqrt(price) ~ lotsize + bedrooms + bathrms + stories +
    driveway + recroom + gashw + airco + garagepl +
    prefarea, data = housing)
#remove fullbase
anova(reduced_model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(price) ~ lotsize + bedrooms + bathrms + stories + driveway +
##     recroom + gashw + airco + garagepl + prefarea
## Model 2: sqrt(price) ~ lotsize + bedrooms + bathrms + stories + driveway +
##     recroom + fullbase + gashw + airco + garagepl + prefarea
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    535 416526
## 2    534 404041  1     12485 16.501 5.591e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#P-value < 0.05 => choose full model
```

Figure 15

```
reduced_model2 <- lm(sqrt(price) ~ lotsize + bathrms + stories +
    driveway + recroom + gashw + airco + garagepl +
    prefarea, data = housing)
#remove bedrooms
anova(reduced_model2, model2)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(price) ~ lotsize + bathrms + stories + driveway + recroom +
##     gashw + airco + garagepl + prefarea
## Model 2: sqrt(price) ~ lotsize + bedrooms + bathrms + stories + driveway +
##     recroom + fullbase + gashw + airco + garagepl + prefarea
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    536 422051
## 2    534 404041  2     18010 11.901 8.774e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#P-value < 0.05 => choose full model
```

Figure 16

```
reduced_model3 <- lm(sqrt(price) ~ lotsize + bathrms + stories +
                       recroom + gashw + airco + garagepl +  prefarea, data = housing)
#remove driveway
anova(reduced_model3, model2)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(price) ~ lotsize + bathrms + stories + recroom + gashw +
##     airco + garagepl + prefarea
## Model 2: sqrt(price) ~ lotsize + bedrooms + bathrms + stories + driveway +
##     recroom + fullbase + gashw + airco + garagepl + prefarea
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    537 433370
## 2    534 404041  3     29329 12.921 3.686e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#P-value < 0.05 => choose full model
```

Figure 17

```
reduced_model4 <- lm(sqrt(price) ~ lotsize + bathrms +  stories +
                       gashw + airco + garagepl + prefarea, data = housing)
#remove recroom
anova(reduced_model4, model2)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(price) ~ lotsize + bathrms + stories + gashw + airco + garagepl +
##     prefarea
## Model 2: sqrt(price) ~ lotsize + bedrooms + bathrms + stories + driveway +
##     recroom + fullbase + gashw + airco + garagepl + prefarea
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    538 449768
## 2    534 404041  4     45727 15.109 1.04e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#P-value < 0.05 => choose full model
```

Figure 18

```
reduced_model5 <- lm(sqrt(price) ~ lotsize + bathrms + stories +
                       gashw + airco + garagepl, data = housing)
#remove prefarea
anova(reduced_model5, model2)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(price) ~ lotsize + bathrms + stories + gashw + airco + garagepl
## Model 2: sqrt(price) ~ lotsize + bedrooms + bathrms + stories + driveway +
##     recroom + fullbase + gashw + airco + garagepl + prefarea
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    539 498009
## 2    534 404041  5     93968 24.838 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#P-value < 0.05 => choose full model
```

Figure 19

```
mint <- lm(sqrt(price)~1,data=housing)

forwardAIC <- step(mint,scope=list(lower=~1, upper=~lotsize+bedrooms+bathrms
                              +stories+driveway+recroom+fullbase+gashw+airco+
                                garagepl+prefarea),direction="forward", data=housing)
```

```
## Start:  AIC=4242.83
## sqrt(price) ~ 1
##
##            Df Sum of Sq     RSS    AIC
## + lotsize   1    382053  907362 4053.0
## + bathrms   1    328959  960456 4084.0
```

```
## + airco     1    272098 1017316 4115.4
## + stories   1    231045 1058370 4137.0
## + garagepl  1    185448 1103967 4160.0
## + bedrooms  1    178022 1111393 4163.7
## + prefarea  1    146778 1142636 4178.8
## + driveway  1    129160 1160254 4187.2
## + recroom   1     93389 1196026 4203.8
## + fullbase  1     53251 1236163 4221.8
## + gashw     1     10880 1278535 4240.2
## <none>                  1289415 4242.8
##
## Step:  AIC=4052.96
## sqrt(price) ~ lotsize
##
##            Df Sum of Sq    RSS    AIC
## + bathrms   1    213917 693445 3908.2
## + stories   1    185297 722065 3930.2
## + airco     1    155533 751829 3952.3
## + bedrooms  1    110168 797194 3984.3
## + prefarea  1     59947 847415 4017.6
## + garagepl  1     51591 855771 4023.0
## + recroom   1     48861 858501 4024.7
## + fullbase  1     40658 866704 4029.9
## + driveway  1     35700 871662 4033.0
## + gashw     1     12100 895262 4047.6
## <none>                  907362 4053.0
##
## Step:  AIC=3908.16
## sqrt(price) ~ lotsize + bathrms
##
##            Df Sum of Sq    RSS    AIC
## + airco     1    108493 584952 3817.3
## + stories   1     90049 603396 3834.2
## + prefarea  1     55654 637791 3864.5
## + driveway  1     38365 655080 3879.1
## + bedrooms  1     32171 661274 3884.2
## + recroom   1     30450 662995 3885.6
## + garagepl  1     29951 663494 3886.0
## + fullbase  1     25024 668421 3890.1
## + gashw     1      6020 687424 3905.4
## <none>                  693445 3908.2
##
## Step:  AIC=3817.26
## sqrt(price) ~ lotsize + bathrms + airco
##
##            Df Sum of Sq    RSS    AIC
## + stories   1     49986 534966 3770.5
## + prefarea  1     46218 538734 3774.3
## + driveway  1     32495 552456 3788.1
## + bedrooms  1     23056 561896 3797.3
## + fullbase  1     22814 562138 3797.5
## + garagepl  1     22673 562279 3797.7
## + recroom   1     20640 564312 3799.6
## + gashw     1     15962 568990 3804.2
```

```
## <none>                  584952 3817.3
##
## Step:  AIC=3770.49
## sqrt(price) ~ lotsize + bathrms + airco + stories
##
##            Df Sum of Sq    RSS    AIC
## + prefarea  1     45982 488983 3723.4
## + fullbase  1     43781 491184 3725.9
## + garagepl  1     25735 509231 3745.6
## + driveway  1     24862 510103 3746.5
## + recroom   1     22413 512553 3749.1
## + gashw     1     14068 520898 3757.9
## + bedrooms  1      7320 527646 3765.0
## <none>                  534966 3770.5
##
## Step:  AIC=3723.42
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea
##
##            Df Sum of Sq    RSS    AIC
## + fullbase  1     27045 461938 3694.3
## + garagepl  1     25472 463511 3696.2
## + gashw     1     16860 472124 3706.3
## + driveway  1     16644 472339 3706.5
## + recroom   1     15254 473729 3708.1
## + bedrooms  1      6046 482938 3718.6
## <none>                  488983 3723.4
##
## Step:  AIC=3694.35
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase
##
##            Df Sum of Sq    RSS    AIC
## + garagepl  1   24668.1 437270 3666.4
## + gashw     1   15922.7 446015 3677.2
## + driveway  1   15524.2 446414 3677.7
## + recroom   1    4949.0 456989 3690.5
## + bedrooms  1    3044.1 458894 3692.7
## <none>                  461938 3694.3
##
## Step:  AIC=3666.39
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase + garagepl
##
##            Df Sum of Sq    RSS    AIC
## + gashw     1   12953.2 424317 3652.0
## + driveway  1   11380.0 425890 3654.0
## + recroom   1    5926.4 431344 3660.9
## + bedrooms  1    2047.9 435222 3665.8
## <none>                  437270 3666.4
##
## Step:  AIC=3651.97
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase + garagepl + gashw
##
```

```
##            Df Sum of Sq    RSS    AIC
## + driveway  1   11590.6 412726 3638.8
## + recroom   1    5920.9 418396 3646.3
## + bedrooms  1    1830.1 422487 3651.6
## <none>                  424317 3652.0
##
## Step:  AIC=3638.84
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase + garagepl + gashw + driveway
##
##            Df Sum of Sq    RSS    AIC
## + recroom   1    5352.7 407373 3633.7
## + bedrooms  1    3153.3 409573 3636.7
## <none>                  412726 3638.8
##
## Step:  AIC=3633.72
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase + garagepl + gashw + driveway + recroom
##
##            Df Sum of Sq    RSS    AIC
## + bedrooms  1    3332.4 404041 3631.2
## <none>                  407373 3633.7
##
## Step:  AIC=3631.23
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase + garagepl + gashw + driveway + recroom + bedrooms
```

Figure 20

```
n <- 546 #Number of observations
forwardBIC <- step(mint,scope=list(lower=~1,upper=~lotsize+bedrooms+
                  bathrms+stories+driveway+recroom+fullbase+gashw+airco+
                   garagepl+prefarea), direction="forward", data=housing,k=log(n))
```

```
## Start:  AIC=4247.13
## sqrt(price) ~ 1
##
##            Df Sum of Sq     RSS    AIC
## + lotsize   1    382053  907362 4061.6
## + bathrms   1    328959  960456 4092.6
## + airco     1    272098 1017316 4124.0
## + stories   1    231045 1058370 4145.6
## + garagepl  1    185448 1103967 4168.6
## + bedrooms  1    178022 1111393 4172.3
## + prefarea  1    146778 1142636 4187.4
## + driveway  1    129160 1160254 4195.8
## + recroom   1     93389 1196026 4212.4
## + fullbase  1     53251 1236163 4230.4
## <none>                  1289415 4247.1
## + gashw     1     10880 1278535 4248.8
##
## Step:  AIC=4061.57
```

```
## sqrt(price) ~ lotsize
##
##             Df Sum of Sq    RSS    AIC
## + bathrms    1    213917 693445 3921.1
## + stories    1    185297 722065 3943.1
## + airco      1    155533 751829 3965.2
## + bedrooms   1    110168 797194 3997.2
## + prefarea   1     59947 847415 4030.5
## + garagepl   1     51591 855771 4035.9
## + recroom    1     48861 858501 4037.6
## + fullbase   1     40658 866704 4042.8
## + driveway   1     35700 871662 4046.0
## + gashw      1     12100 895262 4060.5
## <none>                   907362 4061.6
##
## Step:  AIC=3921.06
## sqrt(price) ~ lotsize + bathrms
##
##             Df Sum of Sq    RSS    AIC
## + airco      1    108493 584952 3834.5
## + stories    1     90049 603396 3851.4
## + prefarea   1     55654 637791 3881.7
## + driveway   1     38365 655080 3896.3
## + bedrooms   1     32171 661274 3901.4
## + recroom    1     30450 662995 3902.9
## + garagepl   1     29951 663494 3903.3
## + fullbase   1     25024 668421 3907.3
## <none>                   693445 3921.1
## + gashw      1      6020 687424 3922.6
##
## Step:  AIC=3834.47
## sqrt(price) ~ lotsize + bathrms + airco
##
##             Df Sum of Sq    RSS    AIC
## + stories    1     49986 534966 3792.0
## + prefarea   1     46218 538734 3795.8
## + driveway   1     32495 552456 3809.6
## + bedrooms   1     23056 561896 3818.8
## + fullbase   1     22814 562138 3819.1
## + garagepl   1     22673 562279 3819.2
## + recroom    1     20640 564312 3821.2
## + gashw      1     15962 568990 3825.7
## <none>                   584952 3834.5
##
## Step:  AIC=3792
## sqrt(price) ~ lotsize + bathrms + airco + stories
##
##             Df Sum of Sq    RSS    AIC
## + prefarea   1     45982 488983 3749.2
## + fullbase   1     43781 491184 3751.7
## + garagepl   1     25735 509231 3771.4
## + driveway   1     24862 510103 3772.3
## + recroom    1     22413 512553 3774.9
## + gashw      1     14068 520898 3783.8
```

```
## + bedrooms  1      7320 527646 3790.8
## <none>                  534966 3792.0
##
## Step:  AIC=3749.23
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea
##
##            Df Sum of Sq    RSS    AIC
## + fullbase  1     27045 461938 3724.5
## + garagepl  1     25472 463511 3726.3
## + gashw     1     16860 472124 3736.4
## + driveway  1     16644 472339 3736.6
## + recroom   1     15254 473729 3738.2
## + bedrooms  1      6046 482938 3748.7
## <none>                  488983 3749.2
##
## Step:  AIC=3724.47
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase
##
##            Df Sum of Sq    RSS    AIC
## + garagepl  1   24668.1 437270 3700.8
## + gashw     1   15922.7 446015 3711.6
## + driveway  1   15524.2 446414 3712.1
## <none>                  461938 3724.5
## + recroom   1    4949.0 456989 3724.9
## + bedrooms  1    3044.1 458894 3727.2
##
## Step:  AIC=3700.81
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase + garagepl
##
##            Df Sum of Sq    RSS    AIC
## + gashw     1   12953.2 424317 3690.7
## + driveway  1   11380.0 425890 3692.7
## + recroom   1    5926.4 431344 3699.7
## <none>                  437270 3700.8
## + bedrooms  1    2047.9 435222 3704.5
##
## Step:  AIC=3690.69
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase + garagepl + gashw
##
##            Df Sum of Sq    RSS    AIC
## + driveway  1   11590.6 412726 3681.9
## + recroom   1    5920.9 418396 3689.3
## <none>                  424317 3690.7
## + bedrooms  1    1830.1 422487 3694.6
##
## Step:  AIC=3681.87
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase + garagepl + gashw + driveway
##
##            Df Sum of Sq    RSS    AIC
## + recroom   1    5352.7 407373 3681.0
```

```
## <none>                  412726 3681.9
## + bedrooms  1    3153.3 409573 3684.0
##
## Step:  AIC=3681.05
## sqrt(price) ~ lotsize + bathrms + airco + stories + prefarea +
##     fullbase + garagepl + gashw + driveway + recroom
##
##            Df Sum of Sq    RSS    AIC
## <none>                  407373 3681.0
## + bedrooms  1    3332.4 404041 3682.9
```

Figure 21

```
backAIC <- step(model2, direction = "backward", data = housing)
```

```
## Start:  AIC=3631.23
## sqrt(price) ~ lotsize + bedrooms + bathrms + stories + driveway +
##     recroom + fullbase + gashw + airco + garagepl + prefarea
##
##            Df Sum of Sq    RSS    AIC
## <none>                  404041 3631.2
## - bedrooms  1      3332 407373 3633.7
## - recroom   1      5532 409573 3636.7
## - driveway  1     12352 416393 3645.7
## - fullbase  1     12485 416526 3645.8
## - gashw     1     12872 416913 3646.4
## - garagepl  1     17397 421438 3652.3
## - prefarea  1     24663 428704 3661.6
## - stories   1     41058 445099 3682.1
## - airco     1     53911 457952 3697.6
## - bathrms   1     62775 466816 3708.1
## - lotsize   1     84500 488541 3732.9
```

Figure 22

```
backBIC <- step(model2,direction="backward", data=bridge, k=log(n))
```

```
## Start:  AIC=3682.86
## sqrt(price) ~ lotsize + bedrooms + bathrms + stories + driveway +
##     recroom + fullbase + gashw + airco + garagepl + prefarea
##
##            Df Sum of Sq    RSS    AIC
## - bedrooms  1      3332 407373 3681.0
## <none>                  404041 3682.9
## - recroom   1      5532 409573 3684.0
## - driveway  1     12352 416393 3693.0
## - fullbase  1     12485 416526 3693.2
## - gashw     1     12872 416913 3693.7
## - garagepl  1     17397 421438 3699.6
```

```
## - prefarea  1     24663 428704 3708.9
## - stories   1     41058 445099 3729.4
## - airco     1     53911 457952 3744.9
## - bathrms   1     62775 466816 3755.4
## - lotsize   1     84500 488541 3780.3
##
## Step:  AIC=3681.05
## sqrt(price) ~ lotsize + bathrms + stories + driveway + recroom +
##     fullbase + gashw + airco + garagepl + prefarea
##
##             Df Sum of Sq    RSS    AIC
## <none>                   407373 3681.0
## - recroom   1      5353 412726 3681.9
## - driveway  1     11022 418396 3689.3
## - gashw     1     13154 420527 3692.1
## - fullbase  1     14677 422051 3694.1
## - garagepl  1     18665 426039 3699.2
## - prefarea  1     25063 432436 3707.3
## - airco     1     53651 461024 3742.3
## - stories   1     56825 464199 3746.0
## - bathrms   1     73208 480581 3765.0
## - lotsize   1     87811 495184 3781.3
```

Figure 23

```
summary(backAIC) #full model has a higher adjusted R^2
```

```
##
## Call:
## lm(formula = sqrt(price) ~ lotsize + bedrooms + bathrms + stories +
##     driveway + recroom + fullbase + gashw + airco + garagepl +
##     prefarea, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.632 -16.701   0.351  15.353  95.826
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.219e+02  6.081e+00  20.054  < 2e-16 ***
## lotsize     6.602e-03  6.248e-04  10.568  < 2e-16 ***
## bedrooms    3.919e+00  1.867e+00   2.099  0.03632 *
## bathrms     2.420e+01  2.657e+00   9.109  < 2e-16 ***
## stories     1.216e+01  1.650e+00   7.366 6.69e-13 ***
## driveway    1.474e+01  3.648e+00   4.040 6.12e-05 ***
## recroom     9.162e+00  3.389e+00   2.704  0.00707 **
## fullbase    1.150e+01  2.832e+00   4.062 5.59e-05 ***
## gashw       2.367e+01  5.739e+00   4.125 4.31e-05 ***
## airco       2.341e+01  2.773e+00   8.441 2.97e-16 ***
## garagepl    7.188e+00  1.499e+00   4.795 2.11e-06 ***
## prefarea    1.700e+01  2.977e+00   5.709 1.88e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.51 on 534 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6802
## F-statistic: 106.4 on 11 and 534 DF,  p-value: < 2.2e-16
```

```
summary(backBIC)
```

```
##
## Call:
## lm(formula = sqrt(price) ~ lotsize + bathrms + stories + driveway +
##     recroom + fullbase + gashw + airco + garagepl + prefarea,
##     data = housing)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75.09 -16.47  -0.35  15.48  98.83
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.295e+02  4.909e+00  26.383  < 2e-16 ***
## lotsize     6.708e-03  6.247e-04  10.739  < 2e-16 ***
## bathrms     2.546e+01  2.597e+00   9.805  < 2e-16 ***
## stories     1.338e+01  1.549e+00   8.639  < 2e-16 ***
## driveway    1.382e+01  3.633e+00   3.805 0.000158 ***
## recroom     9.011e+00  3.399e+00   2.651 0.008255 **
## fullbase    1.235e+01  2.812e+00   4.390 1.36e-05 ***
## gashw       2.392e+01  5.755e+00   4.156 3.77e-05 ***
## airco       2.335e+01  2.782e+00   8.394 4.22e-16 ***
## garagepl    7.425e+00  1.500e+00   4.951 9.91e-07 ***
## prefarea    1.713e+01  2.986e+00   5.737 1.61e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.59 on 535 degrees of freedom
## Multiple R-squared:  0.6841, Adjusted R-squared:  0.6782
## F-statistic: 115.8 on 10 and 535 DF,  p-value: < 2.2e-16
```

Figure 24

```
summary(model2)
```

```
##
## Call:
## lm(formula = sqrt(price) ~ lotsize + bedrooms + bathrms + stories +
##     driveway + recroom + fullbase + gashw + airco + garagepl +
##     prefarea, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.632 -16.701   0.351  15.353  95.826
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.219e+02  6.081e+00  20.054  < 2e-16 ***
## lotsize     6.602e-03  6.248e-04  10.568  < 2e-16 ***
## bedrooms    3.919e+00  1.867e+00   2.099  0.03632 *
## bathrms     2.420e+01  2.657e+00   9.109  < 2e-16 ***
## stories     1.216e+01  1.650e+00   7.366 6.69e-13 ***
## driveway    1.474e+01  3.648e+00   4.040 6.12e-05 ***
## recroom     9.162e+00  3.389e+00   2.704  0.00707 **
## fullbase    1.150e+01  2.832e+00   4.062 5.59e-05 ***
## gashw       2.367e+01  5.739e+00   4.125 4.31e-05 ***
## airco       2.341e+01  2.773e+00   8.441 2.97e-16 ***
## garagepl    7.188e+00  1.499e+00   4.795 2.11e-06 ***
## prefarea    1.700e+01  2.977e+00   5.709 1.88e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.51 on 534 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6802
## F-statistic: 106.4 on 11 and 534 DF,  p-value: < 2.2e-16
```