Sta 220 Final Project

Richard Ly

3/18/2022

Contents

1 Introduction **1**

2 Material and Methods **1**

2.1 Exploratory Data Analysis 2

2.2 Modeling Price Due to Number of Beds, Baths, Garages, Sqft, Lotsize 12

2.3 Modeling Sqft Under the Effects of Number of Beds, Baths, Garages as Factors 16

3 Conclusions **20**

4 Appendix **20**

1 Introduction

We are interested in knowing how several recorded features of a house may affect the house's sale price. To answer this question of interest, we we will explore a dataset concerning the price of houses sold in Gainesville, Florida between September 2019 and March 2020. In doing so, we aim to gain insight into which features significantly influence a house's sale price, particularly in Gainesville, Florida. Such insights may allow both buyers and sellers of houses in Gainseville to make informed decisions based on previous, relatively recently collected data. To phrase the question of interest precisely: Which key variables recorded in the dataset are significant in determining house sale price in Gainesville, Florida? An additional question of interest: How is house size (area measured in square feet) affected by key variables? We list the key variables recorded in the dataset here: number of bedrooms, number of bathrooms, number of garages, house area (square feet), and lot size. To begin, in the Material and Methods section we will explore a number of these variables, obtaining summary statistics and visualizing their distribution and pairwise relationships. In addition, we will perform linear regression to model house sale price as a function of the aforementioned key variables. The results of our work will be summarized in the Conclusion section.

2 Material and Methods

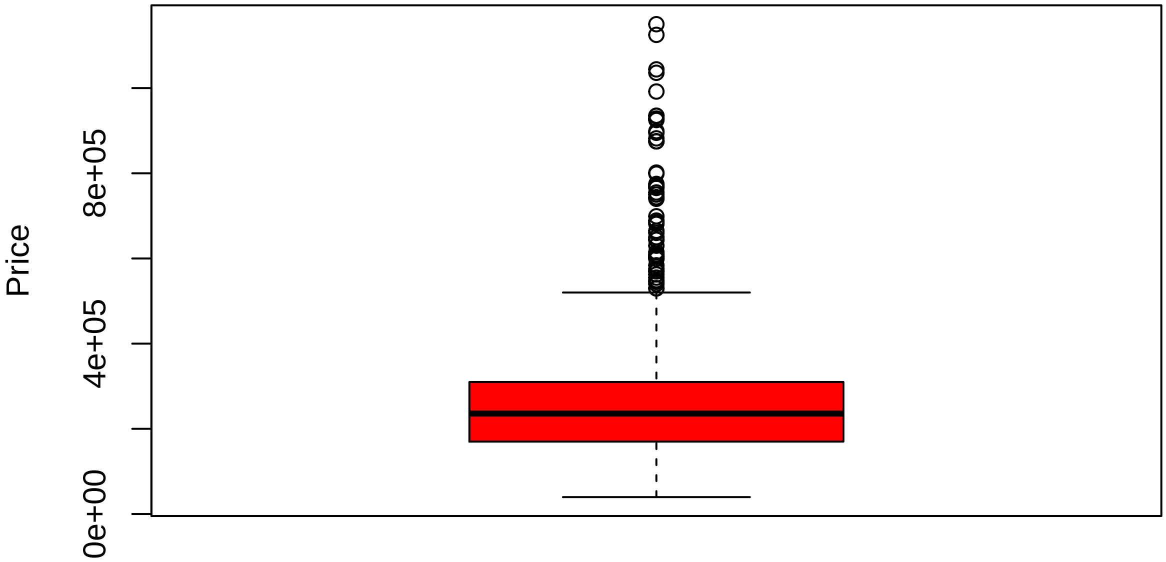
We begin by exploring the variables of interest in the dataset. We perform exploratory data analysis on these variables to obtain insight on their distributions. In addition, we may visualize the data to observe pairwise relationships, correlation, between the quantitative variables. In further visualizing the data, we may observe trends that may influence the model equation to be formulated.

2.1 Exploratory Data Analysis

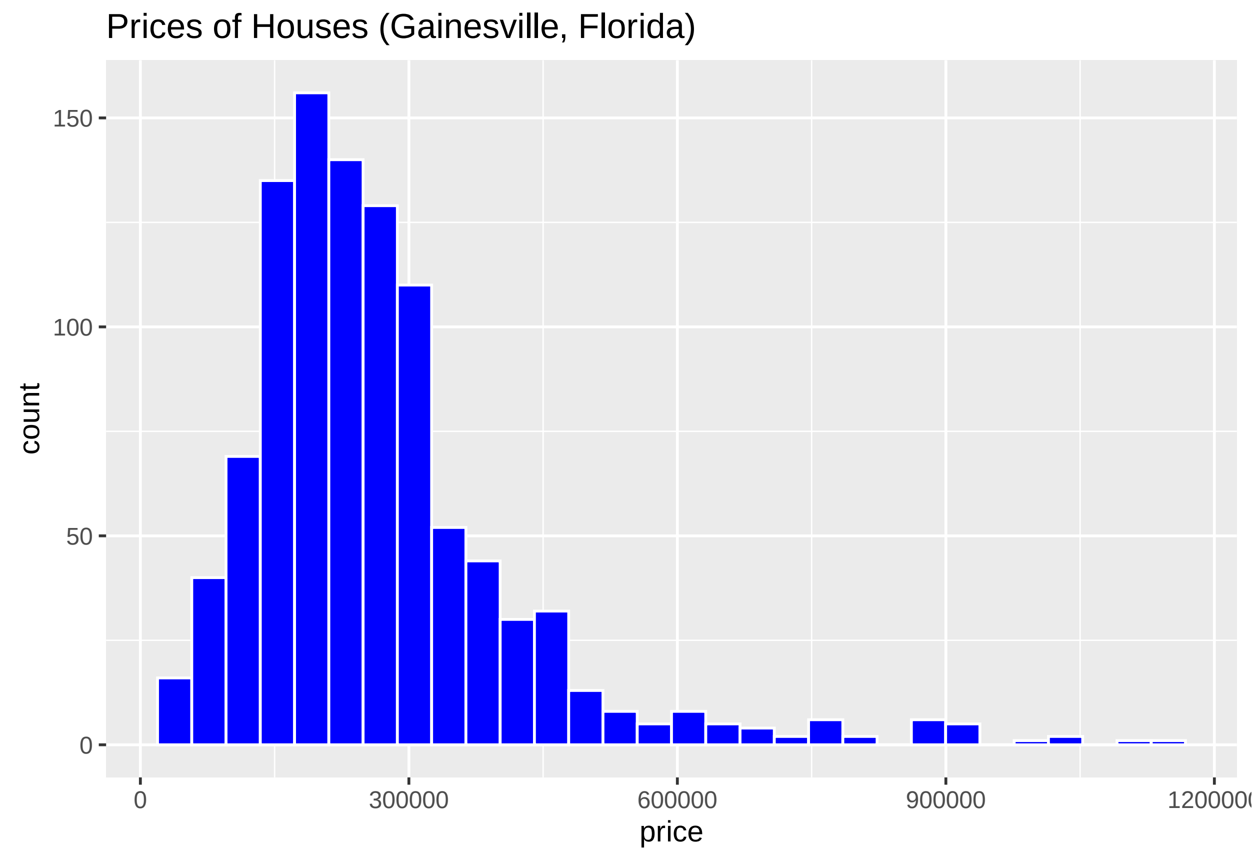
2.1.1 Summary Statistics

After cleaning the data, removing NA values and extreme outliers, we examine the summary statistics. We begin with house prices. There are a number of high price outliers, relative to the majority of the values in the data. We plot the prices in a histogram and observe a unimodal distribution. It has a right skew, matching the previous observation of high outliers. As the `price` is not normally distributed, due to being skewed, it is likely that a transformation of price will need to be performed when fitting a model that will rely on the assumption of normality, such as linear regression.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   39700  170000  235800  266638  310000 1150000
```

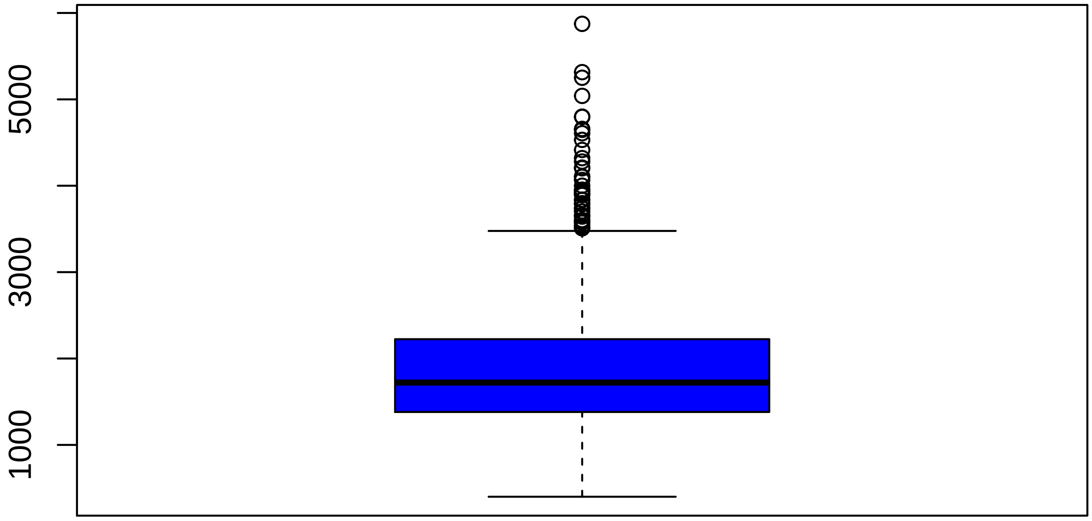


Boxplot of House Prices



We list the summary statistics and create a boxplot to visualize the distribution of house area, measured in square feet.

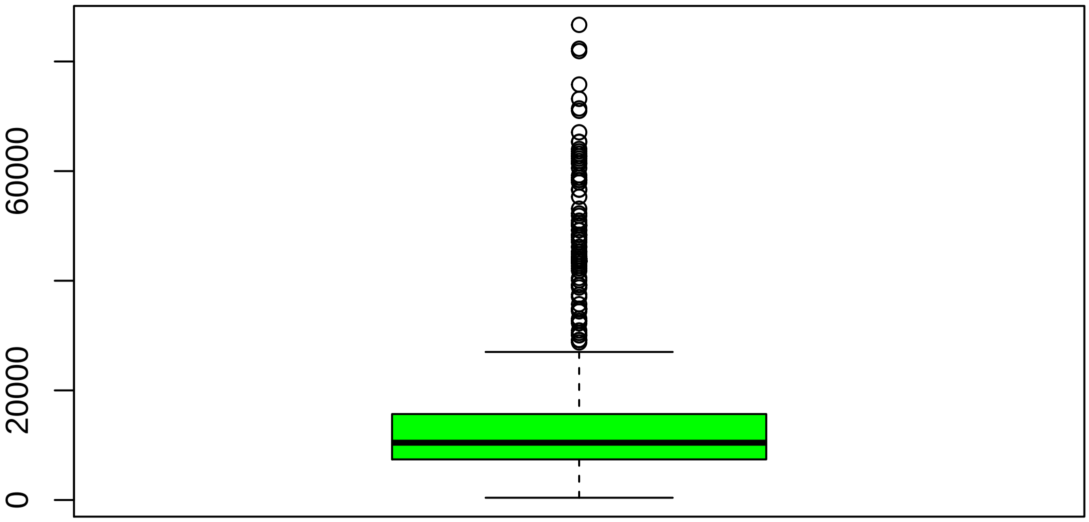
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     400    1381    1721    1891    2221    5874
```



Boxplot of House Area (Sq. Ft.)

Similary we look at the summary statistics of lotsize. However, we notice that the majority of the values for lotsize are relatively small, compared to a number of extremely large outliers. Instead, we look at the summary statistics and distribution of lotsize after subsetting the data to only include lotsizes between 300 and 90,000 square feet (very roughly 2 acres) with a boxplot. This allows us to magnify and better view the distribution of the majority of the data.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     413    7405   10454   14893   15682   86684
```



Boxplot of Lot Size (Sq. Ft.)

We can display a count of how many houses have a number of bedrooms, as we treat number of bedrooms as a categorical variable.

```
##      Count
## 0 Bd     2
## 1 Bd     7
## 2 Bd    98
## 3 Bd   553
## 4 Bd   311
## 5 Bd    45
## 6 Bd     5
## 7 Bd     1
```

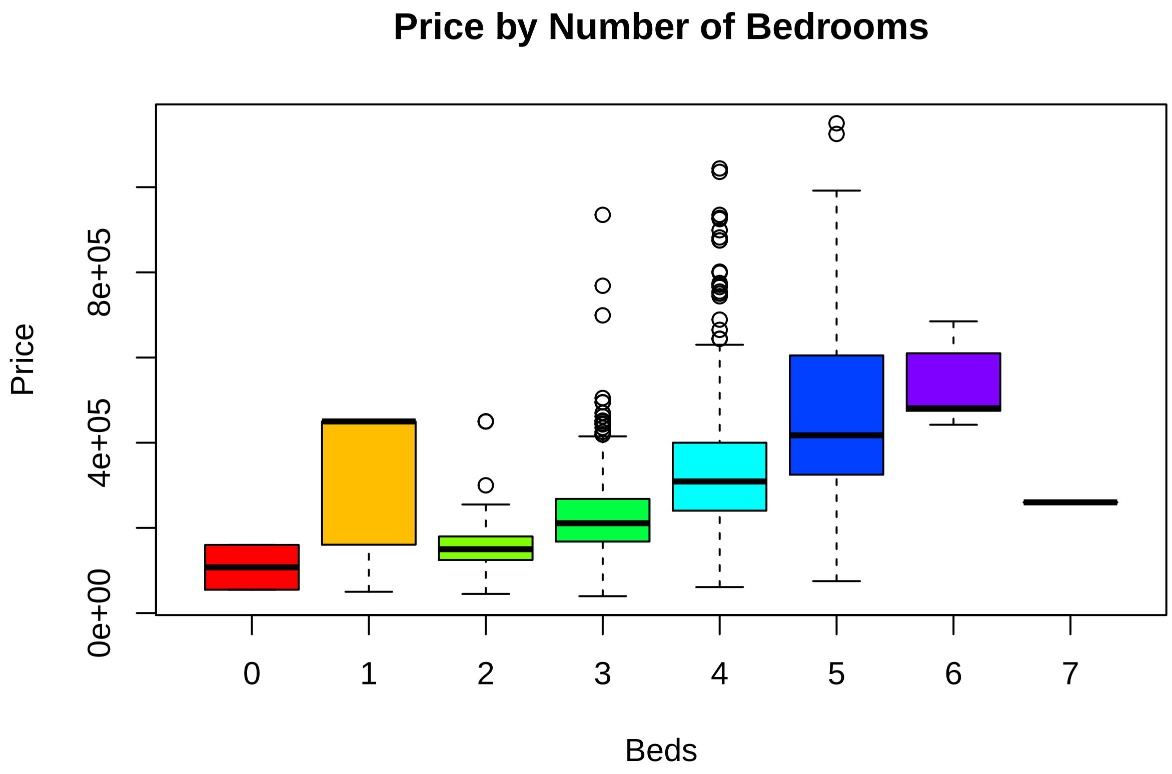
We do the same for number of bathrooms, and treat it as a categorical variable. Below we display the count of number of bathrooms in each house.

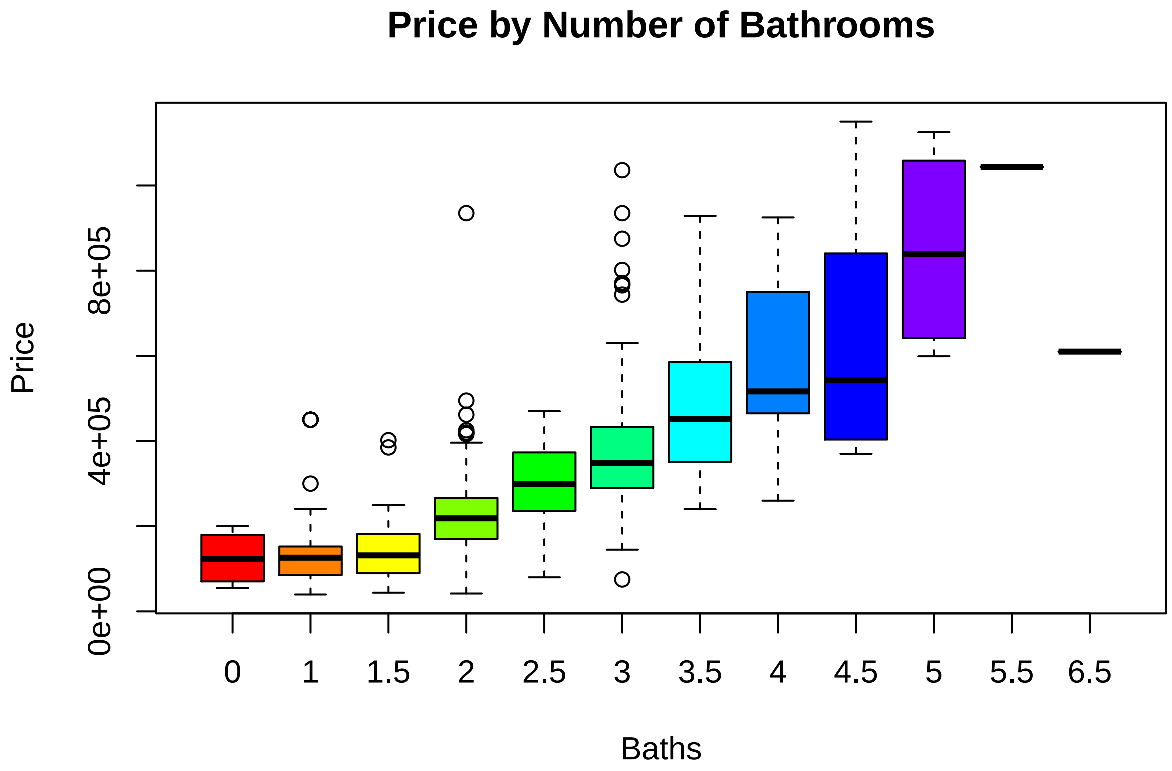
```
##        Count
## 0 Ba       4
## 1 Ba      86
## 1.5 Ba    19
## 2 Ba     622
## 2.5 Ba    71
## 3 Ba     137
## 3.5 Ba    40
## 4 Ba      21
## 4.5 Ba    16
## 5 Ba       4
## 5.5 Ba     1
## 6.5 Ba     1
```

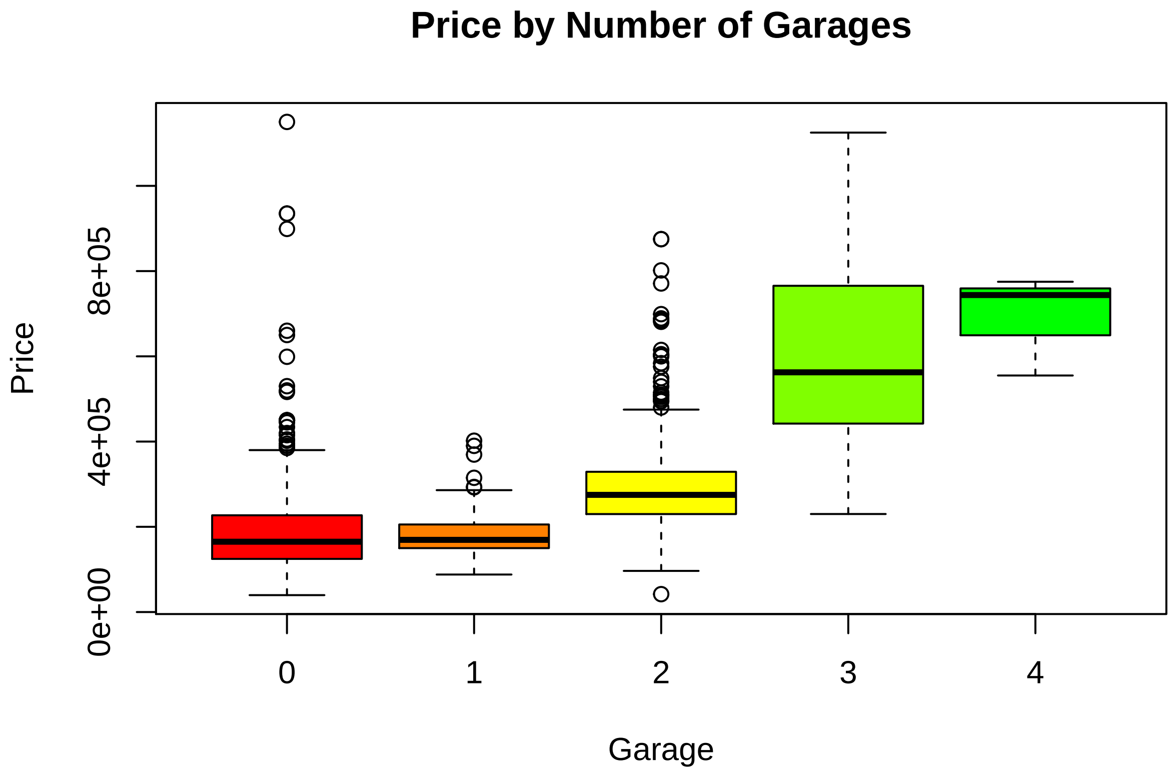
Lastly we look at number of garages in each house, again treated as a categorical variable.

```
##            Count
## 0 Garages   377
## 1 Garage     67
## 2 Garages   526
## 3 Garages    49
## 4 Garages     3
```

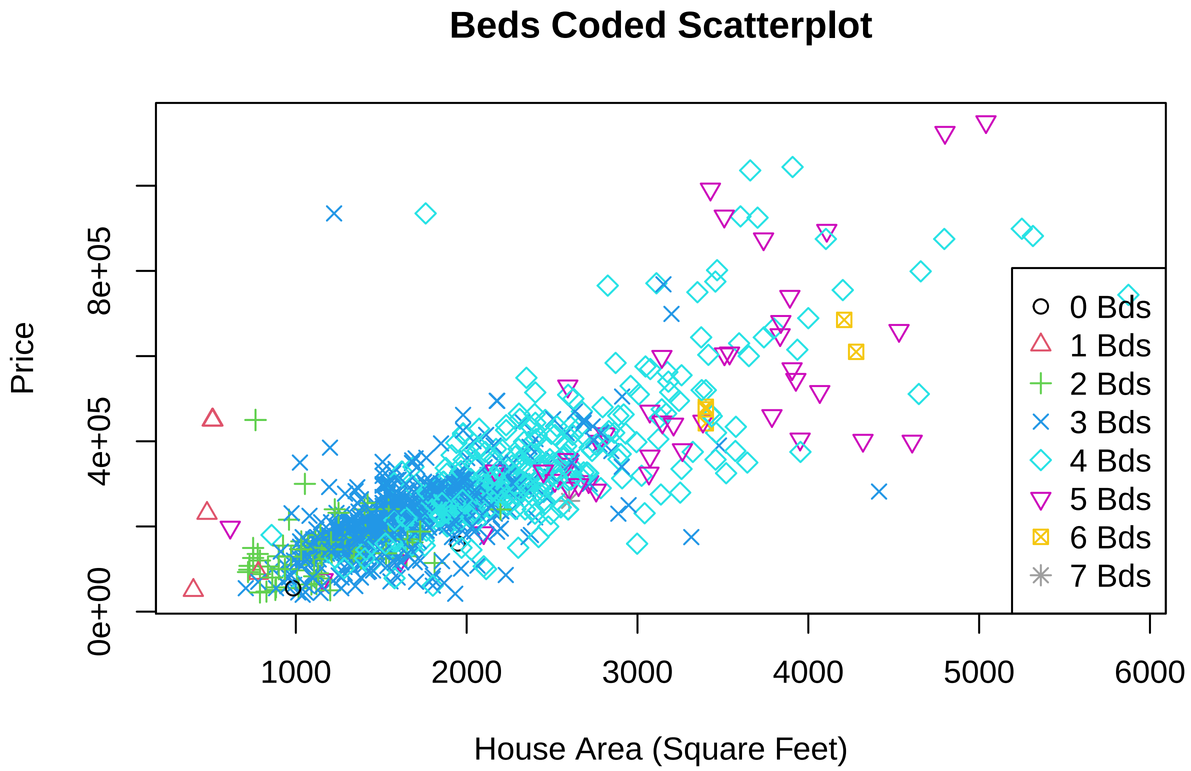
Since we decided to treat number of `bedrooms`, `bathrooms`, and `garages` as categorical, we can examine the distribution of price due to each of these variables by their factor level. We do this by constructing side-by-side boxplots of price, for each categorical variable. We observe that price tends to increase as number of `bedrooms` increases, number of `bathrooms`, or number of `garages` increases. We note that there exist outliers, such as houses with 7 bedrooms being a relatively low price, or houses with 6.5 bathrooms being relatively low in price. Such factor levels are low in frequency and an exception to the general trend. Notably, houses with 1 bedroom have a wide range of values, and are also an exception to the general trend.

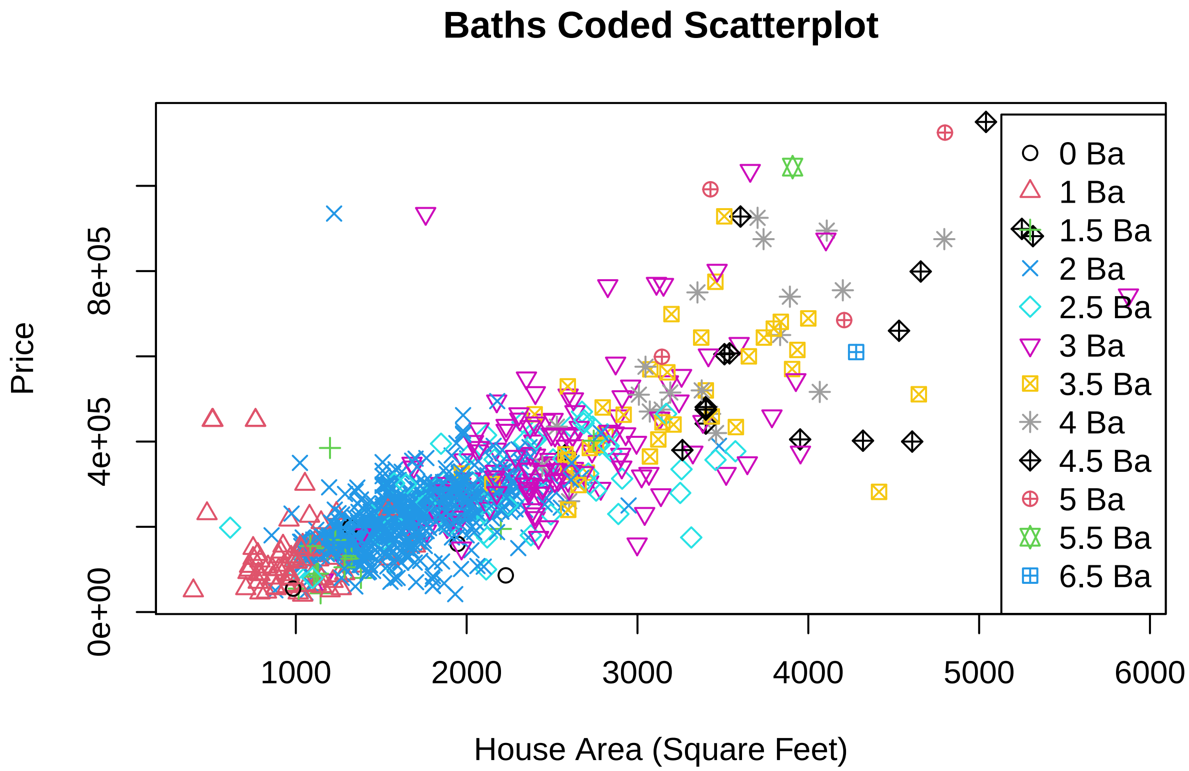


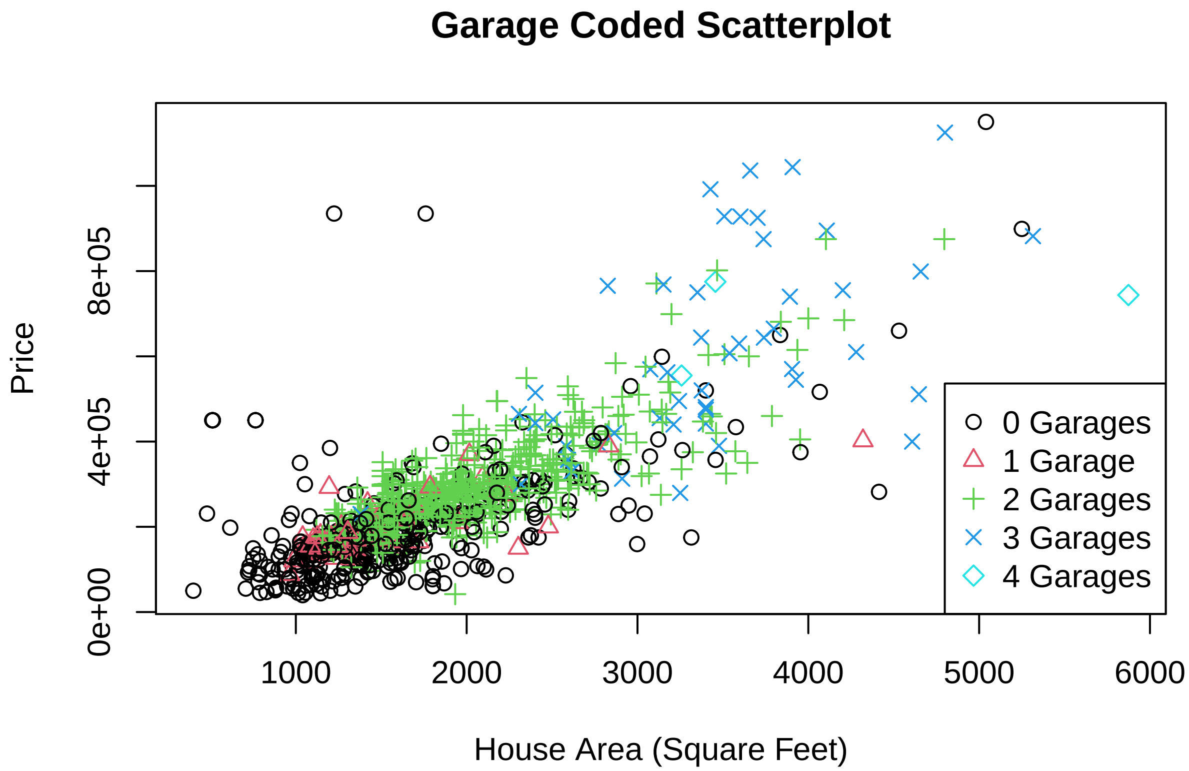

Price by Number of Bathrooms
Price
0e+00
4e+05
8e+05
0
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6.5
Baths



We further visualize the distribution of each categorical variable, as well as its relationship towards price, with scatter plots. The scatter plots are color coded and use various point symbols for factor levels to become more identifiable. We note that houses with 0 or 2 garages high a fair number of high outliers.

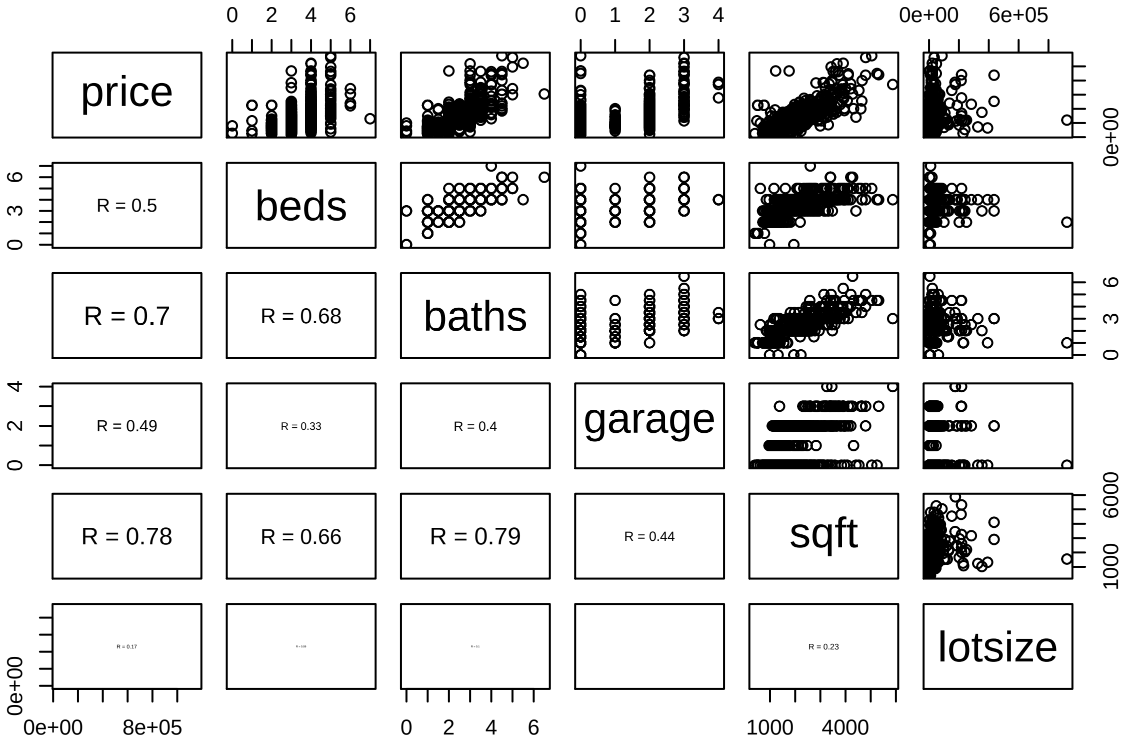

Beds Coded Scatterplot
Price
0e+00
4e+05
8e+05
1000
2000
3000
4000
5000
6000
House Area (Square Feet)
○ 0 Bds
△ 1 Bds
+ 2 Bds
× 3 Bds
◇ 4 Bds
▽ 5 Bds
⊠ 6 Bds
✳ 7 Bds


Baths Coded Scatterplot
Price
0e+00
4e+05
8e+05
1000
2000
3000
4000
5000
6000
House Area (Square Feet)
○ 0 Ba
△ 1 Ba
+ 1.5 Ba
× 2 Ba
◇ 2.5 Ba
▽ 3 Ba
☒ 3.5 Ba
✳ 4 Ba
◇ 4.5 Ba
⊕ 5 Ba
✡ 5.5 Ba
⊞ 6.5 Ba



2.1.2 Pairwise Correlation and Relationships

We are interested in the intervariable relationships, as they may indicate linear or quadratric relationships that should be included in our model. Because of this, we examine the pairwise correlation and pairwise scatter plots in the below table. Note that there appears to be a linear trend in which the price of the house increases alongside an increase in house size (`sqft`), as well as the number of bedrooms (`beds`), bathrooms (`baths`), and garages (`garage`). Interestingly, `lotsize` does not appear to have a pairwise relationship with the `price` of the house. We similarly observe a linear trend for `sqft` increasing pairwise alongside an increase of each of the following key variables: `beds`, `baths`, and ‘garages. These observations concerning the questions of interest will inform our model choice and equation.



2.2 Modeling Price Due to Number of Beds, Baths, Garages, Sqft, Lotsize

We would like to model price as a function of key variables: number of bedrooms, number of bathrooms, number of garages, house size (sqft), and lotsize. From the previous exploratory data analysis, it seems appropriate to use a first order linear regression model, with categorical variables, due to the linear pairwise relationships between various key variables.

2.2.1 Linear Regression Using Bayesian Information Criterion (BIC)

For practical use, we decide to find a model with minimal BIC. The BIC is a measurement of goodness of fit and model complexity. It is used to avoid underfitting (high bias) and overfitting (high variance). For the final model found, we obtain that `sqft`, `beds`, `baths`, `garage` are in the model that minimizes BIC. By minimizing BIC, it can improve the model's capability in terms of prediction. After performing stepwise bi-directional regression, we have obtained a model with highest prediction capability, in terms of the BIC criterion.

$$Y = b_0 + b_{1,1}X_{1,1} + \cdots + b_{1,8}X_{1,8} + b_{2,1}X_{2,1} + \cdots + b_{2,12}X_{2,12} + b_{3,1}X_{3,1} + \cdots + b_{3,5}X_{3,5} + b_4X_4 + \epsilon$$

We define the variables in our model as such:

For the q^{th} qualitative variable with classes, labeled as $C_1, \cdots, C_r$ is represented by $r-1$ dummy variables:

$$Y_{qp} = \begin{cases} 1 \text{ if } C_p \\ 0 \text{ otherwise} \end{cases} \quad , p \in [1, \ldots, r-1]$$

and the r^{th} class corresponds to $Y_{q1} = Y_{q2} = \cdots = Y_{qp} = 0$

- Y: The response variable, house sale `price`, for each observation
- b_0: The intercept of the model equation
- b_{st}: The coefficient for a predictor s with class C_t
- X_{1j}: A dummy variable for the number of bedrooms, $j \in [1, \ldots, 8]$ for 8 classes
- X_{2k}: A dummy variable for the number of bathrooms, $k \in [1, \ldots, 12]$ for 12 classes
- $X_{3\ell}$: A dummy variable for the number of garages, $\ell \in [1, \ldots, 5]$ for 5 classes
- X_4: The house size, measured in square feet
- b_4: The coefficient for house size
- ϵ: the statistical error for each observation, $\epsilon \overset{iid}{\sim} N(\mu, \sigma^2)$ is assumed

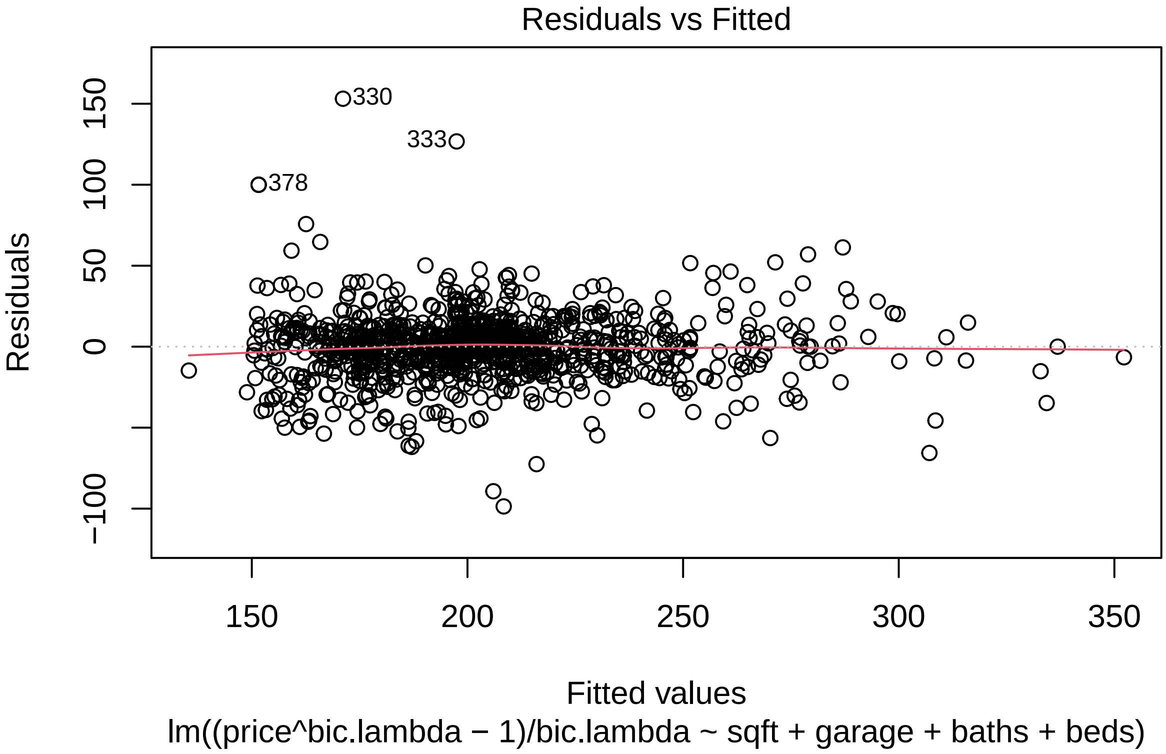
```
#Linear Regression Model: beds, baths, garage, sqft, lotsize on price
none_mod = lm(price~1,data=houses)
full_mod = lm(price~beds+baths+garage+sqft+lotsize,data=houses)
BIC = stepAIC(none_mod, scope = list(upper = full_mod, lower = ~1), direction = "both",
    k = log(length(houses)), trace = FALSE)
summary(BIC)
plot(BIC)

bc.bic = boxcox(BIC)
bic.lambda = bc.bic$x[which.max(bc.bic$y)]
BIC2 = update(BIC,(price^bic.lambda-1)/bic.lambda~.,data=houses)
```

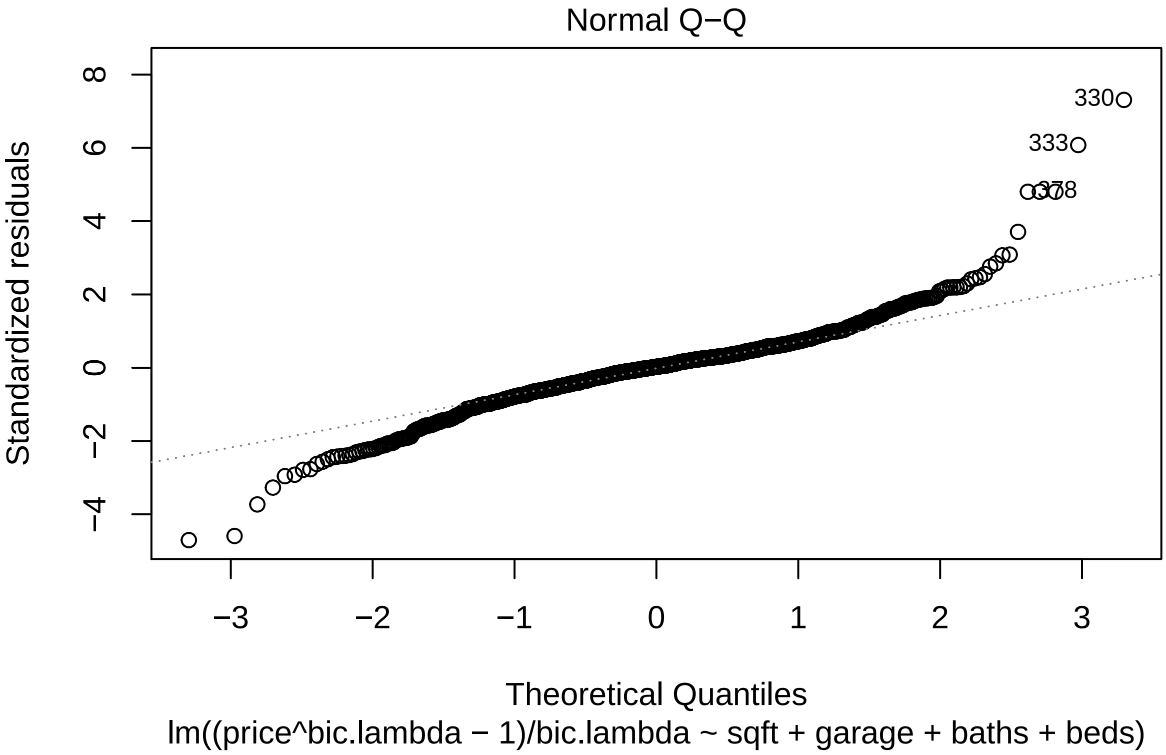
```
summary(BIC2)
```

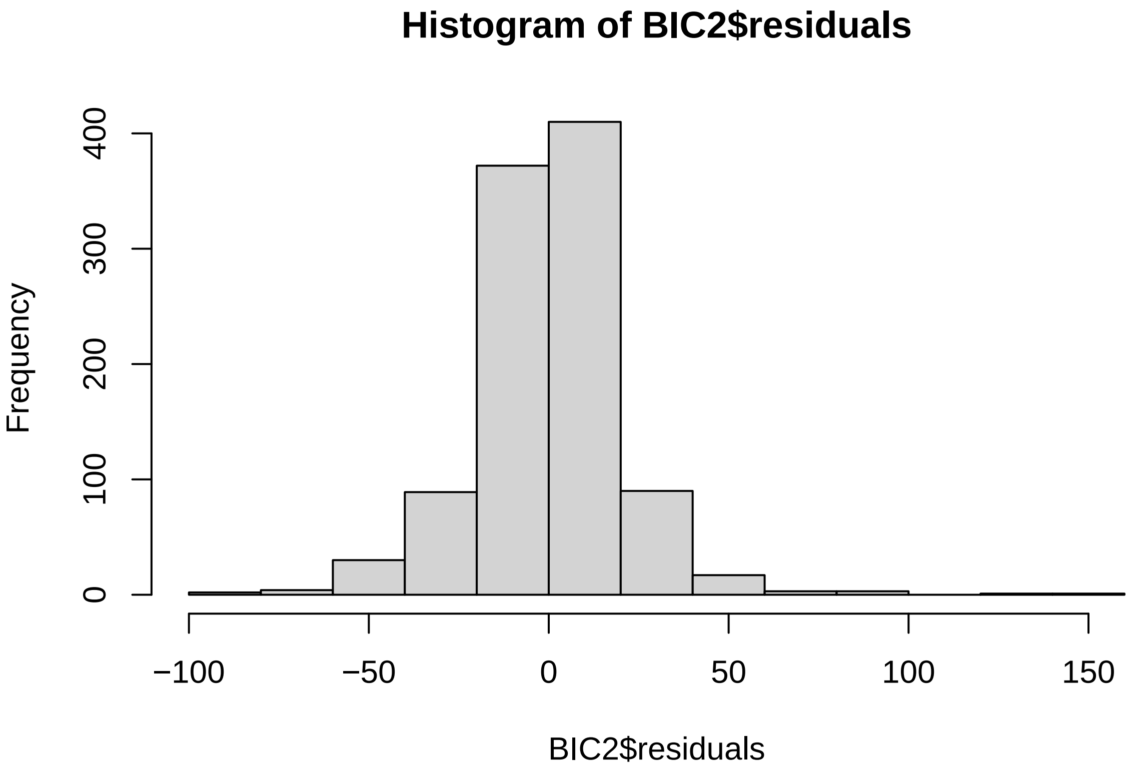
```
##
## Call:
## lm(formula = (price^bic.lambda - 1)/bic.lambda ~ sqft + garage +
##     baths + beds, data = houses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.614 -10.439   0.483   9.784 153.094
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 109.488993  15.055387   7.272 7.13e-13 ***
## sqft          0.026331   0.001695  15.539  < 2e-16 ***
## garage1       6.004166   2.825521   2.125  0.03383 *
## garage2      18.600313   1.609350  11.558  < 2e-16 ***
## garage3      31.457216   3.852024   8.166 9.53e-16 ***
## garage4      28.481923  12.639329   2.253  0.02445 *
## baths1       11.506697  15.117873   0.761  0.44676
## baths1.5     12.233612  15.637639   0.782  0.43421
## baths2       20.153753  14.922046   1.351  0.17713
## baths2.5     22.792622  15.136602   1.506  0.13244
## baths3       31.676173  15.101443   2.098  0.03620 *
## baths3.5     35.237369  15.461912   2.279  0.02288 *
## baths4       46.696240  15.908341   2.935  0.00341 **
## baths4.5     42.017489  16.524363   2.543  0.01115 *
## baths5       82.067526  19.092839   4.298 1.89e-05 ***
## baths5.5     83.023036  26.132234   3.177  0.00153 **
## baths6.5     46.846868  28.823073   1.625  0.10441
## beds1        74.483419  22.607399   3.295  0.00102 **
## beds2        10.477496  21.145294   0.496  0.62036
## beds3         9.265445  21.006412   0.441  0.65925
## beds4         9.982137  21.100999   0.473  0.63627
## beds5         2.819551  21.415756   0.132  0.89528
## beds6       -20.888741  24.303231  -0.860  0.39027
## beds7       -16.770123  30.184994  -0.556  0.57862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21 on 998 degrees of freedom
## Multiple R-squared:  0.6954, Adjusted R-squared:  0.6883
## F-statistic: 99.04 on 23 and 998 DF,  p-value: < 2.2e-16
```

2.2.2 Final Model Diagnostics

We examine the residuals vs fitted value plot to check that the assumption of constant variance in errors is maintained. The residuals vs fitted values plot seems to indicate residuals with constant variance, and that assumption has not been violated. We examine the normal Q-Q plot to check for the assumption of normally distributed errors. We find that the residuals for the final BIC model is somewhat normal in appearance, although having heavy tails, even after a BoxCox transformation.




Normal Q−Q
Standardized residuals
8
6
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
330
333
378
Theoretical Quantiles
lm((price^bic.lambda − 1)/bic.lambda ~ sqft + garage + baths + beds)



2.3 Modeling Sqft Under the Effects of Number of Beds, Baths, Garages as Factors

Given that we treat number of `beds`, `baths`, and `garages` as categorical variables, we note that an ANOVA model seems appropriate to model `sqft`. We give the final model's summary below. The exact procedure used to arrive at this final model may be found in the Appendix section. An initial model was fitted, and was updated in iterations. In each iteration, insignificant factors were removed. A transformation was performed on the response variable, `price`, to obtain a final model that better fit the model assumptions of normally distributed residuals with constant variance.

2.3.1 ANOVA

For an obserbation ℓ:

$$Y_{ijk\ell} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + \epsilon_{ijk\ell}$$

We define the variables as such:

- Y_{ijkl}: The house price
- μ: The population mean for house price
- α_i: The factor effect of number of bedrooms
- $i = 1, \ldots, 8$: indexes the number of classes for bedrooms
- β_j: The factor effect of number of bathrooms

- $j = 1, \ldots, 12$: indexes the number of classes for bathrooms
- $(\alpha\beta)_{ij}$: The interaction effect between factors α under class i and β under class j, of which there are $i * j$ terms
- γ_k: The factor effect of number of garages
- $k = 1, \ldots, 5$: indexes the number of classes for garages
- $\epsilon_{ijk\ell}$: The statistical error for each observation ℓ, $\epsilon_{ijk\ell} \overset{iid}{\sim} N(\mu, \sigma^2)$ is assumed

```
#ANOVA model: beds, baths, garage as factors on sqft
anova.fit = aov(sqft~beds*baths*garage,data=houses)
summary(anova.fit)
anova.fit2 = update(anova.fit,.~.-beds:garage,data=houses)
summary(anova.fit2)
plot(anova.fit2)
bc.anova = boxcox(anova.fit2)
anova.fit3 = update(anova.fit2,log(price)~.,data=houses)
summary(anova.fit3)
anova.fit4 = update(anova.fit3,log(price)~.-baths:garage-beds:baths:garage,data=houses)
summary(anova.fit4)
```

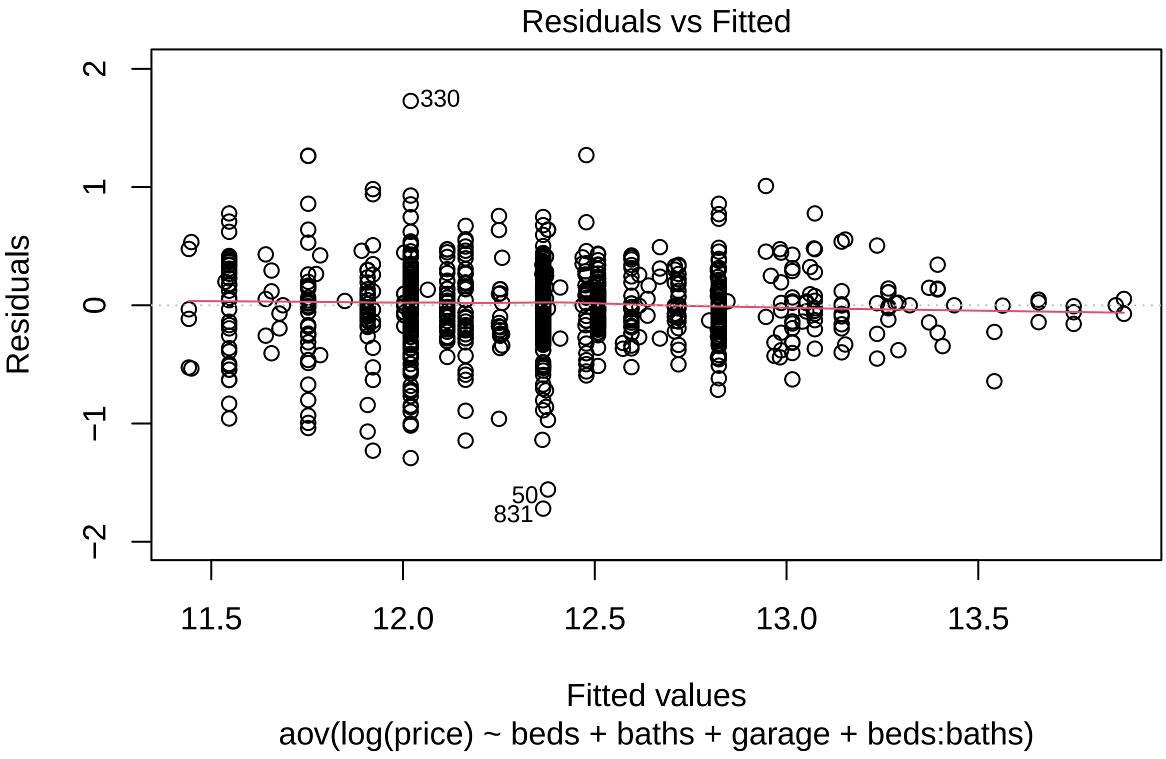
```
summary(anova.fit4)
```

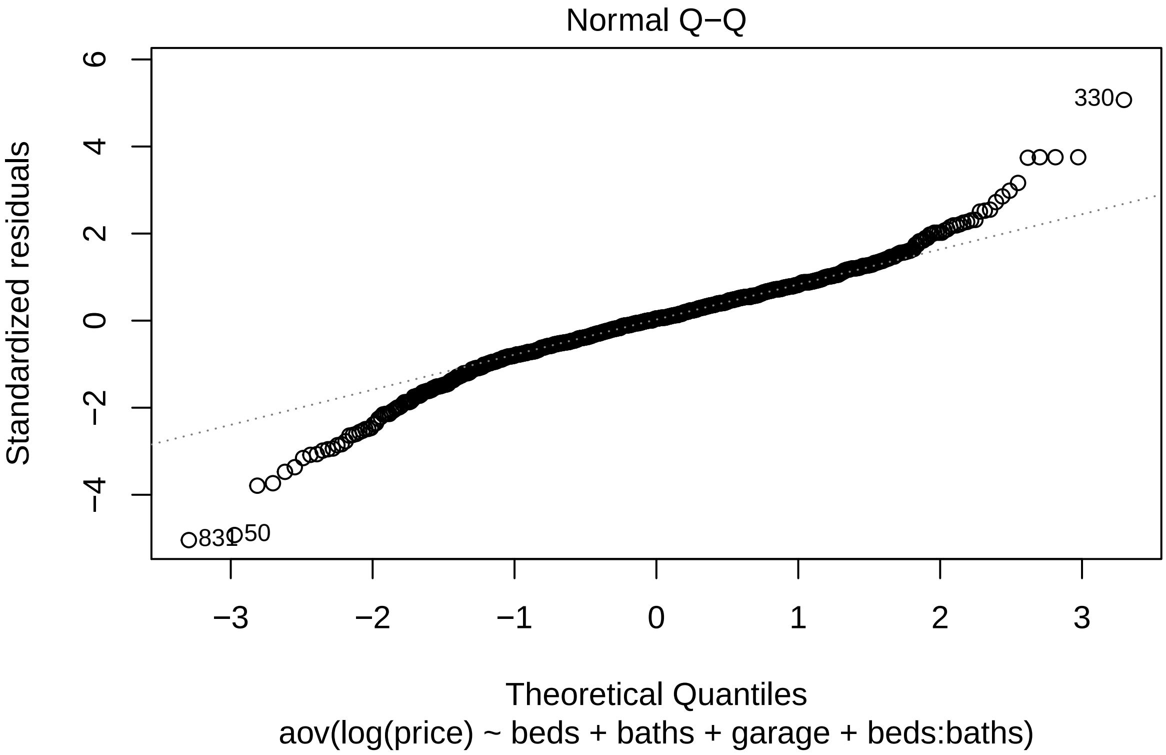
```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## beds          7  82.08  11.726 100.430  < 2e-16 ***
## baths        11  60.60   5.509  47.183  < 2e-16 ***
## garage        4  27.83   6.957  59.587  < 2e-16 ***
## beds:baths   13   4.13   0.317   2.718 0.000874 ***
## Residuals   986 115.13   0.117                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

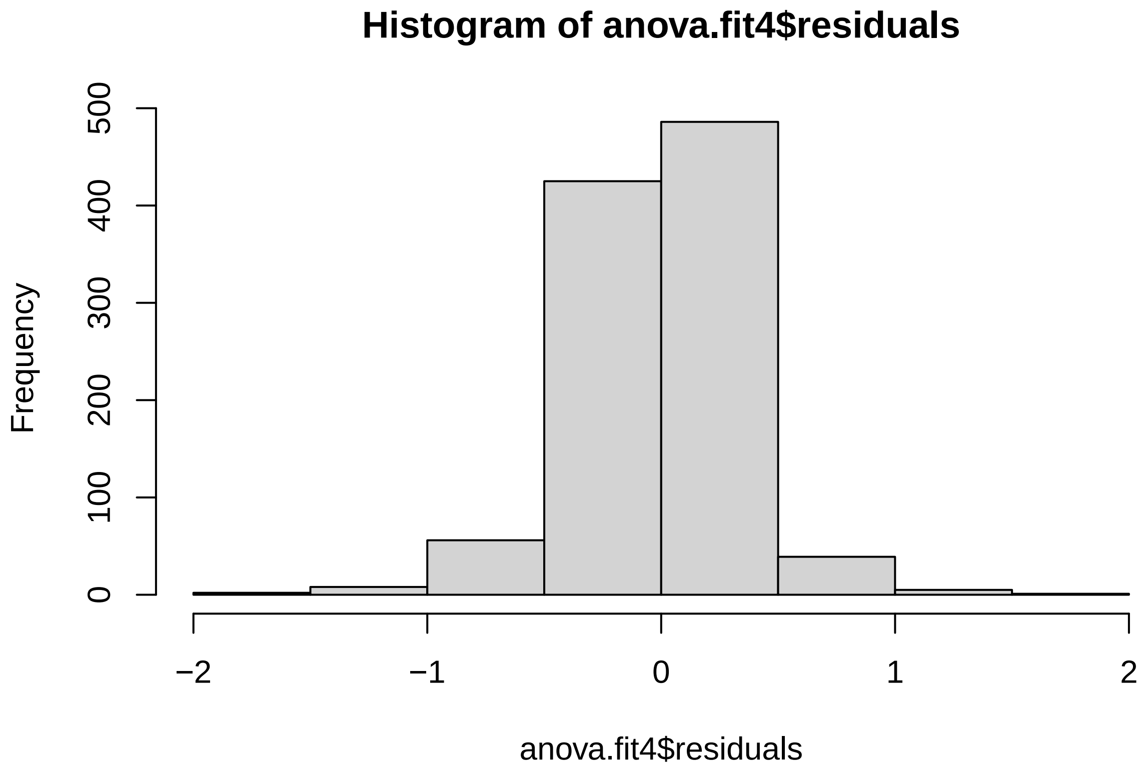
From the summary output, we find the the final ANOVA model contains only significant factors. These factors are `beds`, `baths`, `garage`, and the interaction effect `beds:baths`.

2.3.2 Final Model Diagnostics

We inspect the model diagnostics for the final ANOVA model. The ANOVA model assumes normally distributed errors with constant variance. To check for constant variance, we examine the fitted values vs residuals plot. The variance appears to be fairly constant, although it begins to decrease in variance as `log(price)` increases, a potential concern. To check for normally distributed errors, we can instead examine the normal Q-Q plot for the residuals. We find that the residuals are fairly normal, but has heavy left and right tails. This can be seen by plotting the residuals i na histogram and noting the heavy tails.


Residuals vs Fitted
Residuals
2
1
0
−1
−2
330
50
831
11.5
12.0
12.5
13.0
13.5
Fitted values
aov(log(price) ~ beds + baths + garage + beds:baths)


Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
330
831
50
−3
−2
−1
0
1
2
3
Theoretical Quantiles
aov(log(price) ~ beds + baths + garage + beds:baths)



3 Conclusions

We found, through exploratory data analysis, that there was reason to believe house price is determined by some of its key variables. It was decided to use linear regression on the sale price of houses because of indications of linear relationships between key variables. Through stepwise regression (using the BIC criterion), we found a model that balances bias and variance, having a high capability in prediction. In this final BIC model, we determined that the number of bedrooms, the number of bathrooms, the number of garages, and the house size (square feet), are significant in predicting the price of a house. In addition to this, after arriving at our final ANOVA model, we have found that house size is significantly determined by the number of bedrooms, the number of bathrooms, the number of garages, and the interaction effect between bedrooms and bathrooms. The inclusion of the interaction effect is plausible, as the pairwise scatterplot between the two indicates a linear relationship. The model diagnostics of both of these final models do not reveal any severe departure from the model assumptions of constant, normally distributed errors. It is important to note that the data used to determine these models is restricted to recent years of house sales within Gainesville, Florida. Because of this, the applicability of this model to general house prices can not be claimed. This may be a future question of interest to be explored: Model validation and applicability to house sale prices in various other cities.

4 Appendix

```
#Data Cleaning and Subsetting
library(ggplot2)
```

```r
library(MASS)
houses <- readRDS("houseSalesGNV2020.rds")

#Subset to remove NA values for necessary variables
sum(is.na(houses$lotsize))
houses = subset(houses,!is.na(houses$lotsize))
sum(is.na(houses$lotsize))

sum(is.na(houses$sqft))
houses = subset(houses,!is.na(houses$sqft))
sum(is.na(houses$sqft))

#Check for and Change NA values to 0 for variables than can be 0
sum(is.na(houses$garage))
houses$garage[is.na(houses$garage)] = 0
sum(is.na(houses$garage))

sum(is.na(houses$beds))
houses$beds[is.na(houses$beds)] = 0
sum(is.na(houses$beds))

sum(is.na(houses$baths))
houses$baths[is.na(houses$baths)] = 0
sum(is.na(houses$baths))

sum(is.na(houses$price))

sum(is.na(houses$lotunits))

#Convert lotsize to use same units
acre = which(houses$lotunits == "acres")
houses$lotsize[acre] = houses$lotsize[acre] * 43560
houses$lotunits[acre] = "sqft"

#Remove potential outliers in price
summary(houses$price)
houses <- subset(houses, houses$price > 5000 & houses$price < 2e6 & houses$lotsize > 300)




#Subset filtered data
houses_data = houses[,c("price","beds","baths","garage","sqft","lotsize")]

#Convert categorical variables to factors
houses$beds = as.factor(houses$beds)
houses$baths = as.factor(houses$baths)
houses$garage = as.factor(houses$garage)
```

```r
#Exploratory Data Analysis
#Summary Statistics and Variable Distribution
summary(houses$price)
boxplot(houses$price,col = "red",xlab="Boxplot of House Prices",ylab="Price")
boxplot(houses$price,col = "red",xlab="Boxplot of House Prices",ylab="Price")

summary(houses$lotsize)
boxplot(houses$lotsize,col="green",xlab="Boxplot of Lot Size (Sq. Ft.)")
houses_lotsize_subset = subset(houses,houses$lotsize < 150000)
boxplot(houses_lotsize_subset$lotsize,col="green",xlab="Boxplot of Lot Size (Sq. Ft.)")


beds = as.data.frame(summary(houses$beds))
rownames(beds) = c("0 Bd","1 Bd", "2 Bd","3 Bd","4 Bd","5 Bd","6 Bd","7 Bd")
colnames(beds) = "Count"
beds
#ggplot(houses,aes(x=as.integer(beds)-1)) + geom_histogram()
  #+ xlab("Number of Bedrooms") + ylab("Count")

baths = as.data.frame(summary(houses$baths))
rownames(baths) = c("0 Ba","1 Ba","1.5 Ba", "2 Ba",   "2.5 Ba", "3 Ba",   "3.5 Ba", "4 Ba"
                    , "4.5 Ba" ,"5 Ba",   "5.5 Ba", "6.5 Ba")
colnames(baths) = "Count"
baths

garages = as.data.frame(summary(houses$garage))
rownames(garages) = c("0 Garages","1 Garage", "2 Garages", "3 Garages", "4 Garages")
colnames(garages) = "Count"
garages

#pairs(houses_data)
#cor(houses_data)

panel.cor <- function(x, y) {
    # usr <- par('usr') on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- round(cor(x, y, use = "complete.obs"), 2)
    txt <- paste0("R = ", r)
    cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(houses_data, lower.panel = panel.cor)
```

```
ggplot(houses, aes(x=price)) + geom_histogram(color="white",fill="blue")
+ ggtitle("Prices of Houses (Gainesville, Florida)")

#ggplot(houses, aes(x=sqft,y=price)) + geom_point() + ggtitle("Square Feet vs Price for Houses")

boxplot(houses$price~houses$beds,main='Price by Number of Bedrooms',xlab='Beds'
        ,ylab='Price',col=rainbow(8))

boxplot(houses$price~houses$baths,main='Price by Number of Bathrooms',xlab='Baths'
        ,ylab='Price',col=rainbow(12))

boxplot(houses$price~houses$garage,main='Price by Number of Garages',xlab='Garage'
        ,ylab='Price',col=rainbow(12))
```

```
plot(houses$sqft,houses$price,pch=as.integer(houses$beds),
      col=as.integer(houses$beds),main='Beds Coded Scatterplot',
        ylab='Price',xlab='House Area (Square Feet)')
legend('bottomright',legend=c('0 Bds','1 Bds','2 Bds','3 Bds','4 Bds'
                              ,'5 Bds','6 Bds','7 Bds'),
        pch=c(1,2,3,4,5,6,7,8),col=c(1,2,3,4,5,6,7,8))

plot(houses$sqft,houses$price,pch=as.integer(houses$baths),
      col=as.integer(houses$baths),main='Baths Coded Scatterplot',
        ylab='Price',xlab='House Area (Square Feet)')
legend('bottomright',legend=c("0",'1 Ba','1.5 Ba','2 Ba','2.5 Ba','3 Ba','3.5 Ba'
                              ,'4 Ba','4.5 Ba','5 Ba','5.5 Ba','6.5 Ba'),
        pch=c(seq(1:12)),col=c(seq(1:12)))
```

```
#Linear Regression Model: beds, baths, garage, sqft, lotsize on price
none_mod = lm(price~1,data=houses)
full_mod = lm(price~beds+baths+garage+sqft+lotsize,data=houses)
BIC = stepAIC(none_mod, scope = list(upper = full_mod, lower = ~1), direction = "both",
    k = log(length(houses)), trace = FALSE)
summary(BIC)
plot(BIC)

bc.bic = boxcox(BIC)
bic.lambda = bc.bic$x[which.max(bc.bic$y)]
BIC2 = update(BIC,(price^bic.lambda-1)/bic.lambda~.,data=houses)
summary(BIC2)
plot(BIC2,1:2)
hist(BIC2$residuals)
```

```
#ANOVA model: beds, baths, garage as factors on sqft
anova.fit = aov(sqft~beds*baths*garage,data=houses)
summary(anova.fit)
anova.fit2 = update(anova.fit,.~.-beds:garage,data=houses)
summary(anova.fit2)
plot(anova.fit2)
bc.anova = boxcox(anova.fit2)
anova.fit3 = update(anova.fit2,log(price)~.,data=houses)
summary(anova.fit3)
anova.fit4 = update(anova.fit3,log(price)~.-baths:garage-beds:baths:garage,data=houses)
summary(anova.fit4)
plot(anova.fit4)
```