Statistical Report on Home Prices and the Predictors

Chad Huntebrinker

2024-12-15

# Title

Statistical Report on Home Prices and the Predictors.

# Abstract

These past couple of years, the United States housing market has gotten more and more expensive. Goldman’s Sach’s reports that housing prices will rise by 4.5% this year and 4.4% the following year which is an increase from the previous estimation of 4.2% and 3.2%. They also reported that affordability for these houses are at an all time low since they started collecting that data in the 1980s [1]. And while there is some hope with lowering prices with Forbes stating there was a decrease in August this year in monthly prices by 0.13% (the first decrease in August since 2022), the majority of new home buyers are struggling to find houses for an affordable price [2]. CNN reported that only 2.5% of houses in the United States actually changed hands this year in the first 8 months; that is the lowest turnover rate in the past 30 years [3]. What would be beneficial to these buyers would be figuring out what factors of real estate have the biggest impact on the price of a house. So theis leads to the following question: what parts of a house have the most influence on the price? Is it the number of bedrooms, the number of acres of land it’s on, or maybe the location? The analysis focuses on determining how these factors impact the price of homes. By exploring the relationships between these variables, the report aims to provide valuable insights for home-buyers when making their next purchase.

# Introduction

To consider what impacts the price of homes, a wide range of variables need to be considered. Variables like the number of beds, number of baths, number of acres the house is on, and the square footage of the house itself. There’s also variables like zip code, city, and state where the house is located as well. With these variables, this report aims to investigate whether price of a home can be predicted.

## Primary Analysis Objectives

The Primary Analysis Objective is to investigate whether there is a linear association between price of a house and variables related to the house to see if accurate predictions can be made when given a set of values for these variables.

# Methods

## Data Sources

The dataset has 12 variables (one of them being the dependent variable, price). It also has quantitative (number of bedrooms, number of bathrooms, lot size in acres, etc.) and categorical (housing status, city, state, etc.) predictor variables. Data was collected from realtor.com. It’s a real estate listing website operated by the News Corp subsidiary Move, Inc. and based in Santa Clara, California. It is the second most visited real estate listing website in the United States as of 2024, with over 100 million monthly active users. The data was collected via web scraping using python libraries. The data was originally collected November, 2023.

## Statistical Analysis

The data is available in .csv (CSV) format. The data analysis is completed using the software R (version 4.4.1) and this project focuses on multiple linear regression. The following variables were not included when creating the model: brokered\_by, status, street, city, state, and prev\_sold\_date. Removal of borkered\_by, prev\_sold\_date, and status was due to them having no impact on the price of a house. The removal of street, city, and state was done to help simplify the overall model as it’s difficult to predict the price of a house by all the different cities and streets. And a new variable was created: expensive\_state. The purpose of this variable is to see whether the house is in one of the 25 most expensive states in the USA or the least 25 expensive states (according to NADAQ [4]). This investigation also removed the capital city or any US territories from the dataset (Puerto Rico, Virgin Islands, Guam and Washington D.C.). Missing values were found in the dataset and were removed from the data used. The data was originally cleaned to remove outliers from the beginning. For example, in 2021, the average square footage of houses was 2,480 square feet [5]. So the report only included houses with a square footage between 0 and 4,000. Also, the the national average lot size is 25,240 square feet or just over half an acre [3]. So, the report only included houses on lot sizes between 0 to 1 acre. In a similar way, number of beds included were between 0 and 10, number of baths were between 0 and 10, and home prices were between $100,000 and $1,000,000. This gave a data\_frame with over 900,000 results. This was narrowed down to 50 houses from each state randomly selected while also setting the seed so the same houses would be selected every time. Automatic model selection methods were used to arrive to the final model and the model assumptions are compared and confirmed while reviewing and using the final model.

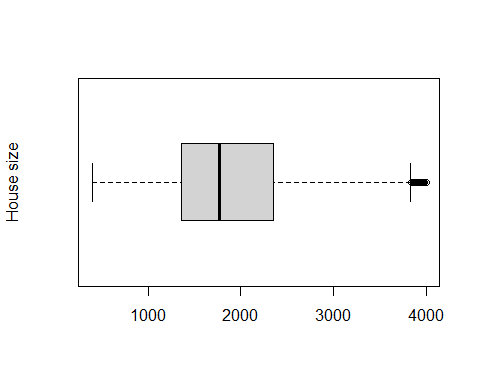
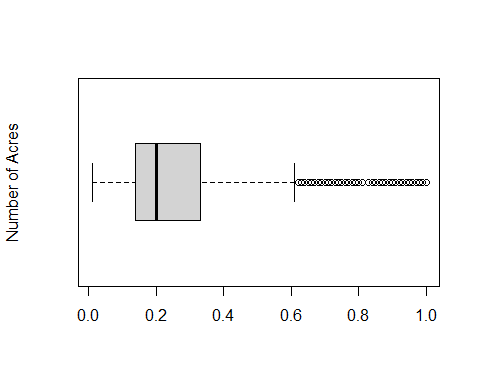
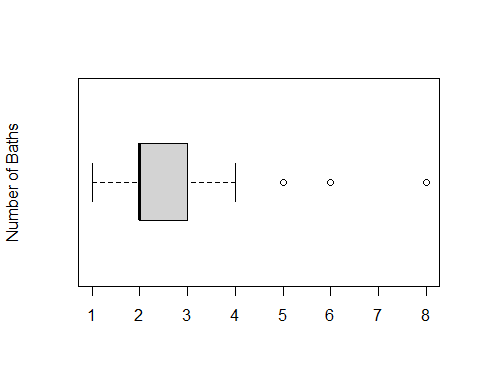
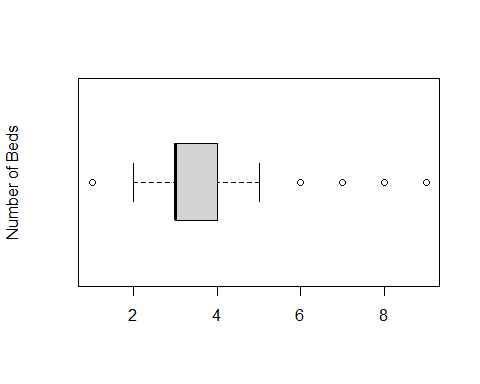
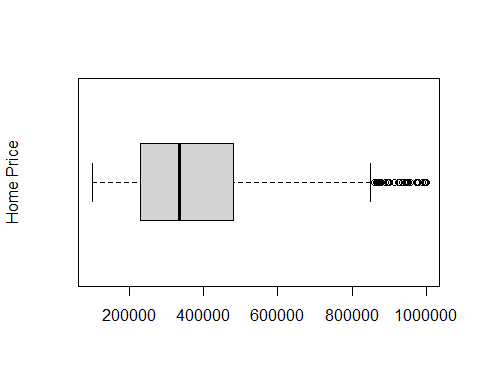
## Model Assumtions

All interfences are done using a significance value of 0.05.

## Primary Objective Analysis

First, this report will investigate the individual predictors and the response variable to verify the skewness and identify any outliers. It will then complete a check for multicollinearity before utilizing automatic model selection to find the most effective predictor variables to include in the model. After that, the report will complete other checks for this final model (like homorskedasticity, normality of the errors, and independence of error terms) before seeing how accurate the model predicts.

### Individual Predictor and Response Variable Investigation



The first step is to review the boxplot for the home price. While the overall spread is pretty good, there are a large amount of outliers on the far right side. This will need to be checked on later on when the model is made. Number of beds and baths had a couple of outliers to the right, but overall seemed okay. Finally, the house size and the number of acres had a large mount of outliers to the right (once again, another thing to check the impact on the model later).

### Multicollinearity

## price bed bath acre\_lot zip\_code  
## price 1.00000000 0.25349589 0.411879772 -0.01780871 0.23121384  
## bed 0.25349589 1.00000000 0.515476680 0.09851121 -0.04929462  
## bath 0.41187977 0.51547668 1.000000000 0.03731894 -0.02201731  
## acre\_lot -0.01780871 0.09851121 0.037318941 1.00000000 -0.15047424  
## zip\_code 0.23121384 -0.04929462 -0.022017309 -0.15047424 1.00000000  
## house\_size 0.42725443 0.62334652 0.695858271 0.17099349 -0.01420820  
## expensive\_state 0.25297310 -0.01123575 -0.002511952 -0.12385574 0.08827275  
## house\_size expensive\_state  
## price 0.4272544 0.252973095  
## bed 0.6233465 -0.011235749  
## bath 0.6958583 -0.002511952  
## acre\_lot 0.1709935 -0.123855744  
## zip\_code -0.0142082 0.088272745  
## house\_size 1.0000000 -0.053341897  
## expensive\_state -0.0533419 1.000000000

The next step is to look at both the relationship between price of the house and all the predictor variables along with the multicolineratiy between the predictor variables. First, there seems to be no or very little multicolineratiy between the predictor variables. There might be some correlation between the house size and the number of beds and the number baths (which would make sense as bigger houses usually have more bedrooms and bathrooms) but it doesn’t seem to be of too much concern. When it comes to the relationship between price and the predictor variables, there may be a weak correlation between price and number of baths or the house size; but overall, there doesn’t seem to be much of a linear relationship between the variables and price individually.

### Automatic Model Selection

## [1] 631.856419 271.918428 136.490845 31.283140 7.929952 7.000000

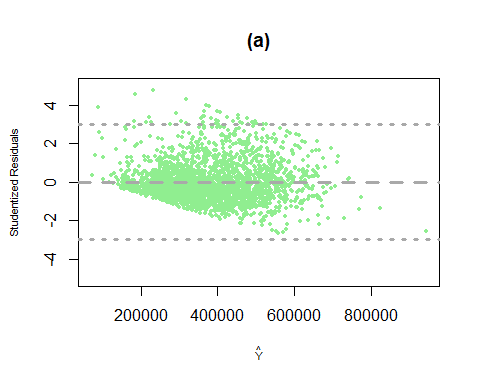
## [1] 0.2048054 0.2968209 0.3317591 0.3590145 0.3654600 0.3662049

## [1] -557.2730 -856.8869 -976.4701 -1072.7508 -1090.1931 -1085.3056

## (Intercept) bed bath acre\_lot   
## TRUE FALSE TRUE FALSE   
## house\_size zip\_code expensive\_stateTRUE   
## TRUE TRUE TRUE

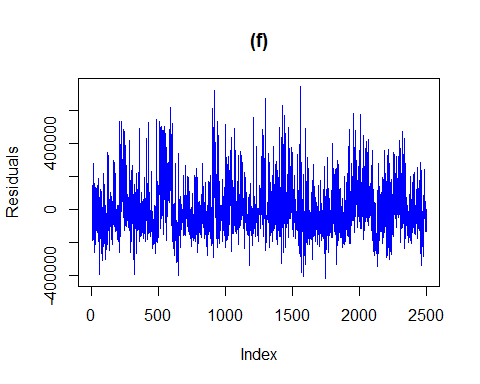
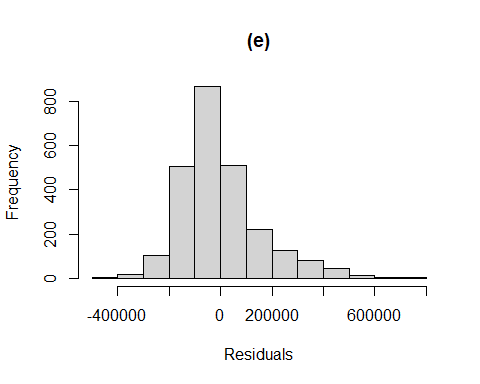
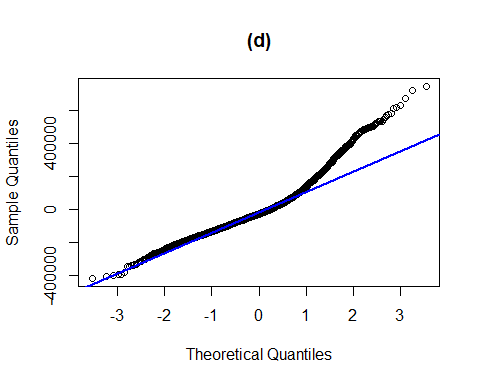
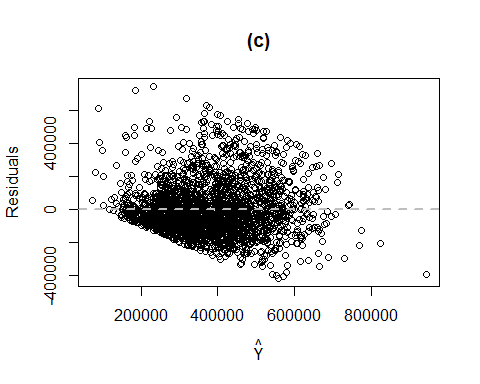
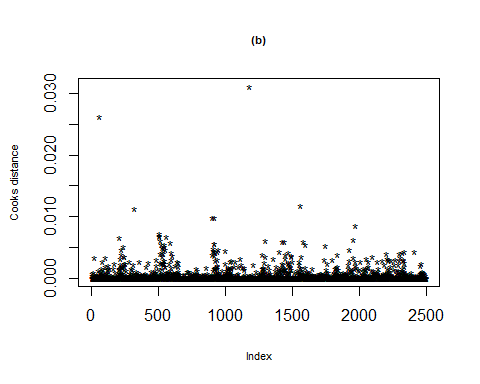
The next step is to find a model via automatic variable selection. This report utilized the “regsubsets” function from the “leap” package in R. This report also utilizes Mallow’s Cp, adjusted R^2, and BIC. Both for Cp and BIC, lower values mean a better model, while a higher adjusted R^2 means a better model. When looking through the scores, this investigation found that the fourth model had the best results. The jump from the fourth model’s Cp score to the fifth model’s Cp is less than the 3rd model’s to the fourth model’s. It’s the same case with the adjusted R^2 score. And with the fourth model’s BIC score being the best, there can be confidence in saying the fourth model should be selected as the final model. After looking at what the fourth model includes, it was found that including number of baths, the house size, the zip code, and whether or not the house was in an expensive state along with the X-intercept would lead to the best model.

### Final Checks on the Model



## [1] "Value to use for Cook's Distance:"

## [1] 0.5660568



## [1] TRUE

## [1] TRUE

## [1] TRUE

## [1] TRUE

## fit lwr upr  
## 1 215319.1 -91886.88 522525.1

## [1] "Actual value: 169900"

## fit lwr upr  
## 1 356716 49564 663868

## [1] "Actual value: 184900"

## fit lwr upr  
## 1 444594.2 137394.7 751793.7

## [1] "Actual value: 513990"

## fit lwr upr  
## 1 356755.6 49234.01 664277.2

## [1] "Actual value: 279000"

## fit lwr upr  
## 1 420842.3 113544.2 728140.3

## [1] "Actual value: 369000"

## fit lwr upr  
## 1 303206.7 -3969.224 610382.6

## [1] "Actual value: 299900"

The first thing the report checked with this new model is the residuals. A graph is created for the studentized residuals (graph a) and Cook’s Distance (graph b) to see if any data points have a huge influence on the model. After looking for any points within a range of ±3 for graph a, the report found some that were above +3. However, when looking for any points with a yellow mark (indicating they’re outliers) in graph b with Cook’s Distance, no such points were found. As a result, the decision made was to continue with the report and include all the data points that were previously there when the model was created.

The next checks this report looks at is the equal variance (graph c), the normaility of the errors (graph d and e), and the independence of the residuals (graph f). All of these graphs indicate that everything is okay. The investigation did find a heavy right tail on graph d, but continued with the model investigation.

The last checks required are the goodness of fit of the model and the accuracy. First, the p-test is done on the four different coefficients. A comparision between their p-value and 0.05 is done and it’s found that each of their values are less than or equal to 0.05. Thus, confirming the goodness of fit with the model. The final check this investigation does is with predictions. The model tries to predict what the price of a house by giving data that’s in the data. What’s found is most of the predictions are decently accurate to the original values (off by $50,000 or less) and they all are within the upper and lower bounds. As a result, it seems to indicate decent accuracy of the final model.

# General Discussion and Conclusions

So the final estimated regression function from this investigation would be:

## Ŷ = -19575.7 + 49710.31 X1 + 83.81 X2 + 1.18 X3 + 122549 X4

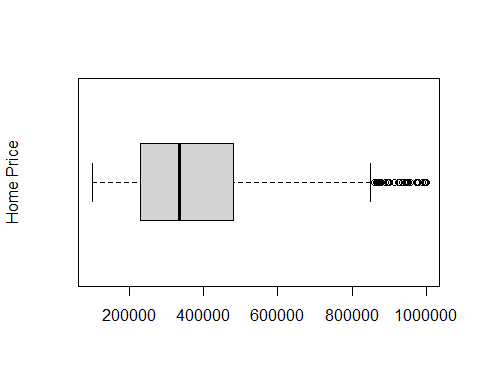
Where Ŷ is the price of a house, X1 represents the number of baths, X2 is the house size, X3 is the zip code where the house is located, and X4 indicates whether the house is in an expensive state or not. The report found that a multiple linear regression model does a decent job explaining the relationship between the predictor variables and the price. The report also looked at different tests and checks like residual plots, Cook’s Distance, and homorskedasticity to verify the model is both accurate and reliable. The model also predicted some of the values in the data and was found to be decently accurate as well. Overall, the model will need some work to better predict the prices of houses. But this model does a decent job of predicting that right now; so homeowners can use it to help assist them when buying a house.

# Appendix: R-code

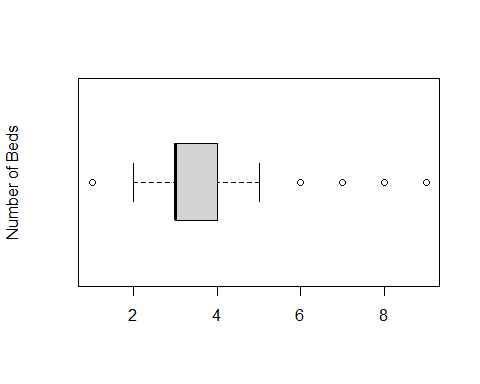
#Chad Huntebrinker  
library(readxl)  
library(leaps)  
library(psych)  
library(lmtest)  
library(dplyr)  
  
#Load the library and data in to R  
excel\_data <- read.csv("realtor-data.zip.csv")  
  
#Remove NA values  
realtor\_data\_no\_na <- na.omit(excel\_data)  
  
#Since we only want the 50 states, we need to remove Puerto Rico, Virgin Islands, Guam and Washington D.C.  
#So we will have all 50 states.  
  
realtor\_data\_no\_na <- realtor\_data\_no\_na[realtor\_data\_no\_na$state != "Puerto Rico", ]  
realtor\_data\_no\_na <- realtor\_data\_no\_na[realtor\_data\_no\_na$state != "Virgin Islands", ]  
realtor\_data\_no\_na <- realtor\_data\_no\_na[realtor\_data\_no\_na$state != "Guam", ]  
realtor\_data\_no\_na <- realtor\_data\_no\_na[realtor\_data\_no\_na$state != "District of Columbia", ]  
  
unique(realtor\_data\_no\_na$state)

## [1] "Massachusetts" "Connecticut" "New Jersey" "New York"   
## [5] "New Hampshire" "Vermont" "Rhode Island" "Wyoming"   
## [9] "Maine" "Pennsylvania" "West Virginia" "Delaware"   
## [13] "Ohio" "Maryland" "Virginia" "Colorado"   
## [17] "North Carolina" "Kentucky" "South Carolina" "Tennessee"   
## [21] "Georgia" "Alabama" "Florida" "Mississippi"   
## [25] "Texas" "Missouri" "Arkansas" "Louisiana"   
## [29] "Indiana" "Illinois" "Michigan" "Wisconsin"   
## [33] "Iowa" "Minnesota" "South Dakota" "Nebraska"   
## [37] "North Dakota" "Montana" "Idaho" "Kansas"   
## [41] "Oklahoma" "New Mexico" "Utah" "Nevada"   
## [45] "Washington" "Oregon" "Arizona" "California"   
## [49] "Hawaii" "Alaska"

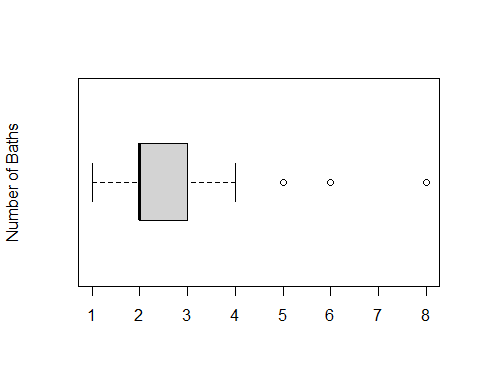
#Remove duplicates  
realtor\_data\_no\_na <- realtor\_data\_no\_na[!duplicated(realtor\_data\_no\_na), ]  
data\_frame\_1 <- realtor\_data\_no\_na  
  
#Remove outliers  
data\_frame\_2 <- subset(data\_frame\_1, data\_frame\_1$price >= 100000 & data\_frame\_1$price <= 1000000)  
data\_frame\_2 <- subset(data\_frame\_2, data\_frame\_2$bed >= 1 & data\_frame\_2$bed <= 10)  
data\_frame\_2 <- subset(data\_frame\_2, data\_frame\_2$bath >= 1 & data\_frame\_2$bath <= 10)  
data\_frame\_2 <- subset(data\_frame\_2, data\_frame\_2$acre\_lot >= 0 & data\_frame\_2$acre\_lot <= 1)  
data\_frame\_2 <- subset(data\_frame\_2, data\_frame\_2$house\_size >= 0 & data\_frame\_2$house\_size <= 4000)  
  
#Most and least expensive states to live in:  
most\_expensive\_states <- c("Pennsylvania", "Delaware", "Utah", "Alaska", "North Carolina", "Arizona",  
 "Vermont", "Texas", "New Hampshire", "Florida", "Georgia", "Illinois", "Virgina",  
 "Maryland", "Colorado", "Nevada", "Washington", "Oregon", "Rhode Island",  
 "Massachusetts", "Connecticut", "Hawaii", "New Jersey", "New York", "California")  
  
#Add variable expensive\_state to see if it is a 25 most expensive state to live in or not  
data\_frame\_2$expensive\_state <- NA  
data\_frame\_2$expensive\_state <- ifelse(data\_frame\_2$state %in% most\_expensive\_states, TRUE, FALSE)  
  
#Take a sample of the data by tampling 50 houses per state.  
set.seed(1)  
sampled\_df <- data\_frame\_2 %>%  
 group\_by(state) %>%  
 sample\_n(50) %>%  
 ungroup()  
  
#Box plots of some of the variables to see if there are any outliers  
options(scipen = 999)  
boxplot(sampled\_df$price, ylab = "Home Price", horizontal = TRUE)



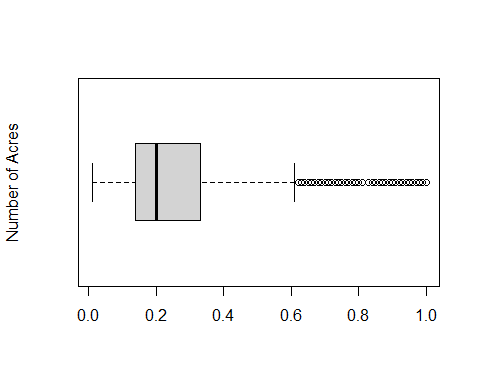
boxplot(sampled\_df$bed, ylab = "Number of Beds", horizontal = TRUE)



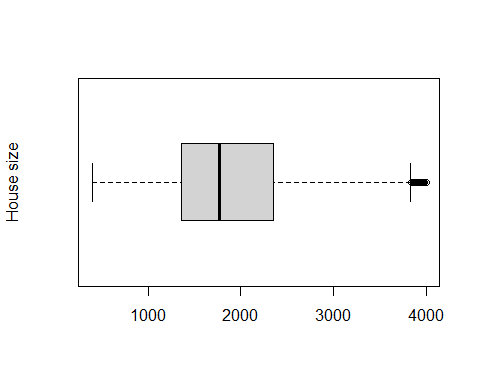
boxplot(sampled\_df$bath, ylab = "Number of Baths", horizontal = TRUE)



boxplot(sampled\_df$acre\_lot, ylab = "Number of Acres", horizontal = TRUE)



boxplot(sampled\_df$house\_size, ylab = "House size", horizontal = TRUE)



#Multicollinearity  
#Remove the following predictor variables: brokered\_by, status, street, city, state, and prev\_sold\_date  
temp <- c(1, 2, 7, 8, 9, 12)  
cor(data\_frame\_2[,-temp])

## price bed bath acre\_lot zip\_code  
## price 1.00000000 0.25349589 0.411879772 -0.01780871 0.23121384  
## bed 0.25349589 1.00000000 0.515476680 0.09851121 -0.04929462  
## bath 0.41187977 0.51547668 1.000000000 0.03731894 -0.02201731  
## acre\_lot -0.01780871 0.09851121 0.037318941 1.00000000 -0.15047424  
## zip\_code 0.23121384 -0.04929462 -0.022017309 -0.15047424 1.00000000  
## house\_size 0.42725443 0.62334652 0.695858271 0.17099349 -0.01420820  
## expensive\_state 0.25297310 -0.01123575 -0.002511952 -0.12385574 0.08827275  
## house\_size expensive\_state  
## price 0.4272544 0.252973095  
## bed 0.6233465 -0.011235749  
## bath 0.6958583 -0.002511952  
## acre\_lot 0.1709935 -0.123855744  
## zip\_code -0.0142082 0.088272745  
## house\_size 1.0000000 -0.053341897  
## expensive\_state -0.0533419 1.000000000

#Check to see what is the best model  
#Remove the following predictor variables: brokered\_by, status, street, city, state, and prev\_sold\_date  
mr <- regsubsets(price~bed + bath + acre\_lot + house\_size + zip\_code + expensive\_state,  
 data=sampled\_df)  
sum\_of\_mr <- summary(mr)  
  
#Check Cp, R^2, and BIC  
sum\_of\_mr$cp

## [1] 631.856419 271.918428 136.490845 31.283140 7.929952 7.000000

sum\_of\_mr$rsq

## [1] 0.2048054 0.2968209 0.3317591 0.3590145 0.3654600 0.3662049

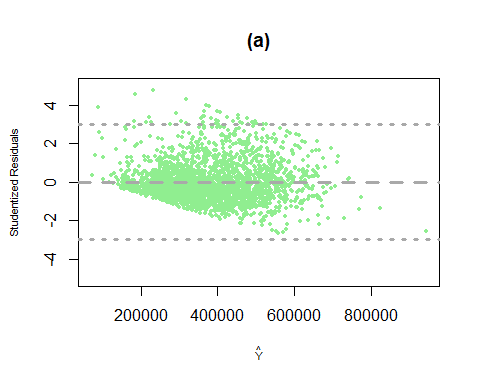
sum\_of\_mr$bic

## [1] -557.2730 -856.8869 -976.4701 -1072.7508 -1090.1931 -1085.3056

#Model 4 is our best model  
sum\_of\_mr$which[4,]

## (Intercept) bed bath acre\_lot   
## TRUE FALSE TRUE FALSE   
## house\_size zip\_code expensive\_stateTRUE   
## TRUE TRUE TRUE

#Thus, we will include bath + house\_size + zip\_code + expensive\_state  
  
final\_model <- lm(price~bath + house\_size + zip\_code + expensive\_state, data = sampled\_df)  
  
plot(fitted(final\_model), rstudent(final\_model),  
 col = "lightgreen", pch = 16,  
 xlab = expression(hat(Y)),  
 ylab = "Studentized Residuals",  
 ylim = c(-5, 5), main = "(a)",  
 cex.lab = 0.7, cex = 0.5)  
abline(h = 0, lty = 2, lwd = 3,  
 col = "darkgray")  
abline(h = 3, lty = 3, lwd = 3,  
 col = "darkgray")  
abline(h = -3, lty = 3, lwd = 3,  
 col = "darkgray")



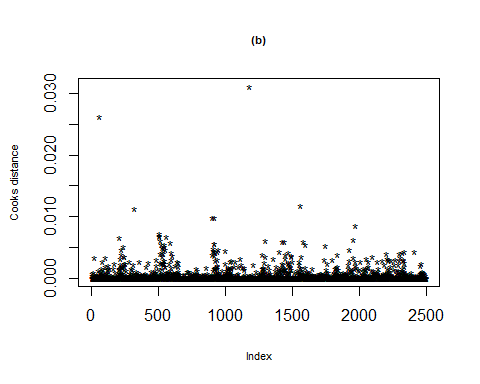
#Cook's Distance  
print("Value to use for Cook's Distance:")

## [1] "Value to use for Cook's Distance:"

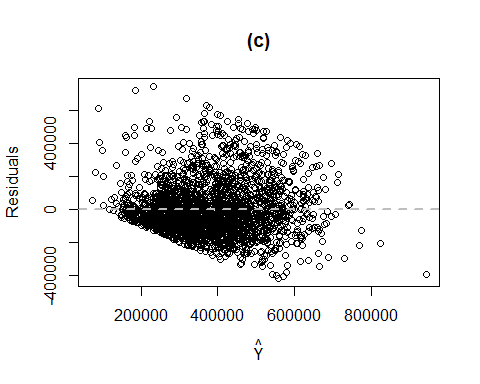
pf(0.95, df1 = 4, df2 = 2500)

## [1] 0.5660568

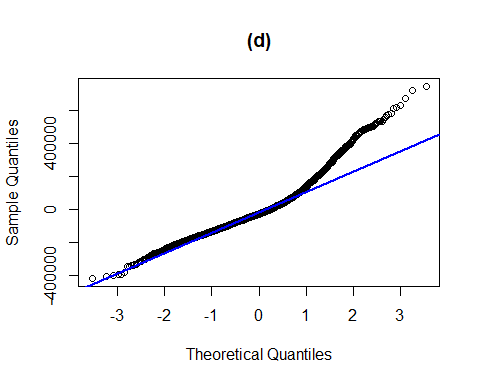
cooks\_distance <- cooks.distance(final\_model)  
plot(cooks\_distance, pch = "\*", cex.main = 0.7, main = "(b)",  
 ylab = "Cooks distance", cex.lab = 0.7)  
text(x = 1:length(cooks\_distance) + 1, y = cooks\_distance, labels = ifelse(cooks\_distance > 0.5660568,  
 names(cooks\_distance), ""),col = "yellow")



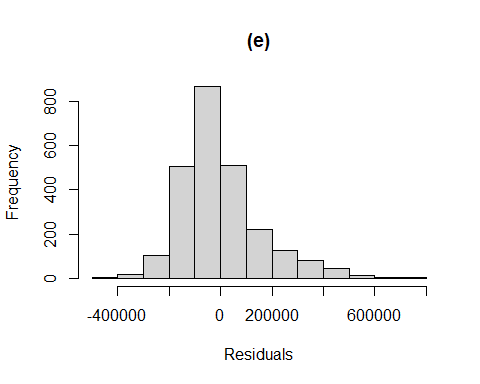
#Homorskedasticity  
plot(fitted(final\_model), residuals(final\_model), main = "(c)",  
 xlab = expression(hat(Y)), ylab = "Residuals")  
abline(h = 0, col = "gray", lwd = 2,lty = 2)



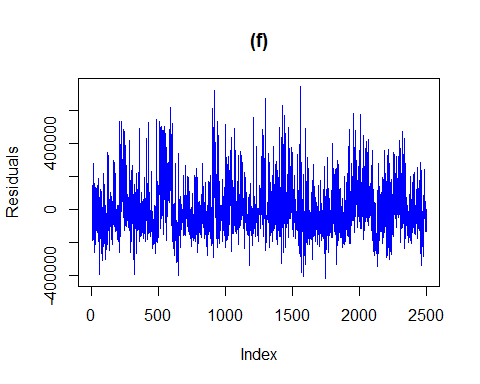
#Normality  
qqnorm(residuals(final\_model), main = "(d)")  
qqline(residuals(final\_model), col = "blue", lwd = 2)



hist(residuals(final\_model), main = "(e)", xlab = "Residuals")



#Independence of error terms  
plot(residuals(final\_model), type = "l", col = "blue", main = "(f)", ylab = "Residuals")



#Check p-score  
p\_score <- coeftest(final\_model)  
  
abs(p\_score[2,4]) <= 0.05

## [1] TRUE

abs(p\_score[3,4]) <= 0.05

## [1] TRUE

abs(p\_score[4,4]) <= 0.05

## [1] TRUE

abs(p\_score[5,4]) <= 0.05

## [1] TRUE

#Validate the model  
predict(final\_model, data.frame(bath = 2, house\_size = 1124,  
 zip\_code = 35055, expensive\_state = FALSE),  
 level = 0.95, interval = "prediction")

## fit lwr upr  
## 1 215319.1 -91886.88 522525.1

print("Actual value: 169900")

## [1] "Actual value: 169900"

predict(final\_model, data.frame(bath = 2, house\_size = 1398,  
 zip\_code = 31558, expensive\_state = TRUE),  
 level = 0.95, interval = "prediction")

## fit lwr upr  
## 1 356716 49564 663868

print("Actual value: 184900")

## [1] "Actual value: 184900"

predict(final\_model, data.frame(bath = 3, house\_size = 1993,  
 zip\_code = 21619, expensive\_state = TRUE),  
 level = 0.95, interval = "prediction")

## fit lwr upr  
## 1 444594.2 137394.7 751793.7

print("Actual value: 513990")

## [1] "Actual value: 513990"

predict(final\_model, data.frame(bath = 3, house\_size = 1124,  
 zip\_code = 8873, expensive\_state = TRUE),  
 level = 0.95, interval = "prediction")

## fit lwr upr  
## 1 356755.6 49234.01 664277.2

print("Actual value: 279000")

## [1] "Actual value: 279000"

predict(final\_model, data.frame(bath = 3, house\_size = 3066,  
 zip\_code = 29150, expensive\_state = FALSE),  
 level = 0.95, interval = "prediction")

## fit lwr upr  
## 1 420842.3 113544.2 728140.3

print("Actual value: 369000")

## [1] "Actual value: 369000"

predict(final\_model, data.frame(bath = 2, house\_size = 1500,  
 zip\_code = 82941, expensive\_state = FALSE),  
 level = 0.95, interval = "prediction")

## fit lwr upr  
## 1 303206.7 -3969.224 610382.6

print("Actual value: 299900")

## [1] "Actual value: 299900"

# Reference

[1] US house prices are forecast to rise more than 4% next year. Goldman Sachs. (2024, September 11). <https://www.goldmansachs.com/insights/articles/us-house-prices-are-forecast-to-rise-more-than-4-percent-next-year>

[2] Forbes Magazine. (2024, November 8). Housing market predictions for 2025: When will home prices drop?. Forbes. <https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/>

[3] Delouya, S. (2024, September 30). A new report illustrates just how stuck the housing market is | CNN business. CNN. <https://www.cnn.com/2024/09/30/economy/housing-market-home-sales-redfin-report/index.html>

[4] Borrelli, L. (2024, January 10). What’s the average square footage of a house? Bankrate. <https://www.bankrate.com/real-estate/average-square-feet-of-a-house/>

[5] Rawlings, T. (2024, February 9). How much land do you need to build a home?. Build On Your Land. <https://buildonyourlandllc.com/how-much-land-do-you-need-to-build-a-home/#>:~:text=The%20size%20of%20a%20lot,you’re%20looking%20to%20build.