Paris Housing
Exploratory Analysis

Name, smahmood@bellarmine.edu
Name, jbuie@bellarmine.edu

I. DATA SET DESCRIPTION

We found the dataset from Kaggle. It has a total of 17 variables. Our dataset contained no missing values. All the variables were integer while our Y variable “price,” being a double class (double-precision floating point number). A visualization of the various variables with their corresponding data types can be referenced in Table 1.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
squareMeters	*Int, Ratio*	*0%*
numberOfRooms	*Int, Ratio*	*0%*
hasYards	*Int, Nominal*	*0%*
hasPools	*Int, Nominal*	*0%*
floors	*Int, Ratio*	*0%*
cityCode	*Int, Nominal*	*0%*
cityPartRange	*Int, Ratio*	*0%*
numPrevOwners	*Int, Ratio*	*0%*
made	*Int, Interval*	*0%*
isNewBuilt	*Int, Nominal*	*0%*
hasStormProtector	*Int, Nominal*	*0%*
basement	*Int, Ratio*	*0%*
attic	*Int, Ratio*	*0%*
garage	*Int, Ratio*	*0%*
hasStorageRoom	*Int, Nominal*	*0%*
hasGuestRoom	*Int, Ratio*	*0%*
price	*dbl, Interval*	*0%*

II. EDA FINDINGS

The data seemed to not be very distributed which can be seen in the standard deviations for each of the variables within Table 2.

Table 2: Summary Statistics for Paris Housing

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
squareMeters	*10000*	*49,870*	*2.877438e+04*	*89*	*25,098*	*50,106*	*74,610*	*99,999*
numberOfRooms	*10000*	*50.36*	*2.881670e+01*	*1*	*25*	*50*	*75*	*100*
hasYard	*10000*	*.5087*	*4.999493e-01*	*0*	*0*	*1*	*1*	*1*
hasPool	*10000*	*0.4968*	*5.000148e-01*	*0*	*0*	*0*	*1*	*1*
floors	*10000*	*50.28*	*2.888917e+01*	*1*	*25*	*50*	*76*	*100*
cityCode	*10000*	*50225*	*2.900668e+04*	*3*	*24694*	*50693*	*75683*	*99,953*
cityPartRange	*10000*	*5.51*	*2.872024e+00*	*1*	*3*	*5*	*8*	*10*
numPrevOwners	*10000*	*5.522*	*2.856667e+00*	*1*	*3*	*5*	*8*	*10*
made	*10000*	*2005*	*9.308090e+00*	*1990*	*1997*	*2006*	*2014*	*2021*

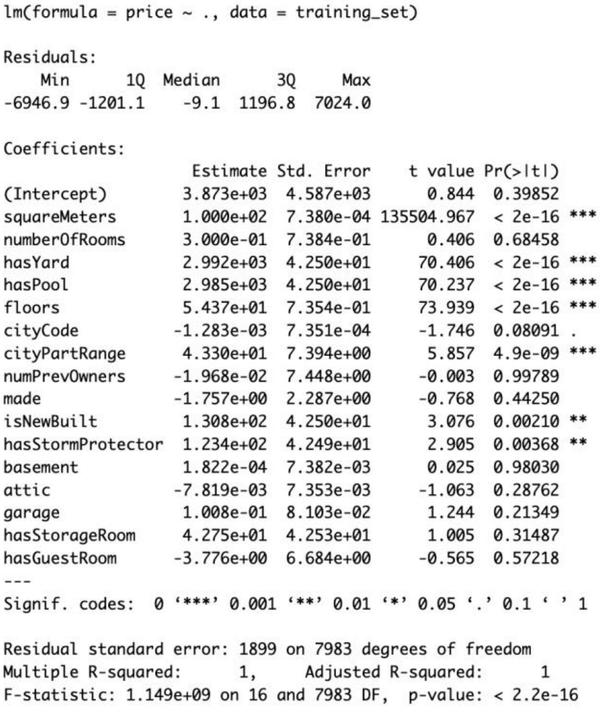
isNewBuilt	*10000*	*.4991*	*5.000242e-01*	*0*	*0*	*0*	*1*	*1*
hasStormProtector	*10000*	*.4999*	*5.000250e-01*	*0*	*0*	*0*	*1*	*1*
basement	*10000*	*5033*	*2.876730e+03*	*0*	*2560*	*5092*	*7511*	*10000*
attic	*10000*	*5028*	*2.894332e+03*	*1*	*2512*	*5045*	*7540*	*10000*
garage	*10000*	*553.1*	*2.620502e+02*	*100*	*327.8*	*554*	*777.2*	*1000*
hasStorageRoom	*10000*	*.503*	*5.000160e-01*	*0*	*0*	*1*	*1*	*1*
hasGuestRoom	*10000*	*4.995*	*3.176410e+00*	*0*	*2*	*5*	*8*	*10*
price	*10000*	*4993448*	*2.877424e+06*	*10314*	*2516402*	*5016180*	*7469092*	*10006771*

We focused on the relationship between the price of a house in Paris and its size in square meters. Through this, we found that larger houses tended to be more expensive. This can also be referenced in Figure 15. Based on this, we decided to filter our data to look at the top 50% largest homes in Paris. This resulted in us confirming that bigger houses were often higher in price. This can be referenced in Figure 16.

Next, we looked at the relationship between the price and whether the home has a pool and a yard. We found that homes with both a pool and a yard tended to be more expensive compared to homes that did not have a pool and yard.

Then, we focused on looking at the relationship between newly built homes and their prices. We found that homes that are newly built tended to be more expensive in comparison with houses that were older.

III. REGRESSION ANALYSIS



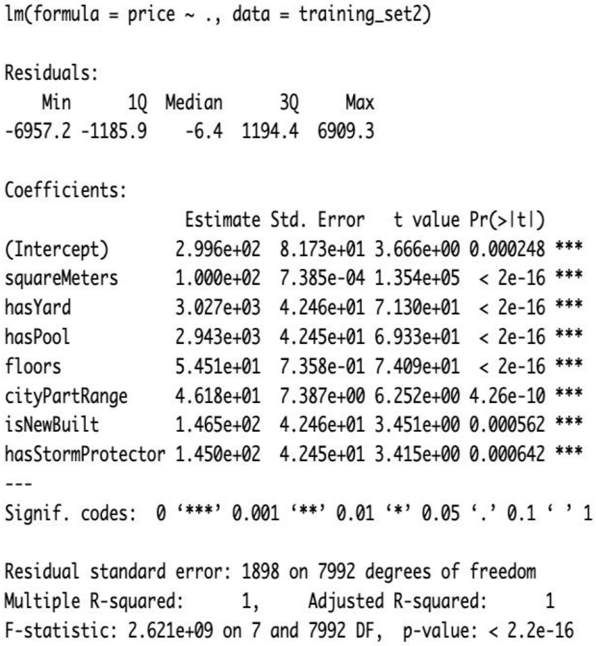
```
lm(formula = price ~ ., data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max 
-6946.9 -1201.1    -9.1  1196.8  7024.0 

Coefficients:
                    Estimate Std. Error    t value Pr(>|t|)    
(Intercept)        3.873e+03  4.587e+03      0.844  0.39852    
squareMeters       1.000e+02  7.380e-04 135504.967  < 2e-16 ***
numberOfRooms      3.000e-01  7.384e-01      0.406  0.68458    
hasYard            2.992e+03  4.250e+01     70.406  < 2e-16 ***
hasPool            2.985e+03  4.250e+01     70.237  < 2e-16 ***
floors             5.437e+01  7.354e-01     73.939  < 2e-16 ***
cityCode          -1.283e-03  7.351e-04     -1.746  0.08091 .  
cityPartRange      4.330e+01  7.394e+00      5.857  4.9e-09 ***
numPrevOwners     -1.968e-02  7.448e+00     -0.003  0.99789    
made              -1.757e+00  2.287e+00     -0.768  0.44250    
isNewBuilt         1.308e+02  4.250e+01      3.076  0.00210 ** 
hasStormProtector  1.234e+02  4.249e+01      2.905  0.00368 ** 
basement           1.822e-04  7.382e-03      0.025  0.98030    
attic             -7.819e-03  7.353e-03     -1.063  0.28762    
garage             1.008e-01  8.103e-02      1.244  0.21349    
hasStorageRoom     4.275e+01  4.253e+01      1.005  0.31487    
hasGuestRoom      -3.776e+00  6.684e+00     -0.565  0.57218    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1899 on 7983 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 1.149e+09 on 16 and 7983 DF,  p-value: < 2.2e-16
```

MLR Figure 1



```
lm(formula = price ~ ., data = training_set2)

Residuals:
    Min      1Q  Median      3Q     Max 
-6957.2 -1185.9    -6.4  1194.4  6909.3 

Coefficients:
                   Estimate Std. Error   t value Pr(>|t|)    
(Intercept)       2.996e+02  8.173e+01 3.666e+00 0.000248 ***
squareMeters      1.000e+02  7.385e-04 1.354e+05  < 2e-16 ***
hasYard           3.027e+03  4.246e+01 7.130e+01  < 2e-16 ***
hasPool           2.943e+03  4.245e+01 6.933e+01  < 2e-16 ***
floors            5.451e+01  7.358e-01 7.409e+01  < 2e-16 ***
cityPartRange     4.618e+01  7.387e+00 6.252e+00 4.26e-10 ***
isNewBuilt        1.465e+02  4.246e+01 3.451e+00 0.000562 ***
hasStormProtector 1.450e+02  4.245e+01 3.415e+00 0.000642 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1898 on 7992 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 2.621e+09 on 7 and 7992 DF,  p-value: < 2.2e-16
```

MLR Figure 2

The first linear model pictured here in (MLR Figure 1) is the summary of the first multiple regression model that we created in R studio. It features all of the variables as predictor variables for the response variable of price. The R squared value that we got was 1, which indicates that the linear regression model did an extremely efficient job of predicting the price when all the factors of a house in Paris are considered. The predictor values with stars beside them mean that we have statistically significant evidence that these values are important predictors for price, while variables with no stars do not have much effect on the price of the house.

Because of this, we created a second multiple linear regression model that used only the significant variables as predictors. This can be visualized in (MLR Figure 2). We still have an R squared value of 1. but the median of the residuals was closer to zero, which means that for the entire data set, our predictions were closer to the actual values than they were with the first model which had a higher residual median. Additionally, using less predictor variables makes the program run more efficiently and prevents other complications from occurring in the model.

IV. DATA SET GRAPHICAL EXPLORATION

Figures 1 – 9 show the distribution of the data using histograms.

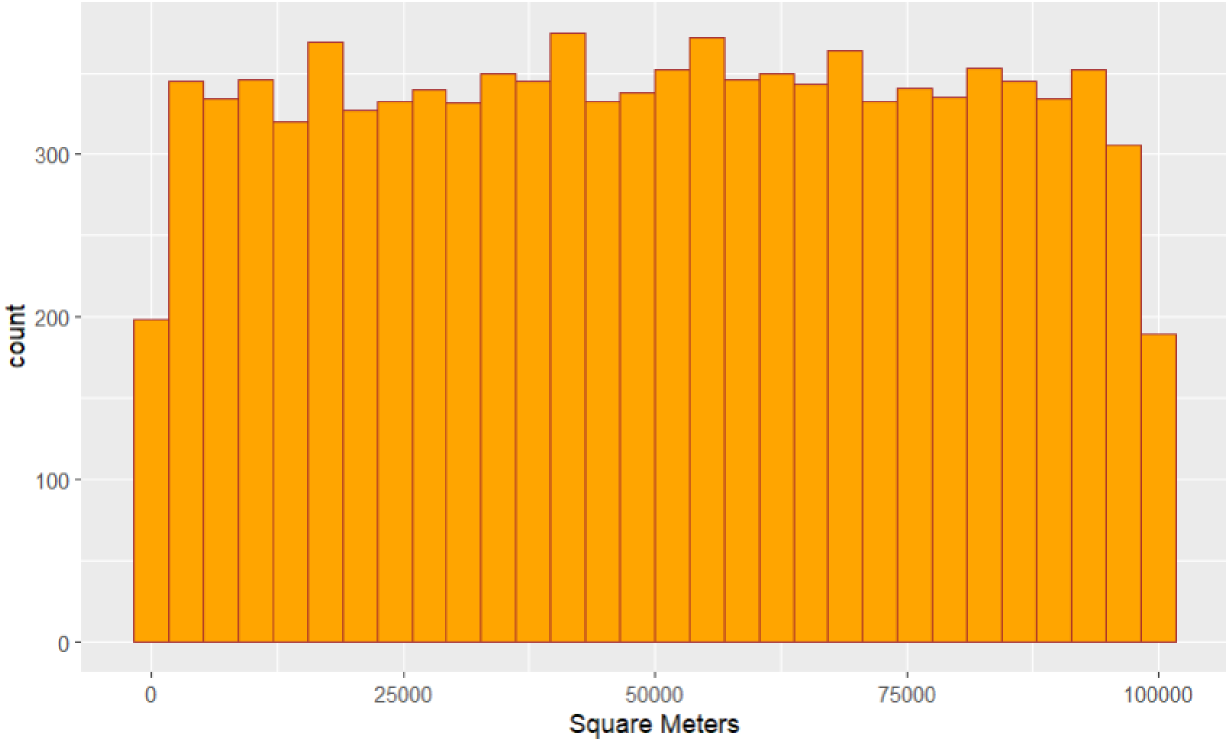


Figure 1 shows the histogram of the square meters of the homes.

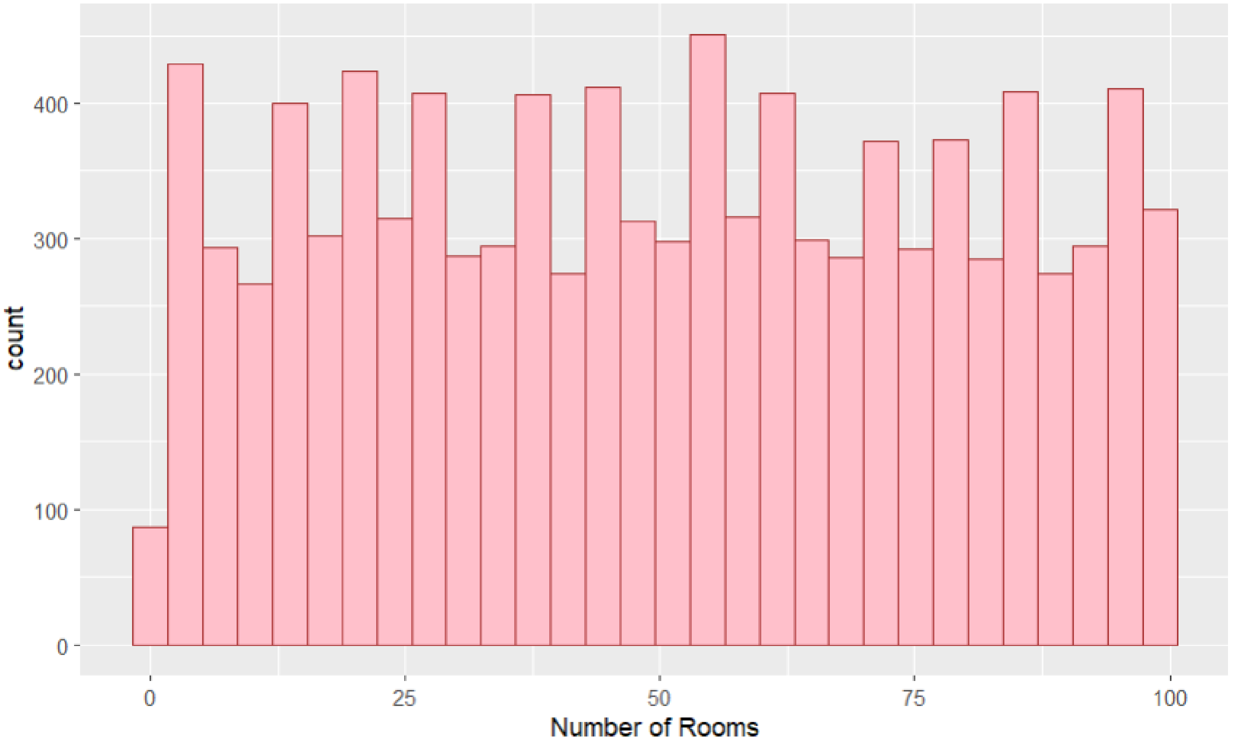


Figure 2 shows the histogram of the number of rooms for houses.

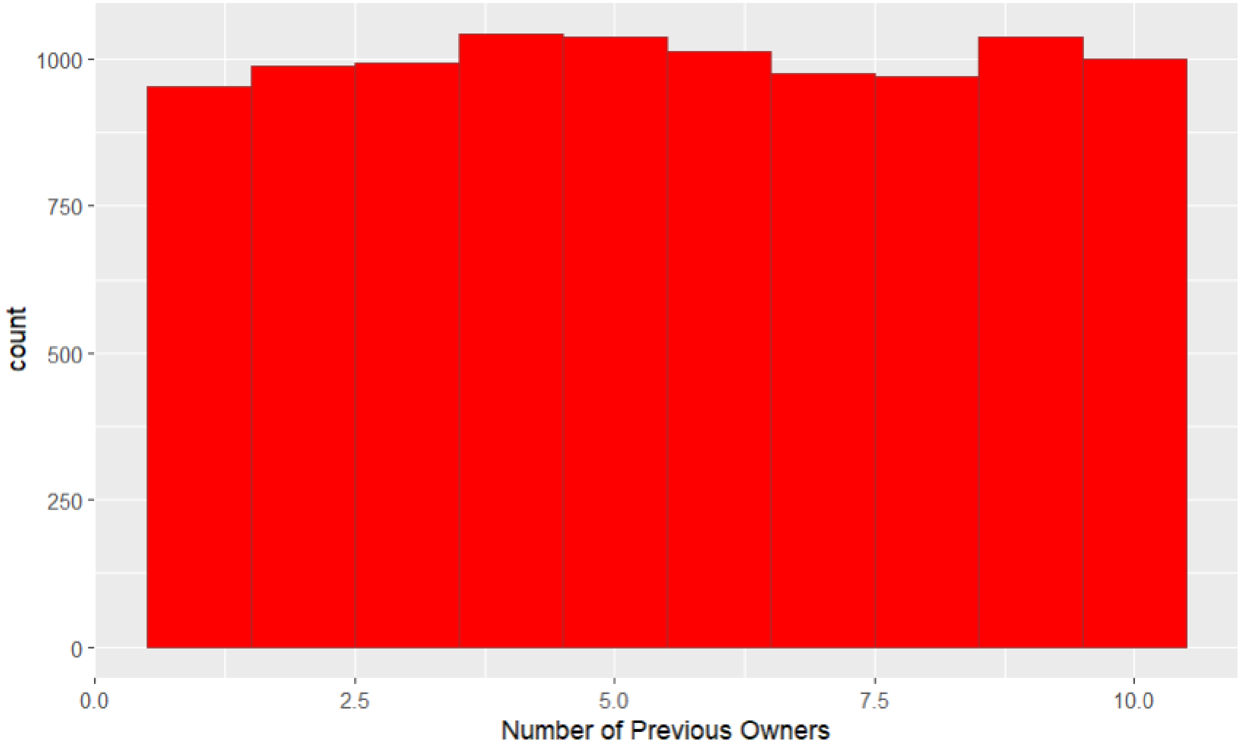


Figure 3 shows the histogram of the number of previous owners for the houses.

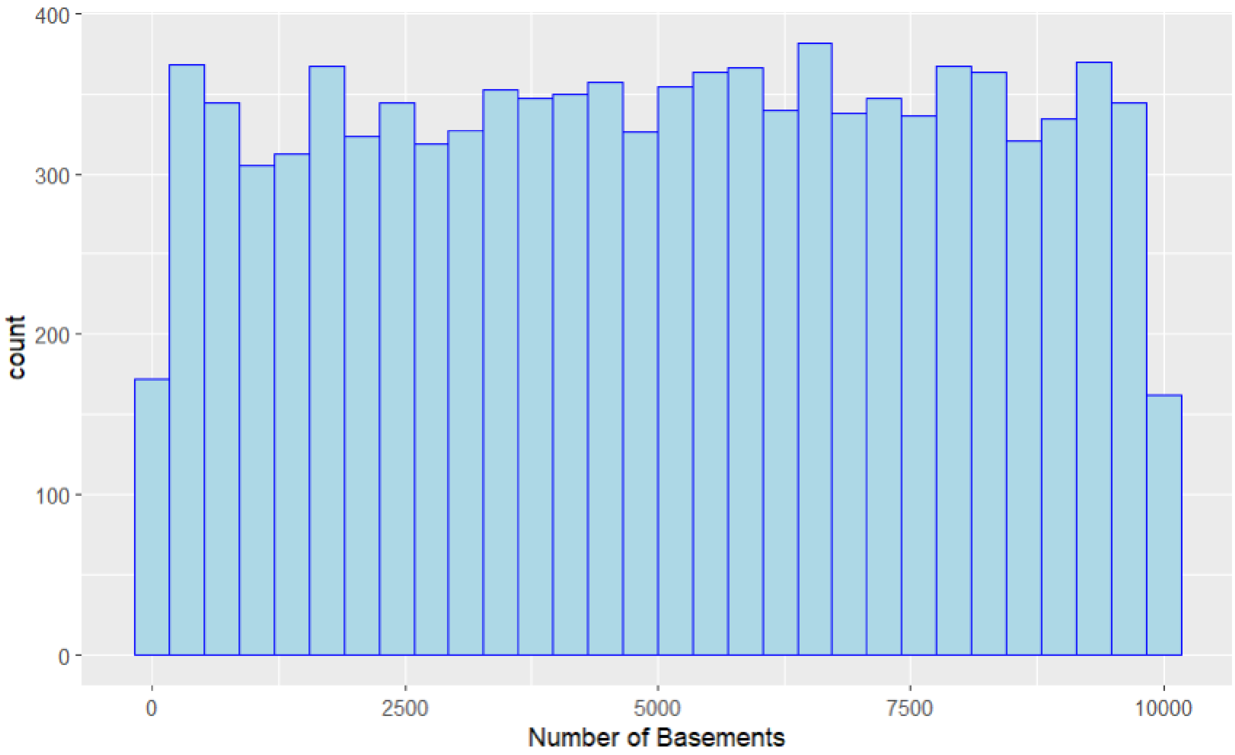


Figure 4 shows the histogram of the number of basements for the houses.

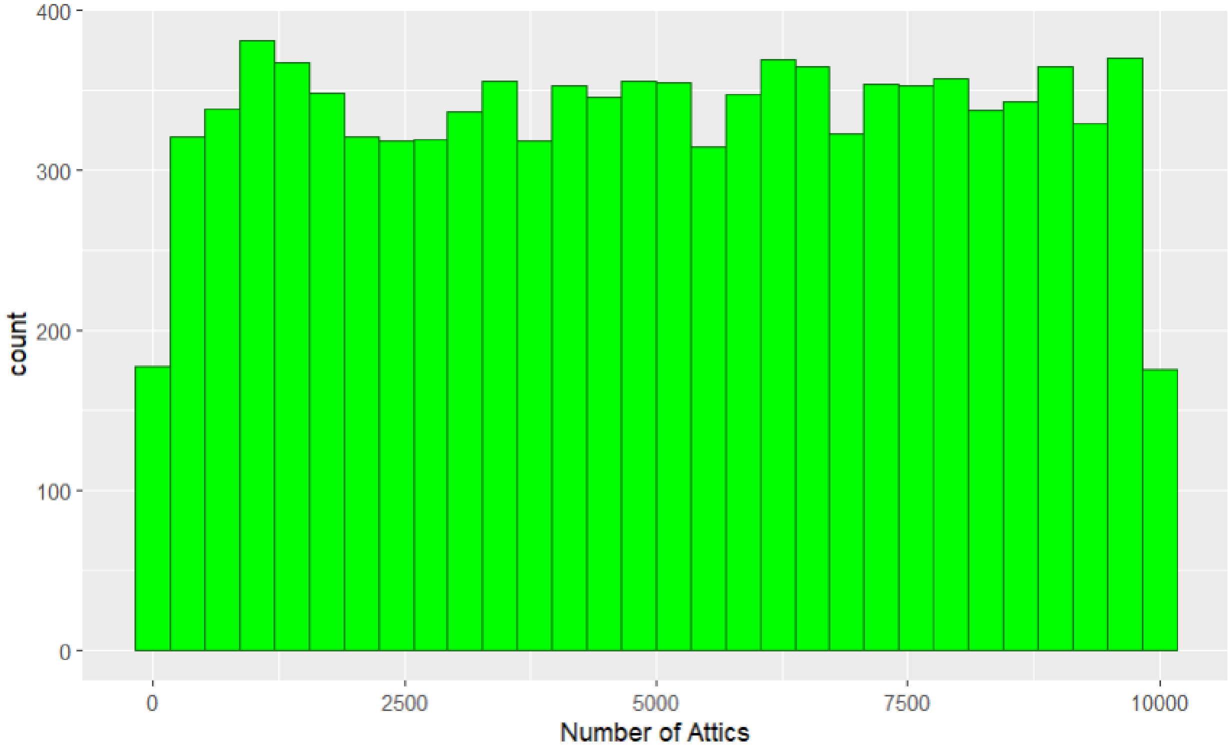


Figure 5 shows the histogram of the number of attics for the houses.

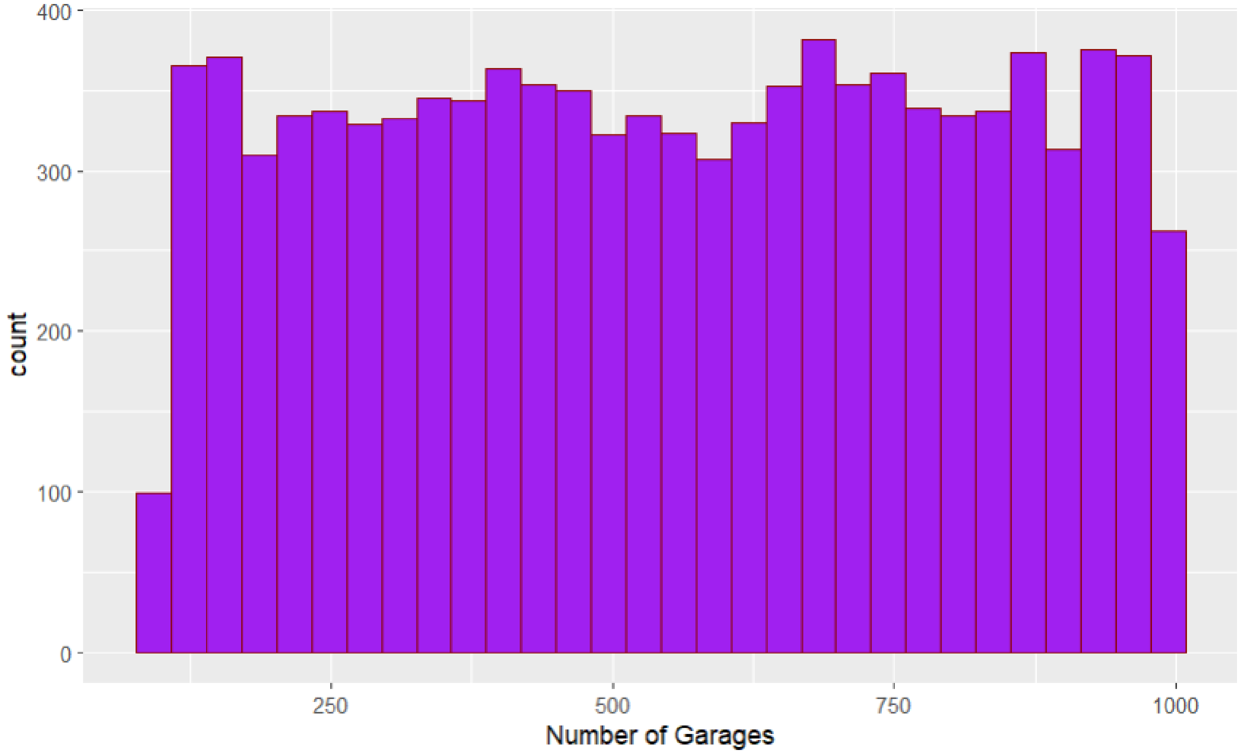


Figure 6 shows the histogram of the number of garages for the houses.

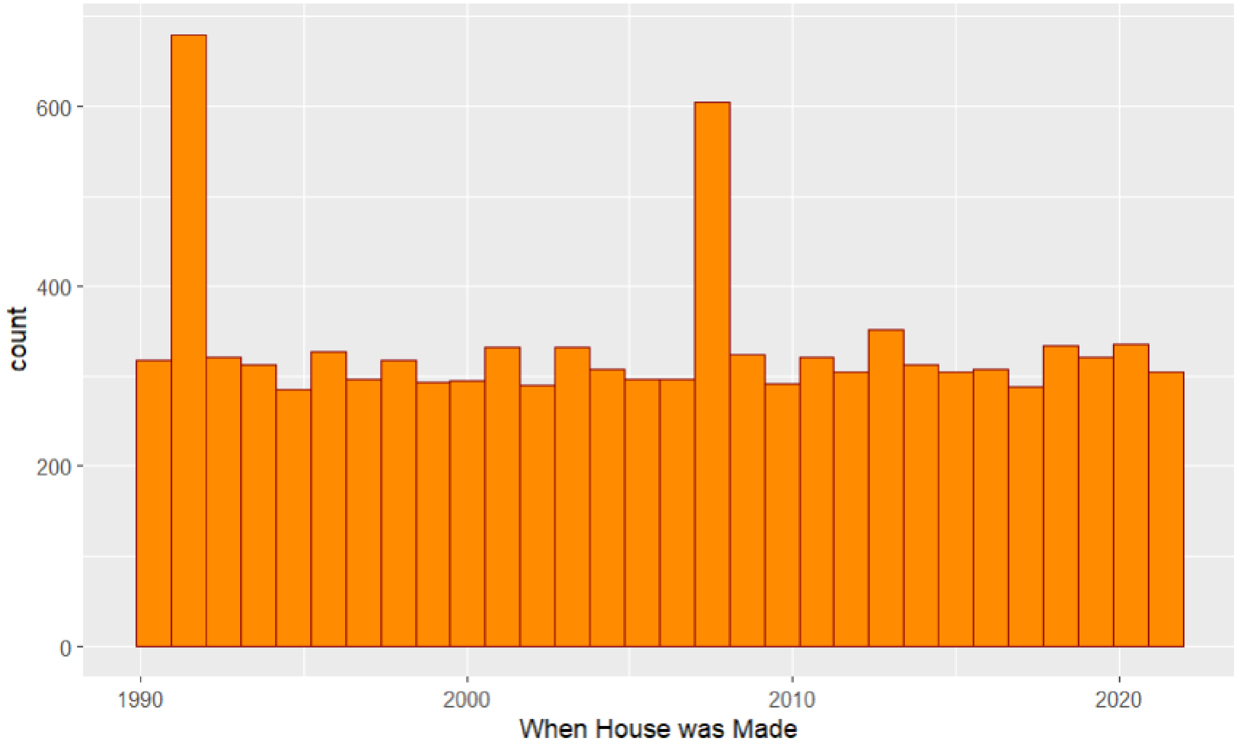


Figure 7 shows the histogram for the year the houses were made.

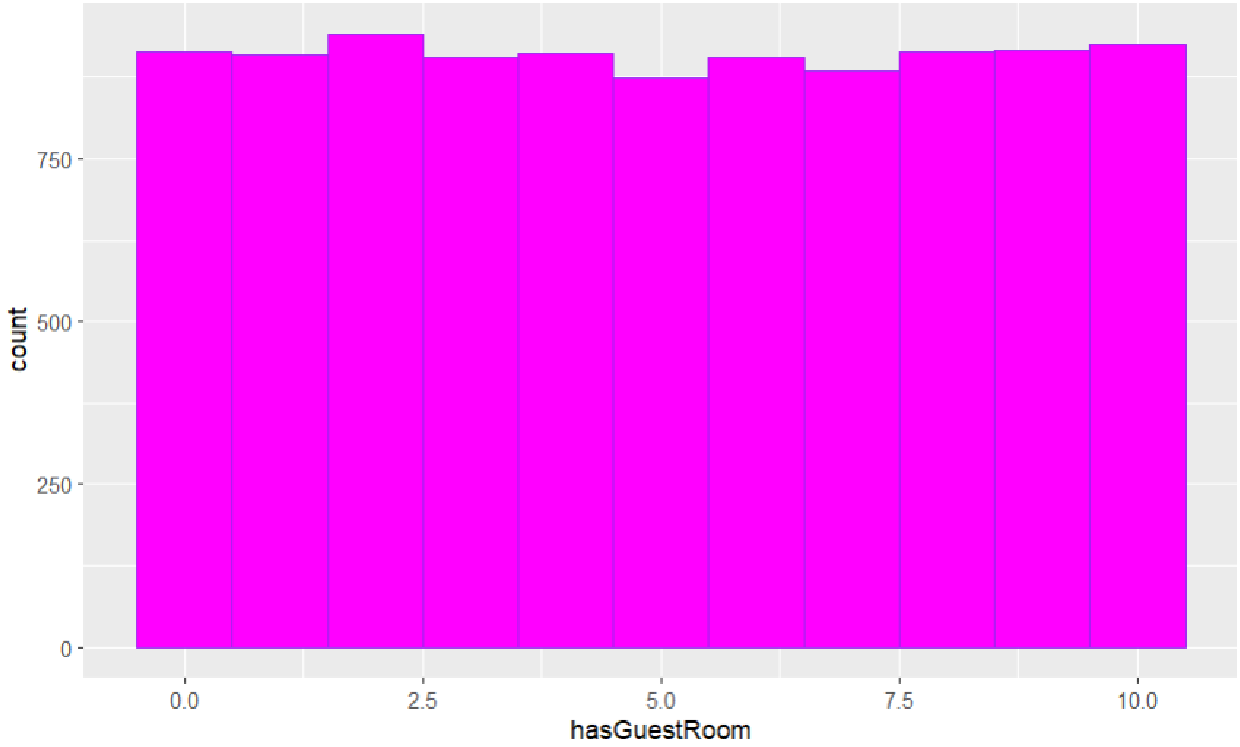


Figure 8 shows the histogram of the number of guest rooms the houses have.

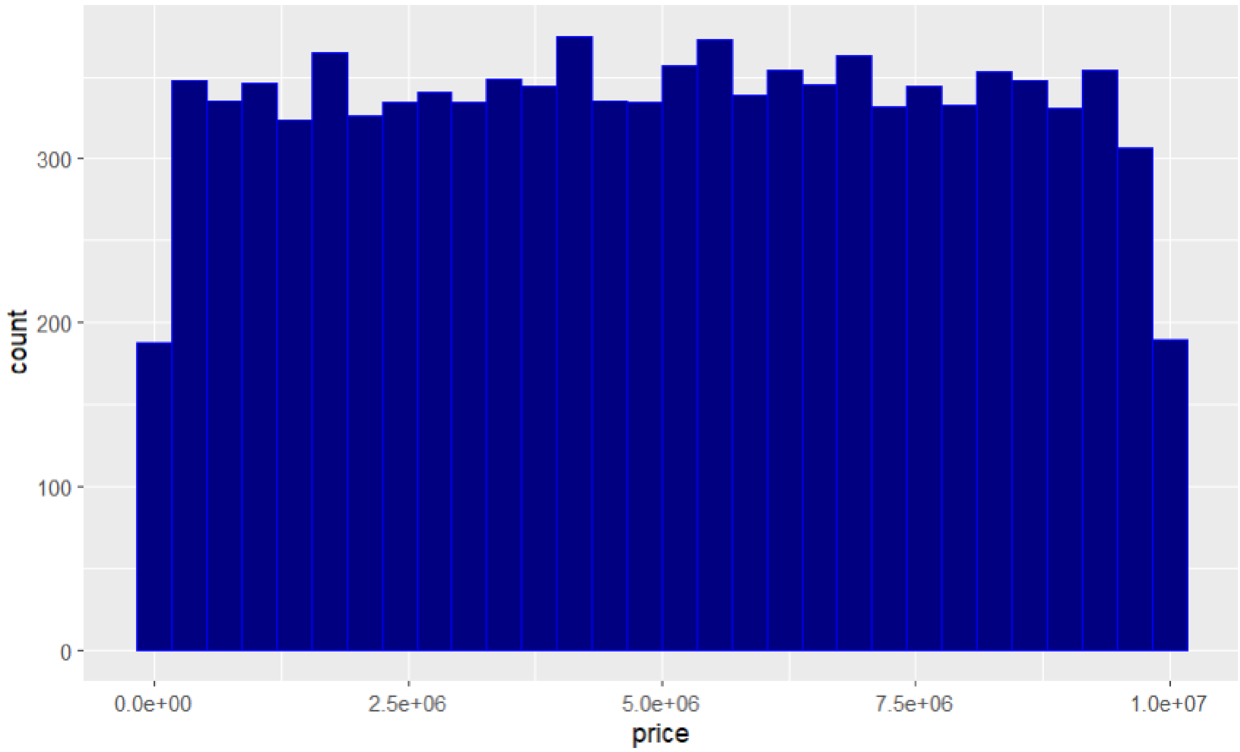


Figure 9 shows the histogram of the prices for the houses.

Figures 10 – 14 show the distribution of data using boxplots.

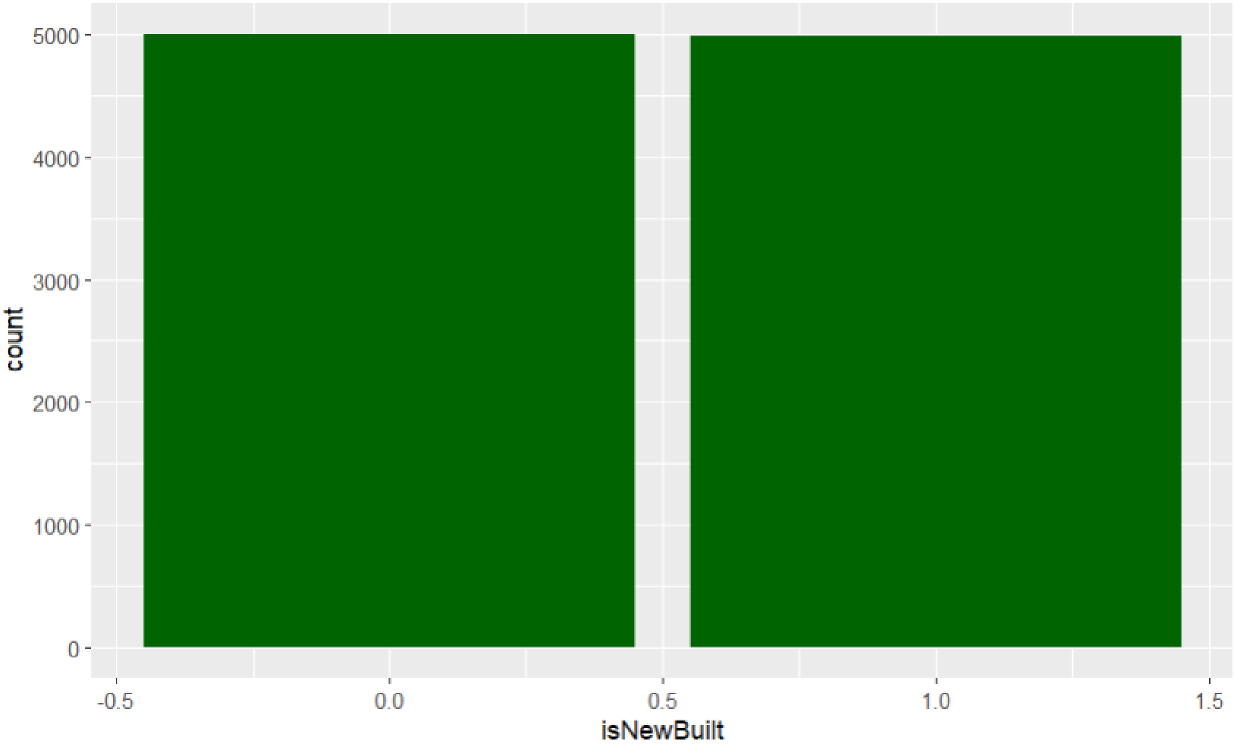


Figure 10 shows the bar plot of whether the house is newly built or not.

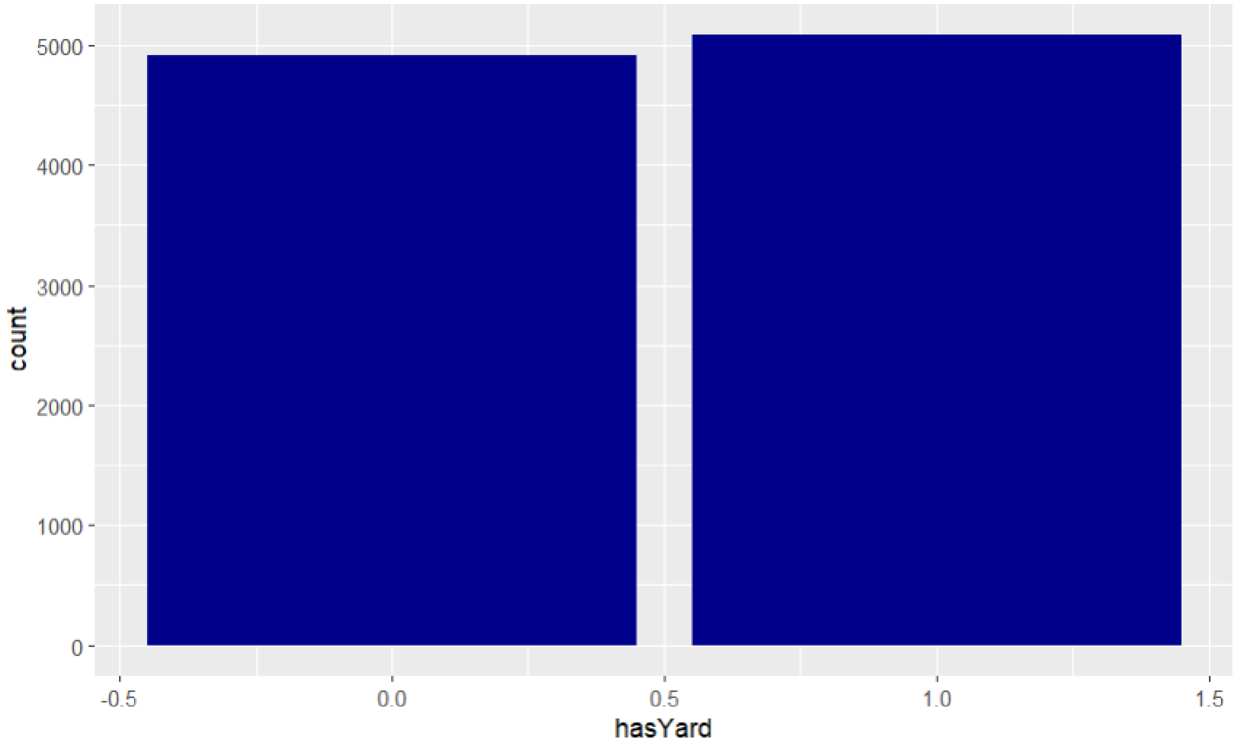


Figure 11 shows the bar plot of whether the house has a yard or not.

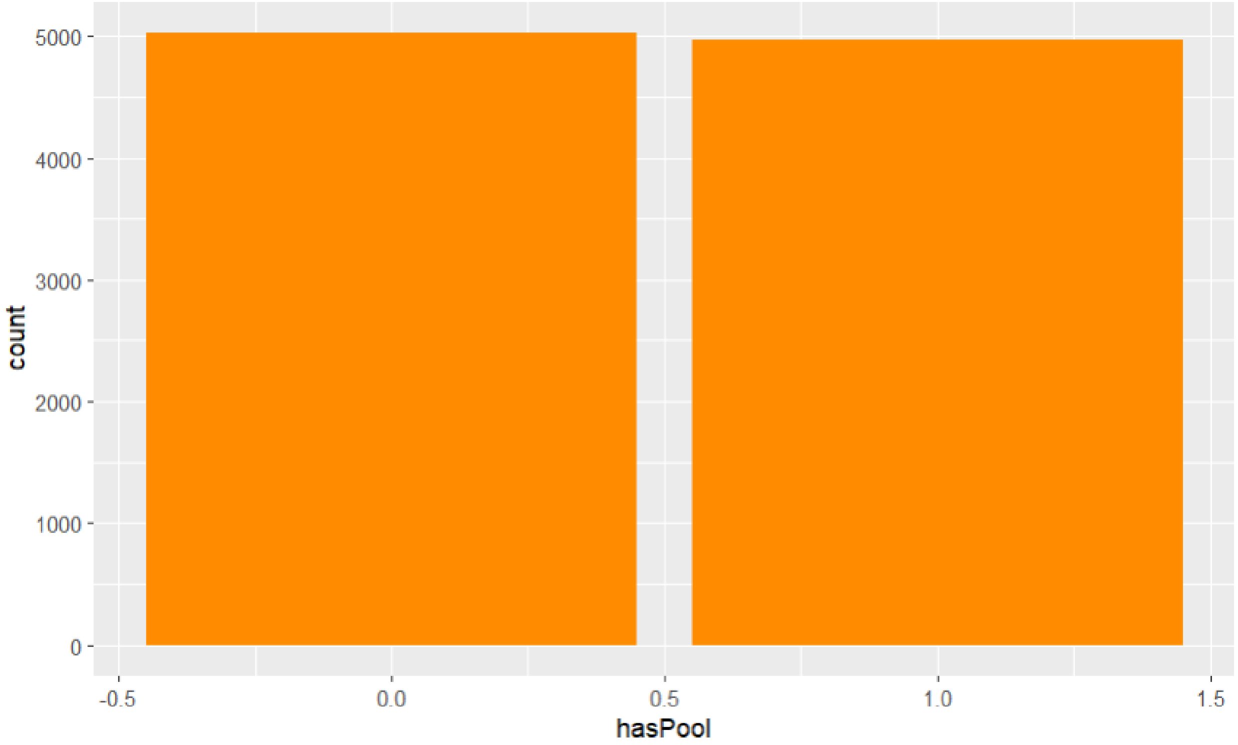


Figure 12 shows the bar plot of whether the house has a pool or not.

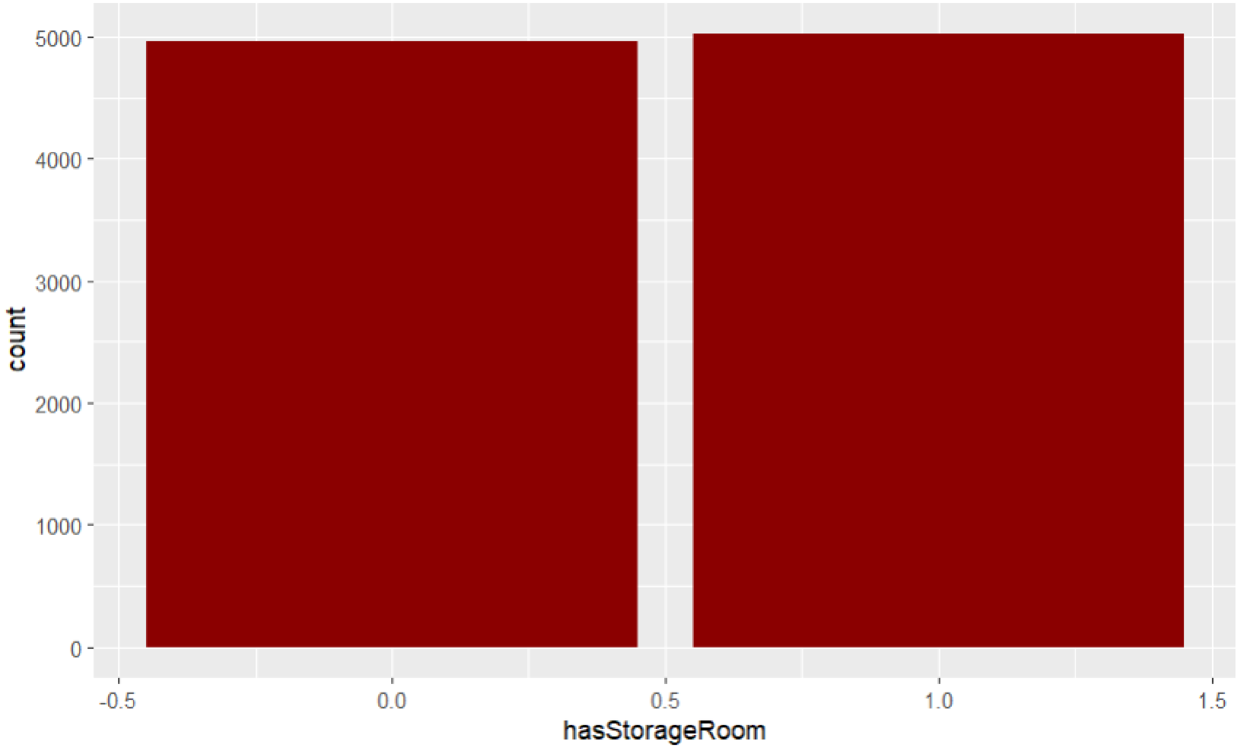


Figure 13 shows the bar plot of whether the house has a storage room or not.

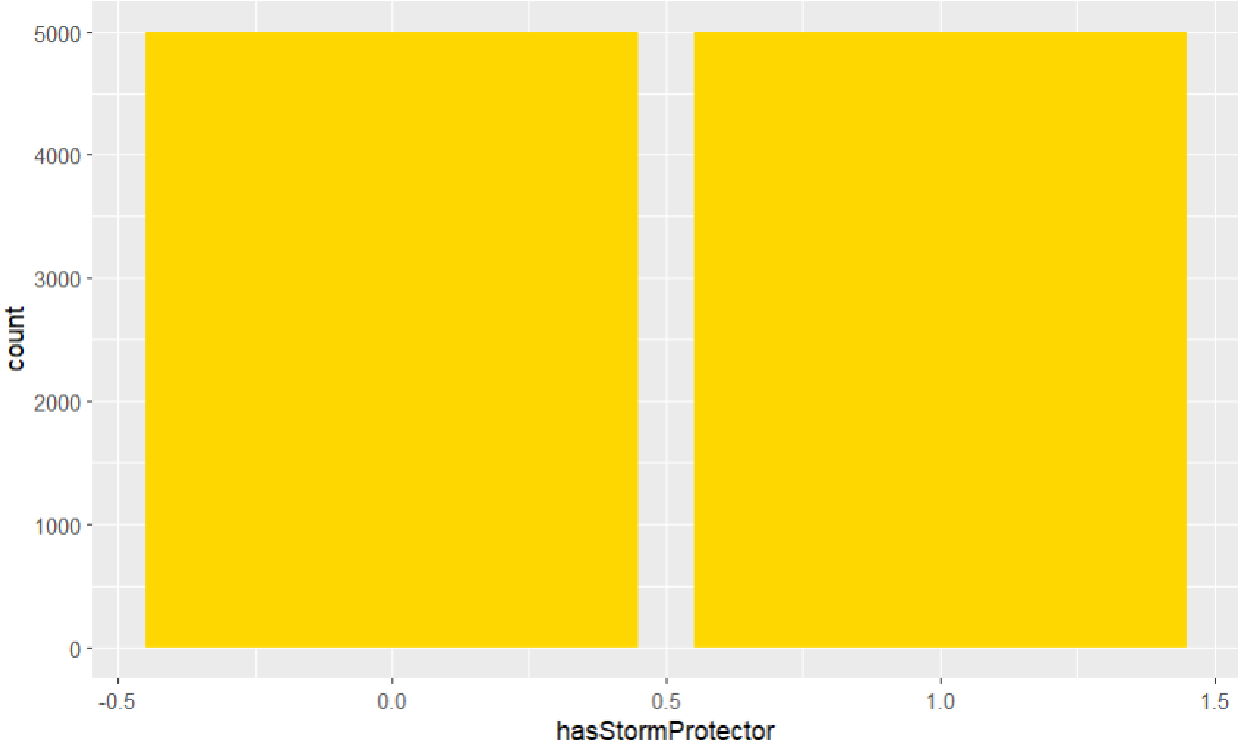


Figure 14 shows the bar plot of whether the house has a storm protector or not.

Figures 15 – 17 show the correlation between the price of the house and how large the house is in square meters.

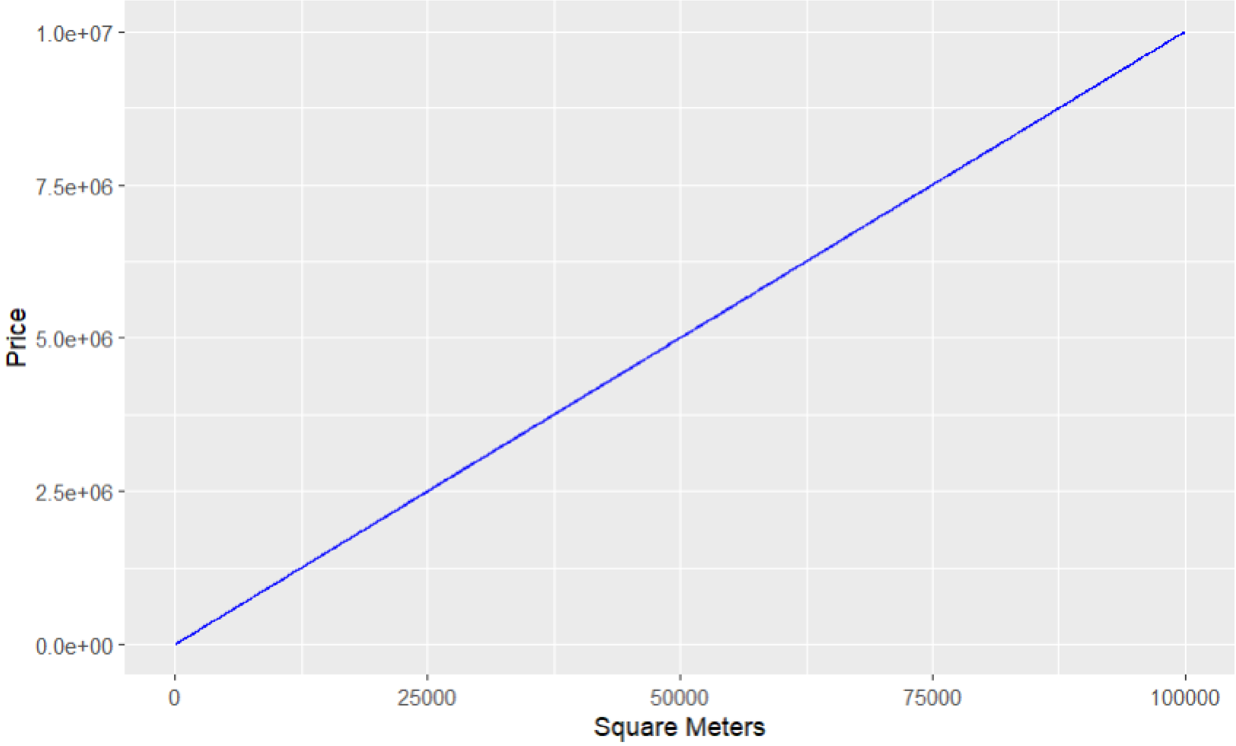


Figure 15 shows the line plot of the price of the house vs. the square meters of the house for all the houses within the dataset.

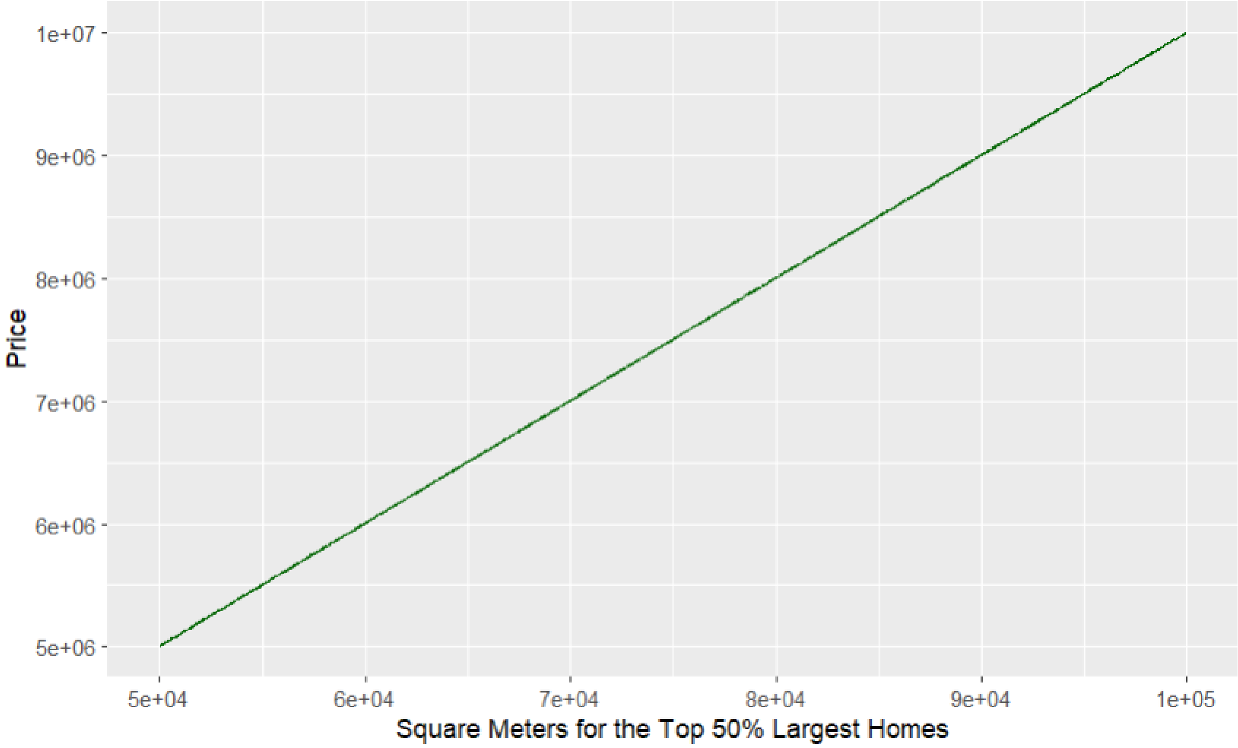


Figure 16 shows the line plot of the price of the house vs. the square meters of the house for the top 50% largest homes.

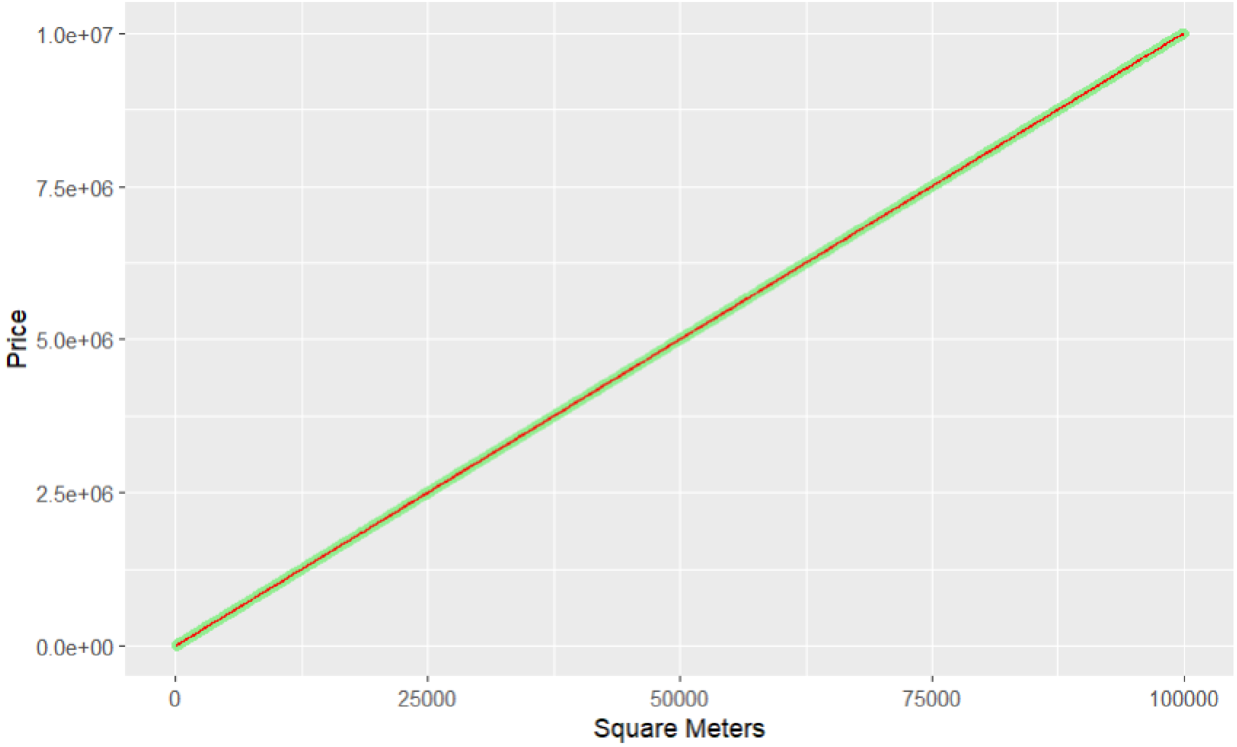


Figure 17 shows the scatter plot of the price of the house vs. the square meters of the house for all the houses within the dataset.

V. SUMMARY OF FINDINGS

Based on the distribution of the data from both the histograms (Figures 1 – 9) and bar plots (Figures 10 – 14), it is clear there is no skewedness. This implies the data is not very dispersed and does not deviate much from the mean. This was confirmed when we looked at the summary statics and standard deviation for all the data within the dataset which can be referenced in Table 2.

Based on the line plots (Figure 15 and Figure 16), there is a strong correlation between price and how large the house is in square meters. This is confirmed by our scatter plot (Figure 17) which shows the densely packed points along the line of best fit. Additionally, when we filtered our data to focus on the top 50% largest houses, we found they had significantly larger prices which was expected.

For our first regression analysis, we set the "price" of the homes as our dependent variable. Based on our Multiple Linear Regression model, we found that the following variables: size of home in square meters (squareMeters), if the house has a yard (hasYard), if the house has a pool (hasPool), number of floors (floors), part of city range (partCityRange), if it is newly built (isNewBuilt), were all very statistically significant. Thus, these variables were the most valid predictor values. If the house has a storm protector (hasStormProtector) was also somewhat statistically significant, and number of garages (garage) was a little statistically significant.

We got an R squared value of 1 which implies that our Multiple Linear Regression model performed very well. This can be contributed to the fact that the data within our dataset did not have much variance. The adjusted R squared value was 1. For our root mean squared value, we got 1948.7. For our mean absolute error value, we got 1529.6. For the valid predictor values, we got a P value of less than 2.2 x 10^-16 which is quite impressive.

For our second regression analysis, we ran another Multiple Linear Regression model using only the valid predictor values. The valid predictor values are decided based on small p-values in original regression analysis. We confirmed that these selected variables were statistically significant, and we got an R squared value of 1. Once again, this implied that our Multiple Linear Regression model was accurate. For the adjusted R squared value we got 1. For the valid predictor values, we got a P value of less than 2.2 x 10^-16 which was expected.