Housing Price Regression Analysis

Sarah Thuman

Abstract

Being able to predict housing prices is a field that draws a lot of interest and study, since so many people invest in real estate, whether as a source of income or to use as a source of equity. Being able to predict how much a house is worth based on facts about the house that can be easily found from websites such as Zillow, would allow buyers to determine if the house is being over- or under- priced. In this paper four hypotheses were tested to determine if certain variables, unit price, basement present, number of bedrooms, and location, contributed to the unit price of houses. Regression analysis was also performed to determine which type of regression model was best at predicting house prices and if any variables were unnecessary and could be removed from the model. Some basic statistics were also performed to provide an understanding of the data being analyzed. This analysis shows that based on the given data set, a multiple linear regression model is best at predicting house prices and that all the variables were necessary to do so.

Introduction

Since the COVID-19 pandemic, house prices have experienced a bubble that has yet to come down [1]. While there are some houses that haven't been overpriced to be unprofitable in the long run, many have which makes it tricky to purchase a home or investment property in the housing market today. Being able to determine if a house is priced over or below its value would help anyone looking to purchase property to determine if it is worth investing in now or waiting for the price to come down.

Previous studies that have investigated this field, have ranged from attempting to determine which factors most affect housing prices [2] to using regression to predict housing prices [3] to machine learning algorithms [4].

The aim of this study is to determine if it is possible to predict housing prices based on factors that can be found on any number of real estate websites and which regression model is best. The data set used for analysis [5] contained thirteen variables: price, area, number of bedrooms, number of bathrooms, number of stories, number of parking spots, located on a main road, has a guestroom, has a basement, has hot water heating, has AC, and located in a preferred area. The furnishing status of the property was also included in the data set but was excluded from analysis in an attempt to cut back on the number of dummy variables. Some assumptions made about the data is that all prices are in USD and the houses included are located in the United States.

Methodology

Basic Statistics

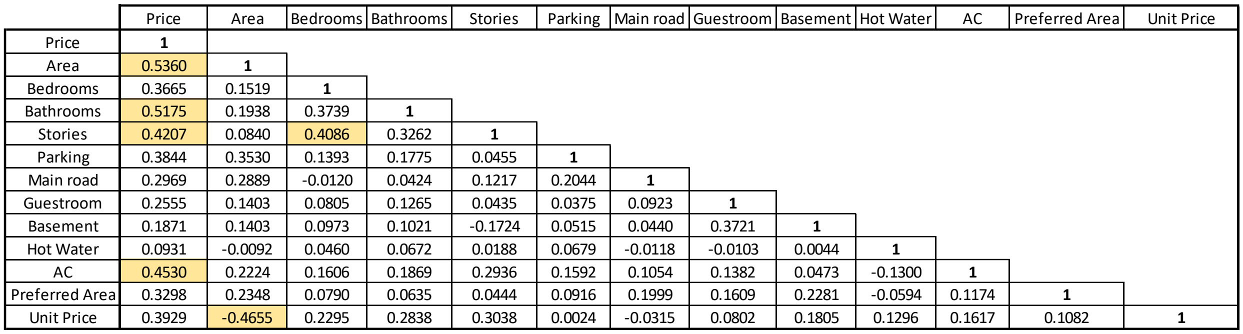
Before conducting any analysis on the housing data, it was first cleaned. All columns were checked for any missing data, dummy variables were created for six categorical columns, and extra columns were removed. Regarding the dummy variables, for all columns 1 represents a "yes" response and 0 represents a "no" response. An additional column was added to the data, unit price, where $unit\ price = \frac{price}{area}$. Once the data was cleaned, basic statistics were conducted on the housing data to better understand the data. Mean and standard deviation were calculated for all variables and is summarized in Table 1.

	Price	Area	Bedrooms	Bathrooms	Stories	Parking
Mean	4,766,729.25	5,150.54	2.97	1.29	1.81	0.69
Standard Deviation	1,870,439.62	2,170.14	0.74	0.50	0.87	0.86

	Main road	Guestroom	Basement	Hot Water	AC	Preferred Area	Unit Price
Mean	0.86	0.18	0.35	0.05	0.32	0.23	993.33
Standard Deviation	0.35	0.38	0.48	0.21	0.47	0.42	346.54

Table 1: Summary of basic statistics

A correlation matric was also calculated to determine the relationship between variables. It can be found in Table 2. Almost all the variables have some positive correlation. None of the variables were highly correlated, however six were mildly correlated and are highlighted in Table 2. The price of the house and the area in square feet have the highest correlation at 0.536, with the price of the house and number of bathrooms being the second highest at 0.518.



	Price	Area	Bedrooms	Bathrooms	Stories	Parking	Main road	Guestroom	Basement	Hot Water	AC	Preferred Area	Unit Price
Price	**1**												
Area	0.5360	**1**											
Bedrooms	0.3665	0.1519	**1**										
Bathrooms	0.5175	0.1938	0.3739	**1**									
Stories	0.4207	0.0840	0.4086	0.3262	**1**								
Parking	0.3844	0.3530	0.1393	0.1775	0.0455	**1**							
Main road	0.2969	0.2889	-0.0120	0.0424	0.1217	0.2044	**1**						
Guestroom	0.2555	0.1403	0.0805	0.1265	0.0435	0.0375	0.0923	**1**					
Basement	0.1871	0.1403	0.0973	0.1021	-0.1724	0.0515	0.0440	0.3721	**1**				
Hot Water	0.0931	-0.0092	0.0460	0.0672	0.0188	0.0679	-0.0118	-0.0103	0.0044	**1**			
AC	0.4530	0.2224	0.1606	0.1869	0.2936	0.1592	0.1054	0.1382	0.0473	-0.1300	**1**		
Preferred Area	0.3298	0.2348	0.0790	0.0635	0.0444	0.0916	0.1999	0.1609	0.2281	-0.0594	0.1174	**1**	
Unit Price	0.3929	-0.4655	0.2295	0.2838	0.3038	0.0024	-0.0315	0.0802	0.1805	0.1296	0.1617	0.1082	**1**

Table 2: Correlation matrix

Scatter plots were also created at looked at but because most of the variables are discrete, the only scatter plot that was useful to look at was between price and area and can be found in Figure 1. Histograms were also produced from the data. An example can be found in Figure 2. All other histograms can be found in Appendix A. From the histograms, none of the variables appeared to be normally distributed. To verify this, a Shapiro-Wilk Normality test was conducted for each variable. The results from the test confirmed that none of the variables are normally distributed. The results are summarized in Table 3. Based on the results from the Shapiro-Wilk tests and the large difference in scales of the variables, log normalization was used to standardize the data so that the regression analysis would not be overly affected by the variables with large scales.

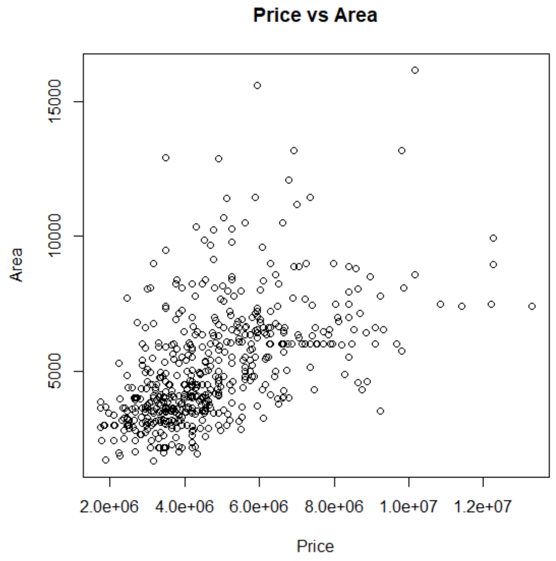


Figure 1: Scatter plot Price vs Area

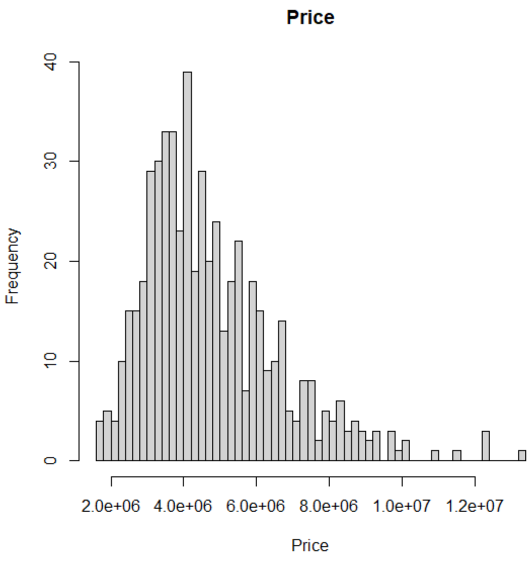


Figure 2: Histogram of price

Shapiro-Wilk	Price	Area	Bedrooms	Bathrooms	Stories	Parking
p-value	3.16E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16

Shapiro-Wilk	Main road	Guestroom	Basement	Hot Water	AC	Preferred Area	Unit Price
p-value	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.93E-10

Table 3: Summary of Shapiro-Wilk tests

Hypothesis Testing

As mentioned in the basic statistics section, the housing data was log normalized. Four hypotheses were tested using this data. Four hypotheses were tested. The first hypothesis was if the mean price of houses with low unit prices was less than the mean price of houses with high unit prices. The data was split into two groups: the houses with unit prices less than or equal to the mean unit price and the houses with unit prices greater than the mean unit price. A random sample of 100 entries was taken with replacement from each group and a significance level of 0.05 was used. The test was conducted using one-sided two sample t-test. The null hypothesis was $H_0: \mu_{price\ with\ low\ unit\ price} = \mu_{price\ with\ high\ unit\ price}$ and the alternative hypothesis was

$H_0: \mu_{price\ with\ low\ unit\ price} < \mu_{price\ with\ high\ unit\ price}$. Since the p-value of the test was 2.2e-16, the null hypothesis was rejected.

The second hypothesis was that the mean unit price of houses with basements is greater than the mean unit price of houses without basements. The data was split into two groups: houses with a basement and houses without a basement. A random sample of 100 entries was taken with replacement from each group and a significance level of 0.05 was used. The test was conducted using one-sided two sample t-test. The null hypothesis was $H_0: \mu_{unit\ price\ with\ basement} = \mu_{unit\ price\ without\ basement}$ and the alternative hypothesis was $H_0: \mu_{unit\ price\ with\ basement} > \mu_{unit\ price\ without\ basement}$. Since the p-value of the test was 0.001, the null hypothesis was rejected.

The third hypothesis was that the unit price of houses located in a preferred area is greater than the unit price of houses not located in a preferred area. The data was split into two groups: houses located in preferred areas and houses not located in preferred areas. A random sample of 100 entries was taken with replacement from each group and a significance level of 0.05 was used. The test was conducted using one-sided two sample t-test. The null hypothesis was $H_0: \mu_{unit\ price\ in\ preferred\ area} = \mu_{unit\ price\ not\ in\ preferred\ area}$ and the alternative hypothesis was $H_0: \mu_{unit\ price\ in\ preferred\ area} > \mu_{unit\ price\ not\ in\ preferred\ area}$. The p-value was 0.002 and therefore the null hypothesis was rejected.

The fourth and last hypothesis tested was that the unit price of houses where the number of bedrooms is less than the mean is less than the unit price of houses where the number of bedrooms is greater than the mean. The data was split into two groups: houses with less than the mean number of bedrooms (2.96) and those with more bedrooms than the mean. A random sample of 100 entries was taken with replacement from each group and a significance level of 0.05 was used. The test was conducted using one-sided two sample t-test. The null hypothesis was $H_0: \mu_{unit\ price\ with < 2.96\ bedrooms} = \mu_{unit\ price\ with > 2.96\ bedrooms}$ and the alternative hypothesis was $H_A: \mu_{unit\ price\ with < 2.96\ bedrooms} < \mu_{unit\ price\ with > 2.96\ bedrooms}$. Since the p-value of the test was 2.5e-06, the null hypothesis was rejected. All hypothesis tests are summarized in Table 4.

Hypothesis	Test	p-value	Reject?	Confidence level	Significance Level
1	two sample, one-sided t-test	2.2E-16	Yes	0.95	0.05
2	two sample, one-sided t-test	0.001	Yes	0.95	0.05
3	two sample, one-sided t-test	0.002	Yes	0.95	0.05
4	two sample, one-sided t-test	2.50E-06	Yes	0.95	0.05

Table 4: Summary of hypothesis tests

Regression Analysis

Regression analysis was performed on the data set. Because the goal of the study was to see if it is possible to predict the price of a house based on the other variable, the dependent variable for all repression analysis was the price variable and the remaining variables were treated as independent variables. To find the model that best fit the data, multiple regression techniques were performed and then the best was selected based on the Akaike information criterion (AIC). The log normalized data was used and split into 80/20 training and test sets. The types of regression models tested were simple linear regression, multiple linear regression, stepwise forward selection, LASSO regression, and polynomial regression. The regression models will be discussed in this order.

The first regression model tested was a simple linear regression model where price was the dependent variable and area was the independent variable. The simple linear regression equation from the model was $\hat{y} = 10.72 + 0.54x_{area}$. The AIC was 188.73. For the multiple linear regression model, price was again the dependent variable, and all other variables, excluding unit price, were used as the independent variables. The equation from this model was $\hat{y} = 12.10 + 0.32x_{area} + 0.12x_{bedrooms} + 0.25x_{bathrooms} + 0.15x_{stories} + 0.03x_{parking} + 0.13x_{mainroad} + 0.04x_{guestroom} + 0.09x_{basement} + 0.22x_{hotwater} + 0.18x_{AC} + 0.14x_{preferred}$. The AIC was -108.00.

The next two models were used to see if any variable reduction could be done to the multiple linear regression model. The first variable reduction model was the stepwise forward selection model. The equation from this model did not result in a reduction in variables and therefore was the same as the multiple linear regression model equation. The AIC was -108 which was also the same as the multiple linear regression model. The next variable reduction technique tested was LASSO regression. The equation from this model was also the same as the multiple linear regression model. The AIC was unable to be calculated for the model but assumed to be the same as the multiple linear regression model since the regression equations were the same. Because both variable reduction techniques failed to reduce the number of variables needed to fit a line to the data, it was concluded that all the variables included in the data set were needed for multiple linear regression.

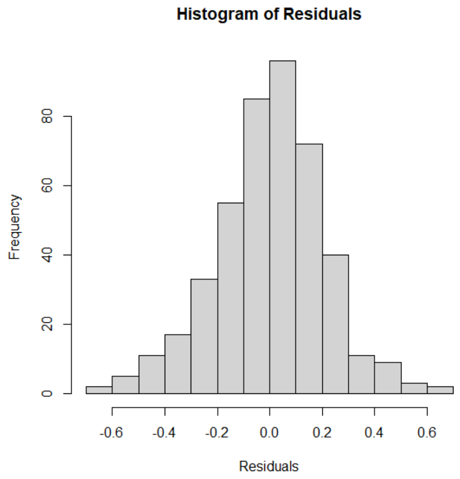


Figure 3: Histogram of Residuals

The last method tested was polynomial regression. Two models were fit using polynomial regression: one linear and one logistic. For both models, the only independent variable used was area and a polynomial of degree 4 was selected. Like all other models, price was the dependent variable. The equation for the linear polynomial regression was $\hat{y} = 15.29 + 4.56x_{area} - 0.35x_{area}^2 - 0.87x_{area}^3 + 0.23x_{area}^4$ and the AIC was 183.99. The equation for the logistic polynomial regression was, $\hat{y} = 0.48 + 5.45x_{area} - 0.43x_{area}^2 - 1.28x_{area}^3 + 0.15x_{area}^4$ and the AIC was 502.73.

Since the multiple linear regression model had the best AIC, it was chosen as the model to be used to predict house prices. The model was used on the test data to predict the price variable and residuals were computed. A histogram of the residuals is shown in Figure 3.

PCA

Component	Cumulative Proportion
1	26.70%
2	46.00%
3	61.30%
4	70.80%
5	77.90%
6	83.70%
7	88.80%
8	92.80%
9	96.10%
10	98.30%
11	1.00%

Table 5: PCA cumulative proportion

Principle component analysis was also performed on the data to try and reduce the number of variables needed to explain the data. The scaled data was used for the analysis and a correlation matrix based of it was computed in order to perform the analysis. A scree plot, biplot, and contribution of each variable chart were created to be able to examine the data. While there was no clear break in the scree plot to indicate how many components should be kept, assessing the cumulative proportions, shown in Table 5, led to the decision that 8 components, which explained 92.8% of the variance, would be enough. However, despite being better at reducing the number of variables needed than the other two variable reduction techniques, PCA was still only able to reduce the number of variables from 11 to 8. An interpretation of the 8 PC loadings and all plots concerning PCA can be found in Appendix B. In the PC loadings, the variables that positively contributed to the loading by greater than 0.3 are highlighted in green and the variables that negatively contributed by less than 0.3 are highlighted in red.

Conclusion/Discussion

In this experiment, housing data was explored to find a best fit model that could be used to predict housing prices. Dimension reduction techniques were also used to see if the number of variables needed to predict housing prices could be reduced. The results from the regression analysis showed that a multiple linear regression was the best choice to fit the data and that all variables are needed for the best fit model. It is interesting to note that the four hypothesis tests that were conducted appear to show a trend that all variables were needed to predict house price, since the hypotheses that there was no difference if the feature was included or not was rejected every time. Further testing like what was done in this study could be done to confirm. For future studies it could be interesting to determine if including different variables, such as the year the house was built, could improve the regression model.

One aspect of the data that may have skewed the results, is that based on the assumption that price was provided in USD, the data set only looked at multi-million-dollar homes. This could account for the fact that, except for PCA, dimension reduction techniques were ineffective in reducing the number of variables needed in the analysis. Another area that could have skewed the results is that the location of the houses was not included in the data set.

Code for the analysis can be found in Appendix C which is included as a separate file.

References

[1] Campisi, N. (2022, May 4). Will The Housing Market Crash? Experts Give 5-Year Predictions. Forbes Advisor. https://www.forbes.com/advisor/mortgages/real-estate/will-housing-market-crash/

[2] Wang, Y., & Jiang, Y. (2016). An Empirical Analysis of Factors Affecting the Housing Price in Shanghai. Asian Journal of Economic Modelling, 4(2), 104–111. https://ideas.repec.org/a/asi/ajemod/v4y2016i2p104-111id853.html

[3] Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. Scientific Programming, 2021, 1–9. https://doi.org/10.1155/2021/7678931

[4] Ho, W. K. O., Tang, B.-S., & Wong, S. W. (2020). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 1–23. https://doi.org/10.1080/09599916.2020.1832558

[5] Housing Price Prediction. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction

Appendices

Appendix A

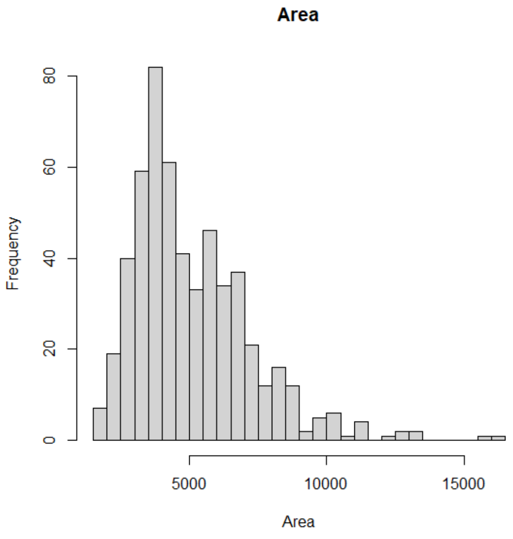


Figure 4: Histogram of area

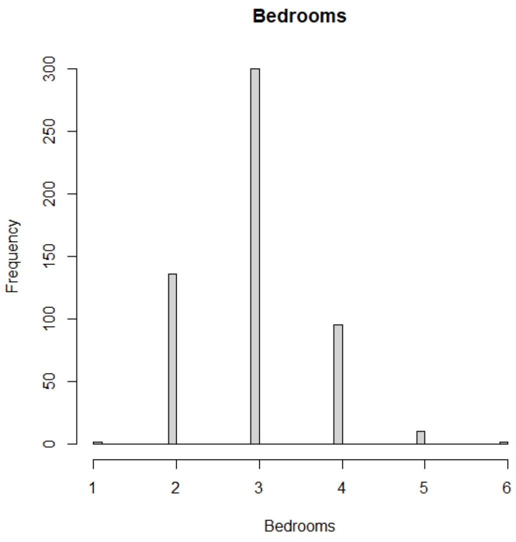


Figure 5: Histogram of bedrooms

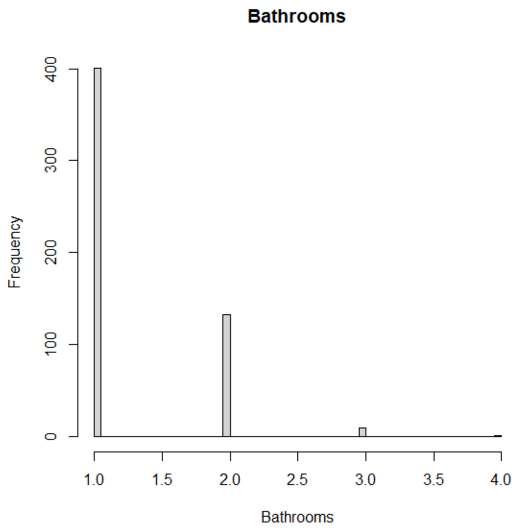


Figure 6: Histogram of bathrooms

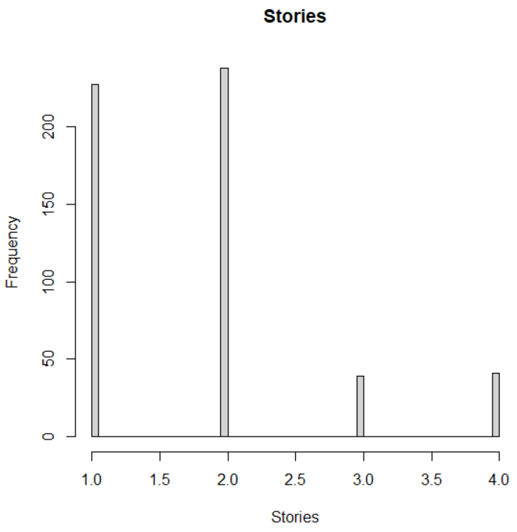


Figure 7: Histogram of stories

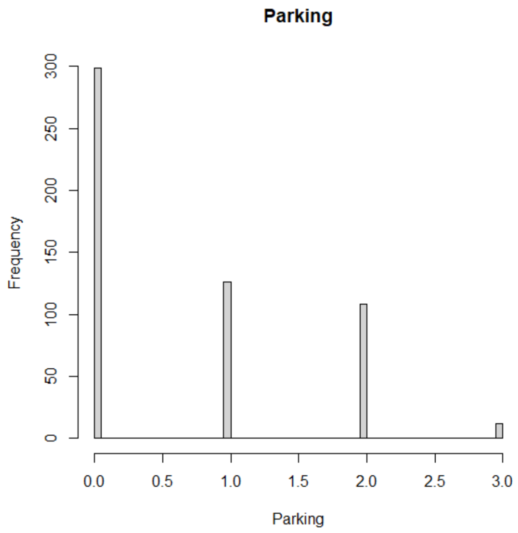


Figure 8: Histogram of parking

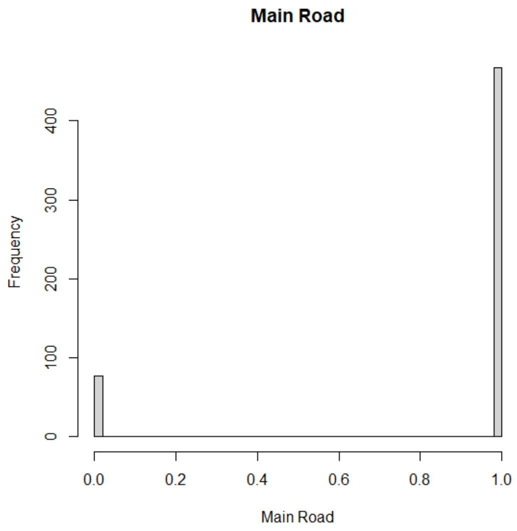


Figure 9: Histogram of main road

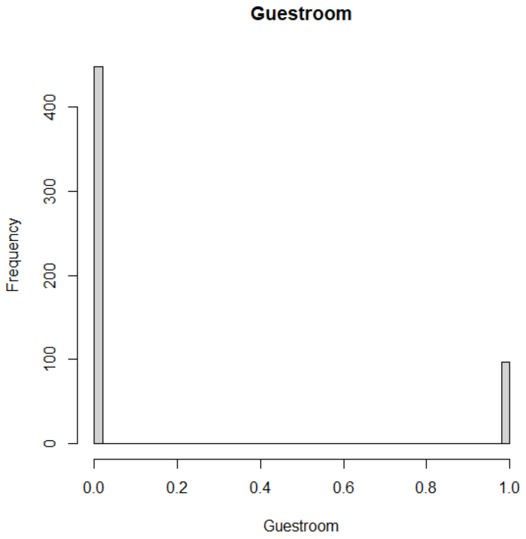


Figure 10: Histogram of guestroom

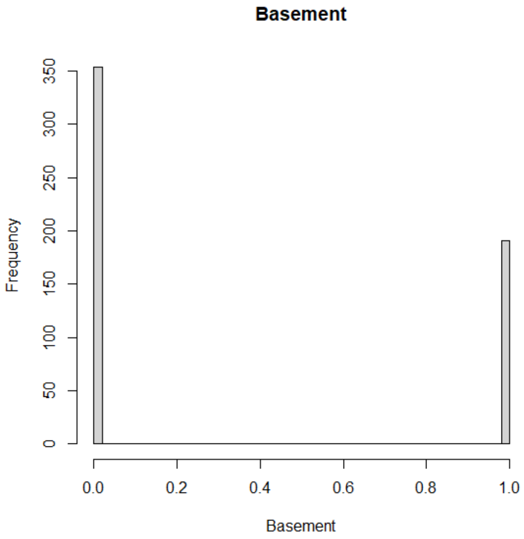


Figure 11: Histogram of basement

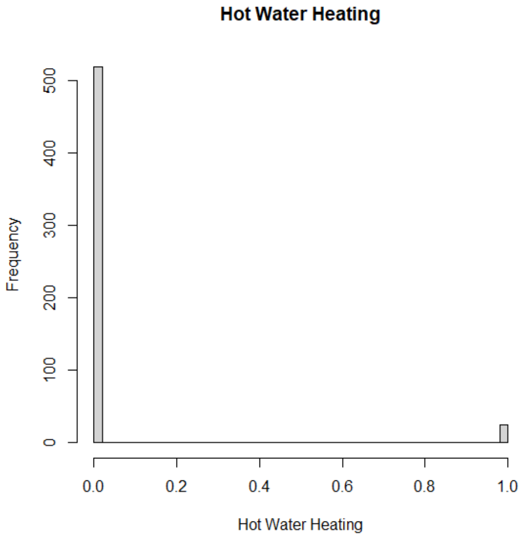


Figure 12: Histogram of hot water heating

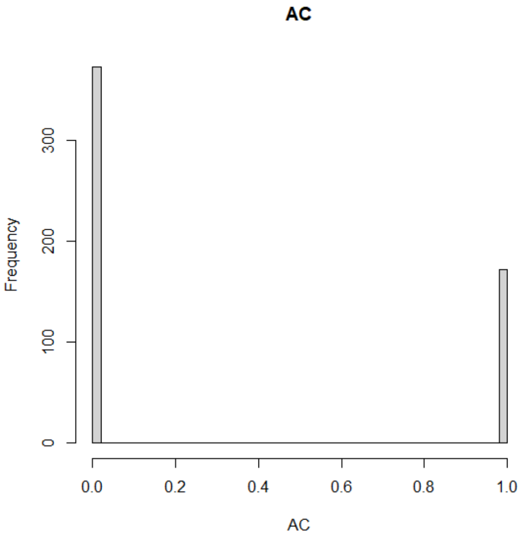


Figure 13: Histogram of AC

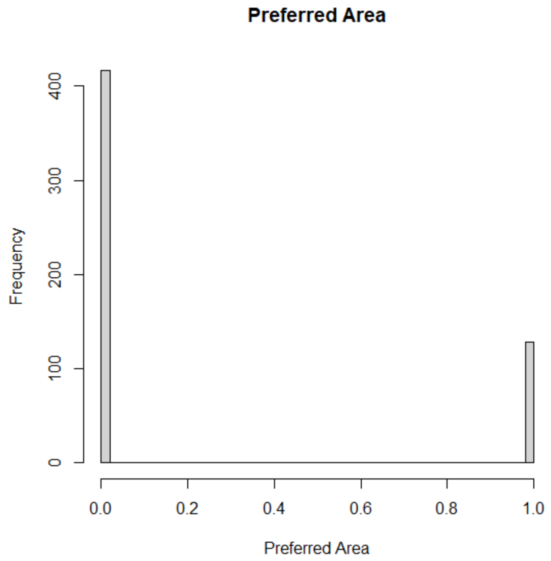


Figure 14Histogram of preferred area

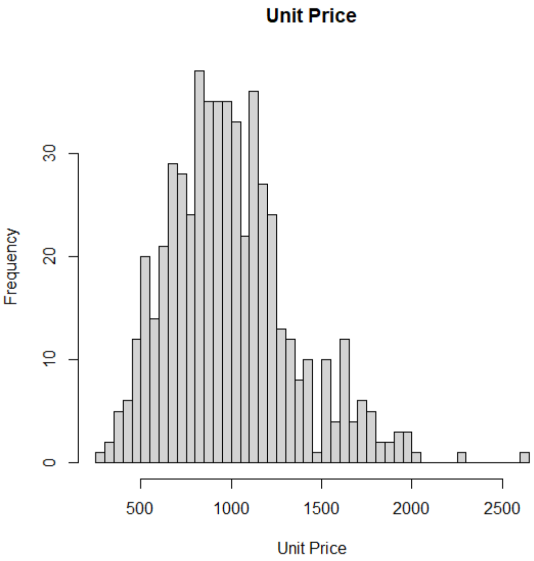


Figure 15: Histogram of unit price

Appendix B

Variable	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6	Component 7	Component 8
Price	0.317	0.253	0.078	0.207	0.058	0.054	0.329	0.105
Area	0.148	0.458	-0.192	0.294	0.061	0.076	0.321	-0.398
Bedrooms	0.355	-0.231	0.239	0.176	0.317	0.016	-0.392	-0.366
Bathrooms	0.353	-0.114	0.203	0.384	0.158	0.284	0.224	0.540
Stories	0.515	-0.170	0.103	-0.365	0.044	0.137	0.007	-0.133
Parking	0.110	0.260	-0.350	0.496	-0.175	-0.139	-0.418	-0.042
Main road	-0.024	0.373	-0.318	-0.362	0.078	0.574	-0.173	0.272
Guestroom	-0.241	0.098	0.443	0.052	-0.308	0.417	0.256	-0.425
Basement	-0.400	0.070	0.435	0.283	0.041	0.008	-0.191	0.298
Hot Water	-0.143	-0.478	-0.416	0.141	-0.013	-0.082	0.476	0.014
AC	0.301	0.245	0.217	-0.209	-0.548	-0.458	0.097	0.205
Preferred Area	-0.139	0.346	0.138	-0.188	0.657	-0.391	0.207	-0.008

Table 6: PCA correlation matric

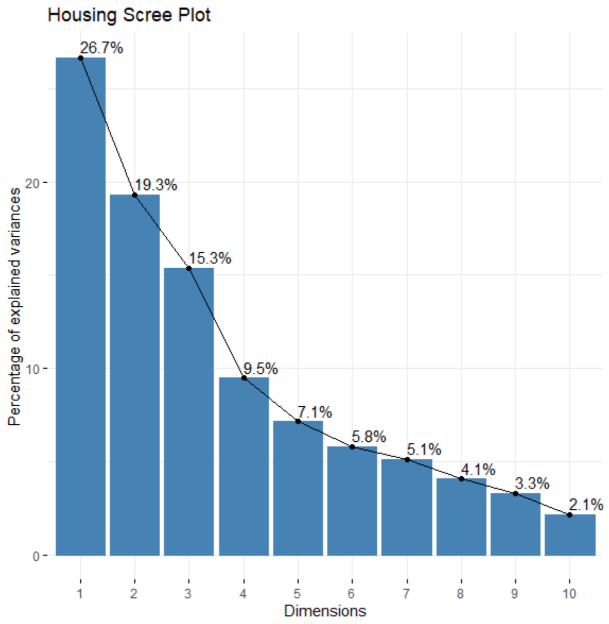


Figure 16: PCA scree plot

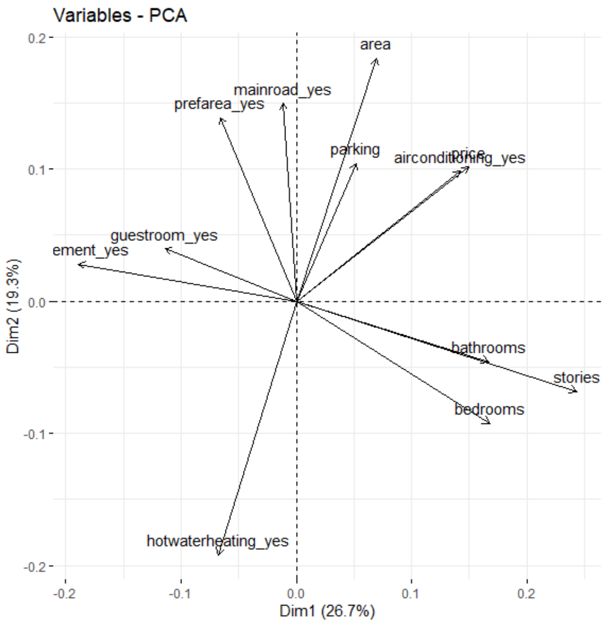


Figure 17: PCA biplot

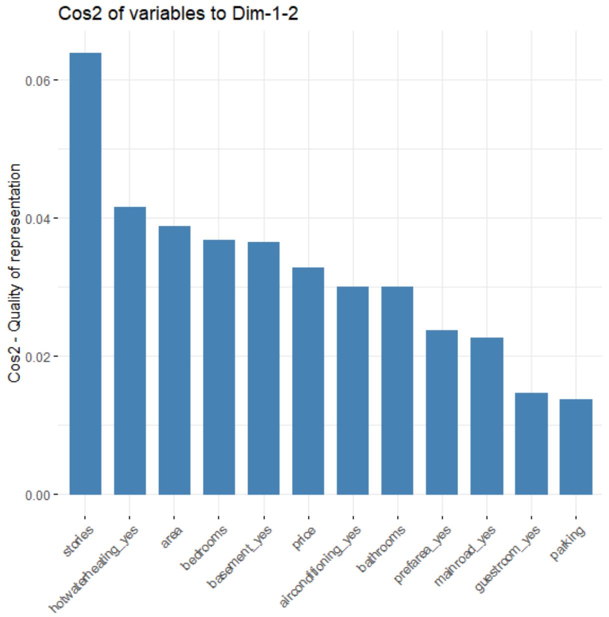


Figure 18: PCA Cos2 plot

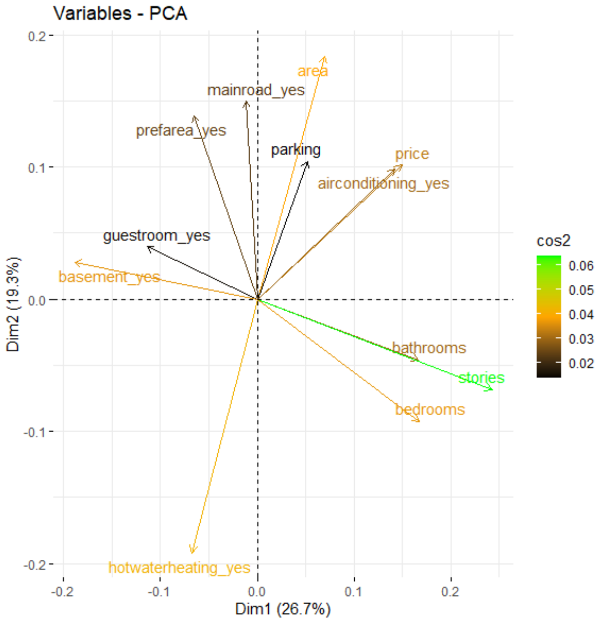


Figure 19: PCA biplot Cos2 plot hybrid