Predicting California Housing Price using Regression Analysis

Rekha Mani
California State University, East Bay
June 2018

I Introduction

Buying or selling house in California can be exciting, but it is an intimidating process. Buyer wants to know whether the price of house is rational or not. A seller wants to forecast the demand and select rational price so seller can sell the house quickly. There are many factors that influence the housing price, but we do not know what factors will influence the price of the house and to what extent these factors will influence the price. The purpose of this paper is to predict the price of the house in California using regression techniques and R software. Through this paper, we attempt to develop relatively good regression equation for predicting the price of the house in California based on the data set collected from San Luis Obispo County in 2009. We also determine the factors influencing the housing price and to what extent they affect the price.

II Data Description

Dataset used in this paper is a collection of real estate listings from San Luis Obispo Country, California, and some locations around it from the year 2009. For more information go to following link: https://www.statcrunch.com/5.0/index.php?dataid=2188686. Data set contains 781 observations and 8 variables that are listed below.

Variables	**Description**
Price	Price of house in thousand dollars
MLS#	Multi listing service ID assigned to each home in numbers
Location	Location from San Luis Obispo county in characters
Bedrooms	Number of Bedrooms in the house
Bathrooms	Number of Bathrooms in the house
SQFT	Size of the house in square feet
Price/SQFT	Price per square foot in dollars
Status	Short sale/Regular/Foreclosure in characters

Based on our research question, Price was selected as response variables and remaining seven variables are predictors.

2.1 Sample of six Observations from the dataset:

Price	MLS#	Location	Bedrooms	Bathrooms	SQFT	Price/SQFT	Status
795000	132842	Arroyo Grande	3	3	2371	335.3	Short Sale
399000	134364	Paso Robles	4	3	2818	141.59	Short Sale
545000	135141	Paso Robles	4	3	3032	179.75	Short Sale
909000	135712	Morro Bay	4	4	3540	256.78	Short Sale
109900	136282	Santa Maria-Orcutt	3	1	1249	87.99	Short Sale
324900	136431	Oceano	3	3	1800	180.5	Short Sale

2.2 Data Summary:

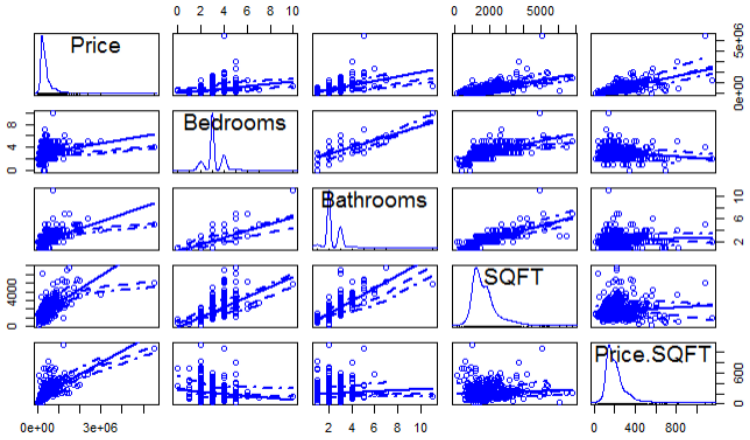
```
      MLS.                      Location       Price
 Min.   :132842   Santa Maria-Orcutt:278   Min.   :  26500
 1st Qu.:149922   Paso Robles       :100   1st Qu.: 199000
 Median :152581   Atascadero        : 68   Median : 295000
 Mean   :151225   Arroyo Grande     : 40   Mean   : 383329
 3rd Qu.:154167   Nipomo            : 37   3rd Qu.: 429000
 Max.   :154580   Lompoc            : 27   Max.   :5499000
                  (Other)           :231
    bedrooms        bathrooms          sqft      price_per_sqft
 Min.   : 0.000   Min.   : 1.000   Min.   : 120   Min.   :  19.33
 1st Qu.: 3.000   1st Qu.: 2.000   1st Qu.:1218   1st Qu.: 142.14
 Median : 3.000   Median : 2.000   Median :1550   Median : 188.36
 Mean   : 3.142   Mean   : 2.356   Mean   :1755   Mean   : 213.13
 3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.:2032   3rd Qu.: 245.42
 Max.   :10.000   Max.   :11.000   Max.   :6800   Max.   :1144.64

         status
 Foreclosure:162
 Regular    :103
 Short Sale :516
```

Based on the above data summary, there are no missing values in the data set. We can proceed further with correlation/visual data analysis.

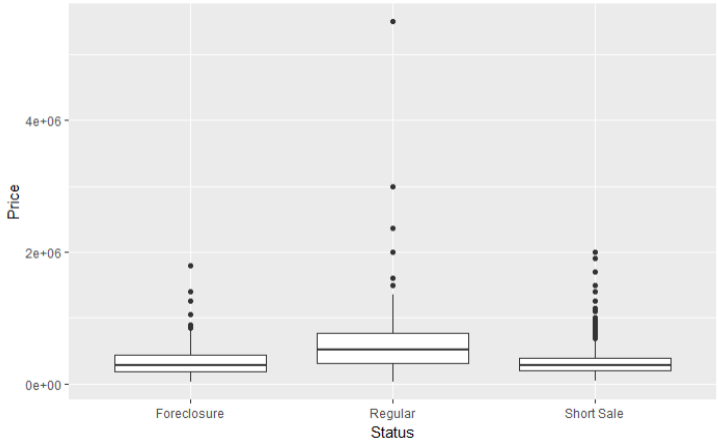
2.3 Correlation:

	MLS.	Price	Bedrooms	Bathrooms	SQFT	Price.SQFT
MLS.	1.000000000	-0.0170353	0.005199098	-0.02062101	-0.07294926	0.01085974
Price	-0.017035296	1.0000000	0.253162417	0.52011005	0.66472361	0.68280273
Bedrooms	0.005199098	0.2531624	1.000000000	0.58837048	0.59707015	-0.15338743
Bathrooms	-0.020621009	0.5201100	0.588370476	1.00000000	0.76154262	0.07963340
SQFT	-0.072949259	0.6647236	0.597070150	0.76154262	1.00000000	0.09842064
Price.SQFT	0.010859743	0.6828027	-0.153387428	0.07963340	0.09842064	1.00000000

2.4 Scatterplot Matrix



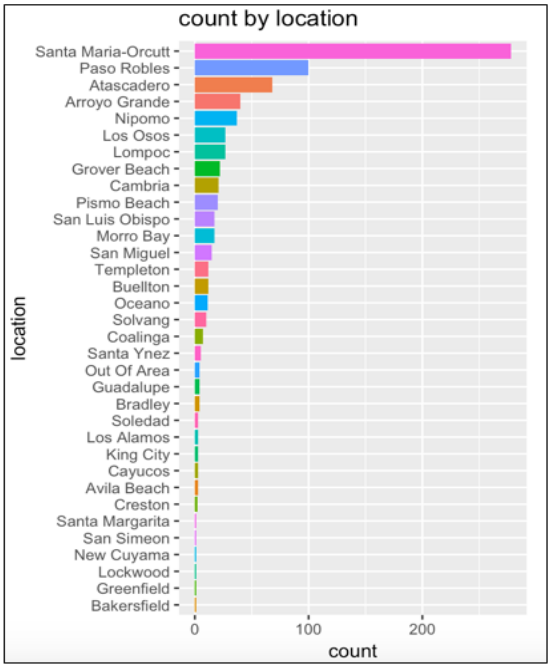
2.5 Boxplot



Following information drawn by observing the correlation, scatterplots and Boxplots:

- Correlation coefficient between Price and Bedroom is 0.25 and the scatterplot shows the positive relationship
- Correlation coefficient between Price and Bathroom is 0.52 and the scatterplot shows the positive relationship
- Correlation coefficient between Price and SQFT is 0.66 and it shows significant positive relationship
- Correlation coefficient between Price and Price per sqft is high (0.68)
- When trying to predict price, the features needs to be independent of the price. Price per sqft is dependent on the Price and should not considered as predictor so excluded Price per sqft variable.
- MLS# is an ID that will not contribute to our prediction so excluded MLS#
- Boxplot between Price and Status shows that houses with Foreclosure and Short sale are in the lower Price range while house with Regular status are ranges from lower to higher price range. All of them have outliers.

2.6 Bar graph showing number of observations in cities



- Based on the graph, data is sparse for some location so removing observation with low frequency to avoid errors that can result when the test set contains categorical data not seen before or the model is not optimal
- After analyzing the data completely, We consider Price as response variable and following variables: Bedrooms, Bathrooms, Sqft, Location, Status as potential predictors for designing the regression model

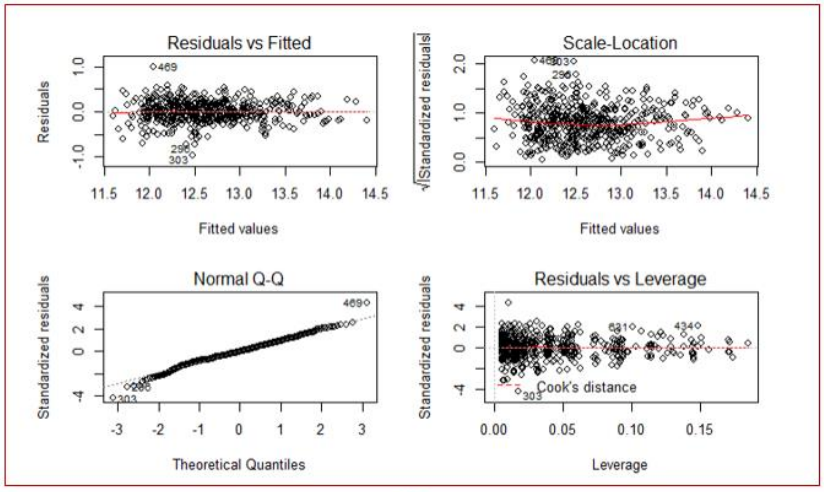
III Statistical Analysis and Result:

3.1 Models

No	Model form	R^2	Non-significant	Outlier removal	AIC	Regression tests Results
1	Price ~ bedrooms +sqft+ bathrooms + status +Location	62%	bathrooms	No	16153.38	Fail
2	Price ~ bedrooms + sqft + Location+Status	62%		No	16151.94	Fail
3	Price~bedrooms+sqft+Location+Status+sqft:bedrooms	62%	sqft:bedrooms	No	16152.03	Fail
4	log(Price) ~ log(sqft) + Location + bedrooms+Status	74%		No	354.42	Log transform -Linearity (not passed) -homoscedasticity(Passed)
5	log(Price) ~ log(sqft) + Location + bedrooms+Status	83%		Yes	-27.27	Log trans and outlier removed - Linearity(Passed) - Homoscedasticity(Not Passed)

As shown above, We started with full additive model(Model 1) and noticed that bathrooms were insignificant. Also Linearity and Non constant variance test failed for Model 1. So proceeded with Model 2 by removing Bathroom variable. Although all variables returned significant, linearity and Non constant variance test failed for Model 2. Next, we tried the interaction model 3 by adding Bedroom and sqft as interaction term. Interaction term was not significant and all regression test failed for Model 3. So removed the interaction term and did log transformation on Model 4. R-Squared value improved to 74% however Linearity failed although non-constant variance test passed. In addition to log transformation, we removed the outliers and ran the model 5. R-Squared increased to 84% and AIC value got lesser. Considering this as final model, proceeded with Diagnostic test on Model 5

3.2 Diagnostic results for Final model 5:




```
        Shapiro-Wilk normality test

data:  train_ro$residual
W = 0.98735, p-value = 0.0001298
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 8.051771    Df = 1     p = 0.004545919
```

Following observed from the diagnostics results for Model 5:

- Linearity: The relationship between X and the mean of Y is linear
- Homoscedasticity: Based on the p-value (0.0045) from non-constant variance test, there is indeed heteroscedasticity
- Normality: Based on the QQ plot, Y is normally distributed for any fixed value of X
- Independence: Observations does appear independent.
- It does seem to have outliers

3.3 Model Performance

To evaluate the final model, we ran the model using the test data and obtained 25% of mean error rate, which is low. This indicates that estimated regression model is valid. To confirm this further, we ran the model on the entire data set and obtained similar results.

3.4 Statistical summary of final model (Using entire data set):

```
Call:
lm(formula = log(Price) ~ log(SQFT) + Status + Bedrooms + Location,
    data = entire_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95312 -0.13164 -0.00237  0.14239  0.99955

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 5.955098   0.211504  28.156  < 2e-16 ***
log(SQFT)                   0.952673   0.031700  30.053  < 2e-16 ***
StatusRegular               0.176606   0.035202   5.017 6.77e-07 ***
StatusShort Sale            0.021019   0.023000   0.914 0.361116
Bedrooms                    0.001078   0.015551   0.069 0.944736
LocationAtascadero         -0.295197   0.047266  -6.245 7.61e-10 ***
LocationBuellton           -0.165672   0.078155  -2.120 0.034400 *
LocationCambria             0.264007   0.074611   3.538 0.000431 ***
LocationGrover Beach       -0.105476   0.062168  -1.697 0.090241 .
LocationLompoc             -0.576129   0.058922  -9.778  < 2e-16 ***
LocationLos Osos           -0.308943   0.062181  -4.968 8.63e-07 ***
LocationMorro Bay           0.187059   0.073005   2.562 0.010621 *
LocationNipomo             -0.329615   0.055967  -5.889 6.20e-09 ***
LocationOceano             -0.195971   0.119440  -1.641 0.101331
LocationPaso Robles        -0.388784   0.044160  -8.804  < 2e-16 ***
LocationPismo Beach         0.259957   0.066639   3.901 0.000106 ***
LocationSan Luis Obispo     0.053353   0.068015   0.784 0.433074
LocationSan Miguel         -0.501112   0.073376  -6.829 1.95e-11 ***
LocationSanta Maria-Orcutt -0.653039   0.040824 -15.996  < 2e-16 ***
LocationSolvang            -0.102869   0.088279  -1.165 0.244334
LocationTempleton          -0.047269   0.078541  -0.602 0.547495
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2251 on 654 degrees of freedom
Multiple R-squared:  0.8388,    Adjusted R-squared:  0.8339
F-statistic: 170.2 on 20 and 654 DF,  p-value: < 2.2e-16
```

Following observed from the statistical summary of final model

- Following independent variables: log (Sqft), Status (Regular), some of the locations are significant.
- With p-value being 2.2e-16, estimated regression model is significant
- R-Square value is fairly high
- Residual standard error is low

IV Discussions

Regression model 5 that was log transformed and outlier removed with following predictors: SQFT, Bedrooms, Location and Status passed normality test including cross validations performed on split data (train and Test) and entire data set. Based on this, we can conclude that this to be estimated regression model for predicting housing price in California. Conclusions cannot be generalizable because Homoscedasticity failed and data set was recorded from the year 2009, which is recession period. Alternative to fix the Heteroscedasticity would be applying weighted least square regression model. In addition, if we have data set that represent all the California locations and obtained in the recent years, we can fine tune the model and provide better model for estimating the housing price of California.

V Reference link: https://www.statcrunch.com/5.0/index.php?dataid=2188686

VI Appendix


Rcode.Rmd


HTML output.zip