Predicting California Housing Price using Regression Analysis 1

**Predicting California Housing Price using Regression Analysis**

Predicting California Housing Price using Regression Analysis 2

**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 |

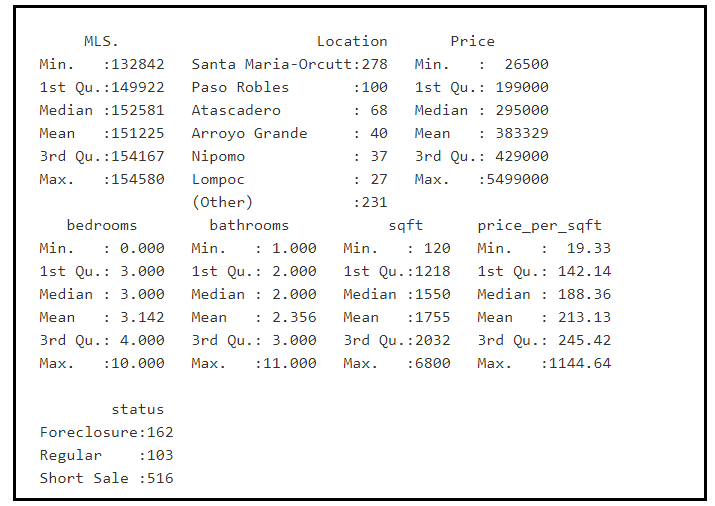
Based on our research question, Price was selected as response variables and remaining seven variables are predictors.

Predicting California Housing Price using Regression Analysis 3

**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 |

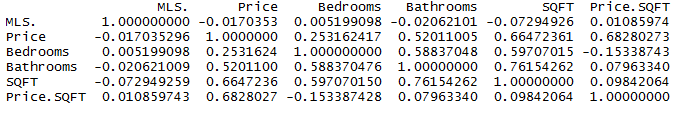
**2.2** **Data Summary**:



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

Predicting California Housing Price using Regression Analysis 4

**2.3 Correlation**:

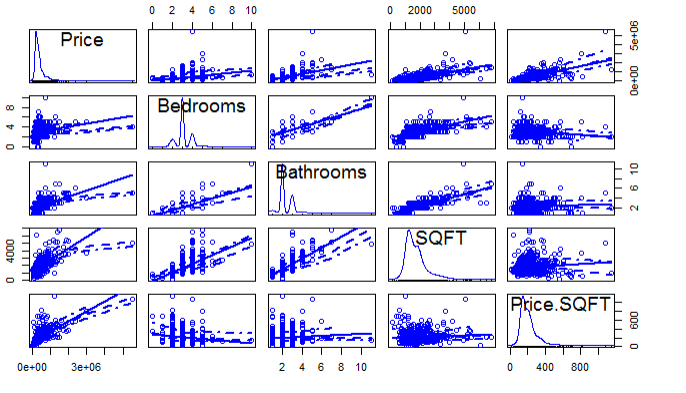


Following information drawn by observing the correlation, scatterplots and Boxplots:

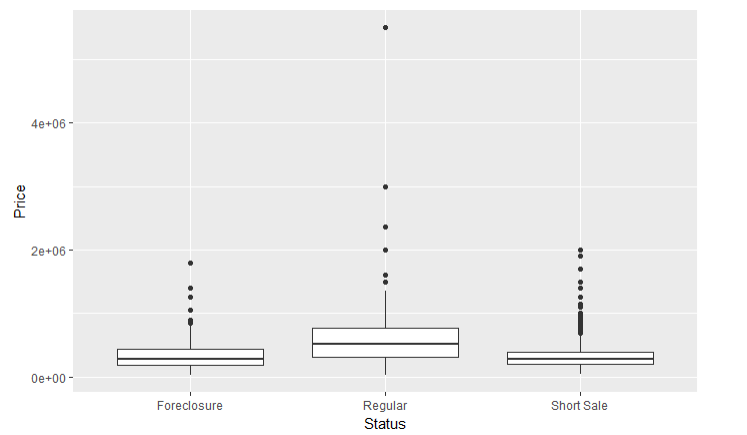
* Correlation coefficient between Price and Bedroom is 0.25 and the scatterplot shows the positive relationship
* Correlation coefficient between Price and Bathroom is 0.52 and the scatterplot shows the positive relationship
* Correlation coefficient between Price and SQFT is 0.66 and it shows significant 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.

**2.4 Scatterplot Matrix**

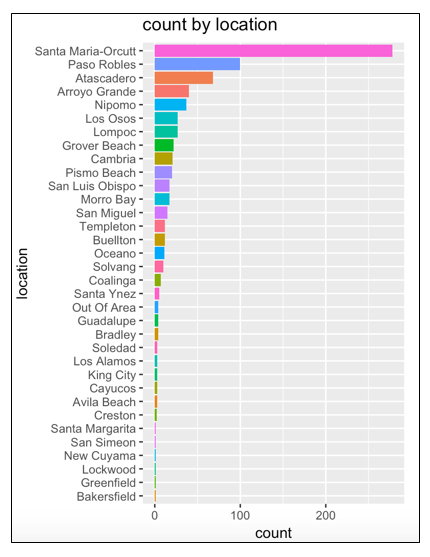


**2.5 Boxplot**



Predicting California Housing Price using Regression Analysis 5

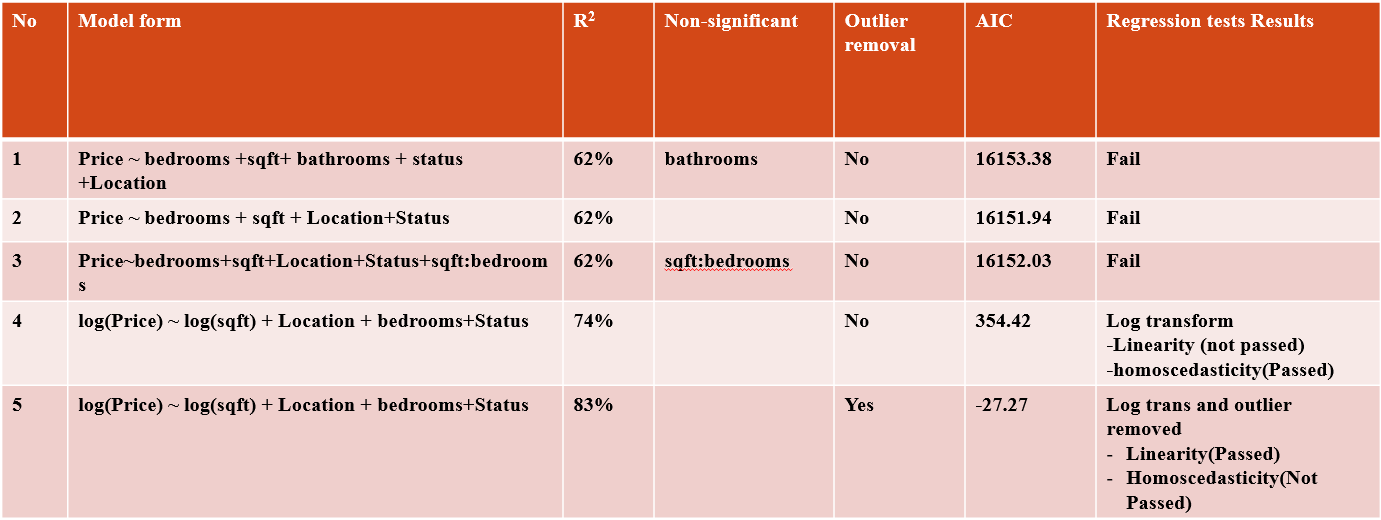
**2.6** **Bar graph showing number of observations in cities**



* 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**



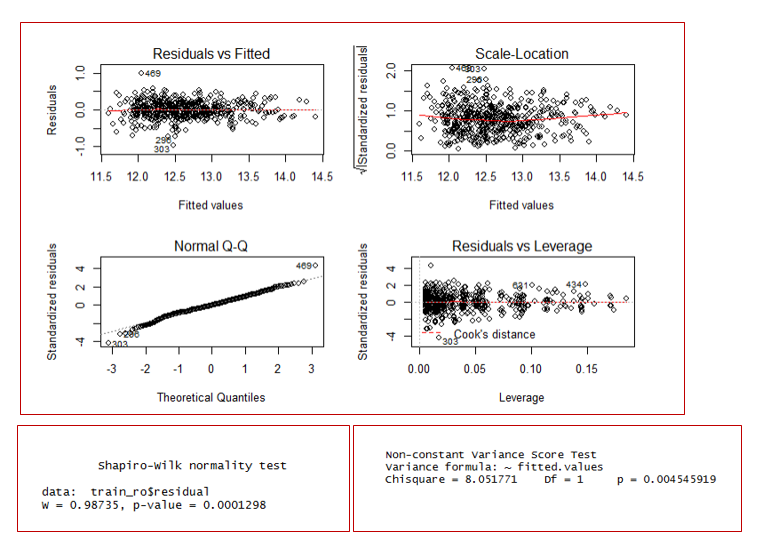
Predicting California Housing Price using Regression Analysis 6

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**:

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

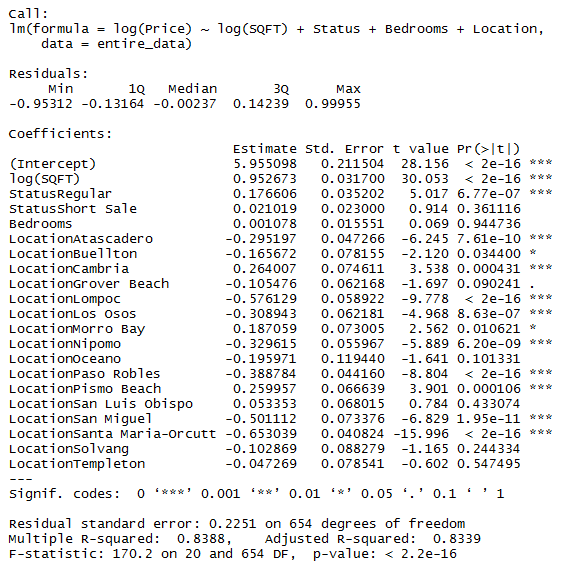


Predicting California Housing Price using Regression Analysis 7

**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):**



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**

****