Multiple Regression Analysis of Real Estate Listings on Redfin

A Data Analysis Project using Python and SQL

Jessie Lenarz

March 17, 2022

Introduction

If you have ever bought or sold a home, you may have wondered how exactly the listing price was set. We would like to predict the list price of a home based on features of the home in order to set a realistic list price for a home we would like to place on the market. We will use public data from the Redfin website to create a predictive model and determine the list price of a recently sold home.

Methodology

Data Sources and Collection

Data was collected by using the search feature at Redfin.com for all property listings in Dakota County, Minnesota on March 15, 2022. The data set was exported as a csv and contained information including address, property type, square footage, number of bedrooms, number of bathrooms, HOA fees, year built, and so on.

Data Cleaning

The data set was imported to IBM's DB2 database and queried using SQL to determine basic structure of the set. One of the features, property type, had a value of "other" for two of the properties. After further investigation, it was determined that both properties had been mis-classified. One was a townhome and the other was a co-op/condo. The values were corrected. Other queries showed a large range in square footage, lot size, year built, and price.

One property did not have location data and was dropped from the data set. Fifty-two properties did not have a lot size listed; the values were replaced by the mean lot size. One hundred seventy-three properties did not have an HOA/month fee listed. Our assumption is that those properties are not in an HOA and the value was set to zero for any missing values.

The following features were dropped from the data set:

- "sale type" – was MLS for all properties
- "State or Province" – was MN for all properties
- "$/square feet" – since price is our target, we will not include
- "source" – was MLS for all properties
- "url" – did not provide any useful information for the model

The remaining features were then analyzed to determine which features should be used in creating a multiple regression model.

Exploratory Data Analysis with Data Visualization

The seaborn library was used to create a heat map (see Figure 1) of the quantitative variables in the cleaned data set.

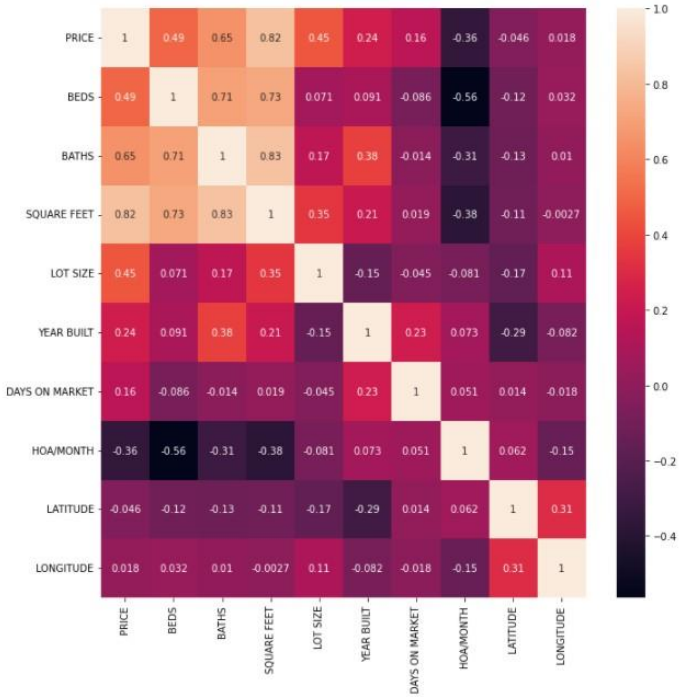

Figure 1: Heatmap of ten quantitative variables in dataset

Based on the heat map, we identified features that are likely related to price: number of beds, number of baths, square feet, lot size, year built, and HOA fees per month. Scatter plots for price versus each of those variables were created. All showed at least a weak correlation (positive for all but HOA fee, which was negatively correlated). A selected sample are shown below in Figure 2. All scatterplots are available in the appendix.

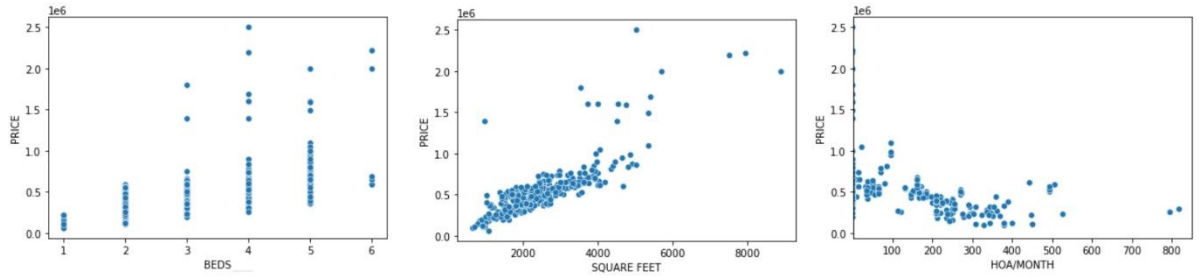

Figure 2: Scatterplots of Number of Bedrooms, Square Feet, and HOA fees vs Price

Boxplots were created to examine the possible relationship of categorical variables to price. Property type and city were investigated and appear in Figure 3. Both appear to be related to the price, but not the presence of several outliers.

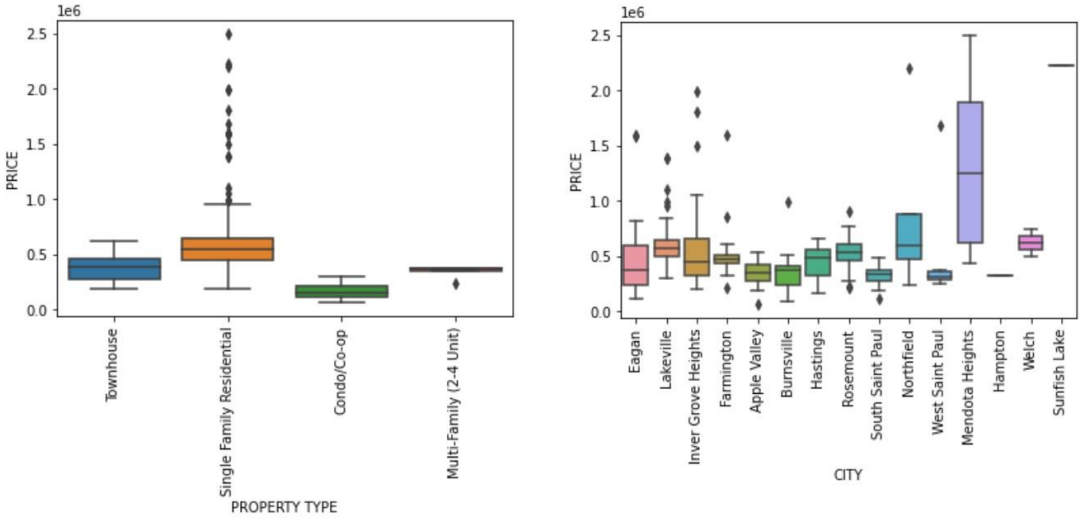

Figure 3: Boxplots for Price by Property Type and City

Based on the database queries, heat map, and charts, the following feature set was determined:

Feature	**Description**
BEDS	Number of bedrooms
BATHS	Number of bathrooms
SQUARE FEET	Square footage of home
LOT SIZE	Size of lot (in square feet)
YEAR BUILT	Year Built
DAYS ON MARKET	Number of days since first listed
HOA/MONTH	HOA fees per month
PROPERTY TYPE	Type of structure: Single-family, townhouse, condo,multi-family
LOCATION	Subdivisions within cities

It was reasonable to suspect that we have clusters in our data set, so a clustering analysis was performed using Scikit Learn in Python. After scaling the data, k-means and DBSCAN were used to determine potential clusters. For the k-means clustering, we determined the number of clusters by looping through all values from one to fifteen. The elbow method indicated that the appropriate number of clusters was five, as shown in Figure 4.

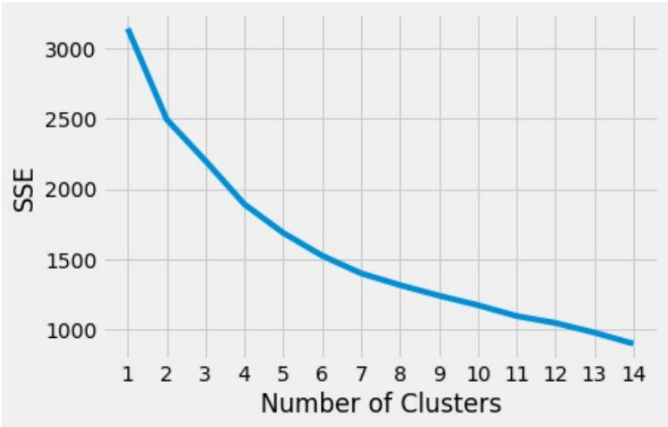

Figure 4: Elbow graph for k-means algorithm

Using five clusters in the k-means algorithm, we had the following results. Note that most properties were in a single cluster.

Cluster Number	Number of Properties
0	338
1	1
2	7
3	2
4	1

Using the DBSCAN algorithm with an epsilon value of 0.5, we had the following results.

Cluster Number	Number of Properties
-1 (outliers)	324
0	13
1	6
2	6

Adjusting epsilon values all the way up to 2 maintained a majority of the points as outliers. Therefore, we determined it was reasonable to assume no clustering.

The data set was then divided into training and testing sets and a multiple regression model was fit on the nine selected features to predict the target variable, price, using the training set. A second multiple regression model was fit on seven of the features (with “Year built” and “Days on market”) using the training set. Both models were analyzed by making predictions based on the model for the testing set and computing the R-squared value, the mean square error, and the root mean square error.

Results

After cleaning the data and determining the data were not clustered, two separate multiple regression models were fit. The first multiple regression used all nine features and had an R-squared value of 0.56 with a root mean square error of 179,138.34, which is 33.6% of the mean price. The second multiple

regression had an R-squared value of 0.50 and a root mean square error of 189,456.13, which is 35.6% of the mean price. Both of these models seem to be less than ideal.

We reran the model on the entire data set and predicted the list price for each model on a recently sold single-family home in Lakeville with 4 bedrooms, 2 bathrooms, 2010 square feet, a 10890 square foot lot, built in 1977 that is not part of an HOA. The model using all nine features predicted a list price of $399,636.96 while the model using only seven features predicted a list price of $444,093.43. The actual list price was $363,010. Based on all of this information, the model with nine features is the better of the two, but it still does not perform well.

Discussion

The model could be improved by identifying further features that may impact list price. Some of those features may include school district, distance to nearest grocery store, distance to nearest park, home energy costs, high speed internet access availability, last known renovations, garage size/style, deck, pool, or fencing. We should also take a second look at the outliers that were identified in each community. It may warrant not including those values in the model. Another possible adjustment would be to replace missing square footage in the data set with the median square footage rather than the mean square footage since there appear to be outliers in the set.

Conclusion

We applied multiple regression algorithms to our dataset after determining there was no clustering to arrive at a model that moderately predicts list price. The model could be improved by the inclusion of more features or by removing outliers from the data set.

Appendix

Code and data files can be accessed at https://github.com/jklenarz/Real-Estate

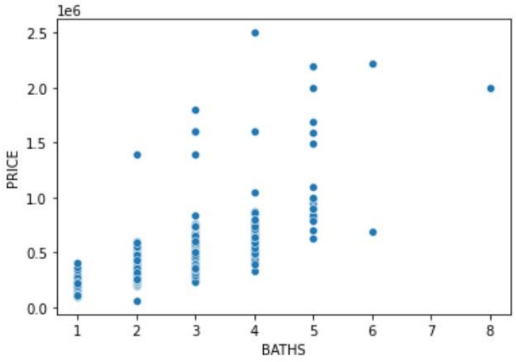

Figure A.1: Number of Bathrooms vs. Price

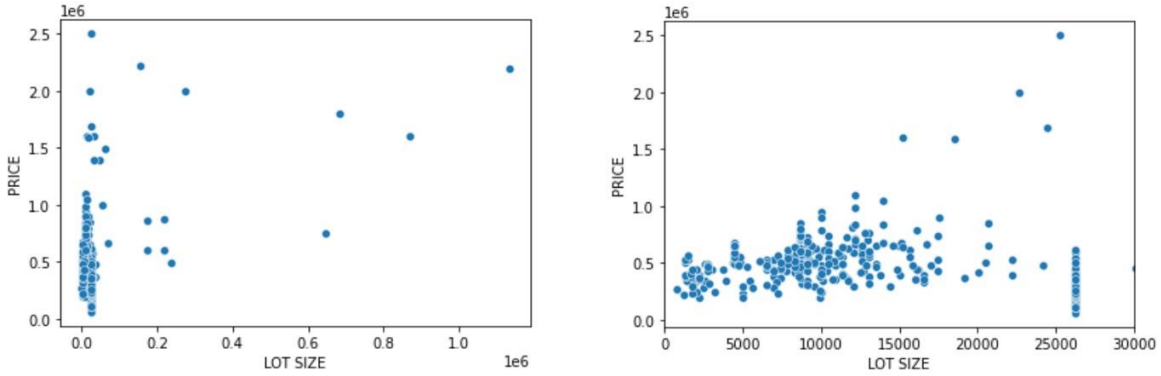

Figure A.2: Lot Size vs Price

Left: all data; Right: Lot Size <30000 sq. ft

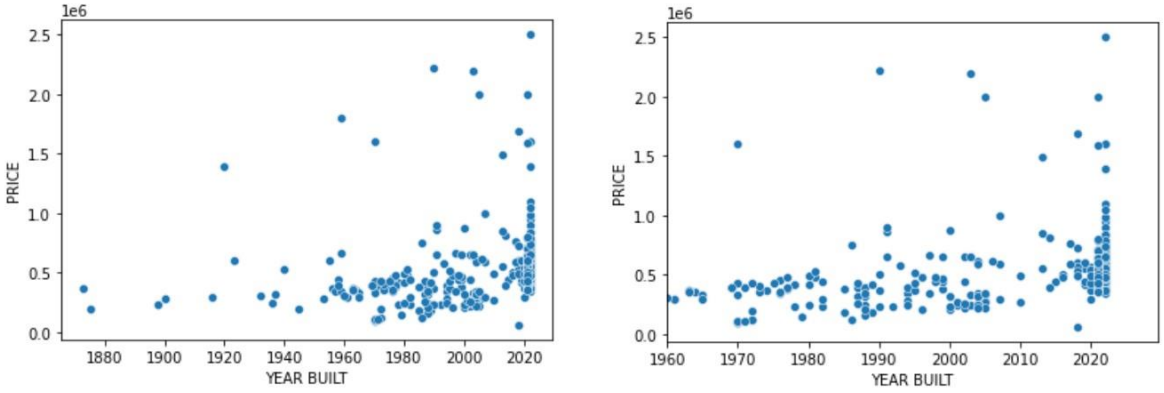

Figure A.3: Year Built vs. Price

Left: all data; Right: Year Built > 1960