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 the home's list price is determined. Is the list price determined by location, number of bedrooms, or some other feature or combination of features? We want to predict the list price of a home based on the home's features in order to set a realistic list price. 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

We collected data 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 file and contained information on 351 properties including address, property type, square footage, number of bedrooms, number of bathrooms, HOA fees, year built, etc.

Data Cleaning

The data set was imported to IBM's DB2 database and queried using SQL to determine the basic structure of the set. One of the features, property type, had a value of "other" for two of the properties. After further investigation, we determined that both properties had been misclassified. 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. We assume that those properties are not in an HOA, and the value was set to zero for any missing values.

We dropped the following features 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 helpful information for the model

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

Exploratory Data Analysis with Data Visualization

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

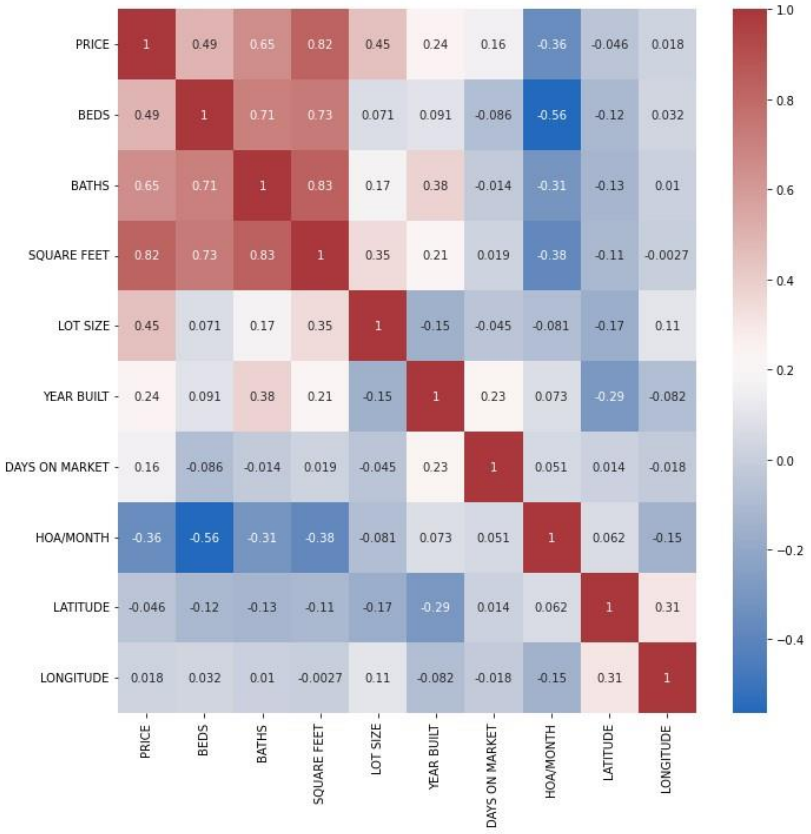

Figure 1: Heatmap of correlation of ten quantitative variables in our dataset

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. We created scatter plots for price versus each of those variables. 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 of the scatterplots are available in the appendix.

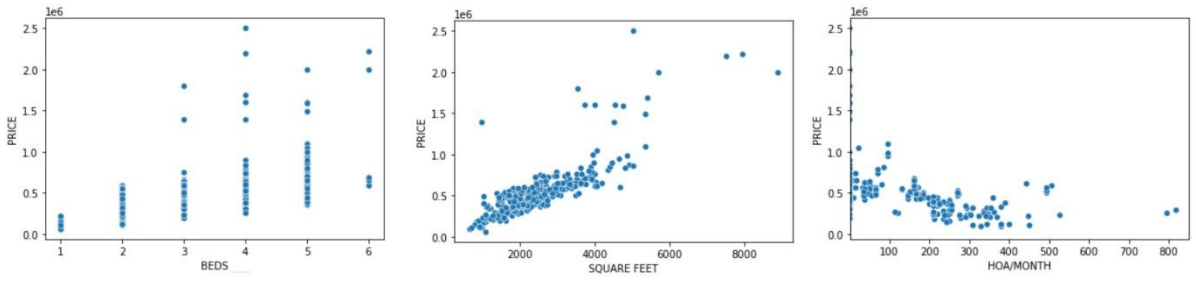

Figure 2: Scatterplots of Number of Bedrooms, Square Feet, and HOA fees vs. Price

We generated boxplots 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 there are a few things that stand out. Single family residential homes have a large number of outliers. Mendota Heights has a larger median and interquartile range (IQR) than the other cities. Ten of the thirteen cities have outliers; of those ten, Lakeville, Inver Grove Heights, and Farmington have three outliers.

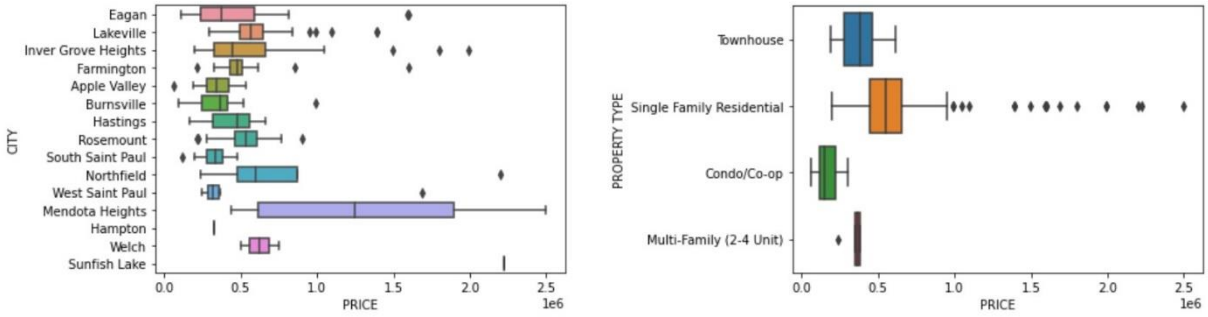

Figure 3: Boxplots for Price by Property Type and City

Based on the database queries, heat map, and charts, we decided upon the following feature set:

Feature	**Description**
BEDS	Number of bedrooms
BATHS	Number of bathrooms
SQUARE FEET	Square footage of home
LOT SIZE	Size of the 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 we performed a clustering analysis using Scikit Learn in Python. After scaling the data, we employed the k-means and DBSCAN algorithms to determine potential clusters. For the k-means clustering, we chose 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.

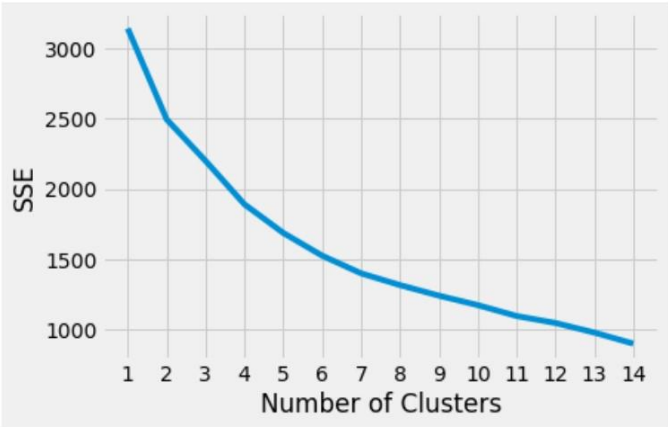

Figure 4: Elbow graph for the k-means algorithm

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

We had the following results using the DBSCAN algorithm with an epsilon value of 0.5.

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.

We then divided the data set into training (80% of the data set) and testing sets (20% of the data set). Using the training set, we fit a multiple regression model on the nine selected features to predict the target variable, price. We fit a second multiple regression model on seven of the features (with "Year built" and "Days on market" left out) 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, we fit two separate multiple regression models using the training set. The first multiple regression used all nine features and an

analysis using the test sest 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 an analysis using the test set had 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 four bedrooms, two 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

We could improve the model by identifying additional features that may impact list price. Some of those features may include the school district, distance to the nearest grocery store, distance to the 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 identified outliers in each community. It may warrant not including those values in the model. Another possible adjustment would be replacing missing square footage in the data set with the median square footage rather than the mean square footage since there appear to be outliers. Alternatively, we could use stochastic regression imputation in replacing the missing data values.

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 are available at https://github.com/jklenarz/Real-Estate

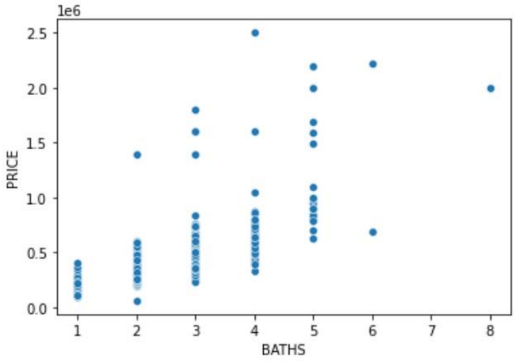

Figure A.1: Number of Bathrooms vs. Price

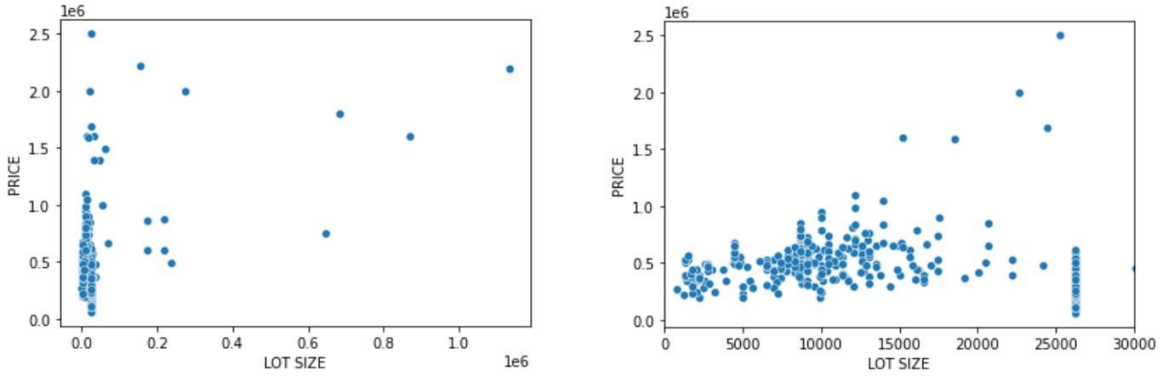

Figure A.2: Lot Size (in sq. ft) vs. Price

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

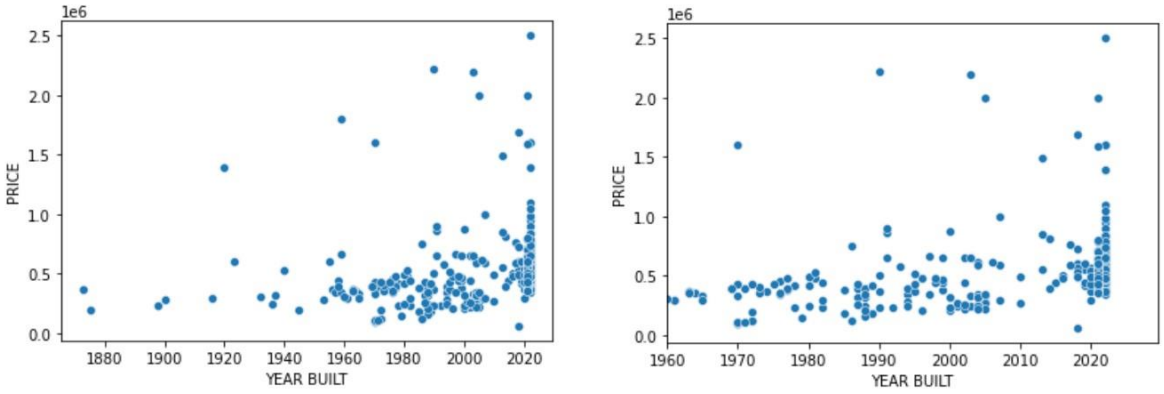

Figure A.3: Year Built vs. Price

Left: all data; Right: Year Built > 1960