**Predicting Housing Prices by City**

**Nick Miller**

**Bellevue University**

**Summer 2020**

**Abstract**

This project takes data from an API and runs it through exploratory data analysis to determine an equation that can estimate the value of a house, given some features of this house. The data comes from the Realtor.com API, and I have chosen to explore my hometown of Gainesville, FL. I have limited the results to 200 homes, so that each iteration of code does not consume too many computer resources.

**Introduction**

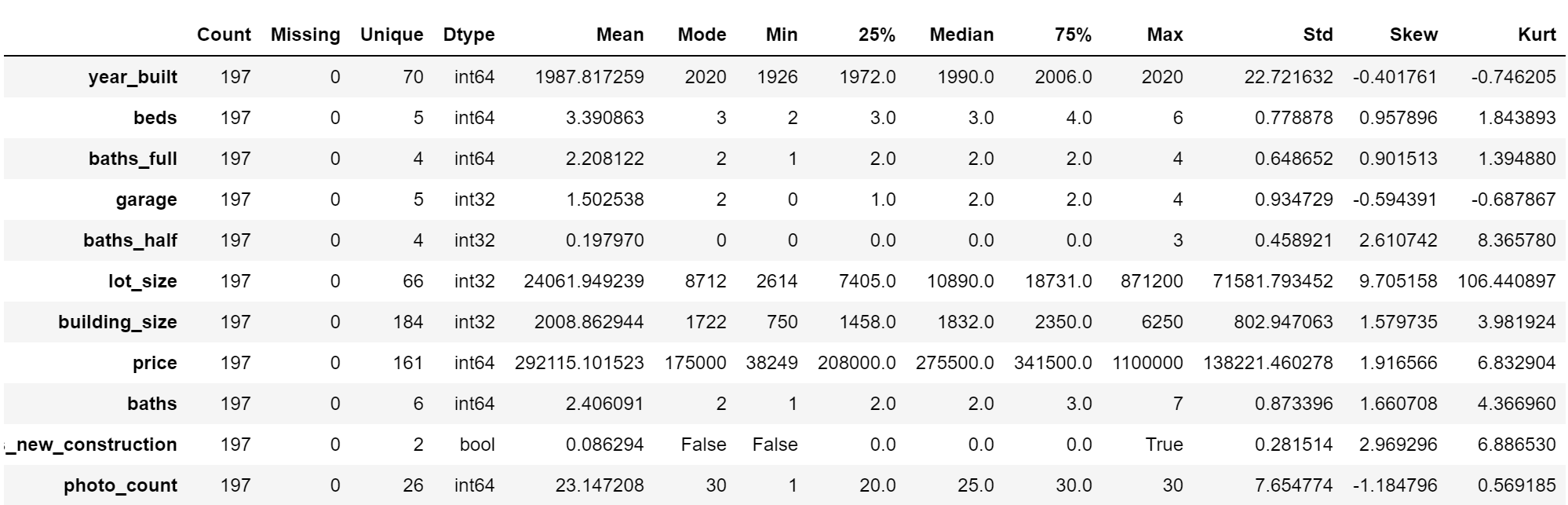
I had seen some examples of this type of project in my recommended readings during my master’s program, and thought it could be interesting. I was looking at buying a new house in the area, and selling my current home. At first, I thought it would only be helpful when purchasing a home, but later I found that it could help when selling as well. Many people reported that Zillow would be a good place to start when gathering data. After many frustrating hours, I believe that the Zillow API is now reserved for realtors, or paying customers. I finally stumbled upon the Realtor API, and it is incredibly easy to use. It even creates the query to use in Python or R. I used Python and Jupyter Notebook for my project. All data and code can be found in the Appendix.

**Data**

The data from the API contained many identifier columns, which were removed. I removed the property type column, since all of my houses are single family homes. I kept columns for year built, number of bedrooms, number of full bathrooms, number of half bathrooms, number of total bathrooms (full plus half), size of garage measured in cars, lot size in square feet, building size in square feet, price in dollars, number of photos in the listing, and if the house was new construction or not. All variables were integers, except the last one, which was a Boolean variable. The data looked good, but was missing some important features that I thought could be important, such as if the property has a pool, a homeowners association with amenities, or even solar panels for electricity. Also, there were a few rows that had some missing values, so I removed those. Also, I changed the entries that were blank to zeroes. I also saved the file to disk, so that I had the same data to work with throughout my project. Then, I had to trim some of the columns, since I wanted only numbers, and the size columns included units. Afterwards, I was ready to start my analysis.

**Univariate Analysis**

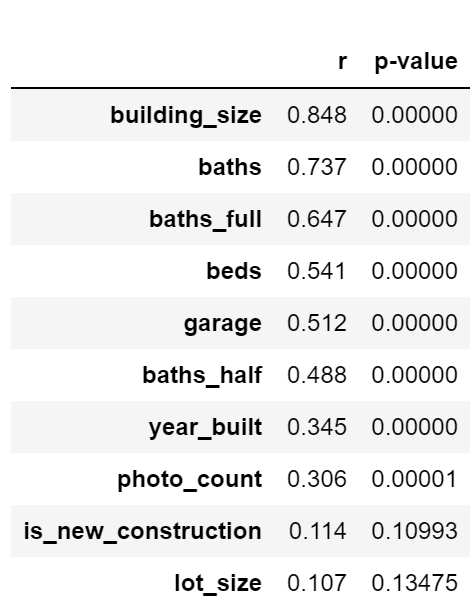
I wanted to explore each variable individually, to get a sense of the raw data, and to see if I could spot any irregularities.



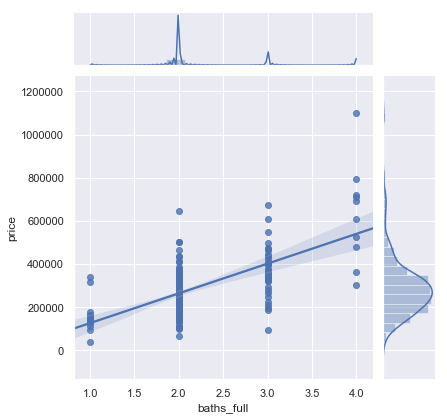
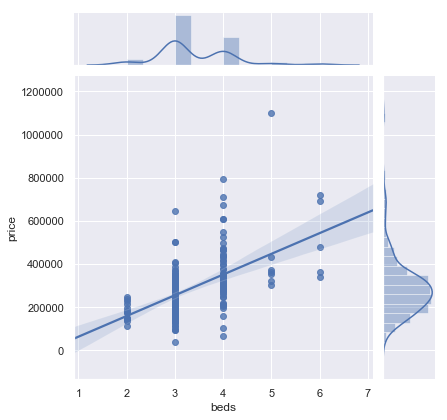
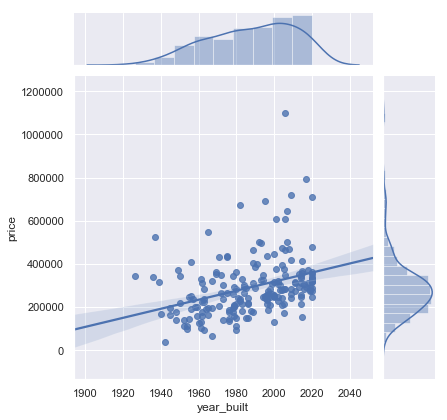
One variable, lot size, had a large skew and kurtosis. I would take note of this for later consideration. Another thing I noticed, was the lot size max was significantly larger than the other data. The max was also larger than the mode by a factor of exactly 100. It could be a typo, or it could be a large plot of land. Further analysis would have to be done. Now on to Bivariate Analysis.

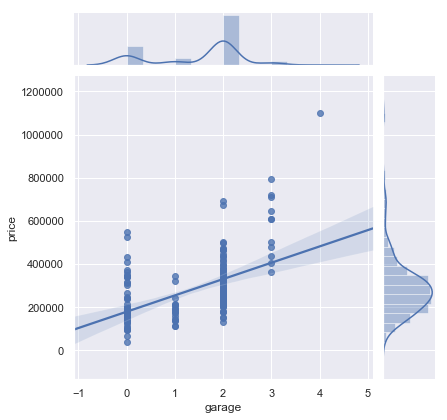
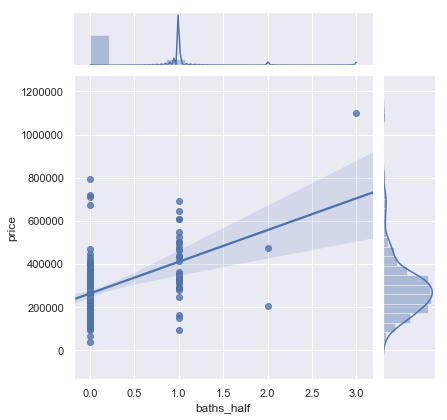
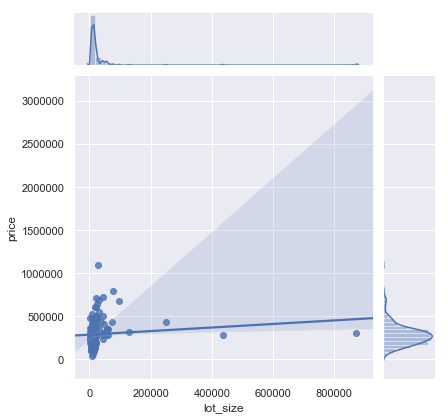
**Bivariate Analysis**

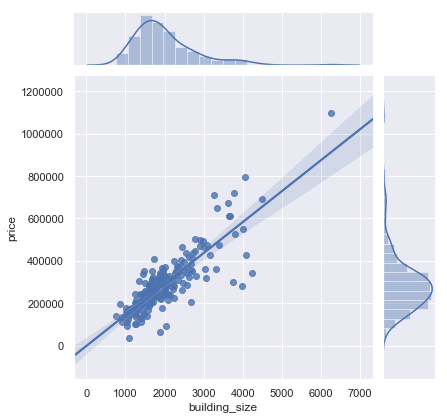
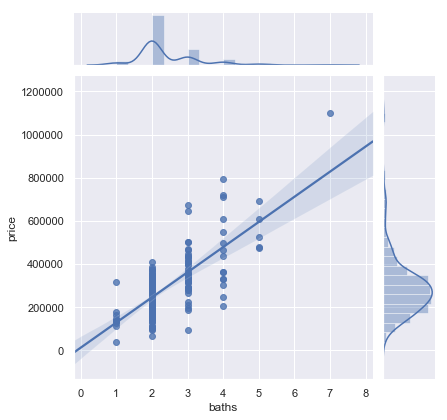
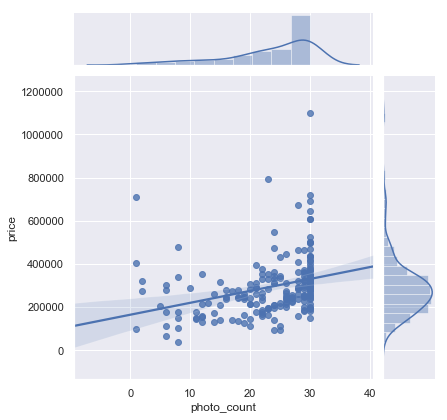
I used the Pearson Correlation test to measure the linear dependence between each independent variable and the price (dependent variable).



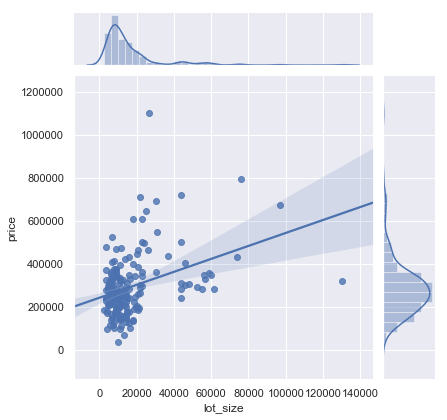
The results of this test are the correlation coefficient (r) and the significance level (p-value). Sorted in decreasing order, we can easily see the building\_size and baths variables have a high correlation with price, while lot\_size and is\_new\_construction have a low correlation. I have decided to drop the new construction column, since it does not appear to influence the price very much, and the p-value is above .05, which means it is not significant. Lot\_size, however, has a high p-value as well. But, before I drop this one, I wanted to explore it more with a scatter plot. Here are the scatter plots for all variables, with histograms.



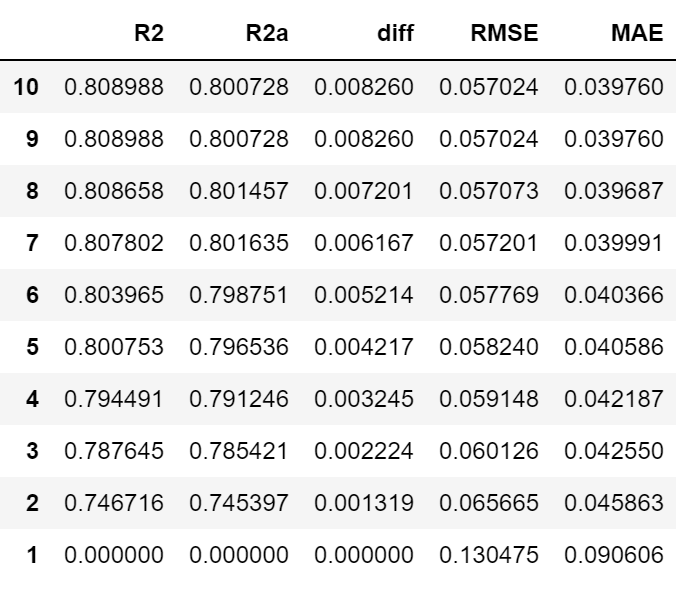
All of the histograms look fairly good, except for the lot\_size vs. price graph. The regression line is not near the middle of the cone, so there is most likely an outlier throwing off the numbers. Looking at the scatter lot and histogram, I decided to remove the 3 outliers above 200000 from lot\_size. After that, I ran another correlation test and scatter plot.

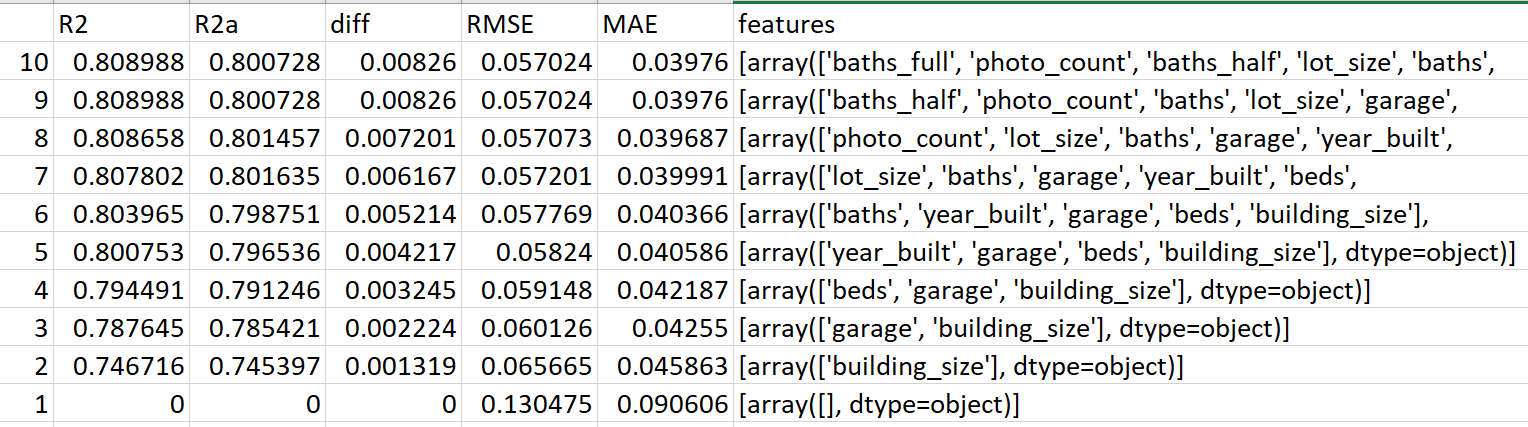
The results are much better. All variables are significant and have positive correlation to the price variable. It is now time for multivariate analysis.

**Multivariate Analysis**

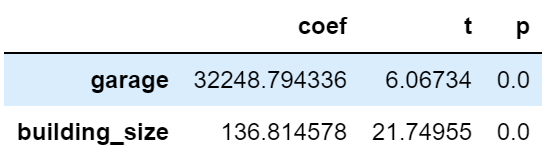
In order to normalize the weight of each variable on the effect on price, I transformed each variable into a range of 0 to 1. The minimum value in each variable range gets assigned to 0, and the max gets assigned to 1. This helps with variables like year, where the range is quite small given the actual numbers, which have only an arbitrary value. Next, I run a machine learning iteration, passing in results from an Ordinary Least Squares regression. This loop sorts by increasing t-test scores, and removes the weakest predictor variable, one at a time. The goal is to find the fewest variables that can predict the price accurately, by removing weak variables, without sacrificing precision too much.

The charts above show the t-value, p-value, R-squared, adjusted R-squared, the difference in the R-squared and the adjusted R-squared, Root Mean Square Error, and the Mean Absolute Error, for the predictions and the actual values of the home prices, using the available variables. It starts with all 9 predictors, then drops the weakest one at each loop. It then runs the correlation test again, and determines the next weakest variable to remove. The R-squared value does not decrease very much, until we get to only 1 variable! Here is the table that was exported with the iterations, showing which variables were used for each loop.



Here we can see that when using the last 2 variables (building\_size and garage) we can calculate the price without much difference than if we used all 9 variables. Next we get our coefficients for our regression line. Using MLR, we get these coefficients.



So, if we have a 2-car garage on a building with 1600 square feet, the estimated price would be 2\*32249 + 1600\*137 = $283698. That value is actually very close to my latest appraisal on my current home, which is a 3 bedroom, 2 bathroom, 2-car garage, 1600 square foot house.

**Conclusion**

I have tested some of the entries in the data, and every one has been within 5% of the predicted value. That is pretty good, but I would like to see if there is more data out there for each of the homes. Many more factors can contribute to the home price, including location, school zones, and even cell phone coverage. I would like to explore this further in another project. But, for this data, I believe we have a satisfactory solution.

**References**

Raghavan, Shreyas. (2017, Jun 17). Create a Model to Predict House Prices Using Python. From <https://towardsdatascience.com/create-a-model-to-predict-house-prices-using-python-d34fe8fad88f>

**Appendix**

API link: <https://rapidapi.com/apidojo/api/realtor>

Data and Code: <https://github.com/nickmiller1023/Projects/tree/master/Home%20Prices>