26 September 2014

General Assembly Data Science
Data Exploration 01
Mark Holt

Question: What is the price of 1342 N Leavitt Street?
Answer: $834333
Confidence: 1 standard deviation $51636

Approach.
My approach to trying to estimate the likely price of the target house is essentially a "nearest neighbor" approach, but *without* using formal statistical methods. My hypothesis, therefore, is to try and establish a selection of features available from the dataset that correlate strongly with price, match those features to those of our target property, and estimate the price based on the prices of the matching houses. I have only used simple statistical measures, consisting of the mean, standard deviation and correlation.

Data Cleaning.
Prior to any analysis I cleaned the dataset. I restricted my analysis to Single-Family-Homes (SFH). With a subset of the data representing SFHs I then removed all records that did not contain a full feature set. In short I removed records that had missing values.

This left 4412 complete SFH records with a complete set of features.

I normalized each of the features with a simple linear transformation. I converted all features to have zero mean unit variance. This allowed different variables with different scales to be compared. It allowed for integer scales (number of baths for example) to be converted to a continuous scale and it equally weights the features, i.e. assumes initially all features are equally important. Features should not be given importance merely because of the scale on which they are measured.

Feature Extraction and Dimensionality Reduction.
I performed a correlation of each feature with the stated house price, which is shown in Fig. 1.

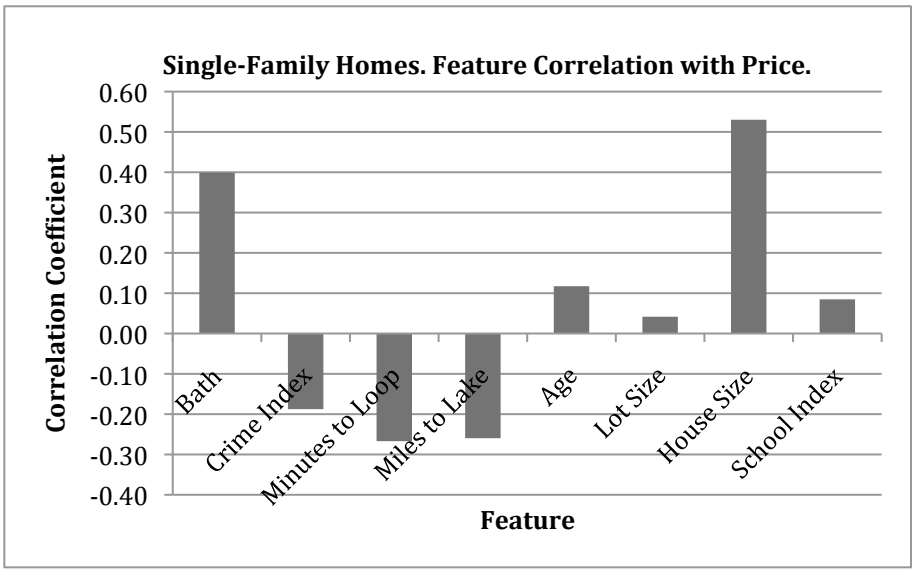

Figure 1. Data for all Single-Family Homes from the Chicago House Prices Database - only records with complete data used.

For the population of SFHs in Chicago the number of baths, and the size of the house appear to be the main features that correlate with price.

Domain knowledge, derived from purchasing houses over my lifetime, tells me that local neighborhood geography is also going to play a significant part in determining the house price.

Looking specifically at the neighborhood of Wicker Park the dataset yielded 22 complete records for SFHs. Two of these records I discarded on the basis of being outliers (records 21 and 22).

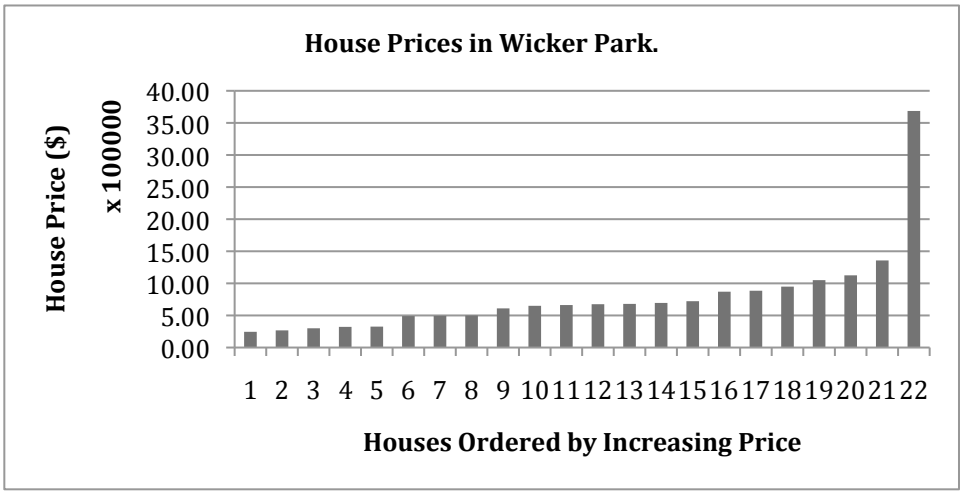

Figure 2. House Prices for Single-Family Homes in Wicker Park, Chicago.

The records shown in Fig. 2. Illustrate the house prices for the 22 records in order of ascending house price. The last record is clearly going to have undue influence on measures such as mean and standard deviation (and indeed that was the case).

Fig. 3. Below shows the correlation coefficients for each feature with house price for the Wicker Park subset. It is interesting to compare these results with that for the Chicago area as a whole. Number of baths and house size still remain positively correlated with price, but in addition house age has emerged as being negatively correlated with price (meaning more recent builds associate with a higher price and older builds with a lower price). Perhaps Wicker Park is an area with many older houses and as such when purchasing a house in this area it is something that buyers must pay attention to.

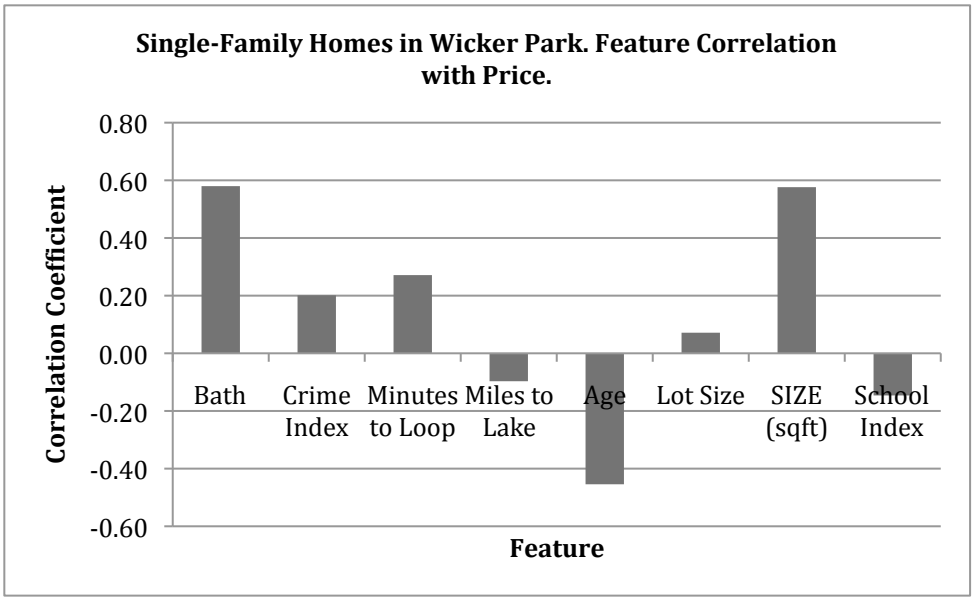

Figure 3. Data for all Single-Family Homes for the Wicker Park Neighborhood.

Using the information above I chose to keep only 3 of the features plus geographical location to attempt to predict the price of the target house. This constitutes a simple means of reducing the dimensionality of the input data. If I were constructing a formal statistical model for the data it would be unreasonable to use 8 dimensional data (with respect to the problem often referred to as the "curse of dimensionality") in any case. A more formal statistical tool to both rank and reduce dimensions is that of Principal Components Analysis (PCA). Also note that I am using a simple correlation of a single feature with the house price. There may be many more complicated correlations going on between the features that appear to have influence on price. These are unaccounted for.

This pruning of the feature set leaves 4 characteristics that appear to be important with respect to determining house price:

1. Number of baths,
2. Age of the house,
3. Size of the house,
4. Geographical location.

<u>The "Nearest Neighbors" Dataset.</u>
Having decided which features to utilize I now tried to find a close match to the target house. Given I had an interest in only 3 features I went back to the original database to ensure I had a complete set of data. I did not want to exclude houses that contained complete data for the 3 selected features. Previously I had excluded any homes with incomplete data for any of the original 8 features. I searched for other houses in Wicker Park that had 2 baths, were 116 years old, and had a house size of 3154 sq ft. Interestingly no other exemplars matching all of these criteria closely existed in the Wicker Park data!

The target house has a number of unusual features. Firstly, it seems impressive to have a house of over 3000 square feet and *only* have 2 baths, and secondly, it seems impressive to have a house size that exceeds the lot size. Within the dataset this house is relatively unusual. Given these observations the obvious question is how these peculiarities might affect price. Would potential buyers reward (or discount) a very old big house sitting on a relatively common lot size (3049 sq ft) with only 2 baths?

If this is the case then these peculiarities may well determine the price quite specifically (as opposed to other houses in the local neighborhood).

Further, if the target house is in fact an outlier, in terms of its characteristics, then it may be difficult to construct a model to accurately predict its price. Multidimensional logistic regression, and indeed Neural Networks (that offer non-linear mapping capability) should not be used to extrapolate. Ideally enough good data need exist to allow a smooth continuous mapping to be formed that will accurately interpolate an answer. This, of course, assumes that the selected features are, in fact, important determiners of house price.

I reexamined the main dataset to look at a different subset of data. Of all the SFHs in Chicago are there other instances of big, old homes with few baths that exceed their lot size? I added two derived features to the data that I called "Size/Bath" and "Size/Lot". "Size/Bath" = house size divided by number of baths, and "Size/Lot" = house size divided by lot size. The target house has a Size/Bath ratio of 1577 and a Size/Bath ratio of 1.03. For the entire collection of data for SFHs there were only 23 examples of houses with a larger Size/Bath ratio, and there were only 5 instances with Size/Lot ratios between 1.00 and 1.05! Intuition suggests that home buyers might offer less for a big home with too few baths, and given building code regulations nowadays perhaps having a house size bigger than the lot size is quite coveted.

From the dataset the following were the candidates to be considered "nearest neighbors" to the target house, based on the features discussed above. Note that none were in the Wicker Park neighborhood directly. As stated above no comparable houses (based on the features used) were found. The target house is listed as the fourth house in the table below.

Address	Neighborhood	House Type	Price	Bath	Age	Lot Size	House Size	Size/Bath	Size/Lot	Distance
1710 W Winnemac Ave, Chicago IL	Ravenswood	Single-Family Home	825000	2	114	3049.00	2943.00	1471.50	0.97	5.3
1330 W Wolfram St, Chicago IL	North Center	Single-Family Home	788000	2	126	3049.00	2950.00	1475.00	0.97	2.9
5338 N Paulina St, Chicago IL	Edgewater	Single-Family Home	890000	2	129	3049.00	3049.00	1524.50	1.00	5.7
1342 N Leavitt St, Chicago, IL	Wicker Park	Single-Family-Home		2	116	3049.00	3154.00	1577.00	1.03	

Table 1. Candidate "Nearest Neighbor" houses listed above the target house of 1342 N Leavitt.

Using Google Maps I measured the walking distance to 1342 Leavitt St from the houses listed above, and recorded that distance in the last column of the Table.

These 4 properties are well matched and take into account the unusual metrics found for the target property. Only 1 house, that in North Center, is somewhat close. I would have preferred to use just local data derived from Wicker Park, but overall I felt the unusual nature of this house was more likely to determine its price. Whilst the Edgewater property is farther away the demographics, particularly median income, were not too different from Wicker Park ($35766 median income for Edgewater, $38915 median income for Wicker Park). There was one additional comparable home that was in DePaul. The home in question was considerably more expensive and the median income in DePaul is listed as $96092. I discounted this house as being part of the nearest neighbor group. The fact that the properties listed above had prices relatively close to each other gave some confidence in the assertion that the peculiarities of the target house more accounted for its price than geographical closeness.

The Prediction.
To predict the price of the target house I chose to take the mean of the 3 nearest neighbors.

Predicted price is, therefore: $834333. The standard deviation is $51636.

Time permitting I would have liked to have tested this "informal" methodology on other randomly picked houses whose prices are known to get a better idea if the methodology it likely to produce any accuracy at all.