# “Bursting Bubbles”: Calculating Fair Housing Price

**Arjun Khandelwal, Patrick Insinger, Philip Sun, Uma Roy**

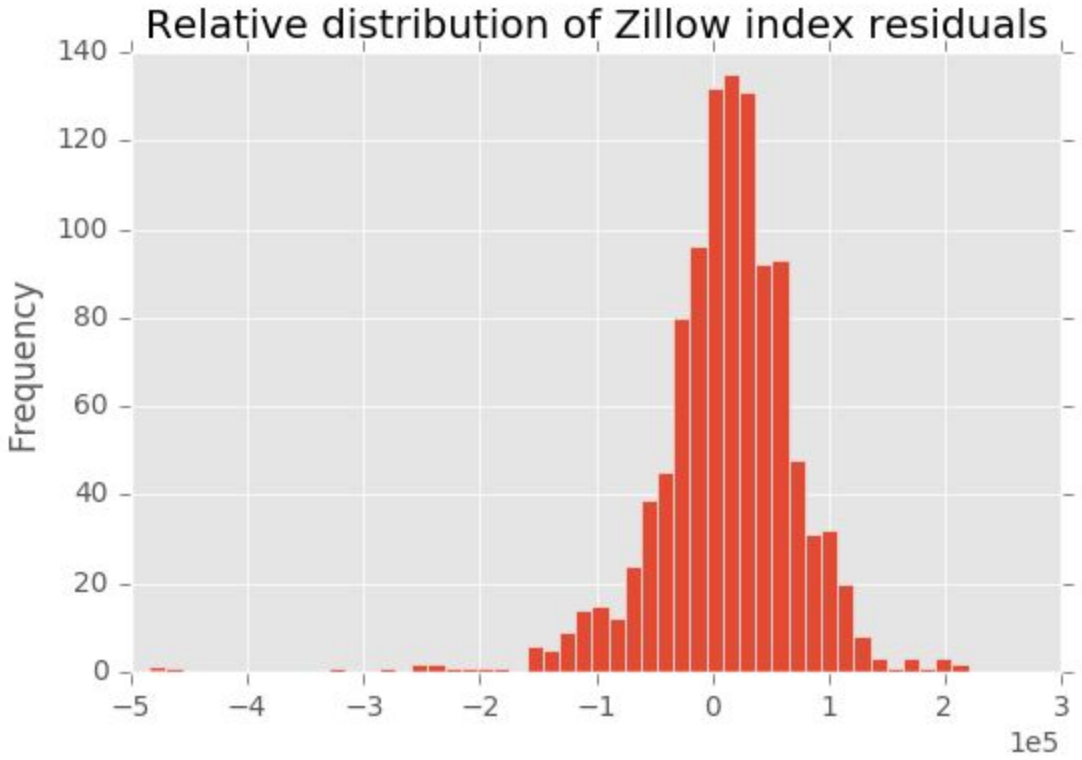
## Topic Question

The San Francisco housing market is notorious for being very expensive. Recently, a report by the UBS Group AG said that “San Francisco has the highest risk of any U.S. city of being in a bubble” (UBS group report). The UBS Housing Bubble Report said that the San Francisco housing market that was “overvalued” by analyzing several fundamental metrics and comparing them to historical levels. San Francisco’s zoning laws are very strict compared to other highly populated urban areas, restricting the heights of buildings, which limits the amount of housing that can be built in the city. As a group of college students who have been affected personally by this phenomenon, often struggling to find inexpensive housing for summer internships or jobs after college, in the ever-popular Bay Area, we wanted to investigate whether San Francisco housing prices were overpriced relative to the market at-large. With this impetus, we attempted to create a model which when given an area (provided there is sufficient data), to determine whether real estate market values are in line with the economic and demographic factors present in the area. We used the demographics, education, industries, jobs, and real estate datasets in our analysis, as well as a dataset that we found on the United States census website that provides the area of each city that we were looking at (found here). Also by analyzing the coefficients of the regression relating housing market prices to other factors, we gained fundamental insight into what drives housing prices up or down. In addition to examining San Francisco, we also examined several other major metropolitan cities, and were able to examine the top overpriced and underpriced housing areas in the country with our model.

## Non-technical Executive Summary

We were interested in predicting the value of Zillow Home Index for each city based on certain economic and demographic factors, including education level, job postings by sector, population demographics, and income demographics, with the goal of examining residuals to determine the presence of housing bubbles. Much of the data that we found was from the United States Census from 2011-2015, so we used the average Zillow Home Index from the time period as the variable that we were trying to predict based on the factors we had identified from that period. The final model that we arrived on was able to predict the Zillow Home Index fairly accurately. Based on this model, we examined our predicted value for the home index minus the Zillow Home Index to determine whether housing was overpriced or underpriced in the city. A negative residual value means that the housing is overpriced and a positive residual value means that the housing is underpriced.

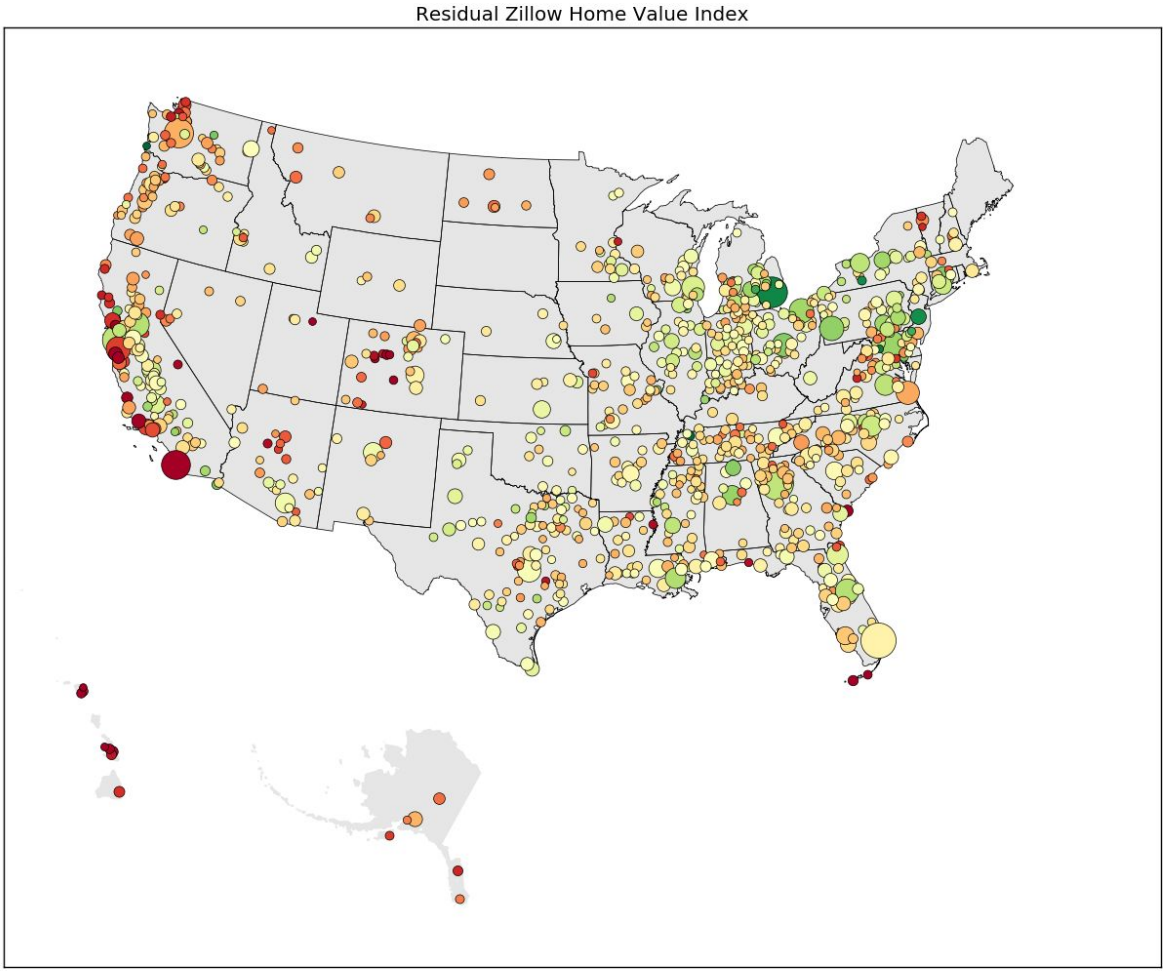
Below is a histogram of the residuals for each each of the cities, showing approximately how many cities fall into the corresponding overpriced/underpriced category.




We were particularly interested in San Francisco, which is reported to be overpriced by many different experts, who claim that San Francisco is currently in a housing bubble. We find that according to our model that San Francisco is the second most overpriced city that we examined (by $450,000). Similarly, we find Miami is overpriced by around $118,000 and New York is overpriced by around $90,000. We find that cities such as Seattle are underpriced by $23,000. We see that many of the cities that are currently being accused of being in housing bubbles, are, according to our model, overpriced, confirming our hypothesis.

We see that this model provides an interesting way of analyzing the economic and demographic trends that influence housing prices and provides a method of quantitatively determining housing prices in an area and comparing them to the current market. Our coefficient analysis (done in the technical section below) also provides intuitive explanations for the coefficients of our model, which also grounds our model in interpretable economic basis.

Finally, below is a map of the cities that we examined, plotting the residuals of our analysis (our model's predicted value minus the Zillow Index).




Circle Size: City population
Circle color: red for negative residual (overpriced)
green for positive residual (underpriced)

**Technical Executive Summary**

Our overall approach was a regression based analysis, where the features were cleaned and bucketed from the demographics, education, jobs and the United States city information data set, and the variable we were trying to predict was Zillow average home price for particular cities and states.

The first important part of our analysis was data cleaning and feature extraction. Although all of the data we were looking at was relatively clean, there were several features that we wanted to bucket and combine to get a better signal for our regression. We also note that the data for education and demographics were taken from the US Census from 2011-2015. We describe our feature extraction and selection process below for each dataset:

**Education:**

The granularity of the education dataset was very fine--for each city and state, there were the number of individuals who had completed various grades (from kindergarten to 12th grade) and the levels of college degrees that they had gotten. We bucketed levels of education into 5 categories based on highest level of education attained: no school, some schooling through high school but no diploma, high school diploma/GED, college diploma, advanced graduate degree (masters, PhD). There were many categories of what sector people studied--we bucketed these categories into 2 distinct areas: science/technology (which includes engineering, business, math, physical sciences, etc.) and humanities/liberal arts sectors (including visual performance, literature, education, etc.). The motivation behind bucketing was to extract the important features from the education level and education area, as to not overwhelm our regression with features.

**Demographics:**

The demographics dataset, similar to the education dataset, had very granular data that we had to bucket. We bucketed age categories per city by the following age brackets: 0-19, 20-34, 35-54, 55+. We also bucketed household income by the following income brackets: $0-$35,000, $35,000-$100,000, $100,000+. We also had a feature for median household income.

**Jobs:**

From the jobs dataset, we extracted the number of job postings by city for each sector from 2011 to 2015 (since the US census only has data between those years). For each city, we included the percentage of job postings in a given sector in the feature vector. We decided to normalize by the total number of job postings to make the features less correlated to population.

**US Census Data:**

From US Census Data, we extracted the area of the city and also the population of the city to get the population density of a city as a feature.

Our final feature vector for each city encompassed all of the categories discussed above, and the final feature count was 62. The number of cities we examined that had data for all the features that we were examining was 1090. Because of the high dimensionality of our feature space relative to the number of datapoints, we used linear regression with L1 regularization (more commonly known as LASSO) for feature extraction--since LASSO drives insignificant coefficients to 0, we only used the features with the non-zero coefficients as our final feature set. For running our LASSO we used SKLearn LassoCV model, which uses cross validation to automatically select the most suitable penalization parameter. To enable easy comparison of regression coefficients between features, we also applied an affine transformation to each feature to shift the empirical mean and variance to 0 and 1, respectively.

The features that had largest negative coefficients with LASSO were the following (listed from largest absolute value to smallest): Percentage of population age 5-19, graduated high school but no college degree, percentage of jobs (POJ) in banquet catering and events jobs, POJ in the legal sector, percentage of population 54+, POJ in nursing.

The features that the largest positive coefficients with LASSO were the following (listed from smallest coefficient to largest): median household income, POJ in hotel culinary and kitchen, POJ in internet, POJ in banking and financial services, POJ in therapy and rehab, POJ in salon/spa/fitness, percent of population with college degree, population density, POJ in concierge and guest service, percentage of population with income over 100,000.

The remaining features had 0 coefficients and were not used in our final model.

**Predicted Value:**

The predicted value was the Zillow home index value from the real estate data set averaged from 2011-2015. Although this is not ideal, and we noted a large disparity in housing prices from 2011 to 2015, since the census data was collected over 4 years and had no granularity about when the information was collected, using the Zillow average from the time period of data collection seemed most appropriate.

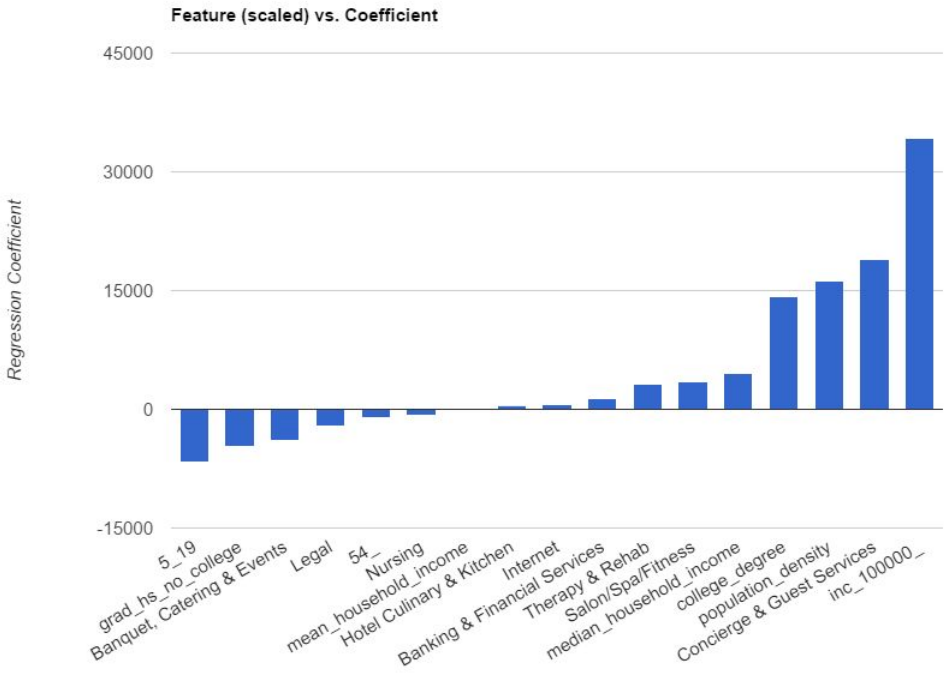
**Final model:**

With these list of final features, we tried various models to predict the Zillow home index value from the features. We tried various non-parametric models, include random forest regression and k-nearest neighbors regression, but we settled on using a simple linear regression model for the interpretability of the coefficients and negligible loss in $R^2$ compared to the more complex models. The out-of-sample $R^2$ of our final model was 0.663.

**Model Analysis:**

The feature selection that we got with LASSO was surprisingly intuitive and lined up with a lot of expected economic indicators for housing. In particular, the negative correlations between housing price and the percentage of population either in the 5-19 or 54+ age bracket are expected, because the young and the old population generally aren't in a position where they are generating income and spending it on housing. Similarly, an increased percentage of people that graduated high school with no college degree is linked to a lower demand for expensive housing.

For the features that were positively correlated with housing price, income and jobs in lucrative sectors (internet, banking and financial services) are sensible as predictors of high housing prices. Similarly, jobs in service subsectors that the wealthy would take advantage of (salon/spa/fitness, hotel/culinary/kitchen) also have an intuitive explanation for why house prices in an area are high. Income over 100,000 had the largest positive coefficient in our LASSO analysis, which has an obvious correlation with expensive housing.




Interestingly, all of the datasets we used proved informative. Without the jobs dataset, for example, the model R^2 dropped down to 0.414, showing the jobs dataset led to great incremental improvement of our R^2.

**Analysis of Major Cities and most overpriced/underpriced areas**

By analysis of the residuals (predicted value - Zillow value) of our regression, we can point to cities that are overpriced, underpriced, or at true market value. Let us look at an example:

For Miami, the Zillow home index average between 2011-2015 was ~$228,000. Our model predicts this value should be $110,000. Hence Miami is overpriced according to our model. For our model, here are top 12 overpriced cities (i.e. the cities with the most negative residuals):

| City | State | Zillow Index | Residual (Predicted - Zillow Index) |
|---|---|---|---|
| Vail | CO | 709,038.3 | -495,084 |
| San Francisco--Oakland | CA | 839,360 | -448,020 |
| Santa Barbara | CA | 835,548.3 | -324,181 |
| Edwards | CO | 701,401.7 | -317,497 |

| | | | |
|---|---|---|---|
| San Luis Obispo | CA | 540,161.7 | -268,691 |
| Princeville | HI | 538,510 | -258,823 |
| Kahului | HI | 414,103.3 | -244,258 |
| Santa Cruz | CA | 622,763.3 | -230,941 |
| Lanai City | HI | 346,555 | -218,799 |
| Half Moon Bay | CA | 776,468.3 | -210,531 |
| Lahaina | HI | 530,275 | -209,708 |
| San Diego | CA | 438,756.7 | -196,318 |

It is interesting to note that many properties in Hawaii show up as well as Vail, CO. According to our model, properties in this area are overpriced according to demographic and economic factors that we were provided with. However, tourism (for tropical vacations and skiing respectively for HI and CO) weren't provided as data into our analysis, which is probably why real estate in this area is seen as overpriced by our model, since tourism generally tends to drive property prices up. As predicted, San Francisco is second amongst the overpriced cities, as predicted by our initial intuition. Another article (here) hinted at San Francisco and Miami both being large housing bubbles, in addition to New York. We found that New York had a residual of -$87551, meaning that it is still overpriced, but not to the extent that Miami or San Francisco is.

Here are the top 10 underpriced cities (i.e. cities with the most positive residuals):

| City | State | Zillow Index | Residual (Predicted - Zillow Index) |
|---|---|---|---|
| Monett | MO | 91,093.33 | 166,944.8 |
| Concord | CA | 378,205 | 167,472.3 |
| Frederick | MD | 233,990 | 172,104.1 |
| Trenton | NJ | 83,740 | 176,687.6 |
| Hornell | NY | 52,461.67 | 180,589.9 |
| Woodstown | NJ | 174,501.7 | 187,180.6 |
| Ocean Park | WA | 149,981.7 | 205,031.8 |
| Goodrich | MI | 151,911.7 | 244,669.2 |

| | | | |
|---|---|---|---|
| Williamston | MI | 156,276.7 | 254,140.8 |
| Poolesville | MD | 383,751.7 | 257,096 |
| Mount Vernon | IN | 106,390 | 508,059.9 |

Most of these cities have small populations are not major metropolitan areas, so it is not surprising that they are underpriced because there is not much demand for housing in these areas.

**Conclusion:**

Overall, we see that our model lines up with our intuition that certain markets in major metropolitan areas such as San Francisco, Miami and New York are overpriced. With this model, we get easily interpretable coefficients that show the direct impact of demographic and economic factors on housing price. This model is able to quantitatively determine whether cities are underpriced or overpriced, and with this information, provides intuition and insight into whether certain areas are in a housing bubble or not. In particular, since our project started with the idea of examining whether housing prices in San Francisco were overpriced, it was a very interesting to see that our model pointed to San Francisco being the second most overpriced city in America. We think that our model could be used in other housing markets where we have similar data to predict home prices and determine the fair market value of houses for the future.