Housing in the San Francisco Bay Area: deal hunting using machine learning

Michael A. Boles
August 2019

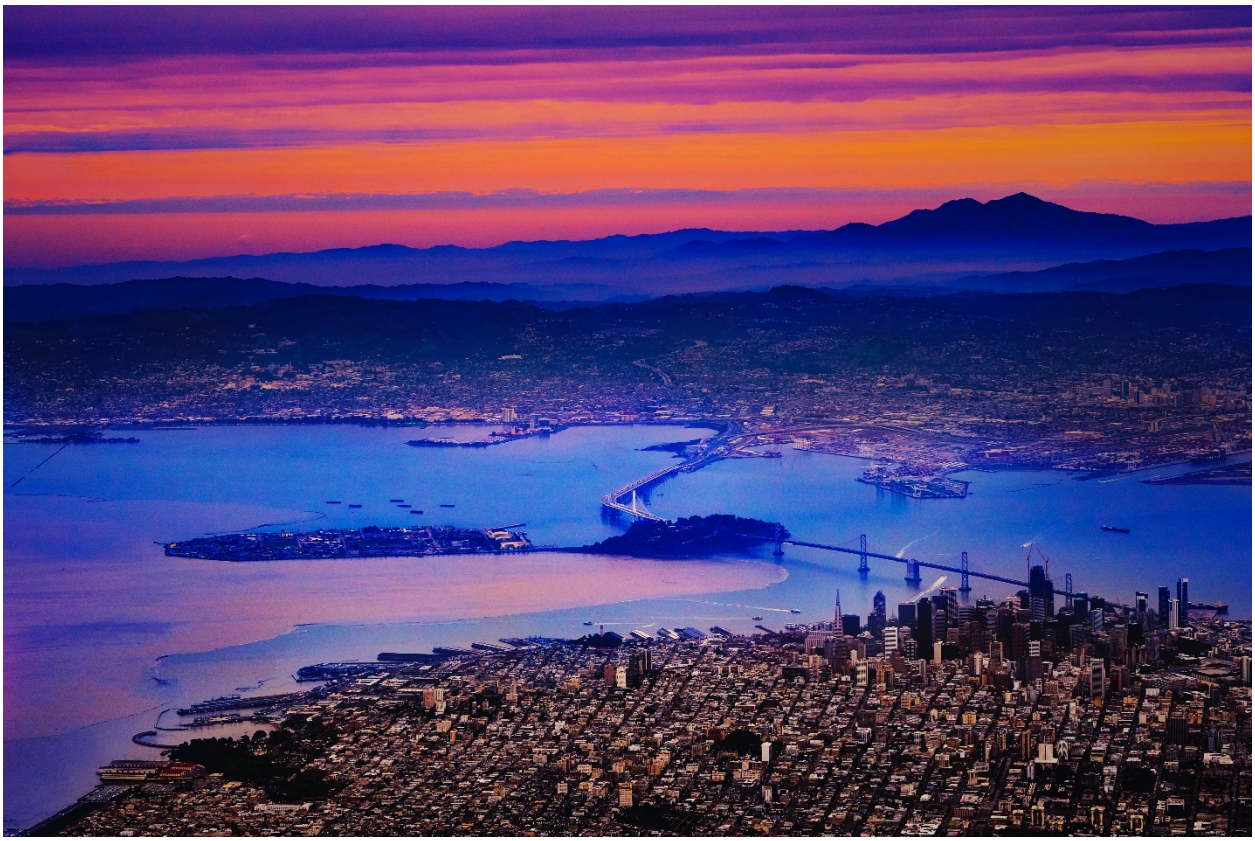


Photo credit: Unsplash (https://unsplash.com/photos/6b9rqGl_w1s)

Abstract
This article describes the collection, visualization, and modeling of single-family home prices listed across the San Francisco Bay Area in June 2019. Complementing listing information (number of bedrooms and bathrooms, home size, and lot size) with location data (school quality and commute times) was found to significantly improve explanatory power of the model, and enabled the identification of undervalued listings and neighborhoods.

Introduction

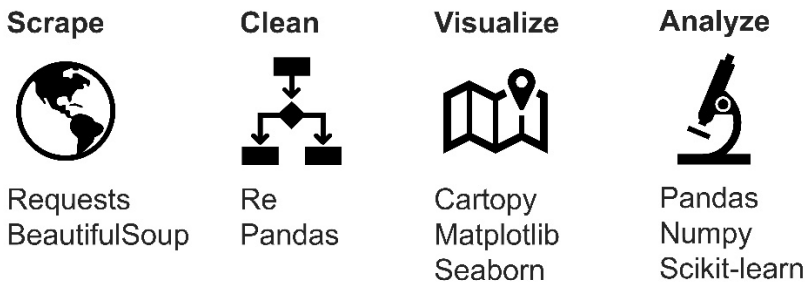
While the performance of the S&P 500 and the overall US housing market has been nearly identical since 2000 (both up ~100%), home price indices in the San Francisco Bay Area have risen by approximately 167% (St. Louis Fed[1]). As such, Bay Area homeowners have enjoyed an opportunity to build wealth through real estate in a way that is not necessarily accessible to most of the rest of the country.

For those already bought into the market, this near-tripling of real estate values since 2000 has undoubtedly been a good thing. However, for those newly relocated to the

region, saving towards a down payment and choosing where to buy can be a daunting task. Inspired by discussions I've had with friends and family, and basic concepts in investing (i.e., buy underpriced assets) I set out to gather as much information about current prices of single-family homes in the Bay Area, apply machine learning techniques to tease out factors driving home values, and identify areas that may be appealing for investment.

Methods

Data from single family home listings (address, beds, baths, home size, lot size, latitude/longitude coordinates, and price) across the Bay Area was scraped in June 2019 from a real estate webpage using the Requests and BeautifulSoup Python libraries, and cleaned and processed using Regex and Pandas. Complementing this listing data, commute times were obtained from Google Maps and school quality data pulled from the 2018 California Assessment of Student Performance and Progress (CAASPP). The resulting information was overlaid onto maps using Cartopy, Matplotlib, and shapefiles (town, zip code, and neighborhood borders) from Stanford Earthworks. Box/strip plots and pairwise relationships between variables were visualized using Seaborn. Ordinary least squares (OLS) regression analysis was applied to the data using Statsmodels and scikit-learn libraries. The full source code for this project is available on my GitHub page (github.com/mboles01).



Results

This study began with the collection of listing data for all single family homes on a popular real estate webpage (www.mlslistings.com). Looping over a list of Bay Area zip codes, location details (address, latitude and longitude coordinates), property characteristics (number of bedrooms, number of bathrooms, home size, and lot size), and list price were scraped for each listing using the Python Requests and BeautifulSoup libraries. Across 214 zip codes, location, property, and price information was collected for 7151 listings in June 2019. The data was cleaned using regular expressions to remove extra whitespace, unwanted characters, and entries with missing values.

I. Visualizing home and land prices

A. Geographic trends

With information for several thousand properties on the market across the Bay Area, the location of all listings was plotted on a terrain map and color-coded by price (Figure 1) using the Python Cartopy package together with city border information obtained from Stanford Earthworks. This map reveals clear geographic trends of the relative cost of homes across the region. For instance, San Francisco, Marin County, and the Peninsula typically contain the most expensive 20% of listings (dark blue data points), while those in Oakland, San Leandro, and Richmond are typically in the least expensive 20% (dark red data points). Similarly, in the South Bay, homes closer to the Santa Cruz Mountains are typically more expensive than those running alongside the Diablo Range.

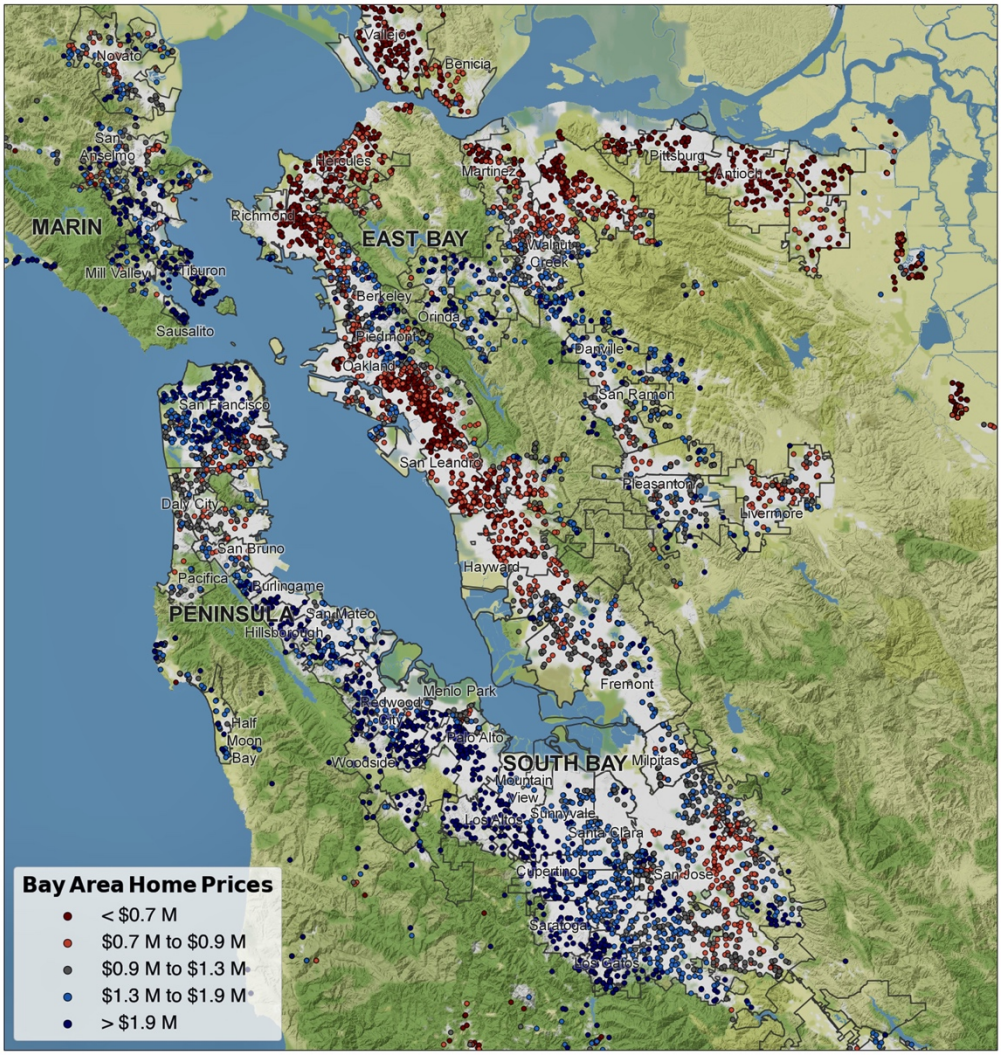


Figure 1. Overview of single-family homes listed for sale in the Bay Area in June 2019. The 7153 entries are split into quintiles by price, with list prices falling within the bottom and top 20% colored red and blue, respectively.

A natural extension of this analysis involves splitting the listings into subregions. Zooming in on San Francisco, the East Bay, the Peninsula, and the South Bay, geographic price trends again emerge (Figure 2). The most expensive single-family home listings in San Francisco fall between downtown and the Presidio, while the least expensive are found in the southern portion of the city, from the Sunset to Bayview districts. In the East Bay, homes on either side of the Oakland Hills (including Piedmont, Berkeley, and Orinda) are the most expensive, while Richmond, South Oakland, and San Leandro are the least expensive. On the Peninsula, homes in Palo Alto, Atherton, and Hillsborough are the most expensive, while San Mateo and East Palo Alto are the least expensive. In the South Bay, homes closest to the Santa Cruz Mountains (Los Altos, Saratoga, and Los Gatos) are the most expensive, while nearly everything east of that is less expensive. Relative cost across regions is also interesting to note: while the top price quintile in the East Bay starts at $1.5 M, that same price point falls at the bottom of the scale on the Peninsula.

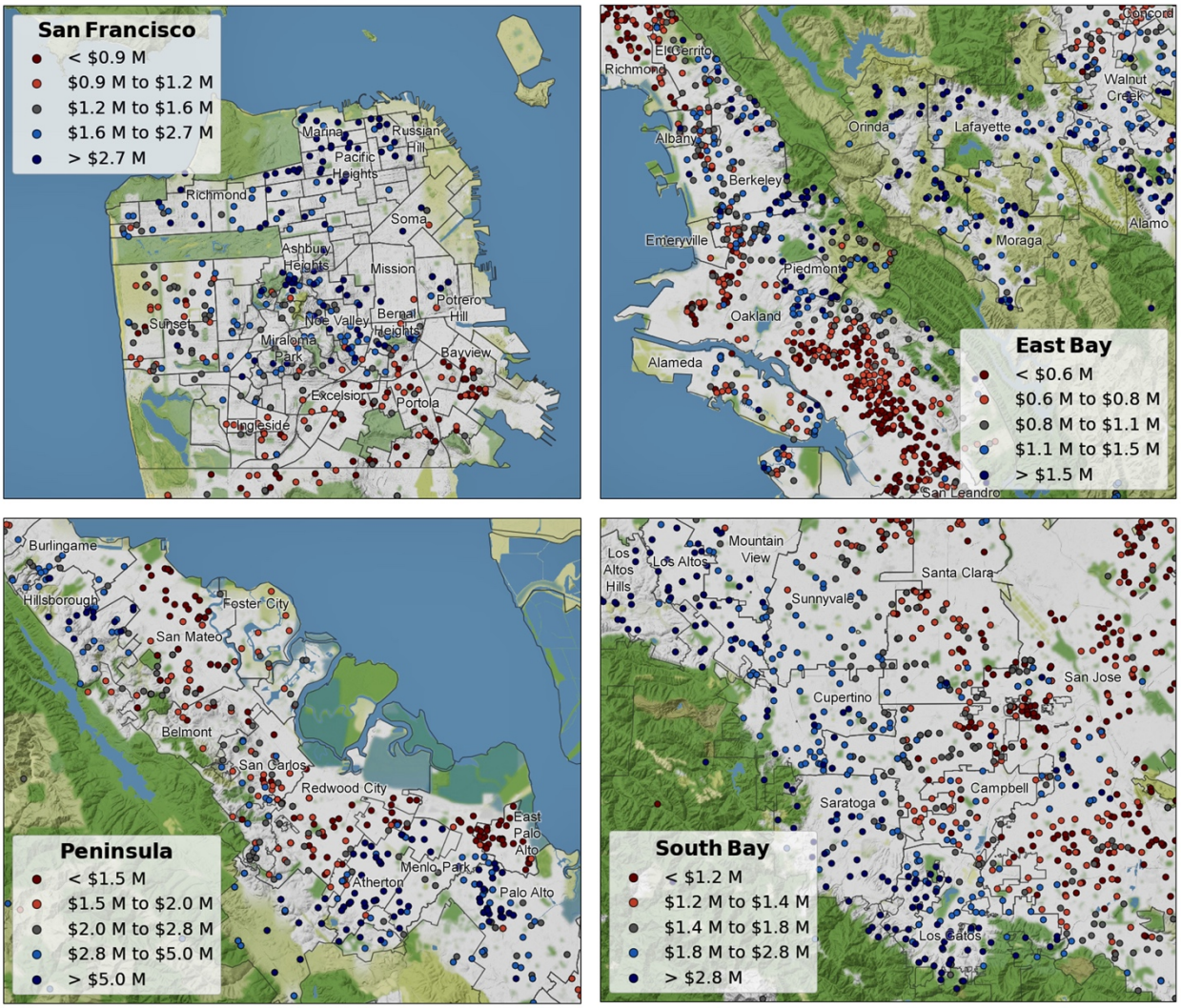


Figure 2. Zoom showing detail of single-family home list prices in the San Francisco, East Bay, Peninsula, and South Bay regions. In each case, price quintiles have been recalculated to reflect the distribution of prices within the highlighted region.

B. Comparing home and land prices across cities and towns

To better understand how prices vary across the Bay Area, box plots were constructed using the Python Seaborn library. This representation enables comparison of distributions across categorical variables (here: cities/towns), where the colored box denotes the bounds of the interquartile range (25th to 75th percentile), the line within it shows the median price, and the whiskers encompass the remainder of the distribution minus outliers.

From this perspective, sorting selected localities from least- to most expensive median home price offers some interesting insights (Figure 3). For example, the lowest median home prices are largely found in the northeast Bay (Vallejo, Antioch, and Richmond), while the highest are found on the Peninsula (Hillsborough, Palo Alto, and Woodside). In addition, the overlay of scatter points on top of the boxplots speaks to where most people live: Oakland, San Jose, and San Francisco each have several hundred current listings, while the more expensive dozen communities have just a handful of homes currently listed for sale.

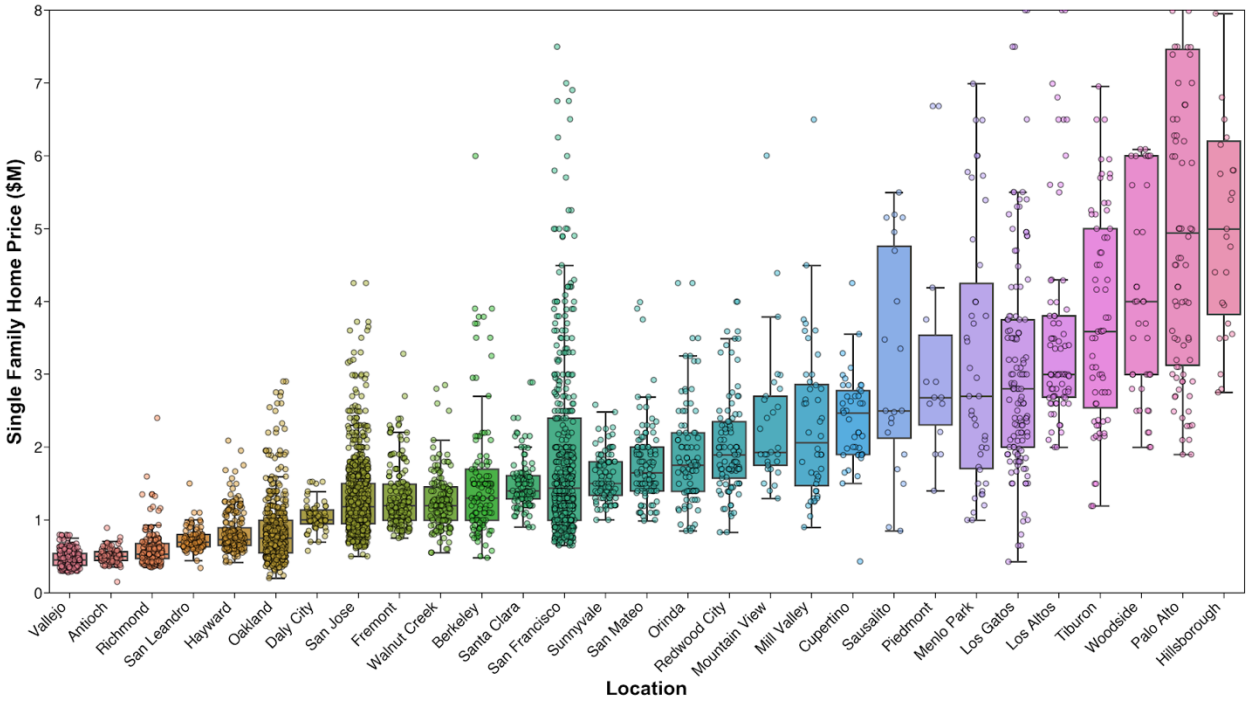


Figure 3. Box plot displaying home price and for selected Bay Area cities, with individual observations superimposed to reveal sample size and distribution.

Interestingly, application of the same analysis to the price of the land the house sits on (Figure 4) leads to somewhat different conclusions. For instance, while the median home price in Woodside ranks third highest of the 29 cities and towns depicted in Figure 3, the median cost of land per unit area in this community (located mostly west of I-280 and known for horseback riding) is in fact the lowest of those 29. Similarly, other towns with high median home prices, such as Orinda and Los Gatos, appear to be more affordable from a land cost perspective. At the other end of the spectrum, Palo Alto and San Francisco are by far the most expensive places to buy land, reflecting their status as major hubs of economic activity in the region.

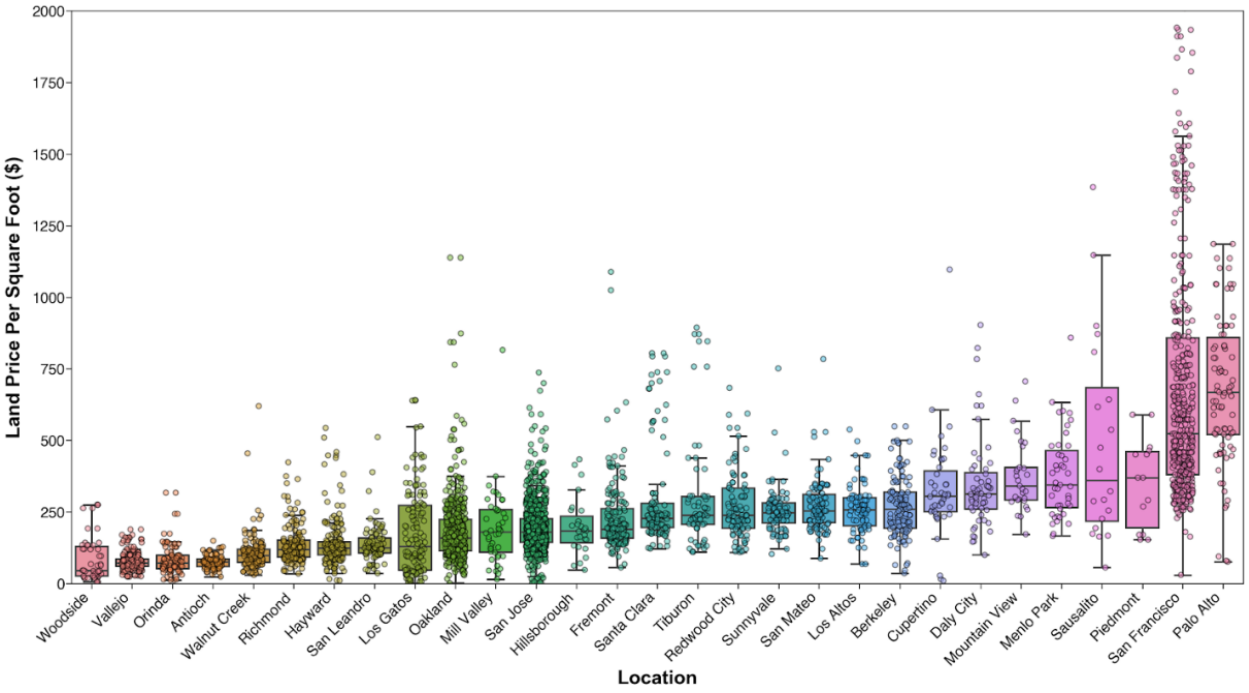


Figure 4. Box plot displaying land price for selected Bay Area cities.

II. Modeling home prices

A. Assessing listing data as predictor of home price

Beyond simply describing the geographic distribution of current Bay Area housing and land prices, the data set was used to model home prices using Python Statsmodels and Scikit-learn libraries.

To begin, prices of the 7151 homes in the data set were plotted individually against property data contained within the listing (number of bedrooms, number of bathrooms, home size, and lot size) and fitted using ordinary least squares (OLS) regression to assess pairwise correlations via Seaborn pairplots (Figure 5, top). Of these four features, the strongest individual predictor of list price is the home size (R^2 = 0.56), while the weakest correlation was found to be lot size (R^2 = 0.30).

On the other hand, when the full data set is narrowed to a single zip code (I chose my own: 95126), the correlation between these individual features and the home price is much stronger (Figure 5, bottom). Home size is again the strongest individual predictor of price, and in fact explains 90% of the variation within a single zip code.

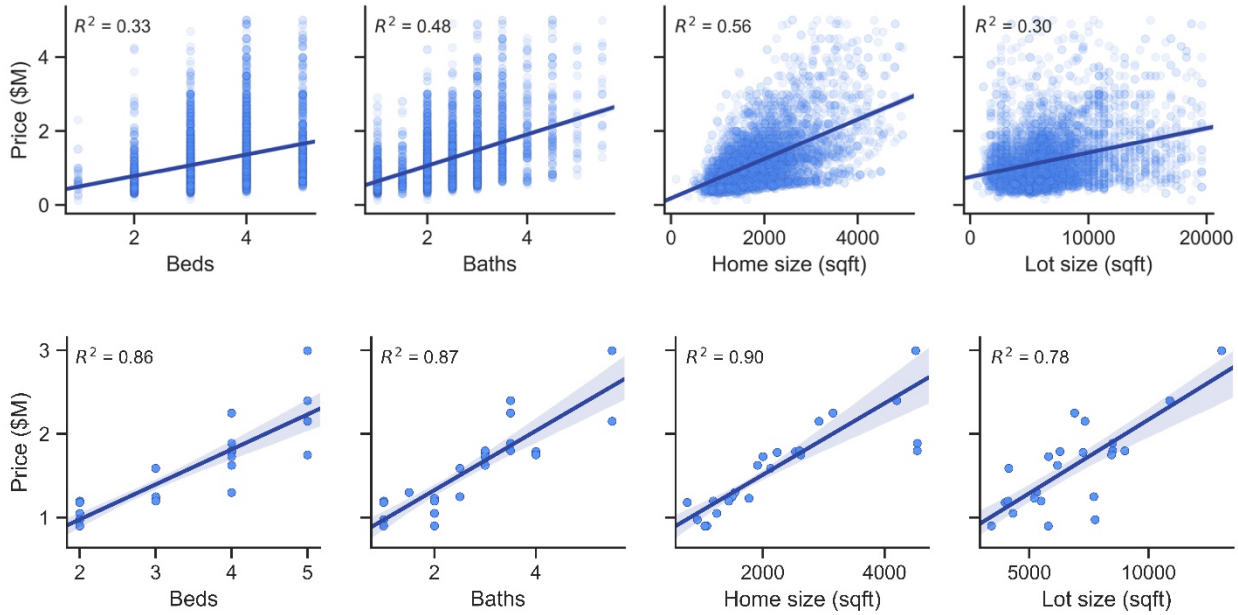


Figure 5. Correlation between price and listing factors is improved when full set of listings (top) is narrowed to a single zip code (bottom).

B. Adding new features: commute time and school quality

The modest correlation between home size and price over the entire region, and strong correlation within a single zip code, raises the very intuitive possibility that location-specific factors also contribute to price. To bring “place” into the equation, additional

data reflecting convenience and privilege was introduced: commute time and public school quality.

Recognizing San Francisco and Palo Alto as the two primary economic centers of the region, the commute time was measured via Google Maps as travel time by car (at 8am on Wednesday) from each zip code to the closer of the two destinations (Figure 6, left). The result indicates that homes up and down the Peninsula, from San Francisco to San Jose, offer the possibility of a commute time shorter than one hour. Similarly, Marin and Oakland have convenient access to San Francisco, as does Fremont to Palo Alto. On the other hand, commute times to the nearer hub are often more than 1.5 hours from homes in the outer East Bay (Richmond, Antioch, San Ramon).

The quality of public schools is also likely to play a role in setting the price of single family homes in a particular area. Conveniently, the California Assessment of Student Performance and Progress (CAASPP) standardized testing authority has made available a large set of student proficiency data for every public school in California. Using the 2018 data set,[3] averaging proficiency measurements across grades for each school, and across schools for each zip code, school quality was quantified as the fraction of students deemed proficient in reading and math within a particular zip code (Figure 6, right). Even after such averaging, this map points to a nearly bimodal split in student outcomes across the Bay Area, with Marin County, the Peninsula, and the Tri-Valley region enjoying excellent schools ($\geq$ 75% of students proficient) while public school students in Antioch, Vallejo, Richmond, San Leandro, and San Jose struggle by comparison.

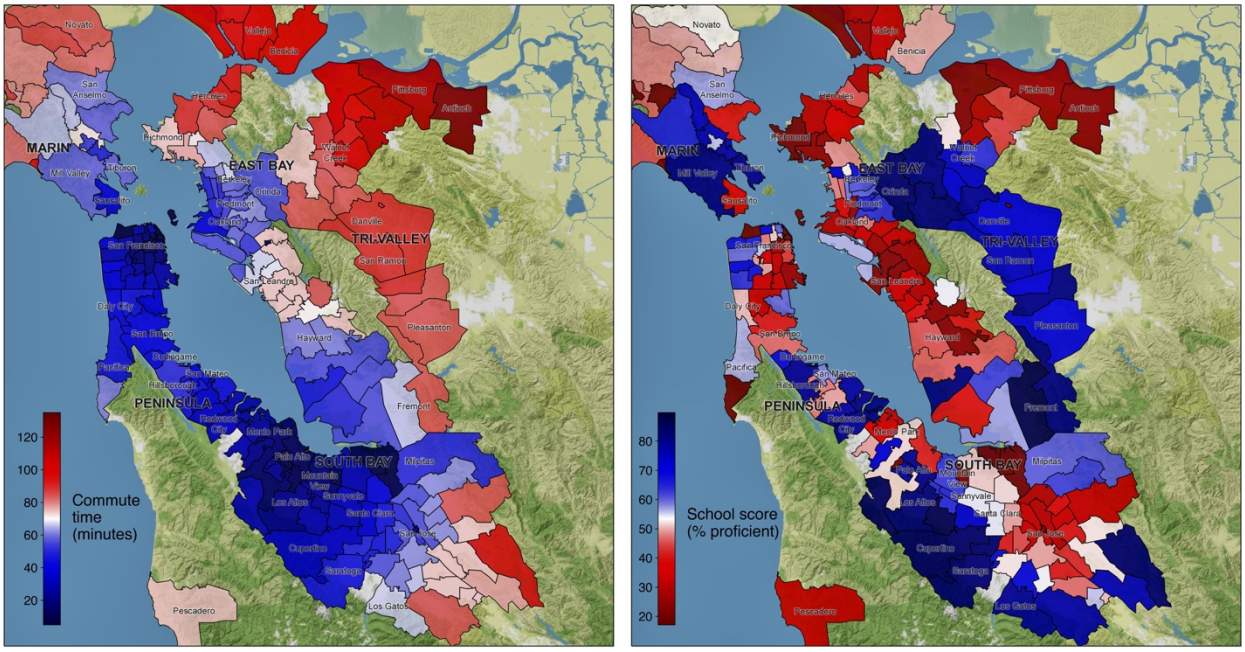


Figure 6. Commute times (left) and school quality (right) for zip codes across the Bay Area.

These new parameters were also evaluated as independent predictors of list price. Indeed, commute time and school score show similar strength of correlation (R^2 ~ 0.55) as home size, the previous best predictor (Figure 7). Intuitively, the lines of best fit imply a positive relationship between price and school quality, and negative relationship between price and commute time.

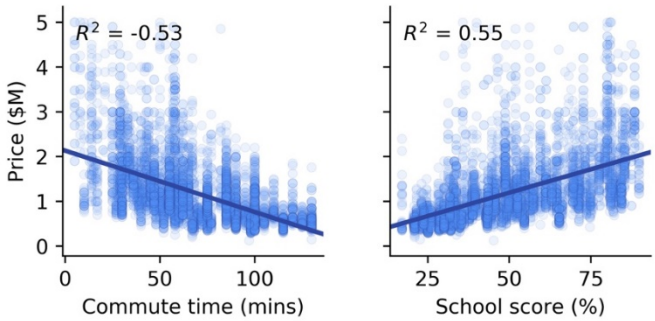


Figure 7. Correlation between price and commute times (left) or school quality (right) across the Bay Area.

C. Selecting features to include in the model

To avoid the pitfalls of multicollinearity in multiple regression fitting,[4] the correlation between all independent variables was assessed via Seaborn pairplots (Figure 8). In fact, three parameters from the listing data (number of bedrooms, bathrooms, and square feet) are all strongly correlated with one another (0.66 < R^2 < 0.79), while school quality and commute time show no significant correlation to each other or the listing data (0.07 < R^2 < 0.31).

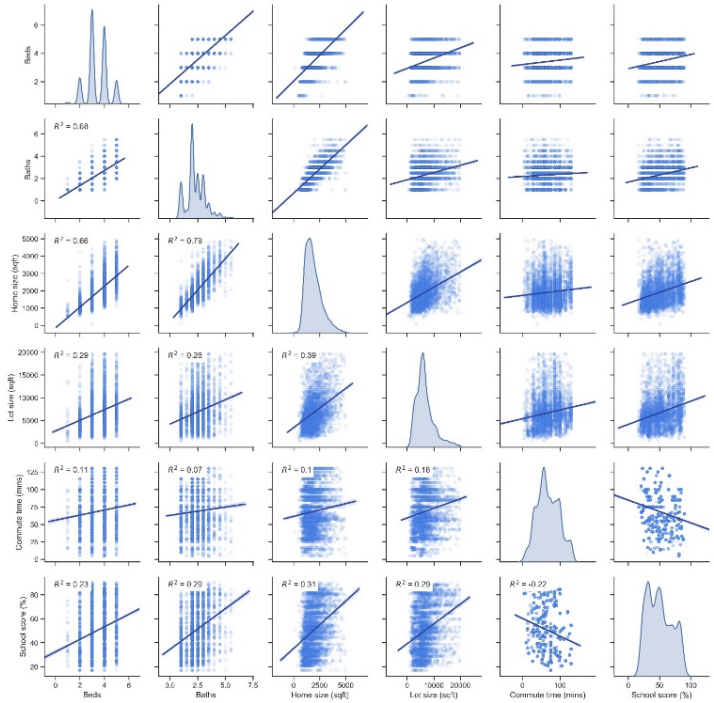


Figure 8. Pairplot assessing correlations among independent variables.

As the strongest independent predictor of price, home size (R^2 = 0.56, Figure 5, top) was kept in the model while number of bedrooms and bathrooms was discarded. Bringing in the commute time and school quality data, the linear regression model was constructed as:

Price ~ *Home size* + *Lot size* + *Commute time* + *School score* (**1**)

D. Fitting, interpreting, and assessing the model

Using these four independent variables, the data was fit to an ordinary least squares (OLS) multiple linear regression model with the Statsmodels package (Figure 9). The data was then filtered to remove outliers (homes of over 5000 sqft, 6 beds or baths, and $5M), bringing the sample size down to 5771.

Using only the four parameters offered in the original listing data (home and lot size, number of beds and baths) offered a model with relatively weak explanatory power (R^2 = 0.33, not pictured). In contrast, replacing the number of beds and baths with commute time and school score, two new parameters not significantly related to listing parameters, more than doubled the fit quality (R^2 = 0.73). It appears as though all four independent variables are important in the model: adjusted R^2 is identical to R^2, and *P*-values are vanishingly small for each.

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.734
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	3977.
Date:	Sun, 28 Jul 2019	Prob (F-statistic):	0.00
Time:	12:23:58	Log-Likelihood:	-82369.
No. Observations:	5771	AIC:	1.647e+05
Df Residuals:	5766	BIC:	1.648e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.344e+05	2.15e+04	29.503	0.000	5.92e+05	6.77e+05
Home_size	462.6851	7.297	63.405	0.000	448.380	476.990
Lot_size	29.6605	1.666	17.799	0.000	26.394	32.927
Commute_time	-1.418e+04	185.436	-76.491	0.000	-1.45e+04	-1.38e+04
School_score	8958.8506	291.200	30.765	0.000	8387.989	9529.712

Omnibus:	2204.409	Durbin-Watson:	0.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15452.518
Skew:	1.658	Prob(JB):	0.00
Kurtosis:	10.298	Cond. No.	3.31e+04

Figure 9. Regression summary for the fit based on Equation 1.

The coefficients for these four features are also interesting to examine: this model suggests that Bay Area homebuyers should, all else being equal, be prepared to pay $460 for each additional square foot of interior space, $30 per additional square foot for exterior space, $14k for every minute of commute time saved, and $9k for each percentage point gain in fraction of students proficient within their local public schools.

Plugging in the coefficients obtained from the multiple linear regression fit, the list price of a home can be predicted using the four inputs of home size, lot size, commute time, and school score:

$$\begin{aligned} \textit{Predicted list price (\$)} \;\approx\; & \textit{\$63,440} \\ & + \textit{(\$463 / sqft)} \cdot \textit{Home size (sqft)} \\ & + \textit{(\$30 / sqft)} \cdot \textit{Lot size (sqft)} \\ & + \textit{(-\$14,180 / min)} \cdot \textit{Commute time (min)} \\ & + \textit{(\$8,959 / \% proficient)} \cdot \textit{School score (\% proficient)} \end{aligned} \qquad (\mathbf{2})$$

For each listing, difference between the actual and predicted list price was calculated. This collection of price differences follows a normal distribution, which tightens up around zero after incorporation of location-specific (commute, schools) data (Figure 10).

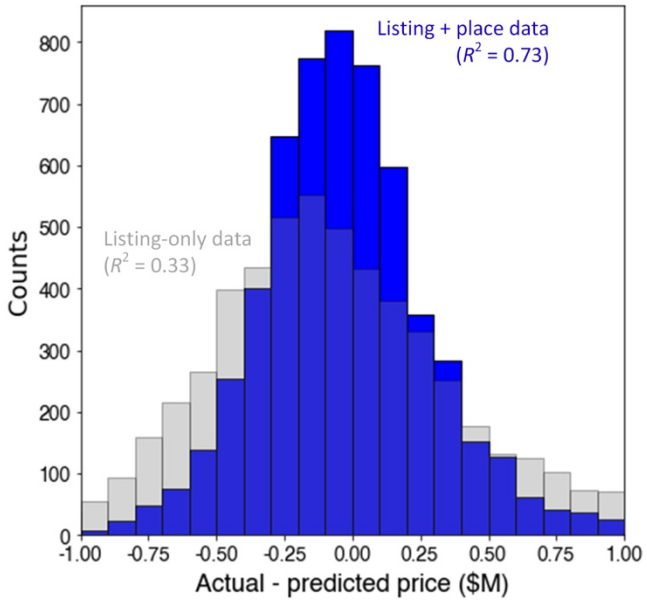


Figure 10. Difference between list price and predicted price before (grey) and after (blue) inclusion of commute time and school quality.

In the limit of a perfect model (R^2 = 1), this spread would collapse into a delta function at zero. In practice, the distribution width is determined by two factors: (1) aspects of a home and its surroundings that may affect price but are not included in this model, and (2) mismatch between intrinsic home value and list price set by agent or homeowner.

To address the first, one might add additional features (e.g., local crime rate, aesthetics of neighborhood, state of (dis)repair of the property) to the data set and re-run the regression. Trulia has made available high-resolution maps of crime risk,[5] but unfortunately, to my knowledge, the underlying data has not been released. Neighborhood aesthetics might be approximated by quantifying desirable attributes such as tree coverage or elevation.

E. Identifying undervalued listings

Buying underpriced assets is the core tenet of value investing.[6] From the perspective of an institutional investor or savvy homebuyer, a home value model offers the attractive prospect of identifying areas (and ultimately, individual properties) that may be mispriced with respect to the rest of the market. To this end, the location of each listing was plotted on a terrain map with marker color now denoting the difference between actual and predicted price (Figure 11). Underpriced listings (shown in red) are those that offer a combination of the four features (home and lot size, commute time and school quality) that the hypersurface of best fit anticipates should be priced above the observed list price. According to the model, underpriced areas can be found in San Francisco (Sunset and Bayview districts), along the Peninsula (Daly City and East Palo Alto), in the East Bay (Alameda, Orinda, Hayward, and Fremont), and Marin (San Anselmo). Aside from a few listings in San Jose, the South Bay looks to be generally overpriced.

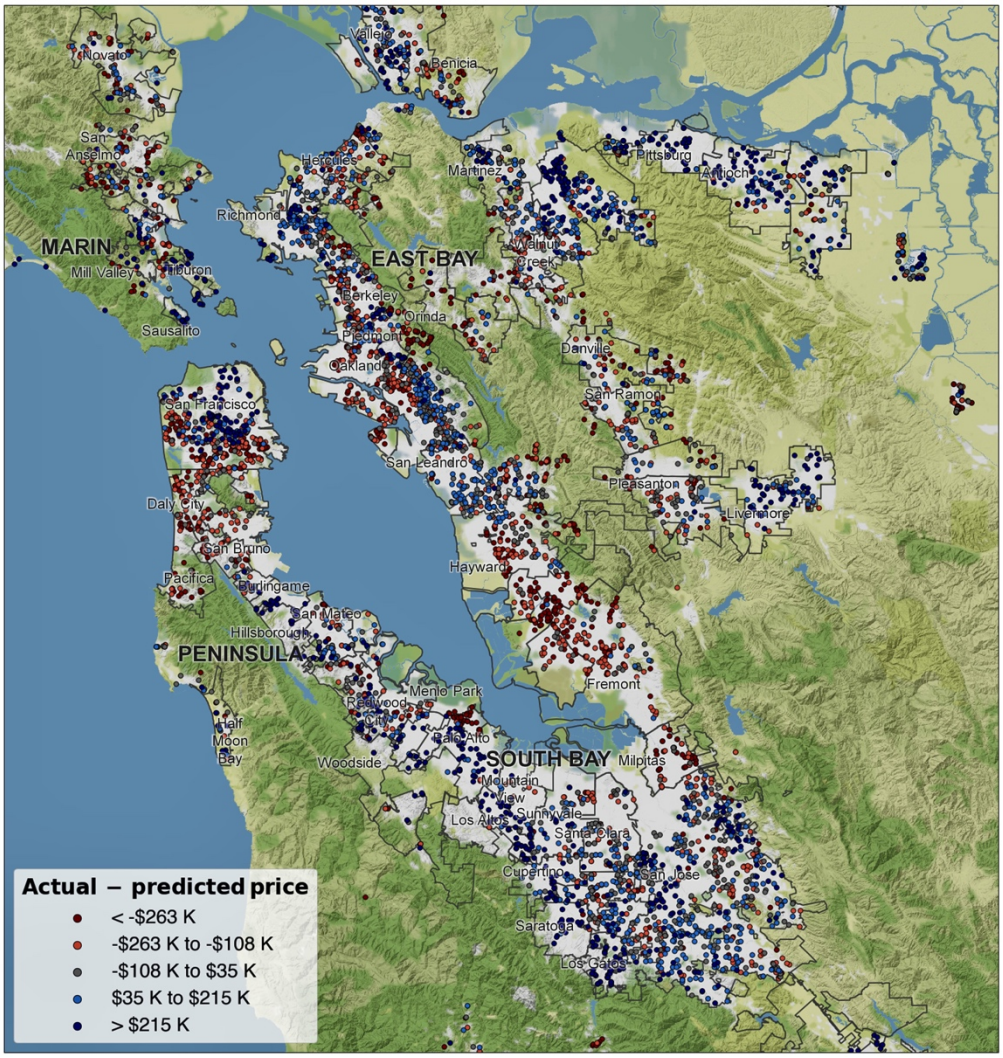


Figure 11. Difference between list price and price predicted by the model described in Equation 2. Listings deemed by the model to be undervalued and overvalued are shown in red and blue, respectively.

Conclusion

Scraping several thousand current real estate listings, the price of single-family housing across the San Francisco Bay Area was explored graphically and statistically. Bringing home and lot size, commute time, and school quality into a multiple linear regression fit enabled modeling of prices and identification of undervalued areas and listings that may be appealing for investment. The application of data science techniques to inform real estate transactions need not be pursued only by professional investors – free, open-source packages for use with Python empower the individual to gather large data sets, visualize key metrics, and apply machine learning to identify deals that may be overlooked by other market participants.

References

[1]Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/series/SFXRSA)
[2]Shapefiles: Stanford Earthworks (https://earthworks.stanford.edu/catalog/stanford-vj593xs7263)
[3]California Assessment of Student Performance and Progress (https://caaspp.cde.ca.gov/sb2018/ResearchFileList)
[4]Statistics How To (https://www.statisticshowto.datasciencecentral.com/multicollinearity/)
[5]Trulia crime data (https://www.trulia.com/real_estate/San_Francisco-California/crime/)
[6]Wikipedia (https://en.wikipedia.org/wiki/Value_investing)