**Abstract**

This article describes recent work mapping out single-family home prices across the San Francisco Bay Area, and adding commute time, school quality, and crime data to model for home prices across the region and ultimately identify undervalued listings.

**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). As such, Bay Area homeowners have enjoyed an opportunity to build wealth through real estate in a way that is not accessible to most of the rest of the country. Whether growth in SF-area real estate values will continue to outstrip other investment opportunities remains to be seen, of course, but the rising fortunes of large technology companies and unique, California-specific geographic constraints (water, mountains, desert) on the expansion of metro areas would appear to tilt the supply-demand balance toward the latter.

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, basic concepts in investing (i.e., buy undervalued assets), and a desire to hone my data science skillset, I set out to gather as much information about current prices of single-family homes in the Bay Area and apply machine learning techniques to tease out the most important factors driving home values.

**Methods**

Single-family home listings (address, beds, baths, home size, lot size, latitude/longitude, and price) across the Bay Area were scraped in June 2019 from a popular real estate webpage using the Requests and BeautifulSoup Python libraries, and cleaned and tabulated using Regex and Pandas. Commute times were obtained from Google Maps, school quality data pulled from the California Assessment of Student Performance and Progress (CAASPP), and crime data was retrieved from \_\_\_\_. The data was plotted on top of maps using Cartopy, Matplotlib, shapefiles from Stanford Earthworks. Box/strip plots, and pairwise relationships between variables, were visualized using Seaborn. Ordinary least squares regression analysis was applied to the data using Statsmodels.

**Results**

a. Maps describe home price trends across the region (Figures 1,2)

b. Box/strip plots enable ranking cities by cost of house, cost of land (Figure 3)

c. Plots of list price vs. listing data (beds, baths, home size, lot size) show weak positive correlations that narrow upon zooming into one zip code (Figure 4)

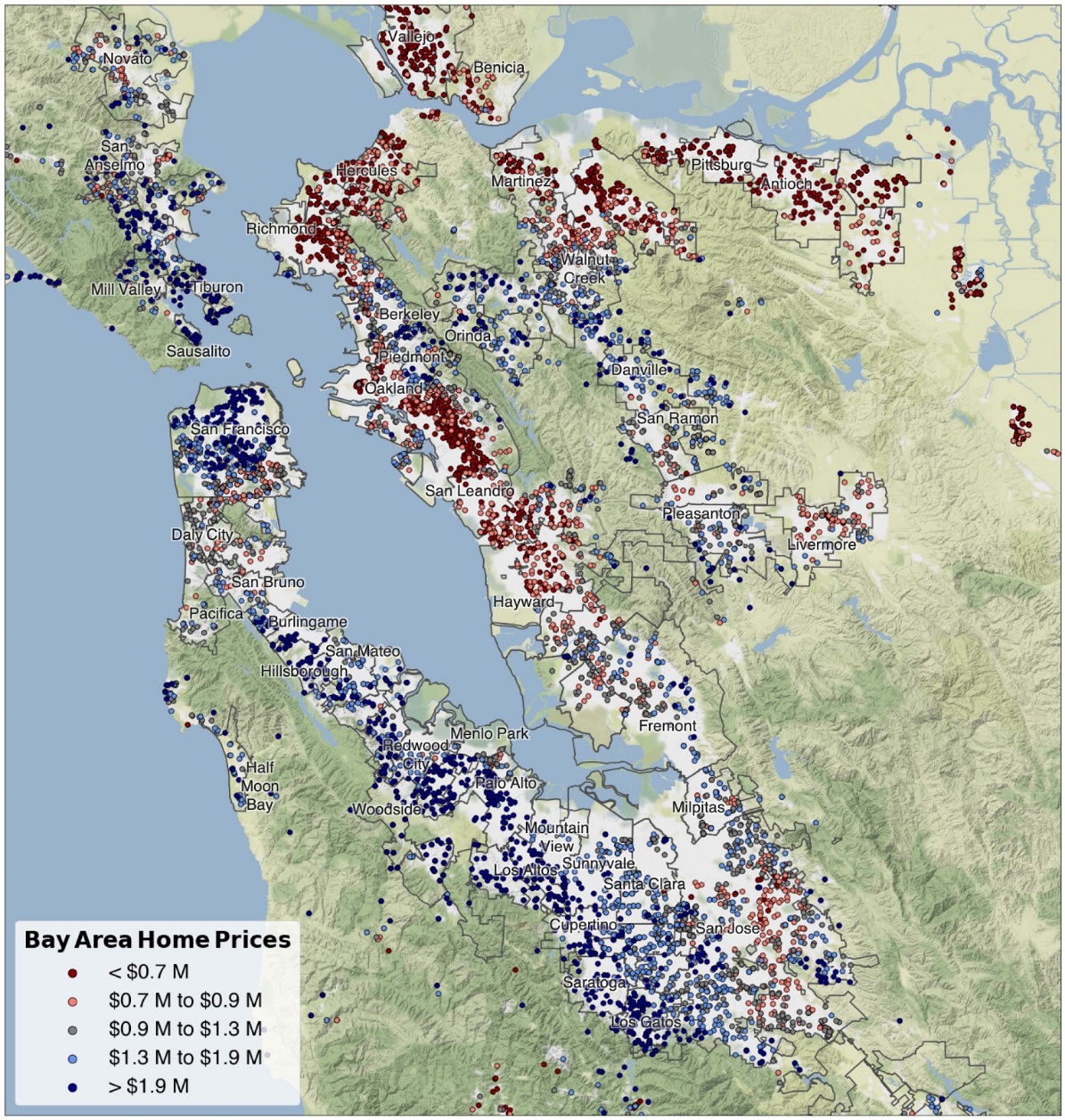
d. Aiming to capture other aspects of “place”, collected data on commute times, school quality, and crime rate (Figure 5?) – rationale for incorporating such metrics

e. Ran linear regression fit on the full data set, evaluating the effect (coefficients) and statistical significance (P-values) of the inputs to devise a model for home prices (Figure 6? Equation 1?)

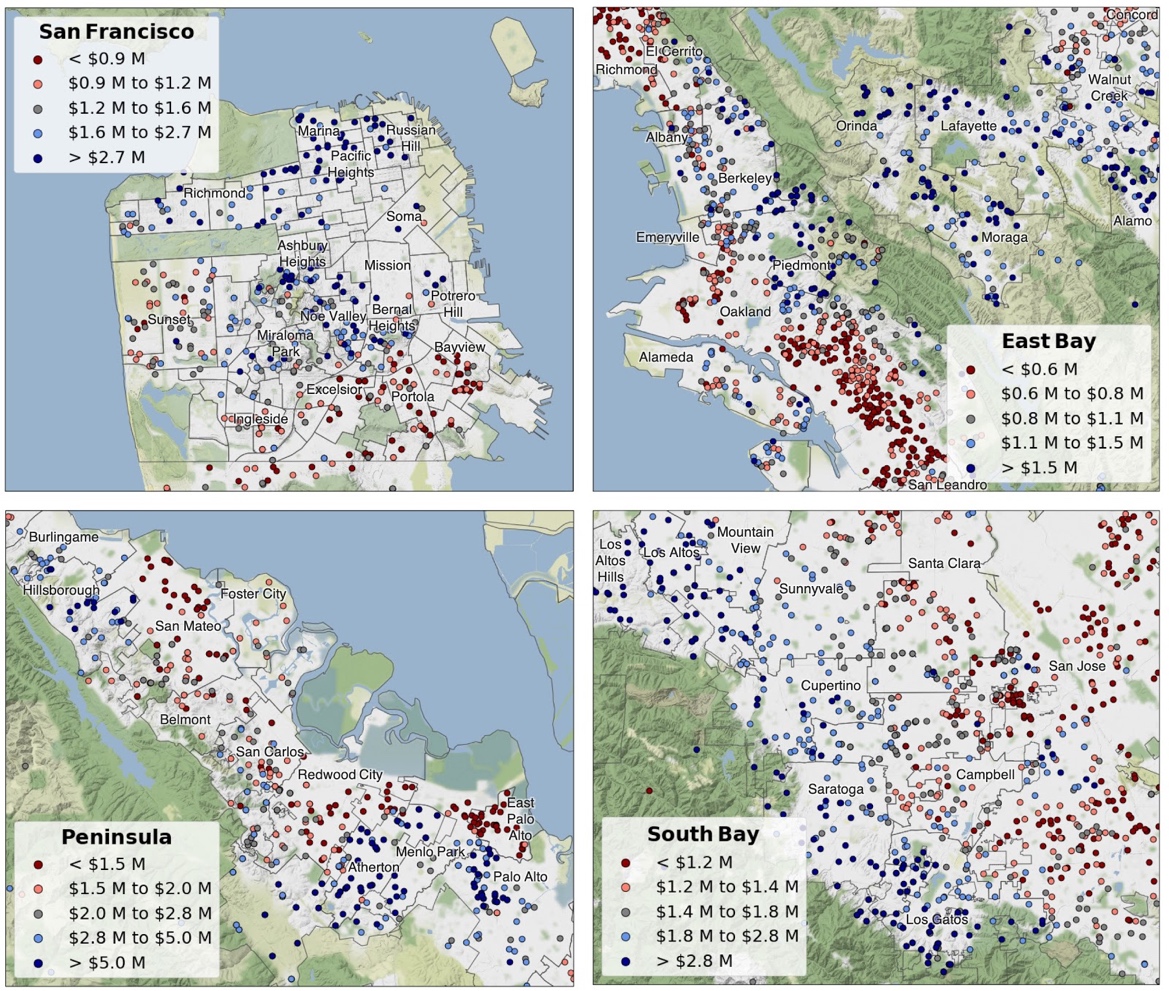
f. Calculated the difference between predicted and actual list prices, used this to identify potentially undervalued homes (Figure 7? – histogram of *P*pred – *P*list, Table 1 – 10 undervalued, 10 overvalued houses?)

**Conclusions**

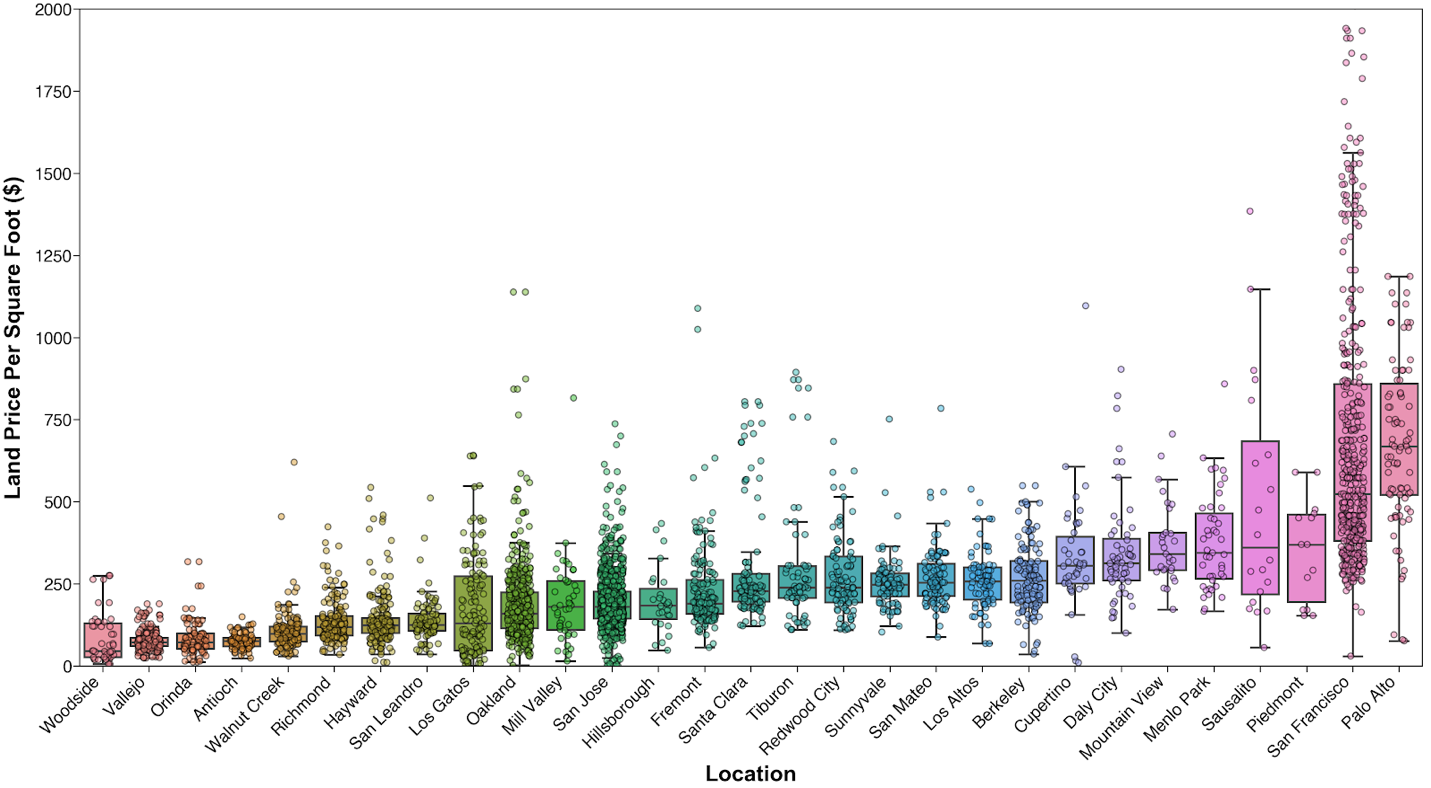
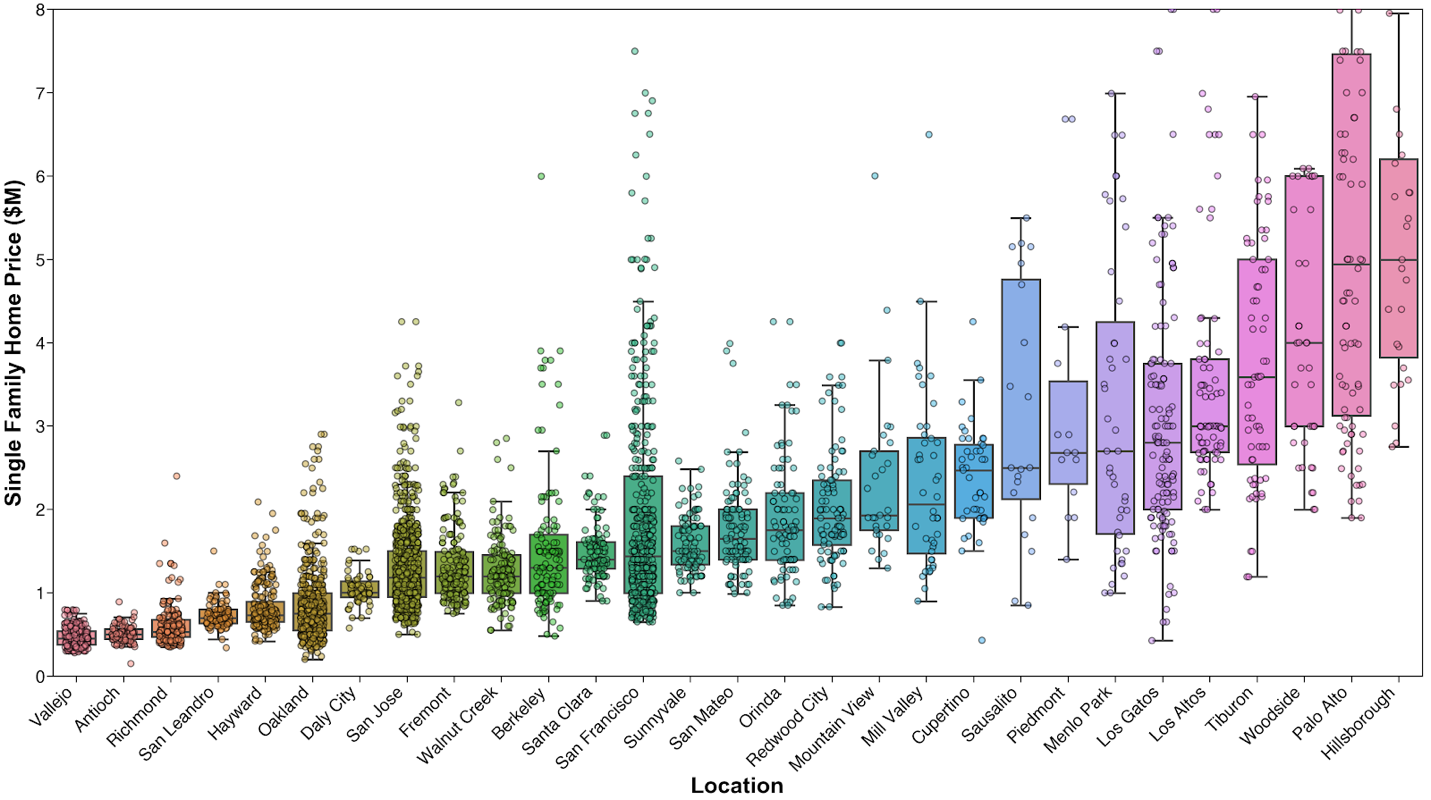
The application of data science techniques to inform real estate investment decisions need not be only pursued by corporate investors – free, open-source packages for use with Python empower the individual to scrape webpages, visualize data, apply machine learning to identify deals that may be overlooked by other market participants.



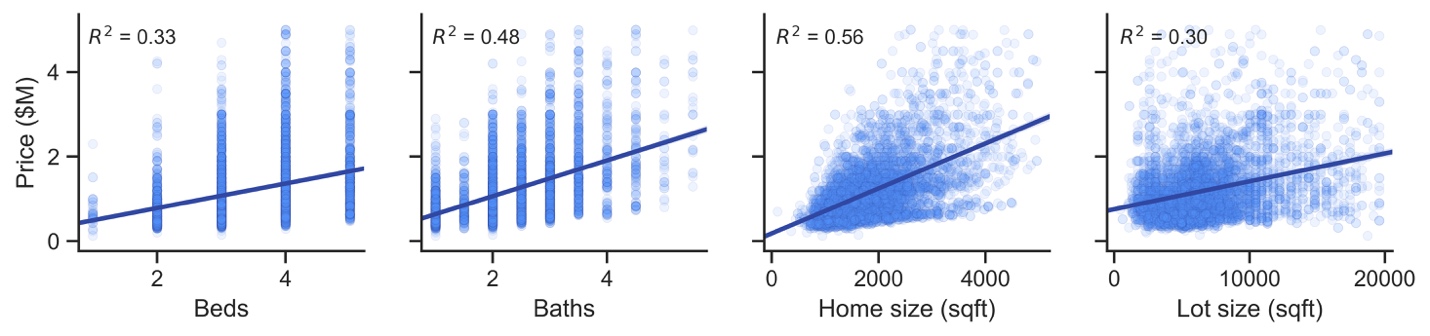
**Figure 1.** Overview of single-family homes listed for sale in the Bay Area in June 2019. The 7153 entries are split into quintiles by price, with list prices falling within the bottom and top 20% colored red and blue, respectively.

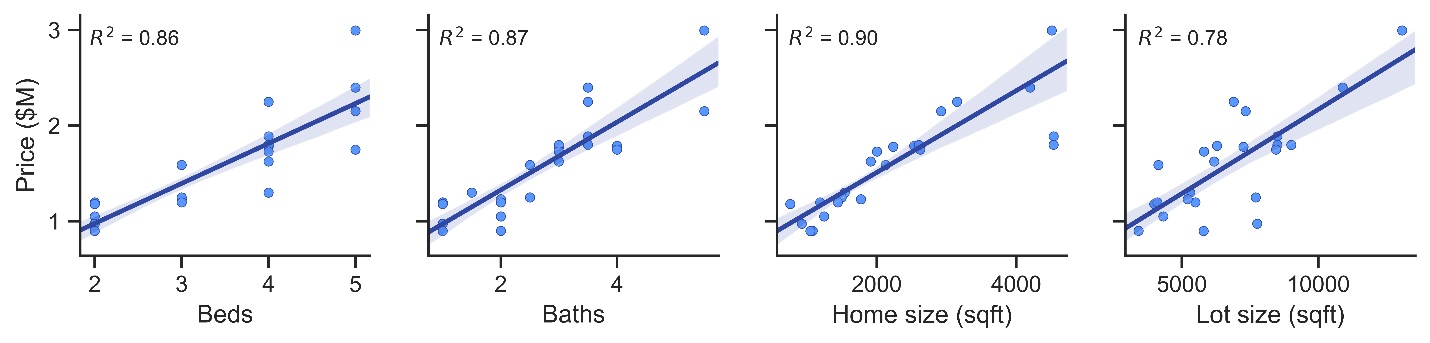


**Figure 2.** Zoom showing detail of single-family home list prices in the San Francisco, East Bay, Peninsula, and South Bay regions. In each case, price quintiles have been recalculated to reflect the distribution of prices within the highlighted region.



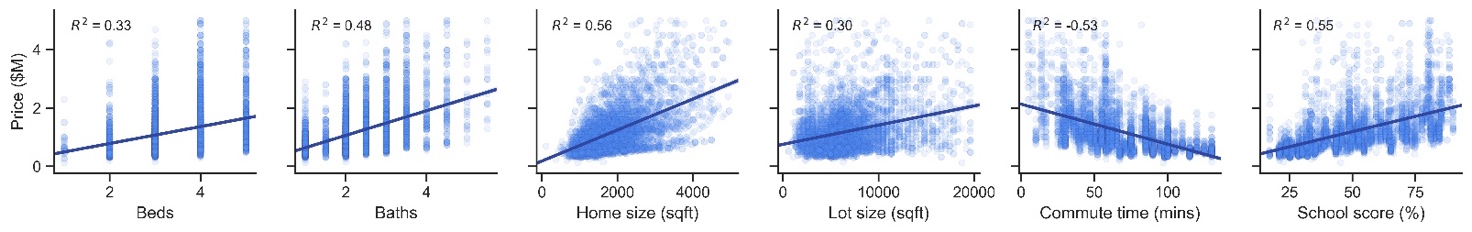
**Figure 3.** Box plots displaying home price (top) and land price (bottom) for selected Bay Area cities, with individual observations superimposed to reveal sample size and distribution.





**Figure 4**. Correlation between price and listing factors is improved when full set of listings (top) is narrowed to a single zip code (bottom).

New parameters commute time, school score are significantly related to home price



New parameters are not related to previous set – home, lot size. Three variables are however related – beds, baths, and home size (sqft). Because home size is most strongly correlated (R2 = 0.56) with price, beds and baths were discarded from the OLS model.

