Springboard Capstone2 Project

Buying Family Home in San Francisco Bay Area. Suggestions and Price Predictions for the First Time Buyers.

Agne Adukaite Aklifazla

Mentor: Ricardo D. Alanis-Tamez

INTRODUCTION

California has beed a Golden State, place of deams to many americans as well as foreigners. However, in the past decade Californians experience sressfull times in higher taxes, higher cost of living and steeping real estate prices. The year 2019-2022 facing global pandemic, loss of jobs, and inflation many people deicided to move out of state, out of city to rural area, or at least thinking to do so. Is it possible for a fimly to buy real estate in Bay Area? Is it worth it? While the question wether California is overrated or not lies most to each person subjective opinion, the question of possibility to buy real estate in Bay Area and the best places to invest can be answered with market analysis.

Data

In this document there will be presented real estate prices analysis using this data:

- Bay Area Real Estate supply from Zillow data. Condo, Townhouse and Single Family Properties.
- Unemployment rates.

- Crime rates.
- School ratings.
- Housing affordability index (HAI).
- Trends data to evaluate home appreciation.

About each data collection:

The goal of this project is to find best family house for the first time buyers. One of the major factors for the family house despite the good **location** is the **rating of the school districts**. Often, good school district is also a good indicator of safer, family orientated neighbourhood. Using [niche website] information and exploring the Bay Area school district maps, I have created a table with top school districts, rating A+, A or A-.

District	City or Town	County	Rating
Tamalpais	Larkspur	Marin	A+
San Francisco	San Francisco	San Francisco	A-
San Mateo	San Mateo	San Mateo	A+
Sequoia	Redwood	San Mateo	A
Palo Alto	Palo Alto	Santa Clara	A+
Mountain View	Mountain View	Santa Clara	A+
Fremont	Sunnyvale	Santa Clara	A+
Santa Clara	Santa Clara	Santa Clara	A
Los Gatos	Los Gatos	Santa Clara	A+
Campbell	San Jose	Santa Clara	A-
Milpitas	Milpitas	Santa Clara	A-
Fremont	Fremont	Alameda	A
Pleasanton	Pleasanton	Alameda	A+
Livermore	Livermore	Alameda	A-
Dublin	Dublin	Alameda	A+
Castro Valley	Castro Valley	Alameda	A
San Ramon	San Ramon	Contra Costa	A+
Alameda	Alameda	Alameda	A
Piedmont	Piedmont	Alameda	A+
Albany	Berkeley	Alameda	A+
Berkeley	Berkeley	Alameda	A+
Acalanes	Lafayette	Contra Costa	A+
Martinez	Martinez	Contra Costa	A-
Benicia	Benicia	Solano	A-
Brentwood	Brentwood	Contra Costa	A-

Table 1. Top School Dstricts in San Francisco Bay Area.

After finding out highest rated school districts in the Bay Area it was time to crack the actual real estate market in the region.

Zillow data was collected using Zillow API via Rapid API. One of the major constrains collecting this data, was the limit of 40 items per inquiry. I chose to collect data in segments by city, and real estate price. When filtering features of the properties using rapid api, I paid attention to:

- Location - city
- Type of property. Selected SINGLE_FAMILY (which included condos, townhouses, and single family houses).
- Minimum number of bedrooms = 2.
- Maximum living area sqf of 3000.

Data consisted 2430 rows and 13 columns before cleaning process.

Unemployment data was collected from U.S. BUREAU OF LABOR STATISTICS, and saved to .csv file for future analysis. Unemployment data was collected for California state overall, and each county.

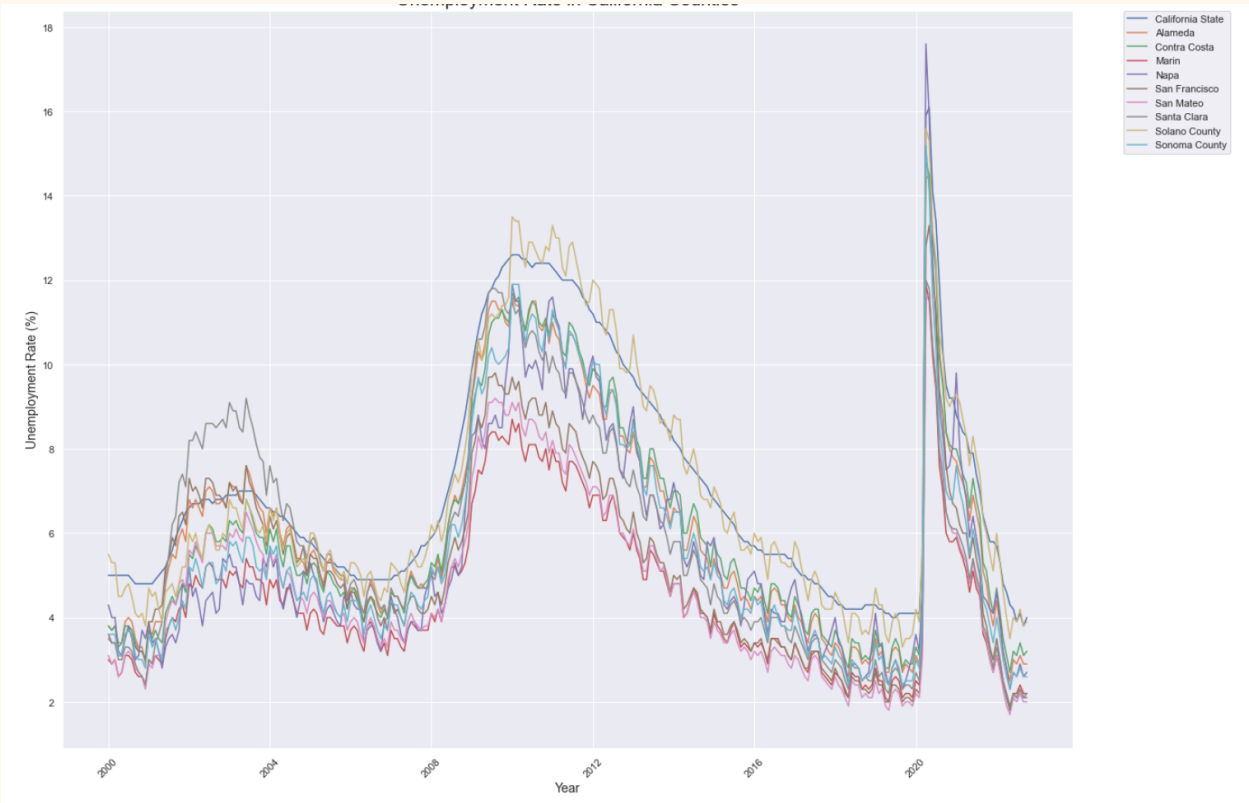


Chart 1. Unemployment Rate in Bay Area (year 2000 to 2022)

From the unemployment data we can see that Solano county is the one which had higher unemployment rate than California State. *Lowest unemployment rates are in Marin, San Mateo and San Francisco counties.*

Crime data was collected from Neighbourhoodscout for each Bay Area City, first to excel file, then converted to .csv for easier manipulation in Jupyter Notebook.

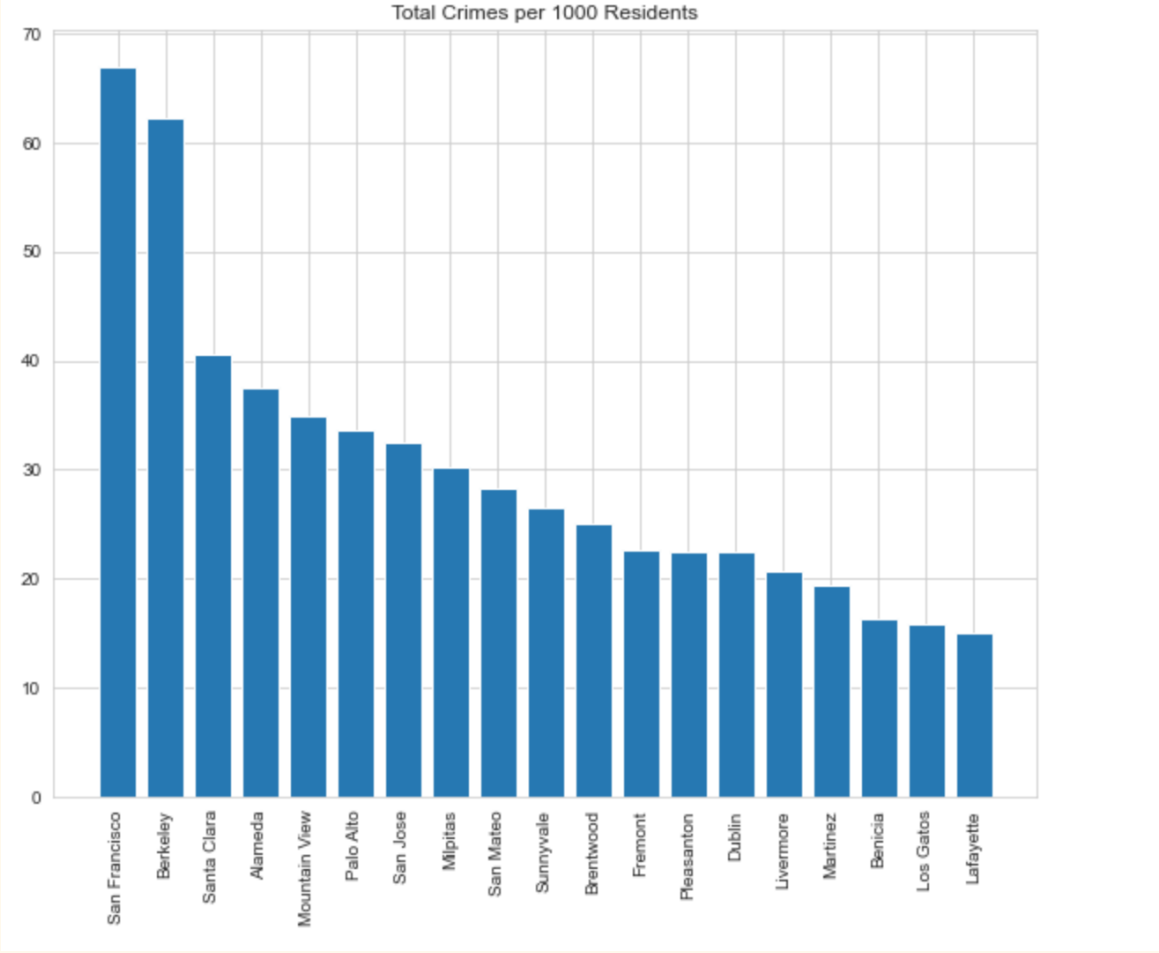


Chart 2. Crime Rates

As we can see from the chart above, most crimes per 1000 residents in cities considered in this analysis have San Francisco and Berkeley. *Three safest cities are Benicia, Los Gatos, Lafayette.*

Housing Affordability Index information was collected from California Association of Realtors Webpage. *C.A.R.'s Firts-time Buyer Housing Affordability Index (FTB_HAI) measures the percentage of households that can afford to purchase an entry-level home in California. C.A.R. also*

reports first time buyer indexes for regions and select counties within the state. The index is the most fundamental measure of housing well-being for first-time buyers in the state. (California Association od Realtors)

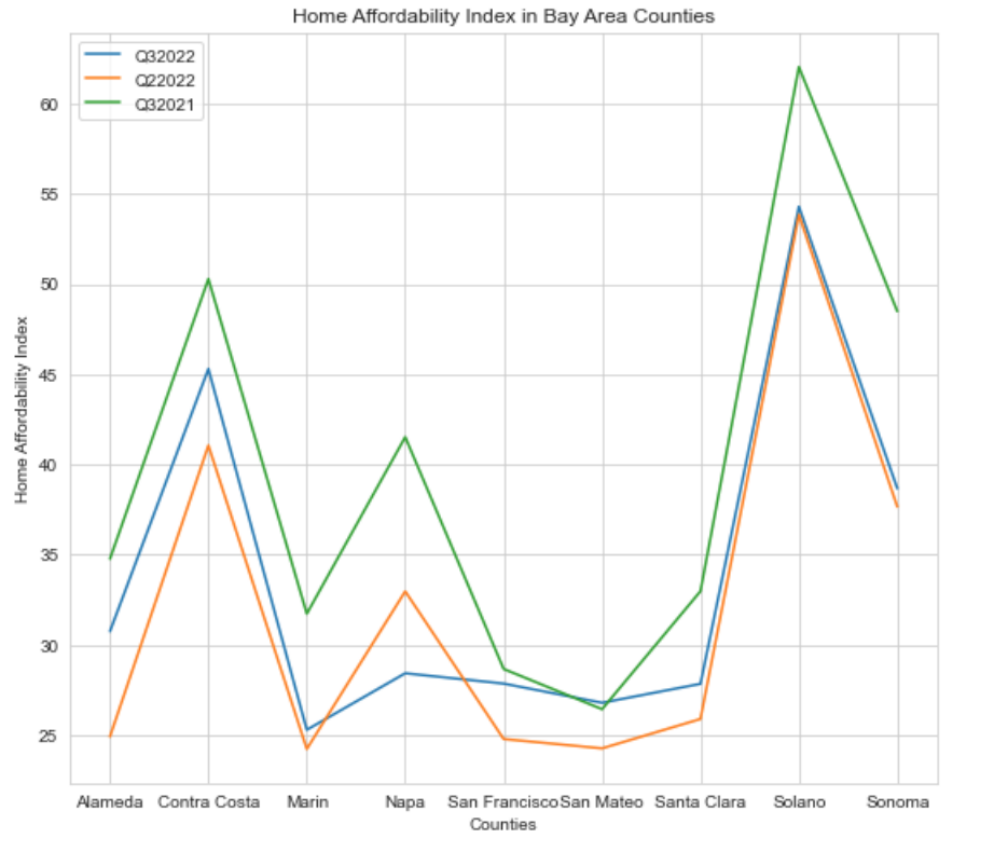


Chart 3. Housing Affordability Index in Each County for 2021 third quarter, 2022 second and third quarters.

Comparing HAI data of 2021 and 2022 affordability index dropped in all counties. Marin, Solano and Sonoma counties seems to recovering slower than others in HAI. The only county which HAI dropped significantly is Napa county. *From HAI perspective the best looking are Contra Costa and Solano counties.*

The last evaluated data is **Property Price Growth (%).**

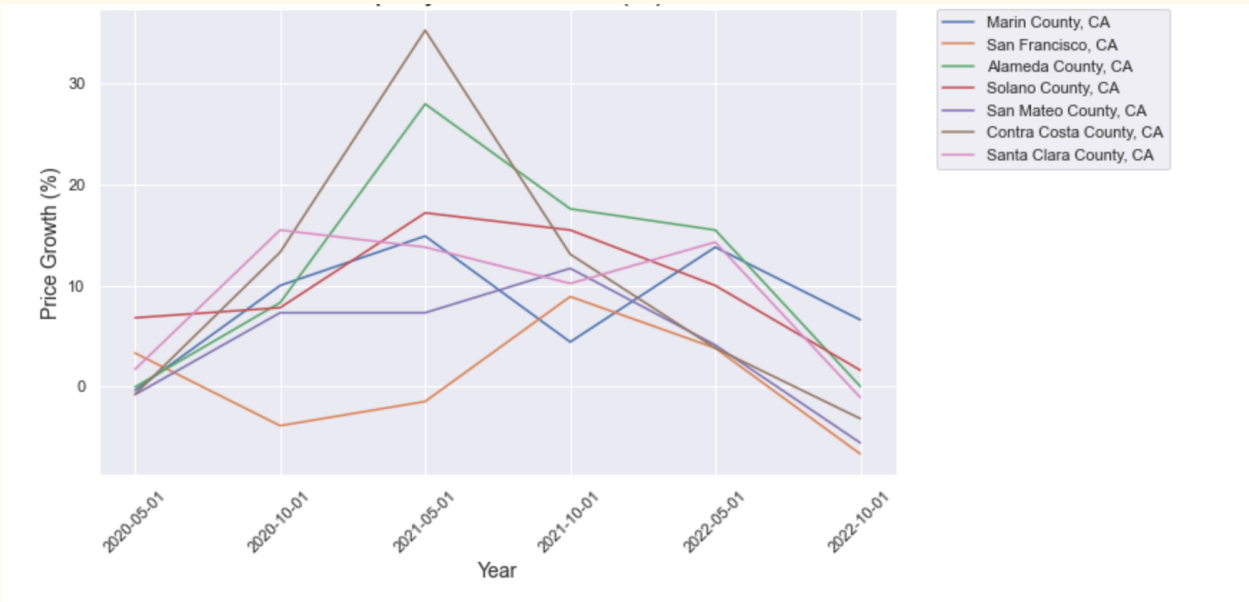


Chart 4. Property Price Change (%).

Interesting data is from Santa Clara county, where in the year of 2021 when most counties' property prices had increased, Santa Clara property prices kept dropping. In *San Mateo county* property prices stayed similar with low increase in 2021 October. From 2022 May to October all counties experienced price drop. Contra Costa county had the highest increase in 2021 May, however largest price drop in 2022 October.

Data Cleaning

Most cleaning required Zillow data. Cleaning was performed in these steps:

1. Removing columns not important for the price prediction and analysis. Columns 'daysOnZillow','dateSold', 'imgSrc', 'listingStatus', 'contingentListingType','country', 'currency', 'lotAreaUnit', 'hasImage', 'listingSubType.is_FSBA', 'listingSubType.is_openHouse' were removed.
2. Checking how many different values are in Construction Type and Property Type Columns. Construction type has three different type of values. Property type has 6 unique values. However, there are properties in the categories 'Lot', 'Multi Family', 'Manufactured'. Those columns were removed from data since we are interested only in

houses, not land, and not multi family houses. Manufactured category was removed due to highly different price than usual property.

3. Removed ‘foreclosre’ listings, since there were only 7 entires, as well as 3 entries of ‘bank owned’, 1 entry ‘for auction’, and 1 entry ‘listing is coming soon’.
4. ‘Lot Area’ values were non numeric, changed it to numeric. Columns ‘new construction type’ and ‘new home’, had nan values. Type from object was replaced to boolean and Nan values replaced to False.
5. After checking if all living area values are > 0 (needed for new column calculations), adding ‘Price per sqf’ column to the data frame.

Exploratory Data Analysis - EDA

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques which are mostly graphical to maximize insight into a data set, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions, develop parsimonious models, and determine optimal factor settings (Engineering Statistics Handbook. National Institute of Standards and Technology).

First, in the EDA checked the distribution of data by property type and cities.

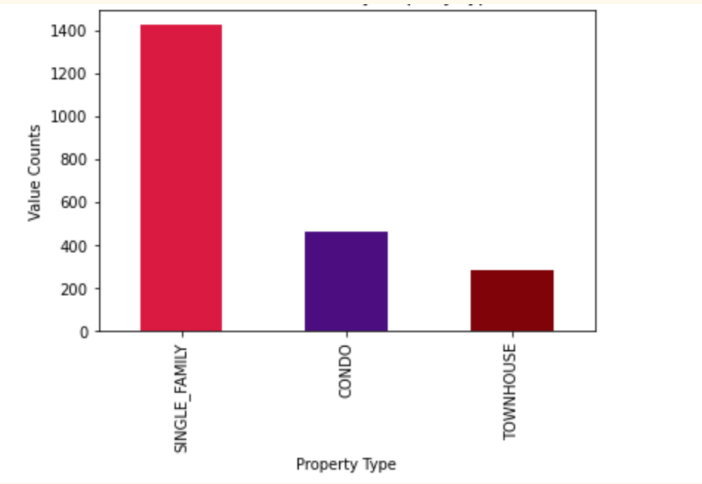


Chart 5. Distribution of Data by Property Type.

When checking distribution of data by property type, noticed that most of the data falls into single family category. Later in the EDA it is useful to see how price correlates with property types, and wether or not that correlation is different fromone another.

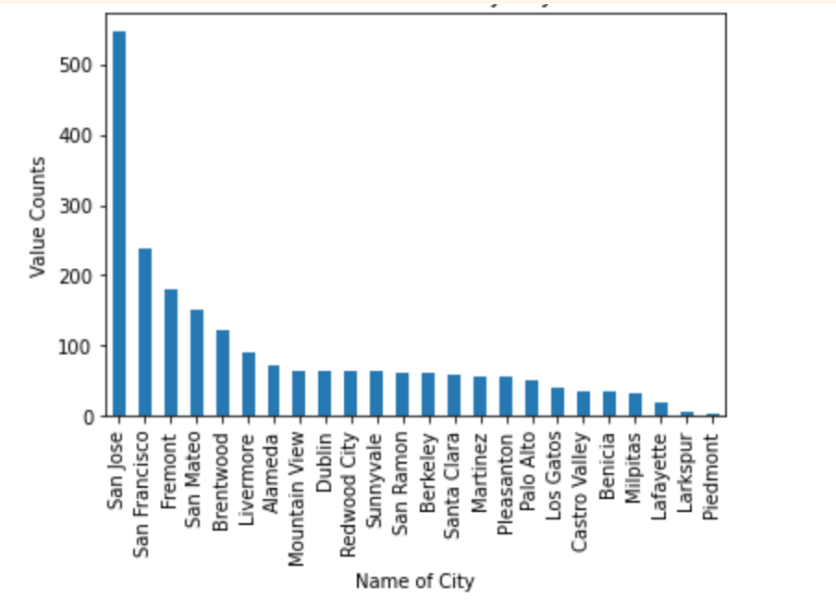
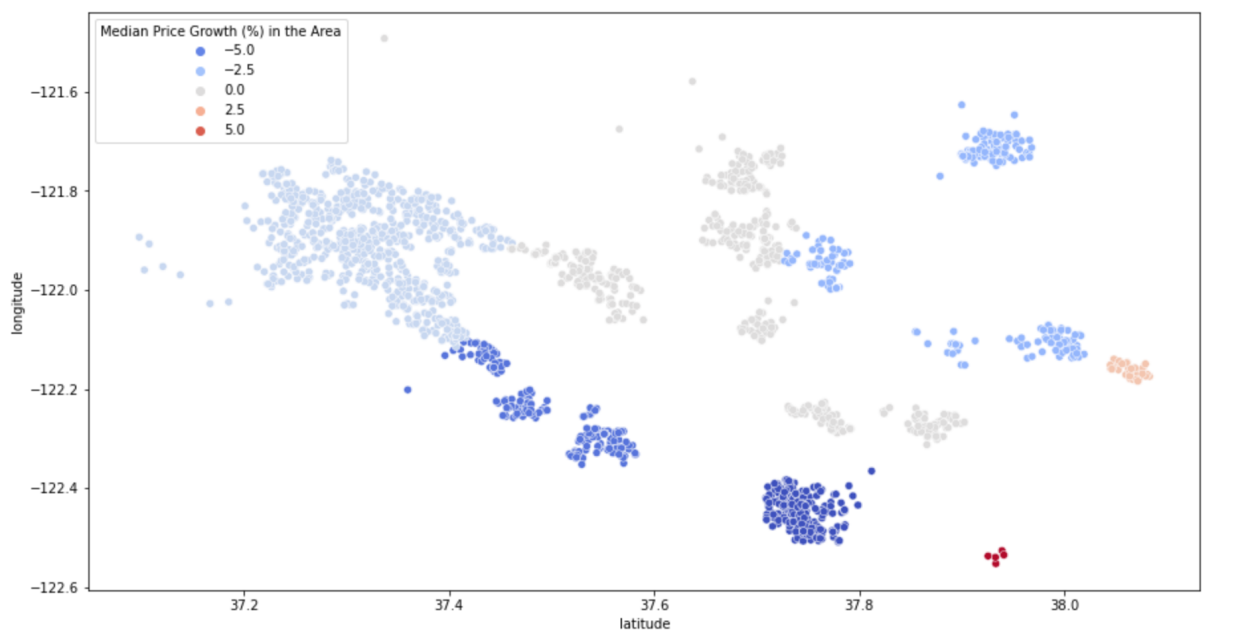
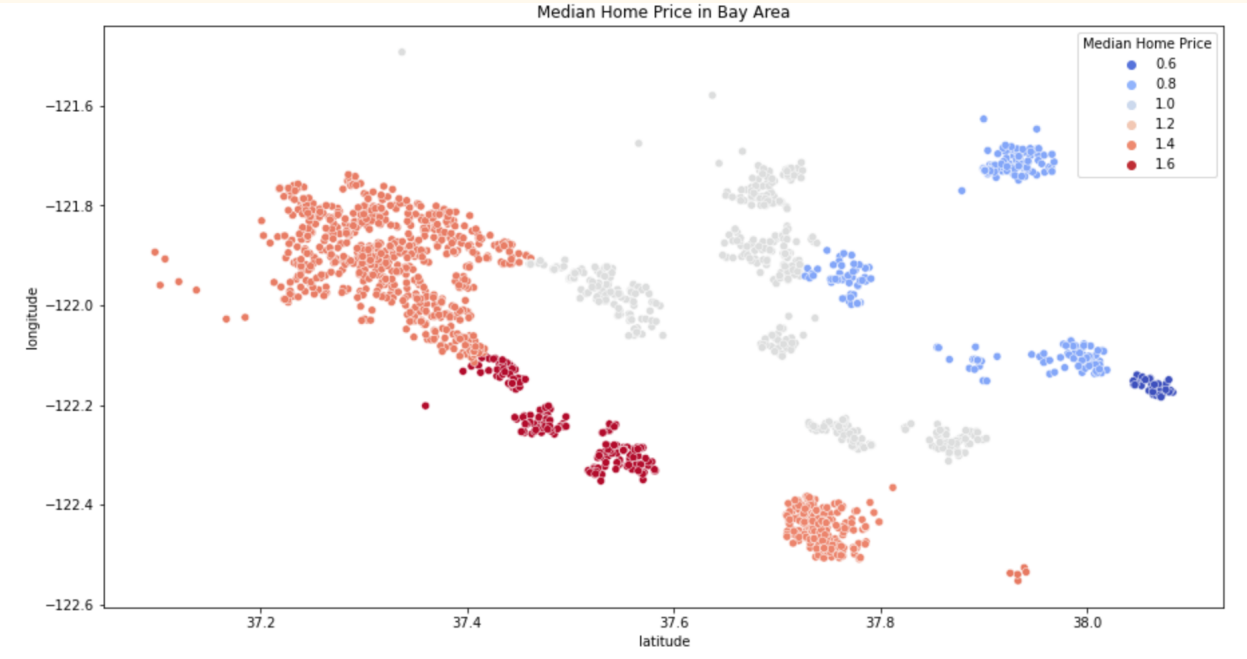


Chart 6. Distribution of Data by City.

Similarly, when checking distribution of data by city, one category has significantly more values than others. In this case San Jose city takes majority of properties listed on Zillow, more than 500. Piedmont, Lafayette and Larkspur having very few values. Decided to keep this in data frame, and not remove, because it other features of these properties may be useful for calculations. However, having in mind that those cities feature importance or correlation may be incorrect due to the amount of values.

Location is the factor which determines real estate price. Geographical distribution of real estate prices in San Francisco Bay Area give interesting results. While median home prices are highest in Redwood City, Palo Alto and San Mateo, highest price growth is recorded in San Jose. However, looking at the price per sqf range we can not see that significant difference.


Median Price Growth (%) in the Area
−5.0
−2.5
0.0
2.5
5.0
longitude
−121.6
−121.8
−122.0
−122.2
−122.4
−122.6
37.2
37.4
37.6
37.8
38.0
latitude


Median Home Price in Bay Area
Median Home Price
0.6
0.8
1.0
1.2
1.4
1.6
longitude
−121.6
−121.8
−122.0
−122.2
−122.4
−122.6
37.2
37.4
37.6
37.8
38.0
latitude

Picture 1. Median Price Growth, Median Home Price, and Price per sfq Distributions in the Area.

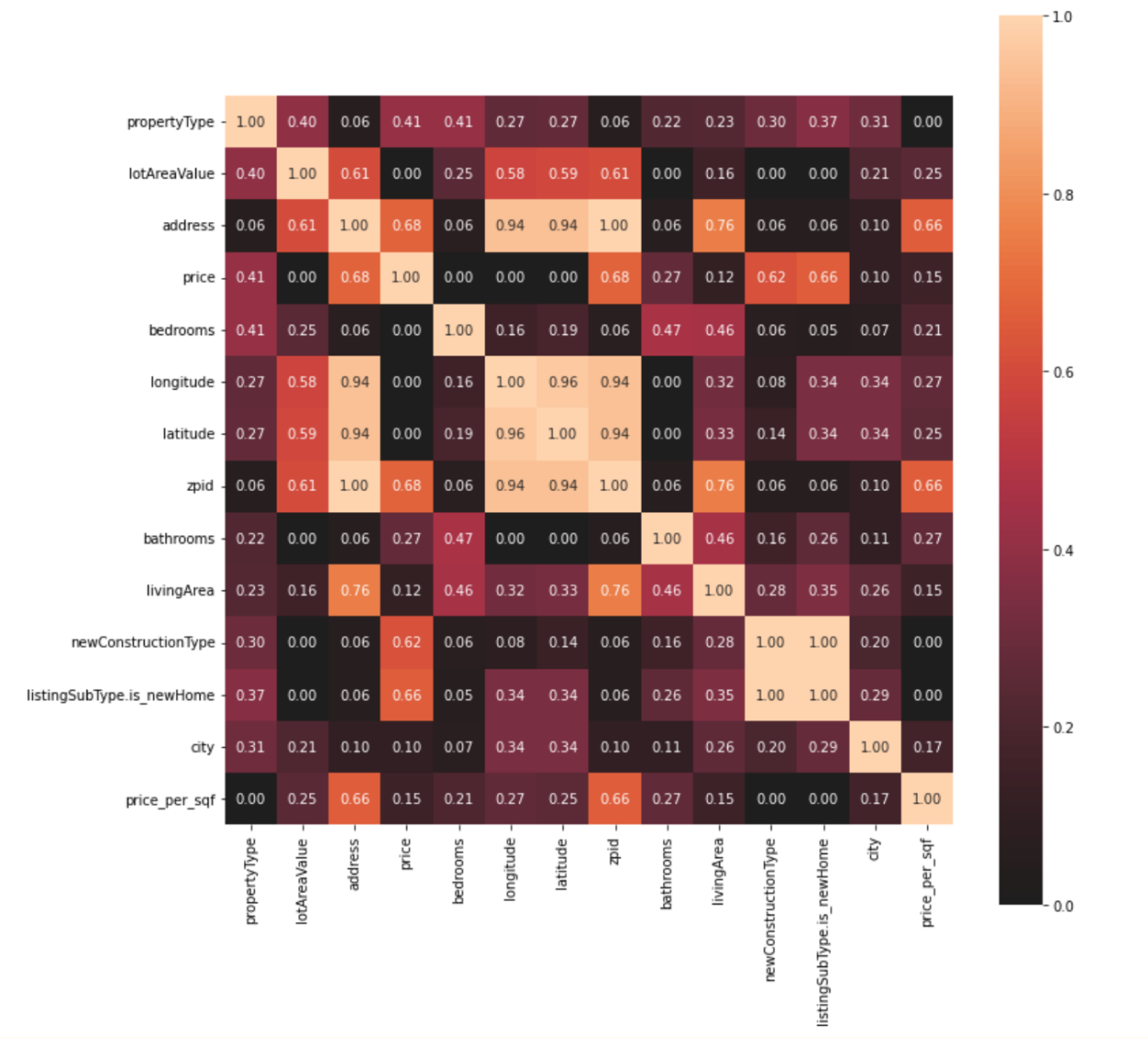


Chart 7. Zillow Data Correlation Matrix.

One of the most useful charts which can give most information at once is correlation matrix. First, was compared only Zillow data features, then all combined. Looking at Zillow correlation matrix most important numerical values correlations are:

price/living area = 0.57

price/bathrooms = 0.41

price/bedrooms = 0.46

bathrooms/living area = 0.7

bathrooms/bedrooms = 0.56

bedrooms/living area = 0.69

Highest categorical values correlations with price:

price/listing sub type is new home = 0.66

price/new construction type = 0.62

price/zpid(zillow property id)=0.68

price/address = 0.68

price/property type = 0.42

Algorithms and Machine Learning

For the machine learning model preparation data needed to be cleaned once more. Data had a lot of price related values, such as: price per sqf, property price, price growth.

Next step, was to create **dummies for categorical values**. Categorical values for dummies: 'Property Type', 'New Construction Type', 'New Home','City', 'School Rating'. *Dummy variables qualitative variables or discrete variables that represent categorical data and can take the values as 0 or 1 to indicate the absence or presence of a specified attribute respectively. Dummy variables are also known as indicator variables, design variables, and binary basis variables.* [source]

Price prediction considering multiple features requires to apply regression machine learning model. To prepare data better used StandardScaler method from scikit learn preprocessing module. *Standard scaler standardize features by removing the mean and scaling to unit variance. The standard score of a sample x is calculated as :*

z = (x-u)/s

Where u *is the mean of the training samples or zero if* `with_mean=False`*, and s in the standard deviation of the training samples or one if* `with_std=False` [source].

Data was split into training (80%) and testing set (20%). First model selected is Random Forest Regressor. *A random forest is a meta estimator that fits a number of classifying decision trees on*

various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree [source].

Training data on Random Forest Regressor gave good results, with r2 score 0.71 (71%). Features, having most significance to the price prediction: Palo Alto city, longitude, latitude, living area, lot area, single_family property type and unemployment rate.

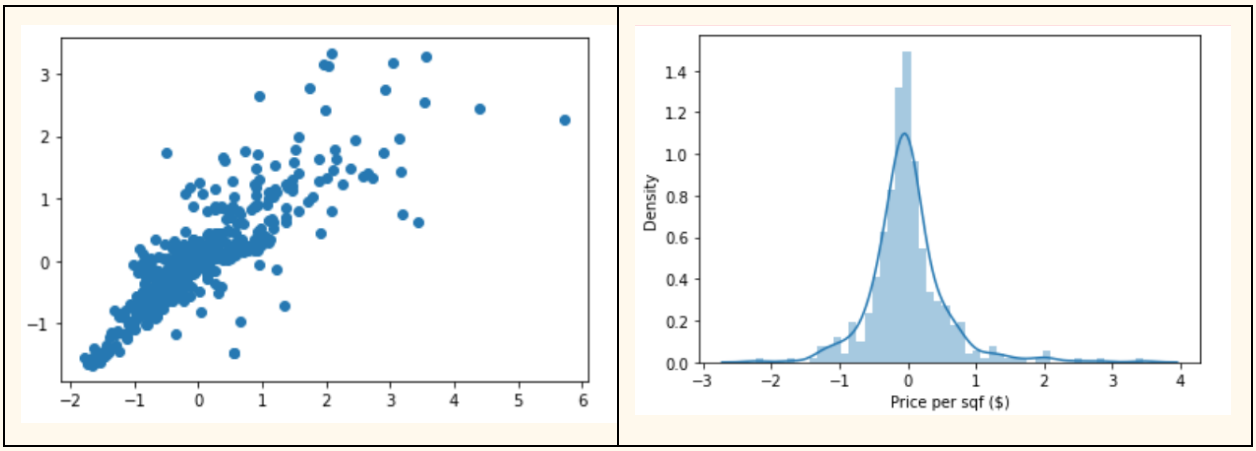


Chart 8. Random Forest Regressor results. Predicted price vs actual price, and predicted price normalization.

Using Grid Search CV and checking on best estimators, results were almost same.

Using XGBoostRegressor r2 score was also similar, however, predicted prices normalization chart showed better results than Random Forest Regressor.

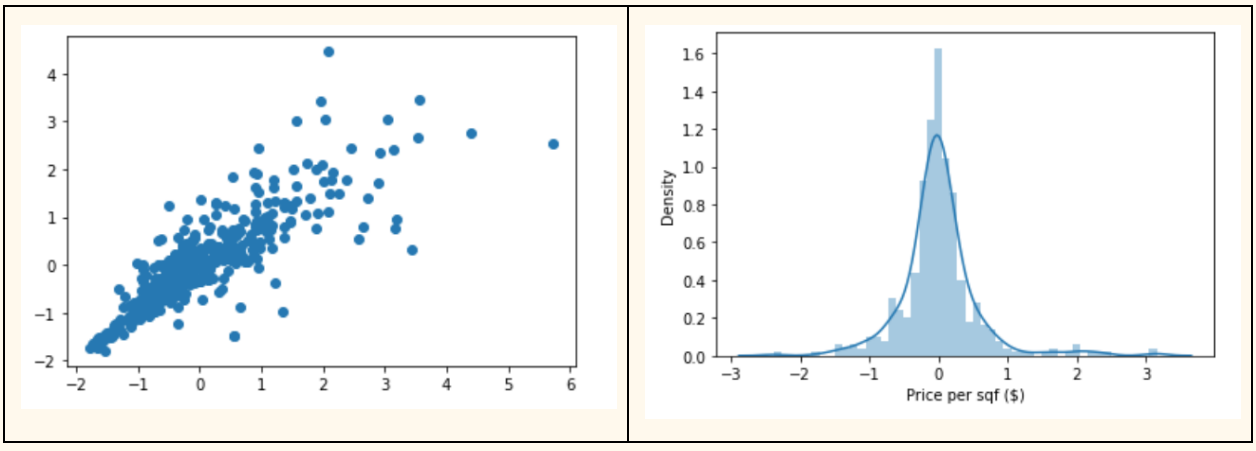


Chart 9. XGBoost Regressor results. Predicted price vs actual price, and predicted price normalization.

Tried manipulate with a data a bit more by removing most important features of longitude and latitude, trying models only on city and only on counties data. Those data sets did not give as good results as the first dataset. R2 scores varied between 0.49 and 0.62.

Although it is unlikely that linear regression would give better results than Random Forest Regressor I wanted to give a shot trying it on dataset. LInear Regressor results: r2 score=0.57. Also, tried Linear Regressor using 6 folds. Best results were from 6 splits of data, and random state=42. R2 for 6 folds Linear Regressor is 0.65.

Choosing a Model

Despite the manual calculations of the scores and errors of machine learning model, I chose to use cross validation score for model selection.

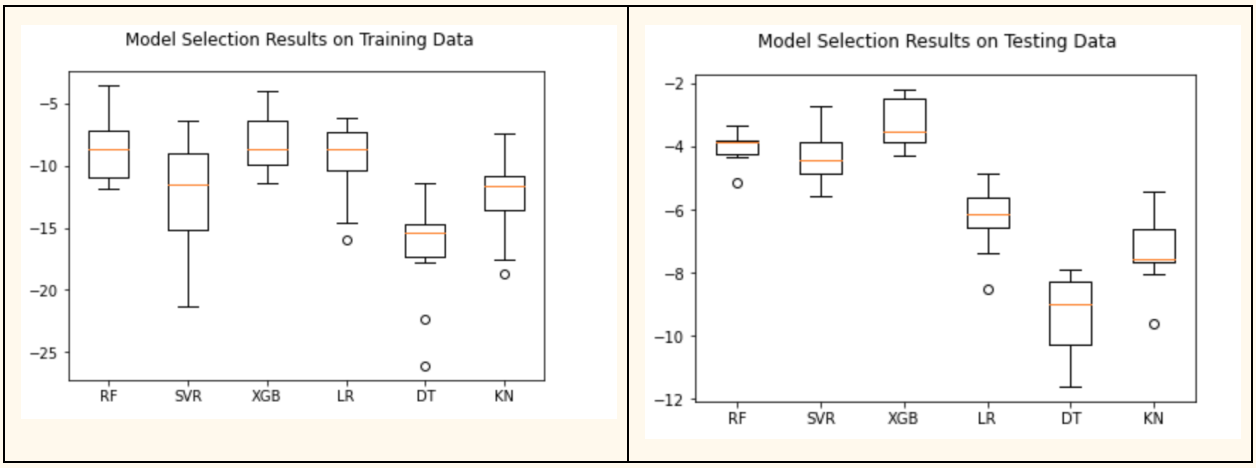


Chart 10. Machine Learning Model's Negative Mean Squared Error.

Best results with the smallest mean squared error was given by XGBoost Regressor. With the r2 score for this method also having best results, model selected for machine learning on the data set.

Predictions

Predicted Price per sqf	Lot Area (sqf)	Property Price ($)	Bedrooms	Bathrooms	Living Area (sqf)	Median Price Growth (%) in the Area	Median Home Price	Monthly Payment ($)[Tax&Ins])	Minimum Qualifying Income ($)	Total Crimes per 1000	Property Type	School Rating	City	County
$549	2.2	1040000	3	3	2131	-3.2	753310	4320	129600	25.09	Single Family	A-	Brentwood	Contra Costa
$924	8507.268	2850000	4	3	2860	-5.6	1641350	9420	282600	28.29	Single Family	A+	San Mateo	San Mateo
$518	3049.2	825000	3	3	1515	-3.2	753310	4320	129600	19.44	Townhouse	0	Martinez	Contra Costa
$758	2996.928	1850000	4	3	2508	-6.7	1411000	8100	243000	66.99	Single Family	A-	San Francisco	San Francisco
$548	4791.6	999000	4	2	1804	0	1071000	6150	184500	62.26	Single Family	A+	Berkeley	Alameda
$1,419	5928.516	2398000	3	2	1610	-5.6	1641350	9420	282600	28.29	Single Family	A+	San Mateo	San Mateo
$943	2613.6	949000	2	2	1035	0	1071000	6150	184500	37.5	Single Family	A	Alameda	Alameda
$419	566.28	1025000	5	3	2522	-3.2	753310	4320	129600	19.44	Single Family	A-	Martinez	Contra Costa
$1,049	5662.8	2800000	4	3	2623	0	1071000	6150	184500	37.5	Single Family	A	Alameda	Alameda
$750	0	759850	2	1	947	-1.1	1434800	8230	246900	24.46	Condo	A+	Sunnyvale	Santa Clara

Table 2. 10 random predicted results.

Lowest price per sqf - 418 USD, in for 566 sqf lot area, 5 bedrooms/3 bathrooms single family house in city of Martinez, Contra Costa county, with school rating A-. Total value $1,025,000

Highest predicted price per sqf - 1419 USD, in 5928 sqf lot area, 3 bedrooms/2 bathrooms, single family house in city of San Mateo, San Mateo county, with school rating A+. Total value: $ 2,398,000

Mid range predicted price per sqf:

750 USD, in 0 sqf lot area, 2 bedrooms/1 bathrooms, condo in city of Sunnyvale, Santa Clara county, with school rating A+. Total value: $ 759,850.

758 USD, in 2996.928 sqf lot area, 4 bedrooms/3 bathrooms, single family house in city of San Francisco, San Francisco county, with school rating A-. Total value: $ 1,850,000.

Future Improvements

- For the better results 2 types of zillow data could be evaluated separately: sold properties, properties currently listed, and later compared to zillow zestimate values.
- To have better distributions between cities, and property types more data could be selected to have similar distributions.