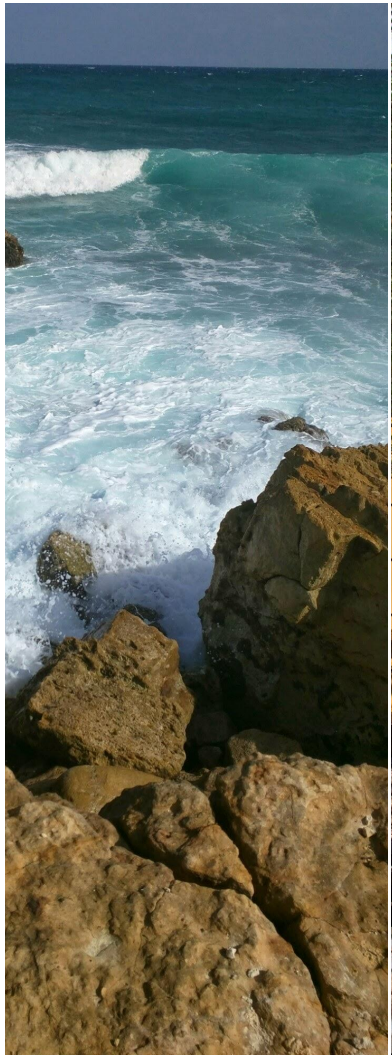
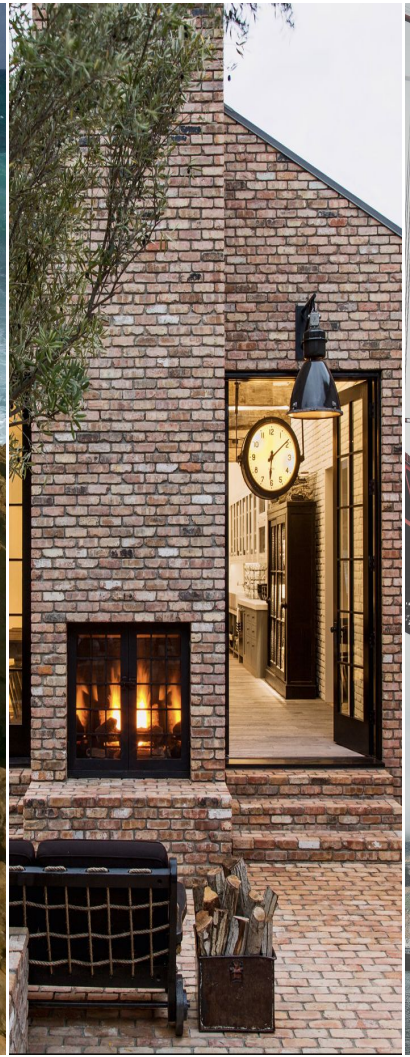




Bid Your Dream House

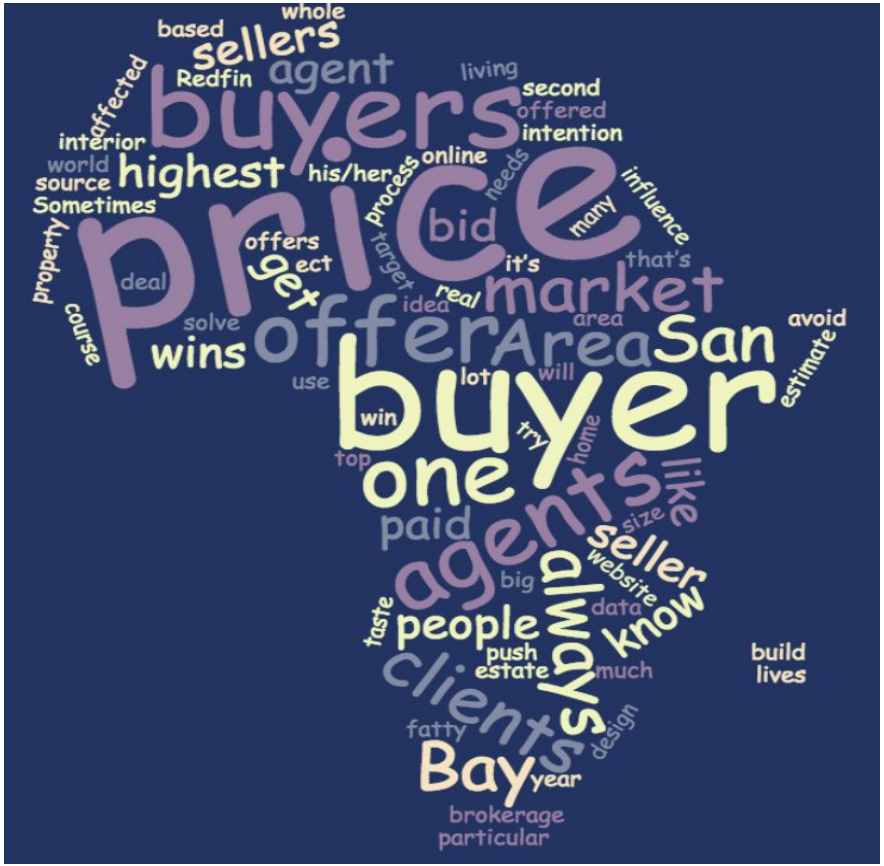
--San Francisco Bay Area House Price Prediction

Leilei Liu • June 2018

Contents

Introduction

Home buyer has limited resources to offer a reasonable bid price



Dataset

- **Property data is from Redfin.com**
- **City boundary shapefile is from ca.gov**

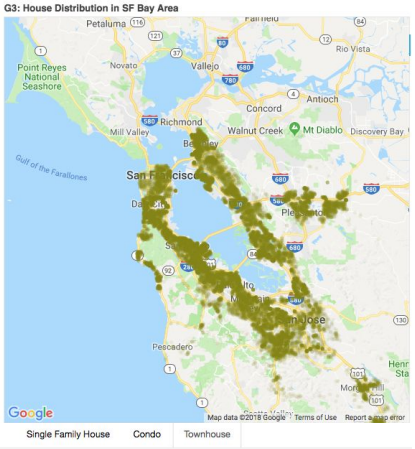
Data Facts

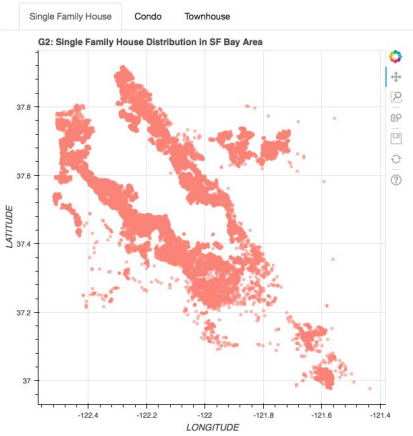
- **Data volume: After data cleaning, there are 32416 rows and 16 columns in the dataset.**
- **Data Dictionary**

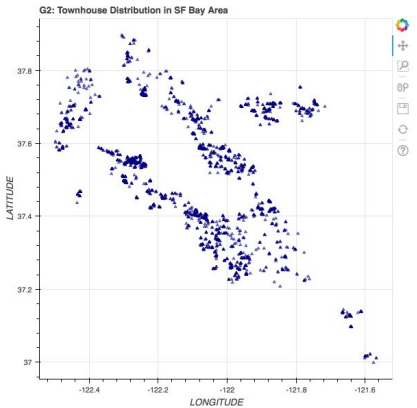
Column Name	Description	Type
SOLD DATE	On which date the property was sold	object
PROPERTY TYPE	Single family house/ townhouse/ condo	object
ADDRESS	Property address	object
CITY	City of the property	object
ZIP	Zip code of the property	object
PRICE IN K	Sold price of the property in thousand	float64
BEDS	Number of bedrooms	float64
BATHS	Number of bathrooms	float64
SQUARE FEET	Living space of the property	float64
LOT SIZE	Lot size of the property	float64
YEAR BUILT	Year of the property was built	float64
$/SQUARE FEET	Sold price/square feet	float64
HOA/MONTH	HOA per month	float64
LATITUDE	Latitude of the property	float64
LONGITUDE	Longitude of the property	float64
COUNTY	County of the property	object

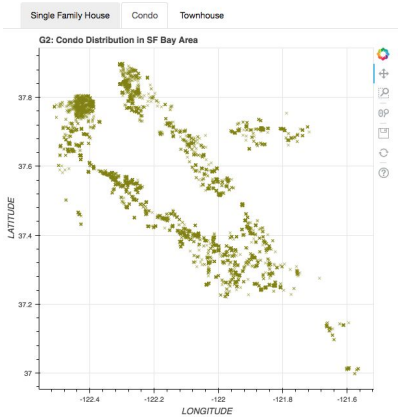
Exploratory Data Analysis

Property distribution



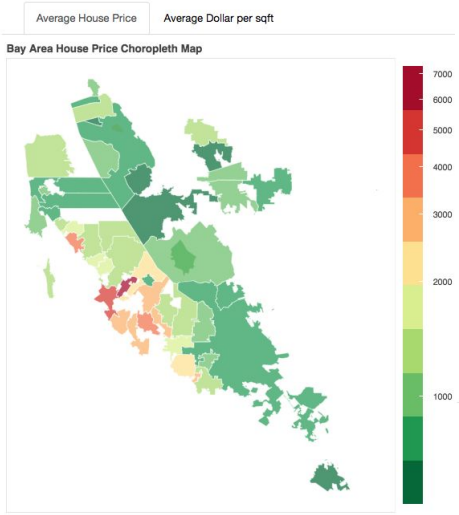


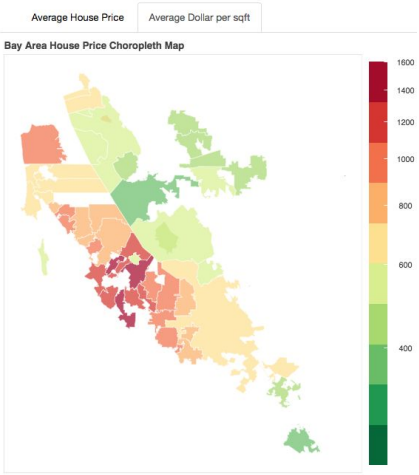




Exploratory Data Analysis

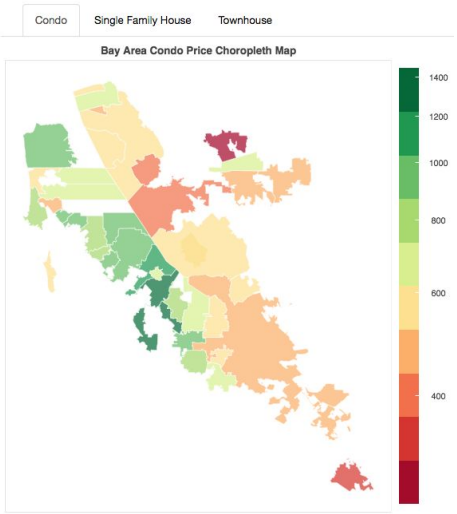
Average total price vs. Average Unit Price ($ per sqft)

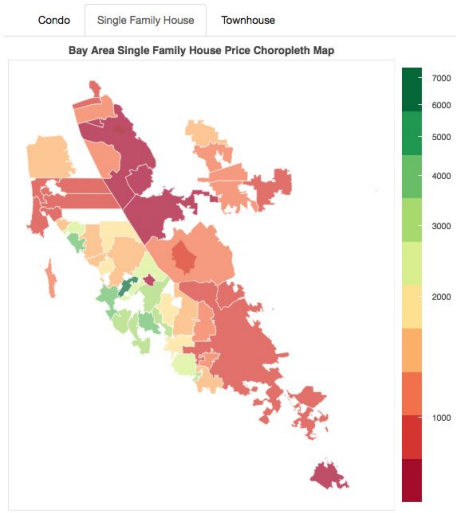


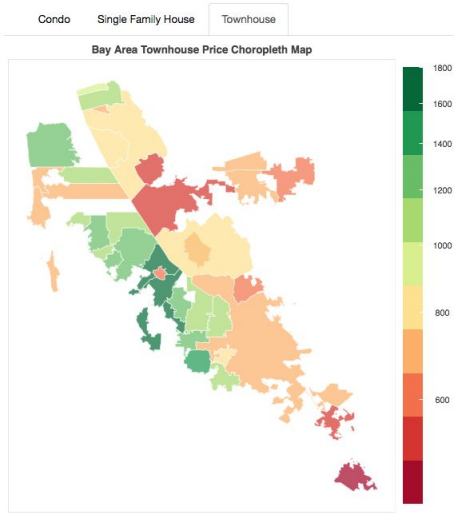


Exploratory Data Analysis

Price Based on Property Type

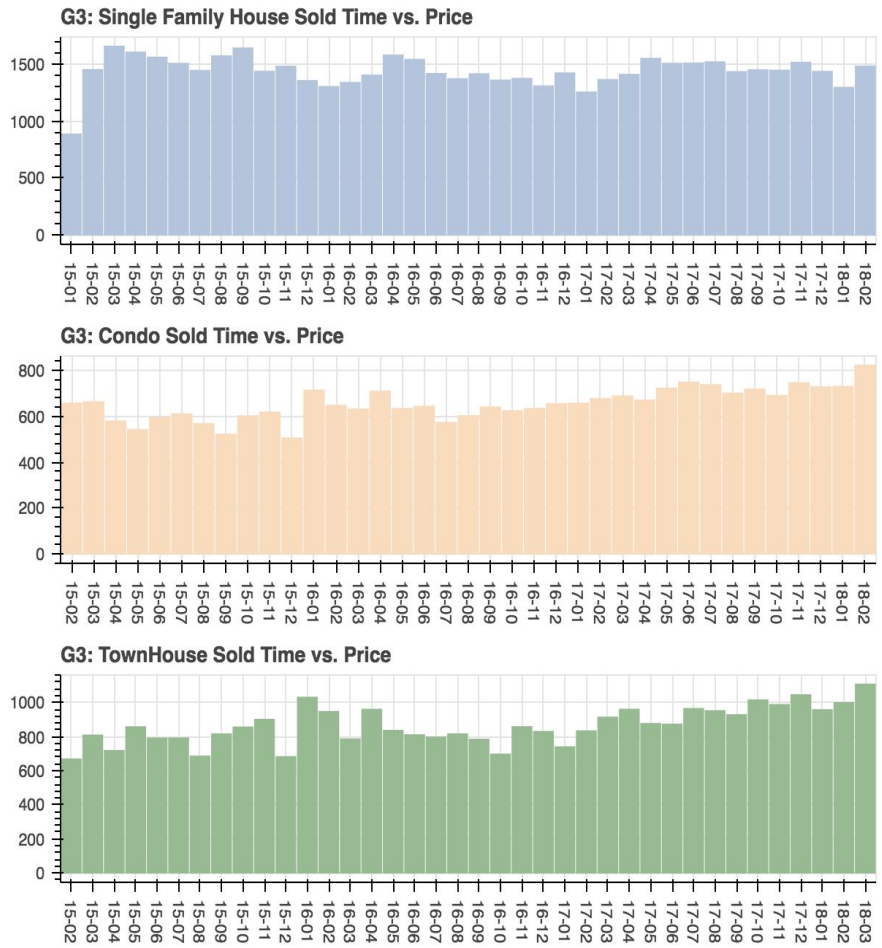






Exploratory Data Analysis

Average Price Trend by Month



Exploratory Data Analysis

Price by County and Property Type

Machine Learning

Linear Regression

- Train three different models for each property type (single family house, townhouse and condo);
- Train three different models based on three different types of location (zip, city, county);
- General model contain different types of property and all types of location.

Conclusion

Analysis

Best model is using three Ridge regression model on different property types.

		SFH	TH	Condo
LinearRegression()	RMSE	615.9830	163.7677	143.7450
	R Squared	0.7042	0.8461	0.8499
RidgeCV()	RMSE	**615.1894**	**163.2328**	**143.7823**
	R Squared	**0.7042**	**0.8460**	**0.8496**
	Alpha	**0.0050**	**0.0116**	**0.0202**
LassoCV()	RMSE	615.5369	163.4276	144.0289
	R Squared	0.7041	0.8457	0.8486
	Alpha	0.0066	0.0202	0.0202

Conclusion

Limitation

- Some cities with few sold property in the dataset may have less accurate prediction.

- If the dataset contains neighborhood name and remodel year, it would help us improve the accuracy.