Capstone Project 1 Milestone Report

The Investment Property Deal You Will Never Miss

Mark Zhao

02/14/2020

Executive Summary

Using scraped and cleaned data from Redfin and Airbnb, I designed a machine learning model that can predict home price and rental income. Given a home's features like beds, baths, type, and location, the model is able to predict the Return on Investment (ROI) and find the best deals of investment property.

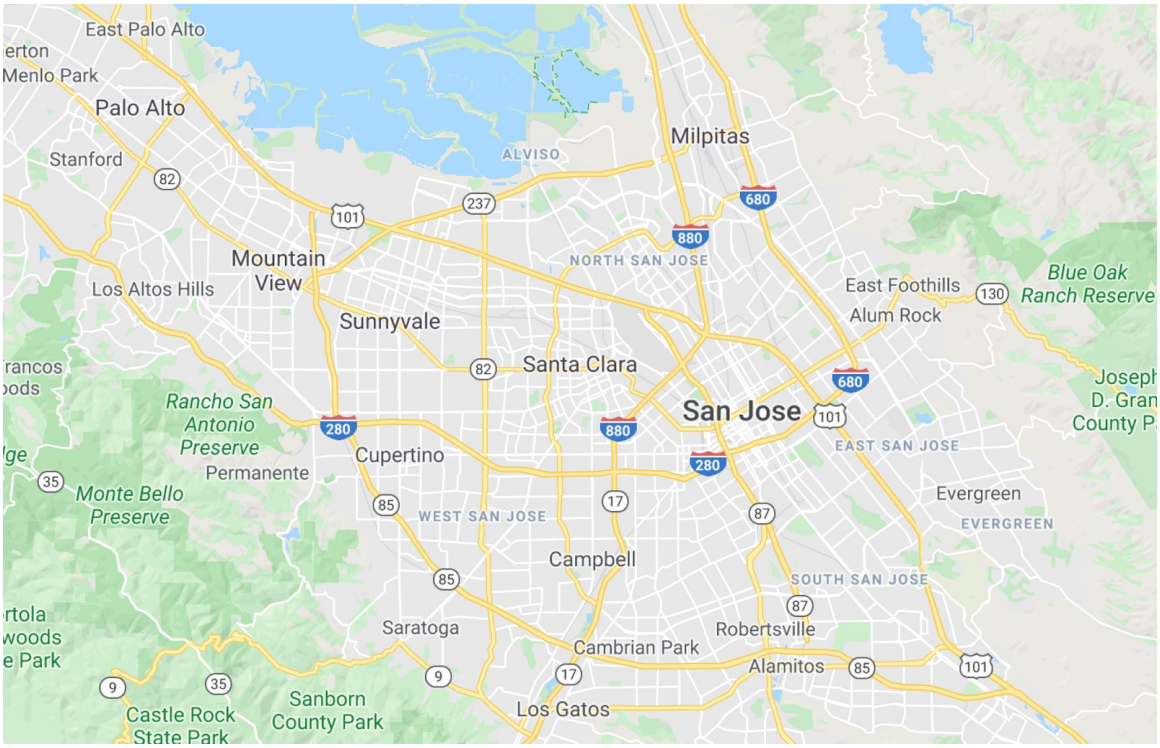
- Data suggest that as an rental property investor, you should avoid neighborhoods with good schools and focus on the ones with great location and size.

- Linear regression, Ridge, LASSO, and Random Forest models were trained and tested with Random Forest being the winner yielding test R^2 = 92.5% for home price and 75.9% for rental income.

- Compared to the median ROI which is around 7%, the model can identify investment properties with yields > 14%, or twice the return on investment.

- The 90th percentile of all predicted cap rates is 11.6%, about 4% higher than the median. Compounded for 10 years with an initial investment of $1 million, this means $900,000 more wealth which makes quite a big difference.

Problem Statement



If I have a budget of $400,000, just by looking at this map, can you tell me where I should buy a property to get the best return on investment?

The old fashioned way, you ask your friends and they tell you to go buy a home in Cupertino because the school is great and the bubble tea there is awesome. Or, you ask your realtors and they tell you to go buy a home in Austin because well, everyone is doing it.

Now, I'm going to introduce a revolutionary tool that could change the game of rental investment. Imagine you can browse thousands of homes that are sorted by return on investment and tell you which are the best deals to grab? This is a whole new experience that gives you many more choices, more flexibility, and more information so you can get a clear and personalized recommendation of investment property.

Sure, there were some attempts to solve this issue but their rental income values rely on zip code level data which can be very inaccurate. Plus, as Airbnb has become more and more a mainstream way for lodging, more homeowners are considering listing their vacant rooms on Airbnb as a source of passive income. As a homeowner myself, I'm looking for an investment property that I can get the highest ROI. Where and what kind of property should I buy so I can get a higher rental income with lower initial investment? Are distances to work or landmarks

more important than schools and crime? There are no such tools available so I decided to make my own!

There are some other interesting topics that I would like to explore. For example, anecdotes suggest you can get the best deals if you buy homes in Q4. Is it a myth or actually is there risk premia that can explain it?

Data

Home Sale Price and Features

- Data Collection and Cleaning
  29919 properties sale price and characteristics in Santa Clara County from 09/2017 to 09/2019 were collected from Redfin.

  City names were a mixture of lowercase and uppercase and thus duplicate observations were removed. Zip codes were substring to 5 letters and the duplicates were removed. After removing irrelevant features we are left with: sale price, sold date, property type, zip code, beds, baths, square feet, lot size, year built, HOA, and coordinates.

- New Features
  To test the hypothesis that home sale prices are affected by seasons, 4 dummy variables extracted from the sale date were added to indicate whether the sale happens in Q1, Q2, Q3, or Q4. The total size of living square feet and lot was added to capture the combination of living size and lot size. The number of sales in each zip code was added to proxy the supply of inventory.

- Outliers and Missing Values
  Multi-Family (5+ Units) and land were removed because they tend to be a commercial office or big apartments with extremely high prices. Only property types SFH, Townhouse, Condo and Multi-Family (2-4 Unit) were included. Observations with price < 100K, or beds/baths < 1, or square feet < 10 were removed due to data errors.

  Since beds, baths, square feet, lot size, year built are the main features, observations with missing values were removed (<5% of total population). For HOA/MONTH, 70% of the observations are NaNs and the median HOA by property type were filled then.

Airbnb Rental Price and Features

- Data Collection and Cleaning
  7196 Airbnb listings as of 07/09/2019 in Santa Clara County were collected from inside-airbnb.com. There are more than 100 features for each observation and only the

relevant ones were left. Host listings count was capped so it's at least 1.Prices were converted to numeric values.

- New Features
  The # of days between host since date and the scraped date is created to measure the # of days hosted as a proxy of host experience. The # of days since the first review and since the last review were calculated respectively. The amenities text string was converted to individual dummy variables and the irrelevant ones were removed. Also, the # of amenities is created as another new feature. A new dummy variable was created to indicate if a listing has review scores or not.

- Outliers and Missing Values
  Observations with 0 bed (which is super unlikely) were changed to be the same as the # of bedrooms or at least 1. Extremely high maximum nights were winsorized with 2000 days.

  Host response time and response rate missing values were replaced by the mode since these two features are categorical. Missing zip codes (5%) were reverse geocoded from Texas A&M geocoder given their coordinates. The missing review scores were replaced with median scores. Missing review dates and reviews per month were filled in with 0. Missing cleaning fees and deposits were replaced with median values. Observations with missing baths and beds were removed (very rare).

Landmark Coordinate Data

17 landmarks' coordinates data were collected from Google Map to calculate the geodesic distance from each property/listing to these landmarks including the big tech companies like Google, Apple, Facebook, downtowns like San Francisco, San Jose, Palo Alto, Mountain View, Sunnyvale, airports like SFO and SJC, and others like Levi's stadium.

School Score Data

The school scores data (percentage standard met and above) for all students' Math and English tests were extracted from CAASPP and were averaged on the zip code level.

Crime Data

The violent crime (including murder, rape, robbery, and aggravated assault) and property crime (theft, burglary, and arson) index data were collected from bestplaces.net on zip code level. Ideally, a big matrix of distances from each property to each crime can be calculated but the computing efforts are too big to be justified for this project.

Exploratory Analysis (Home Price Data)

Exploratory Data Analysis

A histogram of all homes sold in Santa Clara County from 09/2017 to 09/2019 as in Figure 1 suggests the distribution is normal but skewed to the right. In other words, there are many extremely high prices and if you look at the summary statistics below, the max sold price is $23,495,000!

count	mean	std	min	25%	50%	75%	max
24280	$ 1,392,570	$ 896,466	$ 107,000	$ 860,000	$ 1,165,000	$ 1,600,000	$ 23,495,000

Figure 1.

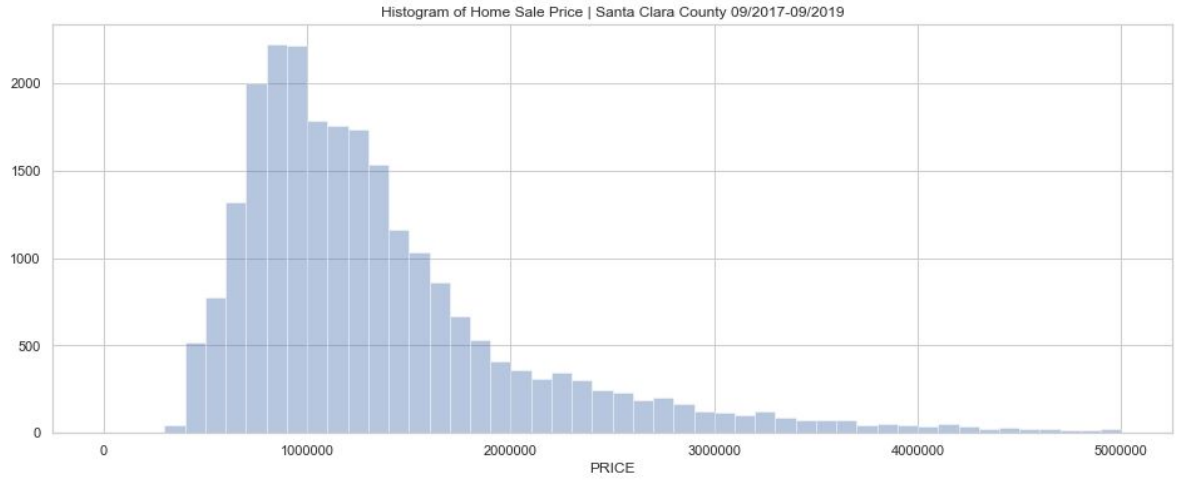
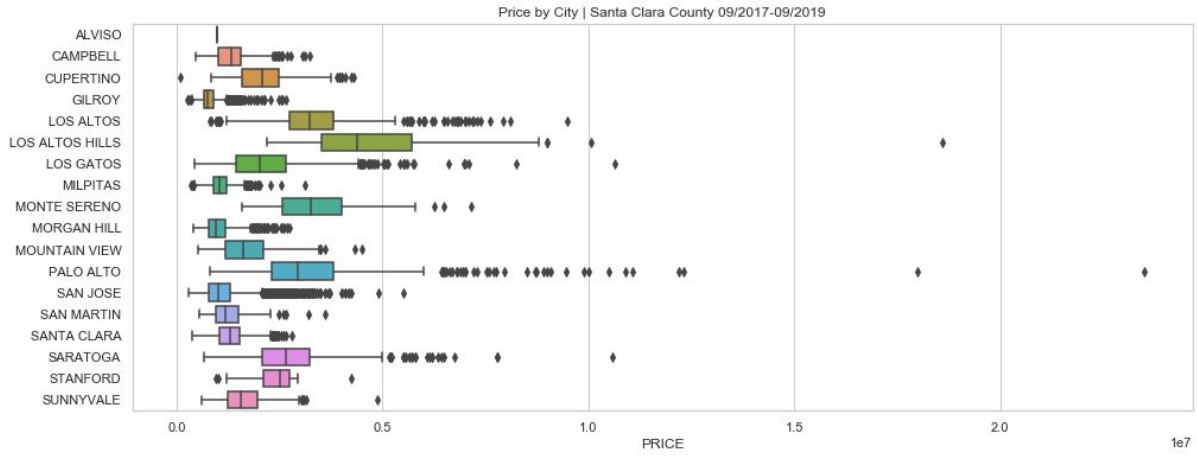


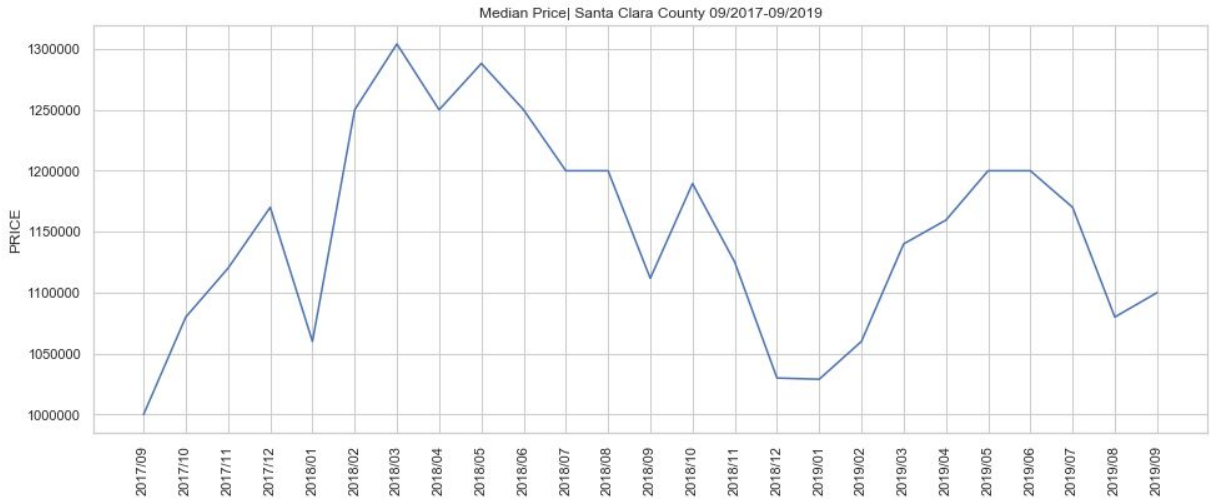
Figure 2 illustrates the distribution of sold prices by city and neighborhoods like Los Altos, Los Altos Hills, Palo Alto, Saratoga, and Monte Sereno exhibit higher median price and outliers which is consistent with the expectation.

Figure 2.



A time series plot of the median price in Figure 3 suggests there might indeed be a seasonality effect: sale prices tend to rise when it's getting warmer and then fall as winter comes.

Figure 3.



Story

Anecdotes suggest if you buy a home at the end of the year you tend to get a deal. On the other hand, you pay a premium (probably through a bidding war) in late spring to early summer for your home. Is it a myth or is it really true?

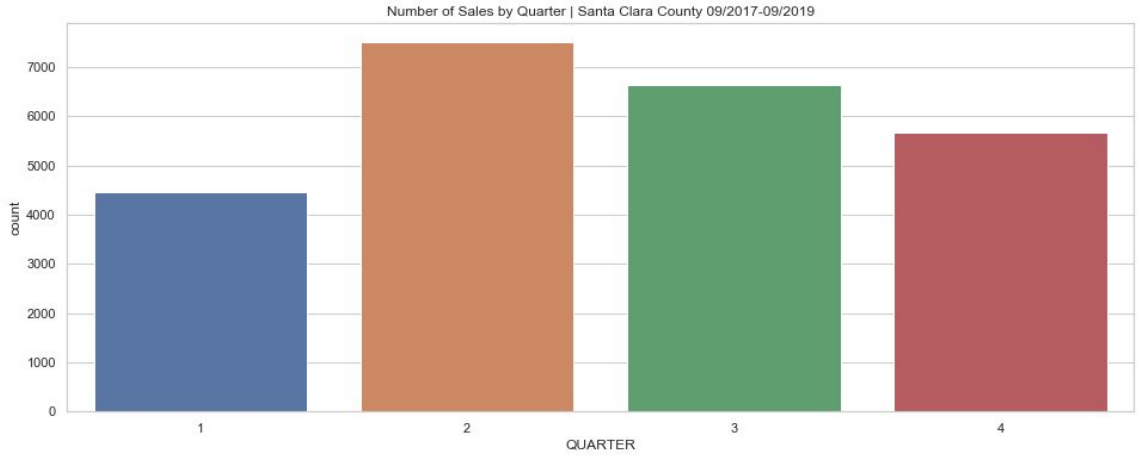
Hypothesis

Null Hypothesis: There are no differences in price between homes sold in Q2 and Q4.

Exploratory Data Analysis

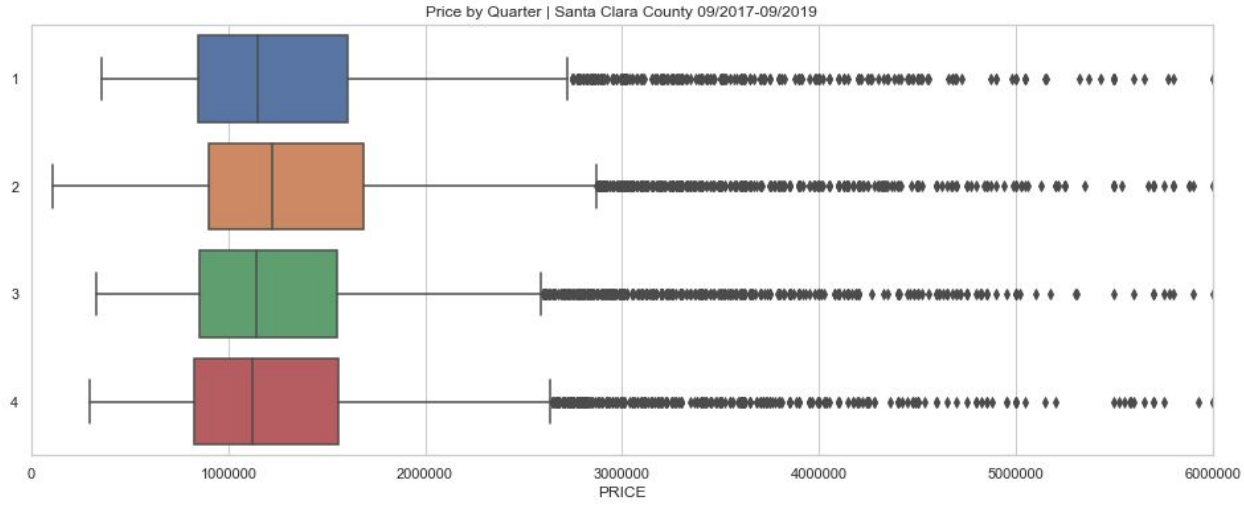
First, we would want to learn the number of sales in each quarter. Not surprisingly, there are more sales in Q2 than any other quarter as shown in Figure 4.

Figure 4.



If we look at price by quarters, the median, 25th and 75th percentile price is higher for Q2 than any other quarter (see Figure 5). This suggests the anecdotes might be true.

Figure 5.



But wait! Before we jump to conclusions, we need to think if this is because we failed to control other factors? Maybe there are more condos/townhouses sold in Q4 than in Q2 and thus lower prices? To make our analysis a little more rigorous, I limit my two samples by property type and zip code. For example, if we only compare single-family houses price in zip code 95123, the prices are still lower in Q4 and not just mean but also median, 25th, 75th, and max.

	count	mean	std	min	25%	50%	75%	max
Q2	264	$ 1,070,864	$ 147,728	$ 750,000	$ 960,000	$ 1,060,000	$ 1,171,250	$ 1,575,000
Q4	185	$ 979,967	$ 122,915	$ 750,000	$ 901,000	$ 950,000	$ 1,050,000	$ 1,498,000

Statistical Data Analysis

This boosts our confidence a little bit. But we still need to run some more analysis. The histogram and probability density functions of the two samples also suggest Q4 is cheaper as shown in Figures 6 and 7.

Now, it's time to run a t-test to determine if there's a statistically significant difference between the means of two samples. It turns out that the t value is 7.09 and the p-value is close to 0. **This suggests the null hypothesis is rejected with a level of significance level of 0.01.** On average you get an average discount of almost $100K if you buy a house in Q4 vs. Q2! This sounds like a huge deal! And by limiting the sample to other property types (like townhouses) in other zip codes it still suggests the same story.

Figure 6.

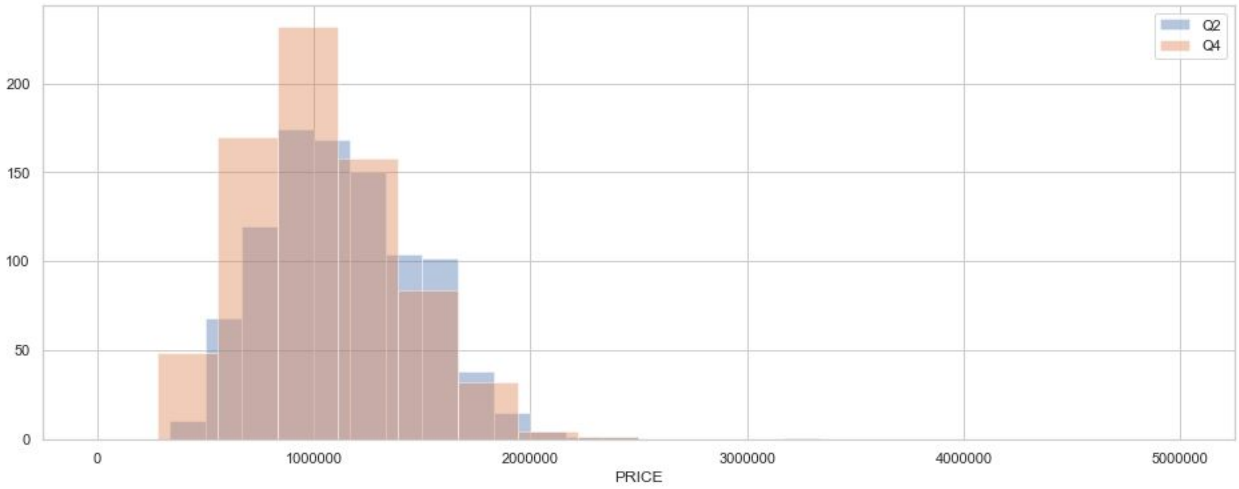
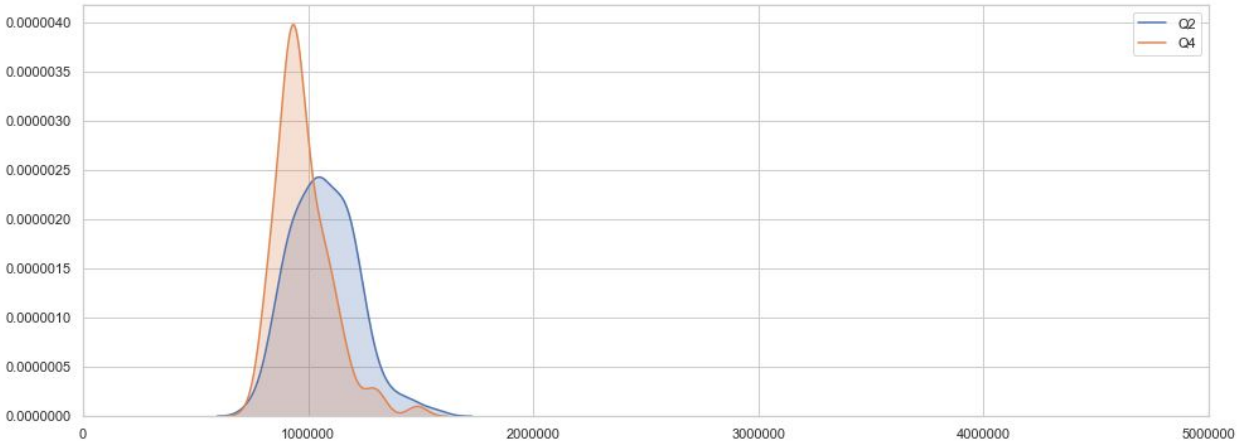


Figure 7.



Caveat

We still only control two features. Ideally, we should run a multivariate regression to properly control for all other features and see if there is still a significant difference between the means of Q2 and Q4 price. Also, this is only 2 years of data and we may need longer data for this analysis.

Conclusion

A quick statistical data analysis suggests you should wait till the end of the year to buy your dream home (if you are not in a hurry)! There could be different reasons for this arbitrage opportunity. People relocate a lot in spring/summer when there are more job offers and they need to relocate to a new area and settle down quickly so they pay a liquidity premium. Also, people pay an education premium if they have kids going to schools. Another reason could be, although on average you pay less in Q4, those houses might be subject to different issues and were listed for a long time (maybe from Q2!). So really you are not getting deals you're just paying the fair price given the conditions of the home that are not captured here. Bummer!

Exploratory Analysis (Airbnb Data)

Exploratory Data Analysis

A histogram of all listed Airbnb prices in Santa Clara County as of 07/2019 as in Figure 8 suggests the distribution is normal but skewed to the right, similarly to the Redfin home prices. In other words, there are many extremely high prices and if you look at the summary statistics below, the max one-day Airbnb price is $10,000!

count	mean	std	min	25%	50%	75%	max
7169	$ 169	$ 292	$ 10	$ 65	$ 102	$ 195	$ 10,000

Figure 8.

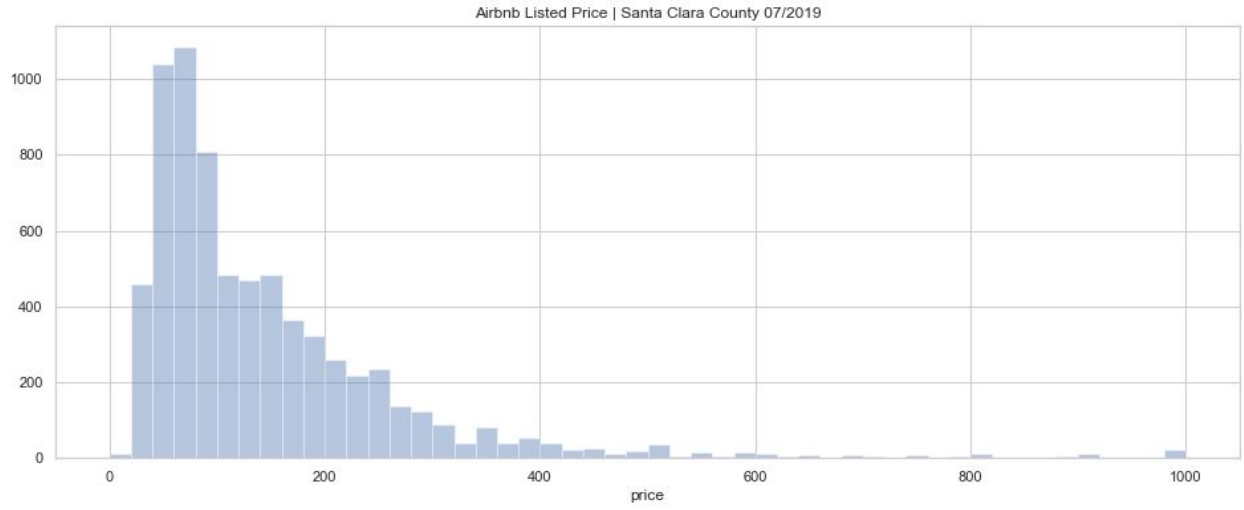


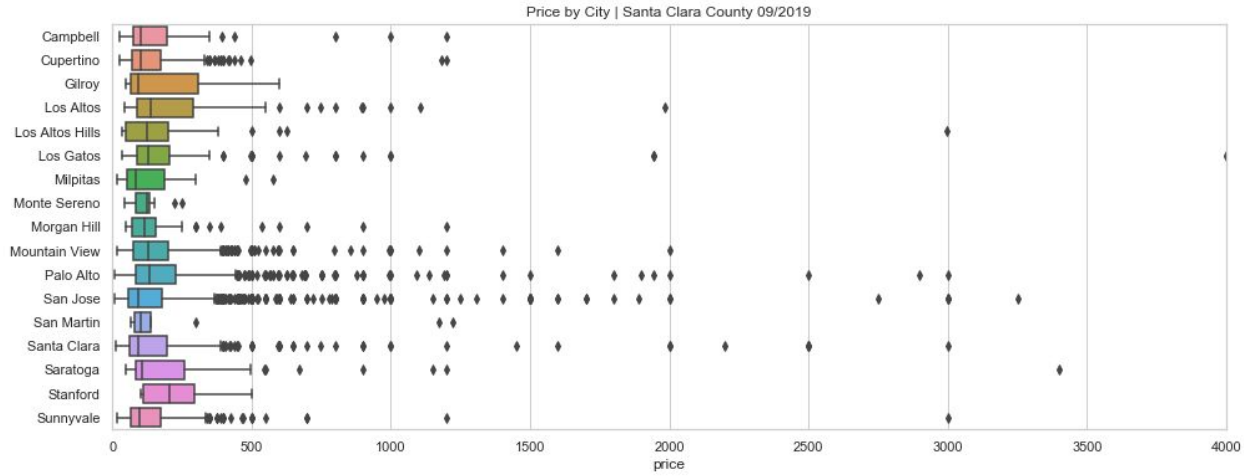
Figure 9 illustrates the distribution of listed prices by the city. Compared to the home price chart in Figure 2, one interesting case is Mountain View which is not necessarily the most expensive neighborhood in terms of home prices but charges one of the highest Airbnb prices. This might suggest that location is very important in Airbnb prices. Another reason could be the homes listed on Airbnb in the premier neighborhoods could be from the lower distributed homes in terms of home prices. In other words, millionaires don't want to rent their homes to strangers.

In-depth Analysis and Machine Learning

Goal

The goal of the in-depth analysis is, given the features of a potential property (such as number of beds/baths, latitude/longitude, etc.), to predict 1) the purchase price of the property and 2) the

Figure 9.



monthly rental income if listed on Airbnb and based on 1) and 2) to calculate the capitalization rate (ROI without considering mortgages).

Home Price Prediction

Training data and testing data were split from the original data with 70% and 30% respectively. Given the list of features: Beds, Baths, Square feet, Lot size, Year, HOA, Latitude, Longitude, Dummy variables for 1) the quarter of the sale 2) city 3) zip code 4) property type, Distances to different locations such as (Google, Apple, San Jose Downtown, Palo Alto Downtown, etc.), School, Crime, a variety of supervised learning models were tested.

Home prices with more than $3 million tend to be outliers and were therefore removed from the analysis. Given the goal is to find the deals for investment property, this cap is considered reasonable. There are 16141 observations in the training data and 6918 in the testing data with 110 features which includes dummy variables.

Linear Regression and LASSO

A simple linear regression model shows a quite decent performance with around 84% R^2. To prevent overfitting and reduce features, a LASSO model was tested with 8 variables eliminated and a final R^2 of 83%. However, as seen from Figure 10, the relationship between residuals and fitted values shows the heteroscedasticity of the errors which is undesirable.

Random Forest

A random forest model with n_estimators=100 and max_depth=20 was tested and the testing score improved to around 92% R^2. The RMSE is around 148000 and the median absolute error is around $70,000. Figure 11 shows the list of features ordered by importance with the most important ones including School, Square feet, Property type SFH and then some location features. This is in line with my hypothesis that schools and the size of the house are important when purchasing homes. Figures 12 and 13 show the errors are close to randomly distributed and there are very few outliers.

Figure 10.

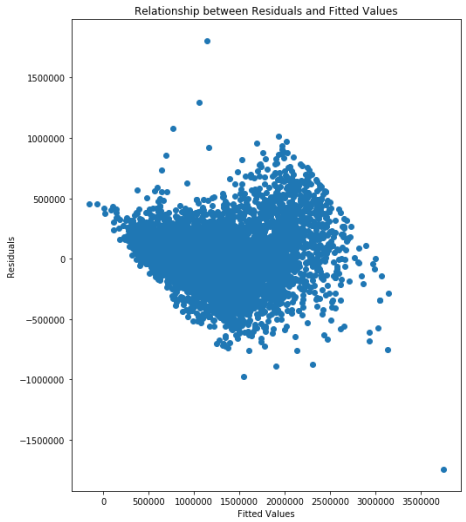
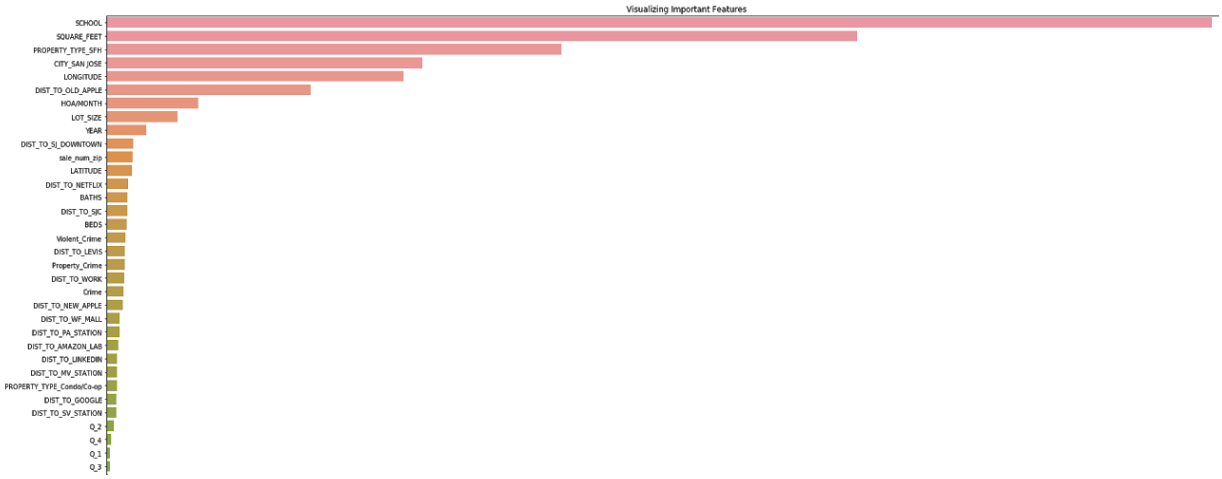


Figure 11.



A cross-validation with 5 folds to tune hyper-parameters such as n_estimators, max_features, max_depth, min_samples_split, and min_samples_leaf didn't improve the model significantly and will not be shown here.

Rental Income Prediction

Training data and testing data were split from the original data with 70% and 30% respectively. Given the list of features in the Airbnb data that overlapped with the Redfin data: Beds, Baths, Latitude, Longitude, Dummy variables for 1) city 2) zip code 3) property type, Distances to different locations such as (Google, Apple, San Jose Downtown, Palo Alto Downtown, etc.), School, Crime, a variety of supervised learning models were tested. Note that one more dummy variable is added which indicates if the entire home is listed or not. This is an existing feature in the Airbnb data but not in the Redfin data. But for prediction purposes, it can be turned on or off

Figure 12.

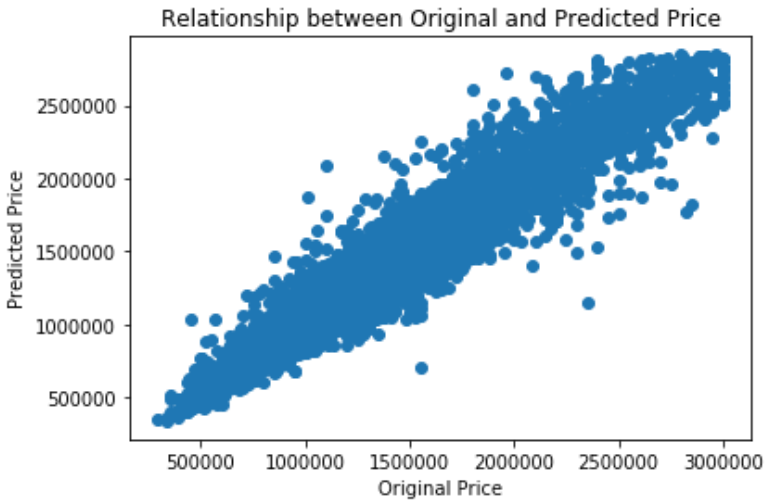
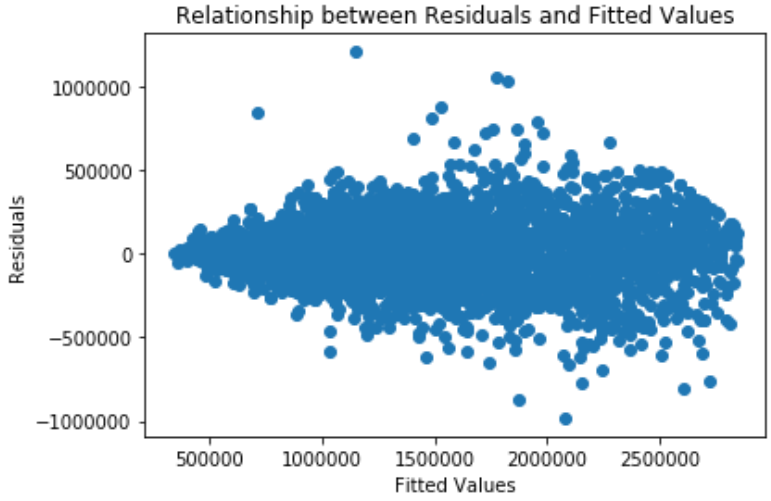


Figure 13.



depending on the property owner's investment goal. For simplicity, it was turned on here which means the whole property is rented. This can be personalized further if the property owner wants to rent out each room separately. Lastly, the target variable is simply the average price times the number of booked days in the next 30 days.

The goal is to find the "equilibrium" rental price, and so listings with extreme prices and/or with very few bookings and/or listings that are inactive or new were removed from the analysis. The filters include the vacant days in the next 30 days <=7, the number of reviews in the last 12 months >=5, and the average daily price <=$500. In the end, there are 1655 observations in the training data and 710 observations in the testing data.

Random Forest

Since linear regression and LASSO have similar issues as in the home price prediction, a random forest model is tested with a testing score around 76% R^2 and the median absolute error $565. The most important features include beds, rent_entire_home, and baths which seem reasonable (see Figure 14). Interestingly, school and crime are less important than location features. This is in line with the hypothesis that since renters mostly seek shorter-term residence they care less about the school and crime and more about the location of the place. Also, as seen in Figures 15 and 16, both the predicted vs actual target and residuals vs. fitted analysis are quite decent. Again, cross-validation didn't improve the performance much so it is not shown here.

Figure 14.

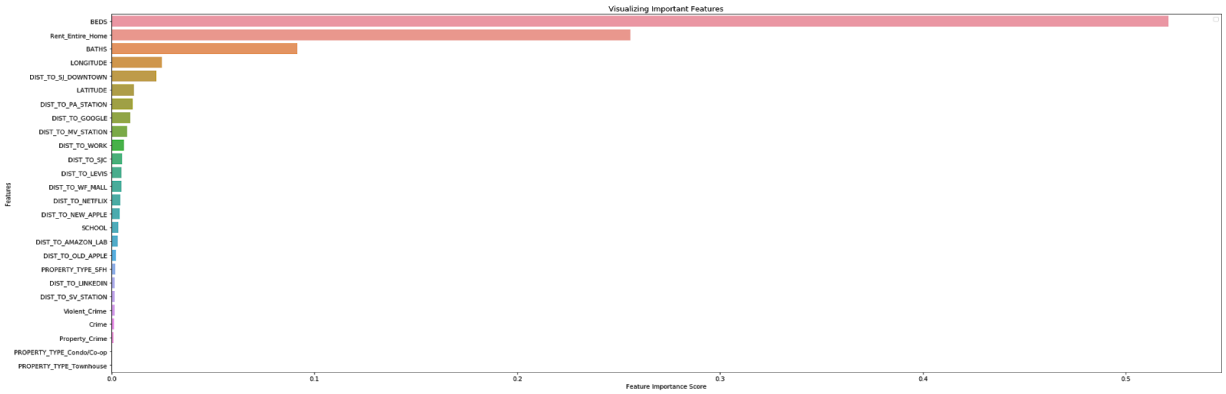


Figure 15.

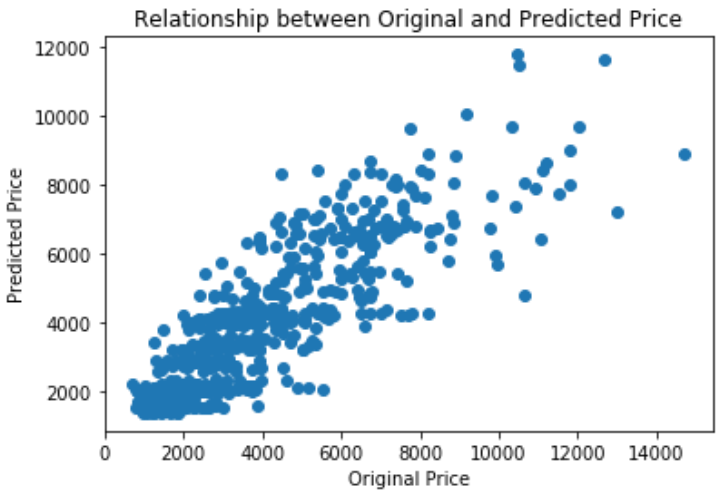
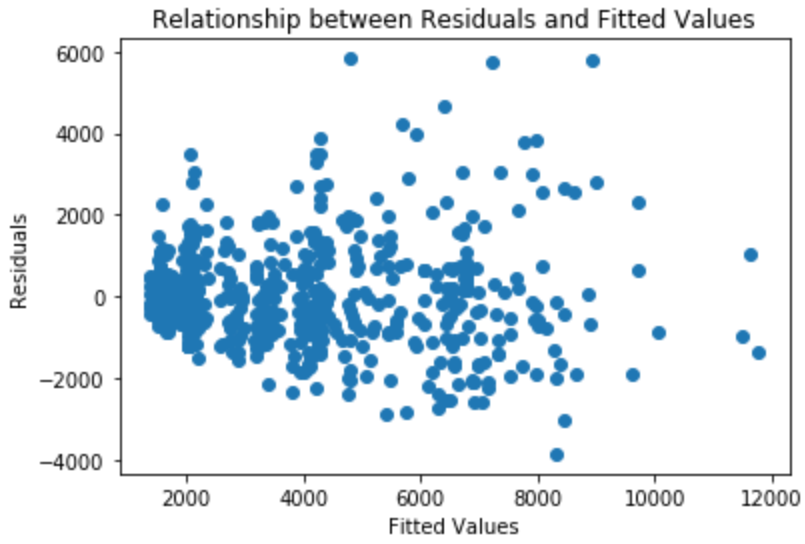


Figure 16.



Capitalization Rate

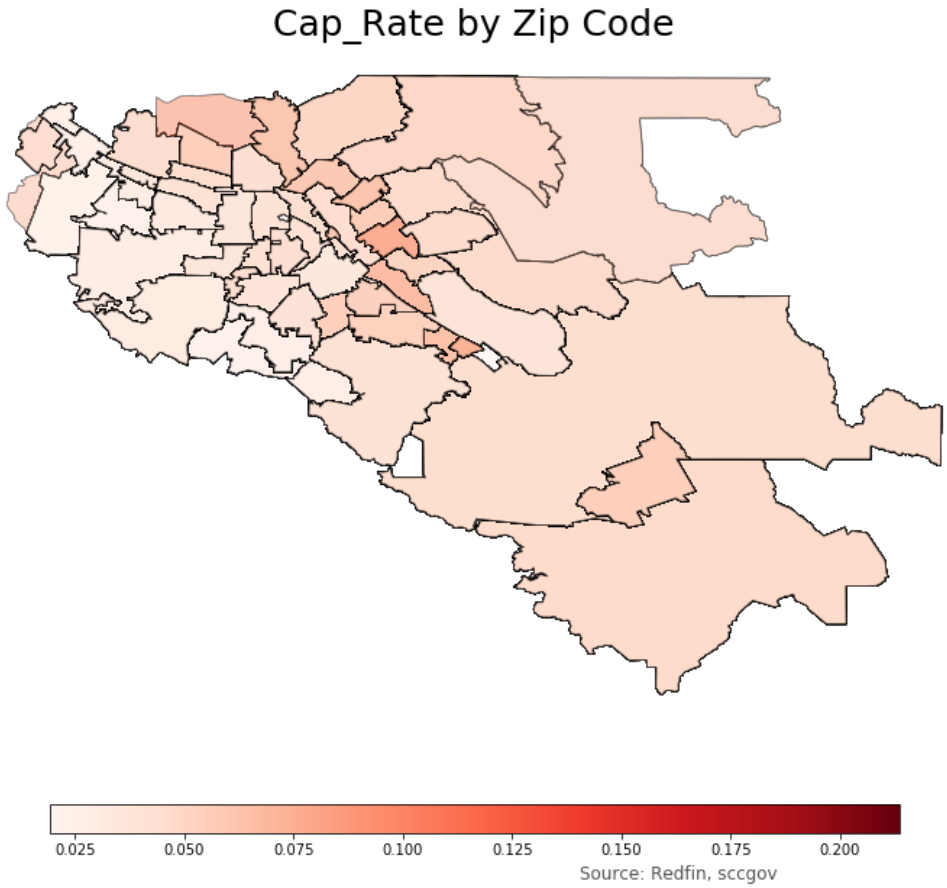
Now, given the two random forest models, one predicting the property price and the other predicting the monthly rental income generated by the property, I can run predictions given the features in the Redfin data and then calculate the capitalization rate as monthly income * 12/ property purchase price. Note that for simplicity costs such as mortgages, interests, maintenance, taxes, insurance, utilities, HOA, management, etc. are not included in the analysis and all-cash purchase is assumed here. The average predicted cap rate is 7.5% which is quite reasonable. Also, home price appreciation is not considered here which could be another project to explore.

Even with the simplified calculation, this model yields some very interesting and important results that can help in property investment. In summary, a better deal (high capitalization rate) can be found in neighborhoods with below-average schools but in a great location (close to work hubs) and the initial investment is lower (cheaper to buy). This can be found by sorting the cap rate by the city as seen in Table 1.

The 90th percentile of cap rates is 11.6%, about 4% higher than the median. Compounded for 10 years with an initial investment of $1 million, this means $900,000 more wealth which makes a big difference.

In Figure 17, the deeper the red, the higher the cap rate and the better the deal. Buying homes in nicer neighborhoods such as Cupertino, Los Altos and Saratoga might give you a higher dollar amount of monthly income but you also paid a (much) higher price for the property and thus they're the worst deals in terms of lower cap rate. Another interesting observation is Palo Alto: although the home price is quite high there it's also close to work and school so the cap rate is one of the highest in the nicer neighborhoods.

Figure 17.



To confirm this finding, a different data source is analyzed: Zillow Rental Index which has average rental income (mostly long-term rentals instead of short-term rentals in the Airbnb data). The data is on the zip code level and the income is average income so it's not as personalized as using the Redfin data but nonetheless still good data to confirm my story. As seen in Tables 2 and 3, sorted by cap rates, the results are very similar to the Redfin data which is comforting.

Caveat

The number of observations for Airbnb data is relatively small. Data from other counties and cities can be merged to get more observations.

To get a more realistic calculation of the capitalization rate, monthly operating costs can be estimated and added in the analysis. Plus, the flexibility of purchasing properties with loans can be added as well.

There are apparently some important missing features such as the home's pictures, home amenities, and finer school and crime data. For example, two homes can look exactly the same

Table 1.

city	Cap_Rate
GILROY	0.138016
MORGAN HILL	0.111843
ALVISO	0.097710
SAN MARTIN	0.090730
MILPITAS	0.084892
SAN JOSE	0.081702
SANTA CLARA	0.069048
CAMPBELL	0.066819
MOUNTAIN VIEW	0.063056
SUNNYVALE	0.058751
STANFORD	0.053086
LOS GATOS	0.052011
PALO ALTO	0.049308
CUPERTINO	0.048751
SARATOGA	0.045557
LOS ALTOS HILLS	0.044901
LOS ALTOS	0.043854
MONTE SERENO	0.039121

based on the current features but one is newly renovated and the other is not. The quality of the house can play an important role in both purchase price and rental price.

Conclusions

Although there are still rooms to improve, this model is a good starting point for people to invest in rental property. Give some basic features of a home, the model is able to predict the purchase price and rental income. Based on the calculated capitalization rate together with the monthly cash flows and initial investment amount, people can make a much more informed decision in purchasing an investment property.

Table 2.

City	Cap_Rate
Gilroy	0.044863
Stanford	0.042369
Morgan Hill	0.039581
San Jose	0.037822
Lexington Hills	0.036556
Milpitas	0.036497
San Martin	0.032600
Santa Clara	0.030857
Campbell	0.030089
Sunnyvale	0.029427
Los Gatos	0.026827
Saratoga	0.024152
Mountain View	0.023098
Cupertino	0.020802
Los Altos	0.019523
Palo Alto	0.015794

Table 3.

City	cap_rate_zillow3beds
San Jose	0.040034
Milpitas	0.039842
Mountain View	0.038844
Campbell	0.036671
Santa Clara	0.033299
Cupertino	0.028229