Calanit Kamala

December 2021

CAPSTONE III FINAL REPORT

THE IMPACT OF ON-LINE HOME LISTINGS ON REAL-ESTATE MARKET OUTCOMES

Over the last decade, the use of on-line platforms while selling a home, has been established as the norm. With this change, the involvement of photos, videos and drone-footage, has increased significantly. This is especially true over the last 2 days, since the spread of COVID-19 has constrained in-person home visitation. As a result, the information provided on on-line home profiles has increased in importance.

Real-estate agents spend considerable amount of time and resources on creating enticing on-line profiles that would attract buyers and offers. However, the variation of investment on on-line home profiles is significantly high. Some agents provide little to no information, some provide excruciating level of details. Some provide one or two photos, while other provide dozens. This variation provides an opportunity to analyze the relationship between market outcomes and profile features. The objective of this analysis is to learn whether different features of on-line home profiles features are associated with final home sale price and length of sale.

Data Collection & Date Cleaning

The MLS (Multiple Listing Service) is a regional data base of homes for sale, only accessible to licensed realtors with the purpose of sharing information on houses for sale. On-line web-sites that provide listing information such as Zillow, Redfin etc, populate home information from data they get from the MLS. The MLS data base relies on realtors to input home information, a fact that makes the data myriad with errors and inconsistencies. Some data fields are required, while others are not. Some data fields are not clearly defined and result in different information being entered. This added a significant layer of complexity to the data cleaning process. Nonetheless, the majority of listings have accurate pricing information as well as detailed home information. In addition, MLS data includes comments made by realtor – both public and private (only available to other realtors), as well as the number of photos uploaded and whether the listing had an additional media link. The addition of a media link normally indicates that a realtor has provided a video description of the home on a separate link.

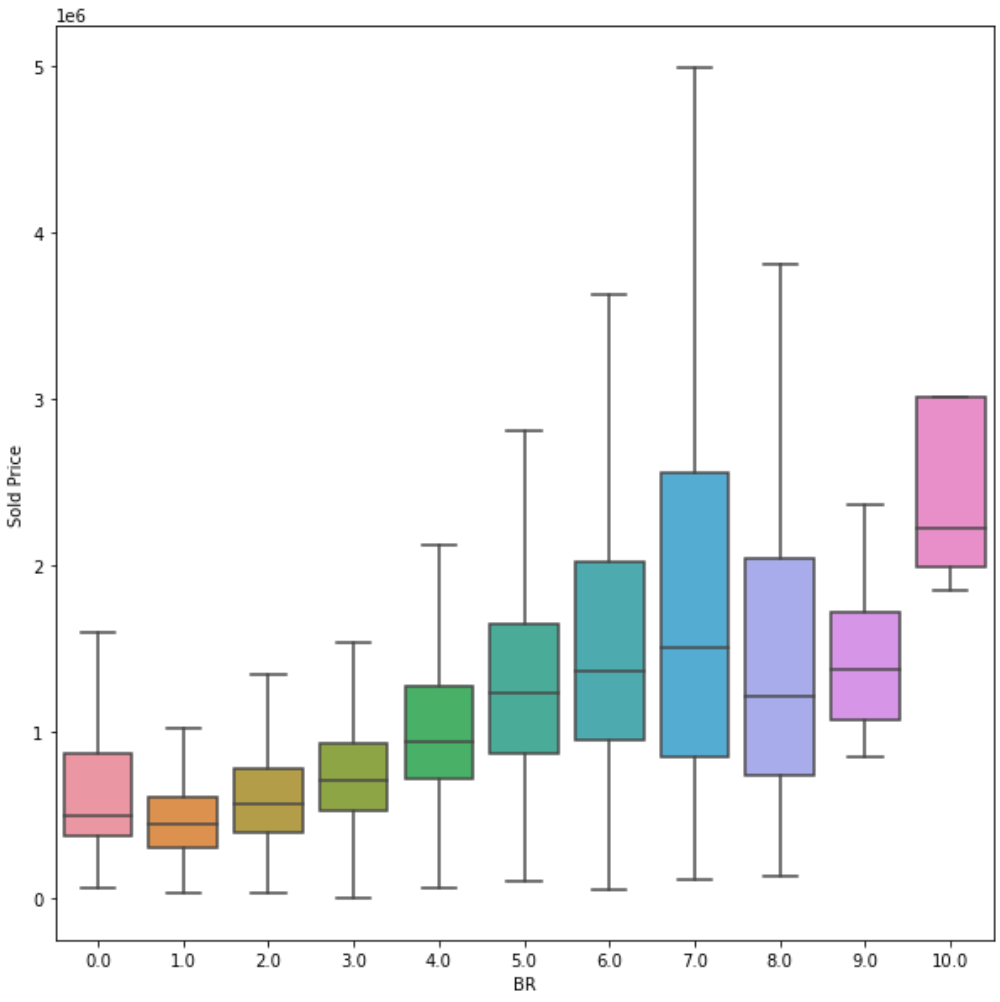
The data for the analysis was collected from the MLS and include detached, single-family homes sold in Alameda county, CA, starting 01/01/2010 up to October 2021. Data excludes sales of bankrupt or short-sale properties.

After removing duplicate observations, null observations and limiting the sample to homes that were on the market for less than 400 days as well as homes with 10 or less bedrooms, 115740 observations remained.

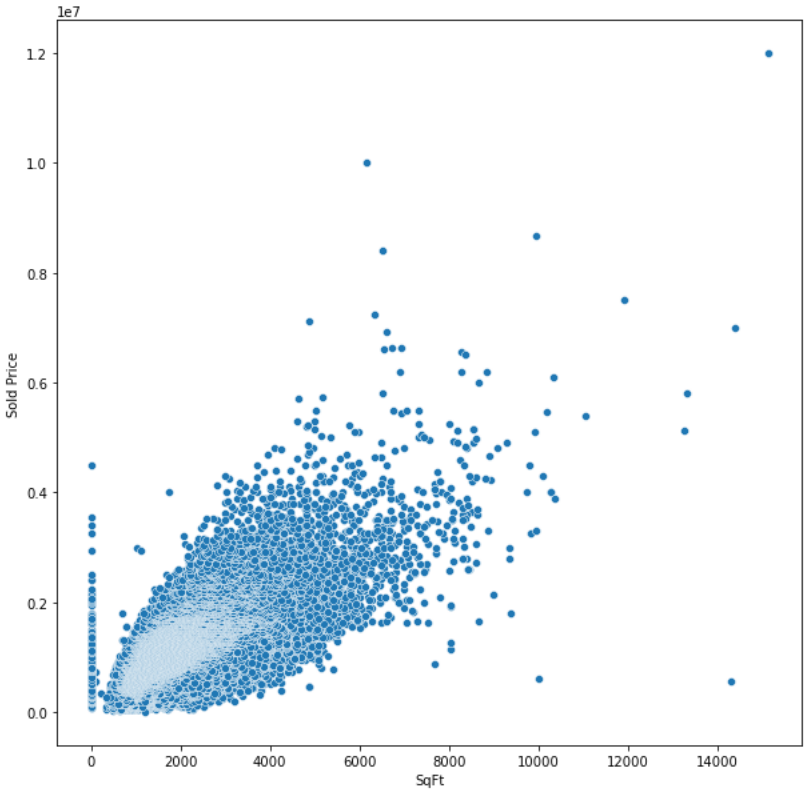
Exploratory Data Analysis

The purpose of EDA was to confirm the data follows expected relationships, as well as to ensure that the data cleaning process was complete (for example, no negative observations or 32 bathrooms in a single property). Relationship between home price and the number of bedrooms (plot #1)/home area (plot #2) were as expected. A time-series of the data was plotted and two trends stand out: home sales are very seasonal and are rising with time, in both volume (plot #3) and prices (plot #4)

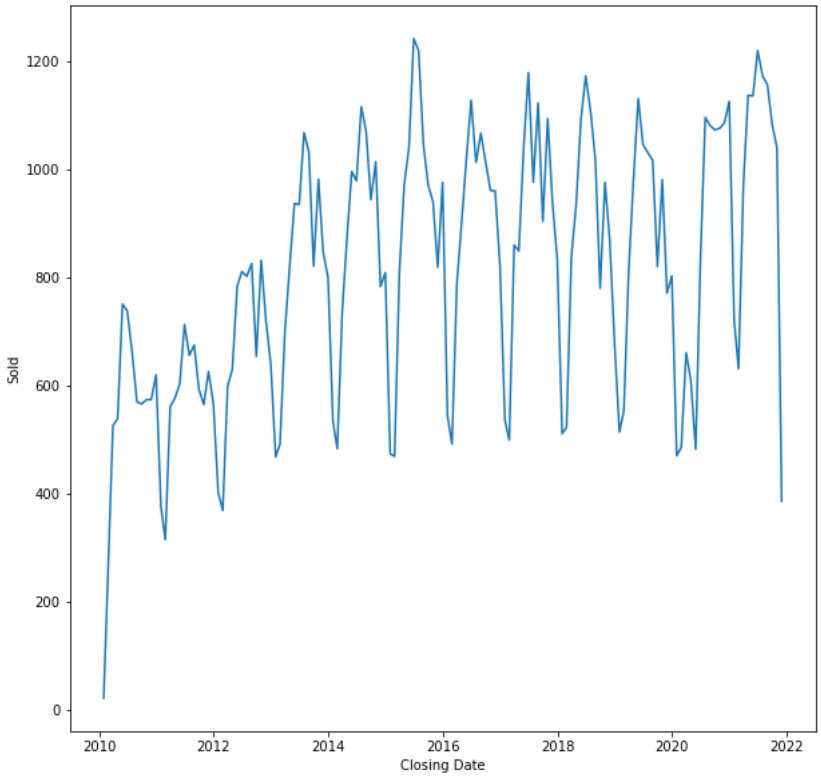
Plot 1. Boxplot of final home price by number of bedrooms



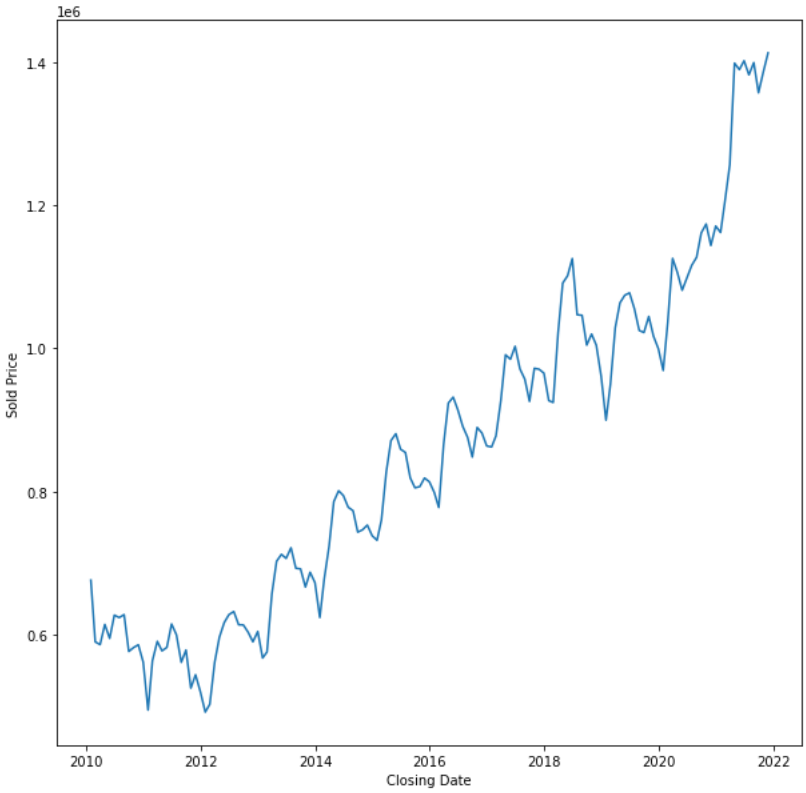
Plot 2. Scatterplot of home area in sqft and home sale price



Plot 3. Time series of the monthly number of homes sold

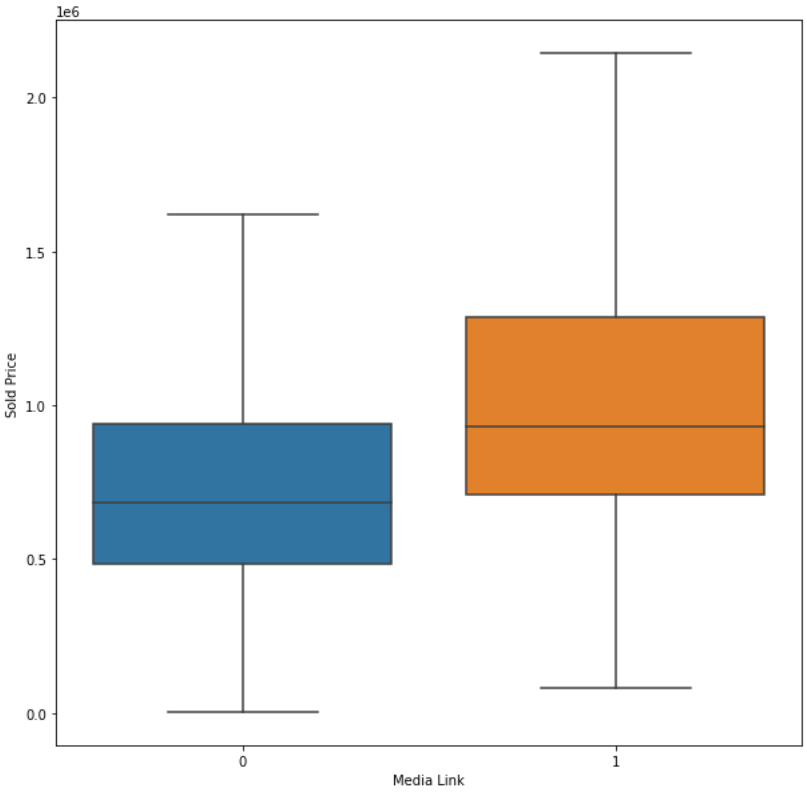


Plot 4. Time series of the monthly average price of homes sold

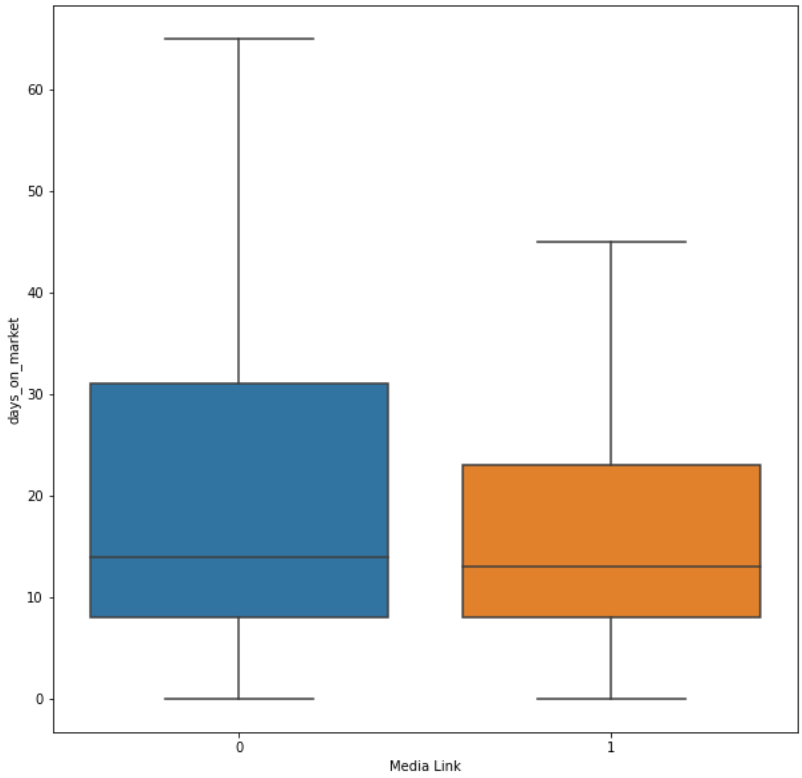


Explanatory data analysis also confirmed visually that there is an association between home prices and the presence of a media link – homes with a media link tend to sell for higher prices (plot #5) and faster (plot #6)

Plot 5. Boxplot of home sale prices by presence of a media link



Plot6. Boxplot of number of days on market presence of a media link

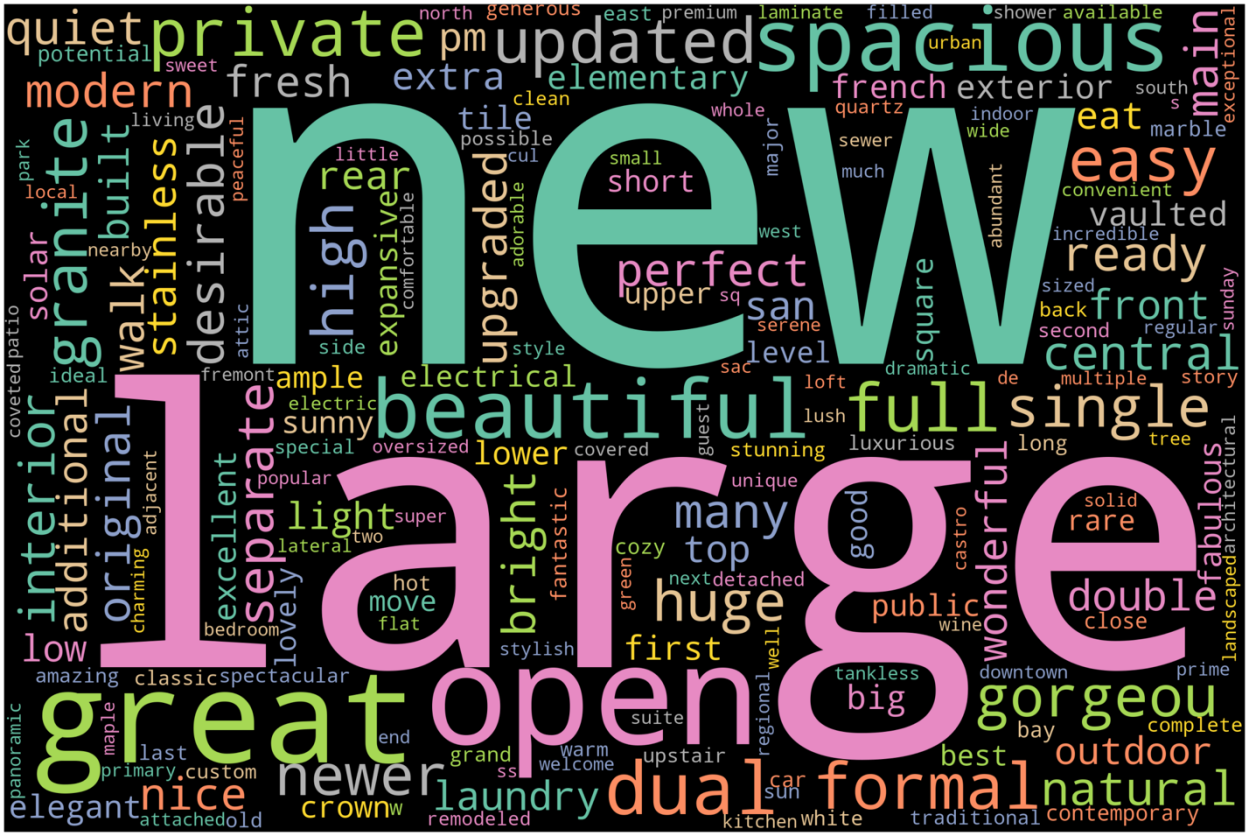


Data Pre-processing

Since a linear regression model is being employed, I did not scale the data, in order to preserve the interpretability of the coefficients.

The total number of words and the total number of adjectives were extracted from public and private realtor remarks, in order to be included as numerical features.

Below is a word cloud created from adjectives used in all public remarks. In order to create it, I combined all remarks associated with each home listing to a single text. The word cloud indicates the frequency of the most used adjectives.



Modeling

Modeling home sale prices based on home features is a fairly common analysis and many examples of such analysis are available on different on-line sources. This analysis includes the following home features: home age, number of bedrooms, number of bathrooms, home area, lot area, whether the house has a garage and how the house was sold (mortgage, cash or government loan). After the baseline model was evaluated, the coefficient on number of bedrooms was negative, which is unlikely to accurately reflect reality. In order to address this issue, I included a non-linear term (number of bedroom squared), which immediately addressed the issue. Having a non-linear relationship between price and number of bedrooms is logical – at first home price is increasing with the number of bedrooms, but if the number of bedrooms is increasing beyond a certain point, it may have a detrimental impact on price.

In order to address the seasonality of the data, both annual and monthly fixes-effects were added to the analysis.

Lastly, there are the features of interest – the features that describe the characteristics of the home's on-line listing. Those features include: number of photos included, the availability of a media link, number of words in the public and private remark section, number of adjectives in the public and private remark section.

The baseline linear regression was estimated using a statmodels OLS. After estimating a simple linear regression in statmodel, a few linear regressors algorithms were tested: Ridge, Lasso and elastic net regression, to test for the effect of regularization as well as random forest, K-neighbors and gradient boosting regressors.

Among all models, Random Forest regressor resulted in highest R^2 and lowest RMSE. A Grid Search Crosss validation algorithm was used for hyperparameters tuning. After hyperparameter tuning, the model's R^2 increased to 0.82 and RMSE decreased further.

Results

The first regression estimated included home final sale price as dependent variable and included all features. According to the baseline OLS regression, only three featuers are not statistically significant (at 95%): lot area, the number of adjectives used in private remarks and how the house sold – other category. All other coefficients are statistically significant at 99%. All coefficients have expected signs. The baseline R^2 is 0.7. The baseline regression reveals a few interesting trends:

- The real estate market has recovered from 2008 real-estate crisis by 2013. Since then market prices consistently increased.
- May, April and June are the best months to sell a house, while February is the worst month.

Investigating the relationship between listing features and home prices, a few insights emerge:

1. The presence of a media link is associated with an increase in final house price of $60,000.
2. Longer public remarks with more adjectives seem to be ***positively*** associated with home prices, whereas longer *private* remarks are ***negatively*** associated with home prices.
3. Number of pictures is ***positively*** associated with home prices.

The second regression estimated includes the number of days on market as the dependent variable, with the same set of features as home price regression. Here are a few of the insights provided by the analysis:

- Similar to trends observed for home prices, on April, May and June homes sell the fastest, while in February market times are the longest.
- Homes with more bedrooms and a garage sell significantly faster than other homes.
- A listing that includes a media link is associated with a faster sale with days on market being 3 days shorter than listings without media link.
- While the number of words in both public and private remarks is associated with slightly longer market times, number of adjectives in both public and private is negatively associated with market times.
- Every 10 photos are associated with a shorter market presence by 1.5 days.

The simple OLS model indicated that all included features are important for the analysis. After searching for the best regressor estimator and preforming hyper parameter tuning, a random forest regressor was the model with highest explanatory power. Looking at the top 10 most important features, we can see that the number of pictures, number of words and adjectives in public remarks are all in the top 10 features.

Feature	**Feature Importance**
SqFt	0.217699
Bth	0.113149
Picture_Count	0.087890

Lot_SqFt	0.074253
word_count_public	0.070198
public_adj_num	0.062552
BR	0.061483
Age	0.050221
Year_sold_2021	0.047762
BR_squared	0.047658

Summary

The data indicates that on-line feature listing are associated with higher home prices and faster sale times. Looking at home prices, listing remarks and adjectives, as well as picture count and the availability of a media link, are all positively associated with higher home prices. In addition, longer private remarks are associated with lower prices.

Even through this association is strong, correlation does not imply causation. For example, it is possible that the owners of more expensive homes hire realtors who invest more effort on on-line advertising, explaining the strong association between home listing and market outcomes. Inferring whether listing features provide a signal for a more effective realtor or drive market outcomes directly, requires a more rigorous statistical approach, such as instrumental variables or a diff-in-diff.